R TUTORIAL 2

Siria Angino
Federica Romei

1. The dataset *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 706 individuals.

   (a) Test the hypothesis that clericals sleeps on average the same minutes per night (in a week) as non-clericals against the hypothesis that clericals sleep more time using a significance level of 5%:

$$H_0 : Sleep_{clericals} = Sleep_{non-clericals} \; H_1 : Sleep_{clericals} > Sleep_{non-clericals}$$

   Solution:

   The first set of hypothesis is a 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 R-Script file, then download the 'foreign' library, import the dataset *sleep75.dta* and use the attach command, so that the objects in the dataset can be accessed by simply giving their names:

```
sleep75=read.dta("sleep75.dta")
attach(sleep75)
```

   Now we can compute the mean and the standard error for the minutes of sleep for clericals and non-clericals (remember to divide the standard deviation by the square root of the number of observations):

```
mean(sleep[clerical==1])
[1] 3311.557
sd(sleep[clerical==1])/sqrt(sum(clerical==1))
[1] 42.50877
```

and

```
mean(sleep[clerical==0])
[1] 3247.851
sd(sleep[clerical==0])/sqrt(sum(clerical==0))
[1] 20.77451
```

Under the following regularity conditions :

(a) Observation I.I.D.

(b) $E(sleep_{i,clerical}) < \infty$

(c) $E(sleep_{i,non-clerical}) < \infty$

(d) $Var(sleep_{i,clerical}) < \infty$

(e) $Var(sleep_{i,non-clerical}) < \infty$

we can apply the Central Limit Theorem, so that:

$$\Delta \hat{sleep} = \frac{\overline{sleep}_{clerical} - \overline{sleep}_{non-clerical}}{\sqrt{SE(sleep_{clericals})^2 + SE(sleep_{non-clericals})^2}} \rightarrow N(0,1)$$

We can compute easily

$$\Delta \hat{sleep} = \frac{3311.55 - 3247.851}{\sqrt{42.5^2 + 20.77^2}} = 1,3467$$

Otherwise, with a one-line command:

```
t.test (sleep[clerical==1], sleep[clerical==0])$statistic
t   1.3465
```

Since t$<$1.64 we cannot reject the null hypothesis.

b) Using the same dataset test the hypothesis that:

$$H_0 : sleep_{male} = sleep_{female} \; H_1 : sleep_{male} \neq sleep_{female}$$

Solution:

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[male==1])
1] 3252.407
sd(sleep[male==1])/sqrt(sum(male==1))
[1] 21.75999
mean(sleep[male==0])
[1] 3284.588
sd(sleep[male==0])/sqrt(sum(male==0))
[1] 26.0821
```

Knowing this,

$$\Delta \hat{sleep} = \frac{3252.407 - 3284.588}{\sqrt{21.76^2 + 26{,}082^2}} = -0.9474052$$

Otherwise, with a one-line command:

```
t.test (sleep[male==1], sleep[male==0])$statistic
t  -0.9474052
```

Since |t|<1.96 we cannot reject the null hypothesis.

2. Using the dataset *sleep75.dta*, estimate the relationship between variables *sleep,* minutes slept at night per week, and *totwrk,* minutes worked per week, using OLS and comment on the direction of the relationship. Using the R-squared reported, explain how much of the variation in *sleep* is actually explained by *totwrk.*

Solution:

We run the following regression:

$$sleep_i = \beta_0 + \beta_1 totwork_i + u$$

Type the command:

```
summary(lm(sleep~totwrk))
```

to get (some parts of the output are omitted):

```
Coefficients:               Estimate Std. Error t value Pr(>|t|)
(Intercept)               3586.37695   38.91243  92.165   <2e-16
totwrk                      -0.15075    0.01674  -9.005   <2e-16


Residual standard error: 421.1 on 704 degrees of freedom
Multiple R-squared: 0.1033,Adjusted R-squared: 0.102
F-statistic: 81.09 on 1 and 704 DF,  p-value: < 2.2e-16
```

The first column indicates the names of the independent variable, the predictor *totwork*, and the intercept of the regression line with the y axis, the constant.

The second column shows the estimated coefficients. The coefficient of *totwork* indicates the predicted variation (a decrease in this case) in minutes slept per week for a 1 unit increase in the predictor, i.e. one minute worked more per week.

The third column shows the standard error of the coefficients. Dividing the estimated parameter by the standard error you get the t-value to test whether the estimate is significantly different from 0. The standard error is also used to estimate confidence intervals and to test other hypothesis.

The fourth column shows the t-statistics under the null hypothesis H_0 that $\beta_k = 0$. The distribution used is a t-distribution with $n - k$ degrees of freedom, where k is the number of regressors. When $n - k > 30$, the t-student distribution converges to a normal distribution; hence, we use its critical values and consider the coefficient significantly different from 0 at 5% if $|t| < 1.96$. Remember, however, that if your sample is very small, you need to compare $|t|$ with different critical values.

The fifth column shows the p-values under the null hypothesis that $\beta_k = 0$. The 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 a α significance level (i.e. α probability to make a type I error), you should reject the null hypotesis if $p-value < \alpha$. For example, if α is 0.05, you reject the null hypothesis if $p - value < 0.05$, or, equivalently, the coefficient is significantly different from 0 at a α level.

We can evaluate confidence intervals using the following command (the default confidence level is 95%):

```
confint(lm(sleep~totwrk))
```

and get:

```
                    2.5 %     97.5 %
(Intercept) 3509.9786507 3662.775252
totwrk        -0.1836126   -0.117879
```

If 0 is not contained in such intervals, the coefficient is significant at a 5% level. The confidence interval, the t-test and the p-value are three different ways to perform the same test. They are absolutely equivalent. The numbers obtained in this specific case tell that there is a negative relationship between sleep and work. The estimated β_1 is -0.15, meaning that, for each unitary increase in minutes worked per week, a decrease of 0.15 minutes slept is predicted. Moreover, this coefficient is statistically different from 0 at a 5% level since:

(a) the t-statistic is -9, hence higher, in absolute terms, than 1.96

(b) the p-value is practically 0

(c) the confidence interval does not contain 0.

3. The data set bwght.dta contains data on births for women in the United States. We are interested in the relationship between the infant birth weight in ounces (the dependent variable, *bwght*) and the average number of cigarettes smoked by the mother per day during pregnancy (the independent variable, *cigs*).

   The following regression is estimated using data on 1388 births:

$$bwght_o = 119.77 - 0.51 cigs_i$$
$$R^2 = 0.02272912$$

   a) What is the predicted birth weight if the mother did not smoke?

   Solution:

   If the mother did not smoke, then *cigs*=0; using the results provided in the text of the exercise, the weight of the infant is 119.77 ounces.

   b) What is the birth weight when one packed per day was smoked (one pack = 20 cigarettes)? Comment on the difference.

Solution:

We need to change the unit of measurement of the dependent variable and run a new regression. First, import the dataset and use the attach command, generate the new variable *packs* and run the regression with the latter as independent variable:

```
databwght=read.dta("bwght.dta")
attach(databwght)
packs=cigs/20
summary(lm(bwght~packs))$coef
```

```
            Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 119.77190  0.5723407 209.266802 0.000000e+00
packs       -10.27544  1.8098186  -5.677609 1.661538e-08
```

```
summary(lm(bwght~packs))$r.squared
[1] 0.02272912
```

As you may notice, the new $\beta_{1,new}$ is exactly equal to the former β_1 multiplied by 20. What we are calculating, actually, is the effect of 20 cigarettes more. The intercept stays constant: packs=0 has the same effect on birth as cigs=0.

We could have calculated the coefficients without having to run a new regression. Stated in general term, if $X_{new} = bX$, then

$$\beta_{1,new} = \frac{Cov(X_{new},Y)}{Var(X_{new)}} = \frac{Cov(bX,Y)}{Var(bX)} = \frac{bCov(X,Y)}{b^2 Var(X)} = \frac{1}{b}\beta_1$$

and

$$\beta_{0,new} = Y - \beta_{1,new}X_{new} = Y - \frac{1}{b}\beta_1 bX = \beta_0$$

In this case, since $packs = \frac{1}{20}cigs$, $\beta_{1,new} = 20 * \beta_1$.

c) What about when weight is measured in grams and cigarettes in packs (1 ounce≈0.3 grams)? What happens to the R^2?

Solution:

Again, we generate a new variable, *wgrams*, to express the infant's birth weight in grams, and run the new regression:

```
wgrams=0.03*bwght
summary(lm(wgrams~packs))$coef

              Estimate Std. Error    t value     Pr(>|t|)
Intercept)   3.5931570 0.01717022 209.266802 0.000000e+00
packs       -0.3082633 0.05429456  -5.677609 1.661538e-08

summary(lm(wgrams~packs))$r.squared
[1] 0.02272912
```

Generally, when $Y_{new} = aY$ and $X_{new} = bX$:

$$\beta_{1,new} = \frac{Cov(X_{new},Y_{new})}{Var(X_{new})} = \frac{Cov(bX,aY)}{Var(bX)} = \frac{abCov(X,Y)}{b^2Var(X)} = \frac{a}{b}\beta_1$$

and

$$\beta_{0,new} = Y_{new} - \beta_{1,new}X_{new} = Y_{new} - \frac{a}{b}\beta_1 bX = a\beta_0$$

In this case, $wgrams = 0.03bwght$ and $packs = \frac{1}{20}cigs$, so:

$$\beta_{1,new} = \frac{0.03}{20}\beta_1$$

and

$$\beta_{0,new} = 0.03 * \beta_0$$

What about the R^2? As you may see, it does not change when the unit of measurement - either of the independent either of the dependent variable - changes. Here it is intuitive, for example, that the variation in the infant's weight explained by the quantity of cigarettes smoked should not depend on the whether cigarettes are measured in units or packs. This intuition can be verified mathematically: using the definition of R^2, it can be shown that it is, in fact, invariant to changes in the units of Y or X.