ECON220B Discussion Section 4

Midterm Review

Lapo Bini

Roadmap

1. Question 1: Algebraic Properties of OLS

2. Question 2: Properties Empirical CDF Estimator

Exercise 1

Consider the following linear model:

$$y_i = \alpha + \mathbf{x}_i^T \beta + u_i$$

where $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^d$ and $u_i \in \mathbb{R}$. Note that the intercept is captured by α and it is not included in $\mathbf{x}_i$. Suppose that we have an iid sample $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$.

We will assume $E[u_i] = 0$ and $E[\mathbf{x}_i u_i] = 0$.

Exercise 1 - Question 1

Write down the sample moment conditions for $\hat{\alpha}$ and $\hat{\beta}$

$$E[u_i] = 0 \qquad \Longrightarrow \qquad \frac{1}{n}\sum \hat{u}_i = 0$$

$$E[u_i x_i] = 0 \qquad \qquad \frac{1}{n}\sum \hat{u}_i x_i = 0 \quad ///$$

EXTRA - USEFUL FOR LATER

(1) DERIVE $\hat{\alpha}$

$$\frac{1}{n}\sum \hat{u}_i = 0 \implies \frac{1}{n}\sum (y_i - \hat{\alpha} - x_i^T \hat{\beta}) = 0$$

$$\frac{1}{n}\sum y_i - \frac{1}{n}\sum \hat{\alpha} - \frac{1}{n}\sum x_i^T \hat{\beta} = 0$$

$$\bar{y} - \frac{n}{n}\hat{\alpha} - \left(\frac{1}{n}\sum x_i^T\right)\hat{\beta} = 0 \quad \therefore \quad \hat{\alpha} = \bar{y} - \bar{x}^T \hat{\beta}$$

Exercise 1 - Question 1

Write down the sample moment conditions for $\hat{\alpha}$ and $\hat{\beta}$

(2) DERIVE $\hat{\beta}$

$$\frac{1}{n}\sum x_i \hat{u}_i = 0 \quad \Rightarrow \quad \frac{1}{n}\sum x_i (y_i - \hat{\alpha} - x_i^T \hat{\beta}) = 0$$

$$\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i \hat{\alpha} - \frac{1}{n}\sum x_i x_i^T \hat{\beta} = 0$$

$$\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i (\bar{y} - \bar{x}^T \hat{\beta}) - \frac{1}{n}\sum x_i x_i^T \hat{\beta} = 0$$

$$\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i \bar{y} + \frac{1}{n}\sum x_i \bar{x}^T \hat{\beta} - \frac{1}{n}\sum x_i x_i^T \hat{\beta} = 0$$

$$\frac{1}{n}\sum x_i y_i - \left(\frac{1}{n}\sum x_i\right)\bar{y} + \left(\frac{1}{n}\sum x_i\right)\bar{x}^T \hat{\beta} - \frac{1}{n}\sum x_i x_i^T \hat{\beta} = 0$$

$$\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y} + \bar{x}\bar{x}^T \hat{\beta} - \frac{1}{n}\sum x_i x_i^T \hat{\beta} = 0$$

$$\hat{\beta} = \left(\frac{1}{n}\sum x_i x_i^T - \bar{x}\bar{x}^T\right)^{-1}\left(\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}\right) \quad ///$$

Exercise 1 - Question 2

Let $\hat{u}_i$ be the regression residual, write down its expression in terms of y_i, $\mathbf{x}_i$ and the estimated coefficient

$$\hat{u}_i = y_i - \alpha - x_i^T\beta \quad ///$$

Exercise 1 - Question 3

Now regress $\hat{u}_i$ on an intercept and $\mathbf{x}_i$. Find the estimated coefficients.

WHAT WE KNOW ABOUT $\hat{u}_i$: $\sum \hat{u}_i = 0$

$\sum x_i \hat{u}_i = 0$

- PROPOSED SOLUTION 1 :

$\hat{u}_i = \gamma_0 + x_i^T \gamma_1 + \varepsilon_i$ THIS IS THE REGRESSION WE WANT TO ESTIMATE. NOW DEFINE :

$\mathbb{X}_i = \underset{(d+1)\times 1}{\begin{vmatrix} 1 \\ x_i \end{vmatrix}}$ $\quad \Gamma = \underset{(d+1)\times 1}{\begin{vmatrix} \gamma_0 \\ \gamma_1 \end{vmatrix}}$ Let's estimate $\hat{u}_i = \mathbb{X}_i^T \Gamma + \varepsilon_i$

$$\therefore \quad \hat{\Gamma} = \left(\sum \mathbb{X}_i \mathbb{X}_i^T\right)^{-1} \left(\underline{\sum \mathbb{X}_i \hat{u}_i}\right) = \begin{vmatrix} \sum 1 \cdot u_i \\ \sum x_i \hat{u}_i \end{vmatrix} = 0$$

THEN $\hat{\gamma}_0 = 0 \quad \hat{\gamma}_1 = 0$ ///

Exercise 1 - Question 3

Now regress $\hat{u}_i$ on an intercept and $\mathbf{x}_i$. Find the estimated coefficients.

- PROPOSED SOLUTION 2: we want to estimate

$\hat{u}_i = \gamma_0 + x_i^T \gamma_1 + \varepsilon_i$. FROM OUR DERIVATION BEFORE

BY PROPERTY OLS.

$$\hat{\gamma}_1 = \left(\frac{1}{m}\sum x_i x_i^T - \bar{x}\bar{x}^T\right)^{-1}\left(\underbrace{\frac{1}{m}\sum x_i \hat{u}_i}_{=0} - \bar{x}\underbrace{\bar{\hat{u}}}_{=\frac{1}{m}\sum \hat{u}_i = 0}\right)$$

$$\hat{\gamma}_1 = \left(\frac{1}{m}\sum x_i x_i^T - \bar{x}\bar{x}^T\right)^{-1}(0 - 0) = 0$$

$$\hat{\gamma}_0 = \bar{\hat{u}} - \bar{x}^T\hat{\gamma}_1 = 0 - \bar{x}^T 0 = 0$$

Then $\hat{\gamma}_1 = \hat{\gamma}_0 = 0$ ///

Exercise 1 - Question 4

Let $\bar{x}$ be the sample mean of the regressors, and define $\check{\mathbf{x}}_i \equiv \mathbf{x}_i - \bar{x}$. Find the estimated coefficient. Now regress y_i on an intercept and $\check{\mathbf{x}}_i$. Find the estimated coefficients $\check{\alpha}$ and $\check{\beta}$. How are they related to the estimates $\hat{\alpha}$ and $\hat{\beta}$?

$$y_i = \alpha + x_i^T\beta + u_i$$

$$y_i = \alpha + x_i^T\beta + \bar{x}^T\beta - \bar{x}^T\beta + u_i$$

$$y_i = \alpha - \bar{x}^T\beta + (x_i^T - \bar{x}^T)\beta + u_i$$

$$y_i = (\alpha - \bar{x}^T\beta) + \tilde{x}_i^T\beta + u_i$$

$$y_i = \check{\alpha} + \tilde{x}_i^T\beta + u_i \qquad \therefore \check{\beta} = \hat{\beta}$$

BUT INTERCEPT IS CHANGING.

TO BETTER SEE THIS, WE WILL DERIVE BOTH $\check{\alpha}$ AND $\check{\beta}$ IN THE NEXT SLIDE.

Exercise 1 - Question 4

Let $\bar{x}$ be the sample mean of the regressors, and define $\check{\mathbf{x}}_i \equiv \mathbf{x}_i - \bar{x}$. Find the estimated coefficient. Now regress y_i on an intercept and $\check{\mathbf{x}}_i$. Find the estimated coefficients $\check{\alpha}$ and $\check{\beta}$. How are they related to the estimates $\hat{\alpha}$ and $\hat{\beta}$?

$$= \left[\frac{1}{m} \sum (x_i^T - \bar{x}^T) \right] \beta = 0$$

(1) $E[u_i] = 0 \Rightarrow \quad \frac{1}{m} \sum y_i = \hat{\gamma}_0 + \left(\frac{1}{m} \sum \check{x}_i^T \right) \beta$

$$\frac{1}{m} \sum y_i = \hat{\gamma}_0 \qquad \hat{\gamma}_0 = \check{\alpha} = \bar{y}$$

(2) $E[x_i u_i] = 0 \qquad \frac{1}{m} \sum x_i y_i = \frac{1}{m} \sum x_i \hat{\gamma}_0 + \frac{1}{m} \sum x_i \check{x}_i^T \check{\beta} = 0$

NOW NOTICE THAT $\sum x_i \check{x}_i^T = \sum x_i (x_i - \bar{x})^T = \sum x_i x_i^T - \sum x_i x^T$

$$= \sum x_i x_i^T - m \bar{x} \bar{x}^T$$

$$\therefore \quad \frac{1}{m} \sum x_i y_i = \frac{1}{m} \sum x_i \hat{\gamma}_0 + \left(\frac{1}{m} \sum x_i x_i^T - \bar{x} \bar{x}^T \right) \check{\beta} = 0$$

Exercise 1 - Question 4

Let $\bar{x}$ be the sample mean of the regressors, and define $\check{\mathbf{x}}_i \equiv \mathbf{x}_i - \bar{x}$. Find the estimated coefficient. Now regress y_i on an intercept and $\check{\mathbf{x}}_i$. Find the estimated coefficients $\check{\alpha}$ and $\check{\beta}$. How are they related to the estimates $\hat{\alpha}$ and $\hat{\beta}$?

LASTLY SUBSTITUTE $\hat{\gamma}_0$ AND $\check{\beta} = \left(\frac{1}{n} \sum x_i x_i^T - \bar{x}\bar{x}^T \right)^{-1} \left(\frac{1}{n} \sum x_i y_i - \bar{x}\bar{y} \right) = \hat{\beta}$ ///

Exercise 1 - Question 5

Claim: $\breve{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$. True or false? TRUE

WE HAVE $y_i = (\alpha + \bar{x}^T\beta) + \tilde{x}_i^T\beta + \hat{u}_i$

FROM $\frac{1}{n}\sum \hat{u}_i = 0 \implies \frac{1}{n}\sum y_i - \breve{\alpha} - \tilde{x}_i^T\hat{\beta} = 0$

$$\frac{1}{n}\sum y_i - \frac{1}{n}\sum \breve{\alpha} + \frac{1}{n}\sum \tilde{x}_i^T\hat{\beta} = 0$$

$$\bar{y} - \breve{\alpha} + \Big(\underbrace{\frac{1}{n}\sum(x_i^T - \bar{x}^T)}_{=0}\Big)\hat{\beta} = 0$$

$$\therefore\ \breve{\alpha} = \bar{y} = \frac{1}{n}\sum y_i$$

Exercise 2

Assume $\{x_i\}_{i=1}^n$ iid sample from a univariate distribution, $x_i \in \mathbb{R}$. Denote by $F(\cdot)$ the cumulative distribution which is defined as:

$$F(x) = \mathbb{P}(x_i \leq x)$$

For simplicity we will assume that x_i is continuously distributed on the unit interval such that:

- $F(x) = 0$ for all $x \leq 0$.
- $F(x) = 1$ for all $x \geq 1$.
- $F(x)$ continuous and strictly increasing for all $x \in (0, 1)$.

Define the empirical CDF as:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}$$

- THIS IS A POINT ESTIMATOR WHEN YOU PLUG x
- YOU GET ESTIMATE CDF ITERATING OVER SUPPORT OF X

Exercise 2 - Question 1

Assume $0 < x < 1$, find the asymptotic distribution of $\hat{F}(x)$

STEP 1 : WHERE WILL THE ASYMPTOTIC DISTRIBUTION BE CENTERED?

$\mathbb{1}\{\cdot\} \sim \text{BERNOULLI}(\mathbb{P}(X_i \leq x))$

$$\frac{1}{n}\sum \mathbb{1}\{X_i \leq x\} \underbrace{\xrightarrow{P}}_{\text{by WLLN}} E[\mathbb{1}\{X_i \leq x\}] = 1 \cdot \mathbb{P}(X_i \leq x) + 0\,(1 - \mathbb{P}(X_i \leq x))$$

SINCE $\{X_i\}_{i=1}^{n}$ IID FINITE MEAN AND VARIANCE

$$\therefore\ \hat{F}(x) = \frac{1}{n}\sum \mathbb{1}\{X_i \leq x\} \xrightarrow{P} \mathbb{P}(X_i \leq x) = F(x)$$

Exercise 2 - Question 1

Assume $0 < x < 1$, find the asymptotic distribution of $\hat{F}(x)$

STEP 2 : APPLY CLT

$$\sqrt{n}\left(\hat{F}(x) - F(x)\right) = \sqrt{n}\left(\frac{1}{n}\sum \mathbb{1}\{X_i \le x\} - \frac{n}{n}F(x)\right)$$

$$= \sqrt{n}\left\{\frac{1}{n}\sum\left(\mathbb{1}\{X_i \le x\} - F(x)\right)\right\}$$

$$= \frac{1}{\sqrt{n}}\sum\left(\mathbb{1}\{X_i \le x\} - F(x)\right) \xrightarrow{d} N(0, V)$$

Exercise 2 - Question 1

Assume $0 < x < 1$, find the asymptotic distribution of $\hat{F}(x)$

STEP 3: FIND ASYMPTOTIC VARIANCE

$$V = \mathrm{Var}\left(\sqrt{m}\left(\hat{F}(x) - F(x)\right)\right) = \mathrm{Var}\left(\frac{1}{\sqrt{m}}\sum\left(\mathbb{1}\{X_i \le x\} - F(x)\right)\right)$$

$$= \frac{1}{m}\mathrm{Var}\left(\sum\left(\mathbb{1}\{X_i \le x\} - F(x)\right)\right)$$

$\{x_i\}_{i=1}^{m}$ IID $\Rightarrow$ $\{\mathbb{1}\{X_i \le x\}\}_{i=1}^{m}$ IID

$\therefore$ ALL CROSS COVARIANCES $= 0$, WE CAN MOVE SUMMATION OUTSIDE

$$= \frac{1}{m}\sum \mathrm{Var}\left(\mathbb{1}\{X_i \le x\} - F(x)\right)$$

IDENTICALLY DISTRIBUTED

$$= \frac{m}{m}\mathrm{Var}\left(\mathbb{1}\{X_i \le x\} - F(x)\right) = \mathrm{Var}\left(\mathbb{1}\{X_i \le x\} - F(x)\right) =$$

JUST A CONSTANT

$a \in \mathbb{R}$, X RANDOM VAR.

$\mathrm{Var}(X + a) = \mathrm{Var}(X)$

$$= \mathrm{Var}\left(\mathbb{1}\{X_i \le x\}\right)$$

$\sim$ BERNOULLI

$$= \mathbb{P}(X_i \le x)\left(1 - \mathbb{P}(X_i \le x)\right) = F(x)\left(1 - F(x)\right)$$

Exercise 2 - Question 2

Assume $0 < x \neq x' < 1$, find the asymptotic distribution of $\hat{F}(x)$ and $\hat{F}(x')$

STEP 1 : FIND PLIM. WE KNOW FROM BEFORE THAT

$\hat{F}(x) \xrightarrow{p} F(x)$ Then we get $\begin{vmatrix} \hat{F}(x) \\ \hat{F}(x') \end{vmatrix} \xrightarrow{p} \begin{vmatrix} F(x) \\ F(x') \end{vmatrix}$

STEP 2 : APPLY CLT (MULTIVARIATE) : $\{X_i\}_{i=1}^{\infty}$ IID, and ASYMPTOTIC VARIANCE $\hat{F}(x)$ FINITE SINCE $F(x) \in [0,1]$ THEN COVARIANCE $\hat{F}(x)$ AND $\hat{F}(x')$ FINITE AND

$$\sqrt{n}\left(\begin{vmatrix} \hat{F}(x) \\ \hat{F}(x') \end{vmatrix} - \begin{vmatrix} F(x) \\ F(x') \end{vmatrix} \right) \xrightarrow{d} \mathcal{N}\left(\begin{vmatrix} 0 \\ 0 \end{vmatrix} ; \begin{vmatrix} F(x)(1-F(x)) & Cov \\ Cov & F(x')(1-F(x')) \end{vmatrix} \right)$$

LET'S DERIVE THIS

Exercise 2 - Question 2

Assume $0 < x \neq x' < 1$, find the asymptotic distribution of $\hat{F}(x)$ and $\hat{F}(x')$

$$\mathrm{Cov} = \mathrm{Cov}\left(\sqrt{m}\left(\hat{F}(x) - F(x)\right); \sqrt{m}\left(\hat{F}(x') - F(x')\right)\right)$$

$$= \mathrm{Cov}\left(\frac{1}{\sqrt{m}}\sum\left(\mathbb{1}\{X_i \leq x\} - F(x)\right); \frac{1}{\sqrt{m}}\sum\left(\mathbb{1}\{X_i \leq x'\} - F(x')\right)\right)$$

$a, b \in \mathbb{K}$

X, Y RVs

$\mathrm{Cov}(aX + bY) = ab\,\mathrm{Cov}(X, Y)$

$$= \frac{1}{m}\mathrm{Cov}\left(\sum\left(\mathbb{1}\{X_i \leq x\} - F(x)\right); \sum\left(\mathbb{1}\{X_i \leq x'\} - F(x')\right)\right)$$

$\mathrm{Cov}(a + Y, b + Y) = \mathrm{Cov}(X, Y)$

$$= \frac{1}{m}\mathrm{Cov}\left(\sum \mathbb{1}\{X_i \leq x\}; \sum \mathbb{1}\{X_i \leq x'\}\right)$$

- m TIMES $\mathrm{Cov}(\mathbb{1}\{X_i \leq x\}, \mathbb{1}\{X_i \leq x'\})$ SINCE IDENTICALLY DISTRIBUTED
- $m(m-1)$ TIMES $\mathrm{Cov}(\mathbb{1}\{X_i \leq x\}, \mathbb{1}\{X_j \leq x'\})$ WITH $i \neq j$ BUT ALL ZEROS SINCE $X_i \perp\!\!\!\perp X_j \ \forall i \neq j$

$$= \frac{m}{m}\mathrm{Cov}\left(\mathbb{1}\{X_i \leq x\}, \mathbb{1}\{X_i \leq x'\}\right)$$

Exercise 2 - Question 2

Assume $0 < x \neq x' < 1$, find the asymptotic distribution of $\hat{F}(x)$ and $\hat{F}(x')$

$$\mathrm{Cov} = \mathbb{E}[\underbrace{\mathbb{1}\{X_i \le x\}\mathbb{1}\{X_i \le x'\}}_{:= K \;=\; \begin{cases} 1 & \text{IF } \{X_i \le x\} \cup \{X_i \le x'\} \\ 0 & \text{OTHERWISE.} \end{cases} \;=\; \begin{cases} 1 & X_i \le \min\{x, x'\} \\ 0 & \text{OTHERWISE} \end{cases}}] - \mathbb{E}[\mathbb{1}\{X_i \le x\}]\mathbb{E}[\mathbb{1}\{X_i \le x'\}]$$

$$\mathrm{Cov} = \mathbb{E}[\underbrace{\mathbb{1}\{X_i \le \min\{x, x'\}}_{\sim \text{BERNOULLI}}] - F(x)F(x')$$

$$\mathrm{Cov} = \mathbb{P}(X_i \le \min\{x, x'\}) - F(x)F(x')$$

$$= F(\min\{x, x'\}) - F(x)F(x')$$

Exercise 2 - Question 3

(*) CALL THIS x AND PROOF IS COMPLETE.

Consider the hypothesis $\mathsf{H}_0 : F(x) = G(x)$ and define the following statistic:

$$KS = \sup_{x \in [0,1]} \sqrt{n} |\hat{F}(x) - G(x)|$$

Show that under the null, the KS statistic can be rewritten as:

$$KS = \sup_{x \in [0,1]} \sqrt{n} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}\{F(x_i) \leq x\} - x \right) \right|$$

DEFINE $y := F(x)$

KOLMOGOROV'S AXIOM OF PROB

WE KNOW $y \in [0,1]$. SINCE $F(\cdot)$ STRICTLY INCREASING AND CONTINUOUS $\exists F^{-1}(\cdot)$, THEN

$$\sup_{x \in [0,1]} \sqrt{n} \left| \frac{1}{n} \sum \mathbb{1}\{X_i \leq x\} - F(x) \right| = \sup_{y \in [0,1]} \sqrt{n} \left| \frac{1}{n} \sum \mathbb{1}\{X_i \leq F^{-1}(x)\} - y \right| \quad (*)$$

$$= \sup_{x \in [0,1]} \sqrt{n} \left| \frac{1}{n} \sum \{ F(x_i) \leq F(F^{-1}(y)) \} - y \} = \sup_{y \in [0,1]} \sqrt{n} \left| \frac{1}{n} \sum \mathbb{1}\{F(x_i) \leq y\} - y \right| \quad ///$$

Exercise 2 - Question 4

Consider the hypothesis $\mathsf{H}_0 : F(x) = G(x)$ and show that the KS statistic does not depend on the underlying distribution $F(\cdot)$.

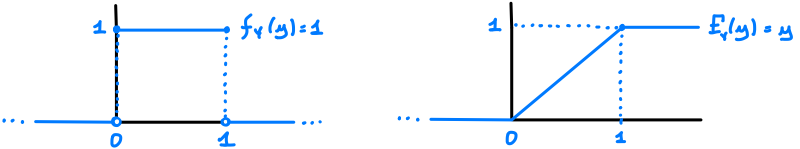
$$\sup_{x\in[0,1]} \sqrt{m} \left| \frac{1}{m} \sum \mathbb{1}\{X_i \le x\} - F(x) \right|$$

NOW CALL $Y := F(x)$

WHAT IS ITS CDF ?

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(F(x) \le y) = \mathbb{P}(X \le F^{-1}(y)) = F(F^{-1}(y)) = y$$

$F_Y(y) = y \Rightarrow$ CDF OF UNIFORM OVER UNIT INTERVAL $[0,1]$



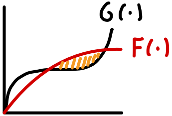
WE DON'T CARE ABOUT $F(\cdot)$, BY CHANGE OF VARIABLE ALWAYS UNIFORM

Exercise 2 - Question 5

Discuss how you can modify the KS statistic to test FOSD.

$H_0 : G(\cdot)$ FOSD $\hat{F}(\cdot)$, 2 OBSERVATIONS

INTERESTED IN THE AREA BETWEEN THE CURVES



IF H_0 $\hat{F}(\cdot) \leq G(\cdot)$ WE INTERESTED IN ORIENTED DEVIATIONS : WE NEED TO REMOVE ABSOLUTE VALUE

$$\therefore \; t = \int_0^1 \sqrt{m}\left(\hat{F}(x) - G(x)\right)^+ dx$$

$\Rightarrow$ IF $t \geq 0$ THEN EVIDENCE TO REJECT H_0

$$t = \int_0^1 \sqrt{m}\left(\hat{F}(x) - G(x)\right) \mathbb{1}\{\hat{F}(x) - G(x) \geq 0\} dx$$