

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)

```
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 counties 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

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

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 decreases

NOW FOR URBAN

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

$Urban(0.11953 - 0.03484 \times \log(Dist))$ is always positive for any value of Distance or Urban, suggesting as urbanisation increases so does the log of apprentices migrating and thus the expected value of apprentices migrating