L3 - 3.2 Least Squares Regression
Lesson 3 – 3.2 Least Squares Regression

Introduction to Least Squares Regression

RECAP:

$r = \textbf{Correlation:}$ A measure of the direction and strength of the linear relationship between the two variables (variables must be quantitative)

$r^2 = \textbf{Correlation Coefficient:}$ Measures the strength of the linear association between two quantitative variables called $r$.

\[
 r = \frac{1}{n-1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right),
\]

the “average” of the product of the z-scores

(don’t need to use—get $r$ from calculator)

A correlation greater than 0.8 is generally described as \textit{strong}, whereas a correlation less than 0.5 is generally described as \textit{weak}. These values can vary based upon the "type" of data being examined and the context. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Always plot the data first to make sure it looks reasonably linear and there are no extreme outliers.
Least Squares Regression

Regression: used to make predictions about the data/situation.

Example #1 – Does Fidgeting Keep You Slim? (Pg 164 Tpse ed4)

Some people don’t gain weight when they overeat. Perhaps fidgeting and other ‘non exercise activity’ (NEA) explains why – some people may spontaneously increase non exercise activity when fed. Researchers deliberately over fed 16 healthy young adults for 8 weeks. They measured fat gain (in kg) as the response variable and change in energy use (in calories) from activity other than deliberate exercise – fidgeting, daily living and the like –as the explanatory variable.

Here is the data:

<table>
<thead>
<tr>
<th>NEA Change (kg)</th>
<th>-94</th>
<th>-57</th>
<th>-29</th>
<th>135</th>
<th>143</th>
<th>151</th>
<th>245</th>
<th>355</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat Gain (kg)</td>
<td>4.2</td>
<td>3.0</td>
<td>3.7</td>
<td>2.7</td>
<td>3.2</td>
<td>3.6</td>
<td>2.4</td>
<td>1.3</td>
</tr>
<tr>
<td>NEA Change (kg)</td>
<td>392</td>
<td>473</td>
<td>486</td>
<td>535</td>
<td>571</td>
<td>580</td>
<td>620</td>
<td>690</td>
</tr>
<tr>
<td>Fat Gain (kg)</td>
<td>3.8</td>
<td>1.7</td>
<td>1.6</td>
<td>2.2</td>
<td>2.0</td>
<td>0.4</td>
<td>2.3</td>
<td>1.1</td>
</tr>
</tbody>
</table>

This scatter plot shows fat gain over 8 weeks of overeating against increase in non exercise activity over the same period with a regression line added.

1. Describe the scatterplot: (be sure you mention direction, form and strength)

The plot shows a moderately strong, negative, linear relationship between NEA change and fat gain with no outliers.

Interpreting a Regression Line

Suppose that y is our response variable (on vertical axis) and x is the explanatory variable (on horizontal axis). A Regression Line relating to y and x has the equation:

\[ \hat{y} = a + bx \]

- \( \hat{y} \) (read “y hat”) is the **predicted value** of y for a given value of x
- b is the **slope**, the amount by which y is predicted to change when x increases
- a is the **y-intercept**, the predicted value of y when x=0
1. What is our regression line from the ‘Does Fidgeting Keep You Slim’ example?

\[ \hat{y} = 3.505 - 0.00344x \]

\[ \hat{y} = \text{fat gain} \quad \quad \quad x = \text{NEA change} \]

2. What do the slope and y-intercept represent?

Slope: \( b = -0.00344 \) tells us the amount of fat gain is predicted to go down

\( 0.0024 \text{kg} \) for each added calorie of NEA.

y-intercept: \( a = 3.505 \) is the fat gain estimated by the model if NEA doesn’t change when a person overeats.

**Question:** Can we say how important a relationship is simply by looking at the size of the slope?

No! Because we can use different units

\( \text{Kg} \rightarrow \text{g} \quad \text{S} \rightarrow \text{min.} \)

**Predicting with a Regression Line**

For the NEA and fat gain data, the equation of the regression line is:

\[ 2nd \rightarrow \text{calc} \rightarrow \text{value} \rightarrow x = 400 \]

1. If a person’s NEA increases by 400 calories when she overeats, what is the predicted fat gain?

\[ \text{fat gain} = 3.505 - 0.00344 (400) \]

\[ \hat{y} = 2.13 \text{kg} \]

**Extrapolation:** is the use of a regression line for prediction far outside the interval of values of explanatory variable \( x \) used to obtain the line. We need to be careful because these are often not accurate

- Eg: NEA of 1500 calories

**Residuals and the Least-Squares Regression Line**

*A good regression line makes the vertical distances of the points from the line as small as possible.*

**Residual:** the difference between an observed value of the response variable and value predicted by the regression line.

\[ \text{Residual} = \text{observed } y - \text{predicted } y \]

\[ = y - \hat{y} \]

1. Find and interpret the residual for the hiker who weighed 187 pounds.

(187 lb Hiker with 30 lb pack) pack weight = 16.3 + 0.0908 (hiker weight)

\[ = 33.28 \]

\[ \text{Residual} = y - \hat{y} \]

\[ = 30 - 33.28 \quad [= -3.28] \]

Homework: Pg. 158 #27–32 & Pg. 191# 37, 39, 41, 43, 45, 47, 53