STATISTICAL MODELLING Part IIC
Example Sheet 2 (of 4) RDS/Lent 2015

In all the questions that follow, X is an n by p design matrix with full column rank and P is the orthogonal projection on to 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 $\hat{\beta} := (X^T X)^{-1} X^T Y$, $\tilde{\sigma}^2 := \|(I - P)Y\|^2/(n - p)$ and $\hat{\varepsilon} := Y - X\hat{\beta}$.

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. Show that the AIC in a normal linear model is
$$n\{1 + \log(2\pi\hat{\sigma}^2)\} + 2(p + 1).$$

3. 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.*

4. 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 the leverage, p_i, of the i^{th} observation satisfies
$$p_i = \frac{1}{n} + \frac{(X_{i2} - \bar{X}_2)^2}{\sum_{k=1}^n (X_{k2} - \bar{X}_2)^2},$$
where $\bar{X}_2 := \frac{1}{n}\sum_{k=1}^n X_{k2}$. *Hint: Why can we assume that the i^{th} component of the second column is $X_{i2} - \bar{X}_2$ rather than X_{i2}?*

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

```
> file_path <-
+ "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> BrainSize <- read.csv(paste0(file_path, "BrainSize.csv"))
> attach(BrainSize)
> BrainSizeLM2 <- lm(PIQ ~ MRI_Count + Height)
```

In this question we will 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, 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. Now look at the correlation between the estimates of these coefficients using

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

and compare this to the correlation between the corresponding variables

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

What do you notice? Explain.

6. 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)
```

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), ]
```

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.

7. Assume X has full column rank. Show that the leverage for the i^{th} observation, p_i is $x_i^T(X^TX)^{-1}x_i$. Deduce the $p_i > 0$.

8. (a) Let A be a $p \times p$ non-singular matrix and let $b \in \mathbb{R}^p$. Prove that if $b^TA^{-1}b \neq 1$, then $A - bb^T$ is invertible with inverse given by

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

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

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

prove that writing $\hat{\beta}_{(-i)}$ for the OLS estimate of β when the i^{th} observation has been removed, the difference

$$\mathrm{Var}(\hat{\beta}_{(-i)}) - \mathrm{Var}(\hat{\beta})$$

is positive semi-definite. Here $\hat{\beta}$ is the usual OLS estimate of β based on all n observations.

(c) Show that

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

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

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

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

9. (a) (Continuation) The *externally studentised residual* of the i^{th} observation may be defined as

$$\tilde\eta_i := \frac{\hat\varepsilon_i}{\tilde\sigma_{(-i)}\sqrt{1-p_i}},$$

where $\tilde\sigma_{(-i)}$ is the equivalent of $\tilde\sigma$ but calculated omitting the i^{th} observation, so

$$\tilde\sigma^2_{(-i)} = \frac{1}{n-p-1}\|Y_{(-i)} - X_{(-i)}\hat\beta_{(-i)}\|^2,$$

where $Y_{(-i)}$ is the response Y without the i^{th} component. Show that $\tilde\eta_i \sim t_{n-p-1}$. *Hint: It may help to first show that*

$$\hat\varepsilon_i = (1-p_i)(Y_i - x_i^T\hat\beta_{(-i)})$$

using (1). How can we construct a hypothesis test based on $\tilde\eta_i$ to test whether the i^{th} observation is an outlier?

(b) Another dataset in the MASS package is `mammals` which gives the body and brain masses of 62 mammals. Log transform both variables and then fit a linear model with `log(brain)` as the response. Then apply your hypothesis test to check whether the observation corresponding to humans is an outlier. The function `rstudent` that calculates externally studentised residuals may be of help. What is the p-value you obtain? (You can also discuss whether a one- or two-sided test is most appropriate here).

(c) Now download the Cambridge Colleges data with

```
> file_path <-
+ "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> Colleges <- read.csv(paste0(file_path, "Colleges.csv"))
```

(note you may have already inputted the file path in an earlier question). Fit a linear model with the percentage of firsts as the response and the logarithm of the wine budget as a covariate. Pick a college (possibly your own) and test whether it is an outlier. Looking at a plot of the data, what appears to be the most outlying college? Note you can add the names of the colleges to the plot by inputting

```
text(log($WineBudget), PercFirsts, rownames(Colleges), cex=0.6, pos=3)
```

after plotting the data (provided the data frame `Colleges` is attached). What is the issue with using your test to now determine whether this college is an outlier?

10. Show that

$$p_i + \frac{\hat\varepsilon_i^2}{\|(I-P)Y\|^2} \le 1,$$

so if p_i is close to 1, the i^{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.*

11*. 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^{th} stage, only the intercept term is in the model. Now suppose that the design matrix for the model fitted in the k^{th} stage for $k < p$ is $X^{(k)} := (1_n \; X_{j_1} \; \cdots \; X_{j_k})$, where X_j denotes the j^{th} column of X. Show that the next variable to enter the model is X_{j^*} where

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

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