Problem 1: \([1 + 1 + 1 + 3 = 5 \text{ points}]\) Consider \(n\) observations \((i = 1, \ldots, n)\) from the following model:

\[
y_i = \beta_0 + \beta_1 X_i + \epsilon_i,
\]

\[
E[\epsilon_i | X_i] = 0 \quad \text{and} \quad E[\epsilon_i^2 | X_i] = X_i^2.
\]

Now answer the following questions:

(a) What is the OLS estimator of \(\beta_1\)?

(b) Is this estimator unbiased? Is it consistent?

(c) What is the OLS standard error of this estimator? [Note that this is not the correct standard error.]

(d) Suggest a transformation of the model such that the OLS standard error of the estimator of \(\beta_1\) based on the transformed model is the correct one.

Answer 1:

(a) The OLS estimator of \(\beta_1\) is

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n) y_i}{\sum_{i=1}^{n} (X_i - \overline{X}_n) X_i}.
\]

(b) The only problem with this model seems to be the violation of the conditional homoskedasticity assumption. Therefore, unbiasedness and consistency are still going to hold for this estimator. [UP TO THIS SHOULD BE SUFFICIENT TO GET FULL CREDIT.]

To see this note that

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n) y_i}{\sum_{i=1}^{n} (X_i - \overline{X}_n) X_i} = \beta_1 + \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n) \epsilon_i}{\sum_{i=1}^{n} (X_i - \overline{X}_n) X_i}.
\]
Therefore, unbiasedness follows from

\[ E[\hat{\beta}_1] = E_X[E[\hat{\beta}_1|X]] = \beta_1 + E_X \left[ \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)E[\epsilon_i|X]}{\sum_{i=1}^{n}(X_i - X_n)X_i} \right] = \beta_1. \]

Consistency follows from

\[ \hat{\beta}_1 = \beta_1 + \frac{1}{n} \sum_{i=1}^{n}(X_i - \bar{X}_n)\epsilon_i = \beta_1 + \frac{\text{Cov}(X_i, \epsilon_i)}{\text{Var}(X_i)} \overset{p}{\to} \beta_1 + \frac{\text{Cov}(X_i, \epsilon_i)}{\text{Var}(X_i)} = \beta_1 \]

because \( \text{Cov}(X_i, \epsilon_i) = E[X_i \epsilon_i] - E[X_i]E[\epsilon_i] = E_X[X_i E[\epsilon_i|X_i]] - E[X_i]E_X[E[\epsilon_i|X_i]] = 0. \)

(c) The OLS standard error of this estimator is \( \hat{\sigma}_{\epsilon}\epsilon \) where \( \hat{\sigma} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2 \) is an estimator of the variance of \( \epsilon \) under the wrong assumption of conditional homoskedasticity. [Note that \( \hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{X}_n \).] Therefore, the OLS standard error is wrong!

(d) Note that we know that \( E[\epsilon^2|X_i] = X_i^2 \). Consider dividing both sides of the model by \( X_i \). Then you have a transformed model

\[ \left( \frac{y_i}{X_i} \right) = \beta_0 \left( \frac{1}{X_i} \right) + \beta_1 + \left( \frac{\epsilon_i}{X_i} \right) \]

and note that the regressors still only depend on \( X_i \)'s (although the constant intercept term now has coefficient \( \beta_1 \)). Now check that the new error, i.e., \( \frac{\epsilon_i}{X_i} \), is conditionally homoskedastic by noting that

\[ E \left[ \left( \frac{\epsilon_i}{X_i} \right)^2 |X_i \right] = \frac{1}{X_i^2} E \left[ \epsilon_i^2 |X_i \right] = \frac{X_i^2}{X_i^2} = 1 \text{ (constant)}. \]

In HW-4 you learned one way of transforming the model to remove heteroskedasticity. This is another way (in the textbook this is given as weighted least squares).
Problem 2: \[1 + 1 + 2 + 2 + (2+2) + (2+2) + 1 = 15 \text{ points}\] Please consult the attached log file to answer the following questions. The commands in the log file are given in bold. The output of the first command ”desc” describes the variables. Note that ”byte” stands for a dummy variable whereas ”float” stands for the usual continuous random variables. The data contain information on married women aged 21-35 with two or more children.

(a) Consult REGRESSION-1: On average, do women with more than 2 children work less than women with 2 children? How much less?

(b) Explain why the OLS regression, i.e., REGRESSION-1 is inappropriate for estimating the causal effect of fertility (morekids) on labor supply (weeks_mw).

(c) Consult REGRESSION-2: Interpret the coefficient of samesex. Please note that stating something like ”this is the marginal effect of samesex” is not an adequate answer and will not be awarded any credit.

(d) Consult the definition of interaction–1, interaction–2, interaction–3 and interaction–4. What do these interaction terms stand for? Please be precise.

(e) Consult REGRESSION-3: Interpret the coefficients of interaction–1 and interaction–2. Please note that stating something like ”this is the marginal effect of interaction – 1/2” is not an adequate answer and will not be awarded any credit. Do women prefer to have another child if the first two are boys or if the first two are girls? Is there any significant difference in this preference?

(f) Compare the results from REGRESSION-4 and REGRESSION-5. Is there any statistical evidence that the regressor morekids is endogenous? Intuitively explain why the IV estimate of the coefficient of morekids is more negative.
(g) Compare the results from REGRESSION-6 and REGRESSION-7. Is there any statistical evidence that the regressor \textit{morekids} is endogenous? Intuitively explain why the IV estimate of the coefficient of \textit{morekids} is less negative.

(h) Intuitively explain why the opposite directions in OLS bias in part (f) and part (g).

Answer 2:

(a) REGRESSION-1 suggests than, on average, women with more than 2 children work 6.23 weeks (per year) less than women with 2 children. The standard error of this estimate is .09 weeks and a 95 percent confidence interval for the number of weeks worked less can vary from 6.05 to 6.40 weeks.

(b) The OLS regression is inappropriate because \textit{morekids} is a potentially endogenous regressor. Various sources of endogeneity are discussed answer 2(f) and 2(g). In addition, one can also argue that it is not clear if having more than 2 children forced these women to work less or if they had more than 2 children because they were working less. In other words, without further assumption on the exogeneity of having more than 2 children, this coefficient cannot be interpreted as a causal estimate.

(c) From REGRESSION-2 we can interpret the coefficient of \textit{samesex} as

\[
\beta_{ss} = E[morekids|samesex = 1] - E[morekids|samesex = 0] = P[morekids|samesex = 1] - P[morekids|samesex = 0]
\]

because \textit{morekids} is a dummy variable (i.e. expectation = probability that the variable is 1). This means that having the first 2 children of same sex increases the probability of having a third child by .067 (i.e., 6.75 percent). The standard error is .2 percent and a 95
percent confidence interval indicates that the increase in probability can range from .064 to .071 (i.e., from 6.4 percent to 7.1 percent).

(d) The interaction terms can be interpreted as follows:

\( \text{interaction} - 1 = 1 \) if the first child is a boy and the second child is a boy; and 0 otherwise.

\( \text{interaction} - 2 = 1 \) if the first child is a girl and the second child is a girl; and 0 otherwise.

\( \text{interaction} - 3 = 1 \) if the first child is a girl and the second child is a boy; and 0 otherwise.

\( \text{interaction} - 2 = 1 \) if the first child is a boy and the second child is a girl; and 0 otherwise.

(e) REGRESSION-3 is richer than REGRESSION-2 because it helps us to compare the differential effect in having further kids when the first two are boys versus when the first two are girls. In other words, it helps us to form an estimate of the gender preference for the mother (probably both parents). As in answer 2(d), this regression implies that if the first 2 children are boys, the probability of having a third child increases by 5.77 percent (s.e. .2 percent and a 95 percent confidence interval ranges from 5.33 percent to 6.23 percent).

On the other hand, if the first 2 children are girls, the probability of having a third child increases by 7.83 percent (s.e. .2 percent and a 95 percent confidence interval ranges from 7.37 percent to 8.30 percent). These increases in probabilities are with respect to the omitted category of mothers whose first 2 children are of different genders.

A quick look at the two confidence intervals show that there is no intersection, and thus gives an indication that the preference to have a third child increases significantly more when the first 2 children are girls as compared to when they are boys. A formal F test strongly rejects (p-value = 0) equality of these two probability increases and thus indicates a strong relative preference for having at least one boy as opposed to having at least one girl.

(f) REGRESSION-4 implies that having more than 2 children decreases the number of weeks
worked by 5.38 (s.e. .089 and a 95 percent confidence interval ranges from 5.56 to 5.21 weeks of less work). On the other hand, REGRESSION-5 implies that having more than 2 children decreases the number of weeks worked by 6.31 (s.e. 1.27 and a 95 percent confidence interval ranges from 8.81 to 3.82 weeks of less work). Of course the two estimates are different, the latter indicating a 20 percent more decrease as compared to the former.

Intuitively this difference could be explained by the presence of an omitted variable, say, mother’s age (age_m). Typically mother’s age is positively related to both – morekids and weeks_mw. Since the age group of the mothers is 21–35, more age means more chance of having kids and also more chance of having a permanent job (and hence more weeks_mw). Therefore, leaving it out from the OLS, i.e., REGRESSION-4 will lead to an endogeneity of morekids and thus cause a positive omitted variable bias. (More on this in answer 2(h).)

A related question is how important is this bias (or inconsistency) in OLS. REGRESSION-5 is the preferred model if morekids is an endogenous regressor because in that case it produces consistent estimate whereas REGRESSION-4 produces inconsistent estimates. On the other hand if morekids is exogenous then REGRESSION-4 is the preferred model because it produces more efficient (precise) estimates - compare the s.e. from the two regressions. Therefore, we can use the Hausman test to compare between the two models. The Hausman statistic is

\[
H(\hat{\beta}_{mk}^{REG5} - \hat{\beta}_{mk}^{REG4})^2 / \text{s.e.}(\hat{\beta}_{mk}^{REG5})^2 - \text{s.e.}(\hat{\beta}_{mk}^{REG4})^2 = 0.539 < 3.814 \text{ (i.e. the } \chi^2 \text{ critical value)}.
\]

Therefore, it seems that the data do not provide enough evidence against REGRESSION-4 when compared against REGRESSION-5. So our concern for precision becomes more important and lead us to choose REGRESSION-4.

\[\text{1In the latter case, both regressions produce consistent estimates.}\]
(g) **REGRESSION-6** implies that having more than 2 children decreases the number of weeks worked by 6.23 (s.e. .088 and a 95 percent confidence interval ranges from 6.40 to 6.06 weeks of less work). On the other hand, **REGRESSION-7** implies that having more than 2 children decreases the number of weeks worked by 5.82 (s.e. 1.25 and a 95 percent confidence interval ranges from 8.26 to 3.38 weeks of less work). Of course the two estimates are different, the former indicating a 10 percent more decrease as compared to the latter.

Intuitively this difference could be explained by the presence of an omitted variable, say, mother’s education. Typically mother’s education will be positively related to weeks$_{mw}$ and negatively related to morekids. Therefore, leaving it out from the OLS, i.e., REGRESSION-4 will lead to a endogeneity of morekids and thus cause a negative omitted variable bias.

A related question is how important is this bias (or inconsistency) in OLS. REGRESSION-7 is the preferred model if morekids is an endogenous regressor because in that case it produces consistent estimate whereas REGRESSION-6 produces inconsistent estimates. On the other hand if morekids is exogenous then REGRESSION-6 is the preferred model because it produces more efficient (precise) estimates - compare the s.e. from the two regressions. Therefore, we can use the Hausman test to compare between the two models. The Hausman statistic is

$$H(\text{REG6, REG7}) = \frac{(\hat{\beta}_{mk}^{\text{REG7}} - \hat{\beta}_{mk}^{\text{REG6}})^2}{\text{s.e.}(\hat{\beta}_{mk}^{\text{REG7}})^2 - \text{s.e.}(\hat{\beta}_{mk}^{\text{REG6}})^2} = .108 < 3.814 \text{ (i.e. the } \chi^2 \text{ critical value).}$$

Therefore, it seems that the data do not provide enough evidence against REGRESSION-6 when compared against REGRESSION-7. So our concern for precision becomes more important and leads us to choose REGRESSION-6.

There are two important things to notice here when reading the answers 2(f) and 2(g). It is almost always possible to tell a plausible story to validate or invalidate any result – notice

---

\[2\text{In the latter case, both regressions produce consistent estimates.}\]
how the different omitted variables affect the results in different directions. So it is important to control for the variables that can cause such biases. This is even more important because the use of the instrumental variables (which are typically not well correlated with the endogenous regressor) causes the IV s.e. to be large and as a result the Hausman test often fails to distinguish between the competing models. So avoiding the omitted variable bias (and thus use of IV) by adequately controlling for the relevant variables is a good empirical strategy. (What other variables should be controlled for here? Family Income (reduces the cost of extra kid)? Religion (may forbid use of contraceptives)?)

(h) The opposite direction in bias could probably be explained through the omission of important variables – mother’s age and education in REGRESSION-4 and mother’s education in REGRESSION-6. As explained before in REGRESSION-6 this causes a negative bias. In REGRESSION-4 there are two competing biases operating in opposite directions (age causing positive bias, education causing negative bias). Given the strongly significant effect of age_m (see REGRESSION-6,7) it seems that the omission of mother’s age from REGRESSION-4 is the dominant source of bias in that regression. This can explain the opposite directions of OLS biases in REGRESSION-4 and REGRESSION-6.