SOCY498C—Introduction to Computing for Sociologists

Neustadtl

# Regression Post-Estimation: adjust

After you have created a dataset, examined your variables, constructed and estimated a model, the real work begins. Stata has a number of *post-estimation* commands that can be used to assess model assumptions as well as provide additional results to support your analysis. One useful post-estimation command is adjust. In Stata version 11 adjust has been replaced with margins, but adjust still works, for now.

adjust (help adjust)

* After an estimation command, adjust provides adjusted predictions of **xb** (the means in a linear-regression setting), probabilities (available after some estimation commands), or exponentiated linear predictions. The estimate is computed for each level of the by() variables, setting the variables specified in [var[= #]...] to their mean or to the specified number if the = # part is specified. If by() is not specified, adjust produces results as if by() defined one group. Variables used in the estimation command but not included in either the by() variable list or the adjust variable list are left at their current values, observation by observation.
* There are many options including:
  + xb (linear prediction; the default),
  + se (display standard error of the prediction),
  + stdf (display standard error of the forecast),
  + ci (display confidence or prediction intervals), and
  + level(#) (set confidence level).

## Creating the Dataset

Assuming that you have downloaded the GSS subset data file (GSS-98-08.dta) from the course Web page and placed it in a directory called “C:\data” the following program creates a dataset used for these examples. You will probably need to make some changes to reflect how your computer is setup.

|  |
| --- |
| /\* Create subset of the GSS data for this example \*/  #delimit ;  use year  prestg80  educ  age  sex  race  marital using "C:\data\GSS-98-08.dta" if year==2008 & race<3, clear  ;  #delimit cr  /\* Create 0/1 indicator variable \*/  rename sex female  rename race black  drop year |

Now, we can 1) estimate our regression model, 2) create a new variable containing the predicted values for each observation , and 3) examine the predicted values. To estimate this model I used the xi: prefix command to create dummy variables. This is a great Stata shortcut for working with dummy variables and interaction terms in regression models. (see help prefix or help xi).

|  |
| --- |
| Figure 1. |
|  |

The variables in this model collectively explain approximately 28% of the variation in occupational prestige. All of the independent measures are statistically significant (*p*<0.05) except for respondent gender and race.

What does the keep command in Figure 1 do? For the rest of my analysis I only want to analyze observations that were used in my model, i.e. all cases with complete data for all variables. There are many ways to do this but I will use the keep command with e(sample). All estimation commands like regress save something called e(sample) that indicate with 0’s and 1’s which observations were used in the last estimation. A values of 1 means the observation was used in the last estimation (i.e. no missing data) and a value of 0 means the observation was excluded from the model due to missing data (i.e. missing data on at least one variable). This can then be used with almost any Stata command after estimation to restrict that command to the estimation sample (see help postest).

Combining this with keep I can isolate the two classes of observations—those included in the model or excluded from the model. The following Stata commands are equivalent ways to keep (or drop) cases used (or not used) in the estimated model: keep if e(sample) or drop if !e(sample)

We will use adjust to look at the expected values or means of the predicted values so we need to create this variable. There are many ways to create predicted values and I will show you three just so you learn a little more about Stata.

Method one generates by using the parameter estimates from the regression model. Method two uses the same values that are stored automatically by Stata after estimating the model (see help \_variables). The third method, my preferred method, used the predict command. The predict command can also be used to calculate residuals, the error term (see help regress   
postestimation##predict).

## Method 1:

gen yhat=5.707446 + ( 2.458468 \*educ) + ///

( .0995691\*age) + ///

( .3035981\*\_Ifemale\_2) + ///

(-1.443287 \*\_Iblack\_2)

## Method 2:

gen yhat1=\_b[\_cons] + (\_b[educ]\*educ) + ///

(\_b[age]\*age) + ///

(\_b[\_Ifemale\_2]\*\_Ifemale\_2) + ///

(\_b[\_Iblack\_2]\*\_Iblack\_2)

## Method 3:

Predict yhat, xb

|  |  |
| --- | --- |
| After creatingwe can examine summary statistics using normal summary commands and adjust.  In this example you can see that the results are the same! When you use the adjust command without specifying any variables, it simply summarizes the linear predictions, the expected values, of the regression as does adjust. | Figure 2. |
| In this example I use the table command (though tabstat would have produced the same results). This command is very powerful and worth some time looking at the documentation (help table).  Again, the results are the same because without specifying any variables adjust simply summarizes the linear predictions of the regression by *marital*. | Figure 3. |

|  |  |
| --- | --- |
| This example demonstrates how adjust (and table) produce the average predicted values for two discrete measures, in this case the intersection of marital status and gender.  As you might suspect, the results are the same. So, you see that adjust easily provides average expected values for categorical variables. Up to seven variables can be used with the by() option. | Figure 4. |
| This example is silly but shows some of the power of the adjust command. In the first example we see the expected values for the intersection of marital status and gender. The second example shows the same thing except that the race dummy variable, specified as part of the adjust command, not as a by() variable, is set to the mean of race, 0.15025907.  The silliness here is that the race dummy is either 0 or 1 and unless we can somehow justify viewing people as 15% black and 85% white, this doesn’t make sense, though there are analysis situations where examining “mixtures” would be interesting.  In the second example you see that *age*, *educ*, and *\_Ifemale\_2* are left “as is” meaning the value of each observation. The race dummy, *\_Iblack\_2*, on the other hand, is set to the mean.  Furthermore, it is possible to set values that are theoretically interesting. | Figure 5. |

|  |  |
| --- | --- |
| This table shows the predicted values for respondents with average education (13.547496) and age (48.75475) by marital status and gender. | Figure 6. |
| Even crazier, you can evaluate the expected values under situations like if “all of observations are white males.” (*\_ifemale\_2* and *\_Iblack\_2* are both equal to 0) | Figure 7. |
| The adjust command has other useful options including calculating confidence and prediction (forecast) intervals around the predicted values. The ci option and stdf options are used to produce these results.  Here we see the 95% confidence intervals for for different marital categories. | Figure 8. |
| This example shows how you can use adjust to provide confidence intervals around the predicted values for different ages (*Xh*).  The if statement uses the mod() function to calculate values only for people with ages divisible by 5 with no remainder. For example, it includes people who are 25 years old (25/5=5 so no remainder) and excludes people who are 26 years old (26/5 leaves a remainder of 1). Google “modulus” for more details.  Of course you can use the if statement to determine these values for any age (e.g. adjust if age==23). | Figure 9. |

|  |  |
| --- | --- |
| Finally, we can extending this syntax and calculate the prediction or forecast standard errors and intervals for the age groups defined by the mod() function by specifying the stdf option. | Figure 10. |

# Problems

Use the General Social Survey for 1988 and create a dataset with the following variables: *year sexfreq, sex, race, educ, marital, age, childs, reliten,* and *attend*. Recode *sexfreq* and *attend* to reflect yearly numbers and recode *reliten* to fix the problem with the order of the responses. Drop all observations where race is equal to “other”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *sexfreq* | | *reliten* | | *attend* | |
| Original | New | Original | New | Original | New |
| 0  1  2  3  4  5  6 | 0  2  12  36  52  156  208 | 1  2  3  4 | 4  2  3  1 | 0  1  2  3  4  5  6  7  8 | 0.0  0.5  1.0  6.0  12.0  30.0  45.0  52.0  104.0 |

1. Regress the sexual frequency measure on *sex, race, educ, marital status,* and *childs.* The variables sex and race are dummy variables. Code them so that 1= female and 1=black, respectively. The variable marital requires four separate dummy variables since there are five categories (married, widowed, divorced, separated, and never married). Exclude the married people from the regression model. *Nota bene*: you can (maybe should) use the xi: prefix for your regressions to make life easier. (see help xi).
2. Use adjust to calculate the following for cases that were used in the regression model:
   1. What is the average predicted value for the entire sample?
   2. What are the average predicted values for the intersection of religious intensity and sex. Interpret this table.
   3. Hold the variables age and education constant at their means and calculate the average predicted values for the intersection of religious intensity and sex.
   4. Calculate the average predicted value of yearly sexual frequency for the intersection of age and sex for ages between 18 and 89 that end in 0 or 5 (e.g. 20, 25,…,85).