Chapter 10: Inference for Regression

MATH 560-01
Statistical Data Analysis

March 22, 2021

These slides are based on material from *Introduction to the Practice of Statistics* by David S. Moore, George P. McCabe, and Bruce A. Craig, 9th edition.

Sections

10.1 Simple Linear Regression

After completing this section, students should be able to:

- Make a *scatterplot* for two quantitative variables.

10.1 Simple Linear Regression

After completing this section, students should be able to:

- Make a *scatterplot* for two quantitative variables.
- Compute and interpret the *correlation coefficient*.

10.1 Simple Linear Regression

After completing this section, students should be able to:

- Make a *scatterplot* for two quantitative variables.
- Compute and interpret the *correlation coefficient*.
- Find the equation of the *regression line* for predicting a response variable based on an explanatory variable and use it for predictions or for estimation of the mean response values.

10.1 Simple Linear Regression

After completing this section, students should be able to:

- Make a *scatterplot* for two quantitative variables.
- Compute and interpret the *correlation coefficient*.
- Find the equation of the *regression line* for predicting a response variable based on an explanatory variable and use it for predictions or for estimation of the mean response values.
- Compute and interpret r^2.

10.1 Simple Linear Regression

After completing this section, students should be able to:

- Make a *scatterplot* for two quantitative variables.
- Compute and interpret the *correlation coefficient*.
- Find the equation of the *regression line* for predicting a response variable based on an explanatory variable and use it for predictions or for estimation of the mean response values.
- Compute and interpret r^2.
- Compute the *least squares estimates* of the *intercept*, *slope*, and *standard deviation* in the simple linear regression model with normal errors.

10.1 Simple Linear Regression

Scatterplots

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.

10.1 Simple Linear Regression

Scatterplots

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.
- **Response variable**: measures the outcome of the study
- **Explanatory variable**: explains or causes a change in the response

10.1 Simple Linear Regression

Scatterplots

- Two variables are **associated** if knowing the value of one of the variables for a case tells you something about the value of the other variable for that case.
- **Response variable**: measures the outcome of the study
- **Explanatory variable**: explains or causes a change in the response
- Two variables have a **positive association** when high values of one variable tend to accompany high values of the other variable and low values of one variable tend to accompany low values of the other variable.
- Two variables have a **negative association** when high values of one variable tend to accompany low values of the other variable and vice versa.

10.1 Simple Linear Regression

Scatterplots

- A **scatterplot** shows the relationship between two quantitative variables measured on the same cases. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each case in the data set appears as a point in the plot.

10.1 Simple Linear Regression

Scatterplots

- A **scatterplot** shows the relationship between two quantitative variables measured on the same cases. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each case in the data set appears as a point in the plot.
- When variables can be distinguished as explanatory and response, the explanatory variable is placed on the horizontal axis and the response variable is placed on the vertical axis.

10.1 Simple Linear Regression

Scatterplots

- A **scatterplot** shows the relationship between two quantitative variables measured on the same cases. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each case in the data set appears as a point in the plot.
- When variables can be distinguished as explanatory and response, the explanatory variable is placed on the horizontal axis and the response variable is placed on the vertical axis.
- Scatterplots can be useful for visually identifying **overall patterns**, **direction**, and **strength** of association between two variables as well as **outliers** (individual cases which differ significantly from the overall pattern).

10.1 Simple Linear Regression

Correlation

- The correlation coefficient is a measure of the **direction** and **strength** of a **linear** relationship.

10.1 Simple Linear Regression

Correlation

- The correlation coefficient is a measure of the **direction** and **strength** of a **linear** relationship.
- It is calculated using the mean and the standard deviation of both the x and y variables.

10.1 Simple Linear Regression

Correlation

- The correlation coefficient is a measure of the **direction** and **strength** of a **linear** relationship.
- It is calculated using the mean and the standard deviation of both the x and y variables.
- Correlation can only be used to describe **quantitative** variables. Categorical variables don't have means and standard deviations.

10.1 Simple Linear Regression

Computing the correlation coefficient:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

where $\bar{x}$ and s_x are the mean and standard deviation of the x-values and $\bar{y}$ and s_y are the mean and standard deviation of the y-values.

10.1 Simple Linear Regression

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})$$

10.1 Simple Linear Regression

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}\bar{y}
\end{aligned}
$$

10.1 Simple Linear Regression

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x} n \bar{y} - \bar{y} n \bar{x} + n\bar{x}\bar{y}
\end{aligned}
$$

10.1 Simple Linear Regression

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y}) \\
&= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\bar{y} \\
&= \sum_{i=1}^{n}x_iy_i-\bar{x}n\bar{y}-\bar{y}n\bar{x}+n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}
\end{aligned}$$

10.1 Simple Linear Regression

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y}) \\
&= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\bar{y} \\
&= \sum_{i=1}^{n}x_iy_i-\bar{x}n\bar{y}-\bar{y}n\bar{x}+n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}
\end{aligned}
$$

Alternate form of r:

$$
r=\frac{1}{n-1}\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{s_xs_y}
$$

10.1 Simple Linear Regression

Properties of r:

- r does not distinguish between x and y

10.1 Simple Linear Regression

Properties of r:

- r does not distinguish between x and y
- r has no units of measurement

10.1 Simple Linear Regression

Properties of r:

- r does not distinguish between x and y
- r has no units of measurement
- $-1 \leq r \leq 1$; Values of r close to -1 or 1 indicate that the points lie close to a straight line

10.1 Simple Linear Regression

Properties of r:

- r does not distinguish between x and y
- r has no units of measurement
- $-1 \leq r \leq 1$; Values of r close to -1 or 1 indicate that the points lie close to a straight line
- $r = -1$ indicates all points lie exactly on a straight line with negative slope
- $r = 1$ indicates all points lie exactly on a straight line with positive slope

10.1 Simple Linear Regression

Properties of r:

- r does not distinguish between x and y
- r has no units of measurement
- $-1 \leq r \leq 1$; Values of r close to -1 or 1 indicate that the points lie close to a straight line
- $r = -1$ indicates all points lie exactly on a straight line with negative slope
- $r = 1$ indicates all points lie exactly on a straight line with positive slope
- r is not resistant to outliers

10.1 Simple Linear Regression

Example 10A: Here are the average temperatures (°F) during the months of February through May in Lafayette, IN.

Month	February(1.5)	March(2.5)	April(3.5)	May(4.5)
Temperature	30	41	51	62

Compute the correlation coefficient between Month and Temperature for this data set.

10.1 Simple Linear Regression

- *Answer*:

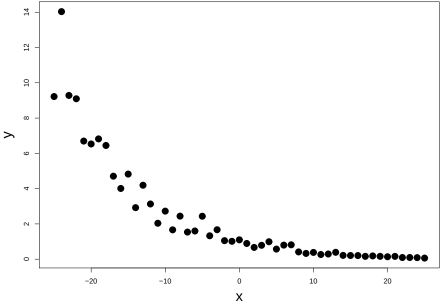
10.1 Simple Linear Regression

Cautions about Correlation

- The correlation coefficient only measures linear association.

Cautions about Correlation

- The correlation coefficient only measures linear association.
- Example: scatterplot of x vs. y

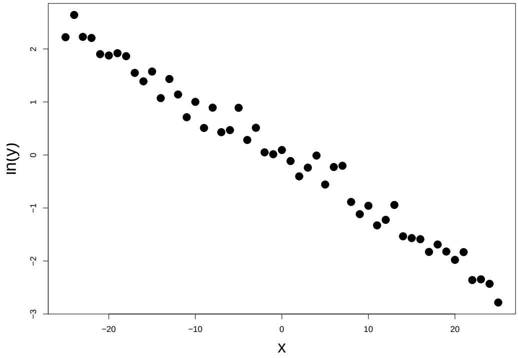


$r = -0.828$

10.1 Simple Linear Regression

Cautions about Correlation

- log transformation: scatterplot of x vs. $\ln(y)$

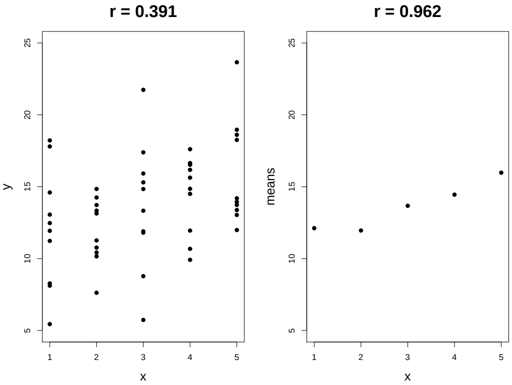


$r = -0.990$

10.1 Simple Linear Regression

Cautions about Correlation

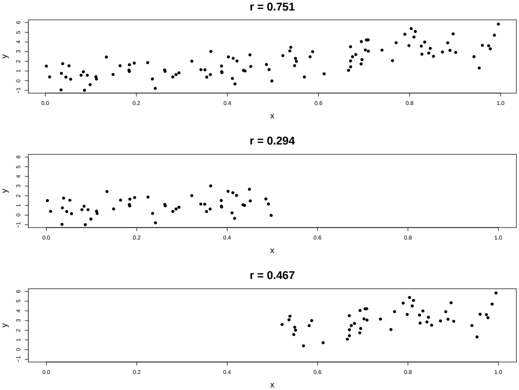
- Correlations **based on averages** are usually higher than if we used data for individuals.
- Example:



10.1 Simple Linear Regression

Cautions about Correlation

- Correlations **based on data with a restricted range** are often lower than it would be with the full range.
- Example:



10.1 Simple Linear Regression

Least-Squares Regression

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.

10.1 Simple Linear Regression

Least-Squares Regression

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.
- We often use a regression line to predict the value of y for a given value of x.

10.1 Simple Linear Regression

Least-Squares Regression

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.
- We often use a regression line to predict the value of y for a given value of x.
- In regression, the distinction between the explanatory and the response variable is important.

10.1 Simple Linear Regression

Least-Squares Regression

The regression line has the form $\hat{y} = b_0 + b_1 x$ where

- $\hat{y}$ is the predicted value of y at x

10.1 Simple Linear Regression

Least-Squares Regression

The regression line has the form $\hat{y} = b_0 + b_1 x$ where

- $\hat{y}$ is the predicted value of y at x
- b_0 is the intercept; this gives the predicted value of y when $x = 0$

10.1 Simple Linear Regression

Least-Squares Regression

The regression line has the form $\hat{y} = b_0 + b_1 x$ where

- $\hat{y}$ is the predicted value of y at x
- b_0 is the intercept; this gives the predicted value of y when $x = 0$
- b_1 is the slope; this gives the amount by which y is predicted to change when x **increases** by 1 unit

10.1 Simple Linear Regression

Least-Squares Regression

- The most common method to obtain the coefficients for the regression line is the *method of least squares*.

10.1 Simple Linear Regression

Least-Squares Regression

- The most common method to obtain the coefficients for the regression line is the *method of least squares*.
- This method obtains the line which makes the sum of the squares of the vertical distances of the data points to the line as small as possible; that is, it finds the values of b_0 and b_1 which minimize $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

10.1 Simple Linear Regression

Least-Squares Regression

- The most common method to obtain the coefficients for the regression line is the *method of least squares*.
- This method obtains the line which makes the sum of the squares of the vertical distances of the data points to the line as small as possible; that is, it finds the values of b_0 and b_1 which minimize $\displaystyle\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.
- The least squares regression line is $\hat{y} = b_0 + b_1 x$ where

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

Derivation of Least Squares Estimates

Let $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. To minimize this function, we differentiate with respect to β_0 and β_1 as follows:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\
\frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i).
\end{aligned}$$

10.1 Simple Linear Regression

Derivation of Least Squares Estimates

Let $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$. To minimize this function, we differentiate with respect to β_0 and β_1 as follows:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\
\frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i).
\end{aligned}$$

Setting the partial derivatives equal to 0 and letting b_0 and b_1 denote the solution, we obtain

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0 \tag{1}$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0. \tag{2}$$

10.1 Simple Linear Regression

Using properties of sums, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \tag{3}$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2. \tag{4}$$

10.1 Simple Linear Regression

Using properties of sums, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \tag{3}$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2. \tag{4}$$

Dividing both sides of equation (3) by n, it follows that

$$\bar{y} = b_0 + b_1 \bar{x}. \tag{5}$$

10.1 Simple Linear Regression

Using properties of sums, we get

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \tag{3}$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2. \tag{4}$$

Dividing both sides of equation (3) by n, it follows that

$$\bar{y} = b_0 + b_1 \bar{x}. \tag{5}$$

Solving (5) for b_0, we get

$$b_0 = \bar{y} - b_1 \bar{x}. \tag{6}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1 \bar{x}) n \bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1 \bar{x}) n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

Equation (7) can be solved for b_1 as follows:

$$b_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) \quad = \quad \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1\bar{x})n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

Equation (7) can be solved for b_1 as follows:

$$\begin{aligned} b_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ b_1 \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) &= \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}) \end{aligned}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1\bar{x})n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

Equation (7) can be solved for b_1 as follows:

$$\begin{aligned}
b_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
b_1 \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \\
b_1 s_x^2 &= \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{s_x s_y} s_x s_y
\end{aligned}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1\bar{x})n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

Equation (7) can be solved for b_1 as follows:

$$\begin{aligned}
b_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
b_1 \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) &= \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}) \\
b_1 s_x^2 &= \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{s_x s_y} s_x s_y \\
b_1 s_x^2 &= r s_x s_y
\end{aligned}$$

10.1 Simple Linear Regression

Next, we substitute (6) into (4) to obtain

$$\sum_{i=1}^{n} x_i y_i = (\bar{y} - b_1\bar{x})n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2. \tag{7}$$

Equation (7) can be solved for b_1 as follows:

$$\begin{aligned}
b_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
b_1 \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \\
b_1 s_x^2 &= \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{s_x s_y} s_x s_y \\
b_1 s_x^2 &= r s_x s_y \\
b_1 &= r\frac{s_y}{s_x}.
\end{aligned}$$

10.1 Simple Linear Regression

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

10.1 Simple Linear Regression

$$
\begin{aligned}
Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} \{(y_i - b_0 - b_1 x_i) + (b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2
\end{aligned}
$$

10.1 Simple Linear Regression

$$
\begin{aligned}
Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} \{(y_i - b_0 - b_1 x_i) + (b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 + 2(b_0 - \beta_0) \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \\
&+ 2(b_1 - \beta_1) \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) + \sum_{i=1}^{n} \{(b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2
\end{aligned}
$$

10.1 Simple Linear Regression

$$
\begin{aligned}
Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} \{(y_i - b_0 - b_1 x_i) + (b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 + 2(b_0 - \beta_0) \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \\
&+ \quad 2(b_1 - \beta_1) \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) + \sum_{i=1}^{n} \{(b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 + \sum_{i=1}^{n} \{(b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2
\end{aligned}
$$

10.1 Simple Linear Regression

$$
\begin{aligned}
Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} \{(y_i - b_0 - b_1 x_i) + (b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 + 2(b_0 - \beta_0) \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \\
&+ \quad 2(b_1 - \beta_1) \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) + \sum_{i=1}^{n} \{(b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 + \sum_{i=1}^{n} \{(b_0 - \beta_0) + (b_1 - \beta_1) x_i\}^2
\end{aligned}
$$

So, $Q(\beta_0, \beta_1)$ is minimized when $\beta_0 = b_0$ and $\beta_1 = b_1$ since the second sum on the right is 0 for this choice of β_0 and β_1 but positive for all other values of β_0 and β_1.

10.1 Simple Linear Regression

Properties of Least-Squares Regression

- The equation for the slope $b_1 = r\frac{s_y}{s_x}$ tells us that y is predicted to change by r standard deviations in y if x changes by one standard deviation in x.

10.1 Simple Linear Regression

Properties of Least-Squares Regression

- The equation for the slope $b_1 = r\frac{s_y}{s_x}$ tells us that y is predicted to change by r standard deviations in y if x changes by one standard deviation in x.
- The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.

10.1 Simple Linear Regression

Properties of Least-Squares Regression

- The equation for the slope $b_1 = r\frac{s_y}{s_x}$ tells us that y is predicted to change by r standard deviations in y if x changes by one standard deviation in x.
- The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.
- The distinction between the explanatory variable and the response variable is important in regression since we are predicting y based on x (not x based on y).

10.1 Simple Linear Regression

Properties of Least-Squares Regression

- The equation for the slope $b_1 = r\frac{s_y}{s_x}$ tells us that y is predicted to change by r standard deviations in y if x changes by one standard deviation in x.
- The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.
- The distinction between the explanatory variable and the response variable is important in regression since we are predicting y based on x (not x based on y).
- r^2 gives a measure of success of regression in explaining the response; specifically, $r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ is the fraction of variation in the values of y explained by the regression line compared with just using the mean to predict y.

- **Example 10B**: Here are the average temperatures (°F) during the months of February through May in Lafayette, IN.

Time	1.5	2.5	3.5	4.5
Temperature	30	41	51	62

(a) What is the equation of the regression line for predicting Temperature based on Month for this data set?
(b) Based on the equation of the regression line, what is the predicted average temperature for March? How much does it differ from the observed value 41?
(c) Based on the equation of the regression line, what is the predicted average temperature for December? Does this prediction make sense?
(d) What proportion of variation in Temperatures is explained by the regression line?

10.1 Simple Linear Regression

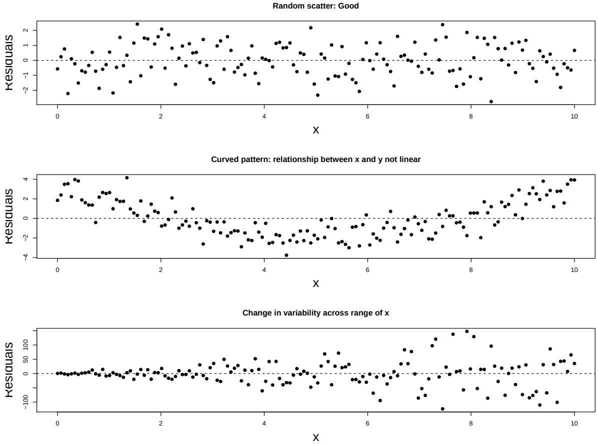
- *Answers*:

<u>Residual Plots</u>

- A **residual plot** is a scatterplot with the explanatory variable on the x-axis and the corresponding residuals on the y-axis. The residual plot is a visual tool to help assess the linearity assumption needed for regression; when valid, the residual plot should show no pattern.

10.1 Simple Linear Regression

Examples of residual plots



Cautions about Correlation and Regression

- **Outliers**: observations that lie outside the overall pattern of other observations; outliers in the y-direction have large residuals; outliers in the x-direction may not have large residuals

10.1 Simple Linear Regression

Cautions about Correlation and Regression

- **Outliers**: observations that lie outside the overall pattern of other observations; outliers in the y-direction have large residuals; outliers in the x-direction may not have large residuals
- **Influential points**: points which change the fit of the regression line compared with what it would be without the point; outliers in the x-direction are often potentially influential points

10.1 Simple Linear Regression

Cautions about Correlation and Regression

- **Lurking variables**: a third variable which may influence the interpretation of relationships among the explanatory and response variables

10.1 Simple Linear Regression

Cautions about Correlation and Regression

- **Lurking variables**: a third variable which may influence the interpretation of relationships among the explanatory and response variables
- Example: What is wrong with the following reasoning: "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage."

10.1 Simple Linear Regression

Cautions about Correlation and Regression

- **Lurking variables**: a third variable which may influence the interpretation of relationships among the explanatory and response variables
- Example: What is wrong with the following reasoning: "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage."
- *Answer: There is a lurking variable severity of fire which is associated with both observed variables. More severe fires will likely have the more firefighters present and more damage. However, increasing the number of firefighters at a particular fire likely reduces the damage for that fire.* **Association does not imply causation.**

10.1 Simple Linear Regression

Inference for Linear Regression

- Now we consider how to make inferences for the slope of the regression line and for the fitted and predicted values of the response variable for a given value of the explanatory variable.

10.1 Simple Linear Regression

Inference for Linear Regression

- Now we consider how to make inferences for the slope of the regression line and for the fitted and predicted values of the response variable for a given value of the explanatory variable.
- To make inferences, we need to distinguish between the sample and the population and make some assumptions about the distribution of the errors.

10.1 Simple Linear Regression

Inference for Linear Regression

- Now we consider how to make inferences for the slope of the regression line and for the fitted and predicted values of the response variable for a given value of the explanatory variable.
- To make inferences, we need to distinguish between the sample and the population and make some assumptions about the distribution of the errors.
- We assume that the population follows a **simple linear regression model** with true slope β_1, true intercept β_0, and true standard deviation σ.

10.1 Simple Linear Regression

Inference for Linear Regression

- Now we consider how to make inferences for the slope of the regression line and for the fitted and predicted values of the response variable for a given value of the explanatory variable.
- To make inferences, we need to distinguish between the sample and the population and make some assumptions about the distribution of the errors.
- We assume that the population follows a **simple linear regression model** with true slope β_1, true intercept β_0, and true standard deviation σ.
- Mathematically, this can be written as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where ε follows a $N(0, \sigma)$ distribution.

10.1 Simple Linear Regression

Inference for Linear Regression

- Then, we assume the observations in the sample

$$(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$$

are independent observations from the simple linear regression model.

10.1 Simple Linear Regression

Inference for Linear Regression

- Then, we assume the observations in the sample

$$(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$$

  are independent observations from the simple linear regression model.

- Specifically, this means that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

  where the errors ε_i are independent and Normally distributed with mean 0 and standard deviation σ.

10.1 Simple Linear Regression

Inference for Linear Regression

- Then, we assume the observations in the sample

$$(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$$

  are independent observations from the simple linear regression model.

- Specifically, this means that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

  where the errors ε_i are independent and Normally distributed with mean 0 and standard deviation σ.

- Our goal in regression analysis is to make inferences about the unknown parameters β_0, β_1, and σ and about the model.

10.1 Simple Linear Regression

Estimates of the Slope and Intercept

- Recall that the least-squares regression line is

$$\hat{y} = b_0 + b_1 x$$

where $b_1 = r\dfrac{s_y}{s_x}$ estimates the slope β_1 and $b_0 = \bar{y} - b_1\bar{x}$ estimates the intercept β_0.

10.1 Simple Linear Regression

Estimates of the Slope and Intercept

- Recall that the least-squares regression line is

$$\hat{y} = b_0 + b_1 x$$

  where $b_1 = r\dfrac{s_y}{s_x}$ estimates the slope β_1 and $b_0 = \bar{y} - b_1\bar{x}$ estimates the intercept β_0.
- The fitted value for the ith observation is $\hat{y}_i = b_0 + b_1 x_i$.

10.1 Simple Linear Regression

Estimates of the Slope and Intercept

- Recall that the least-squares regression line is

$$\hat{y} = b_0 + b_1 x$$

where $b_1 = r\dfrac{s_y}{s_x}$ estimates the slope β_1 and $b_0 = \bar{y} - b_1\bar{x}$ estimates the intercept β_0.

- The fitted value for the ith observation is $\hat{y}_i = b_0 + b_1 x_i$.
- Since the errors are Normally distributed and b_0 and b_1 are linear combinations of the y's (and thus the errors), the estimates b_0 and b_1 are also Normally distributed.

10.1 Simple Linear Regression

Estimates of the Slope and Intercept

- Recall that the least-squares regression line is

$$\hat{y} = b_0 + b_1 x$$

where $b_1 = r\dfrac{s_y}{s_x}$ estimates the slope β_1 and $b_0 = \bar{y} - b_1\bar{x}$ estimates the intercept β_0.

- The fitted value for the ith observation is $\hat{y}_i = b_0 + b_1 x_i$.
- Since the errors are Normally distributed and b_0 and b_1 are linear combinations of the y's (and thus the errors), the estimates b_0 and b_1 are also Normally distributed.
- Even if the ε_i's are not exactly Normal, a general form of the central limit theorem tells us that b_0 and b_1 will be approximately Normal when n is large.

10.1 Simple Linear Regression

Estimate of the Standard Deviation

- The residual for the ith observation is

$$
\begin{aligned}
e_i &= \text{observed response} - \text{predicted response} \\
&= y_i - \hat{y}_i.
\end{aligned}
$$

10.1 Simple Linear Regression

Estimate of the Standard Deviation

- The residual for the ith observation is

$$
\begin{aligned}
e_i &= \text{observed response} - \text{predicted response} \\
&= y_i - \hat{y}_i.
\end{aligned}
$$

- Property: $\displaystyle\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$

10.1 Simple Linear Regression

Estimate of the Standard Deviation

- The residual for the ith observation is

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i. \end{aligned}$$

- Property: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$
- The estimate of σ^2 depends on the residuals. For the simple linear regression model, the estimate of σ^2 is

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}.$$

Then $s = \sqrt{s^2}$ estimates σ.

10.1 Simple Linear Regression

Estimate of the Standard Deviation

- The residual for the ith observation is

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i. \end{aligned}$$

- Property: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$
- The estimate of σ^2 depends on the residuals. For the simple linear regression model, the estimate of σ^2 is

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}.$$

  Then $s = \sqrt{s^2}$ estimates σ.
- The quantity $n-2$ is called the degrees of freedom for s^2.

10.1 Simple Linear Regression

- **Example 10C**: Real estate is typically reassessed annually for property tax purposes. This assessed value is not necessarily the same as the fair market value of the property. A data set includes a SRS of assessment values (x) and prices (y) for 30 homes sold recently in a midwestern city (in thousands of dollars). The summary statistics for the data set are given below:
  $\bar{x} = 184.133, s_x = 45.428, \bar{y} = 195.840, s_y = 47.179, r = 0.91169.$
  (a) Compute the least squares estimates of the intercept, slope, and standard deviation of the simple linear regression model.
  (b) Based on the regression line, estimate the mean sale price of a home that is assessed at \$180,000.
  (c) Based on the regression line, predict the price of a home that is assessed at \$180,000.

10.1 Simple Linear Regression

- *Answers*:

10.2 More Detail about Simple Linear Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the slope and intercept of a simple linear regression model with normal errors

10.2 More Detail about Simple Linear Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the slope and intercept of a simple linear regression model with normal errors
- perform *one-sided* or *two-sided tests* for significance of the slope in a simple linear regression model with normal errors

10.2 More Detail about Simple Linear Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the slope and intercept of a simple linear regression model with normal errors
- perform *one-sided* or *two-sided tests* for significance of the slope in a simple linear regression model with normal errors
- compute *confidence intervals for the mean response* given the value of the explanatory variable for a simple linear regression model with normal errors

10.2 More Detail about Simple Linear Regression

After completing this section, students should be able to:

- compute *confidence intervals* for the slope and intercept of a simple linear regression model with normal errors
- perform *one-sided* or *two-sided tests* for significance of the slope in a simple linear regression model with normal errors
- compute *confidence intervals for the mean response* given the value of the explanatory variable for a simple linear regression model with normal errors
- compute *prediction intervals for a future observation* given the value of the explanatory variable for a simple linear regression model with normal errors

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- The formula for s^2 as well as many other important formulas for making inferences on the regression model depend on an analysis of estimate of the variance of the response variable.

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- The formula for s^2 as well as many other important formulas for making inferences on the regression model depend on an analysis of estimate of the variance of the response variable.
- We start by decomposing the overall deviation from the mean of a single observed response:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- The formula for s^2 as well as many other important formulas for making inferences on the regression model depend on an analysis of estimate of the variance of the response variable.
- We start by decomposing the overall deviation from the mean of a single observed response:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

- A very important algebraic identity holds:

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

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- Sum of squares total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- Sum of squares total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum of squares error = SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- Sum of squares total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum of squares error = SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Sum of squares model = SSM = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

10.2 More Detail about Simple Linear Regression

Analysis of Variance for Regression

- Sum of squares total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum of squares error = SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Sum of squares model = SSM = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $$\text{SST} = \text{SSE} + \text{SSM}$$

10.2 More Detail about Simple Linear Regression

Derivation of SST = SSE + SSM

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

10.2 More Detail about Simple Linear Regression

Derivation of SST = SSE + SSM

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)^2 \\
&= \sum_{i=1}^{n}\left((y_i-\hat{y}_i)^2+2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+(\hat{y}_i-\bar{y})^2\right)
\end{aligned}
$$

10.2 More Detail about Simple Linear Regression

Derivation of SST = SSE + SSM

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}\left((y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right) \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2
\end{aligned}
$$

10.2 More Detail about Simple Linear Regression

Derivation of SST = SSE + SSM

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}\left((y_i-\hat{y}_i)^2+2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+(\hat{y}_i-\bar{y})^2\right) \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2
\end{aligned}
$$

since $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$ as shown on the next slide

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x}))$$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum_{i=1}^{n}(y_i-\bar{y}-b_1(x_i-\bar{x}))(b_1(x_i-\bar{x})) \\ &= \sum_{i=1}^{n}\left(b_1(x_i-\bar{x})(y_i-\bar{y})-b_1^2(x_i-\bar{x})^2\right)\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x})) \\
&= \sum_{i=1}^{n}\left(b_1(x_i - \bar{x})(y_i - \bar{y}) - b_1^2(x_i - \bar{x})^2\right) \\
&= b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x})) \\
&= \sum_{i=1}^{n}\left(b_1(x_i - \bar{x})(y_i - \bar{y}) - b_1^2(x_i - \bar{x})^2\right) \\
&= b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= b_1(n-1)s_x s_y r - b_1^2 s_x^2(n-1)
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x})) \\
&= \sum_{i=1}^{n}\left(b_1(x_i - \bar{x})(y_i - \bar{y}) - b_1^2(x_i - \bar{x})^2\right) \\
&= b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= b_1(n-1)s_x s_y r - b_1^2 s_x^2(n-1) \\
&= \left(r\frac{s_y}{s_x}\right)(n-1)s_x s_y r - \left(r\frac{s_y}{s_x}\right)^2 s_x^2(n-1)
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Demonstration of $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$\hat{y}_i = b_0 + b_1 x_i = (\bar{y} - b_1\bar{x}) + b_1 x_i = \bar{y} + b_1(x_i - \bar{x})$

so

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - \bar{y} - b_1(x_i - \bar{x}))(b_1(x_i - \bar{x})) \\
&= \sum_{i=1}^{n}\left(b_1(x_i - \bar{x})(y_i - \bar{y}) - b_1^2(x_i - \bar{x})^2\right) \\
&= b_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= b_1(n-1)s_x s_y r - b_1^2 s_x^2 (n-1) \\
&= \left(r\frac{s_y}{s_x}\right)(n-1)s_x s_y r - \left(r\frac{s_y}{s_x}\right)^2 s_x^2(n-1) \\
&= (n-1)r^2 s_y^2 - (n-1)r^2 s_y^2 = 0
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

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

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ (n-1)s_y^2 &= (n-2)s^2+\sum_{i=1}^{n}b_1^2(x_i-\bar{x})^2\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
(n-1)s_y^2 &= (n-2)s^2+\sum_{i=1}^{n}b_1^2(x_i-\bar{x})^2 \\
(n-1)s_y^2 &= (n-2)s^2+b_1^2(n-1)s_x^2
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\
(n-1)s_y^2 &= (n-2)s^2 + \sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2 \\
(n-1)s_y^2 &= (n-2)s^2 + b_1^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2 + \left(r\frac{s_y}{s_x}\right)^2(n-1)s_x^2
\end{aligned}
$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
(n-1)s_y^2 &= (n-2)s^2+\sum_{i=1}^{n}b_1^2(x_i-\bar{x})^2 \\
(n-1)s_y^2 &= (n-2)s^2+b_1^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+\left(r\frac{s_y}{s_x}\right)^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+r^2(n-1)s_y^2
\end{aligned}$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
(n-1)s_y^2 &= (n-2)s^2+\sum_{i=1}^{n}b_1^2(x_i-\bar{x})^2 \\
(n-1)s_y^2 &= (n-2)s^2+b_1^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+\left(r\frac{s_y}{s_x}\right)^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+r^2(n-1)s_y^2 \\
(n-1)s_y^2(1-r^2) &= (n-2)s^2
\end{aligned}
$$

10.2 More Detail about Simple Linear Regression

Derivation for s^2

This formula can be used to obtain the simplied version of the formula for s^2.

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\
(n-1)s_y^2 &= (n-2)s^2+\sum_{i=1}^{n}b_1^2(x_i-\bar{x})^2 \\
(n-1)s_y^2 &= (n-2)s^2+b_1^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+\left(r\frac{s_y}{s_x}\right)^2(n-1)s_x^2 \\
(n-1)s_y^2 &= (n-2)s^2+r^2(n-1)s_y^2 \\
(n-1)s_y^2(1-r^2) &= (n-2)s^2 \\
\frac{(n-1)s_y^2(1-r^2)}{n-2} &= s^2
\end{aligned}
$$

10.2 More Detail about Simple Linear Regression

Analysis of Variance (ANOVA) Table

- The calculations for the estimate of the variance are summarized as part of an ANOVA table.

10.2 More Detail about Simple Linear Regression

Analysis of Variance (ANOVA) Table

- The calculations for the estimate of the variance are summarized as part of an ANOVA table.

Source	Degrees of freedom	Sum of Squares	Mean square
Model	1	SSM	$\text{MSM} = \dfrac{\text{SSM}}{1}$
Error	$n-2$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{n-2} = s^2$
Total	$n-1$	SST	$\dfrac{\text{SST}}{n-1} = s_y^2$

10.2 More Detail about Simple Linear Regression

Analysis of Variance (ANOVA) Table

- The calculations for the estimate of the variance are summarized as part of an ANOVA table.

Source	**Degrees of freedom**	**Sum of Squares**	**Mean square**
Model	1	SSM	$\text{MSM} = \frac{\text{SSM}}{1}$
Error	$n-2$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-2} = s^2$
Total	$n-1$	SST	$\frac{\text{SST}}{n-1} = s_y^2$

- An additional column will be added in the next chapter to perform an F test. This is not needed for simple linear regression since it is equivalent to a two-sided test that $\beta_1 = 0$.

10.2 More Detail about Simple Linear Regression

Confidence Intervals and Significance Tests

- For confidence intervals and significance tests, we need to derive formulas for the standard errors of the estimates.

10.2 More Detail about Simple Linear Regression

Confidence Intervals and Significance Tests

- For confidence intervals and significance tests, we need to derive formulas for the standard errors of the estimates.
- In the simple linear regression setting, statisticians are often interested in confidence intervals for:
  1. β_0
  2. β_1
  3. mean response for y when $x = x^*$
  4. predicted response for y when $x = x^*$

10.2 More Detail about Simple Linear Regression

Confidence Intervals and Significance Tests

- For confidence intervals and significance tests, we need to derive formulas for the standard errors of the estimates.
- In the simple linear regression setting, statisticians are often interested in confidence intervals for:
  1. β_0
  2. β_1
  3. mean response for y when $x = x^*$
  4. predicted response for y when $x = x^*$
- For significance testing, the most typical hypothesis of interest is $H_0 : \beta_1 = 0$ which implies that the mean of y does not vary with x.

10.2 More Detail about Simple Linear Regression

Confidence interval for β_0 in the simple linear regression model

A level C confidence interval for β_0 is

$$b_0 \pm t^* SE_{b_0}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

10.2 More Detail about Simple Linear Regression

Confidence interval for β_1 in the simple linear regression model

A level C confidence interval for β_1 is

$$b_1 \pm t^* SE_{b_1}$$

where t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

$$SE_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s}{s_x\sqrt{n-1}}$$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x}$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$
- $\sigma_{\bar{y}}^2 = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma_{y_i}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$
- $\sigma^2_{\bar{y}} = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma^2_{y_i} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
- $b_1 = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})y_i$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$
- $\sigma_{\bar{y}}^2 = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma_{y_i}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
- $b_1 = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})y_i$
- $\mu_{b_1} = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})\mu_{y_i-\bar{y}}$
  $= \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})\beta_1(x_i-\bar{x})$
  $= \frac{1}{(n-1)s_x^2}\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2 = \beta_1$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$
- $\sigma_{\bar{y}}^2 = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma_{y_i}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
- $b_1 = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})y_i$
- $\mu_{b_1} = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\mu_{y_i-\bar{y}}$
  $= \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x})$
  $= \frac{1}{(n-1)s_x^2}\beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 = \beta_1$
- $\sigma_{b_1}^2 = \frac{1}{(n-1)^2 s_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_{y_i}^2 = \frac{1}{(n-1)^2 s_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2$
  $= \frac{1}{(n-1)^2 s_x^4}(n-1)s_x^2\sigma^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{y_i} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\bar{x}$
- $\sigma_{\bar{y}}^2 = \left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sigma_{y_i}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
- $b_1 = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})y_i$
- $\mu_{b_1} = \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\mu_{y_i-\bar{y}}$
  $= \frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x})$
  $= \frac{1}{(n-1)s_x^2}\beta_1\sum_{i=1}^{n}(x_i - \bar{x})^2 = \beta_1$
- $\sigma_{b_1}^2 = \frac{1}{(n-1)^2 s_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_{y_i}^2 = \frac{1}{(n-1)^2 s_x^4}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2$
  $= \frac{1}{(n-1)^2 s_x^4}(n-1)s_x^2\sigma^2 = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
  $\Rightarrow \sigma_{b_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
- So, b_1 follows a Normal$(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}})$ distribution.

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{b_0} = \mu_{\bar{y}} - \bar{x}\mu_{b_1} = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{b_0} = \mu_{\bar{y}} - \bar{x}\mu_{b_1} = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$
- $\bar{y}$ and b_1 are independent

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{b_0} = \mu_{\bar{y}} - \bar{x}\mu_{b_1} = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$
- $\bar{y}$ and b_1 are independent
- $\sigma^2_{b_0} = \sigma^2_{\bar{y}} + \bar{x}^2\sigma_{b_1} = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^n(x_i-\bar{x})^2}$
  $= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}\right)$
  $\Rightarrow \sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n(x_i-\bar{x})^2}}$

10.2 More Detail about Simple Linear Regression

Justification of SE_{b_0} and SE_{b_1}

- $\mu_{b_0} = \mu_{\bar{y}} - \bar{x}\mu_{b_1} = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0$
- $\bar{y}$ and b_1 are independent
- $\sigma^2_{b_0} = \sigma^2_{\bar{y}} + \bar{x}^2\sigma_{b_1} = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
  $= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$
  $\Rightarrow \sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- So, b_0 follows a Normal$\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$ distribution.

10.2 More Detail about Simple Linear Regression

- **Example 10D**: Real estate is typically reassessed annually for property tax purposes. This assessed value is not necessarily the same as the fair market value of the property. A data set includes a SRS of assessment values (x) and prices (y) for 30 homes sold recently in a midwestern city (in thousands of dollars). The summary statistics for the data set are given below:
  $\bar{x} = 184.133, s_x = 45.428, \bar{y} = 195.840, s_y = 47.179, r = 0.91169.$
  (a) Construct a 99% confidence interval for the intercept of the regression line of price on assessment value.
  (b) Construct a 99% confidence interval for the slope of the regression line of price on assessment value.

10.2 More Detail about Simple Linear Regression

- *Answer*:

10.2 More Detail about Simple Linear Regression

Two-sided significance test for the slope in the simple linear regression model

1. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
2. Test statistic: $t = \dfrac{b_1}{SE_{b_1}}$,
   $$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x\sqrt{n-1}}$$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \frac{\alpha}{2}$ where T follows a t distribution with $n-2$ degrees of freedom
4. Reject H_0 is $|t| > t^*$

10.2 More Detail about Simple Linear Regression

One-sided (left-sided) significance test for the slope in the simple linear regression model

1. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 < 0$
2. Test statistic: $t = \dfrac{b_1}{SE_{b_1}}$,

   $$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x \sqrt{n-1}}$$
3. Use the t-table to find the value t^* such that $P(T < -t^*) = P(T > t^*) = \alpha$ where T follows a t distribution with $n - 2$ degrees of freedom
4. Reject H_0 is $t < -t^*$

10.2 More Detail about Simple Linear Regression

One-sided (right-sided) significance test for the slope in the simple linear regression model

1. Test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$
2. Test statistic: $t = \dfrac{b_1}{SE_{b_1}}$,
   $$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s}{s_x\sqrt{n-1}}$$
3. Use the t-table to find the value t^* such that $P(T > t^*) = \alpha$ where T follows a t distribution with $n-2$ degrees of freedom
4. Reject H_0 is $t > t^*$

10.2 More Detail about Simple Linear Regression

- **Example 10E**: Real estate is typically reassessed annually for property tax purposes. This assessed value is not necessarily the same as the fair market value of the property. A data set includes a SRS of assessment values (x) and prices (y) for 30 homes sold recently in a midwestern city (in thousands of dollars). The summary statistics for the data set are given below:
  $\bar{x} = 184.133, s_x = 45.428, \bar{y} = 195.840, s_y = 47.179, r = 0.91169.$

  Let β_1 denote the slope of the regression line of price on assessment value. Test $H_0 : \beta_1 = 0$ versus the alternative $H_a : \beta_1 > 0$ at level .05. Calculate an appropriate test statistic, compute the critical value, and state your conclusion.

10.2 More Detail about Simple Linear Regression

- *Answer*:

10.2 More Detail about Simple Linear Regression

Confidence interval for mean response when $x = x^*$ in the simple linear regression model

A level C confidence interval for the mean response $\mu = \beta_0 + \beta_1 x^*$ when $x = x^*$ is

$$\hat{\mu} \pm t^* SE_{\hat{\mu}}$$

where $\hat{\mu} = b_0 + b_1 x^*$ and t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

The probability that this method produces an interval that contains $\mu = \beta_0 + \beta_1 x^*$ is C.

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

10.2 More Detail about Simple Linear Regression

Prediction interval for a future observation when $x = x^*$ in the simple linear regression model

A level C prediction interval for a future observation from the subpopulation corresponding to $x = x^*$ is

$$\hat{\mu} \pm t^* SE_{y^*}$$

where $\hat{\mu} = b_0 + b_1 x^*$ and t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

The probability that this method produces an interval that contains the future observation is C.

$$SE_{y^*} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

10.2 More Detail about Simple Linear Regression

Justification of $SE_{\hat{\mu}}$

- $\mu_{\hat{\mu}} = \mu_{b_0} + \mu_{b_1} x^* = \beta_0 + \beta_1 x^* = \mu$

10.2 More Detail about Simple Linear Regression

Justification of $SE_{\hat{\mu}}$

- $\mu_{\hat{\mu}} = \mu_{b_0} + \mu_{b_1} x^* = \beta_0 + \beta_1 x^* = \mu$
- $\hat{\mu} = \bar{y} + b_1(x_* - \bar{x})$

10.2 More Detail about Simple Linear Regression

Justification of $SE_{\hat{\mu}}$

- $\mu_{\hat{\mu}} = \mu_{b_0} + \mu_{b_1} x^* = \beta_0 + \beta_1 x^* = \mu$
- $\hat{\mu} = \bar{y} + b_1(x_* - \bar{x})$
- $\sigma^2_{\hat{\mu}} = \sigma^2_{\bar{y}} + (x_* - \bar{x})^2 \sigma^2_{b_1} = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

  $\Rightarrow \sigma_{\hat{\mu}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- So, $\hat{\mu}$ follows a Normal$\left(\mu, \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$ distribution.

10.2 More Detail about Simple Linear Regression

Justification of $SE_{\hat{\mu}}$

- $\mu_{\hat{\mu}} = \mu_{b_0} + \mu_{b_1} x^* = \beta_0 + \beta_1 x^* = \mu$
- $\hat{\mu} = \bar{y} + b_1(x_* - \bar{x})$
- $\sigma^2_{\hat{\mu}} = \sigma^2_{\bar{y}} + (x_* - \bar{x})^2 \sigma^2_{b_1} = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
  $\Rightarrow \sigma_{\hat{\mu}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- So, $\hat{\mu}$ follows a Normal$\left(\mu, \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$ distribution.
- $\sigma^2_{y^*} = \sigma^2_{\bar{y}} + (x_* - \bar{x})^2 \sigma^2_{b_1} + \sigma^2_{\varepsilon}$
  $= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2$
  $\Rightarrow \sigma_{y^*} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

10.2 More Detail about Simple Linear Regression

Justification of $SE_{\hat{\mu}}$

- $\mu_{\hat{\mu}} = \mu_{b_0} + \mu_{b_1} x^* = \beta_0 + \beta_1 x^* = \mu$
- $\hat{\mu} = \bar{y} + b_1(x_* - \bar{x})$
- $\sigma^2_{\hat{\mu}} = \sigma^2_{\bar{y}} + (x_* - \bar{x})^2 \sigma^2_{b_1} = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

  $\Rightarrow \sigma_{\hat{\mu}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- So, $\hat{\mu}$ follows a Normal$\left(\mu, \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$ distribution.
- $\sigma^2_{y^*} = \sigma^2_{\bar{y}} + (x_* - \bar{x})^2 \sigma^2_{b_1} + \sigma^2_{\varepsilon}$

  $= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \sigma^2$

  $\Rightarrow \sigma_{y^*} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- So, y^* follows a Normal$\left(\mu, \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$ distribution.

10.2 More Detail about Simple Linear Regression

- **Example 10F**: Real estate is typically reassessed annually for property tax purposes. This assessed value is not necessarily the same as the fair market value of the property. A data set includes a SRS of assessment values (x) and prices (y) for 30 homes sold recently in a midwestern city (in thousands of dollars). The summary statistics for the data set are given below:
  $\bar{x} = 184.133, s_x = 45.428, \bar{y} = 195.840, s_y = 47.179, r = 0.91169.$
  (a) Construct a 99% confidence interval for the mean sale price for homes that are assessed at \$180,000.
  (b) Construct a 99% prediction interval for the sale price of a new observation that is assessed at \$180,000.

10.2 More Detail about Simple Linear Regression

- *Answer*: