Lecture 13
Model Diagnostics

STAT 8020 Statistical Methods II
September 18, 2019

Whitney Huang
Clemson University

Notes

Agenda

1. Leverage Values
2. Studentized & Studentized Deleted Residuals
3. DFFITS
4. Variance Inflation Factor
5. Non-Constant Variance & Transformation

Notes

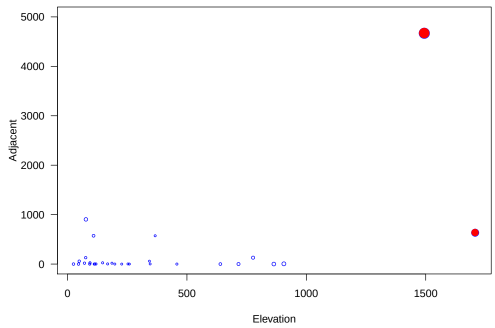
Leverage

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

- The leverage value for the i_{th} observation is defined as:
$$h_i = \boldsymbol{H}_{ii}$$
- Can show that Var$(e_i) = \sigma^2(1 - h_i)$, where $e_i = Y_i - \hat{Y}_i$ is the residual for the i_{th} observation
- $\frac{1}{n} \leq h_i \leq 1, \quad 1 \leq i \leq n$ 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

Notes

Leverage Values of `Species` $\sim$ `Elev` + `Adj`



Notes

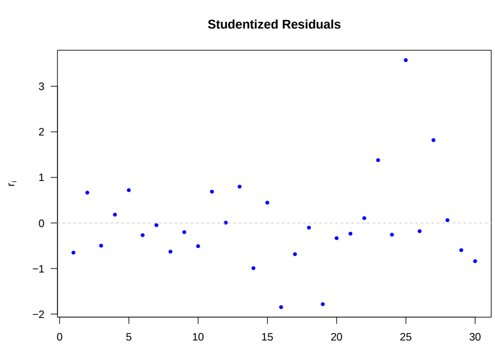
Studentized Residuals

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

- 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 Var$(r_i) = 1$ and Corr(e_i, e_j) tends to be small

Notes

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



Notes

Studentized Deleted Residuals

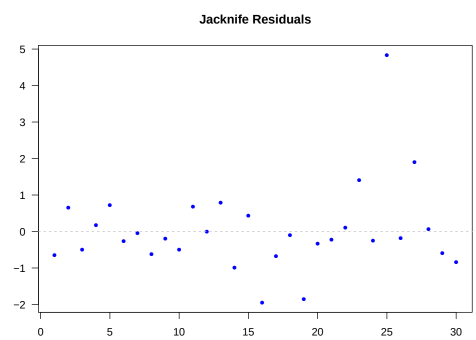
- For a given model, exclude the observation i and recompute $\hat{\beta}_{(i)}$, $\hat{\sigma}_{(i)}$ to obtain $\hat{Y}_{i(i)}$
- The observation i is an outlier if $\hat{Y}_{i(i)} - Y_i$ is "large"
- Can show $\text{Var}(\hat{Y}_{i(i)} - Y_i) = \sigma^2_{(i)} \left(1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T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

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

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

Notes

Jackknife Residuals of `Species` $\sim$ `Elev` $+$ `Adj`



Model Diagnostics
CLEMSON UNIVERSITY
Leverage Values
Studentized & Studentized Deleted Residuals
DFFITS
Variance Inflation Factor
Non-Constant Variance & Transformation

Notes

Influential Observations

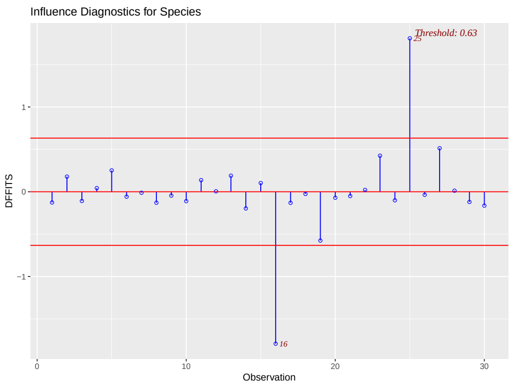
DFFITS

- 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

Model Diagnostics
CLEMSON UNIVERSITY
Leverage Values
Studentized & Studentized Deleted Residuals
DFFITS
Variance Inflation Factor
Non-Constant Variance & Transformation

Notes

DFFITS of `Species` ∼ `Elev` + `Adj`



Notes

Variance Inflation Factor (VIF)

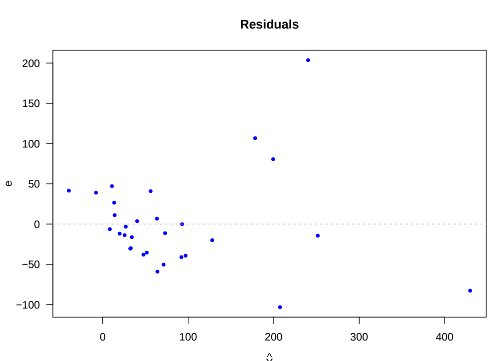
$$\text{VIF}_k = \frac{1}{1 - R_k^2},$$

where R_k^2 is the coefficient of determination when X_k is regressed on the remaining $p-2$ other predictors.

```
> vif(step_gala)
Elevation  Adjacent
 1.404074  1.404074
> vif(full)
     Area Elevation   Nearest
 2.928145  3.992545  1.766099
    Scruz  Adjacent
 1.675031  1.826403
```

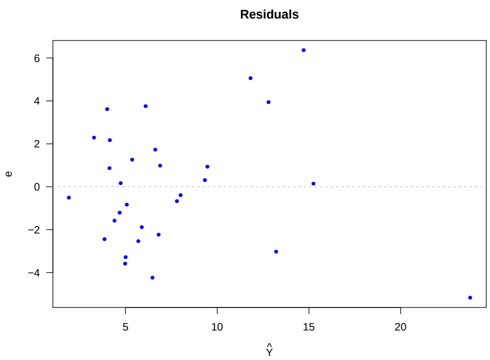

Notes

Residual Plot of `Species` ∼ `Elev` + `Adj`



Notes

Residual Plot After Square Root Transformation



Notes

Notes

Notes