STATA TUTORIAL

Antonio Pacifico
Federica Romei

March 6, 2012

1. The data set sleep75.dta contains cross sectional data on the time spent sleeping per week and the time spent in paid work in 1975 for a sample of individuals.
   (a) Test the hypothesis that clericals sleeps on average the same minutes per night as non-clerical against the hypothesis that clericals sleep more time using a significance level of 5%:
   $$H_0 : Sleep_{clericals} = Sleep_{non-clericals}$$
   $$H_1 : Sleep_{clericals} > Sleep_{non-clericals}$$
   (b) Using the same dataset test the hypothesis that:
   $$H_0 : sleep_{male} = sleep_{female}$$
   $$H_1 : sleep_{male} \neq sleep_{female}$$

Solution:

The first sets of hypothesis is an one-tailed test, since an extreme value on only one side of the sampling distribution would cause the rejection of the null hypothesis. We can rewrite the problem as

$$H_0 : \Delta sleep = sleep_{clericals} - sleep_{non-clericals} = 0$$
$$H_1 : \Delta sleep = sleep_{clericals} - sleep_{non-clericals} > 0$$

Set the directory, open a log file and name your log file "ex1" or something similar:

```
cd "directory" log using ex1.txt, replace
```

Then upload the data set

```
use sleep75, clear
```

Now we can compute the mean for the minute of sleep for clerical and non-clerical

```
mean sleep if clerical==1

Mean estimation                          Number of obs    =      97

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       sleep |   3311.557   42.50877      3227.177    3395.936
--------------------------------------------------------------



mean sleep if clerical==0

Mean estimation                          Number of obs    =     435

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       sleep |   3247.851   20.77451      3207.019    3288.682
--------------------------------------------------------------
```

Under regolarity condition like

1. Observation I.I.D.
2. $E(sleep_{i,clerical}) < \infty$
3. $E(sleep_{i,non-clerical}) < \infty$
4. $Var(sleep_{i,clerical}) < \infty$
5. $Var(sleep_{i,non-clerical}) < \infty$

we can apply the LL Central Limit Theorem so that

$$\Delta \hat{s}leep = \frac{\bar{s}leep_{i,clerical} - \bar{s}leep_{i,non-clerical}}{\sqrt{SE(sleep_{clerical})^2 + SE(sleep_{non-clerical})^2}} \xrightarrow{D} N(0,1)$$

We can compute easily

$$\Delta \hat{s}leep = \frac{3311{,}55 - 3247{,}851}{\sqrt{42{,}5^2 + 20{,}77^2}} = 1,3467$$

So we cannot reject the null hypothesis

The second set of hypothesis is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause the rejection of the null hypothesis. In the same way

```
mean sleep if male==1

Mean estimation                          Number of obs    =     400

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       sleep |   3252.407   21.75999      3209.629    3295.186
--------------------------------------------------------------

mean sleep if male==0

Mean estimation                          Number of obs    =     306

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       sleep |   3284.588    26.0821      3233.265    3335.912
--------------------------------------------------------------
```

$$\Delta \hat{s}leep = \frac{3252{,}40 - 3284{,}58}{\sqrt{21{,}75^2 + 26{,}08^2}} = -0,94$$

So we cannot reject the null hypothesis.

2. Using the data set sleep75.dta estimate the relationship between variables sleep and totwrk using OLS and comment on the direction of the relationship. Using the R-squared reported for this equation, explain how much of the variation in sleep is actually explained by the totwrk.

Solution:

We run the following regression

$$sleep = \beta_0 + \beta_1 totwrk + u$$

Type the command

`reg sleep totwrk, r`

```
Linear regression                                      Number of obs =     706
                                                       F(  1,   704) =   65.69
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1033
                                                       Root MSE      =  421.14


------------------------------------------------------------------------------
             |               Robust
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1507458   .0185992    -8.10   0.000    -.1872623   -.1142294
       _cons |   3586.377   41.98156    85.43   0.000     3503.953    3668.801
------------------------------------------------------------------------------
```

First column shows the dependent variable at the top (sleep) with the predictor variables below it (totwrk). The last variable (cons) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis.

Second column shows the estimate coefficients. This estimate indicates the amount of increase in sleep that would be predicted by a 1 unit increase in the predictor.

Third column shows the standard error of coefficient. The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a $t - value$. The standard errors can also be used for the confidence interval.

Fourth column shows the $t - statistic$ under the Null Hypothsis that $\hat{\beta}_1 = 0$. Stata uses *Student* t distribution with $n - k$ degree of freedom to compute the $t - statistic$ (k is the number of regressors). As you know when $n - k > 30$ *Student t* distribution converge to a *Normal* distribution. Hence usually we say that the coefficient is significantly different from zero at 5% if $|t| < 1.96$. Remember that if the sample you have is very small you need to compare $|t|$ with other value.

Fifth column shows $p - value$ under the Null Hypothesis that $\hat{\beta}_1 = 0$. $p - value$ is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If you are performing a test at $\alpha\%$ significant level (i mean you have α probability to do Type I Errror), you would reject the Null Hypothesis if $p - value < \alpha$. For example, if $\alpha = 5\%$, you will reject the Null Hypothesis if $p - value < 0.05$. Note: When you reject the Null Hypothesis you can say that the coefficient is significantly different from zero at $\alpha\%$.

Sixth column shows the Confidence Interval at 95%. If zero is not contained in the coefficient interval, you can say that your coefficient is significant at 5% level. Such confidence intervals help you to put the estimate from the coefficient into perspective by seeing how much the value could vary.

Column four, five and six test in different way the same Hypothesis.

The regression shows a negative relationship between sleep and work. The coefficient (parameter estimate) is -0,15. So, for every unitary increase in totwork, a -0,15 decrease in sleep is predicted.

R-Square is the proportion of variance in the dependent variable (sleep) which can be predicted from the independent variable (totwrk). This value indicates that 10% of the variance in sleep can be predicted from the variable totwrk.

3. The data set BWGHT.RAW contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (bwght), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (cigs).

   The following simple regression was estimated using data on $n = 1388$ births:

$$bwght = 119.77 - 0.514cigs$$

   (a) What is the predicted birth weight when cigs = 0?

   (b) What about when we change the regressor from cigarettes smoked per day to packages smoked per day? Comment on the difference.

   (c) What about when weight is measured in kgs?

Solution:

```
use bwght, clear

reg bwght cigs, r


Linear regression                                      Number of obs =    1388
                                                       F(  1,  1386) =   34.29
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0227
                                                       Root MSE      =  20.129


------------------------------------------------------------------------------
             |               Robust
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0877334    -5.86   0.000    -.6858767   -.3416675
       _cons |   119.7719   .5745494   208.46   0.000     118.6448     120.899
------------------------------------------------------------------------------



generate cigspack=cigs/20

.reg bwght cigspack, r

Linear regression                                      Number of obs =    1388
                                                       F(  1,  1386) =   34.29
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0227
                                                       Root MSE      =  20.129


------------------------------------------------------------------------------
             |               Robust
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cigspack |  -10.27544   1.754668    -5.86   0.000    -13.71753    -6.83335
       _cons |   119.7719   .5745494   208.46   0.000     118.6448     120.899
------------------------------------------------------------------------------


generate gram=bwght*28

reg gram cigspack, r
```

```
Linear regression                                      Number of obs =    1388
                                                       F(  1,  1386) =   34.29
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0227
                                                       Root MSE      =   563.6

------------------------------------------------------------------------------
             |               Robust
        gram |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cigspack |  -287.7124   49.13071    -5.86   0.000     -384.091   -191.3338
       _cons |   3353.613   16.08738   208.46   0.000     3322.055    3385.171
------------------------------------------------------------------------------
```

We can also use the bwght example to see what happens when we change the units of measurement of the variables. For example we can change the measure of cigarettes smoked and unit of mass: from unit per day we can measure it in pack (consider 20 cigarettes per single pack) and we could use grams instead of ounces.

We can compute the new coefficients without performing a new regression, i.e.:

$$\beta_1 new = \frac{Cov(cigs/20; bwght*28)}{Var(cigs/20)} = \frac{\frac{(28)}{20}Cov(cigs; bwght)}{\frac{1}{20^2}Var(cigs)} = 28 * 20 * \beta_1 old$$

$$\beta_0 new = \bar{bwght} * 28 - \bar{cigs}/20 \beta_1 new$$

$$\beta_0 new = \bar{bwght} * 28 - \frac{\bar{cigs}}{20} 28 * 20 * \beta_1 old$$

$$\beta_0 new = 28 * \beta_0 old$$

We defined R-squared as a goodness-of-fit measure for OLS regression. We can also ask what happens to R-squared when the unit of measurement of either the independent or the dependent variable changes. Without doing any algebra, we should know the result: the goodness-of-fit of the model should not depend on the units of measurement of our variables. For example, the amount of variation in bwght, explained by the quantity of cigarettes smoket, should not depend on whether ciggarettes are measured in units or in packets. This intuition can be verified mathematically: using the definition of R-squared, it can be shown that R2 is, in fact, invariant to changes in the units of y or x.