Dummy Variable Regression

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- Dummy variables are ones that take on either a 1 or a zero, where 1 indicates the presence of some attribute.
  - Sex: 1 = female and 0 = male
- With a categorical variable with j classes or categories, we will generate j-1 dummy variables
- Regression can handle dummy variables in the regression equation as independent variables
- When all the independent variables in regression are dummy variables, there is a special interpretation.
  - Regression will estimate the same relationship as ANOVA
  - The coefficients generate difference of means tests

Creating Dummy Variables

- 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
  - Represented as one variable coded 1 if female and 0 if male
- Example Drug Treatment (Control, 1 mg/l, 2mg/l)
  - Dummy 1 (X_1) = 1 if 1 mg/l, 0 if not
  - Dummy 2 (X_2) = 1 if 2 mg/l, 0 if not
  - If the Control group, it will have a value of zero on Dummy 1 (X_1) and Dummy 2 (X_2)
- The other class is called the reference category and is captured in the intercept term

Creating Dummy Variables

- **In Excel:** use the example of AGE = 1, 2, 3; and AGE is in column B
  - If the sample size is small you can do it by hand
  - For larger data sets, use IF statements
    - Create new column variables for AGE1
      - **=IF(b2=1,1,0)** if the value in B2 = 1, then give the new column a one, otherwise give it a zero
    - Create a new column for AGE2
      - **=IF(b2=2,1,0)** if the value in B2 = 2, then give the new column a one, otherwise give it a zero
    - Copy the formulas for each column
  - In a program like JMP you can use IF statements in a formula or a recode statement

Example: Sorption Rate Regressed on Solvent Type

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

ANOVA Results from Excel

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 Test for Sorption Data

- **Ho:** **Ho: $\mu_1 = \mu_2 = \mu_3$**
- **Ha:** **Ha: At least two means are different**
- **Assumptions** **Equal variances, normal distribution**
- **Test Statistic** **$F^* = 24.512$ $p < .001$**
- **Rejection Region** **$F_{.01, 2, 29} = 3.328$**
- **Conclusion:** **$F^* > F_{.01, 2, 29}$**

or $p < .001$

Reject Ho: $\mu_1 = \mu_2 = \mu_3$

There are differences across the organic chemicals

Regression approach for the Sorption Data

- 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**
  - pick two of the three classes to be included in the model
- You can only include j-1 dummy variables in the model!
  - You must pick 2 of the 3 dummy variables
  - The other one becomes the **Reference Group**

SORPTION	Aromatics	Cloro	Esters
1.06	1	0	0
0.79	1	0	0
0.82	1	0	0
0.89	1	0	0
1.05	1	0	0
0.95	1	0	0
0.65	1	0	0
1.15	1	0	0
1.12	1	0	0
1.58	0	1	0
1.45	0	1	0
0.57	0	1	0
1.16	0	1	0
1.12	0	1	0
0.91	0	1	0
0.83	0	1	0
0.43	0	1	0
0.29	0	0	1
0.06	0	0	1
0.44	0	0	1
0.61	0	0	1
0.55	0	0	1
0.43	0	0	1
0.51	0	0	1
0.10	0	0	1
0.34	0	0	1
0.53	0	0	1
0.06	0	0	1
0.09	0	0	1
0.17	0	0	1
0.60	0	0	1
0.17	0	0	1

Regression Output

- The following is the top half of the Excel Regression output of a regression of Sorption on Aromatics and Cloro (Esters is the Reference Category)
- The ANOVA Table is identical to the previous ANOVA results

Regression of Sorption on Aromatics and Cloro Dummy Variables

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

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 more

- The Regression output gives us the coefficients for our independent dummy variables, Aromatics and Cloro
- Along with a t-test for each one
- **What are the meanings of these coefficients and the t-tests?**

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.3300	0.0670	4.9221	0.0000	0.1929	0.4671
Aromatics	0.6122	0.1095	5.5919	0.0000	0.3883	0.8361
Cloro	0.6763	0.1137	5.9487	0.0000	0.4437	0.9088

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

Estimated values from our equation

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

- Since our independent variables are dummy variables, it is easy to solve the equation
- **When Aromatics = 1**
  - **est. Sorption = .33 + .6122(1) + .6763(0) = .9422**
- **When Cloro = 1**
  - **est. Sorption = .33 + .6122(0) +.6763(1) = 1.006**
- **When Aromatics and Cloro = 0**
  - **est. Sorption = .33 + .6122(0) + .6763(0) = .3300**
  - **This represents Esters!**

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

The t-test for the coefficients

	Coefficients	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.3300	0.0670	4.9221	0.0000	0.1929	0.4671
Aromatics	0.6122	0.1095	5.5919	0.0000	0.3883	0.8361
Cloro	0.6763	0.1137	5.9487	0.0000	0.4437	0.9088

- The coefficients for Aromatics and Cloro represent the difference in means for each type from the Reference Group (Esters)
- We have called this the “effect”
- The t-test is still whether each coefficient is significantly different from zero
- But the interpretation of the t-test is a difference of means test for Aromatics and Cloro compared to the reference group!!
- We can do a formal test, but based on the t-tests we can we that both coefficients are significantly different from zero

The Hypothesis Test for the Aromatics Coefficient

- **Ho:** Ho: $\beta_1 = 0$
- **Ha:** Ha: $\beta_1 \neq 0$
- **Assumptions** Equal variances, normal distribution
- **Test Statistic** $t^* = (.6122-0)/.1095 = 5.591 \quad p < .001$
- **Rejection Region** $t_{.05/2,\, 29} = 2.045$
- **Conclusion:** $t^* > t_{.05,\, 29}$
  or $p < .001$
  Reject Ho: $\beta_1 = 0$

Now we can say that the means for Aromatics and Esters are significantly different from each other

And based on t* = 5.9487 for Cloros, we can also say that the means for Cloros and Esters are significantly different from each other

What happens if another group is the Reference Group?

Regression of Sorption on Cloro and Esters, Aromatics is the Reference Group

Regression Statistics	
Multiple R	0.793
R Square	0.628
Adjusted R Square	0.603
Standard Error	0.260
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	3.305	1.653	24.512	0.000
Residual	29	1.955	0.067		
Total	31	5.261			

	Coef.	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.9422	0.0866	10.8858	0.0000	0.7652	1.1192
Cloro	**0.0640**	**0.1262**	**0.5075**	**0.6157**	-0.1940	0.3221
Esters	**-0.6122**	**0.1095**	**-5.5919**	**0.0000**	-0.8361	-0.3883

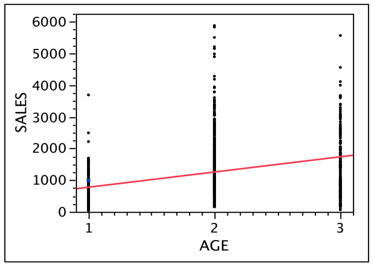
- Much is the same:
  - R^2 is the same; Standard Error is the same
  - Sums of Squares are the same; F-test is the same
- The coefficients are different, but in a predictable way
  - The intercept now represents the mean of Aromatics
  - The coefficients represent a difference of means test with Aromatics
  - **Notice that the t-test for Cloro shows that it is not significantly different from Aromatics**

A few notes

- If you solve the equation for the second regression, you will also perfectly predict the mean level for each group
  - It does not matter which category you make the Reference Category
  - But you might choose the reference category to make the best test
- The F-test is a perfect first test for a dummy variable regression
  - It tells us that at least 1 coefficient is significantly different from zero
  - Which also says that at least two means are different from each other
- Always remember that although we can multiple dummy variables in the equation, we are still representing one concept - the organic solvent

Catalog Sales Data: Regress SALES on AGE

- Example: Age measured with three categories - <31 years; 31 to 55 years; 56 and over
  - Coded as 1, 2, 3
- We could think of this as an ordinal variable
- Or we can represent it with two dummy variables
  - AGE1 has a 1 if <31 and zero for all else;
  - AGE2 has a 1 if 31 to 55, zero for all else;
- The left out category, 56 and over, is called the Reference Category



If we estimate the model with AGE as an ordinal variable (1, 2, 3)

- What do you see?
  - Our model has a weak fit: The model R^2 = .121
  - Age is significant in the model, $p < .001$
- **est SALES = 295.72 + 480.21*AGE**
- If we predict Sales for the three age groups we get the following:

Linear Fit

SALES = 295.7203 + 480.21361*AGE

Summary of Fit

RSquare	0.121278
RSquare Adj	0.120398
Root Mean Square Error	901.3584
Mean of Response	1216.77
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	111907126	111907126	137.7408
Error	998	810822099	812446.99	Prob > F
C. Total	999	922729225		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	295.7203	83.49462	3.54	0.0004*
AGE	480.21361	40.91694	11.74	<.0001*

Age = 1 est Sales = 295.72 + 480.21(1) = $775.93

Age = 2 est Sales = 295.72 + 480.21(2) = $1,256.15

Age = 3 est Sales = 295.72 + 480.21(3) = $1,736.36

Regression of SALES on AGE as two dummy variables

- What do you see?
- R^2 is increases to .190
- F-test is significant
- t-test for AGE1 is significant, coefficient is negative: the youngest age group has significantly less sales than the oldest age group
- t-test for AGE2 is not significant, coefficient is positive

Whole Model

Summary of Fit

RSquare	0.189723
RSquare Adj	0.188098
Root Mean Square Error	865.9769
Mean of Response	1216.77
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	175062951	87531475	116.7217
Error	997	747666275	749916.02	Prob > F
C. Total	999	922729225		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1432.1268	60.48245	23.68	<.0001*
AGE1	-873.5031	79.19012	-11.03	<.0001*
AGE2	69.564116	71.65431	0.97	0.3319

Regression results

$$\hat{Y} = 1432.13 - 873.51(\text{AGE1}) + 69.56(\text{AGE2})$$

- This is a very simple model
  - We expect age to be related to sales, but not explain everything!
  - R-Square is .190 – about 19% of the variability in Sales is explained by the customer's age
- **Focus on the F-test first**
  - F = 116.72 $p < .001$
  - **Age is related to Sales**
- **Focus on coefficients and t-tests**
  - The intercept is $1,432.13
  - The coefficient for AGE1 is negative (-$873.51)
  - The t-test for AGE1 is large (-11.030) and the p-value is small (< .001). We can reject a null hypothesis that the coefficient is zero
  - The coefficient for AGE2 is positive ($69.56), but not significant (t = .971 and p-value = .332)

The Hypothesis Test for AGE2

- **Ho:** **Ho: $\beta_1 = 0$**
- **Ha:** **Ha: $\beta_1 \neq 0$**
- **Assumptions** **Equal variances, normal distribution**
- **Test Statistic** **$t^* = (69.564-0)/71.654 = .97$ p = .332**
- **Conclusion:** **Cannot Reject Ho: $\beta_1 = 0$**

We do not have evidence to suggest that the coefficient for AGE2 is any different than zero.

In other words, there is no difference in sales for those 31 to 55 and those over 55!

It is easy to solve the equation

$$\hat{Y} = 1432.13 - 873.51(\text{AGE1}) + 69.56(\text{AGE2})$$

- **When AGE1 = 1 AGE2 = 0**
  - **est. SALES = 1,432.13 – 873.51(1) + 69.56(0) = $558.62**
- **When AGE2 = 1 AGE1 = 0**
  - **est. SALES = 1,432.13 – 873.51(0) + 69.56(1) = $1,501.69**
- **When AGE1 and AGE2 = 0**
  - **est. SALES = 1,432.13 – 873.51(0) + 69.56(0) = $1,432.13**
  - This is the mean for the Reference Group, AGE3

Compare to AGE as Ordinal

Age = 1 est Sales = 295.72 + 480.21(1) = $775.93

Age = 2 est Sales = 295.72 + 480.21(2) = $1,256.15

Age = 3 est Sales = 295.72 + 480.21(3) = $1,736.36

Dummy Variable Regression

- The equation predicts the mean level for each group
- The intercept represents the mean level for the Reference Category
- The t-tests represent whether the other categories are significantly different from the Reference Category – a difference of means test!
- It does not matter **much** which category is the Reference Group!

Same Model with AGE2 and AGE3

Whole Model

Summary of Fit

RSquare	0.189723
RSquare Adj	0.188098
Root Mean Square Error	865.9769
Mean of Response	1216.77
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	175062951	87531475	116.7217
Error	997	747666275	749916.02	Prob > F
C. Total	999	922729225		<.0001*

Parameter Estimates

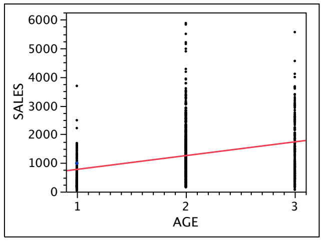
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	558.62369	51.117	10.93	<.0001*
AGE2	943.06725	63.94654	14.75	<.0001*
AGE3	873.50314	79.19012	11.03	<.0001*

Dummy Variable Regression

- A dummy variable regression will always predict the mean levels of each group (when no other variables are in the model)
- Whenever we consider a categorical variable with several dummy variables, we still need to think of it as one variable
  - With AGE, we had three categories, two dummy variables, but only one thing we are measuring – the age of the respondent!
  - The F-test gives the overall effect of AGE
- You can take any continuous variable and convert it into dummy variables. Why?
  - To see if the effect is linear over the range of the variable
  - Here you can observe if the effects for each level are uniform or near uniform

Compare the results from the two models

- In the first model we left AGE as a single variable and continuous
  - The coefficient was 408.214, which modeled a constant effect of AGE on SALES
  - R^2 was very low, .121
- In the second model we created two dummy variables to represent AGE
  - R^2 increased to .1997
  - And we fit the mean level of each group with our model



Let's Add a Second Dummy Variable to the Model

- Run a regression of Sales on:
  - AGE1
  - AGE2
  - GENDER (1=male)
- Can you guess what the components of the model will represent?
  - The Reference Category will now represent older females, a combination of both variables.
  - The coefficients will no longer predict exact means
  - The coefficients will be adjusted for the other variable in the model

Regression results from JMP

- What do you see?
  - R^2 increases slightly to .210
  - Degrees of freedom for regression reflects 3 independent variables
  - The F-test still shows overall significance - some of the coefficients are different from zero
  - AGE1 and GENDER are significant in the model
  - The coefficient for AGE2 is even smaller once we account or "control" for Gender

Whole Model

Summary of Fit

RSquare	0.212411
RSquare Adj	0.210039
Root Mean Square Error	854.1953
Mean of Response	1216.77
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	195998200	65332733	89.5399
Error	996	726731025	729649.62	Prob > F
C. Total	999	922729225		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1322.496	63.073	20.968	<.0001*
AGE1	-883.395	78.135	-11.31	<.0001*
AGE2	3.396	71.751	0.047	0.9623
GENDER	295.715	55.207	5.357	<.0001*

Ready to solve the equation?

$$\hat{Y} = 1322.50 - 883.40(\text{AGE1}) + 3.39(\text{AGE2}) + 295.71(\text{Gender})$$

- When AGE1 = 1 Gender =1 AGE2 = 0
  - est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(1) = $734.81
- When AGE1 = 1 Gender =0 AGE2 = 0
  - est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(0) = $439.10
- The difference is the amount due to Gender!
  - Men spend on average, $295.71 more than women
  - That difference remains across all age levels (based on our model)
  - The difference is statistically significant in the model
- Can you solve for the rest of AGE and GENDER?

Solve the Equation

$$\hat{Y} = 1322.50 - 883.40(\text{AGE1}) + 3.39(\text{AGE2}) + 295.71(\text{Gender})$$

- When AGE1 = 1 Gender =1 AGE2 = 0
  - est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(1) =
- When AGE1 = 1 Gender =0 AGE2 = 0
  - est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(0) =
- When AGE2 = 1 Gender =1 AGE1 = 0
  - est. SALES = 1,322.50 – 883.40(0) + 3.39(1) +295.71(1) =
- When AGE2 = 1 Gender =0 AGE1 = 0
  - est. SALES = 1,322.50 – 883.40(0) + 3.39(1) +295.71(0) =
- When AGE1 =0 Gender =1 and AGE2 =0
  - est. SALES = 1,322.50 – 883.40(0) + 3.39(0) +295.71(1) =
- When AGE1 = 0 Gender =0 and AGE2 =0
  - est. SALES = 1,322.50 – 883.40(0) + 3.39(0) +295.71(0) =

Solve the Equation

$$\hat{Y} = 1322.50 - 883.40(\text{AGE1}) + 3.39(\text{AGE2}) + 295.71(\text{Gender})$$

- **When AGE1 = 1 Gender =1 AGE2 = 0**
  - **est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(1) = $734.81**
- **When AGE1 = 1 Gender =0 AGE2 = 0**
  - **est. SALES = 1,322.50 – 883.40(1) + 3.39(0) +295.71(0) = $439.10**
- **When AGE2 = 1 Gender =1 AGE1 = 0**
  - **est. SALES = 1,322.50 – 883.40(0) + 3.39(1) +295.71(1) = $1,621.60**
- **When AGE2 = 1 Gender =0 AGE1 = 0**
  - **est. SALES = 1,322.50 – 883.40(0) + 3.39(1) +295.71(0) = $1,325.89**
- **When AGE1 =0 Gender =1 and AGE2 =0**
  - **est. SALES = 1,322.50 – 883.40(0) + 3.39(0) +295.71(1) = $1,618.21**
- **When AGE1 = 0 Gender =0 and AGE2 =0**
  - **est. SALES = 1,322.50 – 883.40(0) + 3.39(0) +295.71(0) = $1,322.50**

Summary of our model

- Overall, weak model R^2 = .21 (21% of the variability in Sales is explained by knowing age and gender).
- Not too surprising, there is more to sales than age and sex!
- Younger customers spend less on average than older customers (-$893), middle and older customers spend about the same
- Men, on average, spend about $296 per year more than women
- **All estimates are based on controlling for the other variables in the model!**

Summary

- When independent variables are represented as dummy variables the interpretation takes on a special meaning.
  - The coefficient represents the difference of means between the value of one for the dummy variable and the reference group.
  - The intercept represents the mean level of the reference group.
- We can have multiple variables represented as a series of dummy variables.
- We can convert ordinal and even continuous level variables in a series of dummy variables.