IIA-3 Econometrics: Additional Notes

Emre Usenmez

Easter Break 2025

Notes are based on the following:

Hardy (1993), *Regression with Dummy Variables*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-093, Newbury Park, CA: Sage

Halvorsen, R, and Palmquist, R (1980) *Interpretation of dummy variables in semilogarithmic equations*, American Economic Review 70:474-475.

Interpretation of Regression with Dummy Variables

We will use income difference as an example.

Let's begin the discussion with descriptive statistics.

Descriptive Statistics

Given that dummy variables provide qualitative information, the mena value of a dummy variable indicate the proportion, or relative frequency, of cases in the category coded as 1. This is because the mean value is calculated by adding up all the cases and dividing it by the number of cases. Since all the cases are either 1 or 0, this ends up being the same as the relative frequency. Therefore, for dummy variables, the formulas for the mean and for a proportion are equivalent.

Similarly, the formula for the variance of a dummy variable can be related to the more general variance formula for continous measures. When X_i is continuous:

$$Var(X_i) = \frac{\sum X_i^2}{N} - \Big(\frac{\sum X_i}{N}\Big)^2$$

and when X_i is a dummy variable, $\sum X_i^2$ reduces to n_j, the number of cases coded as 1. Thus, if we denote the proportion of cases coded as 1 as $p_j = \frac{n_j}{N}$ then the variance for a dummy variable becomes:

$$Var(X_i) = p_j - p_j^2 = p_j(1 - p_j).$$

That is, the variance for a dummy variable is the product of the proportion of cases coded as 1 and its complement. This means, the maximum variability in a dummy variable is obtained when the number of cases are evenly split between 1 and 0.

Dummy Explanatory Variables

Suppose we are interested in looking at potential income differences between left and right handed people.

Consider the model

$$\text{Model 1: } Y_i = f(\text{handedness}) = \beta_0 + \beta_1 \text{ left}_i + u_i.$$

When the independent variable is continuous, the distribution of predicted values of the dependent variable are also continuous and so the regression coefficient indicates a slope. On the other hand when it is a dummy variable, the predicted value of the dependent variable changes by the estimated coefficient each time membership in a specified category is switched on or off.

Using some data we obtained from a survey of 3,211 respodents (2,290 of which are right handed), suppose we estimate the following relationship:

$$\begin{aligned} Y_i &= 7,821.9 - 3,202.9 \text{ left} \\ se &: (91.9) \quad (171.6) \\ R_1^2 &: 0.09792 \\ F_1 &: 348.3 \end{aligned}$$

This means predicted income for respondents who are left handed is £4,619.00 which is £3,202.90 less than predicted income of £7,821.9 for right handed people. These are also the same as the mean values of each group. That is, if we average the income for all left handed people in the survey, we would get £4,619.00, and averaging right handed people would yield £7,821.90.

The coefficient of the independent variable measures the effect of being left handed on income. Since that is the case, the standard error of this coefficient is then the standard error of the difference between expected income for right handed people and expected income for left handed people.

When testing against a null hypothesis that there are no difference in expected income between the two groups, i.e. zero effect, the t test reduced to the ratio of the coefficient of the standard error. Since this model has only one independent variable, the F test for the model is a test of the same null hypothesis of zero effect. The value of F is the square of t value. This means, the null hypothesis that $\beta_1 = 0$ is equivalent to the null hypothesis $\mathbb{H}_0 : \mu_{\text{left}} - \mu_{\text{right}} = 0$. Both t test for $\hat{\beta}_1$ and the F test for the model itself are essentially difference of means tests.

Suppose now we instead want to estimate income as a function not of handedness but occupational categories: upper white-collar jobs (*upwc*), lower white-collar (*lowwc*), skilled craftsmen (*skill*), operatives such as welders, stitchers in manufacturing etc. (*oper*), service workers such as barbers, janitors (*serv*), and laborers (*labor*):

Model 2: $Y_i = f(\text{handedness, occupation}) = \beta_0 + \beta_1 \text{ lowwc}_i + \beta_2 \text{ skill}_i + \beta_3 \text{ oper}_i + \beta_4 \text{ serv}_i + \beta_5 \text{ labor}_i + u_i$.

Notice that we are using upper white-collar workers as the reference group and regress without it to avoid perfect multicollinearity. Suppose again that we use data from a survey with 3211 sample points and we now estimate the following relationship:

$$\begin{aligned} Y_i &= 10,702.1 - 3,021.2 \text{ lowwc} - 3,757.1 \text{ skill} - 5,148.2 \text{ oper} - 6,267.7 \text{ serv} - 6,612.1 \text{ labor} \\ se &: (160.8) \quad (274.4) \quad (215.5) \quad (216.8) \quad (289.7) \quad (272.3) \\ R_2^2 &: 0.224 \\ F_2 &: 185.0 \end{aligned}$$

The constant reports the expected income for the reference group, upper white-collar workers, to be £10,702.1. The remaining coefficients report the effect of being in a particular occupational category compared with the reference category. So the coefficient $\hat{\beta}_1$ indicates that on average lower white-collar workers earn £3,021.20 less than upper white-collar workers, i.e. they earn on average £7,680.90. Similarly, laborers earn £6,612.10 less than upper white-collar workers on average, or £4,090.00.

In terms of testing for the effect of occupational category on income, notice that distinctions among occupational groups are captured by the entire set of dummy variables and not by any single variable. Therefore, F test would be appropriate here with the null hypothesis that $\beta_1 = \cdots = \beta_5 = 0$, meaning that F test is a test of the hypothesis that the expected value of income (Y_i) for all occupational groups is the same.

Also notice that F test can be a test of significance of R^2 here. This is because the F-test can be expressed as the ratio of R^2/k to $(1-R^2)/(N-k-1)$. Therefore, if null hypothesis is rejected then a nonzero amount of variation in income is explained by the respondent's occupational category. In this example,

$$F_{5,3205} = \frac{0.224/5}{(1-0.224)/(3211-5-1)} = 185$$

which is significant at better than .001 level. Thus occupation is significant.

Since occupation is significant, we can now test if the expected income for each occupational category is significantly different from that of the reference group using t test. That is, with t test on the coefficients we are checking if the effect of being in the designated category rather than in the reference group is significant.

One other thing to check here is whether the occupational categories are different from each other. That is, we need to check if, for example, the expected income for laborers ($\hat{\beta}_6$) is different, or indeed smaller, than the expected income for operatives ($\hat{\beta}_4$) for example. To test this, notice that $\beta_j = \mathbb{E}(Y_i \mid DUMMY_j = 1) - \mathbb{E}(Y_i \mid REFERENCE)$. Because of this, the difference in expected income for included categories is equal to the difference between their coefficients (i.e. $\beta_j - \beta_k$). So, for example to test for a difference in the effects of laborer and operatives, we'd use a t test for the difference in regression coefficients:

$$t = \frac{\beta_j - \beta_k}{\sqrt{Var(\beta_j) + Var(\beta_k) + 2Cov(\beta_j\beta_k)}} = \frac{-6,612.1 - (-5,148.2)}{\sqrt{Var(\beta_5) + Var(\beta_3) + 2Cov(\beta_5\beta_3)}}.$$

We can do similar operations for other comparisons.

We can put together these two qualitative measures, handedness and occupation, to see if handedness differences in income persists when we control for income differences in occupation. The model then becomes:

$$\begin{aligned}\text{Model 3: } Y_i &= f(\text{handedness, occupation}) \\ &= \beta_0 + \beta_1\ \text{left}_i + \beta_2\ \text{lowwc}_i + \beta_3\ \text{skill}_i + \beta_4\ \text{oper}_i + \beta_5\ \text{serv}_i + \beta_6\ \text{labor}_i + u_i.\end{aligned}$$

where the upper white-collar is still the reference occupation category.

Once more suppose that we use data from a survey with 3211 respondents. Our estimation now becomes:

$$\begin{array}{lccccc} Y_i = 10,811.4 & -1,676.0\ \text{left} & -2,842.1\ \text{lowwc} & -3,566.4\ \text{skill} & -4,604.5\ \text{oper} & \\ se: (158.9) & (172.4) & (271.1) & (213.3) & (220.9) & \\ & -5,512.7\ \text{serv} & -5,647.8\ \text{labor} & & & \\ & (295.9) & (286.2) & & & \end{array}$$

$$R_3^2 : 0.24624$$

$$F_3 : 174.4$$

The constant indicates an expected income when all independent variables are set to zero, so it tells us that the expected income for right handed upper white-collar workers is £10,811.40.

The coefficient for `left` indicates that once the variation in income linked to occupational category is taken into account along with the fact that handedness is not uniformly distributed across all occupational categories, left handed people still average £1,676.00 less income than right handed people. This is a reduction from the difference in earnings of £3,202.90 when handedness was the only explanatory variable. The reduction in the magnitude of the coefficient of `left` suggests that one reason left handed workers average lower incomes is because they are concentrated in occupations that in general commands lower salaries or earnings.

Similarly, the partial regression coefficients associated with occupation dummy variables estimate the effect on expected income of membership in each of the designated categories rather than the reference upper white-collar group, controlling for handedness differences in both income and the distribution of respondents across occupational categories.

To decide whether the partial effects of handedness or the partial effects of occupation, while controlling for other variables, are statistically significant, an F test would be appropriate. However, we would not use the F test for the model as a whole but instead use the incremental F test. Suppose we want to assess whether occupational categories contribute to this model. For this we can compare the R^2s of this third model, and the first model where `left` was the only explanatory variable. The null hypothesis is that once handedness differences are controlled for in both income and occupational category, the expected value of earned income is the same across occupational categories; i.e. $\beta_2 = \cdots = \beta_6 = 0$. The test for the significance of occupation controlling for handedness is:

$$F_{5,3204} = \frac{\frac{R_3^2 - R_1^2}{k_3 - k_1}}{\frac{1 - R_3^2}{N - k_3 - 1}} = \frac{\frac{0.24624 - 0.09792}{6 - 1}}{\frac{1 - 0.24624}{3211 - 6 - 1}} = 126.1$$

Here, the numerator calculates the increments to R^2 that results from specifying the effects of occupational category relative to the difference in the number of independent variables between the two models.

The denominator is the proportion of variance left unexplained when both handedness and occupation are included divided by the degrees of freedom.

Notice that this last model estimates 12 different values for the predicted income. This is because handedness has 2 categories and occupation has 6 categories, together they generate 12 distinct subgroups. With one mean per each handedness-by-occupation subgroups, these predicted values correspond to 12 subgroup means. However we are making a simplifying assumption that the estimated income difference between left handed people and right handed ones (i.e. the effect of `left`) is the same across all occupational groups. In other words, we assume that the income differences across occupational groups are the same for left handed and right handed people. In our model so far, the difference between left handed and right handed workers is always $\hat{\beta}_1 = 1,676$ regardless of occupation. So for example a left handed skilled worker is expected to earn $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 =$ £5,569.00 while a right handed skilled worker is expected to earn $\hat{\beta}_0 + \hat{\beta}_3 =$ £7,245.00 on average. The difference is $\hat{\beta}_1 =$ £1,676.

We will tackle this later on but first lets see what happens when we add some quantitative variables to our model.

Dummy and Quantitative Explanatory Variables

Suppose now income is a function of not only handedness and occupation but also education and employment duration which are measured in years:

$$\begin{aligned}\text{Model 4: } Y_i &= f(\text{handedness, occupation, education, tenure}) \\ &= \beta_0 + \beta_1\ \text{left}_i + \beta_2\ \text{lowwc}_i + \beta_3\ \text{skill}_i + \beta_4\ \text{oper}_i + \beta_5\ \text{serv}_i + \beta_6\ \text{labor}_i + \beta_7\ \text{educ}_i + \beta_8\ \text{dur}_i + u_i.\end{aligned}$$

Suppose again that after estimating using data from a survey with 3211 respondents, our estimation now becomes:

$$\begin{array}{lccccc} Y_i = 5,761.1 & -1,188.1\ \text{left} & -2,316.1\ \text{lowwc} & -2,343.7\ \text{skill} & -3,166.6\ \text{oper} \\ se: (359.0) & (169.4) & (261.8) & (223.7) & (237.5) \\ & -3,918.5\ \text{serv} & -3,606.8\ \text{labor} & +282.0\ \text{educ} & +84.7\ \text{dur} \\ & (299.9) & (306.4) & (23.1) & (6.6) \end{array}$$

$$R_4^2 : 0.31459$$

$$F_4 : 183.7$$

The intercept is now the expected income for right handed upper white-collar workers who have 0 years of schooling and 0 years of working in their occupation. Once the variation in income due to occupation,

education, and duration of employment is partialled out, the expected income for left handed and right handed people differ by £1,888.10. Coefficients for each occupational dummy variables estimate the net difference in expected income for each occupational group relative to the reference group. So, for example, skilled workers earn £2,343.70 less than upper white-collar workers on average. Similarly, with handedness, occupation, and education held constant, each additional year on the job translates into another £84.70 in earnings. Under similar conditions, an additional year of education is associated with an increase of £282.00 in expected income.

Assessing Group Differences

The models thus far share an assumption that the effect of any single explanatory variable is the same across the range of other explanatory variables. That is, as highlighted earlier, there is an assumption that the effect of `left` is the same across all occupational groups. To test the validity of this assumption we can introduce an interaction term. In order to test for interaction effects, we add five interaction terms to the model that multiplies `left` with each of the occupational dummy variables:

$$\begin{aligned}\text{Model 5: } Y_i &= f(\text{handedness, occupation, education, tenure})\\ &= \beta_0 + \beta_1\ \text{left}_i + \beta_2\ \text{lowwc}_i + \beta_3\ \text{skill}_i + \beta_4\ \text{oper}_i + \beta_5\ \text{serv}_i + \beta_6\ \text{labor}_i + \beta_7\ \text{educ}_i + \beta_8\ \text{dur}_i\\ &\quad + \beta_9\ \text{leftlow}_i + \beta_{10}\ \text{leftskill}_i + \beta_{11}\ \text{leftop}_i + \beta_{12}\ \text{leftserv}_i + \beta_{13}\ \text{leftlab}_i + u_i.\end{aligned}$$

Here, the interaction term `leftop` for example, would be coded as 1 if the respondent to the survey is both left handed and an operative such as a welder, stitcher in manufacturing, etc.

Now suppose our estimation becomes:

$$\begin{aligned}Y_i = 5,794.8 &- 3,793.3\ \text{left} - 2,274.9\ \text{lowwc} - 2,418.4\ \text{skill} - 3,427.2\ \text{oper}\\ se: (358.7)\quad &(610.1)\quad (280.2)\quad (232.7)\quad (256.3)\\ &- 4,513.4\ \text{serv} - 4,202.8\ \text{labor} + 292.9\ \text{educ} + 84.0\ \text{dur}\\ &(372.5)\quad (399.0)\quad (23.1)\quad (6.6)\\ &+ 1,501.2\ \text{leftlow} + 2,326.2\ \text{leftskill} + 2,984.8\ \text{leftop} + 3,528.0\ \text{leftserv}\\ &(823.0)\quad (705.0)\quad (672.5)\quad (761.0)\\ &+ 3,383.9\ \text{leftlab}\\ &(747.3)\end{aligned}$$

$$\begin{aligned}R_5^2 &: 0.32138\\ F_5 &: 116.46\end{aligned}$$

First, we would like to check if our approach of allowing for differential effects of handedness and occupation resulted in statistically significant improvement in model's fit. Just as we did earlier when we assessed whether the partial effects of occupation improved the model, we use incremental F-test where we compare the R^2s of this model and the model without the interaction terms. The null hypothesis is that once all the other variables are controlled for, the expected value of earned income is the same across all interaction terms; i.e. $\beta_9 = \cdots = \beta_13 = 0$. The test therefore is:

$$F_{5,3197} = \frac{\frac{R_5^2 - R_4^2}{k_5 - k_4}}{\frac{1 - R_5^2}{N - k_5 - 1}} = \frac{\frac{0.32138 - 0.31459}{13 - 8}}{\frac{1 - 0.32138}{3211 - 13 - 1}} = 6.4$$

which is statistically significant at better than 0.001 level. Although the increment to explanatory power is far from overwhelming, the F-test suggests that the large sample size has enabled us to estimate the differential effects with reasonable accuracy.

Interpretation:

- Intercept still has similar interpretation as Model 4 whereby the expected income for right handed upper white-collar workers who have 0 years of education and 0 years of working in their occupation is £5,794.80.
- In order to have a better understanding of the coefficients of the occupational groups - and to simplify things a bit - suppose the coefficients for 'educ' and 'dur' are 0. Let's map out the 12 handedness-by-occupation subgroups as follows:

	Right	Left
Upper white-collar	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_1$
Lower white-collar	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_9$
Skilled	$\hat{\beta}_0 + \hat{\beta}_3$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_{10}$
Operative	$\hat{\beta}_0 + \hat{\beta}_3$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 + \hat{\beta}_{11}$
Service	$\hat{\beta}_0 + \hat{\beta}_5$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_5 + \hat{\beta}_{12}$
Laborer	$\hat{\beta}_0 + \hat{\beta}_6$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_6 + \hat{\beta}_{13}$

  Notice that the coefficient for 'left' no longer provides an estimate of the average effect of left handedness across all occupational groups as it did in Model 4. Recall in that model the expected income for left and right handed people differed by £1,888.10 once the variation in income due to occupation, education, and duration of employment are partialled out. Here in Model 5 $\hat{\beta}_1$ instead estimates the difference in income between left and right handed upper white-collar workers. The t-test for this coefficient is then a test of the null hypothesis that 'left' has no significant effect on expected income - net of other variables - for upper white-collar workers. That is, the null hypothesis is that expected income for left handed upper white-collar workers is equal to the expected income for right handed upper white-collar workers, after controlling for variation in income due to education and duration of employment. If this null hypothesis is rejected, then we learn that left handed workers among upper white-collar workers average lower incomes than the right handed workers among upper white-collar workers, controlling for all other factors specified in the model.
- Similarly, the coefficients for the occupation dummy variables no longer provide an estimate of the average effect of being in a particular occupational category versus the reference group as it did in Model 4. Recall that in that model the expected incomes for skilled workers were £2,343.70 less than the upper white-collar workers on average, irrespective of handedness (ie left and right handed workers together in each category). Here in Model 5,

  $\hat{\beta}_2$ instead estimates the difference in expected earnings between lower white-collar and upper white-collar workers who are right handed. Right handed lower white-collar workers average £2,274,90 less in income than right handed upper white-collar workers.

  $\hat{\beta}_3 = -2,418.4$ means, right handed skilled laborers earn on overage £2,418.40 less income than right handed upper white-collar workers.

  That is, once interaction terms are specified, the coefficients for the original set of variables (in this example, 'left', 'lowwc', 'skill', 'oper', 'serv', and 'labor') refer to comparisons involving the reference categories:

  $\hat{\beta}_1$ measures the effect of being left handed for the reference category for occupation, i.e. upper white-collar workers;

  $\hat{\beta}_2$ through $\hat{\beta}_6$ measure the effects of being in an occupational category other than upper white-collar for the right handed workers, i.e. the reference category for handedness.

  The t tests associated with the regression coefficients $\hat{\beta}_2$ through $\hat{\beta}_6$ are therefore tests for significant differences among occupational groups for right handed workers that can be generalized to the population.
- The coefficients for the interaction terms estimate the differential effect of occupation by handedness. Alternatively, these coefficients can also be seen as estimates of the differential effect of being left handed

by occupational category.

This is because the difference in predicted income between, say, lower white-collar workers and upper white-collar workers is captured by $\hat{\beta}_2 = -2,274.9$ for right handed workers and by $\hat{\beta}_2 + \hat{\beta}_9 = -2,274.9 + 1,501.2 = -773.7$ for left handed workers. Therefore, $\hat{\beta}_9 = 1,501.2$ estimates the difference in the effect of being lower white-collar worker for left handed workers relative to right handed ones. So, the earnings gap between lower white-collar workers and upper white-collar workers is £1,501.20 narrower for left handed workers than right handed ones, or -£773.70 rather than -£2,274.90.

- The difference in expected income between left and right handed workers who are upper white-collar workers is $\hat{\beta}_1 = -3,793.3$. On the other hand, the left-right difference among lower white-collar workers is $\hat{\beta}_1 + \hat{\beta}_9 = -3,793.3 + 1,501.2 = -2,292.1$.

  Here $\hat{\beta}_9$ estimates the difference in the effect of being left handed for lower white-collar workers relative to upper white-collar workers. So the difference in expected income between left and right handed workers who are lower white-collar workers is -£2,292.10.

  This illustrates that the differences in expected income by ocucpation for left handed workers are captured by the sum of two coefficients: the coefficient of an occupation dummy variable, β_j plus the coefficient of the relevant interaction term, β_{jk}. The connection between β_j and β_{jk} can be defined as follows:

$$\beta_j = \mathbb{E}(Y_i \mid \text{right, occ}_j) - \mathbb{E}(Y_i \mid \text{right, occ}_{\text{ref}})$$

  which is the difference between the expected income given a right handed worker in occupation category j and the expected income given a right handed worker in the reference occupation category. Similarly,

$$\begin{aligned}\beta_{jk} &= \big[\mathbb{E}(Y_i \mid \text{left, occ}_j) - \mathbb{E}(Y_i \mid \text{left, occ}_{\text{ref}})\big] - \big[\mathbb{E}(Y_i \mid \text{right, occ}_j) - \mathbb{E}(Y_i \mid \text{right, occ}_{\text{ref}})\big] \\ &= \big[\mathbb{E}(Y_i \mid \text{left, occ}_j) - \mathbb{E}(Y_i \mid \text{left, occ}_{\text{ref}})\big] - \beta_j\end{aligned}$$

  Rearranging this then gives:

$$\beta_j + \beta_{jk} = \mathbb{E}(Y_i \mid \text{left, occ}_j) - \mathbb{E}(Y_i \mid \text{left, occ}_{\text{ref}}).$$

- The t tests for the coefficients of the interaction terms are therefore testing whether the net income differential between specific occupational groups and the reference group is the same for left and right handed workers.
- If the coefficients for the interaction terms had been negative, we would have had evidence that the earnings differences between upper white collar-workers and remaining occupational groups were larger for left handed workers than right handed workers.

  That is, the occupational differences in earning had been identified for right handed workers through the negative coefficients for the occupation dummy variables. These would have been even larger for left handed workers because of the extra negative effect captured by coefficients for the interaction terms.
- But these coefficients of interaction terms are positive. Therefore, it appears that the differences in earnings across occupational groups are more pronounced for right handed workers, and more compact for left handed workers. In fact, there may very well be no significant occupational differences in income among left handed workers.

  It also seems that the left/right difference in expected income becomes narrower as we move down the occupational scale.

Should the partial effects of education and duration of employment be the same for all subgroups?

This question effectively means testing the hypotheses that $\beta_{\text{educ(left)}} = \beta_{\text{educ(right)}}$ and $\beta_{\text{dur(left)}} = \beta_{\text{dur(right)}}$. Since it is testing the variability of relationships, we would add two new interaction terms to Model 5 that interacts education with left handedness, and duration with left handedness:

Model 6: $Y_i = f(\text{handedness, occupation, education, tenure})$

$$= \beta_0 + \beta_1 \text{ left}_i + \beta_2 \text{ lowwc}_i + \beta_3 \text{ skill}_i + \beta_4 \text{ oper}_i + \beta_5 \text{ serv}_i + \beta_6 \text{ labor}_i + \beta_7 \text{ educ}_i + \beta_8 \text{ dur}_i + \beta_9 \text{ leftlow}_i + \beta_{10} \text{ leftskill}_i + \beta_{11} \text{ leftop}_i + \beta_{12} \text{ leftserv}_i + \beta_{13} \text{ leftlab}_i + \beta_{14} \text{ lefteduc}_i + \beta_{15} \text{ leftdur}_i + u_i.$$