Advanced High School Statistics

First Edition

David M Diez
david@openintro.org
Christopher D Barr
Yale School of Management
chris@openintro.org
Mine Çetinkaya-Rundel
Duke University
mine@openintro.org
Leah Dorazio
San Francisco University High School
leah@openintro.org

Chapter 8

Introduction to linear regression

8.4 Inference for the slope of a regression line

In this section we discuss uncertainty in the estimates of the slope and y-intercept for a regression line. Just as we identified standard errors for point estimates in previous chapters, we first discuss standard errors for these new estimates. However, in the case of regression, we will identify standard errors using statistical software.

8.4.1 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2010, with the exception

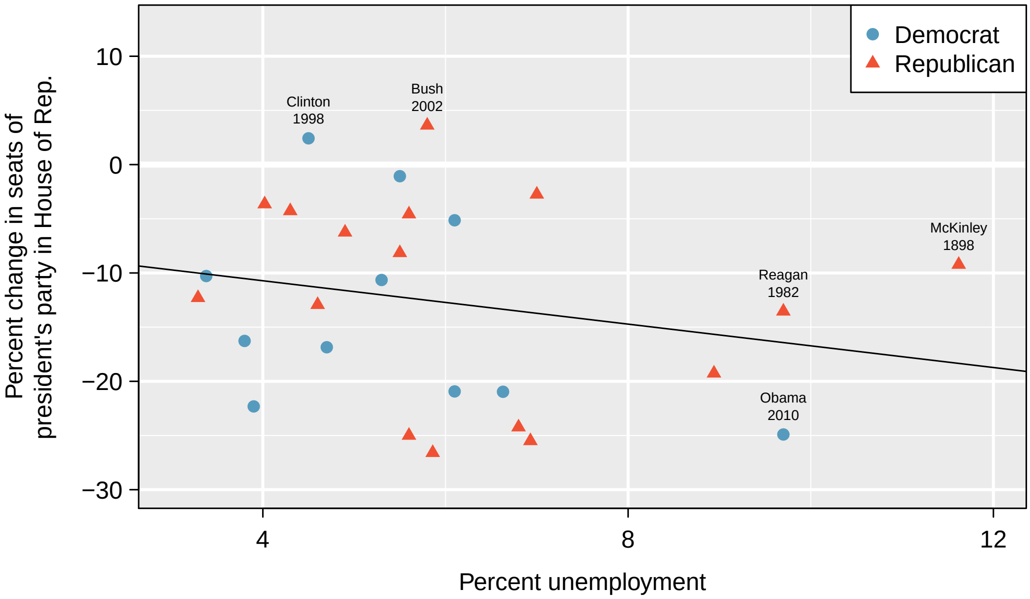


Figure 8.21: The percent change in House seats for the President's party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been fit to the data.

of those elections during the Great Depression. Figure 8.21 shows these data and the least-squares regression line:

$$\begin{aligned}&\%\text{ change in House seats for President's party}\\&\qquad= -6.71 - 1.00 \times (\text{unemployment rate})\end{aligned}$$

We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Democrats in 2010) against the unemployment rate.

Examining the data, there are no clear deviations from linearity, the constant variance condition, or in the normality of residuals (though we don't examine a normal probability plot here). While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found.

⊙ **Guided Practice 8.27** The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?[17]

[17] We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a final analysis.

There is a negative slope in the line shown in Figure 8.21. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the "true" linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate? We can frame this investigation into a one-sided statistical hypothesis test:

H_0: $\beta_1 = 0$. The true linear model has slope zero.

H_A: $\beta_1 < 0$. The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.

We would reject H_0 in favor of H_A if the data provide strong evidence that the true slope parameter is less than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

8.4.2 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1. We will generally label the test statistic using a T, since it follows the t-distribution.

TIP: Hypothesis tests on the slope of the regression line
Use a t-test with $n - 2$ degrees of freedom when performing a hypothesis test on the slope of a regression line.

We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. The table below shows software output for the least squares regression line in Figure 8.21. The row labeled *unemp* represents the information for the slope, which is the coefficient of the unemployment variable.

```
The regression equation is

Change = -6.7142 - 1.0010 unemp

Predictor      Coef   SE Coef       T         P
Constant    -6.7142    5.4567   -1.23    0.2300
unemp       -1.0010    0.8717   -1.15    0.2617

S = 9.624    R-Sq = 0.03%    R-Sq(adj) = -3.7%
```

● **Example 8.28** What do the first and second columns of numbers in the regression summary represent?

The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate.

We previously used a T test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope

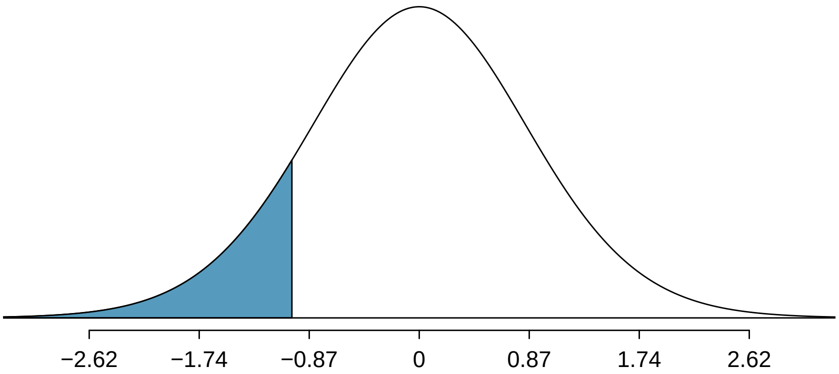


Figure 8.22: The distribution shown here is the sampling distribution for b_1, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

is 0, so we can compute the test statistic using the T score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-1.0010 - 0}{0.8717} = -1.15$$

We can look for the one-sided p-value – shown in Figure 8.22 – using the probability table for the t-distribution in Appendix B.3 on page 452.

● **Example 8.29** In this example, the sample size $n = 27$. Identify the degrees of freedom and p-value for the hypothesis test.

The degrees of freedom for this test is $n - 2$, or $df = 27 - 2 = 25$. Looking in the 25 degrees of freedom row in Appendix B.3, we see that the absolute value of the test statistic is smaller than any value listed, which means the tail area and therefore also the p-value is larger than 0.100 (one tail!). Because the p-value is so large, we fail to reject the null hypothesis. That is, the data do not provide convincing evidence that a higher unemployment rate has any correspondence with smaller or larger losses for the President's party in the House of Representatives in midterm elections.

We could have identified the T test statistic from the software output of the regression model, shown in the `unemp` row and third column (t value). The entry in the `unemp` row and last column represents the p-value for the two-sided hypothesis test where the null value is zero. The corresponding one-sided test would have a p-value half of the listed value.

Inference for regression

We usually rely on statistical software to identify point estimates and standard errors for parameters of a regression line. After verifying conditions hold for fitting a line, we can use the methods learned in Section 7.1 for the t-distribution to create confidence intervals for regression parameters or to evaluate hypothesis tests.

Caution: Don't carelessly use the p-value from regression output

The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area. If neither of these scenarios match your hypothesis test, be cautious about using the software output to obtain the p-value.

Example 8.30 Examine Figure 8.16 on page 390, which relates the Elmhurst College aid and student family income. How sure are you that the slope is statistically significantly different from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. This suggests the hypothesis test will reject the null claim that the slope is zero.

Recall that $b_1 = r\frac{s_y}{s_x}$. If the slope of the true regression line is zero, the population correlation coefficient must also be zero. The linear regression test for $\beta_1 = 0$ is equivalent, then, to a test for the population correlation coefficient $\rho = 0$.

Guided Practice 8.31 The regression summary below shows statistical software output from fitting the least squares regression line shown in Figure 8.16. Use this output to formally evaluate the following hypotheses. H_0: The true slope of the regression line is zero. H_A: The true slope of the regression line is not zero.[18]

```
The regression equation is

aid = 24.31933 - 0.04307 family_income

Predictor        Coef        SE Coef    T         P
Constant         24.31933    1.29145    18.831    < 2e-16
family_income    -0.04307    0.01081    -3.985    0.000229

S = 4.783    R-Sq = 24.86%    R-Sq(adj) = 23.29%
```

TIP: Always check assumptions

If conditions for fitting the regression line do not hold, then the methods presented here should not be applied. The standard error or distribution assumption of the point estimate – assumed to be normal when applying the T test statistic – may not be valid.

[18] We look in the second row corresponding to the family income variable. We see the point estimate of the slope of the line is -0.0431, the standard error of this estimate is 0.0108, and the T test statistic is -3.98. The p-value corresponds exactly to the two-sided test we are interested in: 0.0002. The p-value is so small that we reject the null hypothesis and conclude that family income and financial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated and the true slope parameter is indeed less than 0, just as we believed in Example 8.30.

8.4.3 Summarizing inference procedures for linear regression

Hypothesis test for the slope of regression line

1. State the name of the test being used.
   - Linear regression t-test
2. Verify conditions.
   - The residual plot has no pattern.
3. Write the hypotheses in plain language. No mathematical notation is needed for this test.
   - H_0: $\beta_1 = 0$, There is no significant linear relationship between [x] and [y].
   - H_A: $\beta_1 \neq$, or $<$, or > 0, There is a significant or significant negative or significant positive linear relationship between [x] and [y].
4. Identify the significance level α.
5. Calculate the test statistic and df: $T = \frac{\text{point estimate}-\text{null value}}{\text{SE of estimate}}$
   - The point estimate is b_1
   - SE can be located on regression summary table next to value of b_1
   - $df = n - 2$
6. Find the p-value, compare it to α, and state whether to reject or not reject the null hypothesis.
7. Write the conclusion in the context of the question.

Constructing a confidence interval for the slope of regression line

1. State the name of the CI being used.
   - t-interval for slope of regression line
2. Verify conditions.
   - The residual plot has no pattern.
3. Plug in the numbers and write the interval in the form

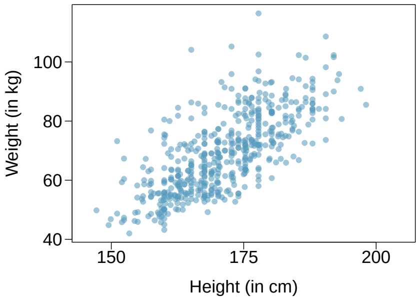
   $$\text{point estimate } \pm t^{\star} \times \text{SE of estimate}$$

   - The point estimate is b_1.
   - $df = n - 2$
   - The critical value t^* can be found on the t-table at row $df = n - 2$
   - SE can be located on regression summary table next to value of b_1
4. Evaluate the CI and write in the form (_ , _).
5. Interpret the interval: "We are [XX]% confident that this interval contains the true average increase in [y] for each additional [unit] of [x].
6. State the conclusion to the original question.

8.6.4 Inference for the slope of a regression line

In the following exercises, visually check the conditions for fitting a least squares regression line, but you do not need to report these conditions in your solutions.

8.35 Body measurements, Part IV. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-105.0113	7.5394	-13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

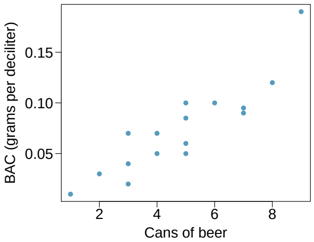
(a) Describe the relationship between height and weight.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for height and weight is 0.72. Calculate R^2 and interpret it in context.

8.36 Beer and blood alcohol content. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.[27] The scatterplot and regression table summarize the findings.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

(a) Describe the relationship between the number of cans of beer and BAC.

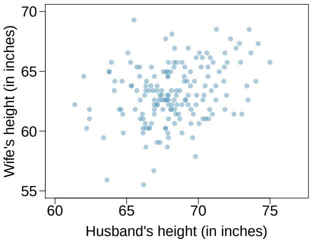
(b) Write the equation of the regression line. Interpret the slope and intercept in context.

(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

8.37 Husbands and wives, Part II. The scatterplot below summarizes husbands' and wives' heights in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Summary output of the least squares fit for predicting wife's height from husband's height is also provided in the table.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43.5755	4.6842	9.30	0.0000
height_husband	0.2863	0.0686	4.17	0.0000

(a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting wife's height from husband's height.

(c) Interpret the slope and intercept in the context of the application.

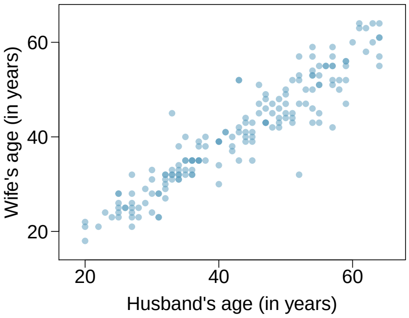
(d) Given that $R^2 = 0.09$, what is the correlation of heights in this data set?

(e) You meet a married man from Britain who is 5'9" (69 inches). What would you predict his wife's height to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict his wife's height? Why or why not?

[27] J. Malkevitch and L.M. Lesser. *For All Practical Purposes: Mathematical Literacy in Today's World.* WH Freeman & Co, 2008.

8.38 Husbands and wives, Part III. Exercise 8.6 presents a scatterplot displaying the relationship between husbands' and wives' ages in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Given below is summary output of the least squares fit for predicting wife's age from husband's age.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5740	1.1501	1.37	0.1730
age_husband	0.9112	0.0259	35.25	0.0000
				df = 168

(a) We might wonder, is the age difference between husbands and wives consistent across ages? If this were the case, then the slope parameter would be $\beta_1 = 1$. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages differs for different ages.

(b) Write the equation of the regression line for predicting wife's age from husband's age.

(c) Interpret the slope and intercept in context.

(d) Given that $R^2 = 0.88$, what is the correlation of ages in this data set?

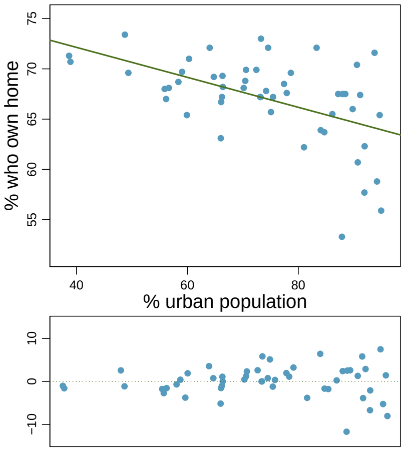
(e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain.

8.39 Urban homeowners, Part II. Exercise 8.33 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

(a) For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?

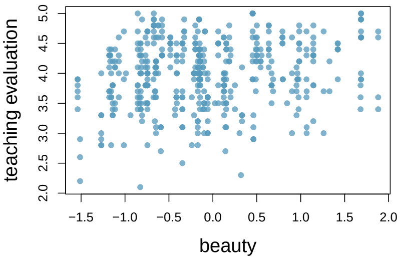
(b) Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?



8.40 Rate my professor. Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors.[28] The scatterplot below shows the relationship between these variables,

[28] Daniel S Hamermesh and Amy Parker. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". In: *Economics of Education Review* 24.4 (2005), pp. 369–376.

and also provided is a regression output for predicting teaching evaluation score from beauty score.

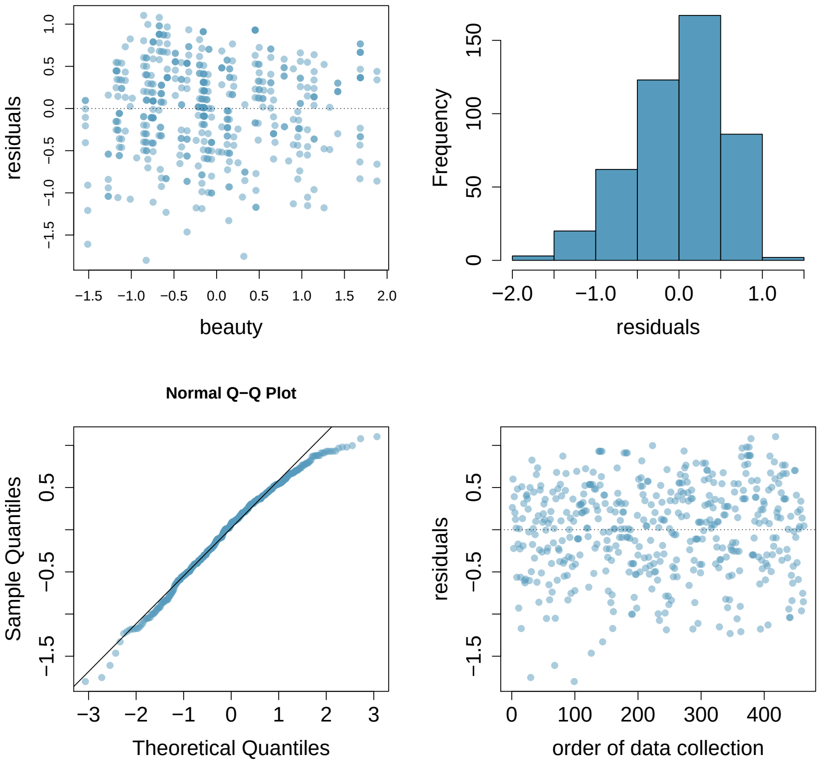


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.010	0.0255	157.21	0.0000
beauty	□	0.0322	4.13	0.0000

(a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table.

(b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.

(c) List the conditions required for linear regression and check if each one is satisfied for this model based on the following diagnostic plots.



8.41 Murders and poverty, Part II. Exercise 8.29 presents regression output from a model for predicting annual murders per million from percentage living in poverty based on a random sample of 20 metropolitan areas. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000
$s = 5.512$	$R^2 = 70.52\%$		$R^2_{adj} = 68.89\%$	

(a) What are the hypotheses for evaluating whether poverty percentage is a significant predictor of murder rate?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of poverty percentage, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

8.42 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-five low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circumference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is

$$\widehat{head\_circumference} = 3.91 + 0.78 \times gestational\_age$$

(a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?

(b) The standard error for the coefficient of gestational age is 0.35, which is associated with $df = 23$. Does the model provide strong evidence that gestational age is significantly associated with head circumference?

8.43 Murders and poverty, Part III. In Exercises 8.41 you evaluated whether poverty percentage is a significant predictor of murder rate. How, if at all, would your answer change if we wanted to find out whether poverty percentage is positively associated with murder rate. Make sure to include the appropriate p-value for this hypothesis test in your answer.

8.44 Cats, Part II. Exercise 8.30 presents regression output from a model for predicting the heart weight (in g) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cat. The model output is also provided below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000
$s = 1.452$	$R^2 = 64.66\%$		$R^2_{adj} = 64.41\%$	

(a) What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?

(b) State the conclusion of the hypothesis test from part (a) in context of the data.

(c) Calculate a 95% confidence interval for the slope of body weight, and interpret it in context of the data.

(d) Do your results from the hypothesis test and the confidence interval agree? Explain.

Appendix A

End of chapter exercise solutions

8.35 (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential. (b) $\widehat{weight} = -105.0113 + 1.0176 \times height$. Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds). Intercept: People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is obviously not possible. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$). H_0: The true slope coefficient of height is greater than zero ($\beta_1 > 0$). A two-sided test would also be acceptable for this application. The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so the p-value for the one-sided hypothesis will be even smaller. That is, we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0. (d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

8.37 (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 > 0$. A two-sided test would also be acceptable for this application. The p-value, as reported in the table, is incredibly small. Thus, for a one-sided test, the p-value will also be incredibly small, and we reject H_0. The data provide convincing evidence that wives' and husbands' heights are positively correlated. (b) $\widehat{height}_W = 43.5755 + 0.2863 \times height_H$. (c) Slope: For each additional inch in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average. Intercept: Men who are 0 inches tall are expected to have wives who are, on average, 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line. (d) The slope is positive, so r must also be positive. $r = \sqrt{0.09} = 0.30$. (e) 63.2612. Since R^2 is low, the prediction based on this regression model is not very reliable. (f) No, we should avoid extrapolating.

8.39 (a) $r = \sqrt{0.28} \approx -0.53$, we know the correlation is negative due to the negative association shown in the scatterplot. (b) The residuals appear to be fan shaped, indicating non-constant variance. Therefore a simple least squares fit is not appropriate for these data.

8.41 (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million. (d) Yes, we rejected H_0 and the confidence interval does not include 0.

8.43 This is a one-sided test, so the p-value should be half of the p-value given in the regression table, which will be approximately 0. Therefore the data provide convincing evidence that poverty percentage is positively associated with murder rate.