

**DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816**

MULTIPLE REGRESSION WITH CATEGORICAL DATA

- I. AGENDA:
- A. Multiple regression with categorical variables
 - 1. Coding schemes
 - 2. Interpreting coefficients
 - 3. Interaction
 - B. Reading: Agresti and Finlay *Statistical Methods in the Social Sciences*, 3rd edition, Chapter 12, pages 449 to 462.
- II. CATEGORICAL INDEPENDENT VARIABLES:
- A. Problem: what does sex discrimination in employment mean and how can it be measured?
 - B. To answer these questions consider these artificial data pertaining to employment records of a sample of employees of Ace Manufacturing:

i	Sex	Merit Pay	i	Sex	Merit Pay
Bob	M	9.6	Tim	M	6.0
Paul	M	8.3	George	M	1.1
Mary	F	4.2	Alan	M	9.2
John	M	8.8	Lisa	F	3.3
Nancy	F	2.1	Anne	F	2.7

- C. Data: here the dependent variable, Y, is merit pay increase measured in percent and the "independent" variable is sex which is quite obviously a nominal or categorical variable.
- D. Our goal is to use categorical variables to explain variation in Y, a quantitative dependent variable.
 - 1. We need to convert the categorical variable gender into a form that “makes sense” to regression analysis.
- E. One way to represent a categorical variable is to code the categories 0 and 1 as follows:

let X = 1 if sex is "male"
0 otherwise

- 1. Example: Bob is scored "1" because he is male; Mary is 0.
- F. Called **dummy variables**, data coded according this 0 and 1 scheme, are in a sense arbitrary but still have some desirable properties.
 - 1. A dummy variable, in other words, is a numerical representation of the categories of a nominal or ordinal variable.
- G. Interpretation: by creating X with scores of 1 and 0 we can transform the above table into a set of data that can be analyzed with regular regression. Here is what the "data matrix" would look like prior to using, say, MINITAB:

i (c1)	Sex (c2)	Merit Pay (c3)	i (c1)	Sex (c2)	Merit Pay (c3)
Bob	1	9.6	Tim	1	6.0
Paul	1	8.3	George	1	1.1
Mary	0	4.2	Alan	1	9.2
John	1	8.8	Lisa	0	3.3
Nancy	0	2.1	Anne	0	2.7

- H. Except for the first column, these data can be considered numeric: merit pay is measured in percent, while gender is "dummy" or "binary" variable with two values, 1 for "male" and 0 for "female."
 - 1. We can use these numbers in formulas just like any data.
 - 2. Of course, there is something artificial about choosing 0 and 1, for why couldn't we use 1 and 2 or 33 and 55.6 or any other pair of numbers?
 - 3. The answer is that we could. Using scores of 0 and 1, however, leads to particularly simple interpretations of the results of regression analysis, as we'll see below.

III. INTERPRETATION OF COEFFICIENTS:

- A. If the categorical variable has K categories (e.g., region which might have K = 4 categories--North, South, Midwest, and West) one uses K - 1 dummy variables as seen later.
- B. Using regular OLS analysis the parameter estimators can be interpreted as usual: a one-unit change in X leads to β_1 change in Y.
- C. But given the definition of the variables a more straight forward interpretation is possible. The model is:

$$E(Y_i) = \beta_0 + \beta_1 X_1$$

1. The model states that the expected value of Y--in this case, the expected merit pay increase--equals β_0 plus β_1 times X. But what are the two possible values of X?
2. First consider males; that is, $X = 1$. Substitute 1 into the model:

$$\begin{aligned} E(Y_i) &= \beta_0 + \beta_1 X_1 \\ &= \beta_0 + \beta_1(1) \\ &= \beta_0 + \beta_1 \end{aligned}$$

- i. The expected merit pay increase for males is thus $\beta_0 + \beta_1$.
3. Now consider the model for females, i.e., $X = 0$. Again, make the substitution and reduce the equation.

$$\begin{aligned} E(Y_i) &= \beta_0 + \beta_1 X_1 \\ &= \beta_0 + \beta_1(0) \\ &= \beta_0 \end{aligned}$$

- i. We can see from these equations that β_0 is the expected value of Y (remember it's merit pay increase in this example) for those subjects or units coded 0 on X--in this instance it is the expected pay increase of females. Stated differently but equivalently, β_0 is the mean of Y (% pay increase) in the population of units coded 0 on X (i.e., females).
 - 1) That is, β_0 is μ_0 where μ_0 is the mean of the dependent variable for the group coded 0.
 - 2) Remember: the expected value of a random variable is its mean or $E(Y_i) = \mu$
- ii. β_1 is the "effect," so to speak, of "moving" or changing from category 0 to category 1--here of changing from female to male--on the dependent variable.
 - 1) Specifically, if $\beta_1 > 0$, then the expected value of Y is higher for the group 1 members (e.g. males) than group 0 cases (e.g., females). Thus, if $\beta_1 > 0$ then men get higher increases on average than women do.
 - 2) On the other hand, if $\beta_1 < 0$, then group 1 people (units) get less Y than do group 0 individuals. If $\beta_1 < 0$, in other words,

then females receive higher pay increases.

- 3) If $\beta_1 = 0$, then both groups have the same expected value on Y.
4. Hence, knowing the values of β_0 and β_1 tells us a lot about the nature of the relationship.
- D. Using 0 and 1 to code gender (or any categorical variable) thus leads to particularly simple interpretations.
 1. If we used other pairs of numbers, we would get the “correct” results but they would be hard to interpret.
- E. Example: to put some "flesh" on these concepts suppose we regressed merit pay (Y) on gender.
 1. We obtain the estimated equation:

$$\hat{Y} = 3.08 + 4.09X_1$$

$$R^2 = .428$$

2. The estimate of the average merit pay increase for women in the population is 3.08 percent. (Let $X = 0$ and simplify the equation.)
3. Men, on average, get 4.09 percent more than women. hence their average increase is:

$$\begin{aligned} \hat{Y}_{men} &= 3.08 + 4.09(1) \\ &= 3.09 + 4.09 = 7.17 \end{aligned}$$

4. The "effect" of being male is 4.09 percent greater merit pay than what women get.
5. Here is a partial regression ANOVA table:

Source	df	SS	MS	F _{obs}
Regression (sex)	1	40.18	40.18	5.89
Residual	8	54.581	6.823	
Total variation in Y	9	94.761		

6. At the .05 level, the critical value of F with 1 and 8 degrees of freedom is 5.32. Thus, the observed F is barely significant. Since the critical F at the .01 level is 11.26, the result (the observed "effect" of Y that is) has a probability of happening by chance of between .05 and .01.

IV. ANOTHER EXAMPLE:

- A. Here is a problem to consider. A law firm has been asked to represent a group of women who charge that their employer, GANGRENE CHEMICAL CO., discriminates against them, especially in pay. The women claim that salary increases for females are consistently and considerably lower than the raises men receive. GANGRENE counters that increases are based entirely on job performance as measured by an impartial "supervisor rating of work" evaluation which includes a number of performance indicators. You have been asked by the law firm to make a preliminary assessment of the merits of the claim. To begin with, you draw a random sample from the company's files:

File No.	Sex	Quality of work score	Years Experience	Division	Salary Increase
1	F	10	9	Production	\$21
2	F	90	1	Production	96
3	F	20	4	Production	47
4	F	80	1	Production	128
5	F	30	4	Research	64
6	F	70	1	Research	52
7	F	10	4	Sales	73
8	F	15	7	Production	19
9	M	20	6	Research	128
10	M	80	3	Sales	474
11	M	50	3	Research	342
12	M	70	2	Sales	330
13	M	30	7	Sales	185
14	M	70	7	Sales	331
15	M	40	1	Sales	267
16	M	90	6	Production	517
17	M	50	8	Production	390

- B. The data:
1. Salary increase is measured in extra dollars per month.
 2. The quality of work index ranges from 0 (lowest) to 100 (highest) rating.
 3. “Division” is divisions within the company
- C. Questions:
1. Even a cursory glance at the data reveal that men get higher increases than women. But the real question is why?
 2. Notice for example that men differ from women on other factors such as experience, division, and job performance evaluations. Are the differences in salary increases due to these factors?
 3. Another problem: suppose raises are tied solely to performance ratings. Is there discrimination in these evaluations.
- D. Preliminary model:

$$E(Y_i) = \beta_0 + \beta_X X + \beta_Z Z + \beta_W W$$

1. X is gender coded:
 1 if female
 0 otherwise
 2. Z is job performance evaluation (i.e., quality of work)
 3. W = XZ, an **interaction** term (see below)
- E. Interaction (W):
1. The interaction term has this meaning or interpretation: consider the relationship between Y and Z. So far in this course, this relationship has been measured by β_Z , the regression coefficient of Y on Z. This coefficient is a partial coefficient in that it measures the impact of Z on Y when other variables have been held constant. But suppose the effect of Z on Y depends on the level of another variable, say X. Then, β_Z by itself would not be enough to describe the relationship because there is no simple relationship between Y and Z. It depends on the level of X. This is the idea of **interaction**.
 2. To be more specific, suppose the relationship between work performance (Z) and pay increase (Y) depends on a worker's sex. That is, suppose a one-unit increase in quality of work performance evaluations for women brings a \$1.00 increase in salary but a one unit increase in Z brings a \$2.00 increase for men. In this case, the effect of Z, quality of work, depends on or is affected by sex.
 3. To test this idea we have to create an "interaction" variable by multiplying Z by X:

$$W = XZ = (\text{Gender}) \times (\text{Work index})$$

4. The variable can be added to the model. If it turns out to be non-significant or does not seem to add much to the model's explanatory power, then it can be dropped. Dropping the interaction term in this context amounts to saying that the job performance rating has the **same** impact on salary increases for both sexes. If, on the other hand, there is a difference in effects of Z, the interaction term will explain some of the variation in Y.
5. The important point is that W has a substantive meaning. In this context it would indicate the presence of a second kind of discrimination:
 - i. One type of discrimination is that on average women get smaller raises.
 - ii. Another type is that extra increments in job evaluations bring less rewards for women than men.

F. Initial OLS estimates of the model:

1. This model is called **complete** because it contains X, Z, and W:

$$\hat{Y} = 59.94 - 29.71X + 4.84Z - 4.05W$$

- i. The R^2 for the complete model is .941.
2. Interpretation of the parameters:
 - i. The coefficients can be interpreted in the usual way. But using the logic presented earlier in conjunction with the definition of the dummy variables, there are more straightforward interpretations.
 - ii. For men the estimated equation is

$$\hat{Y} = 59.94 + 4.84Z$$

*because $X = W = 0$
(remember $X = 0$ for men and hence W also equals 0)*

3. Thus when Z (job performance) is zero, men can expect on average to get a salary increase of \$59.94. For each one-unit increase in their job evaluations they get an extra \$4.84.
4. Now look at the equation for women:

$$\begin{aligned} \hat{Y} &= (59.94 - 29.71) + (4.84 - 4.05)Z \\ &= + 30.23 + .79Z \end{aligned}$$

- i. For women $X = 1$ and hence $-29.71X = -29.71$ and so forth. Work through the equations yourself to make sure you grasp what is going on.

5. Therefore, women can expect an average salary increase of only \$30.23 when Z is zero **and** a one-unit increase in the job evaluation index only brings an extra 79 cents in raises.
6. These figures suggest that there are indeed two types of discrimination working in the company:
 - i. For men β_0 is \$59.94 whereas for women it is \$30.23, a difference of almost \$30.
 - ii. For men, furthermore, β_Z , which measures the return on work performance, is \$4.85 while for women the return is only \$.79.
7. Here is a diagram that illustrates these ideas. (It is not drawn to scale.)

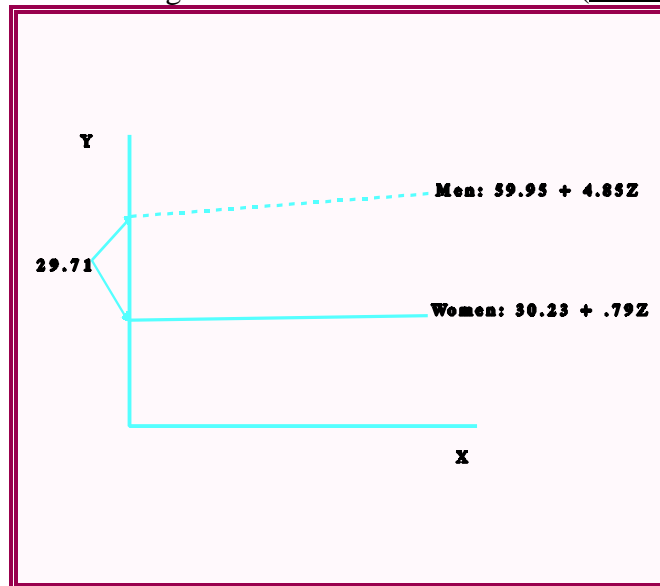


Figure 1: Interaction Effect

V. TESTING FOR INTERACTION:

- A. The estimates are different but since this is a small sample one will want evidence that the differences are not due to sampling error. We thus need to test the significance of β_w .
 1. If it is not significant we might want to drop it from the model.
 2. That would mean that at least one type of discrimination was not operative.
- B. The strategy is this:
 1. We use an “extra” or “added” sum of squares logic.
 2. Does the inclusion in a model of a variable or set of variables significantly increase the explained sum of squares over what we would expect from chance.
 3. So we will compare two models’ R^2 ’s:
 - i. The first will be for the model containing all of the variables, a model we call **complete**.
 - ii. The second is the R^2 for the model without the variables of interest.

It's called the **reduced** model.

- C. One way to do this is to estimate a **complete** model as above and obtain an observed F for it. Here, the F for the complete model is 69.0668 with 3 and 13 degrees of freedom. Next, estimate a **reduced** model, that is a model in which W has not been included:

$$E(Y) = \beta_0 + \beta_X X + \beta_Z Z$$

- i. This model will also have an observed F. The **difference** between the F's can be evaluated by the formula:

$$F_{obs} = \left(\frac{(R_{com}^2 - R_{reduced}^2)}{(1 - R_{com}^2)} \right) \times \left(\frac{(N - K - 1)}{g} \right)$$

- D. Here R_{com}^2 is the R^2 for the **complete** model and $R_{reduced}^2$ is the R^2 for the **reduced** model, K is the number of independent variables in the complete model counting interaction (here 3), and g is the number of variables left out of the reduced model (here 1).

1. Believe it or not, the previous formula is just an application of the ratio of mean squares we discussed earlier in the semester when describing the F test.
 - i. The formula is the ratio of two estimators of the error variance.
 - 1) The first is the residual sum of squares divided by its degrees of freedom.
 - 2) The error or residual sum of squares is equivalent to $1 - R_{comp}^2$
 - 3) When divided by degrees of freedom $N - K - 1$, we have an estimator of the error standard deviation.
 - ii. It can be shown that the component $R_{comp}^2 - R_{reduced}^2$ divided by its degrees of freedom, g, is an independent and unbiased estimator of the error standard deviation, **under** the hypothesis that the extra explained sum of squares added by the additional variables is nil or zero.
2. The R's are obtained from MINITAB (or SPSS) in the usual way.
 - i. I just regress Y on all of the variables including the interaction term and noted the R^2

- ii. I then drop the interaction term and get the R^2 for the “reduced” model.
- 3. For this example the F is, where $g = 1$ and $N - K - 1 = 13$:

$$F_{obs} = \left(\frac{(.9409 - .8340)}{(1 - .9409)} \right) \times \left(\frac{(17 - 3 - 1)}{1} \right) = 23.51$$

- i. The critical F for 1 and 13 degrees of freedom is 17.81 at the .001 level.
- ii. Therefore, include interaction term measured by β_w is statistically significant (at the .001 level) and should be included in the model.
 - 1) That is, the observed F exceeds the critical F at the .001 level.
 - 2) In other words, the interaction term adds “significantly” to the explanatory power of the model.
 - 3) This means in turn that there is evidence of the second kind of discrimination.

VI. NEXT TIME:

- A. More on dummy variables and multiple regression.
- B. Transformations
- C. Regression “diagnostics”

Go to Notes page

Go to Statistics page