THE UNIVERSITY OF MELBOURNE

Student number

Semester 2 Assignment 3, 2024

School of Mathematics and Statistics

MAST90104 A First Course in Statistical Learning

Submission deadline: 6:00 pm on Friday 4 October 2024

This assignment consists of 5 pages (including this page) with 3 questions and 20 total marks

Instructions to Students

Writing

- Please submit a scanned or other electronic copy of your work via the Learning Management System. Your submission should be a single PDF file. You may submit an additional .R file of your code.
- You can type to handwrite your answer. If you handwrite your answer, write on A4 paper. Write on one side of each sheet only. Follow the instructions below for scanning and submitting of your assignment.
- If you use R, please include the R commands/ output in your answer.
- Avoid only showing the R code and/or output as your answer without any explanation. You may lose marks if your answer is unclear.
- Page 1 should only have your student number, the subject code and the subject name. Each question should be on a new page. The question number must be written at the top of each page.

Scanning and Submitting

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Your scanned assignment must be a single PDF file. Check that all pages are present and readable before submitting.

Blank page

Question 1 (8 marks) The file *fat.csv* contains records of the average butterfat content (percentages) of milk for random samples of twenty cows (ten two-year old and ten mature (greater than four years old)) from each of five breeds. The variables are:

- **Butterfat**: butter fat content by percentage
- **Breed**: a factor with levels "Ayrshire", "Canadian", "Guernsey", "Holstein-Fresian" and "Jersey"
- **Age**: a factor with levels "2year" and "Mature"

Hint: When importing the data to R, you will need to specify `Breed` *and* `Age` *as factor*

(a) Fit a two-way ANOVA model (without interaction) to the data. *Hint*: use `lm`.

(b) According to your model in part (a), is the main effect of Breed significant (controlling for Age)?

(c) From your model output in (a), estimate the difference between the butter fat content of two-year old cows and mature cows (controlling for Breed).

(d) Fit a two-way ANOVA model (with interaction) to the data. Compute a confidence interval for the difference between butterfat content of two-year old Jersey cows and two-year old Guernsey cows. Hint: you may use the `gmodels` library.

(e) Fom your fitted model in (d), test the hypothesis that the butter fat content of Canadian cows and Jersey cows have no dependence on age. What is the null distribution of your test statistic?

(f) Test for the presence of interaction between age and breed of cows.

Question 2 (6 marks) You should not use R's `glm()` function for this question. The data frame anaethestic contains data on the concentrations, `conc`, of anaesthetic given to groups of patients of size, `m`, and the subsequent number of patients that responded satisfactorily to the anaesthetic, `satis`. Here is a table of the data frame:

conc	m	satis
0.8	7	1
1.0	5	1
1.2	6	4
1.4	6	4
1.6	4	4
2.5	2	2

We are interested in building a model for patients' response

(a) Fit a binomial regression model to the data using a logit link.

(b) Construct the 95% CIs for the parameter estimates.

(c) Perform a likelihood ratio test for the significance of the dose coefficient.

(d) Estimate the probability of satisfactory response for a patient who received anaesthetic of concentration 2.0, together with a 95% CI.

Question 3 (6 marks) A school principal is interested in the academic performance of students with varying learning paces and whether their teachers have an impact on their performance. The dataset was analysed with `R` and stored in the principal's computer hard-drive. Unfortunately, a computer virus has corrupted the analysis output, leading to some missing numbers. The corrupted output is as follows:

```
#####BEGIN CORRUPTED OUTPUT###############
> data(ExamScore)
> names(ExamScore)
[1] "Score"      "Teacher"    "LearnSpeed"
> levels(ExamScore$Teacher)
[1] "Bev"    "Sue"    "Winnie"
> levels(ExamScore$LearnSpeed)
[1] "Fast-learner" "Neither"      "Slow-learner"
> AdditiveTwoWay <- lm(Score~Teacher+LearnSpeed,data=ExamScore)
> MultiplicativeTwoWay <- lm(Score~Teacher*LearnSpeed,data=ExamScore)
> summary(AdditiveTwoWay)

Call:
lm(formula = Score ~ Teacher + LearnSpeed, data = ExamScore)

Residuals:
Min     1Q Median     3Q    Max
-9.589 -3.000  0.263  2.926  7.798

Coefficients:
                       Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)             53.0676     0.7305    OUTPUT-A   < 2e-16 ***
TeacherSue               OUTPUT-B   0.8035    5.120      9.56e-07 ***
TeacherWinnie           -3.7878     0.8035    -4.714     5.64e-06 ***
LearnSpeedNeither       -3.3305     OUTPUT-C  -4.145     5.76e-05 ***
LearnSpeedSlow-learner  OUTPUT-D     0.8035   -9.553    < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.017 on OUTPUT-E degrees of freedom

> summary(MultiplicativeTwoWay)

Call:
lm(formula = Score ~ Teacher * LearnSpeed, data = ExamScore)

Residuals:
Min      1Q  Median      3Q     Max
-5.4108 -1.2244 -0.0318  1.2681  5.5556

Coefficients:
                                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                         48.8889     0.4806 101.717   <2e-16 ***
TeacherSue                          11.8896     0.6797  17.492   <2e-16 ***
TeacherWinnie                        1.0091     0.6903   1.462    0.146
LearnSpeedNeither                    0.9607     0.6797   1.413    0.160
```

```
LearnSpeedSlow-learner                  0.8231     0.6903   1.193     0.235
TeacherSue:LearnSpeedNeither          -12.3154     0.9688   -12.713   <2e-16 ***
TeacherWinnie:LearnSpeedNeither        -0.8550     0.9688   -0.883    OUTPUT-F
TeacherSue:LearnSpeedSlow-learner     -11.5258     OUTPUT-G -11.897   <2e-16 ***
TeacherWinnie:LearnSpeedSlow-learner -13.7535     0.9762   -14.089   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: OUTPUT-H on OUTPUT-I degrees of freedom

> anova(AdditiveTwoWay,MultiplicativeTwoWay)
Analysis of Variance Table

Model 1: Score ~ Teacher + LearnSpeed
Model 2: Score ~ Teacher * LearnSpeed
   Res.Df     RSS    Df   Sum of Sq     F     Pr(>F)
1    145   OUTPUT-J
2    141   OUTPUT-K  4    1785.8    OUTPUT-L   < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#####END OF CORRUPTED OUTPUT###############
```

Note that the missing numerical output are: `OUTPUT-A`, `OUTPUT-B`, `OUTPUT-C`, `OUTPUT-D`, `OUTPUT-E`, `OUTPUT-F`, `OUTPUT-G`, `OUTPUT-H`, `OUTPUT-I`, `OUTPUT-J`, `OUTPUT-K`, and `OUTPUT-L`.

Your task: Help the school principal recover all missing numerical output.

End of Assignment — Total Available Marks = 20