
Problem set

Please submit your answers and Stata code for computer exercises by 1pm on Monday **1 November**. If you study in person, turn in your answer at the tutorial. If you study online, use Moodle assignment tab to upload your answers.

This problem set is expected to help you understand material covered during the lectures. You are encouraged to discuss the problems and work them out together with you colleagues. It is essential though that in the end you write your own solutions, mainly because the only way to learn how to do proofs is to go through all steps on your own.

1 Analytical exercises

Exercise 1 (Linear transformation of the regressors)

Consider the OLS regression of the $n \times 1$ vector \mathbf{y} on the $n \times k$ matrix \mathbf{X} . Suppose that the matrix of regressors \mathbf{X} was transformed to create an alternative set of regressors $\mathbf{Z} = \mathbf{XC}$, where \mathbf{C} is a $k \times k$ non-singular matrix. Thus, each column of \mathbf{Z} is a mixture of some of the columns of \mathbf{X} .

1. Explain the intuition behind this transformation. Propose an example where it can be useful (you may focus on the case $k = 1$).
2. Compare the OLS estimator from the regression of \mathbf{y} on \mathbf{X} to the OLS estimates from the regression of \mathbf{y} on \mathbf{Z} . In the next two questions refer to these estimators as $\hat{\beta}$ and $\tilde{\beta}$, respectively.
3. Compare the OLS residuals for estimators $\hat{\beta}$ and $\tilde{\beta}$.
4. Compare the variance-covariance matrices of the OLS estimators $\hat{\beta}$ and $\tilde{\beta}$. What can you say about the standard error of these two estimators?

Exercise 2 (Method of moments estimator)

Let X_1, \dots, X_n be a random sample from a uniform distribution on the interval $[\theta - 1, \theta + 1]$.

1. Find a method of moments estimator for the parameter θ . Refer to this estimator as $\hat{\theta}_{MM}$ below.
2. Is your estimator $\hat{\theta}_{MM}$ unbiased?
3. Find the variance of the estimator $\hat{\theta}_{MM}$.

Exercise 3 (Model specification)

Consider two regression models:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (1)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e} \quad (2)$$

where \mathbf{y} and \mathbf{e} are $n \times 1$, \mathbf{X} is $n \times k_1$, \mathbf{Z} is $n \times k_2$, $\boldsymbol{\beta}$ is $k_1 \times 1$, and $\boldsymbol{\gamma}$ is $k_2 \times 1$. Assume that $\mathbb{E}(\mathbf{e}) = 0$ and $\mathbb{E}(\mathbf{e}\mathbf{e}') = \sigma^2 \mathbf{I}_n$.

1. Show that in the second model

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{y} \text{ and } \hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{y},$$

where \mathbf{M}_X and \mathbf{M}_Z are annihilator matrices from projection of \mathbf{y} on \mathbf{X} and \mathbf{Z} respectively. [Hint: instead of following the textbook that derives these expressions from scratch, pre-multiply the “long” equation by $\mathbf{X}'\mathbf{M}_Z$ and use the properties of \mathbf{M} to simplify and derive $\hat{\boldsymbol{\beta}}$. Then apply similar approach to get $\hat{\boldsymbol{\gamma}}$.]

2. Under the assumption that (2) is correctly specified, what are the properties of the OLS estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$?
3. Under the assumption that (1) is the correct model, what are the properties of the OLS estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ if (2) is estimated? [Irrelevant variables]
4. Under the assumption that (2) is the correct model, what are the properties of the OLS estimator $\hat{\boldsymbol{\beta}}$ if (1) is estimated? [Omitted variables]

Exercise 4 (The properties of statistical estimators)

Consider drawing independently two random samples from a population distributed as $N(\mu, \sigma^2)$. The first sample has n_1 observations, and its sample mean is given by $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i}$. The second sample has n_2 observations, and its sample mean is $\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i}$. Consider two estimators of the population mean μ :

$$\hat{\mu}_1 = \frac{1}{2}(\bar{x}_1 + \bar{x}_2)$$
$$\hat{\mu}_2 = (n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2).$$

Compare the properties of these estimators (unbiasedness, efficiency, consistency).

Exercise 5 (Estimation of the error variance)

Show that the estimators

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2$$

and

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{e}_i^2$$

are both consistent for σ^2 .

Exercise 6 (Hypotheses tests)

You have two independent samples $(\mathbf{y}_1, \mathbf{X}_1)$ and $(\mathbf{y}_2, \mathbf{X}_2)$. Both samples have the same number of observations n , and both \mathbf{X}_1 and \mathbf{X}_2 have k columns. The samples satisfy linear projection models $\mathbf{y}_1 = \mathbf{X}_1\beta_1 + \mathbf{e}_1$ and $\mathbf{y}_2 = \mathbf{X}_2\beta_2 + \mathbf{e}_2$, where $\mathbb{E}(\mathbf{x}_{1i}e_{1i}) = 0$ and $\mathbb{E}(\mathbf{x}_{2i}e_{2i}) = 0$. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the OLS estimates of β_1 and β_2 .

1. Find the asymptotic distribution of $\sqrt{n} \left((\hat{\beta}_2 - \hat{\beta}_1) - (\beta_2 - \beta_1) \right)$ as $n \rightarrow \infty$.
2. Find an appropriate test statistic for $H_0 : \beta_2 = \beta_1$.
3. Find the asymptotic distribution of this statistic under H_0 .

2 Computer exercises

In this exercise you will compute the least squares estimates of a linear regression model in Stata. You will also have to interpret your results. Please write your code in do-file and submit it along with your answers to the questions. The dataset for estimation is uploaded on Moodle - cps09marV12.dta.

We are going to work with the following regression model:

$$\log(\text{Wage}) = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Experience} + \beta_3 \cdot 0.01 \cdot \text{Experience}^2 + e, \quad (3)$$

where $\text{Wage} = \text{Earnings}/(\text{Hours} \cdot \text{Weeks})$ is hourly wage rate in current US dollars, Education is the number of accomplished years of schooling, and Experience is approximation of labor market experience computed as $\text{Experience} = \text{Age} - \text{Education} - 6$. Check sections 3.22 and 3.23 in Hansen if you need more detailed data description.

1. Compute descriptive statistics for the dataset and discuss your results. You may use any analytical or graphing commands that in your opinion would give the best description of the data.
2. Use command `regress` to estimate equation (3) by least squares for males only. How can you interpret estimated relationship between wages and education? What about wages and experience? Explain the role of the multiplier 0.01 at the third term: how would the estimates differ if you dropped it?
3. Add a gender dummy to the equation (3) and estimate it by least squares using the entire sample. Compare your results to those you obtained for the previous question, and try to explain any differences that you find. Explain why this model might not be the best way to estimate female wage equation.
4. Numerically calculate the following quantities (\hat{e}_i are OLS residuals and \hat{y}_i are the fitted values from the regression equation (1)):

(a) $\sum_{i=1}^n \hat{e}_i$

(b) $\sum_{i=1}^n \hat{y}_i \hat{e}_i$

Are these calculations consistent with any of the OLS properties? Explain.

5. Estimate wage equation with only one explanatory variable - education. Compute residual variance and R-squared for this model. Explain your results. Now include education and experience. How did residual variance and R-squared change with inclusion of a new regressor? Is it consistent with the theory? What happens if you add the square of experience?
6. Recall the structure of the omitted variable bias (Hansen, section 2.24). Use it to explain why dropping the square of experience in the previous question affected the coefficient on experience more than the coefficient on education. Use Stata to compute any relevant statistics required to support your statements.
7. Compute homoskedastic covariance matrix of the least squares estimator for Equation (3) and report standard errors (make sure to include covariance terms).
8. Try to implement as many heteroskedasticity robust estimators for covariance matrix of the least squares estimator given by equations 4.37-4.40 in the text as possible, and report standard errors (look at `robust` and `vce` commands). How do they compare, is the magnitude of the standard errors obtained under different methods consistent with your expectations?