R TUTORIAL 4

Siria Angino
Federica Romei

1. The dataset smoke.dta contains 808 observations for the following variables:

   - *cigs*: cigarettes smoked per day;
   - *educ*: years of schooling;
   - *cigpric*: state cigarette price per pack, in cents;
   - *age*: in years;
   - *income*: annual income in US dollars;
   - *white*: dummy variable, 1 if white, 0 otherwise.

   (a) Regress *cigs* on *white*. Explain the coefficients and their significance.

   Solution:

   Under these assumptions,

   - *cigs* and *white* are i.i.d;
   - $E[u|white] = 0$;
   - fourth moments are well defined (nonzero, finite fourth moments, or "large outliers are unlikely");

   we can write the following linear model:

   $$cigs_i = \beta_0 + \beta_1 white_i + u_i$$

   In this case $\beta_0 = E[cigs|white = 0] \Rightarrow \hat{\beta}_0 = \overline{cigs_{white=0}}$, while $\beta_1 = E[cigs|white = 1] - E[cigs|white = 0] \Rightarrow \hat{\beta}_1 = \overline{cigs_{white=1}} - \overline{cigs_{white=0}}$, where $\overline{cigs_{white=0/1}}$ is the sample average of the smoked cigarettes for non-white and white respectively.

   ```
   summary(lm(cigs~ white))
   ```

```
Coefficients:             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.4082     1.3869   6.063 2.06e-09
white            0.3168     1.4797   0.214    0.831

Residual standard error: 13.73 on 805 degrees of freedom
Multiple R-squared: 5.694e-05,Adjusted R-squared: -0.001185
F-statistic: 0.04584 on 1 and 805 DF,  p-value: 0.8305
```

or, equivalently, exploiting our previous reasoning:

```
mean(cigs[white==1], na.rm=T)
[1] 8.724965
mean(cigs[white==0], na.rm=T)
[1] 8.408163
```

$\beta_0 = 8.41$ means that a non-white person smokes about, on average, 8 cigarettes per day, while white people smoke $\beta = 0.31$ cigarettes more, hence 8.41+0.31=8.72 in total (always on average). β_0 is significantly different from 0 (as you may desume either by the small p-value or the $|t|>1.96$), while β_1 is not. This means that white and not-white people smoke the same number of cigarettes.

(b) Regress *cigs* on *income*. Explain the coefficient and their significance. What happens to β_0 and β_1 if we change the *income* from dollars to pounds and *cigs* from cigarettes to packs, where 1 pack= 20 cigarettes and 1 dollar= 0.63 pounds?

Solution:

As before:

$$cigs_i = \beta_0 + \beta_1 income_i + u_i$$

```
summary(lm(cigs ~ income))

Coefficients:              Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.145e+00  1.128e+00   6.333 3.99e-10
income                   7.987e-05  5.282e-05   1.512    0.131

Residual standard error: 13.71 on 805 degrees of freedom*
Multiple R-squared: 0.002832,Adjusted R-squared: 0.001593
F-statistic: 2.286 on 1 and 805 DF,  p-value: 0.1309
```

(*You can decide to omit NA observations with the command na.omit) The coefficient for *cigs* is very close to zero; the t-statistic smaller than 1.96 and the p-value higher than 0.05 tell that β_1 is not significantly different from 0 at a 5% level. β_0, even though statistically significant, has not a proper meaning in this case. Look at the R^2: it is nearly 0. Income seems to explains cigarettes consumption very little, at least in linear terms. Of course coefficients change when we change unit of measurement of both variables, but their significance and the R^2 stay the same.

```
pack=cigs/20
incpound=income*0.63
summary(lm(pack ~ incpound))

Coefficients:              Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.572e-01  5.641e-02   6.333 3.99e-10
incpound                  6.339e-06  4.192e-06   1.512    0.131

Residual standard error: 0.6855 on 805 degrees of freedom
Multiple R-squared: 0.002832, Adjusted R-squared: 0.001593
F-statistic: 2.286 on 1 and 805 DF,  p-value: 0.1309
```

We could have also exploited results from previous classes. Here $pack = a * cigs = \frac{1}{20} cigs$ and $incpound = b * income = 0.63 income$, so:

$$\beta_{1,new} = \frac{1}{0.63*20}\beta_1 = \frac{1}{0.63*20} * 7.987e^{-05} = 6.338607e^{-06}$$

and

$$\beta_{0,new} = \frac{1}{20}\beta_0 = \frac{1}{20}7.144685 = 0.3572342$$

(c) Create a new variable *cigsan* as *cigs**365, then regress *cigsan* on *cigpric, white, income and age.* Describe the coefficients and their significance.

Solution:

Under the usual assumptions (i.i.d. observations, E[u|cigpric,white,income,age]=0,fourth moments well defined), and after the change in the unit of measurement for the dependent variable, the model becomes:

$$cigsan_i = \beta_0 + \beta_1 cigpric_i + \beta_2 white_i + \beta_3 income_i + \beta_4 age_i + u_i$$

```
cigsan=cigs*365
summary(lm(cigsan ~ cigpric+white+income+age))
```

```
Coefficients:                Estimate Std. Error t value Pr(>|t|)
(Intercept)                3783.95202 2383.69324   1.587    0.113
cigpric                     -13.06148   37.36868  -0.350    0.727
white                        99.02365  541.15088   0.183    0.855
income                        0.02821    0.01937   1.457    0.146
age                         -11.08889   10.39367  -1.067    0.286

Residual standard error: 5010 on 802 degrees of freedom
Multiple R-squared: 0.004496,Adjusted R-squared: -0.0004693
F-statistic: 0.9055 on 4 and 802 DF,  p-value: 0.4601
```

$\hat{\beta}_1 \simeq -13.06$ means that, for each cent of price increase, a person would smoke 13 cigarettes less in one year. It is the unitary increase in the number of cigarettes smoked per year for a 1 cent change in the price of cigarettes, *ceteris paribus*.

$\hat{\beta}_2 \simeq 99$ means that, given two identical individuals (same income and age, and they can buy cigarettes at the same price) except for the fact that one is white and one is not, the white person will smoke 99 cigarettes more per year than the non-white one. Hence $\hat{\beta}_2$ is the difference between the cigarettes smoked in a year by a white individual and those smoked by a non white one, other things being equal.

$\hat{\beta}_3 \simeq 0.03$ means that, given two identical individuals A and B (same "race" and age, and they can buy cigarettes at the same price), if A has one dollar more than B, he/she will smoke 0.02 cigarettes more per year. Stated in more significant terms, if A has 100 dollars more than B, he/she will smoke 2 cigarettes more. $\hat{\beta}_3$ is thus the unitary increase in the number of cigarettes smoked per year for a 1 dollar change in income, *ceteris paribus*.

$\hat{\beta}_4 \simeq -11.1$ means that, given two identical individuals, if one is 1 year older than the other, he/she will smoke 11 cigarettes less. $\hat{\beta}_4$ is thus the change in cigarettes smoked per year due to a unitary change in age, other things being equal.

As you see, none of these coefficients is significant at a 5% level. The R^2, moreover, is very low. This is not a good model to predict cigarettes consumption.

(d) Add to the (c) regression *educ*. Does something change?

Solution:

The new model is:

$$cigsan_i = \beta_0 + \beta_1 cigpric_i + \beta_2 white_i + \beta_3 income_i + \beta_4 age_i + \beta_5 educ_i + u_i$$

```
summary(lm(cigsan ~ cigpric+white+income+age+educ))

Coefficients:               Estimate Std. Error t value Pr(>|t|)
(Intercept)               5268.48196 2470.49065   2.133   0.0333
cigpric                    -11.28301   37.28657  -0.303   0.7623
white                       97.04046  539.83746   0.180   0.8574
income                       0.04305    0.02045   2.105   0.0356
age                        -15.04847   10.52130  -1.430   0.1530
educ                      -137.38210   62.00305  -2.216   0.0270

Residual standard error: 4997 on 801 degrees of freedom
Multiple R-squared: 0.01056,Adjusted R-squared: 0.004384
F-statistic:  1.71 on 5 and 801 DF,  p-value: 0.1298
```

$\beta_5 \simeq -137.38$ means that, given two identical individuals, if one has one year more of education (a 1-year master, for example), he/she will smoke 137 cigarettes less than the other.

The only significant regressors are income and education (in addition to the intercept, which has no meaning in this case). R^2 increases w.r.t. the previous case (even if the number is still low): education seems to be a good explanatory variable for cigarettes consumpion.

2. The dataset *affairs.dta* contains cross section data from a survey conducted by Psychology Today in 1969. The dataset contains 601 observations on 9 variables:

   - *affairs*: how often engaged in extramarital sexual intercourse during past years;
   - *gender*: dummy variable, 1 if male, 0 if female;
   - *age*: coding years:
     - 17.5 if under 20;
     - 22 if 20-24
     - 27 if 25-29
     - 32 if 30-34;
     - 37 if 35-39;
     - 42 if 40-44;
     - 47 if 45-49;
     - 52 if 50-54;
     - 57 if over 55.

- *yearmarried*: coding number of years married:
  - 0.125 if 3 months or less;
  - 0.417 if 4-6 months;
  - 0.75 if 6 months-1year;
  - 1.5 if 1-2 years;
  - 4 if 3-5 years;
  - 7 if 6-8 years;
  - 10 if 9-11 years;
  - 15 if 12 or more.
- *children*: children: dummy variable, 1 if the married couple has children, 0 otherwise;
- *religiousness*: variable coding religiousness:
  - 1 if anti;
  - 2 if not at all;
  - 3 if slightly;
  - 4 if somewhat;
  - 5 if very.
- *education*: coding years of education as a proxy for level:
  - 9 if grade school;
  - 12 if high school
  - 14 if some college;
  - 16 if college graduate;
  - 17 if some graduate school
  - 18 if master degree
  - 20 PhD, MD or other advanced degree.
- *occupation*: coding kind of occupation according to the Hollingsead classification;
- *rating*: coding self-rating of marriage:
  - 1 if very unhappy;
  - 2 if somewhat unhappy;
  - 3 if average;
  - 4 if happier than the average;
  - 5 if very happy.

(a) Run a multiple regression for *affairs* on what you want. Explain to the class your results.

(b) Whatever regressors you use, do you think that there is a causal effect between *affairs* and the first considered regressor?

Solution:

These questions have not a proper answer. It depends on the selected regressor.