Intro to Stata 5

1. **Before We Begin This Tutorial...**

   This tutorial is designed to provide students with first-hand experiences of analytical skills which are often used in economics research. Specifically, we are going to apply mean comparison tests and regression analysis to data from Tonga. Our research question in this tutorial is whether household wealth reduces the risk of typical chronic illnesses such as diabetes, heart disease, high blood pressure, and cancer. Note that we talk about causality running from wealth to chronic illnesses, not just correlation between them. As we discussed in class, establishing causality is directly useful to policy formation, but simply showing correlation could be useless for policy application. Good economists consider innovative study designs to show a causal relationship between two variables. Economists sometimes use instrumental variables and fixed-effects regressions to alleviate the problems of reverse causality and omitted variables.

   Unfortunately, many of you have not yet learned what are innovative study designs and what econometric skills are available to alleviate the problems of reverse causality and omitted variables,[1] so this tutorial covers basic analytical skills (mean comparison tests and OLS) which economists often use before applying more advanced skills such as instrumental variables and fixed-effects regressions. Actually, many published papers start their analysis with mean comparison tests and OLS and then proceed to fixed-effects regressions perhaps with instrumental variables. (Let me emphasize again that many of good papers are good because of their innovative study designs. If a study design is appropriate, even OLS produces a quite convincing result for establishing a causal relationship between the two variables in non-experimental settings.)

   Although chronic illnesses were considered “affluent diseases” in the past, quite a few studies show that chronic diseases are recently more prevalent among the less affluent than among the more affluent

[1] You can learn the concepts of fixed-effects regressions and instrumental variables in the courses *Research Methodology* offered in a spring term and *Cross-Sectional and Panel Data Analysis* offered in a fall term. Further, if you are interested in how previous good papers established causality (rather than correlation) and how to apply these econometric skills using Stata, *Health Economics* offered in a winter term would do it.

in developing countries[2] as well as in developed countries.[3] However, the causal relationship from wealth to health is more controversial.[4] Examining a causal link from wealth to health is important from the policy perspective. If wealth causes better health, income-generating policies would help improve health. If not, policy makers need different types of policies to address health concerns of the people. Considering the direct relevance to policy formation, researchers eventually want to examine causality running from wealth to health, but this tutorial covers only basic analytical skills (mean comparison tests and OLS) as a starting point of further research.

2. **The Objectives of Stata Tutorial 5**

   This tutorial covers how to perform two basic skills in economics research using Stata. Specifically, we demonstrate some examples of mean comparison tests and OLS regressions using health data from Tonga, a relatively small island country in the Pacific with the total population approximately 100,000 as of 2003.[5]

3. **Preparing for Tutorial 5**

   Let's create a directory called "tutorial5" under "C:\temp\stata-tutorial\" (under which you should find the directories "tutorial1" to "tutorial4" unless you erased them). We use a cross-sectional data set from Tonga in this tutorial. The survey interviews were fielded in 2003. The data are rich in health information and cover 10,994 individuals in 2,089 households in Tonga (approximately 11% of the total population in 2003). The data file pcexp.dta contains individual-level data such as sex, age, marital status, and educational attainment as well as household-level data, such as per-capital annual household expenditure. Each row in pcexp.dta represents a single individual. The data

[2] For example, see Adams, Alayne M., Timothy G. Evans, Rafi Mohammed, and Jennifer Farnsworth. 1997. "Socioeconomic Stratification by Wealth Ranking: Is It Valid?" *World Development* Vol.25 (7): 1165-1172.

[3] For example, see Hayward, Mark D., Toni P. Miles, and Eileen M. Crimmins, and Yu Yang. 2000. "The Significance of Socioeconomic Status in Explaining the Racial Gap in Chronic Health Conditions" *American Sociological Review* Vol.65 (6): 910-930.

[4] For example, look at Meer, Jonathan, Douglas L. Miller, and Harvey S. Rosen. 2003. "Exploring the health-wealth nexus" *Journal of Health Economics* Vol.22 (5): 713-730.

[5] For county information about Tonga, see, for example, the country profile available in the CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/geos/tn.html).

file chronic.dta contains individual-level data on chronic illnesses, such as history of having ever been diagnosed with a chronic illness. Each row in chronic.dta represents a single individual.

First, let's merge the two data sets. The matching variables should be *s00key* (household id) and *s10key* (household member id). Confirm that the merge ends up with perfect match, yielding 10,994 individuals with *_merge*=3 and no individuals with either *_merge*=1 or *_merge*=2. For your reference, the *do* file tutorial5.do (available in the course folder) lists all commands we use in this tutorial, so tutorial5.do may be a useful guide in the rest of this tutorial.

4. **Mean Comparison Test**

   Our research question in this tutorial is whether wealth reduces chronic illnesses or not. Let's start with examining a simple correlation between wealth and the likelihood of suffering from chronic illnesses. Of course, correlation does not imply causality, but correlation may accompany causality, so it is not a bad idea to start our research by checking the correlation between wealth and chronic illnesses.

   Before performing mean comparison tests (and other statistical analysis), it is very important to check that there are no irrelevant observations in the variables you are going to use. For our case, the variable *s4001* contains answers to the question "Have you ever been diagnosed with diabetes, heart disease, high blood pressure, cancer, or any other chronic illness? Yes=1 and No=2." Also, the variable *pcannualexp* contains per-capita annual household expenditure. Let's *summarize* these two variables to see whether all non-missing observations are relevant. You can confirm that there are no obviously irrelevant observations such as negative per-capita annual household expenditure. Next, you may want to see the distribution of per-capita annual household expenditure. It is known that the distributions of both income and expenditure are typically skewed to the right, meaning that there are a small number of households whose incomes or expenditures are much larger in comparison with other households. To see this, the command *histogram* suffices, which shows that *pcannualexp* is actually skewed to the right. Let's take logarithm of *pcannualexp* to make the distribution more balanced. You can create a new variable *log_pcannualexp* which is just the logarithm of *pcannualexp* where the base of the logarithm is *e*. Use the command *histogram* again to see that *log_pcannualexp* has a more symmetric distribution.

Mean comparison tests statistically examine whether two groups have different means on some characteristic. We will look at whether mean log per-capita annual household expenditure is statistically different between those who have ever been diagnosed with the chronic illnesses and those who have not. The command *ttest* does this. Specifically, you can issue the following command:

```
ttest log_pcannualexp, by (s4001) unequal
```

The option *unequal* means that an equal variance of *log_pcannualexp* is not assumed between the two groups ("Yes" and "No" in *s4001*). We do not have any plausible reason why the variances of *log_pcannualexp* are the same across the two groups, so we use the option *unequal* in our case.

Stata responds to your command, saying that mean *log_pcannualexp* are 7.346176 and 7.170124 for those who have ever been diagnosed with chronic illnesses and for those who have not, respectively.[6] Further, the mean comparison test rejects the null hypothesis ($H_0 : \mu_{yes} - \mu_{no} = 0$) in favor of the alternative hypothesis ($H_A : \mu_{yes} - \mu_{no} \neq 0$) with p-value 0.0000 and in favor of the alternative hypothesis ($H_A : \mu_{yes} - \mu_{no} > 0$) with p-value 0.0000. With these results, we can claim that Tongans with chronic illnesses have *higher* per-capita annual household expenditure than Tongans without chronic illnesses on average and that the difference is statistically significant at the conventional levels of significance. This is not consistent with our hypothesis that wealth reduces the incidence of chronic diseases if the causality creates the correlation. However, we use an observational data set, so we must address both omitted variables and reverse causality before saying something about the hypothesis.

We will refine our analysis. Perhaps, you suspect that age may be a critical omitted variable here. Usually, as people age, income (thus expenditure) increases and the likelihood of suffering from chronic illnesses also increases. Thus, after controlling for age, the correlation between expenditure and chronic illness could become negative. To check this, let's move to regression analysis.

5. OLS Regressions

[6] Without logarithm, these numbers correspond to 1,550 pa'anga (local currency unit) and 1,300 pa'anga, respectively.

First, let's create a new variable *chronic* which equals to zero if the individual has never been diagnosed with chronic illnesses before and equals to one if the individual has ever been diagnosed with chronic illnesses before. That is,

```
gen chronic=0 if s4001==2
replace chronic=1 if s4001==1
```

Next, let's start with a simple OLS regression where *chronic* is regressed on a constant and *log_pcannualexp*. To make sure that there are no irrelevant observations, it is a good idea to *summarize* both *chronic* and *log_pcannualexp*. The Stata command for OLS is *regress*, so the following command runs an OLS where the dependent variable is *chronic* and the independent variables are *log_pcannualexp* and a constant.

```
regress chronic log_pcannualexp
```

Note that Stata automatically includes a constant as one of independent variables unless you state otherwise. Please use *help* for details of the command *regress*.

Consistent with the mean comparison test, the OLS coefficient on *log_pcannualexp* is positive and highly statistically significant. If you literally interpret the OLS coefficient on *log_pcannualexp*, it says that a 10% increase in per-capita annual household expenditure is, on average, associated with approximately a 0.002 or 0.2% point increase in the likelihood of having ever been diagnosed with chronic illnesses, so the impact may be small.

Next, let's control for age in the regression. We could include age and age squared in the regression, but here we control for age by using a series of age-group dummies. The variable *s1006y* represents the age of a sample individual. You need to *summarize s1006y* to check the observations. Create dummy variables as follows: *age_minor* is equal to one if the individual is younger than 20 years old and zero otherwise; *age20*, *age30*, *age40*, *age50*, and *age60* is equal to one if the individual is in her/his 20's, 30's, 40's, 50's, and 60's and zero otherwise, respectively. *age70plus* is equal to one if the individual is in her/his 70's or older and zero otherwise. Then, regress *chronic* on a constant, *log_pcannualexp* and all the age-group dummies except *age20* which is the reference group in the regression.

The regression results are consistent with our expectations: First, as people age, people are more likely to have been diagnosed with chronic illnesses in general.[7] The exception is that people in their 70's or older may be less likely to have been diagnosed with chronic illnesses than people in their 60's, if we just compare the magnitude of the coefficient estimates on *age60* and *age70plus*. This result could be due to selection of healthier individuals more likely to survive in their 70's and older. If this is the case, many unhealthy individuals with chronic illnesses die before they become 70, leading to the observation that among those who have survived until 70 or older, the likelihood of chronic illnesses is smaller in comparison with people in their 60's a part of whom may die due to poor health before they become 70.

Second, with control for age, the coefficient estimate on *log_pcannualexp* is still positive but has become smaller in magnitude (was 0.019 and now 0.003) and statistically insignificant at the conventional levels (p-value 0.295). This is consistent with our conjecture that unobserved age was positively correlated with both *chronic* and *log_pcannualexp* in the previous OLS regression.

Next, we continue to add more control variables. It is plausible that higher educational attainment is positively correlated with higher income thus higher expenditure. Also, education may affect the behavior of people in seeking medical services. If educated individuals are more likely to seek medical services than less educated individuals (keeping the levels of wealth the same across educated and less educated individuals), educated individuals may be more likely to have been diagnosed with chronic illnesses than less educated individuals. This story implies that unobserved educational attainment may bias upward the coefficient estimate on *log_pcannualexp* in the previous regression. On the contrary, educated individuals may have more knowledge of how to prevent chronic illnesses than less educated individuals. If this effect is more prevalent or stronger, omitted education in the previous regression may bias downward the coefficient estimate on *log_pcannualexp*.

We also control for the sex of a sample individual. The likelihood of chronic illnesses could be systematically different across males and females. If that is the case, controlling for sex would reduce the

[7] Just for your review, the interpretation of the coefficient on *age60*, for example, is that in comparison with people in their 20's, people in their 60's are approximately 0.30 or 30% points more likely to have been diagnosed with chronic illnesses.

standard error of regression (thus, the standard errors of the coefficient estimates) even if sex is not correlated with *log_pcannualexp*.

Let's create appropriate variables to regress *chronic* on *log_pcannualexp* while controlling for age, educational attainment, and sex. Create a series of dummies for education attainment. The variable *s1010* represents the educational attainment of the sample individual (Some or full primary=1, Some or full secondary=2, Some or full tertiary=3, No schooling=4, Don't know=5). Create the following dummies for educational attainment: The new dummy *noeduc* is equal to one if the individual has no formal education and zero otherwise. Similarly, the new dummy *primary* is equal to one if the individual has some or full primary education but no secondary education and zero otherwise. Finally, the new dummy *secondaryplus* is equal to one if the individual has at least some secondary education and zero otherwise. Next, create a dummy for male. The variable *s1002* contains the sex of the sample individual (Male=1, Female=2). Create a new dummy *male* which is one if the individual is male and zero otherwise.

Summarize alll the variables we are going to use for the regression to check the observations. Regress *chronic* on *log_pcannualexp*, *age_minor*, *age30*, *age40*, *age50*, *age60*, *age70plus primary*, *secondaryplus*, and *male* where people in their 20's, people with no formal education, and females are the reference groups. The results tell you that the likelihood of having been diagnosed with chronic illnesses does not change across the educational groups. In comparison with those without formal education, people with primary education and people with at least secondary education have no statistically different likelihood of having chronic illnesses. In contrast, the results tell you that females are more likely to have chronic illnesses than males and that the difference is statistically significant at the conventional levels.

In terms of the coefficient estimate on *log_pcannualexp*, there is not much difference in the coefficient estimate and its standard error with and without the additional control variables.

6. **Robust Standard Errors**

If heteroskedasticity exists, the estimates of the standard errors would be biased, possibly leading to the over-rejection of the null hypothesis that the population coefficient is zero ($H_0 : \beta_j = 0$). Remember that heteroskedasticity does not bias the coefficient estimates but does bias the estimated standard errors of the coefficient estimates. Below we will see whether the standard error of the residuals from the last regression varies, depending on the level of per-capita annual household expenditure. If that is the case, heteroskedasticity exists and the estimated standard errors in the previous OLS regression would be biased.

It is easy in Stata to estimate the residuals after running a regression. In our case, we can estimate the residuals $\hat{\epsilon} = chronic - (\hat{\beta_0} + \hat{\beta_1} log\_pcannualexp + \hat{\beta_2} age\_minor + \hat{\beta_3} age30 + \hat{\beta_4} age40 + \hat{\beta_5} age50 + \hat{\beta_6} age60 + \hat{\beta_7} age70plus + \hat{\beta_8} primary + \hat{\beta_9} secondaryplus + \hat{\beta_{10}} male)$ by issuing the following command immediately after the OLS regression.

```
predict epsilon_hat, residuals
```

where *epsilon_hat* is the name of the new variable which contains the residual of each observation. The option *residuals* tells Stata that you need to estimate the residuals from the most recent regression you have run. Please use *help* for details of the command *predict.*

To have a rough idea about whether the standard error of the residuals varies, depending on the level of per-capita annual household expenditure, let's divide sample individuals into three groups (low-expenditure, middle-expenditure, and high expenditure) of approximately the same frequency. To do so, you can use the following command.

```
egen pcannualexp3=cut(pcannualexp), group(3)
```

where the new variable *pcannualexp3* is a category variable indicating either low-expenditure individual (0) or middle-expenditure individual (1) or high-expenditure individual (2). The option *group(3)* in the above command instructs Stata to create three groups of (approximately) equal frequency (*Tabulate pcannualexp3* to confirm that the number of sample individuals in each group is approximately equal across the three groups).

Let's *summarize epsilon_hat* by *pcannualexp3.* That is, issue the following command.

```
bysort pcannualexp3: su epsilon_hat
```

The results show that the standard deviation of the residuals seems to increase as per-capita annual household expenditure increases, which would make you suspect that heteroskedasticity exists in the last OLS regression you have run.[8]

In the presence of heteroskedasticity, many researchers in economics use heteroskedasticity-robust standard errors[9] which is easily available in Stata. Practically, heteroskedasticity-robust standard errors would be, in many cases, the only feasible way to address heteroskdasticity.[10] Heteroskedasticity-robust standard errors are valid in the presence of heteroskedasticity of unknown form including homoskedasticity. Many regression commands in Stata come with the option for heteroskedasticity-robust standard errors. Just use the option *vce(robust)* with the command *regress*. Then, Stata calculates heteroskedasticity-robust standard errors rather than standard errors valid only under the assumption of homoskedasticity. Issue the following command (in one line).

```
regress chronic log_pcannualexp age_minor age30 age40
age50 age60 age70plus primary secondaryplus male, vce(robust)
```

You find no differences in the coefficient estimates with and without the option *vce(robust)*. This should be the case. The option *vce(robust)* affects only the estimates of the standard errors, not the coefficient estimates. You see some differences in the standard errors. Heteroskedasticity-robust standard errors are not necessarily larger than homeskedasticity-assumed standard errors although the former is often larger than the latter in general. With heteroskedasticity-robust standard errors, your statistical tests are valid even in the presence of heteroskedasticity.

7. **Econometric Identification Problems We Have Not Addressed**

As we learn in class, OLS with observational data is typically subject to two econometric identification problems. Our research in

[8] A formal test of heteroskedasticity is available in Stata after running a regression. Use *help* to look up *estat hettest* or *regress postestimation*. However, as usual, econometric tests are prone to many errors. You should not blindly believe econometric tests. A theory or prior knowledge is very important.

[9] Many econometrics textbooks discuss heteroskedasticity-robust standard errors. See, for example, *Introductory Econometrics: A Modern Approach* by Jeffrey Wooldridge 3rd Edition, Section 8.2.

[10] Weighted Least Squares (WLS) is rarely used because researchers rarely know the form of heteroskedasticity. Further, Feasible Generalized Least Squares (FGLS) is also of limited use because researchers rarely know the functional form of heteroskedasticity, thus it is difficult to estimate the form of heteroskedasticity from available data.

this tutorial faces the two problems: reverse causality and omitted variables. Let me discuss one by one.

It is likely that people with chronic illnesses earn less due to health problems in comparison with people without chronic illnesses. According to this story, the causality runs from chronic illnesses to income (thus to expenditure). This reverse-causality problem biases downward the OLS coefficient estimate on *log_pcannualexp*.

There are many omitted variables that are potentially correlated with both per-capita annual household expenditure and chronic illnesses, so our OLS results in this tutorial would be subject to omitted-variable problems. You can tell many possible stories of omitted-variable problems. For example, if richer individuals are more likely to see doctors than poorer individuals, then richer individuals are more likely to have been diagnosed with chronic illnesses than poorer individuals given identical health conditions. In this case, the tendency to see a doctor is the omitted variable which is positively correlated with both per-capita annual household expenditure and the likelihood of having been ever diagnosed with chronic illnesses, giving an upward bias to the OLS coefficient estimate on *log_pcannualexp*.

Without addressing these two problems, we cannot say anything about our research question in this tutorial: does wealth causally reduce chronic illnesses? To better answer this question, students, on one hand, need to study more advanced econometric skills, in particular, fixed-effects regressions and instrumental variables, and on the other hand, need to consider an innovative research design so that you can evaluate available data in a way that resembles an experiment in natural science.

End of Tutorial 5
End of This Series of Tutorials