MAST90104: A First Course in Statistical Learning

Week 9 Practical and Workshop

1 Practical questions

1. We revisit the milk data last week. We study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

   (a) Input this data into R.

   (b) Test for the presence of interaction.

   (c) What is the degrees of freedom used for the interaction test?

   (d) From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3?

   (e) Find a 95% confidence interval under the interaction model, for the amount of milk produced from breed 2 and diet 3.

2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

   (a) This question use a data set in package `faraway`. Load the package and read the help file (`?pima`) to get a description of the predictor and response variables, then use `pairs` and `summary` to perform simple graphical and numerical summaries of the data.

   There are some obvious irregularities in the data. Take appropriate steps to correct the problems.

   (b) Fit a model with `test` as the response and all the other variables as predictors.

   Odds are sometimes a better scale than probability to represent chance. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}$$

   In a binomial regression model with a logit link we have

$$\text{logit}(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \eta_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_q x_{q,j}.$$

   That is $\log o_j = \eta_j$, where o_j are the odds for the j-th observation.

   (c) By what proportion do the odds of testing positive for diabetes change for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

   (d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

   (e) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a confidence interval for your prediction.

2 Workshop questions

1. Verify that for the binomial regression model with logistic link

$$\begin{aligned}\mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i} &= 0\\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i \partial \theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}\right)\end{aligned}$$

2. A orthopaediecs surgeon is investigating the functioning score of post-knee replacement surgery patients that underwent three different post-surgical intervention under three particpating physiotherapists. The data collected includes the response `FunctionScore` and two factor variables `Intervention` and `Therapist`. Unfortunately, a computer virus corrupted the output file. The corrupted output is as follows:

```
str(DataF)
'data.frame': 20 obs. of  3 variables:
$ FunctionScore: num  12.35 11.82 7.22 13.69 13.05 ...
$ Intervention : Factor w/ 3 levels "1","2","3": 1 1 3 2 2 1 1 1 2 2 ...
$ Therapist    : Factor w/ 3 levels "A","B","C": 3 1 1 3 3 2 2 2 1 1 ...
> summary(imodel)

Call:
lm(formula = FunctionScore ~ Intervention * Therapist, data = DataF)

Residuals:
Min      1Q  Median      3Q     Max
-0.8285 -0.1937  0.0000  0.3317  0.6139

Coefficients:
                        Estimate  Std. Error t value  Pr(>|t|)
(Intercept)              11.7250     0.3467  33.823 1.81e-12 ***
Intervention2            -3.2753     0.4903  -6.681 3.46e-05 ***
Intervention3            -4.1635     0.4475  -9.303 1.51e-06 ***
TherapistB                2.5520     0.4246   6.011 8.79e-05 ***
TherapistC                0.2161     0.4903   0.441    0.668
Intervention2:TherapistB -0.8392     0.7354  -1.141    0.278
Intervention3:TherapistB -0.2600     0.6169  -0.421    0.682
Intervention2:TherapistC  4.9021     0.6638   7.385 1.39e-05 ***
Intervention3:TherapistC  6.1056     0.7489   8.153 5.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: (b) on (a) degrees of freedom
Multiple R-squared:  0.9788,Adjusted R-squared:  0.9634
F-statistic: 63.49 on 8 and (a) DF,  p-value: 4.075e-08

> sum(imodel$residuals^2)
[1] 2.643836

summary(amodel)

Call:
lm(formula = FunctionScore ~ Intervention + Therapist, data = DataF)

Residuals:
Min       1Q   Median       3Q      Max
-2.91203 -0.68993  0.07554  1.02621  2.34695
```

```
Coefficients:
              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)    10.7563     0.7389  14.558 2.96e-10 ***
Intervention2  -1.7656     0.8253  -2.139 0.049253 *
Intervention3  -2.9097     0.8053  -3.613 0.002555 **
TherapistB      2.7526     0.8020   3.432 0.003705 **
TherapistC      3.6897     0.8361   4.413 0.000504 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.455 on (c) degrees of freedom
Multiple R-squared:  0.7455,Adjusted R-squared:  0.6777
F-statistic: 10.99 on 4 and (c) DF,  p-value: 0.0002298

> anova(amodel,imodel)
Analysis of Variance Table

Model 1: FunctionScore ~ Intervention + Therapist
Model 2: FunctionScore ~ Intervention * Therapist
   Res.Df  RSS  Df   Sum of Sq      F    Pr(>F)
1     (c)  (e)
2     (a)  (d)  (f)    (g)         (h)   6.991e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Deduce the missing output (a) to (h).