Module 4: Introduction to Regression

Dr. Tom Ilvento  
Dr. Mugdim Pašić

January 2006

Regression

- We are looking at the relationship between two or more variables
  - One is called the Dependent Variable (Y), which is to be modeled or predicted
  - The others are called Independent Variables (X or a set of Xs), which are used to predict Y

Types of Regression Models

- In a bivariate (two variable) case, one way to express the relationship is in terms of covariance and correlation
  - Expressed as a linear relationship
  - Regression is an extension of correlation/covariance

Let’s look at a quick example

- I recently went to Europe and had to deal with temperatures in Celsius
- How do I convert from C to F?
  - A friend once told me a quick “rule of thumb” was to double C and add 30

I made a quick data set using my calculator

<table>
<thead>
<tr>
<th>F</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-12.2222</td>
</tr>
<tr>
<td>15</td>
<td>-9.4444</td>
</tr>
<tr>
<td>20</td>
<td>-6.6667</td>
</tr>
<tr>
<td>25</td>
<td>-3.889</td>
</tr>
<tr>
<td>30</td>
<td>-1.1111</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>1.6667</td>
</tr>
<tr>
<td>45</td>
<td>7.2222</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>55</td>
<td>12.7778</td>
</tr>
<tr>
<td>60</td>
<td>15.5556</td>
</tr>
<tr>
<td>65</td>
<td>18.3333</td>
</tr>
<tr>
<td>70</td>
<td>21.1111</td>
</tr>
<tr>
<td>75</td>
<td>23.8889</td>
</tr>
<tr>
<td>80</td>
<td>26.6667</td>
</tr>
<tr>
<td>85</td>
<td>29.4444</td>
</tr>
<tr>
<td>90</td>
<td>32.2222</td>
</tr>
<tr>
<td>95</td>
<td>35</td>
</tr>
</tbody>
</table>
And a quick graph of the data

The correlation between F and C is 1.0 – a perfect linear relationship.

I ran a regression of F on C using Excel

Regression generated an equation \( F = 32 + 1.8C \)

Requirements of Regression

- **Dependent variable** Y is measured as a continuous variable - not a dichotomy or ordinal
- The independent variables can be continuous, dichotomies, or ordinal
- **Linear relationship** in the parameters in the form of: \( Y = a + b_1X_1 + b_2X_2 \)

Nonlinear relationships that are linear in their parameters

- It is possible to represent a nonlinear relationship with a linear approach, such as a Polynomial or Log function
- Log function - take the log of both sides
  \[ Y = aX^b \]
  \[ \log(Y) = a + b\log(X) \]
- Polynomial of the kth order
  \[ Y = a + b_1X + b_2X^2 + b_3X^3 + \ldots + b_kX^k \]

Equation of a line

- The equation of a line is given as:
  \[ Y = a + bX \]
  Where a is the intercept and b is the slope
- We specify a dependent variable Y, and independent variable X
  Note: in multiple regression there may be more than one x:
  \[ Y = a + b_1X_1 + b_2X_2 \]
- When referring to the population I will use Greek terms:
  \[ Y_i = \mu + \beta_1X_{i1} \]

Equation of a line

- \( Y = 5 + .5X \)
  - \( X=0 \) then \( Y=5 \) The intercept
  - \( X=10 \) then \( Y=10 \)
  - \( X=20 \) then \( Y=15 \)
  - \( X=30 \) then \( Y=20 \)
- The slope shows how much Y changes for a unit change in X
- This is a **deterministic model**
In reality, we often have a random component

- **A Probabilistic Model** has a deterministic component and a random error component, denoted as \( e_i \) or \( \epsilon_i \).
  - \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)

- Our Expectation of \( Y \) is the deterministic component
  - \( E(Y) = \beta_0 + \beta_1 X_i \)

The Error Term

- The error component is very important.
  - Observed in population/sample: \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)
  - Predicted from model: \( \hat{Y}_i = \beta_0 + \beta_1 X_i \)
  - The difference between what we predict and what we observe: \( \epsilon_i = \hat{Y}_i - Y_i \)

Have we seen the error term before? YES!!!!!

- Consider the following model for the mean
  - \( Y_i = \mu + \epsilon_i \)
  - A simple model based on the mean
  - Deviations about the mean: \( \epsilon_i = Y_i - \mu \)
  - Sum of Squared deviations: \( \sum \epsilon_i^2 = \sum(Y_i - \mu)^2 \)
  - Mean squared deviation: \( \sum \epsilon_i^2/n = \sum(Y_i - \mu)^2/n \)
  - Population Variance: \( \sum \epsilon_i^2/n = \sigma^2 \)

The Error Term

- The error term in regression is a measure of the:
  - Variance
  - Standard Deviation
  - And ultimately contributes to the estimate of the Standard Error for our coefficients

- We will assume equal variances for \( Y \) (dependent variable) across each level of \( X \) (independent variable)

- In essence we will pool the measure of the variance in regression

How to fit a line to our data?

- We will use the property of **Least Squares**
  - We will find estimates for \( \beta_0 \) and \( \beta_1 \) that will minimize the squared deviations about the fitted line
  - First an example, and then the details

Model Regressing \( Y \) on \( X \): Catalog.xls

- \( Y \) is the Dependent Variable
  - **SALES**
- \( X \) is the Independent Variable
  - **SALARY**

- The correlation between SALES and SALARY is .700
**Excel Steps – see handout**
- Organize data in columns
  - First column contains Y (dependent)
  - Remaining columns contain contiguous Xs (independent)
- TOOLS Data Analysis Regression
- Specify Y variable
- Specify X variables – need to be contiguous columns
- Remember to specify if first row has labels
- Specify output
- Modify the output
  - How many decimal places are showing (3 to 4)
  - Change headings to make them fit
  - Bold headers

**Example – CATALOGS.xls**
- Run the regression of SALES on SALARY
- We will walk through the output

**Excel will add a Trend Line which is based on regression**

**Estimated Regression Equation is**

\[ \text{SALES} = -15.332 + .022(\text{SALARY}) \]

- If SALARY = 0
  - \( \text{SALES} = -15.332 + .022(0) \)
  - \( \text{SALES} = -15.332 \)
- A unit change in SALARY ($1) results in a .022 change in SALES
  - This is better expressed as: $1,000 change in SALARY results in Sales of $22.00

**Regression of SALES on SALARY**
- Our prediction of SALES for a person with a SALARY of $50,000 is:
  - \( \text{SALES} = -15.332 + .022($50,000) \)
  - \( \text{SALES} = $1,084.67 \)
- I refer to this as solving the equation for a person with a salary of $50,000

**EXCEL Output**

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coef</th>
<th>Std Error</th>
<th>t Stat</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-15.332</td>
<td>0.374</td>
<td>-40.688</td>
</tr>
<tr>
<td>Salary</td>
<td>0.022</td>
<td>0.001</td>
<td>30.331</td>
</tr>
</tbody>
</table>
Least Squares Formulas for Bi-variate Regression (two variables $Y$ and $X$, one independent variable - $X$)

\[ \hat{\beta}_1 = \frac{SS_{XY}}{SS_X} \]

\[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \]

where $SS_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$

$SS_X = \sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$

A Few Points

- It is possible to predict outside the range of the data
  - When Salary = 0: $SALES = -15.332 + 0.022(0) = -$15.33
  - When Salary = 1,000,000: $SALES = -15.332 + 0.022(1,000,000) = $21,985
- The model parameters should be interpreted only within the sampled range of the independent variables
- The prediction part of our model is deterministic, but we know we will have some error - our prediction won’t match the data exactly

How to fit a line to our data?

- We will use the property of Least Squares
- We will find estimates for $B_0$ and $B_1$ that will minimize the squared deviations about the fitted line

Least Squares

- ‘Best Fit’ Means Difference Between Actual $Y$ Values & Predicted $Y$ Values Are a Minimum
  \[ SSE = \sum (Y_i - \hat{Y})^2 = \sum \hat{e}^2 = \text{minimum} \]
- Least Squares generates a set of coefficients that minimizes the Sum of the Squared Errors (SSE)

EXCEL Output

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

| Coef | Std Error | t Stat | P(>|t|) |
|------|-----------|--------|--------|
| Intercept | 15.332 | 48.054 | 0.330 | 0.740 |
| Salary   | 0.022 | 0.001 | 30.931 | 0.000 |

Regression Statistics

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>
**Parts of the Output – Regression Statistics**

- **Multiple R** – in a bivariate regression this is the absolute value of the correlation coefficient |r|. In a multivariate regression in is the square root of R².
- **R-Square** – same as we talked about. It varies from zero to one, where one indicates a perfect linear fit.
- **Adjusted R Square** – R square is adjusted for the number of independent variables in the model.
- **Standard Error** – The standard error of the model - the square root of the MSE.
- **Observations** - the number of observations.

**Excel Regression Output – ANOVA Table**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Sig F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>451624335.7</td>
<td>451624335.7</td>
<td>956.71</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>998</td>
<td>471117860.1</td>
<td>472062.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>999</td>
<td>922742195.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parts of the Output – ANOVA: Analysis Of Variance**

- Total Sum of Squares for Y: \[ SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

We decompose the Total Sum of Squares for Y into a part due to:

- **Regression** (or Model) – think of this as explained
  - 451624335.68
- **Residual** (or Error) – think of this as unexplained
  - 471117860.07

**Measures of Variation for Regression**

- \[ SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \] \( n-1 \) df
- \[ SSR = \sum_{i=1}^{k} (\hat{Y}_i - \bar{Y})^2 \] \( k \) df
- \[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \] \( n-k-1 \) df

Where \( n \) is the sample size.

**k** is the number of independent variables in the model.

**Diagram of the decomposition of the Total Sum of Squares**

- Total sum of squares \( (Y_i - \bar{Y})^2 \)
- Explained sum of squares \( (\hat{Y}_i - \bar{Y})^2 \)
- Unexplained sum of squares \( (Y_i - \hat{Y}_i)^2 \)
Diagram of the decomposition of the Total Sum of Squares

- Total sum of squares \((Y_i - \bar{Y})^2\)
- Unexplained sum of squares \((Y_i - \hat{Y})^2\)
- Explained sum of squares \((\hat{Y}_i - \bar{Y})^2\)

A note about R Square
- \(R^2 = \frac{SSR}{SSY}\)
- The Sum of Squares due to Regression divided by the Sum of Squares for Y (aka SS Total)
- What part of the total variability in Y is "explained" by knowing something about the independent variable(s)
- \(R^2 = 1 - \frac{SSE}{SSY}\)
- Shows the linear "fit of the model"
- Ranges from 0 to 1

Parts of the Output – df (Degrees of Freedom): \(k; n-k-1; n-1\)
- Overall, the degrees of freedom are \(n-1\)
- Think of \(k\) as the number of independent variables in the model
- The degrees of freedom for Regression is \(k\)
  - In our example, \(k = 1\) because we only have Salary as an independent variable
  - So, d.f. Regression = 1
- The degrees of freedom for Residual is \(n-k-1\)
  - The sample size minus the number of parameters estimated by the model (intercept and slope coefficients)
  - In our example, d.f. Residual = \(1000 - 1 - 1 = 998\)
  - The d.f. Regression + d.f. Residual = d.f. Total
  - In our example, 1 + 998 = 999

Parts of the Output – MS (Mean Squares)
- MS – Mean Squares
  - We divide the Sums of Squares by their respective degrees of freedom
  - MS Regression = \(\frac{SSR}{k}\)
    - \(451624335.68/1 = 451624335.68\)
  - MS Residual = \(\frac{SSE}{n-k-1}\)
    - \(471117860.07/(1000-2) = 472061.98\)
  - Think of these as "average deviations"

Mean Measures of Variation for Regression
- Mean \(SS_Y = \frac{\sum (Y_i - \bar{Y})^2}{n-1}\) Sample Variance
- Mean \(SSR = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{k}\) Mean Square Regression (MSR)
- Mean \(SSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1}\) Mean Square Error (MSE)

Root Mean Square Error
- The Root Mean Square Error is the Square Root of the MSE
  \[-\frac{\sum (Y_i - \hat{Y}_i)^2}{(n-k-1)}\] 
  \[= \sqrt{472061.98} = 687.068\]
- Excel calls this the "Standard Error" under Regression Statistics
Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple R</strong></td>
<td>0.700</td>
</tr>
<tr>
<td><strong>R Square</strong></td>
<td>0.489</td>
</tr>
<tr>
<td><strong>Adjusted R Square</strong></td>
<td>0.489</td>
</tr>
<tr>
<td><strong>Standard Error</strong></td>
<td>687.068</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>1000</td>
</tr>
</tbody>
</table>

Parts of the Output – ANOVA

- **F** - a very general test that none of the independent variables are significantly different from zero
  - It is the ratio of the MS due to Regression divided by the MS due to Residual
  - If the two MS’s are equal to each other, the ratio should be about or near one
  - If there is only one independent variable, the F-Test = (t-test)$^2$, i.e., $F = t^2$
  - The null and alternative hypothesis for the F-test is
    - $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
    - $H_a$: at least one $\beta_i \neq 0$
  - Focus on the Significance F (p-value) < .05

Excel Regression Output

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std 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>-15.332</td>
<td>45.374</td>
<td>-0.338</td>
<td>0.736</td>
<td>-104.373</td>
<td>73.708</td>
</tr>
<tr>
<td>Salary</td>
<td>0.022</td>
<td>0.001</td>
<td>30.931</td>
<td>0.000</td>
<td>0.021</td>
<td>0.023</td>
</tr>
</tbody>
</table>

\[
\hat{Y} = -15.33 + .022(\text{Salary})
\]

The Last Part shows

- **Coefficients** that we estimate
  - The Intercept of the line
  - The slope coefficient of each independent variable
  - Their Standard Error of each coefficient
  - The t-statistic or $t$
    - This test is based on a Null Hypothesis where the slope coefficient is equal to zero
    - The p-value associated with $t$
      - Probability of finding a value of $t$ or greater given a Null Hypothesis of the coefficient equal to zero for a two-tailed test
    - A 95% Confidence Interval around each coefficient

Regression Coefficient Confidence Interval

- The confidence interval for a coefficient is similar to the C.I. for the mean
- We have an estimate, plus or minus a component that is a function of a t-value and a standard error of the estimate
- It places a Bound of Error around our estimate
- Excel gives the upper and lower bound, based on a $t$-value

\[
\beta \pm t_{\alpha/2, (n-k) df} \cdot (\text{standard error})
\]

The meaning of our coefficients

- Intercept or estimated $\beta_0$
  - The value of the Dependent variable if all independent variables equal zero
  - When using dummy variables, the intercept is the mean of the reference category
  - If a customer has no salary, the sales are \(-\$15.33\)
**The meaning of the coefficients**

- Slope or estimated $\beta_1$
  - The change in Y for a unit change in X

- For every dollar increase in Salary, sales increase by $.022

- For every $1,000 increase in Salary, Sales increase by $22

**Linear Regression Assumptions**

- Mean of Probability Distribution of Error is 0
- Probability Distribution of Error Has Constant Variance $= \sigma^2$
- Probability Distribution of Error is Normally distributed
- Errors Are Independent – they are uncorrelated with each other

**Symmetry**

- A correlation coefficient is a symmetrical measure of association
  - The correlation between Y and X is the same as the correlation between X and Y
  - The order doesn’t matter and neither is established as the dependent or the independent variable
- A regression coefficient is **not symmetrical**
  - The slope and intercept resulting from a regression of Y on X
  - Is not the same as a regression of X on Y

**Inference and Regression**

- We have two main inferential tests with regression
  - F-test for the overall model
  - t-test for individual coefficients

- We can also generate a confidence interval for our coefficients - Excel will give us this without even asking
- We will focus on general conclusions by looking at p-values

**Try it and see – run the regression of Salary on Sales**

- $SALES = -15.33 + .022(SALARY)$
- $SALARY = 28,986.27 + 22.29(SALES)$

- The correlation between them is .700, regardless of which is first or second.
- R Square also remains the same in the two models.
- But the coefficients change
- In regression, it does matter which is the dependent variable and which is the independent variable
- We typically say we "regress Y on a set of X independent variables"

**F-test**

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Sig F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>451624335.7</td>
<td>451624335.7</td>
<td>956.71</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>998</td>
<td>471117860.1</td>
<td>472062.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>999</td>
<td>922742195.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \frac{451624335.7}{472062.0} = 956.71$$
**F-Test – a first stop test**

- **F-Test**
  
  \[ F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} \]

  - very general test that none of the independent variables are significantly different from zero
  - If there is only one independent variable, the F-Test = (t-test)²  i.e., \( F = t^2 \)

**Excel Regression Output**

<table>
<thead>
<tr>
<th>Coef.</th>
<th>Std 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>-15.332</td>
<td>4.374</td>
<td>-3.538</td>
<td>0.038</td>
<td>-30.243</td>
</tr>
<tr>
<td>Salary</td>
<td>0.022</td>
<td>0.001</td>
<td>20.931</td>
<td>0.000</td>
<td>0.021</td>
</tr>
</tbody>
</table>

\[ \hat{Y} = -15.33 + .022(\text{Salary}) \]

The estimated \( \beta \) is .022

**Root Mean Square Error**

- The Root Mean Square Error is the Square Root of the MSE:
  
  \[ \text{MSE} = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n - k - 1} \]

**Excel calls this the "Standard Error" under Regression Statistics**

**Excel Regression Output from Catalog Data**

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th>687.068</th>
<th>( \sqrt{472062} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.700</td>
<td></td>
</tr>
<tr>
<td>R Square</td>
<td>0.489</td>
<td></td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.489</td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td>687.068</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>( F )</th>
<th>Sig F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>451624335.7</td>
<td>451624335.7</td>
<td>956.71</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>998</td>
<td>471117660.1</td>
<td>472062.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>999</td>
<td>922742195.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Standard Error of the Estimated Regression Equation**

- Remember we said the error term of our model is related to the variance (thus the standard deviation) and the standard error
  - And that we assumed constant error variance across all levels of the independent variable X
  - So the **Standard Error** of the Model is given as
    \[ s = \sqrt{\frac{\text{SSE}}{(n-k-1)}} = \text{Root MSE} \]
Standard Error of the Slope in a Bivariate Regression

- It is based on the Root MSE
- And the total sum of squares for the independent variable (the variability of X)

\[ \sigma_{\hat{\beta}_1} = \sigma \div \sqrt{SS_X} \]

Standard Error for \( \hat{\beta}_1 = \frac{\text{Root MSE}}{\sqrt{SS_X}} \)

Test of Slope Coefficient

- Is there a Linear Relationship Between X & Y?
- Involves testing the sample estimate of the slope coefficient, \( \beta_1 \)
- Hypotheses
  - \( H_0: \beta_1 = 0 \) (No Linear Relationship)
  - \( H_1: \beta_1 \neq 0 \) (Linear Relationship)
- The theoretical basis of the test is the Sampling Distribution of the slope coefficient

Sampling Distribution of Sample Slopes

- All Possible Sample Slopes
  - Sample 1: 2.5
  - Sample 2: 1.6
  - Sample 3: 1.8
  - Sample 4: 2.1
- Very large number of sample slopes

The expected value of \( \hat{\beta}_1 \) is \( \beta_1 \)

Hypothesis Test for a slope coefficient for SALARY, use \( \alpha = .05 \)

- Null hypothesis
- Alternative
- Test Statistic
  - Calculation: \( t = \frac{(.02196-0)}{.00071} \)
  - \( t = 30.9296 \)
- P-value: \( P = .000 \)
- Conclusion
  - Reject \( H_0: \beta_1 = 0 \)

Excel Regression Output

<table>
<thead>
<tr>
<th>Coef.</th>
<th>Std Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-15.332</td>
<td>45.374</td>
<td>-0.338</td>
</tr>
<tr>
<td>Salary</td>
<td>0.02196</td>
<td>0.00071</td>
<td>30.93066</td>
</tr>
</tbody>
</table>

\( t = \frac{(.02196-0)}{.00071} = 30.93066 \)
We are looking for a t-value that is larger than 2!
We are looking for a p-value less than .05!

Dummy Variable Regression

Customer Sales By Age Categories

\[ y = 480.21x + 295.72 \]
\[ R^2 = 0.323 \]

<table>
<thead>
<tr>
<th>AGE Category</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5000</td>
</tr>
<tr>
<td>2</td>
<td>6000</td>
</tr>
<tr>
<td>3</td>
<td>7000</td>
</tr>
<tr>
<td>4</td>
<td>8000</td>
</tr>
<tr>
<td>5</td>
<td>9000</td>
</tr>
</tbody>
</table>

We are looking for a t-value that is larger than 2!
We are looking for a p-value less than .05!


Dummy Variable Regression

- Independent variables in regression can be measured on a continuous, ordinal, or categorical level
- Dummy variable regression uses dummy variables
  - With k categories of a variable
  - We will use k-1 dummy variables, each using a zero or one to represent the presence of an attribute
- Example: Age measured with three categories - <31 years, 31 to 55 years, 56 and over - is represented with two dummy variables
  - AGE1 has a 1 if <31 and zero for all else;
  - AGE2 has a 1 if 31 to 55, zero for all else;
  - The left out category, 56 and over, is called the Reference Category

You try it with me

- Open up CATALOGS REV.xls
- Run a regression of SALES on AGE1 and AGE2
- Put the output in a new worksheet – Dummy Reg
- Dress up the output

Excel Output – what do you see?

SUMMARY OUTPUT

R Square is .19
F-test is significant

Regression Statistics

Multiple R 0.436
R Square 0.190
Adjusted R Square 0.188
Standard Error 865.985
Observations 1000

ANOVA
df SS MS F Sig F
Regression 2 175061377.66 87530688.83 116.72 0.000
Residual 997 747680818.08 749930.61
Total 999 922742195.74

Focus on the coefficients

\[ \hat{Y} = 1432.13 - 873.51(\text{AGE1}) + 69.56(\text{AGE2}) \]

Focus on the coefficients

<table>
<thead>
<tr>
<th></th>
<th>Coef</th>
<th>Std Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1432.13</td>
<td>60.483</td>
<td>23.678</td>
<td>0.000</td>
</tr>
<tr>
<td>Age1</td>
<td>-873.51</td>
<td>79.191</td>
<td>-11.030</td>
<td>0.000</td>
</tr>
<tr>
<td>Age2</td>
<td>69.56</td>
<td>71.655</td>
<td>0.971</td>
<td>0.332</td>
</tr>
</tbody>
</table>

1. The intercept is $1,432.13
2. The coefficient for AGE1 is negative (-$873.51)
3. The t-test for AGE1 is large (-11.030) and the p-value is small (< .000) – we can reject a null hypothesis that the coefficient is really zero
4. The coefficient for AGE2 is positive ($69.56), but not significant (t = .971 and p-value = .332)

What do you see?

- This is a very simple model
  - We expect age to be related to sales, but not explain everything!
  - R Square is .190 – about 19% of the variability in Sales is explained by the customer’s age
- Focus on the F-test
  - F = 116.72 p < .000
  - Age is related to Sales
- Focus on coefficients and t-tests

Hypothesis Test for a slope coefficient for AGE2,

- Null hypothesis \( H_0: \beta_2 = 0 \)
- Alternative \( H_1: \beta_2 \neq 0 \) two-tailed test
- Test Statistic \( t = (69.56-0)/71.66 = .971 \)
- p-value \( p = .332 \)
- Conclusion Cannot Reject \( H_0: \beta_2 = 0 \)

We do not have evidence to suggest that the coefficient for AGE2 is any different than zero.
Since our independent variables are dummy variables, it is easy to solve the equation:

- When \( \text{AGE1} = 1 \) and \( \text{AGE2} = 0 \)
  - \( \text{Estimate} = 1,432.13 - 873.51(1) + 69.56(0) = $558.62 \)
- When \( \text{AGE1} = 0 \) and \( \text{AGE2} = 1 \)
  - \( \text{Estimate} = 1,432.13 - 873.51(0) + 69.56(1) = $1,501.69 \)
- When \( \text{AGE1} = 0 \) and \( \text{AGE2} = 0 \)
  - The Reference Group, \( \text{AGE3}! \)
  - \( \text{Estimate} = 1,432.13 - 873.51(0) + 69.56(0) = $1,432.13 \)
  - This represents \( \text{AGE3}! \)

**Dummy Variable Regression**
- The equation predicts the mean level for each group.
- The intercept represents the mean level for the Reference Category.
- The t-tests represent whether the other categories are significantly different from the Reference Category.
- It does not matter much which category is the Reference Group!

**Excel Output**
- \( \hat{Y} = 1,501.69 - 69.56(\text{AGE3}) - 943.06(\text{AGE1}) \)
- Since our independent variables are dummy variables, it is easy to solve the equation:
  - When \( \text{AGE3} = 1 \) and \( \text{AGE1} = 0 \)
    - \( \text{Estimate} = 1,501.69 - 69.56(1) - 943.06(0) = $1,432.13 \)
  - When \( \text{AGE1} = 1 \) and \( \text{AGE3} = 0 \)
    - \( \text{Estimate} = 1,501.69 - 69.56(0) - 943.06(1) = $558.63 \)
  - When \( \text{AGE1} = 0 \) and \( \text{AGE3} = 0 \)
    - The Reference Group, \( \text{AGE2}! \)
    - \( \text{Estimate} = 1,501.69 - 69.56(0) - 943.06(0) = $1,501.69 \)
  - This represents \( \text{AGE2}! \)

Let's run it with two independent variables, both of which are dummy variables:
- Run a regression of Sales on:
  - Age1
  - Age2
  - Gender (1=male)
- The Reference Category will now represent older females, a combination of both variables.

**Excel Output**
- \( \hat{Y} = 1322.50 - 883.40(\text{AGE1}) + 3.39(\text{AGE2}) + 295.71(\text{Gender}) \)
  - Degrees of freedom for regression reflects 3 independent variables.
Since our independent variables are dummy variables, it is easy to solve the equation:

When \( AGE1 = 1 \) \( Gender = 1 \) \( AGE2 = 0 \)
- Estimate = 1,322.50 – 883.40(1) + 3.39(0) + 295.71(1) = $734.81

When \( AGE1 = 1 \) \( Gender = 0 \) \( AGE2 = 0 \)
- Estimate = 1,322.50 – 883.40(1) + 3.39(0) + 295.71(0) = $439.10

When \( AGE2 = 1 \) \( Gender = 1 \) \( AGE1 = 0 \)
- Estimate = 1,322.50 – 883.40(0) + 3.39(1) + 295.71(1) = $1,621.60

When \( AGE2 = 1 \) \( Gender = 0 \) \( AGE1 = 0 \)
- Estimate = 1,322.50 – 883.40(0) + 3.39(1) + 295.71(0) = $1,325.89

When \( AGE1 \) \( Gender = 1 \) and \( AGE2 = 0 \)
- Estimate = 1,322.50 – 883.40(0) + 3.39(0) + 295.71(1) = $1,618.21

When \( AGE1 \) \( Gender = 0 \) and \( AGE2 = 0 \)
- Estimate = 1,322.50 – 883.40(0) + 3.39(0) + 295.71(0) = $1,322.50

Summary of our model
- Overall, weak model \( R^2 = 0.21 \) (21% of the variability in Sales is explained by knowing age and gender).
- Not too surprising, there is more to sales than age and sex!
- Younger customers spend less on average than older customers (-$893), middle and older customers spend about the same.
- Men, on average, spend about $296 per year more than women.
- All estimates are based on controlling for the other variables in the model!

Summary of our model
- Overall, weak model \( R^2 = 0.21 \) (21% of the variability in Sales is explained by knowing age and gender).
- Not too surprising, there is more to sales than age and sex!
- Younger customers spend less on average than older customers (-$893), middle and older customers spend about the same.
- Men, on average, spend about $296 per year more than women.
- All estimates are based on controlling for the other variables in the model!

Excel Output

<table>
<thead>
<tr>
<th>SUMMARY OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression Statistics</td>
</tr>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>ANOVA</td>
</tr>
<tr>
<td>df</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Coef.</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>YEAR</td>
</tr>
</tbody>
</table>

Trend Analysis
- A simple approach to data over time is a trend analysis.
- The independent variable is a time element - year, quarter, or simply a numerical fill (1 to n).
- We are looking to see if the trend is linear.
- We should expect:
  - A high R Square (.8 or higher shows a good fit)
  - A significant t-test for the trend.

Open ManSales.xls
- Manufacturing sales from 1950 to 1991
- Plot SALES versus YEAR
  - Add a linear Trend Line
  - Add the equation and R Square
  - Run the regression of SALES on YEAR.
Look at the Chart

Our model fits well, but...
At times we miss with a linear trend

Nonlinear approach - Polynomial

- Click on the chart line and delete it.
- Grab the whole chart and under the Chart menu, add a new trendline with the equation and R square.
  - Polynomial 2nd order
  - This adds a Year squared term to the model.

- We will run a new regression.
  - Add a new column before SALES.
  - Label the column YRSQ.
  - Insert the square of YEAR in this column.
  - Now run a regression of SALES on Year and YRSQ.

The fit is much better – visually and R Square increases to .99

Excel regression output

Key Questions in Polynomial Regression

Does the model fit better?
- Yes, R² increases to .991

Does the model need YRSQ?
- Yes, because the coefficient for YRSQ is significantly different from zero.