ARE212 FINAL EXAM

ETHAN LIGON

This is the final exam for ARE212, covering material from the second half of the course taught in Spring 2020. The exam is "take-at-home"; you may consult any resources you wish in completing it (notes, textbooks, lecture videos, etc.) except for other people. This last restriction isn't easily enforceable; I rely on you to approach this as principled adults who adhere to the Berkeley Honor Code.

More guidance:

- The exam is due at 11am on Tuesday May 12.
- In completing the exam you should develop written arguments (e.g., expressed using LaTeX or pencil and paper). In some cases you may wish to supplement these written arguments with computation, such as Monte Carlo experiments. Should you do so, please provide me with your working, open source, well-documented code. (This last could be links to a github repo, a Jupyter notebook attached to an email, or similar). In any case please be sure that materials you submit are well-organized and clearly documented—if I overlook some file you've sent or can't run it that's on you.
- You are welcome to use arguments developed in our `bcourses` discussion, but in this case please clearly cite the person and discussion (e.g., "As argued by Ligon in a discussion 'Tests of Normality' (`https://bcourses.berkeley.edu/courses/1487913/discussion_topics/5746331`), the optimal weighting matrix has a great deal of structure than can be exploited.")
- Please email files or links to `ligon@berkeley.edu`.
- If you have questions about the final I will look for these on ligonltd.slack.com, but I do not intend to be continuously available on-line, so much better if you can ask questions early!

I. Cross-Validation & Bootstrap Standard Errors

Consider estimation of a linear model $y = X\beta + u$, with the identifying assumption that $\mathrm{E}(u|X) = 0$.

When we compute K-fold cross-validation of a tuning parameter λ (e.g., the penalty parameter in a LASSO regression), then for each value of λ we obtain K estimates of any given parameter, say β_i; denote the estimates of this parameter by $b_i^\cdot = (b_i^1, \dots, b_i^K)$. If our total sample (say D_1) comprises N iid observations, then each of our K estimates will be based on a sample D_1^k of roughly $N\frac{K-1}{K}$ observations.

(1) How can you use the estimates $b_i^\cdot$ to estimate the variance of the estimator?
(2) What can you say about the variance of your estimator of the variance? In particular, how does it vary with K?

. *Date*: May 11, 2020.

(3) Suppose we use $\bar{b}(\lambda) = K^{-1}\sum_{k=1}^K b^k$ as our preferred estimate of β at a given value of the tuning parameter λ. Construct an R^2 statistic which maps a sample D and a parameter vector b into $[0, 1]$. Compare the following:
   (a) $R^2(D_1, \bar{b}(\lambda))$ and $R^2(D_1, b_{OLS})$, where b_{OLS} denotes the OLS estimator estimated using the entire sample D_1, so that $R^2(D_1, b_{OLS})$ corresponds to the usual least-squares R^2 statistic.
   (b) $R^2(D, \bar{b}(\lambda))$ and $R^2(D, b_{OLS})$, where b_{OLS} and $\bar{b}(\lambda)$ are estimated using D_1 as described above, but where D is some other iid sample from the same data-generating process.
   (c) $K^{-1}\sum_{k=1}^K R^2(D_1^k, \bar{b}(\lambda))$ and $K^{-1}\sum_{k=1}^K R^2(D_1^k, b_{OLS})$;
   (d) $K^{-1}\sum_{k=1}^K R^2(D_1^k, \bar{b}(\lambda))$ and $K^{-1}\sum_{k=1}^K R^2(D_1^k, b^k(\lambda))$;
   (e) $R^2(D, \bar{b}(\lambda))$ and $R^2(D, \beta)$;
   (f) $R^2(D, b_{OLS})$ and $R^2(D, \beta)$;

(4) How do the R^2 statistics you worked with above compare with various notions of mean-square error? The statistics which rely on β are typically infeasible, so setting these aside, how might you use these statistics to choose a "best" estimator?

II. Weighting of Linear IV Estimators

Consider the Linear IV model

$$y = X\beta + u \qquad \mathrm{E}(Z^\top u) = 0.$$

(1) Exploiting the moment condition, under what conditions does the estimator b_{IV} satisfying $Z^\top y = (Z^\top X)b_{IV}$ consistently estimate β?

(2) Suppose that Z has ℓ columns. Construct a symmetric, $\ell \times \ell$ full rank matrix W, and a corresponding estimator b_W satisfying $WZ^\top y = W(Z^\top X)b_W$. Under what conditions will this estimator consistently estimate β?

(3) Describe the GMM criterion function that b_W minimizes.

(4) Consider Hansen's description of the two-stage least squares estimator (Section 12.12). What is W for this estimator? Under what conditions is this the optimally weighted GMM estimator?

(5) $W = I$ for the b_{IV} estimator described above. Under what conditions is b_{IV} the optimally weighted GMM estimator?

(6) Describe a feasible GMM estimator for this model which is optimally weighted given an iid sampling assumption and a regularity condition that second moments of (y, X, Z) be finite.

III. RCT Design

When designing an RCT (randomized control trial), one important element of the experimental design involves *power calculations*; these in turn rely on pre-specification of the regression one proposes to estimate; this regression is generally supposed to identify one or more parameters of interest; often the question the experiment is designed to answer boils down to whether or not this parameter is different from zero, which suggests a test statistic (typically a t statistic).

So, one thing that needs to be settled early is how large the experiment needs to be to make the probability of a type II error less than some benchmark (typically 20%), holding fixed the probability of a type I error (typically 5%). A large number of examples can be found at the AEA registry; one interesting case is the registration `https://www.socialscienceregistry.org/trials/1558`, which eventually led to publication as Bandiera et al. 2020. This involves some randomly assigned *treatment*; in the example given this is a *community-level* treatment involving the establishment of clubs for adolescent girls; the (alternative) hypothesis of the study is that the establishment of such clubs will lead to greater "economic empowerment" for participating girls and "greater control over their bodies".

This particular study involved 150 communities, fifty of which were randomly assigned to be "controls", while 100 were randomly chosen to have clubs established within them. Suppose that whether a girl j lives in a community with a club depends on a binary treatment variable T_j.

(1) Suppose that we're interested in the effect of clubs on some outcome y, and so wish to estimate the parameter β_1 in $y = \beta_0 + T\beta_1 + u$. The random assignment of T implies that it is independent of u. Suggest a moment condition that could be exploited to estimate β_1.

(2) Suppose it is known in advance that the variance of y is one. The registration for this experiment indicates that about 4000 girls lived in treatment communities, while about 2000 lived in control communities. Under an iid sampling assumption, construct a t-statistic which could be used to test the hypothesis that the OLS estimate of β_1, say b_{OLS}, was significantly different from zero.

(3) Still using the OLS estimator and the iid sampling assumption, what is the "minimum detectable effect size" allowing for a probability of type I error of 5% and a probability of type II error of 20% (where the absolute value of β_1 is interpreted as the "effect size")?

(4) It is unlikely that all girls in treatment communities will actually join the "club"; instead, each will make a decision about whether to join or not; denote this by D_j equal to one if girl j joins the club, and zero otherwise. If we're interested in the effects of club participation on outcome y rather than the effects of having a club in the community, this suggests that the equation of interest ought to be something like $y = \gamma_0 + D\gamma_1 + v$. The treatment T is still randomly assigned, though of course D is not; how can this be exploited to obtain estimates of γ_1?

(5) In the registration for the RCT, the researchers proposed using the randomly assigned treatment as an instrument for girls' participation decision, and construct a just-identified two-stage least squares estimator of the coefficient corresponding to our γ_1 (in the application there are other controls, with which we won't concern ourselves). What can we say about the distribution of this estimator and distribution of the test statistic you employed to handle the power calculations? If γ_1 is the coefficient of interest, how would you go about re-doing the power calculations? What are the critical issues, and how could they be addressed?

(6) In addition to the moment conditions which identify the two-stage least squares estimator, the independence of T implies that there are many more moment conditions which could be exploited. Suggest a *sequence* of possible moment conditions, and indicate a practical estimation strategy which could make efficient use of these.

(7) The treatment T is randomly assigned to different *communities*; obviously it is not randomly assigned to different *girls* (the correlation between $(T_j, T_{j'})$ is one for girls (j, j') in the same community). Sketch a causal diagram (a directed graph) illustrating a set of assumptions sufficient for the two-stage least squares estimator to consistently estimate the model parameter γ_1. Comment on the plausibility of these assumptions; are any of these testable?

IV. Nested Samples

Consider the linear model $y = X\beta + u$, where X is thought to depend on u, but where we have a set of instruments Z such that $\mathrm{E}Z^\top u = 0$. In this case our observations on y are limited, in that we don't always observe y even when we do observe (X, Z). We can think of this as having two samples, nested in the following way. We have N_1 iid observations on the triple $(y, X, Z)_1$ but $N_2 > N_1$ iid observations on $(X, Z)_2$, with $(X, Z)_1$ (i.e., the observations on X and Z in the first dataset) a subset of $(X, Z)_2$. How can we best make use of all these data?

(1) One econometrician suggests an augmented sort of two-stage-least squares approach, using the richer dataset to estimate a linear relationship $X_2 = Z_2\pi + v$, and thus constructing a "first-stage" prediction equation $\hat{X} = Z\hat{\pi}_2$ which is more precisely estimated that it would be in the usual case in which only data in $(y, X, Z)_1$ was exploited.
   (a) Continue the argument by substituting into the second stage. What can you say about the properties of the augmented estimator compared to the properties of the usual two-stage least squares estimator?
   (b) Under what conditions would the augmented estimator be preferred to two-stage least squares on just the sample of N_1 observations?

(2) A second econometrician suggests using the smaller sample to construct a sample moment condition $(Z_1^\top y_1) = (Z_1^\top X_1)b$, and argues that if b in this condition identifies β, then it should be possible to construct $\hat{u}_2 = \hat{y}_2 - X_2 b$, and that for this larger set of observations we must have $\mathrm{E}Z_2^\top \hat{u}_2 = 0$. She argues that these two sets of moment conditions could then be combined into an over-identified optimally-weighted GMM estimator.
   (a) How would you construct the optimal GMM weighting matrix for this approach? Derive an expression for the asymptotic variance matrix for the estimator b. How does it depend on the larger sample?
   (b) Comment on this approach. Does the second set of moment conditions add useful information?
   (c) If you also knew that u was homoskedastic how could you exploit this information? How would the resulting estimator compare with two-stage-least squares? What can you say about the relative efficiency of this estimator versus two-stage least squares?