تحليل الانحدار الخطي المتعدد

Multiple Linear Regression Analysis

المحاضـــــــرة الرابعة

مثال (02): اعتمادا على المثال السابق (01) أحسب مصفوفة التباين / التغاير لمعاملات نموذج وزن الطفل، ومن ثم حدد تقديرات الأخطاء (الانحرافات) المعيارية لهذه المعاملات.

الحل:

أولا نقوم بحساب التباين (S_e^2):

$$\mathbf{S_e^2} = \frac{\mathbf{Y^T.Y - b^T.X^T.Y}}{\mathbf{n-(k+1)}}$$

$$(Y^TY) = (11.5 \quad 16 \quad . \quad . \quad . \quad 1.4) . \begin{pmatrix} 11.5 \\ 16 \\ . \\ . \\ . \\ 1.4 \end{pmatrix} = 6653.3853$$

وعليه نجد:

$$[Y^T.Y - b^T.X^T.Y] = \left[(6653.3853) - (-2.1819 \quad 1.2008 \quad 0.12457).\begin{pmatrix} 507.7 \\ 1739.2 \\ 45081.8 \end{pmatrix}\right]$$

$$[Y^T.Y - b^T.X^T.Y] = [(6653.3853) - (6596.6948)] = 56.6904$$

وبالتالي فإن:

$$S_e^2 = \frac{56.6904}{50-(2+1)} = 1.20618$$

بضرب التباين (S_e^2) في المصفوفـــة $(X^TX)^{-1}$ التي حسبت في المثال السابق نتحصل على مصفوفة التباين / التغاير الخاصة بمقدرات معالم نموذج الانحدار (b) كما يلي:

$$\mathbf{S_b^2} = \underset{\mathbf{Var(b)}}{\wedge} = S_e^2.(X^TX)^{-1}$$

$$= 1.20618.\begin{pmatrix} 0.789948 & 0.144239 & -0.014509 \\ 0.144239 & 0.036936 & -0.003023 \\ -0.014509 & -0.003023 & 0.000283 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0.95282} & 0.173978 & -0.017501 \\ 0.173978 & \mathbf{0.044551} & -0.003646 \\ -0.017501 & -0.003646 & \mathbf{0.000341} \end{pmatrix}$$

وبالتالي من خلال هذه المصفوفة نجد:

$$\underset{\mathbf{Var(b_0)}}{\wedge} = 0.95282$$

$$\underset{\mathbf{Var(b_1)}}{\wedge} = 0.044551$$

$$\underset{\mathbf{Var(b_2)}}{\wedge} = 0.000341$$

ومنه فان الانحرافات المعيارية لمعاملات نموذج الانحدار هي:

$$\underset{\boldsymbol{\sigma}}{\wedge}(b_0) = \sqrt{0.95282} = \mathbf{0.976125}$$

$$\underset{\boldsymbol{\sigma}}{\wedge}(b_1) = \sqrt{0.044551} = \mathbf{0.21107}$$

$$\hat{\sigma}(b_2) = \sqrt{0.000341} = \mathbf{0.018466}$$

II-6. خصائص البواقي أو الأخطاء

البواقي هي القيم المقدرة لحد الخطأ العشوائي (ε_i) وهي عبارة عن الفروق بين القيم الفعلية والقيم المقدرة للمتغير التابع، وتتميز البواقي بعدة خصائص أهمها:

1-القيمة المتوقعة لأي عنصر من عناصر متجه البواقي تساوي الصفر، أي:

$$\mathbf{E(e) = 0} \qquad \mathbf{(14)}$$

2-استقلال البواقي عن المتغيرات المستقلة أي:

$$\mathbf{X^T.e = 0} \qquad \mathbf{(15)}$$

3-استقلال البواقي عن القيم المقدرة للمتغير التابع أي:

$$\mathbf{\hat{y}^T.e = 0} \qquad \mathbf{(16)}$$

II-7. معامل التحديد المتعدد

معامل التحديد في الانحدار الخطي المتعدد له نفس التفسير كما رأينا في الانحدار الخطي البسيط. فهو يقيس نسبة التغيرات أو الاختلافات في قيم المتغير التابع التي تفسر بواسطة المتغيرات المستقلة في معادلة انحدار المربعات الصغرى.

وكما رأينا في الانحدار الخطي البسيط فان معامل التحديد في هذه الحالة يحسب بنفس الطريقة أي:

$$\mathbf{R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST}} \qquad \mathbf{(17)}$$

وعند استخدام رموز المصفوفات يمكن كتابة: SST ، SSR ، SSE على النحو التالي:

$$\mathbf{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2 = \mathbf{Y^T.Y - n\overline{Y}^2} \qquad \mathbf{(18)}$$

$$\mathbf{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2 = SST - SSE = \mathbf{b^T.X^T.Y - n\overline{Y}^2} \qquad \mathbf{(19)}$$

$$\mathbf{SSE} = \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 = e^T.e = \mathbf{Y^T.Y - b^T.X^T.Y} \qquad \mathbf{(20)}$$

ومنه يمكن كتابة معامل التحديد كما يلي:

$$R^2 = \frac{b^T . X^T . Y - n\bar{Y}^2}{Y^T . Y - n\bar{Y}^2} \qquad (21)$$

ويتصف معامل التحديد المتعدد كما هو الحال في نموذج الانحدار الخطي البسيط بالخاصية التالية:

$$0 \leq R^2 \leq 1$$

ملاحظة:

كلما كانت قيمة معامل التحديد صغيرة، كلما كان الجزء الأكبر من التباين في المتغير التابع راجع الى متغيرات لا توجد في نموذج الانحدار، أما إذا كانت قيمة معامل التحديد قريبة من الواحد الصحيح فان هذا يدل على أن الجزء الأكبر من التباين في المتغير التابع قد تم تفسيره بواسطة المتغيرات المستقلة الموجودة في نموذج الانحدار.

-الاستخدام المُعيب لــ R^2:

في سياق الحديث عن تحليل الانحدار الخطي المتعدد عادة ما يُساء فهم (R^2) أو يُساء استخدامه. ومن الأهمية أن نعلم أن قيمة (R^2) لا يمكن أن تقل عندما تُضاف متغيرات مفسرة الى نموذج الانحدار، حتى ولو كانت هذه المتغيرات لا تساهم بمعلومات إضافية للتنبؤ بقيمة (Y). وهذا صحيح لأن الاختلاف غير المفسر في العينة لقيم (Y)، كما قيست بواسطة SSE تتناقص بوضوح عندما يكون هناك حد إضافي، قد أضيف لنموذج الانحدار بينما يظل مجموع المربعات الكلي SST ثابتا بصرف النظر عن عدد المكونات في النموذج (لأن SST تكون محددة كليا بواسطة قيم X). لهذا فان مجموع مربعات الانحدار SSR (الاختلاف المفسر) يجب أن يزيد على الأقل عندما تَضاف حدود جديدة (متغيرات مستقلة أو مفسرة) الى النموذج.

فإذا استخدمنا (R^2) لتحديد ما إذا كان يجب إضافة عناصر جديدة للنموذج أم لا. إذن السؤال لا يكون ما إذا كانت هناك زيادة في قيمة (R^2) عند إضافة متغيرات جديدة ولكن بكم تزيد (R^2) ؟

(R^2) الكبيرة لا تعني بالضرورة نموذج أفضل، في الحقيقة (R^2) الكبيرة بدرجة كافية يمكن تحقيقها ببساطة بإضافة متغيرات تفسيرية، البعض منها ربما يُساهم في تفسير القليل من التغيرات في قيم (Y) بالعينة. [بعض المحللين يُخطئون بضم عدد كبير من المتغيرات المفسرة في النموذج كأساس للحصول على قيمة عالية لقيمة (R^2)].

- معامل التحديد المعدل (The Adjusted Coefficient of Determination)

كما سبق وأن ذكرنا فان قيمة (R^2) تزيد بزيادة العناصر المضافة لنموذج الانحدار؛ فبإضافة عناصر كافية (متغيرات مفسرة) للنموذج يمكن أن تقترب (R^2) من الواحد الصحيح. لهذا السبب فان هناك علاقة بديلة لقياس جودة التوفيق قد أُقترحت لتأخذ عناصر النموذج في الحسبان، وهذا المقياس الوصفي لجودة توفيق معادلة المربعات الصغرى يسمى **"معامل التحديد المعدل أو المصحح"** ويحسب وفق العلاقة التالية:

$$\mathbf{R_a^2} = 1 - \left(\frac{SSE/(n-k-1)}{SST/(n-1)}\right) = \mathbf{1} - \left[\frac{\mathbf{n-1}}{\mathbf{n-k-1}}\right] . \frac{\mathbf{SSE}}{\mathbf{SST}} \qquad \mathbf{(22)}$$

ونعلم أن:

$$R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}} = \frac{\mathbf{SST-SSE}}{\mathbf{SST}} = \mathbf{1} - \frac{\mathbf{SSE}}{\mathbf{SST}} \Rightarrow \frac{\mathbf{SSE}}{\mathbf{SST}} = 1 - R^2$$

وعليه فان العلاقة رقم (22) تصبح كما يلي:

$$\mathbf{R_a^2} = \mathbf{1} - \left[\frac{\mathbf{n-1}}{\mathbf{n-k-1}}\right] . (\mathbf{1} - \mathbf{R^2}) \qquad \mathbf{(23)}$$

والشيء الملاحظ على معامل التحديد المعدل هو:

-أنه يأخذ قيما أقل من قيم معامل التحديد (قيمة R_a^2 دائما أقل من قيمة R^2).
-أنه يمكن أن يأخذ قيما سالبة، في حين نجد أن قيم R^2 تكون دائما موجبة.
-قيمة (R_a^2) يمكن أن تتناقص عندما تُضاف متغيرات مفسرة غير مناسبة لنموذج الانحدار.
لهذا فانه في تحليل الانحدار الخطي المتعدد يُفضل استخدام معامل التحديد المعدل على معامل التحديد للمقارنة بين نماذج الانحدار المتنافسة.