رگرسیون خطی چندگانه

Y متغیر وابسته

X_1

$\vdots$

X_k

متغیرهای مستقل

$$Y = b_0 + b_1 X_1 + \dots + b_k X_k$$

فرض کنیم Y یک متغیر وابسته، X_1 و X_2 دو متغیر مستقل باشند

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \qquad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = (X'X)^{-1} X'Y$$

$$X'X = \begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} \\ \sum X_{2i} & \sum X_{1i}X_{2i} & \sum X_{2i}^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum Y_i \\ \sum X_{1i}Y_i \\ \sum X_{2i}Y_i \end{bmatrix}$$

اگر دترمینان یک ماتریس 3×3

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\det(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$\det(A) = a_{11}[a_{22}a_{33} - a_{23}a_{32}] - a_{12}[a_{21}a_{33} - a_{23}a_{31}] + a_{13}[a_{21}a_{32} - a_{22}a_{31}]$$

$$\operatorname{adj}(A) = \begin{bmatrix} \Delta_{11} & \Delta_{21} & \Delta_{31} \\ \Delta_{12} & \Delta_{22} & \Delta_{32} \\ \Delta_{13} & \Delta_{23} & \Delta_{33} \end{bmatrix}$$

$$\Delta_{ij} = (-1)^{i+j} \det(A_{ij})$$

و A_{ij} ماتریسی است که از حذف سطر i و ستون j از ماتریس A حاصل می‌گردد

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$$

مثال: معکوس ماتریس $A = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 3 \\ 1 & 4 & -2 \end{bmatrix}$ را بدست آورید.

$$\det(A) = 2\begin{vmatrix} 2 & 3 \\ 4 & -2 \end{vmatrix} - (-1)\begin{vmatrix} 0 & 3 \\ 1 & -2 \end{vmatrix} + 0\begin{vmatrix} 0 & 2 \\ 1 & 4 \end{vmatrix}$$

$$= 2(-4-12) + 1(0-3) + 0(0-2) = -35$$

$$\det(A) = -35$$

$$\Delta_{11} = (-1)^{1+1}|A_{11}| = \begin{vmatrix} 2 & 3 \\ 4 & -2 \end{vmatrix} = -4-12 = -16$$

$$\Delta_{12} = (-1)^{1+2}|A_{12}| = -\begin{vmatrix} 0 & 3 \\ 1 & -2 \end{vmatrix} = 3$$

$$\Delta_{13} = (-1)^{1+3}|A_{13}| = \begin{vmatrix} 0 & 2 \\ 1 & 4 \end{vmatrix} = -2$$

$$\Delta_{21} = (-1)^{2+1}|A_{21}| = -\begin{vmatrix} -1 & 0 \\ 4 & -2 \end{vmatrix} = -2$$

$$\Delta_{22} = (-1)^{2+2}|A_{22}| = \begin{vmatrix} 2 & 0 \\ 1 & -2 \end{vmatrix} = -4$$

$$\Delta_{23} = (-1)^{2+3}|A_{23}| = -\begin{vmatrix} 2 & -1 \\ 1 & 4 \end{vmatrix} = -9$$

$$\Delta_{31} = (-1)^{3+1}|A_{31}| = \begin{vmatrix} -1 & 0 \\ 2 & 3 \end{vmatrix} = -3$$

$$\Delta_{32} = (-1)^{3+2}|A_{32}| = -\begin{vmatrix} 2 & 0 \\ 0 & 3 \end{vmatrix} = -6$$

$$\Delta_{33} = (-1)^{3+3}|A_{33}| = \begin{vmatrix} 2 & -1 \\ 0 & 2 \end{vmatrix} = 4$$

$$\operatorname{adj}(A) = \begin{bmatrix} -16 & -2 & -3 \\ 3 & -4 & -6 \\ -2 & -9 & 4 \end{bmatrix}$$

$$A^{-1} = \frac{1}{-35}\begin{bmatrix} -16 & -2 & -3 \\ 3 & -4 & -6 \\ -2 & -9 & 4 \end{bmatrix}$$

$$A^{-1}A=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

امتحان درستی معکوس

$$A^{-1}A=\frac{1}{-35}\begin{bmatrix}-16&-2&-3\\3&-4&-6\\-2&-9&4\end{bmatrix}\begin{bmatrix}2&-1&0\\0&2&3\\1&4&-2\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

مثال: تعداد جلسات غیبت ها، تعداد ساعات مطالعه و نمره درس ریاضی 7 دانشجو به صورت زیر است.

غیبت X_1	2	1	0	4	3	0	1
مطالعه X_2	3	5	3	1	3	0	1
نمره Y	14	18	16	7	13	6	10

معادله خط رگرسیون نمره بر حسب تعداد ساعات غیبت و مطالعه را بدست آورید.

$$Y=\begin{bmatrix}14\\18\\16\\7\\13\\6\\10\end{bmatrix}\qquad X=\begin{bmatrix}1&2&3\\1&1&5\\1&0&3\\1&4&1\\1&3&3\\1&0&0\\1&1&1\end{bmatrix}\qquad X'X=\begin{bmatrix}n&\sum X_{1i}&\sum X_{2i}\\\sum X_{1i}&\sum X_{1i}^2&\sum X_{1i}X_{2i}\\\sum X_{2i}&\sum X_{1i}X_{2i}&\sum X_{2i}^2\end{bmatrix}$$

$$X'Y=\begin{bmatrix}\sum Y_i\\\sum X_{1i}Y_i\\\sum X_{2i}Y_i\end{bmatrix}\qquad X'X=\begin{bmatrix}7&11&16\\11&31&25\\16&25&54\end{bmatrix}\qquad X'Y=\begin{bmatrix}84\\123\\236\end{bmatrix}$$

X_{1i}	X_{2i}	Y_i	X_{1i}^2	X_{2i}^2	$X_{1i}X_{2i}$	$X_{1i}Y_i$	$X_{2i}Y_i$	Y_i^2
2	3	14	4	9	6	28	42	196
1	5	18	1	25	5	18	90	324
0	3	16	0	9	0	0	48	256
4	1	7	16	1	4	28	7	49
3	3	13	9	9	9	39	39	169
0	0	6	0	0	0	0	0	36
1	1	10	1	1	1	10	10	100
11	16	84	31	54	25	123	236	1130

$$X'X = \begin{bmatrix} 7 & 11 & 16 \\ 11 & 31 & 25 \\ 16 & 25 & 54 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 84 \\ 123 \\ 236 \end{bmatrix}$$

$$\det(X'X) = 7\begin{vmatrix} 31 & 25 \\ 25 & 54 \end{vmatrix} - 11\begin{vmatrix} 11 & 25 \\ 16 & 54 \end{vmatrix} + 16\begin{vmatrix} 11 & 31 \\ 16 & 25 \end{vmatrix}$$

$$= 7(1049) - 11(194) + 16(-221) = 1673$$

$$\operatorname{adj}(X'X) = \begin{bmatrix} 1049 & -194 & -221 \\ -194 & 122 & 1 \\ -221 & 1 & 96 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{1673}\begin{bmatrix} 1049 & -194 & -221 \\ -194 & 122 & 1 \\ -221 & 1 & 96 \end{bmatrix}$$

$$B = (X'X)^{-1}X'Y = \frac{1}{1673}\begin{bmatrix} 1049 & -194 & -221 \\ -194 & 122 & 1 \\ -221 & 1 & 96 \end{bmatrix}\begin{bmatrix} 84 \\ 123 \\ 236 \end{bmatrix} = \begin{bmatrix} 6.84 \\ -0.63 \\ 2.52 \end{bmatrix}$$

$$Y = 6.84 - 0.63X_1 + 2.52X_2$$

ثابت مدل

ضریب مربوط به غیبت

ضریب مربوط به مطالعه

6.84 نمره دانشجو اگر بدون غیبت و بدون مطالعه

-0.63 به ازای هر جلسه غیبت 0.63 نمره کم می‌شود

2.52 به ازای هر ساعت مطالعه 2.52 نمره اضافه می‌شود

ضریب تعیین مدل

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

$$SSTO = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \qquad SSTO = 1130 - \frac{(84)^2}{7} = 122$$

$$SSR = B'X'y - \frac{(\sum y_i)^2}{n} \qquad SSR = \begin{bmatrix} 6.84 & -0.63 & 2.52 \end{bmatrix} \begin{bmatrix} 84 \\ 123 \\ 236 \end{bmatrix} - \frac{(84)^2}{7}$$

$$= 83.79$$

$$SSE = SSTO - SSR \qquad SSE = 122 - 83.79 = 38.21$$

$$R^2 = \frac{SSR}{SSTO} = \frac{83.79}{122} = 0.69$$

69٪ تغییرات نمره توسط عینیت مطالعه تبیین شد و 31٪ بقیه به متغیرات دیگر مربوط است

آزمون خطی بودن مدل

$$\begin{cases} H_0: & \text{مدل خطی نیست} \\ H_1: & \text{" " هست} \end{cases}$$

منبع تغییرات	SS	df	MS	F^*
رگرسیون	SSR 83.79	k 2	$MSR = \frac{SSR}{k}$ 41.9	$\frac{MSR}{MSE} = 4.39$
خطا	SSE 38.21	n-k-1 4	$MSE = \frac{SSE}{n-k-1}$ 9.55	
کل	SSTO 122	n-1 6		

$\not>$ $f_{0.05}[2,4] = 6.94$ چون

پس H_0 رد نمی‌شود یعنی مدل خطی نیست

H_0 را در سطح خطای α رد می‌کنیم اگر

$$F^* > f_\alpha[k, n-k-1]$$

آیا در سطح خطای مجاز می‌توان گفت که تعداد ساعات غیبت تأثیری در تولید داشته؟

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

$b_1 \neq 0$

x_1 تأثیر دارد.

آزمون پارامترهای مدل رگرسیون

$Y = b_0 + b_1 X_1 + b_2 X_2$ معادله رگرسیون نمونه

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ 〃 〃 جامعه

$$\begin{cases} H_0 : \beta_i = 0 & \text{متغیر } x_i \text{ تأثیر ندارد} \\ H_1 : \beta_i \neq 0 & \text{〃 〃 دارد} \end{cases}$$

برای انجام آزمون آماره

$t^* = \dfrac{b_i}{s(b_i)}$ را از رابطه به دست آورید

H_0 را در سطح خطای α رد می‌کنیم اگر

$$|t^*| > t_{\alpha/2,\, n-k-1}$$

$s(b_i)$ از ماتریس واریانس-کوواریانس به دست آید

$$s^2(B) = (X'X)^{-1} MSE = \begin{bmatrix} s^2_{b_0} & & \text{کوواریانس} \\ & s^2_{b_1} & \\ \text{کوواریانس} & & s^2_{b_k} \end{bmatrix}$$

آیا در سطح خطای ۵٪ می‌توان گفت که تعداد ساعات غیبت تاثیری در نمره داشته؟

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

تعداد ساعات غیبت تاثیر ندارد

" " دارد

$$t^* = \frac{b_1}{s(b_1)}$$

$$s^2(B) = (X'X)^{-1} MSE = \frac{1}{1673}\begin{bmatrix} 1049 & -194 & -221 \\ -194 & 122 & 1 \\ -221 & 1 & 96 \end{bmatrix} (9.55)$$

$$= \begin{bmatrix} \underset{s^2_{b_0}}{5.99} & -1.11 & -1.26 \\ -1.11 & \overset{s^2_{b_1}}{0.70} & 0.006 \\ -1.26 & 0.006 & \overset{s^2_{b_2}}{0.55} \end{bmatrix}$$

$$s^2(b_1) = 0.70 \Rightarrow s(b_1) = 0.84$$

$$t^* = \frac{b_1}{s(b_1)} = \frac{-0.63}{0.84} = -0.75$$

چون $|t^*| = |-0.75| \not> t_{0.025, 4} = 2.776$ پس

H_0 رد نمی‌شود یعنی تعداد ساعات غیبت در نمره تاثیر نداشته است.

آیا تعداد ساعات مطالعه در نمره تاثیر داشته است در سطح ۵٪؟

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

ساعات مطالعه تاثیر نداشته

" " داشته

$b_2 = 2.52 \quad s^2(b_2) = 0.55$

$$t^* = \frac{b_2}{s(b_2)} = \frac{2.52}{\sqrt{0.55}} = 3.4$$

چون $|3.4| > 2.776$

پس H_0 رد می‌شود

یعنی ساعات مطالعه در نمره تاثیر داشته است.

آیا در سطح خطای ۵٪ ضریب ثابت را بگوییم در مطالعه ما نقش دارد و معنی‌دار است؟

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases} \qquad t^* = \frac{b_0}{s(b_0)} = \frac{6.84}{\sqrt{5.99}} = 2.79$$

جواب

چون $|2.79| > 2.776$ پس H_0 رد می‌شود

یعنی متغیر ثابت در مطالعه ما نقش معنی‌دار مؤثر است.

$$Y = b_0 + b_2 X_2$$

تنها پیش‌بینی مدل رگرسیون به صورت خواهد بود

مدل

درست است؟ $Y = 6.84 + 2.52 X_2$

خیر زیرا با حذف متغیر X، ضرایب را دوباره محاسبه کنیم

$$b_2 = \frac{\sum x_2 y - \frac{(\sum x_2)(\sum y)}{n}}{\sum x_2^2 - \frac{(\sum x_2)^2}{n}} \qquad a = \bar{Y} - b\bar{x}_2$$

$$b_2 = \frac{236 - \frac{16(84)}{7}}{54 - \frac{(16)^2}{7}} = 2.52 \qquad a = \frac{84}{7} - 2.52\left(\frac{16}{7}\right) = 6.24$$

$$\boxed{Y = 6.24 + 2.52 X_2}$$

روشهای انتخاب متغیرهای مستقل در مدل چندگانه

① روش پیشرو : Forward

در این روش متغیرها به ترتیب تأثیرشان در مدل وارد می‌شوند ابتدا متغیر مؤثرتر یعنی بیشتر [illegible] متغیر با بیشترین [illegible]

[illegible] اول و ... ——— تا جایی که متغیر دیگری تأثیر نداشته باشد.

② روش پسرو Backward

در این روش ابتدا همه کامل به مدل وارد می‌شوند سپس یک به یک متغیرهای کم‌اثر [illegible] حذف می‌شوند

[illegible]

③ روش گام به گام Stepwise

در این روش مرحله اول، روش پیشرو است یعنی متغیر مؤثرتر وارد مدل می‌شود

از این پس با ورود هر متغیر جدید چک می‌کنیم که آیا متغیرهای از قبل موجود باید از

مدل حذف شوند [illegible] .