تذکر:این جزوه با هدف فعالیت بیشتر دانشجو، کامل نیست.

گروه آمار، محمدحسین دهقان

Advance Linear Regression Models

مدلهای رگرسیون خطی پیشرفته

فهرست مطالب:

فصل سوم : معیارهای انتخاب مدل

۱,۴. مقدمه

۲/۴. معیارهای کلاسیک انتخاب مدل

۱/۲/۴. ضریب تعیین

۲/۲/۴.روشهای مبتنی بر توان پیش بینی

۱/۲/۲/۴. روش CV

۳/۲/۴. باقیمانده های PRESS

۴/۲/۴. معیار AIC

۵/۲/۴. معیارهای الگوریتمی

۱/۵/۲/۴. معیار پیشرو

۲/۵/۲/۴. معیار پسرو

۳/۵/۲/۴. معیار گام بگام

فصل چهارم: هم خطی چندگانه

۱) مقدمه: مفهوم همخطی

۲) روشهای رفع همخطی

۳) متغیرهای گروهی یا اسمی در رگرسیون

۴) اثر متقابل دو یا چند متغیر در مدل

فصل پنجم : مدلهای غیر خطی

۱. انواع ضریب همبستگی

۲. مقدمه ای بر مدلهای غیر خطی

۳. تبدیل مدلهای غیر خطی به مدلهای خطی

۴. رگرسیون ذاتا غیر خطی

۵. رگرسیون لوجیستیک

۶. رگرسیون وزنی

۷. **رگرسیون بریج**

فصل ششم: مدلهای رگرسیون تعمیم یافته

مراجع:

۱. جزوه درسی

۲. Applied_Regression(Bookos.org) [Norman_R._Draper,_Harry_Smith]

۳. جبرخطی و کاربردهای آن (دکتر بزرگ نیا و رضایی)

(۰-۱-۰) **یاد آوری جبر خطی کاربردی**:

۰,۱- **نماد و مفاهیم جبرخطی ماتریسها**:

فرض کنید ماتریسهای p,N,M دلخواه باشند و ماتریسهای B,A مربعی و V یک بردار باشد.

۱- M'را ترانهاده M گویند هرگاه جای سطر و ستون M عوض شود به عبارت دیگر $M` =(a_{ij})`=a_{ji}$

2 - $(M+N)'=M'+N'$, $(MN)'=N'M'$, $(M')'=M$

۳- دو ماتریس M, N جمع پذیر ند (جبری) هر گاه هم اندازه (دارای ابعاد یکسان) باشند

۴- ضرب ماتریس : P, N ضرب پذیرند هر گاه تعداد سطرهای P با تعداد ستونهای N یکسان (برابر) باشند

$N_{lk}*P_{ks}=C_{ls}=(\sum_k n_{lk}P_{ks})$, $Np \neq pN$

5 - $P(M+N)=PM+PN$

6 - اگر A متقارن باشد آنگاه A' نیز متقارن است و$A=A'$

۷- ماتریسهای AA' و $A'A$ متقارن هستند و بطورکلی$P'P, PP'$ متقارن هستند.

۸- طول بردار V را برابر با فاصله مبدا تا نقطه V گویند و به صورت زیر تعریف می شود.

$$\text{طول } V = \sqrt{V'V} = \sqrt{\sum_{i=1}^{p} v_i^2}$$

۹- ضرب عدد (اسکالر) در ماتریس

$CM=(cm_{ij})$

- 10 تمرین (ثابت کنید)

$(PM)'=M'P'$

- 11 **ماتریسهای افراز شده:**

ماتریس A را می توان به زیر ماتریسهایی چون $A_{11}، A_{12}، A_{21}، A_{22}$ افراز کرد بطوریکه زیر افرازها می توانند مربع یا مربع مستطیل باشند.

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

مثال $A=\begin{pmatrix} 2 & 3 & 5 & 8 & 4 \\ 4 & -3 & 0 & 7 & 2 \\ 9 & 3 & 6 & 5 & -2 \\ 3 & 1 & 2 & 1 & 6 \end{pmatrix}$

۱۲- برای تعریف رتبه یک ماتریس ابتدا مفهوم استقلال خطی و وابستگی را تعریف می کنیم.

مجموعه ای از بردارهای $a_1,a_2,...,a_n$ را وابسته خطی گوییم اگر اسکالرهای $c_1,....,c_n$ (که همگی باهم صفر نیستند) پیدا شوند بطوریکه $c_1a_1+....+c_na_n=0$ (E_1)

چنانچه رابطه فوق فقط به ازای تمام c_i ها تواما صفر حاصل شود آنگاه بردارهای $a_1,...,a_n$ را مستقل خطی گویند . به بیان دیگر ستونهای ماتریس A مستقل خطی هستند اگر $AC=0$ فقط به ازای $C=0$ حاصل شود (اگر مجموعه ای از بردار ها شامل صفر باشند این مجموعه وابسته خطی هستند).

چنانچه رابطه (E_1) برقرار باشد آنگاه حداقل یکی ازبردارهای a را می توان بصورت یک ترکیب خطی از بردارهای دیگر مجموعه نوشت. درمیان بردارهای مستقل خطی بردارهای زایدی به این شکل وجود دارد.

رتبه هرماتریس مربع یامستطیل A بصورت زیر تعریف می شود: (Rank)

{تعداد ستونهای مستقل A }=رتبه(A)

{تعداد سطرهای مستقل A}=رتبه(A)

ثابت می شود که تعداد ستونهای مستقل هر ماتریس برابر تعداد سطرهای مستقل خطی آن است.

تذکر۱: اگر ماتریس A فقط دارای یک عضو مخالف صفر باشد آنگاه Rank(A)=1، یعنی بردار0 و ماتریس0 دارای رتبه صفرند.

ماتریس دلخواه N با $n\times p$ رادرنظربگیرید N را دارای رتبه کامل (full rank) گویند هرگاه $R(N)=\min\{n, p\}$ باشد.

تذکر ۲- درماتریس مستطیلی N سطرها یا ستونها یا هر دو وابسته خطی هستند.

مثال: ملاحظه می شود که سطرهای N مستقل خطی هستند $N=\begin{pmatrix}1 & -2 & 3\\ 5 & 2 & 4\end{pmatrix}$

اما ستونهای آن وابسته خطی اند مطلوبست تعیین بردارC بطوریکه AC=0

می دانیم که $A \neq o, C \neq 0$

$$C=\begin{pmatrix}C_1\\ C_2\\ C_3\end{pmatrix}=\begin{pmatrix}14\\ -11\\ -12\end{pmatrix}$$

این رابطه رامی توان برای حاصل ضرب ماتریسها بسط داد .

$A=\begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix}$

$B=\begin{pmatrix}2 & 6\\ -1 & -3\end{pmatrix}$

AB=0

اگر $C=\begin{pmatrix}2\\ -1\end{pmatrix}$ فرض شود، داریم

AB=AC=O پس می توان نتیجه گرفت که ($B \neq C$)

در حالت کلی از رابطه AB=CB نمی توان نتیجه گرفت که A=C؟

مثال:$A=\begin{pmatrix}1 & 3 & 2\\ 2 & 0 & -1\end{pmatrix}, B=\begin{pmatrix}1 & 2\\ 0 & 1\\ 1 & 0\end{pmatrix}$ و $C=\begin{pmatrix}2 & 1 & 1\\ 5 & -6 & 4\end{pmatrix}$

ملاحظه می شود که $A \neq 0 \neq C$ ولی داریم: $A \neq C$; $AB=CB=\begin{pmatrix} 3 & 5 \\ 1 & 4 \end{pmatrix}$

ماتریس ناویژه: ماتریس A (مربعی) بارتبه کامل را ناویژه گویند. ماتریس ناویژه A دارای معکوس منحصر به فرد است که آنرا با A^{-1} نشان می دهند و داریم : $AA^{-1}=A^{-1}A=I$ ،در صورتیکه A دارای رتبه کامل نباشد آنرا ویژه نامند و معکوس ندارد.

تذکر۳- ماتریس مستطیلی بار تبه کامل نیز معکوس پذیر نیست.

قضیه ۱- اگر ماتریسهای A, B سازگار باشند آنگاه

$R(AB) \leq R(A)$

(اگر AB تعریف شده باشد)

$\leq R(B)$

$\leq \min\{R(A), R(B)\}$

قضیه ۲- فرض کنید ماتریسهای B, C ناویژه باشند آنگاه ضرب آنها در ماتریس A ، R(A) را تغییر نمی دهد یعنی $R(AB)=R(CA)=R(A)$

نتیجه ۱- برای هرماتریس A: $R(A'A)=R(AA')=R(A')=R(A)$

برهان: به مدلهای خطی برای آمار(الوین –رنچر) ترجمه بزرگ نیا مراجعه شود

اگرB ناویژه باشد و از AB=CB می توان نتیجه گرفت که $A= C$.

برهان: از طرف راست معادله فوق را در B^{-1} ضرب می کنیم.

چنانچه ماتریس B مستطیلی باشد یا ویژه باشد آنگاه نمی توان تساوی A, C را نتیجه گرفت.

اگر A ناویژه باشد آنگاه دستگاه معادلات AX=C دارای جواب منحصر بفرد است.

قضیه ۳- اگر A ناویژه باشد آنگاه A' نیز ناویژه است و معکوس آن (A') بصورت زیر محاسبه می شود

$(A')^{-1}=(A^{-1})'$

اگر A, B ناویژه باشند بااندازه های مساوی آنگاه AB ناویژه است و $(AB)^{-1}=B^{-1}A^{-1}$.

حالتهای خاص و معکوس آنها:

فرض کنید A متقارن و ناویژه باشد و بصورت زیر افراز شده باشد داریم:

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

اگر $B=A_{22}-A_{21}A_{11}^{-1}A_{12}$ آنگاه به شرط آنکه A_{11}^{-1} , B^{-1} وجود داشته باشند معکوس آن به صورت:

$$A^{-1}=\begin{pmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}B^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}B^{-1} \\ -B^{-1}A_{21}A_{11}^{-1} & B^{-1} \end{pmatrix}$$

به عنوان حالت خاص فرض کنید $A=\begin{pmatrix} A_{11} & a_{12} \\ a'_{12} & a_{22} \end{pmatrix}$

که درآن A_{11} مربع و a_{22} اسکالر و a_{12} بردار است .

اگر A_{11}^{-1} وجود داشته باشد آنگاه A^{-1} را می توان نوشت:

$$A^{-1} = \frac{1}{b}\begin{pmatrix} bA_{11}^{-1} + A_{11}^{-1}a_{12}a'_{12}A_{11}^{-1} & -A_{11}^{-1}a_{12} \\ -a'_{12}A_{11}^{-1} & 1 \end{pmatrix}$$

که درآن $b = a_{22} - a'_{12}A^{-1}_{11}a_{12}$ است.

اگر یک ماتریس مربع بصورت $B+cc'$ ناویژه باشد که در آن C یک بردار و B یک ماتریس ناویژه است

آنگاه:

$$(B+CC')^{-1} = B^{-1} - \frac{B^{-1}CC'B^{-1}}{1+C'B^{-1}C}$$

تعریف : ماتریس متقارن A را معین مثبت گویند هرگاه به ازای تمام مقادیر ممکن $Y \neq 0$ داشته باشیم:

$y'Ay > 0$ (صورت درجه دوم مثبت)

اگر برای تمام مقادیر $y \neq 0$ داشته باشیم $y'Ay \geq 0$ آنگاه $y'Ay$ و A را نیمه معین مثبت گویند.

مثال: برای معین مثبت $A=\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ که

$$y'Ay = 2y_1^2 + 3y_2^2 - 2y_1y_2$$

$$= 2(y_1 - \frac{1}{2}y_2)^2 + \frac{5}{2}y_2^2 > 0$$

مثال ۲- برای تشریح نیمه معین مثبت

$$(2y_1 - y_2)^2 + (3y_1 - y_3)^2 + (3y_2 - 2y_3)^2$$

و

$A=\begin{pmatrix} 13 & -2 & -3 \\ -2 & 10 & -6 \\ -3 & -6 & 5 \end{pmatrix}$ ، واضع است که اگر

$2y_1 = y_2$, $3y_1 = y_3$, $3y_2 = 2y_3$ فرض شود داریم:

y_1, y_2, y_3 مخالف صفرند درصورتیکه $y'Ay=0$

$y=(1,2,3)$ داریم: یعنی برای هر مضرب $y'Ay=0$

در غیر اینصورت $y'Ay>0$ مگر $y=0$ باشد.

قضیه ۴- اگر A معین مثبت باشد آنگاه تمام اعضاء قطری A یعنی a_{ii} مثبت هستند.

اگر A نیمه معین مثبت باشد آنگاه تمام اعضاء قطری Aیعنی $a_{ii} \geq 0$ مثبت هستند.

برهان: فرض کنید $y' =(0,\ldots,1,\ldots,0)$ که فقط مولفه iام آن 1 است در نتیجه :

$$y'Ay = a_{ii}>0$$

پس باید داشته باشیم:

ادامه بحث به کتاب مدلهای خطی برای آمار، صفحه ۴۶ موکول می شود.

دترمینان یک ماتریس:

دترمینان یک ماتریس مربع $n*n$ مانندA یک تابع اسکالر از A است.

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

Det(A)=$|A|=a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}-a_{23}a_{32}a_{11}-a_{33}a_{12}a_{21}$

قضیه۴: اگر D یک ماتریس قطری diagonal باشد:

D=$\begin{pmatrix} d_1 & \dots & 0 \\ & \cdots & \\ & \cdots & \\ 0 & \dots & d_n \end{pmatrix} \Longrightarrow det(D)=|D|=\pi_{i=1}^{n} di$

۲- اگر A یک ماتریس مثلثی باشد دترمینان آن برابر حاصل ضرب اعضای قطر اصلی است .

۳- اگر A ویژه باشد آنگاه det (A)=0 و اگر A ناویژه باشد det(A)$\neq 0$

۴- اگر A معین مثبت باشد آنگاه det(A)>0

۵- همیشه داریم $|A|=|A||A'|$

6- اگرA ناویژه باشد $\det(A)^{-1}=\frac{1}{|A|}$

۷- اگر C=diag(c,...,c) آنگاه $C=cI_n$, $\det(C)=c^n$

نیز A=CB$\Longrightarrow \det(A)=c^n|B|$

فرض کنید A مربعی و$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ A_{11} A_{22} ناویژه باشند و

$=|A_{11}||A_{22}-A_{2i}A_{11}^{-1}A_{12}||A|$

$=|A_{22}||A_{11}-A_{12}A_{22}^{-1}A_{21}||A|$

حالت خاص: $\begin{vmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{vmatrix}$=$a_{11}a_{22}$- $a_{12}a_{21}$

اگر $A=\begin{pmatrix}A_{11} & 0\\ A_{21} & A_{22}\end{pmatrix}$ که در آنA_{11} , A_{22} مربعی نه لزوما به یک اندازه هستند .

$$A=\begin{pmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{pmatrix}$$

آنگاه$|A|=|A_{11}||A_{22}|$

قضیه ۵- اگر A , B ماتریسهای مربعی به یک اندازه باشند

1- $=|A||B|$

$|AB|$

2- $|AB|=|BA|$

3- $|A^2|=|A|^2$

2 -1- 0 بردارها و ماتریسها:

دو بردار a, b را متعامد (عمود بر هم) گویند هر گاه حاصل ضرب داخلی آنها صفر باشد $a'b=0$

اگر θ کسینوس بردارهای a , b باشد، ملاحظه می شود که زمانی که $\theta=90$ باشد

$COS(\theta)=\frac{a\prime b}{\sqrt{(a'a)(b'b)}}$

بردار نرمال: بردار a را نرمال شده گویند هر گاه $a'a=1$ باشد لذا هر بردار دلخواه C را می توان با تقسیم بر طولش یعنی $a=\frac{C}{\sqrt{C\prime C}}$ نرمال کرد.

حال اگر مجموعه ای از بردارهای نرمال $c_p,.....c_1$ که $ci'ci=1$ و $(i\neq j)\ ci'cj=0$ آنگاه این مجموعه را مجموعه متعامد یکه از بردارها گویند. حال اگر ماتریس C دارای ستونهای متعامد یکه باشند انگاه C یک ماتریس متعامد با خاصیت $C'C=I$ است می توان نشان داد که $CC'=I$ برقرار است.

یعنی یک ماتریس متعامد دارای سطرها و ستونهای متعامد یکه می باشد. از رابطه فوق ملاحظه می شود که $C'=C^{-1}$ اگر C متعامد باشد.

قضیه۶- اگر $C_{P,P}$ متعامد باشد و یک $A_{P,P}$ دلخواه باشد آنگاه :

۱- $|C|$ یا $+1=-1$

۲- $|C'AC|=|A|$

۳- $-1\leq c_{ij}\leq 1$ که در آن c_{ij} عضوی دلخواه از C است.

اثر(Trace) یک ماتریس مربع A بصورت زیر تعریف می شود

$Tr(A)=\sum_{i=1}^{n} a_{ii}$

قضیه ۷- اگر A , B ماتریسهای n در n باشند آنگاه:

۱- $\mathrm{Tr}(A \pm B) = Tr(A) \pm Tr(B)$

۲- اگر $A_{n.P}$, $B_{P.n}$ باشد آنگاه

$\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$

۳- اگر A_{np} باشد آنگاه : $\mathrm{Tr}(A'A)=\sum_{j=1}^{P} a'_j a_j$

که در آن a_j سطر j ام A است.

۴- $\mathrm{Tr}(A'A)=\mathrm{Tr}(AA')=\sum_{i=1}^{n}\sum_{j=1}^{p} a^2{}_{ij}$

۴- اگر A یک ماتریس n.n , P_{nn} یک ماتریس ناویژه باشد آنگاه $\mathrm{Tr}(p^{-1}AP)=\mathrm{Tr}(A)$

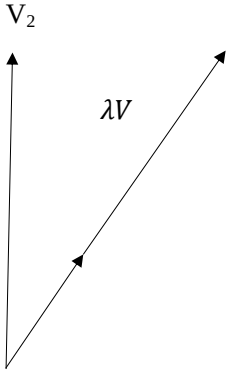
۶- اگر A یک ماتریس n.n , c یک ماتریس متعامد n.n باشد آنگاه:

$\mathrm{Tr}(c'Ac)=\mathrm{Tr}(A)$

بردار و مقدار ویژه یک ماتریس:

برای هر ماتریس مربع A یک مقدار اسکالر λ و یک بردار غیر صفر V می توان یافت بطوریکه $AV=\lambda V$

و λ مقدار ویژه و V را بردار ویژه ماتریس A نامند که گاهی به ترتیب ریشه و بردار مفسر نامیده می شوند.



v

اگر $AV=\lambda V$ را بصورت $\lambda V = (A - \lambda I)V = 0$ -AV

یعنی A- $\lambda I)V$ یک ترکیبی خطی از A- λI است یعنی A- λI یک ماتریس ویژه و داریم

$|A - \lambda I| = 0$ که از حل آن می توان λ مقدار ویژه را بدست آورد و معادله فوق را معادله مفسر یا مشخصه می نامند.

اگر A n*n باشد آنگاه معادله فوق دارای n ریشه خواهد بود یعنی $\lambda_1, \dots, \lambda_n$ که λ_i ها لزوما هم متمایز یا مخالف صفر یا حقیقی نیستند.

تذکر۴- مقادیر ویژه یک ماتریس متقارن حقیقی هستند.

توابعی از یک ماتریس :صحفه71

قضیه12.2 : فرض کنید A یک ماتریس متقارن n*n باشد:

۱- مقادیر ویژه $\lambda_1, \dots, \lambda_n$ حقیقی هستند.

۲- بردارهای ویژه $x_1 \dots x_n$ ماتریس A تواما متعامد هستند یعنی $x'_i x_j = 0$; $v \neq j$

۳- انگاه A را می توان نوشت : $A=CDC'=\sum_{i=1}^{n} \lambda_i x_i x_i'$

که در آن $D=diag(\lambda_1, \dots, \lambda_n)$ و $C=(x_1,\dots,x_n)$ ماتریس متعامد است و رابطه اخیر را تجزیه طیفی ماتریس A گویند.

۴- آنگاه A را می توان قطری کرد: $C'AC=D$

-۵

$$|A| = \prod_{i=1}^{n} \lambda i \ , \ tr(A) = \sum_{i=1}^{n} \lambda_i =$$

ماتریس خود توان :A را ماتریس خود توان نامند هرگاه $A^2=A$

قضیه ۱۳,۲- تنها ماتریس ناویژه خود توان ماتریس همانی I است

چون $A^2=A \Longrightarrow A = AA^{-1} = I$

قضیه ۲- ۱۳ب: اگر A یک ماتریس ویژه متقارن خود توان باشد آنگاه A نیمه معین مثبت است.

مشتق توابع خطی و صورتهای درجه دوم:

فرض کنید $u=f(x)\ ;\ x'=(x_1,\dots x_p)$ آنگاه:

$$\partial u/\partial x = \begin{pmatrix} \partial u/\partial x_1 \\ \vdots \\ \partial u/\partial xp \end{pmatrix} \ ; \quad \partial u/\partial xi \quad , \quad i = 1, \dots, p$$

مشتقات جزئی هستند.

اگر u= a′x باشد آنگاه $\partial u/\partial x = a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}$ ،چون ستونی است پس مشتق آن ستونی خواهد بود.

قضیه۲,۴ت:فرض کنید u=x′Ax که در آن A یک ماتریس متقارن از ثابتها باشد آنگاه

$$\partial u/\partial \mathrm{X} = \frac{\partial(\mathrm{X'AX})}{\partial \mathrm{X}} = 2\mathrm{AX}$$

اثبات : نشان میدهیم که برای $x'=[x_1,x_2,x_3]$

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

$x'Ax= a_{11}x_1^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + a_{22}x_2^2 + 2a_{23}x_2x_3 + a_{33}x_3^2$

$$=2a_i'x \Longrightarrow \frac{\partial(x'Ax)}{\partial x}=2Ax\,\frac{\partial(x'Ax)}{\partial X_i}$$

که a_i' سطر iام A است.

1 -2 -0 یادآوری متغیرهای تصادفی و توزیع ها:

متغیر تصادفی : تابعی از فضای نمونه به مجموعه اعداد حقیقی است.

مثال۱- دو سکه نا اریب را پرتاب می کنیم، متغیر تصادفی Y را برابر ۱ تعریف می کنیم اگر هر دو سکه یکسان ظاهر شوند در غیر اینصورت ۳ تعریف می شود.

X .م.ت. عدد $\frac{1}{3}$ می گیرد اگر حداقل یک شیر بیاید در غیر اینصورت $\frac{8}{3}$ می گیرد

Result	Y	X	Z
HH	1	$\frac{1}{3}$	$\frac{4}{3}$
TH	3	$\frac{1}{3}$	$\frac{10}{3}$
HT	3	$\frac{1}{3}$	$\frac{10}{3}$
TT	1	$\frac{8}{3}$	$\frac{11}{3}$

می توان متغیر تصادفی Z را اینطور تعریف کرد: Z=X+Y

تعریف۱: متغیر تصادفی X گسته است اگر مقادیری که می پذیرد شما را باشد (متناهی یا نا متناهی)

تعریف۲- متغیر تصادفی X پیوسته است اگر مقادیری که می پذیرد شامل یک یا چند فاصله یا اجتماعی از چند فاصله باشد (مجموعه مقادیر X نا شمارا باشد)

تابع توزیع: فرض کنید X یک متغیر تصادفی باشد

$F(x)=p(X\leq x)$

تابع احتمال: فرض کنید X گسسته باشد

$f(x)=p(X=x)$ p.f

تابع چگالی احتمال: فرض کنید X پیوسته باشد.

$f(x)=\frac{\partial F(x)}{\partial x}$, $p(a\leq X\leq b)=\int_a^b f(x)dx\ ;\ a<b$

$F(x)=\int_{-\infty}^{x} f(t)dt$

فرض کنید X_1,….,X_n متغیر تصادفی باشند تابع توزیع توام cdf آنها بصورت زیر تعریف می شود.

$F(x_1,\dots.,x_n)=p(X_1\leq x_1,\dots X_n\leq x_n)$

$f(x_1,\dots,x_n)=\frac{\partial^n F(x_1,\dots x_n)}{\partial x_1\dots\dots\dots\partial x_n}$

تابع توزیع شرطی چند گانه:

$F(x_1,\dots x_k \mid X_{k+1}\in A_{k+1},\dots..,X_n\in A_n)=\frac{p(X_1\leq x_1,\dots,X_k\leq x_k,X_{k+1}\in A_{k+1},\dots.,X_n\in A_n)}{p(X_{k+1}\in A_{K+1},\dots X_n\in A_n)}$

اگر شرطهای فوق اثری روی بقیه متغیرها نداشته باشند آنگاه آنها را دو به دو مستقل گویند.

تعریف: x_n,……..,x_1 را مستقل از یکدیگر گویند هرگاه:

$F(x_1,\dots.,x_n)=F_{x1}(x_1)\dots\dots F_{xn}(xn)\ ;\ \ \forall x_1,\dots.,x_n$

تذکر: توابع $F_{Xn}(x_n)$….,$F_{X1}(x_1)$ را cdf های کناری گویند.

1- $f(x)=\int f(x,y)dy,$

2 − $f(u,v)=\iint f_{u,v,x,y}(u,v,x,y)dxdy$

3- $F_X(x)=F_{X,Y}(x,+\infty)$

4- $f_{x|y}(x|y)=\frac{f(x,y)}{fy(y)}=\frac{f(x,y)}{\int f_{x,y}(x,y)dx}$

5- $f(x,y)=f_x(x)f_y(y)$; $x \perp y$

فرض کنید X,Y متغیرهای تصادفی باشند، آنگاه :

$\mu_x=E(X)=\int xf(x)dx$

$V(x)=\sigma^2{}_X=E(X-\mu_X)^2=EX^2-\mu_X^2$

$\sigma_{x,y}=cov(x,y)=E(X-\mu_x)(Y-\mu_y)$

تابع مولد گشتاور X : $M_X(t)=E(e^{tX})$

فرض کنید **X** یک بردار از متغیر های تصادفی باشد

$X=(X_1,\ldots,X_n)'$

$$E(X)=\begin{pmatrix} EX_1 \\ \vdots \\ EX_n \end{pmatrix}$$

$$V(X)=\begin{pmatrix} V(X_1) & COV(X_1,X_2) & \ldots & COV(X_1,X_n) \\ COV(X_2,X_1) & V(X_2) & . & \ddots \\ \vdots & . & \ddots & . \\ COV(X_n,X_1) & . & \ldots & V(X_n) \end{pmatrix}$$

تابع مولد گشتاور توام:

$M(t_1,\ldots,t_n)=E[\exp(t_1x_1+\ldots.+t_nx_n)]$

امید ریاضی شرطی :

$E(x|A)=\int xf(x|A)dx$

چند خاصیت مفید دیگر:

1- $E(g(x))=\int g(x)f(x)dx$

2- $Cov(x,y)=E(xy) - E(x)E(y)$

3- $X\perp Y \Longrightarrow E(XY)=E(X)E(Y) \Longrightarrow cov(X,Y)=0$

4- $E(x^k)=d^kM_x(t)/\ dt^k\ |t=0 \quad ; k\geq 1$

5- $V(ax+by)=a^2v(x)+b^2v(y)+2ab\ cov(x,y)$

6- $M_{X1,\dots Xn}(t_1,\dots t_n)=E(\ e^{t_1x_1+\cdots+t_nx_n})$

7- $E(g(x))=E_Y(E(g(x)|y))$

8- $V(g(x))=Ey[v(g(x)|y)]+Vy[E(g(x)|y)]$

9- $\sum = V(X) \Longrightarrow$

$\sum$ یک ماتریس متقارن و نامنفی است.

10- $\mu = E(X)\ ,\ \sum = V(X)$

و گر M ماتریسی با عناصر ثابت باشد، آنگاه

$E(MX)=ME(X)=M\mu$

$V(MX)=M\sum M'$

3 -2 -0 توزیع های مهم :

۱- نرمال:

$$X\sim N(\mu,\sigma^2)$$

$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{1}{2\sigma^2}\ (x-\mu)^2\}$ $,x\epsilon R$

$\theta=\{(\mu,\sigma^2)|-\infty<\mu<\infty\ ,\sigma^2>0\}$

$M_x(t)=exp\{\mu t+{\sigma^2t^2}/{2}\}$

2- کی دو یا خی دو یا Chi- squari

$X\sim\chi_k^2$

$$f_x(x)=\frac{1}{\Gamma\left({k}/{2}\right)2^{{k}/{2}}}\ x^{({k}/{2}-1)}\exp\left(-{x}/{2}\right)\ ;\ x>0\ ,k>0\Longrightarrow$$

$E(X)=K,\ V(X)=2K\quad,\quad M_x(t)=(1-2t)^{-{k}/{2}}$

به بیان دیگر می توان:

If $X_1,\ldots X_{n,}$; $X_i\sim N(0,1)$

یک نمونه تصادفی باشند

If $x=\sum_{i=1}^{n} x_i^2 \Longrightarrow x \sim X_n^2$

۳- توزیع استودنت : فرض کنید $x \sim t_k$

که در آن k درجه آزادی توزیع نامیده می شود

$$f(x)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{(k\pi)^{1/2}\Gamma(k/2)}(1+t^2/k)^{-\frac{k+1}{2}} \ , -\infty < x < \infty \ \& \ k>0$$

به بیان دیگر اگر

$Z \sim N(0,1)$,$V \sim X^2_K$, $X=\frac{Z}{\sqrt{V/K}}$

تعریف شود آنگاه: $X \sim t_k$

۴- توزیع فیشر$(x \sim F_{m,n})$

که در آن m, n درجه های آزادی توزیع نامیده می شود.

$$F(x)=\frac{\Gamma(\frac{m+n}{2})\Gamma(m/n)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(1+\frac{m}{n}x)^{\frac{(m+n)}{2}}} \ ; \ x>0 \ ; m,n>0$$

به بیان دیگر

$x=\frac{y/m}{z/n}$, $y \sim x_m^2$, $z \sim x_n^2 \Longrightarrow x \sim F_{m;n}$

۵- توزیع چند متغیره نرمال: $x \sim N_n(\mu, \sum)$

$f_x(x)=\frac{1}{\sqrt{(2\pi)^n \|\sum\|}} exp\{-\frac{1}{2}(x-\mu)'\sum^{-1}(x-\mu)\}$

که در آن $\|\sum\|$ نماینده قدر مطلق دترمینان ماتریس واریانس - کواریانس $\sum$ است

$$\mu = E(X), \sum = V(X), V = (V_1, \ldots . V_n)'$$

$$E(e^{v'x}) = m_x(v) = \exp\{\mu' v + \frac{v'\sum v}{2}\}$$

۴- ۲- ۰ استنباط – آماری Statistical Infrence

۱- point estimation

فرض کنید $x_1,\ldots..x_n$ یک نمونه تصادفی باشد و θ یک پارامتر مجهول جامعه باشد اگر $\widehat{\theta} = \widehat{\theta}(x_1,\ldots x_n)$

یک برآوردگر θ باشد آنگاه خواص زیر را داریم:

a) $\widehat{\theta}$ را نااریب گویند هرگاه (unbiased)

$$E(\widehat{\theta}) = \theta$$

b) Efficiency (کارایی , موثر) یا برآوردگر کارا efficient estimator

یعنی $\widehat{\theta}$ دربین تمام برآوردگرهای نااریب دارای کمترین واریانس است به عبارت دیگر $v(\widehat{\theta})$ می نیمم در بین تمام بر آوردگرها برای (θ).

c - همگرایی در احتمال:

$$\forall\ \varepsilon \geq 0 \quad \rightarrow \quad \lim_{n\to\infty} p(|\theta - \widehat{\theta}| \leq \varepsilon) = 1$$

به بیان دیگر

$$\lim_{n\to\infty} Var(\hat{\theta}) = 0$$

روش درستنمایی ماگزیموم (MLE): $\max \{f(x_1, \cdots, x_n)\}$

برآوردگر های این روش حداقل دو خاصیت اخیر را دارند؟

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + [bias(\hat{\theta})]^2 \qquad -D$$

$Bias(\hat{\theta})=E(\hat{\theta}) - \theta$

فرض کنید $x_1, \dots x_n \sim^{iid} f(\mu, \sigma^2)$

$$\overline{X} = \frac{1}{n}\sum_{1}^{n} xi \quad , \quad s_n^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x}_n)^2$$

آنگاه

1- $E(\bar{x}_n)=\mu_1$, $v(\bar{x}n) = \sigma^2/n$

2- $E(s_n^2) = \sigma^2$, $V(S_n^2) = \frac{2\sigma^4}{n-1}$

3- if $x_1, \cdots, x_n \sim N(\mu, \sigma^2)$

آنگاه: $\bar{x}_n$, s_n^2 مستقل از یکدیگرند و

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) , \frac{(n-1)s_n^2}{\sigma^2} \sim X^2_{n-1}$$

$$\frac{\bar{x}n-\mu}{\sqrt{\frac{S_n^2}{n}}}\sim t_{(n-1)}$$

بر آوردگرهای فاصله ای اطمینان : Confidence Interval Estimators

فاصله تصادفی $(\hat{\theta}_1\ (x_1,\cdots,x_n),\ \hat{\theta}_2\ (x_1,..x_n))$ را یک فاصله اطمینان با ضریب

$100\times(1-\alpha)\%$ برای θ گویند هر گاه:

$$P(\hat{\theta}_1\leq\theta\leq\hat{\theta}_2)=1-\alpha$$

آزمون فرض آماری: Statistical Hypothesis Testing

α سطح معنی داری از قبل معلوم می باشد، $1-\alpha$ را ضریب اطمینان نامیم.

p- value : سطح معنی دارای مشاهده شده (حاصل) از داده ها. فرض H_0 دلالت بر درستی شرایط قبل دارد و فرض H_1، یا جایگزینalternativ دلالت بر تغییر شرایط پیشین یا وجود شرایط جدید و متفاوت از قبل دارد.

۱- مشاهده آماره آزمون

۲- محاسبه احتمال رد فرض H_0 در مقابل H_1 یا محاسبه سطح معنی داری مشاهده شده که آنرا با α^* نشان می دهند و p-value می نامند.

۳- رد یا عدم رد فرض پیشین H_0 (تصمیم در مورد فرض)

قانون انگشتی Rule of Thumb برای تصمیم گیری در رد/قبول فرضیه‌ها:

Decision	α^* سطح معنی داری مشاهده شده
دلیلی بر رد H_0 نیست	$\alpha^* \geq 0.10$
به اکراه فرض H_0 را رد می کنیم	$0.05 \leq \alpha^* < 0.1$
فرض H_0 رد می شود	$0.01 \leq \alpha^* < 0.05$
قویاً فرض H_0 رد می شود	$\alpha^* < 0.01$

فصل ۱ - ۱ یادآوری رگرسیون خطی ساده :

هدف رگرسیون : تعیین رابطه بین متغیر های آزاد و متغیر پاسخ است که در زمینه های گوناگون کاربرد دارد و به منظور پیش بینی مقدار پاسخ با استفاده از متغییرهای آزاد می باشد.

مثال:

۱- مطالعه اثر در آمد روی امید زندگی یا طول عمر در کشورهای گوناگون.

۲- اثر لایه ازون ozone روی سلامتی و بیماریهایی پوستی روی انسان.

۴- اثرآلاینده های صوتی روی سلامت روانی افرادجامعه(اعصاب).

۵- اثر میزان مطالعه روی نمرات پایانی (موفقیت دانشجو).

هدف از آنالیز رگرسیون مطالعه رابطه بین متغیرهای موجود یا فاکتورهای اندازه پذیر پس از مشاهده مقدار متغیرها می باشد و موضوعات مطرح در آنالیز رگرسیون عبارتند از:

۱- تعیین مدل رگرسیون

۲- بر آورد پارامترها

۳- انتخاب متغیرهای موثر (اصلاح الگو)

۴- پیش بینی

فرض کنید $x_1, \ldots, x_p$ متغیرهای آزاد (فاکتور –توصیفی –کمکی) و y متغیر پاسخ باشند.

اگر مدل یا رابطه بین $x_1, \cdots, x_p$ و y تعیینی باشد (دقیق) می توان آنرا این‌طور نوشت

$$Y = f\ (x_1, \cdots, x_p)$$

ولی در عمل همیشه این‌طور نیستند یعنی خطاهای اندازه گیری یا عوامل غیر قابل کنترل دیگر دخالت می کنند که اجبارا مدل را باید با خطای ناشی از آن بصورت زیر نوشت:

تعریف۱- ۱ $$y = f\ (\underline{x}, \beta) + \text{Random flactuation} = f(\underline{x}, \beta) + \underline{\varepsilon}$$

در واقع هدف آنالیز رگرسیون پیدا کردن تابعی تقریبی برای f است که بخشی از y توسط f و بخش کوچکی از y توسط ε تعین می شود که در کل می توان رگرسیون خطی چندگانه را بصورت زیر نوشت :

$$y = f\ (\underline{x}, \beta) \quad (1.2)$$

یعنی تغیرات y بطور کامل با x بیان می شود $g(y)$

تعریف رگرسیون خطی : $y=\beta_0+\beta_1 f_1(x_1)+\cdots+\beta_p f_p(xp)+e$

که در رگرسیون خطی ساده داریم:

$Y=\beta_0+\beta_1 x_1+\varepsilon$

$Y_i=\beta_0+\beta_1 x_{1i}+\varepsilon_i \ ; i=1,2,\ldots,n$

تذکر۱: توابع g, f معلومند و رابطه فوق خطی نامیده می شود هر گاه رابطه مذکور نسبت به $\underline{\beta}$ خطی باشد.

به بیان دیگر مشتق رابطه فوق نسبت به $\underline{\beta}$ ، نباید تابعی از β باشد.

چند مثال دیگر از مدلهای خطی:

$Yi=\beta' x_i+\varepsilon_i \ ; i=1,2,\ldots,n$

$Lny_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\varepsilon_i,\ldots.$

$=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_{12} x_{i1}\, x_{i2}+\beta_3 e^{x_{i1}}+\varepsilon_i\sqrt{y_i}$

مثالی از مدل غیر خطی ($nonlinear$) $y_i=\beta_0\exp\{-\beta_1 x_i^{\beta 2}\}+\varepsilon_i$

تعریف ۱,۲: مدل خطی تعمیم یافته یک مدل رگرسیون است که در آن y متغیر پاسخ endogene از توزیع خانواده نمایی پیروی کند(قبل از این از توزیع) بطوریکه یکی از پارامترهای کانونیک یک تابع خطی از β باشد.

مثال ۱,۱ $yi \sim b(1, \theta_i)$, $\eta_i = \ln\left(\frac{\theta i}{1-\theta i}\right) x_i = \beta' x_i$

بعدا مطالعه و بررسی خواهد شد.

خانواده نمایی:

$F(x|\theta) = h(x)\ g(\theta)\ \exp(\eta(\theta)T(x))$

$= h(x)\ \exp(\eta(\theta)T(x) - A(\theta))$

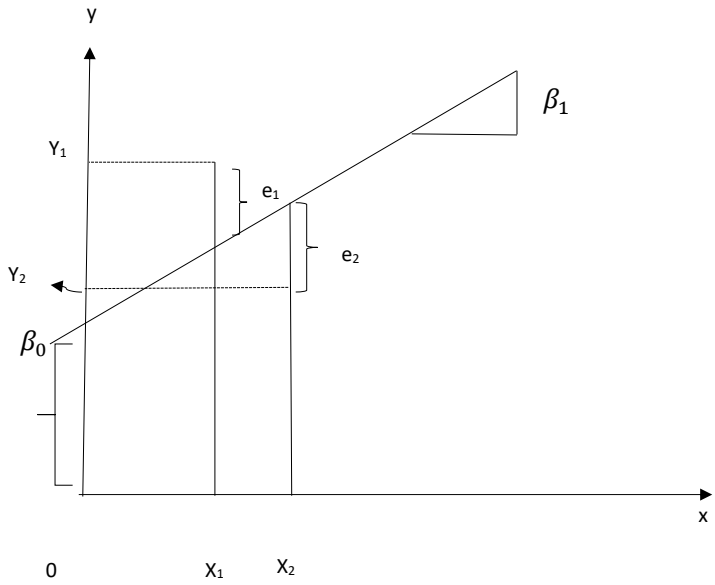
$= \exp\{\eta(\theta)T(x) - A(\theta) + B(x)\}$

تعریف ۱,۳- مدلی را که به فرم زیر باشد

$\theta(y_i) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i \ ; i = 1,2,\dots,n$

را یک مدل جمعی (additive) گویند که در آن توابعی $f_1, \dots, f_p$ دلخواه هستند که پس از مشاهده داده ها باید برآورد شوند. θ می تواند معلوم باشد یا توسط داده ها بر آورد شود.

مثال ۱,۲-



تذکر: مدل جمعی additiv یک مدل یا ابزار مناسب برای بر آورد f در مدلهای خطی می باشد.

مدل رگرسیون ساده:

$$Y_i=\beta_0+\beta_1 x_i+\varepsilon i\ ;\ \ i=1,\dots.,n$$

فرضهای لازم:

خطی بودن: $$I)\ \ E(y_i;x_i)=\beta_0+\beta_1 x_i \Leftrightarrow E(\varepsilon_i)=0$$

هم واریانس یا یکنواختی واریانس εi: $$II)\ \ Var(\varepsilon_i)=\sigma^2;\ i=1,\dots,n$$

نا همبسته بودن: $$III)\ \ cov(\varepsilon_i,\varepsilon_j)=0; i\neq j\quad \varepsilon_1,\dots,\varepsilon_n$$

تفسیر پارامترها :

$$\beta_0 : \mathrm{E}(y, \mathrm{x} = 0) = \beta_0 + 0 * \beta_1 + 0 = \beta_0$$

میانگین y وقتی x=0 :

$$\beta_1 : E(y; x = x + 1) - E(y; x = x)$$

$$= \beta_0 + \beta_1(x + 1) + E(\varepsilon) - (\beta_0 + \beta_1 x) - E(\varepsilon) = \beta_1$$

میانگین تفاضل Y وقتی که x به اندازه یک واحد افزایش یابد برابر β_1 می‌باشد.

مثال ۱,۲- فرض کنید y وزن و x قد باشد اگر مدل بر آورد شده عبارت باشد از y = − 5+50x

سوال : میانگین وزن وقتی که قد صفر است چقدر است ؟

بطور میانگین اثر افزایش قد به اندازه ۱ سانتیمتر روی وزن y چقدر است 0. 5kg

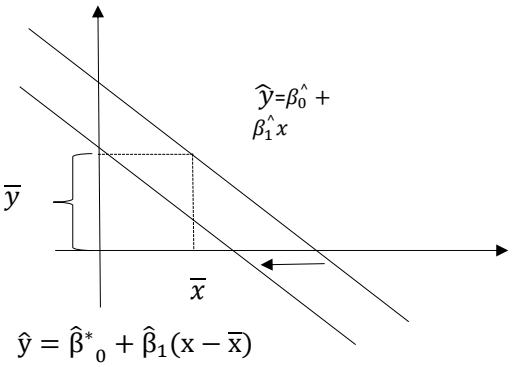
تذکر۱: در مثال فوق ملاحظه می شود که E(y|x=0) در دامنه y قرار ندارد.

می توان مدلی بصورت زیر نوشت که کاملا معادل مدل قبل است

$$Y_i = \beta_0^* + \beta_1(x_i - \overline{x}) + \varepsilon_i \; ; i = 1,2, \dots, n$$

$$\bar{X} = \frac{1}{n}\sum x_i \; , \qquad \beta_0^* = \beta_0 + \beta_1 \overline{x}$$

که با تفسیر β_1 هیچ تغیری نمی کند و



$$\hat{y} = \hat{\beta}^*_0 + \hat{\beta}_1(x - \overline{x})$$

$$E(y \mid x=\overline{x}) = \beta_0^* + \beta_1(\overline{x} - \overline{x}) + E(\varepsilon) = \beta_0^*$$

مثال ۱,۲- مدل فوق را دوباره به فرم جدید می نویسم:

$$\beta_0^* = -5 + 50 \ .1.80 = 85 \quad ; \ \overline{x} = 1.80m$$

$$Y_i = 85+50(x_i-1.8)+\varepsilon_i$$

حال β_0^* برابر میانگین وزن برای فردی که میانگین وزن آن برابر ۱,۸ متر است، می باشد. در حالت کلی داریم:

$$Or \quad \beta_0 = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y}$$

تذکر ۱,۲- واحد اندازه گیری x , y هیچ اثری روی روابط x, y ندارد اگر چه ظاهرا ممکن است مقدار $\hat{\beta}_1, \hat{\beta}_0$ تغییر کنند ولی

kg→ $pounds(2.2)$

cm→ m

yi =− 11+1.1xi+εi

1.2.2

برآورد پارامترها (روش میانگین توانهای دوم خطا MSE)

Yi=$\hat{\beta}_0 + \hat{\beta}_1 x_i$

Min $SS_{Res}(\hat{\beta}_0, \hat{\beta}_1) = min \sum (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)$

= min$\sum e_i^2$

$$\Longrightarrow \partial \frac{SS_{Res}(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_0} = 0, \qquad \partial \frac{SS_{Res}(\hat{\beta}_0,\hat{\beta}_1)}{\partial \hat{\beta}_1} = 0$$

$\hat{\beta}_1 = \frac{\sum y_i(x_i-\overline{x})}{\sum(x_i-\overline{x})^2} =$ واز قبل داریم

$\frac{s_{xy}}{s_{xx}}$

$\hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \bar{x} \Leftrightarrow \hat{\beta}_1^{\ *} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}_n$

مثال ۱,۳.- یک نمونه تصادفی 10 تایی از ماهی های صید شده در عمق های متفاوت و اندازه قد آنها موجود است. هدف بررسی این است که بدانیم آیا عمق محل زیست ماهی‌ها روی اندازه قد آنها تاثیر دارد یا خیر؟

Fish	Profond(depth)	length
1	9.96	10.7
2	7.98	13.1
3	11.8	7.5
4	7.23	12.9
5	9.06	12.2
6	8.39	10.8
7	11.8	8.9
8	7.69	12.0
9	10.2	9.55
10	10.5	10.0

۱- Scatleplot

$$\sum y_i = 107.71, \quad \sum xiyi = 994.32 \quad , \quad \sum x_i = 94.62$$

$$\sum x_i^2 = 919.82, \quad \sum y_i^2 = 1189.9$$

$$S_{xx} = \sum (xi - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = 24.526$$

$$S_{xy} = \sum y_i(x_i - \bar{x}) = \sum (y_i - \bar{y})(x_i - \bar{x}) = \sum x_i y_i - n\bar{x}\ \bar{y} = -24.83$$

$$\hat{\beta}_1 = \frac{sxy}{sxx} = -1.01\ , \hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x} = 20.3$$

Y_i = 20.3 - 1.01x_i = 10.77 - 1.01 (x_i - 9.462)

تفسیر پارامترها : اولا طول(قد) ماهیها با افزایش عمق آب کم می شود. یعنی به ازای هرمتر افزایش عمق آب بطور متوسط طول(قد) ماهیها cm1.01 کاهش می یابند و متوسط قد ماهیها وقتی که x=0 یعنی در سطح آب برابر 2o.3cm می باشد.

۱,۳- رگرسیون خطی چند گانه: Multiple Linear regression

۱,۳,۱- مدل رگرسیون خطی چند گانه : فرض کنید $x_1,...x_p$ متغیرهای آزاد و y متغیر پاسخ یا وابسته باشد.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \ ; i = 1, \dots, n$$

$$\begin{cases} y_1 = \beta_0 + \beta_1 xu + \cdots + \beta_p x_{ip} & + \varepsilon_1 \\ \vdots & \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} & + \varepsilon_n \end{cases}$$

فرم ماتریس مدل: $\boldsymbol{\varepsilon} = (e_1, \dots, e_n)'$ $\mathbf{y} = (y_1,\dots,y_n)'$

$\mathbf{1}=(1,\dots,1)'$ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$

$$\mathbf{X}=(\mathbf{1},\mathbf{x_1},\dots\dots\mathbf{x_p})=\begin{pmatrix}\boldsymbol{x'}_1\\ \vdots \\ \boldsymbol{x'}_n\end{pmatrix}=\begin{pmatrix}1 & x_{11} & \dots & x_{1p}\\ \vdots & \ddots & & \vdots \\ 1 & x_{n1} & \dots & x_{np}\end{pmatrix}$$

$X_i = (1,x_{i1},\dots. X_{ni})'$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (\mathbf{1.3.2})$$

این مدل براساس ماتریسها به فرم ماتریسی رگرسیون چند متغیره خطی می باشد. این فرم برای سادگی محاسبات تئوری ، آزمون فرضها و خواص بر آوردگرها می باشد .

فرضهای حاکم بر مدل چند متغیره خطی مثل فرضهای لازم برای مدل خطی ساده است که بصورت زیر و به فرم ماتریسی ارایه می شود.

۱- خطی بودن $\mathrm{E}(y;x) = \mathbf{x}\boldsymbol{\beta} \Leftrightarrow E(\varepsilon) = \mathbf{0}$

۲- هم واریانسی باقیمانده ها

۳- وابسته نبودن با قیمانده ها $\begin{Bmatrix} Var(\varepsilon) = \sigma^2 \mathrm{I} \\ COV(\varepsilon_i, \varepsilon_j) = 0;\ \forall i \neq j \end{Bmatrix}$

۵- $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$

که در آن **I** یک ماتریس واحد است.

تفسیر β_j : β_j برابر میانگین افزایش Y وقتی که متغیر آزاد X_j به اندازه یک واحد افزایش یابد و بقیه متغیرهای آزاد ثابت باشند.

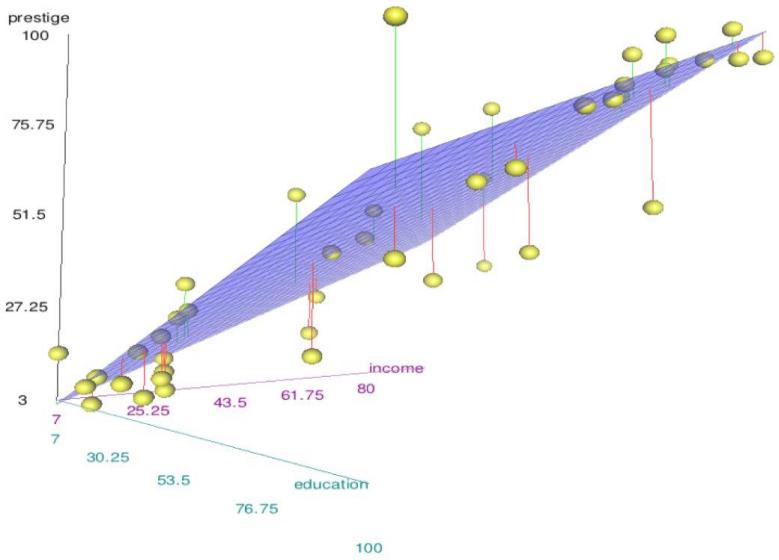
۱۰۳۰۲- **روش کمترین توانهای دوم خطا**: Mean Squar of Erorr Method

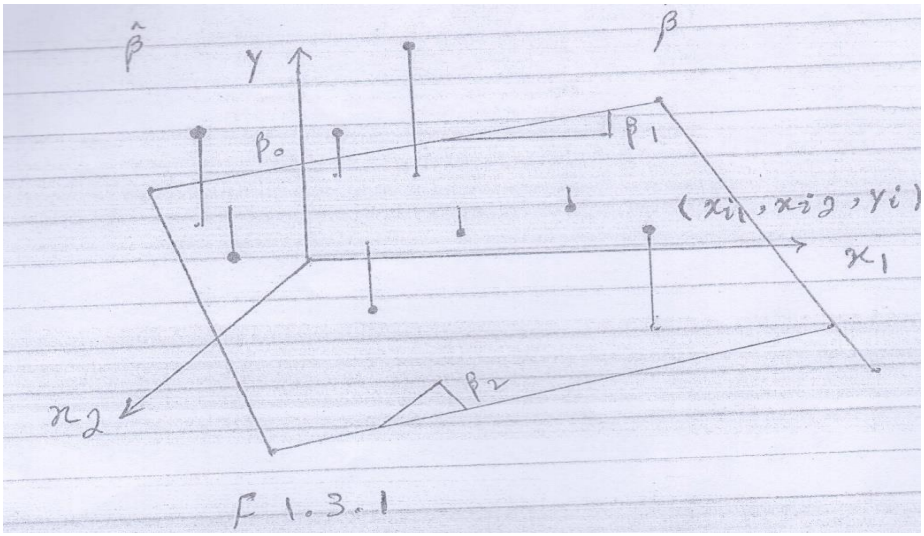
مثل قبل $\min ss_{Res}(\hat{\beta}) = \min\{y_i - (\hat{\beta}_1 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{ip} x_{ip})\}^2$=

Min {y - $x\hat{\beta})'(y - x\hat{\beta})\}$

=min{(y - $\hat{y})'$(y- $\hat{y})$}=min{e′.e } (1.3.3)

صفحه رگرسیون: **Regression space**





F 1.3.1

یعنی $\hat{y}$ یک تابع تعدیل شده و ε باقیمانده ست $ss_{Res}(\hat{\beta})$ تابعی پیوسته نرم (مشتق پذیر) و محدب است که براحتی می نیموم می شود و ماتریس **x** دارای رنک کامل full rank فرض می شود و ماتریس x'x معکوس پذیر فرض می شود. تحت این شرایط کافی است معادله زیر را حل کنیم

$$\frac{\partial}{\partial\hat{\beta}} ss_{Res}(\hat{\beta}) = \begin{pmatrix} \frac{\partial ss_{Res}(\hat{\beta})}{\partial\hat{\beta}_0} \\ \vdots \\ \frac{\partial ss_{Res}(\hat{\beta})}{\partial\hat{\beta}_{p'}} \end{pmatrix} = \mathbf{0}$$

$= \frac{\partial}{\partial\hat{\beta}}\{(y - x\hat{\beta})'(y - x\hat{\beta})\}$

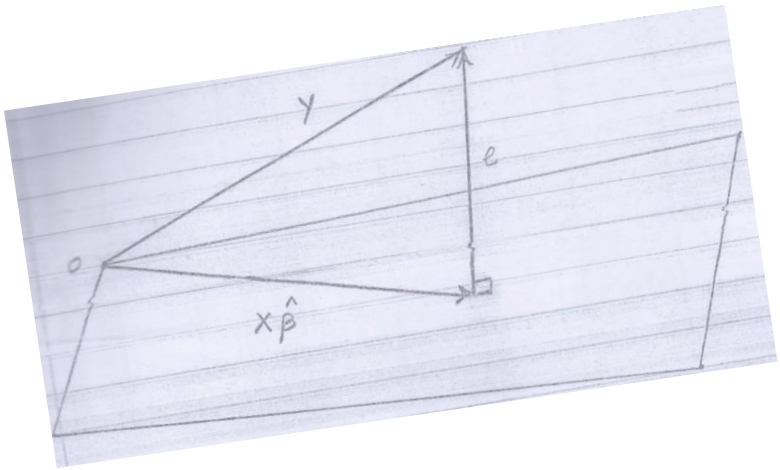
$= \frac{\partial}{\partial\hat{\beta}}\{y'y - 2\hat{\beta}x'y + \hat{\beta}'\, x'x\hat{\beta}\}$

$= 0 - \; 2x'y + 2x'x\hat{\beta} = 0$

$$\Rightarrow 2X'X\hat{\beta} = 2x'y \Longrightarrow \hat{\beta} = (x'x)^{-1}x'y \qquad (1.3.4)$$

برآورد $\boldsymbol{\beta}$ بروش هندسی:

شکل زیراکه در نظر بگیرید:



span(x) (1.302)

$(x\hat{\beta}\,)'\hat{e} = (x\hat{\beta}\,)'\left(y - x\hat{\beta}\,\right) = 0$ (1.3.5)

$$=\hat{\beta}\,'\left(x'y - x'x\hat{\beta}\,\right) = 0 \Longrightarrow \begin{Bmatrix} \hat{\beta}^{\,=0} \\ x'y - x'x\hat{\beta}\;\; = 0 \end{Bmatrix} \rightarrow \hat{\beta}\;\; = (x'x)^{-1}x'y$$

برآورد واریانس عبارت خطا:

$H_{nn} = x\,(x'x)^{-1}\,x'$

خواص ماتریس H:

1. $H = H'$

2. $HH = H^2 = H$

3. $HX=X$ ماتریس همانی

4. trace(H) = p′ = p+1 ، Hاثر ماتریس (جمع قطر اصلی)

و

1. I-H =(I- H) ′

2.(I-H)(I-H) =(I-H) 2 =I-H

3. (I-H)X=0

H را ماتریس تبدیل گویند و

$$\Rightarrow \hat{e} = (Y - X\hat{\beta}) = (Y - X(X'X)^{-1}X'Y) = (Y - HY) = (I - H)Y = (I - H)\varepsilon$$

$$; \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

ماتریس H نقش مهمی را بازی می کند که در زیر ملاحظه خواهید کرد.

در اینجا از H برای بر آورد واریانس σ^2 از عبارت خطا استفاده می کنیم:

آخرین معادله بر آورد واریانس را می دهد:

$$S^2 = \frac{ss_{res}(\hat{\beta})}{n - p'} = \frac{e'e}{n - p'} = \frac{\sum_1^n (y_i - \hat{y}_i)^2}{n - p - 1} \qquad (1.3.6)$$

ملاحظه می شود با برآورد واریانس در رگرسیون خطی ساده مطابقت دارد چون

P=1, p′=2 $\Longrightarrow s^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/(n-2)$

مثال :

$(10.7, 13.1, 7.5, 12.9, 12.2, 10.8, 8.9, 12.0, 9.55, 10.0) = y'$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 9.96 & 7.98 & 11.8 & 7.23 & 9.06 & 8.39 & 11.8 & 7.69 & 10.2 & 10.5 \end{pmatrix} = x'$$

$$= x'x = \begin{pmatrix} 10.0 & 94.62 \\ 94.62 & 919.8 \end{pmatrix}, \ (x'x)^{-1} = \begin{pmatrix} 3.749 & -0.386 \\ -0.386 & 0.0408 \end{pmatrix}$$

$$= x'y = \begin{pmatrix} 107.7 \\ 994.3 \end{pmatrix}, \ \hat{\beta} = (x'x)^{-1}x'y = \begin{pmatrix} 20.3 \\ -1.01 \end{pmatrix}$$

$$SS_{Res} = \sum_{1}^{10} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 4.612 \quad , \ s^2 = \frac{4.612}{10-2} = 0.577$$

1.3.3 روش درستنمایی ماگزیمم Maximum Likelihood Method

در این روش پارامترها طوری برآورد می شود که تابع چگالی توام (پیوسته) مشاهدات ماگزیمم شوند.بدین منظور باید (در رگرسیون) یک توزیع برای عبارت خطا در نظر گرفته شود.

تحت فرض نرمال بودن رگرسیون خطی ساده چگالی متغیر تصادفی پاسخ به ازای متغیرهای غیر تصادفی آزاد، نرمال هستند.

تحت فرض های اولیه روی خطاها در رگرسیون خطی ساده داریم:

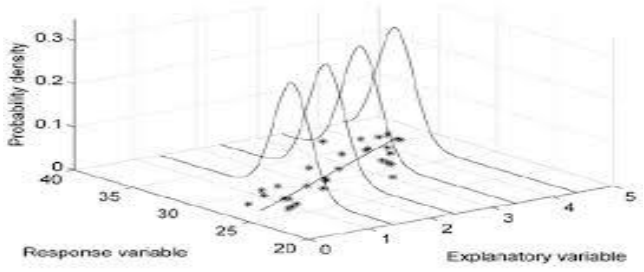
I) $E(e_i)=0 \iff E(y_i|x_i) = \beta_0+\beta_1 x_i$

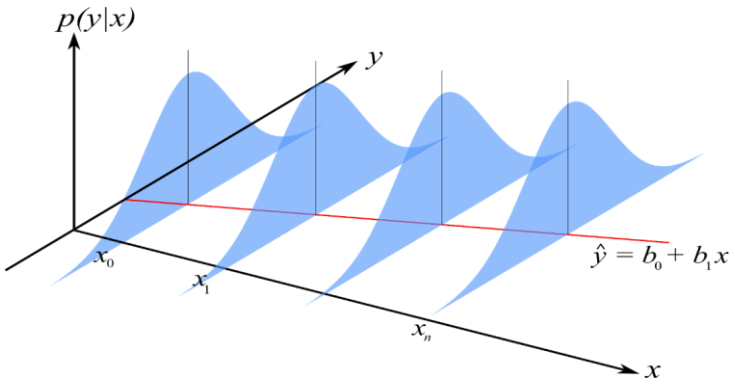
II) $V(\varepsilon_i)= \sigma^2$; i=1,....,n

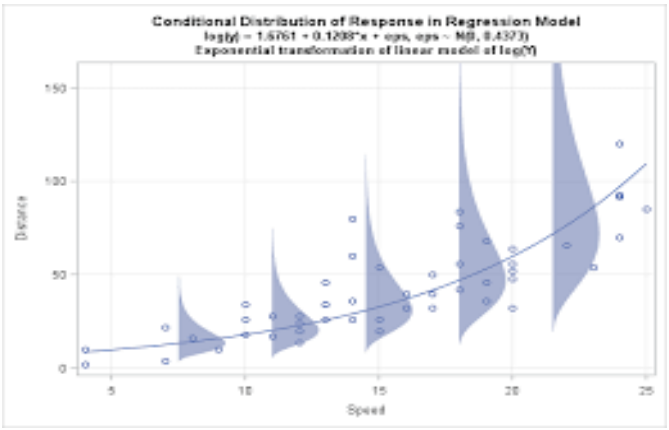
III) $COV(\varepsilon_i , \varepsilon_j)=0$; $i \neq j$

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

$\Rightarrow y_i | x_i \sim N(\beta_0+\beta_1 x_i ; \sigma^2)$; $i=1,...,n$







$f(y_1,...,y_n ; x_1,.....,x_n) = L(\beta_0,\beta_1, \sigma^2)= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\{- \frac{1}{2\sigma^2} (y_i - \hat{y}_i)^2\}$

$=(2\pi \sigma^2)^{- n/2} \exp\{- \frac{1}{2\sigma^2} \sum(y_i - \beta_0 - \beta_1 x_i)^2\}$

(1.3.7)

$L(\beta_0,\beta_1, \sigma^2)= - \frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum(y_i - \beta_0 - \beta_1 x_i)^2$

(1.3.8)

$$\Rightarrow \begin{cases} \frac{\partial L(\widehat{\beta_0},\widehat{\beta_1},\widehat{\sigma^2})}{\partial \widehat{\beta_0}} = 0 & (1.3.9) \\ \frac{\partial L(\widehat{\beta_0},\widehat{\beta_1},\widehat{\sigma^2})}{\partial \widehat{\beta_1}} = 0 & (1.3.10) \\ \frac{\partial L(\widehat{\beta_0},\widehat{\beta_1},\widehat{\sigma^2})}{\partial \widehat{\sigma^2}} = 0 & (1.3.11) \end{cases}$$

$$\widehat{\beta_1} = \frac{\sum y_i(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (1.3.12)$$

$$\widehat{\beta_0} = \overline{y_n} - \widehat{\beta_1}\bar{x} \quad (1.3.13)$$

$$\widehat{\sigma^2} = \frac{\sum(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n} = \frac{SS_{ReS}(\widehat{\beta}_0,\widehat{\beta}_1)}{n} \quad (1.3.14)$$

رگرسیون خطی چند متغیره :

با توجه به فرض نرمال بودن (خاصیت ۴) داریم

(

$$\boldsymbol{y \sim N_n(x\beta, \sigma^2 I_n)}$$

$$f(y;x) = l(\beta,\sigma^2) = \frac{1}{\sqrt{(2\pi)^n ||\sigma^2 I||}} \exp\{-\frac{1}{2}(y-x\beta)'(\sigma^2 I)^{-1}(y-x\beta)\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\frac{(y-x\beta)'(y-x\beta)}{\sigma^2}\}$$

(1.3.15)

لگاریتم تابع درستنمایی ماگزیمم:

$$l(\beta,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}ln\sigma^2 - \frac{1}{2\sigma^2}(y-x\beta)'(y-x\beta) \qquad (1.3.16)$$

$$\Rightarrow \begin{cases} \frac{\partial L(\widehat{\beta},\widehat{\sigma^2})}{\partial\widehat{\beta}} = 0 & (1.3.17) \\ \frac{\partial L(\widehat{\beta},\widehat{\sigma^2})}{\partial\widehat{\sigma^2}} = 0 & (1.3.18) \end{cases}$$

$$\Rightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y} \qquad (1.3.19)$$

$$(1.3.20) \qquad \widehat{\sigma^2} = \frac{(y-x\widehat{\beta})'(y-x\widehat{\beta})}{n} = \frac{e'e}{n} = \frac{SS_{ReS}(\widehat{\beta})}{n}$$

ملاحظه می شود که برآوردگرهای β در روش ML مثل روش MSE است ولی برآوردگر σ^2 با استفاده از ML متفاوت(n) در مخرج کسر می باشد. n بجای p' -n جایگزین شده است.

۱,۳,۴- خواص برآوردگرها : Properties of Estimators

بدلیل سادگی محاسبات ، فرم ماتریسی رگرسیون خطی چندگانه را در نظر می گیریم.

قضیه ۱,۲- فرض کنید مدل رگرسیون Y=Xβ+ E تحت فرض I

E(ε)=0 , (II- III) : V(ε)=$\sigma^2 I_{nn}$

۱. $\hat{\beta}$ نااریب است.

۲. $V(\widehat{\beta}) = \sigma^2 (x'x)^{-1}$

۳- تمام برآوردگرهای خطی y مانند $V'\hat{\beta}$ برآوردگرهای نا اریب برای $V'\beta$ با واریانس می نیمال برای بردار ثابت V است.

۴- $\hat{\sigma}^2 = S^2$ نا اریب است.

۵- $V(\hat{\sigma}^2) = \frac{2\sigma^4}{(n-p')}$

با افزودن فرض IV داریم : $\hat{\beta} \sim N_{\acute{p}}(\beta, \sigma^2(x'x)^{-1})$, $\varepsilon \sim N_n$

اثبات :

1- $E(\hat{\beta}) = E\left((x'x)^{-1}x'y\right) = (x'x)^{-1}x'E(y) = (x'x)^{-1}x'x\beta = \beta$

2- $V(\hat{\beta}) = V\left((x'x)^{-1}x'y\right) = (x'x)^{-1}x'V(y)x\left[(x'x)^{-1}\right]'$

$$= (x'x)^{-1}x'x(x'x)^{-1}\sigma^2 = \sigma^2(x'x)^{-1}$$

۲- قضیه گوس- مارکف Gauss Markov Theorem

۳- $\hat{\beta}$ در بین تمام برآوردگرهای نااریب β که تابعی خطی از $y_1, \dots, y_n$ هستند دارای کمترین واریا نس می باشند. بنابراین $\hat{\beta}$ *برای* β ، $BLUE$ هستند.
(Best Linear Unbiased Estimator).

تمرین با استفاده از H -I

4- $X^2 = \frac{(n-p')S^2}{\sigma^2} \sim X^2_{n-p'} \Rightarrow \begin{cases} EX^2 = n - p' \\ V(X^2) = 2(n-p') \end{cases}$

مثال - ۱،۲- یک فاصله اطمینان ۹۵ درصدی برای β_1 بدست آورید؟ اگر :

$$CI_{1-\alpha}(\beta_1)$$

Confidence Interval for: β_1

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

$$\widehat{\beta_1} = -1.01 \quad , \; S^2 = 0.577 \quad , \quad (x'x)^{-1} = \begin{pmatrix} 3.749 & -0.386 \\ -0.386 & 0.0408 \end{pmatrix}$$

با استفاده از جدول t $t_{(0.025\,,\,10-\,2)}=2.306$

$Cl_{0.95}\,(\beta_1)=[-\ 1.364\ ,-\ 0.656]$

با استفاده از لم ۱،۴ نیز می توان فاصله اطمینان یا آزمون فرض زیر را انجام داد:

$$\begin{cases} H_0 : \beta_j = \beta_{j,0} \\ H_1 : \begin{cases} \beta_j < \beta_{j,0} \\ > \\ \neq \end{cases} \end{cases}$$

که در آن $\beta_{j,0}$ یک مقدار انتخابی (جایگزین) برای β_j است که در عمل معمولا 0 فرض می شود.

آماره آزمون عبارتست از :

$$t = \frac{\widehat{\beta_j} - \beta_{j,0}}{\sqrt{S^2 (x'x)^{-1}_{jj}}} \qquad (1.32)$$

جهت سادگی کار جدول زیر را فراهم می کنیم که در آن $H_0 : \beta_j = \beta_{j,0}$ است و فرض های جایگزین عبارتند از:

حال اگر S_{xx} بزرگ باشد آنگاه $V(\hat{\beta}_1)$ کوچک خواهد بود.

برای مطالعه بیشتر به Draper & smith (1998) *صفحه* ۸۹- ۸۶ مراجعه کنید.

۱،۵- تبدیل Box- Cox برای متغیر پاسخ y:

مثال –

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x^{2}_{i3} + \varepsilon i$$

که

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x^{*}_{i3} + \varepsilon i$$

حتی این تبدیل می تواند برای y_i ها باشد.

این تبدیل معمولا برای وقتی که اشکال در فرض نرمال بودن یا یکنواختی واریانس داشته باشیم ، یکی از روش های معمولی و کار آمد روش Box- Cox می باشد.

این روش بهترین تبدیل ممکن را جستجو می کند برای متغیر تصادفی {y} endogene.

در حالت کلی فرمول تبدیل به صورت زیر تعریف می شود و الگوریتم زیر انجام می شود.

1- $$g(y;\lambda)=\begin{cases}\frac{y^{\lambda}-1}{\lambda} & ;\ \lambda \neq 0\\ \ln(y) & ;\ \lambda \neq 0\end{cases}$$

۲- مقدار λ را تعیین می کنیم (λ^{*}).

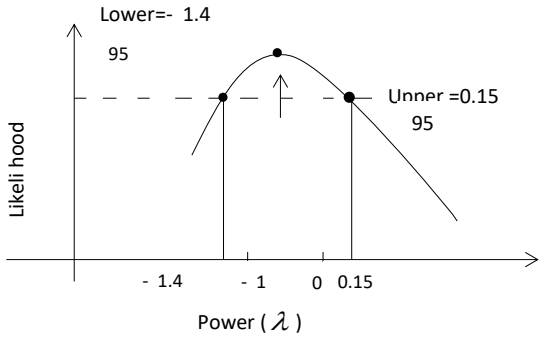
۳- $$y^{*}_{\lambda}=g(y;\lambda^{*})$$

را بدست آورده مدل را برآزش می دهیم.

۴- تابع لگاریتم ماگزیمم درستنمایی را حساب می کنیم (برای مدل جدید) : $l_{\lambda}(\hat{\beta}(\lambda))$

5- Repeat step 1- 4 (for number of values of λ)

مراحل ۱ تا ۴ را برای چندین مقدار از λ مثل (2 , 2-) با گام ۰,۱ تکرار می کنیم.



۶- فرض کنید $\hat{\lambda}$ مقدار برآورد λ باشد که $l_{\lambda}(\hat{\beta}(\lambda))$ را ماگزیمم می کند. لذا براحتی می توان از فاصله اطمینان ۹۵٪ برای λ استفاده کرد.

$$\lambda : \left\{ \lambda : 2\{ l_{\hat{\lambda}}(\hat{\beta}(\hat{\lambda})) - l_{\lambda}(\hat{\beta}(\lambda)) : 3.841 \right\}$$

مثال ۱,۳- اگر این روش را برای مثال مصرف بنزین بکار ببریم همانطور که شما در شکل فوق ملاحظه می کنید (لگاریتم تابع درستنمایی ماگزیمم برای چندین مقدار λ محاسبه و رسم شده است) Box- Cox با توجه به فاصله اطمینان داده شده در فوق می توان پیشنهاد کرد که

$y^{*} = \ln(y)$ معادل $\lambda = 0$ یا $y^{*} = \frac{1}{y}$ معادل $\lambda = -1$ و یا $y^{*} = \frac{1}{\sqrt{y}}$ معادل

$\lambda = -0.5$.

۱,۶- پیش بینی Forecasting

نتایج آنالیز رگرسیون معمولا برای حل مشکل پیش بینی استفاده می شود.

پیش بینی متغیر پاسخ y یا E(y) یا یکی از متغیرهای آزاد مثل $x_1, \dots, x_p$.

در کل پیش بینی به صورت نقطه ای یا فاصله های اطمینان حاصل می شود.

۱,۶,۱- تفسیر E(y|x) :

فرض کنید می خواهیم مقدار متوسط متغیر پاسخ را برای یک ترکیب خاص مانند

$x^{*\prime} = (1, x_1^*, \dots, x_p^*)$ ازمتغیرهای غیرتصادفی آزاد بدست آوریم.

$$E(y ; x^*) = E(x^{*\prime}\beta + \varepsilon) = x^{*\prime}\beta$$

بنا به قضیه گوس- مارکف بهترین برآوردگر خطی برای عبارت فوق عبارتست از $x^{*\prime}\hat{\beta}$ و

$$Var(x^{*\prime}\hat{\beta}) = x^{*\prime} V(\hat{\beta}) x^* = \delta^2 x^{*\prime}(x'x)^{-1} x^*$$

در نتیجه داریم :leme۱,۴

$$\frac{x^{*\prime}\hat{\beta} - x^{*\prime}\beta}{\sqrt{s^2\, x^{*\prime}(x'x)^{-1}\, x^*}} \sim t_{n-p'} \qquad (1.33)$$

فاصله اطمینان $100\times(1-\alpha)\%$ برای $E(y ; x^*)$

$$CI_{1-\alpha}\big(E(y; x^*)\big) = x^{*\prime}\hat{\beta} \pm t_{\frac{\alpha}{2}, n-p'} \sqrt{s^2\, x^{*\prime}(X'X)^{-1}\, x^*} \qquad (1.34)$$

نتیجه ۱,۶,۱- برای رگرسیون خطی ساده داریم :

(1.35)

$$CI(E(Y;x^*)) = \widehat{\beta_0} + \widehat{\beta_1}x^* \pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\,(\frac{1}{n} + \frac{(x^* - \bar{x}_n)^2}{s_{xx}})}$$

۱,۶,۲- استنباط روی یک مقدار y به شرط x مقدار.

فرض کنید x^* یک بردار ثابت باشد از نظر برآورد نقطه ای y به ازای x^* هیچ تغییری بوجود نمی آید. در صورتی تغییر حاصل می شود که واریانس بزرگ داشته باشیم و در سطح فاصله اطمینان مورد نظر باشد.

در عمل می خواهیم $y = \chi^{*'}\beta + \varepsilon$ را برآورد کنیم. با توجه به قضیه گوس- مارکف برآورد $\chi^{*'}\beta$ عبارتست از $\chi^{*'}\hat{\beta}$. مثل E(ε)=0 یعنی بهترین آوردگر برآورد می کند ۰ برای ε.

پس بهترین برآورد نقطه ای y به ازای x^* با $\chi^{*'}\hat{\beta} + 0 = \chi^{*'}\hat{\beta}$ داده می شود.

برای بدست آوردن تابع توزیع برآوردگر :

؟

در نتیجه

$$\frac{x^{*\prime}\widehat{\beta}-(x^{*\prime}\beta+\varepsilon)}{\sqrt{s^2[1+x^{*\prime}(X'X)^{-1}x^*]}}\sim t_{n-p'} \qquad (1.36)$$

$$CI_{1-\alpha}(y;x^*):$$

$$\sqrt{s^2[1+\ x^{*\prime}(X'X)^{-1}\ x^*]}x^{*\prime}\widehat{\beta}\ \pm t_{\frac{\alpha}{2},n-p}$$

نتیجه ۱,۷,۱- در رگرسیون خطی ساده :

$$CI_{1-\alpha}(y;x^*):\widehat{\beta_0}+\widehat{\beta_1}x^*\pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\ (1+\frac{1}{n}+\frac{(x^*-\bar{x}_n)^2}{s_{xx}})}$$

(1.38)

تذکر ۱ : فاصله های اطمینان ۱,۳۷ و ۱,۳۸ فاصله های پیش بینی نیز نامیده می شوند.

مثال ۱,۳- در مثال مصرف بنزین $y_i=\beta_0+\beta_1x_{i1}+.....+\beta_4x_{i4}+\varepsilon_i$

یعنی

$$\hat{y} = \boldsymbol{X}^{*'}\hat{\beta} = 622 \Leftarrow \begin{cases} \boldsymbol{X}^{*'} = (1, 8, 604, 5 \\ \hat{\beta}' = (377.29, -34.79, 13.36, -66.59, -2.43) \end{cases}$$

$$CI_{95}\big(E(y; x^*)\big) = 622 \pm 2.02\sqrt{4396.51\, x^{*\prime}(X'X)^{-1}\, x^*}$$

$$= 622 \pm 2.02\sqrt{4396.51 \times 0.0407}$$

$$= (595, 649)$$

و نیز فاصله اطمینان برای یک مقدار پیش بینی از y در شرایط فوق عبارتست از :

$$CI_{95}(y; x^*) = 622 \pm 2.02\sqrt{4396.51\,(1 + 0.047)} = (485, 759)$$

۱،۶،۳- پیش بینی یک مقدار برای متغیر پاسخ تبدیل یافته :

فرض کنید $y^*=g(y)$ یک مقدار تبدیل یافته y باشد. اگر بخواهیم یک پیش بینی نقطه ای یا فاصله ای (اطمینان) یا برآورد برای متغیر تصادفی تبدیل یافته پیدا کنیم چکار باید کرد؟

فرض کنید تابع تبدیل g(.) یکتا و معکوس پذیر باشد یعنی $g^{-1}(.)$ موجود باشد.

بنابراین براحتی می توان یک فاصله اطمینان برای y بدست آورد و فرض کنید این فاصله (a,b) باشد. چون

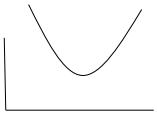
$$y^* = g(y) \Leftrightarrow g^{-1}(y^*) = y$$

$$CI_{1-\alpha}(y^* = g(y)) = (g(a), g(b)) \quad ; \quad .$$

$$CI_{1-\alpha}(y^* = g(y)) = (g(b), g(a)) \quad ;$$

$$CI_{1-\alpha}(y) = (a,b)$$

اما متاسفانه برای E(g(y)) چنین کاری ممکن نیست چون در حالت کلی داریم :



اگر g(.) محدب یا Con Vex $E(g(y)) \geq g(E(y))$;

اگر g(.) مقعر یا Con Cave $E(g(y)) \leq g(E(y))$;

چون در این حالت $E(g(y)) \neq g(Ey)$ لذا نمی توان از یک مقدار پیش بینی نقطه ای برای y^* یا

$E(y^*)$ به پیش بینی نقطه ای برای y یا E(y) رسید. در حالتهای خاص این کار ممکن است مثلا اگر

$$y^* = \ln(y) \Rightarrow y = \exp(y^*)$$

و بنابراین $E(y; x^*) = E(e^{y^*}; x^*)$?

تحت فرض نرمال بودن داریم، این امید یک لگ نرمال خواهد بود یا با استفاده از تابع مولد گشتاور در حالتی

t=1 است داریم :
$$E(y;x^*) = \exp(x^*\beta + \frac{\delta^2}{2})$$

که برآورد آن خواهد بود : $\exp(x^*\hat{\beta} + \frac{s^2}{2})$

۱,۷- آزمون فرضیه و آنالیز واریانس : Hypathesis Testing and Analysis of Variance

چندین سوال درمورد تغییرات متغیر پاسخ که توسط متغیرهای آزاد (Covariables) بیان می شود مطرح است.

دراین بخش روشهای متفاوت آنالیز واریانس و آزمون فرضها را مورد بررسی قرار می دهیم.

۱,۷,۱- آنالیز واریانس : Analysis of Variance

در کل مقدار های $y_1,....,y_n$ متفاوت هستند که و یکی از هدفهای مهم آنالیز رگرسیون، بررسی بخش از تغییرات y است که توسط X بیان می شود. از طرفی اگر تغییرات y را بتوان به صورت (۱,۳۸)

=

+

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

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

$$= \sum(\hat{y}_i - \bar{y}_n)^2 + \sum(y_i - \hat{y}_i)^2 \qquad (1.39)$$

اگر به فرم ماتریسی بنویسیم داریم :

؟

اگر به فرم ۱,۳۹ بنویسیم داریم :

Source	DF	SS	MS	F_o
Reg	p	SS_{Reg}	$\frac{SS_{Reg}}{p}$	$F_{p,n-p'} = \frac{MS_{Reg}}{MS_{Res}}$
Res	$n-p'$	SS_{Res}	$\frac{SS_{Res}}{n-p'}$	
Tot(corrected)	n-1	SS_{Tot}		

جدول آنالیز واریانس شامل β_0 :

Source	d.l	SS	MS	F_o
β_0	1	$n\,\bar{Y}_n^{\,2}$	$n\,\bar{Y}_n^{\,2}$	$F_{1,n-p'}=\dfrac{n\,\bar{Y}_n^{\,2}}{MS_{Res}}$
Reg	p	SS_{Reg}	$\dfrac{SS_{Reg}}{p}$	$F_{p,n-p'}=\dfrac{MS_{Reg}}{MS_{Res}}$
Res	$n-p'$	SS_{Res}	$\dfrac{SS_{Res}}{n-p'}$	
Tot	n	$SS_{Tot}+n\,\bar{Y}_n^{\,2}$		

در مثال مصرف بنزین جدول آنالیز واریانس به صورت زیر می باشد:

$$SS_{\mathrm{Re}\,g}=588366-189050=399316$$
$$\sum y_i^{\,2}=SS_{Tot}+n\bar{y}_n^{\,2}=16556267$$

Source	d.l	SS	MS	F_o
Reg	4	399316	99829	22.70
Res	43	189050	4397	
Tot(corrected)	47	588366		

جدول آنالیز واریانس شامل β_0:

Source	d.l	SS	MS	F_o
β0	1	15967901	$\frac{15967901}{4397}$	3631
Reg	4	399316	99829	22.70
Res	43	189050	4397	
Total	48	16556267		

۱,۷,۲- F- Test برای آزمون رگرسیون چند متغیره کلی :

Hypothesis Testing In Multiple Linear Regression Model

یک آزمون فرض مهم در آنالیز رگرسیون شامل حداقل یکی از متغیرهای آزاد است که بخشی از تغییرات y را بیان می کند. در این آزمون فرض H_0 , یعنی هیچ کدام از متغیرهای آزاد (Co- Var) موثر نیستند (یا

بخش مهمی از تغییرات y را بیان نمی کنند) و فرض جانشین alternative یعنی حداقل یکی از متغیرهای آزاد موثر می باشد یا بخش مهمی از تغییرات y را بیان می کند.

چنانچه $F = \frac{\frac{SS_{\mathrm{Re}g}}{p'}}{\frac{SS_{\mathrm{Re}s}}{(n-p)}}$ یا $F = \frac{SSReg}{p's^2}$ کوچک باشد یعنی مدل نمی تواند بخش بزرگی از تغییرات y

؟

در مثال مصرف بنزین F_0=22.7 است که $P(F_{4,43} \geq 22.7) \approx 0$ در نتیجه فرض H_0 رد و فرض H_1 جایگزین آن می شود یعنی حداقل یکی از متغیرهای آزاد می تواند تغییرات y را بیان کند یا به بیان دیگر حداقل یکی از βi ها مخالف صفر است.

۱,۷,۳- اصل مجموع مربعات باقیمانده های خطای اضافی:

The Rule of sum Squar of Erorrs

به ندرت پیش می آید که فرض اولیه H_0 آزمون کلی ۱,۴۳ رد نشود و مدل رگرسیون ناکارآمد باشد.

در صورت رد شدن فرض H_0 ، می خواهیم بدانیم کدام زیر مجموعه از متغیرهای آزاد (از بین تمام متغیر ها) می تواند به اندازه کافی بخش بزرگی از تغییرات y را بیان کند و در نتیجه سطح معنی داری آن چقدر است؟

بجای مدل کامل یا پیچیده تر یعنی $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p'} x_{ip'} + \varepsilon i$ ، فرض کنید مدل رگرسیون چندگانه باشد ولی با متغیرهای ($k < p$)کمتری نسبت به اولیه فوق باشد.

اگر برای بیان تغییرات y فقط متغیرهای $x_1, \ldots, x_k$ لازم باشد و بقیه متغیرها یعنی p-k متغیر دیگر غیر ضروری باشد آنگاه براحتی معلوم می شود که باید دو مدل زیر را باهم مقایسه کنیم.

$H_0 : \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon i$ مدل کاهش یافته (۱,۴۴)

$H_1 : \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \beta_{k+1} x_{ik+1} + ... + \beta_{p'} x_{ip'} + \varepsilon i$ مدل کامل

؟

و همیشه این رابطه صحیح است که $SS_{Res}^{H_1} \leq SS_{Res}^{H_0}$ زمانی که اختلاف این دو زیاد باشد شرایط به نفع H_1 خواهد بود و اگر این تفاضل کوچک باشد شواهد به نفع H_0 یعنی درستی مدل کاهش یافته خواهد بود.

اینجا آماره آزمون به صورت زیر خواهد بود:

$$F = \frac{(n-p')(SS_{Res}^{H_0} - SS_{Res}^{H_1})}{\Delta df \; SS_{Res}^{H_1}} \sim F_{\Delta df,(n-p')} \qquad (۱,۴۵)$$

روشهای محاسبه Δdf :

I. $\Delta df = df_{SS_{Res}^{H_0}} - df_{SS_{Res}^{H_1}}$

II. $\Delta df =$ {تعداد پارامترهای تحت H_1} - {تعداد پارامترهای تحت H_0}

III. $\Delta df =$ { تعداد محدودیت اعمال شده روی پارامترها برای رسیدن به مدل کاهش یافته }

تذکر ۲- تحت فرض نرمال بودن باقیمانده ها اصل ۱,۷,۳ (آزمون مجموع توانهای دوم خطای اضافی) یک آزمون نسبت درستنمایی است.

مثال ۱,۴- در مثال مصرف بنزین، اقتصاد داان معتقد است که نرخ دارندگان اجازه رانندگی و طول مسیر جاده های فدرال اثری بر مصرف بنزین ندارند. لذا :

$$H_0 \ : \ y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon i$$

و جدول آنالیز واریانس :

Source	d.l	SS	MS	F
Reg	2	153478	76739	7.94
Res	45	434889	9664	
Tot	47	588366		

$$F_0 = \frac{SS_{Res}^{H_0} - SS_{Res}^{H_1}}{\Delta dl\ S^2} = \frac{434.889 - 189050}{(45-43)4397} = 27.96$$

یعنی مدلی که توسط اقتصاددان پیشنهاد شده است رد می شود.

چون $F_0 = 27.96 > 3.21 \Rightarrow RH_0$ و $F_{0.05,2,43} = 3.21$

برای استاندارد کردن مقدار تفاضل فوق ، آنرا بر واریانس تحت فرض H_1 تقسیم می کنیم.

آزمون فرضیه روی ترکیب خطی از پارامترها:

$$\begin{cases} H_0: C\beta = d \\ H_1: C\beta \neq d \end{cases}$$ ۱,۴۶

؟

(۱,۴۷)

$$F = \frac{\left(C\hat{\beta} - d\right)' \left[C\left(x'x\right)^{-1}C'\right]^{-1} \left(C\hat{\beta} - d\right)/r}{SS_{Res}^{H_1}/(n - p')} = \frac{"}{rs_{H_1}^2}$$

مثال ۱,۴- (مصرف بنزین)

فرض کنید : اثر افزایش ۱٪ نرخ فروش روی میانگین مصرف بنزین برابر اثر کاهش $۵۰۰ درآمد روی میانگین مصرف بنزین خواهد بود.

اولا اثر افزایش یک درصدی نرخ فروش در میانگین عبارتست از :

$$E[y; x_1 + 1] - E[y; x_1] = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4] = \beta_1$$

بطور مشابه اثر کاهش 500$ در حقوق یا درآمد (x_3 به اندازه ۰,۵ کاهش می یابد) روی میانگین مصرف بنزین برابر $-0.5\beta_3$ خواهد بود. بنابراین فرض ها عبارتند از :

$$H_0 : \beta_1 + 0.5\beta_3 = 0 \quad \text{یا} \quad H_0 : \beta_1 = -0.5\beta_3$$

یعنی

$$c = (0,1,0,0.5,0), d = 0$$
$$c(x'x)^{-1}c' = 20.32 \Rightarrow c\hat{\beta} - d = -68.1$$
$$F_0 = \frac{(-68.1)(20.32)(-68.1)}{1 \times 4397} = 21.4 \quad ?$$
$$p(F_{1.43 \geq 21.4}) < 0.0001 \Rightarrow RH_0$$

$$F = \frac{\frac{C^2}{\Delta dl}}{\frac{\beta^2}{(n-p)}} = \frac{\frac{(A^2 - \beta^2)}{\Delta dl}}{\frac{SS_{Res}^{H_1}}{(n-p')}} = \frac{(SS_{Res}^{H_0} - SS_{Res}^{H_1})}{\Delta dl \ \delta_{H_1}^2}$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i$$

معادل آزمون زیر می باشد:

$$\begin{cases} H_0: \beta_2 = 0 \ , \beta_4 = 0 \\ H_1: \ \beta_2 \neq 0 \ \ or/and \ \ \beta_4 \neq 0 \end{cases}$$

$$C = \begin{pmatrix} 0,0,1,0,0 \\ 0,0,0,0,1 \end{pmatrix} \quad , \quad d = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} H_0 : \text{Reduced mod} \\ H_1 : \text{Complete mod} \end{cases}$$

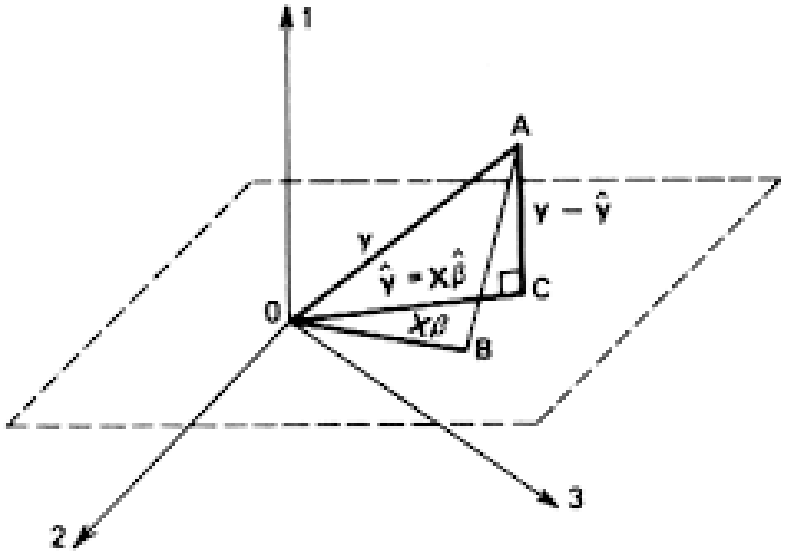


Figure 4.2 A geometrical interpretation of least squares.

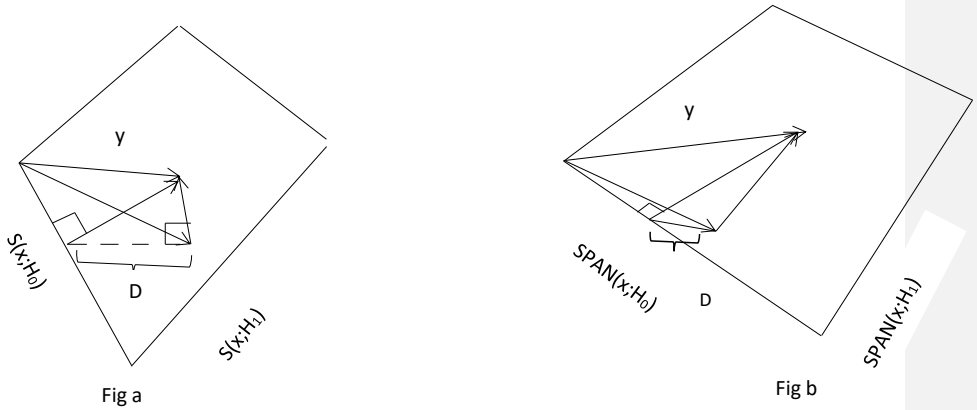


Fig 1.9

توضیح : در شکل a ملاحظه می شود که احتمالا H_0 رد می شود چون D که بیانگر اختلاف توانهای دوم خطا(باقیمانده ها)

می باشد تحت فرض های H_0 و H_1 زیاد است. اما در شکل b به دلیل کوچک بودن D احتمالا این فرض یعنی H_0 رد نشود.

۱،۸: روشهای انتخاب مدل Methods of Model Selection:

۱،۸،۱- روشهای مقایسه ای کلاسیک Classic Comparing Methods:

پس از تجزیه تغییرات کل $SS_T=SS_{Reg}+SS_{Res}$ ما فکر می کنیم که یک مدل رگرسیون را خوب آزمون کرد مگر اینکه کمی با استفاده از ابتکار و خلاقیت !

اصول Principal of Cross Validation اعتبار سنجی متقابل

اصل اعتبار سنجی متقابل توان یک مدل مفروض را برای پیش بینی مشاهدات جدید اندازه می گیرد و به صورت الگوریتم زیر بیان می شود:

۱- حذف i امین مشاهده از داده ها.

۲- پارامترهای مدل را بااستفاده از n- 1 مشاهده باقیمانده برآورد می کنیم و مدل را تشکیل می دهیم.

۳- با استفاده از مدل ۲ مشاهده y_i را با استفاده از x_i پیش بینی می کنیم و آنرا $\hat{y}_{i,-i}$ می نامیم.

۴- گام های ۳- ۱ را برای تمام i=1,...,n تکرار می کنیم.

۵- مجموع توانهای دوم خطای پیش بینی $PRESS = \sum(y_i - \hat{y}_{i,-i})^2 = \sum e^2_{i,-i}$ محاسبه می کنیم.

Predicted Residual Error Sum of Squares

براساس معیار PRESS می توان ضریب تعین پیش بینی

Determination Coefficient Of Prediction پیش بینی مدل را به صورت زیر تعریف کرد.

$$R^2_{Press} = 1 - \frac{Press}{SS_T} = 1 - \frac{\sum_{1}^{n} (y_i - y_{i,-i})^2}{\sum (y_i - \hat{y}_n)^2} \qquad (۱,۵۰)$$

$$0 \le R^2_{Press} \le 1$$

؟

قضیه ۱,۸ : i امین باقیمانده PRESS را می توان به کمک iامین باقیمانده معمولی عضو موقعیت (i,i) از

ماتریس تبدیل H به صورت زیر محاسبه کرد.

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}$$

؟

نتیجه ۱،۸،۱ :

$$PRESS = \sum_{i=1}^{n} (\frac{e_i}{1-h_{ii}})^2$$

در فصل ۲ دیدیم که h_{ii} مبین(معادل) طول فاصله پیش بینی می باشد و $0 \le h_{ii} \le 1$ نزدیک بودن h_{ii} به ۱ نشان دهنده فاصله بزرگ پیش بینی است و این کاملا واضح می باشد هرگاه h_{ii} نزدیک به ۱ باشد $e_{i,-i}$ خیلی بزرگ و به سمت بی نهایت میل می کند و درنتیجه پیش بینی y_i بسیار مشکل خواهد بود.

مثال ۱،۵- در مثال مصرف بنزین برای متغیر y تبدیل نیافته آماره PRESS

$$PRESS = 235401 \Rightarrow R^2_{Press} = 1 - \frac{235401}{588366} = 0.60$$

۱،۸،۳- C_p (Mallows)

این معیار انتخاب مدل ، از بین دو مدل یکی با تعداد متغیرهای کم و دیگری با تعداد متغیرهای زیاد مدل بهتر را پیشنهاد می کند. برای بدست آوردن این معیار، فهمیدن و بررسی نتایج دو مدل با متغیرهای کم و زیاد مهم است. بدین منظور فرض کنید مدل درست با m پارامتر باشد. $E(y) = x_1\beta_1 + x_2\beta_2$ که در آن β_1 دارای p پارمولفه (امتر) و β_2 شامل m- p مولفه (پارامتر) باشد.

حال فرض کنید که ما مدل $E(y)=x_1\beta_1$ را به داده ها برازش داده ایم.

در نتیجه :

$$\hat{\beta}_1 = (x_1'x_1)^{-1}x_1'y$$

برآوردگر β_1 می باشد. حال سوال این است که آیا این برآوردگر خوب است؟

بررسی نااریبی :

$$E(\hat{\beta}_1) = E((x_1'x_1)^{-1}x_1'y) = (x_1'x_1)^{-1}x_1'E(y)$$
$$= (x_1'x_1)^{-1}x_1'(x_1\beta_1 + x_2\beta_2)$$
$$= \beta_1 + (x_1'x_1)^{-1}x_1'x_2\beta_2 = \beta_1 + A\beta_2$$

بنابراین برآوردگر β_1 اریب است. حال اثر این اریبی را روی پیش بینی ملاحظه می کنیم.

اگر $x_1^{*'}\hat{\beta}_1$ را به عنوان برآوردگر $E(y; x^*)$ استفاده کنیم داریم :

$$\text{Bias}(x_1^{*'}\hat{\beta}_1) = E(x_1^{*'}\hat{\beta}_1) - (x_1^{*'}\beta_1 + x_2^{*'}\beta_2) = (x_1^{*'}A - x_2^{*'})\beta_2$$

و بعدا استفاده می کنیم از $(\text{Bias})^2 = \beta_2'(x_1^{*'}A - x_2^{*'})'(x_1^{*'}A - x_2^{*'})\beta_2$ اگر S_p^2 برآوردگر σ^2

باشد که از مدل با p پارامتر بدست آمده باشد آنگاه داریم :

$$E(S_p^2) = \delta^2 + \frac{1}{n-p}\sum[\text{Bias}(\hat{y}(x_i^*))]^2$$

می توان فهمید که هرگاه مدل شامل همه متغیرهای آزاد نباشد(متغیرهای موثر از مدل حذف شوند).

اولا برآوردگرهای پارامترهای مدل اریب خواهند بود و ثانیا $S_p{}^2$ مقدار بیشتری را برای σ^2 تخمین خواهد

زد.

. $S_p{}^2$ is supper estimator for the variance σ^2

برعکس اگر مدل با بیشترین پارامتر را به داده ها برازش دهیم بنابراین اریبی نداریم ولی این ممکن است که واریانس برآوردگرهای پارامترها و پیش بینی از واریانس مدل واقعی بیشتر باشد.

؟

(۱,۵۱)

منطقی است مدلی را پیدا کنیم که دارای MSE پیش بینی می نیمم باشد. از آنجائیکه نمی دانیم چه متغیرهایی مدل واقعی را خواهند ساخت و این کار تقریبا غیر ممکن است، ولی خوشبختانه تقریب زیر این عمل را ممکن می سازد یعنی

چون σ^2 مجهول است از برآورد آن $\hat{\sigma}^2$ استفاده می کنیم.

$$C_p = p + \frac{(s_p^2 - \hat{\delta}^2)(n-p)}{\hat{\delta}^2} = \frac{SS_{Res;p}}{\hat{\delta}^2} - (n-2p) \qquad (1.52)$$

چرا؟

در آن $\hat{\sigma}^2$ میانگین توانهای دوم باقیمانده ها برای مدل کامل (بابیشترین متغیر آزاد) می باشد.

در عمل مدلی خوب است که در آن C_p =P یا به بیان دیگر $C_p - P \approx 0$ وبرای این کار فرض کنید مدل دارای k پارامتر باشد یعنی باید 2^{k+1} مدل ممکن را برازش داد.

و از این بین مدلی را انتخاب کرد که دارای کوچکترین C_p- P باشد.

یعنی باید مدل هایی را که منطقی و قانع کننده هستند در نظر گرفت و از بین آنها که دارای C_p- P حول صفر هستند (مثبت یا منفی کوچک) را برگزید و سپس با استفاده از معیارهایی چون PRESS و R_a^2 .... مدل مطلوب را برگزید.

مثال ۱,۵- مثال مصرف بنزین را در نظر بگیرید.برای متغیر تبدیل نیافته (اولیه) y با استفاده از معیارهای C_p و R^2 و R_a^2 و PRESS مدل های :

۱- مدل کامل (با تمام متغیرها۴) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon i$

۲- مدل شامل تکس(مالیات) فروش(x_1) و درآمد(x_3).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon i$$

۳- مدل شامل فروش(x_1)- نرخ مجاز رانندگی(x_2) و درآمد(x_3).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon i$$

در حالت اول (مدل کامل) داریم :

$$I\begin{cases} PRESS = 235401, C_p = P\ ? \\ R^2 = 0.679 \\ R_a^2 = 0.649 \end{cases}$$

در حالت دوم :

$$II\begin{cases} C_p = \dfrac{SS_{Res;p}}{\hat{\delta}^2} - (n-2p) = \dfrac{434880}{4397} - (48 - 2\times 3) = 56.9 \\ R^2 = 0.261, R_a^2 = 0.228, PRESS = 490323 \end{cases}$$

در حالت سوم :

$$III\begin{cases} C_p = 3.512, R^2 = 0.675, R_a^2 = 0.653 \\ PRESS = 229999 \end{cases}$$

تفسیر : در حالتی $C_p \approx P$ باشد این معیار به نفع مدل مورد نظر رای می دهد. پس معیار C_p مدل سوم III را پیشنهاد می کند و همچنین معیار R_a^2 و PRESS نیز مدل III را توصیه می کنند.

در این مثال به خوبی نشان می دهد که R^2 معیار خوبی نیست چون R^2 مدل کامل (I) را توصیه می کند.

تمرین : برای این مثال و y^* تبدیل یافته معیارهای فوق را محاسبه و باهم مقایسه کنید.

۱،۸،۴- روشهای الگوریتمی :

روش های الگوریتمی در مقایسه با معیارهای R_a^2 و PRESS و C_p کمتر توصیه می شوند اما بعضی وقتها امکان محاسبه تمام (2^{k+1} مدل) مدلها ممکن نیست وقتی که مدل شامل k متغیر باشد.

از طرف دیگر وقتی که با نوع دیگری از رگرسیون کار می کنیم معیار R_a^2 و PRESS و C_p لزوما ممکن نیستند ولی روشهای الگوریتمی همیشه قابل استفاده هستند.

۱- روش پیشرو Forward Selection Method

در این روش با مدل بدون متغیر شروع می کنیم α_{in} را سطح معنی داری در نظر می گیریم و سپس x_j ها را یکی یکی به مدل اضافه و سپس آزمون می کنیم تازمانیکه نتوان متغیر دیگری را به مدل اضافه کرد (اضافه کردن متغیر جدید معنی دار نباشد).

$$\begin{cases} H_0 : y_i = \beta_0 + \varepsilon i \\ H_1 : y_i = \beta_0 + \beta_1 x_{ij} + \varepsilon i \end{cases}$$

به کمک آماره آزمون F و اصل مجموع توانهای دوم باقیمانده های اضافی هر مرحله را بررسی و آزمون می کنیم. یعنی متغیری را به مدل اضافه می کنیم که بیشترین اثر را در افزایش آماره F داشته باشد. فرض می کنیم این متغیر $x_{i(1)}$ باشد مادامیکه سطح معنی داری مشاهده شده آماره F کمتر از α_{in} باشد این کار را ادامه می دهیم. و این معادل افزایش R^2 نیز خواهد بود.

$$H_0 : y_i = \beta_0 + \beta_1 x_{i(1)} + \varepsilon i$$
$$H_1 : y_i = \beta_0 + \beta_1 x_{i(1)} + \beta_2 x_{ij} + \varepsilon i$$

این فرآیند را تا زمانی که همه متغیرها وارد مدل شوند یا اینکه از مرحله ای به بعد دیگر نتوان متغیری دی گری را به مدل اضافه کرد، ادامه می دهیم.

۲- روش پسرو(Backward Selection Method) :

اول α_{out} را مشخص می کنیم.

کار را با مدل کامل (شامل تمام متغیر های آزاد شروع) می کنیم و α_{out} را مشخص می کنیم. سپس متغیری که کمترین اثر را روی R^2 می گذارد از مدل حذف می کنیم . فرض کنید x_{ij} متغیر مورد نظر باشد.

$$\begin{cases} H_0 : y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{j-1} x_{i,j+1} + ... + \beta_{p'} x_{ip'} + \varepsilon i \\ H_1 : y_i = \beta_0 + \beta_1 x_{ij} + ... + \beta_{p'} x_{ip'} + \varepsilon i \end{cases}$$

این عمل را تا زمانی که سطح معنی داری مشاهده شده α^* کمتر از α_{out} شود تکرارمی کنیم.

تذکر مهم : یک اشتباه رایج در عمل این است که ابتدا مدل کامل در نظر گرفته می شود و پس از حذف یک یا دومتغیر از مدل ممکن است متغیر یا متغیرهای بروز کنند (خودنمایی کنند) که اصلا مهم نیستند.

۳- روش گام به گام (Stepwise Selection Method) :

این روش در حقیقت اجرای تکرار همزمان روشهای پیشرو و پسرو تواماً می باشد.

ابتدا α_{in} و α_{out} را مشخص می کنیم.

با مدل بدون متغیر و فقط شامل β_0 شروع می کنیم و

$$\begin{cases} H_0 : y_i = \beta_0 + \varepsilon i \\ H_1 : y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon i \end{cases}$$

پس از آزمون کردن (محاسبه F) در صورت معنی داری $\alpha_{in}^{*} < \alpha_{in}$ آنگاه

$$\begin{cases} H_0' : y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon i \\ H_1 : y_i = \beta_0 + \varepsilon i \end{cases}$$

روش پسرو را اجرا می کنیم و اگر معنی دار بود یعنی $\alpha_{out}^{*} < \alpha_{out}$.

این روش (استراتژی) تا زمانی تکرار می شود که نتوان دیگر هیچ متغیری را به مدل اضافه کرد و دیگر نتوان هیچ متغیری را ازمدل حذف کرد.

در حالت کلی $\alpha_{out} \in (0.05, 0.15)$, $\alpha_{in} \in (0.20, 0.30)$ در نظر گرفته می شود.

برای در نظر گرفتن تعدادی زیادی از مدل های ممکن خطاها را به صورت فوق در نظر می گیریم.

۱,۹- روشهای بررسی کردن مدل The Methods of Comparison Model :

سطوح معنی داری آزمون فرض ها وقتی معتبر هستند که شرایط اولیه (i- iv) مدل درست یا حداقل قابل قبول باشند. در این فصل شرایط درستی آزمون ها را بررسی می کنیم. ابتدا یادآوری می شود :

$I - E(Y|x) = \boldsymbol{x\beta} \equiv E(\boldsymbol{\varepsilon}) = \mathbf{0}$ شرط خطی بودن

هم واریانسی وناهمبسته بودن خطاها:

$$II - V(\varepsilon) = \delta^2 I,\ III - Cov(\varepsilon_i, \varepsilon_j) = 0; i \neq j$$

که درآن $I_{n\times n}$ یک ماتریس واحد (همانی) می باشد(قطر اصلی)(Identity matrix)

$$(IV) - Normality \quad \equiv \quad \varepsilon \sim N_n(0, \sigma^2 I)$$

۱,۹,۱- انواع باقیمانده ها :

مانند یادداشت فوق می توان همه آزمون ها راروی توزیع $\varepsilon_1,\ldots,\varepsilon_2$ انجام داد.

روشن است که بررسی شرایط فوق (I- IV) روی خطاها استوار خواهند بود و برآورد خطاها Residuals

که باقیمانده ها نامیده می شوند و به گونه های زیر تقسیم می شوند :

۱- باقیمانده های معمولی:

$$ev = y_i - \hat{y}_i = y_i - x_i'\hat{\beta}$$
$$e = (I - H)y = (I - H)\varepsilon$$

Commented [T1]:

Commented [T2]: hjhjg

یا به شکل ماتریسی

تحت فرض I :E(e)=0 . تحت فرض های IIIوII داریم :

$$var(e) = \delta^2(I - H)$$

$$v(e_i) = (1 - h_{ii})\delta^2 \quad , \quad cov(e_i, e_j) = -h_{ii}\delta^2$$

تحت فرض IV :

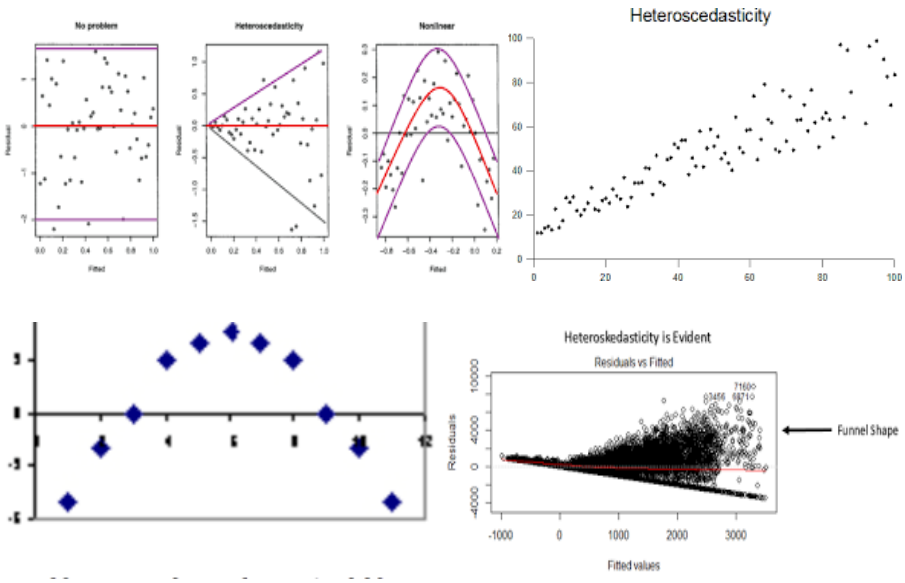
$$e \sim N_n(0, \sigma^2(I - H))$$

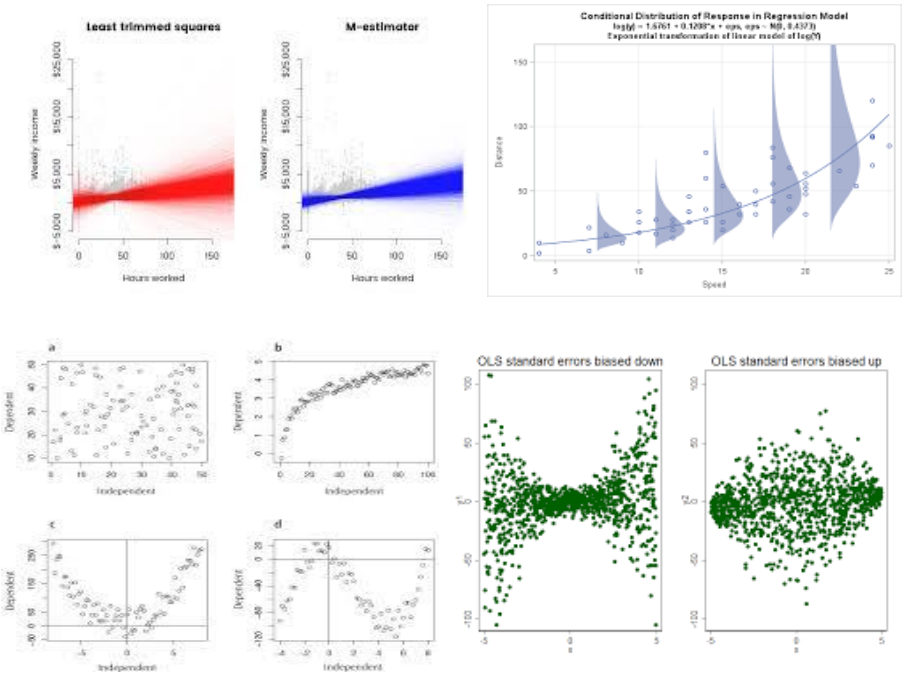
۲- باقیمانده ها Studentized Residuals :

شناسایی کردن identify

$$S_i = \frac{e_i}{S\sqrt{1-h_{ii}}}$$

انواع دیگری از باقیمانده ها رادر فصلهای بعدی بررسی خواهیم کرد که بطور مستقیم فرضهای مدل را بررسی نمی کند اما برای شناسایی (مشخص کردن) مدل های پرتوان در پیش بینی و شناسایی داده های پرت و موثر بکار می رود.





۱,۹,۶- گرافیک باقیمانده ها Residuals plots:

این نمودارها اغلب نشان می دهندکه کدامیک ازفرض های IV -I دارای مشکل هستند.

۱- رسم باقیمانده ها درمقابل برآورد y_i ها $(\hat{y}_i, e_i)$ ← lab 5

این نمودارمعمولا برای تشخیص خطی بودن فرض (I) مدل استفاده می شود. اگر فرض خطی بودن قابل e_i ها نسبت به $\hat{y}_i$ قبول است نمودار بصورت ابری از داده ها دراطراف y=0 را نشان می دهند. یعنی ها نزول یارشد نمی کنند به بیان دیگر e_i از $\hat{y}_i$ ها مستقل اند.

۲- باقیمانده های استودنت شده در مقابل $\hat{y}_i$ ها $(\hat{y}_i, s_i)$

این گرافیک مانند قبل مشکل غیر خطی بودن را بیان می کند و نیز اگر e_i ها هم واریانس (همگن) نباشند را نشان می دهد (اگر واریانس باقیمانده ها ثابت نباشد)

بعلاوه در صورت عدم اشکال s_i ها باید بین (۳و۳-) قرارگیرند.

اگر مقدار قدر مطلق بعضی از s_i ها بزرگتر از ۳ باشد بیان گر غیر نرمال بودن باقیمانده ها یا گویای داده های پرت (Outliers or Extrem Values) می باشد.

۳- نمودار باقیمانده های اولیه(معمولی) e_i در مقابلندیس مشاهدات متناظر (i,e_i).

این گرافیک مشکل همبسته بودن یا همبسته نبودن باقیمانده ها را بررسی می کند.

اگر باقیمانده های با مقادیر بزرگ(کوچک) پیرو باقیمانده های با مقادیر بزرگ (کوچک) باشند آنگاه مشکل خود همبستگی مثبت را بیان می کند.

اگرباقیمانده ها با مقادیر بزرگ(کوچک) پیرو باقیمانده های با مقادیر کوچک(بزرگ) باشند آنگاه مشکل خود همبستگی منفی را بیان می کند.

۴- باقیمانده های استودنت شده در مقابل چندکهای توزیع نرمال (u_i , s_i) این نمودار مشکل نرمال نبودن را بیان می کند.

بدین منظور از QQ- Plot نرمال یا خط راست هنری Henry استفاده می کند.

که در آن u_i ها چندکهای توزیع نرمال استاندارد می باشد(QQ norm) و برای رسم خط راست Henry از امید آماره های مرتب(ترتیبی) نرمال استاندارد استفاده می شود.

در هردو حالت اگر فرض نرمال بودن قابل قبول باشد نمودار باید خط راست با شیب مثبت باشند. اگر نمودار منحنی محدب یا مقعر باشند نشان دهنده این است که توزیع باقیمانده ها متقارن نیستند. بنابراین اگر نمودار به شکل انتگرال چپه یا خوابیده باشد نشان می دهد که باقیمانده ها می توانند از توزیعی آمده باشند که دمهای آن بسیار پهن تر (ضخیمتر) از توزیع نرمال است.

۵- Box- Plot باقیمانده ها :

این نمودار نشان می دهد که باقیمانده ها (معمولی یا استودنت شده) دارای توزیع متقارن هستند یا نیستند و داده پرت داریم یا نداریم.

۶- نمودار متغیر اضافه شده : ص۶۲

مثال-بررسی فایل Data-Boxcox.txt - In Example 14

مثال- در مثال مصرف بنزین و مدل $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon i$ را در نظر بگیرید و پس از بررسی مدل نمودار های :

۱- $(\hat{y}_i, e_i)$ ۳- (i, e_i)

۲- $(\hat{y}_i, s_i)$ ۴- (u_i, s_i)

همانطور که از نمودار $(\hat{y}_i, s_i)$ ملاحظه می شود مشکل نرمال نبودن یا عدم واریانس ثابت را داریم . با توجه Box Cox تبدیل $y^* = \frac{1}{\sqrt{y}}$ را انجام داده و سپس برای مدل جدید نمودار های فوق را رسم کنید. چه تفاوتی را ملاحظه می کنید؟

۱,۹,۳- آزمون فرضیه های معمول Formal Hypothesis Testing :

در بخش قبل گرافهای مفید را بررسی کردیم. علارغم اینکه این گرافها وجود مشکلات در فرض های مدل را بیان می کنند و نشان می دهند دور از انتظار نیست که به مواردی برخوریم که ابهام آمیز باشد. لذا آزمون های ذیل به ما کمک خواهند کرد تا ابهام در گرافها برطرف شوند.

۱- آزمون Shapiro- wilk برای بررسی نرمال بودن باقیمانده ها

`shapiro.test(x)`, `ks.test(x,"pgamma", 3, 2)#alternative: two-sided, g,l`

این آزمون عدم نرمال بودن باقیمانده ها را بررسی می کند. این آزمون در عمل آزمون خط راست Henry است. در این تست فرض H_0 (نرمال بودن باقیمانده ها) و $E(e_i) = \mu + \delta u_i$ که درآن u_i امید آماره ترتیبی دریک نمونه تصادفی به حجم n از یک توزیع نرمال استاندارد می باشد.

فرض کنید $V_{n\times n}$ ماتریس کواریانس آماره ترتیبی با عناصر $v_{ij} = \mathrm{cov}(u_{(i)}, u_{ij})$

و فرض کنید

$$S^2 = \sum_{i=1}^{n} (R_i - \bar{R})^2 \ , \ R^2 = u^T v^{-1} u \quad ,$$

$$C^2 = u^T v^{-1} v^{-1} u \quad , \quad \delta^2 = \frac{(u^T v^{-1} R)}{(u^T v^{-1} u)}$$

$$u^T = (u_{(1)}, ..., u_{(n)})$$

که Shapiro و wilk تعریف می کنند :

$$w = \frac{(\tilde{R}^2)^2 \hat{\delta}^2}{C^2 S^2}$$

باقیمانده ها از توزیع نرمال پیروی می کنند : H_0

فرض H_0 رد می شود هرگاه W کوچک باشد.

در مثال مصرف بنزین (Gas Consumaption)

اگر آزمون SW را روی باقیمانده های استودنت شده قبل از تبدیل داده ها انجام دهیم به

$p - value = 0.0165$ یعنی به RH_0 می رسیم. (Consummation نتیجه گیری)

اگر پس از تبدیل داده ها این آزمون را روی خطاهای باقیمانده استودنت شده انجام دهیم به

$p - value = 0.99$ یعنی AH_0 یا نرمال بودن خطاها می رسیم.

آزمون براساس تبدیل Box- Cox :

براحتی می توان از فاصله اطمینان برای مقداری از λ که تابع درستنمایی را ماگزیمم می کند فهمید که

آیا تبدیل داده ها لازم است یا خیر ؟

یعنی اگر عدد ۱ به این فاصله تعلق داشته باشد تبدیل لازم نیست و چنانچه $1 \notin \underset{1-\alpha}{CI(\lambda)}$ بیانگر عدم نرمال

بودن خطاها یا واریانس غیر ثابت خواهد بود.

در مثال مصرف بنزین : قبل از تبدیل $1 \notin \underset{0.95}{CI(\lambda)} = (-1.4, 0.5)$ پس از تبدیل (Box- Cox) داریم

$y^* = y^{-\frac{1}{2}} = \frac{1}{\sqrt{y}}$ داریم :

$$1 \in \underset{0.95}{CI(\lambda)} = [-0.35, 2.7] \Rightarrow$$

لزومی برای تبدیل مجدد داده نیست :

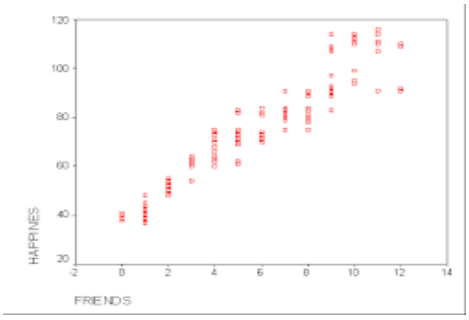
آزمون عدم تناسب Test For Lack of Fit (اساساً p=1 فرض می شود)

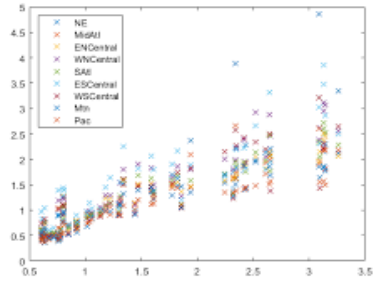
فرض کنید داده های $(x_1, y_1),\ldots,(x_n, y_n)$ در رگرسیون (خطی ساده p=1) که به m گروه تقسیم شوند.

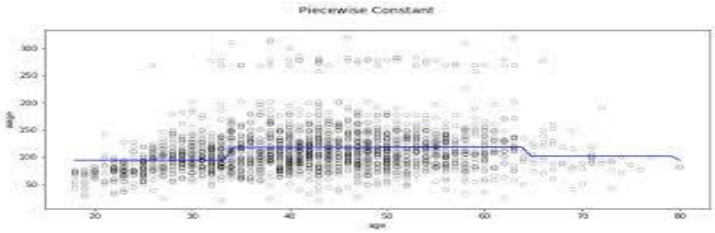
یعنی

$n = \sum_1^m n_i$ فراوانی

به بیان دیگر
$$\begin{cases} y_1 \quad y_2 \ldots y_{n1} , \ldots , y_{n-m+1} , \ldots , y_n \\ x_1 \quad \underbrace{x_1 \ldots x_1}_{n_1} , \ldots , \underbrace{x_m , \ldots , x_m}_{n_m} \end{cases}$$







خطای خالص (واقعی) Pure errors و خطای lack حفره (Lack Of Fit Errors)

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji} \quad ; \quad x = x_j$$

$$\text{Pure error} : SS_{Pure} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2$$

$$d.f(SS_{Pure}) = \sum_{j=1}^{m}(n_j - 1) = n - m$$

می توان نشان داد که :

$$\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ji} - \hat{y}_{ji})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{m} n_j(\hat{y}_j - \bar{y}_j)^2$$

$$<=> SS_{Res} = SS_{Pure} + SS_{LF}$$

اثبات (1998- Draper & Smith) Page47

حال اگر مدل خوب است باید SS_{Pure} بخش بزرگی از SS_{Res} را بیان کنند.

H_0 : مدل خوب است

H_1 : مدل خوب نیست

که $df(SS_{lf})=m-2$

$$F=\frac{\frac{SS_{LF}}{(m-2)}}{\frac{SS_{Pure}}{(n-m)}} \geq F_{\alpha\,;\,m-2\;,\;n-m} \Leftrightarrow RH_0$$

جدول آنالیز واریانس Table of Variance Analysis :

منبع تغییرات	df	SS	MS	F
Regression	1	SS_{Reg}	MS_{Reg}	$\frac{SS_{Reg}}{S^2}$
Residuls	n- 2	SS_{Res}	S^2	
Lake Of Fit	m- 2	SS_{LF}	MS_{LF}	$\frac{MS_{LF}}{S_P^2}$
Pure Error	n- m	SS_{Pure}	$MS_P=S_P^2$	
Total	n- 1	SS_T		

1.9.4 رفع مشکلات عدم همخوانی با فرضیات Solve the problems with the Hypothesis

در کل عدم تشخیص فرم درست بین متغیر پاسخ و متغیرهای کمکی (آزاد) هستند.

این مشکل ممکن است با یکی یا چندین گزینه زیر حل شود Solutions for the Case:

۱- تغییر متغیر (تبدیل) متغیرهای خاصی از متغیرهای کمکی یا اضافه کردن $x_{ij}^2, x_{ij}^3, \ldots$

۲- افزودن متغیر یا متغیر های کمکی جدید به مدل .

۳- افزودن اثرات متقابل به مدل مثل $(\beta_{jj'}, x_{ij}x_{ij'})$

۴- تبدیل متغیر پاسخ

واریانس غیر ثابت

اغلب تبدیل Box- Cox برای متغیر پاسخ جهت ثابت کردن واریانس استفاده می شود. در غیر اینصورت رگرسیون وزنی ممکن است به حل مسئله کمک کند.

علاوه بر این مدل های مختلط (بعدا ارائه خواهد شد) به حل مسئله کمک خواهد کرد.

خود همبستگی خطاها Autocorrelation in erorrs:

این مشکل به سختی قابل حل است . بعضی وقتها افزودن یک متغیر توصیفی که بیان می کند چرا این خود همبستگی وجود دارد مفید است. بعنوان مثال ارائه دلایل انتخاب مقادیر اولیه توسط فرد A ، انتخاب مقادیر بعدی توسط فرد B و ....،که گاهاً افزودن چنین متغیری به حل مشکل خود همبستگی کمک می کند.

اما در کل این مشکل با مراجعه مدل های خطی مختلط (بعداً ارائه می شود) یا سری های زمانی حل خواهد شد.

عدم نرمال بودن Non Normality:

همانطورکه قبلا نیزگفته شد تبدیل داده ها (متغیر پاسخ) توسط Box- Cox به حل مسئله کمک خواهد کرد . بعضی وقتها غیر نرمال بودن داده ها به دلیل چند داده پرت exterem data رخ می دهد، در این حالت از روشهای زیر استفاده می شود، یا روشهای دیگری مثل رگرسیون توانمند (تنومند) Robust Regression یا رگرسیون نا پارامتری non-parametric Reg استفاده کرد.

۱,۱۰- داده های پرت یا موثر Outliers and influential Data

این داده ها اثرات بسیارموثر و مهم روی برآورد پارامترهای رگرسیون و واریانس آنها خواهند داشت. ممکن است این داده ها مضر نیز باشند درنتیجه تشخیص این داده ها خیلی مهم هستند.

۱,۱۰,۱- lever یا فاصله هرداده یا مشاهده از مرکز آن متغیر کمکی مانند x_i از میانگین می باشد، یعنی $x_i - \overline{x}$.

یکبار دیگر عناصر قطر اصلی ماتریس H یعنی h_{ii} را در نظر می گیریم و در نتیجه $x_i - \overline{x}$ به صورت هم ارز با h_{ii} تغییر می کند. (برای توضیح بیشتر به sec.8.2.1 و Sen & Srivastava را ملاحظه کنید).

توجه داشته باشد که $1/n < h_{ii} < 1$ ولی نمی توان گفت که h_{ii} به مفهوم یک فاصله واقعی است.

با این وجود بعضی وقت ها از h_{ii} به عنوان فاصله x_i از $\overline{x}$ استفاده می شود به این منظور که هر چه h_{ii} بزرگتر باشد یعنی x_i از $\overline{x}$ دورتر است تفسیر می شود.

۱,۱۰,۲- داده های پرت Outlier Data

داده های پرت احتمالاً خیلی کم تحت فرضهای مدل مشاهده می شوند.

دررگرسیون خطی ساده نمودار (x_i,y_i) برای شناسایی کردن چنین داده هایی بکار می رود اگر چه چنین نموداری برای رگرسیون چندمتغیره مفید نخواهد بود در عوض نمودار (s_i,y_i) و خط راست هنری روی باقیمانده های استودنت شده برای شناسایی داده های پرت بکار می رود یعنی داده های متناظر خطاهای (باقیمانده های)استودنت شده که خارج از بازه [3,3 -] قرار می گیرند، داده های پرت محسوب می شوند.

نوع دیگری از خطاهای مفید برای تشخیص داده های پرت یا موثر عبارتست از RStudent که مانند باقیمانده های استودنت شده تعریف می شوند فقط واریانس حاضر و مخرج کمی متفاوت است.یعنی i امین واریانس بدون i امین مشاهده به صورت زیر محاسبه می گردد.

$$e_i^* = \frac{e_i}{S_{-i}\sqrt{1-h_{ii}}}$$

$$S_{-i}^2 = \frac{\dfrac{(n-p)s^2 - e_i^2}{(1-h_{ii})}}{n-p-1}$$

۱,۱۰,۳- داده های موثر: داده موثرعبارتست از مشاهداتی که به تنهایی اثری بزرگی روی برآورد پارامترها یا پیش بینی می گذارد.

حتی اگر این داده ها لزوماً پرت نباشند برای شناسایی و بررسی مفید هستند چون خطای حاصل از یک داده کاملاً اثر زیادی درانحراف نتیجه گیری خواهد داشت.

لزوماً مولفه های یک مشاهده دو اثر دارند :

I- فاصله از مرکز یا فاصله اهرمی (گشتاوری)

II- یک مقدارمتغیر پاسخ (y) خیلی بزرگ یا خیلی کوچک نسبت به بقیه مشاهدات با متغیرهای توصیفی مشابه.

لذا پیشنهاد می شود که چگونه باید اثر داده های موثر یا پرت را اندازه بگیریم.

اثر داده موثر روی پیش بینی :

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{S_{-i}\sqrt{h_{ii}}} = e_i^*\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$$

اگر $DFFITS_i \notin [-2,2]$ نشان دهنده این است که داده مربوطه موثر است.

اثر روی مقدار برآورد پارامترها: اثر **i** امین مشاهده روی **j** امین برآورد پارامتر $(\hat{\beta}_j)$ به صورت زیر اندازه گیری می شود.

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{S_{-i}\sqrt{(x'x)^{-1}_{jj}}}$$

اگر قدر مطلق $DFBETAS_{ij}$ از $2/\sqrt{n}$ بزرگترباشد آنگاه این مشاهده (i امین) اثر قابل توجهی روی برآورد β_j دارد .

توجه: اگر مجموعه داده ها بزرگ باشد محاسبه اثر تک تک مشاهدات روی p پارامتر طولانی و خسته کننده خواهد بود.

اما می توان با توجه به فاصله استاندارد شده بین $\hat{\beta}, \hat{\beta}_{-i}$ یعنی D_i را به صورت زیر بیان کرد.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})'(X'X)(\hat{\beta} - \hat{\beta}_{-i})}{ps^2} \qquad \text{(فاصله کوک)}$$

مادامی که D_iها جداشده ازبقیه باشند مشاهدات متناظر آنها رابه عنوان داده موثردر برآوردپارامترها درنظر می گیریم و شایان ذکراست کهDFBETAS را ملاحظه کنیم.

اثر روی واریانس برآوردگرها :

اندازه اثر به صورت زیر داده می شود:

$$COVRATIO_i = \frac{\|S^2_{-i}(X'_{-i}X_{-i})^{-1}\|}{\|S^2(X'X)\|} = \frac{(s^2_{-i})^p}{(S^2)^p}\left(\frac{1}{1-h_{ii}}\right)$$

که درآن II.II نشان دهنده ترمینان است.

اگر مقدار COVRATIO نزدیک ۱باشد یعنی این مشاهده اثر کمی روی واریانس دارد واگر نزدیک صفر یا مقداری بزرگ داشته باشد بیانگر اثر زیاد است.

به بیان دیگر مشاهداتی که $\text{COVRATIO}_i \notin (1 \pm {}^{3p}/_{n})$

بعنوان مشاهدات موثر روی واریانس در نظر گرفته می شوند.

Residuals , rstudent , covratio, dffits , dfbetas (mod)diag hat H?

۱. ۱۰. ۴ – با داده های موثر یا پرت چکار باید کرد؟

اغلب این داده با خطا و بر خلاف قاعده حاصل می شوند. که باید روی این مشاهدات اطلاعات کافی جمع آوری کرد.درصورت امکان داده رااصلاح کرد درغیراینصورت آنرا حذف کرد.

متاسفانه این اعداد یا مشاهدات پیش می آید و تا جایی که موجب انحراف در تغییر و نتیجه گیری نشود آن ها را از داده ها حذف نمی کنیم.

۱۱. ۱- هم خطی چند گانه (چند هم خطی) Multicolinear

چند هم خطی مشکلی است که به ساختار ماتریس X بستگی دارد. لذا لازم است قبل آنالیز رگرسیون وجود هم خطی را بررسی کنیم. لازم به ذکر است که تا کنون فرض کردیم ماتریس X دارای رنگ کامل است یعنی ماتریس X′X معکوس پذیر است و چند هم خطی وجود ندارد. گاهاً برای رفع این مشکل ممکن است متغیری را به X اضافه کرد. چون اگر X دارای رنک کامل باشد تضمین می کند که بردار یکتای $\hat{\beta}$ مجموع توان های دوم خطای رگرسیون یعنی SS_{Res} را مینیمم می کند.

اگر یکی از ستون های X تابعی خطی از بقیه باشد آنگاه ماتریس X′X معکوس پذیر نبوده وبیشتر ازاین نمی توان جلو رفت. برعکس اگر بعضی از ستون های X دقیقاً تابعی خطی از بقیه نباشد(ولی خیلی نزدیک باشد) وحتی بتوان یک برآوردگر یکتای $\hat{\beta}$ نیز پیدا کرد واریانس بعضی از مولفه های $\hat{\beta}$ بسیار بزرگ خواهد بود ودر نتیجه پیش بینی های بسیار متغیر و دور از واقعیت رابه همراه خواهدداشت و حتی می تواند باوجود مهم بودن یک متغیر آن را معنی دار در مدل تشخیص ندهد.

۱ . ۱۱ .۱- تشخیص هم خطی بودن:

اولاً ماتریس همبستگی بین متغیرهای کمکی (آزاد) را محاسبه می کنیم.

فرض کنید :

$$x_{ij}^{*} \coloneqq \left(x_{ij} - \bar{x}_j\right)/s_j \quad , \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$s_j^2 = \frac{1}{n-1}\sum(x_{ij} - \bar{x}_j)^2$$ باشد.

آنگاه ماتریس ضرایب همبستگی نمونه ای داده ها عبارتند: $X^{*'}X^{*}$.

اگر دومتغیر از متغیرهای آزاد همبسته خطی باشند آنگاه ضریب همبستگی بین آنها بزرگ (نزدیک به ۱) است. که دراین صورت با دو مشکل مواجه هستیم:

۱- چه مقدار از ضریب همبستگی را بزرگ تلقی کنیم .

۲- دراکثر حالتهای هم خطی ، هم خطی بین بیش از دو متغیر رخ می دهد (چندهم خطی). برای ملاحظه این مشکل می توان P متغیر، که کاملاً همبسته خطی هستند را تولیدکرد درحالی که ضرایب همبستگی بین زوج متغیر ها کمتراز خواهند بود.

یک روش که به نظر می آید اندازه گیری سطح همبستگی بین هرمتغیر توصیفی با بقیه متغیرها را محاسبه کرد، برقراری یک رگرسیون خطی بین j امین X_j و بقیه متغیرهای توصیفی است. چنانچه R_j^2 بزرگ باشد نشان می دهد که بخش بزرگی از تغییرات رگرسیون توسط متغیرهای غیر از X_j تامین می شود.(p- 1 متغیر دیگر)

این معیارتشخیص هم خطی را در(VIF) بصورت زیرارائه می کنیم.

Variance Inflation

$$VIF_j = \frac{1}{1-R_j^2}$$

$$Var\left(\widehat{\beta_j}\right) = \frac{\sigma^2}{x'x_{jj}}\ VIF_j$$

یعنی اگر R_j^2 نزدیک عدد یک باشد آنگاه VIF_j عددی بزرگ خواهد شد و در نتیجه واریانس $\widehat{\beta_j}$ نیز بزرگ خواهد شد.

یکی از پیشنهادها برای تشخیص هم خطی بین x_j و بقیه این است که $VIF_j >$ ۱۰ باشد.

از نقاط ضعف VIF یکی این است که چند هم خطی ستون :

۱- از ماتریس X را نمی تواند تشخیص دهد.

۲- ناتوان در تشخیص متغیرهای همبسته خطی است.

۳- مقدار دقیقی برای VIF به عنوان چند هم خطی مشخص نمی کند.

معیار دیگری برای محاسبه تورم واریانس به عنوان Tolerance یا TOL_j بصورت:

$$TOL_j = 1 - R_j^2 = \frac{1}{VIF_j}$$

تعریف می شود.

همانطور که ملاحظه می شود TOL_j بخش یا درصدی از تغییرات را که توسط متغیرهای توصیفی دیگر بیان نمی شود را بیان می کند. پر واضح است که اگر VIF_j >۱۰ آنگاه ۱۰٪ >TOL_j (معادل) این است که حداقل ۹۰ ٪ تغییرات توسط بقیه متغیرهای توصیفی بیان می شود.

رهیافت اندازه گیری اثر چند هم خطی براساس خواص ماتریس $X^{*'}X^*$ بیان می شود بصورت زیر تعریف می گردد.

تعریف: چند هم خطی وقتی استفاده می شود که یک ترکیب خطی از ثابت های $C_1, \ldots, C_p$ بطوریکه همه همزمان صفرنیستند و $\sum_{j=1}^{p} c_j x_j^* \approx 0$

بنابه تعریف ماتریس استاندارد X^* داریم : $\text{tr}\left(X^{*'}X\right) = p = \sum_{j=1}^{p} \lambda_j$

یا

$$V'\left(X^{*'}X^*\right)V = diag\ (\lambda_1, \ldots, \lambda_p)$$

که در آن λ_j ها مقادیر ویژه ماتریس $X^{*'}X^*$ بوده و اگر یک چندهم خطی وجود داشته باشد دترمینان $X^{*'}X^*$ که برابر $p = \sum_{j=1}^{p} \lambda_j$ و $\cong \prod_{j=1}^{p} \lambda_j$ ؟

که در آن p تعداد متغیرهای آزاد می باشد. با وجود چند هم خطی $\cong \prod_{j=1}^{p} \lambda_j$ 0 یعنی بعضی از λ_j ها تقریبا صفر خواهد بود.

$$\emptyset_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}} \quad ; \quad j = 1,2,\dots,p$$

اگر $\emptyset_j > 30$ باشدآنگاه نشان دهنده یک همبستگی خطی بین X_j و بقیه متغیرهای توصیفی می باشد. فرض کنید V_j بردار ویژه متناظر با مقدار ویژه λ_j باشد آنگاه عناصر بردار ویژه می توانند ما را در شناسایی متغیرهایی که با *j*امین X همبستگی خطی دارندکمک کنند. می توان نسبت (احتمال) واریانس براوردگر $\hat{\beta}_l$ *که بوسیله j*امین همبستگی خطی بیان می شود را به صورت زیر بدست آورد.

$$P_{lj} = \frac{{V_{lj}}^2/\lambda_j}{C_{jj}} > 0.6$$

به طوریکه $C_{jj} = \sum_{l=1}^{p} \frac{{V_{lj}}^2}{\lambda_j}$

در نتیجه الگوریتم زیر برای شناسایی چند همخطی بکار می رود.

۱. شنساسایی یا مشخص کردن $\emptyset_j$ هایی که مقدارشان از ۳۰ بزرگتر هستند.

۲. برای هرکدام از Φ_jها که بزرگتراز ۳۰ هستند p_{lj}ها متناظر l =1,...p را محاسبه کنید.

۳. متغیرهای توصیفی متناظربا p_{ij}های بزرگتراز *0.60* (معمولاً) درگیر چند هم خطی هستند.

مثال – برای مسئله بنزین چند هم خطی رابررسی کنید (توجه داشته باشید که برای ستون ۱ماتریس *X* نمی توان این موضوع را بررسی کرد.)

بررسی چند هم خطی: *VIF(mod)*

if sqrt (vif(mod))>2⟹problem?

چنانچه با ستون ۱ این آزمون را (بررسی)انجام داد مشکل هم خطی پیش می آید. ص ۸۲ *T*

تذکر :اگر در داده ها وجود *VIF*های بزرگ مشاهده می شود اما آنالیز مقادیرویژه ماتریس $X^{*'}X^{*}$ مشکلی رانشان نمی دهد بااحتمال زیاد ستون۱ ماتریس *X*درگیر یک همبستگی خطی بین ستونها می باشد دراین صورت از الگوریتم فوق می توان استفاده کرد اما بجای ماتریس $X^{*'}X^{*}$ از $\widetilde{X'}\tilde{X}$ که درآن $\tilde{X}$ همان ماتریس *X*است که استانداردشده ولی مرکزی نشده است . یعنی $\tilde{X}_{ij} = {x_{ij}}/{S_j}$

راه حل های ممکن:

اگر مشکل چند هم خطی پیش آید چکارباید کرد؟ اولین و ساده ترین راه حل حذف شبه همبستگی ها خطی است یعنی کاهش ابعاد ماتریس *X* .

ونیز میتوان متغیرهای دارای چند هم خطی را جایگزین کرد. فرض کنید x_6 , x_5, x_1 همبسته خطی باشند.

در این صورت به جای سه متغیر میتوان دو متغیر رابه صورت های ذیل جایگزین کرد.

مثلاً: x_1 , x_5 یا x_1 , $\frac{x_5+x_6}{2}$ یا x_1 , $\frac{x_1+x_5}{2}$, $\frac{x_1+x_6}{2}$

و یا...

که بعضی از محققین خاص هم متغیرهای توصیفی در مدل نگه می دارند. که دراین صورت لازم است یک تبدیل غیر خطی روی متغیرهای توصیفی انجام دهیم(که مشکل هم خطی دارند) یا یک روش برآورد غیراز کمترین توان های دوم استفاده کنیم. اما دراین روش ها به راحتی برآوردگرهای نااریب تولید نمی شود اما برعکس واریانس آنها خیلی کم متاثر از وجود چند هم خطی خواهند بود.

نمونه هایی ازاین روش ها می تواند رگرسیون مولفه های اصلی یا رگرسیون ریج (پل) می باشد.

رگرسیون ریج

توسط رامین خاورزاده/ در مدل سازی و رگرسیون/دیدگاه ۰/فروردین ۶, ۱۳۹۵

هارل و کنارد اولین بار در مقاله ای در مجلهٔ مهندسی شیمی، بیان کردند زمانی که در مدل هم خطی رگرسیون ریج ، می توان به جای برآوردگر های کمترین مربعات باقیمانده از برآوردگر زیر β وجود داشته باشد، برای برآورد :استفاده کرد

$$\widehat{\beta^*}=[X'X+KI]^{-1}X'Y$$

به دست می آیند، شباهت های ریاضی زیادی با تعریف توابع پاسخ درجه دوم دارند. به این $K\geq 0$ برآوردهایی که با .انجام می شود رگرسیون ریج نامیده می شوند $\widehat{\beta^*}$ دلیل، برآوردها و تحلیل هایی که با

:رابطهٔ برآورد ریج با برآورد کمترین مربعات باقیمانده، به شکل زیر است

$$\tilde{b} = \widehat{\beta^*} = Z\widehat{\beta}$$

به عبارتی برآوردگر ریج یک ترکیب خطی از برآوردگر کمترین مربعات باقیمانده است، به طوری که

$$Z = [X'X + KI]^{-1}X'X$$

است و خاصیت گاوس مارکوف، β برآوردی نا اریب برای $\widehat{b}$ در روش کمترین مربعات باقیمانده، فرض کردیم .مینیمم بودن واریانس این برآوردگر در کلاس برآوردگرهای خطی نا اریب را بیان می کند

به دست می آید که β در محاسبهٔ برآوردگر ریج از فرض ناااریب بودن چشم پوشی شده و برآوردگر اریبی برای .دارای واریانس کمتری نسبت به برآوردگر کمترین مربعات باقیمانده است

بنابراین مقدار برآورد در این حالت پایدارتر خواهد بود، یعنی با حذف یا افزودن یک متغیر پیشگوی جدید به مدل و یا .... ضرایب برآورد شدهٔ مدل تغییر چندانی نخواهند کرد

استاندارد شده باشند، از حل معادلات نرمال x ، با فرض اینکه ستون های ماتریس ($\tilde{b}$) در محاسبهٔ برآوردگر ریج :داریم

$$\widehat{b}(k) = (x'x + kI)^{-1}x'y$$

مقداری مثبت است که بایستی توسط تحلیل گر با توجه به یک سری معیارها تعیین شود. با توجه به k که در آن مقدار این نکته که $MSE(\tilde{\ }) = \text{var}iance + (bias)^2$

مقداری را برمی گزینیم که کاهش در واریانس برآوردگر اریب، بیش از افزایش مربع اریبی k بنابراین برای انتخاب .آن کمتر از واریانس برآوردگر ناااریب خواهد بود MSE باشد. در نتیجه

۱.۱۲ – متغیرهای توصیفی(گروهی)

۱.۱۲.۱– متغیرهای توصیفی دومقداری یاچندمقداری(کیفی)

تاکنون با متغیرهایی سروکارداشتیم که پیوسته بودند.ولی درعمل با متغیرهایی برخوردمیکنیم که بطورگسسته

مقادیری را می پذیرند.

مثال:

$$X_{1i} = \begin{cases} \text{H}; & \text{اگر مرد باشد} \\ F; & \text{اگر زن باشد} \end{cases}$$

$$X_{1i} = \begin{cases} 1; & \text{اگر شهر باشد} \\ -1; & \text{اگر روستا باشد} \end{cases}$$

$$X_{1i} = \begin{cases} 1; & \text{اگر روشن باشد} \\ 0; & \text{اگر خاموش باشد} \end{cases}$$

که همیشه می توان این چنین درنظر گرفت(تعریف کرد)

$$X_i = \begin{cases} 1; & \text{اگر صفیت را دارد} \\ 0; & Sinon\,, o.\omega. \end{cases}$$

بنابراین مدل $y_i = \beta_0 + \beta_1 + x_i + \varepsilon_i$ را در نظر می گیریم داریم:

اگر $X_i = \begin{cases} -5; & \text{مرد} \\ 5; & \text{زن} \end{cases}$ فرض شود خیلی مفید نخواهدبود و تغییر پارامترها مشکل می شود.

$$E\left(y_i; \quad \text{مرد } i\right) = \beta_0 - 5\beta_1$$

$$E\left(y_i; \quad \text{زن } i\right) = \beta_0 - 5\beta_1$$

$$\Longrightarrow \beta_1 = 0.1\ d_1 ff\ between\ H\ and\ F$$

$$\beta_1 = 0.1\ DHF$$

ولی اگر $X_i = \begin{cases} 1; & \text{مرد} \\ 0; & \text{زن} \end{cases}$ در نظر بگیریم داریم:

$$E\left[y_i\ ;\ i \text{ مرد باشد}\right] = \beta_0 + \beta_1$$

$$E\left[y_i\ ;\ i \text{ زن باشد}\right] = \beta_0$$

؟

اگر متغیری چند مقدار داشته باشد مثلاً C مقدار که C ≥2. مثال:

$$x_i = \begin{cases} 1 \ ; \ i \ is\ blue \\ 2 \ ; \ i \ is\ red \\ 3 \ ; \ i \ is\ yellow \\ 4 \ ; \ i \ is\ green \end{cases} \qquad polytomic$$

دراین حالت C - 1 متغیر دومقداری برای آن تعریف می کنیم:

$$x_{iB} = \begin{cases} 1 \ ; \ i \ is\ blue \\ 0 \ ; \quad o.\omega \end{cases}$$

$$x_{iR} = \begin{cases} 1 \ ; \ i \ is\ red \\ 0 \ ; \quad o.\omega \end{cases}$$

$$x_{iy} = \begin{cases} 1 \ ; \ i \ is\ yellow \\ 0 \ ; \quad o.\omega \end{cases}$$

$$x_{1i} = blue \implies \left(x_{iB}, x_{iR}, x_{iy}\right) = (1,0,0)$$

$$x_{1i} = red \implies \left(x_{iB}, x_{iR}, x_{iy}\right) = (0,1,0)$$

$$x_{1i} = yellow \implies \left(x_{iB}, x_{iR}, x_{iy}\right) = (0,0,1)$$

$$x_{1i} = green \implies (x_{iB}, x_{iR}, x_{iy}) = (0,0,0)$$

$$y_i = \beta_0 + \beta_B \; x_{iB} + \beta_R \, x_{iR} + \beta_{y¢} x_{iG} + \varepsilon_i$$

$$E[y_i; \quad i = Blue] = \beta_0 + \beta_B$$

$$E[y_i; \quad i = Red] = \beta_0 + \beta_R$$

$$E[y_i; \quad i = Yellow] = \beta_0 + \beta_Y$$

$$E[y_i; \quad i = Green] = \beta_0$$

دراینجا می توان آزمون های زیر را درنظر گرفت .

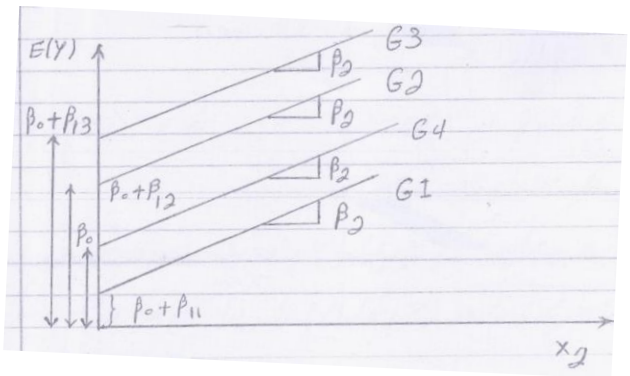
؟

مثال ۹ .۱- فرض کنید x_{i1} انواع محصولات و x_{i2} غلظت نمک در محصول i ام باشد و y_i نشان دهنده کیفیت محصول i باشد. y_i و x_{i2} پیوسته و متغیر x_{i1} چند مقداری شامل $\{1,2,3,4\}$ می باشد به کمک نمودار $E(y_i)$ تابعی از x_{i2} است می توان پارامترهای مدل را تغییر کرد.

$$y_i = \beta_0 + \beta_{11} x_{11i} + \beta_{12} x_{12i} + \beta_{13} x_{13i} + \beta_2 x_{2i} + \varepsilon_i$$

بطوریکه

$$x_{11i} = \begin{cases} 1 & ; \quad x_{1i} = 1 \\ 0 & ; \quad x_{1i} \neq 1 \end{cases}$$



$$x_{12i} = \begin{cases} 1 & ; \quad x_{1i} = 2 \\ 0 & ; \quad x_{1i} \neq 2 \end{cases}$$

$$x_{13i} = \begin{cases} 1 & ; \quad x_{1i} = 3 \\ 0 & ; \quad x_{1i} \neq 3 \end{cases}$$

یعنی y تابعی از x_2 است در هر گروه x_1

آزمون اثر یک متغیر چند گانه (چند مقداری):

درکل برای آزمون یک متغیر چندمقداری(C) باید($C-1$)پارامتر را همزمان صفر قرار داد.

$$\begin{cases} H_0: \ \beta_{11} = \ \dots = \beta_{1c-1} = 0 \\ H_1: \quad o.\omega \end{cases}$$

انتخاب مدل :قانون یا معیار انتخاب مدل متناظر β در یک متغیر چند مقداری است یعنی اینکه همه پارامترها باید در مدل باشند یا هیچکدام در مدل نباشند مثال قبل را در نظر بگیرید. فرض کنید بهترین مدل

$$y_i = \beta_0 + \beta_{11}x_{11i} + \beta_{13}x_{13i} + \varepsilon_i$$

اگر بخواهیم$\beta_{12}x_{12i}$ را اضافه کنیم چون $\beta_{11},\beta_{12},\beta_B$ از طریق متغیری"گروه یا کلاس" به هم وابسته هستند پس وجود یکی معادل وجود همه متغیرها می باشد.

پس در فرآیند آزمون در مواجه با یک متغیرتوصیفی چند مقداری مانند یک متغیرتکی عمل می کنیم یعنی یاهمه متغیرهای نشانگر (چند مقداری)را درمدل یکی قرارمی دهیم یا هیچکدام.

اثر متقابل: Interaction

برای نشان دادن اثرمتقابل دومتغیر مثل x_{1i} , x_{2i} رابا $\beta_{12}x_{1i}x_{2i}$ درنظر می گیریم.اثرمتقابل وقتی لحاظ می شود که تاثیر روی میانگین y داشته باشد.

مثال- *L10*- فرض کنید $y_{10} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$

اثر افزایش یک واحد x_{1i} روی میانگین y_i چقدر است؟

$$1 - E[y_i;\ x_{1i} + 1] = \beta_0 + \beta_1 (x_{1i} + 1) \beta_2 x_{2i} + \beta_{12} (x_{1i} + 1) x_{2i}$$

$$2 - E[y_i;\ x_{1i}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i}$$

$$\Longrightarrow 1 - 2 = \beta_1 + \beta_2 x_{2i}$$

؟

مثال- فرض درفوق x_{1i} دارای سه مقدار باشد بنابراین معادله رگرسیون:

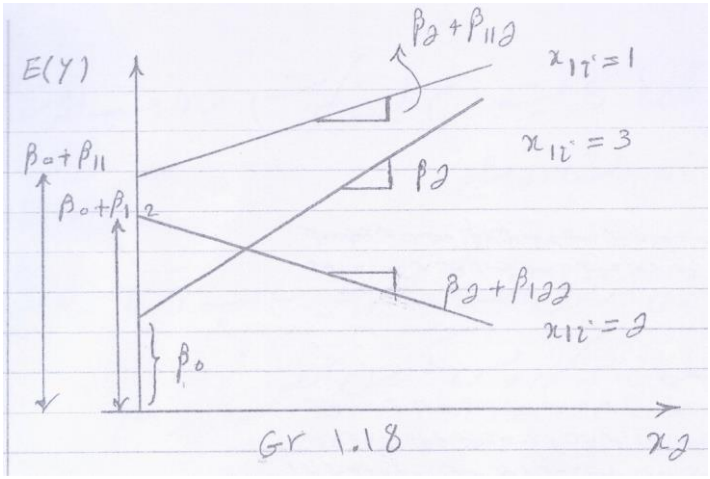
– ۱. ۱۱

$$y_i = \beta_0 + \beta_{11} x_{11i} + \beta_{12} x_{12i} + \beta_2 x_{2i} + \beta_{112} x_{11i} x_{2i} + \beta_{122} x_{12i} x_{2i} + \varepsilon_i$$

بطوریکه

$$x_{11i} = \begin{cases} 1 & ; \quad i = 1 \\ 0 & ; \quad o.\omega \end{cases}$$

$$x_{12i} = \begin{cases} 1 & ; \quad i = 2 \\ 0 & ; \quad o.\omega \end{cases}$$



همانطورکه ملاحظه میشود اثرات متقابل بین یک متغیر پیوسته و یک متغیر چندمقداری جالب و مهم می باشد .

انتخاب اول: اگر اثرمتقابل چندین متغیرمستقل یا کمکی در مدل است آنگاه درکل میتوان اثرات متغیرهای متناظر را در مدل قرار داد.مثال زیر را در نظر بگیرید:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_{12} x_{1i} + \beta_{13} x_{1i} x_{3i} + \beta_{23} x_{2i} x_{3i} + \varepsilon_i$$

چنانچه پس ازآزمون پارامترهای β_{23} , β_1 مهم باشند(فقط) آنگاه به دلیل β_{23} باید متغیرهای متناظر β_2 , β_3نیز در مدل قرار گیرند. یعنی

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_{23} x_{2i} x_{3i} + \varepsilon_i$$

اثرمتقابل (به دومتغیر)محدود نیست یعنی اگر متغیرهای مستقل عبارتند از x_i ; i=1,...,4

آنگاه مدل مناسب می تواند :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \cdots +$$

$$\beta_{1234} x_{1i} x_{2i} x_{3i} x_{4i} + \varepsilon_i$$

تذکر :وقتی که تعداد داده ها زیادنیست (حجم داده ها کم است)به ندرت اثرمتقابل بیش از دو متغیر مستقل استفاده می شود .

آزمون خطی بدون : برای آزمون اینکه آیا رابطه خطی مناسب است یا خیر ، می توان اثرات متقابل را به مدل اضافه کرد و سپس آزمون نمود.

مثال ۱,۱۲:

فرض کنید مدل $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ آیاخطی بودن(رابطه) بین x_{1i} , x_{2i} معتبر است؟

میتوان اثرمتقابل آنها را به صورت زیر آزمون کرد:

$$\begin{cases} H_0: \beta_{12} = 0 \\ H_1: \beta_{12} \neq 0 \end{cases}$$

برای مدل

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

اگر RH_0 آیا مدل قبلی (بدون اثرمتقابل)مناسب نیست .

۱ .۱۳

مدل رگرسیون پیشرفته:

۱ .۱۳ .۱ – مدل چند جمله ای:

آخرین ایده "رگرسیون خطی" یک مدل خطی یک تقریب خوب از رابطه واقعی بین متغیرهای مستقل و پاسخ می باشد که :

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

دربیشتر حالات یک مدل خطی «رگرسیون خطی» یک تقریب مناسب از $f(x_1, \dots, x_k)$خواهد بود. برعکس درحالات خاصی مدل خطی «رگرسیون خطی» به اندازه کافی نزدیک به $f(x_1, \dots, x_k)$ نخواهد بود. دراین حالت اصولاً میتوان گفت که برای بدست آوردن یک تابع هموار «نرم» از بسط چندجمله ای استفاده شود تا برای براورد تابع «رابطه» حقیقی$f(x_1, \dots, x_k)$ مناسب شود. امامشکل انتخاب بیشترین درجه برای x_i هاست. یعنی حداکثر چه درجه ای مناسب خواهد بود.

یعنی :

$$y_i = f(x_1, \dots, x_k) + \varepsilon_i$$

$$\approx \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

$$+\beta_{11} x_{1i}^2 + \cdots + \beta_{kk} x_{ki}^2$$

$$+\beta_{111} x_{1i}^3 + \cdots + \beta_{kkk} x_{ki}^3$$

⋮

$$+\beta_{1\dots1}x_{1i}^{d}+\cdots+\beta_{k\dots k}x_{ki}^{d}+\cdots+\varepsilon_i$$

درعمل به ندرت درجه x_i ها از ۲ بیشتر می شوند یعنی

$$y_i=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}+\beta_{11}x_{1i}^2+\cdots+\beta_{kk}x_{ki}^2+\beta_{12}x_{1i}x_{2i}+\cdots$$
$$+\beta_{k-1;k}x_{k-1,i}x_{ki}+\varepsilon_i$$

این ممکن است که وقتی ازمدل چندجمله ای استفاده می کنیم به شکل عددی یا چند هم خطی برخورد کنیم. دراین حالتها توصیه می شود که با متغیرهای مستقل استاندارد کارکنیم. یعنی بجای x_{ji} از $x_{ji}^*=\frac{(x_{ji}-\bar{x}_j)}{s_j}$ استفاده کنیم که درآن

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ji}\quad,\quad s_j^2=\frac{1}{n-1}\sum(x_{ji}-\bar{x}_j)^2$$

انتخاب مدل:مانند حالت اثرات متقابل، اگرعبارتی بادرجه d درمدل است پس باید تمام درایج کمتراز d نیز درمدل قرار گیرد.

تذکر : ترجیح داده می شود که برای گرفتن نتیجه خوب با کمترین پارامتر یک تبدیل برای y یا بعضی از xها انجام داد بجای استفاده از یک مدل چند جمله ای .

مثال- مدل $\ln y_i=\beta_0+\beta_1x_{1i}+\varepsilon_i$

یا $y_i=\beta_0+\beta_1e^{x_{1i}}+\varepsilon_i$

ممکن است بر مدل $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \varepsilon_i$ ترجیح داده شود که دراین صورت آزمون باقیمانده ها یا مقدارمعیارهای انتخاب مدل اغلب برای تعیین مدل مناسب کمک می کند «با توجه به داده ها»

آزمون خط لیون:

مانند حالت اثرات متقابل عبارتهای بادرجه بیشتر ممکن است برای بردازش یا بررسی یک مدل به آن اضافه شود.

مثال ۱۳. ۱- فرض کنید مدل $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

برای بررسی خطی بودن مدل فوق می توان برای مدل زیر:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{i1}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

فرض های $\begin{cases} H_0: \beta_{11} = \beta_{22} = \beta_{12} = 0 \\ H_1: o.w \end{cases}$

حداقل یکی از پارامترها صفر نباشد.

اگر RH_0 آنگاه مدل اولیه مناسب نیست.

فصل سوم: مدل های عمومی «کلی» رگرسیون خطی با اثرات آمیخته

مقدمه :همانطورکه درفصول قبل ملاحظه کردید مدلهای رگرسیون خطی دربسیاری ازموارد انعطاف پذیر و مناسب برای بررسی ارتباط بین متغیرپاسخ و متغیرهای آزاد بودند. اما موارد زیادی هم پیش خواهدآمد که

این مدل ها مناسب نیستند یعنی اگرچه فرض می کردیم که متغیرهای پاسخ باید ناهمبسته باشند و واریانس ثابت ولی درعمل این چنین نیستند که در این حالت مدلهای قبل جواب گوی مسئله نیستند.

مثال ۱ .۳-

فرض کنید ۶ موش از ۳ خانواده و ازهر خانواده ۲ موش داریم.

فرض کنید در مدت ۱۰ روز به موش iام مقدار x_i پروتئین می دهیم و مقدار افزایش وزن این موش را y_i می نامیم.

$$(y_i, x_i); \quad i = 1,2, \dots ,10$$

درنگاه اول مدل خطی ساده راکه درفصول قبل آموختیم پیشنهاد می کنیم.

مانند $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; i = 1, \dots ,10$ و روشهای استنباط فصل قبل را درنظر می گیریم یا اعمال می کنیم.

اما به احتمال زیاد این روش ها دراینجا مناسب نیستند. چون واکنش موش ها درهر خانواده نسبت به پروتئین اضافه داده شده از همگنی بیشتری برخورداراست. به نظر می آید ولی مناسب خواهد بود که همبستگی داخل خانواده ها رانیز درنظر داشته باشد یاشامل متغیرهای توصیفی که تفاوت بین خانواده ها رابیان می کند باشد.

مثال۲,۲- فرض کنید یک کمپانی جنگلبانی متقاضی دو نوع درخت از نوع A و B رادارد بطوریکه بیشترین رشد قدی را در اسرع وقت دراین سرزمین داشته باشد.بدین منظور ما ۴ نهال ازنوع A و ۴نهال ازنوع B انتخاب و بصورت تصادفی می کاریم وسپس به مدت ۳ سال و هرسال یکبار قد آنهارا اندازه می گیریم.

در نتیجه مدل $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t_j + \beta_3 x_i t_j + \varepsilon_{ij}$

که درآن t_j زمان اندازه گیری قددرختان و y_{ij} اندازه قد درخت iام در زمان t_j و $x_i=1$ اگر درخت iام از نوع A باشد در غیر اینصورت $x_i = 0$ یعنی درخت از نوع B است.

اگر $\beta_3 > 0$ بنابراین درختان نوع A دارای نرخ رشد بیشتر نسبت به نوع B هستند اما باوجود این روش ممکن است مشکلات زیر رخ دهد:

۱- همه درختان با قدیکسان کاشته نمی شوند.«یعنی β_0های متفاوتی برای هردرخت وجوددارد»

۲- درختان نرخ رشد متفاوتی دارند. «یعنی β_2 برای هردرخت متفاوت است»

۳- ε_{ij}ها ممکن است همبسته باشند.

مثال- ۲۳- می خواهیم اثر درجه حرارات آب یک دریاچه را بر روی مقدار PH آب آن بررسی کنیم . بدین منظور نمونه هایی از آب سطح دریاچه راجمع آوری کرده و بلافاصله درجه حرارت و مقدار PH آنرا اندازه می گیریم. فرض کنید x_i اندازه درجه حرارت iامین مشاهده و y_{ij} اندازه PH iامین مشاهده توسط jامین دستگاه اندازه گیری PH باشد.

بنا به فصل قبل مدل پیشنهادی عبارت است از:

$$y_{i,j} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$$

اما دراینجا باید مواظب باشیم چون متغیرهای تصادفی $\varepsilon_{i,j}$ و $\varepsilon_{i,j'}$ ممکن است همبسته باشند حتی برای نمونه های یکسان.

-۲,۲

مدل مختلط خطی تعمیم یافته

فرض کنید

$Y_{n\times 1}$ $\beta_{p\times 1}$

$X_{n\times p'}$ $\varepsilon_{n\times 1}$

و مدل رگرسیون خطی قبلی $Y = X\beta + \varepsilon$ بطوریکه $\varepsilon \sim N_n(0, \sigma_I^2)$

اما مدل رگرسیون مختلط خطی تعمیم یافته، یک مدل رگرسیون خطی است که اجازه می دهد پارامترهای «ضرایب» رگرسیون تصادفی وهمبستگی بین متغیرهای پاسخ داشته باشیم که بصورت زیر تعریف می گردد:

$$Y = X\beta + Z\gamma + \varepsilon \quad «۲.۱»$$

که در آن y وx وβ مانند آنچه دررگرسیون خطی چند گانه دیدیم می باشند. و Z یک ماتریس معلوم و γ یک بردار از متغیرهای تصادفی است که اثرات تصادفی نامیده می شود که قابل مشاهده «اندازه گیری » نیستند.

فرضهای مدل بصورت زیر مطرح می گردد:

$I)\ E(\gamma) = 0$

$II)\ E(\varepsilon) = 0$

$III)\ V(\varepsilon) = V$

$IV)\ V(\gamma) = D$

$V)\ Var\begin{pmatrix}\varepsilon\\ \gamma\end{pmatrix} = \begin{pmatrix}V & 0\\ 0 & D\end{pmatrix}$

ازمدل فوق بافرضهای داده شده خواهیم داشت:

$$E(Y) = X\beta\ ,\quad Var(Y) = ZDZ' + V \quad \text{«۲,۲»}$$

ملاحظه می شود که اضافه کردن اثرات تصادفی داخل مدل خطی مختلط اثری روی میانگین y ندارد. بعلاوه مدل فوق «۲,۲» به ما اجازه می دهد که همبستگی بین متغیرهای پاسخ را مدل بندی کنیم به اشکال زیر:

مستقیماً میتوان داخل تمرین v یا با تعیین ساختار اثرات تصادفی «Dو Z».

درحالت کلی ساختار داده ها و نیازهای تغییرپذیری باید ما را درانتخاب مان راهنمائی کنند. در بخش 2.4 توضیحات بیشتری را خواهیم داد.

مثال ۱ . ۲ .۲- «۱ .۲»

در مثال قبل ۶ موش که از ۳ خانواده بدست آمده بودند. فرض می کنیم که همبستگی بین اعضای یک خانواده وجود دارد.

اضافه وزن jامین موش ازخانواده iام = y_{ij}

مقداراضافی پروتئین موش jام ازخانواده iام = x_{ij}

فرض می کنیم که اثرات مقداراضافی پروتئین برای تمام موشها یکسان هستند.

یعنی

$$E(y_{ij}) = \beta_0 + \beta_1 x_{ij} \quad \text{و} \quad V(\varepsilon_{ij}) = \sigma^2 + \sigma_1^2$$

و

$$cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_1^2 \ ; \ j \neq j'$$

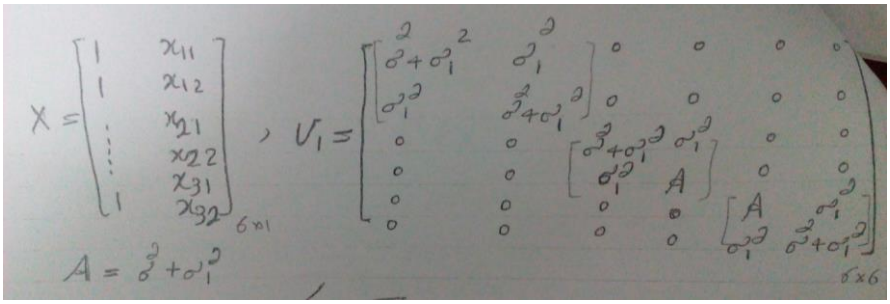
و

$$cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \ ; \ i \neq i'$$

و عناصر این مثال را به اشکال مختلف میتوان بیان کرد:

$$y' = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32})$$

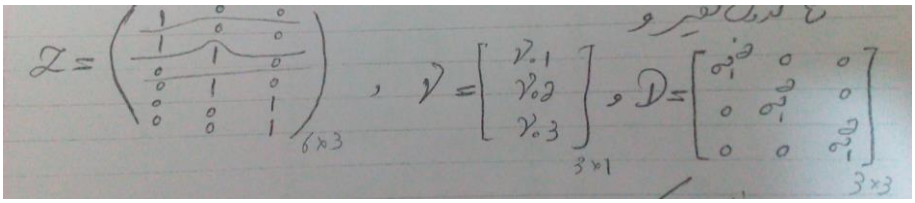
$$\beta' = (\beta_0, \beta_1) \ , \quad \varepsilon' = (\varepsilon_{11}, \varepsilon_{12}, \varepsilon_{21}, \varepsilon_{22}, \varepsilon_{31}, \varepsilon_{32})$$



مدل داده شده فوق شرایط مسئله «۱ .۲» را برآورده می کند. بعلاوه میتوان اثرات خانواده راکه تصادفی اند و غیر قابل اندازه گیری به مدل بصورت زیر اضافه کرد.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0j} + \varepsilon_{ij}$$

با $v_2 = \sigma^2 I$ و ماتریسهای Y و X و β و ε بدون تغییر و



براحتی ملاحظه می شود که

$$E\ y_{ij} = \beta_0 + \beta_1 x_{ij}$$

دقیقاً همان ماتریس فوق است:

$$V(y_{ij}) = ZDZ^{'} + v_2 = v_1$$

اگر مدل را اینطور بنویسیم:

$$y_{ij} = (\beta_0 + \gamma_{0i}) + \beta_1 x_{ij} + \varepsilon_{ij}$$

که درآن $\beta_0 + \gamma_{0i}$ یک متغیر تصادفی یا میانگین β_0 و شامل اثر خانواده می باشد.

مثال۲,۲-

۴ درخت ازنوع A

۴ درخت ازنوع B

$$x_i = \begin{cases} 1\ ; & \text{درخت } \in A \\ 0\ ; & o.w \end{cases}$$

t_1 , t_2 ,t_3 زمان های اندازه گیری است.

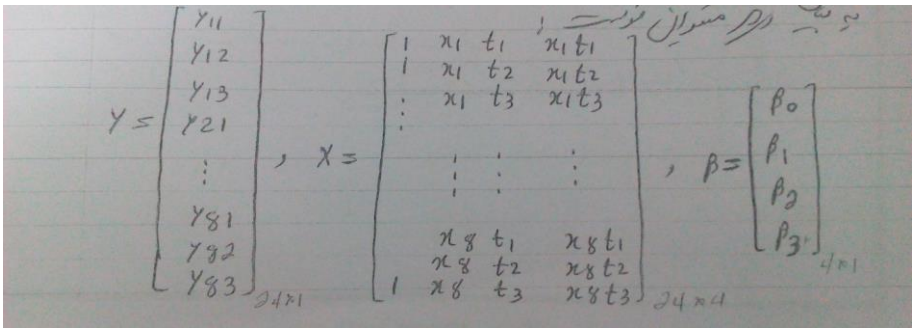
اگر فرض کنید که قد درختان درزمان کاشت متفاوت اند ونرخ رشددرختان نیز متفاوت اند در نتیجه داریم:

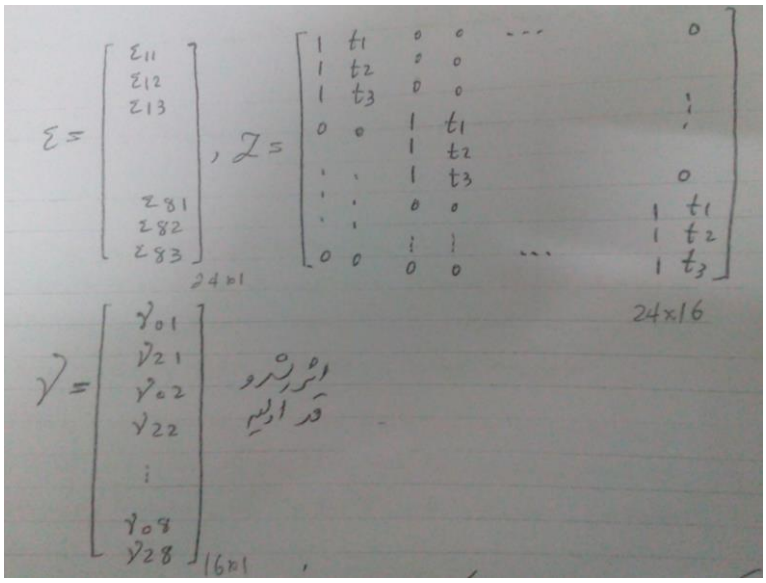
$$y_{ij} = (\beta_0 + \gamma_{0i}) + \beta_1 x_i + (\beta_2 + \gamma_{2i}) t_j + \beta_3 x_i t_j + \varepsilon_{ij} \qquad (2.3)$$

توجه:قد درختان درزمان o متغیری تصادفی بامیانگینβ_0 برای نوع B و برای نوع A ، $\beta_0 + \beta_1$است.

نرخ رشد قدی درختان نیز متغیر تضادفی با میانگین β_2 برای نوع B و $\beta_2 + \beta_3$ برای نوع A.

به بیان دیگر می توان نوشت:





ازآنجاییکه اثرات تصادفی دارای همبستگی هستند برای انتخاب ماتریس های D.V چنین عمل می کنیم.

ماتریس V را می توان $v = \sigma^2 I$ در نظر گرفت ولی انتخاب ساختار ماتریس D بسیارمشکل خواهد بود.

میتوان فرض کردکه γ_{0i} و γ_{2i} دارای واریانس متفاوت یا مساوی هستند و نیز می توان فرض کرد که

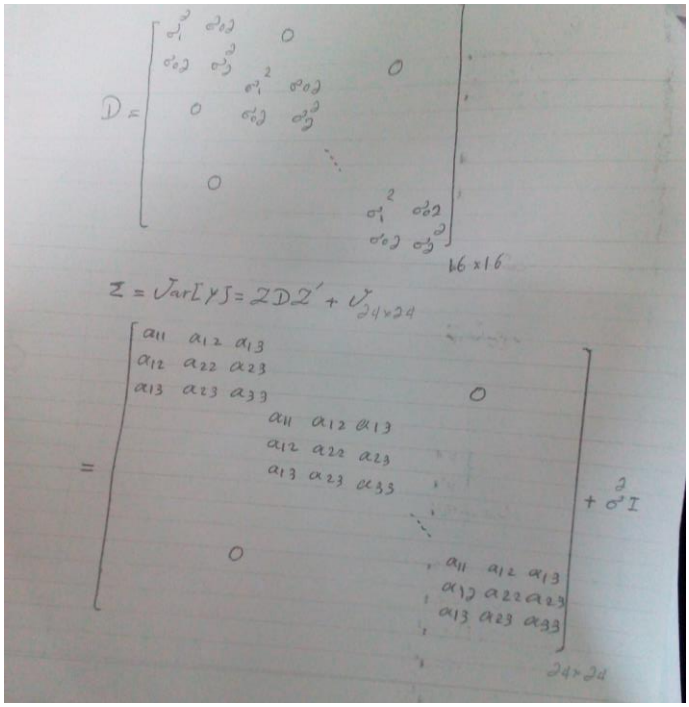
$cov(\gamma_{2i}, \gamma_{0i})$ صفر یامخالف صفر هستند.

حال حالت کلی زیر را می توان در نظر گرفت یعنی

$$V(\gamma_{0i}) = \sigma_1^2$$

$$V(\gamma_{2i}) = \sigma_2^2 \quad \text{و} \quad cov(\gamma_{2i}, \gamma_{0i}) = \sigma_{02}$$

که دراین حالت ماتریس Dخواهد بود:



که در آن $a_{kl} = \sigma_1^2 + \sigma_{02}(t_k + t_l) + \sigma_2^2 t_k t_l$

که این مدل معادل $y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 t_j + \beta_3 x_i t_j + \varepsilon_{ij}$

$$V(y_{ij}) = \sigma^2 + a_{jj} \quad , \quad Cov(y_{ij}, y_{ik}) = a_{jk}$$

به بیان دیگر واریانس قد درختان با تغییر زمان تغییر می کند و اندازه های قد یک دارای همبستگی (خود) همبستگی می باشد.

مثال ۲.۳- در مثال درجه حرارت آب و ph آن اگر فرض کنیم که:

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$$

میتوان یک ساختار همبستگی بین مقادیر اندازه گیری شده توسط یک دستگاه در نظر گرفت یا میتوان این طور فرض کرد.

$$y_{ij} = \beta_0 + \beta_1 x_i + \gamma_{0i} + \varepsilon_{ij}$$

که درآن ε_{ij} ناهمبسته اند و متغیرتصادفی γ_{0i} بیانگر یک همبستگی بین اندازه های گرفته شده توسط یک دستگاه می باشد.

۱ .۲ .۲- فرمهای ممکن ماتریس واریانس:

دوعنصر مهم درمدل مختلط وجود دارد که عبارتنداز ماتریس های D,V

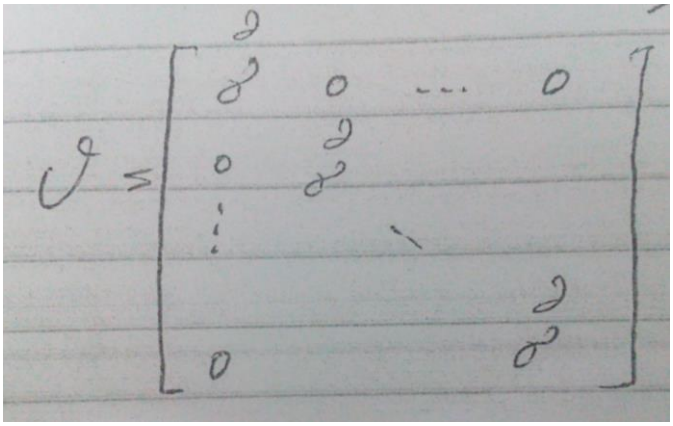
درکل این ماتریس ها بصورت بلوکی–قطری هستند. که هربلوک متناظر همبستگی های مشاهدات آن گروه بین اعضاء آن گروه می باشند.

بعنوان مثال: در نمونه موشها هرخانواده یک بلوک ،در مثال درختان نوع A یا B یک بلوک و در مثال phآب هر دستگاه یک بلوک می باشد.

هرگاه یک مدل مختلط را درنظر می گیریم باید فرم بلوک ماتریس را نیز مشخص کنیم؟

که چندین مدل ممکن برای ماتریسها بصورت زیر درنظر گرفته می شود.

مولفه های واریانس: هرگاه مشاهدات هم واریانس و همبستگی وجود نداشته باشد:



ماتریس های V_2 و D در مثال ۱ .۲ نیز به همین فرم هستند.

این فرم برای ماتریس بلوک باقیمانده های V وماتریس اثرات تصادفی بسیار بکار می رود.

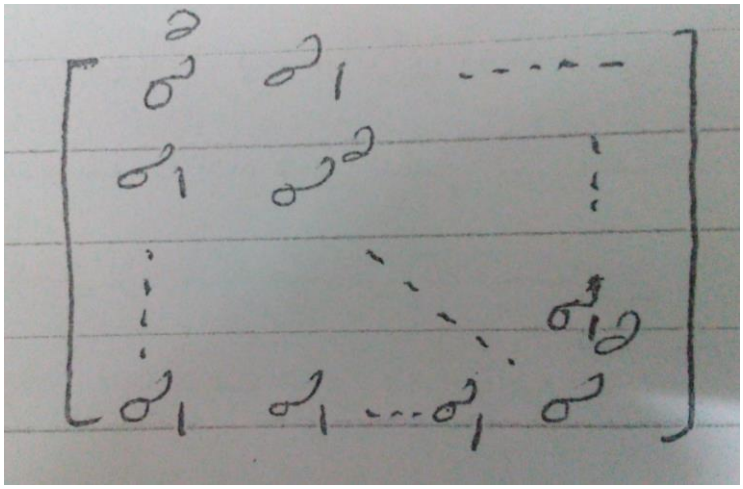
درعمل به ندرت ضرایب .... , β_0, β_1 هم واحد هستند.

اما میتوان فرض کرد که تغییرات عرض ازمبدأ و ضریب زاویه هم واریانس باشند. مثل: ترکیب متقارن

(Compound Symmetry)

دراین حالت فرض می شود که تمام عناصر هم واریانس و کواریانسها مساوی هستند که فرم ماتریس V_1 در

مثال۱ .۲ میباشند.

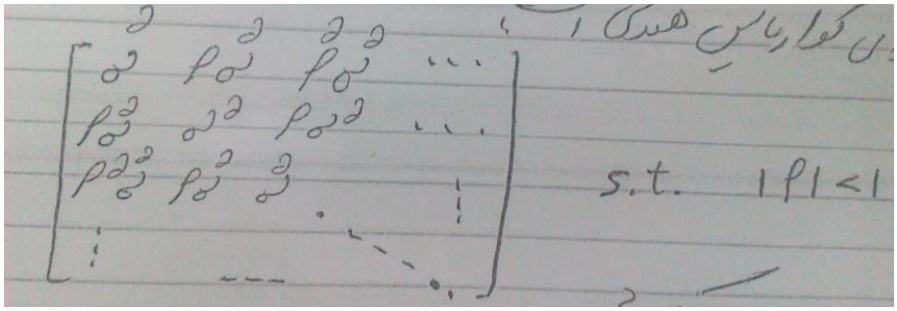


که این فرم دربسیاری از حالتهای این فصل اتفاق می افتد.

یادآوری می کنیم که هرگاه yها هم واریانس و yها هم کواریانس باشند این ساختار برای ماتریس بلوک V خیلی مناسبتر است تا ماتریس بلوک D.

اتورگرسیو مرتبه ۱: «۱»AR

هم واریانس و دارای کواریانس هندسی است.



برای اثبات این ازقضیه B56کتاب Christensen2002استفاده شده است.

و ادامه دستنوشته ها ....