Andrew ElHabr
ECON 7023
1/30/18

Homework 1

Wooldridge Introductory Textbook

6.4)

(i) From multivariable calculus we know that

$$\frac{\partial \log(wage)}{\partial educ} = \beta_1 + \beta_2 pareduc.$$

I would expect that the sign of β_2 to be positive since it is likely that an additional year of individual education yields a higher increase in %wages for individuals with more educated parents when experience and tenure are fixed.

(ii) The coefficient of the interaction term (0.00078) indicates that for every extra year of total parent's education, the effect of a one year increase in the individual's education on wages is exaggerated by 0.078%.

(iii) No, the estimated return to education now depends negatively on parent education since $\beta_2 = -0.0016 < 0$. However, this relationship is not significant since the 95% confidence interval of β_2, i.e. $-0.0016 \pm 1.96(0.0012) = [-0.003952, 0.000752]$, includes 0. Thus, we fail to reject the null hypothesis of $H_0 : \beta_2 = 0$, i.e. the return of education does not depend on parent education.

7.3)

(i) Yes, there is strong evidence that $hsize^2$ should be included in the model since its estimated coefficient indicates that the size of the high school graduating class has a significant diminishing effect on the combined SAT score. From this equation, we can see that the optimal high school size is

$$hsize^* = \frac{19.30}{2(2.19)} \approx 4.41$$

in hundreds.

(ii) When holding $hsize$ fixed, the estimated difference in SAT score between nonblack females and nonblack males is -45.09, which is clearly very statistically significant since the 95% confidence interval of this coefficient, i.e. $-45.09 \pm 1.96(4.29) = [-53.4984, -36.6816]$, does not come even close to including 0.

(iii) When holding $hsize$ fixed, the estimated difference in SAT score between nonblack males and black males is -169.81, which is clearly statistically significant since the 95% confidence interval of this coefficient, i.e. $-169.81 \pm 1.96(12.71) = [-194.7216, -144.8984]$, does not include 0. Thus, we reject the null hypothesis that there is no difference between the scores of nonblack and black males.

(iv) When holding $hsize$ fixed, the estimated difference in SAT score between black females and nonblack females is $-169.82 + 62.31 = -107.51$. In order to test if this difference is statistically different, we would need to compute the standard error of this difference, which we are not directly given with the estimated model.

8.5)

(i) There do not appear to be any significant differences in the two sets of standard errors.

(ii) Holding other factors fixed, increased education by four years will decrease the probability of smoking by (4)(.029) = 0.116.

(iii) Since

$$\frac{\partial \widehat{smokes}}{\partial age} = .020 - 2(.00026)age = .020 - .00052age,$$

the age at which the probability of smoking starts decreasing is $.020/.00052 \approx 38.46$.

(iv) The coefficient of *restaurn* indicates that the probability that a person smokes is .101 smaller if the person lives in a state with restaurant smoking restrictions when compared to the case of the person not living in one of these states.

(v) Using this estimated linear probability model, person number 206 smokes with probability

$$.656 - .069\log(67.44) + .012\log(6,500) - .029(16) + .020(77) - .00026(77)^2 - .101(0) - .026(0) \approx 0.110.$$

Since we know that the state of a person's smoking habits is binary, we would likely say that that this person does not smoke since his or her probability of smoking is relatively small. We see that this matches the true state of the person, who is not a smoker according to our given data.

C6.12)

(i) The youngest age of people in the sample is 25, and there are 211 people at this age.

(ii) The literal interpretation of β_2 is that it is the marginal effect that one unit of age has on net total financial assets (measured in \$1000) at the age of 0. No, it is not of much interest by itself in this model because the model includes the age^2 term.

(iii)

$$\widehat{nettfa} = \underbrace{4.680}_{(10.081)} + \underbrace{0.978}_{(0.025)} inc - \underbrace{2.231}_{(0.490)} age + \underbrace{0.038}_{(0.006)} age^2.$$

No, I am not concerned that the coefficient of *age* is negative because the coefficient of age^2 is positive, which indicates that net total financial assets is actually increasing with age, which is logical.

(iv) I estimated the model

$$netffa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$$

and got

$$\widehat{nettfa} = -\underbrace{18.896}_{(6.746)} + \underbrace{0.978}_{(0.025)} inc - \underbrace{0.345}_{(0.213)} age + \underbrace{0.038}_{(0.006)} age^2,$$

so $\hat{\theta}_2 = 0.345$ and has a t-statistic of -1.619, which indicates that it is not significant at a 5% significance level.

(v) The adjusted R^2 value of the model that does not include *age* is 0.1727, and the adjusted R^2 value of the model that does include *age* is 0.1728, so there is very little difference in terms of this goodness-of-fit metric between the two models. Thus, from a practical perspective, we would prefer to use the model without *age* as a covariate since we prefer more parsimonious models when selecting between models that explain approximately the same amount of variance in the response variable.

(vi) The graph show that net total financial assets increases quadratically with age as expected.

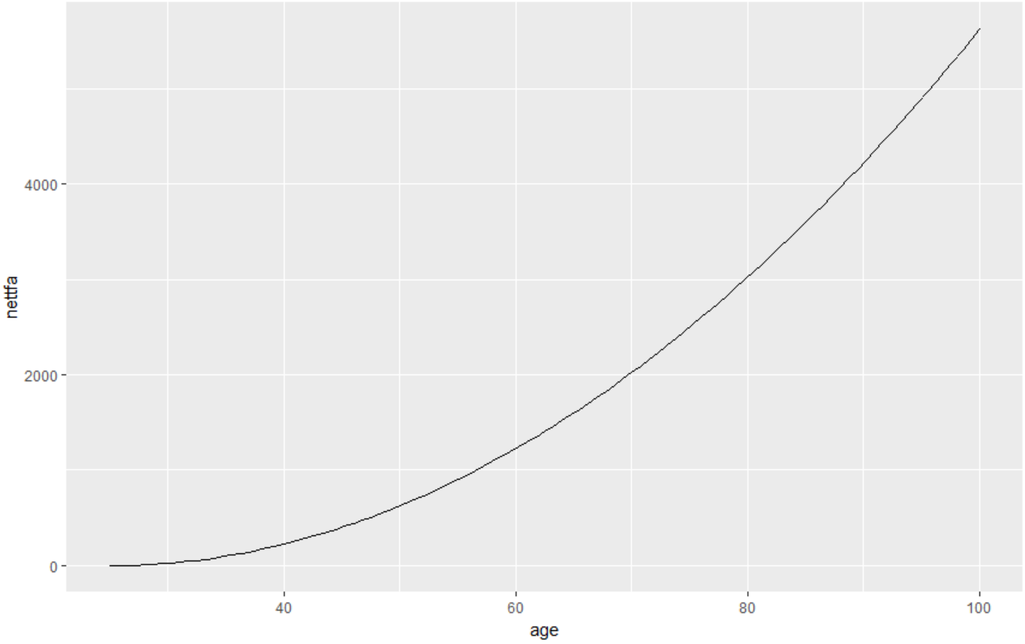

Figure 1: Net total financial assets as a function of *age* for the estimated equation in part (v) for $inc = 30$.

(vii) It does appear that we should include inc^2 in the original model since its coefficient estimate and standard error indicate that it is very significant (t-statistic of 16.561).

C7.12)

(i) Approximately 29% of men and 33% of women are described as having above average looks. More people are classified as having above average looks than below average looks. In fact, about 2.5 times as many people are rated as above average than below average.

(ii) In this case, $H_0 : \alpha_1 = 0$ and $H_1 : \alpha_1 > 0$ for the linear probability model

$$abvavg = \alpha_0 + \alpha_1 female + v.$$

The estimated model is

$$\widehat{abvavg} = \underbrace{0.290}_{(0.016)} - \underbrace{0.040}_{(0.027)} female,$$

so the t-statistic for the coefficient of $female$ is 1.477, which yields a one-sided p-value that the fraction of above average-looking women is higher than the fraction of above average-looking men of approximately 0.070.

(iii) For males, we get the estimated model

$$\log(\widehat{wages}) = \underbrace{1.884}_{(0.024)} - \underbrace{0.199}_{(0.025)} belavg - \underbrace{0.044}_{(0.042)} abvavg,$$

and for females, we get the estimated model

$$\log(\widehat{wages}) = \underbrace{1.309}_{(0.034)} - \underbrace{0.138}_{(0.076)} belavg + \underbrace{0.034}_{(0.055)} abvavg.$$

For males, the coefficient for *belavg* indicates that there is a $100|(e^{-0.199} - 1)| \approx 18\%$ reduction in wages for men that have below average looks. For females, the coefficient for *belavg* indicates that there is a $100|(e^{-0.138} - 1)| \approx 13\%$ reduction in wages for females that have below average looks. The hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 < 0$ means that we are testing only whether below average looks has a significantly negative effect on $\log(wages)$. The p-values for men and women are 0.00048 and 0.03579, respectively, which means that we will reject the null hypothesis in both cases.

(iv) There is not convincing evidence that women with above average looks earn more than those with average looks since the coefficient estimate for *abvavg* is not significant for women (t-statistic of 0.607).

(v) No, the effects of the "looks" variables do not change in important ways once all of the new covariates are introduced in the model for neither males nor females.

(vi) The Chow F statistic for the test of whether the slopes of the regression functions differ across men and women, allowing for an intercept shift under the null, is

$$\begin{aligned} F =& \frac{SSR_p - (SSR_{male} + SSR_{female})}{SSR_{male} + SSR_{female}} \times \frac{n - 2(k+1)}{k} \\ & \frac{293.6291 - (166.0841 + 83.69328)}{166.0841 + 83.69328} \times \frac{1260 - 2(13+1)}{13} \\ \approx & 16.638. \end{aligned}$$

This yields a p value of approximately 0, which indicates that we reject the null hypothesis that the slopes for the two groups are the same, i.e. the slopes for the two groups are different with high probability.

C8.9)

(i) After applying OLS, we find that

$$\begin{aligned} \widehat{cigs} =& \underbrace{-3.64}_{(24.08)} + \underbrace{0.880}_{(0.728)} \log(income) - \underbrace{0.751}_{(5.773)} \log(cigpric) - \underbrace{0.501}_{(0.167)} educ \\ & + \underbrace{0.771}_{(0.160)} age - \underbrace{0.0090}_{(0.0017)} age^2 - \underbrace{2.83}_{(1.11)} restaurn. \end{aligned}$$

(ii) Using the WLS outlined on page 260, we find the same fitted model that is shown in (8.36), i.e.

$$\begin{aligned} \widehat{cigs} =& \underbrace{5.64}_{(17.80)} + \underbrace{1.30}_{(0.44)} \log(income) - \underbrace{2.94}_{(4.46)} \log(cigpric) - \underbrace{0.462}_{(0.120)} educ \\ & + \underbrace{0.482}_{(0.097)} age - \underbrace{0.0056}_{(0.0009)} age^2 - \underbrace{3.46}_{(0.80)} restaurn. \end{aligned}$$

(iii) We find an F statistic of 11.15, which implies that the p value is approximately 0.00002, so we reject the null hypothesis that the weighted residuals are homoskedastic, i.e. we find that the weighted residuals are heteroskedastic.

(iv) The finding in part (iii) indicates that the proposed form of heteroskedasticity used in obtaining (8.36) is incorrect.

(v) The valid standard errors for the WLS estimates that allow the variance function to be misspecified, i.e. robust standard errors, are show in standard form:

$$\begin{aligned} \widehat{cigs} =& \underbrace{5.64}_{(25.70)} + \underbrace{1.30}_{(0.59)} \log(income) - \underbrace{2.94}_{(6.05)} \log(cigpric) - \underbrace{0.462}_{(0.162)} educ \\ & + \underbrace{0.482}_{(0.135)} age - \underbrace{0.0056}_{(0.0014)} age^2 - \underbrace{3.46}_{(1.02)} restaurn. \end{aligned}$$

It happens to be the case that all of the robust standard errors are larger than their normal counterparts.

Wooldridge Panel Data Textbook

3.5)

(a) We know that

$$\text{Var}(\bar{y}_N) = \frac{\text{Var}(y_i)}{N} = \frac{\sigma^2}{N}.$$

Therefore,

$$\text{Var}(\sqrt{N}(\bar{y}_N - \mu)) = (\sqrt{N})^2\text{Var}(\bar{y}_N) = (N)\frac{\sigma^2}{N} = \sigma^2.$$

(b) From the CLT, we know that $\sqrt{N}(\bar{y}_N - \mu) \xrightarrow{d} N(0, \sigma^2)$. Thus, $\text{Avar}(\sqrt{N}(\bar{y}_N - \mu)) = \sigma^2$.

(c)

$$\begin{aligned} \sqrt{N}(\bar{y}_N - \mu) &\xrightarrow{d} N(0, \sigma^2) \\ \implies \bar{y}_N - \mu &\xrightarrow{d} N(0, \frac{\sigma^2}{N}) \\ \implies \bar{y}_N &\xrightarrow{d} N(\mu, \frac{\sigma^2}{N}) \\ \implies \text{Avar}(\bar{y}_N) &= \frac{\sigma^2}{N}. \end{aligned}$$

This is exactly the same as $\text{Var}(\bar{y}_N)$.

(d) We have previously shown that

$$\bar{y}_N \xrightarrow{d} N(\mu, \frac{\sigma^2}{N}).$$

Thus, it must be that the asymptotic standard deviation of $\bar{y}_N$ is

$$\frac{\sigma}{\sqrt{N}}.$$

(e) We need a consistent estimator of *sigma* in order to properly estimate the asymptotic standard error of $\bar{y}_N$. We often take the unbiased estimator of variance, i.e. sample variance, to find $\hat{\sigma}$.

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_{i=1}^N (y_i - \bar{y}_N)^2}{N-1} \\ \implies \hat{\sigma} &= \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y}_N)^2}{N-1}}, \end{aligned}$$

and the asymptotic standard error of $\bar{y}_N$ is

$$\frac{\hat{\sigma}}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y}_N)^2}{N-1}} \Big/ \sqrt{N} = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y}_N)^2}{N(N-1)}}.$$

4.1)

(a) First,

$$\begin{aligned}\log(wage) =&\beta_0 + \beta_1 married + \beta_2 educ + \boldsymbol{z\gamma} + u\\ \implies wage =& e^{\beta_0+\beta_1 married+\beta_2 educ+\boldsymbol{z\gamma}+u}\\ =&e^u e^{\beta_0+\beta_1 married+\beta_2 educ+\boldsymbol{z\gamma}}\\ =&ce^{\beta_0+\beta_1 married+\beta_2 educ+\boldsymbol{z\gamma}}\end{aligned}$$

since $\mathbb{E}[u|married, educ, \boldsymbol{z}) = 0$ implies that $\mathbb{E}[e^u|married, educ, \boldsymbol{z}) = c$ for some real number c. As the hint suggests,

$$\begin{aligned}\frac{\theta_1}{100} =&\frac{\mathbb{E}[wage|married = 1, educ, \boldsymbol{z}] - \mathbb{E}[wage|married = 0, educ, \boldsymbol{z}]}{\mathbb{E}[wage|married = 0, educ, \boldsymbol{z}]}\\ =&\frac{ce^{\beta_0+\beta_1+\beta_2 educ+\boldsymbol{z\gamma}} - ce^{\beta_0+\beta_2 educ+\boldsymbol{z\gamma}}}{ce^{\beta_0+\beta_2 educ+\boldsymbol{z\gamma}}}\\ =&e^{\beta_1} - 1\\ \implies \theta_1 =&100(e^{\beta_1} - 1).\end{aligned}$$

(b) Note that $g(\beta_1) = \theta_1 = 100(e^{\beta_1} - 1)$ and $g(\hat{\beta}_1) = \hat{\theta}_1 = 100(e^{\hat{\beta}_1} - 1)$. Since

$$\frac{\partial g(\hat{\beta}_1)}{\partial \hat{\beta}_1} = 100e^{\hat{\beta}_1},$$

using the delta method we can see that

$$se(\hat{\theta}_1) = \frac{\partial g(\hat{\beta}_1)}{\partial \hat{\beta}_1} se(\hat{\beta}_1) = 100e^{\hat{\beta}_1} se(\hat{\beta}_1).$$

(c) Let $\Delta educ = educ_2 - educ_1$ where $educ_1$ and $educ_2$ are two different education levels. Now we can estimate θ_2 just as we estimated θ_1.

$$\begin{aligned}\frac{\theta_2}{100} =&\frac{\mathbb{E}[wage|married, educ = educ_2, \boldsymbol{z}] - \mathbb{E}[wage|married, educ = educ_1, \boldsymbol{z}]}{\mathbb{E}[wage|married = 0, educ = educ_1, \boldsymbol{z}]}\\ =&\frac{ce^{\beta_0+\beta_1 married+\beta_2 educ_2+\boldsymbol{z\gamma}} - ce^{\beta_0+\beta_1 married+\beta_2 educ_1+\boldsymbol{z\gamma}}}{ce^{\beta_0+\beta_1 married+\beta_2 educ_1+\boldsymbol{z\gamma}}}\\ =&e^{\beta_2(educ_2-educ_1)} - 1\\ =&e^{\beta_2\Delta educ} - 1\\ \implies \theta_2 =&100(e^{\beta_2\Delta educ} - 1).\end{aligned}$$

Similarly, we notice that $h(\beta_2) = \theta_2 = 100(e^{\beta_2\Delta educ} - 1)$ and $h(\hat{\beta}_2) = \hat{\theta}_2 = 100(e^{\hat{\beta}_2\Delta educ} - 1)$. Since

$$\frac{\partial h(\hat{\beta}_2)}{\partial \hat{\beta}_2} = 100(\Delta educ)e^{\hat{\beta}_2\Delta educ},$$

using the delta method we can see that

$$se(\hat{\theta}_2) = \frac{\partial h(\hat{\beta}_2)}{\partial \hat{\beta}_2} se(\hat{\beta}_2) = 100(\Delta educ)e^{\hat{\beta}_2\Delta educ} se(\hat{\beta}_2).$$

(d) From page 70 of the book, we are given that

$$\begin{aligned}\log(\hat{wage}) =&5.40 + 0.14exper + 0.012tenure + .199married - .091south + .184urban\\ &- .188black + .065educ,\end{aligned}$$

where the standard errors of *married* and *educ* are .039 and .006, respectively. Thus when $\Delta educ = 4$,

$$\begin{aligned}
\hat{\theta}_1 =& 100(e^{\beta_1} - 1) = 100(e^{.199} - 1) \approx 22.018, \\
se(\hat{\theta}_1) =& 100e^{\hat{\beta}_1} se(\hat{\beta}_1) = 100e^{.199}(.039) \approx 4.759, \\
\hat{\theta}_2 =& 100(e^{\beta_2 \Delta educ} - 1) = 100(e^{.065(4)} - 1) \approx 29.693, \\
se(\hat{\theta}_2) =& 100(\Delta educ)e^{\hat{\beta}_2 \Delta educ} se(\hat{\beta}_2) = 100(4)e^{.065(4)}(.006) \approx 3.113.
\end{aligned}$$

4.2)

(a) First, we assume that $\mathbb{E}[u|\boldsymbol{x}] = 0$ and $\boldsymbol{X'X}$ is nonsingular. Then, it must be that

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] =& \mathbb{E}[(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}|\boldsymbol{X}] \\
=& (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\
=& (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{X\beta} + u|\boldsymbol{X}] \\
=& (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathbb{E}[u|\boldsymbol{X}] \\
=& \boldsymbol{\beta}.
\end{aligned}$$

(b) Now we also assume that $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2$. First, note that

$$\sigma^2 = \mathrm{Var}(\boldsymbol{u}|\boldsymbol{X}) = \mathbb{E}(\boldsymbol{uu}'|\boldsymbol{X}) - (\mathbb{E}(\boldsymbol{u}|\boldsymbol{X}))^2 = \mathbb{E}(\boldsymbol{uu}'|\boldsymbol{X}).$$

Then, it must be that

$$\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] =& \mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}])(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}])'] \\
=& \mathbb{E}[((\boldsymbol{X'X})^{-1}\boldsymbol{X'y} - \boldsymbol{\beta})((\boldsymbol{X'X})^{-1}\boldsymbol{X'y} - \boldsymbol{\beta})'|\boldsymbol{X}] \\
=& \mathbb{E}[((\boldsymbol{X'X})^{-1}\boldsymbol{X'y} - (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta})((\boldsymbol{X'X})^{-1}\boldsymbol{X'y} - (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta})'|\boldsymbol{X}] \\
=& \mathbb{E}[((\boldsymbol{X'X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}))((\boldsymbol{X'X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\beta}))'|\boldsymbol{X}] \\
=& \mathbb{E}[((\boldsymbol{X'X})^{-1}\boldsymbol{X'u})((\boldsymbol{X'X})^{-1}\boldsymbol{X'u})'|\boldsymbol{X}] \\
=& \mathbb{E}[(\boldsymbol{X'X})^{-1}\boldsymbol{X'uu'X}(\boldsymbol{X'X})^{-1}|\boldsymbol{X}] \\
=& (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{uu}'|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\
=& (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathrm{Var}[\boldsymbol{u}|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\
=& \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1} \\
=& \sigma^2(\boldsymbol{X'X})^{-1}
\end{aligned}$$

4.4)

$$\begin{aligned}
\frac{\sum_{i=1}^N \hat{u}_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} =& \frac{\sum_{i=1}^N (u_i^2 - 2\boldsymbol{x}_i u_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + [\boldsymbol{x}_i((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2)\boldsymbol{x}_i'\boldsymbol{x}}{N} \\
=& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x} - 2\boldsymbol{x}_i u_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{x}_i'\boldsymbol{x} + [\boldsymbol{x}_i((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2\boldsymbol{x}_i'\boldsymbol{x}}{N} \\
=& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} - \frac{\sum_{i=1}^N 2\boldsymbol{x}_i u_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{x}_i'\boldsymbol{x}}{N} + \frac{\sum_{i=1}^N [\boldsymbol{x}_i((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2\boldsymbol{x}_i'\boldsymbol{x}}{N} \\
=& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} - \frac{\sum_{i=1}^N \boldsymbol{a}_i}{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{\sum_{i=1}^N b_i}{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2,
\end{aligned}$$

where $\{\boldsymbol{a}_i\} = \{\boldsymbol{x}_i u_i(\boldsymbol{x}_i'\boldsymbol{x}) : i = 1, 2, ..., N\}$ and $\{b_i\} = \{\boldsymbol{x}_i'\boldsymbol{x}(\boldsymbol{x}_i'\boldsymbol{x}) : i = 1, 2, ..., N\}$. Thus, now using the facts that sample averages are $O_p(1)$ and that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$, which implies that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 = o_p(1)$ by Slutsky's theorem, and taking the limit as $N \to \infty$ of both sides, we have that

$$\begin{aligned}\frac{\sum_{i=1}^N \hat{u}_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} =& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} - O_p(1)o_p(1) + o_p(1) \\ =& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} - o_p(1) + o_p(1) \\ =& \frac{\sum_{i=1}^N u_i^2 \boldsymbol{x}_i'\boldsymbol{x}}{N} + o_p(1).\end{aligned}$$

This gives us the results that the estimator $\hat{\boldsymbol{B}} = N^{-1}\sum_{i=1}^N \hat{u}_i^2 \boldsymbol{x}_i'\boldsymbol{x}$ is consistent for $\boldsymbol{B} = \mathbb{E}[u^2\boldsymbol{x}'\boldsymbol{x}]$.

4.9)

(a)

$$\begin{aligned}\log(y) =& \beta_0 + \boldsymbol{x\beta} + \alpha_1 \log(y_{-1}) + u \\ \implies \log(y) - \log(y_{-1}) =& \beta_0 + \boldsymbol{x\beta} + (\alpha_1 - 1)\log(y_{-1}) + u.\end{aligned}$$

Therefore, it is clear that the estimator of $\boldsymbol{x}$ is still $\boldsymbol{\beta}$ for both cases.

(b) Note that by definition

$$\alpha_1 = \frac{\text{Cov}(\log(y), \log(y_{-1}))}{\text{Var}(\log(y_{-1})} = \frac{\text{Cov}(\log(y), \log(y_{-1}))}{\text{sd}(\log(y_{-1}))\text{sd}(\log(y_{-1}))}.$$

Since we assume that the distributions of y and y_{-1} are identical, it must be that $\text{Var}(\log(y)) = \text{Var}(\log(y_{-1}))$, which implies that $\text{sd}(\log(y)) = \text{sd}(\log(y_{-1}))$. Therefore,

$$\begin{aligned}\alpha_1 =& \frac{\text{Cov}(\log(y), \log(y_{-1}))}{\text{sd}(\log(y_{-1}))\text{sd}(\log(y_{-1}))} = \frac{\text{Cov}(\log(y), \log(y_{-1}))}{\text{sd}(\log(y))\text{sd}(\log(y_{-1}))} = \text{Corr}(\log(y), \log(y_{-1})) \\ \implies |\alpha_1| \leq& 1\end{aligned}$$

since the absolute value of correlation is bounded by 1.

Other

Show that for a regression model, if a regressor x_j is measured with error, then it will be endogenous.

Consider the model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{j-1}x_{j-1} + \beta_j x_j + u.$$

However, the true x_j is unobservable but can be measured with some error, i.e.

$$\tilde{x}_j = x_j + e_j,$$

where $\tilde{x}_j$ is the observed variable, $\mathbb{E}[e_j] = 0$, and e_j has some finite variance $\text{Var}(e_j) < \infty$. Then, the model can be shown to actually be

$$\begin{aligned}y =& \beta_0 + \beta_1 x_1 + ... + \beta_{j-1}x_{j-1} + \beta_j x_j + u \\ =& \beta_0 + \beta_1 x_1 + ... + \beta_{j-1}x_{j-1} + \beta_j(\tilde{x}_j - e_j) + u \\ =& \beta_0 + \beta_1 x_1 + ... + \beta_{j-1}x_{j-1} + \beta_j\tilde{x}_j + (u - \beta_j e_j).\end{aligned}$$

We assume that u is not correlated with $\tilde{x}_j$, which implies that u is also not correlated with e_j. Another reasonable assumption is that e_j is not correlated with x_i for $i = 1, ..., j-1$. In addition, we make the classical errors-in-variables assumption that $\text{Cov}(x_j, e_j) = 0$. If this holds, then it must be that $\tilde{x}_j$ and e_j are correlated since

$$\text{Cov}(\tilde{x}_j, e_j) = \text{Cov}(x_j + e_j, e_j) = \text{Cov}(e_j, e_j) = \text{Var}(e_j).$$

Then, it is clear that the model is endogenous as long as neither β_j nor $\text{Var}(e_j)$ are equal to 0 because

$$\text{Cov}(\tilde{x}_j, u - \beta_j e_j) = -\beta_j \text{Cov}(\tilde{x}_j, e_j) = -\beta_j \text{Var}(e_j).$$