MATH 38172 Generalised Linear Models - Coursework

Chen Bo Calvin Zhang

Question 1

First we import the dataset into R.

```
data = read.csv("cwkdata.csv")
print(str(data))
```

```
## 'data.frame':    120 obs. of  4 variables:
##  $ Anorexia   : chr  "Yes" "No" "Yes" "Yes" ...
##  $ Lymphocyte : num  1.107 1.643 1.161 0.913 1.092 ...
##  $ Temperature: num  37.8 38.6 38 37.4 37.7 ...
##  $ Severe     : int  1 0 0 0 1 0 0 1 0 1 ...
## NULL
```

```
print(summary(data))
```

```
##    Anorexia           Lymphocyte     Temperature        Severe
##  Length:120         Min.   :0.492   Min.   :37.03   Min.   :0.00
##  Class :character   1st Qu.:1.031   1st Qu.:37.76   1st Qu.:0.00
##  Mode  :character   Median :1.218   Median :38.02   Median :0.00
##                     Mean   :1.197   Mean   :37.99   Mean   :0.25
##                     3rd Qu.:1.377   3rd Qu.:38.25   3rd Qu.:0.25
##                     Max.   :1.880   Max.   :38.64   Max.   :1.00
```

We first notice that the dataset contains 120 observations, with 4 variables each. In this coursework, `Severe` will be the response variable and `Anorexia`, `Lymphocyte` and `Temperature` will be explanatory variables. `Severe` indicates whether a patient has a severe condition (1) or not (0). `Anorexia` indicates whether a patient is anorexic (Yes) or not (No). `Lymphocyte` is the number of lymphocytes (in billions of cells per litre) in the patient's blood. Lastly, the `Temperature` variable indicates the patient's body temperature on admission in Celsius degrees.

Both `Severe` and `Anorexia` are indicator variables despite one of them storing integer values and the other storing strings of characters. `Lymphocyte` and `Temperature` are both numerical variables. The former ranges from 0.492 to 1.880, with a mean of 1.197; the latter has a minimum temperature of 37.03 and a maximal one of 38.64, with a mean of 37.99.

Question 2

Part (i)

Because we are trying to predict the probability of patients having severe respiratory disease, we have will use a simple logistic regression. This can be performed in R using the `glm()` function using the `binomial`

family. In this case, the response variable is `Severe` and the single explanatory variable `Anorexia` (`Severe` ~ `Anorexia` with the Wilkinson-Rogers notation). The code implementation is shown below.

```
fitA = glm(Severe~Anorexia, family=binomial, data=data)
print(coef(fitA))
```

```
## (Intercept) AnorexiaYes
##   -1.812379    1.189849
```

```
print(summary(fitA))
```

```
##
## Call:
## glm(formula = Severe ~ Anorexia, family = binomial, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.92689  -0.92689  -0.54997  -0.04983   1.98172
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8124     0.3813  -4.753 2.01e-06 ***
## AnorexiaYes   1.1898     0.4640   2.565   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.96  on 119  degrees of freedom
## Residual deviance: 127.75  on 118  degrees of freedom
## AIC: 131.75
##
## Number of Fisher Scoring iterations: 4
```

As per the SBA example in Lab 1, here we do not need to specify the `weights` parameter, because the response is binary and R assumes a single trial by default.

Part (ii)

The fitted model is

$$\begin{aligned} \texttt{Severe}_i = Y_i &\sim Bernoulli(\mu_i) \\ \log \frac{\mu_i}{1-\mu_i} &= \eta_i = \beta_0 + \beta_1 x_i \\ \hat{\beta}_0 = -1.812379, &\quad \hat{\beta}_1 = 1.189849 \end{aligned}$$

where

- Y_i indicates whether patient i has a severe respiratory disease and follows a Bernoulli distribution with mean μ_i,
- μ_i is the probability that patient i has severe conditions,

- η_i is is the log-odds of the probability of severe disease,
- x_i is the indicator variable for `Anorexia`, where 1 indicates that the patient has anorexia and 0 indicates that the patient does not.

Interpretation:

- β_0 is the log-odds of a patient suffering from a severe respiratory disease if they do not have anorexia,
- β_1 is the difference in log-odds of the probability of severe disease between an anorexic individual and one who does not suffer from anorexia.

Part (iii)

The `summary()` function provides us with useful information to test whether anorexia is associated with the probability of severe disease. In particular, it provides the z-values and p-values for a Wald test of the hypothesis

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$

One asterisk ('*') indicates that we can reject the null hypothesis ($H_0 : \beta_1 = 0$) at a 5% significance level. Hence, in this model, there is significant evidence that anorexia status is associated with the probability of severe disease.

Question 3

Part (i)

This is similar to Question 2 Part (i), but instead of having `Anorexia` as explanatory variable, only `Temperature` will be used. The code implementation is very similar and is provided below.

```
fitT = glm(Severe~Temperature, family=binomial, data=data)
print(coef(fitT))
```

```
## (Intercept) Temperature
##   58.493140   -1.570614
```

```
print(summary(fitT))
```

```
##
## Call:
## glm(formula = Severe ~ Temperature, family = binomial, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.32183  -0.76332  -0.62229  -0.05724   1.96388
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  58.4931    23.7535   2.463   0.0138 *
## Temperature  -1.5706     0.6267  -2.506   0.0122 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.96  on 119  degrees of freedom
## Residual deviance: 128.35  on 118  degrees of freedom
## AIC: 132.35
##
## Number of Fisher Scoring iterations: 4
```

Part (ii)

This is again similar to Question 2 Part (i), but with `Lymphocyte` as an explanatory variable instead of `Anorexia`. The implementation is provided below.

```
fitL = glm(Severe~Lymphocyte, family=binomial, data=data)
print(coef(fitL))
```

```
## (Intercept)  Lymphocyte
##    2.164116   -2.841617
```

```
print(summary(fitL))
```

```
##
## Call:
## glm(formula = Severe ~ Lymphocyte, family = binomial, data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.30593  -0.74452  -0.59293   0.00415   2.16640
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1641     1.0002   2.164  0.03048 *
## Lymphocyte   -2.8416     0.8805  -3.227  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.96  on 119  degrees of freedom
## Residual deviance: 122.99  on 118  degrees of freedom
## AIC: 126.99
##
## Number of Fisher Scoring iterations: 4
```

Conclusions

Similarly to Question 2 Part (iii), we will use the information provided by the `summary()` function to assess whether there is significant evidence that temperature and lymphocytes level are associated with the probability of severe disease. For temperature, we can make the same observation as for anorexia, as one

asterisk indicates that we reject the null hypothesis at a 5% significance level. However, lymphocyte is to be rejected at a 1% significance level because of the two asterisks.

Therefore, there is significant evidence that both temperature and lymphocyte level are associated with the probability of severe disease for their respective models.

Question 4

Part (i)

Let us now use R to fit a model including all three explanatory variable (without interactions) using logistic regression. This easily done using the Wilkinson-Roger notation as `Severe` $\sim$ `Anorexia` + `Temperature` + `Lymphocyte` and the `glm()` function.

```
fit = glm(Severe~Anorexia+Temperature+Lymphocyte, family=binomial, data=data)
print(coef(fit))
```

```
## (Intercept) AnorexiaYes Temperature  Lymphocyte
## -24.6035390   0.6170150   0.7020943  -3.0514753
```

```
print(summary(fit))
```

```
##
## Call:
## glm(formula = Severe ~ Anorexia + Temperature + Lymphocyte, family = binomial,
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.35911  -0.75744  -0.56489   0.02519   2.24274
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.6035    41.0654  -0.599   0.5491
## AnorexiaYes   0.6170     0.5376   1.148   0.2511
## Temperature   0.7021     1.1161   0.629   0.5293
## Lymphocyte   -3.0515     1.5415  -1.980   0.0478 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.96  on 119  degrees of freedom
## Residual deviance: 121.41  on 116  degrees of freedom
## AIC: 129.41
##
## Number of Fisher Scoring iterations: 4
```

Again, using the p-values provided by the `summary()` function, we can perform a Wald test on each explanatory variable and observe that we reject the hypothesis that the lymphocyte level does not contribute to the probability at a 5% significance level, whilst we cannot reject the hypothesis that the other explanatory

variables (including the intercept) do not contribute to the probability of severe disease. Therefore, we can conclude that only `Lymphocyte` has a significant association with the probability of severe disease in this model.

Part (ii)

Comparing the results from Part (i) and Questions 2 and 3, we can say that only the level of lymphocytes significantly contributes to the probability of severe disease. This is seemingly in contrast with the conclusions from Question 2 Part (iii) and Question 3.

However, the reason why `Anorexia` and `Temperature` have a significant association with the probability of severe disease in the simpler models was that they are the only variable in the models fitted above (excluding the intercept) and a model with only the intercept would be too simple. Moreover, fitting a GLM using only `Anorexia` or only `Temperature` tends to find some correlation between the previously mentioned explanatory variables and the severity of the disease. However, these models lacked the level of lymphocytes variable, which, from the model fitted in Part (i), is the only significant explanatory variable, as it is the only variable for which we can reject the hypothesis of it equating to zero at a 5% significance level. Hence, once the `Lymphocyte` variable was introduced, the other two lost their significance in the prediction.

Part (iii)

We can now compare the simple models with the more complex one using the GLRT as the former are all nested in the latter. This is easily done in R with the `anova()` function.

```r
print(anova(fitA, fit, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model 1: Severe ~ Anorexia
## Model 2: Severe ~ Anorexia + Temperature + Lymphocyte
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       118     127.75
## 2       116     121.41  2   6.3418  0.04197 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(anova(fitT, fit, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model 1: Severe ~ Temperature
## Model 2: Severe ~ Anorexia + Temperature + Lymphocyte
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       118     128.35
## 2       116     121.41  2   6.9376  0.03115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(anova(fitL, fit, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model 1: Severe ~ Lymphocyte
## Model 2: Severe ~ Anorexia + Temperature + Lymphocyte
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       118     122.99
## 2       116     121.41  2   1.5735   0.4553
```

The output of the comparisons confirms what has been said in Part (ii). We can see that both the model where only `Anorexia` is used as an explanatory variable and the one where only `Temperature` is used are to be rejected at a 5% significance level in favour of the more complex model (containing the lymphocytes level variable). The only model we cannot reject is the one with only the lymphocytes level variable. Therefore, this model will be used as a final one as there is no statistical evidence to choose the more complex one.

Question 5

As stated in Question 4 Part (iii), the model with only `Lymphocytes` as explanatory variable will be used for prediction. The probability of severe disease for a non-anorexic patient with a temperature of 38°C and lymphocyte level of 1.3 billion cells per litre of blood is Approximately 17.80%. This can be calculated using the R function `predict.glm()`.

```
print(predict.glm(fitL, data.frame(Anorexia="No", Temperature=38, Lymphocyte=1.3),
                  type="response"))
```

```
##         1
## 0.1779958
```

```
print(predict.glm(fit, data.frame(Anorexia="No", Temperature=38, Lymphocyte=1.3),
                  type="response"))
```

```
##         1
## 0.1311448
```

For completeness, we also compared the results with the ones produced by the model with all variables, and as we can see the probabilities are not much different, meaning that the conclusion of Question 4 was indeed correct.