تطبيق أسلوب معادلات التقدير المعممة لتقدير نموذج الانحدار اللوجستي الشرطي للقياسات المكررة

أ.م.د. انتصار عريبي فدعم / كلية الادارة والاقتصاد / جامعة بغداد
الباحث/ يوسف خليل الخفاجي

المستخلص

غالباً ما يستخدم الانحدار اللوجستي الشرطي لدراسة العلاقة بين نتائج حدث ما وعوامل تشخيصية محددة من اجل تطبيق الانحدار اللوجستي والاستفادة من قدراته التنبؤية في الدراسات البيئية. ويهدف هذا البحث إلى اثبات نهجاً جديداً لتطبيق الانحدار اللوجستي الشرطي في البحوث البيئية من خلال طرق الاستدلال المبنية على البيانات الطولية. وبذلك يتطلب التحليل الإحصائي للبيانات الطولية الأساليب التي يمكن أن تراعي بشكل سليم للترابط داخل العناقيد لقياسات الاستجابة. إذا تم تجاهل هذا الارتباط فان استدلالات مثل الاختبارات الإحصائية وفترات الثقة يمكن أن تكون غير صالحة إلى حد كبير.

ولغرض تقدير نموذج الانحدار الشرطي بتحليل تلوث البيئة كدالة لإنتاج النفط والعوامل البيئية تم استخدام نهج معادلة التقدير المعممة (GEE) في صياغة طرق الاستدلال التي من شأنها تسهيل نموذج الانحدار اللوجستي الشرطي بالاستفادة من الارتباطات الفعلية بين الاستجابات في البيانات وكذلك بنية الارتباط المحددة من خلال مقدرات الشطيرة الحصينة (RSE)، فضلاً عن تطبيق العديد من معايير اختيار النموذج المختلفة. لان كفاءة التقديرات تتوقف على مواصفات مصفوفة الارتباط العاملة، وان الاختيار المناسب لمصفوفة الارتباط العاملة يمكن ان يحرز تقدم كبير في كفاءة الاستدلال الإحصائي للـ GEE. وبعد مقارنة أداء المعايير المحددة تبين لنا أن QIC هو معيار الاختيار الأنسب لطريقة GEE، حيث أظهرت نتائج التطبيق أن QICله فقدان المعلومات الأدنى في ضمن طريقة GEE الذي كان الهدف لتطوير نموذج تنبؤي من مجموعة النماذج المرشحة، وقد ثبت أيضاً من خلال هذا البحث، ان شرط الانحدار اللوجستي يكون أداة فعالة يمكن استخدامها في دراسات أخرى لاستكشاف العلاقات بين متغيرات الاستجابة والتفسيرية.

المصطلحات الرئيسية للبحث/ الانحدار اللوجستي الشرطي، القياسات المكررة، طريقة معادلات التقدير المعممة (GEE)، مقدرات الشطيرة الحصينة (RSE)، معايير اختيار النموذج، بنية الارتباط العاملة، التحليل العنقودي.

مجلة العلوم
الاقتصادية والإدارية
العدد 96 المجلد 23
الصفحات 363- 383

*بحث مستل من رسالة ماجستير

1. المقدمة

أصبحت طرق الانحدار جزءاً لا يتجزأ من أي تحليل للبيانات المعنية بوصف العلاقة بين متغير الاستجابة وواحد أو أكثر من المتغيرات التفسيرية. وان متغير الاستجابة في كثير من الأحيان يكون متقطعا، آخذاً اثنين أو أكثر من القيم المحتملة. ففي الدراسات الإيكولوجية والبيئية عادةً ما تستكشف العلاقات المتبادلة المحتملة بين الخصائص البيئية ومتغير الاستجابة مع القيم أما صفر أو واحد (0، 1)، وأن الهدف من استخدام تحليل هذا النموذج لطبيعة ونوعية البيانات هو تحديد نوع النموذج المطلوب في حالة البيانات الطولية بوجود المشاهدات المكررة عبر فترة من الزمن. وانه في حالة الاستجابة الثنائية مع القياسات المكررة يكون النموذج الملائم لها هو نموذج الانحدار اللوجستي الشرطي. حيث يمكن استخدام نظرية معادلات التقدير المعممة Generalized Estimating Equations (GEE) لتنفي الآثار السلبية للمشاهدات المستقلة المترتبة من مجموعات متطابقة. ويعد اختيار نموذج ملائم من مجموعة نماذج مرشحة أحد اهم المشاكل الأساسية في نمذجة البيانات ويتطلب اختيار النموذج الذي يميز البيانات تحديد أفضل نموذج هيكلي ملائم وصريح للنموذج. كما أن تحديد بنية الارتباط العاملة working correlation structure (WCS) من خلال معايير اختيار بنية الارتباط العاملة مفيدة إلى حد كبير عند تحديد أكثر بنية قابلة للتطبيق بشكل معقول. حيث انها تؤثر بشكل كبير على كل من مقدرات النموذج والتنبؤ، ناهيك عن التفسيرات العلمية.

2. مشكلة البحث Research problem:

الانحدار اللوجستي لمتغير استجابة ثنائية (وجود تلوث 1، عدم وجوده 0) مع نمذجة البيانات الطولية المصنفة الى طبقات ومن ثم الى مجموعات فرعية أكثر تجانساً يعتبر من النماذج التي ستعالج مثل هكذا بيانات وان تقدير مثل هكذا نماذج يكون من الصعوبة حلها لوجود ارتباطات داخل المشاهدات ضمن الطبقات او العناقيد. وتحديد بنية الارتباط الصحيحة يسهم في إيجاد مقدرات موثوقة وغير متحيزة وأكثر تقارباً.

3. هدف البحث Research Goal:

الهدف الرئيس لهذا البحث هو:

أ- تحديد بنية الارتباط الصحيحة باستخدام أفضل معيار من بين عدة معايير مختلفة، مثل:
1. معيار معلومات شبه الامكان في ظل نموذج الاستقلال QIC
2. معيار معلومات الارتباط CIC
3. معيار روتنتيزكي وجيويل RJC
4. معيار غوشو DEW
5. معيار الامكان القياسي الزائف GPC
6. معايير معلومات أكايكي وبيز التجريبية EAIC و EBIC

ب- تقدير نموذج الانحدار اللوجستي باستخدام طريقة معادلات التقدير المعممة (GEE) لوضع نموذج تنبؤي مثالي يحتوي على أفضل متغيرات ذات اقل فقدان للمعلومات.

الفصل الاول /الجانب النظري

4. تعريف الانحدار اللوجستي الشرطي

ان الانحدار اللوجستي الشرطي في حالة البيانات الطولية هو تحليل البيانات الطولية الثنائية التي تحتوي على متنبئ واحد أو عدة متنبئات predictors، مع المشاهدات التي هي ليست مستقلة لكنها تكون مُجَمَّعة grouped.

بينما تَستَخدِم نماذج الانحدار الخطي العام التوزيع الطبيعي القياسي لمعالجة أخطاء الانحدار، وتستفيد نماذج الانحدار اللوجستي الشرطي من توزيعات ثنائي الحدين لمعالجة أخطاء الانحدار لبيانات ثنائية، مع نماذج ثنائي الحدين مثل normit / probit ، فضلاً عن Clog-log المستخدمة بشكل أقل شيوعاً. [1]

نظراً لان الارجحية تعطى بالصيغة (2.2)، حيث p هو الاستجابة النسبية، في أنها الاستجابة r لـ n من الاستجابات، حيث p = r /n ، فان Logit(p) = log (p/(1-p)).

ويمكن إثبات أهمية نماذج الانحدار اللوجستي باستخدام الفرق بين اثنين من اللوجيت:

$$logit_a - logit_b = \log\left(\frac{\hat{\pi}_a}{1-\hat{\pi}_a}\right) - \log\left(\frac{\hat{\pi}_b}{1-\hat{\pi}_b}\right)$$

$$logit_a - logit_b = \log\left(\frac{\hat{\pi}_a}{1-\hat{\pi}_a} \Big/ \frac{\hat{\pi}_b}{1-\hat{\pi}_b}\right)$$

$$logit_a - logit_b = odds\ ratio \qquad \dots (1.1)$$

حيث $\hat{\pi}_a$ هي احتمالات وقوع الحدث للمحاولات a و $\hat{\pi}_b$ احتمالات وقوع الحدث للمحاولات b. [22]

5. البيانات الطولية Longitudinal data

الخاصية الرئيسة للبيانات الطولية هي أنها تُسَهّل قياس التغييرات الحساسة للوقت فيما بين العينات، مما تتيح قياس الأحداث الدورية، فضلاً عن توقيت الأحداث لتحديد التحسينات المتأصلة بسبب تسجيل الأحداث الواقعة. [11] كذلك فان استقراء بيانات طولية في الماضي والحاضر يسمح لتحليل التأثيرات المترتبة على الحدث قبل وبعد وقوعه. من جهة أخرى فان التحقق من البيانات في الدراسات الاستطلاعية prospective studies يقلل التحيز بالتأكد من دقة قياس التغيرات في النتائج.

ايضاً تسهل البيانات الطولية تقسيم التأثيرات الوقتية وفقاً لعناقيد وفترات، عبر نطاقات زمنية متعددة. ويتطلب تحليل العلاقات بين البيانات الطولية الأساليب الإحصائية التي تقيس دقة العلاقات داخل المتغيرات التفسيرية مع متغيرات الاستجابة وبالتالي تؤثر في دقة الاستدلال للنموذج. ولهذا يتم اللجوء الى تقدير ما يسمى بالأخطاء المعيارية الحصينة robust standard errors أو تباينات الشطيرة Sandwich variances وهي تقنية تقدير تباين استدلال شبه معلمي التي تميز المعلمات وفقاً لجذور معادلات التقدير، وهو تقدير معلمة التباين الافتراضية لمعادلات التقدير المعممة (GEE).

وان تقديرات خطأ النموذج المعيارية تكون كفوءة بشكل عام عند مقارنتها بتقديرات على أساس الشطيرة في اشارة الى افتراضات النموذج على الرغم من أن معظم الافتراضات غير قابلة للاختبار، وقد تكون غير منطقية في العينات الكبيرة. ونتيجة لذلك، يتم استخدام تباين الشطيرة كبدائل لتحديد افتراضات النموذج. [12] وبالتالي يكون تقدير معلمة متعدد المتغيرات في معادلات التقدير غير متحيز بسبب تباين الشطيرة. وباستبدال تباين الشطيرة مع تقديرات التباين التجريبية Empirical variance estimates لمعادلات التقدير القائمة على أساس الامكان سوف يكشف الاختلافات التجريبية للاستدلال القائم على أساس النموذج.

6. التحليل العنقودي Cluster Analysis

هو عبارة عن نوع من الأساليب الإحصائية التي يمكن تطبيقها على البيانات التي تعكس أنماط مجموعات "طبيعية". اذ يتولى التحليل العنقودي إفراز البيانات الأولية وتجميعها في مجموعات عنقودية. اما العنقود فهو عبارة عن مجموعة من الحالات أو الملاحظات المتجانسة نسبيًا. وتتميز العناصر المكونة للعنقود الواحد بأنها متشابهة مع بعضها. كما أنها تختلف عن العناصر الأخرى، خصوصًا العناصر المكونة للعناقيد الأخرى. وهناك عدة طرائق تستخدم في التحليل العنقودي وتعمل هذه الطرائق على الرغم من اختلافها بمرحلية التجميع نفسها، وهي:

أ- طريقة التحليل العنقودي (K-Means).
ب- التحليل العنقودي الهرمي Hierarchical cluster analysis.
ت- التحليل العنقودي ذو خطوتين Two-step cluster analysis.

7. طريقة معادلات التقدير المعممة GEE

بحسب تصور Liang وZeger (1986)، تسعى GEE إلى نموذج متوسط الاستجابات بدلاً من نمذجة هياكل التباين المشترك بين المواضيع (العناقيد) inter-subject. كما لا يتعين بالضرورة على بنية التباين المشترك في GEE ان تكون محددة بشكل صحيح للحصول على استدلال معقول عن تقديرات معاملات الانحدار أو على الأخطاء المعيارية. والمبدأ الأساسي لافتراض الـ GEE ان μ_{it} هو نموذج المتوسط، بينما بنية التباين هي V_i عندئذ يمكن صياغة معادلة التقدير على النحو الاتي:

$$U(\beta) = \sum_{i=1}^{k} \left(\frac{\partial \mu_{it}}{\partial \beta_p}\right)' V_i^{-1}\{Y_i - \mu_i(\beta)\} \qquad \dots (2.1)$$

الصيغة $U(\beta)=0$ يتم حلها من خلال تقديرات المعلمة، حيث عادةً ما تستخدم خوارزمية نيوتن -رافسون لبلوغ تقديرات المعلمة. فضلا عن ذلك، فان بنية التباين تكون مهمة نظراً لأن اختيارها هو أمر أساسي لتحسين كفاءة تقديرات المعلمة. [7] وان شروط هذه الطريقة هي:

1. تتطلب طريقة GEE مواصفات محدودة للمتوسطات الحدية marginal means، اضافة الى التوزيع المشترك لهياكل التباين المشترك العاملة للقياسات المتكررة، مما يجعلها أكثر دقة بالمقارنة مع افتراضات دوال الإمكان.
2. تتطلب GEE المواصفات لمصفوفة الارتباط العاملة عند تحليل البيانات الطولية، بوصفها الترشيحات للتوزيع الاحتمالي المشترك.
3. تتطلب GEE العينات ذات الأحجام الكبيرة جداً نظراً لان الأخطاء المعيارية التجريبية بشكل عام تكون اقل من تقدير (underestimate) الأخطاء المعيارية الفعلية في العينات المعتدلة. [23]

8. مقدر الشطيرة Sandwich Estimator

في البداية تم وضع هذا المقدر في اطار التصور والادراك من قبل Huber (1967) وطورَ بعد ذلك بواسطة White (1980)، في حين استخدماه Liang وZeger (1986) في البيانات الطولية. [8] ويعرف أيضاً بأسم "مقدر التباين الحصين Robust variance estimator (RVE) خلافاً لتباين العينة له أداء جيد مع البيانات مع مختلف التوزيعات الاحتمالية غير الطبيعية، ومن ثم يمكن قياس التشتت الإحصائي لمجموعة البيانات العددية.

ويستخدم كإحصاءة حصينة فيما يتعلق بتقدير معلمة القياس، مما يجعله فعال للغاية خاصة مع بيانات ملوثة نظراً لتلوث واحد أو عدد قليل من المشاهدات لا تؤثر على التقديرات. [17] ويعبر عنه:

$$\left[X^T \widehat{W} X\right]^{-1} \left[\sum_i X_i^T (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T X_i\right] \left[X^T \widehat{W} X\right]^{-1} \qquad \dots (2.2)$$

هنا $\widehat{W}$ هو مصفوفة التباين المشترك المفترضة للتقدير الحصين او الشطيرة. وان مقدر الشطيرة يعطي تقدير Cov($\widehat{\boldsymbol{\beta}}$)متفوق ولاسيما في العينات الكبيرة على الرغم من القيمة الحقيقية لـ Cov($\boldsymbol{y_i}$).

لشبه الامكان، إذا كانت استجابات الـ $y_1, y_1, y_3, \ldots, y_N$ ذات صلة بالمتغيرات التوضيحية Covariates ، يمكن التعبير عن نموذج الانحدار الخطي بـ: $y \sim N(X_i^T\beta, \sigma^2)$ مع X_i كمتجه المتغيرات التوضيحية، و β متجه المعاملات ، و σ^2 تباين الخطأ. (21)

يمكن أن يعمم النموذج عن طريق إدخال دالة ربط متوسط $E(y_i) = \mu_i$ الناتجة عن دالة المتوسط $g(\mu_i) = X_i^T\beta$ وبالتالي $E(y_i) = \mu_i(\beta)$ بواسطة تضمين المتغيرات التفسيرية وكذلك الربط لـ μ_i. بعد ذلك يمكن إدخال عدم التجانس Heteroscedasticity عن طريق دالة التباين $\mathrm{Var}(y_i) = V_i$ وان V_i يتوقف على μ_i و β، ومعلمات مجهولة محتملة مثل عامل القياس □. ويمكن الحصول على تقدير أعظم امكان للـ β عن طريق تهديف فيشر Fisher scoring اذ ان تقديرات ML تكون غير متحيزة وكفوءة، وتتوقف على العزوم الأولى والثانية بدلاً من حالة y_i الطبيعية. (19)

وهذا يعني أنه إذا كانت دوال المتوسط والتباين صحيحة، مع توزيع y_i غير الطبيعي، فان تقدير الـ β من تعظيم لوغاريتم الإمكان الطبيعي سوف يبقى غير متحيز وكفوء من منظور تقاربي. وان تحقيق الحد الاعظم من لوغاريتم امكان الحالة الطبيعية normality log-likelihood بافتراض التوزيع غير الطبيعي للاستجابة، وان تقدير الـ β هو تقدير شبه الإمكان QL الذي يتم حسابه بشكل تكراري من خلال شبه التهديف Quasi-scoring.

9. تطبيق طريقة معادلات التقدير المعممة

يتم تطبيق طريقة GEE باستخدام دالة الكثافة الحدية للـ y_{it} التي يمكن التعبير عنها على النحو:

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - \alpha(\theta_{it}) + \mathrm{b}(y_{it})\}\phi] \quad \ldots (2.3)$$

حيث ان α وb تمثل صيغ دوال معرفة و ϕ قد تمثل معلمة القياس وربما تكون مجهولة، باعتبار ان قيم المتغيرات التوضيحية للمصفوفة $n_i \times k$ هي $X_i = (x_{i1}, x_{i2}, x_{i3} \ldots, x_{ini})^{\mathrm{T}}$ متصلة بمتجه نتائج الـ $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, \ldots, Y_{ini})^{\mathrm{T}}$ للعناقيد $i = 1,2,3, \ldots, K$ وأيضاً ملاصقة للمعلمات $\theta_{i1}, \theta_{i2}, \theta_{i3}, \ldots,$ ضمن الكثافة الحدية للـ Y_{it} عندما $\theta_{it} = \mathrm{h} * (x_{it}^{\mathrm{T}}\beta)$ ، حيث ان T هو المبدلة بينما $h *$ دالة محددة للربط، مع الاخذ بالحسبان حقيقة أن القيمة المتوقعة لـ $\mu_{it} = \mathrm{E}(Y_{it})$، فضلا عن التباين $\sigma_{it}^2 = var(Y_{it})$ والتي يمكن أيضاً التعبير عنها على التوالي $\mu_{it} = \alpha'(\theta_{it})$ و $\sigma_{it}^2 = \frac{\alpha''(\theta_{it})}{\phi} = v(\mu_{it})/\phi$ حيث ان v هو دالة التباين التي تشير إلى العلاقة بين المتوسط والتباين. (16)

كذلك يتم استخدام مصفوفة هيسيان Hessian matrix لحل الـ GEE ضمن فضاء المعلمة لحساب تقديرات الخطأ المعياري الحصينة (RSE)، في حين أن بنية التباين هي مصفوفة التباين المشترك الجبرية لنتائج الـ Y في العينة.

آخذاً في الاعتبار افتراض بان التوزيع المستقل للـ Y_i $(i = 1, 2,3 \ldots ,K)$ له متجه المتوسط $\mu_i =$ $(\mu_{i1}, \mu_{i2}, \mu_{i3}.., \mu_{ini})^T$، وكذلك مصفوفة التباين Σi التي تحتوي على العناصر القطرية $\sigma_{i1}^2, \sigma_{i2}^2, \sigma_{i2}^2, \ldots, \sigma_{in_i}^2$ جنبا إلى جنب مع العناصر خارج القطر $\rho_{itt'}\sigma_{it}\sigma_{it'}$ حيث ان $(t, t' = 1,2, 3, \ldots, n_i; t \neq t')$ ، فانه يمكن تعريف المصفوفة القطرية التي هي تباين Y_i من درجة $n_i \times n_i$ على النحو: $A_i = \text{diag}\{\alpha''(\theta_{it})\}$

لنفرض مصفوفة الارتباط R_i التي تحتوي على العناصر خارج القطر المشار لها بالرمز $\rho_{itt'}$؛ عندها يمكن التعبير عن مصفوفة التباين على النحو ادناه:

$$\Sigma_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}} / \phi \qquad \ldots (2.4)$$

فإذا افترضنا أن معامل الانحدار β (متجه $p \times 1$) هو المعلمة المقدرة، و φ هي المعلمة المزعجة nuisance parameter في حين ان المصفوفة المتماثلة n x n هي مصفوفة الارتباط التي يشار إليها بالرمز $R(\alpha)$ و α التي هي المتجه $s \times 1$ يمكن ان تميز $R(\alpha)$ تماماً في ان s هو عدد صحيح موجب مناسب، عندها يمكن اعتبار $R(\alpha)$ لتكون مصفوفة الارتباط العاملة.[9]

وبصفة عامة، فان بنية (IN) تحتوي على مصفوفة الوحدة Identity matrix، والبنية (EX) تحتوي على $\rho_{itt'} = \alpha$، وكذلك بنية AR-1 مع $\rho_{itt'} = \alpha^{|t-t'|}$ كلها تنطبق كهياكل ارتباط عاملة.[20] ووفقاً لطريقة GEE، المقدر $\hat{\beta}$ للمعلمة β هو النتيجة للمعادلة:

$$U(\beta, \alpha) \equiv \sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0 \qquad \ldots (2.5)$$

حيث يدل D_i على المصفوفة $n_i \times p$ التي يمكن التعبير عنها بالاتي:

$$D_i = \frac{\partial \mu_i}{\partial \beta}$$ ، و $V_i = \phi A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$ ، و $S_i = Y_i - \mu_i$ ، و $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})^T$.

الحل للمعادلة (2.5) يتوقف على استبدال الـ α مع $\hat{\alpha}(Y, \beta, \phi)$ الذي هو مقدر $K^{\frac{1}{2}}$ المتسق قبل استبدال الـ φ مع مقدر $K^{\frac{1}{2}}$ المتسق ألذي هو $\hat{\phi}(Y, \beta)$. ونتيجة لذلك، يمكن التعبير عن تقدير الـ φ بالصيغة ادناه:

$$\hat{\phi}^{-1} = \frac{1}{(\sum_{i=1}^{K} n_i - p)} \sum_{i=1}^{K} \sum_{t=1}^{n_i} \hat{r}_{it}^2 \qquad \ldots (2.6)$$

حيث بموجبه يمكن أيضاً التوسع بالتعبير عن بواقي بيرسون $\hat{r}_{it}$ بالصيغة الاتية:

$$\hat{r}_{it} = \frac{Y_{it} - \hat{\mu}_{it}}{\sqrt{v(\hat{\mu}_{it})}} \qquad \ldots (2.7)$$

في الحالات التي يتم فيها تحديد بنية الارتباط العاملة (WCS) لتكون بنية الـ EX مع φ المعرفة، عندها يمكن التعبير عن تقدير معلمة الارتباط α بالصيغة الاتية:

$$\hat{\alpha}_{EX} = \frac{\phi}{\left\{\frac{1}{2} \sum_{i=1}^{K} n_i (n_i - 1) - p\right\}} \sum_{i=1}^{K} \sum_{t<t'} \hat{r}_{it} \hat{r}_{it'} \qquad \ldots (2.8)$$

من ناحية أخرى، في الحالات التي يتم فيها تحديد WCS لتكون بنية الـ AR-1 مع φ المعرفة، فانه يمكن التعبير عن تقدير الـ α بالصيغة الاتية:

$$\hat{\alpha}_{\mathrm{AR-1}} = \frac{\phi}{\{\sum_{i=1}^{K} (n_i - 1) - p\}} \sum_{i=1}^{K} \sum_{t \le n_i^{-1}} \hat{r}_{it}\, \hat{r}_{i,t+1} \qquad \dots (2.9)$$

لذلك ووفقاً لطريقة GEE، مصفوفة تباين الـ$\hat{\beta}$ التي تتم الإشارة إليها بالرمز V_r هي التباين الحصين او المصحح تجريبياً (الشطيرة)، التي يمكن التعبير عنها بالصيغة الاتية:

$$V_r = \left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{K} D_i^T V_i^{-1}\, var(Y_i) V_i^{-1} D_i\right\} \left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1} \quad \dots (2.10)$$

المقدار يمثل الخبز المقدار يملئ الجبن واللحم المقدار يمثل الخبز

وان تقدير تباين الـ$\hat{\beta}$ الذي تتم الإشارة إليه بالرمز $\hat{V}_r$ يمكن أن يتحقق عن طريق استبدال $S_i S_i^T$ للـ $var(Y_i)$ وكذلك استبدال β، φ ، α للتقديرات الخاصة بكل منها في المعادلة رقم (2.10) اعلاه.

إذا أخذنا بعين الاعتبار ان $\mathrm{x}_{it} = (x_{it1}, x_{it2}, x_{it3}, \dots, x_{itp})^T$ متجه المتغيرات التوضيحية المقابلة للاستجابة t^{th} لمتغير التلوث i^{th} حيث $x_{it1} = 1$ لكل i وt ، وكذلك النظر في الافتراض بأن y_{it} لها توزيع العائلة الأسية، واعتماد دالة المتوسط $\mu_{it} = \Pr(y_{it} = 1)$ في إشارة الى مجموعة المتغيرات التوضيحية x_{it} وبالتالي يمكن التعبير عن استخدام دالة الربط المشار إليها بالرمز $h(\cdot)$ لتكون $\mu_{it} = h^{-1}(\mathrm{x}_{it}^{\mathrm{T}}\beta)$ ، حيث ان متجه المعلمة β يمكن التعبير عنه على النحو: $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)^{\mathrm{T}}$ وان هياكل الارتباط الواسعة الانتشار $\mathrm{R}_i(\alpha)$ تحدد الاعتمادية ضمن العناقيد، على افتراض المعلمات $\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_s)^{\mathrm{T}}$ تصف بنية مصفوفة الارتباط العاملة بشكل تام. [12] على سبيل المثال: يمكن الحصول على:

1. بُنية (IN) عن طريق افتراض: $\mathrm{R}_i(\alpha) = \mathrm{I}_{ni}$
2. بُنية (EX) عن طريق افتراض: $\mathrm{R}_i(\alpha) = (1-\rho)\mathrm{I}_{ni} + \rho \mathrm{J}_{ni}$

الذي بموجبه سيكون $\rho = \mathrm{corr}(y_{it}, y_{it'})$، حيث ان $t, t' = 1, 2, 3 \dots, n_i$ و $t \neq t'$ ، بينما رتبة مصفوفة الوحدة $n_i \times n_i$ هي I_{ni} مع J_{ni} كون المصفوفة $n_i \times n_i$ جميع عناصرها مساوية للواحد. [2] على النحو الذي اقترحه Liang وZeger (1986)، لتقدير معلمات الانحدار يتطلب معادلات التقدير رقم (2.1) حيث بموجبها تتم الاشارة إلى مصفوفة التباين المشترك العاملة للعنقود i^{th} بالرمز V_i التي يمكن أيضاً أن تعد دالة لمصفوفة الارتباط العاملة حيث يمكن أن تكون بالصيغة الاتية:

$$\mathrm{V}_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \qquad \dots (2.11)$$

مع $A_i = \mathrm{diag}(\mathrm{var}(y_{i1}), \mathrm{var}(y_{i2}), \mathrm{var}(y_{i3}), \dots, \mathrm{var}(y_{ini}))$ بينما $\mathrm{var}(y_{it}) = \alpha(\varphi)\mu_{it}(1-\mu_{it})$ هو دالة الـ μ_{it} التي حددت دالة المتوسط فضلا عن معلمة التشتت φ Dispersion parameter. [12] وكنتيجة لذلك، معادلة التقدير رقم (2.1) هي دالة لمعلمات الانحدار β، فضلا عن معلمات التشتت φ و α. [5]

ولتقدير معلمات الانحدار، فمن الممكن اختزال معادلات التقدير من خلال اعتبارها دالة لـ β التي يمكن أن تتحقق عن طريق استبدال معلمات التشتت α وφ مع $\hat{\alpha}(Y,\beta,\phi)$ لمعلمة التشتت α و $\hat{\phi}(Y,\beta)$ لمعلمات التشتت φ التي سوف تؤدي بالنتيجة في معادلات التقدير التي يمكن التعبير عنها بالصيغة التالية:

$$U(\beta,\hat{\alpha}(\beta,\hat{\phi}(\beta))) = \sum_{i=1}^{k}\left(\frac{\partial\mu_i}{\partial\beta}\right)' V_i^{-1}(\beta,\hat{\alpha}(\beta,\hat{\phi}))(y_i - \mu_i) = \sum_{i=1}^{k}\mathbf{C_i B_i A_i} \quad \dots (2.12)$$

مراعاةً لاستجابات ثنائية لعدة متغيرات، يمكن البرهنة على ان معادلة التقدير $\sum_{i=1}^{k}\mathbf{C_i B_i A_i}$ تكون دالة الدرجة score function وهي مشتقة من توزيع لوجستي متعدد المتغيرات وأن لها عزوم ثابتة من الرتبة الثالثة 3rd order وكذلك عزوم ثابتة من الرتبة الرابعة 4th order. [14] كما أكد Liang و Zeger (1986)، وعندما يتم توفير المقدرات لـ α و φ، فان مقدر معلمة الانحدار $\hat{\beta}$ يتسق مع عدة متغيرات تقاربية طبيعية محتوية على متوسط الـ β، فضلاً عن مصفوفة التباين المشترك التي يتم التعبير عنها بالصيغة الاتية :

$$V = \left(\sum_{i=1}^{k}\mathbf{C_i B_i C_i'}\right)^{-1}\left(\sum_{i=1}^{k}\mathbf{C_i B_i A_i A_i' B_i' C_i'}\right)\left(\sum_{i=1}^{k}\mathbf{C_i B_i' C_i'}\right)^{-1} \quad \dots (2.13)$$

وهذا يعني أن معلمة الانحدار $\hat{\beta}$ تُحَل من خلال مساواة المعادلة رقم (2.12) بالصفر:

$$U\left(\beta,\hat{\alpha}\left(\beta,\hat{\phi}(\beta)\right)\right) = 0^{(13)}$$

ومن ثم يمكن إجراء اختيار نموذج مع كل ما يتناسب مع طريقة GEE من المعايير AIC وQIC. [15]

10. أنواع هياكل الارتباط العاملة

بما ان المشاهدات n_i لكل موضوع من الموضوعات i حيث ان $i = 1, 2, 3 \dots, K$ تكون مترابطة فيما بينها بشكل عام، وترتبط أيضاً مع مصفوفة الارتباط العاملة المشار إليها بالرمز R(α) كما تم تعريفها من خلال نهج GEE الذي وضعه Liang و Zeger (1986) مما يعني أن المتجه $\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_s)^T$ ذي البعد s يمكن استخدامه لتحديد R(α).

وبالنظر الى افتراض أن $n_i = n$، فمن الممكن تحديد الهياكل الشائعة الأربعة التي سيتم استخدامها في الجانب التطبيقي، وهي:

(1) بنية الاستقلالية (IN)، حيث ان، $R(\alpha) = I_n$، يشير I_n الى مصفوفة الوحدة أي عندما لا يكون هناك أي ارتباط فيما بين العناقيد.
(2) بنية قابلة للصرف (EX)، حيث ان α المعلمة المجهولة.
(3) بنية الانحدار الذاتي من الرتبة AR(1)، حيث ان α المعلمة المجهولة.
(4) البنية الثابتة Stationary structure (ST)، حيث ان α لها $n - 1$ من المعلمات المجهولة. لذلك يمكن التعبير عن المصفوفات لهياكل كل منها على النحو الاتي:

$$\text{IN} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \text{AR(1)} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n-1} & \alpha^{n-2} & \alpha^{n-3} & \cdots & 1 \end{pmatrix}$$

$$EX = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix} \quad ST = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} \\ \alpha_1 & 1 & \alpha_1 & \cdots & \alpha_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{n-2} & \alpha_{n-3} & \cdots & \ddots & \alpha_1 \\ \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_1 & 1 \end{pmatrix}$$

وأن توقع Y_{it} الشرطي وفقاً لطريقة GEE هو $E(Y_{it}|X_{it}) = \mu_{it}(\beta)$ الذي يمكن أيضاً التعبير عنه بالشكل: $g^{-1}(X_{it}^T \beta)$ حيث ان g هي دالة الربط لنموذج الـ GLM بينما β هي معلمة الانحدار المجهولة للمتجه ذي البعد p.

فضلا عن ذلك، يمكن أيضاً افتراض بأن $Var(Y_{it}) = v(\mu_{it}(\beta))\phi$ ، التي يمكن أيضاً التعبير عنها بـ: $\phi\sigma_{it}^2$ ، بينما $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}\phi$ هو بنية التباين المشترك العاملة لـ Y_i مع مصفوفة A_i القطرية التي تتكون من σ_{it}^2، حيث ان $t = 1,2,3, \ldots, n_i$ للقطر في حين ان ϕ هي معلمة الزيادة في التشتت overdispersion.

11. معايير اختيار بنية الارتباط العاملة

للتخفيف من سوء مواصفات بنية الارتباط، يطرح بعض الباحثين معايير لاختيار بنية ارتباط عاملة جديدة. فعلى سبيل المثال فسر Rotnitzky و Jewell (1990) نهج لتقييم مدى كفاية مصفوفة الارتباط غير المحددة باستخدام معيار Rotnitzky-Jewell (RJC) مراعاةً للتوزيع التقاربي المبني على إحصاءة والد العاملة المعدلة Adjusted working Wald statistic التي هي في الواقع من التجميع الخطي للمتغيرات العشوائية المستقلة χ_1^2. [16] كما اعتمد Pan (2001) على إطار شبه الإمكان لتطوير الـ QIC الذي هو بمثابة تعديل لـ AIC بغية تعزيز كفاءة مقدرات GEE، فضلاً عن اختيار مصفوفة الارتباط العاملة. [15]

12. معيار شبه الإمكان في ظل نموذج الاستقلال الملائم (QIC)

اقترح Pan (2001) معيار شبه الإمكان المعروف بـ QIC، الذي يمكن أن يستخدم في اختيار نموذج المتوسط المناسب μ_{ij}، فضلاً عن بنية الارتباط العاملة. [15] ووفقاً لـ Hardin و Hilbe (2003)، يمكن التعبير عن دالة شبه الإمكان على النحو الاتي:

$$Q(\mu, \phi; y) = \int_y^{\mu} \frac{\phi(y - \mu^*)}{v(\mu^*)} d\mu^* \qquad \ldots(2.14)^{(7)}$$

حيث ϕ هي معلمة القياس و μ^* هو مقدر معلمة الانحدار μ.

ويقول Pan (2001) [15]، أنه عندما يفترض أن العناقيد، فضلاً عن النقاط الزمنية تكون مستقلة، ينبغي أن يحسب شبه الامكان لبيانات طولية على النحو التالي:

$$Q(\mu, \phi) = -2 \sum_{i=1}^{K} \sum_{t=1}^{n_i} Q(\mu, \phi; Y_{it}) \qquad \ldots(2.15)$$

وكنتيجة لذلك يمكن التعبير عن دالة شبه الإمكان للعنقود i في المشاهدة t التي يتم تقييمها باستخدام معلمات الانحدار β على النحو التالي: $Q(\beta, \phi; Y_{it}, x_{it}) = Q_{it}/\phi$، باعتبار ان Q_{it} تعود إلى توزيع Binomial أي ان $\mu_{it}(1-\mu_{it})$ مع دالة تباين $Q_{it} = y_{it} \ln\{\mu_{it}/(1-\mu_{it})\} + \ln(1-\mu_{it})$ و دالة الربط ln $\{\mu_{it}/(1-\mu_{it})\}$ وإذا كان الافتراض العملي هو أن كلاً من العناقيد، وكذلك المشاهدات تكون مستقلة، عندها يمكن الإشارة إلى QIC على النحو الاتي:

$$\mathrm{QIC}(\mathbf{R}) = -2\sum_{i=1}^{K}\sum_{t=1}^{n} Q(\beta,\phi;Y_{it},x_{it}) + 2\mathrm{tr}\{\Omega \mathrm{V_T}(\mathbf{R})\} \quad \ldots(2.16)$$

حيث يدل **tr** على أثر المصفوفة في حين ان: $\boldsymbol{\Omega} = \sum_{i=1}^{K} \mathrm{D}_i^{\mathrm{T}}\, \mathrm{A}_i^{-1}\, \mathrm{D}_i$

وهي عبارة عن المصفوفات ***p* x *p*** حيث **p** المقابلة لعدد معلمات الانحدار. و$\mathrm{V_T}(\mathbf{R})$ هو مصفوفة التباين المشترك القائمة على النموذج لمعلمات الانحدار المقدرة باستخدام مصفوفة التباين المشترك المستقلة.

13. **معيار غوشو** (DEW)

تماشياً مع اقتراح **Hin** و**Wang** (2009)، يشير أيضاً غوشو واخرون (2011) إلى اختيار بنية الارتباط مع الحد الأدنى لـ **DEW(R)** ، وفقاً لبنية الارتباط العاملة التي يعبر عنها غوشو واخرون (2011) على النحو الاتي:

$$\mathrm{DEW}(\mathrm{R}) = \mathrm{tr}\left[\left\{\left(\frac{1}{K}\sum_{i=1}^{K}\mathrm{S}_i\mathrm{S}_i^{\mathrm{T}}\right)\left(\frac{1}{K}\sum_{i=1}^{K}\mathrm{V}_i\right)^{-1} - I\right\}^2\right] \quad \ldots(2.18)$$

يتم الاشارة إلى مصفوفة الوحدة بالرمز ***I*** في المعادلة **DEW** في حين ان **DEW(R)** هو المعيار الذي يستخدم لقياس الفرق بصورة مباشرة بين مقدر مصفوفة التباين المشترك ومصفوفة التباين المشترك العاملة المحددة. [5]

14. **معيار معلومات الارتباط** Correlation Information Criterion (CIC)

اقترح **Hin** و **Wang** (2009) تعديل الـ **QIC** بغية تحسين أدائه للحصول على معيار معلومات الارتباط، الذي يمكن التعبير عنه على النحو الاتي:

$$\mathrm{CIC}(\mathrm{R}) = \mathrm{tr}\{\Omega \mathrm{V_r}(\mathrm{R})\} \quad \ldots(2.19)$$

في الحقيقة ان **(CIC)** يتوقف على الحد الثاني فقط من **QIC** وفقاً لمعادلة **QIC** رقم **(2.16)**، حيث يمثل الحد الأول مجموع دوال شبه الامكان لجميع المشاهدات، مع افتراض أن العناقيد المعروضة، فضلاً عن النقاط الزمنية تكون مستقلة. [10]

أن الهدف من بناء **CIC** هو لتحسين أداء **QIC** عن طريق تصغير **CIC(R)** في اختيار بنية الارتباط، ومن ثم يمكن التعبير عن بنية الارتباط العاملة لـ **(CIC)** على النحو التالي:

$$\mathrm{CIC}(\mathrm{R}) = \mathrm{tr}\left(\widehat{\Omega}_I \widehat{\mathrm{V}}_{\mathrm{r}}\right) \quad \ldots(2.20)^{(10)}$$

ومن الممكن الحصول على $\widehat{\Omega}_I$ عن طريق استبدال **β** و **φ** و **α** مع تقديراتها الخاصة بكل منها في معادلة **QIC** رقم **(2.17)** التي هي مماثلة لحساب **QIC**، مع $\widehat{\mathrm{V}}_{\mathrm{r}}$ كتقدير للتباين الحصين الذي يناظر معادلة مصفوفة التباين التي يتم الحصول عليها باستخدام طريقة **GEE**.

15. **معيار روتنيتزكي جيويل** The Rotnitzky-Jewell's Criterion (RJC)

وفقاً لـ **Rotnitzky** و**Jewell** (1990)، يمكن استخدام إحصاءات الاختبار لدعم فرضية أن متجه معاملات الانحدار مساوي لـ **β** محددة. [16] ويعرف **Rotnitzky** و**Jewell** (1990) كل من Ψ_0 وΨ_1 و**Ψ** مع الاشارة الى النظرية ذات الصلة بإحصاءات الاختبار على النحو الاتي:

$$\Psi_0 = \frac{1}{K}\sum_{i=1}^{K} D_i^T\ V_i^{-1} S_i S_i^T\ V_i^{-1} D_i \qquad \dots (2.21)$$

$$\Psi_1 = \frac{1}{K}\sum_{i=1}^{K} D_i^T\ V_i^{-1} D_i \qquad \dots (2.22)$$

$$\Psi = \Psi_0^{-1}\Psi_1 \qquad \dots (2.23)$$

حيث Ψ_0 , Ψ_1 هي هياكل الارتباط العاملة المفترضة، و $S_i = Y_i - \mu_i$ وفي هذه الحالة، يكون Ψ مكافئ لمصفوفة الوحدة إذا تم تحديد بنية الارتباط العاملة بشكل صحيح. وتماشياً مع تقييم Hin وآخرون (2007)، وبناءً على ذلك يمكن التعبير عن (RJC) لاختيار بنية الارتباط العاملة على النحو الاتي:

$$RJC(R) = \left[\left\{1 - \frac{tr(\Psi)}{p}\right\}^2 + \left\{1 - \frac{tr(\Psi^2)}{p}\right\}^2\right]^{\frac{1}{2}} \qquad \dots (2.24)$$

16. معيار الامكان القياسي الزائف

Gaussian Pseudo-Likelihood Criterion (GPC)

يستند معيار الامكان القياسي الزائف (GPC) الذي بُنيَ من قبل Carey وWang (2011) على دالة شبه الامكان الموسعة Extended QL التي تم تطويرها سابقاً من قبل Hall و Severini (1998).[6] ان دالة QL الموسعة توفر دالة التقدير لكلاً من β و α، كما ان لها ميزات مماثلة إلى دالة الإمكان القياسي لمتعدد المتغيرات. ومن ثم يمكن الإشارة إلى V_i على النحو الاتي:

$$W_i(\alpha) = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}} \phi \qquad \dots (2.25)$$

بينما $E(Y_i|X_i) = \mu_i$، حيث ان $W_i(\alpha)$ تمثل بنية التباين المشترك العاملة المقدرة من معادلات التقدير. ومن ثم يمكن التعبير عن دالة QL الموسعة على النحو ادناه:

$$LG = -\frac{1}{2}\sum_i \left\{(Y_i - \mu_i)^T W_i^{-1}(Y_i - \mu_i) + \log(|W_i|)\right\} \qquad \dots (2.26)$$

وبالتالي فان معيار الإمكان القياسي الزائف الذي اقترحته Carey وWang (2011) يكون GPC = $-2LG$ ، وإذا تم استخراج تقديرات β و α من GEE ، عندها يمكن مطابقة كل من $\mu_i(\hat{\beta})$ فضلا عن $R_i(\hat{\alpha})$. ولذلك يمكن أن تكتمل عملية اختيار نموذج عن طريق تحديد النموذج المرشح الذي يمتلك الحد الادنى من GPC.[3] مع الاخذ بالحسبان حقيقة أن GEE تفتقر إلى دالة الإمكان.

وبالإضافة إلى ذلك، قام Pan (2001) بتعديل هذا التباعد من خلال تكييف حد عقوبة QIC، مما يعني أنه يمكن تحسين المعايير المعدلة عن طريق استبدال عامل الامكان في AIC مع عامل شبه الإمكان الموسع. كنتيجة لذلك، يمكن التعبير عن المعيار المعدل القائم على أساس AIC بالصيغة الاتية:

$$AGPC = -2LG + 2\dim(\theta) = GPC + 2\dim(\theta) \qquad \dots (2.27)$$

اما المعيار القائم على أساس بيز BIC الذي طوره Schwarz (1978) وهو مشابه لـ AIC مع الفارق الوحيد الذي يحدث في حد العقوبة. [18] بالحفاظ على حد عقوبة مماثل لهذا الذي في BIC، ومن ثم يمكن تعديل BIC لتشكيل المعيار BGPC، الذي يتم التعبير عنه بالصيغة الاتية:

$$\mathrm{BGPC} = -2\mathrm{LG} + \log(K)\dim(\theta) = \mathrm{GPC} + \log(K)\dim(\theta) \quad \dots(2.28)$$

حيث θ هو متجه المعلمات الحرة المقدرة التي يمكن أيضاً التعبير عنها بـ: $(\beta^T, \alpha^T)^T$.

17. معايير الامكان التجريبي لأكايكي و بيز

Empirical Likelihood Criteria **(EAIC) (EBIC):**

يوضح النهج المتبع من قبل Chen وLazar (2012) الذي يستبدل الامكان التجريبي (EL) في AIC وBIC مع الإمكان المعلمي لتشكيل معيارين إضافيين لتحديد مصفوفة الارتباط العاملة ان هذين المعيارين تم بناؤهما ليكونا أكثر فعالية إذا ما قورنا بـ QIC وCIC. [4] ونتيجة لذلك، فمن الضروري مواصلة التحقيق وتقييم أداء المعايير المستندة إلى EL في ضوء المعايير الأخرى التي تمت مناقشتها. عند تقييم EAIC وEBIC، فان Chen وLazar (2012) يركز بالدرجة الأساس على اشتقاق نسبة الامكان التجريبي Empirical likelihood ratio (ELR) $R_F(\alpha)$ لنموذج كامل تحت افتراض أن مصفوفة الارتباط العاملة المستقرة (ST) لها $(p+n-1)$ من المعلمات الحرة مدرجة في $\theta^T = (\beta^T, \alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1})$. [4]

في ظل النموذج المفترض، فان الحصول على ELR للنموذج يتطلب التعريف الأولي لدالة التقدير $\left(g^F(\cdot)\right)$ على النحو الاتي:

$$g^F\left((Y_i, X_i),\ \beta,\ \alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1};\ R_F(\alpha)\right) = \begin{pmatrix} (\partial\mu_i/\partial\beta^T)^T \mathrm{A}_i^{-\frac{1}{2}} R_F^{-1}(\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1}) \mathrm{A}_i^{-\frac{1}{2}} (Y_i - \mu_i) \\ \sum_{t=1}^{n-1} e_{it}(\beta) e_{i,t+1}(\beta) - \alpha_1 \hat{\phi}(\beta)(n - 1 - p/K) \\ \vdots \\ \sum_{t=1}^{1} e_{it}(\beta) e_{i,t+n-1}(\beta) - \alpha_{n-1} \hat{\phi}(\beta)(1 - p/K) \end{pmatrix}_{(p+n-1)\mathrm{X}\,1} \quad \dots(2.29)$$

مع البواقي بيرسون معبراً عنها بالصيغة ادناه:

$$e_{it}(\beta) = \frac{(Y_{it} - \mu_{it}(\beta))}{\sqrt{v(\mu_{it}(\beta))}}$$

$$\hat{\phi}(\beta) = \sum_{i=1}^{K} \sum_{t=1}^{n} e_{it}^2 / (K\,n - p) \quad \dots(2.30)$$

ومن ثم هذا يجعل من الممكن التعبير عن دالة ELR عن طريق استبدال معادلة التقدير الجديدة مع ما ينتج من حد GEE الذي تم بناءه في الصيغة ادناه:

$$\mathcal{R}^F(\beta, \alpha^T) = \sup\left\{\prod_{i=1}^{K} K\omega_i : \omega_i \geq 0, \sum_{i=1}^{K} \omega_i = 1, \sum_{i=1}^{K} \omega_i g^F\left((Y_i, X_i), \beta, \alpha^T;\ R_F(\alpha)\right) = 0\right\} \dots(2.31)$$

ولذلك يكون لمعادلات تقدير الهياكل المرشحة الخاصة بكل منها مصفوفة ارتباط عاملة مماثلة مع بنية ST التي هي مع دالة التقدير $\left(g^F(\cdot)\right)$ سوف تسهل تحقيق ELRs مع كل من القيم المختلفة وكذلك القيم المتوافقة لمختلف مصفوفات الارتباط العاملة. وهذا يتوقف على حقيقة ان مقدرات الامكان التجريبي الأعظم تكون مماثلة لمقدرات GEE باستثناء التعديل الذي سيكون مع النتيجة في قيم ELR التي تساوي 1. بعد ذلك يتم الحصول على قيم ELR عن طريق إدراج تقديرات GEE في دالة ELR المعبر عنها بالرمز $\mathcal{R}^F(\beta,\alpha^T)$. كما قام Chen وLazar (2012) بمواصلة تطوير معايير اختيار مصفوفة الارتباط العاملة مع أعلى قيم ELR الناتجة ، إلا وهي: EAIC وEBIC ، حيث ان:

$$\text{EAIC} = -2\log\mathcal{R}^F\left(\hat{\theta}_G\right) + 2\dim(\theta) \qquad \ldots(2.32)$$

$$\text{EBIC} = -2\log\mathcal{R}^F\left(\hat{\theta}_G\right) + \log(K)\dim(\theta) \qquad \ldots(2.33)$$

حيث dim (θ) هو عنصر من عناصر المعلمات الحرة $(\beta^T,\alpha^T)^T$ التي يتم تقديرها، وان الحد الأدنى لقيم EAICوEBIC تدل على النموذج المحتمل. (4)

الفصل الثاني / الجانب التطبيقي

1.وصف البيانات

تم الحصول على بيانات التلوث من شركة مصافي الوسط في العراق للقياسات اليومية للمدة من أيلول/سبتمبر 2011 إلى كانون الأول/ديسمبر 2013وخلال كل مرحلة قياس على طول المدة المعروضة قيد الدراسة، سجلت مختلف المقاييس الرئيسية، هي: التاريخ والوقت، وقيست كل من الجسيمات ($PM_{2.5}$)، وكبريتيد الهيدروجين (H_2S)، وأكاسيد النيتروجين (NO_x)، والأمونيا (NH_3)، وأول أوكسيد الكاربون (CO)، وثاني أوكسيد الكاربون (CO_2) والأوزون (O_3) بالمايكرو غرام لكل متر مكعب ($\mu g/m^3$) كمتغيرات استجابة عددها (7)، وكذلك متوسط درجة الحرارة لكل ساعة بالـ (°C)، ومتوسط نقطة الندى ومتوسط الرطوبة لكل ساعة بالدرجة المئوية (%)، ومتوسط سرعة الرياح بالـ (كم/ساعة)، ومتوسط كمية النفط الخام المستخدم في عمليات التصفية بالـ (م 3/ساعة) كمتغيرات تفسيرية (توضيحية) عددها (5).

وان تحليل البيانات الاستكشافية Exploratory data analysis (EDA) يتسق مع شروط GEE نظراً لأنه يستخدم لتقييم البيانات الطولية بغية تحديد أنماط الاختلاف المنهجي عبر المجموعات وكذلك جوانب الاختلاف العشوائي الذي تميزه المشاهدات الفردية. ويمكن بعد ذلك استخدام GEEs لفهم قوة الارتباط ونمط الارتباطات عبر الزمن من خلال تمييز الارتباط بغية رسم خطوط مكونات الاختلاف ولتحديد التباين أو نموذج الارتباط.

2.تطبيق معايير اختيار النموذج

تعد نماذج منحنى النمو معتمدة على الوقت في التلوث لمقارنة النموذج، وأنواع بنية التباين المشترك الأربعة المحددة لنماذج منحنى النمو التجميعية. اذ يستخدم النهج GEE في مطابقة النماذج التجميعية، مع قيم معايير اختيار النماذج المطابقة المبينة في الجدول رقم (1) وقد تم تحديد مجموعة نماذج مرشحة[1] باستخدام تحليل البيانات الاستكشافية من خلال إنشاء المتغيرات التفسيرية التي تمثل أفضل تنبؤ لمعدلات التلوث، وتم استخدام برنامج MATLAB في الحصول على نتائج المعايير لتقييم مجموعة النماذج المرشحة المدرجة فيما يأتي تستند الى التوزيع اللوجستي.

[1] النموذج المرشح (النموذج التقريبي): هو النموذج المستخدم لتمثيل نموذج التوليد، مع النموذج المفضل كونه النموذج الذي يقلل من فقدان المعلومات من خلال وجود أصغر اختلاف لـ (Kullback -Leibler) الذي هو السبب في اختيار النموذج مع أدنى قيمة عند استخدام معايير المعلومات مثل AIC. علماً ان نموذج التوليد هو النموذج الذي يمكن استخدامه لتمثيل الاختلافات العشوائية في المشاهدات من حيث المتغيرات التفسيرية (النموذج الحقيقي).

Candidate Models	QIC	CIC	GPC	RJC	AGPC	BGPC	EAIC	EBIC	DEW
Particulate matter (PM2.5) ug/m3 Model_1	0.113	0.038	0.232	0.02	0.088	0.033	0.052	0.036	0.072
Hydrogen Sulfide (H2S) ug/m3 Model_2	0.008	0.044	0.038	0.063	0.104	0.022	0.014	0.075	0.002
Nitrogen Oxides (NOx) ug/m3 Model_3	0.02	0.043	0.149	0.121	0.087	0.099	0.049	0.061	0.044
Ammonia (NH3) ug/m3 Model_4	0.027	0.043	0.043	0.002	0.014	0.046	0.029	0.02	0.032
Carbon monoxide (CO) ug/m3 Model_5	0.076	0.132	0.109	0.031	0.006	0.018	0.024	0.096	0.021
Carbon dioxide (CO2) ug/m3 Model_6	0.009	0.035	0.039	0.019	0.043	0.054	0.025	0.052	0.126
Ozone (O3) ug/m3 Model_7	0.042	0.007	0.028	0.134	0.036	0.157	0.007	0.029	0.036
Total loss information	0.295	0.342	0.638	0.39	0.378	0.429	0.2	0.369	0.333

الجدول رقم (1): ملخص مطابقة مختلف معايير اختيار النموذج

ووفقا للجدول رقم (1)، تُسَجل EAIC أدنى مستوى لفقدان المعلومات.

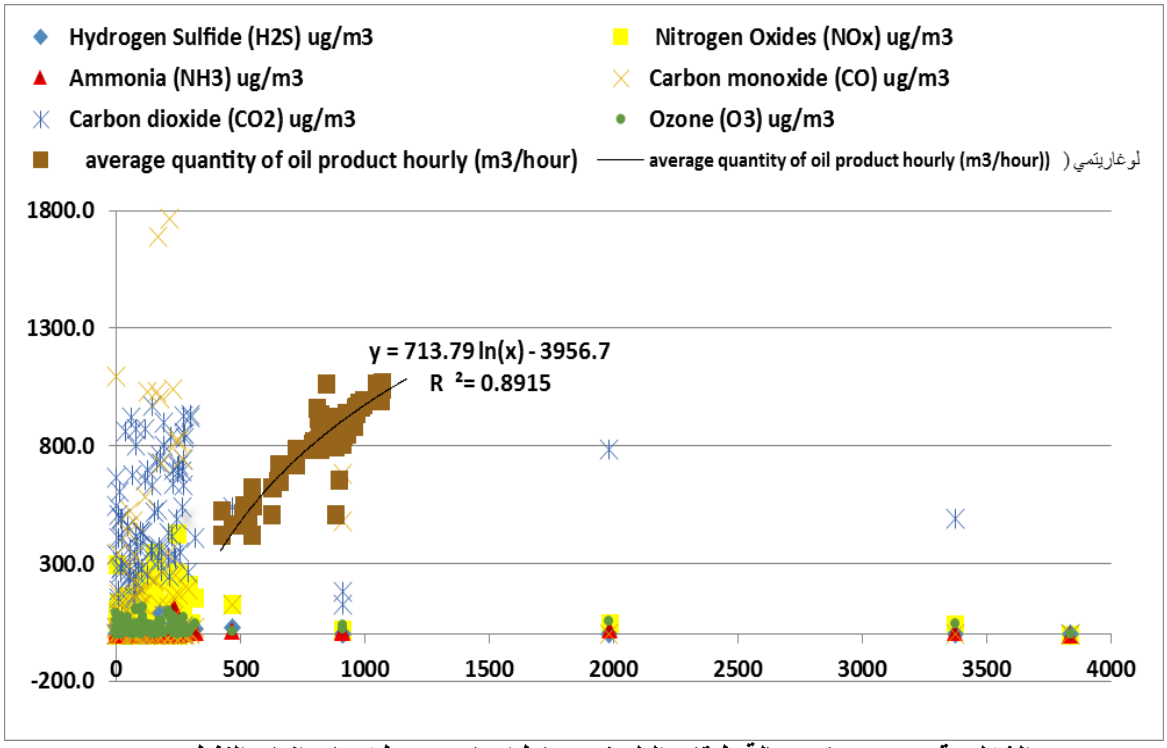
3.تقييم مستويات التلوث

لغرض التعرف على العوامل المثلى لتلك النماذج تم اجراء تحليل نماذج السلاسل الزمنية المبينة في الجدول رقم (2):

الجدول رقم (2): يبين إحصاءات النماذج

Model	Number of Predictors	Model Fit statistics			Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	R-squared	MaxAPE	Statistics	DF	Sig.	
Particulate matter (PM2.5) ug/m3-Model_1	2	.606	.606	8900.851	18.282	17	.371	0
Hydrogen Sulfide (H2S) ug/m3-Model_2	1	.268	.268	4537.344	11.106	17	.851	0
Nitrogen Oxides (NOx) ug/m3-Model_3	3	.257	.257	544.496	22.042	17	.183	0
Ammonia (NH3) ug/m3-Model_4	2	.188	.188	840.238	10.170	17	.896	0
Carbon monoxide (CO) ug/m3-Model_5	0	.676	.270	407391.492	22.672	16	.123	0
Carbon dioxide (CO2) ug/m3-Model_6	3	.327	.327	189862.060	16.264	14	.298	0
Ozone (O3) ug/m3-Model_7	2	.436	.436	3843.192	25.194	16	.066	0

يفترض النموذجين 3 و6 فيما يتعلق بالجدول رقم (2) أن يختلف مستوى التلوث تربيعياً، اما النماذج 1 و2 و4 و7 فتمثل GLMs الخطية المفترضة للبيانات، من ناحية أخرى، يفترض النموذج 5 الاستقرار في مستوى التلوث على مر الزمن بالاعتماد على معيار MaxAPE وكذلك عدد المتنبئات. ولغرض تقييم مستويات التلوث بالنسبة لكميات انتاج النفط مع مرور الوقت تم اجراء الرسم في برنامج Excel في الشكل رقم (1).



الشكل رقم (1): يوضح حالة طبقات التلوث مع خط اتجاه متوسط كميات انتاج النفط.

4.تحديد هياكل التباين المشترك

تم تحديد هياكل التباين المشترك لقياسات التلوث المتكررة كما هو مبين في الجدول رقم (3):

الجدول رقم (3): يبين إحصاءات مطابقة النموذج لهياكل التباين المشترك

Fit Statistic	Mean	SE	Minimum	Maximum	AR-1	IN	EX	ST
Stationary R-squared	0.394	0.186	0.188	0.676	0.188	0.257	0.327	0.606
R-squared	0.336	0.142	0.188	0.606	0.188	0.257	0.27	0.436
RMSE	144.649	142.028	10.507	340.218	10.507	10.624	100.247	283.948
MAPE	1206.496	2171.741	55.076	5943.61	55.076	83.419	137.181	1725.717
MaxAPE	87988.53	157022.9	544.496	407391.5	544.496	840.238	4537.344	189862.1
MAE	95.988	92.547	5.826	199.296	5.826	6.605	65.813	192.403
MaxAE	793.401	884.533	58.43	2333.314	58.43	65.334	653.447	1647.71
EAIC	8.528	3.177	4.775	11.927	4.775	4.824	9.363	11.345

وقد تم تحديد بنية التباين المشترك AR-1 كأفضل نموذج تجميعي نظراً لأنه يمتلك أقل قيمة لـ EAIC، مع حساب النموذج لحوالي 19% في تباين متغير الاستجابة (FIT = 0.188). وهذا يعني أن هناك ارتباط خطي إيجابي لمعدل التلوث مع مرور الوقت.

وكذلك يوحي تحليل الـ GEE أن التباينات المشتركة ضمن القياسات المتكررة للتلوث تُكون مهَيْكَلة على الأرجح في نمط الانحدار الذاتي من الرتبة الأولى، أو النمط الموسمي البسيط كما هو مبين من خلال متغير أول أوكسيد الكاربون (CO) في الجدول رقم (2). وهذا يعني أن معدل التلوث عند الزمن t يتحدد مباشرة عن طريق معدل التلوث عند الزمن $t-1$.

	Initial	Robust standard errors	Logit	Robust standard errors
Intercept	481.348	51.973	156.998	16.662
Carbon dioxide (CO_2)	-0.009	0.067	0.224	0.067
Nitrogen Oxides (NO_x)	-0.02	0.068	0.284	0.066

ومن ثم يمكن تلخيص تقديرات GEE وفقاً لمعاملات أفضل نموذج تجميعي مطابق، وما يقابلها من الأخطاء المعيارية الحصينة (RSE) على النحو الاتي:

يُظهر العمود Initial تأثيرات المتغيرات التفسيرية المناسبة غير المتحسسة للوقت المختارة على التلوث في حين يُظهر العمود Logit نتائج المتغيرات التفسيرية على معدل النمو اللوجستي للتلوث مع مرور الوقت.

قيمة الثابت β_0 المقدر في العمود Initial يساوي 481.35 (SE = 51.97) في حين أن متوسط معدل التلوث هو 157 (SE = 16.66)، مما يعني أن هناك قدراً كبيراً من التلوث ونسبة تلوث مع مرور الوقت في القياس الأولي. وكان التأثير الأولي لثاني أوكسيد الكاربون (CO_2) على التلوث 0.009- (SE =0.067) مما يشير إلى أن ثاني أوكسيد الكاربون لديه مستوى تلوث ادنى بالمقارنة مع غيره من المركبات.

وتأثير ثاني أوكسيد الكاربون على معدل التلوث هو0.224 (SE =0.067) الذي يشير إلى أن ثاني أوكسيد الكاربون يُزَيد التلوث بمعدل أعلى بالمقارنة مع المركبات الأخرى على الرغم من أن التأثير ليس كبيراً.

من ناحية أخرى، فان التأثيرات الأولية لأكاسيد النيتروجين (NO_x) هي 0.02- (SE =0.068) التي تشير إلى أن أكاسيد النيتروجين لها تأثيرات أقل على المستويات الأولية للتلوث على الرغم من أنها تسهم في معدل التلوث بنسبة 0.284 (SE. =0.066). ولغرض التعرف على طبيعة مجموعات البيانات ومن اجل ان يؤخذ في الاعتبار عدم التجانس على مستوى العناقيد، عن طريق تصنيفها بطريقة فعالة يجب اجراء تحليل عنقودي للبيانات وكما يأتي:

5. التحليل العنقودي المتعدد Multiple Cluster Analysis

بإجراء المزيد من التحليل للبيانات تم تطبيق مناهج العنقدة المتدرجة clusterwise، وتحديداً النهج العنقودي ذو خطوتين Two-step cluster analysis في SPSS، لغرض مطابقة النموذج اللوجستي الشرطي مع بنية التباين المشترك (AR-1) بالاعتماد على نتائج المرحلة الأخيرة من تحليل البيانات الاستكشافية.

ويبين الجدولين رقم (4) و(5) نتائج التحليل العنقودي ذو خطوتين لتقييم البيانات الذي بموجبه تم اعتبار المتغيرات التفسيرية مستمرة بينما كانت متغيرات الاستجابة المشفرة فئوية. وكشف أن طبيعة البيانات تتكون من 3 عناقيد تختلف فيما يتعلق بالمتوسط. وكما موضح في الشكل رقم (8).

الجدول رقم (4): يبين التوزيع العنقودي للبيانات

		N	% of Combined	% of Total
Cluster	1	29	26.4%	12.4%
	2	31	28.2%	13.3%
	3	50	45.5%	21.5%
	Combined	110	100.0%	47.2%
Excluded Cases		122		52.8%
Total		232		100.0%

الجدول رقم (5): يبين نبذة مختصرة عن المتغيرات التفسيرية العنقودية

		Average hourly temperature (°C)		Average hourly Dew Point (°C)		Average hourly humidity (%)		Average hourly wind speed (km/h)		average quantity of oil product hourly (m3/hour)	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	22.01	9.12	7.65	3.94	44.84	20.00	10.25	6.91	865.87	137.64
	2	26.76	10.11	6.60	3.19	33.65	19.25	8.97	3.25	794.57	120.42
	3	28.54	9.09	6.97	4.48	29.39	15.57	14.46	7.98	958.58	74.82
	Combined	26.31	9.69	7.04	3.99	34.66	18.82	11.80	7.07	887.92	127.56

تبين ان للبيانات 3 عناقيد قياسية، بينما أشارت التحليلات متعددة العنقودية أن GEE حققت نقاط تقارب بسرعة مراعاةً لأعداد العناقيد المختلفة. فعلى سبيل المثال، عندما $C = 2$ (حيث C هو عدد العناقيد) يتقارب النموذج في (10) تكرارات في حين كان يتطلب (100) تكرار عندما $C = 231$ وفقاً لبدء التشغيل الاولي لمعلمات العضوية العنقودية. وأن هناك زيادة تدريجية في احتواء $C = 2$، مما يعني أنه ليس هناك أي تحسن ملموس يمكن تحقيقه في المطابقة عن طريق زيادة عدد العناقيد لأكثر من 3.

ولذلك تم اعتماد $C = 3$ لتسهيل التحليلات الاضافية، نظراً لأن النموذج ثلاثي العنقود يمثل 47.2% من الاختلافات في متغيرات الاستجابة (FIT = 0.472) في حين يؤدي استبعاد متغير الجسيمات $PM_{2.5}$ بسبب فقدان 121 مشاهدة الى زيادة إمكانية تفسير تقلبية النموذج إلى 99.6% (FIT = 0.996). ولذلك تم استبعاد المتغير $PM_{2.5}$ لزيادة موثوقية النموذج، من إجراء التحليل الاضافي لمجموعة البيانات.

6. تقديرات GEE لنموذج ثلاثي العنقود

يبين الجدول رقم (6) تقديرات معاملات GEE التي تم الحصول عليها من تحليل ثلاثي العنقود، فضلاً عن الأخطاء المعيارية الحصينة المقابلة لها.

الجدول رقم (6): يوضح تقديرات معاملات GEE والأخطاء المعيارية الحصينة لنموذج تلوث ثلاثي العنقود

Clusters	Explanatory variables	Coefficient	Standard error
Cluster 1	Intercept	15.598	1.584
(n_1 = 97)	Temperature (°C)	10.005	1.016
	Dew Point (°C)	4.243	0.431
	Humidity (%)	18.100	1.838
	Wind speed (km/h)	6.523	0.662
	Oil product hourly (m3/hour)	131.134	13.315
Cluster 2	Intercept	13.649	1.367
(n_2 = 106)	Temperature (°C)	9.416	0.915
	Dew Point (°C)	4.173	0.405
	Humidity (%)	17.737	1.723
	Wind speed (km/h)	6.928	0.673
	Oil product hourly (m3/hour)	110.944	10.776
Cluster 3	Intercept	17.736	3.293
(n_3 = 29)	Temperature (°C)	9.025	1.676
	Dew Point (°C)	2.658	0.494
	Humidity (%)	13.502	2.507
	Wind speed (km/h)	5.560	1.033
	Oil product hourly (m3/hour)	163.334	30.330

وتم تأكـــيد الخصائص المميزة للعناقيد الثلاثة من خـــلال معلمات β_0 المـقدرة التي كانت 15.60 (SE 1.58 =، و13.65 (SE = 1.37)، و 17.74 (SE = 3.29) للعناقيد 1 و 2 و 3 على التوالي. فمن الواضح أن العناقيد الثلاثة تختلف بشكل جوهري، وتمثل أنماط مختلفة من التلوث.

الظروف السائدة في العنقود 1 من المحتمل أن تظهر مستويات تلوث معتدلة كما تظهر تغييرات معتدلة في مستوى التلوث عبر السلسلة الزمنية. ومن المرجح أن تُظهر الظروف السائدة في العنقود 2 مستويات تلوث منخفضة نسبياً وتُظهر أيضاً تغييرات منخفضة في مستوى التلوث عبر السلسلة الزمنية. من ناحية أخرى، يشارك العنقود 3 بدرجة عالية من التلوث وزيادة كبيرة في معدل التلوث عبر السلسلة الزمنية.

وان كمية النفط المنتج في كل ساعة كان له تأثير هام وإيجابي على مستوى تلوث جميع العناقيد، مشيراً إلى أن هذه الزيادة في كميات إنتاج النفط أدت إلى ارتفاع مستوى التلوث. وأيضاً أظهرت درجة الحرارة والرطوبة تأثير كبير وإيجابي على مستوى التلوث مما يوحي بان الزيادة في نسبة الرطوبة ودرجة الحرارة يؤدي الى مستويات تلوث أعلى. وتشير جميع العناقيد الثلاثة إلى أن كلاً من نوعيات إنتاج النفط، والرطوبة ودرجات الحرارة لها أثراً معنوياً في معدل التلوث على مر الزمن.

وبمطابقة نماذج متعددة العناقيد لبيانات التلوث باستخدام طريقة GEE من خلال بنية التباين المشترك (AR-1) تم انشاء C = 3 كعدد العناقيد الأمثل وفقاً للأساليب التجريبية الصحيحة للبيانات العنقودية. وكشف النموذج ثلاثي العنقود المقدر عن طريق GEE مسارات جوهرية متميزة في العوامل المسببة للتلوث داخل العناقيد الثلاثة. وهذا يعني أن المتغيرات التفسيرية متفاوتة الوقت لها تأثيرات مختلفة على معدل التلوث عبر العناقيد الثلاثة على الرغم من أن جميع العناقيد تشير إلى تأثير إيجابي على التلوث.

7. **الاستنتاجات**: Conclusions

1.تمت مقارنة أداء المعايير المحددة ودراستها مع تطبيقها على البيانات التي تشير إلى أن QIC هو معيار الاختيار الأنسب لأطر GEE التي تتطلب النموذج المرشح لضمان مطابقته بشكل صحيح، حيث أظهرت نتائج البحث أن QIC له فقدان المعلومات الأدنى في أطر GEE الذي كان الهدف من المجموعة المرشحة لتطوير نموذج تنبؤي.

2.سَجل ايضاً معيار اكايكي التجريبي EAIC أدنى مستوى لفقدان المعلومات مع حساب النموذج الاجمالي لحوالي 19% في تباين متغير الاستجابة.

3.ببساطة يتحكم الانحدار اللوجستي الشرطي بحالة عدم التجانس بتوفير إطار استدلالي حصين بشكل فعال. عن طريق نمذجة الاستجابة لكل من H_2S وNO_x و NH_3 التي هي المتغيرات ذات الأهمية لواحدة من أهم السمات البيئية وإنتاج النفط، فمن الممكن التوصل الى استنتاجات بشأن استجابة المتغيرات لدرجة الحرارة وإنتاج النفط.

4.من الواضح باستخدام تقديرات GEE لأفضل نموذج تجميعي مطابق، أن ثاني أوكسيد الكاربون (CO_2) وأكاسيد النيتروجين (NO_x) هي متغيرات الاستجابة الرئيسة التي يمكن التنبؤ بها باستخدام المتغيرات التفسيرية. في حين يشير التحليل العنقودي المتعدد إلى كميات إنتاج النفط، والرطوبة ودرجة الحرارة.

8. **التوصيات**: Recommendations

يمكن تلخيص التوصيات الرئيسة لهذه البحث كما يأتي:

1.يظهر QIC له فقدان المعلومات الأقل ولذلك يستخدم في تطوير النماذج. وبالمثل، فإنه من الضروري أن تتبنى الدراسات الأخرى إجراء مماثل لضمان تطوير نماذج تنبئية موثوق بها إحصائياً باستخدام طريقة GEE.

2.استخدام معيار اكايكي التجريبي EAIC في تحديد نمط مصفوفة الارتباط العاملة التي تتسم بها القياسات المكررة للبيانات الطولية بوصفه يسَجل أدنى مستوى لفقدان المعلومات الإجمالي للنموذج في تباين متغير الاستجابة.

المصادر

1. Armitage P, Berry G, & Matthews J. (2001). Statistical Methods in Medical Research (4th ed). Oxford: Blackwell Science.

2. Cantoni, E., Flemming, J. & Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, 61, 507-514.

3. Carey, V.J. and Wang, Y.-G., (2011) Working covariance model selection for generalized estimating equations, Statistics in Medicine, 30(26), 3117–3124.

4. Chen, J. and Lazar, N.A., (2012) Selection of working correlation structure in generalized estimating equations via empirical likelihood, Journal of Computational and Graphical Statistics, 21(1), 18–41.

5. Gosho, M., Hamada, C. & Yoshimura, I. (2011). Criterion for the Selection of a Working Correlation Structure in the Generalized Estimating Equation Approach for Longitudinal Balanced Data. *Communications in Statistics – Theory and Methods*, 40(21), 3839–3856.

6. Hall, D.B. and Severini, T.A., (1998) Extended generalized estimating equations for clustered data, Journal of the American Statistical Association, 93(444), 1365–1375.

7. Hardin, J. & Hilbe, J. (2003). Generalized Estimating Equations. London: Chapman and Hall/CRC.

8. He, X., Simpson, D. & Portnoy, S. (1990). Breakdown Robustness of Tests, Journal of the American Statistical Association, 85(40), 446-452

9. Hin L. & Wang, Y. (2009). Working-Correlation-Structure Identification in Generalized Estimating Equations. Statistics in Medicine, 28(4), 642–658.

10.Hin, L. Carey, V. & Wang, Y. (2007), Criterion for working-correlation-structure selection in GEE: Assessment via simulation, The American Statistician 61, 360–364.

11.Koepsell, T.D., and Weiss, N.S. (2003). Epidemiological Methods: Studying the Occurrence of Illness, Oxford University Press, New York, NY.

12.Liang, K-Y & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73 (1): 13–22..

13.Linhart, L., & Zucchini, W. (1986). Model Selection. New York: John Wiley and Sons.

14.Molenberghs, G., & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. Journal of the American Statistical Association, 89, 633-44.

15.Pan W (2001). "Akaike's Information Criterion in Generalized Estimating Equations." Biometrics, 57(1), 120–125.

16.Rotnitzky, A. and Jewell, N. P. (1990), Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data, Biometrika 77(3), 485–497.

17.Rousseeuw, P. & Croux, C. (1993). Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association 88, 1273.

18.Schwarz, G. (1978), Estimating the dimension of a model, The Annals of Statistics, 6, 461–464.

19.Stigler, S. (2010). The Changing History of Robustness, The American Statistician, 64(4): 277-281.

20.Sutradhar, B. C. and Das, K. (2000). On the accuracy of efficiency of estimating equation approach. Biometrics 56(2), 622–625.

21.Wilcox, R. (2001). Introduction to Robust Estimation & Hypothesis Testing, Academic Press.

22.Fleiss, J. (1993). The statistical basis of meta-analysis. Statistical Methods in Medical Research, 2:121-145.

23.Diggle, P., Heagerty, P., Liang, K-Y & Zeger, S. (2002). Analysis of Longitudinal Data. Oxford Statistical Science Series.

Application the generalized estimating equation Method (GEE) to estimate of conditional logistic regression model for repeated measurements

ABSTRACT

Conditional logistic regression is often used to study the relationship between event outcomes and specific prognostic factors in order to application of logistic regression and utilizing its predictive capabilities into environmental studies. This research seeks to demonstrate a novel approach of implementing conditional logistic regression in environmental research through inference methods predicated on longitudinal data. Thus, statistical analysis of longitudinal data requires methods that can properly take into account the interdependence within-subjects for the response measurements. If this correlation ignored then inferences such as statistical tests and confidence intervals can be invalid largely.

For estimating the conditional regression model in the analysis of environment pollution as a function of oil production and environmental factors using the generalized estimating equation (GEE) in the formulation of inference methods that facilitate the conditional logistic regression model taking advantage of the actual correlations between responses in the data, as well as the specific correlation structure through robust sandwich estimators (RSE) as well as application many of various model selection criteria. Because the efficiency of estimates is contingent on the working correlation matrix specification, the appropriate selection of a working correlation matrix can significantly advance the GEE statistical inference efficiency. After comparing the performance of specific criteria indicating that QIC is the selection criterion that is most suited for GEE method. The application results showed that QIC had the lowest information loss in GEE method in which the objective to develop a predictive model of the candidate set, Through this research, condition logistic regression has also been demonstrated to be an effective tool that can be used in other studies to explore the relationships between response and explanatory variables.

Key Words: **Conditional logistic regression, repeated measurements, generalized estimation equations method (GEE), Robust Sandwich Estimators (RSE), Model Selection Criteria, working correlation structure, Cluster analysis.**