MATH38172 - Generalised Linear Models
Coursework 2022

Robin Curnow - 10171119

April 16, 2022

(a)

The summary of the regression model can be found in Table 1. It shows the estimate of the maximum likelihood estimate of each variable in the birthweight.csv data set calculated using fisher scoring (4 iterations). It was assumed that the weight at birth was affected independently by each of the variables. Two race coefficients are shown, one for black and one for "other" mothers. These show the change in log odds of a mother in each group of giving birth to an underweight baby compared to a white mother.

Birth Weight Model 1

	Estimate	Std. Error	p-val
(Intercept)	0.48062	1.19689	0.68801
Age	−0.02955	0.03703	0.42489
Mother's Weight	−0.01542	0.00692	0.02580*
race: Black	1.27226	0.52736	0.01584*
race: Other	0.88050	0.44078	0.04576*
Smoke	0.93885	0.40215	0.01957*
Hypertension	1.86330	0.69753	0.00756**
Premature	0.54334	0.34540	0.11571
UI	0.76765	0.45932	0.09467
Visits	0.06530	0.17239	0.70484
AIC	221.28		
BIC	253.70		
Log Likelihood	−100.64		
Deviance	201.28		
Num. obs.	189		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Summary of Logistic regression model for predicting the probability of a child being underweight at birth using all variables from the data as parameters

Code for part (a)

```
#MATH38172 - GLM Coursework
#Robin Curnow, 10171119
#13/04/2022

install.packages("texreg")
```

[1]

```
library(texreg)
```

```
#a)
#sets working directory and reads excel file into R
setwd("~/GLM")
birthweight <- read.csv("birthweight.csv")

#shows the names and types of the variables in the data
str(birthweight)

#lists the different categories for Race used in the data set
levels(factor(birthweight$Race))
#ensures "White" is used as the baseline in the model
race1 <- factor(birthweight$Race, levels = c("White", "Black", "Other"))

#fits a logistic regression model to the data and outputs a summary of it to the console
bw1 <- glm(Low ~ Age+MotherWeight+race1+Smoke+Hypertension+Premature+UI+Visits,
family=binomial, data = birthweight)
summary(bw1)
#creates LATEX syntax for a table showing the information in the summary table
texreg(bw1, digits=5, single.row = TRUE)
```

(b)

From table 1, the mother's weight before pregnancy, race, smoking status and hypertension have a statistically significant impact on the log odds of a baby being born under weight. Here, this is defined as having a p-value less than 0.05 which implies that there is at least a 95% probability that this sample is not taken from a population where the true parameter value is zero, based on a Wald statistic test. That is, for the i$^{\text{th}}$ parameter,

$$H_0 : B_i = 0 \; vs \; H_1 : B_i \neq 0$$
$$\mathbb{P}(|Z_i| > Z_{0.975}|H_0) < 0.05$$

Where,

$$Z_i = \frac{\hat{\beta}_i}{\sqrt{I_{ii}^{-1}(\hat{\boldsymbol{\beta}})}}$$

and $I_{ii}^{-1}(\hat{\boldsymbol{\beta}})$ is the Fisher information matrix for vector of estimated parameters in the model.

(c)

A summary of the new model is shown in table 2.
Code for part (c)

```
#c)
#fits a logistic regression model to the data and outputs a summary of it to the console
bw2 <- glm(Low ~ MotherWeight+race+Smoke+Hypertension, family=binomial, data = birthweight)
summary(bw2)
#creates LATEX syntax for a table showing the information in the summary table
texreg(bw2, digits=5, single.row = TRUE)
```

(d)

The new model has response variable, Y_i, indicating that the child of the i$^{\text{th}}$ mother has been born underweight. It has explanatory variables x_1 indicating the mother's weight before pregnancy, x_2 which has a value of 1 if the child is classed as "black" and 0 otherwise, similarly x_3 has a value of 1 if

the child is classed as "other" and 0 otherwise, x_4 indicates if the mother has smoked during pregnancy and finally x_5 indicates a history of hypertension. The equations for the model are

$$Y_i \backsim Bern(\mu_i)$$

$$\log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$$

The parameters $\beta_0, ..., \beta_5$ are those listed in the estimate column of table 2. The value of β_0 shows the log odds of a baby being born under weight, if all of the other parameter values were zero. The value of each of the other parameters corresponds to the effect on the log odds ($\log(\frac{\mu}{1-\mu})$) of the chance that a child will be born underweight associated with the value of that explanatory variable increasing by one. Note that variables related to race, smoking and hypertension are categorical and so can only affect the log odds by 1 times the value of the parameter. As opposed to the weight parameter which increases/decreases the log odds by the value of β_1 for each pound difference in weight between two mothers. For example, if one mother has log odds, η, of giving birth to an underweight child, then a mother with the same data except she is 1lb heavier will have log odds $\eta + \beta_1$, from table 2, a decrease of 0.0179. If the difference was 2lb, the difference in log odds would be $2\beta_1$ or 0.0358 Similarly if one mother is black and the other is "other", the difference in log odds will be $\beta_2 - \beta_3$, or 0.344. Two white mothers of the same weight, one who smokes and has hypertension and one who doesn't smoke or have hypertension, will have a difference in log odds of $\beta_3 + \beta_4$ or 2.821. Suffering from hypertension increases the log odds, and therefore the probability of a low birth weight, the most followed by being black and then smoking. Although weight can also affect the log odds as dramatically if there is a big enough weight difference, 97lb to have the same affect as having a history of hypertension.

Birth Weight Model 2

	Estimate	Std. Error	p-val
(Intercept)	0.35205	0.92444	0.70333
Mother's Weight	−0.01791	0.00680	0.00844**
race: Black	1.28766	0.52164	0.01357*
race: Other	0.94365	0.42338	0.01584*
Smoke	1.07157	0.38751	0.04576**
Hypertension	1.74916	0.69082	0.01134*
AIC	220.25		
BIC	239.70		
Log Likelihood	−104.12		
Deviance	208.25		
Num. obs.	189		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Summary of logistic regression model for predicting the probability of a child being underweight at birth using only the parameters from the data that were significant in table 1

(e)

$$\eta = \log(\frac{\mu}{1-\mu}) = \begin{pmatrix} 1 & 130 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 1 & 130 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.35204934 \\ -0.01790657 \\ 1.28766233 \\ 0.943644851.07156635 \\ 1.74916259 \end{pmatrix}$$

$$\hat{\mu} = \frac{\exp(\eta)}{1+\eta} = \frac{\exp(-0.6881429)}{1-0.6881429} = 0.3344463$$

Code for part (e)

```
#e)
a <- coef(bw2)
eta <- c(1, 130, 1, 0, 0, 0)%*%a
mu <- exp(eta)/(1+exp(eta))
mu
```

(f)

$$\eta = \log(\frac{0.2}{1-0.2}) = -1.386294$$

$$\hat{x}_1 = \frac{\eta - \beta_0}{\beta_1} = \frac{1.386294 + 0.35204934}{0.01790657} = 97.07852\ lb$$

$$h(\boldsymbol{\beta}) = \frac{\log(\frac{0.2}{1-0.2}) - \beta_0}{\beta_1} = \frac{-\log(4) - \beta_0}{\beta_1}$$

so $h(\hat{\boldsymbol{\beta}}) = \hat{x}_1$. And

$$\nabla h(\boldsymbol{\beta}) = \begin{pmatrix} -\frac{1}{\beta_1} \\ \frac{\beta_0 + \log(4)}{\beta_1^2} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Confidence interval is then,

$$\hat{x}_1 \pm z_{0.975}\sqrt{\nabla h(\hat{\boldsymbol{\beta}})^T I^{-1}(\hat{\boldsymbol{\beta}}) \nabla h(\boldsymbol{\beta})}$$

and so $x \in [51.00679, 143.15026]$. This is a very wide interval and so for over a wide interval of weights, it is possible for a white non-smoker with uterine irritability, but no hypertension or physician visits to have a 20% chance of having a baby born underweight.

Code for part (f)

```
    #f)
p <- 0.2
#provides a point estimate of the weight using the MLEs
eta1 <- log(p/(1-p))
x_1 <- (eta1 - a[1])/a[2]
x_1

#defines the gradient function of the function of the regression parameters
gradh <- c((-1)/a[2],(a[1]-eta1)/a[2]^2, 0, 0, 0, 0)
x_1CI <- c(x_1) + c(-1,1)*qnorm(0.975)*c(sqrt(t(gradh)%*%(vcov(bw2)%*%gradh)))
x_1CI
```

(g)

To use 'history of premature birth' as a categorical variable, all values of one or more in the 'Premature' column of the data had to be changed to ones, indicating that there is 'some' history. The R code below shows that I did this using a for loop and an if statement. For 'physician visits' to be categorical, the 'factor()' function was used so that a dummy variable would be created for each amount of visits. A summary of the third model from part (g) is shown in table 3. It shows that for this model, only being black, having a history of hypertension or a history of premature birth affect the chances of giving birth to an underweight baby. Table 4 compares the the first and third models side by side. Having the premature birth data as a continuous variable was not only not statistically significant, but

also less of an effect on the outcome. This may be because having *any* history may be a bad thing, but having had more than one occurrence may not add any effect. Smoking was also statistically significant in the first model but not the third. As Shown in table 5, the proportion of smokers with a history is $\frac{18}{30}$ or 60% where as amongst non-smokers it is only $\frac{56}{159}$ or 35% and so these two may be linked.

Code for part (g)

```
    #g
#defines a dummy variable for 'history of premature labour' called prematureHist
prematureHist <- rep(0,nrow(birthweight))
#sets the data for a mother who has had one or more premature birth to 1 in PrematureHist and other c
for(i in c(1:nrow(birthweight))){
  if(birthweight$Premature[i] == 0) {
    prematureHist[i] <- 0
  } else {
    prematureHist[i] <- 1
  }
}
#checks that the prematureHist data matches the original data.
table(PrematureHist, birthweight$Premature)

#sets visits data a categorical variable
visits1 <- factor(birthweight$Visits)

#checks the different number of times that people have had physician visits (there are 6)
levels(visits1)


#defines a third glm with these new categorical factors, summaries it and creates LATEX
output for the summary
bw3 <- glm(Low ~ Age+MotherWeight+race1+Smoke+Hypertension+prematureHist+UI+visits1,
family=binomial, data = birthweight)
summary(bw3)
texreg(bw3, digits=5, single.row = TRUE)

table(birthweight$Smoke, prematureHist)
```

References

[1] Philip Leifeld. "texreg: Conversion of Statistical Model Output in R to LaTeX and HTML Tables". In: *Journal of Statistical Software* 55.8 (2013), pp. 1–24. URL: http://dx.doi.org/10.18637/jss.v055.i08.

Birth Weight Model 3

	Estimate	Std. Error	p-val
(Intercept)	0.640468	1.276771	0.61593
Age	−0.037719	0.038853	0.33165
Mother's Weight	−0.014021	0.007321	0.05548
race: Black	1.186893	0.541908	0.02851*
race: Other	0.760423	0.468856	0.10483
Smoke	0.720645	0.433427	0.09638
Hypertension	1.781318	0.735445	0.01543*
History of Premature	1.446991	0.498000	0.00367**
UI	0.657068	0.466656	0.15912
1 visit	−0.459865	0.482282	0.34033
2 visits	−0.005881	0.536360	0.99125
3 visits	1.155177	0.863338	0.18088
4 visits	−0.600850	1.329353	0.65128
6 visits	−12.126178	882.744101	0.98904
AIC	221.39024		
BIC	266.77470		
Log Likelihood	−96.69512		
Deviance	193.39024		
Num. obs.	189		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Summary of logistic regression model for predicting the probability of a child being born underweight using all the parameters from the data, except 'number of premature births' is replaced with the categorical variable 'history of premature birth' and similarly 'number hospital visits' is now categorical.

	Model 1	Model 3
(Intercept)	0.48	0.64
	(1.20)	(1.28)
Age	−0.03	−0.04
	(0.04)	(0.04)
Mother's Weight	−0.02*	−0.01
	(0.01)	(0.01)
race: Black	1.27*	1.19*
	(0.53)	(0.54)
race: Other	0.88*	0.76
	(0.44)	(0.47)
Smoke	0.94*	0.72
	(0.40)	(0.43)
Hypertension	1.86**	1.78*
	(0.70)	(0.74)
Premature	0.54	
	(0.35)	
UI	0.77	0.66
	(0.46)	(0.47)
Visits	0.07	
	(0.17)	
history of premature		1.45**
		(0.50)
1 visit		−0.46
		(0.48)
2 visits		−0.01
		(0.54)
3 visits		1.16
		(0.86)
4 visits		−0.60
		(1.33)
6 visits		−12.13
		(882.74)
AIC	221.28	221.39
BIC	253.70	266.77
Log Likelihood	−100.64	−96.70
Deviance	201.28	193.39
Num. obs.	189	189

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Comparison of models 1 and 3. (standard error in brackets)

	Smokers	Non-Smokers
No History	12	103
History	18	56

Table 5: Comparison of smokers/non- smokers with mothers with/without a history of premature birth