The University of Georgia

Regression With a Categorical Independent Variable

ERSH 8320

Today's Lecture

- Regression with a single categorical independent variable.
- Coding procedures for analysis.
  - Dummy coding.
- Relationship between categorical independent variable regression and other statistical terms.

Regression with Continuous Variables

- Linear regression regresses a continuous-valued dependent variable, Y, onto a set of continuous-valued independent variables **X**.
- The regression line gives the estimate of the mean of Y conditional on the values of **X**, or $E(Y|\mathbf{X})$.
- But what happens when some or all independent variables are categorical in nature?
- Is the point of the regression to determine $E(Y|\mathbf{X})$, across the levels of Y?
- Can’t we just put the categorical variables into SPSS and push the “Continue" button?

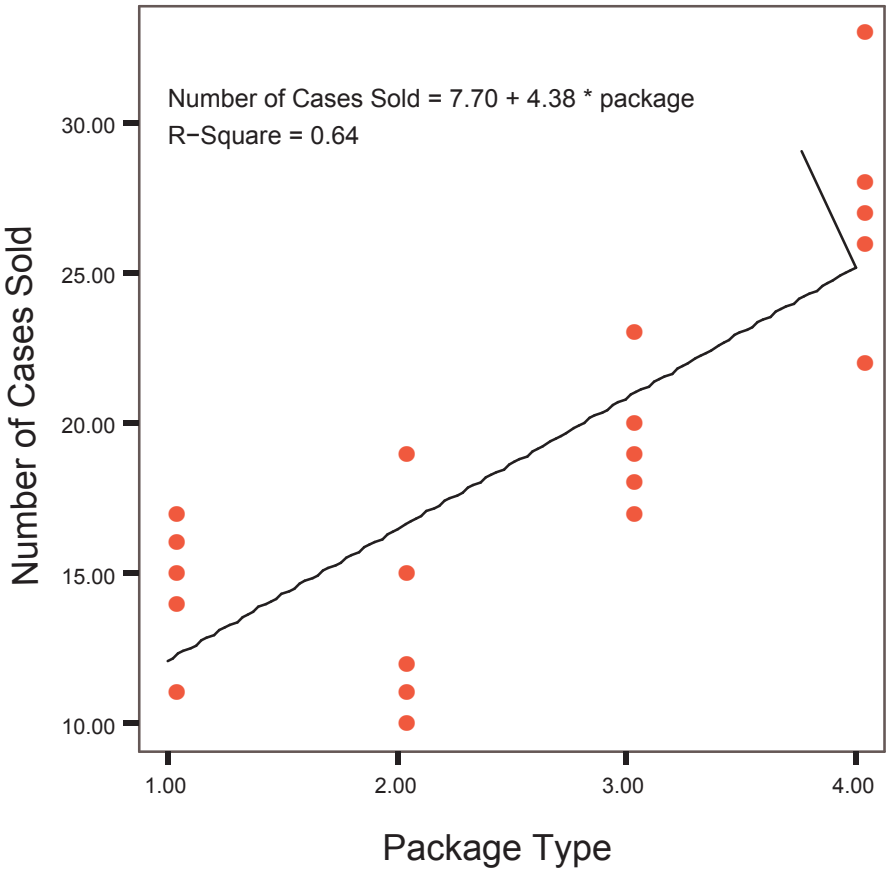
Example Data Set

Neter (1996, p. 676).

- "The Kenton Food Company wished to test four different package designs for a new breakfast cereal.
- "Twenty stores, with approximately equal sales volumes, were selected as the experimental units.
- "Each store was randomly assigned one of the package designs, with each package design assigned to five stores.
- "The stores were chosen to be comparable in location and sales volume.
- "Other relevant conditions that could affect sales, such as price, amount and location of shelf space, and special promotional efforts, were kept the same for all of the stores in the experiment."

Cereal

A Regular Regression?



What is wrong with this picture?

Categorical Variables

- Categorical variables commonly occur in research settings.
- Another term sometimes used to describe for categorial variables is that of qualitative variables.
- A strict definition of a qualitative or categorical variable is that of a variable that has a finite number of levels.
- Continuous (or quantitative) variables, alternatively, have infinitely many levels.
  - Often this is assumed more than practiced.
  - Quantitative variables often have countably many levels.
  - Level of precision of an instrument can limit the number of levels of a quantitative variable.

Research Design

- Categorical variables can occur in many different research designs:
  - Experimental research.
  - Quasi-experimental research.
  - Nonexperimental/Observational research.
- Such variables can be used with regression for:
  - Prediction.
  - Explanation.

Analysis Specifics

- Because of nature of categorical variables, emphasis of regression is not on linear trends but on differences between means (of Y) at each level of the category.
  - Not all categorical variables are ordered (like cereal box type, gender,etc...).
- When considering differences in the mean of the dependent variable, the type of analysis being conducted by a regression is commonly called an ANalysis Of VAriance (ANOVA).
- Combinations of categorical and continuous variables in the same regression is called ANalysis Of CoVAriance (ANCOVA - Chapters 14 and 15).

Example Variable: Two Categories

- From Pedhazur (1997; p. 343): "Assume that the data reported [below] were obtained in an experiment in which E represents an experimental group and C represents a control group.

	E	C
	20	10
	18	12
	17	11
	17	15
	13	17
$\sum Y$	85	65
$\bar{Y}$	17	13
$\sum (Y - \bar{Y})^2 = \sum y^2$	26	34

Old School Statistics: The t-test

- As you may recall from an earlier course on statistics, an easy way to determine if the means of the two conditions differ significantly is to use a t-test (with $n_1 + n_2 - 2$) degrees of freedom.

$H_0 \;\; \mu_1 = \mu_2$

$H_A \;\; \mu_1 \neq \mu_2$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\Sigma y_1^2 + \Sigma y_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Old School Statistics: The t-test

$$t = \frac{17 - 13}{\sqrt{\frac{26+34}{5+5-2}\left(\frac{1}{5} + \frac{1}{5}\right)}} = \frac{4}{\sqrt{3}} = 2.31$$

- From Excel ("=tdist(2.31,8,2)"), $p = 0.0496$.
- If we used a Type-I error rate of 0.05, we would reject the null hypothesis, and conclude the means of the two groups were significantly different.
- But what if we had more than two groups?.
- This type of problem can be solved equivalently from within the context of the General Linear Model.

Variable Coding

- When using categorical variables in regression, levels of the categories must be recoded from their original value to ensure the regression model truly estimates the mean differences at levels of the categories.
- Several types of coding strategies are common:
  - Dummy coding.
  - Effect coding.
- Each type will produce the same fit of the model (via R^2).
- The estimated regression parameters are different across coding types, thereby representing the true difference in approaches employed by each type of coding.
- The choice of method of coding does not differ as a function of the type of research or analysis or purpose (explanation or prediction) of the analysis.

Variable Coding

- Definition: “a code is a set of symbols to which meanings can be assigned” (Pedhazur, 1997; p. 342).
- The assignment of symbols follows a rule (or set of rules) determined by the categories of the variable used.
  - Typically symbols represent the respective levels of a categorical variable.
- All entities within the same symbol are considered alike (or homogeneous) within that category level.
- Categorical levels must be predetermined prior to analysis.
  - Some variables are obviously categorical - gender.
  - Some variables are not so obviously categorial - political affiliation.

Dummy Coding

- The most straight-forward method of coding categorical variables is dummy coding.
- In dummy coding, one creates a set of column vectors that represent the membership of an observation to a given category level.
- If an observation is a member of a specific category level, they are given a value of 1 in that category level's column vector.
- If an observation is not a member of a specific category, they are given a value of 0 in that category level's column vector.

Dummy Coding

- For each observation, a no more that a single 1 will appear in the set of column vectors for that variable.
- The column vectors represent the predictor variables in a regression analysis, where the dependent variable is modeled as a function of these columns.
  - Because of linear dependence with an intercept, one category-level vector is often excluded from the analysis.
- Because all observations at a given category level have the same value across the set of predictors, the predicted value of the dependent variable, Y', will be identical for all observations within a category.
- The set of category vectors (and a vector for an intercept) are now used as input into a regression model.

Dummy Coded Regression Example

Y	X_1	X_2	X_3	Group
20	1	1	0	E
18	1	1	0	E
17	1	1	0	E
17	1	1	0	E
13	1	1	0	E
10	1	0	1	C
12	1	0	1	C
11	1	0	1	C
15	1	0	1	C
17	1	0	1	C
Mean	15	1	0.5	0.5
SS	100	0	2.5	2.5
	$\sum yx_2 = 10$		$\sum yx_3 = -10$	

Dummy Coded Regression

- The General Linear Model states that the estimated regression parameters are given by:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- From the previous slide, you can see what our entries for $\mathbf{X}$ could be, but...
  - Notice that $X_1 = X_2 + X_3$.
- This linear dependency means that:
  - $(\mathbf{X}'\mathbf{X})$ is a singular matrix - no inverse exists.
  - Any combination of two of the columns would rid us of the linear dependency.

Dummy Coded Regression - X_2 and X_3

- For our first example analysis, consider the regression of Y on X_2 and X_3.

$$Y = b_2 X_2 + b_3 X_3 + e$$

- $b_2 = 17$
- $b_3 = 13$
- $\sum y_2 = 100$
- $SS_{res} = \mathbf{X'X} = 60$
- $SS_{reg} = 100 - 60 = 40$
- $R^2 = \frac{40}{100} = 0.4$

Dummy Coded Regression - X_2 and X_3

- $b_2 = 17$ is the mean for the E category.
- $b_3 = 13$ is the mean for the C category.
- Without an intercept, the model is fairly easy to interpret.
- For more advanced models, an intercept will prove to be helpful in interpretation.

Dummy Coded Regression - X_1 and X_2

- For our second example analysis, consider the regression of Y on X_1 and X_2.

$$Y = a + b_2 X_2 + e$$

- $a = 13$
- $b_2 = 4$
- $\sum y_2 = 100$
- $SS_{res} = \mathbf{X'X} = 60$
- $SS_{reg} = 100 - 60 = 40$
- $R^2 = \frac{40}{100} = 0.4$

Dummy Coded Regression - X_2 and X_3

- $a = 13$ is the mean for the C category.
- $b_2 = 4$ is the mean difference between the E category and the C category.
- The C category is called reference category.
- For members of the C category:

$$Y' = a + b_2 X_2 = 13 + 4(0) = 13$$

- For members of the E category:

$$Y' = a + b_2 X_2 = 13 + 4(1) = 17$$

- With the intercept, the model parameters are now different from the first example.
- The fit of the model, however, is the same.

Dummy Coded Regression - X_1 and X_3

- For our third example analysis, consider the regression of Y on X_1 and X_3.

$$Y = a + b_3 X_3 + e$$

- $a = 17$
- $b_3 = -4$
- $\sum y_2 = 100$
- $SS_{res} = \mathbf{X'X} = 60$
- $SS_{reg} = 100 - 60 = 40$
- $R^2 = \frac{40}{100} = 0.4$

Dummy Coded Regression - X_1 and X_3

- $a = 17$ is the mean for the E category.
- $b_3 = -4$ is the mean difference between the C category and the E category.
- The E category is called reference category.
- For members of the E category:

$$Y' = a + b_3 X_3 = 17 - 4(0) = 17$$

- For members of the E category:

$$Y' = a + b_3 X_3 = 17 - 4(1) = 13$$

- With the intercept, the model parameters are now different from the first example.
- The fit of the model, however, is the same.

Hypothesis Test of the Regression Coefficien

- Because each model had the same value for R^2 and the same number of degrees of freedom for the regression (1), all hypothesis tests of the model parameters will result in the same value of the test statistic.

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{0.4/1}{(1-0.4)/(10-1-1)} = 5.33$$

- From Excel ("=fdist(5.33,1,8)"), $p = 0.0496$.
- If we used a Type-I error rate of 0.05, we would reject the null hypothesis, and conclude the regression coefficient for each analysis would be significantly different from zero.

Hypothesis Test of the Regression Coefficien

- Recall from the t-test of the mean difference, $t = 2.321$
- For the test of the coefficient, notice that $F = t^2$.
- Also notice that the p-values for each hypothesis test were the same, $p = 0.0496$.
- The test of the regression coefficient is equivalent to running a t-test when using a single categorical variable with two categories.

Multiple Categories (> 2)

- Generalizing the concept of dummy coding, we revisit our first example data set, the cereal experiment data.
- Recall that there were four different types of cereal boxes.
- A dummy coding scheme would involve creation of four new column vectors, each representing observations from each box type.
- Just as was the case with two categories, a linear dependency is created if we wanted to use all four variables.
- Therefore, we must choose which category to remove from the analysis.

One-Way Analysis of Variance

- Just as was the case for the example with two categories, a multiple category regression model with a single categorical independent variable has a direct link to a statistical test you may be familiar with.
- The regression model tests for mean differences across all pairings of category levels simultaneously.
- Testing for a difference between multiple groups (> 2) equates to a one-way ANOVA model (for a model with a single categorical independent variable).

Y	X_1	X_2	X_3	X_4	X_5	Type
11	1	1	0	0	0	1
17	1	1	0	0	0	1
16	1	1	0	0	0	1
14	1	1	0	0	0	1
15	1	1	0	0	0	1
12	1	0	1	0	0	2
10	1	0	1	0	0	2
15	1	0	1	0	0	2
19	1	0	1	0	0	2
11	1	0	1	0	0	2
23	1	0	0	1	0	3
20	1	0	0	1	0	3
18	1	0	0	1	0	3
17	1	0	0	1	0	3
19	1	0	0	1	0	3
27	1	0	0	0	1	4
33	1	0	0	0	1	4
22	1	0	0	0	1	4
26	1	0	0	0	1	4
28	1	0	0	0	1	4

Breakfast Cereal Example

- To make things interesting, let's drop X_5 from our analysis.

$$Y = a + b_2 X_2 + b_3 X_3 + b_4 X_4 + e$$

- Because X_5 (representing box type four) was omitted from our model, the estimated intercept parameter now represents the mean for group X_5.
- All other parameters represent the difference between their respective category level and category level four with respect to the dependent variable.
- $a = 27.2$
- $b_2 = -12.6$
- $b_3 = -13.8$
- $b_4 = -7.8$

Breakfast Cereal Example

- Therefore:

$$\bar{Y}_A = Y'_A = a + b_2(1) + b_3(0) + b_4(0) = 27.2 - 12.6 = 14.6$$

$$\bar{Y}_B = Y'_B = a + b_2(0) + b_3(1) + b_4(0) = 27.2 - 13.8 = 13.4$$

$$\bar{Y}_C = Y'_C = a + b_2(0) + b_3(0) + b_4(1) = 27.2 - 7.8 = 19.4$$

$$\bar{Y}_D = Y'_D = a + b_2(0) + b_3(0) + b_4(0) = 27.2$$

Hypothesis Test

- To test that all means are equal to each other ($H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$) against the hypothesis that at least one mean differs (H_1 : At least one $\mu_{\cdot} \neq \mu_{\cdot'}$), called an omnibus test, the same hypothesis test from before can be used:

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{0.4/1}{(1-0.4)/(10-1-1)} = 5.33$$

- $\sum y^2 = 1013.0$
- $SS_{res} = 158.4$
- $SS_{reg} = 1013.0 - 158.4 = 854.6$
- $R^2 = 854.6/1013.0 = 0.844$

Hypothesis Tests

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{0.844/3}{(1-0.844)/(20-3-1)} = 28.77$$

- From Excel ("=fdist(28.77,3,16)"), $p = 0.000001$.
- If we used a Type-I error rate of 0.05, we would reject the null hypothesis, and conclude that at least one regression coefficient for this analysis would be significantly different from zero.
- Having a regression coefficient of zero means having zero difference between two means (reference and specific category being compared).
- Having all regression coefficients of zero means absolutely no difference between any of the means.

Final Thought

- Regression with categorical variables can be accomplished by coding schemes.
- Differing ways of coding (or inclusion of certain coded column vectors) may change the interpretation of the model parameters, but will not change the overall fit of the model.
- But what can be said for comparing pair-wise mean differences in a simultaneous regression model...I will leave you in suspense...

Next Time

- Chapter 15: ANCOVA.