Use Principle Component Regression Method In Addressing Linear Multiplicity Problem

استعمال طريقة انحدار المركبات الرئيسية في معالجة مشكلة التعدد الخطي

أ.م. د. شروق عبد الرضا السباح زينب كاظم مزهر القريشي
جامعة كربلاء / كلية الادارة والاقتصاد

(بحث مستل من رسالة ماجستير)

المستخلص:-

عند بناء نموذج انحدار خطي متعدد غالبا ما تواجه الباحث مشكلة التعدد الخطي Co Multi Linearity ، فتصبح تقديرات طريقة المربعات الصغرى غير كفؤة ، وذو قدرة تنبؤيه ضعيفة ومن البدائل لتخطي هذه المشكلة هي طريقة انحدار المركبات الرئيسية Principal components Regression

وبهدف الوصول الى طريقة تمتلك قيم تقديرية دقيقة سحبنا عينة عشوائية بسيطة حجمها 100 شخص من المرضى المصابين بالجلطة الدماغية والقلبية الراقدين في مستشفى الحسين في محافظة كربلاء وباستعمال البرامج الاحصائية ((SPSS,NCSS) ، لتحليل البيانات والكشف عن وجود مشكلة التعدد الخطي ومعالجتها بطريقة الانحدار PCR، وقد تبين من خلال النتائج ان طريقة المركبات الرئيسية تكون طريقة ناجحة لمعالجة مشكلة التعدد وذات نتائج دقيقة .

Abstract:-

When constructing a linear linear regression model, the researcher often encounters the problem of multi-linearity. Thus, the estimation of the method of the least squares becomes inefficient, with low predictive capacity .

Thus as alternatives to overcome this problem are the regression method of the **Principal components Regression**.

The aim is to reach the best method with estimated values, from which a simple random gamble of 100 patients from patients with stroke and cardiac arrest at Al-Hussein hospital in Karbala governorate were taken we used statistical programs (SPSS, NCSS), for analysis the data.

The results showed that the the Principal components method was a successful method to address the problem of multiplicity with accurate results.

المقدمة:

تستعمل النماذج الخطية بشكل واسع في مختلف مجالات العلم وان الانحدار الخطي المتعدد هو احد النماذج الخطية التي يكثر استعماله في تحليل بيانات العديد من البحوث الطبية والصحية والاقتصادية والاجتماعية والعلوم التطبيقية الاخرى، عند تطبيق تحليل الانحدار الخطي المتعدد تواجهنا مشكلة التعدد الخطي التي تحدث عند وجود علاقة ارتباط قوية بين المتغيرات التوضيحية ويؤدي هذا الارتباط الى زيادة تباين مقدرات معلمات الانحدار الامر الذي يجعل نتائج طريقة المربعات الصغرى الاعتيادية (OLS) غير موثوق بها والتي من احد الفرضيات الاساسية لها هي وجوب استقلالية المتغيرات التوضيحية (عدم وجود ارتباط فيما بينها).

ولحل مشكلة التعدد الخطي تم اقتراح عدة طرق من قبل الكثير من الباحثين ومنها طريقة انحدار المركبات الرئيسية Principle Component Regression هو كارل بيرسون (Pearson,1901 Karl) و يعد احدى الطرق البديلة لطريقة المربعات الصغرى الاعتيادية

مشكلة البحث:

لغرض تقدير نموذج انحدار خطي لابد ان يتمتع هذا النموذج بخصائص معينة تعتمد على عدة افتراضات وفي حالة غياب احد تلك الافتراضات فأن النموذج سيعاني جملة من المشكلات والتي تجعل عملية التقدير خاطئة او في بعض الاحيان غير ممكنة ومن هذه المشكلات هي مشكلة التعدد الخطي اي عدم وجود استقلالية بين المتغيرات التوضيحية تكون البيانات مترابطة فيما بينها

اهمية البحث:

تكون اهمية هذا البحث محاولة الوصول الى فهم أعمق لظاهرة مشكلة التعدد الخطي وتناول طرق الكشف عنها واهم المظاهر الدالة على وجودها في نموذج الانحدار الخطي، ودراسة طريقة انحدار المركبات الرئيسية كصيغة مقترحة لمعالجة مشكلة التعدد الخطي.

منهجية البحث:

تم العمل على المنهج الاستقرائي وفيه يبدأ بملاحظة المشكلة ثم وضع الفروض لها وبعد ذلك اختبارها، وقد تم استعمال الاسلوب الاحصائي وفق هذا المنهج.

هدف البحث:

- تشخيص وجود مشكلة التعدد الخطي ومدى تأثيرها على نموذج الانحدار .
- دراسة مدى ارتباط المتغيرات فيما بينها وتعين المتغيرات التوضيحية ذات الاهمية الاحصائية وتأثيرها على المتغير المعتمد .
- التخلص من الارتباط الموجود بين المتغيرات المستقلة (التعدد الخطي) بطريقة انحدار المركبات الرئيسية .

الجلطة الدماغية او السكتة الدماغية [1]

السكتة الدماغية أو الجلطة الدماغية وتعرف سابقا باسم حادثة وعائية دماغية (Cerebrovascular accident – CVA) تحدث عندما يتوقف، أو يتعرقل بشدّة، تدفق الدم إلى أحد أجزاء الدماغ، مما يحرم أنسجة المخّ من الأوكسجين الضروري جدا ومواد التغذية الحيوية الأخرى. ومن جراء ذلك، تتعرض خلايا المخ للموت خلال دقائق قليلة.

السكتة الدماغية هي حالة طوارئ طبية، والعلاج الفوري لها أمر بالغ الحيوّية والأهمية، إذ يمكن من خلاله تقليل الأضرار للدماغ ومنع المضاعفات المحتملة ما بعد السكتة.

الجلطة القلبية:

وتُعرف أيضا باسم النوبة القلبية وهي مرض قلبي حاد مهدِّد للحياة يحدث بسبب احتباس الدم نتيجة انسداد أحد الشرايين التاجية مما يؤدي إلى ضرر أو موت كامل لجزء من عضلة القلب. غالباً ما تكون النوبة طارئاً طبياً يهدد حياة المريض ويستدعي الرعاية الطبية الفورية. تُشَخَّص الحالة بواسطة تاريخ المريض الطبي ونتائج فحص تخطيط القلب وإنزيمات القلب في الدم. أكثر الإجراءات اللازم اتخاذها فوراً هي إعادة تدفق الدم إلى القلب، لذا يجب الإسراع بنقل المصاب فوراً إلى مستشفى أو إلى طبيب أو إحضار الطبيب لموقع المصاب لمعالجته؛ حيث يلعب الوقت هنا دوراً هاماً جداً فتجب السرعة كي يتم إعادة تدفق الدم في الشريان التاجي بأحد أمرين أو كليهما: إذابة الخثرة (وهي كدرة الدم المسببة لانسداد الشريان) بمضادات التخثر، ورأب الوعاء (وتسمى أيضا توسعة الشرايين) وهو إيلاج وتد على متنه بالون إلى الوعاء الدموي المنسدّ، بحيث ينتفخ البالون حين يكون بمحاذاة الخثرة فتنحسر الخثرة نحو الجوانب ويتّسع فضاء الوعاء بعدما ضاق، مُتيحاً للدم الانسياب عبره. وينبغي على وحدة العناية التاجية مراقبة المريض عن كثب تحسباً لمختلف التطورات،

التعدد الخطي:-

تعريف مختصر للتعدد الخطي [4]

ان مصطلح التعدد الخطي او الارتباط الخطي المتعدد هو مصطلح مركب من (Multi))متعدد و(Co) مشترك او متداخل او مرتبط و(Linearty) خطي ويشير المصطلح الى وجود علاقة خطية تامة او غير تامة بين اثنين او اكثر من المتغيرات التوضيحية في نموذج الانحدار مما يؤدي الى احد فروض التحليل.

ويمكن تعريف التعدد الخطي من خلال مفهوم التعامد اي عندما تكون مصفوفة التصميم تامة الرتبة وجميع القيم الذاتية في مصفوفة التصميم تساوي واحد ويدل هذا على ان التعامد موجود . واذا كان على الاقل واحد من القيم الذاتية مساويا للصفر او يقترب من الصفر يعني ان التعامد غير موجود.

انواع العلاقات الخطية بين المتغيرات التنبؤية [5]

العلاقة الخطية التامة

تتحقق هذه العلاقة عندما يكون هناك علا قة خطية بين قيم اثنين او اكثر من المتغيرات التوضيحية فتكون نتيجة محدد المصفوفة تساوي صفر $|X'X| = 0$ وهذا يؤدي الى انتهاك شرط الرتبة Rank الذي يعرف بانه رتبة المصفوفة عندما المحدد لا يساوي صفر ، اذ تكون المصفوفة غير كاملة الرتبة اي ان Rank(X)<P الرتبة اقل من عدد المتغيرات وعليه فانه لا يمكن ايجاد معكوس مصفوفة المعلومات وبالتالي لا يمكن تقدير معلمات النموذج وهنا يظهر التعدد الخطي التام .

العلاقة الخطية الغير تامة (الجزئية) [5]

تظهر هذه الحالة عندما يكون محدد مصفوفة المعلومات لا يساوي صفر $|X'X| \approx 0$ وانما قريب منه.
وتميل المتغيرات للتحرك سوية بالزيادة او النقصان كما في حالة استخدام المتغيرات المرتدة ففي هذه الحالة يكون تقدير معلمات النموذج غير دقيقة وغير ممثلة لواقع المشكلة المدروسة لان تباين المعلمات سيكون كبيرا نتيجة لكبر حجم الاخطاء المعيارية وهذا يؤثر على ظهور بعض او جميع قيم إحصاءه الاختبار T-test لمعلمات النموذج صغيرة نسبيا وهذا يؤدي الى عدم معنوية هذه المقدرات وان كانت قيمة معامل التحديد R^2 للنموذج كبيرة .

الكشف عن وجود مشكلة التعدد الخطي [6]

1. مقياس تضخم تباينات المعاملات Variance Inflation Factor (VIF)

يعتبر مقياس تضخم التباين من الطرق الاساسية للكشف عن وجود مشكلة التعدد الخطي وهو يقيس مدى تضخم تباينات معاملات الانحدار المقدرة عند وجود ارتباط خطي بين المتغيرات التوضيحية ،تكون قيمته دائما اكبر او تساوي الواحد ويمكن ايجاده بالاعتماد على معامل التحديد كما بالصيغة التالية :

$$VIF=1/(1-R^2_j) \quad .....(1)$$

او من خلال البرنامج الجاهز تكون قيم معامل VIF اكبر او تساوي واحد فاذا كانت القيم اكبر من 5 دل على انه توجد مشكلة تعدد خطي .

رقم الحالة Condition Number (C.N)

هناك عدة صيغ لحساب رقم الحالة التي تشير إلى درجة التعدد الخطي وابرزها هي

$$C.N=\sqrt{\frac{\text{اكبر قيمة ذاتية}}{\text{اصغر اقل ذاتية}}} \quad .....(2)$$

اذ اقترح Johnston اذا كانت قيمة C.N تتراوح بين 20 الى 30 تكون مؤشر لوجود تعدد خطي مرتفع بينما اقترح Belsley انه اذا كانت ذا كانت قيمة C.N تتراوح بين 30 الى 100 يكون مؤشر الى وجود تعدد خطي مرتفع جدا .

2. مؤشر الحالة Condition Index (C.I)

يستخدم هذا المؤشر للكشف عن وجود التعدد الخطي ومن خلال القيم المميزة يتم حساب هذا المؤشر وفق الصيغة الاتية :

$$CI_j = \sqrt{\frac{\text{أكبر قيمة عينية}}{\text{القيمة العينية رقم}j}} \quad , \quad j = 1,2,\ldots..p \quad \ldots\ldots\ldots(3)$$

فاذا كانت قيمة C.I اكبر من 30 هذا دليل على وجود مشكلة تعدد خطي .

4.مصفوفة معاملات الارتباط Correlation Matrix

يعد اختبار مصفوفة معاملات ارتباط المتغيرات التوضيحية من ابسط طرق الكشف عن التعدد الخطي فاذا كان الارتباط بين المتغيرات التفسيرية X_1 , X_2عالي فان القيمة المطلقة ل$|r_{12}|$ تكون قريبة من الواحد مما يعرض النموذج للتعدد الخطي ولكن ضعف الارتباط بين المتغيرات لا يعني عدم وجود تعدد خطي .

5. اختبار فايرار وكلوبير Farrar& Glauber

وضع هذا الاختبار في عام (1967) ويستند بشكل اساسي على اختبار مربع كاي (χ^2)والصيغة الرياضية له :

$$\chi^2=-[n-1-1/6(2K+5)]\ln|M| \quad .....(4)$$

اذ ان
M: تمثل مصفوفة الارتباط
K:تمثل عدد المتغيرات
n:تمثل حجم العينة
وتقارن مع القيمة الجدولية ل χ^2 ونرفض الفرضية العدم اذا كانت القيمة المحسوبة اكبر من القيمة الجدولية.

معالجة مشكلة التعدد الخطي[3]

1. تضخيم البيانات (زيادة حجم العينة) يؤدي الى تقليل الارتباط بين المتغبرات المستقلة .
2. حذف بعض المتغيرات التي يزداد ارتباطها بالمتغيرات الاخرى(ارتباط عالي).
3. التحويل المعياري للمتغيرات التوضيحية.
4. تطبيق معلومات لدراسات سابقة تبين العلاقات الثابتة بين بعض المعلمات الانحدارية ويتم توضيح ذلك باستخدام دالة انتاج كوب_دوكلاص وكما يلي

$$Y= \alpha_0 + X^{\alpha 1}{}_1 X^{\alpha 2}{}_2 e + e \qquad \ldots\ldots (5)$$

وعندما $\alpha 1=\alpha 2$ تصبح المعادلة

$$Y=\alpha_0 + X^{\alpha 1}{}_1 X_2 \qquad \ldots\ldots (6)$$

وعنداخذ اللوغارتم ينتج

$$ln_y = ln\alpha_0 + \alpha_1 lnX_1 + 1 - \alpha \qquad \ldots\ldots (7)$$

$$ln_y = ln_{\alpha 0} + \alpha_1 lnX_2 - \alpha_1 lnX_2 + lne \qquad \ldots\ldots (8)$$

$$(\ln Y - \ln X_2) = ln\alpha_0 + \alpha_1 lnX_1 - \alpha_1 lnX_2 + lne \qquad \ldots\ldots (9)$$

$$(\ln Y - \ln X_2) = ln\alpha_0 + \alpha_1 (lnX_1 - lnX_2) + lne \qquad \ldots\ldots (10)$$

وعند ايجاد القيم المقدرة الى α فان $\hat{\alpha}_2$ تكون

$$\hat{\alpha}_2 = 1 - \hat{\alpha} \qquad \ldots\ldots (11)$$

5. استخدام طرائق التقدير وهي:
طريقة انحدار الحرف أ و طريقة المركبات الرئيسية أو طريقة المربعات الصغرى الجزئية

طرق معالجة التعدد الخطي:-

انحدارالمركبات الرئيسية (Principal Componets Regrsion)

1 -مقدرات المركبات الرئيسية [5,10] :

تعد مقدرات المركبات الاساسية من ابسط و اكثر طرق الانحدار التي تعالج مشكلة التعدد الخطي في البيانات واول من اقترح تقنية لتحليل المركبات الرئيسية او الوحدات الرئيسية هو كارل بيرسون (Pearson,1901 Karl) وحسب اعتقاده في ذلك الوقت انه هذا هو الحل الصحيح لمعظم المسائل ذات الاهمية الكبيرة لمن يعمل في مجال الاحصاءات الحيوية، علما انه لم يقوم باقتراح طريقة حساب لمتغيرين او ثلاث متغيرات.

اما هوتلنج (Hotelling,1933) فقد وصف طرق حساب عملية لاكثر من متغيرين او ثلاثة ومع ذلك فانه لم ينتشر استعمال هذه التقنية لحساب المتغيرات الا بعد توفر الحاسبات الالكترونية لان عملها يتطلب حسابات يدوية كثيرة ومملة جدا.

ومن الجدير بالذكر ان تحليل المكونات الرئيسية يهتم بشرح وتفسير هيكل تباينات وتغايرات المتغيرات الاصلية باستعمال توليفات خطية قليلة من هذه المتغيرات ويمكن توضيح ذلك من خلال ما يأتي (بفرض لدينا نموذج الانحدار الخطي العام):

$$Y = X\beta + \varepsilon \qquad \ldots\ldots (12)$$

اذ يتم قياس كل من (X,Y) حول متوسطاتها بحيث ان $(X'X)$ و $(X'Y)$ تمثل مصفوفات معاملات الارتباط .

ويعتبرهذا النموذج بدلالة المتغيرات المستقلة المرتبطة (غير المتعامدة) ولغرض تحويها الى متغيرات غير مرتبطة (P) ولتكن لدينا Vمصفوفة تعامدة وتحقق الشروط التالية :

$V'V=1$

$V'(X'X)V=Z$

اذ ان :

Z: هي مصفوفة قطرية للجذور المميزة لمصفوفة المعلومات $X'X$

V: مصفوفة متعامدة تمثل اعمدتها المتجهات المميزة المعدلة للمصفوفة $X'X$

وبالاعتماد على مصفوفة V يمكننا الحصول على مجموعة جديدة من المتغيرات التوضيحية تكون على شكل تراكيب خطية تسمى ب(المركبات الرئيسية) ويكون تمثيلها بالمعادلة الاتية :

$$P_j = \sum_{j=1}^{k} V_j X_k \quad \text{.....(13)}$$

وتكتب بصيغة المصفوفات كالاتي :

$$P=XV \quad \text{.....(14)}$$

اذ ان :

P: هي مصفوفة المركبات الرئيسية.

X: تمثل مصفوفة المتغيرات المستقلة

V: تمثل مصفوفة المتجهات المميزة المعدلة للمصفوفة $(X'X)$.

وبدلالة المتغيرات المستقلة المرتبطة تكون التراكيب الخطية على هيئة دوال خطية ويتم من خلالها الحصول على متغيرات جديدة مستقلة (غيرمرتبطة) يرمز لها ب P

لنفرض ان لدينا نموذج الانحدار الخطي العام:

$$Y=X\alpha+\mu \quad \text{.....(15)}$$

وبما ان: $P=XV$

وعليه يمكن كتابة النموذج بدلالة المركبات الرئيسية (P) اي بدلالة المتغيرات المستقلة المتعامدة الجديدة بالشكل الاتي :

$$Y=P\gamma+\varepsilon \quad \text{.....(16)}$$

يهدف تحليل المركبات الرئيسية الى اخذ q متغير X1,X2,.....Xq وايجاد تركيب من هذه المتغيرات لانتاج متغيرات غير مرتبطة هي Z1,Z2,.....Zq وان ضعف الارتباط يكون خاصية مفيدة جدا لان هذا يدل على ان المؤشرات تقيس ابعادا مختلفة للبيانات ولكن تكون هذه المؤشرات مرتبة بحيث ان Z1 تعرض اكثر كمية للتغير وتعرض Z_2 ثاني اكبر كمية للتغير .....وهكذا، اذ يرمز لتباين المتغير Zi ب Var(Zi) لمجموعة البيانات المدروسة، ويكون $Var(Z1) \geq Var(Z2) \geq \ldots\ldots\ldots \geq Var(Zq)$.

وضرورة التاكيد على ان تحليل المركبات الرئيسية لا يعمل دائما اي لا يمكن خفض عدد كبير من المتغيرات الاصلية الى عدد صغير من المتغيرات المحولة.

عندما تكون المتغيرات الاصلية غير مرتبطة فان تحليل المركبات لا يضيف شيئاً اي نحصل على افضل نتائج التحليل عندما يكون الارتباط عالي جدا سواء كان موجبا اوسالبا.

ومن الجدير بالذكر ان تحليل المكونات الرئيسية يهتم بشرح وتفسير هيكل تباينات وتغايرات المتغيرات الاصلية باستعمال توليفات خطية قليلة من هذه المتغيرات من خلال الحصول على نفس التباين الكلي يتطلب استعمال P من المكونات الرئيسية وللحصول على الجزء الاكبر من التباين الكلي فان استعمال عدد قليل من هذه المكونات يكون كافيا مثل K واذا تحقق ذلك فان المعلومات التي يمكن الحصول عليها من K من المكونات الرئيسية تكون مطابقة للمعلومات التي يتم الحصول عليها باستعمال المتغيرات الاصلية وعددها P وان (P>K) في هذه الحالة يمكن استعمال K من المكونات الرئيسية بدلا من المتغيرات الاصلية وبالتالي يتم تخفيض عدد البيانات الاصلية المتكونة من n من القياسات عن p من المتغيرات الى عدد بيانات يتكون من n من القياسات و k من المكونات الرئيسية.

يؤدي استعمال طريقة المكونات الرئيسية دائما الى الكشف عن علاقات لم تكتشف بعد ويمكن التعبير عن المكونات الرئيسية (جبريا) بانها توليفات خطية من المتغيرات العشوائية الاصلية X1,X2,.....Xp و(جبريا) تمثل هذه التوليفات الخطية نظام احداثيات جديدة يتم الحصول عليها بتدوير rotating محاور النظام الاصلي X1,X2,.....Xp اذ تمدنا المحاور الجديدة بوصف اكثر بساطة واختصارا لهياكل تشتت المتغيرات الاصلية كما تمدنا باكبر قدر من التشتت.

ويعد تحليل المكونات الرئيسية (PCA) هو احد التقنيات المنتشرة مع العديد من التطبيقات في الهندسة وعلم الاحياء والعلوم الاجتماعية وكذلك تشمل بعض الأمثلة المثيرة للاهتمام كتصنيف الرمز البريدي المكتوب بخط اليد والتعرف على الوجوه البشرية.

وقد تم استخدام ال (PCA) مؤخرا في تحليل بيانات التعبير الجيني ما يسمى تقنيات الحلاقة الجينية باستعمال (PCA)لتجميع جينات متغيرة للغاية ومترابطة في مجموعات البيانات المجهرية.

ويمكن تحليل المركبات الرئيسية بالخطوات الاتية:
1. نبدأ بتحويل المتغيرات $X_1, X_2, \ldots\ldots\ldots\ldots, X_P$ لتكون اوساطها صفرية وتبايناتها متساوية.
2. يتم حساب مصفوفة الارتباط بعد الانتهاء من اجراء الخطوة السابقة.
3. ايجاد القيم المميزة $\lambda_1, \lambda_2, \ldots\ldots\ldots, \lambda_P$ والمتجهات المميزة المناظرة $a_1, a_2, \ldots\ldots a_p$، اذ تكون معادلات المركبة الرئيسية i تعطى بالمتجه a_i وتبايناتها ب λ_i
4. اهمال المركبة التي لا تأخذ في الحسبان الا جزء يسير من التشتت في البيانات اي تشتتها قليل .

2 - أهمية المركبات الرئيسية [2]:

للمركبات الرئيسية وتحليلها فوائد مختلفة فقد يمكن من خلالها تخفيض أوتقليص عدد المتغيرات الكبير من خلال قياس الاهمية لكل متغير إلى عدد أقل دون فقدان كمية من المعلومات ،هذا الهدف يمكن الحصول عليه في تحليل المكونات الأساسية، وكذلك تسهم في حل مشكلة تعدد العلاقة الخطية (Multicolinearity) بين المتغيرات التفسيرية.

ويستخدم تحليل المركبات الرئيسية في تحليل الانحدار كوسيلة لإيجاد معادلة الانحدار المناسبة وخاصة إذا كانت المتغيرات الأصلية تعاني من مشكلة تعدد العلاقة الخطية في هذه الحالة يكون الانحدار لمتغير الاستجابة y على المكونات الأساسية PCj أي (Y / PC)بدلاً من (Y / X)، ويمكن للباحث اختبار إذا كانت بياناته تتوزع توزيعاً طبيعياً، استُدِلّ على أن متغيراته الأصلية ذات توزيع طبيعي وبالعكس ،في هذه الحالة وبدلاً من اختياره المتغيرات كلها يلجأ إلى المركبات الرئيسية المستنتجة من تلك المتغيرات فيختبرها.

3 - خواص المركبات الرئيسية: [3,7]

1. إن جميع الجذور المميزة لمصفوفة الارتباط (R) هي قيم موجبة وذلك لان هذه المصفوفات موجبة التعريف (positive definite)
2. إن مجموع الجذور المميزة يساوي مجموع العناصر القطرية للمصفوفة الستخدمة أي أن :

$$\sum_{j=1}^{p} \lambda_j = P \sum_{j=1}^{p} V_j \qquad \ldots\ldots\ldots (17)$$

3. محدد المصفوفة المستخدمة يساوي

$$|V| = |R| = \lambda_1, \lambda_2, \ldots\ldots\ldots\ldots, \lambda_P \qquad \ldots\ldots (18)$$

4. المتجهات المميزة aj متعامدة فيما بينها أي أن:

$$a_j a_j = \begin{cases} 1 if j = j \\ 0 if j \neq j \end{cases}$$

5. $\lambda = Ava$

اذ ان a هو المتجه المميز للجذر المميز λ

6. التغاير المشترك بين أي مكونيين رئيسين يساوي صفر أي أن:

$Cov(PC_j, PC_j) = 0; j \neq j$

7. التغاير بين أي مكون رئيسي وبين المتغيرات X هو

$cov(X, pc_j) = cov(X, a_j) = v\, a_j$

8. تحويل المتغيرات الاصلية الي المكونات الرئيسية $\frac{\lambda_i}{\sum \lambda_i}$

4 -كيفية ايجاد مقدرات المركبات الرئيسية: [5]

يمكن ايجاد مقدرات المركبات الرئيسية لموجه المعالم (α) لنموذج الخطي العام الذي يعاني من مشكلة تعدد خطي نقوم باجراء الخطوات الاتية:

-1 حذف واحد او اكثر من المركبات الرئيسية (aj).
-2 نطبق طريقة المربعات الصغرى (OLS) على النموذج الجديد بعد الحذف.
-3 اجراء التحويل الخلفي الى مجال المعلمة الاصلي.

لغرض توضيح الخطوات اعلاه نفرض ان المصفوفة (X´X)هي (r) اذ ان اخر (k-r)من عناصر المصفوفة القطرية مساوية الى الصفر او قريبة منه اذا كانت المصفوفة (X´X) قريبة من الاحادية لذا يتم تجزئة المصفوفة المتعامدة (V) والمصفوفة (Z) الى الشكل الاتي:

$$V=(Vr \quad Vk-r) \qquad .....(19)$$

$$Z=\begin{bmatrix} Zr & 0 \\ 0 & Zk-r \end{bmatrix} \qquad .....(20)$$

اذ ان :

Z_r: مصفوفة (r*r)

Z_{k-r} : مصفوفة (k-r)*(k-r)

فتكون مقدرات المركبات الرئيسية باستعمال طريقة (OLS) كالاتي:

$$\mu'\mu = (Y-P\gamma)'(Y-P\gamma) \qquad .....(21)$$

$$=Y'Y-Y'P\gamma-\gamma'P'Y+\gamma'P'P\gamma \qquad .....(22)$$

$$=Y'Y-2\gamma'P'Y+\gamma'P'P\gamma \qquad .....(23)$$

$$\frac{\partial\mu'\mu}{\partial\gamma'}=-2P'Y+2P'P\hat{Y}=0 \qquad .....(24)$$

$$\hat{Y}=(PP)^{-1}P'Y \qquad .....(25)$$

بما ان :

$$P=XV \qquad .....(26)$$

$$\therefore \hat{Y}=Z_r{}^{-1}V_rXY \qquad .....(27)$$

بتعويض عن المصفوفتين (Z وV) من المعادلة 36 و37 نحصل على :

$$\begin{bmatrix} \hat{\gamma}_r \\ \hat{\gamma}_{k-r} \end{bmatrix}=\begin{bmatrix} Z_r & 0 \\ 0 & Z_{k-r} \end{bmatrix}^{-1}\begin{bmatrix} V_r \\ V_{k-r} \end{bmatrix}X'Y \qquad .....(28)$$

وبافتراض $(Z^{-1}{}_{k-r})$ تكون مساوية للصفر نكتب مقدرات المربعات الصغرى (OLS) الى (γ_r)بالشكل الاتي:

$$\hat{\gamma}_r=Z^{-1}{}_rV_rX'Y \qquad .....(29)$$

وبما انه في الجانب التقديري

$$\hat{\gamma}_r=V'{}_r\hat{\beta} \qquad .....(30)$$

ويمكن القول ان مقدرات المركبات الرئيسية تعطى بالشكل الاتي:

$$\hat{\alpha}_{pc}=V_r\hat{\gamma}_r \qquad .....(31)$$

وبتعويض المعادلتين (2-45) و (2-47) نحصل على :

$$\hat{\alpha}_{pc}=V_rZ^{-1}{}_rV_rX'Y \qquad .....(32)$$

تمثل المعادلة السابقة مقدرات متجه المعالم باستعمال طريقة المركبات الرئيسية ويمكن كتابتها بالشكل الاتي:

$$\hat{\alpha}_{pc}=\sum_{i=1}^{k}\sigma^{-1}{}_rV_i'X'YV_i-\sum_{i=r+1}^{k}\sigma^{-1}{}_rV'{}_iX'YV_i \qquad .....(33)$$

$$\hat{\alpha}_{PC}=\hat{\alpha}_{LS}-\sum_{i=r+1}^{k}\sigma_r{}^{-1}V_i'X'YV_i$$

إن المكون الأول يفسر أقصى ما يمكن من التباين بين المتغيرات الأصلية، وثاني مكون (لا يرتبط بالمكون الأول) يفسر أعلى قدر للتباين المتبقي... وهكذا حتى يتم تفسير كل التباين. وان تباين كل المكونات مساوي إلى مجموع تباين المتغيرات الأصلية، ويمكن حساب المكونات بطريقتين:

- استعمال مصفوفة التباين المشترك لمتغيرات الاستجابة وفي هذه الحالة فان المتغيرات تكون مقاسه بالانحرافات عن الوسط الحسابي.
- استعمال مصفوفة الارتباطات لمتغيرات الاستجابة وفي هذه الحالة تستعمل المتغيرات المعيارية ويكون ذلك ضروريا في حالة اختلاف وحدات القياس لمتغيرات الاستجابة.

5 -اختيار عدد المركبات الرئيسية [10]

توجد اساليب مختلفة لتحديد واختيار عدد المركبات الرئيسية الداخلة في التحليل ومن هذه الاساليب اقتراح بعض المعايير منها :

اسلوب kaiser

اقترح هذا المعيار من قبل (**Guttman**) وقد تم تطويره من قبل (**kaiser**) وان تطبيق هذا المعيار بسيط للغاية ويعتبر ان المركبات الرئيسية (P.C) عندما تكون جذورها المميزة اكبر من الواحد الصحيح تبقى في التحليل اي :

اذا كان($\lambda_J > 1$) فان P_J يبقى في التحليل وكذلك يفضل استعمال هذا المعيار عندما يكون عدد المتغيرات التوضيحية (X) بين (20-50).

التطبيق العملي :

وصف البيانات:-

تم جمع عينة عشوائية بسيطة حجمها (100) مريض من المصابين بالجلطة الدماغية والقلبية الراقدين في مستشفى الحسين في محافظة كربلاء المقدسة خلال فترة ثلاثة اشهر واعتمدت الدراسة في نتائجها على البرامج الاحصائية (SPSS version25 , NCSS) .

وشملت الدراسة (23) متغيرا، رمزنا للمتغير المعتمد ب Y والمتغيرات المستقلة بـ (X_i , i=1,2,….,. 22) والجدول رقم (3-1) ادناه يوضح المتغيرات على التوالي:

متغيرات البحث :-

ان المتغير التابع هو من المستوى الفئوي (مقياس الفترة او مقياس المسافة) وهو عدد يدل على كم له مسافات موزونة وهنا يمثل مكان حصول الخثرة الدموية ،فمن المعلوم ان درجة خطورة حصول الخثرة في القلب تختلف عن الدماغ او القدم او أي عضو اخر والمتغيرات المستقلة يمكن توضيحها بالجدول الاتي:

جدول (1) يبين رمز كل متغير من المتغيرات المستقلة

المتغيرات	الرمز	المتغيرات	الرمز
عامل الوراثة	X13	الجنس	X1
درجة القرابة	X14	العمر	X2
نوع العمل	X15	انواع علاج الضغط	X3
طبيعة مزاج المريض	X16	وجود الضغط	X4
المستوى التعليمي	X17	انواع علاج السكر	X5
عدد ساعات نوم الليل	X18	وجود السكر	X6
عدد ساعات نوم النهار	X19	تكرار حصول الجلطة	X7
تناول الادوية	X20	وقت حصول الجلطة	X8
ممارسة الرياضة	X21	وقت مراجعة الطبيب	X9
وجود امراض اخرى	X22	حصول الخثرة	X10
اين وجود الخثرة	Y	ساعة الاستيقاظ صباحا	X11
		نوع الجلطة	X12

قبل اختبار البيانات يتطلب معرفة معنوية النموذج ومعامل التحديد للنموذج كما في الجدول الاتي

جدول (2) يبين تحليل التباين

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Model	22	82.1133	3.732423	56.0588	0.000000
Error	77	5.126696	0.06658046		
Total(Adjusted)	99	87.24	0.8812121		

نلاحظ ان قيمة ال sig هي (0.00) اقل من مستوى الدلالة الاحصائية (0.05) وهذا يعني ان النموذج معنوي وان قيمة معامل التحديد R^2 مساوية الى 0.9412 و الانحراف المعياري 0.2580.

الكشف عن وجود مشكلةالتعدد الخطي

تم الكشف عن وجود التعدد الخطي بين المتغيرات التوضيحية للمصابين بالجلطة القلبية والجلطة الدماغية بطرائق عدة، وذلك باستخدام البرنامج الجاهز SPSS ، اما الطرائق المعتمدة في الكشف عن وجود مشكلة تعدد العلاقة الخطية هي :

مقياس تضخم تباين المعاملات (VIF)

يعد هذا المقياس كما ذكر من اهم مقايس الكشف عن وجود مشكلة التعدد الخطي اذ يمكن قياسهما بطريقتين هما:
الطريقة الاولى :بالاعتماد على قيمة معامل التحديد كما في الصيغة (1)

VIF=1/(1-0.9 6)=25

وبما ان قيمة VIFكبيرة تدل على وجود مشكلة التعدد الخطي.
والطريقة الثانية : من خلال قيم ال VIF الجاهزة بالبرنامج اي انه نلاحظ وجود معاملات ال VIFمعامل تضخيم التباين اكبر من (5) فانه دليل على ان المتغير X_j يعاني من تضخم تباين معامله اي وجود مشكلة تعدد العلاقة الخطية كما موضح في الجدول :

جدول (3) يبين قيم VIF لكل متغير

الرمز	Inflation	الرمز	Inflation
X1	1.7622	X12	1.3157
X2	1.5423	X13	6.3688
X3	6.9525	X14	6.0494
X4	6.8970	X15	2.0457
X5	10.0161	X16	8.8192
X6	9.7689	X17	2.0136
X7	1.3922	X18	1.5452
X8	2.9609	X19	1.4149
X9	3.1193	X20	1.5325
X10	8.3605	X21	1.4791
X11	1.3451	X22	1.3835

* تعني ان هذه المتغيرات تعاني من تضخم تباين معاملاتها.
وبما ان قيم تضخم تباين المعاملات VIF في الجدول (3) للمتغيرات التوضيحية لكل من وجود الضغط وانواع علاج الضغط وانواع علاج السكر و وجود السكر و حصول الخثرة وعامل الوراثة و درجة القرابة وطبيعة مزاج المريض اكبر من (5) فان هذا يعني ان هذه المتغيرات تعاني من مشكلة تعدد العلاقة الخطية بين المتغيرات التوضيحية .

رقم الحالة *Condition Number (C.N)*)

يمكن حساب رقم الحالة بالاعتماد على القيم الذاتية في الجدول اعلاه وفق الصيغة (2) التي ذكرت سابقا.

مؤشر الحالة (C.I) Condition Index

$$CN=\sqrt{\frac{18.522}{0.00000124}} = 3864.85$$

نلاحظ بما ان قيمة CN اكبر بكثير من 1000 هذا يدل على وجود تعدد خطي كبير وخطير

يعد مؤشر الحالة من المؤشرات التي تبرز ظاهرة مشكلة التعدد الخطي وبالاعتماد على القيم الذاتية في الجدول رقم (2-5) ويكون حسابه وفق الصيغة (6) :

$$C.I_1 = \sqrt{\frac{18.522}{18.522}} = 1 \ ,$$

$$C.I_2 = \sqrt{\frac{18.522}{.942}} = 4.433,$$

جدول (5) يبين قيم مؤشر الحالة

القيمة	مؤشر الحالة	القيمة	مؤشر الحالة
λ 1	1.000	**λ13**	17.718
2λ	4.433	**λ14**	19.247
3λ	4.791	**λ15**	21.518
4λ	6.105	**λ16**	22.374
5λ	6.528	**λ17**	23.340
6λ	7.327	**λ18**	26.191
7λ	7.447	**λ19**	29.016
8λ	8.731	**λ20**	41.034
9λ	9.013	**λ21**	55.560
10λ	11.225	**λ22**	60.864
11λ	11.846	**λ 23**	96.234
12λ	16.093		

نجد أن الجذور المميزة (من الجذر المميز رقم 1 إلى الجذر المميز رقم 19) لم تظهر قيمة لدليل الحالة المقابلة لتلك الجذور اكبر من 30 ، في حين أن قيم دليل الحالة من 20 الى 23 قد تجاوزت 30 مما يدل على أن البيانات للظاهرة المدروسة تعاني من مشكلة التعدد الخطي أي توجد علاقة خطية تربط بين المتغيرات التوضيحية

اختبار فراير وكلوبير

ان هذا الاختبار يستخدم لاختبار الفرضية الآتية :

عدم وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية : H_0

وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية : H_1

ومن الصيغة (4) فان قيمة مربع كاي χ^2_{cal} المحسوبة مساوية الى (951.9098) ومن جداول توزيع مربع كاي فان القيمة الجدولية لـ مربع كاي χ^2_{tab} بدرجة حرية $k(k-1)/2$ هي (231) عند مستوى معنوية 0.01 تكون مساوية الى (183.955) عندما n=100,k=23 وبمقارنة χ^2_{cal} مع قيمة χ^2_{tab} نلاحظ ان $\chi^2_{tab} < \chi^2_{cal}$ لذا ترفض فرضية العدم وتقبل الفرضية البديلة التي تدل على وجود مشكلة التعدد الخطي بين المتغيرات التوضيحية .

معالجة مشكلة التعدد الخطي باستعمال طريقة انحدار المركبات الرئيسية :-

تم اعتمادا على البرنامج الإحصائي NCSS في تطبيق انحدار المركبات الرئيسية ومع خلال الصيغ التي ذكرت في الجانب النظري تم ايجاد الجذور المميزة والمتجهات المميزة ولغرض تفسير الجذور المميزة الاكبر من الواحد طبق معيار Kaiser الذي يشترط ان تكون عدد المتغيرات التوضيحية ضمن المدى (20-50) .

وعدد متغيراتنا 22 وهي ضمن العدد المسموح وتبين ان قيم الجذور المميزة من λ1 و لغاية λ9 اكبر من واحد ، و هي التي ستبقى في التحليل ، وتهمل البقية والجدول التالي يوضح ذلك .

جدول (6) يبين نسبة تباين المركبات ونسبة تجمع التباين

المركبات	الجذور المميزة	نسبة التباين	تجمع نسبة التباين
1	**2.848212**	**12.95**	**12.95**
2	**2.544829**	**11.57**	**24.51**
3	**2.288195**	**10.40**	**34.91**
4	**2.035193**	**9.25**	**44.17**
5	**1.853707**	**8.43**	**52.59**
6	**1.670356**	**7.59**	**60.18**
7	**1.461773**	**6.64**	**66.83**
8	**1.195671**	**5.43**	**72.26**
9	**1.000478**	**4.55**	**76.81**
10	0.812797	3.69	80.51
11	0.737635	3.35	83.86
12	0.691872	3.14	87.00
13	0.618814	2.81	89.82
14	0.509397	2.32	92.13
15	0.485350	2.21	94.34
16	0.395833	1.80	96.14
17	0.310838	1.41	97.55
18	0.222175	1.01	98.56
19	0.135987	0.62	99.18
20	0.087415	0.40	99.58
21	0.054618	0.25	99.82
22	0.038855	0.18	100.00

نلاحظ من خلال الجدول اعلاه ان الجذر المميز الاول يفسر (12.95) من التباين الكلي ثم يتناقص ليكون (11.57) للجذر الثاني وهكذا . وباخذ المقدار التراكمي لما تفسره المكونات الاساسية نجد انه الى حد الجذر المميز التاسع ويكون لدينا تقريبا 77% من تباين العينة قد تم تفسيره من قبل الجذور المميزة التسعة من اصل ثلاثة وعشرون جذرا مميزا .في حين ان 13 جذرا مميز الاخيرة لا يفسر اكثر من (25%) من التباين الكلي وعليه يمكن الاعتماد على هذه النتيجة باخذ المكونات الاساسية التسعة الاولى .

عندما يكون الجذر اكبر من واحد حسب معيار كيسير تحديد المركبات الرئيسية الداخلة في التحليل كما في الجدول الاتي:

جدول (7) يبين المركبات الرئيسية الداخلة في التحديد

المعاملات / المركبات	C1	C2	C3	C4	C5	C6	C7	C8
1	-0.0565	-0.0477	0.0525	-0.0494	0.0538	-0.0593	-0.0136	-0.0175
2	-0.0733	-0.0664	0.0688	0.0777	0.0573	0.0505	0.0773	0.0190
3	-0.0879	-0.0689	-0.0895	0.0831	-0.0645	0.0608	0.0660	-0.0733
4	0.0307	-0.0747	-0.0661	0.0637	0.0354	-0.0335	0.0639	-0.1506
5	0.0310	-0.0513	-0.0811	0.0781	0.0463	-0.0448	0.0595	-0.1135
6	-0.0037	-0.0114	-0.0933	0.0924	0.0400	-0.0334	0.0725	-0.1227
7	0.0201	-0.0340	-0.0319	0.0503	0.0049	0.0145	0.0061	-0.0711
8	0.0103	-0.0293	-0.0201	0.0398	0.0065	0.0110	0.0260	-0.0667
9	0.0100	-0.0262	-0.0165	0.0323	0.0014	0.0148	0.0302	-0.0638

تابع الى جدول (7)

المعاملات / المركبات	C9	C10	C11	C12	C13	C14	C15	C16
1	-0.0128	-0.0355	-0.0020	-0.0080	0.0064	-0.0073	-0.0642	-0.0435
2	0.0440	-0.1401	-0.0007	0.0515	-0.1015	0.0917	-0.0574	-0.1580
3	-0.0543	-0.2333	-0.0736	0.0506	-0.0159	0.0037	-0.0725	-0.2423
4	-0.1171	-0.3400	-0.0320	0.1357	0.0065	-0.0394	0.0394	-0.3416
5	-0.0820	-0.3475	-0.0443	0.1378	0.0472	-0.0787	0.0313	-0.3505
6	-0.1006	-0.3712	-0.0089	0.1235	0.0433	-0.0771	-0.0067	-0.3663
7	-0.0491	-0.4070	-0.0747	0.1594	0.0198	-0.0453	0.0166	-0.3971
8	-0.0481	-0.4074	-0.0812	0.1803	0.0209	-0.0460	0.0100	-0.3976
9	-0.0459	-0.4078	-0.0719	0.1756	0.0214	-0.0470	0.0110	-0.3989

تابع الى جدول (7)

المعاملات / المركبات	Cl 17	Cl 18	C19	C20	C21	C22
1	0.0766	-0.0090	-0.0037	0.0229	-0.0571	-0.0536
2	0.1011	-0.0666	-0.0075	0.0965	-0.1183	-0.0498
3	0.1066	-0.1118	-0.0614	0.1139	-0.1259	-0.0382
4	0.0069	-0.0341	0.0045	0.1529	-0.0751	-0.0584
5	0.0085	-0.0335	0.0102	0.1576	-0.0756	-0.0669
6	0.0341	0.0283	0.0702	0.1181	-0.0622	-0.0407
7	0.0345	0.0639	0.1238	0.0877	-0.1096	-0.0076
8	0.0330	0.0612	0.1156	0.0894	-0.0971	0.0108
9	0.0296	0.0583	0.1208	0.1003	-0.0996	0.0103

جدول (8) يبين قيم VIFلكل مركبة حسب طريقة انحدار المركبات الرئيسية

رمز المتغيرات المستقلة	VIF	رمز المتغيرات المستقلة	VIF
X1	0.3811	**X12**	1.0522
X2	0.8882	**X13**	0.2723
X3	0.3289	**X14**	0.3111
X4	0.3689	**X15**	0.2943
X5	0.3262	**X16**	0.2758
X6	0.3251	**X17**	0.2774
X7	0.6466	**X18**	0.3849
X8	0.2927	**X19**	0.4714
X9	0.2813	**X20**	0.8720
X10	0.2830	**X21**	0.9185
X11	0.6987	**X22**	1.0273

نلاحظ من خلال الجدول (8) ان جميع قيم VIF هي اقل من 5 وهذا يعني اختفاء مشكلة التعدد الخطي

الاستنتاجات :-

1. أستنتج ان مشكلة التعدد الخطي تؤثر على معنوية بعض المتغيرات المستقلة و تظهر عدم معنويتها رغم اهميته.

2. هناك عدة اختبارات ومعايير للكشف عن وجود مشكلة التعدد الخطي بين المتغيرات التفسيرية ومعالجتها بعدة طرق وان طريقة المركبات الرئيسية هي الطريقة الابسط تطبيقا والاكثر استيعابا .

التوصيات :-

1- إعطاء موضوع أمراض الجلطة إهتماماً أكبر ، لكثرة انتشاره بين افراد المجتمع ، بحيث تكون هناك مستشفيات خاصة بذلك وتوفير الكوادر الطبية اللازمة .

2- يوصى بالمراجعة الدورية الى المستشفيات للفحص والتأكد من سلامة أبداننا ،وأخذ الحيطة والحذر عند ظهور أحد الاعراض التي ذكرت .

3- يمكن جعل هذه الدراسة كأساس لدراسات مستقبلية موسعة وذلك في حالة وجود ارتباط ذاتي بين الأخطاء العشوائية او في حالة ظهور عدم تجانس التباين بين المتغيرات التوضيحية بالإضافة الى المشكلة المدروسة في هذا البحث

4- ضرورة استعمال الاختبارات الاحصائية للكشف عن وجود مشكلة التعدد الخطي .

المصادر :-

1. ابو حامد ،سمير "الجلطة الدماغية " الطبعة الاولى ،دمشق خطوط النشر والتوزيع ،2009.
2. الجراح ،ريم علي ،"تحليل المكونات الاساسية باستخدام الشبكات العصبية والاصطناعية مع التطبيق "رسالة ماجستير مقدمة الى مجلس كلية علوم الحاسبات والرياضيات ،جامعة الموصل 2003.
3. الراوي ،خاشع محمود ،"المدخل الى تحليل الانحدار " مديرية دار الكتب للطباعة والنشر ، جامعة الموصل العراق ،1987.
4. السيفو ،وليد اسماعيل ،"المدخل الى الاقتصاد القياسي "وزارة التعليم العالي والبحث العلمي ،1988.
5. العزاوي ،دجلة ابراهيم مهدي مع نذير عباس ابراعيم "الاقتصاد القياسي اسلوب كمي باستعمال SPSS Minitab,Eviews ، الطبعة الاولى 2014.
6. القصيمي،عزة مصطفى عبد القادر" استخدام اسلوب المحاكاة في مقارنة مقدرات انحدار الحرف " رسالة ماجستير ،كلية علوم حاسبات ورياضيات جامعة الموصل ،2000
7. حمزة ، حمزة ابراهيم ،"تقدير وتحليل دوال الاقتصاد السوداني باستخدام المكونات الرئيسية " بحث مقدم الى جامعة السودان 2006.

8- Alin, A., Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics,. **2**(3): p. 370-374(2010).
9-Bryan F.J.Manly "Multivariat Statistical Methods Aprimer "fourth Edition,University of otago,(2012).
10-Ramzan,S., Khan,M.I"Dimension Reduction and reduced of multicollinearity using latent variable regression methods "world applied sciences Journal (814):404-410., (2010).
11- Jacquelynne R.King and Donald A.Jackson, " variable selection in large environmental data sets using principal component analysis ",www.zoo.utoronto.ca/jackson /king%20&%20jacson.pdf, (1999).
12 Ying Li, "A Comparison Study of Principle Component Regression, Partial Least Squares Regression and Ridge Regression with Application to Ftir Data" University, Sweden (2010).