SLR: Checking Assumptions and Transformations

Keegan Korthauer

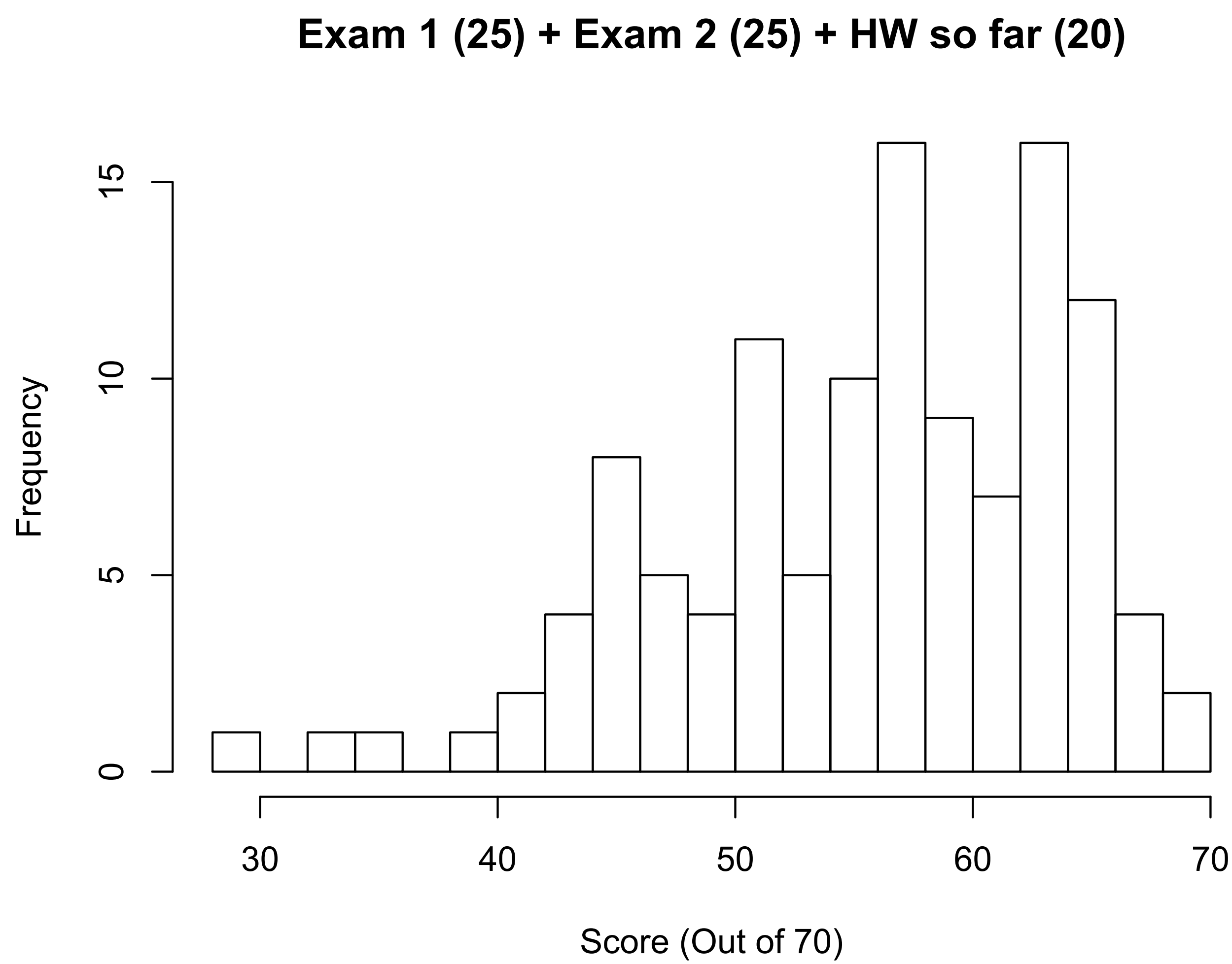
Department of Statistics

UW Madison

Exam 2 Summary Stats

- Mean: 42.9 (85.7%)
- Median: 44.5 (89%)
- Standard deviation: 5.7 (11.5%)
- Most missed questions:
  - **Problem 1**: What is a p-value and general form of CI
  - **Problem 6b**: Stating the null/alternative hypotheses for a Chi-square test of multinomial trial
  - **Throughout**: Forgetting to check assumptions

Histogram of Scores so Far



Unofficial* Letter Grades So Far

Possible points = 25 (Exam 1) + 25 (Exam 2) + 20 (Average of Homework 1-9) = 70

Percentage (Points divided by 70)	Score (Out of 70 Points)	Tentative Letter Grade
90.5% or higher	63.4 or higher	A
[85% – 90.5%)	[59.5 – 63.4)	AB
[78% – 85%)	[54.5 – 59.5)	B
[73% – 78%)	[51 – 54.5)	BC
[65% – 73%)	[45.5 – 51)	C
[57% – 75%)	[40 –45.5)	D
below 57%	below 40	F

***Any official curve will depend on overall final exam performance**

Recap –Simple Linear Regression

- The simple linear regression model assumes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- The least-squares line is:

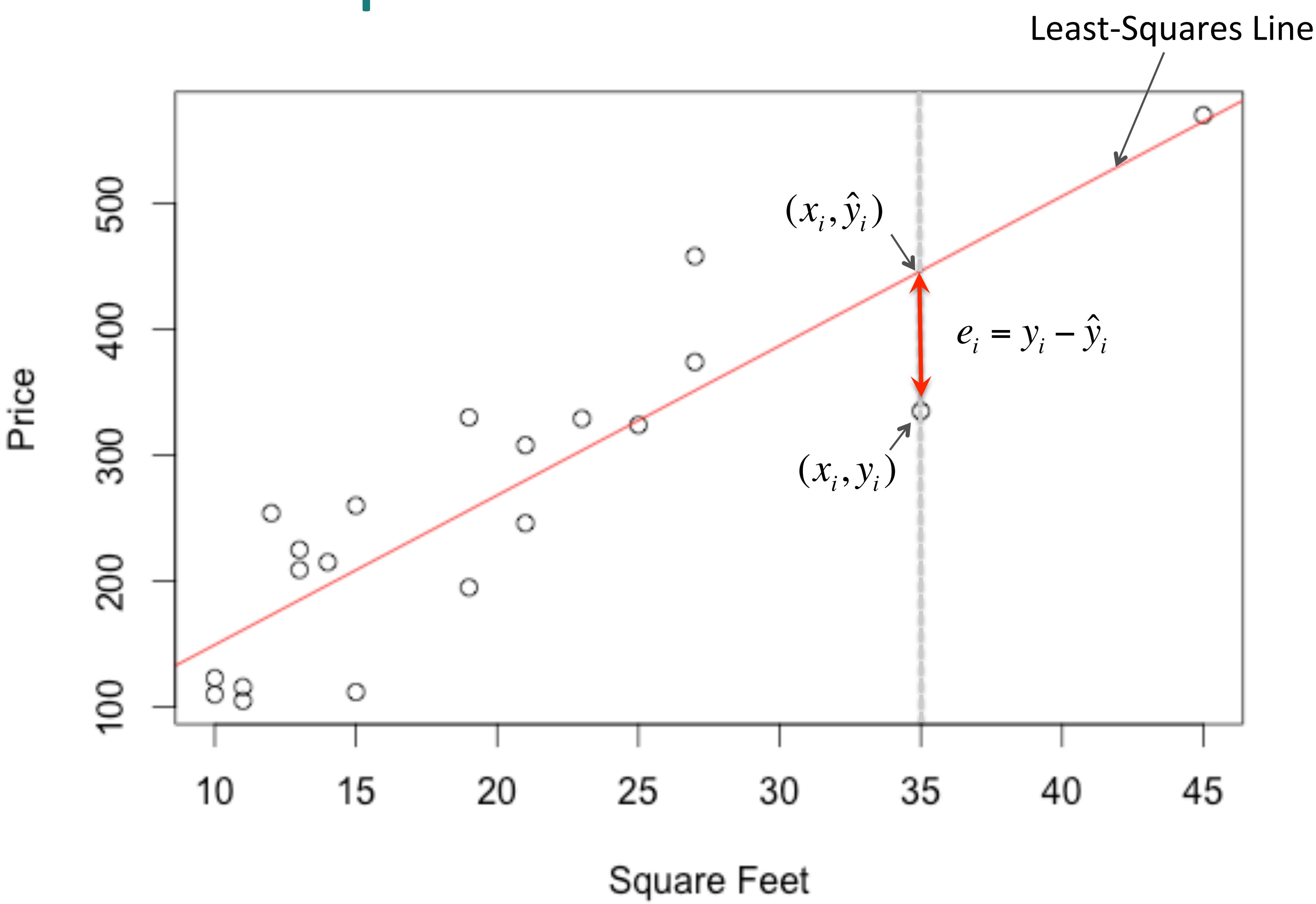
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- Only applies when relationship is **linear**
- Be wary of extrapolation

Least-Squares Line Minimizes SSE



Recap - Assumptions for Errors in Linear Models

1. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **random** and **independent**. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}

2. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have **mean 0**

3. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have the **same variance** denoted by σ^2

4. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **normally distributed**

Questions to Answer Today

1. How do we check the model assumptions?

2. What can we do if the relationship between x and y is not linear?

3. What are outliers and influential points? How do we deal with them?

DIAGNOSTIC PLOTS FOR CHECKING ASSUMPTIONS

Residual plot

Q-Q plot

Residual Plot

- Plot of fitted values versus residuals
  - Used to check assumption 3
- When the linear model is valid and assumptions are satisfied, the plot will show no substantial trend and no heteroscedasticity (unequal variance)
  - There should be no curve to the plot, and the vertical spread of the points should not vary too much over the range of fitted values
- A good residual plot does not by itself prove that the linear model is appropriate

A "Good" Residual Plot

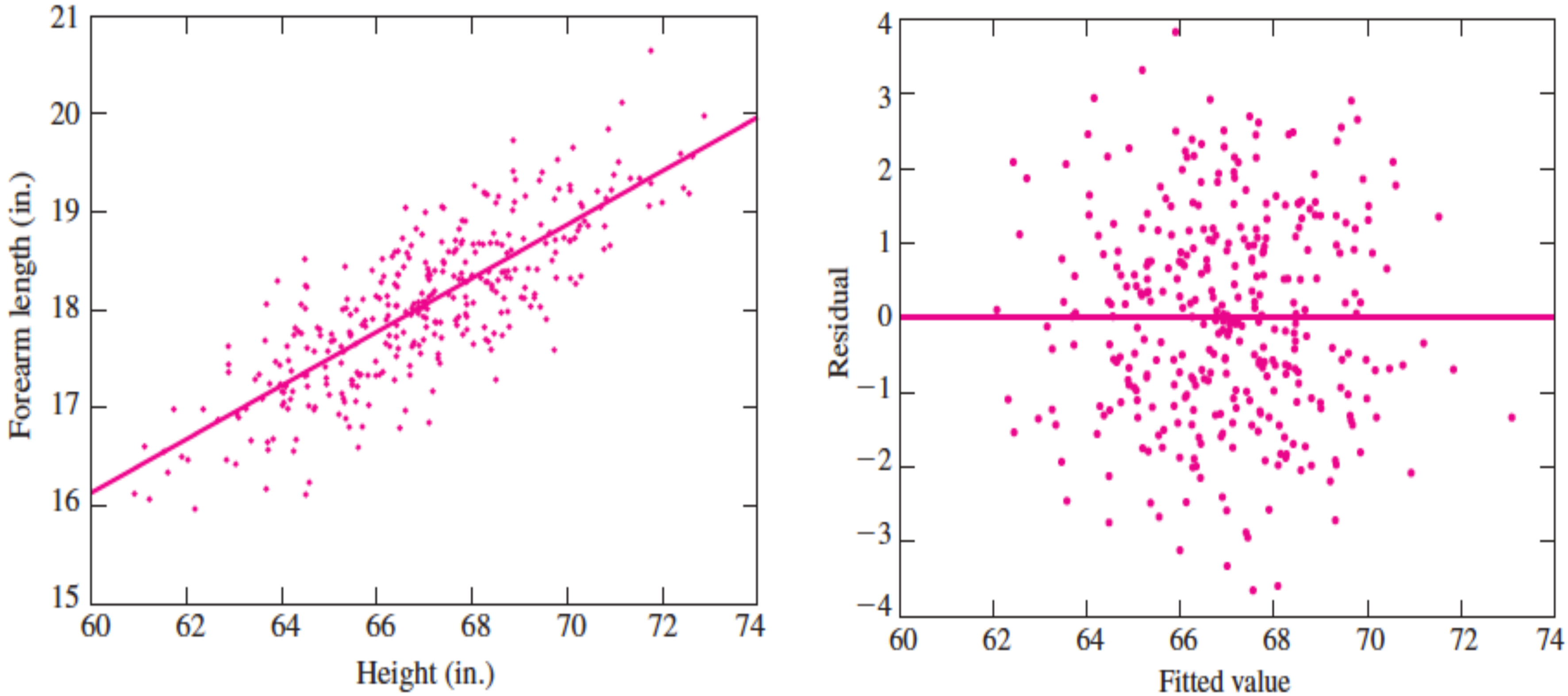


FIGURE 7.1 Heights and forearm lengths of 348 men.

Heteroscedasticity – “Megaphone” Shape

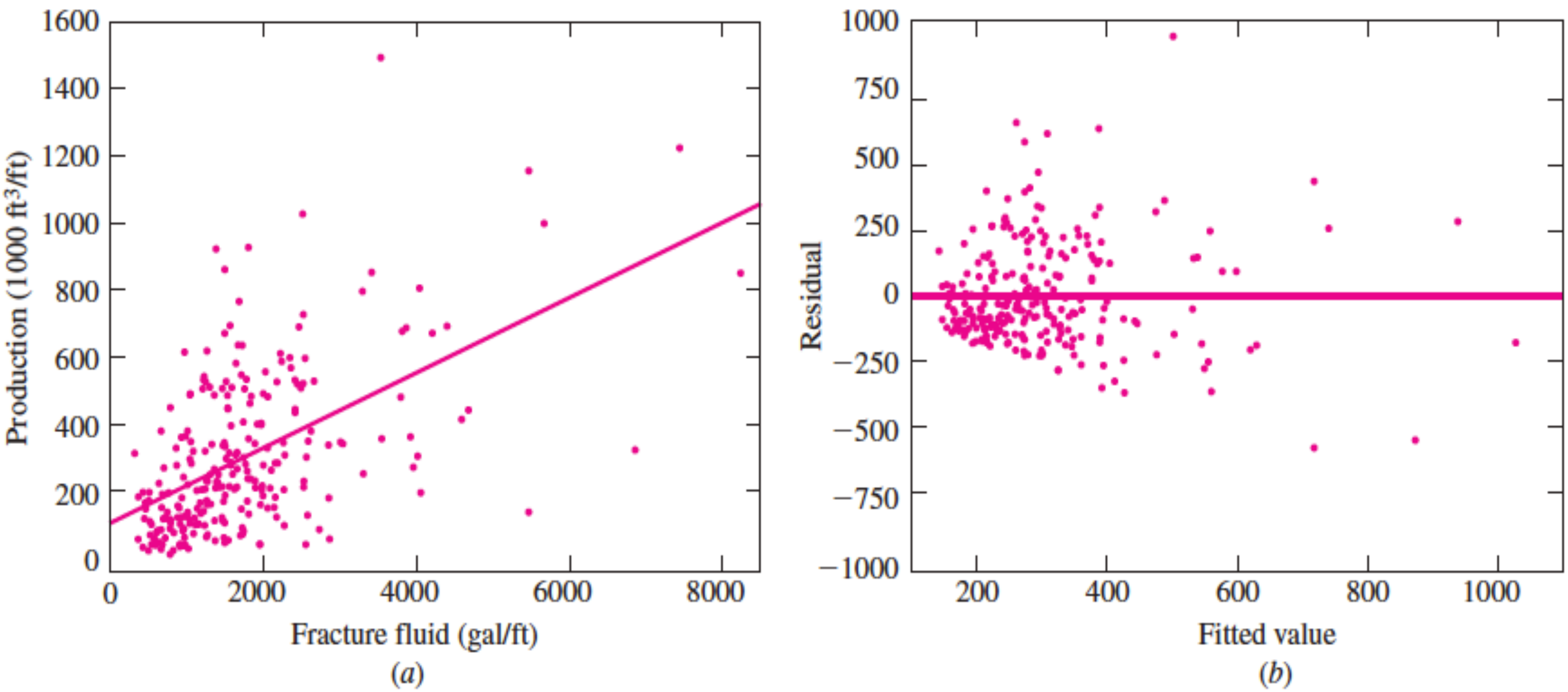


FIGURE 7.17 ***(a)*** Plot of monthly production versus volume of fracture fluid for 255 gas wells. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for the gas well data. The vertical spread clearly increases with the fitted value. This indicates a violation of the assumption of constant error variance.

Curvilinear Trend

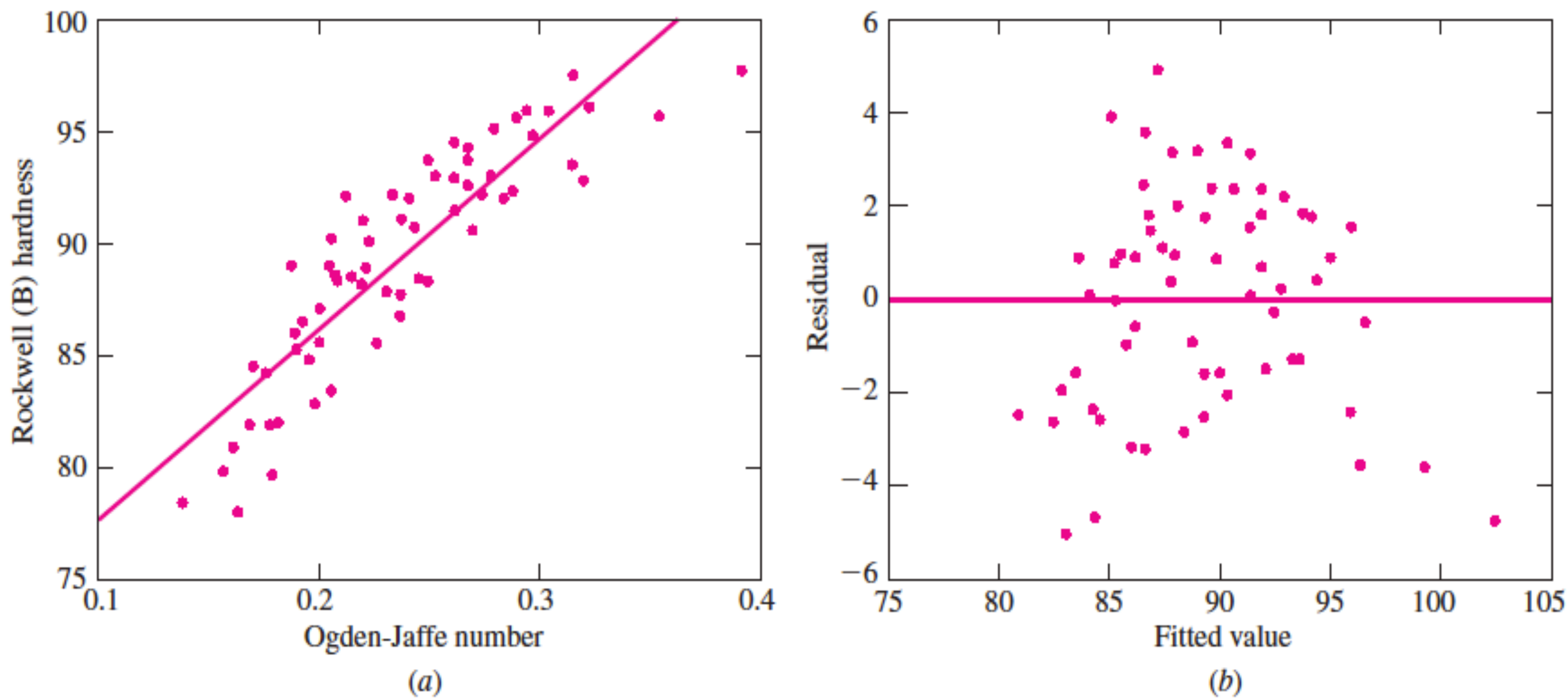
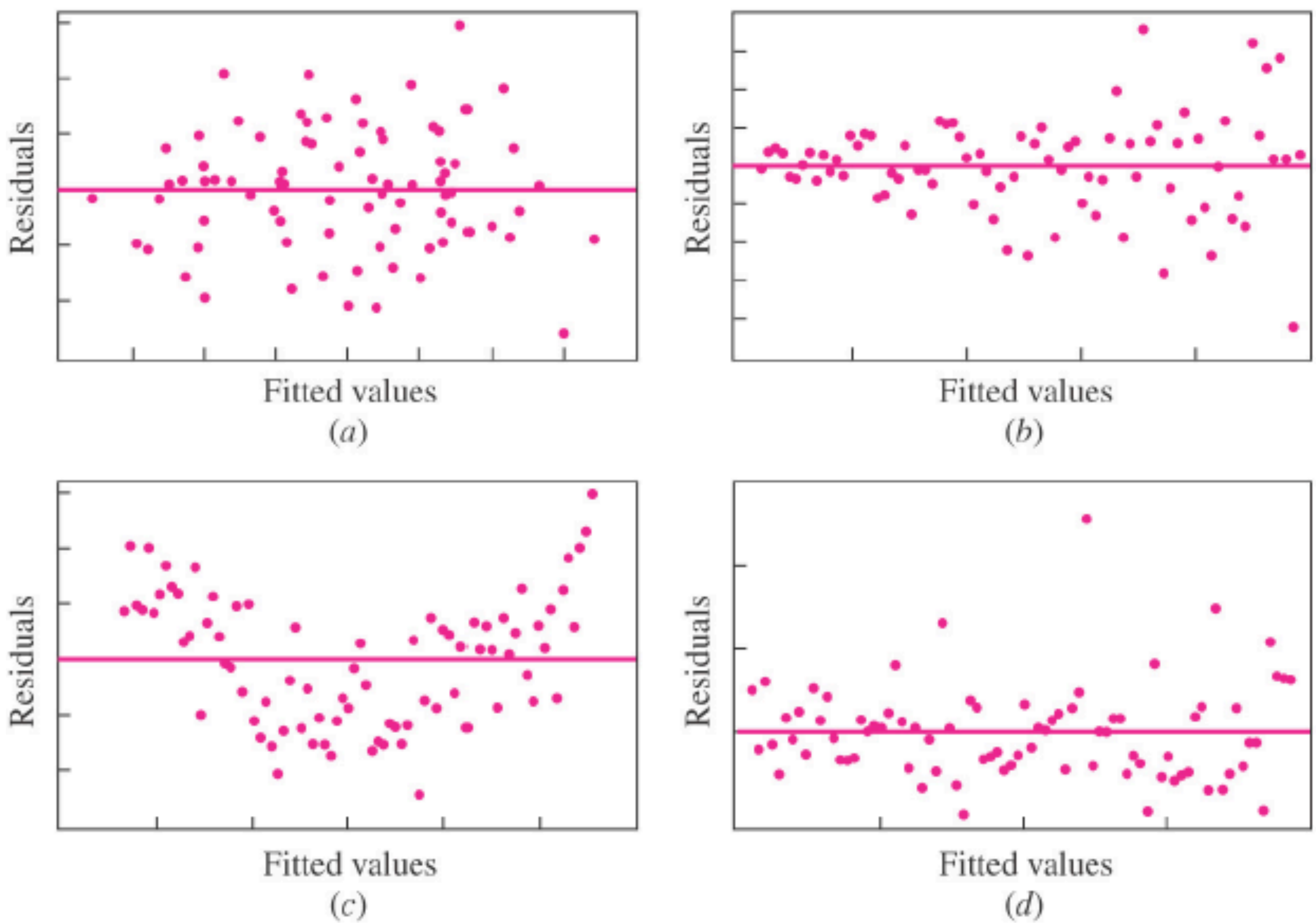


FIGURE 7.16 ***(a)*** Plot of Rockwell (B) hardness versus Ogden–Jaffe number. The least-squares line is superimposed. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for these data. The residuals plot shows a trend, with positive residuals in the middle and negative residuals at either end.


Residuals
Fitted values
(a)
Residuals
Fitted values
(b)
Residuals
Fitted values
(c)
Residuals
Fitted values
(d)

Interpreting Residual Plots

- If no clear trend or pattern in variance, we have no evidence that the assumptions are violated

- With a **small sample size**, residual plots can be difficult to interpret
  - Just like in the case of probability plots
  - If it looks OK except for a couple of suspect points, can proceed with linear model with caution – best practice is to collect more data points

Residual Plots in R

```
# read in data
housing <- read.table("housing.txt", header=TRUE)
attach(housing)

# fit the linear model with the lm(y~x) function
fit1 <- lm(Price~Sqft)

# plot residuals
plot(fit1$fitted, fit1$residuals, xlab="Fitted Values",
     ylab="Residual", main="Residual Plot for the
     Housing Data")

# add the y=0 line
abline(h=0, col="red")
```

Recall the Housing Data Example

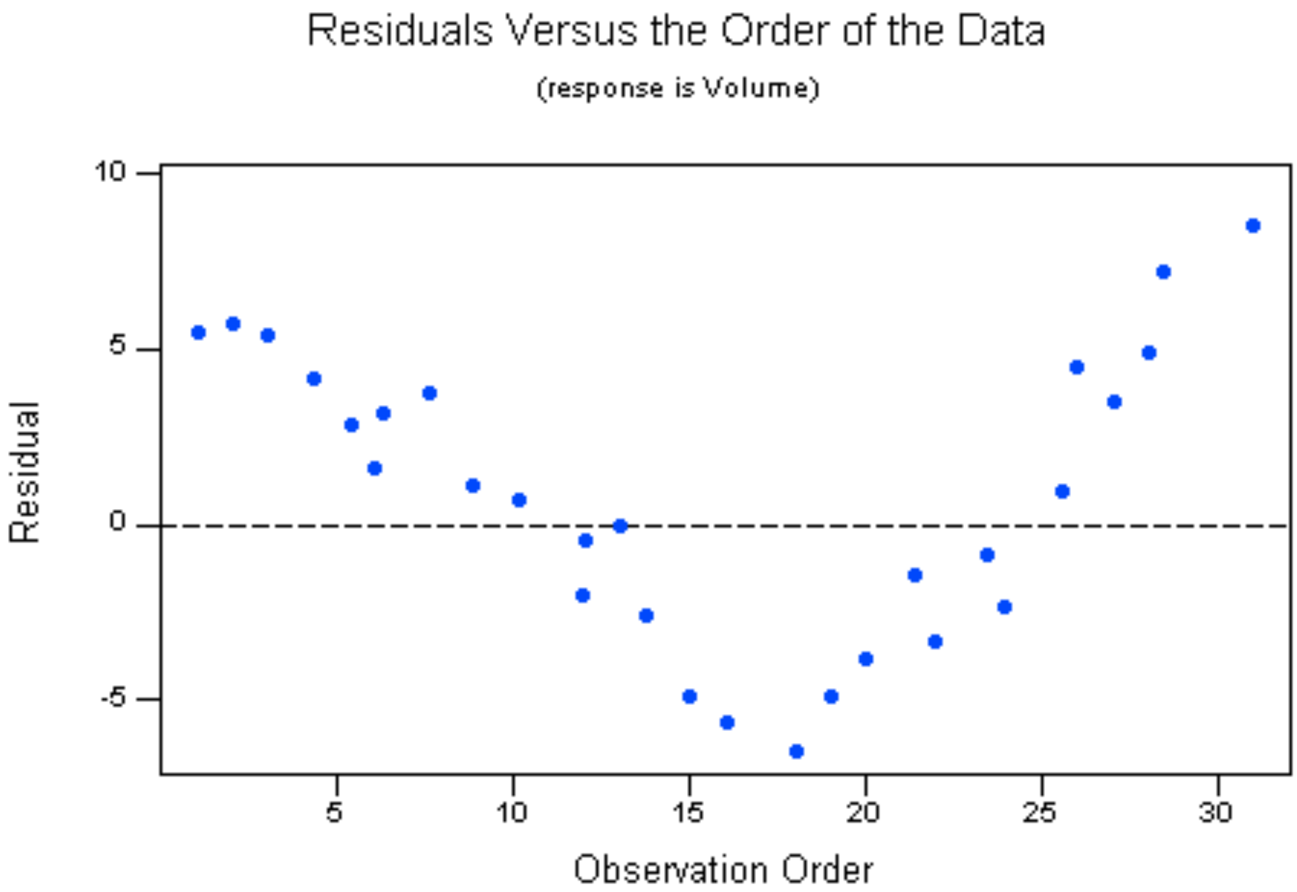
Checking Independence (Assumption 1)

- If the residual plot looks good, move on to check other assumptions

- To check for violations in the assumption of independence, plot the **residuals against the time order** of the observations (if applicable)

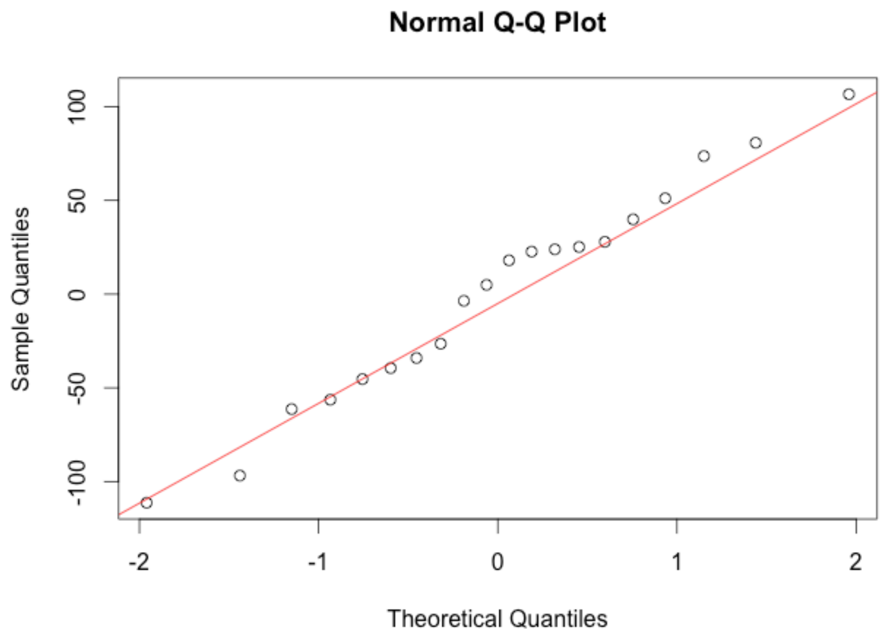
- A pattern/trend suggests that the errors are **not independent**

Example of Time Trend



Checking Normality (Assumption 4)

- A normal probability plot (also called a QQ plot) of the residuals can be used to check assumption 4
- Revisit Section 4.10 for a refresher on probability plots



- Obtained with in R with:

```
qqnorm(fit1$residuals)
qqline(fit1$residuals, col="red")
```

TRANSFORMATIONS

Power

Log

Square root

Fixing the Violations in the Linear Model

- **Transformation** is a useful tool to correct for violations

- We can transform a variable by replacing it with a one-to-one function of itself

- Commonly used functions:
  - Power transformation: raising a variable to a power

$$y^a = \beta_0 + \beta_1 x^b + \varepsilon$$

  - Log transformation: taking the natural logarithm of a variable

$$\log(y) = \beta_0 + \beta_1 \log(x) + \varepsilon$$

  - Square root transformation: special case of power transformation

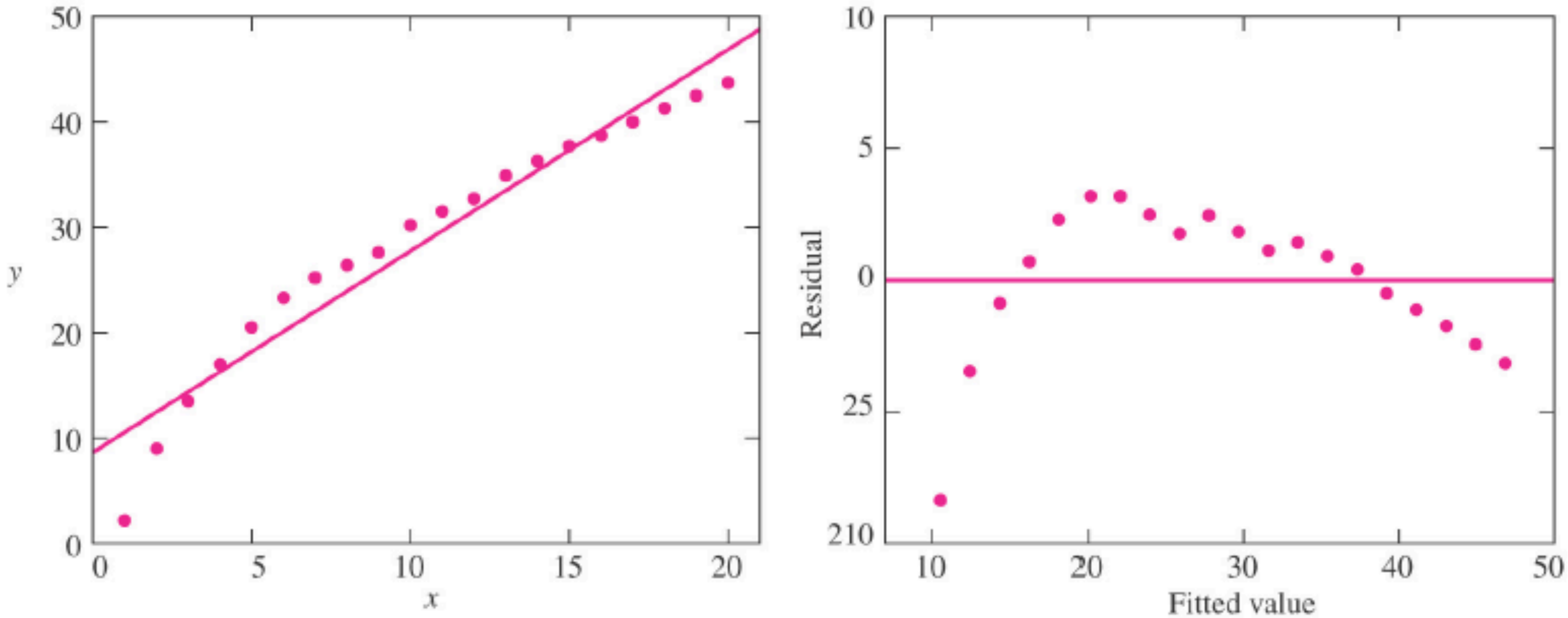
Applying Transformations

- Usually trial and error – apply a transformation and re-check the diagnostic plots (residual, QQ, time order)

- Can transform y, x, or both

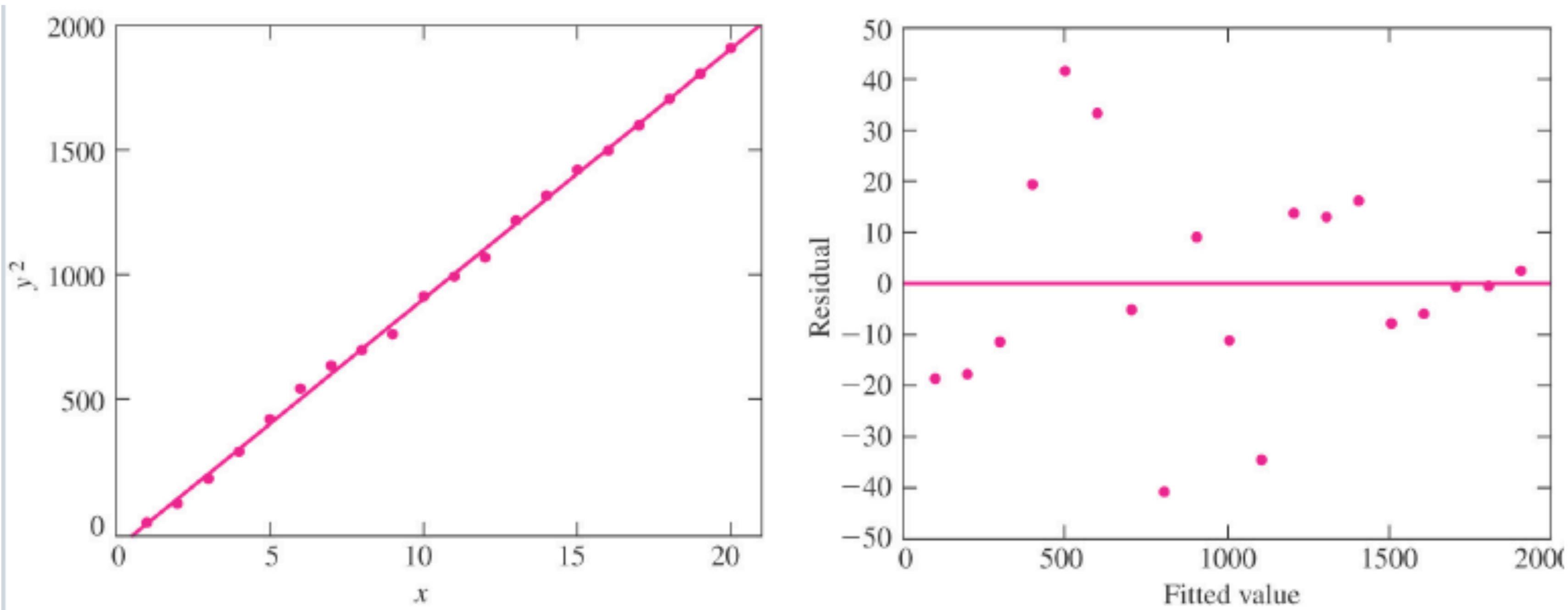
- You aren't guaranteed to find a remedy
  - Sometimes the violations are due to a confounding variable – in that case it is best to use **multiple regression** to include it in the model
  - Nonlinear regression might be more appropriate
  - Other 'flavors' of linear regression, such as weighted least-squares might be more appropriate

An Example of Transformation



Before Transformation, SLR model: $y = \beta_0 + \beta_1 x + \varepsilon$

Example Continued



After Transformation, SLR model: $y^2 = \beta_0 + \beta_1 x + \varepsilon$

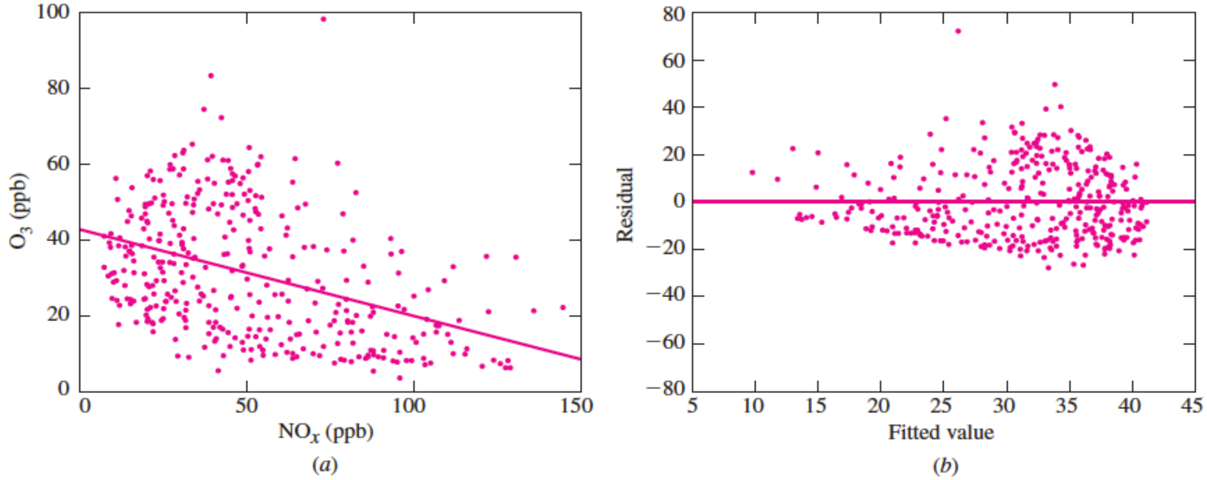


FIGURE 7.15 ***(a)*** Plot of ozone concentration versus NO_x concentration. The least-squares line is superimposed. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for these data. The vertical spread clearly increases with the fitted value. This indicates a violation of the assumption of constant error variance.

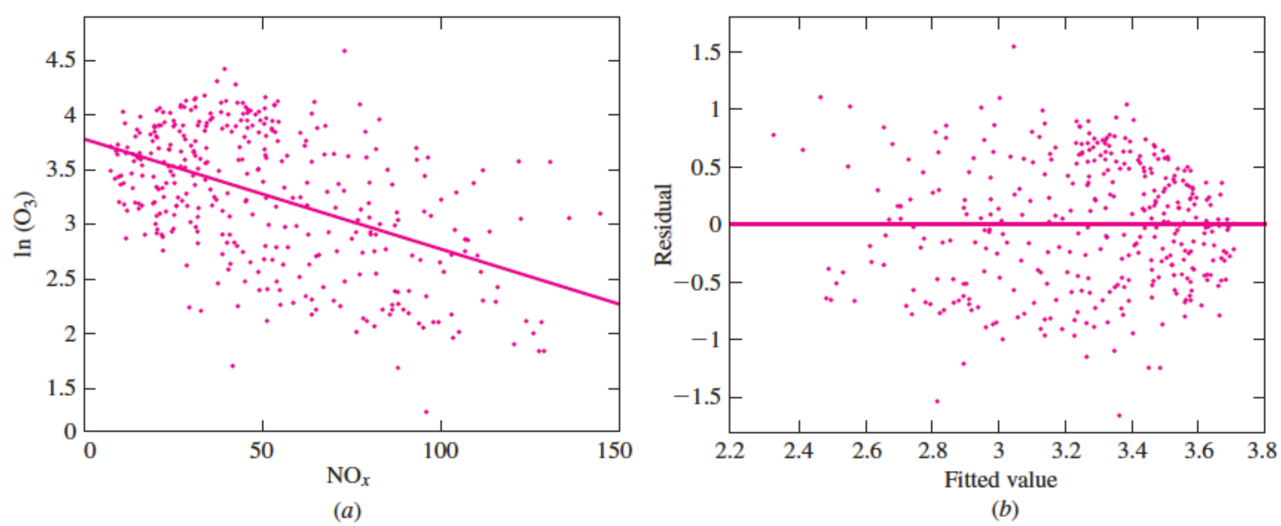


FIGURE 7.20 ***(a)*** Plot of the natural logarithm of ozone concentration versus NO_x concentration. The least-squares line is superimposed. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for these data. The linear model looks good.

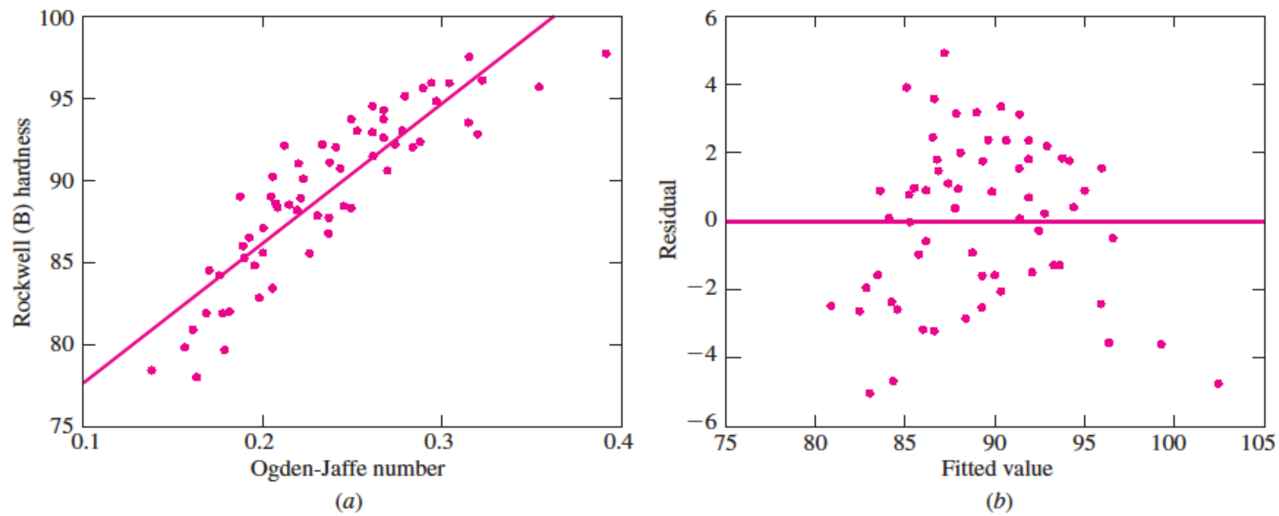


FIGURE 7.16 ***(a)*** Plot of Rockwell (B) hardness versus Ogden–Jaffe number. The least-squares line is superimposed. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for these data. The residuals plot shows a trend, with positive residuals in the middle and negative residuals at either end.

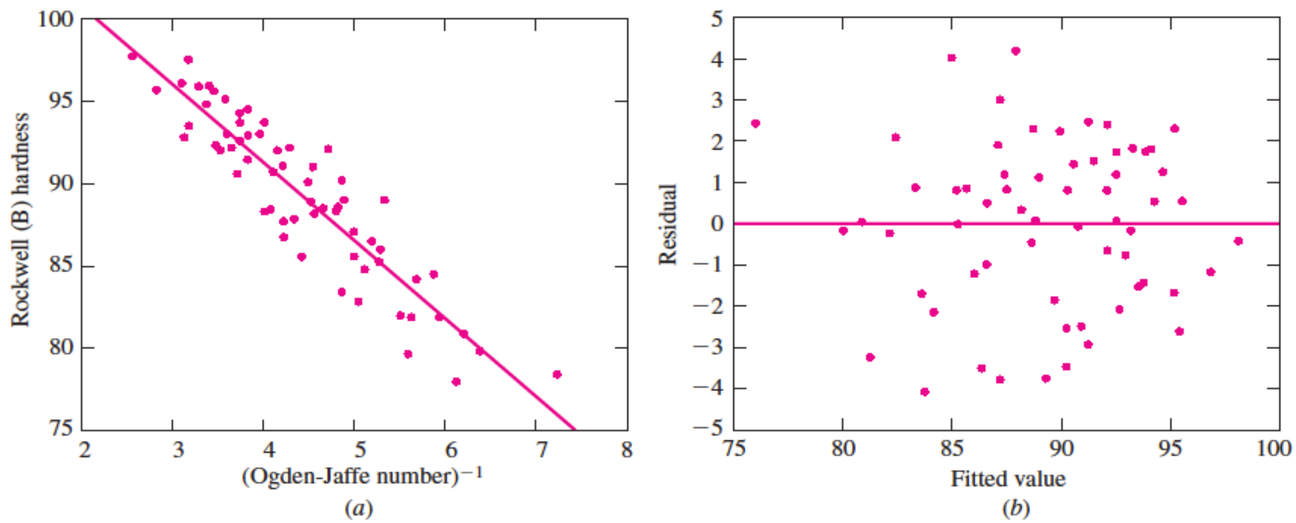
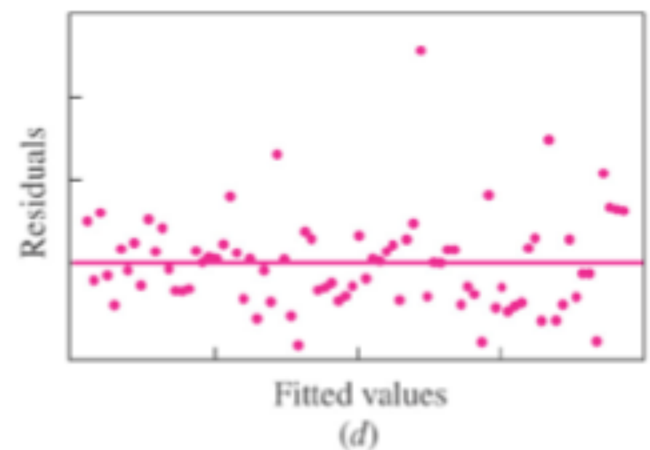


FIGURE 7.21 ***(a)*** Plot of hardness versus (Ogden–Jaffe number)$^{-1}$. The least-squares line is superimposed. ***(b)*** Plot of residuals (e_i) versus fitted values ($\widehat{y}_i$) for these data. The linear model looks good.

OUTLIERS AND INFLUENTIAL POINTS

Outliers

- Outliers are points that are detached from the bulk of the data – in SLR, we can find these visually
- First thing to do: try to find a cause for the extreme value to support its removal from the dataset
  - Data entry error?
  - Different machine operator?
  - Especially warm day?
  - etc…



- If you can't explain it, don't delete it
  - Fit the model with and without the outlier
  - If results change substantially, report both

Influential Points

- Outliers that cause a substantial change in the least-squares line when they are included are called **influential points**

- When influential points are present and you do not have justification to remove them, **avoid computing CIs or PIs or HTs** since the true nature of the linear relationship between x and y is unknown

Examples of Outliers and Influential Points

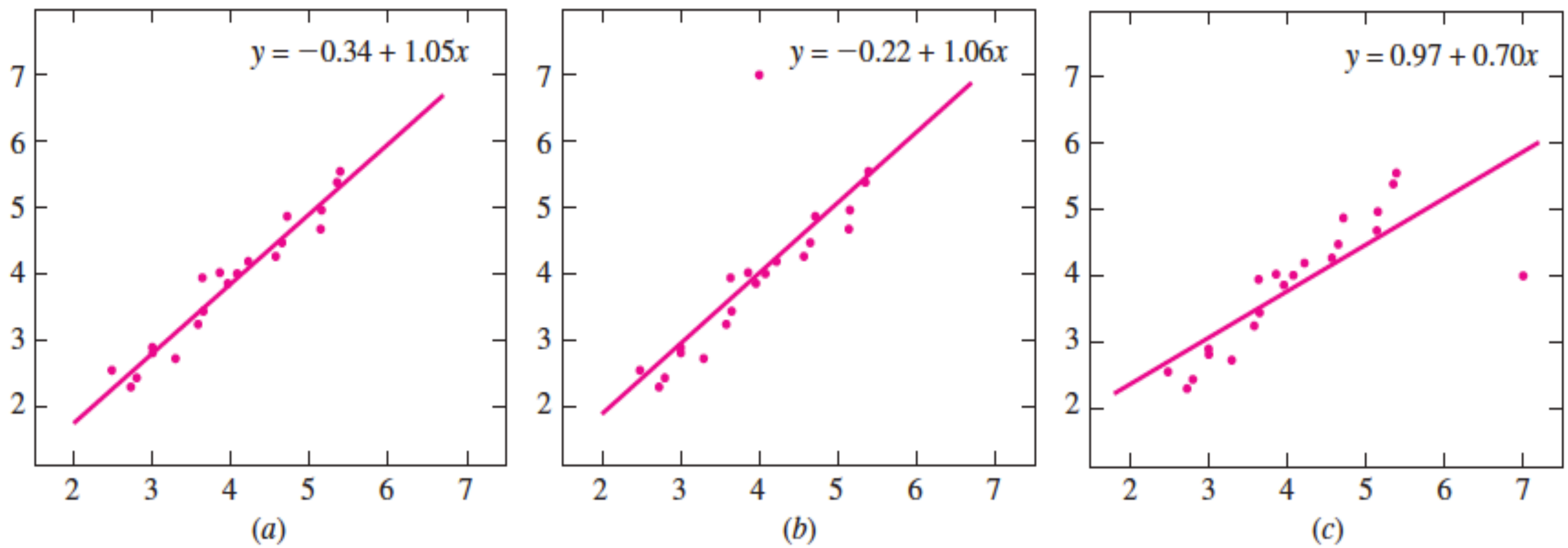


FIGURE 7.23 ***(a)*** Scatterplot with no outliers. ***(b)*** An outlier is added to the plot. There is little change in the least-squares line, so this point is not influential. ***(c)*** An outlier is added to the plot. There is a considerable change in the least-squares line, so this point is influential.

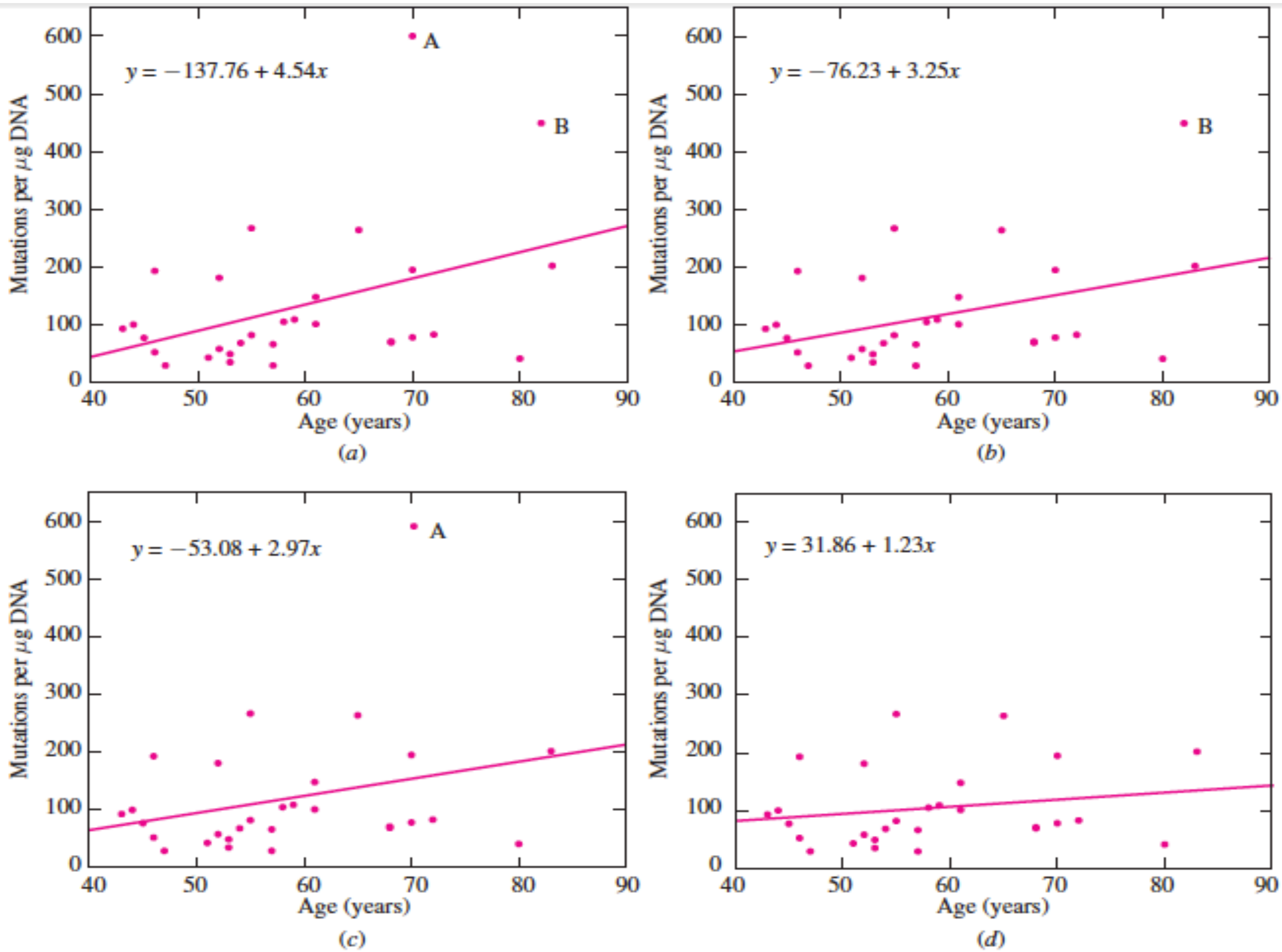
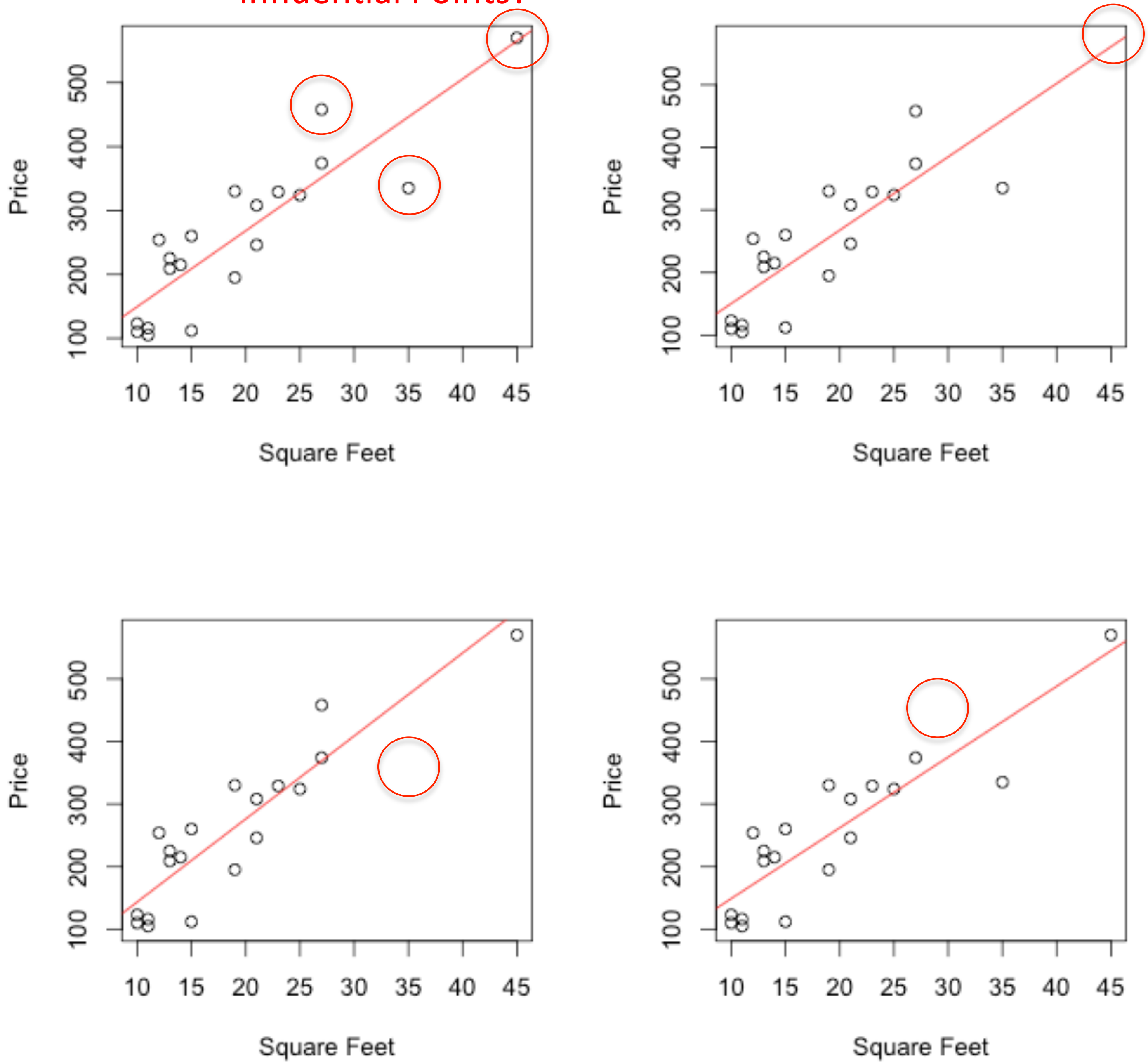
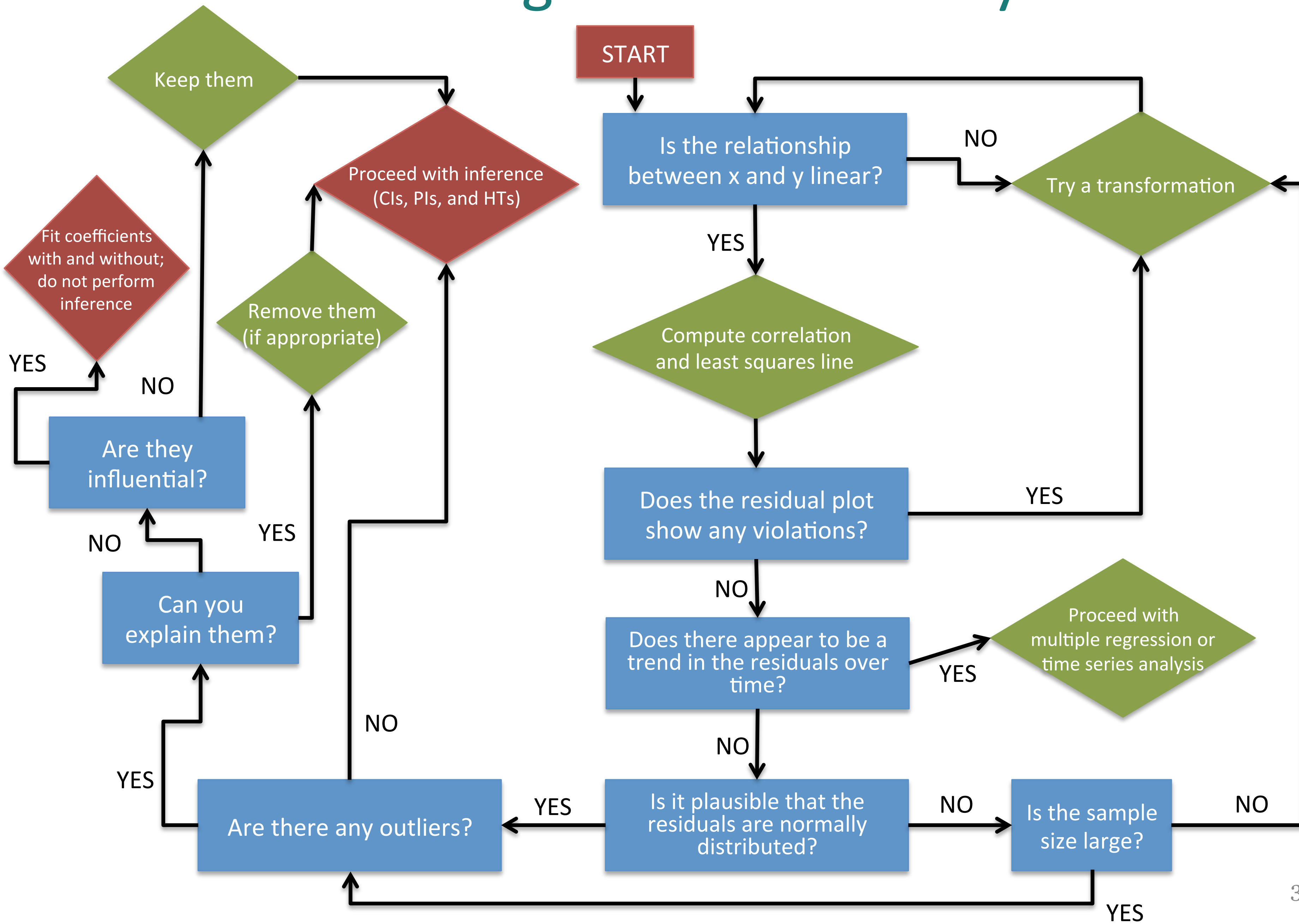


FIGURE 7.24 Mutation frequency versus age. *(a)* The plot contains two outliers, A and B. *(b)* Outlier A is deleted. The change in the least-squares line is noticeable although not extreme; this point is somewhat influential. *(c)* Outlier B is deleted. The change in the least-squares line is again noticeable but not extreme; this point is somewhat influential as well. *(d)* Both outliers are deleted. The combined effect on the least-squares line is substantial.

Influential Points?



SLR Diagnostics Summary



"ALL MODELS ARE WRONG… BUT SOME ARE USEFUL"

-George E. P. Box, 1919-2013

Empirical Models vs. Physical Laws

- **Empirical:**
  - based on observation or experience
  - valid only for data to which it is fit
  - may or may not be useful to predict future outcomes

- **Physical Law:**
  - accepted universal truth
  - applies to all future observations

Example: Triangle Areas vs Perimeters

- We want to predict the area of a triangle from its perimeter
- Sample 20 triangles, measure, and find least-squares line

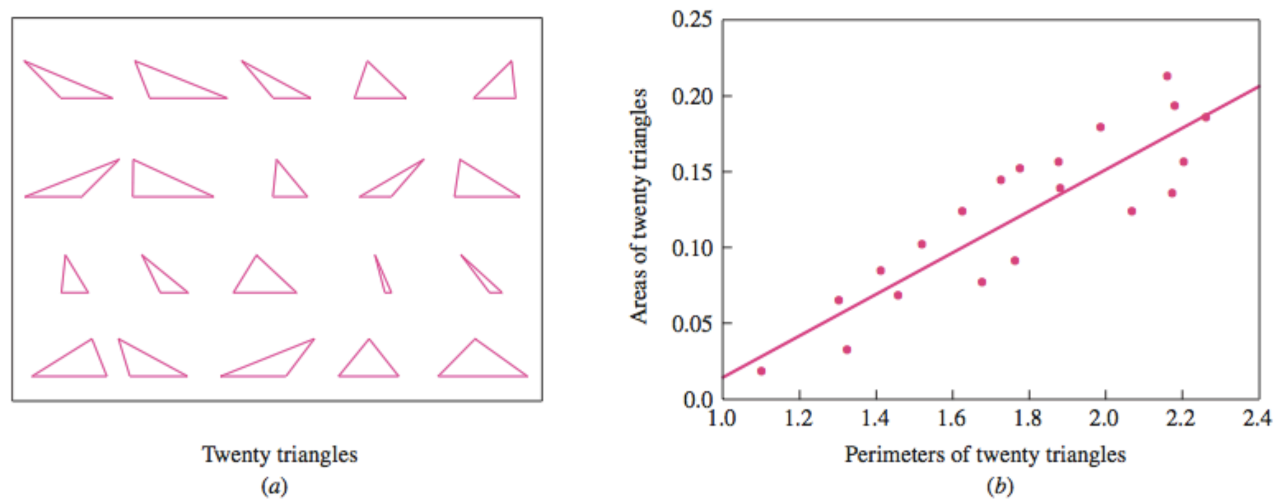


FIGURE 7.27 *(a)* Twenty triangles. *(b)* Area versus perimeter for 20 triangles. The correlation between perimeter and area is 0.88.

- **Empirical model:** Area = -1.232 + 1.373*Perimeter ← WRONG USEFUL?
- **Physical law:** ?????

Next

- Multiple regression – explaining the variation in a dependent variable with more than one independent variable

- HW 10 Due on Friday