EXERCISES (CROSS-VALIDATION)

ETHAN LIGON

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

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?
(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?

. *Date*: May 2, 2022.