Problem Set # 2

Econ 722

Spring, 2018

Instructions: Answer each of the following. Question 4 requires numerical calculations. For these, please submit full source code, commented and cleanly formatted, along with your answers. All solutions must be submitted electronically on Canvas by 11:59pm on Thursday, March 29th. Late problem sets will not be accepted: it is much better to turn in partial solutions rather than nothing at all. You may discuss these problems with your classmates, but if you work together please list the names of the students with whom you have collaborated at the top of your solutions.

1. In this question you'll derive a computational shortcut for leave-one-out cross-validation in the special case of least-squares estimation. (The same basic idea holds for any linear smoother.) Let $\widehat{\beta}$ be the full-sample least squares estimator, and $\widehat{\beta}_{(t)}$ be the estimator that leaves out observation t. Similarly, let $\widehat{y}_t = \mathbf{x}_t'\widehat{\beta}$ and $\widehat{y}_{(t)} = \mathbf{x}_t'\widehat{\beta}_{(t)}$.

   (a) Let X be a $T \times p$ design matrix with full column rank, and define

   $$A = X'X = \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' = \mathbf{x}_t\mathbf{x}_t' + \sum_{k \neq t} \mathbf{x}_k\mathbf{x}_k' = A_{(t)} + \mathbf{x}_t\mathbf{x}_t'$$

   Show that

   $$A^{-1} = A_{(t)}^{-1} - \frac{A_t^{-1}\mathbf{x}_t\mathbf{x}_t'A_{(t)}^{-1}}{1 + \mathbf{x}_t'A_{(t)}^{-1}\mathbf{x}_t}$$

   where you may assume that $A_{(t)}$ is also of rank p.

   (b) Let $\{h_1, \ldots, h_T\} = diag\{\mathbf{I}_T - X(X'X)^{-1}X'\}$. Show that

   $$h_t = 1 - \mathbf{x}_t'A^{-1}\mathbf{x}_t = \frac{1}{1 + \mathbf{x}_t'A_{(t)}^{-1}\mathbf{x}_t}$$

(c) Let $\mathbf{w} = \sum_{k \neq t} \mathbf{x}_k y_k$. Now, note that we can write $\widehat{\beta} = (A_{(t)} + \mathbf{x}_t \mathbf{x}_t')^{-1}(\mathbf{w} + \mathbf{x}_t y_t)$ and $\mathbf{x}_t' \widehat{\beta}_{(t)} = \mathbf{x}_t' A_{(t)}^{-1} \mathbf{w}$. Use these facts along with the results you proved in the preceding parts to show that $(y_t - \widehat{y}_{(t)}) = (y_t - \widehat{y}_t)/h_t$.

(d) Suppose that we wanted to carry out leave-one-out cross-validation under squared error loss:

$$CV_1 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \widehat{y}_{(t)})^2$$

In light of the preceding parts, explain how we could carry out this calculation *without* explicitly calculating $\widehat{\beta}_{(t)}$ for each observation t.

2. This question asks you to derive some simple results for concerning influence functions.

(a) A functional that takes the form $\mathbf{T}(G) = \int_{-\infty}^{\infty} u(z) \, \mathrm{d}G(z)$ for some function u is called a *linear functional*. Derive the influence function of a linear functional.

(b) The mean μ of a distribution G can be expressed as a linear functional. Using part (a), show that the influence function of the mean equals $y - \mu$.

(c) Let $\mathbf{T}$ be a $\mathbb{R}$-valued functional that depends on two *other* $\mathbb{R}$-valued functionals $\mathbf{T}_1$ and $\mathbf{T}_2$ according to $\mathbf{T}(G) = h\left(\mathbf{T}_1(G), \mathbf{T}_2(G)\right)$ where h is a continuously differentiable function from $\mathbb{R}^2$ to $\mathbb{R}$. Derive an expression for the influence function $\psi(G, y)$ of $\mathbf{T}$ in terms of h and the influence functions $\psi_1(G, y), \psi_2(G, y)$ of $\mathbf{T}_1, \mathbf{T}_2$. Hint: the influence function is defined as a limit but is equivalent to a partial derivative.

(d) Use parts (a)–(c) to show that the influence function of the *variance* σ^2 of a distribution equals $(y - \mu)^2 - \sigma^2$.

3. This question asks you to fill in some of the missing details from the example comparing AIC and BIC in Lecture #4. Suppose that $Y_1, \ldots, Y_T \sim$ iid $N(\mu, 1)$. Let $\ell_T(\mu)$ denote the sample log-likelihood function evaluated at μ, where $\text{Var}(Y_i) = 1$ is assumed known.

(a) Show that $\sum_{t=1}^{T} (Y_t - \mu)^2 = T(\bar{Y} - \mu)^2 + T\widehat{\sigma}^2$ and use this result to establish that $\ell_T(\mu) = \text{Constant} - \dfrac{T}{2}(\bar{Y} - \mu)^2$.

(b) Suppose that g is a $N(\mu, 1)$ density while h is a $N(0, 1)$ density. Show that $KL(g; h) = \mu^2/2$. (If you like, you can simply apply the formula from Problem Set #1, although a direct argument is very simple.)

(c) Let $Z \sim N(0, 1)$, and $X = \mathbf{1}\{A\}$ where $A = \left\{|\sqrt{T}\mu + Z| \geq \sqrt{d_T}\right\}$. Show that

$$\mathbb{E}\left\{\left[\left(\sqrt{T}\mu + Z\right) X - \sqrt{T}\mu\right]^2\right\} = \mathbb{P}(A)\, \mathbb{E}\left[Z^2 | X = 1\right] + \left[1 - \mathbb{P}(A)\right] T\mu^2$$

(d) Continuing from the preceding part, argue that the conditional density of Z given $X = 1$ is $\mathbf{1}(A)\varphi(z)/\mathbb{P}(A)$. Using this, along with $\mathbb{E}[Z^2] = 1$, show that

$$\mathbb{P}(A)\mathbb{E}[Z^2|X = 1] = 1 - \int_a^b z^2\varphi(z)\,dz$$

(e) Continuing from the preceding part, show that

$$\int_a^b z^2\varphi(z)\,dz = a\varphi(a) - b\varphi(b) + \Phi(b) - \Phi(a)$$

(f) Combine the three preceding parts and calculate $\mathbb{P}(A)$ to show that

$$R(\mu, \widehat{\mu}) \;\;=\;\; 1 + [b\phi(b) - a\phi(a)] + (T\mu^2 - 1)\,[\Phi(b) - \Phi(a)]$$

where $a = -\sqrt{d_T} - \sqrt{T}\mu$ and $b = \sqrt{d_T} - \sqrt{T}\mu$.

4. Consider a collection of AR(p) models for $p = 1, 2, \ldots, 6$. In this question you will choose the lag order p using AIC, BIC, and cross-validation under two different true data generating processes:

$$\begin{aligned} \text{DGP1:} \quad & y_t = 0.7y_{t-1} + \varepsilon_t \\ \text{DGP2:} \quad & z_t = \varepsilon_t + 0.6\varepsilon_{t-1} \end{aligned}$$

where $\varepsilon_t \sim$ iid $\text{N}(0, 1)$ for $t = 1, \ldots, T$ and $T = 100$. Note that DGP1 is among the candidate AR(p) specifications under consideration while DGP2 is not. To answer this question, you will need to consult some papers from the shared Dropbox folder for the course: Burman, Chow & Nolan (1994); Racine (2000), Ng & Perron (2005); and Bergmeir, Hyndman & Koo (2015). In all of calculations below, carry out estimation via least-squares (the conditional maximum likelihood estimator). Note that when you estimate an AR model in this fashion, you will need to drop the first p observations in your sample, meaning that the different AR models will be use different sample sizes.

(a) Carry out a simulation study to calculate the one-step-ahead predictive MSE of each of the six AR specifications under both data generating processes. Briefly discuss your findings. In particular, you will need to carry out the following steps:

(i) Generate $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{100}, \varepsilon_{101} \sim$ iid $N(0, 1)$.

(ii) Set $y_0 = 0$ and $y_t = 0.7y_{t-1} + \varepsilon_t$ for $t = 1, 2, \ldots, 100, 101$.

(iii) Set $z_t = \varepsilon_t + 0.6\varepsilon_{t-1}$ for $t = 1, 2, \ldots, 100, 101$. (For z_1 you will need to use ε_0.)

(iv) Fit AR(p) models for $p = 1, 2, \ldots, 6$ via conditional maximum likelihood to $\{y_1, \ldots, y_{100}\}$. For each lag-length p construct an out-of-sample forecast $\widehat{y}_{101}(p)$ of y_{101} based on your fitted model.

(v) Do the same as in the preceding step for z: fit AR(p) models for $p = 1, \ldots, 6$ using $\{z_1, \ldots, z_{100}\}$ and construct an out-of-sample forecast $\widehat{z}_{101}(p)$ of z_{101}.

(vi) For each AR(p) model and each DGP, calculate the squared forecast error: $[y_{101} - \widehat{y}_{101}(p)]^2$ and $[z_{101} - \widehat{z}_{101}(p)]^2$.

(vii) Repeat the above steps 10,000 times, storing the squared forecast errors for each DGP and AR lag length in each replication. Use the sample mean of the squared forecast errors across replications to approximate one-step-ahead sample predictive MSE.

(b) Based on the discussion in Ng and Perron (2005), what are the complications in defining AIC and BIC for AR(p) models? On the basis of their simulation results, what formulas do you suggest using for AIC and BIC in this setting?

(c) Based on Burman, Chow & Nolan (1994); Racine (2000); and Bergmeir, Hyndman & Koo (2015) what are the complications in applying cross-validation to AR(p) models? How do you suggest using cross-validation to select the AR lag order?

(d) Given your choices in parts (b) and (c), carry out a simulation study with 10,000 replications comparing AIC, BIC and cross-validation under each of the two DGPs. For each DGP, calculate the fraction of replications in which a particular criterion (AIC, BIC, or Cross-Validation) selects each lag order. Briefly discuss your findings.