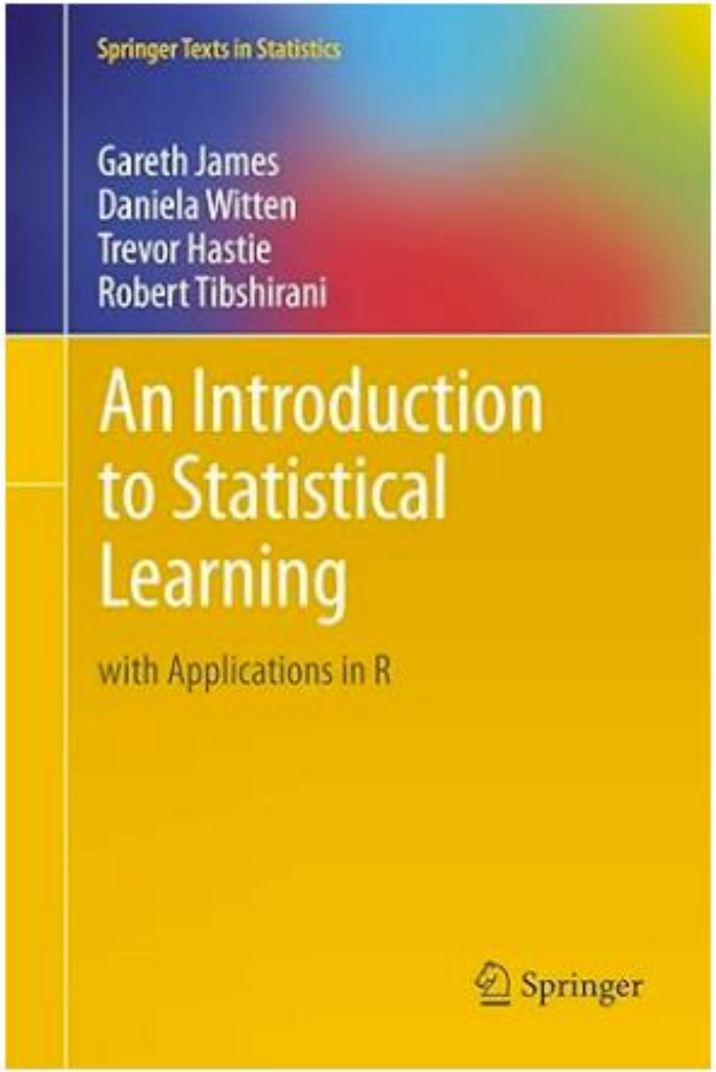
WEEK 07

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

SPRING 2025

AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN R



7.1
REGRESSION CRITISIM

ROLE OF ASSUMPTIONS

- **Simplify the complexity** by imposing constraints
  - E.g. Relationship between dependent and independent variable is linear
- This simplification accelerates our capabilities to analyze the data
- Whenever possible, the plausibility of assumptions for real world data needs to be evaluated.
- “Regression Criticism” is about questioning whether the model and its assumptions **truly fit *all* the data**.

WHY CHECKING ASSUMPTIONS MATTERS

- Even if a regression model **fits the sample data well**, we must ensure that it holds for the broader population.
- We need to evaluate whether our OLS (Ordinary Least Squares) results are trustworthy and **generalizable**.
- **Reasoning:** If assumptions are violated, estimates might be **biased** or **inefficient**.

RECAP: ORDINARY LEAST SQUARES (OLS)

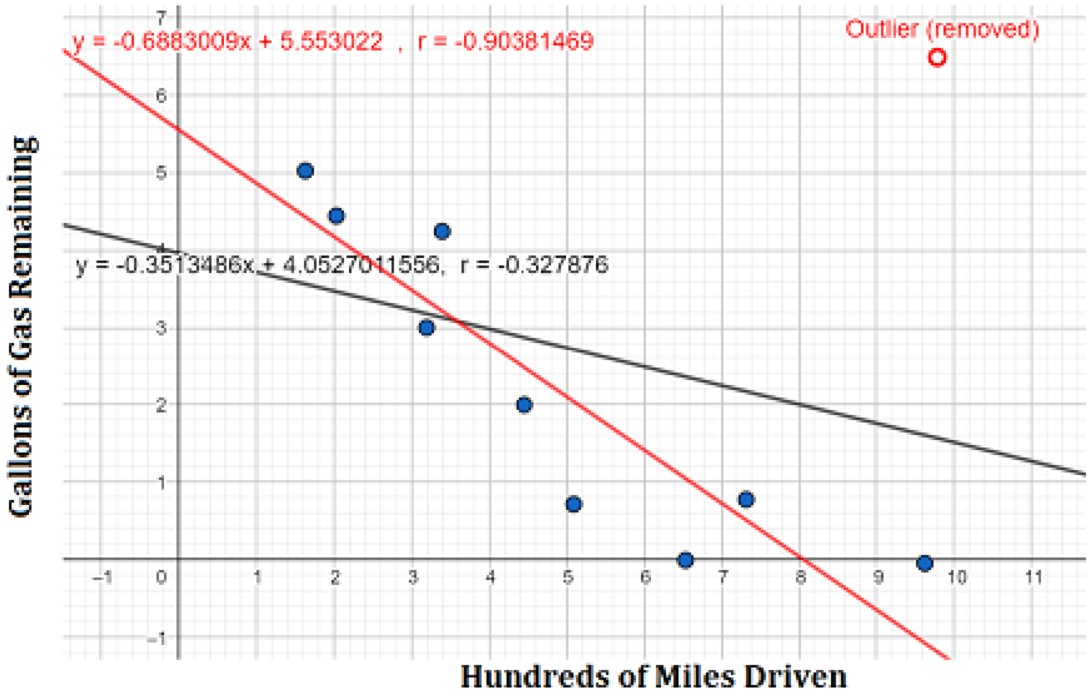
- Predicted Values

$$\hat{y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1}$$

Which $\hat{\beta}$ minimizes the sum of squared residuals, i.e.

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$

- Unusual data points ("outliers") can heavily affect the fit (Power effect)



ASSUMPTION - MULTICOLLINEARITY

- **Multicollinearity**
  - linear relationship among x_i
  - It causes larger standard error for β_j (coefficient estimate) and insignificant t-statistics.
  - **Difficulty interpreting** individual coefficient estimates

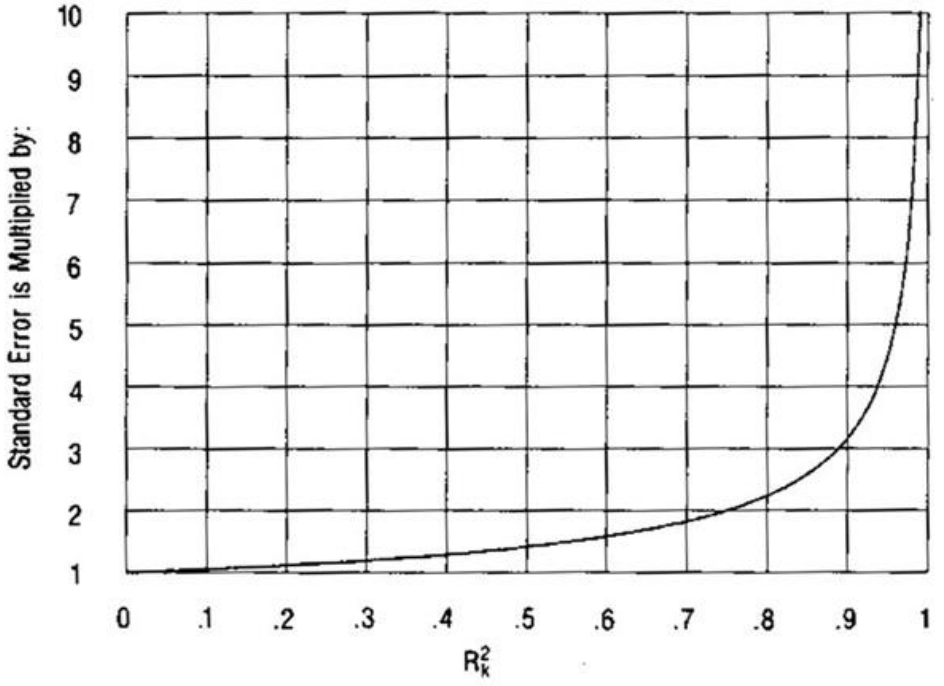


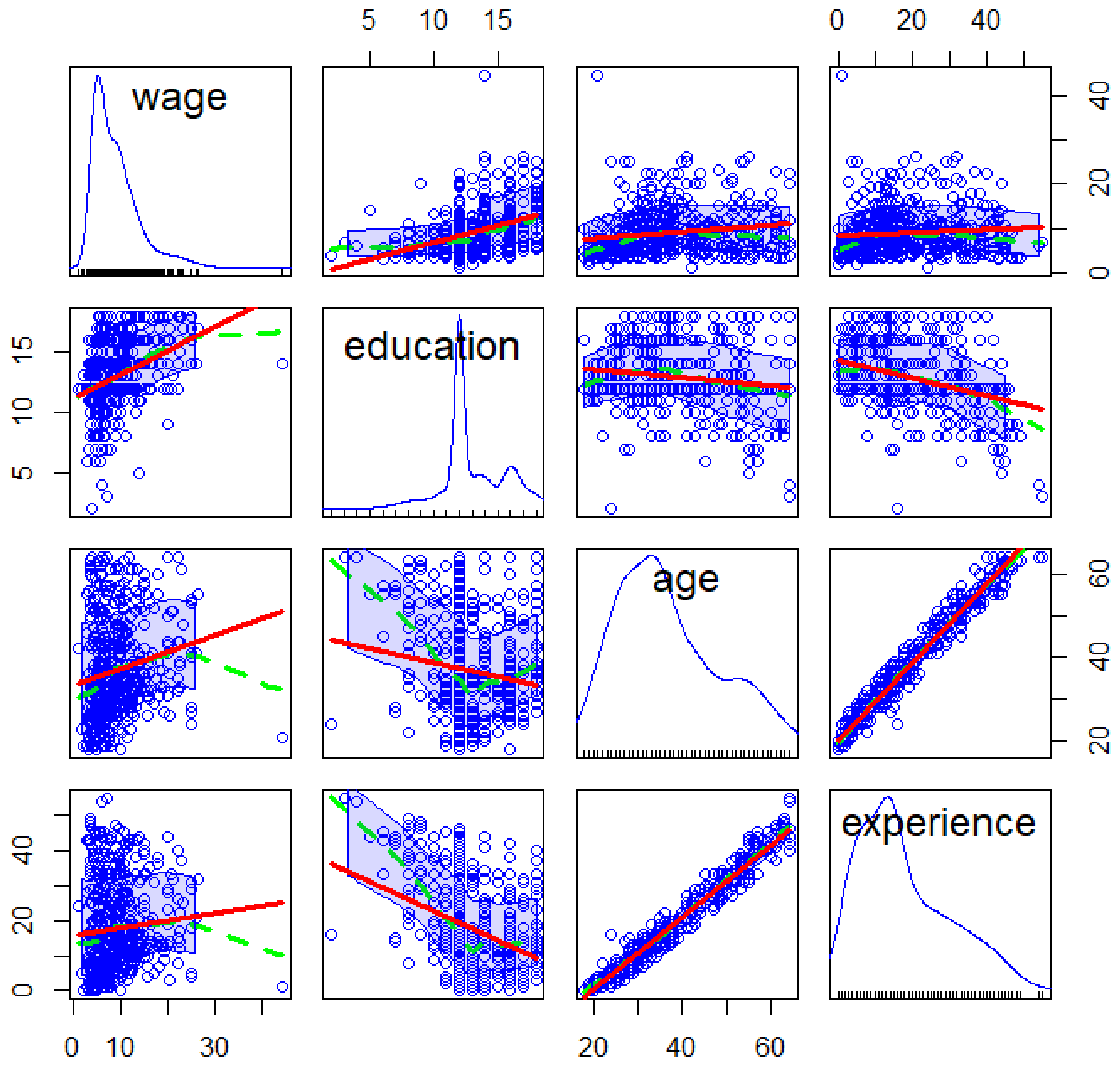
Figure 4.15 Effect of multicollinearity on standard errors (simplified).

HAMILTON, L. C. (1992). REGRESSION WITH GRAPHICS: A SECOND COURSE IN APPLIED STATISTICS.

ASSESS COLLINEARITY

A simple way to detect collinearity is to look at the correlation matrix of the x_i

Which variables look like have collinearity issue?



MODEL WITH MULTICOLLINEARITY AND WITHOUT MULTICOLLINEARITY

$wage \sim education + \boldsymbol{age} + experience$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.76987    7.04271  -0.677    0.499
education    0.94833    1.15524   0.821    0.412
experience   0.12756    1.15571   0.110    0.912
age         -0.02241    1.15475  -0.019    0.985

Residual standard error: 4.604 on 530 degrees of freedom
Multiple R-squared:  0.202,	Adjusted R-squared:  0.1975
F-statistic: 44.73 on 3 and 530 DF,  p-value: < 2.2e-16
```

$wage \sim education + experience$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.9045     1.2189  -4.024 6.56e-05 ***
education     0.9260     0.0814  11.375  < 2e-16 ***
experience    0.1051     0.0172   6.113 1.89e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.599 on 531 degrees of freedom
Multiple R-squared:  0.202,	Adjusted R-squared:  0.199
F-statistic: 67.22 on 2 and 531 DF,  p-value: < 2.2e-16
```

How the standard error change across two models?

How the significance of t-test change for the estimated parameters of the model?

ASSUMPTION: HOMOSCEDASTICITY

- Homoscedasticity: $Var(\varepsilon) = c$ (constant)
- Heteroscedasticity: $Var(\varepsilon) \neq c$ (constant)

The standard errors, confidence intervals, and hypothesis tests reply upon this assumption

- **Residuals versus Predicted Y Plots**
- If data do violate assumptions :
  - The variance of the error term (ε_i) is not constant across observations
  - Look for a "fan" or "cone" shape in Residual vs. Predicted Values Plot

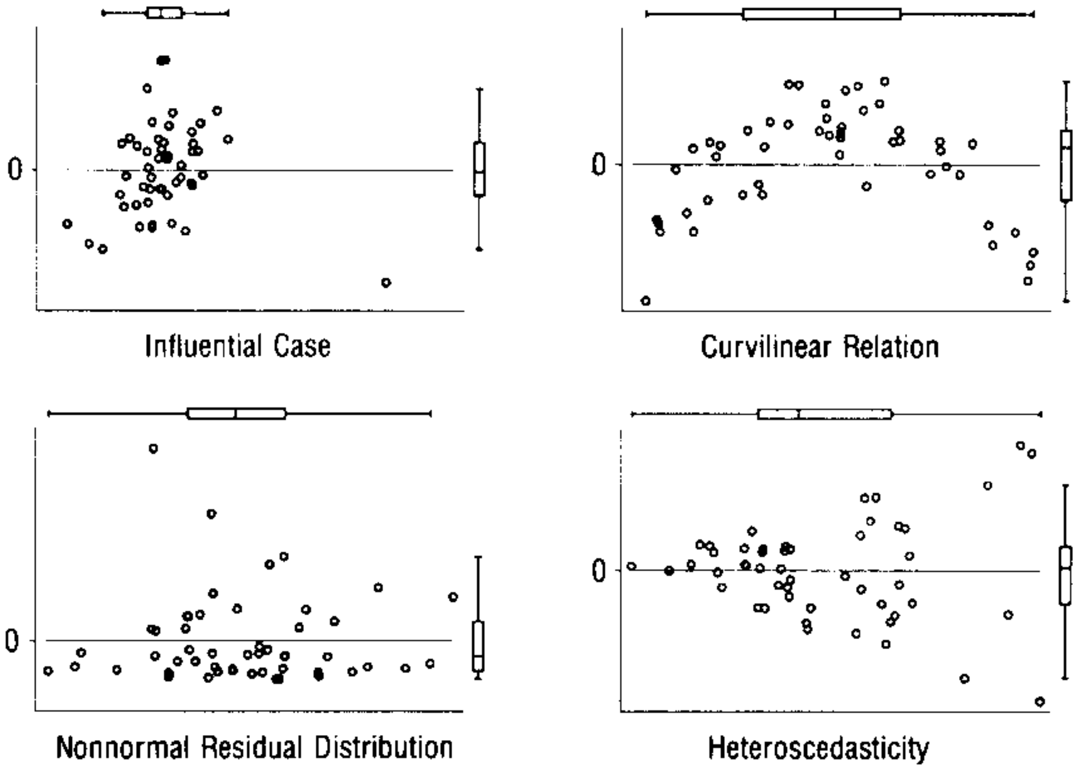
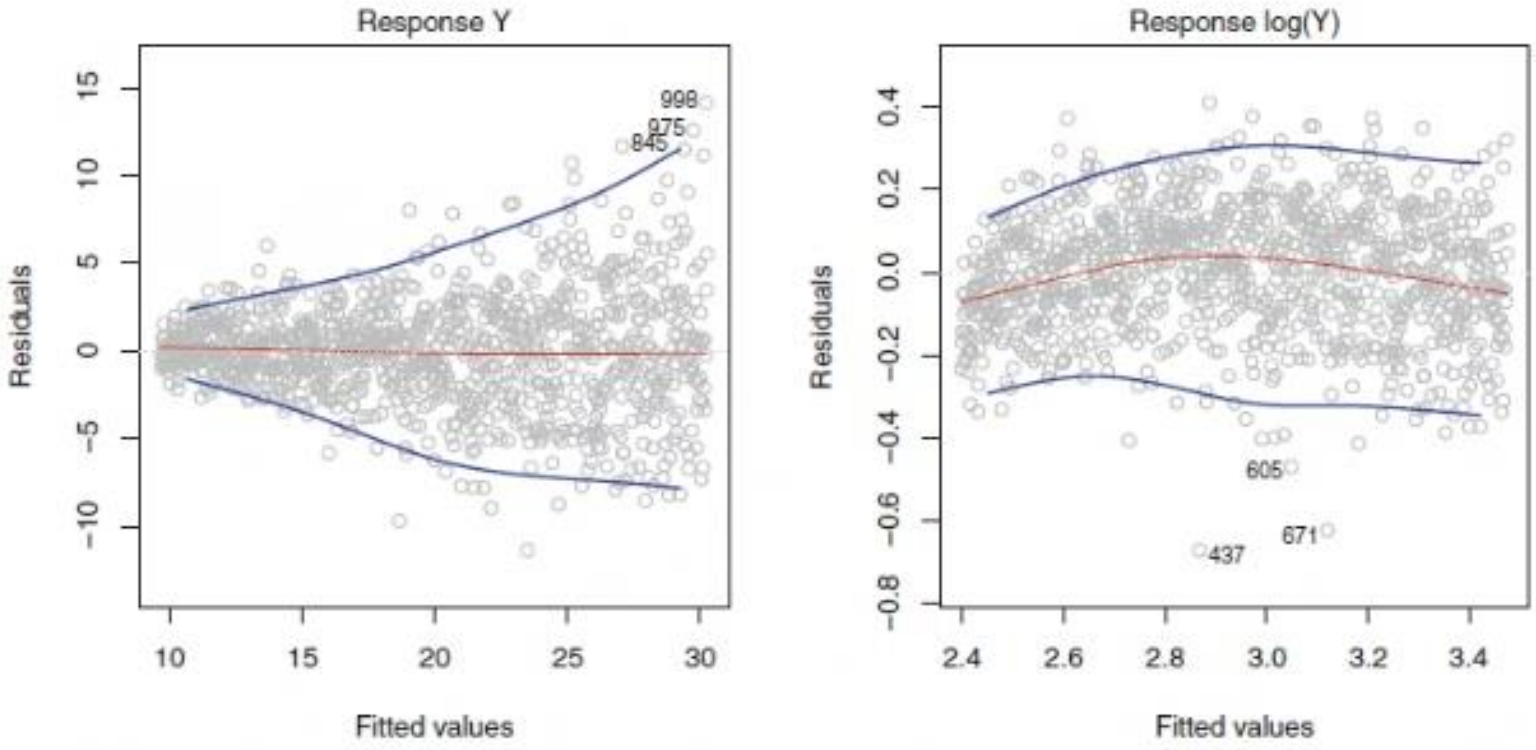


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

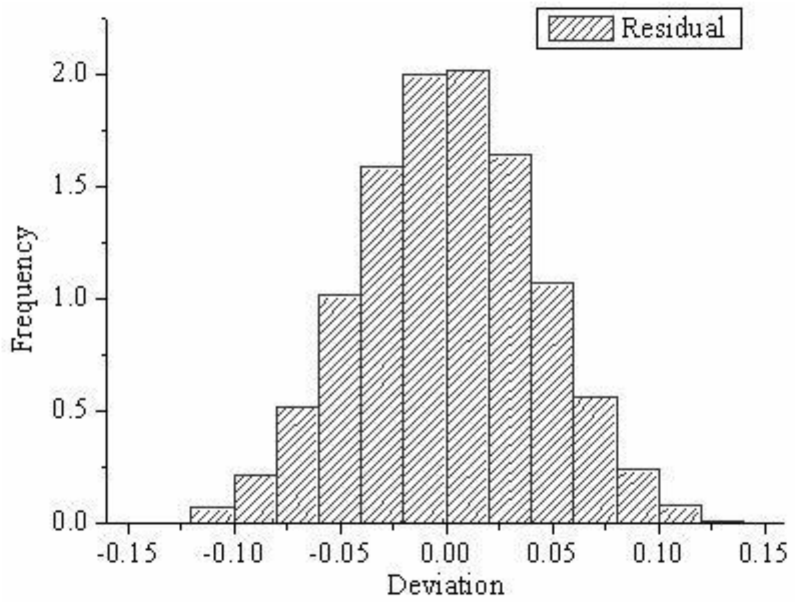
SOLUTION FOR HETEROSCEDASTICITY

- **Possible Remedies**
  - Transform the Data (e.g., $Log(y)$)
    - Logarithmic or other functional transformations can stabilize variance if the relationship is multiplicative

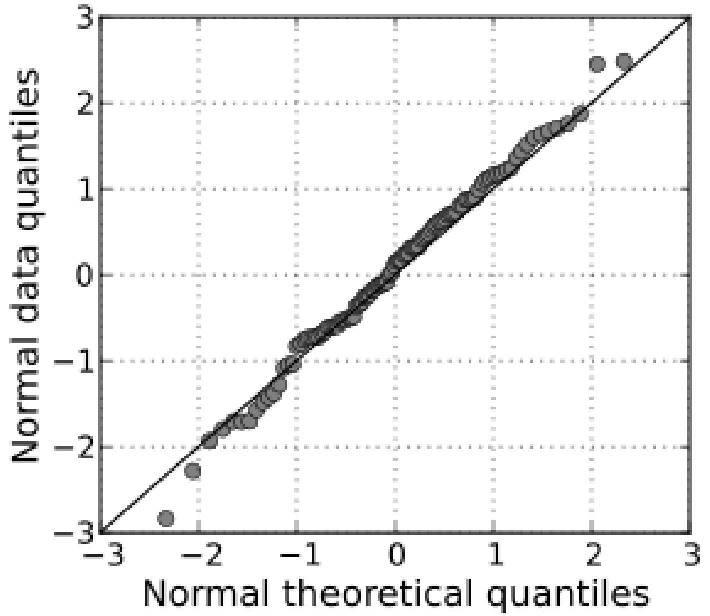


ASSUMPTION – NORMALITY OF THE ERROR

- **Residual Histogram**
  - centered around zero, roughly bell-shaped
  - Severe skewness or heavy tails can invalidate standard inference methods.
  - Distribution of small sample is often not normal

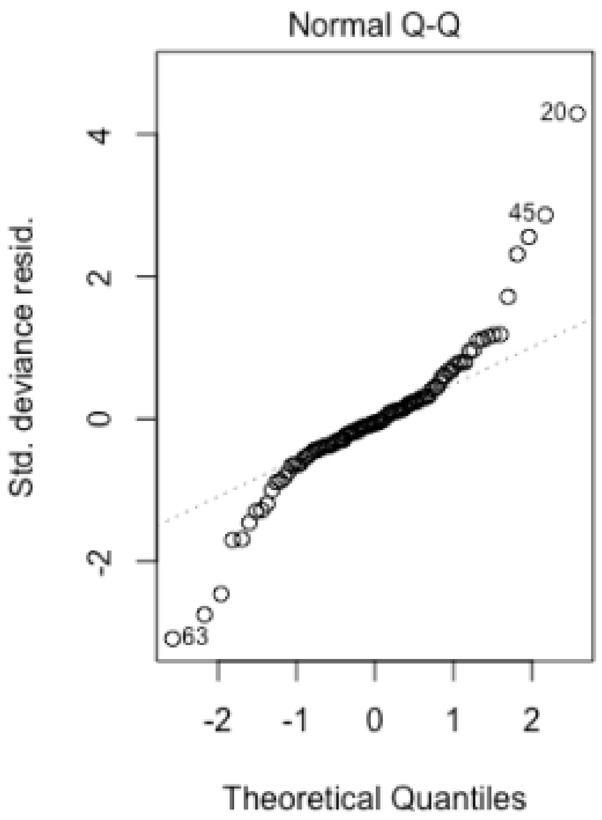


- **Q–Q Plot (Quantile–Quantile Plot)**
  - Plots the quantiles of residuals against the quantiles of a normal distribution
  - If points lie on or near the 45-degree line, the residual distribution is approximately normal
  - Deviations (e.g., “S” shape) can indicate skewness or kurtosis.

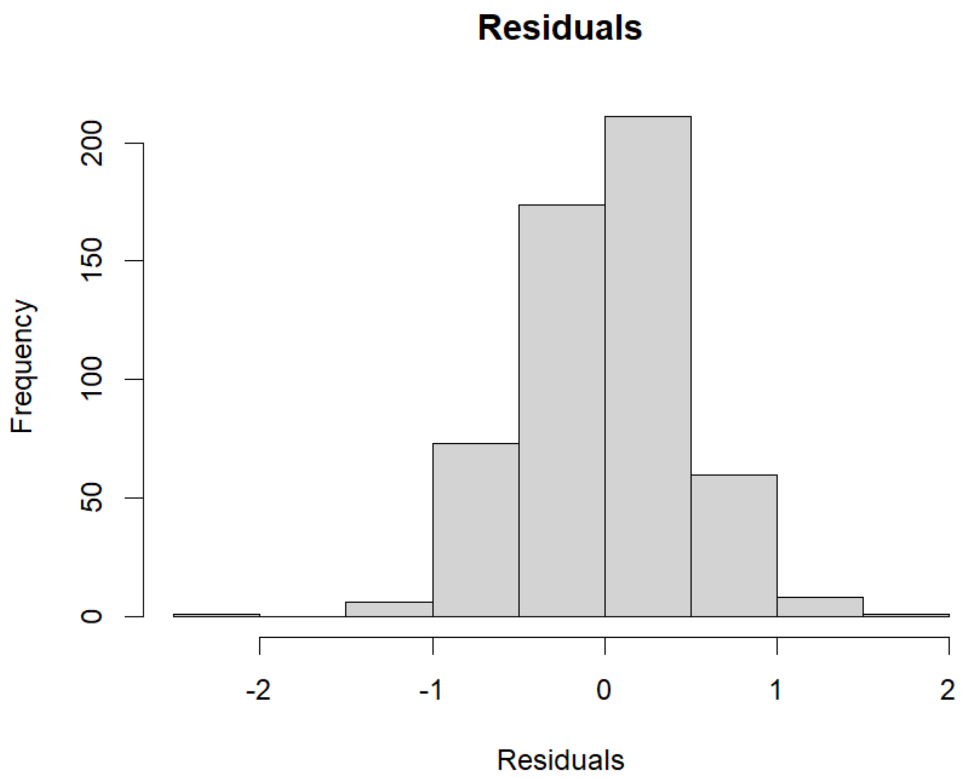


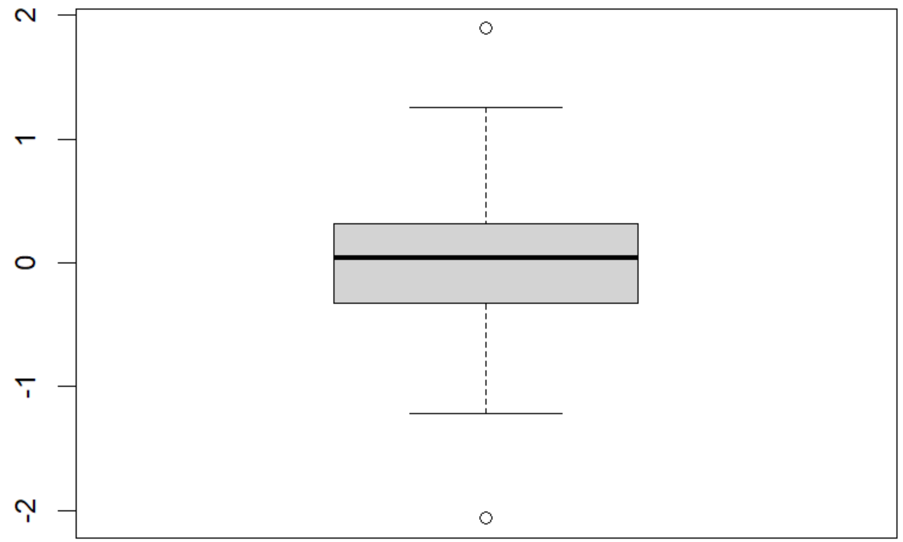
SOLUTION FOR NON-NORMALITY

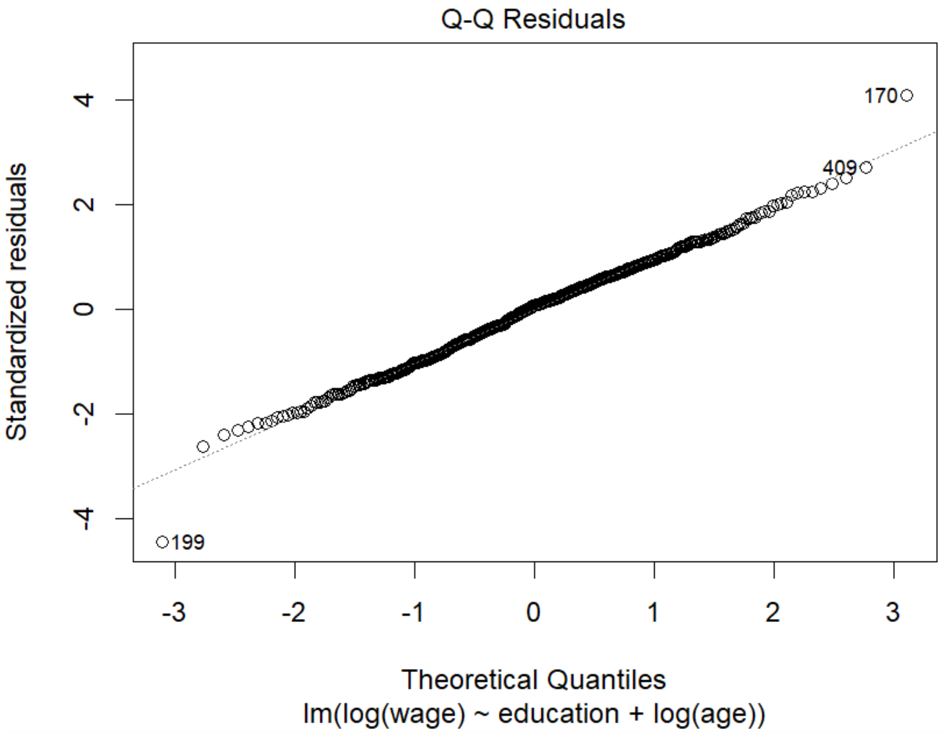
- **Possible Remedies**
  - Transform the Data (e.g., $Log(y)$)
    - Can reduce skewness and stabilize variance.
  - Bootstrap for More Accurate Standard Errors
    - Especially useful for smaller samples or when distributional assumptions are in doubt
    - Resampling techniques can provide inference that does not rely on strict normality assumptions.
      - Generate Bootstrap simulations to obtain the distribution of the estimated parameters



CHECK THE NORMALITY ASSUMPTION



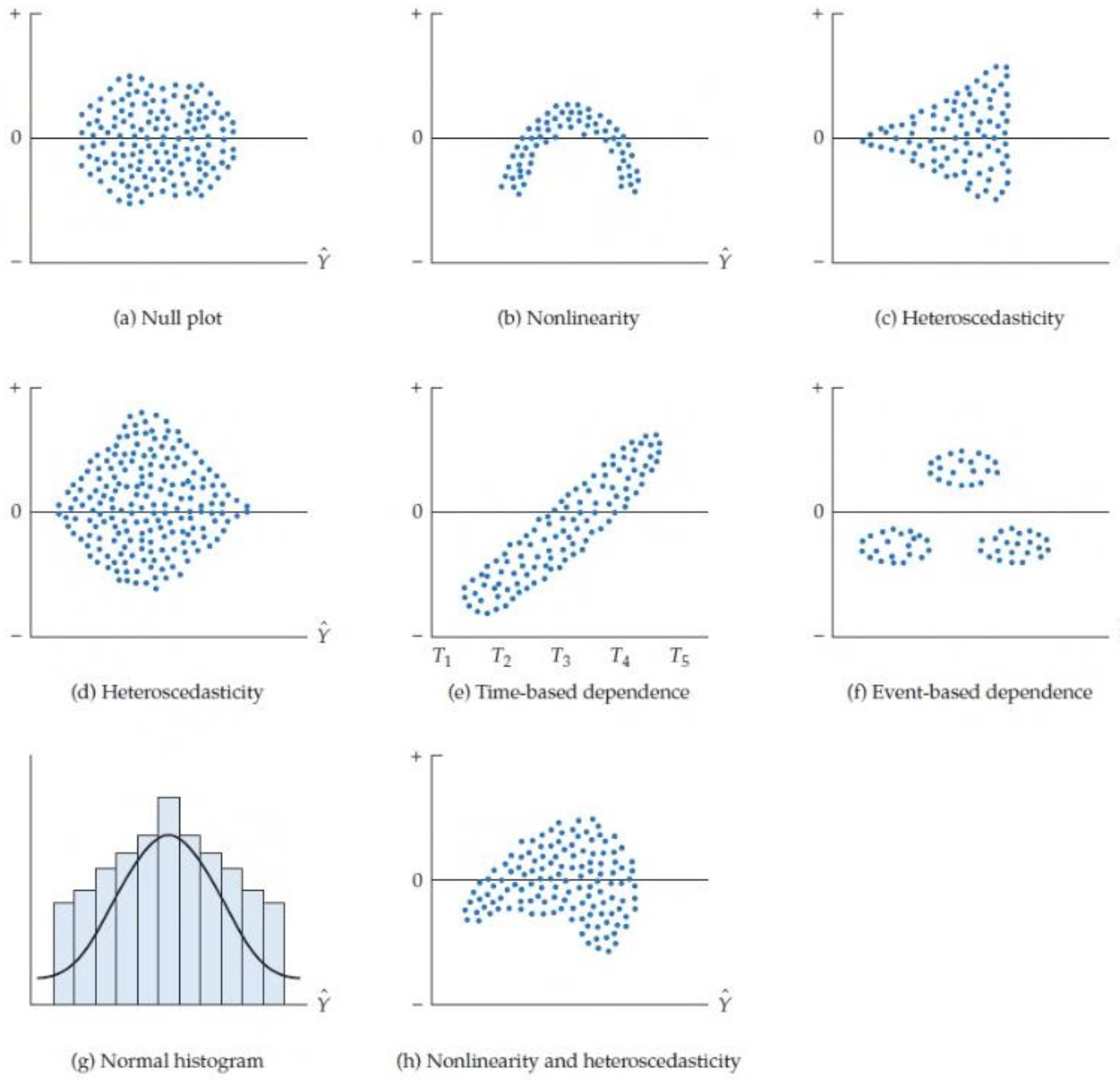




ASSUMPTION – LINEARITY

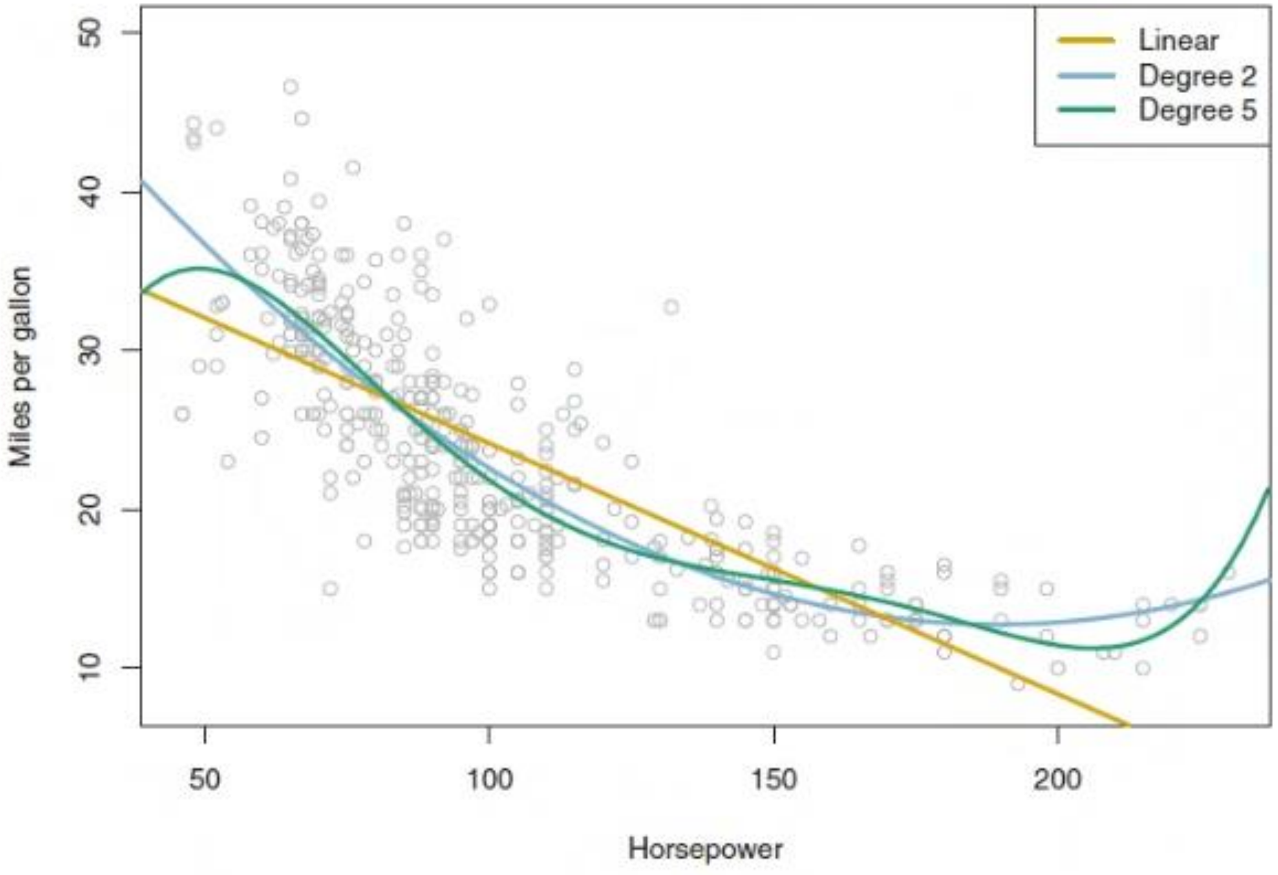
BASIC RESIDUAL PLOT

- **Linear Relationship**: The core premise of multiple linear regression is the existence of a linear relationship between the dependent (outcome) variable and the independent variables
- Linearity of any bivariate relationship is easily examined through **residual plots.**



SOLUTION FOR NON-LINEARITY

POLYNOMIAL REGRESSION



Linear model (R^2 =0.606):

$$mpg = \beta_0 + \beta_1 * horsepower + \ \varepsilon$$

Degree 2 model ($R^2 = 0.688$):

$$mpg = \beta_0 + \beta_1 * horsepower + \beta_1 * horsepower^2 + \ \varepsilon$$

Degree 5 model (not recommend):

$$mpg = \beta_0 + \beta_1 * horsepower + \beta_2 * horsepower^2 + \beta_3 * horsepower^3 + \beta_4 * horsepower^4 + \beta_5 * horsepower^5 + \ \varepsilon$$

SOLUTION FOR NON-LINEARITY

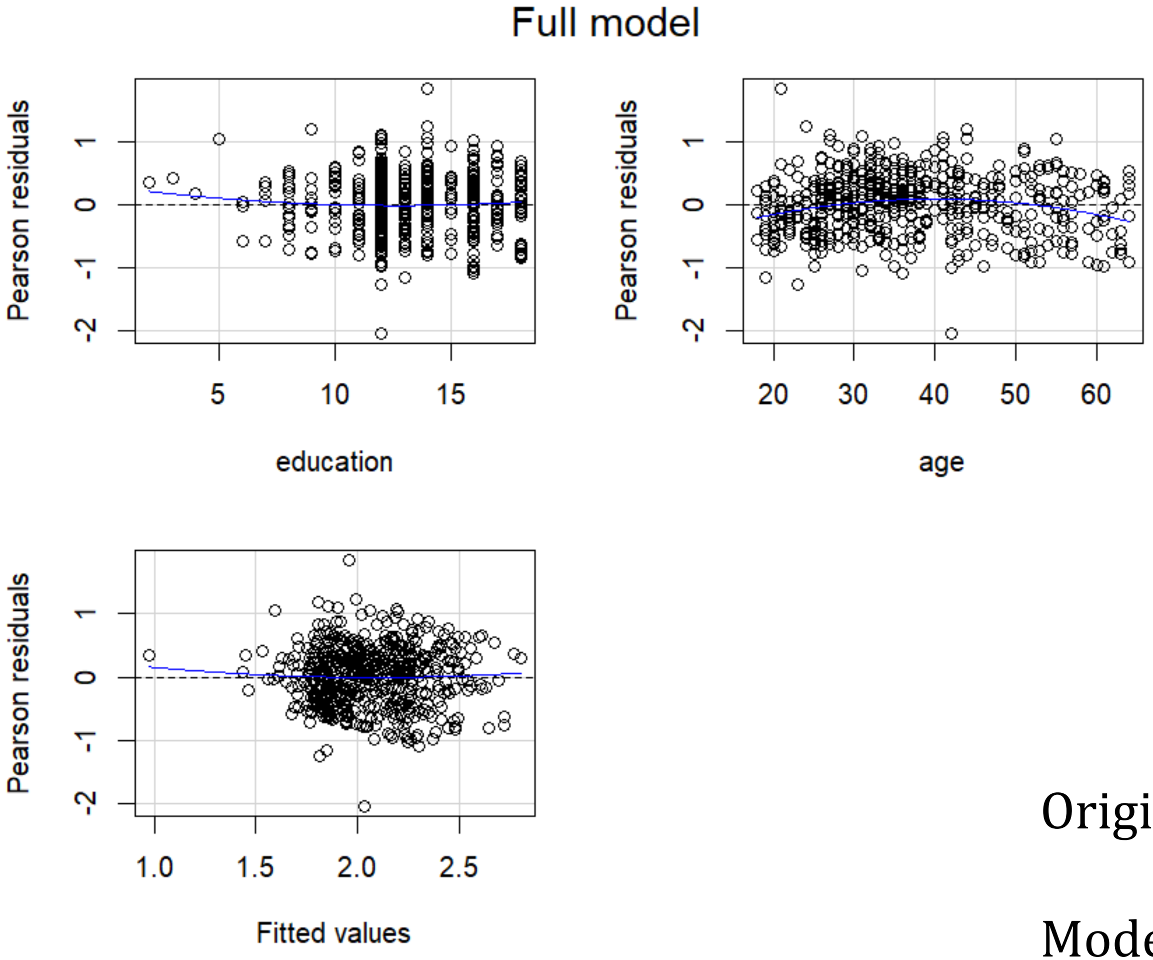
- Plots of residuals versus fitted values and versus each of the regressors in turn are the most basic diagnostic graphs.

- The Tukey test assesses whether including a squared term of an independent variable, already present in the model, enhances model fit.

  - It is implemented in the function **car::residualPlots()**, which:
  - Generates a plot comparing the quadratic function to the residuals.
  - Conducts a t-test to determine the significance of the quadratic term.

ASSUMPTION – LINEARITY

RESIDUAL PLOT



Test for education:

H_0: There is no evidence of nonlinearity.
H_1: There is a evidence of nonlinearity

```
> car::residualPlots(full_model, main="Full model")
           Test stat Pr(>|Test stat|)
education     1.1086           0.2681
age          -4.5541        6.533e-06 ***
Tukey test    0.6012           0.5477
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

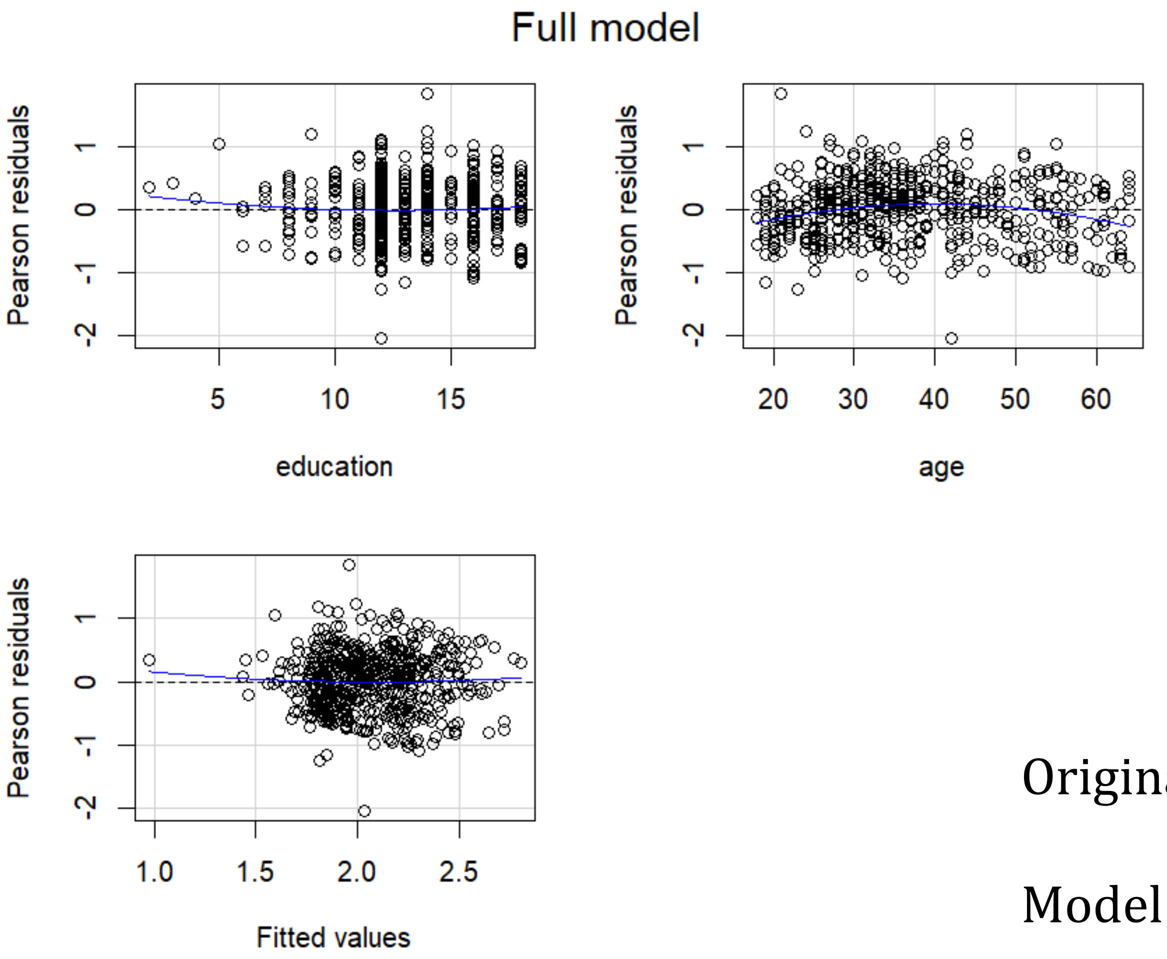
Original full model: $\log(wage) = educaton + age$

Model with quadratic term :

$$\log(wage) = educaton + education^2 + \ \ age$$

ASSUMPTION – LINEARITY

RESIDUAL PLOT



Test for age:

H_0: There is no evidence of nonlinearity.
H_1: There is a evidence of nonlinearity

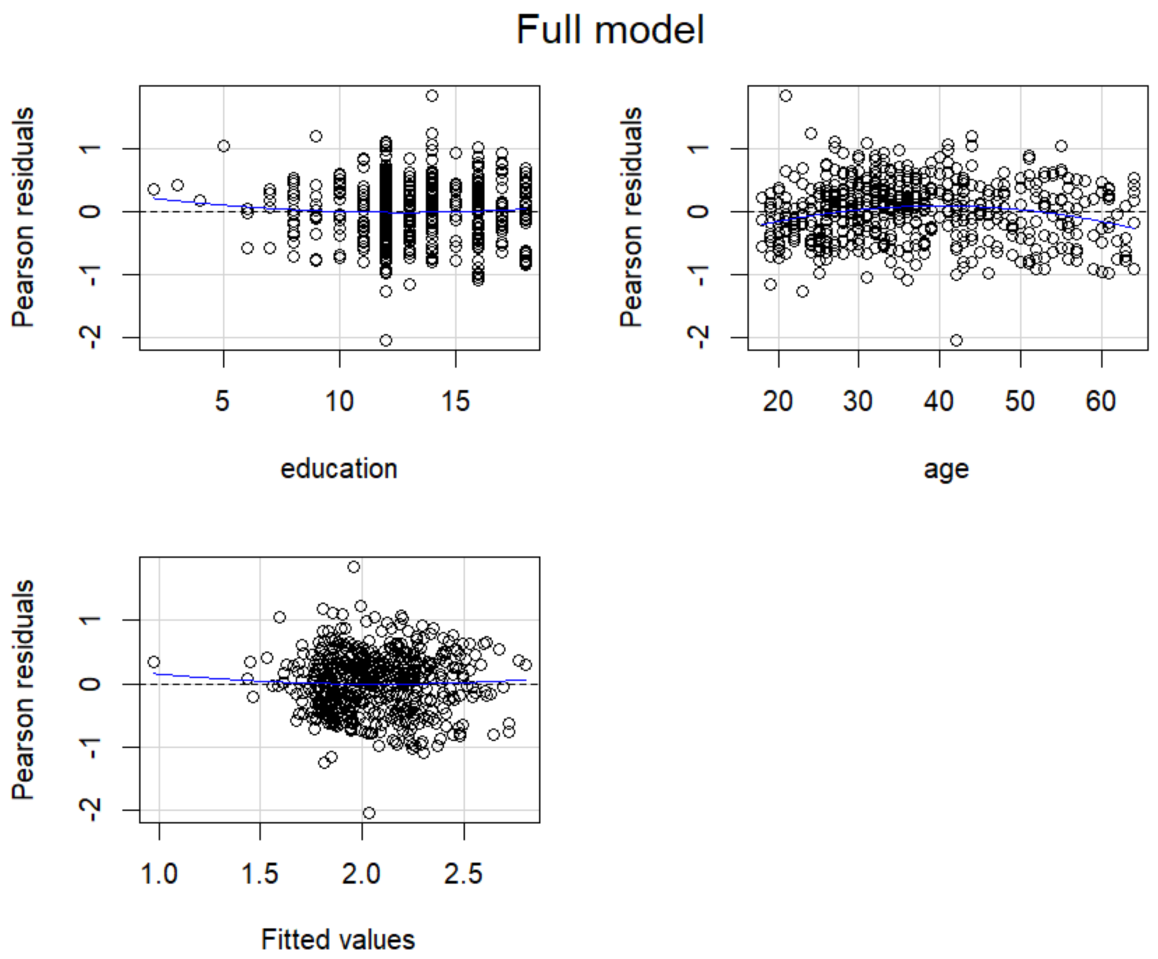
```
> car::residualPlots(full_model, main="Full model")
           Test stat Pr(>|Test stat|)
education     1.1086           0.2681
age          -4.5541        6.533e-06 ***
Tukey test    0.6012           0.5477
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Original full model: $\log(wage) = educaton + age$

Model with quadratic term :

$\log(wage) = educaton + age + age^2$

NON-LINEARITY




```
> car::residualPlots(full_model, main="Full model")
           Test stat Pr(>|Test stat|)
education     1.1086           0.2681
age          -4.5541        6.533e-06 ***
Tukey test    0.6012           0.5477
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Original model: $\log(wage) = educaton + age$

Model with quadratic term：

$$\log(wage) = educaton + education^2 + \ age + age^2$$

ANOVA

- ANOVA (Analysis of Variance) is a statistical method used to compare multiple models by evaluating whether a **simpler model** fits the data significantly worse than a **more complex model**. This helps determine whether adding variables (or transforming them) improves the model's explanatory power.
- In regression analysis, ANOVA is commonly used to compare:
  - **Nested models**, where one model is a subset of another.
  - **Different functional forms**, such as linear vs. quadratic models.

ANOVA RESULT

```
> anova(full_model, model_quadratic)
Analysis of Variance Table

Model 1: log(wage) ~ education + age
Model 2: log(wage) ~ education + age + I(age^2)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    531 117.07
2    530 112.66  1    4.4086 20.74 6.533e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

OUTLIERS

- Identify additional explanatory variables that may clarify why certain cases appear extreme.
- Address measurement errors by:
  a) (a) Refining or correcting the measurement,
  b) (b) Assigning a lower weight to the affected case (Hamilton, Chapter 6), or
  c) (c) Removing the case if necessary.
- Assess population differences:
  a) If an observation belongs to a different population, either exclude it or account for structural differences using methods such as dummy variables or interaction terms.
- Consider data transformation to stabilize variance and improve model fit.

CHECK OUTLIERS – RESIDUAL PLOT

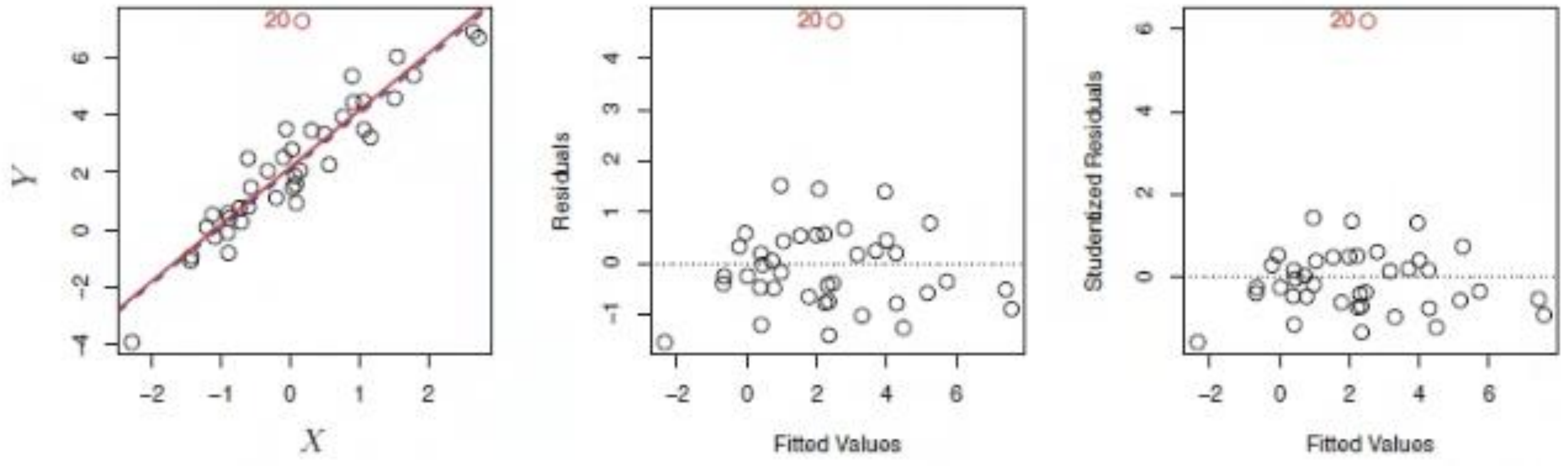


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

R^2 increase from 0.805 to 0.892 after we exclude the outlier (observation 20)

STUDENTIZED RESIDUALS

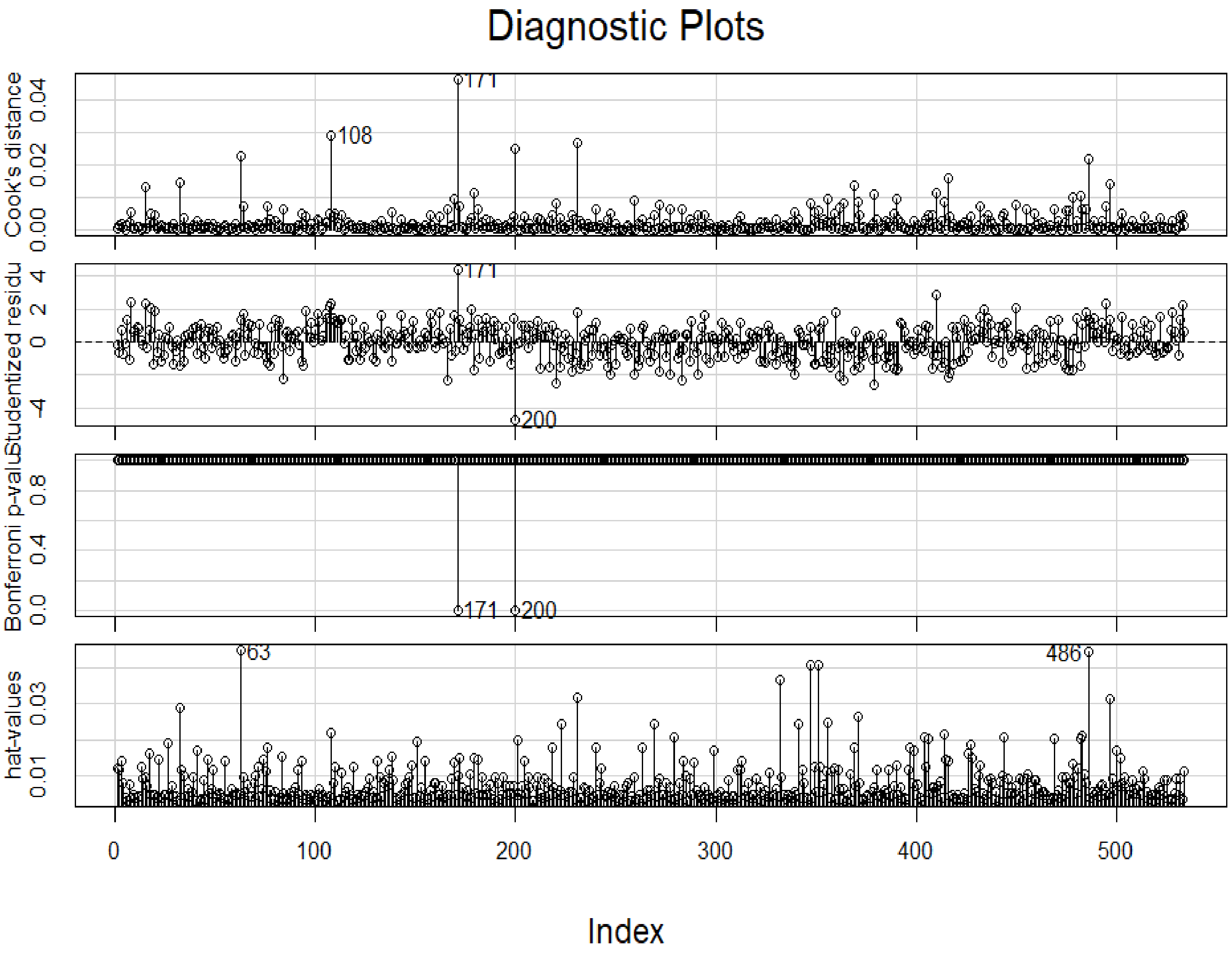
- The key idea behind studentized residuals is to delete each observation one at a time, refit the regression model on the remaining $n - 1$ observations, and compute the deleted residuals. These deleted residuals are then standardized, resulting in studentized residuals.

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

- If an observation has a studentized residual that is **larger than 3 (in absolute value)** we can call it an outlier.

STUDENTIZED RESIDUALS

	wage	education	age
108	14.0	5	55
171	44.5	14	21
200	1.0	12	42
63	7.0	3	64
486	22.2	18	64



COOK'S DISTANCE

- Cook's Distance measures how much removing an observation **changes** the estimated regression coefficients.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{ps^2}$$

a) $D_i > 1$: The observation may be influential.

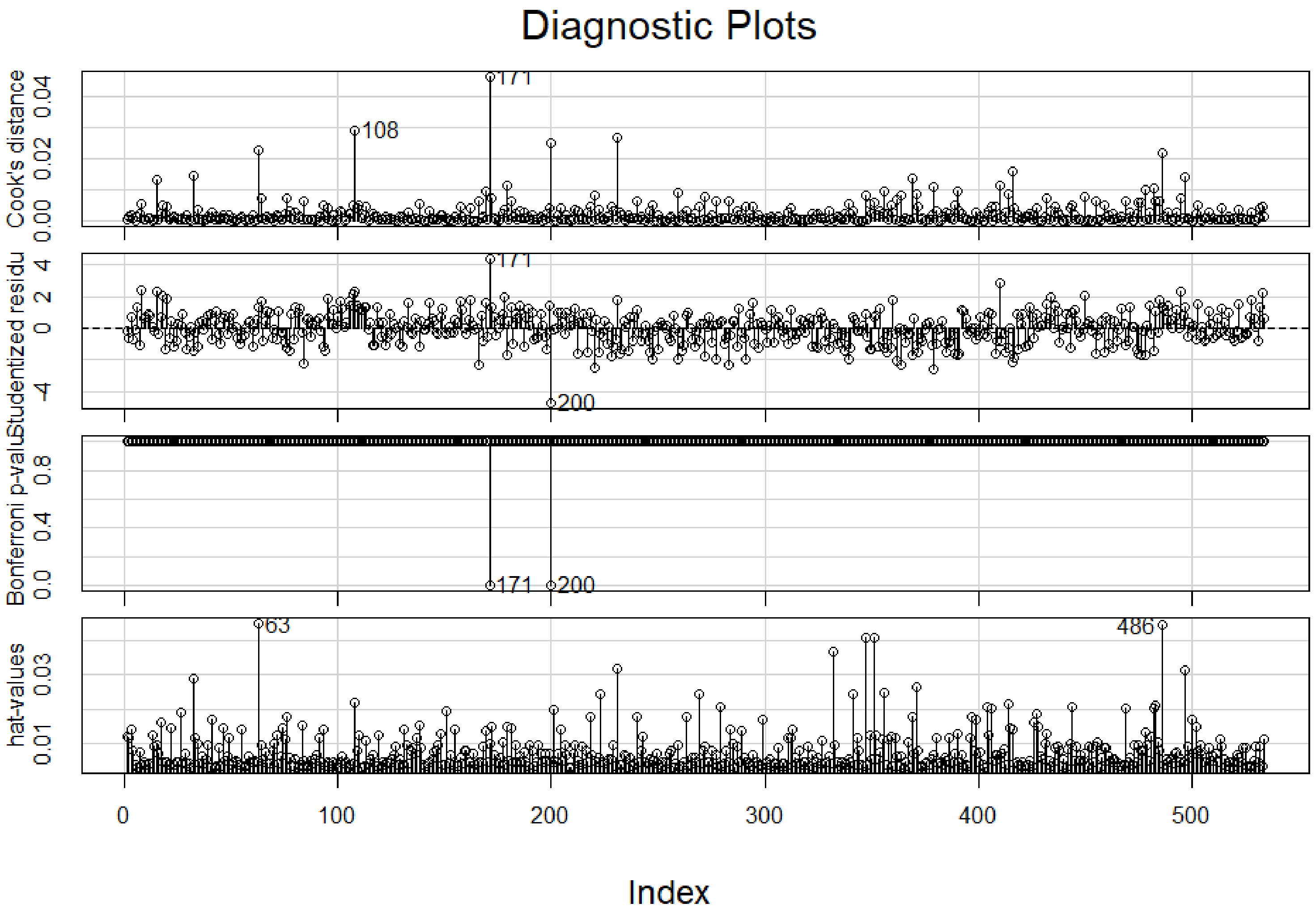
b) Higher values indicate larger impact on model estimates.

1. COOK, R. D. W. S. (1982). RESIDUALS AND INFLUENCE IN REGRESSION.

COOK'S DISTANCE

$$\log(wage) \sim education + age + age^2$$

	wage	education	age
108	14.0	5	55
171	44.5	14	21
200	1.0	12	42
63	7.0	3	64
486	22.2	18	64

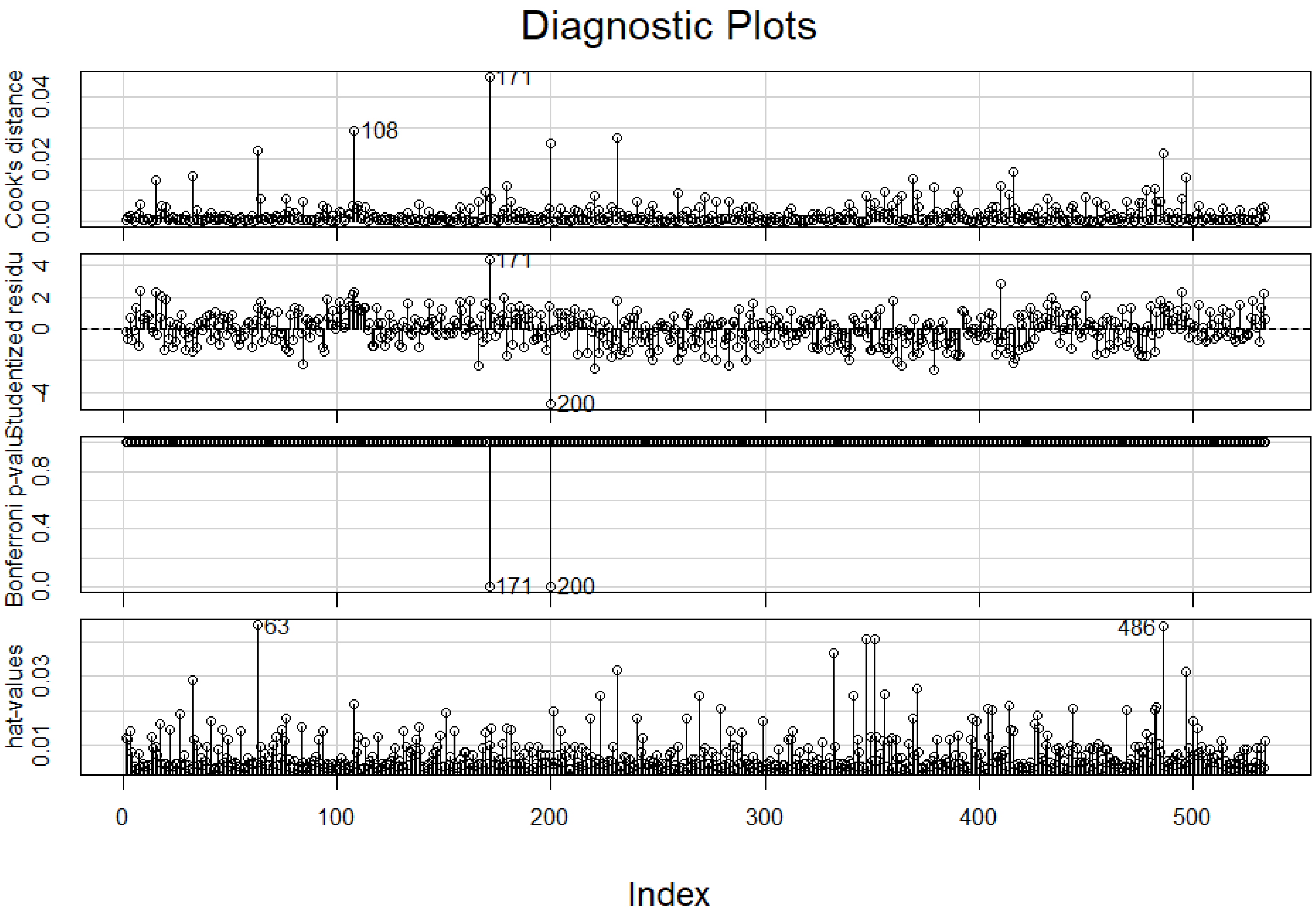


BONFERRONI P-VALUE

- The **Bonferroni correction** adjusts p-values when testing multiple hypotheses to reduce the risk of false positives.
  a) It tests whether an individual **studentized residual** is **statistically significant** after correcting for multiple comparisons.
  b) Used to **flag extreme residuals** that are unlikely due to random chance.
- **If Bonferroni p-value <0.05** → The observation is **significantly different** and may be an **outlier**.

BONFERRONI P-VALUE

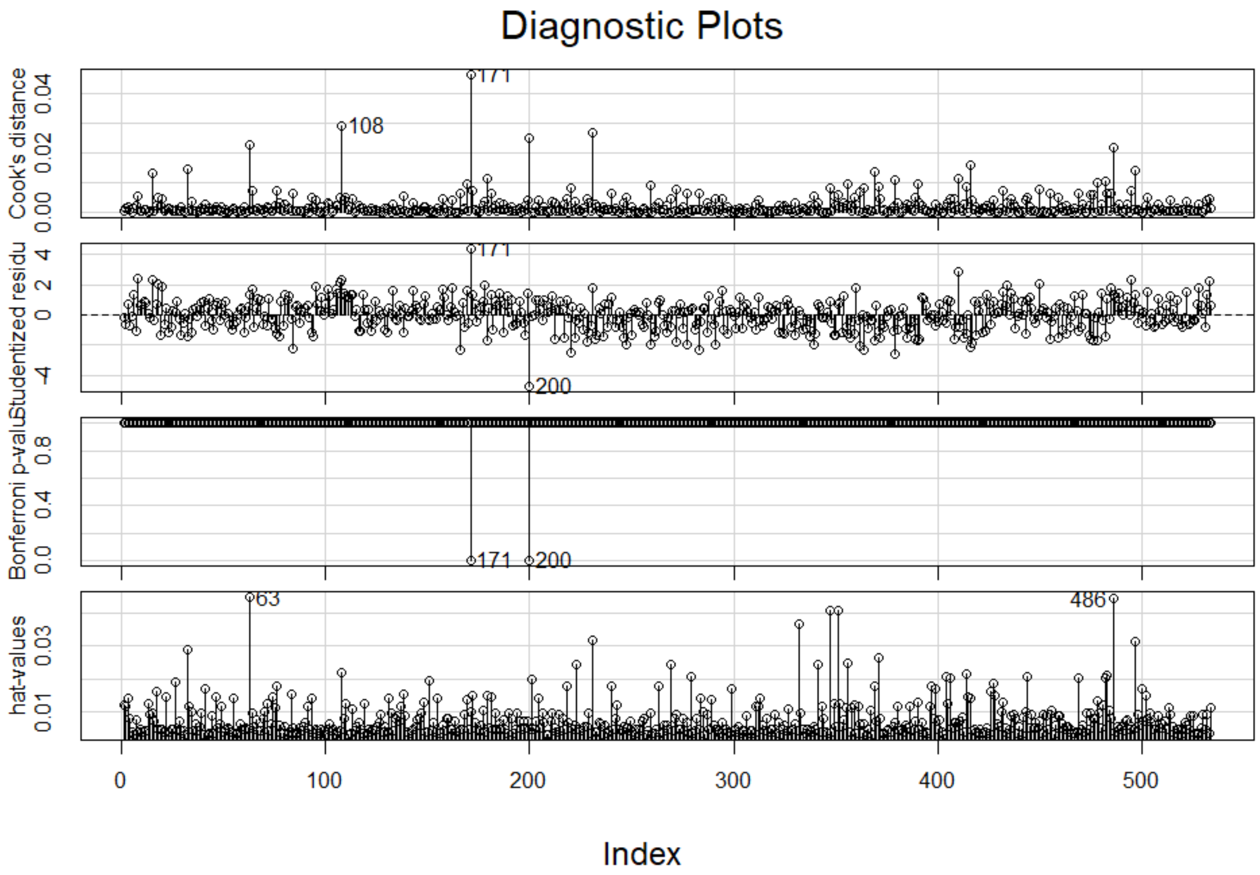
	wage	education	age
108	14.0	5	55
171	44.5	14	21
200	1.0	12	42
63	7.0	3	64
486	22.2	18	64



HAT-VALUE(h_i) - LEVERAGE

- Hat-values measure **leverage**, observation with $high\ leverage$ have an unusual value for x_i.
- Points with high leverage **pull the regression line**, possibly distorting estimates.

	wage	education	age
108	14.0	5	55
171	44.5	14	21
200	1.0	12	42
63	7.0	3	64
486	22.2	18	64



WEEK 03

CODE DEMO SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025