

Regression III: Dummy Variable Regression

Tom Ilvento
FREC 408

Linear Regression Assumptions about the error term

- Mean of Probability Distribution of the Error term is zero
- Probability Distribution of Error Has Constant Variance = σ^2
- Probability Distribution of Error is Normal
- Errors are Independent – they are uncorrelated with each other

(page553)

Regression Applications - we will look at

- Dummy variable regression as an alternative to ANOVA
- Multiple Regression

Dummy Variable Regression

- Regression using a dichotomous independent variable or set of variables
- Regression will estimate the same relationship as ANOVA
 - There will be a few important changes in Excel
 - The Data need to be in columns with matched data for all the variables – no missing values for any variable

For Any Categorical Variable

- I can represent any categorical variable with j classes
- With j-1 dummy variables, coded as 0 and 1
- For example,
 - Sex has 2 classes – male and Female
 - Represent as one variable coded 1 if female and 0 if male

Dummy Variables

- Example **Religion** (Protestant, Catholic, Jewish)
 - Dummy 1 (X1) = 1 if Protestant, 0 if not
 - Dummy 2 (X2) = 1 if Catholic, 0 if not
 - If you are Jewish, then you will have a value of zero on Dummy 1 (X1) and Dummy 2 (X2)
- The other class is called the **reference category** and is captured in the intercept term

Example Problem

- Examines the Sorption Rate of three different hazardous organic solvents
 - Aromatics
 - Chloroalkanes
 - Esters
- Asks if there are differences among the three?
- Sample of 32 sorption rates across the three classes of organic hazardous solvents
- The dependent variable is Sorption Rate

ANOVA Results

ANOVA: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Aromatics	9	8.48	0.942	0.028		
Cloro	8	8.05	1.006	0.161		
Esters	15	4.95	0.330	0.043		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3.305	2	1.653	24.512	0.000	3.328
Within Groups	1.955	29	0.067			
Total	5.261	31				

ANOVA Hypothesis Test for the Factor

- Null hypothesis
 - $H_0: \mu_1 = \mu_2 = \mu_3$
- Alternative
 - H_a : At least two means differ
- Assumptions
 - Equal variances, normal distribution
- Test Statistic
 - $F^* = 24.512$
- Rejection Region
 - $F_{.05, 2, 29 \text{ d.f.}} = 3.328$
- Conclusion
 - $F^* > F$
 - $24.512 > 3.328$
 - Reject $H_0: \mu_1 = \mu_2 = \mu_3$

There are differences across the organic chemicals

Regression Approach

- For Excel, reorganize the data
 - Dependent variable is in a single column
 - Classes are coded as 0/1 in contiguous columns
- Run Tools, Data Analysis, Regression and pick two of the three classes to be included in the model

Let's look at Excel



Regression Output with Esters as the reference category

SUMMARY OUTPUT					
Regression Statistics					
Multiple R		0.7927			
R Square		0.6283			
Adjusted R Square		0.6027			
Standard Error		0.2597			
Observations		32			
ANOVA					
	df	SS	MS	F	Signif. F
Regression	2	3.3054	1.6527	24.5115	0.0000
Residual	29	1.9553	0.0674		
Total	31	5.2608			

ANOVA Table is the same!

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Aromatics	9	8.48	0.942	0.028		
Cloro	8	8.05	1.006	0.161		
Esters	15	4.95	0.330	0.043		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3.305	2	1.653	24.512	0.000	3.328
Within Groups	1.955	29	0.067			
Total	5.261	31				

Regression gives us Coefficients

	Coefficients	Std Error	t Stat	P-value
Intercept	0.3300	0.0670	4.9221	0.0000
Aromatics	0.6122	0.1095	5.5919	0.0000
Cloro	0.6763	0.1137	5.9487	0.0000

$$\hat{Y} = .3300 + .6122(\text{Aromatics}) + .6763(\text{Cloro})$$

Estimated values from our equation

- Since our independent variables are dummy variables, it is easy to solve the equation

$$\hat{Y} = .3300 + .6122(\text{Aromatics}) + .6763(\text{Cloro})$$

- When **Aromatics = 1**
 - $= .33 + .6122(1) + .6763(0) = \mathbf{.9422}$
- When **Chloroalkanes = 1**
 - $= .33 + .6122(0) + .6763(1) = \mathbf{1.006}$
- When **Aromatics and Chloroalkanes = 0**
 - $= .33 + .6122(0) + .6763(0) = \mathbf{.3300}$
 - This represents Esters!**

The model estimated the mean levels for each solvent

	Aromatics	Cloro	Esters
Mean	0.942	1.006	0.330
Standard Error	0.056	0.142	0.054
Median	0.950	1.015	0.340
Mode	#N/A	#N/A	0.060
Standard Deviation	0.168	0.401	0.208
Sample Variance	0.028	0.161	0.043

The t-test for the dummy coefficients

	Coefficients	Std Error	t Stat	P-value
Intercept	0.3300	0.0670	4.9221	0.0000
Aromatics	0.6122	0.1095	5.5919	0.0000
Cloro	0.6763	0.1137	5.9487	0.0000

$$\hat{Y} = .3300 + .6122(\text{Aromatics}) + .6763(\text{Cloro})$$

The test-test for Aromatics and Cloro represent a test if each are significantly different from Esters, i.e., a difference of means test!!

Hypothesis Test for a slope coefficient for Cloro,

- Null hypothesis $H_0: \beta_2 = 0$
- Alternative $H_a: \beta_2 \neq 0$ two-tailed test
- Assumptions
 - Large sample, normal
- Test Statistic $t^* = (.6763 - 0) / .1137$
- Calculation $t^* = 5.9487$
- P-value $P = .000$
- Conclusion
 - Reject $H_0: \beta_2 = 0$

Regression Output with Aromatics as the reference category

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.7927				
R Square	0.6283				
Adjusted R Square	0.6027				
Standard Error	0.2597				
Observations	32				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Signif F</i>
Regression	2	3.3054	1.6527	24.5115	0.0000
Residual	29	1.9553	0.0674		
Total	31	5.2608			

Regression Output with Aromatics as the reference category

	<i>Coefficients</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	0.9422	0.0866	10.8858	0.0000
Cloro	0.0640	0.1262	0.5075	0.6157
Esters	-0.6122	0.1095	-5.5919	0.0000

The coefficients reflects the difference between Cloro and Esters from Aromatics

Notice that the t-test for Cloro shows that it is not significantly different from Aromatics

MN Apartment Sales Example

- A real estate agent wants to use regression analysis to explore the relationship between the sale prices of apartment buildings and various characteristics of the apartments, including:
 - # of apartments
 - age of structure
 - lot size
 - # of on-site parking spaces
 - gross building area
 - condition of apartment building

We will focus on Condition

Price	Condition
\$79,300	F
\$90,300	F
\$93,600	F
\$108,750	G
\$110,000	G
\$134,400	E
\$155,700	E
\$157,500	G
\$162,500	G

Condition has 3 levels or categories

Fair
Good
Excellent

This is just part of the data. We need to create the dummy variables

MN Apartment Sales Example - dummy variables

- First, look at the relationship between PRICE and Condition of apartment building
- There are three categories – Excellent, Good and Fair.
- Need $(3-1) = 2$ dummy variables
 - $X_1 = \begin{cases} 1, & \text{if condition is Excellent} \\ 0, & \text{if condition is NOT excellent} \end{cases}$
 - $X_2 = \begin{cases} 1, & \text{if condition is Good} \\ 0, & \text{if condition is NOT good} \end{cases}$

MN Apartment Sales Example - dummy variables

- The last category "Fair" is the **reference category**
- When $X_1=0$ and $X_2=0$, the condition of apartment building is Fair
- I will label the two dummy variables (X_1 and X_2) as **Excellent** and **Good**

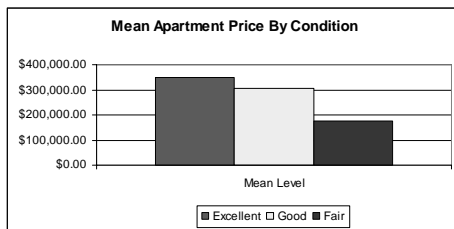
Regression Analysis

- In Excel, we want the regression of **Price** on Conditions by using two dummy variables X_1 and X_2
 - I will label them as **Excellent** and **Good**
- Reorganize the data
 - Dependent variable Price is in a single column
 - Excellent** and **Good** are coded as 0/1
- Run Tools, Data Analysis, Regression and pick **Excellent** and **Good** to be included in the model

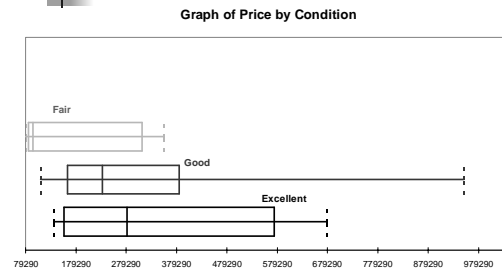
I create two dummy variables to represent Condition

Price	Condition	Excellent	Good
\$79,300	F	0	0
\$90,300	F	0	0
\$93,600	F	0	0
\$108,750	G	0	1
\$110,000	G	0	1
\$134,400	E	1	0
\$155,700	E	1	0
\$157,500	G	0	1
\$162,500	G	0	1

There does seem to be differences across the groups



There also is a lot of spread across the groups



Regression Output with two dummy variables

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.288				
R Square	0.083				
Adjusted R Square	0.000				
Standard Error	211572.694				
Observations	25				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Sig F</i>
Regression	2	89083840373.383	44541920186.692	0.995	0.386
Residual	22	984786106940.857	44763004860.948		
Total	24	1073869947314.240			
	<i>Coeff</i>	<i>Std Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	176940.000	94618.185	1.870	0.075	
Excellent	173310.000	128113.628	1.353	0.190	
Good	128641.286	110226.850	1.167	0.256	

Regression Output with two dummy variables

Regression Statistics	
Multiple R	0.288
R Square	0.083
Adjusted R Square	0.000
Standard Error	211572.694
Observations	25

**R² is only 0.083. Is this number too small?
Is our model is a good one?...**
... We're going to do some tests on it later.

Regression gives us Coefficients and Estimated Model between Price and Condition

	Coeff	Std Error	t Stat	P-value
Intercept	176940.00	94618.185	1.870	0.075
Excellent	173310.00	128113.628	1.353	0.190
Good	128641.29	110226.850	1.167	0.256

Estimated equation/model:

$$\hat{Y} = 176940 + 173310 * (\text{Excellent}) + 128641 * (\text{Good})$$

Estimated values from our equation

- When **Excellent = 1**
 - Price = 176940 + 173310(1) + 128641(0)
 - = \$350,250
- When **Good = 1**
 - Price = 176940 + 173310(0) + 128641(1)
 - = \$305,581
- When **Excellent and Good = 0, Fair**
 - Price = 176940 + 173310(0) + 128641(0)
 - = \$176,940

Regression/ANOVA Output

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.288				
R Square	0.083				
Adjusted R Square	0.000				
Standard Error	211572.694				
Observations	25				
ANOVA					
	df	SS	MS	F	Sig F
Regression	2	89063840373.383	4451920186.692	0.995	0.386
Residual	22	984786106940.857	44763004860.948		
Total	24	1073869947314.240			
	Coeff	Std Error	t Stat	P-value	
Intercept	176940.000	94618.185	1.870	0.075	
Excellent	173310.000	128113.628	1.353	0.190	
Good	128641.286	110226.850	1.167	0.256	

F Test for the estimated model

- Null hypothesis
 - Alternative
 - Assumptions
 - Test Statistic
 - Rejection Region
 - Conclusion
- $H_0: \mu_1 = \mu_2 = 0$
 - H_a : At least one mean differs from 0
 - Equal variances, normal distribution
 - $F^* = 0.995$
 - $F_{.05, 2, 22 \text{ d.f.}} = 3.44$
 - $F^* < F$
 - $0.995 < 3.44$
 - Can't reject $H_0: \mu_1 = \mu_2 = 0$

It seems the conditions have no statistically significant impact on apartment price.

Is that true in real life?

The t-test for the dummy coefficients

	Coeff	Std Error	t Stat	P-value
Intercept	176940.00	94618.185	1.870	0.075
Excellent	173310.00	128113.628	1.353	0.190
Good	128641.29	110226.850	1.167	0.256

$$\hat{Y} = 176940 + 173310(X_1) + 128641(X_2)$$

The t-tests for X1 (Excellent) and X2 (Good) represent a test if each are significantly different from "Fair"

Hypothesis Test for a slope coefficient for Excellent

- Null hypothesis
 - Alternative
 - Assumptions
 - Test Statistic
 - Calculation
 - P-value
 - Conclusion
- $H_0: \beta_1 = 0$
 - $H_a: \beta_1 \neq 0$ **two-tailed test**
 - small sample, normal
 - $t^* = (173310)/128113$
 - $t^* = 1.353$
 - $P = .190$
 - Can't reject $H_0: \beta_1 = 0$

Based on our model and t-test, it indicates that the condition of the apartment makes no difference on Price!

