اختيار أفضل معادلة انحدار Selecting the best Regression equation

نحتاج في اغلب الأحيان البحث عن المتغيرات المستقلة التي يقصد بأنها تؤثر بشكل معنوي على المتغير المعتمد تحت الدراسة ،وبهذا تبتعد بعض المتغيرات المستقلة لكونها غير أساسية في العلاقة أو فيها أخطاء كبيرة في القياس أو أن يكون تأثيرها مماثل لمتغيرات أخرى في الدراسة.

ولكن يجب التوفيق بين تقليل المتغيرات لتقليل تكلفة الدراسة وبين زيادة المتغيرات المستقلة للحصول على نتائج تنبؤية أفضل.

وسنستخدم ثلاثة مقاييس للحكم على النموذج المفضل

أ- معامل التحديد المتعدد Coefficient of Multiple Determination R_P^2

$$R_P^2 = \frac{SSR}{SST} = 1 - \frac{SSe}{SST}$$

ب- قيمة متوسط مربعات الخطأ MSe: Mean Square error

$$MSe_p = \frac{SSe}{n-p}$$

ج-قيمة إحصائية مالو : Mallowss Cp ststistics

$$C_p = \frac{SSe}{MS} - (n-2p)$$

ومن الطرق التي تستخدم لاختبار أفضل معادلة انحدار:

1- طريقة كل الانحدارات الممكنة All possible regression

نحتاج في هذا الأسلوب إيجاد كل المعادلات الممكنة بالنسبة للبيانات المعطات بوجود k من المتغيرات المستقلة نحتاج إلى معادلات عددها 2k بضمنها معادلة تحتوي كل المتغيرات ومعادلة لاتحتوي على متغيرات $(\hat{y}_i = \bar{y})$، وعندما يكون عدد المتغيرات كبيرا فان إيجاد كل المعادلات الممكنة قد لايكون ملائما عمليا لان زيادة عدد المتغيرات يزيد عدد المعادلات بشكل واضح وسريع فمثلا عندما يكون عدد المتغيرات $(k = 10)$ فيصبح العدد الكلي للمعادلات $2^{10} = 1024$.

فعلى سبيل المثال لو كان لدينا أربعة متغيرات (كتاب خاشع الراوي ص 264) فنقوم أولاً بعمل إنحدار خطي بسيط بين y مع كل من $X_1X_2X_3X_4$

وبعد ذلك نقوم بعمل نموذج ثاني يحتوي على متغيرين مستقلين

$$y \begin{cases} X_1X_2 \\ X_1X_3 \\ X_1X_4 \\ X_2X_3 \\ X_2X_4 \\ X_3X_4 \end{cases}$$

وبنفس الطريقة نقوم بعمل نماذج انحدار لy مع ثلاثة متغيرات مستقلة

$$y \begin{cases} X_1X_2X_3 \\ X_1X_2X_4 \\ X_1X_3X_4 \\ X_2X_3X_4 \end{cases}$$

واخيرا نقوم بعمل نموذج الانحدار لy على كل المتغيرات المستقلة $y \to X_1X_2X_3X_4$

وفي كل خطوة من الخطوات نختار النموذج الذي يحتوي على أعلى R^2 أو اقل MSe أو Cp فمثلا لو اخترنا R^2 كمعيار للمقارنة فسيكون لدينا:

المجموعة	عدد المتغيرات المستقلة	معادلة الانحدار	قيمة R^2
A	1	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_4 X_{i4}$	0.67
B	2	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$	0.97
C	3	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_4 X_{i4}$	0.9823
D	4	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4}$	0.9824

أفضل الجميع هو المعادلة الثانية لأن فرق قيمة معامل التحديد لها عن الأولى كبير وعن المعادلتين الاخيرتين بثلاثة واربعة متغيرات قليل.

ان عملية ادخال متغير مستقل جديد إلى نموذج عملية صعبة جدا تحتاج إلى تفسير وتكاليف ويجب التاكد من أن المتغير الذي تم ادخاله يؤثر على التغيرات الحاصلة في Yفعليا لان ادخال أي متغير يزيد كمن قيمة R^2 لانه يمتلك مجموع مربعات أكبر من الصفر فيزداد معامل التحديد ويؤدي إلى تفسيرات منطقية وكما هو موضح:-

$$R^2 = \frac{SSR}{SST}$$

But: $SSR(x_1x_2)+SSR(x_3/x_1x_2)=SSR(x_1x_2x_3)$

وجميعها كميات موجبة

$$\therefore SSR(x_1x_2)<SSR(x_1x_2x_3)$$
$$R^2_{(x_1x_2)}<R^2_{(x_1x_2x_3)}$$

وبهذا عندما يكون X3 غير مهم ويضاف إلى معادلة الانحدار تزداد قيمة R^2 ويحكم على معادلة الانحدار بأنها جيدة في حين أن زيادة R^2 كانت نتيجة إضافة متغير غير مهم.

في الجدول السابق ليس شرطاً أن تكون المجموعة B تمتلك معامل تحديد أكبر من المجموعة A ولكن الثالثة يجب أن تكون اكبر من الثانية والأولى لان المتغيرات في المجموعة الثالثة تحتوي على المتغيرات في لمجموعة الأولى والثانية ولهذا R^2 لها اكبر

مثال:

لنفس بيانات المثال السابق يطلب إجراء طريقة كل الانحدارات الممكنة لإيجاد أفضل مجموعة من المتغيرات التي تؤثر على Y

SOL:

$$\underline{\hat{\beta}} = (X'X)^{-1}X'Y = \begin{bmatrix} 2.1 \\ -1.4 \\ 5.3 \\ 3.2 \end{bmatrix}$$

All possible regression $= 2^m = 2^3 = 8$

$$\hat{Y}_i = \hat{\beta}_0 + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_{i2} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_3 X_{i3} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_3 X_{i3} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + e_i$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + e_i$$

$$SSR(x_1x_2/x_3)=\begin{bmatrix}\hat{\beta}_1, & \hat{\beta}_2\end{bmatrix}*\begin{bmatrix}C_{11} & C_{12}\\ c_{12} & C_{22}\end{bmatrix}^{-1}*\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=405.421$$

$$SSR(x_2x_3/x_1)=\begin{bmatrix}\hat{\beta}_2, & \hat{\beta}_3\end{bmatrix}*\begin{bmatrix}C_{22} & C_{23}\\ c_{32} & C_{33}\end{bmatrix}^{-1}*\begin{bmatrix}\hat{\beta}_2\\ \hat{\beta}_3\end{bmatrix}=320.958$$

$$SSR(x_1x_3/x_2)=\begin{bmatrix}\hat{\beta}_1, & \hat{\beta}_3\end{bmatrix}*\begin{bmatrix}C_{11} & C_{13}\\ c_{13} & C_{33}\end{bmatrix}^{-1}*\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_3\end{bmatrix}=290.093$$

$$\left.\begin{aligned}SSR(x_1/x_2x_3)&=\frac{\hat{\beta}^2{}_1}{C_{11}}=36.607\\ SSR(x_2/x_1x_3)&=\frac{\hat{\beta}^2{}_2}{C_{22}}=269.66\\ SSR(x_3/x_1x_2)&=\frac{\hat{\beta}^2{}_3}{C_{33}}=277.73257\end{aligned}\right\}$$

أعلاه مجموع المربعات الإضافي للمتغيرات $x_1x_2x_3$

$$SSR(x_1)=SSR(x_1x_2x_3)-SSE(x_2x_3/x_1) \quad =85.257$$
$$SSR(x_2)=SSR(x_1x_2x_3)-SSE(x_1x_3/x_2) \quad =116.1133$$
$$SSR(x_3)=SSR(x_1x_2x_3)-SSE(x_1x_2/x_3) \quad =0.795$$

$$SSR(x_1x_2)=SSR(x_1x_2x_3)-SSR(x_3/x_1x_2) \quad =128.4834$$
$$SSR(x_1x_3)=SSR(x_1x_2x_3)-SSR(x_2/x_1x_3) \quad =136.55$$
$$SSR(x_2x_3)=SSR(x_1x_2x_3)-SSR(x_1/x_2x_3) \quad =369.608$$

$$SSE(x_1)=SST(x_1x_2x_3)-SSR(x_1)$$

$$SST(x_1x_2x_3)=\underline{y}'\underline{y}-n\bar{y}^2=418.716$$

$$SSR(x_1x_2x_3)=\underline{\hat{\beta}}'\underline{X}'\underline{y}-n\bar{y}^2=406.216$$

$$SSE(x_1x_2x_3)=SST(x_1x_2x_3)-SSR(x_1x_2x_3)=12.5$$

$$R^2=\frac{SSR(x_1x_2x_3)}{SST(x_1x_2x_3)}=0.97$$

$$SSE(x_1)=SST(x_1x_2x_3)-SSR(x_1)=333.45$$
$$SSE(x_2)=SST(x_1x_2x_3)-SSR(x_2)=302.59$$
$$SSE(x_3)=SST(x_1x_2x_3)-SSR(x_3)=417.921$$

$SSE(x_1x_2) = 290.23$

$SSE(x_2x_3) = 49.107$

$SSE(x_1x_3) = 282.1657$

$SSE(x_1x_2x_3) = 12.5$

$MSe = \frac{12.5}{6} = 2.083333$

ولغرض معرفة أي معادلة انحدار أفضل من بقية المعادلات سنقوم باستخدام المعايير الثلاثة وكما هو موضح من الجدول أدناه

$R_P^2 = \frac{SSR}{SST} = 1 - \frac{SSe}{SST}$	$MSe_p = \frac{SSe}{n-p}$	$C_P = \frac{SSe}{MS} - (n-2p)$
1-0.2036	41.688	154.054
2-0.277	37.82	139.24
3-0.00189	52.24	194.602
4-0.3069	41.46	135.31
5-0.326	40.309	131.439
6-0.8827	7.015	19.57
7-0.97	2.083	4

فلو استخدمنا R^2 كمقياس للمفاضلة فنختار من كل مجموعة المعادلة التي تحتوي على أعلى قيمة R^2 وكما يلي :

المعادلة	R^2
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_{i2} + e_i$	**0.277**
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + e_i$	**0.88**
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + e_i$	**0.97**

واحسن المعادلات لدينا هي المعادلة التي تحتوي على X2وx3 لان لها كبير وكذلك Mseو Cp قليل لبقية المعادلات وهذا يعني أن أضافة المتغير X1 للنموذج الحاوي على X2 وx3 له تاثير ضئيل على معدل Y.