ISSN: 2663-3264

مجلة اتحاد الاحصائيين العرب

مجلة علمية محكمة

Journal of Arab Statisticians Union (JASU)

رئيس التحرير
الدكتور حيدر محمد فريحات

مدير التحرير
الاستاذ الدكتور جاسم محمد التميمي

الناشر
جامعة ميسان بالتعاون مع اتحاد الاحصائيين العرب

المجلد السابع / العدد الأول لسنة ٢٠٢٤

رقم الايداع في دار الكتب والوثائق الوطنية ببغداد(٢٤٣٨) لسنة ٢٠٢٠م

اتحاد الاحصائيين العرب ، مبنى دائرة الاحصاءات العامة ، عمان المملكة الاردنية الهاشمية
Union of Arab Statisticians , Building , Amman , Jordan
P.O. Box 2892Amman 11941 Jordan, Tel: +962 6 5300700 ,
Fax : +962 6 5300710 Emails : info@uarabs.org ,
secgen@uarabs.org , Website: www.uarabs.org

هيئة تحرير المجلة العلمية لاتحاد الاحصائيين العرب

1	أ.د. حيدر فريحات - مدير عام دائرة الاحصاءات الاردنية - الاردن	رئيسا
2	أ.د. غازي ابراهيم رحو – الامين العام لاتحاد الاحصائيين العرب	نائبا للرئيس
3	أ.د. جاسم محمد التميمي – جامعة ديالى – العراق	مدير التحرير
4	أ.د. مهدي العلاق – رئيس الجمعية العراقية للعلوم الاحصائية - العراق	عضوا
5	أ.د. علا فرح عوض – رئيس الجهاز المركزي للإحصاء الفلسطيني - دولة فلسطين	عضوا
6	أ.د.عادل مانع داخل - مساعد رئيس جامعة ميسان للشؤون العلمية والدراسات العليا - العراق	عضوا
7	أ.د. مزهر شعبان العاني – الامين العام المساعد لاتحاد الاحصائيين العرب	عضوا
8	أ.د. عوض حاج علي – جامعة النيلين - جمهورية السودان	عضوا
9	أ.د. حسين عبد العزيز السيد – جمهورية مصر العربية	عضوا
10	ا.د. اسماعيل بن قانة – الجزائر – يعمل في جامعة الملك فيصل	عضوا
11	أ.د. محمد ابو صالح – جامعة عمان العربية – الاردن	عضوا
12	د. سالم محمد القزيزي – دولة ليبيا	عضوا
13	د. وصفي طاهر صالح قهوةچي - جامعة تيشك الدولية - العراق	عضوا
14	د. عبد اللطيف فراخ – المغرب	عضوا

هيئة مستشاروا المجلة العلمية لاتحاد الاحصائيين العرب

١.	أ.د. مزهر شعبان العاني - الأمين العام المساعد لاتحاد الإحصائيين العرب	رئيسا
٢.	أ.د. محمود أبو شعير - جمهورية العراق عميد كلية الرافدين الجامعة	عضوا
٣.	أ.د. أحسن الطيار - جامعة ٢٠ أوت ١٩٥٥ - سكيكدة – سكيكدة – الجزائر	عضوا
٤.	أ.د.عبد الوهاب محمد جواد الموسوي - جامعة الكوفة - جمهورية العراق	عضوا
٥.	أ.د. مناف يوسف حمود - جمهورية العراق	عضوا
٦.	أ.د. أحلام أحمد جمعة - كلية الآداب - جامعة بغداد - جمهورية العراق	عضوا
٧.	د. يحيى بن خميس الحسيني - مركز مسقط للاستشارات الإحصائية - سلطنة عمان	عضوا
٨.	د. علاء عبد السلام مصطفى - جامعة ميسان - جمهورية العراق	عضوا
٩.	د. نبيل جورج ناسي - جامعة صلاح الدين - اربيل - العراق	عضوا
١٠	د.محمد خليل إبراهيم - عميد كلية الإدارة والاقتصاد - جامعة ميسان - جمهورية العراق	عضوا
١١	د.عامر المقدسي رئيس المركز الدولي للتنمية الإدارية (IDS - USA)	عضوا
١٢	د. علي خالد عبدا لله- جامعة ميسان	عضوا
١٣	د. ولاء جودت الجاف – جمهورية العراق	عضوا
١٤	د. سلوى محمد نجرس علاوي - وزارة التعليم العالي والبحث العلمي – جمهورية العراق	عضوا
١٥	د. زرزار العياشي - الجمهورية الجزائرية الديمقراطية الشعبية	عضوا
١٦	د. مكين الدين أحمد عبد الله - جمهورية السودان	عضوا
١٧	د. محمود الصروي - الجهاز المركزي للتعبئة والإحصاء المصري - مصر	عضوا
١٨	السيد أيوي أيوب – دولة فلسطين	عضوا

شروط النشر

١– تنشر المجلة البحوث والدراسات العلمية في المجالات الإحصائية والمعلوماتية المكتوبة باللغة العربية أوالانكليزية أو الفرنسية على أن لا يكون البحث المقدم للنشر قد نشر او قدم للنشر في مجلات او دوريات أخرى او قدم ونشر في دوريات لمؤتمرات أو ندوات.

٢- ترسل نسخة إلكترونية من البحوث والدراسات الى أمين و سكرتير التحرير على أن تتضمن اسم الباحث او الباحثين وألقابهم العلمية وأماكن عملهم مع ذكر عنوان المراسلة وأرقام الهواتف والفاكس والبريد الالكتروني علما أن أجور النشر اي بحث يكلف مبلغ (١٠٠) دولار. او ما يعادله بالدينار العراقي.

٣- يرسل البحث المراد نشره بالمجلة مطبوعًا الكترونيا على أما على دسك او بالايميل على عنوان سكرتارية التحرير وفق المواصفات أدناه:

أ. باللغة العربية باستخدام حرف نوع (Simplified Arabic)

ب. الانكليزية (Times New Roman) وبخط حجم (14) وباستخدام نظام Microsoft Word

ج. أن يتم ترك مسافة 2.5 سم لكافة أبعاد الصفحة.

د. يرفق الباحث ملخصًا عن بحثه باللغة الانكليزية أو العربية أو الفرنسية حسب لغة البحث وبما لايزيد عن صفحة واحدة .

ه. يتم الإشارة الى المصادر العلمية في متن البحث وفي نهايته ، مع مراعاة أن لا تتضمن الصفحة الاخيرة سوى المصادر التي تم الإشارة إليها في المتن ووفق الاصول المعتمدة في ذلك (اسم المؤلف ، سنة النشر ، عنوان المصدر ، دار النشر ،البلد)

و.ترقم الجداول والرسوم التوضيحية وغيرها حسب ورودها في البحث ، كما توثق المستعارة منها بالمصادر الأصلية.

ز.أن لايزيد عدد صفحات البحث اوالدراسة عن (20) صفحة A4. وضرورة إرسال النسخة الالكترونية على عنوان سكرتارية التحرير.

ح. أن يكون البحث مرتب على شكل عمودين في الصفحة.

٤- سيتم إشعار الباحث باستلام بحثه خلال مدة لا تتجاوز اسبوعين من تاريخ الاستلام.

٥- تخضع كل البحوث المرسلة للنشر في المجلة للتقويم العلمي و الموضوعي ويبلغ الباحث بنتيجة التقويم والتعديلات المقترحة إن وجدت خلال مدة لا تتجاوز اسبوعين من تاريخ استلام الردود من كل المقيمين.

٦- لهيئة تحرير ا لمجلة الحق في قبول او رفض البحث ولها الحق في إجراء أي تعديل او إعادة صياغة جزئية للمواد المقدمة للنشر بما يتماشى و النسق المعتمد في النشر لديها بعد موافقة الباحث.

٧- يصبح البحث المنشور ملكًا للمجلة ولا يجوز إعادة نشره في أماكن أخرى.

٨- تعبر المواد المنشورة بالمجلة عن آراء أصحابها ، ولا تعكس بالضرورة وجهة نظر المجلة او اتحاد الاحصائيين العرب.

٩- ترسل البحوث على العناوين أدناه:

العنوان البريدي للاتحاد : ص.ب. ٨٥١١٠٤ عمان ١١١٨٥ الأردن.

البريد الالكتروني : uarabs@gmail.com

أو مدير التحرير : jasimtimimi@yahoo.com

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

كلمـــة العـــدد

انه من دواعي السرور ان تتواصل مجلة اتحاد الإحصائيين العرب بإصدار أعدادها بشكل مستمر وبالتعاون مع جامعة ميسان وكما هو مخطط لها وقد بذل الاتحاد والأمانة العامة وهيئة التحرير والمختصين جهودا كبيرا لإظهار هذه المجلة من اجل ان تكون نبراسا للباحثين وطريقا لإيصال بحوثهم الى اكبر عدد من الاختصاصيين حيث يتم توزيع هذه المجلة على الأجهزة الإحصائية والمؤسسات المنضوية ضمن الاتحاد بالإضافة الى أعضاء الاتحاد ومن خلال دار النشر التي تقوم بعرضها من خلال المعارض والمؤتمرات العلمية . هيئة التحرير والهيئة الاستشارية يقدمون شكرهم وتقديرهم الى الأستاذ الدكتور عادل مانع داخل - مساعد رئيس جامعة ميسان للشؤون العلمية والدراسات العليا - العراق لجهوده المثمرة ومتابعته المستمرة في دعم التعاون بين جامعة ميسان ومجلة اتحاد الإحصائيين العرب مع التقدير و الاحترام.

نتقدم لكم جميعا بالتهنئة ونبارك لكم صدور المجلد السابع / العـــدد الأول لـــسنة ٢٠٢٣ من مجلتنا الغراء JASU وهي تصدر بحلتها الجديدة وقد حصـــلت المجلة على الـترقيم الدولـــي (ISSN 2663-3264) ومما يزيد من رصانة المجلة العلمية وتحتوي المجلة على بحوث محكمة وحديثة وتعالج مشاكل احصائية جديدة وتقدم افكار بحثية متطورة نتمنى ان يكون العدد القادم غزيرا بالبحوث القيمة وان يصدر في موعده المحدد بفضل دعمكم ومشاركتكم الفعالة مدعما بالبحوث المتطورة وخاصة ان الاحصاء الان يعتبر العنصر المهم والاساسي بتداخله مع كافة العلوم مما يفسح المجال لتطبيقات عديدة ومساهمات متنوعة .

ولكم الموفقية والنجاح للجميع شاكرين كل من وضع لبنة في بناء ونشر هذا العدد من المجلة وتقبلوا وافر التقدير والاحترام

مدير التحرير
الاستاذ الدكتور جاسم محمد التميمي

الرؤية والرسالة والاهداف

الرؤية Vision

الريادة في نشر البحوث الاحصائية والسعي للوصول لتصنيف عالمي متقدم بين المجلات العلمية المحكمة ، وان تكون مجلة اتحاد الاحصائيين العرب نبراسا للعلم والمعرفة وواجهة علمية وثقافية واكاديمية مشرقة لاتحادنا الموقر ومركز علم خلاق يجمع بين الاصالة والحداثة.

الرسالة Mission

اثراء العلوم الاحصائية بأجود انواع البحوث والدراسات الاحصائية التي تربط بين الاصالة والحداثة ضمن اطار علمي بناء باستثارة همم الباحثين وتنمية قدراتهم في النشر العلمي الاصيل باللغتين العربية والانكليزية وبما يسهم في ايصال الاحصاء بكل انواعه لكل شعوب العالم واتاحة الفرصة للباحثين لتقديم الصورة الحقيقية الناصعة للمجتمع .

الاهداف Goals

تسعى مجلة اتحاد الاحصائيين العرب الى تحقيق الاهداف الاتية :

١- تشجيع البحوث والدراسات الاحصائية التي تربط الاصالة بالحداثة وصولا الى تنمية الاعتزاز بماضينا الجميل والاختيار الواعي لما في الحداثة من توجيهات تنفع الجيل الجديد.

٢- تنشيط البحث العلمي التخصصي في الاحصاء وتكنولوجيا المعلومات.

٣- تنمية الوعي الاحصائي والبحث العلمي لدى الجيل الجديد من خلال استعراض البحوث الاحصائية وتكنولوجيا المعلومات التي تساهم في انماء روح الاحترام للأصالة .

٤- التواصل العلمي والبحثي الهادف مع المراكز العلمية والعلماء والباحثين للإبراز دور الاحصاء في رفد المجتمع وتطويره.

٥- المساهمة في حل بعض الاشكاليات الاحصاء والمجتمع من خلال البحوث الاحصائية التي تساهم في مساعدة المجتمع لحل اشكالياته ورفد المؤسسات الاحصائية بالبحوث العلمية الرصينة.

محتويات العدد

التنبؤ بالرقم القياسي ومعدل التضخم لإقليم كوردستان العراق بإستخدام نموذج حركي للشبكات العصبية مع السلاسل الزمنية

أ. د. طه حسين علي taha.ali@su.edu.krd
قسم الإحصاء والمعلوماتية / كلية الإدارة والاقتصاد – جامعة صلاح الدين/ أربيل
أ.م. د. نظيرة صديق كريم nazeera.kareem@su.edu.krd
قسم الإحصاء والمعلوماتية/ كلية الإدارة والاقتصاد– جامعة صلاح الدين/ أربيل
أ.م. د. هيام عبد المجيد الحياوي heyamhayawi@gmail.com
قسم الإحصاء والمعلوماتية/ كلية علوم الحاسوب والرياضيات – جامعة الموصل
الباحثة شهلة هاني علي العبيدي Shahla.hani@krso.gov.krd
هيئة إحصاء إقليم كوردستان العراق

الملخص

تمَ في هذا البحث التنبؤ بالرقم القياسي العام والأقسام الرئيسية لأقليم كوردستان العراق والذي من خلاله يتم حساب معدل التضخم العام السنوي والأقسام الأساسية من خلال إستخدام النماذج الحركية للشبكات العصبية (مرشحات غير خطية) مع السلاسل الزمنية في تكوين نماذج خطية للتنبؤ بالفترة المستقبلية (2023–2025) إعتماداً على بيانات مؤخوذة من هيئة إحصاء إقليم كوردستان للفترة الزمنية (2008–2022) وذلك بإستخدام لغة ماتلاب، وتوصل البحث إلى إمكانية إستخدام تلك النماذج في التنبؤ بالأرقام القياسية التي تتضمن تقلبات كبيرة في مقاديرها وهنالك إرتفاع في مستوى الرقم القياسي العام وبعض الأقسام الرئيسية وإنخفاض في بعضها الآخر للسنوات المتنبأ بها والذي أدى إلى إرتفاع معدل التضخم العام السنوي وإرتفاع في بعض الأقسام الأساسية وإنخفاض في بعضها الآخر.

Predicting the Consumer price index and inflation average for the Kurdistan Region of Iraq using a dynamic model of neural networks with time series

Nazeera Sedeek Kareem
Department of Statistics and Informatics/College of Administration and Economics-Salahaddin University
Taha Hussein Ali
Department of Statistics and Informatics/College of Administration and Economics-Salahaddin University
Heyam A.A.Hayawi
Department of Statistics and Informatics/College of Computer and Mathematical Science- Mosul University
Shahla Hani Ali
Kurdistan Region Statistics Office / Iraq

Abstract

In this research, the Consumer price index and the main sections of the Kurdistan Region of Iraq were predicted, in which the annual inflation rate and the main sections are calculated by using dynamic models of neural networks (non-linear filters) with time series in the formation of linear models to predict the future interval (2023-2025). Based on data from the KRSO for the interval (2008-2022) using MATLAB language, the research found that these models can be used to predict consumer price indexes that include large fluctuations in their amounts. There is an increase in the level of the consumer price and some of the main sections and a decrease in some of the other predicted, which led to an increase in the annual inflation rate a rise in some basic sections and a decrease in others.

1: المقدمة:

بيانات السلسلة الزمنية هي سلسلة من قيم البيانات المقدمة خلال فترة زمنية. تعتبر تعداد البقع الشمسية وبيانات المد والجزر وحركة أسعار الأسهم أو بيانات حركة المؤشر ومعظم البيانات الفيزيائية والكيميائية والطبية بيانات سلسلة زمنية. إن التنبؤ للسلسلة الزمنية هو عملية التنبؤ بالقيم المستقبلية على معلمة تستند إلى القيم (المشاهدات) السابقة للأنظمة الفيزيائية أو الكيميائية أو المالية أو الطبية من خلال بناء نموذج تنبؤي حركي. تصبح عملية التنبؤ مشكلة معقدة بسبب السلوك غير الخطي والحركي للمعلمات الموجودة في النظام المادي. في مشكلة السلاسل الزمنية، يتم التنبؤ بالقيم المستقبلية للسلسلة الزمنية y(t) بناءً على القيم السابقة. هذا التنبؤ يدعى الانحدار غير الخطي (nonlinear autoregressive) أو NAR، ويمكن صياغتها على النحو التالي:

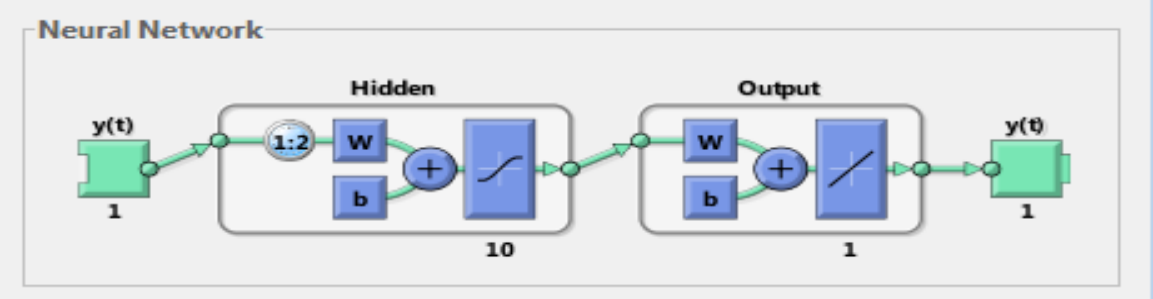
$$y_t = f(y_{t-1}, y_{t-2}, \cdots, y_{t-d}) \quad \cdots \quad (1)$$

من جانب آخر بدأ استخدام الشبكات العصبية في الاقتصاد عند بداية التسعينات وكان هناك اهتمام كبير بتطبيقات الشبكات العصبية في علوم الاقتصاد، وخاصة في مجالات الإحصاءات المالية وأسعار الصرف في المقابل، طبقت دراسات قليلة نسبياً أساليب الشبكة العصبية على العملات المشفرة، مثل (Bit coin). قدم الباحثان (Kuan and H. White) عام (1994) التعريف النهائي للشبكات العصبية لعلم الإقتصاد القياسي، حيث خطط الباحثون العديد من أوجه الشبه بين الاقتصاد القياسي والشبكات العصبية. وقد تمت متابعة المساهمة النظرية ببعض الأعمال التطبيقية من قبل الباحثون (Maaoumi, et al. 1994) الذين استخدموا (14) سلسلة زمنية إقتصادية مع الشبكات العصبية وبرهنوا جودتها. الباحث (Swanson and H. White, 1997) قدمَ محاولة رئيسة أخرى لإستخدام الشبكات العصبية للتنبؤ بمتغيرات الإقتصاد الكلي. الباحثان (Catania and Grassi, 2018) بحثا سلوك السلاسل الزمنية للعملات المالية المشفرة والتي تعتبر (Bit coin) من أبرز الأمثلة عليها. حركية تلك السلسلة معقدة للغاية تعرض مشاهدات متطرفة، عدم تناسق، [2] والعديد من الخصائص غير الخطية التي يصعب تصميمها لذلك طوروا نموذجاً حركياً جديداً قادراً على حساب الذاكرة الطويلة وعدم التناسق في عملية التقلب وعلى هذا الأساس تمَ في هذا البحث إستخدام هذه التقنية الملائمة لبيانات الأرقام القياسية المتقلبة في إقليم كوردستان لتقدير نموذج ملائم للتنبؤ المستقبلي.

2: الجانب النظري:

التنبؤ هو نوع من المرشح الحركي، حيث يتم استخدام القيم السابقة لسلسلة زمنية واحدة أو أكثر للتنبؤ بالقيم المستقبلية من خلال الشبكات العصبية الحركية، والتي تتضمن إستغلال خطية إعتمادية المشاهدة الحالية على المشاهدة التي سبقتها مستخدمةً مرشحات غير خطية والتنبؤ بها. هناك العديد من التطبيقات للتنبؤ، على سبيل المثال قد يرغب المحلل المالي في التنبؤ بالقيمة المستقبلية لسهم أو سند أو أداة مالية أخرى أو قد يرغب المهندس في توقع الفشل الوشيك لمحرك نفاث وهكذا من الأمثلة التطبيقية الأخرى [3].

تُستخدم النماذج التنبؤية أيضاً لتعريف النظام (أو النمذجة الحركية)، حيث تنشئ نماذج حركية للأنظمة المادية [4]. هذه النماذج الحركية مهمة للتحليل والمحاكاة والمراقبة والتحكم في مجموعة متنوعة من الأنظمة بما في ذلك أنظمة التصنيع والعمليات الكيميائية وأنظمة الروبوتات والفضاء [5]. وتتيح هذه التقنية حل ثلاثة أنواع من مشاكل السلاسل الزمنية غير الخطية أحدها الإنحدار الذاتي غير الخطي (Nonlinear Autoregressive).



الشكل (1): مخطط الإنحدار الذاتي غير الخطي للشبكات العصبية

الشكل (1) يمثل مخطط التغذية العكسية (Feedback) للشبكة العصبية وعدد الخلايا العصبية المخفية وعدد (lag) أو (Delays) يرمز له d معتمداً على الصيغة (1) وإذا لم تعمل الشبكة بشكل جيد بعد التدريب يتم تغيرها [6].

سيتم إنشاء وتدريب الشبكة في شكل حلقة مفتوحة (خطوة واحدة) والتي لها أكثر فعالية من التدريب مع حلقة مغلقة (متعدد الخطوات) لأنه يسمح لنا بتزويد الشبكة بمدخلات التغذية العكسية الصحيحة حتى أثناء قيامنا بتدريبها لحساب مخرجات تنبؤية لصحيحة. بعد التدريب، يمكن تحويل الشبكة إلى نموذج حلقة مغلقة، أو أي شكل آخر، يتطلبه التطبيق [7] .

يتم تقسيم مشاهدات السلسلة الزمنية إلى ثلاث مجموعات وكما يلي [8] :

• يتم إستخدام أول 70٪ من البيانات للتدريب، ويتم تقديمها إلى الشبكة أثناء التدريب وضبط الشبكة وفقا لخطأها.

• 15٪ من البيانات تستخدم للتحقق من تعميم الشبكة وإيقاف التدريب قبل overfitting، وتستخدم هذه لقياس تعميم الشبكة، ووقف التدريب عندما يتوقف التعميم عن التحسن [9] .

• يتم إستخدام آخر 15٪ من البيانات بشكل مستقل تماماً لتعميم الشبكة وإختبارها، وهذه توفر مقياس مستقل لأداء الشبكة أثناء وبعد التدريب.

الخوارزمية المستخدمة في هذا البحث تعمل على استخدام دالة تحويل السلسلة الزمنية للجزء المخفي من الشبكة الافتراضية (Levenberg–Marquardt) التي يرمز لها (L–M) ولها دالة التحويل الخطي لجزء المخرجات لتلك الشبكة فضلاً عن تدريب الشبكة العصبية لعدد مختلف من قيم (d) التي لها عشرة خلايا عصبية مخفية كعدد مفترض يمكن تغييره إذا لم يحصل الباحث على نتائج جيده (متوسط الخطأ التربيعي النهائي صغير وخطأ مجموعة الاختبار ومجموعة اختبار التحقق لهما خصائص مماثلة) [10].

1.2: بنية الشبكة العصبية (Neural network architecture) :

في هذا البحث، أستخدم المستقبلات متعدد الطبقات (multilayer perceptron) التي يرمز لها (MLP)، وتنظيم الشبكة العصبية البيزية (Bayesian regularization neural

(network، التي تكتب إختصاراً (BRNN) لنمذجة العلاقات غير الخطية بين متجهات الإدخال وميزات السرعة المستخرجة، ومخرجات الشبكة مع دالة تحويل غير الخطية [11]. تم تصميم شبكة (MLP) الأساسية من خلال ترتيب الوحدات في بنية ذات طبقات، حيث تأخذ كل خلية عصبية في طبقة مدخلاتها من إخراج الطبقة السابقة أو من مدخلات خارجية [12]. يوضح الشكل (1) مخططاً لهيكل (MLP) الخاص بدوال التحويل للطبقة المخفية في شبكة التغذية العكسية وهي تمثل الدالة اللوجستية (Guerino, et al. 2005) في الصيغة (2).

$$F(\eta)=\frac{1}{1+\exp(-\eta)} \quad \cdots \quad (2)$$

حيث تمثل η مدخلات الشبكة ونظراً لإستخدام (MLP) كتقنية إنحدار، فيجب أن تنتج قيم إخراج معقولة خارج نطاق [−1،1] وبالتالي في طبقة الإخراج تستخدم دالة التحويل الخطي لذلك قد تستخدم هذا النوع من الشبكات كمقارب للدالة العامة كدالة لميزات السرعة [13]. النموذج الرياضي لحساب هذا النوع [Pedram, et al., (2017)] موضح في الصيغة (3):

$$y=b_0+\sum_{j=1}^{H} w_j.F\left(\sum_{i=1}^{N} w_{ij}+b_j\right) \quad \cdots \quad (3)$$

w_{ij} عبارة عن أوزان غير خطية تربط الخلايا العصبية المدخلة (N) بالخلايا العصبية المخفية (H) في الطبقة، والأوزان w_j الخطية التي تربط الخلايا العصبية المخفية بطبقة المخرجات.

2.2: التدريب (The training):

تم اقتراح العديد من خوارزميات التدريب وقواعد التعلم لتحديد الأوزان والمعلمات في الشبكات العصبية؛ ومع ذلك لا يمكن تحديد حل أدنى عام. لذلك يعد تدريب الشبكة أحد أهم الخطوات لتصميم الشبكة العصبية التي هي في الأساس تقنية تحسين النسب المتدرجة وهي تقنية أساسية لتدريب الشبكات العصبية العكسية [14] ومع ذلك فإنه يحتوي على بعض القيود مثل التقارب البطيء، طبيعة البحث المحلي وتجهيز البيانات. ولكن الإفراط في تدريبها يؤدي إلى فقد قدرة الشبكات على تقدير المخرجات بشكل صحيح (Burden and Winkler, 2009). نتيجة لذلك، يمكن أن يكون التحقق من صحة النماذج المشكلة، فضلاً عن أن تحسين بنية الشبكة يستغرق وقتًا طويلاً. هناك بعض التعديلات على (backpropagation) مثل خوارزميات التدرج المتزامن و(L–M)، والتي هي أسرع من أي معالجة من خوارزمية (backpropagation), (Mastersm, 1995). خوارزمية (L–M) هي لتقليل مقدار الخطأ التربيعي (Gavin, 2013) والتغلب على بعض القيود في (backpropagation) الخوارزمية القياسية، مثل مسألة دقة التوفيق. يمكن تجنب مشكلة دقة التوفيق في بنيان الشبكة وهو تحدياً

خطيراً،[15] لأنه يحاول تحقيق تقدير دقيق للدالة المنمذج بواسطة شبكة عصبية مع الحد الأدنى لعدد بيانات المدخلات والمعاملات. إن وجود عدد كبير جداً من الخلايا العصبية في الطبقة المخفية يمكن أن يتسبب في حدوث تلاشي دقة التوفيق،[16] نظراً لأن ضوضاء البيانات مصاغة مع الاتجاهات. فضلاً عن إمكانية أن يسبب عدد غير كافٍ من الخلايا العصبية في الطبقة المخفية مشاكل في بيانات التعلم. لغرض العثور على العدد الأمثل من الخلايا العصبية في الطبقة المخفية، وعادة يتم إجراء (10) تجارب اختيار نموذج مع عدد مختلف من الخلايا العصبية تتراوح بين (5 إلى 60) وحساب خطأ التحقق من صحة النموذج لكل تجربة. باستخدام نظرية التقريب الشامل [17]، فقد ثبت نظرياً أن الشبكة العصبية ذات طبقة مخفية واحدة فقط باستخدام دالة تنشيط مقيدة ومحددة يمكنها تقريب أي دالة (Hornik, et al. 1989). وبالتالي في جميع التكوينات في التجربة والاختبارات، يتم إستخدام طبقة مخفية واحدة فقط. اقترح ماكاي (Mackay, 1991) خوارزمية تنظيم بيز لمواجهة مثل هذا التحدي فضلاً عن تعدد المعلمات غير ذات الصلة والمترابطة للغاية مشكلة أخرى يمكن أن تؤدي إلى تدهور قدرة الشبكة على تقريب الدالة والتي يمكن حلها عن طريق النظر في التنظيم (Burden and Winkler, 2009). تنظيم يمكن دمجه على غرار إحصاء بيز [18]. من خلال هذه الطريقة، يمكننا إزالة معظم عيوب الشبكة العصبية العكسية. في هذا البحث تم إستخدام الشبكة العصبية للإنحدار البيزي (BRNN) لتحليل الانحدار والتي تعد بمثابة تعبير لخوارزمية (L–M).

نظرية بيز الافتراضية (L–M) (Foresee and Hagan, 1997) هي دالة تدريب على الشبكة تعمل على تحديث قيم الوزن والتحيز وفقاً لتحسين أمثل لتلك الطريقة، وهذه طريقة بسيطة لتقريب الدالة الذي يحتاج إلى تخزين كبير لبعض المصفوفات وهي طريقة مناسبة لتقدير وقت إستخدام عدد كبير من المدخلات للحصول على أفضل إخراج وتقليل مجموع مربعات الخطأ والأوزان ثم تحديد التركيبة الصحيحة لتركيب شبكة تعميم جيده [19] .

خوارزمية (L–M) هي خوارزمية تكرارية تجد الحد الأدنى من دالة متعدد المتغيرات والتي تعتمد على مجموع مربعات الدوال الحقيقية غير الخطية (Lourakis, 2005)، يستخدم (L–M) على نطاق واسع لحل مشاكل المربعات الصغرى غير الخطية [20] وعادةً ما تعتبر تقنية قياسية للقيام بذلك وهذه الخوارزمية هي طريقة مناسبة للمنحنى، وهي مزيج من تحديث النسب المتدرج وتحديث (Gauss–Newton) اللتان تمثلان طريقتان للتصغير المكافئة للصيغة (4) التي تمثل معادلات النسب المتدرجة والمعادلة الطبيعية لتحديث (Gauss–Newton) التي تمثلها الصيغة (5)، (Pedram, et al. 2017) :

$$h_{gd} = \alpha \, \mathrm{J}' W (y - \hat{y}) \quad \cdots \quad (4)$$

$$\mathrm{J}' W \, \mathrm{J} . h_{gd} = \mathrm{J}' W (y - \hat{y}) \quad \cdots \quad (5)$$

حيث α تمثل نسبة الإضمحلال ولدينا:

$$\mathrm{J}=\begin{bmatrix} \frac{\partial e_1(w)}{\partial w_1} & \frac{\partial e_1(w)}{\partial w_2} & \cdots & \frac{\partial e_1(w)}{\partial w_n} \\ \frac{\partial e_2(w)}{\partial w_1} & \frac{\partial e_2(w)}{\partial w_2} & \cdots & \frac{\partial e_2(w)}{\partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial e_N(w)}{\partial w_1} & \frac{\partial e_N(w)}{\partial w_2} & \cdots & \frac{\partial e_N(w)}{\partial w_n} \end{bmatrix}$$

تبين الصيغة (5) أن خوارزمية (L–M) هي عباره عن مزيج خطي من تحديث النسب المتدرج وتحديث (Gauss–Newton) حيث تختلف المعلمة بشكل متكيف بينهما. وقيمة λ تحدد هذا الاختلاف [21]، وكلما كانت قيمة λ صغيرة ، فإنها تميل نحو تحديث (Gauss–Newton) في حين عندما تكون قيمة λ كبيرة فستكون أقرب إلى تحديث النسب المتدرج لذلك سيتم البدأ بقيمة كبيرة وبالتالي فإن التحديثات الأولى ستكون صغيرة القيم في إتجاه الهبوط الأكثر حدة، تماماً كما ينحدر التدرج اللوني والصيغة (7) تمثل ذلك:

$$[\mathrm{J}'W\,\mathrm{J}+\lambda\,\mathrm{I}]h_{lm}=\mathrm{J}'W(y-\hat{y}) \quad \cdots \quad (7)$$

3.2: دالة الإرتباط الذاتي للخطأ (The error autocorrelation function):

تستخدم دالة الارتباط الذاتي لخطأ التنبؤ للحصول على نموذج تنبؤ مثالي، حيث يصف كيف ترتبط أخطاء التنبؤ في الوقت المناسب للحصول على نموذج تنبؤ مثالي [22] يجب أن يكون هناك قيمة غير صفرية واحدة فقط لدالة الارتباط الذاتي، ويجب أن تحدث عند تأخر صفري. هذا يعني أن أخطاء التنبؤ كانت غير مرتبطة تماماً مع بعضها البعض (White noise). إذا كان هناك ارتباط معنوي في أخطاء التنبؤ، فيجب أن يكون من الممكن تحسين التنبؤ– ربما عن طريق زيادة عدد الفجوات في خطوط التأخير المشاهدة وإذا كانت هناك حاجة إلى نتائج أكثر دقة [23]، فيمكن إعادة تدريب الشبكة الذي سيؤدي ذلك إلى تغيير الأوزان والتحيزات الأولية للشبكة، وقد ينتج شبكة محسنة بعد إعادة التدريب [24].

4.2: الأرقام القياسية ومعدل التضخم السنوي:

تعتبر الأرقام القياسية لأسعار المستهلك أحد المؤشرات الهامة المستخدمة في الدراسات الاقتصادية وخطط التنمية كونها تعكس التغيرات التي تطرأ على هيكل القطاعات الإنتاجية والاستهلاكية في المجتمع. حيث يستخدم الرقم القياسي [25] ، على نطاق واسع، كمؤشر يقيس اتجاهات التضخم والانكماش الاقتصادي إضافة إلى استخدامه كمقياس للتغيرات في القوة الشرائية للعملة الوطنية وفي الحسابات القومية لتركيب تقديرات معظم الأنشطة والإنفاق الخاص

وما يتعلق به من مكونات بالأسعار الثابتة (الحقيقية) [26]. ويمكن حساب معدل التضخم السنوي من خلال الصيغة الآتية:

$$U = \left(\frac{CPI_1}{CPI_0} - 1 \right) \times 100 \quad \cdots \quad (8)$$

حيث يمثل CPI_1 الرقم القياسي للسنة الحالية في حين يمثل CPI_0 الرقم القياسي للسنة الماضية. ويتم إحتساب الرقم القياسي لمجموعات رئيسية وفرعية والأقسام وفي النهاية نحصل على الرقم القياسي العام لكل سنة [27].

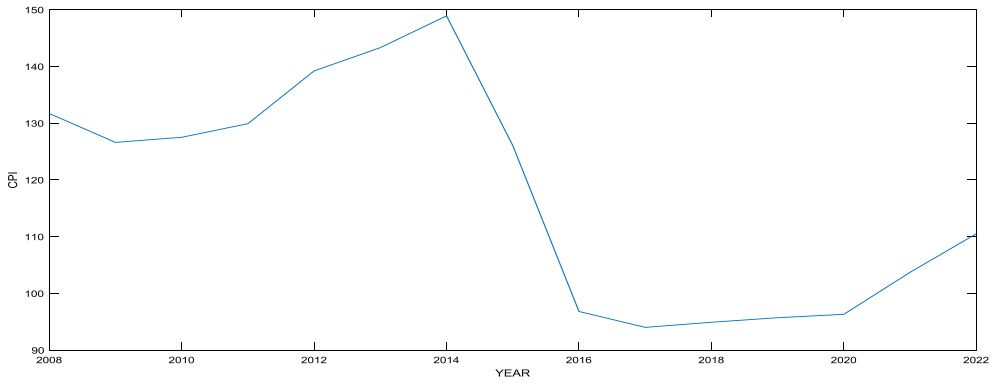
3: الجانب التطبيقي:

تم إستخدام الطريقة الهجينة بين الشبكات العصبية والسلاسل الزمنية في تقدير النموذج الخطي المستخدم للتنبؤ بالرقم القياسي للفترة الزمنية (2023–2025) لأقليم كوردستان إعتماداً على بيانات مؤخوذة من هيئة إحصاء إقليم كوردستان / قسم الأرقام القياسية للفترة الزمنية (2008–2022)، علماً أن الهيئة بدأت بعملية حساب الرقم القياسي الخاص بإقليم كوردستان منذ عام (2008). كما هو موضح في الجدول الآتي:

الجدول (1): السلسلة الزمنية للرقم القياسي في إقليم كوردسان العراق

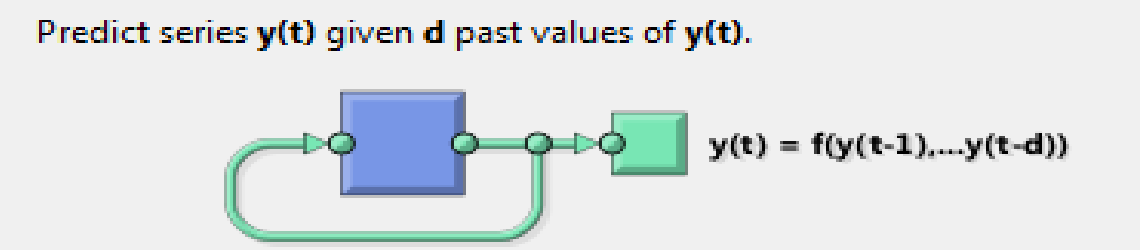
السنة	2008	2009	2010	2011	2012	2013	2014	2015
الرقم القياسي	131.7	126.6	127.5	129.9	139.2	143.3	148.9	126.05
السنة	2016	2017	2018	2019	2020	2021	2022	
الرقم القياسي	96.8	94.0	94.9	95.7	96.3	103.7	110.5	

الشكل الآتي يوضح قيم الأرقام القياسية للفترة الزمنية (2008–2022) لأقليم كوردستان:



الشكل (2): السلسلة الزمنية للأرقام القياسية

تم إستخدام الشبكات العصبية للتغذية العكسية مع السلاسل الزمنية في تقدير النموذج الخطي الملائم لبيانات الأرقام القياسية في إقليم كوردستان إعتماداً على البرنامج الجاهز (software) للغة ماتلاب وتحديد النموذج غير الخطي الحركي للشبكات العصبية الملائم لتوزيع السلسلة في الشكل (2) إعتماداً على المعادلة (1) وذلك لوجود سلسلة واحدة فقط حيز التطبيق مع عدد قليل من المشاهدات [28]. القيم المستقبلية لسلسلة زمنية $y(t)$ يتم التنبؤ بها فقط من القيم السابقة لتلك السلسلة ويمكن أيضًا استخدام هذا النموذج للتنبؤ بالأدوات المالية، ولكن دون استخدام سلسلة مصاحبة [29]. والشكل الآتي يوضح ذلك:



الشكل (3): مخطط الإنحدار الذاتي غي الخطي (NAR)

من خلال البرنامج سيتم تقسيم متجهات الإدخال والمتنبأ بها بشكل عشوائي إلى ثلاث مجموعات وكما يلي:

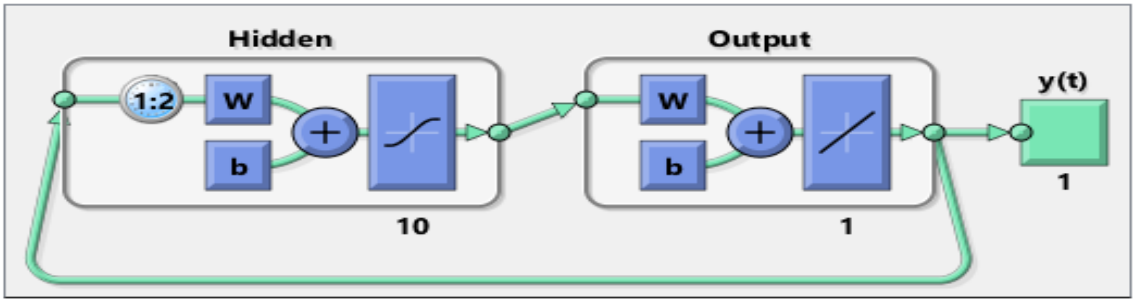
المجموعة الأولى: سيتم استخدام 70 ٪ للتدريب.

المجموعة الثانية: سيتم استخدام 15٪ للتحقق من تعميم الشبكة وإيقاف التدريب قبل تركيبها.

المجموعة الثالثة: سيتم استخدام آخر 15٪ كاختبار مستقل تماماً لتعميم الشبكة.

لذلك ستقسم البيانات إلى (11، 2 و2) للمجموعات الثلاث على التوالي.

الشبكة لها مدخل واحد فقط. في وضع الحلقة المغلقة، يتم ربط هذا الإدخال بالإخراج، ولمحاكاة خطوات الشبكة 20 مرة إلى الأمام، يتم إدخال طابور خلايا فارغ بطول 20. لذلك فإن الشبكة العصبية مع طبقة واحدة من العصبونات المخفية يمثلها الشكل الآتي:

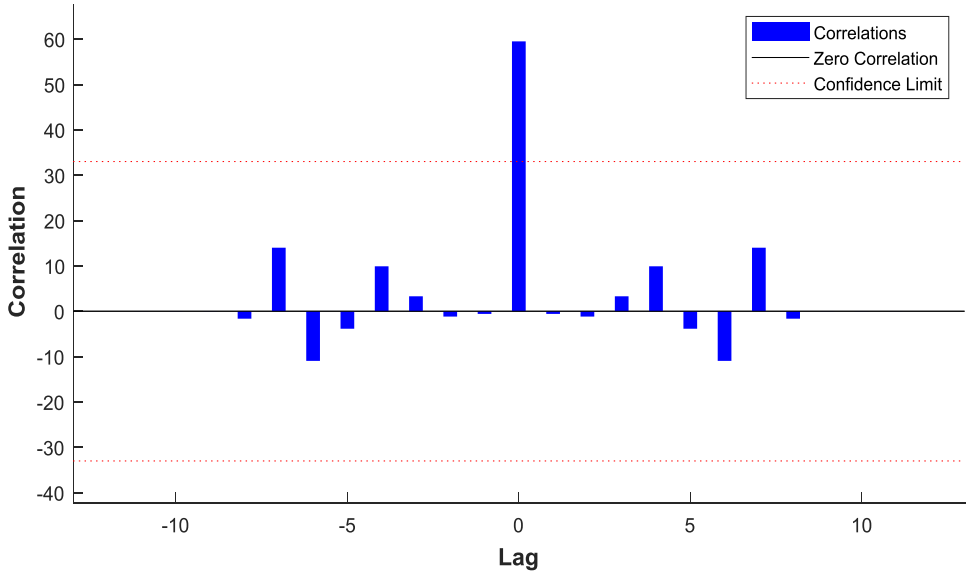


الشكل (4): مخطط التغذية العكسية للشبكة العصبية

الشكل (4) يمثل مخطط التغذية العكسية للشبكة العصبية وعدد خلايا العصبونات المخفية (التي تساوي 10) وعدد الفجوات الزمنية (lag) يساوي (2) معتمداً على الصيغة رقم (1) وإذا لم تعمل الشبكة جيدًا بعد التدريب يتم تغيرها[30].

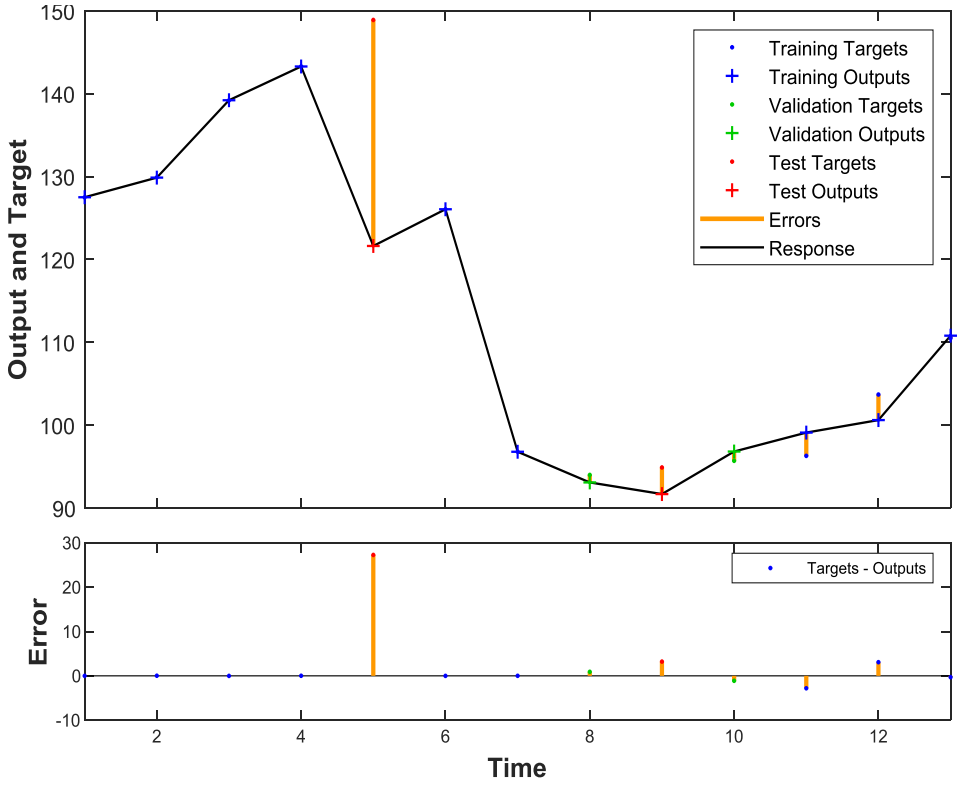
سيتم إنشاء وتدريب الشبكة في شكل حلقة مفتوحة (التدريب بإستخدام التغذية العكسية بطريقة L–M) كما هو موضح أدناه. بعد التدريب، يمكن تحويل الشبكة إلى نموذج حلقة مغلقة [31]، أو أي شكل آخر، يتطلبه التطبيق، وهنا سيتم تدريب الشبكة لتناسب مجموعة بيانات السلاسل الزمنية، وذلك باستخدام التطبيق الديناميكي لسلسلة الشبكة العصبية المتمثلة بالأرقام القياسية. وبعد تكرار تدريب الشبكة إلى أن يتم الحصول على أكبر معامل تحديد (أو تفسير) وأقل متوسط خطأ تربيعي (MSE) ويتم عرض نتائج التحليل كما يلي:

1- الإرتباط الذاتي للأخطاء: تقع الارتباطات، باستثناء العلاقة التي تكون عند تأخر الصفر [32]، تقريباً ضمن حدود الثقة البالغة 95٪ حول الصفر، لذلك يبدو أن النموذج مناسب. كما يلاحظ في الشكل الآتي:



الشكل (5): دالة الإرتباط الذاتي للأخطاء

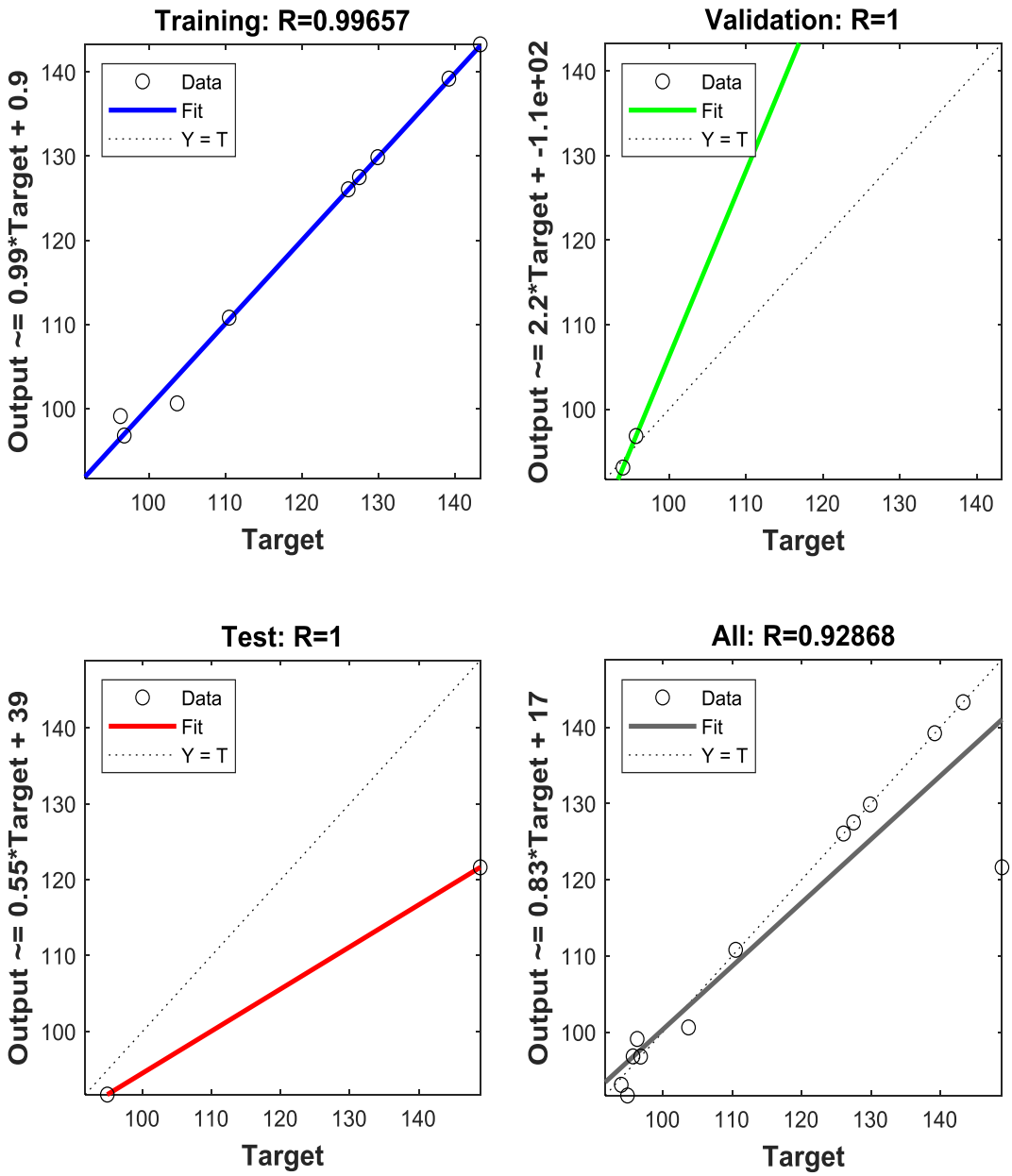
2- إستجابة السلسلة الزمنية: الشكل (6) يوضح إستجابة السلسلة الزمنية التي تعرض المدخلات (الأرقام القياسية الحقيقية) والقيم المتنبأ بها (القيم المتوقعة) والأخطاء مقابل الزمن. كما يشير إلى النقاط الزمنية التي تم اختيارها للتدريب والاختبار والتحقق من الجودة.



الشكل (6): إستجابة السلسلة الزمنية

3- تقييم الشبكة العصبية: في هذه المرحلة، يمكن اختبار الشبكة مقابل بيانات جديدة. إذا كان الأداء في مجموعة التدريب جيدًا، ولكن أداء مجموعة الاختبار أسوأ كثيراً، مما قد يشير إلى زيادة التحمل، فإن تقليل عدد الخلايا العصبونية يمكن أن يحسن النتائج. وهنا تم الحصول على نتائج مرضية بمعامل تحديد بلغ 92.868% ومتوسط خطأ تربيعي بلغ (59.50796).

4- نماذج الإنحدار الذاتية: تعرض مخططات الانحدار التالية (الشكل-7) مخرجات الشبكة فيما يتعلق بأهداف التدريب [33]، التحقق من الصحة ومجموعات الاختبار للحصول على ملائمة مثالية، حيث يجب أن تقع البيانات على طول خط 45 درجة حيث تساوي مخرجات الشبكة الأهداف. بالنسبة لهذا التحليل، تكون الملائمة جيدة بشكل معقول لجميع مجموعات البيانات، حيث تكون قيم معامل التحديد في حالة التدريب والكل 0.92868 أو أعلى، وهنا كانت قيمتها (0.99657 و 0.92868) على التوالي وهي قيم مقبولة تسمح بإعتماد النموذج الكلي المقدر [34] بالتنبؤ بالقيم المستقبلية.



الشكل (7): نماذج الإنحدار الذاتية

وتلخيص القيم المقدرة وخطأ التقدير في الجدول الآتي:

الجدول (2): السلسلة الزمنية للرقم القياسي والمقدرة وخطأ التقدير

السنة	الرقم القياسي	القيم المقدرة	خطأ التقدير
2008	131.7	-----	-----
2009	126.6	-----	-----
2010	127.5	127.5064	-0.0064
2011	129.9	129.8772	0.0228
2012	139.2	139.2146	-0.0146
2013	143.3	143.2878	0.0122
2014	148.9	121.6282	27.2718
2015	126.05	126.0598	-0.0098
2016	96.8	96.79520	0.0048
2017	94.0	93.09440	0.9056
2018	94.9	91.68950	3.2105
2019	95.7	96.82220	-1.1222
2020	96.3	99.10660	-2.8066
2021	103.7	100.6188	3.0812
2022	110.5	110.8017	-0.3017

النموذج الكلي المقدر هو كمايلي:

$$CPI_{2023} = 17 + 0.83 CPI_{2022}$$

من خلال النموذج المقدر أعلاه يمكن التنبؤ بالرقم القياسي لعام (2023)، (2024) و(2025) والملخصة في الجدول الآتي:

الجدول (3): القيم التنبؤية للرقم القياسي في إقليم كوردسان العراق

السنة	2023	2024	2025
الرقم القياسي	108.715	107.233	106.004

من خلال الجدول (3) نلاحظ أن هنالك إرتفاع عام بالقيم التنبؤية في إقليم كوردستان للسنوات القادمة المحددة بالفترة (2023-2025) والتي تمّ إستخدامها في حساب معدل التضخم العام لتلك الفترة إعتماداً على المعادلة (8) كما يوضحة الجدول الآتي:

الجدول (4): معدلات التضخم التنبؤية في إقليم كوردستان العراق

السنة	2023	2024	2025
معدل التضخم العام	-1.615%	-1.363%	-1.146

الجدول (4) يوضح أن هنالك إرتفاع عام بمعدل التضخم للسنوات القادمة وبدرجات متفاوتة والتي يمكن الإعتماد عليها في رسم الخطط الإقتصادية والمالية لتلك السنوات من قبل وزارة التخطيط في إقليم كوردستان.

كما تم تقدير نماذج الأرقام القياسية إلى (12) قسم رئيسي موضح في (جدول I في الملحق) وحساب التنبؤ المستقبلي لها للفترة (2023–2025) مع تقدير معامل التحديد ومتوسطات الخطأ التربيعي والتي من خلالها تم حساب معدلات التضخم لتلك الفترة والملخصة في (الجدول II في الملحق) وبينت أن هنالك ارتفاع في بعض الأقسام وانخفاض في أقسام أخرى، حيث كان هنالك ارتفاع في المواد (الاغذية والمشروبات غير الكحولية، السكن ، المياه ، الكهرباء، الغاز ، الصحة، النقل، الترفيه والثقافة، التعليم والمطاعم) في حين كان هنالك انخفاض في المواد (المشروبات الكحولية و التبغ، الملابس والاحذية، التجهيزات والمعدات المنزلية والصيانة، الاتصال و السلع والخدمات المتنوعة).

4: الإستنتاجات والتوصيات:

من خلال الجانب التطبيقي توصلت الباحثون إلى أهم الاستنتاجات الآتية:

1– إمكانية إستخدام النماذج الحركية للشبكات العصبية مع السلاسل الزمنية للتنبؤ بالأرقام القياسية التي تتضمن تقلبات كبيرة في مقاديرها.

2– هنالك إنخفاض في مستوى الرقم القياسي العام للسنوات المتنبأ بها والذي أدى إلى إنخفاض معدل التضخم العام السنوي.

3– هنالك إرتفاع في مستوى الرقم القياسي لبعض الأقسام الرئيسة وإنخفاض في بعضها الآخر للسنوات المتنبأ بها والذي أدى إلى إرتفاع أو إنخفاض معدل التضخم السنوي.

كما يوصي الباحثون بما يلي:

1– إستخدام نتائج الرقم القياسي العام والأقسام الرئيسية المتنبأ بها ومعدل التضخم السنوي العام والأقسام الرئيسية للفترة (2023–2025) في رسم السياسات الإقتصادية والمالية من قبل وزارة التخطيط في إقليم كوردستان.

2- إجراء دراسات وطرائق أخرى للتنبؤ بالأرقام القياسية ومعدلات التضخم مثل النماذج الحركية التي تعتمد على المدخلات كمتغيرات مستقلة والمخرجات كمتغير تابع.

5. References

[1] C. M. Kuan and H. White. Artificial neural networks: An econometrics perspective. Econometric Review, 13:1–91, 1994.

[2] Ali, Taha Hussein & Awaz Shahab M. "Uses of Waveshrink in Detection and Treatment of Outlier Values in Linear Regression Analysis and Comparison with Some Robust Methods", Journal of Humanity Sciences 21.5 (2017): 38-61.
[3] H. P. Gavin, The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems, Department of Civil and Environmental Engineering, Duke University (2013) 1–17 doi:10.1080/10426914.2014.941480.
[4] Ali, Taha Hussein & Mardin Samir Ali. "Analysis of Some Linear Dynamic Systems with Bivariate Wavelets" Iraqi Journal of Statistical Sciences 16.3 (2019): 85-109.
[5] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural networks 2 (5) (1989) 359–366.
[6] Ali, Taha Hussein & Qais Mustafa. "Reducing the orders of mixed model (ARMA) before and after the wavelet de-noising with application." Journal of Humanity Sciences 20.6 (2016): 433-442.
[7] D. J. C. Mackay, Bayesian Methods for Adaptive Models, thesis, CIT (1991) 98.
[8] Ali, Taha Hussein and Jwana Rostam Qadir. "Using Wavelet Shrinkage in the Cox Proportional Hazards Regression model (simulation study)", Iraqi Journal of Statistical Sciences, 19, 1, 2022, 17-29.
[9] Guerino Mazzola, Gerard Milmeister and Judy Weissmann, (2005), Comprehensive Mathematics for Computer Statistics 2.
[10] Ali, Taha Hussein, "Estimation of Multiple Logistic Model by Using Empirical Bayes Weights and Comparing it with the Classical Method with Application" Iraqi Journal of Statistical Sciences 20 (2011): 348-331.
[11] Pedram Gharani, Brian Suoletto, Tammy Chung, and Hassan Karimi, 92017), An Artificial Neural Network for Gait Analysis to Estimate Blood Alcohol Content Level, arXiv:1712.01691v3.
[12] Ali, Taha Hussein, 2018, Solving Multi-collinearity Problem by Ridge and Eigenvalue Regression with Simulation, Journal of Humanity Sciences, 22.5: 262-276.
[13] Ali, Taha Hussein, and Dlshad Mahmood Saleh. "COMPARISON BETWEEN WAVELET BAYESIAN AND BAYESIAN ESTIMATORS TO REMEDY CONTAMINATION IN LINEAR REGRESSION MODEL" PalArch's Journal of Archaeology of Egypt/Egyptology 18.10 (2021): 3388-3409.
[14] Ali, Taha Hussein, and Saleh, Dlshad Mahmood, "Proposed Hybrid Method for Wavelet Shrinkage with Robust Multiple Linear Regression Model: With Simulation Study" QALAAI ZANIST JOURNAL 7.1 (2022): 920-937.

[15] Ali, Taha Hussein, Avan Al-Saffar, and Sarbast Saeed Ismael. "Using Bayes weights to estimate parameters of a Gamma Regression model." Iraqi Journal of Statistical Sciences 20.1 (2023): 43-54.
[16] Ali, Taha Hussein, Heyam Abd Al-Majeed Hayawi, and Delshad Shaker Ismael Botani. "Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level." Communications in Statistics-Simulation and Computation 52.4 (2023): 1476-1489.
[17] E. Maaoumi, A. Khotanzad, and A. Abaye. Artificial neural networks for some macroeconomic series: A first report. Econometric Reviews, 13:105–122, 1994.
[19] Ali, Taha Hussein, Nasradeen Haj Salih Albarwari, and Diyar Lazgeen Ramadhan. "Using the hybrid proposed method for Quantile Regression and Multivariate Wavelet in estimating the linear model parameters." Iraqi Journal of Statistical Sciences 20.1 (2023): 9-24.
[20] Ali, Taha Hussein, Nazeera Sedeek Kareem, and mohammad, Awaz Shahab "Construction robust simple linear regression profile Monitoring" journal of kirkuk University for Administrative and Economic Sciences, 9.1. (2019): 242-257.
[21] Ali, Taha Hussein, Rahim, Alan Ghafur, and Saleh, Dlshad Mahmood. "Construction of Bivariate F-Control Chart with Application" EURASIAN JOURNAL OF SCIENCE AND ENGINEERING (EAJSE), 4.2 (2018): 116-133.
[22] Ali, Taha Hussein, Saman Hussein Mahmood, and Awat Sirdar Wahdi. "Using Proposed Hybrid method for neural networks and wavelet to estimate time series model." Tikrit Journal of Administration and Economics Sciences 18.57 part 3 (2022).
[23] Ali, Taha Hussein. "Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study)." Communications in Statistics-Simulation and Computation 51.2 (2022): 391-403. https://doi.org/10.1080/03610918.2019.1652319
[24] Ali, Taha Hussein; Saleh, Dlshad Mahmood; Rahim, Alan Ghafur. "Comparison between the median and average charts using applied data representing pressing power of ceramic tiles and power of pipe concrete", Journal of Humanity Sciences 21.3 (2017): 141-149.
[25] Ali, Taha Hussien, (2017), "Using Proposed Nonparametric Regression Models for Clustered Data (A simulation study)." Journal of Humanity Sciences, 29.2: 78-87.
[26] Ali, Taha Hussien, Nazeera Sedeek Kareem, and Awaz shahab mohammad, (2021), Data de-noise for Discriminant Analysis by using Multivariate Wavelets (Simulation with practical application), Journal of Arab Statisticians Union (JASU), 5.3: 78-87

[27] Kareem, Nazeera Sedeek, Taha Hussein Ali, and Awaz shahab M, "De-noise data by using Multivariate Wavelets in the Path analysis with application", Kirkuk University Journal of Administrative and Economic Sciences, 10.1 (2020): 268-294.
[28] N. R. Swanson and H. White. A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. The Review of Economics and Statistics, 79:4:540–550, 1997.
[29] Mustafa, Qais, and Ali, Taha Hussein. "Comparing the Box Jenkins models before and after the wavelet filtering in terms of reducing the orders with application." Journal of Concrete and Applicable Mathematics 11 (2013): 190-198.
[30] Omar, Cheman, Taha Hussien Ali, and Kameran Hassn, Using Bayes weights to remedy the heterogeneity problem of random error variance in linear models, IRAQI JOURNAL OF STATISTICAL SCIENCES, 17, 2, 2020, 58-67.
[32] Raza, Mahdi Saber, Taha Hussein Ali, and Tara Ahmed Hassan. "Using Mixed Distribution for Gamma and Exponential to Estimate of Survival Function (Brain Stroke)." Polytechnic Journal 8.1 (2018).
[33] L. Catania and S. Grassi. Modelling crypto-currencies financial time series. Technical Report, SSRN Working paper, 2018.
[34] L. Catania, S. Grassi, and F. Ravazzolo. Predicting the volatility of cryptocurrency time-series. Technical Report, MIMEO, 2018.
[35] F.D. Foresee and M.T. Hagan. Gauss-newton approximation to Bayesian regularization. In Proceedings of the 1997 International Joint Conference on Neural Networks, pages 1930–1935, 1997.
[36] M. I. a. Lourakis, A Brief Description of the Levenberg-Marquardt Algorithm Implemented by levmar, Matrix 3 (2005) 2. doi:10.1016/j.ijinfomgt.2009.10.001.URL
[37] F. Burden, D. Winkler, Bayesian regularization of neural networks, Artificial Neural Networks: Methods and Applications (2009).
[38] T. Masters, Advanced algorithms for neural networks: a C++ sourcebook, John Wiley & Sons, Inc., 1995.

الملحق

الجدول (I): الأرقام القياسية لأسعار (12) قسم رئيسي خلال الفترة (2008-2018)

الاقسام الرئيسية	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
الاغذية والمشروبات غير الكحولية	138.9	144.1	147.5	150.8	160.1	161.7	165.8	130.3	98.3	90.9	91.9	92.7	91.8	90.4	90.4
المشروبات الكحولية و التبغ	119.8	116.4	115.9	114.8	117.3	120.3	124.8	124.2	118.7	120	122.9	119.2	120.8	123.9	123.9
الملابس والاحذية	122.8	125.6	124.4	127.7	136.1	143.4	148	128.7	112.2	92.3	90.5	93.8	93.6	100.4	100.4
السكن (الإيجار ، المياه ، الكهرباء، الغاز)	234	112.4	104.3	92.5	94.8	89.9	99.7	142.8	128	108.3	96.1	93.4	92.9	96.5	96.5
التجهيزات والمعدات المنزلية والصيانة	122.3	120.2	121.1	117.6	121.1	124.4	128.5	112.7	98.7	92.1	93.6	93.0	90.8	98.1	98.1
الصحة	174.2	162.9	169.7	201.5	208.7	210.6	214.1	167.9	111	107.7	117	117.4	117.2	124.4	124.4
النقل	108.7	107.6	108.3	107.6	109.6	112.6	111.9	99.2	96.3	87.2	89.1	90.8	98.8	112.8	112.8
الاتصال	98.1	99.2	97.5	91.7	93.1	77.5	74.2	84.9	93.3	93.4	102.9	113.1	115.5	114.8	114.8
الترفيه والثقافة	120.4	112.7	98.3	95.6	93.9	95.4	94	107.7	109	80.9	92	94.3	92.3	103.5	103.5
التعليم	206.3	173.2	147.1	181.7	208.5	241.6	258.7	178.4	110.7	118.2	115.9	115.2	108.2	104.7	104.7
المطاعم	127.6	132.8	135.2	138.1	147.7	151.3	158	134.5	106.2	96.4	98.1	97.5	89.0	85.5	85.5
السلع والخدمات المتنوعة	109.5	120.5	136.4	166.2	180	164	156.7	120.1	90.5	90.4	90.4	95.5	110.6	127.3	127.3

الجدول مأخوذ من هيئة إحصاء إقليم كوردستان العراق

الجدول (II): النماذج المقدرة مع بعض المعايير والقيم التنبؤية المستقبلية مع معدل التضخم إلى (12) قسم رئيسي للفترة (2023-2025)

المادة	الأنموذج	بعض المعايير		الرقم القياسي التنبؤي			معدل التضخم السنوي		
		R^2	MSE	2023	2024	2025	2023	2024	2025
الاغذية والمشروبات غير الكحولية	$CPI_1 = 14 + 0.90CPI_{10}$	94.25	97.99	100.92	104.83	108.35	4.5%	3.9%	3.4%
المشروبات الكحولية والتبغ	$CPI_2 = 0.69 + 0.99CPI_{20}$	99.46	0.124	122.33	121.80	121.27	-0.4%	-0.4%	-0.4%
الملابس والاحذية	$CPI_3 = 3.2 + 0.97CPI_{30}$	98.44	0.639	122.34	121.87	121.41	-0.4%	-0.4%	-0.4%
السكن (الإيجار، المياه ، الكهرباء،	$CPI_4 = -7.8 + 1.1CPI_{40}$	99.36	7.363	91.88	93.88	94.80	-0.4%	2.2%	1%

الغاز)									
التجهيزات والمعدات المنزلية والصيانة	$CPI_5 = -21 + 1.2CPI_{50}$	93.43	95.22	91.08	88.30	84.96	-2.5%	-1.3%	-3.8%
الصحة	$CPI_6 = 39 + 0.77CPI_{60}$	91.90	292.4	129.09	138.4	145.57	10.3%	7.2%	2.5%
النقل	$CPI_7 = 8.3 + 0.92CPI_{70}$	96.29	6.241	90.28	91.36	92.35	1.3%	1.2%	1.1%
الاتصال	$CPI_8 = -14 + 1.1CPI_{80}$	86.03	38.62	99.19	95.11	90.62	--3.6%	1.4%	-4.7%
الترفيه والثقافة	$CPI_9 = 4.7 + 0.97CPI_{90}$	88.77	17.72	93.96	95.84	97.67	1.2%	2%	1.9%
التعليم	$CPI_{10} = 1.5 + 0.99CPI_{100}$	99.43	32.53	116.24	116.58	116.91	0.3%	0.3%	0.3%
المطاعم	$CPI_{11} = 0.18 + 1CPI_{110}$	99.98	1.475	98.28	98.46	98.64	0.2%	0.2%	0.2%
السلع والخدمات المتنوعة	$CPI_{12} = 0.42 + 0.98CPI_{120}$	95.88	108.76	94.15	92.68	91.25	1.6%	-1.6%	-1.5%

الجدول من إعداد الباحثين حسب مخرجات لغة MATLAB

القيادة بالبيانات تجربة المؤسسات الحكومية في سلطنة عمان باستخدام الأساليب الإحصائية في صنع القرارات

د. يحيى بن خميس الحسيني
خبير إحصاء مركز مسقط للاستشارات الإحصائية
yahya@mstat.org

الملخص

تهدف الدراسة الى قياس تجربة القيادات الإدارية في بعض المؤسسات الحكومية بسلطنة نحو استخدام الأساليب الإحصائية في عملية صنع القرار ، وقد تكون مجتمع الدراسة من القيادات الإدارية من وظيفة رئيس قسم واعلى بمؤسسات الخدمة المدنية في سلطنة عمان، واشتملت عينة الدراسة على (236) موظف، وتبنى الباحث المنهج الوصفي التحليلي كمنهج للدراسة، واستعان بالاستبانة كأداة للدراسة، وتوصلت الدراسة إلى العديد من النتائج أهمها أن أسلوب الخبرات الشخصية احتل المرتبة الأولى في عملية صنع القرار في حين ان أسلوب استخدام الأساليب الإحصائية جاء في الرتبة الرابعة من اصل خمسه أساليب ، وان هنالك رغبة متوسطة لدى القيادات الإدارية بالتعرف على الأساليب الإحصائية ، وقلة البرامج التدريبية الإحصائية تعتبر العائق الرئيسي لهم لعدم استخدام الأساليب الإحصائية في عملية صنع القرار، وأوصت الدراسة بالعديد من التوصيات أهمها الاهتمام بتطوير ممارسات القيادات الإدارية للأساليب الإحصائية من خلال تكثيف البرامج التدريبية والاستفادة من تجارب المؤسسات الأخرى في استخدام الأساليب الإحصائية في عملية صنع القرار.

Abstract

This study aimed to study the measuring the experience of administrative leaders in some government institutions in the Sultanate towards the use of statistical methods in the decision-making process. The researcher used the analytical descriptive approach as a method for the study and used the questionnaire as a tool for the study. The study reached many results, the most important of which is that the method of personal experiences ranked first in the decision-making process, while the method of using statistical methods came in the fourth rank out of five methods, and that there is a moderate desire The administrative leaders have to learn about statistical methods, and the lack of statistical training programs is considered the main obstacle for them to not use statistical

methods in the decision-making process. Statistical methods in the decision-making process.

مقدمة

يعتبر علم الإحصاء من العلوم ذات الأهمية الكبيرة لكل من يستخدمه استخداما علميا في اتخاذ القرارات لذلك نجد أن لعلم الإحصاء دوراً بارزاً في دعم واتخاذ القرارات البسيطة وصولاً بالقرارات ذات الأهمية الكبيرة، كما يتضح أن المؤسسات والهيئات والشركات المختلفة المتقدمة والناجحة تستخدم الاحصاء بشكل كبير في تقييم منتجاتها ودعم اتخاذ القرارات السليمة.

إن أي دراسة علمية هادفة سليمة هي تلك التي تنتهي باتخاذ قرارات عملية صالحة للعمل بها غير أن صنع القرار ليس بالأمر السهل إلا إن الأسلوب الإحصائي وما يحمله من قوانين ونظريات إحصائية متطورة قد ساهم بقدر عظيم في صنع القرارات بدرجة من الثقة العالية وبنسب خطأ عند حدودها الدنيا، لقد أصبحت وظيفة صنع القرارات هي أساس العمل الإحصائي وأصبح علم الإحصاء يعرف من خلال وظيفة صنع القرارات.

هناك كثير من الدلائل على الاهتمام بالإحصاء واستخدامه منذ زمن بعيد بأغراض التنظيم والتخطيط. وأستخدم الإحصاء في عصره الأول في جمع البيانات عن السكان وحصرهم من قبل الدولة لأهداف معينة تتمثل في استخدامهم في الجيوش أو توجيههم لتنفيذ بعض المباني أو لتوزيع الأراضي على المزارعين بطريقة عادلة. وباختصار نجد أن الإحصاء قبل القرن العشرين كان يتمتع بالبساطة بحيث لم توفر المقومات الكافية لأن يصبح علما. وبظهور نظرية الاحتمالات في القرن الثامن عشر التي كان لها الدور الكبير في تطور هذا العلم حيث أصبح علما مستقلا وبدأ الاهتمام من قبل العلماء في تطبيق النظريات والأساليب الإحصائية في الكثير من العلوم باعتباره الطريقة الصحيحة والأسلوب الأمثل إتباعه في البحث العلمي.

مشكلة الدراسة

اتخذت خلال السنوات الأخيرة في سلطنة عمان عده قرارات التي تلامس المجتمع وافراده وتؤثر بشكل كبير على اساليب المعيشة وأسلوب حياتهم مثل رفع الدعم عن المحروقات، تخفيض الانفاق الحكومي، محدودية التوظيف في القطاع الحكومي وتوجيه الباحثين عن عمل الى القطاع الخاص، فرض بعض الضرائب البلدية وقرارات مختلفة، فهل يتم اتخاذ هذه القرارات على أسس علمية مبنيه على بيانات ومعلومات دقيقه وأساليب وطرق صحيحة، وتتضح مشكلة الدراسة في التساؤلات التالية

- هل يتم اتخاذ القرارات وفقا لدراسات علمية ذات جودة ودقة عالية،

– هل لدى متخذي القرار المام باستخدام أساليب إحصائية في دراسة الظواهر المختلفة قبل بناء أي قرار.

وعليه تقوم هذه الدراسة بالبحث في الطرق التي يتبعها متخذي القرار في سلطنة عمان ومدى استخدامهم للأساليب والطرق الإحصائية.

أهداف الدراسة

تهدف الدراسة إلى معرفة تأثير استخدام البيانات في اتخاذ القرارات من خلال متخذي القرار. وهل القرارات التي تم اتخاذها مبنيه على علم الإحصاء وماهي المعوقات التي تواجه متخذي القرار في عدم استخدام الإحصاء وماهي المقترحات والحلول من وجهه نظرهم.

أسئلة الدراسة

1–ما هو واقع استخدام القيادات الادارية في سلطنة عمان البيانات (الأساليب الإحصائية) في اتخاذ القرارات؟

2–مدى رغبة القيادات الإدارية في سلطنة عمان للمعرفة بالأساليب الإحصائية؟

3–ماهي المعوقات التي تواجه القيادات الادارية في استخدام البيانات لاتخاذ القرارات؟

4–ماهي المقترحات لدى القيادات الادارية لاستخدام البيانات والأساليب الإحصائية في اتخاذ القرارات؟

أهمية الدارسة

تنبع أهمية هذه الدارسة في مساهمتها بدور كبير في توضيح الواقع المستخدم في أساليب اتخاذ القرار لدى القيادات الإدارية بالوحدات الحكومية في سلطنة عمان، ودراسة مدى رغبة القيادات الإدارية في استخدام الإحصاء في عملية صنع القرار وتوضيح المعوقات التي تواجه القيادات الإدارية في عدم استخدام الإحصاء والحلول لمعالجتها، حيث تعتبر هذه الدراسة من أوائل الدراسات في سلطنة عمان التي تتناول هذا الموضوع على حسب علم الباحث.

حدود الدارسة

الحدود الموضوعية تقتصر هذه الدراسة على الأساليب الإحصائية واتخاذ القرار.

الحدود المكانية الجهات الحكومية في محافظة مسقط ذات العلاقة المباشرة مع المجتمع.

مصطلحات الدراسة

الإحصاء

الإحصاء هو مجموعة أساليب تستخدم في عمليات جمع وعرض وتلخيص وتحليل البيانات بهدف الوصول الى اتخاذ قرارات سليمة تتعلق بتفسير الظاهرة محل البحث والوقوف على سلوك تطورها وإمكانية التنبؤ الدقيق بما ستكون عليه فى المستقبل.

عملية صناعة القرار

القرار ما هو إلا سلوك واع منطقي يستند إلى المفاضلة بين عدة بدائل لحل مشكلة معينة واختيار البديل الأمثل الأفضل والفعال من البدائل المتاحة، الربضي (2016) وأن عملية صناعة القرار تقوم أساسا على وجود بديلين أو أكثر، وأن عملية المفاضلة بينها تتطلب وجود مجموعة من المعايير الموضوعية التي يعتمد عليها صانع القرار، كما تتطلب هذه العملية نظاما فعال للمعلومات يستند إليه صانع القرار.

اتخاذ القرار

إن عملية اتخاذ القرارات تتم لمعالجة مشكلات قائمة أو لمواجهة حالات أو مواقف معينة محتملة الوقوع أو لتحقيق أهداف مرسومة. وقد تكون المشكلات القائمة واضحة ومعروفة الأبعاد والجوانب أو قد تكون غامضة بالنسبة لعمقها وأبعادها والأسباب المكونة لها، أو قد تكون غير موجودة في الأساس لكن حذر الإدارة واستطلاعها للظروف المحيطة تجعلها تتنبأ بتوقع حدوثها. لذلك تقوم الإدارة في كل الحالات التي تستدعي اتخاذ القرارات بتجميع كل ما يلزمها من بيانات ومعلومات وتحليل ما يحيط بها من ظواهر وعوامل مختلفة لتساعدها في الوصول إلى القرار الرشيد بعد تحديد البدائل وتقييمها من أجل أن يكون القرار مناسبا لتحقيق الهدف الذي اتخذ من أجله.

مراحل عملية صنع القرار

صنع القرار يمر بعده مراحل حتى تساعد على اتخاذ قرارات تساعد في عمليات التطوير وحل المشكلات، وتتلخص المراحل في خمس مراحل أساسية يتبعها القائد لصناعة القرار وهي كالتالي:

المرحلة الأولى: تشخيص المشكلة وتحديد أبعادها.

المرحلة الثانية: جمع البيانات والمعلومات، وتصنيفها، وتبسيطها.

المرحلة الثالثة: تحديد ووضع البدائل المتاحة.

المرحلة الرابعة: المقارنة بين البدائل واختيار البديل الأمثل.

المرحلة الخامسة: اتخاذ القرار ومتابعه تنفيذه.

وترتبط هذه المراحل ارتباطا وثيقا مع الإحصاء واستخدام الأساليب الاحصائية ، حيث ان في مرحلة تشخيص المشكلة وتحديد ابعادها يحتاج الى تصميم أدوات جمع بيانات موثوقة وهذا ما تقوم بها الأدوات الإحصائية مثل الاستبانة والمقابلة والملاحظة وغيرها، استخدام البرامج الإحصائية في تبويب وتصنيف البيانات وتحويلها الى معلومات يساعد صانع القرار في التعرف على ابعاد الظواهر والمشاكل التي يتطرق لها، كما ان التحليل الاحصائي المعمق يساعد على صنع سيناريوهات مختلفة لصنع القرار ويساعد على إيجاد افضل الحلول مما يسهل على صانع القرار في اجراء المقارنة والمفاضلة في اختيار البديل الأمثل وبالتالي اتخاذ القرار ومتابعته، عملية صناعة القرار تتضمن كل مراحل القرار التي تبدأ عادة بتحديد المشكلة وتحليل أسبابها، وتعيين متغيراتها بما في ذلك جمع البيانات من مصادرها واستعراض الحلول الممكنة وبناء النماذج أو تصميم الحلول والمفاضلة بينها، ومن ثم اختيار البديل الأمثل وإصدار القرار وتنفيذه.

وعليه تتضح العلاقة بين القيادة وصناعة القرار والأساليب الإحصائية الكمية، حيث إنه لا غنى للقيادات الإدارية عن صناعة القرار، وهي تحتاج لأساليب مساعدة للوصول إلى قرار ناجح ودقيق ولا يخضع للمحاولة أو التخمين وإضاعة الوقت دون الحصول على القرار المناسب، إن الأساليب الإحصائية الكمية تسهم في إعطاء صورة واضحة أمام القيادات الإدارية للاختيار المناسب من بين البدائل المطروحة لحل مشكلة أو ظاهرة إدارية تحتاج لمعالجة من خلال معلومات دقيقة يطبق عليه أساليب إحصائية أكثر دقة وتخصصية وتعطي مؤشرات واضحة، ويسعى الباحث من خلال هذه الدراسة الى استعراض تجارب صنع القرار في المؤسسات الحكومية بسلطنة عمان.

منهج الدراسة

استخدم الباحث المنهج الوصفي التحليلي الذي يقوم على دراسة الظاهرة كما هي في الواقع ووصفها وصفا دقيقا والتعبير عنها كيفيا وكمياً، وتحليل الظاهرة وعمل استنتاجات ومقترحات، وهو ما بعرفه ملحم (2000) بانه احد اشكال التحليل والتفسير العلمي المنظم، لوصف ظاهرة او مشكلة محددة وتصويرها كميا عن طريق جمع بيانات ومعلومات مقننة عن الظاهرة او المشكلة وتصنيفها وتحليلها واخضاعها للدراسة الدقيقة، كما يعرفها الرشيدي (2000) بانها مجموعة الإجراءات البحثية التي تتكامل لوصف الظاهرة او الموضوع اعتمادا على جمع الحقائق والبيانات وتصنيفها ومعالجتها وتحليلها تحليلا كافيا ودقيقا لاستخلاص دلالتها والوصول الى نتائج او تعليمات عن الظاهرة او الموضوع محل البحث.

مجتمع الدراسة

مجتمع الدراسة يشمل وحدات الخدمة المدنية في سلطنة عمان ذات العلاقة المباشرة بإصدار القرارات التي تؤثر في المجتمع على الافراد والموظفين، ويبلغ عدد الوحدات (10) جهات وعليه يبلغ حجم مجتمع الدراسة (1801) موزعا حسب القيادات الإدارية والتي تشمل على (مدراء العموم ومساعديهم، مدراء الدوائر ومساعديهم ورؤساء الأقسام) العاملين في محافظة مسقط، ويوضح الجدول توزيع القيادات الإدارية حسب الجهات.

جدول رقم (1) حصر مجتمع الدراسة حسب المستوى الوظيفي وجهة العمل عن الموقف في نهاية عام 2017

م	جهة العمل	رئيس قسم	مدير دائرة	مدير عام	الإجمالي الكلي
1	وزارة العمل	119	124	25	268
2	وزارة الصحة	164	47	17	228
3	وزارة النقل والاتصالات وتقنية المعلومات	75	54	7	136
4	وزارة الاسكان والتخطيط العمراني	106	58	15	179
5	وزارة التنمية الاجتماعية	62	59	7	128
6	وزارة الاوقاف والشؤون الدينية	47	55	5	107
7	وزارة التربية والتعليم	216	159	33	408
8	وزارة التعليم العالي والبحث العلمي والابتكار	71	73	12	156
9	وزارة التجارة والصناعة وترويج الاستثمار	79	34	6	119
10	وزارة الداخلية	43	24	5	72
	الإجمالي الكلي	982	687	132	1801

المصدر: إحصاء موظفي الخدمة عن الموقف في 31 ديسمبر 2017 م، وزارة الخدمة المدنية يونيو 2018.

عينة الدراسة

نظرا لاختلاف أعداد القيادات الادارية وطبيعة المهام والمسؤوليات التي تتطلب صناعة قرارات معينة من جهة حكومية وأخرى، قام الباحث باختيار عينة الدراسة بالطريقة العشوائية الطبقية. وعليه قام الباحث في هذه الدراسة أولاً: بتقسيم المجتمع إلى طبقات على أساس متغير واحد وهو (جهة العمل) والذي يتكون في 10 مستويات يعبر كل مستوى عن جهة حكومية واحدة، ويمثل طبقة من طبقات العينة بحيث تكون الوحدات متجانسة داخل كل طبقة. ثانياً:

الاختيار العشوائي في داخل كل طبقة حسب أسلوب التوزيع النسبي بقدر الأماكن وهو: كما ذكرها القحطاني (2015) وفيها تكون أحجام العينات المختارة من الطبقات حسب نسبتها في المجتمع، استخدم الباحث معادلة ستيفن ثامبسون لتحديد حجم العينة المناسبة، وعليه يكون حجم العينة (236) ويوضح الجدول رقم (2) توزيع العينة حسب جهات العمل والمستوى الوظيفي.

جدول رقم (2) توزيع عينة الدراسة حسب جهة العمل والمستوى الوظيفي

م	جهة العمل	رئيس قسم	مدير دائرة	مدير عام	الإجمالي الكلي
1	وزارة العمل	16	16	2	34
2	وزارة الصحة	22	6	2	30
3	وزارة النقل والاتصالات وتقنية المعلومات	10	7	1	18
4	وزارة الاسكان والتخطيط العمراني	14	8	2	24
5	وزارة التنمية الاجتماعية	8	8	1	17
6	وزارة الاوقاف والشؤون الدينية	6	7	1	14
7	وزارة التربية والتعليم	28	21	4	53
8	وزارة التعليم العالي والبحث العلمي والابتكار	9	10	2	21
9	وزارة التجارة والصناعة وترويج الاستثمار	10	4	1	15
10	وزارة الداخلية	6	3	1	10
	الإجمالي الكلي	129	90	17	236

أداة الدراسة

استخدم الباحث الاستبانة لتجميع البيانات من افراد عينة الدراسة وسوف تكون مبينه على قسمين، يشمل القسم الأول الخصائص الديموغرافية لإفراد عينة الدراسة والقسم الثاني محاور الدراسة المرتبطة باتخاذ القرارات والأساليب الإحصائية.

قام الباحث بعد الاطلاع على الأدب النظري والدراسات السابقة بناء استبيان تضمن (4) محاور و (19) عبارة وهي كالتالي:

-البعد الاول: محور الاساليب المساعدة لصناعة القرار ويتضمن (5) عبارات.

-البعد الثاني: محور الرغبة بالمعرفة الإحصائية ويتضمن (4) عبارات.

-البعد الثالث: محور معوقات تطبيق الأساليب الإحصائية الكمية ويتضمن (5) عبارات.
-البعد الرابع: محور مقترحات تطبيق الأساليب الإحصائية الكمية ويتضمن (5) عبارات.
ويتم الإجابة عليها حسب مقياس الثلاثي (دائماً-أحياناً-أبداً) ونأخذ الدرجات (3-2-1) على التوالي.
يوضح الجدول رقم (3) مقياس الحكم على فقرات الاستبانة ومحاورها لقياس مستوى الموافقة.

جدول رقم (3) الحدود الدنيا والعليا لمقياس لكيرت الثلاثي

مستوى الموافقة	المتوسط الحسابي
منخفضة	**من 1 الى اقل من 1.66**
متوسطة	**من 1.66 الى اقل من 2.33**
عالية	**من 2.33 الى 3**

الأساليب الإحصائية

استخدم الباحث برنامج التحليل الاحصائي (SPSS) لتحليل البيانات مع استخدام الأساليب الإحصائية التي سوف تناسب بيانات الدراسة ويتوقع ان يتم استخدام الاختبارات التالية:

- معامل الثبات (الفا كرونباخ). لحساب ثبات الاستبانة (Reliability).
- المتوسطات الحسابية والانحرافات المعيارية.
- اختبار تحليل التباين (ANOVA).

التحليل الاحصائي للدراسة

- معامل الثبات

يتضح من الجدول رقم (4) تمتع الاستبانة بمعامل ثبات بلغ (0.692) ، معمل الثبات مقبول لتطبيق الاستبانة على عينة الدراسة.

جدول رقم (4) معامل الثبات الكلي للاستبيان

المحور	عدد الفقرات	معامل الثبات
الكلي	**19**	**0.692**

- تحليل البيانات الديموغرافية

يتضح من الجدول رقم (3) تركز افراد عينه الدراسة حسب المؤهل العلمي بالمؤهلات الجامعية بنسبة (57%)، وحسب المسمى الوظيفي بوظيفة رئيس قسم بنسبه (67%).

جدول رقم (5) الخصائص الديموغرافية لعينة الدراسة

المتغير	المستوى	العدد	النسبة (%)
المؤهل العلمي	اقل من المؤهلات الجامعية	22	9.3
	المؤهلات الجامعية	135	57.2
	الماجستير	45	19.1
	الدكتوراة	34	14.4
	المجموع	236	100.0
المسمى الوظيفي	رئيس قسم ومن في حكمهم	158	66.9
	مدير دائرة	55	23.3
	مدير عام مساعد و أعلى	23	9.7
	المجموع	236	100.0

- النتائج المتعلقة بالإجابة على سؤال الدراسة الأول (ما هو واقع استخدام القيادات الادارية في سلطنة عمان البيانات (الأساليب الإحصائية) في اتخاذ القرارات).

للإجابة على هذا السؤال استخدم الباحث تحليل عبارات فقرات الاستبانة حسب المتوسط الحسابي والانحراف المعياري وذلك لمعرفه مستوى الاستجابة على محاور الاستبانة وفقراتها وترتيب الفقرات حسب الرتبة بناء على المتوسط الحسابي، ومن خلال استجابات القيادات الإدارية بوحدات الخدمة المدنية في سلطنة عمان.

جدول رقم (6) المتوسطات الحسابية والانحرافات المعيارية ودرجة الموافقة لفقرات المحور الاول

رقم الفقرة	الرتبة	الفقرة	المتوسط الحسابي	الانحراف المعياري	درجة الموافقة
1	1	الخبرات الشخصية	2.66	.474	عالية
2	3	الاجتماع بالموظفين	2.33	.563	عالية
3	5	تشكيل لجان	1.63	.580	منخفضة
4	2	الاستشارة	2.43	.496	عالية
5	4	الاساليب الاحصائية	1.91	.612	متوسطة
المتوسط العام للمحور (الأساليب الإحصائية)			2.19	.305	متوسطة

يتضح من نتائج الجدول رقم (6) لمحور الدراسة الأول **(الأساليب الاحصائية)** ان الخبرات الشخصية هو الأسلوب الأول الذي يعتمده القيادات الإدارية بوحدات الخدمة المدنية في صنع القرار وذلك بمستوى عالي وبمتوسط حسابي بلغ (2.66)، يليه أسلوب الاستشارة بمستوى عالي وبمتوسط حسابي بلغ (2.43).

اما الأساليب الإحصائية احتلت الرتبة الرابعة من اصل خمسه أساليب مساعدة لصنع القرار وبمستوى موافقة متوسط حيث بلغ الوسط الحسابي للأسلوب (1.19).

- النتائج المتعلقة بالإجابة على سؤال الدراسة الثاني (مدى رغبة القيادات الإدارية في سلطنة عمان للمعرفة بالأساليب الإحصائية؟).

للإجابة على هذا السؤال استخدم الباحث تحليل عبارات فقرات الاستبانة حسب المتوسط الحسابي والانحراف المعياري وذلك لمعرفه مستوى الاستجابة على محاور الاستبانة وفقراتها وترتيب الفقرات حسب الرتبة بناء على المتوسط الحسابي.

جدول رقم (7) المتوسطات الحسابية والانحرافات المعيارية ودرجة الموافقة لفقرات المحور الثاني

رقم الفقرة	الرتبة	الفقرة	المتوسط الحسابي	الانحراف المعياري	درجة الموافقة
1	2	اساليب وصف البيانات	2.09	.612	متوسطة
2	4	اساليب العلاقات الارتباطية	1.91	.423	متوسطة
3	3	اساليب التنبؤ	1.96	.579	متوسطة
4	1	اساليب المقارنة	2.11	.685	متوسطة
المتوسط العام للمحور (المعرفة بالأساليب الإحصائية)			2.01	.410	متوسطة

من خلال النتائج في جدول رقم (7) يتضح ان هنالك رغبة متوسطة لدى القيادات الإدارية في التعرف على الأساليب الإحصائية التي تساعد على عملية صنع القرار، حيث اختار المستجيبين أساليب المقارنة في الرتبة الأولى لاستخدامه في صنع القرار، ثم أسلوب وصف البيانات من خلال التحاليل الوصفية المتربطة بمقاييس النزعة المركزية والتشتت، وفي الرتبة الثالثة أساليب التنبؤ وأخيرا أساليب العلاقات الارتباطية.

- النتائج المتعلقة بالإجابة على سؤال الدراسة الثالث (ماهي المعوقات التي تواجه القيادات الادارية في استخدام البيانات لاتخاذ القرارات).

للإجابة على هذا السؤال استخدم الباحث تحليل عبارات فقرات الاستبانة حسب المتوسط الحسابي والانحراف المعياري وذلك لمعرفه مستوى الاستجابة على محاور الاستبانة وفقراتها وترتيب الفقرات حسب الرتبة بناء على المتوسط الحسابي، حيث يوضح الجدول رقم (8) اهم المعوقات التي تواجه القيادات الإدارية في استخدام الأساليب الإحصائية لصنع القرارات.

حيث جاء في الرتبة الأولى قلة البرامج التدريبية لتطبيق الأساليب الإحصائية الكمية، ثم عدم توافر اشخاص متخصصين في اداراتهم في مجال الإحصاء، والتحدي الثالث من وجه نظرهم صعوبة استخدام الأساليب الإحصائية الكمية، ثم عدم تشجيع الادارة العليا على تطبيق الاساليب الاحصائية الكمية، وأخيرا انه لا يوجد وقت كافي اثناء صنع القرار في استخدام الأساليب الإحصائية كأسلوب مساعد.

جدول رقم (8) المتوسطات الحسابية والانحرافات المعيارية ودرجة الموافقة لفقرات المحور الثالث

رقم الفقرة	الرتبة	الفقرة	المتوسط الحسابي	الانحراف المعياري	درجة الموافقة
1	2	عدم توفر اشخاص متخصصين في مجال الاحصاء	2.24	.525	متوسطة
2	4	عدم تشجيع الادارة العليا على تطبيق الاساليب الاحصائية الكمية	1.95	.719	متوسطة
3	3	صعوبة التعامل مع الاساليب الاحصائية الكمية[	2.05	.654	متوسطة
4	5	لا يوجد وقت كافي لتطبيق الاساليب الاحصائية الكمية	1.81	.585	متوسطة
5	1	ندرة البرامج التدريبية للأساليب الاحصائية الكمية	2.52	.501	عالية
		المتوسط العام للمحور(المعوقات)	2.11	.340	متوسطة

- النتائج المتعلقة بالإجابة على سؤال الدراسة الرابع (ماهي المقترحات لدى القيادات الادارية لاستخدام البيانات والأساليب الإحصائية في اتخاذ القرارات؟

للإجابة على هذا السؤال استخدم الباحث تحليل عبارات فقرات الاستبانة حسب المتوسط الحسابي والانحراف المعياري وذلك لمعرفه مستوى الاستجابة على محاور الاستبانة وفقراتها وترتيب الفقرات حسب الرتبة بناء على المتوسط الحسابي، حيث يوضح الجدول رقم (9) اهم المقترحات التي تساعد القيادات الإدارية في استخدام الأساليب الإحصائية لصنع القرارات.

جدول رقم (9) المتوسطات الحسابية والانحرافات المعيارية ودرجة الموافقة لفقرات المحور الرابع

رقم الفقرة	الرتبة	الفقرة	المتوسط الحسابي	الانحراف المعياري	درجة الموافقة
1	4	أنشاء وحدة متخصصة (دائرة) في الاحصاء بكل ادارة	2.06	.789	متوسطة
2	3	تعيين موظفين متخصصين في الاحصاء	2.53	.662	عالية
3	1	تدريب الموظفين على الاساليب الاحصائية الكمية	2.91	.291	عالية
4	5	الاستعانة بمراكز الاستشارات الاحصائية الخاصة	1.96	.849	متوسطة
5	2	الاستفادة من تجارب المؤسسات الاخرى في صناعة القرار بإستخدام الاساليب الاحصائية الكمية	2.58	.582	عالية
المتوسط العام للمحور (المقترحات)			2.40	.397	عالية

من خلال النتائج في الجدول رقم (9) افاد المستجيبين الى ان تدريب الموظفين على الأساليب الإحصائية الكمية يعد المقترح الأول لديهم للمساعدة في تطبيق الأساليب الإحصائية لعملية صنع القرار، ثم الاستفادة من تجارب المؤسسات الأخرى التي تستخدم الأساليب الإحصائية في عملية صنع القرار، يليها تعيين موظفين متخصصين في الإحصاء بالدوائر التي يعملون بها.

النتائج المتعلقة بالإجابة على فرضية الدراسة الأولى (هل توجد فروق ذات دلالة إحصائية عند مستوى (0.05) في استجابات افراد عينة الدراسة تعزى الى متغير المؤهل العلمي).

استخدم الباحث اختبار تحليل التباين One way ANOVAللإجابة على هذه الفرضية مثل ما يوضح الجدول رقم (10)، ومن نتائج الجدول انه هنالك دلاله إحصائية لمحاور الدراسة الاربعة، مما يعني ان المؤهل العلمي له تأثير في استجابات افراد عينه الدراسة حول استخدام الأساليب الإحصائية في صنع القرار والرغبة في المعرفة بالأساليب الإحصائية والتحديات والمقترحات لاستخدام الأساليب الإحصائية في صنع القرار، ولمعرفه الفروق لصالح حملة أي مؤهل استخدم الباحث اختبار شيفيه مثل ما توضحه الجداول من (11) الى (14).

جدول رقم (10) نتائج اختبار تحليل التباين (ANOVA) حول اثر المؤهل العلمي على استجابات افراد عينة الدراسة

المحور	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	قيمة (ف)	مستوى الدلالة	اتجاه الدلالة
الأساليب المساعدة لصناعة القرار	بين المجموعات	4.355	3	1.452	19.105	.000	دال احصائيا
	داخل المجموعات	17.628	232	.076			
	المجموع	21.983	235				
الرغبة بالمعرفة الاحصائية	بين المجموعات	10.219	3	3.406	26.824	.000	دال احصائيا
	داخل المجموعات	29.463	232	.127			
	المجموع	39.682	235				
معوقات تطبيق الأساليب الإحصائية الكمية	بين المجموعات	4.525	3	1.508	15.433	.000	دال احصائيا
	داخل المجموعات	22.676	232	.098			
	المجموع	27.202	235				
مقترحات تطبيق الأساليب الإحصائية الكمية	بين المجموعات	4.629	3	1.543	11.019	.000	دال احصائيا
	داخل المجموعات	32.485	232	.140			
	المجموع	37.114	235				

جدول رقم (11) اختبار شيفيه للمحور الأول والمؤهل العلمي

المستوى التعليمي	العدد	Subset for alpha = 0.05		
		1	2	3
الماجستير	45	1.9467		
اقل من المؤهلات الجامعية	22		2.2	
المؤهلات الجامعية	135		2.2178	
الدكتوراة	34			2.4059
مستوى الدلالة		1	0.994	1

من خلال الجدول أعلاه يتضح ان الفروق في هذا المحور لصالح القيادات الإدارية العليا الحاصلين على مؤهل الدكتوراة.

جدول رقم (12) اختبار شيفيه للمحور الثاني والمؤهل العلمي

المستوى التعليمي	العدد	Subset for alpha = 0.05		
		1	2	3
اقل من المؤهلات الجامعية	22	1.625		
الماجستير	45	1.7611		
المؤهلات الجامعية	135		2.0889	
الدكتوراة	34			2.3235
مستوى الدلالة		0.427	1	1

من نتائج الجدول رقم (12) يتضح ان الفروق لصالح القيادات الإدارية العليا الحاصلين على مؤهل الدكتوراة.

جدول رقم (13) اختبار شيفيه للمحور الثالث والمؤهل العلمي

المستوى التعليمي	العدد	Subset for alpha = 0.05		
		1	2	3
الدكتوراة	34	1.8706		
المؤهلات الجامعية	135		2.08	
اقل من المؤهلات الجامعية	22			2.3
الماجستير	45			2.3022
مستوى الدلالة		1	1	1

من نتائج الجدول رقم (13) يتضح ان الفروق لصالح القيادات الإدارية العليا الحاصلين على مؤهل الماجستير.

جدول رقم (14) اختبار شيفيه للمحور الرابع والمؤهل العلمي

المستوى التعليمي	العدد	Subset for alpha = 0.05	
		1	2
الماجستير	45	2.1467	
المؤهلات الجامعية	135		2.4252
اقل من المؤهلات الجامعية	22		2.5
الدكتوراة	34		2.6059
مستوى الدلالة		1	0.219

من نتائج الجدول رقم (14) يتضح ان الفروق لصالح القيادات الإدارية العليا الحاصلين على مؤهل الدكتوراة.

النتائج المتعلقة بالإجابة على فرضية الدراسة الثانية (هل توجد فروق ذات دلالة إحصائية عند مستوى (0.05) في استجابات افراد عينة الدراسة تعزى الى متغير المسمى الوظيفي).

استخدم الباحث اختبار تحليل التباين One way ANOVAللإجابة على هذه الفرضية مثل ما يوضح الجدول رقم (15)، ومن نتائج الجدول انه هنالك دلاله إحصائية لمحاوري الدراسة الثالث والرابع، مما يعني ان المسمى الوظيفي له تأثير في استجابات افراد عينه الدراسة حول التحديات والمقترحات لاستخدام الأساليب الإحصائية في صنع القرار، ولمعرفه الفروق لصالح حملة أي مسمى وظيفي استخدم الباحث اختبار شيفيه مثل ما توضحه الجداول من (16) الى (17).

جدول رقم (15) نتائج اختبار تحليل التباين (ANOVA) حول اثر المؤهل العلمي على استجابات افراد عينة الدراسة

المحور	مصدر التباين	مجموع المربعات	درجة الحرية	متوسط المربعات	قيمة (ف)	مستوى الدلالة	اتجاه الدلالة
الأساليب المساعدة لصناعة القرار	بين المجموعات	.401	2	.201	2.167	.117	غير دال احصائيا
	داخل المجموعات	21.582	233	.093			
	المجموع	21.983	235				
الرغبة بالمعرفة الاحصائية	بين المجموعات	.033	2	.017	.098	.906	غير دال احصائيا
	داخل المجموعات	39.649	233	.170			
	المجموع	39.682	235				
معوقات تطبيق الأساليب الإحصائية الكمية	بين المجموعات	3.383	2	1.691	16.544	.000	دال احصائيا
	داخل المجموعات	23.819	233	.102			
	المجموع	27.202	235				
مقترحات تطبيق الأساليب الإحصائية الكمية	بين المجموعات	2.042	2	1.021	6.784	.001	دال احصائيا
	داخل المجموعات	35.072	233	.151			
	المجموع	37.114	235				

جدول رقم (16) اختبار شيفيه للمحور الثالث والمسمى الوظيفي

المسمى الوظيفي	العدد	Subset for alpha = 0.05	
		1	2
مدير عام مساعد وأعلى	23	1.9043	
مدير دائرة	55	1.96	
رئيس قسم ومن في حكمهم	158		2.1962
مستوى الدلالة		0.716	1

من خلال نتائج اختبار شيفيه بالجدول رقم (16) نلاحظ ان الفروق لصالح رؤساء الأقسام ومن في حكمهم

جدول رقم (17) اختبار شيفيه للمحور الرابع والمسمى الوظيفي

المسمى الوظيفي	العدد	Subset for alpha = 0.05	
		1	2
مدير دائرة	55	2.24	
رئيس قسم ومن في حكمهم	158		2.4468
مدير عام مساعد وأعلى	23		2.513
مستوى الدلالة		1	0.726

من نتائج الجدول رقم (17) الفروق لصالح مدير عام مساعد واعلى.

التوصيات:

بناء على النتائج التي توصلت إليها الدراسة الحالية، يوصي الباحث بما يأتي:

- توعية القيادات العليا بمؤسسات وحدات الخدمة المدنية بأهمية استخدام البيانات والأساليب الإحصائية في عمليه صنع القرار.
- الاستفادة من الخبرات لدى المؤسسات في كيفية تطبيق الأساليب الإحصائية في صنع القرار.
- الاهتمام بتطوير القيادات العليا والموظفين باستخدام البيانات والأساليب الإحصائية من خلال عقد ورش تدريبية متخصصه في تحليل البيانات.

- توفير الكوادر البشرية المؤهلة والمتخصصة في محال الإحصاء لمؤسسات الخدمة المدنية.
- الاهتمام بتطبيق القيادات العليا في مؤسسات الخدمة المدنية الأساليب الإحصائية الكمية في صنع القرار.

المقترحات:

في ضوء النتائج التي تم التواصل إليها، تقترح الدراسة عمل دراسات مستقبلية عن:

- دراسة مقارنه استخدام البيانات الضخمة في صنع القرار لدول عربية مختارة.
- بناء تصور مقترح لتطبيق الأساليب الإحصائية في صنع القرار بالمؤسسات الحكومية في سلطنة عمان.

المراجع:

1. العزب، حسين؛ الحنيطي دوخي، العكعاك، عبد الله (2009). أثر العوامل الاجتماعية والوظيفية على مستوى الثقة الإدارية رؤساء الأقسام في الأجهزة الحكومية في محافظة ظفار في سلطنة عُمان دراسة تطبيقية. مجلة المنارة. المجلد 17. العدد5
2. الريضى، ريما وليد حنا (2016) نماذج التنبؤ الإحصائي واستخداماتها في ترشيد القرارات الإدارية والمالية والاقتصادية في المنظمات. القاهرة: المنظمة العربية للتنمية الإدارية.
3. القحطاني، سعد بن سعيد (2015). الإحصاء التطبيقي. الرياض: معهد الإدارة العامة.
4. رشاد، ساهر محمد، فريدون محمد نجيب (2005)، الأساليب الإحصائية واستخدامها في دعم اتخاذ القرار–الجزء الأول –القيادة العامة لشرطة دبي، مركز دعم واتخاذ القرار، جامعة نايف العربية للعلوم الأمنية، المكتبة الأمنية.
5. المركز الوطني للإحصاء والمعلومات، سلطنة عمان، الكتاب الإحصائي السنوي، أعداد مختلفة.
6. وزارة العمل، سلطنة عمان، التقرير الاحصائي السنوي، أعداد مختلفة.
7. الزهراني، إبراهيم بن حنش سعيد. (2018). القيادة الإستراتيجية وأثرها في تطوير قدرات التعلم التنظيمي "دراسة ميدانية بجامعة أم القرى". المجلة الدولية للأبحاث التربوية، 42(2)، 189–238.
8. الجنابي، صاحب عبد مرزوك. (2019). *استراتيجيات القيادة والإشراف*. عمان: دار اليازوري العلمية للنشر والتوزيع.

9. المربع، صالح بن سعد. (2012). *القيادة الإستراتيجية ودورها في تطوير الثقافة التنظيمية بالأجهزة الأمنية*. المملكة العربية السعودية: جامعة الملك نايف العربية للعلوم الأمنية.

10. عبد الدليمي، عدنان رشيد عواد. (2017). *أثر أنماط القيادة الإستراتيجية على الميزة التنافسية: دراسة ميدانية في الجامعات الأردنية الخاصة*. (رسالة ماجستير غير منشورة)، جامعة آل البيت، الأردن.

دراسة المتغيرات المؤثرة في سمنة النساء باستخدام التحليل العاملي

م.زينب يوسف داود
zainabassfor@uomustansiriyah.edu.iq
الجامعة المستنصرية – كلية الآداب – قسم علم النفس

الخلاصة

أُستخدم الأسلوب الإحصائي المعروف بالتحليل العاملي factor analysis لدراسة مجموعة من المتغيرات الغير قابلة للقياس وأثرها في سمنة النساء في سن الانجاب وانقطاع الطمث للاعمار من 20–50 سنة من خلال جمع البيانات لأكثر من 100 إمراءة تم استبعاد الاجابات الغير مكتملة وتم تحليل 100 استمارة استبيان تم جمعها من مختلف طبقات المجتمع الفقيرة والمتوسطة والغنية افرز التحليل 4 عوامل تشبعت بها متغيرات الدراسة الاثني عشر اذ تم التوصل إلى ان هناك 6 متغيرات تشبعت بالعامل الأول وتم تصنيفها ضمن محور (نوع وكمية الطعام وحرق السعرات) والذي يمكن فيه علاج السمنة و4 متغيرات تشبع بها العامل الثاني تم تصنيفها ضمن محور(الوراثة والهرمونات) التي يكون من الصعب السيطرة عليها وقد أُهملت المتغيرات التي تشبعت بالعاملين الثالث والرابع لعدم أهميتها ضمن الدراسة وتم وضع بعض التوصيات نأمل من الجهات المختصة بهذا الجانب الأخذ بها محاولة لتقليل حالات السمنة التي أصبحت آفة تفتك بالمجتمع .

Studying the variables affecting women's obesity using factor analysis

zainab yousif dawood

Abstract

The statistical method known as factor analysis was used to study a group of non-measurable variables and their impact on obesity in women of childbearing age and menopause for ages 20-50 years by collecting data for more than 100 women. Incomplete answers were excluded, and 100 questionnaire forms collected from various poor, middle, and rich classes of society were analyzed. The analysis revealed 4 factors that saturated the twelve variables of the study. It was concluded that there were 6 variables that saturated the first factor and were classified under the topics (controlling the type and quantity of food and burning calories), in which obesity can be treated, and 4 variables that satisfy the second factor have been classified under the topics (genetics and hormones), which are

difficult to control. The variables that were saturated with the third and fourth factors were neglected because they were not important within the study ،Some recommendations have been made that we hope the authorities concerned with this aspect will take into consideration in an attempt to reduce the cases of obesity, which has become a scourge that is devastating society.

المقدمة

تعد السمنة من أكبر المشكلات الصحية التي يواجهها العالم في الوقت الحاضر فهي حالة طبية خطيرة تحدُث نتيجة التراكم المفرط للدهون أو الأنسجة الدهنية في الجسم مما قد يضر بالصحة، وتعتبر السمنة السبب الأكثر شيوعاً للوفاة حول العالم بعد التدخين حيث تضاعف عدد المصابين بالسمنة الى ثلاث أضعاف منذ عام 1975، ووفقاً لمؤشرات منظمة الصحة العالمية فهناك أكثر من ملياري شخص مصاب بالسمنة ، ولأهمية وخطورة السمنة قامت اكثر من 50 جمعية إقليمية ووطنية تضم أعضاء محترفين من الاوساط الطبية والعلمية والباحثين بانشاء إتحاد سميً (بالإتحاد العالمي للسمنة) وقام الإتحاد بتحديد يوم 4 مارس من كل عام للاحتفال سنوياً وعالمياً لتسليط الضوء على هذا المرض بما يتناسب مع المخاطر الصحية المترتبة علية ، و تُعرّف منظمة الصحة العالمية السمنة بأنها تراكم مفرط أو غير طبيعي للدهون يلحق الضرر بصحة الفرد ، ويعدّ أختلال توازن الطاقة بين السعرات الحرارية التي تدخل الجسم والسعرات الحرارية التي يحرقها هو السبب الرئيسي لزيادة الوزن والسمنة ووفق مقياس مؤشر كتلة الجسم Body Mass Index (BMI) والمتمثل بوزن الشخص بالكيلوغرام مقسوماً على مربّع الطول بالمتر (كغم / متر^2) الذي تعتمدة منظمة الصحة العالمية للبالغين ، فمؤشر كتلة الجسم (BMI) الذي يبلغ 25 فاكثر يصنف انه زيادة في الوزن Overweight والذي يبلغ 30 فاكثر يصنف انها سمنة Obesity ، اما بالنسبة للاطفال دون الخمس سنوات **فزيادة الوزن** هو زيادة نسبة الوزن إلى الطول على انحرافين معياريين فوق قيمة المتوسط المعياري لنمو الطفل الذي تعتمده المنظمة ؛ **والسمنة** هي زيادة نسبة الوزن إلى الطول على 3 انحرافات معيارية فوق قيمة المتوسط المعياري لنمو الطفل الذي تعتمده المنظمة ، اما الأطفال الذين تتراوح أعمارهم بين 5 أعوام و19 عاماً فزيادة الوزن هو زيادة نسبة كتلة الجسم حسب السن على انحراف معياري واحد فوق قيمة المتوسط المرجعي للنمو الذي تعتمده المنظمة ؛ والسمنة هي زيادة بأكثر من انحرافين معياريين فوق قيمة المتوسط المرجعي للنمو الذي تعتمده المنظمة. [1] [3]

مشكلة البحث

السمنة ليست مجرد مشكلة تتعلق بالمظهر الجمالي فقط ، بل إنها مشكلة طبية تزيد من عوامل خطر الإصابة بكثير من الأمراض والمشكلات الصحية الأخرى، كامراض القلب وداء السكري وأرتفاع ضغط الدم وأرتفاع مستوى الكوليستيرول والعديد من الامراض الاخرى والتي قد تؤدي بعض منها الى الوفاة . وقد ارتفع اعداد المصابين بزيادة الوزن والسمنة من الاطفال والبالغين والمسنين بالاونة الاخيرة في جميع بلدان العالم الغنية منها والفقيرة ، وكان فرط الوزن والسمنة يُعدان في السابق من مشكلات البلدان المرتفعة الدخل، ولكنهما الآن يتزايدان في البلدان المنخفضة والمتوسطة الدخل، ولاسيما في البيئات الحضرية . أصبحت السمنة الآن واحدة من أهم الأزمات الصحية العامة التي تواجه جيلنا اليوم في جميع أنحاء العالم ، اذ يستمر عدد الأشخاص الذين يعانون من زيادة الوزن او السمنة في الأرتفاعوإذا استمرت في الاتجاهات الحالية ، فتقدر منظمة الصحة العالمية أنه بحلول عام 2025 سيكون هناك مليار شخص يعانون من البدانة المفرطة و 2.7 مليار من البالغين يعانون من زيادة الوزن ، أصبحت السمنة عند البالغين أكثر شيوعاً من نقص التغذية على مستوى العالم ، وفي العديد من البلدان تقتل السمنة عدد اكثر من عدد الاشخاص الذين يتوفون بسبب نقص الوزن [2].

هدف البحث

نظراُ لأهمية موضوع السمنة وما تسببه من مضاعفات خطيرة تهدد حياة المصاب ولقلة البحوث العلمية في هذا المجال ، هدفت هذه الدراسة الى محاولة لإيجاد أهم العوامل المؤثرة على السمنة باستخدام الاسلوب الاحصائي التحليل العاملي (Factor Analysis) لتصنيف المتغيرات التي تؤثر على زيادة الوزن والسمنة وفق محاور لمعرفة نوع السمنة التي تصاب بها النساء للأعمار من 20–50 سنة حصراً وهو سن الحمل والإنجاب والذي تزاد فيه فرص كسب الوزن الزائد والسمنة محاولة لإيجاد الحلول والتوصيات للعلاج من هذا المرض الخطير، وقد تم استبعاد الاطفال والرجال من هذه الدراسة كون الاطفال دون سن ال5 سنوات والاطفال من سن 5–19 سنة هناك مؤشرات تختلف يجب ان تشمل بها الدراسة .

فرضية البحث

استندت فرضية البحث الى ان هناك تاثير كبير للعوامل الاثني عشر التي ادرجت في استبانة الدراسة على زيادة الوزن والسمنة للنساء في سن الانجاب وانقطاع الطمث .

المبحث الأول : الإطار النظري للدراسة

أولاً : مفهوم السمنة Definition of obesity

تعرف السمنة بأنها تلك الحالة الطبية التي تتراكم فيها الدهون الزائدة بالجسم إلى درجة تتسبب في وقوع آثار سلبية على صحة الفرد ، مؤدية بذلك إنخفاض متوسط عمر الفرد أو إلى الوقوع في مشاكل صحية متزايدة وتعرف أيضاً السمنة او البدانة على أنها تراكم مفرط أو غير طبيعي للدهون والذي يلحق الضرر بصحة الفرد ، ويعدّ السبب الرئيسي لزيادة الوزن (وهي المرحلة ماقبل السمنة) أو السمنة هو اختلال توازن الطاقة بين السعرات الحرارية التي تدخل الجسم والسعرات الحرارية التي يحرقها ، ويمكن ان تقاس السمنة نسبة الى الوزن الطبيعي للبالغين بواسطة مؤشر قياس أرتفاعالدهون في الجسم وبالتالي يتم تصنيف زيادة الوزن أو السمنة من خلال حساب نسبة وزن الشخص بالكيلوغرام إلى مربع طوله بالمتر (كغم / م ^{2}) Body Mass Index (BMI) وفقاً لما حددته منظمة الصحة العالمية عام 1997 ونشرته عام 2000 الا ان هذا المقياس لايمكن اعتماده اذ لايؤخذ مؤشر كتلة الجسم للبالغين بعين الاعتبار بصورة نهائية نتيجة لعدد من العوامل المختلفة ، مثل الجنس والعمر والكتلة العضلية ، فقد يكون مؤشر كتلة الجسم غير دقيق في حالة زيادة الكتلة العضلية، كما هو الحال بالنسبة للرياضيين ولاعبي كمال الأجسام ، حيث يكون مؤشر كتلة الجسم لديهم مرتفعاً ويتم تصنيف الشخص كمصاب بالسمنة رغم انخفاض نسبة الدهون في جسمه. كما ان فقدان الكتلة العضلية مع تقدم العمر، في هذه الحالة يكون مؤشر كتلة الجسم طبيعياً رغم زيادة نسبة الدهون لذا هناك طرق اخرى يلجئ اليها الاطباء والباحثون لقياس السمنة منها محيط الخصر ونسبة الخصر الى الورك (Waist – to–hip ratio) وفرجار ثنية الجلد (Skinfold calipers) وهناك طرق اخرى أكثر تعقيداً منها مقياس امتصاص الاشعة السينية ثنائي البواعث (Dual energy X–ray Absorptiometry) لتقدير نسبة دهون الجسم باستخدام الاشعة السينية والماسح الضوئي (Body 3D Scanner) ثلاثي الأبعاد وهو من الطرق الحديثة المعتمدة على الاشعة تحت الحمراء وطرق اخرى عديدة بالاضافة الى الفحوصات الطبية والمختبرية مثل نسبة الكوليسترول والكلكوز بالدم ووضائف الكبد والغدة الدرقية وغيرها . [2] [3]

ثانياً : أنواع السمنة Types of obesity

يمكن تصنيف السمنة بناءاً على [3]:

أ / **مؤشر كتلة الجسم** : وهو صيغة رياضية للتعرف على الوزن الطبيعي للشخص وهو ناتج قسمة الوزن بالكغم على مربع الطول بالمتر وكما موضح في الجدول ادناه

كتلة الجسم(BMI)	الحالة
أقل من 18.5	أقل من الوزن الطبيعي
18.5-24.9	الوزن الطبيعي
25-29.9	زيادة في الوزن
30-34.9	سمنة الفئة الأولى
35-39.9	سمنة الفئة الثانية
40 49.9-	سمنة شديدة (الفئة الثالثة)
50-59.9	سمنة مفرطة (الفئة الرابعة)
60 فاكثر	سمنة خطيرة (الفئة الخامسة)

تعتبر مرحلة زيادة الوزن هي مرحلة ما قبل السمنة ، ويكون مؤشر كتلة الجسم من 25 إلى 29.9

- النوع الأول من السمنة (منخفض الخطورة) : مؤشر كتلة الجسم ما بين 30-34.9
- النوع الثاني من السمنة (متوسط الخطورة) : مؤشر كتلة الجسم ما بين 35-39.9
- النوع الثالث من السمنة (عالي الخطورة) : مؤشر كتلة الجسم 40-49.9 ، ويسمى هذا النوع بالسمنة المفرطة ويحتاج المريض من هذا النوع إجراء إحدى عمليات السمنة (Bariatric Surgery).
- النوع الرابع من السمنة (شديد الخطورة): مؤشر كتلة الجسم 50-59.9 وتكون اكثر من السمنة المفرطة
- النوع الخامس من السمنة (خطير جداً): مؤشر كتلة الجسم 60 فما فوق، وهي الحالة الأشد خطورة.

ب / موقع تراكم الدهون في الجسم

يمكن تصنيف السمنة بناءاً على منطقة ترسب الدهون وتراكمها كلآتي [5][4]

1- السمنة الطرفية وتسمى أيضاً السمنة الوريدية Venous Circulation Obesity

وهي أحد أنواع السمنة الأكثر شيوعاً عند النساء، وهي حالة تراكم الدهون الزائدة في الوركين، والأرداف، والفخذين، وتعرف هذه الحالة باسم جسم الكمثرى ، ويحدث هذا النوع غالباً نتيجة لأسباب وراثية أو في الأشخاص الذين يعانون من تورم في الساق، خاصةً اثناء الحمل .

2- السمنة المركزية وتسمى أيضاً السمنة الناتجة عن التوتر Nervous stomach Obesity وهي حالة تراكم الدهون في منطقة البطن، ويعتبر هذا النوع الأكثر خطورة لأن الأعضاء الحيوية تقع في هذه المنطقة وقد تؤثر السمنة على إمدادات الدم التي تصل للأعضاء

الحيوية . ويعرف هذا الجسم باسم جسم التفاحة ، وهو أحد أنواع السمنة المنتشرة عند الرجال وتنشر أيضاً لدى النساء نتيجة التوتر والإجهاد والقلق، ويطلق عليها المِعدة العصبية ، وتحُدث عادةً لدى الأشخاص الذين يعانون من مرض القولون العصبي

3- السمنة في المنطقة العلوية من الجسم وتسمى أيضاً السمنة الغذائية Food Obesity

وهي أكثر أنواع السمنة انتشاراً ، حيثُ تتراكم الدهون نتيجة عادات الأكل السيئة والإفراط في تناول أطعمة غير صحية، خاصةً تلك التي تحتوي على نسبة عالية من السكريات والدهون الضارة.

4- السمنة الأيضية الوراثية Genetic metabolic Obesity

يتميز هذا النوع بتراكم الدهون في المعدة وانتفاخها كالبالون ، مما يسبب صعوبة التنفس ويمكن أن يؤثر على اعضاء الجسم الاخرى ، اذ يرتبط هذا النوع عادةً بمشاكل في التمثيل الغذائي وغالباً مايكون نتيجة الإفراط في تناول المشروبات الكحولية .

5- السمنة الناتجة عن حساسية الكلوتين Gluten diet Obesity

يتميز هذا النوع من السمنة بتركيز الدهون في الجزء السفلي من الجسم البطن والفخذين ، وغالباً ما ينتج عن زيادة استهلاك الأطعمة الغنية ببروتين الغلوتين، مثل الخبز والحبوب الكاملة، والتي يصعب على الجسم امتصاصها ، يحدُث هذا النوع أيضاً للنساء بعد انقطاع الطمث أو اللاتي يعانين من اختلالات هرمونية .

6- السمنة الناتجة عن الخمول Inactivity Obesity

يتميز هذا النوع من السمنة بتراكم الدهون في المعدة وأعلى الظهر، وغالبًا ما يتطور نتيجة قلة النشاط البدني والخمول .

ج / القياسات الانثروبومترية[6]

المقصود بمصطلح الانثروبومترية هي مقاييس التي تحدد المعايير المختلفة من الجسم البشري ومنها تقدير سمك طبقات تحت الجلد في أماكن مختلفه من الجسم من خلال قياس ثنيات الجلد في مناطق العضلات ومنتصف الذراع العلوي ومنطقه الصدر والبطن والفخذ وتحت عظمة الكتف ، وهناك العديد من القياسات الانثروبومترية ولكن الأكثرها شيوعاً هي طريقه تقدير سمك الجلد خلف منتصف الذراع بواسطه الفرجار لقياس ثنيات الجلد Skinfold Caliper ، ويلاحظ أن سمك الجلد بين فكي الجهاز عبارة عن سمك طبقتين من الجلد والدهن المخزن تحتها من هذه القياسات يمكن حساب كمية الدهن الإجمالي ونسبته في الجسم الخال من الدهن Lean Body Mass . وقد وجد أن حوالي 50% من دهون الجسم تتجمع تحت الجلد ، لذلك فإن قياس سمك طبقه الدهون تحت الجلد تعتبر مقياساً جيداً لمعرفه رصيد الفرد من الدهون (السمنة) . متوسط

نسبة الدهن في جسم البالغ حوالي 12% في الذكور ، 22% في الإناث . يعتبر الرجل سميناً إذا احتوى جسمه على أكثر من 20% من وزنه دهون ، وتعتبر المرأه بدينه إذا إحتوي جسمها علي أكثر من 30% من وزنها دهون .

ثالثاً : أسباب السمنة Causes of obesity

1- الوراثة (التاريخ المرضي للعائلة) : قد تؤثر الجينات الموروثة عن الأهل في مقدار الدهون التي يخزنها الجسم ، وأماكن توزيع تلك الدهون، وقد تؤدي الخصائص الوراثية أيضاً دوراً في مدى كفاءة الجسم في تحويل الغذاء إلى طاقة ، وكيفية تحكم الجسم في الشهية للطعام، وكيفية حرق الجسم السعرات الحرارية أثناء ممارسة الرياضة ، وتكون السُمنة متوارثة غالباً بين أجيال الأسرة. ولا يرجع سبب ذلك إلى الجينات فحسب، بل غالباً يتشارك أفراد الأسرة الواحدة أيضاً في العادات الغذائية ذاتها وممارسة الأنشطة نفسها .

2- طبيعة النمط الغذائي للفرد أو الأسرة : إن إتباع نظام غذائي غير صحي عالي السعرات الذي يفتقر إلى الفاكهة والخضروات، والمليء بالوجبات السريعة والمُثْخَم بالمشروبات عالية السعرات وحِصَص الطعام الكبيرة للغاية ، بالاضافة الى السعرات الحرارية السائلة التي قد يستهلكها الأشخاص دون شعور بالشبع، وبخاصة تلك الموجودة في الكحوليات والمشروبات الغازية المحلّاة والعصائر كل هذه تسهم في زيادة الوزن والسمنة.

3- قلة النشاط والحركة : إذا كان نمط الحياة غير نشط ، فمن الممكن أن يدخل إلى الجسم بسهولة كل يوم قدر من السعرات الحرارية أكبر مما تحرقه بممارسة الأنشطة اليومية الروتينية كالمشي وصعود السلالم ، وترتبط زيادة عدد الساعات التي يقضيها الشخص أمام الشاشات واستخدام السيارة والنظر باستمرار إلى شاشات الكمبيوتر والأجهزة اللوحية والهواتف ارتباطًا كبيرة بزيادة وزنه.

4- غياب أو قلة ممارسة الرياضة : تُعتبر ممارسة الأنشطة البدنية ، مثل المشي والتمارين الهوائية، أمراً ضرورياً للتحكم في الوزن لأنه يساعد الجسم على حرق السعرات الحرارية إذ إنّ من أهمّ الفوائد التي توفرها ممارسة الرياضة تعود إلى تحسين تركيبة الجسم ورشاقته وصحة عمليّات الأيض، فحتى ان لم يخسر الشخص الوزن عند ممارسة الرياضة ؛ فإنّه سيخسر الدهون، ويبني العضلات في جسمه بدلاً من ذلك.

5- خلل في هرمونات الجسم التي تفرزها الغدد الصماء :

- **هرمونات الغدة الدرقية** والتي تعمل على تنظيم عمليات الأيض والنوم ومعدل نبضات القلب اذ تصاب بعض النساء بقصور الغدة الدرقية الأمر الذي يساهم في زيادة الوزن.

- **هرمون الانسولين الذي يُنتج من قبل البنكرياس** وهو يعمل على نقل الكلوكوز إلى خلايا الجسم لاستخدامها كطاقة أو تخزينها على شكل دهون، للحفاظ على مستوى السكر في الدم ، لكن تناول الكثير من الطعام غير الصحي والمشربات السكرية من شأنه أن يصيب الإنسان بما يعرف باسم مقاومة الأنسولين وتصبح الخلايا غير قادرة على الاستجابة للأنسولين والاستفادة من الكلوكوز، بالتالي يزيد تراكمه في مجرى الدم مؤدياً إلى أرتفاعمستوى السكر، وزيادة الوزن والإصابة بمرض السكري من النوع الثاني .
- **هرمون اللبتين (Leptin)** : وهو الهرمون المسؤول عن شعور الانسان بالشبع للتوقف عن تناول الطعام، بشكل عام تقوم الخلايا الدهنية بإفراز هرمون اللبتين، ولكن تناول الأطعمة الغنية بالسكر وتراكم الدهون بشكل أكبر في الجسم يؤدي إلى إفراز كميات أعلى من هذا الهرمون وهذا يسبب تبلد الجسم وتوقف الدماغ عن تمييز شعور الشبع لتناول كمية أكبر من الدهون من ثم زيادة الوزن.
- **هرمون الغريلين (Ghrelin)** : يدعى أيضاً باسم هرمون الجوع ، تفرزه المعدة والأمعاء الدقيقة والدماغ والبنكرياس، اذ أن إفراز مستويات عالية من هذا الهرمون يسبب زيادة الوزن ، فكلما شعر الانسان بالجوع توجه لتناول الطعام بشكل أكبر.
- **هرمون الكورتيزول (Cortisol)**: الذي تفرزه الغدية الكظرية عندما يكون مستوى هذا الهرمون عالي بالجسم ولفترة طويلة تحدث الإصابة بمتلازمة كوشينغ . ويرجع السبب في ذلك إلى إفراز الجسم للكثير من الكورتيزول أو تناول أدوية تُعرف بالغلوكوكورتيكويد لها نفس تأثير الكورتيزول على الجسم.
- **هرمون الاستروجين** (Estrogen) : ويُفرز من قبل المبيضين إذ يساعد هذا الهرمون على تنظيم التمثيل الغذائي ووزن الجسم، لذا فإن انخفاض مستوى الأستروجين الناتج عن مرحلة ما قبل انقطاع الطمث قد يؤدي إلى زيادة الوزن وبالأخص في الجزء السفلي من الجسم كما أن زيادة مستوى الأستروجين نتيجة اتباع نظام غذائي غني بالأستروجين أو تناول أدوية معينة يمكن أن يسبب أرتفاعمستوى السكر في الدم وزيادة الوزن أيضاً.

6- تناول بعض الأدوية : تؤدي بعض الادوية مثل مضادات الاكتئاب، أدوية السكري ، بعض أدوية الصرع وكذلك بعض وسائل منع الحمل إلى زيادة الوزن.

7- اختلال نظام النوم : عدم الحصول على قسط كافٍ من النوم أو العكس، يمكن أن يسبب تغيرات في الهرمونات التي تزيد من شهية الشخص لتناول المزيد من الطعام .

8- الحمل : زيادة الوزن شائعة أثناء الحمل. وتَجِدُ بعض النساء صعوبة في التخلص من هذا الوزن بعد الولادة . وقد تُسهم زيادة الوزن هذه في إصابة النساء بالسُمنة.

9- الإقلاع عن التدخين: يكون الإقلاع عن التدخين مصحوباً غالباً بزيادة الوزن لان النيكوتين يؤدي الى قلة الشهية للطعام وعند الاقلاع عن التدخين تزداد الشهية للطعام. وبالنسبة للبعض يمكن أن يؤدي إلى زيادة الوزن بما يكفي لتشخيصها على أنها سُمنة .

10- العوامل النفسية : العواطف السلبية لها تأثير قوي على عادات الطعام فكثير من الناس يأكلون بنهم أستجابة لمشاعر معينة مثل الملل أو الحزن أو التوتر او الخوف .

11- العوامل الاجتماعية : يمكن ان يكون العيش مع اشخاص يتبعون نظام غذائي غير صحي او أن نقص المال لشراء الأطعمة الصحية أو عدم وجود أماكن آمنة للمشي أو ممارسة الرياضة إلى زيادة خطر الإصابة بالسمنة.

12- الإصابة ببكتريا الامعاء والمعدة : هناك **أجزاء** من بكتيريا الأمعاء اسمها السموم الداخلية "إيندوتوكسين" (endotoxins) يمكنها التسرب إلى مجرى الدم وإلحاق الضرر بالخلايا الدهنية مما يؤدي لزيادة الوزن.

رابعاً : الآثار المترتبة على السمنة Effects of obesity

للدهون الزائدة آثار سلبية على الجسم كله ، مما يتسبب في الإصابة بالامراض وظهور علامات وأعراض غير مريحة منها :

- الإصابة بأرتفاع ضغط الدم وأرتفاع الكوليسترول إلى مستويات غير صحية، وهي عوامل خطورة تُسبب الإصابة بأمراض القلب والسكتات الدماغية.
- الإصابة بداء السكري من النوع الثاني ، يمكن أن تؤثر السمنة في طريقة توظيف الجسم للأنسولين من أجل التحكم في مستويات السكر في الدم. وهو الأمر الذي يزيد خطر مقاومة الأنسولين والإصابة بالسكري.
- الإصابة بانواع معينة من السرطان. كسرطان الرحم وعنق الرحم وبطانة الرحم والمبيض والثدي والقُولون والشرج والمريء والكبد والمرارة والبنكرياس والكلى والبروستات.
- ضيق التنفس وصعوبة التنفس نتيجة ضغط وزن البطن على الرئتين.
- آلم وضغط على المفاصل وعلى أجزاء الجسم المختلفة ، وخاصة في الظهر والساقين والركبتين والكتفين نتيجة الجهد المفرط الذي يبذله الجسم لدعم الوزن .
- صعوبة في بذل المجهود أو المشي بسبب زيادة الوزن .
- التهاب الجلد والالتهابات الفطرية ، نتيجة تراكم العرق والأوساخ في ثنايا الجسم .
- ظهور بقع داكنة على الجلد ، وخاصة الرقبة والإبط وهو رد فعل ناتج عن مقاومة الأنسولين ، أو مقدمات السكري والتي تسمى الشواك الأسود.

- العجز الجنسي والعقم ، بسبب التغيرات الهرمونية وصعوبة تدفق الدم في الأوعية.
- الشخير الليلي وانقطاع التنفس أثناء النوم نتيجة تراكم الدهون في الرقبة والممرات الهوائية .
- زيادة الميل إلى الدوالي والقرحة الوريدية بسبب التغيرات في الأوعية الدموية والدورة الدموية .
- القلق والاكتئاب بسبب عدم الرضا عن صورة الجسم والشراهة عند الأكل .

المبحث الثاني : الجانب التطبيقي

أولاً : جمع البيانات

تم جمع بيانات الدراسة بواسطة أستمارة أستبيان أُعدت من قبل الباحث تم توزيعها على الفئة المستهدفة من النساء للأعمار من 20–50 سنة حصرياً المصابات بزيادة الوزن والسمنة بعد التاكد من حساب مؤشر الكتلة BMI من مختلف الطبقات الاجتماعية الفقيرة والمتوسطة والغنية ، اذا تم توزيع اكثر من 100 استمارة تم استبعاد الاستمارات الغير مكتملة الاجابة وتم تحليل 100 أستمارة إحصائياً باستخدام التحليل العاملي Factor Analysis في برنامج SPSS كونه التحليل المناسب للبيانات التي تم جمعها للتحقق من فرضية البحث بايجاد أهم العوامل المؤثرة في سمنة النساء في سن الانجاب وانقطاع الطمث كون المتغيرات الداخلة في الدراسة يصعب قياسها .

ثانياً : وصف البيانات

تم ادخال البيانات بعد ترميزها في برنامج SPSS وكانت ترمز كالاتي

رمز المتغير	وصف المتغير	رمز المتغير	وصف المتغير
X1	**النشاط والحركة اليومية**	**X7**	**خلل في الهرمونات**
X2	**ممارسة النشاط الرياضي**	**X8**	**الوراثة**
X3	**نمط الغذاء**	**X9**	**الاقلاع عن التدخين**
X4	**نظام النوم**	**X10**	**الحمل**
X5	**العوامل الاجتماعية**	**X11**	**تناول الادوية**
X6	**العوامل النفسية**	**X12**	**الإصابة ببكتريا الامعاء**

تحليل النتائج

بعد تبويب البيانات وادخالها في برنامج SPSS افرز البرنامج النتائج التالية

1- أختبار KMO لجودة القياس : نلاحظ قيمة أختبار (Kaiser- Meyer-Olkin) تساوي 0.760 وهي قيمة مقبولة اي ان التحليل العاملي قام باختزال العوامل بجودة جيدة اذ بلغت القيمة المعنوية للاختبار 0.000

جدول رقم 1

اختبار KMO اختبار جودة قياس التحليل العاملي

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	452.792
	df	66
	Sig.	.000

2- الجذر الكامن (Eigen Value) : ويمثل مجموع مربعات إسهامات كل المتغيرات على كل عامل من عوامل المصفوفة كلاً على حدة ، والعوامل الأولى هي ذات الجذر الكامن الأكبر مما يليها ويجب ان يكون اكبر من الواحد .

جدول رقم 2

قيم عوامل الجذر الكامن

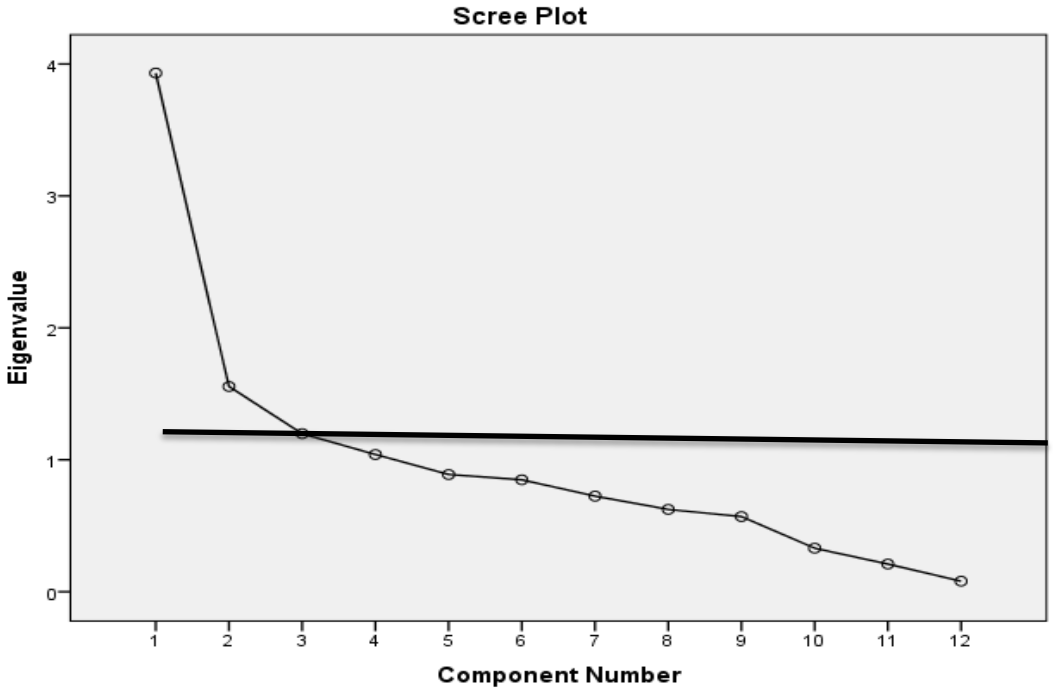
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.932	32.763	32.763	3.932	32.763	32.763	3.809	31.743	31.743
2	1.556	12.963	45.726	1.556	12.963	45.726	1.551	12.926	44.669
3	1.198	9.984	55.710	1.198	9.984	55.710	1.275	10.622	55.291
4	1.040	8.668	64.379	1.040	8.668	64.379	1.091	9.088	64.379
5	.888	7.404	71.783						
6	.848	7.071	78.853						
7	.725	6.039	84.893						
8	.624	5.198	90.091						
9	.570	4.746	94.837						
10	.330	2.750	97.587						
11	.210	1.749	99.336						
12	.080	.664	100.000						

Extraction Method: Principal Component Analysis.

نلاحظ ان التحليل افرز 4 عوامل ، فالعامل الأول كانت قيمة جذره الكامن 3.932 وبنسبة تباين 31.743% من التباين الكلي بعد تدوير العوامل عمودياً بطريقة Varimax ، اما العامل الثاني فكانت قيمه جذره الكامن 1.556 وبنسبة تباين قام بتفسيرها 12.926% من التباين الكلي ليصبح مجموع تفسير العاملين الأول والثاني بنسبة تباين 44.669% من التباين الكلي ، اما العامل الثالث فقيمة جذره الكامن كانت 1.198 ونسبة التباين التي قام بتفسيرها 10.622% من التباين الكلي ليبصح مجموع التباين الكلي للعوامل الأول والثاني والثالث 55.291% من التباين الكلي ، في حين كانت قيمة الجذر الكامن للعامل الرابع 1.040 ونسبة التباين التي قام بتفسيرها 9.088% من التباين الكلي فاصبح مجموع التباين المفسر من قبل العوامل الاربعة مساوي الى 64.379% من مجموع التباين الكلي الحاصل في بيانات الدراسة وتعد نسبة جيدة لتفسير العوامل الاربعة .

3- **الرسم البياني Scree plot** : يبين قيم الجذور الكامنة لكل عامل على المحور الصادي Y-axis ورقم المكون Component على المحور السيني X-axis ويعتبر معيار اخر يمكن استخدامة بالاضافة الى معيار الابقاء على العوامل التي يزيد جذرها الكامن عن الواحد صحيح لتحديد العوامل في التحليل العاملي والابقاء فقط على العوامل التي تكون في المنطقة شديدة الانحدار .



قيم الجذور الكامنة بالنسبة الى العوامل

يتضح من الرسم ان هناك أربع عوامل أكبر من الواحد صحيح وبقية العوامل أقل من الواحد صحيح

4 – مصفوفة العوامل بعد التدوير : ومن خلالها يمكن معرفة تشبع كل متغير على اي عامل من العوامل التي افرزتها نتائج التحليل واي عامل لدية علاقات اكبر من 0.30 مع ثلاث متغيرات او أكثر يمكن أعتباره مكون جيد للأخذ به وفي حالات (Over load) ناخذ القيمة الأكبر.

جدول رقم 3

مصفوفة تشبع المتغيرات بعد تدوير العوامل

Rotated Component Matrix[a]				
	Component			
	1	2	3	4
حركة	.954			
رياضة	.893			
نوم	.865			
غذاء	.842			
اجتماعية	.602		-.334	
تدخين	-.420	-.411		
حمل		.913		-.315
وراثة		.722		
ادوية		.720	.470	
هرمونات		.657		
نفسية			.600	
بكتريا				-.394
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.[a] a. Rotation converged in 8 iterations.				

نلاحظ من مصفوفة التدوير ان

العامل الأول لديه علاقات قوية مع 6 متغيرات من اصل 12 متغير

العامل الثاني لديه علاقات قوية مع 4 متغيرات من اصل 12 متغير

العامل الثالث لديه علاقة قوية مع متغير واحد فقط من اصل 12 متغير

العامل الرابع لدية علاقة قوية مع متغير واحد فقط من اصل 12 متغير
وعلية يمكن ترتيب المتغيرات الاكثر تاثيراً على السمنة وفقاً للتحليل العاملي كما في الجدول ادناه

جدول رقم 4

ترتيب المتغيرات حسب الاهمية وفقاً للعوامل

العامل	وصف المتغير	قيمة التشبع بعد التدوير
1	روتين الحركة و النشاط اليومي	0.954
1	ممارسة النشاط الرياضي	0.893
1	نظام النوم	0.865
1	نمط الغذاء	0.842
1	العوامل الاجتماعية	0.602
1	الاقلاع عن التدخين	0.420
2	الحمل	0.913
2	الوراثة	0.722
2	تناول بعض الادوية	0.720
2	خلل هرمونات	0.657
3	العوامل النفسية	0.600
4	الإصابة ببكتريا الامعاء	0.394

يتبين من أعلاه ان جميع المتغيرات ذات تاثير واضح على السمنة عند النساء في مرحلة الانجاب وانقطاع الطمث اذا يتم تحديد أولوية تاثير هذه المتغيرات بالإعتماد على مقدار تشبع المتغيرات بالعوامل .

الاستنتاجات

من خلال ماتم سرده في الدراسة عن السمنة من مصادر متعددة تم اعتمادها من دراسات وتقارير علمية مختلفة ومن تحليل البيانات احصائياً وفق التحليل العاملي ل100 سيدة يعانين من الوزن الزائدة والسمنة للأعمار من 20–50 سنة تم الوصول الى النتائج الاتية :

1– قلة الحركة والمشي والنشاط اليومي بالاعتماد على السيارات والمصاعد الكهربائية والجلوس لساعات طويلة امام اجهزة التلفاز والهاتف هو **المتغير الأول الأهم في العامل الأول** يليه عدم ممارسة النشاط الرياضي **المتغير الثاني الأهمية في العامل الأول** ، عدم أنتظام النوم من خلال السهر لساعات الليل المتاخرة تزيد من تناول اطعمة او مشروبات عالية السعرات هو **المتغير**

الثالث في الأهمية للعامل الأول ، الإفراط في تناول الطعام وتفضيل الأطعمة المشبعة بالدهون والوجبات السريعة والغنية بالسكريات والاملاح والأبتعاد عن تناول الخضروات والفواكه كان **المتغير الرابع في الاهمية للعامل الأول** ، العوامل الاجتماعية كالسير على خطى الاهل والاصدقاء في البيت والعمل باتباع عادات غذائية سيئة هي **المتغير الخامس في الأهمية للعامل الأول** ، الاقلاع عن التدخين عالرغم من انه من العادات الجيدة الا انه قد يسبب زيادة الوزن والسمنة بسبب نقص النيكوتين في الجسم الذي يقلل من الشهية للاكل هو **المتغير السادس والأخير في الأهمية المؤثر في العامل الأول** .

يمكن إدراج المتغيرات ال6 المشبعة بالعامل الأول تحت محور (**نوع وكمية الطعام وحرق السعرات**) والتي يمكن السيطرة عليها من قبل المصاب بدون اللجوء الى الأدوية والتداخلات الجراحية

2- الحمل هو **المتغير الأول الاكثر أهمية في العامل الثاني** اما الوراثة والجينات المسببة لزيادة الوزن والسمنة هي **المتغير الثاني في الأهمية في العامل الثاني** ، تناول بعض الادوية كمضادات الاكتئاب والكورتزون هي **المتغير الثالث في الأهمية في العامل الثاني** ، الخلل في انتاج بعض هرمونات الجسم هي **المتغير الرابع والأخير المؤثر في العامل الثاني** .

يمكن ادراج المتغيرات ال4 المشبعة بالعامل الثاني تحت محور (**الهرمونات والوراثة**) والتي يصعب السيطرة عليها من قبل المصاب .

3- العاملين الثالث والرابع قد تشبع كل منهما بعامل واحد لذا يمكن اهمالها .

التوصيات

1- الاهتمام بنشر الوعي الصحي من خلال البرامج والبوسترات ووسائل التواصل الاجتماعي فيما يخص علاج و مكافحة السمنة للسيطرة و الحد من تزايدها.

2- الاهتمام بحصة الرياضة في المدارس والتوعية بخطورة السمنة من خلال إدراج حصة توعوية وارشادية في المدارس والجامعات.

3- فتح مراكز للتوعية بمخاطر السمنة مابعد الانجاب للنساء وقاعات للانشطة الرياضية والاهتمام بها وجعلها من اولويات الاهتمام بالمراءة.

4- تناول البحث فئة معينة من المصابين بالسمنة وهي النساء ، يوصي الباحث بتناول دراسات الإصابة بالسمنة لدى الاطفال كونه موضوع يستدعي القلق للحد منه.

المصادر والمراجع

1- الإتحاد العالمي للسمنة . مقدمة للسمنة .(8 /7/ 2023)
https://www.worldobesity.org/ar/patient-portal/resources

2 - كتاب السمنة . الصحة والسكري . مجلة دورية تصدر عن المركز الوطني للسكري والغدد الصم والوراثة.عمان . الاردن

3- منظمة الصحة العالمية . السمنة وفرط الوزن . 9/ 6 / 2021
https://www.who.int/ar/news-room/fact-sheets/detail/obesity- and- overweight

4- Anderson , T.W .(1984).(An introduction to multivariate statistical analysis) , 2nd edition , john Wiley & sons ,new York.

5- Lumish, H. S., et al. (2020). Sex Differences in Genomic Drivers of Adipose Distribution and Related Cardio metabolic Disorders: Opportunities for Precision Medicine.

6- Obesity: Preventing and managing the global, (Who) (2000) *(PDF)*. Geneva: World Health Organization. *ISBN*:*92-4-120894-5*. *الأصل* مؤرشف من *(PDF)*في 01-05-2015 .

7- Power, M., & Schulkin, J. (2008). Sex differences in fat storage, fat metabolism, and the health risks from obesity: Possible evolutionary origins. British Journal of Nutrition.

8- Rencher, A. C (2002) ,"Methods of Multivariate Analysis" , Second Edition, John Wiley & sons, New York , USA,..

Survival analysis of Brain Cancer in Erbil-Kurdistan/Iraq

Prof. Dr. Kurdistan Ibrahim Mawlood
kurdistan.mawlood@su.edu.krd
Chnar Smko Abdullah chnar.abdullah@su.edu.krd
Salahaddin University – Erbil College of Administration and Economics/ Statistics Department

The statistical methods for survival analysis of data have found applications in wide range of fields especially in medical researches during the past few decades. The majority of medical researches focus on the event of time to death. However, another crucial factor in cancer is the amount of time that passes between a treatment response and a recurrence or period of disease-free time. This research is aimed to study the most important factors affecting the brain cancer in Erbil city using the log rank test to detect a difference in a risk of an event between factors group. Using cox proportional model for modeling and identifying the most affecting factors of brain cancer in our data. The data used in this research was obtained from Nanakali Hospital for Cancer in the Kurdistan Region of Iraq – Erbil. The results for our data showed that Morphology, Behavior and age are the most important factors affecting the brain cancer.

Key words: Survival Analysis, Log Rank Test, Cox Proportional Model, Hazard Function, Brain Cancer.

تحليل البقاء لسرطان الدماغ في أربيل – كوردستان / العراق

أ. د. كوردستان ابراهيم مولود
چنار سمكۆ عبدالله
جامعة صلاح الدين –اربيل كلية الادارة و الاقتصاد/ قسم الاحصاء والمعلوماتية

الأساليب الإحصائية لتحليل بقاء البيانات وجدت تطبيقات في مجموعة واسعة من المجالات خاصة في البحوث الطبية خلال العقود الماضية. تركز غالبية الأبحاث الطبية على حدث الوقت حتى الموت. ومع ذلك ، هناك عامل حاسم آخر في الإصابة بالسرطان وهو مقدار الوقت الذي يمر بين استجابة العلاج وتكرار أو فترة خالية من المرض. يهدف هذا البحث إلى دراسة أهم العوامل التي تؤثر على سرطان الدماغ في مدينة أربيل باستخدام اختبار التصنيف اللوغاريتمي لاكتشاف الاختلاف في خطر وقوع حدث بين مجموعة العوامل. استخدام نموذج كوكس النسبي لنمذجة وتحديد العوامل الأكثر تأثيراً لسرطان الدماغ تم الحصول على البيانات المستخدمة في هذا البحث من مستشفى نانكلي للسرطان في إقليم كوردستان العراق - أربيل. أظهرت النتائج أن التشكل والسلوك والعمر من أهم العوامل التي تؤثر على سرطان الدماغ.

الكلمات المفتاحية: تحليل البقاء ، اختبار التصنيف اللوغاريتمي ، نموذج كوكس النسبي ، دالة المخاطرة ، سرطان الدماغ.

1. Introduction

The study of time-to-event data is referred to as survival analysis. The time to event data displays the period of time from a well specified time origin to a clearly defined end point of interest (event). Although time-to-event analysis and time-to-event data are often used more frequently than survival analysis and survival data, the latter word is clearer and more exact.

The beginning and end of time must be clearly identified. For instance, the diagnosis of a specific type of cancer is chosen as the time origin and the death carried on by that specific cancer would be the time end point in a research of that type of cancer. Alternately, a research might monitor subjects from the time of their birth (time of origin) until the start of a disease (end point). The measurement of time is performed in this way. The data on the time to occurrence is typically gathered prospectively over time, such as when data was collected for a clinical experiment or a future cohort study. Sometimes information can be gathered retrospectively by consulting medical records or speaking with people who have a particular disease. (Khawar, 2019).

The World Health Organisation (WHO) reported in 2018 that an estimated reason of one in six deaths is due to cancer. Generally, cancer is a type of disease that causes uncontrollable cell growth, division, and uncontrollably. Brain cancer develops when abnormal cells grow within the brain. All types of brain tumors may cause symptoms that vary depending on the size of the tumor and the portion of the brain that is affected.

In (2019) Mawlood used two advanced statistical methods for studding the most important factors affecting the leukemia in Erbil city, logistic regression and Cox regression. The results indicated that the surgery is the most important factor affecting the leukemia survival patients in both methods.

Mawlood et al (2019) used two Survival models (Cox-Proportional Hazard and Accelerated Failure Time Models) to detect the significant factors effecting on chest cancer. They found significant difference between levels of treatments, namely: Surgery, Radio, Age and Gender

2. Background Information

This section presents the brain cancer which is the first important health issue. Two functions related to survival analysis are used to describe the data; the survival function, which measures the possibility that a patient

will survive to time t and the hazard rate also known as the hazard function, measures the possibility that the patient will die in the future instant of time. Moreover, exploration, description and the basic principles of two models (Cox regression models and Poisson regression models) given and the Log rank test to compares survival of two different groups of individuals

2.1 Survival Analysis

The study of how often and when events happen is known as survival analysis. To find out how covariates affect the length of survival, covariates are studied. Survival analysis is the only method that uses censoring and time-dependent covariates (time-varying explanatory variables). (Mahdi, 2016)

Data involving time till the occurrence of a specific event has been referred to as "survival data" in a wide sense. The occurrence of a tumor, the development of a disease, its recurrence, conception, the decision to stop smoking, and other events may fall under this category. Applications of the statistical methods for survival data analysis have been expanded in recent years beyond biomedical and reliability research to other fields, including longevity of electronic devices, components, or systems (reliability engineering), length of first marriage (sociology), length of newspaper or magazine subscription (marketing), health insurance practice, business, and economics. In recent years, risk and/or prognostic factors connected to response, survival, and the onset of a disease have all been identified. (Inger, 2002)

2.1.1: Survival Function

Let T represent an individual's survival length with density f. The distribution function or cumulative distribution function of T, $F(x) = \int_0^x f(u)$ du and the density function the chances of surviving at a specific time point are not very well-explained at a given time point. Instead, the density and distribution functions are employed along with the survival, hazard, and cumulative hazard functions.

The survival function denotes S(t), is the probability that an individual will survive past time t: (Ameri, 2015)

$$S(t) = p_r(T \geq t) = 1 - F(t) = \int \mathrm{f(x)dx} \qquad \ldots 1$$

Where :

T is a non-negative random variable denotes the time of occurring an event. Then, we can express the p.d.f as:

$$f(t) = \frac{\partial F_{(t)}}{\partial t} = -\frac{\partial S_{(t)}}{\partial t} \qquad \ldots 2$$

And the mean life time denote μ is defined as follows:

$$\mu = \int_0^{\infty} tf(t)\ \mathrm{dt}. \qquad \ldots 3$$

Integrating by parts and taking into account that $\frac{\partial S_{(t)}}{\partial t} = -f(t)$ and also $S(0)=1$ and $S(\infty)= 0$, the mean life time can be re-expressed as follows (Sawadogo, 2018):

$$\mu = \int_0^{\infty} S(t)\ \mathrm{dt}. \qquad \ldots 4$$

2.1.2: Main Goal of Survival Analysis

The main motivation behind the study of survival analysis is the following:

• Provide an estimation of the survivor and hazard functions

• Compare the survival time between group so find visuals using significance tests.

• Fitting survival models using the relevant covariates in the data by using parametric or semi-parametric methods. (Sawadogo, 2018)

2.1.3: Hazard Function

A technique to model the distribution of data in a survival analysis is to use the hazard function, also known as the force of mortality, instantaneous failure rate, instantaneous death rate, or age-specific failure rate. The function's most typical application is to model how an individual's risk of death changes with age. To model any other interesting time-dependent event, though, it can be applied. (Der & Everitt, 2007)

There are various interesting advantages of the Hazard function. In the beginning, it states whether the events happen and, if so, when. It is possible to determine directly the risk of an event happening at a specific time. Higher risk is implied by higher hazard. Second, the computations take both censored and uncensored cases into account. Third, unlike Cox regression, discrete-time survival analysis does not discard information about variations in the event's timing.

suppose the survival time T is such that $t \leq T$, $t + \delta t$, then this probability can be expressed as:

$P(t \leq T < t + \delta t | T \geq t)$

Dividing by the interval length δt and by evaluating the limit of this conditional probability at δt approaches zero, we obtain a rate which defines the hazard function.

That is, the Hazard is given by:

$$h_{(t)} = \lim_{\Delta t \to 0} \left[\frac{pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right]$$

$$h_{(t)} = \lim_{\Delta t \to 0} \left[\frac{pr(t \leq T \leq t + \Delta t) | pr(T \geq t)}{\Delta t} \right]$$

$$h_{(t)} = \lim_{\Delta t \to 0} \left[\frac{[F(t+\Delta t) - F(t)]|\Delta t)}{S(t)} \right]$$

$$h_{(t)} = \frac{\partial F(t)|\partial t}{s(t)}$$

$$h_{(t)} = \frac{f(t)}{s(t)} \qquad \ldots 5$$

The Hazard function is also called conditional failure rate. The Hazard function gives the failure Hazard per unit time during the operation and plays an important role in the data. Nevertheless, in practice, when there is no controlled observation, the hazard function is the percentage of patients who die per unit time, knowing that they have survived to the beginning of the period: (Hout, 2017)

$$\mathrm{h(t)} = \frac{\text{number of patients dying per unit time of the interval}}{\text{number of patients surviving at t}} \qquad \ldots 6$$

2.2: Censoring

In survival analysis, the survival time of an individual is said to be censored if the event of interest has not been observed for that individual. This may be due to the fact that the survival data is analyzed at a point in time when some individuals are still alive. It could also be due to the fact that some individuals have voluntarily left the experimental or clinical trial without notice. A survival time could also be considered as censored if event of interest death for example was due to a cause that is known to be unrelated to the treatment. Suppose an individual has been recruited into an experimental study at an initial time t_0 dies at time t_0 +t with t unknown due to the fact that the individual is still alive or due to not following-up. If the individual was still alive at time t_0 + c, c > 0, then the time c is called a censored survival time. (Sawadogo, 2018)

- **2.2.1: Type I Censoring:** Type I censoring occurs when the censoring time is fixed and under the control of the investigator. In that case, even observations that are not censored are said to have a censoring time. (Ihwah, 2015)

- **2.2.2: Type II Censoring:** Type II censoring occurs when observation is terminated after a predetermined number of events have occurred.

- **2.2.3: Right Censoring:** A survival time is said to be right censored if the time of the unobserved event of interest is known to be greater than a fixed time. That is the time of the event of interest is to the right of the censored time. This type of censoring often occurs when the

study ends without observation of the event of interest for all the individual in the study.

- **2.2.4: Left Censoring:** Left censoring happens when the actual survival time of an individual is less than some fixed value. This type of censoring is most likely to occur when you begin observing a sample at a time when some of the individuals may have already experienced the event. (Samartzis, 2006)

- **2.2.5: Interval Censoring:** is applied to the data in the sense that some transition periods are not seen but are known to fall within a particular time interval. The beginning of dementia, for example, is latent, but when longitudinal data is provided, the onset may be determined to occur within the time span given by two consecutive observations. (Harrell, 2017)

- **2.2.6: Random Censoring:** censoring occurs when observations are terminated for reasons that are not under the control of the investigator. (Sawadogo, 2018)

2.3: The Log Rank Test:

The log-rank test (also known as the Mantel log-rank test, the Cox Mantel log-rank test, and the Mantel Haenszel test) is one commonly used non-parametric test for comparing two or more survival distributions of the patients. This approach is also helpful to identify differences between groups when the risk of an event is consistently higher for one group than another.

Setting up the survival time for both the censored and observed times is the first step in completing this procedure. The log rank test, a one-degree-of-freedom variant of the Chi-square test distribution (Singh, and Mukhopadhyay, 2011) calculates a test statistic used for testing a null hypothesis. The null hypothesis, according to which all groups' survival curves are identical, was put to the test using the log rank test. (Dakhil, et al., 2012)

H_0 : There are no differences between survival curves.
H_1 : There are differences between survival curves.

$$e_{1j} = \frac{n_{1j}}{n_{1j}+n_{2j}} * \left(m_{1j} + m_{2j}\right) \quad \dots 7$$

$$e_{2j} = \frac{n_{2j}}{n_{1j}+n_{2j}} * \left(m_{1j} + m_{2j}\right) \quad \dots 8$$

where:
e_{1j} : represents the expected number of events in group one.
e_{2j} : represents the expected number of events in group two.
n_{1j} : represents the number of individual at risk in group one.
n_{2j} : represents the number of individual at risk in group two.
m_{1j} :is the number of failures in group one.
m_{2j} :is the number of failures in group two.
Here, the data is separated into J categories, denoted by the j= 1, 2,..., J. The additional discrepancy between the observed and expected number of fails in each category is represented by

$$O_i - E_i = \sum_{j=1}^{j} \left(m_{ij} - e_{ij}\right) \quad \ldots . 9$$

with variance

$$var(O_i - E_i) = \sum_{j=1}^{j} \frac{n_{1j}n_{2j}\left(m_{1j}+m_{2j}\right)\left(n_{1j}+n_{2j}-m_{1j}-m_{2j}\right)}{\left(n_{1j}+n_{2j}\right)^2\left(n_{1j}+n_{2j}-1\right)} \quad \ldots 10$$

Here:

$E_1 = \sum_{j=1}^{j} \left(e_{1j}\right)$, represents the expected number of events in group one,

$E_2 = \sum_{j=1}^{j} \left(e_{2j}\right)$, represents the expected number of all events in group two,

O_1 :is the observations number in the first group,

O_2 :is the observations number in the second group,
and J is the end time of the study.
Based on the equations (7, 8, 9 and 10) then the log-rank test method gives:

$$\text{Log-rank test statistic} = \frac{(O_1-E_1)^2}{var(O_1-E_1)} + \frac{(O_2-E_2)^2}{var(O_2-E_2)} \quad \ldots 11$$

The log-rank test is used to determine whether the survival times between two groups differ statistically significantly, but it does not examine the impact of the other independent variables. (Mahdi, 2016)

2.4: Cox Regression Model

The Cox model proposed by Cox (1972) the most popular multivariate method for examining survival time data in medical researchs. The Cox's model enables an analysis in which the explanatory variables can take

either a continuous scale or a categorical form and survival time is considered as a continuous variable. The Cox model is a technique for analyzing the impact of several variables on the period of time it takes for an event to occur. The baseline hazard function and the impact of the covariates on the hazard are both simple multiplicative factors in the Cox's model. The definition of the baseline hazard is the hazard function for that particular person with zero for all factors. Because semi-parametric models require fewer assumptions than parametric models, researchers in the medical sciences commonly prefer them. (Singer & Willett, 1991)

The Cox model can be either one of the following forms: (Fox, 2002)

$h_i(t,x) = h_0(t)e^{\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}$, $t \geq 0$, x = 0,1; $-\infty < x, \beta < \infty$ … 12

Or

$$h_i(t,x) = h_0(t)\, e^{\beta^T x_i}. \quad \ldots 13$$

When;

$$e^{\beta x} = \frac{h_1(t)}{h_0(t)}, \forall t \geq 0. \quad \ldots 14$$

where regression coefficient $\beta = (\beta_{1i}, \beta_{2i}, \ldots\ldots, \beta_{pi})^T$, i = 1, 2, ..., d baseline hazard h_0 (t) is the hazard with the covariates equal to zero (x_{1i}, x_{2i},....., x_{pi} = 0). If we have two patients with the same score on all covariates except covariate m then

h_0 (t): denotes the baseline hazard which may vary over time.

x : : denotes the covariate.

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ refer to the covariate coefficients. (Khawar, 2019)

The fact that the baseline hazard is not stated, nevertheless it is possible to derive respectably accurate estimates of regression coefficients, hazard ratios of interests, and modified survival curves for a range of data scenarios, is a major factor contributing to the Cox model's popularity. (Kleinbum & Klein, 2012).

Therefore, the survival probability function for Cox ph model can be formulated as:

$$S(t|x) = S_{o(t)} \exp^{(\beta x)}. \quad \ldots 15$$

Or

$$S(t|x) = S_{o(t)} \exp\left(\sum_{i=1}^{p} B_i X_i\right) \quad \ldots 16$$

When;

$$s_0(t) = e^{-\int_0^t h_0(x)\, dx}. \quad \ldots 17$$

For any two sets of predictors, x and x^*, the hazard ratio (HR) is constant over time.

$$\frac{h(t|x^*)}{h(t|x)} = \frac{h_0(t)\, exp(\sum_{k=1}^{p} \beta_k x^*)}{h_0(t)\, exp(\sum_{k=1}^{p} \beta_k x)} = exp\left(\sum_{k=1}^{p} \beta_k (x^* - x)\right) \quad \ldots 18$$

2.3.1: The Assumption of Proportional Hazards

Here, some key assumptions can be made.

l. First of those assumptions is that the proportional hazard, needs to be fixed from one patient to another.

2. The second assumption is that there needs to be a linear relationship between the natural log of the hazard function and the explanatory variables.

3. The third assumption is that the explanatory variable, in any case, does not need to depend on time.

4. Another key assumption that can be made is that statistical distributions should not be distributed by any response variable involved in the study.

5. Finally, another assumption is that the rate of hazard needs to increase in a linear pattern with time. (Collet, 2003)

2.3.2: Estimating the Coefficients in the Cox PH Model

The standard likelihood function cannot be used as we do not have any knowledge about baseline hazard $h_0(t)$, it does not have any specific form(unspecified), also we do not model the censoring distribution and is therefore removed out of the formula by Cox. That is why Cox model likelihood function is called "partial likelihood Function". Regression parameter β for Cox model is obtained by maximizing the partial likelihood and first we find out the equation for partial likelihood. Assume that $t_j = t_1, t_2, \ldots, t_p$ be the true failure times with one failure at each time and $R(t_j)$ is the risk set consisting of the subjects under observation i.e have not been censored or have not failed by time t_j, j = 1,2,....,p. Then the full likelihood is:

Then the partial likelihood function for the Cox PH model is given by:

$$L_{(\beta)} = \prod_{i=1}^{k} L_i = \prod_{i=1}^{k} \frac{\exp(\beta' x_i t_j)}{\sum_{k \in R(t_{(j)})} \exp(\beta' x_k t_j)}$$

$$L_{(\beta)} = \prod_{i=1}^{k} \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{\sum_{i=1}^{k} e^{(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}}$$

$$l(\beta) = \log(L_{(B)}) = \log\left[\prod_{i=1}^{k} \frac{\exp(\beta' x_i t_j)}{\sum_{k \in R(t_{(j)})} \exp(\beta' x_k t_j)}\right]$$

$$l(\beta) = \sum_{i=1}^{k} [\beta' x_i - \log\{\sum_{k \in R} exp\ \beta' x_i\}]$$

$$U(\beta) = \frac{\partial}{\partial \beta}\, l(\beta) = x_i - \frac{\sum_{k \in R} x_i exp\, \beta' x_i}{\sum_{k \in R} exp\, \beta' x_i}$$

$$\beta' = x_i - \frac{\sum_{k \in R} x_i exp \beta' x_i}{\sum_{k \in R} exp \beta' x_i}$$

$$I(\beta) = -[\frac{\partial}{\partial \beta}\left(U(\beta)\right)] = -\frac{\partial}{\partial \beta}\left[x_i - \frac{\sum_{k \in R} x_i exp \beta' x_i}{\sum_{k \in R} exp \beta' x_i}\right]$$

$$I(\beta) = -\left[\frac{\sum_{k\in R} x_i x_i{}' exp\beta' x_i}{\sum_{k\in R} exp\beta' x_i} - \frac{[\sum_{k\in R} x_i exp\beta' x_i]\,[\sum_{k\in R} x_i{}' exp\beta' x_i]}{\left(\sum_{k\in R} exp\beta' x_i\right)^2}\right] \quad \ldots 20$$

Equation (20) also known as minus the Hessian Matrix is used to produce the standard errors for the regression coefficients. After we obtain maximum partial likelihood estimator. then asymptotically,

$\hat{\beta} \sim N\,(B_0,\ I^{-1}(\hat{B})\ )$

where $I^{-1}(\hat{\beta})$ is the inverse of information matrix at $\beta = \hat{\beta}$ and $\beta_0$is a true value. This approximate distribution is used to construct confidence interval and test the hypothesis H_0: $\beta = \beta_0$, (Cameron & Trivedi, 2012).

3. Results and Discussions:

This section includes an applied analysis of survival data for patients with brain cancer using (log rank test and Cox proportional hazard model) to identify the factors affecting on the patient using statistical programs for analyzing the data (SPSS V. 25. and STATA V. 16).

3.1 Data Collection

The data used in this research was obtained from the official database of the Nanakali Main Hospital for Cancer in the Kurdistan Region of Iraq - Erbil, where these data were collected by patients through direct contact between the specialist doctor and patients. In this study. Data were collected during (6) years. Starting from January 1, 2016 until December 31, 2021 for all brain cancer patients. During the study period (274) patients died and (53) survived under censored, with a follow-up period until July 31, 2022, the survival time was measured in months and the data contained (10) variables shown below:

Table 1 The Response Variables Measured for these patients at Diagnosis

Variable	Name	Categorization	No.
Gender	**Gender**	**Female = (1)** **Male = (2)**	**181** **146**
Morphology	**Morphology is the study of the size, shape, and structure of cancer.**	**Neoplasm = (1)** **Germinoma = (2)** **Germ cell tumor = (3)** **Hemangioblastoma = (4)** **Ewing sarcoma = (5)** **Glioma = (6)** **Mixed glioma = (7)** **Ependymoma = (8)** **Ependymoma, anaplastic = (9)** **Astrocytoma = (10)** **Astrocytoma, anaplastic = (11)** **Gametocytes astrocytoma = (12)** **Fibrillary astrocytoma = (13)** **Pilocytic astrocytoma = (14)** **Pleomorphic xanthoastrocytoma = (15)** **Glioblastoma = (16)** **Giant cell glioblastoma = (17)** **Gliosarcoma = (18)** **Oligodendroglioma = (19)** **Oligodendroglioma, anaplastic = (20)** **Oligodendroblastoma = (21)** **Medulloblastoma = (22)** **Desmoplastic nodular medulloblastoma = (23)** **Primitive neuroectodermal tumor = (24)** **Large cell medulloblastoma = (25)** **Cerebellar sarcoma = (26)** **Central neurocytoma = (27)** **Atypical teratoid/rhabdoid tumor = (28)** **Meningioma = (29)** **Meningothelial meningioma = (30)** **Transitional meningioma = (31)** **Neurinomatosis = (32)** **Malignant lymphoma = (33)** **Malignant lymphoma, non-Hodgkin = (34)** **Hodgkin lymphoma = (35)** **Malignant lymphoma, large B-cell, diffuse = (36)**	**2** **2** **1** **3** **1** **16** **2** **8** **5** **48** **11** **5** **1** **8** **5** **122** **3** **4** **13** **5** **2** **24** **2** **1** **3** **1** **1** **3** **4** **4** **1** **1** **1** **3** **9** **2**
Behavior	**Behavior of stomach cancer is the ability to grow, invade other areas.**	**Benign = (1)** **Uncertain = (2)** **Malignant = (3)**	**10** **10** **307**

Grade	**grade describes how normal or abnormal cancer cells look under a microscope**	**Grade I = (1)** **Grade II = (2)** **Grade III = (3)** **Grade IV = (4)** **B-Cell = (5)** **Un known = (6)**	**12** **66** **99** **185** **2** **29**
Extent	**extent Means tumor extention beyond limits of organ of origin.**	**Localized = (1)** **Regional by direct extension = (2)** **Regional lymph nodes = (3)** **Regional direct extension and lymph nodes = (4)** **Distant metastasis = (5)** **Not applicable = (6)** **Un known = (7)**	**52** **101** **46** **1** **2** **8** **117**
Surgery	**Surgical operations to remove cancerous tumors.**	**Made surgery = (1)** **Does not make surgery = (2)**	**304** **23**
Radiotherapy	**radiotherapy of patient.**	**Took Radiotherapy= (1)** **Does not take Radiotherapy = (2)**	**242** **85**
Chemotherapy	**Chemotherapy of patient.**	**Injected Chemotherapy = (1)** **Does not inject Chemotherapy = (2)**	**211** **116**
Age-groups	**Age of patient at time of diagnosis**	**1- 10 = 1** **11 – 20 = 2** **21 – 30 = 3** **31 – 40 = 4** **41 – 50 = 5** **51 – 60 = 6** **61 – 70 =7** **71 – 80 =8** **81 – 90 = 9** **91 – 100 = 10**	**35** **29** **39** **47** **68** **52** **39** **15** **2** **1**

3.2: Log Rank Test

The non-parametric log rank test compares two or more independent estimated time to event curves based on censored data. The log rank test is testing the null hypothesis that there is no difference in the overall survival distributions between the groups in the population.

Table 2 the Results of Log Rank Test for Brain Cancer

Overall Comparisons			
Log Rank (Mantel-Cox)	**Chi-Square**	**Df**	**P-value**
Gender	0.935	1	0.334
Morphology	142.431	35	0.000
Behavior	16.682	2	0.000
Grade	6.831	5	0.234
Extent	14.579	6	0.024
Surgery	0.121	1	0.728
Radiotherapy	4.557	1	0.033
Chemotherapy	5.412	1	0.020
Age-group	35.746	9	0.000
Test of equality of survival distributions for the different levels of (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radiotherapy, Chemotherapy, Age-group)			

In the table 3.3 the P-value of the log-rank tests of (gender, Grade, Surgery) greater than 0.05, this means that we accept the null hypothesis. It means that; statistically, the survival curves of the (gender, Grade, Surgery) do not differ.

Morphology: The P-value for the log-rank test is less than 0.05. We therefore reject the null hypothesis. That is there is a significant difference in survival probability between the type groups of morphology Brain cancer.

Behavior: The P-value for the log-rank test is equal to 0.000, or less than 0.05, therefore; we reject the null hypothesis and it means that there is difference in survival between t the three groups of Behavior in Brain cancer.

Extent: The results in table indicates that the P-value is less than 0.05, and this means that we reject the null hypothesis. That is there is a significant difference in survival probability between the seven groups of Extent Brain cancer.

Radiotherapy, Chemotherapy: The P-value for the log-rank test both of them less than 0.05, therefore; we reject the null hypothesis and it means that there is difference in survival between the patients who underwent (Radiotherapy or chemotherapy) and those without it, i.e. the patient who took (Radiotherapy or chemotherapy) has an increased chance of survival.

Age-group: The p-value is $0.000 \leq 0.05$ we reject the null hypothesis, which indicates that there is a significant difference between the classes of age.

3.3 Application of Cox-Proportional Hazard Model

The model building process of Cox-Proportional Hazard Model in this study occurs in nine variables (Gender, Morphology, Behavior, Grade, Extent, Surgery, Radio, Chemotherapy and Age group)

Table 3 Case Processing Summary in Cox-PH Available in Analysis

Case Processing Summary			
		N	Percent
Cases available in analysis	Event [a]	274	83.8%
	Censored	53	16.2%
	Total	327	100.0%
a. Dependent Variable: time			

Table 2 shows the case processing summary in Cox PH available in the analysis that determines whether the event occurred for a specific case or not, the number of cases available in the event analysis is 327 cases, the analysis shows that there are 274 deaths, 83.8% are event data and 53 cases, 16.2%, is the number of patients who are still alive under observation.

Omnibus tests are a type of statistical test for all variables, sometimes called the chi-square test. It is a statistical test carried out on a general hypothesis that tends to find general significance between the variance of parameters. The hypothesis is:

H_0: The model includes explanatory variables.
H_1: The model not includes explanatory variables.

Table 4 Omnibus Tests of Model Coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-Square	D.F	P-Value	Chi-Square	D.F	P-Value	Chi-Square	D.F	P-Value
2764.795	46.301	9	0.000	46.932	9	0.000	46.932	9	0.000

Table 3 shows that the value of chi-square = 46.301 at the degree of freedom of 9 and the P-Value 0.000, which means that the statistical model is statistically significant, which indicates that the variables in the model have importance and effect. Thus, we accept the null hypothesis, which states that the explanatory variables are included in the statistical model.

Table 5 Variables in the Equation for Cox Regression

Variables in the Equation								
X	B	S.E	Wald	Df	P-Value	Exp(B)	99.0% CI for Exp(B)	
							Lower	Upper
Gender	-0.086	0.125	0.472	1	0.492	0.918	0.665	1.267
Morphology	-0.001	0.000	6.380	1	0.009	0.999	0.998	1.000
Behavior	0.452	0.147	9.517	1	0.002	1.572	1.077	2.293
Grade	0.006	0.035	0.025	1	0.875	1.006	0.918	1.102
Extent	0.022	0.018	1.439	1	0.230	1.022	0.975	1.072
Surgery	-0.034	0.232	0.021	1	0.884	0.967	0.531	1.759
Radiotherapy	-0.155	0.189	0.671	1	0.413	0.857	0.526	1.394
Chemotherapy	-0.071	0.171	0.172	1	0.678	0.932	0.600	1.447
Age (Binned)	0.129	0.031	16.783	1	0.000	1.138	1.049	1.234

Table 4 shows estimates of the model's coefficients, standard error and degree of freedom, in addition to Wald's test. It also shows the covariates within the model that have no effect or effect by comparing the value of the covariate with the other categories for each of the variables using Exp (B) are called hazard ratios (HR), shows that the event hazard increases as the value of the ith covariate increases, and therefore the duration of survival decreases., if it is equal one, that is, there is no effect on the event,

in summarize:

- HR = 1: No effect
- HR < 1: decrease in the hazard
- HR > 1: Increase in Hazard

According to our data from the table 4 there is only three significant covariates in the model, we explain each of them and their effects on patients with stomach cancer as follows:

- Morphology is considered as one of the variables that have an impact on increasing the event risk of the patient's survival, a value of Exp(B) = 0.999, which is decrease in the risk of death for to the patient and p-value = 0.000 which indicates a statistically significant effect on the brain cancer patient.
- behavior variable is considered a factor in increasing the survival of the patient with brain cancer with a value of Exp(B) = 1.572. furthermore, the significant of P-value = 0.002 less than 0.01, this indicates that the variable is statistically significant.
- Another significant factor in the model is age. The estimated risk in the age group is Exp(B) = 1.138, which is an increase in the risk of death for patients. the p-value is 0.000 which is is statistically significant.

- Gender, Grade, Extent, Surgery, Radiotherapy and Chemotherapy are not significant factors because their p-value are greater than (0.01).

Then the Cox-PH model with significant factor as follows:

$$h_i(t) = h_0(t)\exp(-0.001\ Morphplogy + 0.452\ Behavior + 0.129\ age)$$

4. Conclusions

After studying the data on brain cancer in Erbil city and from the results in the
practical part, the following conclusions have been reached:

1. The result of the log rank test which is used for testing the null hypothesis that there is no difference in the overall survival probability between the groups showed that; for the gender variable there is no significant difference in survival probability between male and female patients. The log rank test gives statistically significant result for survival distributions of different levels of Morphology and behavior of brain cancer., while statistically, the survival curves of the surgery and Extent groups do not differ. And for different levels of Radiotherapy the results indicated that there is a significant difference in survival between them.
2. The Omnibus test of model effects for Cox models have demonstrated that the model fits the chosen variables however when their p- values are smaller than (0.01), which means that the model is statistically significant, which indicates that the variables in the model have importance and effect.
3. According to the results of the cox model, by the value of P-Value of the Wald Chi square test, the most significant variables that have an impact on brain cancer disease are (Morphology, Behavior and Age group).

– AMERI, S. (2015). Survival Analysis Approach For Early Prediction Of Student Dropout. Wayne State University. pp. 17-18.

– CAMERON, A. C. & TRIVEDI, P. K. (2012). Regression analysis of count data. In: Event history analysis with R. s.l.:Cambridge university press.Brostrom.

– COLLET, D. (2003). Modling Survival Data for Medical research. s.l.:s.n.

– DAKHIL, N. K., AL–MAYALI, Y. M., Y. M. & AL–A'BIDY, M. A. (2012). Analysis of Breast Cancer Data using Kaplan–Meier Survival Analysis. Journal of Kufa for Mathematics and Computer.

– DER, G., & EVERITT, B. (2007). Statistical Analyses Using SAS. Highly Influential Citations.

– FOX, J. (2002). Cox proportional-hazards regression for survival data.. In: An R and S-PLUS companion to applied regression. Gainesville: Sage Publications.
– HARRELL, F. E. (2017). multi-state survival models for interval censored data. U.S.: taylor & francis group.
– HOUT, A. V. D. (2017). Multi-State Survival Models for Interval-censored data. London: Taylor & Francis Group CRC Press.
– IHWAH, A. (2015). The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. Agriculture and Agricultural Science Procedia.
– INGER, P. (2002). Essays on the Assumption of Proportional Hazards in Cox Regression. Uppsala University. p. 11.
– KHAWAR , I., (2019). Overall and Relative Survival of Cancer Patients. p. Faculty of Science and Technology.
– KLEINBUM, D. G. & KLEIN, M. (2012). Survival analysis. In: A self-learning text (3rd ed.). New york: Springer.
– MAHDI, R. S. (2016). Using survival analysis to investigate breast cancer in the Kurdistan region of Iraq. City, University of London Institutional Repository. pp. 51-52.
– MAWLOOD, K. I. (2019). *Using Logistic Regression and Cox Regression Models to Studying the Most Prognostic Factors for Leukemia patients*. Qalaai Zanist Scientific Journal, 4(2), pp. 705-724.
– MAWLOOD, K. I. & OBED, S. A. (2019). *Study and Analysis of the Chest Cancer Data Using Survival Models*. Qalaai Zanist Scientific, 4(2), pp. 2518-6558.
– SAMARTZIS, L. (2006). Survival and censored data. first edition.
– SAWADOGO, S. (2018). Application of Cox Proportional Hazard Model in case of cirrhosis patients and Extension. Minnesota State University, Mankato. p. 11.
– SINGER, J. D. & WILLETT, J. B. (1991). Using Survival Analysis When Designing and Analyzing Longitudinal Studies of Duration and the timing of Events. s.l.:Psychological Bulletin.

Studying COVID19 Data in Erbil-Kurdistan/Iraq: Incidence, Survival, and Treatments in males and females

Prof. Dr. Kurdistan Ibrahim Mawlood
kurdistan.mawlood@su.edu.krd
Sarween Asaad Othman
Sarween.a.othman@su.edu.krd
Salahaddin University – Erbil College of Administration and Economics/ Statistics Department

The basic idea of this study focused on using of some advanced statistical methods for studding the most important factors affecting the Covid19 in Erbil city.
Utilizing Relative Risk Ratio (RRR) to compare the efficiency of the different affecting factor parameters in males and females. Using Kaplan Meier estimator to estimate the mean and median survival time data for all affecting factors to comparing the levels of treatment. Using Exponential parametric survival model for modeling and estimating affecting factor parameters of Covid19 patient's. The data set of this study was obtained from Arzheen private hospital in Erbil city. The results indicated that, all the death cases that have been recorded have had a high blood inflammation and a high D Dimer. Moreover, the results for our data indicated that the Chronic diseases are the most important factors affecting the Covid19 survival patients in the Exponential parametric survival model.
Key words: Survival Analysis, Relative Risk Ratio, Kaplan Meier, Covid19, parametric survival model.

دراسة بيانات كوفيد 19في أربيل – كوردستان / العراق: معدل الإصابة والبقاء والعلاج لدى الذكور والإناث

أ. د. كوردستان ابراهيم مولود
سروين أسعد عثمان
جامعة صلاح الدين –اربيل كلية الادارة و الاقتصاد/ قسم الاحصاء والمعلوماتية

ركزت الفكرة الأساسية لهذه الدراسة على استخدام بعض الأساليب الإحصائية المتقدمة لدراسة أهم العوامل التي تؤثر على كوفيد 19 في مدينة أربيل. استخدام نسبة المخاطر النسبية لمقارنة كفاءة معاملات العوامل المؤثرة المختلفة لدى الذكور والإناث. استخدام مقدر كابلان ماير لتقدير

بيانات متوسط ووقت البقاء على قيد الحياة لجميع العوامل المؤثرة لمقارنة مستويات العلاج. استخدام نموذج البقاء على قيد الحياة الأسي للنمذجة وتقدير العوامل المؤثرة في مريض كوفيد 19. تم الحصول على مجموعة بيانات هذه الدراسة من مستشفى أرزين الخاص في مدينة أربيل. أشارت النتائج إلى أن جميع حالات الوفاة التي تم تسجيلها كانت مصابة بارتفاع التهاب الدم وارتفاع D Dimer ، و أشارت نتائج بياناتنا إلى أن الأمراض المزمنة هي أهم العوامل التي تؤثر على مرضى البقاء على قيد الحياة لـ كوفيد 19 في نموذج البقاء على قيد الحياة الأسي.
الكلمات المفتاحية: تحليل البقاء ، نسبة المخاطر النسبية ، كابلان ماير ، كوفيد 19 ، نماذج البقاء على قيد الحياة.

1. Introduction

survival analysis is a statistical approach for data analysis where the outcome variable of interest is the time until the occurrence of an event, often referred to as a failure time or survival time. In a variety of disciplines, including engineering and medicine, survival analysis is applied. In pharmacological studies, the time to death is modeled, while in engineering, the time to mechanical system failure is studied.(EKMAN, 2017).

In clinical research, the survival time is employed. Depending on the sector of application, survival time may also be referred to as time to event, life time, duration time, or failure time. These techniques are widely used in a variety of fields, including public health, epidemiology, the social sciences, economics, and engineering, in addition to medical research. In terms of theory, methodology, and application, the survival data analysis has seen rapid developments. (LAWLESS, 2003).

2. Background Information

This section reviews the foundation of survival data analysis with an essential issue in health, which is covid-19 disease including the fundamental concepts and basic methods in modeling survival data, some of the key ideas and elements which form the foundation of this research; Kaplan-Meier, Relative Risk.

2.1: Covid-19 disease

Covid-19 is the disease caused by the emerging coronavirus called SARS-CoV-2. This novel virus was first detected by who on December 31, 2019, after a cluster of cases of viral pneumonia were reported in Wuhan, People's Republic of China.

Coronaviruses are a widespread family known to cause illnesses ranging from the common cold to more severe illnesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS).

The most common symptoms of covid-19 are: Fever, dry cough, stress. Other less common symptoms that may affect some patients include: loss of taste and smell, Nasal congestion, conjunctivitis, Sore throat, headache, muscle or joint pain, various rash patterns, nausea or vomiting, diarrhea, tremors or dizziness.
Other less common symptoms include: irritability, confusion, decreased level of consciousness (sometimes associated with seizures), anxiety, depression, sleep disorders, more severe, severe and rare neurological complications such as strokes, encephalitis, delirium and nerve damage. (WHO, 2019)

2.2: Survival analysis

In the past few decades, applications of the statistical techniques for survival data analysis have expanded beyond biomedical research to other fields like criminology, sociology, marketing, institutional research, and health insurance practice. Previously, survival analysis was only associated with the investigation of mortality rates. The first life table was created by John Graunt in 1662. Survival analyses have been used for data involving time until a given event, such as death, the development of an illness, or relapse of a condition.(CAMILLERI, 2019)
The duration of a subject's survival from one point to another is measured by their survival time. The concept of survival need not be taken literally. Here, survival indicates that a person is in a situation that corresponds to the default situation. The situation won't change until an interesting thing happens. Failure is the important occurrence that signifies the end of the time of survival. Failure usually involves dying or going through a bad experience. Failure, however, can sometimes have a beneficial outcome, such a disease remission. Failure may also be known as the event or, or death when death represents the failure. (PINTO, 2015)

2.3: Survival Function

Survival analysis deals with the implementation of certain statistical techniques to model and analyses survival time data. The probability density function (pdf) denoted by,$f(t)$ which can be written as: (LEE & WANG, 2003)

$$\boldsymbol{f(t) = \frac{dF(t)}{dt}} \qquad \ldots \mathbf{1}$$

and Cumulative Distribution Function (C.D.F.) denoted by $F(t)$ describes the probability the time to event (T) is smaller or equal compared to a fixed time (t) and is given as:

$$\boldsymbol{F\,(t) = p(T \leq t)} \qquad \ldots \mathbf{2}$$

From this, the survival function, the probability the time to event (T) is larger compared to a fixed time (t), can be derived as:

$$\mathbf{S\ (t) = Pr\ (T\ >\ } t) = \int_t^{\infty} \boldsymbol{f(u)du} = \mathbf{1 - F(t)}, \quad \mathbf{t \geq 0} \qquad \mathbf{\ldots 3}$$

Which means, the probability that an individual survives beyond time (t). Note that the survival function *S*(t) is a monotonic non-creasing continuous function with S(t) = 1 and $S(\infty) = \lim_{t\to\infty} S(t) = 0$. (DACWEN, 2002)

2.4: Hazard Function

That represents an individual the probability condition of death at time t after survival time, the hazard function that is denoted by $h(t)$, can be written as:

$$\boldsymbol{h(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T + \Delta t | T > t)}{\Delta t}} \qquad \mathbf{\ldots 4}$$

Representing the probability that an individual fails within a small interval $(t, t + \Delta t)$, given that the individual survived to the beginning of the interval.

That relationship between $s(t)$ and $h(t)$ is shown as:

$$\boldsymbol{h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{-d \log s(t)}{dt}} \qquad \mathbf{\ldots 5}$$

$f(t)$ is the density function which is the fraction of the original group for whom the event occurs during the time interval at t adjusted for the width of the time interval (LAWLESS , 2002)

2.5: Relative risk ratio (RRR)

Relative Risk (RR) is the most widely used measures of association in diseases. The direct computation of relative risks is feasible if meaningful prevalence or incidences are available, disease data may serve to calculate relative risks from prevalence (STARE & BOULCH, 2016)

Relative Risk is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group, relative risk measures the association between exposure and outcome along with the risk difference and odds ratio.

In the population under study, relative risk refers to a ratio between individuals of the population who express the trait of interest (e.g. disease), with attention for whether or not those members have previously been exposed to a relative risk (MORRIS, & GARDNER, 1988)

2.7: Kaplan-Meier estimator

Kaplan Meier is obtained out from name of two statisticians, Edward L. Kaplan and Paul Meier, who worked together and released a paper on how to deal with timings to event data in 1958. As a result, they developed the Kaplan-Meier estimator, a method for calculating the frequency or total number of people who survive medical treatment.

Additionally, estimates of survival data and Kaplan-Meier curves have improved data analysis in cohort studies. (ABUBAKAR & ALKASSIM, 2017).

Let $t_1 < t_2 < \cdots < t_k$ be the ordered observed survival times from a sample size n individual. And let r_j be the number of individuals who are at risk of failure time $t = t_j$, and d_j the number of individuals who experience the event at time $t = t_j$ and c_j the number of individuals with censoring times in (t_j, t_{j+1}), where $j = 0, 1, \ldots, k, t_0 = 0$ and $t_{k+1} = \infty$, it's clear in this setting that $r_j = d_j + c_j + d_{j+1} + c_{j+1} + \cdots + d_k + c_k$, the Kaplan-Meier estimator of the survival function is given by (LAWLESS, 2003).

$$\mathbf{KM}(\boldsymbol{t}) = \widehat{\boldsymbol{S}}(\boldsymbol{t_i}) = \prod_{\boldsymbol{tj<t}} \left(\mathbf{1} - \frac{\boldsymbol{d_j}}{\boldsymbol{r_j}}\right) \qquad \ldots \mathbf{6}$$

Where:

$\boldsymbol{r_j}$: The number of individuals alive at the start of the interval.

$\boldsymbol{d_j}$:The number of individuals who died. Is the following.

This mean that the conditional probability of the occurrence of an event at each observed time t_j (i.e.,d_j/r_j). Remember that if a censoring time and a lifetime are recorded as equal, the general convention is to consider the censoring time as being infinitesimally greater in the definition of $\widehat{\boldsymbol{S}}(\boldsymbol{t})$**.** (OBED & MAWLOOD, 2019)

2.8: Parametric distribution for time to event

Survival analysis deals with the analysis of times to events, or lifetimes. Parametric models are used to represent the distributions of lifetimes, and their relationship to explanatory variables, or covariates, parametric models such as the exponential distribution, Weibull distribution, and lognormal distribution, Log-logistic distribution and associated methods of inference. Parametric regression analysis using linear models for log lifetimes (Accelerated failure time models) is described and illustrated (LAWLESS, 2003).

The accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. (LEE & GO, 1997)

For $i = 1, \ldots, n$ let T_i be the failure time for the ith subject and let X_i be the associated p-vector of covariates. The accelerated failure time model specifies that

$$\text{Log } T_i = B_0 + B_1 X_1 + \cdots + B_P X_P + \varepsilon_i \qquad \ldots \mathbf{7}$$

where B_0 is a p-vector of unknown regression parameters and ε_i $(i = 1, \dots, n)$ are independent error terms with a common, but completely unspecified, distribution. (Qi, 2009). the exponential AFT model was class included in this research.

2.8.1: Exponential Distribution

The exponential distribution is the simplest model for lifetime data. It has only one parameter and therefore, not flexible enough to describe commonly encountered hazard shapes for time-to-event data. The exponential distribution has since continued to play a role in lifetime studies analogous to that of the normal distribution in other areas of statistics (LEE & WANG, 2003).

When the survival time T follows the exponential distribution with a parameter λ, the probability density function (pdf) denoted by $f(t)$, is defined as

$$\boldsymbol{f(t) = \lambda \exp(-\lambda t)} \qquad \dots \mathbf{8}$$

The cumulative distribution function is

$$\boldsymbol{F\ (t) = 1 - exp(-\lambda t)} \qquad \boldsymbol{t \geq o} \quad \dots \quad \mathbf{9}$$

and the survival function denoted by *S* (t) is then

$$\boldsymbol{S\ (t) = \exp(-\lambda t)} \qquad \boldsymbol{t \geq 0} \qquad \dots \mathbf{10}$$

So that, the hazard function denoted by $h(t)$ is defined as

$$\boldsymbol{h(t) = \lambda} \qquad \boldsymbol{t \geq 0} \qquad \dots \mathbf{11}$$

for t ≥ 0, where $\lambda > 0$ is the scale parameter (the rate parameter $\frac{1}{\lambda}$); a large value of λ indicates high risk and short survival, whereas a small value indicates low risk and long survival. The distribution with $\lambda = 1$ is called the standard exponential distribution. Since the hazard rate is constant, the exponential distribution frequently found to be inadequate to describe time-to-event data. This makes the applicability of this distribution fairly limited (LIU, 2012).

3. Results and Discussions:

Is in this section Relative Risk used to compare the efficiency of the different affecting factor parameters in males and females, the mean and median survival time estimated for all affecting factors to comparing the levels of treatment using Kaplan Meier estimator. Exponential parametric survival model used for modeling and estimating affecting factor parameters of Covid19 patient's. The following programs were used to analyze the data:

1.Mat-lab. 2. Stata. 3.STATGRAPHICS.

3.1 Data Collection

The data for this study of covid-19 have been collected from Arzheen private hospital in Erbil city. The data consisted of 350 cases for all patients with covid-19 who were registries and treated at Arzheen private hospital, corona department, during 1st September 2020 through 30th June 2021, of those patients 44 died during the study and 306 survivals alive. The survival time are measured in days from the first day that patient admitted to hospital to the date of death or the last visit to the hospital

The following covariate were included as prognostic factors in the study had been collected for all patient:

The patient related variables (Age, Gender, Smoker).

Clinical related variables (Peripheral oxygen saturation (SPO2), White blood cell (WBC), Lymphocyte, Monocyte, Hemoglobin (Hb), Red blood cell (RBC), Platelet (PLT), C reactive protein (CRP), Ferritin, Lactate dehydrogenase (LDH), Heart beat (HR), Blood Pressure, D Dimer).

Chronic diseases (Hypertension, Diabetes mellitus, Chronic lung disease, Cardiovascular disease).

Dependent variable, this is the outcome of treatment of a patient enrolled at a corona department in Arzheen hospital, these outcomes were either died or completed treatment and survival alive. Specific variables used and their categorization is shown in table 1 below.

Table 1 variable categorization

Variable names	Categorization	N	No. of Alive	No. of Death
Age grouped	**<=18** **19 - 39** **40 - 59** **60- 79** **80+**	**2 (0.6%)** **61 (17.4%)** **163 (46.6%)** **121 (34.6%)** **3 (0.9%)**	**2** **56** **138** **107** **3**	**0** **5** **25** **14** **0**
Gender	**1=male** **2=female**	**196 (56%)** **154 (44%)**	**174** **132**	**22** **22**
Smoker	**0=non-smoker** **1=smoker**	**261 (74.6%)** **89 (25.4%)**	**241** **65**	**20** **24**
SPO_2	**1=Low** **2=Normal** **3=High**	**232 (66.3%)** **118 (33.7%)** **0 (0%)**	**193** **113** **0**	**39** **5** **0**
WBC	**1=Low** **2=Normal** **3=High**	**9 (2.6%)** **192 (54.9%)** **149 (42.6%)**	**4** **183** **119**	**5** **9** **30**
Lymphocyte	**1=Low** **2=Normal** **3=High**	**137 (39.1%)** **213 (60.9%)** **0 (0%)**	**117** **189** **0**	**20** **24** **0**

Monocyte	**1=Low 2=Normal 3=High**	**13 (3.7%) 332 (94.9%) 5 (1.4%)**	**7 296 3**	**6 36 2**
Hb	**1=Low 2=Normal 3=High**	**341 (97.4%) 8 (2.3%) 1 (0.3%)**	**297 8 1**	**44 0 0**
RBC	**1=Low 2=Normal 3=High**	**103 (29.4%) 195 (55.7%) 52 (14.9%)**	**91 164 51**	**12 31 1**
PLT	**1=Low 2=Normal 3=High**	**15 (4.3%) 333 (95.1%) 2 (0.6%)**	**8 296 2**	**7 37 0**
CRP	**1=Normal 2=High**	**3 (0.9%) 347 (99.1%)**	**3 303**	**0 44**
Ferritin	**1=Low 2=Normal 3=High**	**12 (3.4%) 100 (28.6%) 238 (68%)**	**12 99 195**	**0 1 43**
LDH	**1=Low 2=Normal 3=High**	**27 (7.7%) 93 (26.6%) 230 (65.7%)**	**27 85 194**	**0 8 36**
HR	**1=Low 2=Normal 3=High**	**55 (15.7%) 73 (20.9%) 222 (63.4%)**	**50 67 189**	**5 6 33**
Blood Pressure	**1=Low 2=Normal 3=High**	**122 (34.9%) 213 (60.9%) 15 (4.3%)**	**109 188 9**	**13 25 6**
D Dimer	**1=Normal 2=High**	**116 (33.1%) 234 (66.9%)**	**116 190**	**0 44**
hypertension	**0=No 1=Yes**	**146 (41.7%) 204 (58.3%)**	**132 174**	**14 30**
diabetes mellitus	**0=No 1=Yes**	**214 (61.1%) 135 (38.9%)**	**195 110**	**19 25**
chronic lung disease	**0=No 1=Yes**	**206 (58.9%) 144 (41.1%)**	**192 114**	**14 30**
cardiovascular disease	**0=No 1=Yes**	**290 (82.9%) 60 (17.1%)**	**265 41**	**25 19**
Status	**0= Alive (censored) 1= Death**	**306 (87.4%) 44 (12.6%)**		
Treatment Duration	**The number of days of treatment duration**			

Table 1 shows the age of diagnosis ranged from 15 to 85years, most of the patients (46.6%) were at the age group of 40 to 59, out of a total of 163 cases including 25 patients died and 138 cases remained alive.

A total of 196 patients (56%) were male and 154 patients (44%) were female. Of the 350 patients with covid-19, elevated (SPO2 and Lymphocyte) not observed, most of the patients had low SPO2 (66.3%), and 39 patients with low SPO2 are dead out of 44 cases. Higher death rate observed in patients with high WBC (42.6%). Only 3 patients were recorded to have normal CRP and non-of them had died from the disease, all the death cases that have been recorded have had high blood inflammation, which means CRP elevated in patients with covid-19. In addition, our result showed that elevated Ferritin observed in 238 patients (68%), 230 (65%) had elevated LDH while 222 (63.4%) had elevated HR. In total of 44 death cases in the study 43 of them had elevated Ferritin, 36 had elevated LDH and 33 had abnormally high HR. Patients that had a normal blood pressure were 213 patients (60.9%) and 25 are died out of 44 cases.

Out of 350 patients 333 patients (95.1%) had a normal PLT and 37 patients of them are died. Regarding HR and D Dimer the results show all the dead cases were 44 patients all had low Hb and high D Dimer. The results show that all the dead cases that were 44 patients all had a high D Dimer in a total of 234 patients (66.9%). The result shows that 261 patients (74.6%) are non-smokers and 89 patients (25.4%) were smokers and 24 patients that died were smokers. the cases that had high hypertension were 204 patients (58.3%) and 30 of them are died. It seems like 25 patients that died had diabetes in a total of 135 patients (38.9%). The patients that had chronic lung diseases were 144 patients (41.1%) and 30 patients have died in total of 44 dead cases. The majority of the patients that had cardiovascular disease were 290 patients (82.94%) and 25 patients died out of them.

3.2 Application of Relative Risk Ratio

Relative risk ratio (calculated for male versus female (M / F)) shown in table (2, 3, 4) expresses which gender is at higher risk of having covid-19 for different levels of all covariates were included in the study.

Table 2 Relative Risk Ratio for (Age, Smoker)

Variable names	Categorization	Gender		Relative Risk (M/F)
		Male	Female	
Age grouped	<=18	1 (50%)	1 (50%)	0.786
	19 – 39	33 (54.1%)	28 (45.9%)	0.926
	40 – 59	89 (54.6%)	74 (45.4%)	0.945
	60- 79	71 (58.7%)	50 (41.3%)	1.116
	80+	2 (66.7%)	1 (33.3%)	1.571
Smoker	0=non-smoker	137 (52.5%)	124 (47.5%)	0.868
	1=smoker	59 (66.3%)	30 (33.7%)	1.545

Our result showed that incidence of covid-19 among female was higher than male up to age 60, and 60 or older males were at a higher risk to have covid-19. In smoker variable males had a higher risk than female by relative of (1.545) it means that smoker meals (1.5) times more likely to be diseased, and for the non-smoker patients female are at higher risk.

Table 3 Relative Risk Ratio of (M/F) for clinical related variables

Variable names	Categorization	Gender		Relative Risk (M/F)
		Male	Female	
SPO_2	1=Low	127 (54.7%)	105 (45.3%)	0.950
	2=Normal	69 (58.5%)	49 (41.5%)	1.106
	3=High	0 (0%)	0 (0%)	0
WBC	1=Low	8 (88.9%)	1 (11.1%)	6.286
	2=Normal	109 (56.8%)	83 (43.2%)	1.032
	3=High	79 (53%)	70 (47%)	0.887
Lymphocyte	1=Low	78 (56.9%)	59 (43.1%)	1.039
	2=Normal	118 (55.4%)	95 (44.6%)	0.976
	3=High	0 (0%)	0 (0%)	0
Monocyte	1=Low	11(84.6%)	2 (15.4%)	4.321
	2=Normal	184 (55.4%)	148 (44.6%)	0.977
	3=High	1 (20%)	4 (80%)	0.196
Hb	1=Low	193 (56.6%)	148 (43.4%)	1.025
	2=Normal	3 (37.5%)	5 (62.5%)	0.471
	3=High	0 (0%)	1 (100%)	0
RBC	1=Low	78 (75.7%)	25 (24.3%)	2.451
	2=Normal	96 (49.2%)	99 (50.8%)	0.762
	3=High	22 (42.3%)	30 (57.7%)	0.576

PLT	**1=Low**	**10 (66.7%)**	**5 (33.3%)**	**1.571**
	2=Normal	**184 (55.3%)**	**149 (44.7%)**	**0.97**
	3=High	**2 (100%)**	**0 (0%)**	**0**
CRP	**1=Normal**	**2 (66.7%)**	**1 (33.3%)**	**1.571**
	2=High	**194 (55.9%)**	**153 (44.1%)**	**0.996**
Ferritin	**1=Low**	**7 (58.3%)**	**5 (41.7%)**	**1.1**
	2=Normal	**72 (72%)**	**28 (28%)**	**2.02**
	3=High	**117 (49.2%)**	**121 (50.8%)**	**0.759**
LDH	**1=Low**	**17 (63%)**	**10 (37%)**	**1.336**
	2=Normal	**43 (46.2%)**	**50 (53.8%)**	**0.676**
	3=High	**136 (59.1%)**	**94 (40.9%)**	**1.137**
HR	**1=Low**	**28 (50.9%)**	**27 (49.1%)**	**0.815**
	2=Normal	**42 (57.5%)**	**31 (42.5%)**	**1.065**
	3=High	**126 (56.8%)**	**96 (43.2%)**	**1.031**
Blood Pressure	**1=Low**	**61 (50%)**	**61 (50%)**	**0.786**
	2=Normal	**126 (59.2%)**	**87 (40.8%)**	**1.138**
	3=High	**9 (60%)**	**6 (40%)**	**1.179**
D Dimer	**1=Normal**	**67 (57.8%)**	**49 (42.2%)**	**1.074**
	2=High	**129 (55.1%)**	**105 (44.9%)**	**0.965**

Our data showed that females had a higher risk compared to males when they had low SPO2, and patients who had a normal SPO2 males are slightly in a higher risk than females. The patients that had a normal or low WBC males are at a higher risk to have covid-19. However, the patients that had a low Monocyte male were more than 4 times at a higher risk but higher risk had observed in females for normal and high Monocyte. Higher risk rate was observed in males' patients with low (lymphocyte, Hb and PLT).

Females have a higher risk when they have a high or normal RBC but males are in a higher risk by 2.5 times than females when their RBC were low. It shows that the patients that had a normal (D Dimer and CRP) males are in more risk than females but when the (D Dimer and CRP) where high females were at more risk. The results show that females are in more risk by a relative of (0.759) compared to males when their ferritin was high. It is clear that the patients that have a low or high LDH females are at a lower risk to getting covid-19. In both (HR and blood pressure) variables males are at a higher risk when their HR and blood pressure is high or normal.

Table 4 Relative Risk Ratio of (M/F) for Chronic diseases

Variable names	Categorization	Gender		Relative Risk (M/F)
		Male	Female	
Hypertension	0=No	78 (53.4%)	68 (46.6%)	0.901
	1=Yes	118 (57.8%)	86 (42.2%)	1.078
Diabetes mellitus	0=No	117 (54.7%)	97 (45.3%)	0.948
	1=Yes	79 (58.1%)	57 (41.9%)	1.089
Chronic lung disease	0=No	123 (59.7%)	83 (40.3%)	1.164
	1=Yes	73 (50.7%)	71 (49.3%)	0.808
Cardiovascular disease	0=No	164 (56.6%)	126 (43.4%)	1.023
	1=Yes	32 (53.3%)	28 (46.7%)	0.898

It is clear that for patient (Hypertension and Diabetes) diseases that meals are at higher risk than females to have covid-19. Higher risk rates have been noted in females related to meals to have covid-19 for patients with Chronic lung disease and Cardiovascular disease.

3.4: Kaplan-Meier Test

The Kaplan-Meier process is a nonparametric technique for estimating one of the better criteria for estimating the number of patients that live for a particular period of time after treatments. The effect of an intervention is measured in clinical trials or community trials by estimating the number of individuals who survived or were saved after the intervention over a period of time, this can be measured for two independent groups, as well as the statistical difference in survival time between them. When comparing two different study populations, this can be used in this research. The result of KM test applied to data set 350 cases for the (Gender, Smoker, SPO_2, Blood Pressure, Hypertension, Diabetes mellitus, Chronic lung disease, Cardiovascular disease) variables show in tables (5, 6, 7, 8, 9, 10, 11, 12).

Table 5 the Means for Survival Time for (Gender) in each group

Means for Survival Time				
			95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound
M	39.482	7.093	25.579	53.384
FM	33.901	5.196	23.716	44.085
Overall	35.892	4.473	27.124	44.660

Table 5 gives the estimated mean time to death for male is greater than female which the estimated mean of the survival times to death for male is (39.482) and for female is (33.901) with the confidence interval

(25.579, 53.384) for male and (23.716, 44.085) for female under probability 95%.

Table 6 the Means for Survival Time for (smoker) in each group

Means for Survival Time				
			95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound
non-smoker	29.670	1.840	26.064	33.276
Smoker	34.174	5.743	22.918	45.431
Overall	35.892	4.473	27.124	44.660

Table 6 showed that the estimated mean time until death for non-smoker is less than smoker which the estimated mean time until death for smoker equal to (34.174) days with the confidence interval (22.918, 45.431) and the estimated mean time until death for non-smoker equal to (29.670) with the confidence interval (26.064, 33.276) under probability 95%.

Table 7 the Means for Survival Time for (SPO_2) in each group

Means for Survival Time				
			95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound
Low	34.008	4.652	24.889	43.126
Normal	32.523	1.657	29.276	35.770
Overall	35.892	4.473	27.124	44.660

Table 7 shows that the estimated mean time to death for low SPO_2 is (34.008) days while for normal SPO_2 is (32.523) days with confidence interval (24.889, 43.126) for low SPO_2, (29.276, 35.770) for normal SPO_2 under probability 95%.

Table 8 the Means for Survival Time for (Blood Pressure) in each group

Means for Survival Time				
			95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound
Low	27.233	1.789	23.726	30.740
Normal	41.021	7.301	26.712	55.330
High	35.416	10.973	13.908	56.923
Overall	35.892	4.473	27.124	44.660

Table 8 displays the estimated mean time until death for normal Blood Pressure is (41.021) days which is the largest, while for low Blood Pressure is (27.233) days and for high Blood Pressure is (35.416) days

with confidence interval (26.712, 55.330) for normal, (23.726, 30.740) for low and (13.908, 56.923) for high under probability 95%.

Table 9 the Means for Survival Time for (Hypertension) in each group

Means for Survival Time				
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
No	25.708	1.849	22.085	29.331
Yes	38.272	5.368	27.750	48.794
Overall	35.892	4.473	27.124	44.660

Table 9 explains the estimated mean time until death for patients who do not have Hypertension is less than who have Hypertension which the estimated mean time until death for have Hypertension equal to (38.272) days with the confidence interval (27.750, 48.794) while who do not have Hypertension equal to (25.708) days with the confidence interval (22.085, 29.331) under probability 95%.

Table 10 the Means for Survival Time for (Diabetes mellitus) in each group

Means for Survival Time				
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
No	28.662	1.856	25.025	32.299
Yes	36.479	5.778	25.155	47.804
Overall	35.892	4.473	27.124	44.660

From table 10 It is clear that the estimated mean time until death for patients who have Diabetes mellitus is greater than who do not have Diabetes mellitus which the estimated mean time until death for have Diabetes mellitus equal to (36.479) days while who do not have Diabetes mellitus is equal to (28.662) days with the confidence interval (25.155, 47.804) for have Diabetes mellitus and (25.025, 32.299) for do not have Diabetes mellitus under probability 95%.

Table 11 the Means for Survival Time for (chronic lung disease) in each group

Means for Survival Time				
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
No	30.069	2.244	25.670	34.467
Yes	35.475	5.017	25.643	45.308
Overall	35.892	4.473	27.124	44.660

Table 11 shows the estimated mean of the survival times to death for patients who have Chronic lung disease is greater than who do not have Chronic lung disease which the estimated mean time until death for have Chronic lung disease equal to (35.475) days and the estimated mean of the survival times to death for patients who do not have Chronic lung disease is (30.069) days with confidence interval (25.643, 45.308) for have Chronic lung disease and (27.124, 44.660) for do not have disease under probability 95%.

Table 12 the Means for Survival Time for (Cardiovascular disease) in each group

Means for Survival Time				
Cardiovascular disease	Mean			
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
No	43.776	6.809	30.431	57.121
Yes	23.190	1.911	19.444	26.936
Overall	35.892	4.473	27.124	44.660

Table 12 showed the estimated mean time until death for patients who have Cardiovascular disease is less than who do not have Cardiovascular disease which the estimated mean time until death for have Cardiovascular disease equal to (23.190) days with the confidence interval (19.444, 26.936) and the estimated mean time until death for do not have Cardiovascular disease equal to (43.776) with confidence interval (30.431, 57.121) under probability 95%.

3.5: Kaplan Meier Curve.

The survival plot of Kaplan Meier curve, used to compare different groups of subjects and analyze time to event data. The survival curve is used to determine the percentage (ratio) of patients who survive a specific occurrence, such as death over a period of time this can be computed for two groups of patients or subjects, and it can also be calculated for a three group of patients or subjects their survival rates differed statistically. Below Kaplan Meier survival curve for two factors (Gender and Smoker).

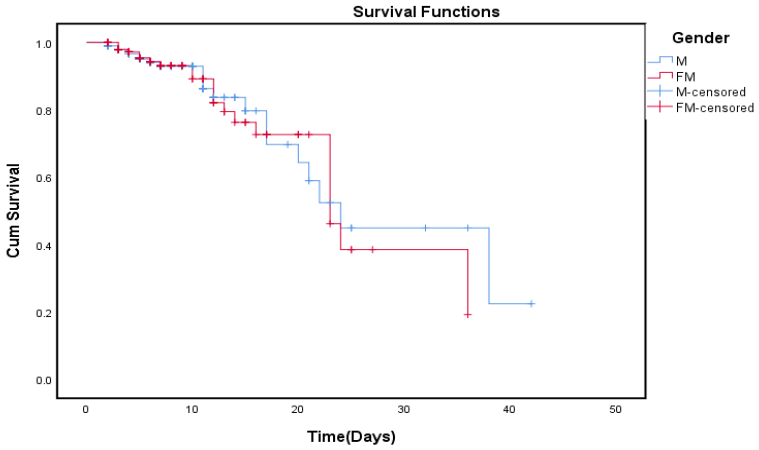


Figure 1 survival plot for Gender

In Figure 1 represent the survival curve for gender the vertical axis represents the cumulative of survival and the horizontal axis represent the time to event, the blue line is males and red line is females, both lines decreasing with time sometimes, the red line is little settled but blue line is constantly decreasing and then both lines little settled. We can see clearly from table 5 the estimated mean time for meals is (39.48) days which is more than females.

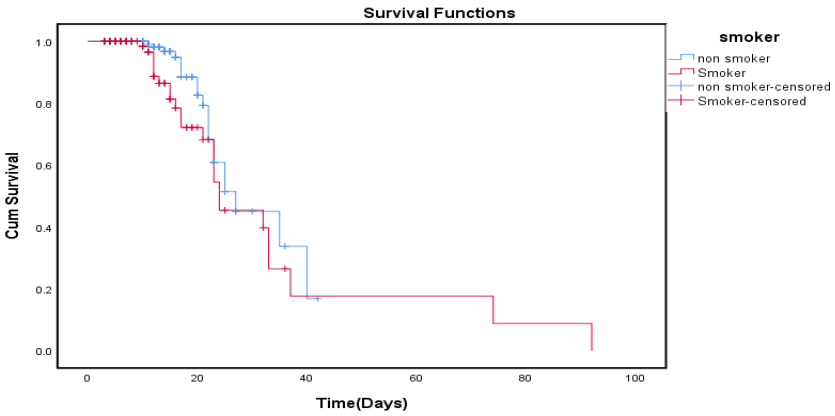


Figure 2 survival plot for Smoker

In Figure 2 represent the survival curve for smoker the vertical axis represents the cumulative of survival and the horizontal axis represent the time to event, the blue line is non-smoker and red line is smoker shows both lines decreasing but blue line is settled also the red line decreasing then little settled again decreasing. We can see clearly from table 6 the estimated mean time for non-smoker is (29.67) days which is less than smoker.

3.6: Fitting Model

The aim from using the Exponential Survival Model is to show the prognostic factor impact on survival in study. The survival function of the fitted rate data was used to validate the fit of the model.

Model fitting is a procedure a function that takes in a set of parameters and returns a predicted data set and 'error function' that provides a number representing the difference between data and the model's prediction for any given set of model parameters. We applied the model to our data using 20 treatments (patient related variables, clinical related variables and chronic diseases).

β: is a Coefficient regression describes the size and direction of the relationship between a predictor and the response variable, coefficients are the numbers by which the values of the term are multiplied in a regression equation. The sign of a regression coefficient tells us whether there is a positive or negative correlation between each independent variable and the dependent variable.

Table 13 Analysis of Fitting Exponential Model

Parameter Estimates in Exponential Model							
Parameter	**β**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Df**	**P-Value**
			Lower	**Upper**			
Constant	**2.642**	**1.616**	**-0.525**	**5.809**			
Age	**-0.228**	**0.089**	**-0.403**	**-0.053**	**6.579**	**1**	**0.010**
Gender	**0.154**	**0.129**	**-0.099**	**0.407**	**1.423**	**1**	**0.233**
Smoker	**0.304**	**0.153**	**0.004**	**0.605**	**4.094**	**1**	**0.043**
SPO2	**-0.026**	**0.125**	**-0.271**	**0.220**	**0.042**	**1**	**0.838**
WBC	**0.122**	**0.116**	**-0.105**	**0.349**	**1.107**	**1**	**0.293**
Lymphocyte	**-0.088**	**0.123**	**-0.329**	**0.152**	**0.520**	**1**	**0.471**
Monocyte	**-0.691**	**0.300**	**-1.280**	**-0.102**	**5.681**	**1**	**0.017**
Hb	**-0.182**	**0.299**	**-0.768**	**0.404**	**0.345**	**1**	**0.557**
RBC	**0.031**	**0.090**	**-0.146**	**0.207**	**0.116**	**1**	**0.733**
PLT	**-0.467**	**0.304**	**-1.063**	**0.128**	**2.469**	**1**	**0.116**
CRP	**0.363**	**0.593**	**-0.799**	**1.526**	**0.337**	**1**	**0.561**
Ferritin	**-0.046**	**0.115**	**-0.271**	**0.178**	**0.163**	**1**	**0.686**
LDH	**0.128**	**0.091**	**-0.0492**	**0.306**	**1.955**	**1**	**0.162**
HR	**0.061**	**0.079**	**-0.093**	**0.215**	**0.596**	**1**	**0.440**
Blood Pressure	**0.263**	**0.117**	**0.033**	**0.492**	**5.022**	**1**	**0.025**
D Dimer	**0.348**	**0.135**	**0.084**	**0.612**	**6.534**	**1**	**0.011**
Hypertension	**0.271**	**0.130**	**0.016**	**0.526**	**4.305**	**1**	**0.038**
diabetes mellitus	**0.398**	**0.126**	**0.151**	**0.645**	**10.226**	**1**	**0.001**
chronic lung disease	**0.401**	**0.130**	**0.147**	**0.656**	**9.749**	**1**	**0.002**
cardiovascular disease	**0.403**	**0.178**	**0.054**	**0.751**	**5.542**	**1**	**0.019**

The survival function for exponential model is:

$S\ (t;\ X) = \exp\ (-t\ [\exp\ (-b_0 - b_1 x_1 - b_2 x_2\ \ldots\ - b_n\ x_n)])$

However, we can write the Exponential Distribution equation with only significant variables as follows:

S (*t;* X) = exp (-t [exp (- 2.642 + 0.228 Age – 0.263 Blood Pressure – 0.304 Smoker + 0.691Monocyte - 0.348 D Dimer – 0.271 Hypertension – 0.398 diabetes mellitus

– 0.401 chronic lung disease– 0.403 cardiovascular disease)]).

The survival model is as follows:

$$\text{Log } T_i = B_0 + B_1 X_1 + \cdots + B_P X_P + \varepsilon_i$$

We fit the survival model above to the data in covid-19 disease

Log T_i = 2.642 – 0.228 Age + 0.154 Gender + 0.304 Smoker– 0.026 SPO_2 + 0.122 WBC

– 0.088 Lymphocyte – 0.691 Monocyte– 0.182 Hb + 0.031 RBC – 0.467 PLT

+ 0.363 CRP – 0.046 Ferritin + 0.128 LDH + 0.061 HR + 0.263 Blood Pressure

+ 0.348 D Dimer+ 0.271 Hypertension + 0.398 diabetes mellitus + 0.401 chronic lung

disease + 0.403 cardiovascular disease + ε_i

- The two patient related variables (Age and Smoker).

Age is one of the variables affecting to the risk in covid-19 diseases decrease by coefficient (β = - 0.228), which is decrease in the risk of the death for patient. The significant value is (0.010< = 0.05), so there is significant effect on covid-19 with chi-Square test value equal (6.579).

While, Smoker one of the significant variables because their p-value is significant (0.043 < = 0.05), the variable affecting in disease by coefficient (β = 0.304), which is increase in the risk of the death for patient.

- The significant variables in clinical related variables (Blood Pressure, Monocyte and D Dimer)

Our results showed that (Blood Pressure, D Dimer and Monocyte) clinical related variables are significant because their p-values are (0.025, 0.011 and 0.017 < = α = 0.05).

Blood Pressure and D Dimer are two factors affecting in covid-19 disease increase by coefficient (β =0.263 and 0.348) respectively, which is increase in the risk of the death for patient, with chi-Square test values equal (5.022 and 6.54)

But, Monocyte factor affect in the disease decrease by coefficient (β = - 0.691), which is decrease in the risk of the death for patient, the chi-Square test value is equal to (5.681) for Monocyte factor.

- The result show in chronic diseases (Hypertension, diabetes mellitus, chronic lung disease and cardiovascular disease).

All factors of chronic disease are significant because their p-value are less than (0.05), with coefficients equal (β = 0.271, 0.398, 0.401 and 0.403) and chi-square test values equal to (4.305, 10.226, 9.749 and 5.542) respectively, so all factors which are increase in the risk of the death for patient.

In addition, diabetes mellitus will be one of the significant factors in our study; because it has a greater value in chi-square test column (10.226) with significant value of (0.001 <= 0.05).

- The variables that do not affect to the risk of the death for patient are (Gender, SPO_2, WBC, Lymphocyte, Hb, RBC, PLT, CRP, Ferritin, LDH, HR) they are not significant factors because their p-value are greater than (0.05).

Conclusions

The following conclusions have been reached after studying the data on covid-19 in Erbil city:

1. Only three patients were found to have normal CRP levels, and none of them had died from the disease. All other death cases were discovered to have high blood inflammation, which indicates that patients with COVID-19 had raised CRP levels.
2. In the study's 44 death cases overall, 33 had abnormally high HR, 36 had raised LDH, and 43 had elevated Ferritin. 213 patients (60.9%) had normal blood pressure, while 25 of 44 cases resulted in deaths. The results for HR and D Dimer indicate that 44 patients who died all had low hemoglobin and high D Dimer. The results indicate that, out of all 44 deceased cases had high D Dimers.
3. Our results indicate that, up to the age of 60, females had a higher incidence of covid-19 than males, and that males 60 years of age or older had a greater risk of getting covid-19. In the smoker variable, men were at a higher risk than women by a ratio of (1.545), which indicates that smoker meals are (1.5) times more likely to covid-19, and for the non-smoker patient's female are at higher risk. According to our results, female patients with low SPO2 had a higher risk than males, and those with normal SPO2 males had slightly higher risk than females.
4. Male patients with low monocyte levels were more than 4 times more at risk, although larger risks for normal and high monocyte levels were seen in female patients.
5. Females are more at risk than males when their RBC is high or normal, but when their RBC is low, men are 2.5 times more at risk than women. It illustrates that in patients with normal (D Dimer and CRP),

males are at higher risk than females, whereas females are at higher risk when the (D Dimer and CRP) are high. When their ferritin levels were high, females were at an increased relative risk of (0.759) compared to males.

6. For Chronic diseases; patients with (Hypertension and Diabetes) diseases meals are at higher risk than females to have covid-19. Higher risk rates have been noted in females related to meals to have covid-19 for patients with Chronic lung disease and Cardiovascular disease.

7. The result of KM test shows the estimated mean time to death for male is greater than female. the estimated mean time to death for low SPO_2 is greater than normal SPO_2. The estimated mean time until death for normal Blood Pressure is greater than other groups (Low and High). For the patients who have Chronic diseases (Hypertension, Diabetes mellitus, Chronic lung disease), the estimated mean of the survival times to death are greater than who do not have (Hypertension, Diabetes mellitus, Chronic lung disease). While, the estimated mean time until death for patients who have Cardiovascular disease is less than who do not have Cardiovascular disease.

8. According to the Exponential model, identified that the most prognostic factors that influenced in covid-19 patient's survival are (Age, Blood Pressure, Smoker, Monocyte, D Dimer, Hypertension, diabetes mellitus, chronic luge disease, cardiovascular disease).

References

1. CAMILLERL, L. (2019). *History of survival analysis. the sunday times of malta*, p. 53.

2. EKMAN, A. (2017). Variable selection for the Cox proportional hazard model. Umea, John Wiley & Sons, Inc.

3. LAWLESS, J. F. (2002). Statistical Models and Methods for Lifetime Data. 2nd ed. Canada: A John Wiley & Sons, Inc.

4. LEE, E. T. & WANG, J. W. (2003). Statistical Methods for Survival Data Analysis. 3rd ed. Canada: John Wiley & Sons, Inc.

5. MORRIS, J. A. & Gardner, M. J., (1988). *Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates*. British Medical journal (Clinical research ed.), 296(6632), p.1313.

6. ABUBAKAR, S. & ALKASSIM, R. (2017). *The Kaplan Meier Estimate in Survival Analysis*. Biom Biostatistics Int J, 5(2), p.128.

7. LAWLESS, J. F. (2003). Statistical Models and Methods for Lifetime Data. 2nd ed. Canada: A John Wiley & Sons, Inc.

8. LIU, X. (2012). Survival Analysis Models and Applications. 1st ed. United Kingdom: A John Wiley & Sons, Ltd.

9. PINTO, J. D. (2015). *Outlier Detection in Survival Analysis based on the Concordance C-index*. In Bioinformatics, pp. 75 -82.
10. QI, J. (2009). Comparison of Proportional Hazards and Accelerated Failure Time Models. (Doctoral dissertation).
11. STARE, J. & BOULCH, D. M. (2016). *Odds Ratio, Hazard Ratio and Relative Risk*. Advances in Methodology and Statistics, 13(1), pp. 56-67.
12. WHO, (2019) World Health Organazation . [Online] Available at: https://www.who.int/ar/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19/ [Accessed 24 May 2022].
13. OBED, S. A. & MAWLOOD, K. I., (2019). Study and Analysis of the Chest Cancer Data Using Survival Models. Qalaai Zanist Scientific, 4(2), pp. 2518-6558.
14. LEE, E. T. & GO, O. T., (1997). Survival Analysis in public Health Research. Annu. Rev. Public Health, Volume 18, p. 105–34.
15. Daowen, Z. (2002). Modeling Survival Data with Parametric Regression Model. ST 745, 100-119.

Assessing Logistic and Poisson Regression Model for Analyzing Data Count of Patients with Tuberculosis Disease in Erbil, Iraqi Kurdistan Region

Asst. Prof. Dr. Paree khan Abdulla Omer
Pareekhan.omer@su.edu.krd
07504702219
College of Administration and Economics
Department of Statistics & informatics
Salaheddin University-Erbil

Abstract

Many studies choose to analyze and classify count variables as binary. The main purpose of this research is to determine how several predictor variables and a response variable relationship to each other's. To achieve this, logistic and Poisson regression models were being used. The review of patient data count for TB disease in Erbil, Iraqi Kurdistan Region, is the main focus of this research. where Gender (male, female) is a categorical response variable and there are multiple predictor variables. The data collected on this disease indicates that it might be a big problem in our society, as it affects a wide range of people for a number of reasons. The total number of cases during that time was (1346). (2012 to 2018). To assess each model's goodness of fit, two criterions were used. The results showed that the Logistic regression model is the best fit in modeling binary response variable in the form of a count data based on the two evaluation criteria used [Akaike Information Criterions (AIC) and Bayesian Information Criterions (BIC)]. Microsoft Excel, Stata 16, and SPSS 25 are the statistical application software used.

Keywords: logistic regression, Poisson regression model, model selection criteria.

ا.م. د. برى خان عبدالله عمر
Pareekhan.omer@su.edu.krd
07504702219
كلية الادارة والاقتصاد – قسم الاحصاء والمعلوماتية
جامعة صلاح الدين -اربيل

الملخص

تختار العديد من الدراسات تحليل وتصنيف متغيرات العد على أنها ثنائية .الغرض الرئيسي من هذا البحث هو تحديد عدد متغيرات التوقعة وعلاقته بمتغير الاستجابة مع بعضها

البعض .لتحقيق ذلك ، تم استخدام نماذج الانحدار اللوجستي وبواسون .تعد مراجعة بيانات المرضى المصابين بمرض السل في أربيل، إقليم كردستان العراق المحورالرئيسي لهذا البحث. حيث يعتبر الجنس (ذكر، أنثى) متغير استجابة فئوية وهناك متغيرات توقع متعددة .تشير البيانات التي تم جمعها حول هذا المرض إلى أنه قد يكون مشكلة كبيرة في مجتمعنا ، حيث أنه يؤثر على مجموعة واسعة من الناس لعدد من الأسباب. وبلغ العدد الإجمالي للمشاهدات خلال تلك الفترة (1346) حالة من (2012-2018) .لتقييم ملاءمة كل نموذج ، تم استخدام معيارين .أظهرت النتائج أن نموذج الانحدار اللوجستي هو الأنسب لنمذجة متغير الاستجابة الثنائية في شكل بيانات تعداد بناءً على معياري التقييم المستخدمين معايير معلومات (AIC) ومعايير المعلومات (BIC). Microsoft Excelو Stata 16 و SPSS 25 هي برامج التطبيقات الإحصائية المستخدمة.

1. Introduction

Regression analysis as a statistical methodology utilizes the relation between two or more quantitative variables. that is, a response variable can be predicted from the other(s). This methodology is widely used in commercial, social, behavioral, and biological sciences among other disciplines. Regression may be of two types: linear and nonlinear. Simple and multiple linear regression are the different types of linear regression (Nduka, 1999), while log-linear, quadratic, cubic, exponential, Poisson, logistic, and power regression are nonlinear regressions. Notably, Poisson and Logistic regression are of interest for us in this research.

Statistical methods with the many variables are commonly used in general health science literature. In the literature, the terms "multivariate analysis" and "multivariable analysis" are often used interchangeably. The relationship between two or more predictor (independent) variables and one outcome (dependent) variable is explored using multivariable methods. The outcome variable's predicted value is represented by the relationship model as a sum of products, each of which is generated by multiplying the independent variable's value by its coefficient. (Park, (2013)).

In many cases research focuses on models with a categorical response variable. Considering that a number of the conditions for this technique's assumptions will not be met, we could not perform a multiple linear regression in this situation. We would instead perform a logistic regression analysis. Logistic regression may thus be seen as a method similar to multiple linear regression that also takes into consideration the categorical nature of the response variable. The logistic regression framework may be used to examine a response structure that has an outcome variable and a set of explanatory variables (one or more). There are so many ways in which proportions and probabilities differ from continuous variables. They have a range of possible values between (0

and 1), whereas continuous variables can potentially take any value between plus or minus infinity. As a result, we cannot assume that a proportion seems to have a normal distribution and we must recognize that proportions have a binomial distribution. The mean and variance of the binomial distribution are not independent, in opposed to the normal distribution. (Park, (2013)).

When the outcome is a count, Poisson regression is useful. Similarly, to how logistic regression is used to estimate odds ratios in comparing various exposure groups, it is used to estimate rates or counts in comparing various exposure groups. Further, logistic and Poisson regression are used to determine the most important variables and the direction of each variable's effect. These models help researchers to account for such data contained in a series of observations between the dependent and independent variables (Armstrong, 2012), (Ijomah, et al., 2018).

Logistic regression and multiple regression models are similar to the general Poisson regression model. The Poisson distribution has many keeping count applications, include: 1) Telecommunication, 2) Biology, 3) Radioactivity, etc. Similar to the previous example, logistic regression is an essential model to take account when a response variable has two possible outcomes, such as the financial performance of the company (profit or loss), blood pressure, etc. For the analysis of data from either observational or experimental investigations, both models are suitable (Michael et al, 2005). Given that the response results are discrete (or binary response variables), Poisson and Logistic Regression models were taken into consideration in this article. (Ijomah et al., 2018).

2. Logistic Regression models with Binary Response Variable

Logistic regression is foremost used to model a binary variable based on one or more other variables, called predictors. The binary variable being modeled is generally referred to as the response variable, or the dependent variable. used the term “response” for the variable being modeled since it has now become the preferred way of designating it. For a model to ft the data well, it is assumed that (Hilbe, 2015):

The predictors are uncorrelated with one another.

they are significantly related to the response.

the observations or data elements of a model are also uncorrelated.

I have emphasized that binary response logistic regression is based on the Bernoulli probability distribution, which consists of a distribution of (1 and 0). The probability function can be expressed for a random sample as:

$$f(y;p)\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \qquad (1)$$

where the joint PDF is the product, Π, of each observation in the data being modeled, symbolized by the subscript (i). then characterize the Bernoulli distribution for a single observation as:

$$f(y_i; p_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \qquad (2)$$

Logistic regression models a relationship between predictor variables and a categorical response variable, used when the response has two possible outcomes. It is sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical (Park, 2013). Examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high or low blood pressure. Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , Y_i = o, 1 \qquad (3)$$

where the outcome Y_i is binary, taking on the value of either 0 or 1. The expected response $E(Y_i)$ has a special meaning in this case. Since $E(\varepsilon_i)$ we have:

$$E(Y_i) = \beta_0 + \beta_i X_i \qquad (4)$$

Consider (Y_i) to be a discrete random variable for which we can state the probability distribution as follows:

Y_i	Probability
1	$P(Y_i = 1) = p_i$
0	$P(Y_i = 0) = 1 - p_i$

Thus, p_i is the probability when $Y_i = 1$ and $1 - p_i$ is the probability that $Y_i = 0$. With the logistic model, estimates of π_i from equations like the one above will always be between 0 and 1. By definition of expected value of a random variable in equation (4), we obtain

$$E(Y_i) = 1(p_i) + 0(1 - p_i) = p_i = P(Y_i = 1) \qquad (5)$$

Equaling equation (4) and (5). we thus have

$$E(Y_i) = \beta_0 + \beta_i X_i = p_i \qquad (6)$$

Then, the logistic mean response function is:

$$E(Y_i) = p_i = \frac{\exp(\beta_0 + \beta_i X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \qquad (7)$$

3. Maximum Likelihood Estimation (logistic regression estimators)

Let the discrete random variable Y_i be Bernoulli random variable and each Y_i observation is an ordinary Bernoulli random variable where:

$$\left.\begin{matrix} P(Y_i = 1) = p_i \\ P(Y_i = 0) = 1 - p_i \end{matrix}\right\} \qquad (8)$$

Then, its probability distribution is representing as follows:

$f_i(Y_i) = p_i^{Y_i}(1-p_i)^{1-Y_i} \quad , \quad Y_i = 0,1 \quad , \; i = 1,2,3,\dots,n$ (9)

Note that $f_i(1) = p_i$ and $f_i(0) = 1 - p_i$

hence, $f_i(Y_i)$ simply represents the probability that, since the Y_i observation are independent. Their joint probability function is:

$$L(Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} \qquad (10)$$

Taking logarithm of equation (10), then the joint probability function:

$$log_e L(Y_1, Y_2, \dots, Y_n) = log_e \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

$$= \sum_{i=1}^{n} \left[Y_i log_e \left(\frac{p_i}{1-p_i} \right) \right] + \sum_{i=1}^{n} log_e(1-p_i) \qquad (11)$$

Since $E(Y_i) = p_i$ for a binary variable, it follows from equation (7) that:

$1 - p_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1}$ (12)

From equation (7), we obtain

$$log_e \left(\frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 X_i \qquad (13)$$

And equation (11) can be expressed as follows:

$$log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i[\beta_0 + \beta_1 X_i] - \sum_{i=1}^{n} log_e\left(1 + (\beta_0 + \beta_1 X_i)\right) \qquad (14)$$

Where $L(\beta_0, \beta_1)$ replaces $L(Y_1, Y_2, \dots, Y_n)$, to show explicitly that this function is now viewed as the likelihood function of the parameter to be estimated, given the sample observation (Ijomah et al. ,2018).

The maximum likelihood estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function in Equation (15). Computer intensive numerical search procedures are therefore required to find the maximum likelihood

estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$. Once the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found, we substitute these values into the response function in Equation (7) to obtain the fitted response function. We shall use $\hat{p}_\mathrm{i}$ to denote the fitted value for the i^{th} case

$$\hat{p}_\mathrm{i} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{X_i})}{1 + \exp(\mathrm{b_0} + \mathrm{b_1 X_i})} \qquad (15)$$

The fitted logistic response function is as follow:

$$\hat{p}_\mathrm{i} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{X_i})}{1 + \exp(\mathrm{b_0} + \mathrm{b_1 X_i})} \qquad (16)$$

In fact, from Multiple Logistic Regression Model there are several predictor variables usually required to obtain adequate description and useful predictions. In extending the simple logistic regression model, we simply replace $\beta_0 + \beta_1 X_i$ in Equation by $\beta_0 + \beta_1 X_i + \beta_2 X_2 + \cdots + \beta_k X_k$. To simplify the formula, we use matrix notation and the following three vectors:

$X'\beta = \beta_0 + \beta_1 X_i + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (17)$

From equation (17) the simple logistic function (7) extends to the multiple logistic response function as follows:

$$E(Y) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)} \qquad (18)$$

4. Poisson Regression model with Binary Response Variable

Poisson regression is frequently applied to count data. Count data is defined as "the number of occurrences of a behavior in a specific period of time". Integers must only be non-negative in count data (Karazsia et al, 2008). Hence A generalized linear regression model with a logarithmic link function is called Poisson regression (Durrant, 2016). The Poisson distribution from the discrete distribution family can be expressed to represent variables with such asymmetric right-slope distributions (Moksony and Hegedus, 2014). Let x_i and y_i be observations from a data set. Here, the numbers xi and yi are respectively a vector of independent and dependent variables. Poisson regression analysis assumes that the y_i shows the Poisson distribution. The probability density function for the Poisson distribution with the parameter λ_i is given in the following formula;

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \quad , \quad y_i = 0, 1, 2, \ldots \qquad \ldots (19)$$

The number of events occurring is denoted by the symbol y_i and the ratio of events occurring per unit of time is denoted by the symbol λ_i. In other words, λ_i provide the distribution's average. The probability here changes

as a function of λ_i. The Poisson probability distribution has a right-angled skew. However, when λ_i increase, the distribution gets closer to the normal distribution. The Poisson regression model's equal mean and variance is its most important feature. Because distortions are apparent in the assumption that the conditional expected value is equal to the variance and the assumption is not met, over- or under-dispersed data sets cannot be described by the Poisson distribution. In this case, updating the data set or starting the analysis with different methods may be a solution. The expected value and variance of y_i are given:

$$\lambda_i = E(y_i|x_i) = Var(y_i|x_i) \qquad \dots(20)$$

The link function illustrating the relationship between the expected value and the independent variables must have the form specified in Equation (21) in order to ensure that the expected value of y_i does not take negative values (Cameron and Trivedi, 1998).

$$\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots., \beta_k x_k \qquad \dots(21)$$

In this equation, λ_i is an exponential function of the arguments. λ_i is the same as given in Equation below:

$$\lambda_i = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots., \beta_k x_k) = e^{x_i'\beta} \qquad \dots(22)$$

Where β_0, β_1,..,β_k represent the unknown parameters.

5. Maximum Likelihood Estimation (Poisson Regression estimators)

There are many methods to calculating β estimators in the Poisson regression analysis based on the distribution of the dependent variable y_i. Maximum likelihood (MLE) method, artificial maximum likelihood (PMLE) method, and generalized linear models (GLM) are the most commonly applied and well-known of these techniques. The most used method for regression models is (Newton Raphson iteration) approach is typically employed in the likelihood method (MLE). The Poisson regression model's log likelihood function is as follows given an observation set:

$$L(\beta|(y,x)) = \sum_{i=1}^{n} P(y_i|\lambda_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \qquad \dots.(23)$$

When the logarithm of this function is taken, Equation below is obtained.

$$lnL(\beta) = \sum_{i=1}^{n} (y_i \ln(\lambda_i) - \lambda_i - lny_i!) \qquad \dots(24)$$

Accordingly, the Poisson MLE of (β) value is calculated from the expression in Equation (25), (Durmuş and Güneri, 2020).

$$\sum_{i=1}^{n}(y_i - \lambda_i)x_i = 0 \qquad \ldots(25)$$

6. Wald statistic

The Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The Wald statistic tests the hypothesis that the respective parameter β_j is equal to zero:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

The Wald statistic is asymptotically distributed as a Chi-square distribution:

$$W_j = \frac{\beta_j^2}{SE_{\beta j}^2} \qquad (26)$$

Each Wald statistic is compared with a Chi-square with 1 degree of freedom. Wald statistics are easy to calculate but their reliability is questionable. (Bewick et al., 2005)

7. Model Selection

The Model selection criteria considered in this research are Akaike Information Criterions (AIC) and Bayesian Information Criterions (BIC), the small values of BIC and AIC model will be chosen as the best model for the selected models.

7.1 Akaike Information Criterion (AIC)

When comparing statistical models fitted by maximum likelihood (ML) to the same data, the Akaike Information Criterion (AIC) is used to measure the relative superiority of each model for the given data set. The statistic penalizes for the number of predictors employed in the model and takes into account model parsimony, as a result:

$$AIC = -2\ln(L) + 2P \qquad \ldots(27)$$

Where $\ln(L)$: In is the natural logarithm (Likelihood) is the value of the likelihood

P: is the number of parameters in the model.

AIC can be calculated using residual sum of squares from regression (Henry, 2010):

$$AIC = n\ln(RSS/n) + 2P \qquad \ldots(28)$$

Where:

n: is the number of data points (observations). RSS: is the residual sum of squares, the AIC values for given data are meaningless, but when they are

compared to the AIC values of competing models, they take on meaning. The model with the smallest value of AIC among competing models is the best (ideal) model for the given data set, the AIC is used to choose a model that fits the data well but has a small number of parameters; as a result, the AIC penalizes the addition of parameters (Adeti, 2016).

7.2 Bayesian Information Criterion (BIC)

Another estimator evaluating model fit for a given data among different types of non-nested model is the Bayesian information criterion (BIC), and its formula is as follows:

$$BIC = -2logL + klog\, n \qquad \dots(29)$$

Where:

L: The model's maximum likelihood function.

k: Number of model parameters.

n: Number of observations (sample size).

The best model to fit the data is the one with the minimum value of BIC (Cameron and Trivedi, 2013).

8. Data collection

The data were collected on the Tuberculosis disease from (Chest and Health Center / Health Ministry - Erbil). This disease might be a big problem in our society where caused many people and there are many reasons affected on the human. Tuberculosis (TB) is caused by germs that are spread from person to person through the air. TB usually affects the lungs, but it can also affect other parts of the body, such as the brain, the kidneys, or the spine. A person with TB can die if they do not get treatment. TB disease can be treated by taking several drugs for 6 to 12 months. It is very important that people who have TB disease finish the medicine, and take the drugs exactly as prescribed. If they stop taking the drugs too soon, they can become sick again; if they do not take the drugs correctly, the germs that are still alive may become resistant to those drugs.

The total data of this disease for this study was (1346) case during (2012 to 2018). The data contains these variables (Years, Gender, Age, Transported or not Transported, Permanent Residence, Nationality, Types of TB disease and Number of times of infection).

Y: Gender (male or female), Response variable encoding in binary response for the criterion variable, in this case it will classify as:

$$Y = \begin{cases} 0 & \text{Female} \\ 1 & \text{Male} \end{cases}$$

Explanatory variables = Ages, transported or not transported, Permanent Residence, Nationalty, Types of TB disease and Number of times of infection.

9. Results of two regression models

This section is divided into two parts; 1) Fitted Logistic Regression and Poisson Regression models. 2) Comparison of the estimated model parameters and identification of the optimal model using the two criterions.

9.1 Fitted Logistic Regression and Poisson Regression models

In this section lists the regression coefficients for both logistic regression model and Poisson regression model along with their standard error values and Wald test statistics for the coefficients of related parameters. There is important to put the hypothesis for both models:

H_0: The model adequately describes the data

H_1: The model does not adequately describe the data

We used the goodness of fit tests, if p-value of Wald test is less than accepted ($\propto = 0.05$) level, the test would reject the null hypothesis of an adequate fit, and if it is not then accepted the null hypothesis. illustrated the output of both models in table accordingly of the software used. As can be observed, some parameters' maximum likelihood coefficients are statistically significant at 5%; statistically significant coefficients. The log-likelihood function (maximum likelihood estimators) of the Logistic regression is used to estimate the parameters denoted as $\beta_0, \beta_1, \dots, \beta_6$. the response variable is (Gender), as categorical variable. The usual way to do this is with an indicator variable. In simple linear regression, we modeled the mean μ_y of the response variable (y) as a linear function of the explanatory variable: $\mu = \beta_0 + \beta_1 x_i$. When y is just (male or female), the mean is the probability (p) of a man. In Logistic regression model the mean (p) in terms of an explanatory variable (x_i). We might try to relate (p) and (x_i) , as in simple linear regression: $p = \beta_0 + \beta_1 x_i$. Unfortunately, this is not a good model. Whenever $(\beta_1 \neq 0)$, extreme values of (x_i) will give values of $\beta_0 + \beta_1 x_i$ that fall outside the range of possible values of (p), $(0 \leq p \leq 1)$. The logistic regression solution to this difficulty is to transform the odds $(\frac{p}{1-p})$ using the natural logarithm. We use the term log odds or logit for this transformation. A model is typically thought of as a simplification of a more complex situation.

The fitted Logistic response function and the fitted values (estimates for the Model) in Table (1) can be expressed as;

$$\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6)} \qquad (30)$$

$$\hat{p}_i = \frac{\exp(3.3793 - 1.4030X_5 - 1.0992X_6)}{1 + \exp(3.3793 - 1.4030X_5 - 1.0992X_6)} \qquad (31)$$

Table (1) Estimated Parameters, Standard Errors, and Wald Test Statistic

	Logistic regression model					Poisson regression model				
	Coef.	**S.E.**	**Wald**	**Sig.**	**Exp(B)**	**Coef.**	**S.E.**	**Wald**	**Sig.**	**Exp(B)**
Cons,	**3.3793**	**0.6814**	**4.96**	**0.000**	**29.350**	**1.3967**	**0.5473**	**2.55**	**0.011**	**4.0418**
X1	**-0.1468**	**0.1475**	**-0.99**	**0.320**	**0.8634**	**-0.0975**	**0.1075**	**-0.91**	**0.364**	**0.9071**
X2	**-0.0050**	**0.0035**	**-1.44**	**0.149**	**0.9950**	**-0.0023**	**0.0026**	**-0.87**	**0.384**	**0.9977**
X3	**0.0852**	**0.1607**	**0.53**	**0.596**	**1.0889**	**0.0418**	**0.1164**	**0.36**	**0.720**	**1.0426**
X4	**0.0586**	**0.1082**	**0.54**	**0.588**	**1.0603**	**0.0311**	**0.0728**	**0.43**	**0.669**	**1.0315**
X5	**-1.4030**	**0.0866**	**-16.19**	**0.000**	**0.2458**	**-0.9064**	**0.0639**	**-14.17**	**0.000**	**0.4039**
X6	**-1.0992**	**0.5127**	**-2.14**	**0.003**	**0.3331**	**-0.7481**	**0.0452**	**-1.66**	**0.098**	**0.4732**

From the results, it is seen that both availability of (X_5) and (X_6) played major roles on the (TB disease)', When the coefficients are examined by the Wald Chi square test, it is seen that (X_5, X_6) of them are significant because their (p-values) of the test less than 5% level, with intercept too. It implies that availability of (types of TB disease) and (Number of times of infection) are the major determining factors on the TB. while (X_1 , X_2 , X_3 , X_4) of estimated coefficient are not significant because their p-values are greater than 5% level.

Considering for the Poisson regression, the log-likelihood function for Poisson regression to estimate the maximum likelihood estimators denoted as (β_0,$\beta_1, \dots, \beta_6$) and it is obtained as in table (1), The Poisson Regression Model with significant parameters estimates as follows:

$$Ln\hat{\lambda} = 1.3967 - 0.9064X_5 \qquad (32)$$

$$\hat{\lambda} = \exp(1.3967 - 0.9064X_5) \qquad (33)$$

Table above also shows that only on availability (X_5: types of TB disease) has main effect on the TB with significance (p-value = 0.000 < 0.05). while (X_6: Number of times of infection) was seen not to have any effect, hence, insignificant.

9.2 Comparison and identification of the optimal model

In this part we present a table of model selection of two types of regression (Logistic and Poisson regression) models by criterions as well as some concepts that mentioned before, their values can be used for

performing model selection based on comparison to models that fit the same data. the table supports the research as follows:

Table (2) Model selection by some criterions for both Logistics and Poisson regression models

	Iteration 0: log likelihood = -607.68524	**Iteration 0: log likelihood = -737.24221**
	Iteration 1: log likelihood = -584.47878	**Iteration 1: log likelihood = -693.97713**
	Iteration 2: log likelihood = -584.24745	**Iteration 2: log likelihood = -693.48056**
	Iteration 3: log likelihood = -584.24728	**Iteration 3: log likelihood = -693.48002**
	Iteration 4: log likelihood = -584.24738	**Iteration 4: log likelihood = -693.48002**
	Logistic regression model	**Poisson regression model**
Number of obs.	**1,346**	**1,346**
Null Deviance	**568.4945**	**682.9600**
Pearson χ^2	**9245.709**	**9335.277**
AIC	**0.878524**	**1.040832**
BIC	**-8478.857**	**-8964.391**
Log likelihood	**-584.24727**	**-693.4800**
Variance function	$\mathbf{V}(\boldsymbol{p}) = \mathbf{p} * (\mathbf{1} - \mathbf{p})$	$\mathbf{V}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}$
Link function	$\mathbf{g}(\mathbf{u}) = \mathbf{ln}(\mathbf{u}/(\mathbf{1} - \mathbf{u}))$	$\mathbf{g}(\boldsymbol{\lambda}) = \mathbf{ln}(\boldsymbol{\lambda})$

From the results of table (2) show the logistic regression model values of their criterions have the lowest of all. Then, it can be seen that the value of Null Deviance, Pearson χ^2 and Log likelihood for logistic regression model equal to (568.4945, 9245.709, -584.24727) respectively, and their values less than Poisson regression model. where the Deviance statistic is distributed approximated a Chi-square distribution. The null Deviance equal as Chi square distributed with the model degree of freedom (1). To assesses the fit of the two models the best model is the one which has the lowest AIC and BIC and here equal to (0.878524, -8478.857) for logistic regression model. The Poisson regression model of the dataset had the largest values of all criterions and indicating a poor fit to the data. Using logistic regression model for this dataset and in the case of binary response variable is a good alternative of Poisson regression model. it had the smaller values of that criterions when compared with Poisson regression model. This indicating that logistic regression model better goodness of fit.

10. Conclusion

In the last section we have studied the performance of each model, the last comparison reaches some conclusions that differ from the results:

1. The best fit for the data of TB disease is the logistic regression model.
2. The best logistic regression model identified is when the explanatory variables are (X_5: Types of TB Disease) and (X_6:Number of times of infection) (since, there are coefficients that are significant at 5% with p-value 0.000 and 0.003 respectively.
3. The Poisson regression model also identified that had coefficient significant at5% with p-value 0.000 for the explanatory variable are (X_5: Types of TB Disease).
4. Results showed that the Logistic regression model is the best fit in modeling binary response variable in form of a count data; based on the two assessment criteria employed Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)].
5. The research identified Logistic regression model to be more suitable for the observation considered.

References:

Adeti, F. (2016): "Modelling Count Outcomes from Dental Caries in Adults: A Comparison of Competing Statistical Models", A thesis Submitted to the department of Mathematics, Kwame Nkrumah University of Science and Technology, Kumasi.

Armstrong, J. S. (2012), Illusions in Regression Analysis, International Journal of forecasting. 28 (3), 689.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. Critical Care (London, England), 9(1), 112-118.

Cameron, A. C.; Trivedi, P. K. (1998): " Regression analysis of count data", Cambridge University Press. ISBN 978-0-521-63201-0.

Durrant, G. (2016): "Poisson Regression Modes for Count Data", Available from: https://www.slideshare.net/synchrony/poisson-regression-models-for-count-data63688148

Durmuş, B. and Güneri, Ö. İ. (2020): "An Application of the Generalized Poisson Model for Over Dispersion Data on The Number of Strikes Between 1984 and 2017", Journal of Operations Research, Statistics, Econometrics and Management Information Systems Volume 8, Issue 2

Henry, G. A. (2010). Comparison of Akaike Information Criterion (AIC) and Bayesian Criterion (BIC) in Selection of an Asymmetric Price Relationship.

Journal of Development and Agricultural Economics, Vol 2 (1), page 001-006.

Hilbe, Joseph M. (2015). Practical Guide to Logistic Regression, International Standard Book Number-13: 978-1-4987-0958-3, California Institute of Technology, USA and Arizona State University, USA.

Ijomah, M. A., Biu, E. O., & Mgbeahurike, C. (2018). Assessing Logistic and Poisson Regression Model in Analyzing Count Data, International Journal of Applied Science and Mathematical Theory ISSN 2489-009X Vol. 4 No. 1

Karazsia, B. and Van Dulmen M. (2008): "Regression Models for Count Data: illustrations using longitudinal predictors of childhood injury", 33(10):1076-84 Available from: https://www.ncbi.nlm.nih.gov/pubmed/18522994. [Accessed on 14 July 2017].

Moksony, F. and Hegedus, R. (2014): "The Use of Poisson Regression in the Sociological Study of Suicide", Corvinus Journal of Sociology and Social Policy, 5(2), 97-114.

Michael, H. K.; Christopher, J. N., John N., and William. L (2005). "Applied Linear Statistical Model", fifth Ed.; 555-623., McGraw Hill International, New York.

Nduka, E.C., (1999), Principles of Applied Statistics 1, Crystal Publishers, Okigwe.

Park, Hyeoun-Ae (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain, journal of Korean Acad. Nurs. Vol.43 No.2

Using Neural Networks to Forecast the Electricity Generation in Kurdistan Region-IRAQ

Wasfi T. Saalih Kahwachi
Research Center Director
Tishk International University Erbil, Iraq
wasfi.kahwachi@tiu.edu.iq
Samyia Khalid Hasan
Statistics and Informatics Department
Salahaddin University-Admin. & Economics Erbil, Iraq
saiya.hasan@su.edu.krd

Abstract

Forecasting the future behavior of time series is an essential issue in statistical sciences since it is required in many aspects of life. Iraq and its region are facing a real problem in electrical power generation. For this purpose, forecasting the power generation has drawn the attention. Electric power generation in the Kurdistan Regain Iraqi (KRI) is crucial because it distributes the power in the region, Kirkuk and Mosul. Data were collected from (Iraqi-Kurdistan Regional Government Ministry of Electricity, General Directorate of Control & Communication, Kurdistan Dispatch Control Center).

in this research were used Artificial neural networks were used by the method of inverse propagation of error and the selection of the best model using to forecasting elect the minimum value of MSE (Mean Square Error) by using the statistical criteria to forecasting this data.

أ.م.د. وصفي طاهر صالح
جامعة تيشك الدولية أربيل، العراق
wasfi.kahwachi@tiu.edu.iq

أ. م.د.سامية خالد حسن
جامعة صلاح الدين-اربيل /كلية الادارة والاقتصاد
saiya.hasan@su.edu.krd

الملخص

يعتبر التنبؤ بالسلوك المستقبلي للسلاسل الزمنية مسألة أساسية في العلوم الإحصائية حيث أنها مطلوبة في العديد من جوانب الحياة ، حيث يواجه العراق ومنطقته مشكلة حقيقية في توليد الطاقة الكهربائية. لهذا الغرض ، لفت التنبؤ بتوليد الطاقة . يعد توليد الطاقة الكهربائية في كردستان-العراق أمرًا بالغ الأهمية لأنه يوزع الطاقة الى محافظة كركوك والموصل. تم جمع البيانات من (وزارة الكهرباء في حكومة إقليم كردستان العراق ، المديرية العامة للرقابة والاتصالات ، مركز مراقبة الإرسال في كردستان).

في هذا البحث تم استخدام الشبكات العصبية الاصطناعية بطريقة الانتشار العكسي للخطأ واختيار أفضل نموذج باستخدام المعيار الاحصائي أقل قيم للمجموع مربعات الخطأ التنبؤ بهذه البيانات.

1. NTRODUCTION

ANN forecasting is a technology that has attracted significant interest in a range of fields, including currency rates, financial resources, meteorological conditions, river flow, and so on. ANN is commonly used to explain the behavior of non-linear data because they do not need rigid and precise parameters for prediction.

2. Theory Part

2.1 Characteristics of ANN

ANN have a few basic properties. Among them are the following; Competitive learning takes place among the output layer's neurons, i.e., only one neuron wins the competition when an input pattern is presented. In most cases, the neurons are organized in a two-dimensional lattice. The neurons are selectively tuned to different input patterns and their positions are organized in relation to one another, resulting in a meaningful coordinate system for distinct input characteristics across the lattice. A distinctive feature is the formation of a topographic map of the input patterns. The neurons (coordinates) are the fundamental statistical properties present in the input patterns. The creation of this paradigm was sparked by the discovery of topologically organized computational maps in the human brain. However, it is unique in its ability to receive and process the transmission of an electrical signal along the length of the nerve that makes up the brain's association system. The neuron is made up of components dendrite and is a group of entrance through which cell receives information from neighboring cells. The human brain is composed of a huge number of neurons that have complex internal connections that have complex internal connections that make up a large network of nerves (neurons) that share some characteristics with each other with other cells of the body. Each cell contains three parts: the cell body and the dendritic branches (Dendrites) and axon (Axon). Dendrites extend from the cell body to other nerves to form a neural network, and the neural ganglion (Synapse) represents a pathway or gateway to connect the dendritic branching coming from another nerve. The inputs reach the cell body, some excite the cell and stimulate it and others inhibit it and the ganglion (nerve) works to integrate or accumulate signals in the cell body, and when it exceeds the required level (Threshold) of the cell it will be inhibited and passed to the other cell along the axon, and branching The dendritic modulation (changing) the amplitude of the signal transmitted through it, and this modulation changes with time, as in the

learning process of an artificial neural network. It can be seen in Fig. 1 and Fig. 2 is shown the ANN to simulate the basic properties of a biological neuron, where the input connections are represented by lines corresponding to the dendritic branches, which is the output to another nerve. Biological and weighted inputs are collected in the summation box that corresponds to the body of the neuron to determine its level of influence to produce the output signal representing the input to other cells associated with it. In terms of NET, computed with the following formula. Net=*WX*, where *X*: $x_1, x_2, \ldots, x_n$ represented the input vector, $w_1, w_2, \ldots, w_m$ represented the weights vector (Hagan et al., 1996; Poznyak et al, 2001; Stergiou and Siganos, 1996).

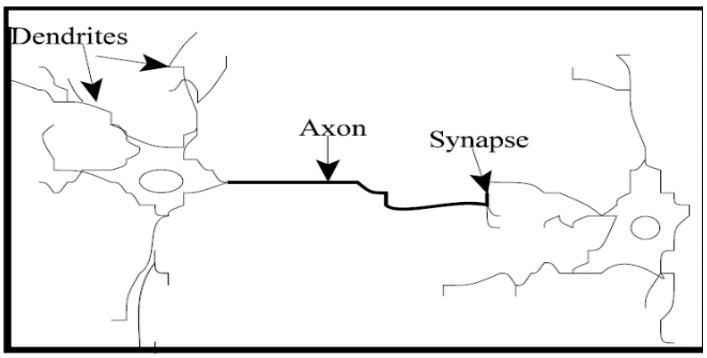
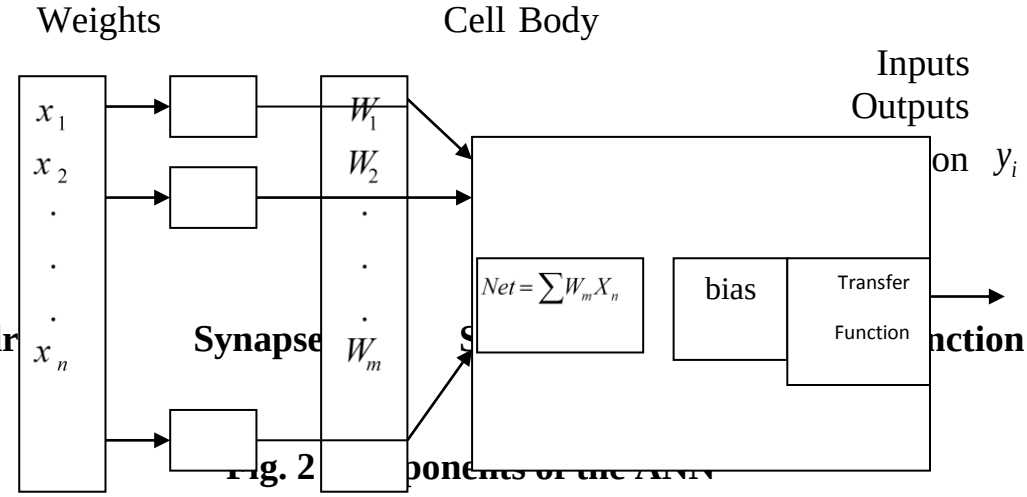


Fig. 1 Components of a Nerve Cell



Dendr Synapse S nction

Fig. 2 ponents of the ANN

2.2 Neural Network using Back Propagation

In the 1970s, the main cause of the decline in interest in neural networks is due to the limitations of a single-layer network. In these networks a collection of inputs is used to determine a set of outputs. These networks employ the supervised teaching approach. The objective of the training process for these networks is to compromise between the network's

ability to respond properly to the training input samples stored in the network to respond effectively to input that is similar but not identical to the training input samples.

2.3 The Back Propagation of Input Training Samples Stage

1. The stage of forward feeding of the input training samples.
2. Stage computation and back propagation of output error.
3. Stage synthesis (determination) of weights.

The network testing phase comes after the training phase, and the forward propagation phase is only one of the stages. The error reverse propagation network training by using the minimum value of the total error square of the outputs computed by the network, which is dependent on the magnitude of the error. Weights are updated between layers until the ideal weights are identified that provide the best fit for the model. The basic purpose of a back propagation network is to minimize error until it has learnt through training. Training tries to adjust the network starting with random weights until the error is as reduced as feasible.

2.4 Back Propagation ANN in Time Series Forecasting

The following stages outline the ANN prediction process

The first step: Selection of Variables

The observations that represent the problem whose values to be forecasted are carefully selected in this stage.

The second stage: is to process the data.

On the observations used, some operations are conducted, such as assessing the final trend, focusing on the relationships between the observations, and determining the data distribution.

The third step: Divide data into sets:

The data is divided into the following groups:

a. **Training Sets:** It is the learning set and the identification of the data model**.**

b. **Testing Sets:** It is a set of estimating the skill of the virtual network and its ability to use it in general.

c. **Validation Set:** It is a set to perform a final test of network performance.

The fourth step: ANN Paradigms:

When defining a neural network model, the following must be selected:

1. The number of input neurons, which must be invariably equal to the number of variables.
2. The number of hidden layers which depends on the error value used in the network.
3. Experiment determines the amount of hidden neurons.

4. A single output neuron.

5. The fifth step: Transfer Function

Typically, the mathematical method is used to calculate the output that prevents the output from surpassing a certain threshold. This formula usually uses one of the following functions:

a. Linear Function
b. Threshold Function
c. Sigmoid Function

The sixth step: Evaluation Criteria

At this step, the total of the error squares is used to evaluate the network error (MSE).

The seven step train with ANN

The following are included in this step:

a. Choosing the set of weights between neurons that minimizes the error is the first step in teaching the model.
b. The regression training algorithm is employed as the second algorithm (slope reduction).

The eighth step: Implementation: This is a critical stage since it evaluates the network's ability to adapt to the state of the change in the cycle, as well as its ability to retrain and access the data with the least amount of mistake possible.

2.5 ANN Performance

The quality of future predictions for a certain phenomenon may be determined by the ANN's training evaluation (Hangan et al., 1996).

1- Learning Rate (α)

The learning rate is one of the factors impacting the process of updating the weights in the neural network since it sets the size of the step in the learning process and the amount of weight change.

2-Momentum (η)

It is one of the important factors that balances the learning process and makes the amount of weight change relatively balanced and stable.

3- Number of ANN Exemplars

The number of vectors (Exemplars) has a direct impact on the network's performance since it reflects the explanatory variables. If the number of vectors is suitable, the ANN may create a model that describes the data. If the network inputs are too complicated, the number of vectors must be raised for the network to learn the behavior of the data.

4- Number of Hidden Nodes

The best way to find out how many hidden nodes a neural network needs to start with a few hidden nodes, watch the results, and then increase the number of hidden nodes until the lowest possible error is achieved and the best results in the comparison criteria.

5- Number of ANN Exemplars

The important factors in the efficiency of neural network training are that it starts with a single hidden level. The attributes, i.e. the data attributes of the neural network, are trained or learned until the lowest possible error is attained. Another hidden level is linked to the network if the neural network does not learn the most of the data's characteristics.

2.6 Elements of ANN

Using the characteristics listed above, the following basic elements of any ANN must be extracted from.

a. Element Processing
b. Topology
c. Algorithm for Learning

3. Application Part

3.1 Using Neural Networks for Forecasting

ANN was applied to alter nonlinear time series and improve prediction accuracy. The first step is to find out the neural network's hidden nodes. It is determined through extensive training, which includes several computer experiments. The following formula was used to compute the number of concealed nodes:

$$N_{hidden} \leq \frac{N_{train} E_{tolerance}}{N_{pts} + N_{output}} \qquad (1)$$

Whereas:

N_{hidden} : The number of hidden nodes.

N_{train} : The number of training time.

$E_{tolerance}$: The amount of the probability error.

N_{pts}: The number of data which is training was performed.

N_{output} : The number of output nodes.

In this dissertation, $E_{tolerance}$the amount of the error 0.01 and the training times N_{train}=1000* N_{pts}=168 and N_{output}=1. By using equation (1), the number of hidden nodes was 4.

1- Choosing the number of hidden neurons in the network

Use all of the observations in Table 1 and the back propagation error network, which was trained using several hidden neurons to choose the

best one. It was found out that it lies between (3-9) and the number of iterations is (1000) to get the minimum value of MSE.

Table 1 Selecting the Number of Hidden Neurons for ANN

Number of Iteration	Test (MSE)	Validation (MSE)	Training (MSE)	ALL(MSE)	Number of Hidden Neurons
1000	1.51E-03	1.32E-03	9.53E-04	4.66E-03	3
1000	4.22E-03	2.60E-05	2.15E-05	2.00E-03	4
1000	2.63E-01	3.61E-03	2.52E-03	6.10E-03	5
1000	1.87E-01	9.04E-03	2.52E-01	9.04E-03	6
1000	5.10E-01	7.18E-03	2.26E-02	6.80E-01	7
1000	1.57E-02	1.40E-02	8.30E-03	1.40E-02	8
1000	8.77E-01	2.97E-04	5.37E-05	3.92E-03	9

2- The performance of the neural network was tested at different percentages of data division according to neuron 4, with the best model being the minimum value of MSE, as shown in Table 1.

Table 2 Performance of Hidden Neurons ANN at Different Data Partition Ratios

Number of Hidden Neurons	Test%	Validation%	Train%	Training (MSE)	Validation (MSE)	Test (MSE)	MSE All
4	20%	15%	65%	2.29E-01	6.68E-01	2.91E+05	6.83E-02
4	15%	20%	65%	8.90E-03	3.94E-02	3.17E-02	3.94E-02
4	10%	10%	80%	1.24E-02	1.48E-02	3.30E-03	1.48E-02
4	20%	20%	60%	1.68E-02	1.63E-02	8.54E-01	1.63E-02
4	25%	5%	70%	3.87E-04	2.15E-04	7.49E-01	2.15E-03
4	10%	15%	75%	2.29E-01	6.68E-01	2.91E+05	6.83E-02
4	15%	15%	70%	2.15E-05	2.60E-05	4.22E-03	2.00E-03

The ratio of testing, validation, and training of the neural network's performance at Number of Hidden Neurons is shown in Table 2. The data used in this dissertation consists of 168 observations, which have been processed using the sigmoid function in the hidden layer and many empirical tests, with a vector ratio of 15% (25 observations) for the test set, 15% (25 observations) for the validation set, and 70% for the training set (118 observations). To get the best possible result.

Fig. 3 shows the performance of the training, Validation, and test groups stooping on the point of tiring iteration (1000), with MSE of Validation being 0.0044661.

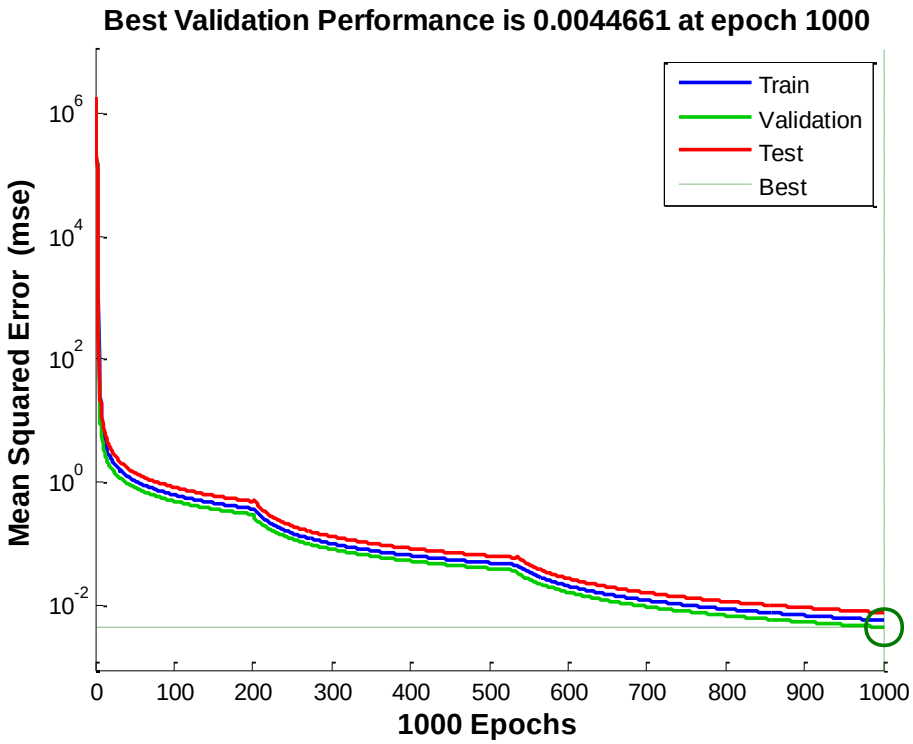


Fig. 3 Performance of the ANN (Training, Validation, and Test)

3.2 Conclusions

1. The best model of the neural network models is the (1, 4, 1) model among the others models because it has the smallest value MSE in test, Validation, Training and ALL(MSE) in the differences hidden of Number of Neurons.
2. By using the sigmoid function in the hidden layer and many empirical tests, with 25 observations for the test set, 25 observations for the validation set, and 118 observations for the training set, it is the best possible result.

3.3 Recommendations

1. Depending on the back propagation neural networks to forecasting value to generate electricity power in Kurdistan Region.
2. Using back propagation neural networks with multivariate time series.

3.4 **References**

1. BISHOP, C. (1995) Neural Networks for Pattern Recognition. Oxford University Press.
2. GERSHENSON, C. (1998) Artificial Neural Networks for Beginners. Sussex Academy, UK.
3. HAGAN, M., Demuth, H. & BEALE, M. (1996) Neural Network Design. PWS Publishing Co., USA.
4. ISSA, Z. (2000) *Neural Network Architecture Algorithms Applications*.
5. PONZNYAK, A., SANCBEZ, E. & YU, W. (2001) Differential Neural Networks for Robust Nonlinear Control. World Scientific publishing Co., Singapore.
6. SARLE, W. (1994) Neural Networks and Statistical Models. Proceeding of the Nineteenth Annual SAS Vsers Group International Conference, AS Institute Inc., Cary, NC, USA.
7. SINHA, H. (2002) Designing a Neural Network for Forecasting Finan- cial and Economic Time Series. June 4th, ind E 593.
8. S
TERGIOU, CHR. & SIGANOS, D. (1996) Neural Networks. www.dse.doc.ic.ac.uk/~nd/ Suprise Journal, Vol. l4 / sll/report.html
9. XU, S., CHIN, L. (2008). A Novel Approach for Deterring the Optimal Number of Hidden Layer Neurons for FNNs and its Application in Data Ming. 5th International Conference on information Technology and Application (ICITA).
10. ZURADA, M. & CHOLEW, J. (1994) Introduction Artificial Neural Systems. In Proceedings of the IEEE International Joint Conference on Neural Networks, Houston, Texas, USA.

Appendix
KURDISTAN REGION POWER GENERATION (2006-2019)
(X1) Government

Date	Erbil Average Load	Suly Average Load	Duhok Average Load	Cement Fac. Average Load	Delta Fac. Average Load	Iron Fac. Average Load	Azmeer Steel Fac. Average Load
	MW	MW	MW	MW	MW	MW	MW
2006							
Jan	114	166	153	0	0	0	0
Feb	74	185	169	0	0	0	0
Mar	120	184	167	0	0	0	0
Apr	110	168	161	0	0	0	0
May	106	182	156	0	0	0	0
Jun	112	290	155	0	0	0	0
Jul	202	274	159	0	0	0	0
Aug	202	283	156	0	0	0	0
Sep	124	173	153	0	0	0	0
Oct	100	143	148	0	0	0	0
Nov	105	145	156	0	0	0	0
Dec	106	176	145	0	0	0	0
2007							
Jan	98	142	123	0	0	0	0
Feb	101	191	67	0	0	0	0
Mar	119	149	140	0	0	0	0
Apr	123	150	175	0	0	0	0
May	110	136	170	0	0	0	0
Jun	183	227	174	0	0	0	0
Jul	186	230	151	0	0	0	0
Aug	187	234	169	0	0	0	0
Sep	172	202	165	0	0	0	0
Oct	161	197	165	0	0	0	0
Nov	140	172	137	0	0	0	0

Dec	**148**	**179**	**105**	**0**	**0**	**0**	**0**
2008							
Jan	**97**	**114**	**38**	**0**	**0**	**0**	**0**
Feb	**81**	**99**	**62**	**0**	**0**	**0**	**0**
Mar	**94**	**116**	**125**	**0**	**0**	**0**	**0**
Apr	**98**	**120**	**132**	**0**	**0**	**0**	**0**
May	**128**	**149**	**107**	**0**	**0**	**0**	**0**
Jun	**170**	**210**	**109**	**0**	**0**	**0**	**0**
Jul	**195**	**236**	**103**	**0**	**0**	**0**	**0**
Aug	**187**	**228**	**113**	**0**	**0**	**0**	**0**
Sep	**165**	**190**	**120**	**0**	**0**	**0**	**0**
Oct	**208**	**225**	**132**	**0**	**0**	**0**	**0**
Nov	**252**	**261**	**133**	**0**	**0**	**0**	**0**
Dec	**284**	**287**	**137**	**0**	**0**	**0**	**0**
2009							
Jan	**281**	**285**	**140**	**0**	**0**	**0**	**0**
Feb	**273**	**284**	**142**	**0**	**0**	**0**	**0**
Mar	**277**	**274**	**139**	**0**	**0**	**0**	**0**
Apr	**242**	**274**	**141**	**0**	**0**	**0**	**0**
May	**291**	**299**	**150**	**0**	**0**	**0**	**0**
Jun	**292**	**301**	**148**	**0**	**0**	**0**	**0**
Jul	**337**	**373**	**131**	**0**	**0**	**0**	**0**
Aug	**355**	**361**	**133**	**0**	**0**	**0**	**0**
Sep	**355**	**331**	**132**	**0**	**0**	**0**	**0**
Oct	**333**	**312**	**135**	**0**	**0**	**0**	**0**
Nov	**386**	**375**	**136**	**0**	**0**	**0**	**0**
Dec	**414**	**407**	**140**	**0**	**0**	**0**	**0**
2010							
Jan	**425**	**448**	**149**	**0**	**0**	**0**	**0**
Feb	**397**	**462**	**197**	**0**	**0**	**0**	**0**
Mar	**404**	**462**	**237**	**0**	**0**	**0**	**0**
Apr	**357**	**421**	**237**	**0**	**0**	**0**	**0**
May	**339**	**399**	**223**	**0**	**0**	**0**	**0**
Jun	**397**	**450**	**237**	**0**	**0**	**0**	**0**

Jul	**404**	**458**	**238**	**0**	**0**	**0**	**0**
Aug	**412**	**468**	**233**	**0**	**0**	**0**	**0**
Sep	**369**	**420**	**220**	**0**	**0**	**0**	**0**
Oct	**337**	**388**	**206**	**0**	**0**	**0**	**0**
Nov	**368**	**448**	**236**	**0**	**0**	**0**	**0**
Dec	**420**	**470**	**253**	**0**	**0**	**0**	**0**
2011							
Jan	**449**	**455**	**259**	**42**	**0**	**0**	**0**
Feb	**440**	**460**	**281**	**35**	**0**	**0**	**0**
Mar	**486**	**496**	**296**	**48**	**0**	**0**	**0**
Apr	**439**	**461**	**284**	**48**	**0**	**0**	**0**
May	**472**	**481**	**272**	**44**	**0**	**0**	**0**
Jun	**536**	**540**	**320**	**43**	**0**	**0**	**0**
Jul	**570**	**581**	**345**	**47**	**0**	**0**	**0**
Aug	**608**	**608**	**358**	**49**	**0**	**0**	**0**
Sep	**548**	**530**	**322**	**56**	**0**	**0**	**0**
Oct	**448**	**443**	**293**	**57**	**0**	**0**	**0**
Nov	**640**	**648**	**426**	**66**	**0**	**0**	**0**
Dec	**660**	**687**	**454**	**65**	**0**	**0**	**0**
2012							
Jan	**671**	**739**	**453**	**49**	**0**	**0**	**0**
Feb	**676**	**774**	**484**	**59**	**0**	**0**	**0**
Mar	**672**	**755**	**492**	**91**	**0**	**0**	**0**
Apr	**543**	**522**	**335**	**90**	**0**	**0**	**0**
May	**580**	**541**	**334**	**100**	**0**	**0**	**0**
Jun	**687**	**642**	**393**	**100**	**0**	**0**	**0**
Jul	**774**	**715**	**435**	**102**	**0**	**0**	**0**
Aug	**771**	**746**	**442**	**81**	**0**	**0**	**0**
Sep	**701**	**603**	**387**	**107**	**0**	**0**	**0**
Oct	**573**	**490**	**337**	**94**	**0**	**0**	**0**
Nov	**666**	**622**	**402**	**91**	**0**	**0**	**0**
Dec	**787**	**783**	**511**	**92**	**0**	**0**	**0**
2013							
Jan	**838**	**821**	**522**	**91**	**0**	**0**	**0**

Feb	**850**	**791**	**510**	**94**	**0**	**0**	**0**
Mar	**771**	**675**	**445**	**105**	**0**	**0**	**0**
Apr	**638**	**529**	**336**	**104**	**0**	**0**	**0**
May	**654**	**538**	**363**	**96**	**0**	**0**	**0**
Jun	**842**	**691**	**450**	**113**	**0**	**0**	**0**
Jul	**989**	**791**	**528**	**104**	**0**	**0**	**0**
Aug	**967**	**784**	**512**	**108**	**0**	**0**	**0**
Sep	**802**	**663**	**424**	**122**	**0**	**0**	**0**
Oct	**639**	**533**	**373**	**118**	**0**	**0**	**0**
Nov	**784**	**701**	**463**	**111**	**0**	**0**	**0**
Dec	**1007**	**891**	**673**	**104**	**0**	**0**	**0**
2014							
Jan	**998**	**907**	**672**	**104**	**0**	**19**	**0**
Feb	**948**	**849**	**623**	**118**	**0**	**28**	**3**
Mar	**824**	**753**	**500**	**104**	**0**	**26**	**4**
Apr	**758**	**657**	**433**	**106**	**0**	**22**	**4**
May	**805**	**639**	**437**	**120**	**0**	**24**	**3**
Jun	**905**	**747**	**508**	**101**	**0**	**19**	**3**
Jul	**984**	**834**	**565**	**56**	**0**	**11**	**3**
Aug	**976**	**882**	**592**	**48**	**0**	**2**	**2**
Sep	**890**	**718**	**499**	**89**	**0**	**21**	**2**
Oct	**749**	**617**	**432**	**90**	**0**	**21**	**3**
Nov	**862**	**769**	**576**	**83**	**0**	**19**	**3**
Dec	**973**	**967**	**668**	**79**	**3**	**22**	**4**
2015							
Jan	**1022**	**896**	**691**	**0**	**0**	**0**	**0**
Feb	**989**	**883**	**657**	**0**	**0**	**0**	**0**
Mar	**960**	**825**	**615**	**0**	**0**	**0**	**0**
Apr	**854**	**695**	**509**	**0**	**0**	**0**	**0**
May	**858**	**663**	**466**	**0**	**0**	**0**	**0**
Jun	**949**	**728**	**530**	**0**	**0**	**0**	**0**
Jul	**1012**	**791**	**576**	**0**	**0**	**0**	**0**
Aug	**1013**	**776**	**565**	**0**	**0**	**0**	**0**
Sep	**948**	**720**	**534**	**0**	**0**	**0**	**0**

Oct	**797**	**636**	**450**	**0**	**0**	**0**	**0**
Nov	**919**	**787**	**566**	**0**	**0**	**0**	**0**
Dec	**960**	**768**	**619**	**0**	**0**	**0**	**0**
2016							
Jan	**880**	**603**	**764**	**0**	**0**	**0**	**0**
Feb	**786**	**550**	**691**	**0**	**0**	**0**	**0**
Mar	**801**	**482**	**655**	**0**	**0**	**0**	**0**
Apr	**803**	**465**	**667**	**0**	**0**	**0**	**0**
May	**835**	**492**	**664**	**0**	**0**	**0**	**0**
Jun	**930**	**530**	**717**	**0**	**0**	**0**	**0**
Jul	**1018**	**582**	**663**	**123**	**0**	**0**	**0**
Aug	**1009**	**588**	**668**	**129**	**0**	**0**	**0**
Sep	**909**	**530**	**640**	**107**	**0**	**0**	**0**
Oct	**852**	**470**	**555**	**70**	**0**	**0**	**0**
Nov	**834**	**499**	**592**	**77**	**0**	**0**	**0**
Dec	**805**	**561**	**600**	**104**	**0**	**0**	**0**
2017							
Jan	**840**	**582**	**640**	**108**	**0**	**0**	**0**
Feb	**845**	**554**	**621**	**107**	**0**	**0**	**0**
Mar	**884**	**528**	**629**	**93**	**0**	**0**	**0**
Apr	**833**	**493**	**591**	**79**	**0**	**0**	**0**
May	**804**	**463**	**556**	**96**	**0**	**0**	**0**
Jun	**923**	**502**	**597**	**117**	**0**	**0**	**0**
Jul	**918**	**507**	**606**	**124**	**0**	**0**	**0**
Aug	**856**	**476**	**603**	**117**	**0**	**0**	**0**
Sep	**847**	**476**	**621**	**106**	**0**	**0**	**0**
Oct	**772**	**431**	**554**	**87**	**0**	**0**	**0**
Nov	**802**	**454**	**596**	**85**	**0**	**0**	**0**
Dec	**841**	**537**	**639**	**106**	**0**	**0**	**0**
2018							
Jan	**832.8**	**539.1**	**678.4**	**112.4**	**0**	**0**	**0**
Feb	**858.7**	**556.2**	**699.9**	**112.2**	**0**	**0**	**0**
Mar	**879**	**513.2**	**685.1**	**92.3**	**0**	**0**	**0**
Apr	**792.5**	**453**	**600.5**	**85.2**	**0**	**0**	**0**

May	**830.6**	**481.8**	**608.4**	**105.9**	**0**	**0**	**0**
Jun	**897**	**459**	**587**	**119**	**0**	**0**	**0**
Jul	**868**	**451**	**570**	**108**	**0**	**0**	**0**
Aug	**878**	**454**	**575**	**102**	**0**	**0**	**0**
Sep	**811.2**	**454.9**	**594.3**	**98.2**	**0**	**0**	**0**
Oct	**795.7**	**448.6**	**576.5**	**89.7**	**0**	**0**	**0**
Nov	**943.7**	**527.4**	**722.3**	**96.5**	**0**	**0**	**0**
Dec	**985**	**636.5**	**794.1**	**111.4**	**0**	**0**	**0**
2019							
Jan	**1042**	**641**	**823**	**126**	**0**	**0**	**0**
Feb	**1032**	**575**	**809**	**127**	**0**	**0**	**0**
Mar	**1070**	**594**	**816**	**120**	**0**	**0**	**0**
Apr	**1090**	**570**	**847**	**99**	**0**	**0**	**0**
May	**1029**	**551**	**691**	**122**	**0**	**0**	**0**
Jun	**1137**	**530**	**736**	**148**	**0**	**0**	**0**
Jul	**1105**	**582**	**717**	**142**	**0**	**0**	**0**
Aug	**1132**	**594**	**734**	**135**	**0**	**0**	**0**
Sep	**1072**	**585**	**741**	**129**	**0**	**0**	**0**
Oct	**913**	**521**	**652**	**102**	**0**	**0**	**0**
Nov	**924**	**583**	**744**	**90**	**0**	**0**	**0**
Dec	**1164**	**663**	**847**	**127**	**0**	**0**	**0**

(X2) Kurdistan Region Government to Mousul and Kirkuk

Date	Inter-Connections (MW)	Inter-Connections (MW)	Date	Inter-Connections (MW)	Inter-Connections (MW)
	From KRG to Mousel	From KRG to Kurkuk		From KRG to Mousel	From KRG to Kurkuk
2006			2013		
Jan	32	0	Jan	51	267
Feb	26	0	Feb	50	264
Mar	30	0	Mar	49	238
Apr	24	0	Apr	45	221
May	34	0	May	42	221
Jun	37	0	Jun	53	224
Jul	38	0	Jul	91	234
Aug	40	0	Aug	91	236
Sep	44	0	Sep	98	235
Oct	38	0	Oct	97	216
Nov	25	0	Nov	87	224
Dec	26	0	Dec	74	257
2007			2014		
Jan	13	0	Jan	65	255
Feb	4	0	Feb	74	253
Mar	0	0	Mar	84	216
Apr	11	0	Apr	53	222
May	25	0	May	78	187
Jun	18	0	Jun	88	223
Jul	18	0	Jul	65	211
Aug	1	0	Aug	11	195
Sep	0	0	Sep	2	215
Oct	0	0	Oct	2	209
Nov	0	0	Nov	4	218
Dec	0	0	Dec	6	247
2008			2015		
Jan	0	0	Jan	6	253
Feb	0	0	Feb	6	246
Mar	0	0	Mar	11	239

Apr	**0**	**0**	**Apr**	**18**	**255**
May	**0**	**0**	**May**	**21**	**222**
Jun	**0**	**0**	**Jun**	**23**	**221**
Jul	**0**	**0**	**Jul**	**-3**	**217**
Aug	**0**	**0**	**Aug**	**-4**	**215**
Sep	**0**	**0**	**Sep**	**-3**	**203**
Oct	**0**	**0**	**Oct**	**-3**	**218**
Nov	**0**	**0**	**Nov**	**20**	**217**
Dec	**0**	**0**	**Dec**	**34**	**259**
2009			**2016**		
Jan	**0**	**0**	**Jan**	**33**	**215**
Feb	**0**	**0**	**Feb**	**17**	**218**
Mar	**0**	**0**	**Mar**	**10**	**188**
Apr	**0**	**0**	**Apr**	**3**	**192**
May	**0**	**0**	**May**	**12**	**204**
Jun	**0**	**0**	**Jun**	**17**	**205**
Jul	**0**	**0**	**Jul**	**18**	**204**
Aug	**0**	**0**	**Aug**	**19**	**205**
Sep	**0**	**0**	**Sep**	**12**	**199**
Oct	**0**	**0**	**Oct**	**7**	**191**
Nov	**0**	**0**	**Nov**	**28**	**194**
Dec	**0**	**0**	**Dec**	**38**	**211**
2010			**2017**		
Jan	**0**	**0**	**Jan**	**41**	**213**
Feb	**0**	**0**	**Feb**	**41**	**208**
Mar	**0**	**0**	**Mar**	**34**	**202**
Apr	**0**	**0**	**Apr**	**17**	**190**
May	**0**	**0**	**May**	**5**	**202**
Jun	**0**	**0**	**Jun**	**8**	**210**
Jul	**0**	**0**	**Jul**	**82**	**211**
Aug	**0**	**0**	**Aug**	**364**	**212**
Sep	**0**	**0**	**Sep**	**357**	**203**
Oct	**0**	**0**	**Oct**	**224**	**147**
Nov	**0**	**0**	**Nov**	**274**	**181**
Dec	**0**	**0**	**Dec**	**368**	**118**

2011			2018		
Jan	0	0	Jan	53	281
Feb	0	0	Feb	9	119
Mar	0	0	Mar	0	0
Apr	0	0	Apr	83	519
May	0	0	May	99	375
Jun	0	20	Jun	106	381
Jul	0	150	Jul	106	343
Aug	0	199	Aug	103	291
Sep	0	189	Sep	107	281
Oct	0	154	Oct	64	266
Nov	0	201	Nov	8	399
Dec	0	201	Dec	102	349
2012			2019		
Jan	0	181	Jan	264	101
Feb	0	241	Feb	289	103
Mar	0	242	Mar	297	102
Apr	0	193	Apr	282	93
May	0	228	May	320	107
Jun	44	216	Jun	357	108
Jul	47	219	Jul	318	107
Aug	45	225	Aug	329	107
Sep	44	232	Sep	338	110
Oct	44	200	Oct	241	102
Nov	46	210	Nov	262	103
Dec	48	264	Dec	292	112

(X3) Factories

Date	Mass Cement Fac.	Bazyan Cement Fac.	Taslija Cement Fac.	Delta Cement Fac.	Polteks Steel Fac.	Ezmeer Steel Fac.	Super Steel Fac.	Mass Steel Fac.	KAR Oil Refinery	Gasn Cement	Qarachokh Cement
2006											
Jan	0	0	0	0	0	0	0	0	0	0	0
Feb	0	0	0	0	0	0	0	0	0	0	0
Mar	0	0	0	0	0	0	0	0	0	0	0

Apr	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
May	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jun	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jul	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Aug	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Sep	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Oct	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Nov	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Dec	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
2007											
Jan	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Feb	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Mar	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Apr	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
May	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jun	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jul	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Aug	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Sep	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Oct	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Nov	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Dec	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
2008											
Jan	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Feb	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Mar	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Apr	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
May	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jun	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Jul	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Aug	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Sep	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Oct	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Nov	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Dec	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Comparison Multivariate Time Series Model VAR(P) and wavelet Transformation to forecast Water Supply of Iraqi Rivers

Dr. Nabeel G. Nacy Nabeel.sulaiman@su.edu.krd
Dr. Mohammed A. Badal Mohammed.badal@su.edu.krd
Shaymaa M. Shakir
Statistical office Region
Assistant. Professor- Sallahaddin University-Statistics Department

In this paper, the wave filter was used on the water resource crisis in Iraq, represented by the Tigris and Euphrates rivers, which is one of the major crises and has an important economic impact since 1980, through the study and analysis of the Tigris and Euphrates rivers of water resources in Iraq for the years 1980-2021 and the application of the autocorrelation model VAR(P) in the multivariate time series, which is the intersection model between the variables. The model has been modified by wavelet analysis. The model gives a detailed and good analysis by testing the hypotheses that must be provided in the time series using the STATAv.17 program and for more than one variable and through statistical measures that give The lowest value of FPE, SBIC, HQIC, AIC and the maximum value of R^2. Forecasting the imports of the Euphrates and Tigris waters during the next five years from 2022-2026, we conclude that the values of the annual rate of water imports from the two rivers are decreasing after 2021.

Key words: VAR(P), FPE, SBIC, HQIC, R^2, AIC, Haar wavelet, Stationary, Wavelet transformation, Wavelet denoising, Daubechies Wavelet.

المستخلص

في هذا البحث استخدم مرشح المويجة على ازمة موارد المياه في العراق والمتمثلة بنهري دجلة والفرات ، وهي من الازمات الكبيرة والتي لها تاثيرها الاقتصادي المهم منذ عام 1980، من خلال دراسة وتحليل موارد هذين النهرين في العراق للاعوام 1980-2021 وتطبيق نموذج الارتباط الذاتي VAR(P) في متعدد المتغيرات للسلسلة الزمنية وهو نموذج التقاطع بين المتغيرات وتم تعديل النموذج بالتحليل المويجي ويعطي النموذج تحليل مفصل و جيد من خلال اختبار الفروض الواجب توفرها في السلسلة الزمنية باستخدام برنامج STATAv.17 و لاكثر من متغير وعن طريق المقاييس الاحصائية التي تعطي اقل قيمة FPE, SBIC, HQIC, AIC والقيمة القصوى لـ R^2 التنبؤ بالواردات لمياه دجلة والفرات خلال االخمسة سنوات القادمة من 2022-2026 ، نستنتج أن قيم المعدل السنوي للواردات المائية من النهرين في تناقص مستمر بعد عام 2021.

1. Introduction:

On the premise that succeeding values in the data file represent successive measurements conducted at uniformly spaced time intervals, time series analysis is predicated. Time series analysis has two basic objectives:

a) determining the phenomenon represented by the sequence of observations and

b) forecasting (predicting future values of the time series variable).

The pattern of the observed time series data must be recognized and more or less properly stated in order to achieve both of these objectives. Once the pattern has been identified, we can analyze it and combine it with further data (i.e., use it in our theory of the investigated phenomenon). Regardless of the level of our comprehension and the accuracy of our theory explaining the phenomenon, we can extrapolate the found pattern to forecast upcoming occurrences.

As in most other analyses, in time series analysis it is assumed that the data consist of a systematic pattern (usually a set of identifiable components) and random noise (error) which usually makes the pattern difficult to identify. Most time series analysis techniques involve some form of filtering out noise in order to make the pattern more salient.

The two primary types of components that make up the majority of time series patterns are trend and seasonality. The former depicts an all-encompassing systematic linear or (most frequently) nonlinear component that evolves through time and does not repeat, or at least does not repeat within the time frame covered by our data. The latter may be formally similar, but it repeats itself over time at regular intervals. In actual data, those two broad types of time series components might coexist.

In several forecast comparisons, large-scale econometric models were found to be outperformed by univariate time series models. Vector Auto Regressive (VAR) processes are a suitable model class for describing the Data Generation Process (DGP) of a small or moderate set of time series variables. The failure of the larger models can be attributed to the insufficient representation of the dynamic interactions in a system of variables. These models frequently regard all variables as a priori endogenous and allow for complex dynamics. If two or more variables share a stochastic trend, they are said to be cointegrated. The VAR form is not the most practical model configuration when cointegrating relations are included in a system of variables. In that situation, taking into account particular parameterizations that aid in the understanding of the cointegration structure is useful. Vector Error Correction Models (VECMs) or vector equilibrium correction models are the end models.

Because Iraq was and still is an agricultural nation, the water crisis there is one of the most significant challenges facing the Iraqi economy, particularly the agricultural sector. The scale of the water issue is significant, and if it does not surface now, the future will be stormy and dangerous. In fact, some academics and political analysts have dubbed the new millennium the third of the water cycle, as opposed to the era of oil that characterized the previous years.

Due to their unrealistic goals of food self-sufficiency, people in Syria and Iraq have a high population growth rate of 2.31% and 2.44%, respectively. As a result, the agricultural sector is currently the largest user of water. Significant issues with water quality have been brought about by the increase of agricultural operations, particularly in the Tigris and Euphrates basins. Take note that the Middle East's water challenges are intimately tied to the socio-economic stability at the national and regional levels. Five Middle Eastern and Southwest Asian nations make up the catchments of the Tigris and Euphrates Rivers. Turkey, Syria, Iran, Saudi Arabia, and Iraq are among them (Table 1). The two rivers' combined catchment area is 917103 km2, and according to ESCWA (2013), there were around 46 million people living there in 2013. Due to the fact that the two rivers originate in Turkey's southeast, Turkey has taken control of the two basins' riparian parties. (Al-Ansari et al. 2018).

In southwest Asia, the Tigris River is the second-longest river. It travels through Turkey for about 400 km, forms the border with Syria for about 47 km, and then travels into Iraq (ESCWA, 2013). Records of its flow along the Turkish-Syrian-Iraqi border show that it has an annual flow of 21 BCM (Billion Cubic Meter). Syria wants to take 4.5% of the water from the Tigris, whereas Turkey intends to use 14.1% of the water to which it supplies 52%. Finally, Iraq gives around 23% of its resources and wants to utilize 92.5% of the river's water.

Table 1: Tigris and Euphrates Basins

Countries	Tigris River		Euphrates River	
	Catchment area		Catchment area	
	Km^2	%	Km^2	%
Turkey	57719	12.2	125000	28.2
Syria	946	0.2	76000	17.1
Iraq	274400	58.0	177000	39.9
Iran	140038	29.6	-	-
Saudi Arabia	-	-	66000	14.9
Total	473103	100	444000	100

The Tigris River has five major tributaries in Iraq (Table 2). In total, the river irrigates 4 million hectares of agriculture (ESCWA, 2013). Iraq

began concentrating on land reclamation in the 1980s as a result of salinity and water logging issues. A total of 1.5 million ha of land were partially reclaimed, and about 1 million ha of agricultural land were reclaimed. Future plans are for the irrigation of 134,000 acres and the reclamation of 920,000 ha (ESCWA, 2013).

The Tigris River has five major tributaries in Iraq (Table 2). In total, the river irrigates 4 million hectares of agriculture (ESCWA, 2013). Iraq began concentrating on land reclamation in the 1980s as a result of salinity and water logging issues. A total of 1.5 million ha of land were partially reclaimed, and about 1 million ha of agricultural land were reclaimed. Future plans are for the irrigation of 134,000 acres and the reclamation of 920,000 ha (ESCWA, 2013).

Table 2: Main shared Tributaries of the River Tigris in Iraq.

Tributary	Mean Annual Flow (BCM)	Catchment Area (km^2)
Khabour	**2.00**	**6143 (Turkey 57%, Iraq 43%)**
Greater Zab	**12.70**	**26310 (Turkey 35%, Iraq 65%)**
Lesser Zab	**7.80**	**19780 (Iran 24%, Iraq 76%)**
Adhaim	**0.79**	**12965 (Iraq 100%)**
Diyala	**4.60**	**33240 (Iran 25%, Iraq75%)**
Tib	**1.00**	**Iran, Iraq**
Dwairij	**1.00**	**Iran, Iraq**

As agricultural activity in Turkey increased, more dams were built, and the climate changed, the Tigris' flow gradually began to decline. Due to a rise in salt and other pollutants, the water quality is degrading along with the decrease in flow. Long-term records show that the river's water quality is satisfactory within Turkey. Once the river enters Iraq, where a steady increase can be seen along the river, deterioration becomes alarming. The quality degrades to an unusable level for irrigation at Baghdad. The situation worsens downstream of Baghdad where the increase in salinity is more noticeable.

2. Material and method:

2.1 The model of vector autoregression (VAR): One of the most effective, adaptable, and simple methods for the study of multivariate time series is the vector autoregression (VAR) model. Dynamic multivariate time series are a logical extension of the univariate autoregressive model. The VAR model has shown to be particularly effective for forecasting and characterizing the dynamic behavior of economic and financial time series. It frequently offers forecasts that are

better than those from complex simultaneous equations models and univariate time series models. Because they can be made conditional on the likely future courses of certain model variables, forecasts from VAR models can be highly flexible. The VAR model is utilized not only for data description and forecasting but also for structural inference and policy research. The causal consequences of unexpected shocks or innovations to defined variables on the variables in the model are summarized as a result of particular assumptions about the causal structure of the data under consideration. Typically, forecast error variance decompositions and impulse response functions are used to summarize these causal effects.

Let $Y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})^T$ indicate a $(n \times 1)$ vector of time series variables. This is the stationary vector autoregression model. The fundamental p-lag vector autoregressive (VAR(p)) model has the formula

$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + \varepsilon_t, \quad t = 1,2, \ldots, T$ (1) where Π_i are $(n \times n)$ coefficient matrices and ε_t is a $(n \times 1)$ unobservable zero mean white noise vector process (serially uncorrelated or independent) with time invariant covariance matrix Σ. or A bivariate $VAR(2)$ model, for instance, has the following form:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \quad (2)$$

Or

$$y_{1t} = c_1 + \pi_{11}^1 y_{1t-1} + \pi_{12}^1 y_{2t-1} + \pi_{11}^2 y_{1t-2} + \pi_{12}^2 y_{2t-2} + \varepsilon_{1t}$$

$$y_{2t} = c_2 + \pi_{21}^1 y_{1t-1} + \pi_{22}^1 y_{2t-1} + \pi_{21}^2 y_{1t-2} + \pi_{22}^2 y_{2t-2} + \varepsilon_{2t}$$

Where $cov(\varepsilon_{1t}, \varepsilon_{2s}) = \begin{cases} \sigma_{12} & \text{if } t = s \\ 0 & o/w \end{cases}$.

The identical regressors, lagged values of y_{1t} and y_{2t}, are present in each equation. With lagged variables and deterministic terms serving as common regressors, the VAR(p) model is thus merely a seemingly unrelated regression (SUR) model.

The VAR(p) is represented by the lag operator notation as

$\Pi(L) Y_t = c + \varepsilon_t$ (3)

where $\Pi(L) = I_n - \Pi_1 L - \ldots - \Pi_p L^p$. The $VAR(p)$ is stable if the roots of $det(I_n - \Pi_1 z - \cdots - \Pi_p z^p) = 0$

lie outside the complex unit circle (have modulus greater than one), or, equivalently, if the eigenvalues of the companion matrix

$$F = \begin{pmatrix} \Pi_1 & \Pi_2 & \ldots & \Pi_{n-1} & \Pi_n \\ I_n & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & I_n & 0 \end{pmatrix} \quad (4)$$

have a modulus below one. A stable VAR(p) process is stationary and ergodic with time invariant means, variances, and autocovariances if the process was begun in the infinite past.
The unconditional mean is given by if Y_t is covariance stationary.

$$\mu = \left(I_n - \Pi_1 - \Pi_2 - \cdots - \Pi_p\right)^{-1} c \qquad (5)$$

The mean-adjusted form of the $VAR(p)$ is then

$$Y_t - \mu = \Pi_1(Y_{t-1} - \mu) + \Pi_2(Y_{t-2} - \mu) + \cdots + \Pi_p\left(Y_{t-p} - \mu\right) + \varepsilon_t \ldots (6)$$

The fundamental VAR(p) model might be too limited to accurately capture the key aspects of the data. To adequately represent the data, additional deterministic factors, such as a linear temporal trend or seasonal dummy variables, may be needed. Furthermore, stochastic exogenous factors might also be necessary. Given by is the general version of the VAR(p) model with exogenous variables and deterministic terms.

$$Y_t = \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + \Phi D_t + G X_t + \varepsilon_t \qquad (7)$$

where D_t is a $(l \times 1)$matrix of exogenous variables, X_t denotes a $(m \times 1)$ matrix of deterministic components, and Φ and G denote parameter matrices.

2.2. Wavelet: Wavelet theory can be used in a variety of fields. Harmonic analysis is related to all wavelet transforms since they are all types of time-frequency representation for continuous-time (analog) signals. Almost all discrete wavelet transforms that are actually useful employ discrete-time filter banks. In wavelet nomenclature, these filter banks are referred to as wavelet and scaling coefficients. Filters with finite impulse response (FIR) or infinite impulse response (IIR) may be found in these filter banks. The Fourier analysis's uncertainty principle and corresponding sampling theory apply to the wavelets that make up a continuous wavelet transform (CWT): One cannot simultaneously assign an exact time and frequency response scale to an event included in a signal. There is a lower bound on the product of the time and frequency response scale uncertainty. As a result, rather of merely marking a single point in the time-scale plane, such an event marks an entire region in the scaleogram of a continuous wavelet representation of this signal. Discrete wavelet bases may also be taken into account in relation to different variations of the uncertainty principle.
The three classes of wavelet transforms are continuous, discrete, and multiresolution-based.
2.2.1. *Continuous shift and scale parameters for wavelet transforms:*A given signal with finite energy is projected onto a continuous family of frequency bands (or analogous subspaces of the L^p function space $L^2(\mathbb{R})$) in continuous wavelet transforms. For any positive frequencies $f > 0$, the

signal, for example, may be represented on every frequency band with the form $[f, 2f]$. The original signal can then be recreated by appropriately integrating all of the output frequency components.

A subspace of scale 1 is scaled down to create the frequency bands or subspaces (sub-bands). In most circumstances, the shifts of one generating function ψ in $L^2(\mathbb{R})$, the *mother wavelet*—generate this subspace in turn. This function is given for the scale one frequency band [1,2].

$$\psi(t) = 2\,sinc(2t) - sinc(t) = \frac{sin(2\pi) - sin(\pi t)}{\pi t} \qquad (8)$$

with the (normalized) *sinc* function.

The functions (also referred to as *child wavelets*)

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \qquad (9)$$

where a is positive and specifies the scale and b is any real number and determines the shift are used to construct the subspace of scale an or frequency band $[1/a, 2/a]$. A point is defined by the pair (a, b) in the right halfplane $\mathbb{R}_+ \times \mathbb{R}$. The form

$$x_a(t) = \int_{\mathbb{R}} WT_\psi\{x\}(a,b).\psi_{a,b}(t)db \qquad (10)$$

is the projection of a function x into the subspace of scale a with wavelet coefficients

$$WT_\psi\{x\}(a,b) = \langle x, \psi_{a,b}\rangle = \int_{\mathbb{R}} x(t).\psi_{a,b}(t)dt. \qquad (11)$$

A scaleogram of the signal can be created using the wavelet coefficients for the analysis of the signal x.

2.2.2. *Discrete wavelet transforms with discrete shift and scale parameters:* One can wonder if it is enough to choose a discrete subset of the upper halfplane to be able to reconstruct a signal from the associated wavelet coefficients since it is computationally unfeasible to analyze a signal using all wavelet coefficients. The affine system for some real parameters $a > 1$, $b > 0$ is an example of such a system. All the points $(a^m, na^m b)$ with m, n in $\mathbb{Z}$ make up the matching discrete subset of the halfplane. The child wavelets are now specified as

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a^m}}\psi\left(\frac{t-nb}{a^m}\right) \qquad (12)$$

a prerequisite for any signal's reconstruction using the equation

$$x(t) = \sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} \langle x, \psi_{m,n}\rangle.\psi_{m,n}(t) \qquad (13)$$

is that an orthonormal basis of $L^2(\mathbb{R})$ is formed by the functions $\{\psi_{m,n}: m,n \in \mathbb{Z}\}$.

2.3. Mother wavelet: Continuously differentiable functions with compact support are preferred as the mother (prototype) wavelet for practical applications and efficiency considerations (functions). In contrast, one selects the wavelet functions from a subspace of the space $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$in order to satisfy analytical constraints (in the continuous WT) and generally for theoretical reasons (R) It is possible to square-integrate measurable functions in this space (in absolute value).

$$\int_{-\infty}^{\infty} |\psi(t)|dt < \infty \;\; and \;\; \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \qquad (14)$$

Being in this area makes it possible to establish the zero mean and square norm one conditions:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \quad \text{is the condition for zero mean, and}$$
$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1 \quad \text{is the condition for square norm one.} \qquad (15)$$

The mother wavelet must meet an admissibility criterion (roughly speaking, a kind of half-differentiability) in order to be a wavelet for the continuous wavelet transform ψ.

2.3.1. *Wavelet denoising:* Consider measuring the noisy signal $x = s + v$. Assume $v \sim N(0, \sigma^2 I)$ and that s has a sparse representation in a particular wavelet basis.

$$y = W^T x = W^T s + W^T v = p + z \qquad (16)$$

The majority of the elements in p are 0 or almost 0 and $z \sim N(0, \sigma^2 I)$. The estimation issue is the recovery of a signal in *iid* Gaussian noise since W is orthogonal. One approach is to use a Gaussian mixture model for p because p is sparse. Assume that there is a prior $p \sim aN(0, \sigma_1^2) + (1 - a)N(0, \sigma_2^2)$. The variance for "significant" coefficients is given by σ_1^2, and the variance for "insignificant" coefficients is given by σ_2^2.

The shrinkage factor, which depends on the prior variances σ_1^2 and σ_2^2, is then denoted by $\tilde{p} = E(p/y) = \tau(y)y, \tau(y)$. The shrinkage factor has the effect of setting tiny coefficients to 0 early while leaving large coefficients unaffected. Large coefficients contain the genuine signal while small coefficients are primarily made up of noise. In order to produce $\tilde{s} = W\tilde{p}$, use the inverse wavelet transform last.

2.3.2. *Haar Wavelet:* In mathematics, the Haar wavelet is a sequence of rescaled "square-shaped" functions which together form a wavelet family or basis. Wavelet analysis is similar to Fourier analysis in that it allows a

target function over an interval to be represented in terms of an orthonormal basis.

The Haar wavelet is also the simplest possible wavelet. The technical disadvantage of the Haar wavelet is that it is not continuous, and therefore not differentiable. This property can, however, be an advantage for the analysis of signals with sudden transitions, such as monitoring of tool failure in machines.

The Haar wavelet's mother wavelet function ψ(t) can be described as

$$(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2} \\ -1 & \frac{1}{2} \le t < 1 \\ 0 & otherwise \end{cases} \quad (17)$$

Its scaling function $\varphi(t)$ can be described as

$$\varphi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & otherwise \end{cases} \quad (18)$$

The Haar function $\psi_{n,k}$ is defined by the formula

$$\psi_{n,k}(t) = 2^{n/2}\psi(2^n t - k), \quad t \in \mathbb{R} \quad (19)$$

for each pair of integers *n*, *k* in Z.

The orthonormal basis in $L^2(\mathbb{R})$ of the Haar system on the line.

1. There are several noteworthy features of the Haar wavelet. Any continuous real function with compact support can be uniformly approximated by linear combinations of the shifted functions $\varphi(t), \varphi(2t), \varphi(4t), \dots, \varphi(2^n t), \dots$. This also applies to function spaces where each function can be roughly represented by continuous functions.
2. Any continuous real function on [0, 1] can be uniformly approximated by linear combinations of the constant function $1, \varphi(t), \varphi(2t), \varphi(4t), \dots, \varphi(2^n t), \dots$, the shifted functions and (n t).
3. The form's orthogonality

Haar Matrix: The 2×2 Haar matrix that is associated with the Haar wavelet is

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

The orthogonal and real Haar transform matrix is real. As a result, the following equations can be used to construct the inverse Haar transform.

$$H = H^*, \qquad H^{-1} = H^T, \qquad i.e. HH^T = I$$

where *I* is the identity matrix. For instance, n=4

$$H_4^T H_4 = \frac{1}{2}\begin{bmatrix} 1 & 1 & \sqrt{2} & 0 \\ 1 & 1 & -\sqrt{2} & 0 \\ 1 & -1 & 0 & \sqrt{2} \\ 1 & -1 & 0 & -\sqrt{2} \end{bmatrix} \cdot \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{2} & -\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Thus, the inverse Haar transform is

$$x_n = H^T y_n \qquad (20)$$

2.3. Daubechies wavelet: The Daubechies wavelets are often selected to have the maximum number A of vanishing moments for a given support width (this does not imply the optimum smoothness) (2A - 1). The two naming conventions in use are DN, which stands for the length or number of taps, and dbA, which stands for the quantity of vanishing moments. The wavelet transforms D4 and db2 are so identical.

The one whose scaling filter has extremal phase is chosen from the 2^{A-1} feasible solutions of the algebraic equations for the moment and orthogonality requirements. Using the quick wavelet transforms, the wavelet transform is also simple to use. Daubechies wavelets are frequently employed to solve a wide variety of issues, such as signal discontinuities, fractal issues, or the self-similarity characteristics of a signal.

In fact, the Daubechies wavelets cannot be expressed in closed form and are not described in terms of the resulting scaling and wavelet functions. The graphs below were created using the cascade algorithm, a numerical method that involves simply performing an appropriate number of inverse transformations on the values [1 0 0 0 0...].

Notably, the spectra displayed here are the amplitudes of the continuous Fourier transforms of the scaling and wavelet functions rather than the frequency response of the high and low pass filters.

Construction: In this case, the wavelet sequence (a band-pass filter) and the scaling sequence (a low-pass filter) will both be normalized to have a sum and sum of squares of two. They are sometimes normalized to have sum $\sqrt{2}$ such that any shifts of the sequences by an even number of coefficients are orthonormal to one another.

When writing the orthogonality condition using the general representation for a scaling sequence of an orthogonal discrete wavelet transform with approximation order A.

The Daubechies wavelet transform is implemented via a pair of linear filters. This filter pair must possess a characteristic known as a quadrature

mirror filter. The coefficient values for filter of order 4 can be determined by solving the coefficient of the linear filter c_i using the quadrature mirror filter property, which yields the solution shown below.

$$c_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \qquad c_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \qquad c_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad c_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

2.5. Fitting the model: Akaike Information Criteria (AIC) is then a means to choose the model that best balances these downsides, as under-fitting a model may not reflect the true nature of the variability in the outcome variable, while an over-fitted model lacks generality. AIC criteria asymptotically overestimates the order with positive probability, or classical null-hypothesis testing can be applied on the best model to ascertain the association between particular variables and the outcome of interest:

$$\mathrm{AIC} = 2\mathrm{K} - 2\log(\mathrm{L}(\hat{\theta}/\mathrm{y})) \qquad (21)$$

K is the number of estimable parameters (or degrees of freedom), and $\log(\mathrm{L}(\hat{\theta}/\mathrm{y}))$ is the estimated model's maximum log-likelihood. This estimate has been further updated to account for tiny data samples:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2\mathrm{K}(\mathrm{K}+1)}{\mathrm{n}-\mathrm{K}-1} \qquad (22)$$

If n is the sample size, K and AIC are as previously described, and This correction is negligible and AIC is sufficient if n is big relative to K. However, AICc is more broadly used and frequently used in place of AIC. The model with the lowest AICc (or AIC) score is thereafter the best model. The AIC and AICc ratings are ordinal and have no meaning on their own, it is vital to remember this.

A model is more likely to be the true model if the Bayesian information criterion (BIC) is lower, according to an estimate of the posterior probability of a model being true, given a certain Bayesian setup:

$$\mathrm{BIC} = 2\log \mathrm{n} - 2\log\left(\mathrm{L}\left(\frac{\hat{\theta}}{\mathrm{y}}\right)\right) \qquad (23)$$

A diagnostic technique for evaluating a time series model's lack of fit is the Box-Ljung test. For a series of vector auto-regressions of orders 1,..., max lag(p), Varsoc reports the final prediction error (FPE), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and Hannan and Quinn information criterion (HQIC) lag order selection statistics. BIC and HQ criteria estimate the order consistently under fairly general conditions if the true order p is less than or equal. lso presented is a series of likelihood-ratio test data for all entire VARs with orders lower than or equal to the highest lag order. The maximum lag and

estimation parameters in the post estimation version are dependent on the model that was just fitted or the model that was specified in estimates (estname), although the lag order for a vector error correction model can also be chosen using the pre-estimation version of varsoc (VECM). The lag-order selection statistics covered here can be applied when there are I(1) variables, as demonstrated by Nielsen (2001). To help researchers fit a VAR with the right order, numerous selection-order statistics have been established. The output of the trend stationary [TS] var contains several of these selection-order statistics. While retaining a common sample and option specification, the varsoc command computes these statistics over a range of lags p. It also performs a series of likelihood ratio (LR) checks and computes four information criteria. The FPE, AIC, HQIC, and SBIC are among the information criteria. The LR test contrasts a VAR with p lags with one with p-1 lags for a certain lag p.

To choose a lag order using this series of LR tests, the null hypothesis is that all the coefficients on the p-th lags of the endogenous variables are zero.

Hamilton (1994, 295-296) demonstrated that the log likelihood for a VAR(p) is:

$$LL = \frac{T}{2}\left[ln\left|\hat{\Sigma}^{-1}\right| - K.ln(2\pi - K)\right] \qquad (24)$$

where T is the number of observations, K is the number of equations, and $\hat{\Sigma}$ is the maximum likelihood estimate of $E(u_t u'_t)$. The log likelihood can be expressed as follows, where u_t is the K ×1 vector of disturbances:

$$LL(j) = 2[LL(j) - LL(j-1)] \qquad (25)$$

2.6. ***Model-order statistics:*** According to Lutkepohl (2005, 147), the formula for the FPE (final prediction error) is as follows:

$$FPE = |\Sigma u|\left(\frac{T + KP + 1}{T - KP - 1}\right)^{K}$$

by whom wrote

$$FPE = |\Sigma u|\left(\frac{T + \bar{m}}{T - \bar{m}}\right)^{K} \qquad (26)$$

where the average number of parameters across the K equations is represented by $\bar{m}$. This implementation takes collinearity-related variables out of the equation.

The constant term should be removed from the log likelihood, according to Lutkepohl (2005), as it has no bearing on inference. The Lutkepohl versions of the information requirements are:

$$AIC = ln|\Sigma u| + \frac{2PK^2}{T} \qquad (27)$$

$$SBIC = ln|\Sigma u| + \frac{ln(T)}{T}PK^2 \qquad (28)$$

$$HQIC = ln|\Sigma u| + \frac{2ln[ln(T)]}{T} PK^2 \qquad (29)$$

where PK^2 is the total number of parameters in the model and LL is the log likelihood.

2.7. The Stationarity*:* A time series with stationarity has characteristics that are independent of the observational time. The variables in y_t are covariance stationary if their initial two moments exist and are independent of time, which is a need for inference a y_t fter var and *svar*. A variable is covariance stationary, more specifically, if:

a- $E[y_t]$ is finite and independent of t, for all t.

b- For any t, $Cov[y_t, y_s]$ is a finite function of |t-s| but not of t or s by themselves.

c- For all t, $Var[y_t]$ is a finite number and is independent of t.

Only when all of the Eigen values for A_p are smaller than 1 in absolute value is a VAR(P) process considered stationary.

3. Results and discussion: Iraq's water shortage was the result of both internal and external factors. Climate change, global warming-related rains, the Tigris and Euphrates rivers coming from neighboring nations rather than Iraq, political unrest, and a lack of international law were the external causes (Lack of water dams and lakes on the surface rivers in Iraq, non-creeping vegetation from river banks, lack of sedimentation, lack or weakness of water supply, lack of modern irrigation methods, low water cost, population increase).

The VAR(p) model to forecast the annual rate of water imports of the Tigris and Euphrates rivers periods for the future years, using the VAR(2) model by equation, it takes the data of water imports in Iraq governorate for the period between 1980 to 20 21table 3. As a result, the predicted values of water imports from the Tigris and Euphrates rivers for next five years in Iraq (2022-2026).

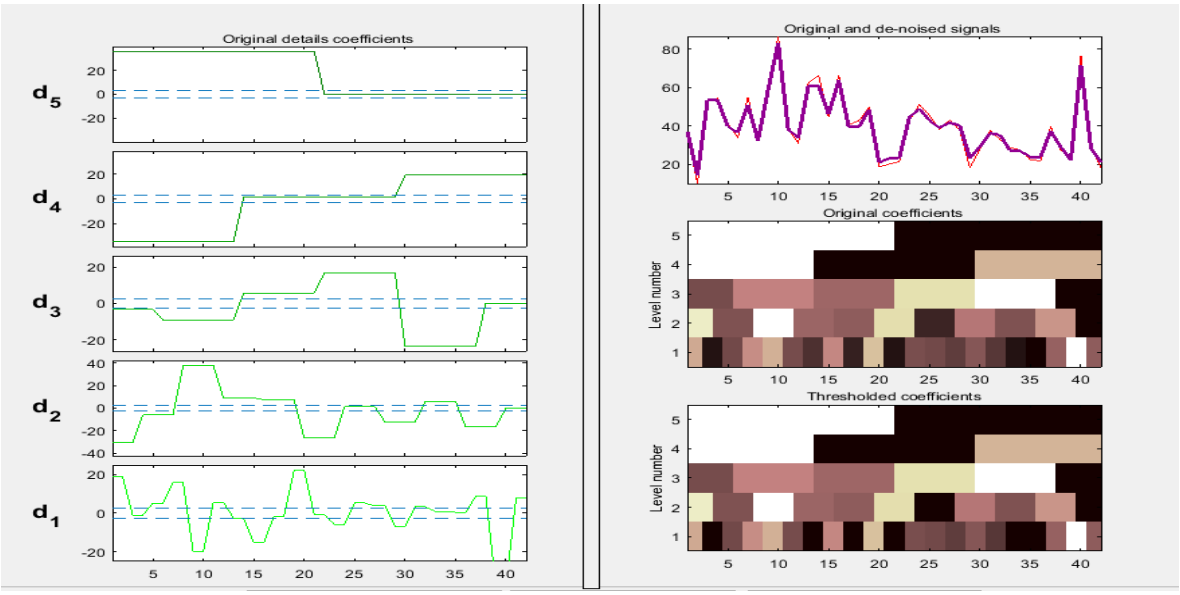
wavelet model: De-nosing Tigris Harr wavelet and De-nosing Tigris db(2) wavelet are used after utilizing discrete wavelet transforms from analyzing Matlab v.14 programming.

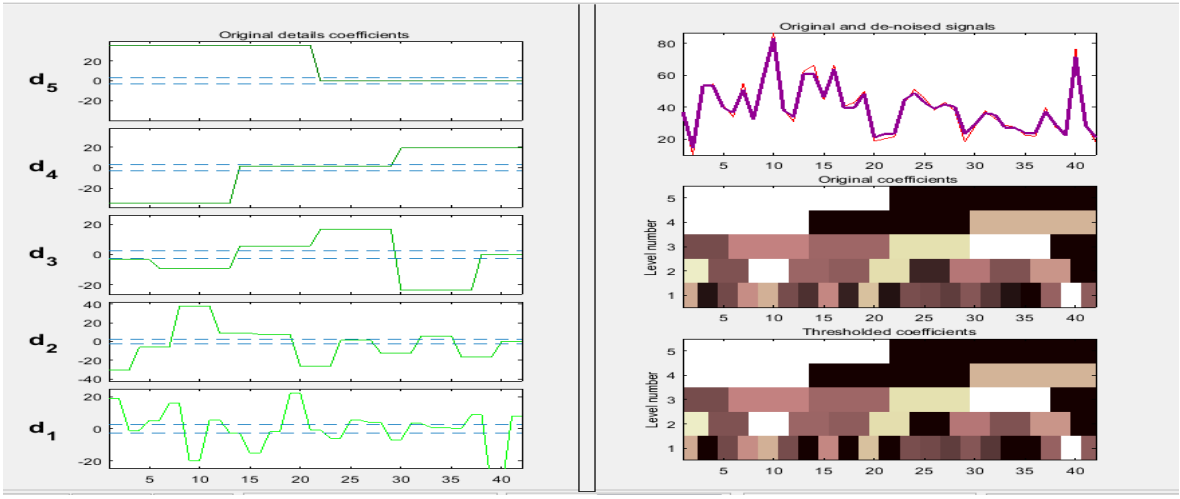
Table 3: Water imports in Iraq of the Tigris and Euphrates rivers for Years (1980-2021) billion cubic meters

Year	Euphrates	Tigris
1980	36.36	36.60
1981	29.80	9.99
1982	30.56	52.93
1983	31.43	54.40
1984	27.18	41.27
1985	37.22	34.00
1986	23.65	54.96
1987	17.22	32.60
1988	19.58	58.54
1989	46.73	86.66
1990	9.05	38.80
1991	12.40	30.87
1992	12.15	62.72
1993	12.37	66.36
1994	15.29	44.85
1995	23.90	66.40
1996	30.12	40.60
1997	27.64	42.75
1998	28.95	49.87
1999	18.61	18.60
2000	17.23	20.10
2001	9.59	21.28
2002	10.67	42.98
2003	15.71	51.13
2004	20.54	45.51
2005	17.57	38.07
2006	20.64	43.17
2007	19.33	37.76
2008	14.70	18.27
2009	13.57	27.97
2010	12.45	37.68
2011	14.62	32.94
2012	20.50	28.70
2013	17.85	27.46
2014	14.85	22.45
2015	13.54	21.80
2016	15.15	39.60
2017	13.16	27.37
2018	9.58	23.40
2019	16.95	76.56
2020	20.24	29.43
2021	13.00	18.24

Source: Ministry of Water Resources in Iraq

From analysis Matlab v.14 programing by De-nosing Tigris Harr wavelet and De-nosing Tigris db(2) wavelet applied De-nosing Haar wavelet *Tigris* & De-nosing db(2) wavelet *Tigris*





And applied De-nosing Haar wavelet Euphrates& De-nosing db(2) wavelet *Euphrates*

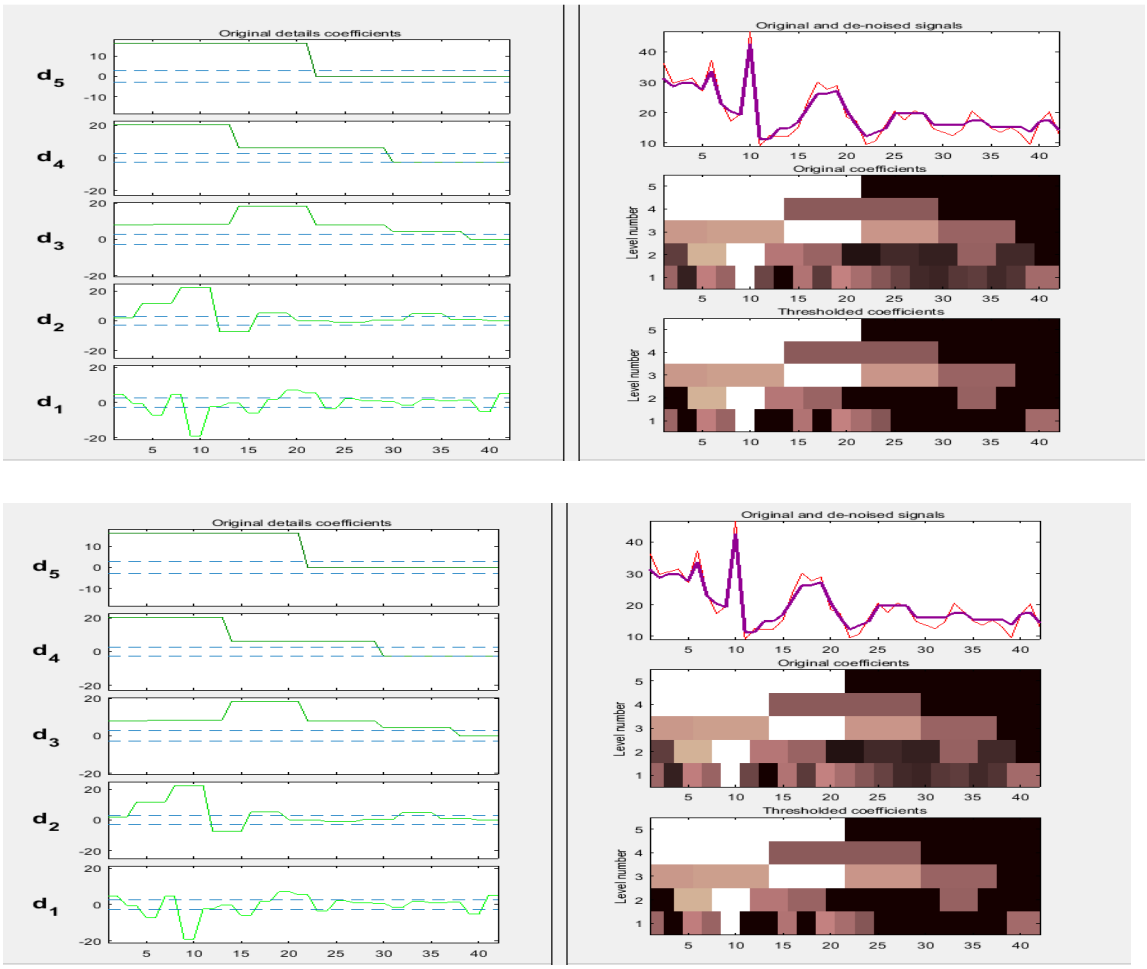


Table 4: Test of regression model (VAR) for variable (Tigris and Euphrates)

Equation	No. Parms.	RMSE	R^2	χ^2	$P > \chi^2$
Tigris	4	19.7732	0.8133	174.269	0.0000
Euphrates	4	7.8875	0.8699	267.505	0.0000

Table 5: Vector Auto regression (VAR) for variables haar wavelet (Tigris and Euphrates)

Years: 1982 – 2021	Number of obs. = 40
Log likelihood = -287.8013	AIC = 14.79006
FPE = 9096.566	HQIC = 14.91219
Det (Sigma_ml) = 5089.437	SBIC = 15.12784

From the table (5) the VAR model is the best because AIC (14.79006) value is less than HQIC, SBIC values.

Before forecasting we must check weather our VAR model has satisfied all assumption or not (Diagnostic checking for residually):
Null: there is no serial correlation (no autocorrelation)
Alt: there is serial correlation (autocorrelation)

Table 6: Lagrange-multiplier test at lag order

Lag	chi2	df	Prob. > chi2
1	12.9412	4	0.01157
2	25,3275	4	0.00004

Null: Residuals are normally distribution
Alt: Residuals are not normally distribution
We accepted assumption for Null because the residuals are normal in Lag(1) and Lag(2).

Table 7: Jarque-Bera test of VAR for normality

Equation	chi2	df	Prob. > chi2
Tigris	6.244	2	0.24058
Euphrates	6.407	2	0.46175
ALL	12.650	4	0.35520

From the table (7) the single-equation and overall Jarque–Bera statistics do not reject the null of normality, the null hypothesis is that the disturbance term in that equation has a univariate normal distribution, for all equations jointly, the null hypothesis is that the K disturbances come from a K-dimensional normal distribution.

Table 8: analysis the matrix var(p) and test of stationary time series

Rank	Parameters	LL	Eigen value	Trace statistic	Critical value 5%
0	6	-262.14105		23.2977	15.41
1	9	-254.52086	0.36125	8.0573	3.76
2	10	-250.49221	0.21099		

From the table (8) we see the series in model var(p) for Lags(2) it stationary because Eigen value less than one.
Now the forecasting for rivers Tigris and Euphrates rivers for years (2022-2026)

Table 9 Forecasting Water imports in Iraq of the Tigris and Euphrates rivers for years (2022-2026) billion cubic meters

Years	Forecasting Euphrates	Forecasting Tigris
2022	13.3759	23.8325
2023	11.9416	22.2784
2024	11.4039	22.2171
2025	10.7589	21.3063
2026	10.2494	20.5107

Table 10 Test of regression model (VAR) for variables haar wavelet (Tigris and Euphrates)

Equation	No. Parms.	RMSE	R^2	χ^2	$P > \chi^2$
Haar Tigris	4	17.3937	0.8500	226.729	0.0000
Haar Ephrates	4	6.2682	0.9156	433.7568	0.0000

Table 11 Test of Auto regression model (VAR) for variables haar wavelet (Tigris and Euphrates)

		Coef.	Std. Err.	z	$P > \|z\|$	[95%Conf.	Interval]
Haar Tigris	Haar Tigris L1.	.4985729	.1828621	2.73	0.006	.1401698	.856976
	L2.	.1127973	.1860474	0.61	0.544	-.251849	.4774436
	Haar Euphrates L1.	.1655264	.5016588	0.33	0.741	-.817707	1.14876
	L2.	.5218336	.4468163	1.17	0.243	-.353911	1.39758
Haar Euphrates	Haar Tigris L1.	.133567	.065898	2.03	0.043	-.004407	.262726
	L2.	-.029369	.0670467	-0.4	0.661	-.160779	.102039
	Haar Euphrates L1.	.339829	.180786	1.88	0.060	-.014505	.694161
	L2.	.403098	.161022	2.50	0.012	.087502	.718694

Table 12 Vector Auto regression (VAR) for variables db wavelet (Tigris and Euphrates)

Years: 1982 – 2021	Number of obs.= 40
Log likelihood = -287.8013	AIC = 14.79006
FPE = 9096.57	HQIC = 14.91219
Det (Sigma_ml) = 6089.438	SBIC = 15.12784

From the table (12) the VAR model is the best because AIC (14.79006) value is less than HQIC, SBIC values.

Table 13 Test of regression model (VAR) for variables db wavelet (Tigris and Euphrates)

Equation	No. Parms.	RMSE	R^2	χ^2	$P > \chi^2$
db Tigris	4	17.3937	0.8500	266.729	0.0000
db Ephrates	4	6.26826	0.9156	433.757	0.0000

Table 14 Test of Auto regression model (VAR) for variables db wavelet (Tigris and Euphrates)

		Coef.	Std. Err.	z	$P > \|z\|$	[95%Conf.	Interval]
db Tigris	db Tigris						
	L1.	.498573	.182863	2.73	0.006	.140169	.856976
	L2.	.112798	.186047	0.61	0.544	-.251849	.477444
	db Euphrates						
	L1.	.165526	.501659	0.33	0.741	-.817707	1.14876
	L2.	.521834	.446816	0.17	0.243	-.353911	1.39758
db Euphrates	db Tigris	.133567					
	L1.	-	.065899	2.03	0.043	.004407	.262726
	L2.	.029369	.067047	-0.4	0.661	-.160779	.102039
	db Euphrates						
	L1.	.339829	.180786	1.88	0.060	-.014504	.694161
	L2.	.403098	.161022	2.50	0.012	.087502	.718694

4. Conclusions:

1- We come to the conclusion that using the vector autoregression VAR(p) model, which is one of the most effective, flexible, and user-friendly models for the analysis of multivariate time series. By comparing the expected annual rate of water imports for the Tigris and Euphrates rivers during the next five years with their previous annual rates, we conclude that the values of the annual rate of water imports from the two rivers are decreasing after 2021.

2- The best statistic by AIC (14.79006) is provided by the model VAR(p), which is less than HQIC and SBIC.

3- From diagnostic testing of all underlying assumptions to screening for residuals, the optimum VAR(p).

4- The value of coefficient of determination(R^2) increased after using wavelet filter with the model VAR(p).

5- Based on the foregoing, we conclude that the answer must come from within Iraq and involve the building of dams and water storage facilities rather than letting the water flow out to sea

6- The project aims to encourage farmers to use modern irrigation methods in Iraqi agriculture and to disseminate experiments that proved the role of modern irrigation methods in reducing the consumption of water, in addition to the Cree rivers and continuously to prevent the growth of plants that consume water by the water and lining the drains in order to reduce the waste water resulting from the project.

References:

1. Akaike, H. (1973), "*Information theory and an extension of the maximum likelihood principle*." In Second International Symposium on Information Theory, ed. B. N. Petrov and F. Csaki, 267–281. Budapest: Akailseoniai–Kiudo.

2. Al-Ansari, N.; AlJawad, S.; Adamo, N.; Sissakian, V.K. & Laue, J. (2018) "*Water Quality within the Tigris and Euphrates Catchments"* Journal of Earth Sciences and Geotechnical Engineering, vol. 8, no. 3, 2018, 95-121 ISSN: 1792-9040 (print version), 1792-9660 (online) Scienpress Ltd.

3. Amemiya, T. (1985), "*Advanced Econometrics*". Cambridge, MA: Harvard University Press.

4. Badal, M. A. (2018) "*Using Multivariate Time Series Model VAR(P) to forecast Water Supply of Tigris and Euphrates Rivers in Iraq*" vol.22, no. 6, P.P. 220-228.

5. *ESCWA annual report 2013: 40 years with the Arab world ESCWA*, (2014); E/ESCWA/OES/2014/1

6. Franses, Ph. H. and Dijk, D.V. (2000), "*Nonlinear Time Series Models in Empirical Finance*" Cambridge University Press, ISBN 0 511 01100 8 virtual.

7. Gikungu, S.W., Waititu, A.G. (2015),"*Forecasting Inflation Rate in Kenya Using SARIMA Model"* American Journal of the oretical and Applied Statistics 4,15-18.

8. Hamilton, J. D. (1994), *"Time Series Analysis"*. Princeton: Princeton University Press.

9. Kaya, I. (1998) "*The Euphrates-Tigris basin: An overview and opportunities for cooperation under international law*" Office of Arid Lands Studies, College of Agriculture and Life Sciences, The university of Arizona; No. 44, Fall/Winter 1998; ISSN: 0277-9455 | E-ISSN: 1092-5481,Conflict Resolution and Transboundary Water Resources.

10. Luo, C.S., Zhou, L., Qingfeng, W. (2013)," *Application of SARIMA Model in Cucumber Price Forecast*", Applied Mechanics and Materials, Vols.373-375, pp.1686-1690.
11. Lutkepohl, H. (1993), "*Introduction to Multiple Time Series Analysis"*. 2nd ed. New York: Springer.
12. Lutkepohl, H. (2005), "*New Introduction to Multiple Time Series Analysis"*. New York: Springer.
13. Michal, S. (2001),"*Time Series Analysis* "log linear publishing, Canada.
14. Mira, S.K., Ahmad M.R. (2015),"*Time Series Models for Average Monthly Solar radiation in Malaysia*", Research and Education in Mathematics, International Conference Kuala Lumpur, Malaysia.
15. Nielsen, B. (2001), *"Order determination in general vector autoregressions".* Working paper, Department of Economics, University of Oxford and Nuffield College.
16. Paulsen, J. (1984), "*Order determination of multivariate autoregressive time series with unit roots"*. Journal of Time Series Analysis 5: 115–127. 8-Tsay, R. S. (1984), Order selection in nonstationary autoregressive models. Annals of Statistics 12: 1425–1433.

A Statistical Study of Dermatological Diseases for β-Thalassemia Major Patients using ANCOVA

Mardin Sameer Ali (mardin.ali@pti.edu.krd)
Dr. Delshad Shaker Ismael Botani (delshd.botani@su.edu.krd)

Abstract

β-thalassemia has two types called major and minor, β-thalassemia major is more dangerous than minor and it leads to blood disorder that reduces the production of hemoglobin. Hemoglobin is an iron-containing protein found in red blood cells and is responsible for transporting Oxygen to all of the body's cells. The objective of this study is to evaluate and compare the effects of gender and living place of patients (Inside city, outside city) on degree of dermatological diseases for patients who have β-thalassemia major, and taking into consideration that the duration of blood transfusion (in years) was studied as a covariate for degree of dermatological diseases. Two-way ANCOVA was used to analysis the data by applying SPSS v.25 and MS Excel. The most important findings in this study were the gender and place of living of β-thalassemia major patients did not have significant effects on dermatological diseases (place of living has slight effects on it), at the same time duration of blood transfusion have a significant effects on dermal diseases and should be taken into consideration by dermatologists and hematologists. Also, it is recommended to have more attention to patients living outside the cities, because there may be lack of medical centers related to thalassemia.

Keywords: Dermatological Diseases, ANCOVA, β-thalassemia major.

دراسة إحصائية للأمراض الجلدية لمرضى الثلاسيميا بيتا الكبرى باستخدام ANCOVA

د. دلشاد شاكر اسماعيل بوتاني (delshd.botani@su.edu.krd)[1]
ماردين سمير علي (mardin.ali@pti.edu.krd)[2]

الملخص

الثلاسيميا بيتا لها نوعان يسمى الثلاسيميا الكبرى والثانوية، الثلاسيميا بيتا الكبرى أخطر من الثلاسيميا الثانوية وتؤدي إلى اضطراب الدم الذي يقلل من إنتاج الهيموجلوبين. الهيموجلوبين هو بروتين يحتوي على الحديد ويوجد في خلايا الدم الحمراء وهو مسؤول عن نقل الأكسجين إلى جميع خلايا الجسم. الهدف من هذه الدراسة هو تقييم ومقارنة تأثيرات الجنس والمكان المعيشي للمرضى (داخل المدينة ، خارج المدينة) على درجة الأمراض الجلدية للمرضى المصابين بالثلاسيميا بيتا الكبرى، مع الأخذ بنظر الاعتبار أن مدة نقل الدم (بالسنوات) تمت دراستها

1 Statistics & Informatics Department, Adm. & Eco. College, Salahaddin-Erbil University, Erbil, Iraq.
2 Accounting Dep., Paitaxt Technical Institute, Erbil, Iraq.

كمتغير مشترك لدرجة الأمراض الجلدية. تم استخدام تحليل التباين المسترك (ANCOVA) ثنائي الاتجاه لتحليل البيانات من خلال تطبيق SPSS v.26 و MS Excel. كانت أهم النتائج في هذه الدراسة أن الجنس ومكان معيشة مرضى الثلاسيميا بيتا الكبرى لم يكن لهما تأثيرات معنوية على الأمراض الجلدية (مكان المعيشة كانت له تأثيرات طفيفة)، وفي نفس الوقت كان لمدّة نقل الدم تأثيرات معنوية على الأمراض الجلدية ويجب أن تؤخذ في الاعتبار من قبل أطباء الامراض الجلدية وأمراض الدم. كما يوصى بالمزيد من الاهتمام بالمرضى الذين يعيشون خارج المدن، لأنه قد يكون هنالك نقص في المراكز الطبية المتعلقة بمرض الثلاسيميا.

الكلمات الدالة: الأمراض الجلدية ، تحليل التباين المشترك (ANCOVA) ، الثلاسيميا بيتا الكبرى.

1. Introduction

Thalassemia is a hemoglobin composition disorder marked by a lack of or reduced compound of globin chains. It is most commonly found in malarial, tropical and sub-tropical regions of Mediterranean countries, the Middle East, Transcaucasia, Central Asia, the Indian Subcontinent (South Asia), and Southeast Asia. α and β-thalassemia were the most common kind of thalassemia [1].

Thalassemia (Hereditary Anemia) is a very common genetic disorders in the world where around 4.83% of the people in the world carry globin heterozygous for α-thalassemia and β-thalassemia [2]. It caused by inheritance of an affected allele from father and mother, and the fundamental abnormality in thalassemia is impaired production of the globin chain, so thalassemia is named by reference to the affected globin chain: α-thalassemia involves the α-chain, β-thalassemia the β-chain [3].

β-thalassemia is a blood condition that causes a decrease in hemoglobin production. Hemoglobin is an iron-containing protein in red blood cells that transports oxygen to all of the body's cells. Low hemoglobin levels cause an oxygen shortage in many regions of the body in patients with β-thalassemia. A lack of red blood cells (anemia) affects those who are affected, resulting in pale skin, weakness, weariness, and other major problems. People with β-thalassemia are more likely to have irregular blood clots [4].

β-thalassemia is divided into two categories based on the severity of symptoms: β-thalassemia major (also known as Cooley's anemia) and β-thalassemia minor. People with significant β-thalassemia will require regular blood transfusions and may not live a normal life. Pale skin, yellowish skin (jaundice), poor appetite, infections, belly (abdominal) enlargement, and other symptoms are common [5] [6]. There appears to be a strong link between dermatological illnesses and -thalassemia major [2] [7]. As a result, the authors try to find a relationship and effects of some factors on dermatological diseases for β-thalassemia major patients.

The great importance of this study is to any Private or governmental institutions who have direct or indirect relationship with Hereditary Anemia such as hematological hospitals, University of medicine, private centers and hospitals of dermatology and hematology, etc.

The body of this study is divided into 5 sections, first section is devoted for introduction and basic concepts, second section is related to objective of the study, methods and materials of the study is the address of third section, fourth section is concerning the results which consists of an overview of descriptive statistics and Analysis of Covariance (ANCOVA), and the last section is devoted for conclusions and recommendations. Each section may be divided into sub-sections depending on the subjects and requirements of the study.

2. Study Objectives

The main objective of this study is to evaluate and compare the effects of two factors: gender and living place of patients (Inside city, outside city) on degree of dermatological diseases for patients who have β-thalassemia major, and taking into consideration that the duration of blood transfusion (in years) was studied as a concomitant (covariate) for degree of dermatological diseases. The appropriate experiment for this situation is called Two-Way ANCOVA and had been used for data analyses.

Depending on the above objectives, the authors stated 4 main scientific hypotheses concerning this study and they are as follows:

First:

H_0: There is no significant difference in degree of dermatological diseases among males and females of β-thalassemia major patients.

H_1: There is significant difference in degree of dermatological diseases among males and females of β-thalassemia major patients.

Second:

H_0: There is no significant difference in degree of dermatological diseases among of β-thalassemia major patients who are living inside and outside cities.

H_1: There is significant difference in degree of dermatological diseases among of β-thalassemia major patients who are living inside and outside cities.

Third:

H_0: There is no significant interactions in degree of dermatological diseases between gender and place of living (address) of β-thalassemia major patients.

H_1: There is significant interactions in degree of dermatological diseases between gender and place of living (address) of β-thalassemia major patients.

Fourth:
H_0: There is no significant effects of duration of blood transfusion on degree of dermatological diseases for β-thalassemia major patients.
H_1: There is significant effects of duration of blood transfusion on degree of dermatological diseases for β-thalassemia major patients.

3. Methods and Materials

3.1. Data Collection

The data that used in this study is belonging to an MSc student from Medicine College – Hawler Medical University, after the owner's permission and completion of the MSc study in 2015. The data is concerning 176 β-thalassemia major patients who were visiting Erbil Thalassemia Center for a period starting from April, 2013 until January, 2014. The collected data is related to patients with ages starting from new born babies until 34 years old, the studied variables are consisting of gender, address (living inside or outside cities), degree of dermal symptoms (response variable), and duration of blood transfusion (covariate variable). The data entering and analyzing were done using SPSS V.23 and Microsoft Excel 2013.

3.2. Methodology

Since this study trying to understand the dermatological diseases effects for β-thalassemia major patients, solving specific practical medical questions related to dermatological diseases, and data had been collected, the applied research is being taken into consideration. The ANCOA has been used to analyze the dermal diseases related to β-thalassemia major patients because there are many cases in KRG with hereditary anemia.

3.2.1 Analysis of Covariance (ANCOVA)

In studying ANOVA, an analyst applies several treatments or treatment combinations to randomly selected experimental units and then wants to compare the treatment means for response (dependent) variable. In ANOVA, linear models are used to facilitate a comparison of these means. Then the model is consisting of one continuous response variable with one categorical independent variable or more. In ANCOVA, the situation is the same except that there are one categorical independent variable or more with one continuous independent variable or more. These extra continuous independent variables may affect the response variable, these are then known as covariates or concomitant variables. Analysis of covariance is sometimes described as a **blend of ANOVA and regression** [8] [9].

Before conducting any statistical analyses especially ANCOVA, one must know under which conditions should be implemented this analysis. Therefore, below are some important assumptions required for conducting this analysis [8] [9] [10]:

1) The response (dependent) variable and covariates should be measured at the continuous level (scale variable).
2) The independent (explanatory) variables should contain categorical variables and continuous variables (covariates).
3) The response variable is linearly related to the covariate. If this assumption holds, part of the error in the model is predictable and can be removed to reduce the error variance.
4) The slopes of the groups (treatments) are the same. This assumption is called the assumption of homogeneity of regression slopes. If we computed and drew the regression line for each of these scatterplots, they should look more or less the same if the assumption is met (i.e., the values of β in each group or treatment should be equal).
5) Independence of the covariate and treatment effect. The covariate shares its variance only with the unexplained variations (i.e. there is no interaction between the covariate and the independent variable).
6) The dependent variable (or residuals) should be approximately normally distributed for each treatment.
7) Homogeneity of error variances for each treatment, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

3.2.2 Two-Way ANCOVA

The linear statistical model of two-way ANCOVA is as follows [11] [12]:

$$\boldsymbol{y_{ijk} = \mu + \alpha_i + \gamma_j + \delta_{ij} + \beta x_{ijk} + \epsilon_{ijk}} \ldots (\mathbf{1})$$

Where: $i = 1,2,\ldots,a \quad ; \quad j = 1,2,\ldots,b \qquad k = 1,2,\ldots,n$

μ = the overall mean score.

α_i = the true effect of the i^{th} level of factor *A*.

γ_j = the true effect of the j^{th} level of factor *B*.

δ_{ij}= the effect of the interaction between α_i and δ_j.

x_{ijk}= the covariate measured on the same experimental unit as y_{ijk}.

ϵ_{ijk} = random errors component and approximately distributed $NIID(0, \sigma^2)$.

The above model can be written in matrix form as follows:

$$\boldsymbol{Y = Z\alpha + X\beta + \varepsilon} \ldots (\mathbf{2})$$

Where Z contains 0s and 1s, α contains μ and parameters such as α_i, γ_j, and δ_{ij} representing factors and interactions; X contains the covariate values; and β contains coefficients of the covariates. If there are 2 levels

for each factor with one covariate, then equation 2 can be rewrite as follows:

$$\begin{pmatrix} y_{111} \\ y_{112} \\ \vdots \\ y_{11n} \\ y_{121} \\ y_{122} \\ \vdots \\ y_{12n} \\ y_{211} \\ y_{212} \\ \vdots \\ y_{21n} \\ y_{221} \\ \vdots \\ y_{22n} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \\ \delta_{11} \\ \delta_{12} \\ \delta_{21} \\ \delta_{22} \end{pmatrix} + \begin{pmatrix} x_{111} \\ x_{112} \\ \vdots \\ x_{11n} \\ x_{121} \\ x_{122} \\ \vdots \\ x_{12n} \\ x_{211} \\ x_{212} \\ \vdots \\ x_{21n} \\ x_{221} \\ \vdots \\ x_{22n} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \vdots \\ \epsilon_{11n} \\ \epsilon_{121} \\ \epsilon_{122} \\ \vdots \\ \epsilon_{12n} \\ \epsilon_{211} \\ \epsilon_{212} \\ \vdots \\ \epsilon_{21n} \\ \epsilon_{221} \\ \vdots \\ \epsilon_{22n} \end{pmatrix} \quad \dots (3)$$

The estimation of the parameters are as follows [9]:

$$\widehat{\boldsymbol{\beta}} = [\acute{\boldsymbol{X}}(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}\acute{\boldsymbol{Z}})^{-1}\acute{\boldsymbol{Z}})\boldsymbol{X}]^{-1}\acute{\boldsymbol{X}}(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}\acute{\boldsymbol{Z}})^{-1}\acute{\boldsymbol{Z}})\boldsymbol{Y} \dots (\mathbf{4})$$

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Z}\acute{\boldsymbol{Z}})^{-1}\acute{\boldsymbol{Z}}\boldsymbol{Y} - (\boldsymbol{Z}\acute{\boldsymbol{Z}})^{-1}\boldsymbol{Z}\acute{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} \dots (\mathbf{5})$$

If both factors and their interactions have fixed effects, then the *ANCOVA* table for model in equation 1 will be as follows (table 1):

Table 1: Two-Way ANOVA Table (Factorial CRD)

S.O.V.	d.f.	S.S.	M.S.	F	η^2
Treatments	ab-1	SS_{tr}	MS_{tr}		
A	a-1	SS_A	MS_A	$\frac{MS_A}{MS_E}$	$\frac{SS_A}{SS_T}$
B	b-1	SS_B	MS_B	$\frac{MS_B}{MS_E}$	$\frac{SS_B}{SS_T}$
AB	(a-1) (b-1)	SS_{AB}	MS_{AB}	$\frac{MS_{AB}}{MS_E}$	$\frac{SS_{AB}}{SS_T}$
Covariate	1	SS_{cov}	MS_{cov}	$\frac{MS_{cov}}{MS_E}$	$\frac{SS_{cov}}{SS_T}$
Error	ab(n-1)	SS_E	MS_E		
Total	abn-1	SS_T			

Where:

$$SS_{tr} = \sum_{ij} \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{abn} + \frac{[\sum_{ijk}(x_{ijk}-\bar{x}_{ij.})(y_{ijk}-\bar{y}_{ij.})]^2}{\sum_{ijk}(x_{ijk}-\bar{x}_{ij.})^2} - \frac{[\sum_{ijk}(x_{ijk}-\bar{x}_{...})(y_{ijk}-\bar{y}_{...})]^2}{\sum_{ijk}(x_{ijk}-\bar{x}_{...})^2} \quad ...(6)$$

$$SS_{error} = \sum_{ijk} y_{ijk}^2 - \frac{y_{ij.}^2}{n} - \frac{[\sum_{ijk}(x_{ijk}-\bar{x}_{ij.})(y_{ijk}-\bar{y}_{ij.})]^2}{\sum_{ijk}(x_{ijk}-\bar{x}_{ij.})^2} \quad ...(7)$$

$$SS_A = bn \sum_{i=1}^{a} (\bar{x}_{i..} - \bar{x}_{...})(\bar{y}_{i..} - \bar{y}_{...}) \quad ...(8)$$

$$SS_B = an \sum_{j=1}^{b} (\bar{x}_{.j.} - \bar{x}_{...})(\bar{y}_{.j.} - \bar{y}_{...}) \quad ...(9)$$

$$SS_{AB} = n \sum_{ij} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) \quad ...(10)$$

$$SS_x = bn \sum_{ijk} (x_{ijk} - \bar{x}_{...})^2 \quad ...(11)$$

$$SS_T = SS_{tr} - SS_x - SS_{error} \quad ...(12)$$

$$Eta\ Squared = \eta^2 = \frac{SS_{effect}}{SS_{Total}} \quad ...(13)$$

In the next section, all the above equations and ANCOVA table will be used to understand the nature of the data and how they are correlated. The assumptions of the two-way ANCOVA are going to be checked and then the raw data will be analyzed.

4. Results

This section will be divided into 2 main sub-sections related to data analysis for the comparison of the effects of some categorical and continuous variables on the degree of dermatological diseases. The

obtained data were analyzed through SPSS V.23 and Microsoft Excel 2013. Depending on the hypotheses that were referenced in the third section, two-way ANCOVA methods had been used in this section to know these effects.

4.1. Descriptive Statistics

Figure 1 shows that 28% of patients are females living outside city and 18% – females living inside city, 31% – males living outside city, 23% – males living inside city, i.e. that percentage of males who live outside the city is highest and females who live inside city is lowest.

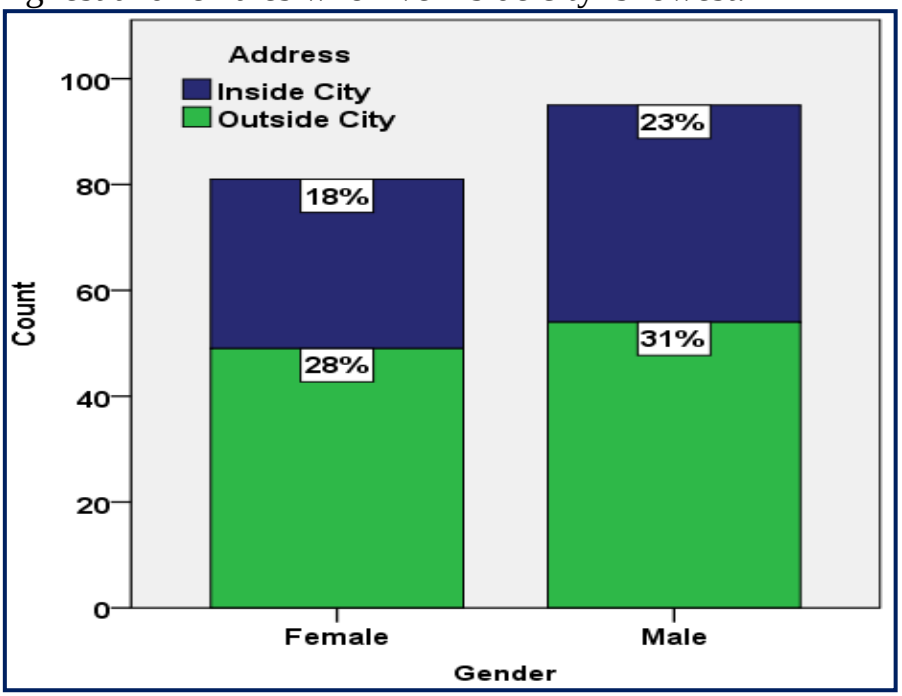


Figure 1: Gender and living place of β-thalassemia major patients

An important descriptive statistics are presented in table 2 which consist of number of patients, mean, standard deviation, and Standard error of mean according to gender and address (living place). The degree of dermatological diseases is approximately the same for males and females and for patients who live inside or outside cities (they have almost the same means). Concerning the duration of blood transfusion, there is a difference between patients who live inside and outside the cities, but there are approximately the same means for males and females.

Table 2: Descriptive statistics for degree of dermatological diseases and duration of blood transfusion according to gender and living place (address)

Variables		Duration of Blood Transfusion (Years)				Dermatological Diseases Level			
		N	$\bar{x}$	SD	SE	N	$\bar{x}$	SD	SE
Address	Inside City	73	10.0	6.2	.7	73	4.2	1.8	.2
	Outside City	103	9.1	5.7	.6	103	4.4	1.6	.2
	Total	176	9.4	5.9	.4	176	4.3	1.7	.1
Gender	Female	81	9.5	5.9	.7	81	4.2	1.8	.2
	Male	95	9.4	5.9	.6	95	4.4	1.6	.2
	Total	176	9.4	5.9	.4	176	4.3	1.7	.1

Table 3 is a very important table which illustrates the above table in another way using an interaction between gender and duration of blood transfusion. The results are differ with table 2 because there are differences between patients who live inside cities and outside cities for male and female, but there are little differences between patients who live inside cities and outside cities according to male and female for the degree of dermatological diseases.

Table 3: Descriptive statistics for degree of dermatological diseases and duration of blood transfusion according to combination between gender and living place

Variables		Duration of Blood Transfusion (Years)				Degree of Dermatological Diseases			
		N	$\bar{x}$	SD	SE	N	$\bar{x}$	SD	SE
Female	Inside City	32	9.9	6.7	1.2	32	3.9	2.0	.4
	Outside City	49	9.2	5.3	.8	49	4.4	1.6	.2
	Total	81	9.5	5.9	.7	81	4.2	1.8	.2
Male	Inside City	41	10.0	5.8	.9	41	4.4	1.7	.3
	Outside City	54	8.9	6.0	.8	54	4.5	1.6	.2
	Total	95	9.4	5.9	.6	95	4.4	1.6	.2

It is interesting to have a look at the histogram and Q-Q plot of dependent variable without any restriction. Figure 2 show these two figures and it appears that the response variable distributed normal.

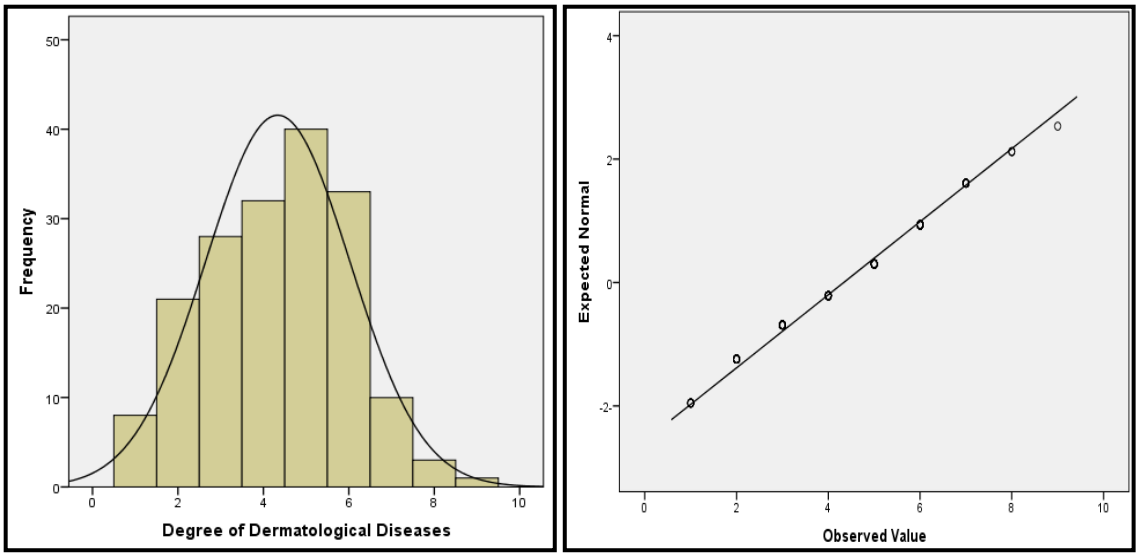


Figure 2: Histogram and Q-Q plot of degree of dermatological diseases (Response variable)

4.2. Data Analysis (Two-Way ANCOVA)

This section is the most critical section in whole study because its results will lead to have conclusions about the study hypotheses. Therefore, right results of this section will be a basis for right conclusions and recommendations. ANCOVA is a statistical method that can be easily used for studying phenomena that have quantitative outcome and two (or more) categorical independent variables with one (or more) continuous independent variable. As mentioned in the previous sections that ANCOVA method have some assumptions that must be tested and checked. The familiar assumptions are equality of error variances and normality of dependent variable.

To test the equality of error Variances, Levene's Test was used and it shows significant results ($F_{(3,172)}$ = 1.72, P-value = 0.165), i.e. non-violation of an assumption. However, in this case the non-significant result suggests the variance is equal across the various combinations of gender and place of living.

Concerning the normality assumption of the residuals of response variable, table 4 shows that errors are distributed normally depending on Shapiro-Wilk test or Kolmogorov-Smirnov test. It is worth to say that there is no relationship between covariate variable and the other two categorical variables.

Table 4: Test of normality for residuals of response variable

Variable	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Residuals of response variable	.048	176	.200*	.993	176	.555

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Concerning the relationships between response variable and covariate, there is a significant correlation between them ($r = 0.49, p - value = 0.00$). Also, there is no significant interaction between covariate variable and the two factors.

All the assumptions of two-way ANCOVA are satisfied, so it can be producing table 1 using equations 6 to 13 in order to construct equation 1 which is related to linear statistical model of this kind of experiment. Table 5 describe the two-way ANCOVA table with all calculations of finding the effects of first factor, second factor, interaction, and the covariate variable.

Table 5: Two-Way ANCOVA table for gender (Factor A) and living place of patients (Factor B) effects on the degree of dermatological diseases

S.O.V.	d.f.	S.S.	M.S.	F	P-Value	Eta Squared (η^2)
A	**1**	**3.483**	**3.483**	**1.611**	**.206**	**0.007**
B	**1**	**7.468**	**7.468**	**3.453**	**.065**	**0.015**
AB	**1**	**.720**	**.720**	**.333**	**.565**	**0.001**
Covariate	**1**	**122.742**	**122.742**	**56.761**	**.000**	**0.250**
Error	**171**	**369.778**	**2.162**			
Total	**175**	**499.222**				

The above table shows that there are no statistically significant effects of the two factors, also, there is no statistically significant effects of the interaction between the levels of two factors because P-Value < 0.05. Figure 3 clearly describe the interaction between the gender and living place of patients for increasing and decreasing of dermatological diseases degree, it seems that there is a weak effects of the two factors on the dermatological diseases.

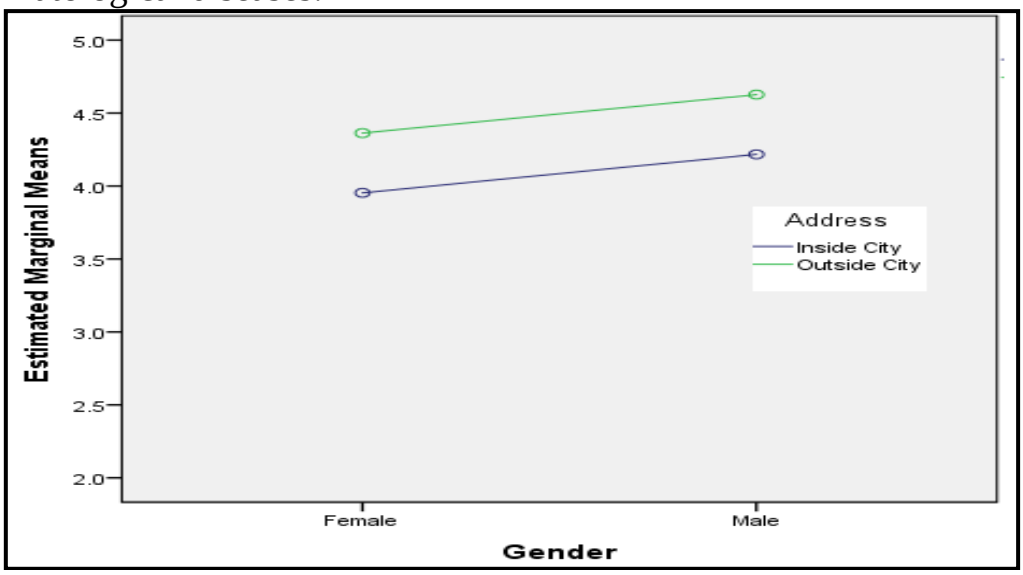


Figure 3: Interaction effects of gender and living place (Address)

As a general, β-thalassemia major patients who live outside cities are more exposed than patients inside cities. The duration of blood transfusion has a great effect on the response variable as shown in figure 4.

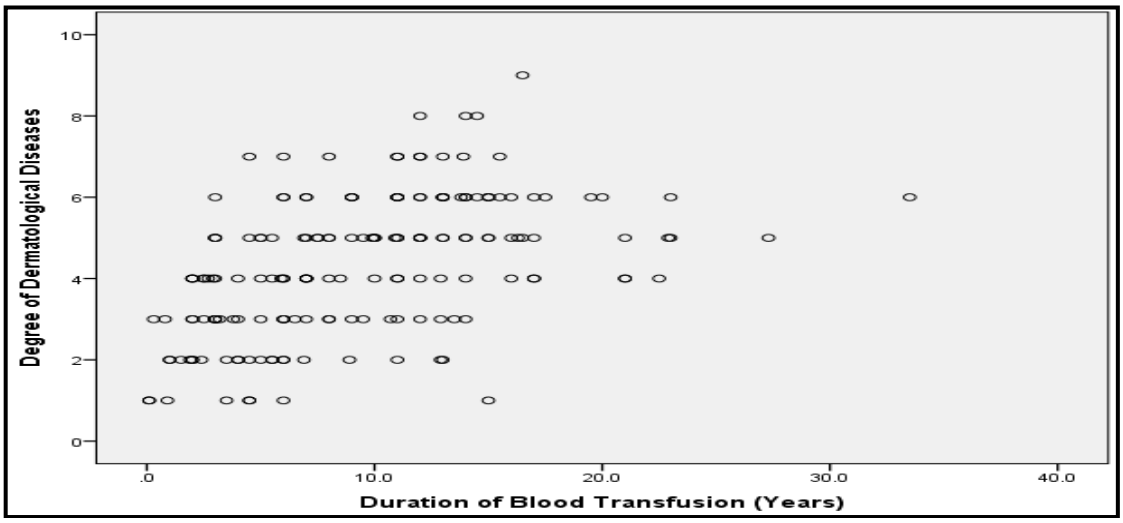


Figure 4: Relationship between duration of blood transfusion (Covariate) and degree of dermatological diseases (Response variable)

Finally, the study had been finished and remained just the conclusions and recommendation to be discussed that are the most important findings in any scientific research.

5. Conclusions

Depending on the results of the analyses in chapter three, the most important conclusions are as follows:

1) There are no significant effects of gender and living place of patients on the degree of dermatological diseases for β-thalassemia major patients.

2) The effect of duration of blood transfusion (Covariate) on the degree of dermatological diseases for β-thalassemia major patients is statistically significant.

3) There is a slightly difference between the degree of dermatological diseases of patients who living in outside cities and inside cities.

Therefore, depending on the study results and conclusions, the most important factors that had effects on the degree of dermatological diseases for β-thalassemia major patients is the total years of blood transfusion and should be taken into consideration by dermatologists and hematologists. Also, it is recommended to have more attention to patients living outside the cities may be there is few medical centers related to thalassemia.

References

1. D. J. Weatherall, The Thalassemias: Disorders of Globine, 7th ed., New York: McGrow Hill, 2006.
2. S. I. Kerim and K. M. Hasan, "Cutaneous manifestation among patients with β-thalassemia major," *Iraqi Journal of Hematology*, vol. 3, no. 2, pp. 98-107, 2014.
3. G. F. Salih and H. A. Hamakarim, "IDENTIFICATION OF β-GLOBIN MUTATIONS WHICH PRODUCED β-THALASSEMIA BY ARMS-PCR ASSAY AND DIRECT SEQUENCING," *Journal of Sulaimani Medical College,* vol. 6, no. 2, pp. 123-134, 2016.
4. "Health conditions," 20 7 2021. [Online]. Available: https://ghr.nlm.nih.gov/condition/beta-thalassemia.
5. "Health Library," 15 9 2017. [Online]. Available: http://www.hopkinsmedicine.org/healthlibrary/conditions/hematology_and_blood_disorders/beta_thalassemia_cooleys_anemia_85,P00081/

A. M. Saud, *Molecular and Biochemical Study on β-Thalassemia Patients in Iraq (PhD Thesis),* Baghdad: University of Baghdad, 2012.

A. Dogramaci, N. Savas, B. Ozer and N. Duran, "Skin diseases in patients with beta-thalassemia major," *International Journal of Dermatol,* vol. 48, no. 10, 2009.

A. Field, Discovering Statistics Using IBM SPSS Statistics, 4th ed., Sage, 2013.

A. C. Rencher and G. B. Schaalje, Linear Models in Statistics, 2nd ed., New Jersey: John Wiley & Sons, 2008.

6. University-of-Sheffield, "Stats tutor," University of Sheffield, 26 9 2021. [Online]. Available: http://www.sheffield.ac.uk/polopoly_fs/1.531229!/file/MASH_ANCOVA_SPSS.pdf.
7. D. C. Howell, Statistical Methods for Psychology, 7th ed., California: WADSWORTH , 2009.
8. M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, Applied Linear Statistical Models, 5th ed., New York: McGraw-Hill, 2005.

Comparison of Some Censored Regression Models with Application

Raaed Fadhil Mohammed
Lecturer, College of Administration & Economics, University of Mustansiriya, Baghdad, Iraq,
ORCID ID: 0000-0002-7309-8702,
E-mail: raad@uomustansiriyah.edu.iq

- ABSTRACT:

Censored regression models are among the critical statistical models used in many studies and research, mainly in which the data are restricted in one part and not another. Hence, it is difficult to use and apply traditional regression models to those data to analyze and study the relationship between the variables affecting them. Many well-known censored regression models, such as the Tobit model, are the most widely used among these models. In this paper, we used extended models of statistical distributions transformed into regression models capable of analyzing survival and reliability functions within the censored data. They can be considered as censored regression models, including the Log-BXII Weibull model from the Weibull distribution and the Log-BXIIEE model from the Exponentiated Exponential distribution in addition to the Tobit model and estimating the parameters of these models by the (MLE) method and applying these models to the data of (259) patients with renal failure (156 males and 103 females) of the dependent variable (Urea) and comparing the results of the three models through the use of comparison criteria (AIC, and H-QIC). Using SAS software showed that the three models are equivalent and parallel in importance, meaning researchers can use any new model rather than the Tobit, the most widely used censored regression model.

Keywords: Tobit Model, Log-BXIIEE Model, Log-BXIIW Model.

- INTRODUCTION:

Studying regression models, their data structure, and parameter estimation of those models by conducting statistical inference are essential issues and the focus of many researchers' attention which are compatible with regression models and considerable progress in technology that can estimate the models. Linear regression models have been and still are the most significant share because they are one of the most influential and well-established statistical tools in many applied research fields. They are the most common and used for analysing

experimental problems in social, economic, and life sciences. Because it allows estimating the effect of explanatory variables on the dependent variable by returning point estimates and standard errors to calculate confidence intervals and p-values. Various deductive models and tools exist, from traditional linear models to nonparametric regression.

The censored regression model is concerned with the study and analysis of the censored data where it is possible to define censoring in statistics as a case in which the value of the measure or the value of observation is partially known meaning that the information is incomplete. At the same time, there is a case in which the information is entirely concerning (uncensored data). The (censored data) can be defined as the data or values that have not been observed, or it is impossible to measure their viewing or complete record information about the viewing during the study or research period. Furthermore, we can say that censored data is the number of failed units for a particular period of the experiment or the period specific to the investigation and then knowing the number of failed units, and it can be said that the value or observation occurs outside the parameters of the study or measurement tool. While addressing the issue of censored data, we must demonstrate that the data points to a gap in the sample's knowledge, making it challenging to determine or assess whether the distribution of those data is known.

The most critical problems facing statisticians or those interested in applied statistics are the estimation of unknown parameters of the statistical model and when there is more than one model representing the phenomenon studied. Therefore, the problem of this paper is described in the estimate of the parameters of the three models used in our study, which are the Tobit Model, Log-BXIIW, and Log-BXIIEE, and comparing those models in the presence of censored data, then selecting the best model. Also, we reviewed censored data, the MLE method, the Tobit model, the LBXIIW model, and the LBXIIEE model; the comparison criteria (Akaike Information Criteria, Hannan-Quinn Information Criteria) were also applied in the practical aspect of the study on sample 259 patients.

- CONCEPT OF CENSORED:

When data or observations are constrained in one defined part and undefined in the other, this is called censored data; therefore, using the traditional regression model will result in biased and inconsistent estimates. As a result, a suitable model for that data must be chosen, and this model is known as the censored regression model.

Censored data are values or data that are not observed, the inability to measure them, or the failure to record complete information about the

observation during the study period. It may be characterized as figuring out how many units failed within a given experiment, deciding on a specified time for research, and then knowing how many units failed. There is another definition, which is the value or observation that occurs outside the study scope at the measurement instrument. The types of data subject to censoring include: (Lawless, 2011)

1- Type I Censored Data:

It is also called right censoring when a sample of size (n) and a predefined period will be the number of observations subject to (r). The observed value of (k) is an unexpected result, and the time of the experiment is fixed (time is a fixed amount); the number of observed failures is a random variable because failure times on the right are missing, and (r) is a random variable that cannot be determined until after time expires. (Balakrishnan, 2013)

2- Type II Censored Data:

It is also called left censoring when the sample size (n) and the number of observable observations (r) are fixed and predetermined". Therefore, the experimental time will be time (t), the random variable that cannot be determined. This type of censoring is used in event time studies. (Time-To Event). The main difference between the first and second types of observation is the random outcome under the censored data. (Rinne, 2008)

3- Type III Censored Data:

In this type of censoring, data or units of observations are located somewhere on a time interval between two points.

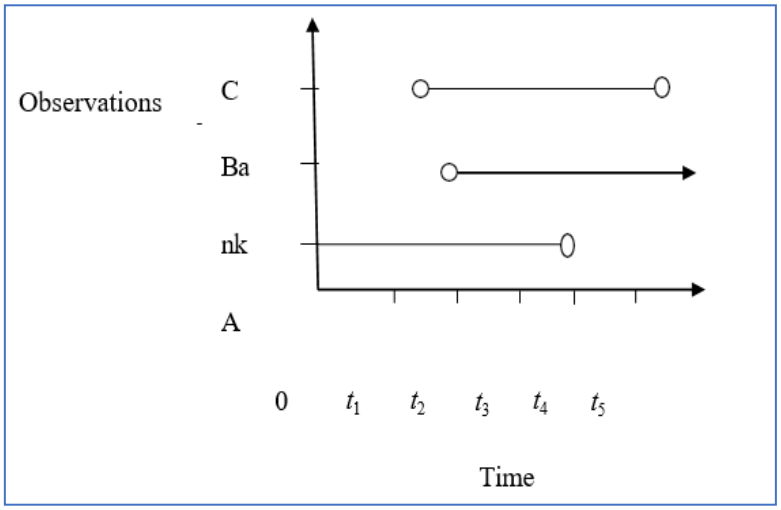


Figure 1. Description of the three types of censoring A, B, and C. (Salih, 2015)

- MAXIMUM LIKELIHOOD ESTIMATION METHOD (MLE):

MLE Method is one of the most important and widely used parametric estimation methods. In this method, we find the estimation of the parameters that have the maximum likelihood function, and the estimators of this method are characterized by their excellent accuracy compared to other methods. It expresses the likelihood function as follows:

$$L(x_1, \dots, x_n; \vartheta) = f(x_1; \vartheta).f(x_2; \vartheta) \dots f(x_n; \vartheta) = \Pi_{i=1}^{n} f(x_i; \vartheta) \quad \forall = 1,2, \dots, n \qquad (1)$$

$$L(x_1, \dots, x_n; \vartheta) = [\vartheta^{x_1}(1-\vartheta)^{1-x_1}] \dots [\vartheta^{x_n}(1-\vartheta)^{1-x_n}] = \vartheta^{\sum_{i=1}^{n} x_i}(1-\vartheta)^{n-\sum_{i=1}^{n} x_i} \qquad (2)$$

by taking the natural logarithm of (2)

$$lnL = \sum_{i=1}^{n} x_i \, ln\, \vartheta + (n - \sum_{i=1}^{n} x_i)\, ln(1-\vartheta) \qquad (3)$$

When taking the partial derivation of the equation (3) for ϑ, the following is obtained:

$$\frac{\partial lnL}{\partial \vartheta} = (\sum_{i=1}^{n} x_i)\frac{1}{\vartheta} + (n - \sum_{i=1}^{n} x_i)\frac{1}{(1-\vartheta)}(-1) = \frac{\sum_{i=1}^{n} x_i}{\vartheta} \qquad (4)$$

By equaling the above equation to zero get the maximum likelihood estimate. (Bhuyan, 2010)

- CENSORED REGRESSION MODELS:

It is necessary to model some distributions to generate new shapes and models that can be used to find new regression models that can be utilized in the study of survival analysis and reliability within censored data, resulting in censored regression models.

Here, the focus is on the log-BXII Weibull model and the log-BXIIEE model using the BurrXII system, which is a distribution that emerges from the Burr distribution and because it is used to transform distributions, it is called the BurrXII system. The BurrXII distribution, which contains Logistic and Weibull figures, is a prevalent and special model for survival data modeling and phenomenology modeling. (Arellano et al., 2012)

In 1942 Burr distribution was created by (Irving W. Burr) and the name of the distribution is derived from it; since the probability density function (P.D.F) has a set of shapes, Burr distribution is useful for approximating the graphs of those shapes, especially when a mathematical environment is needed for the (C.D.F). Burr distribution

includes different distributions such as normal, log-normal, and logistic distributions. (Paranaiba, et al., 2011)

1- Tobit Model:

The Tobit model is the result of a combination of multiple regression model and Probit analysis, where the values of the dependent variable (y*) are not observed for some observations. However, the values of the explanatory variables (X's) can be observed for all observations. The Tobit model was proposed by James Tobin (1958) to describe the relationship between the dependent variable and the explanatory variables through his study of household expenditures on durable goods using a regression model, taking into account the fact that expenses (as a dependent variable for his model) are a positive value that does not can be negative. The Tobit model deals with dependent variable data because it is divided into two sections, each with its distribution function. Observations with values equal to or close to (0) take the (C.D.F), while observations with values greater than (0) take the (P.D.F), and we get the mixed function that expresses the Tobit model by multiplying the functions (C.D.F) and (P.D.F). The Tobit model has the following general form: (Tobin, 1958)

$$y_i^* = \gamma + \mathcal{g}x_i + \varepsilon_i \qquad (5)$$

$$y_i = \begin{cases} y^* & if \;\; y^* > \lambda \\ 0 & if \;\; y^* \le \lambda \end{cases} \qquad (6)$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad , \quad y^* \sim N(x\mathcal{g}, \sigma^2)$$

Where:

γ: a constant.

λ: restriction point.

y_i: the dependent variables.

y^*: the latent variable.

$\mathcal{g}$: the model's parameters.

x_i: the explanatory variables.

We derive the mixed function for the following by multiplying equations (2) and (3):

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y - x\mathcal{g})^2}{2\sigma^2}\right)\right]\left[1 - \vartheta\left(\frac{\lambda - x\mathcal{g}}{\sigma}\right)\right] \qquad (7)$$

When $\lambda = 0$

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y - x\mathcal{g})^2}{2\sigma^2}\right)\right]\left[1 - \vartheta\left(\frac{-x\mathcal{g}}{\sigma}\right)\right] \qquad (8)$$

And it can be expressed as follows

$$f(y) = \left[\frac{1}{\sigma}\phi\left(-\frac{(y_i - x\not{g})}{\sigma}\right)\right]\left[1 - \vartheta\left(\frac{-x\not{g}}{\sigma}\right)\right] \quad (9)$$

Where:

$\phi(.)$: the probability density function (P.D.F).

$\vartheta(.)$: the cumulative function (C.D.F).

Applying the maximum likelihood method, get the following:

$$L\big(f(y)\big) = \prod_{i=1}^{n}\left[\frac{1}{\sigma}\phi\left(-\frac{(y_i - x\not{g})}{\sigma}\right)\right]\left[1 - \vartheta\left(\frac{-x\not{g}}{\sigma}\right)\right] \quad (10)$$

by taking the natural logarithm of (3), we get:

$$ln\, L = \sum_{i=1}^{n}\left[-ln\,\sigma + ln\,\phi\,\frac{(y_i - x\not{g})}{\sigma}\right] + ln\left[1 - \vartheta\left(-\frac{x\not{g}}{\sigma}\right)\right] \quad (11)$$

We utilize numerical approaches to acquire estimates of the Tobit model since it is difficult to discern the estimation of the most significant possibilities of the above equation. (Carson, 2007)

2- Log-BXII Weibull Model:

This distribution was first utilized by French scientist Maurice in 1927. However, the Swedish scientist Weibull explicitly classified it as one of the continuous distributions and one of the most commonly used failure models in recent years. It gained significance and a significant position in life testing and reliability. When the rigorous randomness constraints of the exponential distribution are not satisfied, this distribution is generally appropriate. As a result, this distribution increases versatility and comes in three variants based on the number of parameters: one, two, and three. (Thomopoulos, 2017)

Suppose (x) is a random variable that represents the failure time. The Weibull distribution's probability density function (P.D.F.) and cumulative distribution function (C.D.F.) are then stated as follows:

$$f(x;\gamma^*,\not{g}^*) = \gamma^*\not{g}^{*-1}\,(x/\not{g}^*)^{\gamma-1}\,exp\left(-\left(\frac{x}{\not{g}^*}\right)^{\gamma^*}\right) \quad (12)$$

$$F(x;\gamma^*,\not{g}^*) = 1 - exp\left(-\left(\frac{x}{\not{g}^*}\right)^{\gamma^*}\right) \quad (13)$$

Where: $x \geq 0, \gamma^* \geq 0, \not{g}^* \geq 0$
x: a random variable.
γ^*: a shape parameter.
$\not{g}^*$: a scale Parameter.

Several regression models have been presented in recent years, such as the log-BurrXII Weibull model, built using the location regression model, and the Burr distribution with Weibull. LBXIIW may be stated mathematically using the following formula: (Yousof, et al., 2018)

$$f(x) = \tau b \gamma^* \not{g}^{*-1} \left(\frac{x}{\not{g}^*}\right)^{\gamma^*-1} exp\left(-\frac{x}{\not{g}^*}\right)^{\gamma^*} \frac{\left[1-exp\left(-\frac{x}{\not{g}^*}\right)^{\gamma^*}\right]^{\tau-1}}{\left[1-exp\left(-\frac{x}{\not{g}^*}\right)^{\gamma^*}\right]^{\tau+1}} \left(1 + \left(\frac{1+exp\left(-\frac{x}{\not{g}^*}\right)^{\gamma^*}}{1-exp\left(-\frac{x}{\not{g}^*}\right)^{\gamma^*}}\right)^{\tau}\right)^{-(b+1)} \quad (14)$$

$$x \sim BXIIW\ (\gamma^*, \not{g}^*, \tau, b)$$

Where (τ) and (b) are the parameters of the Burr distribution, location regression models can be distinguished by the random variable $y = log(x)$ having a distribution where position $\mu(v)$ leans on the independent variable vector (v) in practice; this means that for different levels or values of the independent variable (x), the position parameter has different values. The following is the location regression model:

$$y = \mu(v) + \sigma z \quad (15)$$

Where z has a non-dependent distribution on (v), and the random variable (y) has a probability function for each $(y \in R)$. Given that the random variable (x) has a probability function (13) and is distributed BXIIW, and that the random variable $y = log(x)$, $\gamma^* = e^{-\mu}$ and $\not{g}^* = 1/\sigma$.
The probability function (11) is as follows:

$$f(y, \tau, b, \mu, \sigma) = \frac{\tau b}{\sigma} \frac{[1 - exp\left(-e^{(y-\mu)/\sigma}\right)]^{\tau-1}}{exp\left(-\tau e^{\frac{y-\mu}{\sigma}} - \left(\frac{y-\mu}{\sigma}\right)\right)} [1 + [exp\left(e^{\frac{y-\mu}{\sigma}}\right) - 1]^{\tau}]^{-(b+1)} \quad (16)$$

Where $\tau > 0$, $b > 0$, are the shape parameters and that $(\mu \in R)$ is the location parameter and $\sigma > 0$ is the scale parameter. (Thomopoulos, 2017)

Given that the equation (16) reflects the LBXIIW distribution and that $y\sim LBXIIW\ (\tau, b, \mu, \sigma)$ and $Y\sim LBXIIW\ (\tau, b, \mu, \sigma)$, the standard random variable $Z = (Y - \mu)/\sigma$, the (P.D.F) for (Z) is as follows:

$$f(z;\tau,b) = \tau b\ \frac{[1-exp(-e^z)]^{\tau-1}}{exp(-e^z - z)}\ [1 + [exp(e^z) - 1]^{\tau}]^{-(b+1)} \qquad (17)$$

The linear regression model that links (y_i) and the independent variable vector $x_i^T = (v_{i1}, \ldots, v_{ip})$, where T has a BXIIW distribution as a random variable.

$$y = \mu_i + \sigma z_{\ i} \qquad (18)$$

Where:
z_i: the random error has a (P.D.F) function (14).
μ_i: the location parameter of y_i.
$\mu_i = {v_i}^T\tau,\ \tau = (\tau_1, \ldots, \tau_p)^T$ the vector related to independent variables.
$\tau > 0,\ b > 0$, σ: unknown parameters.
The LBXIIW model can also be used to fit a variety of data types where independent variables influence the average response to Y significantly. Suppose that D and E are data sets representing log-lifetime and log-censoring, respectively. The MLE approach and the natural logarithm of equation (18) are used to estimate the vector ($\psi = (\tau, b, \kappa^T, \sigma)$) parameters. In that case, the following results will be obtained:

$$u_i = e^{z_i} \qquad , \qquad z_i = \frac{(y_i - \mu_i)}{\sigma}$$

Where (r) represents the number of failed observations, the equation will be as follows:

$$ln(\psi) = r\,ln(\tau b) - r\,ln\,\sigma + (\tau - 1)\sum_{i\in G} ln(1 - e^{-u_i}) + \tau\sum_{i\in G}(u_i + z_i) - (b+1)\sum_{i\in G} ln(1 + (e^{u_i - 1})^{\tau}) - b\sum_{i\in G} ln\left(1 + \left(\frac{1 - e^{-u_i}}{e^{-u_i}}\right)^{\tau}\right) \qquad (19)$$

The estimation for the parameters vector $\hat{\psi}$ can be discovered using statistical software, including the NLMixed library in SAS software. (Lawless, 2011)

3- Log-BXII Exponentiated Exponential Model:

This distribution was discussed by Gupta (1998). It is a novel continuous distribution of exponential distribution families. It has to scale and shape parameters comparable to those found in the Weibull distribution family and the gamma distribution. Features of the Weibull and Gamma distributions are identical to the characteristic of this distribution, which may also be employed as an alternative. The Exponentiated Exponential distribution function (P.D.F) can be represented as follows:

$$f(x;\gamma^{**},\lambda^{*}) = \gamma^{**}\,\lambda^{*}\left(1-e^{-\lambda^{*}x}\right)^{\gamma^{**}-1} e^{-\lambda^{*}x} \qquad (20)$$

The (C.D.F) can be represented as follows:

$$F(x;\gamma^{**},\lambda^{*}) = (1-e^{-\lambda^{*}x})^{\gamma^{**}} \qquad (21)$$

Where:
λ^{*}: a scale parameter of the exponentiated exponential distribution.
γ^{**}: a shape parameter of the exponentiated exponential distribution.
The mathematical formula for BXIIEE is: (Ibrahim, et al., 2020)

$$f(x) = \tau b\gamma^{*}\lambda^{*}\, e^{-\lambda^{*}x}\, \frac{\left[1-e^{-\lambda^{*}x}\right]^{\tau\gamma^{**}-1}}{\left[1-\left(1-e^{-\lambda^{*}x}\right)^{\gamma^{**}}\right]^{\tau+1}} \left(1+\left(\frac{\left[1-e^{-\lambda^{*}x}\right]^{\gamma^{**}}}{1-\left(1-e^{-\lambda^{*}x}\right)^{\gamma^{**}}}\right)^{\tau}\right)^{-b-1} \qquad (22)$$

$$x\sim LBXIIEE\,(\tau,b,\gamma^{**},\lambda)$$

The model may be stated using the location-scale regression model, which relates the independent variable vector $v_i^T = (v_{i1},\ \dots\ , v_{ip})$ with the response variable average, as follows:

$$y = v_i^T \not{g} + \sigma z_i \qquad ,\forall\, i = 1,\dots,n \qquad (23)$$

T is a random variable that tracks the BXIIEE distribution, whereas y follows the LBXIIEE distribution. (Kundu, et al., 2001)
Assuming that M and N are sets of items for y_i representing log-censoring or log-lifetime, respectively, and that censoring times and lifetimes are independent. The LBXIIEE regression model will be as

follows when using the (MLE) method to estimate the parameters of the vector $\xi = (\tau, b, \gamma^{**}, \lambda^{*}, \not{g}^{T})$ and taking the natural logarithm: (Lawless, 2011)

$$ln(\xi) = r\, ln\left(\frac{\tau b \gamma^{**} \lambda^{*}}{\sigma}\right) - \lambda^{*} \sum_{i \in M} u_i + (\gamma^{**} - 1) \sum_{i \in M} ln(1 - exp(-\lambda^{*} u_i))$$
$$+ (\gamma^{**} - 1) \sum_{i \in M} ln\,(1 - exp(-\lambda^{*} u_i)^{\tau}$$
$$+ (\gamma^{**} + 1) \sum_{i \in M} ln\,(1 - (1 - exp(-\lambda^{*} u_i))^{\gamma^{**}}$$
$$-(b+1) \sum_{i \in M} ln\left(1 + \left(\frac{(1 - exp(-\lambda^{*} u_i))^{\gamma^{**}}}{1 - (1 - exp(-\lambda^{*} u_i))^{\gamma^{**}}}\right)^{\tau}\right)$$
$$+ \sum_{i \in M} ln\left(1\right.$$
$$-\left(1\right.$$
$$\left.\left. + \left(\frac{(1 - exp(-\lambda^{*} u_i))^{\gamma^{**}}}{1 - (1 - exp(-\lambda^{*} u_i))^{\gamma^{**}}}\right)^{\tau}\right)^{-b}\right) \qquad (24)$$

Where:

$$u_i = e^{z_i} \qquad , \qquad z_i = \frac{(y_i - v_i^{T} \not{g})}{\sigma}$$

The feature vector $(\hat{\xi})$ can be estimated using the NLMixed library in SAS.

- COMPARISON CRITERIA:

It is known statistically that several models are used to study a particular phenomenon to reach the best model that represents it. There must be a comparison criterion based on which comparisons are made between these models, and there are many statistical criteria through which comparisons are made in choosing the best model. We will address some of these criteria, which are AIC and HQIC, and below, we will review these standards.

1- Akaike Information Criterion (AIC):

Named after Japanese statistician Hirotsugu Akaike coined in 1974, the AIC rates the quality of each model when there is a set of data

models, and provides a means for selecting the best model; its mathematical formula is as follows:

$AIC = -2\hat{\rho} + 2p$ (25)

Where:

p: the number of model parameters.

$\hat{\rho}$: the maximum likelihood function value.

The model that has the lowest value for this criterion is the best. (Takane, et al., 1987)

2- Hannan-Quinn Information Criterion (H-QIC):

It is a criterion proposed by Hannan & Quinn in 1979. This criterion is characterized as being objective and automatic and is highly proportional to small sample sizes and its formula:

$HQIC = -2\hat{\rho} + 2p\ log\ [log(n)]$ (26)

Where:

n: sample size

p: the number of model parameters.

$\hat{\rho}$: the highest value of the maximum likelihood function.

The model that has the lowest value for this criterion is selected. (So, et al., 2009)

- RENAL FAILURE:

Renal failure disease is a deficiency in the work and functions of the renal, which leads to a general imbalance in the body. It is also known as the deterioration of the ability of the renal to filter impurities in the blood, where the renal becomes unable to perform its typical job of filtering the body's by-products from the blood. Renal failure often results from persistently high blood pressure or diabetes, in addition to infections and genetic diseases of the renal that can cause permanent failure. Renal failure takes two forms, acute and chronic.

1- Acute Renal Failure:

It is the sudden cessation of renal function from working for a few hours, days or weeks in a critical manner, where toxins and metabolic products accumulate in the body in the blood, which leads to a sudden rise in the proportion of urea and creatine in the blood. The renal usually returns to total efficiency when the cause is gone.

2- Chronic Renal Failure:

Its symptoms may not begin to appear until after the renal efficiency is less than 25% of its function volume, and the renal does not return to its function even after the removal of the cause. This disease was chosen as a practical aspect of this study for presenting actual data for patients

with renal failure and the absence of a previous statistical survey dealing with kidney failure in Kirkuk governorate-Iraq. Data on this disease was collected from the patient records in Kirkuk General Hospital / Industrial Renal Unit for 259 patients, 156 patients from male and 103 female patients.

The stages of renal failure are five steps and are calculated using the filtration rate:

First stage: Renal function decreases with few symptoms.

Second and Third stage: The need for care increases to relieve and treat renal dysfunction.

Fourth and Fifth stage: The patient needs treatments, and the disease is considered severe in these stages and requires dialysis or a kidney transplant if possible.

- DATA ANALYSIS AND ESTIMATION:

In this section, we review the tables for estimating the three censored regression models used in this paper and the criteria used according to the variables (gender - age - Urea - Creatinine). Where the analysis and estimation of three models (Tobit, LBXIIW, and LBXIIEE) are done when Urea is a dependent variable, analyzing these models includes extracting T-test values, standard error values (S.E), and P-values based on variables that significantly affect the dependent variable. In addition to the table of comparison criteria between models above to know the best model, the scales of Urea (0.5-1.0) mg-dl and creatinine (0.5-1.1) mg-dl.

- Urea As a Dependent Variable:

The three models (Tobit, LBXIIW, LBXIIEE) are estimated for the data of patients, considering that the dependent variable is the Urea, after the same variable is calculated for the data in general in the previous paragraphs. The following tables show the results of the analysis and estimation of the mentioned models for patients.

1- Practical Application Using the Tobit Model:

Table 1. Tobit model parameter estimation for urea variable.

Step	Variable	$\hat{g}$	S.E	T	P-value
1	constant	18.540	6.773	3.514	0.0119
	age	-0.057	0.113	-0.301	0.6163
	creatinine	20.679	5.134	4.534	0.0004
2	constant	15.933	4.9808	3.831	0.0046
	creatinine	20.437	5.1211	4.502	0.0005

Table (1) shows that the Tobit model estimates are calculated based on the dependent variable urea; In the first step, we notice that the value of the t-test for the creatinine variable was significant, amounting to (4.534), with a significant percentage of (0.0004). In the second stage, the age variable is eliminated, and the creatinine variable is the explanatory variable with the most influence on the dependent variable.

Table 2. Tobit model criterion values for urea variable.

Step	AIC	H-QIC
1	102.7	106.5
2	100.9	103.8

Table (2) shows the Tobit model criteria table for the variable Urea for patients. The values of criterion AIC and H-QIC change depending on the stages; in Table (1), each step indicates the importance of the two criteria for the same stage, and each step represents the values of the two criteria for the same stage. The graphic below shows the Tobit model's projected value for the dependent variable.

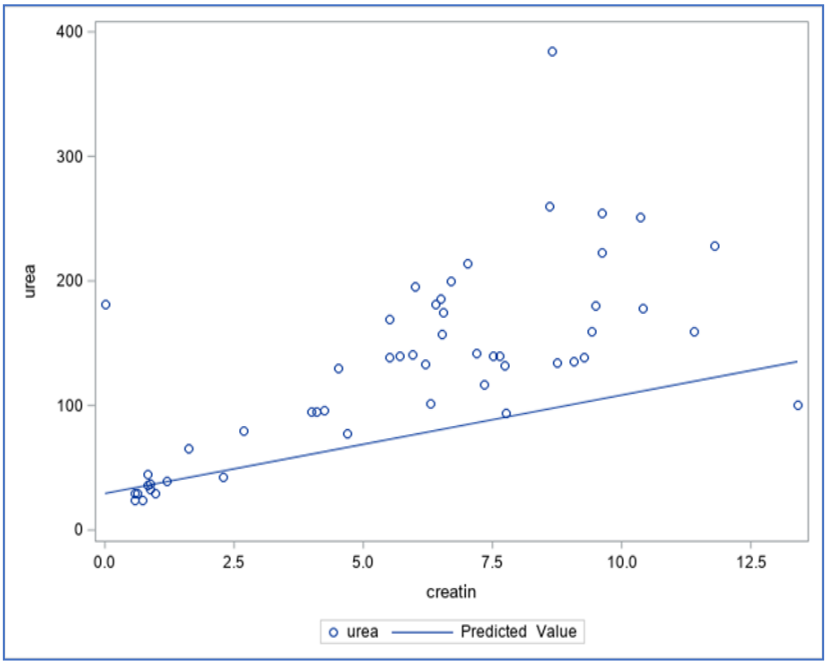


Figure 2. Predicted value for Tobit model.

2- Practical Application Using the LBXIIW Model:

Table 3. LBXIIW model parameter estimation for urea variable.

Step	Variable	g	S.E	T	P-value
1	constant	22.216	54.902	0.42	0.5791
	age	0.0284	0.3296	0.93	0.8317
	creatinine	55.518	14.402	4.67	0.0033
	γ	16.699	18.058	0.93	0.2559
	β	0.1573	0.1431	1.17	0.1459
	σ	147.36	123.17	1.19	0.1357
2	constant	23.486	49.317	0.49	0.5279
	creatinine	55.853	14.978	4.56	0.0038
	γ	16.398	17.591	0.94	0.2522
	β	0.1613	0.1433	1.27	0.1356
	σ	147.07	120.13	1.29	0.1252

Table (3) shows that the LBXIIW model estimates are calculated based on the dependent variable urea; In the first step, we notice that the value of the t-test for the creatinine variable was significant, amounting to (4.67), with a significant percentage of (0.0033). In the second stage, the age variable is eliminated, and the creatinine variable is the explanatory variable with the most influence on the dependent variable.

Table 4. LBXIIW model criterion values for urea variable.

Step	AIC	H-QIC
1	138.8	144.5
2	135.3	139.1

Table (4) displays the LBXIIW model's criteria for the dependent variable Urea for patients. We see that the values of the criteria AIC and H-QIC fluctuate depending on the steps. These criteria attain their lowest value in the final phase, and each step indicates the importance of the three criteria for the same stage in Table (3). The figure below shows the predicted value for the dependent variable for the LBXIIW model.

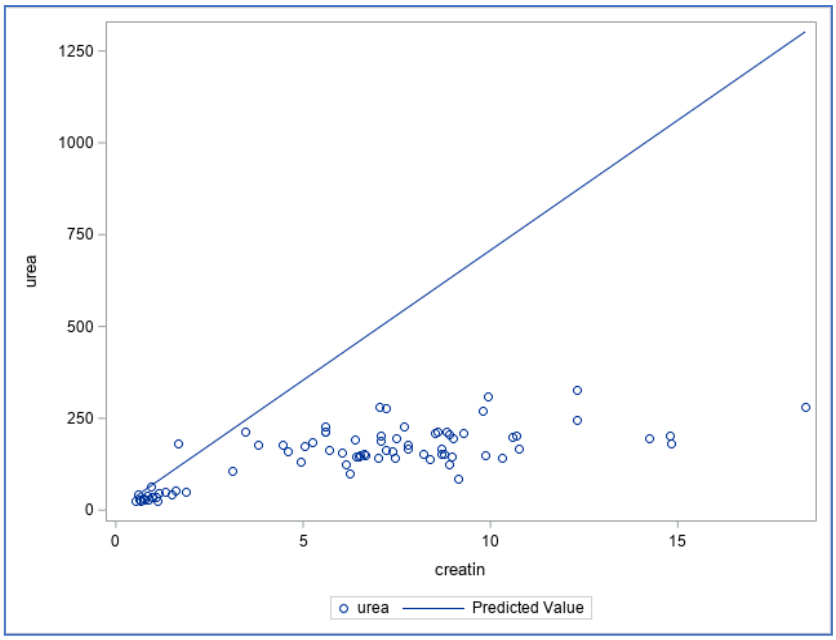


Figure 3. Predicted value for LBXIIW model.

3- Practical Application Using the LBXIIEE Model:

Table 5. LBXIIEE model parameter estimation for urea variable.

Step	Variable	g	S.E	T	P-value
1	constant	30.342	5.4984	5.82	0.0000
	age	-0.0829	0.0696	-1.09	0.2373
	creatinine	4.4287	2.1717	2.71	0.0909
	τ	21.730	3.2020	6.43	0.0000
	b	2.6620	0.0880	43.64	0.0000
	γ	48.872	0.9641	52.73	0.0000
	β	0.0925	0.0800	2.06	0.2513
	σ	150.01	165.85	1.03	0.3682
2	constant	41.197	15.197	3.61	0.0110
	creatinine	5.4913	0.9152	8.09	0.0000
	τ	23.399	7.9206	3.74	0.0077
	b	3.3951	0.0001	3.99	0.0000
	γ	25.994	18.582	1.48	0.1718
	β	0.1015	0.0753	1.55	0.1814
	σ	108.08	93.144	1.26	0.2501

In the first step of Table 5, we see that the t-test results for the variables (τ, b, and γ) are all significant, with values of (6.43, 43.64, and 52.73). The variable (γ) became non-significant in the second step. Thus, it was

eliminated. The t-test values for the variables (creatinine, τ, b) are all significant at the level of significance (p-value ≤ 0.0001), with values of (8.09, 3.74, and 3.99). Table (6) shows the LBXIIEE model's criteria for the dependent variable Urea in renal failure patients. The values of criterion AIC and H-QIC change depending on the stages. In the final phase, these criteria achieve their lowest value, and each step reflects the values of the three criteria for the same stage in Table (5).

Table 6. LBXIIEE model criterion values for urea variable.

Step	AIC	H-QIC
1	97.2	104.8
2	94.7	101.3

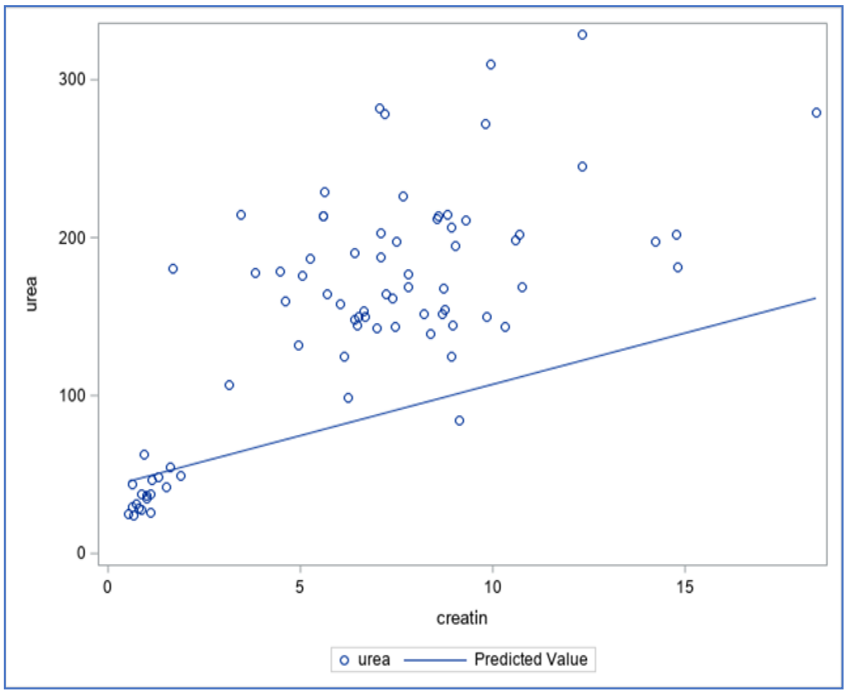


Figure 4. Predicted value for LBXIIEE model.

In figures (2), (3), and (4), notice that the three models are almost equivalent to each other, as some estimated values approach the actual values. Also, we note that the model (LBXIIEE) is the best model for data.

The following table (7) summarizes the criteria values for renal failure patients' data, which shows that the three models are close in preference data used in this paper. The LBXIIEE model is the best of the two models (TOBIT and LBXIIW), where it has the lowest values of criteria (AIC

and H-QIC). Furthermore, the TOBIT model is better than the LBXIIW model, as it has the lowest values of criteria compared with the LBXIIW model for data of renal failure patients for the same dependent variable.

Table 7. Comparative AIC and H-QIC values for the three models.

Model	AIC	H-QIC
TOBIT	101.9	104.8
LBXIIW	136.3	140.1
LBXIIEE	95.7	102.3

- CONCLUSIONS:

- Statistical distributions transformed into regression models can be used as censored regression models from the BurrXII system for distributions applied to survivals.
- Through the practical application, we found all three censored regression models can use in dealing with the censored data instead of relying on the Tobit model, which is the most common model in most cases.
- We notice that the three models are almost equivalent, as some estimated values approach the real values. Also, we note that the model (LBXIIEE) is the best model for renal failure patients' data, where it has the lowest criteria (AIC and H- QIC).

- REFERENCES:

[1] Arellano-Valle, R. B., Castro, L. M., González-Farías, G., & Muñoz- Gajardo, K. A. (2012). Student-t censored regression model: properties and inference. Statistical Methods & Applications, 21(4), 453-473.

[2] Balakrishnan, N., & Kundu, D. (2013). Hybrid censoring: Models, inferential results and applications. Computational Statistics & Data Analysis, 57(1), 166-209.

[3] Bhuyan, K. C. (2010) "Probability Distribution Theory and Statistical inference ". American International University, Bangladesh.

[4] Carson, R. T., & Sun, Y. (2007). The Tobit model with a non-zero threshold. The Econometrics Journal, 10(3), 488-502.

[5] Gupta, R. D., & Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 43(1), 117-130.

[6] Ibrahim, M., Ea, E. A., & Yousof, H. M. (2020). A new distribution for modeling lifetime data with different methods of estimation and censored regression modeling. *Statistics, Optimization*

& Information Computing, *8*(2), 610-630.
[7] Lawless, J. F. (2011). Statistical models and methods for lifetime data (Vol.362). John Wiley & Sons.
[8] Paranaíba, P. F., Ortega, E. M., Cordeiro, G. M., & Pescim, R. R. (2011). The beta Burr XII distribution with application to lifetime data. Computational Statistics & Data Analysis, 55(2), 1118-1136.
[9] Rinne, H. (2008). The Weibull distribution: a handbook. CRC press.
[10] Salih, M. A. (2015) "Parameters Estimation for Pareto Type – II Distribution "PhD Thesis-College of Administration and Economics-Al- Mustansiriya University.
[11] So, I. (2009). Comparison of criteria for estimating the order of autoregressive process: a Monte Carlo approach. European Journal of Scientific Research, 30(3), 409-416.
[12] Takane, Y., & Bozdogan, H. (1987). Akaike Information Criterion (Aic)- Introduction.
[13] Thomopoulos, N. T. (2017). Statistical distributions. Applications and Parameter Estimates. Cham, Switzerland: Springer International Publishing.
[14] Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society, 24-36.
[15] Yenilmez, İ., Kantar, Y. M., & Acitas, S. (2018). Estimation of censored regression model in the case of non-normal error. Sigma Journal of Engineering and Natural Sciences-Sigma Muhendislik ve Fen Bilimleri Dergisi, 36, 513-521.

- APPENDIX:

```
%%% SAS CODE
%%% Tobit Model
proc import datafile='C:\Users\dell\Desktop\dataset.xlsx' DBMS =xlsx
Out=dataset;
run;
proc qlim data = dataset;
model urea = age gender creatin;
endogenous urea ~ censored (ub=45);
run;
%% % LBXIIW Model
proc import datafile='C:\Users\dell\Desktop\dataset.xlsx' DBMS =xlsx
Out=dataset;
run;
```

```
proc nlmixed data=dataset;
parms b0= 10  b1=2  b2=.5        b3= 1  a=2 b=1  s=  150;
m = b0 + (b1*age) + (b2*gender) + (b3*creatin);
y=urea;
if status_u=1 then f =((a*b)*((1-exp(-exp((y-m)/s)))**(a-1)) *
((1+((exp(exp((y-
m)/s))-1)**(a)))**(-(b+1)))) /(s*exp((-a*exp((y-m)/s))-((y-m)/s))) ;
else f = (1+(((1-exp(-exp((y-m)/s)))/(exp(-exp((y-m)/s))))**a))**-b;
ll=log(f);
model y ~ general(ll);
run;
%%% LBXIIEE Model
proc import datafile='C:\Users\dell\Desktop\dataset.xlsx' DBMS =xlsx
Out=dataset;
run;
proc nlmixed data=dataset;
parms b0=1 b1=1 b2=1 b3=1 a=1 b=2.5 aa=2 bb=1 s=150;
m =b0 + (b1*age) + (b2*gender) + (b3*creatin);
y=urea;
if status_u=1 then  f=((aa*bb*a*b)/s)*(exp(-b*exp((y-m)/s)))*((1-exp(-
b*exp((y- m)/s)))**(a-1))*((((1-exp(-b*exp((y-m)/s)))**a)**(aa-1))/((1-
((1-exp(-b*exp((y-
m)/s)))**a))**(aa+1)))*((1+((((1-exp(-b*exp((y-m)/s))) **a)/(1-((1-exp(-
b*exp((y- m)/s)))**a)))**aa))**(-bb-1)) ;
else  f= 1-((1+(((1-exp(-b*exp((y-m)/s)))**a)/(1-(1-exp(-b*exp((y-m)/s)))
**a))**aa)**-bb);
ll=log(f);
model y ~ general(ll);
run;
```