تصحيح التحيز المخفي لخطأ التصنيف في نماذج الانحدار مع التطبيق

فاطمة كاظم موسى ابتسام كريم عبدالله

جامعة بغداد / كلية الادارة والاقتصاد / قسم الإحصاء

المستخلص

تناولت هذه الرسالة مشكلة خطأ التصنيف في نماذج الانحدار، وهي إحدى الإشكاليات الشائعة في التحليل الإحصائي، حيث يؤدي تصنيف البيانات بشكل غير صحيح إلى استعمال انموذج غير ملائم، مما ينعكس سلبًا على دقة التقديرات والتنبؤات. يظهر هذا الخطأ عندما يتم التعامل مع المتغيرات بطريقة غير مناسبة، مثل تصنيف متغير مستمر على أنه متغير فئوي، مما يؤدي إلى تحيز في النتائج وعدم دقة في الاستدلالات الإحصائية المستخلصة من الانموذج.

تهدف الرسالة إلى دراسة أثر خطأ التصنيف على دقة النماذج الإحصائية، واستعمال الأساليب المناسبة لاكتشاف هذا الخطأ، وتقدير المعلمات بدقة، وتحليل تأثير الخطأ على جودة النتائج المستخلصة من النماذج الإحصائية.

تناولت الرسالة خطأ التصنيف، حيث تم مناقشة تأثير التصنيف الخاطئ على تقدير المعلمات الإحصائية، ويستعرض الأساليب المختلفة المستعملة لتصحيح هذا التحيز. كما يتم تسليط الضوء على الفرضيات الأساسية لنماذج الانحدار المصنفة بشكل خاطئ، وتأثير أخطاء القياس على العلاقات الإحصائية بين المتغيرات.

اعتمدت الدراسة على مراجعة الأدبيات العلمية المذكورة آنفاً التي تناولت هذه المشكلة، بالإضافة إلى إجراء تطبيقات عملية على بيانات تم جمعها من مستشفى الطفل المركزي التعليمي في محافظة بغداد والتي تتضمن 100 حالة مصابة بمرض ابيضاض الدم الليمفاوي الحاد (Acute Lymphoblastic Leukemia – ALL) ، حيث تم تحليل هذه البيانات بأستعمال نماذج إحصائية متقدمة للتحقق من مدى تأثير خطأ التصنيف (Classification Error) على دقة النتائج والتنبؤات. بالإضافة إلى استعمال المحاكاة الإحصائية لدراسة تأثير أخطاء التصنيف في ظروف مختلفة، مما سمح بفحص أداء النماذج المختلفة تحت سيناريوهات متنوعة من الخطأ والتباين.

أظهرت نتائج التحليل والمحاكاة أن التصنيف غير الصحيح يؤدي إلى تحيزات واضحة في تقديرات المعلمات، مما يؤثر في كفاءة الانموذج في التنبؤ واتخاذ القرارات الإحصائية. وكما تبين أن إهمال معالجة هذا الخطأ قد يؤدي إلى استنتاجات مضللة تؤثر في دقة النماذج المستعملة في الدراسات التطبيقية. بناءً على ذلك، توصي الدراسة بضرورة التدقيق في طبيعة البيانات قبل تصنيفها، واستعمال أساليب تحليل متطورة لتصحيح الأخطاء المحتملة، بما يضمن تحسين جودة النماذج الإحصائية وزيادة موثوقية نتائجه.

مفاتيح الكلمات: الانحدار المصنف بشكل خاطئ ، تأثيرات التصنيف الخاطئ ،طريقة المربعات الصغرى المعدلة لتقليل التحيز ، طريقة المربعات الصغرى المعدلة خلال طريقة المربعات الصغرى العادية.

1-المقدمة

يُعتبر خطأ التصنيف من المشاكل الإحصائية الشائعة التي تنشأ عند تصنيف البيانات بشكل غير صحيح، مما يؤدي إلى استعمال نماذج انحدار غير دقيقة. يحدث هذا الخطأ نتيجة لعدة أسباب، منها عدم وضوح معايير التصنيف، وجود أخطاء في عملية جمع البيانات، أو تبني افتراضات غير صحيحة حول طبيعة المتغيرات المستخدمة في الانموذج. وعندما يكون هناك خطأ في تصنيف البيانات، يصبح تقدير معلمات الانموذج متحيزاً ، مما يؤثر على دقة التقديرات الإحصائية ويقلل من موثوقية النتائج.

قد ينتج خطأ التصنيف في نماذج الانحدار غالبًا عن سوء فهم طبيعة البيانات، حيث يتم التعامل مع المتغيرات بشكل غير صحيح، مثل تصنيف متغير مستمر على أنه متغير فئوي، أو اختيار انموذج غير ملائم لنوع البيانات المتاحة. يؤدي هذا الخطأ إلى ضعف دقة التنبؤات، حيث تصبح التوقعات أقل موثوقية، وقد تعطي تحليلات غير دقيقة أو حتى مضللة، مما قد يؤثر على عملية اتخاذ القرار الإحصائي وزيادة معدل الأخطاء في التقديرات.

لمعالجة خطأ التصنيف، يتم استعمال تقنيات تصحيح مثل تقدير احتمالات التصنيف الخاطئ (Probabilities of Misclassification) وتطبيق طرق مثل المربعات الصغرى المعدلة، التي تهدف إلى تقليل التحيز وتحسين جودة التقديرات. علاوة على ذلك، فإن التنبؤ بأستعمال انموذج متأثر بخطأ التصنيف يكون أقل دقة، مما قد يؤدي إلى اتخاذ قرارات خاطئة بناءً على نتائج مضللة. لذلك، من الضروري فحص البيانات بعناية قبل تصنيفها والتأكد من ملاءمة النموذج لطبيعة البيانات المستخدمة.

2-الاستعراض المرجعي

- في عام (2006) درس الباحث (Mahajan) مشكلة التعريف والتقدير في نماذج الانحدار غير المعيارية عندما يكون هناك متغير ثنائي مصنف بشكل خاطئ، حيث قد يرتبط خطأ القياس بالمتغيرين. أظهرت الدراسة أن دالة الانحدار تصبح غير معيارية في حال وجود متغير عشوائي إضافي يرتبط بالمتغير الأساسي الحقيقي غير المرصود، ولكنه لا يرتبط بخطأ القياس. تبع ذلك تعريف وظائف الانحدار شبه المعيارية والمعيارية استنادًا إلى النتيجة الأساسية التي تم التوصل إليها. اقترح الباحث تقديراً أساسياً يعتمد على استراتيجية التعريف، ودرس خصائص العينة الكبيرة لهذا التقدير، كما ناقش إجراءات تقدير بديلة. بالإضافة إلى ذلك، اقترح اختباراً لتصنيف الأخطاء في الانموذج يعتمد على قيد الاستبعاد الذي يسهل تنفيذه.[3]
- في عام (2008) اقترح الباحث (Brachet) نهجًا يعتمد على الأساليب المعلمية الحديثة لتصحيح تحيز التصنيف الخاطئ في المتغير التابع الثنائي، مما يسمح بالحصول على تقديرات متسقة لمعاملات انموذج المرحلة الثانية.[1]
- في عام (2017) استنتج الباحثان (Meyer& Mittag) ان التحيز المقارب ضمن النماذج المعلمية التي تسمح بربط أخطاء التصنيف بالعوامل القابلة وغير القابلة للملاحظة. أظهرت المحاكاة وبيانات التحقق أن تقديرات التحيز دقيقة حتى في العينات المحدودة، وغالبًا ما تؤدي إلى تحيز منحدر الانحدار الخطي نحو الصفر. كما تناولت الدراسة تحليل التحيز في تطبيق عملي، حيث تبين أن تقييد تباين التصنيف الخاطئ والمتغيرات المشتركة يؤدي إلى تفاقم التحيز لجميع المقدرات التي تم فحصها. وتوصلت النتائج إلى أن استخدام مقدرات تأخذ في الاعتبار انموذج التصنيف الخاطئ أو بيانات التحقق يساعد في تقليل التحيز وتحسين دقة التقديرات.[2]

3-انموذج الانحدار المصنف بشكل خاطئ

نفترض انموذج الانحدار الخطي المتعدد ، للمشاهدة i في العينة عشوائية بحجم n ، يكتب في الشكل التالي:

$$Y_i = a + \beta D_i^* + X_i'\delta + \varepsilon_i,\ \mathbb{E}[\varepsilon_i|X_i,D_i^*] = 0 \qquad (1)$$

حيث ان :

Y_i: متغير الاستجابة.

X_i: موجه المتغيرات التفسيرية و المقاسة بشكل صحيح وابعاده (K*1) وهو متغير السيطرة.

D_i^*: هو متغير ثنائي ويكون$D_i^* \in \{0,1\}$. يأخذ 1 عند الاستجابة للعلاج وصفر في حال عدم الاستجابة للعلاج.

ε_i: حد الخطأ العشوائي بمتوسط صفر وغير مرتبط مع X_i و D_i^* .

a : معلمة الانموذج وهي معلمة التقاطع.

β: معلمة الانموذج وهي تقيس تأثير العلاج.

δ: معلمة الانموذج.

لا يلاحظ الباحث الوضع الحقيقي من خلال D_i^* ولكن من المحتمل ان يكون بديل ثنائي قد تم تصنيفه بشكل خاطئ D_i .

يمكن ان نكتب D_i وفق الصيغة :

$$D_i = D_i^* + U_i \qquad (2)$$

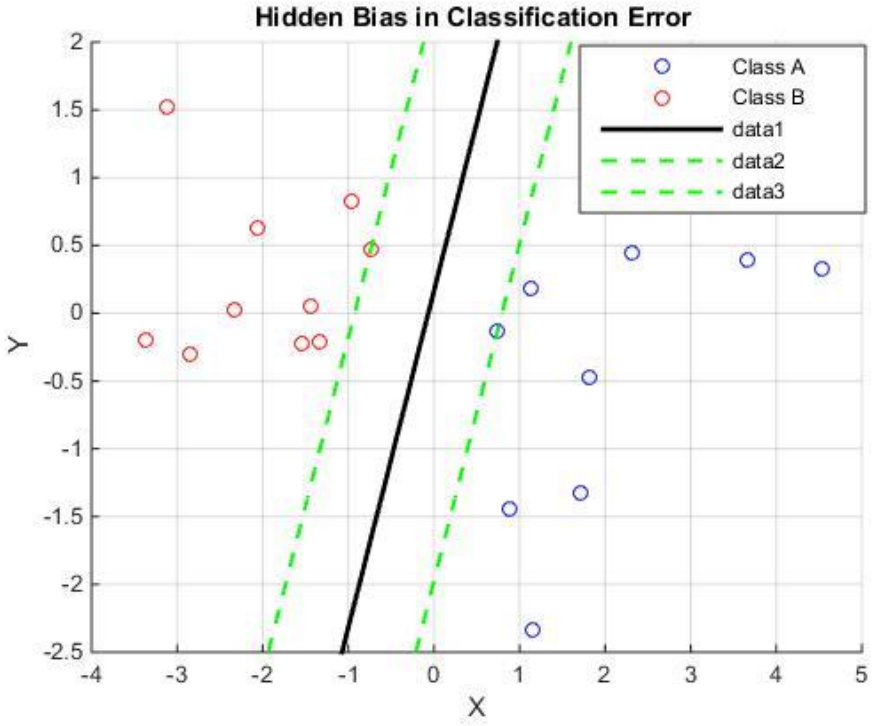
بحيث ان :

U_i: يدل على خطأ القياس ويكون مستقل عن ε_i بحيث ان $U_i \in \{-1,0,1\}$.

ان D_i^* يرتبط بالضرورة سلبيا بخطأ القياس U_i، بالتعويض المعادلة (1) في المعادلة (2) نحصل على:

$$Y_i = a + \beta D_i + X_i'\delta + (\varepsilon_i - \beta U_i) \qquad (3)$$

ان الطبيعة غير الكلاسيكية لخطأ القياس في D_i تجعل تقدير معلمات الانموذج امرا صعبا. ويمكننا توضيحه بالرسم التالي :



شكل رقم (1) يوضح التحيز المخفي لخطا التصنيف

حيث ان الخط الفاصل باللون الاسود يفصل بين البيانات المصنفة بشكل خاطئ والبيانات المصنفة بشكل صحيح وتشير الخطوط المقطعه باللون الاخضر الى هامش حد الخطأ بينما النقاط باللون الاحمر هي النقاط التي تقع داخل الهامش ولكن تم تصنيفها بشكل خاطئ.

4- تقدير المربعات الصغرى المعدلة حسب التحيز

اقترح انه مع وجود مقدرين متسقين ل $Cov(D_i, U_i)$ و $Cov(X_i, U_i)$ يمكننا بسهولة تصحيح تحيزات المربعات الصغرى العادية عن طريق عكس الحدود الموجودة على الجانب الايمن من هذه المعادلات حول معلمة β و معلمة δ .

يتم اعطاء التحيزات لمقدري المربعات الصغرى العادية للمعلمة β و للمعلمة δ و كالاتي:

$$plim\begin{bmatrix}\hat{\beta}_{ols}-\beta\\ \hat{\delta}_{ols}-\delta\end{bmatrix}=\beta\begin{bmatrix}Var(D_i) & Cov(D_i,X_i)\\ Cov(X_i,D_i) & Var(X_i)\end{bmatrix}^{-1}\begin{bmatrix}-\varphi Var(D_i)\\ \theta(X_i,D_i)\end{bmatrix} \quad (4)$$

حيث ان φ و θ يعتمدان على α_0 و α_1.

تبين المعادلة (4) التحيزات في تقدير المربعات الصغرى العادية للانموذج في المعادلة (3). عندما يكون $Cov(D_i,X_i)=0$ او اذا لم تكن هناك متغيرات مساعدة في الانموذج، فأن $\hat{\delta}_{ols}$ تكون متسقة مع δ لكن $plim\,\hat{\beta}_{ols}=\beta\left(1-\frac{Cov(D_i,U_i)}{Var(D_i)}\right)=\beta(1-\varphi)$. [1](Aigner,1973)

ثم اقترح الباحث (Aigner,1973) مقدر المربعات الصغرى المعدلة لتصحيح تحيز مقدر المربعات الصغرى العادية والذي يتكون من قسمة المربعات الصغرى العادية على التحيز. ولكن اذا كان $Cov(D_i,X_i)\neq 0$ سوف يكون $Cov(D_i^*,X_i)\neq 0$ سوف يكون كل من مقدر المربعات الصغرى العادية ومقدر المربعات الصغرى المعدلة لمعلمة β و لمعلمة δ غير متسقة بسبب الحدود التي تتضمن المعلمة θ في الجانب الايمن من المعادلة (4). حيث تولد هذه الحدود تحيزا تم تجاهله الى حد كبير ويسمى بالتحيز المخفي. ومن المثير للاهتمام ان هذا المكون لا يولد فقط تحيزا لمقدر المربعات الصغرى العادية ، ولكنه قد يؤدي الى تحيز عكس الاشارة لكل من مقدر المربعات الصغرى العادية ومقدر المربعات الصغرى المعدلة وهذا واضح بشكل خاص لتقدير معلمة δ . ولإثبات ذلك نفترض ان β موجبة حيث يمكن كتابته بالشكل التالي :[3]

$$plim\,\hat{\delta}_{ols}=\delta+\beta(\theta+\varphi)\Delta^{-1}Cov(X_i,D_i) \quad (5)$$

$$plim\,\hat{\delta}_{Mls}=\delta+\beta\varphi\Delta^{-1}Cov(X_i,D_i) \quad (6)$$

حيث ان:

$$\Delta=Var(X_i)-Cov(X_i,D_i)Var(D_i)^{-1}Cov(D_i,X_i) \quad (7)$$

للحصول على مستويات عالية بما فيه الكفاية من التقارير الخاطئة والارتباط الايجابي بين X_i و D_i^* ، قد يكون $plim\,\hat{\delta}_{ols}$ موجبا حتى لو δ تكون سالبة ، على سبيل المثال في انموذج العودة الى الكلية اذا كان التخرج من الجامعة مرتبطا ارتباطا وثيقا بالاناث وكان خطأ التصنيف مرتفعا ، فقد تستنتج مقدر المربعات الصغرى العادية ومقدر المربعات الصغرى المعدلة خطأ ان النساء يكسبن اكثر من الرجال في المتوسط (وهو تناقض مع فجوة الاجور المعروفة في اقتصاد العمل).

سوف نحدد مقدرات المربعات الصغرى المعدلة بالتحيز وكالتالي:

$$\begin{bmatrix}\hat{\beta}\\ \hat{\delta}\end{bmatrix}=\begin{bmatrix}(1-\varphi)S_{DD} & S_{DX}\\ (1+\theta)S_{XD} & S_{XX}\end{bmatrix}^{-1}\begin{bmatrix}S_{YD}\\ S_{YX}\end{bmatrix} \quad (8)$$

حيث ان كل من φ و θ يعتمد على α_0 و α_1 وعليه:

1-تكون هذه التقديرات متسقة اي ان $\hat{\beta}\xrightarrow{p}\beta$ و $\hat{\delta}\xrightarrow{p}\delta$.

2-بالاضافة الى ذلك ان $\mathbb{E}[\varepsilon_i^2|X_i,D_i^*]=\sigma^2$ حيث تكون هذه المقدرات بشكل شبه طبيعية .

ويمكننا ايضا كتابة مقدر المربعات الصغرى المعدلة من خلال مقدر المربعات الصغرى العادية وكالتالي:

$$\begin{bmatrix}\hat{\beta}_{MLS}\\ \hat{\delta}_{MLS}\end{bmatrix}=\begin{bmatrix}\hat{\beta}_{ols}-\beta\\ \hat{\delta}_{ols}-\delta\end{bmatrix}+\hat{\beta}_{ols}\begin{bmatrix}(1-\varphi)S_{DD} & S_{DX}\\ S_{XD} & S_{XX}\end{bmatrix}^{-1}\begin{bmatrix}\varphi S_{DD}\\ 0\end{bmatrix} \quad (9)$$

عندما تكون المقدرات كما في المعادلات السابقة يمكننا بسهولة ملاحظة عندما لا يكون هناك خطأ في القياس اي ان $\alpha_0=0$ حيث ان $\varphi=0$ و $\theta=0$ بحيث تكون مقدرات المربعات الصغرى المعدلة بالتحيز ومقدرات المربعات الصغرى

المعدلة ومقدرات المربعات الصغرى العادية متساوية مع بعضها اي ان $\hat{\beta} = \hat{\beta}_{MLS} = \hat{\beta}_{ols}$ و $\hat{\delta} = \hat{\delta}_{MLS} = \hat{\delta}_{ols}$ وتكون كل هذه المقدرات متسقة. عندما يوجد خطأ التصنيف يكون X و D^* غير مرتبطين ، ان مقدر فرضية المربعات الصغرى المعدلة بالتحيز ومقدر مربعات الصغرى المعدلة الخاصة بالباحث (Aigner,1973) كلاهما متسق ومكافئ بشكل مقارب ، لكن مقدرات المربعات الصغرى العادية غير متسق. عندما يكون هناك خطأ التصنيف(*classification error*) و يكون X و D^* مرتبطين سوف يكون فقط مقدر المربعات الصغرى المعدلة بالتحيز متسق وان مقدر المربعات الصغرى المعدلة ومقدر المربعات الصغرى العادية غير متسقة. اذا تم تقدير β و δ بشكل متسق فيمكن الحصول على مقدر متسق لمعلمة التقاطع a . سوف يتم تقدير معلمة التقاطع a بأستعمال طريقة المربعات الصغرى المعدلة بالتحيز وكما يلي:

$$\hat{a} = \bar{Y} - \hat{\beta}\frac{\bar{D}-\alpha_0}{1-\alpha_0-\alpha_1} - \bar{X}'\hat{\delta} \qquad (10)$$

حيث ان $\hat{a}$ متسق وطبيعي بشكل تقريبي اي ان $\hat{a} \xrightarrow{p} a$ و $\sqrt{n}(\hat{a}-a) \xrightarrow{d} N(0,\sigma_a^2)$

5- الجانب التجريبي

***المحـاكاة**

في عصر البيانات الضخمة، يُعدُّ توليد البيانات عملية أساسية تُستخدم في مجالات متعددة، مثل الذكاء الاصطناعي، وتحليل البيانات، والبحث العلمي. يهدف توليد البيانات إلى إنشاء مجموعات بيانات اصطناعية أو محاكاة بيانات واقعية لاختبار النماذج، تحسين أداء الأنظمة، أو تعويض نقص البيانات الفعلية.

تم توليد بيانات محاكاة بمستويات مختلفة من التباين ($\sigma^2 = 0.3, 0.6, 0.9$) وأحجام عينات ($n = 50, 100, 150$) بأستعمال برنامج (MATLAB, 2012a) وبتكرار 1000 مرة للانموذج ($Y_i = a + \beta D_i^* + X_i'\delta + \varepsilon_i$) لفحص سلوك كل طريقة. بالإضافة إلى ذلك، تم تحليل بيانات فعلية تتعلق بـ 100 حالة مرضية من مستشفى الطفل المركزي التعليمي.

***نتائج المحاكاة بأستخدام معيار RMSE**

1- اظهرت طريقة المربعات الصغرى الاعتيادية (OLS) تفوقا واضحا عندما كان حجم العينة 50 او 100 في معظم الحالات حيث سجلت اقل قيم RMSE مقارنة بالطرائق الاخرى، خصوصا عند $\sigma^2 = 0.3$ و $\sigma^2 = 0.6$.

2- ان طريقة المربعات الصغرى المعدلة بالمربعات الصغرى الاعتيادية (MLS) لم تكن الافضل اداء في اي من الحالات ، حيث سجلت RMSE اعلى مقارنة بالطرائق الاخرى .

3- اظهرت طريقة المربعات الصغرى المعدلة حسب التحيز ذات المرحلتين (BALS) تطوراً ملحوظاً عند زيادة حجم العينة الى 150 حيث اصبحث اقل قيمة RMSE .

6-الجانب التطبيقي

يعتبر مرض ابيضاض الدم الليمفاوي الحاد (Acute Lymphoblastic Leukemia - ALL) أحد أكثر أنواع السرطانات شيوعًا في الأطفال، ويتميز بنمو غير طبيعي للخلايا الليمفاوية في الدم ونخاع العظم. يتميز هذا المرض بنمو خلايا الدم البيضاء بشكل مفرط مما يعيق قدرة الجسم على إنتاج خلايا دم طبيعية، ويؤدي إلى مجموعة متنوعة من الأعراض مثل فقر الدم، نزيف، وعدوى متكررة. على الرغم من التقدم الكبير في التشخيص والعلاج، لا يزال مرض ابيضاض الدم الليمفاوي الحاد يشكل تحديات كبيرة في الرعاية الصحية، وخاصة في ما يتعلق بالتشخيص المبكر، التنبؤ بالتطورات المرضية، وتحديد أفضل استراتيجيات العلاج.

سوف نتطرق في هذا الفصل الى الجانب التطبيقي ، الذي تم في ايجاد المقدرات وتحديد خطا التحيز المخفي. باستعمال بيانات تم جمعها من مستشفى الطفل المركزي التعليمي وتشمل المتغيرات الاتية:

كمية الجرعة (Cytosar) هو دواء مضاد للسرطان يستخدم لعلاج انواع معينة من سرطان الدم مثل ابيضاض الدم الليمفاوي الحاد (ALL) تعتمد جرعة السايتوسار على عدة عوامل منها العمر ونوع السرطان المعالج وبروتوكول العلاج المستخدم، ويعتبر المتغير المعتمد في بحثنا هذا.

عدد خلايا الدم البيضاء (WBC) الطبيعي للأطفال الذين تتراوح اعمارهم بين السنة و 5 سنوات يتراوح بين 5،000 الى 15,000 خلية لكل ميكرولتر في الدم والذين تتراوح اعمارهم بين 6 سنوات الى 17 سنة حوالي 4،500 الى 13،500 خلية لكل ميكرولتر في الدم وهذه القيم قد تختلف احصائياً بناءً على المختبرات والمصادر الطبية المستخدمة، وهو المتغير المستقل.

عدد الصفائح الدموية (Platelet Count) الطبيعي للأطفال الذين تتراوح أعمارهم بين السنة و15 سنة يتراوح عادةً بين: 150,000 إلى 450,000 صفِيحة لكل ميكرولتر من الدم قد تختلف القيم الطبيعية للصفائح الدموية بين المختبرات لكن النطاق المرجعي الطبيعي يتراوح بين 150,000 و450,000 صفِيحة لكل ميكرولتر من الدم. في حال كان العدد خارج هذا النطاق ، يُوصَى بمراجعة الطبيب لتقييم الأسباب المحتملة وتحديد الحاجة إلى متابعة طبية استنادا إلى التوزيع الإحصائي للقيم الطبيعية، وهو متغير مستقل.

العلاج وهو المتغير الثنائي ويقصد فيه هل استجاب المريض الى العلاج او لا.

الجنس وهو متغير مستقل يتم تحديد فيه جنس المريض.

العمر وهو متغير مستقل يتم تحديد فيه عمر المريض. وباستعمال برنامج MATLAB توصلنا الى الاتي:

جدول رقم (3-1) يوضح فيه الوسط الحسابي والانحراف المعياري للمتغير المعتمد والمتغيرات المستقلة وقيمة الارتباط لكل متغير مع المتغير الثنائي

المتغير	الوسط الحسابي	الانحراف المعياري	الارتباط مع العلاج
الجرعة	24.42	6.74	**0.0022**
العلاج	0.34	0.47	**1**
الخلايا	6023.07	1.51	**0.10**
الصفائح	166940.96	1.16	**0.47**
الجنس	1.5	0.50	**0.16**
العمر	6.12	3.77	**0.05**

نلاحظ من الجدول اعلاه ان كمية الجرعة والعمر لها ارتباط منخفض مع استجابة العلاج بينما نلاحظ كمية الصفائح والجنس وكمية الخلايا لها ارتباط اقوى مع استجابة العلاج، وبما ان التحيز المخفي مرتبط بمدى الارتباط بين المتغير الثنائي المصنف بشكل خاطئ والمتغيرات الاخرى، فان هذه البيانات تكون ملائمة لبحثنا هذا.

جدول رقم (3-2) يوضح فيه المعلمات المقدرة ودرجة المعنوية لكل معلمة

المتغير	ols	SE	mls	SE	bals	SE
العلاج	1.84	0.93	2.76	0.97	2.36	**0.75**
الخلايا	0.00011	0.000069	0.00012	0.000077	-0.00005	**0.000067**
الصفائح	-0.0000041	0.000009	-0.0000043	0.37	-0.00000026	**0.0000087**
الجنس	-0.78	0.43	0.63	**0.26**	-0.94	**0.37**
العمر	0.12	0.27	-0.0079	**0.025**	0.11	**0.26**
الحد الثابت	16.52	0.000064	13.13	0.000071	16.38	**0.000062**

العلاج: القيم المقدرة عبر الطرائق الثلاثة قريبة من بعضها، لكن طريقة BALS أظهرت أقل خطأ معياري (SE = 0.75)، مما يشير إلى دقة أعلى في التقدير.

الخلايا: القيم المقدرة سالبة وصغيرة جداً، مما يدل على تأثير ضعيف أو غير معنوي إحصائيًا، مع اختلاف بسيط بين الطرائق الثلاثة.

الصفائح: القيم المقدرة سالبة في بعض الطرائق، مما يشير إلى احتمال وجود تحيز في البيانات، حيث أن التأثير الحقيقي المتوقع إيجابي لكن يظهر سالباً بسبب التحيز.

الجنس: القيم المقدرة سالبة في معظم الطرائق، مما يشير إلى أن التحيز قد يؤدي إلى عكس التأثير الحقيقي.

العمر: هناك تباين بسيط بين الطرائق الثلاثة، لكن التأثير الفعلي يبدو ضعيف نسبياً

الحد الثابت: يظهر تباين بسيط بين الطرائق، مما قد يعكس تأثير المتغيرات المفقودة أو التحيز في الانموذج.

7- الاستنتاجات

1- أظهرت الدراسة أن خطأ التصنيف في نماذج الانحدار يؤدي إلى تحيزات واضحة في تقديرات المعلمات، مما يؤثر بشكل مباشر في دقة النماذج الإحصائية المستخدمة في التنبؤ والتحليل.

2- طريقة المربعات الصغرى الاعتيادية (OLS) كانت الأكثر كفاءة عند أحجام عينات صغيرة ومتوسطة، حيث سجلت أقل قيم لمعيار RMSE مقارنةً بطرق التقدير الأخرى.

3- عند زيادة حجم العينة إلى 150، تفوقت طريقة المربعات الصغرى المعدلة بالتحيز ذات المرحلتين (BALS) في تحقيق أقل قيم لـ RMSE، مما يشير إلى أنها أكثر استقرارًا عند العينات الكبيرة.

4- طريقة المربعات الصغرى المعدلة حسب المربعات الصغرى الاعتيادية (MLS) لم تكن الأفضل في أي حالة، مما يشير إلى أنها تتأثر بشكل كبير بتغيرات التباين وتصنيف البيانات.

8-التوصيات

1- عند التعامل مع حجم عينة صغير أو متوسط، يُفضل استعمال طريقة المربعات الصغرى المعدلة بالتحيز نظرًا لكفاءتها في تحقيق أقل خطأ تقديري.

2- عند زيادة حجم العينة، يُوصى باستعمال طريقة المربعات الصغرى المعدلة بالتحيز ذات المرحلتين BALS نظرًا لدقتها الأعلى في التقدير وتقليل التحيز الناتج عن التصنيف الخاطئ.

3- تجنب استعمال طريقة المربعات الصغرى المعدلة بالمربعات الصغرى الاعتيادية كطريقة تقدير رئيسية دون معالجة أخطاء التصنيف، حيث أظهرت أداءً ضعيفًا مقارنة بالطرائق الأخرى.

4- تعزيز التعاون بين الباحثين والكوادر الطبية لضمان فهم طبيعة البيانات المطلوبة للأبحاث، بحيث يكون هناك إرشادات واضحة لكيفية تصنيف البيانات الطبية بطريقة متوافقة مع أساليب التحليل الإحصائي الحديثة.

9-المصادر

[1] Brachet, T. (2008), 'Maternal Smoking, Misclassification, and Infant Health.'

Available at SSRN: https://ssrn.com/abstract=1009781.

[2] Meyer, B. D. & Mittag, N. (2017), 'Misclassification in binary choice models,'

Journal of Econometrics 200(2), 295–311.

[3] Mahajan, A. (2006), 'Identification and estimation of regression models with misclassification', Econometrica 74(3), 631–665.

Journal of [4] Meyer, B. D. & Mittag, N. (2017), 'Misclassification in binary choice models', Econometrics 200(2), 295–311.

[5] Aigner, D. J. (1973), 'Regression with a binary independent variable subject to errors of observation', Journal of Econometrics 1(1), 49–59

Abstract

This research deals with the problem of misclassified regression, which is one of the common problems in the field of statistical analysis. This error occurs when data is incorrectly classified, leading to the selection of an inappropriate regression model, thus affecting the accuracy of estimates and predictions. The research aims to analyze the effects of misclassification of statistical models and explore the tools and techniques that can be used to detect and correct it. In this study, previous literature on misclassification in regression models was reviewed, and a statistical model was developed to identify and correct this error using advanced statistical methods, such as estimating the probability of misclassification and applying modified least squares to reduce bias in estimates. The thesis also provides a theoretical and methodological analysis of how this error affects the conclusions drawn from statistical models. The results showed that misclassification leads to deviations in parameter estimates, which affects the accuracy of models used in prediction and analysis. Based on these results, the study recommends the need to scrutinize the nature of the data and its correct classification before choosing the appropriate statistical model, and it also recommends the use of advanced statistical techniques to analyze and correct potential errors in classification.