MOEC0021
Empirical Methods
University of Zurich HS 2020
PROF: Emanuele Dicarlo / TA: A. Jenni, M.R. Greco, S. Bagagli, L. Schubert

---

Problem Set 2 - R-squared and Hypothesis Testing

This problem set is due on the **9th of November** at **23:59**. Solutions should be turned in on OLAT in a single PDF file. Include in your file any code you wrote to answer the questions. Please name your file as GroupName_PS2.pdf.
One of your goals is to communicate efficiently. Please keep your answers succinct; lengthy answers will be marked down.

---

1. Theory – Playing around with R^2

Consider the model

$$y = \beta X + \epsilon, \qquad \epsilon \sim i.i.d(0, \sigma^2 I)$$

Where X includes a constant term.

(a) Show that $TSS = ESS + RSS$. Also, show you can write $R^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}}$ where $\tilde{y}_i = y_i - \bar{y}$.

(b) Show that $R^2 = corr^2(\mathbf{y}, \hat{y})$. What is the intuition behind it?

(c) Suppose you decided to measure all of your X variables in different units such that your new X variable, call it $\tilde{X}$, is exactly double your old one, i.e. $\tilde{X} = 2X$. Suppose you run the regression of y on $\tilde{X}$; call the resulting estimate $\tilde{\beta}$. You showed in Exercise Set 1 that $\tilde{\beta} = \frac{1}{2}\hat{\beta}$. Is the R^2 different in the two models? Provide an intuitive answer.

(d) Intuitively discuss the fact that including another regressor in the linear model always decreases the RSS.

(e) Provide a formal proof of point (d).

(f) Can you suggest a problem of interpreting the R^2 as a measure of how "good" the model is? If you think the model might not be "good", why might it nevertheless have a high R^2?

2. Empirical Question - Hypothesis Testing

We will now try to investigate the role of class size on educational outcomes using a sample of schools. When asked about their views on class size in surveys, parents and teachers generally report that they prefer smaller classes. This may be because those involved with teaching believe that smaller classes promote student learning, or simply because smaller classes offer a more pleasant environment for the pupils and teachers who are in them. Class size is often thought to be easier to manipulate than other school inputs, and it is a variable at the heart of policy debates on school quality and the allocation of school resources in many countries. In this exercise we will try to learn more about the role of class size, while testing you on your econometric toolkit. For this question, assume that Assumption 2 (Mean-zero Error) holds so that you can make causal statements in your answers.[1] Download the dataset *class_size.dta* from OLAT and import it into Stata or R. The dataset includes 4th grade (primary school) students from several schools of different sizes and located in 6 geographical areas. For each school you have the following information (variable names in *italic*):

- region code: *regioncode*
- school identifier: *schlcode*
- school size (i.e. total number of pupils): *sc_size*
- number of boys in school: *sc_boys*
- number of girls in school: *sc_girls*
- number of classes in school: *n_classes*
- class size: *classize*
- class identifier: *class_id*
- average marks in grammar tests (points): *mrkgrm*
- percentage of disadvantaged kids in class: *pct_dis*

(a) Generate a new dummy named *big_school* which is equal to 1 if *n_classes*> 2.

i. First regress grammar scores on class size. Interpret the coefficient on class size.

ii. Regress grammar scores on class size and the *big_school* dummy. How does the coefficient on class size change with respect to (a).i? How do you interpret it in the second specification?

[1] Note that this is a very strong assumption that is unlikely to hold, but we want to focus on other aspects of econometrics for the moment.

(b) Drop the *big_ school* dummy.

i. Now regress grammar scores on class size and *pct_ dis*. Also, generate natural log of grammar scores (*ln_ mrkgrm*) and regress it on class size and *pct_ dis*. What are their coefficients? How do you interpret them? How do they compare? (Note: Make sure to compare approximately equivalent objects from each specification.)

ii. In the regression of grammar scores on class size and *pct_ dis* how do you interpret the coefficient on *pct_ dis*?

(c) Generate now a dummy *small_ size* equal to 1 if $classize <= 10$. Now regress grammar scores on *small_ size* and *pct_ dis*. You estimate thus the following model:

$$mrkgrm_i = \beta_1 + \beta_2 small\_size_i + \beta_3 pct\_dis_i + \varepsilon_i$$

i. How do you interpret the coefficient on *small_ size*, regardless of whether it is statistically significant? Is it *economically* significant in your opinion? Test both in R/Stata and "by hand" the hypothesis that $\beta_2 = 0$. Should you use a one-sided or two-sided test? Do the one you think most appropriate.

ii. Use R/Stata to get $\hat{\beta}_2$ (the coefficient on *small_ size*) using partitioned regression as we did in lecture.

iii. Use R/Stata to show that $\hat{\beta}_1 = \bar{y} - \bar{X}_{-1}\hat{\beta}_{-1}$.

iv. Regress *ln_ mrkgrm* on *small_ size* and *pct_ dis*. What is the interpretation of the coefficient on *small_ size*, regardless of whether it is statistically significant?

(d) Generate a dummy *many_ dis* equal to 1 if the percentage of disadvantaged pupils is larger than 10%. Estimate the following model:

$$mrkgrm_i = \beta_1 + \beta_2 classize_i + \beta_3 many\_dis_i + \beta_4 many\_dis_i * class\_size + \varepsilon_i$$

i. Test in R/Stata the individual hypotheses that $\beta_3 = 0$ and $\beta_4 = 0$. Test in R/Stata the joint hypothesis that they are both zero. Compute the F-statistic with R/Stata using the formula from slide 103, *top01d*. What do you conclude?

ii. What is the effect of having 10 additional students in a class with less than 10% disadvantaged pupils?

(e) Run the model in a.i separately for classes with a high percentage of disadvantaged pupils (many_dis= 1) and those with a low one. How do coefficients compare to those found in (d)? Show all three estimation results in one table and explain.

(f) Generate a dummy for each region. Can you include all of them in your model? Why or why not?

(g) Now regress grammar scores on *classize* and *pct_dis* for each regional subsample (i.e. perform the regression for each region). Comment on the pattern of your class size estimates across regions. Is the effect of class size *statistically* different across regions? Provide support to your conclusions (*hint*: propose and estimate an appropriate model).

(h) Consider the subsample of schools with *only one* class.

i. Regress grammar scores on *classize*, *sc_boys*. How do you interpret the coefficient on *sc_boys*? And that on *classize*?

ii. Regress grammar scores on *sc_girls*, *sc_boys*. How do you interpret the coefficient on *sc_boys*?

iii. From the estimation in (h).ii can you say anything about the effect of increasing the class size by one person? Would this effect be different if the additional pupil is female or male? Provide formal arguments to your conclusions.

(i) Throughout this question we have assumed that Assumption 2 holds. What do you think about this assumption? Can you think about other factors we did not take into consideration in our model that could bias the conclusion that we are measuring the true effect of class size on students' performance?