Problem 1 (Automatic model selection) We consider the dataset on cement properties used in the lectures. The residual sum of squares (RSS) and Mallows C_p for the models with intercept are:

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
– – – –	2715.8	442.58	1 2 – –	57.9		1 2 3 –	48.1	
			1 – 3 –	1227.1	197.94	1 2 – 4	48.0	
1 – – –	1265.7	202.39	1 – – 4	74.8	5.49	1 – 3 4	50.8	
– 2 – –	906.3		– 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

(a) Use forward selection and backward elimination based on F statistics with significance level 5% to choose variables.

(b) Another selection criterion is Mallows C_p, i.e.,

$$C_p = \frac{\mathrm{RSS}_p}{s^2} + 2p - n,$$

where RSS_p is the residual sum of squares for a model with p covariates and s^2 is the variance estimate for the full model.

(i) How do we use this criterion? Compute the values of C_p missing from the table above.

(ii) What models are selected with C_p, using forward selection and backward elimination? What is the overall best model?

Problem 2 (AIC and C_p)

(a) Show that the log-likelihood for the model $y \sim \mathcal{N}_n(X_{n\times p}\beta, \sigma^2 I_n)$, where $n > p$ and X is of rank p, is

$$\ell(\beta, \sigma^2) \equiv -\tfrac{1}{2}\left\{n \log \sigma^2 + (y - X\beta)^{\mathrm{T}}(y - X\beta)/\sigma^2\right\}, \quad \beta \in \mathbb{R}^p, \sigma^2 > 0,$$

where $\equiv$ means that additive constants have been ignored, deduce that the maximum likelihood estimates are

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \quad \widehat{\sigma}^2 = (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})/n = y^{\mathrm{T}}(I_n - H)y/n = \mathrm{RSS}_p/n,$$

say, and hence verify that

$$\mathrm{AIC} \equiv n \log \widehat{\sigma}^2 + 2p.$$

(b) If $\widehat{\sigma}_0^2$ is the unbiased estimate $\mathrm{RSS}_q/(n-q)$ under some fixed correct model with q covariates, show that minimising AIC is equivalent to minimising $n\log\{1 + (\widehat{\sigma}^2 - \widehat{\sigma}_0^2)/\widehat{\sigma}_0^2\} + 2p$, and that this last expression is roughly equal to $n(\widehat{\sigma}^2/\widehat{\sigma}_0^2 - 1) + 2p$. Deduce that model selection using C_p approximates that using AIC.

Problem 3 ($\mathrm{AIC_c}$ for the linear model) Consider data generated by a true model g under which the responses $Y_j \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$, let $\mathrm{E}_g(\cdot)$ denote expectation with respect to this model, and suppose we choose a candidate model $f(y;\theta)$ to minimize the loss when predicting a new sample $Y_j^+ \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma^2)$ independent of the old one,

$$\mathrm{E}_g\left(\mathrm{E}_g^+\left[\sum_{j=1}^n 2\log\left\{\frac{g(Y_j^+)}{f(Y_j^+;\widehat{\theta})}\right\}\right]\right);$$

here E_g^+ denotes expectation over $Y_1^+,\dots,Y_n^+$, and $\widehat{\theta}$ is the maximum likelihood estimator of $\theta = (\mu_1,\dots,\mu_n,\sigma^2)$ based on $Y_1,\dots,Y_n$.

(a) Show that the sum in the expectation above may be written

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

and deduce that the inner expectation equals

$$\sum_{j=1}^n \left\{\log\widehat{\sigma}^2 + \frac{\sigma^2}{\widehat{\sigma}^2} + \frac{(\mu_j - \widehat{\mu}_j)^2}{\widehat{\sigma}^2} - \log\sigma^2 - 1\right\}.$$

(b) Suppose that a candidate linear model with full-rank $n \times p$ design matrix X is correct, that is, $\mu = X\beta$ for some $\beta_{p\times 1}$. Deduce that in this case $\sum(\mu_j - \widehat{\mu}_j)^2 = (\widehat{\mu} - \mu)^{\mathrm{T}}(\widehat{\mu} - \mu) \sim \sigma^2\chi_p^2$ independent of $n\widehat{\sigma}^2 \sim \sigma^2\chi_{n-p}^2$, and use the facts that for $\nu > 2$ the expected values of a χ_ν^2 variable and of its reciprocal are ν and $(\nu-2)^{-1}$ to show that the loss above is

$$n\mathrm{E}_g\left(\log\widehat{\sigma}^2\right) + \frac{n^2}{n-p-2} + \frac{np}{n-p-2} - n\log\sigma^2 - n,$$

or equivalently,

$$n\mathrm{E}_g\left(\log\widehat{\sigma}^2\right) + \frac{n(n+p)}{n-p-2}.$$

(c) Show that this loss is estimated unbiasedly by

$$\mathrm{AIC_c} = n\log\widehat{\sigma}^2 + n\frac{1+p/n}{1-(p+2)/n},$$

and that $\mathrm{AIC_c} \doteq n\log\widehat{\sigma}^2 + n + 2(p+1) + O(p^2/n)$, so for large n and fixed p minimising $\mathrm{AIC_c}$ will select the same model as minimising $\mathrm{AIC} = n\log\widehat{\sigma}^2 + 2p$, but that when p is comparable with n, $\mathrm{AIC_c}$ penalizes model dimension more severely.