Problem Set # 6

Econ 722

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. 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:

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

   where $\varepsilon_t \sim$ iid $\mathrm{N}(0, 1)$ for $t = 1, \ldots, T$ and $T = 100$. Note that DGP1 is among the candidate AR(p) specifications while DGP2 is not. To answer this question, you will need to consult several 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 your answers below, estimate via least-squares (the conditional maximum likelihood estimator).

   (a) Which AR lag length minimizes one-step-ahead predictive MSE under under DGP2?

   (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 apply cross-validation to AR(p) models? What approach do you suggest for using cross-validation to select the lag order in this problem?

(d) Given your choices in parts (b) and (c), carry out a simulation study with 10,000 replications comparing AIC, BIC and cross-validation in 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.