Chapter 12: More About Regression

Section 12.1
Inference for Linear Regression

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Introduction

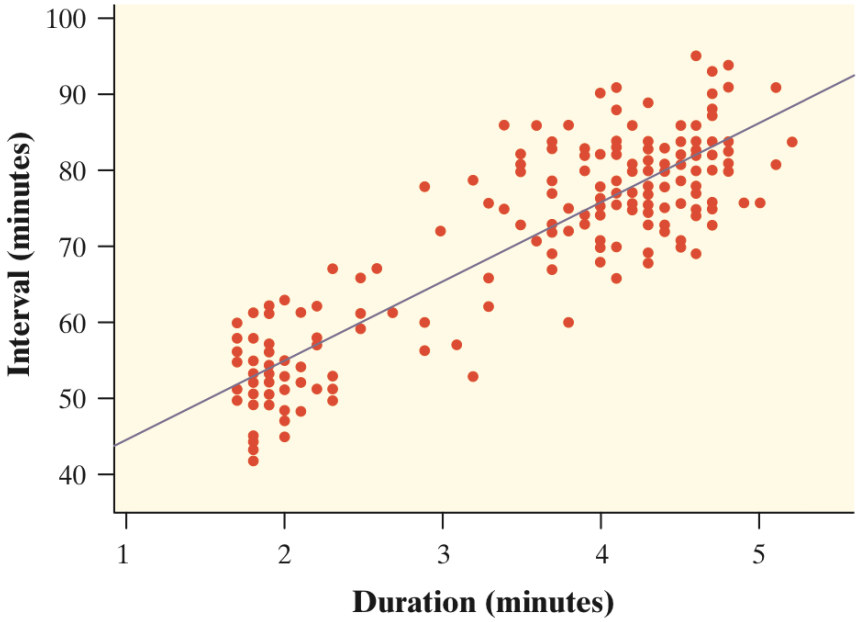
When a scatterplot shows a linear relationship between a quantitative explanatory variable *x* and a quantitative response variable *y*, we can use the least-squares line fitted to the data to predict *y* for a given value of *x*. If the data are a random sample from a larger population, we need statistical inference to answer questions like these:

- Is there really a linear relationship between *x* and *y* in the population, or could the pattern we see in the scatterplot plausibly happen just by chance?
- In the population, how much will the predicted value of *y* change for each increase of 1 unit in *x*? What's the margin of error for this estimate?

In Section 12.1, we will learn how to estimate and test claims about the slope of the population (true) regression line that describes the relationship between two quantitative variables.

Inference for Linear Regression

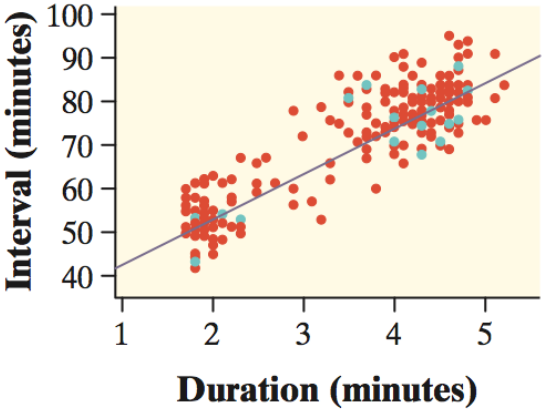
In Chapter 3, we examined data on eruptions of the Old Faithful geyser. Below is a scatterplot of the duration and interval of time until the next eruption for all 222 recorded eruptions in a single month. The least-squares regression line for this population of data has been added to the graph. It has slope 10.36 and *y*-intercept 33.97. We call this the **population regression line** (or true regression line) because it uses all the observations that month.



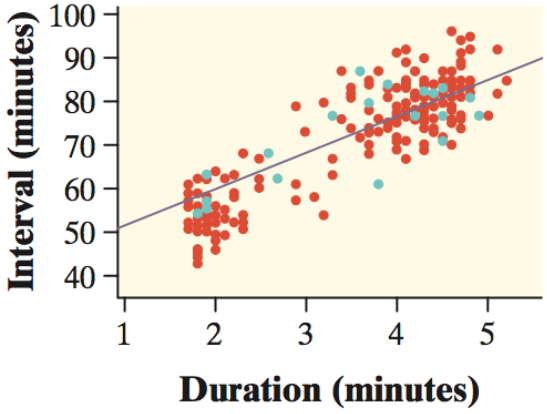
Suppose we take an SRS of 20 eruptions from the population and calculate the least-squares regression line $\hat{y} = a + bx$ for the sample data. How does the slope of the **sample regression line** (also called the estimated regression line) relate to the slop of the population regression line?

Sampling Distribution of *b*

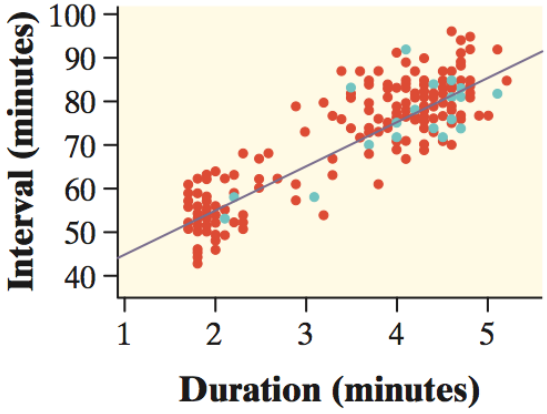
The figures below show the results of taking three different SRSs of 20 Old Faithful eruptions in this month. Each graph displays the selected points and the LSRL for that sample.



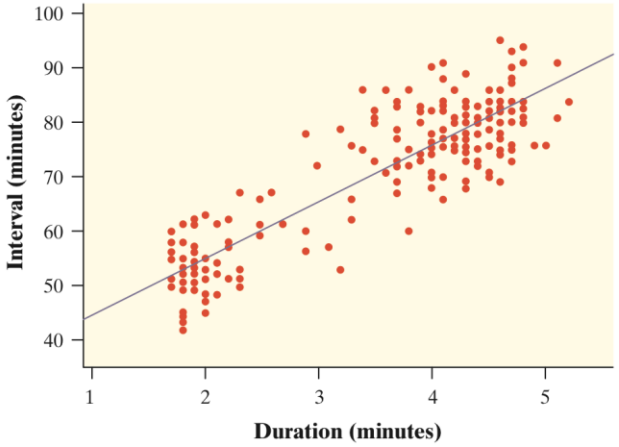
Sample 1: $\hat{y} = 32.8 + 10.2x$



Sample 2: $\hat{y} = 44.0 + 7.7x$



Sample 3: $\hat{y} = 36.0 + 9.5x$

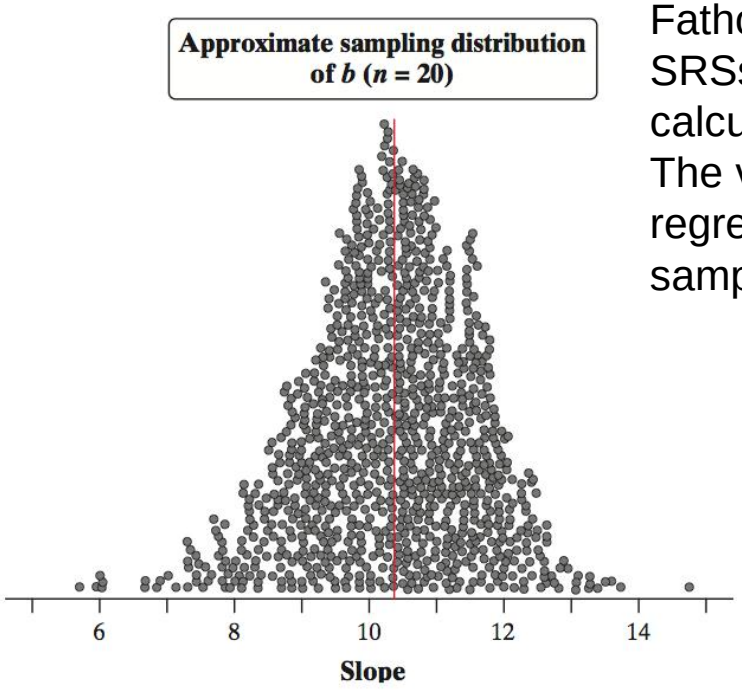


Notice that the slopes of the sample regression lines – 10.2, 7.7, and 9.5 – vary quite a bit from the slope of the population regression line, 10.36.

The pattern of variation in the slope *b* is described by its sampling distribution.

Sampling Distribution of *b*

Confidence intervals and significance tests about the slope of the population regression line are based on the sampling distribution of *b*, the slope of the sample regression line.



Fathom software was used to simulate choosing 1000 SRSs of $n = 20$ from the Old Faithful data, each time calculating the equation of the LSRL for the sample. The values of the slope *b* for the 1000 sample regression lines are plotted. Describe this approximate sampling distribution of *b*.

Shape: We can see that the distribution of *b*-values is roughly symmetric and unimodal. A Normal probability plot of these sample regression line slopes suggests that the approximate sampling distribution of *b* is close to Normal.

Center: The mean of the 1000 *b*-values is 10.32. This value is quite close to the slope of the population (true) regression line, 10.36.

Spread: The standard deviation of the 1000 *b*-values is 1.31. Later, we will see that the standard deviation of the sampling distribution of *b* is actually 1.30.

Condition for Regression Inference

The slope *b* and intercept *a* of the least-squares line are *statistics*. That is, we calculate them from the sample data. These statistics would take somewhat different values if we repeated the data production process. To do inference, think of *a* and *b* as estimates of unknown parameters α and β that describe the population of interest.

Conditions for Regression Inference

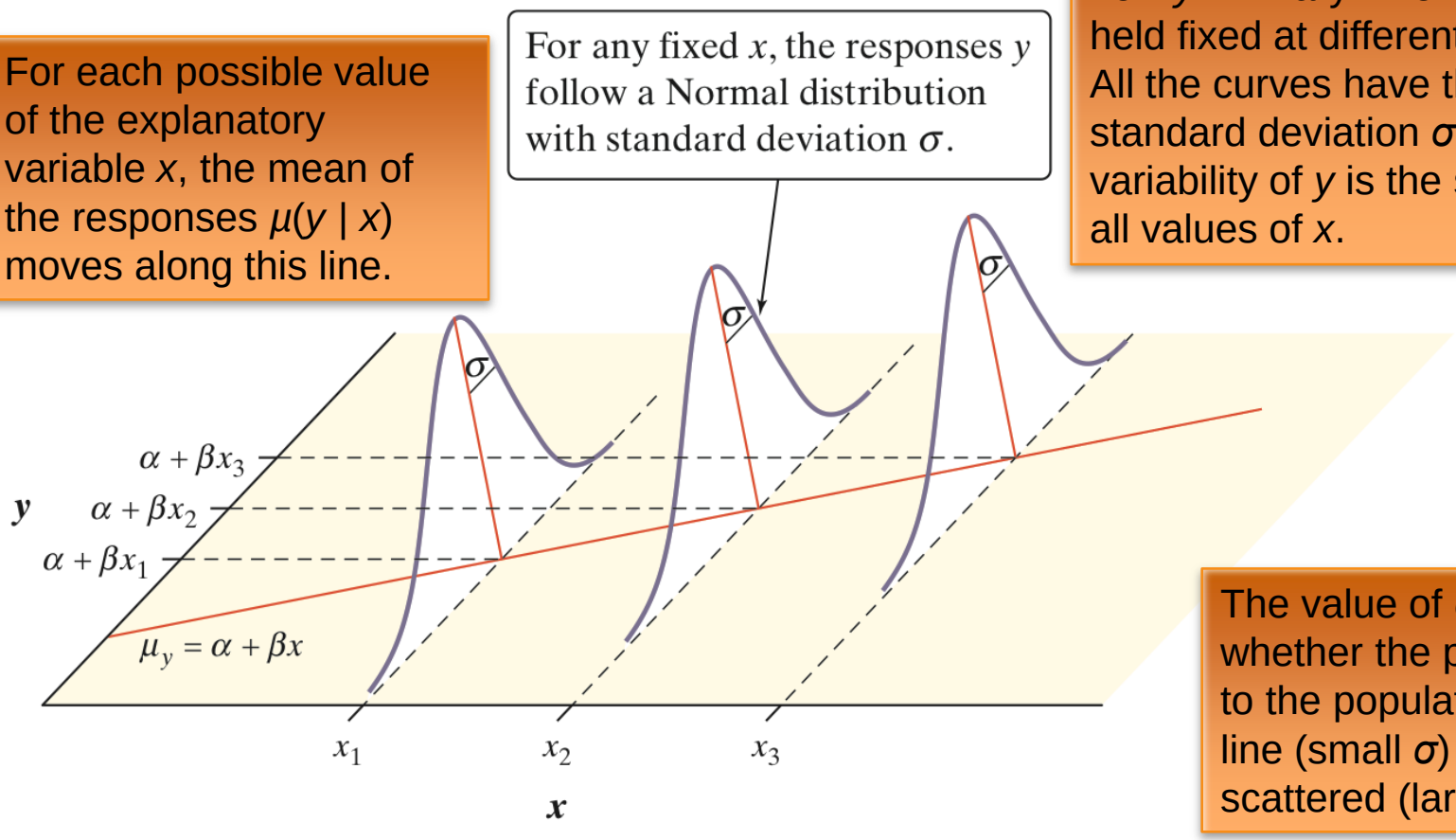
Suppose we have *n* observations on an explanatory variable *x* and a response variable *y*. Our goal is to study or predict the behavior of *y* for given values of *x*.

- **Linear** The (true) relationship between *x* and *y* is linear. For any fixed value of *x*, the mean response μ_y falls on the population (true) regression line $\mu_y = \alpha + \beta x$. The slope *b* and intercept *a* are usually unknown parameters.
- **Independent** Individual observations are independent of each other.
- **Normal** For any fixed value of *x*, the response *y* varies according to a Normal distribution.
- **Equal variance** The standard deviation of *y* (call it σ) is the same for all values of *x*. The common standard deviation σ is usually an unknown parameter.
- **Random** The data come from a well-designed random sample or randomized experiment.

Condition for Regression Inference

The figure below shows the regression model when the conditions are met. The line in the figure is the population regression line $\mu_y = \alpha + \beta x$.

For each possible value of the explanatory variable x, the mean of the responses $\mu(y \mid x)$ moves along this line.



The Normal curves show how y will vary when x is held fixed at different values. All the curves have the same standard deviation σ, so the variability of y is the same for all values of x.

The value of σ determines whether the points fall close to the population regression line (small σ) or are widely scattered (large σ).

How to Check the Conditions for Inference

You should always check the conditions before doing inference about the regression model. Although the conditions for regression inference are a bit complicated, it is not hard to check for major violations.

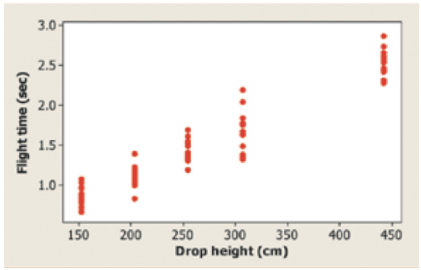
Start by making a histogram or Normal probability plot of the residuals and also a residual plot. Here's a summary of how to check the conditions one by one.

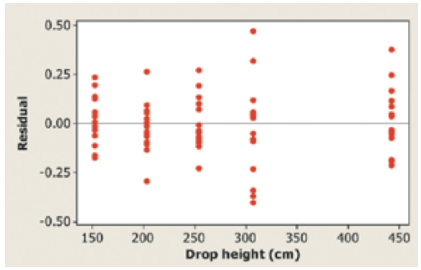
How to Check the Conditions for Regression Inference

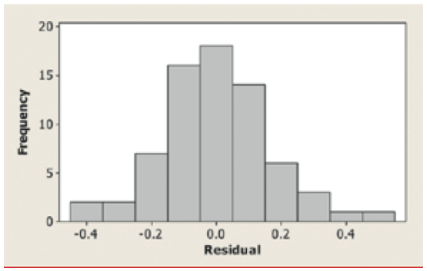
- **L** **Linear** Examine the scatterplot to check that the overall pattern is roughly linear. Look for curved patterns in the residual plot. Check to see that the residuals center on the "residual = 0" line at each *x*-value in the residual plot.
- **I** **Independent** Look at how the data were produced. Random sampling and random assignment help ensure the independence of individual observations. If sampling is done without replacement, remember to check that the population is at least 10 times as large as the sample (10% condition).
- **N** **Normal** Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality.
- **E** **Equal variance** Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The amount of scatter should be roughly the same from the smallest to the largest *x*-value.
- **R** **Random** See if the data were produced by random sampling or a randomized experiment.

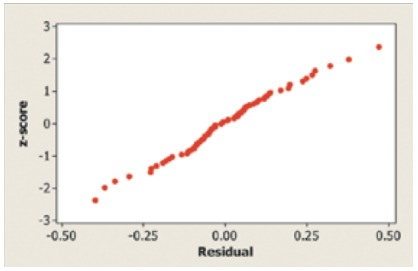
Example: The Helicopter Experiment

Mrs. Barrett's class did a variation of the helicopter experiment on page 738. Students randomly assigned 14 helicopters to each of five drop heights: 152 centimeters (cm), 203 cm, 254 cm, 307 cm, and 442 cm. Teams of students released the 70 helicopters in a predetermined random order and measured the flight times in seconds. The class used Minitab to carry out a least-squares regression analysis for these data. A scatterplot, residual plot, histogram, and Normal probability plot of the residuals are shown below.









✓ **Linear** The scatterplot shows a clear linear form. For each drop height used in the experiment, the residuals are centered on the horizontal line at 0. The residual plot shows a random scatter about the horizontal line.

✓ **Equal variance** The residual plot shows a similar amount of scatter about the residual = 0 line for the 152, 203, 254, and 442 cm drop heights. Flight times (and the corresponding residuals) seem to vary more for the helicopters that were dropped from a height of 307 cm.

✓ **Normal** The histogram of the residuals is single-peaked, unimodal, and somewhat bell-shaped. In addition, the Normal probability plot is very close to linear.

✓ **Independent** Because the helicopters were released in a random order and no helicopter was used twice, knowing the result of one observation should give no additional information about another observation.

✓ **Random** The helicopters were randomly assigned to the five possible drop heights.

Except for a slight concern about the equal-variance condition, we should be safe performing inference about the regression model in this setting.

Estimating the Parameters

When the conditions are met, we can do inference about the regression model $\mu_y = \alpha + \beta x$. The first step is to estimate the unknown parameters.

- ✓If we calculate the least-squares regression line, the slope b is an unbiased estimator of the population slope β, and the y-intercept a is an unbiased estimator of the population y-intercept α.
- ✓The remaining parameter is the standard deviation σ, which describes the variability of the response y about the population regression line.

The LSRL computed from the sample data estimates the population regression line. So the residuals estimate how much y varies about the population line.

Because σ is the standard deviation of responses about the population regression line, we estimate it by the standard deviation of the residuals

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Example: The Helicopter Experiment

Computer output from the least-squares regression analysis on the helicopter data for Mrs. Barrett's class is shown below.

Regression Analysis: Flight time (sec) versus Drop height (cm)

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%

The least-squares regression line for these data i

$$\widehat{flight\ time} = -0.03761 + 0.0057244(drop\ height)$$

The slope β of the true regression line says how much the average flight time of the paper helicopters increases when the drop height increases by 1 centimeter.

Because $b = 0.0057244$ estimates the unknown β, we estimate that, on average, flight time increases by about 0.0057244 seconds for each additional centimeter of drop height.

Example: The Helicopter Experiment

Computer output from the least-squares regression analysis on the helicopter data for Mrs. Barrett's class is shown below.

Regression Analysis: Flight time (sec) versus Drop height (cm)

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%

The least-squares regression line for these data i

$$\widehat{flight\ time} = -0.03761 + 0.0057244(drop\ height)$$

We need the intercept a = -0.03761 to draw the line and make predictions, but it has no statistical meaning in this example. No helicopter was dropped from less than 150 cm, so we have no data near $x = 0$.

We might expect the actual y-intercept α of the true regression line to be 0 because it should take no time for a helicopter to fall no distance.

The y-intercept of the sample regression line is -0.03761, which is pretty close to 0.

Example: The Helicopter Experiment

Computer output from the least-squares regression analysis on the helicopter data for Mrs. Barrett's class is shown below.

Regression Analysis: Flight time (sec) versus Drop height (cm)

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%

The least-squares regression line for these data i

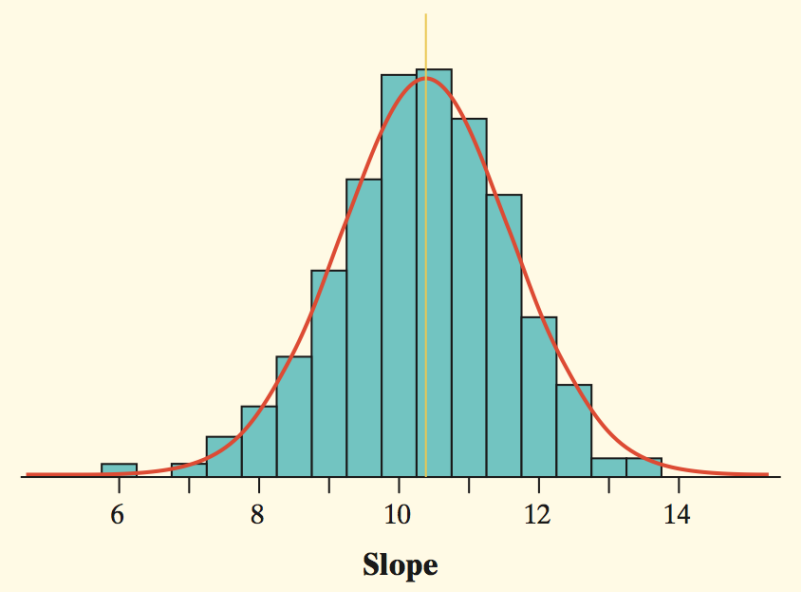
$$\widehat{flight\ time} = -0.03761 + 0.0057244(drop\ height)$$

Our estimate for the standard deviation σ of flight times about the true regression line at each x-value is $s = 0.168$ seconds.

This is also the size of a typical prediction error if we use the least-squares regression line to predict the flight time of a helicopter from its drop height.

The Sampling Distribution of *b*

Let's return to our earlier exploration of Old Faithful eruptions. For all 222 eruptions in a single month, the population regression line for predicting the interval of time until the next eruption *y* from the duration of the previous eruption *x* is $\mu_y = 33.97 + 10.36x$. The standard deviation of responses about this line is given by $\sigma = 6.159$.



If we take all possible SRSs of 20 eruptions from the population, we get the actual sampling distribution of *b*.

Shape: Normal

Center : $\mu_b = \beta = 10.36$ (*b* is an unbiased estimator of β)

Spread: $\sigma_b = \dfrac{\sigma}{s_x\sqrt{n-1}} = \dfrac{6.159}{1.083\sqrt{20-1}} = 1.30$

In practice, we don't know σ for the population regression line. So we estimate it with the standard deviation of the residuals, *s*. Then we estimate the spread of the sampling distribution of *b* with the **standard error of the slope**:

$$SE_b = \frac{s}{s_x\sqrt{n-1}}$$

The Sampling Distribution of *b*

What happens if we transform the values of *b* by standardizing? Since the sampling distribution of *b* is Normal, the statistic

$$z = \frac{b - \beta}{\sigma_b}$$

has the standard Normal distribution.

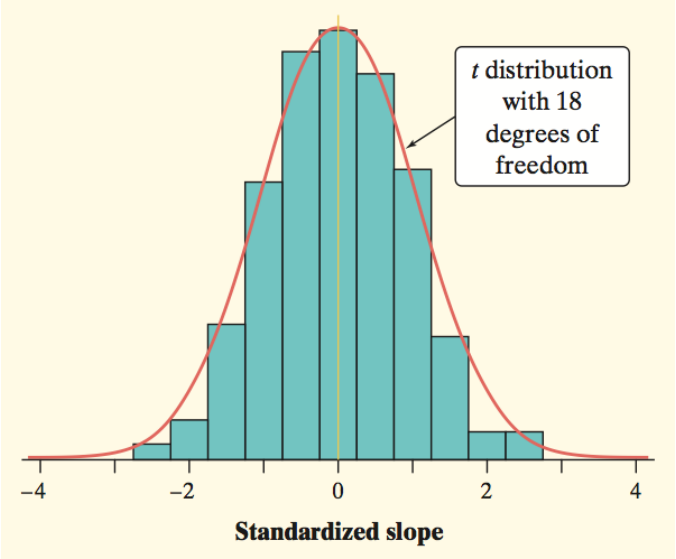
Replacing the standard deviation σ_b of the sampling distribution with its standard error gives the statistic

$$t = \frac{b - \beta}{SE_b}$$

which has a *t* distribution with *n* - 2 degrees of freedom.

The figure shows the result of standardizing the values in the sampling distribution of *b* from the Old Faithful example. Recall, *n* = 20 for this example.

The superimposed curve is a *t* distribution with df = 20 – 2 = 18.



Constructing a Confidence Interval for the Slope

The slope β of the population (true) regression line $\mu_y = \alpha + \beta x$ is the rate of change of the mean response as the explanatory variable increases. We often want to estimate β. The slope b of the sample regression line is our point estimate for β. A confidence interval is more useful than the point estimate because it shows how precise the estimate b is likely to be. The confidence interval for β has the familiar form

statistic ± (critical value) · (standard deviation of statistic)

Because we use the statistic b as our estimate, the confidence interval is

$$b \pm t^* SE_b$$

We call this a ***t* interval for the slope**.

t Interval for the Slope of a Least-Squares Regression Line

When the conditions for regression inference are met, a level C confidence interval for the slope $\beta$of the population (true) regression line is

$$b \pm t^* SE_b$$

In this formula, the standard error of the slope is

$$SE_b = \frac{s}{s_x\sqrt{n-1}}$$

and t^* is the critical value for the t distribution with df = n - 2 having area C between -t^* and t^*.

Example: Helicopter Experiment

Earlier, we used Minitab to perform a least-squares regression analysis on the helicopter data for Mrs. Barrett's class. Recall that the data came from dropping 70 paper helicopters from various heights and measuring the flight times. We checked conditions for performing inference earlier. Construct and interpret a 95% confidence interval for the slope of the population regression line.

Regression Analysis: Flight time (sec) versus Drop height (cm)

Predictor	Coef	SE Coef	T	P
Constant	−0.03761	0.05838	−0.64	0.522
Drop height (cm)	0.0057244	0.0002018	28.37	0.000

S = 0.168181 R-Sq = 92.2% R-Sq(adj) = 92.1%

SE_b = 0.0002018, from the "SE Coef " column in the computer output.

Because the conditions are met, we can calculate a *t* interval for the slope *β* based on a *t* distribution with df = *n* - 2 = 70 - 2 = 68. Using the more conservative df = 60 from Table B gives $t^* = 2.000$.

The 95% confidence interval is

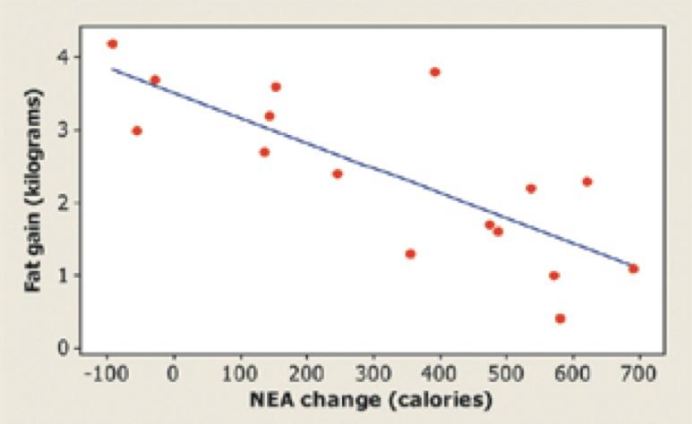
$$b \pm t^* SE_b = 0.0057244 \pm 2.000(0.0002018)$$
$$= 0.0057244 \pm 0.0004036$$
$$= (0.0053208, 0.0061280)$$

We are 95% confident that the interval from 0.0053208 to 0.0061280 seconds per cm captures the slope of the true regression line relating the flight time *y* and drop height *x* of paper helicopters.

Example: Does Fidgeting Keep you Slim?

In Chapter 3, we examined data from a study that investigated why some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explains why. Researchers deliberately overfed a random sample of 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) and change in energy use (in calories) from activity other than deliberate exercise for each subject. Here are the data:

NEA change (cal):	−94	−57	−29	135	143	151	245	355
Fat gain (kg):	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (cal):	392	473	486	535	571	580	620	690
Fat gain (kg):	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1



Regression Analysis: Fat gain versus NEA change

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
NEA change	−0.0034415	0.0007414	−4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq(adj) = 57.8%

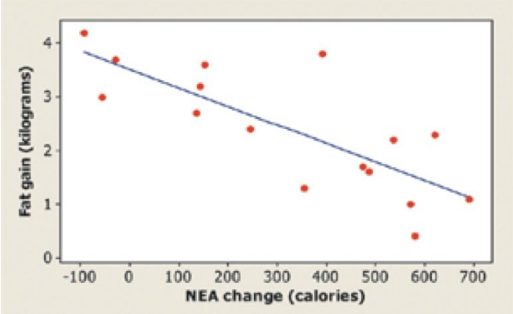
Construct and interpret a 90% confidence interval for the slope of the population regression line.

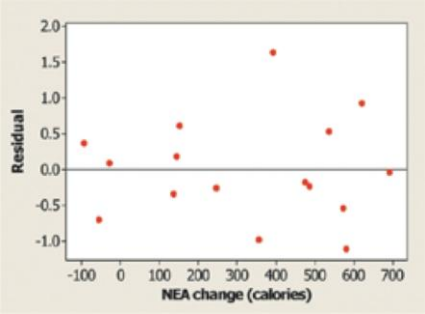
Example: Does Fidgeting Keep you Slim?

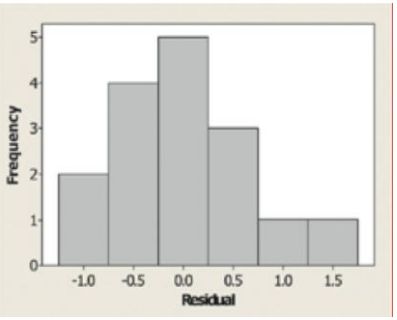
State: We want to estimate the true slope β of the population regression line relating NEA change to fat gain at the 90% confidence level.

Plan: If the conditions are met, we will use a *t* interval for the slope to estimate β.

- **Linear** The scatterplot shows a clear linear pattern. Also, the residual plot shows a random scatter of points about the "residual = 0" line.







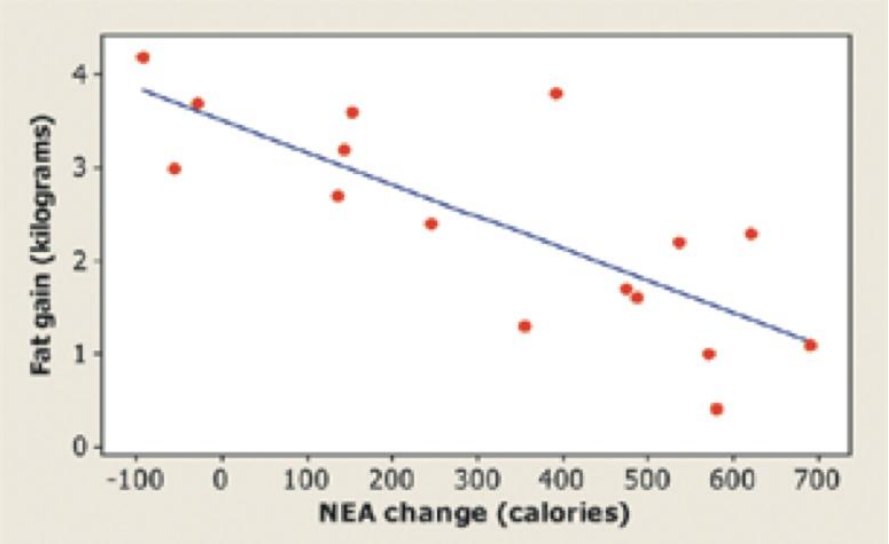
- **Independent** Individual observations of fat gain should be independent if the study is carried out properly. Because researchers sampled without replacement, there have to be at least 10(16) = 160 healthy young adults in the population of interest.
- **Normal** The histogram of the residuals is roughly symmetric and single-peaked, so there are no obvious departures from normality.
- **Equal variance** It is hard to tell from so few points whether the scatter of points around the residual = 0 line is about the same at all *x*-values.
- **Random** The subjects in this study were randomly selected to participate.

Example: Does Fidgeting Keep you Slim?

Do: We use the *t* distribution with 16 - 2 = 14 degrees of freedom to find the critical value. For a 90% confidence level, the critical value is $t^* = 1.761$. So the 90% confidence interval for β is

$$\begin{aligned} b \pm t^* SE_b &= -0.0034415 \pm 1.761(0.0007414) \\ &= -0.0034415 \pm 0.0013056 \\ &= (-0.004747, -0.002136) \end{aligned}$$

Conclude: We are 90% confident that the interval from -0.004747 to -0.002136 kg captures the actual slope of the population regression line relating NEA change to fat gain for healthy young adults.



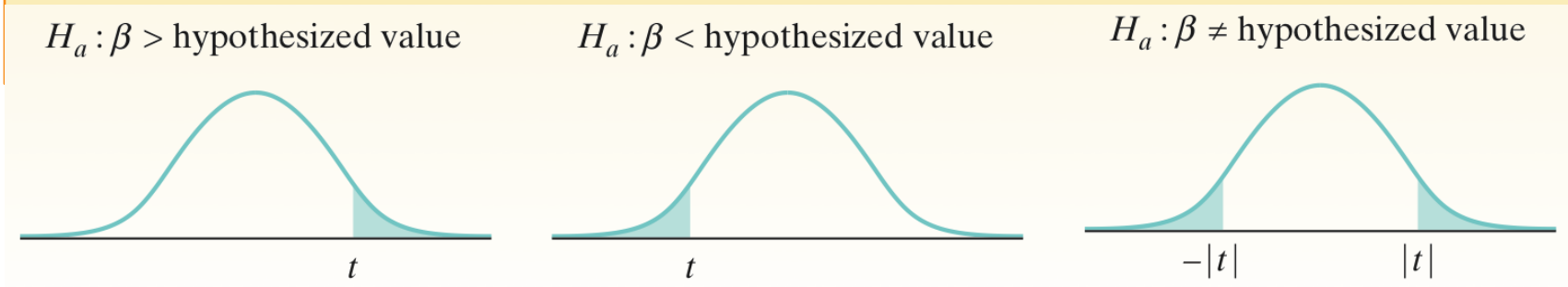
Performing a Significance Test for the Slope

t Test for the Slope of a Least-Squares Regression Line

Suppose the conditions for inference are met. To test the hypothesis $H_0 : \beta =$ hypothesized value, compute the test statistic

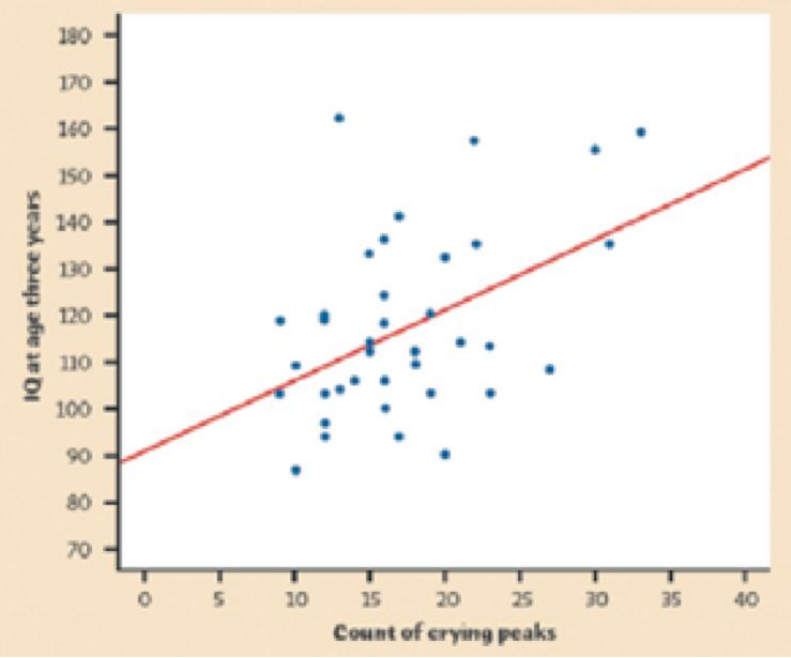
$$t = \frac{b - \beta_0}{SE_b}$$

Find the *P*-value by calculating the probability of getting a *t* statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the *t* distribution with df = *n* - 2.



Example: Crying and IQ

Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test. A scatterplot and Minitab output for the data from a random sample of 38 infants is below.



Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants?

Example: Crying and IQ

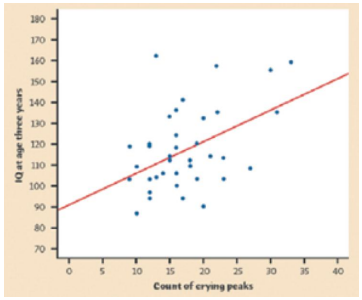
State: We want to perform a test of

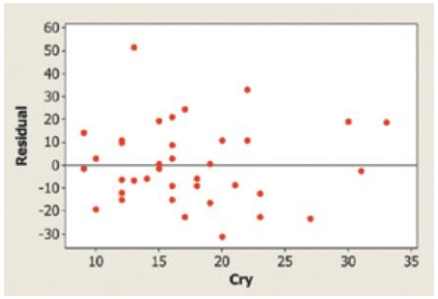
$$H_0 : \beta = 0$$
$$H_a : \beta > 0$$

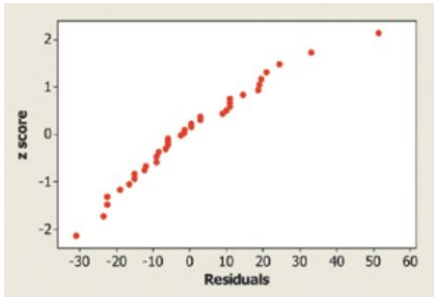
where β is the true slope of the population regression line relating crying count to IQ score. No significance level was given, so we'll use $\alpha = 0.05$.

Plan: If the conditions are met, we will perform a *t* test for the slope β.

- **Linear** The scatterplot suggests a moderately weak positive linear relationship between crying peaks and IQ. The residual plot shows a random scatter of points about the residual = 0 line.







- **Independent** Later IQ scores of individual infants should be independent. Due to sampling without replacement, there have to be at least 10(38) = 380 infants in the population from which these children were selected.
- **Normal** The Normal probability plot of the residuals shows a slight curvature, which suggests that the responses may not be Normally distributed about the line at each *x*-value. With such a large sample size ($n = 38$), however, the *t* procedures are robust against departures from Normality.
- **Equal variance** The residual plot shows a fairly equal amount of scatter around the horizontal line at 0 for all *x*-values.
- **Random** We are told that these 38 infants were randomly selected.

Example: Crying and IQ

Do: With no obvious violations of the conditions, we proceed to inference. The test statistic and *P*-value can be found in the Minitab output.

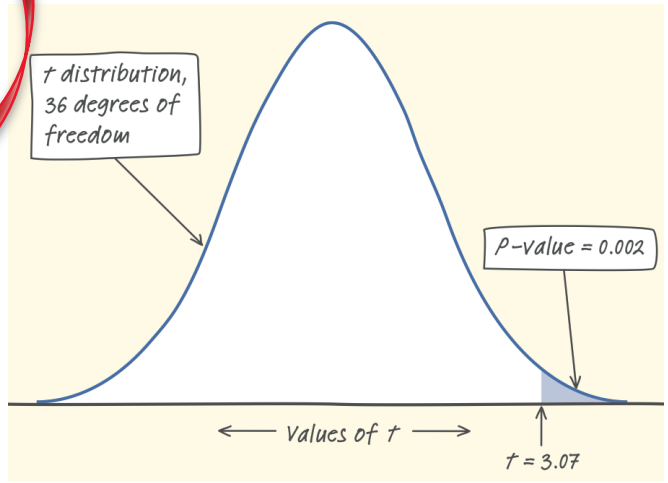
Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

$$t = \frac{b - \beta_0}{SE_b} = \frac{1.4929 - 0}{0.4870} = 3.07$$

The Minitab output gives $P = 0.004$ as the *P*-value for a two-sided test. The *P*-value for the one-sided test is half of this, $P = 0.002$.



Conclude: The *P*-value, 0.002, is less than our $\alpha = 0.05$ significance level, so we have enough evidence to reject H_0 and conclude that there is a positive linear relationship between intensity of crying and IQ score in the population of infants.