تحلیل رگرسیون

رگرسیون در واقع نوعی مدلسازی ریاضی است. در رگرسیون با مطالعه مشاهدات مربوط به یک پدیده، الگویی رسم که با آن الگو شناسایی و تحلیل رفتار پدیده مورد نظر را پیش‌بینی کنیم

$$Y = f(x_1, x_2, \dots, x_n)$$

(... و تناسب شغل با بدن، سابقه کار، سن، درآمد و مرتبه شغلی) f = رضایت شغلی

چون همیشه در پیش‌بینی مقدار خطا (Error) پیش می‌آید و هر خطایی، زیانی دارد لذا باید توابع زیان را بررسی کنیم

انواع توابع زیان در مدلسازی

Y_i : مقدار واقعی $\quad \hat{Y}_i$: پیش‌بینی

۱- حداقل سازی مجموع قدر مطلق خطا

$$\min \sum_{i=1}^{n} |Y_i - \hat{Y}_i| = \min \sum_{i=1}^{n} |e_i|$$

$$e_i = Y_i - \hat{Y}_i$$

فرض کنید برای یک تابع یک متغیره به صورت $Y = \alpha x + \beta$ باشد

$$\hat{Y}_i = \alpha x_i + \beta$$

$$\min \sum |Y_i - \alpha x_i - \beta|$$

$$Y_i - \alpha x_i - \beta = u_i - v_i \quad \rightarrow \quad \text{تغییر متغیر}$$

$$\min \sum_{i=1}^{n} u_i + v_i$$

$$Y_i - \alpha x_i - \beta = u_i - v_i$$

$$u_i, v_i \geq 0$$

۲) الگوی ۲) حداقل حداکثر اثرات

$$\min \quad \max \quad |e:|$$

$$\min \quad \max \quad |y_i - \alpha x_i - \beta|$$

$$\max |y_i - \alpha x_i - \beta| = z$$

$$\min z$$

$$\begin{cases} y - \alpha x_i - \beta \leq z \\ y - \alpha x_i - \beta \geq -z \end{cases}$$

رابطه میان سرمایه و میزان تولید برای شش دوره زمانی بصورت زیر جمع آوری شده است

	۱	۲	۳	۴	۵	۶
سرمایه	۲۵۰	۳۰۰	۴۲۰	۵۰۰	۶۰۰	۷۰۰
تولید	۵۰	۵۵	۸۰	۱۲۲	۱۵۰	۲۰۰

با استفاده از تابع زیان مجموع قدر مطلق خطاها و سپس ماکسیم قدر مطلق خطاها، پارامترهای مدل را برآورد کنید.

$$\min \sum_{i=1}^{6} u_1 + v_1 + u_2 + v_2 + \text{———} + u_6 + v_6$$

$$50 - 250\alpha - \beta = u_1 - v_1$$

$$55 - 300\alpha - \beta = u_2 - v_2$$

$$\alpha = 125$$

$$\beta = 0$$

$$200 - 700\alpha - \beta = u_6 - v_6$$

$$u_i, v_i \geq 0$$

min z

$$d_- - r_d \cdot \alpha - \beta \leq z$$

$$d_- - r_d \cdot \alpha - \beta \geq -z$$

$$r_{--} - v_{..} \alpha - \beta \leq z$$

$$r_{--} - v_{..} \alpha - \beta \geq -z$$

L.P

$\alpha = 0.25$

$\beta = 0$

الگوی حداکثر خطا ، نقطه بدترین حالت توجه می‌کند پس الگوی خوبی نیست (که حالت بدبینانه است)

الگوی مجموع قدر مطلق خطاها ؛ دقیق‌تر هست خود قدر مطلق زیادتوی نیست

الگوی سوم) حداقل مجذورات خطا

$$\min \sum_{i=1}^{n} e_i^2$$

$$Y = f(x_1, \text{——} x_k)$$

$$= b_0 + b_1 x_1 + \text{——} + b_k x_k$$

$$\min \sum (Y_i - b_0 - b_1 x_1 - b_2 x_2 \text{——} - b_k x_k)^2$$

خوشبختانه الگوی فوق یک الگوی شناخته شده است که ماشین سازی آن می‌توانیم پارامترهای الگوی ریاضی مورد نظر را برآورد کنیم مثلاً در یک تابع یک متغیره شکل

$$Y = b_0 + b_1 x_1$$

$$b_1 = \frac{n \sum XY - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

مثال

x	۰	۱	۱
y	۲	۱	۲

$\Sigma xy = (0\times2) + (1\times1) + (1\times2) = 3$

$\Sigma x^2 = 0^2 + 1^2 + 1^2 = 2$

$\Sigma x = 2$

$\Sigma y = 5$

$$b_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{3\times3 - (2)(5)}{3\times2 - 4} = \frac{-1}{2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = \frac{5}{3} + \frac{1}{2}(\frac{2}{3}) = 2/3 + 1/2 = \frac{10+3}{6} = 12/6 = 2$$

$$y = 2 - \frac{1}{2}x$$

کاربرد جبر ماتریسی در برآورد تابع زیان حداقل مجذورات خطا

$$Y = X\beta + U$$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & \vdots & & \vdots \\ 1 & x_{1n} & & x_{kn} \end{pmatrix} \qquad U = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \qquad \hat{\beta} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$$

ماتریس متغیرهای مستقل — متغیرهای مستقل

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

بردار مشاهدات متغیر وابسته

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

x	۰	۱	۱
y	۲	۱	۲

$$X = \begin{pmatrix} ۱ & ۰ \\ ۱ & ۱ \\ ۱ & ۱ \end{pmatrix} \qquad Y = \begin{pmatrix} ۲ \\ ۱ \\ ۲ \end{pmatrix} \qquad X^T = \begin{pmatrix} ۱ & ۱ & ۱ \\ ۰ & ۱ & ۱ \end{pmatrix}$$

$$X^T X = \underset{۲\times۳}{\begin{pmatrix} ۱ & ۱ & ۱ \\ ۰ & ۱ & ۱ \end{pmatrix}} \underset{۳\times۲}{\begin{pmatrix} ۱ & ۰ \\ ۱ & ۱ \\ ۱ & ۱ \end{pmatrix}} = \underset{۲\times۲}{\begin{pmatrix} ۳ & ۲ \\ ۲ & ۲ \end{pmatrix}}$$

$$(X^T X)^{-1} = \frac{۱}{۲} \begin{pmatrix} ۲ & -۲ \\ -۲ & ۳ \end{pmatrix} = \begin{pmatrix} ۱ & -۱ \\ -۱ & ۳/۲ \end{pmatrix}$$

$$X^T Y = \underset{۲\times۳}{\begin{pmatrix} ۱ & ۱ & ۱ \\ ۰ & ۱ & ۱ \end{pmatrix}} \underset{۳\times۱}{\begin{pmatrix} ۲ \\ ۱ \\ ۲ \end{pmatrix}} = \begin{pmatrix} ۵ \\ ۳ \end{pmatrix}_{۲\times۱}$$

$$\underset{۲\times۲}{\begin{pmatrix} ۱ & -۱ \\ -۱ & ۳/۲ \end{pmatrix}} \begin{pmatrix} ۵ \\ ۳ \end{pmatrix}_{۲\times۱} = \begin{pmatrix} ۲ \\ -۱/۲ \end{pmatrix} \begin{matrix} \leftarrow b_0 \\ \rightarrow b_1 \end{matrix}$$

$$Y = ۲ - ۱/۲\, x$$

صـ۵

هزینه تبلیغات	۳	۵	۴	۷	۹	۶	۵	۴	۸
فروش	۱۱	۲۰	۱۶	۲۴	۲۶	۱۵	۲۱	۱۸	۲۷

اعتبارسنجی مدل ← اعتبارسنجی خرد — Micro
اعتبارسنجی کلان — Macro

۱- اعتبارسنجی خرد: عمدتاً نگاه به اجزاء (ضرایب) است. و اینکه وجودشان در مدل ضرورت دارد یا نه؟!
برای پاسخگویی به این سئوال از آزمون t استفاده می‌شود. برای هرکدام از ضرایب

مراحل اعتبارسنجی خرد:

۱- فرض‌های آماری: $\beta_i = 0$ رد اعتبار ضریب
$\beta_i \neq 0$ تأیید اعتبار ضریب

۲- آزمون آماره

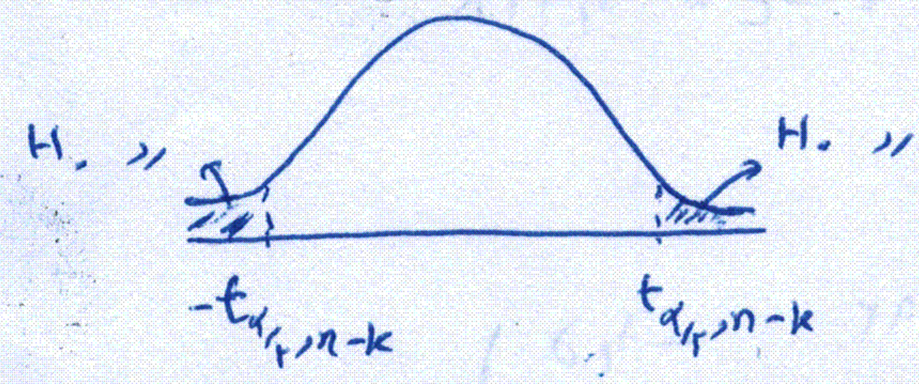
$$t = \frac{b_i - \beta_i}{S_{b_i}}$$

S_{b_i} ← خطای معیار ضریب iام

n: تعداد مشاهدات

k: تعداد ضرایب

۳- مقادیر بحرانی $t_{\alpha/2,(n-k)}$



نحوه بدست آوردن S_{b_i}

مجموع مجذورات خطا

۱) $SSE = (Y^T Y) - (\hat{\beta}^T X^T Y)$

Sum of Square for error

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

حل

۲) $S_e^2 = \frac{SSE}{n-k}$

S_e خطای معیار برآورد

۳) $Var(\hat{\beta}) = S_e^2 (X^T X)^{-1}$

$$\begin{pmatrix} S_{b_0}^2 & & \\ & S_{b_1}^2 & \\ & & S_{b_k}^2 \end{pmatrix}$$

هزینه تبلیغات	3	5	4	7	9	6	5	4	8
هزینه فروش	11	20	16	24	26	15	21	18	27

$$(X^TX)^{-1} = \frac{1}{288}\begin{bmatrix} 321 & -51 \\ -51 & 9 \end{bmatrix} \qquad X^TY = \begin{pmatrix} 178 \\ 1082 \end{pmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 4,179 \\ 2,292 \end{bmatrix} \qquad \hat{y} = 4,179 + 2,292x$$

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy =$$

$$SSE = 3748 - (4,178 \quad 2,292)\begin{pmatrix} 178 \\ 1082 \end{pmatrix} = 59.614$$

$$(Y^TY) - (\hat{\beta}^T X^T Y)$$

$$S_e^2 = \frac{59,614}{9-2} = 8,517 \qquad S_e = 2,918$$

ص ۷

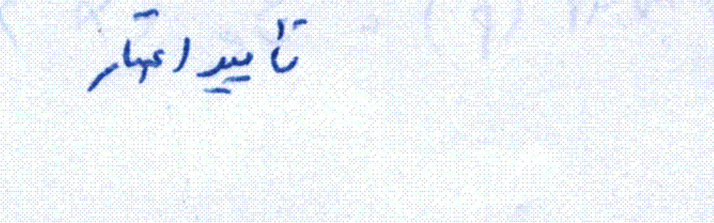
$$Var(\hat{\beta}) = S_e^2 (X^TX)^{-1} = \begin{pmatrix} 9,492 & -1,5 \\ -1,5 & 0,133 \end{pmatrix}$$

$$S_{b_0} = \sqrt{9,492}$$

$$S_{b_1} = \sqrt{0,133}$$

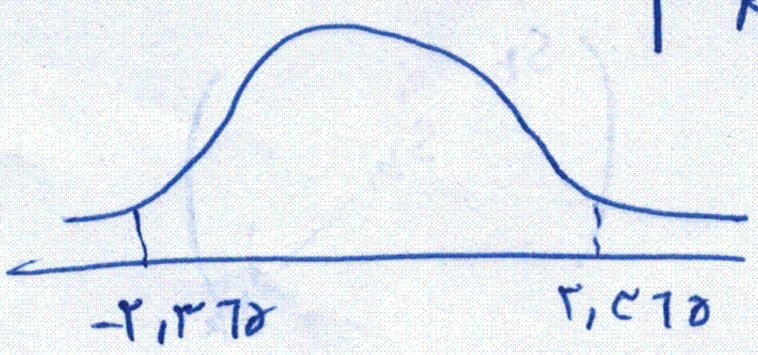
$$t = \frac{3,178}{\sqrt{9,492}} = 1,032$$

$$\begin{cases} \beta_0 = 0 \\ \beta_0 \neq 0 \end{cases}$$



$$t = \frac{2,292}{\sqrt{0,133}} = 6,283$$

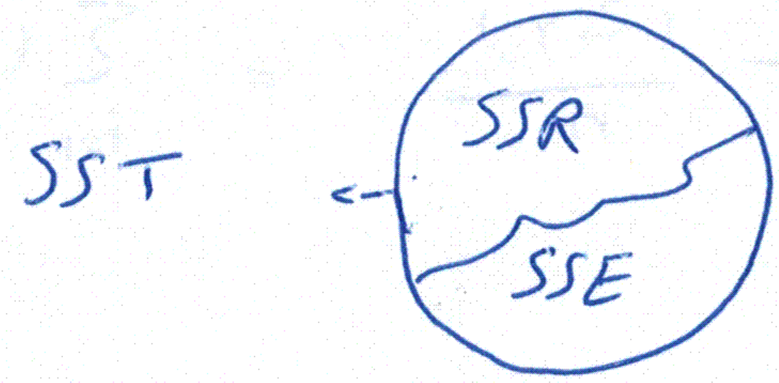
$$\begin{cases} \beta_1 = 0 \\ \beta_1 \neq 0 \end{cases}$$



در H_0 تایید اعتبار

اعتبارسنجی کلان :

برای اعتبارسنجی کلان معادله برآورد شده از تکنیک تحلیل واریانس استفاده می‌شود. در تکنیک تحلیل واریانس، مجموع تغییرات در داده‌های آماری (SST) به دو بخش تفکیک افراز می‌شود. یکی مجموع تغییرات ناشی از مدل (SSR) و دیگری مجموع تغییرات ناشی از خطا (SSE). سپس با استفاده این دو شاخص در قالب یک آزمون آماری به نام F ، معناداری کل مدل مورد بررسی قرار می‌گیرد.



$$SST = SSR + SSE$$

برای اعتبارسنجی کلان جدول تحلیل واریانس رگرسیون به صورت زیر تشکیل می‌گردد.

منبع تغییر	مجموع مجذورات	درجه آزادی df	میانگین مجذورات	F_o	F_t
رگرسیون	SSR مجموع مجذورات رگرسیون	$k-1$	$MSR = \frac{SSR}{k-1}$	$F_o = \frac{MSR}{MSE}$	$F_{\alpha, k-1 \text{ و } n-k}$
خطا	SSE مجموع مجذورات خطا	$n-k$	$MSE = \frac{SSE}{n-k}$		
کل	SST مجموع مجذورات کل	$n-1$			

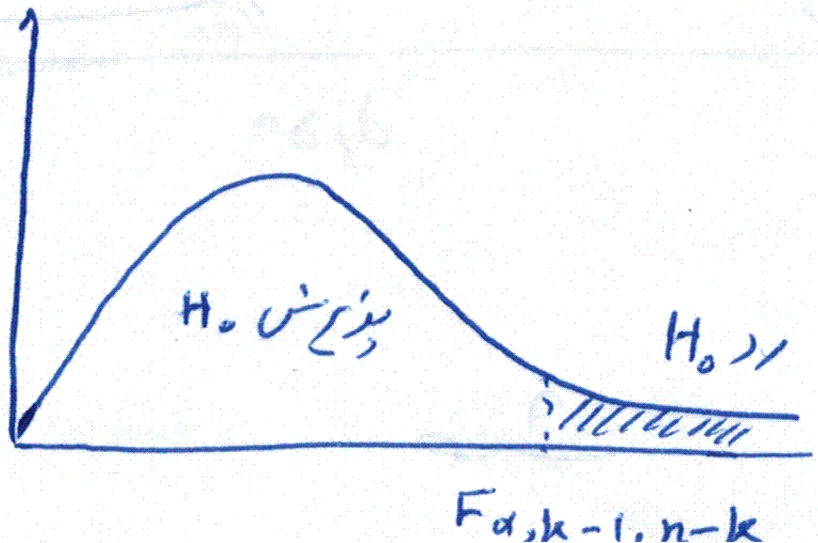
H_0: مدل فاقد اعتبار

H_1: مدل دارای اعتبار

$$F_o = \frac{MSR}{MSE}$$

$F_o > F_t$ اعتبار مدل : H_0 رد

$F_o \leq F_t$ اعتبار : H_0 رد نمی‌شود



x	۳	۵	۴	۷	۹	۶	۵	۴	۸
y	۱۱	۲۰	۱۶	۲۴	۲۶	۱۵	۲۱	۱۸	۲۷

$$SST = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSR = SST - SSE = b_0 \Sigma y + b_1 \Sigma xy - \frac{(\Sigma y)^2}{n} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

حل مثال

$$SST = ۳۷۴۸ - \frac{1}{9}(۱۷۸)^2 = ۲۲۷,۵۵۵$$

$$SSE = ۳۷۴۸ - ۶,۷۸۹(۱۷۸) - ۲,۲۹۲(۱۰۸۲) = ۵۹,۶۱۴$$

$$SSR = SST - SSE = ۱۶۷,۹۴۱$$

$$df_1 = k - 1 = ۲ - 1 = 1$$

$$df_2 = n - k = ۹ - ۲ = ۷$$

$$MSR = \frac{SSR}{k-1} = \frac{۱۶۷,۹۴۱}{1} = ۱۶۷,۹۴۱$$

$$MSE = \frac{SSE}{n-k} = \frac{۵۹,۷۱۴}{۷} = ۸,۵۱۷$$

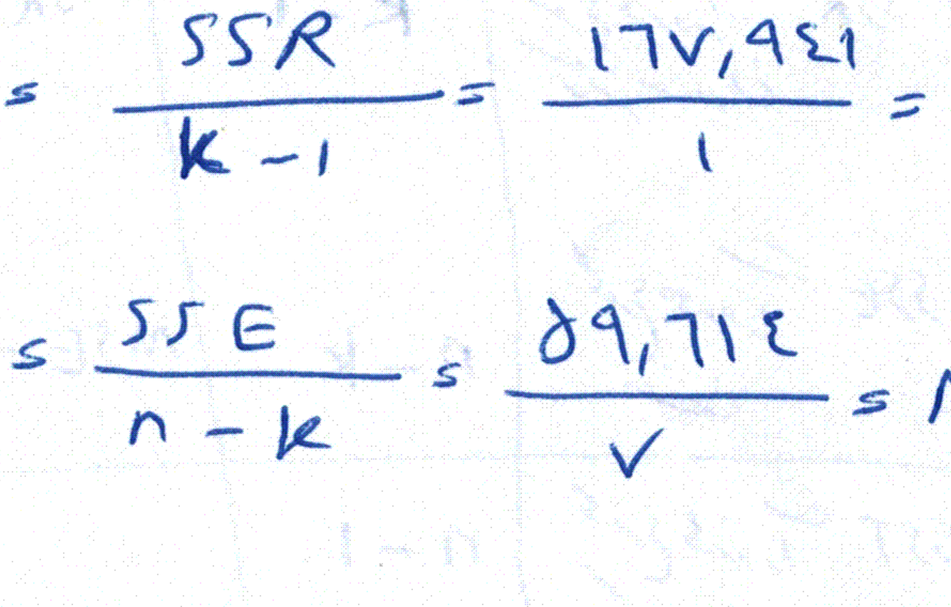
$$F_0 = \frac{MSR}{MSE} = ۱۹,۷۲$$

$$F_t = F_{۰.۰۵,۱,۷} = ۵,۵۹$$

$F_0 > F_t$ H_0 رد

↓

پذیرش اعتبار مدل



ضریب تعیین (R^2)

این ضریب نشان دهنده میزان قدرت تبیین کنندگی مدل است. نشان می دهد تا چه حد تغیرات متغیر وابسته توسط متغیرها مستقل قابل تعبیر و تفسیر است
تغیرات

(۱) $SST = SSE + SSR$

(۲) طرفین بر SST تقسیم شوند $\rightarrow \frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST} \rightarrow R^2$

(۳) $1 = \frac{SSE}{SST} + R^2$

(۴) $R^2 = 1 - \frac{SSE}{SST}$ یا $R^2 = \frac{SSR}{SST}$

مثال قبل $R^2 = \frac{۱۶۸}{۲۲۷/۵} = ۰/۷۳۹$

یعنی تقریباً ۷۴٪ از تغیرات فروش ناشی از تغیرات تبلیغات است و ۲۶٪ بقیه ناشی از عوامل شناخته ای است که هنوز مطالعه نکرده ایم.