

Exercises for the course of Econometrics

Introduction

- 1.1 A researcher is using data for a sample of 30 observations to investigate the relationship between some dependent variable y_i and independent variable x_i . Preliminary analysis of the sample data produces the following information:

$$\begin{aligned} \sum_{i=1}^N y_i &= 7 & \sum_{i=1}^N x_i &= 1 & \sum_{i=1}^N y_i^2 &= 55 & \sum_{i=1}^N x_i^2 &= 3 \\ \sum_{i=1}^N x_i y_i &= 11, \end{aligned}$$

with $N = 30$. Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

- (a) Compute the OLS estimates of the intercept parameter β_1 and the slope parameter β_2 .
 - (b) Compute the value of the R-squared, the coefficient of determination for the estimated OLS sample regression equation. Briefly interpret what the calculated value of this R-squared means.
 - (c) Calculate the estimation of the variance and the standard error of the estimators of β_1 and β_2 .
- 1.2 Let *children* denote the number of children ever born to a woman, and let *education* denote years of education for the mother. A simple model relating fertility to years of education is

$$\text{children} = \beta_1 + \beta_2 \cdot \text{education} + u,$$

where u is the unobserved error term.

- (a) What kinds of factors are contained in u ? Are these likely to be correlated with level of education?
- (b) In your opinion, will a simple regression analysis uncover the causal effect of education on fertility? Explain.

1.3 Consider the simple regression model. Does a change in the unit of measurement for the dependent variable modify the parameter estimates or the R-squared? And a change in the unit of measurement for the regressors? Explain and justify.

1.4 In the wage equation,

$$\log(wage) = \beta_1 + \beta_2 \cdot education + u,$$

if $wage$ is measured in cents rather than in dollars, what difference does it make to the equation? Justify.

1.5 Suppose that the constant is equal to zero, i.e., $\beta_1 = 0$. Then compute the OLS estimator of the slope coefficient β_2 and its variance.

1.6 In the linear consumption function

$$consumption = \beta_1 + \beta_2 \cdot income + u$$

the marginal propensity to consume (MPC) out of income is simply the slope β_2 , while the average propensity to consume (APC) is $consumption/income$. Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained:

$$\widehat{consumption} = -124.84 + 0.853 \cdot income$$

- (a) Interpret the intercept in this equation and comment on its sign and magnitude.
- (b) What is predicted consumption when family income is \$30,000?

(c) With *income* on the x -axis, draw a graph of the estimated MPC and APC.

1.7 Using data from 1988 for houses sold in Andover, MA, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*distance*):

$$\widehat{\log(\text{price})} = 9.40 + 0.312 \cdot \log(\text{distance})$$

$$N = 135, \quad R^2 = 0.162$$

- (a) Interpret the coefficient on $\log(\text{dist})$. Is the sign of this estimate what you expect it to be?
- (b) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of price with respect to distance?
- (c) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

1.8 A researcher considers two regression specifications to estimate the alternative relationships between a variable x and a variable y and, using a sample of N observations, estimates the following regression equations:

$$\widehat{\log y} = \hat{\alpha}_1 + \hat{\alpha}_2 \log x \tag{1}$$

$$\widehat{\log(y/x)} = \hat{\beta}_1 + \hat{\beta}_2 \log x \tag{2}$$

- (a) Write $\hat{\beta}_2$ in terms of $\hat{\alpha}_2$.
- (b) Write $\hat{\beta}_1$ in terms of $\hat{\alpha}_1$. What is the intuition?
- (c) Demonstrate that

$$\widehat{\log y_i} - \log x_i = \widehat{\log(y_i/x_i)}.$$

- (d) Demonstrate that the residuals of regression (1) are the same as those of regression (2).

Chapter 1

- 2.1 Prove that $\mathbf{X}'\mathbf{X}$ is positive definite if $\mathbf{X}$ is of full column rank.
- 2.2 Define the projection matrix as $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the annihilator matrix as $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Show the following.
- (a) $\mathbf{P}$ and $\mathbf{M}$ are symmetric and idempotent.
 - (b) Idempotent matrix (different from $\mathbf{I}$) are singular.
 - (c) $\mathbf{P}\mathbf{X} = \mathbf{X}$, $\mathbf{M}\mathbf{X} = \mathbf{0}$.
 - (d) $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$.
 - (e) $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}\mathbf{u}$.

Note also that the trace of an idempotent matrix — the sum of the elements on its main diagonal — equals the rank of the matrix and thus is always an integer.

- 2.3 Prove that the R-squared is the square of the simple correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- 2.4 Prove that if a regression is fitted without a constant term, the residuals will not necessarily sum to zero, and the R-squared if calculated as $1 - \sum_{i=1}^N \hat{u}_i^2 / (\sum_{i=1}^N y_i^2 - N\bar{y}^2)$, may be negative.
- 2.5 Define uncentered sum of squares as:

$$\text{USST} = \sum_{i=1}^N y_i^2, \quad \text{USSE} = \sum_{i=1}^N \hat{y}_i^2, \quad \text{USSR} = \sum_{i=1}^N \hat{u}_i^2$$

where "U" stands for "uncentered". Then prove:

$$\text{USST} = \text{USSE} + \text{USSR}.$$

Show that this identity holds even if there is no constant in the regression. Then define the uncentered R-squared.

2.6 The variable *rdintens* is expenditures on research and development (R&D) as a percentage of sales. The variable *sales* is total sales measured in millions of dollars. The variable *profmarg* is profits as a percentage of sales. Using a data set for 32 firms in the chemical industry, the following equation is estimated:

$$\widehat{rdintens} = 0.472 + 0.321 \cdot \log(sales) + 0.050 \cdot profmarg$$

$$(1.369) \quad (0.216) \quad (0.046)$$

$$N = 32, \quad R^2 = 0.099.$$

- (a) Interpret the coefficient on $\log(sales)$. In particular, if *sales* increases by 10%, what is the estimated percentage point change in *rdintens*? Is this an economically large effect?
- (b) Test the hypothesis that R&D intensity does not change with sales, against the alternative that it does change with sales. Do the test at the 5% and 10% levels.
- (c) Does *profmarg* have a statistically significant effect on *rdintens*?

2.7 Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let *return* be the total return from holding a firm's stock over the four-year period from the end of 1990 to the end of 1994. The efficient markets hypothesis says that these returns should not be systematically related to information known in 1990. If firm characteristics known at the beginning of the period help to predict stock returns, then we could use this information in choosing stocks. For 1990, let *dkr* be a firm's debt to capital ratio, let *eps* denote the earnings per share, let $(\log)netinc$ denote net income, and let $(\log)salary$ denote total compensation for the CEO.

(a) The following equation was estimated:

$$\begin{aligned}\widehat{return} = & 40.44 + 0.952 \cdot dkr + 0.472 \cdot eps \\ & (29.30) \quad (0.854) \quad (.332) \\ & - 0.025 \cdot netinc - 0.003 \cdot salary \\ & (0.020) \quad (0.009)\end{aligned}$$

$$N = 142, \quad R^2 = 0.0285.$$

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant? [Hints: The restricted model is not estimated. How can you proceed?]

(b) Now reestimate the model using the log form for *netinc* and *salary*:

$$\begin{aligned}\widehat{return} = & 69.12 + 1.056 \cdot dkr + 0.586 \cdot eps \\ & (164.66) \quad (0.847) \quad (.336) \\ & - 31.18 \cdot \log(netinc) + 39.26 \cdot \log(salary) \\ & (14.16) \quad (26.40)\end{aligned}$$

$$N = 142, \quad R^2 = 0.0531.$$

Do any of your conclusions from part (a) change?

(c) Overall, is the evidence for predictability of stock returns strong or weak?

2.8 Consider the following model:

$$\log(wage)_i = \beta_1 + \beta_2 education_i + \beta_3 experience_i + \beta_4 experience_i^2 + \beta_5 female_i + u_i$$

where $\log(wage)_i$ is the logarithm of hourly wage of worker i , $education_i$ his/her level of education in years, exp_i his/her experience in years, fem_i a dummy variable equal to one if the worker is a woman and 0 otherwise.

- (a) What are the consequences of the introduction of an additional dummy variable $male_i$ equal to one if the worker is a man and 0 otherwise? Does this pose problems? If yes, which ones?
- (b) This model is estimated with a sample of 526 U.S. workers in 1976:

$$\begin{aligned}\widehat{\log(wage)} = & 0.392 + 0.082 \text{ education} + 0.036 \text{ experience} \\ & (0.102) \quad (0.007) \quad (0.007) \\ & - 0.002 \text{ experience}^2 - 0.329 \text{ female} \\ & (0.001) \quad (0.037)\end{aligned}$$

$$R^2 = 0.402, \quad N = 526$$

Interpret (i) the estimate of the parameter β_2 which measures the effect of education on wages, and (ii) the estimate of the parameters β_2 and β_3 which measure the effect of experience on wages.

- (c) Consider the simplified model estimated on the same data:

$$\begin{aligned}\widehat{\log(wage)} = & 0.401 + 0.079 \text{ education} - 0.340 \text{ female} \\ & (0.100) \quad (0.007) \quad (0.035)\end{aligned}$$

$$R^2 = 0.371, \quad N = 526$$

(i) The results are slightly different. Why? What is the interpretation of these parameters? (ii) Using these estimates and the previous ones, test the following hypothesis: "once the level of education and the sex are maintained constant, experience has no effect on wages". Make a test at the 5% level of significance.

2.9 The following equation was estimated:

$$\begin{aligned}\widehat{sleep} = & 3,638.25 - 0.148 \cdot totwork - 11.13 \cdot educ + 2.20 \cdot age \\ & (112.28) \quad (0.017) \quad (5.88) \quad (1.45)\end{aligned}$$

$$N = 706, \quad R^2 = 0.113.$$

where *sleep* is the number of sleeping hours over the year, *totwork* is the number of the number of working hours, *educ* is the education in years, and *age* is the age in years.

- (a) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.
- (b) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = 3,638.25 - 0.151 \cdot totwork$$

(112.28) (0.017)

$$N = 706, \quad R^2 = 0.103.$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

- (c) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?
- (d) Suppose that the *sleep* equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

2.10 In demand analysis, household demands are usually supposed to be zero homogeneous. That is to say, the demanded quantities do not change when prices and income vary in the same proportion. The econometrician considers J goods, and considers the following equation for the demand of good 1:

$$\log q_1 = \beta_0 + \sum_{j=1}^J \beta_j \log p_j + \beta_{J+1} \log Y + u,$$

where q_1 is the demanded quantity of good 1, $p_1, \dots, p_J$ are the prices of goods 1 to J , and Y is the household income. How can the homogeneity restriction be imposed in the model? How can this restriction be empirically tested?

2.11 The following regression equation is estimated as a Cobb-Douglas production function:

$$\widehat{\log Q} = 1.37 + \underset{(0.257)}{0.632 \log K} + \underset{(0.219)}{0.452 \log L}$$

$$R^2 = 0.98$$

where Q is the level of production of some given good, K the level of capital used in the production process and L the level of labor. Knowing that

$$\text{Cov}(\widehat{\beta}_{\log K}, \widehat{\beta}_{\log L}) = 0.055,$$

test the following hypothesis.

- (a) The capital and labor elasticities are identical.
- (b) There are constant returns to scale.

The problem does not give the number of observations. How can you proceed? Does this lack of information should affect your conclusions?

2.12 Prove that the adjusted R-squared increases with the addition of a variable only if the F statistic for testing the significance of that variable exceeds unity.

2.13 Under the Gauss-Markov hypothesis, does there exist a linear, but not necessarily unbiased, estimator of β that has a variance smaller than that of the OLS estimator? If so, how small can the variance be? Justify.

2.14 Prove that under usual assumptions $\text{Cov}(\widehat{\mathbf{u}}, \widehat{\beta} | \mathbf{X}) = 0$. Explain all the steps.
Hint:

$$\text{Cov}(\widehat{\mathbf{u}}, \widehat{\beta} | \mathbf{X}) = \mathbb{E} \left[(\widehat{\beta} - \beta)(\widehat{\mathbf{u}} - \mathbb{E}(\widehat{\mathbf{u}} | \mathbf{X})) | \mathbf{X} \right].$$

2.15 In the restricted OLS, the sum of squared residuals is minimized subject to the constraint implied by the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. For the Lagrangian as:

$$\min_{\mathbf{b}} \text{SSR}(\mathbf{b}) = \frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{R}\mathbf{b} - \mathbf{r})' \boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the Q -dimensional multiplier. The restricted least square estimator $\hat{\boldsymbol{\beta}}_R$ is the solution to the minimization problem.

(a) Show that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_R &= \hat{\boldsymbol{\beta}}_U - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \\ \boldsymbol{\lambda} &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_U = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the unrestricted least square estimator.

(b) Let $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_U$, the residuals from the restricted regression. Show that

$$\begin{aligned} \text{SSR}_R - \text{SSR}_U &= (\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)' (\mathbf{X}'\mathbf{X}) (\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) \\ &= (\mathbf{R}\hat{\boldsymbol{\beta}}_U - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_U - \mathbf{r}) \\ &= \boldsymbol{\lambda}' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'] \boldsymbol{\lambda} \\ &= \hat{\mathbf{u}}' \mathbf{P} \hat{\mathbf{u}} \end{aligned}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(c) Verify that you have proved that

$$F = \frac{(\text{SSR}_R - \text{SSR}_U)/Q}{\text{SSR}_U/(N - K)}.$$

2.16 Criticize the following statement: "It is a nonsense to test a hypothesis consisting of a large number of equality restrictions, because the t-test will most likely reject at least some of the restrictions".

2.17 Verify the identity

$$\mathbf{I}(\boldsymbol{\theta}_0) = -\mathbf{E} \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$$

for the linear regression model.

2.18 Consider the sample regression function:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_K x_K.$$

Then show that:

$$\hat{\beta}_k = \frac{\sum_{i=1}^N \hat{r}_{ki} y_i}{\sum_{i=1}^N \hat{r}_{ki}^2}, \text{ avec } k = 2, \dots, K,$$

where $\hat{r}_k$ with $k = 1, \dots, K$ is the residual obtained from the regression (with a constant) of the variable x_k on all the other explanatory variables, namely, $x_2, \dots, x_{k-1}, x_{k+1}, \dots, x_K$. Interpret the result and compare it with the OLS estimator of the simple regression model. Why can we assume that the denominator is different from zero?

Chapter 2

3.1 A sequence of real numbers is a trivial example of a sequence of random variables.

Is it true that $\lim_{N \rightarrow \infty} z_N = \alpha$ implies $\text{plim}_{N \rightarrow \infty} z_N = \alpha$? Hint: If $\lim_{N \rightarrow \infty} z_N = \alpha$, $|z_N - \alpha| < \varepsilon$ for N sufficiently large.

3.2 Suppose that $\sqrt{N}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \sigma^2)$. Does it follow that $\text{plim } \hat{\theta} = \theta$? Give a formal proof using the properties of convergence.

3.3 Prove that under assumptions 2.1–2.5,

$$\text{plim } \hat{\sigma}^2 = \sigma^2, \quad \text{where} \quad \hat{\sigma}^2 \equiv \frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2.$$

Hint: $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}} = \mathbf{x}_i' \boldsymbol{\beta} - \mathbf{x}_i' \hat{\boldsymbol{\beta}} + u_i$.

3.4 There is no unique way to write the linear hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, because for any conformable nonsingular matrix $\mathbf{F}$, the same set of restrictions can be represented as $\tilde{\mathbf{R}}\boldsymbol{\beta} = \tilde{\mathbf{r}}$ with $\tilde{\mathbf{R}} = \mathbf{F}\mathbf{R}$ and $\tilde{\mathbf{r}} = \mathbf{F}\mathbf{r}$. Does a different choice of $\mathbf{R}$ and $\mathbf{r}$ affect the asymptotic distribution of the Wald statistics? And its numerical value?

- 3.5 Find the 5% critical values of $F(10, \infty)$ and the 5% critical values of $\chi^2(10)$. What is the relationship between the two? Is it intuitive? Then prove that

$$\frac{SSR_R - SSR_U}{\hat{\sigma}^2} \xrightarrow{d} \chi^2(Q).$$

- 3.6 Let *savings* denote the amount saved by a household over the year, and let *income* denote its total income for the same period. Consider the savings function:

$$savings = \beta_1 + \beta_2 \cdot income + u, \quad u = \sqrt{income} \cdot \varepsilon$$

where ε is a random variable with $E(\varepsilon) = 0$ and $\text{var}(\varepsilon)$. Assume that ε is independent of income.

- (a) Show that $E(u_i | income) = 0$, so that the zero conditional mean assumption is satisfied. Show that $\text{var}(u_i | income) = \sigma^2 income$, so that the homoskedasticity assumption is violated. Provide a discussion that supports the assumption that the variance of savings increases with family income.
 - (b) How can the model be transformed to deal with heteroskedasticity, so that estimators are efficient?
- 3.7 Suppose that the variance matrix of the error term $\mathbf{u}$ is heteroskedastic with $\text{var}(\mathbf{u}) = \mathbf{V}\sigma^2$. Show that $\hat{\beta}_{GLS} = \beta + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{u}$. Moreover, under the classical hypothesis, and the null hypothesis, show that

$$F = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) / Q}{\hat{\mathbf{u}}'\mathbf{V}^{-1}\hat{\mathbf{u}} / (N - K)}$$

is distributed as a Fisher distribution with Q and $N - K$ degrees of freedom, where $\hat{\mathbf{u}}$ are the traditional residuals $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{GLS}$.

3.8 Consider the following estimation of the monthly total expenditure equation, i.e.,

$$\widehat{\text{totexp}} = \underset{(0.155)}{0.350} + \underset{(0.162)}{0.923} \text{ income} - \underset{(0.065)}{0.111} \text{ income}^2 .$$

where variables are in thousands of euros. The estimated variance matrix of the parameter estimates is:

$$\widehat{\text{Var}} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 0.024 & -0.008 & -0.002 \\ -0.008 & 0.026 & -0.005 \\ -0.002 & -0.005 & 0.004 \end{pmatrix} .$$

The average monthly income in the sample is 2.108 thousands of euros and the average monthly total expenditure 1.802 thousands of euros. The elasticity of total expenditure with respect to income computed at the average point of the sample is thus:

$$\epsilon = \frac{2.108}{1.803} (0.923 - 2 \times 0.111 \times 2.108) \approx 0.532..$$

Using the Delta method, compute the standard error or the asymptotic variance of this elasticity.

3.9 Consider the NLS estimation method. Then prove that

- (a) $\mathbb{E}(u_i | \mathbf{x}_i) = 0 \implies \mathbb{E} \left(\frac{\partial f(\mathbf{x}_i, \beta)}{\partial \beta} u_i \right) = \mathbf{0}$ for any $\mathbf{x}_i$.
- (b) $\mathbb{E}(u_i^2 | \mathbf{x}_i) = \sigma^2 \implies \mathbb{E} \left(\frac{\partial f(\mathbf{x}_i, \beta)}{\partial \beta} \frac{\partial f(\mathbf{x}_i, \beta)}{\partial \beta'} u_i^2 \right) = \Sigma_{\nabla \nabla} \sigma^2$ for any $\mathbf{x}_i$.

Chapter 3

4.1 Consider a sample of individuals containing information on wages, education, experience and so on. The econometrician selects the subsample of individuals who participate in the labor market (for who wages can be observed) and, using this sample, regresses the wage on the various explanatory variables. What can you say about the possible bias of the estimators?

4.2 Consider the following very simple keynesian macroeconomic model that explains aggregate consumption and gross national product:

$$\text{consumption function:} \quad C_i = \beta_1 + \beta_2 Y_i + u_i \quad (1)$$

$$\text{GNP identity:} \quad Y_i = C_i + I_i. \quad (2)$$

where Y_i is GNP, C_i is aggregate consumption, and I_i is aggregate investment.

- (a) Derive the reduced-form of the model.
- (b) Prove that Y_i in (1) is endogenous (in the econometrics sense) and that OLS estimator is biased.
- (c) Suggest a consistent estimator.

4.3 Consider the simple regression model $y = \beta_1 + \beta_2 x + u$ under the Gauss-Markov assumptions. For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $z_i = g(x_i)$. Define a slope estimator as

$$\tilde{\beta}_2 = \frac{\sum_{i=1}^N (z_i - \bar{z}) y_i}{\sum_{i=1}^N (z_i - \bar{z}) x_i}.$$

- (a) Show that $\tilde{\beta}_2$ is linear and unbiased. Remember, because $E(u|x) = 0$, you can treat both z_i and x_i as non random.
- (b) Add the homoskedasticity assumption. Show that

$$\text{var}(\tilde{\beta}_2) = \sigma^2 \frac{\sum_{i=1}^N (z_i - \bar{z})^2}{\left(\sum_{i=1}^N (z_i - \bar{z}) x_i \right)^2}.$$

- (c) Show directly that, under the Gauss-Markov assumptions, $\text{var}(\hat{\beta}_2) \leq \text{var}(\tilde{\beta}_2)$, where $\hat{\beta}_2$ is the OLS estimator. From which general theorem is this result a direct consequence?

4.4 Show that $E(u^2|\mathbf{Z}) = \sigma^2$ implies $E(u^2 \mathbf{z}' \mathbf{z}) = \sigma^2 E(\mathbf{z}' \mathbf{z})$.

4.5 Let $\mathbf{A}$ be a $L \times Q$ matrix of full row rank (so $L \leq Q$) such that $\mathbf{A}\Sigma_{\mathbf{z}\mathbf{x}}$ is of full column rank. Let $\mathbf{z}_i^* = \mathbf{A}\mathbf{z}_i$ (so $\mathbf{z}_i^*$ is a vector of L transformed instruments).

(a) Check whether Assumptions 3.3 and 3.4 continue to hold with the set of instruments $\mathbf{z}_i^*$.

(b) Verify that the 2SLS estimator reduces to the traditional OLS estimator if $\mathbf{z}_i = \mathbf{x}_i$.

4.6 Show in the general case that 2SLS estimators and IV estimators coincide when the model is exactly identified.

4.7 Consider the following equation estimated on a sample of individuals:

$$\begin{aligned} health_i = & \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot weight_i + \beta_3 \cdot height_i + \beta_4 \cdot male_i \\ & + \beta_5 \cdot work_i + \beta_6 \cdot exercise_i + u_i \end{aligned}$$

where $health_i$ is an indicator of individual i 's health, age_i is his/her age, $weight_i$ is his/her weight, $height_i$ is his/her height, $male_i$ a dummy equal to one if he/she is male, $work_i$ is his/her weekly working hours, and $exercise_i$ the average number of hours per week he/she dedicates to physical exercises. The following is assumed:

$$\begin{aligned} \mathbb{E}(u_i) &= \mathbb{E}(u_i \cdot age_i) = \mathbb{E}(u_i \cdot weight_i) = \mathbb{E}(u_i \cdot height_i) \\ &= \mathbb{E}(u_i \cdot male_i) = \mathbb{E}(u_i \cdot work_i) = 0. \end{aligned}$$

(a) For what reasons is $exercise_i$ supposed to be endogenous? What type of endogeneity is it?

(b) What kind of instruments would you suggest to estimate this model?

4.8 We consider the following regression to explain household savings:

$$s_i = \beta_1 + \beta_2 y_i + u_i$$

where s_i is the level of savings per months in dollars for household i , y_i is the level of the household income, u_i is the error term, β_1 and β_2 are parameters.

- (a) Do you think that y_i is exogenous? Why? Give some reasons that may explain why this variable is endogenous.
- (b) Would you modify the model? Would you use instruments? What instruments would you suggest?

4.9 The government of a developing country wants to set up a program in which poor families will receive vouchers for food rations. A researcher decides to do an experiment in which 500 families (each with one child) are randomly assigned to a treatment group ($T_i = 1$) who will benefit from food vouchers or a control group ($T_i = 0$) who will not benefit. He performs the following regression:

$$\begin{array}{rcccl} \widehat{H}_i & = & 0.653 & - & 0.209 \quad F_i \\ & & (0.038) & & (0.046) \end{array}$$

where H_i is a dummy variable equal to one if the child in the family is in poor health and 0 otherwise and F_i is a dummy variable equal to 1 if the family received food vouchers.

- (a) Interpret the results. Note: This model is the linear probability model.
- (b) The researcher finds that some families in the control group were mistakenly given food vouchers. In this case, can we consider that the variable F_i is exogenous? Why? Is it a traditional measurement error problem?
- (c) The researcher decides to use the IV method and, for this purpose, uses the assignment T_i to the treatment or control group as an instrumental variable. The results of the first-stage estimation are as follows:

$$\begin{array}{rcccl} \widehat{F}_i & = & 0.376 & + & 0.624 \quad T_i \\ & & (0.030) & & (0.030) \end{array}$$

Do you think that the conditions of relevance and exogeneity that an instrument must satisfy are satisfied?

- 4.10 With data from the United States, Evans and Schwab (in: *Finishing High School and Starting College: Do Catholic Schools Make a Difference?* The Quarterly Journal of Economics 110, 1995) study the effects of attending school in a Catholic high school on the probability of attending university. Let $college_i$ be a binary variable equal to 1 if the student goes to university; $cathHS_i$ a binary variable equal to 1 if the student comes from a Catholic high school. A linear probability model is estimated:

$$college_i = \beta_0 + \beta_1 cathHS_i + \text{autres facteurs} + u_i$$

where other factors include sex, ethnic origins, family income and parent education and u_i is the error term.

- (a) For what reason(s) could the variable $cathHS_i$ be correlated with the error term u_i ?
- (b) Let $cathrel_i$ be a binary variable equal to 1 if the student is from a Catholic family. Discuss the two conditions that this variable must satisfy to be a correct instrument. Which of these conditions can be tested? How to do this test? In the end, based on your intuition, do you think $cathrel_i$ is a valid instrument?
- (c) Evans and Schwab have data on the results in standardized tests done by each student before entering high school. What can they do with this variable to improve the estimates? Can these variables be used as instruments?