Problem Set 5 Solutions

ECON 210 Econometrics A

Evan Zuofu Liao*

November 12, 2015

Question 1

(a)(i) Code will be posted on Chalk. I summarize the results below.

Table 1: Q1 Part (a) (i)

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	2.3262***	(0.2783)	8.357	4.64e-13
x_1	1.5097***	(0.1017)	14.851	<2e-16
x_2	1.9297***	(0.1170)	16.498	<2e-16

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

Comparing the estimates with the true values, we can see that the estimates of β_1 and β_2 are NOT biased.

(ii) We repeat the exercise with x_2 and x_3 being correlated. Here's my results

Table 2: Q1 Part (a) (ii)

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	1.4452***	(0.2766)	5.225	9.95e-07
x_1	1.5435***	(0.1132)	13.640	<2e-16
x_2	2.6002***	(0.1014)	25.643	<2e-16

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

The estimate for the coefficient of x_2 is clearly biased!

*Comments and questions to evanliao@uchicago.edu. This solution draws from answers provided by previous TAs.

(iii) We compare results for $\beta_3 = 2$ and $\beta_3 = 0$. As we can see from the table below, the bias is proportional to β_3. In other words, our regression is not biased if $\beta_3 = 0$, and the bias gets larger as β_3 increases.

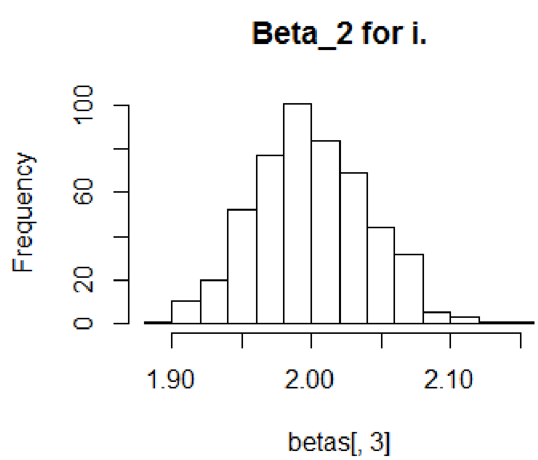
Table 3: Q1 Part (a) (iii)

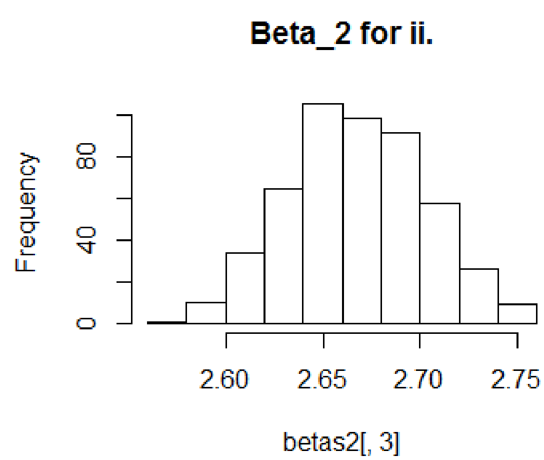
	β_3=2			β_3=0		
Variables	Estimate	Stand. Error	Pr(> |t|)	Estimate	Stand. Error	Pr(> |t|)
Intercept	1.7554***	(0.4259)	7.93e-05	0.9406***	(0.1793)	2.05e-7
x_1	1.5408***	(0.1743)	4.22e-14	1.5013***	(0.0848)	<2e-16
x_2	3.3277***	(0.1561)	2.00e-16	1.9216***	(0.0788)	<2e-16

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

(b) We repeat the same exercise with $n = 1000$. I'll skip the results here. Basically as n gets large, we'll get more precise estimates (i.e., smaller standard errors), but the bias will not disappear.

(c) I attach the histograms below





Question 2

(a) An easy way to see this is to realize that $D^{HD} + D^{HG} + D^{C} = 1$, so we have a problem of perfect multicollinearity. More formally in order to estimate the coefficients we have

$$\widehat{\Gamma} = (\hat{\gamma}_0\ \hat{\gamma}_1\ \hat{\gamma}_2\ \hat{\gamma}_3)' = (\mathbb{D}'\mathbb{D})^{-1}(\mathbb{D}'\mathbb{Y})$$

where $\mathbb{D}_{n\times 4}$ is our matrix of data points. Essentially $(\mathbb{D}'\mathbb{D})$ is not invertible because $\mathbb{D}$ is not linearly independent (you can check that $\exists\ c \neq 0 \in \mathbb{R}^4$ such that $\mathbb{D}c = 0$).

(b) The OLS implies that $\sum_{i=1}^n \hat{\epsilon}_i = 0$ and $\sum_{i=1}^n D_i\hat{\epsilon}_i = 0$ where $D_i = (D_i^{HD}\ D_i^{HG}\ D_i^C)'$. So we know $\sum_{HD} \hat{\epsilon}_i = 0$, $\sum_{HG} \hat{\epsilon}_i = 0$ and $\sum_C \hat{\epsilon}_i = 0$. Therefore,

$$\begin{aligned}
\bar{Y}_{HD} &= \frac{1}{n}\sum_{HD} Y_i = \hat{\alpha}_0 \frac{1}{n}\sum_{HD} D_i^{HD} + \frac{1}{n}\sum_{HD} \hat{\epsilon}_i = \hat{\alpha}_0 \\
\bar{Y}_{HG} &= \frac{1}{n}\sum_{HG} Y_i = \hat{\alpha}_1 \frac{1}{n}\sum_{HG} D_i^{HG} + \frac{1}{n}\sum_{HG} \hat{\epsilon}_i = \hat{\alpha}_1 \\
\bar{Y}_{C} &= \frac{1}{n}\sum_{C} Y_i = \hat{\alpha}_2 \frac{1}{n}\sum_{C} D_i^{C} + \frac{1}{n}\sum_{C} \hat{\epsilon}_i = \hat{\alpha}_2
\end{aligned}$$

(c) Similarly with our orthogonality conditions we know $\sum_{i=1}^n \hat{\nu}_i = 0$, $\sum_{HG} \hat{\nu}_i = 0$ and $\sum_C \hat{\nu}_i = 0$. Subtracting the latter two equations from the first equation gives us $\sum_{HD} \hat{\nu}_i = 0$. Thus we have

$$\begin{aligned}
\bar{Y}_{HD} &= \frac{1}{n}\sum_{HD} Y_i = \hat{\beta}_0 + \frac{1}{n}\sum_{HD} \hat{\nu}_i = \hat{\beta}_0 \\
\bar{Y}_{HG} &= \frac{1}{n}\sum_{HG} Y_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{HG} D_i^{HG} + \frac{1}{n}\sum_{HG} \hat{\nu}_i = \hat{\beta}_0 + \hat{\beta}_1 \\
\bar{Y}_{C} &= \frac{1}{n}\sum_{C} Y_i = \hat{\beta}_0 + \hat{\beta}_2 \frac{1}{n}\sum_{C} D_i^{C} + \frac{1}{n}\sum_{C} \hat{\nu}_i = \hat{\beta}_0 + \hat{\beta}_2
\end{aligned}$$

(d) The difference is clear from the coefficients. In model (1) the coefficients reflect the average income for the various groups. In model (2) the coefficients reflect the difference in average income between the group and high school drop outs.

Question 3

(a) In order to interpret as elasticities we simply take the log of all variables. We have

Table 4: Q3 Part (a)

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
Intercept	4.62092***	(0.25441)	18.163	<2e-16
log(sales)	0.16213***	(0.03967)	4.087	6.67e-05
log(mktval)	0.10671*	(0.05012)	2.129	0.0347

Note: *** significant at the 0.1% level ** significant at the 1% level
* significant at the 5% level

(b) The following table gives us a comparison of running regression with profit at level and with log profits.

Table 5: Q3 Part (b)

	profit (n=177)			log(profit) (n=168)		
Variables	Estimate	Stand. Error	$\Pr(>\|t\|)$	Estimate	Stand. Error	$\Pr(>\|t\|)$
Intercept	4.687e+00***	(3.797e-01)	<2e-16	4.21187***	(0.3312)	<2e-16
log(sales)	1.614e-01***	(3.991e-02)	7.92e-05	0.21142 ***	(0.04893)	2.68e-05
log(mktval)	9.753e-02	(6.369e-02)	0.128	0.17538*	(0.07016)	0.0134
profits	3.566e-05	(1.520e-04)	0.815	-0.10281	(0.07139)	0.1517

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

Note that for log profits we had to remove the negative profit observations since the log does not exist. However, this should not be a big concern since only 9 out of 177 observations were removed. As we can see above, using profits at level reduces all coefficients and makes it hard to interpret. Therefore, if we want to interpret as elasticities it makes more sense to use logs rather than levels.

(c) The correlation matrix between profits and all the other variables is as follows

$$\begin{array}{cccccccccc} & salary & age & college & grad & comten & ceoten & sales & mktval & profmarg \\ profits & 0.3939 & 0.1147 & -0.0459 & 0.0978 & 0.1437 & -0.0216 & 0.7983 & 0.9181 & 0.1255 \end{array}$$

We see that profit is highly correlated with sales and market value. This gives rise to the concern of perfect multicollinearity. High degree of multicollinearity causes the matrix inversion to be almost singular. As a result, our statistical software may not obtain numerically accurate estimates. Also, multicollinearity makes our regression highly unstable.

(d) Running the desired regression gives us

Table 6: Q3 Part (d)

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
Intercept	4.0146940***	0.3305285	12.146	<2e-16
log(sales)	0.2116452***	0.0478575	4.422	1.78e-05
log(mktval)	0.1674818*	0.0687439	2.436	0.01592
log(profits)	-0.0913613	0.0699629	-1.306	0.19345
ceoten	0.0417268**	0.0142714	2.924	0.00395
ceoten^2	-0.0011272*	0.0004759	-2.369	0.01903

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

We can see that an additional year of CEO tenure increases salary by $0.0417 + (-0.0011)^2 \approx 4.17\%$.

Question 4

(a) Results are reported below

Table 7: Q4 Part (a)

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	-0.6357	(1.0854)	-0.586	0.558
age	0.5852***	(0.0362)	16.165	<2e-16
female	-3.6640***	(0.2107)	-17.391	<2e-16
bachelor	8.0830***	(0.2088)	38.709	<2e-16

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

(b) Adding the interaction term we have

Table 8: Q4 Part (b)

	Estimate	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	-0.68442	(1.08630)	-0.630	0.529
age	0.58401***	(0.03622)	16.125	<2e-16
female	-3.43375***	(0.29768)	-11.535	<2e-16
bachelor	8.28387***	(0.27795)	29.804	<2e-16
fem×bach	-0.46136	(0.42135)	-1.095	0.274

Note: *** significant at the 0.1% level ** significant at the 1% level * significant at the 5% level

The coefficient on the interaction term captures the *difference* in the effect of acquiring a bachelor degree for women versus men. In other words, it tells us that a bachelor degree rewards men better than women. (However, note that the coefficient is not statistically significant so the gender specific impact of college education should not be substantial).

(c) We simply read off the results of the coefficient from the R-output. We find that we cannot reject the null that the coefficient is 0 at the 5 or 10% level. Only at 11.9% would we reject the null hypothesis that the coefficient is 0.

(d) Again, I summarize the results below. We can interpret the coefficients for age×bach as the *difference* in the effect of an additional year for college graduates versus nongraduates.

Table 9: Q4 Part (d)

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	7.79166***	(1.50259)	5.185	2.21e-07
female	-3.43495***	(0.29643)	-11.588	<2e-16
bachelor	-9.11482***	(2.15854)	-4.223	2.44e-05
fem×bach	-0.35311	(0.41979)	-0.841	0.4
age	0.29745***	(0.05044)	5.897	3.85e-09
age×bach	0.58640***	(0.07215)	8.127	5.07e-16

Note: *** significant at the 0.1% level ** significant at the 1% level
* significant at the 5% level

(e) Testing whether $\beta_4 = \beta_5$ is equivalent to testing a linear restriction on all coefficients. In other words, we can do the following hypothesis testing

$$H_o : c'\beta = 0 \text{ vs } H_a : c'\beta < 0$$
$$\text{where } \beta = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5)' \text{ and } c = (0\ 0\ 0\ 0\ 1\ -1)'$$

This is convenient because now we can apply the multivariate version of the CLT to derive the limiting distribution of $c'\beta$. Note that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$
$$c'[\sqrt{n}(\hat{\beta}_n - \beta)] = \sqrt{n}(c'\hat{\beta}_n - c'\beta) \xrightarrow{d} c'\mathcal{N}(0, \Omega) = \mathcal{N}(0, c'\Omega c)$$
$$\frac{\sqrt{n}(c'\hat{\beta}_n - c'\beta)}{\sqrt{c'\Omega c}} \xrightarrow{d} \mathcal{N}(0, 1)$$

So a natural test statistic is

$$T_n = \frac{c'\hat{\beta}_n - 0}{\sqrt{c'\widehat{\Omega}_n c}}$$

where $\widehat{\Omega}_n$ is the sample variance-covariance matrix for the estimated coefficients. Note that when computing $\widehat{\Omega}_n$, the sample size n is already taken into consideration. Plugging in these values in R we get $|T_n| = 2.55$, which rejects the null hypothesis at the 1% level.