STA302 Midterm Review

The OLS estimates of β_0 and β_1 in simple regression are the values that minimize the residual sum of squares function.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

set derivatives to 0, and solve:

$$\begin{cases} \dfrac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \dfrac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \beta_0 n + \beta_1 \sum x_i = \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Since $SXX = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$

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

Take back get $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \dfrac{SXY}{SXX} = r_{xy}\dfrac{SD_y}{SD_x} = r_{xy}\left(\dfrac{SYY}{SXX}\right)^{\frac{1}{2}}$

Def of Symbols:

Quantity	Definition	Description
$\bar{x}$	$\sum x_i / n$	Sample average of x
$\bar{y}$	$\sum y_i / n$	______ of y
SXX	$\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) x_i$	Sum of squares for the x's
SD_x^2	$SXX/(n-1)$	Sample variance of the x's.
SD_x	$\sqrt{SXX(n-1)}$	Sample standard deviation
SD_y^2	$SYY/(n-1)$	Sample variance of the y's.
SD_y	$\sqrt{SYY/(n-1)}$	Sample SD of the y's.
SYY	$\sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y}) y_i$	Sum of squares of the y's.
SXY	$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i$	Sum of cross-products.
S_{xy}	$SXY/(n-1)$	Sample covariance
r_{xy}	$S_{xy}/(SD_x SD_y)$	Sample correlation

estimating variance:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \quad df = n-2 \quad \text{, (as 2 parameters)}$$

$\swarrow$

residual mean square.

$$RSS = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX$$

$\hat{\sigma}$ is called the standard error of regression.

$$\boxed{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)}$$

Means and variances of least square estimates

$\hat{\beta}_1 = \sum c_i y_i$ for $\underbrace{c_i = \frac{(x_i - \bar{x})}{SXX}}_{\text{fixed}}$ by (last page)

$$E(\hat{\beta}_1 | X) = E(\sum c_i y_i | X = x_i) = \sum c_i E(y_i | X = x_i)$$

$$= \sum c_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \underbrace{\sum c_i}_{0} + \beta_1 \underbrace{\sum c_i x_i}_{1}$$

$$= \beta_1$$

unbiased, similarly $\hat{\beta}_0 = \beta_0$

$$Var(\hat{\beta}_1 | X) = Var(\sum c_i y_i | X = x_i) = \sum c_i^2 Var(Y | X = x_i)$$

$$= \sigma^2 \sum c_i^2$$

$$= \frac{\sigma^2}{SXX}$$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x} | X)$$

$$= Var(\bar{y} | X) + \bar{x}^2 Var(\hat{\beta}_1 | X) - 2\bar{x} Cov(y, \hat{\beta}_1 | X)$$

Since $Cov(\bar{y}, \hat{\beta}_1 | X) = Cov(\frac{1}{n} \sum y_i, \sum c_i y_i)$

$$= \frac{1}{n} \sum c_i (y_i, y_i)$$

$$= 0$$

So $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$

Finally, $Cov(\hat{\beta}_0, \hat{\beta}_1 | X) = Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 | X)$

$$= Cov(\bar{y}, \hat{\beta}_1) - \bar{x} Cov(\hat{\beta}_1, \hat{\beta}_1)$$

$$= 0 - \sigma^2 \frac{\bar{x}}{SXX} = -\sigma^2 \frac{\bar{x}}{SXX}$$

Estimating σ^2

- $\hat{\sigma}^2$ is the average size of $\hat{e}_i^2 = (y_i - \hat{y}_i)^2$
- $\hat{\sigma}^2$ can be obtained by dividing $RSS = \sum \hat{e}_i^2$ by its degrees of freedom (df)

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- $\hat{\sigma}^2$ is called "standard error of regression".
- if e_i are iid from $N(0, \sigma^2)$ then $\frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$

- RSS can be calculated by its definition.

$$RSS = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_i [y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})]^2$$
$$= SYY + \hat{\beta}_1^2 SXX - 2\hat{\beta}_1 SXY$$
$$= SYY + \frac{SXY^2}{SXX^2} SXX - 2\frac{SXY^2}{SXX}$$
$$= SYY - \frac{SXY^2}{SXX} = \cancel{SX} SYY - \hat{\beta}_1^2 SXX$$

Properties of Least Squares Estimates

- $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written in a linear combination of y_i's.
- let $c_i = \frac{x_i - \bar{x}}{SXX}$ (free of y_i's), note that $\sum_i \cancel{(x_i-\bar{y})} (x_i - \bar{x})\bar{y} = 0$

$$\hat{\beta}_1 = \sum_i \frac{(x_i - \bar{x})}{SXX} y_i = \sum_i c_i y_i$$

- the fitted line passes through $(\bar{x}, \bar{y})$
- estimators are unbiased. denote $X = \{x_1, \cdots, x_n\}$

  ($\hat{\theta}$ is unbiased for θ if $E(\hat{\theta}) = \theta$)

$$E(\hat{\beta}_0 | X) = \beta_0, \quad E(\hat{\beta}_1 | X) = \beta_1, \quad E(\hat{\sigma}^2 | X) = \sigma^2$$

WHY?

- recall $\hat{\beta}_1 = \frac{SXY}{SXX} = \sum_i \frac{x_i - \bar{x}}{SXX} y_i = \sum_i c_i y_i$

$$E(\hat{\beta}_1 | X) = E(\sum c_i y_i | X = x_i) = \sum_i c_i E(y_i | X = x_i)$$
$$= \sum_i c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum C_i + \beta_1 \sum_i C_i x_i$$

- $\sum_i C_i = \sum_i (x_i - \bar{x}) = 0$, $\sum_i C_i x_i = \frac{\sum_i (x_i - \bar{x}) x_i}{SXX} = 1$

$$E(\hat{\beta}_1 | X) = \beta_1$$

- Since $\cancel{E(\bar{y}|\bar{x})}$ $E(\bar{y} | X) = \beta_0 + \beta_1 \bar{x}$ we have

$$E(\hat{\beta}_0 | X) = E(\bar{y} | X) - \beta_1 \bar{x} = \beta_0$$

Variances of LSE

· variances of the estimates

$$Var(\hat{\beta}_1|X) = \frac{\sigma^2}{SXX}$$

$$Var(\hat{\beta}_0|X) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1|X) = -\sigma^2\frac{\bar{x}}{SXX}$$

$$\rho(\hat{\beta}_0, \hat{\beta}_1|X) = \frac{-\bar{x}}{\sqrt{SXX/n + \bar{x}^2}} = \frac{-\bar{x}}{\sqrt{(n-1)SD_x^2/n + \bar{x}^2}}$$

- in all expressions, σ^2 are unknown
- to estimate $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$, replace σ^2 by $\hat{\sigma}^2$

$$\widehat{Var}(\hat{\beta}_1|X) = \hat{\sigma}^2\frac{1}{SXX}$$

$$\widehat{Var}(\hat{\beta}_0|X) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$$

- the square root of an estimated variance is called a standard error (se):

$$se(\hat{\beta}_1|X) = \sqrt{\widehat{Var}(\hat{\beta}_1|X)} \qquad se(\hat{\beta}_0|X) = \sqrt{\widehat{Var}(\hat{\beta}_0|X)}$$

Deriving Variances of LSE

- recall y_i's are assumed independent given x_i's.

$$Var(\hat{\beta}_1|X) = Var(\sum_i c_i y_i|X) = \sum_i c_i^2 Var(y_i|X=x_i)$$

$$= \sigma^2\sum_i c_i^2 = \sigma^2\sum_i (x_i - \bar{x})^2\frac{1}{SXX^2}$$

$$= \sigma^2\frac{1}{SXX}$$

$$Var(\hat{\beta}_0|X) = Var(\bar{y} - \hat{\beta}_1\bar{x}|X)$$

$$= Var(\bar{y}|X) + \bar{x}^2 Var(\hat{\beta}_1|X) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1|X)$$

Since $Cov(\bar{y}, \hat{\beta}_1|X) = Cov(\frac{1}{n}\sum_i y_i, \sum c_i y_i|X)$

$$= \frac{1}{n}\sum c_i Cov(y_i, y_i|X)$$

$$= \frac{\sigma^2}{n}\sum c_i$$

$$= \frac{\sigma^2}{n}\sum(x_i - \bar{x}) = 0$$

And $Var(\bar{y}|X) = \frac{\sigma^2}{n}$, so $Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{SXX} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$

Deriving Cov. of LSE

- now for cov. between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$
\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1 | X) &= Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 | X) \\
&= Cov(\bar{y}, \hat{\beta}_1 | X) - \bar{x}\, Cov(\hat{\beta}_1, \hat{\beta}_1 | X) \\
&= 0 - \sigma^2 \frac{\bar{x}}{SXX} \\
&= -\sigma^2 \frac{\bar{x}}{SXX}
\end{aligned}
$$

- $$\rho(\hat{\beta}_0, \hat{\beta}_1 | X) = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1 | X)}{\sqrt{Var(\hat{\beta}_0 | X) Var(\hat{\beta}_1 | X)}}$$

Lecture 2 : part ii

Comparing Models

- Analysis of Variance (ANOVA)
- a simple example : comparing 2 reg models:

$E(Y|X=x) = \beta_0$ v.s. (slope 0, cannot predict Y given X. or X, Y not related)

$E(Y|X=x) = \beta_0 + \beta_1 x$

β_0 can be estimated by minimizing $\Sigma(y_i - \beta_0)^2$, by OLS with only 1 parameter

thus $\hat{\beta}_0 = \bar{y}$

- residual sum of squares is

$$\Sigma(y_i - \hat{\beta}_0)^2 = \Sigma(y_i - \bar{y})^2 = SYY \text{ with } n-1 \text{ df.}$$

Which one to use?

- call $\widehat{E(Y|X)} = \hat{\beta}_0$ fitted model 1

$\widehat{E(Y|X)} = \hat{\beta}_0 + \hat{\beta}_1 x$ fitted model 2

- compare RSS's

$RSS_1 = SYY$

$RSS_2 = SYY - \frac{(SXY)^2}{SXX}$

$RSS_1 \geq RSS_2$

- idea: if adding the slope β_1 does not help much, then RSS_2 should not be much smaller than RSS_1.

- How small is small?
- "sum of squares due to regression" (SSreg):

$$SSreg = RSS_1 - RSS_2$$
$$= SYY - (SYY - \frac{SXY^2}{SXX})$$
$$= \frac{SXY^2}{SXX}$$

df for SSreg = df for RSS_1 − df for RSS_2
$= (n-1) - (n-2) = 1$

ANOVA

- compare "standardized version of SSreg" v.s. "standardized version of RSS_2".

sum of sqs → SS; $\frac{SS}{df}$ (mean sqr. error) → MSE

-

Source	df	SS	MSE	F	p-value
Regression	1	SSreg	SSreg/1	$MSreg/\hat{\sigma}^2$	
Residual	n−2	RSS	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	n−1	SYY			

- If slope helps, $RSS_2 \ll RSS_1 \Rightarrow SSreg = RSS_1 - RSS_2$ large enough.

F-test for Regression

$$\Rightarrow F = \frac{SSreg/1}{RSS/n-2} \text{ will be large}$$

F ≠ chi-square since → is not true variance, it's just est of var.

- F is a rescaled version of ~~#~~ $SSreg = RSS_1 - RSS_2$

key assumption for F-test: e_i are iid $N(0, \sigma^2)$ then

$$\frac{SSreg}{\sigma^2} \sim \chi^2_1 \text{ (if } \beta_1 = 0), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}, \quad SSreg \perp RSS$$

- recall F-distribution: $F \sim F(1, n-2)$, given $\beta_1 = 0$.
- what we are doing is a stat test:

NH: $E(Y|X=x) = \beta_0$ v.s. AH: $E(Y|X=x) = \beta_0 + \beta_1 x$

- compare "the observed value of F" calculated from the sample to the critical value $F_{(\alpha, 1, n-2)}$, the upper-α quantile or $100(1-\alpha)$th percentile of $F_{(1,n-2)}$

- if $F_{obs} > F_{(\alpha,1,n-2)}$, reject NH, use AH
- if $F_{obs} \leq F_{(\alpha,1,n-2)}$, don't reject NH

p-value and interpretation.

- Assuming NH is true, the prob. that the test statistic is at least as extreme as was observed in the sample. e.g. in the previous F-test, p-value $= P(F \geq F_{obs} | \beta_1 = 0) \approx 0$.
- a measure of the strength of the evidence against NH in favor of AH, not the probability that NH is true.

- compare p-value with significance level α, say $\alpha = 0.05$

(NH is true) then in 5% of tests where NH is true we will get a p-value smaller than or equal to 0.05.

(NH is false) expect to see small p-values more often. ~~The~~

Coefficient of Determination, R^2

- $R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY}$ tells you how useful your slope is.

  say $R^2 = 0.995$, 99.5% of the variability ~~of~~ in the data ...

  SSreg: variability explained by the slope

  SYY: variability in the data.

- $R^2 = \frac{SSreg}{SYY} = \frac{(SXY)^2}{SXX\,SYY} = r_{xy}^2$ (square of sample correlation)

Confidence intervals and tests.

- for "simple problems", if $\hat{\theta}$ is an estimate for θ, then a $100(1-\alpha)\%$ confidence interval (CI) for θ is $(\hat{\theta} - t_{(\frac{\alpha}{2},d)} se(\hat{\theta}), \hat{\theta} + t_{(\frac{\alpha}{2},d)} se(\hat{\theta}))$ where $se(\hat{\theta})$ is standard error for θ, $t_{(\frac{\alpha}{2},d)}$ is the value that cuts off $\frac{\alpha}{2} \cdot 100\%$ in the upper tail of the t-distribution with $df = d$.
- key assumption: e_i's are iid $N(0, \sigma^2)$

For β_0 (intercept)

C.I. $(\hat{\beta}_0 - t_{(\frac{\alpha}{2}, n-2)} se(\hat{\beta}_0), \hat{\beta}_0 + t_{(\frac{\alpha}{2}, n-2)} se(\hat{\beta}_0))$

where $se(\hat{\beta}_0) = \hat{\sigma} \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)^{\frac{1}{2}}$

- Hypothesis test : NH $:\beta_0 = \beta_0^*$, β_1 arbitrary
  AH: $\beta_0 \neq \beta_0^*$, β_1 arbitrary

- t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$ and compare to $t_{(\frac{\alpha}{2}, n-2)}$

For β_1 (slope)

C.I. : $\hat{\beta}_1 \pm t_{(\frac{\alpha}{2}, n-2)} se(\hat{\beta}_1)$
$= \hat{\beta}_1 \pm t_{(\frac{\alpha}{2}, n-2)} \frac{\hat{\sigma}}{\sqrt{SXX}}$

- doing the t-test NH: $\beta_1 = 0$ vs. AH : $\beta_1 \neq 0$ is the same as comparing $y = \beta_0$ and $y = \beta_0 + \beta_1 x$ with an F-test.

- t-statistic : $t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SXX}}$

- $t^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/SXX} = \frac{\hat{\beta}_1^2 SXX}{\hat{\sigma}^2}$ = F-statistic from ANOVA

- a one-to-one correspondence here.
- from the fact that the square of t_d is $F_{(1,d)}$
- F --- global test , t --- for single parameter.

Prediction and Fitted values

- prediction: predict the value of y given a new value of x.
- denote new values : x_*, y_*.
- $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$
- $Var(\tilde{y}_* | x_*) = \sigma^2 + \sigma^2 (\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})$
- ~~se~~ $sepred(\tilde{y}_* | \tilde{x}_*) = \hat{\sigma}^2 (1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})^{\frac{1}{2}}$
- construct a prediction interval for y_*: $\tilde{y}_* \pm t_{(\frac{\alpha}{2}, n-2)} sepred(\tilde{y}_* | X, x_*)$
- fitted values (estimation, but not prediction)
- estimated by fitted values.
  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  standard error is $sefit(\hat{y} | X, x) = \hat{\sigma} (\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX})^{\frac{1}{2}}$

  CI. $(\hat{\beta}_0 + \hat{\beta}_1 x) \pm sefit(\hat{y} | X, x) [2F(\alpha, 2, n-2)]$

Residuals

- $\hat{e_i} = y_i - \hat{y_i}$
- useful plot: residuals v.s. fitted values

Lecture 3. Multiple Linear Regression.

e.g. $E(Y|X_1=x_1, X_2=x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$Var(Y|X_1=x_1, X_2=x_2) = \sigma^2$

- adding a term good or not?

- e.g. UN.txt

Fertility, PPgdp, Purban, locality ...

Y: log(Fertility) X_1: log(PPgdp), X_2: ~~log~~ Purban

$R_1^2 = 46\%$ for $\widehat{E(Y|X_1)} = 2.703 - 0.153x_1$

$R_2^2 = 35\%$ for $\widehat{E(Y|X_2)} = 1.750 - 0.013x_2$

Now $46\% \le R^2 \le 46\% + 35\%$

- how much additional explanation offered by X_2?
- let $\hat{e}_{Y|X_1}$ be the residuals of regressing Y on X_1: the part of variability of Y not explained by X_1, or variability of Y after the effect of X_1 is removed
- similarly for $\hat{e}_{Y|X_2}$

Multiple Linear Regression (MLR)

- in general, multiple linear model:

$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

$Var(Y|X) = \sigma^2$

  - a linear function of parameters $\{\beta_0, \cdots, \beta_p\}$

Terms and Predictors

- predictors: the original data that you collect.
- terms: created from predictors, the X-variable in MLR.

Matrix notation for MLR

- observed values $(x_{11}, x_{12}, \cdots, x_{1p}, y_1)$
  $(x_{21}, x_{22}, \cdots, x_{2p}, y_2)$
  $\vdots$
  $(x_{n1}, x_{n2}, \cdots, x_{np}, y_n)$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- ith row of $X \to \vec{x}_i'$

$$\vec{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \vec{x_i} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \vec{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

- there are $(p+1)$ parameters, including the intercept β_0
- multiple linear regression in matrix notation.

$$\vec{Y} = \vec{X}\vec{\beta} + \vec{e}$$

- ith row is $y_i = \vec{x}_i'\vec{\beta} + e_i$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

- about the vector of errors e:

$E(\vec{e}) = \vec{0} \qquad Var(\vec{e}) = \sigma^2 \vec{I}_n$

- add normality assumption: $\vec{e} \sim N(\vec{0}, \sigma^2 \vec{I}_n)$

OLS for MLR

- $RSS(\vec{\beta}) = \sum_i^n (y_i - \vec{x}_i'\vec{\beta})^2 = (\vec{Y} - \vec{X}\vec{\beta})'(\vec{Y} - \vec{X}\vec{\beta})$
  - if $(\vec{X}'\vec{X})^{-1}$ exists, RSS is minimized by

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

  - $\vec{X}'\vec{X}$, $\vec{X}'\vec{Y}$ similar to SXX, SXY
  - Residual $\hat{e} = Y - \hat{Y}$
  - $RSS = \hat{e}'\hat{e} = (Y - X\hat{\beta})'(Y - X\hat{\beta})$
  - $\sigma^2 = Var(Y|X)$ is estimated with $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$
  - with normality assumption: $(n-p+1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$

OLS using ~~matrix~~ matrices

- model : $E(Y|X=x) = \beta' x$ and $Var(Y|X=x) = \sigma^2$

Properties of OLS Estimates

- assume $E(e)=0$, $Var(e) = \sigma^2 I_n$, $\hat{\beta}$ is unbiased

$$E(\hat{\beta}|X) = E((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'E(Y|X)$$
$$= (X'X)^{-1}X'X\beta$$
$$= \beta$$

- for variance, need $Var(B'Z) = B'Var(Z)B$

$$Var(\hat{\beta}|X) = Var((X'X)^{-1}X'Y|X) = (X'X)^{-1}X'[Var(Y|X)]X(X'X)^{-1}$$
$$= (X'X)^{-1}X'[\sigma^2 I_n]X(X'X)^{-1} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

Residuals sum of squares

- $RSS = RSS(\hat\beta) = (Y - X\hat\beta)'(Y - X\hat\beta) = Y'Y + \hat\beta'(X'X)\hat\beta - 2Y'X\hat\beta$
- $\hat\beta'(X'X)\hat\beta = \hat\beta'(X'X)(X'X)^{-1}X'Y = \hat\beta'X'Y = Y'X\hat\beta$
- $RSS = Y'Y - \hat\beta'X'X\hat\beta = Y'Y - \hat Y'\hat Y$, with $\hat Y = X\hat\beta$

ANOVA

- Comparing $E(Y|X=x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ with $E(Y|X=x) = \beta_0$
- similar to SLR

Source	df	SS	MS	F	p-value
Regression	p	SSreg	SSreg/p	$MSreg/\hat\sigma^2$	
Residual	$n-(p+1)$	RSS	$\hat\sigma^2 = \frac{RSS}{n-(p+1)}$		
Total	$n-1$	SYY			

- coefficient of determination

$$R^2 = \frac{SSreg}{SYY} = \frac{SYY - RSS}{SYY} = 1 - \frac{RSS}{SYY}$$

Sequential analysis of variance tables

- order matters.

Predictions & Fitted values

- similar to SLR
- prediction: given a new x_*, predict y_* with $\hat{y_*} = x_*'\hat\beta$

$$\text{sepred}(\hat y_* | x_*) = \hat\sigma\sqrt{1 + x_*'(X'X)^{-1}x_*}$$

- fitted value: given a value x, want to estimate the mean function at x.
- $\hat E(Y|X=x) = \hat y = x'\hat\beta$
- $\text{sefit}(\hat y|x) = \hat\sigma\sqrt{x'(X'X)^{-1}x}$
- so $\text{sepred}(\hat y_*|x_*) = \sqrt{\hat\sigma^2 + \text{sefit}(y_*|x_*)^2}$

Lecture 4 Drawing Conclusions

...

Sampling from Normal population

- data: $(x_1,y_1),\cdots,(x_n,y_n)$
- $\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$
- conditional distribution of y_i given x_i?
  - $y_i|x_i \sim N(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i-\mu_x), \sigma_y^2(1-\rho_{xy}^2))$
  - define $\beta_0 = \mu_y - \beta_1\mu_x$, $\beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$, $\sigma^2 = \sigma_y^2(1-\rho_{xy}^2)$
  - $y_i|x_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$
    - $\hat{\mu}_x = \bar{x}$, $\hat{\mu}_y = \bar{y}$, $\hat{\sigma}_x^2 = \frac{SXX}{n-1}$, $\hat{\sigma}_y^2 = \frac{SYY}{n-1}$, $\hat{\rho}_{xy} = \frac{SXY}{\sqrt{SXX \cdot SYY}}$

      $\hat{\rho}_{xy} = r_{xy}$
    - $\hat{\beta}_0 = \bar{y} - r_{xy}\frac{\cancel{SX}SYY}{SXX}\bar{x}$

      $\hat{\beta}_1 = r_{xy}\frac{SYY}{SXX} = \frac{SXY}{SXX}$

Missing Data?

unrecorded ..
- missing at random (MAR): prob. of missing ~~data~~ does not depend on its value.
  - strategies: deleting & guessing

    (better): imputation - need stat model

Computationally ~~intuitive~~ intensive methods

- $y_1,\cdots,y_n \sim$ dist G, construct a 95% CI for median
- I. G is known

  4 steps:
  1. obtain a sample $y_1^*,\cdots,y_n^*$
  2. compute median, store it.
  3. repeat 1, 2 --- times
  4. say 1000 times, then 95% CI is (25th smallest, 25th largest)

- II G unknown.

Lec 5 Weights, lack of fit and more.

WLS and LOF

Weighted Least squares (WLS)

- change $Var(Y|X) = \sigma^2$ to $Var(Y|X=x_i) = Var(e_i) = \frac{\sigma^2}{w_i}$
  where $w_1, \dots, w_n$ are known + numbers
- in matrix form $Y = X\beta + e$ $\quad Var(e) = \sigma^2 W^{-1}$
  W is a diagonal matrix with elements $w_1, \dots, w_n$
- β is the minimer of

$$RSS(\beta) = \sum w_i (y_i - x_i'\beta)^2$$
$$= (Y - X\beta)' W (Y - X\beta)$$

WLS Solution

- solution is $\hat{\beta} = (X'WX)^{-1} X'WY$
- translate WLS to an OLS problem.
- calculate

$$Var(W^{\frac{1}{2}} e) = W^{\frac{1}{2}} Var(e) W^{\frac{1}{2}}$$
$$= W^{\frac{1}{2}} (\sigma^2 W^{-1}) W^{\frac{1}{2}}$$
$$= W^{\frac{1}{2}} (\sigma^2 W^{-\frac{1}{2}} W^{-\frac{1}{2}}) W^{\frac{1}{2}}$$
$$= \sigma^2 (W^{\frac{1}{2}} W^{-\frac{1}{2}})(W^{-\frac{1}{2}} W^{\frac{1}{2}})$$
$$= \sigma^2 I$$

then $\cdot$ $\underbrace{W^{\frac{1}{2}} Y}_{Z} = \underbrace{W^{\frac{1}{2}} X}_{M}\beta + \underbrace{W^{\frac{1}{2}} e}_{d}$

$$Z = M\beta + d$$
$$\hat{\beta} = (M'M)^{-1} M'Z$$
$$= ((W^{\frac{1}{2}} X)'(W^{\frac{1}{2}} X))^{-1} (W^{\frac{1}{2}} X)'(W^{\frac{1}{2}} Y)$$
$$= (X'W^{\frac{1}{2}} W^{\frac{1}{2}} X)^{-1} (X'W^{\frac{1}{2}} W^{\frac{1}{2}} Y)$$
$$= (X'WX)^{-1} (X'WY)$$

LOF (lack of fit)

- F-test from ANOVA could only tell if the regression model (i.e. slope in SLR) helps explaining or not. it doesn't tell if the explanation is enough. "lack of fit".
- main idea: model good, $E(\hat{\sigma}^2) \approx \sigma^2$
  not good, $\hat{\sigma}^2$ will be estimating something bigger than σ^2.
  since its size will depend both on the errors and on systematic bias from fitting the wrong mean function.

• Lack OF test : 2 case $\begin{cases} \sigma^2 \text{ known} \\ \text{unknown} \end{cases}$

• σ^2 known NH(no LOF)

assuming normal error

$$X^2 = \frac{RSS}{\sigma^2} = \frac{(n-(p+1))^2 \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-(p+1))}$$

p-value is $P(X^2 \geq X^2_{obs} \mid \text{no LOF})$

• σ^2 unknown, estimate it! (model free manner)

• repeated measurements at some x_i's (replicates)

$y_{ij}, j=1,\cdots,n_i$, corresponding to x_i

Sum of Squares for Pure Error

3 replicates at x_i, calculate the sample variance of these 3 observations use it as an estimate of σ^2 (at x_i). Since $Var(y_{ij}|x_i) = \sigma^2$ is constant at all x_i's. Pool more values of x_i to get a better version of estimate of σ^2.

• SSpe : sum of squares for pure error.

• similar to "pooled sample variance".

the ~~pooled~~ pure error estimate of σ^2 is

$$\hat{\sigma}^2_{pe} = \frac{SSpe \rightarrow \text{add up}}{dfpe \rightarrow \text{add up}}$$

• $RSS = SS_{lof} + SSpe$

↳ sum of sqs due to LOF ($\bar{y}_i \Rightarrow \beta_0 + \beta_1 x_i$)

↳ sum of squares due to pure error ($y_{ij} \Rightarrow \bar{y}_i$)

• SS_{pe} is a saturated model.

Decomposition, $RSS = SS_{pe} + SS_{lof}$

$$RSS_{ols} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_i\sum_j (y_{ij} - \bar{y}_i + \bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum_i\sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$+ 2\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)\right](\bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum\sum(y_{ij} - \bar{y}_i)^2 + \sum n_i(\bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= SS_{pe} + SS_{lof}$$

$$= SS_{pe} + RSS_{wls}$$

$$F \text{ value} = \frac{SS_{lof}/df_{lof}}{SS_{pe}/df_{pe}}$$

General F-testing

- $NH: Y = X_1\beta_1 + e$

  $AH: Y = X_1\beta_1 + X_2\beta_2 + e$
- NH is a subset of AH (i.e. let some para in AH be 0)
- $F = \dfrac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}}$
- compare to critical value $F(\alpha, df_{NH} - df_{AH}, df_{AH})$ or p-value $P(F \geq F_{obs} | NH)$ with $F \sim F(df_{NH} - df_{AH}, df_{AH})$ under NH.

STA302 Recap.

Chapter 6: Polynomials and Factors

$E(Y|X)=\beta_0+\beta_1 X+\beta_2 X^2$ $\qquad \frac{dE(Y|X)}{dx}=0 \Rightarrow x_M=\frac{-\beta_1}{2\beta_2}$

$$E(Y|X_1=x_1,X_2=x_2)=\underbrace{\beta_0}_{\text{intercept}}+\underbrace{\beta_1x_1+\beta_2x_2}_{\text{linear}}+\underbrace{\beta_{11}x_1^2+\beta_{22}x_2^2}_{\text{quadratic}}+\underbrace{\beta_{12}x_1x_2}_{\text{interaction}}$$

$\#$ of $\quad 1 + k + k + \frac{k(k-1)}{2}$ terms

The Delta Method.

For non-linear combination, since we have $x_M=\frac{-\beta_1}{2\beta_2}$, its estimate is $\hat{x}_M=\frac{-\hat\beta_1}{2\hat\beta_2}$

- provides approximate standard errors for nonlinear combinations of parameter estimates.
- suppose $\hat{\vec\theta}\overset{\circ}{\sim}N(\vec\theta,\vec\Sigma)$ and $g(\vec\theta)$ is a function of $\vec\theta$

when n is large, we have

$$E[g(\hat{\vec\theta})]\approx g(\vec\theta)$$
$$Var[g(\hat{\vec\theta})]\approx \dot g(\vec\theta)'\Sigma\dot g(\vec\theta)$$
$$\text{where } \dot g(\vec\theta)=\frac{\partial g}{\partial\vec\theta}=\left(\frac{\partial g}{\partial\theta_1},\cdots\frac{\partial g}{\partial\theta_k}\right)'$$

- back to $\hat{x}_M$
- $\vec\beta=(\beta_0,\beta_1,\beta_2)'$, $\hat{\vec\beta}=(\hat\beta_0,\hat\beta_1,\hat\beta_2)'$, large n $\rightarrow \hat{\vec\beta}\overset{\circ}{\sim}N(\vec\beta,\sigma^2(X'X)^{-1})$

R function gives $\widehat{Cov}(\hat\beta)=\hat\sigma^2(X'X)^{-1}$

- $g(\hat{\vec\beta})=\frac{-\hat\beta_1}{2\hat\beta_2}\Rightarrow \dot g(\hat{\vec\beta})=\left(0,\frac{-1}{2\hat\beta_2},\frac{\hat\beta_1}{2\hat\beta_2^2}\right)$
- $Var(g(\hat{\vec\beta}))=\dot g(\hat{\vec\beta})'\widehat{Cov}(\hat\beta)\dot g(\hat{\vec\beta})$

$$=\frac{1}{4\hat\beta_2^2}\left(Var(\hat\beta_1)+\frac{\hat\beta_1^2}{\hat\beta_2^2}Var(\hat\beta_2)-\frac{2\hat\beta_1}{\hat\beta_2}Cov(\hat\beta_1,\hat\beta_2)\right)$$

e.g. x_1: baking time $\quad x_2$: baking temperature.

$E(Y|x_1,x_2)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^2+\beta_4x_2^2+\beta_5x_1x_2$

solve for optimal baking time (x_1)

$x_M=g(\vec\beta;x_2)=\frac{-(\beta_1+\beta_5x_2)}{2\beta_3}$ $\qquad \frac{\partial x_M}{\partial\vec\beta}=\dot g(\vec\beta;x_2)=\left(0,-\frac{1}{2\beta_3},0,\frac{\beta_1+\beta_5x_2}{2\beta_3^2},0,-\frac{x_2}{2\beta_3}\right)'$

$$Var(\hat{x}_M)=\dot g(\hat\beta;x_2)'\,\widehat{Cov}(\hat\beta)\cdot\dot g(\hat\beta;x_2)$$

$\widehat{Cov}(\hat\beta)\rightarrow\hat\sigma^2(X'X)^{-1}$

$$\hat{x}_M\pm z_{\frac{\alpha}{2}}\sqrt{\dot g(\hat\beta;x_2)'\widehat{Cov}(\hat\beta)\dot g(\hat\beta;x_2)}$$

Factors

- allow qualitative or categorical predictors in the mean function
- different levels: male or female, eye color, etc.
- dummy variables in the regression model
  - e.g. 0 for male, 1 for female or -1, 1.
- give the same outcomes if you are clear.

The factor Rule:

a factor with d levels can be represented by at most d dummy variables. If the intercept is in the mean function, at most d-1 of the dummy variables can be used in the mean function.

- define the jth dummy variable U_j, $j=1,\dots,5$

$$u_{ij} = \begin{cases} 1 & \text{if } D_i = j\text{th category of } D \\ 0 & \text{o.w.} \end{cases}$$

$\underset{\text{total}}{\text{TS}}$: sleep D: Danger

$$E(TS|D) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5 \quad (*)$$

①

Two models for the same thing.

β_j: can be interpreted as the population mean for all species with danger $\underset{\text{index}}{j}$

- no intercept since the sum of U_j is a column of ones, the intercept is implicit in (*)

$$E(TS|D) = \eta_0 + \eta_2 U_2 + \cdots + \eta_5 U_5$$

$\eta_0 = \beta_1$, $\eta_0 + \eta_2 = \beta_2$, $\eta_0 + \eta_3 = \beta_3$, $\eta_0 + \eta_4 = \beta_4$, $\eta_0 + \eta_5 = \beta_5$

② "one-way analysis of variance" model - fits a separate mean for each

- ANOVA for ① NH: all β's are zero or $E(TS|D)=0$
  - ② NH: $E(TS|D) = \eta_0$ (only intercept left there)

adding a term log(BodyWt)

then $\cdot\ E[TS|\log(\text{BodyWt}), D] = \sum_{j=1}^{5} (\beta_{0j} U_j + \beta_{1j} U_j x)$

$\cdot\ E[TS|\log(\text{BodyWt}), D] = \eta_0 + \eta_1 x + \sum_{j=2}^{5} (\eta_{0j} U_j + \eta_{1j} U_j x)$

Chapter 7 Transformations

- messy data

transform (i) predictor (ii) response or (iii) both $\Rightarrow$ so we have $E(Y|X=x) \approx \beta_0 + \beta_1 x$

Power Transformation (Both predictor & response)

- be applied to (i)(ii)(iii)
- U: original variable, strictly positive (+)

$$\psi(U, \lambda) = U^{\lambda}$$

- usual range λ: -2 to 2
- $\lambda = 1 \to$ no transformation

$\lambda = \frac{1}{2} \to \sqrt{U}$

$\lambda = -1 \to$ inverse

$\lambda = 0 \to$ taken as the log transformation

- applying log trans to both the response & predictor

$$\log(\text{BrainWt}) = \beta_0 + \beta_1 \log(\text{BodyWt}) + e$$

- $\text{BrainWt} = \beta_0 \times \text{BodyWt}^{\beta_1} \times e$

Scaled Power transformation (predictor only)

$$\psi_S(X, \lambda) = \begin{cases} (X^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log_e(X) & \text{if } \lambda = 0 \end{cases}$$

$\psi_S(X, \lambda)$ is a continuous function of λ

$\lim_{\lambda \to 0} \psi_S(X, \lambda) = \log_e(X)$

How to choose λ?

fit $(\psi_S(X, \lambda), Y)$ for different λ

note Y is not transformed, so minimize $\text{RSS}(\lambda)$

Box-Cox Transformation for Response. (for response $Y>0$)

modified power transformation

$$\psi_M(Y, \lambda_y) = \psi_S(Y, \lambda_y) \times gm(Y)^{1-\lambda_y}$$

$$= \begin{cases} gm(Y)^{1-\lambda_y} \times (Y^{\lambda_y}-1)/\lambda_y & \text{if } \lambda_y \neq 0 \\ gm(Y) \times \log(Y) & \text{if } \lambda_y = 0 \end{cases}$$

where $gm(Y) = \exp\{\frac{1}{n}\sum_{i=1}^{n} \log_e(y_i)\}$
(geometric mean of Y)

- Box-Cox method assumes
  $$E(\psi_M(Y, \lambda_y) | X=x) = \vec{\beta}'\vec{x}$$

- $gm(Y)^{1-\lambda_y}$ guarantees that the unit of $\psi_M(Y, \lambda_y)$ are the same for all values of λ_y.
- λ_y can be chosen as the one that minimizes $RSS(\lambda_y)$
- goal of B-C : not for linearity, but for normality
- try to make $\hat{e}_i$ as normal as possible.
- R function : boxcox (object, lambda = ...)

Chapter 8 : Diagnostic via Residuals.

Regression Diagnostic (model checking)

check if the linear model assumptions are satisfied or not.

- Residuals

recall: $\hat{\beta} = (\vec{X}'\vec{X})^{-1}\vec{X}'\vec{Y}$

then $\hat{\vec{Y}} = \vec{X}\hat{\beta} = \vec{X}(\vec{X}'\vec{X})^{-1}\vec{X}'\vec{Y}$

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

↓ hat matrix

transforms the data $\vec{Y}$ into fitted values $\hat{\vec{Y}}$.

residuals : $\hat{e} = \vec{Y} - \hat{\vec{Y}} = \vec{Y} - \vec{H}\vec{Y} = (\vec{I}-\vec{H})\vec{Y}$

idempotent projection matrix : $\vec{H}' = \vec{H}$, $\vec{H}\vec{H} = \vec{H}$, $\vec{H}\vec{X} = \vec{X}$

Difference between $\hat{e}$ and e

e (the statistical errors)

$E(e)=0$, $Cov(e)=\sigma^2 I$

$\Rightarrow E(\hat{e})=0$ and $Cov(\hat{e})=\sigma^2(I-H)$

* not that the var of $\hat{e_i}$'s are not the same

· let h_{ii} be the ith diagonal element H

then $Var(\hat{e_i})=\sigma^2(1-h_{ii})$ → leverage value

also $\hat{e_i}$'s are correlated

$\boxed{\text{The Hat Matrix}}$

verify : $HH = X(X'X)^{-1}X' \cdot X(X'X^{-1})X'$

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

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

$= H$

similarly, $I-H$ is also idempotent.

· direct consequences :

$(I-H)X=0 \Rightarrow E(\hat{e})=0$

$H(I-H)=0$

$Cov(\hat{e},\hat{Y})=Cov((I-H)Y,HY)=\sigma^2 H(I-H)=0$

$Cov(Y)=\sigma^2 I$, $Cov(\hat{Y})=\sigma^2 HH'=\sigma^2 H$

$Cov(\hat{e})=\sigma^2(I-H)(I-H')=\sigma^2(I-H)$

note that $Cov(\hat{e})=Cov(Y-\hat{Y})=Cov(Y)-Cov(\hat{Y})$

Diagonal of the Hat Matrix h_{ii}

$Var(\hat{e_i})=\sigma^2(1-h_{ii})$

with an intercept, $\frac{1}{n}\leq h_{ii}\leq\frac{1}{r_i}$, r_i is # of replicates of x_i

so $h_{ii}\uparrow$, $Var(\hat{e_i})\downarrow$

when $Var(\hat{e_i})=0$

then only the ith observation get $\hat{y_i}$.

H is idempotent

$$h_{ii} = h_{ij}^2 \quad \text{i.e. } h_{ii}(1-h_{ii}) = \sum_{j\neq i} h_{ij}^2$$

$$\hat{y}_i = \sum_{i=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j\neq i}^n h_{ij} y_j$$

large h_{ii} $\rightarrow$ unusual values ~~of~~ for x_i (large $h_{ii} \neq$ outliers)

- When doing WLS

- assumption : $Var(\vec{e}) = \sigma^2 W^{-1}$: W: known weights

$$H = W^{\frac{1}{2}} X (X'WX)^{-1} X' W^{\frac{1}{2}}$$

$$\hat{Y} = X\hat{\beta} = HY$$

① $\hat{e}_i = y_i - \hat{y}_i$

or

② $\hat{e}_i = \sqrt{w_i}(y_i - \hat{y}_i)$ (Pearson / weighted residuals)

→ use this

When model correct:

Let U be any of the terms.

- then $E(\hat{e}_i | U_i) = 0$, $Var(\hat{e}_i | U_i) = \sigma^2(1-h_{ii})$
- a plot of residuals has constant mean zero
- var funct. of ~~$\hat{e}_i$~~ $\hat{e}$ is Not constant
- variability will be smaller for large h_{ii}
- look like Null plots

- Incorrect : not null plots

STA302 Final Review.

These are the whole semester's formula and proofs.

Chapter 2: Simple Linear Regression (part I).

SLR model: $E(Y|X=x)=\beta_0+\beta_1 x$ $\qquad Var(Y|X=x)=\sigma^2$

alternative form: $y_i=\beta_0+\beta_1 x_i+e_i$, $E(e_i)=0$, $Var(e_i)=\sigma^2$, e_i's iid.

e_i: statistical error, no negative meaning, vertical distance. y_i and "true value of $E(Y|X=x_i)$"

Parameter estimation:

fitted value for case i: $\hat{y}_i=\hat{E}(Y|X=x_i)=\hat{\beta}_0+\hat{\beta}_1 x_i$

residual: $\hat{e}_i=y_i-\hat{y}_i=y_i-\hat{E}(Y|X=x_i)=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i$ (vertical distance y_i and $\hat{y}_i$)

OLS (ordinary least squares)

RSS (residual sum of squares)

$RSS(\beta_0,\beta_1)=\sum[y_i-(\beta_0+\beta_1 x_i)]^2$

we want RSS minimized, $(\hat{\beta}_0,\hat{\beta}_1)=\arg\min_{\beta_0,\beta_1} RSS(\beta_0,\beta_1)$

• method: differentiate w.r.t. to β_0 & β_1, set them to be 0.

$$\frac{\partial RSS(\beta_0,\beta_1)}{\partial\beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0 \qquad \frac{\partial RSS}{\partial\beta_1}=-2\sum x_i(y_i-\beta_0-\beta_1 x_i)=0$$

$\Rightarrow \beta_0 n+\beta_1\sum x_i=\sum y_i$, $\beta_0\sum x_i+\beta_1\sum x_i^2=\sum x_i y_i$

solve this based on:

$SXX=\sum(x_i-\bar{x})^2=\sum(x_i-\bar{x})x_i$ sum of squares for x's

$SD_x^2=\frac{SXX}{n-1}$ sample variance of the x's.

$SD_x=\sqrt{\frac{SXX}{n-1}}$ sample standard deviation of the x's.

$SXY=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum(x_i-\bar{x})y_i$ ~~sam~~ sum of cross-products

$s_{xy}=\frac{SXY}{n-1}$ sample covariance

$r_{xy}=\frac{s_{xy}}{SD_x SD_y}$ sample correlation

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}, \qquad \hat{\beta}_1=\frac{SXY}{SXX}$$

Estimating σ^2

$\hat{\sigma}^2=\frac{RSS}{n-2}$ (average size of $\hat{e}_i^2=(y_i-\hat{y}_i)^2$) this is called "the residual mean square"

$df=n-2$

$\hat{\sigma}$ is called "standard error of regression"

$RSS=SYY-\hat{\beta}_1^2 SXX$ (by definition)

· Properties of LSE

Let $c_i = \frac{x_i - \bar{x}}{SXX}$, and $\sum(x_i - \bar{x})\bar{y} = 0$, then $\hat{\beta}_1 = \sum \frac{x_i - \bar{x}}{SXX} y_i = \sum c_i y_i$ & fitted line passes $(\bar{x}, \bar{y})$

· estimators are unbiased ($E(\hat{\theta}) = \theta$)

in details: ① $\cancel{E(\hat{\beta}_0|X) = \beta_0 + \beta_1\bar{x}}$ since $E(\bar{y}|X) = \beta_0 + \beta_1\bar{x}$

then $E(\hat{\beta}_0|X) = E(\hat{\bar{y}}|X) - \beta_1\bar{x} = \beta_0$

② $E(\hat{\beta}_1|X) = E(\sum c_i y_i | X = x_i) = \sum c_i E(y_i | X = x_i) = \sum c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i$

$\sum_i c_i = \sum(x_i - \bar{x}) = 0$

$\boxed{\sum c_i x_i = \frac{\sum(x_i - \bar{x})x_i}{SXX} = 1}$

$\cancel{E(\hat{\beta}_1|X)} = \beta_1$

$\boxed{③\ E(\hat{\sigma}^2|X) = \sigma^2}$

$\cancel{E(\hat{\sigma}^2|X) = E(\frac{RSS}{n-2}|X=x_i) = E((SYY - \hat{\beta}_1^2 SXX)|X=x_i)}$

$\cancel{= E(SYY|X=x_i) - \beta_1^2 E(SXX|X=x_i)}$

Variances of the LSE (want them small)

① $Var(\hat{\beta}_1|X) = \frac{\sigma^2}{SXX}$

② $Var(\hat{\beta}_0|X) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$

③ $Cov(\hat{\beta}_0, \hat{\beta}_1|X) = -\sigma^2 \frac{\bar{x}}{SXX}$

④ $\rho(\hat{\beta}_0, \hat{\beta}_1|X) = \frac{-\bar{x}}{\sqrt{SXX/n + \bar{x}^2}} = \frac{-\bar{x}}{\sqrt{(n-1)SD_x^2/n + \bar{x}^2}}$

σ^2 are unknown

to estimate those, replace σ^2 by $\hat{\sigma}^2$ $\Rightarrow$ $\hat{Var}(\hat{\beta}_1|X) = \frac{\hat{\sigma}^2}{SXX}$, $\hat{Var}(\hat{\beta}_0|X) = \hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$

"standard error" is $se(\hat{\beta}_1|X) = \sqrt{\hat{Var}(\hat{\beta}_1)}$

$se(\hat{\beta}_0|X) = \sqrt{\hat{Var}(\hat{\beta}_0)}$

Derivation of ①②③④

① $Var(\hat{\beta}_1|X) = Var(\sum c_i y_i|X) = \sum c_i^2 Var(y_i|X = x_i) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (x_i - \bar{x})^2 / SXX^2 = \frac{\sigma^2}{SXX}$

② $Var(\hat{\beta}_0|X) = Var(\bar{y} - \hat{\beta}_1\bar{x}|X) = Var(\bar{y}|X) + \bar{x}^2 Var(\hat{\beta}_1|X) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1|X)$

since $Cov(\bar{y}, \hat{\beta}_1|X) = Cov(\frac{1}{n}\sum y_i, \sum c_i y_i|X)$

$= \frac{1}{n}\sum c_i Cov(y_i, y_i|X)$

$= \frac{\sigma^2}{n}\sum c_i = \frac{\sigma^2}{n}\sum(x_i - \bar{x}) = 0$

$Var(\bar{y}|\bar{x}) = Var(\frac{1}{n}\sum y_i|X = x_i) = \frac{1}{n} Var(\sum y_i|X = x_i) = \frac{\sigma^2}{n}$?

So $Var(\hat{\beta}_0|X) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{SXX} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$

③ $Cov(\hat{\beta}_0, \hat{\beta}_1|X) = Cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1|X) = Cov(\bar{y}, \hat{\beta}_1|X) - \bar{x}\,cov(\hat{\beta}_1, \hat{\beta}_1|X)$

$= 0 - \sigma^2\frac{\bar{x}}{SXX}$

$= -\sigma^2\frac{\bar{x}}{SXX}$

④ $\rho(\hat{\beta}_0, \hat{\beta}_1|X) = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1|X)}{\sqrt{Var(\hat{\beta}_0|X)Var(\hat{\beta}_1|X)}}$

Chapter 2: SLR (part II)

· Comparing models → Analysis of Variance (ANOVA)

$E(Y|X=x)=\beta_0$ v.s. $E(Y|X=x)=\beta_0+\beta_1 x$

$\downarrow$ $\cancel{\hat{\beta_0}=\Sigma(y_i-\beta_0)^2}$

$RSS=\Sigma(y_i-\beta_0)^2=\Sigma(y_i-\bar{y})^2=SYY$ with $df=n-1$

$\left.\begin{array}{l}\widehat{E(Y|X)}=\hat{\beta}_0 \\ \widehat{E(Y|X)}=\hat{\beta}_0+\hat{\beta}_1 x\end{array}\right\} \Rightarrow$ compare two RSS's

$RSS_1=SYY$

$RSS_2=SYY-\frac{SXY^2}{SXX}$

$RSS_2 \leq RSS_1$

How small is small?

the difference between RSS_1 & RSS_2 is "sum of squares due to regression" (SSreg)

$SSreg = RSS_1 - RSS_2$

$= SYY - SYY + \frac{SXY^2}{SXX}$

$= \frac{SXY^2}{SXX}$

$df=(n-1)-(n-2)=1$

ANOVA (we compare the "standard version of SSreg" v.s. "stand version of RSS_2")

Source	df	SS	MS	F	p-value
Regression	1	SSreg	$\frac{SSreg}{1}$	$\frac{MSreg}{\hat{\sigma}^2}$	
Residual	n-2	RSS	$\hat{\sigma}^2=\frac{RSS}{(n-2)}$		
Total	n-1	SYY			

MS: "mean squares"

· if RSS_2 is good, (adding β_1 helps), $RSS_2 \ll RSS_1$ then SSreg $\uparrow$

the $F=\frac{MSreg}{\hat{\sigma}^2}=\frac{SSreg/1}{\frac{RSS}{n-2}}$ will be $\uparrow$

So in general, comparing two models: → comparing $F_{obs}=\frac{RSS_a/a}{RSS_b/b}$, where a, b are df.

with

$F(\alpha, 1, n-2)$

if $F_{obs} > F_{(\alpha,1,n-2)}$, reject $NH: E(Y|X=x)=\beta_0$

$F_{obs} \leq F_{(\alpha,1,n-2)}$, don't reject NH

• p-value, the strength of the evidence against NH in favor of AH.

Coefficient of Determination, R^2.

$R^2 = \frac{SSreg}{SYY}$ "how useful slope is"

measures the strength of the relationship between x_i & y_i

SYY ~~SSreg~~: variability ~~on~~ in the ~~dat~~ data

~~SYY~~

SSreg: variability in the data explained by the slope.

• note $R^2 = \frac{SSreg}{SYY} = \frac{(SXY)^2}{SXX\,SYY} = r_{xy}^2$ → correlation square.

CI & Tests

$\hat{\theta}$ is an est. of θ, a $100(1-\alpha)\%$ CI for θ is $(\hat{\theta} - t(\frac{\alpha}{2}, d)\,se(\hat{\theta}), \hat{\theta} + t(\frac{\alpha}{2}, d)\,se(\hat{\theta}))$

$d = df$

① For β_0

• assume e_i's are iid $N(0, \sigma^2)$

CI: $(\hat{\beta}_0 - t(\frac{\alpha}{2}, n-2)\,se(\hat{\beta}_0), \hat{\beta}_0 + t(\frac{\alpha}{2}, n-2)\,se(\hat{\beta}_0))$

where $se(\hat{\beta}_0) = \hat{\sigma}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})^{\frac{1}{2}}$

t ~~statistics~~ $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$, say $\beta_0^* = 0$, in NH: $\beta_0 = \beta_0^*$

$\downarrow$

and compare to $t(\frac{\alpha}{2}, n-2)$.

② For β_1

CI. $\hat{\beta}_1 \pm t(\frac{\alpha}{2}, n-2)\frac{\hat{\sigma}}{\sqrt{SXX}})$

~~same~~ Hypothesis test: similar to $\hat{\beta}_0$

CI ~~for~~ & tests - t & F

t-statistic: $t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SXX}}$

$t^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/SXX} = \frac{\hat{\beta}^2 SXX}{\hat{\sigma}^2}$ = F statistic in ANOVA

• Prediction $\neq$ Fitted value.

① $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$ $\quad$ $\tilde{y}_*$ predicts unbiasedly the unobserved y_*.

$$Var(\tilde{y}_* - y_* | X, x_*) = Var(y_* | x_*) + Var(\tilde{y}_* | X, x_*)$$
$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)$$

$$\text{sepred}(\tilde{y}_* - y_* | X, x_*) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)^{\frac{1}{2}}$$

prediction interval: $\quad \tilde{y}_* \pm t_{(\frac{\alpha}{2}, n-2)} \text{sepred}(\tilde{y}_* | X, x_*)$

② Fitted value.

$$\text{sefit}(\hat{y} | X, x) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}\right)^{\frac{1}{2}} \qquad CI: (\hat{\beta}_0 + \hat{\beta}_1 x) \pm \text{sefit}(\hat{y} | X, x)[2F(\alpha; 2, n-2)]$$

Chapter 3: Multiple Linear Regression (MLR)

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
$$Var(Y|X) = \sigma^2$$

• predictors (original data you collect), terms (created from predictors)

Matrix notation

observed values $(x_{11}, x_{12}, \cdots, x_{1p}, y_1)$
$(x_{21}, \cdots, x_{2p}, y_2)$
$\vdots$
$(x_{n1}, \cdots, x_{np}, y_n)$

• $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ $\quad$ ith row of X will be x_i'

• $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ $\quad x_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$ $\quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$

• $Y = X\beta + e$

namely
$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + e_1 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + e_n \end{pmatrix}$$

• $E(e) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\quad Var(e) = \sigma^2 I_n, \quad e \sim N(0, \sigma^2 I_n)$

Ordinary Linear Squares

$$\boxed{\begin{array}{l} RSS(\beta)=\Sigma(y_i-x_i'\beta)^2=(Y-X\beta)'(Y-X\beta) \\ \text{if } (X'X)^{-1} \text{ exists, } RSS(\beta) \text{ is minimized by } \hat{\beta}=\underbrace{(X'X)^{-1}}_{S_{XX}}\underbrace{X'Y}_{S_{XY}} \end{array}}$$

similar to SXX SXY

Residuals: $\hat{e}=Y-\hat{Y}$

$RSS=\hat{e}'\hat{e}=(Y-X\hat{\beta})'(Y-X\hat{\beta})$

$\sigma^2=Var(Y|X)$ is estimated with $\hat{\sigma}^2=\frac{RSS}{n-(p+1)}$ b/c $(p+1)$ parameters

OLS estimates using Matrices

minimize $RSS(\beta)=(Y-X\beta)'(Y-X\beta)=Y'Y+\beta'(X'X)\beta-2Y'X\beta$

similarly setting $\frac{\partial c\beta}{\partial\beta}=c'$ and $\frac{\partial\beta'V\beta}{\partial\beta}=(V+V')\beta$ to be 0.

we have normal equation : $X'X\beta=X'Y$

OLS est. is $\hat{\beta}=(X'X)^{-1}X'Y$

Properties of OLS estimates:

- Assume $E(e)=0$, $Var(e)=\sigma^2 I_n$, $\hat{\beta}$ is unbiased b/c

$E(\hat{\beta}|X)=E((X'X)^{-1}X'Y|X)=(X'X)^{-1}X'E(Y|X)=(X'X)^{-1}X'X\beta=\beta$

- Variance, note $Var(B'Z)=B'Var(Z)B$

$Var(\hat{\beta}|X)=Var((X'X)^{-1}X'Y|X)=\cancel{(X'X)^{-1}X'[\sigma^2 I_n]X(X'X)^{-1}}$

$=(X'X)^{-1}X'[Var(Y|X)X(X'X)^{-1}$

$=(X'X)^{-1}X'[\sigma^2 I_n]X(X'X)^{-1}$

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

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

- $RSS=RSS(\hat{\beta})=(Y-X\hat{\beta})'(Y-X\hat{\beta})=Y'Y+\hat{\beta}'(X'X)\hat{\beta}-2Y'X\hat{\beta}$

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

So $RSS=Y'Y-\hat{\beta}X'X\hat{\beta}=Y'Y-\hat{Y}'\hat{Y}$ with $\hat{Y}=X\hat{\beta}$

Two models: $(E(Y|X)=\beta_0+\Sigma\beta_j x_j$ v.s. $E(Y|X)=\beta_0)$

ANOVA

Source	df	SS	MS	F	p-value
Regression	p	SSreg	SSreg/p	$\frac{MSreg}{\hat{\sigma}^2}$	
Residual	$n-(p+1)$	RSS	$\hat{\sigma}^2=\frac{RSS}{n-(p+1)}$		
Total	$n-1$	SYY			

Coefficient of Determination : (same) $R^2=\frac{SSreg}{SYY}=1-\frac{RSS}{SYY}$

Sequential Analysis of variance.

Two models: $RSS_S - RSS_B \rightarrow$ contribution of Tax.

$SS_{reg} = SS_{others} + SS_{tax}$

Dlic, Income, logMiles — Tax, given Dlic, Income, logMiles.

& the order of decomposition matters, $\_\_$ given $\_\_$

Predictions & Fitted values (similar to ~~lin~~ SLR)

$\hat{y}_* = x_*'\hat{\beta}$, $\mathrm{sepred}(\hat{y}_*|x_*) = \hat{\sigma}\sqrt{1 + x_*'(X'X)^{-1}x_*}$

prediction: given a new x_*, predict y_* with

$\hat{E}(Y|X=x) = \hat{y} = x'\hat{\beta}$

$\mathrm{sefit}(\hat{y}|x) = \hat{\sigma}\sqrt{x'(X'X)^{-1}x}$

$\mathrm{sepred}(\hat{y}_*|x_*) = \sqrt{\hat{\sigma}^2 + \mathrm{sefit}(\hat{y}_*|x)}$

fitted value: given a value x, estimate the mean function at x.

Chapter 4 : Drawing Conclusions

$E(Y|X) = 15 + 3x_1 + 4X_2 - 2X_3$ (change in X_1 affect X_2, X_3? Maybe, because of correlation)

same model, different interpretation $\Rightarrow R^2$, $\hat{\sigma}^2$ are identical, estimates & t-values are not.

Sampling from Normal Population

- $\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$
- $y_i|x_i \sim N\left(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\right)$
- define $\beta_0 = \mu_y - \beta_1\mu_x$, $\beta_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}$; $\sigma^2 = \sigma_y^2(1-\rho_{xy}^2)$

$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

- $\hat{\mu}_x = \bar{x}$, $\hat{\mu}_y = \bar{y}$, $\hat{\sigma}_x^2 = \frac{SXX}{n-1}$, $\hat{\sigma}_y^2 = \frac{SYY}{n-1}$, $\hat{\rho}_{xy} = \frac{SXY}{\sqrt{SXX\,SYY}}$
- plug-in to get $\hat{\beta}_0, \hat{\beta}_1 \Rightarrow$ OLS estimates

Missing data? (missing at random) MAR is easiest to handle.
① deleting ② guessing ③ more advanced method: imputation).

Chapter 5 Weighed Least Squares & Lack of Fit

WLS. so ~~=~~ predictors have weights.

change to $Var(Y|X=x_i) = Var(e_i) = \frac{\sigma^2}{w_i}$, where $w_i > 0$. (known #s)

matrix form: $Y = X\beta + e$, $Var(e) = \sigma^2 W^{-1}$, W is diagonal with $w_1, \dots, w_n$

- the estimator β is defined to be the minimizer of $RSS(\beta) = \sum w_i (y_i - x_i'\beta)^2$
$= (Y - X\beta)' W (Y - X\beta)$

so $\hat\beta = (X'WX)^{-1} \cancel{X'W} X'WY$

How to get this? ① $Var(W^{\frac{1}{2}}e) = W^{\frac{1}{2}} Var(e) W^{\frac{1}{2}} = W^{\frac{1}{2}}(\sigma^2 W^{-1}) W^{\frac{1}{2}} = \sigma^2 I$

② $W^{\frac{1}{2}}Y = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}e$. let $Z = W^{\frac{1}{2}}Y$, $M = W^{\frac{1}{2}}X$, $d = W^{\frac{1}{2}}e$

then $Z = M\beta + d$

$\hat\beta = (M'M)^{-1}M'Z = (X'WX)^{-1}(X'WY)$

- Lack of Fit (LOF)

F test in ANOVA tells only if the regression model is explaining or not.

Lack of Fit idea: ① model good, $E(\hat\sigma^2) \approx \sigma^2$

② not, $\hat\sigma^2 > \sigma^2$

Two cases: σ^2 known/unknown

- known (no lack of fit (NH), $X^2 = \frac{RSS}{\sigma^2} = \frac{(n-(p+1))\hat\sigma^2}{\sigma^2} \sim \chi^2_{(n-(p+1))}$)

p-value is $P(X^2 \geq X^2_{obs} \mid \text{no lack of fit})$

- unknown (we can do this if we have some replicates)

- SS_{pe} : sum of squares for pure error.

x_i	y_{ij}	$\bar y_i$	$\Sigma(y_{ij} - \bar y_i)^2$	$\hat\sigma$	df
1	.	2.6233	0.0243	0.1102	2
1	.				
1	.				
2	2.4	2.4	0	0	0
3	4....	4.4450	0.1301	0.3606	1
3	...				
4		4.0325	2.2041	0.8571	3
4	..				
4					
4	....				
			2.3585		

$SS_{pe} = 2.3585$ with 6 df

$\hat\sigma^2_{pe} = \frac{SS_{pe}}{df} = 0.3951$

$RSS = SS_{lof} + SS_{pe}$

$\downarrow$ SS due to lof $(\bar y_i \Rightarrow \beta_0 + \beta_1 x_i)$

$\downarrow$ SS due to pe $(y_{ij} \Rightarrow \bar y_i)$

Decomposition

$$RSS_{ols} = \sum_i^n \sum_j^n (y_{ij} - \hat\beta_0 - \hat\beta_1 x_i)^2 = \sum_i \sum_j (y_{ij} - \bar y_i + \bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$$
$$= \Sigma\Sigma (y_{ij} - \bar y_i)^2 + \sum_i n_i (\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 + 2\sum_i \left(\sum_j (y_{ij} - \bar y_i)\right)(\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)$$
$$= \Sigma\Sigma (y_{ij} - \bar y_i)^2 + \sum_i n_i (\bar y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$$
$$= SS_{pe} + SS_{lof}$$
$$= SS_{pe} + RSS_{wls}$$

- $F\text{value} = \dfrac{SS_{lof}/df_{lof}}{SS_{pe}/df_{pe}}$ compare with $F(df_{lof}, df_{pe})$

- General F-test NH, AH.

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH} / \cancel{RSS_{AH}}\, df_{AH}} \quad \text{v.s.} \quad F(\alpha, \cancel{dft}\ df_{NH} - df_{AH}, df_{AH})$$

Chapter 6: Polynomials & Factors

Case ~~poly~~ straight lines or quadratic?) By ~~testing~~ F-test

$\boxed{\text{1 predictors}}$

- For $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$, solve $\dfrac{dE(Y|X)}{dx} = 0 \Rightarrow x_M = \dfrac{-\beta_1}{2\beta_2}$ (Max/min)

- $\boxed{\text{General predictors}}$

$$E(Y|X_1 = x_1, X_2 = x_2) = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{linear}} + \underbrace{\beta_{11} x_1^2 + \beta_{22} x_2^2}_{\text{quadratic}} + \underbrace{\beta_{12} x_1 x_2}_{\text{interaction}}$$

~~interc~~

- Highest order 2, k predictors $\Rightarrow$ 1 intercept
  - k linear terms
  - k quadratic terms
  - $\frac{k(k-1)}{2}$ interaction terms

$\boxed{\text{The Delta Method}}$ provides approximate standard errors for nonlinear combinations of parameter estimaties.

Sps $\hat\theta \,\dot\sim\, N(\theta, \Sigma)$ & $g(\theta)$ is a continuous function of θ

when $n \to \infty$, we have

$$E[g(\hat\theta)] \approx g(\theta)$$
$$Var[g(\hat\theta)] \approx \dot g(\theta)' \Sigma \dot g(\theta), \text{ where } \dot g(\theta) = \frac{\partial g}{\partial \theta} = \left(\frac{\partial g}{\partial \theta_1}, \cdots, \frac{\partial g}{\partial \theta_k}\right)'$$

Back to $\hat x_M$

$\beta = (\beta_0, \beta_1, \beta_2)'$, $\hat\beta = (\hat\beta_0, \hat\beta_1, \hat\beta_2)'$

for large n, $\hat\beta \,\dot\sim\, N(\beta, \sigma^2 (X'X)^{-1})$

$$\widehat{Cov}(\hat\beta) \approx \hat\sigma^2 (X'X)^{-1}$$

$$g(\hat\beta) = \frac{-\hat\beta_1}{2\hat\beta_2} \Rightarrow \dot g(\hat\beta) = \left(0, \frac{-1}{2\hat\beta_2}, \frac{\hat\beta_1}{2\hat\beta_2^2}\right)$$
$$Var(g(\hat\beta)) = \dot g(\hat\beta)' \widehat{Cov}(\hat\beta) \dot g(\hat\beta)$$
$$= \frac{1}{4\hat\beta_2^2}\left(Var(\hat\beta_1) + \frac{\hat\beta_1^2}{\hat\beta_2^2} Var(\hat\beta_2) - \frac{2\hat\beta_1}{\hat\beta_2} Cov(\hat\beta_1, \hat\beta_2)\right)$$

$$P\left(\frac{|g(\hat\beta) - g(\beta)|}{\sqrt{\dot g(\hat\beta)' \Sigma \dot g(\hat\beta)}} < z_{\frac{\alpha}{2}}\right) = \cancel{1-}\alpha$$

$$\cancel{\hat x_M \pm z_{\frac{\alpha}{2}} \sqrt{\dot g(\hat\beta; x_2)}}$$

$\boxed{\text{Factors}}$ levels, dummy variables

~~Factors~~ Rule: A factor with d levels can be replaced by at most ~~d~~ d dummy variables.
If the intercept is in the mean function, at most d-1 of the dummy variables can be used in the mean function.

e.g.

U_1	U_2	U_3
1	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	1
1	0	0

sum of U_j's is a column of ones.

jth dummy variable U_j, $j=1,\cdots,5$

$$u_{ij} = \begin{cases} 1 & \text{if } D_i = j\text{th category of } D \\ 0 & \text{o.w.} \end{cases}$$

reg. model: $E(TS|D) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5$ (mean function does not have an intercept)

β_j: population mean for all species with Danger index j.
TS: total hour of ~~#~~ sleep
D: danger index

equivalent model: $E(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$
(choice to delete first dummy variable)
$\eta_0 = \beta_1$, $\eta_0 + \eta_2 = \beta_2$, $\eta_0 + \eta_3 = \beta_3$, $\cdots$, $\eta_0 + \eta_5 = \beta_5$

need to see quiz 2 for example.

"One-way analysis of variance" models

4 mean functions for the sleep data.

	df	RSS	F	P(>F)
Model 1, most general				
2				
3				
4				

$l = 2,3,4$

$$F_l = \frac{(RSS_l - RSS_1)/(df_l - df_1)}{RSS_1/df_1}$$

Chapter 7. Transformation

① tranform predictor ② response ③ both. s.t. we have $E(Y|X=x) \simeq \beta_0 + \beta_1 x$

Power Trans (be applied to ①②③)

U: original variable, positive $\quad \psi(U,\lambda) = U^\lambda \quad \lambda: -2$ to 2

Scaled Power Trans (① only)

$$\psi_S(X,\lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log_e(X) & \text{if } \lambda = 0 \end{cases}$$

continuous with $\lim_{\lambda\to 0} \psi_S(X,\lambda) = \log_e(X)$

How to choose λ?
fit $(\psi_S(X,\lambda), Y)$ to try minimizing RSS
some usual choice
$\lambda \in \{-1, -\frac{1}{2}, 0, \frac{1}{3}, \frac{1}{2}, 1\}$

Box-Cox Trans for Response (2)

modified power trans : for $Y > 0$.

$\psi_m(Y, \lambda_y) = \psi_S(Y, \lambda_y) \times gm(Y)^{1-\lambda_y}$

$= \begin{cases} gm(Y)^{1-\lambda_y} \times (Y^{\lambda_y} - 1)/\lambda_y & \text{if } \lambda_y \neq 0 \\ gm(Y) \times \log(Y) & \text{if } \lambda_y = 0 \end{cases}$ $gm(Y) = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log_e y_i\right)$ geometric mean.

B-C ~~tran~~ method assumes $E(\psi_m(Y, \lambda_y) | X = x) = \beta' x$

$\lambda_y \to$ minimize $RSS(\lambda_y)$

goal of B-C : not for linearity, but for normality.

Chapter 8 : Diagnostics via residuals

"check ~~model~~ whether ~~it~~ the linear model assumptions are satisfied or not"

Residuals :

$\hat{\beta} = (X'X)^{-1}X'Y$, $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$, define $H = X(X'X)^{-1}X'$

H : hat matrix, transform data Y into fitted values $\hat{Y}$.

residuals $\hat{e} = Y - \hat{Y} = Y - HY = (I - H)Y$

H is idempotent proj. matrix. $H' = H$, $HH = H$, $HX = X$

assumptions for e (statistical error)

$E(e) = 0$, $Cov(e) = \sigma^2 I \Rightarrow E(\hat{e}) = 0$ and $Cov(\hat{e}) = \sigma^2(I - H)$

h_{ii} : the ith diagonal element of H "Leverage value"

$Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$

if intercept is included, $\sum \hat{e}_i = 0$ (check SLR case)

HAT MATRIX: ~~H, (I-H) are~~ $(I-H)$ idempotent

some direct consequences.

$(I-H)X = 0 \Rightarrow E(\hat{e}) = 0$, $H(I-H) = 0$

$Cov(\hat{e}, \hat{Y}) = Cov((I-H)Y, HY) = \sigma^2 H(I-H) = 0$

$Cov(Y) = \sigma^2 I$, $Cov(\hat{Y}) = \sigma^2 HH' = \sigma^2 H$

$Cov(\hat{e}) = \sigma^2(I-H)(I-H)' = \sigma^2(I-H)$

note that $Cov(\hat{e}) = Cov(Y - \hat{Y}) = Cov(Y) - Cov(\hat{Y})$

• $h_{ij} = x_i'(X'X)^{-1}x_j = x_j'(X'X)^{-1}x_i = h_{ji}$,

$\sum_{i=1}^{n} h_{ii} = p'$, if mean function includes an intercept, $\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$

$\frac{1}{n} \le h_{ij} \le \frac{1}{r_i}$, r_i is the number of rows of X that are identical to x_i.

$h_{ii} \uparrow$, $\rightarrow Var(\hat{e}_i) \downarrow$

$Var(\hat{e}_i) = 0 \rightarrow$ ith observation is $\underset{\text{only}}{\wedge}$ used to get $\hat{y}_i$.

high $h_{ii} \uparrow \longrightarrow$ unusual value for x_i. (but $\neq$ outliers)

- H idempotent, $\hat{h}_{ii} = h_{ij}^2$ i.e. $h_{ii}(1-h_{ii}) = \sum_{j \neq i} h_{ij}^2$

$$\hat{y}_i = \sum_{i=1}^n h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^n h_{ij} y_j$$

as $h_{ii} \to 1$, $\hat{y}_i \to y_i$, $\hat{y}_i$ is mostly determined by y_i only.

Doing WLS

$Var(e) = \sigma^2 W^{-1}$ (assumption)

$H = W^{\frac{1}{2}} X (X'WX)^{-1} X' W^{\frac{1}{2}}$, fitted values: $\hat{Y} = X\hat{\beta} = HY$

residuals defined as $\hat{e}_i = \sqrt{w_i}(y_i - \hat{y}_i)$ Pearson/weighted residual.

When Model correct!

U be any terms/linear combination of terms, then

$E(\hat{e}_i | U_i) = 0$, $Var(\hat{e}_i | U_i) = \sigma^2 (1 - h_{ii})$

so
- plot of residuals : mean zero, like null plot
- variance function : not constant
- variability $\downarrow$, large $h_{ii} \uparrow$

Testing curvature in residual plot.

test $\hat{e}$ v.s. U

1. refit the data with original model $+ U^2$
2. test the significance of the coefficient of U^2

- if U does not depend on any estimated coefficients, use t-test
  o.w. use z-test

Nonconstant Variance. (issues that residual plots may often have)

① do WLS (problem is to determine weights)

② variance stablizing transformation.

$Var(Y|X=x) = \sigma^2 \implies Var(Y|X=x) = \sigma^2 g(E(Y|X=x))$

where g is usually $\begin{cases} \log(Y) \\ \sqrt{Y} \\ Y^{-1} \end{cases}$

Chapter 9 : Outliers & Influence

Mean shift outlier model

- non-outlier : $E(Y|X=x_i)=x_i'\beta$

  outlier : $E(Y|X=x_i)=x_i'\beta+\delta$

NH: $\delta=0$ (TEST) this

Variance stays the same.

An Outlier Test

sps ith case is an outlier. ~~suppose~~ suspected to be

dummy variable $U: \begin{cases} u_j=0 & \text{for } j\neq i \\ u_j=1 & \text{o.w.} \end{cases}$

then fit the model using least squares.

$$E(Y|X)=X\beta+\delta U$$

$\hat{\delta}$ is estimated mean shift

two sided t-test. NH: $\delta=0$, AH: $\delta\neq 0$.

df : $n-p'-1$.

WHAT IS

P': $\beta_0, \beta_1, \cdots, \beta_p$

P: $X_1, X_2, \cdots, X_p$

$p'=p+1$

Alternative approach :

① ith - suspected outlier

② Step 1 : delete ith case ($n-1$) points left.

③ Step 2 : use reduced data, estimate β, σ^2.

denote $\hat{\beta}_{(i)}$, $\hat{\sigma}^2_{(i)}$, df ~~for~~ is $n-p'-1$.

④ Step 3 : compute the fitted value for the deleted case :

$\hat{y}_{i(i)}=x_i'\hat{\beta}_{(i)}$

$Var(y_i-\hat{y}_{i(i)})=Var(y_i)+Var(\hat{y}_{i(i)})=\sigma^2+\sigma^2 x_i'(X_{(i)}'X_{(i)})^{-1}x_i$

$X_{(i)}$ is the matrix X with ith row deleted.

⑤ Step ~~5~~ 4 : under the model, we have

$E(y_i)=x_i'\beta+\delta$, $E(\hat{y}_{i(i)})=E(x_i'\hat{\beta}_{(i)})=x_i'\beta \Rightarrow E(y_i-\hat{y}_{(i)})=\delta$

and t-statistic for $\delta=0$ is

$$t_i=\frac{y_i-\hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+x_i'(X_{(i)}'X_{(i)})^{-1}x_i}}$$ use $\hat{\sigma}_{(i)}$ to replace $\hat{\sigma}$

with df $=n-p'-1$, this is identical ↑.

Why prefer second approach?

• define "standardized residual" $r_i = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

$$t = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{\frac{1}{2}}$$

• Bonferroni Adjustment: if we ~~want~~ have n data points, apply the above t-tests to all cases, adjust the overall significance level to α.

BA: change α to $\frac{\alpha}{n}$, then the overall sig. $< \alpha$.

Influence Analysis. (remove once a time then compare)

Cook's distance.

$$D_i = \frac{(\hat{\beta}_{(i)}-\hat{\beta})'(X'X)(\hat{\beta}_{(i)}-\hat{\beta})}{p'\hat{\sigma}^2}$$

$$= \frac{(\hat{Y}_{(i)}-\hat{Y})'(\hat{Y}_{(i)}-\hat{Y})}{p'\hat{\sigma}^2}$$

$$= \frac{1}{p'}r_i^2\frac{h_{ii}}{1-h_{ii}}$$

a normalized distance between $\hat{\beta}_{(i)}$ & $\hat{\beta}$

a scaled Euclidean distance b/w $\hat{Y}_{(i)}$ & $\hat{Y}$

$D_i \uparrow \rightarrow$ potential problem. (compare different suspects)

~~Normalit~~ Normal Probability Plots (check for normality of e_i)

• Q-Q plot: have iid random #s $\{x_1, \cdots, x_n\}$

① sort $x_{(1)} \leq \cdots \leq x_{(n)}$

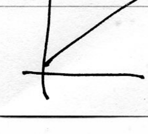
② find expected order statistic $u_{(1)} \leq \cdots u_{(n)}$ from $N(0,1)$, $u_{(i)}$ is the $\frac{100i}{n}$ th percentile.

$P(Z \leq u_{(i)}) = \frac{i}{n}$, $Z \sim N(0,1)$

③ if $x_i \sim N(\mu, \sigma^2)$, then $E(x_{(i)}) = \mu + \sigma u_{(i)}$

$\Rightarrow$ Q-Q plot (sample quantile v.s. population quantile) plot.

normal: "line"



Chapter 10

Variable Selection / model selection

Collinearity (caused by redundant terms)

e.g. $c_1X_1 + c_2X_2 = c_0$ for some c_0, c_1 & c_2.

$c_1X_1 + \cdots + c_pX_p \approx c_0$

collinearity is measured by square of sample correlation

(r_{12}^2 for two terms, max of r_{ij}^2 for multiple terms).

- $r_{12}^2 = 1 \Rightarrow$ collinearity.

  $r_{ij}^2 \approx 1 \Rightarrow$ approximate collinearity.

If collinearity: inverse of $X'X$ not exist

exists so no fitting can be done

to drop some terms!

if approx. collinearity: $var(\hat{\beta})$ ↗ large.

AC & Variance

$E(Y|X_1=x_1, X_2=x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\Downarrow$ $Var(\hat{\beta}_j) = \frac{\sigma^2}{1-r_{12}^2} \frac{1}{SX_jX_j}$

~~Show (10.7)~~

where $SX_jX_j = \sum (x_{ij} - \bar{x}_j)^2$

Fixing Collinearity

minimize $Var(\hat{\beta})$, remove redundant terms

Selection criteria: $f_1(RSS) + f_2(p)$ → penalty.

large models: RSS↓, p↑

small models: RSS↑, p↓

4. Common Criteria

Akaike information Criteria (AIC)

$n\log(\frac{RSS}{n})+2p$

Bayesian Information Criteria (BIC)

$n\log(\frac{RSS}{n})+p\log(n)$

Mallow's C_p

$\frac{RSS}{\hat{\sigma}^2}+2p-n$ ($\hat{\sigma}^2$ is estimated with all terms)

Cross-validation (CV)

$\sum_{i=1}^{n}(y_i-\hat{y}_{i(i)})^2=\sum_{i=1}^{n}\frac{\hat{e}_i^2}{(1-h_{ii})^2}$ (predictive RSS) → (PRESS)

A. tools when you know which model you want to fit

① 3 assumptions of LR

② OLS, WLS

③ CI, tests, t, F

④ se, (delta method)

⑤ pred

⑥ interpretation of fitted model

B. diagnostic Checking

① LOF test

② residual plots (3 components)

③ leverage h_{ii}

④ outlier test

⑤ Cook's distance

⑥ Q-Q plot.

C. Improving your model.

① Trans (3 types)

② adding / dropping terms

③ ridge regression.