Section 9.2

Exercise 1. (a) Let's go back to the hours spent online and the test score data from the last handout. Redraw the scatter plot. Again, here is the data.

Hours spent online, x	0	1	2	3	5	5	6	7	10
Test score, y	96	85	82	74	68	76	58	65	50

.

The above scatter plot indicates a linear relationship between hours spent online and test score. A *regression line* summarizes the relationship between an explanatory variable and a response variable. (In this example, the explanatory variable is the hours spent online and the response variable is the test score.)

Definition: The **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. Let $\hat{y}$ denote the predicted value of y. The equation of the regression line has the form

$$\hat{y} = a + bx.$$

The regression line generally used by statisticians is the line that makes the sum of the squares of the vertical distances of the data points from the line as **small as possible**. Before finding this line, let's look at *another* predicting line that comes close.

(b) Take the line $\hat{y}$ = -4x + 92. Draw the line on the scatterplot in part (a) and find the sum of the squares of the vertical distances of the data points from the line.

The technical term for the vertical distances is the residuals.

Definition: A **residual** is the difference between an observed value of the response variable and the value predicted by the predicting line. That is,

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \end{aligned}$$

(The first predicted value in the following table is found by substituting 0 for x.)

Hours spent online, x	0	1	2	3	5	5	6	7	10
Test score, y	96	85	82	74	68	76	58	65	50
$\hat{y}$									
Residual									
Square of residual									

The sum of the squares of the residuals =

The predicting line in (b) makes the sum of the squares of the residuals small, but there are many lines that do an even better job. The best line is the least squares line.

Definition: The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

A regression line has the form

$$\hat{y} = a + bx.$$

It turns out that for the least squares line,

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \text{ and } a = \bar{y} \text{ - } b\bar{x}.$$

Using the formula above, the least squares regression line is

(c) Find the sum of the squares of the vertical distances of the data points from the least-squares regression line.

Hours spent online, x	0	1	2	3	5	5	6	7	10
Test score, y	96	85	82	74	68	76	58	65	50
$\hat{y}$									
Residual	4.91	-1.84	-0.587	-4.34	-1.83	6.17	-7.58	3.67	1.43
Square of residual	24.10	3.38	0.34	18.79	3.35	38.06	57.44	13.49	2.04

The sum of the squares of the distances =

Note that 160.99 is smaller than the sum found in part (b).

(d) Would it make sense to use the least squares line to predict test scores for different amounts of hours spent online? Why or why not?

(e) Use the least squares line to predict the test score for a student who spends 8 hours online.

Substituting into the equation $\hat{y}$ = -4.252x + 91.091 yields:

$\hat{y}$ =

According to the least squares line, a student who spends 8 hours online during the weekend will score a .

(f) Use the least squares line to predict the test score for a student who spends twenty-six hours online.

Substituting into the equation $\hat{y}$ = -4.252x + 91.091 yields:

$\hat{y}$ =

(g) Is the prediction in part (f) believable?

<u>**Definition**</u>: Predicting $\hat{y}$ values for x values that are **beyond** observed x values in the data set is called **extrapolation**. Extrapolation may produce unrealistic forecasts.

<u>**Definition**</u>: Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation**.

(h) Suppose we computed a least squares line for the donuts and free throws data. Would it make sense to use that line to make predictions? Why or why not?

It to use the least squares line. The data to follow a linear pattern.

Class Exercise 1. Here is the budget (in millions of dollars) and worldwide gross (in millions of dollars) for eight of the most expensive movies ever made.

Budget, x	207	204	200	200	180	175	175	170
Gross, y	553	391	1835	784	749	218	255	433

(a) Find the least squares line for the data. (A smaller number of data points doesn't prevent us from calculating a least squares line. Since the number of points is so small, additional points will have a big impact on the line.) **Answer:** $\hat{y} = -2101.498 + 14.580x$
(b) Based on the correlation, does it make sense to make a prediction using the least squares line?
(c) According to the least squares line, what will be the gross of a movie whose budget is 1 dollar?
(d) Comment on your answer in part (c).

Class Exercise 2. The number of crimes reported (in millions) and the number of arrests reported (in millions) by the U.S. Department of Justice for 14 years.

Crimes, x	1.66	1.65	1.60	1.55	1.44	1.40	1.32
Arrests, y	0.72	0.72	0.78	0.80	0.73	0.72	0.68

Crimes, x	1.23	1.22	1.23	1.22	1.18	1.16	1.19
Arrest, y	0.64	0.63	0.63	0.62	0.60	0.59	0.60

(a) Find the least squares line for the data. **Answer:** $\hat{y} = .226 + .330x$
(b) Does it make sense to use the least squares line to make predictions?

If r is the sample correlation coefficient, then it can be shown that

$$r^2 = \frac{\Sigma(\hat{y}-y)^2}{\Sigma(y-\bar{y})^2}$$

r^2 is called the *coefficient of determination.*

Exercise 2. Let x be the age of a licensed driver in years. Let y be the percentage of all fatal accidents (for a given age) due to failure to yield the right-of-way. For example, the first data pair states that 5% of all fatal accidents of 37-year-olds are due to failure to yield the right-of-way. Here is the data:

x	37	47	57	67	77	87
y	5	8	10	16	30	43

(#12) (a) Display the scatter diagram displaying the data.

(b) Verify that $\Sigma x = 372$, $\Sigma y = 112$, $\Sigma x^2 = 24814$, $\Sigma y^2 = 3194$, $\Sigma xy = 8254$, and $r \approx 0.943$.
(c) Find $\bar{x}$, $\bar{y}$, a, and b. Then find the equation of the least-squares line $\hat{y} = a + bx$.

(d) Graph the least-squares line on your scatter diagram. Be sure to use the point $(\bar{x}, \bar{y})$ as one of the points on the line.
(e) Find the value of the coefficient of determination r^2. What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?

(f) Predict the percentage of all fatal accidents due to failing to yield the right-of-way for 70-year-olds.

Class Exercise 3. You are the foreman of the Bar-S cattle ranch in Colorado. A neighboring ranch has calves for sale, and you are going to buy some to add to the Bar-S herd. How much should a healthy calf weigh? Let x be the age of the calf (in weeks), and let y be the weight of the calf (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used by many ranchers).

x	1	3	10	16	26	36
y	42	50	75	100	150	200

(#8) (a) Display the scatter diagram displaying the data.
(b) Verify that $\Sigma x = 92$, $\Sigma y = 617$, $\Sigma x^2 = 2338$, $\Sigma y^2 =$ 82,389, $\Sigma xy =$ 13,642, and $r \approx 0.998$.
(c) Find $\bar{x}$, $\bar{y}$, a, and b. Then find the equation of the least-squares line $\hat{y} = a + bx$. **Answer:** $\bar{x}$ $=$ **15.33 weeks,** $\bar{y} =$ **102.83 kg;** $a \approx$ **33.696;** $b \approx$ **4.509;** $\hat{y} =$ **33.70** + $4.51x$
(d) Graph the least-squares line on your scatter diagram. Be sure to use the point $(\bar{x}, \bar{y})$ as one of the points on the line.
(e) Find the value of the coefficient of determination r^2. What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*? **Answer:** $r^2 \approx$ **0.995,** 99.5%, **0.5**%
(f) The calves you want to buy are 12 weeks old. What does the least-squares line predict for a healthy weight? **Answer: 87.8 kg**

Class Exercise 4. Data for this problem are based on information from *STATS Basketball Scoreboard.* It is thought that basketball teams that make too many fouls in a game tend to lose the game even if they otherwise play well. Let x be the number of fouls that were more than (i.e., over and above) the number of fouls made the opposing team made. Let y be the percentage of times the team with the larger number of fouls won the game. (#10)

x	0	2	5	6
y	50	45	33	26

(#10) (a) Display the scatter diagram displaying the data.
(b) Verify that $\Sigma x = 13$, $\Sigma y = 154$, $\Sigma x^2 = 65$, $\Sigma y^2 = 6290$, $\Sigma xy = 411$, and $r \approx$ -0.988.
(c) Find $\bar{x}$, $\bar{y}$, a, and b. Then find the equation of the least-squares line $\hat{y} = a + bx$. **Answer:** $\bar{x}$ $=$ **3.25,** $\bar{y} =$ **38.5,** $a =$ **51.286,** $b = -3.934$, $\hat{y} =$ **51.29** $-$ $3.934x$
(d) Graph the least-squares line on your scatter diagram. Be sure to use the point $(\bar{x}, \bar{y})$ as one of the points on the line.
(e) Find the value of the coefficient of determination r^2. What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*? **Answer:** $r^2 =$ **0.975,** 97.5%, 2.5%
(f) If a team had $x = 4$ fouls over and above the opposing team, what does the least-squares equation forecast for y? **Answer:** 35.55%

Homework

C Problems

Section 9.2: 7-17 ODD

B Problems

None

A Problems

None