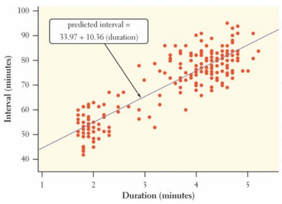
Chapter 12: More About Regression

Section 12.1
Inference for Linear Regression

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.

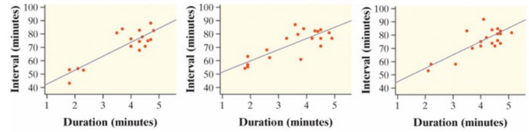


Suppose we take an SRS of 20 eruptions from the population and calculate the least-squares regression line $\hat{y} = a + bx$ for the sample data. How does the slope b of the **sample regression line** relate to the slope of the population regression line?

Feb 28-7:21 PM

Sampling Distribution of *b*

The graphs 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 least-squares regression line for that sample.



(a) Sample 1: $\hat{y} = 32.8 + 10.2x$ (b) Sample 2: $\hat{y} = 44.0 + 7.7x$ (c) 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.

Feb 28-7:26 PM

Sampling Distribution of a Slope

Choose an SRS of n observations (x, y) from a population of size N with least-squares regression line

predicted $y = \alpha + \beta_x$

Let b be the slope of the sample regression line. Then:

The **mean** of the sampling distribution of b is $\mu_b = \beta$.

The **standard deviation** of the sampling distribution of b is $\sigma_b = \dfrac{\sigma}{\sigma_x \sqrt{n}}$ as long as the *10% condition* is satisfied: $n \leq 0.10N$.

The sampling distribution of b will be **approximately** Normal if the values of the response variable y follow a Normal distribution for each value of the explanatory variable x (the *Normal condition*).

Feb 28-7:27 PM

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 actual 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 $\mu_y = \alpha + \beta x$.

Independent: Individual observations are independent of each other. When sampling without replacement, check the *10% condition.*

Normal: For any fixed value of x, the response y varies according to a Normal distribution.

Equal SD: The standard deviation of y (call it σ) is the same for all values of x.

Random: The data come from a well-designed random sample or randomized experiment.

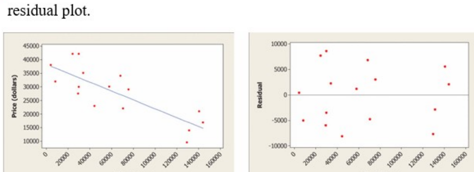
Feb 28-7:30 PM

How to Check the Conditions for Regression Inference

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.

The acronym **LINER** should help you remember the conditions for inference about regression.

L *__Linear:__* Examine the scatterplot to see if the overall pattern is roughly linear. Make sure there are no 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.



Good: Scatterplot has a linear form. Bad: Residual plot shows a curved pattern.

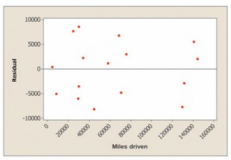
Feb 28-7:30 PM

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*). But there are other issues that can lead to a lack of independence. One example is measuring the same variable at intervals over time, yielding what is known as *time-series data*. Knowing that a young girl's height at age 6 is 48 inches would definitely give you additional information about her height at age 7. You should avoid doing inference about the regression model for time-series data.

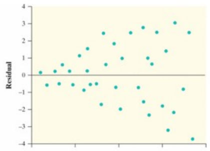
N ***Normal:*** Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality. Ideally, we would check the distribution of residuals for Normality at each possible value of x. Because we rarely have enough observations at each x-value, however, we make one graph of all the residuals to check for Normality.

Feb 28-7:31 PM

E ***Equal SD:** Look at the scatter of the residuals above and below the "residual = 0" line in the residual plot. The vertical spread of the residuals should be roughly the same from the smallest to the largest x-value.



Good: Residuals have roughly equal spread at all x-values in the data set.

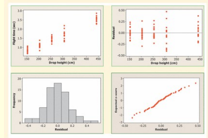


Bad: The response variable y has greater spread for larger values of the explanatory variable x.

R ***Random:** See if the data came from a well-designed random sample or randomized experiment. If not, we can't make inferences about a larger population or about cause and effect.

Feb 28-7:31 PM

Example 1: Mrs. Barrett's class did a helicopter experiment. 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 and residual plot, plus a histogram and Normal probability plot of the residuals are shown below.



PROBLEM: Check whether the conditions for performing inference about the regression model are met.

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

✓ **Independent:** Because the helicopters were released in a random order and no helicopter was used twice, knowing the result of one observation should not help us predict the value of another observation.

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

✓ **Equal SD:** 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 a little more for the helicopters that were dropped from a height of 307 cm.

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

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

Mar 1-12:22 PM

You will always see some irregularity when you look for Normality and equal standard deviation in the residuals, especially when you have few observations. Don't overreact to minor issues in the graphs when checking these two conditions.

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.

least-squares regression line, $\hat{y} = a + bx$

$\mu_y = \alpha + \beta x$

y-intercept a

the slope b

estimator of the population y-intercept α.

population slope β.

standard deviation σ

variability of the response y about the population regression line.

residuals estimate how much y varies about the population (true) line.

standard deviation of the residuals

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

Mar 3-9:19 AM

In a regression setting, we often want to estimate the slope β of the population (true) regression line. 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 gives a set of plausible values for β.

The confidence interval for β has the familiar form

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

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

$$b \pm t^* SE_b$$

We call this a ***t* interval for the slope**. Here are the details.

t INTERVAL FOR THE SLOPE

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

$$b \pm t^* SE_b$$

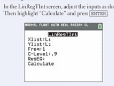
In this formula, the standard error of the slope is

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

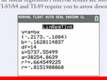
and t^* is the critical value for the t distribution with df = $n - 2$ having $C\%$ of its area between $-t^*$ and t^*.

TI-83/84 with recent OS

- Press STAT, then choose TESTS and LinRegTInt. . . .
- In the LinRegTInt screen, adjust the inputs as shown. Then highlight "Calculate" and press ENTER.



- The linear regression t interval results are shown below. Th
TI-83/84 and TI-89 require you to arrow down to see the re



Mar 3-9:27 AM

Example 2: 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. Some computer output from this regression is shown below. We checked conditions for performing inference earlier.

Regression Analysis: Flight time versus Drop height

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%

α y intercept

b slope

standard error

standard deviation

Write an equation to predict flight time: $\mu_y = \alpha + \beta x$

y = -.03761 +.0057244x

If a helicopter was dropped from 300 cm what would the predicted time be? 1.67839 seconds

About how far off do you expect the predicteion to be? 0.168181 seconds

a) Give the standard error of the slope, SE_b. Interpret this value in context.

We got the value of the standard error of the slope, 0.0002018, from the "SE Coef" column in the computer output. If we repeated the random assignment many times, the slope of the sample regression line would typically vary by about 0.0002 from the slope of the true regression line for predicting flight time from drop height.

Mar 3-9:35 AM

b) Find the critical value for a 95% confidence interval for the slope of the true regression line. Then calculate the confidence interval. Show your work.

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

$$b \pm t^* SE_b = 0.0057244 \pm 2.000(0.0002018)$$
$$= 0.0057244 \pm 0.0004026$$
$$= (0.0053218, 0.0061270)$$

c) Interpret the interval from part (b) in context.

We are 95% confident that the interval from 0.0053218 to 0.0061270 seconds per cm captures the slope of the true regression line relating the flight time y and drop height x of paper helicopters.

d) Explain the meaning of "95% confident" in context.

If we repeat the experiment many, many times, and use the method in part (b) to construct a confidence interval each time, about 95% of the resulting intervals will capture the slope of the true regression line relating flight time y and drop height x of paper helicopters.

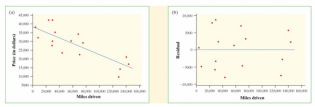
Mar 3-9:36 AM

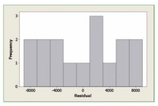
Example 3: Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4 × 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 SuperCrew 4 × 4s was selected from among those listed for sale on autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks.

Here are the data:

Miles driven:	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price (in dollars):	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven:	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price (in dollars):	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

Minitab output from a least-squares regression analysis for these data is shown on the next slide.





Regression Analysis: Price (dollars) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13   R-Sq = 66.4%   R-Sq(adj) = 64.0%

Write an equation to model the price of the trucks and make a **prediction** of the price of a truck if it has **30,000 miles**? How far do you **expect** the price to be off?

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

Mar 3-9:36 AM

STATE: We want to estimate the slope β of the population regression line relating miles driven to price with 90% confidence.

PLAN: If the conditions are met, we will use a t interval for the slope of a regression line.

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

Independent: Because we sampled without replacement to get the data, there have to be at least 10(16) = 160 used Ford F-150 SuperCrew 4 × 4s listed for sale on autotrader.com. This seems reasonable to believe.

Normal: The histogram of the residuals is roughly symmetric and single-peaked, so there are no obvious departures from Normality.

Equal SD: The scatter of points around the residual = 0 line appears to be about the same at all x-values.

Random: We randomly selected the 16 pickup trucks in the sample.

Mar 3-12:50 PM

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

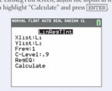
$$b \pm t^* SE_b = -0.16292 \pm 1.761(0.03096) = -0.16292 \pm 0.05452$$
$$= (-0.21744, -0.10840)$$

Using technology: The calculator's LinRegTInt gives (−0.2173, −0.1084) using df = 14.

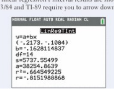
CONCLUDE: We are 90% confident that the interval from −0.2173 to −0.1084 captures the slope of the population regression line relating price to miles driven for used Ford F-150 SuperCrew 4 × 4s listed for sale on autotrader.com.

TI-83/84 with recent OS

- Press STAT, then choose TESTS and LinRegTInt. . . .
- In the LinRegTInt screen, adjust the inputs as shown. Then highlight "Calculate" and press ENTER.



- The linear regression t interval results are shown below. Th… TI-83/84 and TI-89 require you to arrow down to see the re…



Regression Analysis: Price (dollars) versus Miles driven

Predictor	Coef	SE Coef	T	P
Constant	38257	2446	15.64	0.000
Miles driven	−0.16292	0.03096	−5.26	0.000

S = 5740.13   R-Sq = 66.4%   R-Sq(adj) = 64.0%

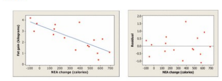
Mar 3-12:50 PM

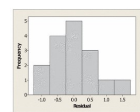
Try:

Does fidgeting keep you slim? Some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explain why—some people may spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed a random sample of 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) as the response variable and change in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like—as the explanatory variable. Here are the data:[4]

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

Minitab output from a least-squares regression analysis for these data is shown below.





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%

Write an equation to make a prediction.

Construct and interpret a 95% confidence interval for the slope of the population (true) regression line.

Correct Answer
S: β = slope of the population regression line relating fat gain to change in NEA. *P:* t interval for the slope. Linear: There is no leftover pattern in the residual plot. Independent: The sample size (n = 16) is less than 10% of all healthy young adults. Normal: The histogram of the residuals shows no strong skewness or outliers. Equal SD: Other than one point with a large positive residual, the residual plot shows roughly equal scatter for all x values. Random: Random sample. *D:* With df = 14, (−0.005032, −0.001852). *C:* We are 95% confident that the interval from −0.005032 to −0.001852 captures the slope of the population regression line relating fat gain to change in NEA.

Mar 3-1:01 PM

Day 2:

Performing a Significance Test for the Slope
When the conditions for inference are met, we can use the slope b of the sample regression line to construct a confidence interval for the slope β of the population (true) regression line. We can also perform a significance test to determine whether a specified value of β is plausible. The null hypothesis has the general form H_0: β = hypothesized value. To do a test, standardize b to get the test statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

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

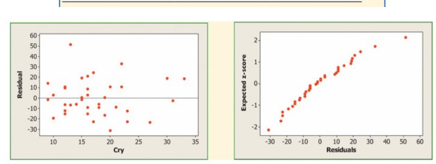
To find the P-value, use a t distribution with $n - 2$ degrees of freedom. Here are the details for the ***t* test for the slope**.

Suppose the conditions for inference are met. To test the hypothesis H_0: $\beta = \beta_0$, compute the test statistic $t = \dfrac{b - \beta_0}{SE_b}$

Mar 3-1:12 PM

Example 5: 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. The table below contains data from a random sample of 38 infants.

Crycount	IQ	Crycount	IQ	Crycount	IQ	Crycount	IQ
10	87	20	90	17	94	12	94
12	97	16	100	19	103	12	103
9	103	23	103	13	104	14	106
16	106	27	108	18	109	10	109
18	109	15	112	18	112	23	113
15	114	21	114	16	118	9	119
12	119	12	120	19	120	16	124
20	132	15	133	22	135	31	135
16	136	17	141	30	155	22	157
33	159	13	162				



Mar 6-10:50 AM

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%

a) What is the equation of the least-squares regression line for predicting IQ at age 3 from the number of crying peaks (crycount)? Interpret the slope and y intercept of the regression line in context.

The equation of the least-squares line is

predicted IQ score = 91.268 + 1.4929 (crycount)

Slope: For each additional crying peak in the most active 20 seconds, the regression line predicts an increase of about 1.5 IQ points.

y intercept: The model predicts that an infant who doesn't cry when flicked with a rubber band will have a later IQ score of about 91.

b) Explain what the value of s means in this setting.

The size of a typical prediction error when using the regression line in part (a) is 17.50 IQ points.

Mar 6-10:53 AM

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

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 do 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: 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 slight curvature, but no strong skewness or obvious outliers that would prevent use of t procedures.

Equal SD: 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.

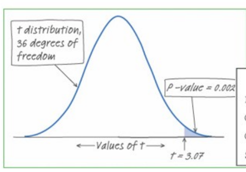
Mar 6-10:53 AM

Do: We can get the test statistic and P-value from the Minitab output.

Test statistic: $t = 3.07$ $\quad t = \frac{b - \beta_0}{SE_b} = \frac{1.4929 - 0}{0.4870} = 3.07$

df = n-2

P-value: The figure below displays the P-value for this one-sided test as an area under the t distribution curve with $38 - 2 = 36$ degrees of freedom. 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, $P = 0.002$. *table .0025< p < .001



Using technology: The calculator's LinRegTTest gives $t = 3.065$ and P-value = 0.002 using df = 36.

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 /2

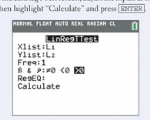
S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

CONCLUDE: Because the P-value, 0.002, is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence of a positive linear relationship between intensity of crying and IQ score in the population of infants.

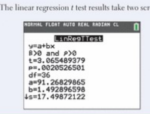
Mar 6-10:54 AM

TI-83/84

- Press STAT, then choose TESTS and LinRegTTest. . . .
- In the LinRegTTest screen, adjust the inputs as shown. Then highlight "Calculate" and press ENTER.



- The linear regression t test results take two screens to presen



AP® EXAM TIP The formula for the test statistic in a t test for the slope of a population (true) regression line often leads to calculation errors by students. As a result, we recommend using the calculator's LinRegTTest feature to perform calculations on the AP® exam. Be sure to name the procedure (t test for slope) and to report the test statistic ($t = 3.065$), P-value (0.002), and df (36) as part of the "Do" step.

Mar 6-10:56 AM

CHECK YOUR UNDERSTANDING

The previous Check Your Understanding (page 752) described some results from a study of nonexercise activity (NEA) and fat gain. Here, again, is the Minitab output from a least-squares regression analysis for these data.

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%

Do these data provide convincing evidence at the $\alpha = 0.05$ significance level of a negative linear relationship between fat gain and NEA change in the population of healthy young adults? Assume that the conditions for regression inference are met.

Correct Answer
S: H_0: $\beta = 0$ versus H_a: $\beta < 0$, where β is the slope of the true regression line relating fat gain to NEA change. *P:* t test for the slope β. *D*: $t = -4.64$. P-value ≈ 0.000/2 ≈ 0. *C*: Because the P-value of approximately 0 is less than $\alpha = 0.05$, we reject H0. There is convincing evidence that the slope of the true regression line relating fat gain to NEA change is negative.

Mar 6-10:57 AM