Problem Set 4 Solutions

ECON 210 Econometrics A

Evan Zuofu Liao*

November 5, 2015

Question 1 (2.7 W)

(i) $\mathbb{E}[u|inc] = \mathbb{E}[e\sqrt{inc}|inc] = \sqrt{inc}\,\mathbb{E}[c|inc] = \sqrt{inc}\,\mathbb{E}[c] = 0$
(ii) $\text{Var}[u|inc] = \text{Var}[e\sqrt{inc}|inc] = inc\,\text{Var}[e|inc] = inc\,\text{Var}[e] = \sigma_e^2 inc$
(iii) Low income families do not have much leeway in spending. They have to spend a large proportion of their income on living expenses such as food and clothing. On the other hand, higher income families have more leeway in deciding how much to spend on consumption and how much to saving. This suggests wider variability in saving among higher income families.

Question 2 (2.8 W)

(i) From equation (2.66) in Wooldridge we have

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \tag{1}$$

$$= \frac{\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n x_i^2} \tag{2}$$

$$= \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1 + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2} \tag{3}$$

Taking the conditional expectation we find

$$\mathbb{E}[\tilde{\beta}_1|x_1, ..., x_n] = \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1$$

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

Thus the bias is zero when $\beta_0 = 0$ or $\sum_{i=1}^n x_i = 0$. Note that in the latter case $\sum_{i=1}^n x_i = 0$ is the same as $\bar{x}_n = 0$, which suggests that regression through the origin is identical to regression with an intercept.

(ii) From equation (3) in part (i), we know

$$\begin{aligned}
\text{Var}\left[\tilde{\beta}_1 | x_1, ..., x_n\right] &= \frac{\text{Var}\left[\sum_{i=1}^n x_i u_i | x_1, ..., x_n\right]}{(\sum_{i=1}^n x_i^2)^2} \\
&= \frac{\sum_{i=1}^n x_i^2 \text{Var}\left[u_i | x_1, ..., x_n\right]}{(\sum_{i=1}^n x_i^2)^2} \\
&= \frac{1}{\sum_{i=1}^n x_i^2} \text{Var}\left[u_i | x_1, ..., x_n\right] \\
&= \frac{\sigma_u^2}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

(iii) From equation (2.57) in Wooldridge, we have (conditional on sample $x_1, ..., x_n$)

$$\text{Var}\left[\hat{\beta}_1\right] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Given the hint, we can easily see that $\text{Var}\left[\tilde{\beta}_1\right] \leq \text{Var}\left[\hat{\beta}_1\right]$ unless $\bar{x}_n = 0$

(iv) For a fixed sample size, the bias of $\tilde{\beta}_1$ increases as $\bar{x}_n$ increases. However as $\bar{x}_n$ increases the variance of $\hat{\beta}_1$ increases relative to the variance of $\tilde{\beta}_1$. So there is a trade-off between bias and variance. Also, the bias of $\tilde{\beta}_1$ is small when β_0 is small. Thus choosing between $\tilde{\beta}_1$ and $\hat{\beta}_1$ requires information on the size of $\bar{x}_n$, β_0 and n

Question 3 (2.9 W)

(i) Replacing y_i and x_i with $c_1 y_i$ and $c_2 x_i$ in the estimators for the slope and the intercept, and noting that $\overline{c_1 y}_n = c_1 \bar{y}_n$ and $\overline{c_2 x}_n = c_2 \bar{x}_n$, we have

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n (c_1 y_i - c_1 \bar{y}_n)(c_2 x_i - c_2 \bar{x}_n)}{\sum_{i=1}^n (c_2 x_i - c_2 \bar{x}_n)^2} \\
&= \frac{\sum_{i=1}^n c_1 c_2 (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n c_2^2 (x_i - \bar{x}_n)^2} \\
&= \frac{c_1}{c_2} \hat{\beta}_1 \\
\tilde{\beta}_0 &= c_1 \bar{y}_n - \tilde{\beta}_1 c_2 \bar{x}_n = c_1 \bar{y}_n - \frac{c_1}{c_2} c_2 \bar{x}_n = c_1 \hat{\beta}_0
\end{aligned}$$

(ii) We use the same approach as in part (i). Again, note that $\overline{(c_1 + y_i)} = c_1 + \bar{y}_n$ and $\overline{(c_2 + x_i)} = c_2 + \bar{x}_n$. Then we know

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n (c_1 + y_i - (c_1 + \bar{y}_n))(c_2 + x_i - (c_2 + \bar{x}_n))}{\sum_{i=1}^n (c_2 + x_i - (c_2 + \bar{x}_n))^2} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\
&= \hat{\beta}_1 \\
\tilde{\beta}_0 &= c_1 + \bar{y}_n - \tilde{\beta}_1 (c_2 + \bar{x}_n) \\
&= c_1 - \hat{\beta}_1 c_2 + \bar{y}_n - \hat{\beta}_1 \bar{x}_n \\
&= c_1 - \hat{\beta}_1 c_2 + \hat{\beta}_0
\end{aligned}$$

as desired.

Question 4 (2.13 W)

(i) We can compute that the average salary is \$957.95 and the average IQ is about 101.28. The standard deviation of IQ is about 15.05.

(ii) We want to estimate the following model

$$Wage = \beta_0 + \beta_1 IQ + U$$

Regression results give us $\hat{\beta}_1 = 8.3$ and $\hat{\beta}_0 = 116.99$. Thus an increase in IQ by 15 increases predicted monthly wage by $15(8.3) = \$124.5$. R^2 obtained from this regression is 0.096, suggesting that IQ score can explain less than 10% of the variation in monthly wage.

(iii) Now the model of interest becomes

$$\log Wage = \beta_0^* + \beta_1^* IQ + U^*$$

Running this regression gives us $\hat{\beta}_1^* = 0.0088$ and $\hat{\beta}_0^* = 5.89$. So the percentage change in monthly wages resulting from an increase in IQ by 15 is about $15(.0088) = .132 = 13.2\%$.

Question 5

(a) Adding the data point $(-Y_i, -X_i)$ for each (Y_i, X_i) to the dataset implies that sample means for the new dataset are zero, i.e. $\bar{Y}_{2n} = 0$ and $\bar{X}_{2n} = 0$. Thus we have

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\sum_{j=1}^{2n}(Y_j - \bar{Y}_{2n})(X_j - \bar{X}_{2n})}{\sum_{j=1}^{2n}(X_j - \bar{X}_{2n})^2} \\ &= \frac{2\sum_{i=1}^{n} X_i Y_i}{2\sum_{i=1}^{n} X_i^2} \\ &= \hat{\beta}\end{aligned}$$

(b) From part (a) we know that $\hat{\alpha}_1 = \hat{\beta}$ and $\hat{\alpha}_0 = \bar{Y}_{2n} - \hat{\alpha}_1 \bar{X}_{2n} = 0$. So we have

$$\begin{aligned}\sum_{j=1}^{2n} \hat{\epsilon}_i^2 &= \sum_{j=1}^{2n}(Y_j - \hat{\alpha}_0 - \hat{\alpha}_1 X_j)^2 \\ &= \sum_{j=1}^{2n}(Y_j - \hat{\beta} X_j)^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{\beta} X_i)^2 + \sum_{i=1}^{n}(-Y_i + \hat{\beta} X_i)^2 \\ &= 2\sum_{i=1}^{n} \hat{u}_i^2\end{aligned}$$

(c) From Q2, we know that the variances of $\hat{\alpha}_1$ and $\hat{\beta}$ are

$$\begin{aligned}\text{Var}\left[\hat{\alpha}_1\right] &= \frac{\sigma_\epsilon^2}{\sum_{j=1}^{2n}(X_j - \bar{X}_{2n})^2} = \frac{\sigma_\epsilon^2}{\sum_{j=1}^{2n} X_j^2} = \frac{\sigma_\epsilon^2}{2\sum_{i=1}^{n} X_i^2} \\ \text{Var}\left[\hat{\beta}\right] &= \frac{\sigma_u^2}{\sum_{i=1}^{n} X_i^2}\end{aligned}$$

(d) In order to compare the variances of the slope estimators, we have to first estimate the error variances. Given $X_1, ..., X_n$, we can use the following unbiased estimators

$$\begin{aligned}\hat{\sigma}_\epsilon^2 &= \frac{1}{2n-2}\sum_{j=1}^{2n} \hat{\epsilon}_j^2 = \frac{1}{2n-2}\sum_{j=1}^{2n}(Y_j - \hat{\alpha}_0 - \hat{\alpha}_1 X_i)^2 \\ \hat{\sigma}_u^2 &= \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta} X_i)^2\end{aligned}$$

From part (b) we know that $|\hat{\epsilon}_j| = \hat{u}_i$. This implies that $\hat{\epsilon}_j^2 = \hat{u}_i^2$. Therefore, we have

$$\frac{\widehat{\text{Var}}\left[\hat{\alpha}_1\right]}{\widehat{\text{Var}}\left[\hat{\beta}\right]} = \frac{\hat{\sigma}_\epsilon^2/2}{\hat{\sigma}_u^2} = \frac{\frac{2}{2(2n-2)}\sum_{i=1}^n \hat{u}_i^2}{\frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2} = \frac{n-2}{2(n-1)} \leq \frac{1}{2}, \text{ for } n > 2$$

Clearly the variance of $\hat{\beta}$ is larger. This result is intuitive since we have more data points in our estimation of α_1, thus we expect the variance to be smaller.

Question 6

(a) I report the results from the dummy regression below

```
Call:
lm(formula = y ~ d)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7731 -0.6385 -0.4654  0.1115  1.7654

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.29231    0.05625 325.185  < 2e-16 ***
d            0.64887    0.22707   2.858  0.00459 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.907 on 275 degrees of freedom
Multiple R-squared:  0.02884,   Adjusted R-squared:  0.02531
F-statistic: 8.166 on 1 and 275 DF,  p-value: 0.004595
```

Testing the null hypothesis that the means are equal is equivalent to testing the significance of the coefficient. We find that the t-value is 2.858. Thus we may reject the null hypothesis at the 1% significance level.

(b) The big mistake is that we have too few observations for post-signage fatalities.

(c) Running the requested regression we have

```
Call:
lm(formula = y ~ d)

Residuals:
         1          2          3          4          5          6
 9.180e+01 -4.020e+01 -2.420e+01 -3.320e+01  5.800e+00 -7.105e-15

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   951.20      24.25  39.219 2.53e-06 ***
d              33.74      59.41   0.568      0.6
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 54.23 on 4 degrees of freedom
Multiple R-squared:  0.07462,   Adjusted R-squared:  -0.1567
F-statistic: 0.3226 on 1 and 4 DF,  p-value: 0.6004
```

Not surprisingly, the problem is that we have only 1 data point for the treatment group. Projecting out is also a problem.

(d) Here's the new table from the simulated numbers

Year	Fatalities
2008	967
2009	919
2010	929
2011	965
2012	984
2013	330

(e) Running the regression with the new dataset gives us

```
Call:
lm(formula = data$deaths ~ data$signs)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-9.4118 -2.3231 -0.3231  1.6769 11.6769

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.323      0.219  83.670   <2e-16 ***
data$signs     1.089      0.884   1.232    0.219
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.531 on 275 degrees of freedom
Multiple R-squared:  0.005485,  Adjusted R-squared:  0.001869
F-statistic: 1.517 on 1 and 275 DF,  p-value: 0.2192
```

(f) With the two additional explanatory variables, the regression gives us

```
Call:
lm(formula = data$deaths ~ data$signs + data$tourists + data$icy_days)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9476 -1.0207  0.1134  1.2307  6.4729

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.3342938  0.5964442  25.710  < 2e-16 ***
data$signs    -2.0867899  0.4806180  -4.342 1.99e-05 ***
data$tourists  0.0015234  0.0006728   2.264   0.0243 *
data$icy_days  2.1230048  0.1058937  20.048  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.852 on 273 degrees of freedom
Multiple R-squared:  0.7284,    Adjusted R-squared:  0.7254
F-statistic:   244 on 3 and 273 DF,  p-value: < 2.2e-16
```

(g) It is good to include these variables since we think they might be correlated with both fatalities and signage. In other words, excluding these variable from the regression might result in estimation of the total impact of signage rather than its partial impact. Adding

them would allow us to get the variation in traffic deaths that is specifically linked to the signs rather than ice or weather.

(h) The change in sign confirms our suspicion that the dummy regression captures the total derivative of fatalities with respect to signage rather than its partial derivative with respect to signage. In other words, the indirect impact of signage on fatalities through icy days outweighs its direct impact on fatalities. Since signage is positively correlated with icy days, and icy days are also positively correlated with fatalities, we see that the dummy regression gives us a positive sign.

To see that signage is positively correlated with icy days, look at the plot below. What you can see from the plot is that the number of icy days which is plotted in the y-axis is highly correlated with early-in-the-year week data. Since the post-signage data is observed from the first 17 weeks of 2013 and that time of the year is typically associated with freezing winter, so we know that we have a sample selection bias! This explains why adding the variable of icy days when our data for the treatment group is focused early in the year sways our results that much.