10.3 Regression

Definition

Regression Equation

Given a collection of paired data, the regression equation
Regression

Definition

Regression Equation
Given a collection of paired data, the regression equation

\[ \hat{y} = b_0 + b_1 x \]

algebraically describes the relationship between the two variables.

Regression Line
(line of best fit or least-squares line)

The graph of the regression equation.
The Regression Equation

\[ y = \hat{y} = b_0 + b_1x \]

- \( x \) is the independent variable (predictor variable)
- \( y \) is the dependent variable (response variable)
The Regression Equation

- $x$ is the independent variable (predictor variable)
- $\hat{y}$ is the dependent variable (response variable)

\[
\hat{y} = b_0 + b_1 x
\]

Assumptions

1. We are investigating only linear relationships.
2. For each $x$ value, $y$ is a random variable having a normal (bell-shaped) distribution. All of these $y$ distributions have the same variance. Also, for a given value of $x$, the distribution of $y$-values has a mean that lies on the regression line. (Results are not seriously affected if departures from normal distributions and equal variances are not too extreme.)

Regression Line Plotted on Scatter Plot
Notation for Regression
Equation

<table>
<thead>
<tr>
<th>Population Parameter</th>
<th>Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>( b_0 )</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>( b_1 )</td>
</tr>
</tbody>
</table>

\( y = \beta_0 + \beta_1 x \)

Formula for \( b_1 \) and \( b_0 \)

\[
\begin{align*}
\text{Formula 10-2} \quad & b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \\
\text{(slope)}
\end{align*}
\]

\[
\begin{align*}
\text{Formula 10-3} \quad & b_0 = \bar{y} - b_1 \bar{x} \\
\text{(y-intercept)}
\end{align*}
\]

calculators or computers can compute these values
Rounding the y-intercept \( b_0 \) and the slope \( b_1 \)

- Round to three significant digits
- If you use the formulas 10-2, 10-3, try not to round intermediate values or carry to at least six significant digits.

Example: Lengths and Weights of Male Bears

<table>
<thead>
<tr>
<th>x Length (in.)</th>
<th>53.0</th>
<th>67.5</th>
<th>72.0</th>
<th>72.0</th>
<th>73.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>y Weight (lb)</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>262</td>
<td>360</td>
<td>332</td>
</tr>
</tbody>
</table>

- \( b_0 = -352 \) (rounded)
- \( b_1 = 9.66 \) (rounded)
Example: Lengths and Weights of Male Bears

- $x$ Length (in.) 53.0 67.5 72.0 72.0 73.5 73.0 37.0
- $y$ Weight (lb) 80 344 416 348 262 360 332 34

\[ b_0 = -352 \text{ (rounded)} \]
\[ b_1 = 9.66 \text{ (rounded)} \]
\[ y = -352 + 9.66x \]

Scatter Diagram of Paired Data

The regression line fits the sample points best.
Predictions

In predicting a value of $y$ based on some given value of $x$ ...

1. If there is not a significant linear correlation, the best predicted $y$-value is $\bar{y}$.

2. If there is a significant linear correlation, the best predicted $y$-value is found by substituting the $x$-value into the regression equation.

Predicting the Value of a Variable

1. Calculate the value of $r$ and test the hypothesis that $r = 0$.
2. Is there a significant linear correlation?
   - Yes: Use the regression equation to make predictions. Substitute the given value in the regression equation.
   - No: Given any value of one variable, the best predicted value of the other variable is its sample mean.
Guidelines for Using The Regression Equation

1. If there is no significant linear correlation, don’t use the regression equation to make predictions.

2. When using the regression equation for predictions, stay within the scope of the available sample data.

3. A regression equation based on old data is not necessarily valid now.

4. Don’t make predictions about a population that is different from the population from which the sample data was drawn.

---

Example: Lengths and Weights of Male Bears

<table>
<thead>
<tr>
<th>x Length (in.)</th>
<th>53.0</th>
<th>67.5</th>
<th>72.0</th>
<th>72.0</th>
<th>73.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>y Weight (lb.)</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>262</td>
<td>360</td>
<td>332</td>
</tr>
</tbody>
</table>

\[ \hat{y} = -352 + 9.66x \]

What is the weight of a bear that is 60 inches long?

Since the data does have a significant positive linear correlation, we can use the regression equation for prediction.
Example: Lengths and Weights of Male Bears

<table>
<thead>
<tr>
<th>x Length (in.)</th>
<th>53.0</th>
<th>67.5</th>
<th>72.0</th>
<th>72.0</th>
<th>73.5</th>
<th>68.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>y Weight (lb.)</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>262</td>
<td>360</td>
<td>332</td>
<td>34</td>
</tr>
</tbody>
</table>

\[ \hat{y} = -352 + 9.66 \times 60 \]

\[ \hat{y} = 227.6 \text{ pounds} \]

A bear that is 60 inches long will weigh approximately 227.6 pounds.
Example: Lengths and Weights of Male Bears

<table>
<thead>
<tr>
<th>x Length (in.)</th>
<th>53.0</th>
<th>67.5</th>
<th>72.0</th>
<th>72.0</th>
<th>73.5</th>
<th>68.5</th>
<th>73.0</th>
<th>37.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>y Weight (lb.)</td>
<td>80</td>
<td>344</td>
<td>416</td>
<td>348</td>
<td>262</td>
<td>360</td>
<td>332</td>
<td>34</td>
</tr>
</tbody>
</table>

If there were no significant linear correlation, to predict a weight for any length:

- Use the average of the weights (y-values)
  \[ \bar{y} = 272 \text{ lbs} \]
Residuals and the Least-Squares Property

Definitions

- **Residual**
  for a sample of paired $(x, y)$ data, the difference $(y - \hat{y})$ between an observed sample $y$-value and the value of $\hat{y}$, which is the value of $y$ that is predicted by using the regression equation.

- **Least-Squares Property**
  A straight line satisfies this property if the sum of the squares of the residuals is the smallest sum possible.

---

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>4</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>
Residuals and the Least-Squares Property

\[ \hat{y} = 5 + 4x \]

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>4</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
</tbody>
</table>

Residuals:
- Residual = 7
- Residual = -13
- Residual = -5
- Residual = 11

What is the best predicted size of a household that discards 0.50 lb of plastic?

Data from the Garbage Project

<table>
<thead>
<tr>
<th>x Plastic (lb)</th>
<th>0.27</th>
<th>1.41</th>
<th>2.19</th>
<th>2.83</th>
<th>2.19</th>
<th>1.81</th>
<th>0.85</th>
<th>3.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>y Household</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Using a calculator:
- \( b_0 = 0.549 \)
- \( b_1 = 1.48 \)
- \( y = 0.549 + 1.48 \cdot (0.50) \)
- \( y = 1.3 \)

A household that discards 0.50 lb of plastic has approximately one person.