Sect 2.3: Least–Squares Regression

If a linear relationship exists between two variables we would like to summarize (or model) this overall pattern by drawing a straight line on the scatterplot.

A regression line is a _____________________________________________________________________________________.

A regression line is typically used to __________ the value of ___ for a particular value of ___.

Recall the equation of a line:

\[ y = a + bx \]

where \( x \) 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 \( x = 0 \).

Example: Blood Alcohol Content

Q: Describe the relationship between number of beers consumed and BAC.

Q: What would you guess the correlation to be?

Q: How does the sign (+ or -) of the correlation relate to the sign of the slope \( b \)?
Suppose we are given the following regression line for the BAC data:

\[ \text{BAC} = -0.065 + \text{______________________________} \times (\text{# of beers}) \]

Draw this line on the scatterplot.

1) Choose \text{______________________________} \text{r # of beers}.

2) Predict the BAC for the \text{______________________________} chosen in step 1 by plugging these values into “\text{# of beers}” in the above equation and then calculating BAC.

3) Plot these points on the graph and draw a line through these two points.

We can use the regression line to \text{predict} the value of \text{y} for a given value of \text{x}.

Q: Using the line you drew on the graph, what would we predict a person’s BAC to be if they drank 8 beers?

Q: What was the observed BAC for the person who actually drank 8 beers?

- When the data points lie close to the line, we can be confident our prediction is accurate.
- If the points are fairly scattered about the line, we put less confidence in the predictions.

\textbf{Extrapolation} is using the \text{______________________________} to predict \text{___} values for values of \text{___} far outside the \text{______________________________}.

These predictions are often \text{______________________________}.

Example

For the age-height example, what does the LS line predict the height to be when the age is 30 years (360 months).

\[ Y = 64.9 + \text{__________}360 = \text{_____________} \text{cm} = \text{_______} \text{in=}\text{____feet} \]

Example: 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.

(a) According to the regression line, how much did the farm population decline each year on the average during this period?

(b) What percent of the observed variation in farm population is accounted for by linear change over time?

(c) Use the regression equation to predict the number of people living on farms in 1990. Is this reasonable?
Least-Squares Regression

Many lines can be drawn through the data and used for prediction. Which line is “best”? We want a line that yields the most accurate predictions so we want to minimize the vertical distance between the line and the actual data points.

The most common method of choosing the best line is the least-squares regression line of $y$ on $x$. This line makes ______________________________
__________________________________________________
__________________________________________________

Equation of the Least-Squares Regression Line

How do we compute the slope $b$ and intercept $a$ for the least-squares regression line?

1) Use computer software (e.g. Minitab) OR

2) Use summary statistics of the data.
Interpreting Minitab Output

MTB > Regress 'BAC' 1 'BEERS';
Regression Analysis: BAC versus BEERS

The regression equation is
\[ \text{BAC} = -0.0127 + 0.0180 \times \text{BEERS} \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.01270</td>
<td>0.01264</td>
<td>-1.00</td>
<td>0.332</td>
</tr>
<tr>
<td>BEERS</td>
<td>0.017964</td>
<td>0.002402</td>
<td>7.48</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ S = 0.02044 \quad \text{R-Sq = 80.0\%} \quad \text{R-Sq(adj) = 78.6\%} \]

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>0.023375</td>
<td>0.023375</td>
<td>55.94</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>14</td>
<td>0.005850</td>
<td>0.000418</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>0.029225</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>BEERS</th>
<th>BAC</th>
<th>Fit</th>
<th>SE Fit</th>
<th>Residual</th>
<th>St Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9.00</td>
<td>0.19000</td>
<td>0.14897</td>
<td>0.01128</td>
<td>0.04103</td>
<td>2.41R</td>
</tr>
</tbody>
</table>

R denotes an observation with a large standardized residual

Below the first table, the important value \[ \text{________________________________________} \]
\[ \text{____________________________________________________________________} \]

From reading the output, the correlation between BAC and beers is \[ \text{_________} \], and the percent of variation in BAC explained by the regression on beers is \[ \text{_________} \].

The slope and intercept for the regression line, are given directly, or can be found in the column labeled \[ \text{_________} \] in the first table. The value next to the word \[ \text{_________} \] is the intercept of the regression line (a). The value below that is the slope of the regression line (b).

The equation for this regression line is \[ \text{________________________________________} \]

Minitab can also give you plots of the data with the regression line and without it and residual plots.
Using 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. The correlation between \( x \) and \( y \) is denoted by \( r \).

The slope is calculated as: \( b = \).

The intercept as: \( a = \bar{y} - b\bar{x} \).

Then, \( \hat{y} = a + bx \) is the least squares regression line.

Notation:

* \( \hat{y} \) = predicted value
* \( y \) = observed data value

A residual is

\[
\text{residual} = \]

Example BAC Data

<table>
<thead>
<tr>
<th>Beers</th>
<th>BAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>4.81</td>
</tr>
<tr>
<td>standard deviation</td>
<td>2.20</td>
</tr>
<tr>
<td>correlation (r)</td>
<td></td>
</tr>
</tbody>
</table>

Q: Calculate the least-squares regression line for this data and plot it on the graph.

Q: Use the least-squares regression line to predict what the BAC would be for a person who drank 1 beer.

Q: Use the other regression line to predict what the BAC would be for a person who drank 1 beer. How do they compare?

Q: Use the least-squares regression line to predict what the BAC would be for a person who didn’t drink any beers. Does this make sense?

Q: What is the residual for the observation where the number of beers is 1.
Facts about Least-Squares Regression:

1) The distinction between the explanatory and response variable is important! Different regression lines will result.

2) The L.S. Regression line always goes through the point ________________ and you can write it:

   \[ \frac{y - \bar{y}}{s_y} = \]

3) The square of the correlation, \( r^2 \), is the __________________________ of the \( y \) values explained by the least-squares regression of \( y \) on \( x \).

   \[ r^2 = \frac{\text{variance of the } y \text{ values}}{\text{variance of the } y \text{ values}} = \frac{\text{var } \hat{y}}{\text{var } y} \]

   Thus, \( 0 \leq r^2 \leq 1 \) and the closer \( r^2 \) is to 1, the more linear the relationship between \( x \) and \( y \) and the better \( x \) predicts \( y \).

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 you were given the regression equation \( \hat{y} = 0.3 - 1.2x \) and that \( r^2 = 0.25 \). What is \( r \)?