Simple Linear Regression and Correlation

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients $b_0$ and $b_1$
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values
Correlation vs. Regression

- A scatter diagram can be used to show the relationship between two variables.
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables:
  - Correlation is only concerned with strength of the relationship.
  - No causal effect is implied with correlation.
Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

**Dependent variable:** the variable we wish to predict or explain (i.e. runoff)

**Independent variable:** the variable used to explain the dependent variable (i.e. rainfall)
**Simple Linear Regression Model**

- Only **one** independent variable, \( X \)
- Relationship between \( X \) and \( Y \) is described by a linear function
- Changes in \( Y \) are assumed to be caused by changes in \( X \)
Types of Relationships

Linear relationships

Curvilinear relationships
Types of Relationships (continued)

**Strong relationships**

![Strong relationships graph 1](image1)

![Strong relationships graph 2](image2)

**Weak relationships**

![Weak relationships graph 1](image3)

![Weak relationships graph 2](image4)
Types of Relationships

No relationship

Y

X

Y

X
Simple Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

- Dependent Variable
- Population Y intercept
- Population Slope Coefficient
- Independent Variable
- Random Error term

**Linear component**

**Random Error component**
Simple Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

\( \varepsilon_i \sim N(0, \sigma^2) \)

- Observed Value of \( Y \) for \( X_i \)
- Predicted Value of \( Y \) for \( X_i \)
- Intercept = \( \beta_0 \)
- Slope = \( \beta_1 \)
- Random Error for this \( X_i \) value

\( \sigma \) is the standard deviation of the error term.
Simple Linear Regression Model

$\epsilon_i \sim N(0, \sigma^2)$

$y = \beta_0 + \beta_1 x$
Simple Linear Regression Model

(a) Normal, mean 0, standard deviation $\sigma$

(b) $\varepsilon_i \sim N(0, \sigma^2)$

$\beta_0 + \beta_1 x_1$

$\beta_0 + \beta_1 x_2$

$\beta_0 + \beta_1 x_3$

Line $y = \beta_0 + \beta_1 x$
The simple linear regression equation provides an estimate of the population regression line.

\[ \hat{Y}_i = b_0 + b_1 X_i \]

- Estimated (or predicted) Y value for observation i
- Estimate of the regression intercept
- Estimate of the regression slope
- Value of X for observation i

The individual random error terms \( e_i \) have a mean of zero.
Least Squares Method

- $b_0$ and $b_1$ are obtained by finding the values of $b_0$ and $b_1$ that minimize the sum of the squared differences between $Y$ and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$
Finding the Least Squares Equation

- Computational formula for the slope $b_1$:

\[
b_1 = \frac{SS_{XY}}{SS_{XX}} \quad b_1 \Rightarrow \hat{\beta}_1 \text{ in the text book}
\]

where

\[
SS_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})
\]

\[
SS_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
Finding the Least Squares Equation

- Computational formula for the Y intercept $b_0$:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$b_0 \Rightarrow \hat{\beta}_0$ in the textbook.
The coefficients $b_0$ and $b_1$, and other regression results in this chapter, will be found using Excel.
Interpretation of the Slope and the Intercept

- $b_0$ is the estimated average value of $Y$ when the value of $X$ is zero.

- $b_1$ is the estimated change in the average value of $Y$ as a result of a one-unit change in $X$. 
Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet).

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet
Sample Data for House Price Model

<table>
<thead>
<tr>
<th>House Price in $1000s (Y)</th>
<th>Square Feet (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
</tr>
<tr>
<td>312</td>
<td>1600</td>
</tr>
<tr>
<td>279</td>
<td>1700</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
</tr>
</tbody>
</table>
Graphical Presentation

- **House price model: scatter plot**

![House price model: scatter plot](image-url)
Regression Using Excel

- Tools / Data Analysis / Regression
Excel Output

**Regression Statistics**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.76211</td>
</tr>
<tr>
<td>R Square</td>
<td>0.58082</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.52842</td>
</tr>
<tr>
<td>Standard Error</td>
<td>41.33032</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>

The regression equation is:

\[
\text{house price} = 98.24833 + 0.10977 \text{ (square feet)}
\]
Graphical Presentation

- House price model: scatter plot and regression line

\[
\text{house price} = 98.24833 + 0.10977 \text{ (square feet)}
\]

Slope = 0.10977
Intercept = 98.248
Interpretation of the Intercept, $b_0$

- $b_0$ is the estimated average value of $Y$ when the value of $X$ is zero (if $X = 0$ is in the range of observed $X$ values)

- Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, $98,248.33$ is the portion of the house price not explained by square feet
Interpretation of the Slope Coefficient, $b_1$

House price = 98.24833 + 0.10977 (square feet)

- $b_1$ measures the estimated change in the average value of $Y$ as a result of a one-unit change in $X$

- Here, $b_1 = 0.10977$ tells us that the average value of a house increases by $0.10977($1000) = $109.77, on average, for each additional one square foot of size.
Predict the price for a house with 2000 square feet:

\[
\hat{\text{house price}} = 98.25 + 0.1098 \text{ (sq. ft.)}
\]

\[
= 98.25 + 0.1098(2000)
\]

\[
= 317.85
\]

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850
**Interpolation vs. Extrapolation**

- When using a regression model for prediction, only predict within the relevant range of data.

![Graph showing interpolation and extrapolation]

- **Relevant range for interpolation**

- Do not try to extrapolate beyond the range of observed X’s.
Measures of Variation

- Total variation is made up of two parts:

\[ \text{SST} = \text{SSR} + \text{SSE} \]

- Total Sum of Squares
- Regression Sum of Squares
- Error Sum of Squares

where:

- \( \bar{Y} \) = Average value of the dependent variable
- \( Y_i \) = Observed values of the dependent variable
- \( \hat{Y}_i \) = Predicted value of \( Y \) for the given \( X_i \) value

\[ \text{SST} = \sum (Y_i - \bar{Y})^2 \]
\[ \text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2 \]
\[ \text{SSE} = \sum (Y_i - \hat{Y}_i)^2 \]
Measures of Variation (continued)

- **SST = total sum of squares**
  - Measures the variation of the $Y_i$ values around their mean $Y$

- **SSR = regression sum of squares**
  - Explained variation attributable to the relationship between $X$ and $Y$

- **SSE = error sum of squares**
  - Variation attributable to factors other than the relationship between $X$ and $Y
Measures of Variation

\[
\text{SST} = \sum (Y_i - \bar{Y})^2
\]

\[
\text{SSE} = \sum (Y_i - \hat{Y}_i)^2
\]

\[
\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2
\]
The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

The coefficient of determination is also called r-squared and is denoted as $r^2$.

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Note: $0 \leq r^2 \leq 1$
Examples of Approximate $r^2$ Values

Perfect linear relationship between $X$ and $Y$:

$100\%$ of the variation in $Y$ is explained by variation in $X$.
Examples of Approximate $r^2$ Values

0 < $r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X
Examples of Approximate $r^2$ Values

No linear relationship between $X$ and $Y$:  

The value of $Y$ does not depend on $X$. (None of the variation in $Y$ is explained by variation in $X$)
Excel Output

Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.76211</td>
</tr>
<tr>
<td>R Square</td>
<td>0.58082</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.52842</td>
</tr>
<tr>
<td>Standard Error</td>
<td>41.33032</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

\[ r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082 \]

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficient Estimates

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>1.69296</td>
<td>0.12892</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>
Standard Error of Estimate

(continued)

\[ \sigma \sim N(0, \sigma^2) \]

Standard Error Variance

Normal, mean 0, standard deviation \( \sigma \)

\( \varepsilon_i \sim N(0, \sigma^2) \)
The standard deviation of the variation of observations around the regression line is estimated by

\[
\hat{\sigma} \quad S_{YX} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}}
\]

Where

- \( \text{SSE} \) = error sum of squares
- \( n \) = sample size
### Excel Output

#### Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.76211</td>
</tr>
<tr>
<td>R Square</td>
<td>0.58082</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.52842</td>
</tr>
<tr>
<td>Standard Error</td>
<td>41.33032</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

#### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
<td>0.12892</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>
Comparing Standard Errors

$S_{YX}$ is a measure of the variation of observed $Y$ values from the regression line.

The magnitude of $S_{YX}$ should always be judged relative to the size of the $Y$ values in the sample data.

i.e., $S_{YX} = $41.33K is moderately small relative to house prices in the $200 - $300K range.
Assumptions of Regression

Use the acronym LINE:

- **Linearity**
  - The underlying relationship between X and Y is linear

- **Independence of Errors**
  - Error values are statistically independent

- **Normality of Error**
  - Error values (ε) are normally distributed for any given value of X

- **Equal Variance (Homoscedasticity)**
  - The probability distribution of the errors has constant variance
Residual Analysis

The residual for observation $i$, $e_i$, is the difference between its observed and predicted value.

- Check the assumptions of regression by examining the residuals:
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of $X$ (homoscedasticity)

**Graphical Analysis of Residuals**
- Can plot residuals vs. $X$
Residual Analysis for Linearity

- **Not Linear**
  - Residuals show a clear pattern, indicating non-linearity.

- **Linear**
  - Residuals are randomly scattered around the zero line, indicating linearity.
Residual Analysis for Independence

Not Independent

Independent
Residual Analysis for Normality

- A normal probability plot of the residuals can be used to check for normality:

![Graph showing a normal probability plot with residual values on the x-axis and percent on the y-axis. The points follow a straight line, indicating normality.]
Residual Analysis for Equal Variance

- **Non-constant variance**
- **Constant variance**
Excel Residual Output

RESIDUAL OUTPUT

<table>
<thead>
<tr>
<th>Predicted House Price</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 251.92316</td>
<td>-6.923162</td>
</tr>
<tr>
<td>2 273.87671</td>
<td>38.12329</td>
</tr>
<tr>
<td>3 284.85348</td>
<td>-5.853484</td>
</tr>
<tr>
<td>4 304.06284</td>
<td>3.937162</td>
</tr>
<tr>
<td>5 218.99284</td>
<td>-19.99284</td>
</tr>
<tr>
<td>6 268.38832</td>
<td>-49.38832</td>
</tr>
<tr>
<td>7 356.20251</td>
<td>48.79749</td>
</tr>
<tr>
<td>8 367.17929</td>
<td>-43.17929</td>
</tr>
<tr>
<td>9 254.6674</td>
<td>64.33264</td>
</tr>
<tr>
<td>10 284.85348</td>
<td>-29.85348</td>
</tr>
</tbody>
</table>

House Price Model Residual Plot

Does not appear to violate any regression assumptions
Inferences About the Slope

- The standard error of the regression slope coefficient \( b_1 \) is estimated by

\[
S_{b_1} = \frac{S_{YX}}{\sqrt{SS_{XX}}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}
\]

where:
- \( S_{b_1} \) = Estimate of the standard error of the least squares slope
- \( S_{YX} = \sqrt{\frac{SSE}{n-2}} \) = Standard error of the estimate
### Excel Output

**Regression Statistics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.76211</td>
</tr>
<tr>
<td>R Square</td>
<td>0.58082</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.52842</td>
</tr>
<tr>
<td>Standard Error</td>
<td>41.33032</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Coefficients**

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>58.03348</td>
<td>1.69296</td>
<td>0.12892</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>
Comparing Standard Errors of the Slope

$S_{b_1}$ is a measure of the variation in the slope of regression lines from different possible samples.
Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
  
  \[
  H_0: \beta_1 = 0 \quad \text{(no linear relationship)} \\
  H_1: \beta_1 \neq 0 \quad \text{(linear relationship does exist)}
  \]

- Test statistic

\[
t = \frac{b_1 - \beta_1}{S_{b_1}}
\]

where:

- \(b_1\) = regression slope coefficient
- \(\beta_1\) = hypothesized slope
- \(S_{b_1}\) = standard error of the slope

\[
d.f. = n - 2
\]
Inference about the Slope: 
t Test

Simple Linear Regression Equation:

\[
\text{house price} = 98.25 + 0.1098 \text{ (sq.ft.)}
\]

The slope of this model is 0.1098
Does square footage of the house affect its sales price?
Inferences about the Slope: t Test Example

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

From Excel output:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>1.69296</td>
<td>0.12892</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>3.32938</td>
<td>0.01039</td>
</tr>
</tbody>
</table>

\[
t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938
\]
Inferences about the Slope: t Test Example

Test Statistic: \( t = 3.329 \)

**Hypotheses:**

- \( H_0: \beta_1 = 0 \)
- \( H_1: \beta_1 \neq 0 \)

From Excel output:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>1.69296</td>
<td>0.12892</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>3.32938</td>
<td>0.01039</td>
</tr>
</tbody>
</table>

Decision: Reject \( H_0 \)

Conclusion: There is sufficient evidence that square footage affects house price.

d.f. = 10 - 2 = 8

\( \alpha/2 = 0.025 \)

\( t_{\alpha/2} = 2.3060 \)

\( -t_{\alpha/2} = 2.3060 \)

\( t = 3.329 \)
Inferences about the Slope: t Test Example

H₀: β₁ = 0
H₁: β₁ ≠ 0

P-value = 0.01039

From Excel output:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
</tr>
</tbody>
</table>

This is a two-tail test, so the p-value is
P(t > 3.329)+P(t < -3.329) = 0.01039
(for 8 d.f.)

Decision: P-value < α so Reject H₀

Conclusion: There is sufficient evidence that square footage affects house price
F Test for Significance

- **F Test statistic:**

\[
F = \frac{\text{MSR}}{\text{MSE}}
\]

where

\[
\text{MSR} = \frac{\text{SSR}}{k}
\]

\[
\text{MSE} = \frac{\text{SSE}}{n - k - 1}
\]

where \(F\) follows an F distribution with \(k\) numerator and \((n - k - 1)\) denominator degrees of freedom

\((k = \text{the number of independent variables in the regression model})\)
**Excel Output**

**Regression Statistics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.76211</td>
</tr>
<tr>
<td>R Square</td>
<td>0.58082</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.52842</td>
</tr>
<tr>
<td>Standard Error</td>
<td>41.33032</td>
</tr>
<tr>
<td>Observations</td>
<td>10</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848
\]

With 1 and 8 degrees of freedom

**P-value for the F Test**

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
<td>0.12892</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>
**F Test for Significance**

**Test Statistic:**

\[
F = \frac{\text{MSR}}{\text{MSE}} = 11.08
\]

**Decision:**

Reject \( H_0 \) at \( \alpha = 0.05 \)

**Conclusion:**

There is sufficient evidence that house size affects selling price.
Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1}$$

Excel Printout for House Prices:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>

d.f. = n - 2

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)
Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size.

This 95% confidence interval does not include 0.

**Conclusion:** There is a significant relationship between house price and square feet at the .05 level of significance.
The Sample Covariance

- The sample covariance measures the strength of the linear relationship between \textit{two variables} (called bivariate data).

- The sample covariance:

\[
\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}
\]

- Only concerned with the strength of the relationship
- No causal effect is implied
Interpreting Covariance

- **Covariance** between two random variables:

  - $\text{cov}(X, Y) > 0 \rightarrow X$ and $Y$ tend to move in the **same** direction
  - $\text{cov}(X, Y) < 0 \rightarrow X$ and $Y$ tend to move in **opposite** directions
  - $\text{cov}(X, Y) = 0 \rightarrow X$ and $Y$ are independent
Coefficient of Correlation

- Measures the relative strength of the linear relationship between two variables
- Sample coefficient of correlation:

\[
r = \frac{\text{cov}(X, Y)}{S_X S_Y}
\]

where

\[
\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}
\]

\[
S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}}
\]

\[
S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}}
\]
Features of Correlation Coefficient, $r$

- Unit free
- Ranges between $-1$ and $1$
- The closer to $-1$, the stronger the negative linear relationship
- The closer to $1$, the stronger the positive linear relationship
- The closer to $0$, the weaker the linear relationship
Scatter Plots of Data with Various Correlation Coefficients

- $r = -1$
- $r = -0.6$
- $r = 0$
- $r = 0$
- $r = +0.3$
- $r = +1$
Using Excel to Find the Correlation Coefficient

- Select Tools/Data Analysis
- Choose Correlation from the selection menu
- Click OK...
Using Excel to Find the Correlation Coefficient

- Input data range and select appropriate options
- Click OK to get output
Interpreting the Result

- $r = .733$

- There is a relatively strong positive linear relationship between test score #1 and test score #2

- Students who scored high on the first test tended to score high on second test, and students who scored low on the first test tended to score low on the second test.
t Test for a Correlation Coefficient

- **Hypotheses**
  - $H_0: \rho = 0$ (no correlation between X and Y)
  - $H_A: \rho \neq 0$ (correlation exists)

- **Test statistic**
  - $t = \frac{r - \rho}{\sqrt{1 - r^2}} \sqrt{\frac{1}{n - 2}}$
    - (with $n - 2$ degrees of freedom)
    - where
      - $r = +\sqrt{r^2}$ if $b_1 > 0$
      - $r = -\sqrt{r^2}$ if $b_1 < 0$
Example: House Prices

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

\[ H_0: \rho = 0 \quad \text{(No correlation)} \]
\[ H_1: \rho \neq 0 \quad \text{(correlation exists)} \]
\[ \alpha = .05 \, , \, df = 10 - 2 = 8 \]

\[
t = \frac{r - \rho}{\sqrt{1 - r^2}} = \frac{.762 - 0}{\sqrt{1 - .762^2}} = 3.329
\]
Example: Test Solution

\[
t = \frac{r - \rho}{\sqrt{1 - r^2}} = \frac{.762 - 0}{\sqrt{1 - .762^2}} = 3.329
\]

d.f. = 10-2 = 8

\[
\alpha/2 = .025
\]

Decision:
Reject H₀

Conclusion:
There is evidence of a linear association at the 5% level of significance.
Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around $Y$ to express uncertainty about the value of $Y$ for a given $X_i$

\[ \hat{Y} = b_0 + b_1 X_i \]

Confidence Interval for the mean of $Y$, given $X_i$

Prediction Interval for an individual $Y$, given $X_i$
Confidence Interval for the Average $Y$, Given $X$

Confidence interval estimate for the mean value of $Y$ given a particular $X_i$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, $\bar{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$
Prediction Interval for an Individual Y, Given X

Confidence interval estimate for an Individual value of Y given a particular $X_i$

Confidence interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{1 + h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case.
Estimation of Mean Values:  
Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i = 317.85$ ($1,000s$)

$$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from $280,660$ to $354,900$
Estimation of Individual Values: Example

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i = 317.85 \ (\$1,000s)$

\[
\hat{Y} \pm t_{n-1}S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}} = 317.85 \pm 102.28
\]

The prediction interval endpoints are 215.50 and 420.07, or from $215,500$ to $420,070$
Finding Confidence and Prediction Intervals in Excel

- In Excel, use
  PHStat | regression | simple linear regression …

- Check the
  “confidence and prediction interval for X=”
  box and enter the X-value and confidence level desired
## Finding Confidence and Prediction Intervals in Excel (continued)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Confidence Interval Estimate</strong></td>
<td></td>
</tr>
<tr>
<td><strong>X Value</strong></td>
<td>2000</td>
</tr>
<tr>
<td><strong>Confidence Level</strong></td>
<td>95%</td>
</tr>
<tr>
<td><strong>Sample Size</strong></td>
<td>10</td>
</tr>
<tr>
<td><strong>Degrees of Freedom</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>t Value</strong></td>
<td>2.306006</td>
</tr>
<tr>
<td><strong>Sample Mean</strong></td>
<td>1715</td>
</tr>
<tr>
<td><strong>Sum of Squared Difference</strong></td>
<td>1571500</td>
</tr>
<tr>
<td><strong>Standard Error of the Estimate</strong></td>
<td>41.33032</td>
</tr>
<tr>
<td><strong>t Statistic</strong></td>
<td>0.151666</td>
</tr>
<tr>
<td><strong>Average Predicted Y (YHat)</strong></td>
<td>317.7838</td>
</tr>
<tr>
<td><strong>For Average Predicted Y (YHat)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interval Half Width</strong></td>
<td>37.11952</td>
</tr>
<tr>
<td><strong>Confidence Interval Lower Limit</strong></td>
<td>280.6643</td>
</tr>
<tr>
<td><strong>Confidence Interval Upper Limit</strong></td>
<td>354.9033</td>
</tr>
<tr>
<td><strong>For Individual Response Y</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interval Half Width</strong></td>
<td>102.2813</td>
</tr>
<tr>
<td><strong>Prediction Interval Lower Limit</strong></td>
<td>215.5025</td>
</tr>
<tr>
<td><strong>Prediction Interval Upper Limit</strong></td>
<td>420.0651</td>
</tr>
</tbody>
</table>

### Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Input values

### Prediction Interval Estimate for $Y_{X=X_i}$
Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range
Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter diagram of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
  - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
  - Use a histogram, stem-and-leaf display, box-and-whisker plot, or normal probability plot of the residuals to uncover possible non-normality
Strategies for Avoiding the Pitfalls of Regression

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range