مدلهای رگرسیون خطی
Linear Regression Models

Dr. Dehghan
Department of Statistics, University of Sistan and Baluchestan

مدلهای رگرسیون خطی 1

فهرست مطالب:

14.1. خطای معیار
15.1. پیش بینی یک مشاهده جدید
16.1. ارزیابی شرایط اولیه رگرسیون پس از برازش مدل
17.1. انواع باقیمانده ها
18.1. بررسی خطی بودن
19.1. بررسی همگن بودن واریانس
20.1. بررسی استقلال مشاهدات
21.1. بررسی نرمال بودن مشاهدات
22.1. تبدیل داده ها
23.1. پیش بینی تحت تبدیل

فصل دوم : مدلهای غیر خطی

1.2. مقدمه ای بر مدلهای غیر خطی
2.2. تبدیل مدلهای غیر خطی به مدلهای خطی

فصل سوم : رگرسیون چند متغیره

1.3. مقدمه و روابط ماتریسی
2.3.یادآوری رگرسیون ساده با روش ماتریسی
3.3. مدل و نمادها
4.3. پارامترهای رگرسیون چندمتقیره
1.4.3. تسیرپارامترها
2.4.3. برآوردپارامترها
3.4.3. توزیع برآوردگرها
4.4.3. توزیع $\hat{Y}$, $\widehat{\in}$

5.3. تفسیر هندسی رگرسیون
6.3. براورد نااریب σ^2 و آزمون فرضها

فصل چهارم : معیارهای انتخاب مدل

فصل پنجم: هم خطی چندگانه

Linear Regression Models

مدلهای رگرسیون خطی

1. مقدمه:

کلمه رگرسیون از علم ژنتیک genetic و سالها قبل توسط Francis Galton (1822-1911) مطرح شد، وی پسر خاله Charles Darwin است. دقیقاً این اصطلاح توسط Galton و در سال 1886 مطرح و مورد استفاده قرار گرفت.

1.1 هدف رگرسیون:

هدف رگرسیون پی بردن به اثرات متغیر ها روی همدیگر است که در زمینه های گوناگون کاربرد فراوان دارد.

مثال 1:

1. اثر ازون ozone روی سلامتی انسان چیست؟
2. اثر درجه حرارت هوا روی مصرف انرژی در زمستان؟
3. اثر میانگین درآمد روی امید زندگی در کشورهای مختلف چیست؟
4. آیا تعداد ساعات مطالعه آزاد روی نمره پایان ترم اثر دارد؟
5. آیا رابطه ای بین وزن نوزاد و مادر وجود دارد؟

و هدف از آنالیز رگرسیون، مطالعه و بدست آوردن رابطه بین متغیرهای موجود یا فاکتورهای اندازه پذیر پس از مشاهده مقدار متغیرها می باشد.

موضوعات مطرح شده در یک آنالیز رگرسیون چند متغیره عبارتند از:

1. تعیین یا مشخص کردن مدل رگرسیون؟؟
2. برآورد پارامترها
3. انتخاب متغیرها

4. پیش بینی Forecasting

و نتیجه گیری روی مدل رگرسیون انجام خواهد شد.

فرض کنید چندین متغیرآزاد (exogenes) مانند $X_1,\ldots, X_p$ و Y متغیر پاسخ (endogen) که تابعی از $X_1,\ldots, X_p$ است داشته باشیم. یعنی

$$Y= f(x_1, \ldots ,x_p) \qquad (1)$$

به بیان دیگر:

$X_1,\ldots, X_p$ متغیرهای آزاد (کمکی – فاکتور یا توصیفی) (exogenes) می نامند.

در بیشتر حالات روابطه فوق دقیق یا تعینی نیستند.

مثال2: اگر خطای اندازه گیری را در نظر بگیریم می‌توان نوشت :

یا بطور کل می توان نوشت:

$$Y= f(\underline{x})+ \text{Random fluctuation}$$

$$= f(\underline{x}) + \underline{\varepsilon} \qquad (2)$$

رگرسیون خطی چند متغیره (چند گانه): $y= \beta_0+ \beta_1 f_1(x_1)+\ldots + \beta_p f_p(x_p)+\text{fluctuation}$

Y : endogen variable

$x_1,\ldots ,x_p$: exogene variables.

$f_1, \ldots ,f_p$: known transformation

$\beta_0,\ldots , \beta_p$: unknown parameters

که باید برآورد شوند.

اگر مدل رگرسیون مورد نظر فقط یک متغیر آزاد (exogene) داشته باشد مدل را رگرسیون خطی ساده و در غیر این صورت (بیش از یک متغیر آزاد) رگرسیون خطی چند گانه می نامند.

2.1 آنالیز رگرسیون:

اگر رگرسیون خطی ساده مورد نظر باشد سوالات اساسی زیر مطرح می شوند:

1- آیا رابطه منطقی و مستدل بین x, y برقرار است؟
2- آیا مدل $y= \beta_0+\beta_1x_1+\varepsilon$ مناسب است؟
3- چگونه پارامتر های β_0 , β_1 را انتخاب (تعیین) کنیم؟ برای برآورد کردن پارامترها از کدام روش (کمترین توانهای خطا یا یا درستنمایی ماگزیمم) استفاده کنیم؟
4- آیا β_0, β_1 واقعاً لازم است که در مدل باشند؟
بدین منظور لازم است که از آزمون فرض ها جهت معنی داری پارامترها ( ) استفاده شود.
5- چگونه از مدل برازش شده برای پیش بینی استفاده کنیم؟
6- چگونه اثر یک متغیر داخل مدل را اندازه بگیریم؟

برای تفسیر پارامترهای برآورد شده رگرسیون خطی ساده زیر:

$$y_i= \beta_0+ \beta_1+ x_i+ \varepsilon_i \quad ; i=1,2,\dots,n$$

در نظرگرفته و داریم:

$$(x_1, y_1)، \dots ،(x_n, y_n)$$

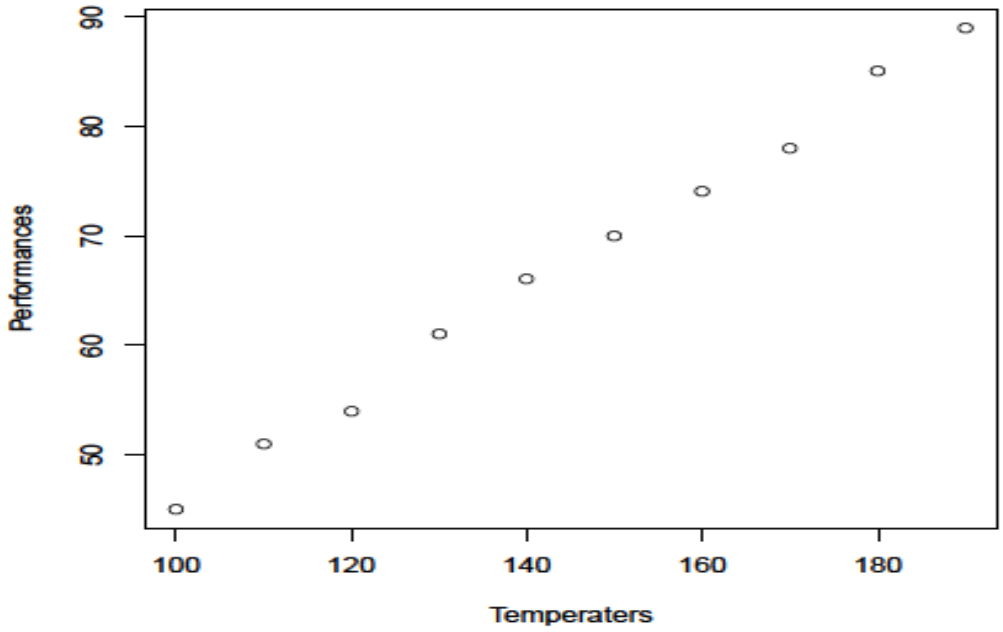
تذکر1:

1- در مدل قبل (رگرسیون) ساده، متغیر پاسخ (endogen) y_n , $y_1,\dots$ متغیر تصادفی هستند.
2- $x_n \dots x_1,$ مقادیر ثابت شناخته شده و غیر تصادفی اند.
3- β_1, β_0 پارامترهای مدل به ترتیب عرض از مبدأ و ضریب زاویه، ناشناخته هستند که باید برآورد شوند.
4- e_n , $e_1, \dots$ مقادیر واقعی ناشناخته و یک متغیر تصادفی(خطا) هستند.

مثال 3- یک سیم رسانا را در نظر گرفته و را ند مان (میزان رسانایی به درصد) آن را بر حسب مقدار درجه حرارت به سانتی گراد اندازه گرفته ایم.

Temperature: 100, 110,120,130,140,150,160,170,180,190

Performances: 45 51 54 61 66 70 74 78 85 89

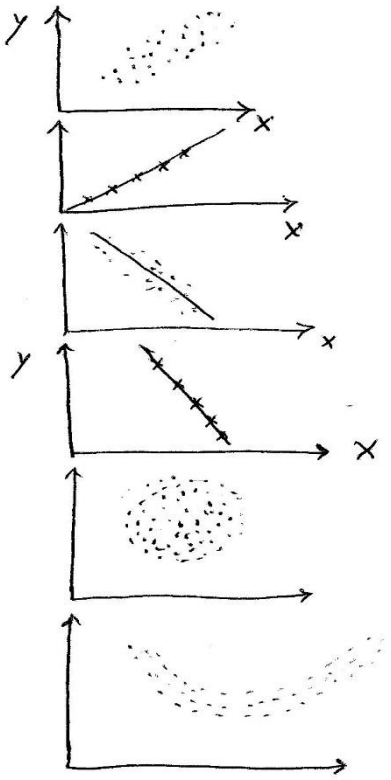


تذکر2: می توان مجموعه داده های $\underline{y}$ و $\underline{x}$ را به صورت ($\underline{y}$ و $\underline{x}$) نشان داد.

$$(\underline{\mathbf{x}}\ ,\ \underline{\mathbf{y}}\ ) =[(\ x_1,y_1),\ \dots\ (\ x_n,\ y_n)]'$$

3.1 بررسی نمودار پراکنش داده ها:

با رسیم نمودار پراکنش داده ها می توان فهمید که همبستگی بین داده های x, y مثبت ، منفی، کامل و یا ناقص هستند.



1- همبستگی مثبت ناقص

2- همبستگی مثبت کامل

3- همبستگی منفی ناقص

4- همبستگی منفی کامل

5- نا همبسته

6- همبستگی غیر خطی

همانطور که ملاحظه کردید در مثال قبل نمودار (گراف) پراکنش (Scatter plot-division) رابطه بین درجه حرارت و راند مان سیم رسانا را توصیف می کند.

برای اینکه بتوان از رگرسیون خطی استفاده کرد لازم است شرایط زیر را در نظر بگیریم:

1- فرض می شود رابطه بین متغیر وابسته endogene ومتغیر آزاد exogene خطی است.
2- Y دارای واریانس ثابت Homogeneous می باشد.
3- مشاهدات متغیر y (پاسخ) مستقل از یکدیگرند.

بدین منظور فرضهای لازم را به صورت زیر بیان می کنیم:

H_1: $E(\varepsilon_i)=0$; $i=1,\dots,n$

H_2: $v(\varepsilon_i)=\sigma^2$; $i=1,\dots,n$

H_3: $cov(\varepsilon_i,\varepsilon_j)=0$ for $i\neq j$

H_4: $\varepsilon_i \sim N(0,\sigma^2)$; $j=1,\dots,n$

یعنی تحت فرضهای فوق می توان فاصله اطمینان و آزمون فرض ها را بنا کرد و روش درستنمایی ماگزیمم را بکار برد.

تفسیر β_0 و β_1 :

در رابطه خطی $Y_i=\beta_0+\beta_1x_i+\varepsilon_i$

β_0 عرض از مبدأ را می توان میانگین y وقتی که x مساوی صفر است در نظر گرفت و β_1 برابر میانگین y است مادامی که x به اندازه یک واحد افزایش داشته باشد.

اگر $\beta_0=0$ باشد مدل فوق شامل عرض از مبدأ نبوده یعنی از مبدأ می گذرد.

مدل فوق را می توان نسبت به x مرکزی کرد یعنی:؟

بطوریکه ؟

4.1 بهترین تابع پیش بینی کننده:

فرض کنید X و Y دو متغیر تصادفی با چگالی های $f_X(x)$ و $f_Y(y)$ چگالی توام $f_{X,Y}(x,y)$ و چگالی شرطی $f_{Y|X}(y|x)$ باشد. حال می خواهیم Y را از روی مقدار داده شده X پیش بینی کنیم. حال فرض کنید D(x) تابع پیش بینی کننده Y در X=x باشد. در صورتیکه قدر مطلق خطای تصادفی این کار |Y-D(x)| است، D(x) را طوری انتخاب می کنیم که $E_{X,Y}(|Y-D(x)|^2)$ مینیمم شود. برای محاسبه آن از فرمول زیر استفاده می کنیم:

$$E_{X,Y}(|Y-D(x)|^2)=E_X(E_{Y|x}(|Y-D(x)|^2)|X)$$

حال $E_{Y|x}(|\mathrm{Y}-\mathrm{D}(\mathrm{x})|^2$(وقتی مینیمم می شود که $\mathrm{D}(\mathrm{x})=E(\mathrm{Y}|\mathrm{x})$.

تمرین 1: رابطه فوق را ثابت کنید.

5.1 **بهترین تابع پیش بینی کننده خطی**:

برای تعیین بهترین تابع پیش بینی کننده باید چگالی توام در دست باشد، ولی اغلب چنین نیست. از این رو در عمل تابع پیش بینی کننده را به صورت خطی زیر در نظر می گیریم:

$$=l(x)=a+bx\hat{Y}$$

که نیازی به چگالی توام نباشد. برای تعیین بهترین تابع پیش بینی کننده خطی باید a , b را طوری اختیار کنیم که MSE یعنی

مینموم شود. با توجه به اینکه داریم:

بنابر این بهترین تابع پیش بینی کننده خطی وقتی که پارامترهای X و Y معلوم باشند عبارتست از:

می دانیم هرگاه X و Y دارای توزیع توام نرمال باشند، $E(Y|X)$ هم دارای فرم خطی $l(x)$ می باشد و انتخاب فرم خطی برای تابع پیش بینی کننده توجیه پذیر است. خلاصه اینکه هرگاه پارامترهای میانگین، واریانس و ضریب همبستگی X و Y را داشته باشیم، بدون نیاز به توزیع توام، می توانیم بهترین تابع پیش بینی کننده خطی را بدست آوریم.

6.1 برآورد بهترین تابع پیش بینی کننده خطی

پارامترها و برآورد آنها در مدل رگرسیون خطی ساده:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1, \dots, n$$

در اینجا علاوه بر برآورد β_0 و β_1 باید واریانس باقیمانده های تصادفی σ^2 را نیز برآورد کرد که بدین منظور از روش کمترین توانهای دوم خطا و روش درستنمایی ماگزیمم استفاده می کنیم.

الف) روش کمترین توانهای دوم خطا

فرض کنید $\widehat{\beta_0}$ و $\widehat{\beta_1}$ برآورگردهای β_0 و β_1 باشند. $\hat{e} = Y_i - \hat{Y}_i$

ب) روش درستنمایی ماگزیمم:

با فرض مستقل بودن مشاهدات y_i تابع درستنمایی بصورت زیر خواهد بود:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i)$$

بدین منظور باید تابع توزیع y_i ها را بدانیم و از طرفی داریم:

$$y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i, \quad i = 1,2, \dots, n$$

با توجه به چهارمین شرط قبلی یعنی

$$H_4: \quad e_i \sim N(0, \sigma^2)$$

و ثابت بودن $(\beta_0 + \beta_1 x_i)$ خواهیم داشت:

(4)

که به منظور ماگزیمم کردن این عبارت بهتر است اول لگاریتم گرفته شود :

تذکر3:همانطورکه ملاحظه می شود جمله ی اول بهβ_1وβ_0 بستگی ندارد و فقط لازم است از جمله دوم نسبت به β_1 وβ_0 مشتق گرفته برابر صفرقراردهیم . ملاحظه می شود که جوابهای روش ML و روش MSE دقیقا یکی هستند .

مزیت روش ML بر MSE این است که اجازه می دهد واریانس σ^2 را نیز برآورد کنیم

و نیز اجازه میدهد که توزیع پارامترها و پیش بینی و آزمون فرض ها را انجام دهیم.

از رابطه (1) ملاحظه می شود که خط رگرسیون الزاماً از نقطه $(\bar{x}, \bar{y})$ (مرکز ثقل) می گذرد.

و سرانجام از رابطه (2) خواهیم داشت.

مثال4- در مثال قبل پارامترهای رگرسیون خطی ساده را برآورد کنید.

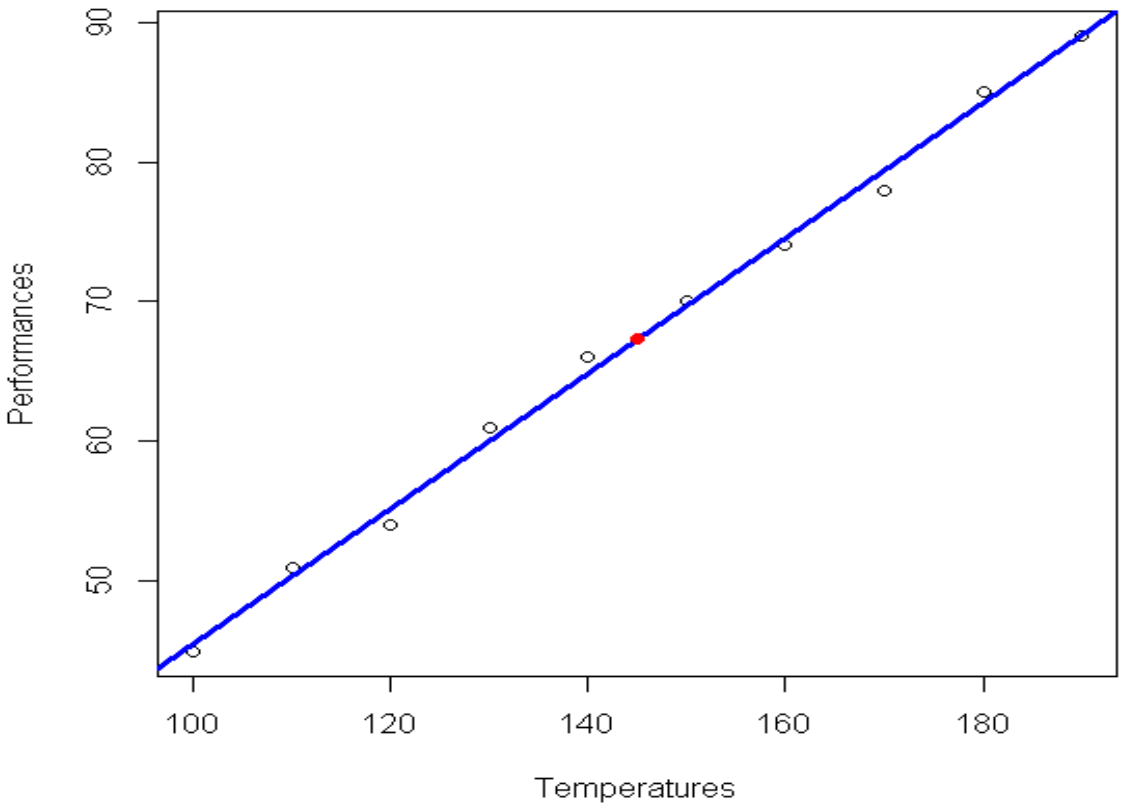
$$n = 10, \quad \sum_{1}^{10} x_i = 1450 \ , \ \sum_{1}^{10} x_i^2 = 218500 \ , \quad \sum_{1}^{10} y_i = 673 \ , \sum_{1}^{10} x_i y_i$$

$$\sum_{1}^{10} x_i y_i = 101570$$

و در نتیجه داریم:

در نتیجه مدل برازش شده عبارتست از:

و شکل ذیل خط برازش و پراکنش داده ها را نشان می دهد.



مثال 5: متغیر پاسخ را بر آورد و خطاها را محاسبه کنید:

i	Y_i	$\hat{y}_i$	$\hat{\varepsilon}_i$
1	45	45.561	-0.561
2	51	50.391	0.609
3	54	55.221	-1.221
4	61	60.051	0.949
5	66	64.881	1.119

6	66	69.711	0.289
7	70	74.541	-0.541
8	74	79.371	-1.371
9	78	84.20	0.799
10	85	89.031	-0.031

مثال 6: نمودار خطاها را رسم کنید؟

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\hat{e}_i = y_i - \hat{y}_i$ اگر

معادله نرمال خط رگرسیون: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

تمرین 2 - ثابت کنید :

a) $\sum e_i = 0$

b) $\sum e_i x_i = \sum e_i x_i = 0$

c) $\bar{\hat{y}} = \bar{y}$

d) $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i$

تعریف:

پراکندگی Y : $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$

پراکندگی $\hat{Y}_i$: $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2$

پراکندگی باقیمانده ها : $\sum \hat{e}_i^2 = \sum (Y_i - \hat{Y}_i)^2$

برآورد واریانس σ^2:

با توجه به معادله (4) درستنمایی ماگزیمم، پارامتر باقیمانده یعنی σ^2 را برآورد می کنیم .

با توجه به $\hat{\beta}_1 , \hat{\beta}_0$ داریم:

مثال7- با توجه به مثال قبل σ^2 را برآورد کنید.

$$\hat{\sigma}^2 = 0.7224$$

7.1 توزیع پارامترها:

در مدل دگرسیون خطی ساده و تحت و فرض H4 داشتیم:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$$

و اینکه $\beta_0 + \beta_1 x_i$ مقداری ثابت است داریم $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

I) توزیع $\hat{\beta}_1$: (قبلاً $\hat{\beta}_1$ را بدست آورده ایم)

اگر فرض کنیم ؟

واریانس $\hat{\beta}_1$:

؟

=؟

بنابراین:

؟

II) توزیع $\hat{\beta}_0$: ار آنجائیکه می دانیم، $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$

و رابطه قبلی:

؟

ترکیبی خطی از یک متغیر تصادفی نرمال ؟

است پس نتیجه نیز دارای توزیعی نرمال می باشد که می بایست امید و واریانس آنرا بدست آوریم.

تمرین 3 :

$\text{cov}\,(\hat{\beta}_0, \hat{\beta}_1) = ?$

$$\sigma_{\hat{\beta}_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

8.1 تجزیه و تحلیل واریانس:

یکی از پرسشهای مهم در این درس این است که دقت برآورد چه میزان است؟

از طرفی داشته باشیم

اگر $\sum(y_i - \bar{y})^2$ مجموع مربعات خطای کل بنامیم SS_T

اگر $\sum(\hat{y}_i - \bar{y})^2$ مجموع مربعات خطای بیان شده بنامیم SS_R(چرا؟)

اگر $\sum e_i^2$ مجموع مربعات خطای بیان نشده بنامیم SS_E :

$$SS_T = SS_R + SS_E$$

تذکر 4: *بنا به قضیه گاوس- مارکف* ،

$\hat{\beta}_1, \hat{\beta}_0$ در بین تمام برآوردگرهای نااریب β_1وβ_0 که تابعی خطی از $y_1, \dots, y_n$ هستند دارای کمترین واریا نس می باشند. بنابراین $\hat{\beta}_1, \hat{\beta}_0$ *برای* β_0 و β_1 ، $BLUE$ هستند.(Best Linear unbiased Estimator).

تمرین 4 : چرا $\hat{\beta}_1, \hat{\beta}_0$ *خطی نامیده می شوند؟*

$\hat{\sigma}^2$ یا همان MSE برآورد نااریب σ^2 می باشد چون ...

از طرفی؟

دارای n-2 درجه آزادی است چون دارای دو معادله زیر هستیم که از تعداد (n) کم می شود.

؟

از طرفی داریم:

$$e_i = y_i - \hat{y}_i \sim N(o, \sigma^2)$$

؟

؟

9.1 ضریب تعیین The coefficient of determination

ضریب تعیین که به درصد بیان می شود را با R^2 نشان می داده و بصورت زیر تعریف می کنیم:

$$R^2 = \frac{SS_R}{SS_T}$$

تمرین 5: نشان دهید، $0 < R^2 < 1$.

در نتیجه می توان نوشت:

تذکر5: ضریب همبستگی با n یا n-1 تعریف می شود.

اگر R1= همبستگی خطی دقیق (کامل) و مثبت است.

اگر R^* 1- = همبستگی خطی دقیق(کامل) و منفی است.

اگر $0 < R^* < 1$ همبستگی خطی مثبت (ناقص).

اگر $-1 < R^* < 0$ همبستگی خطی منفی (ناقص).

اگر R^* 0= باشد آنگاه x ,y ناهمبسته خطی هستند.

تذکر6: اگر x ,y مستقل باشند آنگاه x ,y نا همبسته اند ولی عکس آن ممکن است درست نباشد.

هر چه R^2 بزرگتر باشد نشان می دهند که بخش مهمی (بزرگی) از تغییرات توسط x_i ها بیان می شود و قابل کنترل است. به بیان دیگر اگر R^2 نزدیک به 1 باشد، مبین این است که با داشتن مقادیر x، می توان y را به وسیله خط رگرسیون برازش شده بطور دقیق تر پیش بینی کرد.

$$SS_T = \sum_1^n (y_i - \bar{y})^2$$

قضیه کاکران: اگر Q_2, Q_1, Q سه مجموع مربعات باشند،

بطوریکه $Q = Q_1 + Q_2$، در این صورت درجه آزادی Q برابر است با مجموع درجه های آزادی Q_1 و Q_2.

جدول آنالیز واریانس(ANOVA) :

منبع تغییرات (پراکندگی)	SS	df	MS	F
پراکندگی بیان شده	$SS_R = \hat{\beta}_1^{\,2} \sum (x_i - \bar{x})^2$	1	$\frac{SS_R}{1}$	$\frac{MS_R}{MS_E}$
پراکندگی بیان نشده	SS_E	n-2	$\frac{SS_E}{n-2}$	
پراکندگی کل	SS_T	n-1		

مثال 8: جدول تجزیه واریانس را برای داده های زیر کامل کنید و بگوئید این رابطه خطی دارای چه دقتی است؟ (R^2)

X	3	2	1	4	2	6	8	$\bar{x} = 3.71$
Y	12	12	12	19	15	24	25	$\bar{y} = 17$

$$S_X = 2.5 \quad , \quad S_y = 5.72 \quad , \quad S_y^2 = 32.67 \quad , \quad R^2 = 0.87 \quad , \quad n = 7$$

جدول تجزیه واریانس با استفاده از R^2:

؟

مثال 9 جدول آنالیز واریانس، R^2 و خط رگرسیون را برای داده های زیر محاسبه و رسم کنید:

X قد	165	167	160	161	158	172	169	163	167	168
Y وزن	59	68	55	65	61	75	71	58	58	66

$$\bar{x} = 165 \quad , \quad S_x^2 = \frac{1}{(n-1)}\sum(x_i - \bar{x})^2 = 19.56$$

$$\bar{y} = 63.6 \ , \quad S_y^2 = 41.82$$

؟

یعنی این رابطه چندان منطقی به نظر نمی رسد.؟

جدول آنالیز واریانس:

؟

10.1 فاصله های اطمینان و آزمون فرض:

می دانیم:

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{S_{XX}}\right) \qquad , \quad S_{xx} = \sum(x_i - \bar{x})^2$$

از طرفی مستقل از $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$ می باشد،

آنگاه:

؟

بنابراین فاصله اطمینان (1-α) % برای β_1 خواهد بود.(Confidence Interval)

$$\left[\hat{\beta}_1 \mp \frac{S}{\sqrt{S_{xx}}} t_{(n-2),1-\frac{\alpha}{2}}\right]$$

مثال 10 جدول آنالیز واریانس، R^2 و خط رگرسیون را برای داده های زیر محاسبه و رسم کنید:

X	38	35	80	71	99	69	94	73	86	88	87	77	89	67	65	60	31	36	76 53
Y	40	33	78	66	96	61	102	70	93	75	82	79	94	66	65	64	34	39	83

$\bar{x} = 1374/20$, $S_x^2 = \frac{1}{(n-1)}\sum(x_i - \bar{x})^2 = 429.38$

, $S_{X,Y}$ =-93363.3 $\bar{y} = 1379/20$, $S_y^2 = 425.63$

11.1 آزمون فرضها

فرض کنید β_{10} مقداری داده شده باشد آنگاه:

$$\begin{cases} H_0: \beta_1 = \beta_{10} \\ H_1: \beta_1 \neq \beta_{10} \end{cases}$$

در این صورت تحت فرض H_0 آماره آزمون t :

؟

که اگر $|t| > t_{(n-2),\frac{\alpha}{2}}$ آنگاه فرض H_0 رد می شود.

؟

به همین ترتیب برای $\hat{\beta}_0$ داریم:

$$\hat{\beta}_0 \sim N(\beta_0, \quad \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right))$$

بنابراین $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$ و

؟

بنابراین فاصله اطمینان (1-α)% برای β_0 عبارتست از: $CI_{1-\alpha}(\beta_0)$

$$\left[\hat{\beta}_0 \mp t_{\frac{\alpha}{2},\ n-2} \ .\ S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right]$$

آزمون فرض برای $\beta_0 = \beta_{00}$:

؟

آنگاه آماره آزمون تحت فرض H_0 عبارت است از:

؟

اگر $t_{\frac{\propto}{2},n-2}$ مقدار آماره در سطح α باشد آنگاه اگر

؟

12.1 ضریب همبستگی خطی پیرسون

این ضریب همبستگی خطی X و Y را با ρ نشان می دهیم و به صورت زیر تعریف می شود.

$$\rho_{X,Y} = \frac{cov(x,y)}{\sigma_x . \sigma_y}$$

و آنرا به صورت زیر بآورد می کنیم:

؟

و براحتی ثابت می شود $-1 \leq R^* \leq 1$ ؟

تذکر5: ضریب همبستگی با n یا n-1 تعریف می شود.

اگر r =1 همبستگی خطی دقیق (کامل) و مثبت است.

اگر r = -1 همبستگی خطی دقیق(کامل) و منفی است.

اگر $0 < r < 1$ همبستگی خطی مثبت (ناقص).

اگر $-1 < r < 0$ همبستگی خطی منفی (ناقص).

اگر r =0 باشد آنگاه x ,y ناهمبسته خطی هستند.

تذکر6: اگر x ,y مستقل باشند آنگاه x ,y نا همبسته اند ولی عکس آن ممکن است درست نباشد. در توزیع نرمال؟

تمرین6: ثابت کنید $a)\ R^2 = r^2$

$b)\ r_{e,x} = 0$

$c)\ r_{e,y} = \sqrt{1 - R_{x,y}^2}$

تمرین 7- نشان دهید، اگر $X \to ax + b$ و $y \to cy + d$ تبدیل کنیم، ضریب همبستگی تغییر نمی کند.

13.1 توزیع ضریب همبستگی (R^*)

فیشر آماردان مشهور انگلیسی توزیع تقریبی R^* را به صورت زیر بدست آورد:

$$W = \frac{1}{2}\ln\left(\frac{1+\mathrm{R}}{1-\mathrm{R}}\right) = tanh^{-1}(\mathrm{R})$$

وی نشان داد برای (n>25) , W تقریبا دارای توزیع نرمال با

$$\mu_W = E(W) \triangleq \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = tanh^{-1}(\rho)$$

و

$\sigma^2{}_W = Var(W) \triangleq \frac{1}{n-3}$ است، در نتیجه داریم:

$$\frac{W-\mu_W}{\sigma_W} \approx N(0.1)$$

14.1 آزمون معنی داری ضریب همبستگی و فاصله اطمینان برای ρ:

$$\begin{cases} H_0: \rho = \rho_0 \\ H_1: \rho \neq \rho_0 \end{cases}$$

اگـر

$$|z| > z_{1-\frac{\alpha}{2}} \Rightarrow RH_o$$

مثال 11: برای یک نمونه تصادفی 103 تایی از (X, Y) داریم r=0.5 آزمونهای زیر را انجام دهید:

-1 $\begin{cases} H_0: \rho = 0.6 \\ H_1: \rho \neq 0.6 \end{cases}$

-2 $\begin{cases} H_0: \rho = 0.8 \\ H_1: \rho \neq 0.8 \end{cases}$

-3 $\begin{cases} H_0: \rho = 0.6 \\ H_1: \rho \geq 0.6 \end{cases}$

-4 $\begin{cases} H_0: \rho = 0.6 \\ H_1: \rho \leq 0.4 \end{cases}$

حل 1:

15.1 انحراف معیار $\hat{y}$:

اولاً داریم: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$

برای بدست آوردن $V(\hat{y})$ اولاً ثابت می کنیم که $\bar{y}$, $\hat{\beta}_1$ ناهمبسته هستند. بدین منظور فرض کنید a_i و c_i مقادیر ثابتی باشند و

؟

؟

با توجه به اینکه y_i, y_j ها ناهمبسته اند، داریم:

؟

؟

؟

خطای معیار برآورد شده $\widehat{y}_i$ عبارت است از:

؟

همانطور که ملاحظه می شود انحراف معیار $\hat{y}_i$ *در نقطه* $\bar{x}$ می نیموم می شود به عبارت دیگر هر چه از $\bar{x}$ دور می شویم پیش بینی ها بدتر میشوند. در نتیجه داریم:

؟

و $\frac{(n-2)S^2}{\delta^2} \sim \chi^2_{n-2}$ *مستقل از* Z_k *می باشند و داریم:*

؟

لذا می توان فاصله اطمینان (1-α)% *برای* $E(y|x = x_k)$ *را بدست آورد:*

؟

*این فا*صله اطمینان را می توان چنین تعبیر کرد:

اگر نمونه گیری از $\hat{y}_i$ *ها بارها (به تعداد زیاد) و به حجم نمونه* یکسان تکرار شود و یک مقدار ثابت برای x در نظر گرفته شود و خط برازش شده برای تعیین y مورد استفاده قرار گیرد. آنگاه 100 ×(1-α)% تمام فاصله های اطمینانی که برای $E(y|x = x_k)$ *بدست خواهند آمد در برگیرنده مقدار واقعی میانگین* y *ها در ازای* $x = x_k$ است.

مثال 11. در مثال3 صفحه 3 یک فاصله اطمینان 95 درصد برای β_1 بدست آورده و فرض $H_0: \beta_1 = 0$ را در مقابل $H_0: \beta_1 \neq 0$ آزمون کنید و $\propto = 0.05$.

$$S_{\widehat{\beta 1}} = \frac{S}{\sqrt{S_{xx}}} = ؟$$

$$CI_{0.95}\ (\beta_1) = ؟$$

$CI_{0.95}\ (\beta_1)$؟

آزمون فرض:

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

؟

از طرفی می توان این نتیجه را از فاصله اطمینان فوق گرفت؟

مثال 12. در مثال قبل، فرض $H_0: \beta_0 = 13$ را آزمون کنید:

$$S_{\widehat{\beta}_0} = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = ؟$$

$$CI_{0.95}(\beta_0) = ؟, \qquad t_{0.25,23} = 2.0690$$

$CI_{0.95}(\beta_0)$ =(?)

یا فرض H_0 را رد نمی کنیم.

از طرفی $t_0 = \frac{\widehat{\beta}_0 - \beta_{00}}{s\widehat{\beta}_0}$ مشاهده شده عبارتست از:

$t_0 = \frac{13.6230-13}{0.5816} = 1.0716$ چون $|t_0| < t_{1-\frac{\alpha}{2}}$ پس دلیلی بر رد فرض H_0 وجود ندارد.

مثال 13. با مراجعه به مثال 3 $CI_{0.95}(.)$ *برای* $E(y/x = 28.6)$ *حساب کنید؟*

15.1 انحراف معیار $\widehat{y}_i$:

اولاً داریم: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1(x - \bar{x})$

برای بدست آوردن $V(\widehat{y}_i)$ اولاً ثابت می کنیم که $\bar{y}$, $\hat{\beta}_1$ ناهمبسته هستند. بدین منظور فرض کنید a_i و c_i مقادیر ثابتی باشند و

$U=a_1y_1 + \cdots + a_ny_n$

$W=c_1y_1 + \cdots + c_ny_n$

$Cov(U,W)=cov(\sum a_iy_i, \sum c_iy_i) = \sum a_ic_iv(y_i) + \sum_{i\neq j} a_ic_j\, cov(y_i, y_j)$

با توجه به اینکه y_i, y_j ها ناهمبسته اند، داریم:

$$)= \sigma^2(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$$

خطای معیار برآورد شده $\widehat{y}_i$ ***عبارت است از:***

$$S_{\hat{y}_i} = S * (\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}})^{0.5}$$

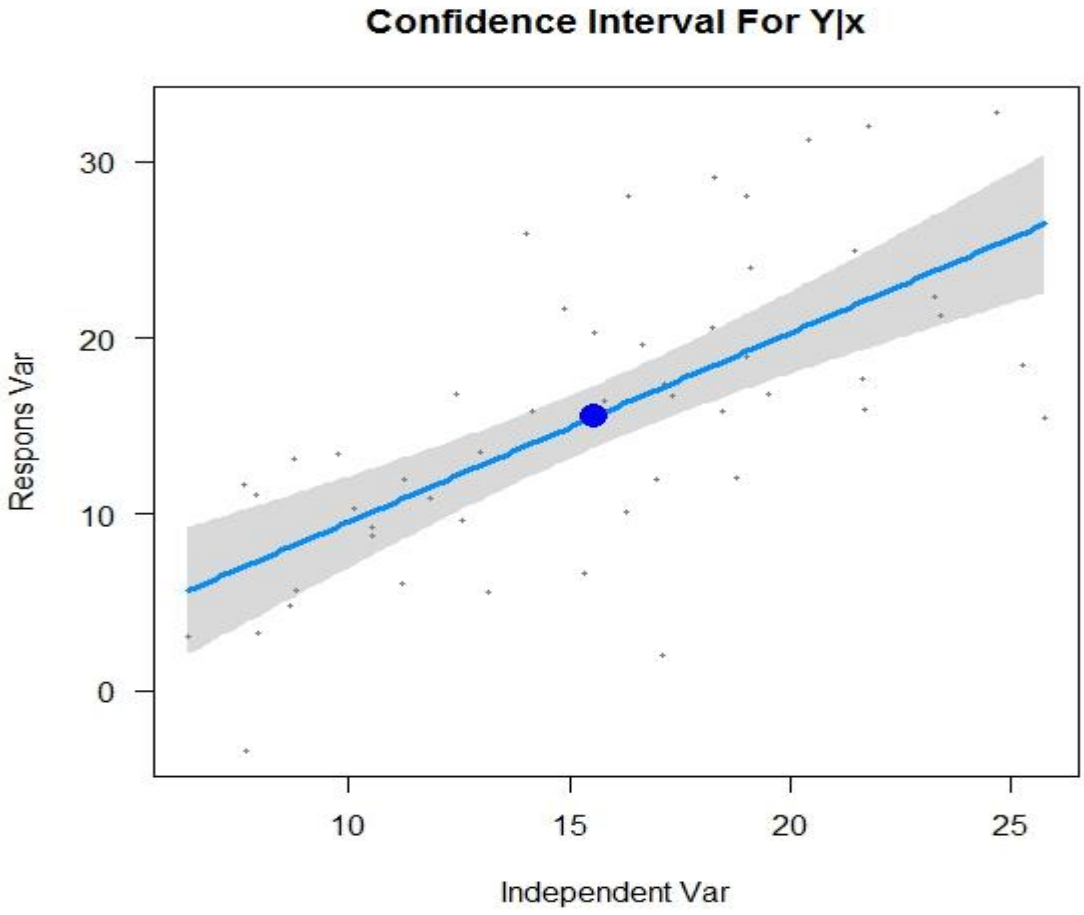
همانطور که ملاحظه می شود انحراف معیار $\hat{y}_i$ *در نقطه* $\bar{x}$ می نیمم می شود به عبارت دیگر هر چه از $\bar{x}$ دور می شویم پیش بینی ها بدتر میشوند. در نتیجه داریم:

لذا می توان فاصله اطمینان *%100(1-α)* برای $E(y|x = x_i)$ را بدست آورد.

$$CI(E(y|x = x_i)) = \left[\hat{y}_i \mp t_{\frac{\propto}{2},(n-2)} * s\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}}\right]$$

Example:

```
x<-runif(50,5,26)
y<-x+rnorm(50,0,6)
f<-lm(y~x)
install(visreg)
visreg(f,col="green",main="Confidence Interval For Y|x",xlab="Independent Var",
ylab="Respons Var")
points(mean(x),mean(y), pch=16, col="blue", cex=2)
```



این فاصله اطمینان را می توان چنین تعبیر کرد:

اگر نمونه گیری از $\hat{y}_i$ *ها بارها (به تعداد زیاد) و به حجم نمونه* یکسان تکرار شود و یک مقدار ثابت برای x در نظر گرفته شود و خط برازش شده برای تعیین y مورد استفاده قرار گیرد، آنگاه 100 ×(1-α)% تمام فاصله های اطمینانی که برای $E(y|x = x_i)$ *بدست خواهند آمد در برگیرنده مقدار واقعی میانگین* y *ها در ازای* $x = x_i$ است.

مثال 11. در مثال3 صفحه 3 یک فاصله اطمینان 95 درصد برای β_1 بدست آورده و فرض $H_0: \beta_1 = 0$ را در مقابل $H_0: \beta_1 \neq 0$ آزمون کنید و $\propto = 0.05$.

آزمون فرض:

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

آیا می توان این نتیجه را از فاصله اطمینان فوق گرفت؟

مثال 12. در مثال قبل، فرض $H_0: \beta_0 = 13$ را آزمون کنید:

$$S_{\widehat{\beta}_0} = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} = \sqrt{0.7926(\frac{1}{25} + \frac{(56.6)^2}{7154.42})} = 0.5816$$

$CI_{0.95}(\beta_0)$ =(12.4197 , 14.8263)

مثال 13. با مراجعه به مثال 3، $CI_{0.95}(.)$ *برای* $E(y/x = 28.6)$ *حساب کنید.*

Temperature: 100, 110,120,130,140,150,160,170,180,190

Performances: 45 51 54 61 66 70 74 78 85 89

16.1 پیش بینی یک مشاهده جدید:

از آنجائی که یکی از اهداف رگرسیون پیش بینی *forecasting* است، این بخش بسیار مهم می باشد. یعنی می خواهیم y را در نقطه ای غیر از داده های موجود (x_i) ها برآورد کنیم. بدین منظور فرض می کنیم کهx_k نقطه ای جدید و غیر از مشاهدات باشد و متغیر پاسخ جدید را y_{new} یا y می نامیم و فرض می کنیم که مدل رگرسیون مورد نظر برای مشاهده جدید مناسب است. آنگاه $\hat{y}_{new} = \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ برآورد نقطه ای y خواهد بود. در ضمن y (متغیر تصادفی) مستقل از y_i ها و $\hat{y}_k$ خواهد بود و دارای توزیع نرمال با واریانس σ^2 می باشد.

بنابراین برای بدست آوردن فاصله اطمینان *%100(1-α)* برای امید ریاضی متغیر پاسخ $y|x = x_k$ داریم:

$$y \sim N(\beta_0 + \beta_1 x_k, \delta^2)$$

و

بنابراین فاصله اطمینان $(1-\alpha)$ 100 درصد برای y خواهد بود:

$$CI_{1-\propto}(y) = \left[\hat{y}_k \mp t_{\frac{\propto}{2}}, (n-2)\sqrt{1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{S_{xx}}}\right]$$

اگر مایل باشیم که میانگین m مشاهده جدید را برای یک سطح داده شده ای از متغیر تصادفی پیش بینی کنیم و میانگین مقادیر m مشاهده جدید را با $\bar{y}_{k\,(new)}$ نشان دهیم آنگاه.

$$\bar{y}_{k\,(new)} \sim N(\beta_0+\beta_1 x_k \;\; , \; V(\bar{y}_{k\,(new)}))$$

$$\bar{y}_k \sim N(\beta_0+\beta_1 x_k \;\; , \; V(\bar{y}_k))$$

$$Z = \frac{w_m}{\sqrt{V(w_m)}} \sim N(0,1)$$

و چون مستقل از متغیر تصادفی $\chi = \frac{(n-2)S^2}{\delta^2} \sim \chi^2_{(n-2)}$ می باشد، داریم.

بنابراین فاصله اطمینان (1-α) درصد برای $\bar{y}_{k\,(new)}$ عبارتست از:

$$C_{1-\propto} I\left(\bar{y}_{k\,(new)}\right) = \left[\hat{y}_k \mp t_{\frac{\propto}{2},(n-2)}\, S\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{S_{xx}}}\right]$$

16.1 ارزیابی کیفیت مدل رگرسیون:

مقدمه:

a) فرض کنید $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$ اگر فرض H_0 پذیرفته شود یعنی اینکه متغیر x_iها اثری روی y_iها ندارد. و اگر H_0 رد شود یعنی اینکه مدل مورد نظر معنی دار است.

b) **ارزیابی شرایط در نظر گرفته شده در مدل رگرسیون: (شرایط ایده آل)**

شرایط عبارتند از:

H_1: $E(\varepsilon_i) = 0$	(نشان دهنده خطی بودن مدل)
H_2: $v(\varepsilon_i) = \sigma^2$	(واریانس ثابت است)
H_3: $cov(\varepsilon_i, \varepsilon_j) = 0 \quad , i \neq j$	ناهمبسته بودن خطاها
H_4: $\varepsilon_i \sim N(0, \sigma^2)$	توزیع خطاها نرمال باشند

اگر تمام شرایط فوق برقرار نباشند آنگاه مدل خطی ساده فوق مناسب نخواهد بود و به عنوان مثال می بایست از تبدیل داده ها استفاده کرد یا مدل غیر خطی دیگری بر آن برازش داد.

17.1 انواع باقیمانده ها:

همانطور که ملاحظه کردید شرایط فرض شده روی مدل خطی رگرسیون، روی باقی مانده ها (خطاها) بنا شده اند و در نتیجه خطاها و بررسی آنها از اهمیت خاصی برخوردار هستند. در اینجا برآورد گرهای ε_i ها یعنی $\hat{\varepsilon}_i$ باقی مانده ها نامیده می شوند.

دو نوع باقی مانده قابل تشخیص هستند:

1. (برآورد) باقی مانده های عادی
2. (برآورد) باقی مانده های استودنت شده (studentized residuals)

1.17.1 باقی مانده های عادی عبارتند از: $\hat{\varepsilon}_i = y_i - \hat{y}_i$ و داریم:

2.17.1 باقی مانده های استودنت شده که r_i نامیده می شوند:

این روش برای بررسی نا هم واریانسی یا واریانس غیرثابت باقی مانده های $r_i = \frac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}$

بکار می رود. (heterogeneous ناهمگن)

که $r_i \sim N(0, {\sigma^2}/{S^2})$ پیروی می کند. یعنی r_iها دارای توزیع نرمال تعدیل شده می باشند.

18.1. بررسی خطی بودن

برای سنجش این موضوع، 3 معیار نموداری موجود است که خطی بودن یا نبودن مدل را بررسی می کند.

1. $Scatter\ plot$ نمودار پراکنش داده ها (x_i, y_i) اطراف یک خط پراکنده شده باشند.
2. نمودار پراکنش $\hat{\varepsilon}_i$ ها برحسب x_iها حول محور o پراکنده شده باشند.
3. نمودار پراکنش $\hat{\varepsilon}_i$ ها یا r_iها برحسب y_iها حول o و بدون روند پراکنده شده باشند.

برای بررسی خطی بودن مدل رگرسیون بررسی 1و2 یا 1و3 لازم است.

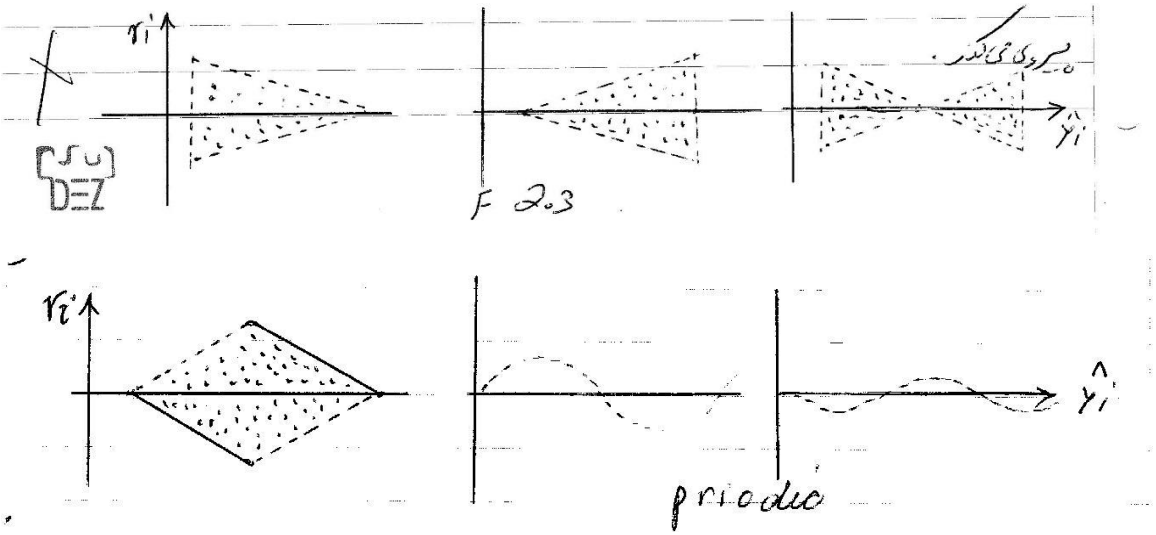
چنانچه خطی بودن مدل تأئید نشود آنگاه باید روش تبدیل داه ها را مورد بررسی قرار گیرد.

تذکر: برای حذف واریانس غیر همگن (r_iها) میتوان از رگرسیون موزون یا تبدیل y_i ها استفاده کرد.

19.1 بررسی همگن بودن واریانس (homogeneity)

در این بخش نا همگن بودن واریانس باقیمانده ها را بررسی می کنیم. (واریانس برحسب y_iها کم یا زیاد می شوند)

نمودار باقیمانده استودنت شده r_iها علاوه بر اینکه خطی بودن مدل را نشان می داد همگن بودن یا نبودن واریانس باقیمانده ها(r_iها) را نیز نشان میدهد. در حالت نرمال (ثابت بودن واریانس r_iها) داریم: $-3 \leq r_i \leq 3$ اگر $|r_i|$ خارج از این فاصله باشند نشان دهنده غیرنرمال بودن داده ها یا داده های پرت (extreme values) می باشند.



20.1 بررسی استقلال و نرمال بودن داده ها

در حالت کلی آزمون کردن استقلال داده ها مشکل است. اگر داده ها تابعی از زمان باشند (سری زمانی) میتوان نمودار داده ها (پراکنش داده ها) را که بصورت تابعی از اندیس داده هاست، رسم کرد. اگر این نمودار (پراکنش داده ها) هیچ رابطه ای را نشان ندهد آنگاه فرض می شود که استقلال داده ها محفوظ است. در غیر این صورت ممکن است داده ها خود همبسته باشند (Auto correlation) یا هم بستگی های دیگری بین داده ها موجود باشند.

1.20.1 آزمون تصادفی بودن یک دنباله از داده ها:

آزمون والد- ولفویتز (Wald - Wolfwitz) ، فرض کنید دنباله x_n , ... ,x_1 موجود باشد. اعضاء این دنباله را با میانه (میانگین) مقایسه می کنیم که در آن n_1 تعداد مثبت ها و n_2 تعداد منفی ها و R تعداد تغییرات مثبت و منفی باشند. اگر $n_1, n_2 > 10$ آنگاه:

$$Z = \frac{R-\mu}{\sigma} \sim N(0,1); \quad \mu = \frac{2n_1n_2}{n_1+n_2} + 1, \qquad \sigma^2 = \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}$$

یا

$$\sigma^2 = \frac{(\mu-1)(\mu-2)}{(n-1)}$$

می توان اصلاح پیوستگی را به صورت زیر انجام داد:

$$Z = \frac{R - \mu \pm 0.5}{\sigma}$$

بنابراین اگر $|Z_o| > Z_{(1-\alpha/2)}$، آنگاه فرض H_0: independent را رد می کنیم.

تذکر8: اگر $n_1 < n_2 < 10$ باشد، بدین منظور از جدول مربوطه استفاده می کنیم.

مثال:

14 1 23 10 3 6 7 12 21 11 16 8 17 22 24 15 13 25 9 18 2 4 5 20 19

N1=14, n2=11, R=13, mean=13.32, var=2.6488, Zo=-0. 1966 so the sequence is independent

2.21.1 بررسی **نرمال بودن داده ها:**

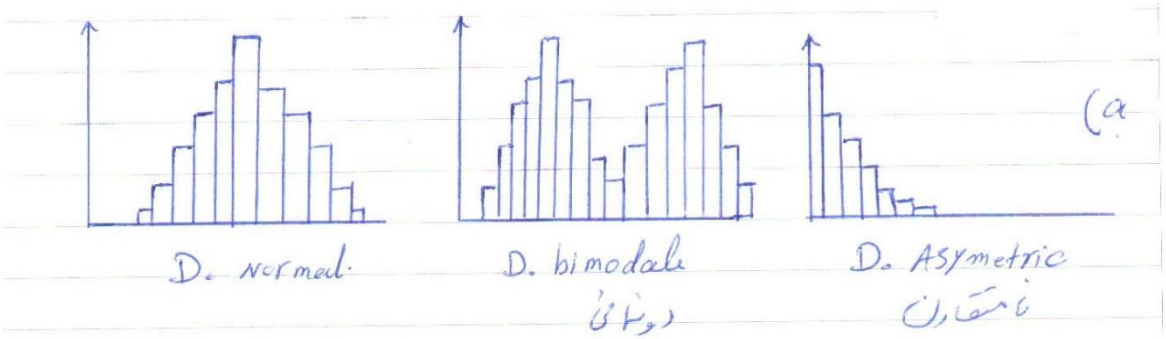
این معیار سنجش خیلی حساسcritical نیست (مقایسه نموداری).

با این حال اگر داده ها نرمال نباشند تمام فاصله های اطمینان و آزمونها (در سطح معنی داری) مشاهده شده اریب هستند. بدین منظور از روش نمودار فراوانی داده ها و آزمون های مربوطه استفاده می کنیم:

1. نمودار فراوانی داده ها (باقیمانده ها).
2. نمودار شاخه – برگ (Stem- leaf)
3. نمودار جعبه ای (Box- plot)
4. QQ_Plot
5. خط راست هنری line- Henry
6. آزمون شپیرو- ویلک (Shapiro- Wilk test)
7. آزمون کلموگرو – اسمیرنو (Kolmogrov – Smirnov Test)

این آزمونها بر روی باقیمانده های studentized بنا شده اند (r_iها)، مزیت این روش این است که حتی اگر r_iها وابسته (همبسته) باشند واریانس ارائه شده ثابت خواهد بود.

1. نمودار فراوانی داده ها (باقیمانده ها).



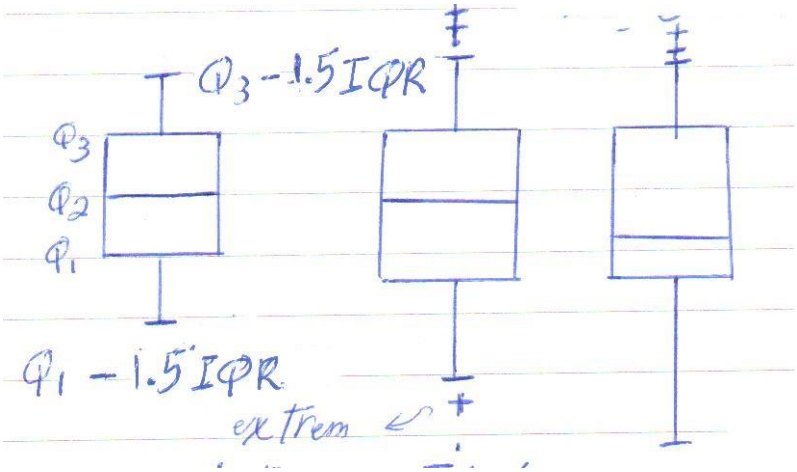
2. نمودار شاخه و برگ Stem and leaf plot

1,21,31,45,4,19,17,25,36,19,26,27,3,7

(*stem*)شاخه	(*leaf*) برگ
0	1 4
1	*7 9 9*
2	1 5 6 7
3	1 6 7
4	5

در این فرم داده های پرت *(extreme values)* مشخص می شوند.

3. نمودار *moustach* برای باقیمانده ها



4. *QQ- plot*

این نمودارها بر روی z_i ها ساخته می شوند.(*QQ- plot*)

5. برای خط *Henry*:

(*a*) باقیمانده ها را مرتب می کنیم.

(*b*) امید ریاضی آماره ترتیبی نرمال $(z_{(i)})$ را با فرمول زیر بدست می آوریم:

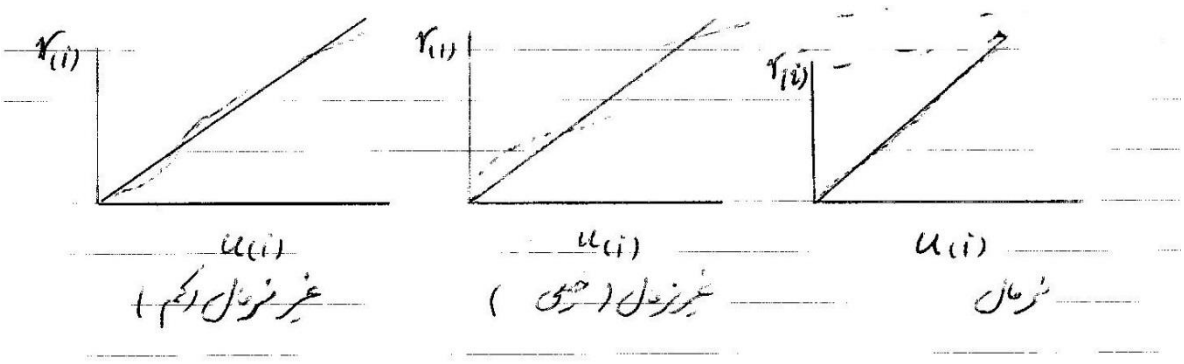
$$u_{(i)} = E(z_{(i)}) = \int_{-\infty}^{\infty} z\, n\binom{n-1}{i-1}\phi^{i-1}(z)\varphi(z)dz$$

بطوریکه $\varphi(z) = \frac{1}{\sqrt{2\pi}} exp\left(-z^2/2\right)$, $\phi(z) = \int_{-\infty}^{z} \Phi(t)dt$

چنانچه n به اندازه کافی بزرگ باشد، میتوان $E(z_{(i)}) = \phi^{-1}(\frac{i-\frac{3}{8}}{n-\frac{1}{4}})$ بدست آورد.

(c) نمودار $r_{(i)}$ها را برحسب $u_{(i)}$ها رسم می کنیم. اگر باقیمانده ها r_i(ها) نرمال باشند آنگاه $r_{(i)}$ باید روی خط راست قرار گیرند.

اما اگر $r_{(i)}$ها یک منحنی محدب (convex) یا مقعر (concave) تشکیل دهند آنگاه باقیمانده ها متقارن نیستند.



6. آزمون Shapiro-Wilk:

همانطور که ملاحظه می شود تغییر یا نتیجه گیری از روی نمودارها خیلی آسان نیست.

برای تشخیص نرمال بودن باقیمانده ها، آزمون های عددی Shapiro-Wilk و Kolmogorov-Smironov را ارائه می کنیم:

آماره آزمون sw عبارت است از:

$$w = \frac{(\sum_{i=1}^{n} a_i r_i)^2}{\sum_{1}^{n}(r_i - \bar{r})^2}$$

بطوریکه

$$a_i = \frac{\bar{z}^T v^{-1}}{\sqrt{\bar{z}^T v^{-1} \bar{z}}}$$

و

$$\bar{z} = (E(z_{(1)}), \dots, E(z_{(n)}))$$

و V ماتریس واریانس- کواریانس $z_{(i)}$ ها می باشد.

آماره W تقریباً معادل ضریب تعیین بین متغیرهای $r_{(i)}$, $u_{(i)}$ می باشد.

و نیز آماره W و خط راست *Henry* به هم مربوط هستند بطوریکه اگر مقدار W نزدیک 1 باشد، آنگاه نمودار خط *Henry* نشان دهنده اینست که داده ها روی خط می باشند. مزیت آزمون sw این است که به ما اجازه می دهد که یک آزمون تحت فرض اولیه $H_o: w = 1$ انجام دهیم. نرم افزارهای آماری یک سطح معنی داری *p-value* برای این آزمون ارائه می دهند.

با وجود این آزمون sw در مورد نامتقارن بودن یا مقادیر پرت (*extrems Values*) چیزی به ما نخواهد گفت (برخلاف روشهای گرافیکی).

7. آزمون *ks: Kolmogorov – smirnov*

در حالت کلی این آزمون *k.s* اجازه می دهد که یک نمونه داده شده را که از توزیع *F* (هرچه باشد*F*) پیروی می کند بررسی می کنیم. فرض کنید تحت فرض $H_o: F = F_o$

این آزمون k.s بر اساس فاصله بین F_0, F بنا شده است اگر F_n تابع توزیع تجربی از F باشد بطوری که

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x) \ , \ x \in IR$$

$$n \to \infty \Rightarrow F_n(x) \to F_0(x) \qquad \text{وقتی}$$

در بحث رگرسیون فرض بر این است که نمونه مورد نظر از توزیع نرمال پیروی کند $H_0 = F_0$(نرمال) ، یعنی فرض می شود که F_0 توزیع نرمال باشد و آماره آزمون عبارت است از:

$$D_n = sup_x|F_n(x) - F_0(x)|$$

بنابراین تحت فرض H_0 ، D_n نزدیک صفر خواهد بود. در این مورد نیز نرم افزارهای آماری یک سطح معنی داری p-value برای ks می دهند. اگر چه این آزمون زیاد استفاده می شود اما نسبت به sw از توان کمتری بهره مند است و به جای آن آزمون sw ترجیح داده می شود.

22.1 تبدیل داده ها :

در حالتی که شرایط چهارگانه رگرسیون نقض شود یعنی رگرسیون خطی، مناسب نباشد ممکن است تبدیل داده ها راهی مناسب برای استفاده از رگرسیون خطی باشد. اغلب یک تابع مثل $g_{(.)}$ موجود است که تبدیل داده ها را برای حل مسئله غیرخطی، ناهمگنی واریانس یا عدم نرمال بودن بکار رود.

در این بخش، 3 روش برای انتخاب روش تبدیل استفاده می شود. اولین روش وقتی استفاده می شود که شرط همگنی واریانس باقیمانده ها نقض شود، که در اینصورت شامل یافتن یک تقریب از واریانس متغیر تبدیل یافته می باشد.

دو روش دیگر یعنی روش Box-Cox و Mosteller - Tukey برای رفع مشکل غیرخطی بودن استفاده می شود. و نیز این روش برای حل مشکل عدم نرمال استفاده می شود.

1.22.1 واریانس تقریبی متغیر تصادفی تبدیل یافته:

مادامیکه نمودار پراکنش باقیمانده ها بر حسب $\hat{y}$ نشان دهنده تغییرات صعودی یا نزولی و... باشد، در این حالت (نامتجانس بودن واریانس) برای همگن کردن واریانس باقیمانده ها باید تبدیلی مناسب از y پیدا کرد.

حال واریانس تقریبی متغیر تبدیل را با استفاده از سری تیلر(مرتبه اول) بدست می آوریم.

$$g(y) = g(a) + g'(a)(x-a) + \cdots + g^{(n)}(a)\,(x-a)^n/n_i + \cdots$$

بنابراین:

$$Var(g(y)) \approx \{g'(E(y))\}^2 Var(y)$$

به عنوان مثال اگر y متغیر تصادفی باشد که در آن $\sigma_y = \mu^{\frac{1}{2}}$ ، داریم:

$$V(g(y)) \approx ?$$

اگر $g(y) = \sqrt{y}$ آنگاه $g'(\mu) = \frac{1}{2\sqrt{\mu}}$

چنانچه متغیر تصادفی y مثبت باشد آنگاه تبدیلات رایج را که در جدول زیر گرد آورده ایم ملاحظه کنید.

جدول 1.6.2 برای حل نامتناجنس بودن واریانس

موقعیت position	تبدیل

$\sqrt{y}$	داده هایی از نوع poisson $v(\varepsilon_i) \propto E(y_i)$
$\log(y)$	وقتی که y بزرگ و گسترده باشند، خیلی مؤثر است. $v(\varepsilon_i) \propto (E(y_i))^2$
$\frac{1}{y}$	$v(\varepsilon_i) \propto [E(y_i)]^4$
$arc\sin(\sqrt{y})$	$y \in [0,1], v(\varepsilon_i) \propto E(y_i)\big(1 - E(y_i)\big)$ $y \sim Benulli$

2.22.1 روش Mosteller and Tukey

پس از بدست آوردن تابع $g_{(.)}$ مناسب، اولین گام این است که بسادگی نمودار پراکنش ازواج $(x_i, g(y_i))$ را برای توابعی $g_{(.)}$ (چندین تابع) رسم کنیم تا اینکه بهترین نمودار خطی ممکن را پیدا کنیم. در واقع روشهای Mosteller و Tukey روشهای گرافیکی هستند که به سادگی نشان می دهند کدام تابع $g_{(.)}$ مناسب است.

3.21.1 روش Box – Cox

ایده این روش توسط George Box و David cox در سال 1964 ثبت شد. که شامل بسط یا گسترش مدل رگرسیون با معرفی یک پارامتر اضافه (کمکی) مثل λ می باشد.

$$g(y_i, \lambda) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

که در آن λ پارامتر ناشناخته می باشد و $g_{(0)}$ به صورت زیر تعیین می شود.

$$g(y, \lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & , \quad \lambda \neq 0 \\ \log(y) & , \quad \lambda = 0 \end{cases}$$

و

حال با فرض نرمال بودن باقیمانده های جدید، به روش درستنمایی ماگزیمم پارامترهای $\beta_0, \beta_1, \sigma^2, \lambda$ را برآورد می کنیم. بنابراین فرض می شود که :

$$w_i(\lambda) = \beta_0 + \beta_1 x_i + \varepsilon_i \;\; ; \;\; w_i \sim N(\beta_0 + \beta_1 x_i . \sigma^2)$$

مثال 14:

X: 23.4 27.0 27.5 25.6 26.7 25.7 26.9 23.1 24.1 26.8 24.1 27.9 24.1 26.6 26.3

Y: 23.2 28.2 28.6 25.6 27.2 26.7 27.1 23.4 24.6 27.2 23.7 27.1 23.3 25.8 26.1

X: 29.7 28.7 20.5 22.9 27.6

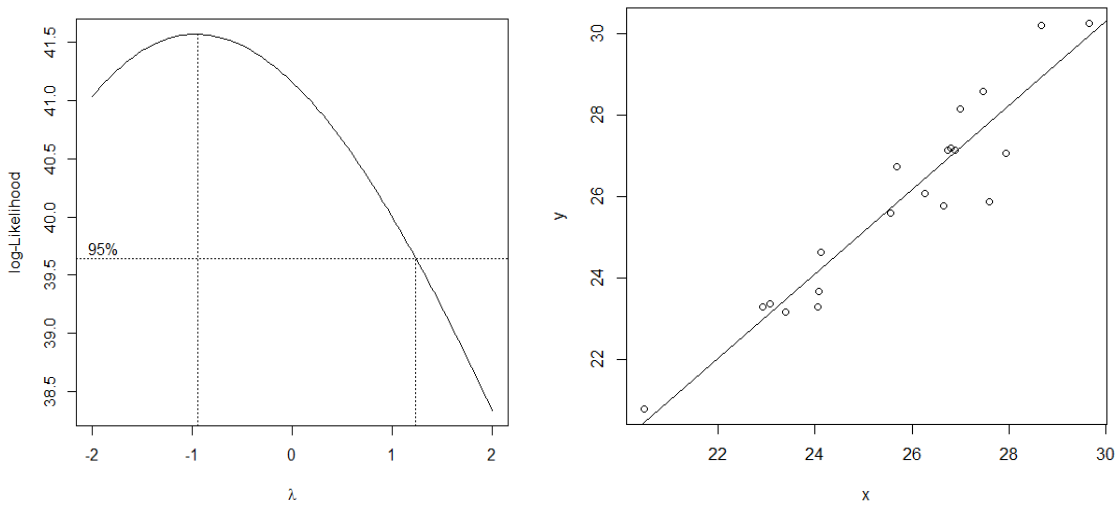
Y: 30.3 30.2 20.8 23.3 25.9

Sum(x)=515.05 , Sum(y)=520.51 ,sum(xy)= 13513.51, n=20, sd(x)=2.27, sd(y)=2.72,

1- مدل رگرسیون برازش شده قبل از تبدیل:

$$\hat{y} = -0.77 + 1.036x$$

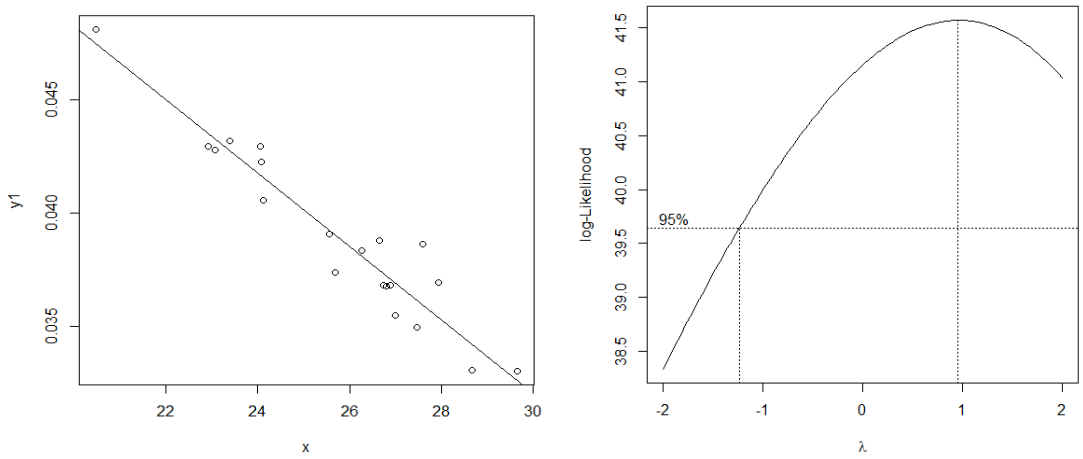
با دستور Boxcox(y~x)



مدل رگرسیون برازش شده بعد از تبدیل:

$$\hat{y}_1 = \ 0.081 - 0.0017x$$

با دستور Boxcox(Y_1~x)



23.1 پیش بینی تحت تبدیل:

باید از کاربرد متغیر تصادفی y مطمئن بود و از طرفی می دانیم در حالت کلی تساوی زیر برقرار نیست:

$$E(g(y)) \neq g(E(y))$$

و این درست نیست که به سادگی تبدیل معکوس را برای پیش بینی روی y (بدون تبدیل- اولیه) بکار ببریم. اما زمانی که $g_{(.)}$ تابع یکنوا (صعودی) است داریم:

$$p(a \leq g(y) \leq b) = p(\, g^{-1}(a) \leq y \leq g^{-1}(b))$$

مثال 15: اگر

$$g(y_i) = y^*{}_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \Rightarrow \quad g(y) = \sqrt{y} \;\Rightarrow y = y^{*2}$$

آنگاه:

$$C.I\left(\hat{y}^*{}_k\right) = (a\,, b) \Rightarrow C.I\,(\hat{y}_k) = (a^2, b^2)$$

مثال 16: برآورد پارامترها – بررسی شرایط (4 شرط) روی رگرسیون ماگزیموم غلظت concentration ازن Ozone$O3$: $\mu g|ml$

بر حسب درجه حرارت T_{12} و داده ها $max03 = \beta_0 + \beta_1 T_{12} + 4$

1- خطی بودن رابطه بین $T_{12}, 03$.
2- متجانس بودن (همگن بودن) واریانس.
3- استقلال (باقیمانده ها)
4- نرمال بودن باقیمانده ها

از آنجائیکه هدف از رگرسیون برآورد (پیش بینی است) پس باید نرمال بودن باقیمانده ها حفظ شود.

مثال 17: برای آزمون کامل یک مدل می توان فایل $Reg - complete$ را انجام داد.

MASS

Install(leaps)

library(leaps)

تمرین 8: فرض کنید نمونه های

$(x_1, y_1, \dots, (x_m, y_m)$ بطوری که $\begin{cases} \bar{x}, & s_x^2 \\ \bar{y}, & s_y^2 \end{cases}$

$$\begin{cases} \bar{x}', & s_{x'}^2 \\ \bar{y}', & s_{y'}^2 \end{cases} \quad , \quad (x_1', y_1'), \dots, (x_n', y_n')$$

داده های رگرسیونی باشند. با فرض رگرسیون خطی ساده ، پارامترها ی رگرسیون خطی ساده دو نمونه تلفیق شده را بدست آورید.

میانگین دو نمونه تلفیق شده : $\bar{x}'' = \dfrac{m\bar{x} + n\bar{x}'}{m+n}$

$$\bar{y}'' = \frac{m\bar{y} + n\bar{y}'}{m+n}$$

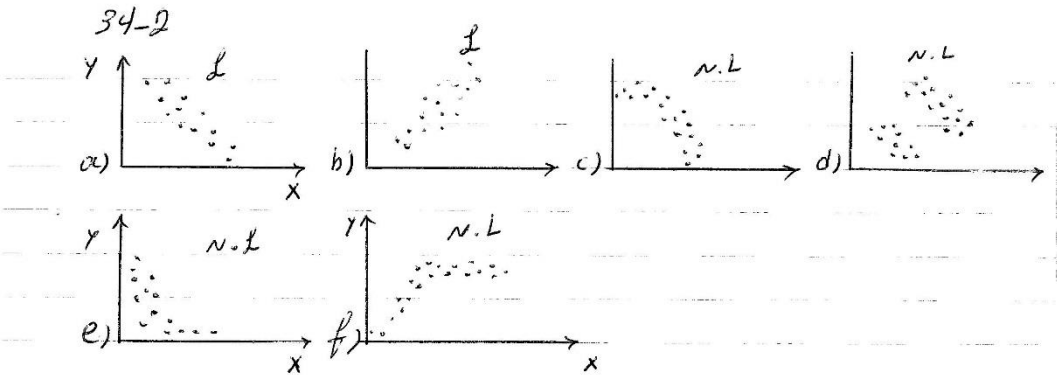
$$s_3^2 = \text{واریانس تلفیق شده} = \frac{1}{n}\left[ms_x{}^2 \times + ns_{x'}^2 + m\bar{x}^2 + n\bar{x}'^2\right] - \left[(m\bar{x} + n\bar{x}')/N\right]^2$$

کواریانس دو نمونه تلفیق شده :

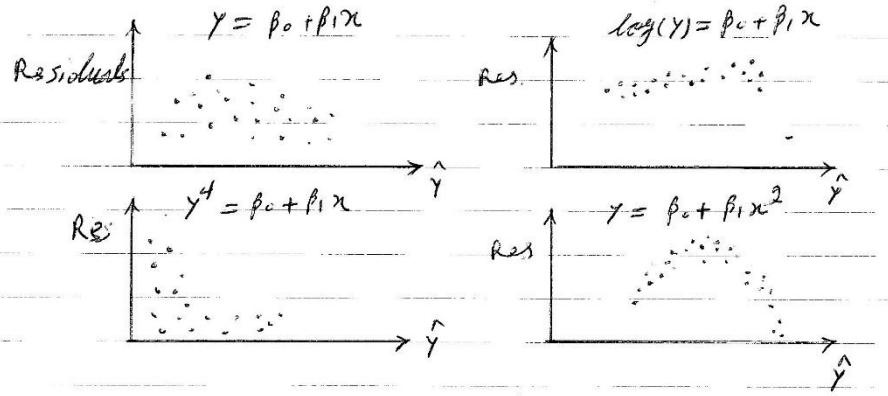
$$S_{xy} = \frac{1}{N}\sum x_i y_i - \bar{x}\bar{y}$$

$$= \frac{1}{N}\left\{\left[\left(m\, S_{xy} + m\bar{x}\bar{y}\right) + \left(n\, S_{x'y'} + n\bar{x}'\,\bar{y}'\right)\right] - (m\bar{x} + n\bar{x}') \times \left(m\bar{y} + n\bar{y}'\right)\right\}$$

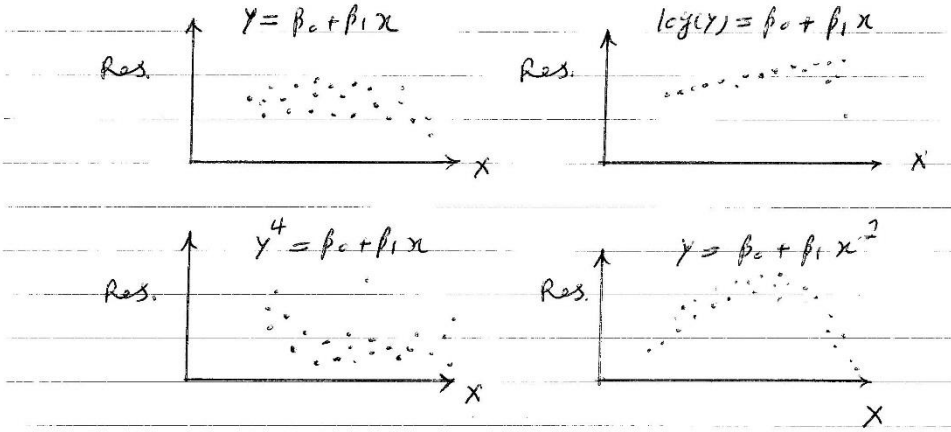
$$;\ N = m + n$$


34-2
Y
X
a)
b)
c)
d)
e)
f)
N.L

(2.2.1)


Residuals
$Y = \beta_0 + \beta_1 x$
$\log(Y) = \beta_0 + \beta_1 x$
$Y^4 = \beta_0 + \beta_1 x$
$Y = \beta_0 + \beta_1 x^2$
Res.
$\hat{Y}$

(2.2.2)


Res.
$Y = \beta_0 + \beta_1 x$
$\log(Y) = \beta_0 + \beta_1 x$
$Y^4 = \beta_0 + \beta_1 x$
$Y = \beta_0 + \beta_1 x^2$
X

(2.2.3)

فصل دوم : مدلهای غیر خطی

1.2 مختصری در مورد مدلهای غیر خطی:

مدلهای غیر خطی کدامند؟

چرا ترجیح می دهیم از مدلهای خطی استفاده کنیم؟

فرض کنید $T(x)$ یک تابع غیر خطی برحسب x با چند پارامتر مجهول باشد. عموما مدلها به صورت زیر فرض می شوند:

$$Y|x=T(x)+E$$

یا

$$Y|x=T(x)*E \quad ; \quad E > 0$$

که در اولی خطا جمعی و در دومی خطا ضربی خواهد بود. بر خلاف مدلهای خطی، در مدلهای غیر خطی یک قانون کلی برای برآورد پارامترها نمی توان بیان کرد. یعنی هر مورد را باید جدا گانه بررسی نمود، و یا در صورت امکان با یک تبدیل مناسب آن را به صورت مدل خطی درآورد. بدین منظور مثالهای زیر را مورد بررسی قرار می دهیم:

مثال 20. اگر $Y|x=T(x)=a+bx^2+E$

مثال 21. $Y|x=T(x)=a*x^b*E \quad ; \ a, b, E > 0$

در این صورت برای برآورد پارامترها، به صورت زیر عمل می کنیم:

حال با فرض : $z = \log(Y)\,,\ \beta_0 = \log(a)\,,\ \beta_1 = \log(b), U = \log(E)\,,$

خواهیم داشت: $Z = \beta_0 + \beta_1 x + U.$

مثال 22. فرض کنید،

$$Y|x = T(x) = \frac{1}{1 + a * \exp(bx + E)}$$

برای برآورد پارامترها چنین عمل می کنیم:

$$\log\left(\frac{1}{Y}-1\right)=\log(a)+bx+E$$

مثال 23. فرض کنید،

$$Y|x=T(x)=\frac{\exp(\beta_0+\beta_1 x+E)}{1+\exp(\beta_0+\beta_1 x+E)}$$

مدل خطی آن را بدست آورید:

$$Y'=\ln(\frac{1-Y}{Y})=\beta_0+\beta_1 x+E$$

مثال 24. فرض کنید: $Y|x=T(x)=a+b*exp(-cx)+E$،

این مدل را نمی توان به مدل خطی تبدیل کرد، چنانچه به خواهیم به روش کمترین توانهای دوم، پارامترها را براورد کنیم، $\sum_{i=1}^{n}{E_i}^2=\sum_{i=1}^{n}(Y_i-a-bexp(-cx))^2$

ملاحظه می شود که حل آن منوط به حل دستگاه معادلات سه مجهولی، غیر خطی، پیچیده وعموما روشهای عددی می باشد.

مثال 25.

$$\frac{1}{y_i}=\beta_0+\beta_1(1/x_i)+\varepsilon_i$$

فصل سوم : رگرسیون خطی چند متغیره:

1.3. مقدمه

در این بخش برای بررسی و تجزیه و تحلیل رگرسیون ازماتریسها و روابط ماتریسی و جبر خطی استفاده می شود. بدین منظور یادآوری جبر ماتریسی را بطور خلاصه لازم می دانیم. قبل از پرداختن به رگرسیون خطی چندگانه، ابتدا رگرسیون خطی ساده را بطریق ماتریسی مورد بررسی قرار می دهیم.

روابط ماتریسی:

فرض کنید P, N, M ماتریسهای دلخواه، B, A ماتریسهای مربع باشند و v یک بردار ستونی باشد، بنا براین روابط زیر را داریم:

1. $(M')' = M, (MN)' = N'M', N'+M'=(N+M)'$.
2. A متقارن است $\Leftrightarrow A' = A$;
3. AA' و $A'A$ متقارن هستند;
4. $P(M + N) = PM + PN$;
5. $A^{-1}A = I = AA^{-1}$;
6. A متقارن باشد $\Leftrightarrow$ A^{-1} متقارن است;
7. $(AB)^{-1} = B^{-1}A^{-1}$;
8. $|A'| = |A|, |A^{-1}| = 1/|A|$
9. $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B), \mathrm{tr}(AB) = \mathrm{tr}(BA)$
10. $(A - vv')^{-1} = A^{-1} + \frac{A^{-1}vv'A^{-1}}{1-v'A^{-1}v}$
11. اگرv یک بردار ویژه ماتریس A باشد و اگر λ مقداری ثابت و موجود باشد بطوریکه $Av = \lambda v$ **آنگاه** λ **مقدار ویژه ماتریس** A **است**.
12. برای بدست آوردن مقدار ویژه ماتریس A از معادله $|A - \lambda I| = 0$ استفاده می کنیم.
13. اگر A متقارن باشد تمام مقادیر ویژه A حقیقی هستند.
14. اگر A ماترسی مربع $n \times n$ و متقارن باشد , فرض کنید $\lambda_1, \ldots, \lambda_n$ مقادیرویژه و $v_1, \ldots, v_n$ بردارهای ویژه متناظرباشند. فرض کنید V با iامین ستون v_i، بنابراین

$$V'V = VV' = I \ (V \ \textit{is} \ \text{orthogonale}) \ , V'AV = diag(\lambda_1, \ldots, \lambda_n)$$

15. ماتریس A معین مثبت نامیده می شود اگر برای تمام مقادیر $v \neq 0$ ، $v'Av > 0$، ماتریس A نیمه معین مثبت نامیده می شود اگر برای تمام مقادیر $v \neq 0$، $v'Av \geq 0$.

تمرین 2-1 :

اگر $Var[X] = 10, Var[Y] = 15, Var[Z] = 20$,

$Cov(X, Y) = -2, Cov(Y, Z) = 5$ و X و Z مستقل هستند. موارد زیر را محاسبه کنید:

- آنگاه $Var[X + Z], Var[X - Y]$ et $Var[2X + 3Z - Y]$.
- $Cov(X + Y, 2X - 3Z)$.
- $\sum = Var\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$

با استفاده از خواص ماتریسها، ماتریس واریانس کواریانس $\sum$ را حساب کنید.

2.3. در مورد رگرسیون خطی ساده

از قبل داریم.

و می دانیم $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , i = 1, \dots, n}$

$$\boldsymbol{y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1}, \ x = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n\times 2}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}} \quad (1.2.E)$$

یا $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$

می توان نوشت:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_1} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x_n} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \end{pmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta_0 + \beta_1 x_1 + \varepsilon_1} \\ \vdots \\ \boldsymbol{\beta_0 + \beta_1 x_n + \varepsilon_n} \end{bmatrix} \quad ، \ (\mathbf{2.2.}\boldsymbol{E})$$

حال به راحتی میتوان نوشت: $y = x\beta + \varepsilon$ ، $(\mathbf{3.2}\boldsymbol{E})$

مثال 2-1: اگر حالت روانی افراد ($psychologiccal\ position$) را با y نشان دهیم و متغیرهای آزاد : جنس x_5, درآمد x_4 , قد x_3 , وزن x_2 و سن x_1 باشند. می توان یک رابطه خطی بین متغیرهای آزاد و متغیر وابسته نوشت.

همانطور که ملاحظه می کنید در رگرسیون چند متغیره (خطی) متغیرهای از نوع پیوسته، گسسته، دوگانه و کیفی (جنس) استفاده می شود.

3.3 مدل و نمادها

مدل رگرسیون خطی چندمتغیره:

فرض y متغیر وابسته و $x_p, \dots, x_1$ متغیرهای آزاد یا کمکی باشند.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad ، \ (\mathbf{4.2.}\boldsymbol{E})$$

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_{ip} + \varepsilon_i \quad , \quad i = 1, \dots, n$$

تذکر9: متغیرهای $x_1, \dots, x_p$ غیر تصادفی اند و متغیرy تصادفی است.

و $\varepsilon_i \sim N(0,\ \delta^2)$

و $\beta_p, \dots, \beta_0$ پارامترهای ثابت نامشخص هستند که باید برآورد شوند.

تذکر10: n تعداد مشاهدات باید از تعداد پارامترها بیشتر باشد یعنی $n > p + 1$.

مانند رگرسیون خطی ساده، شرایط لازم برای رگرسیون خطی چندگانه نیزعبارتند از:

$$H_1: \ E(\varepsilon_i) = 0 \ , \ i = 1, \dots, n$$

$$H_2:\ v(\varepsilon_i) = \delta^2$$

$$H_3:\ cov(\varepsilon_i, \varepsilon_j) = 0\ ,\ i \neq j$$

$$H_4:\ \varepsilon_i \sim N(0, \delta^2)$$

مثال 2-2:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i\ ,\ i = 1, \dots, n$$

که میتوان آن را به صورت:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i\ ,$$

مثال 2-3- مدل سه جمله ای$(polynomial\ cubic)$

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x^2{}_{1,i} + \beta_3 x^3{}_{1,i} + \varepsilon_i$$

که می توان به صورت زیر نوشت:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

فرض کنید $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i\ ,\ i = 1, \dots, n$

و شکل ماتریسی $(\mathbf{4.2.}\boldsymbol{E})$ آن عبارتست از:

$$\begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x_{11}}, \dots, \boldsymbol{x_{1p}} \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x_{n1}}, \dots, \boldsymbol{x_{np}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_0} \\ \vdots \\ \boldsymbol{\beta_p} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_p x_{1p}} + \boldsymbol{\varepsilon_1} \\ \vdots \\ \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \cdots + \boldsymbol{\beta_p x_{np}} + \boldsymbol{\varepsilon_n} \end{bmatrix}$$

$$y_{n\times 1} = x_{n\times p'} * \beta_{p'\times 1} + \varepsilon_{n\times 1} \quad , \quad p' = p + 1 \quad , \quad (\mathbf{5.2.}\boldsymbol{E})$$

$n = \text{observetion number}$

$y = \text{endogenes} \quad (n \times 1)$

$x = \text{exogenes} \quad (n \times p')$

$\beta = \text{parameters} \ (p' \times 1)$

$\varepsilon = \text{errors} \qquad (n \times 1)$

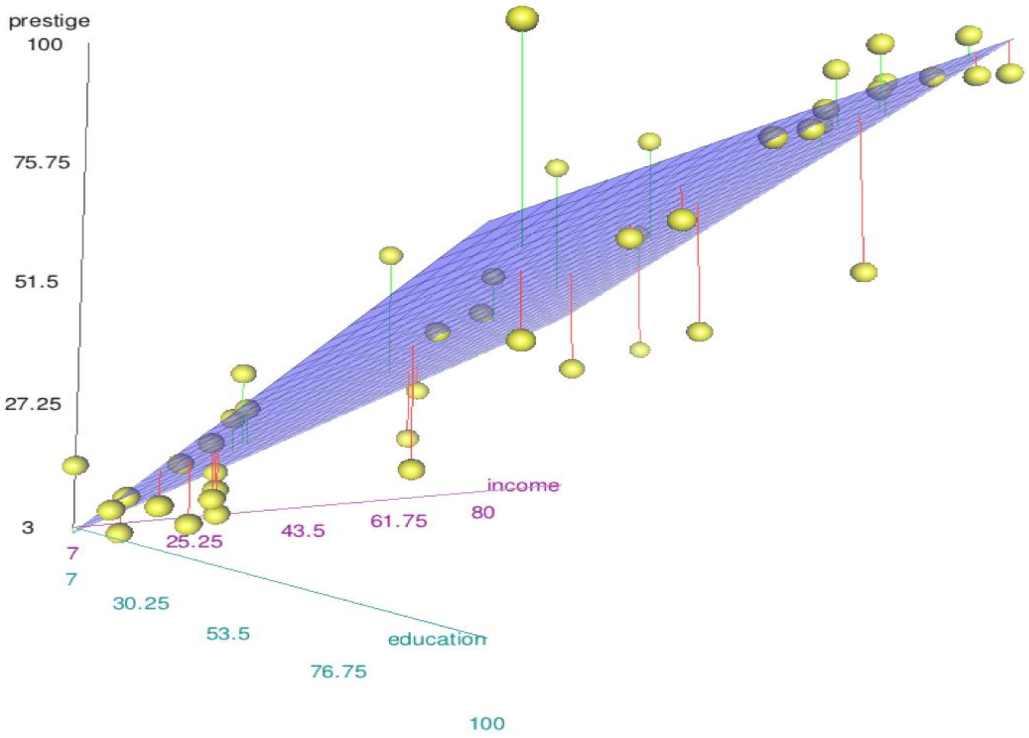
مثال 2-4. ادعای هاینز و همکاران (2005)، به عنوان یک راهنما در این فصل مورد استفاده قرارمی گیرد، شرح تجزیه و تحلیل، پیدا کردن یک مدل برای پیش بینی اندازه کبودی سیب در مرتب سازی و بسته بندی کارگاه، آزمایش های آزمایشگاهی نشان داده اند که قطر کبودی به میلی متر[*MM*] را می توان بر حسب ارتفاع سقوط و تراکم میوه (به [g/cm^3] مدل بندی کرد. جدول زیر این داده های نمونه را نشان می دهد.

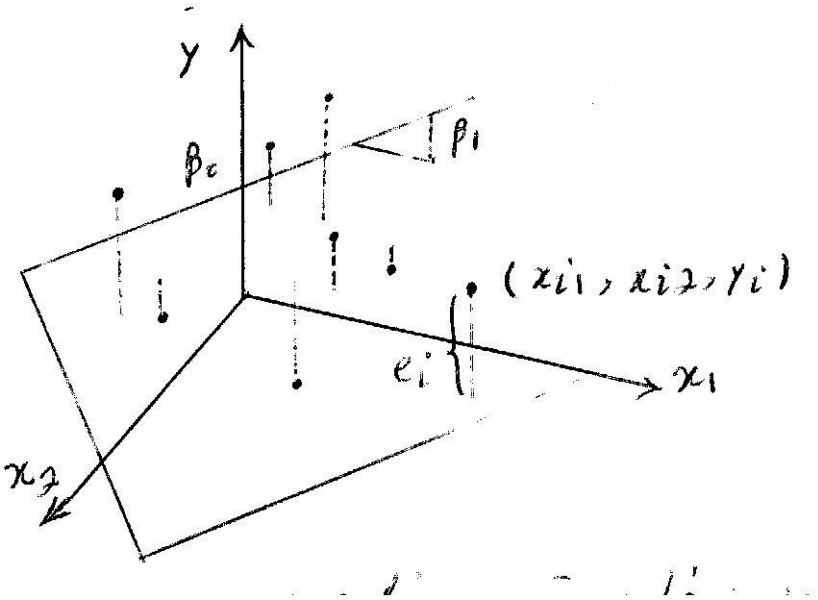
$number$	$bruising\ Diameter\ [mm]$	$Height\ of\ fall\ \ [mm]$	$density\ [^{g}/_{cm^3}]$
1	3.62	303.7	0.9
2	7.27	366.7	1.04
3	2.66	336.8	1.01
4	1.53	304.5	0.95
5	4.91	346.8	0.98
6	10.36	600.0	1.04

7	5.26	369.0	0.96
8	6.09	418.0	1.00
9	6.57	269.0	1.01
10	4.24	323.0	0.94
11	8.04	562.2	1.01
12	3.46	284.2	9.97
13	8.50	558.6	1.03
14	9.34	415.0	1.01
15	5.55	349.5	1.04
16	8.11	462.8	1.02
17	7.32	333.1	1.05
18	12.58	502.1	1.1
19	0.15	311.4	0.91
20	5.23	351.4	0.96

4.3. پارامترهای رگرسیون چند متغیره

1.4.3 تفسیر پارامترها





نمایش هندسی مدل رگرسیون خطی چندمتغیره (Geometric Graphic)

2.4.3 برآورد پارامترهای رگرسیون چند متغیره:

مانند رگرسیون خطی ساده، برای برآورد پارامترهای رگرسیون خطی چند متغیره از روش کمترین توانهای دوم خطا استفاده می شود. در نتیجه برای برآورد بردار $\underline{\boldsymbol{\beta}}$ باید

$$\sum_{i=1}^{n} \hat{e}_i^2$$

مجموع توانهای دوم خطا (باقیمانده ها) می نیمم شود.

با استفاده از روابط ماتریس داریم:

$(p+1)$ معادله و از طرفی داریم:

$$\frac{\partial f(\hat{\beta})}{\partial \beta} = -2x'y + 2x'x\hat{\beta} = 0$$

(6.2.E)

$$\Rightarrow x'x\hat{\beta} = x'y \Rightarrow \hat{\beta} = (x'x).x.y$$

در صورتیکه ماتریس $x'x$ معکوس پذیر باشد (full rank دارای رتبه کامل).

مثال 2-6- اگر P=1 باشد مطلوبست برآوردهای β توسط ماتریسها:

$$X = \begin{bmatrix} 1 & x_1 \\ : & : \\ . & . \\ 1 & x_n \end{bmatrix}, x' = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}$$

$$X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix}$$

$$\det(X'X) = n\sum {x_i}^2 - (n\bar{x})^2 = n(\sum {x_i}^2 - \bar{x})$$

$$= n\sum(x_i^2 - \bar{x}^2) = nS_{xx}$$

$$(X'X)^{-1} = \frac{1}{n\,S_{xx}}\begin{bmatrix} \sum {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}x'y = \frac{1}{n\,S_{xx}}\begin{bmatrix} \sum {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right)$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\,S_{xx}}\begin{bmatrix} \sum {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \frac{1}{n\,S_{xx}}\begin{bmatrix} n\bar{y}\sum x_i^2 & -n\bar{x}\sum x_i y_i \\ -n\bar{x}\sum y_i & +n\sum x_i y_i \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - n\bar{x}\bar{y}}{n\;S_{xx}} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{n\sum {x_i}^2 - n\bar{x}\sum x_i y_i}{n\;S_{xx}}$$

$$= \frac{\bar{y}\sum {x_i}^2 - ny\bar{x}^2 - (\bar{x}\sum x_i y_i - ny\bar{x}^2)}{S_{xx}}$$

$$= \frac{\bar{y}\sum(x_i - \bar{x})^2 - \bar{x}Sxy}{S_{xx}} = \bar{y}\frac{S_{xx}}{S_{xx}} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

همانطوریکه ملاحظه می شود دقیقا همان برآوردگرهای بدست آمده در رگرسیون خطی ساده هستند و بطور کلی داریم:

$$\hat{\beta}_0 = \bar{y} - \bar{x}_*'\hat{\beta}_*$$

E. 3.3.2.2

که در آن :

$$\begin{cases} \hat{\beta}' = (\hat{\beta}_1, \dots, \hat{\beta}_P) \\ \bar{x}_*' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_P) \end{cases}$$

مثال 2-7: رگرسیون خطی سه متغیره (مدل برازش شده را بدست آورید).

$$\hat{y} = -33.831 + 0.0134x_1 + 34.890x_2$$

X’ X=

		HF	Density
	20.00	7767.80	28.9300
HF	7767.80	3201645.78	10349.678
Density	28.93	10349.68	118.3677

X’ Diameter=

	120.7900
HF	51129.1690
Density	153.8365

XX inverse=

24, 63666 0, 005321 −26, 74679
0, 005321 0, 0000077 −0, 008353
−26, 74679 −0, 008353 30, 096389

B0=-33, 8310
B1=0, 0134
B2=34, 8900

3.4.3. توزیع برآورد گرها $\hat{\beta}$

برای بدست آوردن توزیع برآوردگرها، باید در نظر داشت که ماتریس باقیمانده ها ε دارای توزیع نرمال چند متغیره می باشد و n تعداد مشاهدات می باشد.

محاسبه امید و واریانس برآوردگرها براساس (خواص) توزیع نرمال چند متغیره می باشند.

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)' \sim N_n(\underline{O}, \sigma^2 I_n)$ که می باشد:

در نتیجه:

$$\underline{y} \sim N_n(X\beta, \sigma^2 I_n)$$

تعیین توزیع $\hat{\beta}$ مانند روش رگرسیون خطی ساده می باشد. یعنی نشان می دهیم که $\hat{\beta}$ ترکیبی خطی از y_i ها می باشند.

$$\hat{\beta} = (X'X)^{-1}X'y$$

بنابراین $\hat{\beta} \sim N_P\left(E(\hat{\beta}), V(\hat{\beta})\right)$

$$E(\hat{\beta}) = \beta$$

نکته:

$$V(AX) = A * V(X) * A'$$

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

$$\Rightarrow \hat{\beta} \sim N_P(\beta, \sigma^2(X'X)^{-1})$$

مثال 8-2 : واریانس و کوارایانس $\hat{\beta}$ وقتی که P=1 .

$$(X'X)^{-1} = \frac{1}{n\,S_{xx}}\begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$$

$$V(\hat{\beta}_1) = \frac{n\sigma^2}{n\,S_{xx}} = \sigma^2/S_{xx}$$

$$V(\beta_0) = \frac{\sigma^2 \sum x_i^2}{n\,S_{xx}} = \sigma^2 \frac{\sum x_i^2}{n\,S_{xx}} \pm \frac{\sigma^2\bar{x}^2}{S_{xx}} = \sigma^2(1 + \frac{\bar{x}^2}{S_{xx}})$$

$$COV(\hat{\beta}_1, \hat{\beta}_0) = \frac{n\bar{x}\,\sigma^2}{n\;S_{xx}} = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

تذکر11: علامت $\mathrm{COV}(\hat{\beta}_1, \hat{\beta}_0)$ مخالف علامت $\bar{x}$ است و مقدار کواریانس صفر است هر گاه $\bar{x} = 0$ باشد.

4.4.3 توزیع $\hat{y}$ و $\hat{\varepsilon}$

مدل رگرسیون خطی چند متغیره

$$y = X\beta + \varepsilon$$

با توجه به $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'y$$

آنگاه :

$$\hat{y} = X\hat{\beta} = X\,(X'X)^{-1}X'y = Hy \qquad (E.3.4.1)$$

که در آن H ماتریس تبدیل نامیده می شود.

$$H = X(X'X)^{-1}X'$$

ممکن است باقیمانده ها را بصورت تابعی از ماتریس H بیان کنیم.

$$\hat{\varepsilon} = \hat{\varepsilon} = (I_n - H)y\,.\,(E.3.4.2)$$

ماتریسH ممکن است مانند یک ماتریس تبدیل (خطی) عمل کند یعنی ماتریس H یا $I_n - H$ مثل یک عملگر خطی عمل می کند. خواص زیر برای این دو ماتریس به صورت زیر برقرار است:

1. $H' = H$ (H symetric) متقارن
2. $HH = H$ (idempotence) خودتوان
3. $HX = X$ همانی (identity)
4. $I_n - H = (I_n - H)'$ (symetric)
5. $(I_n - H)(I_n - H) = I_n - H$ خودتوان
6. $(I_n - H)X = X$

با استفاده از خواص ماتریس H می توان امید و واریانس $\hat{y}$ و $\hat{\varepsilon}$ را محاسبه کرد.

$$V(\hat{y}) = (I_n - H)\sigma^2$$

بنابراین توزیع های $\hat{y}$ و $\hat{\varepsilon}$ بترتیب عبارتند از :

$$\hat{y} \sim N(X\beta,\ \sigma^2 H) \qquad (E.3.4.3)$$

$$\hat{\varepsilon} \sim N(o, \sigma^2(I_n - H)) \qquad (E.3.4.4)$$

5.3. تعبیر هندسی رگرسیون:

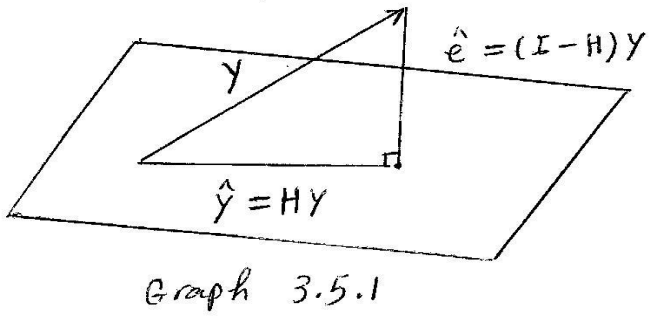
در مدل رگررسیون خطی چند متغیره به دنبال این هستیم که y را بصورت ترکیب خطی از ستونهار X به اضافه خطاها بنویسیم.

$$y_i = \beta_0 + \beta_1 x_{il} + \cdots + \beta_P x_{ip} + \varepsilon_i \text{ و } \quad i = 1,2,..,n$$

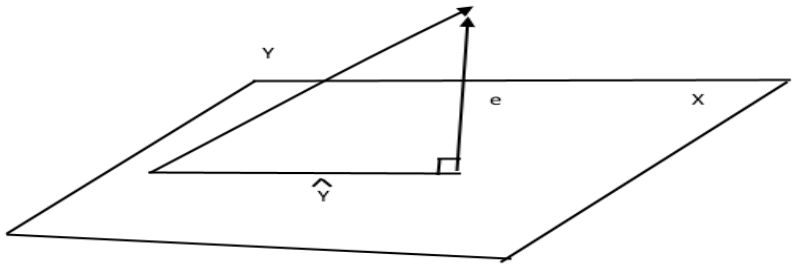
یعنی اگر X معکوس پذیر است (full rank) فضای تولید شده P+1 بعدی خواهد بود. در این حالت ماتریس H تصویرکننده y در زیر فضای X خواهد بود .

و بردار باقیمانده ها $\hat{\varepsilon} = y - \hat{y} = (I_n - H)y$ عمود بر فضای تولید شده توسط بردارهای X خواهد بود.

شکل 3.5.1 را برای رگرسیون خطی سه متغیره ملاحظه کنید (P=2).



Graph 3.5.1



اگر X (full rank) نباشد آنگاه ستونهای خاصی (متغیرهای خاصی) از X همبسته خطی هستند. در این حالت کافی است یک یا چند متغیر را از مدل رگرسیون حذف کنیم تا مدل جدید با ماتریس جدید X دارای رتبه کامل شود.

از طرفی باقیمانده ها، تصویر y در فضای عمود بر X است. (Perpendicular) ، همیشه X را Full rank با بعد n-P در نظر می گیریم.

زاویه قائمه بین بردار برازش $(\hat{y})$ predicted و بردار باقیمانده ها $\hat{\varepsilon}$ در نتیجه برآورد β $(\hat{\beta})$ توسط روش کمترین مجموع توانهای دوم خطا یعنی $\hat{\varepsilon}'\hat{\varepsilon}$ می باشد، از متعامد بودن $\hat{y}$ و $\hat{\varepsilon}$ میتوان گفت که

$\mathrm{COV}(\hat{y},\hat{\varepsilon}) = 0$. بنابراین در عمل بردارهای متعامد، بردارهای نا همبسته هستند. در فضای نرمال بردارهای ناهمبسته، مستقل نیز هستند.

این روش اجازه میدهد که مجددا β را برآورد کنیم یعنی وقتی $\hat{\beta} \times \hat{\varepsilon}$ متعامد هستند، آنگاه :

$$(X\hat{\beta})'\hat{\varepsilon} = (X\hat{\beta})'(y - x\hat{\beta}) = 0$$

اگر $\hat{\beta} \neq 0$ باشد (در غیر اینصورت قابل حل نیست) یعنی :

$$\hat{\beta} = 0$$

یا

$$\hat{\beta} \neq 0 \Rightarrow X'y - X'X\hat{\beta} = 0 \Rightarrow \hat{\beta} = ؟$$

6.3 برآوردهای ناادیب σ^2 و آزمون فرض روی β.

فرض نرمال بودن ، نتیجه می دهد که $\hat{\beta}\sim N_P(\beta,\sigma^2(X'X)^{-1})$ و برای β_i خاصی داریم.

$$\frac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{V_{ii}}}\sim N(0,1)$$

که در آن V_{ii} عنصر (i,i) از ماتریس$(X'X)^{-1}$

اگر σ^2 معلوم باشد آنگاه می توان فاصله اطمینان برای β_i را بدست آورد. از طرفی در عمل σ^2 نامعلوم است و باید برآورد شود که بصورت زیر می توان σ^2 را برآورد کرد:

$$\hat{\sigma}^2=\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-P'} \text{ و } P'=P+1<n$$

برای فهمیدن اینکه چرا بر $n-P'$ تقسیم شده است به قضیه 1 زیر مراجعه کنید:

قضیه 1: فرض کنید Z بردار تصادفی دارای توزیع نرمال استاندارد چند متغیره باشد . یعنی $Z\sim N_n(o,\sigma^2 I_n)$ و $W=AZ$

آنگاه :

$$w'(AA')^{-1}w\ \sim\chi^2_l$$

که در آن l رتبه ماتریس A است.

و با استفاده از قضیه 1 داریم :

بنابراین:

$$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\delta^2} \sim \chi^2_{n-P'}$$

$$E = \left(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}\right) = n - P'$$

مثال 2-9: برای مثال (رگرسیون سه متغیره) σ^2 را برآورد کنیم .

$$\frac{38.21}{20-3} = 2.247$$

از طرفی $\hat{\sigma}^2$ از $\hat{\beta}$ مستقل است پس می توان نوشت:

$$C.I(\beta_i) = \left[\hat{\beta}_i \pm t_{n-P',1-\propto/2}\sqrt{S^2 V_{ii}}\right] \qquad (E.3.6.2)$$

مثال 2-10، یک فاصله اطمینان برای β_1 مدل رگرسیون سه متغیره بیابید.

$$\begin{cases} \hat{\beta}_1 0.01314 \\ V_{22} = 0.0000077\,, \qquad S^2 = \hat{S}^2 = 2.247 \\ n - P' = 20 - 3 = 17 \\ t_{17,0.975} = 2.110 \end{cases}$$

$$C.I_{0.95}\,(\beta_1) = (0.00436, 0.0219)$$

a. **پیش بینی *forecasting***

فرض کنید می خواهیم یک ترکیب خطی از $\underline{\beta}$ داشته باشیم که ما آنرا با θ نشان می دهیم.

$$\theta = \sum_{o}^{P} a_i \beta_i = a'\beta \qquad a = (a_o, a_1, \dots, a_p)'$$

می دانیم که $a'\beta \sim N\ (a'\beta,\ \sigma^2 a'(X'X)^{-1}a)$

این برآوردگر نااریب است و ما نشان می دهیم بهترین آماره نااریب نیز می باشد .

1.7.3. (قضیه) نظریه *Gauss- Markov*

این قضیه بیان می کند که:

آماره $a'\hat{\beta} = a'(X'X)^{-1}X'y$ نا اریب و با کمترین واریانس در بین تمام برآوردگرهای نااریب است یعنی آماره $a'\hat{\beta}$ بهترین آماره است.

اثبات: فرض کنید $C'y$ یک آماره نااریب دیگری از $a'\beta$ باشد. نشان می دهیم که $Var\ (C'y) \geq Var(a'\hat{\beta})$.

می دانیم که :

2.7.3. آزمون فرضها برای $a'\beta$:

اگر σ^2 معلوم باشد آنگاه می توان فاصله اطمینان و آزمون فرض برای θ انجام داد.

که $\frac{a'\hat{\beta}-\theta}{\sqrt{\sigma^2 a'(X'X)^{-1}a}} \sim N\ (0,1)$

اما همیشه σ^2 معلوم نیست و بجای آن از برآوردگر آن یعنی $\hat{\sigma}^2 = \left(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-P'}\right)$ استفاده می کنیم و داریم :

حال با توجه به این نتیجه می توان فاصله اطمینان برای θ یا فرض $H_0: \vartheta = \theta^*$ را آزمون کرد.

3.7.3- تفسیر *E(y|x)*:

می خواهیم مقدار میانگین *y* برای ترکیب خطی *?* $x^{*'} = (x_1^*, \dots, x_p^*)$ را برآورد کنیم یعنی :

$$E(y: x^*) = ?$$

که این برآوردی از ترکیب خطی β است. با استفاده از قضیه *Gauss – Marhov* کافی است که $x^{*'}$ را با a' جایگزین کنیم.

در نتیجه برای بدست آوردن یک فاصله اطمینان $(1-\propto)\%$ برای $E(y|x^*)$ از :

$$CI(E(y: x^{*'}) = [x^{*'}\hat{\beta} \pm t_{n-P',(1-\propto/2)}\sqrt{\hat{\sigma}^2 x^{*'}(\mathrm{X'X})^{-1}\mathrm{X}^*]}$$

(E.3.7.1)

مثال 2-11 :*3-1* را ملاحظه کنید.

در این مثال رگرسیون سه متغیره یک فاصله اطمینان برای *y* به شرط اینکه

$$X \sim N_P\left(\mu_1, \quad \sum\right)$$

$$f_x(x_1, \dots, x_p) = \frac{1}{(2\pi)^{p/2}\sum^{\frac{1}{2}}} \exp\left(-{}^1/_2 (X-\mu)^T \sum^{-1} (X-\mu)\right)$$

در نتیجه برای بدست آوردن یک فاصله اطمینان $(1-\propto)\ 100\%$ برای $E(y|x^{*'})$ می توان از :

$$C.I\left(E: (y: x^{*'})\right) = \left[x^{*'}\hat{\beta} \pm t_{n-p\ ,(1-\propto/2)}\sqrt{\hat{\sigma}^2 x^{*'}(x'x)^{-1}x^*}\right]$$

(E. 3.7.1)

مثال 12-2: *Lab 6 example 3-1* را ملاحظه کنید.

در این مثال رگرسیون سه متغیره یک فاصله اطمینان برای ***y*** به شرط زیر بدست آورید:

$$\begin{cases} x_1 = 325/m/s^2) \\ x_2 = 0.98(g/cm^3) \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i 1 + x_{i2} + \varepsilon_i$$

اگر فرض شود یعنی بهترین برآورد (قضیه *Gauss- Markov*)

$$x^* = \begin{pmatrix} 1 \\ 325 \\ 0.98 \end{pmatrix} \text{ و } \mathrm{x}^{*'}\beta = \vartheta$$

$$\mathrm{x}^{*'}\hat{\beta} = (1, 325\ , 0.98)\begin{pmatrix} -33.831 \\ 0.01314 \\ 31.890 \end{pmatrix} = 4.63$$

$$var\ (\mathrm{x}^{*'}\hat{\beta}) = \hat{\delta}^2 x^{*'}(x'x)^{-1}x^* = 2.217(\ 1\ 325098)$$

$$\begin{pmatrix} 24.667 & 0.0053 & -26.747 \\ 0.0053 & 0.0000077 & -0.0084 \\ -26.747 & -0.0084 & 30.0964 \end{pmatrix}\begin{pmatrix} 1 \\ 325 \\ 0.98 \end{pmatrix} = 2.247 \times 0.0718$$

$$= 0.1613 \Rightarrow CI(E/yx^*) = (3.78\ , 5.48)$$

3.7.3. تفسیر و نتیجه گیری روی مقدار ***y*** به شرط ***x*** :

فرض کنید ***y*** مقدار پیش بینی بر روی ترکیب خطی x^* باشد.

یعنی ابتدا می خواهیم $y = x^{*'}\beta + \varepsilon$ را برآورد کنیم. بنا به قضیه *Gauss- Markov* بهترین برآوردگر $x^{*'}\beta$ عبارتست از $x^{*'}\hat{\beta}$

از آنجائیکه $E(\varepsilon) = 0$ در نتیجه برآوردگر نقطه ای عبارتست از $x^{*'}\hat{\beta} = x^{*'}\hat{\beta} + 0$

متغیری جدید توصیف می شود:

$$\omega_* = x^{*'}\hat{\beta} - (x^{*'}\hat{\beta} + \varepsilon)$$

$$\mathrm{V}(\omega_*) = \mathrm{V}\left(x^{*'}\hat{\beta}\right) + \mathrm{V}(\in)$$

در نتیجه داریم :

$$\frac{x^{*'}\hat{\beta} - (x^{*'}\beta + \varepsilon)}{\sqrt{\hat{\sigma}^2 x^{*'}(x'x)^{-1}x^*}} \sim \tau_{n-p"}$$

از 1 فاصله اطمینان $100 \times (1-\propto)\ \%$ برای پیش بینی y به شرط x^* عبارتست از :

$$CI\left(\left(y/x^{*'}\right)\right) = \left[x^{*'}\hat{\beta} \pm t_{n-p',(\propto/2)}\sqrt{\hat{\sigma}^2 x^{*'}(x'x)^{-1}x^*}\right]$$

مثال 2-13: در مثال قبل، یک فاصله اطمینان 0.95 برای پیش بینی y به شرط x^* بدست آورید :

$$x^{*'} = (1, 325, 0.98)$$

$$x^{*'}\hat{\beta} = 4063^{(mm)}, \quad x^{*'}(x'x)^{-1}x^* = 0.0718$$

بنا به فاصله اطمینان فرض داریم :

8.3. آنالیز واریانس ***Variance Analysis***

$$y = f(x_1, \dots, x_p) + \varepsilon$$

$$\begin{pmatrix} Response\ variablels \\ y_1, \dots, y_n \end{pmatrix} = \begin{pmatrix} explication \\ variables \\ x_1, \dots, x_n \end{pmatrix} + \begin{pmatrix} Random\ Variables \end{pmatrix}$$

(3.18)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$+(2\sum \hat{e}_i(\hat{y}_i - \bar{y}) = 0) = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

یعنی داریم:

$$SS_{Tot} = SS_{Reg} + SS_E$$

و فرم ماتریس آن خواهد بود:

$$SS_{Tot} = \sum(y_i - \bar{y})^2 = y'y - n\bar{y}^2$$

$$SS_{Reg} = \sum(\hat{y}_i - \hat{y})^2 = \hat{y}'\,\hat{y} - n\bar{y}^2 = \hat{\beta}'x'x\hat{\beta} - n\bar{y}^2$$

$$SS_{Res} = \sum(y_i - \hat{y}_i)^2 = (y - \hat{y})'(y - \hat{y}) = \hat{\varepsilon}'\varepsilon$$

$$SS_T = SS_{Reg} + SS_{Res}$$

$$SS_T = \left(\hat{\beta}'x'\,x\beta - n\bar{y}^2\right) + \hat{\varepsilon}'\varepsilon \qquad (3.19)$$

که بنا به قضیه پیتاگور *pythagore* می توان نوشت (3.19):

$$\|y\|^2 = \|x\hat{\beta}\|^2 + \|y - x\hat{\beta}\|^2$$

که در آن $\|\,.\,\|$ بیان گر نرم یا طول بردار خواهد بود.

از طرفی داریم:

$$y'y = (\,x\hat{\beta})' \times \hat{\beta} + (y - x\hat{\beta})'(y - x\hat{\beta})$$

اگر این فرمول را باز کنیم به $\hat{\beta}' \times y'$ یا $(\,x\hat{\beta})' \times y$ که بدلیل متقارن بودن $\hat{y}, y$ صفر خواهند بود .

$$\Rightarrow y'y - n\bar{y}^2 = \hat{\beta}'x'x\hat{\beta} - n\bar{y}^2 + \hat{\varepsilon}'.\hat{\varepsilon}$$

که همان $SS_T = SS_{Reg} + SS_{Res}$ خواهد بود.

و جدول آنالیز واریانس را می توان نوشت:

منبع تغییرات	*DF*	*SS*	*MS*	$F_{P,n-p'}$
رگرسیون	*p*	$\sum(\hat{y}_i - \bar{y})^2$	*SSR/p*	$\frac{M_{SR}}{MS_E}$
خطا	$n - p'$	$\sum(y_i - \hat{y}_i)^2$	*SSE/(n − p′)*	
Total	*n-1*	$\sum(y_i - \bar{y})^2$		

نیز می توان جدول آنالیز واریانس شامل اثرات β_0 را بصورت زیر تنظیم کرد:

منبع تغییرات	*DF*	*SS*	*MS*	*F*

β_0	*1*	$n\bar{y}^2$	$n\bar{y}^2$	$n\bar{y}^2/S^2$
model	*p*	SS_R	*SSR/p*	$\frac{MS_R}{MS_E}$
Error	$n-p'$	SS_E	$SS_E/(n-p')$	
Total	*n*	$SS_T + n\bar{y}^2$		

مثال 2-14– جدول آنالیز واریانس را برای مثال قبل محاسبه نمائید.

$$SS_T = y'y - n\bar{y}^2 = 904.60 - 20\left(\frac{12.179}{20}\right) = 175.089$$

$$SS_{Res} = \hat{\varepsilon}'.\hat{\varepsilon} = \delta^2(n-p') = 38.21$$

$$SS_{Reg} = SS_T - SS_{Res} = 136.88$$

منبع تغییرات sourco	*DF*	*SS*	*MS*	*F*
model	*2*	*136.88*	$\frac{136.88}{8} = 68.44$	*30.46*
Error	*17*	*38.21*	$\frac{38.21}{17} = 2.247$	
Total	*19*	*175.09*		

3.9- آزمون عمومی *F* برای رگرسیون خطی چند متغیره

این آزمون در رگرسیون چند متغیره به منظور بررسی معنی دار بودن یا نبودن مدل رگرسیون است.

$$\begin{cases} H_0: \ \beta_1 = \beta_2 = \cdots = \beta_P = 0 \\ H_1: \text{حداقل یکی از ضرایب مخالف صفر باشد} \end{cases}$$

تحت فرض H_0 با مقدار :

$$F_o = \frac{MS_{Reg}}{MS_E} : \ F_{1-\alpha,P,n-P'}$$

مقدار آماره آزمون را با تابع فیشر و درایج آزادی مربوط(جدول)، مقایسه می کنیم.

مثال 2-15 : در مثال قبل داریم:

$$F_{\mathrm{o}} = \frac{\mathrm{MS}_{\mathrm{Reg}}}{\mathrm{MS}_{\mathrm{E}}} = \frac{68.44}{2.247}$$

$$= 30.46$$

$$\begin{cases} \mathrm{H}_0: \ \beta_1 = 0 \text{و} \mathrm{i} = 1 \text{و} \dots, \mathrm{p} \\ H_1: if\ not \end{cases}$$

$$\mathrm{F}_{1-\propto,\mathrm{P},\mathrm{n}-\mathrm{P}'} = \mathrm{F}_{2,17}(0.95) = 3.59$$

چون

$$F_O > F_{1-\propto} \Rightarrow RH_o$$

اما در عمل همیشه اینطور نیست یعنی ممکن است k تا$(k < P)$ از متغیرهای آزاد $\mathrm{x}_1, \dots, x_k$ برای مدل لازم باشد. در این صورت باید مدلهای زیر را آزمون کنیم.

$$\begin{cases} H_0: y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \to \hat{\varepsilon} \\ H_1: y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \to \hat{\varepsilon}^* \end{cases}$$

اگر $SS_{\mathrm{Res}}^{(0)}$ تحت فرض H_0 و $SS_{\mathrm{Res}}^{(1)}$ تحت فرض H_1 باشد و همیشه داریم :

$$SS_{\mathrm{Res}}^{(0)} - SS_{\mathrm{Res}}^{(1)}$$

که این تفاوت می تواند کم یا زیاد باشد، و بهتر است :

در نظر بگیریم ومی توان تابع معیارF را اینطور تعریف کرد :

$$F = \frac{(SS_{\mathrm{Res}}^{(0)} - SS_{\mathrm{Res}}^{(1)})/\Delta\mathrm{dd}}{SS_{\mathrm{Res}}^{(1)}/(\mathrm{n} - \mathrm{p}')} = \frac{SS_{\mathrm{Res}}^{(0)} - SS_{\mathrm{Res}}^{(1)}}{\Delta df \ S_{H_1}^2}$$

که در آن $l = \Delta_{df}$ اختلاف بین درایج آزادی تحت $\mathrm{H_o}$ و H_1 است. بنابراین تحت فرض $\mathrm{H_o}$، هرگاه :

روشهای محاسبه Δdf:

$$1 - \Delta df = df\left(SS_{\text{Res}}^{(0)}\right) - \mathrm{d}f\left(SS_{\text{Res}}^{(1)}\right)$$

$$2 - \Delta df = \#\,(H_1) - \#\,(H_0)$$

$3 - H_o$ تعداد شرط ها برای ازرسیدن مدل تحت H_1 به مدل

به بیان دیگر :

$$1 - df\ SS_{\text{Res}}^{(0)} = \mathrm{n} - (\mathrm{k} + 1) = \mathrm{n} - \mathrm{k} - 1$$

$$df\left(SS_{\text{Res}}^{(1)}\right) = \mathrm{n} - \mathrm{P}'$$

$$\Rightarrow \Delta df = n - k - 1 - n + P'$$

$$\Delta df = p - k$$

$$2 - \#\ parameter\ under\ \ H_1 = P + 1$$

$$\#\ parameter\ under\ \ H_0 = K + 1$$

$$\Delta df = P + 1 - K - 1 = P - K$$

برای رسیدن از H_1 به H_0 داریم:

$$\beta_{\mathrm{i}} = 0\ , i = k + 1\ , \ldots , P$$

پس p-k شرط روی پارامترهای (معادلات مساوی 0) ؟.

تذکر12: مدلهای جدید تحت فرض $\mathrm{H_o}$ و H_1 قابلیت پوشش فرض های اولیه را نیز دارا می باشد . اگر $\Delta df =$ P یا $k=0$ فرض شود.

جدول آنالیز واریانس کلی (تعمیم یافته) برای مدل تعدیل یافته :

منبع تغییرات	*DF*	*SS*	*MS*	*F*

(comR – RedR	$P-K$	$SS_{Reg}=\|\|\hat{y}-\hat{y}^{*}\|\|^{2}$	$SS_{Reg}/(P-K)$	$\frac{MS_{Reg}}{MS_{Res}}$
E(complete)	$n-P'$	$SS_{Res}=\|\|\hat{\varepsilon}\|\|^{2}$	$\frac{SS_{Reg}}{n-P'}$	
E (Reduced)	*n-K'*	$SS_{Res}=\|\|\hat{\varepsilon}^{*}\|\|^{2}$		

آزمون فرض کلی مدلهای خطی رگرسیون چندگانه

بدین منظور فرض کنید . $C_{\mathrm{r}\times\mathrm{p}'}$ یک ماتریس و $d_{\mathrm{r}\times 1}$ یک بردار باشد آنگاه آزمون فرض کلی را می توان بصورت زیر نوشت:

$$\begin{cases}\mathrm{H}_0\colon\ C\beta=d\\ \mathrm{H}_1\colon \mathrm{C}\beta\neq \mathrm{d}\end{cases}$$

یا به عبارت دیگر می توان آزمون فوق را به صورت زیر بیان کرد:

$$\begin{cases}H_0: y=Z\,\alpha+\varepsilon & \text{مدل کاهش یافته :}\\ H_1\quad \mathrm{y}=X\,\beta+\varepsilon & \text{مدل کامل :}\end{cases}$$

از معادله $C\beta=d$ تحت فرض H_0 , $\hat{\beta}$ بدست آمده و داخل سیستم با معادله $y=\mathrm{X}\beta+\varepsilon$ در فرض H_1 قرار می دهیم.

و می توان نشان داد که:

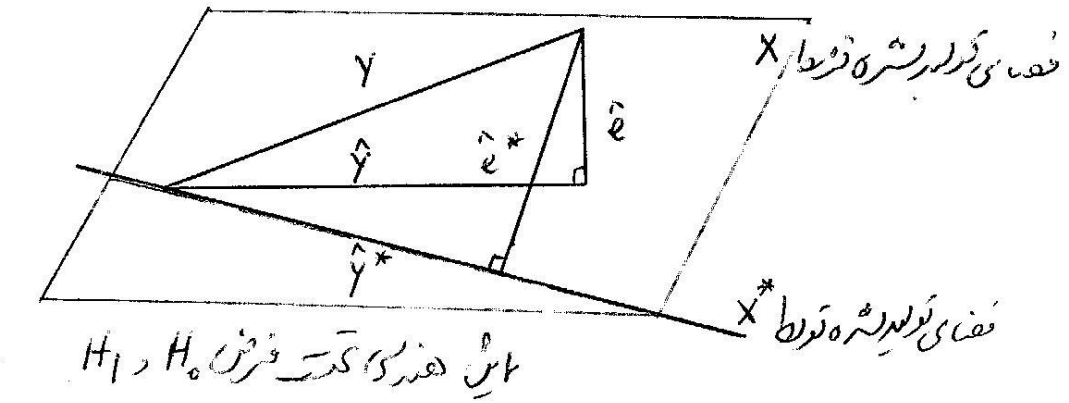
$$\left(SS_{\mathrm{Res}}^{(0)}\right)-\left(SS_{\mathrm{Res}}^{(1)}\right)=\ ?$$

که در آن $\sigma^2[\mathrm{C}(X'\mathrm{X})^{-1}\mathrm{C}']$ واریانس $\mathrm{C}\hat{\beta}-\mathrm{d}$ می باشد.

یعنی اگر $\mathrm{C}\hat{\beta},\mathrm{d}$ اختلاف کمی داشته باشند آنگاه فرض H_0 رد نخواهد شد .

و آزمون F :

که با $F_{r,n-P'}$ مقایسه می شود.



فصل چهارم : معیارهای انتخاب مدل

1.4.مقدمه :

در این حالت فرض می شود که تعداد زیادی متغیر آزاد (کمکی) داریم. درواقع می خواهیم مدلی را انتخاب کنیم که تمام متغیرهای داخل مدل (ضرایب) معنی دار باشند.

لذا فرض می شود مدل خوب عبارت باشد از :

$$y_i = \beta_0 + \beta_1 x_{i1} + \ \dots + \ \beta_h x_{ih} + \varepsilon_i$$

اگر بخواهیم متغیر $x_{i\,h+1}$ را به مدل اضافه کنیم خواهیم داشت :

$$y_i = \beta_0^* + \beta_1^* x_{i1} + \ \dots + \beta_h^* x_{ih} + \beta_{h+1}^* x_{i\,h+1} + \varepsilon_i$$

اگر $V(\hat{\beta}^*) \geq v(\hat{\beta})$ بیانگر این است که دقت مدل جدید کمتر شده است .

برای توصیف اریبی مدل، مثال زیر را در نظر می گیریم :

$$y = X_1 \beta_1 + \varepsilon \qquad (4.1)(\,wrong\ model)$$

$$y = \ X_1 \beta_1 + \ X_2 \beta_2 + \varepsilon = \ X\beta_z + \varepsilon \ \ (4.2)(\,correct\ model)$$

فرض کنید ? شامل $P' = P + 1$ عنصر باشد، و β_2 شامل $m - P'$ پارامتر . فرض می شود که بجای مدل خوب، ما مدل بد (4.1) را برگزیده ایم *(m=ntp)* .

$$E(\hat{\beta}_1) = \beta_1 + \ (X_1'X_1)^{-1}X_1'\,X_2\beta_2$$

بنابراین ملاحظه می شود که $\hat{\beta}_1$ یک برآوردگر اریب برای β_1 می باشد.

اگر کسی مایل باشد که y را در نقطه (بردار) x^* پیش بینی کند بطوری که مدل (4.1) دارای P' پارامتر باشد:

$$\hat{y}(x^*) = x_1^{*'}\hat{\beta}_1$$

$$E\big(\hat{y}(x^*)\big) = x_1^{*'}E\,(\hat{\beta}_1)$$

$$x^{*\prime}\{\beta_1 + (X_1'X_1)^{-1}X_1'\,X_2\beta_2\}.$$

بنابراین مدل زیر دارای m پارامتر بوده و

$$\hat{y}(x^*) = x_1^{*'}\hat{\beta}_1 + x_2^{*'}\hat{\beta}_2$$

$$E\big(\hat{y}(x^*)\big) = x_1^{*'}\beta_1 + x_2^{*'}\beta_2$$

آنگاه اریبی مدل برای پیش بینی عبارتست از :

$$Bias\,\hat{y}(x^*) = x_1^{*'}\{\beta_1 + (X_1'X_1)^{-1}X_1'\,X_2\beta_2\} - x_1^{*'}\beta_1 - x_2^{*'}\beta_2$$

$$Bias\,\hat{y}(x^*) = x_1^{*'}\big\{(X_1'X_1)^{-1}X_1'\,X_2 - x_2^{*'}\big\}\beta_2$$

نتیجه؟؟:

2.4. معیارهای مقایسه کلاسیک

با فرض اینکه مدل مورد نظر یک مدل خوب می باشد. معیار ضریب تعیین، بخوبی میزان تغییرات بیان شده توسط x_i ها را نشان می دهد.

$$\mathrm{R}^2 = \frac{\mathrm{SS_{Reg}}}{\mathrm{SS_T}} = 1 - \frac{\mathrm{SS_{Res}}}{\mathrm{SS_T}} \qquad (4.2.1)$$

بطوریکه می دانیم $0 \le R^2 \le 1$ و هر چه R^2 به یک نزدیک باشد بیانگر این است که مدل برازش شده بسیار مناسب تر است از آنجائیکه $\mathrm{SS_T}$ به مدل بستگی ندارد و با اضافه کردن ...

1.2.4. ضریب تعیین تعدیل شده :

$$R_a^2 = 1 - \frac{MS_{Res}}{MS_T} = 1 - \frac{SS_{Res}/(n - P')}{SS_T/(n-1)} = 1 - (n-1)\frac{S^2}{SS_T}$$

مثال 2-16: در مثال رگرسیون سه متغیره قبل

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{136.88}{175.09} = 78.2\ \%$$

$$R_a^2 = 1 - (n-1)\frac{S^2}{SS_T} = 1 - 19 \times \frac{2.247}{175.09} = 75.6\%$$

2.2.4. روشهای مبتنی بر توان پیش بینی مدل.

1.2.2.4. اصول *Cross – Validation*

این اصول توسط الگوریتم زیر بیان می شود:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

این معیار برای مقایسه کردن تمام مدل ها حتی مدلهای تبدیل یافته *BOX – COX* نیز استفاده می شود.

مطمئنا هر مدلی که دارای $PRESS$ کمتر باشد نشان دهنده دقت پیش بینی بیشتر می باشد.

براساس این معیار می توان یک ضریب تعیین پیش بینی به شکل زیر تعریف کرد:

$$R_P^2 = ؟$$

بنابراین R_P^2 هر چه به یک نزدیک باشد نشان دهنده این است که مدل مورد نظر نیز خیلی خوب پیش بینی می کند.

3.2.4. باقیمانده های PRESS :

برآورد y_i توسط PRESS عبارتست از $\hat{y}_{i,-i}$ پس :

برای مقایسه مدل ها می توان از مجموع توانهای دوم خطای PRESS استفاده کرد و نیز می توان تک تک $\hat{\varepsilon}_i$ را بصورت انفرادی آزمون کرد . چون در یک مدل خوب، یک پیش بینی بد می تواند روی PRESS اثر منفی بزرگی بگذارد.

بنابراین محاسبات آن خیلی خسته کننده به نظر می رسد که خوشبخانه با استفاده از خواص ماتریس H محاسبات خیلی ساده تر خواهد شد.

قضیه 3- i امین باقیمانده PRESS می تواند به کمک i امین باقیمانده معمولی در موقعیت عنصر (i , i) از ماتریس H بدست آید.

$$\hat{\varepsilon}_{i,i} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

نتیجه عملی قضیه فوق این است که :

آماره PRESS و باقیمانده های PRESS بدون انجام n رگرسیون از روش cross- validation قابل دستیابی است.

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1 - h_{ii}} \right)^2$$

که در آن h_{ii} قبلا تعریف شده است یعنی h_{ii} طول فاصله پیش بینی را نشان می دهد . $0 \le h_{ii} \le 1$. یعنی یک مقدار h_{ii} نزدیک 1 نشان دهنده فاصله پیش بینی بزرگ می باشد.

مادامی که مخرج $\hat{\varepsilon}_{i,i}$ نزدیک 0 است یعنی h_{ii} نزدیک 1 است در نتیجه $\hat{\varepsilon}_{i,i}$ نزدیک $+\infty$ است و در نتیجه پیش بینی مقدار y_i خیلی مشکل است.

4.4- معیار C_P (Mallows) :

این روش براساس میانگین توانهای دوم خطای برآوردگر بنا شده است

Mean of Square of Error (MSE)

فرض کنید $\hat{y}(x^*)$ برآوردگر $y(x^*)$ باشد.

$$MSE\left(\hat{y}(x^*)\right) = \mathrm{V}\left(\hat{y}(x^*)\right) + bias^2\left(\hat{y}(x^*)\right)$$

$$E\left(\hat{y}(x^*)\right) = \mathrm{y}(x^*) = \mathrm{E}\left(\overline{\hat{y}}(x^*)\right), \vartheta = \mathrm{y}(x^*)$$

$$MSE\left(\hat{y}(x^*)\right) = \delta^2\, x_1^{*\prime}(X_1'X_1)^{-1}X_1^* + \left\{ x_1^{*\prime}\left\{(X_1'X_1)^{-1}X_1'\, X_2 - x_2^{*\prime}\right\}\beta_2\right.^2$$

براساس مدل صفحه قبل:

حال برای می نیموم کردن *MSE* با دو مشکل مواجه هستیم . اولا، بردار x^* را نمی شناسیم (نا معلوم است) .

دوم اینکه *MES* به β_2 نامعلوم بستگی دارد.

اما *Mallows* یک معیار دیگری را معرفی کرده است که به ما اجازه می دهد تا اولین مشکل را حل کنیم.

لذا می توان بجای *MSE*:

$$\sum_{\mathrm{i}=1}^{n} \frac{\mathrm{MSE}\left(\hat{\mathrm{y}}(\mathrm{x_i})\right)}{\sigma^2}$$

را می نیمم کرد . که در آن تقیسم بر σ^2 نقشی جز ساده کردن محاسبات ندارد.

محاسبه بخش واریانس در *MSE* :

$$\sum_{\mathrm{i}=1}^{n} \frac{\mathrm{V}\left(\hat{\mathrm{y}}(\mathrm{x_i})\right)}{\sigma^2} = tr[(x_1'x_1)(X_1'X_1)^{-1}]$$

$$= tr(I_{P'}) = P', \qquad tr(AB) = tr(BA)$$

محاسبه بخش *bias* اریبی برآوردگر:

$$EC(S_P^2) = \sigma^2 + \frac{\sum_1^n bais\,[\hat{y}(x_i)]}{n - P'}$$

که در آن S_P^2 واریانس مدل با P پارامتر است . اگر σ^2 معلوم باشد آنگاه میتوان $\sum bais\,[\hat{y}(x_i)]^2$ را با $(S_P^2 - \sigma^2)(n - P')$ برآورد کرد .

با اینکه σ^2 نامعلوم است آنگاه از برآوردگر آن یعنی $\hat{\sigma}^2$ واریانس داخل مدل کامل (مدل بزرگتر) استفاده کرد .

بنابراین معیار سنجش پیشنهاد شده توسط *Mallows*:

$$C_P = ؟$$

در عمل مدلی خوب است که در آن :

$$C_P - P' \approx o \quad \text{یا} \quad C_P \approx P'$$

اگر k متغیر آزاد (کمکی) موجود باشند آنگاه به تعداد 2^{k+1} مدل رگرسیون خواهیم داشت که در نتیجه 2^{k+1} تا C_P قابل محاسبه و مقایسه می باشند که از میان آنها می بایست مدلی را انتخاب کرد که دارای $C_P - P'$ کوچکتر باشد.

اما در عمل باید بین تعدادی مدل قابل قبول (توجیه پذیر)، آنی را که دارای کوچکترین $C_P - P'$ مقدارمی باشد انتخاب کنیم. اما تعدادی از مدل های قابل توجیه دارای $C_P - P'$ تقریبا کوچک باشند باید برای جدا کردن مدل خوب از معیارهای دیگری مثل $(PRESS\ , R_a^2, \dots)$ استفاده کرد.

از طرفی می توان C_P را به شکل زیر ایجاد کرد:

$$C_P = \frac{SS_{Res}}{\sigma^2} + 2P' - n$$

i. **معیار اطلاعات آکائیک *(AIC)*** :

Akaike Information Citer

این روش توسط آماردان ژاپنی *Hirotsugu Akaike* در سال (1974) ارائه شد. این معیار در عمل استفاده فراوان دارد و برای بررسی کیفیت مدل بکار می رود.

$$AIC = -2lnL\left(\hat{\beta}, \delta_{ML}^2\right) + 2P'$$

$$= n + ln\,(2\pi) + ln\left(\frac{SS_{Res}}{n}\right)$$

چون

$$\widehat{\sigma}^2_{ML} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n = \frac{SS_{Res}}{n}$$

اگر مؤلفه ثابت فوق را کنار بگذاریم داریم :

$$AIC = ln\left(\frac{SS_{Res}}{n}\right) + 2P'$$

همانطور که ملاحظه می شود AIC براساس SS_{Res} بنا شده یعنی هر چه پیش بینی و دقیقتر باشد، آنگاه AIC مقداری کوچکتر (ضعیفتر) را بخود اختصاص می دهد. لذا هر چه مقدار AIC به $2P'$ نزدیک تر باشد بیانگر این است که مدل از پیش بینی قابل قبولی بهره مند است.

5.2.4. معیار (روشهای) های **اگوریتمی *Algorithimic method***

1.5.2.4. **روش پیشرو *forward method***

در این روش با مدل

1- $y = \beta_o$ شروع می کنیم.

2- متغیری از متغیرها را به مدل اضافه می کنیم که بیشترین اثر را روی کاهش SS_{E} داشته باشد.

3- سپس آزمون زیر را برای متغیر اضافه شده به مدل انجام می دهیم.

$$F_0 = ؟$$

که درآن $l = n - P'$ و فرض می شود که متغیر افزوده درست است هر گاه

$$\mathrm{F}_0 > \mathrm{F}_{1,l,(1-\alpha)}$$

4- اگر متغیر افزوده صحیح باشد می توان گام 2 را برای متغیر بعدی بکار برد و آنرا تا زمانی ادامه می دهیم که اضافه کردن هیچ متغیری معنی دار نباشد و معمولا سطح معنی داری $\alpha = 0.05$ فرض می شود.

$$0.05 \leq \alpha \leq 0.4$$

معمولا ابتدا (در شروع) α را بزرگ در نظر می گیریم تا مدل شامل متغیر باشد.

2.5.2.4. **روش پسرو *backward method***

این روش برعکس روش پیشرو خواهد بود

1. مدل کامل را در نظر می گریم (شامل متغیرها).
2. سپس متغیری از متغیرها را از مدل برمی داریم بطوریکه کمترین اثر روی افزایش خطا را داشته باشد.

3. و سپس آنرا آزمون می کنیم بطوریکه معنی دار نباشد یعنی

$$F_0 < F_{1,l,(1-\alpha)}$$

$$F_0 = \frac{\{SS_{\text{Res}}(\text{R}) - SS_{\text{Res}}(\text{C})\}}{SS_{\text{Res}}(\text{C})/l}$$

4. اگر متغیر حذف شده از مدل معنی دار نبود آنگاه برای متغیرهای بعدی روش 2 و 3 را ادامه می دهیم تا اینکه هیچ متغیری از مدل را نتوان حذف کرد و معمولا $\alpha = 0.1$ یا $\alpha = 0.2$فرض می شود .

تذکر 13 (مهم):

3.5.2. روش گام بگام ***(stepwise)***

این روش مخلوطی از روش های پیشرو و پسرو می باشد و با روش پیشرو شروع می شود.

مثال 2-17 – در این مثال روشهای پیشرو – پسرو و گام بگام را بررسی می کنیم .

متغیرهای داخل ها مدلها	SS_{Res}
x_0=*constant*	*30*
x_1	*25*
x_2	*24*
x_3	*20*
x_1 , x_2	*6*
x_1 , x_3	*15*
x_2 , x_3	*12*
x_1 , x_2, x_3	*4*

روش پیشرو : 1- با x_3 شروع می کنیم و

$$F_0 = \frac{30-20}{20/(10-20)} = 4 > F_{0.5.1,8} = 0.5$$

2 – حال x_2 را اضافه می کنیم و داریم :

$$F_0 = \frac{20-12}{12\ (10-3)} = 4.67 > F_{0.5\ ,1\ ,7}$$

پس x_2 و x_3 را داخل مدل نگه میداریم.

3- حال x_1 را به مدل اضافه می کنیم :

$$\mathrm{F_o} = \frac{12-4}{4/(10-4)} = 12 > F_{0.5\,,1,6}$$

با این روش همه متغیرها را داخل مدل نگه می داریم.

روش پسرو:

در این روش از مدل کامل شروع می کنیم . حال باید متغیری را حذف کرد که کمترین اثر را روی *Error* داشته باشد پس x_3 را حذف می کنیم .

$$\mathrm{F_o} = \frac{6-4}{4/(10-4)} = \frac{2\times 6}{4} = 3 < F_{0.1\,,1,8} = 3.46$$

چون اثرش معنی دار نیست پس می توان x_3 را حذف کرد. حال فقط داخل مدل متغیرهای x_1 و x_2 را داریم. که از این بین x_2 را حذف می کنیم.

$$\mathrm{F_o} = \frac{24-6}{6/(10-3)} = \frac{18\times 7}{6} = 21 < F\ (0.1)$$

لذا نمی توان x_2 یا x_1 را حذف کرد و مدل با این دو متغیر خواهد بود.

روش گام به گام:

1- با مدل x_0شروع می کنیم.

2- x_3 را به مدل اضافه می کنیم .

$$\mathrm{F_o} = \frac{30-20}{20/8} = 4 >\ 0.5$$

3- پس x_3 را نگه می داریم .

4- x_2 را به مدل اضافه می کنیم .

$$\mathrm{F_o} = \frac{20-12}{12/7} = \frac{8\times 7}{12} = 4.67 > 0.5$$

پس مدل دارای $\mathrm{x_3}$ و x_2 می باشد. و نمی توان هیچ کدام از آنها یعنی $\mathrm{x_3}$ و x_2 را حذف کرد چون هر دو معنی دار هستند (روش پسرو) مقایسه با *F(0.1)*

5- روش پیشرو: حال x_1 را به مدل اضافه می کنیم :

$$\frac{12-4}{4/6}=\frac{8\times 6}{4}=12>F_{0.5}$$

6- روش پسرو: x_3 را حذف می کنیم :

$$\frac{6-4}{4/6}=\frac{2\times 6}{4}=3<F_{0.1\ ,1,6}$$

پس x_3 را حذف می کنیم و در نهایت متغیرهای x_2 و x_1 را نگه می داریم که همان نتیجه روش پسرو می باشد.

4.7- گرافهای متغیرهای اضافه شود: (رگرسیون پیشرفته)

فصل پنجم : هم خطی چندگانه (چند هم خطی) یا همبسگی داخلی

5.1- تعریف:

آنگاه این چند هم خطی را کامل گویند. چنانچه اسکالی های $d_1, \dots d_p$ (که تماما صفر نیستند) و چند همخطی را غیرکامل یا به مفهوم بزرگ نامند.

$$\left(i \text{ دقیق}\right) \quad \sum_j d_j x_j$$

در صورتیکه چند همخطی کامل داشته باشیم آنگاه ماتریس $(x'x)$ معکوس پذیر نیست.

و در نتیجه برازش خط رگرسیون غیرممکن است. که این مشکل قابل حل است و میتوان متغیری را که تابعی از بقیه می باشد را پیدا و حذف کرد و با متغیرهای باقیمانده که چند همخطی کامل وجود ندارد خط رگرسیون را برازش داد . (مثالی از چند هم خطی کامل بزنید)

اما چند همخطی بزرگ(غیرممکن) یا غیردقیق که معمولا اتفاق می افتد و تشخیص آن مشکل است.

که این مشکلات زیر را ایجاد می کند:

ممکن است اشکالات زیر بر اثر هم خطی داخلی بوجود آید:

1- ایجاد بی ثباتی در $(X'X)^{-1}$ یعنی تغییرات بزرگ در $\hat{\beta}$ ها باعث تغییرات بسیار کمی در $\hat{y}$ شود .

2- مشاهده $\hat{\beta}_i$ هایی دور از انتظار یا برخلاف درک شهودی.

3- ایجاد واریانس بزرگ $\hat{\beta}_i$ ها یا $\hat{y}_i$ ها.

4- روشهای انتخاب متغیرها با هم مطابقت ندارند.

5- بعضی از متغیرها گرچه معنی دار نیستند ولی دارای همبستگی بزرگ با *y* هستند.

3.2- تشخیص چند همخطی داخلی نادقیق:

5.2.1- از طریق ماتریس همبستگی :*contre-intuitif* غیرمنتظره درک شهودی

بدین منظوره پس از محاسبه و مشاهده ماتریس همبستگی بین متغیرهای آزاد *(exogon)* چند همخطی را مشخص می کنیم:

$$x_j^* = \frac{x_j - \overline{x}_j}{S_j} \quad , j \in \{1,2,..,P\}$$

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}\,, \qquad \overline{x}_j = \left(\overline{x}_{j1}\,, \ldots, \overline{x}_{jn}\right)$$

$$S_j = \sqrt{\sum \left(x_{ji} - \overline{x}_j\right)^2/(n-1)}$$

و ماتریس همبستگی :

$$x^{*\prime}x^* = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1P} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \vdots \\ r_{P1} & & \cdots & 1 \end{bmatrix}$$

$$X^* = \left(\frac{x_1 - \overline{x}_1}{S_1} \ldots \frac{x_P - \overline{x}_P}{S_P}\right)$$

$$R_{jh} = \sum \frac{\left(x_{ji} - \overline{x}_j\right)\left(x_{hi} - \overline{x}_h\right)}{S_j S_k}, j,k \in \{1, \ldots, P\}$$

اگر بین دو متغیر آزاد هم خطی وجود داشته باشد می بایستی یک ضریب همبستگی خطی بزرگ بین آنها وجود داشته باشد.

اما در این بین دو مشکل بزرگ وجود دارد:

1- مشکل می توان گفت که بین کدام ها یک همبستگی بزرگ خطی وجود درد.

2- اکثرا چند همخطی بین بیش از دو متغیر آزاد *(exogene)* پیش می آید.

اگر در یک مثال P متغیر آزاد موجود باشند که بطور کامل همبسته خطی باشند آنگاه ضرایب همبستگی زوج متغیر باید کمتر از $\frac{1}{P-1}$ باشند.

5.3- عاملی فاکتور تورم یا بزرگ شدن واریانس *(VIF)*

Variance Infolation factor