مقارنة طرائق تقدير معالم انموذج SCHEFF'E الخليط باستخدام Generalized Inverse وانحدار الخطوات المتسلسل لمعالجة مشكلة التعدد الخطي

أ.د. دجلة ابراهيم مهدي /كلية الادارة والاقتصاد/ جامعة بغداد
الباحث / حلا سلمان فرحان

تاريخ التقديم:2018/7/3
تاريخ القبول:2018/8/20

المستخلص:

تجارب الخليط عبارة عن متغيرات استجابة تعتمد على نسب المكونات لهذا الخليط . في بحثنا هذا سنعمل مقارنة بين نموذج **Schaffer** ونموذج **Kronecker** لتجارب الخليط خاصة عندما تكون المنطقة التجريبية مقيدة . ونظرا لما تعانيه تجارب الخليط من مشكلة الارتباطات العالية ومشكلة التعدد الخطي بين المتغيرات التوضيحية لما له اثر على حساب مصفوفة المعلومات **(Fisher Information Matrix)** $(\acute{X}X)$ لنموذج الانحدار .

لتقدير معالم أنموذج الخليط اعتمدنا في بحثنا على استخدام صيغة المعكوس المعمم **(Generalized Inverse)** وتوظيفها في عملية التقدير، فضلا عن اسلوب انحدار الخطوات المتسلسلة **(The Stepwise Regression procedure)** .

باستخدام مقياس عامل تضخم التباين **(VIF) (Variance Inflation Factor)** لبيان مدى التباينات العالية في كلا النموذجين، فضلا عن استخدام تحويل المكونات الزائفة للحدود الدنيا **(L-Pseudo Component)** ، وتمثيله باستخدام المحاكاة بلغة **R** بمعيار المقارنة متوسط مطلق الخطأ النسبي **Mean Absolute Percentage Error (MAPE)** .

المصطلحات الرئيسة للبحث/ نماذج الخليط ، عامل تضخم التباين ،المكونات الزائفة، المعكوس المعمم ، انحدار الخطوات المتسلسلة ، متوسط مطلق الخطأ النسبي .

*بحث مستل من اطروحة دكتوراه .

1.المقدمة:

يعد نموذج الخليط من النماذج المهمة التي تعنى بدراسة المكونات الكيميائية من الناحية الاحصائية وكيفية التعامل مع هذه المكونات وتاثير التفاعلات عليها. لدراسة تجارب الخليط يقوم الباحث باختيار عدد من الخلطات التي تختلف في نسب مكونين او اكثر في كل من المخاليط تحت الدراسة . يتم قياس نسب المكونات عن طريق الحجم اوالوزن او الكسر المولي (تعبير كيميائي (الكسر الجزئي) يصف نسبة تواجد مادة معينة في مخلوط من المواد) ، على فرض ان مقياس الاستجابة في الخليط يعتمد على نسب المكونات في الخليط وليس على كمية الخليط. يكون التعبير عن مكونات الخليط من خلال النسب الغير سالبة والتي مجموعها مساوي للواحد الصحيح .[III] PP.4−5

يتضمن هذا البحث مفهوم عام عن الخليط وعرض للنماذج الخطية المستخدمة في تقدير سطح الاستجابة ، النموذج الاول مقدم من قبل [XI] (1958) Scheff' e (S-Model) من الدرجة الثانية والنموذج الثاني مقدم من قبل [V](1998) (Draper and pukelsheim) (Kronecker Model) (K-Model) من الدرجة الثانية، خاصة عندما تكون المنطقة التجريبية مقيدة .

2.هدف البحث:

نظرا لما تعانيه تجارب الخليط من مشكلة الارتباطات العالية ووجود التعدد الخطي بين اعمدة مصفوفة المتغيرات التوضيحية نتيجة لوجود قيد الوحدة لمكونات الخليط $\sum_{i=1}^{p} x_i = 1$ والتفاعلات بين المكونات، هدفنا في هذا البحث معالجة هذه المشاكل في نموذجي الخليط (S-Model) و (K-Model) التي تعطي كل منها نوع مختلف من المعلومات لمكونات الخليط بهدف الوصول الى افضل استجابة للخليط .

3. الصيغة العامة لنماذج الخليط: Generalized Formula for Mixture Model

في اغلب الاحيان بعض الحالات من نسب الخليط ، تكون مقيدة اي يكون لها حد اعلى وحد ادنى وتكون هذه القيود بالشكل الاتي :

$$0 \le L_i \le X_i \le U_i \le 1 \qquad ,1 \le i \le p$$

اذ :

L_i : تمثل الحد الاعلى للمكون i .

U_i : تمثل الحد الادنى للمكون i .

هنا سوف ناخذ بنظر الاعتبار النموذج الذي يمثل متغيرات الاستجابة التي تكون بدورها متمثلة بالعلاقة النسبية بين المكونات وتفاعلاتها مع بعضها بحيث نقوم بتحديد الخليط من المكونات التي تعطي نتائج مرغوب فيها من قيم الاستجابة.[III] PP.7

عندما يكون النموذج الملائم للبيانات نموذج من ضمن النقاط $\{p, m\}$، علما ان m تمثل درجة النموذج و p عدد المكونات في الخليط.

لتكن لدينا مجموعة من المكونات التي تمثل سطح الاستجابة ممثلة بنسب من المتغيرات التوضيحية $(x_1, x_2, x_3, \dots, x_p)$ وان هذه النسب تتمثل بكميات موجبة ومجموعها مساوي الى الواحد الصحيح وعادة نرمز لنسبة المكون (i) بالرمز (x_i) ،

فاذا فرضنا ان (i) تمثل عدد من المكونات في نظام خليط قيد الدراسة ولتكن x_i تمثل النسب الجزئية للمركبة (i) في الخليط فان :

$$0 \leq x_i \leq 1 , \qquad i = 1,2,3, \dots \dots . . , p \qquad (3-1)$$

And

$$\sum_{i=1}^{p} x_i = x_1 + x_2 + \cdots + x_p = 1 \qquad (3-2)$$

وليكن $(x_1 , x_2 , x_3 , \dots . . , x_p)$ متجه نسب يحوي p من المكونات و y_i الاستجابة المقابلة للخليط ، والفضاء للمتجه هو المجال : [IV] PP.2

$$X = \left\{ x = (x_1, x_2, \dots . . , x_3) : 0 \leq x_i \leq 1, i = 1,2, \dots . , p; \sum_{i=1}^{p} x_i = 1 \right\} \qquad (3-3)$$

قدم Scheff’ e (1958) نماذج الخليط بصيغ عامة بدرجات مختلفة ليحدد متوسط الاستجابة لدالة الخليط ومنها نموذج الخليط من الدرجة الثانية (S-Model) : [XI] PP.346

$$y_k = \sum_{i=1}^{p} \beta_i x_i + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \varepsilon_k \qquad (3-4)$$

$$\forall \ i, j = 1,2, . . , p \qquad k = 1,2, . . , n$$

ويمكن التعبير عنه بالصيغة العامة لنماذج الانحدار :

$$Y = X\beta + \varepsilon$$

Y: متجه متغير الاستجابة من الدرجة $n \times 1$ ، n حجم العينة تحت الدراسة.

X: مصفوفة المتغيرات التوضيحية وتفاعلاتها من الدرجة $n \times (p+1)$ ، p عدد المعالم المقترنة بمصفوفة المتغيرات التوضيحية .

β : متجه معالم النموذج من الدرجة $(p+1) \times 1$.

ε : متجه الخطأ التجريبي من الدرجة $n \times 1$.

على فرض ان الخطأ التجريبي (ε) يتخذ توزيع طبيعي مستقل :

$$\varepsilon \sim IND(0, \sigma^2 \mathrm{I})$$

لاحظ انه خلافا لما هو معروف في نماذج الانحدار المعتادة سوف يتم اسقاط الحد الثابت β_0 من جميع نماذج الخليط لانها تمثل الميل الحدي (Intercept) الذي ليس له دور مهم وواضح في تفسير عملية دراسة تجارب الخليط ، التي تعنى اكثر بدراسة تاثير المتغيرات التوضيحية والتفاعلات الثنائية والثلاثية ... الخ ، على متغيرات الاستجابة. [IV] PP.11-12 [VIII] PP.888 ,

هو نموذج Kronecker (1998) (Draper and pukelsheim)اما النموذج الثاني مقدم من قبل وهو اقل عرضة لخاصية النموذج الغير مستقر نتيجة الارتباط (K-Model) من الدرجة الثانية Model العالي بين المعلمات والخطأ التجريبي العالي عند العمل على نماذج من الدرجة الثانية ، فان النموذج البديل (K-Model): يحتوي حدود من الدرجة الثانية فقط والنموذج العام له: [V]$^{PP.304-305}$

$$y_k = \sum_{i=1}^{p} \beta_i x_i^2 + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \varepsilon_k \quad , \qquad (3-5)$$

$$i,j = 1,2,..,p \qquad k = 1,2,..,n$$

4. مشكلة التعدد الخطي: Multicollinearity Problem

نتيجة لوجود التفاعلات في نموذج (S-Model) و(K-Model) فان اعمدة المصفوفة X تكون معتمدة خطيا وتعاني من تعدد خطي ، لذلك ومن متطلبات حساب مقدرات β ان تكون مصفوفة المعلومات $(\acute{X}X)$ غير مفردة (non-Singular) وهذا غير متوفر في مصفوفة المعلومات في نماذج الخليط التي تكون في معظمها مفردة (Singular) ومن النتائج المترتبة على هذه الخاصية فان مقدرات المربعات الصغرى (OLS) للمعالم تمتلك خطأ تجريبي كبير ، وارتباط عالي ، بالاضافة الى ذلك ان هذه المقدرات معتمدة بشكل كبير على دقة موقع النقطة في التصميم. [VII]$^{PP.348-350}$, [I]$^{PP.338}$

فيما يلي سوف نوضح بعض القياسات التشخيصية التي يمكن ان تساعد على كشف وتحديد التعدد الخطي.

4-1 عامل تضخم التباين (Variance Inflation Factor): (VIF)

ان العلاقة الخطية المتداخلة تتمثل في الإرتباط بين متغيرين أو أكثر من المتغيرات التوضيحية. عندما يكون هناك ارتباط عالي بين أثنين من المتغيرات التوضيحية ويتم بناء نموذج الإنحدار فتكون نتيجة معاملات الإنحدار غير دقيقة والخطأ المعياري كبير في معالم النموذج وبالتالي لا يمثل النموذج القيم الصحيحة التي نهدف إليها. نستطيع تقدير العلاقة الخطية المتداخلة وتحديد فيما اذاوجد تعدد خطي في مصفوفة المتغيرات التوضيحية في النموذج بأستخدام معامل تضخم التباين بما يتعلق بتقدير معالم نموذج الانحدار وفق الصيغة التالية : $[X]^{PP.325}$

$$VIF(\hat{\beta}_j) = (1 - R_j^2)^{-1} \qquad , \forall\, j = 1,2,\dots,p \quad (4-1)$$

حيث ان :

$$R_j^2 = \frac{x_j' X_j (X_j' X_j)^{-1} X_j' x_j}{x_j' x_j} \qquad \forall\, j = 1,2,\dots,p \qquad (4-2)$$

R_j^2 : معامل الارتباط المتعدد للعمود j نتيجة ارتداد العمود x_j من مصفوفة X_j.

p : عدد المعالم في النموذج .

X_j: مصفوفة المتغيرات التوضيحية X محذوفا منها العمود j من درجة $n \times (p-1)$.

x_j : المتجه j من درجة $n \times 1$ لمصفوفة المتغيرات التوضيحية X .

من اهم طرق التحويل (Transformation) المستخدمة في تجارب الخليط :

5. طريقة المكونات الزائفة للحدود الدنيا :(L-Pseudo component)

الجزء المهم في تجارب الخليط ان هناك قيود حقيقية على نسب المكونات عندما $0 \leq x_i \leq 1 \ , \forall i = 1,2,..,p$ و $\sum_{i=1}^{p} x_i = 1$ ، في كثير من الاحيان ليس هناك حرية كاملة في دراسة المجال بشكل كامل بسبب بعض القيود الاضافية التي يتم وضعها على بعض نسب المكونات التي يجب ان تكون لها حدود دنيا لا يكون الخليط جيدا اذا كانت نسب هذه المكونات دون الحد الادنى منها. [III] PP.133

لذلك ان هذه القيود التي وضعت على بعض المكونات من شأنه ان يحدد الخلطات المرغوبة في منطقة جزئية من المجال ، وكذلك الحال عند وضع حدود عليا على بعض نسب المكونات ، او الاثنين معا .
لبيان كيف يتم حساب المكونات الزائفة :

$$0 \leq L_i \leq x_i \quad \forall\, i = 1,2,\dots,p$$

حيث

L_i: الحدود الدنيا لنظام يحتوي على p من المكونات ، وبالتالي نحصل على مكونات زائفة لكل متغير توضيحي كما يلي: [III] PP.134

$$x_i^* = \frac{x_i - L_i}{1 - L} \qquad (5-1)$$

عندما $L = \sum_{I=1}^{p} L_i < 1$

x_i^* : المكون الزائف للمتغير التوضيحي x_i .

ان اتجاه مجال المكونات الزائفة هو نفسه اتجاه المكونات الاصلية [X] PP.327−328 ، الخطوة التالية هي لحساب قيم الاستجابة لنموذجي الخليط تحت الدراسة: النموذج الاول (S-Model) بدلالة المكونات الزائفة :

$$\begin{aligned} y_i^* = {} & \tau_1 x_{i1}^* + \tau_2 x_{i2}^* + \tau_3 x_{i3}^* + \tau_4 x_{i4}^* + \tau_5 x_{i5}^* + \tau_{12} x_{i1}^* x_{i2}^* + \tau_{13} x_{i1}^* x_{i3}^* \\ & + \tau_{14} x_{i1}^* x_{i4}^* + \tau_{15} x_{i1}^* x_{i5}^* + \tau_{23} x_{i2}^* x_{i3}^* + \tau_{24} x_{i2}^* x_{i4}^* \\ & + \tau_{25} x_{i2}^* x_{i5}^* + \tau_{34} x_{i3}^* x_{i4}^* + \tau_{35} x_{i3}^* x_{i5}^* + \tau_{45} x_{i4}^* x_{i5}^* \\ & + \varepsilon_i \qquad (5-2) \end{aligned}$$

اما النموذج الثاني (K- Model) باستخدام المكونات الزائفة يأخذ الشكل الاتي:

$$\begin{aligned} y_i^* = {} & \tau_1 {x_{i1}^*}^2 + \tau_2 {x_{i2}^*}^2 + \tau_3 {x_{i3}^*}^2 + \tau_4 {x_{i4}^*}^2 + \tau_5 {x_{i5}^*}^2 + \tau_{12} x_{i1}^* x_{i2}^* \\ & + \tau_{13} x_{i1}^* x_{i3}^* + \tau_{14} x_{i1}^* x_{i4}^* + \tau_{15} x_{i1}^* x_{i5}^* + \tau_{23} x_{i2}^* x_{i3}^* \\ & + \tau_{24} x_{i2}^* x_{i4}^* + \tau_{25} x_{i2}^* x_{i5}^* + \tau_{34} x_{i3}^* x_{i4}^* + \tau_{35} x_{i3}^* x_{i5}^* \\ & + \tau_{45} x_{i4}^* x_{i5}^* + \varepsilon_i \qquad (5-3) \end{aligned}$$

اذ ان :

y_i^* : متغير الاستجابة لنموذج الانحدار بدلالة المكونات الزائفة.

τ_i ، τ_{ij} : معالم النموذج بدلالة المكونات الزائفة .

x_{ij}^* : المكون الزائف لكل من المتغيرات التوضيحية لنموذج الخليط .

$$, i \neq j \qquad \forall\ i = 1,2,\dots..p \qquad , j = 1,2,..,n.$$

6. **اسلوب انحدار الخطوات المتسلسلة**: The Stepwise Regression procedure (S.T):

من الاساليب الاكثر شيوعا من خلالها نستطيع الاستغناء عن بعض المتغيرات التوضيحية في نموذج الانحدار التي يكون تاثيرها غير معنوي في النموذج وعندها يتم حذفه مع المعلمة المرتبطة به في النموذج ،حيث ان معيار الاختيار للمتغيرات التوضيحية سوف يكون على اساس قوة الارتباط مع متغير الاستجابة من خلاله يتم ترشيح اي المتغيرات التي سوف تكون ضمن الاختبار عن طريق استخدام مصفوفة الارتباطات البسيطة (Correlation Matrix) بين المتغيرات التوضيحية والتفاعلات ومتغير الاستجابة: $[I]^{PP.251-252}$

$$\rho = \begin{pmatrix} 1 & \varrho_{12} & \varrho_{13} & \varrho_{14} & \varrho_{15} \cdots\cdots & \varrho_{1y} \\ & 1 & \varrho_{23} \cdots\cdots & \varrho_{25} \cdots\cdots & & \varrho_{2y} \\ & & & \vdots & & \vdots \\ & & & 1 \cdots\cdots & & \varrho_{y5} \\ & & & \vdots & & \vdots \\ & & & & & 1 \end{pmatrix}$$

حيث نبدأ بادخال المتغيرات التي تمتلك اعلى ارتباط مع متغير الاستجابة ، اما مقياس الذي سوف نعتمد عليه بالتثبيت او حذف المتغيرات التوضيحية سوف يكون باستخدام اختبار (F) :

$$F_1 = \frac{SSR(x_1)}{MSE(x_1)} = \frac{\beta_1^2 \sum_{i=1}^{n} x_{i1}^2}{MSE(x_1)} \qquad (6-1)$$

وتقارن قيمة F_1 مع قيمة F_t الجدولية بمستوى $(1, n-p, 1-\alpha)$ ، فاذا كانت F_1 غيرمعنوية نحذف x_1 من النموذج ونتوقف عن الاختيار، واذا كانت معنوية نثبت x_1 في النموذج، ، وبعد ثبوت المتغير التوضيحي x_1 نقوم بحساب معامل الارتباط الجزئي (Partial Correlation Coefficient) لجميع المتغيرات التوضيحية الباقية بما فيها التفاعلات مع متغير الاستجابة y بثبوت المتغير التوضيحي x_1 اي نستخرج: $[IX]^{PP.122}$

$$\varrho_{2y.1} = (\varrho_{2y} - \varrho_{12}\varrho_{1y}) \Big/ \sqrt{(1-\varrho_{12}^2)(1-\varrho_{1y}^2)} \qquad (6-2)$$

وبعدها نختار المتغير التوضيحي او التفاعل الذي يعطي اعلى ارتباط جزئي ، وليكن المتغير التوضيحي x_2 ، نستخرج فيما بعد F_2 الجزئية لكل من المتغيرات x_1 و x_2 كالاتي:

$$F_1 = \frac{SSR(x_1/x_2)}{MSE(x_1,x_2)} = \frac{\hat{\beta}_1^2}{S^2(\hat{\beta}_1)}$$

$$F_2 = \frac{SSR(x_2/x_1)}{MSE(x_1,x_2)} = \frac{\hat{\beta}_2^2}{S^2(\hat{\beta}_2)}$$

وتقارن قيمتي F_1, F_2 الجزئية مع F_t الجدولية بمستوى $(1, n-p, 1-\alpha)$ حيث ان p عدد المتغيرات الداخلة في النموذج ، ونختار اقل F جزئية، فاذا كانت F_2 هي الصغرى تقارن مع F_t فاذا كانت :

a) معنوية: نثبت x_1 و x_2 في النموذج ونكرر العملية ،بادخال متغير ثالث حسب اقل معامل ارتباط جزئي بثبوت كل من المتغيرات الداخلة في الاختبار بالخطوة السابقة.

b) غير معنوية: نحذف x_2 ثم نتوقف نهائيا عن الاختيار ويبقى النموذج بدلالة المتغير التوضيحي x_1 فقط.

اما اذا كانت F_1 هي الصغرى تقارن مع F_t فاذا كانت:

a) معنوية : نثبت x_1 و x_2 في النموذج ونكرر العملية ،بادخال متغير ثالث حسب اقل معامل ارتباط جزئي بثبوت كل من المتغيرات الداخلة في الاختبار بالخطوة السابقة.

b) غير معنوية: نحذف x_1 من النموذج ، ويبقى المتغير التوضيحي x_2 في النموذج وبعدها نكرر العملية بادخال متغير جديد للنموذج بوجود المتغير التوضيحي x_2 في النموذج وهكذا نبدأ بحساب معامل الارتباط الجزئي بثبوت المتغير التوضيحي x_2 ، بنفس الخطوات السابقة.[I]$^{PP.254-255}$

ونستمر بهذا الاسلوب الى ان يحوي النموذج النهائي على جميع المتغيرات التوضيحية وتفاعلاتها التي تكون لها تاثير معنوي على النموذج مجتمعة، وبهذه الطريقة نحصل على افضل نموذج يمكن ان يعبر عن متغير الاستجابة ،مع ملاحظة امكانية حذف متغير توضيحي او تفاعل تم ادخاله الى النموذج في خطوة سابقة وذلك نتيجة للتاثير المتداخل بين المتغيرات وتفاعلاتها.

7. طريقة مقترحة باستخدام Generalized Invers for Matrices:

ليكن النموذج الخطي العام $Y = X\beta + \varepsilon$ عندما Y هي متجه عشوائي يمثل الاستجابة $n \times 1$ و Xمصفوفة من الدرجة $n \times p$ والرتبة k لمتغيرات التوضيحية الغير عشوائية (عندما $n > p > k$) ، β متجه المعالم الغير معلومة من الدرجة $p \times 1$ ، ε متجه عشوائي من الدرجة $n \times 1$ يتوزع $\varepsilon \sim N(0, \sigma_\varepsilon^2\, I)$ هذا النموذج يسمى نموذج تصميم (**Design Model**) .

اذا كانت $p = k$ فان هذا النموذج يسمى نموذج خطي عام **(Generalized Linear model)** وتنطبق عليه جميع نظريات النموذج الخطي العام ، ونظرا لما تعانيه نماذج الخليط من مشكلة التعدد الخطي والتباين الكبير بين المتغيرات التوضيحية ووجود التفاعلات والعلاقات الخطية بين تلك المتغيرات في النموذج، مما له اثر كبير على رتبة المصفوفة X وهذا بدوره سوف ينعكس على عملية التقدير وصعوبة ايجاد معكوس المصفوفة $(X\acute{X})$، هنا في هذا البحث سوف نستخدم المعكوس المعمم لمصفوفة غير مفردة **(non-singular)** كحل وحيد لمجموعة محددة من المعادلات ، ان المعكوس العام موجود لاي مصفوفة مستطيلة او مربعة غير مفردة من المصفوفات على الاطلاق . [VI]$^{PP.480}$

<u>تعريف</u> : لتكن $(X\acute{X})$ مصفوفة مفردة من الدرجة $p \times p$ لها رتبة k حيث ان $(k < p)$ ، $(X\acute{X})^g$ المعكوس المعمم **"g-invers"** لمصفوفة المعلومات للنموذج من شروط المعكوس المعمم : [II]$^{PP.32}$

$$\left.\begin{array}{l} \bullet\ (\acute{X}X)(\acute{X}X)^g \ \ is\ symmetric \\ \bullet\ (\acute{X}X)^g(\acute{X}X) \ \ is\ symmetric \\ \bullet\ (\acute{X}X)(\acute{X}X)^g(\acute{X}X) = (\acute{X}X) \\ \bullet\ (\acute{X}X)^g(\acute{X}X)(\acute{X}X)^g = (\acute{X}X)^g \end{array}\right\} \quad (7\text{-}1)$$

نظرية : لكل مصفوفة $(\acute{X}X)$ تمتلك معكوس عام وحيد $(\acute{X}X)^g$ يحقق الشروط الاربعة في الصيغ (7-1) $[\text{VI}]^{PP.25-26}$.

عند تطبيق هذه الطريقة على الصيغة الطبيعية (Normal Equation) :

$$(\acute{X}X)\beta = \acute{X}Y \qquad (7-2)$$

يجب اولا ان نثبت ان هذه الصيغة تمتلك خاصية الاتساق (Consistent) اي ان $rank(\acute{X}X) = rank(\acute{X}X:\beta)$.

نتيجة لهذا الصيغة (7-2) تمتلك خاصية الاتساق ، فان الحلول العامة الى $\hat{\beta}$ مساوية الى :

$$\hat{\beta}^g = (\acute{X}X)^g\acute{X}Y + [I - (\acute{X}X)^g(\acute{X}X)]h \qquad (7-3)$$

h: هي متجه اختياري من القيم حيث $h \in E_p$.

ولاثبات ان $\hat{\beta}^g$ هي تقدير غير متحيز :

$$E(\hat{\beta}^g) = (\acute{X}X)^g\acute{X}EY + [I - (\acute{X}X)^g(\acute{X}X)]h$$
$$= (\acute{X}X)^g\acute{X}X\beta + (\acute{X}X)^g\acute{X}E\varepsilon + [I - (\acute{X}X)^g(\acute{X}X)]h$$

بضرب طرفي المعادلة من اليسار بـ $(\acute{X}X)$:

$$= (\acute{X}X)(\acute{X}X)^g\acute{X}X\beta + [(\acute{X}X) - (\acute{X}X)(\acute{X}X)^g(\acute{X}X)]h$$

وحسب الشرط الثالث من شروط المعكوس المعمم فان:

$$= \acute{X}X\beta + [(\acute{X}X) - (\acute{X}X)]h$$

اذا:

$$(\acute{X}X)E(\hat{\beta}^g) = (\acute{X}X)\beta \quad \rightarrow E(\hat{\beta}^g) = \beta \qquad (7-4)$$

اما بالنسبة الى تباين $\hat{\beta}^g$ وبالاعتماد على نظرية المبدلة لمصفوفة المعكوس المعمم فان $((\acute{X}X)^g\acute{)} = ((\acute{X}X\acute{)})^g$:

$$Var - Cov(\hat{\beta}^g) = (\acute{X}X)^g\acute{X}Var - Cov(Y)X(\acute{X}X)^g$$
$$= \sigma_\varepsilon^2(\acute{X}X)^g\acute{X}X(\acute{X}X)^g$$

وحسب الشرط الرابع من شروط المعكوس المعمم فان:

$$Var - Cov(\hat{\beta}^g) = \sigma_\varepsilon^2(\acute{X}X)^g \qquad (7-5)$$

8 . متوسط مطلق النسبة المئوية للخطأ Mean Absolute Percentage Error (MAPE) .

هو من المقاييس الاكثر شيوعا للاخطاء التقديرية ،يعمل (MAPE) بشكل افضل عندما لايكون هناك قيم متطرفة تظهر في البيانات وخاصة عندما تكون مساوية للصفر ، فان بوجود الاصفار او الاصفار القريبة يمكن (MAPE) ان يعطي صورة مشوهة عن الخطأ ، ونظرا لان في هذا البحث نريد معرفة تاثير جميع مكونات الخليط بما فيها التفاعلات على مقدار الاستجابة للخليط لذلك تكون جميع قيم معالم النموذج اكبر من الصفر (موجبة او سالبة). $[\text{XII}]^{PP.1-2}$

اما الصيغة الرياضية لحسابه:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| 100\%$$

N : عدد المشاهدات في التجربة.

9. جانب المحاكاة:

يعتبر الجانب التجريبي من افضل جوانب البحث العلمي لانه يعتمد بالاساس على التجربة العلمية مما يوفر فرصة عملية لمعرفة الحقائق وتطبيق النظريات ومعرفة الطرائق السليمة للتعامل مع الظواهر وتفسيرها عن طريق هذه التجارب .

نظرا لخصوصية هذه التجارب وماتحتاجه من وقت طويل لاجراءها ،لذلك تم توليد هذه البيانات باستخدام اسلوب المحاكاة (Monte Carlo Simulation) لما يوفره هذا الاسلوب من امكانية تكوين بيانات تجارب الخليط تعمل على محاكاة الواقع العملي ومايعطيه هذا الاسلوب من مساحة اكبر للتجريب وبالتالي القدرة على التقدير والاختبار ، وذلك من خلال تكرار تلك العملية لمرات عديدة تتحدد بقيم المدخلات الخاصة بالتجربة . وتاتي اهمية المحاكاة هو ان كل تجربة او كل تكرار هو عملية عشوائية مستقلة عن التجربة التي تليها ، وبالتالي يمكننا استخدام اسلوب المحاكاة للحصول على بيانات تطبيق شرط نماذج الخليط بتوليد خمسة متغيرات عشوائية ، $\sum_{i=1}^{5} x_i = 1$ ، $0 \le x_i \le 1, \ \ \forall\, i = 1,2,\dots 5$ مع وجود التفاعلات بين هذه المكونات وبالتالي الحصول على بيانات مرتبطة خطيا وتعاني من مشكلة التعدد الخطي مما يصعب على الباحث عملية التقدير ويتطلب تطبيق طرق خاصة تتعامل مع هذه المشاكل بشكل مباشر او غير مباشر، وبعد اخذ المتغيرات التوضيحية واجراء عليها تحويل (L-Pseudo component) وبهذا نحصل على مصفوفتان لكل نموذج للمتغيرات التوضيحية (مصفوفة النموذج الاصلي ،مصفوفة المكونات الزائفة للحدود الدنيا).

تم توليد مشاهدات المتغيرات التوضيحية وفق القيد $\sum_{i=1}^{5} x_i = 1$ ، $0 \le x_i \le 1, \ \ \forall\, i = 1,2,\dots 5$ وبتوزيع (Uniform Distribution) ولعدة حالات باحجام عينات مختلفة ($n = 30\ ,50\ ,75\ ,100\ ,150$) فتصبح لدينا مصفوفتان لنموذجي (S- Model)، (k-Model) ، وبعد ذلك نعمل على المتغيرات الناتجة بتحديد قيم افتراضية للمعالم:

جدول (1) القيم الافتراضية لمعالم نموذجي (S- Model) و (k-Model) لتجارب المحاكاة

المعلمات	β_1	β_2	β_3	β_4	β_5	β_{12}	β_{13}	β_{14}	β_{15}	β_{23}	β_{24}	β_{25}	β_{34}	β_{35}	β_{45}
القيم الافتراضية	55	64	90	70	90	60	60	50	44	80	-59	54	80	60	50

تم اعتماد هذه القيم الافتراضية للمعلمات باحجام عينات ، وانحراف معياري مختلف لمعرفة سلوك كل من النموذجين وطرق التقدير في حال تغير هذه القيم وتاثيرها على مقدار الخطأ في تقدير معالم النموذجين .

اضافة الى توليد الخطأ التجريبي يتوزع طبيعيا بمتوسط صفر وانحراف معياري ($\sigma = 2\ , 2.50$) بمستوى معنوية $\alpha = 0.05$ ، بعد ذلك يتم توليد البيانات (Y) عن طريق نموذجي (S- Model) (k- model), ، الخطوة اللاحقة هي تقدير معلمات كل من انموذج (S- Model) وانموذج (K- Model) وفق الصيغ (3-4) و (3-5) وذلك باستخدام طرائق التقدير (Stepwise Regression و Generalized Invers) ،علما ان تم تكرار هذه العملية (1000) مرة.

10. تحليل البيانات:

نلاحظ من الجدول (2) من خلال مقياس (VIF) ان بعد تطبيق تحويل المكونات الزائفة للحدود الدنيا على مصفوفة كل من نموذجي (S- Model) و (K-model) انخفاض قيمة VIF ولكن وفق معيار (VIF>10) فانها تبقى تعاني من تعدد خطي لذلك نلجأ الى استخدام الطرائق التي تعالج هذا الارتباط الكبير بين تلك المتغيرات لكي تعطي نتائج افضل :

جدول (2) قيمة (VIF) لنموذجي الخليط للمصفوفة الاصلية ومصفوفة التحويل بطريقة L-Pseudo component

parameters	S-Model	psd(S-Model)	K-Model	psd(K-Model)
β_1	1239.73	313.20156	63.42871	35.193551
β_2	635.4051	181.5565	33.30961	20.394533
β_3	2098.597	395.80171	136.2711	46.23862
β_4	1202.105	230.56821	73.44289	30.974267
β_5	1137.691	188.60041	215.3876	21.457923
β_{12}	71.18995	65.22564	14.51438	12.919842
β_{13}	94.70623	68.40277	21.41244	13.820179
β_{14}	75.71038	72.11581	16.89905	16.454047
β_{15}	404.893	26.63105	75.84496	12.129768
β_{23}	183.0283	129.8205	34.40801	28.102936
β_{24}	22.9516	20.85802	8.186428	5.064963
β_{25}	252.0141	10.79339	85.71208	9.048993
β_{34}	70.54294	51.38564	57.73043	14.430421
β_{35}	1864.708	126.02282	447.7984	36.226161
β_{45}	1271.163	173.35892	287.2271	44.760852

الجداول (3),(4) يوضح قيم MAPE لطرائق التقدير المستخدمة في هذا البحث وحسب حجم العينة وقيمة الانحراف المعياري :

جدول (3) يبين قيم MAPE لنموذجي الخليط لكل طرق التقدير باستخدام للمصفوفة الاصلية ومصفوفة التحويل L-Pseudo component عندما $(\sigma = 2)$

	n=30		n=50		n=75	
Method	S- model	K- model	S- model	K- model	S- model	K- model
g-inv	0.017866876	0.029074887	0.016617045	0.026023098	0.01492257	0.02533951
ST	0.023024848	0.028706853	0.015400641	0.025585981	0.014477631	0.024919871
psd-g-inv	0.012424083	0.025256419	0.008514957	0.024689794	0.006863756	0.024495885
psd-ST	0.012767	0.026167145	0.009891758	0.025573252	0.008862809	0.025497405

جدول (4) يبين قيم MAPE لنموذجي الخليط لكل طرق التقدير باستخدام للمصفوفة الاصلية ومصفوفة التحويل L-Pseudo component عندما $(\sigma = 2.50)$

	n=30		n=50		n=75	
Method	S- model	K- model	S- model	K- model	S-model	K- model
g-inv	0.031807953	0.03696881	0.026712208	0.028649137	0.021533385	0.026556647
ST	0.036084596	0.035362173	0.02227793	0.029119398	0.017764899	0.026436739
psd-g-inv	0.021793482	0.026916516	0.014551539	0.025273988	0.010929177	0.024949196
psd-ST	0.019177249	0.02829267	0.013593832	0.026418397	0.011337768	0.026079128

نلاحظ من جدول (3) عندما $(\sigma = 2)$:

1. ان كلا من طريقة (g-inv) و (ST) عند تطبيقها على (S-Model) تعطي قيم اقل لمعيار MAPE بالمقارنة مع (K-Model) لجميع احجام العينات ، في حال تطبيق كلا الطريقتين على مصفوفة البيانات الاصلية او مصفوفة المكونات الزائفة للحدود الدنيا .
2. قيم MAPE تقل بزيادة حجم العينة (لكلا النموذجين) في كلا الطريقتين.
3. اما في حالة تطبيق كلا الطريقتين على مصفوفة تحويل المكونات الزائفة للحدود الدنيا (psd-g-inv) و(psd-ST) فانهما تمتلكان اقل MAPE لنموذج (S-Model) عنه بالنسبة الى نموذج (K-Model) ولجميع احجام العينات .
4. بالمقارنة بين تطبيق الطريقتين على مصفوفة النموذج الاصلية ووتطبيقها على مصفوفة التحويل للمكونات الزائفة للحدود الدنيا فان كلا الطريقتين تمتلكان MAPE اقل في حالة تطبيقها على مصفوفة تحويل المكونات الزائفة بالمقارنة عند تطبيقها على مصفوفة النموذج الاصلية (لكلا النموذجين).

اما من الجدول (4) عندما $(\sigma = 2.50)$:

1. ان كلا من طريقة (g-inv) و (ST) عند تطبيقها على (S-Model) تعطي قيم اقل لمعيار MAPE بالمقارنة مع (K-Model) لجميع احجام العينات ، في حال تطبيق كلا الطريقتين على مصفوفة البيانات الاصلية او مصفوفة المكونات الزائفة للحدود الدنيا .
2. لكن بمقارنة طريقة (g-inv) مع طريقة (ST) في حالة تطبيقها على مصفوفة النموذج الاصلية فان طريقة (g-inv) تمتلك اقل MAPE بالنسبة الى حجم العينة n=30 فقط ولكن بالنسبة الى باقي احجام العينات فان طريقة (ST) افضل بامتلاكها اقل MAPE لكلا النموذجين.
3. بالنسبة الى تطبيق الطرائق على مصفوفة المكونات الزائفة للحدود الدنيا فان طريقة (g-inv) اقل MAPE من (ST) بالنسبة الى نموذج (K-Model)
4. اما بالنسبة الى نموذج (S-Model) فان طريقة (ST) تمتلك اقل MAPE بالمقارنة مع (g-inv) في حالة حجوم العينات n=30,50 ،ماعدا حجم العينة n=75 على العكس منهم .

واخيرا للمقارنة بين الطرائق بصورة عامة بالنسبة الى الانحراف المعياري $(\sigma = 2, 2.50)$ فانها تزداد فيها قيمة MAPE كلما زاد الانحراف المعياري .

فان اقل MAPE تم الحصول عليه لنموذج (S-Model) في حالة تطبيق طريقة (g-inv) على مصفوفة المكونات الزائفة للحدود الدنيا عند حجم العينة n=75 .

وبهذا نتوصل الى ان زيادة حجم العينة واقل انحراف معياري للخطأ التجريبي مع اجراء تحويل المكونات الزائفة على مصفوفة (S-Model) وتطبيق الطريقة المقترحة (g-inv) تعطي اقل MAPE .

11. الاستنتاجات:

1. ان معامل تضخم التباين يقل باستخدام تحويل المكونات الزائفة للحدود الدنيا لكلا النموذجين (S-Model) و (K-Model) ، وعند تطبيق الطرائق على المستخدمة في البحث فقد توصلنا الى اقل قيم الى معيار MAPE بصورة عامة عند تطبيقها على مصفوفة المكونات الزائفة للحدود الدنيا بالمقارنة مع تطبيقها على مصفوفة النموذج الاصلية.
2. نلاحظ عدد مرات الافضلية لكلا الطريقتين ولكل احجام العينات وباختلاف الانحراف المعياري فان :

عدد مرات الافضلية لنموذج (S-Model) بالنسبة الى المصفوفة الاصلية

Method	عدد مرات الافضلية	النسبة
g-inv	2	0.33
ST	4	0.67

عدد مرات الافضلية لنموذج (K-Model) بالنسبة الى المصفوفة الاصلية

Method	عدد مرات الافضلية	النسبة
g-inv	1	0.17
ST	5	0.83

عدد مرات الافضلية لنموذج (S-Model) بالنسبة الى مصفوفة PSd

Method	عدد مرات الافضلية	النسبة
g-inv	4	0.67
ST	2	0.33

عدد مرات الافضلية لنموذج (K-Model) بالنسبة الى مصفوفة PSd

Method	عدد مرات الافضلية	النسبة
g-inv	6	0.100
ST	0	0

نلاحظ مماسبق ان نسبة الافضلية الكلية لطريقة (g-inv) تبلغ (0.54) اما طريقة (ST) تبلغ نسبتها (0.46) وبهذا تكون طريقة g-inv افضل نسبيا .

12. التوصيات:

بناءاً على ماتم التوصل اليه في الاستنتاجات يوصي الباحث بالاتي:

1. استخدام طريقة (g-inv) بعد اجراء التحويلات المناسبة لمصفوفة المتغيرات التوضيحية في التخلص من مشكلة التعدد الخطي .
2. دراسة نماذج خليط بصيغ اخرى مثلا (التي تحوي تفاعلات ثلاثية) وغيرها من النماذج الكثيرة .
3. اجراء دراسات احصائية في حالة وجود قيم شاذة بالاضافة الى التعدد الخطي بعد ان تتم معالجتها من وجود الشواذ واستخدام احد الطرائق الحصينة.
4. الاخذ بنظر الاعتبار تاثير كمية المكونات على نماذج الخليط .
5. استخدام طرق تحويلات اخرى مثل تحويل المركبات الزائفة للحدود العليا (U- Pseudo component) وغيرها من التحويلات ومعالجة التعدد الخطي في نماذج الخليط .

13. المصادر:

I) كاظم ،اموري هادي & الدليمي ، محمد مناجد عيفان(1988) " مقدمة في تحليل الانحدار الخطي" كلية الادارة والاقتصاد ، جامعة بغداد.

II) **Baksalary,J.K. & Baksalary,O.M.(2004)," Relationships between generalized inverses of a matrix and generalized inverses of its rank-one-modifications" .Linear Algebra and its Applications 388 ,31–44.**

III) **Cornell,J,A.(2002)." Experiments with Mixtures Designs, Models, and the Analysis of Mixture Data" Third Edition. New York: Wiley.**

IV) **Cornell,J,A.(2011)" A Primer on Experiments with Mixtures " New Jersey: Wiley.**

V) **Draper,N.R. & Pukelsheim,F.(1998)." Mixture models based on homogeneous polynomials" Journal of Statistical Planning and Inference .71 , 303–311.**

VI) **Graybill,F.A,(1976)"Theory and application of the linear model" Duxbury Press.**

VII) **Gujarati,D.N.(2004)." Basic Econometrics" Fourth Edition. Statistics.**

VIII) **Khuri, A. I. (2005). "Slack-variable Models Versus Scheff' e Mixture Models". Journal of Applied Statistics, 9, 887–908.**

IX) **Montgomery, D. C. (2013). "Introduction to Linear Regression".**

X) **Salgado,J.C & Alonso,S.(2014)."Kronecker Models Versus Scheffé's Mixture Models". International Conference on Industrial Engineering and Operations Management, January 7 – 9.**

XI) **Scheff' e, H. (1958), "Experiments With Mixtures". J. Roy. Statist Sot., Ser. B, 20, 344-360.**

XII) **Tofallis .C. (2014). "A better measure of relative prediction accuracy for model selection and model estimation". Journal of the Operational Research Society,1–11.**

Compare Estimate Methods of Parameter to Scheff' e Mixture Model By Using Generalized Inverse and The Stepwise Regression procedure for Treatment Multicollinearity Problem

Abstract :

Mixture experiments are response variables based on the proportions of component for this mixture. In our research we will compare the scheff' e model with the kronecker model for the mixture experiments, especially when the experimental area is restricted.

Because of the experience of the mixture of high correlation problem and the problem of multicollinearity between the explanatory variables, which has an effect on the calculation of the Fisher information matrix $(\acute{X}X)$ of the regression model.

to estimate the parameters of the mixture model, we used the (generalized inverse) And the Stepwise Regression procedure, as well as the use of the(Variance Inflation Factor) (VIF) scale to demonstrate the high variances in both models, as well as the use of the (L-Pseudo component) , by Using the R-language simulation To compare them. with critical for compare mean absolute percentage error (MAPE).

Key words: mixture models, variance inflation factor, L-Pseudo component, generalized inverse, Stepwise Regression procedure, mean absolute percentage error