

التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي (دراسة تطبيقية على مرض فقر الدم)

أ.م.د. رباب عبد الرضا صالح البكري/جامعة بغداد /كلية الإدارة والاقتصاد
الباحث /محمد شاكر محمود العزي/ جامعة بغداد /كلية الإدارة والاقتصاد

تاريخ التقديم: 2017/1/15
تاريخ القبول: 2017/1/26

المستخلص

تعد طريقة الانحدار اللوجستي الثنائي **regression logistic Binary** والذالة المميزة الخطية **Linear discriminant function** من اهم الطرائق الاحصائية المستخدمة في التصنيف والتنبؤ، عندما تكون البيانات من النوع الثنائي (0,1) فانه لا يمكن استخدام الانحدار الاعتيادي فلذلك نلجأ الى الانحدار اللوجستي الثنائي والذالة المميزة الخطية في حالة وجود مجموعتين، وفي حالة وجود مشكلة التعدد الخطي **Multicollinearity** بين البيانات (ان البيانات يوجد فيها ارتباطات عالية بين المتغيرات) اصبح عدم الامكان في استخدام الانحدار اللوجستي والذالة المميزة الخطية، ولحل هذه المشكلة نلجأ الى طريقة انحدار المربعات الصغرى الجزئية **Partial least square regression** لحل مشكلة التعدد الخطي.

وقد جرى في هذه البحث المقارنة بين الانحدار اللوجستي الثنائي **regression logistic binary** والذالة المميزة الخطية **linear discriminant function** عن طريق خطأ التصنيف. حيث تم جمع بيانات عن مرض فقر الدم بمتغيرين هما فقر الدم الحاد بالرمز (0)، وفقر الدم المزمن بالرمز (1) وبعده متغيرات حول المرض. جمعت البيانات من عدة مستشفيات عراقية، وجمعت عينة من المرضى الراقدين في المستشفى وحالات سابقة رقدت في المستشفى بعينة قدرها (140) مريضاً مصاباً بهذا المرض. وعند اختبار البيانات الصغرى الجزئية **Partial least square**.

وتوصل البحث الى ان الذالة المميزة الخطية **linear discriminant function** هي أفضل في تصنيف البيانات من الانحدار اللوجستي الثنائي **binary logistic regression**، اذ صنفت الذالة المميزة البيانات بشكل صحيح وأكثر دقة من الانحدار اللوجستي الثنائي.

المصطلحات الرئيسية للبحث/ الذالة المميزة الخطية- الانحدار اللوجستي الثنائي- المربعات الصغرى الجزئية - مشكلة التعدد الخطي - نسبة التصنيف.



مجلة العلوم

الاقتصادية والإدارية
العدد 99 المجلد 23
الصفحات 397-373

البحث مستل من رسالة ماجستير



1- المقدمة Introduction

في اي بلد في العالم توجد هناك الكثير من الامراض والعراق واحد من البلدان الموجودة في العالم التي طالما تعاني الكثير من الامراض المتنوعة التي تصيب الانسان سواءً كانت مزمنة او مكتسبة او متوارثة عن طريق الاجيال. ان فقر الدم واحد من كثير من الامراض التي تعاني منها المجتمعات في العالم ولطالما اخذ الكثير من الباحثين دراسات حول اسباب هذا المرض والوقاية منه. ففي موضوع البحث سيتم اخذ المرض من اتجاهين الاتجاه الاول (فقر دم حاد) والاتجاه الثاني (فقر دم مزمن). وعندما يكون المتغير المعتمد من النوع الثنائي فأننا نلجأ الى الانحدار اللوجستي سيتم اعتماد طريقتين للعمل على هذا المرض وهما طريقة الانحدار اللوجستي الثنائي وطريقة الدالة المميزة الخطية باستعمال المربعات الصغرى الجزئية.

ان طريقة الانحدار اللوجستي من الطرائق المهمة التي تدخل في تحليل البيانات التي يكون فيها المتغير المعتمد (Y) يكون فيه البيانات ثنائية (0,1)، ويكون الهدف الاساس من هذه الطريقة هو ايجاد أفضل نموذج يصف الحالة بين المتغير المعتمد والمتغير التوضيحي (المتغيرات التوضيحية). وان طريقة التحليل المميز هي من التحليلات الاحصائية التي تهتم بمتعدد المتغيرات حيث يتم بهذا التحليل استعمال مجموعة من التغيرات للتمييز بين مجموعتين او أكثر بواسطة عدة دوال تمييزية.

2- مشكلة البحث problem of research

لغرض توظيف بيانات فقر الدم ثنائي الاستجابة والتي تعاني من وجود مشكلة التعدد الخطي من اجل تكوين نموذج احتمالي للمريض بفقر الدم والذي يتم على اساس تصنيف او تحديد المريض بفقر الدم (الحاد او المزمن) ثم استعمال اسلوبين من اساليب التصنيف لغرض ايجاد أفضل نموذج احتمالي بأقل خطأ تصنيف ممكن.

3- هدف البحث objective of research

يتضمن هذا البحث المقارنة بين اسلوبين من اساليب التصنيف وهما الانحدار اللوجستي الثنائي والدالة المميزة الخطية بوجود مشكلة التعدد الخطي ولمعرفة مدى قابليتهم لتصنيف البيانات بأقل احتمال خطأ للتصنيف وذلك باستعمال بيانات حقيقية حول مرض فقر الدم (فقر الدم الحاد، فقر الدم المزمن) بوجود مشكلة التعدد الخطي بعد معالجتها بطريقة المربعات الصغرى الجزئية ومن ثم تطبيق اسلوبي الانحدار اللوجستي الثنائي الاستجابة والدالة المميزة الخطية.

4- الجانب النظري

من الأساليب الإحصائية المهمة في متعدد المتغيرات والتي تستعمل في تحليل وتقويم العلاقات بين مجموعة من المتغيرات يكون فيه المتغير التابع (متقطع)، ليس دائماً يكون فيها المتغير التابع مستمراً وذلك لغرض الحصول على انموذج رياضي يوضح العلاقة بين مثل هكذا بيانات تستعمل اسلوبي التحليل المميز والانحدار اللوجستي. وان هناك بعض الافتراضات الخاصة للأسلوبين المذكورين آنفاً وان تكون المتغيرات التوضيحية مستقلة ولا يوجد أي ارتباط بينها ولمعالجة مثل هكذا حالات تستعمل طرائق عدة منها طريقة المركبات الرئيسية وطريقة المربعات الصغرى الجزئية وطريقة الحرف وسوف نركز على طريقة المربعات الصغرى الجزئية.

1-4 المربعات الصغرى الجزئية⁽¹⁾⁽⁵⁾⁽⁶⁾ Partial least square

تعد طريقة المربعات الصغرى الجزئية أكثر الطرائق أهمية في الانحدار فهي تستعمل لتقليص عدد المتغيرات التوضيحية المرتبطة في الانموذج الى مركبات غير مرتبطة (خطية، متعامدة)، او عندما يكون عدد المتغيرات التوضيحية أكثر من عدد المشاهدات في التجربة. وطريقة المربعات الصغرى الجزئية مشابهة لطريقة المركبات الرئيسية وأسلوب انحدار الحرف لمعالجة مشكلة التعدد الخطي ولكنها تختلف في الحسابات فحوارزمية (Partial least square) تأخذ بالحسبان التباين المشترك ما بين متغير (متغيرات) الاستجابة والمتغيرات التوضيحية، اما طريقة المركبات الرئيسية (PCA) تأخذ بنظر الاعتبار التباين بين المتغيرات التوضيحية فقط. يقوم بتحويل المتغيرات التوضيحية المرتبطة الى مركبات رئيسية تختلف في الحسابات فحوارزمية الحل بطريقة PLS طريقة تكرارية عند استعمالها تنتج سلسلة من النماذج ويتوقف الحل التكراري عندما نصل الى العدد الكلي من المركبات في الانموذج او عندما تكون البواقي مساوية للصفر.



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

ففي حال تساوي عدد المركبات مع عدد المتغيرات التوضيحية فإن النتائج ستكون متطابقة مع طريقة المربعات الصغرى. ولأجل تحديد عدد المركبات التي تصغر من خطأ التنبؤ نستعمل طريقة العبور الشرعي Cross-validation وبعد تحديد عدد المركبات تقدر معالم النموذج الانحدار لكل متغير. وان طريقة PLS اول من طبقها الباحث Wold عندما يكون هناك ارتباط عالي ما بين المتغيرات التوضيحية او عندما يكون عدد المتغيرات التوضيحية تفوق عدد المشاهدات. وتوجد عدة خوارزميات وأكثرها تداولاً خوارزمية NIPALS عام 1973 للباحث Wold وخوارزمية SIMPLS 1993 المنسوبة للعالم De Jong وخوارزمية KERNEL المنسوبة للعالم DAYAL وخوارزمية orthogonal projection to latent structures عام 2002 المنسوبة الى TRYGG وكلها الأسلوب نفسه (الهدف) ولكن الاختلاف يكون في حساب الخوارزمية، خوارزمية NIPALS او تسمى (PLS1, PLS2) حيث ان خوارزمية PLS1 تستعمل عندما يكون متغير الاستجابة متجه اما خوارزمية PLS2 تستعمل عندما يكون متغير الاستجابة مصفوفة، وخوارزمية PLS1 تعطي نتائج خوارزمية SIMPLS نفسها ويكون الاختلاف في كيفية تفرغ البيانات، ففي الأول يتم عن طريق X و Y اما في خوارزمية SIMPLS يكون لمصفوفة التباين والتباين المشترك. وسنركز في هذا البحث على الخوارزميتين NIPALS و SIMPLS.

2-4 انحدار المربعات الصغرى الجزئية (1) (5) (6) partial least square regression

ان طريقة المربعات الصغرى الجزئية تعتمد على خطوتين اساسيتين لإيجاد المتغيرات الكامنة Latent variable بين Y و X من خلال تعظيم مصفوفة التباين والتباين المشترك اما الخطوة الثانية فهي انحدار Y على المركبات. فلو فرضنا ان لدينا مصفوفة $X_{n,p}$ والمتجه $Y_{n,1}$ فطريقة المربعات الصغرى الجزئية تعتمد على النموذج الثاني بين X و Y الاتي:

$$X = TP + E \dots \dots \dots (1)$$

$$Y = Uq + F \dots \dots \dots (2)$$

حيث ان:

- T: مصفوفة درجات - X-score X ببعده $n*r$
- U: مصفوفة درجات - Y-score Y ببعده $n*r$
- P: مصفوفة تحميلات - X-loading X ببعده $p*r$
- q: متجه تحميلات - Y-loading Y ببعده $1*r$
- E: مصفوفة البواقي - X-residual X ببعده $n*p$
- F: متجه البواقي - Y-residual Y ببعده $n*1$

والمصفوفة P والمتجه q له r من الاعمدة ويكون عدد بما يأتي:

$$r < \min(n,p) \dots \dots \dots (3)$$

حيث ان:

- P: عدد المتغيرات
- n: عدد المشاهدات
- r: عدد المركبات

والعلاقة الداخلية تكون كالآتي:

$$U = TD + H \dots \dots \dots (4)$$

حيث ان:

- D: هي مصفوفة قطرية ذات بعد $r*r$
- H: مصفوفة البواقي ذات بعد $n*r$

ملخص طريقة المربعات الصغرى الجزئية هي إيجاد w من مجال X والمتجه c من مجال Y حيث ان:

$$\text{Max cov}(X_w, Y_c) \quad \text{with} \quad \|X_w\| = 1 \quad \|Y_c\| = 1 \dots \dots \dots (5)$$



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

حيث ان $cov(X_w, Y_c)$ هو تقدير التباين المشترك وتنفذ الطريقة بصورة تكرارية متسلسلة وهذا يعني إيجاد المتجهات Scores الواحدة بعد الأخرى حتى يتم استخراج كافة المتجهات الى r تحت قيد عدم الارتباط بين المتجهات r .

اما لحل المعادلة المذكورة أنفأ توجد عدة خوارزميات سنركز على خوارزمية NIPALS(PLS1) وخوارزمية SIMPLS في هذه الاطروحة.

3-4 خوارزمية NIPALS Algorithm NIPALS⁽¹⁾⁽⁵⁾⁽⁶⁾

1- نختار عمود من أعمدة المصفوفة Y بحيث ان $U_1=Y$.

2- حساب اوزان X (X-weight) باستخدام الصيغة التالية:

$$W_1 = \hat{X}U_1/\hat{U}_1U_1\text{.....(6)}$$

وان W_1 هي متجه ببعد $(p*1)$.

3- W_1 تكون normalized بالشكل الاتي:

$$W_1 = W_1/\|W_1\|\text{.....(7)}$$

1-نبدا بحساب درجات X (X-score) وهي اسقاطات للبيانات او المشاهدات X على اوزان X (weight).
X-)

$$t_1 = XW_1\text{.....(8)}$$

حيث ان t_1 هي متجه ببعد $(n*1)$

2-حساب اوزان Y (y-weight)

$$C_1 = \hat{U}t_1/\hat{t}_1t_1\text{.....(9)}$$

حيث ان C_1 متجه ببعد $(1*1)$.

3- C_1 يكون normalized بالشكل الاتي:

$$C_1 = C_1/\|C_1\|\text{.....(10)}$$

4-حساب درجات Y (y-score)

وهي تراكيب خطية لمتغير الاستجابة وهي استقطاعات لبيانات Y على اوزان Y (Y-weight).

$$U_1^* = YC_1\text{.....(11)}$$

حيث ان U_1^* متجه ببعد $(n*1)$

5-إيجاد U كالآتي:

$$U = U_1^* - U_1\text{.....(12)}$$

$$\Delta U=(U\Delta)' * (U\Delta)\text{.....(13)}$$

$$\Delta U < \epsilon$$

فإذا كانت $\Delta U < \epsilon$ حيث ϵ قيمة صغيرة وجدنا اول مركبة نتوقف، عدا ذلك نذهب الى الفقرة الأولى وتستهمل U_1^* بدل U_1 ونستمر بالخطوات.

6-إيجاد تحميلات X (X-loading) وهي معالم خطية تربط المتغيرات التوضيحية X مع درجات X (score).
X-)

$$P_1 = \hat{X}t_1/\hat{t}t_1\text{.....(14)}$$



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد
الخطي [دراسة تطبيقية على مرض فقر الدم]

7- إيجاد تحميلات Y - (Y -loading) وهي معالم خطية تربط متجه الاستجابات الى درجات Y - (Y -score) وهي بأبعاد 11^* . ويتم حسابها من الصيغة التالي:

$$q = \hat{Y}U_1/\hat{U}_1U_1.....(15)$$

حيث ان q متجه ببعد (1^*1) .

8- بعد ذلك يتم إيجاد التداخل الخطي للمعالم بواسطة انحدار OLS بالصيغة الآتية:

$$d_1 = \hat{U}_1t_1/t_1t_1.....(16)$$

حيث ان d_1 متجه ببعد (1^*1)

9- عمل تفرغ الى بيانات X و Y

$$X_1 = X_1 - t_1\hat{p}_1.....(16)$$

$$Y_1 = Y - d_1t_1\hat{c}_1.....(17)$$

ثم نستمر بالخطوات من (1-12) مرات عدة باستعمال البيانات المفرغة الى X و Y حتى نحصل على عدد المركبات المحدد، ثم نحدد معاملات الانحدار من خلال المعادلة الآتية:

$$\beta = W(\hat{P}W)^{-1}C.....(18)$$

حيث ان W مصفوفة ببعد p^*r وان P مصفوفة ببعد p^*r وان C مصفوفة ببعد r^*r .

5- العبور الشرعي Cross-Validation⁽¹⁾⁽⁵⁾⁽⁶⁾

ان تحديد عدد المركبات h في انموذج المربعات الصغرى الجزئية ضروري لأجل بناء الانموذج وتوجد عدة طرائق وهي جذر متوسط مربعات الخطأ الشرعي (Leave-one-out-cross-validation) وطريقة M -field-cross validation وهذا الأسلوب يستعمل عندما يكون حجم البيانات كبير حيث تقسم البيانات الى مجموعتين مجموعة اختيارية وهي مستقلة عن المجموعة الثانية وهي مجموعة التدريب والخطوات المتبعة لأجل تحديد عدد المركبات كالآتي:

1- حذف مشاهدة واحدة بالاعتماد على طريقة العبور الشرعي المختار.

2- يتم حساب الانموذج بدون المشاهدة المحذوفة.

3- التنبؤ بقيمة الاستجابة $X_{ij.h}$ والتي لا تتضمن المشاهدة المحذوفة باستعمال طريقة PLS ل h مركبة

$$\hat{Y}_{ij.h}$$

4- إيجاد جذر متوسطات مربعات الخطأ التشريعي ل h مركبة وكالآتي:

$$PRESS = \sum_{i=1}^h (y_i - \hat{y}_{-i.(h)})(19)$$

حيث ان $\hat{y}_{-i.(h)}$ هي القيمة التقديرية الى Y الحاصل عليها من البيانات بدون المشاهدة المحذوفة i من

انموذج المربعات الصغرى الجزئية PLS للمركبة h .

وان عدد المركبات النهائية تحدد من اقل PRESS او من خلال اقل قيمة الى جذر متوسط مربعات الخطأ الشرعي:

$$RMSECV(h) = \sqrt{\frac{1}{n} \sum_{i=1}^h (y_i - \hat{y}_{-i.(h)})}.....(20)$$

$$= \sqrt{\frac{1}{n} RRESS(h)}.....(21)$$

حيث ان n العدد الكلي.



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

تكرار الخطوات من (1-3) حيث في كل مرة يتم حذف مشاهدة واحدة وإيجاد الانموذج والتنبؤ. وان احصاءة لا يمكن الاعتماد عليها في حالة وجود القيم الشاذة في البيانات لذا يتوجب إزالة القيم الشاذة الموجودة في البيانات قبل البدء بعملية تحديد المركبات. وان تحديد منحني $RMSECV(h)$ من الطرائق الشائعة لتحديد العدد الأقل من المركبات.

6- الانحدار اللوجستي logistic regression⁽²⁾⁽⁴⁾⁽⁸⁾

يعد انموذج الانحدار اللوجستي من النماذج الإحصائية المهمة في تحليل البيانات اذ ان الهدف الأساسي من معظم الدراسات هو تحليل وتقويم العلاقات بين مجموعة من المتغيرات للحصول على صيغة نستطيع من خلالها ان نصف النموذج ويستعمل انموذج الانحدار اللوجستي لوصف العلاقة بين متغير الاستجابة من النوع المتقطع والمتغيرات التوضيحية ويكون على نوعين انموذج الانحدار اللوجستي ثنائي الاستجابة وانموذج الانحدار اللوجستي متعدد الاستجابة وسنركز على النوع الأول من الانحدار اللوجستي.

1-6 انموذج الانحدار اللوجستي ثنائي الاستجابة⁽²⁾⁽⁴⁾⁽⁸⁾

ان من خصائص الانحدار اللوجستي ثنائي الاستجابة ان المتغير التابع (y) متغير الاستجابة يتبع توزيع برنولي ويأخذ القيم (0) و (1) اي بمعنى ان (1) باحتمال قدره (p) واحتمال فشل (1-p) قدره (0) وبمعنى اخر احتمال حدوث الاستجابة (1) واحتمال عدم حدوث الاستجابة (0). ان نموذج الانحدار الخطي المتعدد MLR يكون كالآتي:

$$Y = XB + \epsilon \dots \dots \dots (22)$$

حيث ان:

Y : موجه المتغير المعتمد

X : مصفوفة المتغيرات التوضيحية

B : موجه معلمات دالة الانحدار

E : موجه الأخطاء العشوائية والذي يشترط به تحقق الشروط الآتية:

1- تتوزع توزيع طبيعي ϵ_i

2- $E(\epsilon_i) = 0$

3- $Cov(\epsilon_i, \epsilon_j) = 0$

4- $Var(\epsilon_i) = \sigma^2$

ففي حالة وجود متغير توضيحي واحد فان متوسط قيم المشاهدة Y عند متجه معين للمتغير x فهو $E(y|x)$ وبذلك يمكن كتابة النموذج على النحو التالي:

$$E(y|x) = \beta_0 + \beta_1 x \dots \dots \dots (23)$$

ان الطرف الايمن لهذا النموذج يأخذ قيم $(-\infty, +\infty)$ ، لكن عندما يكون المتغير (y) ثنائيا فان الانموذج اعلاه لا يكون ملائما لأن:

$$E(y/x) = P_r(y = 1) = p \dots \dots \dots (24)$$

وفي هذه الحالة يكون الطرف الايمن محصور بين (0,1)، وهذا يعني ان الانموذج يكون غير قابل للتطبيق احصائيا. وللتخلص من هذه المشكلة سنقوم بإدخال تحويل رياضي على المتغير التابع y. وبما ان (0 ≤ p ≤ 1) وان $\left(\frac{p}{1-p}\right)$ هو مقدار موجب محصور بين (0, ∞) أي $(0 \leq \frac{p}{1-p} \leq \infty)$ وبأخذ



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

اللوغاريتم الطبيعي للأساس (e) للمقدار $\left(\frac{p}{1-p}\right)$ فان مجال القيمة تصبح محصور بين $(-\infty, +\infty)$ وتكون كالاتي $(-\infty \leq \log_e\left(\frac{p}{1-p}\right) \leq \infty)$ ، وبالنهاية يمكن كتابة انموذج الانحدار في حالة وجود متغير توضيحي واحد وكالاتي:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \dots \dots \dots (25)$$

اما اذا كان لدينا اكثر من متغير توضيحي فتصبح صيغته كالاتي:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \dots \dots \dots (26)$$

اذ ان:

$$i = 1, 2, 3, \dots, n \quad j = 1, 2, 3, \dots, k$$

موجه للمعالم المطلوب تقديرها: $\beta_1, \beta_2, \dots, \beta_k$

x_{ij} : متغيرات توضيحية.

اما بالنسبة $\left(\frac{p}{1-p}\right)$ نسبة افضلية النجاح (odds of success) او نسبة الافضية للحدث المرغوب به وصيغته الرياضية هي كالاتي:

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\beta_0 + \sum_{j=1}^k \beta_j x_{ij}} \dots \dots \dots (27)$$

وصيغة احتمالات الاستجابة لأنموذج الانحدار اللوجستي تكتب كالاتي:

$$p = \frac{1}{1 + (e^{\beta_0 + \beta_1 x_i}) - 1} \dots \dots \dots (28)$$

وان المقدار $\text{Log}_e\left(\frac{p}{1-p}\right)$ يسمى لوغاريتم نسبة افضلية النجاح (logs odds of success).

2-6 الافتراضات الخاصة بالانحدار اللوجستي

ان الانحدار اللوجستي لا يتطلب افتراضات كثيرة فقط يتطلب عدم وجود ارتباط بين المتغيرات التوضيحية وان حجم المشاهدات كبيرة في كل مجموعة يفترض انها تكون اكبر من خمس مرات من عدد المعالم المستعملة في الانموذج النهائي.

3-6 تقدير احتمالات الاستجابة Estimation of response probabilities (2)(4)(8)

ان صيغة احتمال الاستجابة لأنموذج الانحدار اللوجستي الثاني يمكن الحصول عليها وذلك بإدخال ال (e) على المعادلة الاتية:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \dots \dots \dots (29)$$

وذلك صيغة احتمال الاستجابة تصبح كالاتي:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_i} \dots \dots \dots (30)$$

ويمكن ايضا ان نكتب المعادلة المذكورة أنفأ بالصيغة الاتية:

$$\frac{1}{\frac{1}{p} - 1} = e^{\beta_0 + \beta_1 x_i} \dots \dots \dots (31)$$



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

وباستعمال الطرائق الجبرية للاشتقاق تصبح صيغة احتمال الاستجابة لـ نموذج الانحدار اللوجستي الثنائي يكون كما يأتي:

$$P = \frac{1}{1+(e^{\beta_0+\beta_1x_i})-1} \dots \dots \dots (32)$$

وهذا يعني ان احتمال متغير الاستجابة (y) تأخذ القيمة (1) ويكون كالآتي:

$$P(y = 1 | x) = \frac{1}{1+(e^{\beta_0+\beta_1x_i})-1} \dots \dots \dots (33)$$

وعند القيمة (0) فان احتمال متغير الاستجابة (y) يكون على النحو الآتي:

$$P(y = 0 | x) = \frac{1}{1+e^{\beta_0+\beta_1x_i}} \dots \dots \dots (34)$$

وبما ان مجموع الاحتمالات يساوي (1) فأن:

$$P(y = 1|x) + P(y = 0|x) = 1 \dots \dots \dots (35)$$

7- بعض الاختبارات الإحصائية المهمة (8)(4)(2)

هناك بعض الاختبارات الإحصائية المهمة منها:

1- اختبار wald

ليبيان أهمية ومعنوية معاملات الانحدار المقدره بطريقة الإمكان الأعظم بالنسبة للأنموذج اللوجستي وتختبر احصاءة wald حسب الفرضية الآتية:

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0$$

واحصاءة الاختبار هي كالآتي:

$$wald = t^2 = \left(\frac{\hat{b}_i}{S.E(\hat{b}_i)} \right)^2 \dots \dots \dots (36)$$

: هي تمثل معلمة المتغير المعتمد \hat{b}_i

: الخطأ القياسي للمعلمة $S.E(\hat{b}_i)$.

فاذا كانت قيمة p-value اقل من 0.05 نرفض فرضية العدم H_0 أي ان معاملات المتغير التوضيحي معنوية.

2- اختبار الدرجة او المعيارية Score Test

وهو اختبار إحصائي يختبر المعلمات في اطار فرضية العدم ($H_0 : =0\beta$) ، إذ يعد هذا الاختبار من أقوى الاختبارات عندما تقترب قيمة المعلمات من قيمة (β_0) ، فالميزة الرئيسية لهذا الاختبار هو عدم حاجته الى تقدير المعلومات تحت إطار الفرضية البديلة ($H_1 : \beta \neq 0$) ، إذ تقترب أحصاءة هذا الاختبار من توزيع مربع كاي (χ_p^2) وتقارن مع قيمتها الجدولية.

3- اختبارات حسن المطابقة Goodness OF FIT

جودة الملاءمة تعني كم إن الإنموذج الإحصائي هو ملائم لبيانات عينة الدراسة، فمقاييس جودة الملاءمة تقيس التقارب بين القيم المشاهدة والمتوقعة للإنموذج وفيما يأتي بعض الاختبارات المهمة لجودة الملاءمة.



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

4-اختبار wilk's Lambd

في هذا الاختبار يمكن الكشف عن وجود علاقة خطية بين المتغيرات وتكون صيغته كالآتي:

$$A = \frac{|w|}{|T|} = \frac{|w|}{|w-B|} \dots \dots \dots (37)$$

حيث ان:

W: مصفوفة التباين والتباين المشترك داخل المجاميع.

B: مصفوفة التباين والتباين المشترك بين المجاميع.

T: مصفوفة التباين والتباين المشترك الكلي.

ويمكن حساب مصفوفة B بالشكل الآتي:

$$B = \sum_{i=1}^k (\bar{X}_j - \bar{X})(\bar{X}_j - \bar{X}) \dots \dots \dots (38)$$

ويمكن حساب مصفوفة T بالشكل الآتي:

$$T = \sum_{i=1}^k \sum_{j=1}^m (\bar{X}_{ij} - \bar{X})(\bar{X}_{ij} - \bar{X}) \dots \dots \dots (39)$$

ويمكن حساب مصفوفة W بالشكل الآتي:

$$W=T-B \dots \dots \dots (40)$$

ويمكن ملاحظة ان قيمة A محصورة بين الصفر والواحد.

5-اختبار Hosmer-Lemeshow

لقياس جودة المطابقة Goodness of Fit لنموذج الانحدار اللوجستي يتم استعمال اختبار Hosmer -Lemeshow ويستعمل هذا الاختبار بشكل واسع لتقدير جودة الملاءمة للنموذج ويسمح بأي عدد من المتغيرات التوضيحية والتي قد تكون مستمرة او متقطعة وهذا الاختبار

يشابه الى حد ما اختبار χ^2 لجودة الملاءمة QUALITY OF CONVENIENCE إذ

يقوم هذا الاختبار بتجميع حالات العينة بناء على قيم الاحتمالات المتوقعة، وقد اقترح Hosmer

Lemeshow- استعمال احدى استراتيجيتين للتجميع في هذا الاختبار هما:

أ-تجميع الحالات بناء على المنينيات للاحتتمالات المتوقعة.

ب-تجميع الحالات بناء على قيم ثابتة للاحتتمالات المتوقعة.

وتفضل الاستراتيجية الاولى على الثانية لاسيما عندما يكون هناك العديد من الاحتمالات

المتوقعة صغيراً (اقل من 0.2) ووفقاً لهذه الاستراتيجية يتم تجميع الحالات بناء على المنينيات

للاحتمالات على عشر مجموعات (G=10) إذ يكون عدد الحالات في كل مجموعة (N/10) وإذ

توضع في المجموعة الاولى الحالات ذات اقل قيمة للاحتمالات المتوقعة وتوضع في المجموعة

الاخيرة الحالات ذات القيم الاعلى للاحتمالات المتوقعة وكذلك مع بقية المجموعات بالترتيب.

ويتم جمع القيم المشاهدة والمتوقعة للحالات وفقاً لقيمتي المتغير التابع Y (0 , 1) وكذلك في

كل فئة من المجموعات العشر وبعد ذلك يتم حساب احصاءة Hosmer -Lemeshow التي يرمز

لها بالرمز H وفقاً لإحصاء مربع كاي χ^2 للتكرارات المشاهدة والمتوقعة ، إذ ان الاحصاءة H

تتبع توزيع مربع كاي بدرجات حرية تساوي (d-2) .



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

6- معامل التحديد R^2 The Coefficient Of Determination

كثيراً ما يعرف R^2 بأنه نسبة التباين او الاختلاف في المتغير التابع التي يمكن تفسيرها من

خلال المتغيرات التوضيحية في الإنموذج، ويمكن أن يعرف R^2 بطرائق أخرى، كل منها ينتج قيمة متطابقة لمعامل التحديد في إنموذج الانحدار الخطي (LRM)، ومع ذلك، عندما يتم تطبيق هذه الصيغ للنماذج الفئوية فإنها تنتج قيم مختلفة، وبالتالي توفر قياسات مختلفة لملاءمة النماذج، ان احصاءة الاختبار (R^2) لا تفسر جودة الملاءمة للإنموذج كما في الانحدار الخطي ولكنها تعد مؤشراً لأهمية المتغيرات التوضيحية للتنبؤ بمتغير الاستجابة ومن ثم قياس حجم التأثير.

ويمكن أن نذكر صيغة R^2 Cox & Snell كالتالي :

$$R^2_{cs} = 1 - \left[\frac{\hat{L}(B_{(0)})}{\hat{L}(B)} \right]^{2/N} \dots\dots\dots(41)$$

ان هذا الصيغة إذ $\hat{L}(B_{(0)})$ هو دالة الإمكان للإنموذج الذي يحتوي فقط على معلمة الحد الثابت (دالة الامكان للإنموذج تحت فرضية العدم) اي وإن $\hat{L}(B)$ دالة الإمكان للإنموذج الذي يتضمن معلمات الميل (دالة الامكان للإنموذج تحت الفرضية البديلة) N: العدد الكلي للملاحظات

أما صيغة Nagelkerke R^2 فيمكن توضيحها بالصيغة الآتية:

$$R^2_N = \left[\frac{R^2_{cs}}{1 - \{\hat{L}(B_{(0)})\}^{2/N}} \right] \dots\dots\dots(42)$$

ان حجم العينة الكبير يعطي قوة لتحليل الاختبار لنموذج اللوجستيك إذ انه كلما ازداد حجم العينة كلما ازدادت قوة الاختبار للنموذج.

7- جودة توفيق النموذج

لقياس جودة توفيق النموذج سنجري اختبارا لا معلما يكون اعتماده على اختبار نسبة الامكان الاعظم (likelihood ratio test) والذي يتبع احصاءة او توزيع مربع كاي وكما في الصيغة الآتية:

$$\chi^2 = 2[\log_e H_0 - \log_e H_1] \dots\dots\dots(43)$$

اذ أن

H_0 : قيمة دالة الامكان الاعظم تحت فرضية العدم.

H_1 : قيمة دالة الامكان الاعظم تحت الفرضية البديلة.



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

8- التحليل التمييزي: Discriminant analysis (9)(7)(3)

يعد التحليل التمييزي من الأساليب الإحصائية المهمة في متعدد المتغيرات التي تهتم بتفريق (تمييز) بين مجتمعين (مجموعتين) أو أكثر من خلال إيجاد توافق خطية للمتغيرات التوضيحية تعرف بدالة التمييز التي يمكن عن طريقها الفصل أو تمييز بين مجموعتين أو أكثر ثم تأتي بعدها عملية تصنيف المفردات الجديدة إلى أحد المجموع بأقل خطأ ممكن. وهناك عدة دوال تمييزية منها الدالة المميزة الخطية والتي قدمت من قبل العالم فيشر عام (1936) والتي تستعمل في حالة تساوي مصفوفة التباينات المشتركة للمجاميع كافة وإن كل مجموعة تعبر عن مجتمع طبيعي لمتعدد المتغيرات إما في حالة عدم تساوي مصفوفة التباينات المشتركة لنجا إلى الدالة المميزة التربيعية.

1-8 الدالة المميزة الخطية ل مجموعتين The linear discriminant function for two groups

(9)(7)(3) groups

تسمى هذه الدالة بدالة فيشر (Fisher Function) تتوزع فيها المشاهدات توزيعاً طبيعياً. لنفترض أن لدينا مجتمعين ونريد المقارنة بينهما ولنفرض أن هذين المجتمعين لهما نفس مصفوفة التباين والتباين المشترك ($\Sigma_1 = \Sigma_2$) ولهما نفس موجه متوسطات (μ_1, μ_2) بالتتابع، وتم اختيار عينتين عشوائيتين ($x_{21}, x_{22}, x_{23}, \dots, x_{2n_2}$) و ($x_{11}, x_{12}, x_{13}, \dots, x_{1n_2}$) لكل من المجتمعين. وتكون صيغة الدالة المميزة الخطية ليفشر تكتب بالشكل الآتي:

$$Z = \underline{a}' \underline{x} = \left(\bar{x}_1 - \bar{x}_2 \right)' S_p^{-1} \underline{x} \dots \dots \dots (44)$$

حيث أن:

$$\underline{a} = S^{-1} \left(\bar{x}_1 - \bar{x}_2 \right) \dots \dots \dots (45)$$

وإن مقياس (Mahalanobis) يعتمد بالأساس على القياسات ذات المسافة القليلة بين قيم المتغيرات للمشاهدات الجديدة وقيم المتوسطات للمتغيرات لكل مجموعة يمكن كتابتها بالشكل الآتي:

$$D_i^2 = \left(\bar{x}_1 - \bar{x}_i \right)' S_p^{-1} \left(\bar{x} - \bar{x}_i \right) \dots \dots \dots (46)$$

إذ أن :

\bar{x}_i : موجه متوسطات المتغيرات لكل قيمة من المجموعة (i)

S_p^{-1} : هو معكوس مصفوفة التباين والتباين المشترك المقدرة داخل العينتين.

2-8 الافتراضات الخاصة بالتحليل المميز (9)(7)(3)

إن من الافتراضات الخاصة بالتحليل المميز بأن المجاميع المدروسة ذات توزيع طبيعي متعدد متغيرات في حالة الدالة المميزة الخطية يشترط فيها أن تكون مصفوفة التباين متساوية للمجاميع كافة وفي حالة عدم التساوي نستعمل الدالة المميزة التربيعية، كذلك يتطلب أن تكون حجم العينة كبيرة حيث إن المجاميع المختلفة تتضمن على الأقل 20 مشاهدة لكل متغير توضيحي وأيضاً يشترط عدم وجود ارتباط بين المتغيرات التوضيحية وعدم وجود قيم شاذة بينها.

9- تصنيف البيانات classification of data (10)

لنفرض أن:

لدينا عينة بحجم (n).

إن عدد المشاهدات من النوع (0) هي n_1

وإن عدد المشاهدات من النوع (1) هي n_2



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

وان لدينا بيانات من نوع ثنائي (binary) وكانت لدينا تصنيف البيانات كما مبين في الجدول رقم (1) الاتي:
جدول رقم (1) يبين تصنيف البيانات

| البيانات الثنائية | (0) | (1) |
|-------------------|-----------------|-----------------|
| (0) | A ₁₁ | A ₁₂ |
| (1) | A ₂₁ | A ₂₂ |

وتكون معايير التصنيف بالشكل الاتي:
1-نسبة التصنيف الصحيحة تحسب بالشكل الاتي:

$$= \left(\frac{A_{11}+A_{22}}{n} \right) * 100\% \dots \dots \dots (47)$$

وبالتالي يمكن حساب معيار خطأ التصنيف الصحيح بالشكل الاتي:

$$100\% \dots \dots \dots - \text{نسبة التصنيف الصحيح} = \text{معيار خطأ التصنيف} \dots \dots \dots (48)$$

2-نسبة التصنيف الجزئية على مستوى (0) و (1) وتحسب بالشكل الاتي:

$$p(0) = \left(\frac{A_{11}}{n_1} \right) * 100\% \dots \dots \dots (49)$$

$$p(1) = \left(\frac{A_{22}}{n_2} \right) * 100\% \dots \dots \dots (50)$$

1-نسبة التصنيف الكلية على مستوى (0) و (1) وتحسب بالشكل الاتي:

$$p(0) = \left(\frac{A_{11}+A_{21}}{n} \right) * 100\% \dots \dots \dots (51)$$

$$p(1) = \left(\frac{A_{12}+A_{22}}{n} \right) * 100\% \dots \dots \dots (52)$$

10- الجانب التطبيقي

ان انتشار الامراض في العراق خاصة والعالم بصورة عامة على تنوعها وتعدد اساليبها تعتبر ظاهرة جديدة بالبحث والدراسة وضرورة الوقوف على مسبباتها وتحديد طرق معالجتها من الواجبات الاساسية التي تقع على عاتق الدارسين والباحثين كل حسب اختصاصه على حد سواء. بعد البحث والتقصي الامراض الاكثر شيوعا وخطورة على فئات المجتمع الانساني وتبين ان امراض فقر الدم هي الاكثر شيوعا. ومن ضمن امراض كثيرة شائعة والاكثر تأثيرا على انتشار اجيال جديدة ترفد مجتمعنا بطاقات متجددة على صعيد العقل والجسد. وفي موضوع الرسالة تم تناول فقر الدم على مستويين وهما فقر الدم الحاد بالرمز (0)، وفقر الدم المزمن بالرمز (1).

10-1 فقر الدم

هو حالة مرضية تحدث بسبب انخفاض في نسبة تركيز الهيموكلوبين عن المستوى الطبيعي وبسبب الهبوط في مستوى الهيموغلوبين تعاني الأجهزة من عدم الحصول على ما يكفي من الأوكسجين وبالتالي يشكو المريض من اعراض الإرهاق والصداع وعدم التركيز والخمول وغيرها.
هناك ثلاث أنواع رئيسية لفقر الدم: فقر الدم الناجم عن فقدان الدم، وفقر الدم الناجم عن خلل في إنتاج كريات الدم الحمراء، وفقر الدم الوراثي. ويمكن اعتبار فقر الدم هو الحالة المرضية الأكثر شيوعا في أمراض الدم.
اما انواع فقر الدم فتعدد تصنيفاتها لكن الأبرز والأشمل من هذه الانواع هي:

1- فقر الدم الحاد

2- فقر الدم المزمن



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

والذي تم على هذا الاساس تم اعتماد انواع فقر الدم في هذا البحث.
اما علامات ظهور هذا المرض فهي على النحو الاتي:

- 1- الاعياء العام والارهاق
 - 2- شحوب الجلد
 - 3- تسارع او عدم انتظام في ضربات القلب.
 - 4- انقطاع أو ضيق النفس.
 - 5- الم في الصدر.
 - 6- دوام.
 - 7- صعوبة في التركيز او التفكير.
 - 8- احساس بالبرودة في الأطراف .
 - 9- صداع.
 - 10- تكسر الأظافر.
- قد يكون فقر الدم غير ظاهر ولكن بشكل عام تزداد الأعراض مع تقدم المرض.

2-10 متغيرات الدراسة:

- 1- انواع فقر الدم المتمثل بالمتغير (Y)، هناك نوعان من فقر الدم وهما:
أ- فقر الدم الحاد: وهو نوع من انواع المرض الذي يحدث غالبا بسبب نزف دم يسمى في بعض المصادر بنزف الدم الداخلي (يحدث داخل جسم الانسان) ناتج بسبب اصابة خارجية للجسم او بسبب النزف الحاد للدورة الشهرية عند النساء وغيرها، ونزف دم خارجي (خارج جسم الانسان) الذي يحدث غالبا بسبب الاصابة بجرح على الجسم يسبب هذا النزيف.
ب- فقر الدم المزمن: وهو النوع الثاني من مرض فقر الدم المصاحب للالتهاب هو شكل من أشكال فقر الدم التي تحدث مع الأمراض المزمنة مثل: **العدوى المزمنة** وتنشيط المناعة المزمن **والسرطان**. كما تقترح الاكتشافات الجديدة أن هذا النوع من الأنيميا هو نتيجة إنتاج الجسم مادة **الهيبيدين (hepcidin)** وهو المتحكم الرئيسي بعمليات أيض الحديد في جسم الإنسان. ومن انواع فقر الدم المزمن فقر دم البحر الأبيض المتوسط او تكسر كريات الدم الحمراء.
- 2- الجنس المتمثل بالمتغير (X1).
- 3- العمر المتمثل بالمتغير (X2): ان للعمر تأثير كبير على مرض فقر الدم ففي فقر الدم الحاد غالبا يصاب الكبير في السن والشباب سواء كان ذكر او انثى بسبب تعرضهم لمسببات النزيف اكثر من صغار السن، اما مرض فقر الدم الوراثي فغالبا يصيب جميع الفئات العمرية.
- 4- نسبة الهيموغلوبين (hp) المتمثل بالمتغير (X3): الهيموغلوبين هو بروتين موجود داخل كريات الدم الحمراء، وهو الذي يكسب خلايا الدم اللون الأحمر، وتتخلص وظيفته في نقل الأوكسجين من الرئة إلى مختلف أعضاء الجسم حتى تقوم بوظائفها على أكمل وجه، ويقوم بنقل ثاني أكسيد الكربون من أنحاء الجسم إلى الرئة لطرده خارج الجسد، ويلعب الهيموجلوبين دوراً مهماً في الحفاظ على شكل خلايا الدم الحمراء، ففي شكله الطبيعي يحافظ على ثبات الكريات بحيث يكون شكلها دائرياً مقعر الوجهين، وعندما يكون غير طبيعياً فيدمر الشكل الرئيس للكريات مما يعطل وظيفتها وجرياتها في الأوعية الدموية.
اما نسب فقر الدم الطبيعية فتكون كالآتي:
يختلف المعدل الطبيعي للهيموغلوبين من شخص لآخر حسب عمره وجنسه كما يلي:
النسبة الطبيعية للرجال: من 13.5-17.5 جرام/ديسيلتر.
النسبة الطبيعية للإناث: من 12-16 جرام/ديسيلتر.
اما خارج هذه المعدلات فيعتبر اصابة بفقر الدم .



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

- 5- نسبة ferritin في الدم المتمثل بالمتغير (X4): ومخزن الحديد الأكبر في الجسم. عمليا، يشكل الفيريتين تركيبة من الحديد والبروتين يتيح تخزين الحديد في الانسجة. مستوى الفيريتين في المصل يمثل، عمليا، حجم مخزون الحديد في الجسم، وبواسطة هذا الفحص يمكن تشخيص سبب وجود فقر الدم (الانيميا)، ان انخفاض مستوى الفيريتين في الدم يشير الى فقدان الحديد ومن ثم حالة فقر الدم.
- 6- نسبة retic count المتمثلة بالمتغير (X5): تنقسم الخلايا في الدم إلى ثلاث مجموعات: خلايا الدم الحمراء المسؤولة عن نقل الأوكسجين من الرئتين إلى مختلف أنسجة الجسم، خلايا الدم البيضاء التي تعتبر جزءاً من جهاز المناعة وتحارب الجراثيم التي تدخل للجسم من الخارج، وطبعا الصفائح الدموية التي تلعب دوراً مهماً في عملية تخثر الدم هو فحص روتيني يقيس كمية وسماوات الأنواع الثلاثة من خلايا الدم.
- 7- نسبة MCV المتمثل بالمتغير (X6): معدل حجم كريات الدم الحمراء والتي يتراوح حجم خليتها لجميع الفئات العمرية للذكر والأنثى، ما بين (80.5 إلى 99.7) فيكولتر، ويزداد معدل حجم كريات الدم الحمراء في حال كان هنالك خلل ما في النسيج النخاعي، والذي يتزامن معه نقص حاصل في فيتامين (B12)، وحمض الفوليك، أما النقص الحاصل في معدل حجم كريات الدم الحمراء فينتج عن النقص الحاصل في الحديد والبيريدوكسين، والذي يتزامن مع عدد من الأمراض مثل التلاسيميا، وفقر نشاط الغدة الدرقية، وغيره من الأمراض المزمنة.
- 8- نقص الحديد في الدم Iron deficiency Anemia المتمثل بالمتغير (X7): يعتبر نقص الحديد في الجسم هو أكثر اسباب فقر الدم (ضعف الدم) انتشاراً في العالم، و يُصيب تقريباً 30% من سكان العالم (حوالي 500 مليون نسمة حالياً) ، و يرجع ذلك إلى قدرة الجسم المحدودة لامتصاص الحديد من الغذاء و فقد الحديد عن طريق فقد الدم بالنزف.
- 9- نسبة transferrin في الدم والمتمثل بالمتغير (X8): هو بروتين يقوم بنقل الحديد للخلايا، وبشكل اساسي الى النخاع العظمي، حيث يتم إنتاج خلايا الدم الحمراء. فيرتبط بروتين آخر يدعى الفيريتين (Ferritin). ويقيس الترانسفيرين مستوى الحديد في الدم المرتبط بالترانسفيرين، فضلا عن فحص كمية الترانسفيرين الاجمالية وكمية الفيريتين. تشير هذه الفحوص الى كمية الحديد في جسم الانسان، واذا ما كان الشخص يعاني من نقص او فائض في عنصر الحديد. في حالات الإصابة بنقص الحديد، ينخفض مستوى الحديد والفيريتين، بينما يرتفع، بالمقابل، مستوى الترانسفيرين.
- 10- سبب الفقر هو نزف الدم المتمثل بالمتغير (X9): تعد الاصابات المباشرة بالجسم التي تؤدي الى نزيف حاد نوع من انواع فقر الدم وكذلك النزيف الداخلي الحاد الحاصل لدى بعض المرض بسبب صدمة خارجية على الرأس وغيرها من الامراض وكذلك النزف الحاد للدورة الشهرية لدى المرأة . كل هذه الاسباب تسبب فقر الدم الحاد.
- 11- فقر دم الامراض المزمنة المتمثلة بالمتغير (X10): يطلق عليه أيضا فقر الدم المصاحب للالتهاب هو شكل من أشكال فقر الدم التي تحدث مع الأمراض المزمنة مثل: العدوى المزمنة وتنشيط المناعة المزمن أو السرطان. كما تقترح الاكتشافات الجديدة أن هذا النوع من الأنيميا هو نتيجة إنتاج الجسم مادة الهيبسيدين (hepcidin) وهو المتحكم الرئيسي بعمليات أيض الحديد في جسم الإنسان.
- 12- فقر الدم هو نقصان في كريات الدم الحمراء المتمثل بالمتغير (X11): انخفاض نسبة خلايا الدم الحمراء (كريات الدم الحمراء) في الدم. فان هناك العديد من العوامل التي يمكن ان تسهم في خفض نسبة RBC اكثر من المعتاد والتي تتأثر بالعمر والالتهابات الفيروسية وبعض الامراض المزمنة.

وقد جمعت البيانات من المستشفيات الاتية:

- 1-مستشفى بعقوبة العام / محافظة ديالى
- 2-مستشفى كلار العام / محافظة السليمانية
- 3-مستشفى الطوارئ المركزي / محافظة أربيل



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

3-10 التحليل الاحصائي للبيانات:

بعد اجراء التحليل الاولي للبيانات في برنامج ال SPSS وجد الاتي:

1- عدد المصابين بفقر الدم الحاد هم (67) مريضاً وبنسبة 47.9%، اما المصابين بفقر الدم المزمن هم (73) مريضاً وبنسبة 52.1% ما بين راقلين في المستشفى وقت جمع البيانات وما بين الملفات لمرضى رقدوا سابقا في المستشفى (طبقات المرضى). وكما مبين في الجدول رقم (2) ادناه:

جدول (2) يبين عدد المرضى المصابين بأنواع فقر الدم

| انواع فقر الدم | عدد المصابين بأنواع بالمرض | نسبة المصابين |
|----------------|----------------------------|---------------|
| فقر دم حاد | 67 | 47.9 |
| فقر دم مزمن | 73 | 52.1 |
| Total | 140 | 100.0 |

2- اما عدد الذكور والاناث في العينة فكانوا كما في الجدول رقم (3) كالآتي:

جدول رقم (3) يبين عدد الذكور والاناث في العينة

| | عدد الذكور والاناث | نسبة الذكور والاناث |
|-------|--------------------|---------------------|
| ذكر | 83 | 59.3 |
| انثى | 57 | 40.7 |
| Total | 140 | 100.0 |

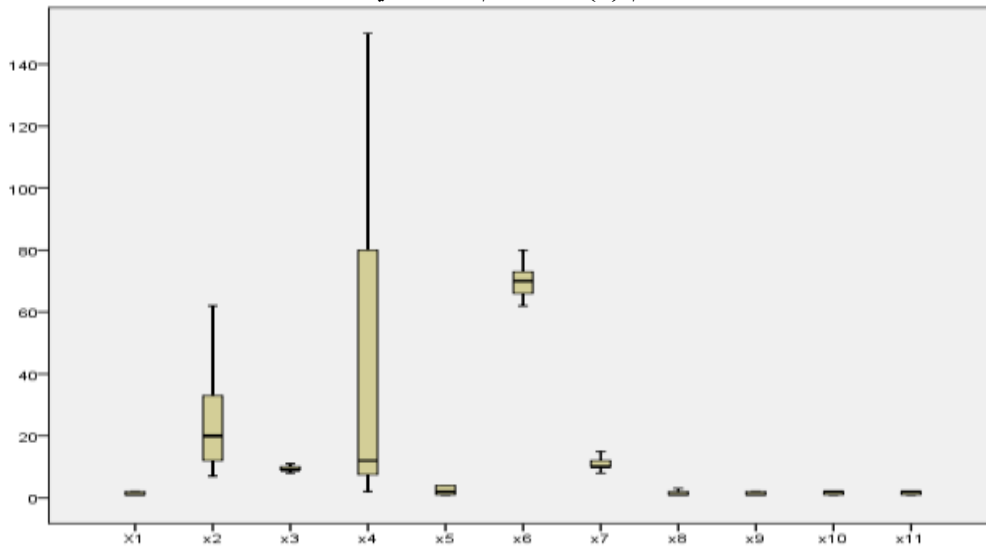
تبين ان عدد الذكور في العينة المختارة كانوا (83) ذكر وبنسبة (59.3%) وكانوا عدد الاناث ضمن العينة (57) وبنسبة (40.7%).

4-10 طريقة المربعات الصغرى الجزئية (partial least square method)

نبدأ بأجراء التحليل الاحصائي لطريقة المربعات الصغرى الجزئية وتكون كالآتي:

1- قبل البدء بتطبيق طريقة انحدار المربعات الصغرى الجزئية (partial least square regression) نتأكد من عدم وجود قيم شاذة وذلك من اختبار Box plot حيث ان طريقة المربعات الصغرى الجزئية لا يمكن ان تطبق في حالة وجود قيم شاذة وذلك كما في الشكل رقم (1) الاتي:

شكل رقم (1) يبين القيم الشاذة في البيانات





التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

نلاحظ من الشكل رقم (1) أعلاه ان المتغيرات التوضيحية لا تحتوي على أي قيمة شاذة خارج حدود (box plot).
2- نقوم بالتأكد من وجود ارتباط ذاتي بين المتغيرات من خلال اختبار VIF وكما مبين في الجدول رقم (4) الاتي:

جدول (4) يبين مشكلة التعدد الخطي

| Model | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. | Collinearity Statistics | |
|--------------|-----------------------------|------------|---------------------------|---------|------|-------------------------|--------|
| | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 (Constant) | 1.155 | 1.452 | | .795 | .428 | | |
| X1 | -.115- | .120 | -.113- | -.963- | .337 | .284 | 3.520 |
| x2 | .005 | .004 | .181 | 1.406 | .162 | .237 | 4.218 |
| x3 | -.019- | .061 | -.025- | -.310- | .757 | .605 | 1.654 |
| x4 | -.002- | .001 | -.215- | -2.551- | .012 | .555 | 1.802 |
| x5 | -.279- | .055 | -.718- | -5.042- | .000 | .194 | 5.151 |
| x6 | .028 | .014 | .299 | 1.984 | .049 | .174 | 5.757 |
| x7 | -.013- | .038 | -.043- | -.333- | .740 | .236 | 4.238 |
| x8 | -.411- | .146 | -.514- | -2.809- | .006 | .117 | 8.517 |
| x9 | .212 | .216 | .212 | .981 | .329 | .084 | 11.859 |
| x10 | -.092- | .191 | -.082- | -.483- | .630 | .138 | 7.254 |
| x11 | -.060- | .135 | -.054- | -.443- | .659 | .269 | 3.714 |

من الجدول رقم (42) اعلاه نلاحظ بان قيمة ال $VIF > 5$ للمتغيرات (x5-X6-X8-X9-X10) وهذا يدل على وجود مشكلة التعدد الخطي. فلذلك سوف نلجأ الى حل هذه المشكلة استخدام طريقة المربعات الصغرى الجزئية (partial least square method).

3- ولمعالجة مشكلة التعدد الخطي نطبق طريقة المربعات الصغرى الجزئية وكما في الجدول رقم (5) الاتي:
جدول رقم (5) يبين اختبار لمعنوية المربعات الصغرى الجزئية

| Source | DF | SS | MS | F | P |
|----------------|-----|---------|--------|-------|------|
| Regression | 7 | 17.255 | 2.465 | 18.40 | 0.00 |
| Residual Error | 132 | 17.6805 | 0.1339 | | |
| Total | 139 | 34.9357 | | | |

نلاحظ من الجدول رقم (أعلاه) جدول تحليل التباين لمتغير الاستجابة والذي يتضمن الانموذج الذي يحتوي على 7 مركبات حيث ان قيمة p-value المحسوبة لمتغير الاستجابة بلغت قيمتها (0.00) وهي اقل من قيمة $\alpha=0.05$ وهذا يدل على ان المركبات السبعة في الانموذج معنوية.

4- تم اختبار الانموذج بطريقة cross-validation لسبع مركبات ويمتلك معامل تحديد $R^2=0.49$ ومعامل تنبؤي 0.40 وان X-variance تشير الى مقدار التباين وقد ل 7 مركبات 0.94 من تباين المركبات المستقلة وكما مبين في الشكل رقم (6) الاتي:



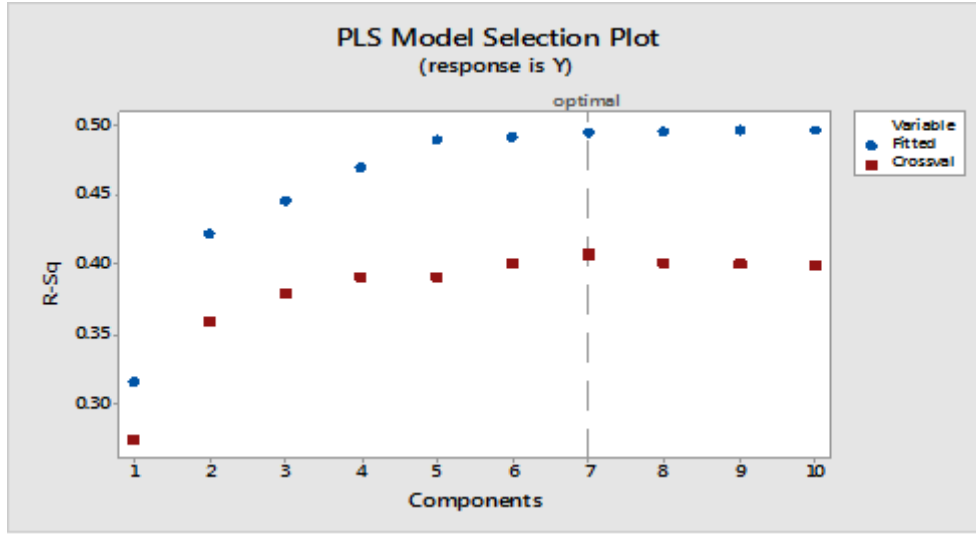
التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

جدول رقم (6) يبين معاملات التحديد ومعاملات التحديد التنبؤية

| Components | X Variance | Error | R ² | R ² (pred) | PRESS |
|------------|------------|--------|----------------|-----------------------|---------|
| 1 | 0.334 | 23.908 | 0.315 | 0.274 | 25.3604 |
| 2 | 0.481 | 20.193 | 0.421 | 0.359 | 22.3847 |
| 3 | 0.641 | 19.375 | 0.445 | 0.378 | 21.7079 |
| 4 | 0.730 | 18.518 | 0.469 | 0.391 | 21.2640 |
| 5 | 0.763 | 17.851 | 0.489 | 0.391 | 21.2659 |
| 6 | 0.861 | 17.752 | 0.491 | 0.401 | 20.9165 |
| 7 | 0.943 | 17.680 | 0.493 | 0.407 | 20.7064 |
| 8 | | 17.621 | 0.495 | 0.400 | 20.9360 |
| 9 | | 17.608 | 0.495 | 0.399 | 20.9757 |
| 10 | | 17.608 | 0.495 | 0.398 | 21.0105 |

5- ان افضل انموذج لطريقة المربعات الصغرى الجزئية حسب طريقة العبور الشرعي هل ل 7 مركبات حيث ان المحور العمودي يشير الى قيم معامل التحديد التنبؤي اما المحور الافقي فيشير الى المركبات حيث ان قيم المركبات تتزايد حتى المركبة السابقة وبعدها تتناقص. وكما مبين في الشكل رقم (2) ادناه:

شكل رقم (2) يبين عدد المركبات المختارة من المتغيرات المستقلة

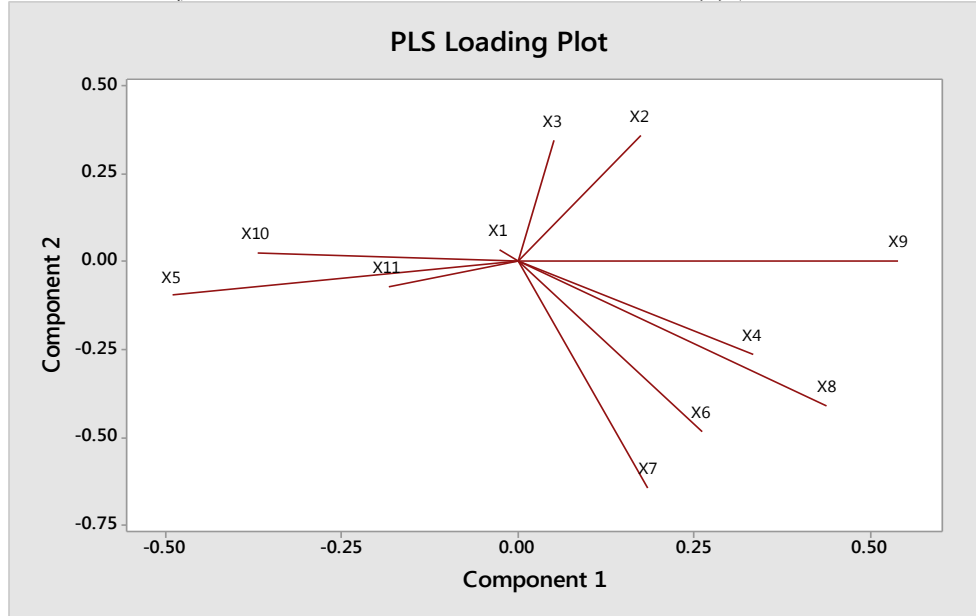


6- ان الشكل البياني لمركبات التحميل (Pls Loading) حيث ان X1 المتمثل (بالجنس) و X11 المتمثل (سبب فقر الدم هو نقصان في كريات الدم الحمر) يمتلك خط قصير جدا وهذا يدل على امتلاكها X-Loading منخفضة وهي ليست ذات علاقة مع متغير الاستجابة (فقر الدم) وان المتغيرات X2 المتمثل (بالعمر) و X3 المتمثل (بنسبة الهيموكلوبين hp) و X4 المتمثل (بنسبة الفرتين) و X5 المتمثل (نسبة retic count) و X6 المتمثل (نسبة MCV) و X7 المتمثل (نقص الحديد في الدم) و X8 المتمثل (نسبة ترانسفيرين) و X9 المتمثل (سبب فقر الدم هو نزف الدم) و X10 المتمثل (فقر الدم هو بسبب الامراض المزمنة) تمتلك خطوط طويلة أي ان لها احتمالات عالية أي ترتبط بعلاقة معنوية مع متغير الاستجابة. وكما مبين في الشكل رقم (3) الاتي:



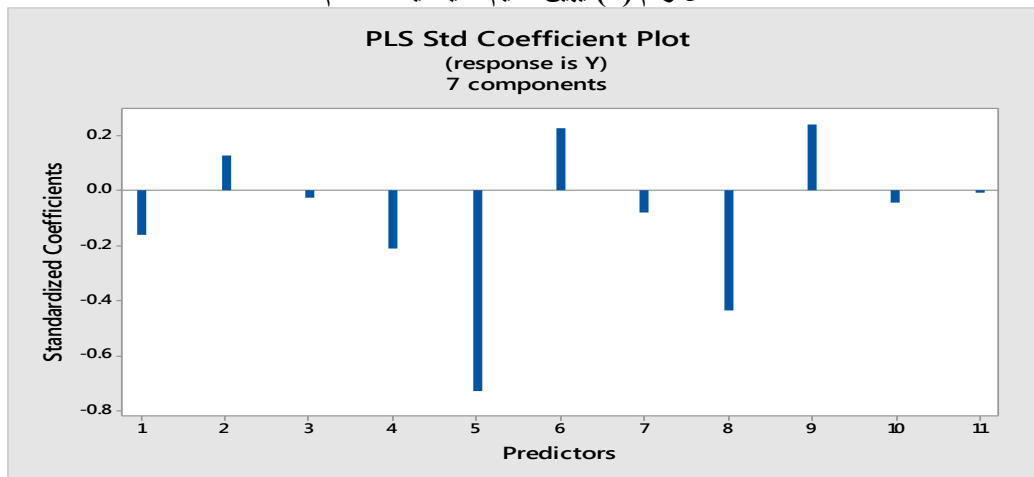
التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

شكل رقم (3) يبين علاقة المتغيرات المستقلة بالمتغير التوضيحي



7- الشكل رقم (4) يوضح القيم القياسية للمعالم (Standardized coefficients) وللمتغيرات التوضيحية حيث نجد ان x_5 المتمثل (نسبة ترانسفيرين) و x_9 المتمثل (سبب فقر الدم هو نزف الدم) و x_6 المتمثل (نسبة MCV) و x_4 المتمثل (بنسبة الفرتين) x_2 المتمثل (بالعمر) x_1 المتمثل (بالجنس) تمتلك اكبر معاملات قياسية حيث ان x_9 المتمثل (سبب فقر الدم هو نزف الدم) و x_6 المتمثل (نسبة MCV) و x_3 المتمثل (بنسبة الهيموغلوبين hp) و x_7 المتمثل (نقص الحديد في الدم) و x_2 المتمثل (بالعمر) ترتبط بصورة ايجابية مع متغير الاستجابة نزف الدم المتمثل (Y) والمتغيرات x_1 الجنس و x_5 و x_8 ترتبط بصورة سلبية مع متغير الاستجابة.

شكل رقم (4) يبين القيم القياسية للمعالم





التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

8- ان النموذج الرياضي لطريقة المربعات الصغرى الجزئية هو كالآتي:

$$y = 1.54 - 0.16x_1 + 0.003x_2 - 0.01x_3 - 0.001x_4 - 0.28x_5 + 0.02x_6 - 0.02x_7 - 0.34x_8 + 0.24x_9 - 0.04x_{10} - 0.006x_{11}$$

5-10 الدالة المميزة الخطية باستعمال المربعات الصغرى الجزئية

تم استعمال طريقة المربعات الصغرى الجزئية للتخلص من مشكلة التعدد الخطي. وتم استعمال طريقة العبور الشرعي لتحديد عدد المركبات وقد بلغ عدد المركبات الداخلة في انموذج المربعات الصغرى الجزئية (7) مركبات وسيتم تطبيق الاختبارات الإحصائية الخاصة بالدالة المميزة على هذه المركبات وتكون الاختبارات كالآتي:

1- من الجدول رقم (7) ادناه ومن خلال اختبار Wilk's نجد ان المركبة الأولى والثاني لها اثر معنوي ومهم في تكوين الدالة المميزة الخطية ولها تاثير كبير في التفرقة بين المجموعتين.

جدول رقم (7) يبين معنوية المركبات

| | Wilks' Lambda | F | df1 | df2 | Sig. |
|----|---------------|--------|-----|-----|------|
| C1 | .684 | 63.647 | 1 | 138 | .000 |
| C2 | .894 | 16.422 | 1 | 138 | .000 |
| C3 | .977 | 3.310 | 1 | 138 | .071 |
| C4 | .975 | 3.470 | 1 | 138 | .065 |
| C5 | .981 | 2.687 | 1 | 138 | .103 |
| C6 | .997 | .390 | 1 | 138 | .534 |
| C7 | .998 | .286 | 1 | 138 | .594 |

2- من الجدول رقم (8) ادناه تم اختبار معنوية الدالة المميزة الخطية حيث توجد قوة للتمييز من خلال التباين بين المجموعتين والتي قد فسرت 100% من التباين.

جدول رقم (8) يبين معنوية الدالة المميزة الخطية

| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
|----------|------------|---------------|--------------|-----------------------|
| 1 | .976 | 100.0 | 100.0 | .703 |

3- من خلال الجدول رقم (9) ادناه تم اختبار وجود علاقة خطية بين المتغيرات حيث ان قيمة Wilk's lambda 0.506 ونجد معنوية chi-square حيث بلغ قيمة sig. 0.00 وهي اقل من قيمة $\alpha=0.05$ وهذا يعني وجود العلاقة الخطية بين المجموعتين في المركبات.

جدول رقم (9) يبين اختبار Wilks' Lambda

| Test of Function(s) | Wilks' Lambda | Chi-square | Df | Sig. |
|---------------------|---------------|------------|----|------|
| 1 | .506 | 91.601 | 7 | .000 |

4- من خلال الجدول رقم (10) نجد ان معاملات الدالة المميزة الخطية للمركبات وهي كالآتي:

$$y = 0.626pc_1 + 0.550pc_2 + 0.307pc_3 + 0.363pc_4 + 0.504pc_5 + 0.164pc_6 + 0.105pc_7$$



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

جدول رقم (10)
يبين معاملات الدالة المميزة الخطية

| | Function |
|------------|----------|
| | 1 |
| C1 | .626 |
| C2 | .550 |
| C3 | .307 |
| C4 | .363 |
| C5 | .504 |
| C6 | .146 |
| C7 | .105 |
| (Constant) | .000 |

5- ومن خلال الجدول رقم (11) نجد ان:

$$\hat{M} = 1/2 (-1.024 + 0.940)$$

اذا كانت المفردة الجديدة اكبر من \hat{M} تصنف على انها تعود الى المجموعة الأولى (فقر الدم الحاد) واذا كانت المفردة الجديدة اقل من \hat{M} تصنف على انها تعود الى المجموعة الثانية (فقر الدم المزمن).

جدول رقم (11)
يبين مجاميع الدالة المميزة الخطية

| | Function |
|-------------|----------|
| Y | 1 |
| فقر دم حاد | -1.024- |
| فقر دم مزمن | .940 |

من خلال الجدول رقم (12) ادناه نجد ان 67 صنفت بشكل صحيح من 67 أي ان 100% صنفت بشكل صحيح من فقر الدم الحاد وان 50 مشاهدة من مجموع 73 صنفت على انها مرض مزمن أي بنسبة 68.5% وان نسبة التصنيف الصحيحة بلغت 83.6% وهذا يعني ان احتمال خطأ التصنيف 16.4%.
جدول رقم (12) يبين نسب التصنيف للدالة المميزة

| | Y | Predicted Group Membership | | Total |
|-------|-------------|----------------------------|-------------|-------|
| | | فقر دم حاد | فقر دم مزمن | |
| Count | فقر دم حاد | 67 | 0 | 67 |
| | فقر دم مزمن | 23 | 50 | 73 |
| % | فقر دم حاد | 100.0 | .0 | 100.0 |
| | فقر دم مزمن | 31.5 | 68.5 | 100.0 |



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

6-10 الانحدار اللوجستي باستعمال المربعات الصغرى الجزئية

ان اول خطوة في التحليل الاحصائي للانحدار اللوجستي هو ان نقوم بتضمين الحد الثابت فقط في الانموذج في الخطوة الصفرية ثم بعدها يتم إضافة المتغيرات التوضيحية وذلك لتحديد كفاءة الانموذج.
1-ولإجراء اختبار لمعنوية النموذج بوجود علاقة بين المتغير المعتمد والمركبات كانت النتائج قيم χ^2 -square بان قيمة sig. اصغر من $\alpha=0.05$ أي ان النموذج ملائم عند ادخال المركبات والحد الثابتز وكما مبين في الجدول رقم (13) الاتي:

جدول رقم (13) يبين معنوية النموذج

| | Chi-square | Df | Sig. |
|-------------|------------|----|------|
| Step 1 Step | 88.215 | 7 | .000 |
| Block | 88.215 | 7 | .000 |
| Model | 88.215 | 7 | .000 |

2-ومن خلال الجدول رقم (14) ادناه نلاحظ ان قيمة سالب ضعف لوغارتم دالة الإمكان الأعظم في الدورة الثانية كانت 193.824 وتوقفنا لان التغير في معاملات الحد الثابت أصبح اقل من 0.001.

جدول رقم (14)

يبين سالب ضعف لوغارتم دالة الإمكان الأعظم

| Iteration | -2 Log likelihood | Coefficients | |
|-----------|-------------------|--------------|--|
| | | Constant | |
| Step 0 1 | 193.824 | .086 | |
| 2 | 193.824 | .086 | |

3-ان النسب المنوية للتصنيف الصحيح لمشاهدات المتغير المعتمد في الخطوة الصفرية حيث بلغت 52.1%. وان جميع المشاهدات ضمن المجموعة الاولى صنفت بشكل خاطئ أي بمعنى ان نسبة التصنيف الصحيحة 0% وان معيار خطأ التصنيف 100%. وان جميع المشاهدات صنفت بشكل صحيح ضمن المجموعة الثانية حيث بلغت نسبة التصنيف الصحيحة 100% وان معيار خطأ التصنيف 0%. وكما مبين في الجدول رقم (15) ادناه:

| جدول رقم (15) يبين نسبة التصنيف للخطوة الصفرية | | | | |
|--|-------------|------------|-------------|--------------------|
| | Observed | Predicted | | |
| | | Y | | Percentage Correct |
| | | فقر دم حاد | فقر دم مزمن | |
| Step 0 Y | فقر دم حاد | 0 | 67 | .0 |
| | فقر دم مزمن | 0 | 73 | 100.0 |
| Overall Percentage | | | | 52.1 |

4-ان قيمة سالب ضعف لوغارتم دالة الإمكان الأعظم بلغت 105.609 وهي اقل من متجه الانموذج الذي يتضمن الحد الثابت وهذا يدل على جودة النموذج. واتضح أيضا ان قيمة R^2 تشكل 0.624 من التباين في متغير الاستجابة تم تفسيره من قبل المركبات. وكما مبين في الجدول رقم (16) ادناه:



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

جدول رقم (16)

يبين سالب ضعف لو غارتم دالة الإمكان الأعظم

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|------|----------------------|----------------------|---------------------|
| 1 | 105.609 ^a | .467 | .624 |

5- ان النسب المنوية للتصنيف الصحيح لمشاهدات المتغير المعتمد في الخطوة رقم (1) حيث بلغت 80%. وان 54 من المشاهدات ضمن المجموعة الاولى صنفت بشكل صحيح و13 صنفت بشكل خاطئ. وان 58 من البيانات ضمن المجموعة الثانية صنفت بشكل صحيح و15 صنفت بشكل خاطئ. وكما مبين في الجدول رقم (17) ادناه:

جدول رقم (17) يبين نسبة التصنيف للخطوة رقم (1)

| | Observed | Predicted | | |
|--------------------|-------------|-------------|------------|--------------------|
| | | Y | | Percentage Correct |
| | | فقر دم مزمن | فقر دم حاد | |
| Step 1 | فقر دم حاد | 54 | 13 | 80.6 |
| | فقر دم مزمن | 15 | 58 | 79.5 |
| Overall Percentage | | | | 80.0 |

6- ان احصاءة مربع كاي لهذا النموذج قد بلغت 8.922 وان قيمة مستوى المعنوية اكبر من $\alpha=0.05$ أي ان النموذج ملائم عند الدخال المتغيرات التوضيحية اذ تقاربت القيم المشاهدة من القيم المشابهة لها أي ان النموذج يمثل البيانات بشكل جيد. وكما مبين في الجدول رقم (18) الاتي:
جدول رقم (18) يبين احصاءة مربع كاي

| Step | Chi-square | Df | Sig. |
|------|------------|----|------|
| 1 | 8.922 | 8 | .349 |

7- ان النموذج الرياضي للانحدار اللوجستي الثنائي كان بالشكل الاتي:
ان النموذج الرياضي للانحدار اللوجستي الثنائي كان كالاتي:

$$\log_e \left(\frac{p}{1-p} \right) = 0.833 + 1.195pc_1 + 1.094pc_2 + 0.954pc_3 + 0.674pc_4 + 1.018pc_5 - 0.045pc_6 + 0.542pc_7$$

ونلاحظ ان المعلمات المقدرة باستعمال دالة الإمكان الأعظم واحصاءة Wald لهذه المعلمات نلاحظ ان اهم المركبات المعنوية هي المركبة الأولى والثانية. وكما مبين في الجدول رقم (19) ادناه:



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

جدول رقم (19) يبين المعلمات المقدرة للنموذج

| | | B | S.E. | Wald | Df | Sig. | Exp(B) | 95% C.I. for EXP(B) | |
|---------------------|----|--------|------|--------|----|------|--------|---------------------|-------|
| | | | | | | | | Lower | Upper |
| Step 1 ^a | C1 | 1.195 | .249 | 23.015 | 1 | .000 | 3.302 | 2.027 | 5.380 |
| | C2 | 1.094 | .296 | 13.641 | 1 | .000 | 2.987 | 1.671 | 5.338 |
| | C3 | .954 | .576 | 2.741 | 1 | .098 | 2.595 | .839 | 8.023 |
| | C4 | .674 | .345 | 3.825 | 1 | .050 | 1.963 | .999 | 3.859 |
| | C5 | 1.018 | .655 | 2.420 | 1 | .120 | 2.769 | .767 | 9.987 |
| | C6 | -.045- | .364 | .016 | 1 | .901 | .956 | .468 | 1.949 |
| | C7 | .542 | .369 | 2.159 | 1 | .142 | 1.719 | .835 | 3.541 |
| Constant | | .833 | .404 | 4.246 | 1 | .039 | 2.300 | | |

7-10 مقارنة بين الدالة المميزة الخطية والانحدار اللوجستي

سنقوم بأجراء مقارنة بين الدالة المميزة الخطية والانحدار اللوجستي الثنائي من حيث احتمال خطأ التصنيف. فوجد ان احتمال خطأ التصنيف للدالة المميزة الخطية كان (17.1%) اقل من خطأ التصنيف للانحدار اللوجستي الثنائي الذي بلغ (20%) وهذا يعني ان الدالة المميزة الخطية أفضل في تصنيف مشاهدات عينة فقر الدم من الانحدار اللوجستي الثنائي.

11- الاستنتاجات

نستنتج من هذا البحث الآتي:

- 1- نلاحظ ان معامل تضخم التباين للبيانات قد قل بعد استخدام انحدار المربعات الصغرى الجزئية ولجميع المتغيرات حيث أصبحت جميعها اقل من (5) وهذا يعني انه تم التخلص من مشكلة التعدد الخطي.
- 2- عند المقارنة بين الطريقتين باستخدام معيار خطأ التصنيف تم التوصل الى ان الدالة المميزة الخطية هي أفضل في تصنيف البيانات من الانحدار اللوجستي الثنائي حيث بلغ معيار خطأ التصنيف للدالة المميزة الخطية (17.1%) وهو اقل من معيار خطأ التصنيف للانحدار اللوجستي الثنائي الذي بلغ (20%).
- 3- عند التوصل الى ان الدالة المميزة الخطية أفضل في التصنيف من الانحدار اللوجستي الثنائي هذا يعني ان نموذج الدالة المميزة أفضل في التنبؤ من الانحدار اللوجستي الثنائي لأنها أعطت اقل خطأ تصنيف.

12- التوصيات:

بناءً على ما تم التوصل اليه من الاستنتاجات نوصي بالآتي:

- 1- نوصي الى استخدام طريقة انحدار المربعات الصغرى الجزئية في التخلص من مشكلة التعدد الخطي لأنها عالجت الارتباط الخطي بين المتغيرات التوضيحية.
- 2- اجراء دراسات إحصائية في حالة وجود قيم شاذة ومشكلة تعدد خطي واجراء دراسات مقارنة بين الانحدار اللوجستي والتحليل المميز بعد ان تتم المعالجة من وجود الشواذ بأحد الطرائق الحصينة والارتباطات بين المتغيرات (بطريقة المربعات الصغرى الجزئية)
- 3- نوصي بتوسيع انحدار المربعات الصغرى الجزئية وذلك في حالة عدد المتغيرات التوضيحية اكبر من عدد المشاهدات ومن ثم استعمالها في المقارنة بين الانحدار اللوجستي والدالة المميزة الخطية سواء كانت خطية او تربيعية.



التحليل المميز والانحدار اللوجستي بوجود مشكلة التعدد الخطي [دراسة تطبيقية على مرض فقر الدم]

13-المصادر

- 1-البكري، رباب عبد الرضا (2015)، مقارنة بعض الطرائق الخطية لمعالجة مشكلة التعدد الخطي في النماذج مع تطبيق عملي، رسالة دكتوراه- كلية الإدارة والاقتصاد - جامعة بغداد.
- 2-التميمي، رعد فاضل حسن، (2013)، "الانحدار والسلاسل الزمنية أساليب إحصائية تطبيقية متقدمة باستخدام برنامج Minitab، كتاب، كلية الإدارة والاقتصاد، الجامعة المستنصرية.
- 3-الحمداني، بسمة رشيد، 2014، "تميز الملاك الطبي بحسب معرفتهم للتصنيف الدولي ((ICD-10 باستخدام الدالة المميزة"، رسالة ماجستير في جامعة بغداد، كلية الإدارة والاقتصاد.
- 4-عباس، علي خضير، 2012، " استخدام أنموذج الانحدار اللوجستي في التنبؤ بالدوال ذات المتغيرات الاقتصادية التابعة النوعية"، مجلة كركوك للعلوم الادارية والاقتصادية، مجلد 2، العدد 2.
- 5-Abdi, hervi, 2010, " Partial least squares regression and projection on latent structure regression (PLS Regression)", John Wiley & Sons.
- 6-Abdelmounaim Kerkri, Zoubir Zarrouk, Jelloul ALLAL, "A comparison of NIPALS algorithm with two other missing data treatment methods in a principal component analysis" University Mohamed.
- 7-Boaz Nadler, Ronald R.Coifman, 2005, " Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis", Department of Mathematics, Yale University.
- 8-Erik Brorson, Asterios Geroukis, 2014, "A comparison between discriminant analysis and logistic regression using principal components", Department of Statistics, Uppsala University, Uppsala University.
- 9-Leo H. Chiang, Evan L. Russell, Richard D. Braatz,2000, " Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis", Department of Chemical Engineering, university of Illinois.
- 10- Tormod Ness and Bjorn-Helge Mevik, "Understanding the collinearity problem in regression and discriminant analysis", Journal of Chemo metrics, P (413-426), 2001.



**discriminate analysis and logistic regression existence of multicolleniarty
problem(Empirical Study on Anemia)**

Abstract

The method binery logistic regression and linear discriminant function of the most important statistical methods used in the classification and prediction when the data of the kind of binery (0,1) you can not use the normal regression therefore resort to binary logistic regression and linear discriminant function in the case of two group in the case of a Multicollinearity problem between the data (the data containing high correlation) It became not possible to use binary logistic regression and linear discriminant function, to solve this problem, we resort to Partial least square regression.

In this, search the comparison between binary logistic regression and linear discriminant function using error Category. In the practical side in the collection of data on the data on anemia collection Two variables are severe anemia (0) and and chronic anemia (1) and several variables about the disease. The Data were collected from several Iraqi hospitals, where samples collected from patients at the hospital are asleep, and previous cases lay in the hospital a sample of (140) the patient is infected with the disease. When the test data and found that Multicollinearity problem, It has been processed using a method partial least square. The research found that linear discriminant function It is the best in the classification of data from binary logistic regression classified as linear discriminant function the data correctly and more accurate than binary logistic regression.

Keyword: linear discriminant function- binary logistic regression- partial least square– multicollinearity problem – ratio of classification.