Lecture 12

Model Selection and Diagnostics

STAT 8020 Statistical Methods II

September 16, 2019

Whitney Huang

Clemson University

Agenda

1. **Automatic Search Procedures**
2. **Variable Selection Criteria**
3. **Diagnostics in Multiple Linear Regression (MLR)**

Variable Selection

- What is the appropriate subset size?
- What is the best model for a fixed size?

Automatic Search Procedures

- Forward Selection
- Backward Elimination
- Stepwise Search
- All Subset Selection

Mallows' C_p Criterion

$$
\begin{aligned}
(\hat{Y}_i - \mu_i)^2 &= (\hat{Y}_i - \mathrm{E}(\hat{Y}_i) + \mathrm{E}(\hat{Y}_i) - \mu_i)^2 \\
&= \underbrace{(\hat{Y}_i - \mathrm{E}(\hat{Y}_i))^2}_{\text{Variance}} + \underbrace{(\mathrm{E}(\hat{Y}_i) - \mu_i)^2}_{\text{Bias}^2},
\end{aligned}
$$

where $\mu_i = \mathrm{E}(Y_i | X_i = x_i)$

- Mean squared prediction error (MSPE): $\sum_{i=1}^{n} \sigma^2_{\hat{Y}_i} + \sum_{i=1}^{n} (\mathrm{E}(\hat{Y}_i) - \mu_i)^2$
- C_p criterion measure:

$$
\begin{aligned}
\Gamma_p &= \frac{\sum_{i=1}^{n} \sigma^2_{\hat{Y}_i} + \sum_{i=1}^{n} (\mathrm{E}(\hat{Y}_i) - \mu_i)^2}{\sigma^2} \\
&= \frac{\sum \text{Var}_{\text{pred}} + \sum \text{Bias}^2}{\text{Var}_{\text{error}}}
\end{aligned}
$$

C_p Criterion

- Do not know σ^2 nor numerator
- Use $\text{MSE}_{X_1,\cdots,X_{p-1}} = \text{MSE}_\text{F}$ as the estimate for σ
- For numerator:
  - Can show $\sum_{i=1}^n \sigma^2_{\hat{Y}_i} = p\sigma^2$
  - Can also show $\sum_{i=1}^n (\mathrm{E}(\hat{Y}_i) - \mu_i)^2 = \mathrm{E}(\text{SSE}_\text{F}) - (n-p)\sigma^2$

$$\Rightarrow C_p = \frac{\text{SSE} - (n-p)\text{MSE}_\text{F} + p\text{MSE}_\text{F}}{\text{MSE}_\text{F}}$$

C_p Criterion Cont'd

Recall

$$\Gamma_p = \frac{\sum_{i=1}^{n} \sigma^2_{\hat{Y}_i} + \sum_{i=1}^{n} (\mathrm{E}(\hat{Y}_i) - \mu_i)^2}{\sigma^2}$$

- When model is correct $\mathrm{E}(C_p) \approx p$
- When plotting models against p
  - Biased models will fall above $C_p = p$
  - Unbiased models will fall around line $C_p = p$
  - By definition: C_p for full model equals p

Adjusted R^2 Criterion

Adjusted R^2, denoted by R^2_{adj}, attempts to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model.

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

- Choose model which maximizes R^2_{adj}
- Same approach as choosing model with smallest MSE

Predicted Residual Sum of Squares *PRESS* Criterion

- For each observation i, predict Y_i using model generated from other $n-1$ observations
- $PRESS = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$
- Want to select model with small $PRESS$

Other Approaches

- Akaike's information criterion (AIC)

$$n \log(\frac{\text{SSE}_k}{n}) + 2k$$

- Bayesian information criterion (BIC)

$$n \log(\frac{\text{SSE}_k}{n}) + k \log(n)$$

- Can be used to compare **non-nested** models

Residuals and Leverage

Recall in MLR that $\hat{Y} = X(X^TX)^{-1}X^TY = HY$ where H is the hat-matrix

- Can show that $Var(e) = (I - H)\sigma^2$. Therefore $Var(e_i) = \sigma^2(1 - h_i)$, where $h_i = H_{ii}$ are called **leverages**
- $\sum_{i=1}^{n} h_i = p$ and $h_i > \frac{1}{n}, 1 \leq i \leq n \Rightarrow$ a “rule of thumb" is that leverages of more than $\frac{2p}{n}$ should be looked at more closely
- $Var(\hat{Y}) = H\sigma^2 \Rightarrow \hat{Var}\hat{Y}_i = h_i\hat{\sigma}^2$

Studentized Residuals

As we have seen $Var(e_i) = \sigma^2(1 - h_i)$, this suggests the use of $r_i = \frac{e_i}{\hat{\sigma}(1-h_i)}$

- r_i's are called **studentized residuals**
- If the model assumptions are correct then $Var(r_i) = 1$ and $Corr(e_i, e_j)$ tends to be small

Influential Observations

DFFITS

- Difference between the fitted values $\hat{Y}_i$ and the predicted values $\hat{Y}_{i(i)}$
- $\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_i}}$
- Concern if absolute value greater than 1 for small data sets, or greater than $2\sqrt{p/n}$ for large data sets