Chapter 2 Appendices

Appendix 2A. Distributional Results in the Central Case

2A.1. Distributions

Assume all distributions are central. We will take advantage of the well-known property of hierarchical (or nested) models that for $k \geq k_*$ and $L > 0$,

$$\mathrm{SSE}_k - \mathrm{SSE}_{k+L} \sim \sigma_*^2 \chi_L^2, \tag{2A.1}$$

$$\mathrm{SSE}_k \sim \sigma_*^2 \chi_{n-k}^2 \tag{2A.2}$$

and

$$\mathrm{SSE}_k - \mathrm{SSE}_{k+L} \text{ is independent of } \mathrm{SSE}_{k+L}. \tag{2A.3}$$

We will also need the distribution of linear combinations of SSE_k and SSE_{k+L}. Consider the linear combination $a\mathrm{SSE}_k - b\mathrm{SSE}_{k+L}$ where a and b are scalars. It follows from Eq. (2A.2) that

$$E[a\mathrm{SSE}_k - b\mathrm{SSE}_{k+L}] = a(n-k)\sigma_*^2 - b(n-k-L)\sigma_*^2 \tag{2A.4}$$

and

$$\begin{aligned} var[a\mathrm{SSE}_k - b\mathrm{SSE}_{k+L}] &= var[a\mathrm{SSE}_k - a\mathrm{SSE}_{k+L} + a\mathrm{SSE}_{k+L} - b\mathrm{SSE}_{k+L}] \\ &= var[a(\mathrm{SSE}_k - \mathrm{SSE}_{k+L}) + (a-b)\mathrm{SSE}_{k+L}]. \end{aligned}$$

Applying Eqs. (2A.1)–(2A.3), we have

$$var[a\mathrm{SSE}_k - b\mathrm{SSE}_{k+L}] = 2a^2 L\sigma_*^4 + 2(a-b)^2(n-k-L)\sigma_*^4. \tag{2A.5}$$

Since many model selection criteria (such as AIC) use some function of $\log(\mathrm{SSE}_k)$, we will also introduce some useful distributional results involving $\log(\mathrm{SSE}_k)$. It can be shown (Gradshteyn, 1965, p. 576) that

$$\int_0^\infty z^{\nu-1} e^{-\mu z} \log(z)\,dz = \frac{1}{\mu^\nu}\Gamma(\nu)\left[\psi(\nu) - \log(\mu)\right],\ \mu > 0, \nu > 0,$$

where

$$\psi(\nu) = -C - \sum_{j=0}^{\infty}\left(\frac{1}{j+\nu} - \frac{1}{j+1}\right),$$

C = 0.577 215 664 901 is Euler's constant, and ψ is Euler's psi function. For $Z \sim \chi^2_m$ with m degrees of freedom,

$$E\left[\log(Z)\right] = \int_0^{\infty} \log(z) \frac{1}{2^{m/2}\Gamma(m/2)} z^{m/2-1} e^{-z/2} dz$$
$$= \log(2) + \psi\left(\frac{m}{2}\right),$$

which has no closed-form solution. For $\mathrm{SSE}_k \sim \sigma_*^2 \chi^2_{n-k}$,

$$E\left[\log(\mathrm{SSE}_k)\right] = \log(\sigma_*^2) + \log(2) + \psi\left(\frac{n-k}{2}\right). \qquad (2A.6)$$

Although ψ has no closed-form solution, a simple recursion exists which is useful for computing exact expectations in small samples (Gradshteyn, 1965 pp. 943–945):

$$\psi(v+1) = \psi(v) + \frac{1}{v},\ v > 0, \qquad (2A.7)$$

where

$$\psi(\tfrac{1}{2}) = -C - 2\log(2) \text{ and } \psi(1) = -C. \qquad (2A.8)$$

This recursion will be used to check the accuracy of the Taylor expansion derived in Section 2.3 as well as in studying small-sample properties in Section 2.6.

The distribution of differences between $\log(\mathrm{SSE}_k)$ and $\log(\mathrm{SSE}_{k+L})$ is more involved. Since we have differences of logs,

$$\log(\mathrm{SSE}_{k+L}) - \log(\mathrm{SSE}_k) = \log\left(\frac{\mathrm{SSE}_{k+L}}{\mathrm{SSE}_k}\right).$$

Let $Q = \mathrm{SSE}_{k+L}/\mathrm{SSE}_k$. Assuming nested models and applying Eqs. (2A.1)–(2A.3), we have

$$Q \sim \frac{\chi^2_{n-k-L}}{\chi^2_{n-k-L} + \chi^2_L}.$$

In nested models these two χ^2 are independent, and Q has the Beta distribution

$$Q \sim \mathrm{Beta}\left(\frac{n-k-L}{2}, \frac{L}{2}\right).$$

Since $Q = \mathrm{SSE}_{k+L}/\mathrm{SSE}_k$, the log-distribution is

$$\log\left(\frac{\mathrm{SSE}_{k+L}}{\mathrm{SSE}_k}\right) \sim \text{log-Beta}\left(\frac{n-k-L}{2}, \frac{L}{2}\right). \qquad (2A.9)$$

It can be shown (Gradshteyn, 1965 p. 538 and p. 541) that

$$\int_0^1 t^{\mu-1}(1-t^r)^{\nu-1}\log(t)dt = \frac{1}{r^2}B\left(\frac{\mu}{r}+\nu,\nu\right)\left[\psi\left(\frac{\mu}{r}\right)-\psi\left(\frac{\mu}{r}+\nu\right)\right]$$

and

$$\begin{aligned}\int_0^1 t^{\mu-1}(1-t^r)^{\nu-1}\log^2(t)dt =& B\left(\frac{\mu}{r}+\nu,\nu\right)\left[\left(\psi\left(\frac{\mu}{r}\right)-\psi\left(\frac{\mu}{r}+\nu\right)\right)^2\right. \\ &\left. + \psi'(\mu)-\psi'(\mu+\nu)\right],\end{aligned}$$

where

$$\psi'(v)=\sum_{j=0}^{\infty}\frac{1}{(j+v)^2}$$

and $\mu > 0, \nu > 0$. For $Q \sim Beta(\frac{m}{2}, \frac{L}{2})$,

$$\begin{aligned}E\left[\log(Q)\right] &= \int_0^1 \log(t)B^{-1}\left(\frac{m}{2},\frac{L}{2}\right)t^{\frac{m}{2}-1}(1-t)^{\frac{L}{2}-1}dt \\ &= \psi\left(\frac{m}{2}\right)-\psi\left(\frac{m}{2}+\frac{L}{2}\right)\end{aligned}$$

and

$$\begin{aligned}E\left[\log^2(Q)\right] &= \int_0^1 \log^2(t)B^{-1}\left(\frac{m}{2},\frac{L}{2}\right)t^{\frac{m}{2}-1}(1-t)^{\frac{L}{2}-1}dt \\ &= \left(\psi\left(\frac{m}{2}\right)-\psi\left(\frac{m}{2}+\frac{L}{2}\right)\right)^2+\psi'\left(\frac{m}{2}\right)-\psi'\left(\frac{m}{2}+\frac{L}{2}\right).\end{aligned}$$

Hence,

$$var\left[\log(Q)\right]=\psi'\left(\frac{m}{2}\right)-\psi'\left(\frac{m}{2}+\frac{L}{2}\right)$$

and

$$var\left[\log\left(\frac{\mathrm{SSE}_{k+L}}{\mathrm{SSE}_k}\right)\right]=\psi'\left(\frac{n-k-L}{2}\right)-\psi'\left(\frac{n-k}{2}\right). \tag{2A.10}$$

Eq. (2A.10) also has no closed-form, but again, convenient recursions for ψ' exist which are useful in small samples (Gradshteyn, 1965 pp. 945–946):

$$\psi'\left(\frac{m}{2}\right)=\psi'\left(\frac{m}{2}-1\right)-\frac{4}{(m-2)^2}, \tag{2A.11}$$