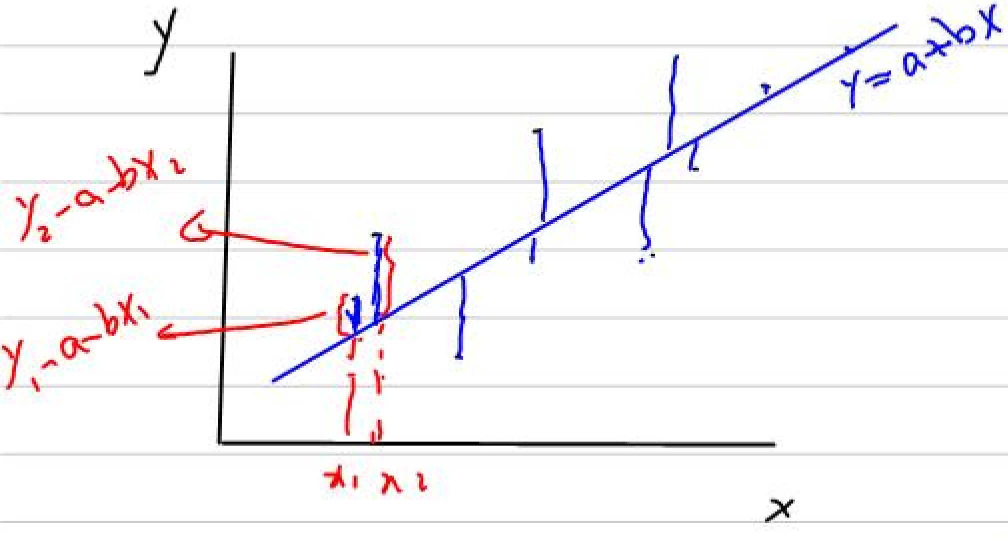
رگرسیون خطی ساده

رگرسیون خطی ساده: فرض کنید x و y دو متغیر می‌باشند و (x_1,y_1)، (x_2,y_2)، ... (x_n,y_n) یک نمونه تصادفی n تایی از زوج متغیر (x,y) باشند.

اگر همبستگی بین x و y شدید باشد می‌توان رابطه بین x و y را به صورت معادله یک خط به شکل $y=a+bx$ بدست آورد. این معادله را خط رگرسیون y برحسب x می‌نامیم. در این معادله y را متغیر وابسته و x را متغیر مستقل می‌نامیم.

از خط رگرسیون برای پیش‌بینی استفاده می‌شود.



روش حداقل مجموع مربعات خطا

هدف یافتن خطی است که مجموع مربعات خطاها در آن حداقل باشد یعنی

$$Q=\sum_{i=1}^{n}(y_i-a-bx_i)^2=\min$$

یعنی بایستی a و b را طوری بیابیم که Q حداقل شود.

پس باید $\frac{\partial Q}{\partial a}=0$ ، $\frac{\partial Q}{\partial b}=0$ باشد.

$$\frac{\partial Q}{\partial a}=\sum -2(y_i-a-bx_i)=0 \quad \Rightarrow \quad n a+b\sum x_i=\sum y_i$$

$$\frac{\partial Q}{\partial b}=\sum -2x_i(y_i-a-bx_i)=0 \quad \Rightarrow \quad a\sum x_i+b\sum x_i^2=\sum x_i y_i$$

پس از حل دستگاه فوق داریم

$$b=\frac{\sum x_i y_i-\frac{\sum x_i \sum y_i}{n}}{\sum x_i^2-\frac{(\sum x_i)^2}{n}}$$

$$a=\bar{y}-b\bar{x}$$

مثال: تعداد ساعات مطالعه و نمره آزمون پایانی 10 دانشجو به صورت زیر آورده شده است.

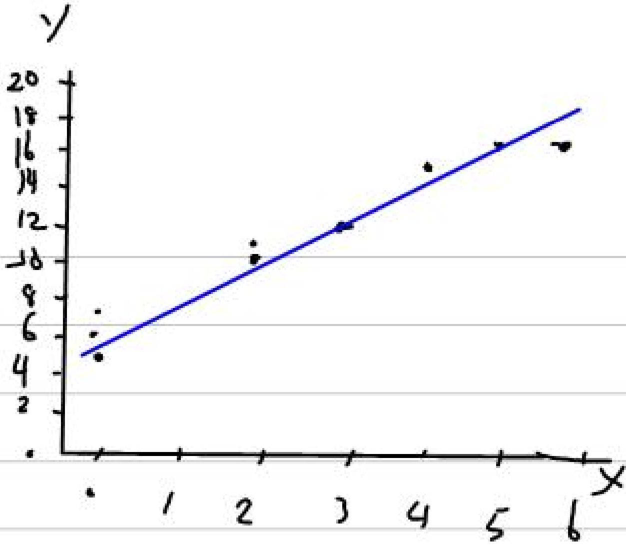
مطالعه x	0	2	3	0	2	4	5	6	2	0
نمره y	5	10	12	6	11	15	16	16	10	7

الف- ابتدا آیا می‌توان رگرسیون را از داده‌ها برآورد کرد؟

ب- معادله خط رگرسیون نمره بر حسب تعداد ساعات مطالعه را به دست آورید.

ج- نمره دانشجویی که 2.5 ساعت مطالعه کرده است را تقریب بزنید.

مطالعه x	0	2	3	0	2	4	5	6	2	0
نمره y	5	10	12	6	11	15	16	16	10	7
x_i^2	0	4	9	0	4	16	25	36	4	0
$x_i y_i$	0	20	36	0	22	60	80	96	20	6
y_i^2	25	100	144	36	121	225	256	256	100	49



$\Sigma x_i = 24 \quad \Sigma y_i = 108 \quad \Sigma x_i^2 = 98 \quad \Sigma x_i y_i = 334 \quad \Sigma y_i^2 = 1312$

$$b = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} = \frac{334 - \frac{24(108)}{10}}{98 - \frac{(24)^2}{10}} = \frac{74.8}{40.4} = 1.85$$

$$a = \bar{y} - b\bar{x} = \frac{108}{10} - 1.85\left(\frac{24}{10}\right) = 6.36$$

پس خط رگرسیون عبارت است از

$$y = 6.36 + 1.85x$$

6.36 ← عرض از مبدأ

1.85 ← شیب خط

6.36 : بیانگر نمره کسی که مطالعه نکند

1.85 : به ازای هر ساعت مطالعه به نمره اضافه می‌شود.

$$x = 2.5 \implies y = 6.36 + 1.85(2.5) = 10.985$$

$$r(x,y) = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}\sqrt{\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}}} = \frac{334 - \frac{24(108)}{10}}{\sqrt{98 - \frac{(24)^2}{10}}\sqrt{1312 - \frac{(108)^2}{10}}} = 0.975$$

شاخص‌های برازش مدل

① ضریب تعیین مدل

$$R^2 = r^2$$

ضریب تعیین یعنی چند درصد تغییرات متغیر وابسته توسط متغیر مستقل تبیین می‌شود.

در مثال فوق

$$R^2 = (0.975)^2 = 0.95$$

یعنی ۹۵٪ تغییرات نمره توسط مطالعه تبیین می‌شود و فقط ۵٪ آن توسط عوامل دیگر تبیین خواهد شد.

② خطی بودن مدل

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

برای انجام آزمون F محاسبات زیر را داریم: تغییرات کل = تغییرات تشریح شده توسط مدل + تغییرات باقیمانده

$$SSTO = SSR + SSE$$

به عبارت دیگر

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

$$SSTO = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$SSR = a\sum Y_i + b\sum X_iY_i - \frac{(\sum Y_i)^2}{n}$$

$$SSE = SSTO - SSR$$

سپس جدولی به نام جدول Anova به صورت زیر تشکیل می‌دهیم

منبع تغییرات	SS	df	MS	F^*
رگرسیون	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
مانده	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
کل	SSTO	$n-1$		

H_0 را در سطح خطای α رد می‌کنیم اگر $F^* > f_\alpha[1, n-2]$

در مثال فوق خطی بودن مدل را در سطح خطای ۵٪ آزمون کنید

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

$$SSTO = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 1312 - \frac{(108)^2}{10} = 145.6$$

$$SSR = a\sum Y_i + b\sum X_i Y_i - \frac{(\sum Y_i)^2}{n} = 6.36(108) + 1.85(334) - \frac{(108)^2}{10} = 138.38$$

$$R^2 = \frac{138.38}{145.6} = 0.95 \qquad SSE = 145.6 - 138.38 = 7.22$$

منبع تغییرات	SS	df	MS	F^*
رگرسیون	138.38	1	138.38	153.33
مانده	7.22	8	0.9025	
کل	145.6	9		

چون $F^* = 153.33 > f_{0.05}[1,8] = 5.32$ پس H_0 رد می‌شود پس مدل خطی هست.

مثال: محققی قصد دارد عملکرد آموزش معلمان را با استفاده از میزان حضور آنها در دوره‌های آموزشی ضمن خدمت پیش‌بینی نماید. برای این کار، نمونه‌ای تصادفی از معلمان انتخاب و تعداد دوره‌های گذرانده شده و نمره عملکرد کسب شده آنها را در جدول زیر ثبت کرده است.

X (تعداد دوره)	3	5	2	0	7	1	6
Y (نمره عملکرد)	15	17	15	10	28	10	27

الف) معادله خط رگرسیون نمره عملکرد برحسب تعداد دوره‌های گذرانده شده را بدست آورید.

ب) عملکرد معلمی که 4 دوره گذرانده را پیش‌بینی کنید.

ج) ضریب مدل را تفسیر کنید.

د) ضریب تعیین را محاسبه، تفسیر کنید.

ه) معنی‌دار بودن مدل را در سطح خطای 5٪ بیازمایید.

$a = 8.41$

$b = 2.63$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2		
3	15	45	9	225		
5	17	85	25	289		
2	15	30	4	225		
0	10	0	0	100		
7	28	196	49	784		
1	10	10	1	100		
6	27	162	36	729		
24	122	528	124	2452		

$$b = \frac{528 - \frac{24(122)}{7}}{124 - \frac{(24)^2}{7}} = \frac{109.7}{41.7} = 2.63$$

$$a = \frac{122}{7} - 2.63\left(\frac{24}{7}\right) = 8.41$$

الف- مدل: $Y = 8.41 + 2.63X$

ب- $X = 4 \Rightarrow Y = 8.41 + 2.63(4) = 18.93$

ج-

8.41 : متوسط نمره عملکرد کسانی که در هیچ دوره‌ای شرکت نکرده‌اند

2.63 : به ازای هر دوره، به نمره عملکرد افزوده می‌شود

د-

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

$$SSTO = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 2452 - \frac{(122)^2}{7} = 325.71$$

$$SSR = a\sum Y_i + b\sum X_iY_i - \frac{(\sum Y_i)^2}{n} = 8.41(122) + 2.63(528) - \frac{(122)^2}{7} = 288.37$$

$$SSE = 325.71 - 288.37 = 37.34$$

$$R^2 = \frac{SSR}{SSTO} = \frac{288.37}{325.71} = 0.89$$

یعنی 89٪ تغییرات نمره عملکرد توسط تعداد دوره‌ها تبیین می‌شود و 11٪ آن به متغیرهای دیگر بستگی دارد.

منبع تغییرات	SS	df	MS	F^*
رگرسیون	288.37	1	288.37	38.6
باقیمانده	37.34	5	7.47	
کل	325.71	6		

چون $F^* = 38.6 > F_{0.05}(1,5) = 6.61$ پس H_0 رد می‌شود یعنی مدل خطی است.

رگرسیون خطی چندگانه

اگر یک متغیر وابسته y و تعداد k متغیر مستقل $x_1, x_2, \ldots, x_k$ در نظر بگیریم و بخواهیم یک رابطه خطی به صورت

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

بدست آوریم گوییم رگرسیون خطی چندگانه داریم.

فرض کنید (x_{11}, x_{21}, y_1) ، (x_{12}, x_{22}, y_2) ، ... ، (x_{1n}, x_{2n}, y_n) یک مجموعه n تایی از سه متغیر x_1 و x_2 و y باشند برای یافتن مدلی به صورت

$$y = b_0 + b_1 x_1 + b_2 x_2$$

به صورت زیر عمل می‌کنیم.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

ماتریس Y

معادلات رگرسیون با توجه به ماتریس‌ها به صورت زیر در می‌آید

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}$$

$$Y = XB$$

که در آن

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

برای کمترین مجموع مربعات خطا نیز B از فرمول زیر به دست می‌آید

$$B = (X'X)^{-1}X'Y$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \qquad X' = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} & x_{2n} \end{bmatrix}$$

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

با بدست آوردن $X'X$ را بدست آورد، $X'Y$ ضرب در ماتریس B به دست آید

طریقه محاسبه کردن دترمینان ماتریس 3x3

فرض کنیم

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

ابتدا دترمینان A را بدست آورید

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

$$= a_{11}\left[a_{22}a_{33} - a_{23}a_{32}\right] - a_{12}\left[a_{21}a_{33} - a_{23}a_{31}\right] + a_{13}\left[a_{21}a_{32} - a_{22}a_{31}\right]$$