Economics 361
Problem Set #6

Jun Ishii *
Department of Economics
Amherst College

Fall 2022

Question 1: Minimax Play at Wimbledon

This question requires you to read "Minimax Play at Wimbledon," by Mark Walker and John Wooders, published in the December 2001 volume of the *American Economic Review*, and the associated lecture handout "Hypothesis Testing and Tennis."

(a) Briefly explain why the authors believe that their test, using data from professional tennis matches, is a "better" test of the minimax theorem than earlier tests using data from minimax experiments (experimental data)?

(b) The intuition underlying the Pearson statistic is that one should reject the null hypothesis when the difference between the observed frequency of some event (O_j) and the expected frequency under the null hypothesis (E_j) is large. In other words, reject when $(O_j - E_j)$ is too positive or too negative. This suggests a "two-sided" hypothesis test. Yet, hypothesis tests involving the Pearson statistic are almost always one-sided: reject if the test statistic is too large (positive). Explain why.

(c) See Figure 4 on p.1532. What is the minimum value of the function being drawn? What does that minimum value indicate? Also, for what value of the "receiver's mixture probability on the left" does the function achieve its minimum? What is the relationship between that (argmin) mixture probability and the underlying hypothesis test?

(d) Use the data in Table 1 on p.1526 to test the null hypothesis of whether Sampras in the 1995 U.S. Open when serving in the "Ad" court utilized a proper mixed strategy. Use the Pearson test statistic and a significance level of 5%. Clearly show your work/steps. Start by showing why the Pearson test statistic value for this test is 1.524 (as stated in Table 1).

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

Question 2: Some Goldberger Problems

Problems **(a) - (c)** are adapted from Goldberger Problem 16.2. You are given a random sample (Y, X) that satisfies the Gauss Markov assumptions. In addition, you are told that $\sigma^2 = 1$ and

$$X'X = \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix}$$

Note that X is a (N x 2) matrix. Let β_1 denote the coefficient before the first variable and β_2 the second variable in X.

(a) Calculate $\text{Var}(b_1^{ols}|X)$, $\text{Var}(b_2^{ols}|X)$, and $\text{Cov}(b_1^{ols}, b_2^{ols}|X)$

Note: If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ then $A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

(b) Calculate $\text{Var}(b_1^{ols} + b_2^{ols}|X)$ and $\text{Var}(b_1^{ols} - b_2^{ols}|X)$

Let $t_1 = b_1^{ols} + b_2^{ols}$ and $t_2 = b_1^{ols} - b_2^{ols}$. Let $\theta_1 = \beta_1 + \beta_2$ and $\theta_2 = \beta_1 - \beta_2$

(c) Now answer Problem 16.2 in Goldberger:

> You are offered the choice of two jobs: estimate $\beta_1 + \beta_2$ or estimate $\beta_1 - \beta_2$. If you choose the former, you will be paid $10 - (t_1 - \theta_1)^2$. If you choose the latter, you will be paid $10 - (t_2 - \theta_2)^2$. To maximize your expected pay, which job should you take? What pay will you expect to receive?

(d) This is a slightly re-worded version of Goldberger Problem 20.1.

You are given a random sample (Y, X) that satisfies the classical normal regression model (Gauss Markov + Multivariate Normality) assumptions. In addition, you are told that $\sigma^2 = 2$ and

$$X'X = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

There are 32 observations in the sample. The OLS estimates are $b_1 = 3, b_2 = 2$

Test at the 5% significance level the following *joint* null hypothesis

- $H_0 : \beta_1 = 3$ and $\beta_2 = 3$

What is the alternative hypothesis against which you are testing?

(e) This is a slightly reworded version of Goldberger Problem 28.3.

$$\text{Suppose that:} \quad \begin{aligned} y_1 &= \theta + \epsilon_1 \\ y_2 &= 2\theta + \epsilon_2 \\ y_3 &= 3\theta + \epsilon_3 \end{aligned}$$

You observe $\{y_1, y_2, y_3\}$ but not $\{\epsilon_1, \epsilon_2, \epsilon_3, \theta\}$. $(\epsilon_1, \epsilon_2, \epsilon_3)$ are distributed **independently** of each other with the same mean of zero but different variances: $\sigma_1^2 = 4, \sigma_2^2 = 6, \sigma_3^2 = 8$.

Find the minimum variance linear unbiased estimator (MVLUE) of θ. Linear refers to linear function of (y_1, y_2, y_3).

Question 3: A Fitness Test for Heteroskedasticity

Consider the following "regression equation" for some observation i of a given sample (Y, X)

$$Y_i \;=\; X_i'\beta + \epsilon_i \qquad \text{note: } X_i \text{ and } \beta \text{ are both } (k \times 1) \text{ vectors}$$

We can express the regression equation in matrix form as

$$Y \;=\; X\beta + \epsilon$$

Let b^{ols} be the OLS estimator of β. X achieves full (column) rank and includes the customary column of ones ("1").

Denote the value of Y predicted by OLS as follows:

$$\hat{Y} \;=\; \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_N \end{pmatrix} \;=\; \begin{pmatrix} X_1'b^{ols} \\ \vdots \\ X_N'b^{ols} \end{pmatrix} \;=\; Xb^{ols}$$

Denote the associated residuals as follows:

$$e \;=\; \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix} \;=\; \begin{pmatrix} Y_1 - \hat{Y}_1 \\ \vdots \\ Y_N - \hat{Y}_N \end{pmatrix} \;=\; Y - \hat{Y}$$

(a) Explicitly show that $\hat{Y}'e = 0$

(b) Explicitly show that $Y'Y = \hat{Y}'\hat{Y} + e'e$

(c) The result shown in **(b)** is often expressed in words as

> "The total sum of squares (TSS) is equal to the explained sum of squares (ESS) plus the residual sum of squares (RSS)."

Briefly explain why.

Let $\bar{Y} = \frac{1}{N}\sum_i Y_i$ and $\bar{\hat{Y}} = \frac{1}{N}\sum_i \hat{Y}_i$ and $\bar{e} = \frac{1}{N}\sum_i e_i$

(d) Explicitly show that $\bar{Y} = \bar{\hat{Y}}$. **Hint:** $\bar{e} \;=\; ?$

(e) Explicitly show that the sample variance of Y is equal to the sample variance of $\hat{Y}$ plus the sample variance of e: $\frac{1}{N}\sum_i (Y_i - \bar{Y})^2 = \frac{1}{N}\sum_i (\hat{Y}_i - \bar{\hat{Y}})^2 + \frac{1}{N}\sum_i (e_i - \bar{e})^2$

———————- CONTINUED ———————-

A common (but not necessarily compelling) practice for researchers using OLS is to report a "goodness of fit" statistic known as R^2, defined as follows:

$$R^2 \equiv \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

Heuristically, R^2 reflects how much of the variation in Y is explained by the variation in the (OLS) predicted values of Y. Recall our earlier motivation of OLS as an estimator of the $BLP_{MSE}(Y|X)$ (and the $BP_{MSE}(Y|X)$ with the Linearity Condition).

(f) Explain why $R^2 \in [0, 1]$. Explain why R^2 may not be bounded between 0 and 1 if OLS is estimated *without* a constant (i.e. exclude column of ones from X).

(g) What must the values of $\{Y_i\}_{i=1}^N$ be in order for $R^2 = 1$? What must the values of $\{\hat{Y}_i\}_{i=1}^N$ be in order for $R^2 = 0$? For both questions, explain why.

Halbert White, in a classic 1980 article in *Econometrica*, proposed a "simple" test to see whether a given sample violated the Homoskedasticity condition (same conditional variance across observations) – i.e. test of $H_o : \sigma_i^2 = \sigma_j^2$ for all (i, j) in sample vs. $H_a : \sigma_i^2 \neq \sigma_j^2$ for some (i, j) in sample.[1] The other two Gauss-Markov conditions are assumed to be satisfied.

The steps for calculating the White test statistic for testing heteroskedasticity are as follows:

1. Run OLS on (Y, X) as usual, save the residuals $\{e_i\}_{i=1}^N$
2. Apply OLS on the following regression equation $(e_i)^2 = Z_i'\gamma + \eta_i$ where $\eta_i \equiv (e_i)^2 - Z_i'\gamma$
   where Z_i includes all elements in X_i, their squares, and their cross-products (excluding redundancies)
3. Calculate R^2 from the above "auxiliary" regression and multiply it by N to get the test statistic: $TS = NR^2$

White is able to show that, under the null hypothesis of homoskedasticity, the above test statistic has an *asymptotic* distribution of χ^2 with $q - 1$ degrees of freedom where q is the number of variables/elements in Z_i: $NR^2 \overset{a}{\sim} \chi^2_{q-1}$

(h) The derivation of this test statistic and its sampling distribution (when H_o is true) is beyond the scope of this course (but not by too much). However, the intuition for this test statistic is not, especially given the discussion earlier in this question. Provide the intuition.

Aside: There are many (including your Econ 361 professor and Arthur Golberger) who believe that R^2 is *much* over-valued by the profession. Note that even if $b^{ols} = \beta$ (perfect estimation), R^2 will generally be less than 1. However, R^2 is useful in deriving/calculating some test statistics.

[1] "A Heteroscedastic-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 1980, pp.817-838

Questions 4 requires you to use econometrics software, presumably STATA or R. The data file required for this problem set, wages.csv, is available on the course website. The other questions do not require econometrics software.

Question 4: Unobserved Ability and Fixed Effects

Suppose you are told that

$$\begin{aligned} E[ln(\text{Wage}_{it}) \mid \text{EXP}, \text{ABIL}] &= \beta_0 + \beta_1 \text{EXP}_{it} + \beta_2 \text{ABIL}_i \\ &\text{where} \\ ln(\text{Wage}_{it}) &\quad \text{log wage for person } i \text{ by year } t \\ \text{EXP}_{it} &\quad \text{years of work experience for person } i \text{ by year } t \\ \text{ABIL}_i &\quad \text{innate ability of person } i \end{aligned}$$

For M workers, you observe their wage and education level for T years, but not their innate ability. So you have a size $N = M \times T$ sample of $\{\ln(\text{Wage}_{it}), \text{EXP}_{it}\}_{i=1,t=1}^{M,T}$.

(a) Briefly explain why this sample is **not** random

(b) Show or explain why

$$E[ln(\text{Wage}_{it}) \mid \text{EXP}] = \beta_0 + \beta_1 \text{EXP}_{it} + \beta_2 E[\text{ABIL}_i \mid \text{EXP}]$$

(c) Suppose you believe that EXP_{it} and ABIL_i are positively correlated with each other; workers with greater innate ability tend to have more work experience. (e.g. Workers with higher innate ability tend to get work earlier and keep their jobs longer) What does this belief suggest about $\frac{d}{dEXP} E[\text{ABIL}_i \mid \text{EXP}]$?

Download the data set wages.csv from the course website. It contains experience, ability, and wage data for 4 workers over 10 years (sample size of 40). The dataset should be read using the following command: infile i t exp abil wage using wages.csv

If you stored wages.csv in a directory other than the STATA data directory, put the directory address in front of wages.csv. e.g. if the file is in C:\projects\data then use:
infile i t exp abil wage using C:\projects\data\wages.csv

The variables are

- i indicates the worker (1 through 4)
- t indicates the year (1 through 10)
- EXP indicates the experience for worker i in year t
- ABIL indicates the innate ability of worker i
- WAGE indicates the wage for worker i in year t

(d) Regress Wage on (EXP, ABIL, constant). Report the estimated coefficients and standard errors

(e) Regress Wage on (EXP, constant) – omitting ABIL. Report the estimated coefficients and standard errors

(f) Use the STATA Command correl EXP ABIL to calculate the sample correlation between EXP And ABIL. How does this sample correlation help explain the difference between estimated coefficient before **EXP** obtained in **(d)** and in **(e)**?

(g) Use the following STATA Commands to create a "dummy variable" for each worker: a variable that takes the value of 1 when the observation is associated with that worker and zero otherwise. Label the dummy variables (d1 - d4). These dummy variables (and their estimated coefficients) are known as "fixed effects."

- gen d1 = 0
- replace d1 = 1 if i == 1
- gen d2 = 0
- replace d2 = 1 if i == 2
- gen d3 = 0
- replace d3 = 1 if i == 3
- gen d4 = 0
- replace d4 = 1 if i == 4

(h) Try to regress Wage on EXP, d1, d2, d3, d4, constant. Explain why this fails. (**HINT:** Full rank)

(i) Regress Wage on EXP, d1, d2, d3, d4 (no constant). Compare these estimates to those of **(d)** and **(e)**. Do fixed effects seem to help address "omitted variables bias" ?

(j) Take a stab at the intuition for the result in **(i)** concerning fixed effects. **Hint:** See **(b)**