Practice Examination 2016: Solutions

1. (a) *Show that for a linear regression model, the parameter estimates for the regression coefficients β derived under the least squares criterion are identical to the maximum likelihood estimates for the parameters under a specific distributional assumption made about the random residual errors, ϵ.*

Under distributional assumption for the residual error vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ that specifies (in the usual notation) $\epsilon|\mathbf{X} \sim \text{Normal}\,(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, the likelihood function for data $\mathbf{y}$ is

$$L(\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\right\}$$

The log-likelihood is

$$\begin{aligned}\ell(\beta, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) + \text{constant} \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\beta) + \text{constant}.\end{aligned}$$

We seek to maximize this function with respect to β and σ^2. It is evident that, in terms of β, the maximum value of $\ell(\beta, \sigma^2)$ is attained (for any σ^2) when $S(\beta)$ is minimized. Thus the maximum likelihood estimate of β is identical to the least squares estimate.

6 MARKS

(b) *Demonstrate the difference between (i) the estimate of σ^2 commonly used under least squares methodology, and (ii) the estimate of σ^2 derived using maximum likelihood.*

From above it is evident by inspecting the score equation

$$\frac{\partial\ell(\beta, \sigma^2)}{\partial\sigma^2} = 0$$

that the ML estimate is

$$\widehat{\sigma}^2_{\text{ML}} = \frac{1}{n}S(\widehat{\beta}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\beta})^\top(\mathbf{y} - \mathbf{X}\widehat{\beta}) = \left(\frac{n-2}{n}\right)\widehat{\sigma}^2$$

and thus the two estimates are different.

4 MARKS

(c) *Suppose that, in a study where a single continuous predictor X is recorded, a bivariate Normal distribution assumption is made about the joint distribution of X and Y.*

By considering this factorization of the joint density of X and Y, show that there is an exact correspondence between the parameters in the joint model and parameters that appear in a standard simple linear regression model.

We may factorize the joint density

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$$

where $X \sim \text{Normal}(\mu_X, \sigma_X^2)$, and

$$Y|X = x \sim \text{Normal}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right)$$

Equating the conditional expectation of Y given $X = x$

$$\mathbb{E}_{Y|X}[Y|x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

with a simple linear regression

$$\mathbb{E}_{Y|X}[Y|x] = \beta_0 + \beta_1 x$$

we identify

$$\beta_0 = \mu_Y - \mu_X\rho\frac{\sigma_Y}{\sigma_X} \qquad\qquad \beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$$

5 MARKS

(d) *Suppose that a simple linear model with intercept β_0 set equal to zero is considered. For this model, derive the least squares estimate of β_1, and the variance of the corresponding estimator under standard assumptions.*

In this setting, we have that

$$\widehat{\beta}_1 = \arg\min_{\beta}\sum_{i=1}^{n}(y_i - x_{i1}\beta_1)^2 = \arg\min_{\beta_1} S(\beta_1).$$

Taking the derivative of $S(\beta_1)$ wrt β_1 and equating to zero, we have to solve

$$-2\sum_{i=1}^{n} x_{i1}(y_i - x_{i1}\beta_1) = 0.$$

This leads to

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_{i1}y_i}{\sum_{i=1}^{n} x_{i1}^2}.$$

The variance of the estimator is easily computed as

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_{i1}Y_i}{\sum_{i=1}^{n} x_{i1}^2} = \sum_{i=1}^{n} c_i Y_i \qquad \text{where} \quad c_i = \frac{x_{i1}}{\sum_{i=1}^{n} x_{i1}^2}$$

and as the Y_i are independent,

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}_1|\mathbf{X}] = \sum_{i=1}^{n} c_i^2 \text{Var}_{X_i|\mathbf{X}}[Y_i|\mathbf{X}] = \sigma^2\sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_{i1}^2}.$$

5 MARKS

2. (a) *Explain the terms in decomposition (written in the standard notation of the course)*

$$\mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y} = \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y} + \mathbf{y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{y}.$$

This is the ANOVA sums of squares decomposition. We have hat matrices

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \qquad \mathbf{H}_1 = \mathbf{1}(\mathbf{1}^\top\mathbf{1})^{-1}\mathbf{1}^\top$$

where $\mathbf{1}$ is an $(n \times 1)$ vector of ones, and

$$\begin{aligned} \text{SS}_\text{T}: &= \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y} &= \sum_{i=1}^n (y_i - \overline{y})^2 \\ \text{SS}_\text{Res}: &= \mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{y} &= \sum_{i=1}^n (y_i - \widehat{y}_i)^2 \\ \text{SS}_\text{R}: &= \mathbf{y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{y} &= \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 \end{aligned}$$

It is a decomposition of the variation of the y_i around their mean $\overline{y}$ (SS_T) into the variation of the y_i around the fitted values $\widehat{y}_i$ (SS_Res) plus the variation of the $\widehat{y}_i$ around the global mean $\overline{y}$ (SS_R). We have

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2$$

as the cross product term in the expansion of the square

$$\sum_{i=1}^n (\widehat{y}_i - \overline{y})(\widehat{y}_i - \overline{y}) = 0.$$

8 MARKS

(b) *Consider the terms in the decomposition from (a) written in their random variable versions. State the distributional results related to the decomposition that are used in the construction of the 'global' Fisher-F test.*

Under the null hypothesis that the non-intercept parameters are in fact zero, we have that

$$\frac{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}}{\sigma^2} \sim \chi^2_{n-1} \qquad \frac{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}}{\sigma^2} \sim \chi^2_{n-p} \qquad \frac{\mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}}{\sigma^2} \sim \chi^2_{p-1}$$

This leads to the Fisher-F test based on the statistic

$$F = \frac{\mathbf{Y}^\top(\mathbf{H} - \mathbf{H}_1)\mathbf{Y}/(p-1)}{\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\mathbf{Y}/(n-p)} = \frac{\text{SS}_\text{R}/(p-1)}{\text{SS}_\text{Res}/(n-p)} \sim \text{Fisher}(p-1, n-p).$$

4 MARKS

(c) *In a simple linear regression analysis of a sample of size $n = 36$, the following sums of squares quantities were computed*

$$SS_R = 678.90 \qquad SS_{Res} = 112.84.$$

Reconstruct, in standard format, the ANOVA table for carrying out the F-test for these data.

The ANOVA table is as follows:

Source	SS	df	MS	F
Regression	678.90	1	678.90	204.56
Residual	112.84	34	3.32	
Total	791.74	35		

6 MARKS

Report the conclusion of the test if it is known that the 0.975 quantile of the Student-t distribution with 34 degrees of freedom lies at 2.032. The F-test can be carried out by realizing that a Fisher$(1, 34)$ random variable is the square of the Student(34) random variable, so that the critical value for the test is $2.032^2 = 4.129$, and the null hypothesis is rejected. 2 MARKS

3. (a) *If* $\mathbf{e} = (e_1, \ldots, e_n)^\top$ *denote the residuals from the fit of a multiple regression model based on an* $(n \times p)$ *matrix* $\mathbf{X}$ *to data* $\mathbf{y} = (y_1, \ldots, y_n)^\top$ *using least squares, show that*

$$\mathbf{X}^\top \mathbf{e} = \mathbf{0}$$

where the right hand side of this equation is a column vector of zeros of length p*.*

We have that $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$. But the normal equations dictate that $\widehat{\beta}$ is a solution to

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$$

so we must have $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$.

6 MARKS

(b) *If the model in (a) includes an intercept, show that*

$$\sum_{i=1}^{n} e_i = 0.$$

If the model contains an intercept, the first column of $\mathbf{X}$ is a column of ones, so therefore, in the calculation from (a) we consider the first row of the solution

$$\mathbf{1}^\top \mathbf{e} = \sum_{i=1}^{n} e_i = 0.$$

2 MARKS

(c) *Suppose that true regression model relating a single predictor* x_1 *to response* y *takes the form*

$$\mathbb{E}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$$

but that a simple linear regression model in x_1 *is fitted. Derive the expectation of the residuals from the fitted model, conditional on* x_1*, in this case.*

If $\mathbf{H}$ is the hat matrix for the **fitted** model, we have that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}|\mathbf{X}] = (\mathbf{I}_n - \mathbf{H})\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}]$$

but in reality

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = [\mathbf{X}\ \underset{\sim}{x}_1^2][\beta_0\ \beta_1\ \beta_2]^\top$$

where $\underset{\sim}{x}_1^2$ is the column vector $(x_{11}^2, \ldots, x_{n1}^2)^\top$. Therefore

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = (\mathbf{I}_n - \mathbf{H})[\mathbf{X}\ \underset{\sim}{x}_1^2] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = (\mathbf{I}_n - \mathbf{H}) \left(\mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \underset{\sim}{x}_1^2 \beta_2 \right)$$

If $\beta_2 = 0$, then we have $\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = 0$, otherwise

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \beta_2(\mathbf{I}_n - \mathbf{H})\underset{\sim}{x}_1^2$$

6 MARKS

(d) *Consider again data generated from the true model in (c), but where the residual errors $\epsilon_i, i = 1, \ldots, n$ are uncorrelated but have different variances, specifically*

$$\text{Var}_{Y_i|X_i}[Y_i|\mathbf{x}_i] = \sigma^2(1 + \exp(x_{i1})).$$

Derive the theoretical variance of the residuals from the fitted model in this case, conditional on x_1, if the correct conditional expectation model is fitted using (ordinary) least squares.

Using similar calculations to above, we have

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \text{Var}_{\mathbf{Y}|\mathbf{X}}[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}|\mathbf{X}] = (\mathbf{I}_n - \mathbf{H})\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}](\mathbf{I}_n - \mathbf{H})^\top.$$

In reality, however,

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{V}(\mathbf{X})$$

say, where $\mathbf{V}(\mathbf{X})$ is the diagonal matrix with entries $(1 + \exp(x_{i1})), i = 1, \ldots, n$. Thus

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H})\mathbf{V}(\mathbf{X})(\mathbf{I}_n - \mathbf{H})^\top.$$

As $\mathbf{V}(\mathbf{X})$ is diagonal and $(\mathbf{I}_n - \mathbf{H})$ symmetric, we may write

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H})(\mathbf{I}_n - \mathbf{H})^\top\mathbf{V}(\mathbf{X}) = \sigma^2(\mathbf{I}_n - \mathbf{H})\mathbf{V}(\mathbf{X})$$

so that, if h_{ii} is the ith diagonal element of $\mathbf{H}$, we have

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[Y_i - \widehat{Y}_i|\mathbf{X}] = \sigma^2(1 - h_{ii})(1 + \exp(x_{i1})).$$

Note: as the errors are still uncorrelated, we may deduce this result directly from standard theory that gives the result (under standard equal variance assumptions) that

$$\text{Var}_{\mathbf{Y}|\mathbf{X}}[Y_i - \widehat{Y}_i|\mathbf{X}] = \sigma^2(1 - h_{ii})$$

by changing the variance from σ^2 to $\sigma^2(1 + \exp(x_{i1}))$.

6 MARKS

4. *A **one way analysis of variance** (ANOVA) is based on a linear regression model for a single factor predictor that takes, say, K levels, which identify subgroups indexed by different levels of the factor. In this question, $K = 4$, and the subgroup means are $\mu_1, \mu_2, \mu_3, \mu_4$. The analysis aims to test the null hypothesis that the subgroups have the same expected value,*

$$\mathrm{H}_0 \ : \ \mu_1 = \mu_2 = \mu_3 = \mu_4,$$

under the standard assumptions of linear regression modelling.

(a) *If the total sample size is 20, and the study is balanced, construct the $\mathbf{X}$ matrix for fitting the one way ANOVA model, and write down the form of $\mathbf{X}^\top\mathbf{X}$, if the standard contrast parameterization is used.*

In this setting, as the study is balanced, we have four groups of five observations each. In the contrast parameterization, we take the first group as the baseline group, and define

$$\beta_0 = \mu_1 \qquad \beta_1 = \mu_2 - \mu_1 \qquad \beta_2 = \mu_3 - \mu_1 \qquad \beta_3 = \mu_4 - \mu_1.$$

so that

$$\mu_1 = \beta_0 \qquad \mu_2 = \beta_0 + \beta_1 \qquad \mu_3 = \beta_0 + \beta_2 \qquad \mu_3 = \beta_0 + \beta_2.$$

If we arrange the data in group order (that is, five from group 1, followed by five from group two etc.), then we have for the β parameterization

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and so

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} 20 & 5 & 5 & 5 \\ 5 & 5 & 0 & 0 \\ 5 & 0 & 5 & 0 \\ 5 & 0 & 0 & 5 \end{bmatrix}$$

6 MARKS

(b) *Write down $\mathbf{X}$ if a parameterization in terms of $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)^\top$ is used, where*

$$\gamma_0 = \overline{\mu} \qquad \gamma_1 = \mu_1 - \overline{\mu} \qquad \gamma_2 = \mu_2 - \overline{\mu} \qquad \gamma_3 = \mu_3 - \overline{\mu}$$

where

$$\overline{\mu} = \frac{1}{4}\sum_{j=1}^{4} \mu_j.$$

In the reparameterized version, we have that

$$\mu_1 = \gamma_0 + \gamma_1 \qquad \mu_2 = \gamma_0 + \gamma_2 \qquad \mu_3 = \gamma_0 + \gamma_3 \qquad \mu_4 = \gamma_0 - \gamma_1 - \gamma_2 - \gamma_3.$$

Thus in the new parameterization

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

6 MARKS

(c) *The following edited* `R` *output records the analysis of a data set under the conditions of the question:* `x1` *is the factor predictor. Do the results indicate that the null hypothesis above should be rejected ? Justify your answer.*

```
1  > summary(lm(y ~ x1-1))
2  Coefficients:
3      Estimate Std. Error t value Pr(>|t|)
4  x11   2.0944     0.2348   8.920 1.31e-07 ***
5  x12   1.9157     0.2348   8.159 4.29e-07 ***
6  x13   2.6391     0.2348  11.240 5.27e-09 ***
7  x14   1.7602     0.2348   7.497 1.27e-06 ***
8  ---
9  Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
10
11 Residual standard error: 0.525 on 16 degrees of freedom
12 Multiple R-squared:  0.9536,    Adjusted R-squared:  0.942
13 F-statistic: 82.17 on 4 and 16 DF,  p-value: 1.86e-10
```

NB: Read the output carefully.

In the first line, and lines 4–7, we can see that this is not the standard contrast parameterization, but instead the parameterization in terms of $\mu_1, \mu_2, \mu_3, \mu_4$ that omits an explicit intercept. Thus the null hypothesis required is not being tested by the F-test on line 13; note, for example, the incorrect degrees of freedom (it should be 3, not 4). The decomposition used here is

$$\overline{\text{SS}}_\text{T} = \text{SS}_\text{Res} + \overline{\text{SS}}_\text{R}$$

but we need $\text{SS}_\text{T} = \text{SS}_\text{Res} + \text{SS}_\text{R}$. However, this decomposition can be computed in this balanced design. SS_Res is preserved under the two parameterizations, and we have from line 11 $\text{SS}_\text{Res} = 16 \times 0.525^2 = 4.41$. Also, by definition, as $\widehat{y}_i = \overline{y}_j$ for data in group j, we have

$$\text{SS}_\text{R} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{j=1}^{4} 5(\overline{y}_j - \overline{y})^2 = 2.20$$

as $\overline{y}_j, j = 1, \ldots, 4$ are given by lines 4-7, and $\overline{y} = (\overline{y}_1 + \overline{y}_2 + \overline{y}_3 + \overline{y}_4)/4 = 2.102$. Hence for the desired test, we have

$$F = \frac{\text{SS}_\text{R}/(p-1)}{\text{SS}_\text{Res}/(n-p)} = \frac{2.20/3}{4.41/16} = 2.661$$

This number would then need to be compared against the Fisher$(3, 16)$ distribution.

8 MARKS

5. (a) *Explain the difference between a model selection procedure based on stepwise (forward and/or backward) selection and a selection procedure based on the information criteria AIC or BIC. Explain specifically the statistical summaries used and compared.*

Stepwise testing relies on sequential comparisons of Full and Reduced models using 'partial' F-tests to test null hypotheses that specify that subsets of parameters in the Full model are equal to zero. The test relies upon SS_{Res} quantities for each model, and compares them in a pairwise fashion, taking into account the number of parameters that are set to zero. If the Full model has p parameters, and r are set to zero to form the Reduced model, we have

$$F = \frac{(\text{SS}_{\text{Res}}(\text{Reduced}) - \text{SS}_{\text{Res}}(\text{Full})/r}{\text{SS}_{\text{Res}}(\text{Full})/(n-p)}$$

which is distributed as Fisher$(r, n-p)$ if the Reduced model is an adequate simplification.

AIC and BIC do not operate in a sequential fashion, but instead can be computed across all models to be compared simultaneously. We have, up to additive constants that depend on n but not the model, we have for a model with p regression coefficients

$$\text{AIC} = n \log\left(\frac{\text{SS}_{\text{Res}}}{n}\right) + 2(p+1).$$

and

$$\text{BIC} = n \log\left(\frac{\text{SS}_{\text{Res}}}{n}\right) + (p+1)\log(n).$$

The model with the smallest AIC or BIC is chosen. The second term in each expression acts as a penalty that increases with increasing model complexity.

5 MARKS

(b) *The following output in* `R` *demonstrates the analysis of a data set where three continuous predictors are used to explain the variation in an observed response:*

```
> fit1<-lm(y ~ x1+x2+x3)
> fit2<-lm(y ~ (x1+x2+x3)^2)
> drop1(fit2,test='F')
Single term deletions
Model:
y  ~  (x1 + x2 + x3)^2
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              103.68 50.656
x1:x2   1    12.310 115.99 51.797  2.4932 0.1292864
x1:x3   1   109.648 213.33 68.858 22.2079 0.0001185 ***
x2:x3   1     1.897 105.58 49.164  0.3842 0.5420344
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
> fit3<-update(fit2, ~ .-x2:x3)
```

```
> drop1(fit3,test='F')
Single term deletions
Model:
y ~ x1 + x2 + x3 + x1:x2 + x1:x3
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              105.58 49.164
x1:x2   1     11.23 116.81 49.994  2.3401    0.1403
x1:x3   1    120.03 225.61 68.425 25.0109 5.255e-05 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
> anova(fit3)
Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1  35.330  35.330  7.3618   0.01269 *
x2         1   1.055   1.055  0.2198   0.64378
x3         1   0.641   0.641  0.1336   0.71820
x1:x2      1   0.777   0.777  0.1620   0.69124
x1:x3      1 120.030 120.030 25.0109 5.255e-05 ***
Residuals 22 105.581   4.799
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
> fit4<-lm(y~x1+x2+x3+x1:x3)
> drop1(fit4,test='F')
Single term deletions

Model:
y ~ x1 + x2 + x3 + x1:x3
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              116.81 49.994
x2      1     3.007 119.82 48.706  0.5921 0.4494427
x1:x3   1   109.577 226.39 66.521 21.5756 0.0001128 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
> fit5<-lm(y~x1+x3+x1:x3)
> drop1(fit5,test='F')
Single term deletions

Model:
y ~ x1 + x3 + x1:x3
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              119.82 48.706
x1:x3   1    106.71 226.53 64.538  21.374 0.0001083 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```

In standard notation, list the five models fitted, and summarize the conclusions to be drawn from the analysis.

The five models fitted are:

- line 1: $X_1 + X_2 + X_3$;
- line 2: $(X_1 + X_2 + X_3)^2 = X_1 + X_2 + X_3 + X_1 : X_2 + X_1 : X_3 + X_2 : X_3$;
- line 14: $X_1 + X_2 + X_3 + X_1 : X_2 + X_1 : X_3$;
- line 37: $X_1 + X_2 + X_3 + X_1 : X_3$;
- line 49: $X_1 + X_3 + X_1 : X_3$

A stepwise procedure leads us to conclude that the model

$$X_1 * X_3 = X_1 + X_3 + X_1 : X_3$$

is the most suitable model, as all the more complicated models can be simplified without significant loss of fit. The single term deletions quickly identify the terms that can be omitted, until the fit of the final model reveals that the interaction $X_1 : X_3$ should not be omitted.

In terms of AIC, we have that amongst the fitted models, this identified model also has the lowest AIC.

10 MARKS

(c) *The formula for the BIC for a model with p regression parameters β as computed by* R *is*

$$\text{BIC}(\beta) = n \log\left(\frac{\text{SS}_{\text{Res}}(\beta)}{n}\right) + (p+1)\log(n) + \text{constant}$$

where the constant depends on the sample size, and is identical for all models fitted.

Compute the BIC for the models fitted in the R *output above, ignoring the constant term. Explain which model is selected using BIC here.*

We note that

$$\text{BIC} = \text{AIC} + (p+1)(\log n - 2).$$

Here we know that $n = 28$ (see the output lines 25–34). Hence $\log n - 2 = 1.3322$). From the output, we can compute as follows:

Model	Line	AIC	$p+1$	BIC
Model 1	46	66.521	5	73.182
Model 2	8	50.656	8	61.314
Model 3	20	49.164	7	58.489
Model 4	44	49.994	6	57.987
Model 5	56	48.706	5	55.367

and Model 5 is selected once more.

5 MARKS