

REGRESSION I: One Independent Variable

Tom Ilvento
FREC 408

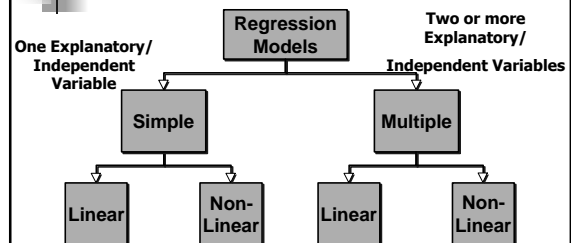
Regression

- 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



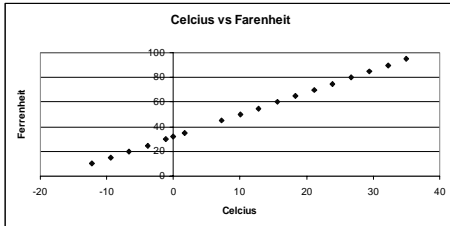
Let's look at a quick example

- 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

F	C
10	-12.2222
15	-9.4444
20	-6.6667
25	-3.8889
30	-1.1111
32	0
35	1.6667
45	7.2222
50	10
55	12.7778
60	15.5556
65	18.3333
70	21.1111
75	23.8889
80	26.6667
85	29.4444
90	32.2222
95	35

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

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	1				
R Square	1				
Adjusted R Square	1				
Standard Error	7.1E-05				
Observations	18				
ANOVA					
	df	SS	MS	F	
Regression	1	12372.94	12372.94	2456783461491.05	
Residual	16	0.00	0.00		
Total	17	12372.94			
		<i>Coef</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	32	0.00	1519489.65		0.00
C	1.8	0.00	1567412.98		0.00

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
 - $Y = aX^b$
 - $\text{Log}(Y) = a + b \cdot \text{Log}(X)$
- 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
 - $Y = a + b_1X_1 + b_2X_2$

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
 - ϵ_{i1}
 - $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_{i1}$
- Our Expectation of Y is the deterministic component
 - $E(Y) = \beta_0 + \beta_1 X_1$

The Error Term

- The error component is very important

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \quad \text{Observed in population/sample}$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} \quad \text{Predicted from model}$$

$$\epsilon_i = \hat{Y}_i - Y_i \quad \text{The difference between what we observe and what we predict}$$

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

- Consider the following model for the mean

$$Y_i = \mu + \epsilon_i$$

$$\epsilon_i = Y_i - \mu \quad \text{Deviations about the mean}$$

$$\sum \epsilon_i^2 = \sum (Y_i - \mu)^2 \quad \text{Sum of Squared deviations}$$

$$\sum \epsilon_i^2 / n = \sum (Y_i - \mu)^2 / n \quad \text{Mean squared deviation}$$

$$\sum \epsilon_i^2 / n = \sigma^2 \quad \text{Population Variance}$$

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

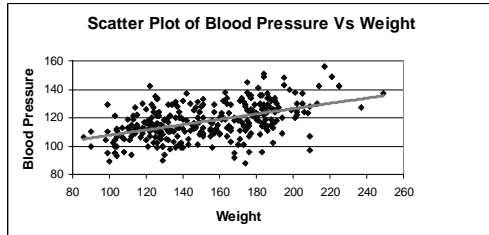
How to fit a line to our data?

- We will use the property of **Least Squares** (page543-45)
- We will find estimates for B_0 and B_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:

$$\text{BLOOD PRESSURE} = 88.894 + .187(175)$$

$$\text{BLOOD PRESSURE} = \mathbf{121.619}$$

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

EXCEL Output

Multiple R	0.4822					
R Square	0.2325					
Adjusted R Square	0.2301					
Standard Error	10.7638					
Observations	312					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	10882.5441	10882.5441	93.9283	0.0000	
Residual	310	35916.6354	115.8601			
Total	311	46799.1795				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	88.8942	3.0087	29.5461	0.0000	82.9742	94.8142
WEIGHT	0.1872	0.0193	9.6917	0.0000	0.1492	0.2252

Least Squares Formulas for Bi-variate Regression

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X}$$

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

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

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

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
- $\hat{\beta}_1 = \frac{100.083}{83.342} = 1.20$
- $\hat{\beta}_0 = 8.508 - 1.20(4.275) = 3.378$
- $\hat{Y}_i = 3.378 + 1.20X_i$

Cocoon Temperature Example

Prediction Table			Predicted Cocoon Temp
Intercept	b	Air Temp	
3.378	1.20	0	
3.378	1.20	2	
3.378	1.20	4	
3.378	1.20	6	
3.378	1.20	8	
3.378	1.20	10	

Cocoon Temperature Example

Prediction Table			Predicted Cocoon Temp
Intercept	b	Air Temp	
3.378	1.20	0	3.378
3.378	1.20	2	
3.378	1.20	4	
3.378	1.20	6	
3.378	1.20	8	
3.378	1.20	10	

Cocoon Temperature Example

Prediction Table			Predicted Cocoon Temp
Intercept	b	Air Temp	
3.378	1.20	0	3.378
3.378	1.20	2	5.778
3.378	1.20	4	
3.378	1.20	6	
3.378	1.20	8	
3.378	1.20	10	

Cocoon Temperature Example

Prediction Table			Predicted Cocoon Temp
Intercept	b	Air Temp	
3.378	1.20	0	3.378
3.378	1.20	2	5.778
3.378	1.20	4	8.178
3.378	1.20	6	10.578
3.378	1.20	8	12.978
3.378	1.20	10	15.378

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

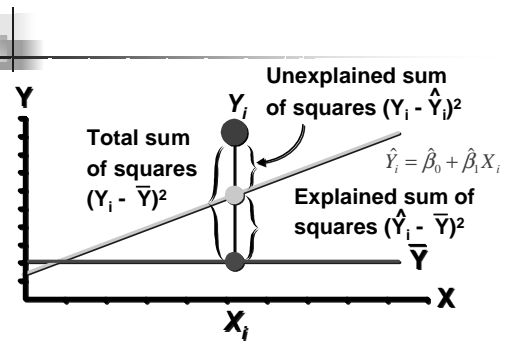
Least Squares

(page543)

- 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum

$$\sum_{i=1}^n (Y_i - \hat{Y})^2 = \sum_{i=1}^n \hat{\epsilon}^2 = \text{minimum}$$

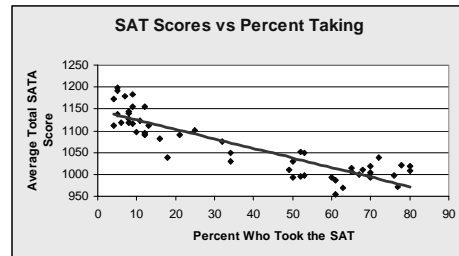
- Least Squares generates a set of coefficients that minimizes the Sum of the Squared Errors (SSE)



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

- Y is the Dependent Variable
 - State average SAT Score in 1999 - **SATOTAL**
- X is the Independent Variable
 - Percent of high school seniors who took the SAT - % **TAKING**
- The correlation between SATOTAL and TAKING is -.89

Scatter Plot of SATOTAL vs TAKING



Descriptive Statistics of SAT TOTAL and % TAKING

	Total	% Taking
Mean	1065.98	37.00
Standard Error	9.51	3.87
Median	1050.00	34.00
Mode	993.00	8.00
Standard Deviation	67.88	27.62
Sample Variance	4607.86	763.00
Kurtosis	-1.12	-1.67
Skewness	0.32	0.17
Range	245.00	76.00
Minimum	954.00	4.00
Maximum	1199.00	80.00
Sum	54365	1887
Count	51	51

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						
Regression Statistics						
Multiple R	0.886					
R Square	0.785					
Adjusted R Square	0.780					
Standard Error	31.813					
Observations	51					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>	
Regression	1	180803	180803	178.652	5.7757E-18	
Residual	49	49589.97	1012.04			
Total	50	230393				
	<i>Coef</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
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

Excel Regression Output

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.886
R Square	0.785
Adjusted R Square	0.780
Standard Error	31.813
Observations	51.000

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					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>
Regression	1	180803.007	180803.007	178.652	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

ANOVA Table in Regression

- Same as in ANOVA
- But the terms might change
 - SST is now Regression or Model
 - SSE is now Residual or Error
- F-Test (using F^*)
 - 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., $F^* = t^{*2}$

F-Test

- The null and alternative hypothesis for the F-test is
 - $H_0: \beta_1 = \beta_2 = \beta_k = 0$
 - H_a : at least one $\beta_i \neq 0$

Excel Regression Output

	Coefficients	Standard Error	t Stat	P-value
Intercept	1146.529	7.494	152.992	0.000
% Taking	-2.177	0.163	-13.366	0.000

$$\hat{Y} = 1146.529 - 2.177(\text{TAKING})$$

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 β_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 β_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 = σ^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 = $3.375 + 1.201 \text{ Air Temp}$
- Air temp = $-2.403 + .785 \text{ Cocoon temp}$
- 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**"