University of Glasgow

December 2016
1 hour 30 mins

EXAMINATION FOR THE DEGREE OF MASTERS
(SOLUTION)

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. FALSE:. α, β, σ^2 are parameters

      ii. TRUE: R-squared does not adjust for the degrees of freedom

      iii. TRUE. If a variable is significant at the 5% level, it is also significant at the 10% level.

   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) Answer (ii)

CONTINUED OVERLEAF/

(c) Answer (iv)

(d) Answer (ii)

(e) Answer (iv)

(f) Answer (i)

2. **Parts (a) and (b) are standard types of question as seen in lectures, tutorials and past papers. However, parts (c) and (d) are unseen variations on examples done in lectures and require combining knowledge together.**

(a)

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\beta} \\ \boldsymbol{\beta}^T &= (\mu_1, \mu_2) \end{aligned}$$

[1 MARK]

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ . \\ . \\ y_{1n_1} \\ y_{21} \\ . \\ . \\ y_{2n_2} \end{pmatrix} \quad \textbf{[1 MARK]} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ . & . \\ . & . \\ 1 & 0 \\ 0 & 1 \\ . & . \\ . & . \\ 0 & 1 \end{pmatrix} \quad \textbf{[1 MARK]}$$

(b) $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ **[1 MARK]**

(c)

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix} \\ \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} &= \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \end{pmatrix} \end{aligned}$$

[3 MARKS]

CONTINUED OVERLEAF/

(d)

$$\begin{aligned} RSS &= \mathbf{Y}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{Y})^T\hat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2 - \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \end{pmatrix}^T \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \end{pmatrix} \\ &= \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2 - n_1\bar{y}_{1.}^2 - n_2\bar{y}_{2.}^2 \end{aligned}$$

[3 MARKS]

Therefore, since df_{err} = sample size minus number of parameters $= n_1 + n_2 - 2$,

$$\hat{\sigma^2} = \frac{RSS}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2 - n_1\bar{y}_{1.}^2 - n_2\bar{y}_{2.}^2}{n_1 + n_2 - 2}$$

[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) If the original measurements are skewed then logarithms can transform the data so that the relationship between the variables is more consistent with a linear relationship **[2 MARKS]**

   (b) There appears to be a moderate positive linear relationship between lcavol and lpsa **[2 MARKS]**

   (c) Since $n = 97$ and $\alpha = 0.05$, from statistical tables (using closest n=95) we get $r_{n,\alpha} = 0.2017$. **[1 MARKS]**

   Therefore, we reject the null hypothesis $\text{H}_0 : \rho = 0$ if the absolute value for the correlation coefficient is greater than 0.2017. **[1 MARKS]**

   Since the sample correlation coefficient is 0.734, reject H_0 and conclude that we do have a statistically significant linear relationship between lcavol and lpsa. **[1 MARKS]**

   The sample correlation coefficient suggests that this is a strong positive linear relationship. **[1 MARK]**

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

CONTINUED OVERLEAF/

SOLUTION: Estimated standard error for the intercept

$$\mathbf{b}^T = [1, 0]$$

$$\sqrt{\frac{RSS}{n-p}\{\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}\}} = \sqrt{\frac{61.421}{97-2}\{0.05832771\}} = 0.194$$

[2 MARKS]

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

SOLUTION: Since the p-value for the coefficient of lpsa is $< 2\times10^{-16}$ **and hence** < 0.05**, we reject the null hypothesis that** $\beta = 0$ **and conclude that there is a statistically significant relationship between lcavol and lpsa.** **[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]**

SOLUTION: 95% **Confidence Interval for the population mean:**

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}\{\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}\}}$$

$$\mathbf{b}^T = (1, 2.5)$$

[1 MARK]

$$-0.50858 + 2.5 \times 0.74992 \pm t(97-2; 0.975)\sqrt{\frac{61.421}{95}\{\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}\}}$$

[1 MARK]

$$\begin{aligned}
\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b} &= (1, 2.5)\begin{pmatrix} 0.05832771 & -0.019374874 \\ -0.01937487 & 0.007817534 \end{pmatrix}\begin{pmatrix} 1 \\ 2.5 \end{pmatrix} \\
&= (0.009890535, 0.000168961)\begin{pmatrix} 1 \\ 2.5 \end{pmatrix} \\
&= 0.0103 \quad \textbf{[2 MARKS]}
\end{aligned}$$

[1 MARK]

95% confidence interval

CONTINUED OVERLEAF/

$$
\begin{aligned}
-0.50858 + 2.5 \times 0.74992 &\pm 1.98\sqrt{\frac{61.421}{95}(0.0103)} \\
1.36622 &\pm 1.98(0.0816852) \\
1.36622 &\pm 0.1617367 \\
(1.204483 &, 1.527957) \quad \textbf{[1 MARK]}
\end{aligned}
$$

The population mean lcavol value is highly likely to lie between 1.2 and 1.5 log ml for lpsa measurement of 2.5. **[1 MARK]**

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

SOLUTION: Sum of squares for the Model = MSModel $\times$ dfModel = 71.938 $\times$ 1 = 71.938 **[1 MARKS]**

CONTINUED OVERLEAF/

(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. **[5 MARKS]**

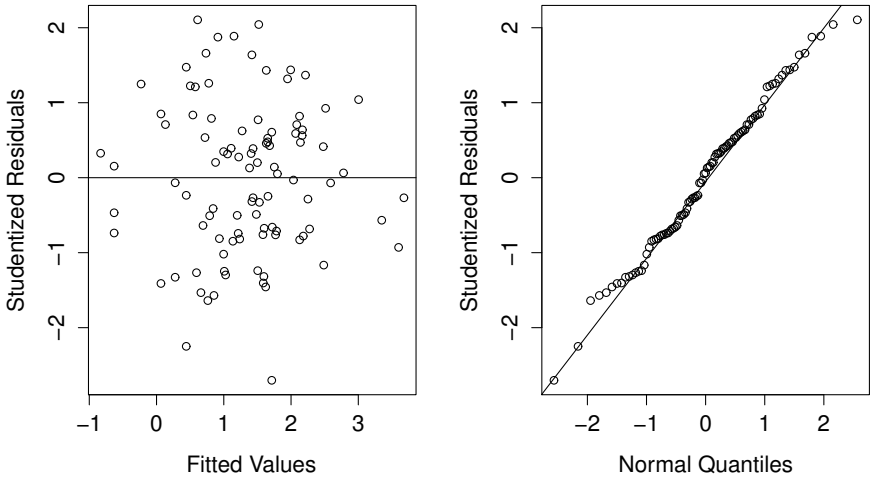

Figure 1: Standardised residuals versus fitted values (left) and Normal Q-Q plot for standardised residuals (right) from Model 1.

SOLUTION: The plot of the (standardised) residuals versus the fitted values would be used to assess whether or not the assumption of constant variance of the errors was plausible, and additionally if the deterministic part of the model was appropriate. Since there is no curvature in the plot the deterministic part appears appropriate. However, as indicated here and in the initial scatter plot for the data there is more variability at lower values than for higher values. The assumption of constant variance may be dubious. However, it may also be simply a feature of less data at higher values. [2 MARKS]

SOLUTION: The normal probability (Q-Q) plot can be used to assess the assumption of normally distributed errors. Here the points follow an approximate diagonal line. However, there is a small amount of weaving around the line which may suggest that the assumption of normality is slightly dubious. [2 MARKS]

CONTINUED OVERLEAF/

4. (a) The plot highlights a positive increasing trend (i.e. relationship) over time for site 1 whereas there is little trend evident for site 2. It appears therefore that the ecologists model with a separate regression line for each site is appropriate. However, the increasing trend for site 1 is fairly small and hence a model with parallel regressions should also possibly be considered.

(b) In the following, $n_1 = n_2 = 21$

$$E(\mathbf{Y}) = \mathbf{X}\beta$$

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ . \\ . \\ y_{1n_1} \\ y_{21} \\ . \\ . \\ y_{2n_2} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & (x_{11} - \bar{x}_1.) & 0 & 0 \\ . & . & . & . \\ . & . & . & . \\ 1 & (x_{1n_1} - \bar{x}_1.) & 0 & 0 \\ 0 & 0 & 1 & (x_{21} - \bar{x}_2.) \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & 1 & (x_{2n_2} - \bar{x}_2.) \end{pmatrix}$$

$$\beta = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}$$

(c)

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n_1 & \sum_j (x_{1j} - \bar{x}_1.) & 0 & 0 \\ \sum_j (x_{1j} - \bar{x}_1.) & \sum_j (x_{1j} - \bar{x}_1.)^2 & 0 & 0 \\ 0 & 0 & n_2 & \sum_j (x_{2j} - \bar{x}_2.) \\ 0 & 0 & \sum_j (x_{2j} - \bar{x}_2.) & \sum_j (x_{2j} - \bar{x}_2.)^2 \end{pmatrix}$$

since $\sum_j (x_{1j} - \bar{x}_1.) = 0$ and $S_{x_1x_1} = \sum_j (x_{1j} - \bar{x}_1.)^2$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n_1 & 0 & 0 & 0 \\ 0 & S_{x_1x_1} & 0 & 0 \\ 0 & 0 & n_2 & 0 \\ 0 & 0 & 0 & S_{x_2x_2} \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & 0 \\ 0 & \frac{1}{S_{x_1x_1}} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{S_{x_2x_2}} \end{pmatrix}$$

CONTINUED OVERLEAF/

(d)

$$\mathbf{b}^T = (\ 0 \quad 1 \quad 0 \quad -1\ )$$

Therefore, required confidence interval is:

$$\hat{\beta}_1 - \hat{\beta}_2 \pm t(n_1 + n_2 - k; 0.975)\sqrt{\frac{r}{n_1 + n_2 - k}(\frac{1}{S_{x_1x_1}} + \frac{1}{S_{x_2x_2}})}$$

$$\begin{aligned} 0.010 - 0.003 &\pm t(38; 0.975)\sqrt{\frac{0.369946}{38}(\frac{1}{0.2852903} + \frac{1}{0.1942308})} \\ 0.007 &\pm 2.024\sqrt{0.0097(3.505 + 5.149)} \\ 0.007 &\pm 0.586 \\ (-0.579 &, \quad 0.593) \end{aligned}$$

Since this confidence interval includes zero, there is insufficient evidence that separate regression lines are required for this model and a model with parallel regression lines could be considered next.

END OF QUESTION PAPER.