HOME ASSIGNMENT 2

DANIIL BUCHKO

Question (1)**.** Suppose that the population model determining y is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

and this model satisfies the Gauss-Markov assumptions. However, we estimate the model that omits x_3. Let $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2. Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$\mathbb{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{i1} x_{i3}}{\sum_{i=1} \hat{r}_{i1}^2}$$

where the $\hat{r}_{i1}$ are the OLS residuals from the regression of x_1 on x_2.

Solution.

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{1i}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i)}{\sum_{i=1}^n \hat{r}_{1i}^2} =$$

$$= \underbrace{\beta_0 \frac{\sum_{i=1}^n \hat{r}_{1i}}{\sum_{i=1}^n \hat{r}_{1i}^2}}_{0 \text{ (from FOC)}} + \beta_1 \frac{\sum_{i=1}^n \hat{r}_{1i} x_{1i}}{\sum_{i=1}^n \hat{r}_{1i}^2} + \beta_2 \underbrace{\frac{\sum_{i=1}^n \hat{r}_{1i} x_{2i}}{\sum_{i=1}^n \hat{r}_{1i}^2}}_{0 \text{ (from FOC)}} + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{1i} x_{3i}}{\sum_{i=1}^n \hat{r}_{1i}^2}$$

$$+ \frac{\sum_{i=1}^n \hat{r}_{1i} u_i}{\sum_{i=1}^n \hat{r}_{1i}^2} =$$

$$= \beta_1 \frac{\sum_{i=1}^n \hat{r}_{1i} x_{1i}}{\sum_{i=1}^n \hat{r}_{1i}^2} + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{1i} x_{3i}}{\sum_{i=1}^n \hat{r}_{1i}^2} + \frac{\sum_{i=1}^n \hat{r}_{1i} u_i}{\sum_{i=1}^n \hat{r}_{1i}^2}$$

Now, by using LIME we obtain the following:

$$\mathbb{E}(\tilde{\beta}_1) = \beta_1 \underbrace{\frac{\sum_{i=1}^n \hat{r}_{1i} x_{1i}}{\sum_{i=1}^n \hat{r}_{1i}^2}}_{\text{cant prove it is 1 :(}} + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{1i} x_{3i}}{\sum_{i=1}^n \hat{r}_{1i}^2}$$

□

Question (2)**.** Consider the simple regression model $y = \beta_0 + \beta_1 x + u$ under the first four Gauss-Markov assumptions. For some function

$g(x)$, for example $g(x) = x^2$ or $g(x) = \ln(1 + x^2)$ define $z_i = g(x_i)$. Define slope as follows:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}$$

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $\mathbb{E}(u|x) = 0$, you can treat both x_i and z_i as nonrandom in your derivation.

(ii) Add the homoskedasticity assumption, MLR.5. Show that

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\left(\sum_{i=1}^n (z_i - \bar{z}) x_i\right)^2}$$

(iii) Show directly that, under the Gauss-Markov assumptions, $\mathrm{Var}(\hat{\beta}_1) \leq \mathrm{Var}(\tilde{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator.

Solution. (i)

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} = \frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (z_i - \bar{z}) x_i} =$$
$$= \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} \implies$$

$$\mathbb{E}(\tilde{\beta}_1) = \beta_1 + \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\middle| x\right]\right] =$$
$$\beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^n (z_i - \bar{z}) \mathbb{E}\left[u_i | x\right]}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right] = \beta_1$$

(ii)

$$\mathrm{Var}(\tilde{\beta}_1) = \mathbb{E}\left[\tilde{\beta}_1^2\right] - \underbrace{\mathbb{E}^2\left[\tilde{\beta}_1\right]}_{\text{calculated}} \implies$$

$$\tilde{\beta}_1^2 = \left(\frac{\sum_{i=1}^n (z_i - \bar{z}) y_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right)^2 = \left(\frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right)^2 =$$
$$\left(\frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_1 x_i + u_i)}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right)^2 = \left(\beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right)^2 =$$
$$\beta_1^2 + 2\beta_1 \frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i} + \underbrace{\left(\frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z}) x_i}\right)^2}_{(*)}$$

$$\left(\frac{\sum_{i=1}^n (z_i-\bar{z})u_i}{\sum_{i=1}^n (z_i-\bar{z})x_i}\right)^2 = \frac{\left(\sum_{i=1}^n (z_i-\bar{z})u_i\right)^2}{\left(\sum_{i=1}^n (z_i-\bar{z})x_i\right)^2} =$$
$$\frac{\sum_{i=1}^n \sum_{m=1}^n (z_i-\bar{z})u_i(z_m-\bar{z})u_m}{\left(\sum_{i=1}^n (z_i-\bar{z})x_i\right)^2} =$$
$$\frac{\sum_{i=1}^n (z_i-\bar{z})^2 u_i^2 + \sum_{i=1}^n \sum_{m=1,m\neq i}^n (z_i-\bar{z})u_i(z_m-\bar{z})u_m}{\left(\sum_{i=1}^n (z_i-\bar{z})x_i\right)^2} = (*)$$

Thus:

$$\tilde{\beta}_1^2 = \beta_1^2 + 2\beta_1 \frac{\sum_{i=1}^n (z_i-\bar{z})u_i}{\sum_{i=1}^n (z_i-\bar{z})x_i} +$$
$$+ \frac{\sum_{i=1}^n (z_i-\bar{z})^2 u_i^2 + \sum_{i=1}^n \sum_{m=1,m\neq i}^n (z_i-\bar{z})u_i(z_m-\bar{z})u_m}{\left(\sum_{i=1}^n (z_i-\bar{z})x_i\right)^2}$$

Incorporating the fact that $\mathbb{E}[u_i u_m] = 0$ for $m \neq i$ and taking use of LIME we get that:

$$\mathbb{E}\left[\tilde{\beta}_1^2\right] = \beta_1^2 + \mathbb{E}\left[\frac{\sum_{i=1}^n (z_i-\bar{z})^2 \mathbb{E}[u_i^2|x]}{\left(\sum_{i=1}^n (z_i-\bar{z})x_i\right)^2}\right] =$$
$$\beta_1^2 + \sigma_u^2 \frac{\sum_{i=1}^n (z_i-\bar{z})^2}{\sum_{i=1}^n ((z_i-\bar{z})x_i)^2}$$

and finally using the formula for $\mathrm{Var}(\tilde{\beta}_1) = \mathbb{E}[\tilde{\beta}_1^2] - \mathbb{E}^2[\tilde{\beta}_1]$:

$$\mathrm{Var}(\tilde{\beta}_1) = \beta_1^2 + \sigma_u^2 \frac{\sum_{i=1}^n (z_i-\bar{z})^2}{\sum_{i=1}^n ((z_i-\bar{z})x_i)^2} - \beta_1^2 = \sigma_u^2 \frac{\sum_{i=1}^n (z_i-\bar{z})^2}{\sum_{i=1}^n ((z_i-\bar{z})x_i)^2}$$

□

Question (3)**.** Consider the standard linear multivariate regression model under the Gauss-Markov assumptions:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

(1) Derive the OLS estimate $\hat{\beta}_2$ in terms of sample variances and sample covariances of random variables x_1, x_2, y.
(2) How would you get the OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ by computing a series of OLS estimates from simple regressions only?
(3) Suppose you have decided to get OLS estimates for $\gamma_0, \gamma_1, \gamma_2$ from
$$x_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 y + v$$
Is $\tilde{\beta}_2 = 1/\hat{\gamma}_2$ an unbiased estimator of β_2?

Solution. (1)
(2) I would use the Frisch-Waugh theorem in the following setup:

(a) Step 1. Estimate $x_1 = \alpha_0 + \alpha_1 x_2 + r_1$ and collect residuals $\hat{r}_1$
(b) Step 2. Estimate $x_2 = \alpha_0 + \alpha_1 x_1 + r_2$ and collect residuals $\hat{r}_2$
(c) Step 3. Estimate $y = \lambda_0 + \lambda_1 \hat{r}_1 + \varepsilon$ and $y = \lambda_0 + \lambda_2 \hat{r}_2 + \varepsilon$. Coefficients $\hat{\lambda}_1$ will be an estimator of β_1 and $\hat{\lambda}_2$ for β_2

(3)

□

Question (4)**.** Consider a simple regression model through the origin under the Gauss-Markov assumptions

$$y = \gamma x + \varepsilon$$

and let

$$\tilde{\gamma} = \frac{\sum_{i=1}^n y_i x_i^3}{\sum_{i=1}^n x_i^4}$$

be an estimator of the slope parameter γ.

(1) Under what condition this estimator is well defined?
(2) Show conditional unbiasness of $\tilde{\gamma}$
(3) Derive conditional variance of the estimator.

Solution. (1)
Estimator is considered good under conditional unbiasness and consistency. Also we can see that estimator is well defined everywhere under the Gauss-Markov assumptions since the $\mathrm{Var}(X) \neq 0$, therefore $\tilde{\gamma} < \infty$.

(2) Conditional unbiasness:

$$\tilde{\gamma} = \frac{\sum_{i=1}^n y_i x_i^3}{\sum_{i=1}^n x_i^4} = \frac{\sum_{i=1}^n (\gamma x_i + \varepsilon_i) x_i^3}{\sum_{i=1}^n x_i^4} = \gamma + \frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}$$

Using LIME we obtain that:

$$\mathbb{E}(\tilde{\gamma}) = \mathbb{E}\left[\gamma + \frac{\sum_{i=1}^n x_i^3 \mathbb{E}\left[\varepsilon_i | x\right]}{\sum_{i=1}^n x_i^4}\right] = \gamma$$

(3) Conditional variance:

$$\mathrm{Var}(\tilde{\gamma}|x) = \mathbb{E}\left[\tilde{\gamma}^2 | x\right] - \underbrace{\mathbb{E}\left[\tilde{\gamma}|x\right]^2}_{\text{calculated}}$$

$$\tilde{\gamma}^2 = \left(\gamma + \frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}\right)^2 = \gamma^2 + 2\gamma \frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4} + \underbrace{\left(\frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}\right)^2}_{(*)}$$

Firstly lets do some algebra with (*):

$$\left(\frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}\right)^2 = \left(\frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}\right)\left(\frac{\sum_{i=1}^n \varepsilon_i x_i^3}{\sum_{i=1}^n x_i^4}\right) =$$
$$\frac{\sum_{i=1}^n \varepsilon_i x_i^3 \sum_{i=1}^n \varepsilon_i x_i^3}{\left(\sum_{i=1}^n x_i^4\right)^2} = \frac{\sum_{i=1}^n \sum_{m=1}^n \varepsilon_i \varepsilon_m x_i^3 x_m^3}{\left(\sum_{i=1}^n x_i^4\right)^2} =$$
$$\frac{\sum_{i=1}^n \varepsilon_i^2 x_i^6 + \sum_{i=1}^n \sum_{m=1, m\neq i}^n \varepsilon_i \varepsilon_m x_i^3 x_m^3}{\left(\sum_{i=1}^n x_i^4\right)^2} = (*)$$

Inserting back and taking mathematical expectation keeping in mind that $\mathbb{E}(\varepsilon_i, \varepsilon_j) = 0$ as a consequence of i.i.d yields:

$$\mathbb{E}\left[\tilde{\gamma}^2\right] = \mathbb{E}\left[\gamma^2 + \frac{\sum_{i=1}^n x_i^6 \mathbb{E}\left[\varepsilon_i^2 | x\right]}{\left(\sum_{i=1}^n x_i^4\right)^2}\right] = \gamma^2 + \sigma_\varepsilon^2 \frac{\sum_{i=1}^n x_i^6}{\left(\sum_{i=1}^n x_i^4\right)^2}$$

Thus conditional variance is the following:

$$\mathrm{Var}(\tilde{\gamma}|x) = \gamma^2 + \sigma_\varepsilon^2 \frac{\sum_{i=1}^n x_i^6}{\left(\sum_{i=1}^n x_i^4\right)^2} - \gamma^2 = \sigma_\varepsilon^2 \frac{\sum_{i=1}^n x_i^6}{\left(\sum_{i=1}^n x_i^4\right)^2}$$

□

Question (5)**.** Use the data in `hprice1.csv` to estimate the model

$$\text{price} = \beta_0 + \beta_1 \text{sqrtf} + \beta_2 \text{bdrms} + u$$

where price is the house price measured in thousands of dollars.

(1) Write out the results in equation form.
(2) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?
(3) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).
(4) What percentage of the variation in price is explained by square footage and number of bedrooms?
(5) The first house in the sample has sqrft $= 2,438$ and bdrms $= 4$. Find the predicted selling price for this house from the OLS regression line
(6) The actual selling price of the first house in the sample was $\$300,000$ (so price $= 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

Solution. (1) Regression result:

$$\text{price} = -19.31 + 0.13 \times \text{sqrtf} + 15.19 \times \text{bdrms}$$

(2)

$$\Delta\text{price} = \hat{\beta}_1 \Delta\text{sqrtf} + \hat{\beta}_2 \Delta\text{bdrms}$$

Since $\hat{\beta}_2 = 15.19$ thus each additional bedroom increases price by $\$15k$ (if we accept significance at 10% level).

(3)

$$\Delta\text{price} = \hat{\beta}_1 140 + \hat{\beta}_2 = 0.13 \times 140 + 15.19 = 33,200$$

(4) Since reporting adjusted $R^2 = 0.6233$ that means that roughly 62% of variation in prices is explained by factors.

(5) Using function `predict` we get that predicted selling price for this house should be \$354.6052 thousands.

(6) Actual price was lower than predicted, thus it suggests that buyer underpaid for the house. Residuals for this house are:

$$300 - 354.6 = -54.6$$

□

Question (6)**.** In this problem set you continue to explore the returns to education and the gender gap in earnings. The data set is the same as for problem set 1 (`cps99_ps1.csv`; see problem set 1 for a description of the variables). Consider the regression of average hourly earnings (*ahe*) on years of education (*yrseduc*). Consider the following three variables that have been omitted from the regression:

(1) *gender*
(2) a binary variable that = 1 if the worker's last name falls in the first half of the alphabet, and = 0 if it falls in the second half.
(3) the worker's native ability (for example *IQ* or some better measure)

For each: Will this omission arguably lead to omitted variable bias? Why or why not? If your answer is that it will, state the sign of the bias (is the effect of a year of education overestimated or underestimated?). Explain.

Solution. Remembering formula for omitted variable bias from lecture:

$$\hat{\beta}_1 \to \beta_1 + \left(\frac{\sigma_u}{\sigma_x}\right) \rho_{xu}$$

To create bias, omitted variable must be correlated with the `yrseduc`. If correlation is positive then bias would be positive and the effect of `yrseduc` will be overestimated. Vice versa is true: if correlation with omitted variable is negative then bias will be negative and the variable will be underestimated.

(1) Starting with `gender`. There is column `female` in data which is slightly correlated with `yrseduc` (coef is 0.028). Other factors fixed such correlation creates overstimation of `yrseduc`.
(2) Unfortunately no data for workers names is provided but probably there is no correlation between the years of education and the family name of the person. Thus omitting such factor will not create bias.
(3) Worker's `IQ` is probably to be correlated positively with years of education which implies existence of positive bias in β_1 estimation and thus overstimation.

□

Question (7)**.** Consider an experimental approach to estimation of the effect on earnings of education.

(1) Describe an experimental protocol for an idealized randomized controlled experiment that would permit unbiased estimation of the effect on earnings of an additional year of education at a typical US educational institution. (Ignore practical and ethical issues.)
(2) Explain why, precisely, your experimental protocol would eliminate omitted variable bias arising from omission of the variables in question 1 (or any other variables).

Solution. (1) Idealized controlled experiment's aim is to separate education effect on earnings from all other possible effets, which may be correlated with education. In order to measure influence of years of education only we need to collect all the data of personality of students and make sure our sample is as representative as possible. By representiveness i mean that we need to select students for model from all possible social groups: students from rich/middle/poor families, representatives of different minorities, students ambitions and everything that can contribute to the level of earnings through years of education.
(2) This approach is indeed going to eliminate omitted variable bias, since we clean unexplained part of the model from all possible factors, which could influence the earnings through years of education.

□

Question (8)**.** Estimate two regressions: (i) `ahe` on `yrseduc` and (ii) `ahe` on `yrseduc` and `female`.

(1) In regression (ii), what is the coefficient on `yrseduc`? Explain what this means.
(2) In regression (ii), test the hypothesis that the population coefficient on `female` in specification (ii) is zero, against the hypothesis that it is nonzero, at the 5% significance level. In everyday words (not statistical terms), what precisely is the hypothesis is that you are testing?
(3) Does the coefficient on `yrseduc` change from regressions (i) to (ii) in a substantively important way, that is, is the difference between the two estimates large in a real–world sense? What is the reason that including `female` in the regression results in a large (or only a small) change in the coefficient on `yrseduc`? (Hint: what is the correlation between `yrseduc` and `female`?)

Solution. (1) We can see in the table below that coefficient on `yrseduc` is 1.341. It means that growth of years of education is positively correlated with the average hourly earnings. Each additional year of education was correlated with the earnings growth by $1.341
(2) From the table we can see that coefficient value is -3.396 which is quite higher than its standart deviation 0.234. So coefficient is statistically different from zero, which implies that men and women have different weekly earnings, having excluded influence of years of education.
(3) Different models provided insignificantly different estimations of `yrseduc` influence on `ahe` (1.32186 vs 1.34147). Since correlation between `female` and `yrseduc` is relatively small (0.028) we didnt notice change in coefficient after estimating model (ii). Small insignificant correlation doesnt contribute to omitted variable bias and thus coefficients didnt change.

□

TABLE 1

	Dependent variable:
	ahe
yrseduc	1.341^{***}
	(0.047)
female	-3.396^{***}
	(0.234)
Constant	-1.384^{**}
	(0.646)
Observations	3,781
R^2	0.211
Adjusted R^2	0.210
Residual Std. Error	7.133 (df = 3778)
F Statistic	504.501^{***} (df = 2; 3778)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question (9)**.** Use multiple linear regression of `ahe` on `yrseduc` and `ba` (a binary variable equaling one for those who attained a bachelor's degree) to estimate the marginal value of a bachelor's degree, more specifically, the value of going to school for 16 years and receiving a bachelor's degree, relative to going to school for 16 years but not receiving a bachelor's degree. Construct a 95% confidence interval for the marginal value of a bachelor's degree.

Solution. After estimation we get the following table:

TABLE 2

	Dependent variable:
yrseduc	1.204^{***}
	(0.075)
ba	0.786^{*}
	(0.466)
Constant	-1.246
	(0.894)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that marginal effect of having bachelor degree is \$0.786. Having robust estimations of std for this coefficient we can construct 95% interval of this coefficient, namely:

TABLE 3

	2.5 %	97.5 %
(Intercept)	-3.297	0.806
yrseduc	1.037	1.371
ba	-0.133	1.705

Since `yrseduc` has interval of $(1.037, 1.371)$ we can calculate the interval for value of going to school 16 years as follows:

$$16 \times (1.037, 1.371) = (16.592, 21.936)$$

and thus if we add having bachelor degree then resulting intervals will be:

$$(16.592, 21.936) + (-0.133, 1.705) = (16.459, 23.641)$$

Interval for relative marginal value of receiving bachelor degree is simply then:

$$(0.99, 1.078)$$

□