MAST90104: A First Course in Statistical Learning

Week 9 Practical and Workshop Solution

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.

   Solution

```
>milk <- data.frame(milk=c(18.8,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8),
diet=factor(c(1,1,2,3,3,1,2,2,3)),
breed=factor(c(1,1,1,1,1,2,2,2,2)))
> with(milk, interaction.plot(diet, breed, milk))
```

   (b) Test for the presence of interaction.

```
> imodel <- lm(milk ~ breed * diet, data=milk)
> anova(imodel)
Analysis of Variance Table

Response: milk
Df Sum Sq Mean Sq F value Pr(>F)
breed       1  0.174  0.1742  0.0312 0.8710
diet        2 36.204 18.1018  3.2460 0.1777
breed:diet  2  1.874  0.9372  0.1681 0.8527
Residuals   3 16.730  5.5767
```

   There is clearly no evidence of interaction.

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

   Solution: The degrees of freedom used are 2 and 3.

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

   Solution:

```
> imodel$coefficients
(Intercept)       breed2        diet2        diet3
20.00         2.30        -3.30         1.85
breed2:diet2 breed2:diet3
-1.45        -2.35
> c(1,1,0,1,0,1)%*%imodel$coefficients
[,1]
[1,] 21.8
```

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

   Solution

```
> estimable(imodel, c(1,1,0,1,0,1), conf.int=0.95)
Estimate Std. Error  t value DF    Pr(>|t|) Lower.CI Upper.CI
(1 1 0 1 0 1)     21.8   2.361497 9.231434  3 0.002689148 14.28466 29.31534
```

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) 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.
   Solution: It is clear that there are missing observations for many variables, which have been recorded as zeros. The easiest (not necessarily the only or best) way to deal with these is to remove the relevant observations from the data set. On the other hand, 0 is a plausible value for insulin, diabetes and test so these 0's are not excluded.

```
> library(faraway)
> data(pima)
> View(pima)
> missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
> pima <- pima[!missing,]
```

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

```
> model <- glm(cbind(test, 1-test)~., family=binomial, data=pima)
> summary(model)

Call:
glm(formula = cbind(test, 1 - test) ~ ., family = binomial, data = pima)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.677562   1.005400  -9.626  < 2e-16 ***
pregnant     0.121235   0.043926   2.760 0.005780 **
glucose      0.037439   0.004765   7.857 3.92e-15 ***
diastolic   -0.009316   0.010446  -0.892 0.372494
triceps      0.006341   0.014853   0.427 0.669426
insulin     -0.001053   0.001007  -1.046 0.295651
bmi          0.085992   0.023661   3.634 0.000279 ***
diabetes     1.335764   0.365771   3.652 0.000260 ***
age          0.026430   0.013962   1.893 0.058371 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 465.23  on 523  degrees of freedom
AIC: 483.23

Number of Fisher Scoring iterations: 5
```

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.

Solution:

```
> quantile(pima$bmi)
0%    25%    50%    75%   100%
18.200 27.875 32.800 36.900 67.100
```

For $i = 1, 3$, let o_i, p_i, η_i be the odds, probability and linear response for a woman with `bmi` at the first and third quartiles respectively (27.87 and 36.90). We have

$$\begin{aligned} \frac{o_1}{o_3} &= \exp(\log(o_1/o_3)) \\ &= \exp(\eta_1 - \eta_3) \\ &= \exp(\beta_{bmi}(27.87 - 36.90)) \end{aligned}$$

A point estimate and 95% CI for $\beta_{bmi}(27.87 - 36.90)$ are

$$-9.03(0.085992 \pm 2 \times 0.023661) = -0.7765078 \pm 0.4273177.$$

Transforming this to the odds scale, we get an odds ratio of $e^{-0.7765078} = 0.4600097$ with 95% CI $(0.3000443, 0.705259)$. So, roughly speaking, all else being equal, the odds of showing evidence of diabetes are between 29 to 70 percent less for a woman with bmi 27.87 compared to a woman with bmi 36.90.

(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.

Solution

First compute the sample correlation of the predictors

```
> # d
> round(cor(pima),3)
          pregnant glucose diastolic triceps insulin   bmi diabetes   age  test
pregnant     1.000   0.125     0.205   0.095  -0.007 0.009    0.007 0.641 0.253
glucose      0.125   1.000     0.219   0.227   0.460 0.247    0.166 0.279 0.504
diastolic    0.205   0.219     1.000   0.226   0.007 0.307    0.008 0.347 0.183
triceps      0.095   0.227     0.226   1.000   0.126 0.647    0.119 0.161 0.255
insulin     -0.007   0.460     0.007   0.126   1.000 0.191    0.152 0.081 0.212
bmi          0.009   0.247     0.307   0.647   0.191 1.000    0.151 0.073 0.301
diabetes     0.007   0.166     0.008   0.119   0.152 0.151    1.000 0.072 0.233
age          0.641   0.279     0.347   0.161   0.081 0.073    0.072 1.000 0.315
test         0.253   0.504     0.183   0.255   0.212 0.301    0.233 0.315 1.000
```

Recall that `model` uses all predictors

```
> summary(model)

Call:
glm(formula = cbind(test, 1 - test) ~ ., family = binomial, data = pima)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.86271  -0.66387  -0.36716   0.63466   2.49423

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -9.6775617  1.0054003 -9.6256 < 2.2e-16 ***
```

```
pregnant     0.1212346  0.0439258  2.7600 0.0057804 **
glucose      0.0374387  0.0047648  7.8574 3.922e-15 ***
diastolic   -0.0093162  0.0104463 -0.8918 0.3724941
triceps      0.0063413  0.0148532  0.4269 0.6694264
insulin     -0.0010529  0.0010068 -1.0458 0.2956512
bmi          0.0859919  0.0236610  3.6343 0.0002787 ***
diabetes     1.3357638  0.3657712  3.6519 0.0002603 ***
age          0.0264297  0.0139625  1.8929 0.0583707 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.788  on 531  degrees of freedom
Residual deviance: 465.230  on 523  degrees of freedom
AIC: 483.23

Number of Fisher Scoring iterations: 5
```

`diastolic` is not significant in the presence of the other variables. There is positive correlation between diastolic and test, yet in the model `diastolic` has a negative coefficient. This is possible because `diastolic` is correlated with other (more significant) variables: the test is more likely to be positive when diastolic is large, but this is because glucose, triceps, bmi and age are all more likely to be large, and these all have the effect of increasing the chance of a positive test.

(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.

Solution

```
> x <- predict(model, newdata = list(pregnant=1, glucose=99, diastolic = 64, triceps = 22,
insulin = 76, bmi=27, diabetes=.25, age=25), type="link", se.fit=TRUE)
> ilogit(c(x$fit-2*x$se.fit, x$fit, x$fit+2*x$se.fit))
1          1          1
0.02632817 0.04435407 0.07378636
```