Chapter 12. Simple Linear Regression and Correlation

12.1 The Simple Linear Regression Model
12.2 Fitting the Regression Line
12.3 Inferences on the Slope Parameter $\beta_1$
12.4 Inferences on the Regression Line
12.5 Prediction Intervals for Future Response Values
12.6 The Analysis of Variance Table
12.7 Residual Analysis
12.8 Variable Transformations
12.9 Correlation Analysis
12.10 Supplementary Problems
12.1 The Simple Linear Regression Model
12.1.1 Model Definition and Assumptions (1/5)

- With the *simple linear regression* model
  \[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]
  the observed value of the dependent variable \( y_i \) is composed of a linear function \( \beta_0 + \beta_1 x_i \) of the explanatory variable \( x_i \), together with an error term \( \varepsilon_i \). The error terms \( \varepsilon_1, \ldots, \varepsilon_n \) are generally taken to be independent observations from a \( N(0, \sigma^2) \) distribution, for some error variance \( \sigma^2 \). This implies that the values \( y_1, \ldots, y_n \) are observations from the independent random variables
  \[ Y_i \sim N (\beta_0 + \beta_1 x_i, \sigma^2) \]
as illustrated in Figure 12.1
12.1.1 Model Definition and Assumptions (2/5)

Figure 12.1 - Simple linear regression model

\[ y = \beta_0 + \beta_1 x \]

- Regression line
- Distribution of \( y_n \)
- Distribution of \( y_1 \)
- Distribution of \( y_2 \)

Variables:
- \( x \): Independent variable
- \( y \): Dependent variable
- \( \beta_0 \): Intercept
- \( \beta_1 \): Slope
12.1.1 Model Definition and Assumptions (3/5)

- The parameter $\beta_0$ is known as the intercept parameter, and the parameter $\beta_0$ is known as the intercept parameter, and the parameter $\beta_1$ is known as the slope parameter. A third unknown parameter, the error variance $\sigma^2$, can also be estimated from the data set. As illustrated in Figure 12.2, the data values $(x_i, y_i)$ lie closer to the line $y = \beta_0 + \beta_1 x$ as the error variance $\sigma^2$ decreases.
12.1.1 Model Definition and Assumptions (4/5)

- The **slope parameter** $\beta_1$ is of particular interest since it indicates how the expected value of the dependent variable depends upon the explanatory variable $x$, as shown in Figure 12.3.
- The data set shown in Figure 12.4 exhibits a quadratic (or at least nonlinear) relationship between the two variables, and it would make no sense to fit a straight line to the data set.

![Figure 12.3: Interpretation of slope parameter $\beta_1$](image1.png)

![Figure 12.4: For this nonlinear relationship a simple linear regression model is not appropriate](image2.png)
• **Simple Linear Regression Model**

The *simple linear regression* model

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

fit a straight line through a set of paired data observations \((x_1, y_1), \ldots, (x_n, y_n)\). The error terms \(\epsilon_1, \ldots, \epsilon_n\) are taken to be independent observations from a \(N(0, \sigma^2)\) distribution. The three unknown parameters, the **intercept parameter** \(\beta_0\), the **slope parameter** \(\beta_1\), and the **error variance** \(\sigma^2\), are estimated from the data set.
12.1.2 Examples (1/2)

- Example 3: **Car Plant Electricity Usage**

  The manager of a car plant wishes to investigate how the plant’s electricity usage depends upon the plant’s production.

  The linear model
  \[ y = \beta_0 + \beta_1 x \]
  will allow a month’s electrical usage to be estimated as a function of the month’s production.

  ![Car plant electricity usage data set](image) 

  **Figure 12.5**
  Car plant electricity usage data set
12.1.2 Examples (2/2)

**FIGURE 12.6**
Graph of car plant electricity usage

<table>
<thead>
<tr>
<th>Electricity usage</th>
<th>Production</th>
<th>Million kWh</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.75</td>
<td>3.5</td>
<td>1.25</td>
</tr>
<tr>
<td>2.50</td>
<td>4.0</td>
<td>3.75</td>
</tr>
<tr>
<td>2.00</td>
<td>4.5</td>
<td>3.50</td>
</tr>
<tr>
<td>1.75</td>
<td>5.0</td>
<td>3.25</td>
</tr>
<tr>
<td>1.50</td>
<td>5.5</td>
<td>3.00</td>
</tr>
<tr>
<td>1.25</td>
<td>6.0</td>
<td>2.75</td>
</tr>
</tbody>
</table>
12.2 Fitting the Regression Line
12.2.1 Parameter Estimation (1/4)

The regression line \( y = \beta_0 + \beta_1 x \) is fitted to the data points \((x_1, y_1), \ldots, (x_n, y_n)\) by finding the line that is "closest" to the data points in some sense.

As Figure 12.14 illustrates, the fitted line is chosen to be the line that \textit{minimizes} the sum of the squares of these vertical deviations

\[
Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2
\]

and this is referred to as the \textit{least squares} fit.
12.2.1 Parameter Estimation (2/4)

With normally distributed error terms, \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are maximum likelihood estimates.

(\( \therefore \)) The joint density of the error terms \( \epsilon_1, \ldots, \epsilon_n \) is

\[
\left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n e^{-\frac{\sum_{i=1}^{n} \epsilon_i^2}{2\sigma^2}}.
\]

This likelihood is maximized by minimizing

\[
\sum \epsilon_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2 = Q
\]

(\( \therefore \)) \( \frac{\partial Q}{\partial \beta_0} = -\sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)) \) and

\( \frac{\partial Q}{\partial \beta_1} = -\sum_{i=1}^{n} 2x_i (y_i - (\beta_0 + \beta_1 x_i)) \)

\( \Rightarrow \) the normal equations

\[
\sum y_i = n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad \text{and}
\]

\[
\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2
\]
12.2.1 Parameter Estimation (3/4)

\[ \beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{S_{XY}}{S_{XX}} \]

and then

\[ \beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \beta_1 \bar{x} \]

where

\[ S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \]

\[ = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \]

and

\[ S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} \]

For a specific value of the explanatory variable \( x^* \), this equation provides a fitted value \( \hat{y}_{x^*} = \beta_0 + \beta_1 x^* \) for the dependent variable \( y \), as illustrated in Figure 12.15.
The error variance $\sigma^2$ can be estimated by considering the deviations between the observed data values $y_i$ and their fitted values $\hat{y}_i$. Specifically, the sum of squares for error SSE is defined to be the sum of the squares of these deviations

$$
SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2
$$

$$
= \sum_{i=1}^{n} y_i^2 - \beta_0 \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i y_i
$$

and the estimate of the error variance is

$$
\hat{\sigma}^2 = \frac{SSE}{n-2}
$$
12.2.2 Examples (1/5)

- **Example 3: Car Plant Electricity Usage**

For this example $n = 12$ and

$$
\sum_{i=1}^{12} x_i = 4.51 + \cdots + 4.20 = 58.62
$$

$$
\sum_{i=1}^{12} y_i = 2.48 + \cdots + 2.53 = 34.15
$$

$$
\sum_{i=1}^{12} x_i^2 = 4.51^2 + \cdots + 4.20^2 = 291.2310
$$

$$
\sum_{i=1}^{12} y_i^2 = 2.48^2 + \cdots + 2.53^2 = 98.6967
$$

$$
\sum_{i=1}^{12} x_i y_i = (4.51 \times 2.48) + \cdots + (4.20 \times 2.53) = 169.2532
$$
12.2.2 Examples (2/5)
12.2.2 Examples(3/5)

The estimates of the slope parameter and the intercept parameter:

\[ \beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \]

\[ = \frac{(12 \times 169.2532) - (58.62 \times 34.15)}{(12 \times 291.2310) - 58.62^2} = 0.49883 \]

\[ \beta_0 = \bar{y} - \beta_1 \bar{x} = \frac{34.15}{12} - (0.49883 \times \frac{58.62}{12}) = 0.4090 \]

The fitted regression line:

\[ y = \beta_0 + \beta_1 x = 0.409 + 0.499x \]

\[ \hat{y}_{5.5} = 0.409 + (0.499 \times 5.5) = 3.1535 \]
12.2.2 Examples (4/5)

Using the model for production values $x$ outside this range is known as extrapolation and may give inaccurate results.
\[ \sigma^2 = \frac{\sum_{i=1}^{n} y_i^2 - \bar{\beta}_0 \sum_{i=1}^{n} y_i - \bar{\beta}_1 \sum_{i=1}^{n} x_i y_i}{n-2} \]

\[
= \frac{98.6967 - (0.4090 \times 34.15) - (0.49883 \times 169.2532)}{10} = 0.0299
\]

\[ \Rightarrow \sigma = \sqrt{0.0299} = 0.1729 \]
12.3 Inferences on the Slope Parameter $\beta_1$

12.3.1 Inference Procedures (1/4)

Inferences on the Slope Parameter $\beta_1$

- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{XX}})$.

- A two-sided confidence interval with a confidence level $1 - \alpha$ for the slope parameter in a simple linear regression model is

$$\beta_1 \in (\hat{\beta}_1 - t_{\alpha/2, n-2} \times s.e.(\hat{\beta}_1), \hat{\beta}_1 + t_{\alpha/2, n-2} \times s.e.(\hat{\beta}_1))$$

which is

$$\beta_1 \in (\hat{\beta}_1 - \frac{\sigma t_{\alpha/2, n-2}}{\sqrt{S_{XX}}}, \hat{\beta}_1 + \frac{\sigma t_{\alpha/2, n-2}}{\sqrt{S_{XX}}})$$

- One-sided $1 - \alpha$ confidence level confidence intervals are

$$\beta_1 \in (-\infty, \hat{\beta}_1 + \frac{\sigma t_{\alpha, n-2}}{\sqrt{S_{XX}}} ) \quad \text{and} \quad \beta_1 \in (\hat{\beta}_1 - \frac{\sigma t_{\alpha, n-2}}{\sqrt{S_{XX}}}, \infty)$$
The two-sided hypotheses

\[ H_0 : \beta_1 = b_1 \quad \text{versus} \quad H_A : \beta_1 \neq b_1 \]

for a fixed value \( b_1 \) of interest are tested with the \( t \)-statistic

\[
t = \frac{\bar{\beta}_1 - b_1}{\hat{\sigma}/\sqrt{S_{XX}}}
\]

The \( p \)-value is

\[ p-value = 2 \times P(X > |t|) \]

where the random variable \( X \) has a \( t \)-distribution with \( n - 2 \) degrees of freedom. A size \( \alpha \) test rejects the null hypothesis if \(|t| > t_{\alpha/2, n-2}\).
12.3.1 Inference Procedures (3/4)

- The one-sided hypotheses
  \[ H_0 : \beta_1 \geq b_1 \text{ versus } H_A : \beta_1 < b_1 \]
  have a \( p \)-value
  \[
  p\text{-value} = P(X < t)
  \]
  and a size \( \alpha \) test rejects the null hypothesis if \( t < -t_{\alpha,n-2} \).

- The one-sided hypotheses
  \[ H_0 : \beta_1 \leq b_1 \text{ versus } H_A : \beta_1 > b_1 \]
  have a \( p \)-value
  \[
  p\text{-value} = P(X > t)
  \]
  and a size \( \alpha \) test rejects the null hypothesis if \( t > t_{\alpha,n-2} \).
12.3.1 Inference Procedures

- An interesting point to notice is that for a fixed value of the error variance $\sigma^2$, the variance of the slope parameter estimate decreases as the value of $S_{XX}$ increases. This happens as the values of the explanatory variable $x_i$ become more spread out, as illustrated in Figure 12.30. This result is intuitively reasonable since a greater spread in the values $x_i$ provides a greater "leverage" for fitting the regression line, and therefore the slope parameter estimate $\beta_1$ should be more accurate.

FIGURE 12.30

The slope parameter $\beta_1$ is estimated more accurately in Scenario II than in Scenario I since the data points are more spread out and $S_{xx}$ is larger.

- $d_1 < d_2$
12.3.2 Examples (1/2)

- Example 3: **Car Plant Electricity Usage**

\[
S_{xx} = \sum_{i=1}^{12} x_i^2 - \frac{(\sum_{i=1}^{12} x_i)^2}{12} = 291.2310 - \frac{58.62^2}{12} = 4.8723
\]

\[
\Rightarrow s.e.(\beta_1) = \frac{\sigma}{\sqrt{S_{xx}}} = \frac{0.1729}{\sqrt{4.8723}} = 0.0783
\]

The \( t \)-statistic for testing \( H_0 : \beta_1 = 0 \)

\[
t \frac{\beta_1}{s.e.(\beta_1)} = \frac{0.49883}{0.0783} = 6.37
\]

The two-sided \( p \)-value

\[
p - \text{value} = 2 \times P(X > 6.37) \approx 0
\]
12.3.2 Examples(2/2)

With $t_{0.005,10} = 3.169$, a 99% two-sided confidence interval for the slope parameter

$$
\beta_1 \in \left( \beta_1 - \text{critical point} \times s.e.(\beta_1), \quad \beta_1 + \text{critical point} \times s.e.(\beta_1) \right)
= \left( 0.49883 - 3.169 \times 0.0783, \quad 0.49883 + 3.169 \times 0.0783 \right)
= \left( 0.251, \quad 0.747 \right)
$$
12.4 Inferences on the Regression Line

12.4.1 Inference Procedures (1/2)

Inferences on the Expected Value of the Dependent Variable

A $1 - \alpha$ confidence level two-sided confidence interval for $\beta_0 + \beta_1 x^*$, the expected value of the dependent variable for a particular value $x^*$ of the explanatory variable, is

$$\beta_0 + \beta_1 x^* \in (\overline{\beta}_0 + \overline{\beta}_1 x^* - t_{\alpha/2, n-1} \times s.e.(\overline{\beta}_0 + \overline{\beta}_1 x^*),$$

$$\beta_0 + \beta_1 x^* + t_{\alpha/2, n-2} \times s.e.(\overline{\beta}_0 + \overline{\beta}_1 x^*))$$

where

$$s.e.(\overline{\beta}_0 + \overline{\beta}_1 x^*) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}}$$
One-sided confidence intervals are

\[ \beta_0 + \beta_1 x^* \in (-\infty, \beta_0 + \beta_1 x^* + t_{\alpha,n-2} \times s.e.(\beta_0 + \beta_1 x^*)) \]

and

\[ \beta_0 + \beta_1 x^* \in (\beta_0 + \beta_1 x^* - t_{\alpha,n-1} \times s.e.(\beta_0 + \beta_1 x^*), \infty) \]

Hypothesis tests on \( \beta_0 + \beta_1 x^* \) can be performed by comparing the \( t \)-statistic

\[ t = \frac{(\beta_0 + \beta_1 x^*) - (\beta_0 + \beta_1 x^*)}{s.e.(\beta_0 + \beta_1 x^*)} \]

with a \( t \)-distribution with \( n-2 \) degrees of freedom.
12.4.2 Examples (1/2)

- Example 3: Car Plant Electricity Usage

\[
s.e.(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}} = 0.1729 \times \sqrt{\frac{1}{12} + \frac{(x^* - 4.885)^2}{4.8723}}
\]

With \( t_{0.025,10} = 2.228 \), a 95% confidence interval for \( \beta_0 + \beta_1 x^* \)

\[
\beta_0 + \beta_1 x^* \in (0.409 + 0.499x^* - 2.228 \times 0.1729 \times \sqrt{\frac{1}{12} + \frac{(x^* - 4.885)^2}{4.8723}}, \]
\[
0.409 + 0.499x^* + 2.228 \times 0.179 \times \sqrt{\frac{1}{12} + \frac{(x^* - 4.885)^2}{4.8723}}
\]

At \( x^* = 5 \)

\[
\beta_0 + 5 \beta_1 \in (0.409 + (0.499 \times 5) - 0.113, 0.409 + (0.499 \times 5) + 0.113)
\]
\[
= (2.79, 3.02)
\]
12.4.2 Examples (2/2)

Figure 12.33
Car plant electricity usage

Confidence Bands for the Fitted Regression Line

Electricity usage

Production

\( \bar{x} = 4.885 \)
12.5 Prediction Intervals for Future Response Values
12.5.1 Inference Procedures (1/2)

- **Prediction Intervals for Future Response Values**

A $1 - \alpha$ confidence level two-sided prediction interval for $y |_{x^*}$, a future value of the dependent variable for a particular value $x^*$ of the explanatory variable, is

$$y |_{x^*} \in (\hat{\beta}_0 + \hat{\beta}_1 x^* - t_{\alpha/2, n-1} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}},$$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* + t_{\alpha/2, n-2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}})$$
One-sided confidence intervals are

\[ y \mid x^* \in (-\infty, \beta_0 + \beta_1 x^* + t_{\alpha, n-2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}) \]

and

\[ y \mid x^* \in (\beta_0 + \beta_1 x^* - t_{\alpha, n-1} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}, \infty}) \]
12.5.2 Examples (1/2)

- Example 3: **Car Plant Electricity Usage**

With \( t_{0.025,10} = 2.228 \), a 95% confidence interval for \( y \mid x^* \)

\[
y \mid x^* \in (0.409 + 0.499x^* - 2.228 \times 0.1729 \times \sqrt{\frac{13}{12}} + \frac{(x^* - 4.885)^2}{4.8723}),
\]

\[
0.409 + 0.499x^* + 2.228 \times 0.179 \times \sqrt{\frac{13}{12}} + \frac{(x^* - 4.885)^2}{4.8723}
\]

At \( x^* = 5 \)

\[
y \mid 5 \in (0.409 + (0.499 \times 5) - 0.401, 0.409 + (0.499 \times 5) + 0.401) = (2.50, 3.30)
\]
12.5.2 Examples (2/2)
12.6 The Analysis of Variance Table
12.6.1 Sum of Squares Decomposition (1/5)

**Figure 12.39**
Sum of squares decomposition for regression analysis

- Total sum of squares (SST)
- Sum of squares for regression (SSR)
- Sum of squares for error (SSE)
12.6.1 Sum of Squares Decomposition (2/5)

FIGURE 12.40
Sum of squares for simple linear regression

\[
\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

\[
\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]
12.6.1 Sum of Squares Decomposition (3/5)

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean squares</th>
<th>$F$-statistic</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>MSR = SSR</td>
<td>$F=\text{MSR}/\text{MSE}$</td>
<td>$P(F_{1,n-2} &gt; F)$</td>
</tr>
<tr>
<td>Error</td>
<td>$N-2$</td>
<td>SSE</td>
<td>$\sigma^2 = \text{MSE} = \text{SSE}/(n-2)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 12.41**
Analysis of variance table for simple linear regression analysis
12.6.1 Sum of Squares Decomposition (4/5)

**Figure 12.42**

The coefficient of determination \( R^2 \) is larger in Scenario II than in Scenario I.
The total variability in the dependent variable, the total sum of squares

\[ \text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

can be partitioned into the variability explained by the regression line, the regression sum of squares

\[ \text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

and the variability about the regression line, the error sum of squares

\[ \text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \]

The proportion of the total variability accounted for by the regression line is the coefficient of determination

\[ R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{1}{1 + \frac{\text{SSE}}{\text{SSR}}} \]

which takes a value between zero and one.
12.6.2 Examples(1/1)

- Example 3: Car Plant Electricity Usage

\[ F = \frac{\text{MSR}}{\text{MSE}} = \frac{1.2124}{0.0299} = 40.53 \]

\[ R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{1.2124}{1.5115} = 0.802 \]
12.7 Residual Analysis

12.7.1 Residual Analysis Methods (1/7)

- The **residuals** are defined to be
  \[ e_i = y_i - \hat{y}_i, \quad 1 \leq i \leq n \]
  so that they are the differences between the observed values of the dependent variable \( y_i \) and the corresponding fitted values \( \hat{y}_i \).

- A property of the residuals
  \[ \sum_{i=1}^{n} e_i = 0 \]

- Residual analysis can be used to
  - Identify data points that are **outliers**,
  - Check whether the fitted model is **appropriate**,
  - Check whether the error variance is **constant**, and
  - Check whether the error terms are **normally** distributed.
12.7.1 Residual Analysis Methods (2/7)

- A nice random scatter plot such as the one in Figure 12.45 ⇒ there are no indications of any problems with the regression analysis
- Any patterns in the residual plot or any residuals with a large absolute value alert the experimenter to possible problems with the fitted regression model.
12.7.1 Residual Analysis Methods (3/7)

- A data point \((x_i, y_i)\) can be considered to be an outlier if it does not appear to predict well by the fitted model.
- Residuals of outliers have a **large absolute value**, as indicated in Figure 12.46. Note in the figure that \(\frac{e_i}{s}\) is used instead of \(e_i\).
- [For your interest only] \(\text{Var}(e_i) = (1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{XX}})s^2\).
12.7.1 Residual Analysis Methods (4/7)

- If the residual plot shows positive and negative residuals grouped together as in Figure 12.47, then a linear model is not appropriate. As Figure 12.47 indicates, a nonlinear model is needed for such a data set.
12.7.1 Residual Analysis Methods (5/7)

- If the residual plot shows a “funnel shape” as in Figure 12.48, so that the size of the residuals depends upon the value of the explanatory variable $x$, then the assumption of a constant error variance $\sigma^2$ is not valid.
• A normal probability plot (a normal score plot) of the residuals
  – Check whether the error terms $\varepsilon_i$ appear to be normally distributed.
• The normal score of the $i$th smallest residual
  $$\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right)$$
• The main body of the points in a normal probability plot lie approximately on a straight line as in Figure 12.49 is reasonable
• The form such as in Figure 12.50 indicates that the distribution is not normal
12.7.1 Residual Analysis Methods

**Figure 12.49**
A normal scores plot of a simulated sample from a normal distribution, which shows the points lying approximately on a straight line.

**Figure 12.50**
Normal scores plots of simulated samples from nonnormal distributions, which show nonlinear patterns.
12.7.2 Examples (1/2)

- Example: Nile River Flowrate

![Residual Plot]

FiguRe 12.51
Plot of standardized residuals for the Nile River flowrate example
12.7.2 Examples(2/2)

\[ x = 3.88 \]
\[ \hat{y}_5 = -0.470 + (0.836 \times 3.88) = 2.77 \]
\[ \Rightarrow e_i = y_i - \hat{y}_i = 4.01 - 2.77 = 1.24 \]
\[ \frac{e_i}{\sigma} = \frac{1.24}{\sqrt{0.1092}} = 3.75 \]

\[ x = 6.13 \]
\[ e_i = y_i - \hat{y}_i = 5.67 - (-0.470 + (0.836 \times 6.13)) = 1.02 \]
\[ \frac{e_i}{\sigma} = \frac{1.02}{\sqrt{0.1092}} = 3.07 \]
12.8 Variable Transformations
12.8.1 Intrinsically Linear Models (1/4)

Model: \( y = \gamma_0 e^{\gamma_1 x} \)
Linear format: \( \ln y = \ln \gamma_0 + \gamma_1 x \)
12.8.1 Intrinsically Linear Models (2/4)

Model: \( y = \gamma_0 x^{\gamma_1} \)

Linear format: \( \ln y = \ln \gamma_0 + \gamma_1 \ln x \)
12.8.1 Intrinsically Linear Models (3/4)
12.8.1 Intrinsically Linear Models (4/4)

Model: \( y = \frac{x}{\gamma_0 + \gamma_1 x} \)

Linear format: \( \frac{1}{y} = \gamma_1 + \gamma_0 \frac{1}{x} \)
12.8.2 Examples (1/5)

- Example: *Roadway Base Aggregates*

![Figure 12.54: Nonlinear transformation for roadway base aggregates example](image)

- **Model:** 
  
  \[ M = y_0 \theta^{y_1} \]

- **Original variables**
12.8.2 Examples (2/5)
12.8.2 Examples (3/5)

FIGURE 12.54  
Nonlinear transformation for roadway base aggregates example (continued)

Resilient modulus $M$  

Fitted model $M = 5014 \theta^{0.304}$
### 12.8.2 Examples (4/5)

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.77516</td>
<td>0.77516</td>
<td>1144.235</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>0.00948</td>
<td>0.00068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>15</td>
<td>0.78464</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Root MSE: 0.02603
- R-square: 0.9879
- Dep Mean: 9.58188
- Adj R-sq: 0.9870
- C.V.: 0.27164

#### Parameter Estimates

| Variable | DF | Estimate | Standard Error | T for H0: | Prob > |T| |
|----------|----|----------|----------------|-----------|---------|---|
| INTERCEP | 1  | 8.516334 | 0.03216518     | 264.769   | 0.0001  |   |
| LNBS     | 1  | 0.304277 | 0.00899522     | 33.827    | 0.0001  |   |
12.8.2 Examples (5/5)
The sample correlation coefficient $r$ for a set of paired data observations $(x_i, y_i)$ is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}}$$

It measures the strength of linear association between two variables and can be thought of as an estimate of the correlation $\rho$ between the two associated random variable $X$ and $Y$. 
Under the assumption that the $X$ and $Y$ random variables have a bivariate normal distribution, a test of the null hypothesis

$$H_0 : \rho = 0$$

can be performed by comparing the $t$-statistic

$$t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

with a $t$-distribution with $n - 2$ degrees of freedom. In a regression framework, this test is equivalent to testing $H_0 : \beta_1 = 0$. 
Figure 12.62. Sample correlation coefficient $r$. The correlation coefficient indicates the strength and direction of the linear relationship between two variables. High values of $r$ (close to 1 or -1) indicate a strong linear relationship, while values closer to 0 indicate a weaker relationship. The signs of $r$ indicate the direction of the relationship: positive for an upward trend and negative for a downward trend.
Figure 12.63
Misleading sample correlation coefficient for a nonlinear relationship

Sample correlation coefficient $r = 0$

$y = \hat{\beta}_0 + \hat{\beta}_1 x$
12.9.2 Examples(1/1)

- Example: *Nile River Flowrate*

\[ r = \sqrt{R^2} = \sqrt{0.871} = 0.933 \]