SMM634 - Group Assignment

(worth 50% of final grading)

Deadline 9 December 2022

1. The data frame `Visits.txt` contains data originating from the 1977–1978 Australian Health Survey. Interest focuses on exploring the effect that several variables may have on the number of doctor visits in past 2 weeks, `visits`. Of particular interest are the variables:

- `gender`, factor indicating gender.
- `age`, age in years divided by 100.
- `income`, annual income in tens of thousands of dollars.
- `illness`, number of illnesses in past 2 weeks.
- `reduced`, number of days of reduced activity in past 2 weeks due to illness or injury.
- `health`, general health questionnaire score using Goldberg's method.
- `private`, factor. Does the individual have private health insurance?
- `freepoor`, factor. Does the individual have free government health insurance due to low income?
- `freerepat`, factor. Does the individual have free government health insurance due to old age, disability or veteran status?
- `nchronic`, factor. Is there a chronic condition not limiting activity?
- `lchronic`, factor. Is there a chronic condition limiting activity?

The aim of the analysis is to model number of doctor visits as a function of the variables described above using approaches seen in the second half of this term.

Using these data, write a report addressing the following points:

(a) Each group member must specify a regression model and explain the reason for choosing that model. [5]

(b) Provide summaries of the fitted models (e.g., using tables and figures), interpret the empirical findings and compare them. [10]

(c) Discuss pros and cons of your analyses and which model specification is more suitable for analysing these data. [10]

Other 5 marks will be for overall report structure (e.g., report neatness, presentation style, extra effort).

The report must be in pdf format. It (excluding the title page) must not be longer than 6 pages (including graphs, tables, etc.) using font size 12pt with one and a half line spacing and at least 2.5 cm margin.

[Total marks: 30]

2. In the Irish education system, the Leaving Certificate is the final examination for high school students and is usually taken between the ages of 16 and 19 years. In a study to investigate educational opportunities in Ireland, data were gathered on 441 school leavers. Variables recorded included:

`Gender`: Male or female, taking values `GenderMale` or `GenderFemale`;

`DVRT`: Drumcondra Verbal Reasoning Test score (a measure of verbal reasoning ability, measured on a continuous scale);

`Prestige`: A measure of the father's occupational prestige (continuous-valued with higher values indicating more prestigious occupations);

`SchoolType`: Secondary school or vocational school taking values `SchoolTypeSecondary` and `SchoolTypeVocational`;

`CertTaken`: Whether or not the Leaving Certificate was taken (0=no, 1=yes).

The following (edited) R output is from an analysis of these data in which logistic regression was used to explore the social and academic factors associated with taking the Leaving Certificate:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.486916   0.923494  -8.107 5.18e-16 ***
DVRT          0.055578   0.008306   6.692 2.21e-11 ***
GenderFemale  0.496952   0.217793   2.282   0.0225 *
Prestige      0.037782   0.007593   4.976 6.50e-07 ***
---
(Dispersion parameter for binomial family taken to be 1)
```

Note: Your answers should be concise and to the point.

(a) Write down the statistical representation of the model summarised by the output above. Give the estimated values of any model parameters. [4]

(b) How do you interpret the coefficient labelled `GenderFemale`? How do you interpret the associated *p*-value? [2]

(c) According to the model above, what is the probability of taking a Leaving Certificate for a boy with a DVRT score of 100 and whose father has a prestige score of 40 (which, incidentally, are both close to the average values in the data set)? [2]

(d) The model above was extended by adding the `SchoolType` variable. The result was as follows:

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.571088   1.028429  -4.445 8.80e-06 ***
DVRT                  0.040374   0.009259   4.361 1.30e-05 ***
GenderFemale         -0.004924   0.252142  -0.020  0.98442
Prestige              0.025694   0.008456   3.039  0.00238 **
SchoolTypeVocational -3.149843   0.421097  -7.480 7.43e-14 ***
---
```

What does this new model tell us about the roles of gender and school type in determining whether or not a student takes the Leaving Certificate? [4]

(e) The following analysis of deviance table was used to compare three models in R (you aren't told explicitly what the models are, but the table gives you all the information you need to figure it out):

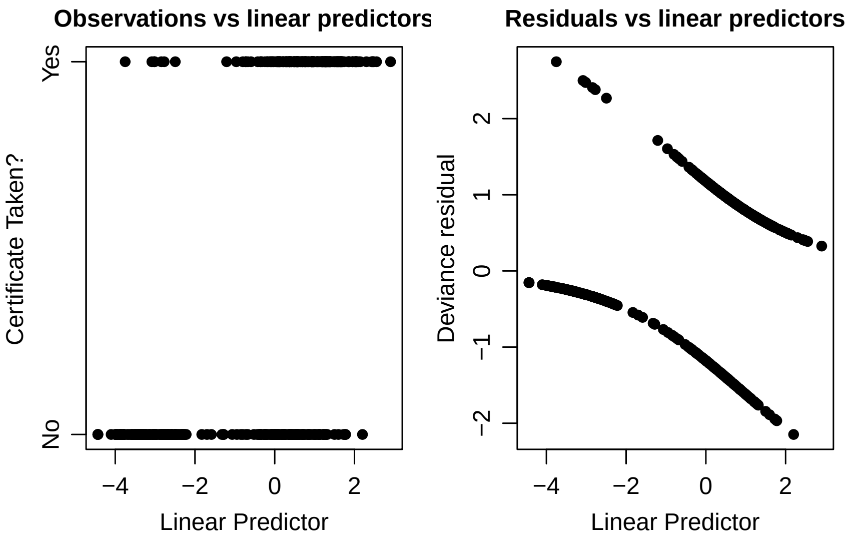
```
Analysis of Deviance Table

Model 1: CertTaken ~ DVRT
Model 2: CertTaken ~ DVRT + Gender + Prestige
Model 3: CertTaken ~ DVRT + Gender + Prestige + SchoolType

    Resid. Df  Resid. Dev  Df  Deviance  P(>|Chi|)
1       439     545.24
2       437     515.01     2   30.227  2.731e-07 ***
3       436     417.82     1   97.187 < 2.2e-16 ***
---
```

What are the hypotheses being tested in this table, and what do the results tell you? [5]

(f) The following figure shows plots of both the observed values of the response variable and the deviance residuals against the linear predictors, for the model in part (e) above.



What does the "residuals versus linear predictors" plot tell you about the (lack of) model fit? Explain why the residuals versus linear predictors plot shows two well-defined curves. [3]

[Total marks: 20]