Solution 1

(a) The $F_{1,8}$ critical value for a test at level 5% is 5.32.

Forward selection At each step we consider adding the variable that most reduces RSS.

- Initial model : $y = \beta_0 + \epsilon$
- Step 1 : $y = \beta_0 + \beta_4 x_4 + \epsilon$, $F = (2715.8 - 883.9)/1 \div 47.9/(13-5) = 305.95 > 5.32$.
- Step 2 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$, $F = 135.13 > 5.32$.
- Step 3: $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 4.47 < 5.32$.

Final model : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

Backward selection For each step, we consider removing the variable inducing the lowest RSS increase.

- Initial model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- Step 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$, $F = (48 - 47.9/1 \div 47.9/(13-5) = 0.0167 < 5.32$.
- Step 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 1.65 < 5.32$.
- Step 3 : $y = \beta_0 + \beta_2 x_2 + \epsilon$, $F = 141.70 > 5.32$.

Final model : $y = \beta_0 + \beta_2 x_2 + \beta_1 x_1 + \epsilon$.

(b) i) Mallows C_p is similar to AIC: the model with lowest C_p is preferred. To compute the missing C_p, we need to know s^2, which can be found using any known C_p, or

$$s^2 = \frac{\|e_{\text{full}}\|^2}{n-p} = \frac{\text{RSS}_{\text{full}}}{13-5} = \frac{47.9}{8} = 5.99.$$

The completed table is:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
$- - - -$	2715.8	442.58	$1\ 2 - -$	57.9	2.67	$1\ 2\ 3 -$	48.1	3.03
			$1 - 3 -$	1227.1	197.94	$1\ 2 - 4$	48.0	3.02
$1 - - -$	1265.7	202.39	$1 - - 4$	74.8	5.49	$1 - 3\ 4$	50.8	3.48
$- 2 - -$	906.3	142.37	$- 2\ 3 -$	415.4	62.38	$- 2\ 3\ 4$	73.8	7.325
$- - 3 -$	1939.4	314.90	$- 2 - 4$	868.9	138.12			
$- - - 4$	883.9	138.62	$- - 3\ 4$	175.7	22.34	$1\ 2\ 3\ 4$	47.9	5

ii) Forward selection leads to $y = \beta_0 + \sum_{i \in \{1,2,4\}} \beta_i x_i$, whereas backward selection gives $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. The latter has the lowest C_p, so can be regarded as best overall.

Solution 2

(a) The log-likelihood function is easily derived from the normal density function, and we ignore the additive constants $-n \log(2\pi)/2$.

$\ell(\beta, \sigma^2)$ is maximised with respect to β by minimising the sum of squares, and we saw in the week 1 lectures that (under the conditions given) that the given formula for $\widehat{\beta}$ is the least squares (and therefore the maximum likelihood) estimate. This does not depend on σ^2, so it is the overall MLE for β, and

$$\ell(\widehat{\beta}, \sigma^2) \equiv -\tfrac{1}{2}\left\{n \log \sigma^2 + (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})/\sigma^2\right\},$$

differentiation of which leads to the given formula for $\widehat{\sigma}^2$ (note that this gives a maximum of the log-likelihood). Moreover $y - X\widehat{\beta} = (I_n - H)y$, giving the other formulae for $\widehat{\sigma}^2$. Finally,

$$\mathrm{AIC} = -2\{\ell(\widehat{\beta}, \widehat{\sigma}^2 - (p+1)\} = n \log \widehat{\sigma}^2 + n + 2p + 2 \equiv n \log \widehat{\sigma}^2 + 2p \equiv n \log \mathrm{RSS}_p + 2p,$$

as required.

(b) We can add any constant we like to AIC and leave the results unchanged, so consider

$$\mathrm{AIC} - n \log \widehat{\sigma}_0^2 = n \log\left\{1 + (\widehat{\sigma}^2 - \widehat{\sigma}_0^2)/\widehat{\sigma}_0^2\right\} + 2p \approx n \frac{\widehat{\sigma}^2 - \widehat{\sigma}_0^2}{\widehat{\sigma}_0^2} + 2p = \frac{\mathrm{RSS}_p}{\widehat{\sigma}_0^2} + 2p - n,$$

with the approximation valid when $(\widehat{\sigma}^2 - \widehat{\sigma}_0^2)/\widehat{\sigma}_0^2$ is not too large. As the last expression is C_p, minimising either this or AIC will tend to give the same model.

Solution 3

(a) This is just (twice) the difference in log-likelihood functions, and as

$$2\sum_{j=1}^{n} \log g(Y_j^+) = -n \log \sigma^2 - \frac{1}{\sigma^2}\sum_{j=1}^{n}(Y_j^+ - \mu_j)^2$$

we easily obtain the first expression. To take the inner expectation we note that

$$\mathrm{E}_g^+\{(Y_j^+ - \mu_j)^2\} = \sigma^2, \quad \mathrm{E}_g^+\{(Y_j^+ - \widehat{\mu}_j)^2\} = \sigma^2 + (\widehat{\mu}_j - \mu_j)^2,$$

and substituting these into the first expression gives the second one.

(b) In this case $\widehat{\beta} \sim \mathcal{N}_p\{\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\}$ is independent of the residual sum of squares S^2, and $n\widehat{\sigma}^2 = (n-1)S^2 \stackrel{\mathrm{D}}{=} \sigma^2\chi^2_{n-p}$. Hence $\widehat{\mu} - \mu = X(\widehat{\beta} - \beta)$ is independent of $\widehat{\sigma}^2$, and

$$\sum_{j=1}^{n}(\widehat{\mu}_j - \mu_j)^2 = (\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu) = (\widehat{\beta} - \beta)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - \beta)\sigma^2\chi^2_p,$$

owing to the general result that if $Z \sim \mathcal{N}_p(\gamma, \Omega)$ then $(Z - \gamma)^{\mathrm{T}}\Omega^{-1}(Z - \gamma) \sim \chi^2_p$. The expectation is

$$\mathrm{E}_g\left[\sum_{j=1}^{n}\left\{\log \widehat{\sigma}^2 + \frac{\sigma^2}{\widehat{\sigma}^2} + \frac{(\mu_j - \widehat{\mu}_j)^2}{\widehat{\sigma}^2} - \log \sigma^2 - 1\right\}\right],$$

and this reduces to

$$n\mathrm{E}_g(\log \widehat{\sigma}^2) + n\sigma^2\mathrm{E}_g(1/\widehat{\sigma}^2) + \mathrm{E}_g\{(\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu)\}\mathrm{E}_g(1/\widehat{\sigma}^2) - n \log \sigma^2 - n,$$

using the independence of $\widehat{\mu}$ and σ^2. Now

$$\mathrm{E}_g(1/\widehat{\sigma}^2) = \mathrm{E}_g\left\{\frac{n}{\sigma^2 V_{n-p}}\right\} = \frac{n}{\sigma^2(n-p-2)}, \quad \mathrm{E}_g\{(\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu)\} = p\sigma^2,$$

where V_ν has the χ^2_ν distribution, and this yields the given expression.

Additive constants not depending on p can be ignored. and dropping such terms yields the final expression.

(c) This also just uses some Taylor series expansions for small p/n.