10.1. The active terms

- Variable selection
  - Aim: Identify the correct model
    - select the useful predictor
    - Ignore the non-informative terms
  - Y v.s. X_1, X_2, … X_{999}
    - Divide $X=(X_1, X_2, \ldots X_{999})$ into two sets, X_A, and X_I,
    - so that $E(Y|X) = E(Y|X_A)=X_A\beta_A$
      - X_A = active terms
      - X_I =inactive terms

$X=(X_A, X_I)$

given X_A

given all $X_1 \cdots X_{999}$

$X = (X_1 \; X_2 \cdots X_5)$ aliased terms : $X_3 = c_0 + c_1X_1 + \cdots + c_5X_5$

10.1. Active terms and multicollinearity

$\Downarrow$ $(X'X)^{-1}$ does not exist $\Downarrow$ $\hat\beta = (X'X)^{-1}X'Y$ does not exist

- Multicollinearity
  - some terms can be approximated by linear combination of the other terms.
    - e.g. $X_3 \approx c_0+c_1X_1+c_2X_2+c_4X_4+c_5X_5$
  - In this case, X'X is close to singular (det=0),
    - $\hat{\beta} = (X'X)^{-1}X'Y$ and $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ can be huge.
    - We should avoid including all variables with multicollinearity in the regression model
      - e.g. set $X_A=(X_1,X_2,X_4,X_5)$,
      - $X_I=(X_3)$, since X_3 can be explained by X_1,X_2,X_4,X_5

$SYY = RSS + SS_{reg}$

10.1. Active terms and multicollinearity

- How to detect multicollinearity?
  - Check $(X'X)^{-1}$?
    - problem: don't know how large is large.
  - A better method: R^2_j, the coefficient of determination for the regression

$$X_j = c_0 + c_1 X_1 + \ldots + c_{j-1} X_{j-1} + c_{j+1} X_{j+1} + \ldots + c_p X_p + e$$

- $R^2_j \approx 1 \rightarrow$ multicollinearity, i.e. some terms can be approximated by linear combination of the other terms.

$$R^2 = \frac{SS_{reg}}{SYY}$$

1 = good fit

0 = bad fit

e.g. $(X_1\ X_2\ X_3\ X_4\ X_5)$

$X_1 = C_0 + C_1 X_2 + C_2 X_3 + C_3 X_4 + C_4 X_5 + e \rightarrow R_1^2$

$X_2 = \cdots(X_1, X_3\ X_4\ X_5) \rightarrow R_2^2$

$X_3 = (X_1, X_2, X_4\ X_5) \rightarrow R_3^2$

$X_4 = \cdots$

$X_5 = \cdots \rightarrow R_5^2$

If all R_j^2 are small $\Rightarrow$ no multi-collinearity

If some $R_j^2 \approx 1 \Rightarrow$ multicollinearity

10.1. Active terms and multicollinearity

- Relationship between $Var(\hat\beta_j)$ and R_j^2
  - Using the idea of Added Variable Plot

$$Let \quad X_O = (1 \; X_1 \, X_2 \ldots X_{j-1} \, X_{j+2} \ldots X_p), \quad H_O = X_O(X_O'X_O)^{-1}X_O'$$

$Y = X\beta + e$

$$For \quad Y = X_O\beta_O + X_j\beta_j + e,$$

$\hat e_{Y|X_O} \quad \hat e_{X_j|X_O}$

$$\hat\beta_j = \text{Regression coefficient between } (I-H_O)Y \text{ and } (I-H_O)X_j$$

$$= \left(X'_j(I-H_O)X_j\right)^{-1}X'_j(I-H_O)Y \quad \leftarrow (X'X)^{-1}X'Y$$

$$Var(\hat\beta_j) = \sigma^2\left(X'_j(I-H_O)X_j\right)^{-1} \quad \leftarrow \quad Var(AX) = A\,Var(X)A'$$

- For the regression

$$X_j = c_0 + c_1X_1 + \ldots + c_{j-1}X_{j-1} + c_{j+1}X_{j+1} \ldots + c_pX_p + e, \qquad or$$

$$X_j = X_Oc_O + e,$$

$$we\text{ have} \quad R_j^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{X'_j(I-H_O)X_j}{SX_jX_j} \quad \leftarrow Y'(I-H)Y$$

$$\frac{SSreg}{SYY} = \frac{SSreg}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

10.1. Variance Inflation Factor (VIF)

- Relationship between $Var(\hat{\beta}_j)$ and R_j^2
  - $$Var(\hat{\beta}_j) = \sigma^2 \left(X'_j (I - H_O) X_j \right)^{-1}$$
    $$R_j^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{X'_j (I - H_O) X_j}{SX_j X_j}$$

$y = X\beta + e$

$y = \beta_0 + \beta_j X_j + e$

(simple linear)

  - Therefore, $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{(1 - R_j^2) SX_j X_j}$

Compare to $Var(\hat{\beta}_j) = \dfrac{\sigma^2}{SX_j X_j}$

If $R_j^2 \approx 1 \Rightarrow \frac{1}{1-R_j^2}$ v.large

If $R_j^2 \approx 0 \Rightarrow \frac{1}{1-R_j^2} \approx 1$

  - Variance Inflation Factor (VIF)

$$\boxed{VIF = \frac{1}{(1 - R_j^2)}}$$

    - Increase in Variance due to multicollinearity
    - Big VIF or R_j^2 implies multicollinearity
    - In practice, $R^2 > 0.7$ is regarded as strong correlation
      - Becareful if $R_j^2 > 0.7$ or VIF > 1/0.3 = 3.33

$\frac{1}{1-0.7}$

10.1. Active terms and multicollinearity

- Example

```
x1=c(1,3,2,4,5,2,3,1,0,5)
x2=c(8,9,7,2,5,9,6,4,4,1)
x3=2*x1-5*x2+rnorm(10,0,0.1)
x4=c(3,1,4,2,7,3,4,5,6,3)
y=3+x1+2*x2+2*x4+rnorm(10,0,0.5)
summary(lm(y~x1+x2+x3))
#Find VIF
    Rj1=summary(lm(x1~x2+x3+x4))$r.squared; VIF1=1/(1-Rj1)
    Rj2=summary(lm(x2~x1+x3+x4))$r.squared; VIF2=1/(1-Rj2)
    Rj3=summary(lm(x3~x1+x2+x4))$r.squared; VIF3=1/(1-Rj3)
    Rj4=summary(lm(x4~x1+x2+x3))$r.squared; VIF4=1/(1-Rj4)
    print(rbind(c("Rj",Rj1,Rj2,Rj3,Rj4),c("VIF",VIF1,VIF2,VIF3,VIF4)))
#Modified fitting by deleting either one of x1,x2,x3
    summary(lm(y~x1+x2+x4))
    summary(lm(y~x2+x3+x4))
    summary(lm(y~x1+x3+x4))
```

$x_3 \approx 2x_1 - 5x_2$ multicollinearity

Y vs $X_1, X_2, \cdots\cdots X_{999}$

Model 1 : $Y = \beta_0 + \beta_1 X_1 \rightarrow RSS_1$ $RSS = \Sigma(Y_i - \hat{Y}_i)^2$ measures the goodness of fit of a regression

Model 2 : $Y = \beta_0 + \beta_2 X_2 \rightarrow RSS_2$

$\vdots \quad \vdots \quad \vdots$

Model k $\quad Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10} + \cdots \rightarrow RSS_k$

Q: Can we use RSS to select a good model?

A: No: The fitting of Regression is based on

$$\min_{\beta_M} RSS(\beta_M)$$

More parameter (more X variables) $\rightarrow$ Small RSS

$\Rightarrow$ Choose the model with most number of variab

Chapter10.pdf - Adobe Reader

File Edit View Window Help

9 / 20 74.4%

Tools Sign Comment

$RSS = \sum (Y_i - \hat{Y}_i)^2$

10.2. Automatic Variable Selection procedure

- For all possible candidate models, we compute
  - Akaike Information Criteria (AIC) — E_{Y^*}(likelihood($Y|X$))

    $$n \log\left(\frac{RSS}{n}\right) + 2p_C$$

    Number of parameters in the model
    e.g. p+1 in regression

  - Bayesian Information Criteria (BIC) — posterior prob(model)

    $$n \log\left(\frac{RSS}{n}\right) + p_C \log(n)$$

  - Mallow's C_p Statistics — $\sum (\hat{Y}_i - E(Y_i|X_i))^2$

    $\hat{\sigma}^2$ using all X

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

  - Predicted residual sum of Square (PRESS)

    $$\sum_{i=1}^{n} \{y_i - \hat{y}_{i(i)}\}^2 = \sum_{i=1}^{n} \left\{\frac{\hat{e}_i}{1 - h_{ii}}\right\}^2$$

$\hat{Y}_i$ — from fitting using all data

$\hat{Y}_{i(i)}$ — is not affected by (X_i, Y_i)

prediction of the Y_i, using reg fitting without (X_i, Y_i)

11.69 x 8.27 in

NEW PAGE TEMPLATES PREV. Page 14 of 23 NEXT PC MODE POWERPOINT BALL PEN HIGH LIGHT SIZE COLOR LINE RECT. CIRCLE SELECT ERASE ALL ERASE REC Recording 00:00:00 SAVE OPEN HELP EXIT

10.2. Model Selection in Practice

- When you have Y and X_1, X_2, … X_p 2 2 2
  - For each possible model (e.g. $y=\beta_o+\beta_1x_1$, $y=\beta_o+\beta_1x_1+\ldots+\beta_px_p$, $y=\beta_{p-1}x_{p-1}+\beta_px_p$ etc)
    - Find AIC, BIC, C_p, PRESS
    - Report the best model (smallest value) w.r.t each criteria
    - How many possible models ? (how many combinations)
      - 2^p
      - If p=20, 2^{20}=1048576 / 60x60x24
      - If a regression takes 1s , how long does it take for model selection?
  - Solutions
    - Forward Selection
    - Backward Selection

10.2. Automatic Variable Selection procedure

lack of fit
is large
⇕
RSS is
large
⇕
poor
fitting

AIC	BIC	C_p	PRESS
$n\log\left(\frac{RSS}{n}\right)+2p_c$	$n\log\left(\frac{RSS}{n}\right)+p_c\log(n)$	$\frac{RSS}{\hat{\sigma}^2}+2p_c-n$	$\sum_{i=1}^{n}\{y_i-\hat{y}_{i(i)}\}^2=\sum_{i=1}^{n}\left\{\frac{\hat{e}_i}{1-h_{ii}}\right\}^2$

- Smaller value implies a better model
  - For AIC, BIC, C_p, they have a common structure:

    lack of fit + penalty for model complexity

    - Larger model → Smaller RSS, Bigger penalty
    - Smaller model → Larger RSS, Smaller penalty
    - When p is fixed, they yield the same result (min RSS)
  - For PRESS, no penalty is need since different data are used for fitting and estimating errors

e.g. $Y \quad X_1, X_2, X_3$

$2 \quad 2 \quad 2$

Model 1: $Y = \beta_0 + \beta_1 x_1 + e \rightarrow AIC_1$

Model 2: $Y = \beta_0 + \beta_2 X_2 + e \rightarrow AIC_2$

(use BIC / C_p / Press instead of AIC)

$\vdots$

Model K: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + e \rightarrow AIC_k$

$\vdots$

Model 10: $\rightarrow AIC_{10}$

123 23
1 int
2
3
12
13

$2^3 = 8$

e.g. If $AIC_5 = \min_{k=1\cdots10} AIC_k$

$\Longrightarrow$ select Model 5

NEW PAGE | TEMPLATES | PREV. | Page 22 of 23 | NEXT | PC MODE | POWERPOINT | BALL PEN | HIGH LIGHT | SIZE | COLOR | LINE | RECT. | CIRCLE | SELECT ERASE | ALL ERASE | REC | Recording 00:00:00 | SAVE | OPEN | HELP | EXIT

Exam

- Ch 1-8 & Ch 10
- 1 page of A4 size (double sided)
- 50 point from Ch 4-10 from MC game
  short question
- 30 points numerical calculation $\overset{\text{data}}{(\vdots\vdots\vdots)}$ $(X'X)^{-1}$ are given
  eg $\hat{\beta}$, CI for β_i, CE for β, prediction interval ...
- 10 point $Cov(\hat{e}, \hat{Y})$ E(Sum of Sq)

$\boxed{\text{10 point "new"}}$