Coursework Submission

University ID: 10667974

April 28, 2023

Report for MATH38172 Generalised Linear Models

The aim of this report is to observe and create a generalised linear model of the relation between smoker and non-smoker women and death probabilities.

Question 1) Data Import in R

(i) Explanation: In **part 1**, we are asked to import the data in R. The data itself consists of categorical variables; the age groups and the status of the women (smokers and non-smokers) and continuous data of the number of alive and dead women 20 years after being at risk. We can import the data in **R**, by using the **read.csv()** command in **R** and subsequently attaching the table with **attach()**. This will enable us to operate on the data
(ii) R Code:

```
rm(list = ls())
options(scipen = 20)
cw.data <- read.csv("smokingdata.csv")
attach(cw.data)
cw.data
```

Question 2) Fitting a Tentative Logistic Regression Model

a) Explaining the Proportion with "Smoking" alone

(i) Explanation: We are asked to create a logistic regression model using only the "Smoking" categorical variable. We do that with the **glm()** command in **R**. When calling the **summary()** command on the newly created model, we can observe the estimates of the coefficients, their standard deviations and importantly, their p-values. The output can be found in the **R** Code section of this question. To note, from the summary, the calculated coefficients are $(-0.78052, -0.37858)$ and their respective standard errors are $(0.07962, 0.12566)$
(ii) R Code:

```
tot.at.risk <- Dead+Alive
proportion.of.dead <- Dead/tot.at.risk
fit1_smoking <- glm(proportion.of.dead ~ Smoking, data = cw.data, family = binomial, weights =
summary(fit1_smoking)
#Call:
#glm(formula = proportion.of.dead ~ Smoking, family = binomial,
#    data = cw.data, weights = tot.at.risk)
#
#Deviance Residuals:
#   Min      1Q  Median      3Q     Max
#-9.052  -5.674  -1.869   5.776  12.173
#
#Coefficients:
#              Estimate Std. Error z value             Pr(>|z|)
#(Intercept)   -0.78052    0.07962  -9.803 < 0.0000000000000002 ***
#SmokingSmoker -0.37858    0.12566  -3.013              0.00259 **
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#(Dispersion parameter for binomial family taken to be 1)
#
#    Null deviance: 641.5  on 13  degrees of freedom
#Residual deviance: 632.3  on 12  degrees of freedom
```

```
#AIC: 683.29
#
#Number of Fisher Scoring iterations: 4
```

b) Discussing the Tentative Model

(i) Explanation:

Having done the summary, we next write out our model, a series of Bernoulli trials, each taking a value of 1 or 0, depending if the woman has died within the next 20 years. **We can write our response** $Y_i \sim$ **Bern**(μ_i).

Where μ, is given by the simple logistic regression and can be interpreted as the probability of death within the next 20 years given by the log-odds. Further, we can also write our model in equation form, using the logit link function, because our response Y_i is binary:

$$\log \frac{\mu_i}{1-\mu_i} = \eta_i = \beta_0 + \beta_1 * x_i \tag{1}$$

Naturally, as we did not specify, **R** set the intercept (β_0) to "non-smoker", and if the i'th individual observed is indeed a smoker, i.e *SmokingSmoker*, the term β_1 is present as well, i.e the dummy variable x_i is:

$$x_i = \begin{cases} 1, & \text{if i'th woman is a smoker} \\ 0, & \text{if i'th woman is a non-smoker} \end{cases} \tag{2}$$

From the **R** output, we know the values of the fitted coefficients. Hence, the fitted model looks like such:

$$\log \frac{\hat{\mu_i}}{1-\hat{\mu_i}} = \hat{\eta_i} = -0.78052 - 0.37858 * x_i \tag{3}$$

Although we see that both estimates have a very significant p-value, we can clearly observe something is not quite right, as the coefficient for "the person is smoking", β_1 is negative. Why is that? Because of the extraneous variation present, we can make a biased conclusion that if a person is smoking, then they have a lower probability of dying.

To show the above statement is true, we can output the proportion of dead smokers versus dead non-smokers in the next 20 years. **We can further interpret the intercept,** β_0**, as the expected log-odds of death within the next** 20 **years of a non-smoker and** β_1**, as the difference in the expected log-odds of a woman smoker dying, compared to those of a non-smoker woman at risk in the next** 20 **years.** From equation (3), with $x_i = 0$, we can solve for $\hat{\mu}$. Doing that, we find the estimated proportion of dead non-smokers from the total women at risk in the next 20 years is: $\hat{\mu}_{non-smoker} = 0.314208$. And if $x_i = 1$ (the woman at risk smokes), the proportion of expected dead women from the total women at risk is in the next 20 years is: $\hat{\mu}_{smoker} = 0.238831$. Clearly, this result shows that women who smoke on average have a lower mortality rate, indicated by the lower proportion of dead women smokers over the total women at risk.

If we observe more closely, we see that there are far fewer smokers than non-smokers, in all of the age groups. More importantly, there are 93 more smokers in the age bracket **65-74**, and 51 more in the **75+** bracket. We see that a big chunk of those elderly women do not live through the first of the two brackets, and all of them die in the second bracket, no matter the status.

From here, there are two ways we can proceed. One way. is to cut the data of elderly women, i.e use **nrows = nrow(read.csv()) - 4)** as an argument when creating the model (or reading the table), which would mostly remove the skewness present in our data, and draw better conclusions from that model or even better; create a new model, with the age-brackets as explanatory variables. This is the topic of the next question.

(ii) R Code:

```
#cw.data.altered <- read.csv("smokingdata.csv", nrows = nrow(read.csv("smokingdata.csv")) - 4)
#Command not ran, only here for the interested reader.
```

Question 3) Fitting a Logistic Regression Model Using All of the Data

a) Explaining the Proportion with Both Categorical Variables

(i) Explanation: We're once more asked to create a model, this time with both categorical variables. We'll do so with aforementioned commands and call the **summary()** on the new model.
(ii) R Code:

```
fit2_proper <- glm(proportion.of.dead ~ Smoking + Age, data = cw.data,
family = binomial, weights = tot.at.risk)
summary(fit2_proper)
#Call:
#glm(formula = proportion.of.dead ~ Smoking + Age, family = binomial,
#    data = cw.data, weights = tot.at.risk)
#
#Deviance Residuals:
#     Min        1Q    Median        3Q       Max
#-0.72545  -0.22836   0.00005   0.19146   0.68162
#
#Coefficients:
#                Estimate Std. Error z value             Pr(>|z|)
#(Intercept)      -3.8601     0.5939  -6.500      0.0000000000805 ***
#SmokingSmoker     0.4274     0.1770   2.414              0.015762 *
#Age25-34          0.1201     0.6865   0.175              0.861178
#Age35-44          1.3411     0.6286   2.134              0.032874 *
#Age45-54          2.1134     0.6121   3.453              0.000555 ***
##Age55-64          3.1808     0.6006   5.296      0.0000001182012 ***
#Age65-74          5.0880     0.6195   8.213 < 0.0000000000000002 ***
#Age75+           27.8073 11293.1430   0.002              0.998035
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#(Dispersion parameter for binomial family taken to be 1)
#
#    Null deviance: 641.4963  on 13  degrees of freedom
#Residual deviance:   2.3809  on  6  degrees of freedom
#AIC: 65.377
#
#Number of Fisher Scoring iterations: 20
```

b) Explaining the Notation of the New Model

(i) Explanation: Defining the response the same as above, the extended model can be written as:

$$\log \frac{\mu_i}{1-\mu_i} = \eta_i = \beta_0 + \beta_1 * x_i + \sum_{j=1}^{6} d_{i,j} * \gamma_j \tag{4}$$

Where β_0 again is the intercept, but this time it represents a woman who is in their **18-24** age and is not a smoker. β_0 is the baseline. β_1 is the coefficient of the smoking status of the woman. $d_{i,1}$ $\dots d_{i,6}$ are the indicators for the categorical variables for the age groups of the women, while γ_j are their respective coefficients. **We interpret the baseline, β_0 as the log-odds of a woman dying within the next 20 years, who is currently of age 18-24 and is a non-smoker. We proceed to interpret β_1 the same, i.e the difference in the log-odds of death between a woman smoker and a non-smoker (provided they're from the same age group). The γ_j's here represent the difference in the log-odds of death within 20 years between a woman in the respective bracket age and one aged $18-25$. For instance, γ_2 represents the difference in the log-odds of death within 20 years between a woman in the age group of 35-44, compared to the baseline of $18-25$, with both having the same smoking status. Similarly, we interpret the other γ_j's for the respective $j'th$ age group.**

Each categorical variable takes a value of 1 or 0, depending on if the observed woman belongs to the age group. Indicators $d_{i,j}$:

$$d_{i,j} = \begin{cases} 1, & \text{if i'th woman belongs to j'th age group} \\ 0, & \text{if i'th woman does not belong to j'th age group} \end{cases} \tag{5}$$

From the **R** summary, we can write the down the estimated coefficients:

$$\log \frac{\hat{\mu_i}}{1-\hat{\mu_i}} = -3.8601+0.4274*x_i+d_{i,1}*0.1201+d_{i,2}*1.3411+d_{i,3}*2.1134+d_{i,4}*3.1808+d_{i,5}*5.0880+d_{i,6}*27.8073 \tag{6}$$

This is:

$$\log \frac{\hat{\mu_i}}{1-\hat{\mu_i}} = \hat{\eta_i} = \hat{\beta_0} + \hat{\beta_1} * x_i + \sum_{j=1}^{6} d_{i,j} * \hat{\gamma_j} \tag{7}$$

For example, if a woman is in the age group **35-44**, with that age group being the *second ordered by age covariate* we have, then the equation for the linear predictor looks like such:

$$\hat{\eta_i} = -3.8601 + 0.4274 * x_i + 1 * 1.3411 \tag{8}$$

Further depending if the they're a smoker, x_i will either take a 1 or 0.

An important observation to make is now that we use all the data available to us, we can see more clearly how age and if being a smoker affects the probability of dying. From the **R** output, we extrapolate that the categorical variable **AGE75+** has a p-value of over 0.99, which means that it is not statistically significant. Also, the same can be said about the variable **Age25-34**. The former of the two, we discussed above is insignificant, because be it smoker or non-smoker- all people die in that age group, hence it does not contribute much in explaining the variation. As for the latter, we can make an observation that people are generally healthy when young, so their longevity cannot be shortened by being exposed to cigarettes for a few years. **Nonetheless, all the parameters, except the intercept, take a positive value, which is a result we should be expecting, as both age and smoking increase the log-odds (even if a little) of dying within the next 20 years. As the age rises of the i'th person at risk, we see an increase in their coefficient, as they change age categories.**

c) Comparison to Q2) b)

(i) Explanation: Contrast to the negative coefficient we got in **Q2) b)** for β_1, we now have a positive one, as discussed above, which aligns with what we expect to see. In **Q2)** we built a model with only a single explanatory variable, that being if a person is a smoker or non-smoker. Because it was the sole explanatory variable, it was statistically significant, yet the model present was not adequate,

because of the aforementioned negative coefficient of β_1. By including the categorical variables for the age groups we were able to explain this extraneous variance. We were able to explain more of the variance, by using all of the existing data into the model we created in **Q3)**. Further, as we hinted above, statistically we do not see the covariate for **AGE75+** contributing to the model, same as the covariate **Age25-34**. From the **R** output, both p-values are high, i.e they are not significant.

d) Nested Model Comparison at the 5% SL

(i) Explanation: Now that we have our full model, we can compare its parameters, if it is proper to drop any of them. We do this, by nested model comparison and the **anova()** function in **R**.

To check if the age categorical variables are a good edition to our model, we compare the full model, our "fit2_proper", with our model from question **1**, "fit1_smoking". The R output results in a deviance of 629.92. The chi-squared value, with 6 degrees of freedom[1] is 12.592. With the null hypothesis for this test being: $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_6 = 0$ versus H_A : *At least one is significant.* We reject this null hypothesis, concluding that at least one of them is significant. **R** supports our conclusion with the p-value of 0.00000000000000022, when using argument **test="Chisq"** in the **anova()** function. Concluding, that the probability of death depends on the covariate **age**.

A similar model comparison can be ran to check if the variable *Smoking* can be dropped. We create a new model, regressing the proportion only on the categorical variable *Age*. Running **anova()** yields a deviance result of 5.946 and we can recall the chi-squared valued at the 5% significance level is $1.96^2 = 3.8416$. Therefore, we once more reject the null hypothesis, which is: $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. Again, if use the above *chisq* argument, when running **anova()**, the p-val of the test is 0.01475, which is once more significant. Hence, we conclude that the parameter β_1 is different from 0. Thereby, the probability of death depends on the smoking status.

Explicit test: Reject H_0 iff $L > \chi^2_{\alpha,df}$ Where L, as the notes in Ch. 15.1, is given by the difference in the log-likelihoods: $L = 2\,(l_B - l_A)$; α is the SL, while *df* is the difference in parameters between the two models.

(ii) R Code:

```
anova(fit1_smoking,fit2_proper)
anova(fit1_smoking,fit2_proper,test="Chisq")#629.92
fit3_check <- glm(proportion.of.dead ~ Age, data = cw.data, family = binomial, weights = tot.a
summary(fit3_check)
anova(fit3_check, fit2_proper)
anova(fit3_check, fit2_proper,test="Chisq")#5.946, Reject
```

Question 4) Further Model Results

a) Estimating the Probability of Death within the Next 20 years of a Woman Aged 55-64 who does not smoke

(i) Explanation: We're instructed to output the estimate of a woman, who is aged **55-64**, and who is not a smoker herself. Recall equation (7), which had the full fitted model:

$$\log \frac{\hat{\mu_i}}{1-\hat{\mu_i}} = \hat{\eta_i} = \hat{\beta_0} + \hat{\beta_1} * x_i + \sum_{j=1}^{6} d_{i,j} * \hat{\gamma_j}$$

[1]The d.f here is the difference in parameters of the two models

We now invert and try to solve for $\hat{\mu}$. The result is the following equation:

$$\hat{\mu} = \frac{e^{\hat{\beta_0}+\hat{\beta_1}*x_i+\sum_{j=1}^{6} d_{i,j}*\hat{\gamma_j}}}{1+e^{\hat{\beta_0}+\hat{\beta_1}*x_i+\sum_{j=1}^{6} d_{i,j}*\hat{\gamma_j}}} \tag{9}$$

As the woman is a non-smoker, the indicator x_i takes a value of 0. Further, they belong to the fourth age-group[2], i.e **55-64**, which means we can simplify equation (9), to:

$$\begin{aligned}\hat{\mu} &= \frac{e^{\hat{\beta_0}+d_{i,4}*\hat{\gamma_4}}}{1+e^{\hat{\beta_0}+d_{i,4}*\hat{\gamma_4}}} = \\ &= \frac{e^{-3.8601+1*3.1808}}{1+e^{-3.8601+1*3.1808}} = \\ &= 0.336407\end{aligned}$$

We interpret the result as follows: The probability of death within the next 20 years for a **non-smoker** in the age group **55-64** is 33.64%.

We can solve this by hand, or by using **R**. In **R**, bind the result of the **coef()** function to a variable, and call from that vector the specific values we need. Example code below:
(ii) R Code:

```
coeff.vector <- coef(fit2_proper)
coeff.vector
mu <- exp(coeff.vector[1]+coeff.vector[6])/(exp(coeff.vector[1]+coeff.vector[6])+1)
mu#[1] 0.336407
```

b) Delta Method for Finding Confidence Intervals

(i) Explanation: Now that we've found our estimate, it's only appropriate we find its confidence interval. Because we're dealing with a function, the course has geared us with a tool, namely the delta method, to finding distributions of functions. Hence, page 23 of the notes states that the confidence interval for a function is given by:

$$h(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2} * \text{s.e.}[h(\hat{\boldsymbol{\theta}})] = h(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2}\sqrt{\nabla h(\hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1} \nabla h(\hat{\boldsymbol{\theta}})} \tag{10}$$

Let: $\hat{\boldsymbol{\theta}} = \left(\hat{\beta_0}, \hat{\beta_1}, \hat{\gamma_1}, \hat{\gamma_2}, \hat{\gamma_3}, \hat{\gamma_4}, \hat{\gamma_5}, \hat{\gamma_6}\right)^{\mathrm{T}}$

Here, we have a choice on how to approach the question. A logical approach is to apply the delta method to the function

$$h(\hat{\boldsymbol{\theta}}) = \hat{\beta_0} + \hat{\gamma_4}$$

This would help us in our calculations as we only do a first derivative to find the gradient of the function. Given that $h(\hat{\boldsymbol{\theta}})$ is of a simple form, we find the partial derivatives of $h(\hat{\boldsymbol{\theta}})$ with respect to $\hat{\beta_0}$ and $\hat{\gamma_4}$ to both equal a constant, 1 (For clarity, $\frac{\partial h(\hat{\boldsymbol{\theta}})}{\partial \hat{\beta_0}} = \frac{\partial h(\hat{\boldsymbol{\theta}})}{\partial \hat{\gamma_4}} = 1$). Hence, the resulting gradient to the function of the vector is:

$$\nabla h(\hat{\boldsymbol{\theta}}) = (1,0,0,0,0,1,0,0)^{\mathrm{T}}$$

R can help us with the calculations from here on. We need to compute the standard error of this function, shown in equation (10), which is done by finding the square root of the gradient times the inverse fisher information matrix at the MLE, given to us by function **vcov()**, times the above

[2]Here we do not count the age group **18-24**, as it is part of the intercept in this model.

gradient. Running the below code yields the endpoints for our function $h(\hat{\boldsymbol{\theta}})$: $[-0.3177507, 0.3177507]$. Now that we have our endpoints, we proceed to solve equation (9) at those two endpoints. This results in the following 95% CI: $[0.2695124, 0.4105731]$.

Explicit calc:

$$
\begin{aligned}
& h(\hat{\boldsymbol{\theta}}) \pm z_{\alpha/2} * \text{s.e.}[h(\hat{\boldsymbol{\theta}})] \\
& -0.6793473 \pm 0.3177507 \\
& (-0.9970980, -0.3615965) \\
& \mu_L = \frac{exp^{(-0.9970980)}}{1 + exp^{(-0.9970980)}} \\
& \mu_U = \frac{exp^{(-0.3615965)}}{1 + exp^{(-0.3615965)}}
\end{aligned}
$$

Hence, $\mu_L = 0.2695124$ and $\mu_U = 0.4105731$. Giving us the aforementioned 95% CI.

(ii) R Code:

```
coeff.vector <- coef(fit2_proper)
delta_vector <- c(1,0,0,0,0,1,0,0)
transpose.delta_vector <- t(delta_vector)
se <- sqrt(transpose.delta_vector %*% vcov(fit2_proper) %*% delta_vector)
endpoints_of_h <- c(-qnorm(0.975)*se, qnorm(0.975)*se)
eq <- c(coeff.vector[1]+coeff.vector[6]+endpoints_of_h)
death.ci.95 <- exp(eq)/(exp(eq)+1)#[1] 0.2695124 0.4105731
```
