تحليل الانحدار الخطي المتعدد
Multiple Linear Regression Analysis

المحاضـــــــرة السادسة

11-II. أسلوب تحليل التباين لنموذج الانحدار الخطي المتعدد

من المحتمل أن بعض أو كل المتغيرات المستقلة في معادلة المربعات الصغرى لا تكون مفيدة في تفسير الاختلاف في قيم المتغير التابع (Y)، ومن الأغراض الأساسية لتقييم النموذج هو تحديد أيُ من هذه المتغيرات المُفسِرة إن وُجِد، يجب أن يتواجد في معادلة الانحدار. لكن يجب أن نبحث أولا ما إذا كانت العلاقة بين (Y) وأيُ من المتغيرات المفسرة المحددة موجودة أم لا. لهذا نستخدم تحليل التباين.

ففي مثال "انحدار وزن الطفل على عمره وطوله" ربما نسأل ما إذا كانت توجد علاقة واضحة بين وزن الطفل (Y)، وأي من المتغيرات المستقلة؛ عمر الطفل (X_1)، وطول الطفل (X_2). نفترض مؤقتا أنه **لا توجد علاقة** بين (Y)، وكل من (X_1)، (X_2). هذا يعني أن المعلمتين (β_1)، (β_2) في نموذج انحدار المجتمع **تساوي الصفر**، سنأخذ هذا على أنه **الفرض العدمي. وإ**ذا كانت العلاقة غير واضحة، إذن **على الأقل واحدة من المعلمتين (β_1)، (β_2) لا تساوي الصفر، ويكون هذا هو شكل الفرض البديل.**

وعلى العموم، في نموذج إنحدار المجتمع يكون أسلوب تحليل التباين قائما على اختبار الفرضيتين التاليتين:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \text{Not All } \beta_i \text{ Equal Zero}, (i = 1, 2, \ldots \ldots \ldots, k) \end{cases}$$

إن إحصاء الاختبار الأساسي في منهج تحليل التباين هو **(F)**، حيث تكون القيمة الموضحة في العمود الأخير من جدول تحليل التباين هي القيمة المحسوبة للاختبار.

وجدول تحليل التباين لنموذج الإنحدار الخطي المتعدد يتم إعداده بنفس طريقة إعداد جدول تحليل التباين لنموذج الإنحدار الخطي البسيط و هو يأخذ الشكل التالي:

جدول (2-2): جدول تحليل التباين لنموذج الانحدار الخطي المتعدد

مصدر التباين	مجموع المربعات (SS)	درجات الحرية (df)	متوسط مجموع المربعات (MSS)	إحصاء الاختبار F
الانحدار	$SSR = b^T.X^T.Y - n\overline{Y}^2$	(k+1)-1	$\frac{b^T.X^T.Y-n\overline{Y}^2}{(k+1)-1}$	$\frac{b^T.X^T.Y-n\overline{Y}^2/(k+1)-1}{Y^T.Y-b^T.X^T.Y/n-(k+1)}$
البواقي	$SSE = Y^T.Y - b^T.X^T.Y$	n-(k+1)	$\frac{Y^T.Y-b^T.X^T.Y}{n-(k+1)}$	
المجموع	$SST = Y^T.Y - n\overline{Y}^2$	n-1		

إن النسبة (F) الموضحة في العمود الأخير من هذا الجدول تتبع **توزيع فيشر (F)** بدرجتي حرية "k" و "n-(k+1)".

وبتحديد مستوى المعنوية أو الدلالة تكون قاعدة القرار كما يلي:

-إذا كانت قيمة (F) المحسوبة أكبر من قيمة (F) الجدولية $[\mathbf{F}_{(1-\alpha;\, k;\, n-k-1)}]$ فإننا نرفض الفرضية الصفرية ونقبل بديلتها، أي ليست جميع المعالم قيمها تساوي الصفر. وبالتالي نحكم بوجود علاقة خطية بين المتغير التابع وبعض المتغيرات المستقلة على الأقل.

-أما إذا كانت قيمة (F) المحسوبة أقل من قيمة (F) الجدولية $[\mathbf{F}_{(1-\alpha;\, k;\, n-k-1)}]$ فإننا نقبل الفرضية الصفرية ونرفض بديلتها، أي جميع المعالم تكون معدومة. وبالتالي نحكم بعدم وجود علاقة خطية بين المتغير التابع والمتغيرات المستقلة.

مثال (05): إعتمادا على بيانات مثال نموذج انحدار وزن الطفل، إختبر مدى ملائمة هذا النموذج لتمثيل العلاقة بين المتغير التابع "وزن الطفل" والمتغيرين المستقلين "عمر الطفل وطوله" عند مستوى المعنوية (0,01).

الحل: للإجابة على هذا السؤال يجب صياغة الفرضيتين التاليتين:

$$\begin{cases} H_0: \beta_1 = \beta_2 = 0 \\ H_1: \text{Not All } \beta_i \text{ Equal Zero}, (i = 1, 2) \end{cases}$$

بعد ذلك نقوم بإعداد جدول تحليل التباين وذلك كما هو موضح في الجدول الموالي:

جدول (2-3): جدول تحليل التباين لنموذج إنحدار وزن الطفل على عمره وطوله

مصدر التباين	مجموع المربعات (SS)	درجات الحرية (df)	متوسط مجموع المربعات (MSS)	إحصاء الاختبار F
الانحدار	SSR =1441.50	2	720.75	597.63
البواقي	SSE = 56.6904	47	1.206	
المجموع	SST= 1498.20	49		

ومن الجداول الإحصائية نجد قيمة (F) الجدولية $[\mathbf{F}_{(1-\alpha;\, k;\, n-k-1)}]$ تساوي:

$$\mathbf{F}_{(1-\alpha;\, k;\, n-k-1)} = \mathbf{F}_{(0.99;\, 2;\, 47)} = \mathbf{5.06}$$

ومن هنا نلاحظ أن قيمة (F) المحسوبة (597.63) أكبر من قيمة (F) الجدولية (5.06)، وعليه نرفض الفرضية الصفرية ونقبل بديلتها، أي ليست جميع المعالم قيمها تساوي الصفر. وبالتالي نحكم بوجود علاقة خطية بين متغير وزن الطفل وكل من عمره وطوله وذلك مستوى المعنوية (0.01).

II-12. التقدير والتنبؤ

كما رأينا في تحليل الإنحدار الخطي البسيط يُستخدم نموذج الإنحدار المقدر في:

-تقدير القيمة المتوسطة للمتغير التابع التي تقابل قيم معينة للمتغيرات المستقلة (x_0).

-التنبؤ بمشاهدة جديدة للمتغير التابع التي تقابل قيم معينة للمتغيرات المستقلة (x_0).

1-تقدير القيمة المتوسطة للمتغير التابع:

يتم تقدير القيمة المتوسطة للمتغير التابع (أي $\hat{y}_0$) بالتعويض في نموذج الإنحدار المقدر وذلك كما يلي:

$$\hat{\mathbf{y}}_{\mathbf{0}} = \mathbf{x}_{\mathbf{0}}.\mathbf{b} \qquad (29)$$

حيث:

(x_0): متجه صفي يحتوي على قيم المتغيرات المستقلة المُراد عندها تقدير القيمة المتوسطة للمتغير التابع أي أن:

$$\mathbf{x}_{\mathbf{0}} = (\mathbf{1} \quad \mathbf{x}_{\mathbf{01}} \quad \mathbf{x}_{\mathbf{02}} \quad . \quad . \quad . \quad \mathbf{x}_{\mathbf{0k}})$$

b: متجه عمودي يحتوي على المعالم المقدرة لنموذج الإنحدار:

$$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{pmatrix}$$

إن هذا التقدير يعتبر أفضل تقدير خطي غير متحيز (Best Linear Unbiased Estimator) للقيمة المتوقعة، كما أن لهذا المقدر أقل تباين وهو:

$$\mathrm{Var}(\hat{y}_0) = \mathrm{Var}(x_0.b) = (x_0.\mathrm{var} - \mathrm{cov}(b).x_0^T)$$

$$\mathrm{Var}(\hat{y}_0) = x_0.\sigma^2(x^Tx)^{-1}.x_0^T$$

وباستيفاء فرضية تبعية حد الخطأ العشوائي للتوزيع الطبيعي نجد أن ($\hat{y}_0$) يتبع **التوزيـــع الطبيعـــي بوسط حسابي:**

$$\mathbf{E}(\hat{\mathbf{y}}_{\mathbf{0}}) = \mathbf{E}(\mathbf{x}_{\mathbf{0}}.\mathbf{b}) = \mathbf{x}_{\mathbf{0}}.\boldsymbol{\beta} \qquad (30)$$

وتباين:

$$\mathbf{Var}(\hat{\mathbf{y}}_{\mathbf{0}}) = \boldsymbol{\sigma}^{\mathbf{2}}.\mathbf{x}_{\mathbf{0}}.(\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-\mathbf{1}}.\mathbf{x}_{\mathbf{0}}^{\mathbf{T}} \qquad (31)$$

وبالطريقة المعتادة لتحويل متغير طبيعي الى متغير طبيعي معياري يمكننا أن نحسب القيمة (Z) حيث:

$$\mathbf{Z} = \frac{\hat{y}_0 - x_0.\beta}{\sigma\sqrt{x_0.(x^Tx)^{-1}.x_0^T}} \sim \mathbf{N(0,1)} \qquad (32)$$

وبما أن التباين (σ^2) مجهول فإننا نقدره بالتباين (S_e^2)، وبالتالي في هذه الحالة لا نستخدم التوزيع الطبيعي بل نستخدم توزيع ستيودنت، إذ نجد أن قيمة (T) التي تساوي:

$$\mathbf{T} = \frac{\hat{y}_0 - x_0.\beta}{S_e\sqrt{x_0.(x^Tx)^{-1}.x_0^T}} \sim \mathbf{t}_{(1-\frac{\alpha}{2};\ n-k-1)} \qquad (33)$$

لها توزيع ستيودنت بدرجات حرية $(n-k-1)$.

انطلاقا مما سبق، نجد أن فترة الثقة للقيمة المتوسطة للمتغير التابع تعطى بالصيغة التالية:

$$\mathbf{x_0.b \mp t_{(1-\frac{\alpha}{2};\ n-k-1)}.S_e\sqrt{x_0.(x^Tx)^{-1}.x_0^T}}$$

أو

$$\hat{y}_0 \mp \mathbf{t_{(1-\frac{\alpha}{2};\ n-k-1)}.S_e\sqrt{x_0.(x^Tx)^{-1}.x_0^T}} \qquad (34)$$

مثال (06): اعتمادا على نموذج إنحدار وزن الطفل على عمره وطوله، قدر القيمة المتوسطة لوزن الطفل عندما يكون عمره ست سنوات ونصف وطوله 120 سنتمتر. بعد ذلك حدد فترة ثقة 95 % لهذا التقدير.

الحل:

1-تقدير القيمة المتوسطة لوزن الطفل:

في البداية نقوم بتحديد متجه الصف (x_0) حيث:

$$\mathbf{x_0} = (1 \quad 6.5 \quad 120)$$

وعليه فإن تقدير القيمة المتوسطة المطلوب هو:

$$\hat{y}_0 = x_0.b = (1 \quad 6.5 \quad 120).\begin{pmatrix} -2.1819 \\ 1.2008 \\ 0.12457 \end{pmatrix}$$

$$\hat{y}_0 = 20.5717 \text{ kg}$$

أي أن وزن الطفل يكون في المتوسط 20.6 كغ عندما يكون عمره ست سنوات ونصف وطوله 120 سنتمتر.

2-إيجاد فترة الثقة 95 % لهذا التقدير:

فترة الثقة المطلوبة تأخذ الشكل التالي:

$$\hat{y}_0 \mp t_{(1-\frac{\alpha}{2};\, n-k-1)} \cdot \sqrt{x_0 . S_e^2 . (x^T x)^{-1} . x_0^T}$$

نعلم مما سبق أن:

$$S_e^2 . (x^T x)^{-1} = \begin{pmatrix} \mathbf{0.95282} & 0.173978 & -0.017501 \\ 0.173978 & \mathbf{0.044551} & -0.003646 \\ -0.017501 & -0.003646 & \mathbf{0.000341} \end{pmatrix}$$

وعليه يكون:

$$\mathbf{x_0 . S_e^2 . (x^T x)^{-1} . x_0^T} =$$

$$(1 \quad 6.5 \quad 120) . \begin{pmatrix} \mathbf{0.95282} & 0.173978 & -0.017501 \\ 0.173978 & \mathbf{0.044551} & -0.003646 \\ -0.017501 & -0.003646 & \mathbf{0.000341} \end{pmatrix} . \begin{pmatrix} 1 \\ 6.5 \\ 120 \end{pmatrix}$$

$$\mathbf{x_0 . S_e^2 . (x^T x)^{-1} . x_0^T = 0.120543}$$

ومن جدول توزيع ستيودنت الموضح في الملحق رقم (02) نجد:

$$t_{(1-\frac{\alpha}{2};\, n-k-1)} = t_{(1-\frac{0.05}{2};\, 50-2-1)} = 2.014$$

ومنه فإن فترة الثقة المطلوبة هي:

$$20.5717 \mp 2.014 . \sqrt{0.120543} = [\mathbf{19.872} \ ; \ \mathbf{21.271}]$$

ومن هنا نستنتج بأن وزن الطفل عندما يكون عمره ست سنوات ونصف وطوله 120 سنتمتر سوف يتراوح ما بين 19.872 كغ كحد أدنى، و 21.271 كغ كحد أقصى وهذا بدرجة ثقة 95 %.