STATISTICAL MODELLING Part IIC / Michaelmas 2025
Example Sheet 2 (of 4)

In questions that follow, by normal linear model we mean $Y = X\beta + \varepsilon$, $\varepsilon \mid X \sim N(0, \sigma^2 I)$, where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$. Let $\hat{\beta} = (X^T X)^{-1} X^T Y$ be the ordinary least squares estimator of β and $P = X(X^T X)^{-1} X^T$ be the orthogonal projection matrix on to the column space of X.

1. Suppose $Y \mid X \sim \mathrm{N}(X\beta, \sigma^2\Sigma)$ where $\Sigma = \mathrm{diag}(w_1^{-1}, \ldots, w_n^{-1})$ and $w_1, \ldots, w_n$ are known. Show that the maximum likelihood estimator of β solves the following weighted least squares problem:
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} w_i (Y_i - X_i^T \beta)^2.$$
Describe how you would modify your `mylm` function from Example Sheet 1 to solve this problem by transforming the input `X` and `Y`.

2. Show that the AIC in a normal linear model (up to additive constants) is
$$n\{1 + \log(2\pi\hat{\sigma}^2_{\mathrm{MLE}})\} + 2(p+1),$$
where $\hat{\sigma}^2_{\mathrm{MLE}} = \|(I-P)Y\|^2/n$ 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.

3. Suppose $Y, Y^* \sim \mathrm{N}(\mu, \sigma^2 I)$ are i.i.d., where $\mu \in \mathbb{R}^n$ is a unknown non-random vector. Let $X \in \mathbb{R}^{n\times p}$ be fixed. Consider the following definition of mean-squared prediction error:
$$\mathrm{MSPE} = \frac{1}{n}\mathsf{E}(\|Y^* - \hat{\mu}\|^2).$$
   (a) Let $\hat{\mu} = X\hat{\beta}$. Show that
$$\mathrm{MSPE} = \sigma^2 + \frac{1}{n}\|(I-P)\mu\|^2 + \frac{\sigma^2 p}{n}.$$
   Compare this to the *bias-variance tradeoff* in the lectures and identify the "bias^2" and "variance" terms.

   (b) Consider any linear estimator of the form $\hat{\mu} = MY$ where $M \in \mathbb{R}^{n\times n}$ can depend on X. Show that
$$C_M = \|Y - \hat{\mu}\|^2 + 2\sigma^2 \cdot \mathrm{trace}(M)$$
   is an unbiased estimator of $n \cdot \mathrm{MSPE}$. Show that this reduces to Mallows' C_p when $M = P$.

4. Consider the normal linear model with fixed $X \in \mathbb{R}^{n\times p}$ and some $1 \le i \le n$. Let X_i^T denote the i^{th} row of X and $X_{(-i)}$ denote the $(n-1)\times p$ matrix obtained by deleting the i^{th} row. Suppose $P_{ii} < 1$ and $X_{(-i)}$ has full column rank. Let $\hat{\beta}_{(-i)}$ be the OLS estimator of β when the i-th observation has been removed.

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

(b) Prove that the difference

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

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

(c) Show that

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{1}{1-P_{ii}}(X^TX)^{-1}X_i(Y_i - X_i^T\hat{\beta}).$$

Use this to deduce the identity

$$\hat{\mu}_i = P_{ii}Y_i + (1-P_{ii})\hat{\mu}_{(-i)},$$

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

$$D_i := \frac{\|X(\hat{\beta} - \hat{\beta}_{(-i)})\|^2}{p\hat{\sigma}^2} = \frac{1}{p}\Big(\frac{P_{ii}}{1-P_{ii}}\Big)\tilde{R}_i^2,$$

where

$$\tilde{R}_i = (Y_i - X_i^T\hat{\beta})/(\hat{\sigma}\sqrt{1-P_{ii}})$$

is the i^{th} studentised residual.

5. Return to the house prices data studied in practical 3.

```
> file_path <- "https://raw.githubusercontent.com/AJCoca/SM19/master/"
> HousePrices <- read.csv(paste0(file_path, "HousePrices.csv"))
> HousePricesLM2 <- lm(Sale.price ~ Living.area + Property.tax, data = HousePrices)
```

In this question we will plot a confidence ellipse for the coefficients for living area and property tax. 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(HousePricesLM2, c(4, 7)), type = "l")
```

Use `confint` and `abline` to 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 2 of Example sheet 1. Save your output by using the `pdf` command (if you are using `Rstudio`, you can also click on "Export" above the plot window). Now look at the correlation between the estimates of these coefficients using

```
> summary(HousePricesLM2, correlation = TRUE)$correlation
```

and compare this to the correlation between the corresponding variables

```
> cor(HousePrices$Living.area, HousePrices$Property.tax)
```

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.*

```
> 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.

7. Consider the setup in question 4.

   (a) Discuss how you can use the studentised residual $\tilde{R}_i$ to test whether the i-th observation is an outlier by replacing $\hat{\sigma}$ in the definition of $\tilde{R}_i$ with another estimator of σ so that $\tilde{R}_i \sim t_{n-p-1}$. *Hint: Recall that* $\mathsf{Var}(Y_i - X_i^T\hat{\beta}) = \sigma^2(1 - P_{ii})$ *from Example Sheet 1.*

   (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).

8. Consider a random vector $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$.

   (a) Assuming that all expressions below are well defined and you can interchange derivative and expectation, show that the solution to the population least squares problem is given by

$$\arg\min_{\beta \in \mathbb{R}^p} \mathsf{E}\{(Y - \beta^T X)^2\} = \{\mathsf{E}(XX^T)\}^{-1}\mathsf{E}(XY).$$

   (b) Consider the partition $X = (X_0, X_1) \in \mathbb{R}^{p_0} \times \mathbb{R}^{p-p_0}$, and suppose $\mathsf{E}(Y \mid X) = \beta_0^T X_0 + \beta_1^T X_1$ for some $\beta_0 \in \mathbb{R}^{p_0}$ and $\beta_1 \in \mathbb{R}^{p-p_0}$. Let

$$\tilde{\beta}_0 = \arg\min_{\beta_0 \in \mathbb{R}^p} \mathsf{E}\{(Y - \beta_0^T X_0)^2\}.$$

   Show that $\beta_0 = \tilde{\beta}_0$ is generally true only if $\mathsf{E}(X_0 X_1^T) = 0$.

9. Let $Z \in \mathbb{R}^{n \times q}$ be the data matrix for a collection of instrumental variables. Let $P_Z = Z(Z^TZ)^{-1}Z^T$ be the projection matrix onto the column space of Z. The two-stage least squares estimator is defined as

$$\hat{\beta}_{\text{TSLS}} = (\hat{X}^T\hat{X})^{-1}\hat{X}^TY,$$

where $\hat{X} = P_Z X$ is the fitted-value of the first-stage regression of X on Z.

(a) This definition assumes that $\hat{X}^T\hat{X}$ is invertible. Show that this implies $q \geq p$.

(b) Now show that

$$\hat{\beta}_{\text{TSLS}} = (X^T P_Z X)^{-1}(X^T P_Z Y).$$

How is this different from the OLS estimator for regressing Y on X? And from the estimated coefficients of X from the OLS estimator for regressing Y on X and Z?

(c) Modify your `mylm` function from Example Sheet 1 so it can also take `Z` ($n \times q$ matrix) as input and return $\hat{\beta}_{\text{TSLS}}$. Design a simulation example in R that demonstrates the TSLS estimator is consistent (converges to the true β) but the (single-stage) OLS estimators in part (b) are not.

(d) Do you thnk you can still use `mylm` to calculate or approximate the standard errors of $\hat{\beta}_{\text{TSLS}}$? Check your answer by designing a simulation study in R. *Hint: Try calculating the variance of $\hat{\beta}_{TSLS}$ by pretending $Z^T X$ is a constant.*

10. (a) 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 $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T Y$.

(b) Suppose you have access to a numerical solver of the so-called "lasso" problem

$$\text{minimize} \quad \|Y - X\beta\|^2 + \lambda\|\beta\|_1$$

for any $\lambda \geq 0$. Describe how you can use it to solve the so-called "elastic net" problem

$$\text{minimize} \quad \|Y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|^2$$

for any $\lambda_1, \lambda_2 \geq 0$. *Hint: Augment X and Y with some additional rows.*

11. Consider the linear model with heteroscedastic noise: $(X_i, Y_i) \in \mathbb{R} \times \mathbb{R}, i = 1, \ldots, n$ are i.i.d. and satisfy

$$Y_i \mid X_i \sim \text{N}(\alpha + \beta X_i, \sigma^2(X_i)).$$

We saw in lectures that the OLS estimator of (α, β) is still consistent and asymptotically normal. Suppose we know

$$\sigma^2(X_i) = \tau^2(1 + \eta X_i^2)$$

for some unknown $\tau^2, \eta > 0$. Could you find an estimator of (α, β) that is more efficient (is consistent but has smaller variance asymptotically) than the OLS estimator? Compare the asymptotic variances of your estimator and the OLS estimator in a simulation study using R. *Hint: Find an estimator of (τ^2, η) using the OLS estimator of β, then solve a weighted least squares problem.*