MAS programmes - Statistical Analysis (Spring Term)
SPRING 2020

Assignment

Deadline: 30th April, 2020

Total marks: [40]. Marks are shown in boxes [ ]. There are 2 questions in this assignment.

1. In a study on infant respiratory disease, data are collected on a sample of 2074 infants. The information collected includes whether or not each infant developed a respiratory disease in the first year of their life; the gender of each infant; and details on how they were fed as one of three categories (breast-fed, bottle-fed and supplement).

   A model is fitted to the data in R using the following command:

```
fit <- glm(disease/(disease + nondisease) ~ gender + food,
family = binomial, weights = disease + nondisease, data=babyfood)
```

   Parts of the commands `summary(fit)` and `anova(fit)` are summarized below.

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6127 0.1124 -14.347 < 2e-16 ***
genderGirl  -0.3126 0.1410 -2.216 0.0267 *
foodBreast  -0.6693 0.1530 -4.374 1.22e-05 ***
foodSuppl   -0.1725 0.2056 -0.839 0.4013
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 26.37529 on 5 degrees of freedom
Residual deviance: 0.72192 on 2 degrees of freedom
AIC: 40.24

Number of Fisher Scoring iterations: 4

       Df Deviance Resid. Df Resid. Dev
NULL                       5    26.3753
gender  1   5.4761         4    20.8992
food    2  20.1772         2     0.7219
```

(a) State the model that has been fitted. [3]

(b) Interpret the significance of the parameters of each explanatory variable. [3]

(c) Does the model in (a) provide a good fit? Justify your answer. [2]

(d) Using R syntax, show how you would calculate the p-value. [2]

(e) Interpret the effect of breast feading and derive a 95% confidence interval for it. [3]

(f) Estimate an unbiased estimator of the scale parameter. [2]

(g) Let fit2 and fit3 the models fitted by the following R commands:

```
> fit2 <- glm(disease/(disease + nondisease) ~ gender + food,
family = binomial(link="identity"), weights = disease + nondisease)

> fit3 <- glm(disease/(disease + nondisease) ~ gender + food,
family = binomial(link="probit"), weights = disease + nondisease)
```

Which model seems the most reasonable to endorse? Justify your answer. [2]

(h) Let D be the deviance on d degrees of freedom of the model fitted by the following R command:

```
> fit4 <- glm(disease/(disease + nondisease) ~ gender*food,
family = binomial, weights = disease + nondisease)
```

Which are the numerical values of D and d? Justify your answer. [3]

2. (a) The classification rule that minimizes the total expected misclassification cost leads to the following form:

"Allocate the individual with observation $\mathbf{x}$ to population 1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geqslant \frac{\pi_2 C(1|2)}{\pi_1 C(2|1)}$$

otherwise allocate it to population 2."

Let us now suppose that observation vectors from the i-th population have the $N_p(\mu_i, \Sigma)$ distribution, $i = 1, 2$.

Derive a simplified version of the above rule. [10]

(b) A random sample of 49 old men participating in a study of aging were classified by psychiatric examination into one of two categories: senile or non-senile. An independently administered adult intelligence test revealed large differences between the two groups in certain subsets of the test. The group means are as follows:

Subtest	Senile ($n_1 = 37$)	Non-senile ($n_2 = 12$)
Information	12.57	8.75
Similarities	9.57	5.35
Arithmetic	11.49	8.50
Picture completion	7.97	4.75

We want to construct a rule that would allow us to distinguish between the two categories on the basis of the results of the tests.

Using the R output below calculate the the squared Mahalanobis distance between the samples and then using R syntax, show how you would calculate the probability of misclassification. [4]

```
> Sp
        [,1]    [,2]    [,3]    [,4]
[1,] 11.2553  9.4042  7.1489  3.3830
[2,]  9.4042 13.5318  7.3830  2.5532
[3,]  7.1489  7.3830 11.5744  2.6170
[4,]  3.3830  2.5532  2.6170  5.8085

> invSp <- round(solve(Sp),4)
invSp
        [,1]    [,2]    [,3]    [,4]
[1,]  0.2591 -0.1358 -0.0588 -0.0647
[2,] -0.1358  0.1865 -0.0383 -0.0144
[3,] -0.0588 -0.0383  0.1510 -0.0170
[4,] -0.0647 -0.0144 -0.0170  0.2112
```

(c) 100 individuals were each given 3 scores based on a series of tests designed to measure an underlying attitude. The unbiased sample covariance matrix S_U is given by:

$$S_U = \begin{pmatrix} 197 & & \\ 88 & 52 & \\ 109 & 36 & 73 \end{pmatrix}$$

with eigenvalues and eigenvectors given by:

	Eigenvectors		
	1	2	3
	.815	.055	-.577
	.360	.733	.577
	.455	-.678	.577
Eigenvalues	296.724	25.276	0

i) Explain why this shows that one variable is redundant. [1]

It was decided to omit the first variable and carry out a principal component analysis on the remaining two variables.

ii) Can you suggest why? Without carrying out any further detailed computation, roughly what do you expect the two new principal components to look like? Explain your answer and briefly discuss possible interpretations and uses for these components. [5]