مقارنة نماذج الانحدار اللوجستي الشرطية مع التأثيرات الثابتة والمختلطة في حالة البيانات الطولية

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

تاريخ التقديم: 2016/6/22
تاريخ القبول: 2016/10/31

المستخلص:

من الواضح أن الانحدار اللوجستي الشرطي مختلط التأثيرات أكثر فعالية في دراسة الاختلافات النوعية في بيانات التلوث الطولية، فضلاً عن آثارها المترتبة على مجموعات فرعية غير متجانسة. حيث يثبت هذا البحث أن الانحدار اللوجستي الشرطي هو طريقة تقييم حصينة للدراسات البيئية، من خلال تحليل تلوث البيئة كدالة لإنتاج النفط والعوامل البيئية. ونتيجة لذلك، فقد ثبت من الناحية النظرية أن الهدف الأساسي لاختيار النموذج في هذا البحث هو تحديد نموذج مرشح ليكون هو الأمثل للتصميم المشروط، وينبغي على النموذج المرشح تحقيق قابلية التعميم، وحسن المطابقة والتقتير parsimony فضلاً عن إقامة التوازن بين التحيز وتقلبه، غير أنه في المجال العملي من الأفضل اختيار أكثر المعلمات معنوية ً لتصميم البحث من خلال مطابقة أفضل نموذج مرشح لهذا البحث. حيث تبين المحاكاة أن الانحدار اللوجستي الشرطي مختلط التأثيرات هو أكثر دقة لدراسات التلوث، مع نماذج الانحدار اللوجستي الشرطية ثابتة التأثيرات من المحتمل ان تولد استنتاجات خاطئة. وهذا لان الانحدار اللوجستي الشرطي مختلط التأثيرات يقدم أفكاراً تفصيلية على العناقيد التي تم تجاهلها إلى حد كبير من قبل الانحدار اللوجستي الشرطي ثابت التأثيرات.

المصطلحات الرئيسة للبحث: طريقة الإمكان الأعظم، الانحدار اللوجستي الشرطي، البيانات الطولية، نماذج التأثيرات المختلطة، معيار شبه الإمكان في ظل نموذج الاستقلال (QIC)، معيار اكايكي التجريبي (EAIC)، التلوث البيئي، التحليل العنقودي.

مجلة العلوم الاقتصادية والإدارية
العدد 98 المجلد 23
الصفحات 406-429

*البحث مستل من رسالة ماجستير.

1. المقدمة Introduction

عند تحليل البيانات المعنية بوصف العلاقة بين متغير الاستجابة وواحد أو أكثر من المتغيرات التفسيرية والذي يتم عن طريق الانحدار الذي أصبح جزءاً لا يتجزأ من هذا الاجراء والذي يكون متغير الاستجابة في كثير من الأحيان متقطعاً، آخذاً اثنين أو أكثر من القيم المحتملة. وان النموذج الملائم لمثل هذه الدراسات هو نموذج الانحدار اللوجستي. مما يعني أن هذا الوصف يجب أن يكون مبسطاً ومثالياً لظاهرة معينة وهو ما يعرف بالنموذج الإحصائي، ومن خلال هذه النماذج يمكننا فهم الظواهر واستخراج المعلومات منها والتحقق من صحتها والتنبؤ بها واستخلاص الاستدلالات الإحصائية منها.

وحيث أن هذه النماذج اللوجستية تعتبر نماذج مرنة لأنها تحتوي على كل من التأثيرات الثابتة والعشوائية، والتي تستخدم على نطاق واسع في حالة بيانات القياسات المتكررة غير المتوازنة. [1]

كما تعرف البيانات الطولية متعددة المتغيرات بانها عبارة عن قياسات متعددة ومتكررة ومتوالية حيث يتم تسجيلها وتحليلها في وقت واحد، مما يطرح تحديات منهجية وحسابية محددة ولاسيما عندما تكون المشكلة في حالة الابعاد العالية. على افتراض أن تشترك القياسات في مجموعة من التأثيرات العشوائية الكامنة (المخفية) من موضوع واحد التي بدورها تستخدم لإنشاء هيكلية الارتباط بين تلك القياسات المتكررة.

وغالباً ما يتم تحديد التأثيرات العشوائية في نماذج التأثيرات المختلطة الخطية وغير الخطية، بهدف السيطرة على العوامل المحددة التي من المتوقع أن تؤدي الى هذا التغيير العشوائي في المعاملات، على سبيل المثال كمية التأثيرات في القياسات المتكررة وتباينها ضمن الموضوع المحدد.

2. مشكلة البحث Research problem

يشكل التلوث تحدياً رئيساً في جميع أنحاء العالم مع معظم المنظمات التي تسعى إلى الحد من معدل التلوث في سبيل التخفيف من ظاهرة الاحتباس الحراري. واحد أهم مسببات التلوث المعروفة هي شركات إنتاج ومعالجة وتكرير النفط التي تعد من بين الدوافع الرئيسة لتلوث الهواء والتربة والمياه على حد سواء بسبب العمليات المكثفة لانتاج الطاقة المطلوبة من إنتاج النفط. ولنمذجة البيانات الخاصة بالتلوث فالانحدار اللوجستي لمتغير استجابة ثنائية (وجود تلوث 1، عدم وجوده 0) مع نمذجة البيانات الطولية المصنفة الى طبقات ومن ثم الى مجموعات فرعية أكثر تجانساً يعد من النماذج التي ستعالج مثل هكذا بيانات وان تقدير مثل هكذا نماذج يكون من الصعوبة حلها لوجود ارتباطات داخل المشاهدات ضمن الطبقات او العناقيد. لهذا يجب تحديد بنية الارتباط الصحيحة مما يسهم في إيجاد مقدرات موثوقة وغير متحيزة وأكثر تقارباً.

3. هدف البحث Research Goal

يسعى هذا البحث إلى إثبات صحة هذه الفكرة الشائعة بفحص مستويات التلوث بالنسبة إلى حجم إنتاج النفط على وجه التحديد من خلال البيانات المقدمة من شركة مصافي الوسط ولتحقيق هذا الهدف احصائياً لابد من:-

أ- تحديد المتغيرات المثالية للنموذج العشوائي باستخدام معيار QIC مع تحديد بنية الارتباط الواقعية للبيانات الفعلية باستخدام معيار اكايكي التجريبي EAIC.

ب- تقدير نموذج الانحدار اللوجستي الثابت والمختلط باستخدام طريقة الإمكان الأعظم (MLE) والغرض من ذلك هو رسم الاستدلال للتأثيرات الثابتة والتأثيرات المختلطة للانحدار اللوجستي الشرطي في حالة بيانات التلوث الطولية.

الفصل الأول / الجانب النظري

1. مفهوم الانحدار اللوجستي Logistic Regression:

هو أسلوب إحصائي احتمالي يستخدم في تحليل مجموعة من البيانات تعتمد على متغير مستقل واحد او مجموعة متغيرات مستقلة. ويتم قياس نتائج الانحدار اللوجستي او اللوجيت من خلال متغير ثنائي التفرع فيها اثنين فقط من النتائج الممكنة. [12] ولذلك يتم ترميز المتغير التابع أما 0 أو1 على سبيل المثال في هذا البحث حيث يتم ترميز بيانات التلوث اما 1 (ملوثة) أو 0 (غير الملوثة).

ان الانحدار اللوجستي الشرطي في حالة البيانات الطولية هو تحليل البيانات الطولية الثنائية التي تحتوي على متنبئ واحد أو عدة متنبئات، مع مشاهدات ليست مستقلة بل مُجَمَّعة. يفترض الانحدار اللوجستي كذلك أن حدود دالة التوزيع اللوجستي التركيبية الخطية $w^T xER$ لـ 0 و 1 نظراً لان الانحدار اللوجستي يتنبأ باحتمال الفرص الإيجابية. كما ان الانحدار اللوجستي ينفي الافتراضات الأساسية الملازمة للانحدار الخطي، فضلاً عن النماذج الخطية العامة التي تستند إلى الـ (OLS) مثل الافتراضات على الحالة الخطية linearity، وحالة تجانس التباين homoscedasticity، ومستويات القياسات، والحالة الطبيعية normality. [10]

ونتيجة لذلك، لا يَعد الانحدار اللوجستي وجود علاقة خطية نظراً لأنه يعتمد على تحويل لوغاريتم غير خطي للتنبؤ بنسبة ارجحية تمكينه لوصف مختلف أنواع العلاقات. أيضاً فان المتغيرات المستقلة والبواقي (حدود الخطأ) لنماذج اللوجيت لا تُعد كمتغيرات متعددة طبيعية، على الرغم من أن المتغيرات المتعددة الطبيعية تولد نتائج أكثر اتساقاً. [5]

2. افتراضات الانحدار اللوجستي Logistic regression assumptions:

يفترض الانحدار اللوجستي ما يأتي:

1.ان تكون المشاهدات والبواقي مستقلة مما يؤدي بالنموذج الى تقليل العلاقات الخطية المتداخلة المتعددة multi-collinearity من خلال اما تصنيف التفاعل بين المتغيرات الفئوية، أو إجراء تحليل عاملي مسبق للانحدار اللوجستي. [9]

2.أن هناك حالة خطية بين المتغيرات المستقلة ولوغاريتم الارجحية log odds، ما عدا ذلك فان الانحدار يُقيّم بأقل من قيمته وينفي العلاقة المهمة كغير معنوية ويدعم فرضية العدم بشكل فعال. هذا التقليل بشكل طبيعي هو عن طريق تصنيف المتغيرات المستقلة لبيانات ترتيبية قبل إدراجها الى النموذج. [4]

وأخيراً، يتطلب الانحدار اللوجستي أحجام عينات كبيرة نظراً لان تقديرات الإمكان الاعظم تكون أقل حصانة مقارنة مع (OLS) في الانحدار الخطي، وتتطلب الحد الأدنى من (10) حالات لكل متغير مستقل، و(30) حالة لكل تقدير معلمة. [10] لذا فان المعاملات التي تم إنشاؤها من الانحدار اللوجستي لتنبؤ تحويل logit باحتمال حدوث متغير او متغيرات الاستجابة تتوقف على الصيغة الاتية:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k \quad \dots (1.1)$$

حيث X_k المتغيرات المستقلة وb_k معلمات الانحدار ويتم الرمز إلى احتمال الحدوث p بينما يتم تسجيل احتمال تحويل اللوجيت فيما يأتي:

$$odds = \frac{p}{1-p} = \frac{\text{احتمال وجود صفة مميزة}}{\text{احتمال غياب صفة مميزة}} \quad \dots (1.2)$$

وكذلك:

$$logit(p) = \ln(\frac{p}{1-p}) \quad \dots (1.3)$$

ونتيجة لذلك، فان الانحدار اللوجستي يحدد معلمات التقدير التي تعظم ارجحية الحدوث الذي هو عكس الانحدار الخطي الذي يحدد المعلمات التي تقلل مجموع الخطأ التربيعي. [15]

3. **البيانات الطولية** Longitudinal data:

تتكون البيانات الطولية من التعَقُّب لعينة في نقاط مختلفة في الزمن. وبصفة عامة، يمكن أن تتألف العينة من الخصائص الفردية والاقتصادية والمعيشية ، أو كما هي الحال في هذا البحث، التي تتكون من الخصائص المناخية والتلوث، ويطلق على بيانات القياسات المكررة بالبيانات الطولية في الدراسات السريرية والبيئية Clinical & Environmental Studies اما في الدراسات الاقتصادية فيطلق عليها تسمية البيانات اللوحية panel data. [3]

جدول رقم (1) يبين تصميم مجموعة البيانات الطولية[1]

Subject/ Individual	Repeated measurement	Y_{it}	X_{itj}
1 1 . . . 1	1 2 . . . n_1	y_{11} y_{12} . . . y_{1n_1}	$x_{111}, x_{112}, \ldots\ldots\ldots, x_{11p}$ $x_{121}, x_{122}, \ldots\ldots\ldots, x_{12p}$ . . . . $x_{1n_1 1}, x_{1n_1 2}, \ldots\ldots\ldots, x_{1n_1 p}$
2 2 . . . 2	1 2 . . . n_2	y_{21} y_{22} . . . y_{2n_2}	$x_{211}, x_{212}, \ldots\ldots\ldots, x_{21p}$ $x_{221}, x_{222}, \ldots\ldots\ldots, x_{22p}$ . . . . $x_{2n_2 1}, x_{2n_2 2}\ldots\ldots\ldots\ldots, x_{2n_2 p}$
…	…	…	…
…	…	…	…
…	…	…	…
K ***K*** . . . ***K***	1 2 . . . n_k	y_{k1} y_{k2} . . . y_{kn_k}	$x_{k11}, x_{k12}, \ldots\ldots\ldots, x_{k1p}$ $x_{k21}, x_{k22}, \ldots\ldots\ldots, x_{k2p}$ . . . . $x_{kn_k 1}, x_{kn_k 2}, \ldots\ldots\ldots, x_{kn_k p}$

[1] الجدول من عمل الباحث.

المتغير المعتمد يمثل التلوث y_{it} هو الاستجابة t^{th} للعنقود i^{th} بينما تمثل المتغيرات المستقلة الخصائص المناخية x_{itj} وهي المتغير j^{th} للقياسات it^{th} حيث ان $i = 1,2,3, \ldots, K$ و $t = 1, 2,$ $j = 1, 2, 3 \ldots, p$ و p أذ ان k هو العدد الكلي للعناقيد (المواضيع)، و n_i هو العدد الكلي للقياسات المتكررة (الطبقات) للعنقود i^{th} و p هو عدد المتغيرات المستقلة.

الدور الرئيس للبيانات الطولية هو أنها تُسَهّل قياس التغييرات الحساسة للوقت فيما بين العينات ، مما تتيح قياس الأحداث الدورية، فضلاً عن توقيت الأحداث لتحديد التغييرات الملازمة لها بسبب تسجيل الأحداث الواقعة. [11] كذلك فان استقراء بيانات طولية في الماضي والحاضر يسهل تحليل 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.

4. 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 رغم اختلافها بمرحلية التجميع نفسها، وهي:

أ- طريقة التحليل العنقودي (K-Means).

ب- التحليل العنقودي الهرمي Hierarchical cluster analysis.

ت- التحليل العنقودي ذو الخطوتين Two-step cluster analysis.

5. طريقة الإمكان الأعظم Maximum Likelihood Method (MLM)

بافتراض المتغير العشوائي X حيث توزيع الـ X في المجموعة S له معلمات مجهولة θ مع قيم ضمن فضاء المعلمة Θ، عندها يمكن الإشارة إلى pdf لـ X على S بالرمز f_θ عندما $\theta \in \Theta$. إذ اعتبرنا أن المتغير X والمعلمة θ هي متجهات مُقيَّمة، عندها يمكن الحصول على دالة الامكان L عندما يتم عكس أدوار الـ x و θ في دالة الكثافة الاحتمالية. [8] هذا يعني أن θ يعد متغيراً بينما x المعلومات المعطاة ومن ثم تكون وجهة النظر هذه في التقدير كما يأتي:

تسعى طريقة الإمكان الاعظم لإيجاد قيمة $u(x)$ للمعلمة θ التي تعظم قيمة $L_x(\theta)$ لكل $x \in S$، مما يعني أن الإحصاءة $u(X)$ هي مقدر الإمكان الاعظم للمعلمة θ . ولذلك تستخدم هذه الطريقة لإيجاد قيم المعلمة الأكثر احتمالاً لإنتاج البيانات المشاهدة. [2] وبالنظر الى أن دالة اللوغاريتم الطبيعي تتزايد على $(0,\infty)$ فان قيمة $L_x(\theta)$ العظمى تحدث في نقاط مماثلة لقيمة $\ln[L_x(\theta)]$ العظمى، وهي دالة لوغاريتم الامكان التي غالباً ما تكون أسهل في الحساب عندما تحسب بالمقارنة مع استخدام دالة الامكان لبُنيَة دالة الكثافة الاحتمالية $f_\theta(x)$ الناتجة. [4]

ويمكن استخدام حساب التفاضل والتكامل عندما تحدث فضاء المعلمات Θ كمجموعة معلمات مستمرة وان $\Theta \subseteq R^k$ عندما يكون المتجه θ لـ k من المعلمات الحقيقية $\theta = (\theta_1, \theta_2, \theta_3 \ldots, \theta_k)$ ، حيث إذا كانت قيمة دالة الامكان العظمى L_x في نقطة المتجه θ ضمن الفضاء Θ، فان الحد الأعظم الموضعي لـ L_x يكون عند النقطة θ. كنتيجة لذلك، ونظراً لكون دالة الإمكان هذه قابلة للاشتقاق يمكن التوصل لحل النقطة العظمى لدالة الامكان L_x، من خلال:

$$\frac{\partial}{\partial \theta_r} L_x(\theta) =, r \in \{1,2,3 \ldots, k\} \qquad \ldots (2.2)$$

وهذا ما يعادل:

$$\frac{\partial}{\partial\theta_r}\ln[L_x(\theta)] =, r \in \{1,2,3 \dots, k\} \quad \dots (2.3)$$

في الحالة التي يكون فيها $X = (X_1, X_2, X_3,\dots,X_n)$ متغير نتيجة في عينة عشوائية للحجم n في توزيع متغير عشوائي X مع القيم في R ودالة الكثافة الاحتمالية g_θ عندما $\theta \in \Theta$، عندها يتم الحصول على قيم X من $S = R^n$. ونتيجة لذلك تتحقق دالة الكثافة الاحتمالية المشتركة لـ X بضرب دوال الكثافة الاحتمالية الحدية. [10] ومن ثم هذا يعطي دالة الامكان في مثل هذه الحالة على النحو الاتي:

$$L_x(\theta) = \prod_{i=1}^{n} g_\theta(x_i);\ x = (x_1, x_2, x_3, \dots, x_n) \in S, \theta \in \Theta \quad \dots (2.4)$$

كنتيجة لذلك، يمكن التعبير عن دالة لوغاريتم الإمكان على النحو الاتي:

$$\ln[L_x(\theta)] = \sum_{i=1}^{n} \ln[g_\theta(x_i)];\ x = (x_1, x_2, x_3, \dots, x_n) \in S, \theta \in \Theta \quad \dots (2.5)$$

6. توظيف طريقة الإمكان الأعظم لتقدير نموذج الانحدار اللوجستي الشرطي

Use the MLE to estimate the conditional logistic regression model

كما هي الحال في الانحدار اللوجستي الاعتيادي، من الممكن ادراج التأثيرات العشوائية في نماذج الانحدار اللوجستي الشرطية إذا تم استبدال معاملات الانحدار الثابتة بمعاملات الانحدار العشوائية. ويرجع ذلك الى تكييف مستوى ما هو مطلوب، وعدم وجود شروط تعترض في النماذج الشرطية في حين تعد التأثيرات العشوائية هي معاملات الانحدار العشوائية. وبالنسبة لحالة النموذج logit في هذا البحث، يعين لكل يوم قيمة، والتي هي قياس المنفعة U بمعنى (Utility) لكافة متغيرات الاستجابة المتاحة وبين المتغيرات في كل وقت من الأوقات، ويتم تحديد المتغير وفقاً لأعلى منفعة لنا. [7]

على افتراض أن $n = 1, 2, 3,\dots, K$ تتوافق مع الأيام، و $t = 1, 2, 3.,\dots,t_n$ هو السلسلة الزمنية لليوم n بينما $j = 1, 2, 3,\dots, J$ تتوافق مع القياسات الملوثة التي هي العينة من كل القياسات الملوثة لليوم n عند الزمن t، فان النموذج logit سيعتبر المنافع (utilities) بمثابة متغيرات عشوائية بينما U_{njt} هو يوم المنفعة n لتسجيل القياس الملوث jth عند الزمن t.

على اساس ان $x_{njt1}, x_{njt2}, x_{njt3}, \dots, x_{njtm}$ هي m من المتغيرات المستقلة مثل السمات البيئية التي يتم قياسها كمياً عند القياس الملوث jth وتسجيلها في اليوم n عند الزمن t، على افتراض أن المنفعة المخصصة للقياس الملوث تتوقف على شكل السمات البيئية، عندها يمكن صياغة المعادلات الاتية:

$$U_{njt} = \beta_1 x_{njt1} + \beta_2 x_{njt2} + \beta_3 x_{njt3} + \cdots \beta_m x_{njtm} + b_{n1} z_{njt1} + \cdots + b_{nq} z_{njtq} + \varepsilon_{njt}$$

$$U_{njt} = \mathrm{x}_{njt'}\beta + \mathrm{z}_{njt},\mathrm{b} + \varepsilon_{njt} \quad \dots (2.6)$$

حيث ان $\beta_1, \beta_2, \beta_3,\dots,\beta_m$ تدل على معاملات الانحدار الثابتة، بينما $b_{n1}, b_{n2}, b_{n3},\dots,b_{nq}$ تمثل التأثيرات العشوائية اليومية، وتمثل $Z_{njt1}, Z_{njt2}, Z_{njt3}, \dots Z_{njtq}$ القيم الثابتة لبُنيَة التأثيرات العشوائية، وهي مساوية لمجموعة المتغيرات المشاركة الفرعية x_{njt1} التي تتكون من معاملات عشوائية. فضلا عن ذلك، يدل ε_{njt} على حدود الخطأ العشوائي المستقلة، حيث ان كل من $\beta = (\beta_1, \beta_2, \beta_3,\dots,\beta_m)'$ ، $\mathbf{x}_{njt} = (x_{njt1}, x_{njt2}, x_{njt3}, \dots, x_{njtm})'$ و $\mathbf{b} = (b_{n1}, b_{n2}, b_{n3},\dots,b_{nq})'$ و $\mathbf{z}_{njt} = (Z_{njt1}, Z_{njt2}, Z_{njt3}, \dots Z_{njtq})'$ تتوزع توزيعاً متطابقاً. [13]

مع الاخذ بعين العناية انه في حالة غياب التأثيرات العشوائية سوف يقلل من الدالة الاحتمالية الأسية لان الأخطاء العشوائية تتوزع توزيعاً متطرفاً Extreme distribution، ومن ثم يمكن توسيع النموذج المحدد نظراً لافتراض ان له تأثيرات طفيفة، مما يجعل عملية تحديد النموذج على درجة عالية من المرونة.[14]

وبافتراض أن التأثيرات العشوائية b لها توزيعات مستقلة وهي متطابقة وتكون لها الكثافة $f(\mathrm{b};\theta)$، فضلاً عن متجه المعلمات المجهولة θ، في تلك الحالة سيكون احتمال ان اليوم له قياس التلوث j في مجموعة قياسات التلوث J هو$\{1, 2, 3,\ldots,J\}$. وهذا يعني أنه يمكن التعبير عن المنفعة U_{njt} التي هي أعلى من المنافع لجميع حالات $i \neq j$ كما يأتي:

$$P(\mathrm{x}_{njt}) = \int \frac{\exp(\mathrm{x'}_{njt}\beta + \mathrm{z'}_{njt}\mathrm{b})}{\sum_{i=1}^{j}\exp(\mathrm{x'}_{njt}\beta + \mathrm{z'}_{njt}\mathrm{b})} f(\mathrm{b};\theta)\mathrm{db} \qquad \ldots (2.7)$$

ويعد توزيع التأثيرات العشوائية عموماً توزيع طبيعي متعدد المتغيرات، والتي لديها متجه متوسط (0) الذي يتطلب تقدير معلمات التباين والتباين المشترك. ومع ذلك، من الممكن استخدام التوزيعات الأخرى على سبيل المثال توزيع lognormal، وتوزيع uniform، فضلاً عن التوزيع المثلثي triangular. [6] إذا كانت كافة قيم الـ Z_{njti} في المعادلة رقم (2.6) صفر وإذا الغي تباين الـ b نظراً لان b = 0، فأنه من الممكن تبسيط المعادلة رقم (2.7) لتشكيل المعادلة الاتية:

$$P(\mathrm{x}_{njt}) = \frac{\exp(\mathrm{x'}_{njt}\beta)}{\sum_{i=1}^{j}\exp(\mathrm{x'}_{njt}\beta)} \qquad \ldots (2.8)$$

المعادلة رقم (2.8) تمثل نموذج الانحدار اللوجستي الشرطي الاعتيادي (ثابت التأثيرات). [13] فضلا عن ذلك، يمنع إدراج تأثيرات عشوائية يومية في العامل الاحتمالي إمكانية تحديد مشاهدة متماثلة، نظراً لأنه بإضافة معامل انحدار على مستوى الحالات العشوائية يسمح للاختلافات بين الحالات في استجابة المتغيرات المشاركة x، مما يعني أن كل حالة من الحالات قد تستجيب بشكل مختلف للتغييرات في x. نظراً لان التأثيرات العشوائية يمكن عدها متغيرات عشوائية غير ملحوظة وهي مشتركة بين جميع المشاهدات لحالة معينة، والنموذج مختلط التأثيرات لا يفترض أن المشاهدات لتلك الحالة غير مرتبطة.. [17]

كما ان الزيادة في معدل التلوث لا تعتمد على الزيادة في الظروف المناخية وإنتاج النفط عند نمذجة logit متعدد الحدود المختلط أي انه سيتم إلغاء خاصية استقلالية البدائل غير ذات الصلة (IIA) Independence of Irrelevant Alternatives بوضوح على مستوى المجتمع، من خلال تحفيز الارتباط بالنسبة لبدائل القياس في مؤشر المنفعة التصادفية utility stochastic. [18] وهذا يمكن أن يتضح من خلال عدَ استجابات يوم واحد لميل التلوث عن طريق افتراض أن كل قياس تلوث هو واحد من فئات القياسات الثلاثة التي يتم ترميزها للمتغيرات المشاركة x_{jP} و x_{jC}.

على اساس ان القياس j هو قياس رطوبة عالية، درجة حرارة مرتفعة، إنتاج نفط مرتفع، حيث يكون $x_{jP} = 1$، و $x_{jC} = 0$؛ وقياس رطوبة منخفضة، درجة حرارة منخفضة، إنتاج نفط منخفض مع $x_{jP} = 0$، و $x_{jC} = 1$؛ وقياس عدم وجود رطوبة، عدم وجود درجة حرارة، عدم وجود إنتاج نفط فيه $x_{jP} = 0$، و $x_{jC} = 0$.

وبافتراض أن $J > 2$ القياسات المتاحة حيث القياس $j = 1$ هو الوقت حيث $x_{1P} = 1$، و $x_{1C} = 0$، بينما القياس $j = 2$ هو في زمن الأساس حيث $x_{2P} = 0$، و $x_{2C} = 0$، ثم على افتراض ان نموذج الانحدار اللوجستي الشرطي ثابت التأثير يحتوي على عامل احتمالي مقسوم على $\exp(\beta_P x_{jP} + \beta_C x_{jC})$، ويمكن التعبير عن نسبة احتمال مشاهدة القياس $j = 1$ ليوم معين الى احتمال مشاهدة القياس $j = 2$ في اليوم نفسه أو في يوم آخر كما يأتي:

$$\frac{\exp(\beta_P)\Big/\sum_{i=1}^{j}\exp(\beta_P x_{jP}+\beta_C x_{jC})}{\exp(0)\Big/\sum_{i=1}^{j}\exp(\beta_P x_{jP}+\beta_C x_{jC})}=\frac{\exp(\beta_P)}{1}=\exp(\beta_P) \qquad \dots(2.9)$$

وان هذا النموذج لا يتوقف على ما إذا كانت هناك نسبة عالية في العوامل البيئية، وإنتاج النفط في القياسات الأخرى المتاحة (بدائل القياس المتاحة). وكذلك يفترض النموذج بدلاً من أن يشير إلى ميل ثابت بالرمز β_P، هناك ميل عشوائي لمتغيرات x_P الذي تتم الإشارة إليه بالرمز $\beta_P + b$. وبما ان b هو ثابت عبر جميع القياسات في يوم معين، عندها يمكن التعبير عن نسبة احتمال مشاهدة القياس $j = 1$ الى احتمال مشاهدة القياس $j = 2$ كما يأتي:

$$\frac{\exp(\beta_P+b)\Big/\sum_{i=1}^{j}\exp(\beta_P x_{jP}+\beta_C x_{jC})}{\exp(0)\Big/\sum_{i=1}^{j}\exp(\beta_P x_{jP}+\beta_C x_{jC})}=\frac{\exp(\beta_P+b)}{1}=\exp(\beta_P+b) \qquad \dots(2.10)$$

هذا النموذج هو أيضاً لا يتوقف على عناصر القياسات البديلة، ولكنه يعتمد بالأحرى على التأثير العشوائي b الخاص بيوم محدد، الذي هو مخفي او غير ملحوظ. وباعتبار أن نسبة احتمال يوم تم اختياره عشوائياً يأخذ القياس $j = 1$ بالنسبة إلى احتمال يوم آخر تم اختياره عشوائياً يأخذ القياس $j = 2$، وهذا يمكن التعبير عنه كما يأتي:

$$\frac{\int\left\{\exp(\beta_P+b)\Big/\sum_{i=1}^{j}\exp\left((\beta_P+b)x_{jP}+\beta_C x_{jC}\right)\right\}f(b)\,db}{\int\left\{\exp(0)\Big/\sum_{i=1}^{j}\exp\left((\beta_P+b)x_{jP}+\beta_C x_{jC}\right)\right\}f(b)\,db}$$

$$=\exp(\beta_P)\left[\frac{\int\left\{\exp(\beta_P+b)\Big/\sum_{i=1}^{j}\exp\left((\beta_P+b)x_{jP}+\beta_C x_{jC}\right)\right\}f(b)\,db}{\int\left\{\exp(0)\Big/\sum_{i=1}^{j}\exp\left((\beta_P+b)x_{jP}+\beta_C x_{jC}\right)\right\}f(b)\,db}\right] \qquad \dots(2.11)$$

يتوقف هذا النموذج الآن على عناصر القياسات البديلة، وإلغاء "استقلالية البدائل غير ذات صلة" بصورة فعالة على مستوى كل ساعة. هذا يعني أنه من خلال إضافة معاملات عشوائية الى نموذج الانحدار اللوجستي الشرطي، فان احتمال المتوسط في كل ساعة لمشاهدة قياس معين يتوقف على بدائل القياس المتاحة.

كذلك مع الأخذ بنظر العناية تقدير الإمكان الاعظم لمعلمات النموذج المحددة في المعادلتين (2.6) و (2.7) استناداً إلى البيانات التي تتكون من تصميم المعاينة المتطابقة[2].

[2] هي طريقة أخذ العينات التي غالباً ما تستخدم للمساعدة في تقييم مدى التأثير السببي لمجموعة المتنبئ على ضوء الاستجابات المقدمة عادةً عندما تكون التجارب العشوائية غير متاحة أو لا يمكن اجراءها.

ويمكن تبسيط الترميز بافتراض أن يتم الإشارة إلى القياسات الملوثة باليوم n عند الوقت t في القياسات المسجلة J بالرمز $j = 1$، مما يعني ان القياسات التي تكون غير ملوثة يشار إليها بـ $j = 2, 3, 4,\ldots,J$ الذي يحافظ على مستوى العمومية الحالي. فمن الممكن الحصول على تقديرات الامكان الاعظم للعامل الاحتمالي فضلاً عن معلمات توزيع التأثيرات العشوائية بحساب قيم الـ β و θ من خلال تعظيم الإمكان في المعادلة (2.12) المذكور انفا:

$$L(\beta,\theta) = \prod_{n=1}^{K} \int \prod_{t=1}^{t_n} \frac{\exp(x'_{n1t}\beta + z'_{n1t}b)}{\sum_{i=1}^{j} \exp(x'_{njt}\beta + z'_{njt}b)} f(b;\theta)db \quad \ldots (2.12)$$

حيث يمكن استخدام العديد من أساليب الاستدلال المستندة إلى أساس الإمكان للمعلمة β عند تطبيق دالة الامكان في المعادلة رقم (2.12) ، على سبيل المثال فترات ثقة Wald ((من خلال عكس لوغاريتم إمكان هيسيان السالبة)) واختبار نسبة الامكان، أو حتى اختيار نموذج عن طريق معايير معلومات أكايكي.[14] وتماشياً مع المتطلبات الأساسية للتقتير، فمن المهم إنشاء الضرورة للتأثيرات العشوائية في الدالة الاحتمالية نظراً لان الانحدار الشرطي ثابت التأثيرات يمكن أن يعزز كفاءة التقدير فضلاً عن قابلية تفسيرها نموذجاً مسبقاً إذا ثبت أنها ليست شرطاً لاستخدام تأثيرات عشوائية.[20] النموذج ثابت التأثيرات مشابه لنموذج الانحدار اللوجستي الشرطي مختلط التأثيرات مع التباين، فضلاً عن معلمات التباين المشترك $f(b;\theta)$ مساوية للصفر. ولذلك فمن الممكن استخدام اختبار نسبة امكان لنموذج متداخل nested model لإثبات ضرورة زيادة تعقيد النموذج باستخدام تأثيرات عشوائية.

7. نهج معادلات التقدير المعممة Generalized Estimating Equations (GEE)

ان نهج معادلات التقدير المعممة (GEE) هو تقنية شائعة الاستخدام لتقييم البيانات الطولية، فضلاً عن البيانات العنقودية. وتتطلب طريقة GEE بنية الارتباط العاملة (WCS)، التي يتم بموجبها تحديد مواصفات الهياكل المطلوبة التي تكون اما مستقلة، او قابلة للصرف exchangeable، او الانحدار الذاتي من الرتبة الأولى Autoregressive (AR-1). 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:

$$U(\beta) = \sum_{i=1}^{k} \left(\frac{\partial \mu_{it}}{\partial \beta_p}\right)' V_i^{-1}\{Y_i - \mu_i(\beta)\} \quad \ldots (2.13)$$

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.[22]

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 العناية 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 $(t, t' = 1,2, 3, \ldots$ $t \neq t'; n_i)$ ، فانه يمكن تعريف المصفوفة القطرية التي هي تباين Y_i من درجة $n_i \times n_i$ على النحو: A_i = 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 الاتي:

$$\Sigma_i = A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}} / \phi \quad \ldots (2.14)$$

الصيغة المذكورة انفا تنتج لـ k+1 من معادلات GEE 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. [23]

8. أنواع هياكل الارتباط العاملة

بما ان المشاهدات n_i لكل موضوع من الموضوعات (العناقيد) i حيث ان $i = 1, 2, 3 \ldots, K$ 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) [24] مما يعني أن المتجه $\alpha = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_s)^T$ 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) هيكل الاستقلالية Independence structure (IN)، حيث ان، $R(\alpha) = I_n$، يشير I_n الى مصفوفة الوحدة أي عندما لا يكون هناك أي ارتباط فيما بين العناقيد.

(2) هيكل قابل للصرف Exchangeable structure (EX)، حيث ان α المعلمة المجهولة.

(3) هيكل الانحدار الذاتي من الرتبة Autoregressive Order 1 structure (AR 1)، حيث ان α المعلمة المجهولة.

(4) الهيكل الثابت Stationary structure (ST)، حيث ان α لها $n - 1$ من المعلمات المجهولة. لذلك يمكن التعبير عن المصفوفات لهياكل كل منها على النحو الاتي:

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

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

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

9. معيار شبه الإمكان في ظل نموذج الاستقلال (*QIC*)

Quasi-likelihood under independence Criterion

واحد من المعايير الأكثر شيوعاً المستخدمة في اختيار النماذج المستندة إلى الإمكان هو **AIC**. حيث اقترح **Pan (2001)** [21] معيار شبه الإمكان المعروف بـ **QIC**، الذي يمكن أن يستخدم في اختيار نموذج المتوسط المناسب μ_{ij}، فضلاً عن بنية الارتباط العاملة. [21] ووفقاً لـ **Hardin** و **Hilbe (2003)** [22]، يمكن التعبير عن دالة شبه الإمكان على النحو الاتي:

$$Q(\mu, \phi; y) = \int_{y}^{\mu} \frac{\phi(y - \mu^*)}{v(\mu^*)} d\mu^* \quad \dots (2.15)^{(19)}$$

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

ويقول **Pan (2001)** [21]، أنه ينبغي أن يحسب شبه الإمكان **Quasi-likelihood** للبيانات الطولية عندما يفترض أن كلاً من العناقيد، وكذلك النقاط الزمنية تكون مستقلة على النحو الاتي:

$$Q(\mu, \phi) = -2 \sum_{i=1}^{K} \sum_{t=1}^{n_i} Q(\mu, \phi; Y_{it}) \quad \dots (2.16)$$

وكنتيجة لذلك يمكن التعبير عن دالة شبه الإمكان للعنقود ***i*** في المشاهدة ***t*** التي يتم تقييمها باستخدام معلمات الانحدار **β** على النحو الاتي: $Q(\beta, \phi; Y_{it}, x_{it}) = Q_{it}/\phi$، باعتبار ان Q_{it} تعود إلى توزيع **Binomial** أي ان $Q_{it} = y_{it} \ln\{\mu_{it}/(1-\mu_{it})\} + \ln(1-\mu_{it})$ مع دالة تباين $\mu_{it}(1-\mu_{it})$ مع 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:

$$\mathrm{QIC}(\mathbf{R}) = -2 \sum_{i=1}^{K} \sum_{t=1}^{n} Q(\beta, \phi; Y_{it}, x_{it}) + 2\mathrm{tr}\{\Omega \mathrm{V_T}(\mathbf{R})\} \quad \dots (2.17)$$

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

هي عبارة عن المصفوفة ***p*** x ***p*** وان **p** تمثل عدد معلمات الانحدار. و$\mathrm{V_T}(\mathbf{R})$ هو مصفوفة التباين المشترك القائمة على النموذج لمعلمات الانحدار المقدرة باستخدام مصفوفة التباين المشترك المستقلة. كما يفترض **Pan (2001)** [21] انه يمكن التعبير عن **QIC** كما يأتي:

$$\mathrm{QIC}(R) = -2Q(\hat{\beta}, \hat{\phi}) + 2\mathrm{tr}(\widehat{\Omega}_I \hat{V}_r) \quad \dots (2.18)$$

الحد الأول في المعادلة رقم **(2.18)** يمثل شبه الامكان، وكذلك يمكن أن يعبر عنه كدالة لـ $\hat{\beta}$ من خلال الاستبدال بـ $\hat{\mu}$. ومن ثم من الممكن الحصول على "مقدر مصفوفة التباين المشترك لمعلمات الانحدار باستخدام مصفوفة التباين المشترك المفترضة " الذي يرمز له $\widehat{\Omega}_I$ عن طريق استبدال كل من **β** و **ϕ** و **α** مع تقديراتها الخاصة بكل منها.

من ناحية أخرى، فان $\hat{V}_r$ هو "تقدير التباين الحصين في ظل بنية الارتباط العاملة المحددة **(R)**" الذي يمكن استخدامه في اختيار بنية الارتباط مع الحد الأدنى لقيمة **QIC(*R*)** وفقاً لبنية الارتباط العاملة بحسب قول **Pan (2001)**. [21]

10. معيار الامكان التجريبي لأكايكي Empirical Likelihood Akaike (EAIC) information Criteria

AIC في EL (2012) [16] الذي يستبدل الامكان التجريبي (Lazarو Chen يوضح النهج المتبع من قبل مع الإمكان المعلمي لتشكيل معيار إضافي لتحديد مصفوفة الارتباط العاملة ان هذا المعيار تم بناؤه ليكون أكثر (2012) [16] يركز بالدرجة الأساسLazarوChen، فانEAIC . عند تقييم QICفعالية إذا ما قورن بـ لنموذج كامل تحت افتراض أن ELR (Empirical likelihood ratio)على اشتقاق نسبة الامكان التجريبي مـن المعلمـات الحـرة مدرجـة فـي $p+n-1$ لهـا ((ST) المسـتقرة (α) مصـفوفة الارتبـاط العاملـة R_F $\theta^T = (\beta^T, \alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1})$[16] .

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

$$g^F\left((Y_i, X_i),\ \beta,\ \alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1};\ R_F(\alpha)\right) =$$

$$\begin{pmatrix} (\partial\mu_i/\partial\beta^T)^T \mathrm{A}_i^{-\frac{1}{2}} R_F^{-1}(\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{n-1}) \mathrm{A}_i^{-\frac{1}{2}} (Y_i - \mu_i) \\ \sum_{t=1}^{n-1} e_{it}(\beta) e_{i,t+1}(\beta) - \alpha_1 \hat{\phi}(\beta)(n - 1 - p/K) \\ \vdots \\ \sum_{t=1}^{1} e_{it}(\beta) e_{i,t+n-1}(\beta) - \alpha_{n-1} \hat{\phi}(\beta)(1 - p/K) \end{pmatrix}_{(p+n-1)\mathrm{X}\,1} \quad \dots (2.19)$$

مع البواقي بيرسون معبراً عنها بالصيغة انفاً:

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

$$\hat{\phi}(\beta) = \sum_{i=1}^{K} \sum_{t=1}^{n} e_{it}^2 / (K n - p) \quad \dots (2.20)$$

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

$$\mathcal{R}^F(\beta, \alpha^T) = \sup\left\{ \prod_{i=1}^{K} K\omega_i :\ \omega_i \geq 0, \sum_{i=1}^{K} \omega_i = 1, \sum_{i=1}^{K} \omega_i g^F((Y_i, X_i), \beta, \alpha^T;\ R_F(\alpha)) = 0 \right\} \dots (2.21)$$

ولذلك يكون لمعادلات تقدير الهياكل المرشحة الخاصة بكل منها مصفوفة ارتباط عاملة مماثلة مع بنية ST التي هي جنباً إلى جنب 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 معادلة التقدير المعممة باستثناء التعديل الذي سيكون مع النتيجة في قيم ELR التي تساوي 1.

من جهة أخرى فان Chen وLazar (2012) [16] قام بحساب تقديرات $\hat{\theta}_G$ التي تتوافق مع كل من هياكل الارتباط العاملة المرشحة الأربعة على النحو الاتي: $(\hat{\beta}_{\mathrm{IN}}, O^T)^T$ لبنية الـ (IN)، و $(\hat{\beta}_{\mathrm{EX}}^T,\ \hat{\alpha}_{\mathrm{EX}}^T)^T$ لبنية الـ (EX)، و $(\hat{\beta}_{\mathrm{AR(1)}}^T,\ \hat{\alpha}_{\mathrm{AR(1)}}^T)^T$ لبنية الـ (AR-1)، و $(\hat{\beta}_{\mathrm{ST}}^T,\ \hat{\alpha}_{\mathrm{ST}}^T)^T$ لبنية الـ (ST).

بعد ذلك يتم الحصول على قيم ELR عن طريق إدراج تقديرات معادلة التقدير المعممة في دالة ELR المعبر عنها بالرمز $\mathcal{R}^F(\beta, \alpha^T)$. كما قام Chen وLazar (2012) [16] بمواصلة تطوير معايير اختيار مصفوفة الارتباط العاملة مع أعلى قيم ELR الناتجة، إلا وهي: EAIC ، حيث ان:

$$EAIC = -2\log \mathcal{R}^F(\hat{\theta}_G) + 2\dim(\theta) \quad \ldots(2.22)$$

هنا، dim (θ) هو عنصر من عناصر المعلمات الحرة[3] $(\beta^T, \alpha^T)^T$ التي يتم تقديرها، وان الحد الأدنى لقيم EAIC تدل على النموذج المحتمل (النموذج المفضل كونه النموذج الذي يقلل من فقدان المعلومات). [16]

الفصل الثاني / الجانب التطبيقي

1. أسلوب جمع البيانات

تضمن البحث الحصول على البيانات المناخية من موقع (Weather History for KQTZ) على شبكة الانترنيت وبيانات التلوث من شركة مصافي الوسط التي تمثل القياسات اليومية التي تستند إلى الوقت للمدة من أيلول/سبتمبر 2011 إلى كانون الأول/ديسمبر 2013. وحيث ان مجموعة البيانات الطولية تتعقب العناصر المسببة للتلوث الجوي من خلال القياسات المتكررة للمركبات مثل مركبات الكاربون (CO_x)، ومركبات الكبريت (SO_x)، ومركبات النيتروجين (NO_x) فضلا عن المتغيرات التفسيرية بما في ذلك سرعة الرياح ودرجات الحرارة وكميات النفط المنتجة.

بالنظر لخلفية البيانات، فقد تألفت مجموعة البيانات بناء على ذلك من (14) سمة مميزة، حيث كانت اثنان فئوية (التاريخ والوقت)، وكانت سبعة ثنائية (متغيرات الاستجابة) هي 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 ($\mu g/m^3$)، وخمسة كانت مستمرة (المتغيرات التفسيرية) هي 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. ولذلك كانت هناك حاجة الى الانحدار اللوجستي لتوليد نموذج تنبؤي لمتغيرات الاستجابة وتصنيف الدقة والحساسية، فضلاً عن الطابع الخاص بالتنبؤ.

2. ترميز البيانات Data coding

تم قياس مستوى بيانات التلوث لفترة أكثر من سنتين (28 شهراً) وعلى مدى 80 يوماً فقط، وبناء على ذلك، تم افتراض نموذج logit للعينة الذي بموجبه تم تحويل البيانات إلى استجابات ثنائية ومتنبئات ثنائية بغية التنبؤ بنتائج التلوث على أساس متغيرات التنبؤ.

وتمت معالجة البيانات استناداً إلى الحدود القصوى المسموح بها لملوثات الهواء المنبعثة من مصادر الاحتراق حيث تعد القيم الأدنى من الحدود القصوى غير ملوثة [0] وتعد القيم الأعلى من الحدود القصوى ملوثة [1]، وكما هو مبين في الجدول رقم (2):

جدول رقم (2) يبين الحدود القصوى المسموح بها لتركيز كل ملوث من الملوثات التي يسمح بطرحها إلى البيئة بموجب المعايير الوطنية[4]

الملوثات	(PM2.5) ug/m3	(H2S) ug/m3	(NOx) ug/m3	(NH3) ug/m3	(CO) ug/m3	(CO2) ug/m3	(O3) ug/m 3
الحدود القصوى	150	10	500	50	500	500	20

[3] **المعلمات الحرة free parameters: هي المتغيرات القابلة للتعديل والمستخدمة في النماذج الرياضية التي يمكن تعديلها من أجل جعل النموذج يحتوي البيانات. بدلا من الثوابت، وخلافاً لغيرها من المعالم التي تقتصر على تمثيل البيانات ذات مغزى، فان المعلمات الحرة يمكن تعديلها للسماح للنماذج لاحتواء البيانات، وأنها تقدم أفكاراً إضافية لتوفير رؤى مفيدة على البيانات. ويمكن الحصول على قيم المعلمات الحرة المستخدمة في النماذج من التجارب السابقة، او مجرد تخمينات، أو بشكل عشوائي.**

[4] **المصدر: المحددات البيئية استناداً الى البند تاسعاً /المادة -2 من قانون حماية وتحسين البيئة رقم (27) لسنة 2009 المنشور في جريدة الوقائع العراقية العدد 4142 في 2010/1/25.**

ثم يمكن استخدام دالة لوجستية لنموذج احتمالات النتائج كدالة للمتغيرات التفسيرية.

1. تحديد النماذج من خلال معايير اختيار النموذج المثلى

تعتمد نماذج منحنى النمو growth curve على الوقت في التلوث لمقارنة النموذج، وأنواع بنية التباين المشترك الأربعة المحددة (ST,EX,AR-1,IN) لنماذج منحنى النمو التجميعية (يكون فيه النمو في البداية بطيئا (طور التعجيل الموجب) ثم يزداد بسرعة وعندئذ يعرف بالطور اللوغاريتمي (logarithmic phase) ثم يتباطأ بالتدريج حيث تزداد المقاومة البيئية بنسبة عكسية (طور التعجيل السالب) حتى يصل الى مستوى متوازن لحد ما ويبقى عنده). وتمت مطابقة النماذج التجميعية، مع قيم معايير اختيار النماذج المطابقة المبينة في الجدول رقم (3):

الجدول رقم (3): ملخص مطابقة معايير اختيار النموذج

Model	QIC	EAIC
Particulate matter (PM2.5) ug/m3 - Model_1	0.113	0.052
Hydrogen Sulfide (H2S) ug/m3 - Model_2	0.008	0.014
Nitrogen Oxides (NOx) ug/m3 - Model_3	0.02	0.049
Ammonia (NH3) ug/m3 - Model_4	0.027	0.029
Carbon monoxide (CO) ug/m3 - Model_5	0.076	0.024
Carbon dioxide (CO2) ug/m3 - Model_6	0.009	0.025
Ozone (O3) ug/m3 - Model_7	0.042	0.007
Total loss information	0.295	0.2

حيث تم استخدام مجموعة نماذج مرشحة تستند الى التوزيع اللوجستي. ووفقا للجدول رقم (3)، تُسَجل EAIC أدنى مستوى لفقدان المعلومات.

لغرض تعزيز رؤية البيانات وتحديد الهياكل الأساسية الرئيسة والمتغيرات، وإنشاء نماذج مثلى فضلاً عن التعرف على العوامل المثلى وتم استخدام تحليل البيانات الاستكشافية في تحليل البيانات الأولية بضمنها تحليل السلاسل الزمنية والتحليل العنقودي المتعدد كالاتي:

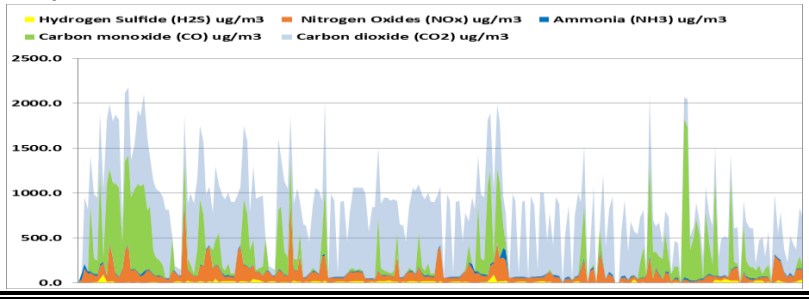
4. تقييم مستويات التلوث

تم إجراء تحليل نماذج السلاسل الزمنية المبينة في الجدول ضمن الملحق رقم (1):

حيث يفترض النموذجين 3 و6 فيما يتعلق بالملحق رقم (1) 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. ولغرض تقييم مستويات التلوث مع مرور الوقت تم اجراء الرسم في برنامج Excel في الشكل رقم (1).

الشكل رقم (1): يوضح حالة مستويات التلوث 3D مع مرور الوقت.

مستويات التلوث ug/m^3



الوقت بالساعات.

5. **تحديد هياكل الارتباط العاملة** Determine the working correlation structure

تم تحديد هياكل التباين المشترك لقياسات التلوث المتكررة كما هو مبين في الجدول رقم (4):

الجدول رقم (4): يبين إحصاءات مطابقة النموذج لهياكل التباين المشترك

Fit Statistic	Mean	SE	Minimum	Maximum	AR-1 structure	IN structure	EX structure	ST structure
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
Max APE	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
Max AE	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 ارتباطاً خطياً إيجابياً لمعدل التلوث مع مرور الوقت.

وكذلك يوحي تحليل البيانات الاستكشافية 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 (4). وهذا يعني أن معدل التلوث عند الزمن t يتحدد مباشرة عن طريق معدل التلوث عند الزمن $t-1$. 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:

6. **التحليل العنقودي المتعدد** Multiple Cluster Analysis

بإجراء المزيد من التحليل للبيانات تم تطبيق مناهج العنقدة المتدرجة clusterwise، وتحديداً النهج العنقودي ذو الخطوتين 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.

الجدول رقم (5): يبين التوزيع العنقودي للبيانات

		N	% of Combined	% of Total
Cluster	1	97	41.8%	41.6%
	2	106	45.7%	45.5%
	3	29	12.5%	12.4%

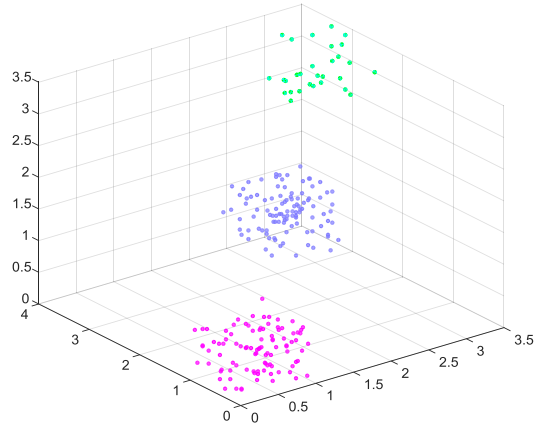
ويبين الجدولين رقم (5) و(6) نتائج التحليل العنقودي ذو الخطوتين لتقييم البيانات الذي بموجبه تم اعتبار المتغيرات التفسيرية مستمرة بينما كانت متغيرات الاستجابة المشفرة فئوية. وكشف أن طبيعة البيانات تتكون من 3 عناقيد تختلف فيما يتعلق بالمتوسط.

الجدول رقم (6): 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	26.28	10.005	6.201	4.2433	32.68	18.0999	12.06	6.5234	891.85	131.133
	2	25.47	9.415	4.781	4.1729	31.93	17.7373	11.85	6.9284	830.44	110.944
	3	29.45	9.0251	5.379	2.6580	25.05	13.5015	10.95	5.5602	713.46	163.333
	Combined	26.30	9.6611	5.450	4.0862	31.39	17.5254	11.83	6.5858	841.49	138.397

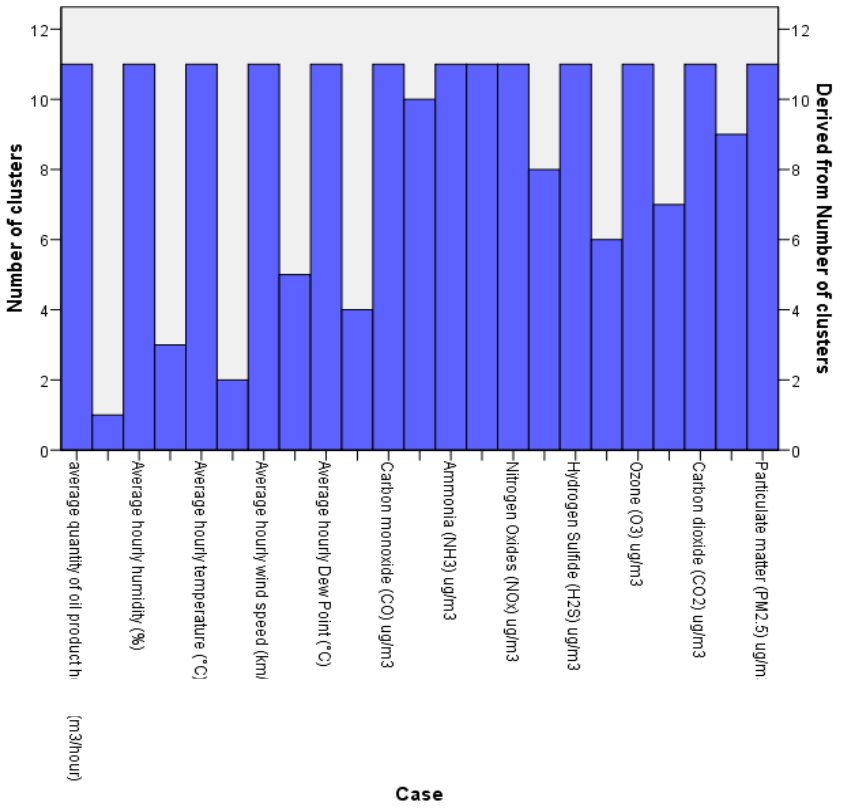
فضلا عن ذلك، تم اجراء تحليل اضافي لعناقيد البيانات باستخدام نهج ترابط وارد Ward's linkage في نظام MATLAB الذي أنتج الشكل رقم (2):

الشكل رقم (2): يوضح التوزيع العنقودي لترابط وارد ثلاثي الابعاد



كلا النهجين أشار إلى أن للبيانات 3 عناقيد قياسية، بينما أشارت التحليلات العنقودية المتعددة حققت نقاط تقارب بسرعة مراعاةً لأعداد العناقيد المختلفة. فعلى سبيل المثال، عندما $C = 2$ يتقارب النموذج في (10) تكرارات في حين كان يتطلب (100) تكرار عندما $C = 231$ وفقاً لبدء التشغيل الاولي لمعلمات العضوية العنقودية. ويبين الشكل رقم (3) مقاييس صحة نماذج متعددة العنقودية المختلفة انسجاماً مع بنية التباين المشترك (AR-1) ووفقاً لهذا الشكل، أن 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 99.6% (FIT = 0.996) لإمكانية تفسير تقلبية النموذج.

الشكل رقم (3): يبين جدول تكتلية العضوية العنقودية



7. تقدير نموذج الانحدار اللوجستي الشرطي باستخدام الإمكان الاعظم

Estimating model the conditional logistic regression using MLE

لغرض تقييم افتراض انبعاثات ملوثات الهواء المتجانسة من خلال تحليل تأثير تصفية النفط على التلوث البيئي. تم تسهيل التباعد لافتراض بان الاختلافات فيما بين الأيام تترك اثراً في انظمة التلوث الجوي عن طريق وضع فروض ($x_{1C} = 1$, $x_{1P} = 0$) على معدلات التلوث التي تخالف افتراض (IIA) Independence of Irrelevant Alternatives. وتم اجراء نمذجة واضحة للقياس على أساس اليوم، عن طريق تطبيق معدلات التلوث للملوثات الفردية عبر (80) يوم فقط لان كل يوم فيه عدة قياسات تلوث متكررة على مدى عدة ساعات ضمن الفترة ما بين 5 - أيلول / سبتمبر 2011 و 19 - كانون الأول / ديسمبر 2013 بالنسبة للظروف المناخية وكميات إنتاج النفط في شركة مصافي الوسط خلال تلك الفترة.

تتألف العوامل البيئية من ثلاث فئات من القياسات موزعة عشوائياً تم تحديدها من خلال التحليل العنقودي وتم اختبار اوضاع القياس الثلاثة استناداً إلى انظمة التلوث المختلفة للأيام في تلك الفترة، مع استنتاجات إحصائية دقيقة هي:

1. القياس C1 له قياسات تلوث عالية (العنقود 1)، ان معدل التلوث يتماشى مع فرضية "IIA" حيث يَفترض نظام تلوث متجانس في النظام البيئي 1.

2. و C2 له قياسات تلوث معتدلة (العنقود 2) ايضاً معدل التلوث يتماشى مع فرضية "IIA" حيث يَفترض النظام البيئي 2 الاختلافات فيما بين الأيام في انظمة التلوث.

3. في حين كان C3 له قياسات تلوث منخفضة (العنقود 3). معدل التلوث في النظام البيئي 3 قد انتهك فرضية "IIA" على المستوى اليومي نظراً لان حدوث C1 بدلاً من C2 كان يتوقف على ما إذا كان هناك حدوث لـ C3 خلال اليوم ولذلك تم تطبيق نظام تلوث متجانس للنظام البيئي 3.

حيث يقول Bhat (2001) [3]، ان استخدم تقنيات المحاكاة التي ترتكز على أرقام هالتون شبه العشوائية Halton quasi-random numbers فعالة في تقييم دالة الامكان.[6] ولذلك فان تعظيم الامكان في المعادلة رقم (2.12) يمكن تحقيقه باستخدام طريقة أرقام هالتون شبه العشوائية من خلال تنفيذ نسخة معدلة من البرنامج [19] لنظام MATLAB. ولذلك يمكن اعتبار تقنية أرقام هالتون شبه العشوائية مستقرة عددياً التي برهنت انها تولد تقدير إمكان أعظم تقريبي دقيق للانحدار اللوجستي الشرطي.

8. تقدير معلمات النموذج مختلط التأثيرات

Estimate of Mixed-effects model parameters

ان متوسط العوامل البيئية وإنتاج النفط في كل ساعة يتطابق بالنسبة للملوثات الأربعة التي تم تحديدها في التحليل التراكمي لعنقود واحد وهي ثاني أوكسيد الكاربون (CO_2)، وأكاسيد النيتروجين (NO_x) وكبريتيد الهيدروجين (H_2S) وأول أوكسيد الكاربون (CO)، وكذلك التحليل العنقودي المتعدد بوصفها تحتوي على اقل فقدان للمعلومات نظراً لأن قياسات C1 – C3 تم تحديدها في جميع القياسات لكل ساعة.

ولغرض تقييم معدل التلوث وفقاً لخطر التلوث تم تطبيق الانحدار اللوجستي الشرطي في صياغة الدوال الاحتمالية، مع دوال مختلطة التأثيرات التي تسهل الاختلافات فيما بين الايام في معامل الـ C1 وفقاً لـ $N(\beta_1,\sigma^2)$، في حين تم تطبيق C2 كقياسات خط الأساس. وتم تسهيل مطابقة النماذج من خلال تعظيم الامكان في المعادلة رقم (2.12) باستخدام نظام MATLAB النسخة R2014a وكانت النتائج في الجدول رقم (8):

الجدول رقم (8): يوضح قيم معاملات معدلات تلوث CO_2 و NO_x المقدرة من خلال الانحدار اللوجستي الشرطي مع التوزيع الطبيعي للمعاملات المتعلقة بالمتغيرات البيئية وكميات إنتاج النفط لثلاث أنظمة تلوث بيئية.

النظام البيئي C1 (جدول 8)				
Variable	ß قيمة معاملات معدلات التلوث	SE الخطأ المعياري	95% CI	
			Lower	Upper
Fixed Coefficients				
H_2S	-0.028	0.047	0.244	0.432
CO	0.119	0.030	0.030	0.150
Random Coefficients				
Oil produced	-3.830	13.315	865.429	918.288
temperature	-262.062	1.016	24.263	28.296
humidity	-140.829	1.838	29.040	36.336
NO_x	0.084	0.015	0.000	0.050
CO_2	0.670	0.551	0.494	0.832
SD of coefficients				
Oil produced	0.914	75.995		
temperature	-96.395	5.760		
humidity	63.555	10.436		
NO_x	-1.007	0.055		
CO_2	0.140	0.710		
Max. log likelihood	-1629.056			
النظام البيئي C2 (تابع جدول 8)				

Variable	ß	SE	95% CI	
			Lower	Upper
Fixed Coefficients				
H_2S	0.005	0.048	0.456	0.648
CO	2.200	0.000	0.000	0.000
Random Coefficients				
Oil produced	-488.248	10.776	809.078	851.811
temperature	-274.524	0.915	23.652	27.279
humidity	-147.855	1.723	28.523	35.355
NO_x	-18.210	0.000	0.000	0.000
CO_2	0.670	0.000	1.000	1.000
SD of coefficients				
Oil produced	-110.770	94.452		
temperature	52.871	5.446		
humidity	37.805	8.475		
NO_x	0.000	0.015		
CO_2	0.000	0.158		
Max. log likelihood	-1630.338			

النظام البيئي C3 (تابع جدول 8)

Variable	ß	SE	95% CI	
			Lower	Upper
Fixed Coefficients				
H_2S	0.157	0.094	0.289	0.676
CO	2.200	0.000	1.000	1.000
Random Coefficients				
Oil produced	-44.478	30.330	651.336	775.594
temperature	-46.699	1.676	26.015	32.881
humidity	-4.609	2.507	19.916	30.187
NO_x	0.000	0.000	0.000	0.000
CO_2	1.000	0.000	1.000	1.000
SD of coefficients				
Oil produced	-12.914	49.688		
temperature	-14.750	28.742		
humidity	0.242	16.455		
NO_x	0.000	0.372		
CO_2	0.000	1.422		
Max. log likelihood	-1626.172			

فقد اعتمد التفاوت الرئيس في أوضاع القياس الثلاثة المختلفة على العناقيد المعنية بأنظمة التلوث، الذي نشأ عما إذا كانت الأيام لها معدلات تلوث عالية أو متوسطة أو منخفضة. ايضاً اختلفت الأوضاع فيما يتعلق بمخالفة افتراض الاستقلالية مع النظام البيئي C2 الذي يُعتبر الأساس لمعدل التلوث لجميع النماذج.

يمثل النظام البيئي C1 قياسات بيئية وإنتاج نفط عالية مع فرضية استقلالية البدائل غير ذات صلة IIA صحيحة، وكما هو متوقع، فأن تقديرات معامل النموذج NO_x وCO_2 هي 0.084 (± SE) و 0.670 (± SE) في الجدول رقم (8) مرتفعة نسبياً وبما يتماشى مع التوقع النظري. وكانت الانحرافات المعيارية (SD) لنموذج الـ C1 لها فرق معنوي بين الصفر فالمعامل العشوائي كان مكمل (مفيداً) للنموذج في الجدول رقم (8).

أيضاً افترض النظام البيئي C2 IIA صحيحة، ويمثل قياسات بيئية وإنتاج نفط معتدلة. نسبة إلى C1، كان النموذج C2 أقل فاعلية تجريبية نظراً لأنه بالرغم من كون المعاملات العشوائية عالية مع NO_x و CO_2 لها 18.21- (SE ±) و 0.67 (SE ±)، كانت الانحرافات المعيارية لاثنين من المعاملات العشوائية ليس لها فرق معنوي بين الصفر مشيراً إلى أن معاملات NO_x و CO_2 لم تكن ضرورية لاحتواء النموذج في الجدول رقم (8).

واما في النظام البيئي C3 كل الأيام سجلت قياسات بيئية وإنتاج نفط منخفضة، مع انتهاك لفرضية IIA وان احتمال حدوث C3 يتوقف على ما إذا كان C1 يحدث. في هذا النظام البيئي، كانت معاملات C3 أقل بكثير بالمقارنة مع معاملات C1، وان NO_x و CO_2 تمتلك معاملات 0.00 و 1.00 ومن ثم فهي ليست تكاملية لمطابقة نموذج الانحدار اللوجستي الشرطي. وان حدوث C1 و C3 بَقيَ ثابتاً نسبياً في كلا النموذجين.

9. تقدير معلمات النموذج ثابت التأثيرات

Estimate of Fixed-effects model parameters

استناداً الى التحليل العنقودي المتعدد الذي ينص على ان نوعيات إنتاج النفط، والرطوبة، ودرجة الحرارة كانت الدوافع البيئية الرئيسية للتلوث. ونتيجة لذلك، يتم تقييم أثر الدوافع البيئية للتلوث على معدل التلوث الفعلي. ولذلك تم تطبيق الانحدار اللوجستي الشرطي في صياغة الدوال الاحتمالية التي تتيح تقييم تأثيرات إنتاج النفط العشوائية إذا أخذنا بعين الاعتبار انها تتوزع حسب التوزيع الطبيعي القياسي $N(0,\sigma^2)$ في حين كان من المفترض أن تكون درجة الحرارة قياس خط الأساس. وغالباً ما تتطلب مطابقة النموذج من خلال تقدير الإمكان الاعظم التقليل من تأثيرات البواقي الكبيرة، وكذلك تحسين المطابقة للبيانات المقدرة بالفعل. وكانت النتائج في الجدول (13).

الجدول رقم (9): يبين نتائج الدوال الاحتمالية لمعامل معدل التلوث استناداً إلى متوسط التلوث البيئي في كل ساعة (بوصفها ذات تأثير ثابت) للقياسات المتكررة مع توزيعات مُعامل طبيعية، للفترة ما بين ايلول /سبتمبر-2011 وكانون الأول /ديسمبر-2013 في شركة مصافي الوسط.

Variable	ß	SE
Fixed Coefficients		
H_2S	-0.083	0.079
NO_x	0.889	0.328
NH_3	-0.028	0.316
Random Coefficients		
Oil produced	-116.725	63.296
temperature	-90.384	0.000
SD of coefficients		
Oil produced	55.210	29.924
temperature	-16.096	0.000
Max. log likelihood	-1626.008	

نسبة الى مصفوفة متوسط درجة الحرارة في كل ساعة، ومتوسط كميات النفط المنتجة في كل ساعة كان لها تأثير أعلى على NO_x، و NH_3 و H_2S بهذا الترتيب (الجدول رقم (9)). بيد ان استجابة الملوثات لدرجة الحرارة كانت أكثر دقة، باعتبار أن كلاً من معامل درجة الحرارة وكذلك الانحراف المعياري للمعامل له خطأ معياري صفر (0).

مع ذلك فقد أشار احتمال متوسط التأثيرات الثابتة في كل ساعة إلى ميل عام للملوثات يكون مدفوعاً بمتوسط كميات النفط المنتجة في كل ساعة على مدى متوسط درجة الحرارة في كل ساعة. وأوضح نموذج الانحدار اللوجستي الشرطي فعلياً محل أثر درجة الحرارة وكميات إنتاج النفط وكذلك تحديد عدم التجانس في الاستجابة لإنتاج النفط ضمن العينة، مع NO_x وجود معامل استجابة سلبي (لأنه ظهر بإشارة موجبه) بينما يظهر NH_3 و H_2S معامل استجابة إيجابي لإنتاج النفط (لأنه ظهر بإشارة سالبة) والتي توحي بأن الزيادة في إنتاج النفط ودرجة الحرارة يَنفي مستوى تلوث NO_x.

10.اختبار الفرضية الإحصائية للنموذج العام

Statistical hypothesis testing for the general model

لغرض اختبار معنوية المعلمات المقدرة بنموذج الانحدار اللوجستي تم احتساب احصاءة والد، حيث يجب ان تكون معنوية المعلمات المناظرة لقبول او رفض فرضية العدم اقل من (0.05) التي تنص على ان:

H_0: الزيادة في معدل التلوث لا تعتمد على الزيادة في الظروف المناخية وإنتاج النفط.

H_1: الزيادة في معدل التلوث تعتمد على الزيادة في الظروف المناخية وإنتاج النفط.

الجدول رقم (10): يبين ملخص النموذج العام

	B	S.E.	Wald	Df	Sig.	Exp(B)
NO_x	-18.205	28382.56	0	1	0.999	0
CO_2	0.67	0.438	2.334	1	0.127	1.954
temperature	0.039	0.043	0.855	1	0.355	1.04
humidity	-0.008	0.024	0.122	1	0.726	0.992
Oil produced	-0.003	0.001	6.234	1	0.013	0.997
Constant	-0.327	2.292	0.02	1	0.887	0.721

واشار اختبار والد في الجدول رقم (10) أن هناك اختلاف معنوي في حدوث C1 في مجموعة البيانات، على اساس أن فرضية العدم قد رفضت في جميع المتغيرات فيما عدا NO_x، الذي كانت له قيمة اختبار والد صفر (0).

11. الاستنتاجات: Conclusions

1. تم تطبيق البيانات باستخدام الانحدار اللوجستي الشرطي ثابت التأثيرات لتحديد النموذج العام استناداً إلى متوسط التلوث البيئي في كل ساعة للقياسات المتكررة مع توزيعات مُعامل طبيعية. وأن هذا النموذج يشير إلى عدم التجانس في تلوث H_2S و NH_3 بسبب العوامل الأخرى التي تشمل درجة الحرارة، وكمية النفط المنتج. وعلى العكس من ذلك، يكون معدل تلوث NO_x هامشياً ويتوقف على درجة الحرارة وكمية النفط المنتج.
2. 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.
3. وتبين ايضاً أن درجة الحرارة لها تأثير أعلى على تلوث H_2S و NH_3 مقارنة بإنتاج النفط. وان النموذج ثابت التأثيرات يُظهر استجابة متميزة من NO_x لدرجة الحرارة وإنتاج النفط، في أن الزيادة في هذه المتغيرات التفسيرية تنتج انخفاضاً في NO_x التي تم دحضها في البداية بواسطة نموذج C2 مختلط التأثيرات. ولكن وفقاً لاختبار والد، النموذج مختلط التأثيرات أكثر دقة مقارنة بالنموذج ثابت التأثيرات الأمر الذي يؤدي من ثم إلى استنتاج مفاده أن التباينات في معدل التلوث تستند بشكل ملحوظ إلى مجموعة فرعية كبيرة من المتغيرات البيئية.
4. النموذج مختلط التأثيرات يبدد الافتراض العام بأن إنتاج النفط هو المسؤول بشكل كبير عن معدل التلوث في شركة مصافي الوسط. من ثم قد تقتضي إدارة معدل التلوث في تلك المنطقة تحديد أصول الملوثات الفردية.

5. تبين انه باستخدام معاملات عشوائية في حالة النماذج يمكن أن تولد استنتاجات خاصة بالملوثات التي توفر تحليلاً دقيقاً جداً حول معدل التلوث استناداً إلى العوامل البيئية، وخلافاً للنماذج ثابتة التأثيرات التي لا نستنتج منها سوى استدلال عن متوسط التلوث. وكثيرا ما تستخدم معدلات تلوث الهواء لقياس نوعية الهواء في المنطقة وتستخدم أيضاً لتحديد ومعاقبة المصادر المطلِقة للانبعاثات. كنتيجة لذلك، فان إجراء تقييم متحيز لمعدل التلوث قد يؤدي إلى عدم كفاية تنظيم الانبعاثات التي يمكن أن تزيد من إدامة تلوث الهواء.

12. **التوصيات**: Recommendations

يمكن تلخيص التوصيات الرئيسة لهذا 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 تنبؤية موثوق بها إحصائياً.

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

3. أن استخدام الانحدار اللوجستي الشرطي مختلط التأثيرات يوفر سياق أكثر تفصيلاً للعلاقات المتغيرة بدلاً من الانحدار اللوجستي الشرطي ثابت التأثيرات. ونتيجة لذلك، أنه باستخدام الانحدار اللوجستي الشرطي مختلط التأثيرات لم يثبت فقط أن يكون أكثر ملاءمة لتحديد الارتباطات المشتركة، وانما هو أيضاً فعال في إنشاء استدلالاً موثوقاً به.

4. إضافة كل من متغيرات استجابة وتفسيرية بغية التخفيف من حالة المؤشرات الخطية المتداخلة المتعددة Multicollinearity.

5. إضافة بيانات من أجل تسهيل تطوير نموذج تنبؤ مثالي، واقترح أن البيانات اليومية للفترة من 10 إلى 20 سنة تسفر عن نموذج أكثر دقة.

6. ينبغي أن تنظر شركة مصافي الوسط أيضاً في تحسين عمليات جمع البيانات، بمواصلة جمع البيانات عن الملوثات المحتملة وفي مختلف الظروف المناخية مثل النفايات السائلة والصلبة التي يتم الحصول عليها من خلال عملية التكرير وذلك لتسهيل تقييم المخاطر على مسار الهواء في الدراسات المستقبلية المماثلة.

7. 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. ولهذا الأمر آثار متنوعة يمكن استخدامها في الدراسات المستقبلية لاستكشاف العلاقة بين المتغيرات.

8. اجراء دراسة عن التلوث لفحص المستويات المستمرة لانبعاثات الكاربون الناجمة عن عمليات التكرير في مصافي النفط بالنسبة لكميات النفط المنتجة. ان تناول هذا الموضوع من شأنه تسهيل التعرف على إمكانية تلوث الكاربون عن طريق مصافي النفط. باعتبار أن المتغيرات البيئية أيضاً تديم التلوث وتحفز التفاعلات بين الملوثات، فمن المهم تحديد معدل الملوثات الفردية على مستوى المُنْشَأة لتحديد الأثر الفعلي من تكرير النفط على معدل التلوث. فضلا عن ذلك، من المهم انشاء ردود الفعل المحتملة بين الملوثات من الناحية النظرية، فضلاً عن المنتجات المشتقة منها من أجل مواصلة دراسة مستويات التلوث بالنسبة لمستوى المنتجات الثانوية في الهواء.

المصادر

1. العزاوي، احمد ذياب احمد (2005) "مقارنة بين بعض طرائق تقدير انموذج انحدار اللوجستك والطرائق الحصينة للتجارب الحياتية ذات الاستجابة الثنائية باستخدام أسلوب المحاكاة" رسالة ماجستير في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

2. شيرين علي حسين، (2009) "مقدرات الامكان الأعظم الموزونة الحصينة ومقارنتها مع طرائق أخرى لانموذج اللوجستك مع تطبيق عملي" رسالة ماجستير في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

3. القيسي، باسم شليبه مسلم عباس (2009) "التحليل البيزي لنماذج الانحدار الخاصة بالبيانات المزدوجة Panel Data" أطروحة دكتوراه في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

4. Agresti, A. (2002). Categorical Data Analysis. New York: Wiley-Interscience.
5. Amemiya, T. (1985). Advanced Econometrics. Harvard: Harvard University Press.
6. Bhat, C. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B-Methodological, 35, 677–693.
7. Cooper, A. & Millspaugh, J. (1999). The application of discrete choice models to wildlife resource selection studies. Ecology, 80, 566–575.
8. Freedman, D. (2009). Statistical Models: Theory and Practice. Cambridge: Cambridge University Press.
9. Greene, William H. (2003). Econometric Analysis. Prentice Hall.
10. Hilbe, Joseph M. (2009). Logistic Regression Models. Florida: Chapman & Hall/CRC Press.
11. Koepsell, T.D., and Weiss, N.S. (2003). Epidemiological Methods: Studying the Occurrence of Illness, Oxford University Press, New York, NY.
Long JS (1997) Regression Models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications
12. McDonald, T., Manly, B., Nielson, R. & Diller, L. (2006). Discrete-choice modelling in wildlife studies exemplified by Northern Spotted Owl nighttime habitat selection. Journal of Wildlife Management, 70, 375–383.
13. McFadden, D. & Train, K. (2000) Mixed MNL models for discrete response. Journal of Applied Econometrics, 15, 447–470.
14. Pampel FC (2000) Logistic regression: A primer. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132. Thousand Oaks, CA: Sage Publications.
15. 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.
16. Revelt, D. & Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics, 80, 647–657.
17. Skrondal, A. & Rabe- Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. Psychometrika, 68, 267–287.
18. Train, K.E. (2006). Mixed Logit Estimation by Maximum Simulated Likelihood. Matlab package. Available at:
http://elsa.berkeley.edu/Software/abstracts/train1006mxlmsl.html, Accessed on 26 June 2015.
19. Verbeke, G. & Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.
20. Pan W (2001). "Akaike's Information Criterion in Generalized Estimating Equations." Biometrics, 57(1), Hardin, J. & Hilbe, J. (2003). Generalized Estimating Equations. London: Chapman and Hall/CRC.
21. Hin L. & Wang, Y. (2009). Working-Correlation-Structure Identification in Generalized Estimating Equations. Statistics in Medicine, 28(4), 642–658.
22. Liang, K-Y & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73 (1): 13–22.

Compare to the conditional logistic regression models with fixed and mixed effects for longitudinal data

ABSTRACT

Mixed-effects conditional logistic regression is evidently more effective in the study of qualitative differences in longitudinal pollution data as well as their implications on heterogeneous subgroups. This study seeks that conditional logistic regression is a robust evaluation method for environmental studies, thru the analysis of environment pollution as a function of oil production and environmental factors. Consequently, it has been established theoretically that the primary objective of model selection in this research is to identify the candidate model that is optimal for the conditional design. The candidate model should achieve generalizability, goodness-of-fit, parsimony and establish equilibrium between bias and variability. In the practical sphere it is however more realistic to capture the most significant parameters of the research design through the best fitted candidate model for this research. Simulation studies demonstrate that the mixed-effects conditional logistic regression is more accurate for pollution studies, with fixed-effects conditional logistic regression models potentially generating flawed conclusions. This is because mixed-effects conditional logistic regression provides detailed insights on clusters that were largely overlooked by fixed-effects conditional logistic regression.

Key Words: **Maximum likelihood method, conditional logistic regression, longitudinal data, mixed effects models, Quasi-likelihood under independence Criterion (QIC), Empirical Akaika Information Criteria (EAIC), environmental pollution, Cluster analysis.**