Qualification Test for Ph.D. Program in Business
Research Methods

4/24/2020

For 1st semester:

1. An economics department at a large state university keeps track of its majors' starting salaries. Does taking econometrics affect starting salary? Let SAL = salary in dollar, GPA = grade point average on a 4.0 scale, $METRICS = 1$ if student took econometrics, and $METRICS = 0$ otherwise. Using the data containing information on 50 recent graduates, we obtain the estimated regression

$$\widehat{SAL} = 24200 + 1643GPA + 5033METRICS \quad R^2 = 0.74$$
(se) (1078) (352) (456)

(a) Interpret the estimated equation. (5%)

(b) How would you modify the equation to see whether women had lower starting salaries than men? (10%)

(c) How would you modify the equation to see if the value of econometrics was the same for men and women? (10%)

2. Please describe the method of testing the equivalence of two regression equations. (Hint: Chow test) (15%)

3. (a) Explain what is meant by (i) an omitted variable and (ii) an irrelevant variable. (10%)

(b) Explain the consequences of omitted and irrelevant variables for the properties of the least squares estimator. (10%)

4. Consider the following estimated regression equation (standard errors in parentheses):

$$\hat{y} = 5.83 + 0.869x \quad R^2 = 0.756$$
(se) (1.23) (0.117)

Rewrite the estimated equation that would result if

(a) All values of x were divided by 20 before estimation (10%)

(b) All values of y were divided by 50 before estimation (10%)

5. In multiple regression analysis, what are the relationships between ***t***- and ***F***-tests? (10%)

In multiple regression analysis, what are the relationship between t- and F-test?

Chapter 5.

$N-K = 51-3 = 48$

Ch 5.

6. Suppose from a sample of **51** observations, the least squares estimates and the corresponding estimated covariance matrix are given by

$K = 3$ ref 5.12

$N = 51$, $K = 3$

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}, \quad \widehat{\text{cov}(b)} = \begin{bmatrix} 3 & -2 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

Test each of the following hypotheses and state the conclusion: $t = \frac{\ldots}{\ldots} \sim t_{(48)}$.

(a) $\beta_1 + 3\beta_2 = 5$ (5%)

(b) $\beta_1 - \beta_2 + 2\beta_3 = 4$ (5%)

$\widehat{var}(c_1 b_1 + c_2 b_2 + c_3 b_3 | X) = c_1^2 \widehat{var}(b_1|X) + c_2^2 \widehat{var}(b_2|X) + c_3^2 \widehat{var}(b_3|X) + 2c_1c_2 \widehat{cov}(b_1, b_2|X) + 2c_1c_3 \widehat{cov}(b_1, b_3|X) + 2c_2c_3 \widehat{cov}(b_2, b_3|X)$

(a) $se(b_1 + 3b_2 - 5) = [var(b_1 + 3b_2 - 5)]^{1/2}$

$[Var(b_1) + Var(3b_2) + 2.3.1 \, cov(b_1, b_2)]^{1/2}$

H_0: $\beta_1 - \beta_2 + 2\beta_3 - 4 = 0$

H_1: $\beta_1 - \beta_2 + 2\beta_3 - 4 \neq 0$

(b) $se(b_1 - b_2 + 2b_3 - 4) = [Var(b_1 - b_2 + 2b_3 - 4)]^{1/2}$

$= [Var(b_1) + Var(b_2) + 2^2 Var(b_3) - 2\, cov(b_1, b_2) - 2.2\, cov(b_2, b_3) + 2.2\, (cov(b_1, b_3))]^{1/2}$

$= [3 + 4 + 4.(3) - 2(-2) - 2(0) + 4(1)]^{1/2}$

$= [7 + 12 + 4 + 4]^{1/2}$

$= (19 + 8)^{1/2}$

$= (27)^{1/2}$

$= 5{,}19$

$t = \frac{b_1 - b_2 + 2b_3 - 4}{5{,}19}$

$t = \frac{2 - 3 + 2(-1) - 4}{5{,}19}$

$t = \frac{-7}{5{,}19}$; $t = -1{,}349$

$-1{,}349 > 0 - t_c$

$b_1 = 2$, $b_2 = 3$, $b_3 = -1$

$Var\, b_1 = 3$, $var\, b_2 = 4$, $var\, b_3 = 3$.

$Cov(b_1, b_2) = -2$.

$Cov(b_2, b_3) = 0$

$Cov(b_1, b_3) = 0$.

Conclusion: do not reject H_0 at level significance 5%

I

Qualification Test for Ph.D. Program in Business

Research Methods

3/17-18/2022

For 1st semester:

1. A large company is accused of gender discrimination in wages. The following model has been estimated from the company's human resource information

$$\ln(WAGE) = 1.439 + .0834\, EDU + .0512\, EXPER + .1932\, MALE$$

where ***WAGE*** is hourly wage, ***EDU*** is years of education, ***EXPER*** is years of relevant experience, and ***MALE*** indicates the employee is male.

(a) What is the marginal effect of experience on wages? (5%) → log 0,0512 → exp (0,0512) − 1 = ?

(b) How much more do men at the firm earn, on average? (5%) → 0,1932 ×100% = 19,32% more than females

(c) What hypothesis would you test to determine if the discrimination claim is valid? (5%)

$H_0: \beta_{male} = 0$; $H_1: \beta_{male} \geq 0$

2. (a) Explain what is meant by (i) an omitted variable and (ii) an irrelevant variable. (5%)

(b) Explain the consequences of omitted and irrelevant variables for the properties of the least squares estimator. (10%)

read textbook maybe Ch. 5

K = 3

3. When using N = 50 observations to estimate the model $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + e_i$, you obtain SSE = 2132.65 and $s_y = 9.8355$. } Chp (4) (7)

(a) Find R^2. (5%) → Find the formula in ch 7

(b) Find the value of the F-statistic for testing $H_0: \beta_2 = 0, \beta_3 = 0$. Do you reject or fail to reject H_0 at a 5% level of significance? (5%)

4. Please describe the method of testing the equivalence of two regression equations. (Hint: Chow test) (10%)

read Ch. 7

3) ⓐ $R^2 = 1 - \frac{SSE}{SST}$

$S_y = \sqrt{\sum (y_i - \bar{y})^2 / (N-1)}$; 50−1

$S_y = 9,8355$

$\sum (y_i - \bar{y})^2 = SST$

$\sum (y_i - \bar{y})^2 = 9,8355^2 \cdot (50-1)$

ⓑ $F_s = \frac{(SSE_R - SSE_U) / J}{SSE_U / (N-K)}$

J is the number of joint hypothesis

J = 2

N = 50

K = 3

III

Qualification Test for Ph.D. Program in Business
Research Methods

3/17-18/2022

For 2nd semester:

✓ 1. Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u$$

where *health* is some quantitative measure of the person's health, *age, weight, height*, and *male* are self-explanatory, *work* is weekly hours worked, and *exercise* is the hours of exercise per week.

(a) Why might you be concerned about *exercise* being correlated with the error term u? (5%)

(b) Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with u. (5%)

2. Please describe a test for the existence of correlation between the error term and the explanatory variables in a model, explaining the null and alternative hypotheses, and the consequences of rejecting the null hypothesis. (15%) Ch 8

represents observation taken over time and the values of the variable in current position may be influenced by their past values.

3. Please explain (a) why lags are important in models that use time-series data, and (b) the ways in which lags can be included in dynamic econometric models. (15%) • Ch. 9 ARDL

4. Please describe the two-stage least squares estimation procedure for estimating an equation in a simultaneous equations model, and explain how it resolves the estimation problem for least squares. (10%) → 2SLS

10.3.6 page 495

✓ time series data represent the observation value of the current variable might be affected by the previous value of the variable

Ch. 8: Breusch - Pagan Test : Number 2

(b) the ways in which lags can be included in dynamic econometric models:
→ auto regressive model : which model a variable as a function of its own past values
→ distributed lags model : which model the relationship between a variable and its past values over a range of time period

In addition to autoregressive model & distributed lags model, lags can also included in dynamic econometric model with model using other techniques such as VAR (vector autoregression) & dynamic regression model

(3a) Lags are important because they can allow us to capture the fact of the current values of a variable may depend on past variable.

time series data: represent observation taken over time; the value of current variable may be influenced by their past values.

lags are used to capture the dependencies and to account for the dynamic nature of the data

lags are used to capture the dependencies and to account for the (dynamic)

2nd semester / 2022

5.2. Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1 \quad (5.53)$$

where *health* is some quantitative measure of the person's health; *age*, *weight*, *height*, and *male* are self-explanatory, *work* is weekly hours worked, and *exercise* is the hours of exercise per week.

a. Why might you be concerned about *exercise* being correlated with the error term u_1?

b. Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with u_1.

c. Now assume that *disthome* and *distwork* are in fact uncorrelated with u_1, as are all variables in equation (5.53) with the exception of *exercise*. Write down the reduced form for *exercise*, and state the conditions under which the parameters of equation (5.53) are identified.

d. How can the identification assumption in part c be tested?

a. Rules of thumb for regressors being correlated to the error term: (i) LHS and RHS variables determined by simultaneous decision (e.g., $Q^D_{chicken}$ as function of Q^D_{beef} and other factors; since chicken and beef are substitutes people's decision on how much to consume is a joint decision), (ii) omitted variable (i.e., regressor left out is captured by error term so if that omitted variable is correlated to any of the regressors in the model, the error term will be correlated to those regressors), (iii) LHS and RHS variables related by a constant (e.g., two equations for Q^D and Q^S, both as function of price; because equilibrium has $Q^D = Q^S$, price is automatically determined).

In this case, one could argue either case (i) or (ii). For the first one, (1) health and exercise could be jointly determined: if a person is not feeling well, he may not work out as much. In the second case, (2) we can easily think of variables that were omitted: family history (for genetic illnesses), occupation (e.g., teachers exposed to more illnesses).

b. I can't think of any reason why *disthome* and *distwork*, would be correlated to the error term. On the other hand, there's probably a strong correlated between these variables and *exercise*, because having a health club or gym nearer to home or work would make it more likely for someone to workout (assuming they workout in a gym... I don't). A better option may be *gymonway* set to 1 if there is a gym located between work and home.

c. Structural equations (full information):

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1$$

$$exercise = \alpha_0 + \alpha_1 age + \alpha_2 weight + \alpha_3 height + \alpha_4 male + \alpha_5 work + \alpha_6 disthome + \alpha_7 distwork + \alpha_8 health + \varepsilon$$

Reduced form (sub health equation into exercise equation):

$$exercise = \pi_0 + \pi_1 age + \pi_2 weight + \pi_3 height + \pi_4 male + \pi_5 work + \pi_6 disthome + \pi_7 distwork + v$$

This reduced form equation is what we use to estimate *exercise* in the first stage of 2SLS. Assuming *disthome* and *distwork* correlated to *exercise* and not correlated to u_1, then the estimate for *exercise* will not be correlated to u_1. When we plug that in for *exercise* in the second stage of 2SLS, the model (5.53) is identified (actually since we have two IVs it'll be over specified).