Simple Linear Regression (SLR)
Linear Regression and Correlation

- Linear regression and correlation are aimed at understanding how two variables are related.
- The variables are called Y and X.
- Y is called the dependent variable.
- X is called the independent variable.
- We want to know how, and whether, X influences Y.
Simple Linear Regression Model

Prediction Equation

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \]

Sample Slope

\[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

where

\[ S_{XY} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \]

\[ S_{XX} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]

Sample Y-intercept

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]
### Simple Linear Regression Model (Example)

Parameter Estimation Solution Table

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$Y_i$</th>
<th>$X_i^2$</th>
<th>$Y_i^2$</th>
<th>$X_iY_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.0</td>
<td>16</td>
<td>9.00</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>5.5</td>
<td>36</td>
<td>30.25</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>6.5</td>
<td>100</td>
<td>42.25</td>
<td>65</td>
</tr>
<tr>
<td>12</td>
<td>9.0</td>
<td>144</td>
<td>81.00</td>
<td>108</td>
</tr>
<tr>
<td>32</td>
<td>24.0</td>
<td>296</td>
<td>162.50</td>
<td>218</td>
</tr>
</tbody>
</table>
Coefficient of Correlation

Values

-1.0  -0.5  0  +0.5  +1.0

Increasing degree of negative correlation
Correlation analysis

Testing the Coefficient of Correlation

$H_0: \rho = 0$ There is no correlation between $x$ and $y$.

$H_1: \rho \neq 0$ There is a correlation between $x$ and $y$.

Reject $H_0$ if:

$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

Test Statistic:

$$t = \frac{r \sqrt{(n-2)}}{\sqrt{1 - r^2}}$$

follows a Student’s $t$ Distribution with $(n-2)$ degrees of freedom. (why d.f.=n-2?)
This coefficient represents the percentage change in the dependent variable explained by the independent variable. The formula for $R^2$ is shown below:

$$R^2 = r^2 \times 100$$

For this example this would be calculated as follows:

$$R^2 = 0.82^2 \times 100 = 67.24\%$$

Based on the coefficient of determination approximately 67.24% of the number of buns sold can be explained by the number of raisins per bun.
Regression Analysis: Pull Strength(y) versus wire length, Die Height

The regression equation is

Pull Strength(y) = 3.31 + 2.41 wire length + 0.0153 Die Height

<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>3.3114</td>
<td>0.9959</td>
<td>3.33</td>
<td>0.008</td>
</tr>
<tr>
<td>wire length</td>
<td>2.4066</td>
<td>0.1260</td>
<td>19.10</td>
<td>0.000</td>
</tr>
<tr>
<td>Die Height</td>
<td>0.015254</td>
<td>0.002710</td>
<td>5.63</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 1.50278   R-Sq = 98.2%   R-Sq(adj) = 97.9%

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>2</td>
<td>1258.83</td>
<td>629.41</td>
<td>278.70</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>10</td>
<td>22.58</td>
<td>2.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>1281.41</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>