ECMA 31000: Solutions to Problem Set 6

Joe Hardwick

Due Nov 30 by 11:59PM

Question 1 Let $x = (1, x_1, x_2')'$, where x_1 is a scalar random variable and x_2 is a random vector. Suppose you observe an iid sample of $\{y_i, x_i\}_{i=1}^n$. Consider the model

$$y = \beta_0 + \beta_1 x_1 + x_2'\beta_2 + u; \qquad \mathrm{E}\left(u|x\right) = 0, Var\left(u|x\right) = \sigma^2.$$

a) Use the Frisch-Waugh-Lovell decomposition to find a formula for the OLS estimator of β_1. Call it $\hat{\beta}_1^{OLS}$.

ANS: Let X_2 be a matrix containing the observations of $(1, x_2')'$, and let X_1 be a column vector containing the observations of x_1. Then

$$\begin{aligned}\hat{\beta}_1^{OLS} &= \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} Y \\ &= \beta_1 + \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} U.\end{aligned}$$

b) Find the mean and variance of $\hat{\beta}_1^{OLS}$, conditional on $\{x_i\}_{i=1}^n$.

ANS: We have

$$\mathrm{E}\left(\hat{\beta}_1|X\right) = \beta_1 + \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} \mathrm{E}\left(U|X\right) = \beta_1,$$

and

$$\begin{aligned}Var\left(\hat{\beta}_1|X\right) &= \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} \mathrm{E}\left(UU'|X\right) M_{X_2} X_1 \left(X_1' M_{X_2} X_1\right)^{-1} \\ &= \sigma^2 \left(X_1' M_{X_2} X_1\right)^{-1} \\ &= \sigma^2 \left(\left(M_{X_2} X_1\right)' M_{X_2} X_1\right)^{-1} \\ &= \frac{\sigma^2}{SSR_1},\end{aligned}$$

where SSR_1 is the sum of squared residuals of a regression of x_1 on $(1, x_2)$.

c) Suppose I omit the entire vector x_2 from the regression, and estimate

$$y = b_0 + b_1 x_1 + \epsilon,$$

using OLS. Suppose $\hat{b}_1^{OLS}$ is treated as an estimator of β_1. Derive the omitted variables bias of $\hat{b}_1^{OLS}$ conditional on $\{x_i\}_{i=1}^n$. Show that $\hat{b}_1^{OLS} \overset{p}{\nrightarrow} \beta_1$ in general.

ANS: Now let c be an $n \times 1$ column vector of ones, let X_1 be a vector containing the observations of x_1, and let X_2 be a matrix containing the observations of x_2. We have

$$\begin{aligned} \hat{b}_1^{OLS} &= \left(X_1' M_c X_1\right)^{-1} X_1' M_c Y \\ &= \beta_1 + \left(X_1' M_c X_1\right)^{-1} X_1' M_c X_2 \beta_2 + \left(X_1' M_c X_1\right)^{-1} X_1' M_c U. \end{aligned}$$

The omitted variables bias is given by

$$\begin{aligned} \mathrm{E}\left(\hat{b}_1^{OLS} | X\right) - \beta_1 &= \left(X_1' M_c X_1\right)^{-1} X_1' M_c X_2 \beta_2 \\ &= \frac{\hat{Cov}\left(x_1, x_2\right)}{\hat{Var}\left(x_1\right)} \beta_2 \end{aligned}$$

We now compute the probability limit of $\hat{b}_1$. You can verify that

$$\hat{b}_1^{OLS} \overset{a.s.}{\to} \beta_1 + \frac{Cov\left(x_1, x_2\right)}{Var\left(x_1\right)} \beta_2$$

d) (Inclusion of irrelevant variables). Now suppose $\beta_2 = 0$. In this case we say x_2 is "irrelevant", because $\mathrm{E}\left(y|x\right)$ does not depend on x_2. What is the bias of $\hat{b}_1^{OLS}$? Compare the variances of $\hat{b}_1^{OLS}$ and $\hat{\beta}_1^{OLS}$ conditional on $\{x_i\}_{i=1}^n$.

ANS: If $\beta_2 = 0$, then the regression $y = b_0 + b_1 x_1 + \epsilon$ is correctly specified, and $\mathrm{E}\left(\epsilon|x\right) = 0$, $Var\left(\epsilon|x\right) = \sigma^2$. From part c), we see that the omitted variables bias is now 0. The variance of $\hat{b}_1^{OLS}$ conditional on X is given by

$$\begin{aligned} Var\left(\hat{b}_1|X\right) &= \sigma^2 \left(X_1' M_c X_1\right) \\ &= \frac{\sigma^2}{SSR_c}, \end{aligned}$$

where SSR_c is the sum of squared residuals of a regression of x_1 on a constant. We see that the variance is lower when x_2 is left out, because $SSR_c \geq SSR_1$.

e) Construct the regression

$$x_1 = \gamma_0 + x_2' \gamma_1 + v; \qquad \mathrm{E}\left(x_2 v\right) = 0, \mathrm{E}\left(v\right) = 0.$$

Show that

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{Var\left(u\right)}{Var\left(v\right)}\right).$$

Also derive the asymptotic distribution of $\hat{b}_1$ under the assumption that $\beta_2 = 0$. Hint: If you use part a), show first that $M_{X_2} X_1 = M_{X_2} V$, where X_2 is a matrix containing the observations of $(1, x_2')'$, and X_1 is a column vector containing the observations of x_1.

ANS: We could find the answer by computing the limiting distribution of $\hat{\beta}$ and extracting the $(2,2)$ entry of the asymptotic variance matrix. Here we present a solution based on FWL. Let X_2

be a matrix containing the observations of $(1, x_2')'$, and let X_1 be a column vector containing the observations of x_1. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) &= \sqrt{n}\left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} U \\
&= \left(\frac{1}{n} V' M_{X_2} V\right)^{-1} \frac{1}{\sqrt{n}} \left(V' M_{X_2} U\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} v_i^2 - \frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}' \left(\frac{1}{n}\sum_{i=1}^{n} x_{i2} x_{i2}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}\right)^{-1} \\
&\quad \times \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i u_i - \frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}' \left(\frac{1}{n}\sum_{i=1}^{n} x_{i2} x_{i2}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{i2} u_i\right].
\end{aligned}$$

Now note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}' \left(\frac{1}{n}\sum_{i=1}^{n} x_{i2} x_{i2}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{i2} u_i &= o_p(1) \cdot O_p(1) = o_p(1); \\
\left(\frac{1}{n}\sum_{i=1}^{n} v_i^2 - \frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}' \left(\frac{1}{n}\sum_{i=1}^{n} x_{i2} x_{i2}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} v_i x_{i2}\right)^{-1} &\xrightarrow{a.s.} \mathrm{E}\left(v_i^2\right)^{-1} = \frac{1}{Var(v)}; \\
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i u_i &\xrightarrow{d} \mathcal{N}\left(0, \mathrm{E}\left(u^2 v^2\right)\right).
\end{aligned}$$

and

$$\mathrm{E}\left(u^2 v^2\right) = \mathrm{E}\left(v^2 \mathrm{E}\left(u^2 | x\right)\right) = Var(u)\, Var(v).$$

It follows that

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) &= \left(\frac{1}{n} V' M_{X_2} V\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i u_i + o_p(1) \\
&\xrightarrow{d} \frac{1}{Var(v)} \mathcal{N}\left(0, Var(u)\, Var(v)\right) \overset{d}{=} \mathcal{N}\left(0, \frac{Var(u)}{Var(v)}\right).
\end{aligned}$$

If we assume $\beta_2 = 0$, the regression model becomes

$$y = \beta_0 + \beta_1 x_1 + u; \qquad \mathrm{E}(u|x) = 0, Var(u|x) = \sigma^2$$

and so $(\beta_0, \beta_1) = (b_0, b_1)$. The formula for $\hat{b}_1^{OLS}$ becomes

$$\hat{b}_1 = b_1 + \left(X_1' M_c X_1\right)^{-1} X_1' M_c U.$$

Similar arguments provide

$$\sqrt{n}\left(\hat{b}_1 - b_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{Var(u)}{Var(x_1)}\right),$$

which is a smaller asymptotic variance than is obtained when including x_2 in the regression. Imposing a restriction ($\beta_2 = 0$) when it is true improves the asymptotic variance without affecting consistency.

f) Use your results in part (e) to explain why, if an irrelevant variable is included that is highly correlated with a relevant variable, it can increase the finite sample and asymptotic variance significantly. Use your answers to (c),(d) and (e) to argue that the choice to include regressors presents a tradeoff between bias and variance.

ANS: If the variables in x_2 are irrelevant, the error variance u does not decrease when they are included in the regression. The variance of v may decrease significantly, however, thus greatly increasing the (asymptotic) variance of $\hat{\beta}_1$, as can be seen by observing that

$$Var\left(x_1\right) \geq Var\left(v\right) = \min_{\gamma_0,\gamma_1} \mathrm{E}\left(x_1 - \gamma_0 - x_2'\gamma_1\right)^2,$$

and noting that the irrelevant variables may be highly correlated with the relevant variables. On the other hand, our answer to (c) shows that omitting a relevant variable causes our estimates to be biased and inconsistent. (d) and (e) show that excluding irrelevant variables lowers the variance. In summary, the choice to include regressors represents a tradeoff between the omitted variables bias caused by omitting any relevant variable, and the inflation of the variance of the estimate of the parameter of interest caused by including such a variable when it irrelevant but correlated with the regressors already in the model.

Question 2 Your goal is to estimate the average treatment effect of offering job training grants on firm productivity. You observe an iid sample of $\{y_i, d_i, x_{i2}\}_{i=1}^n$, where y_i is the productivity of firm i, x_{i2} is a vector of observable characteristics of firm i, and

$$d_i = \begin{cases} 1 & \text{if firm } i \text{ receives a job training grant,} \\ 0 & \text{otherwise.} \end{cases}$$

Define the potential outcomes y_{i0}, y_{i1} as

$$\begin{aligned} y_{i0} &= \text{scrap rate of firm } i \text{ without grant,} \\ y_{i1} &= \text{scrap rate of firm } i \text{ with grant.} \end{aligned}$$

You are able to completely randomize the allocation of grants to firms, so y_{i0}, y_{i1} are independent of d_i. Suppose $\mathrm{P}\left(d = 1\right) = p \in (0, 1)$. Suppose $\mathrm{E}\left(y_i^2\right) < \infty$.

a) Argue that without loss of generality, you can write

$$y_i = \beta_0 + \beta_1 d_i + u_i, \qquad \mathrm{E}\left(u_i|d_i\right) = 0,$$

where $\beta_1 = \mathrm{E}\left(Y_{i1} - Y_{i0}\right)$ equals the average treatment effect in the population of firms.

ANS: As in PSET 5, we have

$$y_i = d_i y_{i1} + \left(1 - d_i\right) y_{i0},$$

so WLOG (because d_i takes 2 values):

$$\begin{aligned} \mathrm{E}\left(y_i|d_i=1\right) &= \beta_0+\beta_1 \\ \mathrm{E}\left(y_i|d_i=0\right) &= \beta_0. \end{aligned}$$

Together these imply

$$y_i = \underbrace{\mathrm{E}\left(y_i|d_i=0\right)}_{\beta_0} + \left(\underbrace{\mathrm{E}\left(y_i|d_i=1\right)-\mathrm{E}\left(y_i|d_i=0\right)}_{\beta_1}\right) d_i + u_i,$$

where $\mathrm{E}\left(u_i|d_i\right)=0$. Random assignment implies that

$$\begin{aligned} \mathrm{E}\left(y_i|d_i=1\right)-\mathrm{E}\left(y_i|d_i=0\right) &= \mathrm{E}\left(y_{i1}|d_i=1\right)-\mathrm{E}\left(y_{i0}|d_i=0\right) \\ &= \mathrm{E}\left(y_{i1}-y_{i0}\right)=ATE. \end{aligned}$$

Let $x=(1,d,x_2)$ and assume $\mathrm{E}\left(xx'\right)$ exists and invertible. Suppose you add in the observable characteristics to your equation, and now you assume

$$y_i = b_0+b_1d_i+x_{i2}'b_2+v_i; \qquad \mathrm{E}\left(x_iv_i\right)=0. \tag{1}$$

Suppose also that d_i is independent of x_{i2} (this would not be the case if x_{i2} contained y_i, for example). Assume $\mathrm{E}\left(v^2xx'\right)$ exists.

b) Is it necessarily the case that $\mathrm{E}\left(u_i|d_i,x_{i2}\right)$=0 or $\mathrm{E}\left(v_i|d_i,x_{i2}\right)=0$?

ANS: No. Just because u has mean 0 conditional on d, we cannot add extra variables into the conditioning without loss of generality. Suppose in fact that

$$\mathrm{E}\left(y_i|d_i,x_{i2}\right)=a_0+a_1d_i+x_{i2}'a_2$$

for some $\left(a_0,a_1,a_2\right)$, with $a_2\neq 0$. Then

$$\mathrm{E}\left(u_i|d_i,x_{i2}\right)=a_0-\beta_0+\left(a_1-\beta_1\right)d_i+x_{i2}'a_2\neq 0.$$

We can't conclude $\mathrm{E}\left(v_i|d_i,x_{i2}\right)=0$ either, because we can't be sure the conditional mean of y is linear in x_{i2} and d_i.

c) Derive the bias of $\hat{b}_1^{OLS}$ conditional on $\{d_i,x_{i2}\}_{i=1}^n$, when considered as an estimator of β_1. Show that leaving out x_{i2} would yield an unconditionally unbiased estimator.

ANS: Let X_2 be a matrix containing the observations of $(1,x_2')'$, and let X_1 be a column vector containing the observations of d.

$$\hat{b}_1=\left(X_1'M_{X_2}X_1\right)^{-1}X_1'M_{X_2}Y$$

$$= \beta_1 + \left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} U;$$
$$\mathrm{E}\left(\hat{b}_1 | D, X\right) = \beta_1 + \underbrace{\left(X_1' M_{X_2} X_1\right)^{-1} X_1' M_{X_2} \mathrm{E}\left(U | D, X\right)}_{\text{Bias}}.$$

If instead we leave out x_{i2}, then letting c be an $n \times 1$ column vector of ones, we obtain

$$\hat{b}_1 = \left(X_1' M_c X_1\right)^{-1} X_1' M_c Y$$
$$= \beta_1 + \left(X_1' M_c X_1\right)^{-1} X_1' M_c U.$$

Since $\mathrm{E}\left(u_i | d_i\right) = 0$, we obtain

$$\mathrm{E}\left(\hat{b}_1 | D\right) - \beta_1 = \left(X_1' M_c X_1\right)^{-1} X_1' M_c \mathrm{E}\left(U | D\right) = 0.$$

The law of iterated expectation implies that $\hat{b}_1$ is also unconditionally unbiased.

d) Despite your answer to c), argue that $b_1 = \beta_1$. Show explicitly where you use the independence assumption.

ANS: From Lecture 11, consider the regression

$$d = \gamma_0 + x_2' \gamma + \epsilon; \quad \mathrm{E}\left(\epsilon\right) = 0, \mathrm{E}\left(x_2 \epsilon\right) = 0.$$

where γ represents the best linear predictor of d given x_2. Since d is independent of x_2, in fact $\mathrm{E}\left(d | x_2\right) = \mathrm{E}\left(d\right)$, so the best linear predictor is in fact the best predictor under square loss, found by setting $\gamma_0 = \mathrm{E}\left(d\right)$ and $\gamma = 0$. Let

$$\tilde{d} = d - \gamma_0 - x_2' \gamma = d - \mathrm{E}\left(d\right).$$

Then

$$b_1 = \frac{\mathrm{E}\left(\tilde{d} y\right)}{\mathrm{E}\left(\tilde{d}^2\right)} = \frac{Cov\left(d, y\right)}{Var\left(d\right)} = \beta_1.$$

e) Argue that $\hat{b}_1^{OLS} \xrightarrow{p} \beta_1$.

ANS: We know from class that if $\hat{b} = \left(\hat{b}_0^{OLS}, \hat{b}_1^{OLS}, \hat{b}_2^{OLS}\right)$ is the OLS estimator of $b = \left(b_0, b_1, b_2\right)$, then

$$\hat{b} \xrightarrow{a.s.} \mathrm{E}\left(x x'\right)^{-1} \mathrm{E}\left(x y\right) = b,$$

so $\hat{b}_1^{OLS} \xrightarrow{a.s.} b_1 = \beta_1$, using part d).

f) Show that

$$\sqrt{n}\left(\hat{b}_1 - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}\left[\left(d_i - p\right)^2 \left(y_i - b_0 - b_1 d_i - x_{i2}' b_2\right)^2\right]}{p^2 \left(1 - p\right)^2}\right).$$

ANS: Repeating work from earlier, but now without simplifying (since we have not assumed homoskedasticity) we know that if we model

$$d = \gamma_0 + x_2'\gamma + \epsilon; \quad \mathrm{E}(\epsilon) = 0, \mathrm{E}(x_2\epsilon) = 0.$$

we obtain

$$\sqrt{n}\left(\hat{b}_1 - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}\left(v^2\epsilon^2\right)}{Var\left(\epsilon\right)^2}\right).$$

Since d_i is independent of x_i, it follows that $\mathrm{E}(d_i|x_i) = \mathrm{E}(d_i)$, so $\gamma = 0$ and $\epsilon = d_i - \mathrm{E}(d_i)$. Therefore,

$$Var(\epsilon) = \left[\mathrm{E}\left(d_i - \mathrm{E}\left(d_i\right)\right)^2\right] = Var\left(d_i\right) = p\left(1-p\right),$$

so the result follows.

g) Specialize your result in f) by assuming $\mathrm{E}(v|d,x) = 0$ and $Var(v|d,x) = Var(v)$. Argue that adding in covariates (as long as they are independent of d) weakly lowers the asymptotic variance. Does it lower the finite sample variance conditional on $\{d_i, x_i\}_{i=1}^n$?

ANS: The asymptotic distribution under homoskedasticity becomes

$$\sqrt{n}\left(\hat{b}_1 - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}\left[\left(y_i - b_0 - b_1 d_i - b_2 x_i\right)^2\right]}{p\left(1-p\right)}\right),$$

which is clearly decreasing as more regressors are added, since the best linear predictor minimizes $\mathrm{E}\left[\left(y_i - b_0 - b_1 d_i - b_2 x_i\right)^2\right]$, which can always be attained by setting the coefficient on additional regressors to 0, and perhaps lowered further if they are not 0. Under these additional assumptions, the OLS estimator is now unbiased, since

$$\mathrm{E}(V|D,X) = 0.$$

To show that the finite sample conditional variance may increase, consider the case where $b_2 = 0$. Then

$$y_i = b_0 + b_1 d_i + v_i; \quad \mathrm{E}(v_i|d_i,x_i) = 0, Var(v_i|d_i,x_i) = Var(v).$$

The conditional variance if x_i is included is given by

$$Var\left(\hat{b}_1|D,X\right) = \sigma^2\left(X_1' M_{X_2} X_1\right)^{-1}.$$

The conditional variance if x_i is excluded is

$$Var\left(\hat{b}_1|D,X\right) = \sigma^2\left(X_1' M_c X_1\right)^{-1}.$$

We know that

$$\sigma^2\left(X_1' M_{X_2} X_1\right)^{-1} \geq \sigma^2\left(X_1' M_c X_1\right)^{-1}$$

where the inequality will usually be strict, since in sample correlation between d, x will not typically be zero even though they are independent. So, including irrelevant regressors can increase finite sample variance even if asymptotically the variance is unaffected.

Question 3 (Efficient estimation when the form of heteroskedasticity is known) Let $\{y_i, x_i\}_{i=1}^n$ be an iid sample of observations of y, x, where y is a scalar random variable and $x \in \mathbb{R}^k$, and suppose

$$y_i = x_i'\beta + u; \qquad \mathrm{E}\left(u_i|x_i\right) = 0, Var\left(u_i|x_i\right) = \sigma\left(x_i\right)^2.$$

In class we proved the Gauss-Markov theorem, in the case that $\sigma\left(x_i\right)^2 = \sigma^2$. In this question we find the best linear unbiased estimator when the function $\sigma\left(\cdot\right)$ is known and $\sigma^2\left(x\right) > 0$ for all values of x. Stack the observations to form a matrix representation of the model $Y = X\beta + U$. Let

$$Var\left(U|X\right) = \begin{pmatrix} \sigma\left(x_1\right)^2 & 0 & 0 & 0 \\ 0 & \sigma\left(x_2\right)^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma\left(x_n\right)^2 \end{pmatrix} := \Omega.$$

a) Pre-multiply this model by $\Omega^{-1/2}$ to obtain a transformed model

$$Y^* = X^*\bar{\beta} + U^*.$$

What are Y^*, X^*, U^* and $\bar{\beta}$? What is the variance covariance matrix of the error term in your newly transformed model?

ANS: We obtain

$$\begin{aligned} \Omega^{-1/2}Y &= \Omega^{-1/2}X\beta + \Omega^{-1/2}U, \text{ or} \\ Y^* &= X^*\bar{\beta} + U^*, \end{aligned}$$

where $Y^* = \Omega^{-1/2}Y$, $X^* = \Omega^{-1/2}X$, $\bar{\beta} = \beta$ and $U^* = \Omega^{-1/2}U$. We have

$$\mathrm{E}\left(U^*|X^*\right) = \mathrm{E}\left(\mathrm{E}\left(U^*|X\right)|X^*\right) = \mathrm{E}\left(\Omega^{-1/2}\mathrm{E}\left(U|X\right)|X^*\right) = 0,$$

and

$$\begin{aligned} Var\left(U^*|X^*\right) &= \mathrm{E}\left(U^*\left(U^*\right)'|X^*\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(U^*\left(U^*\right)'|X\right)|X^*\right) \\ &= \mathrm{E}\left(\Omega^{-1/2}Var\left(U|X\right)\Omega^{-1/2}|X^*\right) \\ &= I_n. \end{aligned}$$

b) What is the Best Linear Unbiased Estimator of $\bar{\beta}$ in your newly transformed model, conditional on X^*?

ANS: Since the errors have mean 0 conditional on X^* and are homoskedastic, by the Gauss Markov Theorem, the BLUE of $\bar{\beta}$ is the OLS estimator:

$$\hat{\bar{\beta}} = \left((X^*)' X^*\right)^{-1} (X^*)' Y^* = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y.$$

c) What is the Best Linear Unbiased Estimator of β in your original model, conditional on X? Hint: Show that the estimator from part b) is a linear and unbiased estimator of β.

ANS: We propose that $\hat{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$ is in fact the BLUE of β. $\hat{\beta}$ is called the generalized least squares (GLS) estimator of β. It is linear in Y, with $A(X) = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}$, unbiased (conditionally and unconditionally), and

$$\begin{aligned} Var\left(\hat{\beta}|X\right) &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} Var\left(U|X\right)\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1} \\ &= \left(X'\Omega^{-1}X\right)^{-1}. \end{aligned}$$

For any other linear unbiased estimator $\tilde{\beta} = AY$, for some matrix $A\left(X\right)$ depending only on X, we have $AX = I$ (since $\tilde{\beta}$ is conditionally unbiased), and

$$Var\left(\tilde{\beta}|X\right) = A\Omega A'.$$

Defining $C = A - \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}$ yields $CX = 0$, so

$$\begin{aligned} A\Omega A' - \left(X'\Omega^{-1}X\right)^{-1} &= \left(C + \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\right)\Omega\left(C + \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\right)' \\ &\quad - \left(X'\Omega^{-1}X\right)^{-1} \\ &= C\Omega C' \end{aligned}$$

which is positive semi-definite. It follows that the variance of any other linear unbiased estimator is weakly greater than that of the GLS estimator.

Question 4 (Ridge regression) Let $\{y_i, x_i\}_{i=1}^n$ be an iid sample of observations of y, x, where y is a scalar random variable and $x \in \mathbb{R}^k$, and suppose

$$y_i = x_i'\beta + u; \qquad \mathrm{E}\left(u_i|x_i\right) = 0, Var\left(u_i|x_i\right) = \sigma^2.$$

Suppose $\mathrm{E}\left(xx'\right) < \infty$ and is invertible. Suppose you use the "ridge regression" estimator of β:

$$\hat{\beta} = \left(\sum_{i=1}^n x_i x_i' + \lambda I_k\right)^{-1} \sum_{i=1}^n x_i y_i,$$

where $\lambda > 0$ is a constant and I_k is the $(k \times k)$ identity matrix.

a) Show that

$$\sum_{i=1}^{n} x_i x_i' + \lambda I_k$$

is always invertible.

ANS: Let $c \in \mathbb{R}^k \backslash \{0\}$ and note that

$$\begin{aligned} c' \left(\sum_{i=1}^{n} x_i x_i' + \lambda I_k \right) c &= c' \left(\sum_{i=1}^{n} x_i x_i' \right) c + \lambda c' c \\ &\geq 0 + \lambda \left\| c \right\|^2 \\ &> 0, \end{aligned}$$

where the first inequality follows because $\sum_{i=1}^{n} x_i x_i'$ is positive semidefinite, and the second because $\lambda > 0$ and $c \neq 0$. It follows that $\left(\sum_{i=1}^{n} x_i x_i' + \lambda I_k\right) c \neq 0$, so $\sum_{i=1}^{n} x_i x_i' + \lambda I_k$ is full rank and hence invertible.

b) Is $\hat{\beta}$ a consistent estimator of β? Is it unbiased?

ANS: We have

$$\begin{aligned} \hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' + \frac{\lambda}{n} I_k \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' + \frac{\lambda}{n} I_k \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \beta \\ &+ \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' + \frac{\lambda}{n} I_k \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i u_i . \end{aligned}$$

Since

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} x_i x_i' &\overset{a.s.}{\to} \mathrm{E}\left(x x' \right); \\ \frac{\lambda}{n} I_k &\overset{a.s.}{\to} 0; \\ \frac{1}{n} \sum_{i=1}^{n} x_i u_i &\overset{a.s.}{\to} \mathrm{E}\left(x u \right) = 0; \end{aligned}$$

the continuous mapping theorem implies that

$$\hat{\beta} \overset{a.s.}{\to} \mathrm{E}\left(x x' \right)^{-1} \mathrm{E}\left(x x' \right) \beta + \mathrm{E}\left(x x' \right)^{-1} \mathrm{E}\left(x u \right) = \beta,$$

so $\hat{\beta}$ is consistent. $\hat{\beta}$ is typically biased, because

$$\hat{\beta} - \beta = - \left(\sum_{i=1}^{n} x_i x_i' + \lambda I_k \right)^{-1} \lambda \beta + \left(\sum_{i=1}^{n} x_i x_i' + \lambda I_k \right)^{-1} \sum_{i=1}^{n} x_i u_i,$$

so

$$\mathrm{E}\left(\hat{\beta}-\beta|X\right)=-\left(\sum_{i=1}^{n}x_ix_i'+\lambda I_k\right)^{-1}\lambda\beta.$$

c) Suppose further that $Var\left(u|x\right)=\sigma^2$. Derive the asymptotic distribution of $\hat{\beta}$.

ANS: We have

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta\right)=&-\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'+\frac{\lambda}{n}I_k\right)^{-1}\frac{\lambda}{\sqrt{n}}I_k\beta\\&+\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'+\frac{\lambda}{n}I_k\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_iu_i.\end{aligned}$$

Since

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_iu_i\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\mathrm{E}\left(xx'\right)\right);\\\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'+\frac{\lambda}{n}I_k\right)^{-1}\frac{\lambda}{\sqrt{n}}I_k\beta=o_p\left(1\right);\end{aligned}$$

we have

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\mathrm{E}\left(xx'\right)^{-1}\right).$$

Question 5 (Computational Question) Download the data PS6.csv from Canvas. The dataset contains data from a simulated experiment. Do not use packages designed for regression except to check your answers. You observe outcomes Y_i, treatment status D_i, and characteristics X_i for each of $n=1643$ individuals. You are interested in studying the average treatment effect $\mathrm{E}\left(Y_{i1}-Y_{i0}\right)$. You may assume that D_i is randomly assigned and independent of X_i. Consider the following regressions:

$$\begin{aligned}Y_i&=\beta_0+\beta_1D_i+U_i;\quad\mathrm{E}\left(U_i\right)=\mathrm{E}\left(D_iU_i\right)=0;\\Y_i&=\gamma_0+\gamma_1D_i+X_i'\gamma_2+V_i;\quad\mathrm{E}\left(V_i\right)=\mathrm{E}\left(D_iV_i\right)=0,\mathrm{E}\left(X_iV_i\right)=0.\end{aligned}$$

a) Compute the OLS estimators of (β_0,β_1) and $(\gamma_0,\gamma_1,\gamma_2)$. Are both estimators consistent for the average treatment effect? Are they unbiased?

ANS: Our work in Question 2 implies $\hat{\beta}_{OLS}$ is unbiased and consistent for the ATE by random assignment, but while $\hat{\gamma}_1$ is consistent due to random assignment and independence of D and X, it may be biased (conditional on D, X) in a finite sample. You should obtain

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 1.5832 \\ 1.5857 \end{pmatrix};\ \begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} = \begin{pmatrix} -5.75517 \\ 2.22952 \\ -0.23829 \\ 1.95474 \\ 0.32981 \end{pmatrix}.$$

b) Compute a consistent estimate of the asymptotic covariance matrix of the OLS estimates of (β_0, β_1) assuming homoskedasticity. Do the same for $(\gamma_0, \gamma_1, \gamma_2)$. What do you notice about the estimated asymptotic variance of $\hat{\beta}_1$ versus that of $\hat{\gamma}_1$? Explain in the context of Questions 1 and 2.

ANS: A consistent estimate of the asymptotic variance of $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ assuming homoskedasticity is given by:

$$A\hat{V}ar\left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}\right) = \frac{SSR}{n}\left(\frac{X'X}{n}\right)^{-1} = (n-2)\begin{pmatrix} 0.2434101 & -0.2434101 \\ -0.2434101 & 0.5907279 \end{pmatrix} = \begin{pmatrix} 399.4 & -399.4 \\ -399.4 & 969.4 \end{pmatrix}.$$

The reason the estimated covariances are the same in magnitude as the estimated variance of $\hat{\beta}_0$ is that $\hat{\beta}_0 = \bar{y}_0$, the sample average of untreated individuals, and $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$, the difference between the average for treated and untreated individuals. Note that $Cov\left(\hat{\beta}_0, \hat{\beta}_1 | D\right) = Cov\left(\bar{y}_1, \bar{y}_0 | D\right) - Var\left(\bar{y}_0 | D\right) = -Var\left(\bar{y}_0\right)$, since the units in each group are independent of each other and treatment is randomly assigned. Alternatively, note that

$$\left(X'X\right)^{-1} = \begin{pmatrix} n & \sum d_i \\ \sum d_i & \sum d_i^2 \end{pmatrix} = \frac{1}{\bar{d}\left(1-\bar{d}\right)}\begin{pmatrix} \sum d_i & -\sum d_i \\ -\sum d_i & n \end{pmatrix},$$

because $d_i^2 = d_i$. Similarly,

$$A\hat{V}ar\left(\begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix}\right) = \frac{SSR}{n}\left(\frac{X'X}{n}\right)^{-1}$$

$$= (n-5)\begin{pmatrix} 0.04973 & -0.03776 & 0.00151 & -0.00161 & -0.00235 \\ -0.03776 & 0.09282 & 0.00010 & 0.00031 & -0.00041 \\ 0.00151 & 0.00010 & 0.00055 & -0.00043 & -0.00032 \\ -0.00161 & 0.00031 & -0.00043 & 0.00106 & -0.00025 \\ -0.00235 & -0.00041 & -0.00032 & -0.00025 & 0.00096 \end{pmatrix}$$

$$= \begin{pmatrix} 81.462 & -61.854 & 2.473 & -2.637 & -3.851 \\ -61.854 & 152.045 & 0.165 & 0.515 & -0.669 \\ 2.473 & 0.165 & 0.893 & -0.699 & -0.519 \\ -2.6374 & 0.515 & -0.699 & 1.742 & -0.405 \\ -3.851 & -0.669 & -0.519 & -0.405 & 1.566 \end{pmatrix}$$

I have written the answers this way because the "lm" package in R computes the estimated variance covariance matrix using the unbiased estimate of $\hat{\sigma}^2$. You can check your answers by verifying that your matrix satisfies the above equality. The estimated variance is smaller for $\hat{\gamma}_1$ than $\hat{\beta}_1$, because adding covariates has reduced the error variance (and greatly improved in-sample fit). Since D is independent of X, adding these covariates does not inflate the finite sample variance much. (You can verify this by regressing D on X and a constant versus regressing D on just a constant). Overall we see a significant reduction in variance.

c) Drop the assumption of homoskedasticity. Now repeat part b) with an estimate of the asymptotic covariance matrix that is robust to heteroskedasticity. Do the estimates seem to be much different?

ANS: You can check your answers using the sandwich package in R to compute the "HC0" standard errors. A consistent estimate under heteroskedasticity is given by

$$A\hat{V}ar\left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}\right) = \left(\frac{X'X}{n}\right)^{-1} \frac{X'\hat{\Omega}X}{n} \left(\frac{X'X}{n}\right)^{-1} \approx \begin{pmatrix} 412.6 & -412.6 \\ -412.6 & 955.8 \end{pmatrix}$$

This is not much different from the estimate assuming homoskedasticity. For $(\gamma_0, \gamma_1, \gamma_2)$ we have a consistent estimate under heteroskedasticity given by

$$A\hat{V}ar\left(\begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix}\right) = \begin{pmatrix} 76.362 & -57.431 & 2.034 & -2.411 & -3.227 \\ -57.431 & 153.439 & 0.903 & 0.236 & -1.991 \\ 2.034 & 0.903 & 0.860 & -0.771 & -0.455 \\ -2.411 & 0.236 & -0.771 & 1.885 & -0.426 \\ -3.227 & -1.991 & -0.455 & -0.426 & 1.543 \end{pmatrix},$$

which is also not much different than the estimate assuming homoskedasticity. Overall, we notice that including the regressors has significantly decreased the standard error of the coefficient on the treatment dummy.