MAST90104: A First Course in Statistical Learning

Assignment 3, 2023 Solution

1. (16 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 an additive two-factor model to the data.
   Solution

```
butterfat <- read.csv('fat.csv',header = T)
str(butterfat)
butterfat$Breed <- factor(butterfat$Breed)
butterfat$Age <- factor(butterfat$Age)
#a.  additive model
amodel <- lm(Butterfat ~ Breed + Age, data=butterfat)
summary(amodel)
amodel$coefficients
# (Intercept)         BreedCanadian         BreedGuernsey
#      4.0077                0.3785                0.8900
# BreedHolstein-Fresian            BreedJersey            AgeMature
#               -0.3905                 1.2325               0.1046
```

   (b) From this model, estimate the difference between the butter fat content of two-year old cows and mature cows.
   Solution: The difference is the coefficient of `AgeMature`, 0.1046 .

   (c) Fit a linear model with interaction to the data. Calculate a confidence interval for the difference between butterfat content of mature Jersey cows and mature Guernsey cows.
   Solution Model with interactions:

```
imodel <- lm(Butterfat ~ Breed * Age, data=butterfat)
imodel$coefficients
#                   (Intercept)                    BreedCanadian
#                         3.966                            0.522
#                 BreedGuernsey            BreedHolstein-Fresian
#                         0.933                           -0.303
#                   BreedJersey                        AgeMature
#                         1.167                            0.188
#        BreedCanadian:AgeMature          BreedGuernsey:AgeMature
#                        -0.287                           -0.086
# BreedHolstein-Fresian:AgeMature          BreedJersey:AgeMature
#                        -0.175                            0.131
```

   The 95% confidence interval for the difference is $(0.08133698, 0.82066302)$.

```
library(gmodels)
ci <- estimable(imodel,c(0,0,-1,0,1,0,0,-1,0,1),conf.int = 0.95)
c(ci$Lower, ci$Upper)
# [1] 0.08133698 0.82066302
```

(d) Test the hypothesis that the butter fat content of Canadian cows has no dependence on age.
Solution This is the same as testing the hypothesis $H_0 : C\boldsymbol{\beta} = 0$, where

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

for the model with interactions in part (c) .

```
library(car)
C <- rep(0,10)
C[6] = 1; C[7] = 1
linearHypothesis(imodel, C)
# Linear hypothesis test
#
# Hypothesis:
# AgeMature  + BreedCanadian:AgeMature = 0
#
# Model 1: restricted model
# Model 2: Butterfat ~ Breed * Age
#
#   Res.Df     RSS  Df  Sum of Sq       F Pr(>F)
# 1     91  15.629
# 2     90  15.580   1   0.049005  0.2831  0.596
```

The F-test statistic is 0.2831, follows an F distribution with degrees of freedom 1 and 90. The p-value is 0.596, so we do not reject the null hypothesis: there is little evidence that the butter fat content of Canadian cows depends on age.

(e) Test for the presence of interaction between age and breed of cows.
Solution: We use anova to test for interaction by comparing the additive and interaction models.

```
anova(amodel, imodel)
# Analysis of Variance Table
#
# Model 1: Butterfat ~ Breed + Age
# Model 2: Butterfat ~ Breed * Age
#   Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     94 16.094
# 2     90 15.580  4   0.51387 0.7421 0.5658
```

The p-value = 0.5658 suggests that the is not enough evidence for the presence of interactions between the 2 factors (or we don't reject the null hypothesis of no interaction).

2. (12 marks) Depletion of the ozone layer allows the most damaging ultraviolet radiation to reach the Earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable). To measure the relationship, researchers sampled the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. The data from this study is given in the file `ozone.csv`. There are 3 variables:

- **Inhibit**: percent inhibition of primary phytoplankton production in water
- **UVB**: UVB exposure
- **Surface**: a factor with levels "Deep" and "Surface"

Hint: When importing the data to R, you will need to specify `Surface` as factor

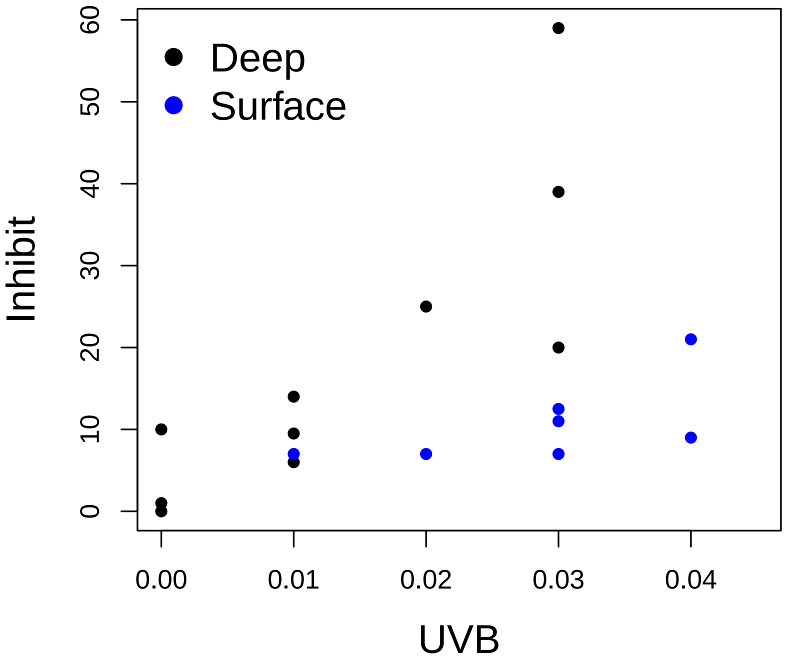

Figure 1: Percentage inhibition versus UVB exposure

(a) Plot the percentage inhibition against UVB exposure, use different colours for observations at the surface and in the deep. Is there any evidence of an interaction?

Solution

See Figure 1 for reference

From the plot, it seems that the effect of UVB on inhibition at the surface is different to the effect in the deep.

For part b and c, you should not use `lm()` function.

(b) Using matrix calculation, fit a model for percentage inhibition with interaction between UVB exposure and depth level. You should use treatment contrasts in this question.

Solution The reparameterised model is of the form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where X is

```
Xtreat
#       [,1] [,2] [,3] [,4]
# [1,]     1    0 0.00 0.00
# [2,]     1    0 0.00 0.00
# [3,]     1    0 0.01 0.00
# [4,]     1    1 0.01 0.01
# [5,]     1    1 0.02 0.02
# [6,]     1    1 0.03 0.03
# [7,]     1    1 0.04 0.04
# [8,]     1    0 0.01 0.00
# [9,]     1    0 0.00 0.00
# [10,]    1    1 0.03 0.03
# [11,]    1    1 0.03 0.03
# [12,]    1    0 0.01 0.00
# [13,]    1    0 0.03 0.00
# [14,]    1    1 0.04 0.04
# [15,]    1    0 0.02 0.00
# [16,]    1    0 0.03 0.00
# [17,]    1    0 0.03 0.00
```

Then using the formula of the least square estimator $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}$, we got the result:

```
y <- ozone$Inhibit
(b <- solve(t(Xtreat)%*%Xtreat,t(Xtreat)%*%y))
#             [,1]
# [1,]    1.180556
# [2,]    1.277778
# [3,] 1226.388889
# [4,] -939.930556
```

(c) Estimate the percent inhibition of primary phytoplankton production in water below the surface when UVB exposure is 0.02.

Solution Once we got $\mathbf{b}$, the the prediction is $(\mathbf{x}^*)^T\mathbf{b}$

```
c(1,0,0.02,0)%*%b
#          [,1]
# [1,] 25.70833
```

The estimated percentage of inhibition is 25.70833.

3. (12 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.

Solution: Following the lecture note of Module 7, we can fit the binomial regression with logit link as follow

```
conc = c(0.8,1,1.2,1.4,1.6,2.5)
m = c(7,5,6,6,4,2)
satis = c(1,1,4,4,4,2)
anaethestic = data.frame(conc = conc,m = m,satis = satis)
# binomial model - logit link
l <- function(tha, y,n, conc) {
eta <- tha[1] + tha[2]*conc
return(sum(y*eta - n*log(1 + exp(eta))))
}
(betahat <- optim(c(0, 1), l, y = anaethestic$satis,
conc = anaethestic$conc,n = anaethestic$m,
control = list(fnscale = -1,reltol=1e-16))$par)
# [1] -6.468675  5.566762
```

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

Solution We first need to compute $\boldsymbol{\mathcal{I}}(\beta)^{-1}$

```
ilogit <- function(x) 1/(1+exp(-x))
#calculate estimated probabilities from parameters
phat = ilogit(betahat[1]+betahat[2]*anaethestic$conc)

I11 <- sum(anaethestic$m*phat*(1 - phat))
I12 <- sum(anaethestic$m*anaethestic$conc*phat*(1 - phat))
I22 <- sum(anaethestic$m*anaethestic$conc^2*phat*(1 - phat))
(Iinv <- solve(matrix(c(I11, I12, I12, I22), 2, 2)))
#           [,1]      [,2]
# [1,]  5.849531 -4.848630
# [2,] -4.848630  4.176661
```

The diagonal elements of $\boldsymbol{\mathcal{I}}(\beta)^{-1}$ are the variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Then the CIs are $(-11.209092, -1.728257)$ and $(1.561133, 9.572390)$ respectively.

```
# estimate of SD
(sdp <- c(sqrt(Iinv[1,1]),sqrt(Iinv[2,2])))
# [1] 2.418580 2.043688
# 95% ci
betahat[1] + c(-1,1)*1.96*sdp[1]
# [1] -11.209092  -1.728257
betahat[2] + c(-1,1)*1.96*sdp[2]
# [1] 1.561133 9.572390
```

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

Solution The (scaled) deviance of the model with `conc` is 1.7321, degrees of freedom is 4. The deviance of the model without `conc` (only intercept) is 15.43341, and the degrees of freedom is 5.

The LR test statistic is $D^A - D^B = 15.43341 - 1.7321 = 13.70127$, follows a χ^2 distribution with d.f 1. The p-value is $0.000214 < 0.01$, so we reject the smaller model: The effect of dose/concentration is significant.

One can also compare the test statistic with the critical value of χ^2_1 instead of computing p-value.

```
# c. likelihood ratio test
y <- anaethestic$satis
n <- anaethestic$m
ylogxy <- function(x, y) ifelse(y == 0, 0, y*log(x/y))

# deviance of model
(D <- -2*sum(ylogxy(n*phat, y)+ ylogxy(n*(1-phat), n - y)))
# [1] 1.732136
(df <- length(y) - length(betahat))
# [1] 4

# deviance of null model (with only intercept)
phatN <- sum(y)/sum(n)
(DN <- -2*sum(ylogxy(n*phatN, y) + ylogxy(n*(1-phatN), n - y)))
# [1] 15.43341
(DfN <- length(y) - 1)
# [1] 5
# p-value
pchisq(DN - D, DfN - df, lower=FALSE)
# [1] 0.0002143095
```

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

Solution The estimated probability of satisfactory response is $0.9906672 \approx 0.991$.

```
dosenew = 2.0
si2 <- matrix(c(1, dosenew), 1, 2) %*% Iinv %*% matrix(c(1, dosenew), 2, 1)
(p2 <- ilogit(betahat[1] + dosenew*betahat[2]))
#[1] 0.9906672
```

To compute the CI, we first compute the CI for $\eta_i = \beta_0 + \beta_1 x_i$, then take the inverse logit of the lower and upper bounds. The CI of the probability is $(0.7649, 0.9997)$.

```
# lower bound
ilogit(betahat[1] + dosenew*betahat[2] - qnorm(0.975)*sqrt(si2))
#          [,1]
# [1,] 0.764917
# upper bound
ilogit(betahat[1] + dosenew*betahat[2] + qnorm(0.975)*sqrt(si2))
#           [,1]
# [1,] 0.9997113
```