

3. [200 points] Suppose that you have just landed a job at a top economic consulting firm and that you are having a disagreement with your boss about an econometric model. You think that the data are generated by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_o + \mathbf{u}$$

where $\boldsymbol{\beta}_o \in \mathbb{R}^k$, \mathbf{X} is $n \times k$, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{u} \stackrel{\text{i.i.d.}}{\sim} (\mathbf{0}, \sigma_o^2 \mathbf{I})$.

On the other hand, your boss says that years of experience point her to the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_o + \mathbf{Z}\boldsymbol{\rho}_o + \mathbf{u}$$

where \mathbf{Z} is also $n \times k$ since, after all, “it can’t hurt to add more variables to the model”. You are not sure about that so you set out to investigate her claim.

A. [65 Points] Suppose that your model is the correct one but that you estimate the parameters by applying OLS to the competing model. Derive an expression for $\hat{\boldsymbol{\beta}}$ and show that this estimator is unbiased, stating clearly any assumptions that you make. Then derive an expression for the variance of this estimator. Apply OLS to the correct model and call the estimator $\tilde{\boldsymbol{\beta}}$. Repeat the previous steps.

Solution. To get $\hat{\boldsymbol{\beta}}$ for the first model we can appeal to the FWL Theorem and run OLS on the regression equation

$$\mathbf{M}_Z \mathbf{y} = \mathbf{M}_Z \mathbf{X} \boldsymbol{\beta} + \text{residuals}$$

where $\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. The usual OLS algebra gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{y} \quad (10 \text{ points}).$$

Replacing for the *correct* model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_o + \mathbf{u}$ gives

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_o + (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{u}.$$

The analysis for both estimators presumes that the assumptions of the classical linear model are satisfied. The estimator $\hat{\boldsymbol{\beta}}$ is unbiased (15 points) provided that $\mathbb{E}(\mathbf{u}|\mathbf{X}, \mathbf{Z}) = \mathbf{0}$ (10 points). The variance of $\hat{\boldsymbol{\beta}}$ is $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma_o^2(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}$ where we have used the fact that \mathbf{M}_Z is idempotent (10 points). By similar calculations the OLS estimator $\tilde{\boldsymbol{\beta}}$ for the

correct model is unbiased and has covariance matrix $\text{Var}(\tilde{\beta}) = \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1}$ (15 points).

Comment. Most students got full marks on this question. Points were lost for not stating the assumptions of the models as required.

B. [30 Points] How can we compare our two estimators? Invoke a well-known theorem to make your argument and then show explicitly the difference in efficiency between $\hat{\beta}$ and $\tilde{\beta}$. Which estimator is more efficient? When will there be no loss of efficiency? Start with the one-parameter case and then extend your argument to the k parameter case.

Solution. The estimator $\hat{\beta}$ was derived by applying OLS to the correct model so by the Gauss-Markov Theorem $\hat{\beta}$ should be more efficient than any other linear unbiased estimator of β_o (10 points). That is, $\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ should be a positive semidefinite matrix.

For the one parameter case ($k = 1$) we have a single regressor, \mathbf{x} . The decomposition $\mathbf{x} = \mathbf{P}_z\mathbf{x} + \mathbf{M}_z\mathbf{x}$ or the simple observation that the hypotenuse of a right-angled triangle is longer than either of the other sides allows us to write $\mathbf{x}'\mathbf{M}_z\mathbf{x} \leq \mathbf{x}'\mathbf{x}$ which in turns allows us to write

$$\sigma_o^2(\mathbf{x}'\mathbf{M}_z\mathbf{x})^{-1} \geq \sigma_o^2(\mathbf{x}'\mathbf{x})^{-1} \quad (5 \text{ points}).$$

This expression will hold with equality if and only if regressing \mathbf{x} on \mathbf{Z} has no explanatory power. Another way to say this is that \mathbf{x} lies in a space orthogonal to \mathbf{Z} or, alternatively, that the angle formed between \mathbf{x} and \mathbf{Z} is 0. (5 points)

For the k parameter case we can invoke the following result without proof.

Claim. Let \mathbf{A} and \mathbf{B} be symmetric positive definite matrices of the same dimensions. Then $\mathbf{A} - \mathbf{B}$ is positive definite if and only if $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ is positive definite.

Then we can proceed by showing that

$$\begin{aligned} \text{Var}(\hat{\beta})^{-1} - \text{Var}(\tilde{\beta})^{-1} &= \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_z\mathbf{X} \\ &= \mathbf{X}'(\mathbf{I} - \mathbf{M}_z)\mathbf{X} \\ &= \mathbf{X}\mathbf{P}_z\mathbf{X} \\ &= (\mathbf{P}_z\mathbf{X})'(\mathbf{P}_z\mathbf{X}) \end{aligned}$$

where the last step exploits the idempotency of \mathbf{P}_z . We have shown previously that a matrix that can be written as the transpose of a matrix times itself must be positive semi-definite so the result follows. This matrix will be equal to $\mathbf{0}$ if and only if $\mathbf{P}_z\mathbf{X} = \mathbf{0}$ which only holds if \mathbf{X} and \mathbf{Z} are mutually orthogonal. That is, if $\mathbf{Z}'\mathbf{X} = \mathbf{0}$. (10 points)

Comment. Points were deducted for failing to invoke the theorem and for failing to derive the one-parameter case. Analyzing the one-parameter often led to a more thorough interpretation of the difference in efficiency between the two estimators. A number of students

did an excellent job by deriving the correlation coefficient between \mathbf{X} and \mathbf{Z} and using it to analyze the difference in efficiency. This is more than I expected but it serves to illustrate the utility of asking you to analyze the one-parameter case in detail.

C. [85 Points] Now, on the contrary, suppose that it is your boss that has the correct model. Repeat the calculations you performed above. Suggest some criteria by which to evaluate the performance of your estimator and derive an expression to compare to the variance of the correctly specified model. Is one estimator unambiguously better than the other? Discuss.

Solution. The DGP is not a special case of the model being estimated so the model is misspecified. Substituting the correct (larger) model into our OLS estimate $\hat{\beta}$ and taking the appropriate conditional expectation gives

$$\mathbb{E}(\hat{\beta}) = \beta_o + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma_o. \quad (15 \text{ points})$$

Unlike the previous case, underspecifying the model will in general lead to a biased estimate. The bias is directly proportional to the magnitude of the parameter γ_o and it is unlikely that it will vanish as $n \rightarrow \infty$ so that the estimator will also be inconsistent. Note that the only cases in which the bias is $\mathbf{0}$ occurs when $\gamma_o = 0$ (the model is not misspecified) or when $\mathbf{Z}'\mathbf{X} = \mathbf{0}$ (the regressors are mutually orthogonal).

The fact that $\hat{\beta}$ is biased precludes us from invoking the Gauss-Markov Theorem to compare the estimators so instead we invoke the Mean Squared Error criterion which gives equal weight to errors arising from randomness and systematic errors arising from bias (*10 points*). The assumption on the conditional independence of the error term leads to the MSE

$$\text{MSE}(\hat{\beta}) = \sigma_o^2(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma_o\gamma_o'\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (20 \text{ points})$$

The first term reflects what the covariance matrix would look like if we were estimating a correctly specified model. The second term reflects the bias of our estimator which naturally depends on the value of γ_o and will only be zero under the conditions mentioned above. (*10 points*)

We are interested in comparing the MSE above to the MSE corresponding to the correctly specified model, $\text{Var}(\tilde{\beta}) = \sigma_o^2(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}$ (recall that this estimator is unbiased) (*10 points*). No unambiguous comparison is possible. The first term cannot be larger (ie, do worse) than $\sigma_o^2(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}$ by the same geometric argument outlined previously. If the second term (the bias) is small enough (due to γ_o or $\mathbf{Z}'\mathbf{X}$ being small) then the incorrectly specified model could be more efficient than its competitor. However, if the bias is large enough, then the incorrectly specified model could be less efficient. One could also construct a two-parameter ($k = 2$) example to show that the effect on efficiency is not uniform across coefficients. (*15 points*)

D. [20 Points] What do you conclude about your boss' claim? Do your conclusions depend in any way on the size of the sample to be analyzed?

Solution. The short answer to your boss' claim is that "it depends". If the sample to be analyzed is very large then although the bias will not disappear, the variance will be consistently estimated so that overspecifying the model is largely innocuous. On the other hand if the sample size is small then the efficiency gains from the misspecified model could possibly offset the bias. That is, the first term in the MSE above may be small enough to compensate for the magnitude of the second term (the bias) relative to the correctly specified model.

Comment. Parts C and D were graded together because students spread out their analysis of the MSE over the two sections. Therefore I was flexible in how I allocated points across questions. The majority of students understood the last two questions well. I was looking for a clear understanding of the implications of biasedness vis-a-vis the Gauss-Markov Theorem. Points were deducted for not bothering to interpret the MSE expression (a surprisingly small number of students took the time to do this). The question 'when will there be no loss of efficiency?' was not meant to be rhetorical and students were penalized for not addressing it. The derivations in this part were very straightforward but the question was worth many points because I was looking for a thorough and thoughtful discussion. Saying 'it is ambiguous' was not enough. Students that made appropriate comparisons and alluded to the large sample properties of the variance of the estimator were rewarded with high marks. Students that did not invoke the MSE criterion lost a lot of points as did students that concluded that the sample size was irrelevant. A few points were deducted for erroneous statements about the asymptotic bias.

4. [200 points] Let \mathbf{y} be $n \times 1$, \mathbf{X} be $n \times (k+1)$ and suppose that $\mathbb{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. Consider the linear programming problems

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right\}$$

and

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2,$$

subject to $\sum_{j=1}^k \beta_j^2 \leq t$

A. [20 Points] Show that there is a one-to-one correspondence between the parameters λ and t above. Offer an interpretation of these parameters and compare the objective functions above to the one for OLS.

Solution. The setup did not restrict λ and t but it should be obvious that the problem is of interest for $\lambda > 0$ and $t > 0$. The relationship between λ and t will be explored in more detail below, but at this point one can note that λ and t are inversely related. A large value of λ in the first problem implies a large penalty on the sum-of-squares of the parameters. A similar effect is obtained by smaller values for t in the second problem.

The objective functions are identical to OLS except for the constraint that penalizes by the sum-of-squares of the parameters. The parameter $\lambda > 0$ controls the amount of the penalty: the penalty increases with λ so greater values of λ lead to smaller “shrunk” values for the parameter. This estimator imposes a size constraint on the coefficients to ameliorate the high variance of coefficients in a model with many correlated variables. Another way to (informally) interpret what the estimator is doing is to recognize that any coefficient that is not shrunk towards zero must be doing a good job of minimizing the residual sum-of-squares. In this sense, λ and t control model complexity by shrinking some of the coefficients to zero.

Comment. I was very very generous with this part.

B. [15 Points] Provide a convincing argument for why the intercept is not penalized in the problems above.

Solution. Penalizing the intercept would make the procedure depend on the origin chosen for \mathbf{y} . Adding the same constant to each of the y_i 's should shift the predictions by the same amount. An estimator that penalizes the intercept would not allow the intercept to 'soak up' this shift in the y_i 's. One way to show this explicitly is to note that if a constant c were to be added to the response vector $\mathbf{y}_1 = \mathbf{y}_0 + \mathbf{c}$, then the coefficient for the intercept takes the form

$$\hat{\beta}_0 = \frac{i'(y_1 - X\beta)}{i'i} = \frac{i'(y_0 - X\beta)}{i'i} + c = \hat{\beta}_0 + c$$

Comment. Most students understood this question well and arguing in words earned you 10 points. Students that went a little further to show explicitly how the intercept 'soaks up' the shift in the response were awarded full marks. Students that simply pointed out that the intercept was not constrained received only 5 points because they didn't address the reason *why* this is done.

C. [60 Points] Now suppose the data has been centered so that the data matrix \mathbf{X} has k columns. Rewrite the first problem above in matrix form and show that the solution $\hat{\beta}$ is a linear function of \mathbf{y} . Be careful to provide conditions that allow $\hat{\beta}$ to be well defined.

Solution. Let us start by tackling the first problem as it is the relevant for the analysis that follows. After centering the first problem becomes

$$\min_{\beta} (\mathbf{y} - \mathbf{X}'\beta)'(\mathbf{y} - \mathbf{X}'\beta) + \lambda\beta'\beta.$$

The first order condition for the problem gives us a modified 'normal equation'

$$\begin{aligned} -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} + 2\lambda\hat{\beta} &= 0 \\ \Leftrightarrow (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)\hat{\beta} &= \mathbf{X}'\mathbf{y} \end{aligned}$$

where \mathbf{I}_k denotes a $k \times k$ identity matrix. To argue that the estimator is well defined one needs to address the invertibility of $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)$. One can assume that \mathbf{X} is of full rank so that $\mathbf{X}'\mathbf{X} > \mathbf{0}$ and $\lambda\mathbf{I}_k > \mathbf{0}$ imply that $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})$ is positive definite. Alternatively we can get away without assuming that \mathbf{X} is of full rank by noting that adding a positive constant to the diagonal of $\mathbf{X}'\mathbf{X}$ before inversion makes the problem nonsingular and thus well-defined even if $\mathbf{X}'\mathbf{X}$ is not of full rank. In fact, this is one of the main motivations for this estimator: when $\mathbf{X}'\mathbf{X}$ is ill-conditioned (nearly singular), this estimation approach is more robust than OLS. The original reference for this approach is Hoerl and Kennard's 1970 which in turn is related to the more general theory developed by Stein (1956) and later by Tibshirani (1996).

The discussion allows us to, one, verify that the problem satisfies the second order condition necessary for a minimum $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k) > \mathbf{0}$ and, two, derive an expression for our estimator

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)^{-1}\mathbf{X}'\mathbf{y}.$$

We immediately note that $\hat{\beta}$ is a linear function of \mathbf{y} . Several of the ideas discussed above are embodied in this expression. Relative to the usual OLS estimator this $\hat{\beta}$ is shrunken. The amount of shrinkage is determined by λ in ways that we will address below. For completeness, let us turn to the Kuhn-Tucker version of the same problem.

After centering the problem and normalizing the constraint by $t > 0$ the Lagrangian for the optimization problem becomes

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}) + \mu(\boldsymbol{\beta}'\boldsymbol{\beta}/t - 1)$$

where μ denotes the Lagrangian multiplier for this problem. The first order conditions with respect to $\boldsymbol{\beta}$ and μ are

$$\begin{aligned} -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\frac{\mu}{t}\hat{\boldsymbol{\beta}} &= 0 \\ \Leftrightarrow (\mathbf{X}'\mathbf{X} + \frac{\mu}{t}\mathbf{I}_k)\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \end{aligned}$$

and

$$\frac{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}{t} \leq 1 \quad \text{with} \quad \mu\left(\frac{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}{t} - 1\right) = 0 \quad \text{by complementary slackness.}$$

We proceed on the assumption that the conditions outlined above that allow the problem to be well defined are satisfied throughout. There are three cases to consider.

Case 1: For completeness we note that if $\hat{\boldsymbol{\beta}} = 0$ then $t > 0$ and the complementary slackness condition together imply that $\mu = 0$. The first order condition reduces to $\mathbf{X}'\mathbf{y} = 0$. In this case the data and the response are orthogonal and the true model has $\boldsymbol{\beta}_o = 0$.

Case 2: If $\hat{\boldsymbol{\beta}} \neq 0$ and $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} < t$ then $\mu = 0$ by complementary slackness. Solving the first order condition for $\boldsymbol{\beta}$ gives $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$. In this case the constraint on the size of $\hat{\boldsymbol{\beta}}$ is not binding so the constraint is irrelevant and the solution is equivalent to regular OLS.

Case 3: If $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} = t \neq 0$ then the complementary slackness condition implies that $\mu \geq 0$. Solving the first order condition for $\hat{\boldsymbol{\beta}}$ gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \frac{\mu}{t}\mathbf{I}_k)^{-1}\mathbf{X}'\mathbf{y}.$$

A couple of observations follow. 1) If the constraint is just binding because t happened to equal the sum-of-squares of $\hat{\boldsymbol{\beta}}$ then the constraint is irrelevant and the 'shadow value' of increasing t (relaxing the constraint) is $\mu = 0$. This knife-edge case leads to the OLS estimator for the coefficient. 2) Otherwise, the OLS solution would pick a $\hat{\boldsymbol{\beta}}$ of greater magnitude but the binding constraint restricts the size of the coefficient so that $\mu > 0$. 3) When we compare this result to the solution of the problem above we can see the inverse relationship between t and λ . One way to interpret this is to view λ as the normalized shadow value of relaxing the constraint.

Comment. This question was very easy for 60 points. You did not even have to solve the Kuhn-Tucker problem. Most students got full marks except those that did not state the conditions that allow the problem to be well defined. The question clearly asked you to state

them.

D. [10 Points] Is the problem well defined if $\mathbf{X}'\mathbf{X}$ is *not* of full rank?

Solution. Yes. See above.

Comment. All students got full marks on this problem except those that did not address the invertibility of $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_K)^{-1}$.

E. [60 Points] Find $\mathbb{E}(\hat{\beta}|\mathbf{X})$. Use the conditions you outlined in part C above to show that $\hat{\beta}$ is biased for β unless $\beta = 0$.

Solution. Replacing for the true model, taking the conditional expectation and adding and subtracting $\lambda\mathbf{I}_k$ gives that

$$\mathbb{E}(\hat{\beta}) = \beta_o - (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)^{-1}\beta_o\lambda.$$

Using the conditions outlined above, we know that $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)^{-1} > \mathbf{0}$ so by definition there is no nonzero vector c that gives $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_k)^{-1}c = \mathbf{0}$. Therefore, $\hat{\beta}$ is biased unless $\beta_o = \mathbf{0}$.

Comment. This question caused considerable difficulty in the sense that people derived the correct expression but did not argue correctly. The previous two questions were intended to focus your attention on the invertibility (and associated implications) of $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}c = \mathbf{0}$ and you were expected to use it in the proof. Without this argument the proof is not really rigorous enough. I gave you a hint by saying “use the conditions you outlined in part C”, but few students seemed to notice. It wasn’t meant to be tricky, quite the opposite.

F. [15 Points] Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$. Is $\hat{\beta}$ consistent for β ? What happens to $\hat{\beta}$ as $\lambda \rightarrow 0$?

Solution. Assume that $\frac{\mathbf{X}'\mathbf{X}}{n} \rightarrow_p \mathbb{E}(\mathbf{X}'\mathbf{X}) = \mathbf{Q}$ where \mathbf{Q} is a non-singular matrix and note that $\frac{\lambda}{n}\mathbf{I}_k \rightarrow_p \mathbf{0}_{k \times k}$. We can rewrite $\hat{\beta}$ and use the CMT and Slutsky theorem to show that

$$\begin{aligned}\hat{\beta} &= \beta_o - \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_k\right)^{-1}\frac{\beta_o\lambda}{n} + \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_k\right)^{-1}\frac{\mathbf{X}'u}{n} \\ &\rightarrow_p \beta_o - \mathbf{Q}^{-1} \cdot \mathbf{0} + \mathbf{Q}^{-1} \cdot \mathbf{0} \\ &= \beta_o\end{aligned}$$

where we have invoked the Law of Large Numbers for i.i.d data so that $\frac{\mathbf{X}'u}{n} \rightarrow_p \mathbb{E}(\mathbf{X}'u) = \mathbf{0}$. Though biased, the estimator is consistent for β_o . As $\lambda \rightarrow 0$, $\hat{\beta} \rightarrow \hat{\beta}^{OLS}$.

G. [20 Points] Now let $\lambda = an$ where $a > 0$ is fixed and $n \rightarrow \infty$. Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$.

Solution. Using the derivation above but replacing for $\lambda = an$ and again invoking the CMT and Slutsky theorem gives

$$\begin{aligned}
 \hat{\beta} &= \beta_o - \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_k\right)^{-1} \frac{\beta_o\lambda}{n} + \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_k\right)^{-1} \frac{\mathbf{X}'\mathbf{u}}{n} \\
 &= \beta_o - \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{an}{n}\mathbf{I}_k\right)^{-1} \frac{\beta_o an}{n} + \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{an}{n}\mathbf{I}_k\right)^{-1} \frac{\mathbf{X}'\mathbf{u}}{n} \\
 &= \beta_o - \left(\frac{\mathbf{X}'\mathbf{X}}{n} + a\mathbf{I}_k\right)^{-1} \beta_o a + \left(\frac{\mathbf{X}'\mathbf{X}}{n} + a\mathbf{I}_k\right)^{-1} \frac{\mathbf{X}'\mathbf{u}}{n} \\
 &\rightarrow \beta_o - (\mathbf{Q} + a\mathbf{I}_k)^{-1} \beta_o a + (\mathbf{Q} + a\mathbf{I}_k)^{-1} \cdot \mathbf{0} \\
 &= \beta_o - (\mathbf{Q} + a\mathbf{I}_k)^{-1} \beta_o a \\
 &\neq \beta_o
 \end{aligned}$$

unless $\beta_o = 0$ for reasons similar to Part C. If the penalty grows at the same rate as the sample size then $\hat{\beta}$ is not consistent for β_o .

Comment. Students lost points for not stating the necessary assumptions and results that allow for algebraic operations on limits (Slutsky, CMT, etc).