Comparison of Some Methods for Estimating the Scheff'e Model of the Mixture

مقارنة بعض طرائق تقدير إنموذج Scheff'e الخليط

أ. د.دجلة ابراهيم مهدي العزاوي/ كلية الادارة والاقتصاد/ جامعة بغداد

الباحث/ حلا سلمان فرحان

OPEN ACCESS

P - ISSN 2518 - 5764
E - ISSN 2227 - 703X

Received:8/8/2018
Accepted: 2/10/2018

المستخلص

نظرا لما تعانيه تجارب الخليط من مشكلة الارتباطات العالية ووجود مشكلة التعدد الخطي بين المتغيرات التوضيحية وذلك لوجود قيد الوحدة والتفاعلات بينها في النموذج مما يزيد من وجود الارتباطات بين المتغيرات التوضيحية وهذا ما يوضحه عامل تضخم التباين **Variance Inflation Vector (VIF)** ، كذلك تم التطرق الى استخدام تحويل المكونات الزائفة للحدود الدنيا **(L-Pseudo component)** للتقليل من الارتباطات بين مكونات الخليط .

لتقدير معالم إنموذج الخليط اعتمدنا في بحثنا على استخدام طرائق تقدير تعمل على زيادة التحيز وتقلل من التباين منها طريقة إنحدار الحرف **Ridge Regression Method** وطريقة تقدير **(Least Absolute Shrinkage and Selection Operator) (LASSO)** فضلا عن طريقة تقدير الشبكة المرنة **Elastic Net** ، وتمثيله باستخدام المحاكاة بلغة **R** بمعيار المقارنة متوسط مطلق الخطأ النسبي **Mean Absolute Percentage Error (MAPE)**.

المصطلحات الرئيسة للبحث/ إنموذج الخليط ، المكونات الزائفة ، عامل تضخم التباين، إنحدار الحرف، **LASSO** ، الشبكة المرنة ، متوسط مطلق الخطأ النسبي.

Journal of Economics and Administrative Sciences
2019; Vol. 25, No.110
Pages: 392- 405

*بحث مستل من اطروحة دكتوراه

1. المقدمة

يعد إنموذج Scheffe الخليط من النماذج المهمة التي تعنى بدراسة المكونات الكيميائية من الناحية الاحصائية وكيفية التعامل مع هذه المكونات وتأثير التفاعلات عليها. وهو يعد من النماذج الخطية المستخدمة في تقدير سطح الاستجابة ، وهو مقدم من قبل [XI] (1958) Scheffe (S-Model) لدراسة تجارب الخليط من الدرجة الثانية، لاسيما عندما تكون المنطقة التجريبية مقيدة .

2. هدف البحث ومشكلته

نظرا لما تعانيه نماذج تجارب Sheffe الخليط من مشكلة الارتباطات العالية ووجود العلاقات الخطية بين مكونات الخليط فضلا عن وجود التعدد الخطي بين اعمدة مصفوفة المتغيرات التوضيحية نتيجة لتأثير قيد الوحدة على مكونات الخليط $\sum_{i=1}^{p} x_i = 1$ والتفاعلات بين المكونات، يهدف الباحث الى معالجة هذه المشاكل في إنموذج Scheffe الخليط (S-Model) والمقارنة بين طرائق التقدير بهدف الوصول الى افضل استجابة لمكونات الخليط .

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

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

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

حيث :

L_i : تمثل الحد الاعلى للمكون i .

U_i : تمثل الحد الادنى للمكون i .

هنا سوف نأخذ بنظر العناية 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. [I] 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 :

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

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: [II] PP.2

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

قدم Scheffe [11] (1958) نماذج الخليط بصيغ عامة بدرجات مختلفة ليحدد متوسط الاستجابة لدالة الخليط ومنها أنموذج الخليط من الدرجة الثانية (S-Model) : $[V]^{PP.346}$

$$y_i = \sum_{i=1}^{p} \beta_i x_i + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \varepsilon_i \qquad i = 1,2,..,p\,, j = 1,2,..,n \quad (3-4)$$

ويمكن التعبير عنه بالصيغة العامة لنماذج الانحدار :

$$Y = X\beta + \varepsilon \qquad (3-5)$$

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 او الثلاثية ... الخ ، على متغيرات الاستجابة. $[II]^{PP.11-12}[IV]^{PP.888}$,

نتيجة لوجود التفاعلات في نموذج (S-Model) فان أعمدة المصفوفة X تكون معتمدة خطياً وتعاني من تعدد خطي ، وتمتلك ارتباطات عالية بين مكونات الخليط ، من النتائج المترتبة على هذه الخاصية فان مقدرات المربعات الصغرى (OLS) للمعالم تمتلك خطأ تجريبي كبير، وارتباط عالي، فضلا عن ذلك ان هذه المقدرات معتمدة بشكل كبير على دقة موقع النقطة في التصميم ، لذلك سوف نستخدم احد طرائق التحويل L-Pseudo component في تجارب الخليط لمصفوفة المتغيرات التوضيحية بهدف التقليل من هذه الارتباطات بين المكونات.

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

الجزء المهم في تجارب الخليط ان هناك قيود حقيقية على نسب المكونات عندما $0 \le x_i \le 1 \;\; ,\forall i = 1,2,..,p$ ، 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. $[V]^{PP.133}$

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 ، ومن ثم نحصل على مكونات زائفة لكل متغير توضيحي كما يأتي :

$$x_i^* = \frac{x_i - L_i}{1 - L} \qquad (4-1)$$

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، الخطوة التالية هي لحساب قيم الاستجابة لنموذج الخليط **(S-Model)** بدلالة المكونات الزائفة :

$$\begin{aligned} y_i^* = {} & \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + \beta_4 x_{i4}^* + \beta_5 x_{i5}^* + \beta_{12} x_{i1}^* x_{i2}^* \\ & + \beta_{13} x_{i1}^* x_{i3}^* + \beta_{14} x_{i1}^* x_{i4}^* + \beta_{15} x_{i1}^* x_{i5}^* + \beta_{23} x_{i2}^* x_{i3}^* \\ & + \beta_{24} x_{i2}^* x_{i4}^* + \beta_{25} x_{i2}^* x_{i5}^* + \beta_{34} x_{i3}^* x_{i4}^* + \beta_{35} x_{i3}^* x_{i5}^* \\ & + \beta_{45} x_{i4}^* x_{i5}^* + \varepsilon_i \qquad (4-2) \end{aligned}$$

فيما يأتي سوف نوضح مقياس تشخيصي يمكن ان يساعدنا على كشف وتحديد التعدد الخطي.

5. 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 ومن ثم 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 الاتية: [X][PP.325]

$$VIF(\hat{\beta}_j) = (1 - R_j^2)^{-1} \qquad , \forall\, j = 1,2,\dots,p \qquad (5-1)$$

حيث ان :

$$R_j^2 = \frac{x_j' X_j (X_j' X_j)^{-1} X_j' x_j}{x_j' x_j} \qquad \forall\, j = 1,2,\dots,p \qquad (5-2)$$

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 .

ولغرض تقدير سطح الاستجابة لنموذج الخليط سوف نستخدم عدة طرائق لتقدير معالم إنموذج **Sheffe** الخليط منها:

6. مقدرات انحدار الحرف: Ridge Regression Estimator(Rid)

نتيجة المشاكل التي تعاني منها نماذج الخليط لذلك فان رتبة مصفوفة المتغيرات التوضيحية اقل من درجتها وبهذا تكون محددة مصفوفة المعلومات (Fisher Matrix) قريبة من الصفر، مما يسبب وجود ارتباطات عالية بين المتغيرات التوضيحية وهذا بدوره يعطي مقدرات لمعالم نماذج الخليط ذات تباين كبير جدا، ويكون من غير الممكن ايجاد المقدرات لمعالم نماذج الخليط وصعوبة التنبؤ والاختبار، لذلك يعد اسلوب انحدار الحرف (Ridge Regression) من ابرز اساليب معالجة مشكلة التعدد الخطي، فاذا كانت β_j غير مقيدة يمكن ان تكون عرضة الى تباينات كبيرة، للتحكم في التباين يمكننا تنظيم معاملات الانحدار (بمعنى التحكم بمدى نمو المعاملات اي تقليص تاثيرها)، لذلك نفرض قيد انحدار الحرف: $[VIII]^{PP.61-64}$, $[VII]^{PP.112-113}$

$$minimize \sum_{i=1}^{n}(y_i - x_i\beta)^2 \quad s.t \sum_{j=1}^{p}\beta_j^2 \leq \epsilon$$

$$\Leftrightarrow minimize\ (Y - X\beta)'(Y - X\beta)\ \ s.t \sum_{j=1}^{p}\beta_j^2 \leq \epsilon \qquad (6-1)$$

ϵ : معلمة التناغم (Tuning Parameter) ، $\epsilon \geq 0$.

وبحسب الفرض ان X قياسية بمتوسط (Zero) وتباين (1) ، و Y مركزية . وبهذا يمكن كتابة قيد الحرف ℓ_1 كما يعرف بمجموع مربعات البواقي (RSS)(Residual Sum of Square) : $[IX]^{PP.55}$

$$RSS(\beta)_{\ell_1} = \sum_{i=1}^{n}(y_i - \acute{\beta}x_i)^2 + \pi\sum_{j=1}^{p}\beta_j^2$$
$$= (Y - X\beta)'(Y - X\beta) + \pi \parallel \beta \parallel^2 \qquad (6-2)$$

اذ ان :

π : تمثل معلمة الضبط (Regularization Parameter) او (معلمة الجزاء (Penalty Parameter).

ان الحل للمعادلة (6-2) يعطي اقل خطأ تنبؤي (Predication error(PE)) بالمقارنة مع مقدرات (OLS)، فضلا عن $RSS(\beta)_{\ell_1}$ محدب و له حل وحيد وبأخذ المشتقة للمعادلة (5-2) : $[VIII]^{PP.160}$

$$\frac{\partial RSS(\beta)_{\ell_1}}{\partial\beta} = -2\acute{X}(Y - X\beta) + 2\pi\beta$$

فأن:

$$\hat{\beta}_{\pi}^{ridge} = (\acute{X}X + \pi I)^{-1}\acute{X}Y \qquad (6-3)$$

من الواضح ان $\hat{\beta}_{\pi}^{ridge}$ متحيزة ولتوضيح ذلك:

لتكن $R = (\acute{X}X)$ فان :

$$\hat{\beta}_{\pi}^{ridge} = (R + \pi I)^{-1}R(R^{-1}\acute{X}Y)$$

$$= [R(I - \pi R^{-1})]^{-1} R\hat{\beta}^{ols}$$
$$= (I - \pi R^{-1})^{-1} \hat{\beta}^{ols}$$

نستنتج ان :

$$E(\hat{\beta}_{\pi}^{ridge}) = (I - \pi R^{-1})^{-1} E(\hat{\beta}^{ols})$$
$$= (I - \pi R^{-1})^{-1} \beta \qquad \text{if } \pi \neq 0 \text{ , is baise} \qquad (6-4)$$

من ميزات طريقة انحدار الحرف **(Ridge Regression)** انها تعطي تقديرات لكل قيم معالم نموذج الانحدار بدون ان تجري عليها عملية حذف للمعالم لذلك هي تمتلك متغيرات بشكل واسع لانها تقع ضمن النموذج بشكل واضح. $[III]^{PP.2350-2351}$

نلاحظ ان الحل يتعلق بقيمة معلمة الضبط π **(Regularization Parameter)**، ان وجود معلمة الضبط π يعالج مشكلة **(Singular)** حتى وان كانت $\acute{X}X$ لاتمتلك معكوساً عاماً وهو الدافع الاصلي لطريقة انحدار الحرف.

فان لكل π يكون لدينا حل فهي تتحكم بمسار الحلول، كذلك تعد π ايضا معلمة التقلص **(Shrinkage Parameter)** لانها تتحكم بحجم المعاملات وكمية التنظيم (تحدد قيم المعاملات لكل متغير توضيحي).

عندما $\pi \downarrow 0$ نحصل على حل المربعات الصغرى ، اما $\pi \uparrow \infty$ فان $\hat{\beta}_{\pi}^{ridge} = 0$ (نموذج يحوي على الميل الحدي فقط) **(Intercept model)** ان قيم π تتبع مجموعة حلول الحرف .

7. **مقدرات** The LASSO: Least Absolute Shrinkage and Selection Operator (L)

قدمت هذه الطريقة من قبل **Tibshirani (1996)** وان معالم هذه الطريقة هي حل لتحسين مشكلة **(Singular)** وذلك عند اضافة قيد **LASSO** ℓ_2 التي تهدف الى تقليل التباين للنموذج وذلك باضافة قيد عبارة عن مجموع القيم المطلقة للمعالم ، وبهذا فان صيغة التقدير لنموذج الانحدار الخطي العام (3-5) مقابل للصيغة الاتية: $[IV]^{PP.17}$

$$minimize \sum_{i=1}^{n} (y_i - \acute{\beta} x_i)^2 \ s.t \ \sum_{i=1}^{p} |\beta| \leq \epsilon$$

$$\Leftrightarrow minimize \ (Y - X\beta\acute{)}(Y - X\beta) \ \ s.t \ \sum_{i=1}^{p} |\beta| \leq \epsilon \qquad (7-1)$$

وهو مكافئ الى دالة الخطأ :

$$PRSS(\beta_{\pi}^{LASSO})_{\ell_2} = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \pi \sum_{i=1}^{p} |\beta|$$
$$= (Y - X\beta\acute{)}(Y - X\beta) + \pi \parallel \beta \parallel_2 \qquad (7-2)$$

نلاحظ هنا ان لدينا معلمة ضبط π **(Regularization Parameter)** ومعلمة تناغمϵ **(Tuning Parameter)** تتحكم بمقدار التنظيم ، التي تتطابق بمقدار **(one to one)** مع القيمة الابتدائية لـ ϵ): $[I]^{PP.27}$

$$\sum_{i=1}^{p}|\beta| \leq \epsilon \qquad (7-3)$$

ومن هنا يتضح وجود اتجاه معين للحلول المؤشرة لكل (ϵ) ،لذلك ان الصيغة القياسية لحساب القيمة الابتدائية لـ ϵ ():

$$\epsilon_0 = \sum_{j=1}^{p}|\hat{\beta}_j^{ols}| \qquad (7-4)$$

فعندما $\pi = 0$ لن نحصل بالنتيجة على تقليص عدد المتغيرات الداخلة في نموذج الانحدار (اي نحصل على حل مشابه لطريقة المربعات الصغرى الاعتيادية OLS). في كثير من الاحيان اتجاه الحلول يكون باعتماد مؤشر عامل التقليص ϵ_0 . $[VI]^{PP.409-410}$

اذا كانت π كبيرة بما فيه الكفاية او (ϵ) صغيرة ، ستضع بعض المعاملات مساوية الى الصفر تماما ، نستنتج مما سبق ان LASSO تعطينا تقليص في عدد معالم النموذج وتقليل دالة الخطأ ومن ثم تعطي حلول مبعثرة (Sparsity Solution) [وضع القيم التقديرية للمعاملات ذات القيم الصغيرة مساوية الى الصفر]، لذلك نستطيع القول ان LASSO تقوم باختيار النموذج الملائم للبيانات تحت الدراسة . $[VIII]^{PP.68-69}$

8. **مقدرات الشبكة المرنة**: Elastic Net (EL)

تعمل الشبكة المرنة Elastic Net على مجموعة من العوامل المؤثرة على نموذج الانحدار الذي تكون فيه المتغيرات التوضيحية مترابطة بقوة (اي تمتلك ارتباطا قويا) حيث تميل هذه المتغيرات المترابطة الى ان تكون داخل او خارج النموذج معا. كذلك تقوم في وقت واحد على اختيار المتغيرات تلقائيا فضلا عن عملية التقليص بشكل مستمر ، التي من خلالها يمكن تحديد المجموعات المترابطة .

ففي حالة النموذج الخطي العام $(3-5)$على فرض ان متغيرات الاستجابة مركزية والمتغيرات التوضيحية قياسية، لأية قيمة ثابتة غير سالبة (π_1,π_2) يمكن تعريف الشبكة المرنة بالصيغة: $[XIII]^{PP.302-303}$

$$L(\pi_1,\pi_2,\beta) = |Y - X\beta|^2 + \pi_2|\beta|^2 + \pi_1|\beta|_1 \qquad (8-1)$$

عندما :

$$|\beta|^2 = \sum_{j=1}^{p}\beta_j^2$$
$$|\beta|_1 = \sum_{j=1}^{p}|\beta|$$

فان مقدرات الشبكة المرنة هي تقليل للدالة (8-1) :

$$\hat{\beta}^{elastic} = argmin_{\beta}\{\, L(\pi_1,\pi_2,\beta)\} \qquad (8-2)$$

لتكن $\theta = \pi_2/(\pi_1 + \pi_2)$ فان حل $\hat{\beta}$ في الصيغة (8-2) يؤدي الى حالة من الامثلية :

$$\hat{\beta}^{elastic} = argmin_{\beta}|Y - X\beta|^2 \;\; s.t\; (1-\theta)|\beta|_1 + \theta|\beta|^2 \leq \epsilon \qquad (8-3)$$

لبعض قيم ϵ .

ان الدالة $(1-\theta)|\beta|_1 + \theta|\beta|^2$ تسمى جزاء الشبكة المرنة (Penalty Elastic Net)، وهي مزيج من جزاء LASSO وجزاء الحرف Ridge .

عندما $\theta = 1$ فان الشبكة المرنة (Elastic Net) تصبح انحدار حرف (Ridge Regression) لذلك فإننا في هذا البحث سوف نفرض ان $\forall\ \theta \in (0,1)$، فان دالة جزاء الشبكة المرنة هي مفردة عند $(\theta = 0)$ (بدون مشتقة اولى) ، وشديدة التحدب عند $\theta > 0$ لذلك نجد فيها خصائص تجمع بين كل من LASSO و Ridge .

ولكي تكون قيمة π منضبطة سوف نذكر أهم الطرائق الشائعة والقياسية التي اعتمدها الباحثون لاختيار قيمة π هي طريقة (التقاطع الشرعي) (Cross-Validation) .

9. التقاطع الشرعي: (Cross-Validation)

اساس اختيار قيمة π يكون على اساس تقليل متوسط مربعات الخطأ وهي اهم خطوة لتحديد النموذج الافضل الذي يمثل البيانات التجريبية تحت الدراسة ففي حالة وجود نموذج جيد يكون التنبؤ (التقدير) جيدا في حالة لدينا بيانات جديدة مستقلة .

الية عمل هذه الطريقة هو تجزئة البيانات التجريبية الى مجموعات متساوية تقريبا عددها (k) وهذا النهج يسمى (K-Fold Cross Validation) وهو الاكثر شيوعا يعمل على : $[VIII]^{PP.241-242}$

- تقسيم البيانات التجريبية (n) (يشمل هذا التقسيم كلا من $(X),(Y)$ الى (k) من المجموعات المنفصلة بنفس الحجم، على فرض ان $n = (n_1, n_2, \dots \dots, n_k)$ ، الاختيار الشائع هو عندما (k=5) , (k=10).
- لكل k (لكل مجموعة) نطبق النموذج $\hat{Y}_{-k}^{\pi}$ القيمة التقديرية لنموذج الانحدار بقيمة π للمجموعة التجريبية باستثناء المجموعة (k_{th} − fold) والتي نرمز لها n_K .
- حساب قيمة المشاهدات المطابقة للمجموعة n_K ، بالاعتماد على البيانات التجريبية التي تقابل هذه المجموعة (المستبعدة).
- حساب خطأ التقاطع الشرعي (Cross Validation error) (MSECV) للمجموعة k_{th} :

$$(MSECV)_k^{(\pi)} = |n_k|^{-1} \sum_{(X,Y)\in N_k} (Y - \hat{Y}_{-k}^{(\pi)})^2 \qquad (9-1)$$

- بعد ذلك يمتلك النموذج خطأ (MSECV) عام لكل المجاميع التي تم استبعادها للنموذج في كل مرة :

$$(MSECV)^{(\pi)} = |k|^{-1} \sum_{i=1}^{k} (CV\ error)_i^{(\pi)} \qquad (9-2)$$

- نختار π^* التي تجعل $(MSECV)^{(\pi)}$ اقل مايمكن .
- نقوم بحساب النموذج النهائي $\hat{Y}^{(\pi^*)}$ على المجموعة التجريبية بالكامل.

10. 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 ولاسيما 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 تأثير جميع مكونات الخليط بما فيها التفاعلات على مقدار الاستجابة للخليط لذلك تكون جميع قيم معالم النموذج اكبر من الصفر (موجبة او سالبة). $[XII]^{PP.1}$

اما الصيغة الرياضية لحسابه:

$$MAPE = \frac{1}{n} \sum_{i=1}^{p} \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\% \qquad (9-1)$$

n : حجم العينة .

y_i : القيمة الحقيقية للمقدر.

$\hat{y}_i$: القيمة التقديرية .

11. جانب المحاكاة:

نظراً لخصوصية هذه التجارب وما تحتاجه 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 وما يعطيه من مساحة اكبر للتجريب ومن ثم 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، ومن ثم يمكننا استخدام اسلوب المحاكاة للحصول على بيانات تطبيق شرط نماذج الخليط بتوليد خمسة متغيرات عشوائية، $\sum_{i=1}^{5} x_i = 1$ ، $0 \leq x_i \leq 1, \ \ \forall i = 1,2, \dots 5$ مع وجود التفاعلات بين هذه المكونات ومن ثم الحصول على بيانات مرتبطة خطيا وتعاني من مشكلة التعدد الخطي مما يصعب على الباحث عملية التقدير ويتطلب تطبيق طرائق خاصة تتعامل مع هذه المشاكل، وبعد اخذ المتغيرات التوضيحية واجراء عليها تحويل (L-Pseudo component) نحصل على مصفوفتان لقيم المتغيرات التوضيحية لنموذج (S- Model) (مصفوفة المتغيرات التوضيحية للنموذج الاصلية، مصفوفة المكونات الزائفة للنموذج باستخدام الحدود الدنيا).

تم توليد مشاهدات المتغيرات التوضيحية وفق القيد $\sum_{i=1}^{5} x_i = 1$، $0 \leq x_i \leq 1$, $\forall i = 1,2, \dots 5$ وبتوزيع (Uniform Distribution) ولعدة حالات بأحجام عينات مختلفة $n = 30, 50, 75, 100, 150)$ فتصبح لدينا مصفوفتين لنموذج (S- Model)، وبعد ذلك نعمل على المتغيرات الناتجة بتحديد قيم افتراضية للمعالم من خلال تقدير معالم النموذج بطريقة المربعات الصغرى الاعتيادية (OLS) واستخدامها كقيم افتراضية في حالة استخدام كل من بيانات المصفوفة الاصلية للمتغيرات التوضيحية وبيانات مصفوفة التحويل (L-Pseudo component) كما في الجدول (1) :

جدول (1) يوضح القيم الافتراضية لمعالم انموذج (S-Model) لتجارب المحاكاة

parameters	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 إنموذج Sheffe الخليط لكل من مصفوفتي المتغيرات التوضيحية الاصلية ومصفوفة التحويل بالاعتماد على طرائق التقدير، في حال اي تغيير على هذه القيم وتأثيرها على مقدار الخطأ في تقدير معالم النموذج.

فضلا عن توليد الخطأ التجريبي الذي يتوزع طبيعياً بمتوسط صفر وقيم انحراف معياري $\sigma = 2, 2.25, 2.50)$ بمستوى معنوية $\alpha = 0.05$ ، بعد ذلك يتم توليد متغيرات الاستجابة (Y) لنموذج (S- Model) كما في الصيغة (4-3) ، الخطوة اللاحقة هي تقدير معلمات الانموذج (S-Model) وذلك باستخدام طرائق التقدير شبه المعلمية (Rid),(L),(EL) والمقارنة بينها، علما انه تم تكرار هذه العملية (r=1000) مرة.

12. **تحليل البيانات**:Analysis of Data

في هذا الجزء سوف يتم عرض نتائج المحاكاة التي تم الحصول عليها باستخدام لغة البرمجة (R-3.3.3) ومن ثم نتطرق الى تحليلها للنموذج وفقا لطرائق التقدير المستخدمة والتي تمتلك اقل MAPE.

<u>اولاً</u> : قيم معامل تضخم التباين (VIF) عند استخدام مصفوفة المتغيرات التوضيحية الاصلية ومصفوفة التحويل (L-Pseudo component) كما مبين في الجدول (2):

جدول (2) يبين قيم (VIF) لنموذج الخليط عند استخدام المصفوفة الاصلية ومصفوفة التحويل L-Pseudo component

parameters	المصفوفة الاصلية (S-Model)	مصفوفة التحويل PSd (S-Model)
β_1	2607.239	523.541
β_2	81618.616	4567.389
β_3	7999.162	900.240
β_4	20435.332	1107.702
β_5	1487.251	378.457
β_{12}	1140.686	90.908
β_{13}	1781.515	521.573
β_{14}	3245.783	269.782
β_{15}	410.946	313.527
β_{23}	14260.804	1134.798
β_{24}	55876.060	1856.111
β_{25}	2374.087	317.686
β_{34}	1686.775	99.041
β_{35}	220.139	80.840
β_{45}	1264.593	147.656

نلاحظ من الجدول (2) من خلال مقياس (VIF) ان بعد تطبيق تحويل المكونات الزائفة للحدود الدنيا على مصفوفة المتغيرات التوضيحية لنموذج (S- Model) بالمقارنة عند استخدام مصفوفة التحويل (L-Pseudo component) 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 .

<u>ثانياً</u>: نتائج استخدام مصفوفة المتغيرات التوضيحية الاصلية التي توضح قيم MAPE لنموذج (S-Model) عند القيم الافتراضية ، لكل طرائق التقدير وكما موضحة في الجدول (3) و (4) و(5):

جدول (3) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة المتغيرات التوضيحية الاصلية عند 2= σ قيمة الانحراف المعياري للخطأ

Method	n=30	n=50	n=75	n=100	n=150
Rid	0.01867008	0.01840622	0.01835375	0.01787551	0.01688512
L	0.01860133	0.01826395	0.01815913	0.01772676	0.01696184
EL	0.01860669	0.01833059	0.0182696	0.0177907	0.01681296

جدول (4) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة المتغيرات التوضيحية الاصلية عند 2.25= σ قيمة الانحراف المعياري

Method	n=30	n=50	n=75	n=100	n=150
Rid	0.02112876	0.02079539	0.02050185	0.02010336	0.01898436
L	0.0210017	0.02066157	0.02035572	0.01998867	0.01909305
EL	0.0210685	0.0207243	0.02041783	0.0200242	0.01891496

جدول (5) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة المتغيرات التوضيحية الاصلية عند 2.50= σ قيمة الانحراف المعياري

Method	n=30	n=50	n=75	n=100	n=150
Rid	0.02333106	0.02309249	0.02276267	0.02224372	0.02115909
L	0.02321711	0.02296087	0.02263177	0.02216885	0.02153492
EL	0.02327199	0.02302975	0.02269256	0.02218024	0.02113326

نلاحظ من الجداول (3) و (4) و (5) عند استخدام مصفوفة المتغيرات التوضيحية الاصلية عندما $\sigma = 2\ , 2.25\ ,\ 2.50$ ولحجوم العينات $n = 30,50,75,100$ ان طريقة التقدير L تمتلك اقل MAPE تاتي بعدها طريقة التقدير EL ، اما عند حجم العينة $n = 150$ فان طريقة التقدير EL تمتلك اقل MAPE تليها طريقة التقدير Rid.

ثالثاً: نتـائج استخدام مصفوفة التحويل L-Pseudo Component التي توضح قيـم MAPE لنموذج (S-Model) عند القيم الافتراضية ،لكل من طرائق التقدير وكما موضحة في الجدول (6) و (7) و (8):

جدول (6) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة التحويل L-Pseudo Component عند 2= σ قيمة الانحراف المعياري للخطأ

	n=30	n=50	n=75	n=100	n=150
Rid	0.02960028	0.02940854	0.02925963	0.0290376	0.02877413
L	0.04055594	0.0394141	0.03807974	0.03805764	0.03713747
EL	0.03056314	0.03037967	0.03027214	0.03004734	0.03002145

جدول (7) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة التحويل L-Pseudo Component عند 2.25= σ قيمة الانحراف

Method	n=30	n=50	n=75	n=100	n=150
Rid	0.0311502	0.03091748	0.03064144	0.03026941	0.02987684
L	0.0419942	0.04053421	0.03962771	0.03955797	0.03841002
EL	0.03212701	0.03176646	0.03167012	0.03143212	0.03119868

جدول (8) متوسط مطلق الخطأ النسبي (MAPE) لتقدير الاستجابة لنموذج (S-Model) للقيم الافتراضية باستخدام مصفوفة التحويل L-Pseudo Component عند 2.50= σ قيمة الانحراف للخطأ

Method	n=30	n=50	n=75	n=100	n=150
Rid	0.0322775	0.03226372	0.03223303	0.03172118	0.03149842
L	0.04382538	0.04192518	0.04118188	0.04060364	0.03973693
EL	0.03328221	0.03324392	0.03313624	0.03288681	0.03283612

نلاحظ من الجداول(6) و(7) و(8) عند استخدام مصفوفة التحويل L-Pseudo Component عندما $\sigma = 2, 2.25, 2.50$ ، ان طريقة التقدير Rid تمتلك اقل MAPE تأتي بعدها طريقة التقدير EL ولجميع حجوم العينات .

بصورة عامة وللمقارنة بين كل من استخدام مصفوفة المتغيرات الاصلية ومصفوفة التحويل L-Pseudo Component في تقدير سطح الاستجابة لنموذج S-Model فان قيم MAPE عند استخدام المصفوفة الاصلية في التقدير اقل منها عند استخدام مصفوفة التحويل L-Pseudo Component لجميع حجوم العينات وقيم الانحراف المعياري .

13. الاستنتاجات:

1. قيمة عامل تضخم التباين (VIF) تقل عند استخدام مصفوفة التحويل L-Pseudo Component للنموذج منها عند استخدام المصفوفة الاصلية للمتغيرات التوضيحية.
2. يتضح من تحليل النتائج ان دراسة مجال الخليط الكلي يعطي أقل MAPE للنموذج الكلي (متغير الاستجابة) بالمقارنة عند دراسة جزء من مجال الخليط باستخدام تحويل المكونات الزائفة L-Pseudo Component.
3. بزيادة حجم العينة واقل انحراف معياري للخطأ التجريبي وعند استخدام مصفوفة المتغيرات التوضيحية الاصلية (المجال الاصلي للخليط) لنموذج (S-Model) وتطبيق طريقة التقدير EL تعطي اقل MAPE.
4. نلاحظ عدد مرات الافضلية لكل طريقة تقديرية ولكل احجام العينات وباختلاف الانحراف المعياري فان:

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

Method	عدد مرات الافضلية	النسبة
L	12	80%
EL	3	20%

عدد مرات الافضلية لطرائق التقدير لنموذج (S-Model) بالنسبة الى مصفوفة التحويل

Method	عدد مرات الافضلية	النسبة
Rid	15	100%

نلاحظ مما سبق ان نسبة الافضلية الكلية لطريقة التقدير Rid تساوي 50% ونسبة الافضلية لطريقة التقدير L تساوي 40% اما بالنسبة الى طريقة التقدير EL بلغت نسبة الافضلية 10% لجميع تجارب الخليط.

14. التوصيات:

بناءً على ماتم التوصل اليه في الاستنتاجات يوصي الباحث بالاتي:

1. استخدام طريقة التقدير (Rid) باستخدام مصفوفة المتغيرات التوضيحية الاصلية (مجال الخليط الكلي) وهي مفيدة في التخلص من العلاقات الخطية بين مكونات الخليط والحصول منها على افضل استجابة.
2. استخدام طريقة التقدير (L) عند استخدام مصفوفة التحويل L-Pseudo Component (دراسة جزء من مجال الخليط) وهي جيدة في حالة حجوم العينات $n = 30,50,75,100$ في التخلص من العلاقات الخطية بين مكونات الخليط والحصول منها على افضل استجابة.
3. دراسة نماذج Sheff'e الخليط بصيغ اخرى مثلا (التي تحوي تفاعلات ثلاثية والتكعيبية) وغيرها من نماذج الخليط .
4. عمل دراسة لمقارنة بين تجارب الخليط (Mixture Experiment) والتجارب العاملية (Factorial Experiment).
5. استخدام طرائق تحويلات اخرى مثل تحويل المركبات الزائفة للحدود العليا(U- Pseudo component) وغيرها من التحويلات ومعالجة التعدد الخطي في نماذج الخليط .

15. المصادر:

I) صالح ، طارق عزيز (2016) "بعض الطرائق شبه المعلمية في تقدير واختيار المتغير لأنموذج المؤشر الواحد ".اطروحة دكتوراه ، كلية الادارة والاقتصاد، جامعة بغداد.

2- Abdullah, M. N. Jeiad, H. A. & Hussein ,R.A. (2017)" Multibiometric Identification System Based on SVD and Wavelet Decomposition" Engineering and Technology Journal, Vol. 35, Part A, No. 1.

3- Breiman, L. (1996)." Heuristics Of Instability and Stabilization In Model Selection". The Annals of Statistics, Vol. 24, No. 6, 2350-2383.

4- Cand`es E. and Tao T.(2006)"The Dantzig selector: Statistical Estimation when p is Much Larger than n". Available at http:// www.acm. caltech. edu/ ~emmanuel/papers/ Dantzig Selector. pdf.

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

6- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression". Annals of Statistics, 32 (2): 409–499.

7- Frank, L.E. and Friedman, J.H. (1993). "A Statistical View of Some Chemometrics Regression Tools". Technometrics, 35, 109–148.

8- Hastie, T. Tibshirani, R., and Friedman, J. (2008). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Second Edition, Springer Series in Statistics.

9- Hoerl, A.E. and Kennard, R.W. (1970). "Ridge regression: Biased estimation for non-orthogonal problems". Technometrics, 12: 55-67

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

11- Scheffe, H. (1958), "Experiments With Mixtures". J. Roy. Statist Sot., Ser. B, 20, 344-360.

12-Tofallis,c.(2014)." A Better measure of Relative Prediction Accuracy for Model Selection and Model Estimation". Journal of the Operational Research Society , 1–11.

13-Zou, H. and Hastie, T. (2005)."Regularization and variable selection via the elastic net". Journal of the Royal Statistical Society, Series B. 67: pp. 301–320.

Comparison of Some Methods for Estimating the Scheff'e Model of the Mixture

Abstract

Because of the experience of the mixture problem of high correlation and the existence of linear MultiCollinearity between the explanatory variables, because of the constraint of the unit and the interactions between them in the model, which increases the existence of links between the explanatory variables and this is illustrated by the variance inflation vector (VIF), L-Pseudo component to reduce the bond between the components of the mixture.

To estimate the parameters of the mixture model, we used in our research the use of methods that increase bias and reduce variance, such as the Ridge Regression Method and the Least Absolute Shrinkage and Selection Operator (LASSO) method as well as the Elastic Net estimation method, In R the comparison criterion is the absolute mean percent error (MAPE).

Key words/ Mixture Models, L-Pseudo component, Variance Inflation Factor, Ridge Regression , LASSO, Elastic Net , Mean Absolut Percentage Error