Assignment 2

630516am and 590049bm

23 Nov 2021

Question 1

```
set.seed(2021)
```

```
b0 <- 3
b1 <- 5
b2 <- 8

gamma <- 1

x1 <- rnorm(5000, mean = 1, sd = 1)
x2 <- rnorm(5000, mean = 2, sd = 1)

z <- rgamma(5000, shape = 1.2, scale = 1.1)

sigma_sq <- 1*exp(gamma*z)
epsilon <- rnorm(5000, mean = 0, sd = sqrt(sigma_sq))

y <- b0 + b1*x1 + b2*x2 + epsilon
```

```
model1 <- lm(y ~ x1 + x2)

modelsummary(model1,
             vcov = c("iid", "HC0"),
             gof_map = gm,
             stars = T)
```

```
lmtest::bptest(formula = y ~ x1 + x2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  y ~ x1 + x2
```

	Model 1	Model 2
(Intercept)	2.609***	2.609***
	(0.375)	(0.315)
x1	5.009***	5.009***
	(0.149)	(0.078)
x2	8.080***	8.080***
	(0.152)	(0.097)
N	5000	5000
Adj. R2	0.44	0.44

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
## BP = 0.50429, df = 2, p-value = 0.7771
```

```r
gamma <- 0

sigma_sq <- 1*exp(gamma*z)
epsilon <- rnorm(5000, mean = 0, sd = sqrt(sigma_sq))

y <- b0 + b1*x1 + b2*x2 + epsilon
```

```r
model1 <- lm(y ~ x1 + x2)
model2 <- lm(y ~ x1 + x2, weights = sigma_sq)
```

Question 2

1. Show that the OLS estimator of the parameter β is not consistent.
2. Derive plim (b) where b is the OLS estimator of β. Determine the sign of the magnitude of the inconsistency when $0 < \beta < 1$, that is, the sign of plim(b) - β when $0 < \beta < 1$.

In order to evaluate consistency, we must derive the probability limit. Hence, we answer two questions at once.

First, we demean the two variables so that the constant-term α equals zero. Then we regress $\tilde{C} = \beta\tilde{D} + \epsilon$. We can do this because of Frisch-Waugh-Lovell. The estimate that we get is:

$$\hat{\beta} = (\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T C_{=}(\tilde{D}^T\tilde{D})^{-1}(\beta\tilde{D} + \epsilon)$$

and

$$\mathbb{E}[\hat{\beta}] = \beta + (\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T\epsilon$$

Evaluating the probability limit gives:

$$\text{plim}_{n\to\infty}(\hat{\beta}) = \beta + \text{plim}(\frac{1}{n}\tilde{D}^T\tilde{D})^{-1} \cdot \text{plim}(\frac{1}{n}\tilde{D}^T\epsilon)$$

which simplifies to:

$$\beta + \frac{1}{\text{Var}(D)} \cdot \frac{1}{1-\beta}\sigma^2$$

by the fact that variances and covariances are the same after demeaning, and by the reduced form equation for D made explicit below. Under $0 < \beta < 1$, since variances are positive, the right term can only be positive and thus the bias is always positive.

Substituting equation (2) into equation (1) and solving for C gives:

$$C = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}Z_i + \frac{1}{1-\beta}\epsilon_i$$

substituting this back in the definition for D gives:

$$D = \frac{\alpha}{1-\beta} + \left(\frac{\beta}{1-\beta} + 1\right)Z_i + \frac{1}{1-\beta}\epsilon_i$$

From this, we can calculate Cov(D, ϵ_i), which is $\frac{1}{1-\beta}\text{Var}(\epsilon) = \frac{1}{1-\beta}\sigma^2$.

3. Find an instrumental variable (IV) for the endogenous variable D i and argue why it could be an IV.

The instrumental variable could be Z, because it is relevant, i.e. Cov(D, Z) $\neq 0$. Also, it is exogenous (valid), as it is exogenously generated and has no correlation with the error term ϵ according to the DGP sketched out here.

4. Derive b_{IV}, the IV estimator of β in terms of the variables C, D, and Z step by step.

First, suppose X is a matrix consisting of a column of 1's and D, so we can write:

$$\mathbb{E}[z_i \epsilon_i] = 0 = \frac{1}{n} \sum z_i (c_i - \alpha - \beta D_i) = \frac{1}{n} \sum z_i (c_i - x_i \beta)$$

Using this moment condition to solve for β, we retrieve the b_{IV} estimator:

$$\hat{b}_{IV} = \left(\sum z_i^T x_i \right)^{-1} \left(z_i^T c_i \right) = (Z^T X)^{-1} Z^T C$$

5. Use the expression of b_{IV} to show that it is consistent.

$$b_{IV} = (Z^T X)^{-1} Z^T C = (Z^T X)^{-1} Z^T (X\beta + \epsilon) = (Z^T X)^{-1} Z^T X \beta + (Z^T X)^{-1} Z^T \epsilon = \beta + (Z^T X)^{-1} Z^T \epsilon$$

Evaluating the plim of this estimator then gives:

$$\text{plim}(b_{IV}) = \beta + \text{plim}(Z^T X)^{-1} \cdot \text{plim}(Z^T \epsilon)$$

where the last factor goes to zero as $n \to \infty$.

Question 3

1. Plot the distribution of the growth rate of employment and of import exposure 1990-2007 across US commuting zones

Question 4

1. Imagine we fit a linear probability model of $y_i = \alpha + \beta x_i + \epsilon_i$. Derive the distribution of the error terms. Will our least-squares parameter estimate β be unbiased? Will it still be the most efficient estimator?

We know that y_i is distributed with probability p. If we use a linear model to estimate a y, we impose that $p(y_i = 1) = \mathbb{E}[y_i] = \alpha + \beta x_i$. Then, we can characterize the distribution of the error term:

$$\epsilon_i = \begin{cases} 1 - \hat{\alpha} - \hat{\beta} x_i & \text{with } p = \alpha + \beta x_i \\ -\hat{\alpha} - \hat{\beta} x_i & \text{with } p = 1 - (\alpha + \beta x_i) \end{cases}$$

Then, since ϵ_i is now a shifted Bernouilli variable, we can calculate the expected value as:

$$\mathbb{E}[\epsilon_i] = (1 - \alpha - \beta x_i) \cdot (\alpha + \beta x_i) + (-\alpha - \beta x_i) \cdot (1 - \alpha - \beta x_i) = 0$$

The fact that $\mathbb{E}[\epsilon] = 0$ also means that the OLS estimator is unbiased. However, the variance σ_ϵ^2 as $p(1-p) = (\alpha + \beta x_i) \cdot (1 - (\alpha + \beta x_i)) = f(x_i)$. This means that the variance of the error term is heteroskedastic! Hence, the estimator will not be the most efficient estimator, as one of the Gauss-Markov assumptions is violated.

2. Now imagine that we want to estimate this regression model for a given distribution of the errors F (e.g the logistic distribution) using maximum likelihood. Write out the distribution of y_i.

In this case, more generally, y_i is distributed as:

$$y_i = \begin{cases} 1 & \text{with } p = F(X_i\beta) \\ 0 & \text{with } 1-p = 1 - F(X_i\beta) \end{cases}$$

The likelihood of one observation (which is the pdf) is then simply:

$$l_1(y_i|x_i) = (F(X_i\beta))^{y_i}(1 - F(X_i\beta))^{1-y_i}$$

3. Use the distribution to write out the log-likelihood function. Then, write out the first-order condition for maximisation with respect to β.

The log-likelihood for n observations is:

$$\mathcal{L}_n(y_i|x_i) = \sum_{i=1}^{n} y_i \log(F(X_i\beta)) + (1 - y_i)\log(1 - F(X_i\beta))$$

Taking the first derivative with respect to the parameters β gives:

$$\frac{\partial \log \mathcal{L}_n(y_i|x_i)}{\partial \beta} = \sum y_i \frac{1}{F(X_i\beta)} f(X_i\beta)X_i - (1 - y_i)\frac{1}{1 - F(X_i\beta)} f(X_i\beta)X_i = 0$$

This can be rewritten as:

$$\sum \frac{y_i - F(X_i\beta)}{F(X_i\beta)(1 - F(X_i\beta))} f(X_i\beta)X_i = 0$$

4. Imagine we assume a logistic distribution of the errors. Show that our expression above simplifies to . . ..

We use the fact from Heij et al., p. 449, that for the logistic distribution, $F(.)(1 - F(.)) = f(.)$. Then, our expression simlifies to:

$$\sum y_i - F(X_i\beta)x_i = 0$$

Then, substituting the logit cdf for F gives:

$$\sum y_i - \left(\frac{1}{1 + \exp^{-x_i\beta}}\right) x_i = 0$$

which is what we were required to show.

5. Finally, use the value of $F(x_i\beta)$ to write the log of the odds ratio as a function of the parameters of the model. Thus, give an interpretation of the value of β.

The log odds ratio is defined as:

$$OR = \frac{\frac{1}{1+\exp^{-x_i\beta}}}{1 - \frac{1}{1+\exp^{-x_i\beta}}} = e^{x_i\beta}$$

The log-odds ratio is then:

$$\log OR = x_i\beta$$

Beta is then equal to the derivative of the log odds ratio with respect to a regressor. This means that the the strength of β is indicative of the relative likelihood of $P(Y_i = 1)$ occurring versus $P(Y_i = 0)$ occurring. In other words, if $\beta > 0$, then an increase in the independent variable makes the event more likely, and a decrease in the independent variable makes the event less likely.