University of Glasgow

**December 2016**
**1 hour 30 mins**

**EXAMINATION FOR THE DEGREE OF MASTERS (SCIENCE)**

# Regression Modelling

*"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: Candidates should attempt all 4 questions. Questions are not equally weighted.**

1. Answer the following questions:

   (a) State whether the following three statements are TRUE or FALSE. Provide a one line explanation for your answer.

      i. The parameters to be estimated in the simple linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$, for $i = 1, \ldots, n$ where $\epsilon_i \sim N(0, \sigma^2)$ are $\alpha, \beta$ and $\epsilon_i$'s.

      **[1 MARKS]**

      ii. Consider two alternative linear models that are nested. If we compare the models on the basis of the adjusted R-squared and the R-squared, the R-squared will prefer the extended model more often than the adjusted R-squared.

      **[1 MARKS]**

**CONTINUED OVERLEAF/**

iii. If a covariate in a model is significant at the 5% level, it is also significant at the 10% level. **[2 MARKS]**

For part (b)-(f) choose the correct answers from the four options. Note that in some cases more than one answer may be correct

(b) In a regression study, a 95% confidence interval for $\beta_1$ was given as: (-5.65, 2.61). What would a test for $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ conclude? **[3 MARKS]**

i. reject the null hypothesis at $\alpha$=0.05 and all smaller $\alpha$

ii. fail to reject the null hypothesis at $\alpha$=0.05 and all smaller $\alpha$

iii. reject the null hypothesis at $\alpha$=0.05 and all larger $\alpha$

iv. fail to reject the null hypothesis at $\alpha$=0.05 and all larger $\alpha$

(c) If a predictor variable $x$ is found to be highly significant we would conclude that: **[2 MARKS]**

i. a change in $y$ causes a change in $x$

ii. a change in $x$ causes a change in $y$

iii. changes in $x$ are not related to changes in $y$

iv. changes in $x$ are associated with changes in $y$

(d) In the regression model $Y_i = \alpha + \beta x_i + \epsilon_i$, for $i = 1, \ldots, n$. a change in $y$ for a one unit increase in $x$: **[2 MARKS]**

i. will always be the same amount, $\alpha$

ii. will always be the same amount, $\beta$

iii. will depend on the error term

iv. will depend on the level of $x$

(e) The following appeared in the magazine Financial Times, March 23, 1995: "When Elvis Presley died in 1977, there were 48 professional Elvis impersonators. Today there are an estimated 7328. If that growth is projected, by the year 2012 one person in four on the face of the globe will be an Elvis impersonator." This is an example of: **[2 MARKS]**

i. extrapolation

ii. dummy variables

**CONTINUED OVERLEAF/**

iii. misuse of causality

iv. multicollinearity

(f) Which of the following ANOVA components are not additive? **[2 MARKS]**

i. Mean squares

ii. Sum of squares

iii. Degrees of freedom

iv. None of the above

2. Consider the following model:

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

Model 1: $Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad \epsilon_{ij}$'s independent

(a) For Model 1, write the model in vector-matrix form $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, identifying clearly the elements of $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$. **[3 MARKS]**

(b) State the general formula in vector-matrix form for the least squares estimators of the vector of parameters $\boldsymbol{\beta}$. **[1 MARK]**

(c) Use the general formula from part (b) to derive estimators for $\mu_1$ and $\mu_2$. **[3 MARKS]**

(d) Use the result from part (c) to state an expression for the residual sum of squares and hence to state an algebraic expression for the estimate of the error variance $\hat{\sigma^2}$. **[3, 3 MARKS]**

3. A study was conducted on 97 men with prostate cancer who were due to receive a radical prostatectomy (a surgical procedure). A doctor was interested in the relationship between the logarithm of the size of the cancer (`lcavol`) and a potential continuous predictor variable, the logarithm of the prostate specific antigen (`lpsa`).

(a) What may be the reason for using logarithmic transformations of the data rather than the original measurements? **[2 MARKS]**

(b) Comment on the scatterplot below with respect to the research interests behind the study. **[2 MARKS]**

**CONTINUED OVERLEAF/**

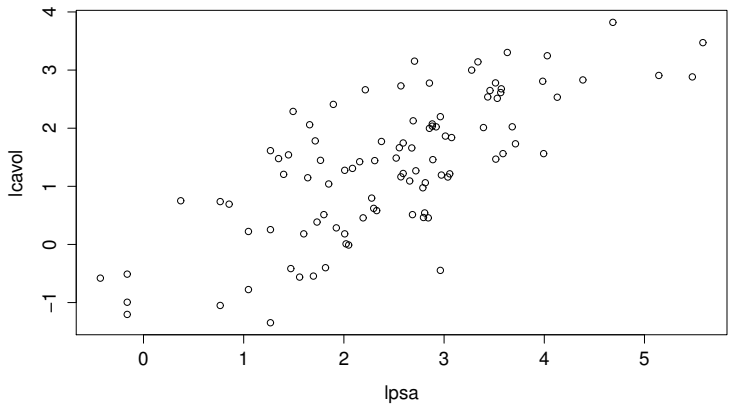



Figure 1: Plot of Log Cancer Volume (`lcavol`) against Log Prostate Specific Antigen (lpsa)

(c) Assuming that a linear relationship is appropriate, the sample correlation coefficient between `lcavol` and `lpsa` was computed to be 0.734. Use the statistical tables to perform a test (at a significance level of $\alpha = 5\%$) of the null hypothesis that the population correlation coefficient $\rho$ is 0 vs $\mathrm{H}_A : \rho \neq 0$. and comment on the results of the test. **[3 MARKS]**

(d) The following linear model, Model 1, was fitted to these data, with the log cancer volume, `lcavol`, as the response ($Y$) and the log prostate specific antigen, `lpsa`, as the explanatory variable ($x$),

Model 1 : $Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, 97, \quad \epsilon_i \sim N(0, \sigma^2), \quad \epsilon_i$'s independent.

A selection of the `R` output from fitting Model 1 is displayed below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50858      A     -2.619   0.0103
lpsa         0.74992    0.07109  10.548   <2e-16

Analysis of Variance Table

Response: lcavol
          Df Sum Sq Mean Sq F value    Pr(>F)
lpsa       1    B    71.938  111.27 < 2.2e-16
Residuals 95 61.421   0.647
```

**CONTINUED OVERLEAF/**

The model can be written in standard vector-matrix form $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, where $(\mathbf{X}^T\mathbf{X})^{-1}$ is

```
 0.05832771 -0.019374874
-0.01937487  0.007817534
```

i. Use the `R` output in the previous page to compute the standard error (`Std. Error`) for the intercept term, labelled as `A` in the output above. **[2 MARKS]**

ii. Comment on the p-value for the coefficient of `lpsa` with regard to what it tells us about the relationship with `lcavol`. **[2 MARKS]**

iii. Use the `R` output to compute a 95% confidence interval for the population mean log cancer volume when the `lpsa` recorded is 2.5. **[4 MARKS]**

iv. Use the `R` output in the previous page to compute the Sum of Squares (`Sum Sq`) for the model, labelled as `B` in the output. **[1 MARK]**

(e) For the two plots provided in Figure 2, explain for each plot which assumption of the normal linear model it is useful for assessing and comment specifically on whether or not the assumptions appear valid in this context. **[4 MARKS]**

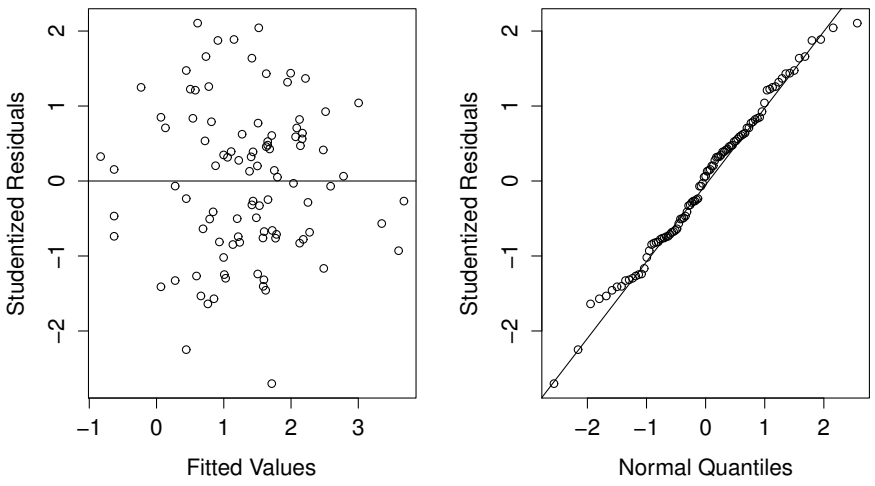


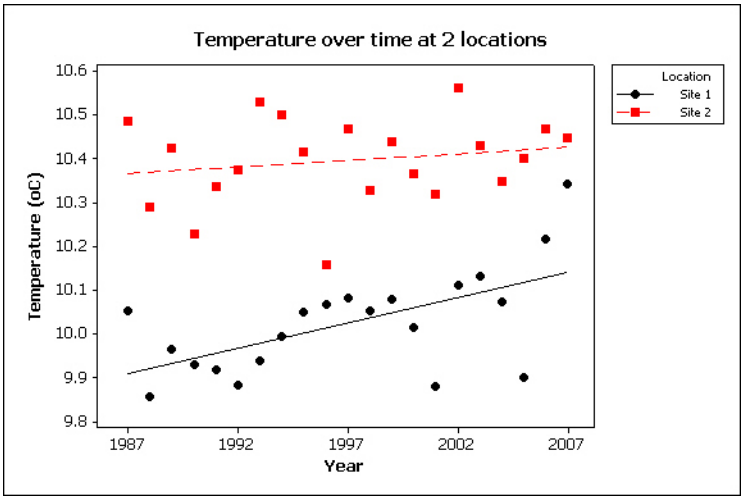

Figure 2: Standardised residuals versus fitted values (left) and Normal Q-Q plot for standardised residuals (right) from Model 1.

**CONTINUED OVERLEAF/**

4. An ecologist is interested in whether the temperature trend over time (from 1987 to 2007 i.e. 21 years) in a large deep loch (lake) in Scotland is different depending on the location that the measurements are taken. Measurements are taken at two sites in the loch, one in the north (site 1) and one in the south (site 2). He believes that the relationship between temperature ($y$) and time (year, $x$) at each site can be described by the following model,

$$Y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_{i\cdot}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad \text{and } \varepsilon_{ij} \text{ independent } i = 1, 2, j = 1, ...21.$$

(a) A plot of the data is provided below with a separate regression line for each site. Comment on whether the ecologist's model appears reasonable. **[3 MARKS]**




(b) Write the ecologist's model in vector-matrix form, $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, clearly identifying $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\beta}$.

**[3 MARKS]**

(c) Using part (b), derive $(\mathbf{X}^T\mathbf{X})^{-1}$ **[2 MARKS]**

(d) The general formula for a confidence interval for a linear combination of $\beta$ can be written as:

$$\mathbf{b}^T\hat{\beta} \pm t(n-k; 0.975)\sqrt{\frac{r}{n-k}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

where $\mathbf{b}$ is a vector of constants, $n$ is the sample size, $k$ is the number of parameters and $r$ is the residual sum-of-squares.

**CONTINUED OVERLEAF/**

Using this formula, part(c) and the summary statistics below, calculate a 95% confidence interval for $\beta_1 - \beta_2$ for the ecologist's model. Comment on whether it is plausible that the two regression lines could in fact be parallel.

Summary statistics:

$$\hat{\beta}_1 = 0.010, \hat{\beta}_2 = 0.003, r = 0.369946$$

$$S_{x_1x_1} = \sum_{j=1}^{21}(x_{1j} - \bar{x}_1.) = 0.2852903, S_{x_2x_2} = \sum_{j=1}^{21}(x_{2j} - \bar{x}_2.) = 0.1942308$$

where $r$ is the residual sum-of-squares for the ecologist's model.

**[4 MARKS]**

**Total: 60**

**END OF QUESTION PAPER.**