Review: Week 3

Read: Chapters 7-10 in the text book

Review: Scatter plots, Association and Correlation

When we think that changes in a variable $x$ explain, or maybe even cause, changes in a second variable $y$, we call $x$ an explanatory variable and $y$ a response variable.

A scatter plot is an efficient way to display the relationship between two quantitative variables. Look at the overall pattern of the plot. In particular study the direction, form, strength and look for outliers. If the plot appears to display a straight line, this indicates a linear relationship between the two variables.

The correlation is a measure of strength and direction of a linear relationship between two quantitative variables.

Suppose we have $n$ observations of data on two variables $x$ and $y$. Further suppose that the mean and standard deviation for $x$ is $\bar{x}$ and $s_x$, and for $y$ they are $\bar{y}$ and $s_y$. Then, the correlation between $x$ and $y$ is given by,

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right) \left(\frac{y_i - \bar{y}}{s_y}\right).$$

The correlation is always between -1 and 1. If $r > 0$, this indicates a positive association. When $r = 1$, this indicates that all of the points lie on a straight line with a positive slope. If $r < 0$, this indicates a negative association. When $r = -1$, this indicates all points lie on a straight line with negative slope. If $r$ is close to 0, this indicates a very weak linear relationship.

Note that the correlation only measures the strength of a linear relationship. There may be other types of association between two variables, but these patterns will not be captured by the correlation.

High correlation does not imply causation.

A lurking variable has an important effect on the relationship among the variables in a study but is not included among the variables studied.
Exercises:

Exercise 1: Scatter plot

What would you expect scatter plots of the data described below look like? In particular mention the direction, form, strength and potential outliers.

(a) The height of a person (in inches) is the explanatory variable and the weight of the same person (in lbs) is the response variable.
(b) The weight of a person (in lbs) is the explanatory variable and the weight of the same person (in kg) is the response variable.
(c) Shoe size is the explanatory variable and GPA is the response variable.
(d) Altitude is the explanatory variable and temperature is the response variable.

Exercise 2: Correlation

In this exercise we will use the following data set:

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

(a) Make a scatter plot for this data.
(b) What is the correlation between $x$ and $y$?

Exercise 3: A study shows that there is a positive correlation between the number of beds in a hospital and the median number of days that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital?
Scatter plot and correlation drill

Draw a scatter plot of the variables $x$ and $y$ so that:

- $x$ and $y$ have positive association
- $x$ and $y$ have negative association
- The correlation between $x$ and $y$ is 0
- The correlation between $x$ and $y$ is 0.5
- The correlation between $x$ and $y$ is -0.5
- The correlation between $x$ and $y$ is 0.2
- The correlation between $x$ and $y$ is -0.2
- The correlation between $x$ and $y$ is 0.9
- The correlation between $x$ and $y$ is -0.9
- The correlation between $x$ and $y$ is 1.0
- The correlation between $x$ and $y$ is -1.0
- The correlation between $x$ and $y$ is 1.5
**Review:** Linear Regression

A regression line is a straight line that describes how a response variable $y$ changes with the explanatory variable $x$. The least squares regression line is the regression line that minimizes the sum of the squared vertical distance from the $y$ data points to the line.

Suppose we have $n$ observations on two variables $x$ and $y$. Further suppose that the mean and standard deviation for $x$ is $\bar{x}$ and $s_x$, and for $y$ they are $\bar{y}$ and $s_y$, and the correlation between the two variables is $r$. The equation of the least-squares regression of $y$ on $x$ is $\hat{y} = a + bx$ with slope $b = r \frac{s_y}{s_x}$ and intercept $a = \bar{y} - b\bar{x}$.

The square of the correlation, $r^2$, represents the fraction of the variation in the values of $y$ that is explained by the least squares line of $y$ on $x$.

A residual is the difference between the observed value of $y$ (the actual data point) and the predicted value of $y$ (as determined by the regression line). A residual plot is a scatterplot of the residuals against the explanatory variables. When the regression line picks up the overall pattern of the data, the residual plot should have no particular pattern.

An influential point is an observation that, if removed, would considerably change the position of the regression line. Influential points are often outliers in the horizontal direction.

**Exercises:**

**Exercise 1:** Linear Regression

A company that manufactures building material measures the length and weight of 10 wooden boards. The length and weight of board $i$ is given by $x_i$ and $y_i$, respectively, where $i = 1, \ldots, 10$. We are given the following summary information about the data:

$$\bar{x} = 73, \quad s_x = \sqrt{\frac{892}{9}}, \quad \bar{y} = 16, \quad s_y = \sqrt{\frac{34}{9}}, \quad \text{and} \quad r = 0.89$$

(a) Calculate the least square regression line $\hat{y} = a + bx$.
(b) Calculate the proportion of variability in the weight of the boards that is explained by the regression line.
(c) Suppose we find a new board that is 60 inches long. Predict its weight.
Exercise 2: Six observations were measured on two variables $x$ and $y$. Fig. 1 shows a scatter plot of the 6 observations. However it turns out that one of the observations was recorded incorrectly. The observation at the point $(7, -5)$ should instead read $(7, 5)$. Fig. 2 shows the correct scatterplot.

![Fig. 1](image1)

![Fig. 2](image2)

(a) For the data set depicted in Fig. 1 we calculated the following summary statistics:

$$
\bar{x} = 3.550, \ s_x = 1.805, \ \bar{y} = 0.233, \ s_y = 2.617 \text{ and } r = -0.935
$$

Calculate the least-square regression line and plot it in Fig. 1.

(b) For the data set depicted in Fig. 2 we calculated the following summary statistics:

$$
\bar{x} = 3.550, \ s_x = 1.805, \ \bar{y} = 1.900, \ s_y = 1.607 \text{ and } r = 0.855
$$

Calculate the least-square regression line and plot it in Fig. 2.

(c) Why do the results in (a) and (b) differ so much?
**Linear regression drill:**

Fit the regression line through the cloud of points.

For each of the plots below, draw the best-fitting line for predicting $y$ from $x$. 