**Simple Linear Regression**

- **Regression analysis** is a tool for building mathematical and statistical models that characterize relationships between variables.
- Finds a linear relationship between:
  - one independent variable $X$ and
  - one dependent variable $Y$
- First prepare a scatter plot to verify the data has a linear trend.
For time series data, use a line chart.

Line chart of historical crude oil prices
Finding the Best-Fitting Regression Line

- For cross-sectional data, use a scatter chart.
- \( Y = a + b \times X \)
- Two possible lines are shown below.
- Line A is clearly a better fit to the data.
- We want to determine the best regression line.
Excel Trendline Tool

- Right click on data series and choose Add trendline from pop-up menu
- Check the boxes Display Equation on chart and Display R-squared value on chart
- \( R^2 \) (R-squared) is a measure of the “fit” of the line to the data.
  - The value of \( R^2 \) will be between 0 and 1.
  - A value of 1.0 indicates a perfect fit and all data points would lie on the line; the larger the value of \( R^2 \) the better the fit.
Modeling a Price-Demand Function

Linear demand function:
Sales = 20,512 - 9.5116(price)
Excel’s *Trendline* tool is used to fit various functions to the data.

- **Exponential**  \( y = 50.49e^{0.021x} \)  \( R^2 = 0.664 \)
- **Logarithmic**  \( y = 13.02\ln(x) + 39.60 \)  \( R^2 = 0.382 \)
- **Polynomial 2°**  \( y = 0.13x^2 - 2.399x + 68.01 \)  \( R^2 = 0.905 \)
- **Polynomial 3°**  \( y = 0.005x^3 - 0.111x^2 + 0.648x + 59.497 \)  \( R^2 = 0.928 \)
- **Power**  \( y = 45.96x^{0.0169} \)  \( R^2 = 0.397 \)
Common Mathematical Functions Used in Predictive Analytical Models

- **Linear**
  \[ y = a + bx \]

- **Logarithmic**
  \[ y = \ln(x) \]

- **Polynomial (2\(^{nd}\) order)**
  \[ y = ax^2 + bx + c \]

- **Polynomial (3\(^{rd}\) order)**
  \[ y = ax^3 + bx^2 + dx + e \]

- **Power**
  \[ y = ax^b \]

- **Exponential**
  \[ y = ab^x \]

*(the base of natural logarithms, \( e = 2.71828 \ldots \) is often used for the constant \( b \))*
Third order polynomial trendline fit to the time series data

\[ y = 0.0052x^3 - 0.1111x^2 + 0.6483x + 59.497 \]

\[ R^2 = 0.9282 \]
Size of a house is typically related to its market value.

\[ X = \text{square footage} \]

\[ Y = \text{market value ($)} \]

The scatter plot of the full data set (42 homes) indicates a linear trend.
Market value = 32,673 + $35.036 \times \text{square feet}

The estimated market value of a home with 2,200 square feet would be: market value = $32,673 + $35.036 \times 2,200 = $109,752

The regression model explains variation in market value due to size of the home. It provides better estimates of market value than simply using the average.
Least-Squares Regression

- We estimate the parameters from the sample data:
  \[ \hat{Y} = b_0 + b_1 X \]

- Residuals are the observed errors associated with estimating the value of the dependent variable using the regression line:
  \[ e_i = Y_i - \hat{Y}_i \]
Least Squares Regression

The best-fitting line minimizes the sum of squares of the residuals.

\[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1X_i])^2 \]

\[ b_1 = \frac{\sum_{i=1}^{n} X_iY_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \]

\[ b_0 = \overline{Y} - b_1\overline{X} \]

Excel functions:
- =INTERCEPT(known_y’s, known_x’s)
- =SLOPE(known_y’s, known_x’s)
Using Excel Functions to Find Least-Squares Coefficients

- Slope = $b_1 = 35.036$
  
  \[ =\text{SLOPE}(C4:C45, \ B4:B45) \]

- Intercept = $b_0 = 32,673$
  
  \[ =\text{INTERCEPT}(C4:C45, \ B4:B45) \]

- Estimate $Y$ when $X = 1750$ square feet
  
  \[ ^\wedge \ Y = 32,673 + 35.036(1750) = $93,986 \]
  
  \[ =\text{TREND}(C4:C45, \ B4:B45, \ 1750) \]
Data > Data Analysis >
Regression
Input Y Range (with header)
Input X Range (with header)
Check Labels

Excel outputs a table with many useful regression statistics.
Multiple R: \(| r |\), sample correlation coefficient, varies from -1 to +1 (\(r\) is negative if slope is negative)

R Square: coefficient of determination, \(R^2\), varies from 0 (no fit) to 1 (perfect fit)

Adjusted R Square - adjusts \(R^2\) for sample size and number of \(X\) variables

Standard Error - variability between observed and predicted \(Y\) values, formally called the standard error of the estimate, \(S_{yx}\).

---

**Home Market Value Regression Results**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Regression Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Multiple R</td>
<td>0.731255223</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.534734202</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.523102557</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>7287.722712</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></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>12</td>
<td>1</td>
<td>2441633669</td>
<td>2441633669</td>
<td>45.97236277</td>
<td>3.79802E-08</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>2124436093</td>
<td>53110902.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>41</td>
<td>4566069762</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<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>17</td>
<td>Intercept</td>
<td>32673.2199</td>
<td>8831.950745</td>
<td>3.699434116</td>
<td>0.000649604</td>
<td>14823.18178</td>
</tr>
<tr>
<td>18</td>
<td>Square Feet</td>
<td>35.03637258</td>
<td>5.16738385</td>
<td>6.780292234</td>
<td>3.79802E-08</td>
<td>24.59270036</td>
</tr>
</tbody>
</table>
53% of the variation in home market values can be explained by home size.

The standard error of $7287 is less than standard deviation (not shown) of $10,553.
Regression as Analysis of Variance

ANOVA conducts an $F$-test to determine whether variation in $Y$ is due to varying levels of $X$.

ANOVA is used to test for *significance of regression*:

$H_0$: slope coefficient = 0. Home size is *not* a significant variable
$H_1$: slope coefficient $\neq$ 0. Home size is a significant variable

Excel reports the $p$-value (*Significance F*).
Rejecting $H_0$ indicates that $X$ explains variation in $Y$.

$p$-value = $3.798 \times 10^{-8}$
- Reject $H_0$: The slope is not equal to zero. Using a linear relationship, home size is a significant variable in explaining variation in market value.
An alternate method for testing whether a slope or intercept is zero is to use a t-test:

\[ t = \frac{b_1 - 0}{\text{standard error}} \]

- Use p-values to draw conclusion
- Neither coefficient is statistically equal to zero.
Confidence Intervals for Regression Coefficients

- Confidence intervals (Lower 95% and Upper 95% values in the output) provide information about the unknown values of the true regression coefficients, accounting for sampling error.

- We may also use confidence intervals to test hypotheses about the regression coefficients.
  - To test the hypotheses, check whether $B_1$ falls within the confidence interval for the slope. If it does, reject the null hypothesis $H_0: \beta_1 = B_1$.

In Home Market Value data, a 95% confidence interval for the intercept is [14,823, 50,523], and for the slope, [24.59, 45.48].

Although we estimated that a house with 1,750 square feet has a market value of $32,673 + 35.036(1,750) =$93,986, if the true population parameters are at the extremes of the confidence intervals, the estimate might be as low as $14,823 + 24.59(1,750) =$57,855 or as high as $50,523 + 45.48(1,750) =$130,113.
Residual Analysis and Regression Assumptions

- **Residual** = Actual $Y$ value − Predicted $Y$ value
- **Standard residual** = residual / standard deviation
- Rule of thumb: Standard residuals outside of ±2 or ±3 are potential outliers.
- Excel provides a table and a plot of residuals.

This point has a standard residual of 4.53
A linear regression model with more than one independent variable is called a **multivariate linear regression model**.

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \]

where

- \( Y \) is the dependent variable,
- \( X_1, \ldots, X_k \) are the independent (explanatory) variables,
- \( \beta_0 \) is the intercept term,
- \( \beta_1, \ldots, \beta_k \) are the regression coefficients for the independent variables,
- \( \varepsilon \) is the error term
We estimate the regression coefficients—called partial regression coefficients — $b_0, b_1, b_2, \ldots b_k$, then use the model:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

The partial regression coefficients represent the expected change in the dependent variable when the associated independent variable is increased by one unit while the values of all other independent variables are held constant.
Predict student graduation rates using several indicators

Regression model below

The value of $R^2$ indicates that 53% of the variation in the dependent variable is explained by independent variables.

All coefficients are statistically significant.
Practical Issues in Trendline and Regression Modeling

- Identifying the best regression model often requires experimentation and trial and error.

- The independent variables selected should make sense in attempting to explain the dependent variable
  - Logic should guide your model development. In many applications, behavioral, economic, or physical theory might suggest that certain variables should belong in a model.

- Additional variables increase $R^2$ and, therefore, help to explain a larger proportion of the variation.
  - Even though a variable with a large p-value is not statistically significant, it could simply be the result of sampling error and a modeler might wish to keep it.

- Good models are as simple as possible (the principle of parsimony).
A Model with Categorical Variables

- Employee Salaries provides data for 35 employees

- Predict Salary using Age and MBA (code as yes=1, no=0)

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \]

where

- \( Y \) = salary
- \( X_1 \) = age
- \( X_2 \) = MBA indicator (0 or 1)
Salary = 893.59 + 1044.15 × Age + 14767.23 × MBA
- If MBA = 0, salary = 893.59 + 1044 × Age
- If MBA = 1, salary = 15,660.82 + 1044 × Age
Procedures to find the best model

- **Backward Elimination** begins with all independent variables in the model and deletes one at a time until the best model is identified.

- **Forward Selection** begins with a model having no independent variables and successively adds one at a time until no additional variable makes a significant contribution.

- **Stepwise Selection** is similar to Forward Selection except that at each step, the procedure considers dropping variables that are not statistically significant.

- **Sequential Replacement** replaces variables sequentially, retaining those that improve performance. These options might terminate with a different model.

- **Exhaustive Search** looks at all combinations of variables to find the one with the best fit, but it can be time consuming for large numbers of variables.
Checking Regression Assumptions

- **Linearity**
  - examine scatter diagram (should appear linear)
  - examine residual plot (should appear random)

- **Normality of Errors**
  - view a histogram of standard residuals
  - regression is robust to departures from normality

- **Homoscedasticity**: variation about the regression line is constant
  - examine the residual plot

- **Independence of Errors**: successive observations should not be related.
  - This is important when the independent variable is time.
Checking Regression Assumptions for the Home Market Value Data

- **Linearity** - linear trend in scatterplot
  - no pattern in residual plot
Checking Regression Assumptions for the Home Market Value Data

Normality of Errors – residual histogram appears slightly skewed but is not a serious departure.
Checking Regression Assumptions for the Home Market Value Data

- **Homoecedasticity** – residual plot shows no serious difference in the spread of the data for different $X$ values.
- **Independence of Errors** – Because the data is cross-sectional, we can assume this assumption holds.
Overfitting means fitting a model too closely to the sample data at the risk of not fitting it well to the population in which we are interested. In multiple regression, if we add too many terms to the model, then the model may not adequately predict other values from the population. Overfitting can be mitigated by using good logic, intuition, theory, and parsimony.