Department of Economics
University of Maryland

John C. Chao
Spring 2014

Economics 422

Midterm Examination

This exam contains 4 questions. You should try to do all 4 questions. Please provide explanations for all your answers. Right answers with no explanation will receive no credit. On the other hand, wrong answers with thoughtful explanations may receive some credit. Partial credit will be given. **GOOD LUCK!**

QUESTION 1 (24 POINTS)

Consider the linear regression

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, ..., n \tag{1}$$

Suppose that the following conditions hold.

A1 (X_i, Y_i), $i = 1, ..., n$ are $i, i.d.$

A2 $E[u_i|X_i] = 0$ with probabiliy one.

A3 Large outliers are rare: $E[X_i^4] < \infty$ and $E[Y_i^4] < \infty$.

Suppose you have collected a data set of size $n = 64$ on the variables (X_i, Y_i) and have estimated the model given by equation (1) using ordinary least squares (OLS) to obtain the following results.

$$\begin{aligned}
\widehat{\beta}_1 &= 0.1, \\
\widehat{\sigma}_v^2 &= \frac{1}{62}\sum_{i=1}^{64} \widehat{v}_i^2 = 9, \\
\widehat{\sigma}_X^2 &= \frac{1}{64}\sum_{i=1}^{64} \left(X_i - \overline{X}\right)^2 = 3
\end{aligned}$$

where $\widehat{v}_i = \left(X_i - \overline{X}\right)\widehat{u}_i$ with $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$.

(a) Given these results, construct a 95% confidence interval for the slope coefficient β_1. Show your work.

(b) Use the confidence interval you have constructed in part (a) to test the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_1 : \beta_1 \neq 0$. What is the significance level of the test you are implementing? Explain why you are able to test hypotheses using a confidence interval.

QUESTION 2 (28 POINTS)

Let $X_1, X_2, ..., X_n$ be an independent and identically distributed random sample. Suppose further that each X_i has a Bernoulli distribution, i.e.,

$$X_i = \begin{cases} 1 & \text{with probabiliy } p \\ 0 & \text{with probabiliy } 1 - p \end{cases}$$

for some unknown parameter $p \in (0, 1)$.

(a) Propose an unbiased estimator of p and show that the estimator you have proposed is indeed unbiased.

(b) It is well known that the variance of X_i in this case is given by the formula

$$\sigma_X^2 = Var\left(X_i\right) = p\left(1-p\right)$$

Show that the estimator

$$\widehat{\sigma}_X^2 = \overline{X}_n\left(1-\overline{X}_n\right), \text{ where } \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is not an unbiased estimator of σ_X^2.

(c) Propose an unbiased estimator of σ_X^2 and show that the estimator you have proposed is indeed unbiased.

QUESTION 3 (24 POINTS)

Consider a linear regression which seeks to explain scores on a particular standardized test based on whether one is left-handed, i.e.,

$$TS_i = \beta_1 LH_i + \beta_2 RH_i + \varepsilon_i, \ \ i = 1, ..., n; \tag{2}$$

where TS_i is the score of individual i on this particular standardized test and LH_i and RH_i are the following dummy variable

$$\begin{aligned} LH_i &= \begin{cases} 1 & \text{if person } i \text{ is left-handed} \\ 0 & \text{if person } i \text{ is right-handed} \end{cases}, \\ RH_i &= \begin{cases} 1 & \text{if person } i \text{ is right-handed} \\ 0 & \text{if person } i \text{ is left-handed} \end{cases}. \end{aligned}$$

(Assume that everyone has a dominant hand and that no one is precisely ambidextrous.) Suppose that the regression above satisfies the following assumptions:

A1 (LH_i, RH_i, u_i) are $i.i.d.$ for $i = 1, 2, ..., n$

A2 $E[u_i | LH_i, RH_i] = 0$ with probability one for $i = 1, ..., n$

A3 $E[u_i^4] < \infty$, i.e., large outliers are rare.

A4 $Var(u_i | LH_i, RH_i) = \sigma_u^2$ for every i, i.e., u_i is homoskedastic.

Suppose further that you have estimated equation (2) above based on data you have collected, and you have obtained the following results

$$\begin{aligned} \widehat{TS}_i &= \widehat{\beta}_1 LH_i + \widehat{\beta}_2 RH_i \\ &= \underset{(3.6)}{5} \times LH_i + \underset{(3.3)}{-5} \times RH_i, \\ n &= 18 \end{aligned}$$

and $Cov\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = 0$. The numbers given within the parentheses are standard errors.

(a) Suppose you are interested in the mean difference in test scores between people who are right-handed and people who are left-handed, i.e., the quantity

$$\tau = E[TS_i | \text{ person } i \text{ is right-handed}] - E[TS_i | \text{ person } i \text{ is left-handed}]$$

In particular, suppose you are interested in testing the null hypothesis $H_0 : \tau = 0$ versus the alternative hypothesis $H_1 : \tau \neq 0$. Using the results you have obtained above, implement such a test at the 5% significance level. Show your calculations.

(b) Suppose that in addition to assumptions A1-A4 above, you know that

$$u_i \text{ is distributed } N\left(0, \sigma_u^2\right).$$

Use this information to construct a better test of $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$ at the 5% signficance level. Show your work. Does your conclusion in part (b) differ from that in part (a)?

QUESTION 4 (24 POINTS)

Consider the linear regression

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, ..., n.$$

Suppose that the following assumptions hold:

A1 (X_i, u_i) are $i.i.d.$ for $i = 1, 2, ..., n$

A2 Both X_i and u_i are continuous random variables

A3 $E[u_i|X_i] = 0$ with probability one for $i = 1, ..., n$

A4 $E[X_i^4] < \infty$ and $E[u_i^4] < \infty$, i.e., large outliers are rare.

A5 $Var(u_i|X_i) = \sigma_u^2$ for every i, i.e., u_i is homoskedastic.

Suppose further that you have decided to estimate the slope coefficient β_1 using the estimator

$$\widetilde{\beta}_1 = \frac{1}{3} \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

(a) Derive a formula for the conditional variance $Var\left(\widetilde{\beta}_1 | X_1, ..., X_n\right)$ of $\widetilde{\beta}_1$.

(b) Compare the conditional variance formula you have derived in part (a) with the conditional variance of the OLS estimator, i.e., $Var\left(\widehat{\beta}_1 | X_1, ..., X_n\right)$, where

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}.$$

Which estimator has the bigger conditional variance $\widetilde{\beta}_1$ or $\widehat{\beta}_1$? Is this a violation of the Gauss-Markov Theorem? Why or why not?