13.6: Exercises

Line tting, residuals, and correlation

7.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

7.2 Trends in the residuals. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

7.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether tting a linear model would be reasonable.

7.4 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether tting a linear model would be reasonable.

7.5 The two scatterplots below show the relationship between nal and mid-semester exam grades recorded during several years for a Statistics course at a university.

(a) Based on these graphs, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the final exam is higher?

7.6 Husbands and wives, Part I. The Great Britain Office of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.16 The scatterplot on the left shows the wife's age plotted against her husband's age, and the plot on the right shows wife's height plotted against husband's height.

(a) Describe the relationship between husbands' and wives' ages.

(b) Describe the relationship between husbands' and wives' heights.

(c) Which plot shows a stronger correlation? Explain your reasoning.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

7.7 Match the correlation, Part I.

Match the calculated correlations to the corresponding scatterplot.

(a) R = -0.7

(b) R = 0.45

(c) R = 0.06

(d) R = 0.92

7.8 Match the correlation, Part II.

Match the calculated correlations to the corresponding scatterplot.

(a) R = 0.49

(b) R = -0.48

(c) R = -0.03

(d) R = -0.85

[16]*D.J. Hand. A handbook of small data sets. Chapman & Hall/CRC, 1994.*

7.9 Speed and height. 1,302 UCLA students were asked to ll out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.

(a) Describe the relationship between height and fastest speed.

(b) Why do you think these variables are positively associated?

(c) What role does gender play in the relationship between height and fastest driving speed?

7.10 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.17

(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

[17]*Source: R Dataset, http://stat.ethz.ch/R-manual/R-patch...tml/trees.html*.

7.11 The Coast Starlight, Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

(a) Describe the relationship between distance and travel time.

(b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?

(c) Correlation between travel time (in miles) and distance (in minutes) is R = 0.636. What is the correlation between travel time (in kilometers) and distance (in hours)?

7.12 Crawling babies, Part I. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.18 Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (^{0}F) and age is measured in weeks.

(a) Describe the relationship between temperature and crawling age.

(b) How would the relationship change if temperature was measured in degrees Celsius (^{0}C) and age was measured in months?

(c) The correlation between temperature in ^{0}F and age in weeks was R = -0.70. If we converted the temperature to ^{0}C and age to months, what would the correlation be?

[18]J.B. Benson. "Season of birth and onset of locomotion: Theoretical and methodological implications". In: Infant behavior and development 16.1 (1993), pp. 69-81. issn: 0163-6383.

7.13 Body measurements, Part I. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.19 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height.

(b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

7.14 Body measurements, Part II. The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described in Exercise 7.13.

(a) Describe the relationship between hip girth and weight.

(b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?

7.15 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

7.16 Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

(a) $5,000 more than women?

(b) 25% more than women?

(c) 15% less than women?

[19]*G. Heinz et al. "Exploring relationships in body dimensions". In: Journal of Statistics Education 11.2 (2003).*

Fitting a line by least squares regression

7.17 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.20 The scatterplot below shows the relationship between these two variables along with the least squares fit.

(a) Describe the relationship between number of tourists and spending.

(b) What are the explanatory and response variables?

(c) Why might we want to t a regression line to these data?

(d) Do the data meet the conditions required for tting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.

7.18 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.21 Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.

(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

(b) In this scenario, what are the explanatory and response variables?

(c) Why might we want to t a regression line to these data?

(d) Do these data meet the conditions required for tting a least squares line?

7.19 The Coast Starlight, Part II. Exercise 7.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 107 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R^2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R^2 in the context of the application.

(d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.

(e) It actually takes the the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value.

(f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

[21]*Source: Starbucks.com, collected on March 10, 2011, www.starbucks.com/menu/nutrition.*

7.20 Body measurements, Part III. Exercise 7.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 108.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

(a) Write the equation of the regression line for predicting height.

(b) Interpret the slope and the intercept in this context.

(c) Calculate R2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

(e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means.

(f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

7.21 Grades and TV. Data were collected on the number of hours per week students watch TV and the grade they earned in a biology class on a 100 point scale. Based on the scatterplot and the residual plot provided, describe the relationship between the two variables, and determine if a simple linear model is appropriate to predict a student's grade from the number of hours per week the student watches TV.

7.22 Nutrition at Starbucks, Part II. Exercise 7.18 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.

7.23 Helmets and lunches. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.

(a) If the R2 for the least-squares regression line for these data is 72%, what is the correlation between lunch and helmet?

(b) Calculate the slope and intercept for the leastsquares regression line for these data.

(c) Interpret the intercept of the least-squares regression line in the context of the application.

(d) Interpret the slope of the least-squares regression line in the context of the application.

(e) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.

Types of outliers in linear regression

7.24 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.

7.25 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.

7.26 Crawling babies, Part II. Exercise 7.12 introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average rst crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53^0F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an inuential point?

7.27 Urban homeowners, Part I. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas in 2010.22 There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.

(a) Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas in 2010.

(b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?

Inference for linear regression

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

7.28 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.23 The scatterplot and regression table summarize the ndings.

	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?

[22]*United States Census Bureau, 2010 Census Urban and Rural Classi cation and Urban Area Criteria and Housing Characteristics: 2010.*

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

7.29 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 R2 and interpret it in context.

7.30 Husbands and wives, Part II. Exercise 7.6 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

(a) We might wonder, is the age difference between husbands and wives constant over time? If this were the case, then the slope parameter would be 1 = 1. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages actually has changed.

(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 R2 = 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.

7.31 Husbands and wives, Part III. 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 t 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 R2 = 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?

7.32 Urban homeowners, Part II. Exercise 7.27 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, R2 = 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?

7.33 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty- ve 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

head circdumference = 3:91 + 0:78 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 signi cantly associated with head circumference?

7.34 Rate my professor. Some college students critique professors' teaching at RateMyProfessors.com, a web page where students anonymously rate their professors on quality, easiness, and attractiveness. Using the self-selected data from this public forum, researchers examine the relations between quality, easiness, and attractiveness for professors at various universities. In this exercise we will work with a portion of these data that the researchers made publicly available.[24]

The scatterplot on the right shows the relationship between 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. Given below are associated diagnostic plots. Also given is a regression output for predicting teaching evaluation score from beauty score.

[24] *J. Felton et al. "Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness". In: Assessment & Evaluation in Higher Education 29.1 (2004), pp. 91-108.*

	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 satis ed for this model.

Contributors

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)