REGRESSION I: One Independent Variable

Tom Ilvento
FREC 408

Regression
- We are looking at the relationship between two or more variables
  - One is called the **Dependent Variable** \((Y)\), which is to be predicted (Def10.1 p536)
  - The others are called **Independent Variables** \((Xs)\), which are used to predict \(Y\) (Def10.1 p537)

**Review**
- 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

Types of Regression Models

- **Simple**
  - Linear
  - Non-Linear
- **Multiple**
  - Linear
  - Non-Linear

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>24.4444</td>
</tr>
<tr>
<td>20</td>
<td>36.6667</td>
</tr>
<tr>
<td>25</td>
<td>48.8889</td>
</tr>
<tr>
<td>30</td>
<td>61.1111</td>
</tr>
<tr>
<td>35</td>
<td>73.3333</td>
</tr>
<tr>
<td>40</td>
<td>85.5556</td>
</tr>
<tr>
<td>45</td>
<td>97.7778</td>
</tr>
<tr>
<td>50</td>
<td>110.0000</td>
</tr>
<tr>
<td>55</td>
<td>122.2222</td>
</tr>
<tr>
<td>60</td>
<td>134.4444</td>
</tr>
<tr>
<td>65</td>
<td>146.6667</td>
</tr>
<tr>
<td>70</td>
<td>158.8889</td>
</tr>
<tr>
<td>75</td>
<td>171.1111</td>
</tr>
<tr>
<td>80</td>
<td>183.3333</td>
</tr>
<tr>
<td>85</td>
<td>195.5556</td>
</tr>
<tr>
<td>90</td>
<td>207.7778</td>
</tr>
<tr>
<td>95</td>
<td>219.9999</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 $C$

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
  - Although it is possible to represent a nonlinear relationship with a linear approach
  - Polynomial, log

Nonlinear relationships that are linear in their parameters

- Log function
  - $Y = aX^b$
  - $\log(Y) = a + b \log(X)$
- Polynomial
  - $Y = a + bX + bX^2$

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 $X$
  - $Y = a_1X_1 + a_2X_2$

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

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

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

The Error Term

- The error term in regression is a measure of the:
  - **Variance**
  - **Standard Deviation**
  - And ultimately the **Standard Error**
- 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** (page 543-45)
- 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 \( X \)

- \( Y \) is the Dependent Variable
  - Blood Pressure
- \( X \) is the Independent Variable
  - Weight
- The correlation between BLOOD PRESSURE and WEIGHT is .482
Excel will add a Trend Line based on a regression.

Estimated Regression Equation is:
- BLOOD PRESSURE = 88.894 + .187(WEIGHT)
- If WEIGHT = 0
  - BLOOD PRESSURE = 88.894 + .187(0)
  - BLOOD PRESSURE = 88.894

- A unit change in WEIGHT results in a .187 change in BLOOD PRESSURE.

Regression of Blood Pressure on Weight:
- Our prediction of Blood Pressure for a person of weight 175 pounds is:
  BLOOD PRESSURE = 88.894 + .187(175)
  BLOOD PRESSURE = 121.619

I refer to this as solving the equation for a person weighing 175 pounds.

Least Squares Formulas for Bi-variate Regression:

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

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

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

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

Cocoon Temperature Example:
- Cocoon Temp = a + b(Air Temp)
- \( SS_{XY} = 100.083 \)
- \( SS_X = 83.24 \)
- Mean Air Temp = 4.275
- Mean Cocoon Temp = 8.508
- \( \hat{\beta}_1 = \frac{100.083}{83.342} = 1.20 \)
- \( \hat{\beta}_0 = 8.508 - 1.20(4.275) = 3.378 \)
- \( \hat{Y}_i = 3.378 + 1.20X_i \)
### A Few Points

- It is possible to predict outside the range of the data (or the experiment).
- When temp = 20: \(3.378 + 1.20(20) = 27.378\)
- The model parameters should be interpreted only within the sampled range of the independent variable.
- 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
  \[ \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum \hat{\epsilon}^2 \text{ = minimum} \]
- Least Squares generates a set of coefficients that minimizes the Sum of the Squared Errors (SSE)

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

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

**Model Regressing SAT (Y) on Percent Taking (X)**

- Y is the Dependent Variable
- X is the Independent Variable
- The correlation between SATOTAL and TAKING is -.89

**Scatter Plot of SATOTAL vs TAKING**

**Descriptive Statistics of SAT TOTAL and % TAKING**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>% Taking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1065.98</td>
<td>37.00</td>
</tr>
<tr>
<td>Standard Error</td>
<td>9.51</td>
<td>3.87</td>
</tr>
<tr>
<td>Median</td>
<td>1050.00</td>
<td>34.00</td>
</tr>
<tr>
<td>Mode</td>
<td>983.00</td>
<td>8.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>67.88</td>
<td>27.62</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>4607.86</td>
<td>763.00</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.12</td>
<td>-1.67</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.32</td>
<td>0.17</td>
</tr>
<tr>
<td>Range</td>
<td>245.00</td>
<td>76.00</td>
</tr>
<tr>
<td>Minimum</td>
<td>954.00</td>
<td>4.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>1199.00</td>
<td>80.00</td>
</tr>
<tr>
<td>Sum</td>
<td>54365</td>
<td>1887</td>
</tr>
<tr>
<td>Count</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

**Excel Steps**

- 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
- I modify the output
  - How many decimal places are showing (3 to 4)
  - Change Headings to make them fit
  - Bold Headers
### Excel Regression 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>

#### ANOVA

<table>
<thead>
<tr>
<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>180803.007</td>
<td>180803.007</td>
<td>178.652</td>
</tr>
<tr>
<td>Residual</td>
<td>49</td>
<td>49589.974</td>
<td>1012.040</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>230393.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Coef Std Error t Stat P-value Lower 95% Upper 95%

| Intercept | 1146.529 | 7.494 | 152.992 | 0.000 | 1131.469 | 1161.589 |
| % Taking  | -2.177   | 0.163 | -13.366 | 0.000 | -2.504   | -1.850   |

### 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 it is the square root of $R^2$.
- **R-Square** – same as we talked about.
- **Adjusted R Square** – 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.

### A note about R Square

- $R^2 = SSR/SSTotal$
- $R^2 = 1 – SSE/SSTotal$
- Shows the linear "fit of the model".
- How much we explain of the dependent variable by knowing something about the independent variable(s).
- Ranges from 0 to 1.

### Excel Regression Output

#### ANOVA Table in Regression

<table>
<thead>
<tr>
<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>180803.007</td>
<td>180803.007</td>
<td>178.652</td>
</tr>
<tr>
<td>Residual</td>
<td>49</td>
<td>49589.974</td>
<td>1012.040</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>230393.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### F-Test (using $F^*$)

- Very general test that none of the independent variables are significantly different from zero.
- It is only one independent variable, the F-Test = (t-test)$^2$ i.e., $F^* = t^2$. 

---

7
F-Test
- The null and alternative hypothesis for the F-test is
  - \( H_0: \beta_1 = \beta_2 = \beta_k = 0 \)
  - \( H_a: \text{at least one } \beta_i \neq 0 \)

Excel Regression 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>1146.529</td>
<td>7.494</td>
<td>0.000</td>
</tr>
<tr>
<td>% Taking</td>
<td>2.177</td>
<td>0.163</td>
<td>-13.366</td>
</tr>
</tbody>
</table>

\[ \hat{Y} = 1146.529 - 2.177(TAKING) \]

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 \( t^* \)
  - 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

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 no one takes the test, the average state SAT score is 1146.53

The meaning of the coefficients
- Slope or estimated \( \beta_1 \)
  - The change in \( Y \) for a unit change in \( X \)
- For every percent increase in the students who take the SAT, the average state SAT score drops by 2.18 points

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 Normal
- 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

Cocoon and Air Temp example

- Cocoon temp = 3.375 + 1.201 Air Temp
- Air temp = -2.403 + .785 Cocoon temp
- The correlation between them is .971, regardless of which is first or second
- 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"