ECON 2023: Introductory Econometrics with Recitation

National Taiwan University

Midterm Exam

April 28, 2021

79

Name: 林[illegible]

Student ID: B089010[illegible]

Instructions:

- This is a closed-book exam. You may use one A4-sized, double-sided sheet of reference notes.
- The exam lasts 150 minutes.
- There are 100 total points available.
- Answers in any languages other than English will NOT be graded.
- Show your work. Partial credits may be awarded if you obtain an incorrect answer but get some parts of the working correct.
- Keep your answers concise. A right answer hidden among irrelevant arguments will not give you the full credit.

Good luck!

1. Answer the following questions.

(a) **(3 points)** Explain the difference between an estimator and an estimate.

an estimator is the relationship between the data and the estimate
(β) (x) (y)

an estimate is what you estimate with the data and the estimator

-3

(b) **(3 points)** Discuss the intuition behind the fact that two random variables being uncorrelated does not mean they are independent.

uncorrelated $E(x)E(y) = E(xy) \not\Rightarrow P(x)P(y) = P(xy)$ independent

e.g. $x = 3,\ P(x) = 0.8$

$y = 4,\ P(y) = 0.2$

$E(x)E(y) = 3 \times 0.8 + 4 \times 0.2 = 3.2$

$= xy \cdot P(xy)$

$\Rightarrow P(xy) = \frac{3.2}{12} \neq P(x)P(y) = 0.16$

(c) **(3 points)** Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. In general, how does the standard error of the OLS estimator $\hat{\beta}_1$ change when the variance of X_i gets larger?

it also gets larger

-3

$\hat{\beta}_1$ is the estimated slope of $\hat{Y}_i$ to X_i, so as X_i has bigger variance, $\hat{Y}_i$ becomes more unpredictable, and the standard error of the slope $\hat{\beta}_1$ becomes larger.

(d) **(3 points)** The Gauss-Markov Theorem states that, under certain assumptions, the OLS estimator is the best linear unbiased estimator (BLUE). Explain what is meant by the OLS estimator being "linear."

-3

It means that the relationship between $\hat{Y}_i$ and X_i is linear, meaning, you can present the equation in the $y=ax+b$ way

i.e. $\frac{dy}{dx}$ is a constant, the change of $\hat{Y}_i$ on a unit change of X_i is always the same

(e) **(3 points)** Explain why the conditional mean independence assumption is weaker than the conditional mean zero assumption. (Hint: What is allowed under the conditional mean independence assumption but not allowed under the conditional mean zero assumption?)

Conditional mean zero : $E(u_i|X_i)=0$ ——①

conditional mean independence : $cov(u_i|X_i)=0$ ——②

① ⇒ ②

② ⇏ ①

$E(u_i|X_i)$ can be $\neq 0$ while $cov(u_i|X_i)=0$

but $cov(u_i|X_i)$ can't be $\neq 0$ while $E(u_i|X_i)=0$

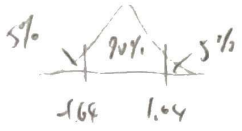
2. A standardized test is administered to 500 randomly selected students in Korea. In this sample, the mean is 250, and the standard deviation is 20.

(a) **(5 points)** Construct a 90% confidence interval for the average test score for Korean students. Interpret the interval in the context.

$\frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{500}} = 0.894$

$250 \pm 0.894 \times 1.64 = [248.534, 251.466]$

$(\Phi(-1.64) = 0.05)$



(b) **(5 points)** Another 500 students are selected at random from Korea. They are given a course before the test is administered. Their average test score is 262 with a standard deviation of 13. Is there statistically significant evidence that the course helped?

$H_0: \bar{Y}_b = \bar{Y}_a$

$H_1: \bar{Y}_b > \bar{Y}_a$

(b for those with the course
a for those without the course)

$SE(\bar{Y}_b - \bar{Y}_a) = \sqrt{\frac{s_b^2}{n_b} + \frac{s_a^2}{n_a}} = \sqrt{\frac{13^2}{500} + \frac{20^2}{500}} = 1.067$

$\frac{\bar{Y}_b - \bar{Y}_a}{SE(\bar{Y}_b - \bar{Y}_a)} = \frac{262-250}{1.067} = 11.246 > 1.64$

∴ reject H_0

There is statistically significant evidence that the course helped at the 5% level #

(c) **(5 points)** The original 500 students are given the course and then are asked to take the test a second time. The average change in their test scores is 5 points, and the standard deviation of the change is 20 points. Is there statistically significant evidence that students will perform better on their second attempt, after taking the course?

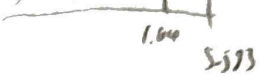
$H_0: \bar{Y}_c = \bar{Y}_a$

$H_1: \bar{Y}_c > \bar{Y}_a$

(c for session avg
a for original init)

$$\frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{500}} = 0.894$$

$$\frac{5}{0.894} = 5.593 > 1.64$$



∴ reject H_0

There is statistically significant evidence that students will perform better on their second attempt, after taking the course at the 5% level

3. Let $\bar{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider an estimator of μ: $W = \left(\frac{n-1}{n}\right)\bar{Y}$.

(a) **(5 points)** Show that W is biased and find the bias.

$$\text{bias} = E(\hat{\mu}) - \mu = E(W) - \mu$$

$$= E\left(\frac{n-1}{n}\bar{Y}\right) - \mu = \frac{n-1}{n}E(\bar{Y}) - \mu \quad \because E(\bar{Y}) = \mu$$

$$= \frac{n-1}{n}\mu - \mu = \mu\left(\frac{n-1}{n} - 1\right)$$

$$= -\frac{\mu}{n} \neq 0$$

$\therefore$ W is biased with the bias $= -\frac{\mu}{n}$

(b) **(5 points)** Show that W is consistent. (Hint: If $X \xrightarrow{p} \theta$ and $A \xrightarrow{p} a$, then $AX \xrightarrow{p} a\theta$.)

consistency: $\hat{\mu}_Y \to \mu_Y$ on large sample

W has a bias of $-\frac{\mu}{n}$

$$\lim_{n\to\infty}\left(-\frac{\mu}{n}\right) = 0$$

$\Rightarrow$ on large sample, bias of $W \to 0$, i.e. $W \to \mu$

$\therefore$ $W \to \mu$ on large sample, so W is consistent

(c) **(5 points)** Find var(W).

$$W = \left(\frac{n-1}{n}\right)\bar{Y} = \left(\frac{n-1}{n}\right)\frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\mathrm{Var}(Y) = \sigma^2$$

$$\mathrm{Var}(W) = \mathrm{Var}\left(\frac{n-1}{n}\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

$$= \left(\frac{n-1}{n^2}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(Y_i)$$

$$= \frac{(n-1)^2}{n^4} n\sigma^2$$

$$= \frac{(n-1)^2}{n^3}\sigma^2 \quad \#$$

4. Consider the regression model $Y_i = \alpha + u_i$.

(a) **(3 points)** Derive a formula for the least squares estimator of α.

$$\frac{\partial \sum_{i=1}^{n} (Y_i - \alpha)^2}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial (Y_i^2 - 2Y_i\alpha + \alpha^2)}{\partial \alpha}$$

$$= \sum_{i=1}^{n} (-2Y_i + 2\alpha)$$

$$= -2\left(\sum_{i=1}^{n} Y_i - n\alpha\right) = 0$$

$$\Rightarrow \hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

#

(b) **(3 points)** Show that the estimated residuals, $\hat{u}_i$, always sum to zero.

$$Y_i = \alpha + u_i$$

$$\hat{Y}_i = \hat{\alpha}$$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

$$= \alpha + u_i - \hat{\alpha}$$

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (\alpha + u_i - \hat{\alpha})$$

$$= \sum_{i=1}^{n} (\alpha + u_i) - \sum_{i=1}^{n} \hat{\alpha}$$

$$= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \quad \text{from (a)}$$

$$= \sum_{i=1}^{n} Y_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i = 0$$

#

5. **(4 points)** Suppose that average salary for college graduates depends on two factors, average GPA and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average):

$$SALARY_i = \beta_0 + \beta_1 GPA_i + \beta_2 BEAUTY_i + u_i.$$

Assume this equation satisfies the assumptions A1, A2, A3, and A4. Suppose that, on average, students with a higher GPA earn more after graduation, that better-looking students have a lower GPA at college, and that better-looking students earn more after graduation. What is the likely bias in the estimator for GPA coefficient obtained from the simple regression of $SALARY_i$ on GPA_i?

the simple regression only has one regressor, which is GPA, however, BEAUTY, an omitted variable, is also correlated with GPA, the included regressor, according to the premise

omitted variable bias will therefor appear in the simple regression

∴ the GPA coefficient would be underestimated in the simple regression, as when GPA goes up, BEAUTY goes down, while both GPA and BEAUTY are positively correlated to SALARY

6. A researcher investigates whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks, using data from fast-food restaurants in New Jersey and Pennsylvania. The level of observation is a fast-food restaurant. Table 1 describes the variables, Table 2 presents the summary statistics of the variables, and Table 3 summarizes the OLS estimation results.

Table 1: Data description

Variable	Description
$PriceSoda$	price of medium soda in USD
$PrpBlack$	proportion black for zip code
$Income$	median family income in USD1,000's for zip code
NJ	= 1 for New Jersey
BK	= 1 if Burger King
KFC	= 1 if Kentucky Fried Chicken
RR	= 1 if Roy Rogers
WEN	= 1 if Wendy's

Table 2: Summary statistics

Variable	Obs	Mean	Std Dev	Min	Max
$PriceSoda$	401	1.04	0.09	0.73	1.49
$PrpBlack$	401	0.11	0.18	0.00	0.98
$Income$	401	47.00	13.22	15.92	136.53
NJ	401	0.81	0.39	0.00	1.00
BK	401	0.41	0.49	0.00	1.00
KFC	401	0.20	0.40	0.00	1.00
RR	401	0.24	0.43	0.00	1.00
WEN	401	0.15	0.36	0.00	1.00

The level of observation is a fast-food restaurant.

Table 3: Estimated regressions

	Dependent variable: *PriceSoda*			
	(1)	(2)	(3)	(4)
PrpBlack	0.0649	0.1150	0.0614	0.0576
	(0.0240)	(0.0288)	(0.0281)	(0.0238)
Income		0.0016	0.0009	0.0002
		(0.0004)	(0.0003)	(0.0003)
NJ			0.0775	0.0778
			(0.0098)	(0.0076)
BK				0.0796
				(0.0093)
KFC				0.0252
				(0.0118)
RR				0.1274
				(0.0108)
Constant	1.0374	0.9563	0.9306	0.8954
	(0.0053)	(0.0191)	(0.0163)	(0.0146)
Obs	401	401	401	401
R^2	0.0181	0.0642	0.1689	0.4068

The level of observation is a fast-food restaurant. Heteroskedasticity robust standard errors are in parentheses.

(a) **(5 points)** What proportion of fast-food restaurants in the data are located in Pennsylvania?

from Table 2, we see that the mean for NJ is 0.81, meaning, 19% of fast-food restaurants in the data are located in Pennsylvania

(b) **(5 points)** Determine if the coefficient on *PrpBlack* in Model 1 is statistically significant. What is the expected change in *PriceSoda* if the share of blacks increases by 10 percentage points?

① $H_0: \beta_{PrpBlack} = 0$

$H_1: \beta_{PrpBlack} \neq 0$

$\frac{0.0649}{0.0243} = 2.7042 > 1.96$

∴ reject H_0

the coefficient on PrpBlack in Model 1 is statistically significant at the 5% level

② $0.0649 \times 0.1 = 0.00649$ (USD) #

(c) **(5 points)** If we measured *Income* in dollars instead of thousands of dollars, what would be the coefficient estimates in $PriceSoda_i = \beta_0 + \beta_1 PrpBlack_i + \beta_2 Income_i + u_i$?

coefficient of Income would become $\frac{1}{1000}$, while others won't change from Model 2

2 ∴ $\hat{\beta}_0 = 0.9563$

$\hat{\beta}_1 = 0.1150$

$\hat{\beta}_2 = 0.0016 \times \frac{1}{1000} = 1.6 \times 10^{-6}$

$u_i = 0$ (random selection) #

(d) **(5 points)** Compare the estimated coefficient on *PrpBlack* between Model 1 and Model 2. Is the coefficient larger or smaller when we control for income? Explain why the coefficient changes in this way.

the coefficient for PrpBlack is larger when we control for income

It's because Black people are more likely to be poorer than non-Blacks

i.e. generally, PrpBlack ↑ Income ↓

while both PrpBlack & Income are positively correlated with PriceSoda

the fall of Income accompanied by the rise of PrpBlack cancels down some of the PriceSoda rise resulting from the rise of PrpBlack, thus the coefficient of PrpBlack is lower in Model 1, when Income isn't controlled

(e) **(5 points)** Determine if the coefficient on NJ in Model 3 is statistically significant. Interpret the coefficient in the context.

$\frac{0.0775}{0.0098} = 7.91 > 1.96$

$H_0: \beta_{NJ} = 0$

$H_1: \beta_{NJ} \neq 0$

$\therefore$ reject H_0

the coefficient on NJ in Model 3 is statistically significant

Meaning, controlling PrpBlack and Income, PriceSoda in New Jersey is estimately 0.0775 USD higher than in Pennsylvania.

(f) **(5 points)** Calculate and interpret the adjusted R^2 for Model 3.

$$\begin{cases} \bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} \\ R^2 = 1 - \frac{SSR}{TSS} \end{cases}$$

$$\Rightarrow \bar{R}^2 = 1 + (R^2 - 1)\frac{n-1}{n-k-1}$$

$$= 1 + (0.1689 - 1)\frac{401-1}{401-3-1}$$

$$= 0.1626$$

R^2 is always higher than $\bar{R}^2$ (the meaningful one) as the more regressor, the proportion of ESS on TSS will certainly became bigger even if the newly added regressors aren't

(g) **(5 points)** Determine if the coefficient on KFC in Model 4 is statistically significant. Interpret the coefficient in the context.

$$\begin{cases} H_0: \beta_{KFC} = 0 \\ H_1: \beta_{KFC} \neq 0 \end{cases}$$

$$\frac{0.0252}{0.0118} = 2.1356 > 1.96$$

$\therefore$ reject H_0

the coefficient on KFC in Model 4 is statistically significant

Meaning, controlling PropBlack, Income, NJ, BK, and RR,

PriceSoda in KFC is estimately 0.0252 higher than that in WEN

(h) **(5 points)** Use Model 4. Write down the null and alternative hypotheses for testing whether prices of medium soda differ across fast-food chains.

H_0: $\beta_{BK} = 0 \wedge \beta_{KFC} = 0 \wedge \beta_{RR} = 0$

H_1: $\beta_{BK} \neq 0 \vee \beta_{KFC} \neq 0 \vee \beta_{RR} \neq 0$

use Model 3 as restricted, Model 4 as unrestricted $\quad q = 3$, $k = 6$

$$F^{act} = \frac{\frac{R^2_4 - R^2_3}{q}}{\frac{1 - R^2_4}{n-k-1}} = \frac{\frac{0.4068 - 0.1689}{3}}{\frac{1-0.4068}{401-6-1}} = 52.67 > F_{3,\infty}$$

$\therefore$ reject H_0

There is statistically significant evidence that the prices of medium soda differ across fast-food-chains

(i) **(5 points)** Consider a regression model without an intercept:

$$PriceSoda_i = \gamma_1 PrpBlack_i + \gamma_2 Income_i + \gamma_3 NJ_i + \gamma_4 BK_i + \gamma_5 KFC_i + \gamma_6 RR_i + \gamma_7 WEN_i + u_i.$$

Can we estimate the model using the data? If so, calculate the estimated coefficients for $\gamma_4, \gamma_5, \gamma_6$, and γ_7. If not, explain.

from Table 2, we see that the mean of BK + KFC + RR + WEN = 1

but now that the intercept is omitted, we can calculate the model without facing multicollinearity

$PriceSoda_i = \beta_1 PrpBlack_i + \beta_2 Income_i + \beta_3 NJ_i + \beta_4 BK_i + \beta_5 KFC_i + \beta_6 RR_i + \beta_0$

$= \beta_1 PrpBlack_i + \beta_2 Income_i + \beta_3 NJ_i + (\beta_4 + \beta_0) BK_i + (\beta_5 + \beta_0) KFC_i + (\beta_6 + \beta_0) RR_i + \beta_0 [1 - (BK_i + KFC_i + RR_i)]$

WEN_i

$\Rightarrow \begin{cases} \gamma_4 = \beta_4 + \beta_0 \\ \gamma_5 = \beta_5 + \beta_0 \\ \gamma_6 = \beta_6 + \beta_0 \\ \gamma_7 = \beta_0 \end{cases}$ #