Empirical Analysis I - Problem Set 3

Timothy Schwieg
Paulo Henrique Ramos
Samuel Barker
Rafeh Qureshi

Question 1

Notice that, because we are interested in the limit distribution of the variance estimator, we can assume that the observations are centered (that is, consider that the mean of the distribution was already substracted from the X_i; i.e. $X_i = Y_i - \mu$ for some Y_i with the distribution we are interested in). This gives us that $\mathbb{E}[X_i] = 0$ (which we can interpret as the first centered moment, and similarly with higher moments: all are interpreted as centered), and it does not change the variance or its estimator.

Also, notice that $S_n^2 = \frac{1}{n-1}\sum (X_i - \overline{X}_n)^2 = \frac{n}{n-1}\left[[\frac{1}{n}\sum X_i^2] - \overline{X}_n^2\right]$. Thus we can write $S_n^2 = g(\frac{n}{n-1}, \overline{X^2}_n, \overline{X}_n)$, where $g(a,b,c) = a[b-c^2]$ is a continuous and differentiable function on the relevant domain.

Because $\frac{n}{n-1} \xrightarrow{p} 1$, $\overline{X^2}_n \xrightarrow{p} \mathbb{E}[X_i^2]$ and $\overline{X}_n \xrightarrow{p} \mathbb{E}[X_i]$ due to the WLLN (which we can apply, because $\mathbb{E}[X_i^4] \leq \infty$), and marginal convergence in probability implies joint convergence in probability, we have $(\frac{n}{n-1}, \overline{X^2}_n, \overline{X}_n) \xrightarrow{p} (1, \mathbb{E}[X_i^2], \mathbb{E}[X_i])$.

Also, when applied to the limits, $g(1, \mathbb{E}[X_i^2], \mathbb{E}[X_i])) = [\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2] = \sigma^2$, the true variance.

Also, because $\mathbb{E}[X_i^4] \leq \infty$, we can apply the CLT to this vector:

$$\begin{aligned}\sqrt{n}\left(\begin{bmatrix}\frac{n}{n-1}\\ \overline{X^2}_n\\ \overline{X}_n\end{bmatrix} - \begin{bmatrix}1\\ \mathbb{E}[X_i^2]\\ \mathbb{E}[X_i]\end{bmatrix}\right) &\xrightarrow{d} N\left(0, \begin{bmatrix}0 & 0 & 0\\ 0 & \mathbb{E}[X_i^4] - \mathbb{E}[X_i^2]^2 & \mathbb{E}[X_i^3] - \mathbb{E}[X_i^2]\mathbb{E}[X_i]\\ 0 & \mathbb{E}[X_i^3] - \mathbb{E}[X_i^2]\mathbb{E}[X_i] & \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\end{bmatrix}\right)\\ \text{(because } \mathbb{E}[X_i] = 0) \ &= N\left(0, \begin{bmatrix}0 & 0 & 0\\ 0 & \mathbb{E}[X_i^4] - \mathbb{E}[X_i^2]^2 & \mathbb{E}[X_i^3]\\ 0 & \mathbb{E}[X_i^3] & \mathbb{E}[X_i^2]\end{bmatrix}\right)\end{aligned} \tag{1}$$

Now we can use the delta method with $g(.)$ (its derivative on (a,b,c) is $(1,1,-2c)$) and

conclude:

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \begin{bmatrix}1 & 1 & -2\mathbb{E}\left[X_i\right](=0)\end{bmatrix} \begin{bmatrix}0 & 0 & 0 \\ 0 & \mathbb{E}\left[X_i^4\right] - \mathbb{E}\left[X_i^2\right]^2 & \mathbb{E}\left[X_i^3\right] \\ 0 & \mathbb{E}\left[X_i^3\right] & \mathbb{E}\left[X_i^2\right]\end{bmatrix} \begin{bmatrix}1 \\ 1 \\ -2\mathbb{E}\left[X_i\right](=0)\end{bmatrix}$$
$$= N(0, \mathbb{E}\left[X_i^4\right] - \mathbb{E}\left[X_i^2\right]^2)) \tag{2}$$

Thus, recalling that $\mathbb{E}\left[X_i^4\right]$ and $\mathbb{E}\left[X_i^2\right]$ are centered moments, we have that $\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \mathbb{E}\left[(Y_i - \mu)^4\right] - \sigma^4)$, if we consider Y_i as the uncentered variable.

Question 2

Show that $o_P(1) + \mathcal{O}_P(1) = \mathcal{O}(1)$.

Note that $X_n = o_P(1)$ if $X_n \xrightarrow{p} 0$ and $X_n = \mathcal{O}_P(1)$ if X_n is tight. Note that $X_n \xrightarrow{p} 0$ implies that $X_n \xrightarrow{d} 0$ and therefore X_n is tight.

Let $X_n = o_P(1)$ and $Y_N = \mathcal{O}(1)$. By the above logic, X_n is tight. So $\forall \epsilon > 0, \exists B_x, B_y$ such that:

$$\inf_n \Pr(|X_n| \leq B_x) \geq 1 - \frac{\epsilon}{2}$$
$$\inf_n \Pr(|Y_n| \leq B_y) \geq 1 - \frac{\epsilon}{2}$$

For any such $\epsilon > 0$, choose M such that

$$\frac{M}{2} > B_x \qquad \frac{M}{2} > B_y$$

Define A and B such that:

$$A := \{|X_n| + |Y_n| > M\} \Rightarrow \left\{|X_n| > \frac{M}{2}\right\} \cup \left\{|Y_n| > \frac{M}{2}\right\} =: B$$

Note that:

$$\Pr(A) \leq \Pr(B) \leq \Pr\left(|X_n| > \frac{M}{2}\right) + \Pr\left(|Y_n| > \frac{M}{2}\right) < \epsilon$$

From the definition of tightness, and our choice of M.

Define $C := \{|X_n + Y_n| > M\}$. From the triangle inequality we know that $|X_n + Y_n| \leq |X_n| + |Y_n|$

Thus:

$$|X_n + Y_n| > M \Rightarrow |X_n| + |Y_n| > M$$
$$\Pr(|X_n + Y_n| > M) \leq \Pr(|X_n| + |Y_n| > M) < \epsilon$$
$$\Pr(|X_n + Y_n| \leq M) \geq 1 - \epsilon$$

This tells us that $X_n + Y_n$ is tight, and therefore $\mathcal{O}_P(1) + \mathcal{O}_P(1) = \mathcal{O}_P(1)$. So:

$$o_P(1) + \mathcal{O}_P(1) = \mathcal{O}_P(1)$$

Question 3

In what sense is $o_P(1) = \mathcal{O}_P(1)$? Is $\mathcal{O}_P(1) = o_p(1)$?

We say that a sequence of random variables, X_n, is $o_P(1)$ if $X_n \xrightarrow{p} 0$. We say that X_n is $\mathcal{O}_P(1)$ if X_n is tight. Since we have that

$$X_n = o_P(1) \implies X_n \xrightarrow{d} 0$$

and

$$X_n \xrightarrow{d} X \implies X_n = \mathcal{O}_P(1),$$

(where X is a random variable) we have,

$$X_n = o_P(1) \implies X_n = \mathcal{O}_P(1).$$

In this sense,

$$o_P(1) = \mathcal{O}_P(1).$$

However, the converse is not true in general. For instance, realize that $X_n \xrightarrow{d} X$ is a sufficient condition for tightness, but not for convergence in probability. Only when X is a constant does it imply convergence in probability, but even then, X must equal 0 for $X_n = o_P(1)$.

An even stronger statement can be said though: in general, tightness does not imply convergence in distribution, and therefore does not imply convergence in probability. Consider, a sequence of random variables, X_n, where $X_{2n} \sim U[0,1]$, and $X_{2n+1} \sim U[2,3]$. It is obvious that X_n does not converge in distribution. However, it is tight. To prove this, take $M_\epsilon = 3$. Then, we have

$$\sup \Pr(|X_n| > 3) < \epsilon, \forall \epsilon > 0.$$

Thus, $X_n = \mathcal{O}_P(1)$, but $X_n \neq o_P(1)$.

Question 4

Suppose $\tau_n \uparrow \infty$ and for all $\epsilon > 0$, there exists $B > 0$, such that

$$\inf_n Pr(|\tau_n(\hat{\theta} - \theta)| \leq B) \geq 1 - \epsilon$$

Equivalently, we have $\inf_n Pr(|\hat{\theta} - \theta| \leq \frac{B}{|\tau_n|}) \geq 1 - \epsilon$.
Now, we can choose some N $\in \mathbb{N}$ such that, for all $n > N$, $\frac{B}{\tau_n} < \delta$ as B is a constant and $\tau_n \uparrow \infty$. Then, we have that, for all $n > N$,

$$\begin{aligned} 1 - \epsilon &\leq \inf_n (Pr(|\hat{\theta} - \theta)| \leq \frac{B}{|\tau_n|}) \\ &\leq \inf_{n>N} (Pr(|\hat{\theta} - \theta)| \leq \frac{B}{|\tau_n|}) \\ &\leq \inf_{n>N} (Pr(|\hat{\theta} - \theta)| \leq \delta) \end{aligned}$$

This equivalently states that tightness of $\tau_n(\hat{\theta} - \theta)$ implies that
$Pr(|\hat{\theta} - \theta)| \leq \delta) \to 1$

Question 5

a

($\Longrightarrow$) If $p(1|1) = p(1|0)$ and $p(0|1) = p(0|0)$, then the ratio $\frac{p(1|1)}{p(1|0)} \Big/ \frac{p(0|1)}{p(0|0)} = 1/1 = 1 = \rho$.

($\Longleftarrow$) If $\rho = 1$, then we have:

$$\begin{gathered}
\frac{p(1|1)}{p(1|0)} = \frac{p(0|1)}{p(0|0)} \\
\Longleftrightarrow \\
\frac{p(0|0)}{p(1|0)} + \frac{p(1|0)}{p(1|0)} = \frac{p(0|1)}{p(1|1)} + \frac{p(1|1)}{p(1|1)} \\
\Longleftrightarrow \\
\frac{p(0|0) + p(1|0)}{p(1|0)} = \frac{p(0|1) + p(1|1)}{p(1|1)} \\
\Longleftrightarrow \\
\frac{1}{p(1|0)} = \frac{1}{p(1|1)} \\
\Longleftrightarrow \\
p(1|0) = p(1|1)
\end{gathered} \tag{3}$$

But, this implies that $1 = \frac{p(0|1)}{p(0|0)}$, which implies $p(0|1) = p(0|0)$.

b

Using bayes rules, we have:

$$\begin{aligned}
\rho &= \frac{\frac{p(1,1)}{p(1)}}{\frac{p(1,0)}{p(0)}} \Big/ \frac{\frac{p(0,1)}{p(1)}}{\frac{p(0,0)}{p(0)}} \\
\rho &= \frac{p(1,1)}{p(1,0)} \Big/ \frac{p(0,1)}{p(0,0)}
\end{aligned} \tag{4}$$

c

Let's first take $\hat{p}_n^{(1,1)} := \frac{1}{n} \sum \mathbb{1}_{\{Y_i=1, X_i=1\}}$. Because $\mathbb{1}_{\{Y_i=1, X_i=1\}}$ is a Bernoulli random variable, it has finite expectation $(p(1,1))$ and variance $((p(1,1)[1 - p(1,1)])$. Thus, since the sample of students is i.i.d., the random variable $\mathbb{1}_{\{Y_i=1, X_i=1\}}$ is also i.i.d., and the WLLN gives us that $\hat{p}_n^{(1,1)} \xrightarrow{p} p(1,1)$. The same rationale applies to $\hat{p}_n^{(1,0)}, \hat{p}_n^{(0,1)}$ and $\hat{p}_n^{(0,0)}$. And because marginal convergence in probability implies joint convergence in probability, we have: $(\hat{p}_n^{(1,1)}, \hat{p}_n^{(1,0)}, \hat{p}_n^{(0,1)}, \hat{p}_n^{(0,0)}) \xrightarrow{p} (p(1,1), p(1,0), p(0,1), p(0,0))$.

Now we know from item (b) that $\rho = \frac{p(1,1)}{p(1,0)} \Big/ \frac{p(0,1)}{p(0,0)}$, and no terms in this expression are zero. Thus, $f(a,b,c,d) = \frac{a}{b} \Big/ \frac{c}{d}$ is a continuous function on the relevant domain, and we can apply the continuous mapping theorem to conclude that:

$$\hat{\rho}_n = f(\hat{p}_n^{(1,1)}, \hat{p}_n^{(1,0)}, \hat{p}_n^{(0,1)}, \hat{p}_n^{(0,0)}) \xrightarrow{p} f(p(1,1), p(1,0), p(0,1), p(0,0)) = \rho \tag{5}$$

Thus our estimator is $\hat{\rho}_n := \frac{\hat{p}_n^{(1,1)}}{\hat{p}_n^{(1,0)}} \Big/ \frac{\hat{p}_n^{(0,1)}}{\hat{p}_n^{(0,0)}}$, and its is consistent.

d

As seen in item (c) above, $\mathbb{1}_{\{Y_i=1, X_i=1\}}$ is bernoulli, with finite expectation and variance (and similarly with the other indicator functions for different outcomes). Therefore, we can apply the multivariate CLT to conclude that:

$$\sqrt{n}\left(\begin{bmatrix} \hat{p}_n^{(1,1)} \\ \hat{p}_n^{(1,0)} \\ \hat{p}_n^{(0,1)} \\ \hat{p}_n^{(0,0)} \end{bmatrix} - \begin{bmatrix} p(1,1) \\ p(1,0) \\ p(0,1) \\ p(0,0) \end{bmatrix} \right) \xrightarrow{d} N(0, \Sigma), \tag{6}$$

where:

$$\Sigma = \begin{bmatrix} p_{(1,1)}[1-p_{(1,1)}] & -p_{(1,1)}p_{(1,0)} & -p_{(1,1)}p_{(0,1)} & -p_{(1,1)}p_{(0,0)} \\ -p_{(1,0)}p_{(1,1)} & p_{(1,0)}[1-p_{(1,0)}] & -p_{(1,0)}p_{(0,1)} & -p_{(1,0)}p_{(0,0)} \\ -p_{(0,1)}p_{(1,1)} & -p_{(0,1)}p_{(1,0)} & p_{(0,1)}[1-p_{(0,1)}] & -p_{(0,1)}p_{(0,0)} \\ -p_{(0,0)}p_{(1,1)} & -p_{(0,0)}p_{(1,0)} & -p_{(0,0)}p_{(0,1)} & p_{(0,0)}[1-p_{(0,0)}] \end{bmatrix} \tag{7}$$

We can then take the function $g(a,b,c,d) = \ln \frac{a}{b} \Big/ \frac{c}{d} = \ln a - \ln b - \ln c + \ln d$, which is continuous and differentiable on $(p(1,1), p(1,0), p(0,1), p(0,0))$, since they are greater than zero. Using the delta method, we obtain that:

$$\begin{aligned} \sqrt{n}(g(\hat{p}_n^{(1,1)}, \hat{p}_n^{(1,0)}, \hat{p}_n^{(0,1)}, \hat{p}_n^{(0,0)}) - g(p(1,1), p(1,0), p(0,1), p(0,0))) \\ \xrightarrow{d} Dg(p(1,1), p(1,0), p(0,1), p(0,0))N(0, \Sigma) \\ \sqrt{n}(\ln \hat{\rho}_n - \ln \rho) \xrightarrow{d} Dg(p(1,1), p(1,0), p(0,1), p(0,0))N(0, \Sigma) \end{aligned} \tag{8}$$

Now we can calculate the variance of $Dg(.)N(0,\Sigma)$ and call it τ^2 as we wanted:

$$\begin{bmatrix} \frac{1}{p_{(1,1)}} \\ -\frac{1}{p_{(1,0)}} \\ -\frac{1}{p_{(0,1)}} \\ \frac{1}{p_{(0,0)}} \end{bmatrix}^T \begin{bmatrix} p_{(1,1)}[1-p_{(1,1)}] & -p_{(1,1)}p_{(1,0)} & -p_{(1,1)}p_{(0,1)} & -p_{(1,1)}p_{(0,0)} \\ -p_{(1,0)}p_{(1,1)} & p_{(1,0)}[1-p_{(1,0)}] & -p_{(1,0)}p_{(0,1)} & -p_{(1,0)}p_{(0,0)} \\ -p_{(0,1)}p_{(1,1)} & -p_{(0,1)}p_{(1,0)} & p_{(0,1)}[1-p_{(0,1)}] & -p_{(0,1)}p_{(0,0)} \\ -p_{(0,0)}p_{(1,1)} & -p_{(0,0)}p_{(1,0)} & -p_{(0,0)}p_{(0,1)} & p_{(0,0)}[1-p_{(0,0)}] \end{bmatrix} \begin{bmatrix} \frac{1}{p_{(1,1)}} \\ -\frac{1}{p_{(1,0)}} \\ -\frac{1}{p_{(0,1)}} \\ \frac{1}{p_{(0,0)}} \end{bmatrix} =$$

$$\begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{p_{(1,1)}} \\ -\frac{1}{p_{(1,0)}} \\ -\frac{1}{p_{(0,1)}} \\ \frac{1}{p_{(0,0)}} \end{bmatrix} = \frac{1}{p_{(1,1)}} + \frac{1}{p_{(1,0)}} + -\frac{1}{p_{(0,1)}} + \frac{1}{p_{(0,0)}} =: \tau^2 \tag{9}$$

e

The estimator is: $\hat{\tau^2}_n := 1/\hat{p}_n^{(1,1)} + 1/\hat{p}_n^{(1,0)} + 1/\hat{p}_n^{(0,1)} + 1/\hat{p}_n^{(0,0)}$. As seen in item (c) above, we have $(\hat{p}_n^{(1,1)}, \hat{p}_n^{(1,0)}, \hat{p}_n^{(0,1)}, \hat{p}_n^{(0,0)}) \xrightarrow{p} (p_{(1,1)}, p_{(1,0)}, p_{(0,1)}, p_{(0,0)})$. Thus, because all terms are greater than zero, the function $f(a,b,c,d) = 1/a + 1/b + 1/c + 1/d$ is continuous on the relevant domain, and we can apply the CMT to conclude that $\hat{\tau^2}_n \xrightarrow{p} \tau^2$.

f

We know from previous items that $\sqrt{n}(\ln \hat{\rho}_n - \ln \rho) \xrightarrow{d} N(0, \tau^2)$. Because the function $g(x) = \exp x$ is continuous and differentiable at the true ρ, we can apply the delta method to obtain:

$$\sqrt{n}(\hat{\rho}_n - \rho) \xrightarrow{d} \rho N(0, \tau^2) = N(0, \tau^2 \rho^2) \tag{10}$$

g

If gender is independent of supporting Obama, the we have that the conditional probabilities equal the unconditional ones: $p(1|0) = p(1|1)$ and $p(0|0) = p(0|1)$. As seen in item (a), this is equivalent to $\rho = 1$, thus we can define $H_0 : \rho = 1$ and $H_1 : \rho \neq 1$.

Now define $T_n := \left| \frac{\sqrt{n}(\hat{\rho}_n - 1)}{\sqrt{\hat{\tau^2}_n}} \right| = \frac{\sqrt{n}(|\hat{\rho}_n - 1|)}{\sqrt{\hat{\tau^2}_n}}$. We will reject the null if $T_n > z_{1-\frac{\alpha}{2}}$.

Under the null hypothesis (i.e., taking $\rho = 1$), and using CMT and slutsky (because $\hat{\tau^2}_n \xrightarrow{p} \tau^2 > 0$, $g(x) = 1/\sqrt{x}$ is continuous on the relevant domain, and τ^2 is constant), we know that: $Tn \xrightarrow{d} |z| \sim N(0,1)$. Therefore, we have:

$$\limsup \Pr(Tn > z_{1-\frac{\alpha}{2}}) \leq \limsup \Pr(Tn \geq z_{1-\frac{\alpha}{2}}) \leq \Pr(|z| \geq z_{1-\frac{\alpha}{2}}) = \alpha. \tag{11}$$

Thus the test is consistent in level α. And we can also calculate the $p-$value using the following:

$$\begin{aligned} p - value &= \inf \{\alpha \in (0,1) | T_n \geq z_{1-\frac{\alpha}{2}}\} \\ &= \inf \{\alpha \in (0,1) | \phi(T_n) \geq \phi(z_{1-\frac{\alpha}{2}})\} \\ &= \inf \{\alpha \in (0,1) | \phi(T_n) \geq 1 - \frac{\alpha}{2}\} \\ &= \inf \{\alpha \in (0,1) | \alpha \geq 2(1 - \phi(T_n))\} = 2(1 - \phi(T_n)) \end{aligned} \tag{12}$$

where T_n is as defined above and $\phi(.)$ is the cdf of the standard normal distribution.

Question 6

$$\mathbb{V}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 | X\right]$$

Let $Z = Y^2$. Then

$$\mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] = \mathbb{E}\left[\mathbb{E}[Z|X]\right] = \mathbb{E}[Z] = \mathbb{E}\left[Y^2\right]$$

a

$$\begin{aligned}\mathbb{V}(Y|X) &= \mathbb{E}\left[Y^2|X\right] - 2\mathbb{E}\left[Y\mathbb{E}\left[Y|X\right]|X\right] + \mathbb{E}\left[\mathbb{E}\left[Y|X\right]^2|X\right] \\ &= \mathbb{E}\left[Y^2|X\right] - 2\mathbb{E}\left[Y|X\right]\mathbb{E}\left[Y|X\right] + \mathbb{E}\left[Y|X\right]^2 \\ &= \mathbb{E}\left[Y^2|X\right] - \mathbb{E}\left[Y|X\right]^2\end{aligned}$$

b

$$\begin{aligned}\mathbb{E}\left[\mathbb{V}(Y|X)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y|X\right]^2\right] \\ \mathbb{V}\left(\mathbb{E}\left[Y|X\right]\right) &= \mathbb{E}\left[\mathbb{E}\left[Y|X\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right]^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\mathbb{V}(Y|X)\right] + \mathbb{V}\left(\mathbb{E}\left[Y|X\right]\right) &= \mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right]^2 \\ &= \mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] - \mathbb{E}\left[Y\right]^2 \\ &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[Y\right]^2 \\ &= \mathbb{V}(Y)\end{aligned}$$

Question 7

To prove this, notice first that it is essentially Jensen's Inequality with conditional expectations. Thus, we will need the Chordal Slope Lemma. Also, (after defining $c := \mathbf{E}[Y|X]$) the following objects will be helpful:

$$\begin{aligned}\Delta_{+,h(c)} &:= \frac{f(c+h) - f(c)}{h} \\ \Delta_{-,h(c)} &:= \frac{f(c) - f(c-h)}{h} \\ D_+(c) &:= \lim_{h \downarrow 0} \Delta_{+,h(c)} \\ D_-(c) &:= \lim_{h \downarrow 0} \Delta_{-,h(c)},\end{aligned}$$

where f is a convex function. It is also easy to see by the Chordal Slope Lemma that $D_-(c)$ and $D_+(c)$ are bounded below and above respectively by $\Delta_{-,h(c)}$ and $\Delta_{+,h(c)}$.

Next, select an $m \in [D_-(c), D_+(c)]$, and define

$$L(x) := f(c) + m(x - c).$$

We we now want to show that $L(x) \leq f(x)$. There are three cases: when $c > x$, $c = x$, and when $c < x$. From this point on, we will replace the previous convex function f with another convex function, call it ϕ.

First consider $c = x$. The inequality holds trivially.

Next, take $c > x = c - h$. Notice that since $m \in [D_-(c), D_+(c)]$, we get:

$$\begin{aligned} m &\geq \frac{\phi(c) - \phi(x)}{c - x} \\ \phi(c) + m(x - c) &\leq \phi(x) \\ L(x) &\leq \phi(x). \end{aligned}$$

For the last case, take $c < x = c - h$. Just like above, we get:

$$\begin{aligned} m &\leq \frac{\phi(x) - \phi(c)}{x - c} \\ \phi(c) + m(x - c) &\leq \phi(x) \\ L(x) &\leq \phi(x). \end{aligned}$$

Thus, $L(x) \leq \phi(x)$.

Next, take, $x = Y$ and recall that $c := \mathbf{E}[Y|X]$. We have that

$$\begin{aligned} L(Y) &\leq \phi(Y) \\ 0 &\leq \phi(Y) - L(Y) \\ 0 &\leq \mathbf{E}[\phi(Y) - L(Y)|X] && 3) \\ 0 &\leq \mathbf{E}[\phi(Y)|X] - \mathbf{E}[L(Y)|X] && 1) \\ \mathbf{E}[L(Y)|X] &\leq \mathbf{E}[\phi(Y)|X] \\ \mathbf{E}[\phi(\mathbf{E}[Y|X])|X] + \mathbf{E}[mY|X] - \mathbf{E}[m\mathbf{E}[Y|X]|X] &\leq \mathbf{E}[\phi(Y)|X] && 1) \\ \phi(\mathbf{E}[Y|X]) + m\mathbf{E}[Y|X] - m\mathbf{E}[Y|X] &\leq \mathbf{E}[\phi(Y)|X] && 1) \ \& \ 2) \\ \phi(\mathbf{E}[Y|X]) &\leq \mathbf{E}[\phi(Y)|X]. \end{aligned}$$

And thus, our result has been obtained. The steps above can be justified from three properties of conditional expectation (the steps have been labeled accordingly). Namely: 1) $\mathbf{E}[Y + Z|X] = \mathbf{E}[Y|X] + \mathbf{E}[Z|X]$; 2) If $Y = f(X)$, then $\mathbf{E}[Y|X] = f(X)$; and 3) we know that if $\Pr(0 \leq Y) = 1$, then $\Pr(0 \leq \mathbf{E}[Y|X]) = 1$.

Question 8

a

Noting that $f(y|x) = 0$ if $f_X = 0$, we know the integral over $\mathbb{R}^{\mathbb{k}} \times \mathbb{R}$ simplifies to the integral over the area where $f_X(x) > 0$ (as it is 0 everywhere else). Therefore, we are integrating over this region, unless stated otherwise.

$$\begin{aligned} E[m^{*2}(X)] &= \int (\int y f(y|x) dy)^2 f_X(x) dx \\ &\leq \int (\int |y| \frac{f(y,x)}{f_X(x)} dy)^2 f_X(x) dx \end{aligned}$$

Knowing that $\int \frac{f(y,x)}{f_X(x)} dy = 1$, we know (i.e by Cauchy -Schwartz):

$$(\int |y| \frac{f(y,x)}{f_X(x)} dy)^2 \leq (\int y^2 \frac{f(y,x)}{f_X(x)} dy)$$

Thus, we can write out

$$\begin{aligned} E[m^{*2}(X)] &\leq \int (\int y^2 \frac{f(y,x)}{f_X(x)} dy) f_X(x) dx \\ &= \int \int (y^2 \frac{f(y,x)}{f_X(x)} f_X(x)) dy dx \\ &\leq \int \int y^2 f(x,y) dy dx \leq E(Y^2) < \infty \end{aligned}$$

where the inequality follows because $y^2 > 0$ and we expanded the integration to the entire region in the last line; and again, f_X is zero everywhere else.

b

Recall, from class that

$$\begin{aligned} E[(y - m(x))^2] =& E[(y - m(x) + m^*(x) - m^*(x))^2] \\ =& E[(y - m^*(x))^2] + 2E[(y - m^*(x))(m^*(x) - m(x))] + E[(m^*(x) - m(x))^2] \\ \geq& E[(Y - m^*(X))^2] \end{aligned}$$

Thus, we found that $\min E[(Y - m^*(X))] \Leftrightarrow E[(Y - m^*(X))m(X)] = 0$ for all m(X). Now, see that

$$\begin{aligned} E[(y - m^*(x))m(x)] &= \int \int (y - m^*(x))m(x) f(y,x) dy dx \\ &= \int (\int (y - m^*(x))m(x) f(y,x) dy) dx \\ &= \int m(x) f_X(x) (\int y f(y|x) - m^*(x) f(y|x) dy) dx \\ &= \int m(x) m^*(x) f_X(x) dx - \int m(x) m^*(x) (\int f(y|x) dy) f_X(x) dx \end{aligned}$$

As $\int f(y|x) dy$ just integrates to 1, these two terms on the left and right are equal (namely $E[(y - m^*(x))m(x)] = 0$

Question 9

Mean independence means that $\mathbb{E}[Y|X] = c$, a constant, and using the law of iterated expectations, we have that $c = \mathbb{E}[c] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$. Thus we have:

$$\begin{aligned} Cov[Y,X] &= \mathbb{E}[YX] - \mathbb{E}[X]\mathbb{E}[Y] \\ \text{(By LIE)} \quad &= \mathbb{E}[\mathbb{E}[YX|X]] - \mathbb{E}[X]\mathbb{E}[Y] \\ \text{(Because X is a function of X)} \quad &= \mathbb{E}[X\mathbb{E}[Y|X]] - \mathbb{E}[X]\mathbb{E}[Y] \\ \text{(Because of mean independence)} \quad &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0 \end{aligned} \tag{13}$$

But $Cov[Y,X] = 0$ does not imply mean independence. Take, for instance, $X \sim N(0,1)$ and $Y = X^2$. Then we have $Cov[Y,X] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0$ (because the mean and third moment of the normal are zero), but $\mathbb{E}[Y|X] = Y = X^2$, because Y is a function of X.

Also, mean independence does not imply independence. Take $Y|X \sim N(0, \sigma^2 X)$, with a non-degenerate random variable. Then we do have $\mathbb{E}[Y|X] = 0$, a constant, but the distributions of Y and X are dependent (by construction, the conditional distribution of Y changes with X, and thus cannot be equal the unconditional distribution for all X).

Question 10

Let (Y,X) be a bivariate normal random variable. Find $\mathbb{E}[Y|X]$.

$$(Y,X) \sim \mathcal{N}\left(\begin{pmatrix}\mu_Y \\ \mu_X\end{pmatrix}, \begin{pmatrix}\sigma_Y^2 & \rho\sigma_Y\sigma_X \\ \rho\sigma_Y\sigma_X & \sigma_X^2\end{pmatrix}\right)$$

Any bivariate normal random variable can be rewritten as:

$$\begin{aligned} X &= \sigma_X Z_1 + \mu_X \\ Y &= \sigma_Y \rho Z_1 + Z_2\sqrt{1-\rho^2} + \mu_Y \end{aligned}$$

This allows us to rewrite Z_1 and then Y.

$$\begin{aligned} Z_1 &= \frac{X - \mu_X}{\sigma_X} \\ Y &= \sigma_Y \rho \left(\frac{X - \mu_X}{\sigma_X}\right) + Z_2\sqrt{1-\rho^2} + \mu_Y \end{aligned}$$

Taking the expectation conditioned on X.

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{E}\left[\sigma_Y \rho \frac{X-\mu_X}{\sigma_X}|X\right] + \mathbb{E}\left[\sqrt{1-\rho^2}Z_2|X\right] + \mathbb{E}[\mu_Y|X] \\ &= \frac{\sigma_Y \rho}{\sigma_X}\mathbb{E}[X|X] - \frac{\sigma_Y \rho \mu_X}{\sigma_X} + \sqrt{1-\rho^2}\mathbb{E}[Z_2|X] + \mu_Y \\ &= \frac{\sigma_Y \rho}{\sigma_X}X - \frac{\sigma_Y \rho \mu_X}{\sigma_X} + \mu_Y\end{aligned}$$

where $\mathbb{E}[Z_2|X] = \mathbb{E}[Z_2] = 0$ by the fact that Z_1, Z_2 are independent, and X is a function of Z_1 only.

Question 11

To answer this question, we are going to need to prove the following fact: that independence of X and Y implies that $\mathbf{E}[Y|X] = \mathbf{E}[Y]$, which is a constant.

Consider the definition of conditional expectation. Since all we are given is that the first moment for Y exists, we have to work from the following definition: $\mathbf{E}[Y|X]$ is any $m^*(X)$ with $\mathbf{E}[|m^*(X)|] < \infty$ such that for any Borel set B in $\mathcal{B}$,

$$\mathbf{E}[(Y - m^*(X))\mathbb{1}_{\{X\in B\}}] = 0.$$

Working from this definition, we can obtain our result. First, let $m^*(X) = \mathbf{E}[Y]$ and B an arbitrary Borel set, then test to see if it solves the following:

$$\begin{aligned}\mathbf{E}[(Y - m^*(X))\mathbb{1}_{\{X\in B\}}] &= 0 \\ \mathbf{E}[(Y - \mathbf{E}[Y])\mathbb{1}_{\{X\in B\}}] &= 0 \\ \mathbf{E}[Y\mathbb{1}_{\{X\in B\}}] &= \mathbf{E}[\mathbf{E}[Y]\mathbb{1}_{\{X\in B\}}] \\ \mathbf{E}[Y]\mathbf{E}[\mathbb{1}_{\{X\in B\}}] &= \mathbf{E}[\mathbf{E}[Y][\mathbb{1}_{\{X\in B\}}] && \text{by } Y \perp\!\!\!\perp X \\ \mathbf{E}[Y]\mathbf{E}[\mathbb{1}_{\{X\in B\}}] &= \mathbf{E}[Y]\mathbf{E}[\mathbb{1}_{\{X\in B\}}] \\ \mathbf{E}[Y]\Pr\{X \in B\} &= \mathbf{E}[Y]\Pr\{X \in B\}.\end{aligned}$$

Since $\mathbf{E}[Y]$ works above, and $\mathbf{E}[Y|X] := m^*(X)$ we have that $\mathbf{E}[Y|X] = \mathbf{E}[Y]$ with probability one, as we wanted (using the result that the conditional expectation is unique up to a set of probability zero). Thus, $\mathbf{E}[Y|X]$ is equal to a constant with probability one, and that constant is $\mathbf{E}[Y]$.

Question 12

a

Take

$$Y = \beta_0 + \beta_1 X + U$$

Now, since we are adopting the best linear predictor interpretation, we have:

$$\beta_1 = \frac{Cov(X,Y)}{\sigma_X^2} \tag{14}$$

$$= \rho_{X,Y}\frac{\sigma_Y}{\sigma_X} \tag{15}$$

Thus, the $|\beta_1| < 1$ does not necessarily mean $\frac{Var(X)}{Var(Y)} < 1$ as we need $\frac{\beta_1}{\rho_{X,Y}} < 1$. Note, you can also see that $|\beta| < 1$ doesn't imply the claim from just writing out (and knowing that under this interpretation the covariance between X and U is zero):

$$\frac{var(Y)}{var(X)} = \frac{\beta_1^2 var(X) + var(U)}{var(X)}$$

b

As $\sigma_X = \sigma_Y$, the above equation (2) implies that we have $\beta_1 = \rho_{X,Y}$ (which has absolute value less than 1, by cauchy-schwarzs inequality). And $\beta_1 = 1$ iff $\rho_{X,Y} = 1$ (which means that one variable is a linear function of the other with probability one).

This can also be seen in the following expression: $\sigma_Y^2 = \beta_1^2\sigma_X^2 + \sigma_U^2$, where, for $\beta_1 = 1$ we require that $\sigma_U^2 = 0$, since $Cov(X,U) = 0$ in the best linear predictor interpretation.

c

Again, as we have

$$\begin{aligned}\beta_1 &= \rho_{X,Y}\frac{\sigma_Y}{\sigma_X} \\ &= \rho_{X,Y}\frac{\sigma_X}{\sigma_Y} \\ &= \alpha_1\end{aligned}$$

as the distributions (and variances) are equal. The equality of α_1 and β_1 requires, either $\rho_{X,Y} = 0$ or $\sigma_X = \sigma_Y$.

Question 13

a

i

Yes. Because we are interpreting the regression as the best linear predictor of Y given X, the vector $\beta := (\beta_0, \beta_1)'$ minimizes $\mathbb{E}\left[(Y - \overline{X}'\beta)^2\right]$, where $\overline{X} = (1, X)$. This results in the

first order conditions:

$$\begin{aligned} -2\mathbb{E}\left[\overline{X}(Y-\overline{X}'\beta)\right] &= 0 \\ \mathbb{E}\left[\begin{pmatrix}1\\X\end{pmatrix}\left(Y-\begin{pmatrix}1 & X\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\end{pmatrix}\right)\right] &= 0 \\ \mathbb{E}\left[\begin{pmatrix}Y-\beta_0-\beta_1 X\\X(Y-\beta_0-\beta_1 X)\end{pmatrix}\right] &= 0 \\ \begin{pmatrix}\mathbb{E}[U]\\\mathbb{E}[XU]\end{pmatrix} &= 0 \end{aligned} \tag{16}$$

Since $Cov[X,U] = \mathbb{E}[XU] - \mathbb{E}[X]\,\mathbb{E}[U] = 0$, we have that U and X are uncorrelated.

ii

From the regression equation $Y = \beta_0 + \beta_1 X + U$ we have:

$$\begin{aligned} \mathbb{E}[Y|X=0] &= \beta_0 + \mathbb{E}[U|X=0] \\ \mathbb{E}[Y|X=1] &= \beta_0 + \beta_1 + \mathbb{E}[U|X=1] \end{aligned} \tag{17}$$

Now we can show that $\mathbb{E}[U|X=0] = \mathbb{E}[U|X=1] = 0$ using the conditions obtained in item (i) above ($\mathbb{E}[U] = 0 = \mathbb{E}[XU]$):

$$\begin{aligned} 0 = \mathbb{E}[XU] &= \mathbb{E}[XU|X=0]\Pr(X=0) + \mathbb{E}[XU|X=1]\Pr(X=1) \\ &= 0 + \mathbb{E}[U|X=1]\Pr(X=1) \\ &\implies \mathbb{E}[U|X=1] = 0 \text{ (because } \Pr(X=1) > 0), \end{aligned} \tag{18}$$

and this also implies:

$$\begin{aligned} 0 = \mathbb{E}[U] &= \mathbb{E}[U|X=0]\Pr(X=0) + \mathbb{E}[U|X=1]\Pr(X=1) \\ &= \mathbb{E}[U|X=0]\Pr(X=0) + 0 \\ &\implies \mathbb{E}[U|X=0] = 0 \text{ (because } \Pr(X=0) > 0). \end{aligned} \tag{19}$$

Therefore, we have that:

$$\begin{aligned} \beta_0 &= \mathbb{E}[Y|X=0] \\ \beta_1 &= \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] \end{aligned} \tag{20}$$

iii

Yes. As seen in item (ii), $\mathbb{E}[Y|X] = 0$, a constant, for all possible values of X. This happens because, when we have binary variables as conditioners, the conditional expectation takes a linear form, and thus our best linear predictor is exactly equal to the conditional expectation. As a consequence, by definition of U as the diference between the the conditional expectation and the linear predictor, it will be constantly zero.

b

i

Not necessarily. The interpretation of causality only assumes that $Y = g(X.U)$, i.e., that Y can be determined as a function of X and U. But it does not assume anything else about the relationship between X and U.

ii

Notice that we can still take conditional expectations on both sides of the regression and get:

$$\begin{aligned}\beta_0 &= \mathbb{E}[Y|X=0] - \mathbb{E}[U|X=0] \\ \beta_1 &= \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] + \mathbb{E}[U|X=0] - \mathbb{E}[U|X=1]\end{aligned} \tag{21}$$

If X and U were in fact uncorrelated, then we could use the same steps in item (ii) of letter (a) above to conclude that $\mathbb{E}[U|X] = 0$ always, and obtain the same values of β. But if this is not case, the values will differ, because $\mathbb{E}[U|X]$ is not constant in X.

Question 14

The best linear predictor of Y conditioned on $\boldsymbol{X}$ is given by:

$$\min_{\boldsymbol{b}\in\mathbb{R}^3} \mathbb{E}\left[[Y - X'\boldsymbol{b}]^2\right]$$

Note that $X_1 b_1 + X_2 b_2 + X_3 b_3 = X_1(b_1 + \alpha_1 b_3) + X_2(b_2 + \alpha_2 b_3) := (\gamma_1, \gamma_2)$.

The best linear predictor of Y given (X_1, X_2) is given by:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \mathbb{E}\left[[Y - X_1\beta_1 - X_2\beta_2]^2\right]$$

It would not be possible to minimize over two dimensions and do better than minimizing over three. One could fix $b_3 = 0$ and then reach the same problem as minimizing over two dimensions.

This tells us that

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \mathbb{E}\left[[Y - X_1\beta_1 - X_2\beta_2]^2\right] \geq \min_{\boldsymbol{b}\in\mathbb{R}^3} \mathbb{E}\left[[Y - X'\boldsymbol{b}]^2\right]$$

One cannot do any worse minimizing over the two dimensions either. For any value of $\boldsymbol{b}$, choose $\boldsymbol{\gamma}$ as above, and $\mathbb{E}\left[[Y - X'\boldsymbol{b}]^2\right] = \mathbb{E}\left[[Y - (X_1, X_2)'\boldsymbol{\gamma}]^2\right]$. Thus the two dimensional case can always do as well as the three dimensional case and:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \mathbb{E}\left[[Y - X_1\beta_1 - X_2\beta_2]^2\right] \leq \min_{\boldsymbol{b}\in\mathbb{R}^3} \mathbb{E}\left[[Y - X'\boldsymbol{b}]^2\right]$$

This leads us to conclude that:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \mathbb{E}\left[[Y - X_1\beta_1 - X_2\beta_2]^2\right] = \min_{\boldsymbol{b}\in\mathbb{R}^3} \mathbb{E}\left[[Y - X'\boldsymbol{b}]^2\right]$$

This means the best linear predictor of Y given $\boldsymbol{X}$ is equivalent to the best linear predictor of Y given (X_1, X_2). Since we know that there is no perfect colinearity between (X_1, X_2) we may apply the standard Linear Regression approach.

$$\boldsymbol{\beta} = \mathbb{E}\left[(X_1, X_2)(X_1, X_2)'\right] \mathbb{E}\left[(X_1, X_2)Y\right]$$

The solution to the minimization problem over all X is any combination of b_1, b_2, b_3 such that $\boldsymbol{\beta} = (b_1 + \alpha_1 b_3, b_2 + \alpha_2 b_3)'$.

Question 15

We are given that $\mathbf{E}[Y|X] = X'\beta$, and that $Y = X'\beta + U$. This implies that $\mathbf{E}[U|X] = 0$. To see this take the conditional expectation of $Y = X'\beta + U$:

$$\begin{aligned} &\mathbf{E}[Y|X] = \mathbf{E}[X'\beta + U|X] \\ &\mathbf{E}[Y|X] = \mathbf{E}[X'\beta|X] + \mathbf{E}[U|X] && 1) \\ &\mathbf{E}[Y|X] = X'\beta + \mathbf{E}[U|X]. && 2) \end{aligned}$$

And since we are given that $\mathbf{E}[Y|X] = X'\beta$, it is immediate that:

$$\mathbf{E}[U|X] = 0.$$

As in Question 7, the steps above can be justified from two properties of conditional expectation (the steps have been labeled accordingly). Namely: 1) $\mathbf{E}[Y + Z|X] = \mathbf{E}[Y|X] + \mathbf{E}[Z|X]$; and 2) If $Y = f(X)$, then $\mathbf{E}[Y|X] = f(X)$.

Although this implies that U is mean independent of X, it does not imply independence. Notice that because Y takes values in $\{0, 1\}$, we have that $Y|X$ is Bernoulli with $p = \mathbf{E}[Y|X]$, i.e.

$$Var[Y|X] = \mathbf{E}[Y|X](1 - \mathbf{E}[Y|X]).$$

We can also observe that

$$Var[U|X] = Var[Y - X'\beta|X] = Var[Y|X] = \mathbf{E}[Y|X](1 - \mathbf{E}[Y|X]) \tag{22}$$

And since it is given that $\mathbf{E}[Y|X] = X'\beta$, we have that:

$$Var[U|X] = X'\beta(1 - X'\beta)$$

which does depend on X, unless $\beta = 0$. So it is not reasonable to assume that $Var[U|X]$ does not depend on X if we believe X has some prediction power over Y.

Question 16

Intuitively, since X and $\hat{X}$ take values in $\{0, 1\}$, we cannot have the measurement error to "cancel out" in the case of classical measurement error as if X =1, the measurement error cannot be positive and if X = 0, the measurement error must be positive, so it must be

negatively correlated with X.

Note that if E(V) = 0,

$$\begin{aligned} Cov(X,V) &= E((X - E(X))(V - E(V))) \\ &= E(XV) - E(X)E(V) \\ &= E(XV) \end{aligned}$$

Now, looking at variance of $\hat{X}$, we see that if $Cov(X,V) = E(XV) = 0$, we must have $Var(\hat{X}) = Var(X) + Var(V)$

$$\begin{aligned} Var(\hat{X}) =& E(X^2) + E(V^2) + 2E(XV) - E(\hat{X})^2 \\ =& E(X^2) - E(X)^2 + E(V^2) \\ =& Var(X) + Var(V) \end{aligned}$$

But here, as $E(X^2) = E(X) = E(\hat{X}) = E(\hat{X}^2)$ (since V has mean zero and X and $\hat{X}$ take values in $\{0, 1\}$) and,

$$var(\hat{X}) = E(\hat{X})(1 - E(\hat{X})) = var(X)$$

we then have $Var(V) = 0$, so V= 0 and $\hat{X}$ is just X.