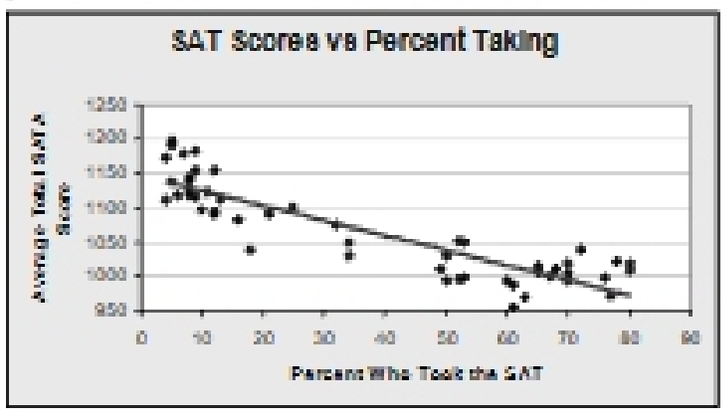
REGRESSION II: Hypothesis Testing in Regression

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



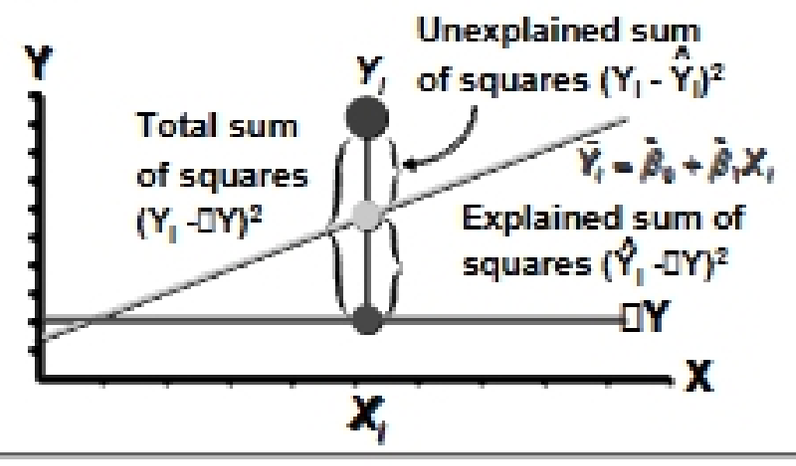
Excel Regression Output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.888					
R Square	0.788					
Adjusted R Square	0.780					
Standard Error	31.813					
Observations	51					
ANOVA						
	df	SS	MS	F	Sig F	
Regression	1	180803	180803	178.052	3.77E-18	
Residual	49	49589.97	1012.04			
Total	50	230393				
	Coef	Std Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1118.328	7.191	155.502	0.000	1103.869	1132.787
% Taking	-2.177	0.163	-13.343	0.000	-2.504	-1.850

Excel Regression Output

ANOVA					
	df	SS	MS	F	Sig F
Regression	1	180803.007	180803.007	178.052	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
  - $\beta_1, \beta_2, \ldots \beta_k$
- Some books use it to represent the number of **parameters estimated,** which includes the intercept coefficient
  - $\beta_0, \beta_1, \beta_2, \ldots$

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	178.662	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	178.662	0.000
Residual	49	49589.974	1012.040		
Total	50	230392.980			

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

F-Test

- F-Test (using F*)
  - $F = \frac{MS_{Regression}}{MS_{Residual}}$
  - 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: 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	0.888
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			

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

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