Econometrics 710 Problem Set #9

April 29, 2015

TA: Cristian Hernandez
cahernandez3@wisc.edu

Exercise 9.3

(a) We are mainly interested in studying the returns of education on wage, after controlling for experience and demographics. Table 1 shows the parameter estimates from three specifications of the outcome equation. Model (1) is the specification in equation (3.12) in chapter 3 that includes a constant, education (ED), experience (EX), and experience squared (EXSQ). Model (2) includes all regressors except the occupation dummies (MANUF to PROF) and Model (3) include all regressors with occupation dummies. Because of collinearity, we dropped age (AGE)—collinear with ED, EX and the constant—, one of the marital status dummies (MARRFE), and one of the occupation dummies (PROF) when we consider specification (3).[1,2]

	Model (1)	Model (2)	Model (3)
Education	0.0726 (0.0072)	0.0725 (0.0069)	0.0707 (0.0067)
Experience	0.0296 (0.0048)	0.0292 (0.0047)	0.0255 (0.0047)
Experience^2	-0.0003 (0.0001)	-0.0004 (0.0001)	-0.0003 (0.0001)
Hispanic	-	-0.0348 (0.0556)	0.0110 (0.0537)
Female	-	-0.3087 (0.0350)	-0.2444 (0.0395)
Non-white	-	-0.1385 (0.0703)	-0.1271 (0.0713)
Union	-	0.2118 (0.0371)	0.2416 (0.0370)
Married	-	0.0270 (0.0366)	0.0305 (0.0354)
# Children in household	-	-0.0317 (0.0138)	-0.0233 (0.0134)
Regional dummy(South)	-	0.0082 (0.0364)	0.0103 (0.0350)
Constant	0.3939 (0.1042)	0.5103 (0.1038)	0.4783 (0.1034)
Adjusted R^2	0.2360	0.3866	0.4287

Table 1: OLS estimates for log(wage) equation. Standard errors are heteroskedastic-robust (Horn-Horn-Duncan formula). Sample size $n = 550$.

[1] Note that the data doesn't exhibit collinearity in the dummies, but if we include all the dummies and we add them up, sometimes we get that the sum is equal to 2, which is a sign that some of the entries are wrong or that the categorical variable is misspecified.

[2] If we include all the dummies, the parameter estimate for coefficient changes to approximately 0.05. It is hard to believe that including those dummies should change the slope coefficient for education in such magnitude. Instead, we would expect that only the constant should be affected.

(b) and (c)

A right/reasonable specification of the model depends on what's the question that you are trying to answer. For example, if I am interested in whether the effect of gender gap differs for the white or non-white group, I should add an interaction term for the female dummy (FE) and non-white dummy (NONWH). We just have to make sure that we include all the relevant variables, so that we minimize the omitted variable bias. If the coefficient of (FE × NONWH) is negative then the wage gap between male and female is bigger among the non-white group compared with the white group. I report two regression results in Table 2 by adding these variable with Model (2), (3) in (a).

	Model (4)	Model (5)
Education	0.0701 (0.0071)	0.0685 (0.0068)
Experience	0.0286 (0.0047)	0.0249 (0.0048)
Experience squared	−0.0004 (0.0001)	−0.0003 (0.0001)
Femald × Non-White	0.2770 (0.1344)	0.2747 (0.1284)
Hispanic	−0.0407 (0.0558)	0.0050 (0.0540)
Female	−0.3383 (0.0366)	−0.2764 (0.0385)
Non-white	−0.2689 (0.1021)	−0.2554 (0.0994)
Union	0.2119 (0.0371)	0.2417 (0.0369)
Married	0.0321 (0.0366)	0.0351 (0.0354)
# Children in household	−0.0338 (0.0139)	−0.0253 (0.0135)
Regional dummy(South)	0.0012 (0.0360)	0.0036 (0.0346)
Constant	0.5585 (0.1089)	0.5198 (0.1074)
Adjusted R^2	0.3927	0.4347

Table 2: OLS Estimates for log(wage) equation. Model (4): Model (2) with interaction dummy. Model (5): Model (3) with interaction dummy. Sample size is $n = 550$. Standard errors are robust (Horn-Horn-Duncan formula.)

For (d) and (e), we consider the Model (2) in (a)

(d) To test if the error variance for men and women is the same, we construct the null hypothesis

$$H_0 : \sigma^2_{men} = \sigma^2_{women}$$

and use a two-sided F-test for equality of variance

$$F = \frac{\widehat{s}^2_{women}}{\widehat{s}^2_{men}} = \frac{0.2255}{0.1364} = 1.6530 > 1.2792 = F_{207-11,343-1,0.975}$$

We reject the null hypothesis that error variances are equal among men and women. Therefore, the data suggests that the error variances are different for men and women.

Another way to approach this problem is to model the variance and perform a test on the coefficients.

(e) We apply the same method in (d) for whites and nonwhites. Since test statistic $F = \frac{\widehat{s}^2_{non-white}}{\widehat{s}^2_{white}} = \frac{0.2557}{0.1369} = 0.5354 < 1.4820 = F_{57-11,493-11,0.975}$, we fail to reject the null hypothesis $H_0 : \sigma^2_{white} = \sigma^2_{non-white}$.

(f) For the remaining problem (f), (g), we will use Model (1) in (a). We use the following model for conditional variance;

$$\begin{aligned} e_i^2 &= \alpha_0 + \boldsymbol{z}'_{1i}\boldsymbol{\alpha_1} + \zeta_i \\ &= \boldsymbol{z}'_i\boldsymbol{\alpha} + \zeta_i \end{aligned}$$

where $\boldsymbol{z_{1i}}$ includes the square of all the regressors x_i in model (1) without the constant. The following equation shows the results of the skedastic regression

$$\widehat{e}_i^2 = \underset{(0.0440)}{0.0925} + \underset{(0.0002)}{0.0002}ED^2 + \underset{(0.0000)}{0.0001}EX^2 + \underset{(0.0000)}{0.0000}EXSQ^2$$

To test heteroskedasticity, the null is $H_0 : \boldsymbol{\alpha_1} = 0$. We use the Wald statistic with heteroskedastic-robust covariance matrix. As a result, $W_n = 10.123 > \chi_{0.95}(3) = 7.82$; therefore, under 5% significance level we reject the null hypothesis that conditional variance is not a function of the regressors.

(g) Using this model for the conditional variance, I re-estimated the model (1) and report the results of FGLS with OLS. Specifically, I used the following FGLS estimator

$$\tilde{\beta}_{FGLS} = (X'\tilde{D}X)^{-1}(X'\tilde{D}y)$$

where

$$\tilde{D} = \text{diag}\{\tilde{\sigma}_1^2, \cdots, \tilde{\sigma}_n^2\}, \quad \tilde{\sigma}_i^2 = \tilde{\alpha}' z_i, \quad \tilde{\alpha} = (Z'Z)^{-1}Z'\hat{e}^2$$

For the estimator of the asymptotic covariance matrix of FGLS, I used a White-type estimator, as proposed by Cragg (1992);

$$\tilde{V}_\beta = \left(\frac{1}{n}X'\tilde{D}^{-1}X\right)^{-1}\left(\frac{1}{n}X'\tilde{D}^{-1}\hat{D}\tilde{D}^{-1}X\right)\left(\frac{1}{n}X'\tilde{D}^{-1}X\right)^{-1}$$

where $\hat{D} = \text{diag}\{\hat{e}_1^2, \cdots, \hat{e}_n^2\}$.

Table 3: OLS and FGLS Estimates log(wage) equation.

	OLS	FGLS
Education	0.0726 (0.0072)	0.0713 (0.0073)
Experience	0.0296 (0.0048)	0.0290 (0.0057)
Experience^2	-0.0003 (0.0001)	-0.0003 (0.0001)
Constant	0.3939 (0.1042)	0.4118 (0.1112)

(h), (i) The parameter estimates are similar between OLS and FGLS. Interestingly, standard errors for OLS are smaller than the one for FGLS for all variables.

Exercise 9.4

By the law of iterated expectations,

$$\mathbb{E}|y_i - g(\mathbf{x}_i)| = \mathbb{E}(\mathbb{E}(|y_i - g(\mathbf{x}_i)| \ \ |\mathbf{x}_i))$$

For any fixed $\mathbf{x}_i = \mathbf{x}$, we will show that $med(y_i|\mathbf{x}_i = \mathbf{x}) = \arg\min_\theta \mathbb{E}(|y_i - \theta| \ \ |\mathbf{x}_i)$.

Let's expand the objective function

$$\mathbb{E}(|y-\theta| \ \ |\mathbf{x}) = \int_{-\infty}^{\infty} |y-\theta| dF_{y|\mathbf{x}}(y|\mathbf{x}) = \int_{-\infty}^{\theta} (\theta - y) dF_{y|\mathbf{x}}(y|\mathbf{x}) + \int_{\theta}^{\infty} (y-\theta) dF_{y|\mathbf{x}}(y|\mathbf{x})$$

Now take derivative with respect to θ applying the Leibniz rule of integration (differentiation under the integral); we get the following first order condition

$$\frac{d\mathbb{E}(|y-\theta| \ \ |\mathbf{x})}{d\theta} = -\int_{-\infty}^{\theta} dF_{y|\mathbf{x}}(y|\mathbf{x}) - \int_{\theta}^{\infty} dF_{y|\mathbf{x}}(y|\mathbf{x}) = -1 + 2F_{y|\mathbf{x}}(\theta|\mathbf{x}) = 0$$

Therefore,

$$F_{y|\mathbf{x}}(\theta|\mathbf{x}) = 1/2$$

Let $Q_{y|x}(\tau|\mathbf{x})$ denote the τ quantile of $y|\mathbf{x}$ (which is the inverse function of the conditional CDF). Apply the conditional quantile function to both sides of the previous equation and get

$$Q_{y|\mathbf{x}}(F_{y|\mathbf{x}}(\theta)|\mathbf{x}) = \theta = Q_{y|\mathbf{x}}(0.5|\mathbf{x}) = med(y|\mathbf{x})$$

Since this holds for any given $\mathbf{x}$, then $m(\mathbf{x}_i) = med(y_i|\mathbf{x}_i)$ minimizes $\mathbb{E}|y_i - g(\mathbf{x}_i)|$.

Exercise 9.5

$$0 = \mathbb{E}g(y_i - \theta) \quad = \quad \mathbb{E}(\tau - 1(y_i - \theta < 0)) = \tau - P(y_i < \theta)$$

Therefore, $P(y_i < \theta) = \tau$. Thus, θ is a τ quantile of the distribution of y_i assuming that y_i is continuous random variable.

Exercise 9.6

For continuous random variable U, and given θ,

$$\begin{aligned}
\mathbb{E}\rho_\tau(U-\theta) &= \mathbb{E}(U-\theta)(\tau - 1(U<\theta)) \\
&= \tau\mathbb{E}U - \theta\tau - \mathbb{E}U1(U<\theta) + \theta\mathbb{E}1(U<\theta) \\
&= \tau\mathbb{E}U - \theta\tau - \int_{-\infty}^{\theta} u\cdot dF_U(u) + \theta F_U(\theta) \qquad (1)
\end{aligned}$$

The FOC, applying the Leibniz rule, for θ is

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta}\mathbb{E}\rho_\tau(U<\theta) &= -\tau - \theta f_U(\theta) + F_U(\theta) + \theta f_U(\theta) \\
&= -\tau + F_U(\theta)
\end{aligned}$$

Thus, $\theta = Q_U(\tau) = F_U^{-1}(\tau)$, which is the τth quantile of U.

Exercise 10.1

$F_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n 1(\mathbf{x}_i \le \mathbf{x}), F_0(\mathbf{x}) = P(\mathbf{x}_i \le \mathbf{x}) = \mathbb{E}(1(\mathbf{x}_i \le \mathbf{x}))$. For a fixed $\mathbf{x}$, $1(\mathbf{x}_i \le \mathbf{x})$ are i.i.d random variables with mean $\mathbb{E}(1(\mathbf{x}_i \le \mathbf{x})) = F_0(\mathbf{x})$ and variance $\mathbb{E}(1(\mathbf{x}_i \le \mathbf{x})^2) - (\mathbb{E}(1(\mathbf{x}_i \le \mathbf{x})))^2 = F_0(\mathbf{x}) - (F_0(\mathbf{x}))^2 = F_0(\mathbf{x})(1-F_0(\mathbf{x}))$. By the central limit theorem, for any $\mathbf{x}$,

$$\sqrt{n}(F_n(\mathbf{x}) - F_0(\mathbf{x})) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n 1(\mathbf{x}_i \le \mathbf{x}) - \mathbb{E}1(\mathbf{x}_i \le \mathbf{x})\right) \xrightarrow{d} N(0, F_0(\mathbf{x})(1-F_0(\mathbf{x})))$$

Exercise 10.2

We know that each bootstrap draw y_i^* can take n possible values $\{y_1, \cdots, y_n\}$ with probability $1/n$ and $\{y_1^*, \cdots y_n^*\}$ are independent draws from the EDF.

$$\begin{aligned}
\mathbb{E}T_n &= \mathbb{E}\bar{y}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}y_i = \mu \\
\text{var}(T_n) &= \mathbb{E}(T_n - \mathbb{E}T_n)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (y_i-\mu)\right)^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(y_i-\mu)^2 = \frac{\sigma^2}{n} \\
\mathbb{E}T_n^* &= \mathbb{E}\bar{y}_n^* = \frac{1}{n}\sum_{i=1}^n \mathbb{E}y_i^* = \frac{1}{n}\sum_{i=1}^n \bar{y}_n = \bar{y}_n \\
\text{var}(T_n^*) &= \mathbb{E}(T_n^* - \mathbb{E}T_n^*)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (y_i^* - \bar{y}_n)\right)^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(y_i^* - \bar{y}_n)^2 = \frac{1}{n^2}\sum_{i=1}^n \left(\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y}_n)^2\right) = \frac{1}{n}\hat{\sigma}^2
\end{aligned}$$

Exercise 10.3

Assume that we use the same random integer i' from $[1, 2, \cdots, n]$ for each bootstrap draw for two different bootstrap procedures. The bootstrap procedure explained in the question creating a bootstrap data set (y^*, x^*) where $x^* = x_{i'}, y^* = {x^*}'\hat{\beta} + \hat{e}_{i'} = y_{i'}$. This is equivalent to the nonparametric bootstrap $(\tilde{y}^*, \tilde{x}^*)$ where $\tilde{x}^* = x_{i'}, \tilde{y}^* = y_{i'}$ for each bootstrap draw, since we get the exact same sample.

Exercise 10.4

The alternative percentile confidence interval is

$$C_2 = [\hat{\theta} - \tilde{q}_n^*(.95), \hat{\theta} - \tilde{q}_n^*(.05)]$$

where $\tilde{q}_n^*(\alpha)$ is the α quantile of the distribution of $\tilde{T}_n^* = (\hat{\theta}^* - \hat{\theta})$.

Since $\tilde{q}_n^*(\alpha) = s(\hat{\theta})q_n^*(\alpha)$ $\left(\alpha = P(T_n^* \leq q_n^*(\alpha)) = P(\frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta})} \leq q_n^*(\alpha)) = P(\tilde{T}_n^* \leq s(\hat{\theta})q_n^*(\alpha))\right)$, this is equivalent to the suggested bootstrap confidence interval,

$$C = [\hat{\theta} - s(\hat{\theta})q_n^*(.95), \hat{\theta} - s(\hat{\theta})q_n^*(.05)]$$

Exercise 10.5

Bootstrap t-statistics should be centered at $\hat{\theta}$, and $q_n^*(0.95)$ should be the 95% quantile of this centered bootstrap t-statistic $T_n^* = \frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)}$. Therefore, the procedure is wrong.

Exercise 10.6

(a) A 95% Efron percentile interval is

$$[0.75, 1.3]$$

(b) A 95% alternative percentile interval is

$$[\hat{\theta} - \hat{q}_n^*(0.975), \hat{\theta} - \hat{q}_n^*(0.025)]$$

where $\hat{q}_n^*(0.025), \hat{q}_n^*(0.975)$ are $2.5\%, 97.5\%$ sample quantiles of the distribution of $T_n^* = \hat{\theta}^* - \hat{\theta}$. By the invariance of the quantile of monotone transformations, the quantiles of T_n^* are equal to the quantiles of $\hat{\theta}^*$ minus $\hat{\theta}$, which are -0.45, 0.1 from the given information. Thus, a 95% alternative percentile confidence interval is [1.1, 1.65].

(c) A 95% percentile-t bootstrap confidence interval is

$$[\hat{\theta} - s(\hat{\theta})\tilde{q}_n^*(0.975), \hat{\theta} - s(\hat{\theta})\tilde{q}_n^*(0.025)]$$

where $\tilde{q}_n^*(0.025)$ and $\tilde{q}_n^*(0.975)$ are the 2.5% and 97.5% quantiles of $\tilde{T}_n^* = \frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)}$ and $s(\hat{\theta}^*)$ is the bootstrap standard error. Since we don't have any additional information of $s(\hat{\theta}^*)$, we can't compute sample quantiles of $\tilde{T}_n^*$. Then, we can't report a 95% percentile-t confidence interval.

Exercise 10.7

We consider the following regression equation;

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 + e_i \tag{2}$$

where y_i is house prices (in thousands of dollars), the vector x_i includes the number of bedrooms, lot size (sq ft), size of the house (sq ft), the colonial style dummy, and a constant. The following equation reports the OLS estimates with Horn-Horn-Duncan heteroskedastic-robust standard errors in parenthesis

$$\begin{aligned}\widehat{\text{House price}} &= \underset{(0.9845)}{1.1004}(\#\text{ of bedrooms}) + \underset{(0.0003)}{0.0002}(\text{lot size}) + \underset{(0.0023)}{0.0124}(\text{house size}) \\ &\quad + \underset{(1.7215)}{1.3716}(\text{Colonial Dummy}) - \underset{(3.9123)}{2.4127}\end{aligned}$$

95% confidence intervals for the regression coefficients (β_j) using asymptotic normal approximation and percentile-t bootstrap confidence interval are;

$$\begin{aligned} CI_{ASY} &= [\hat{\beta}_j - 1.96 s(\hat{\beta}_j), \hat{\beta}_j + 1.96 s(\hat{\beta}_j)] \\ CI_{BT} &= [\hat{\beta}_j - s(\hat{\beta}_j) q_n^*(0.975), \hat{\beta}_j - s(\hat{\beta}_j) q_n^*(0.025)] \end{aligned}$$

where $q_n^*(\alpha)$ is the α quantile of the bootstrap statistic $T_n^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{s(\hat{\beta}_j^*)}$.

Table 4: 95% confidence intervals for the OLS regression coefficients in equation (1). Asymptotic confidence interval (CI_{ASY}) and percentile-t bootstrap confidence interval (CI_{BT}) are reported;

Variables	CI_{ASY}	CI_{BT}
# of bedrooms	[-0.829, 3.030]	[-1.394, 3.134]
lot size (sq.ft)	[-0.0004, 0.0008]	[-0.0021, 0.0015]
size of house (sq.ft)	[0.008, 0.017]	[0.009, 0.024]
colonial dummy	[-2.002, 4.746]	[-4.048, 3.910]
constant	[-10.081, 5.255]	[-12.336, 5.366]