Econometrics I

Written examination

Thursday May 02, 2019, 09:00 - 14:00

Examiner: Andreas Rosenblad, Department of Statistics, Stockholm University

Instructions

Allowed tools:

- Pocket calculator
- Text book: Wooldridge, J.M. *Introductory Econometrics: A Modern Approach.* Cengage Learning, Boston.
- Notes written in the text book are allowed.

Note that no formula sheet is provided.

Passing rate: 50% of overall total, which is 100 points. For detailed grading criteria, see the course description.

The maximum number of points for each problem is given in the right margin. If not indicated otherwise, to obtain the maximum number of points on each problem, detailed and clear solutions are required. Answers may be given in English or Swedish.

Solutions to the exam questions will be uploaded to Athena on the afternoon of May 02, 2019. The corrected exams will be available at the student office of the Department of Statistics three weeks after the date of the exam.

Question 1 (64 points)

The R package `wooldridge` contains the data set BWGHT2, which gives data on birth weights and associated characteristics for a number of children born in the U.S. We are interested in estimating the multiple linear regression model

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 drink + \beta_3 \log(meduc) + \beta_4 fage + \beta_5 mage + \beta_6 mage^2 + u$$

where $bwght$ gives the child's birth weight in grams, $cigs$ and $drink$ give the average number of cigarettes per day and drinks per week, respectively, consumed during pregnancy, $meduc$ gives the mother's education level in years, and $fage$ and $mage$ give the father's and mother's, respectively, age in years. The error term u is assumed to fulfill the usual requirements of normality, homoskedasticity, and independence. The R code and parts of the output for estimating this model using the sample of $n = 1680$ observations with complete cases are given below.

```
> library(wooldridge)
> out.bwght <- lm(bwght ~ cigs + drink + log(meduc) + fage + mage +
I(mage^2), data = bwght2)
> summary(out.bwght)

Call:
lm(formula = bwght ~ cigs + drink + log(meduc) + fage + mage +
    I(mage^2), data = bwght2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)             426.680   5.449 5.82e-08
cigs          -10.671     3.430
drink          -9.699              -0.202   0.8400
log(meduc)     -4.218    89.614  -0.047
fage                      3.385   2.069   0.0387
mage           62.645              2.296
I(mage^2)      -1.082     0.452             0.0168
---

Residual standard error: 565.1 on 1673 degrees of freedom
Multiple R-squared:  0.01419,    Adjusted R-squared:
F-statistic: 4.013 on 6 and 1673 DF,  p-value:
```

Note that the standard error of regression (SER) is called residual standard error in the R output.

(a) Assume that you did not know the sample size n used for this estimation. Show how could calculate n from the R output. (6)

(b) What is the value of the adjusted R-squared for this model? (6)

(c) State the null and alternative hypotheses for testing the overall significance of the regression, and perform the test using a significance level of 1%. What is your conclusion? (6)

(d) With $\hat{\sigma}$ denoting the standard error of regression for this model, let SST_2 denote the total sample variation for the average number of drinks consumed per week during pregnancy, and let R_2^2 denote the R-squared from regressing the latter variable on all other explanatory variables (including an intercept). Calculate the value of (8)

$$\sqrt{\frac{\hat{\sigma}^2}{SST_2\,(1-R_2^2)}}$$

(e) A researcher hypothesizes that if the average number of cigarettes smoked per day during pregnancy increases by one, a child's birth weight will decrease with 10 grams, ceteris paribus. State the null and two-sided alternative hypotheses for testing this research question, and perform the test using a significance level of 5%. What is your conclusion? (8)

(f) What is the functional form for mother's age that is used in this model called? (4)

(g) At what age of the mother does the child's birth weight reach its highest level, ceteris paribus? What is this point called? (8)

(h) Approximately, how much would a child's birth weight increase if the mother's education level increased with 1%, ceteris paribus? (6)

(i) Suppose that we are interested in the effect of the average number of drinks consumed per week during pregnancy on the child's birth weight measured in kilograms. What would the estimated slope coefficient be in this case? (6)

(j) Now, suppose that we measure the mother's education level in weeks instead of years, assuming that one year of education equals 40 weeks. Call this new variable $meducweeks = meduc \times 40$. What would the estimated slope coefficient of $\log(meducweeks)$ be if we used $\log(meducweeks)$ in the multiple linear regression model in place of $\log(meduc)$? (6)

Question 2 (36 points)

Give the correct answer for the following multiple-choice questions. No motivation is needed.

(a) A simultaneous equations model is represented by the equations (3)

$$y_1 = \gamma_1 + \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_2 = \gamma_2 + \alpha_2 y_1 + \beta_2 z_2 + u_2$$

where z_1 and z_2 are exogenous variables. How should the reduced form equation for y_2 be formulated?

A. By expressing y_1 as a linear function of y_2 and an error term.

B. By expressing y_2 as a linear function of y_1 and an error term.

C. By expressing y_1 as a linear function of y_2, exogenous variables and an error term.

D. By expressing y_2 as a linear function of exogenous variables and an error term.

(b) Which one of the following four statements is true for the OLS estimates of the simple linear regression model $y = \beta_0 + \beta_1 x + u$ and their associated statistics? (3)

A. The sum of the OLS residuals is positive.

B. The sample average of the OLS residuals is positive.

C. The sample covariance between the regressors and the OLS residuals is positive.

D. The point $(\bar{x}, \bar{y})$ always lies on the OLS regression line.

(c) Let SSR_r denote the sum of squares for a restricted model and SSR_{ur} the sum of squares for an unrestricted model when using the F-statistic (3)

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

Which one of the following four statements about this F-statistic is true when it is used for testing multiple linear restrictions in a multiple linear regression model?

A. If the calculated value of the F-statistic is higher than the critical value, we reject the alternative hypothesis in favor of the null hypothesis.

B. SSR_r is never smaller than SSR_{ur}.

C. The degrees of freedom of a restricted model is always less than the degrees of freedom of an unrestricted model.

D. The F-statistic is more flexible than the t-statistic for testing a hypothesis with a single restriction.

(d) If $\hat{\beta}_j$ is an unbiased and consistent estimator of β_j in a multiple linear regression model, which one of the following four statements is true? (3)

A. The distribution of $\hat{\beta}_j$ tends toward zero as the sample size n grows.

B. The distribution of $\hat{\beta}_j$ tends toward a standard normal distribution as the sample size n grows.

C. The distribution of $\hat{\beta}_j$ remains unaffected as the sample size n grows.

D. The distribution of $\hat{\beta}_j$ becomes more and more tightly distributed around β_j as the sample size n grows.

(e) Let GDP and FDI denote Gross Domestic Product and foreign direct investment, respectively. For the following equation, (3)

$$\log(GDP) = 2.65 + 0.527\log(bankcredit) + 0.222FDI$$

which one of the following four statements is true?

A. If GDP increases by one percent, bank credit increases by approximately 0.527 percent, ceteris paribus.

B. If bank credit increases by one percent, GDP increases by approximately 0.527 percent, ceteris paribus.

C. If GDP increases by one percent, bank credit increases by approximately log(0.527) percent, ceteris paribus.

D. If bank credit increases by one percent, GDP increases by approximately log(0.527) percent, ceteris paribus.

(f) Which one of the following four statements about the differences between the LPM model and the logit and probit models is true for a model with variables appearing in level form? (3)

A. The LPM assumes constant partial effects for the independent variables, while the logit and probit models imply non-constant partial effects.

B. The LPM assumes non-constant partial effects for the independent variables, while the logit and probit models imply constant partial effects.

C. The LPM assumes constant partial effects for the dependent variable, while the logit and probit models imply non-constant partial effects.

D. The LPM assumes constant partial effects for the independent variables, while the logit and probit models imply non-constant partial effects for the dependent variable.

(g) For the multiple linear regression model (3)

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

let $\hat{r}_{ij}$ denote the i^{th} residual from regressing x_j on all other independent variables and SSR_j the sum of squared residuals from the latter regression. What is the square root of the formula

$$\widehat{Var}\left(\hat{\beta}_j\right) = \frac{1}{SSR_j^2}\sum_{i=1}^{n}\hat{u}_i^2\hat{r}_{ij}^2$$

called?

A. The homoskedasticity-robust variance estimate for $\hat{\beta}_j$.

B. The homoskedasticity-robust standard error for $\hat{\beta}_j$.

C. The heteroskedasticity-robust variance estimate for $\hat{\beta}_j$.

D. The heteroskedasticity-robust standard error for $\hat{\beta}_j$.

(h) Suppose that z is an instrument for x in the simple linear regression model (3)

$$y = \beta_0 + \beta_1 x + u$$

Which one of the following four statements is true?

A. The condition $Cov(z, u) = 0$ can be tested statistically.

B. The condition $Cov(z, x) \neq 0$ cannot be tested statistically.

C. The instrumental variables estimator is biased if $Cov(x, u) \neq 0$.

D. The ordinary least squares estimator is unbiased if $Cov(x, u) \neq 0$.

(i) Which one of the following four statements is true? (3)

A. A variable is said to have a causal effect on another variable if both variables increase or decrease simultaneously.

B. The notion of 'ceteris paribus' plays an important role in causal analysis.

C. The difficulty in inferring causality disappears when analyzing data at high levels of aggregation.

D. The problem of inferring causality arises if experimental data is used in statistical analyses.

(j) Which one of the following four statements about the differences between least absolute deviations (LAD) and ordinary least squares (OLS) estimation is true? (3)

A. OLS is more computationally intensive than LAD.

B. OLS is more sensitive to outlying observations than LAD.

C. OLS is justified for very large sample sizes while LAD is justified for smaller sample sizes.

D. OLS is designed to estimate the conditional median of the dependent variable while LAD is designed to estimate the conditional mean.

(k) For which one of the following four situations will the Gauss-Markov theorem not hold for a multiple linear regression model? (3)

A. The error term has the same variance given any values of the explanatory variables.

B. The error term has an expected value of zero given any values of the independent variables.

C. The independent variables have exact linear relationships among them.

D. The regression model relies on the method of random sampling for collection of data.

(1) In the following model of the annual savings of an individual as a linear function of his/hers education and annual income, (3)

$$savings = \beta_0 + \beta_1 education + \beta_1 income + u$$

the variable education is a binary variable taking the value 1 if the individual is educated and 0 otherwise, while annual income is measured i U.S. dollars. Which one is the base (reference) category in this model?

A. The group of uneducated individuals.

B. The group of educated individuals.

C. The group of individuals with a high annual income.

D. The group of individuals with a low annual income.

Department of Statistics

Stockholms universitet

Correction sheet

Date: 02/05/2019

Room: Brunnsvikssalen

Exam: Econometrics 1

Course: Econometrics

Anonymous code: 0002-VEE

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X								3
Teacher's notes 64	33								

HU

Points	Grade	Teacher's sign.
97	A	AR

SU, DEPARTMENT OF STATISTICS

Room: Brunnsvikssalen **Anonymous code:** 0002-UEE **Sheet number:** 1

QUESTION 1

a) The output shows 1673 degrees of freedom.

Since d.f = n − k − 1, and we have k = 6 variables,

$n = 1673 + k + 1 = 1673 + 6 + 1 = 1680$ observations /6

b) $\text{Adj } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1} = 1 - \frac{(1-0{,}01419)(1679)}{1673}$

$\text{Adj } R^2 = 0{,}01065$ /6

c) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

H_A : At least one parameter is different from zero.

$\alpha = 0{,}01$ Statistic : $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ if H_0 is true.

$\sim F_{(6, 1673)}$

Reject H_0 if $F_{obs} > F_{(6,1673), \alpha=0{,}01}$

$F_{obs} = \frac{0{,}01419/6}{(1-0{,}01419)/(1673)} = 4{,}013$ corresponds with the output in R!

$F_{(6, 1673), \alpha=0{,}01} = 2{,}80$

$F_{obs} > F_{(6,1673), \alpha=0{,}01} \Rightarrow H_0$ is rejected!

The regression is significant on the 1% level. /6

(turn the page.)

d) The formula $\sqrt{\frac{\hat{\sigma}^2}{SST_2(1-R_2^2)}}$ is the standard error of $\hat{\beta}_2$.

Since $se(\hat{\beta}_2)$ is used in R to compute the t statistic, we can use the estimate $\hat{\beta}_2$ and its t value in the output to calculate the standard error.

$$t_{\hat{\beta}_2} = \hat{\beta}_2 / se(\hat{\beta}_2)$$

$$\sqrt{\frac{\hat{\sigma}}{SST_2(1-R_2^2)}} = se(\hat{\beta}_2) = \hat{\beta}_2 / t_{\hat{\beta}_2} = \frac{-9{,}699}{-0{,}202} = \underline{\underline{48{,}015}}$$

/8

e) $H_0 : \beta_1 = -10 \qquad \alpha = 0{,}05$

$H_a : \beta_1 \neq -10$

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \sim t_{1673} \text{ if } H_0 \text{ is true.}$$

Reject H_0 if $|t_{\hat{\beta}_1}| > t_{1673}(\alpha = 0{,}05)$

$$|t_{\hat{\beta}_1}| = \left| \frac{-10{,}671 - (-10)}{3{,}430} \right| = 0{,}1956$$

$$t_{1673, \alpha = 0{,}05} = 1{,}960$$

$|t_{\hat{\beta}_1}|$ is <u>not</u> larger than $1{,}960$,

so we fail to reject the hypothesis that an increase in smoked cigarettes by 1 per day would decrease the birth weight by 10 g ceteris paribus.

/8

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksalen Anonymous code: 0002-VEE Sheet number: 2

QUESTION 1

f) Quadratic form. /4

g) The turning point : $mage^* = \left|\frac{-\hat{\beta}_5}{2\hat{\beta}_6}\right| = \left|\frac{62,645}{2(-1,082)}\right|$

$mage^* = 28,949 \approx 29$

At approx. 29 years of age of the mother, the birth weight is at its highest point. /8

h) $\hat{\beta}_3 = -4,218$

$\Delta bwght = (\hat{\beta}_3/100)\% \Delta meduc$

A 1% increase in mother's education level decreases the birth weight by 0,042 grams. This is likely inaccurate, especially considering the very large standard error of $\hat{\beta}_3$. /6

i) $\hat{\beta}_{2,kilograms} = -0,009699$ so that a 1 drink increase gives a $-0,009699$ kg change, which is $-9,699$ grams. /6

Turn the page

j) The coefficient would be the same, as logarithmic variables are invariant to rescaling.

/6

SU, DEPARTMENT OF STATISTICS

Room: Brunnsviksssalen Anonymous code: 0002-UEE Sheet number: 3

QUESTION 2

a) D /3

b) D /3

c) B /3

d) D /3

e) B /3

f) A /3

g) C /0

h) C /3

i) B /3

j) B /3

k) C /3

l) A /3

Σ 33