Thursday, 7th May 2015
9.30 am - 11.00 am

EXAMINATION FOR THE DEGREE OF M.SCI. (TAUGHT)
(SCIENCE)

Regression Modelling: Level M

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

NOTE: If all four questions are attempted, candidates should clearly indicate which three questions they wish to be marked. Otherwise, only the first three questions in the script book will be marked.

1. (a) For the following linear models, write down the standard vector-matrix form $E(\underline{Y}) = X\underline{\beta}$ clearly identifying $\underline{Y}, X$, and $\underline{\beta}$:

   (i) Data: $y_{ij}, i = 1, 2; j = 1, \ldots, n_i$.

   Model: $E(Y_{ij}) = \mu + \alpha_i$.

   (ii) Data: $(y_i, x_i), i = 1, \ldots, n$.

   Model: $E(Y_i) = \beta_0 + \beta_1(x_i - \bar{x})$.

   (iii) Data: $(y_i, x_i), i = 1, \ldots, 2n$.

   Model: $E(Y_i) = \alpha + \beta x_i + \gamma D_i + \delta x_i D_i$, where $D_i = 1$ for $i = 1, \ldots, n$ and 0 otherwise. **[6 MARKS]**

CONTINUED OVERLEAF/

(b) Define a residual and a studentized residual for the model $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$. **[3 MARKS]**

(c) Show that the residuals for the model in (b) are linear combinations of the responses. **[3 MARKS]**

(d) State the Gauss-Markov assumptions and the Gauss-Markov Theorem and explain why it is important. **[3 MARKS]**

(e) Consider the model $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$ and the least squares method to estimate the parameter vector $\underline{\beta}$. Sketch a diagram that shows the geometric view of the least squares for the linear model above. Add a hyperplane in your diagram and identify the response $\underline{Y}$, the vector of fitted values $\hat{\underline{Y}}$, and the vector of residuals $\hat{\underline{\varepsilon}}$. **[5 MARKS]**

2. (a) Explain how residuals plots can be used to examine the linear model assumptions:

   (i) The errors have constant variance.

   (ii) The errors are normally distributed.

   [4 MARKS]

(b) Define the leverage of the ith observation. If it is large, what does this mean? **[2 MARKS]**

(c) State what the coefficient of determination of a linear model is (for example, for a model with intercept); no need to define it. The following quantities:

- coefficient of determination (for models with/without intercept), the
- adjusted coefficient of determination (for models with/without intercept), the
- mean residual sum of squares, and the
- Akaike Information Criterion

can be used to compare models. Identify which of these quantities can be used to compare:

(i) two linear models with intercept and same number of explanatory variables,

(ii) any two linear models (regardless whether they have intercept or not, or their respective numbers of explanatory variables),

(iii) two linear models without intercept and same number of explanatory variables, and

(iv) two linear models with intercept and different number of explanatory variables, and

(v) two linear models without intercept and different number of explanatory variables.

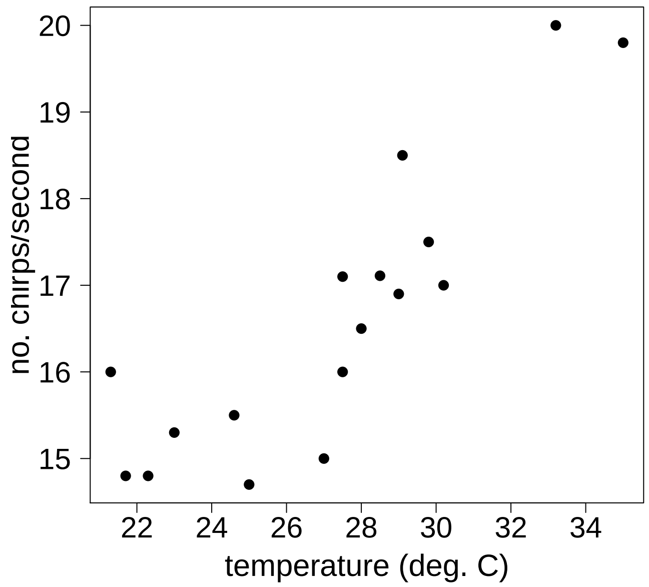
CONTINUED OVERLEAF/

[4 MARKS]

(d) Suppose we have the maximum daily temperature and the amount of ice cream sold in each of 60 days in a town close to Glasgow. The estimated Pearson correlation coefficient between maximum daily temperature and amount of ice cream sold is 0.862. Using statistical tables, construct a 95% confidence interval for the population correlation coefficient and comment on its value. **[5 MARKS]**

(e) Explain what you could do if the assumption of equality of variances (namely, constant variances) in a normal linear model is violated. **[5 MARKS]**

3. (a) A cricket is a small insect that makes a chirp (loud high sound) by rubbing its wings together. An entomologist recorded the temperature in degrees centigrade (°C) and the number of chirps per second made by 17 crickets. His/her aim was to estimate the number of chirps per second depending on the temperature. We can see a scatter plot of the data below.



The entomologist proposed the simple linear regression model:

$$\text{no.chirpspersec}_i = \beta_0 + \beta_1 \text{temperature}_i + \varepsilon_i, i = 1, \ldots, 17$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and ε_is are independent for $i = 1, \ldots, 17$.
Some of the results of fitting this model are shown below:

$$\widehat{\text{no.chirpspersec}} = 6.58 + 0.37\,\text{temperature}.$$

CONTINUED OVERLEAF/

```
                  Coef.   se(Coef)
Intercept         6.58    1.4851
temperature       0.37    0.0541

 Analysis of Variance Table

                  Sum Sq      Df   Mean Sq     F value      Pr(>F)
Model             32.724       1         a           b    5.7e-06 ***
Residuals              c      15     0.702
Total             43.255      16
```

$$(X^T X)^{-1} = \begin{pmatrix} 3.1415776 & -0.113263063 \\ -0.1132631 & 0.004161383 \end{pmatrix}.$$

where X is the design matrix.

(i) Complete the ANOVA table above with the values of a, b and c; calculate and comment on the value of R^2 for this model. **[5 MARKS]**

(ii) Calculate a 95% confidence interval for β_1. **[4 MARKS]**

(iii) Calculate a 95% prediction interval for the number of chirps per second when the temperature was 31 °C. **[4 MARKS]**

(iv) Comment on the adequacy of the model. **[4 MARKS]**

(b) For the linear model $\underline{Y} = X\underline{\beta}+\underline{\varepsilon}$, state the formulae for the least squares estimator (LSE) of $\underline{\beta}$ and for the residual sum of squares in terms of the LSE, both in vector-matrix notation.

[3 MARKS]

4. (a) Data from 21 pieces of synthetic fibre used for manufacturing wigs have been collected. The synthetic fibre has to be shrunk before the manufacture of the wigs. To shrink it, each piece of fibre is put in a solution with a certain chemical and with a very high temperature. We have the values of the concentration of the chemical (x) and temperature (D coded as 1=high or 0=medium) that were used for each piece of fibre and the measured percentage of shrinkage (y) when the fibre was dried. There were 9 pieces assigned to high temperature and 12 to medium temperature.

It is of interest to investigate the extent to which the percentage of shrinkage depends on the chemical concentration and on the temperature. Hence the following model was initially proposed:

$$\text{PercentShrinkage}_s = \alpha + \beta\,\text{concentration}_s + \gamma D_s + \delta\,\text{concentration}_s\,D_s + \varepsilon_s, \quad (1)$$

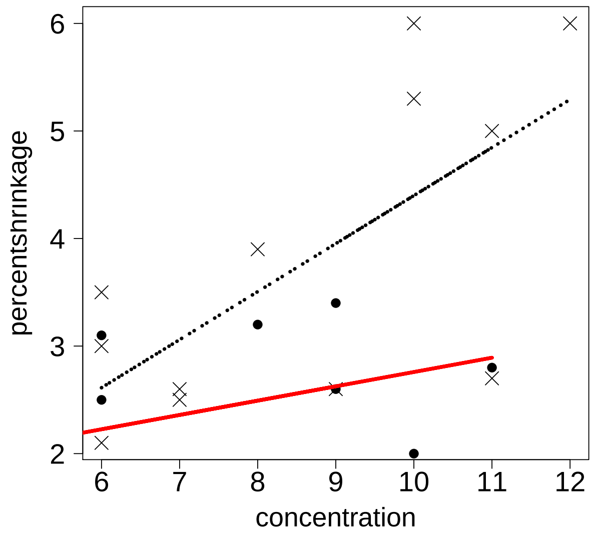
CONTINUED OVERLEAF/

where $D_s = 1$ for $s = 1, \ldots, 9$ (high temperature) and $D_s = 0$ otherwise ($s = 10, \ldots, 21$: medium temperature), and ε_s are assumed to be i.i.d. $N(0, \sigma^2)$. (Additional comment: this model is equivalent to:

$$\text{PercentShrinkage}_{ij} = \mu + \alpha_i + \beta_i \text{concentration}_{ij} + \varepsilon_{ij},$$

$i = 1, 2$; (high and medium temperature resp.) $j = 1, \ldots, n_i$, where $n_1 = 9$, $n_2 = 12$ and ε_{ij} are assumed to be i.i.d. $N(0, \sigma^2)$.)

We also have a scatter plot of the percentage of shrinkage versus concentration with the fitted regression model above for the two temperature groups; see plot below where the crosses represent observations for medium temperature, the black dots those for high temperature, the dashed line is the fitted regression line for medium temperature and the continuous line, that for high temperature.



The results of fitting model (1) are shown below, giving parameter estimates, standard errors and p-values (so the first column with numbers indicates the values for $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ and $\hat{\delta}$ resp.):

```
                          Estimate Std. Error  Pr(>|t|)
(Intercept)               -0.06677    1.18890   0.9559
concentration              0.44661    0.13454   0.0041 **
temperhigh                 1.49447    1.82553   0.4243
concentration:temperhigh  -0.31351    0.21732   0.1673

Multiple R-squared:  0.5472, Adjusted R-squared:  0.4673
```

(i) If we want to know if the effect of concentration on the percentage of shrinkage is different depending on the temperature level, which parameter in the model above indicates this: β, γ or δ? Is it significant in the model fit? **[4 MARKS]**

CONTINUED OVERLEAF/

(ii) Comment on the value for the coefficient of determination. **[3 MARKS]**

(iii) Comment on the adequacy of model (1) which is a two different lines model. Instead of this model, we could consider a second model (2) below.

$$\text{PercentShrinkage}_s = \alpha + \beta\text{concentration}_s + \gamma D_s + \varepsilon_s; \qquad (2)$$

the results of fitting this model are shown below:

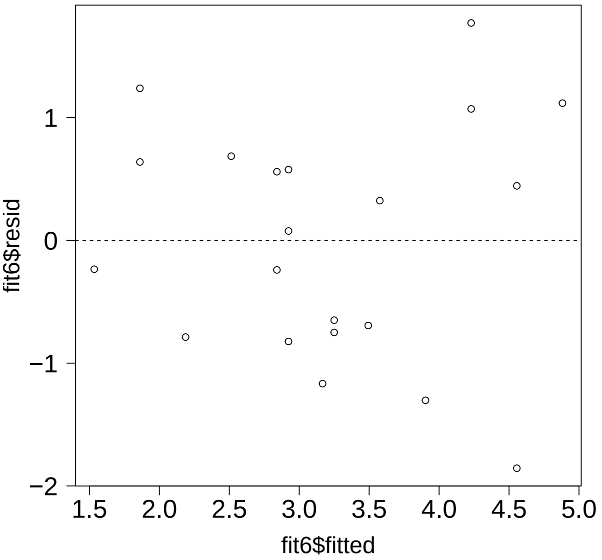
```
              Estimate Std. Error  Pr(>|t|)
(Intercept)     0.9646     0.9780    0.3370
concentration   0.3264     0.1088    0.0077 **
temperhigh     -1.0622     0.4507    0.0300 *

Residual standard error: 1.008 on 18 degrees of freedom
Multiple R-squared:  0.4918,Adjusted R-squared:  0.4353
F-statistic: 8.709 on 2 and 18 DF,  p-value: 0.002261
```

Interpret the output and in light of the results from fitting model (1) and model (2) describe the final model you would select.

[5 MARKS]

(iv) Describe a model selection stepwise method that starts in model (1) above and finds a best fitting model. Specify the criterion that you can use to compare models in the model selection method. **[5 MARKS]**

(v) Comment on the following graph of the residuals versus the fitted values, which corresponds to model (2):



[3 MARKS]

END OF QUESTION PAPER.