

In all the questions that follow,  $X$  is an  $n$  by  $p$  design matrix with full column rank and  $P$  is the orthogonal projection onto the column space of  $X$ . We will assume that  $n - p \geq 2$ . The vector  $Y \in \mathbb{R}^n$  will be a vector of responses and we will define  $\hat{\beta} := (X^T X)^{-1} X^T Y$ ,  $\tilde{\sigma}^2 := \|(I - P)Y\|^2 / (n - p)$ ,  $\hat{\varepsilon} := Y - X\hat{\beta}$  and  $\hat{Y} := PY$ . Let  $p_i = P_{ii}$  be the leverage for the  $i^{\text{th}}$  observation. The average of  $a \in \mathbb{R}^n$  is denoted by  $\bar{a} = n^{-1} a^T \mathbf{1}_n = n^{-1} \sum_{i=1}^n a_i$ .

1. Consider a linear model  $Y = X\beta + \varepsilon$ . Now suppose we reparametrise by letting  $\theta = A\beta$  where  $A \in \mathbb{R}^{p \times p}$  is invertible, so now we have  $Y = XA^{-1}\theta + \varepsilon$  (with  $XA^{-1}$  the new design matrix). Show that the fitted values and predictions based on applying OLS in the reparametrised model will be identical to those in the original model.
2. An  $n$ -vector is called constant, if all its entries are the same. The *sample correlation* of two non-constant  $n$ -vectors  $a, b$  is defined as

$$\text{corr}(a, b) = \frac{(a - \bar{a}\mathbf{1}_n)^T (b - \bar{b}\mathbf{1}_n)}{\|a - \bar{a}\mathbf{1}_n\| \|b - \bar{b}\mathbf{1}_n\|}.$$

If either  $a$  or  $b$  is constant, set  $\text{corr}(a, b) = 0$ .

Consider now a linear model  $Y = X\beta + \varepsilon$ , where the first column of  $X$  is a vector of 1's. Suppose  $Y$  and the fitted values  $\hat{Y}$  are non-constant.

- (a) Show that the coefficient of determination  $R^2$  satisfies

$$R^2 = \frac{\|\hat{Y} - \bar{Y}\mathbf{1}_n\|^2}{\|Y - \bar{Y}\mathbf{1}_n\|^2}.$$

- (b) Show that  $\text{corr}(Y, \hat{Y}) = \sqrt{R^2}$ .
- (c) Show that  $\text{corr}(Y, \hat{Y}) = \sup_{a \in \text{col}(X)} \text{corr}(Y, a)$ , where  $\text{col}(X)$  is the column space of  $X$ .

3. Show that the AIC in a normal linear model is

$$n\{1 + \log(2\pi\hat{\sigma}^2)\} + 2(p + 1).$$

4. Return to the brain sizes data studied in practical 3.

```
> file_path <- "http://www.statslab.cam.ac.uk/~ra591/data/"
> BrainSize <- read.csv(paste0(file_path, "BrainSize.csv"))
> attach(BrainSize)
> BrainSizeLM2 <- lm(PIQ ~ MRI_Count + Height)
```

- (a) We want to plot a confidence ellipse for the coefficients for brain size and height. To do this, first install the `ellipse` package using

```
> install.packages("ellipse")
```

and select a mirror of your choice. Next, load the package with `library(ellipse)`. Look at `?ellipse.lm` and plot a 95% confidence ellipse for the coefficients with

```
> plot(ellipse(BrainSizeLM2, c(2, 3)), type = "l")
```

Using `abline`, add to the plot the end points of 95% confidence intervals for each of the coefficients in red (e.g., using `confint`), and also add in blue the sides of the confidence rectangle in question 8 of Example sheet 1. If you are using `RStudio`, you can output a pdf of your plot by clicking on “Export” above the plot window.

- (b) Look at the correlation between the estimates of the coefficients using

```
> summary(BrainSizeLM2, correlation = TRUE)$correlation
```

which is an estimate of  $\text{Corr}(\hat{\beta}, \hat{\beta})$ . This can also be obtained in R from

```
> cov2cor(vcov(BrainSizeLM2))
```

Compare this to the sample correlation between the corresponding variables

```
> cor(Height, MRI_Count)
```

What do you notice? Explain.

5. Let  $f$  and  $g$  be two densities on  $\mathbb{R}$  with  $S := \{x : g(x) > 0\} = \{x : f(x) > 0\}$ . Show that the Kullback–Leibler divergence,

$$K(g, f) := \int_S [\log\{g(x)\} - \log\{f(x)\}]g(x)dx,$$

is non-negative. *Hint: Use Jensen’s inequality.*

6. Consider forward selection in the linear model  $Y = \beta_0 1_n + X\beta + \varepsilon$ , where  $1_n$  is an  $n$ -vector of 1’s. At the 0<sup>th</sup> stage, only the intercept term is in the model. Now suppose that the design matrix for the model fitted in the  $k^{\text{th}}$  stage for  $k < p$  is  $X^{(k)} := (1_n \ X_{j_1} \ \cdots \ X_{j_k})$ , where  $X_j$  denotes the  $j^{\text{th}}$  column of  $X$ . Show that the next variable to enter the model, leading to largest reduction in the residual sum of squares, is  $X_{j^*}$  where

$$j^* = \operatorname{argmax}_{j \neq j_1, \dots, j_k} \frac{|(X_j^\perp)^T Y|}{\|X_j^\perp\|}.$$

Here,  $X_j^\perp$  denotes the orthogonal projection of  $X_j$  onto the orthogonal complement of the column space of  $X^{(k)}$ .

7. Assume  $X$  has full column rank.

- (a) Show that  $p_i = x_i^T (X^T X)^{-1} x_i$ , where  $x_i$  is the  $i^{\text{th}}$  row of  $X$  (regarded as a column vector). Deduce that  $p_i > 0$ .
- (b) Suppose the design matrix  $X$  consists of just a single variable and a column of 1’s representing an intercept term (as the first column). Show that

$$p_i = \frac{1}{n} + \frac{(X_{i2} - \bar{X}_2)^2}{\sum_{k=1}^n (X_{k2} - \bar{X}_2)^2},$$

where  $\bar{X}_2 := \frac{1}{n} \sum_{k=1}^n X_{k2}$ . *Hint: Why can we assume that the  $i^{\text{th}}$  component of the second column is  $X_{i2} - \bar{X}_2$  rather than  $X_{i2}$ ?*

8. 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
> pairs(hills)
```

- (a) The data contain one known error in the winning time. Identify this error (think carefully!) and subtract an hour from the winning time. *Hint: You can examine the plots and identify observations for which the response and covariates satisfy certain inequalities e.g.*

```
> hills[(hills$time > 50) & (hills$dist < 10), ]
```

- (b) Can you see any reason why we might want to consider taking logarithms of the response and the predictor variables? Explain why we should include an intercept term if we do choose to take logarithms.
- (c) 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.
9. (a) Let  $A$  be a  $p \times p$  non-singular matrix and let  $b \in \mathbb{R}^p$ . Prove that if  $b^T A^{-1} b \neq 1$ , then  $A - bb^T$  is invertible with inverse given by

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

- (b) Consider a linear model  $Y = X\beta + \varepsilon$  with  $\text{Var}(\varepsilon) = \sigma^2 I$ , and let  $x_i^T$  denote the  $i^{\text{th}}$  row of  $X$ . Further, let  $X_{(-i)}$  denote the  $(n-1) \times p$  matrix obtained by deleting the  $i^{\text{th}}$  row of  $X$ , and write  $\hat{\beta}_{(-i)}$  for the OLS estimate of  $\beta$  when the  $i^{\text{th}}$  observation has been removed. Suppose that  $X_{(-i)}$  has full column rank and that the leverage score  $p_i$  is less than 1. By noting that

$$X^T X = \sum_{i=1}^n x_i x_i^T,$$

prove that  $\text{Var}(\hat{\beta}_{(-i)}) - \text{Var}(\hat{\beta})$  is positive semi-definite.

- (c) Show that

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

and hence deduce that the Cook's distance  $D_i$  of the observation  $(Y_i, x_i)$  satisfies

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

where  $\hat{\eta}_i = (Y_i - x_i^T \hat{\beta}) / (\tilde{\sigma} \sqrt{1 - p_i})$  is the  $i^{\text{th}}$  standardised residual. *Hint: Use part (a) and question 7(a).*

10. Show that

$$p_i + \frac{\hat{\varepsilon}_i^2}{\|(I - P)Y\|^2} \leq 1,$$

so if  $p_i$  is close to 1, the  $i^{\text{th}}$  residual is forced to be close to 0. *Hint: First write out an expression for  $\hat{\varepsilon}_i$  involving  $I - P$ . Then make use of the fact that  $I - P$  is an orthogonal projection and use the Cauchy-Schwarz inequality to get something that, after simplification, gives the above.*