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.

Solution: Let's begin by stacking the observations in the $T\times 1$ vector $Y = [y_1, \ldots, y_T]'$

and the $T \times k$ matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_T]'$. Then consider the following:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\widehat{\beta})^2\right] - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\widetilde{y}_t - \widetilde{\mathbf{x}}_t'\widehat{\beta})^2\right] \\
=&\mathbb{E}[(\tilde{Y} - \tilde{X}\hat{\beta})'(\tilde{Y} - \tilde{X}\hat{\beta})] - \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})] \\
=&\mathbb{E}[(\tilde{Y} - \tilde{X}\hat{\beta} - \tilde{X}\tilde{\beta} + \tilde{X}\tilde{\beta})'(\tilde{Y} - \tilde{X}\hat{\beta} - \tilde{X}\tilde{\beta} + \tilde{X}\tilde{\beta})]+ \\
&- \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})] \\
=&\mathbb{E}[((\tilde{Y} - \tilde{X}\tilde{\beta}) + \tilde{X}(\tilde{\beta} - \hat{\beta}))'((\tilde{Y} - \tilde{X}\tilde{\beta}) + \tilde{X}(\tilde{\beta} - \hat{\beta}))]+ \\
&- \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})] \\
=&\mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'(\tilde{Y} - \tilde{X}\tilde{\beta})] - \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})]+ \\
&+ \mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'\tilde{X}(\tilde{\beta} - \hat{\beta})] + \mathbb{E}[(\tilde{\beta} - \hat{\beta})'\tilde{X}'(\tilde{Y} - \tilde{X}\tilde{\beta})]+ \\
&+ \mathbb{E}[(\tilde{\beta} - \hat{\beta})'\tilde{X}'\tilde{X}(\tilde{\beta} - \hat{\beta})]
\end{aligned}
$$

Now note that:

$$
\mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'(\tilde{Y} - \tilde{X}\tilde{\beta})] = \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})]
$$

since both samples have the same distributions. Further consider:

$$
\begin{aligned}
\mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'\tilde{X}(\tilde{\beta} - \hat{\beta})] &= \mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'\tilde{X}\tilde{\beta}] - \mathbb{E}[(\tilde{Y} - \tilde{X}\tilde{\beta})'\tilde{X}\hat{\beta}] \\
&= \mathbb{E}[\tilde{Y}'P_{\tilde{X}}\tilde{Y}] - \mathbb{E}[\tilde{Y}'P_{\tilde{X}}^2\tilde{Y}] - \mathbb{E}[\tilde{Y}'\tilde{X} - \tilde{Y}'\tilde{X}]\mathbb{E}[\hat{\beta}] \\
&= 0
\end{aligned}
$$

because of the property $P_{\tilde{X}}^2 = P_{\tilde{X}}$ and by independence of (X, Y) and $(\tilde{X}, \tilde{Y})$. Similarly:

$$
\mathbb{E}[(\tilde{\beta} - \hat{\beta})'\tilde{X}'(\tilde{Y} - \tilde{X}\tilde{\beta})] = 0
$$

Therefore we can finally write:

$$
\begin{aligned}
&\mathbb{E}[(\tilde{Y} - \tilde{X}\hat{\beta})'(\tilde{Y} - \tilde{X}\hat{\beta})] - \mathbb{E}[(Y - X\hat{\beta})'(Y - X\hat{\beta})] = \\
&\quad = \mathbb{E}[(\tilde{\beta} - \hat{\beta})'\tilde{X}'\tilde{X}(\tilde{\beta} - \hat{\beta})] \geq 0
\end{aligned}
$$

since $\tilde{X}'\tilde{X}$ is a positive semi-definite matrix.

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.

Solution: We want to find $A^{-1} = (A_{(t)} + \mathbf{x}_t\mathbf{x}_t')^{-1}$. The inverse A^{-1} is a matrix such that:

$$AA^{-1} = I \Leftrightarrow (A_{(t)} + \mathbf{x}_t\mathbf{x}_t')A^{-1} = I$$

premultiply by $\mathbf{x}_t'A_{(t)}^{-1}$ to get:

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

noting that $\mathbf{x}_t'A_{(t)}^{-1}\mathbf{x}_t$ is scalar we can find:

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

If, instead we premultiply the initial expression by just $A_{(t)}^{-1}$ we would get:

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

and substituting the expression we found for $\mathbf{x}_t'A^{-1}$ we can write:

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

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

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

Solution: By definition we have that $h_t = 1 - \mathbf{x}_t' A^{-1} \mathbf{x}_t$. Substituting A^{-1} with the expression found at point above we get:

$$\begin{aligned}
h_t &= 1 - \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t + \frac{(\mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t)^2}{1 + \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t} \\
&= \frac{1 + \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t - \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t + (\mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t)^2 - (\mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t)^2}{1 + \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t} \\
&= \frac{1}{1 + \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t}
\end{aligned}$$

(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 $(y_t - \widehat{y}_{(t)}) = (y_t - \widehat{y}_t)/h_t$.

Solution: First note:

$$\mathbf{w} = (A_{(t)} + \mathbf{x}_t \mathbf{x}_t')\widehat{\beta} - \mathbf{x}_t y_t$$

Now:

$$\begin{aligned}
\widehat{y}_{(t)} &= \mathbf{x}_t' A_{(t)}^{-1} \mathbf{w} = \mathbf{x}_t' \widehat{\beta} + \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t \mathbf{x}_t' \widehat{\beta} - \mathbf{x}_t' A_{(t)}^{-1} \mathbf{x}_t y_t \\
&= \widehat{y}_t - (h_t^{-1} - 1)(y_t - \widehat{y}_t)
\end{aligned}$$

Finally:

$$y_t - \widehat{y}_{(t)} = y_t - \widehat{y}_t + (h_t^{-1} - 1)(y_t - \widehat{y}_t) = (y_t - \widehat{y}_t) h_t^{-1}$$

(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.

Solution: Note that:

$$CV(1) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_{(t)})^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2/h_t^2$$

Hence, in order to compute the above quantity one just need to obtain $\widehat{\beta}$ and hence the $\widehat{y}_t$ and finally calculate the h_t all at once as $diag\{\mathbf{I}_T - X(X'X)^{-1}X'\}$.

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.