STATS 330 Final Assessment; University of Auckland, Semester 1

Hasnain Cheena 190411106

Due Date: 1:00pm NZ Time, Thursday 25th June 2020

Question 1

1a)

$$log(\mu_i) = \beta_0 + \beta_1 CullBefore_i + \beta_2 GrazingModerate_i + \beta_3 CullBefore_i GrazingModerate_i$$

$$Y_i \sim Poisson(\mu_i)$$

Where:

- $CullBefore_i = 1$ if the cull of feral animals has not occurred yet on the ith site and 0 if the cull of feral animals has occurred.
- $GrazingModerate_i = 1$ if the ith site has a grazing status of moderate and 0 if the ith site has a grazing status of heavy.
- Y_i is Poisson distributed and is the number of birds heard/observed during the bird survey at the ith site.
- μ_i is the mean number of birds heard/observed at the ith site.

1b)

```
heavy.cull.effect = (exp(-0.6776)-1)*100
heavy.cull.effect
```

```
## [1] -49.21657
```

```
medium.cull.effect = (exp(-0.6776+0.4463+0.8213)-1) * 100
medium.cull.effect
```

```
## [1] 80.39884
```

The effect of *Cull* on the number of birds observed depends on the grazing status of the site.
We estimate that for sites with a grazing status of heavy, the expected number of birds observed at a site prior to culling is 49.2% lower than after culling.
We estimate that for sites with grazing status of medium, the expected number of birds observed at a site prior to culling is 80.4% higher than after culling.

1c)

```
1-pchisq(437.23, 58)
```

```
## [1] 0
```

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Assuming the conditions of the chi-squared approximation are met, yes there is evidence suggesting a lack of fit. This is because the observed deviance is much greater than the degrees of freedom of the chi-squared distribution. Therefore, the observed deviance will be out in the tails leading to a small p-value suggesting a lack of fit.

1d)

Model $conservation.fit2$ assumes the response variable has a negative binomial distribution. This is to deal with the observed overdispersion. It deals with overdispersion by changing the assumption between the mean and variance. The negative binomial assumes the mean and variance have a quadratic relationship instead of equal like the Poisson distribution.

1ei)

Estimates of $conservation.fit2$ will be different to $conservation.fit1$.

1eii)

Standard errors of $conservation.fit2$ will be higher than $conservation.fit1$. This is because the negative binomial model can account for more variance than a Poisson regression model, thus the standard errors will increase.

1f)

$$AIC = -2l + 2k$$

```
aic = -2*-180.06 + 2*5
aic
```

```
## [1] 370.12
```

$$BIC = -2l + klog(n)$$

```
bic = -2*-180.06 + log(62)*5
bic
```

```
## [1] 380.7557
```

1g)

Using AIC as the model selection criterion I would choose $conservation.fit2$. This is because it has evidence supporting it as it has a much lower AIC score than $conservation.fit1$.

1h)

If the second bird survey was carried out at the same sites at a 30-minute period rather than 20-minute period, overall more birds would have been observed. Therefore, we know the effect that the survey time period has on the expected number of birds observed.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Hence, the length of a survey measures its exposure to the number of birds observed. This known effect can be incorporated in our model by fitting the log of the variable $survey.period$ (a new variable created that denotes the length of the survey) as an offset.

Question 2

2a)

```
donald.weight.kg = 243/2.2046
pnorm(donald.weight.kg, 78,13.2, lower.tail = FALSE)

## [1] 0.007318909
```

2b)

```
dnorm(80,78,13.2) / dnorm(80,70.3,16.8)

## [1] 1.486416
```

2c)

```
set.seed(1)
n = 100000
mean((rnorm(n, 78, 13.2) - rnorm(n, 70.8, 16.8)) >= 20)

## [1] 0.27475
```

2d)

```
qnorm(0.25, 70.3, 16.8)

## [1] 58.96857
```

Question 3

3ai)

$hist(means)$ shows the distribution of sample means.

3aii)

As N is increased more random points are sampled of the Poisson distribution. This leads to the distribution of sample means becoming more normally distributed. This is an example of Central Limit Theorem in action.

3aiii)

When $Nsim$ is increased the mean of the sample means will get closer to lambda.

3aiv)

When $lambda$ is increased the distribution of sample means will become more normally distributed.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

3b)

$mean(myvec)$ calculates the average of the mean squared prediction error (MSPE).
It is not truly successful as the mean squared prediction error has been calculated on the data used in the model building process. Therefore, the MSPE estimate is optimistic and smaller than it truly should be.

Question 4

4a)

```
kids.df = transform(kids.df, numberOfPeopleAtHome= (as.numeric(bothparents)+1
) + siblings + 1)

fit.4a = glm(tv ~ numberOfPeopleAtHome, poisson, kids.df)
```

I fitted a Poisson regression model as the response variable is a discrete count. Furthermore, I combined the variables bothparents and siblings into a single numeric variable that contains information about the number of people living at home. I assumed that if bothparents = 0 then there is a single parent at home and if bothparents = 1 then there are two parents living at home. Summing that result to the total number of other siblings plus 1 (accounting for the child itself) gives you the total number of people living at home.

4b)

```
kids.df = transform(kids.df, obesity=weight/(height^2) > 30)

fit.4b = glm(obesity ~ age + sex + tv + dbp + siblings + bothparents + bovs,
binomial, kids.df)
```

I fitted a logistic regression model as the response variable is binary. Furthermore, I calculated BMI and used the WHO BMI figure of 30[1] to identify obese children. All possible explanatory variables were added to the model to determine if any of the regressors can help explain obesity in children.

4c)

```
fit.4c = glm(bothparents ~ siblings * bovs, binomial, kids.df)
```

I fitted a logistic regression model as the response variable is binary. Furthermore, an interaction term is key here to assess the condition whether the effect of family size depends on whether children have an oversea's influence.

4d)

```
kids.df = transform(kids.df, obesity=weight/(height^2) > 30)

fit.4b.gam = gam(obesity ~ s(age) + sex + tv + s(dbp) + siblings + bothparent
s + bovs, binomial, kids.df)
```

In the GAM model above I have chosen not to smooth the variables tv or siblings. This is because tv and siblings may have less than 10 distinct values.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Moreover, I would fit a GAM first. This is because fitting a GAM before a GLM would allow me to do linearity and significance tests to assess the explanatory variables. Furthermore, any variables that were found to fail linearity, GAM plots would be used to make judgements on appropriate non-linear terms to add into the model.

4e)

$fit1$ has aspects of it that are inappropriate:

1. Smoothing age with degrees of freedom of 15 is much too high.
2. *bovs* does not need to be smoothed as it is discrete.
3. Height is a continuous variable and more specifically not a count variable. Thus, a Poisson distribution may not be the best to represent the response.

Question 5

```
p = 0:10
A = -log(exp(3/2 * (p^2)))
B = (p-1)^2
lambda = 2

scores = A + lambda*B

p[which.min(scores)]

## [1] 4
```

Therefore, using the penalty function approach described in the question, 4 parameters is found to be optimal.

Question 6

6a)

Yes, it is important to perform the train test split in a random manner. Firstly, the test dataset is used to judge the model's predictive performance. In order to do this fairly and correctly the training and test sets must be approximately representative samples of the population. This fair representation is a result of random train test partitions.

6b)

Seed the random number generator to make it reproducible. For example, using code $set.seed(1)$.

6c)

Typically, having a training set of 20% relative to a test set of 80% is not proper technique. This is because the training set may not have enough data to model the underlying pattern and thus the model will be underfit. Using 80% training and 20% test would work better.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Question 7

7a)

Jitter was used to separate overlapping points on the scatterplot. This was done so overlapping points would show more clearer, to produce a high-quality scatterplot.

7b)

The dataframe was sorted by $Start$ so that the fitted model could easily and correctly be overlaid on the scatterplot using $lines(fitted(kfit1))$.

7c)

```
data(kyphosis, package = "rpart")

ooo <- with(kyphosis, order(Start))
kyphosis <- kyphosis[ooo, ]
kfit1 = glm(Kyphosis ~ Start, binomial, data = kyphosis)

cfs <- coef(kfit1)
eta = cfs[1] + cfs[2] * with(kyphosis, Start)
n.obs = nrow(kyphosis)

Nsim <- 1e4
devs <- numeric(Nsim)

for (i in Nsim){
  ysim = rbinom(n.obs, size = 1, prob = 1 / (1 + exp(-eta)))
  mod_i = glm(cbind(ysim, 1-ysim) ~ Start, binomial, data = kyphosis)
  devs[i] = deviance(mod_i)
}

round(mean(devs > deviance(kfit1)), 3)

## [1] 0
```

Conditions allowing for the chi-squared distribution approximation are not met as the data is ungrouped binary data, which is extremely sparse. This means each observation is a singular trial. Therefore, a simulation was performed to assess overdispersion with respect to the binomial.

The result above shows that there is evidence suggesting a lack of fit. Therefore, there is evidence of overdispersion with respect to the binomial.

7d)

size denotes the total number of trials that occur. $size = 1$ was set because the response variable being simulated is both ungrouped and binary, where each observation is a singular trial.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

7e)

The bootstrapping method is parametric bootstrapping.

7f)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1\theta$$

$$\theta = \frac{log(\frac{0.25}{1-0.25}) - \beta_0}{\beta_1}$$

$$\theta = \frac{-1.098 - \beta_0}{\beta_1}$$

The above shows where the -1.098 comes from. It is the log-odds of having a 25% probability of presence of kyphosis.

7g)

```
data(kyphosis, package = "rpart")

ooo <- with(kyphosis, order(Start))
kyphosis <- kyphosis[ooo, ]
kfit1 = glm(Kyphosis ~ Start, binomial, data = kyphosis)

#expected values and
cfs <- coef(kfit1)
eta = cfs[1] + cfs[2] * with(kyphosis, Start)
n.obs = nrow(kyphosis)

Nsim = 1e3
est.theta = matrix(0, Nsim)

start.value = data.frame("Start"=c(10))

theta.hat = predict(kfit1, start.value, type="response")

#simulate
for (i in 1:Nsim){
  #simulate responses
  ysim = rbinom(n.obs, size = 1, prob = 1 / (1 + exp(-eta)))
  #refit
  kfit.sim = glm(cbind(ysim, 1-ysim) ~ Start, binomial, data = kyphosis)
  #estimate probability when start is 10
  est.theta[i] = predict(kfit.sim, start.value, type="response")
}

#quantiles to obtain 95% interval
```

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

```
a = theta.hat - quantile(est.theta, prob=0.025)
b = quantile(est.theta, prob=0.975) - theta.hat
ci = c(theta.hat-b, theta.hat+a)
names(ci) = c("2.5%", "97.5%")
ci
```

```
##      2.5%     97.5%
## 0.1166056 0.3149339
```

Therefore, we estimate that when the top ten vertebrae are operated on (Start=10), the probability of presence of kyphosis is between 0.12 and 0.32.

7h)

Assuming mgcv is already attached:

$$fit.kyphosis.gam = gam(kyphosis \sim s(Start), binomial, kyphosis)$$

Question 8

8a)

```
prevelance = 744/1586
prevelance
```

```
## [1] 0.4691047
```

8b)

```
sensitivity = 670/744
sensitivity
```

```
## [1] 0.9005376
```

8c)

```
specificity = 640/842
specificity
```

```
## [1] 0.760095
```

8d)

```
negative.nodisease = 640 / (74+640)
negative.nodisease
```

```
## [1] 0.8963585
```

8e)

From a patient's point of view, it is better to have false positive than false negative. This is because a false positive means the patient doesn't have the disease instead thinks they do. However, in the much worse case of a false negative the patient has the disease and thinks they don't which causes them to not get the required treatement for the disease.

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight