Statistical Methods for Big Data

Exercise Sheet 6

1. In a survey carried out in 1974 each respondent was asked if he or she agreed or disagreed with the statement:

"Women should take care of running their homes and leave running the country up to men".

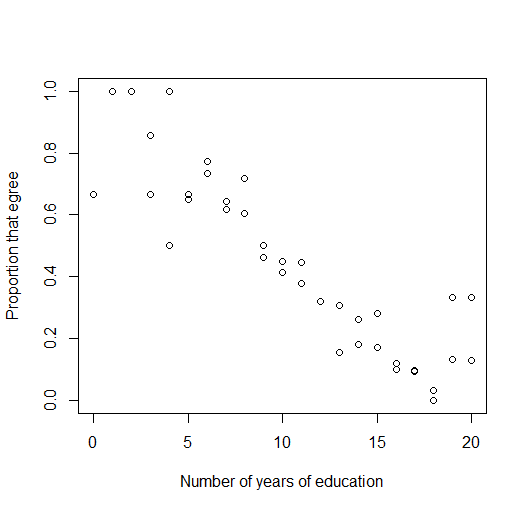
For each respondent, the response "agree" or "disagree" was recorded, along with the number of years’ education received by the respondent. The data is summarised in Table 1 below.

To find out whether years’ of education affect the response to the question, logistic regression analysis was performed on the data and the output obtained is shown in Figure 1 and Table 2.

**Table 1**

|  |  |  |
| --- | --- | --- |
| **Education** | **Agree** | **Disagree** |
| 0 | 8 | 4 |
| 1 | 3 | 0 |
| 2 | 4 | 0 |
| 3 | 12 | 4 |
| 4 | 15 | 5 |
| 5 | 27 | 14 |
| 6 | 42 | 14 |
| 7 | 53 | 31 |
| 8 | 166 | 85 |
| 9 | 59 | 64 |
| 10 | 87 | 112 |
| 11 | 86 | 121 |
| 12 | 305 | 648 |
| 13 | 48 | 162 |
| 14 | 46 | 7981 |
| 15 | 16 | 57 |
| 16 | 28 | 225 |
| 17 | 6 | 57 |
| 18 | 1 | 49 |
| 19 | 3 | 15 |
| 20 | 5 | 24 |

**Figure 1**



**Table 2**

|  |
| --- |
| Deviance Residuals:  Min 1Q Median 3Q Max  -2.74737 -0.88358 -0.07487 0.86240 3.10504  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 2.50334 0.17843 14.03 <2e-16 \*\*\*  education -0.27065 0.01541 -17.56 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 451.722 on 40 degrees of freedom  Residual deviance: 64.025 on 39 degrees of freedom  AIC: 206.09  Number of Fisher Scoring iterations: 4 |

1. What are the null and alternative hypotheses for this question?
2. Figure 1 shows a plot of years’ education on the x-axis and p on the y-axis. What does p represent?

(c) Does the model indicate that there is a relationship between the response to the question and the years of education of the respondent? Explain why you think this.

(d) Write down the logistic model that has been fitted to this data.

(e) Use the equation from part (d) to estimate the proportion of respondents with twelve years of education that agreed with the question.

(f) Use the residual deviance to estimate the goodness of fit of this model. Does the model fit the data adequately?

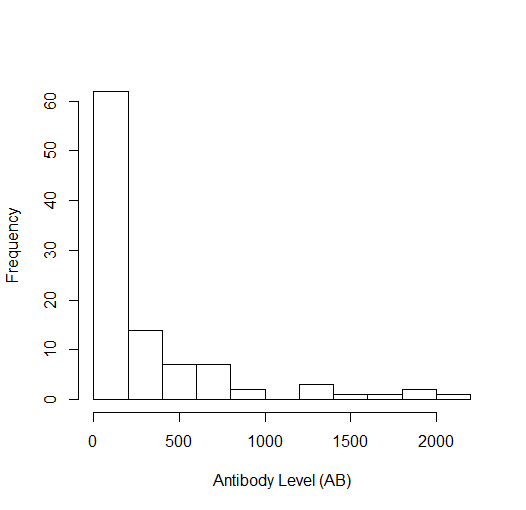
1. In a study of the disease malaria, a random sample of 100 children from a village in Ghana, aged 3-15 years, was followed for a period of 8 months. At the beginning of the study, values of a particular antibody (**AB**) and age (**Age**) were recorded for each child. At the end of the 8 months, the presence (1) or absence (0) of symptoms of malaria (**Mal**) was recorded. The first 8 observations of the data set are shown in Table 1.

**Table 1**

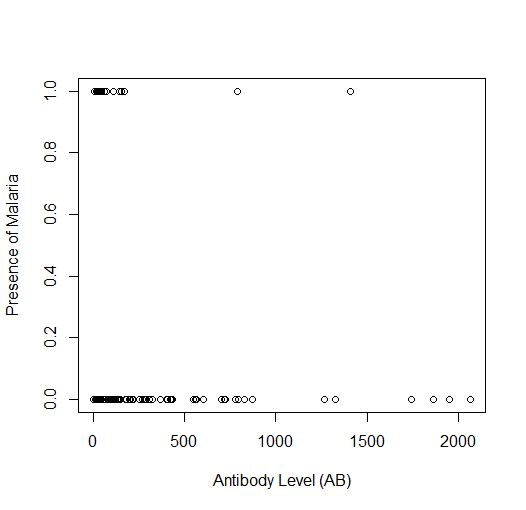
|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Age** | **AB** | **Mal** |
| 1 | 15 | 546 | 0 |
| 2 | 14 | 268 | 0 |
| 3 | 12 | 284 | 0 |
| 4 | 15 | 38 | 0 |
| 5 | 14 | 827 | 0 |
| 6 | 12 | 252 | 0 |
| 7 | 12 | 24 | 1 |
| 8 | 13 | 1740 | 0 |

To find out whether the age and antibody level of a child will affect the likelihood that the child will show symptoms of malaria, some preliminary data analysis was carried out followed by logistic regression analysis. The output obtained is shown in Figures 1-3 and Tables 2-3.

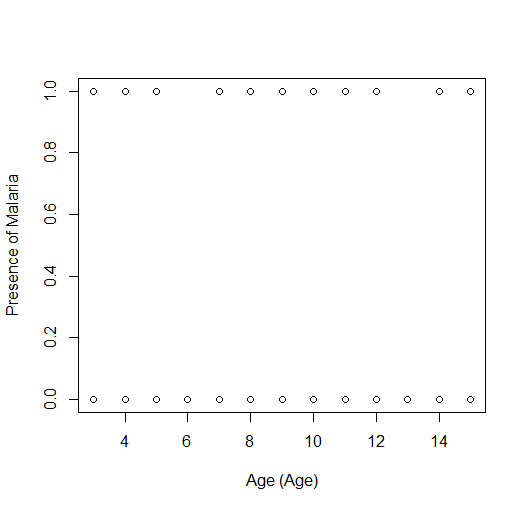
**Figure 1**



**Figure 2**



**Figure 3**



**Table 2: Output from Model 1**

|  |
| --- |
| Call:  glm(formula = Mal ~ log(AB) + Age, family = binomial, data = malaria)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.8492 -0.7536 -0.4838 0.8809 2.5796  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 2.57234 0.95184 2.702 0.006883 \*\*  log(AB) -0.68235 0.19552 -3.490 0.000483 \*\*\*  Age -0.06546 0.06772 -0.967 0.333703  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 116.652 on 99 degrees of freedom  Residual deviance: 98.017 on 97 degrees of freedom |

**Table 3 Output from Model 2**

|  |
| --- |
| Call:  glm(formula = Mal ~ log(AB), family = binomial, data = malaria)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.9159 -0.7339 -0.4854 0.8813 2.4722  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 2.1552 0.8401 2.565 0.010305 \*  log(AB) -0.7122 0.1932 -3.686 0.000228 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 116.652 on 99 degrees of freedom  Residual deviance: 98.968 on 98 degrees of freedom  AIC: 102.97  Number of Fisher Scoring iterations: 4 |

* 1. With reference to the preliminary data analysis, explain why the models specified in Tables 2 and 3 use the log transformed antibody level as an explanatory variable.
  2. Does the output from Model 1 (shown in Table 2) indicate that there is a relationship between the presence of malaria and the antibody level? Explain why you think this.
  3. Use the residual deviance to decide whether Model 1 or Model 2 should be selected.
  4. Write down the equation for the logistic model you selected in part (d).
  5. Use the residual deviance to compare Model 2 with the Null Model. What do you conclude?

1. An important issue in archaeology is the analysis of soil samples to determine their origin and age. It was proposed that the concentrations of trace elements found in soil samples vary significantly depending on the history of the soil. Samples taken from areas of undisturbed soil may exhibit different patterns of trace element composition to samples taken from areas where the soil has been modified by cultivation or dwelling.

A study was carried out on 22 soil samples where the concentrations of 9 trace elements were measured for each sample. The area that the soil sample was taken from was recorded as one of three categories:

Category A – on-site, specific feature e.g. hearth

Category B – on-site, general

Category C – off-site

The categories represent three levels of historical human activity.

A subset of the soil sample data is shown in the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Fe** | **Cu** | **P** | **Mn** | **V** | **Co** | **Zn** | **Cr** | **Ca** |
| A | 2.21 | 36 | 0.914 | 950 | 50 | 7 | 130 | 34 | 2.26 |
| B | 1.98 | 27 | 0.337 | 786 | 45 | 7 | 110 | 29 | 1.38 |
| C | 2.08 | 64 | 1.436 | 1998 | 45 | 7 | 352 | 29 | 3.33 |

A Principal Components Analysis was carried out on the samples to determine whether the soil samples from the three different categories can be distinguished from one another and which variables are important for determining this. The analysis was carried out using **R**, and the output from the analysis is presented on the following pages.

**Table 1**: Summary of model

Importance of components:

PC1 PC2 PC3 PC4 PC5 Standard deviation 2.1115 2.0150 0.46992 0.39963 0.20699

Proportion of Variance 0.4954 0.4511 0.02454 0.01775 0.00476

Cumulative Proportion 0.4954 0.9465 0.97105 0.98879 0.99355

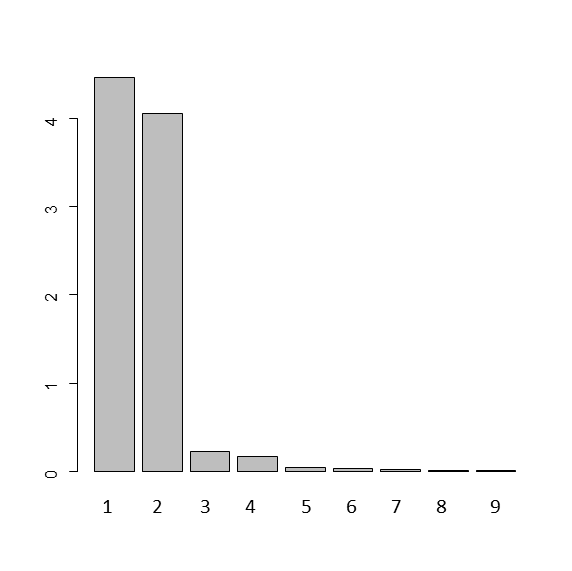
PC6 PC7 PC8 PC9

Standard deviation 0.16886 0.12487 0.10097 0.06114

Proportion of Variance 0.00317 0.00173 0.001130.00042

Cumulative Proportion 0.99672 0.99845 0.99958 1.00000

**Figure 1.** Scree plot from the PCA of the soil samples data set.



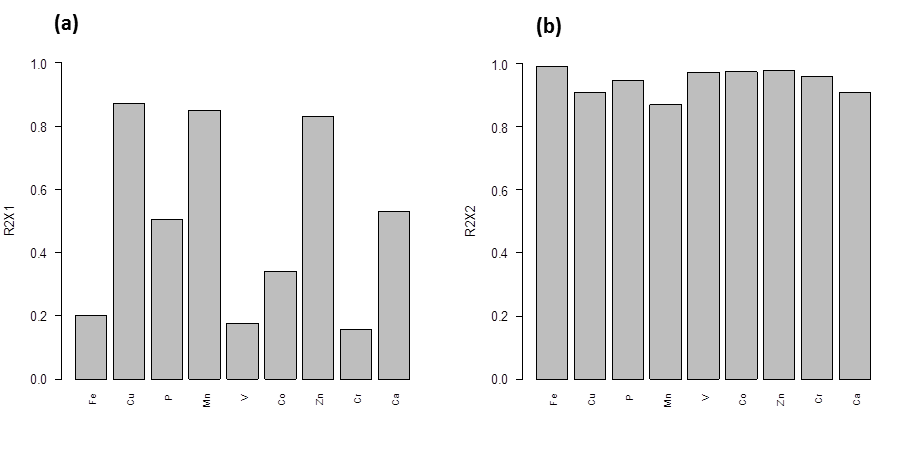
**Figure 2.** Scores plot for the soil samples



**Figure 3.** Loading plot for the soil samples



**Figures 4 (a) and (b).** Barcharts showing the fraction of the explained variation of the variables of the soil samples data set for a PCA model with 4(a) 1 principal component and 4(b) 2 principal components.



1. How many principal components are needed if you wish to account for 95% of variation in the data (please round the output data to 2 decimal places)?
2. Based on the output of the PCA, is it possible to distinguish between the soil samples? Please explain your answer, referencing the output you used to reach your decision.
3. Comment on the scores plot shown in Figure 2.
4. Which category of soil samples has comparatively high levels of the trace elements P, Ca, Zn, Cu and Mn?
5. Comment on the composition of trace elements found in soil samples that were subject to a high level of human activity in the past.
6. Which variables (trace elements) are well explained by the second principal component?
7. The erythrocyte sedimentation rate (ERS) is the rate at which red blood cells (erythrocytes settle out of suspension in blood plasma, when measured under standard conditions. Certain medical conditions such as rheumatic diseases, chronic infections and malignant diseases cause the levels of certain proteins (fibrinogen and globulin) in the blood plasma rise which, in turn, causes the ESR to increase.

To assess whether measuring the levels of the ESR is a useful diagnostic tool, a study collected data on the levels of the proteins fibrinogen and globulin along with the corresponding ESR level.

The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

Download the plasma data set from blackboard.

The ESR variable is a binary variable where “1” indicates an ESR value greater than 20mm/hr and “0” indicates an ESR value less than 20mm/hr.

* 1. Plot the variable ESR as a function of the variable fibrinogen
  2. Fit a logistic regression model to the data, where ESR is the response variable and fibrinogen is the explanatory variable.
  3. Write down the equation for the logistic model that has been fitted to this data.
  4. Sketch (on paper) the shape of the logistic regression curve that fits this data
  5. How will a unit increase in fibrinogen affect the odds of the ESR being greater than 20mm/hr?
  6. Use the equation from part (c) to estimate the probability that the ESR is greater than 20mm/hr if the fibrinogen level is 3.5
  7. Plot the logistic curve onto the data plotted in part (a)
  8. Use the residual deviance to evaluate the goodness of fit for this model.
  9. Use the residual deviance and the null deviance to evaluate whether the variable fibrinogen is explaining a significant amount of the variation in this data.
  10. Repeat the above with the globulin explanatory variable.
  11. Create a model with fibrinogen and globulin as the explanatory variables and use a model selection process to find the best model.