Applied Statistics **Problem Sheet 1** **MT 2018**

1. (a) In the following linear model the response variable Y depends on two explanatory variables x and z, and $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Suppose $\sum_i x_i = \sum_i z_i = \sum_i x_i z_i = 0$ and let Y_* be a future observation that we intend to make, where the explanatory variables have values x_* and z_*. If E_* is the expectation of Y_*, write down the MLE of E_* and calculate its variance.

(b) In an experiment to determine the radioactivity of a particular isotope, radioactivity observations $Y_1, Y_2, \ldots, Y_n$, each normally distributed with variance σ^2, are taken at times $t_l, t_2, \ldots, t_n$. The actual radioactivity at time t_i is given by $\beta_0 + \beta_1(e^{-t_i} - c)$, where $c = \frac{1}{n}\sum_{i=1}^n e^{-t_i}$.

(i) Explain how it is possible to formulate a linear model relating the observations to the actual radioactivity, stating any additional assumptions you make.

(ii) Consider the problem of estimating the expected radioactivity E_t at a new time t. Show that the variance of the maximum likelihood estimator $\widehat{E_t}$ of E_t is given by

$$\mathrm{Var}(\widehat{E_t}) = \sigma^2 \left[\frac{1}{n} + \frac{(e^{-t} - c)^2}{\sum_{i=1}^n (e^{-t_i} - c)^2} \right].$$

2. Consider the linear model $Y = X\beta + \epsilon$ where Y is a random n-vector, X is an $n \times p$ design matrix of rank p (where $p < n$), β is a p-vector of parameters, and ϵ is an n-vector of independent random variables with mean zero and variance σ^2. Derive the least squares estimator $\widehat{\beta}$ for β and determine its variance matrix. Is $\widehat{\beta}$ unbiased?

Let $\widetilde{\beta} = DY$ be a second linear estimator of β, where D is a $p \times n$ matrix. Derive the expectation and variance matrix of $\widetilde{\beta}$.

Now suppose $\widetilde{\beta}$ is an unbiased estimator of β.

(a) Prove that $DX = I_p$, where I_p is the $p \times p$ identity matrix.

(b) Defining the matrix $D^* = D - (X^T X)^{-1} X^T$, show that the variance of $\widetilde{\beta}$ is given by

$$\mathrm{Var}(\widetilde{\beta}) = \sigma^2 \left(D^* D^{*T} + \left(X^T X \right)^{-1} \right).$$

Deduce that for any $i = 1, 2, \ldots, p$, we have $\mathrm{Var}(\widetilde{\beta}_\mathrm{i}) \geqslant \mathrm{Var}(\widehat{\beta}_\mathrm{i})$. Which estimator would you prefer to use to estimate β and why?

3. Consider the linear model $y = X\beta + \epsilon$, with y an $n \times 1$ vector, X an $n \times p$ matrix of rank p, β a $p \times 1$ vector and ϵ an $n \times 1$ multivariate normal random vector $\epsilon \sim N(0, \sigma^2 I_n)$.

Suppose we know the error variance σ^2.

Let $k \in \{1, \ldots, p-1\}$, and let $\widetilde{X} = [X_1, \ldots, X_{p-k}]$ and $\beta^{(0)} = (\beta_1, \ldots, \beta_{p-k})$. What is the maximized log-likelihood $\ell(\widehat{\beta}, \sigma^2; y)$? What is the Likelihood Ratio Test (LRT) statistic for the comparison of the model $H_0 : y = \widetilde{X}\beta^{(0)} + \epsilon$ with $H_1 : y = X\beta + \epsilon$, and what is its distribution?

4. (*TT09 BS1a exam Q1*) The data set `cigarettes` contains measurements of the carbon-monoxide (variable name *CO*), tar and nicotine content and tobacco weight for $n = 25$ cigarettes. The data are plotted below.

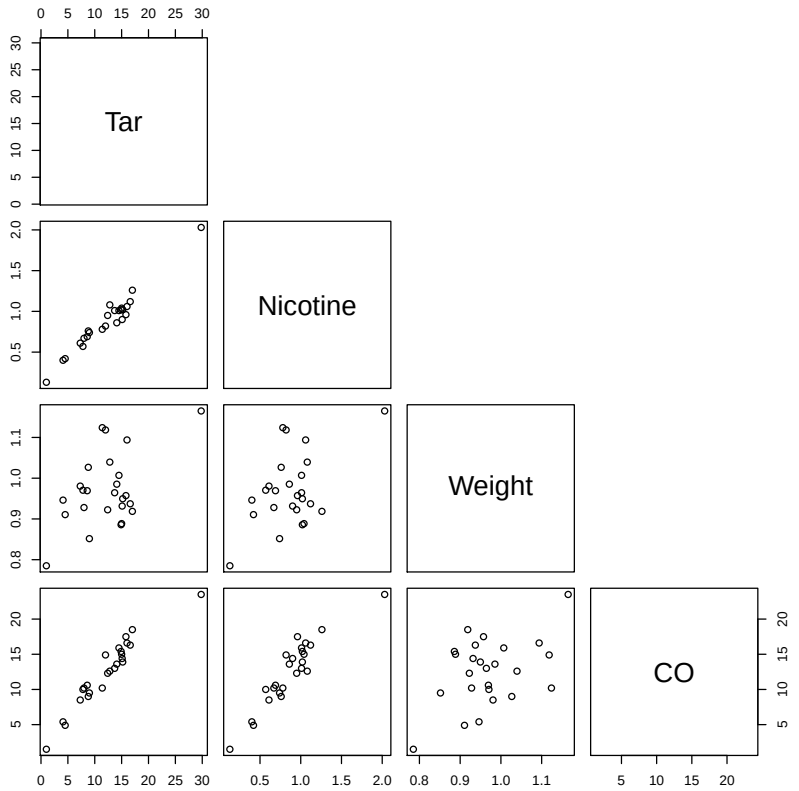

Figure 1: Cigarette data for Q4 and Q5.

In the normal linear model $CO \sim 1 + Nicotine + Tar + Weight$ the response is CO and all the other variables (including an intercept) are explanatory. The following partial R-output gives a standard-format ANOVA table for this model.

```
Response: CO
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
Nicotine   1   462.26   462.26   -B-     1.27e-12
Tar        1    33.00    33.00  15.7883 0.0006923
Weight     1 0.002357 0.002357   -C-         -D-
Residuals -A-   43.89     2.09
```

Give the values of the missing entries `-A-`, `-B-`, `-C-` and `-D-`. What is the residual sum of squares for the model $CO \sim 1$ (i.e. the model with just an intercept)? The residual sum of squares for the model $CO \sim 1 + Tar$ (i.e. an intercept and Tar) is 44.87. Carry out model selection (i.e. given the information available, carry out appropriate F-tests and say which of the models considered you would select as your preferred model).

5. (*Multicollinearity*) Let $y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I_n)$ be a normal linear model with $n \times p$ design matrix X of rank p, let $\widehat{\beta}$ be the MLE for the p parameters β, and let $H = X(X^T X)^{-1} X^T$.

Denote by $\widetilde{X} = X_{1:(p-1)}$ the matrix of the first $p-1$ columns of X.

Consider also a second model $X_p = \widetilde{X}\gamma + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_n)$ and γ is a $(p-1)$-dimensional parameter. In this model the 'response' is X_p, the last column of X, and the explanatory variables are the other $p-1$ columns of X. Let $H_{-p} = \widetilde{X}(\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T$.

(a) Show that

$$\mathrm{var}(\widehat{\beta}_p) = \frac{\sigma^2}{X_p^T(I_n - H_{-p})X_p}.$$

Hint: if A is an invertible symmetric $p \times p$ matrix and we break it into blocks $A = \begin{pmatrix} B & u \\ u^T & A_{pp} \end{pmatrix}$, with u the $(p-1) \times 1$ vector and B the $(p-1) \times (p-1)$ matrix with $u_i = A_{ip}$ and $B_{ij} = A_{ij}$ for $i, j = 1, \ldots, p-1$, then $[A^{-1}]_{pp} = (A_{pp} - u^T B^{-1} u)^{-1}$.

(b) Show that $X_p^T(I_n - H_{-p})X_p$ is the residual sum of squares for the regression of X_p on the remaining columns of X.

(c) What are the implications for regression with near linearly-dependent groups of variables? Use the following R-output (cigarette CO data) to illustrate your point.

```
> cig.lm1 <- lm(CO ~ 1 + Nicotine + Tar + Weight, data = cig)
> summary(cig.lm1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2022     3.4618   0.925 0.365464
Nicotine     -2.6317     3.9006  -0.675 0.507234
Tar           0.9626     0.2422   3.974 0.000692 ***
Weight       -0.1305     3.8853  -0.034 0.973527
...
> cig.lm2 <- lm(CO ~ 1 + Tar + Weight, data = cig)
> summary(cig.lm2)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.11433    3.41620   0.912    0.372
Tar          0.80419    0.05904  13.622 3.36e-12 ***
Weight      -0.42287    3.81299  -0.111    0.913
...
> cor(cig[2:4, 2:4]) # correlation matrix
               Tar  Nicotine    Weight
Tar      1.0000000 0.9960417 0.9290669
Nicotine 0.9960417 1.0000000 0.8925088
Weight   0.9290669 0.8925088 1.0000000
```

(d) (*extra for experts*) Consider how $\mathrm{var}(\widehat{\beta}_p)$ varies as the vector X_p is varied subject to $|X_p| = 1$, and subject also to X_{-p} (the design matrix X with column p excluded) being held fixed. Find the smallest value that $\mathrm{var}(\widehat{\beta}_p)$ can take over choices of the vector X_p, subject to $|X_p| = 1$ with X_{-p} fixed, and show that this minimum is achieved if and only if X_p is orthogonal to all the columns of X_{-p}.

[*Note that* $X_{-p} = \widetilde{X}$.]