Эконометрика — Совбак ВШЭ и РЭШ, 2022 midterm

Совбак ВШЭ и РЭШ | Эконометрика | 2022 | midterm

QUESTION 1

True, false, or maybe true — 20 points

Indicate whether each statement is **true**, **maybe true**, or **false**, with a brief explanation.

(a) (4 points) If the correlation between X and Y is zero, then the slope coefficient from a regression of Y on X is also zero.

(b) (4 points) The slope coefficient from a regression of $y_i + c$ on $x_i + c$ is the same as the slope coefficient from a regression of y_i on x_i.

For parts (c)-(e), suppose

$$E(y_i \mid x_i) = \alpha + \beta x_i,$$

and all standard OLS assumptions hold.

(c) (4 points) The estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is unbiased.

(d) (4 points) In large samples, inference may be based on

$$\frac{\widehat{\beta} - \beta}{\text{s.e.}(\widehat{\beta})}$$

having a normal distribution.

(e) (4 points) The expression

$$\frac{\frac{1}{n}\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad e_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i,$$

is a good large-sample approximation to the sampling variance of $\widehat{\beta}$.

QUESTION 2

Short answers — 40 points

(a) (8 points) Suppose

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2,$$

but you do not know this and regress y_i only on x_i. Is the slope estimator unbiased for β_1?

(b) (8 points) Independent random variables X and Y have variances 9 and 25. You wish to estimate the difference between their means as precisely as possible and can collect at most 200 observations in total. How many observations should be collected for X and how many for Y?

(c) (8 points) Show that sample residuals are uncorrelated with fitted values:

$$\sum_{i=1}^{n} e_i \hat{y}_i = 0,$$

where

$$\hat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i, \qquad e_i = y_i - \hat{y}_i.$$

(d) (8 points) A regression of student test score on a dummy for whether the student's parents have higher education, using 200 students, gives a highly significant coefficient with $t = 10.3$, but $R^2 = 0.08$. Is this possible?

(e) (8 points) Suppose

$$E(y_i \mid x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

Let $\widetilde{\beta}_1$ be the slope from a regression of y_i on x_{1i} alone. Is $\widetilde{\beta}_1$ a consistent estimator of β_1? Suggest a method for consistently estimating β_1 using only simple regressions, each with one regressor and a constant.

QUESTION 3

Grants, maternal literacy, and test scores — 16 points

A regression uses data for 2,384 students in rural Kenya. The regressors are:

- $grants$: a dummy for whether the school received a cash grant that year;
- mom_lit: a dummy for whether the student's mother is literate.

Grant money was used to improve school quality by building and repairing classrooms and purchasing textbooks, desks, blackboards, and other equipment. Test scores range from 0 to 100.

The estimated regression is

$$\widehat{test\_score} = 48.4 + 1.20\, grants + 2.67\, mom\_lit, \qquad R^2 = 0.008,$$

with standard errors

$$(0.6) \qquad (0.66) \qquad (0.72).$$

(a) (4 points) Construct a 95% confidence interval for the effect of receiving a grant. Do students at schools receiving grants perform significantly better than students at schools that do not?

(b) (4 points) The grants in the observed year were approximately \2.50 $per student. What is the predicted effect on test scores of grants worth10 per student?

(c) (4 points) What is the estimated test-score difference between a student whose mother is literate and a student whose mother cannot read or write? Suppose the \10$ could instead be spent on educating the mother, with a 70% probability of making her literate. What is the estimated effect of that policy on test scores?

(d) (4 points) What proportion of variation in test scores is explained by the two regressors? Is this a large or small amount?

QUESTION 4

Computer use and the wage structure — 24 points

In Alan Krueger's paper *How Computers Have Changed the Wage Structure* (Quarterly Journal of Economics, 1993), computer use at work is used as a proxy for computer skills. The data come from the 1984 and 1989 U.S. Current Population Surveys.

Table I reports the percentage of workers in different groups who directly use a computer at work. Table II reports OLS regressions of log hourly wages on computer use and other controls. Standard errors are in parentheses.

Table I. Percentage using a computer at work

Group	1984	1989
All workers	24.6	37.4
Gender		
Men	21.2	32.3
Women	29.0	43.4
Education		
Less than high school	5.0	7.8
High school	19.3	29.3
Some college	30.6	45.3
College	41.6	58.2
Postcollege	42.8	59.7
Race		
White	25.3	38.5
Black	19.4	27.7
Age		
18-24	19.7	29.4
25-39	29.2	41.5
40-54	23.6	39.1
55-65	16.9	26.3
Occupation		
Blue-collar	7.1	11.6
White-collar	33.0	48.4
Union status		
Union member	20.2	32.5
Nonunion	28.0	41.1
Hours		
Part-time	23.7	36.3
Full-time	28.9	42.7

Group	1984	1989
Region		
Northeast	25.5	38.0
Midwest	23.4	36.0
South	23.2	36.5
West	27.0	39.9

Sample sizes are 61,712 in 1984 and 62,748 in 1989.

Table II. OLS estimates of the effect of computer use on pay

Dependent variable: $\ln(hourly\ wage)$.

Regressor	1984 (1)	1984 (2)	1984 (3)	1989 (4)	1989 (5)	19
Intercept	1.937 (0.005)	0.750 (0.023)	0.928 (0.026)	2.086 (0.006)	0.905 (0.024)	1.094 (
Uses computer at work	0.276 (0.010)	0.170 (0.008)	0.140 (0.008)	0.325 (0.009)	0.188 (0.008)	0.162 (
Years of education	—	0.069 (0.001)	0.048 (0.002)	—	0.075 (0.002)	0.055 (
Experience	—	0.027 (0.001)	0.025 (0.001)	—	0.027 (0.001)	0.025 (
Experience squared $/100$	—	-0.041 (0.002)	-0.040 (0.002)	—	-0.041 (0.002)	-0.040 (
Black	—	-0.098 (0.013)	-0.066 (0.012)	—	-0.121 (0.013)	-0.092 (
Other race	—	-0.105 (0.020)	-0.079 (0.019)	—	-0.029 (0.020)	-0.015 (
Part-time	—	-0.256 (0.010)	-0.216 (0.010)	—	-0.221 (0.010)	-0.183 (
Lives in SMSA	—	0.111 (0.007)	0.105 (0.007)	—	0.138 (0.007)	0.130 (
Veteran	—	0.038 (0.011)	0.041 (0.011)	—	0.025 (0.012)	0.031 (
Female	—	-0.162 (0.012)	-0.135 (0.012)	—	-0.172 (0.012)	-0.151 (
Married	—	0.156 (0.011)	0.129 (0.011)	—	0.159 (0.012)	0.143 (
Married $\times$ Female	—	-0.168 (0.015)	-0.151 (0.015)	—	-0.141 (0.015)	-0.131 (
Union member	—	0.181 (0.009)	0.194 (0.009)	—	0.182 (0.010)	0.189 (
Eight occupation dummies	No	No	Yes	No	No	
R^2	0.051	0.446	0.491	0.082	0.451	

Columns (2), (3), (5), and (6) also include three regional dummies. Sample sizes are 13,335 in 1984 and 13,379 in 1989.

(a) (5 points) Column (1) regresses wages on computer use alone. What is the earnings advantage of computer users in 1984? How does it change in 1989, using column (4)?

(b) (5 points) Does the simple regression imply a causal effect of computer use on wages? Why are additional regressors included in columns (2) and (3)?

(c) (9 points) Does Table I help explain the change in the computer-use coefficient across specifications? Is it surprising that the coefficient falls after adding controls?

(d) (5 points) Columns (4) and (6) show a significant earnings advantage from computer use even after adding many controls. Is this evidence of a causal effect, or is there still room for doubt?