Lecture 4

Multiple Linear Regression: Model Selection and Model Checking

Reading: Faraway 2014 Chapters 6, 9.1, and 10

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

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

Model Selection in Multiple Linear Regression

Multiple Linear Regression Model:

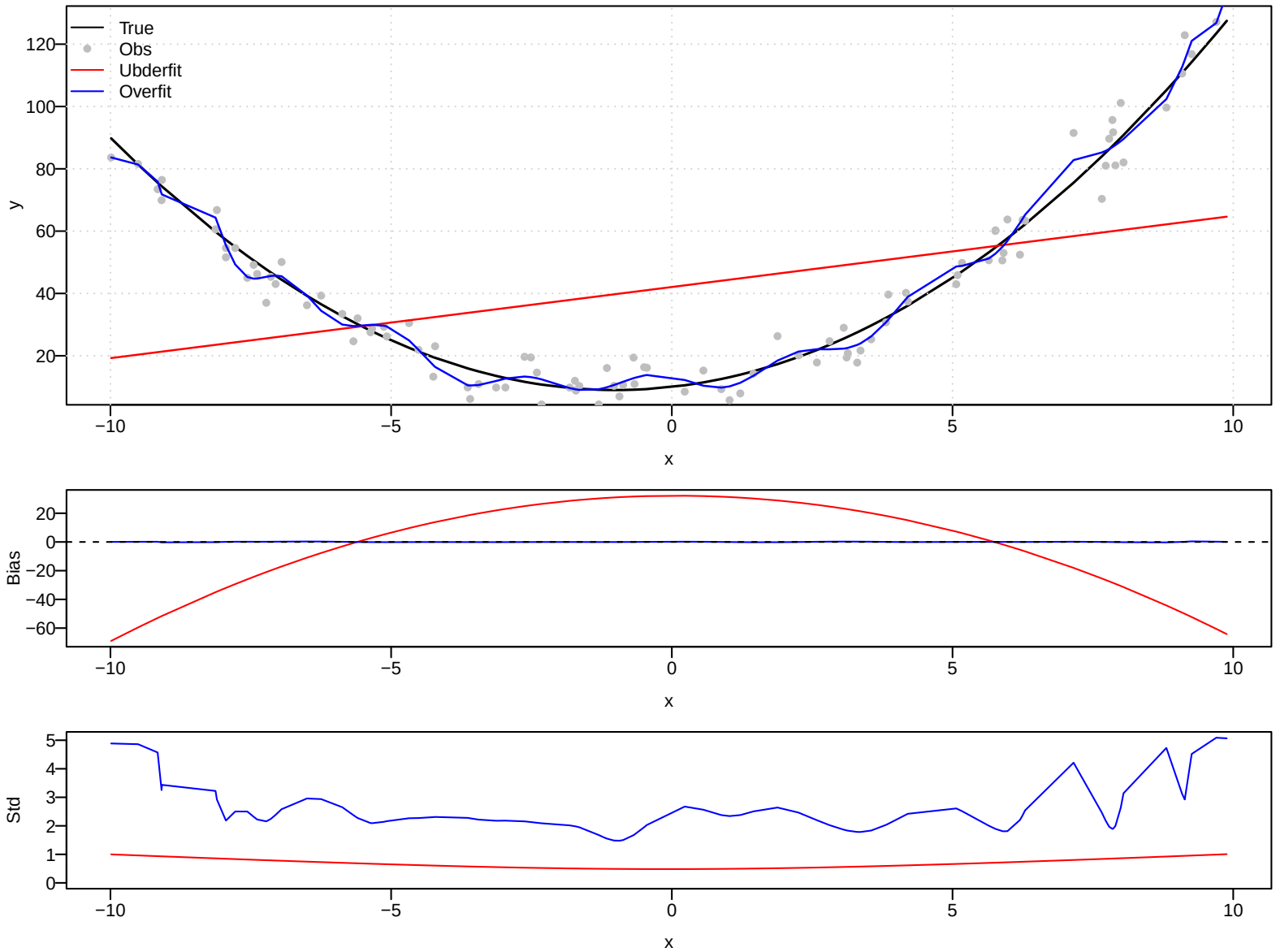
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

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 to balance bias and variance

An Example of Bias and Variance Tradeoff



Balancing Bias And Variance: Mallows' C_p Criterion

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

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

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

C_p statistic:

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

- When model is correct $\text{E}(C_p) \approx p$
- 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. See R session for an example

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

Information criteria

Information criteria are statistical measures used for model selection. Commonly used information criteria include:

- Akaike's information criterion (AIC)

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

- Bayesian information criterion (BIC)

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

Here k is the number of the parameters in the model.

These criteria balance the goodness of fit of a model with its complexity

Automatic Search Procedures

- 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 Assumptions

Model:

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

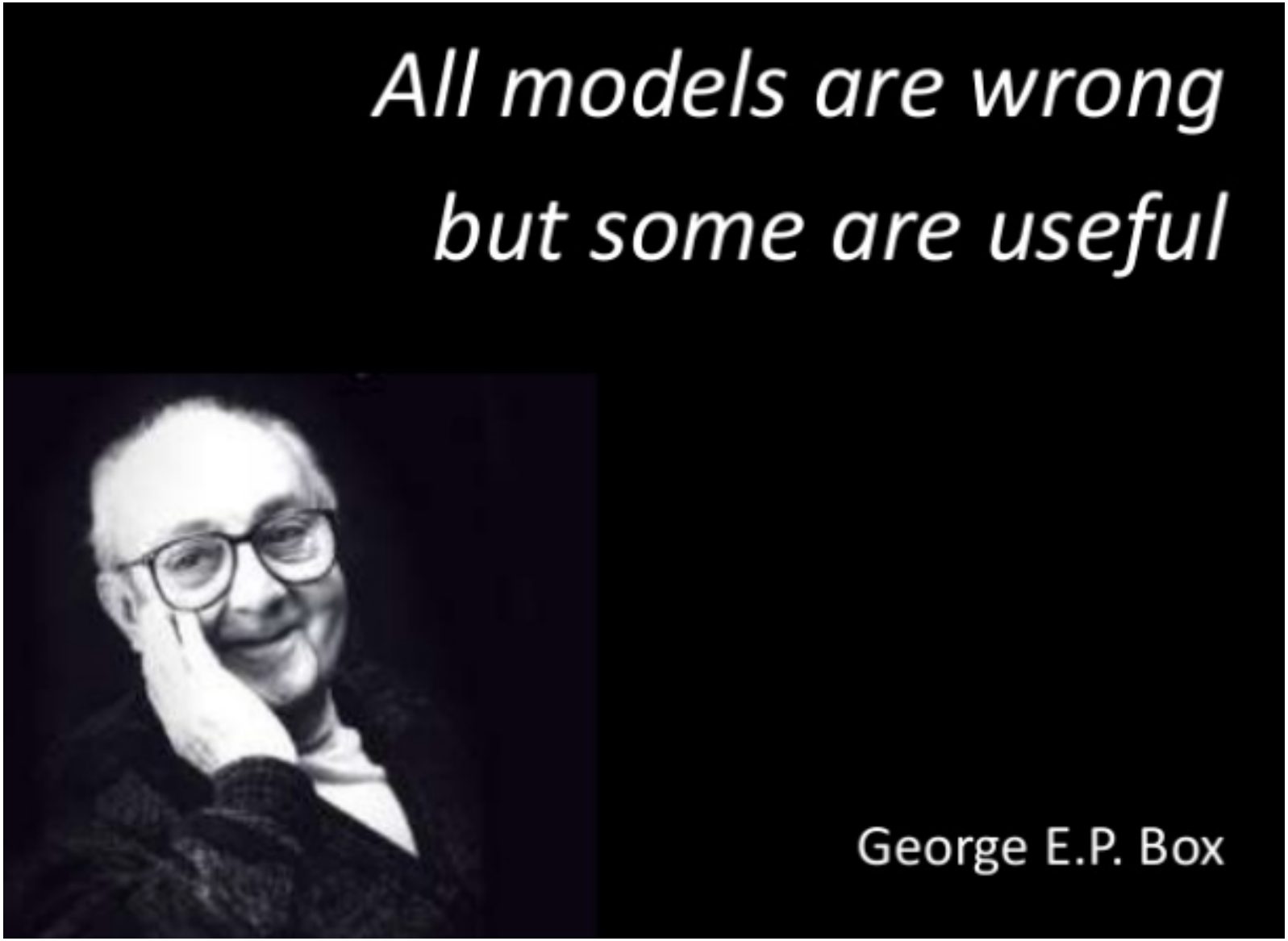
We make the following assumptions:

- Linearity:

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

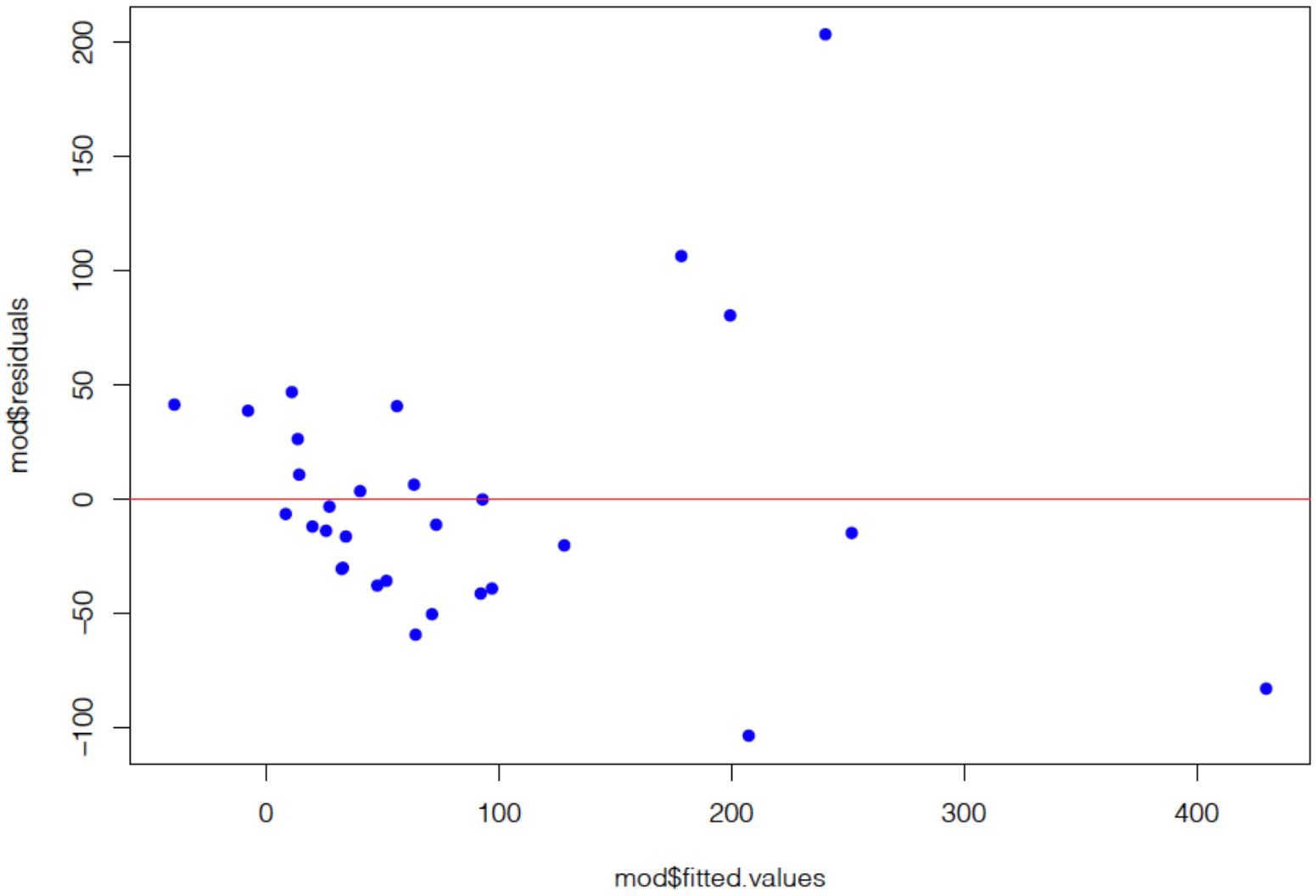
- Errors have constant variance, are independent, and normally distributed

$$\varepsilon \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$



Residuals versus Fits Plot

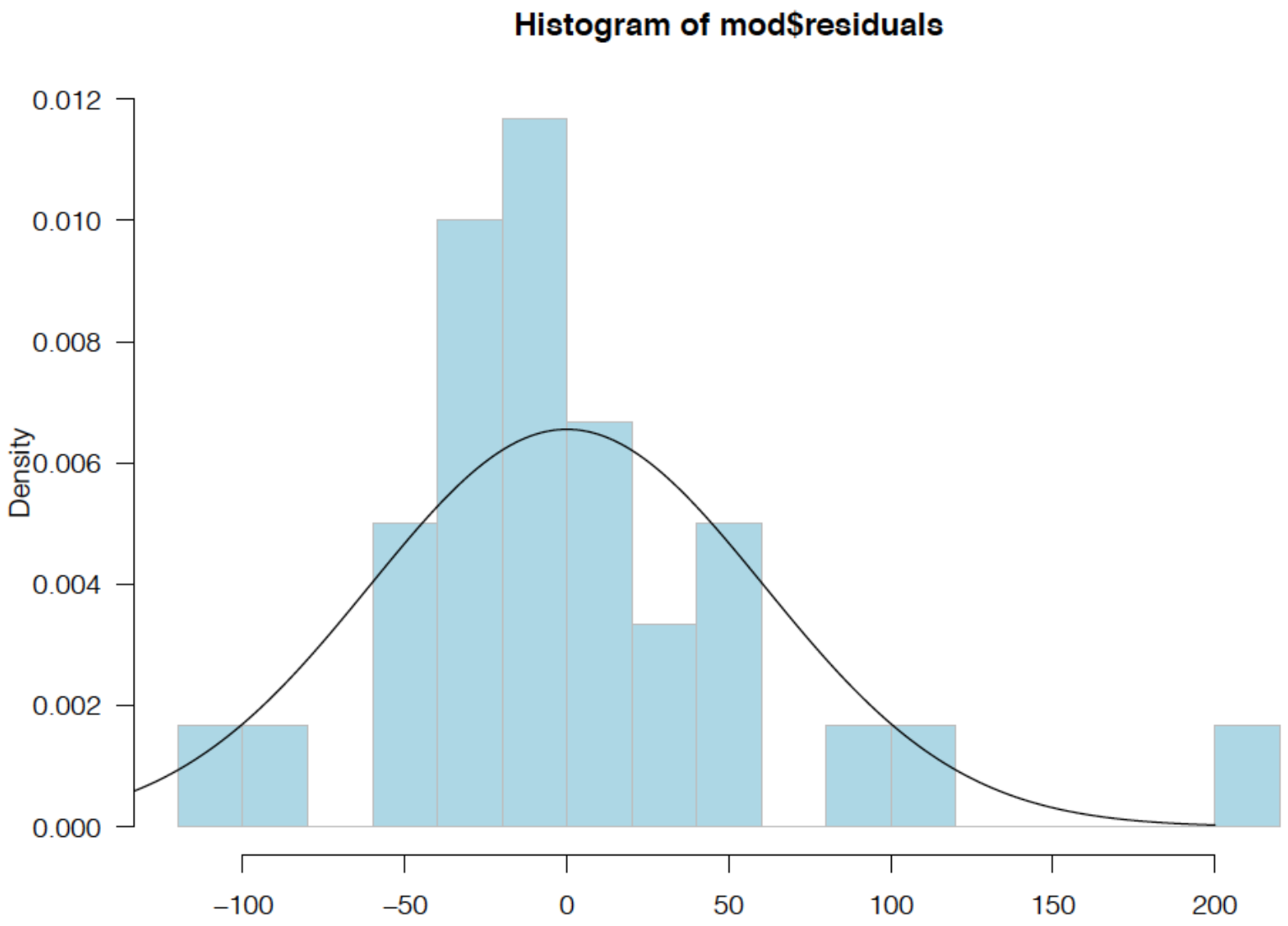
```
plot(mod$fitted.values, mod$residuals, pch = 16, col = "blue")
abline(h = 0, col = "red")
```



We will revisit this in the end of the lecture

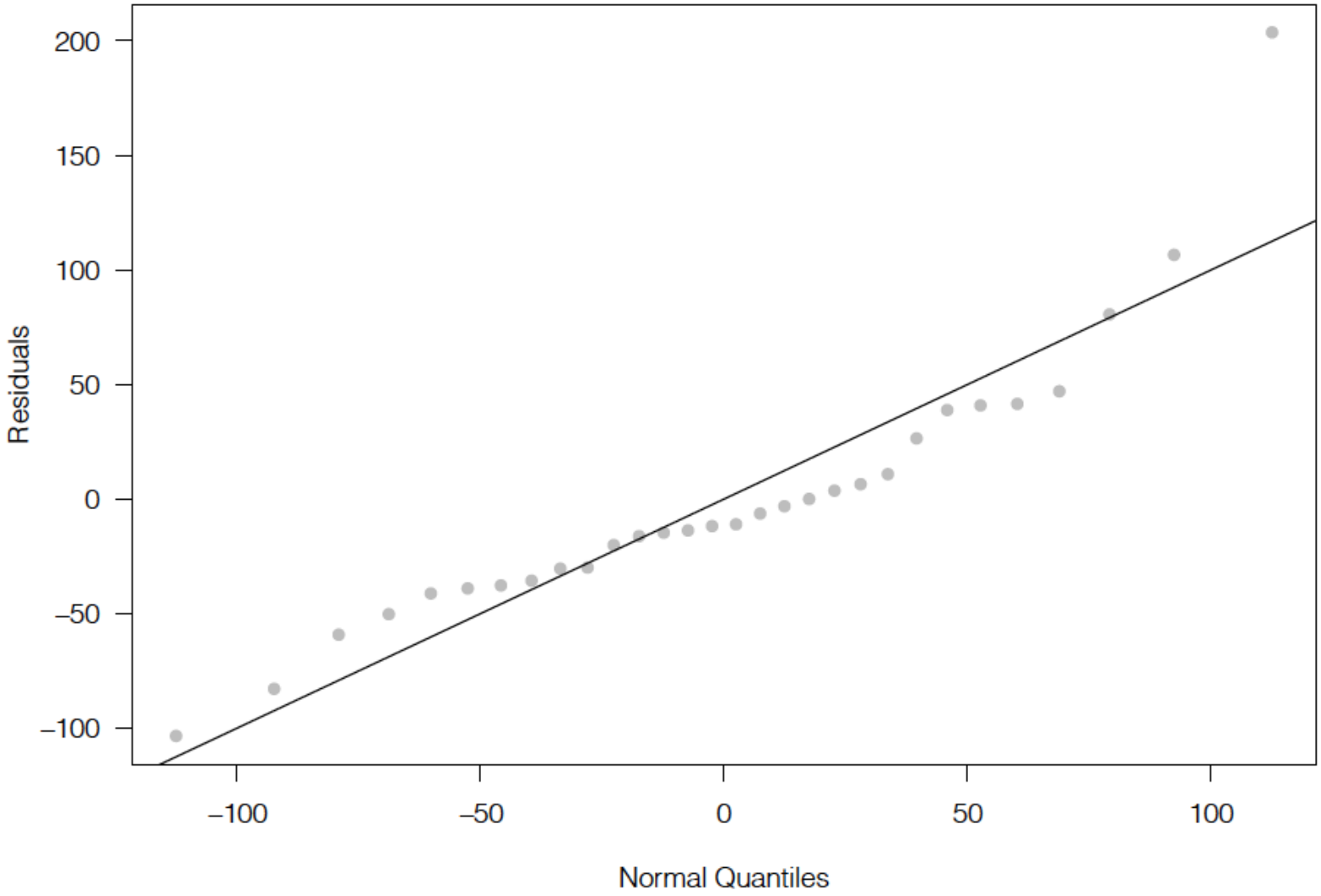
Assessing Normality of Residuals: Histogram

```
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)
```



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)
```



Leverage: Detecting "Extreme" Predictor Values

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

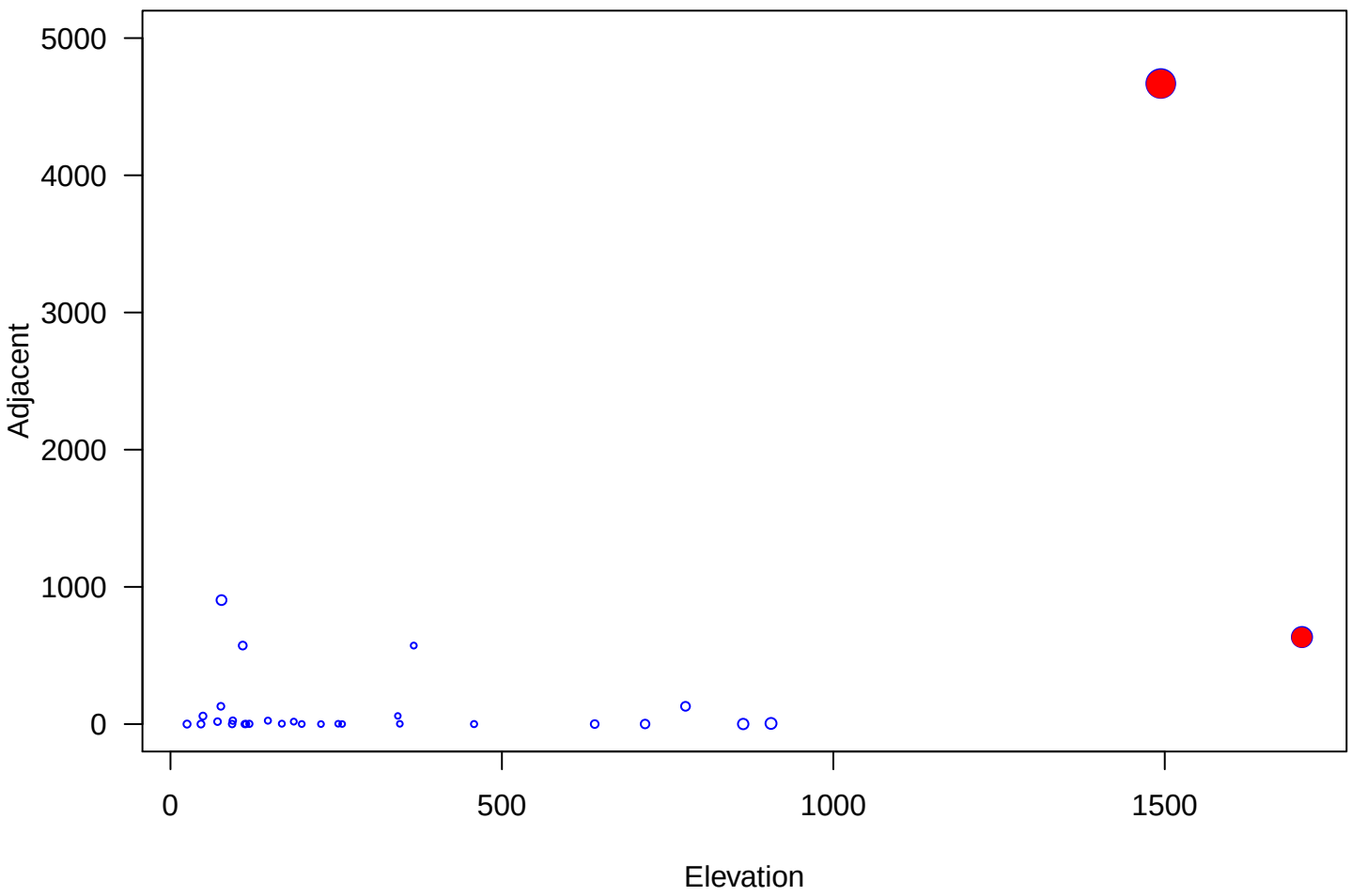
- The leverage value for the i_{th} observation is defined as:

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

- Can show that $\text{Var}(e_i) = \sigma^2(1 - h_i)$, where $e_i = y_i - \hat{y}_i$ is the residual for the i_{th} observation

- $\frac{1}{n} \leq h_i \leq 1, \quad 1 \leq i \leq n$ and $\bar{h} = \sum_{i=1}^{n} \frac{h_i}{n} = \frac{p}{n} \Rightarrow$ a "rule of thumb" is that leverages greater than $\frac{2p}{n}$ should be examined more closely

Leverage Values of `Species ~ Elev + Adj`

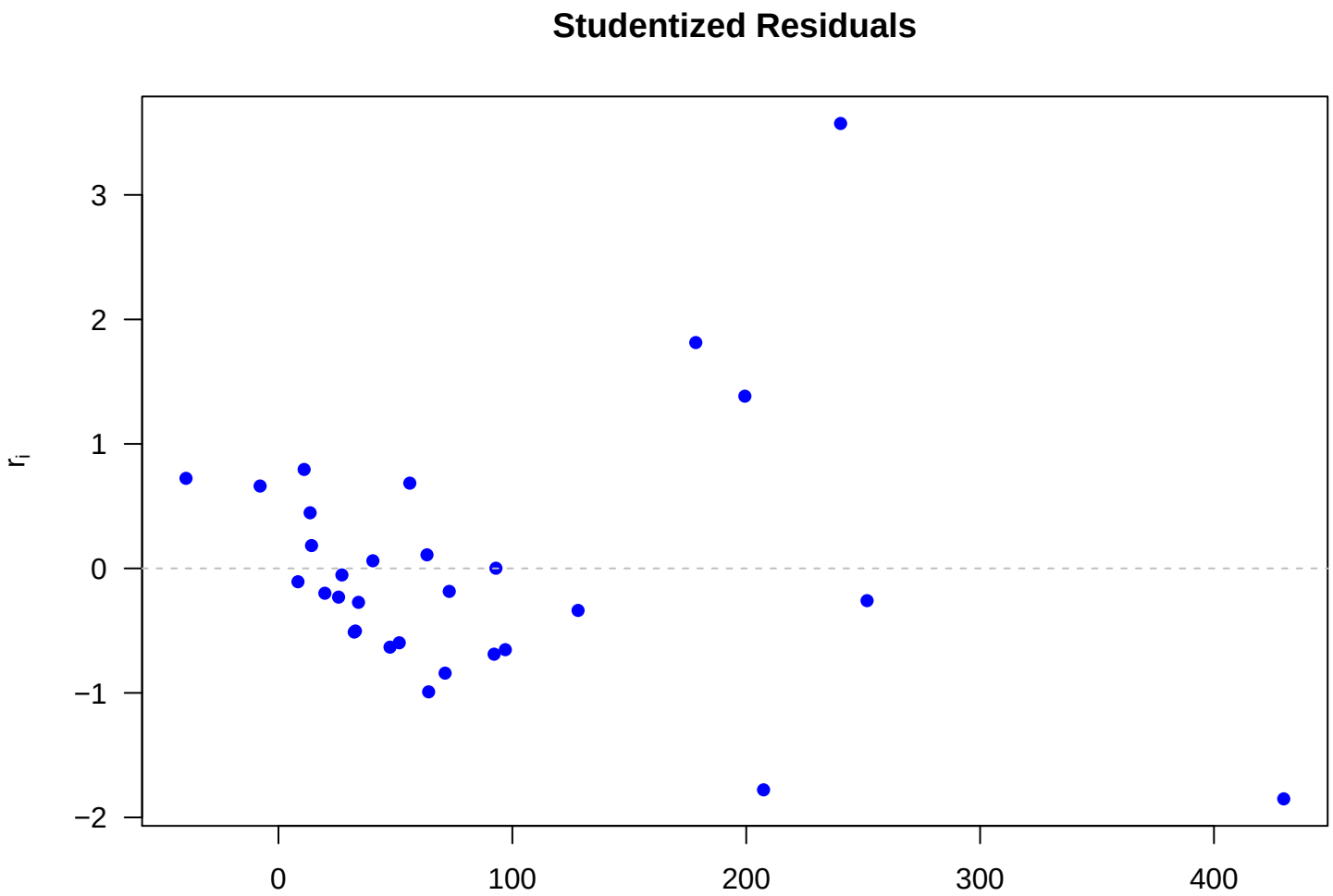


Standardized Residuals

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 **standardized 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 $\text{Corr}(r_i, r_j)$ tends to be small

Standardized Residuals of `Species ~ Elev + Adj`



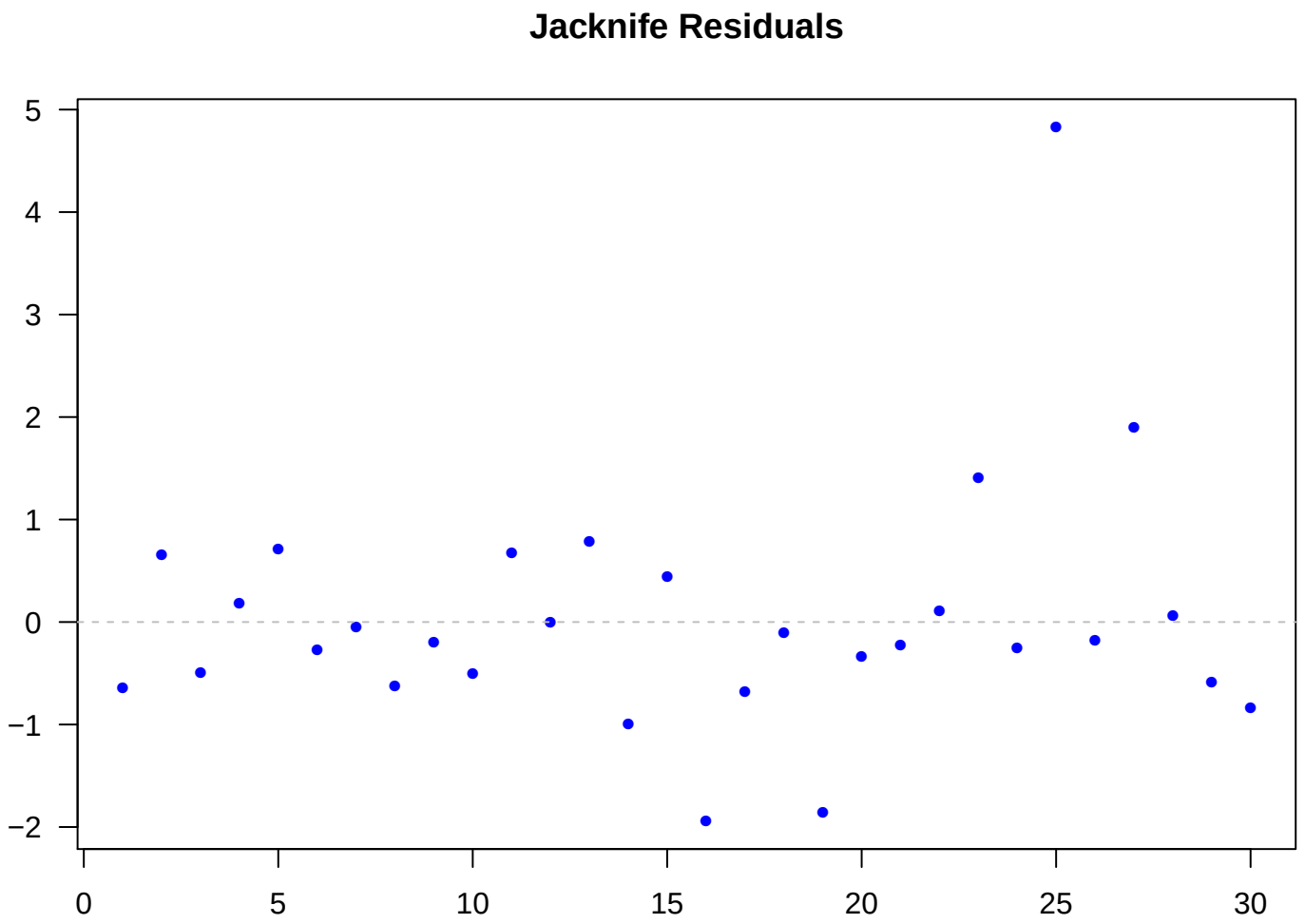
Studentized (Jackknife) Residuals

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

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

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

Studentized (Jackknife) Residuals of `Species ~ Elev + Adj`

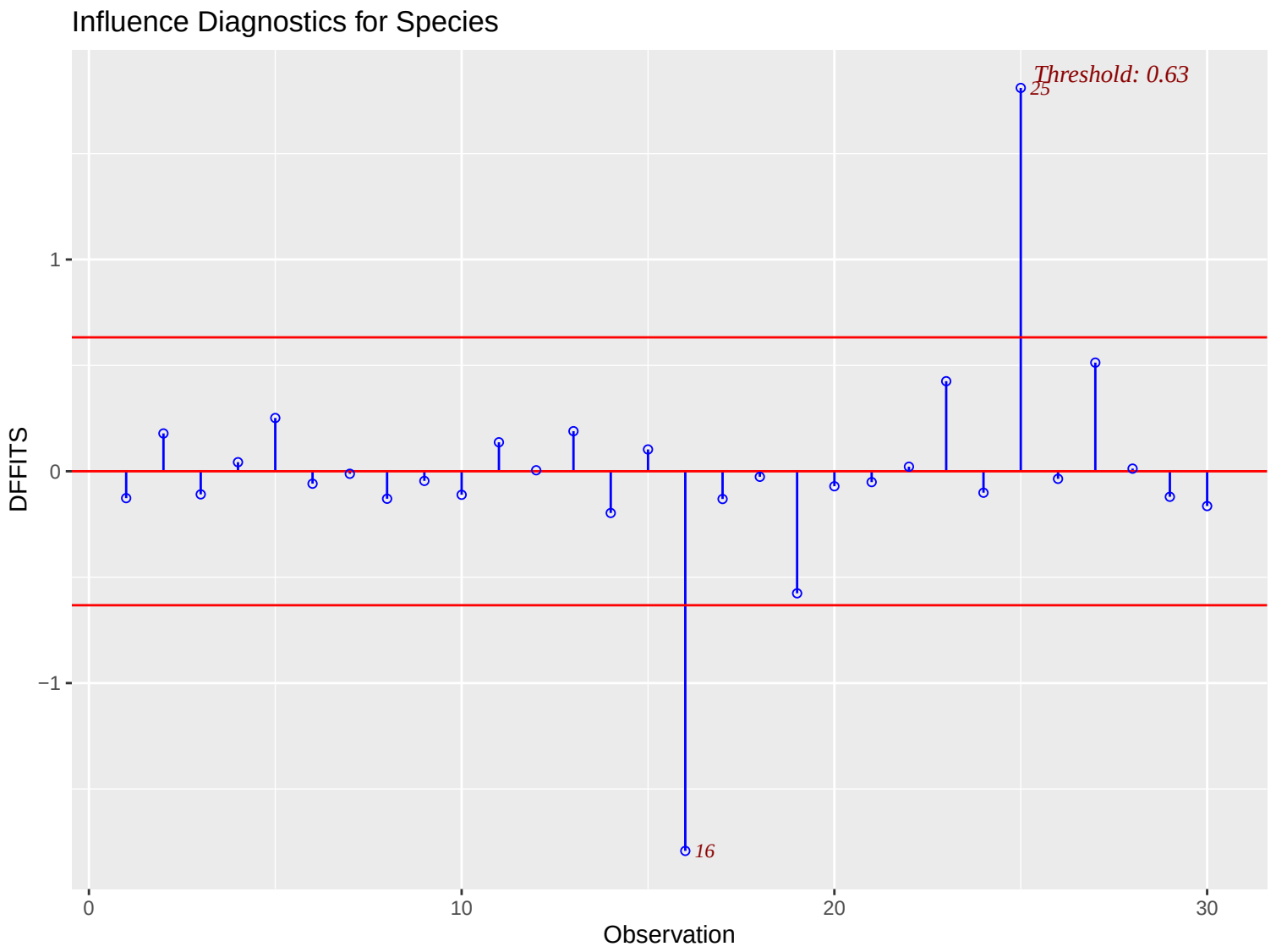


Identifying Influential Observations: DFFITS

DFFITS measures the change in the predicted values for each observation when that observation is omitted.

- 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`



Identifying Influential Observations: Cook's Distance

Cook's Distance quantifies how much the predicted values change when a particular observation is excluded from the analysis.

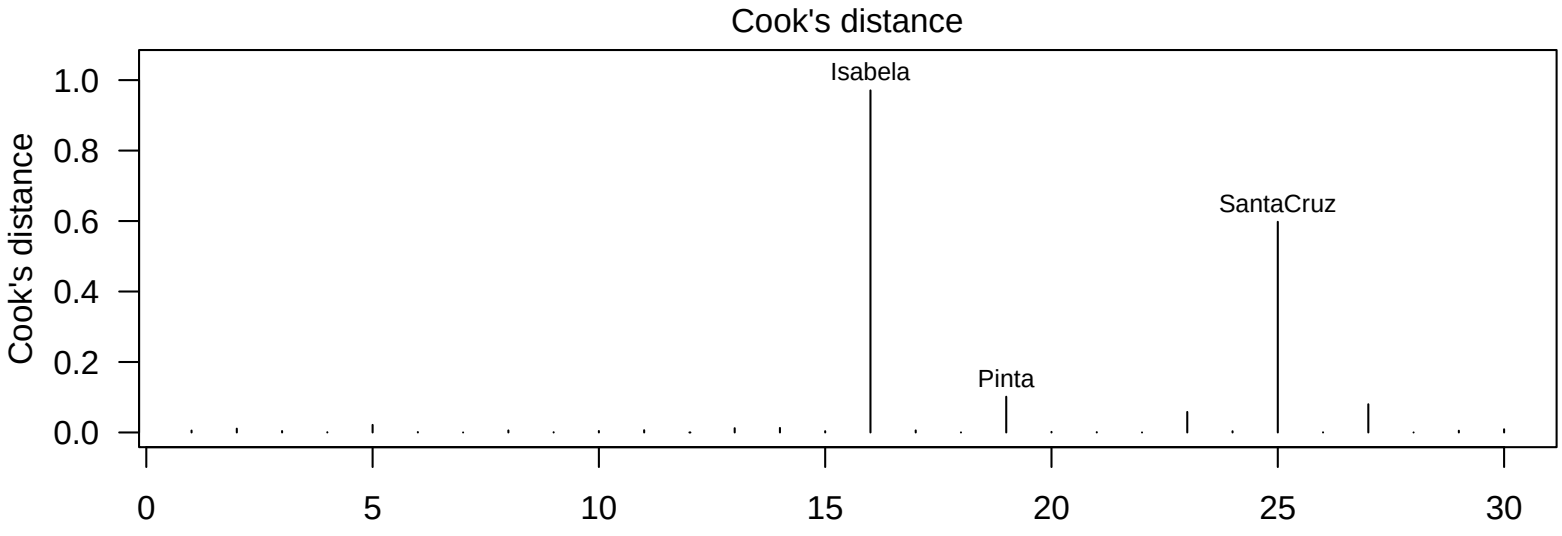
- Cook's distance measure (D_i) is defined as:

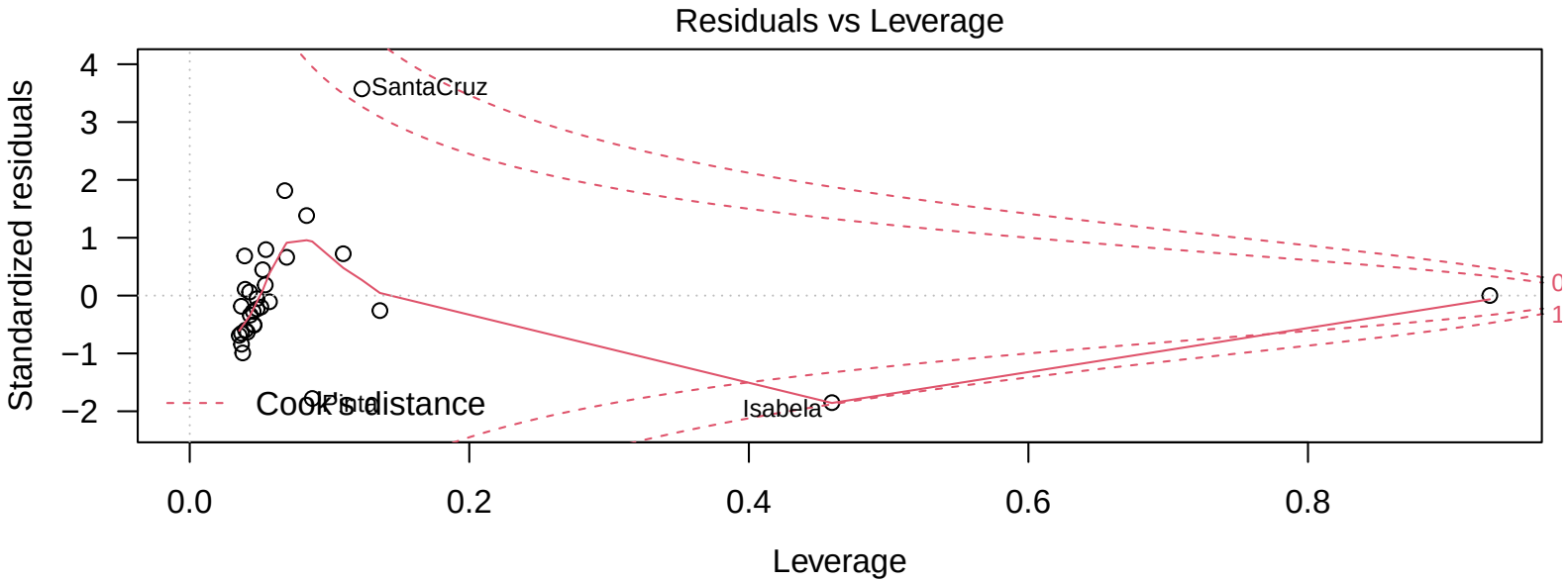
$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times \text{MSE}} \left(\frac{h_i}{(1 - h_i)^2} \right)$$

- Cook's Distance considers both leverage and residual, providing a broader measure of influence
- Here are the guidelines commonly used:

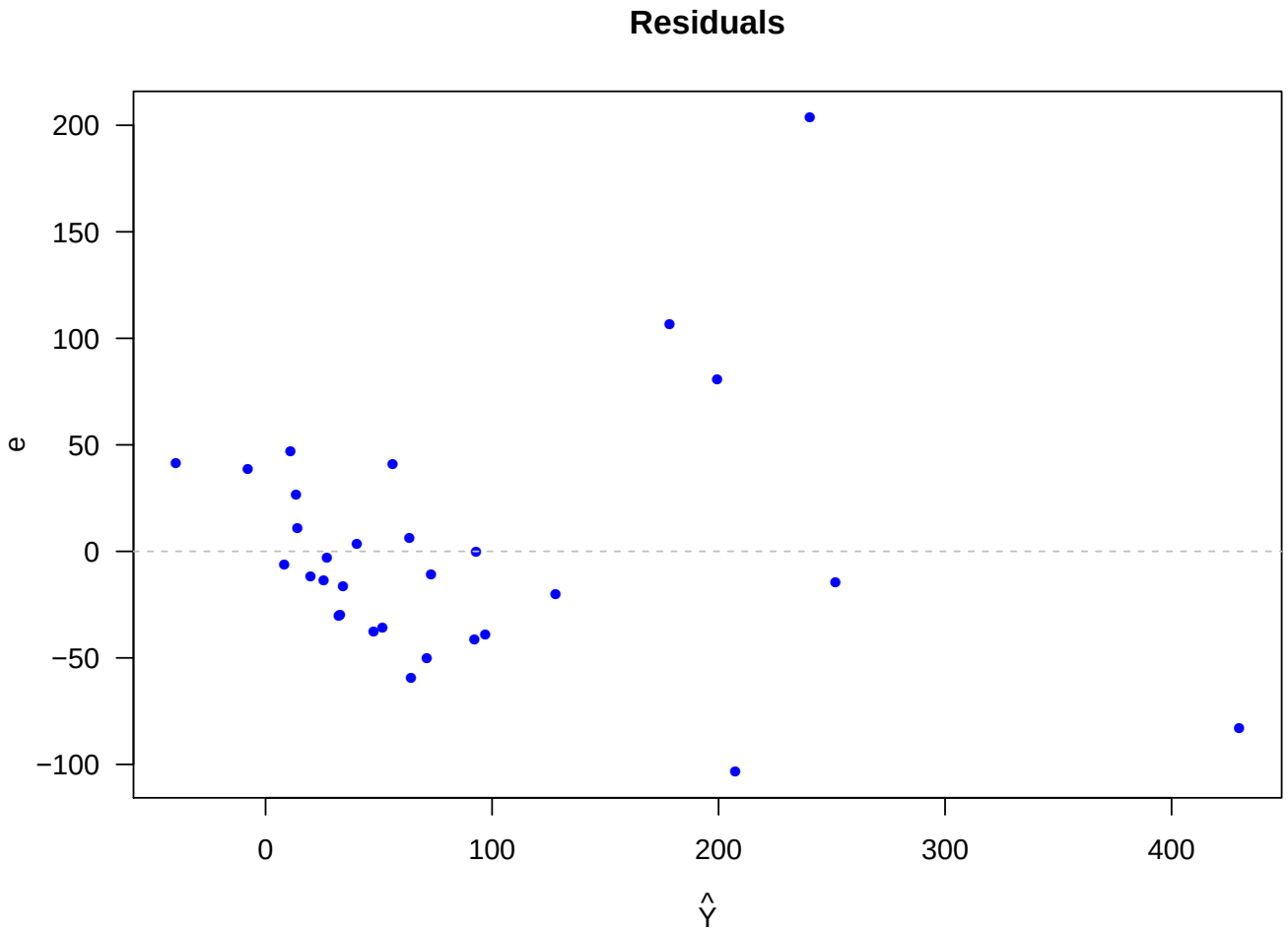
1. If $D_i > 0.5$, then the i$^{\text{th}}$ data point is worthy of further investigation as it may be influential
2. If $D_i > 1$, then the i$^{\text{th}}$ data point is quite likely to be influential

Cook's Distance of `Species ~ Elev + Adj`





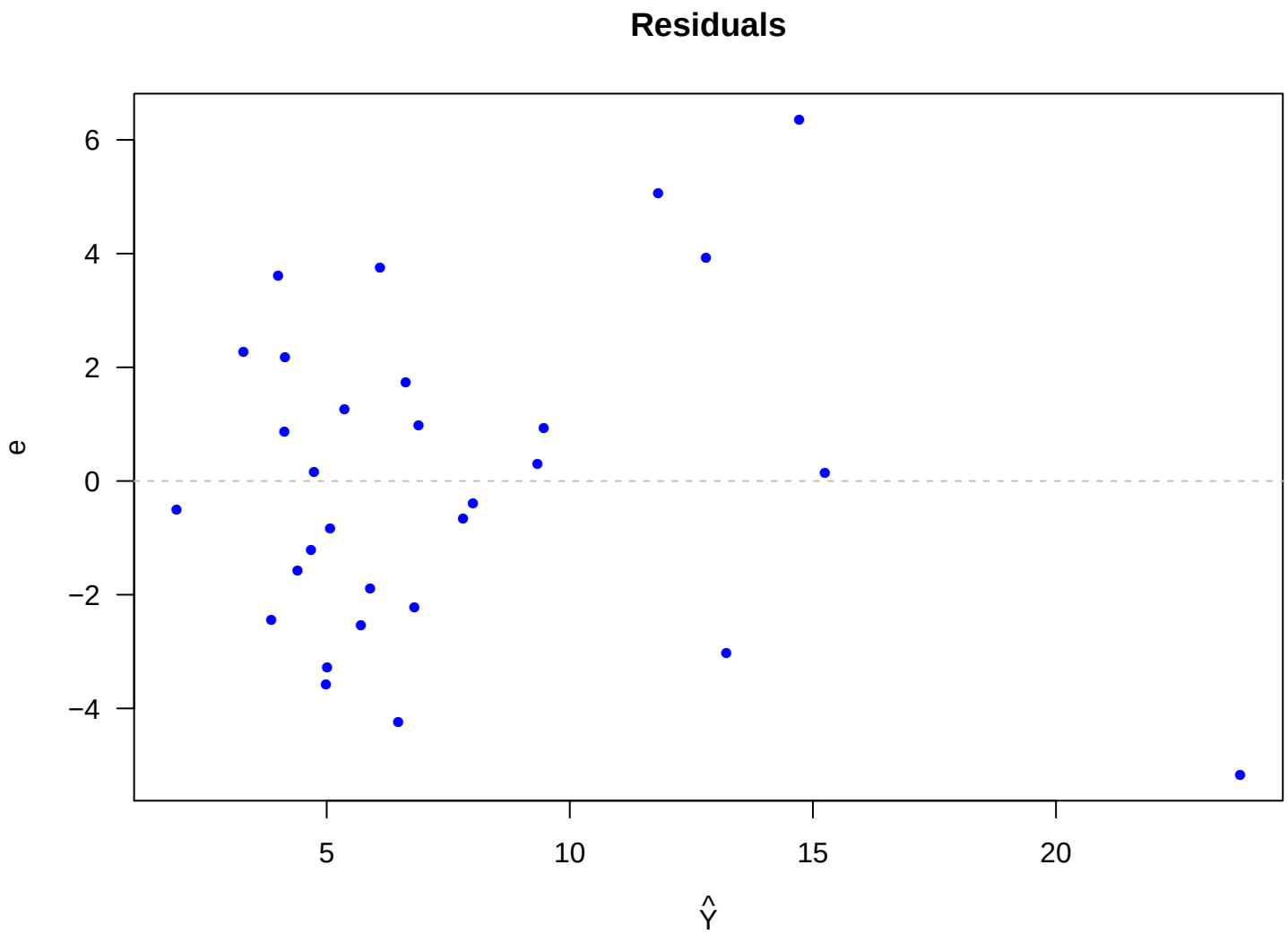
Residual Plot of `Species ~ Elev + Adj`



Such a residual plot suggests a violation of constant variance

Residual Plot After Square Root Transformation

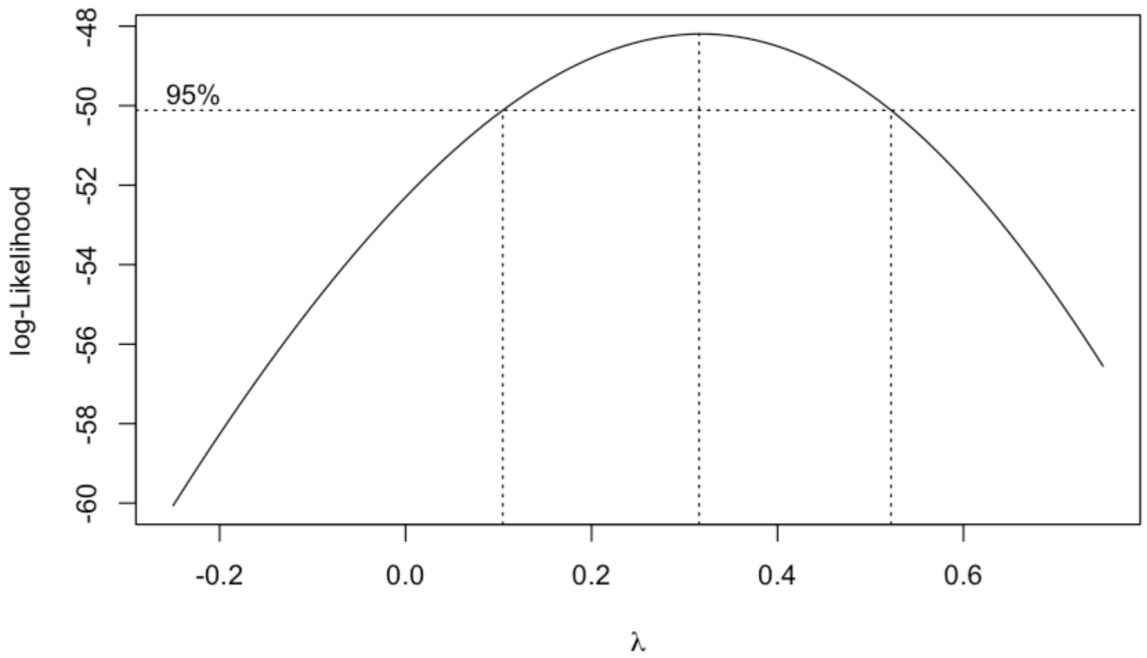
$$\sqrt{\text{Species}} \sim \text{Elev} + \text{Adj}$$



Box-Cox Transformation

The Box-Cox method [Box and Cox, 1964] is a powerful way to determine if a transformation on the response is needed

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$$



In R, we can use the `boxcox` function from the `MASS` package to perform a Box-Cox transformation. The plot suggests a cube root may be needed

Summary

These slides cover:

- Model/variable selection can be done via some criterion-based methods to balance bias and variance
- Model diagnostics is crucial to ensure valid statistical inference
- Box-Cox Transformation can be used to transform the response in order to correct model violations

`R` functions to know:

- `regsubsets` in the `leaps` library and `step` for model selection
- `influence.measures` includes a suite of functions (`hatvalues`, `rstandard`, `rstudent`, `dffits`, `cooks.distance`) for computing regression diagnostics
- `boxcox` in the `MASS` library for performing a Box-Cox transformation