Econometrics
Homework 2 Solutions

Question 1

Determine if each of the following statement is true or false. Briefly explain your answers.

(1) False. pex is positively related to score, but negatively related to attendance. Thus the coefficient on attendance is downward biased.

(2) True, because it reduces $R^2_{k,-k}$ but does not increase s^2.

(3) False. Note that it is possible to have omitted variable bias. For example, people take a busy job (given age and education), earn more and sleep less. It is because they take a busy job to earn more, not because they sleep less to earn more.

(4) False. There is likely omitted variable bias, such as omitted income variables.

(5) False. They cannot happen together, so the product must always be zero.

(6) True. (Here I refer to the usual OLS regression, also known as linear probaility model.) It is because $E(y_i|x_i) = \Pr(y_i = 1|x_i) = x_i'\beta$.

(7) False. It is used for heteroscedasticity with exogenous regressor when OLS estimator is used.

(8) False. Again, it is used to improve efficiency for heteroscedastic (and autocorrelated) errors when regressors are exogenous.

(9) True. It has to affect the endogenous regressor, but not through the error term to satisfy the exogeneity and relevance conditions.

(10) True. It is the FGLS for panel data when the error term ε_{it} are iid mean zero and c_i is uncorrelated to the regressors.

Question 2

(i) In this specification, $E(\ln(income)|city = 1, x_i) = \beta_0+\beta_1+\beta_2 x_i$, $E(\ln(income)|city = 2, x_i) = \beta_0 + 2\beta_1 + \beta_2 x_i$ and $E(\ln(income)|city = 3, x_i) = \beta_0 + 3\beta_1 + \beta_2 x_i$, so the effect between Guangzhou and Beijing is $2\beta_2$ and that between Shanghai and Beijing is β_2. So that's true.

(ii) We should define $Beijing_i = 1$ if the observation is from Beijing, 0 if not. Similarly, $Shanghai_i = 1$ if the observation is from Shanghai, 0 if not. We just need two dummies with a general constant term to avoid dummy variable trap. The equation becomes

$$\ln(income_i) = \beta_0 + \beta_{11} Beijing_i + \beta_{12} Shanghai_i + ... + u_i$$

Question 3

(a) Richer parents pay to have smaller class size ($\delta < 0$), and richer parents' children generally have better results due to better ability and family learning environment ($\beta_2 > 0$), and so the bias is negative.

(b) (Considering past results as the missing variable) If weaker students are assigned to small class ($\delta > 0$) and weaker students generally have low score ($\beta_2 > 0$), so the bias becomes positive.

(c) If there can be randomized experiments, then the missing variable has no correlation with class size ($\delta = 0$), then there is no omitted variable bias problem.

(d) You can put variables that can explain test score but may be related to student-teacher ratio in that specific environment. (You can have your own examples.)

Question 4

(a) For one percent higher in expenditure per student, the conditional mean of math passing rate is increased by 7.75, holding values of other regressors constant. (Positive effect is reasonable as more resources can improve students' score.)

For one percent larger in enrollment, the passing rate is reduced by 1.26, holding other regressors constant. (The negative sign is expected as a larger class may reduce class effectiveness.)

For one unit change in poverty (here indeed the unit is unclear), the math passing rate is reduced by 0.324, holding other regressors constant. (The negative sign is expected, as poorer students perform not as good because of less desirable study environment, family education or with weaker genes.)

The t-values are 2.55, -2.17, -9.0 respectively, so their absolute values are higher than the 5% critical value 1.96, and so all three slope coefficients are statistically significant.

(No need to interpret the intercept term.)

(b) It is because poverty is correlated to both expenditure and enrollment. Poverty is negatively correlated to log(expend) and poverty is negatively affecting math passing rate, so the bias is positive (upward).

On the other hand, poverty is also negatively correlated to log(enrollment). As enrollment has a negative effect on math passing rate, the bias is also positive (upward).

Question 5

Here let us consider the omitted variable bias in a simple case. Here we consider how attendence and previously having an econometrics class can affect the test score. To simply, consider both variables as binary variables. $attn_i = 1$ if one often attend class, and 0 if not. Also $preclass_i = 1$ if one has taken an econometric class before, zero otherwise.

The following table represents the mean score for each combination:

	number of obs	mean score
$attn = 1, preclass = 1$	10	90
$attn = 1, preclass = 0$	40	80
$attn = 0, preclass = 1$	40	80
$attn = 0, preclass = 0$	10	70

By the number of observations for each combination, it is clear that those who have previous econometrics class are less likely to attend classes.

(a) $\overline{score}(attn = 1) = [(10)(90) + (40)(80)]/50 = 82$; $\overline{score}(attn = 0) = [(40)(80) + (10)(70)]/50 = 78$. The effect of attendance obtained by this difference is 4.

(b) $\overline{score}(preclass = 1) = [(10)(90) + (40)(80)] = 82$. $\overline{score}(preclass = 0) = [(40)(80) + (10)(70)]/50 = 78$. The effect of previous economic class obtained by this difference is 4.

(c) To control away the effect of the other variable, one straightforward way is to compare the cases where the other variable is actually fixed.

Holding $preclass = 1$, the score difference becomes $90 - 80 = 10$. Similarly, holding $preclass = 0$, the score difference becomes $80 - 70 = 10$. So the actual effect of attendance is 10. Neglecting preclass under-estimate the effect of attendance because $attn$ and $preclass$ are negatively correlated, while both have positive effect on score.

(This is a made-up example for illustration so they are the same to illustrate the idea. In real data, due to sampling variations or more complicated data generating mechanism they are different.)

(d) Holding $attn = 1$, the score difference becomes $90 - 80 = 10$. Similarly, holding $attn = 0$, the score difference is $80 - 70 = 10$. The true effect is thus 10. Neglecting $attn$ underestimating the effect of $preclass$.

Question 6

Consider a linear model of monthly beer consumption:

$$beer_i = \beta_0 + \beta_1 inc_i + \beta_2 price_i + \beta_3 edu_i + \beta_4 female_i + u_i$$

If there is heteroscedasticity of the form

$$Var(u_i|x_i) = \sigma^2 inc_i^2$$

where x_i refers to the vector of all regressors.

(a) Yes. If the only violation of basic assumption is heteroscedasticity, the OLS estimator for β is unbiased and consistent. (No proof needed here.)

(b) No. In the deviation of the usual variance formula, homoscedasticity is used, so the formula is incorrect under heteroscedasticity. We should use White's heteroscedasticity robust standard errors instead.

(c) The null hypothesis is that there is homoscedasicity and the alternative hypothesis is heteroscedasticity. We should obtain the OLS residual of the above regression, called it e_i, obtain e_i^2 and run the following regression

$$\begin{aligned} e_i^2 \;=\; & \alpha_0 + \alpha_1 inc_i + \alpha_2 inc_i^2 + \alpha_3 price_i + \alpha_4 price_i^2 + \alpha_5 edu_i + \alpha_6 edu_i^2 + \alpha_7 female_i \\ & + \alpha_8 inc_i \times price_i + \alpha_9 inc_i \times edu_i + \alpha_{10} inc_i \times female_i + \alpha_{11} price_i \times edu_i \\ & + \alpha_{12} price_i \times female_i + \alpha_{13} edu_i \times female_i + v_i \end{aligned}$$

Note that $female$ is a binary variable, so its square is itself and so we don't put it again into the regression. (In some alternative forms, we may use only the original regressors or the polynomial of predicted values as regressors in this auxiliary regression, but the original form of White test is to use all regressors, its squares, and cross products.)

The test statistics is nR^2 where R^2 is that of the auxiliary regression. Under the null hypothesis it is distributed as $\chi^2(13)$, and we reject the null if the calculated nR^2 in the sample is at the largest 5% under $\chi^2(13)$.

(d) We can perform Generalized Least Squares by runing OLS on

$$\begin{aligned}\frac{beer_i}{inc_i} &= \beta_0\frac{1}{inc_i}+\beta_1\frac{inc_i}{inc_i}+\beta_2\frac{price_i}{inc_i}+\beta_3\frac{edu_i}{inc_i}+\beta_4\frac{female_i}{inc_i}+\frac{u_i}{inc_i}\\ &= \beta_0\frac{1}{inc_i}+\beta_1+\beta_2\frac{price_i}{inc_i}+\beta_3\frac{edu_i}{inc_i}+\beta_4\frac{female_i}{inc_i}+\tilde{u}_i\end{aligned}$$

Question 7

(a) General sources of problem: measurement error in regressors, omitted variables that are correlated to the regressors, and simultaneous equation (reverse causation). In this case, there may be simultaneous equation problem if hours one is willing to work can affect the wage the employers are willing to pay. Also, measurement error is a common problem when wage is calculated by using earnings divided by wage.

(b) (i) It is not in the equation as a regressor. (ii) It is correlated to wage. (iii) It is uncorrelated to the error term (determinants of hours of work not included as a regressor).

(c) We should run the first stage regression: regress lw on $ax, ax^2, hedu, kl6, k618, edu, loinc$ and age (a constant of course. It can be omitted when it's understood.) We should look at the F statistics of testing whether ax, ax^2 and $hedu$ all have zero coefficients. The rule of thumb is that it's weak if it's below 10. So, here the instruments are weak. Note the p-value is irrelevant here.

(d) We can do the overidentifying restriction test. Obtain the residuals for the IV regression, then regress the residual on all exogenous variables. Here, they are $ax, ax^2, hedu$, $kl6, k618, edu, loinc$ and age. Then, obtain the statistic nR^2. This is distributed χ^2 with degrees of freedom 2 under the null hypothesis of valid over-identification restrictions. (It's not robust to heteroscedasticity, but I don't want to complicate the matters here.) The p-value is a lot higher than 0.05, so it cannot reject the null of valid over-identification restrictions.

(e) Run the first stage regression described in part (c), obtain the residual (lw_res), and do the OLS on the original equation by adding this residual into the equation (i.e. regress lhr on $lw, kl6, k618, edu, loinc, age$ and lw_res and 1 using OLS). The t-statistic on the coefficient of lw_res is our test statistic, which is normally distributed. The p-value is so low that we reject the null that the regressor lw is exogenous.

(f) Using White's robust standard errors only affects the variance estimates, but not the coefficient estimates. T values are affected because it involves the standard errors.

(g) The wage elasticity is 1.761, which is positive. As $t = 1.761/0.599 = 2.94$, so it is statistically significant. It is reasonable, as the higher the wage, the more the women tend to work more, as leisure becomes more expensive and income effect is positive.

(h) Having small kids has a negative impact on hours of work, a smaller impact of the kids are older. More education also tend to work less in terms of hours. Older women work less. Those with higher income from other sources also work less (income effect on

leisure). The coefficients on kids below 6 and education are statistically significant at 5% level. (I am brief here.)

(i) We now test $\beta_{kl6} - \beta_{k618} = 0$ against they are not equal. Given $F = 3.945$, the degrees of freedom here are (1,420) and the 5% critical value is about 3.84, so we reject the null that the effects are the same at 5% level.

(j) One should add an interaction term $kl6 \times lw$. The coefficient would show the difference of wage elasticity on labor supply for those with a small kid than those without a small kid.

Question 8

(a) Since given education level and age (experience), more able people are more likely to join the party, so b_2 obtained this way is likely to also capture the effect of aibility, but not just the effect of being in the party.

(b) Some proxy of ability or human capital investment other than level of education may help. (e.g. parental income, or some past test scores, university GPA.)

(c) Parents' party membership can be valid instruments if it affects the child's probability to become members and has no effect on earnings directly, after controlling for other regressors. (correlated to children's party membership status but uncorrelated to error term of the earnings equation.)

(d) Maybe not, if party membership tends to be determined early on in life. It is useful unless some people change status between the surveys.

Question 9

(a) For consistency of pooled OLS, there should be no correlation between x_{kit} and $c_i + u_{it}$. That means there should be no correlation between individual effect as well as the (idiosyncratic) error term and all the regressors.

(b) Basically the same as part (a). (Note: This is more efficient when c_i differ across individuals.)

(c) No. There is autocorrelation in the error term $v_{it} = c_i + u_{it}$ of the estimated model within individual unit, so the usual OLS formula is not valid. One should be cluster robust standard errors, cluster at the unit of i.

(d) c_i includes anything that is individual specific, but have similar effects over time. For example, ability, family backgroud.

(e) The transformation is

$$\ln wage_{it} - \overline{\ln wage_i} = (x_{it} - \bar{x}_i)'\beta + (d_{it} - \bar{d}_i) + (u_{it} - \bar{u}_i)$$

so that the fixed effect c_i is eliminated before estimation. Consistency requires the error term and the regressors are not correlated for the same individual across different time periods.

Question 10

(a) **FE**: $\bar{y}_i = (y_{i1} + y_{i2})/2$ and $\bar{x}_i = (x_{i1} + x_{i2})/2$, and

$$\begin{aligned} y_{i1} - \bar{y}_i &= (x_{i1} - \bar{x}_i)'\beta + (\varepsilon_{i1} - \bar{\varepsilon}_i) \\ y_{i2} - \bar{y}_i &= (x_{i2} - \bar{x}_i)'\beta + (\varepsilon_{i2} - \bar{\varepsilon}_i) \end{aligned}$$

(Notice that the constant term 1 has also been differenced away. I have omitted the the details here.)

Now, $y_{i1} - (y_{i1} + y_{i2})/2 = (y_{i1} - y_{i2})/2$ and $y_{i2} - (y_{i1} + y_{i2})/2 = (y_{i2} - y_{i1})/2$. Similarly for x_{i1} and x_{i2}. So, for OLS, the objective function becomes

$$\min_{\beta} \sum_{i=1}^{n} \left(\frac{1}{2}(y_{i1} - y_{i2}) - \frac{1}{2}(x_{i1} - x_{i2})'\beta \right)^2 + \left(\frac{1}{2}(y_{i2} - y_{i1}) - \frac{1}{2}(x_{i2} - x_{i1})'\beta \right)^2$$

Note that the term inside the first and second squares are just negative of each other, and thus having the same value after squared, so it is the same as

$$\min_{\beta} \sum_{i=1}^{n} \frac{1}{2} \left((y_{i2} - y_{i1}) - (x_{i2} - x_{i1})'\beta \right)^2$$

Then for **FD**:

$$y_{i2} - y_{i1} = (x_{i2} - x_{i1})'\beta + (\varepsilon_{i2} - \varepsilon_{i1})$$

The objective function of OLS becomes

$$\min_{\beta} \sum_{i=1}^{n} \left((y_{i2} - y_{i1}) - (x_{i2} - x_{i1})'\beta \right)^2$$

Since the objective function of FE and FD differ only by a multiple 1/2, the minimizers are the same. Thus, FD and FE estimators are the same for $T = 2$.

(b) The first differenced model is consistent when $(x_{ki1} - x_{ki2})$ is uncorrelated to $(\varepsilon_{i1} - \varepsilon_{i2})$. So if ε is uncorrelated to x for oneself and between the pair of twins, then it is satisfied. (This allow correlation between c_i and x_{ij} as it is removed from the estimating equation already.)

(c) Since twins must be born on the same day, their age must be the same, so differencing will result in zeros. Thus, coefficients on age cannot be obtained.

(d) The coefficient implies that for a year increase in education, the wage is higher by about 9.2%. This is statistically significant, as $t = 0.092/0.024 > 1.96$.(I have forgotten to put in the number of observations, which is 149 pairs of twins.)

(e) The measurement error in regressors would result in attenuation bias (bias towards zero). (Measurement error is a problem here because we rely only on differences within pairs of twins, without using the differences across different pairs of twins.)

(f) First, though $\widetilde{x}_{i1} - \widetilde{x}_{i2}$ can still be subject to measurement error, if this is not correlated to the error term $\varepsilon_{i2} - \varepsilon_{i1}$ and the original measurement errors, then it is a valid instrument, as it is correlated to $x_{i1} - x_{i2}$.

(g) The FDIV estimate is much higher than FD estimate, which implies the measurement error leads to a serious downward bias. The standard error is higher than in FD case. It means the effect of 1 year of education, controlling effect from family and gene, is 16.7% and is statistically significant.

(From Ashenfelter, Orley and Krueger, Alan (1994) "Estimates of the Economic Return to Schooling from a New Sample of Twins" *American Economic Review* 84(5), 1157-73.)