

# PLSC 503: Solutions to Midterm

Thad Dunning

Department of Political Science

Yale University

Spring 2010

**Question 1:** Draws are being made at random with replacement from a box. The number of draws is getting larger and larger. Say whether each of the following statements is true or false, and explain. (Remember that “converges” means “gets closer and closer.”)

(a) The probability histogram for the sum of the draws (when put in standard units) converges to the standard normal curve.

(b) The histogram for the numbers in the box (when put in standard units) converges to the standard normal curve.

(c) The histogram for the numbers drawn (when put in standard units) converges to the standard normal curve.

(d) The probability histogram for the product of the draws (when put in standard units) converges to the standard normal curve.

(e) The histogram for the numbers drawn converges to the histogram for the numbers in the box.

(f) The variance of the numbers drawn converges to zero.

(g) The variance of the histogram for the numbers drawn converges to zero.

(h) The variance of the average of the draws converges to zero.

**Solutions:**

(a) True, by the central limit theorem.

(b) False; the distribution of the box is what it is. This distribution is unaffected by the size of the sample.

(c) False; the distribution of the numbers drawn will converge to the distribution of the box.

(d) False; the central limit theorem does not apply to the product, only the sum and average.

(e) True; see (c).

(f) False; see (e). If the numbers in the box have a positive variance, so will the numbers drawn.

(g) False; this is just (f) in disguise.

(h) True. The variance of the average is the variance of the box, divided by  $n$ . The numerator is a positive constant; as the denominator goes to infinity, the variance of the average goes to zero.

**Question 2:** An analyst fits a regression to a large data set. True or false, and explain:

(a) If the usual OLS assumptions are violated, the analyst can't calculate the  $R^2$  of the regression.

(b) If the usual OLS assumptions are violated, the  $R^2$  statistic does not have any ready interpretation.

**Solutions:**

(a) False; the  $R^2$  is a mechanical calculation and does not require a statistical model; it is a measure that compares the empirical variance around the regression line (or plane) to the total variance in  $Y$ . The OLS assumptions are not relevant.

(b) False, for the same reason as (a). Comparing the variance around the regression line (or plane) to the total variance in  $Y$  tells us how well the line (plane) fits the data. This could be useful if we're looking at, say, the relationship between fathers' heights and sons' heights. We don't need a statistical model for this.

**Question 3:** A large college course has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues this way:

“If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That’s what happened to me—chance variation.”

Is this a good defense? Answer yes or no, and explain briefly.

**Solution:** To answer the question, it is helpful to construct a box model. The TA says that the scores in his section are like 30 draws at random from a box containing all 900 scores. Under the null hypothesis, the box has a mean of 63 and an SD of 20. Thus, the expected value for the average of the draws is 63, and the SE is  $20/\sqrt{30} \doteq 3.65$ . By the central limit theorem, the average of the sample is normally distributed. So  $z = \frac{\text{observed}-\text{expected}}{\text{SE}} = \frac{55-63}{3.65} \doteq -2.2$ . Thus,  $p \doteq 0.01$  or 1%. (You can find the approximate  $p$ -value from a standard normal table).

It is not very plausible that this difference is due to chance variation: the defense is no good.

**Question 4:** A newspaper article says that on average, college freshmen spend 7.5 hours a week going to parties. One administrator thinks that these figures do not apply at her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours. Is the difference between 6.6 and 7.5 real, or can it be easily explained by chance?

(a) Formulate the null and alternative hypotheses in terms of a box model.

(b) Fill in the blanks. The null says that the average of the box is \_\_\_\_\_. The alternative says that the average of the box is \_\_\_\_\_.

(c) Now answer the question: is the difference real?

**Solution:**

(a) The box has one ticket for each freshman at the university, showing how many hours per week that student spends at parties. So there are about 3000 tickets in the box. The data are like 100 draws from the box.

(b) The null hypothesis says that the average is less than 7.5 hours. The alternative says that the average is less than 7.5 hours.

(c) The observed value for the sample average is 6.6 hours. The SD of the box is not known but can be estimated from the data as 9 hours. On this basis, the SE for the sample average is estimated as  $\frac{9}{\sqrt{100}} = 0.9$  hours. Then  $z \frac{(\text{observed}-\text{expected})}{\text{SE}} \doteq \frac{(6.6-7.5)}{0.9} = -1$ .

So the difference looks like chance.

**Question 5:** In a graduate class, 30 students take a midterm. 10 are left-handed and the other 20 are right-handed. The 10 left-handers score 83 (out of 100) on the exam on average, with an SD of 7, while the right-handers score 89 on average, with an SD of 9. Is the difference between 89 and 83 statistically significant? Explain.

**Solution:**

This is not a problem that lends itself readily to a statistical test. Try to formulate the problem as a box model. The question does not describe a process of sampling from a larger population; we've just got the 30 students at hand, which is at best a convenience sample from some undefined population—but how can we then set up a model for the sampling process? Meanwhile, there's no chance process that divides the students in the class into left-handers and right-handers (so this isn't like an experiment). Thus, the term “statistically significant” doesn't apply.

**Question 6:** A geography test was given to a simple random sample of 250 high school students in a certain large school district. One question involved an outline map of Europe, with the countries identified only by number. The students were asked to pick out Great Britain and France. As it turned out, 65.8% could find France, compared to 70.2% for Great Britain. Is the difference statistically significant? Or can this be determined from the information given? Explain.

**Solution:** This can't be determined from the information given. The responses are correlated. If we wanted to formulate this as a box model, the box should have four kinds of tickets, with values  $\{1, 1\}$ ,  $\{1, 0\}$ ,  $\{0, 1\}$ ,  $\{0, 0\}$ . Here,  $\{1, 1\}$  indicates that the student can find both France and Great Britain,  $\{1, 0\}$  indicates that the student can find France but not Great Britain, and so on. If we want to conduct a statistical test, it should be formulated in terms of this model; for instance, we might want to test the null hypothesis that the proportion who can find France but not Great Britain is equivalent to the proportion who can find Great Britain but not France. But this is a different question, and we would need more data.

**Question 7:** The great French kings of history had mediocre chief ministers, while the great ministers served under kings of lesser talent. Is this a fact of French history? Or of statistics? Explain briefly.

**Solution:** This is due to the laws of statistics: namely, the regression effect. Better-than-average kings may have better-than-average ministers, on average. But some of the greatness of really great kings is due to a lucky shock, and their ministers won't be quite so favored, on average. Same goes for great ministers; that's why on average they serve under kings of lesser talent.

**Question 8:** There is a study group of 10 subjects in a randomized controlled experiment. 7 of the subjects are assigned at random to treatment and 3 are assigned to the control group.

Observed data on the response variable look as follows:

Assigned to Treatment	Assigned to Control
3	–
2	–
5	–
6	–
3	–
4	–
5	–
–	2
–	4
–	3

(a) Construct a box model for this experiment, drawing on our discussion of the Neyman model.

What is in the box?

(b) Define the intention-to-treat parameter in terms of the model you constructed in (a).

(c) Estimate the intention-to-treat parameter, using the data in the table.

(d) Attach a standard error to your estimate in (c). To do this, use the formula for the variance of a difference-of-means of two independent samples. (To estimate the variance of tickets in the box, you should divide by the number of tickets in the sample, minus one.)

(e) Now, suppose an investigator assumes the OLS model:

$$Y_i = \alpha + \beta T_i + \epsilon_i, \tag{1}$$

where  $T_i$  is a 0-1 variable, with 1 indicating that a subject was assigned to treatment. Make a list of the “usual OLS assumptions.”

(f) Under the OLS model, what is  $E(Y_i|T_i = 0)$ ? How about  $E(Y_i|T_i = 1)$ ?

(g) Denote the design matrix as  $X$ . What is a typical row of this matrix? What size is  $X$ ? Denote the response variable as  $Y$ . What size is  $Y$ ?

(h) Calculate  $X'X$ ,  $(X'X)^{-1}$ ,  $X'Y$ , and  $(X'X)^{-1}X'Y$ . Use  $(X'X)^{-1}X'Y$  to estimate  $\alpha$  and  $\beta$ .

(i) Express  $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0)$  in terms of your estimates  $\hat{\alpha}$  and/or  $\hat{\beta}$ . How does this difference compare to your answer in (c)? Comment briefly.

(j) Calculate the OLS residual for each subject, and calculate the sum of squared residuals. (Show and/or describe your work).

(k) Now use the usual OLS formula to attach estimated standard errors to  $\hat{\alpha}$  and  $\hat{\beta}$ .

(l) Attach a standard error to the difference  $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0)$  you found in (i). How does this compare to your estimated standard error in (d)?

(m) Do you think the usual OLS assumptions are satisfied here? Why or why not? Which assumptions are the most plausible? What assumptions might be less plausible? Explain your answers carefully.

**Solution:**

(a) The box has 10 tickets; each ticket has two values,  $T_i$  and  $C_i$ . We will draw seven tickets at random without replacement and put those tickets in the assigned-to-treatment group; the other three tickets go in the control group. We will observe  $T_i$  for the tickets assigned to treatment, and  $C_i$  for the tickets assigned to control. In the table, for instance, we see seven observed values of  $T_i$  (the first seven subjects listed) and three observed values of  $C_i$ .

(b) The intention-to-treat parameter is the difference between two other parameters: (i) the average  $T_i$  if we assigned all 10 subjects to the treatment group, and (ii) the average  $C_i$  if we assigned all 10 subjects to the control group.

(c) We can estimate the average  $T_i$  we would observe if we assigned all 10 subjects to the treatment group by the average of the seven  $T_i$ 's that we did assign at random to the treatment group. The average outcome for subjects assigned to treatment is  $\frac{3+2+5+6+3+4+5}{7} = 4$ . Similarly, we can estimate the average  $C_i$  we would observe if we assigned all 10 subjects to the treatment group by the average of the three  $C_i$ 's that we did assign at random to the treatment group. This is  $\frac{2+4+3}{3} = 3$ . The estimated intention-to-treat parameter is the difference, that is,  $4 - 3 = 1$ .

(d) The number 4 in (c) is a realized value of the average  $T_i$  in the treatment group, which could have turned out differently in a different experiment. We can find the variance of this random variable as follows. The true variance of the mean is the variance of the  $T_i$ 's in the box, divided by  $n$ . Here,  $n$  is seven. But we don't know the variance of the  $T_i$ 's in the box. To estimate this variance, we can use the variance of the  $T_i$ 's in our sample (that is, the treatment group). In the formula for the variance of the  $T_i$ 's in our sample, we should divide by  $n - 1$ , not  $n$ , because we are using the sample mean (not the mean of the box) to find the variance of the  $T_i$ 's in our sample. Dividing by  $n - 1$ , the variance of the  $T_i$ 's in the treatment group an unbiased estimator for the  $T_i$ 's in the box.

Thus, the estimated variance of the  $T_i$ 's in the box is the sum of the squared deviations from the average outcome of 4, for each subject assigned to the treatment group, divided by  $n - 1$ . Here,  $n - 1 = 7 - 1 = 6$ . For the first subject in the table, whose outcome is 3, the squared deviation from average is  $(3 - 4)^2 = (-1)^2 = 1$ . The outcomes for the other treatment subjects are 2, 5, 6, 3, 4, and 5. Thus, the estimated variance of the  $T_i$ 's in the box is

$$\frac{(-1)^2 + (-2)^2 + (1)^2 + (2)^2 + (-1)^2 + (0)^2 + (1)^2}{6} = \frac{12}{6} = 2. \quad (2)$$

Similarly, the estimated variance of the  $C_i$ 's in the box is the sum of the squared deviations from the average outcome of 3, for each subject assigned to the control group, divided by  $n - 1$ . In the

control group,  $n - 1 = 3 - 1 = 2$ . So, our estimator of the variance in the box is

$$\frac{(2 - 3)^2 + (4 - 3)^2 + (3 - 3)^2}{2} = \frac{2}{2} = 1. \quad (3)$$

Now, the variance of the average in the treatment group, minus the average in the control group, is just the sum of the variances. (Recall that if  $A$  is one random variable and  $B$  is another, and  $A$  and  $B$  are independent, then  $\text{var}(A - B) = \text{var}(A) + \text{var}(B)$ .) The sum of the variances is  $1 + 2 = 3$ . The standard error is the square root of the variance, namely,  $\sqrt{3} = 1.73$ .

Note: we are doing these calculations as if we had two independent samples, which we don't. However, footnotes 10, 11, 14, and 21 on pages A-32 to A-36 in Freedman, Pisani, and Purves (2007) show why this is legitimate.

(e) Here are the usual OLS assumptions:

1.  $Y_i = \alpha + \beta T_i + \epsilon_i$ .
2.  $T_i \perp \epsilon_i$ .
3. The  $\epsilon_i$  are independent and identically distributed, with  $E(\epsilon_i) = 0$  and  $\text{var}(\epsilon_i) = \sigma^2$ .
4. The vector of  $T_i$ 's is linearly independent of the constant vector.

If we're inferring causation, we also need a response schedule, which tells us how unit  $i$  responds to being assigned to treatment or control:

$$5. Y_{i,T_i} = \alpha + \beta T_i + \epsilon_i.$$

(f) Under the model,  $E(Y_i|T_i = 0) = E(\alpha + \beta T_i + \epsilon_i|T_i = 0) = \alpha + E(\epsilon_i|T_i = 0) = \alpha$ , because  $T_i = 0$  so  $\beta$  drops out, and  $E(\epsilon_i|T_i) = 0$ . Similarly,  $E(Y_i|T_i = 1) = \alpha + \beta$ .

(g) The design matrix  $X$  is  $10 \times 2$ ; it has typical row  $[1 \quad T_i]$ .  $Y$  is  $10 \times 1$ .

(h) We'll write the second column of  $X$  with the seven subjects assigned to treatment stacked at the top, and the three subjects assigned to control at the bottom. The seven at the top thus have  $T_i = 1$ ; the three at the bottom have  $T_i = 0$ . Thus,

$$(X'X) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \quad (4)$$

$$= \begin{pmatrix} 10 & 7 \\ 7 & 7 \end{pmatrix}. \quad (5)$$

Now, the determinant of  $X'X$  is  $(10)(7) - (7)(7) = 21$ ; the adjoint is

$$\begin{pmatrix} 7 & -7 \\ -7 & 10 \end{pmatrix}. \quad (6)$$

So we have

$$(X'X)^{-1} = \frac{1}{21} \begin{pmatrix} 7 & -7 \\ -7 & 10 \end{pmatrix}. \quad (7)$$

You can verify that  $(X'X)^{-1}(X'X) = I_{2 \times 2}$ , as it should be. Now,

$$Y = \begin{pmatrix} 3 \\ 2 \\ 5 \\ 6 \\ 3 \\ 4 \\ 5 \\ 2 \\ 4 \\ 3 \end{pmatrix}. \quad (8)$$

(Remember, the treatment observations are stacked on top, and the control observations are stacked on the bottom). Thus,

$$X'Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 5 \\ 6 \\ 3 \\ 4 \\ 5 \\ 2 \\ 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 37 \\ 28 \end{pmatrix} \quad (9)$$

Finally,

$$(X'X)^{-1}X'Y = \frac{1}{21} \begin{pmatrix} 7 & -7 \\ -7 & 10 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 5 \\ 6 \\ 3 \\ 4 \\ 5 \\ 2 \\ 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 37 \\ 28 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}. \quad (10)$$

Thus, the OLS estimates of  $a$  and  $b$  are  $\hat{a} = 3$  and  $\hat{b} = 1$ .

(i)  $(\hat{Y}|T_i = 1) = \hat{\alpha} + \hat{\beta}$  and  $(\hat{Y}|T_i = 0) = \hat{\alpha}$ . Thus,  $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0) = \hat{\beta} = 1$ . This is exactly the difference-of-means we calculated in (c).

Lesson: When we have a dummy (0-1) independent variable and no control variables, an OLS regression of the response on a constant and the dummy variable gives the difference of mean outcomes. Here, the estimated constant is the average response in the control group, that is,  $(\hat{Y}|T_i = 0) = \hat{\alpha} = 3$ . The average response in the treatment group is  $(\hat{Y}|T_i = 1) = \hat{\alpha} + \hat{\beta} = 3 + 1 = 4$ .

(j) The residual is the difference between the actual and fitted (“predicted”) value. For example, for the first subject the actual value is 3, while the fitted value is  $\hat{\alpha} + \hat{\beta} = 4$ . (This subject was

assigned to treatment). So the residual is  $3 - 4 = -1$ . The vector of residuals is

$$e = \begin{pmatrix} -1 \\ -2 \\ 1 \\ 2 \\ -1 \\ 0 \\ 1 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \quad (11)$$

and the sum of squared residuals is

$$e'e = \begin{pmatrix} -1 & -2 & 1 & 2 & -1 & 0 & 1 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ 1 \\ 2 \\ -1 \\ 0 \\ 1 \\ -1 \\ 1 \\ 0 \end{pmatrix} = 14.$$

(k) The usual OLS formula for the estimated variance-covariance matrix of  $\hat{\gamma} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$  is

$$\text{cov}(\hat{\gamma}) = \hat{\sigma}^2(X'X)^{-1}, \quad (13)$$

where  $\hat{\sigma}^2 = \frac{e'e}{n-p}$ . Here,  $n - p = 10 - 2 = 8$ , so  $\hat{\sigma}^2 = \frac{14}{8} = 1.75$ .

Taking  $(X'X)^{-1}$  from (h), we have

$$\hat{\sigma}^2(X'X)^{-1} = \frac{1.75}{21} \begin{pmatrix} 7 & -7 \\ -7 & 10 \end{pmatrix} \doteq \begin{pmatrix} 0.583 & -0.583 \\ -0.583 & 0.833 \end{pmatrix} \quad (14)$$

The standard errors of  $\hat{\alpha}$  and  $\hat{\beta}$  are the square roots of the diagonal elements of (14), namely,  $SE_{\hat{\alpha}} = \sqrt{0.583} = 0.764$  and  $SE_{\hat{\beta}} = \sqrt{0.833} = 0.913$ .

(l) We showed in (i) that  $(\hat{Y}|T_i = 1) - (\hat{Y}|T_i = 0) = \hat{\beta}$ , so by (k) the standard error for the difference is 0.913.

This is very different from the standard error we found in (l), namely,  $\sqrt{3} = 1.73$ . Indeed, the standard error calculated under the OLS assumptions is about half as big as the standard error we found using the formula for the variance of a difference of means.

Lesson: If the Neyman statistical model is right, the OLS nominal standard errors (those computed from the usual regression formulas) can be very misleading.

(m) You might quibble with the assumption that  $Y_i = \alpha + \beta T_i + \epsilon_i$ : for example, the existence of an additive random error term is not guaranteed by the experiment or the randomization. However, if you are willing to accept the equation, it seems plausible that  $T_i \perp \epsilon_i$ ; after all, subjects are assigned at random to treatment and control.

On the other hand, the distributional properties of the error term seem more suspect. For example, why is  $\text{var}(\epsilon_i) = \sigma^2$  the same for all subjects? The formula for the variance of a difference-of-means adjusts automatically for heteroskedasticity—that is, unequal variances given  $T_i$ —in the treatment and control group. On the other hand, the usual OLS assumptions imply homoskedasticity. As we saw in (1), this can make a real difference to the estimated standard errors.

(A note on the usual full rank condition for OLS: since there is only one independent variable here, plus the constant,  $T_i$  will be linearly independent of the constant vector as long as there is some variance in  $T_i$ , which is guaranteed by the experimental design: seven subjects go to treatment and 3 to control. So the full rank condition is clearly satisfied).

**Question 9:** Suppose you assume the model  $Y_i = \beta X_i + \gamma Z_i + \epsilon_i$ , with the usual OLS assumptions. Here,  $X_i$  and  $Z_i$  are mean-zero scalar random variables, and  $\epsilon_i$  is the disturbance term. A critic expresses the following concerns. Comment on the validity of the critic's observations.

(a) Omitting the intercept in the model will lead to bias, because we are forcing the regression plane to go through zero.

(b) The disturbance term in the model may be correlated with the independent variables ( $X, Z$ ). Therefore, you should plot the residuals from your regression against  $X$  and  $Z$  to see if either of them show signs of being correlated with the disturbance term.

(c)  $X$  and  $Y$  are highly correlated, causing problems of multicollinearity. This problem, according to the critic, means that the estimates and standard errors are consistent but biased in small samples.

**Solution:**

(a) The critic is confused. It's not a problem that the regression plane goes through zero; the independent variables have zero mean, and when  $X$  and  $Y$  are both at 0,  $Y$  should be as well. (The expectation of the error term is also 0).

(b) The disturbance term in the model might well be correlated with the independent variables ( $X, Z$ ). But a plot of the residuals against  $X$  and  $Z$  isn't going to help; the residuals are orthogonal to the independent variables after any OLS fit. That is a property of OLS. (Here, there's no intercept, so the lack of correlation of the residuals and  $X$  and  $Z$  might not be exact).

(c) Our poor critic is again quite confused. Multicollinearity is about the relationship between  $X$  and  $Z$ , not  $X$  and  $Y$ . Anyway, even if the critic were talking about the correlation between  $X$  and  $Z$ , bias is not the issue. The issue is that the standard errors will be quite inflated.

**Question 10:** Suppose the true model is  $Y_i = \beta X_i + \gamma Z_i + \epsilon_i$ , with the usual OLS assumptions. Here,  $X_i$  and  $Z_i$  are mean-zero scalar random variables, and  $\epsilon_i$  is the disturbance term. We run a regression of  $Y$  on  $X$ , omitting  $Z$ . Show that:

(a) The fitted coefficient on  $X$  is a conditionally unbiased estimator for  $\beta$  if  $X$  and  $Z$  are statistically independent.

(b) The fitted coefficient on  $X$  is a conditionally biased estimator for  $\beta$  if  $X$  and  $Z$  are not statistically independent. Derive an expression for the bias, in terms of the covariance of the random variables  $X$  and  $Z$ .

**Solution:**

(a) The OLS regression of  $Y$  on  $X$  gives the least-squares fit

$$\frac{\text{cov}(Y_i, X_i)}{\text{var}X_i}. \quad (15)$$

Plugging in from the true model for  $Y_i$ , we have

$$\begin{aligned}\frac{\text{cov}(\beta X_i + \gamma Z_i + \epsilon_i, X_i)}{\text{var}X_i} &= \frac{\beta \text{var}(X_i)}{\text{var}(X_i)} + \frac{\beta \text{cov}(X_i, Z_i)}{\text{var}(X_i)} + \frac{\beta \text{cov}(X_i, \epsilon_i)}{\text{var}(X_i)} \\ &= \beta.\end{aligned}\tag{16}$$

Conditionally on  $X$ , the final term of (16) goes to zero, because  $X_i$  and  $\epsilon_i$  are independent.

Moreover, if  $X_i$  and  $Z_i$  are statistically independent, then they have zero covariance (as random variables), and the second term is zero as well. So the conditional expectation of (16) is  $\beta$ .

(b) If  $X_i$  and  $Z_i$  have non-zero covariance, then the second term of (16) is not zero. Thus, unless there is some strange relationship between  $X_i$  and  $Z_i$  (i.e. non-independence but zero covariance as random variables), the fitted coefficient of  $X$  is a conditionally biased estimator for  $\beta$ ; the bias is

$$\frac{\beta \text{cov}(X_i, Z_i)}{\text{var}(X_i)}.\tag{17}$$