ECOM40006/90013 Econometrics 3

Week 5 Extras: Solutions

Question 1: Reviewing the econometric toolkit

Here we provide very informal versions of each of these definitions. It is up to you to add as much additional detail as you like.

(a) (i.) **Continuous Mapping Theorem.** Let $g(\cdot)$ be continuous. Then, if:

$$\begin{aligned} X_n \xrightarrow{p} X \quad &\text{then} \quad g(X_n) \xrightarrow{p} g(X), \\ X_n \xrightarrow{d} X \quad &\text{then} \quad g(X_n) \xrightarrow{d} g(X), \\ X_n \xrightarrow{\text{a.s.}} X \quad &\text{then} \quad g(X_n) \xrightarrow{\text{a.s.}} g(X). \end{aligned}$$

This extends to multivariate formats as well.

(ii.) **Slutsky's Theorem**. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

$$\begin{aligned} X_n + Y_n &\xrightarrow{d} X + c, \\ X_n Y_n &\xrightarrow{d} Xc, \\ \frac{X_n}{Y_n} &\xrightarrow{d} \frac{X}{c} \qquad \text{provided } c \neq 0. \end{aligned}$$

This can also be used in a multivariate context.

(iii.) **Khintchine's WLLN.** Let X_n be i.i.d. with finite mean μ for every i. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}(X_i) = \mu.$$

We can also find variants of this for multivariate use.

(iv.) **Lindeberg-Lévy CLT.** Let X_n be i.i.d. with finite mean μ and variance σ^2 for every i. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0,1).$$

There are many ways this can be written. Another way is to write it in the sample mean format: if you let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of X_n, then

$$\frac{\bar{X}_n - \mu}{\text{sd}(\bar{X}_n)} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

Alternatively you can also write it as

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

As a quick summary, all the weak laws of large numbers effectively say the same thing:

$$\text{Sample averages} \xrightarrow{p} \text{expectations.}$$

Their only differences lie in the *requirements* to use those theorems. For example:

- Khintchine's WLLN: i.i.d. RVs and a finite mean.
- Chebyshev's Theorem: independent RVs and a finite mean.
- Kolmogorov's WLLN: just a sequence of RVs satisfying a really nasty condition.

For the Central Limit Theorems, they effectively all say

$$\text{Appropriately scaled averages} \xrightarrow{d} N(0, 1).$$

In most applications, the scaling factor is usually $\sqrt{n}$ but this can differ depending on the context. Just like the WLLNs, the CLTs differ mainly in their requirements. For example:

- Lindeberg-Lévy CLT: i.i.d. RVs, finite mean and variance.
- Lyapunov CLT: independent RVs with finite third central moments.
- Lindeberg-Feller CLT: independent RVs, first two moments exist, no wild variance in the tails.

Now for the rest of the definitions.

(b) (i.) **Big O.** $X_n = O(Y_n)$ if there exists a $c \neq 0$ such that

$$\lim_{n\to\infty} \left| \frac{X_n}{Y_n} \right| = c.$$

(ii.) **Little o.** $X_n = o(Y_n)$ if

$$\lim_{n\to\infty} \frac{X_n}{Y_n} = 0.$$

(iii.) **Stochastic order.** $X_n = O_p(Y_n)$ if there exists $c_1, c_2, \ldots$ such that every $c_n = O(1)$ and

$$\frac{X_n}{Y_n} \xrightarrow{p} c_n.$$

Also, $X_n = o_p(Y_n)$ if

$$\frac{X_n}{Y_n} \xrightarrow{p} 0.$$

(c) Notice that via the rules of partitioned matrix transposes, we can write

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}' = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \qquad \text{and similarly} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Note that each individual x_i is a $k \times 1$ column vector. This then makes x_i' a $1 \times k$ row vector. For your own purposes, you might also want to confirm for yourself that X is always $n \times k$ in these cases – confirming for yourself will help address multiple points of confusion that you might encounter further down the line.

In any case we can calculate these matrices directly:

$$\begin{aligned} X'X &= \begin{bmatrix} x_1 & x_2 & dots & x_n \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \\ &= x_1x_1' + x_2x_2' + \cdots + x_nx_n' \\ &= \sum_{i=1}^{n} x_i x_i'. \end{aligned}$$

Similarly

$$X'y = \begin{bmatrix} x_1 & x_2 & dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Note that in the grand scheme of things, these calculations are considered relatively straightforward, at least in comparison to the things that we're going to be doing later on. For now, understanding that you can swap between matrices and sums of smaller matrices is critical to getting your head around a number of the tricks and techniques we'll be working with in later weeks!

Question 2: The OLS estimator

(a) Manual expansion gives

$$\begin{aligned} u'u &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta'X')(y - X\beta) && \because (A+B)' = A' + B',\ (AB)' = B'A' \\ &= y'y - y'X\beta + \beta'X'y - \beta'X'X\beta \\ &= y'y - 2y'X\beta - \beta'X'X\beta \end{aligned}$$

where in the last line, we observe that $y'X\beta$ is a scalar expression, so its transpose is also a scalar and the two are equal.

(b) The OLS estimator solves the first-order condition

$$\begin{aligned}\arg\min_{\beta} u'u \implies \frac{\partial u'u}{\partial \beta} &= -2\frac{\partial y'X\beta}{\partial \beta} + \frac{\partial \beta' X'X\beta}{\partial \beta} \\ &= -2X'y + 2X'X\beta = 0\end{aligned}$$

where in order to evaluate this first-order condition, we require two rules of vector calculus: for a (conformable) vector a,

$$\frac{\partial a\beta}{\partial \beta} = a'$$

and for a symmetric matrix A,

$$\frac{\partial \beta' A\beta}{\partial \beta} = 2A\beta.$$

Further rearrangement of the FOC yields

$$X'X\beta = X'y \implies \hat{\beta} = (X'X)^{-1}X'y.$$

(c) Using the expression $y = X\beta + u$ rearrangement gives

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + u) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\ &= \beta + (X'X)^{-1}X'u\end{aligned}$$

When X is non-stochastic (i.e. it is not random), it may pass freely through the expectations operator so that

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}(\beta) + \mathbb{E}((X'X)^{-1}X'u) \\ &= \beta + (X'X)^{-1}X'\mathbb{E}(u) \\ &= \beta\end{aligned}$$

since $\mathbb{E}(u) = 0$. Hence, the conclusion that we can make about the OLS estimator here is that we can consider it unbiased in small samples.

(d) For consistency, multiply by n/n on the rightmost expression:

$$\begin{aligned}\hat{\beta} &= \beta + \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{n}X'u \\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i u_i \\ &\xrightarrow{p} \beta + \mathbb{E}(x_i x_i')^{-1}\mathbb{E}(x_i u_i) \\ &= \beta + Q^{-1}x_i\mathbb{E}(u_i) \\ &= \beta\end{aligned}$$

provided that X is nonstochastic. A couple of notes here:

- The second line changes the representation of $X'X$ to its summation counterpart so that we can use the assumptions given in the question, specifically that $\mathbb{E}(x_i x_i') = Q$.
- The third line simultaneously uses the Weak Law of Large Numbers, Slutsky's Theorem and the Continuous Mapping Theorem. Can you see where each one of these theorems needs to be used?
- The assumption of non-stochastic X is not required to show consistency. In particular, as long as $X'X$ converges to something in probability and $X'u$ has mean zero in expectation, then the same result holds.

(e) From before, we were able to determine that

$$\begin{aligned} \hat{\beta} &= \beta + \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{n}X'u \\ \implies \hat{\beta} - \beta &= \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{n}X'u \\ \implies \sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{\sqrt{n}}X'u. \end{aligned}$$

The left-hand side is now in a format where the Central Limit Theorem can be used. Let's first examine the expression on the right. Firstly, observe that we can rearrange:

$$\begin{aligned} \frac{1}{\sqrt{n}}X'u &= \sqrt{n}\left(\frac{1}{n}X'u\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i - 0\right) \end{aligned}$$

which is in an appropriate format for the Central Limit Theorem to be used (specifically the Lindeberg-Lévy CLT). So all we need to do is figure out the mean and variance of $x_i u_i$. For the mean we have $\mathbb{E}(x_i u_i) = 0$ as we already derived that when showing consistency of $\hat{\beta}$. As for the variance, we can write

$$\begin{aligned} \text{Var}(x_i u_i) &= x_i \text{Var}(u_i) x_i' \\ &= \sigma^2 x_i x_i' \\ &= \sigma^2 \mathbb{E}(x_i x_i') \\ &= \sigma^2 Q \end{aligned}$$

since X is non-stochastic: the expectation of a constant is itself, and we can go both ways with this. Therefore, we can conclude

$$\begin{aligned} \frac{1}{\sqrt{n}}X'u &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i - 0\right) \\ &\xrightarrow{d} N(0, \text{Var}(x_i u_i)) \\ &= N(0, \sigma^2 Q). \end{aligned}$$

Aside: for the case where x_i is stochastic, a bit more work is required, where we have to use variance decomposition:

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X|Y)) + \text{Var}(\mathbb{E}(X|Y))$$

Applying it to this situation,

$$\text{Var}(x_i u_i) = \mathbb{E}(\text{Var}(x_i u_i|x_i)) + \text{Var}(\mathbb{E}(x_i u_i|x_i))$$

Under the assumption of zero conditional mean, $\mathbb{E}(u_i|x_i) = 0$, we can use the Law of Iterated Expectations to write $\mathbb{E}(x_i u_i|x_i) = x_i\mathbb{E}(u_i|x_i) = 0$, getting rid of the term on the right. From there, we can pull out x_i from the conditional variance:

$$\begin{aligned}\text{Var}(x_i u_i) &= \mathbb{E}(x_i \text{Var}(u_i|x_i) x_i') \\ &= \mathbb{E}(x_i \sigma^2 x_i') \\ &= \sigma^2 \mathbb{E}(x_i x_i') \\ &= \sigma^2 Q,\end{aligned}$$

using the assumption that $Var(u_i|x_i) = \sigma^2 < \infty$, i.e. there is no heteroskedasticity. So we can get the same result even when x_i is stochastic, but we have to work more for it.

(f) We can now finish off the problem as follows: using the CMT and Slutsky's Theorem, we obtain

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n} X'X\right)^{-1} \frac{1}{\sqrt{n}} X'u \\ &\xrightarrow{d} \mathbb{E}(x_i x_i')^{-1} N(0, \sigma^2 Q) \\ &\overset{d}{=} Q^{-1} N(0, \sigma^2 Q) \\ &\overset{d}{=} N(0, \sigma^2 Q^{-1} Q Q^{-1}) \\ &\overset{d}{=} N(0, \sigma^2 Q^{-1}).\end{aligned}$$

Note that Q is symmetric by virtue of being positive definite, so the transpose of Q^{-1} is still Q^{-1}. Moving expressions over to the right-hand side gives us an approximate distribution for $\hat{\beta}$, which is:

$$\hat{\beta} \xrightarrow{d} N\left(\beta, \frac{\sigma^2 (X'X)^{-1}}{n}\right)$$

using the standard rules of variance.

Question 3: The Delta method

(a) Note that if $\hat{\beta}$ is consistent for β then

$$\hat{\beta} \xrightarrow{p} \beta \qquad \text{and so} \qquad g(\hat{\beta}) \xrightarrow{p} g(\beta)$$

by the Continuous Mapping Theorem.

(b) A first-order Taylor series expansion of $g(\hat{\beta})$ around the true value β gives

$$g(\hat{\beta}) \approx g(\beta) + g'(\beta)[\hat{\beta} - \beta].$$

It is worth mentioning again that the use of a first-order Taylor series expansion is a rather informal way of verifying the results of the Delta method. A more proper way would be to use the *Mean Value Theorem*, which gives an almost identical expression that holds with equality

$$g(\hat{\beta}) = g(\beta) + g'(\tilde{\beta})[\hat{\beta} - \beta] \qquad \tilde{\beta} \in [\beta, \hat{\beta}].$$

The key thing to note here is that in both cases, $g'(\beta)$ and $g'(\tilde{\beta})$ are both considered non-random so further steps will yield the same result; it is just that one holds with equality and the other one is only approximate.

Continuing with the first-order approximation, we have

$$\begin{aligned} g(\hat{\beta}) &\approx g(\beta) - g'(\beta)\beta + g'(\beta)\hat{\beta} \\ &\approx a + g'(\beta)\hat{\beta} \end{aligned}$$

where $a = g(\beta) - g'(\beta)\beta$ is a constant (and is thus ignored by the variance operator).

From what we found in part (a), $\hat{\beta}$ is consistent so $g(\hat{\beta})$ is consistent for $g(\beta)$, which gives the mean of $g(\hat{\beta})$ as $g(\beta)$. We'll need this for the asymptotic distribution if we wanted to derive it.

Now, the variance of $g(\hat{\beta})$ is simply

$$\begin{aligned} \text{Var}(g(\hat{\beta})) &\approx \text{Var}(a + g'(\beta)\hat{\beta}) \\ &= \text{Var}(g'(\beta)\hat{\beta}) \\ &= [g'(\beta)]^2 \text{Var}(\hat{\beta}), \end{aligned}$$

hence we can say that

$$g(\hat{\beta}) \overset{a}{\sim} N(g(\beta), [g(\beta)]^2 \text{Var}(\hat{\beta})),$$

i.e. we can use a normal distribution to approximate the distribution of $g(\hat{\beta})$. We'll talk more about how this works in the future, but keep in mind that approximate distributions are exactly that – approximate. Asymptotic distributions (e.g. when we use $\xrightarrow{d}$) only hold for infinitely large sample sizes, but the use of the $\overset{a}{\sim}$ symbol indicates that we are attempting to use this for smaller samples. Whether or not the approximation is any *good* is a different matter entirely, but it's better than nothing.

(c) Before starting, we should first check a few features of $g(\hat{\beta})$. In particular,

- $\hat{\beta}$ is also $k \times 1$ so $\hat{\beta} - \beta$ is $k \times 1$.

- $g(\beta)$ is an $m \times 1$ vector, where you can think of m as the number of hypotheses that we wish to test – at least, in a nonlinear hypothesis framework (which you might see later).
- Due to the above point, a derivative needs to be taken with respect to β'. The result is a matrix G, which is a $m \times k$ vector. In particular, let's define

$$G = \frac{\partial g(\beta)}{\partial \beta'}$$

  i.e. it is the derivative of $g(\beta)$ evaluated at the true value(s) β.

In this case, a first-order Taylor series approximation would give

$$g(\hat{\beta}) \approx g(\beta) + G[\hat{\beta} - \beta] = g(\beta) - G\beta + G\hat{\beta} = a + G\hat{\beta},$$

where $a = g(\beta) + G\beta$ is considered a constant. Now if $\hat{\beta} \sim N(0, \Sigma)$ then one has the approximate variance of $g(\hat{\beta})$ as being

$$\begin{aligned} \text{Var}(g(\hat{\beta})) &\approx \text{Var}(a + G\hat{\beta}) \\ &= \text{Var}(G\hat{\beta}) \\ &= G\text{Var}(\hat{\beta})G' \\ &= G\Sigma G', \end{aligned}$$

which is the multivariate version of the Delta method that we obtained in part (b).

Question 4: Variance estimators and confidence intervals

(a) The rank of an idempotent matrix is equal to its trace, so

$$\begin{aligned} \text{rank}(M_X) &= \text{rank}(I_N - P_X) \\ &= \text{tr}(I_N - P_X) \\ &= \text{tr}(I_N) - \text{tr}(P_X) && \because \text{tr}(A+B) = \text{tr}(A) + \text{tr}(B) \\ &= n - \text{tr}(X(X'X)^{-1}X') \\ &= n - \text{tr}(X'X(X'X)^{-1}) && \because \text{tr}(AB) = \text{tr}(BA) \\ &= n - \text{tr}(I_K) \\ &= n - k \end{aligned}$$

Hence, the rank of M_X is $n - k$.

(b) Recall from pre-tutorial 1 that $M_X e = e$. It is okay if you can't remember it, since you can still apply $e = y - X\hat{\beta}$ directly to derive the same result. This immediately gives us

$$e'e = u'M_X u.$$

(c) We write

$$\frac{e'e}{\sigma^2} = \frac{u'M_X u}{\sigma^2} = \left(\frac{u}{\sigma}\right)' M_X \left(\frac{u}{\sigma}\right) = z'M_X z$$

where $z := u/\sigma \sim N(0, I_N)$. This is because since $u \sim N(0, \sigma^2 I_N)$, we have

$$\frac{u}{\sigma} = \frac{1}{\sigma} N(0, \sigma^2 I_N) = N(0, I_n),$$

so z is multivariate standard normal. Furthermore, M_X is symmetric and idempotent with rank $n-k$, so using our results from Question 2, we can immediately conclude that

$$z' M_X z \sim \chi^2_{n-k}.$$

(d) Just take expectations:

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2) &= \mathbb{E}\left(\frac{e'e}{n-k}\right) \times \frac{\sigma^2}{\sigma^2} \\
&= \frac{\sigma^2}{n-k} \mathbb{E}\underbrace{\left(\frac{e'e}{\sigma^2}\right)}_{\sim \chi^2_{n-k}} \\
&= \frac{\sigma^2}{n-k}(n-k) \\
&= \sigma^2
\end{aligned}$$

Hence $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

(e) This one is a bit tricky. What we want to do is to turn $\text{cov}(\hat{\beta}_k, e)$ into an expression where both of the expressions are functions of u, which is where the randomness comes from. To begin with, $\hat{\beta}_k = \ell_k' \hat{\beta}$, so we can work from there:

$$\begin{aligned}
\text{cov}(\hat{\beta}_k, e) &= \text{cov}(\ell_k' \hat{\beta}, e) \\
&= \text{cov}(\ell_k' \beta + \ell_k'(X'X)^{-1} X' u, M_X u) \\
&= \text{cov}(\ell_k'(X'X)^{-1} X' u, M_X u) && \text{(Covariance ignores constants)} \\
&= \text{cov}(Au, Bu)
\end{aligned}$$

where $A = \ell_k'(X'X)^{-1}X'$ and $B = M_X$. Then,

$$\begin{aligned}
\text{cov}(Ae, Be) &= A\text{Var}(u)B' \\
&= \sigma^2 AB' \\
&= \sigma^2 \ell_k'(X'X)^{-1} X' M_X \\
&= 0
\end{aligned}$$

because $X' M_X = 0$. Hence, $\hat{\beta}_k$ and e are uncorrelated. Because they are also jointly normally distributed, we may take this one step further and claim that they are also independent.

(f) Recall that a t_{n-k} distribution satisfies

$$t_{N-K} = \frac{N(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}},$$

where the $N(0,1)$ and χ^2_{n-k} distributions are independent. We should first check for something that is $N(0,1)$ in the numerator. We have

$$\begin{aligned}\hat{\beta}_k &\sim N(\beta_k, \text{Var}(\ell_k'\hat{\beta})) \\ &\sim N(\beta_k, \ell_k'\text{Var}(\hat{\beta})\ell_k) \\ &\sim N(\beta_k, \sigma^2\ell_k'(X'X)^{-1}\ell_k)\end{aligned}$$

Now we can start moving things over to the LHS. Observe that $\text{Var}(\hat{\beta}_k)$ is a scalar so even though it is in matrix form, we can still do things like dividing by it and putting it under a square root:

$$\begin{aligned}\hat{\beta}_k - \beta_k &\sim N(0, \sigma^2\ell_k'(X'X)^{-1}\ell_k) \\ &\sim \sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}N(0,1) \\ \implies \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}} &\sim N(0,1).\end{aligned}$$

Note that we have a problem: in general

$$\underbrace{\sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}}_{\text{sd}\hat{\beta}_k} \neq \underbrace{\sqrt{\hat{\sigma}^2\ell_k'(X'X)^{-1}\ell_k}}_{\widehat{\text{sd}(\hat{\beta}_k)}}$$

and the t-statistic is precisely what we just derived, except that we have $\hat{\sigma}$ in place of σ. So we need to find some way to get σ back into this expression. First note that if we multiply by $\sqrt{\sigma^2}/\sqrt{\sigma^2}$, then

$$\begin{aligned}\frac{1}{\widehat{\text{sd}(\hat{\beta}_k)}} &= \frac{\sqrt{\sigma^2}}{\sqrt{\sigma^2\hat{\sigma}^2\ell_k'(X'X)^{-1}\ell_k}} \\ &= \sqrt{\frac{\sigma^2}{\hat{\sigma}^2}}\frac{1}{\sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}} \\ &= \frac{\left(\frac{1}{\sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}}\right)}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}.\end{aligned}$$

Now observe

$$\begin{aligned}\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} &= \sqrt{\frac{1}{\sigma^2}\frac{e'e}{n-k}} \\ &= \sqrt{\left(\frac{u'M_Xu}{\sigma^2}\right)/(n-k)} \\ &= \sqrt{\chi^2_{n-k}/(n-k)}\end{aligned}$$

Since we have already shown that $\hat{\beta}_k$ and e are independent, then any functions of these variables are also independent. Therefore, we can write

$$\frac{\hat{\beta}_k - \beta_k}{\widehat{\text{sd}(\hat{\beta}_k)}} = \frac{\left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2\ell_k'(X'X)^{-1}\ell_k}}\right)}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \sim t_{n-k}$$

as required.

(g) Generally if we want $\mathbf{P}(a < X < b)$ we can write

$$\mathbf{P}(a < X < b) = \int_a^b f_X(x)\,dx = \int_0^b f_X(x)\,dx - \int_0^a f_X(x)\,dx = F_X(b) - F_X(a).$$

So in the context of a t-distribution with $n-k$ degrees of freedom, let's define the following:

- $q_{0.975}$: the t-statistic is below this number 97.5% of the time.
- $q_{0.025}$: the t-statistic is below this number 2.5% of the time.

To be more precise, these actually represent values on the t-distribution such that the value of the CDF at that point is 0.975 and 0.025 respectively. In particular this implies that

$$\mathbf{P}(q_{0.025} < t < q_{0.975}) = \mathbf{P}(t < q_{0.975}) - \mathbf{P}(t < 0.025) = 0.975 - 0.025 = 0.95$$

so this gives us a candidate for a 95% confidence interval. So we can say that with 95% confidence that

$$\begin{aligned} & t \in (q_{0.025}, q_{0.975}) \\ \implies & \frac{\hat{\beta}_k - \beta_k}{\widehat{\mathrm{sd}(\hat{\beta}_k)}} \in (q_{0.025}, q_{0.975}) \end{aligned}$$

By the symmetry of the t-distribution around zero, we can flip the signs on the LHS:

$$\begin{aligned} & \frac{\beta_k - \hat{\beta}_k}{\widehat{\mathrm{sd}(\hat{\beta}_k)}} \in (q_{0.025}, q_{0.975}) \\ \implies & \beta_k - \hat{\beta}_k \in (\widehat{\mathrm{sd}(\hat{\beta}_k)} q_{0.025}, \widehat{\mathrm{sd}(\hat{\beta}_k)} q_{0.975}) \\ \implies & \beta_k \in (\hat{\beta}_k + \widehat{\mathrm{sd}(\hat{\beta}_k)} q_{0.025}, \hat{\beta}_k + \widehat{\mathrm{sd}(\hat{\beta}_k)} q_{0.975}). \end{aligned}$$

This gives us a 95% confidence interval for β_k.

Question 5: Variances and estimator efficiency

(a) Let's summarize the information we have: essentially

$$A \xrightarrow{p} \beta \text{ and } B \xrightarrow{p} \beta \qquad \text{but} \qquad \mathrm{Var}(A) < \mathrm{Var}(B).$$

Note that since $\mathrm{Var}(A)$ is smaller, all else constant it is the preferred estimator over B. The main point here is: just because something is consistent doesn't necessarily mean it's automatically the best one to use.

Understanding this kind of answer also requires understanding what it means to have the 'distribution' of an estimator. In essence, if we took many samples of data (which come from some underlying distribution driven by u) and computed, say, an OLS estimate for

each one of the samples, we would have a dataset of $\hat{\beta}$'s. If we were to put these in a histogram, it would represent the distribution of $\hat{\beta}$ (which would be dependent on the sample size n).

So if we calculated the two estimators from before for each of our samples and plotted their densities, we'd find the following:

- Both of them would be centered around the same point, particularly for large sample sizes.
- One of the distributions is going to be 'flatter' than the other. This would be the estimator with the higher variance.

Heuristically, if we picked a dataset at random and looked at how close A and B were to their true value, we'd find that A is closer to the true value more frequently than B is.[1]

(b) If the bias of C is small then we would seriously consider using C over D, especially with efficiency gains. In essence we want the estimator to be as close to the true value β as possible. If the bias of C only sends us off by a small amount, the efficiency benefits could still make it such that C is closer to β on average, which would be where we would use C over D.

(c) (i.) Note that Ω differs depending on what X is. The calculations are the same as the unconditional version, but we just have to tack on an $|X$ expression in our variance. So in short we just have

$$\begin{aligned}\text{Var}(u^*|X) &= \text{Var}(\Omega^{-1/2}u|X)\\ &= \Omega^{-1/2}\text{Var}(u|X)\Omega^{-1/2}\\ &= \Omega^{-1/2}\Omega\Omega^{-1/2}\\ &= I_N,\end{aligned}$$

as required. Note that in the second line, any functions of X are treated as constant due to the properties of conditional expectations and Ω is a function of X.

(ii.) OLS estimation implies

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^*\\ &= (X'\Omega^{-1/2}\Omega^{-1/2}X)^{-1}X'\Omega^{-1/2}\Omega^{-1/2}y\\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.\end{aligned}$$

(iii.) First note that if we sub in $y = X\beta + u$ we get

$$\begin{aligned}\hat{\beta}_{GLS} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(X\beta + u)\\ &= \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u.\end{aligned}$$

[1] A couple of things to note is: (i) this doesn't always happen for every single dataset of course, but on average this is what you would expect, and (ii) in practice we never observe the true value, but the theory is good for telling us what tools best suit the situation at hand.

Now the conditional variance of the GLS estimator given X is

$$\begin{aligned}\text{Var}(\hat{\beta}_{GLS}|X) &= \text{Var}(\beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u|X) \\ &= \text{Var}((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u|X) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\text{Var}(u|X)\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1}.\end{aligned}$$

(iv.) For this one, we multiply by N/N as usual and treat the calculations in a very similar manner to how we would show the same thing for the OLS estimator:

$$\begin{aligned}\hat{\beta}_{GLS} &= \hat{\beta} + \left(\frac{1}{N}X^{*\prime}X^*\right)^{-1}\frac{1}{N}X^{*\prime}u^* \\ &\xrightarrow{p} \beta + P^{-1}\mathbb{E}(X'\Omega^{-1}u) \\ &= \beta + P^{-1}\mathbb{E}(\mathbb{E}(X'\Omega^{-1}u|X)) \\ &= \beta + P^{-1}\mathbb{E}(X'\Omega^{-1}\underbrace{\mathbb{E}(u|X)}_{=0}) = \beta.\end{aligned}$$

Question 6 (bonus): OLS and block matrices

(a) This one's a straight set of calculations. In particular, observe that

$$X' = \begin{bmatrix}\ell & X_s\end{bmatrix}' = \begin{bmatrix}\ell' \\ X_s'\end{bmatrix}$$

so that

$$X'X = \begin{bmatrix}\ell' \\ X_s'\end{bmatrix}\begin{bmatrix}\ell & X_s\end{bmatrix} = \begin{bmatrix}\ell'\ell & \ell'X_s \\ X_s'\ell & X_s'X_s\end{bmatrix} = \begin{bmatrix}n & \ell'X_s \\ X_s'\ell & X_s'X_s\end{bmatrix}$$

since $\ell'\ell$ is a sum of n ones. Similarly,

$$\begin{bmatrix}\ell' \\ X_s'\end{bmatrix}y = \begin{bmatrix}\ell'y \\ X_s'y\end{bmatrix}$$

as each of the individual elements of the block matrix are conformable with y.

(b) Use $X'X\hat{\beta} = X'y$ to get

$$\begin{bmatrix}n & \ell'X_s \\ X_s'\ell & X_s'X_s\end{bmatrix}\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_s\end{bmatrix} = \begin{bmatrix}\ell'y \\ X_s'y\end{bmatrix}.$$

In a simultaneous equations format, we have

$$n\hat{\beta}_1 + \ell'X_s\hat{\beta}_s = \ell'y \tag{1}$$

$$X_s'\ell\hat{\beta}_1 + X_s'X_s\hat{\beta}_s = X_s'y. \tag{2}$$

From (1), we have

$$\begin{aligned} n\hat{\beta}_1 &= \ell' y - \ell' X_s \hat{\beta}_s \\ \implies \hat{\beta}_1 &= \frac{1}{n}\ell' y - \frac{1}{n}\ell' X_s \hat{\beta}_s \\ &= \bar{y} - \bar{X}_s' \hat{\beta}_s \end{aligned}$$

as required.

(c) From (2), using what we obtained above for $\hat{\beta}_1$ along with the fact that $1/n = (\ell'\ell)^{-1}$ gives us

$$\begin{aligned} X_s'\ell(\bar{y} - \bar{X}_s'\hat{\beta}_s) + X_s' X_s \hat{\beta}_s &= X_s' y \\ \implies X_s'\ell\bar{y} - X_s'\ell\bar{X}_s'\hat{\beta}_s + X_s' X_s \hat{\beta}_s &= X_s' y \\ \implies X_s\ell(\ell'\ell)^{-1}\ell' y - X_s'\ell(\ell'\ell)^{-1}\ell' X_s\hat{\beta}_s + X_s' X_s \hat{\beta}_s &= X_s' y \\ \implies X_s' P_1 y - X_s' P_1 X_s \hat{\beta}_s + X_s' X_s \hat{\beta}_s &= X_s' y \\ \implies X_s' X_s \hat{\beta}_s - X_s' P_1 X_s \hat{\beta}_s &= X_s' y - X_s' P_1 y \end{aligned}$$

Now some factoring out will occur. On the LHS we factor out X_s' on the left and $X_s\hat{\beta}_s$ on the right. For the RHS we factor out X_s' on the left and y on the right. This gives us

$$\begin{aligned} X_s'(I - P_1)X_s\hat{\beta}_s &= X_s'(I - P_1)y \\ \implies X_s' M_1 X_s \hat{\beta}_s &= X_s' M_1 y \\ \implies \hat{\beta}_s &= (X_s' M_1 X_s)^{-1} X_s' M_1 y, \end{aligned}$$

as required.

(d) Most of the work has already been done for us here. The key thing to note is that $M_1\ell = 0$ using our results from earlier in the preparation sheet. If we premultiply by M_1 then the intercept cancels out and using the definitions we have been given, straight OLS on the transformed data yields

$$\hat{\beta}_s = (X_s^{*\prime} X_s^*)^{-1} X_s^{*\prime} y^*$$

and plugging in our expressions give us

$$\hat{\beta}_s = (X_s' M_1 M_1 X_s)^{-1} X_s' M_1 M_1 y = (X_s' M_1 X_s)^{-1} X_s' M_1 y$$

which is the same as what we had above.