Section 2.3: Least Squares Regression

A Royal Bengal tiger wandered out of a reserve forest. We tranquilized him and want to take him back to the forest. We need an idea of his weight, but have no scale!

Studies have shown that Royal Bengal tigers in this forest weigh, on average 200 lbs with SD 60 lbs.

Q: Our best guess of his weight is then?

We hear that the doctor who prepared the tranquilizer is 6” above average in height. Now our best guess for the tiger’s weight is?

Now we hear that studies have also found means and SD’s of some other variables on these tigers and r between those variables.

r for first claw length & weight is .20
r for neck girth & weight is .80

Which would be more helpful?

Idea: If there is linear relationship (given by r):
- We can use it to predict what y might be if we know x
- The stronger the correlation, the more confident we are in our estimate of y for a given x.

Goal: Fit a line to the scatterplot that can be used to predict y for a given x.
Recall the equation of a line:

In our setting, $x$ is the explanatory variable being used to predict the response variable $y$. In this equation, $b$ is called the ________, the amount $y$ changes on average when $x$ increases by one unit.

$a$ is the ________, the value of $y$ when $x = 0$.

A regression line is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes.

Q: How does the sign (+ or -) of the correlation relate to the sign of the slope $b$?

Idea: If we know how to land on this line, we can go directly to it to predict $y$.

Issue: We can draw lot of lines through a scatter plot depending on what values of “a” and “b” we use. Which line is the best in the sense that it predicts $y$ for a given $x$ most accurately?

Notation:

We define Residuals =
The best line is the one that minimizes the distance between predicted $y$ and observed $y$ in some sense, i.e.,

The method of least squares does this by selecting those values of $a, b$ that minimizes the sum of the squares of the residuals.

The line obtained by this method is called the **Least-Squares Regression Line**.

Computing the slope $b$ and intercept $a$ of the least-squares regression line:

1. Use computer software (e.g. Excel, Minitab) OR
2. Use summary statistics of the data.

Let $\bar{x}$ and $\bar{y}$ be the means of the explanatory variable $x$ and response variable $y$, respectively, and $s_x$ and $s_y$ be the corresponding standard deviations. Suppose the correlation between $x$ and $y$ is denoted by $r$.

The slope is calculated as:

The intercept is:

The least squares regression line is then given by:

Fact: The residuals always add up to zero (due to the math of regression)
Example: Height Data
Phyllis wonders if tall women tend to date tall men and vice versa. She measures herself, her roommate and some other friends. Then she obtains the heights of the next man each woman dates. Below is a scatterplot of the data.

The summary statistics are:

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>65.5</td>
<td>69</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1.05</td>
<td>2.53</td>
</tr>
<tr>
<td>Correlation (r)</td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>

Q: Calculate the least-squares regression line for predicting the date’s height for a given height of woman and plot it on the graph above.

Q: Predict the date’s height given the woman’s height is 67”.

Q: What is the residual for the observation where the woman’s height was 67”?

Q: What is the average of all residuals?
Some facts about Least-Squares Regression:

1. The distinction between the explanatory and response variable is important! Different regression line will result if we switch $x$ and $y$.

2. How successful is the regression line in explaining the response?

   The square of the correlation, $r^2$, is the proportion of the variability of the $y$ values explained by the least-squares regression of $y$ on $x$.

Q: Suppose we have a perfect linear relationship. What is $\hat{y}$? What is $r^2$?

Q: Suppose we have absolutely no linear relationship. What is $\hat{y}$? What is $r^2$?

**Relationship between $r$ and $r^2$**

Q: The correlation between the heights of the men and women in the “height data” is $r = 0.69$. What is $r^2$?

Q: Suppose we have the regression equation $\hat{y} = 0.3 - 1.2x$ and $r^2 = 0.25$. What is $r$?
Outliers and Influential Observations

Outlier:

Influential Observation:

Example: Height Data

Notice the effect on the regression calculations of adding a woman of height 55 inches to the data. This woman is an influential observation.

Typically, points that are outliers in the x direction are influential for the least-squares regression line.