Problem Set 1

Due: 3/6

Part One: Hand-Written Exercise

1. Verify the statement on slide 23, Lecture 1. That is, suppose $y_i = \beta_0 + \beta_1 x_i + u_i$, please show that the OLS estimators are:
   (a) $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
   (b) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

2. Consider the following regression models:

$$\begin{aligned} &\text{Model A:}\ \ y_i = \beta_0 + \beta_1 x_i + u_i \\ &\text{Model B:}\ \ y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + v_i \end{aligned}$$

   where $\bar{x} = \frac{1}{n}\sum x_i$, and $\text{Var}(y_i) = \sigma^2$.

   (a) Find the OLS estimators of β_0 and α_0. Are they identical? Are their variances identical? If not, which variance is larger?
   (b) Find the OLS estimators of β_1 and α_1. Are they identical? Are their variances identical? If not, which variance is larger?

3. (a) Show $SST = SSR + SSE$ when there is an intercept term in the regression.
   (b) Show SST need not be equal to $SSR + SSE$ when there is no intercept term.

Part Two: Computer Exercise

1. (a) Let $x = c(1:150)$
   (b) Select the number in x that is greater than 135 or smaller or equal to 5.
   (c) Select the number in x that is greater than 70 and smaller than 90.
   (d) Select the number in x that is divisible by 4 and 5

2. (a) Draw 150,000 observations from standard normal distribution and name it as "X"
   (b) Evaluate the mean, median, max, min, and variance of X.
   (c) Randomly select 5,000 subsamples from X without replacement, call it Y and calculate its mean and variance.

Problem Set 1: Solution

Part One: Hand-Written Exercise

1. The least-squares(LS) criterion function is

$$Q_n(\beta_0,\beta_1) := \textstyle\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i)^2$$

The first order conditions(FOCs) are

$$\frac{\partial Q_n(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i) = 0 \tag{1}$$

$$\frac{\partial Q_n(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i)x_i = 0 \tag{2}$$

By (1)(2),

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

$$\sum_{i=1}^n x_i\beta_0 + \sum_{i=1}^n x_i^2\beta_1 = \sum_{i=1}^n x_i y_i \tag{4}$$

By (3)(4), we can get

$$\hat\beta_1 = \frac{\sum_{i=1}^n (y_i-\bar y)(x_i-\bar x)}{\sum_{i=1}^n (x_i-\bar x)^2}$$

$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x. \blacksquare$$

2. We already know that for model A:

$$\hat\beta_0 = \bar y - \hat\beta_1\bar x$$
$$\hat\beta_1 = \frac{\sum (x_i-\bar x)(y_i-\bar y)}{\sum (x_i-\bar x)^2}$$

Now for model B, from the F.O.C. of $\sum \left(y_i-\alpha_0-\alpha_1(x_i-\bar x)\right)^2$ we have:

$$\begin{cases} -2\sum\left(y_i-\hat\alpha_0-\hat\alpha_1(x_i-\bar x)\right)=0 \\ -2\sum\left(y_i-\hat\alpha_0-\hat\alpha_1(x_i-\bar x)\right)(x_i-\bar x)=0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum y_i = n\hat\alpha_0 + \hat\alpha_1\sum(x_i-\bar x) \\ \sum y_i(x_i-\bar x) = \hat\alpha_0\sum(x_i-\bar x) + \hat\alpha_1\sum(x_i-\bar x)^2 \end{cases}$$

$\hat{\alpha}_0$ and $\hat{\alpha}_1$ is therefore given by:

$$\begin{aligned}
\hat{\alpha}_0 &= \bar{y} \\
\hat{\alpha}_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

(a) $\hat{\alpha}_0$ and $\hat{\beta}_0$ are not identical, and their variance is given by:

$$\begin{aligned}
\mathrm{Var}(\hat{\alpha}_0) &= \mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n} \\
\mathrm{Var}(\hat{\beta}_0) &= \frac{\sigma^2}{n} \cdot \frac{\sum x_i^2}{\sum(x_i - \bar{x})^2}.
\end{aligned}$$

Since for any sample of data, $\sum x_i^2 \geq \sum(x_i - \bar{x})^2$ (please verify), hence $\mathrm{Var}(\hat{\beta}_0) \geq \mathrm{Var}(\hat{\alpha}_0)$.

(b) $\hat{\alpha}_1$ and $\hat{\beta}_1$ are identical. ■

3. (a)

$$\begin{aligned}
\because \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} \hat{u}_i \hat{y}_i - \sum_{i=1}^{n} \hat{u}_i \bar{y} \\
&= \sum_{i=1}^{n} \hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \sum_{i=1}^{n} \hat{u}_i \bar{y} \\
&= \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{u}_i x_i}_{=0} - \bar{y} \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{=0} \\
&= 0 \\
\therefore \sum_{i=1}^{n} (\hat{u}_i + \hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \blacksquare
\end{aligned}$$

(b)

$$
\begin{aligned}
\because \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} \hat{u}_i\hat{y}_i - \sum_{i=1}^{n} \hat{u}_i\bar{y} \\
&= \sum_{i=1}^{n} \hat{u}_i(\hat{\beta}_1 x_i) - \sum_{i=1}^{n} \hat{u}_i\bar{y} \\
&= \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{u}_i x_i}_{=0} - \bar{y} \underbrace{\sum_{i=1}^{n} \hat{u}_i}_{\neq 0} \\
&\neq 0 \\
\therefore \sum_{i=1}^{n}(\hat{u}_i + \hat{y}_i - \bar{y})^2 &\neq \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \blacksquare
\end{aligned}
$$

4.

$$
\begin{aligned}
\because Cov(\bar{y}, \hat{\beta}_1) &= Cov\left(\frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} Cov\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i(x_i - \bar{x})\right) \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x}) Cov(\sum_{j=1}^{n} y_j, y_i) \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} \sigma^2 \\
&= 0 \\
\therefore Var(\hat{\beta}_0) &= Var(\bar{y} - \hat{\beta}_1\bar{x}) \\
&= Var(\bar{y}) + Var(\hat{\beta}_1\bar{x}) \\
&= \sigma_0^2 \frac{\sum_{i=1}^{n} x_i^2/n}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \blacksquare
\end{aligned}
$$

(d) Randomly select 5,000 subsamples from X with replacement, call it Z and calculate its mean and variance.

(e) Find the 45^{th} percentile in X. Also, find the number z such that $Pr(a \leq z) = 0.45$, where $a \sim N(0, 1)$.

(f) Find the probability of drawing $x \in X$ such that $x \in (-0.55, 1.25]$. Also, find the probability of drawing a, where $a \sim N(0, 1)$ such that $a \in (-0.55, 1.25]$.

3. (a) Create matrix $\mathbf{X} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 7 \\ 1 & 8 & 10 \\ 1 & 10 & 13 \\ 1 & 12 & 16 \end{bmatrix}$

(b) Create matrix $\mathbf{Y} = \begin{bmatrix} 1 & 9 \\ 2 & 8 \\ 3 & 7 \\ 4 & 6 \\ 5 & 5 \\ 6 & 4 \end{bmatrix}$

(c) Create matrix $\mathbf{Z}$, a $6 * 3$ matrix, where

$$Z_{ij} = \begin{cases} X_{i1} + Y_{i1}, & \text{if } j = 1 \\ X_{i2}, & \text{if } j = 2 \\ X_{i3} - 2 * Y_{i2}, & \text{if } j = 3 \end{cases} \quad \text{,for } i = 1, 2, ..., 6$$

Spring 2023

Quantitative Analysis
Chung-Ming Kuan

Problem Set 2

Due: 3/13

Part One: Hand-Written Exercise

1. Verify the statement in slide 18, Lecture 2. That is, let $\hat{r}_{i,1}$ be the OLS residual of regressing x_1 on the constant one and $x_2, ..., x_k$. Show that $\sum_{i=1}^n \hat{r}_{i,1} x_{i,1} = \sum_{i=1}^n \hat{r}_{i,1}^2$.

2. For the multiple linear regression, the data matrix denoted as $\mathbf{X}$ is below:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 1 & 8 \\ 2 & 4 & 5 & 7 \\ 3 & 6 & 2 & 9 \\ 4 & 8 & 2 & 2 \end{bmatrix}$$

For this data matrix, can you calculate the OLS estimators? Why or why not? Please give a brief explanation.

3. Consider the model $y_i = \beta_0 + \beta_1 x_i + u_i$ with $\text{Var}(y_i) = \sigma^2$. Under the Classical Assumptions, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Let $\tilde{\beta}_1$ be the OLS estimator of β_1 by assuming the intercept is zero. That is, $\tilde{\beta}_1$ is obtained under the assumption $\beta_0 = 0$.

   (a) Calculate $\mathbb{E}(\tilde{\beta}_1)$ in terms of x_i, β_0, and β_1.

   (b) If $\beta_0 \neq 0$, is $\tilde{\beta}_1$ unbiased?

   (c) Calculate the variance of $\tilde{\beta}_1$.

   (d) Compare between $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$. Is it true that $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$ in general?

   (e) Does the result in (d) violate the Gauss-Markov Theorem, which states that $\hat{\beta}_1$ should have the smallest variance? Explain.

Part Two: Computer Exercise

1. Let $\mathbf{X} = \begin{bmatrix} 7 & 2 & 3 \\ 4 & 6 & 7 \\ 9 & 2 & 0 \\ 0 & 9 & 0 \\ 5 & 3 & 5 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} 6 \\ 2 \\ 4 \\ 2 \\ 1 \end{bmatrix}$.

   (a) Please construct the OLS estimator $\hat{\beta}$. (Reminder: Don't forget the intercept term.)

   (b) Given a new observation $x^* = (0, 4, 3)'$, please calculate $\hat{y}$.

2. Please load the data set "mtcars" from R using the code `data(mtcars)`. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

   (a) Please show the data for the automobile "Duster 360".

   (b) Please show the `qsec` (1/4 mile time) for all the automobile.

   (c) Please show the data with `cyl` (number of cylinders) = 6.

   (d) Please list the automobiles with `mpg` (miles/gallon) > 15, `vs` (Engine) = 1, and `hp` (horsepower) between 50 and 150.

   (e) Suppose we have the following model:

$$\texttt{drat}_i = \beta_0 + \beta_1 \texttt{wt}_i + \beta_2 \texttt{hp}_i + \beta_3 \texttt{qsec}_i + \beta_4 \texttt{vs}_i + u_i.$$

   Please find $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 without the function `lm()`.

   (f) Following (e), find those estimators with the function `lm()`.

Problem Set 2: Solution

Part One: Hand-Written Exercise

1. Since $\hat{r}_{i,1}$ is the OLS residual of regressing x_1 on the constant one and $x_2, ..., x_k$, we can write:

$$x_{i,1} = \hat{c}_0 + \hat{c}_2 x_{i,2} + ... + \hat{c}_k x_{i,k} + \hat{r}_{i,1},$$

where $\hat{c}_j$ are the OLS estimates. We now continue with the proof:

$$\begin{aligned} \sum \hat{r}_{i,1}^2 &= \sum \hat{r}_{i,1} \left(x_{i,1} - \hat{c}_0 - \hat{c}_2 x_{i,2} - ... - \hat{c}_k x_{i,k} \right) \\ &= \sum \hat{r}_{i,1} x_{i,1} - \hat{c}_0 \sum \hat{r}_{i,1} - \hat{c}_2 \sum \hat{r}_{i,1} x_{i,2} - ... - \hat{c}_k \sum \hat{r}_{i,1} x_{i,k} \\ &= \sum \hat{r}_{i,1} x_{i,1}, \end{aligned}$$

by the fact that $\sum \hat{r}_{i,1} = \sum \hat{r}_{i,1} x_{i,2} = ... = \sum \hat{r}_{i,1} x_{i,k} = 0$. ■

2. We can not calculate the OLS estimators because there is exact multicollinearity among regressors, that is, the 2^{nd} column equals the 1^{st} column times 2. Therefore, $(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$ doesn't exist.

3. (a)

$$\tilde{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$
$$\mathbb{E}(\tilde{\beta}_1) = \frac{\sum x_i \mathbb{E}(y_i)}{\sum x_i^2} = \beta_0 \cdot \frac{\sum x_i}{\sum x_i^2} + \beta_1.$$

(b) If $\beta_0 \neq 0$, then $\tilde{\beta}_1$ is biased iff $\sum x_i \neq 0$.

(c)

$$\mathrm{Var}(\tilde{\beta}_1) = \mathrm{Var}\left(\frac{\sum x_i y_i}{\sum x_i^2} \right) = \mathrm{Var}\left(\frac{\sum x_i u_i}{\sum x_i^2} \right) = \frac{\sigma^2}{\sum x_i^2}$$

(d)

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

As $\sum x_i^2 \geq \sum (x_i - \bar{x})^2$ for any sample of data, $\mathrm{Var}(\hat{\beta}_1) \geq \mathrm{Var}(\tilde{\beta}_1)$ in general.

(e) No, since $\tilde{\beta}_1$ is NOT unbiased in general. In the case where $\tilde{\beta}_1$ is unbiased, then we have $\bar{x} = 0$, causing $\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\tilde{\beta}_1)$. ■

Spring 2023

Quantitative Analysis
Chung-Ming Kuan

Problem Set 3

Due: 03/20

Part One: Hand-Written Exercise

1. We mentioned that the F statistic is given by:

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)},$$

where SSR_r and SSR_{ur} are the residual sums of squares of restricted and unrestricted regressions respectively. $(\mathrm{SSR}_r - \mathrm{SSR}_{ur})$ and SSR_{ur} are independent of each other.

(a) Given the fact that:

$$\frac{(n-k-1+q)\hat{\sigma}_r^2}{\sigma^2} - \frac{(n-k-1)\hat{\sigma}_{ur}^2}{\sigma^2} \sim \chi^2(q),$$

where $\hat{\sigma}_r^2$ and $\hat{\sigma}_{ur}^2$ are the OLS estimators of σ^2 of the restricted and unrestricted regressions respectively. Please show that

$$\frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)} \sim F(q, n-k-1).$$

(b) Show that the F statistic can also be written as the R-squared form

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)},$$

where R_r^2 and R_{ur}^2 are the R^2s of the restricted and unrestricted regressions.

2. Abby and Bob are trying to understand the difference of the health expenditure, y, of a smoker and a non-smoker with different models. Abby adopts Model A while Bob adopts Model B:

$$\text{Model A: } E[y] = \beta_0 + \beta_1 x_1, \text{ where } x_1 = \begin{cases} 1, & \text{for smokers,} \\ 0, & \text{for non-smokers} \end{cases}$$
$$\text{Model B: } E[y] = \alpha_0 + \alpha_1 x_2, \text{ where } x_2 = \begin{cases} 0, & \text{for smokers,} \\ 1, & \text{for non-smokers} \end{cases}$$

(a) Please express β_0 and β_1 with α_0 and α_1.

(b) Are predictions, $\hat{y}$, the same for Model A and Model B? Discuss both $\hat{y}$ for a smoker and a non-smoker.

(c) Chris combines Model A and Model B and get Model C:

$$\text{Model C:} \;\; E[y] = \delta_0 + \delta_1 x_1 + \delta_2 x_2,$$

$$\text{where } x_1 = \begin{cases} 1, & \text{for smokers,} \\ 0, & \text{for non-smokers} \end{cases}, \; x_2 = \begin{cases} 0, & \text{for smokers,} \\ 1, & \text{for non-smokers} \end{cases}$$

Chris claims that Model C has more explaining power than Model A and Model B since it includes more explanatory variables. Is his statement true? Explain it.

3. The following model can be used to study whether campaign expenditures affect election outcomes:

$$\text{voteA} = \beta_0 + \beta_1 ln(\text{expendA}) + \beta_2 ln(\text{expendB}) + \beta_3 \text{prtystrA} + u,$$

where "voteA" is the percentage of the vote received by candidate A, "expendA" and "expendB" are campaign expenditures by candidates A and B, and "prtystrA" is a measure of party strength for candidate A (the percentage of the most recent presidential vote that went to A's party).

(a) What is the interpretation of β_1?

(b) In terms of the parameters, state the null hypothesis that the effect of the increase in A's expenditure will be offset by the increase in B's expenditure.

(c) Write the detailed procedure to do the hypothesis testing in (b).

(d) If someone claims that both candidates' expenditures do not have any effect on the outcome, how can you specify a testing null hypothesis?

(e) Write the detailed procedure to do the hypothesis testing in (d).

Part Two: Computer Exercise

Following Question 2 of the computer exercise in Problem Set 2, consider the following model:

$$\texttt{drat}_i = \beta_0 + \beta_1 \texttt{wt}_i + \beta_2 \texttt{hp}_i + \beta_3 \texttt{qsec}_i + \beta_4 \texttt{vs}_i + u_i,$$

1. Test the hypothesis $H_0 : \beta_1 = 0$.

(a) Please construct the t statistic without the function `lm()`.

(b) Use the function `lm()` to directly obtain the t statistic. Verify that it's identical to (a).

2. Test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$.

   (a) Please construct the constrained and unconstrained model, obtain R^2_{ur} and R^2_r and construct the F statistic.

   (b) Instead of R^2_{ur} and R^2_r, please obtain SSR_{ur} and SSR_r and recalculate the F statistic. Verify that it's identical to (a).

   (c) Use the function `linearHypothesis()` to directly obtain the F statistic. Verify that it's identical to (a).

Problem Set 3: Solution

Part One: Hand-Written Exercise

1. (a) Let

$$\hat{\sigma}_r^2 = \frac{1}{n-(k+1-q)}\sum_{1=1}^{n}\hat{e}_{i,r}^2 = \frac{SSR_r}{n-(k+1-q)}$$
$$\hat{\sigma}_{ur}^2 = \frac{1}{n-(k+1)}\sum_{1=1}^{n}\hat{e}_{i,ur}^2 = \frac{SSR_{ur}}{n-(k+1)},$$

where $\hat{e}_{i,r}^2$ and $\hat{e}_{i,ur}^2$ are the residuals from restricted and unrestricted models respectively. The fact that

$$\frac{(n-k-1+q)\hat{\sigma}_r^2}{\sigma^2} - \frac{(n-k-1)\hat{\sigma}_{ur}^2}{\sigma^2} \sim \chi^2(q)$$

implies

$$\frac{SSR_r - SSR_{ur}}{\sigma^2} \sim \chi^2(q).$$

Moreover, we know that

$$\frac{(n-k-1)\hat{\sigma}_{ur}^2}{\sigma^2} = \frac{SSR_{ur}}{\sigma^2} \sim \chi^2(n-k-1).$$

Finally, since $(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q$ and $\mathrm{SSR}_{ur}/(n-k-1)$ are independent of each other (proof omitted), we have the following result:

$$\frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)} = \frac{SSR_r - SSR_{ur}}{\sigma^2 q}\Big/\frac{SSR_{ur}}{\sigma^2(n-k-1)}$$
$$\sim \frac{\chi^2(q)/q}{\chi^2(n-k-1)/(n-k-1)} \sim F(q, n-k-1). \quad \blacksquare$$

(b)

$$F = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k-1)} = \frac{(\frac{\mathrm{SSR}_r}{\mathrm{SST}} - \frac{\mathrm{SSR}_{ur}}{\mathrm{SST}})/q}{\frac{\mathrm{SSR}_{ur}}{\mathrm{SST}}/(n-k-1)}$$
$$= \frac{(1-R_r^2-1+R_{ur}^2)/q}{(1-R_{ur}^2)/(n-k-1)} = \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}. \quad \blacksquare$$

2. (a) Use the fact that $x_2 = 1 - x_1$ and rewrite Model B as

$$E[y] = \alpha_0 + \alpha_1 x_2 = \alpha_0 + \alpha_1(1 - x_1) = (\alpha_0 + \alpha_1) - \alpha_1 x_1$$

Compare with Model A, we can see that $\beta_0 = \alpha_0 + \alpha_1$ and $\beta_1 = -\alpha_1$.

(b) The fitted value of a smoker is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(1) = (\hat{\alpha}_0 + \hat{\alpha}_1) + (-\hat{\alpha}_1) = \hat{\alpha}_0 = \hat{\alpha}_0 + \hat{\alpha}_1(0)$$

The fitted value of a non-smoker is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(0) = \hat{\beta}_0 = \hat{\alpha}_0 + \hat{\alpha}_1$$

Thus, the predictions are the same for Model A and Model B.

(c) No, it's not true. We only need one dummy variable for two levels. Including x_1 and x_2 in the same model will cause the issue of collinearity, and make the equation unsolvable.

3. (a) An 1% increase in "expendA" will lead to an $0.01\beta_1$ unit increase for "voteA".

(b) $H_0 : \ \beta_1 + \beta_2 = 0$.

(c) Let $\mathbf{R} = (0, 1, 1, 0)$, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$. Then under the null hypothesis, our test statistic t and its distribution is then given by:

$$t = \frac{\mathbf{R}\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\mathbf{R}'}} = \frac{\mathbf{R}\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}} \sim t(n-4)$$

(d) $H_0 : \ \beta_1 = \beta_2 = 0$.

(e) Let $\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$. Then under the null hypothesis, our test statistic F and its distribution is then given by:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}})}{2\hat{\sigma}^2} \sim F(2, n-4).$$

Problem Set 4

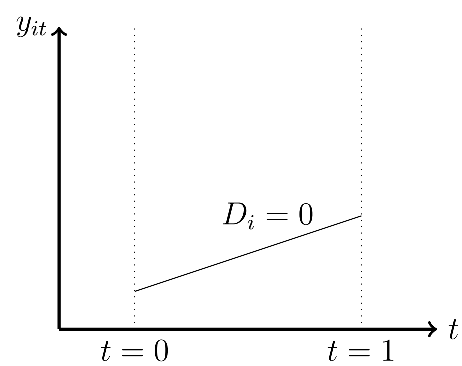
Due: 03/27

Part One: Hand-Written Exercise

1. The government is going to bulid a new railway. Tom wants to understand if a city's GDP will be influenced by the passing of the new railway. He uses a DID model below:

$$y_{it} = \alpha + \beta t + \gamma D_i + \varphi_i + \delta(D_i t) + u_{it},$$

where y_{it} denotes the GDP of city i at time t, and all the other notations are defined the same as in our lecture slides. Suppose both γ and δ are greater than 0, and the regression line for cities which are not passed by the railway ($D_i = 0$) is:



Draw the line for the cities passed by the railway ($D_i = 1$) on the plot above, and indicate β, γ and δ.

2. Consider the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

that satisfies the Modern Assumptions. Moreover, let $\mathrm{Var}(x_1) = \sigma_{x_1}^2$ and $\mathrm{Cov}(x_1, x_2) = \sigma_{x_1 x_2}$. Suppose we exclude an important variable x_2 and obtain the corresponding OLS estimator $\tilde{\beta}_1$. That is, we obtain $\tilde{\beta}_1$ from the model $y_i = \beta_0 + \beta_1 x_{1i} + u_i$.

(a) Is $\tilde{\beta}_1$ consistent for β_1?

(b) As the sample size $n \to \infty$ and $\sigma_{x_1 x_2} > 0$, does $\tilde{\beta}_1$ over- or under-estimate β_1? By how much?

(c) As the sample size $n \to \infty$ and $\sigma_{x_1 x_2} > 0$, does $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ converge to a normal distribution, or any other distributions?

3. Answer the following questions with "True" or "False" and briefly explain them. All notations are defined as in our lecture slides.

   (a) A biased estimator must be inconsistent.

   (b) An unbiased estimator must be consistent.

Part Two: Monte Carlo Simulation

- Simulation design:
  - Sample sizes N:
    (i) 10 (ii) 500
  - Number of replications: 1000
  - Data generating process (DGP):
    (i) $y_i \sim N(0,1)$ (ii) $y_i \sim t(4)$ (iii) $y_i \sim t(1)$
  - The statistics:

$$M_N = \frac{1}{\hat{\sigma}_N \sqrt{N}} \sum_{i=1}^{N} \phi(y_i), \quad \text{where} \quad \hat{\sigma}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\phi(y_i) - \frac{1}{N} \sum_{i=1}^{N} \phi(y_i) \right)^2,$$

    with the moment functions:
    (i) $\phi(y_i) = y_i$ (ii) $\phi(y_i) = y_i^3$ (iii) $\phi(y_i) = sin(y_i)$ (iv) $\phi(y_i) = cos(y_i)$

1. For the total of 24 different ways to construct M_N, please plot their corresponding histograms using 1000 replications. Open a new window and combine the 24 graphs on a single plot and place them as 6×4.

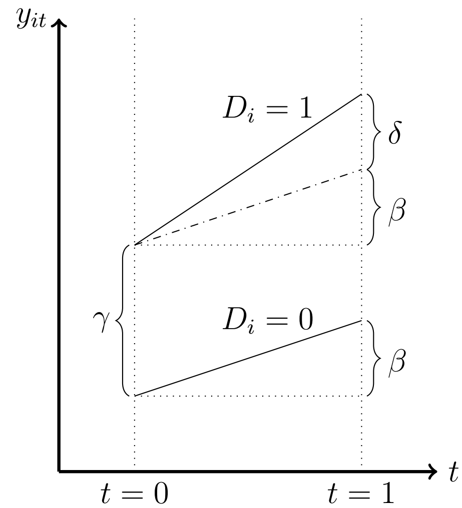
2. Please compute the empirical frequencies of the events: $M_N^2 > 3.8414588$ and $M_N^2 > 6.6348966$ for each simulations. Record them under their corresponding graphs. Check if the frequencies are, respectively, sufficiently close to the 5% and 1% nominal levels.

3. Please add the Gaussian kernel density estimate (KDE) of M_N as well as the probability density function (PDF) of $N(0,1)$ for each simulation graph.

Problem Set 4: Solution

Part One: Hand-Written Exercise

1. .



2. (a)

$$\tilde{\beta}_1 = \frac{\sum(x_{i1} - \bar{x}_1)y_i}{\sum(x_{i1} - \bar{x}_1)^2} = \frac{\sum(x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum(x_{i1} - \bar{x}_1)^2}$$
$$= \beta_1 + \beta_2 \frac{\sum(x_{i1} - \bar{x}_1)x_{2i}}{\sum(x_{i1} - \bar{x}_1)^2} + \frac{\sum(x_{i1} - \bar{x}_1)u_i}{\sum(x_{i1} - \bar{x}_1)^2}.$$

By the given assumptions, we have:

$$\frac{1}{n}\sum(x_{i1} - \bar{x}_1)u_i \xrightarrow{p} \mathbb{E}(x_1 u) - \mu_{x1}\mathbb{E}(u) = 0$$
$$\frac{1}{n}\sum(x_{i1} - \bar{x}_1)^2 \xrightarrow{p} \sigma^2_{x_1}$$
$$\frac{1}{n}\sum(x_{i1} - \bar{x}_1)x_{i2} = \frac{1}{n}\sum(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \xrightarrow{p} \sigma_{x_1 x_2}.$$

Therefore, we have:

$$\tilde{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}}.$$

$\tilde{\beta}_1$ is consistent for β_1 only when $\sigma_{x_1 x_2} = 0$, otherwise $\tilde{\beta}_1$ is not consistent.

(b) Since $\sigma_{x_1 x_2} > 0$, then if $\beta_2 > 0$, $\tilde{\beta}_1$ overestimates β_1 by $\beta_2 \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}}$ as $n \to \infty$. On the other hand, if $\beta_2 < 0$, then $\tilde{\beta}_1$ underestimates β_1 by $-\beta_2 \frac{\sigma_{x_1 x_2}}{\sigma^2_{x_1}}$ as $n \to \infty$.

(c) $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ does not follow any distributions. Since $\sigma_{x_1 x_2} > 0$, thus

$$\tilde{\beta} - \beta_1 \xrightarrow{p} \beta_2 \frac{\sigma_{x_1 x_2}}{\sigma_{x_1}^2} \neq 0.$$

So $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ clearly diverges as $n \to \infty$.

3. (a) False. Recall that $\hat{\sigma}^2_{OLS} = \frac{1}{n-k-1} \sum \hat{e}_i^2$ is an unbiased estimator of σ^2. Hence, for $s \in \mathbb{R}$, $\hat{\sigma}^2(s) = \frac{1}{n-s} \sum \hat{e}_i^2$ is a biased estimator as long as $s \neq k+1$. However, $\forall s \in \mathbb{R}$, $\hat{\sigma}^2(s)$ is consistent for σ^2.

(b) False. Consider a case where we have the data x_i, $i = 1, ..., n$, and the true population mean $\mu_x = 0$. Our estimator is designed as:

$$\hat{\mu}_n = \begin{cases} -1 & \text{with } p = 1/2 \\ 1 & \text{with } p = 1/2 \end{cases}.$$

This estimator, although completely disregards the data x_i, is still unbiased. It is, however, not consistent ($\lim\limits_{n \to \infty} \hat{\mu}_n \neq 0$).

Problem Set 5

Due: 04/03

Reminder: Please upload hand-written part on NTU COOL

Part One: Hand-Written Exercise

1. Verify the statement on slide 40, Lecture 4. That is, write down the 3×3 matrix $\mathbf{R}\tilde{\mathbf{D}}\mathbf{R}'$ using the notation d_{ij}, where $\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$ and $\tilde{\mathbf{D}} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1(k+1)} \\ d_{21} & d_{22} & \cdots & d_{2(k+1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(k+1)1} & d_{(k+1)2} & \cdots & d_{(k+1)(k+1)} \end{bmatrix}$.

2. Verify the statement on slide 10, Lecture 5. That is, $\hat{\beta}_{1,IV} = \dfrac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$.

3. Verify the statement on slide 28, Lecture 5. That is, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{b_o}\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{D_o})$.

Part Two: Part Two: Computer Exercise

1. Please load the dataset `SchoolingReturns` in `R`, which is a cross-section data from the U.S. National Longitudinal Survey of Young Men (NLSYM) in 1976, containing 3,010 observations on 22 variables. The variable we are interested in modelling is "`wage`". However, using the variable "`education`", the years of education, to explain "`wage`" is problematic because it can be argued that schooling is endogenous (and thus "`experience`" is also endogenous since it equals to `age` - `education` - 6). Thus, we conduct 2SLS estimations with the outcome log(`wage`), endogenous regressors "`education`", "`experience`" and the square of "`experience`" with their IV "`nearcollege`", "`age`" and the square of "`age`". Other exogenous regressors are "`ethnicity`", "`smsa`" and "`south`".

   (a) Perform the first stage of 2SLS.

   (b) Perform the second stage of 2SLS. Show the estimated coefficient for "`education`".

   (c) Perform 2SLS with the function "`ivreg`" and show the estimated coefficient for "`education`". Verify that it's identical to (b).

Problem Set 5: Solution

Part One: Hand-Written Exercise

1.

$$\mathbf{R}\tilde{\mathbf{D}}\mathbf{R}' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1(k+1)} \\ d_{21} & d_{22} & \cdots & d_{2(k+1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(k+1)1} & d_{(k+1)2} & \cdots & d_{(k+1)(k+1)} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} d_{21} & d_{22} & \cdots & d_{2(k+1)} \\ d_{31} & d_{32} & \cdots & d_{3(k+1)} \\ d_{41} & d_{42} & \cdots & d_{4(k+1)} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} d_{22} & d_{23} & d_{24} \\ d_{32} & d_{33} & d_{34} \\ d_{42} & d_{43} & d_{44} \end{bmatrix}. \quad \blacksquare$$

2. From slide 6, Lecture 5, we have known that

$$\hat{\beta}_{\mathbf{IV}} = (\mathbf{Z}'\mathbf{X})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y} = \left(\sum_{i=1}^{n} \mathbf{z_i}\mathbf{x_i}'\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{z_i} y_i\right).$$

Letting

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1} & \mathbf{z} \end{bmatrix} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix},$$

we have

$$
\begin{aligned}
\hat{\beta}_{\mathbf{IV}} &= \left(\sum_{i=1}^{n} \mathbf{z_i x_i'}\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{z_i} y_i\right) \\
&= \left(\sum_{i=1}^{n} \begin{bmatrix} 1 \\ z_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix}\right)^{-1} \sum_{i=1}^{n} \begin{bmatrix} 1 \\ z_i \end{bmatrix} y_i \\
&= \begin{bmatrix} n & n\bar{x} \\ n\bar{z} & \sum_{i=1}^{n} z_i x_i \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} z_i y_i \end{bmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n} z_i x_i - n^2 \bar{x}\bar{z}} \begin{bmatrix} \sum_{i=1}^{n} z_i x_i & -n\bar{x} \\ -n\bar{z} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} z_i y_i \end{bmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n} z_i x_i - n^2 \bar{x}\bar{z}} \begin{bmatrix} n\bar{y}\sum_{i=1}^{n} z_i x_i - n\bar{x}\sum_{i=1}^{n} z_i y_i \\ -n^2\bar{z}\bar{y} + n\sum_{i=1}^{n} z_i y_i \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{\bar{y}\sum_{i=1}^{n} z_i x_i - \bar{x}\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i x_i - n\bar{x}\bar{z}} \\ \\ \dfrac{\sum_{i=1}^{n} z_i y_i - n\bar{z}\bar{y}}{\sum_{i=1}^{n} z_i x_i - n\bar{x}\bar{z}} \end{bmatrix}
\end{aligned}
$$

Note that

$$
\begin{aligned}
\hat{\beta}_{1,IV} &= \frac{\sum_{i=1}^{n} z_i y_i - n\bar{z}\bar{y}}{\sum_{i=1}^{n} z_i x_i - n\bar{x}\bar{z}} \\
&= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}. \quad \blacksquare
\end{aligned}
$$

3. Recall that from slide 27, Lecture 5, we have

$$
\hat{\boldsymbol{\beta}}_{\text{GMM}} = \left(\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{W}}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{W}}\boldsymbol{Z}'\boldsymbol{y}\right).
$$

and we can rewrite it in a similar way from slide 24, Lecture 5,

$$
\hat{\boldsymbol{\beta}}_{\text{GMM}} = \boldsymbol{b}_o + \left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)(\boldsymbol{Z}'\boldsymbol{X}/n)\right]^{-1}\left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)(\boldsymbol{Z}'\boldsymbol{\epsilon}/n)\right].
$$

Thus,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{b}_o &= \left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)(\boldsymbol{Z}'\boldsymbol{X}/n)\right]^{-1}\left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)(\boldsymbol{Z}'\boldsymbol{\epsilon}/n)\right] \\
\Rightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{b}_o) &= \left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)(\boldsymbol{Z}'\boldsymbol{X}/n)\right]^{-1}\left[(\boldsymbol{X}'\boldsymbol{Z}/n)(\hat{\boldsymbol{W}}/n)\boldsymbol{V}^{\frac{1}{2}}\underbrace{\boldsymbol{V}^{-\frac{1}{2}}(\boldsymbol{Z}'\boldsymbol{\epsilon}/\sqrt{n})}_{\xrightarrow{D}\mathcal{N}(\mathbf{0},\mathbf{I})}\right] \\
\Rightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GMM}} - \boldsymbol{b}_o) &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{D_o}). \blacksquare
\end{aligned}
$$