Text Exercise 3

(a) Consider the usual linear model, where $y = X\beta + \varepsilon$. We now compare two regressions which differ in how many variables are included in the matrix X. In the full (unrestricted) model p_1 regressors are included. In the restricted model only a subset of $p_0 < p_1$ regressors are included. Show that the smallest model is preferred according to the AIC if $\frac{s_0^2}{s_1^2} < e^{\frac{2}{n}(p_1 - p_0)}$.

Answer:

$$\text{Unrestricted:} \qquad \text{AIC}_1 = \log(s_1^2) + \frac{2p_1}{n}$$

$$\text{Restricted:} \qquad \text{AIC}_0 = \log(s_0^2) + \frac{2p_0}{n}$$

Smallest model preferred :

$$\begin{aligned}
\text{AIC}_0 &< \text{AIC}_1 \\
\text{AIC}_0 - \text{AIC}_1 &< 0 \\
\log(s_1^2) + \frac{2p_1}{n} - \log(s_0^2) - \frac{2p_0}{n} &< 0 \\
\log\left(\frac{s_0^2}{s_1^2}\right) &< \frac{2}{n}(p_1 - p_0) \\
\therefore \frac{s_0^2}{s_1^2} &< e^{\frac{2}{n}(p_1 - p_0)}
\end{aligned}$$

(b) Argue that for very large values of n the inequality of (a) is equal to the condition $\frac{s_0^2 - s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$. Use that $e^x \approx 1 + x$ for small values of x.

Answer:

Since $e^x \approx 1 + x$ for small x :

$e^{\frac{2}{n}(p_1 - p_0)} \approx 1 + \dfrac{2}{n}(p_1 - p_0)$ since n is large $\rightarrow \dfrac{2}{n}(p_1 - p_0)$ is small

From (a):

$$\frac{s_0^2}{s_1^2} < 1 + \frac{2}{n}(p_1 - p_0)$$
$$\frac{s_0^2}{s_1^2} - 1 < \frac{2}{n}(p_1 - p_0)$$
$$\therefore \frac{s_0^2 - s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$$

(c) Show that for very large values of n the condition in (b) is approximately equal to $\frac{e'_R e_R - e'_U e_U}{e'_U e_U} < \frac{2}{n}(p_1 - p_0)$ where e_R is the vector of residuals for the restricted model with p_0 parameters and e_U the vector of residuals for the full unrestricted model with p_1 parameters.

Answer:

$$
\begin{aligned}
&\text{By definition:}\\
&s_1^2 = \frac{1}{p_1} e'_U e_U\\
&s_0^2 = \frac{1}{p_0} e'_R e_R
\end{aligned}
$$

From (b):

$$
\begin{aligned}
\frac{s_0^2}{s_1^2} &< 1 + \frac{2}{n}(p_1 - p_0)\\
\frac{\frac{1}{p_0} e'_R e_R}{\frac{1}{p_1} e'_U e_U} &< 1 + \frac{2}{n}(p_1 - p_0)\\
\frac{p_1}{p_0}\frac{e'_R e_R}{e'_U e_U} &< 1 + \frac{2}{n}(p_1 - p_0)
\end{aligned}
$$

$$
\begin{aligned}
&p_0 < p_1 \to \frac{p_1}{p_0} > 1\\
&\text{Above implies } \frac{p_1}{p_0} = 1 + \epsilon \text{ where } \epsilon > 0
\end{aligned}
$$

$$
\begin{aligned}
&\text{Therefore:}\\
(1+\epsilon)\frac{e'_R e_R}{e'_U e_U} &< 1 + \frac{2}{n}(p_1 - p_0)\\
\frac{e'_R e_R}{e'_U e_U} &< 1 + \frac{2}{n}(p_1 - p_0) \qquad \text{since } 1 < 1 + \epsilon\\
\therefore \frac{e'_R e_R - e'_U e_U}{e'_U e_U} &< \frac{2}{n}(p_1 - p_0)
\end{aligned}
$$

(d) Finally, show that the inequality from (c) is approximately equivalent to an F-test with critical value 2, for large sample sizes.
Answer:

$$F = \frac{(e'_R e_R - e'_U e_U)/(p_1 - p_0)}{e'_U e_U/(n - p_1)}$$

When $n >> p_1 \rightarrow (n - p_1) \approx n :$

$$F \approx \frac{(e'_R e_R - e'_U e_U)/(p_1 - p_0)}{e'_U e_U/(n)}$$

From (c):

$$\begin{aligned}
\frac{e'_R e_R - e'_U e_U}{e'_U e_U} &< \frac{2}{n}(p_1 - p_0) \\
\frac{e'_R e_R - e'_U e_U/(p_1 - p_0)}{e'_U e_U/n} &< 2 \\
F &< 2
\end{aligned}$$