STATISTICAL MODELLING Part IIC.

Example Sheet 2 (of 4) RN/Lent 2010

1. Consider the linear model $Y = X\beta + \epsilon$, where the $n \times p$ matrix X has full rank $p < n$ and $\epsilon \sim N_n(0, \sigma^2 I)$. Let A be an $n \times n$ matrix defined as follows: the top p rows consist of the matrix $(X^\top X)^{-1} X^\top$, while for $i = p+1, \ldots, n$, the i^{th} row $a_i^\top$ of A satisfies $a_i^\top X (X^\top X)^{-1} z = 0$ for all $z \in \mathbb{R}^p$. Such a choice of $n - p$ linearly independent vectors a_i is possible because $X(X^\top X)^{-1} z$ belongs to the p–dimensional subspace $U = \{Xb : b \in \mathbb{R}^p\}$.

   (a) Let $Z = AY$. What is the distribution of Z? Show that for $i = 1, \ldots, p$ we have that $Z_i = (AY)_i = \hat{\beta}_i$, so that $\hat{\beta}$ is a function of $(Z_1, \ldots, Z_p)^\top$, and, incidentally, $(AX\beta)_i = \beta_i$.

   (b) Show that for $i = p+1, \ldots, n$ and $j = 1, \ldots, p$,

   $$(AA^\top)_{ij} = 0$$

   and argue that this implies that $(Z_1, \ldots, Z_p)^\top$ and $(Z_{p+1}, \ldots, Z_n)^\top$ are independent.

   (c) Show that $\hat{\sigma}^2$ is a function of $(Z_{p+1}, \ldots, Z_n)^\top$.

   Conclude that $\hat{\beta}$ and $\hat{\sigma}^2$ are thus independent.

2. Let $Y = X\beta + \epsilon$, where X and β are partitioned as $X = (X_0\, X_1)$ and $\beta^T = (\beta_0^T\, \beta_1^T)$ respectively (where β_0 has p_0 components and β_1 has $p - p_0$ components).

   (a) Show that

   $$\|Y\|^2 = \|P_0 Y\|^2 + \|(P - P_0)Y\|^2 + \|Y - PY\|^2.$$

   (b) Recall that the likelihood ratio statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is a strictly increasing function of $\|(P - P_0)Y\|^2 / \|Y - PY\|^2$. Use Cochran's theorem to find the joint distribution of $\|(P - P_0)Y\|^2$ and $\|Y - PY\|^2$ under H_0. How would you perform the hypothesis test?

   [Hint: from the first example sheet $\text{rank}(P) = p$, *and* $\text{rank}(I - P) = n - p$. *Similar arguments give that* $\text{rank}(P_0) = p_0$. *For brownie points, argue that* $\text{rank}(P - P_0) = p - p_0$*.]*

3. (Continuation) An extension of Cochran's theorem says the following: suppose that $Y \sim N_n(\mu, \sigma^2 I)$ and that we may write $\|Y\|^2 = Y^T A_1 Y + \ldots + Y^T A_k Y$, where $Y^T A_i Y$ is a positive semi-definite quadratic form of rank r_i. If

$r_1 + \ldots + r_k = n$, then $Y^TA_1Y, \ldots, Y^TA_kY$ are independent with $Y^TA_iY \sim \sigma^2\chi^2_{r_i}\left(\frac{1}{\sigma^2}\mu^TA_i\mu\right)$ for $i = 1, \ldots, k$. Deduce the joint distribution of $\|(P - P_0)Y\|^2$ and $\|Y - PY\|^2$ under H_1.

4. Consider the balanced two-way ANOVA model

$$H_3 : Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$$

with an appropriate identifiability constraint. Consider also the submodels H_0 : $\alpha_i = \beta_j = 0$ for all i, j; $H_1 : \alpha_i = 0$ for all i; $H_2 : \beta_j = 0$ for all j. Show directly that

$$RSS(H_1) - RSS(H_3) = RSS(H_0) - RSS(H_2).$$

5. One of the data sets in the *Modern Applied Statistics in S-Plus* (MASS) library is `hills`. You can find out about the data with

```
> library(MASS)
> ?hills
> attach(hills)
> pairs(hills)
```

The data contain one known error in the winning time. Identify this error (think carefully!) and subtract an hour from the winning time. Can you see any reason why we might want to consider taking logarithms of the variables? Explain why we should include an intercept term if we do choose to take logarithms.

Explore at least two linear models for this data, and give estimates with standard errors for your preferred model. Predict the record time for a hypothetical 5.3 mile race with a 1100ft climb, giving a 95% prediction interval.

6. Look at the `cabbages` data in the `library(MASS)` package. Does the planting date have a significant effect on the weight of the cabbage head?

7. (a) Let A be a $p \times p$ non-singular matrix and b be a $p \times 1$ column vector. Show that

$$(A - bb^T)^{-1} = A^{-1} + \frac{A^{-1}bb^TA^{-1}}{1 - b^TA^{-1}b}.$$

(b) Consider the linear model $Y = X\beta + \epsilon$, where X is a known $n \times p$ matrix, $\epsilon \sim N_n(0, \sigma^2I)$, and let x_i^T denote the ith row of X. Further, let $X_{(-i)}$ denote the $(n-1) \times p$ matrix obtained by deleting the ith row of X, and suppose that this matrix is of full rank p. Prove that

$$X_{(-i)}^TX_{(-i)} = X^TX - x_ix_i^T.$$

Let $\hat{\beta} = (X^T X)^{-1} X^T Y$ denote the MLE in the model with all n observations, and let $\hat{\beta}_{(-i)} = (X_{(-i)}^T X_{(-i)})^{-1} X_{(-i)}^T Y_{(-i)}$ denote the MLE in the model with the ith observation deleted. Show that

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{1}{1 - p_i} (X^T X)^{-1} x_i (Y_i - x_i^T \hat{\beta}),$$

where $p_i = x_i^T (X^T X)^{-1} x_i$ is the leverage of the ith observation in the model.

8. (Continuation) Cook's distance of the observation (x_i, Y_i) is defined as

$$D_i = \frac{1}{p\tilde{\sigma}^2} (\hat{\beta}_{(-i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(-i)} - \hat{\beta}), \quad i = 1, \ldots, n,$$

where $\tilde{\sigma}^2 = \|Y - X\hat{\beta}\|^2/(n - p)$, and it may appear that we need to fit $n + 1$ linear models in order to calculate all of the Cook's distances. Deduce, however, that

$$D_i = \frac{1}{p} \Big(\frac{p_i}{1 - p_i} \Big) \hat{\eta}_i^2,$$

where $\hat{\eta}_i = (Y_i - x_i^T \hat{\beta})/(\tilde{\sigma}\sqrt{1 - p_i})$ is the ith studentised fitted residual.

9. In the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, show that β_0, β_1 and β_2 are mutually orthogonal if and only if

$$\sum_{i=1}^n x_{i1} = \sum_{i=1}^n x_{i2} = \sum_{i=1}^n x_{i1} x_{i2} = 0.$$