2019-sem2

2019 Semester 2 Past Exam

Question 1)

a)

Poisson distribution has the feature $E(Y) = Var(Y)$. This is not evident in the data.

b)

i)

$Log(mu_i) = \beta_0 + \beta_1 \times day_i$

Where mu_i refers to the expected number of spam emails for that day And day_i is a dummy variable which takes the value 1 if it is a weekend, 0 otherwise

ii)

```r
beta_0 = 3.0378
beta_1 = 0.5002

#week day
exp(beta_0)
```

```
## [1] 20.8593
```

The fitted is $Y \sim poisson(20.86)$

iii)

```r
exp(beta_0 + beta_1)
```

```
## [1] 34.39805
```

The fitted distribution is $Y \sim poisson(34.40)$

c)

This is because the deviance is extremely large relative to the degrees of freedom. The results the goodness of fit chi-squared test conforms this.

d)

The change in standard error and corresponding t-value and p-value.

It is a quasi-poisson model, with a dispersion parameter that changes the variance.

$Var(Y) = k \times \mu$

Where k is our dispersion parameter, which is greater than 1

Thus we have increased the variance of Y, which in turn increases the standard deviation, standard error, t-value and p-values.

e)

This is because our deviance goodness of fit test (with the null hypothesis H0: the model is appropriate) is 0.298. Therefore we accept the null, the model is appropriate.

f)

Given $AIC = -2\ell + 2k$

```
poisEll <- -1271.2
poisK <- 2

negBinEll <- -799.94
negBinK <- 3

AICPois <- -2*poisEll + 2*poisK
AICNegBin <- -2*negBinEll + 2*negBinK

AICPois
```

```
## [1] 2546.4
```

```
AICNegBin
```

```
## [1] 1605.88
```

THe AIC for the poisson model is 2544. The AIC for the negative binomial model is 1603.88

g)

$E(Y) = \mu = exp(\beta_0 + \beta_1 \times day_i)$

```r
beta_0 = 3.0378
beta_1 = 0.5002
mu = exp(beta_0 + beta_1*0)

mu
```

```
## [1] 20.8593
```

$E(Y) = \mu = 20.86$

$Var(Y) = \mu + \frac{\mu^2}{\theta}$

```r
theta = 3.9364

varY <- mu + (mu^2/theta)

varY
```

```
## [1] 131.3944
```

$Var(Y) = 131.39$

h)

The model with a negative binomial distribution is the best model. It has the smallest AIC relative the poisson model, and is the only distribution to pass a deviance GOF check.

```r
100*(exp(c(0.35027, 0.6502))-1)
```

```
## [1] 41.94507 91.59240
```

Holding other variables constant, the expected number of spam emails for a given day is between 41.95% and 91.60% higher on the weekend than on a weekday.

Question 2)

a)

Due to sparsity it is very unlikely that the residual deviance is approximated by a chi-squared distribution.

b)

i)

First we would determine our number of simulations, typically 10000

Then create an empty vector to store the deviance for each simulation

For each simulation, we would randomly generate a new y that follows a binomial distribution using our estimated probability of success $\hat{p}$ and with number of trials is 1.

Then also fit a new model glm(y ~., family = binomial, data = data)

And store the estimated deviance in our vector.

ii)

First determine number of simulations Then Create empty vector for deviance

For each simulation generate a random sample of the observed data, with replacement.

Fit a model based on this data using glm(dieldrin ~., data, binomial)

Record the deviance and for each simulation put it in the deviance vector

iii)

We are testing the null hypothesis that the logistic regression is an appropriate model for the data. Therefore, we are assessing our *estimated* values. This is exactly what parametric bootstrapping does. Alternatively non-parametric assess the TRUE values, which is inappropriate in this instance.

c)

The p-value when using empirical sampling is the proportion of sampled deviance greater or equal to the actual. Given a deviance of 41.87, and given the median sampling deviance is 42.3. This indicate that the p-value is slightly larger than 0.5

d)

```
#Specificity
20/27
```

```
## [1] 0.7407407
```

```
#Sensitivity
12/16
```

```
## [1] 0.75
```

```
11/43
```

```
## [1] 0.255814
```

The specificity is approximately 0.74, the sensitivity is approximately 0.75

Overall error rate is 0.26

e)

i)

The first number, 0.344, indicates the cut off point that produces the maximum sum of sensitivity + specificity. The numbers inside the brackets are those corresponding figures of sensitivity and specificity.

ii)

AUC standard for Area Under Curve. It is the area below the ROC curve on the ROC plot.

iii)

```
0.667 + 0.875
```

```
## [1] 1.542
```

```
0.8+0.8
```

```
## [1] 1.6
```

NO, the maximized sum of sensitivity and specificity is 1.542 which is smaller than 0.8+0.8 = 1.6

Question 3)

a)

A direct causal effect occurs when change on one variable leads to a direct change in another variable, when holding everything else constant

b)

The variables Dry, Mtemp, Month, Pollen and Pap

c)

(Just include all the variables with direct causal effects)

$asthma \sim dry + mtemp + month + pollen + Pap$ and it is a poisson regression model (given we are modelling a count)

d)

Total causal effect indirect effects on the response variable as a result of the change in an explanatory (ie it includes indirect casual pathways.)

e)

A confounder is variable that a direct effect on all variables of concern (response variable and any explanatory variables). If we are modelling any direct causal from our data, the confounder would be month.

f)

$asthma \sim month$ using poisson regression

g)

$asthma \sim calm + month + dry + mtemp$, using poisson regression