## Lecture 13

## **Model Diagnostics**

STAT 8020 Statistical Methods II September 18, 2019



Leverag

Studentized & Jackknife Residuals

DFFITS

Non-Constant Variance & Transformation

Whitney Huang Clemson University

## **Agenda**

## Model Diagnostics



Leverage

Jackknife Residuals

FFITS

on-Constant ariance &

- Leverage
- Studentized & Jackknife Residuals
- **3 DFFITS**
- 4 Non-Constant Variance & Transformation

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

#### Model Diagnostics



#### Leverage

itudentized & ackknife Residuals

#### DFFITS

on-Constant ariance & ransformation

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

• The leverage score for the  $i_{th}$  observation is defined as:

$$h_i = \boldsymbol{H}_{ii}$$

## **Model Diagnostics**



#### Leverage

Jackknife Residuals

JEFIIS

on-Constant ariance & ansformation

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

• The leverage score for the  $i_{th}$  observation is defined as:

$$h_i = \boldsymbol{H}_{ii}$$

## **Model Diagnostics**



#### Leverage

Jackknife Residuals

JEFIIS

on-Constant ariance & ansformation

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

• The leverage score for the  $i_{th}$  observation is defined as:

$$h_i = \boldsymbol{H}_{ii}$$

• 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  $i_{\text{th}}$  observation





#### Leverage

ackknife Residuals

DEFIIS

lon-Constant ariance & ransformation

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

• The leverage score for the  $i_{th}$  observation is defined as:

$$h_i = \boldsymbol{H}_{ii}$$

• 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  $i_{\text{th}}$  observation





#### Leverage

ackknife Residuals

DEFIIS

lon-Constant ariance & ransformation

# Model Diagnostics CLEMSEN UNIVERSITY

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

• The leverage score for the  $i_{th}$  observation is defined as:

$$h_i = \boldsymbol{H}_{ii}$$

- 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  $i_{\text{th}}$  observation
- $\frac{1}{n} \leq h_i \leq 1$ ,  $1 \leq i \leq n$  and  $\bar{h}_i = \frac{p}{n} \Rightarrow$  a "rule of thumb" is that leverages of more than  $\frac{2p}{n}$  should be looked at more closely

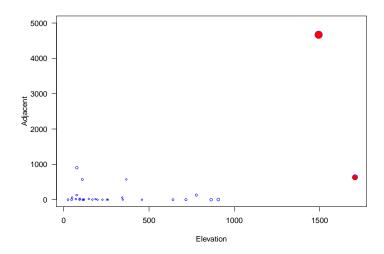
#### everage

ckknife Residuals

DEFIIS

Von-Constant Variance & Transformation

## Leverage Scores of Species $\sim \mathtt{Elev} + \mathtt{Adj}$



#### **Model Diagnostics**



#### Leverage

Studentized &

#### DEFITS

lon-Constant /ariance & fransformation

## Studentized Residuals

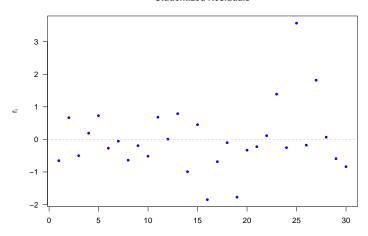
DFFITS

Non-Constant
Variance &
Fransformation

- As we have seen  ${\rm Var}(e_i)=\sigma^2(1-h_i)$ , this suggests the use of  $r_i=rac{e_i}{\hat{\sigma}\sqrt{(1-h_i)}}$ 
  - r<sub>i</sub>'s are called **studentized residuals**. r<sub>i</sub>'s are sometimes preferred in residual plots as they have been standardized to have equal variance.
  - $\bullet$  If the model assumptions are correct then  ${\rm Var}(r_i)=1$  and  ${\rm Corr}(e_i,e_j)$  tends to be small

## Studentized Residuals of Species $\sim$ Elev + Adj

#### Studentized Residuals



#### Model Diagnostics



Leverage

Studentized &

DEFITS

on-Constant ariance &

## **Jackknife Residuals**

• For a given model, exclude the observation i and recompute  $\hat{\beta}_{(i)}$ ,  $\hat{\sigma}_{(i)}$  to obtain  $\hat{Y}_{i(i)}$ 

Model Diagnostics



Leverage

Studentized & Jackknife Residuals

DFFITS

Non-Constant Variance & Transformation

## **Jackknife Residuals**

• For a given model, exclude the observation i and recompute  $\hat{\beta}_{(i)}$ ,  $\hat{\sigma}_{(i)}$  to obtain  $\hat{Y}_{i(i)}$ 

Model Diagnostics



Leverage

Studentized &

DFFITS

Non-Constant Variance & Transformation

DFFITS

Non-Constant /ariance & Fransformation

• 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"

DFFITS

Non-Constant /ariance & Fransformation

• 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"

on-Constant ariance & ransformation

- 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"
- $\bullet \ \, \mathsf{Can} \ \, \mathsf{show} \ \, \mathsf{Var}(\hat{Y}_{i(i)} Y_i) = \sigma^2 \left( 1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_i)^{-1} \boldsymbol{x}_i \right)$

on-Constant ariance & ransformation

- 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"
- $\bullet \ \, \mathsf{Can} \ \, \mathsf{show} \ \, \mathsf{Var}(\hat{Y}_{i(i)} Y_i) = \sigma^2 \left( 1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_i)^{-1} \boldsymbol{x}_i \right)$

Variance & Fransformation

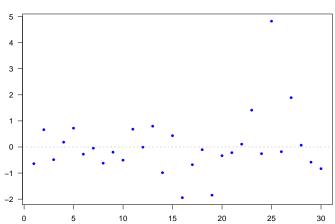
- For a given model, exclude the observation i and recompute  $\hat{\beta}_{(i)}$ ,  $\hat{\sigma}_{(i)}$  to obtain  $\hat{Y}_{i(i)}$
- The observation i is an outlier if  $\hat{Y}_{i(i)} Y_i$  is "large"
- $\bullet \ \, \mathsf{Can} \ \, \mathsf{show} \ \, \mathsf{Var}(\hat{Y}_{i(i)} Y_i) = \sigma^2 \left( 1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_i)^{-1} \boldsymbol{x}_i \right)$
- Define the jackknife residuals as

$$t_i = \frac{\hat{Y}_{i(i)} - Y_i}{\sqrt{\hat{\sigma}^2 \left(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_i)^{-1} \boldsymbol{x}_i\right)}}$$

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

## Jackknife Residuals of Species $\sim \mathtt{Elev} + \mathtt{Adj}$

### Jacknife Residuals



#### Model Diagnostics



Leverage

Studentized &

DEFITS

Non-Constant
Variance &

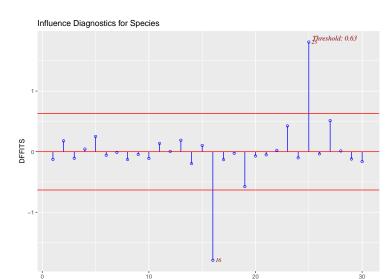
#### DFFITS

Non-Constant Variance & Fransformation

## **DFFITS**

- $\bullet$  Difference between the fitted values  $\hat{Y}_i$  and the predicted values  $\hat{Y}_{i(i)}$
- $\qquad \text{DFFITS}_i = \frac{\hat{Y}_i \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_i}}$
- 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** $\sim$ **Elev** + **Adj**



Observation

#### Model Diagnostics



#### Leverage

Studentized &

#### **DEFITS**

Non-Constant Variance & Transformation

## Residual Plot of Species $\sim$ Elev + Adj

## Model Diagnostics

### CLEMS#N UNIVERSITY

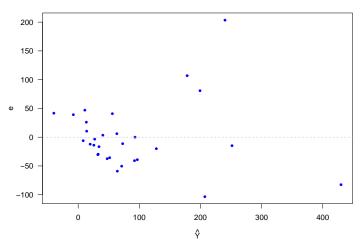
#### Leverage

tudentized &

#### **DEFITS**

Non-Constant Variance & Transformation

## Residuals



## **Residual Plot After Square Root Transformation**

# **Model Diagnostics**

Non-Constant

