Chapter 10

Variable Selection

10.1. The active terms

- Variable selection
  - Aim: Identify the correct model
    - select the useful predictor
    - Ignore the non-informative terms
  - Y v.s. $X_1, X_2, \ldots X_{999}$
    - 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

10.1. Active terms and multicollinearity

- 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

10.1. Active terms and multicollinearity

- Remarks
  - Multicollinearity
    - some terms can be approximated by linear combination of the other terms.
    - (X'X) is close to non-singular. Inverse exists but unstable
    - e.g. $X_3 \approx c_0+c_1X_1+c_2X_2+c_4X_4+c_5X_5$
  - Perfect/Exact multicollinearity or Aliased
    - some term is exactly expressed by linear combination of the other terms.
    - (X'X) is singular. Inverse does not exist.
    - e.g. $X_3 = c_0+c_1X_1+c_2X_2+c_4X_4+c_5X_5$

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 <u>approximated</u> by linear combination of the other terms.

10.1. Active terms and multicollinearity

- Relationship between $Var(\hat{\beta}_j)$ and $\mathrm{R}_{\mathrm{j}}^2$
  - Using the idea of Added Variable Plot

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

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

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

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

$$Var(\hat{\beta}_j) = \sigma^2\left(X'_j(I - H_O)X_j\right)^{-1}$$

  - For the regression

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

$$X_j = X_Oc_O + e,$$

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

10.1. Variance Inflation Factor (VIF)

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

    $$\boxed{\text{Compare to } \quad Var(\hat{\beta}_j) = \frac{\sigma^2}{SX_j X_j}}$$

  - 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

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))
```

10.2. Automatic Variable Selection procedure

- For all possible candidate models, we compute
  - Akaike Information Criteria (AIC)

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

Number of parameters in the model

e.g. p+1 in regression

  - Bayesian Information Criteria (BIC)

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

  - Mallow's C_p Statistics

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

  - Predicted residual sum of Square (PRESS)

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

10.2. Automatic Variable Selection procedure

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

- 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

10.2. Automatic Variable Selection procedure

- AIC — BIC — C_p — PRESS

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

- What to do in practice?
  - Find the 4 best models according to each criteria
  - Usually they yield similar results, otherwise….
    - need to incorporate subject knowledge

10.2. Model Selection in Practice

- When you have Y and X_1, X_2, … X_p
  - For each possible model (e.g. $y=\beta_o+\beta_1 x_1$, $y=\beta_o+\beta_1 x_1+\ldots+\beta_p x_p$, $y=\beta_{p-1} x_{p-1}+\beta_p x_p$ 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
      - If a regression takes 1s , how long does it take for model selection?
  - Solutions
    - Forward Selection
    - Backward Selection

10.2. Model Selection in Practice

- Forward Selection
  - Set-up
    - K=Total number of terms may be added
    - i=1, L=1 (current number of terms, first start with intercept only)
    - V_i (current criterion values)
  - Step i
    - Each of the time, add one term beyond the current model
    - Obtain K+1-L criterion values
    - Stop if
      - All terms are included
      - All K+1-L criterion values are greater than V_i, i.e. additional terms does not improve the fitting.
    - Set V_{i+1} be the minimum of the K+1-L criterion values. Add the corresponding term to the current model, go to Step i+1

10.2. Model Selection in Practice

- Forward Selection
  - Each time add 1 variable, until no improvement
- Example
  - #Data
  - set.seed(1);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)
  - n=10; V_0= n*log(sum(lm(y~1)$residuals^2)/n)+2*1=35.49
  - # Step 1: (current model y= β_o, V_0=35.49)
    - AIC1a=n*log(sum(lm(y~x1)$residuals^2)/n)+2*(1+1)
    - AIC1b=n*log(sum(lm(y~x2)$residuals^2)/n)+2*(1+1)
    - AIC1c=n*log(sum(lm(y~x3)$residuals^2)/n)+2*(1+1)
    - AIC1d=n*log(sum(lm(y~x4)$residuals^2)/n)+2*(1+1)
    - print(c(AIC1a, AIC1b, AIC1c, AIC1d))
    - V_1=29.30, add x2,

AIC

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

10.2. Model Selection in Practice

- Forward Selection
  - Each time add 1 variable, until no improvement
- Example
  - # Step 2: (current model $y= \beta_o + \beta_2 X_2$, V_1=29.30)
    - AIC2a=n*log(sum(lm(y~x2+x1)$residuals^2)/n)+2*(2+1)
    - AIC2b=n*log(sum(lm(y~x2+x3)$residuals^2)/n)+2*(2+1)
    - AIC2c=n*log(sum(lm(y~x2+x4)$residuals^2)/n)+2*(2+1)
    - print(c(AIC2a, AIC2b, AIC2c))
    - V_2=13.85, Add x4
  - # Step 3: (current model $y= \beta_o + \beta_2 X_2 + \beta_4 X_4$, V_2=13.85)
    - AIC3a=n*log(sum(lm(y~x2+x4+x1)$residuals^2)/n)+2*(3+1)
    - AIC3b=n*log(sum(lm(y~x2+x4+x3)$residuals^2)/n)+2*(3+1)
    - print(c(AIC3a, AIC3b))
    - V_3= -9.30, Add x1

10.2. Model Selection in Practice

- Forward Selection
  - Each time add 1 variable, until no improvement
- Example
  - # Step 4: (current model $y= \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$, $V_3 = -9.30$)
    - AIC4a=n*log(sum(lm(y~x2+x4+x1+x3)$residuals^2)/n)+2*(4+1)
    - print(c(AIC4a))
    - $V_4 = -8.56 > -9.30 = V_3$. STOP, without adding x3
    - Conclusion
      - The model $y= \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$ is the best according to AIC using Forward Selection.

10.2. Model Selection in Practice

- Backward Selection
  - Set-up
    - K=Total number of terms may be added
    - i=1, L=K+1 (current number of terms, first fit all terms)
    - V_i (current criterion values)
  - Step i
    - Each of the time, remove one term from the current model
    - Obtain L-1 criterion values
    - Stop if
      - All terms are removed
      - All L-1 criterion values are greater than V_i, i.e. removing terms worsen the fitting.
    - Set V_{i+1} be the minimum of the L-1 criterion values. Remove the corresponding term from the current model, go to Step i+1

10.2. Model Selection in Practice

- Backward Selection
  - Each time delete 1 variable, until no improvement
- Example
  - #Data
  - set.seed(2);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)
  - n=10; V_0= n*log(sum(lm(y~x1+x2+x3+x4)$residuals^2)/n)+2*(4+1)=-4.48
  - # Step 1: (current model $y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, $V_0 = -4.48$)
    - AIC1a=n*log(sum(lm(y~x2+x3+x4)$residuals^2)/n)+2*(3+1)
    - AIC1b=n*log(sum(lm(y~x1+x3+x4)$residuals^2)/n)+2*(3+1)
    - AIC1c=n*log(sum(lm(y~x1+x2+x4)$residuals^2)/n)+2*(3+1)
    - AIC1d=n*log(sum(lm(y~x1+x2+x3)$residuals^2)/n)+2*(3+1)
    - print(c(AIC1a, AIC1b, AIC1c, AIC1d))
    - V_1= -6.46, remove x1,

10.2. Model Selection in Practice

- Backward Selection
  - Each time delete 1 variable, until no improvement
- Example
  - # Step 2: (current model $y= \beta_o + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, $V_0 = -6.46$)
    - AIC2a=n*log(sum(lm(y~x3+x4)$residuals^2)/n)+2*(2+1)
    - AIC2b=n*log(sum(lm(y~x2+x4)$residuals^2)/n)+2*(2+1)
    - AIC2c=n*log(sum(lm(y~x2+x3)$residuals^2)/n)+2*(2+1)
    - print(c(AIC2a, AIC2b, AIC2c))
    - V_2=16.10> -6.46 =V_1. STOP, without removing any variable
    - Conclusion
      - The model $y= \beta_o+\beta_2 x_2+\beta_3 x_3 +\beta_4 x_4$ is the best according to AIC using Backward Selection.

What you have learnt

- Regression: Relationship b/w variables
  - $Y=X\beta+e$
    - Least Sq Est β, Test if $\beta=0$, C.I. for β, Predict Y
    - Compare between models by F-test
  - Diagnostic check – Residual/scatter/AV plot
  - Improvement
    - Non-constant variance -- WLS
    - Curve relationship -- 1) Poly Reg 2) Transform
    - Check Outlier – 1) T-test 2) Cook's distance
  - Variable selection: 1)AIC. 2)BIC. 3)C_p. 4)Press