

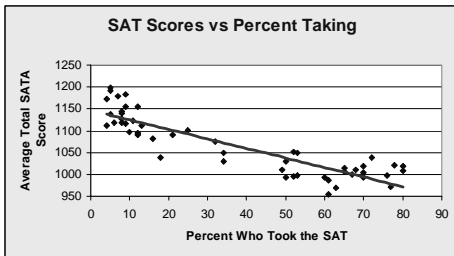
REGRESSION II: Hypothesis Testing in Regression

Tom Ilvento
FREC 408

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



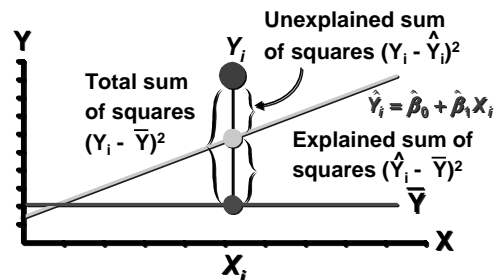
Excel Regression Output

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.886					
R Square	0.785					
Adjusted R Square	0.780					
Standard Error	31.813					
Observations	51					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>	
Regression	1	180803.007	180803.007	178.652	5.7757E-18	
Residual	49	49589.974	1012.040			
Total	50	230392.980				
	<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

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			

A look at the sources of Variation in the Model



Measures of Variation for Regression

- $SS_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$ **n - 1 df**
- $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ **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

Mean Measures of Variation for Regression

- Mean $SS_Y = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$ **Sample Variance**
- Mean $SSR = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{k}$ **Mean Square Regression (MSR)**
- Mean $SSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n-k-1)}$ **Mean Square Error (MSE)**

What is k in the formulas?

- The book uses k as the number of **independent variables**, or the number of slope coefficients
 - $\beta_1, \beta_2, \dots, \beta_k$
- Some books use it to represent the number of **parameters estimated**, which includes the intercept coefficient
 - $\beta_0, \beta_1, \beta_2, \dots$

Be careful with the notation!

So here is what you should remember

- **Total Sum of Squares**
 - SS_Y has n-1 degrees of freedom
- **Sum of Squares Regression**
 - SSR has # independent variables = k degrees of freedom
- **Sum of Squares Error**
 - SSE has n- # parameters estimated = n-(k+1) degrees of freedom

ANOVA Table example from Excel with 1 independent variable

ANOVA	df	SS	MS	F	Sig. F
Regression	1	180803.007	180803.007	178.652	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

Source	Source	Degrees of Freedom
Regression	SSR	# independent variables k
Residual	SSE	n- # parameters estimated (including intercept) n-(k+1)
Total	SSy	n-1

F - test

ANOVA	df	SS	MS	F	Sig. F
Regression	1	180803.007	180803.007	178.652	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

$$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}$$

F-Test

- F-Test (using F*)

- $F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}$

- 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

F-Test

- The null and alternative hypothesis for the F-test is

- $H_0: \beta_1 = \beta_2 = \beta_k = 0$
 - $H_a: \text{at least one } \beta_i \neq 0$
 - T.S. $F^* = 178.652$
 - Compare with table F with 1 and 49 d.f. at a specified α level (e.g., .05)
 - Or look at the p-value of .000
 - Conclusion???

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 estimated β is -2.177

Root Mean Square Error

- The Root Mean Square Error is the Square Root of the MSE =

$$\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{(n - k - 1)}}$$

Excel calls this the "Standard Error" under Regression Statistics

Excel Regression Output from SAT Data

Regression Statistics		
Multiple R	0.886	
R Square	0.785	$31.813 = \sqrt{1012.040}$
Adjusted R Square	0.780	
Standard Error	31.813	
Observations	51.000	

ANOVA					
	df	SS	MS	F	Sig. F
Regression	1	180803.007	180803.007	178.652	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

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

$$s = \sqrt{\frac{SSE}{(n - k - 1)}} = \text{Root MSE}$$

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)

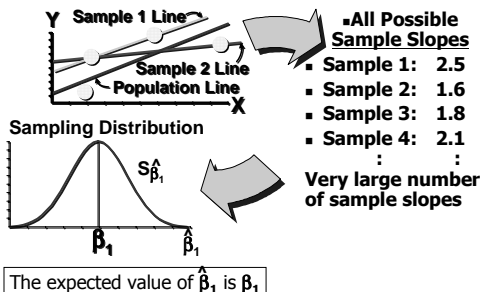
$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_X}}$$

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

Test of Slope Coefficient

- Is there a Linear Relationship Between X & Y
- Involves Population Slope β_1
- Hypotheses
 - $H_0: \beta_1 = 0$ (No Linear Relationship)
 - $H_a: \beta_1 \neq 0$ (Linear Relationship)
- The theoretical basis of the test is the Sampling Distribution of the slope coefficient

Sampling Distribution of Sample Slopes



Hypothesis Test for a slope coefficient for TAKING, use $\alpha = .05$

- Null hypothesis $H_0: \beta_1 = 0$
- Alternative $H_a: \beta_1 \neq 0$ two-tailed test
- Assumptions
 - Large sample, normal
- Test Statistic $t^* = (-2.177 - 0) / .163$
- Calculation $t^* = -13.366$
- P-value $P = .000$
- Conclusion
 - Reject $H_0: \beta_1 = 0$

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

$$t^* = (-2.177 - 0) / .163 = -13.366$$

Excel Regression Output from SAT Data

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

$31.813 = \sqrt{1012.040}$

ANOVA					
	df	SS	MS	F	Sig. F
Regression	1	180803.007	180803.007	178.652	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

Another Example

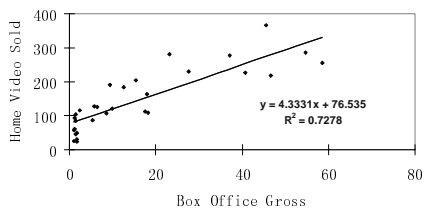
- A company distributing home videos of previously released movies wants to build a model to predict home video sales based on box office gross of the movie.
- They collect information on 30 movies, recording the gross sales and the home video sales
- Dependent Variable Y is
 - Home Video Units Sold in 1,000s
- Independent Variable X is
 - Box Office Gross in \$1,000,000

Descriptives

	Gross	Videos
Mean	15.93	145.54
Standard Error	3.24	16.46
Median	8.98	118.44
Mode	1.53	#N/A
Standard Deviation	17.75	90.14
Sample Variance	314.97	8126.01
Kurtosis	0.22	-0.40
Skewness	1.19	0.62
Range	57.41	341.00
Minimum	1.10	24.14
Maximum	58.51	365.14
Sum	477.76	4366.24
Count	30	30

Scatter Plot

Scatter Plot of Box Office Gross vs Home Video Sales



Excel Regression Output

Regression Statistics	
Multiple R	0.853
R Square	0.728
Adjusted R Square	0.718
Standard Error	47.867
Observations	30.000

Excel Regression Output

ANOVA						
	df	SS	MS	F	Sig. F	
Regression	1.000	171499.778	171499.778	74.851	0.000	
Residual	28.000	61154.424	2184.122			
Total	29.000	232654.202				
	Coef	Std. Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	76.535	11.832	6.469	0.000	52.299	100.772
Gross	4.333	0.501	8.652	0.000	3.307	5.359

Excel Regression Output

	Coef	Std. Error	t Stat	P-value
Intercept	76.535	11.832	6.469	0.000
Gross	4.333	0.501	8.652	0.000

$$\hat{Y} = 76.535 + 4.333(\text{Gross})$$

The estimated β is 4.333

The meaning of coefficients

- Intercept or estimated β_0
 - If the box office gross is zero, the home video sales is about 76 thousand**
- Slope or estimated β_1
 - The change in Y for a unit change in X
 - For every one million dollar increase in box office gross, the home video sales goes up by 4.3 thousand**

F-Test

- The null and alternative hypothesis for the F-test is
 - $H_0: \beta_1 = \beta_2 = \beta_k = 0$
 - $H_a: \text{at least one } \beta_i \neq 0$
- $$F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}$$

F-Test

ANOVA	df	SS	MS	F	Sig. F
Regression	1.000	171499.778	171499.778	74.851	0.000
Residual	28.000	64154.424	2291.229		
Total	29.000	235654.202			

- T.S. $F^* = 74.851$
- Compare with table F with 1 and 28 d.f. at a specified α level (e.g., .05)
- We reject H_0

Test of Slope Coefficient

- Is there a Linear Relationship Between X & Y
- Involves Population Slope β_1
- Hypotheses
 - $H_0: \beta_1 = 0$ (No Linear Relationship)
 - $H_a: \beta_1 \neq 0$ (Linear Relationship)
- The theoretical basis of the test is the Sampling Distribution of the slope coefficient

Hypothesis Test for a slope coefficient for GROSS, use $\alpha = .05$

- Null hypothesis $H_0: \beta_1 = 0$
- Alternative $H_a: \beta_1 \neq 0$ **two-tailed test**
- Assumptions
 - Large sample, normal
- Test Statistic $t^* = (4.333 - 0) / .501$
- Calculation $t^* = 8.652$
- P-value $P = .000$
- Conclusion **Reject $H_0: \beta_1 = 0$**

Excel Regression Output

	Coef	Std Error	t Stat	P-value
Intercept	76.535	11.832	6.469	0.000
Gross	4.333	0.501	8.652	0.000

$$t^* = (4.333 - 0) / 0.501 = 8.652$$

How to use estimated equation to predict Y?

- If the box office gross for a new movie is \$20 million, what is its predicted home video units sales?

Prediction

	Coef	Std Error	t Stat	P-value
Intercept	76.535	11.832	6.469	0.000
Gross	4.333	0.501	8.652	0.000

Estimated equation is

$$\hat{Y} = 76.535 + 4.333(\text{Gross})$$

When Gross = 20,

$$\hat{Y} = 76.535 + 4.333 * 20 = 163.195$$

What About a Multivariate Example?

- This is the case with more than one independent variable
- The Excel output will give a F-test for the model
- And t-tests for the individual coefficients for each independent variable
- This enables us to decide if each independent variable has a significant influence (meaning other than zero) on the dependent variable
- In a multiple regression the estimation of each independent variable is interpreted as *"holding all the other independent variables constant."*

Rainfall Regression.

- **An article in Geography (July 1980) used regression to predict average annual rainfall levels in California. Data on the following variables were collected for 30 meteorological weather stations scattered throughout California. For the group work we will focus on a bi-variate regression of Annual Percip on Latitude. You will have the option of examining all the variables for this problem for the last assignment**

Rainfall Regression

- **Annual Percip** DEPENDENT VARIABLE: Annual Precipitation in inches
- **Independent Variables**
 - **Altitude** The altitude of the station in feet
 - **Latitude** The latitude of the station in degrees
 - **Distance** Distance from the coast in miles
 - **Facing** I made this into a dummy variable. Stations on the Westward facing slopes of the California mountains were coded as 1, whereas stations on the leeward side were coded as 0

Briefly Describe

	Annual Percip	Altitude	Latitude	Distance	Facing
Mean	19.807	1375.300	37.027	78.700	0.433
Standard Error	3.035	382.812	0.487	12.653	0.092
Median	15.345	290.000	36.700	74.500	0.000
Mode	18.200	4152.000	33.800	1.000	0.000
Standard Deviation	16.621	2096.746	2.667	69.301	0.504
Sample Variance	276.264	4396344.631	7.110	4802.631	0.254
Kurtosis	3.051	0.777	-1.091	-1.192	-2.062
Skewness	1.700	1.461	0.228	0.417	0.283
Range	73.210	6930.000	9.200	197.000	1.000
Minimum	1.660	-178.000	32.700	1.000	0.000
Maximum	74.870	6752.000	41.900	198.000	1.000
Sum	594.220	41259.000	1110.800	2361.000	13.000
Count	30	30	30	30	30

Correlation Matrix

	Annual Percip	Altitude	Latitude	Distance	Facing
Annual Percip	1.000				
Altitude	0.302	1.000			
Latitude	0.577	0.231	1.000		
Distance	-0.210	0.574	0.161	1.000	
Facing	0.598	0.050	-0.011	-0.490	1.000

Regression of Average Annual Precipitation on Latitude

SUMMARY OUTPUT					
Regression Statistics					
Multiple R		0.577			
R Square		0.333			
Adjusted R Square		0.309			
Standard Error		13.819			
Observations		30			
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>
Regression	1	2664.887	2664.887	13.956	0.001
Residual	28	5346.766	190.956		
Total	29	8011.654			
	<i>Coeff</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	-113.303	35.721	-3.172	0.004	
Latitude	3.595	0.962	3.736	0.001	

Excel output

CA Rain Regression Summary						
Regression Statistics						
Multiple R		0.959				
R Square		0.738				
Adjusted R Square		0.696				
Standard Error		9.160				
Observations		30				
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>	
Regression	4	5913.817	1478.454	17.619	0.000	
Residual	25	2097.836	83.913			
Total	29	8011.654				
	<i>Coeff</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-113.756	24.311	-4.679	0.000	-163.825	-63.688
Altitude	0.002	0.001	1.948	0.063	0.000	0.005
Latitude	3.454	0.656	5.264	0.000	2.103	4.805
Distance	-0.054	0.039	-1.383	0.179	-0.134	0.026
Facing	15.858	4.371	3.628	0.001	6.855	24.860

Interpretation of Output

- R Square is high - .738 or 74% of the variability in annual precipitation across the measuring stations is due to knowing something about:
 - the longitude of the station
 - the altitude of the station
 - the distance of the station from ocean
 - whether the station is on the west side of the mountains.
- Adjusted R Square is also relatively high (.96)

Interpretation of Output

- Based on the F* test we can conclude that at least one of the independent variables is significant, meaning it is significantly different from zero
 - F* = 17.619,
 - p < .001

Prediction Equation

- What would we predict the annual precipitation for
 - Altitude of 1,000 ft
 - Latitude of 36
 - Distance of 100 miles
 - Facing = 1

Interpretation of Output

- Next we look at the t-tests for the individual coefficients
 - **Altitude** **1.948** **p=.063**
 - **Longitude** **5.264** **p<.001**
 - **Distance** **-1.383** **p=.179**
 - **Facing** **3.628** **p=.001**
- The coefficients for Longitude and Facing are clearly different from zero
- We have a harder time making the same conclusion about Altitude and Distance