Chapter 5 – 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 standard deviation (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 the correlation coefficient $r$ 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>1.0490</td>
<td>2.5300</td>
</tr>
<tr>
<td>Correlation (r)</td>
<td>0.6784</td>
<td></td>
</tr>
</tbody>
</table>

![Scatterplot of height data](chart.png)

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?
Supplementary Note
Calculation of Least Squares Regression Line

**Situation 1:** The original dataset is not available, but \( \bar{x}, \bar{y}, s_x, s_y \) and \( r \) are available. Use the book formulae:

\[
b = r \frac{S_y}{S_x}, \quad a = \bar{y} - b \bar{x}
\]

where
\( \bar{x} \) is the mean of the x-variable (explanatory variable),
\( \bar{y} \) is the mean of the y-variable (response variable),
\( s_x \) is the standard deviation of the x-variable (explanatory variable),
\( s_y \) is the standard deviation of the y-variable (response variable).

**Situation 2:** The original dataset is available. Use the “Table with 7 columns” and either

(i) obtain \( S_{xx}, S_{yy}, S_{xy} \) from the table, then use the formulae in the class notes:

\[
b = \frac{S_{xy}}{S_{xx}}, \quad a = \bar{y} - b \bar{x}
\]

or

(ii) obtain \( \bar{x}, \bar{y}, s_x, s_y \) and \( r \) from the table, then use the book formulae:

\[
b = r \frac{S_y}{S_x}, \quad a = \bar{y} - b \bar{x}
\]
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. The least-squares regression line always passes through the point

3. 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.
Cautions about Correlation and Regression

Correlation and regression are very powerful tools for describing the relationship between two variables. But we need to keep some cautions in mind while using them so that we do not misuse them.

- Correlation and least-squares regression line are not resistant to outliers.
- They only describe __________ patterns.
1. Extrapolation

Extrapolation is

Ex: Farming Population

The number of people living on American farms has declined during this century. Below is a plot of farm population (millions of persons) from 1935 to 1980.

Use the regression equation to predict the number of people living on farms in 1990. Comment on the result.

→ Never use regression line to predict
2. Using Averaged Data

Ex: Year 2007 Stat2331 test results

One lecture with 4 lab sections

<table>
<thead>
<tr>
<th>Section</th>
<th>X</th>
<th>Average (X)</th>
<th>Y</th>
<th>Average (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>57,51,48,53,52</td>
<td>52.2</td>
<td>92,81,78,82,81</td>
<td>82.8</td>
</tr>
<tr>
<td>2</td>
<td>62,46,59,52,55</td>
<td>54.8</td>
<td>77,84,95,79,86</td>
<td>84.2</td>
</tr>
<tr>
<td>3</td>
<td>58,51,52,49,47</td>
<td>51.4</td>
<td>95,70,82,78,80</td>
<td>81.0</td>
</tr>
<tr>
<td>4</td>
<td>70,41,64,46,54</td>
<td>55.0</td>
<td>69,89,101,78,91</td>
<td>85.6</td>
</tr>
</tbody>
</table>

For each section let
X = midterm score (out of 80)
Y = final score (out of 120)

If we plot: (i) the averages

Scatter plot of averages
(ii) the individual scores

$r$ for individual scores was

$r$ for section averages was

$r$ based on averages __________ the strength of $r$ compared to what it would be for the individual values (by not accounting for variation in individual values).

→ Beware of them

3. Lurking Variables

Relationship between two variables often is greatly influenced by other variables.

A lurking variable is
Example: There is a negative correlation between the number of flu cases reported each week through the year and the amount of ice cream sold that week. What is a plausible lurking variable that may account for this association?

When analyzing relationships between variables, we need to be careful about claiming that changes in the explanatory variable cause changes in the response variable.

4. High correlation **DOES NOT** imply causation.

Ex: A study of elementary school children, ages 6 to 11, finds a high positive correlation between shoe size and scores on an IQ test. Does it make sense to claim that bigger shoe sizes cause better IQ? What explains this correlation?

Remarks:
- This does not mean we cannot use a regression line. For example, we have a regression of scores of an IQ test on shoe size; we can still predict a kid’s IQ score from his/her shoe size, even if “shoe size” isn’t causing “IQ”.

The best evidence that an association is due to causation comes from an experiment in which the explanatory variable is directly changed and other influences on the response are controlled.