Chapter 14 Inference on the Least-Squares Regression Model & Multiple Regression

14.1A Using Randomization Techniques on the Slope of the Least-Squares Regression Line

Objective

1. Use randomization to test the significance of the slope of the least-squares regression model

Review of Least-Squares Regression

Example

The following data represent a random sample of three bedroom homes that recently sold in Seattle, Washington. The Zestimate is the predicted selling price of the home. Treat the Zestimate as the explanatory variable and selling price as the response variable.

Zestimate	Sale Price	Zestimate	Sale Price
371,485	355,000	361,111	350,000
459,767	455,000	300,753	280,000
398,718	420,000	319,548	325,000
419,374	400,000	343,201	330,000
554,154	533,230	568,761	594,450
563,803	620,000	426,609	425,000

Source: Zillow.com

(a) Draw a scatter diagram of the data. What type of relation appears to exist between Zestimate and selling price?

(b) Does a linear relation exist between the Zestimate and selling price?

(c) Find the least-squares regression line.

(d) Interpret the slope. Explain why it does not make sense to interpret the *y*-intercept.

(e) Predict the selling price of a home in Seattle whose Zestimate is $410,000.

(f) Can this model be used to predict selling prices of homes in Chicago? Why or why not?

In the least-squares regression equation $\hat{y} = b_1 x + b_0$, the values for the slope, b_1, and intercept, b_0, are statistics. The statistics b_o and b_1 are estimates for the population intercept, β_0, and population slope, β_1. The true linear relation between the explanatory variable, *x*, and the response variable, *y*, is given by

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

So, from the Zillow data we have b_0 = -55137.083 and b_1 = 1.1301. We want to know if selling price is positively associated with the Zestimate. There are two possibilities.

1. There is no linear association between the Zestimate and selling price.
2. There is a positive association between the Zestimate and selling price.

We can state the two possibilities presented above using the language of hypothesis testing. Statement (1) is the statement of "no change" or "no effect" and represents the null hypothesis. Statement (2) is the statement we are trying to demonstrate.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

Use the statement in the null model.

Activity

For the Zillow data.

(a) Randomly assign a selling price to each Zestimate. Draw a scatter diagram of the randomized data. Find the slope of the least-squares regression line for the randomized data.

(b) In StatCrunch, select Applets > Resampling > Randomization test for slope. Click "1000 times" five times. What is the likelihood of obtaining a slope of 1.1301 or higher under the assumption there is no linear relation between the Zestimate and selling price.

It is worth noting that the shape of the distribution in the null model is bell-shaped and is centered at 0. The fact that the center is near 0 should not be surprising because the statement in the null hypothesis is that the slope of the line describing the relation between Sale Price and Zestimate is 0 (and the random assignment behaves as if there is no association between the two variables). Therefore, over the long-term we would expect half the random assignments to result in a slope less than 0, and half to result in a slope greater than 0. There is natural variability in the slopes as a result of random assignment.

Testing Hypotheses Regarding the Slope of the Least-Squares Regression Using Randomization

Step 1 Verify that the explanatory variable and the response variable in the study are quantitative.

Step 2 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways.

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$

Step 3 Build a null model that randomly assigns the response variable in the study to the explanatory variable under the assumption the statement in the null hypothesis is true.

Step 4 Estimate the P-value from the model in Step 3.

Step 5 State the conclusion.

14.1 Testing the Significance of the Least-Squares Regression Model

Objectives

1. State the requirements of the least-squares regression model
2. Compute the standard error of the estimate
3. Verify that the residuals are normally distributed
4. Conduct inference on the slope of the least-squares regression model
5. Construct a confidence interval about the slope of the least-squares regression model

❶ State the Requirements of the Least-Squares Regression Model

In the least-squares regression equation $\hat{y} = b_1 x + b_0$, the values for the slope, b_1, and intercept, b_0, are statistics. Therefore, they have sampling distributions associated with them.

To find the sampling distribution of the slope, b_1, and intercept, b_0, certain requirements must be satisfied.

Requirement 1 for Inference on the Least-Squares Regression Model

For any particular value of the explanatory variable x (such as 32 in Example 1), the mean of the corresponding responses in the population depends linearly on x. That is,

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

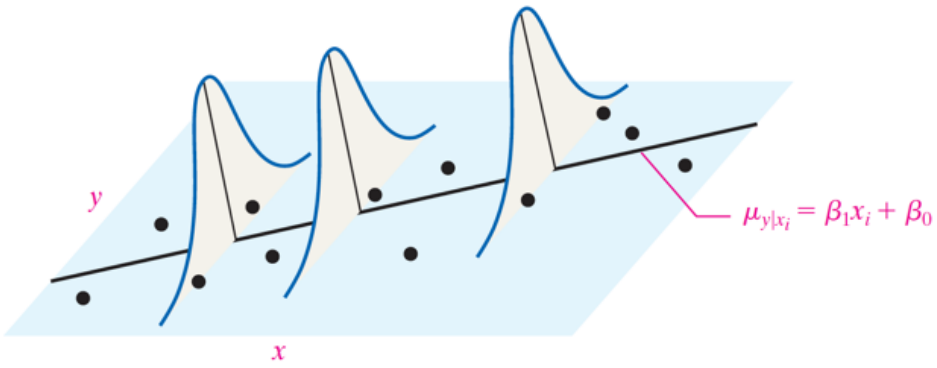
for some numbers β_0 and β_1, where $\mu_{y|x}$ represents the population mean response when the value of the explanatory variable is x.

Requirement 2 for Inference on the Least-Squares Regression Model

The response variable is normally distributed with mean $\mu_{y|x} = \beta_1 x + \beta_0$ and standard deviation σ.

"In Other Words"

When doing inference on the least-squares regression model, we require (1) for any explanatory variable, x, the mean of the response variable, y, depends on the value of x through a linear equation, and (2) the response variable, y, is normally distributed with a constant standard deviation, σ. The mean increases/ decreases at a constant rate depending on the slope, while the standard deviation remains constant.



The **least-squares regression model** is given by

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \quad \textbf{(1)}$$

where

y_i is the value of the response variable for the ith individual
x_i is the value of the explanatory variable for the ith individual
β_0 and β_1 are the parameters to be estimated based on sample data
ε_i is a random error term with mean 0 and standard deviation $\sigma_{\varepsilon_i} = \sigma$, the error terms are independent
$i = 1, \ldots, n$, where n is the sample size (number of ordered pairs in the data set)

❷ Compute the Standard Error of the Estimate

The **standard error of the estimate**, s_e, is found using the formula

$$s_e = \sqrt{\frac{\Sigma (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\Sigma \text{ residuals}^2}{n - 2}} \quad \textbf{(2)}$$

Example Zillow Data

Find the standard error of the estimate for the Zillow data.

❸ Verify That the Residuals Are Normally Distributed

The least-squares regression model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ requires the response variable, y_i, to be normally distributed. Because $\beta_1 x_i + \beta_0$ is constant for any x_i, if y_i is normal, then the residuals, ε_i, must be normal. To perform statistical inference on the regression line, we verify that the residuals are normally distributed by examining a normal probability plot.

Example Verify Residuals Are Normally Distributed

Verify the residuals for the Zillow data are normally distributed.

❹ Conduct Inference on the Slope of the Least-Squares Regression Model

We want to know if the sample data suggest a linear relation exists between the explanatory and response variables. To do this, we conduct one of three hypothesis tests.

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$

Activity Exploring the Sampling Distribution of the Slope

Open the data set "HomeRuns2017" in the SullyStats group within StatCrunch. This data set represents all home runs hit in the 2017 baseball season. Therefore, it is population data. The variable "Distance" represents the distance the ball traveled (in feet). The variable "Speed Off Bat" represents the speed the ball left the bat when hit (in miles per hour).

(a) Draw a scatter diagram between Distance and Speed Off Bat, treating Speed Off Bat as the explanatory variable.

(b) Find the least-squares regression line treating Speed Off Bat as the explanatory variable. What are the values of β_0 and β_1 ?

(c) Using StatCrunch, select Data > Sample. Select the columns Distance and Speed Off Bat. Let the sample size equal 12 and draw 1000 samples. Check the box sample all columns at one time. Check the box open in a new data table. Click Compute! Now find the least-squares regression for each sample and save the estimates in the data table.

(d) Draw a histogram of the values of b_1. Find the mean and standard deviation of b_1. Compare the mean of b_1 to the value of β_1 from part (b).

(e) The formula for the standard deviation of b_1 is $s_{b_1} = \dfrac{s_e}{\sqrt{\Sigma(x_i - \bar{x})^2}}$. Compute this value for the first sample.

Hypothesis Test Regarding the Slope Coefficient, β_1

To test whether two quantitative variables are linearly related, use the following steps provided that

- The sample is obtained using random sampling or from a randomized experiment.
- The residuals are normally distributed with constant error variance.

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$

Step 2 Select a level of significance, α, depending on the seriousness of making a Type I error.

Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

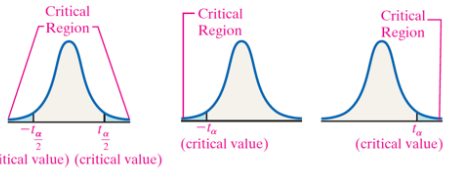
P-Value Approach

By-Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

which follows Student's t-distribution with $n - 2$ degrees of freedom. Use Table VII to determine the critical value.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$	$-t_\alpha$	t_α

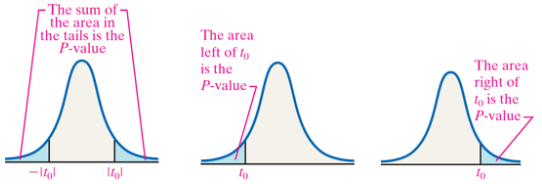


Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

which follows Student's t-distribution with $n - 2$ degrees of freedom. Use Table VII to approximate the P-value.

Two-Tailed	Left-Tailed	Right-Tailed



Two-Tailed	Left-Tailed	Right-Tailed
$P\text{-value} = 2P(t > \lvert t_0 \rvert)$	$P\text{-value} = P(t < t_0)$	$P\text{-value} = P(t > t_0)$

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P-value. The directions for obtaining the P-value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on pages 683–684.

Step 4 If P-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Testing for a Linear Relation

Test whether a linear relation exists between Zestimate and Sale Price using the Zillow data.

❺ Construct a Confidence Interval about the Slope of the Least-Squares Regression Model

Confidence Intervals for the Slope of the Regression Line

A $(1 - \alpha) \cdot 100\%$ confidence interval for the slope of the true regression line, β_1, is given by the following formulas:

$$\text{Lower bound: } b_1 - t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad \text{Upper bound: } b_1 + t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{3}$$

Here, $t_{\alpha/2}$ is computed with $n - 2$ degrees of freedom.

Note: This interval can be computed only if the data are randomly obtained, the residuals are normally distributed, and there is constant error variance.

Example **Constructing a Confidence Interval for the Slope of the True Regression Line**

Determine a 95% confidence interval for the slope of the true regression line for the Zillow data.

14.2 Confidence and Prediction Intervals

Objectives

1. Construct confidence intervals for a mean response
2. Construct prediction intervals for an individual response

If we use the least-squares regression model from the Zillow data to make a prediction where the Zestimate is $410,000, it has two interpretations.

1. It would be the mean selling price of all homes in Seattle whose Zestimate is $410,000.
2. It would be the predicted selling price of a particular home in Seattle whose Zestimate is $410,000.

Confidence intervals for a mean response are intervals constructed about the predicted value of y, at a given level of x, that are used to measure the accuracy of the mean response of all the individuals in the population.

Prediction intervals for an individual response are intervals constructed about the predicted value of y that are used to measure the accuracy of a single individual's predicted value.

Confidence Interval for the Mean Response of y, $\hat{y}$

A $(1 - \alpha) \cdot 100\%$ confidence interval for $\hat{y}$, the mean response of y for a specified value of x, is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}}$$

$$\text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} \qquad \textbf{(1)}$$

where x^* is the given value of the explanatory variable, n is the number of observations, and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

Confidence Interval for the Mean Response of y, $\hat{y}$

A $(1 - \alpha) \cdot 100\%$ confidence interval for $\hat{y}$, the mean response of y for a specified value of x, is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}}$$

$$\text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}} \qquad \textbf{(1)}$$

where x^* is the given value of the explanatory variable, n is the number of observations, and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

Example Confidence and Prediction Intervals

Construct a 95% confidence interval for the mean selling price of all homes in Seattle whose Zestimate is $410,000. Construct a 95% prediction interval for the selling price of a particular home in Seattle whose Zestimate is $410,000.