STATISTICAL MODELLING Part IIC / Michaelmas 2022
Example Sheet 2 (of 4)

In all the questions that follow, X is an n by p design matrix with full column rank and H 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 $\hat{\beta} := (X^T X)^{-1} X^T Y$, $\hat{\sigma}^2 := \|(I - H)Y\|^2/(n - p)$ 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 solution to the ridge regression problem

$$\hat{\beta}_\lambda = \arg\min_\beta \|Y - X\beta\|^2 + \lambda\|\beta\|^2$$

is given by $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T Y$. Try to prove this result using two methods: (1) Matrix derivatives; (2) Treating ridge regression as solving the least squares problem for the following expanded design matrix and responses:

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I_p \end{pmatrix}, \tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}.$$

3. Suppose $Y \sim \mathrm{N}(\mu, \sigma^2 I)$ where $\mu \in \mathbb{R}^n$ is unknown. Consider any linear estimator of the form $\hat{\mu} = MY$ where $M = M(X) \in \mathbb{R}^{n\times n}$ can depend on X. Show that

$$C_M = \|Y - \hat{\mu}\|^2 + 2\sigma^2 \cdot \mathrm{trace}(M)$$

is, condition on X, an unbiased estimator of the mean squared prediction error

$$\mathbb{E}(\|Y^* - \hat{\mu}\|^2 \mid X)$$

where $Y^* = \mu + \epsilon^*$, $\epsilon^* \sim \mathrm{N}(0, \sigma^2 I)$ is independent of Y. Explain why this reduces to Mallow's C_p when $\hat{\mu} = X\hat{\beta}$.

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

$$n\{1 + \log(2\pi\hat{\sigma}^2_{\mathrm{MLE}})\} + 2(p + 1),$$

where $\hat{\sigma}^2_{\mathrm{MLE}} = \|(I - H)Y\|^2/n$ is the maximum likelihood estimator of σ^2. When the noise variance σ^2 is known, re-derive the AIC and show that it is equivalent to Mallows' C_p.

5. 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, H_{ii}, of the i^{th} observation satisfies

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

where $\bar{X}_2 := \frac{1}{n}\sum_{k=1}^n X_{k2}$. Describe what kind of observations may have a large leverage. *Hint: Why can we assume that the i^{th} component of the second column is $X_{i2} - \bar{X}_2$ rather than X_{i2}?*

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

```
> path <- "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> BrainSize <- read.csv(file.path(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 (if prompted). 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 6 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.

7. 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.*

```
> subset(hills, time > 50 & dist < 20)
```

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 the transformed 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, give a 95% prediction interval for both models and explain how and why they differ.

8. Let Y be a random variable with density $f(y;\theta)$ for $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ and some $\theta \in \Theta \subseteq \mathbb{R}^d$, and write $\ell(\theta;Y)$ and $U(\theta;Y)$ for the corresponding log-likelihood and score functions. Assume that the order of differentiation with respect to a component of θ and integration over $\mathcal{Y}$ may be interchanged where necessary. Show that, for $r, s = 1, \ldots, d$,

$$\mathrm{Cov}_\theta\{U_r(\theta;Y), U_s(\theta;Y)\} = -\mathbb{E}_\theta\Big\{\frac{\partial^2}{\partial\theta_r\partial\theta_s}\ell(\theta;Y)\Big\}.$$

9. In the normal linear model, find the Fisher information matrix for the parameters (β, σ^2). Assume $X^TX/n \to \Sigma$. Use the asymptotic theory of the maximum likelihood estimator to obtain the asymptotic covariance matrix of $(\hat\beta, \hat\sigma^2)$ and show that it is indeed the limit of the exact covariance matrix of $(\hat\beta, \hat\sigma^2)$.

10. (a) Let A be a $p \times p$ non-singular matrix and let $b \in \mathbb{R}^p$. Prove that if $v^TA^{-1}u \neq -1$, then $A + uv^T$ is invertible with inverse given by the Sherman-Morrison formula

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}.$$

(b) Consider a linear model $Y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon \mid X) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. 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, H_{ii}, is less than 1. Let $\hat\beta_{(-i)}$ be the OLS estimator of β when the i^{th} observation has been removed. Prove that the difference

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

is positive semi-definite. *Hint: Use* $X^TX = \sum_{i=1}^n x_ix_i^T$ *and* $H_{ii} = X_i^T(X^TX)^{-1}X_i$.

(c) Show that

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

and use this to deduce the identity

$$\hat\mu_i = H_{ii}Y_i + (1 - H_{ii})\hat\mu_{(-i)}, \tag{2}$$

where $\hat\mu_i = X_i^T\hat\beta$ and $\hat\mu_{(-i)} = X_i^T\hat\beta_{(-i)}$.

(d) Show that Cook's distance D_i of the observation (Y_i, X_i) can be expressed as

$$D_i := \frac{\|X(\hat\beta - \hat\beta_{(-i)})\|^2}{p\hat\sigma^2} = \frac{1}{p}\Big(\frac{H_{ii}}{1 - H_{ii}}\Big)\hat\eta_i^2,$$

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

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

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

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

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

where $Y_{(-i)}$ is the response Y without the i^{th} component. Show that $\tilde{\eta}_i \sim t_{n-p-1}$ in the normal linear model, that is, if $\epsilon \mid X \sim N(0, \sigma^2 I)$. *Hint: Use* (2). 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 68 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 t-test is most appropriate here).