

Exercise Set 1: Multiple Linear Regression and Inference

Hand-in date: February 14th

You have to submit by email (TO: `ricmora@eco.uc3m.es`, SUBJECT: MEI Set 1) a STATA do file which contains both the STATA code to answer the questions and your answers (as comments in the do file). The name of the do file should be

SET1yourname.do

(ex: SET1RicardoMora.do)

1. We use Wooldridge's dataset ATTEND.DTA to study the relation between class assistance and final exam results at the university. Assistance is measured as the proportion of classes attended by the student (*atndrte*) and exam results are standardized (*stndfnl*).

- (a) (20 points) Download the data set. Obtain descriptive statistics for *atndrte* and *stndfnl*. Are these two variables correlated?
- (b) (20 points) Consider the relation between the rate of class assistance and final marks,

$$stndfnl = \delta_0 + \delta_1 atndrte + \varepsilon,$$

where δ_0 and δ_1 are unknown parameters and ε includes all variation in *stndfnl* which cannot be captured by $\delta_1 atndrte$. Estimate the model by OLS. Is *atndrte* statistically significant?

- (c) (20 points) Generate $\widehat{stndfnl}_i = \widehat{\delta}_0 + \widehat{\delta}_1 atndrte_i$, $i = 1, \dots, n$, where $\widehat{\delta}_0$ and $\widehat{\delta}_1$ are the OLS estimators for δ_0 and δ_1 . Also, compute residuals $\widehat{\varepsilon}_i = stndfnl_i - \widehat{stndfnl}_i$, $i = 1, \dots, n$.
- (d) (20 points) Show that

$$\sum_{i=1}^n \widehat{\varepsilon}_i = \sum_{i=1}^n \widehat{\varepsilon}_i \cdot atndrte_i = \sum_{i=1}^n \widehat{\varepsilon}_i \cdot \widehat{stndfnl}_i$$

- (e) (20 points) Compute the sample variances $\widehat{Var}(stndfnl)$, $\widehat{Var}(\widehat{stndfnl})$, $\widehat{Var}(\widehat{\varepsilon})$ and check that

$$\widehat{Var}(stndfnl) = \widehat{Var}(\widehat{stndfnl}) + \widehat{Var}(\widehat{\varepsilon}),$$

and that

$$R^2 = \frac{\widehat{Var}(\widehat{stndfnl})}{\widehat{Var}(stndfnl)} = 1 - \frac{\widehat{Var}(\widehat{\varepsilon})}{\widehat{Var}(stndfnl)}.$$

2. Consider now the relation between the final grades for college students and class assistance AND homework. Homework is the rate of homework handed in by the student (*hwrte*). Now the model is

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 hwrte + v \tag{1}$$

where v includes all variation in *stndfnl* which cannot be captured neither by $\beta_1 atndrte$ nor by $\beta_2 hwrte$.

- (a) (20 points) Under what sufficient conditions $\beta_1 = \delta_1$?
- (b) (20 points) Obtain OLS estimates. Is *hwrte* significant?
- (c) (20 points) Compute $\widehat{Cov}(stndfnl, hwrte)$ and show that:

$$\hat{\delta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{\hat{C}(atndrte, hwrte)}{\hat{V}(atndrte)}.$$

Comment the results.

- (d) (20 points) Obtain OLS estimates in the model, $atndrte = \gamma_0 + \gamma_1 hwrte + u$, store the OLS residuals, \hat{u}_i . In view of your answer to question (a), is there any evidence that any of the sufficient conditions for $\beta_1 = \delta_1$ are satisfied?
- (e) (20 points) Now estimate by OLS the model

$$stndfnl = \alpha_0 + \alpha_1 \hat{u} + w$$

Is $\hat{\alpha}_1 = \hat{\beta}_1$? If so, why?

3. Use HPRICE1 data from Wooldridge to estimate the model

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u \quad (2)$$

where

<i>price</i> =	house price
<i>assess</i> =	assessed housing value before sale
<i>lotsize</i> =	size of the lot
<i>sqrft</i> =	square footage
<i>bdrms</i> =	number of bedrooms

We want to test whether the assessed housing price is a rational valuation. If this is the case, then a 1% change in *assess* should be associated with a 1% change in *price*; that is $\beta_1 = 1$. In addition, *lotsize*, *sqrft* and *bdrms* should not help to explain $\log(price)$, once the assessed value has been controlled for.

Testing this hypothesis can be stated as

$$H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0.$$

There are four restrictions to be tested; three are exclusion restrictions, but $\beta_1 = 1$ is not.

- (a) (20 points) Compute the qF -statistic:

$$qF = q \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \frac{n}{q}$$

and compare it with the critical value of the distribution χ_q^2 at the 5%.

- (b) (20 points) Compute the F -statistic:

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \frac{n - k - 1}{q}$$

and compare it with the critical value of the distribution $F_{n-k-1, q}$ at the 5%.

- (c) (20 points) Can we use the R_r^2 and the R_{ur}^2 to compute F (as shown in Slides #2, pp.38)? Explain. (Hint: Use the R^2 's to compute the small sample F statistic and check that the result you obtain does not coincide with the one you gave in (b) or the one STATA gives. Then look at the comments in the slides for an explanation.)

- (d) (20 points) In the model

$$\log(\text{price}) = \delta_0 + \delta_1 \log(\text{lotsize}) + \delta_2 \log(\text{sqrft}) + \delta_3 \text{bdrms} + v \quad (3)$$

test with the F statistic that an extra room together with a reduction by 5% of the footage of the house (without changing *lotsize*) is not going to change the price of the house (Hint: Obtain the linear relation between δ_3 and δ_2 under the null and replace one of the two parameters in the unrestricted model. Then estimate the restricted model with the new controls).

- (e) (20 points) Test the same null using the t statistic and check that $t^2 = F$.