

Exercise sheet **3**
The Multiple Regression Model

Note: In those problems that include estimations and have a reference to a data set the students should check the outputs obtained with Gretl.

1. Let the model be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

where $E(\varepsilon|X_1, X_2) = 0$, and assume that we have a sample of size n .

- a) Derive the first order conditions of the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for the β .coefficients
 b) Show that

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \\ \hat{\beta}_1 &= \frac{1}{D}(s_{22}s_{1y} - s_{12}s_{2y}) \\ \hat{\beta}_2 &= \frac{1}{D}(s_{11}s_{2y} - s_{12}s_{1y})\end{aligned}$$

where

$$\begin{aligned}s_{11} &= \frac{1}{n} \sum_i x_{1i}^2, & s_{12} &= \frac{1}{n} \sum_i x_{1i}x_{2i} = s_{21}, & s_{1y} &= \frac{1}{n} \sum_i x_{1i}y_i \\ s_{22} &= \frac{1}{n} \sum_i x_{2i}^2, & & & s_{2y} &= \frac{1}{n} \sum_i x_{2i}y_i \\ D &= s_{11}s_{22} - s_{12}^2.\end{aligned}$$

and $y_i = Y_i - \bar{Y}$, $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$.

- c) What happens when $D = 0$? Interpret the result.
 d) Show that, when $s_{12} = 0$, the estimator $\hat{\beta}_1$ coincides with the estimator $\hat{\gamma}_1$ in the simple regression

$$\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 X_1,$$

and interpret the results.

2. Which of these situations, if any, does not comply with the assumptions of the classical regression model?

- a) The variable X_2 is the reciprocal of the variable X_1
 b) The variable X_2 is the variable X_1 squared
 c) The variable X_1 is an artificial variable that takes the value 1 for females and 0 for males and the variable X_2 is an artificial variable that takes the value 1 for males and 0 for females.

3. Even though wine is a consumption good, vintage wines can be considered as an investment good given their characteristics. In particular, we have data on the auction prices of thousands of red Bordeaux vintage wines from 1952 to 1980. These wines are stored for a considerable period of time before being consumed, which leads to an increase in the price given the

cost of storage. This entails an opportunity cost given the possibility of investing in other alternatives. Our data file `BORDEAUX.GDT` contains information on LPR (logarithm of the price of wine), $lluvinv$ (Amount of rainfall in the winter preceding the harvest), $tempmed$ (Average degree Centigrade while the grapes ripe), $lluvcos$ (Amount of rainfall while the grapes ripe), $edad$ (Number of years since the harvest)

- a) Estimate the linear projection of lpr on $edad$. Given the results, what would be the annual profitability from keeping the wine?
 - b) Carry out the multiple regression of lpr on $edad$, $lluvinv$, $lluvcos$, $tempmed$. How does the estimation of the profitability rate change? How do you explain this difference? (**Clue:** Which factors can affect the quality of wine?).
4. Assume that $\ln(wage)$ is the logarithm of monthly income and $educ$ is the years of education. Consider the linear model

$$\ln(wage) = \gamma_0 + \gamma_1 educ + v.$$

The file `WAGE2.GDT`, in Blackburn and Neumark (1992), contains data on monthly salaries, education, socioeconomic characteristics and the IQ (Intelligence Quotient) of 935 American males for 1980

- a) Is it reasonable to assume that $E(v|educ) = 0$? What other factors would be included in v ?
- b) Consider the model

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u,$$

where $abil$ is IQ and u satisfies the assumption $E(u|educ, abil) = 0$. Interpret the coefficient β_1 .

- c) Imagine that the initial sample available to us does not contain information on the intelligence quotient, IQ, of the workers, so the following linear projection is estimated

$$\begin{aligned} \ln(\widehat{wage}) &= 5,97 + \underset{(0,006)}{0,06} educ \\ n &= 935, R^2 = 0,097, \sum_i \widehat{v}_i^2 = 149,52. \end{aligned}$$

Explain under which conditions the OLS estimation of the slope of the linear projection is an unbiased estimator of β_1 . You should include a confidence interval of 95% for the slope of the estimated model.

- d) Assume that we get information about the workers' IQ. We obtain the following estimation

$$\begin{aligned} \ln(\widehat{wage}) &= 5,66 + \underset{(0,007)}{0,04} educ + \underset{s_{\widehat{\beta}_2}}{0,0059} abil \\ n &= 935, R^2 = 0,130 \end{aligned}$$

Test at the 5% significance level, that intelligence does not affect salary.

- e) Obtain the sample covariance of the education and the IQ. Given the results, interpret the parameter associated to the education in both regressions.

f) If $E(u^2|educ, abil) = \sigma^2(educ)$, where $\sigma^2(educ)$ is a positive function of $educ$, would this change any of your answers? What would be the consequences of the homoskedasticity assumption not being satisfied?

5. The variable *rdintens* is the expenditure on research and development (R&D) as a percentage of sales. Sales (*sales*) are measured in millions of dollars. The variable *profmarg* is the profits as a percentage of sales. Using the data file **RDCHEM.GDT**, which contains data on 32 firms in the Chemicals sector in the USA, we have obtained the following equation

$$\widehat{rdintens} = 0,472 + \underset{(0,216)}{0,321 \ln(sales)} + \underset{(0,046)}{\hat{\beta}_2} profmarg$$

$$n = 32, R^2 = 0,098$$

- a) Interpret the coefficient of $\ln(sales)$: If sales increase by 10%, what is the average estimated change, *ceteris paribus*, in *rdintens*? Is the effect economically significant? Is the effect statistically significant?
- b) The following alternative model has been estimated using the same database

$$\widehat{rdintens} = 1,104 + \underset{(0,216)}{0,302 \ln(sales)}$$

$$n = 32, R^2 = 0,061$$

Is the coefficient of *profmarg* in the model in section (a) significant? If we know $\hat{\beta}_2$ to be positive, what is its magnitude?

- c) Consider the following model that relates profits to sales

$$\widehat{profmarg} = 7,34 + \underset{s_{\hat{\gamma}_1}}{\hat{\gamma}_1} \ln(sales).$$

$$n = 32, R^2 = 0,0069$$

What is the sign and magnitude of $\hat{\gamma}_1$? Is it significant? Find $s_{\hat{\gamma}_1}$.

6. To study the effect of class attendance on final grades, we have used the data file **ATTEND.GDT** which contains data on 680 students enrolled in a class of Introduction to Microeconomics in an American university, and we estimate the following regression model

$$\begin{aligned} stndfnl = & \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 priGPA^2 \\ & + \beta_4 ACT + \beta_5 ACT^2 + \beta_6 (priGPA \times atndrte) + U, \end{aligned} \quad (1)$$

where *stndfnl* is the results of a standardized final exam, *atndrte* is the percentage of class attendance, *priGPA* is the average grade obtained in previous years and *ACT* is the university entrance grade. The results of the estimation are the following

OLS, using observations 1–680

Dependent variable: *stndfnl*

	Coefficient	Std. Error	t-ratio	p-value
const	2,0503	1.3603	1,5072	0.1322
<i>atndrte</i>	−0,0067	0.0102	−0,6561	0.5120
<i>priGPA</i>	−1,6285	0.4810	−3,3857	0.0008
<i>priGPA</i> ²	0,2959	0.1010	2,9283	0.0035
<i>ACT</i>	−0,1280	0.0985	−1,3000	0.1940
<i>ACT</i> ²	0,0045	0.0022	2,0829	0.0376
<i>priGPA</i> × <i>atndrte</i>	0,0056	0.0043	1,2938	0.1962

Mean dependent var	0.029659	S.D. dependent var	0.989461
Sum squared resid	512.7624	S.E. of regression	0.872872
<i>R</i> ²	0.228654	Adjusted <i>R</i> ²	0.221777
<i>F</i> (6, 673)	33.25018	P-value(<i>F</i>)	3.49e−35

1. a) What is the estimated partial effect of class attendance on the results of the final exam? Can we conclude that the effect is significant for a student whose average grade in previous years was equal to 3 if $\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_6) = 0,0001$?
 - b) If we add the term $\beta_7 (ACT \times atndrte^2)$ to the equation (1), what will be the partial effect of attending class in terms of the unknown parameters?
 - c) Explain how you would test that the model is linear in all the variables. Show clearly all the steps to follow in order to perform this test, that is, which is the null hypothesis and the alternative, statistic to be used (explaining its components and the decision rule)
7. Using the 526 observations in *WAGE1.GDT* on wage earners in the USA for 1976, we consider the following specification

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$$

where *wage* is the wage per hour, *educ* is education in years, *exper* in the work experience (in years) and *tenure* is the tenure in the current job (in years). Also, assume that $E(\varepsilon | educ, exper, tenure) = 0$ holds.

OLS, using observations 1–526

Dependent variable: *ln(wage)*

	Coefficient	Std. Error	t-ratio	p-value
const	0,2844	0.1042	2,7292	0.0066
<i>educ</i>	0,0920	0.0073	12,5552	0.0000
<i>exper</i>	0,0041	0.0017	2,3914	0.0171
<i>tenure</i>	0,0221	0.0031	7,1331	0.0000

Mean dependent var	1.623268	S.D. dependent var	0.531538
Sum squared resid	101.4556	S.E. of regression	0.440862
<i>R</i> ²	0.316013	Adjusted <i>R</i> ²	0.312082
<i>F</i> (3, 522)	80.39092	P-value(<i>F</i>)	9.13e−43

1. a) What is the interpretation of β_1 ? Given the estimation, which is the average wage difference between two workers with the same job experience and tenure if they differ in one year of education? Does it have a ceteris paribus interpretation?

b) The slopes of the multiple regression model measure the effect of the corresponding variable keeping the rest of variables constant. We will check this for the coefficient of education, following these steps:

(i) Estimate the linear projection of *educ* on *exper* and *tenure*.

- Then, save the residuals as a new variable, *res_educ*. For this, in the menu on top where the estimations are shown, choose Save → Residuals. A new window will appear, where you can change the name of the new variable.
- How are the coefficients? How do *exper* and *educ* change when *educ* changes (in which direction)?
- Which information do the residuals provide? (Clue: recall that the linear projection decomposes additively the dependent variable in an explained part and an unexplained part)

1) Estimate the linear projection of $\ln(\textit{wage})$

- Compare the coefficient of *educ* in this last estimation with that of the regression of $\ln(\textit{wage})$ on *educ*, *exper* and *tenure*
- What is the conclusion that we can draw from this?

8 Regression analysis can be used to test whether markets use information efficiently to value shares. Let *return* be the total return of the shares of a firm over a period of 4 years, from the end of 1990 to the end of 1994. The hypothesis of efficient markets says that this return should not be related in a systematic way to the information known in 1990. If the known characteristics of the firm at the beginning of the period could help predict the market return, then we could use that information to choose those shares over others. For 1990, let's define *dkr* as the quotient of the debt of the firm to its capital, *eps* the earnings per share, *netinc* the net income and *salary*, the salary of the CEO

a) Using the data in `RETURN.GDT`, estimate the following equation:

$$\textit{return} = \beta_0 + \beta_1 \textit{dkr} + \beta_2 \textit{eps} + \beta_3 \textit{netinc} + \beta_4 \textit{salary} + u.$$

Test whether the explanatory variables are jointly significant at the 5% level. Test whether *netinc* and *salary* are jointly significant. Is there any explanatory variable which is individually significant?

b) Re-estimate the model taking logarithms of *netinc* and *salary*

$$\textit{return} = \beta_0 + \beta_1 \textit{dkr} + \beta_2 \textit{eps} + \beta_3 \log(\textit{netinc}) + \beta_4 \log(\textit{salary}) + u.$$

How do your conclusions in section (a) change?

c) Why have we not applied logarithms to the variables *dkr* and *eps* in section (b) ?

d) In general terms, is the evidence in favour of the predictability of the share's return strong or weak?

9. Consider the Linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i.$$

Explain how you would test the following hypotheses:

- a) $\beta_1 = 0$.
- b) $\beta_1 = 0$ and $\beta_4 = \beta_5$
- c) $\beta_1 = 0$, $\beta_3 = 2$, and $\beta_4 = \beta_5$.

10 Consider the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad (2)$$

that satisfy the assumptions 1 to 4. We want to test the null hypothesis $H_0 : \beta_1 - 3\beta_2 = 1$.

- a) Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the OLS estimators of β_1 and β_2 respectively. Express $V(\widehat{\beta}_1 - 3\widehat{\beta}_2)$ in terms of $V(\widehat{\beta}_1)$, $V(\widehat{\beta}_2)$ and $C(\widehat{\beta}_1, \widehat{\beta}_2)$.
- b) Write down the t-statistic to test $H_0 : \beta_1 - 3\beta_2 = 1$.
- c) Defining $\theta = \beta_1 - 3\beta_2$ and its estimate (based on the OLS estimators of β_1 and β_2), $\widehat{\theta} = \widehat{\beta}_1 - 3\widehat{\beta}_2$, write down an equivalent specification of (2) where β_0 , θ , β_2 and β_3 are present, and it allows us to directly obtain $\widehat{\theta}$ and its standard error from a data sample.
- d) Explain an alternative strategy to that of section (b) to test $H_0 : \beta_1 - 3\beta_2 = 1$.

11 Consider the following specification for a production function:

$$y_i = \beta_0 + \beta_1 l_i + \beta_2 k_i + \varepsilon_i \quad (i = 1, \dots, n), \quad (3)$$

where $y = \log$ of output, $l = \log$ of labor, $k = \log$ of capital. Additionally assume that $E(\varepsilon_i | l_i, k_i) = 0$ for all l_i, k_i .

We would like to test the hypothesis of constant return to scale, that is, $\beta_1 + \beta_2 = 1$. Explain how you will do it for each of these cases:

2. We have the OLS estimates for the regression (3) and the respective variance and covariance matrix for the estimated parameters.
3. We have the OLS estimates of the regression of $(y_i - k_i)$ on a constant, $(l_i - k_i)$ and k_i .
4. We have the sum of the squared residuals for the OLS regression (3) and the sum of the squared residuals for a regression of $(y_i - k_i)$ on a constant and $(l_i - k_i)$.