Problem Set # 5

Econ 722

1. (Adapted from Hastie, Tibshirani & Friedman, 2008) Suppose we observe a random sample $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ from some population and decide to forecast y from $\mathbf{x}$ using the following linear model:
$$y_t = \mathbf{x}_t'\beta + \varepsilon_t$$
Let $\widehat{\beta}$ denote the ordinary least squares estimator of β based on $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$. Now suppose that we observe a *second* random sample $\{(\widetilde{\mathbf{x}}_t, \widetilde{y}_t)\}_{t=1}^T$ from the sample population that is *independent* of the first. Show that
$$E\left[\frac{1}{T}\sum_{t=1}^T (y_t - \mathbf{x}_t'\widehat{\beta})^2\right] \leq E\left[\frac{1}{T}\sum_{t=1}^T (\widetilde{y}_t - \widetilde{\mathbf{x}}_t'\widehat{\beta})^2\right]$$
In other words, show that the in-sample squared prediction error is an overly optimistic estimator of the out-of-sample squared prediction error.

2. (Adapted from Claeskens & Hjort, 2008) Leave-one-out cross-validation seems extremely computationally intensive at first blush: we need to calculate T *separate* maximum likelihood estimates! In fact, however, for a broad class of estimators that can be expressed as *linear smoothers*, there is a computational shortcut. In this question you'll examine the special case of least-squares estimation. 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.

3. This question is based on Hurvich & Tsai (1993). I will share this paper with you via Dropbox: you should read it before attempting this problem. Don't worry – it's short! Consider a VAR(p) model with no intercept

$$\begin{aligned} \underset{(q\times 1)}{\mathbf{y}_t} &= \underset{(q\times q)}{\Phi_1}\,\mathbf{y}_{t-1} + \ldots + \Phi_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon_t} \\ \boldsymbol{\epsilon}_t &\overset{iid}{\sim} N_q(\mathbf{0}, \Sigma) \end{aligned}$$

where we observe $\mathbf{y}_1, \ldots, \mathbf{y}_N$. In this question we will restrict our attention to the *conditional* maximum likelihood estimator, which reduces the problem to a multivariate regression with effective sample size $T = N - p$, namely

$$\underset{(T\times q)}{Y} = \underset{(T\times pq)}{X}\,\underset{(pq\times q)}{\Phi} + \underset{(T\times q)}{U}$$

where

$$\underset{(T\times q)}{Y} = \begin{bmatrix} \mathbf{y}_{p+1}' \\ \mathbf{y}_{p+2}' \\ \vdots \\ \mathbf{y}_N' \end{bmatrix}, \quad \underset{(pq\times q)}{\Phi} = \begin{bmatrix} \Phi_1' \\ \Phi_2' \\ \vdots \\ \Phi_p' \end{bmatrix}, \quad \underset{(T\times q)}{U} = \begin{bmatrix} \boldsymbol{\epsilon}_{p+1}' \\ \boldsymbol{\epsilon}_{p+2}' \\ \vdots \\ \boldsymbol{\epsilon}_N' \end{bmatrix}$$

and

$$\underset{(T\times pq)}{X} = \begin{bmatrix} \mathbf{y}_p' & \mathbf{y}_{p-1}' & \cdots & \mathbf{y}_1' \\ \mathbf{y}_{p+1}' & \mathbf{y}_p' & \cdots & \mathbf{y}_2' \\ \vdots & \vdots & & \vdots \\ \mathbf{y}_{N-1}' & \mathbf{y}_{N-2}' & \cdots & \mathbf{y}_{N-p-1}' \end{bmatrix}$$

(a) Derive the conditional maximum likelihood estimators for Φ and Σ as well as the maximized log-likelihood for this problem.

(b) Use your answers to the preceding part to show that, up to a scaling factor,

$$\begin{aligned}
\text{AIC} &= \log\left|\widehat{\Sigma}_p\right| + \frac{2pq^2 + q(q+1)}{T} \\
\text{BIC} &= \log\left|\widehat{\Sigma}_p\right| + \frac{\log(T)(pq^2 + q(q+1)/2)}{T}
\end{aligned}$$

(c) Show that, again up to a scaling factor,

$$\text{AIC}_c = \log\left|\widehat{\Sigma}_p\right| + \frac{(T+qp)q}{T-qp-q-1}$$

(d) Replicate rows 1,2 and 4 of Tables I and II from Hurvich & Tsai (1993). (In other words, replicate the AIC, BIC/SIC, and AIC_C results but not the AIC_C^{BD} results.) Rather than 100 simulation replications, use 1000. Note that Hurvich and Tsai use a slightly different scaling than I give in the expressions above and they also treat the constant terms from the AIC and BIC a bit differently. Does this matter for the model selection decision? Why or why not? In answering the final part of this question, you may find it helpful to read Ng & Perron (2005): although they do not consider VAR models, some of the same considerations apply.

4. Show that the influence function for maximum likelihood estimation is given by

$$\text{infl}(G, y) = J^{-1}\left(\frac{\partial \log f(y|\theta_0)}{\partial \theta}\right)$$

where

$$\int \left.\frac{\partial \log f(z|\theta)}{\partial \theta}\right|_{\theta=\mathbb{T}(G)} dG(z) = 0$$

defines the functional $\mathbb{T}$ that yields the solution $\theta_0 = \mathbb{T}$ to the ML problem.

5. In this question you will derive the simplest possible version of the FIC. Consider a linear regression model with two scalar regressors x and z

$$y_t = \theta x_t + \gamma z_t + \epsilon_t$$

where $\{(x_t, z_t, \epsilon_t)\}_{t=1}^T \sim$ iid with means $(0,0,0)$ and variances $(\sigma_x^2, \sigma_z^2, \sigma_\epsilon^2)$. The target parameter is the mean response at a *particular* covariate level (x^*, z^*). In other words we have $\mu(\theta, \gamma) = \theta x^* + \gamma z^*$ where (x^*, z^*) are fixed constants.

(a) Derive the FIC for this problem, where our goal is to choose between the *full* model, which carries out OLS estimation using both x and z, and the *narrow model* which carries out OLS estimation using x only. This corresponds to the restriction $\gamma = 0$, so we consider a DGP in which $\gamma_T = \delta/\sqrt{T}$. *None of the other parameters of the DGP vary with sample size.* The easiest way to proceed is directly from the formulas for the OLS estimators rather than via the results in Claeskens & Hjort (2003). Be sure to explain your asymptotic arguments.

(b) Compare the FIC decision rule for this problem to those of the AIC, BIC, Mallow's C_p, and the t-test of the null hypothesis $H\colon \gamma = 0$ at the $\alpha \times 100\%$ level. Comment on any relationships you uncover.