**SIMPLE LINEAR REGRESSION**

In linear regression, we consider the frequency distribution of one variable (Y) at each of several levels of a second variable (X).

Y is known as the dependent variable. The variable for which you collect data.  
X is known as the independent variable. The variable for the treatments.

**Determining the Regression Equation**

One goal of regression is to draw the “best” line through the data points.  
The best line usually is obtained using means instead of individual observations.

**Example**  
Effect of hours of mixing on temperature of wood pulp

<table>
<thead>
<tr>
<th>Hours of mixing (X)</th>
<th>Temperature of wood pulp (Y)</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>21</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>108</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>174</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>512</td>
</tr>
<tr>
<td>10</td>
<td>86</td>
<td>860</td>
</tr>
<tr>
<td>12</td>
<td>92</td>
<td>1104</td>
</tr>
</tbody>
</table>

\[ \Sigma X = 42 \quad \Sigma Y = 319 \quad \Sigma XY = 2800 \]
\[ \Sigma X^2 = 364 \quad \Sigma Y^2 = 21,967 \quad n = 6 \]
The equation for any straight line can be written as:  \( \hat{Y} = b_0 + b_1 X \)

where:  
- \( b_0 \) = Y intercept, and  
- \( b_1 \) = regression coefficient = slope of the line

The linear model can be written as:  
\[ Y_i = \beta_0 + \beta_1 X + \varepsilon_i \]

where:  
\( \varepsilon_i \) = residual = \( Y_i - \hat{Y}_i \)

With the data provided, our first goal is to determine the regression equation

Step 1. Solve for  \( b_1 \)

\[
b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum XY - \left( \frac{\sum X \sum Y}{n} \right)}{\sum X^2 - \left( \frac{\sum X^2}{n} \right)} = \frac{\text{SS Cross Products}}{\text{SS X}} = \frac{\text{SSCP}}{\text{SS X}}
\]
for the data in this example

\[
\begin{align*}
\sum X &= 42 & \sum Y &= 319 & \sum XY &= 2,800 & \sum X^2 &= 364 & \sum Y^2 &= 21,967 \\
\end{align*}
\]

\[
\begin{align*}
b_1 &= \frac{\sum XY - \left(\frac{\sum X \sum Y}{n}\right)}{\sum X^2 - \left(\frac{\sum X^2}{n}\right)} \\
&= \frac{2800 - \left(\frac{42 \times 319}{6}\right)}{364 - \left(\frac{42^2}{6}\right)} \\
&= \frac{567}{70} \\
&= 8.1
\end{align*}
\]

The number calculated for \(b_1\), the regression coefficient, indicates that for each unit increase in \(X\) (i.e., hours of mixing), \(Y\) (i.e., wood pulp temperature) will increase 8.1 units (i.e., degrees).

The regression coefficient can be a positive or negative number.

To complete the regression equation, we need to calculate \(b_0\).

\[
b_0 = \bar{Y} - b_1 \bar{X} = \frac{319}{6} - 8.1 \left(\frac{42}{6}\right) = -3.533
\]

Therefore, the regression equation is: \(\hat{Y} = -3.533 + 8.1X\)
**Assumptions of Regression**

1. There is a linear relationship between X and Y
2. The values of X are known constants and presumably are measure without error.
3. For each value of X, Y is independent and normally distributed: $Y \sim \text{N}(0, \sigma^2)$. 
4. Sum of deviations from the regression line equals zero: $\sum (Y_i - \hat{Y}_i) = 0$.
5. Sum of squares for error are a minimum.

If you square the deviations and sum across all observations, you obtain the definition formulas for the following sums of squares:

$$\sum (\hat{Y}_i - \overline{Y})^2 = \text{Sum Squares Due to Regression}$$

$$\sum (Y_i - \hat{Y}_i)^2 = \text{Sum Squares Due to Deviation from Regression (Residual)}$$

$$\sum (Y_i - \overline{Y})^2 = \text{Sum Squares Total}$$
Testing the hypothesis that a linear relationship between X and Y exists

The hypotheses to test that a linear relationship between X and Y exists are:

\[ H_0: \beta_1 = 0 \]
\[ H_A: \beta_1 \neq 0 \]

These hypotheses can be tested using three different methods:
1. F-test
2. t-test
3. Confidence interval

Method 1. F-test

The ANOVA to test \( H_0: \beta_1 = 0 \) can be done using the following sources of variation, degrees of freedom, and sums of squares:

<table>
<thead>
<tr>
<th>SOV</th>
<th>df</th>
<th>Sum of Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to regression</td>
<td>1</td>
<td>( \frac{(\sum XY - (\frac{\sum X \sum Y}{n})^2)}{SSCP^2} )</td>
</tr>
<tr>
<td>Residual</td>
<td>n-2</td>
<td>Determined by subtraction</td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>( \sum Y^2 - \left(\frac{\sum Y^2}{n}\right) = SSY )</td>
</tr>
</tbody>
</table>

Using data from the example:

\( \sum X = 42 \quad \sum Y = 319 \quad \sum XY = 2,800 \quad \sum X^2 = 364 \quad \sum Y^2 = 21,967 \)

Step 1. Calculate Total SS =

\[ \sum Y^2 - \left(\frac{\sum Y^2}{n}\right) = 21,967 - \frac{319^2}{6} = 5,006.833 \]
Step 2. Calculate SS Due to Regression =

\[
\left( \frac{\sum XY - \left( \frac{\sum X \sum Y}{n} \right)^2}{\sum X^2 - \left( \frac{\sum X^2}{n} \right)^2} \right) = \left( \frac{2800 - \left( \frac{42x319}{6} \right)^2}{364 - \frac{42^2}{6}} \right) = \frac{321,489}{70} = 4,592.7
\]

Step 3. Calculate Residual SS = SS Deviation from Regression

Total SS - SS Due to Regression

5006.833 - 4592.7 = 414.133

Step 4. Complete ANOVA

<table>
<thead>
<tr>
<th>SOV</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to Regression</td>
<td>1</td>
<td>4592.7</td>
<td>4592.7</td>
<td>Due to Reg. MS/Residual MS = 44.36**</td>
</tr>
<tr>
<td>Residual</td>
<td>4</td>
<td>414.133</td>
<td>103.533</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>5006.833</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The residual mean square is an estimate of \( \sigma^2_{\text{Y|X}} \), read as variance of Y given X. This parameter estimates the statistic \( \sigma^2_{\text{Y|X}} \).

Step 5. Because the F-test on the Due to Regression SOV is significant, we reject \( H_0: \beta_1 = 0 \) at the 99% level of confidence and can conclude that there is a linear relationship between X and Y.

**Coefficient of Determination - \( r^2 \)**

From the ANOVA table, the coefficient of variation can be calculated using the formula

\[
r^2 = \frac{\text{SS Due to Regression}}{\text{SS Total}}
\]

This value always will be positive and range from 0 to 1.0.

As \( r^2 \) approaches 1.0, the association between X and Y improves.

\( r^2 \times 100 \) is the percentage of the variation in Y that can be explained by having X in the model.

For our example: \( r^2 = \frac{4592.7}{5006.833} = 0.917. \)
We can conclude that 91.7% (i.e. 0.917 x 100) of the variation in wood pulp temperature can be explained by hours of mixing.

**Method 2. t-test**

The formula for the t-test to test the hypothesis $H_0: \beta_1=0$ is:

$$
t = \frac{b_1}{s_{b_1}}
$$

where: $b_1$ the regression coefficient, and

$$
s_{b_1} = \sqrt{\frac{s^2_{Y|X}}{SS X}}
$$

Remember that $s^2_{Y|X} = \text{Residual MS} = \frac{[SS Y - (SSCP^2 / SS X)]}{(n-2)}$

For our example:

**Step 1. Calculate $s^2_{b_1}$**

We know from previous parts of this example:

$$
SS Y = 5006.833 \\
SSCP = 567.0 \\
SS X = 70.0
$$

Therefore, $s^2_{b_1} = \frac{s^2_{Y|X}}{SS X}$

$$
\begin{align*}
    SS Y - \frac{SSCP^2}{SS X} \\
    \frac{SS Y - \frac{SSCP^2}{SS X}}{n-2}/SS X \\
    \frac{5006.833 - \frac{567^2}{70}}{6-2}/70 \\
    \frac{5006.833 - 567^2}{70} \\
    = 1.479
\end{align*}
$$
Step 2. Calculate t statistic

\[ t = \frac{b_1}{s_{b_1}} \]

\[ = \frac{8.1}{\sqrt{1.479}} \]

\[ = 6.66 \]

Step 3. Look up table t value

\text{Table } t_{\alpha/2, (n-2) \text{ df}} = t_{0.05/2, 4 \text{ df}} = 2.776

Step 4. Draw conclusions

Since the table t value (2.776) is less that the calculated t value (6.66), we reject \( H_0: \beta_1=0 \) at the 95% level of confidence. Thus, we can conclude that there is a linear relationship between hours of mixing and wood pulp temperature at the 95% level of confidence.

\textbf{Method 3. Confidence Interval}

The hypothesis \( H_0: \beta_1=0 \) can be tested using the confidence interval:

\[ CI = b_1 \pm t_{\alpha/2, (n-2) \text{ df}} (s_{b_1}) \]

For this example:

\[ CI = b_1 \pm t_{\alpha/2, (n-2) \text{ df}} (s_{b_1}) \]

\[ = 8.1 \pm 2.776\sqrt{1.479} \]

\[ = 4.724 \leq \beta_1 \leq 11.476 \]

We reject \( H_0: \beta_1=0 \) at the 95% level of confidence since the CI does not include 0.
Predicting Y Given X

Regression analysis also can be used to predict a value for Y given X.

Using the example, we can predict the temperature of one batch of wood pulp after mixing X hours.

In this case, we predict an individual outcome of \( Y_X \) drawn from the distribution of Y.

This estimate is distinct from estimating mean or average of a distribution of Y.

The value of an individual Y at a given X will take on the form of the confidence interval:

\[
CI = \hat{Y} \pm t_{\alpha/2,(n-2)df} \left( s_{Y|X=X_0} \right)
\]

where \( s_{Y|X=X_0} = \sqrt{s_{Y|X=0}^2} \), and

\[
s_{Y|X=X_0}^2 = s_{Y|X}^2 \left[ 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right]
\]

Remember \( s_{Y|X}^2 \) is the Residual Mean Square

Example 1

We wish to determine the temperature of the one batch of wood pulp after mixing two hours (i.e., \( Y_{X=2} \)).

Step 1. Using the regression equation, solve for \( \hat{Y} \) when X=2.

\[
Y = -3.533 + 8.1X
\]

\[
\hat{Y} = -3.533 + 8.1(2) = 12.667
\]

Step 2. Solve for \( s_{Y|X=2}^2 \)

\[
s_{Y|X=X_0}^2 = s_{Y|X}^2 \left[ 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right]
\]

\[
= 103.533 \left[ 1 + \frac{1}{6} + \frac{(2 - 7)^2}{70} \right]
\]

\[
= 157.765
\]
Step 3. Calculate the confidence interval

\[ \text{CI} = \hat{Y} \pm t_{\alpha/2, (n-2)} \sqrt{\frac{s_{Y|X=X_0}^2}{n}} \]

\[ = 12.667 \pm 2.776 \sqrt{157.65} \]

\[ = 12.667 \pm 34.868 \]

Therefore: LCI = -22.201 and UCI = 47.535

**Note: This CI is not used to test a hypothesis**

This CI states that if we mix the wood pulp for two hours, we would expect the temperature to fall within the range of -22.201 and 47.535 degrees 95% of the time.

We would expect the temperature to fall outside of this range 5% of the time due to random chance.

**Example 2**

We wish to determine the temperature of the one batch of wood pulp after mixing seven hours (i.e., \( Y_{X=7} \)).

Step 1. Using the regression equation, solve for \( \hat{Y} \) when \( X=7 \).

Remember \( \hat{Y} = -3.533 + 8.1X \)

\( \hat{Y} = -3.533 + 8.1(7) = 53.167 \)

Step 2. Solve for \( s_{Y|X=7}^2 \)

\[ s_{Y|X=X_0}^2 = s_{Y|X}^2 \left[ 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS X} \right] \]

\[ = 103.533 \left[ 1 + \frac{1}{6} + \frac{(7 - 7)^2}{70} \right] \]

\[ = 120.789 \]
Step 3. Calculate the confidence interval

\[ \text{CI} = \hat{Y} \pm t_{\alpha/2, (n-2) df} \left( s_{Y|X=X_0} \right) \]

\[ = 53.167 \pm 2.776\sqrt{120.789} \]

\[ = 53.167 \pm 30.509 \]

Therefore: LCI = 22.658 and UCI = 83.676

Note: For \( X=7 \) (i.e., at the mean of \( X \)), the variance \( s^2_{Y|X=X_0} \) is at a minimum.

This CI states that if we mix the wood pulp for seven hours, we would expect the temperature to fall within the range of 22.658 and 83.676 degrees 95% of the time.

We would expect the temperature to fall outside of this range 5% of the time due to random chance.

Predicting \( \bar{Y} \) Given \( X \)

Regression analysis also can be used to predict a value for \( \bar{Y} \) given \( X \).

Using the example, we can predict the average temperature of wood pulp after mixing \( X \) hours.

In this case, we predict an individual outcome of \( \bar{Y}_X \) drawn from the distribution of \( Y \).

This estimate is distinct from distribution of \( Y \) for a \( X \).

The value of an individual \( Y \) at a given \( X \) will take on the form of the confidence interval:

\[ \text{CI} = \hat{Y} \pm t_{\alpha/2, (n-2) df} \left( s_{Y|X=X_0} \right) \]

where \[ s_{Y|X=X_0} = \sqrt{s^2_{Y|X=X_0}}, \] and

\[ s^2_{Y|X=X_0} = s^2_{Y|X} \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SS X} \right] \]
Example 1

We wish to determine the average temperature of the wood pulp after mixing two hours (i.e., $Y_{X=2}$).

Step 1. Using the regression equation, solve for $\hat{Y}$ when $X=2$.

Remember $\hat{Y} = -3.533 + 8.1X$

\[
\hat{Y} = -3.533 + 8.1(2) = 12.667
\]

Step 2. Solve for $s_{\bar{Y}|X=2}^2$

\[
s_{\bar{Y}|X=2}^2 = s_{Y|X}^2 \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SSX} \right]
\]

\[
= 103.533 \left[ \frac{1}{6} + \frac{(2 - 7)^2}{70} \right]
\]

\[
= 54.232
\]

Step 3. Calculate the confidence interval

\[
CI = \hat{Y} \pm t_{\alpha/2,(n-2)df} \left( s_{\bar{Y}|X=2} \right)
\]

\[
= 12.667 \pm 2.776 \sqrt{54.232}
\]

\[
= 12.667 \pm 20.443
\]

Therefore: LCI = -7.776 and UCI = 33.110

Note: This CI is not used to test a hypothesis

This CI states that if we mix the wood pulp for two hours any number of times, we would expect the average temperature to fall within the range of -7.776 and 33.110 degrees 95% of the time.

We would expect the temperature to fall outside of this range 5% of the time due to random chance.
Example 2

We wish to determine the average temperature of wood pulp after mixing seven hours.

Step 1. Using the regression equation, solve for \( \hat{Y} \) when \( X=7 \).

Remember \( \hat{Y} = -3.533 + 8.1X \)

\[
\hat{Y} = -3.533 + 8.1(7) = 53.167
\]

Step 2. Solve for \( s^2_{\hat{Y}|X=7} \)

\[
s^2_{\hat{Y}|X=7} = s^2_Y \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SSX} \right]
\]

\[
= 103.533 \left[ \frac{1}{6} + \frac{(7 - 7)^2}{70} \right]
\]

\[
= 17.256
\]

Step 3. Calculate the confidence interval

\[
CI = \hat{Y} \pm t_{\alpha/2,(n-2)df} \left( s_{\hat{Y}|X=X_0} \right)
\]

\[
= 53.167 \pm 2.776\sqrt{17.256}
\]

\[
= 53.167 \pm 11.532
\]

Therefore: LCI = 41.635 and UCI = 64.669

Note: For \( X=7 \) (i.e., at the mean of \( X \)), the variance \( s^2_{\hat{Y}|X=X_0} \) is at a minimum.
Comparing $s_{Y|X=x_0}^2$ and $s_{F|X=X_0}^2$

$s_{Y|X=x_0}^2$ is always greater than $s_{F|X=X_0}^2$.

Comparing the formulas:

$$s_{Y|X=x_0}^2 = s_{Y|X}^2 \left[ 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right]$$

$$s_{F|X=X_0}^2 = s_{Y|X}^2 \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{SS_X} \right].$$

Notice that in the formula for $s_{Y|X=x_0}^2$, you add one while in the formula for $s_{F|X=X_0}^2$, you do not.

Comparison of $s_{Y|X=x_0}^2$ and $s_{F|X=X_0}^2$.

| $X$ | $s_{Y|X=x_0}^2$ | $s_{F|X=X_0}^2$ |
|-----|-----------------|-----------------|
| 2   | 157.767         | 54.232          |
| 7   | 120.789         | 17.256          |
We can draw the two confidence intervals as “confidence belts” about the regression line.

Notice that:
1. The confidence belts are symmetrical about the regression line
2. The confidence belts are narrowest at the mean of X, and
3. The confidence belts for the distribution based on means is narrower that the distribution based on an individual observation.
Determining if Two Independent Regression Coefficients are Different

It may be desirable to test the homogeneity of two $b_1$'s to determine if they are estimates of the same $\beta_1$.

This can be done using a $t$-test to test the hypotheses:

$H_0$: $\beta_1 = \beta_1'$

$H_A$: $\beta_1 \neq \beta_1'$

Where $t = \frac{b_1 - b_1'}{\sqrt{(\text{Residual 1 MS} + \text{Residual 2 MS}) \left(\frac{1}{X_1SS} + \frac{1}{X_2SS}\right)}}$

The Table $t$ value has $(n_1 - 2) + (n_2 - 2)$ df.

Example

<table>
<thead>
<tr>
<th>X</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>58</td>
</tr>
</tbody>
</table>

$\sum X = 15$  $\sum Y_1 = 82$  $\sum Y_2 = 187$

$\sum X_2 = 55$  $\sum Y_1^2 = 1426$  $\sum Y_2^2 = 7827$

Step 1. Determine regression coefficient for each $Y$

for $Y_1$  $\sum XY = 272$

Thus $b_1 = \frac{272 - (15 \times 82)/5}{55 - 15^2/5} = 2.6$

for $Y_2$  $\sum XY = 651$

Thus $b_1' = \frac{651 - (15 \times 187)/5}{55 - 15/5} = 9.0$
Step 2. Calculate Residual MS for each Y

\[ \text{Residual MS} = \frac{\text{SS Y} - \left( \frac{\text{SSCP}^2}{\text{SS X}} \right)}{n - 2} \]

Remember Residual MS = \frac{\left( 1426 - \frac{8^2}{5} \right) - \left( \frac{26^2}{10} \right)}{5 - 2}

= 4.5

\[ \text{Residual 1 MS} = \frac{\left( 7827 - \frac{187^2}{5} \right) - \left( \frac{90^2}{10} \right)}{5 - 2} \]

= 7.7

Step 3. Solve for t

\[ \frac{2.6 - 9.0}{\sqrt{(4.5 + 7.7) \times \left( \frac{1}{10} + \frac{1}{10} \right)}} \]

= \frac{-6.4}{\sqrt{12.2(0.2)}}

= -4.10

Step 4. Look up table t value with (n_1 - 2) + (n_2 - 2) df

\[ t_{0.05/2, 6 \text{ df}} = -2.447 \]

Step 4. Make conclusions

Because the absolute value of the calculated t-value (-4.10) is greater than the absolute value of the tabular t value (2.776), we can conclude at the 95% level of confidence that the two regression coefficients are not estimating the same \( \beta_1 \).
**Summary - Some Uses of Regression**

1. Determine if there is a linear relationship between an independent and dependent variable.

2. Predict values of Y at a given X
   Most accurate near the mean of X.
   
   **Should avoid predicting values of Y outside the range of the independent variables that were used.**

3. Can adjust Y to a common base by removing the effect of the independent variables (Analysis of Covariance).

4. ANOVA (CRD, RCBD, and LS) can be done using regression

5. Compare homogeneity of two regression coefficients.