Lecture 4

Multiple Linear Regression: Model Selection and Model Checking

Reading: Faraway 2014 Chapters 6, 9.1, and 10

STAT 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

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

Why Model Selection Matters

Multiple Linear Regression Model:

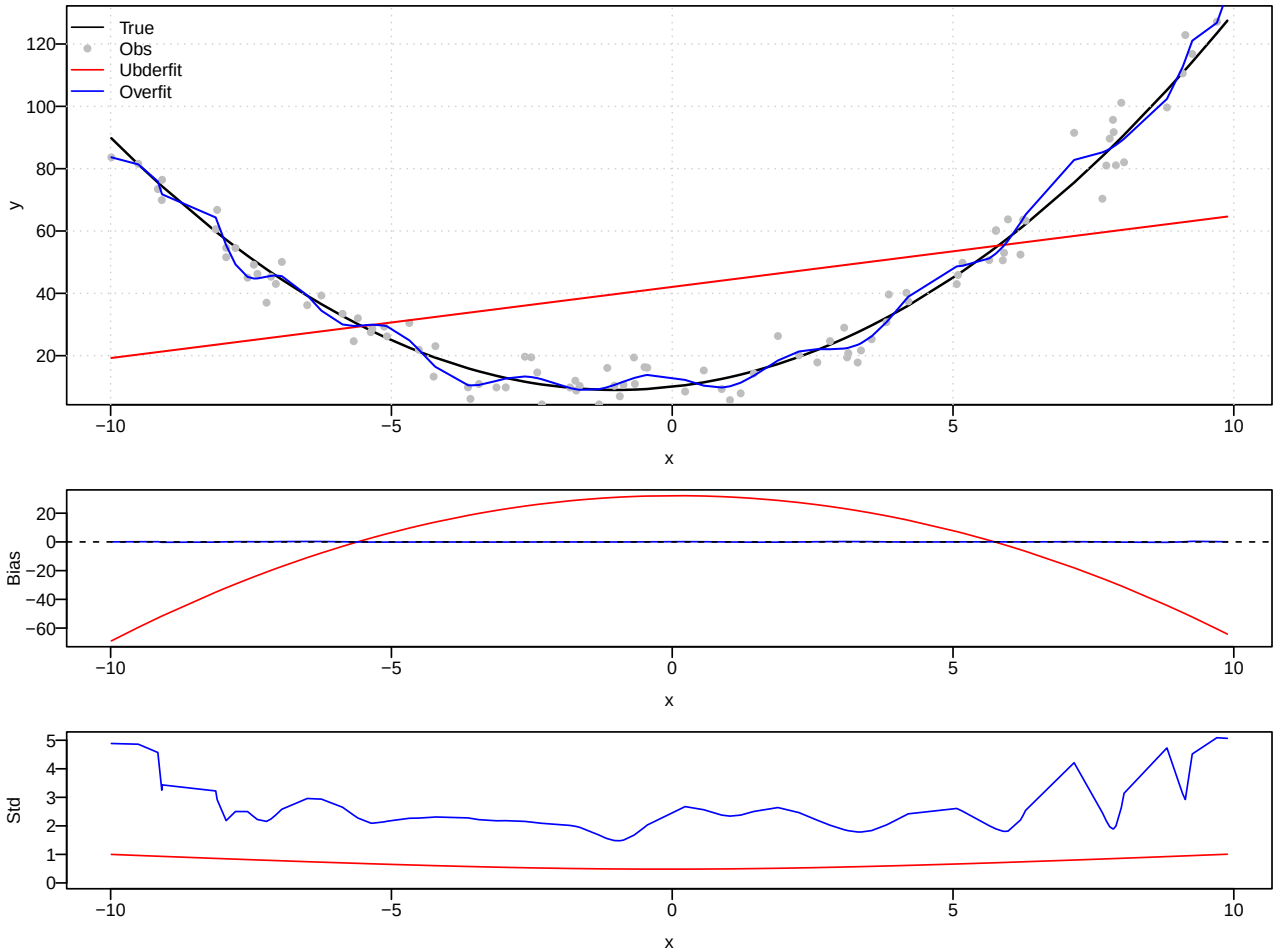
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

Goal: choose a model that predicts well

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



Interpretation:

- Underfit → high bias, low variance
- Overfit → low bias, high variance
- Best model → a balance between bias and variance

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: Model Selection

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

Goal: Choose simple model (i.e., small p) with small C_p. See R session for an example

Adjusted R^2: Penalized Fit

Adjusted R^2 (R^2_{adj}) accounts for the artificial increase in R^2 when additional predictors are added.

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

Rule:

- Choose the model with the largest R^2_{adj}

Interpretation:

- Higher value $\rightarrow$ better balance of fit and complexity

Information criteria (AIC & BIC)

Information criteria are measures for model selection that balance goodness of fit and model complexity

- Akaike's information criterion (AIC)

$$n \log(\frac{\text{SSE}_p}{n}) + 2p$$

- Bayesian information criterion (BIC)

$$n \log(\frac{\text{SSE}_p}{n}) + p \log(n)$$

Here p is the number of the parameters in the model

Smaller AIC/BIC $\rightarrow$ better model; BIC penalizes model complexity more strongly than AIC

Automatic Model Selection

- 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:

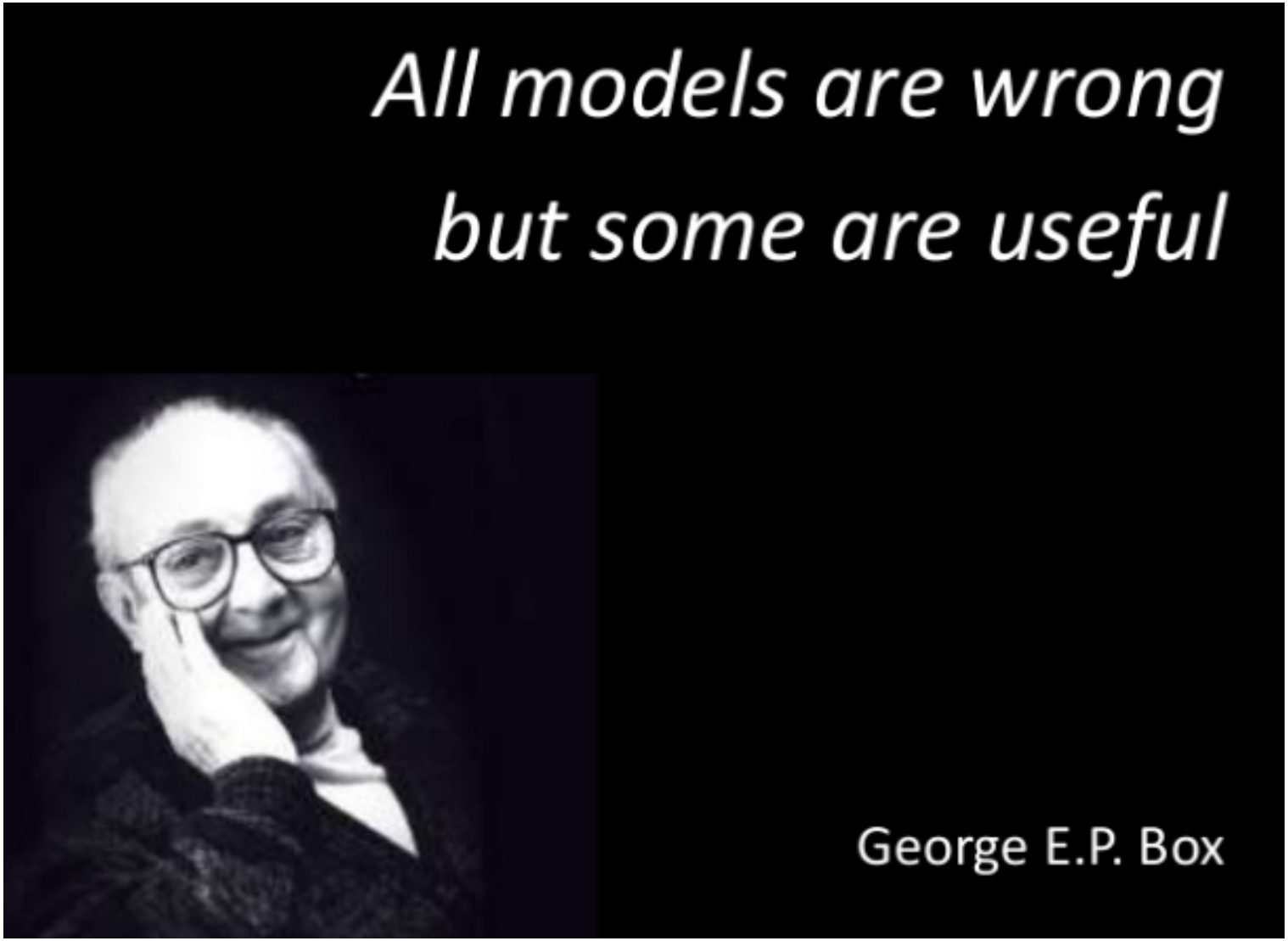
- Linearity: mean is linear in predictors

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

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

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

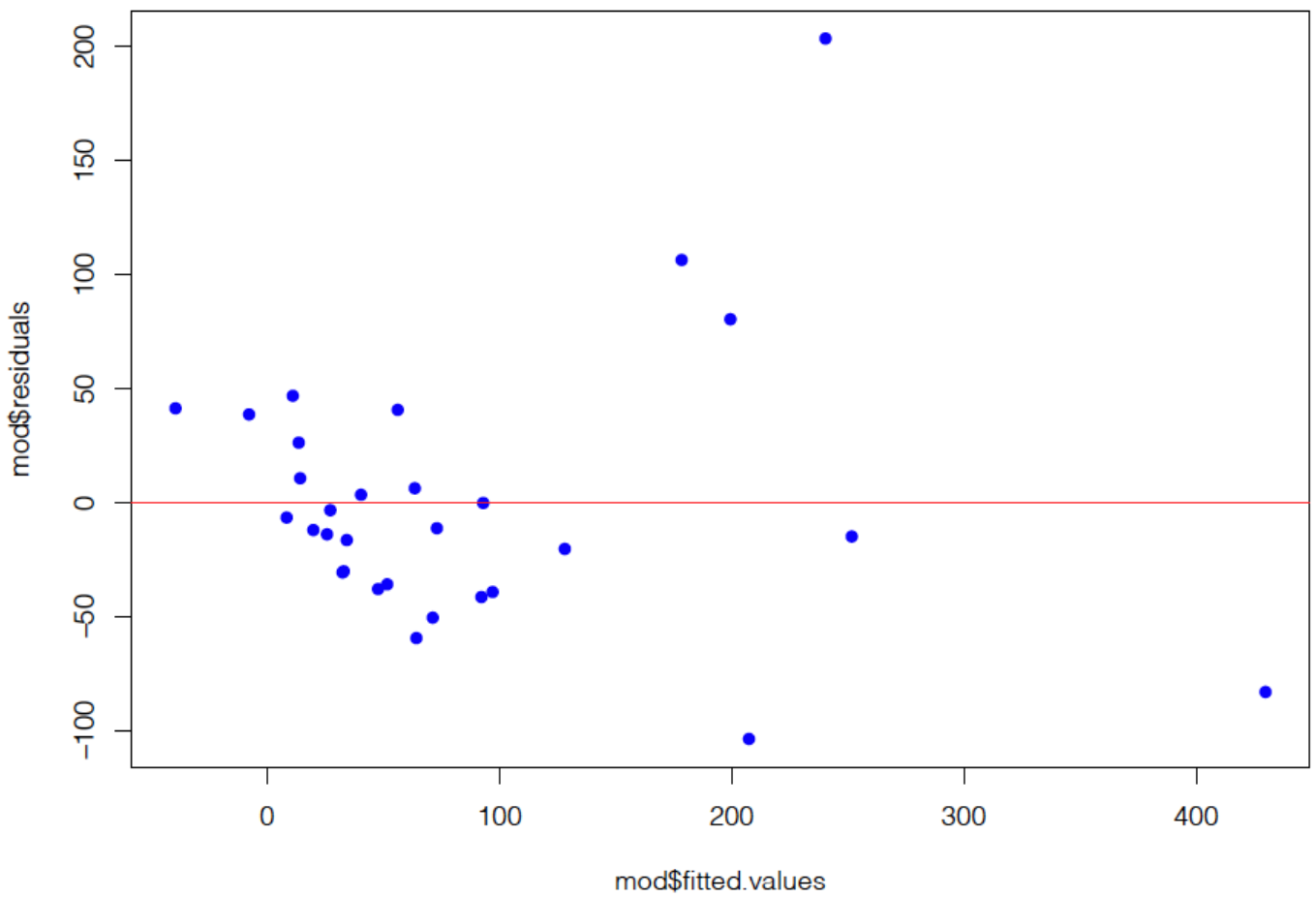
Violations $\Rightarrow$ invalid inference



"All models are wrong, but some are useful" – their usefulness depends on validating assumptions through careful diagnostics

Residuals versus Fits Plot

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

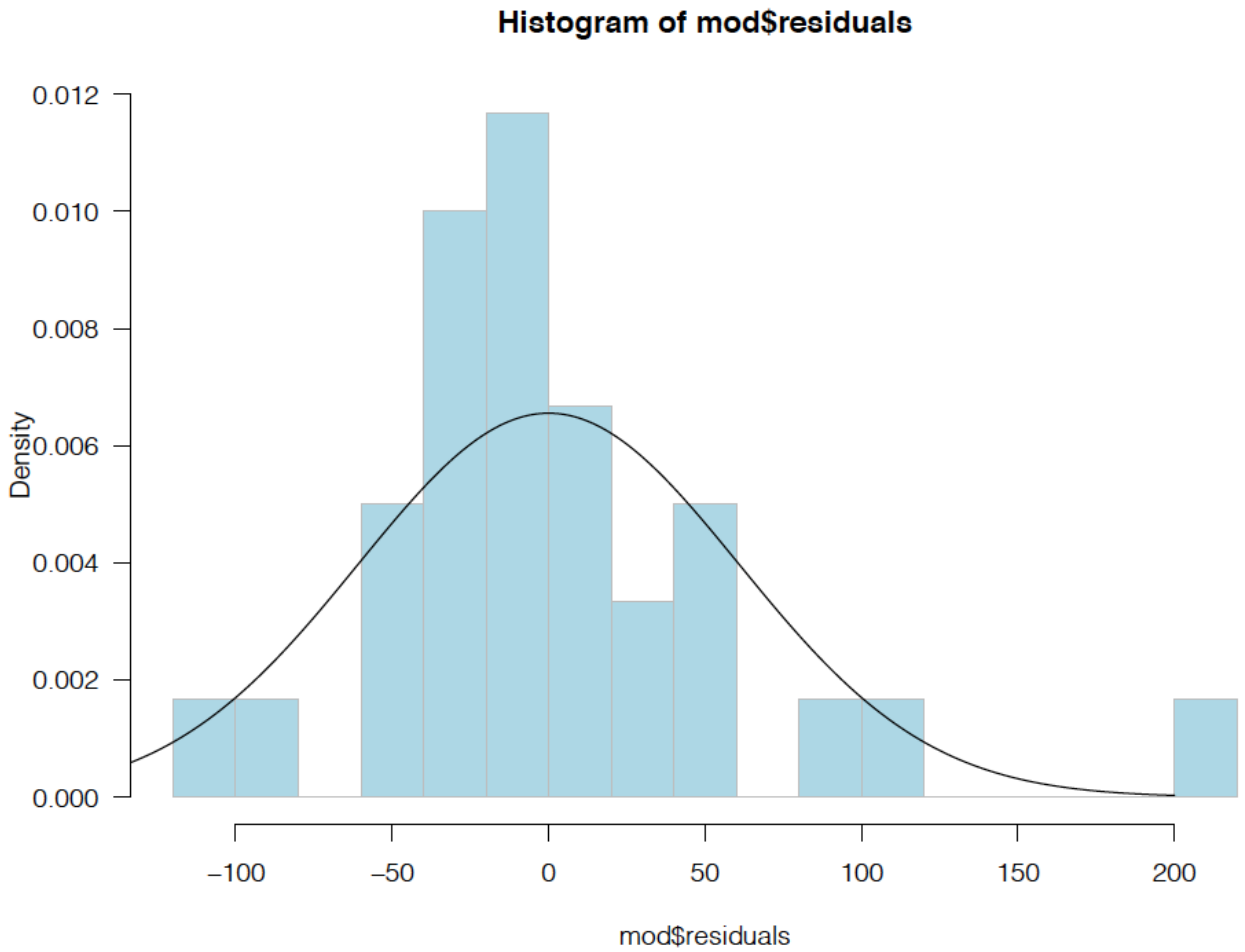


What to check:

- Random scatter ⇒ good
- Pattern ⇒ model misspecification
- Funnel shape ⇒ non-constant variance

Assessing Normality of Residuals: Histogram

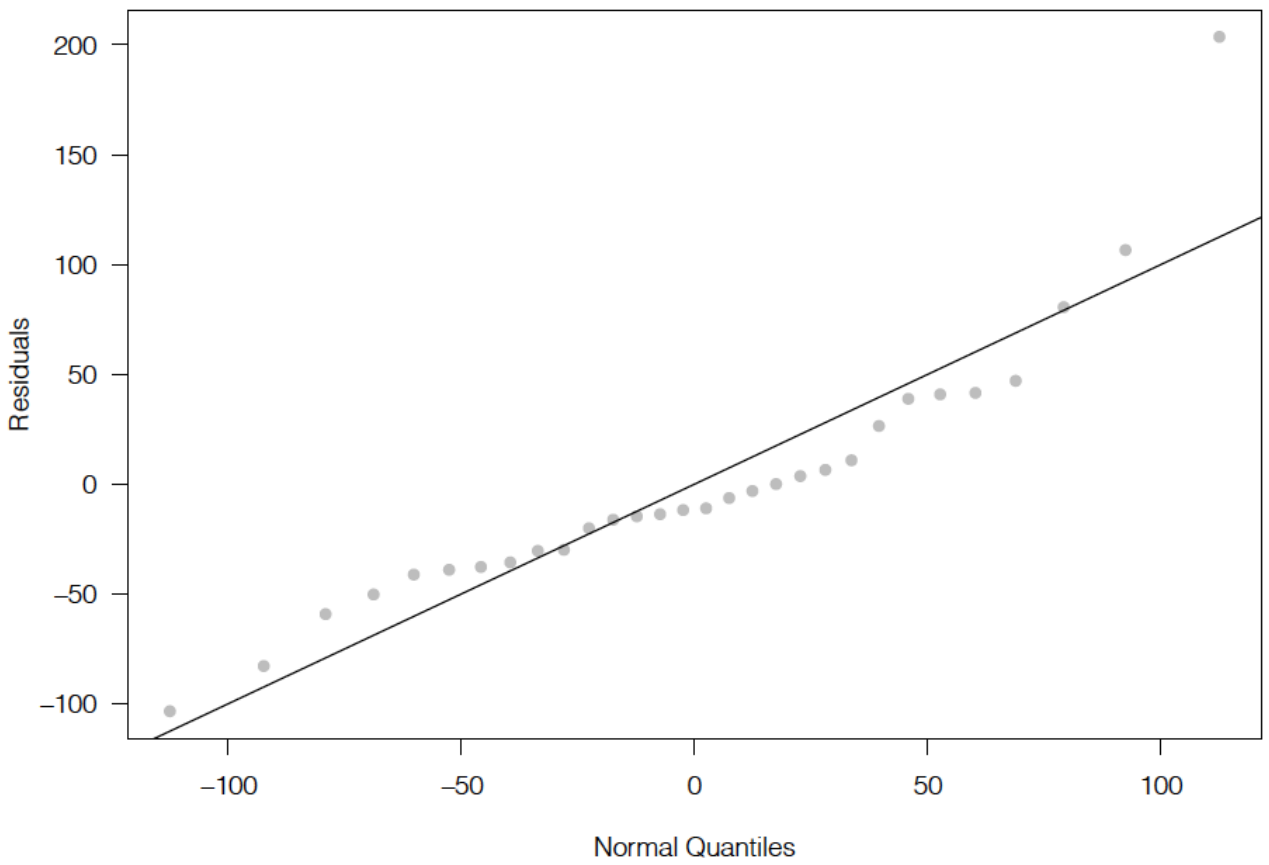
```
par(las = 1)
hist(mod$residuals, 12, prob = T,
     col = "lightblue", border = "gray")
xg <- seq(-200, 200, 1)
yg <- dnorm(xg, 0, 60.86)
lines(xg, yg)
```



Good: approximately symmetric, bell-shaped
Concern: skewness or heavy tails

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)
```



Interpretation:

- Points on line ⇒ normality holds
- Curvature ⇒ skewness
- Extreme deviations ⇒ outliers

Leverage: Detecting Extreme Predictor Values

Definition:

$$h_i = H_{ii}, \quad \text{where } \hat{\mathbf{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

Key facts:

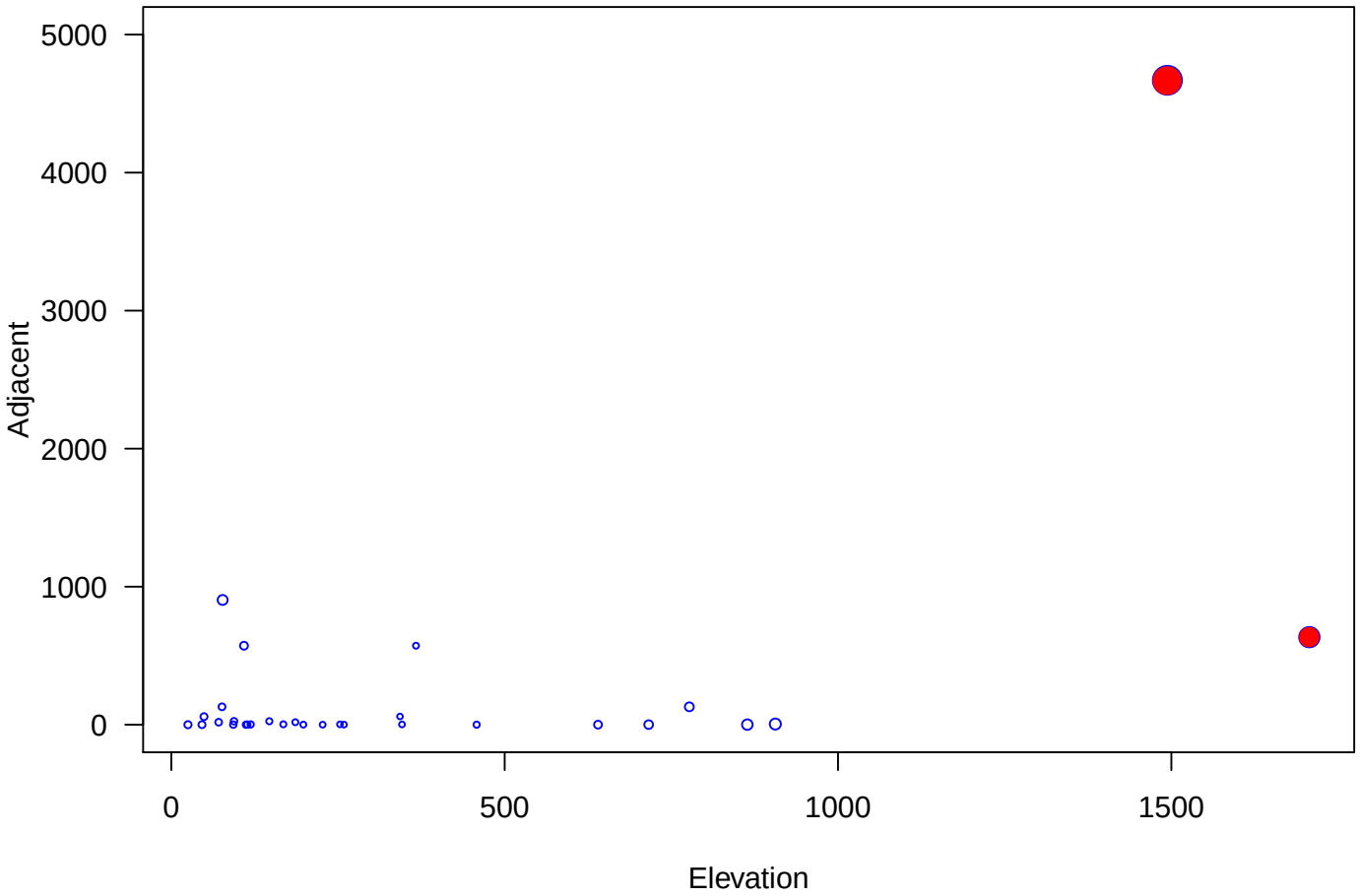
- $\frac{1}{n} \leq h_i \leq 1, \;\; \sum_{i=1}^{n} h_i = p \;\Rightarrow\; \bar{h} = \frac{p}{n}$
- $\text{Var}(e_i) = \sigma^2(1 - h_i) \Rightarrow$ higher leverage (h_i large) $\Rightarrow \hat{y}_i$ is pulled closer to y_i

Rule of thumb:

$$h_i > \frac{2p}{n} \text{ (twice the average)} \;\Rightarrow\; \text{high leverage}$$

Interpretation: Unusual x values $\Rightarrow$ strong influence on model fit

Leverage Values of `Species ~ Elev + Adj`



Highlighted points (larger circles) have high leverage–far from typical x values

Standardized Residuals

Definition:

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

Why?

- Adjusts for unequal variance: $\text{Var}(e_i) = \sigma^2(1-h_i)$
- Makes residuals comparable (approx. unit variance)

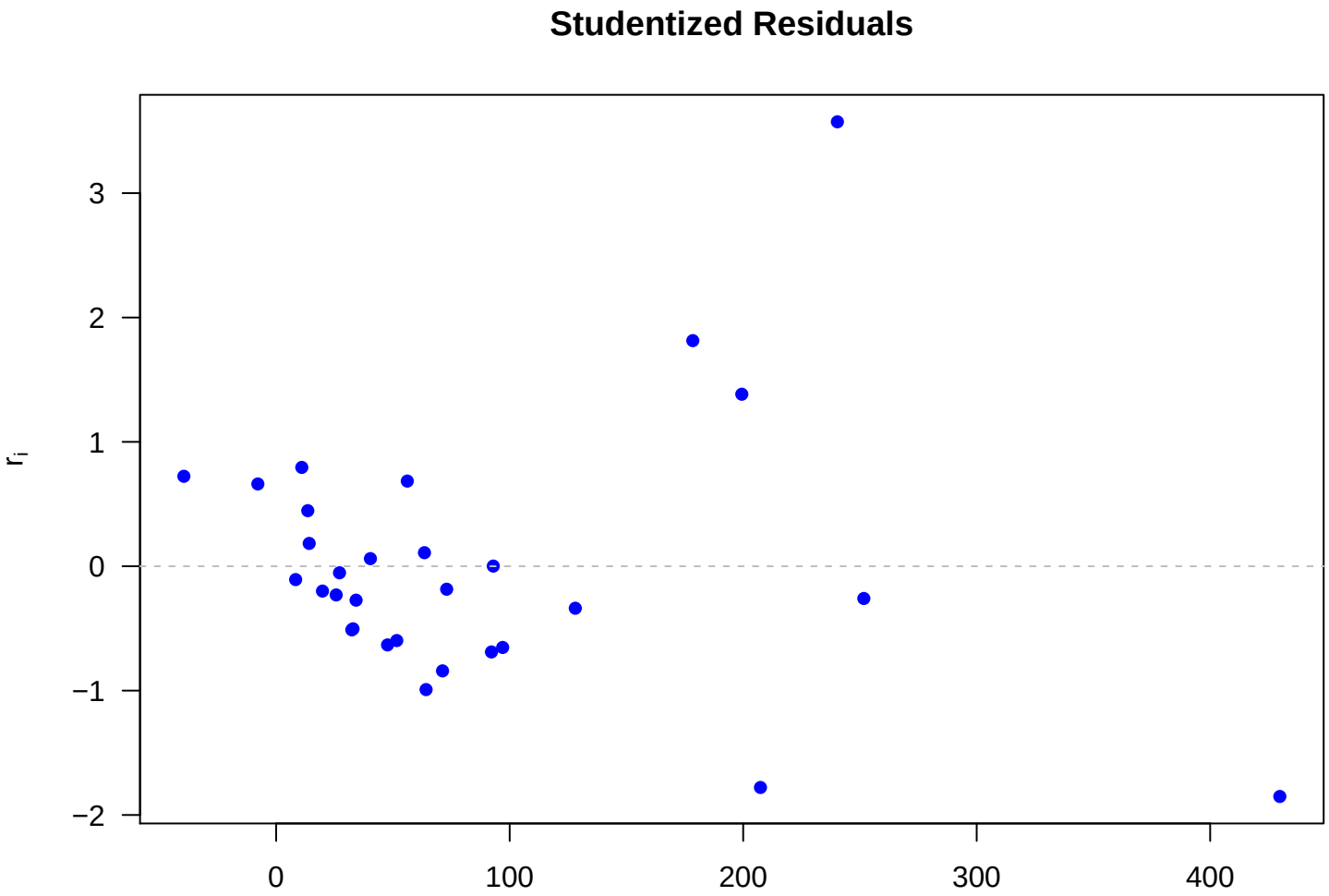
Properties (if model is correct):

- $\text{Var}(r_i) \approx 1$
- weak correlation between r_i's

Use: Better for outlier detection

Rule: Compare to t_{n-p}

Standardized Residuals of `Species ~ Elev + Adj`



In practice, residual plots using standardized residuals typically show nearly the same pattern

Studentized (Jackknife) Residuals

- Remove observation i and refit the model to obtain $\hat{\boldsymbol{\beta}}_{(i)}, \hat{\sigma}_{(i)}, \hat{y}_{i(i)}$. A large deleted-fit residual $y_i - \hat{y}_{i(i)}$ suggests observation i may be an outlier
- The variance of the deleted-fit residual is

$$\text{Var}(y_i - \hat{y}_{i(i)}) = \sigma^2_{(i)}(1 - h_i)$$

- Define the **Studentized (Jackknife) residual**

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\sqrt{\text{MSE}_{(i)}(1 - h_i)}} \sim t_{n-p-1}$$

under the model assumptions

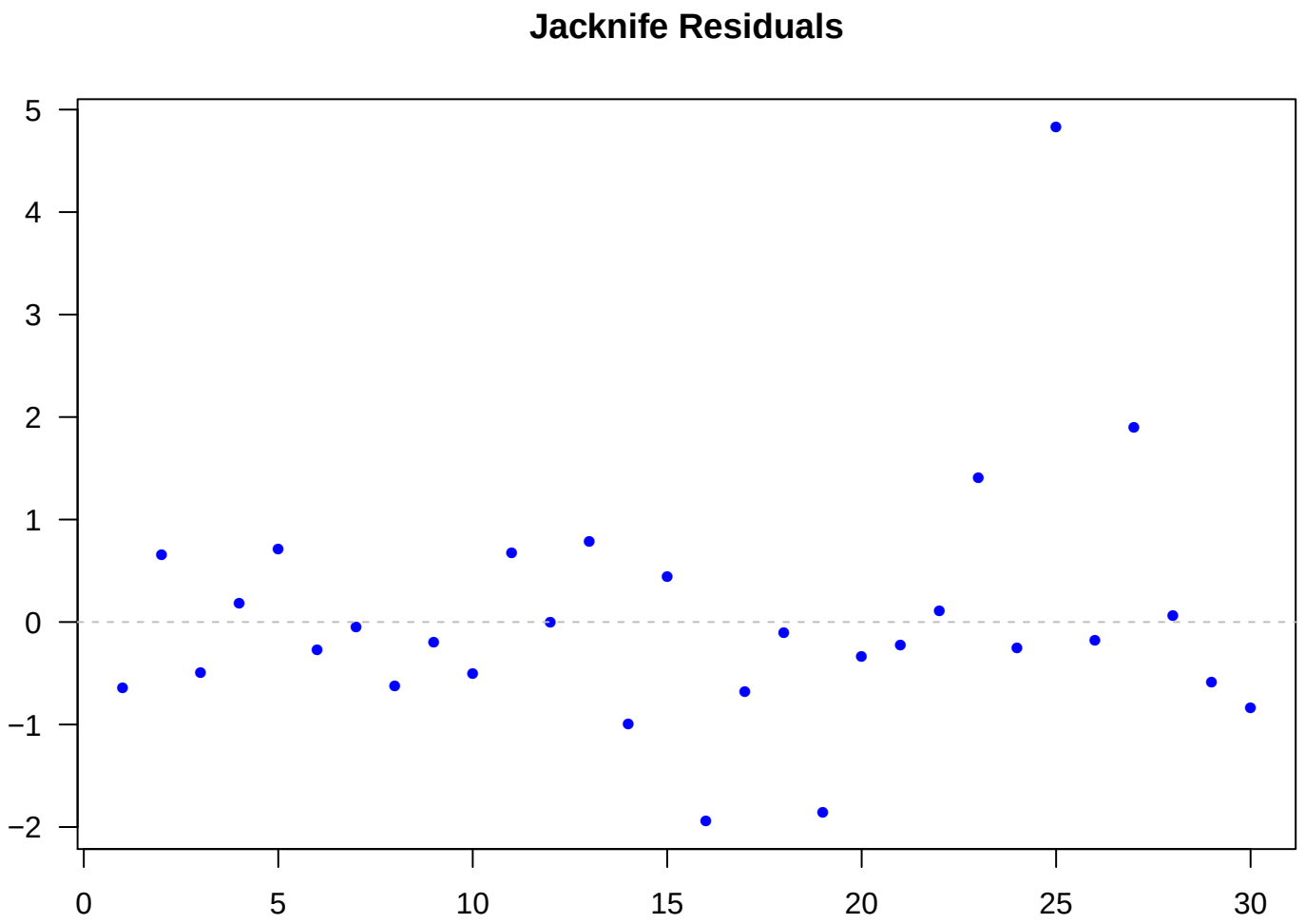
Interpretation

- Large $|t_i|$ indicates a potential outlier
- More reliable than standardized residuals since observation i is excluded when estimating variability

Rule of Thumb

- $|t_i| > 2$: possible outlier
- $|t_i| > 3$: strong evidence of outlier

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



Identifying Influential Observations: DFFITS

DFFITS measures how much the fitted value for observation i changes when observation i is removed.

- Compare the fitted value $\hat{y}_i$ and the deleted-fit prediction $\hat{y}_{i(i)}$
- Definition

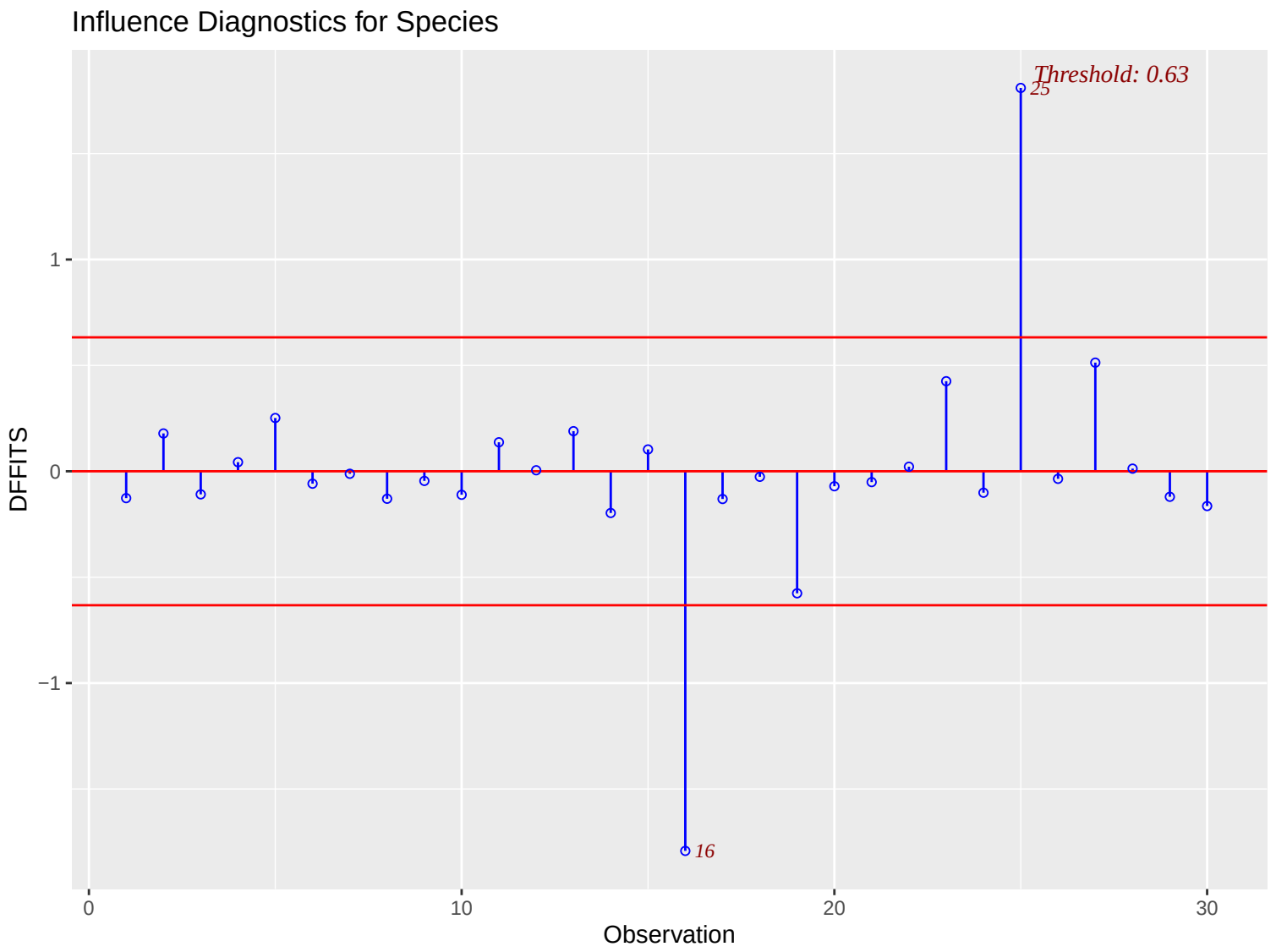
$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_i}}$$

- Large $|\text{DFFITS}_i|$ suggests an influential observation
- Rule of thumb:

$$|\text{DFFITS}_i| > 1 \quad \text{or} \quad |\text{DFFITS}_i| > 2\sqrt{p/n}$$

- Note: Studentized (Jackknife) residuals focus on outlier detection, while DFFITS measures influence by combining residual size and leverage

DFFITS of `Species ~ Elev + Adj`



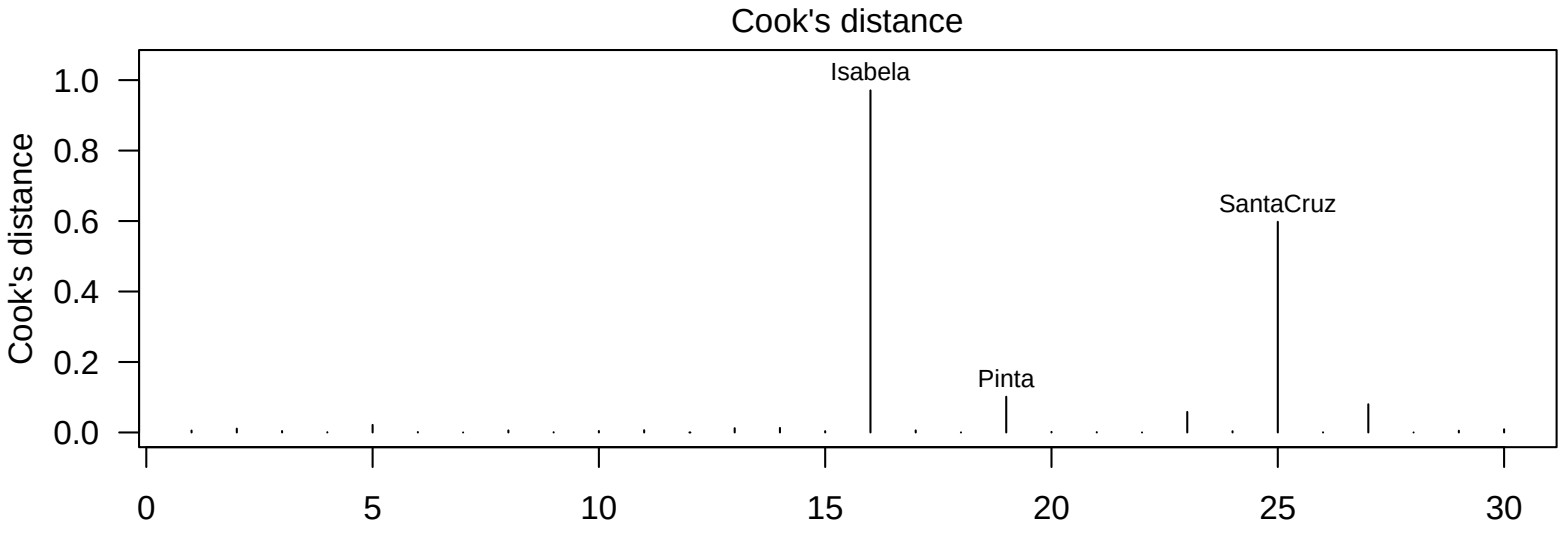
Cook's Distance

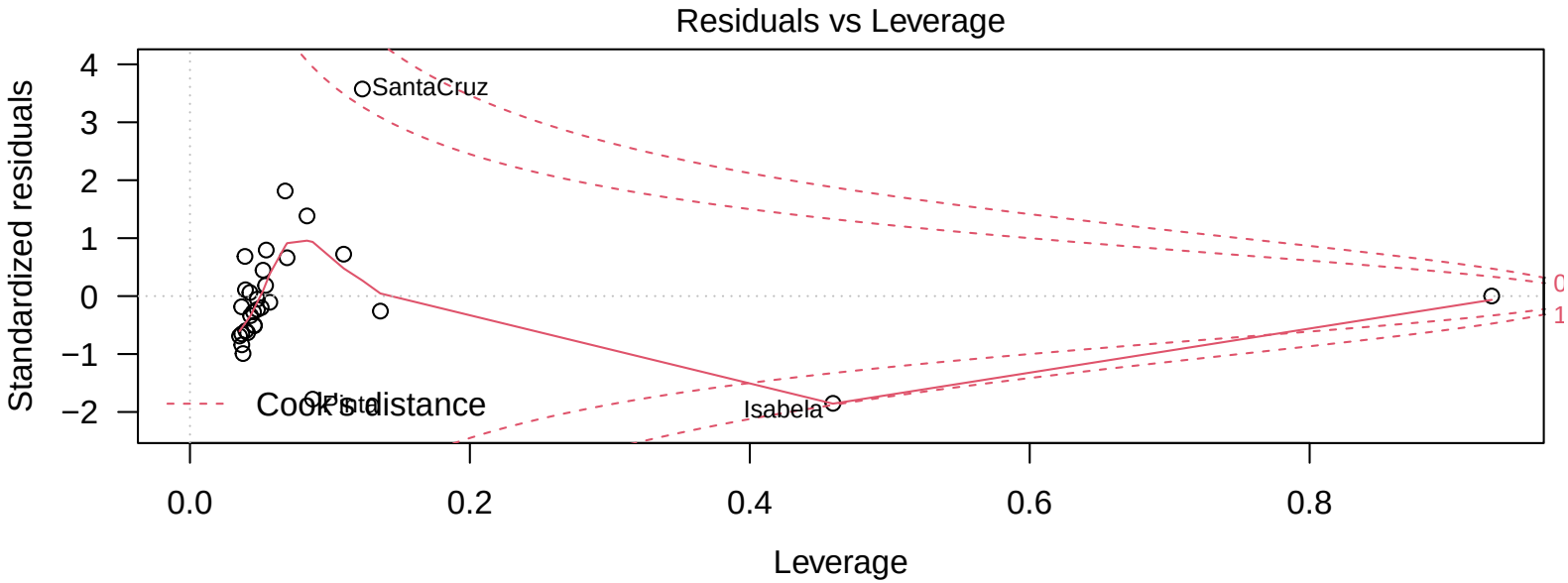
Cook's Distance measures how much the overall regression fit changes when observation i is removed

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

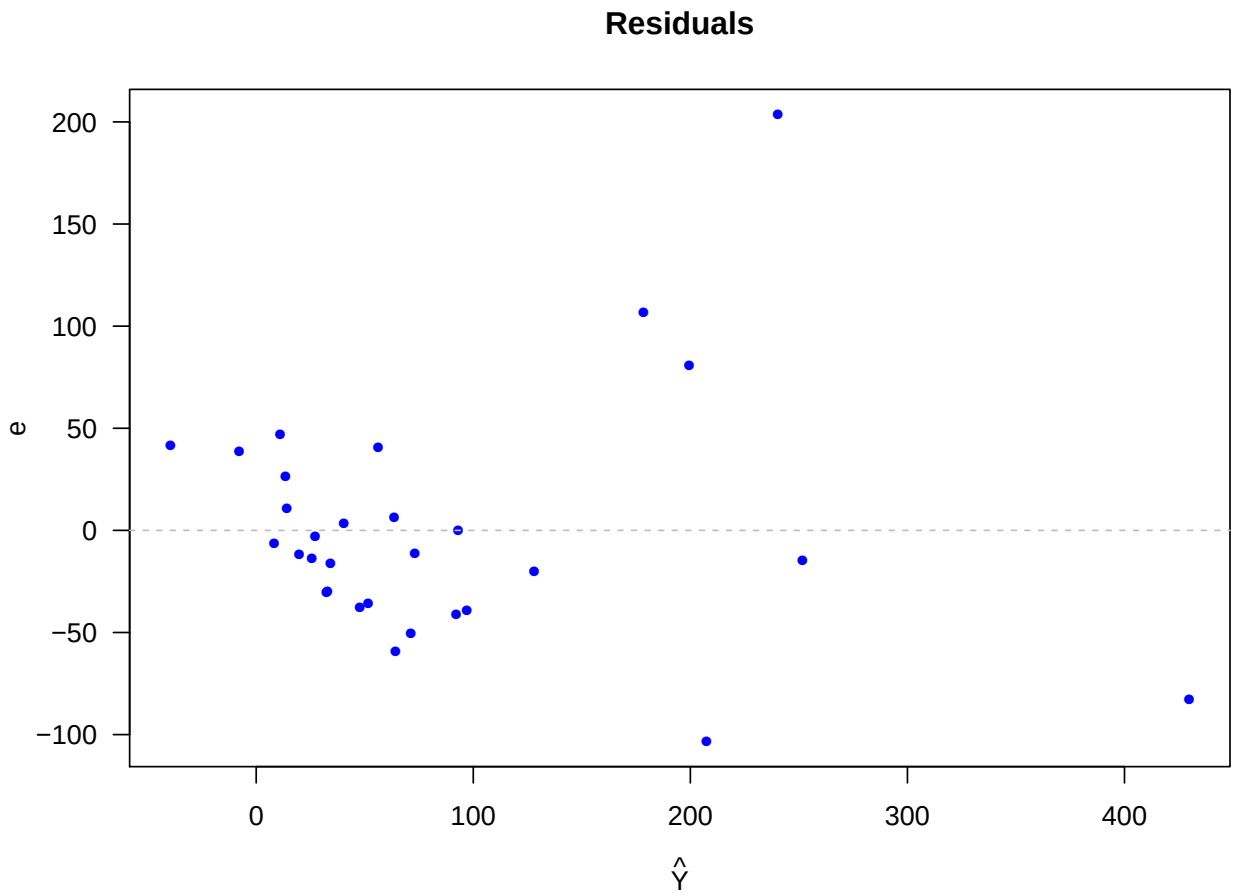
- Combines both residual size and leverage to assess influence
- Unlike DFFITS, which measures the change in the fitted value for observation i, Cook's distance measures the overall impact on the regression model
- Common guidelines:
  - $D_i > 4/n$: potentially influential
  - $D_i > 1$: highly influential
- In R diagnostic plots, observations with the largest Cook's distances are often labeled automatically

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





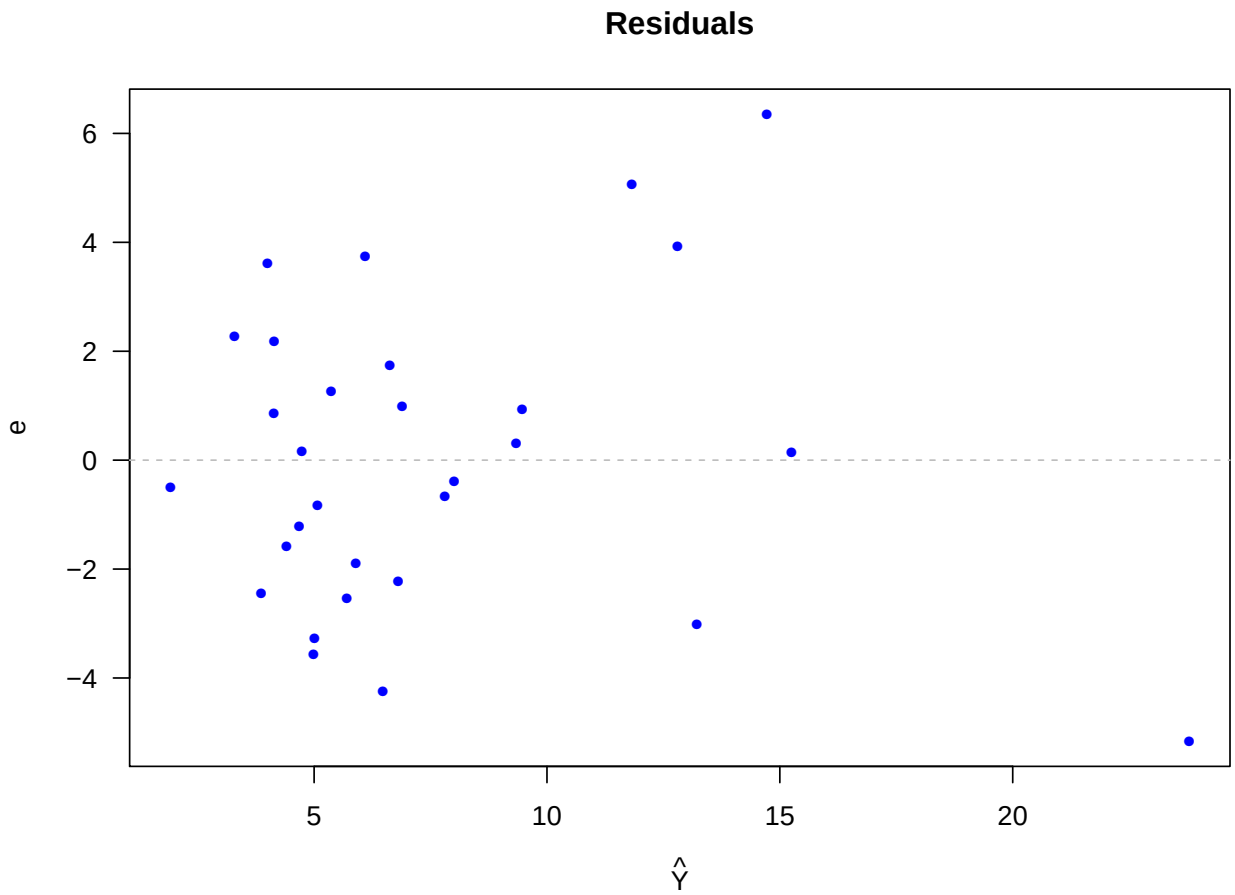
Residual Plot of `Species` ~ `Elev` + `Adj`



Spread increases with fitted values ⇒ violation of constant variance assumption

Residual Plot After Square Root Transformation

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



Variance is more stable $\Rightarrow$ transformation improves model assumptions

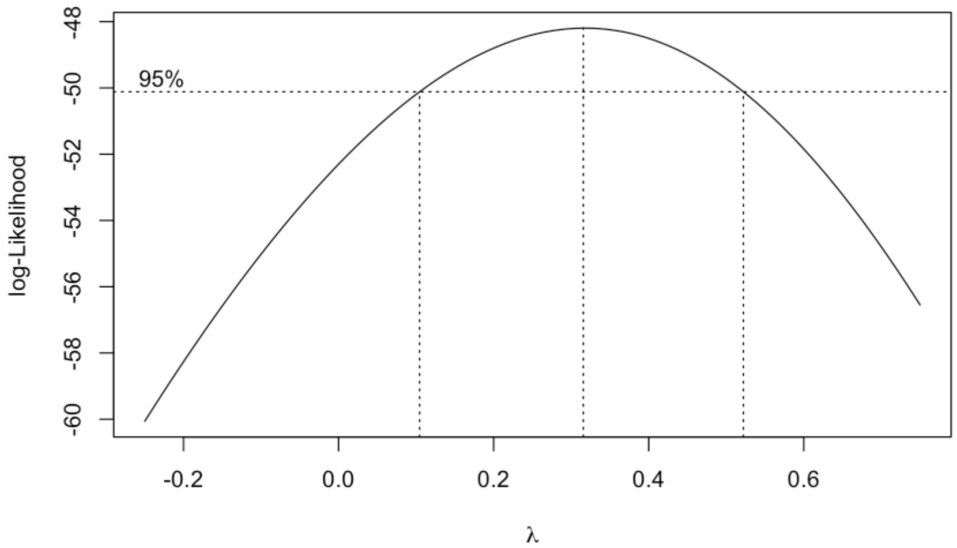
Box-Cox Transformation [Box and Cox, 1964]

Goal: Find transformation to stabilize variance

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

Interpreation:

- $\lambda = 1 \Rightarrow$ no transformation
- $\lambda = 0 \Rightarrow$ log transformation



In R, the `boxcox` function in the `MASS` package can be used to select a Box–Cox transformation. The plot suggests a cube-root transformation

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

MLR Model Selection and Diagnostics Workflow

1. Fit candidate models
2. Compare models using: C_p, R^2_{adj}, AIC/BIC
3. Select a model
4. Check diagnostics:
   - Residual plots
   - Normality
   - Influence
5. Address issues if needed:
   - Transformations
   - Remove or adjust variables

Key Idea: Model selection is not the final step — diagnostic checking is essential