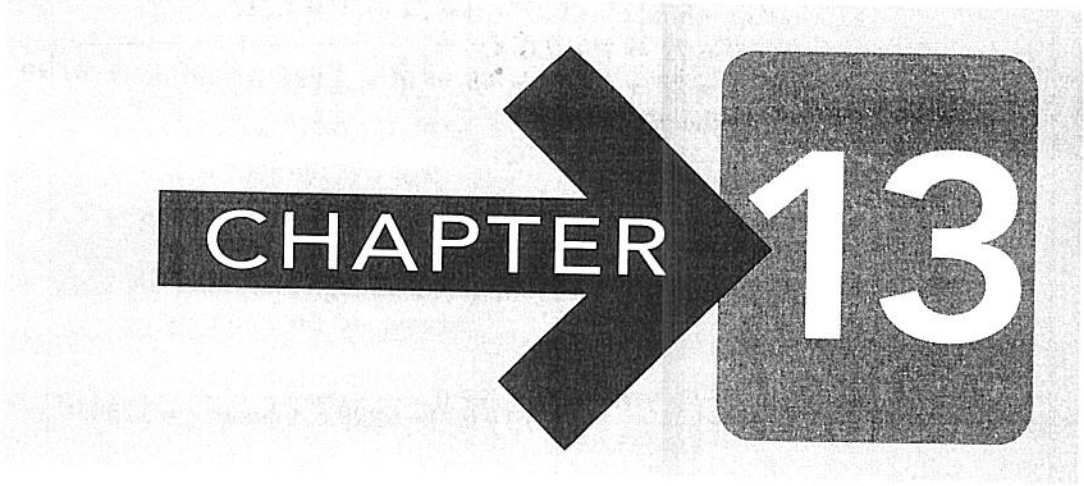


Inference for Regression

IN THIS CHAPTER

Summary: In the last two chapters, we've considered inference for population means and proportions and for the difference between two population means or two population proportions. In this chapter, we extend the study of linear regression begun in Chapter 7 to include inference for the slope of a regression line, including both confidence intervals and significance testing. Finally, we will look at the use of technology when doing inference for regression.

Key Ideas

- Simple Linear Regression (Review)
- Significance Test for the Slope of a Regression Line
- Confidence Interval for the Slope of a Regression Line
- Inference for Regression Using Technology

Simple Linear Regression

When we studied data analysis earlier in this text, we distinguished between *statistics* and *parameters.* Statistics are measurements or values that describe samples, and parameters are measurements that describe populations. We have also seen that statistics can be used to estimate parameters. Thus, we have used $\bar{x}$ to estimate the population mean μ, s to estimate the population standard deviation σ, etc. In Chapter 7, we introduced the least-squares regression line ($\hat{y} = a + bx$), which was based on a set of ordered pairs. $\hat{y}$ is actually a statistic because it is based on sample data. In this chapter, we study the parameter, μ_y, that is estimated by $\hat{y}$.

Before we look at the model for linear regression, let's consider an example to remind us of what we did in Chapter 7:

example: The following data are pulse rates and heights for a group of 10 female statistics students:

Height	70	60	70	63	59	55	64	64	72	66
Pulse	78	70	65	62	63	68	76	58	73	53

a. What is the least-squares regression line for predicting pulse rate from height?
b. What is the correlation coefficient between height and pulse rate? Interpret the correlation coefficient in the context of the problem.
c. What is the predicted pulse rate of a 67″ tall student?
d. Interpret the slope of the regression line in the context of the problem.

solution:

a. *Pulse rate* = 47.17 + 0.302 (*Height*). (Done on the TI-83/84 with *Height* in `L1` and *Pulse* in `L2`, the LSRL can be found `STAT CALC LinReg(a+bx) L1,L2,Y1`.)
b. $r = 0.21$. There is a weak, positive, linear relationship between Height and Pulse rate.
c. *Pulse rate* = 47.17 + 0.302(67) = 67.4. (On the Ti-83/84: `Y1(67)` = 67.42. Remember that you can paste `Y1` to the home screen by entering `VARS Y-VARS Function Y1`.)
d. For each increase in height of one inch, the pulse rate is predicted to increase by 0.302 beats per minute (or: the pulse rate will increase, on average, by 0.302 beats per minute).

When doing inference for regression, we use $\hat{y} = a + bx$ to estimate the true population regression line. Similar to what we have done with other statistics used for inference, we use a and b as estimators of population parameters α and β, the intercept and slope of the population regression line. The conditions necessary for doing inference for regression are:

- For each given value of x, the values of the response variable y-values are independent and normally distributed.
- For each given value of x, the standard deviation, σ, of y-values is the same.
- The mean response of the y-values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$.

example: Consider a situation in which we are interested in how well a person scores on an agility test after a fixed number of 3-oz. glasses of wine. Let x be the number of glasses consumed. Let x take on the values 1, 2, 3, 4, 5, and 6. Let y be the score on the agility test (scale: 1–100). Then for any given value x_i, there will be a distribution of y-values with mean μ_{y_i}. The conditions for inference for regression are that (i) each of these distributions of y-values are normally distributed, (ii) each of these distributions of y-values has the same standard deviation σ, and (iii) each of the μ_{y_i} lies on a line.

Remember that a *residual* was the error involved when making a prediction from a regression equation (residual = actual value of y – predicted value of $y = y_i - \hat{y}_i$). Not surprisingly, the standard error of the predictions is a function of the squared residuals:

$$s = \sqrt{\frac{SS_{RES}}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}.$$

s is an estimator of σ, the standard deviation of the residuals. Thus, there are actually three parameters to worry about in regression: α, β, and σ, which are estimated by a, b, and s, respectively.

The final statistic we need to do inference for regression is the standard error of the slope of the regression line:

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

In summary, inference for regression depends upon estimating $\mu_y = \alpha + \beta x$ with $\hat{y} = a + bx$. For each x, the response values of y are independent and follow a normal distribution, each distribution having the same standard deviation. Inference for regression depends on the following statistics:

- a, the estimate of the y intercept, α, of μ_y
- b, the estimate of the slope, β, of μ_y
- s, the standard error of the residuals
- s_b, the standard error of the slope of the regression line

In the section that follows, we explore inference for the slope of a regression line in terms of a significance test and a confidence interval for the slope.

Inference for the Slope of a Regression Line

Inference for regression consists of either a significance test or a confidence interval for the slope of a regression line. The null hypothesis in a significance test is usually H_0: $\beta = 0$, although it is possible to test H_0: $\beta = \beta_0$. Our interest is the extent to which a least-squares regression line is a good model for the data. That is, the significance test is a test of a linear model for the data.

We note that in theory we could test whether the slope of the regression line is equal to any specific value. However, the usual test is whether the slope of the regression line is zero or not. If the slope of the line is zero, then there is no linear relationship between the x and y variables (remember: $b = r\frac{s_y}{s_x}$; if $r = 0$, then $b = 0$).

The alternative hypothesis is often two sided (i.e., H_A: $\beta \neq 0$). We can do a one-sided test if we believed that the data were positively or negatively related.

Significance Test for the Slope of a Regression Line

The basic details of a significance test for the slope of a regression line are given in the following table:

• HYPOTHESIS • ESTIMATOR • STANDARD ERROR	CONDITIONS	TEST STATISTIC
• Null hypothesis $H_0 : \beta = \beta_0$ (most often: $H_0 : \beta = 0$) • Estimator: b (from: $\hat{y} = a + bx$) • Standard error of the residuals:	• For each given value of x, the values of the response variable y are independent and normally distributed.	

Continued

$$s = \sqrt{\frac{SS_{RES}}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}.$$

(Gives the variability of the vertical distances of the y-values from the regression line)

- Standard error of the slope:

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

(Gives the variability of the estimates of the slope of the regression line)

- For each given value of x, the standard deviation of y is the same.
- The mean response of the y-values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$.

$$t = \frac{b - \beta_o}{s_b} = \frac{b}{s_b} \text{ (if } \beta_o = 0\text{)},$$

$$\text{df} = n - 2$$

example: The data in the following table give the top 15 states in terms of per pupil expenditure in 1985 and the average teacher salary in the state for that year.

STATE/SALARY		PER PUPIL EXPENDITURE
MN	27360	3982
CO	25892	4042
OR	25788	4123
PA	25853	4168
WI	26525	4247
MD	27186	4349
DE	24624	4517
MA	26800	4642
RI	29470	4669
CT	26610	4888
DC	33990	5020
WY	27224	5440
NJ	27170	5536
NY	30678	5710
AK	41480	8349

Test the hypothesis, at the 0.01 level of significance, that there is no straight-line relationship between per pupil expenditure and teacher salary. Assume that the conditions necessary for inference for linear regression are present.

solution:

I. Let β = the true slope of the regression line for predicting salary from per pupil expenditure.

$$H_0\text{: } \beta = 0.$$

$$H_A\text{: } \beta \neq 0.$$

II. We will use the t-test for the slope of the regression line. The problem states that the conditions necessary for linear regression are present.

III. The regression equation is

Salary = 12027 + 3.34 *PPE*

(s = 2281, s_b = 0.5536)

$$t = \frac{3.34 - 0}{0.5536} = 6.04, \text{ df} = 15 - 2$$

$$= 13 \Rightarrow P - \text{value} = 0.0000.$$

(To do this significance test for the slope of a regression line on the TI-83/84, first enter *Per Pupil Expenditure* (the explanatory variable) in `L1` and *Salary* (the response variable) in `L2`. Then go to `STAT TESTS LinRegTTest` and enter the information requested. The calculator will return the values of t, p (the P-value), df, a, b, s, r^2, and r. Minitab will not give the the value of r—you'll have to take the appropriate square root of r^2—but will give you the value of s_b. If you need s_b for some reason—such as constructing a confidence interval for the slope of the regression line—and only have access to a calculator, you can find it by noting that, since $t = \frac{b}{s_b}$, then $s_b = \frac{b}{t}$. Note that Minitab reports the P-value as 0.0000.)

IV. Because $P < \alpha$, we reject H_0. We have evidence that the true slope of the regression line is not zero. We have evidence that there is a linear relationship between amount of per pupil expenditure and teacher salary.

A significance test that the slope of a regression line equals zero is closely related to a test that there is no correlation between the variables. That is, if ρ is the population correlation coefficient, then the test statistic for H_0: $\beta = 0$ is equal to the test statistic for H_0: $\rho = 0$. You aren't required to know it for the AP exam, but the t-test statistic for H_0: $\rho = 0$, where r is the sample correlation coefficient, is

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \text{ df} = n - 2.$$

Because this and the test for a non-zero slope are equivalent, it should come as no surprise that

$$r\sqrt{\frac{n-2}{1-r^2}} = \frac{b}{s_b}.$$

Confidence Interval for the Slope of a Regression Line

In addition to doing hypothesis tests on H_0: $\beta = \beta_0$, we can construct a confidence interval for the true slope of a regression line. The details follow:

• PARAMETER • ESTIMATOR	CONDITIONS	FORMULA
• Population slope: β • Estimator: b (from: $\hat{y} = a + bx$) • Standard error of the residuals: $s = \sqrt{\frac{SS_{RES}}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$ • Standard error of the slope: $s_b = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$	• For each given value of x, the values of the response variable y are independent and normally distributed. • For each given value of x, the standard deviation of y is the same. • The mean response of the y-values for the fixed values of x are linearly related by the equation $\mu_y = \alpha + \beta x$.	$b \pm t^* s_b$, df $= n - 2$ (where t^* is the upper critical value of t for a C-level confidence interval)

example: Consider once again the earlier example on predicting teacher salary from per pupil expenditure. Construct a 95% confidence interval for the slope of the population regression line.

solution: When we were doing a test of H_o: $\beta = 0$ for that problem, we found that *Salary* = 12027 + 3.34 PPE. The slope of the regression line for the 15 points, and hence our estimate of β, is $b = 3.34$. We also had $t = 6.04$.

Our confidence interval is of the form $b \pm t^* s_b$. We need to find t^* and s_b. For $C = 0.95$, df = 15 − 2 = 13, we have $t^* = 2.160$ (from Table B; if you have a TI-84 with the `invT` function, use `invT(0.975,13)`). Now, as mentioned earlier, $s_b = \frac{b}{t} = \frac{3.34}{6.04} = 0.5530$.

Hence, $b \pm t^* s_b = 3.34 \pm 2.160(0.5530) = (2.15, 4.53)$. We are 95% confident that the true slope of the regression line is between 2.15 and 4.53. Note that, since 0 is *not* in this interval, this finding is consistent with our earlier rejection of the hypothesis that the slope equals 0. This is another way of saying that we have statistically significant evidence of a predictive linear relationship between PPE and *Salary*.

(If your TI-84 has the `invT` function in the `DISTR` menu, it also has, in the `STAT TESTS` menu, `LinRegTInt`, which will return the interval (2.146, 4.538). It still doesn't tell you the value of s_b. There is a more complete explanation of how to use technology to do inference for regression in the next section.)

Inference for Regression Using Technology

If you had to do them from the raw data, the computations involved in doing inference for the slope of a regression line would be daunting.

For example, how would you like to compute $s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$ by hand?

Fortunately, you probably will never have to do this by hand, but instead can rely on computer output you are given, or you will be able to use your calculator to do the computations.

Consider the following data that were gathered by counting the number of cricket chirps in 15 seconds and noting the temperature.

Number of Chirps	22	27	35	15	28	30	39	23	25	18	35	29
Temperature (F)	64	68	78	60	72	76	82	66	70	62	80	74

We want to use technology to test the hypothesis that the slope of the regression line is 0 and to construct a confidence interval for the true slope of the regression line.

First let us look at the Minitab regression output for this data.

The regression equation is
Temp = 44.0 + 0.993 Number

Predictor	Coef	St Dev	*t* ratio	*P*
Constant	44.013	1.827	24.09	0.000
Number	0.99340	0.06523	15.23	0.000
s = 1.538	R-sq = 95.9%		R-sq(adj) = 95.5%	

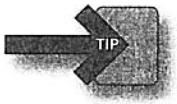
You should be able to read most of this table, but you are not responsible for all of it. You see the following table entries:

- The regression equation, Temp = 44.0 + 0.993 Number, is the Least Squares Regression Line (LSRL) for predicting temperature from the number of cricket chirps.
- Under "Predictor" are the y-intercept and explanatory variable of the regression equation, called "Constant" and "Number" in this example.
- Under "Coef" are the values of the "Constant" (which equals the y-intercept, the a in $\hat{y} = a + bx$; here, $a = 44.013$) and the slope of the regression line (which is the coefficient of "Number" in this example, the b in $\hat{y} = a + bx$; here, $b = 0.99340$).
- For the purposes of this course, we are not concerned with the "Stdev," "t-ratio," or "p" for "Constant" (the last three entries in the first line of printout—only the "44.013" is meaningful for us).
- "Stdev" of "Number" is the standard error of the slope (what we have called s_b, the variability of the estimates of the slope of the regression line, which *equals* here $\frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$ s_b = 0.06523); "t-ratio" is the value of the t-test statistic ($t = \frac{b}{s_b}$,

df = $n - 2$; here, $t = \dfrac{0.99340}{0.06523} = 15.23$); and P is the P-value associated with the test statistic assuming a two-sided test (here, $P = 0.000$; if you were doing a *one*-sided test, you would need to divide the given P-value by 2).

- s is the standard error of the residuals (which is the variability of the vertical distances of the y-values from the regression line; $s = \sqrt{\dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}}$; (here, $s = 1.538$.)
- "R-sq" is the coefficient of determination (or, r^2; here R-sq = 95.9% ⇒ 95.9% of the variation in temperature that is explained by the regression on the number of chirps in 15 seconds; note that, here, $r = \sqrt{0.959} = 0.979$—it's positive since $b = 0.9934$ is positive). You don't need to worry about "R-sq(adj)."

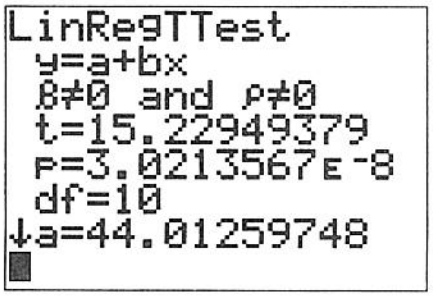
All of the mechanics needed to do a t-test for the slope of a regression line are contained in this printout. You need only to quote the appropriate values in your write-up. Thus, for the problem given above, we see that $t = 15.23 \Rightarrow P$-value = 0.000.

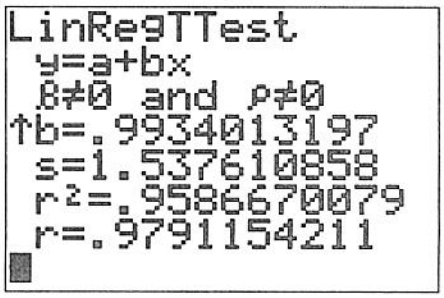


Exam Tip: You may be given a problem that has both the raw data and the computer printout based on the data. If so, there is no advantage to doing the computations all over again because they have already been done for you.

A confidence interval for the slope of a regression line follows the same pattern as all confidence intervals (estimate ± (critical value) × (standard error)): $b \pm t^* s_b$, based on $n - 2$ degrees of freedom. A 99% confidence interval for the slope in this situation (df = 10 ⇒ $t^* = 3.169$ from Table B) is $0.9934 \pm 3.169(0.06523) = (0.787, 1.200)$. (*Note*: The newest software for the TI-84 has a `LinRegTInt` built in. The TI-83/84 and earlier versions of the TI-84 do not have this program. The program requires that the data be available in lists and, unlike other confidence intervals in the `STAT TESTS` menu, there is no option to provide `Stats` rather than `Data`.)

To use the calculator to do the regression, enter the data in, say, `L1` and `L2`. Then go to `STAT TESTS LinRegTTest`. Enter the data as requested (response variable in the Ylist:). Assuming the alternative is two sided (H_A: $\beta \neq 0$), choose β and $\rho \neq 0$. Then `Calculate`. You will get the following two screens of data:





This contains all of the information in the computer printout except s_b. It does give the number of degrees of freedom, which Mini- Tab does not, as well as greater accuracy. Note that the calculator lumps together the test for both the slope (β) and the correlation coefficient (ρ) because, as we noted earlier, the test statistics are the same for both.

If you *have* to do a confidence interval using the calculator and do not have a TI-84 with the `LinRegTInt` function, you first need to determine s_b. Because you know that $t = \frac{b}{s_b} \Rightarrow s_b = \frac{b}{t}$, it follows that $s_b = \frac{0.9934}{15.2295} = 0.0652$, which agrees with the standard error of the slope ("St Dev" of "Number") given in the computer printout.

A 95% confidence interval for the slope of the regression line for predicting temperature from the number of chirps per minute is then given by $0.9934 \pm 2.228(0.0652) = (0.848, 1.139)$. $t^* = 2.228$ is based on $C = 0.95$ and df $= 12 - 2 = 10$. Using `LinRegTInt`, if you have it, results in the following (note that the "s" given in the printout is the standard error of the residuals, not the standard error of the slope):

```
LinRegTInt
 y=a+bx
 (.84806,1.1387)
 b=.9934013197
 df=10
 s=1.537610858
↓a=44.01259748
```

› Rapid Review

1. The regression equation for predicting grade point average from number of hours studied is determined to be GPA = 1.95 + 0.05(Hours). Interpret the slope of the regression line.

   Answer: For each additional hour studied, the GPA is predicted to increase by 0.05 points.

2. Which of the following is *not* a necessary condition for doing inference for the slope of a regression line?

   a. For each given value of the independent variable, the response variable is normally distributed.
   b. The values of the predictor and response variables are independent.
   c. For each given value of the independent variable, the distribution of the response variable has the same standard deviation.
   d. The mean response values lie on a line.

   Answer: (b) is not a condition for doing inference for the slope of a regression line. In fact, we are trying to find out the degree to which they are not independent.

3. True–False: Significance tests for the slope of a regression line are always based on the hypothesis H_0: $\beta = 0$ versus the alternative H_A: $\beta \neq 0$.

   Answer: False. While the stated null and alternative may be the usual hypotheses in a test about the slope of the regression line, it is possible to test that the slope has some particular non-zero value so that the alternative can be one sided (H_A: $B > 0$ or H_A: $\beta < 0$). Note that most computer programs will test only the two-sided alternative by default. The TI-83/84 will test either a one- or two-sided alternative.