Chapter 3 Review Problems

These exercises are designed to help you review the important ideas and methods of the chapter. Relevant learning objectives are provided in bulleted form before each exercise.

- Identify explanatory and response variables in situations where one variable helps explain or influences another.
- Explain why association doesn't imply causation.

R3.1. The risks of obesity A study observes a large group of people over a 10-year period. The goal is to see if overweight and obese people are more likely to die during the study than people who weigh less. Such studies can be misleading, because obese people are more likely to be inactive and poor.

(a) What are the explanatory and response variables in the study?

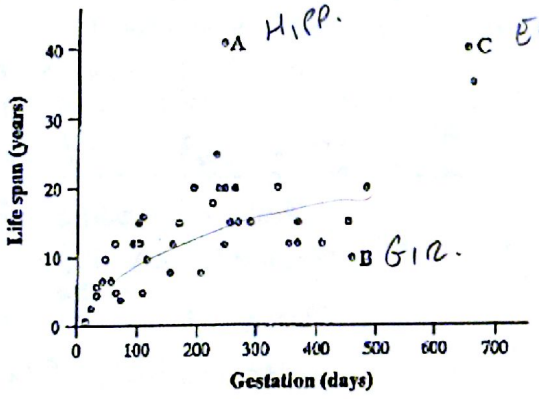
(b) If the study finds a strong association between these variables, can we conclude that increased weight causes greater risk of dying? Why or why not?

- Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.

(A) EXP = WT OF PERSON
RESP = MORTALITY RATE

(B) NO - OBESE PEOPLE TEND TO BE POOR, POOR HEALTH CARE MIGHT BE INVOLVED.

R3.2 Born to be old? Is there a relationship between the gestational period (time from conception to birth) of an animal and its average life span? The figure shows a scatterplot of the gestational period and average life span for 43 species of animals.[24]



(a) Describe the direction, form, and strength of the scatterplot.

(b) Three "unusual" points are labeled on the graph: Point A is for the hippopotamus, Point B is for the giraffe, and Point C is for the Asian elephant. In what way is each of these animals "unusual"?

(A) DIRECTION: POSITIVE.
FORM: SLIGHTLY CURVED
STRENGTH: MODERATE ASSOC.

(B) HIP - HIGHER LIFESPAN THAN OTHER ANIMALS W/COMP. GEST PERIODS
GIR - LOWER L/S THAN OTHERS W/COMP. GEST. PDS
ELE - HIGHER L/S THAN COMP. GEST. PDS

R3.3 Penguins diving A study of king penguins looked for a relationship between how deep the penguins dive to seek food and how long they stay under water.[25] For all but the shallowest dives, there is a linear relationship that is different for different penguins. The study gives a scatterplot for one penguin titled "The Relation of Dive Duration (y) to Depth (x)." Duration y is measured in minutes and depth x is in meters. The report then says, "The regression equation for this bird is: $\hat{y} = 2.69 + 0.0138x$."

(a) What is the slope of the regression line? Explain in specific language what this value says about this penguin's dives.

(b) According to the regression line, how long does a typical dive to a depth of 200 meters last?

(c) Does the y intercept of the regression line make any sense? If so, interpret it. If not, explain why not.

(A) SLOPE = 0.0138
FOR EACH ADD'L METER OF DEPTH PREDICTED CHANGE IN TIME IS AN INCREASE OF 0.0138 MINUTES.

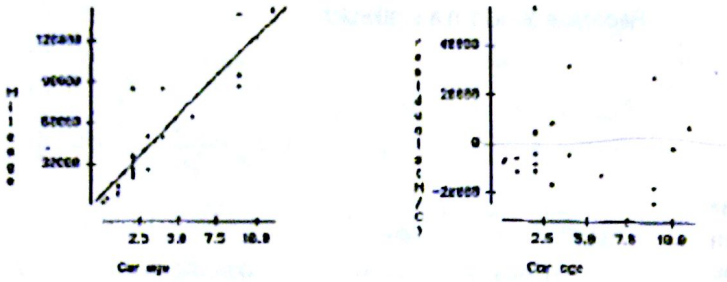
(B) $\hat{y}(200) = 2.69 + 0.0138(200)$
≈ 5.45 MINUTES

(C) A DIVE OF 0m WOULD LAST 2.69 MINUTES - DOES NOT MAKE SENSE.

R3.4 Stats teachers' cars A random sample of AP Statistics teachers was asked to report the age (in years) and mileage of their primary vehicles. A scatterplot of the data, a least-squares regression printout, and a residual plot are provided below.

R-squared = 82.0% R-squared (adjusted) = 81.1%
s = 19280 with 21 - 2 = 19 degrees of freedom

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	7288.54	6591	1.11	<0.2826
Car age	11630.6	1249	9.31	<0.0001



(a) Give the equation of the least-squares regression line for these data. Identify any variables you use.

(b) One teacher reported that her 6-year-old car had 65,000 miles on it. Find its residual.

(c) Interpret the slope of the line in context.

(d) What's the correlation between car age and mileage? Interpret this value in context.

(e) How well does the regression line fit the data? Justify your answer using the residual plot and s.

(A) $\hat{y} = 7288.54 + 11{,}630.6x$

x = CAR AGE (YRS)

y = MILEAGE

(B) $(6, 65000)$

$\hat{y}(6) = 77072.14$

$65{,}000 - 77072.14 = -12072.14$ MILES

(C) FOR EACH ADD'L YEAR THE PREDICTED MILEAGE SHOULD ↑ BY 11,630.6 MILES.

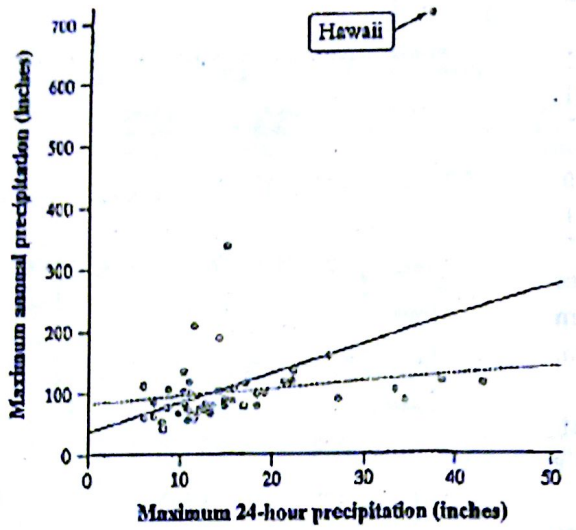
(D) $r = +\sqrt{.82} \approx 0.906$

THIS SHOWS A STRONG LINEAR RELATIONSHIP BETWEEN AGE + MILEAGE.

(E) THE LINE FITS REASONABLY WELL. THE RESIDUAL PLOT SHOWS NO PATTERN. THE TYPICAL RESIDUAL WHEN USING THE PREDICTOR MODEL IS 19,280 MI.

THIS SUGGESTS THAT, ALTHOUGH THE RESIDUAL PLOT IS ACCEPTABLE, THE MODEL'S PREDICTIONS WILL ON AVG BE IN ERROR BY ALMOST 20K MILES WHICH WOULD NOT BE USEFUL.

R3.7 When it rains, it pours The figure below plots the record-high yearly precipitation in each state against that state's record-high 24-hour precipitation. Hawaii is a high outlier, with a record-high yearly record of 704.83 inches of rain recorded at Kukui in 1982.



(a) The correlation for all 50 states in the figure is 0.408. If we leave out Hawaii, would the correlation increase, decrease, or stay the same? Explain.

(b) Two least-squares lines are shown on the graph. One was calculated using all 50 states, and the other omits Hawaii. Which line is which? Explain.

(c) Explain how each of the following would affect the correlation, s, and the least-squares line:
- Measuring record precipitation in feet instead of inches for both variables
- Switching the explanatory and response variables

(A) CORRELATION WOULD DECREASE BECAUSE HI IS ABOVE-AVG FOR BOTH MAX-24HR + MAX-ANNUAL.

(B) THE STEEPER LINE IS WITH ALL 50 STATES + THE OTHER IS W/O HI. HI IS AN INFLUENTIAL POINT + PULLS THE LINE TO IT.

(C) FEET VS INCHES

CORRELATION WILL NOT CHANGE SINCE IT HAS NO UNITS

s WILL DECREASE BECAUSE IT IS NOW IN FEET NOT INCHES.

LEAST SQUARES LINE - SLOPE WOULD NOT Δ BC Y-INT WOULD BECAUSE OF CONVERSION OF UNITS.

SWITCHING EXP/RES VARS

CORRELATION WILL NOT Δ BUT STD ERROR AND L-S LINE WILL.

R3.5 Late bloomers? Japanese cherry trees tend to blossom early when spring weather is warm and later when spring weather is cool. Here are some data on the average March temperature (in °C) and the day in April when the first cherry blossom appeared over a 24-year period:[26]

Temperature (°C) to first bloom:	4.0	5.4	3.2	2.6	4.2	4.7	4.9	4.0	4.9	3.8	4.0	5.1
Days in April:	14	8	11	19	14	14	14	21	9	14	13	11

Temperature (°C) to first bloom:	4.3	1.5	3.7	3.8	4.5	4.1	6.1	6.2	5.1	5.0	4.6	4.0
Days in April:	13	28	17	19	10	17	3	3	11	6	9	11

(a) Make a well-labeled scatterplot that's suitable for predicting when the cherry trees will bloom from the temperature. Describe the direction, form, and strength of the relationship.

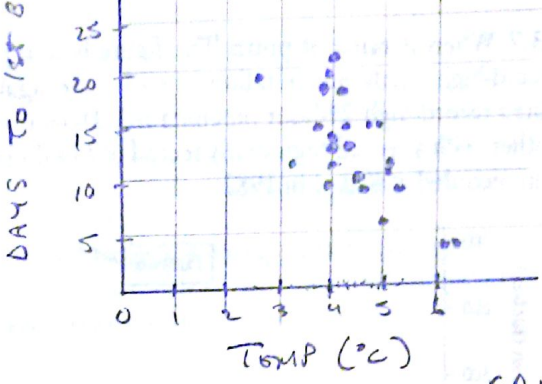
(b) Use technology to find the equation of the least-squares regression line. Interpret the slope and y intercept of the line in this setting.

(c) The average March temperature this year was 3.5°C. When would you predict that the first cherry blossom would appear? Show your method clearly.

(d) Find the residual for the year when the average March temperature was 4.5°C. Show your work.

(e) Use technology to construct a residual plot. Describe what you see.

(f) Find and interpret the value of r^2 and s in this setting.



(A) NEGATIVE LINEAR FAIRLY STRONG

(B) $\hat{y} = 33.12 - 4.69x$

y = # DAYS x = TEMP

SLOPE: FOR AN INCREASE IN 1°C THE MODEL PREDICTS A ↓ OF 4.69 DAYS TO 1ST BLOOM

Y-INT: OUTSIDE OF RANGE OF DATA SO NO MEANING

(C) $\hat{y} = 33.12 - 4.69(3.5)$
$= 16.7$ DAYS ⇒ APRIL 17TH

(D) $(4.5, 10)$
$\hat{y}(4.5) = 12.015$
$y - \hat{y} = 10 - 12.015 = -2.015$

(E) $r^2 = 0.724$ ⇒ 72% OF VAR IN DAYS TO BLOOM IS EXP. BY LIN. REL.
$s = 3.02$ ⇒ ON AVG, LIN MODEL HAS AN ERROR OF APPROX 3 DAYS.

(F) RES.D
NO PATTERN ⇒ GOOD.
TEMP °C

R3.6 What's my grade? In Professor Friedman's economics course, the correlation between the students' total scores prior to the final examination and their final-examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final-exam score from her pre-exam total.

(a) Find the equation for the appropriate least-squares regression line for Professor Friedman's prediction. Interpret the slope of this line in context.

(b) Use the regression line to predict Julie's final-exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Determine r^2. Use this result to argue that her actual score could have been much higher (or much lower) than the predicted value.

$r = 0.6$
$\bar{x} = 280$ $s_x = 30$
$\bar{y} = 75$ $s_y = 8$

(A) $b_1 = (0.6)\frac{8}{30} = 0.16$
$b_0 = 75 - 0.16(280) = 30.2$
$\hat{y} = 30.2 + 0.16(x)$
0.16 ⇒ FOR AN ↑ OF 1 PT ON PRE-EXAM TOTAL, MODEL PREDICTS AN ↑ OF 0.16 PTS.

(B) $\hat{y} = 30.2 + 0.16(300) = \boxed{78.2}$

(C) $r^2 = (0.6)^2 = 0.36$
SO ONLY 36% OF VARIATION IS ACCOUNTED FOR BY THE LIN. RELATIONSHIP. ∴ NOT GOOD EST.