MA 575 – Qualifying Exam

Fall 2011

1. One way to obtain *bootstrap* estimates for the coefficients β from the regression

$$\mathbf{y} = X\beta + \mathbf{e}, \quad \mathrm{E}[\mathbf{e}|X] = 0, \quad \mathrm{Var}[\mathbf{e}|X] = \sigma^2 I_n, \qquad (*)$$

is based on sampling the residuals as in the procedure below:

Step 1. Obtain the LSE $\hat{\beta}$ and residuals $\hat{\mathbf{e}}$ by regressing $\mathbf{y}$ on X as in (*).

Step 2. *Bootstrap the residuals*: sample with replacement and equally likely each of the residuals in $\hat{\mathbf{e}}$ to obtain $\hat{\mathbf{e}}^*$. Note that we can write $\hat{\mathbf{e}}^* = B\hat{\mathbf{e}}$ where B is a matrix where the i-th row "selects" the i-th sampled residual, that is, if the j-th residual was sampled at the i-th time then the i-th row of B has zeros in every position but j, which has one ($B_{ij} = 1$.)

Step 3. Define bootstrap responses $\mathbf{y}^* = X\hat{\beta} + \hat{\mathbf{e}}^*$ using the residuals in the previous step. Note that $\hat{\beta}$ is fixed.

Step 4. Finally, obtain a bootstrap estimate for β as the LSE $\hat{\beta}^*$ from regressing $\mathbf{y}^*$ on X.

Let H and $\hat{\mathbf{y}}$ be the hat matrix and the fitted values from (*).

(a) Show that $\mathbf{y}^*$ is a linear combination of $\mathbf{y}$, that is, find a matrix A that depends on H and B such that $\mathbf{y}^* = A\mathbf{y}$. Is it possible for $\mathbf{y}^*$ to be $\mathbf{y}$? Explain.

(b) Show that (i) $\mathrm{Var}[\hat{\mathbf{e}}^*|X, B] = \sigma^2 B(I - H)B^T$ and so, using the result from the previous item, that

$$\mathrm{Var}[\mathbf{y}^*|X, B] = \mathrm{Var}[\hat{\mathbf{y}}|X] + \mathrm{Var}[\hat{\mathbf{e}}^*|X, B]. \qquad \text{(ii)}$$

What can you conclude from (ii) about the correlation between $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}^*$? How would you explain this result in light of the correlation between $\hat{\mathbf{y}}$ and the original residuals $\hat{\mathbf{e}}$?

(c) Suppose you regress the bootstrap residuals $\hat{\mathbf{e}}^*$ on X with mean function $\mathrm{E}[\hat{\mathbf{e}}^*|X] = X\gamma$ to obtain the LSE $\hat{\gamma}$. Now show, using the fact that regressing the fitted values $\hat{\mathbf{y}}$ on X yields the same LSE $\hat{\beta}$, that $\hat{\beta}^* = \hat{\beta} + \hat{\gamma}$. Is $\hat{\beta}^*$ unbiased for β? Explain.

(d) Show that when β includes an intercept then, on average, the bootstrap estimate for β is the LSE $\hat{\beta}$, that is, show that

$$\mathrm{E}_B[\hat{\beta}^*|X] = \hat{\beta}$$

where the expectation is taken over bootstrap samples.

2. Suppose that in an experimental study you suspect that many observations were tainted by a technician and now you want to test them *jointly* for being outliers. To this end, you organize the suspected observations as the last q observations from a total of n and adopt a *mean shift outlier model* (MSOM) on these last observations:

$$
\begin{aligned}
y_1 &= \mathbf{x}_1^T \beta + e_1 \\
&\vdots \\
y_{n-q} &= \mathbf{x}_{n-q}^T \beta + e_{n-q} \\
y_{n-q+1} &= \mathbf{x}_{n-q+1}^T \beta + \delta_1 + e_{n-q+1} \\
&\vdots \\
y_n &= \mathbf{x}_n^T \beta + \delta_q + e_n
\end{aligned}
$$

This model can be specified in matrix form by

$$
\underbrace{\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} X_1 & 0 \\ X_2 & I_q \end{bmatrix}}_{X} \begin{bmatrix} \beta \\ \delta \end{bmatrix} + \mathbf{e}
$$

where $\mathrm{E}[\mathbf{e}|X] = 0$ and $\mathrm{Var}[\mathbf{e}|X] = \sigma^2 I_n$ (as usual) and $\delta = [\delta_1 \cdots \delta_q]^T$. After some algebra, we can show that

$$
(X^T X)^{-1} = \begin{bmatrix} (X_1^T X_1)^{-1} & -(X_1^T X_1)^{-1} X_2^T \\ -X_2 (X_1^T X_1)^{-1} & I_q + X_2 (X_1^T X_1)^{-1} X_2^T \end{bmatrix}.
$$

Now consider $\hat{\beta}$ and $\hat{\delta}$, the LSE for β and δ under this model, and $\hat{\beta}_1$, the LSE for β when regressing only $\mathbf{y}_1$ on X_1, that is, when ignoring the last q observations.

(a) Show that (i) $\hat{\beta} = \hat{\beta}_1$ and (ii) $\hat{\delta} = \mathbf{y}_2 - X_2 \hat{\beta}_1$, that is, the LSE for δ is the difference between the (removed) observed values and the fitted values for X_2 in the model without the last q suspected observations.

(b) Show that the last q observations are *perfectly* fit by the MSOM: $\hat{\mathbf{y}}_2 \doteq X_2 \hat{\beta} + \hat{\delta} = \mathbf{y}_2$. What can you say about the relation between the LSE $\hat{\sigma}^2$ for σ^2 under the MSOM and the LSE $\hat{\sigma}_1^2$ for σ^2 under the model without the last q observations?

(c) Find the hat matrix for the MSOM and comment on the leverage for the suspected data points in light of the results from the previous item.

(d) Conduct a joint outlier test by testing $\delta_1 = \cdots = \delta_q = 0$. State the test statistic and its distribution under the null.