

Lecture 4

Multiple Linear Regression: Model Selection and Model Checking

Reading: Faraway (2014) Chapters 6 and 10 or Faraway
(2002) Chapters 7 and 10

DSA 8020 Statistical Methods II
January 25-29, 2021

Whitney Huang
Clemson University

Agenda

Multiple Linear
Regression: Model
Selection and Model
Checking



Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

- 1 **Model Selection**
- 2 **Model Diagnostics**
- 3 **Non-Constant Variance & Transformation**

Multiple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

Basic Problem: how to choose between competing linear regression models?

- **Model too “small”:** underfit the data; poor predictions; high bias; low variance
- **Model too big:** “overfit” the data; poor predictions; low bias; high variance

In the next few slides we will discuss some commonly used model selection criteria to choose the “right” model

Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Mallows' C_p Criterion

A good model should balance bias and variance to get good predictions

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

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

- Mean squared prediction error (MSPE):

$$\sum_{i=1}^n \sigma_{\hat{Y}_i}^2 + \sum_{i=1}^n (E(\hat{Y}_i) - \mu_i)^2$$

- C_p criterion measure:

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

C_p statistic:

$$C_p = \frac{\text{SSE}}{\text{MSE}_F} + 2p - n$$

- When model is correct $E(C_p) \approx p$
- 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

We desire models with small p and C_p around or less than p

Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Adjusted R^2 Criterion

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

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

- Choose model which maximizes R_{adj}^2
- 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
- $PRESS = \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$
- PRESS statistic is a form of **cross-validation**
- Want to select model with small *PRESS*

There are two widely used information criteria:

- Akaike's information criterion (AIC)

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

- Bayesian information criterion (BIC)

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

- **Forward Selection:** begins with no predictors and then adds in predictors one by one using some criterion (e.g., p-value or AIC)
- **Backward Elimination:** starts with all the predictors and then removes predictors one by one using some criterion
- **Stepwise Search:** a combination of backward elimination and forward selection. Can add or delete predictor at each stage
- **All Subset Selection:** Comparing all possible models using a selected criterion. Impractical for “large” number of predictors

Model:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

We make the following **assumptions**:

- Linearity:

$$E(Y|x_1, x_2, \dots, x_{p-1}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Model:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

We make the following **assumptions**:

- Linearity:

$$E(Y|x_1, x_2, \dots, x_{p-1}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

- Errors have constant variance, are independent, and normally distributed

$$\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

*All models are wrong
but some are useful*



George E.P. Box

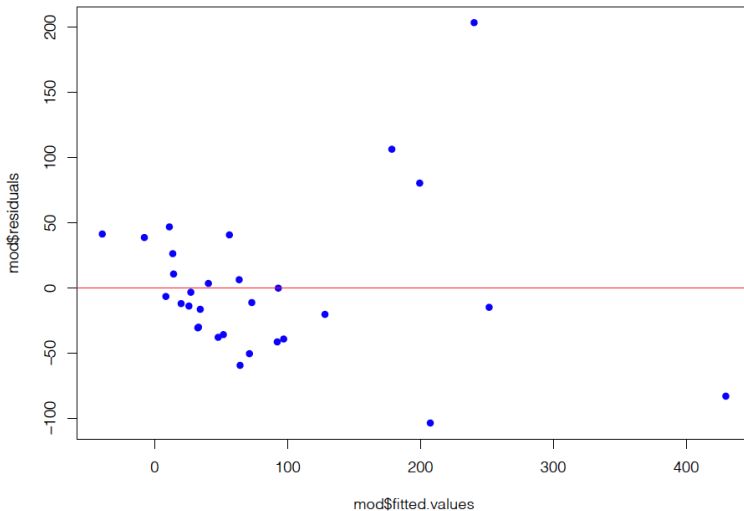
Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Residuals versus Fits Plot

```
plot(mod$fitted.values, mod$residuals, pch = 16, col = "blue")  
abline(h = 0, col = "red")
```

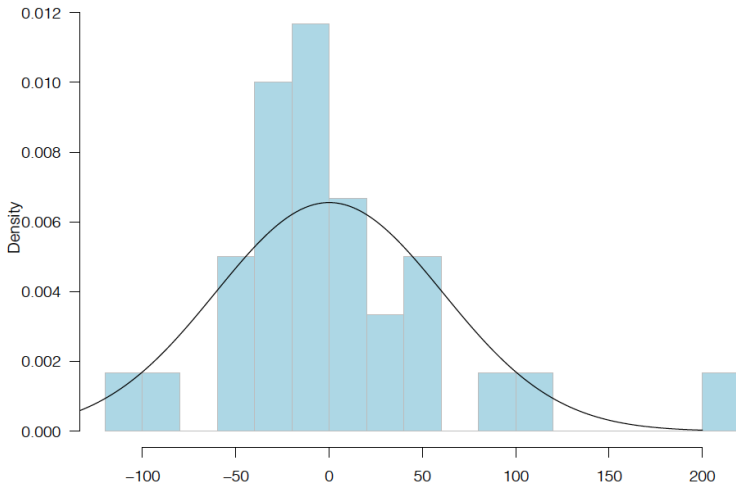


We will revisit this in the end of the lecture

Assessing Normality of Residuals: Histogram

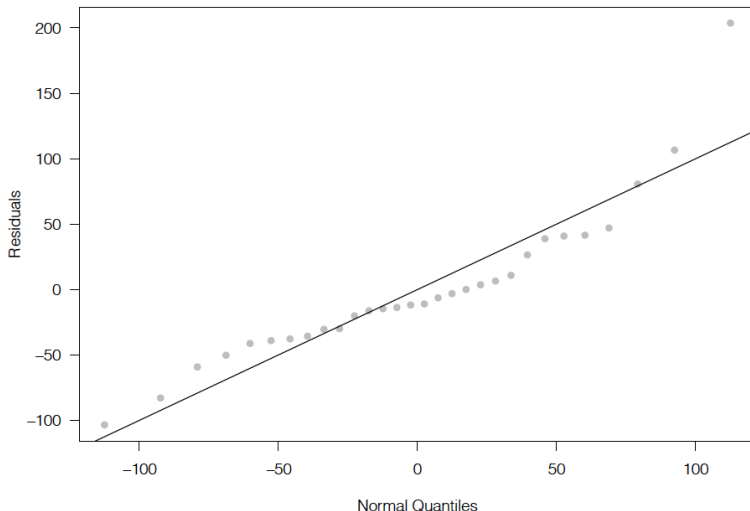
```
par(las = 1)
hist(mod$residuals, 12, prob = T,
     col = "lightblue", border = "gray")
xg <- seq(-200, 200, 1)
yg <- dnorm(xg, 0, 60.86)
lines(xg, yg)
```

Histogram of mod\$residuals



Assessing Normality of Residuals: QQ Plot

```
plot(qnorm(1:30 / 31, 0, 60.86), sort(mod$residuals), pch = 16,  
     col = "gray", xlab = "Normal Quantiles", ylab = "Residuals")  
abline(0, 1)
```



Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Leverage: Detecting “Extreme” Predictor Values

Recall in MLR that $\hat{y} = X(X^T X)^{-1} X^T y = H y$ where H is the hat-matrix

- The leverage value for the i_{th} observation is defined as:

$$h_i = H_{ii}$$

- Can show that $\text{Var}(e_i) = \sigma^2(1 - h_i)$, where $e_i = y_i - \hat{y}_i$ is the residual for the i_{th} observation
- $\frac{1}{n} \leq h_i \leq 1$, $1 \leq i \leq n$ and $\bar{h} = \sum_{i=1}^n \frac{h_i}{n} = \frac{p}{n} \Rightarrow$ a “rule of thumb” is that leverages of more than $\frac{2p}{n}$ should be looked at more closely

Leverage Values of Species ~ Elev + Adj

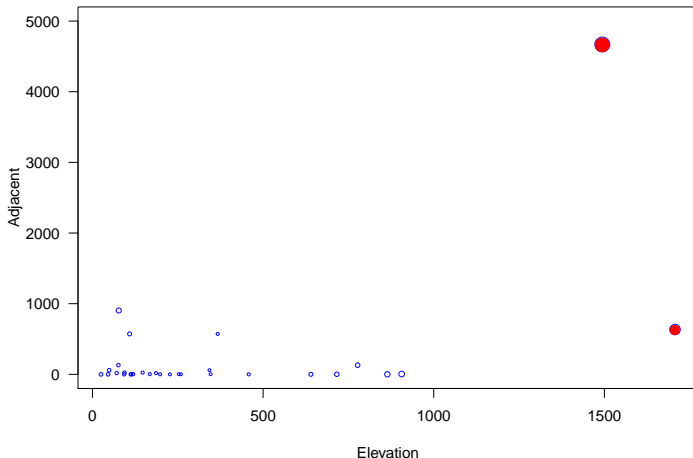
Multiple Linear
Regression: Model
Selection and Model
Checking



Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation



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

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1-h_i)}}$$

- r_i 's are called **studentized residuals**. r_i 's are sometimes preferred in residual plots as they have been standardized to have equal variance.
- If the model assumptions are correct then $\text{Var}(r_i) = 1$ and $\text{Corr}(e_i, e_j)$ tends to be small

Studentized Residuals of Species \sim Elev + Adj

Multiple Linear
Regression: Model
Selection and Model
Checking

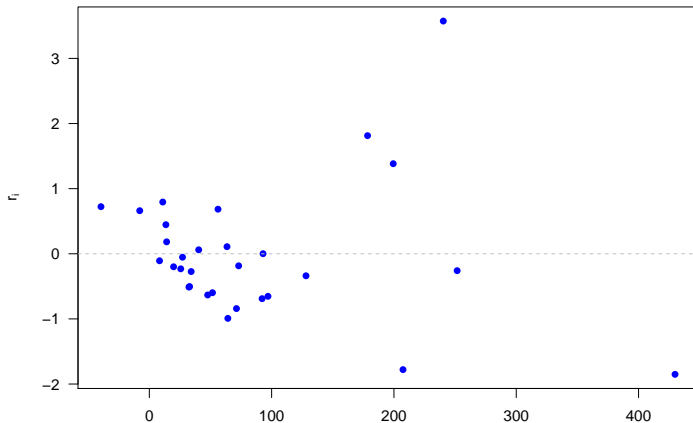


Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Studentized Residuals



- For a given model, exclude the observation i and recompute $\hat{\beta}_{(i)}$, $\hat{\sigma}_{(i)}$ to obtain $\hat{y}_{i(i)}$
- The observation i is an outlier if $\hat{y}_{i(i)} - y_i$ is “large”
- Can show
$$\text{Var}(\hat{y}_{i(i)} - y_i) = \sigma_{(i)}^2 \left(1 + \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i \right) = \frac{\sigma_{(i)}^2}{1 - h_i}$$
- Define the **Studentized Deleted Residuals** as

$$t_i = \frac{\hat{y}_{i(i)} - y_i}{\hat{\sigma}_{(i)}^2 / (1 - h_i)} = \frac{\hat{y}_{i(i)} - y_i}{\text{MSE}_{(i)} (1 - h_i)^{-1}}$$

which are distributed as a t_{n-p-1} if the model is correct and $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$

Jackknife Residuals of Species ~ Elev + Adj

Multiple Linear
Regression: Model
Selection and Model
Checking

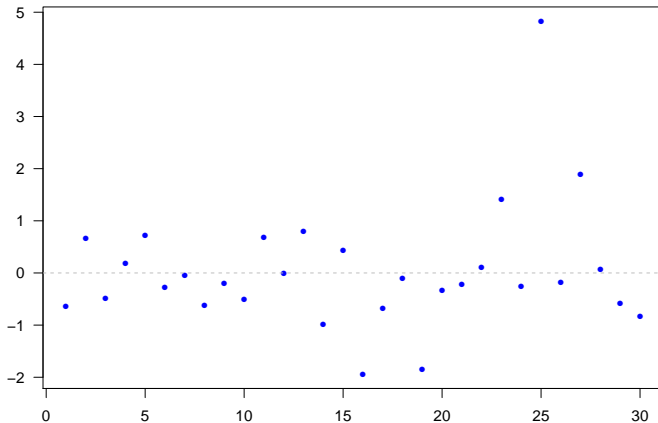


Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Jackknife Residuals



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

DFFITS of Species ~ Elev + Adj

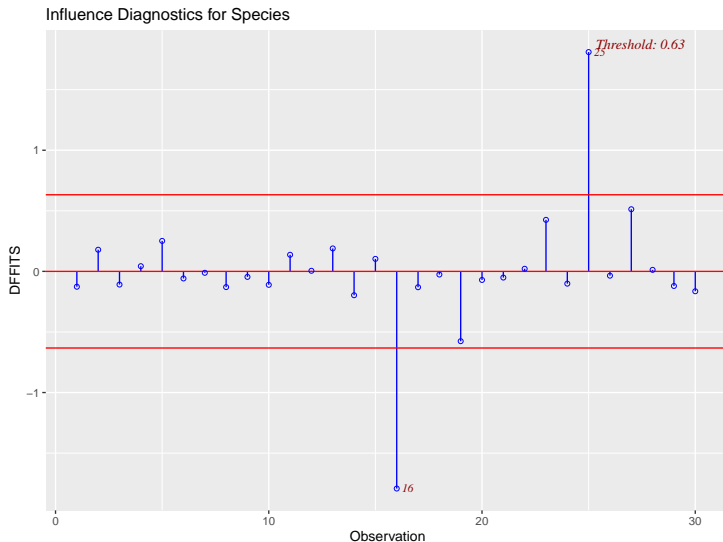
Multiple Linear
Regression: Model
Selection and Model
Checking

CLEMSON
UNIVERSITY

Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation



Residual Plot of Species ~ Elev + Adj

Multiple Linear
Regression: Model
Selection and Model
Checking

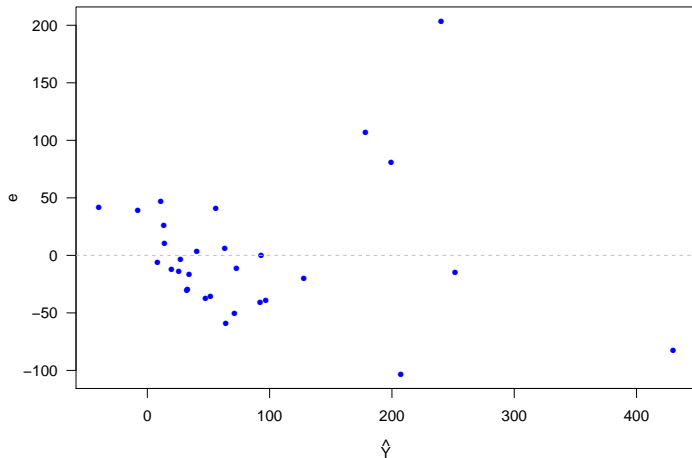
CLEMSON
UNIVERSITY

Model Selection

Model Diagnostics

Non-Constant
Variance &
Transformation

Residuals



Residual Plot After Square Root Transformation

Residuals

