L8. Regression II

**Checking for normality of our residual**

One of the OLS assumptions is that our error term is normally distributed. Since the error term per se is not observable, the closest thing we have is the residual. A residual is the difference between the observed value and the fitted value of our regression. We can test whether it’s normally distributed using the predict command and formal normality tests.

**Example**:

predict r, resid

qnorm r

swilk r

If our error is not normally distributed you might want to think about any potential misspecification errors you may have in your model. In any case, many claim this assumption is robust to violation.

**Influential cases**

One potential problem we might run into is outliers. These are values that are either extremely high or extremely low and have the capacity of introducing bias in our estimates. Often, they can reflect coding errors as well -e.g. missing value as 999999. How do we detect them? At this stage, we have a number of descriptive statistic techniques to identify potential outliers (box plots, scatter plots, summarize, etc.). Another way to look at it is the residual is very high.

Leverage: Outliers can have leverage as well. An observation with high leverage is that with an extreme value on the deviation from the mean of a particular variable. This can have an effect on our estimates.

Influential cases: However, we don’t know how influential these cases are. Hence, influence is a measure of how much these cases are affecting our regression line.

The first thing we can do is look at the studentized residuals to identify potential **outliers**.

Example:

reg eduyrs i.fath\_occ14\_3c male age

predict sresid, rstudent

stem sresid

We should pay attention to studentized residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and even yet more concerned about residuals that exceed +3 or -3. This command will allow us to identify how many outliers we may have, but it won’t tell us what cases they are. We can use the hilo command to identify the cases if we state the residuals against our caseid variable (this is useful if you are analyzing countries or regions). We can also call the residuals against a particular variable.

 

**Leverage**

In order to test for leverage, we can call the predict command once again with the option leverage. Notice we will have to name the variable in a different way, usually referring to the leverage calculation.

Example:

reg eduyrs i.fath\_occ14\_3c male age

predict lev, leverage

sum lev, d

The cut off point for leverage is usually calculated as (2k +2)/ n. Where k is the number of predictors and n is the number of observations. Hence, for us the cut-off point is:

display 8/37181

.00021516

Anything above this cut-off point can be problematic. A way to quickly identify these observations is by using the list command:

list eduyrs fath\_occ14\_3c male age if lev > .00021516

**Influence**

Influence measures combine information on the residuals and the leverage. We can calculate how influential some observations are using Cook’s D. Usually Cook’s D has a cut-off point of 4/n where n is the number of observations. Anything above that threshold is considered to be an influential case.

Example:

reg eduyrs i.fath\_occ14\_3c male age

predict d, cooksd

sum d, d

The cut off point for Cook’s D is usually calculated as 4/ n. Where n is the number of observations. Hence, for us the cut-off point is:

display 4/37,181

.10810811

Anything above this cut-off point can be problematic. A way to quickly identify these observations is by using the list command:

list eduyrs fath\_occ14\_3c male age if lev > .00021516

**DFbetas**

You can also consider more specific measures of influence that assess how each coefficient is changed by deleting the observation. This measure is called DFBETA and is created for each of the predictors. If you run the command, Stata will create a variable for each predictor in the model. This variable captures how much the coefficient would change for that particular variable if we dropped the influential case. The cutoff point here is 2/sqrt(n). We should be concerned if the absolute value of our dfbeta exceeds this cutoff point.



If you have few observations, you can use the hilo command:

hilo idno \_dfbeta\_4 \_dfbeta\_5 \_dfbeta\_6 \_dfbeta\_7, show(20)

If you have too many potential influential cases, you can’t select them one by one. Hence, I suggest you run a regression with the influential cases and a regression without the influential cases as a robustness check. How could we do that? Think about the if command….do the estimates change substantially?

**Weighting your regression estimates:**

What is weighting?

“The adjustment of computations of survey statistics to counteract harmful effects of non- coverage, non-response, or unequal probabilities of selection into the sample.” – Robert Groves

“Weighting refers to a number of techniques that all speak to the question: ‘How much attention should I pay to each case in the dataset?’” – Eric Plutzer

In practice, a survey weight is a value assigned to each case in the data file. In most statistical techniques covered in an introductory course we treat every case the same. This is equivalent to giving every case a weight of 1.00.

The value assigned to each case indicates how much each case will count in a statistical procedure. For example, a weight of 2.00 means that the case counts in the dataset as two identical cases. Weights can and often are fractions, but are always positive and (hopefully) non- zero.

In survey research, weights are designed to increase the representativeness of our sample. In any particular dataset, certain types of people may be under or over represented compared to their proportion in the population. For example, if the percent male in our sample is 40% but is 49% in the population, we assign weights so each male is worth more than one male and each female is worth less than one female to yield the correct population percent. The goal is to allow less biased inferences about the target population.

There are three broad types of weights:

Design weights: adjust for design decisions that result in unequal probabilities of selection. Weights can rebalance the sample if the weights are the inverse of the probability of selection.

Post-stratification weights: The goal of a post-stratification weight is to adjust the data for differential non-response rates on the basis of known characteristics of the population. Post-stratification is only possible when the identical variable is available for both the sample and a census and the variable has identical categories.

**Attrition model weights**: attempt to adjust for the probability of participation vs. dropping out of a longitudinal (follow-up) survey. Individuals may be associated with groups that have higher or lower probabilities of attrition.

Specifying Survey Design Features in Stata

InStata,youcanusethesvyset commandtospecifythestratificationscheme,sampling weights and primary sampling units (used in clustering) for your data. The keywords used here are strata, which specifies the stratification variable, pweight, which specifies the weight variable and psu, which specifies the primary sampling unit.

**Syntax for svyset:**

svyset [pweight=yourpweight], strata(yourstrata) psu(yourpsu)

SVY Commands

The following table lists many of the most commonly used SVY commands for survey data. You can use these commands once you have given Stata the relevant information about the survey design using the commands above. Each of the SVY commands described in this section produces estimates that are corrected for the complex design of the survey you are analyzing.

General/Set-Up

svyset Specify survey design features svydes Describe strata and PSUs

Means, Proportions, Ratios and Totals

 svymean

 svyprop

 svyratio

 svytotal

Cross-Tabulations

svytab

Regression Models

 svyreg

 svyintreg

 svyivreg

 svylogit

 svymlogit

 svyologit

 svyprobit

 svyoprobit

 svypois

Post-Estimation

lincom

test

**Example syntax using svy command**

 svy: mean age sex

 svy: regress grades momeduc female

Another approach is to directly type into Stata the weights we want to use. Stata allows for 4 types of weights (as taken from the Stata manual)

**Fweights:**  Frequency fweights indicate replicated data. The weight tells the command how many observations each observation really represents. Fweights allow data to be stored more parsimoniously. The weighting variable contains positive integers. The result of the command is the same as you duplicated each observation however many times and then ran the command unweighted.

**Pweights**: Sampling pweights indicate the inverse of the probability that this observation was sampled.

**Aweights:** Analytic aweights are typically appropriate when you are dealing with data containing averages. The weighting variable contains the number of persons over which the average was calculated (or a number proportional to that amount).

**Iweights:**  This weight has no formal statistical definition and is a catch-all category. The weight somehow reflects the importance of the observationand any command that supports such weights will define exactly how such weights are treated.

Usually, you would want to apply weights for descriptive statistics and for your analysis. However, there are some scholars that claim that in an analytical setting, getting representational power is not as important as capturing the actual effect of a variable x on y. I suggest you use both, see if the estimates change. If they don’t, mention in your paper you have done robustness checks with weights. If you do, then add a table in the appendix.

As far as Stata goes, we look for the weight variable, and we add it as an element before the option coma. Try to stick to probability weights, as they are directly related to sampling inverse selection probability.

reg eduyrs i.fath\_occ14\_3c male age [pw=dweight]