2021-sem-1

mjon238

14/02/2022

#Question 1

a)

Firstly the residual deviances

```
#Poisson residual deviance
1-pchisq(97.045, 30)
```

```
## [1] 5.405949e-09
```

```
#Negative binomial residual deviance
1-pchisq(40.148, 30)
```

```
## [1] 0.1020294
```

If the model is appropriate we would expect the residual deviance to come from an approximate chi-squared distribution. Assuming our models expected values are not sparse, which is the case in this scenario

If we run chi-squared test for the residual deviance with the null hypothesis: the assumptions for this models distribution are appropriate. We reject the null hypothesis for the poisson distribution, i.e. the distribution is NOT appropriate. We accept the null hypothesis for the negative binomial distribution, the model distribution is appropriate.

The quasi poisson deviance is difficult to assess because it does not have a log likelihood function (R just assumes the same residual deviance as poisson).

The deviance plots also suggest the negative binomial model is a better fit. For negative binomial model, the pearson residuals and deviance residuals are a pattern less band around a mean of zero, and are between the interval (-3,3). This suggests the pearson and deviance residuals are approximately standard Normal, with mean zero, variance 1. This further indicates that the distribution is appropriate.

The pearson and deviance residuals in the poisson/quaispoisson fits have slightly increasing variance and are all observations are NOT inside the interval (-3,3). This suggest the residual distributions are not approximately standard normal with mean zero, variance 1 and the models distribution is not appropriate.

Finally the AIC for the negative binomial model is smaller than the poisson model, with a difference greater than 10. Telling us this model better describes the data.

Also the quasi-poisson has a noticebale outlier, the negative binomial does not.

b)

i)

This offset allows transformation of our response variable into determining the count of apprentices per 1000 members of the population. This is a more appropriate response because it removes the influence of population size in determining the number of apprentices and allows us to better understand the influence of other features in a county.

ii)

we can interpret the model as the expected number of apprentices per 1000 members of the population.

c)

```r
beta_0 = 4.48652
beta_1 = -1.26959
beta_2 = 0.11953
beta_3 = -0.03484

distance = 60
urban = 15
population = 28000

logY = beta_0 + beta_1*log(distance) + beta_2*urban + beta_3*log(distance)*urban

exp(logY)
```

```
## [1] 0.3470198
```

We would expect 9.72 apprentices of the total population

d)

The authors expected countys further from Edinburgh had lower migration and more urbanised areas had lower migration.

We can simplify the model, given the coefficients

$$log(\mu_i) = \beta_0 + \beta_1 \times log(Dist) + \beta_2 \times Urban + \beta_3 \times log(Dist) \times Urban$$

FIRST DISTANCE: SO we can simplify to

$$log(\mu_i) = \beta_0 + \beta_2 \times Urban + log(Dist)(\beta_1 + \beta_3 \times Urban)$$

$\beta_1 + \beta_3 \times Urban$ becomes our coefficient for log Distnace

The Value $log(Dist)(-1.26959 - 0.03484 \times Urban)$ Is always negative for distance > 1 and any value of urban, which is true for ALL counties. Therefore we can confirm as distance increases the expected value of log apprentices and thus apprentices migrating decreasses

NOW FOR URBAN

$$log(\mu_i) = \beta_0 + \beta_1 \times log(Dist) + Urban(\beta_2 + \beta_3 \times \log(Dist)$$

$\beta_2 + \beta_3 \times log(dist)$ becomes our coefficient for Urban

$Urban(0.11953 - 0.03484 \times log(Dist)$ is only positive from SOME values of distance, specifically when Distance is smaller than

When Distnace is greater than 0.309, $(0.11953 - 0.03484 \times \log(Dist)$ is smaller than 0 and any increase in urban will have a negative effect on the log of apprentices moving to Edinburgh.

In conclusion the distance and apprentcies is always negatively related.

Urbanisation and apprencietns is dependntent on the level of distnace, when distance is small the relationship is positive, when distance is big it is negative.

Question 2)

a)

```
exp(-1.1857479)
```

```
## [1] 0.3055176
```

Holding all other factors constant, given an additional child under the age 7 year old of a married woman, the odds that same married woman participates in the labor force is multiplied by 0.31.

b) Didn't do this

c)

i)

A non-parametric bootstrap was used given in line 5 and 6 we take a sample of real data with replacement.

ii)

```
mean <-  0.8280919
quantileLow <- 0.6911
quantileHigh <- 0.9171

c(2*mean - quantileHigh, 2*mean - quantileLow)
```

```
## [1] 0.7390838 0.9650838
```

Using non-parametric bootstrapping, the 95% (inverted) confidence interval is 0.739, 0.965

d)

i)

Specificity is the proportion of women that we both predicted and actually Do NOT participate in the labour force out of the total women we predicted do NOT participate in the labour force

Sensitivity is the proportion of married woman that we both predicted and actually DO participate in the labour force out of the total women we predicted participate in the labour force.

ii)

```
#Specificity
337 /(337+134)
```

```
## [1] 0.7154989
```

```
#Sensitivity
255/(255+146)
```

```
## [1] 0.6359102
```

```
#Error Rate
(134+146)/(134+146+337+255)
```

```
## [1] 0.3211009
```

The specificity is 0.715, the sensitivity is 0.636, the error rate is 0.291

iii)

The estimates above are optimistic because we are using the training data to determine test predictive ability. We can generate an honest estimate by separating the data into training data and test data. Train the model using the training data and then we fit the model using test data.

After we have done this we can assess the predictive ability of the model by comparing the fitted test set with the actual.

We could also use cross-validation but given we have a large number of observations, this is probably unnecessary.

e)

i)

0.467 Represents the cut off point (the value of p_i that determines a success), 0.677 represents the specificity at this cutoff point, 0.713 represents the sensitivity at this cut off point.

ii)

The true positive rate is equal to sensitivity, and the false positive rate is equal to 1- specificty

Therefore in this scenario sensitivity is 0.7, and specificity is 0.8

This is not possible because on the ROC plot that very point is above the ROC curve. I.e. no cut off point exists when TPR is at least 0.7, FPR is at least 0.2.

Question 3)

a)

Variables F, D, and E have direct causal effects on G

b)

A, C, D, should be included

c)

A and C should be included

d)

i)

A and C are the only confounder

ii)

G is the only collider