STATS 2107
Statistical Modelling and Inference II
Assignment 4

Matt Ryan

Semester 2 2022

ASSIGNMENT CHECKLIST

- Have you shown all of your working, including probability notation where necessary?
- Have you included all R output and plots to support your answers where necessary?
- Have you included all of your R code?
- Have you made sure that all plots and tables each have a caption?
- Is the submission a single pdf file - correctly orientated, easy to read? If not, penalties apply.
- Assignments emailed instead of submitted by the online submission on Canvas will not be marked and will receive zero.
- Have you checked that the assignment submitted is the correct one, as we cannot accept other submissions after the due date.
- Assignments submitted no later than 24 hours after the deadline will still receive 100% credit. This is in acknowledgement of that fact that people can get sick or have other commitments occur or have internet difficulties. However, assignments can not be submitted more than 24 hours after the deadline.
- Other variations to the assessment on medical/compassionate grounds will only be considered if suitable documentation is provided. In this case you should email Matt Ryan as soon as possible. Please do not ask your tutor for an extension or try to submit an assignment directly to them. Tutors are not able to grant extensions or accept any assignment submissions.

Q1

This question may be typed or hand written and scanned in as a pdf.

The aim of this question is to familiarise yourself with linearising models and dealing with transformations in linear regression.

The saturated growth equation is useful for modelling the growth of some animal species. It can be written as

$$Y = \frac{\alpha x}{\delta + x},$$

where Y is some physical measurements of animal size, x represents time, and α and δ are unknown parameters.

(a) Linearise the saturated growth equation.

[2 marks]

(b) Find $\hat{\alpha}$ and $\hat{\delta}$ based on the least-squares estimate of the parameters of the linearised model.

[2 marks]

(c) Explain why the approach in part (b) is different to fitting the saturated model *directly* using the method of least squares. **Hint:** What is the function we are minimizing in each case?

[3 marks]

(d) A study was conducted in 2002 to investigate the growth of harbour seals. Carcasses of seals that drifted ashore were collected. Their length (in cm) were measured and their age (in years) were estimated based on teeth. The linearised saturated growth model was fitted to the data using R and the following output was obtained. Note that some entries in the outputs in this question are missing and you are to calculate this. The suffix `.trans` is used to indicate that a transformation was applied to the variable.

```
seal.lm <- lm(Length.trans ~ Age.trans)
summary(seal.lm)
```

```
##
## Call:
## lm(formula = Length.trans ~ Age.trans)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.986e-04 -1.731e-04 -2.973e-05  1.127e-04  5.092e-04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.075e-03  9.825e-05  61.830  < 2e-16 ***
## Age.trans   ?????????  2.897e-04   8.423 1.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002495 on 13 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8332
## F-statistic: 70.94 on 1 and 13 DF,  p-value: 1.267e-06
```

i) How many observations are in the data?

[1 mark]

ii) A few observations from the data are given below. For the following observations, write down the response vector $\boldsymbol{y}$ and the design matrix $\boldsymbol{X}$ for the linearised model.

[2 marks]

Length	Age
489	4
476	3
513	7
382	1

iii) Using Part (b) and the R output above, calculate $\hat{\alpha}$ and $\hat{\delta}$ from the saturated growth curve for this data, and give the best fitting saturated growth curve.

[6 marks]

iv) Find the 90% confidence interval for the intercept term in the linearised model.

[4 marks]

v) The following output gives the 90% prediction interval for `Length.trans` of a seal aged 6 years, based on the **linearised model**. Find the corresponding 90% prediction interval for the **length of a seal** aged 6 years.

[5 marks]

```
x0 <- tibble(Age.trans=??)
predict(seal.lm, newdata = x0, interval = "prediction", level = 0.9)
```

```
##          fit         lwr         upr
## 1 0.006481724 0.006023172 ???????????
```

[Question total: 25]

Q2

This question may be typed or hand written and scanned in as a pdf.

The aim of this question is to explore linear transformations of the least squares estimates. You will get to familiarise yourself with the matrix notation approach to multiple linear regression.

Suppose X is an $n \times p$ matrix with linearly independent columns. Let $X^* = XA$, where A is an invertible $p \times p$ matrix.

(a) Show that the columns of X^* are also linearly independent.
[*Hint*: Prove by contradiction, *i.e.*, start by assuming the columns of X^* are **not** linearly independent.]
[3 marks]

(b) Show that $X^*({X^*}^T X^*)^{-1} {X^*}^T = X(X^T X)^{-1} X^T$.
[3 marks]

(c) Consider two alternative models

$$M: \ \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad M^*: \ \boldsymbol{Y} = X^*\boldsymbol{\beta^*} + \boldsymbol{\epsilon}.$$

Let

- $\hat{\boldsymbol{\eta}} = X\hat{\boldsymbol{\beta}}$, and
- $\hat{\boldsymbol{\eta}}^* = X^* \hat{\boldsymbol{\beta}}^*$

be the vectors of fitted values for models M and M^* respectively. Show that $\hat{\boldsymbol{\eta}}^* = \hat{\boldsymbol{\eta}}$, *i.e.*, the vector of fitted values is the same, whatever the form of the design matrix X.

[4 marks]

[Question total: 10]

Q3

THIS QUESTION IF FOR POSTGRADUATE STUDENTS ONLY. This question may be typed or hand written and scanned in as a pdf.

The aim of this question is to explore the concepts of the Box-Cox transformation. In particular, you will show that the transformation is continuous for all λ.

Let Y be a positive random variable, and consider the Box-Cox transformation for normalising data presented in Week 9 given by

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0\,, \\ \log(Y) & \text{if } \lambda = 0\,, \end{cases}$$

where log denotes the natural logarithm. Show that the transformation $Y^{(\lambda)}$ is continuous in λ.

[5 marks]

[Question total: 5]

Q4

Please submit your answer to this question online using MyUni. You will be asked to answer quiz questions and to upload an R script file.

The aim of this question is to explore the concepts on ANCOVA and the three types of regression models associated to it: identical, parallel, and separate. You will also get to whet you skills on performing some analysis on contrasts.

An agriculture company has developed three new variants of fast growing crop, and wishes to investigate their sensitivity to soil salinity. Seeds of each plant type were planted in different sites with varying levels of soil salinity (measured in decisiemens per meter (dS/m)). The height (in 10cm) of the crop was measured at 15 days after they were planted. The data is given in `crop.csv`, available from MyUni. It contains the following variables for 85 crops:

Variable	Details
Crop_no	Unique identifier for each crop
Type	Variant of crop (Type I, II, or III)
Salinity	Salinity level of the soil (dS/m)
Height	Height of the crop (10cm)

(a) Produce a scatterplot of Height against Salinity level by crop type. Describe the relationship.

[3 marks]

(b) Fit the following linear models in R:

- Model 1 – identical regression lines: Height on Salinity level
- Model 2 – parallel regression lines: Height on Type and Salinity level
- Model 3 – separate regression lines: Height on Type and Salinity level with interaction

Consider Model 3 (separate regression lines). Write down the line of best fit for the relationship between Height on Type and Salinity level for each of Type I, Type II, and Type III crop separately.

[3 marks]

(c) Test for a statistically significant interaction term in Model 3 (separate regression lines) at the 5% significance level. Clearly state the null and alternative hypotheses, the test statistic, the P-value, and your conclusion.

[2 marks]

(d) Calculate the BIC for all three models. Which model has the best fit according to BIC?

[2 marks]

(e) Assess the assumptions for Model 3 (separate regression lines) using appropriate diagnostic plots.

[4 marks]

(f) The company is interested in the difference between crop types II and III. Based on Model 3 (separate regression lines), calculate the estimated difference in mean height between Type II and Type III crops (that is, $\hat{\mu}_{\text{Type II}} - \hat{\mu}_{\text{Type III}}$). Then calculate the corresponding 95% confidence interval. [**Hint: Review the lecture on contrasts from Week 8.**]

[3 marks]

(g) Upload your R script with the answers to these questions via the quiz on MyUni.

[3 marks]

[Question total: 20]

[[Assignment total: 60]]