STAT 2300 NOTES
CHAPTER 12: SIMPLE LINEAR REGRESSION ANALYSIS
Outline

1. 12.1: The Simple Linear Regression Model
   - Introduction
   - The First Step: Determine a Model Form
   - Form and Assumption of a Linear Relationship

2. 12.2: The Least Squares Estimates, and Point Estimation and Prediction
All of the statistical inference we have conducted, has entailed a single variable.

- Within Chapters 7 and 8, we conducted inference for parameter(s) of a single population. The inference encompassed the population parameters, $\mu$, $\rho$, and $\sigma^2$.
- Within Chapter 9, we conducted inference surrounding two populations. The inference encompassed the population parameters, $\mu_i$, $\rho_i$, and $\sigma^2_i$, where $i = 1, 2$.
- Within Chapter 12, we will learn how to examine the linear relationship between two variables, where at least one of the variables is continuous.
- It turns out that the two sample t-test of Chapter 9 (Section 9.2), in the case where $\sigma^2_1 = \sigma^2_2$, is a special case of the simple linear regression model.
Introduction

Overview

- All of the statistical inference we have conducted, has entailed a single variable.
  - Within Chapters 7 and 8, we conducted inference for parameter(s) of a single population. The inference encompassed the population parameters, $\mu$, $\rho$, and $\sigma^2$.
  - Within Chapter 9, we conducted inference surrounding two populations. The inference encompassed the population parameters, $\mu_i$, $\rho_i$, and $\sigma_{i}^2$, where $i = 1, 2$.
  - Within Chapter 12, we will learn how to examine the linear relationship between two variables, where at least one of the variables is continuous.
  - It turns out that the two sample t-test of Chapter 9 (Section 9.2), in the case where $\sigma_{1}^2 = \sigma_{2}^2$, is a special case of the simple linear regression model.
All of the statistical inference we have conducted, has entailed a single variable.

- Within Chapters 7 and 8, we conducted inference for parameter(s) of a single population. The inference encompassed the population parameters, $\mu$, $\rho$, and $\sigma^2$.
- Within Chapter 9, we conducted inference surrounding two populations. The inference encompassed the population parameters, $\mu_i$, $\rho_i$, and $\sigma^2_i$, where $i = 1, 2$.

- Within Chapter 12, we will learn how to examine the linear relationship between two variables, where at least one of the variables is continuous.
- It turns out that the two sample t-test of Chapter 9 (Section 9.2), in the case where $\sigma^2_1 = \sigma^2_2$, is a special case of the simple linear regression model.
All of the statistical inference we have conducted, has entailed a single variable.

- Within Chapters 7 and 8, we conducted inference for parameter(s) of a single population. The inference encompassed the population parameters, $\mu$, $\rho$, and $\sigma^2$.
- Within Chapter 9, we conducted inference surrounding two populations. The inference encompassed the population parameters, $\mu_i$, $\rho_i$, and $\sigma^2_i$, where $i = 1, 2$.

Within Chapter 12, we will learn how to examine the linear relationship between two variables, where at least one of the variables is continuous.

- It turns out that the two sample t-test of Chapter 9 (Section 9.2), in the case where $\sigma^2_1 = \sigma^2_2$, is a special case of the simple linear regression model.
All of the statistical inference we have conducted, has entailed a single variable.

Within Chapters 7 and 8, we conducted inference for parameter(s) of a single population. The inference encompassed the population parameters, $\mu$, $\rho$, and $\sigma^2$.

Within Chapter 9, we conducted inference surrounding two populations. The inference encompassed the population parameters, $\mu_i$, $\rho_i$, and $\sigma_i^2$, where $i = 1, 2$.

Within Chapter 12, we will learn how to examine the linear relationship between two variables, where at least one of the variables is continuous.

It turns out that the two sample t-test of Chapter 9 (Section 9.2), in the case where $\sigma_1^2 = \sigma_2^2$, is a special case of the simple linear regression model.
Introduction

The Setup

- Assume we have a continuous variable, which we denote by \( y \).
- Suppose we want to investigate the (linear) relationship between \( y \) and another variable. Denote this variable by \( x \).
- For each (of the \( n \)) participant of our study, we observe \( x \) and \( y \), and denote the observations by \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).
- Goal: Build a model for predicting \( y \) as a function of \( x \).
  - Example: Is there a linear relationship between MPH (\( x \)) and stopping distance (\( y \))?
Introduction

The Setup

- Assume we have a continuous variable, which we denote by $y$.
- Suppose we want to investigate the (linear) relationship between $y$ and another variable. Denote this variable by $x$.
- For each (of the $n$) participant of our study, we observe $x$ and $y$, and denote the observations by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Goal: Build a model for predicting $y$ as a function of $x$.
  - Example: Is there a linear relationship between MPH ($x$) and stopping distance ($y$)?
Assume we have a continuous variable, which we denote by $y$.

Suppose we want to investigate the (linear) relationship between $y$ and another variable. Denote this variable by $x$.

For each (of the $n$) participant of our study, we observe $x$ and $y$, and denote the observations by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Goal: Build a model for predicting $y$ as a function of $x$.

Example: Is there a linear relationship between MPH ($x$) and stopping distance ($y$)?
Introduction

The Setup

- Assume we have a continuous variable, which we denote by $y$.
- Suppose we want to investigate the (linear) relationship between $y$ and another variable. Denote this variable by $x$.
- For each (of the $n$) participant of our study, we observe $x$ and $y$, and denote the observations by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Goal: Build a model for predicting $y$ as a function of $x$.
  - Example: Is there a linear relationship between MPH ($x$) and stopping distance ($y$)?
Assume we have a continuous variable, which we denote by $y$.

Suppose we want to investigate the (linear) relationship between $y$ and another variable. Denote this variable by $x$.

For each (of the $n$) participant of our study, we observe $x$ and $y$, and denote the observations by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Goal: Build a model for predicting $y$ as a function of $x$.

Example: Is there a linear relationship between MPH ($x$) and stopping distance ($y$)?
The dependent (or response) variable is the variable we wish to understand or predict.

The independent (or predictor) variable is the variable we will use to understand or predict the dependent variable.

Regression analysis is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.

A regression analysis which entails a single independent variable is called simple regression.

When the mean of the response variable is modeled (directly) as a linear combination of independent variables, the regression is called linear regression.
The **dependent** (or response) variable is the variable we wish to understand or predict.

The **independent** (or predictor) variable is the variable we will use to understand or predict the dependent variable.

Regression analysis is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.

A regression analysis which entails a single independent variable is called simple regression.

When the mean of the response variable is modeled (directly) as a linear combination of independent variables, the regression is called linear regression.
The dependent (or response) variable is the variable we wish to understand or predict.

The independent (or predictor) variable is the variable we will use to understand or predict the dependent variable.

Regression analysis is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.

A regression analysis which entails a single independent variable is called simple regression.

When the mean of the response variable is modeled (directly) as a linear combination of independent variables, the regression is called linear regression.
The dependent (or response) variable is the variable we wish to understand or predict.

The independent (or predictor) variable is the variable we will use to understand or predict the dependent variable.

Regression analysis is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.

A regression analysis which entails a single independent variable is called simple regression.

When the mean of the response variable is modeled (directly) as a linear combination of independent variables, the regression is called linear regression.
Introduction

Background

- The **dependent** (or response) variable is the variable we wish to understand or predict.
- The **independent** (or predictor) variable is the variable we will use to understand or predict the dependent variable.
- **Regression analysis** is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.
- A regression analysis which entails a single independent variable is called simple regression.
- When the mean of the response variable is modeled (directly) as a linear combination of independent variables, the regression is called linear regression.
Utility of the Linear Regression Model

Two Reasons to Construct a Simple Linear Regression Model

1. To investigate whether a linear relationship between $x$ and $y$ exists.
2. For any given value of $x$, we want to predict the value of $y$. 

Utility of the Linear Regression Model

Two Reasons to Construct a Simple Linear Regression Model

1. To investigate whether a linear relationship between $x$ and $y$ exists.
2. For any given value of $x$, we want to predict the value of $y$. 
Utility of the Linear Regression Model

Two Reasons to Construct a Simple Linear Regression Model

1. To investigate whether a linear relationship between $x$ and $y$ exists.
2. For any given value of $x$, we want to predict the value of $y$. 
The First Step: Determine a Model Form

Example 12.1: Fuel Consumption (1)

Require a Linear Relationship Between $x$ and $y$

- We want to use temperature ($x$) to predict fuel consumption ($y$).
  - More specifically, for any given value of the independent variable, we want to predict the population mean response value, denoted by $\mu_{y|x}$. The symbol, $\mu_{y|x}$, is read, “The population mean response (in $y$), given the value of $x$.”
  - In order to do this, we need to verify that the mean response value ($\mu_{y|x}$), changes linearly for increasing values of $x$.
- The relationship between $x$ and $y$ may be linear, quadratic, log-linear, etc.
- To investigate the relationship, create a scatterplot of $y$ versus $x$. 
The First Step: Determine a Model Form

Example 12.1: Fuel Consumption (1)

Require a Linear Relationship Between $x$ and $y$

- We want to use temperature ($x$) to predict fuel consumption ($y$).
- More specifically, for any given value of the independent variable, we want to predict the population mean response value, denoted by $\mu_{y|x}$. The symbol, $\mu_{y|x}$, is read, “The population mean response (in $y$), given the value of $x$.”
- In order to do this, we need to verify that the mean response value ($\mu_{y|x}$), changes linearly for increasing values of $x$.
- The relationship between $x$ and $y$ may be linear, quadratic, log-linear, etc.
- To investigate the relationship, create a scatterplot of $y$ versus $x$. 

Example 12.1: Fuel Consumption (1)

**Require a Linear Relationship Between x and y**

- We want to use temperature \( (x) \) to predict fuel consumption \( (y) \).
  - More specifically, for any given value of the independent variable, we want to predict the population mean response value, denoted by \( \mu_{y|x} \). The symbol, \( \mu_{y|x} \), is read, “The population mean response (in \( y \)), given the value of \( x \).”
  - In order to do this, we need to verify that the mean response value \( (\mu_{y|x}) \), changes linearly for increasing values of \( x \).

- The relationship between \( x \) and \( y \) may be linear, quadratic, log-linear, etc.

- To investigate the relationship, create a scatterplot of \( y \) versus \( x \).
Example 12.1: Fuel Consumption (1)

Require a Linear Relationship Between $x$ and $y$

- We want to use temperature ($x$) to predict fuel consumption ($y$).
- More specifically, for any given value of the independent variable, we want to predict the population mean response value, denoted by $\mu_{y|x}$. The symbol, $\mu_{y|x}$, is read, “The population mean response (in $y$), given the value of $x$.”
- In order to do this, we need to verify that the mean response value ($\mu_{y|x}$), changes linearly for increasing values of $x$.
- The relationship between $x$ and $y$ may be linear, quadratic, log-linear, etc.
- To investigate the relationship, create a scatterplot of $y$ versus $x$. 
Example 12.1: Fuel Consumption (1)

Require a Linear Relationship Between \( x \) and \( y \)

- We want to use temperature \((x)\) to predict fuel consumption \((y)\).
  - More specifically, for any given value of the independent variable, we want to predict the population mean response value, denoted by \( \mu_y|x \). The symbol, \( \mu_y|x \), is read, “The population mean response (in \( y \)), given the value of \( x \).”
  - In order to do this, we need to verify that the mean response value \( (\mu_y|x) \), changes linearly for increasing values of \( x \).
- The relationship between \( x \) and \( y \) may be linear, quadratic, log-linear, etc.
- To investigate the relationship, create a scatterplot of \( y \) versus \( x \).
The First Step: Determine a Model Form

**Example 12.1: Fuel Consumption (2)**

### The Data

<table>
<thead>
<tr>
<th>Week</th>
<th>Temperature °F (x)</th>
<th>Fuel Consumption (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.0</td>
<td>12.4</td>
</tr>
<tr>
<td>2</td>
<td>28.0</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>8</td>
<td>62.5</td>
<td>7.5</td>
</tr>
</tbody>
</table>
The First Step: Determine a Model Form

Example 12.1: Fuel Consumption (3)

Scatterplot of the Data (response is labeled along the vertical axis)
The First Step: Determine a Model Form

Example 12.1: Fuel Consumption (3)

Scatterplot of the Data with Least Squares Regression Line (red)
Based on the scatterplot from the previous slide, it appears there is a linear relationship between $x$ and $y$.

So, for any given value of $x$, we model the mean of the response, $\mu_{y|x}$, by the linear equation

$$\mu_{y|x} = \beta_0 + \beta_1 x.$$  \hfill (1)

$\beta_0$ and $\beta_1$ are called regression parameters.

$\beta_0 = \mu_{y|0}$ is called the intercept of the linear model. It carries the same units as $y$.

$\beta_1$ is the slope of the linear model. It describes how the mean response changes for a one-unit change in $x$. It carries the ratio of units $y$, divided by units $x$.

We use (to no surprise, right) statistical inference to estimate the regression parameters.
The Linear Model

- Based on the scatterplot from the previous slide, it appears there is a linear relationship between $x$ and $y$.
- So, for any given value of $x$, we model the mean of the response, $\mu_{y|x}$, by the linear equation

$$\mu_{y|x} = \beta_0 + \beta_1 x.$$  

- $\beta_0$ and $\beta_1$ are called regression parameters.
- $\beta_0 = \mu_{y|0}$ is called the intercept of the linear model. It carries the same units as $y$.
- $\beta_1$ is the slope of the linear model. It describes how the mean response changes for a one-unit change in $x$. It carries the ratio of units $y$, divided by units $x$.
- We use (to no surprise, right) statistical inference to estimate the regression parameters.
Based on the scatterplot from the previous slide, it appears there is a linear relationship between $x$ and $y$.

So, for any given value of $x$, we model the mean of the response, $\mu_{y|x}$, by the linear equation

$$\mu_{y|x} = \beta_0 + \beta_1 x.$$  \hspace{1cm} (1)

$\beta_0$ and $\beta_1$ are called **regression parameters**.

$\beta_0 = \mu_{y|0}$ is called the **intercept** of the linear model. It carries the same units as $y$.

$\beta_1$ is the **slope** of the linear model. It describes how the mean response changes for a one-unit change in $x$. It carries the ratio of units $y$, divided by units $x$.

We use (to no surprise, right) statistical inference to estimate the regression parameters.
Based on the scatterplot from the previous slide, it appears there is a linear relationship between $x$ and $y$.

So, for any given value of $x$, we model the mean of the response, $\mu_{y|x}$, by the linear equation

$$\mu_{y|x} = \beta_0 + \beta_1 x.$$  \hspace{1cm} (1)

$\beta_0$ and $\beta_1$ are called regression parameters.

$\beta_0 = \mu_{y|0}$ is called the intercept of the linear model. It carries the same units as $y$.

$\beta_1$ is the slope of the linear model. It describes how the mean response changes for a one-unit change in $x$. It carries the ratio of units $y$, divided by units $x$.

We use (to no surprise, right) statistical inference to estimate the regression parameters.
Based on the scatterplot from the previous slide, it appears there is a linear relationship between \( x \) and \( y \).

So, for any given value of \( x \), we model the mean of the response, \( \mu_{y|x} \), by the linear equation

\[
\mu_{y|x} = \beta_0 + \beta_1 x.
\]  

(1)

\( \beta_0 \) and \( \beta_1 \) are called regression parameters.

\( \beta_0 = \mu_{y|0} \) is called the intercept of the linear model. It carries the same units as \( y \).

\( \beta_1 \) is the slope of the linear model. It describes how the mean response changes for a one-unit change in \( x \). It carries the ratio of units \( y \), divided by units \( x \).

We use (to no surprise, right) statistical inference to estimate the regression parameters.
Based on the scatterplot from the previous slide, it appears there is a linear relationship between $x$ and $y$.

So, for any given value of $x$, we model the mean of the response, $\mu_{y|x}$, by the linear equation

$$
\mu_{y|x} = \beta_0 + \beta_1 x.
$$

$\beta_0$ and $\beta_1$ are called **regression parameters**.

$\beta_0 = \mu_{y|0}$ is called the **intercept** of the linear model. It carries the same units as $y$.

$\beta_1$ is the **slope** of the linear model. It describes how the mean response changes for a one-unit change in $x$. It carries the ratio of units $y$, divided by units $x$.

We use (to no surprise, right) statistical inference to estimate the regression parameters.
Graphical Representation of the Linear Model

- The y-intercept $\beta_0$
- The line of means: $\mu_{\hat{y}|x} = \beta_0 + \beta_1 x$
- $\mu_{\hat{y}|28} =$ Mean weekly fuel consumption when $x = 28$
- The error term for the first week (a positive error term)
- $\mu_{\hat{y}|45.9} =$ Mean weekly fuel consumption when $x = 45.9$
- The error term for the fifth week (a negative error term)
- $9.4 =$ The observed fuel consumption for the first week
- $12.4 =$ The observed fuel consumption for the first week
- $\beta_1 =$ The slope
- One-degree increase in $x$
Outline

1. 12.1: The Simple Linear Regression Model

2. 12.2: The Least Squares Estimates, and Point Estimation and Prediction
   - Introduction
   - Finding the Best Fitting Line to the Data
   - Example
   - Cautions About Simple Linear Regression
The Setup

- We do not know the values of the regression parameters, so we must estimate these from sample data.
- Now, equation (1), says the mean response is a linear function of $x$. Hence, at any fixed value of $x$, we assume random error in the response about the mean value, $\mu_{y|x}$.
- We introduce the notion of random error into the model, via the residual, denoted by $\epsilon$. Correspondingly, equation (1) is rewritten as

$$y = \mu_{y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon. \quad (2)$$

- Note, for a given value of $x$, $\epsilon$ measures the vertical distance that the value of $y$ is from $\mu_{y|x}$.
- A good fitting model should have small residuals. STAT 5100 will have at least a week of coursework devoted to analysis of residuals.
We do not know the values of the regression parameters, so we must estimate these from sample data.

Now, equation (1), says the mean response is a linear function of $x$. Hence, at any fixed value of $x$, we assume random error in the response about the mean value, $\mu_{y|x}$.

We introduce the notion of random error into the model, via the residual, denoted by $\epsilon$. Correspondingly, equation (1) is rewritten as

$$y = \mu_{y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon.$$  \hfill (2)

Note, for a given value of $x$, $\epsilon$ measures the vertical distance that the value of $y$ is from $\mu_{y|x}$.

A good fitting model should have small residuals. STAT 5100 will have at least a week of coursework devoted to analysis of residuals.
Introduction

The Setup

- We do not know the values of the regression parameters, so we must estimate these from sample data.
- Now, equation (1), says the mean response is a linear function of $x$. Hence, at any fixed value of $x$, we assume random error in the response about the mean value, $\mu_{\theta|x}$.
- We introduce the notion of random error into the model, via the residual, denoted by $\epsilon$. Correspondingly, equation (1) is rewritten as

$$y = \mu_{\theta|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon.$$  \hfill (2)

- Note, for a given value of $x$, $\epsilon$ measures the vertical distance that the value of $y$ is from $\mu_{\theta|x}$.
- A good fitting model should have small residuals. STAT 5100 will have at least a week of coursework devoted to analysis of residuals.
Introduction

The Setup

- We do not know the values of the regression parameters, so we must estimate these from sample data.
- Now, equation (1), says the mean response is a linear function of $x$. Hence, at any fixed value of $x$, we assume random error in the response about the mean value, $\mu_{y|x}$.
- We introduce the notion of random error into the model, via the residual, denoted by $\epsilon$. Correspondingly, equation (1) is rewritten as

$$y = \mu_{y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon. \quad (2)$$

- Note, for a given value of $x$, $\epsilon$ measures the vertical distance that the value of $y$ is from $\mu_{y|x}$.
- A good fitting model should have small residuals. STAT 5100 will have at least a week of coursework devoted to analysis of residuals.
The Setup

- We do not know the values of the regression parameters, so we must estimate these from sample data.
- Now, equation (1), says the mean response is a linear function of \( x \). Hence, at any fixed value of \( x \), we assume random error in the response about the mean value, \( \mu_{y|x} \).
- We introduce the notion of random error into the model, via the residual, denoted by \( \epsilon \). Correspondingly, equation (1) is rewritten as

\[
y = \mu_{y|x} + \epsilon = \beta_0 + \beta_1 x + \epsilon.
\]  

(2)

- Note, for a given value of \( x \), \( \epsilon \) measures the vertical distance that the value of \( y \) is from \( \mu_{y|x} \).
- A good fitting model should have small residuals. STAT 5100 will have at least a week of coursework devoted to analysis of residuals.
How do we find the best equation for \( y \) on \( x \)?

We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as \( b_0 \) and \( b_1 \), respectively.

Once determined, the values of \( b_0 \) and \( b_1 \) are substituted into equation (1), in lieu of \( \beta_0 \) and \( \beta_1 \), respectively, and we obtain the expression

\[
\hat{y} = \hat{\mu}_{y|x} = b_0 + b_1 x.
\]  

"\( \hat{y} \)" is the predicted mean response for a given value of \( x \).

Equation (3) is called the least squares regression line.

The least squares regression line can be used to predict the mean value of the response from given values of \( x \).
The Least Squares Regression Line

How do we find the best equation for $y$ on $x$?

We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as $b_0$ and $b_1$, respectively.

Once determined, the values of $b_0$ and $b_1$ are substituted into equation (1), in lieu of $\beta_0$ and $\beta_1$, respectively, and we obtain the expression

$$\hat{y} = \mu_{y|x} = b_0 + b_1 x. \quad (3)$$

“$\hat{y}$” is the predicted mean response for a given value of $x$.

Equation (3) is called the least squares regression line.

The least squares regression line can be used to predict the mean value of the response from given values of $x$. 

"Finding the Best Fitting Line to the Data"
The Least Squares Regression Line

How do we find the best equation for \( y \) on \( x \)?

We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as \( b_0 \) and \( b_1 \), respectively.

Once determined, the values of \( b_0 \) and \( b_1 \) are substituted into equation (1), in lieu of \( \beta_0 \) and \( \beta_1 \), respectively, and we obtain the expression

\[
\hat{y} = \hat{\mu}_{y|x} = b_0 + b_1 x.
\] (3)

“\( \hat{y} \)” is the predicted mean response for a given value of \( x \).

Equation (3) is called the least squares regression line.

The least squares regression line can be used to predict the mean value of the response from given values of \( x \).
The Least Squares Regression Line

- How do we find the best equation for \( y \) on \( x \)?
- We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as \( b_0 \) and \( b_1 \), respectively.
- Once determined, the values of \( b_0 \) and \( b_1 \) are substituted into equation (1), in lieu of \( \beta_0 \) and \( \beta_1 \), respectively, and we obtain the expression
  \[
  \hat{y} = \mu_{y|x} = b_0 + b_1 x.
  \] (3)
- “\( \hat{y} \)” is the predicted mean response for a given value of \( x \).
- Equation (3) is called the least squares regression line.
- The least squares regression line can be used to predict the mean value of the response from given values of \( x \).
Finding the Best Fitting Line to the Data

The Least Squares Regression Line

- How do we find the best equation for y on x?
- We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as $b_0$ and $b_1$, respectively.
- Once determined, the values of $b_0$ and $b_1$ are substituted into equation (1), in lieu of $\beta_0$ and $\beta_1$, respectively, and we obtain the expression

$$\hat{y} = \hat{\mu}_{y|x} = b_0 + b_1 x.$$  (3)

- “$\hat{y}$” is the predicted mean response for a given value of x.
- Equation (3) is called the least squares regression line.
- The least squares regression line can be used to predict the mean value of the response from given values of x.
How do we find the best equation for $y$ on $x$?

We want to determine the best values (estimates) of the slope and intercept parameters of equation (2). Denote these values as $b_0$ and $b_1$, respectively.

Once determined, the values of $b_0$ and $b_1$ are substituted into equation (1), in lieu of $\beta_0$ and $\beta_1$, respectively, and we obtain the expression

$$\hat{y} = \hat{\mu}_{y|x} = b_0 + b_1 x.$$  \(3\)

“$\hat{y}$” is the predicted mean response for a given value of $x$.

Equation (3) is called the least squares regression line.

The least squares regression line can be used to predict the mean value of the response from given values of $x$. 
The Least Squares Regression Line (cont.)

- There are three methods by which we can obtain the value of $b_0$ and $b_1$, one of which is the method of least squares (it turns out, all three methods are equivalent).
- The least squares regression line is the line fit to the data, which minimizes the sum of the squares of the vertical distances from the observed data values to the fitted line.
- Operationally, this involves minimizing the function

$$
\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 ,
$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, are observed data values for $x$ and $y$, and $\hat{y}_i$ is the predicted value of $y_i$, when $x_i$ is substituted within expression (3).
There are three methods by which we can obtain the value of $b_0$ and $b_1$, one of which is the method of least squares (it turns out, all three methods are equivalent).

The least squares regression line is the line fit to the data, which minimizes the sum of the squares of the vertical distances from the observed data values to the fitted line.

Operationally, this involves minimizing the function

$$
\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2,
$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, are observed data values for $x$ and $y$, and $\hat{y}_i$ is the predicted value of $y_i$, when $x_i$ is substituted within expression (3).
The Least Squares Regression Line (cont.)

- There are three methods by which we can obtain the value of $b_0$ and $b_1$, one of which is the method of least squares (it turns out, all three methods are equivalent).

- The least squares regression line is the line fit to the data, which minimizes the sum of the squares of the vertical distances from the observed data values to the fitted line.

- Operationally, this involves minimizing the function

$$
\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2,
$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, are observed data values for $x$ and $y$, and $\hat{y}_i$ is the predicted value of $y_i$, when $x_i$ is substituted within expression (3).
The Least Squares Regression Line

- Note that the least squares regression line does not pass through every point of our observed data.
- But, it is the unique line which minimizes the sum of the squared residual values.
- It is our best estimate of the line $\mu_{y|x} = \beta_0 + \beta_1 x$. 
The Least Squares Regression Line

- Note that the least squares regression line does not pass through every point of our observed data.
- But, it is the unique line which minimizes the sum of the squared residual values.
- It is our best estimate of the line $\mu_{y|x} = \beta_0 + \beta_1 x$. 
The Least Squares Regression Line

- Note that the least squares regression line does not pass through every point of our observed data.
- But, it is the *unique* line which minimizes the sum of the squared residual values.
- It is our best estimate of the line

\[ \mu_{y|x} = \beta_0 + \beta_1 x. \]
Finding the Best Fitting Line to the Data

**Point Estimate for \( \beta_1, b_1 \)**

- For the least squares regression line of \( y \) regressed on \( x \), \( \hat{y} = b_0 + b_1 x \), the point estimate for \( \beta_1, b_1 \), satisfies

\[
b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}},
\]

where

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.
\]

- Note that the covariance of \( x \) and \( y \) is equal to \( (n - 1)^{-1} SS_{xy} \). The covariance between two random variables measures the degree by which one variable changes with respect to the other variable.

- Also, the formula for \( SS_{xx} \) should look familiar. Namely, \( SS_{xx} \) is equal to \( (n - 1)s_x^2 \), where \( s_x^2 \) is the sample variance for the variable \( x \).
Finding the Best Fitting Line to the Data

**Point Estimate for $\beta_1, b_1$**

- For the least squares regression line of $y$ regressed on $x$, $\hat{y} = b_0 + b_1 x$, the point estimate for $\beta_1, b_1$, satisfies

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}},$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$  

- Note that the **covariance of $x$ and $y$** is equal to $(n - 1)^{-1} SS_{xy}$. The covariance between two random variables measures the degree by which one variable changes with respect to the other variable.

- Also, the formula for $SS_{xx}$ should look familiar. Namely, $SS_{xx}$ is equal to $(n - 1)s_x^2$, where $s_x^2$ is the sample variance for the variable $x$. 
Finding the Best Fitting Line to the Data

### Point Estimate for $\beta_1$, $b_1$

- For the least squares regression line of $y$ regressed on $x$, $\hat{y} = b_0 + b_1 x$, the point estimate for $\beta_1$, $b_1$, satisfies

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}},$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

- Note that the **covariance of $x$ and $y$** is equal to $(n - 1)^{-1} SS_{xy}$. The covariance between two random variables measures the degree by which one variable changes with respect to the other variable.

- Also, the formula for $SS_{xx}$ should look familiar. Namely, $SS_{xx}$ is equal to $(n - 1)s_x^2$, where $s_x^2$ is the sample variance for the variable $x$. 

---

**Bill Welbourn**  
Chapter 12: STAT 2300 – Summer 2009
Point Estimate for $\beta_0, b_0$

For the least squares regression line of $y$ regress on $x$, $\hat{y} = b_0 + b_1 x$, the point estimate for $\beta_0, b_0$, satisfies

$$b_0 = \bar{y} - b_1 \bar{x}.$$
Finding the Best Fitting Line to the Data

Interpreting the Parameter Estimates

- The slope, $b_1$, is the estimated change in mean $y$ per one unit increase in $x$.
- The intercept, $b_0$, is the estimated mean of $y$ when $x = 0$.
- Due to the structure of the assumed model, namely,

$$
\mu_{y|x} = b_0 + b_1 x,
$$

the predicted value of $\mu_{y|0}$, $b_0$, can assume any real number.
- This can lead to a misleading interpretation of $b_0$.
- To remedy this notion, whenever $x$ is a continuous variable, we “center” the values of $x$ about its sample mean, $\bar{x}$. In this case, $b_0$ is the mean value of $y$ given $x = \bar{x}$.

- Of the two parameters for the regression model, we are most interested in the slope parameter, $\beta_1$. It is this parameter which tells us how the mean of $y$ changes, as $x$ changes. If the value of this parameter is not zero, this is suggestive of a relationship between $x$ and $y$. In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
Interpreting the Parameter Estimates

- The slope, $b_1$, is the estimated change in mean $y$ per one unit increase in $x$.
- The intercept, $b_0$, is the estimated mean of $y$ when $x = 0$.
  - Due to the structure of the assumed model, namely,

$$\mu_{y|x} = \beta_0 + \beta_1 x,$$

the predicted value of $\mu_{y|0}$, $b_0$, can assume any real number.
  - This can lead to a misleading interpretation of $\beta_0$.
  - To remedy this notion, whenever $x$ is a continuous variable, we “center” the values of $x$ about its sample mean, $\bar{x}$. In this case, $\beta_0$ is the mean value of $y$ given $x = \bar{x}$.

- Of the two parameters for the regression model, we are most interested in the slope parameter, $\beta_1$. It is this parameter which tells us how the mean of $y$ changes, as $x$ changes. If the value of this parameter is not zero, this is suggestive of a relationship between $x$ and $y$. In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
Interpreting the Parameter Estimates

- The slope, \( b_1 \), is the estimated change in mean \( y \) per one unit increase in \( x \).
- The intercept, \( b_0 \), is the estimated mean of \( y \) when \( x = 0 \).
  - Due to the structure of the assumed model, namely,

\[
\mu_{y|\mathbf{x}} = \beta_0 + \beta_1 x,
\]

the predicted value of \( \mu_{y|0}, b_0 \), can assume any real number.
  - This can lead to a misleading interpretation of \( \beta_0 \).
  - To remedy this notion, whenever \( x \) is a continuous variable, we “center” the values of \( x \) about its sample mean, \( \bar{x} \). In this case, \( \beta_0 \) is the mean value of \( y \) given \( x = \bar{x} \).

- Of the two parameters for the regression model, we are most interested in the slope parameter, \( \beta_1 \). It is this parameter which tells us how the mean of \( y \) changes, as \( x \) changes. If the value of this parameter is not zero, this is suggestive of a relationship between \( x \) and \( y \). In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
The slope, $b_1$, is the estimated change in mean $y$ per one unit increase in $x$.

The intercept, $b_0$, is the estimated mean of $y$ when $x = 0$.

- Due to the structure of the assumed model, namely, $y \mid x = \beta_0 + \beta_1 x$,

the predicted value of $\mu_{y \mid 0}$, $b_0$, can assume any real number.

- This can lead to a misleading interpretation of $\beta_0$.

- To remedy this notion, whenever $x$ is a continuous variable, we “center” the values of $x$ about its sample mean, $\bar{x}$. In this case, $\beta_0$ is the mean value of $y$ given $x = \bar{x}$.

- Of the two parameters for the regression model, we are most interested in the slope parameter, $\beta_1$. It is this parameter which tells us how the mean of $y$ changes, as $x$ changes. If the value of this parameter is not zero, this is suggestive of a relationship between $x$ and $y$. In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
Interpreting the Parameter Estimates

- The slope, $b_1$, is the estimated change in mean $y$ per one unit increase in $x$.
- The intercept, $b_0$, is the estimated mean of $y$ when $x = 0$.
  - Due to the structure of the assumed model, namely,
    \[ \mu_{y|x} = \beta_0 + \beta_1 x, \]
    the predicted value of $\mu_{y|0}$, $b_0$, can assume any real number.
    - This can lead to a misleading interpretation of $\beta_0$.
    - To remedy this notion, whenever $x$ is a continuous variable, we “center” the values of $x$ about its sample mean, $\bar{x}$. In this case, $\beta_0$ is the mean value of $y$ given $x = \bar{x}$.
- Of the two parameters for the regression model, we are most interested in the slope parameter, $\beta_1$. It is this parameter which tells us how the mean of $y$ changes, as $x$ changes. If the value of this parameter is not zero, this is suggestive of a relationship between $x$ and $y$. In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
The slope, $b_1$, is the estimated change in mean $y$ per one unit increase in $x$.

The intercept, $b_0$, is the estimated mean of $y$ when $x = 0$.
- Due to the structure of the assumed model, namely,

$$\mu_{y|x} = \beta_0 + \beta_1 x,$$

the predicted value of $\mu_{y|0}$, $b_0$, can assume any real number.
- This can lead to a misleading interpretation of $\beta_0$.
- To remedy this notion, whenever $x$ is a continuous variable, we “center” the values of $x$ about its sample mean, $\bar{x}$. In this case, $\beta_0$ is the mean value of $y$ given $x = \bar{x}$.

Of the two parameters for the regression model, we are most interested in the slope parameter, $\beta_1$. It is this parameter which tells us how the mean of $y$ changes, as $x$ changes. If the value of this parameter is not zero, this is suggestive of a relationship between $x$ and $y$. In Section 12.4, we will see how to conduct statistical tests for this regression parameter.
Example 12.3

The Parameter Estimate, $b_1$

Consider again the idea of predicting the mean fuel consumption $(y)$, given temperature $(x)$.

For the eight data values given (see Section 12.1 slides), we find

$$SS_{xy} = -179.6475 \quad \text{and} \quad SS_{xx} = 1404.355,$$

for which it holds

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = -0.1279217.$$
Example 12.3

The Parameter Estimate, $b_1$

Consider again the idea of predicting the mean fuel consumption ($y$), given temperature ($x$).

For the eight data values given (see Section 12.1 slides), we find

$$SS_{xy} = -179.6475$$ and $$SS_{xx} = 1404.355,$$

for which it holds

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = -0.1279217.$$
Example 12.3

The Parameter Estimate, $b_0$

We also find

$$\bar{x} = 43.98 \quad \text{and} \quad \bar{y} = 10.2125,$$

so that

$$b_0 = \bar{y} - b_1 \bar{x} = 10.2125 - (-0.1279)(43.98) = 15.8375.$$
Example 12.3

The Least Squares Regression Line

Thus, based on these data, the least squares regression line, $y$ regressed on $x$, is given by

$$\hat{y} = b_0 + b_1 x = 15.84 - 0.1279x.$$  

This equation says, the predicted mean fuel consumption when the temperature is zero degrees Fahrenheit, is 15.84 MMcf.

The equation also says, for every one degree Fahrenheit increase, the predicted mean fuel consumption decreases by 0.13 MMcf.
Thus, based on these data, the least squares regression line, $y$ regressed on $x$, is given by

$$\hat{y} = b_0 + b_1 x = 15.84 - 0.1279x.$$

This equation says, the predicted mean fuel consumption when the temperature is zero degrees Fahrenheit, is 15.84 MMcf.

The equation also says, for every one degree Fahrenheit increase, the predicted mean fuel consumption decreases by 0.13 MMcf.
Example 12.3

The Least Squares Regression Line

Thus, based on these data, the least squares regression line, \( y \) regressed on \( x \), is given by

\[
\hat{y} = b_0 + b_1 x = 15.84 - 0.1279 x.
\]

This equation says, the predicted mean fuel consumption when the temperature is zero degrees Fahrenheit, is 15.84 MMcf.

The equation also says, for every one degree Fahrenheit increase, the predicted mean fuel consumption decreases by 0.13 MMcf.
Example 12.3

The Least Squares Regression Line, $\hat{y} = 15.84 - 0.128x$
Example 12.3

Using the Least Squares Regression Line for Prediction

- At a temperature of 28°F, what is the predicted mean fuel consumption?
- At a temperature of 40°F, what is the predicted mean fuel consumption?
Example 12.3

Using the Least Squares Regression Line for Prediction

- At a temperature of 28°F, what is the predicted mean fuel consumption?
  - It is,

\[ \hat{y} = 15.84 - 0.1279(28) = 12.26 \text{ MMcf}. \]

- At a temperature of 40°F, what is the predicted mean fuel consumption?
Example 12.3

Using the Least Squares Regression Line for Prediction

- At a temperature of 28°F, what is the predicted mean fuel consumption?
  - It is,
    \[ \hat{y} = 15.84 - 0.1279(28) = 12.26 \text{ MMcf}. \]

- At a temperature of 40°F, what is the predicted mean fuel consumption?
Example 12.3

Using the Least Squares Regression Line for Prediction

- At a temperature of 28 °F, what is the predicted mean fuel consumption?
  - It is,

\[ \hat{y} = 15.84 - 0.1279(28) = 12.26 \text{ MMcf.} \]

- At a temperature of 40 °F, what is the predicted mean fuel consumption?
  - It is,

\[ \hat{y} = 15.84 - 0.1279(40) = 10.72 \text{ MMcf.} \]
Example 12.3

Predicting Fuel Consumption at 40°F

The least squares line
\[ \hat{y} = 15.84 - 0.1279x \]

The point estimate of mean fuel consumption when \( x = 40 \)

\[ \hat{y} = 10.72 \]
The least squares regression line only tells us about the linear relationship between the two variables $x$ and $y$.

- Always plot your data to verify a linear relationship exists between your variables.
- The least squares regression line should only be used to predict a mean response, within the region of observed data values (your textbook refers to this region as the experimental region). This is called interpolation.
- Since we do not observe data outside of the experimental region, you should never use the least squares regression line to predict a mean response outside of your experimental region. Doing so is called extrapolation.
The least squares regression line only tells us about the linear relationship between the two variables $x$ and $y$.

- Always plot your data to verify a linear relationship exists between your variables.

- The least squares regression line should only be used to predict a mean response, within the region of observed data values (your textbook refers to this region as the experimental region). This is called **interpolation**.

- Since we do not observe data outside of the experimental region, you should never use the least squares regression line to predict a mean response outside of your experimental region. Doing so is called **extrapolation**.
The least squares regression line only tells us about the linear relationship between the two variables $x$ and $y$.

- Always plot your data to verify a linear relationship exists between your variables.

The least squares regression line should only be used to predict a mean response, within the region of observed data values (your textbook refers to this region as the experimental region). This is called **interpolation**.

- Since we do not observe data outside of the experimental region, you should never use the least squares regression line to predict a mean response outside of your experimental region. Doing so is called **extrapolation**.
The least squares regression line only tells us about the linear relationship between the two variables $x$ and $y$.

- Always plot your data to verify a linear relationship exists between your variables.

The least squares regression line should only be used to predict a mean response, within the region of observed data values (your textbook refers to this region as the experimental region). This is called **interpolation**.

Since we do not observe data outside of the experimental region, you should **never** use the least squares regression line to predict a mean response outside of your experimental region. Doing so is called **extrapolation**.
Example of Extrapolation
Suppose we have collected data within timeframe 2 shown in the figure and construct the least squares regression line to predict mean \( \log(\text{viable cells}) \) from time (hours).
Using the least squares line to predict mean log(viable cells) within timeframe 2 would be appropriate, since a linear relationship seems to exist between log(viable cells) and time.
However, using this least squares regression line to predict mean log(viable cells) from time, within any timeframe other than timeframe 2, would lead to very misleading predictions.
Cautions About Simple Linear Regression

Example of Extrapolation

- Bottom line, never extrapolate outside of your experimental region.
12.4: Testing the Significance of the Slope Parameter

Outline

12.4: Testing the Significance of the Slope Parameter
- Introduction
- Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$
A simple linear regression model is essentially useless unless there is a significant linear relationship between $y$ and $x$.

- If $\beta_1 = 0$, then $\mu_{y|x} = \beta_0$, for all values of $x$. That is, no linear relationship exists between $y$ and $x$.
- On the other hand, if $\beta_1 \neq 0$, then $\mu_{y|x} = \beta_0 + \beta_1 x$, and the mean response value is dependent on the value of $x$. That is, there is a linear relationship between $y$ and $x$.

Within this section, we will investigate statistical inference for the regression parameter, $\beta_1$. 
Introduction

Idea Surrounding Inference for $\beta_1$

- A simple linear regression model is essentially useless unless there is a significant linear relationship between $y$ and $x$.
  - If $\beta_1 = 0$, then $\mu_{y|x} = \beta_0$, for all values of $x$. That is, no linear relationship exists between $y$ and $x$.
  - On the other hand, if $\beta_1 \neq 0$, then $\mu_{y|x} = \beta_0 + \beta_1 x$, and the mean response value is dependent on the value of $x$. That is, there is a linear relationship between $y$ and $x$.
- Within this section, we will investigate statistical inference for the regression parameter, $\beta_1$. 
A simple linear regression model is essentially useless unless there is a significant linear relationship between $y$ and $x$.

- If $\beta_1 = 0$, then $\mu_{y|x} = \beta_0$, for all values of $x$. That is, no linear relationship exists between $y$ and $x$.
- On the other hand, if $\beta_1 \neq 0$, then $\mu_{y|x} = \beta_0 + \beta_1 x$, and the mean response value is dependent on the value of $x$. That is, there is a linear relationship between $y$ and $x$.

Within this section, we will investigate statistical inference for the regression parameter, $\beta_1$. 
Idea Surrounding Inference for $\beta_1$

- A simple linear regression model is essentially useless unless there is a significant linear relationship between $y$ and $x$.
  - If $\beta_1 = 0$, then $\mu_{y|x} = \beta_0$, for all values of $x$. That is, no linear relationship exists between $y$ and $x$.
  - On the other hand, if $\beta_1 \neq 0$, then $\mu_{y|x} = \beta_0 + \beta_1 x$, and the mean response value is dependent on the value of $x$. That is, there is a linear relationship between $y$ and $x$.
- Within this section, we will investigate statistical inference for the regression parameter, $\beta_1$. 
Introduction

Working Definitions/Assumptions

- For any value of the independent variable \(x\) we assume the population of errors \(\epsilon\) is normally distributed with mean zero and unknown variance \(\sigma^2\) (see tan box at bottom of page 486). That is, \(\epsilon \sim N(0, \sigma)\).

- The sum of squared errors, \(SSE\), is defined by
  \[
  SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.
  \]

- Correspondingly, the mean square error, \(MSE\), defined by
  \[
  MSE = \frac{SSE}{n - 2},
  \]
  is the point estimate for \(\sigma^2\).
12.4: Testing the Significance of the Slope Parameter

Introduction

Working Definitions/Assumptions

- For any value of the independent variable \( x \) we assume the population of errors \( (\epsilon) \) is normally distributed with mean zero and unknown variance \( \sigma^2 \) (see tan box at bottom of page 486). That is, \( \epsilon \sim N(0, \sigma) \).

- The **sum of squared errors**, \( SSE \), is defined by

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

- Correspondingly, the **mean square error**, \( MSE \), defined by

\[
MSE = \frac{SSE}{n - 2},
\]

is the point estimate for \( \sigma^2 \).
For any value of the independent variable \( x \) we assume the population of errors \( \epsilon \) is normally distributed with mean zero and unknown variance \( \sigma^2 \) (see tan box at bottom of page 486). That is, \( \epsilon \sim N(0, \sigma) \).

The **sum of squared errors**, \( SSE \), is defined by

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

Correspondingly, the **mean square error**, \( MSE \), defined by

\[
MSE = \frac{SSE}{n - 2},
\]

is the point estimate for \( \sigma^2 \).
The Sampling Distribution of \((b_1 - \beta_1)(s_{b_1})^{-1}\)

Under the regression assumptions (see Section 12.3) – which include the assumption that \(\epsilon \sim N(0, \sigma)\) – it follows that the sampling distribution of the ratio

\[
\frac{b_1 - \beta_1}{s_{b_1}},
\]

has a \(t\) distribution with \(n - 2\) degrees of freedom, where

\[
s_{b_1} = \sqrt{\frac{MSE}{SS_{xx}}}.\]
Define the test statistic, $t$, where

$$t = \frac{b_1}{s_{b_1}}; \quad s_{b_1} = \sqrt{\frac{MSE}{SS_{xx}}}.$$

Suppose the regression assumptions of Section 12.3 hold. Then, at the $\alpha$ level of significance, the null hypothesis, $H_0 : \beta_1 = 0$, calls for rejection in favor of $H_a : \beta_1 \neq 0$, if and only if $|t| > t_{\alpha/2}$, where $t_{\alpha/2}$ is based on the $t$ distribution with $n - 2$ degrees of freedom.
Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

### Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The total source row of the table summarizes the variance in the response variable. The degrees of freedom value for this row is $n - 1$, and the SS (sum of squares) value for this row is $\sum_{i=1}^{n} (y_i - \bar{y})^2$. Let $s_y^2$ denote the sample variance for the response variable. It is,

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{SS}{DF}.$$
Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

**Interpreting Software Output: ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Breakdown of the ANOVA Table**

The regression source row of the table gives the numerical summary of the regression model, corresponding to all regression parameter estimates associated with independent variables. In the case of simple linear regression, this entails summarizing the single regression parameter estimate, $b_1$. 

---

Bill Welbourn

Chapter 12: STAT 2300 – Summer 2009
12.4: Testing the Significance of the Slope Parameter

Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The DF (degrees of freedom) value corresponds to the number of independent variables of the regression model. In the case of simple linear regression, our model entails a single independent variable.
12.4: Testing the Significance of the Slope Parameter

Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The SS (sum of squares) value quantifies the explained variation in the response due to the regression of $y$ on $x$. The closer this value is to the total variation, the better the model is at predicting the mean response across the values of $x$. 
Hypothesis Test, Testing the null hypothesis, \( H_0 : \beta_1 = 0 \)

## Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Breakdown of the ANOVA Table

The MS (mean square error) value is the mean of the sum of squares (SS) due to the regression of \( y \) on \( x \). The value of MS is SS divided by the degrees of freedom (DF).
Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The F test statistic value is the ratio of the mean square error (MS) due to the regression of $y$ on $x$ (i.e., 22.9808), divided by the mean square error due to residual error (i.e., 0.4280). It is,

$$F = \frac{22.9808}{0.4280} = 53.6949.$$
12.4: Testing the Significance of the Slope Parameter

Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The p-value is the probability of obtaining an F test statistic value at least as extreme as we observed (i.e., $F = 53.6949$), under the assumption that the null hypothesis, $H_0 : \beta_1 = 0$, is true. It is calculated based on the upper tail area of the $F$ distribution with numerator degrees of freedom, $df_1 = 1$ (i.e., DF Regression), and denominator degrees of freedom, $df_2 = 6$ (i.e., DF Residual). Its numerical value is given by

$$p\text{-value} = P(F \geq 53.6949) = 0.0003.$$
Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The residual source row gives the numerical summary of the variation in the response not explained by the regression of \( y \) on \( x \) (i.e., unexplained variation in \( y \)). The sum of squares value is \( SSE \) and the degrees of freedom is \( n - 2 \). Correspondingly, the mean square value is \( MSE \).
12.4: Testing the Significance of the Slope Parameter

Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

Interpreting Software Output: ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-Value (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>22.9808</td>
<td>22.9808</td>
<td>53.6949</td>
<td>0.0003</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>2.5679</td>
<td>0.4280</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>25.5488</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breakdown of the ANOVA Table

The coefficient of determination, denoted by $r^2$ ($r$ is the Pearson correlation coefficient for the variables $x$ and $y$), is the proportion of variation in the response explained by the regression of $y$ regressed on $x$. Its numerical value is given by

$$r^2 = \frac{SS \text{ due to regression}}{SS \text{ Total}} = \frac{22.9808}{25.5488} = 0.8995.$$  

Values for the coefficient of determination exceeding 0.90 indicate a good model fit.
Hypothesis Test, Testing the null hypothesis, \( H_0 : \beta_1 = 0 \)

Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
Hypothesis Test, Testing the null hypothesis, $H_0 : \beta_1 = 0$

Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Breakdown of the Parameter Estimate Table

To conduct inference for $\beta_1$, we need only be concerned with the row corresponding to the slope parameter.
Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Breakdown of the Parameter Estimate Table

This value corresponds to the estimate for $\beta_1$, namely $b_1$. 
Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Breakdown of the Parameter Estimate Table

This value corresponds to the standard error of the estimate $b_1$, namely $s_{b_1}$. 
Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>$t$ Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Breakdown of the Parameter Estimate Table

The $t$ test statistic value corresponding to the null hypothesis, $H_0 : \beta_1 = 0$. Its sampling distribution under $H_0$ is the $t$ distribution with $n - 2$ degrees of freedom.
12.4: Testing the Significance of the Slope Parameter

Hypothesis Test, Testing the null hypothesis, \( H_0 : \beta_1 = 0 \)

### Interpreting Software Output: Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.8379</td>
<td>0.8018</td>
<td>19.7535</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.1279</td>
<td>0.0175</td>
<td>-7.3277</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

### Breakdown of the Parameter Estimate Table

The p-value is the probability of obtaining a test statistic value at least as extreme (alternative hypothesis is \( H_a : \beta_1 \neq 0 \)) as we observed, under the assumption that \( H_0 \) is true. Its numerical value is given by

\[
p\text{-value} = 2P(t \geq | -7.3277|) = 0.0003.
\]