UNIVERSITY OF CAMBRIDGE
FACULTY OF ECONOMICS

M.Phil. in Economics
M.Phil. in Economic Research

Subject M300 Econometric Methods

Exercise Sheet 1

1. Consider the following joint probability density

$$f_{X,Y}(x,y) = \begin{cases} (x+xy+y)/4 \text{ if } x \in [0,1], y \in [0,2] \\ 0 \text{ otherwise} \end{cases}$$

(a) What's $E(Y)$? What's $E(Y|X)$?

Solution: Marginal distribution of Y has density zero for $y \notin [0,2]$, and

$$f_Y(y) = \int_0^1 \frac{x+xy+y}{4}dx = (1+y)\frac{x^2}{8} + \frac{yx}{4}\Big|_{x=0}^1 = \frac{1+3y}{8}$$

for $y \in [0,2]$. Marginal distribution of X has density zero for $x \notin [0,1]$, and

$$f_X(x) = \int_0^2 \frac{x+xy+y}{4}dy = (x+1)\frac{y^2}{8} + \frac{xy}{4}\Big|_{y=0}^2 = \frac{1+2x}{2}.$$

The conditional density of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x+xy+y}{2+4x}.$$

Marginal expectation of Y is

$$E(Y) = \int_0^2 \frac{y+3y^2}{8}dy = \frac{4}{16} + \frac{8}{8} = \frac{5}{4}.$$

Marginal expectation of X is

$$E(X) = \int_0^1 \frac{x+2x^2}{2}dx = \frac{1}{4} + \frac{2}{6} = \frac{7}{12}.$$

Conditional expectation of Y is

$$E(Y|X) = \int_0^2 y\frac{X+Xy+y}{2+4X}dy = \frac{2X+\frac{8}{3}(X+1)}{2+4X} = \frac{7X+4}{3+6X}$$

(b) What's the BLP of Y given X? Graph the results from (a) and (b).

Solution: Recall that the bestlinear predictor of Y given X equals $\alpha+\beta X$, where

$$\alpha = E(Y) - \beta E(X),$$

and

$$\beta = \frac{Cov(Y,X)}{Var(X)}.$$

To compute Cov(Y,X) :

$$\begin{aligned} E(XY) &= E[XE(Y|X)] = E\left[\frac{4X+7X^2}{3+6X}\right] = \int_0^1 \frac{4x+7x^2}{6}dx \\ &= \frac{2}{6} + \frac{7}{18} \approx .7222 \end{aligned}$$

Therefore,

$$Cov(Y,X) = E(XY) - E(X)E(Y) \approx .7222 - \frac{7}{12}\frac{5}{4} \approx -0.0070.$$

To compute Var(X) :

$$E\left(X^2\right) = \int_0^1 \frac{x^2+2x^3}{2}dx = \frac{1}{6} + \frac{1}{4} = \frac{5}{12},$$

so that

$$Var(X) = E\left(X^2\right) - E(X)^2 = \frac{5}{12} - \frac{7^2}{12^2} \approx 0.0764.$$

Therefore,

$$\beta \approx \frac{-0.0070}{0.0764} \approx -0.0916,$$

and

$$\alpha \approx \frac{5}{4} + 0.0916 * \frac{7}{12} \approx 1.3034.$$

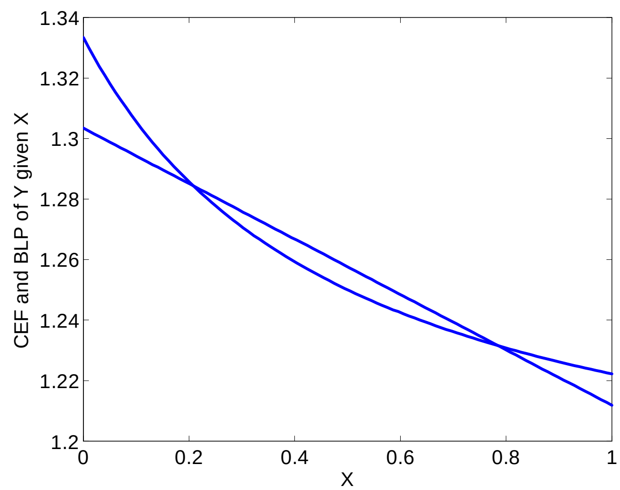
(c) What would OLS of Y on a constant and X estimate? What would OLS of Y on X (no constant) estimate?

Solution: OLS of Y on a constant and X estimates BLP:

$$BLP = 1.3034 - 0.0916 * X.$$

Note that

$$\begin{aligned} \hat{\beta}_{OLS} &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)} \\ \hat{\alpha}_{OLS} &= \bar{y} - \hat{\beta}_{OLS} * \bar{x}. \end{aligned}$$



These equations are the sample analogs of the equations defining the coefficients of the BLP. OLS of Y on X (no constant) estimates the BLP of Y when the intercept is restricted to be zero. Note that the mean squared error of such a constrained linear predictor is

$$MSE = E\left(Y - bX\right)^2.$$

The first order condition for the minimization is

$$E\left(Y - bX\right)X = 0.$$

Hence,

$$b = \frac{E\left(XY\right)}{E\left(X^2\right)} \approx \frac{0.7222}{5/12} \approx 1.7333.$$

When no constant is included in the OLS regression of Y on X, we have

$$\hat{b}_{OLS} = \frac{\sum x_i y_i}{\sum x_i^2},$$

which is the sample analog of b.

2. Suppose that Y and X are $n \times 1$ vectors of data, and the following conditions hold: (1) $Y = X\beta + e$, (2) $rank\left(X\right) = 1$, (3) $E\left(e|X\right) = 0$ and (4) $Var\left(e|X\right) = \sigma^2 I_n$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \sum_{i=1}^n Y_i$.

   (a) Consider the estimator $\tilde{\beta} = \bar{Y}/\bar{X}$. Show that $\tilde{\beta}$ is linear and conditionally unbiased. Calculate its conditional variance and compare it to the conditional variance of the OLS estimator.

Solution: $\tilde{\beta} = \frac{\bar{Y}}{\bar{X}} = \frac{1'Y}{1'X}$. Therefore $\tilde{\beta}$ is linear in y

$$E\left(\tilde{\beta}|X\right) = \frac{1'E\left(Y|X\right)}{1'X} = \frac{1'X\beta}{1'X} = \beta$$

Hence, $\tilde{\beta}$ is conditionally (and unconditionally) unbiased. For the conditional variance, we have

$$Var\left(\tilde{\beta}|X\right) = \frac{1'Var\left(Y|X\right)1}{\left(1'X\right)^2} = \frac{\sigma^2 n}{(1'X)^2}$$

Let $\hat{\beta}_{OLS} = \frac{X'Y}{X'X}$ be the OLS estimator. Then

$$Var\left(\hat{\beta}|X\right) = \frac{X'Var\left(Y|X\right)X}{\left(X'X\right)^2} = \frac{\sigma^2}{(X'X)}.$$

Since

$$X'X = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2 \geq n\bar{x}^2 = \frac{(1'X)^2}{n}$$

we have

$$\frac{\sigma^2}{(X'X)} \leq \frac{\sigma^2 n}{(1'X)^2}$$

and therefore

$$Var\left(\hat{\beta}|X\right) \leq Var\left(\tilde{\beta}|X\right).$$

This is consistent with the Gauss-Markov theorem, which says that OLS is BLUE (note that the conditions of the Gauss-Markov theorem are satisfied in this example).

(b) Suppose that you decide to use the first m ($< n$) observations and do OLS. Show that this estimator $\hat{\beta}_{(m)}$ is linear and conditionally unbiased, but not minimum conditional variance.

Solution: $\hat{\beta}_m = \frac{X'_m y_m}{X'_m X_m}$. therefore $\hat{\beta}_m$ is linear in y. Further

$$E\left(\hat{\beta}_m|X\right) = \frac{X'_m E\left(Y_m|X\right))}{X'_m X_m} = \frac{X'_m X_m \beta}{X'_m X_m} = \beta$$

Hence, $\hat{\beta}_m$ is conditionally unbiased.

$$Var\left(\hat{\beta}_m|X\right) = \frac{X'_m Var\left(Y_m|X\right)X_m}{\left(X'_m X_m\right)^2} = \frac{\sigma^2}{X'_m X_m}$$

Since $X'X = \sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{m} x_i^2 = X'_m X_m$, we have

$$\frac{\sigma^2}{(X'X)} \leq \frac{\sigma^2}{X'_m X_m}$$

and therefore

$$Var\left(\hat{\beta}|X\right) \leq Var\left(\hat{\beta}_m|X\right)$$

so that $\hat{\beta}_m$ is not of minimum variance.

(c) Could you suggest a minimum conditional variance, linear estimator (not necessarily unbiased)?

Solution: Any constant has zero variance and is a minimum variance estimator. It is biased (and rather silly) though.

3. STATA file problemset1.dta contains data from 1990 cross-section of the NLSY (National Longitudinal Survey of Youth). The file contains wage (variable w0), education (variable ed0), and age (variable a0) variables for 392 individuals. Create variables lwage=log(w0), educ=ed0, and age=a0.

(a) Regress lwage on the dummies for all possible combinations of values of educ and age using command *xi: regress lwage i.educ*i.age* (executing this command will automatically create the dummies). This is an example of saturated regression. Why do you think STATA omits many dummy variables from the regression?

Solution: See attached log file. STATA omits many dummy variables because there are no observations where these dummies equal 1. Hence, the dummy variables are identically equal to zero, which creates perfect multicollinearity.

(b) Consider the hypothesis that the conditional expectation function E (lwage|educ,age) is linear in educ, age and $(\text{age})^2$. That is, $E\,(\text{lwage}|\text{educ,age}) = \beta_0 + \beta_1\text{educ}+\beta_2\text{age}+\beta_3\,(\text{age})^2$. Since the linearity imposes a constraint on the CEF, we can call the regression of lwage on constant, wage, age and $(\text{age})^2$ "restricted" regression. What is the "unrestricted" regression? Explain.

Solution: The unrestricted version is the saturated regression. The CEF of dependent variable given explanatory variables in the saturated regression is always linear. In this regression, the coefficient on the dummy variable that equals one iff educ=v_1 and age=v_2 (where v_1 and v_2 are any possible values of education and age) is simply $E\,(\text{lwage}|\text{educ}=v_1,\text{age}=v_2)$. The relationship between $E\,(\text{lwage}|\text{educ}=v_1,\text{age}=v_2)$ and $E\,(\text{lwage}|\text{educ}=v_3,\text{age}=v_4)$ (where v_3 and v_4 are some other possible values of education and age) may be arbitrary. No constraint is imposed.

(c) Using results from the "restricted" and "unrestricted" regressions, compute the (homoskedasticity-only) F statistic to test the hypothesis from (b). Do you accept or reject the null?

Solution: Recall that

$$F = \frac{(SSR_r - SSR_u)\,/q}{SSR_u/\,(n-k)},$$

where q is the number of constraints, n is the number of observations, and k is the number of variables in the unconstrained regression. Let us count the number of constriants. In our data, there are 14 different values of education (only 13 dummies included to avoid multicollinearity). The constrained regression $E\,(\text{lwage}|\text{educ,age}) = \beta_0 + \beta_1\text{educ}+\beta_2\text{age}+\beta_3\,(\text{age})^2$ "sets" all coefficients on the 13 dummies equal to the same unspecified value, which introduces 12 constraints. Next, there are two different values of age (only one dummy is included to avoid multicollinearity). In the constrained regression, age is omitted to avoid multicollinearity with age2. All in all, there is only one coefficient corresponding to the age dummy in the saturated regression, and there is only one coefficient corresponding to age2 in the restricted regression. Hence, no additional restriction results. Finally, there are 11 included interaction dummies in the saturated regression. No interaction effects is specified in the constrained model. Effectively, this imposes 11 constraints (the coefficients on the interaction dummies are set to zero). Hence, the total number of constraints is q =12+11=23. (We could have computed this directly by subtracting "model degrees of freedom df" (given in the ANOVA outcome of the regression command) of constrained regression from that of the unconstrained one. Further, $n = 392$, and $k = 26$. Hence,

$$F \approx \frac{(221.16 - 210.37)\,/23}{210.37/\,(392 - 26)} \approx 0.82.$$

The 95% critical value of $F\,(23, 366)$ is 1.5588. We do not reject the null.

4. Consider the regression model

$$y_i = x_i'\beta + e_i, i = 1, ..., n,$$

where (1) (y_i, x_i) are i.i.d, (2) $E\,(x_i x_i')$ is non-degenerate, (3) $E\,(e_i|x_i) = 0$, and (4) $Var\,(e_i|x_i) = \sigma^2$. Assume that x_i does not contain a constant term. The corresponding uncentered coefficient of determination is defined by

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}$$

where $\hat{y}_i = x_i'\hat{\beta}$, $(i = 1, ..., n)$, with $\hat{\beta}$ the OLS estimator of β.

(a) Show that

i. $\sum_{i=1}^n \hat{y}_i^2/n = \hat{\beta}'(\sum_{i=1}^n x_i x_i'/n)\hat{\beta}$.

Solution:

$$\sum_{i=1}^{n} \hat{y}_i^2/n = \sum_{i=1}^{n}(x_i'\widehat{\beta})^2/n = \sum_{i=1}^{n}(\widehat{\beta}'x_i)(x_i'\widehat{\beta})/n = \widehat{\beta}'(\sum_{i=1}^{n} x_i x_i'/n)\widehat{\beta}.$$

ii. $\sum_{i=1}^{n} \hat{y}_i^2/n \xrightarrow{p} \beta' E[x_i x_i']\beta$.

Solution: By the Law of Large Numbers

$$\sum_{i=1}^{n} x_i x_i'/n \xrightarrow{p} E[xx'].$$

Since, OLS is consistent, $\widehat{\beta} \xrightarrow{p} \beta$. By Continuous Mapping Theorem,

$$\sum_{i=1}^{n} \hat{y}_i^2/n = \widehat{\beta}'(\sum_{i=1}^{n} x_i x_i'/n)\widehat{\beta} \xrightarrow{p} \beta' E[x_i x_i']\beta.$$

iii. $\sum_{i=1}^{n} y_i^2/n \xrightarrow{p} \sigma^2 + \beta' E[x_i x_i']\beta$.

Solution: By the Law of Large Numbers

$$\sum_{i=1}^{n} y_i^2/n \xrightarrow{p} E[y^2] = var[y]+(E[y])^2 = E[var[y|x]]+E[(E[y|x])^2] = \sigma^2+\beta' E[xx']\beta$$

(b) Hence, or otherwise, conclude that

$$R^2 \xrightarrow{p} \frac{\beta' E[x_i x_i']\beta}{\sigma^2 + \beta' E[x_i x_i']\beta}.$$

Solution: This convergence follows from (ii), (iii) and the Continuous Mapping Theorem (actually, it is sufficient to use Slutsky's theorem).

5. Consider regression

$$y_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \tag{1}$$

where the data have a group structure so that

$$E\left(e_{ig} e_{jg}\right) = \rho_e \sigma_e^2.$$

Assume an extreme inter-claster dependence $\rho_e = 1$. Further, let CEF be linear, the size of each cluster be n, and the number of clusters in the sample be G.

(a) A colleague writes (1) using matrix notations as

$$y = X\beta + e.$$

Explain what y, X, β, and e are.

Solution:

$$y = \left(y_{11}, y_{21}, ..., y_{n1}, y_{21}, ..., y_{n2}, ..., y_{G1}, ..., y_{Gn}\right)',$$

$$X = \begin{pmatrix} 1 & 1 & ... & 1 & 1 & ... & 1 & ... & 1 & ... & 1 \\ x_1 & x_1 & ... & x_1 & x_2 & ... & x_2 & ... & x_G & ... & x_G \end{pmatrix}',$$

$$\beta = (\beta_0, \beta_1)',$$

and

$$e = (e_{11}, e_{21}, ..., e_{n1}, e_{21}, ..., e_{n2}, ..., e_{G1}, ..., e_{Gn})'.$$

(b) Let $\mathbf{1}_n = \underbrace{(1, 1, ..., 1)}_{n \text{ times}}'$. Show that $E(ee'|X) = \sigma_e^2 \text{diag}\underbrace{(\mathbf{1}_n \mathbf{1}_n', ..., \mathbf{1}_n \mathbf{1}_n')}_{G \text{ times}}$ (the latter block-diagonal matrix is also denoted as $I_G \otimes \mathbf{1}_n \mathbf{1}_n'$, where $\otimes$ is the Kronecker product)

Solution: The elements of $E(ee'|X)$ outside the $n \times n$ diagonal blocks are zero assuing e_{ig_1} and e_{jg_2} are uncorrelated for $g_1 \neq g_2$. The elements in the $g-$th $n \times n$ block are equal to $E(e_{ig} e_{jg}) = \rho_e \sigma_e^2 = \sigma_e^2$. Hence, the block looks like follows

$$g\text{-th block} = \begin{pmatrix} \sigma_e^2 & \sigma_e^2 & ... & \sigma_e^2 \\ \sigma_e^2 & \sigma_e^2 & ... & \sigma_e^2 \\ ... & ... & ... & ... \\ \sigma_e^2 & \sigma_e^2 & ... & \sigma_e^2 \end{pmatrix} = \sigma_e^2 \mathbf{1}_n \mathbf{1}_n'.$$

(c) Show that $E(X'ee'X|X) = \sigma_e^2 n X'X$

Solution:

$$E(X'ee'X|X) = X'\sigma_e^2 \text{diag}\underbrace{(\mathbf{1}_n \mathbf{1}_n', ..., \mathbf{1}_n \mathbf{1}_n')}_{G \text{ times}} X.$$

On the other hand,

$$X = \begin{pmatrix} \mathbf{1}_n' & ... & \mathbf{1}_n' \\ x_1 \mathbf{1}_n' & ... & x_G \mathbf{1}_n' \end{pmatrix}'.$$

Hence,

$$\begin{aligned} E(X'ee'X|X) &= \sigma_e^2 \begin{pmatrix} \sum_{g=1}^G \mathbf{1}_n' \mathbf{1}_n \mathbf{1}_n' \mathbf{1}_n & \sum_{g=1}^G x_g \mathbf{1}_n' \mathbf{1}_n \mathbf{1}_n' \mathbf{1}_n \\ \sum_{g=1}^G x_g \mathbf{1}_n' \mathbf{1}_n \mathbf{1}_n' \mathbf{1}_n & \sum_{g=1}^G x_g^2 \mathbf{1}_n' \mathbf{1}_n \mathbf{1}_n' \mathbf{1}_n \end{pmatrix} \\ &= \sigma_e^2 n^2 \begin{pmatrix} 1 & \sum_{i=1}^G x_i \\ \sum_{i=1}^G x_i & \sum_{i=1}^G x_i^2 \end{pmatrix} = \sigma_e^2 n X'X \end{aligned}$$

(d) Prove that the square of the Moulton factor, $Var\left(\hat{\beta}_1\right)/Var_{\text{hom}}\left(\hat{\beta}_1\right)$, equals n, which is consistent with the general formula $Var\left(\hat{\beta}_1\right)/Var_{\text{hom}}\left(\hat{\beta}_1\right) = 1 + (n-1)\rho_e$.

Solution: In this problem, we are conditioning on X. I am omitting the conditioning notation, and writing $Var\left(\hat{\beta}_1\right)$ instead of $Var\left(\hat{\beta}_1|X\right)$. We have

$$\begin{aligned} Var\left(\hat{\beta}_1\right) &= Var\left((X'X)^{-1} X'e\right) = (X'X)^{-1} E(X'ee'X)(X'X)^{-1} \\ &= (X'X)^{-1} \sigma_e^2 n X'X (X'X)^{-1} = \sigma_e^2 n (X'X)^{-1}. \end{aligned}$$

Now,

$$Var_{\text{hom}}\left(\hat{\beta}_1\right) = \sigma_e^2 \left(X'X\right)^{-1}$$

Therefore,

$$Var\left(\hat{\beta}_1\right) / Var_{\text{hom}}\left(\hat{\beta}_1\right) = n.$$

6. Consider a poor elderly population that uses emergency rooms for primary care. Let y_i measure health of the i-th randomly selected person, and let d_i be the dummy that equals 1 iff the person is admitted to the hospital. Let the potential outcomes y_{1i} and y_{0i} be defined as

$$y_i = \begin{cases} y_{1i} \text{ if } d_i = 1 \\ y_{0i} \text{ if } d_i = 0 \end{cases} .$$

(a) Explain in words what is the meaning of $E\left(y_{0i}|d_i = 1\right)$ and of $E\left(y_{1i}|d_i = 0\right)$.

Solution: $E\left(y_{0i}|d_i = 1\right)$ is the average of what the health of all those admitted to the hospital would have been, had they been not admitted. Similarly, $E\left(y_{1i}|d_i = 0\right)$ is the average of what health of all those not admitted to the hospital would have been, had they been admitted.

(b) Why it may be the case that $E\left(y_{0i}|d_i = 1\right) \neq E\left(y_{0i}|d_i = 0\right)$?

Solution: d_i is not independent of the potential health outcomes. Those who are more likely having problems if not admitted (low y_{0i}) have a higher chance of being admitted.

(c) What is the likely sign of $E\left(y_{0i}|d_i = 1\right) - E\left(y_{0i}|d_i = 0\right)$?

Solution: Negative (see (b)).

(d) Is the slope of the population regression of y_i on d_i larger or smaller than the causal effect of hospitalization, assuming this effect is the same for everybody?

Solution: Smaller. Negative selection bias.

7. In section 3.2.3 of Angrist and Pischke's textbook "Mostly Harmless Econometrics", there is a discussion of the causal effect of schooling on earnings. The textbook claims that including a white-collar occupational dummy, w_i, into the regression of earnings y_i on schooling s_i is an example of using a "bad control" variable.

(a) Suppose that the schooling was randomly assigned to people, and let $E\left(y_i|s_i\right) = \beta_1 + \beta_2 s_i$. Further, assume that the causal effect of the change of s_i from s to $s+1$ is the same for everybody, and equals ρ. Show that $\beta_2 = \rho$.

Solution: We have

$$E\left(y_i|s_i = s+1\right) = \beta_1 + \beta_2\left(s+1\right),$$

and

$$E(y_i|s_i = s) = \beta_1 + \beta_2 s$$

Therefore,

$$\begin{aligned}
\beta_2 &= E(y_i|s_i = s+1) - E(y_i|s_i = s) \\
&= E(y_{s+1,i}|s_i = s+1) - E(y_{s,i}|s_i = s) \\
&= E(y_{s+1,i}|s_i = s+1) - E(y_{s,i}|s_i = s+1) \\
&\quad E(y_{s,i}|s_i = s+1) - E(y_{s,i}|s_i = s) \\
&= \rho + [E(y_{s,i}|s_i = s+1) - E(y_{s,i}|s_i = s)]
\end{aligned}$$

Since schooling is randomly assigned (by assumption), it is independent of the potential outcomes. Therefore,

$$E(y_{s,i}|s_i = s+1) - E(y_{s,i}|s_i = s) = 0$$

and

$$\beta_2 = \rho.$$

(b) In addition to the assumptions made in (a), assume that $E(y_i|s_i, w_i) = \gamma_1 + \gamma_2 s_i + \gamma_3 w_i$. Show that

$$\gamma_2 = \rho + E(y_{si}|s_i = s+1, w_i) - E(y_{si}|s_i = s, w_i).$$

Solution: We have

$$E(y_i|s_i = s+1, w_i) = \gamma_1 + \gamma_2 (s+1) + \gamma_3 w_i$$

and

$$E(y_i|s_i = s, w_i) = \gamma_1 + \gamma_2 s + \gamma_3 w_i.$$

Subtracting the second equation from the first, we get

$$\begin{aligned}
\gamma_2 &= E(y_i|s_i = s+1, w_i) - E(y_i|s_i = s, w_i) \\
&= E(y_{s+1,i}|s_i = s+1, w_i) - E(y_{s,i}|s_i = s, w_i) \\
&= E(y_{s+1,i}|s_i = s+1, w_i) - E(y_{s,i}|s_i = s+1, w_i) \\
&\quad + E(y_{s,i}|s_i = s+1, w_i) - E(y_{s,i}|s_i = s, w_i) \\
&= \rho + E(y_{s,i}|s_i = s+1, w_i) - E(y_{s,i}|s_i = s, w_i).
\end{aligned}$$

(c) Since s_i is independent from y_{si}, we have $E(y_{si}|s_i) = E(y_{si})$. Explain in words, how is this possible that, despite the latter equality, $E(y_{si}|s_i, w_i) \neq E(y_{si}|w_i)$ in general, so that the selection bias $E(y_{si}|s_i = s+1, w_i) - E(y_{si}|s_i = s, w_i)$ is not equal zero.

Solution: Even though s_i is independent from y_{si}, it is statistically dependent on w_i (because the schooling affects the future choice of the occupation). The occupational dummy, in its turn, may depend on the potential outcomes. Hence, even though $E(y_{si}|s_i) = E(y_{si})$, we may have $E(y_{si}|s_i, w_i) \neq E(y_{si}|w_i)$.

(d) Based on your analysis in (c), explain which variables can be called "bad controls" in an experimental setting.

Solution: Those variables that are affected by the treatment, and cannot be thought as fixed at the time of the treatment.

8. In the setting of problem (5), let g_i be a variable that equals 1 if the i-th randomly chosen person is a woman, and -1 if the person is a man. Suppose that $E(y_i|s_i, g_i) = \delta_1 + \delta_2 s_i + \delta_3 g_i$.

(a) Argue that, if s_i is randomly assigned, then $E(y_{si}|s_i = s+1, g_i) - E(y_{si}|s_i = s, g_i) = 0$, so that there is no selection bias, and $\delta_2 = \rho$.

Solution: Note that if s_i is randomly assigned, it is independent from g_i and y_{si}. Therefore,

$$E(y_{si}|s_i = s+1, g_i) = E(y_{si}|g_i),$$

and similarly

$$E(y_{si}|s_i = s, g_i) = E(y_{si}|g_i).$$

Hence, there is no selection bias.

(b) Suppose that you have a sample of (y_i, s_i, g_i) of size n. Assume that $\sum_{i=1}^n g_i = 0$ (the same number of men and women in the sample), and $\sum_{i=1}^n g_i s_i = 0$ (average education level is the same for men and women in the sample). Show that the OLS estimate of the coefficient on s_i in the "short regression" of y_i on constant and s_i is the same as the OLS estimate of the coefficient on s_i in the "long regression" of y_i on constant, s_i, and g_i.

Solution: Consider a short regression $y_i = \beta_1 + \beta_2 s_i + e_i$.

$$\hat{\beta}_{OLS} = \begin{pmatrix} n & \sum s_i \\ \sum s_i & \sum s_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum s_i y_i \end{pmatrix}.$$

Now, consider the long regression $y_i = \delta_1 + \delta_2 s_i + \delta_3 g_i + u_i$. We have

$$\begin{aligned} \hat{\delta}_{OLS} &= (X'X)^{-1} X'Y \\ &= \begin{pmatrix} n & \sum s_i & \sum g_i \\ \sum s_i & \sum s_i^2 & \sum s_i g_i \\ \sum g_i & \sum s_i g_i & \sum g_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum s_i y_i \\ \sum g_i y_i \end{pmatrix} \\ &= \begin{pmatrix} n & \sum s_i & 0 \\ \sum s_i & \sum s_i^2 & 0 \\ 0 & 0 & \sum g_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum s_i y_i \\ \sum g_i y_i \end{pmatrix} \\ &= \begin{pmatrix} \hat{\beta}_{OLS} \\ \left(\sum g_i^2\right)^{-1} \sum g_i y_i \end{pmatrix}. \end{aligned}$$

In particular, $\hat{\beta}_{2,OLS} = \hat{\delta}_{2,OLS}$.

(c) Suppose that the adjusted R^2 from the "long regression" is $\bar{R}^2_{long} = 0.2$ and that from the short regression is $\bar{R}_{short} = 0.1$. Prove that, under the assumptions made in (b), the homoskedastic standard error estimate for the OLS coefficient on s_i from the "long regression" equals $\sqrt{8/9}$ times the homoskedastic standard error estimate for the OLS coefficient on s_i from the "short regression". Hence, the precision of the estimate in the "long regression" is higher than that in the "short regression".

Solution: The estimate of $SE\left(\hat{\beta}_{2,OLS}\right)$ equals

$$\widehat{SE}\left(\hat{\beta}_{2,OLS}\right) = \sqrt{\frac{SSR_{short}}{n - k_{short}} \left(X'_{short}X_{short}\right)^{-1}_{22}},$$

where $\left(X'_{short}X_{short}\right)^{-1}_{22}$ is the element in the second row and second column of the matrix $\left(X'_{short}X_{short}\right)^{-1}$. Similarly,

$$\widehat{SE}\left(\hat{\delta}_{2,OLS}\right) = \sqrt{\frac{SSR_{long}}{n - k_{long}} \left(X'_{long}X_{long}\right)^{-1}_{22}}.$$

But, as follows from the solution to (b),

$$\left(X'_{short}X_{short}\right)^{-1}_{22} = \left(X'_{long}X_{long}\right)^{-1}_{22}.$$

Hence,

$$\frac{\widehat{SE}\left(\hat{\delta}_{2,OLS}\right)}{\widehat{SE}\left(\hat{\beta}_{2,OLS}\right)} = \sqrt{\frac{SSR_{long}}{n - k_{long}} \frac{n - k_{short}}{SSR_{short}}}.$$

Now, recall that an adjusted R^2 equals

$$\bar{R}^2 = 1 - \frac{n}{n-k}\frac{SSR}{TSS}.$$

Therefore,

$$\frac{SSR_{long}}{n - k_{long}} = \frac{TSS_{long}}{n}\left(1 - \bar{R}^2_{long}\right) = \frac{TSS_{long}}{n} 0.8$$

and

$$\frac{SSR_{short}}{n - k_{short}} = \frac{TSS_{short}}{n}\left(1 - \bar{R}^2_{short}\right) = \frac{TSS_{long}}{n} 0.9.$$

On the other hand, $TSS_{long} = TSS_{short}$ because the dependent variable in both regressions is the same. Hence,

$$\sqrt{\frac{SSR_{long}}{n - k_{long}} \frac{n - k_{short}}{SSR_{short}}} = \sqrt{\frac{8}{9}}.$$

(d) In light of your answers to (a), (b), and (c), explain in words why g_i is a "good control".

Solution: Including g_i into the regression does not introduce selection bias because g_i is fixed at the time of the treatment assigniment. On the other hand, including g_i into regression, reduces uncertainty, and leads to lower standard errors of the OLS estimates.