Econometrics
Homework 3 Suggested Solutions

Question 1

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

(1) False. The missing variable *pex* is positively related to *score*, so $\beta_3 > 0$. Its relation to *attn* is negative, so $\delta_{3.2} < 0$. Therefore the bias is negative. (Intuitively, those who attend on average have weaker background and those who attend have stronger background, so it indirectly captures this effect.)

(2) True. Dropping this variable does not increase σ^2 but it makes $R^2_{k,-k}$ lower, which makes the variance of OLS estimator for other variables lower.

(3) False. There may be omitted variables, so it is not safe to regard this as a causal effect. One possibly is reverse causation. Those who earn more have to work longer hours, and sleep less, not sleeping less to earn more.

(4) False. We should at least clear possible candidates of omitted variables before we can make causal claims. One possibility is that those country with high income invest more on scientific research and also buy more electronic products.

(5) False, because we don't have data with both summer and winter to be true, so it must always be zero and we can never identify the coefficient.

(6) True, because $E(y_i = 1|x_i) = \Pr(y_i = 1|x_i)$

(7) False. White's robust standard error is used for heteroscedasticity.

(8) False. It is only used when there is heteroscedasticity or autocorrelation in the error term to improve efficiency. It cannot cure endogeneity problem.

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.

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

(a) The higher the growth in population, the lower the average gdp growth rate. The absolute t-stastic is $2.62 > 1.98$, so it is statistically significant at 5% level. (Whether this is causal may be debateable.)

(b) The higher the physical capital, the higher the growth rate. The t-value is $2.51 > 1.98$, so it is statistically significant at 5%.

The higher the human capital (measured by average education), the higher the growth rate. The t-value is 5.5, so it is statistically signficant.

(As these are input in the production, the higher the stock, the more an economy can produce.)

(c) For non-OECD:

$$\widehat{gdp\_pw_i} = -0.068 - 0.063 gpop_i + 0.719 s_{k,i} + 0.044 educ_i$$

For OECD:

$$\widehat{gdp\_pw_i} = 0.313 - 8.101 gpop_i + 0.289 s_{k,i} + 0.047 educ_i$$

(d) It can be done with R^2

$$\begin{aligned}
F &= \frac{(R_U^2 - R_R^2)/J}{(1-R_U^2)/(n-K)} \\
&= \frac{(0.845 - 0.775)/4}{(1-0.845)/(104-8)} \\
&= 10.84 > 2.47
\end{aligned}$$

So we can reject the null that the two types of countries share the same equation.

(e) There are three interaction terms, so the degrees of freedom are 3, 96. The F statistic is smaller than the critical value, so we cannot reject the null. (Thus the significance comes mainly from the difference in intercept term.)

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 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

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 $preclass$.

Computer Question

Optional – No need to hand in, but you are welcome to try.

(a)

```
. regress growth rgdp60 tradeshare yearsschool rev_coups assasinations

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  5,    58) =    4.76
       Model |  60.4973376     5  12.0994675           Prob > F      =  0.0010
    Residual |  147.310822    58  2.53984176           R-squared     =  0.2911
-------------+------------------------------           Adj R-squared =  0.2300
       Total |   207.80816    63  3.29854222           Root MSE      =  1.5937

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      rgdp60 |  -.0004613   .0001508    -3.06   0.003    -.0007631   -.0001594
  tradeshare |   1.340819   .9600631     1.40   0.168    -.5809558    3.262594
 yearsschool |   .5642445   .1431131     3.94   0.000     .2777726    .8507165
   rev_coups |  -2.150426    1.11859    -1.92   0.059    -4.389527    .0886756
assasinati~s |   .3225844   .4880043     0.66   0.511    -.6542624    1.299431
       _cons |   .6268915    .783028     0.80   0.427    -.9405093    2.194292
------------------------------------------------------------------------------
```

Per unit (PPP adjusted USD) increase in real gdp in 1960 is associated to a 0.00046 percentage point lower in the average growth rate over 1960-1995. It makes sense when more advanced countries grow slower, while less advanced countries growth faster by catching up. This is statistically significant.

A increase in tradeshare from zero to 1 leads to a 1.341 percentage point higher in average growth rate. This makes sense as the developing countries usually grow faster by massive export. But it is not statistically signficant.

A year higher in average years of schooling increase the growth rate by 0.56 percentage point. This makes sense because education increases human capital and thus productivity. This is statistically significant.

One more revolution is associated to a 2.13 percentage lower in growth rate. This makes sense because the more revolution the less stable the political situation, which is harmful to investment and growth. This is significant at 10% level.

One more assasinations is associated to 0.322 percentage point higher in growth rate. This does not agree with what we expect because assasination is associated with political instability too. Though, its t-value is less than 1, so it's far from statistically significantly different from zero.

(b) Holding values of other regressors constant, the difference in growth rate is $0.56424(8-4) = 2.257\,0$.

(c)

```
. test rev_coups  assasinations

 ( 1)  rev_coups = 0
 ( 2)  assasinations = 0

       F(  2,    58) =    1.87
            Prob > F =    0.1637
```

The F test has a p-value of 0.16, which is larger than 0.10 or 0.05, so it's not statistically significant at 10% or 5% level.

(d) Other coefficients do not change, but the coefficient and standard error on rgdp are multiplied by 1000.

```
. gen rgdp60pt=rgdp60/1000

. reg  growth  rgdp60pt tradeshare yearsschool rev_coups assasinations

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  5,    58) =    4.76
       Model |  60.4973365     5  12.0994673           Prob > F      =  0.0010
    Residual |  147.310823    58  2.53984178           R-squared     =  0.2911
-------------+------------------------------           Adj R-squared =  0.2300
       Total |   207.80816    63  3.29854222           Root MSE      =  1.5937

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rgdp60pt |  -.4612892   .1508007    -3.06   0.003    -.7631497   -.1594288
  tradeshare |   1.340819   .9600631     1.40   0.168    -.5809559    3.262594
 yearsschool |   .5642445   .1431131     3.94   0.000     .2777726    .8507164
   rev_coups |  -2.150426    1.11859    -1.92   0.059    -4.389527    .0886756
assasinati~s |   .3225844   .4880043     0.66   0.511    -.6542624    1.299431
       _cons |   .6268915    .783028     0.80   0.427    -.9405093    2.194292
------------------------------------------------------------------------------
```

(e) (I have used part (d), but they are mostly the same)

```
. reg  growth  rgdp60pt tradeshare yearsschool yearsschoolsq rev_coups assasinations

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  6,    57) =    6.74
       Model |  86.2329295     6  14.3721549           Prob > F      =  0.0000
    Residual |   121.57523    57  2.13289878           R-squared     =  0.4150
-------------+------------------------------           Adj R-squared =  0.3534
       Total |   207.80816    63  3.29854222           Root MSE      =  1.4604

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rgdp60pt |  -.3902437   .1396981    -2.79   0.007    -.6699843   -.1105031
  tradeshare |   .8635399   .8904598     0.97   0.336    -.9195751    2.646655
 yearsschool |   1.388832   .2712044     5.12   0.000     .8457548     1.93191
yearsschoo~q |  -.0957335   .0275602    -3.47   0.001    -.1509218   -.0405452
   rev_coups |  -2.051963    1.02546    -2.00   0.050    -4.105411     .001485
assasinati~s |   .0349798   .4548039     0.08   0.939    -.8757492    .9457088
       _cons |  -.4219499   .7785019    -0.54   0.590    -1.980873    1.136973
------------------------------------------------------------------------------
```

R^2 and $\bar{R}^2$ are much higher, implying a better fit. The coefficients on other variables change slightly. The effect of initial gdp is a bit smaller in magnitude. The turning point of years of schooling is about 7.25, which means it raises fast at lower education level, but it stablizes at higher level.

(f) $(1.388832(8) - 0.0957335(8)^2) - (1.388832(4) - 0.0957335(4)^2) = 0.960\,12$. It is quite a bit smaller, probably because it takes care of the decline in effect for higher years of schooling.