CHAPTER 15
Describing Relationships:
Regression, Prediction, & Causation
THOUGHT QUESTION

Studies have shown a negative correlation between the amount of food consumed that is rich in beta carotene and the incidence of lung cancer in adults. Does this correlation provide evidence that beta carotene is a contributing factor in the prevention of lung cancer? Explain.
THOUGHT QUESTION

From past natural disasters, a strong positive correlation has been found between the amount of aid sent and the number of deaths. Would you interpret this to mean that sending more aid causes more people to die? Explain.
LINEAR REGRESSION

- Objective: To *quantify* the linear relationship between an explanatory variable and a response variable.

  We can then *predict* the average response for all subjects with a given value of the explanatory variable.
**Regression Lines**

- If we notice that a scatterplot shows a straight-line relationship between two quantitative variables, we might want to summarize this overall pattern by drawing a line on the graph.
- A **regression line** is a straight line that describes how a response variable $y$ changes as the explanatory variable $x$ changes.
- A regression line is often used to predict the value of $y$ for a given value of $x$. 
REGRESSION EQUATIONS

- When we have a scatterplot with a strong relationship, it’s relatively easy to draw a line close to the points.

- However, as we get more scatter on the scatterplot, this leads to the possibility of different people drawing different lines when basing it on eyesight alone.

- So we use the computer to help us draw a regression line. The computer will ensure that we end up with the best possible line.
REGRESSION EQUATIONS

- We want a line that is close to the points in the vertical (y) direction since we want to predict y. The most common way to do this is known as the least-squares method.

- 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.

http://hadm.sph.sc.edu/Courses/J716/demos/LeastSquares/LeastSquaresDemo.html
Example 15.1

Swimming time vs. Pulse

Pulse rate (BPM)

Swimming time (minutes)
EXAMPLE 15.1

Swimming time vs. Pulse

Pulse rate (BPM)

Swimming time (minutes)

\[ y = -12.884x + 589 \]

Least-squares regression line

Least-squares regression equation
Recall that the equation of a line has the form \( y=mx+b \). Here, \( x \) stands for the explanatory variable and \( y \) stands for the response variable.

- The number \( m \) is the slope of the line, which is the amount by which \( y \) changes when \( x \) increases by one unit.
- The number \( b \) is the y-intercept, or the value of \( y \) when \( x=0 \).
- To use the equation for prediction, just substitute your x-value into the equation and calculate the resulting y-value.
Example 15.1 (Continued)

- Professor Moore swims 2000 yards regularly in an attempt to undo middle age. We looked at the scatterplot displaying the relationship between the swimming time and his pulse rate afterwards. From the computer, the least-squares regression line is:
  
  \[ \text{pulse rate} = (-12.884 \times \text{swimming time}) + 589 \]

- Say that tomorrow the professor takes 35 minutes to swim 2000 yards. Use the equation to predict his pulse rate.
UNDERSTANDING PREDICTION

- Computers make predictions very easy – which leads to carelessness on the part of the humans using the computer.
- For example, the computer does not decide which is the explanatory variable and which is the response variable. If we put the variables in the wrong place, we would end up with completely different predictions.
- Prediction outside the range of the available data is risky.
A Caution

- Sarah’s height was plotted against her age.
- Can you predict her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?
A CAUTION

- Regression line:
  \[ y = 0.383x + 71.95 \]
- height at age 42 months?
  \[ y = 88 \text{ cm.} \]
- height at age 30 years?
  \[ y = 209.8 \text{ cm.} \]
  - She is predicted to be 6'10.5" at age 30.
Correlation and regression are closely connected.
Both correlation and regression are strongly affected by outliers.
If the correlation is positive, then the slope of the regression line will be _______.
If the correlation is negative, then the slope of the regression line will be _______.

Correlation and regression are strongly affected by outliers.
r Squared

- Measures usefulness of regression prediction
- $r^2$ (the square of the correlation): measures the percentage of the variation in the values of the response variable ($y$) that is explained by the regression line
  - $r=1$, $r^2=1=100%$
    - regression line explains all (100%) of the variation in $y$
  - $r=0.7$, $r^2=0.49=49%$
    - regression line explains almost half (49%) of the variation in $y$
EXAMPLE 15.1 (CONTINUED)

- From example 15.1, suppose the $r = -0.940$. How much of the variation in pulse rates would the least-squares regression line explain?
  - The regression line explains 88.36% of the variation in pulse rates.
EXAMPLE 15.2

- Below is a scatterplot showing the relationship between payroll and the number of victories for the 32 teams in the NFL.

![Scatterplot showing the relationship between payroll and wins for NFL teams](image)

The equation of the best fit line is:

\[ y = 0.0613x + 1.8599 \]
EXAMPLE 15.2

- Suppose a team has a payroll of 103 (million). Predict the number of wins.

- Suppose the correlation between the payroll and the number of wins turns out to be $r=0.276$. How much of the variation in number of wins would the least-squares regression line explain?
THE QUESTION OF CAUSATION

- A strong relationship between two variables does not always mean that changes in one variable causes changes in the other.
- The relationship between two variables is often influenced by other variables which are lurking in the background.
- The best evidence for causation comes from randomized comparative experiments.
ESTABLISHING CAUSATION WITHOUT A RANDOMIZED COMPARATIVE EXPERIMENT

- If all five of the following properties exist, we can usually make a good case for cause-and-effect, even without an experiment:
  1. The relationship is strong (usually $r$ greater than 0.9).
  2. The relationship is consistent (having a strong relationship between the same two variables in study after study on different people).
  3. Higher values in our explanatory response are associated with stronger responses.
  4. The alleged cause precedes the effect in time.
  5. The alleged cause make sense.
REMINDERS

- Chapter 15 questions are just suggested problems. You don’t need to turn them in.
- Please turn in Chapter 14 homework
- Exam 2 is tomorrow