Assignment 3 ST466

Alok Kumar Singh 19250990

03:27:27 PM 10 April, 2020

Q1. School administrators study the attendance behavior of secondary school students. A predictor of the number of days of absence includes a standardized test in math and gender identity. The data can be found in attendance.csv.

a). Fit the Poisson regression model to these data. Provide the Poisson regression equation based on the model output. Provide an interpretation of the coefficients.

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

```
## -- Conflicts ---------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
attend <- read.csv("attendance.csv")
head(attend)
```

```
##   gender math daysabs
## 1   male   63       4
## 2   male   27       4
## 3 female   20       2
## 4 female   16       3
## 5 female    2       3
## 6 female   71      13
```

```r
fit <- glm(daysabs ~ ., family = poisson(), data = attend)
summary(fit)
```

```
##
## Call:
## glm(formula = daysabs ~ ., family = poisson(), data = attend)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1803  -2.5346  -0.8987   0.8441   7.3173
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.4031733  0.0495417  48.508  < 2e-16 ***
## gendermale  -0.2548442  0.0467239  -5.454 4.92e-08 ***
## math        -0.0112160  0.0009297 -12.064  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2217.7  on 313  degrees of freedom
## Residual deviance: 2042.5  on 311  degrees of freedom
## AIC: 2931.9
##
## Number of Fisher Scoring iterations: 5
```

Answer:-

muhat = exp(BetaNot + BetaOneX1 + BetaTwoX2)

The model fit:-
muhat = exp(2.4031733 - 0.2548442gendermale - 0.0112160math)

Interpretations of coefficient:-

For X1 increases from x1 to x1 + 1 and other x's remain fixed. Then mu changes from mu(x) = exp(BetNot + BetaOnex1 + … + BetaKxk) to mu(x + 1) = exp(BetaNot + BetaOne(x1 + 1) + … + BetaKxk) = mu(x)exp(BetaOne) Each unit increase in x1 multiplies the mean response by exp(BetaOne) .

Intercept – for a unit change in the predictor variables, the difference in the logs of expected counts is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.

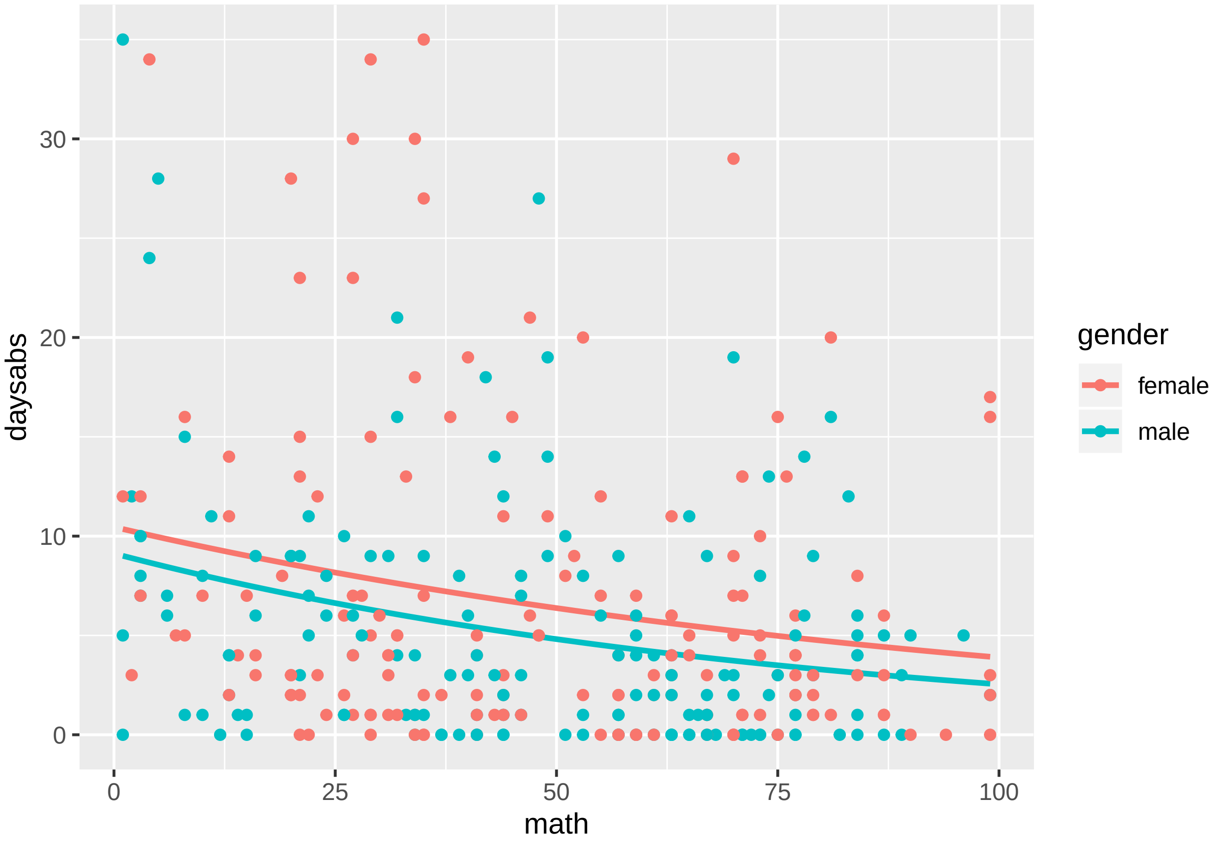
maths-For a increase of one point in maths, the difference in the logs of expected counts would be expected to decrease by 0.0012 unit, while holding the other variables in the model constant.

gendermale – This is the estimated Poisson regression coefficient comparing male to female, given the other variables are held constant in the model. The difference in the logs of expected counts is expected to decreease by 0.2548 unit for males compared to females, while holding the other variables constant in the model.

b). Pot the observed number of absent days vs the math score and distinguish the points based on gender (using colour). Overlay predictions from your model on this plot and comment on the model fit.

```
attend_res <- attend %>% mutate(fit_p = predict(fit, type = "response"),
res_p = residuals(fit))

ggplot(attend_res, aes(x = math, y = daysabs, color = gender)) +
geom_smooth(method = "glm", se = F , aes(fill = fit_p),method.args = list(family = 'poisson')) + geom_p
```



```r
library(AER)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
dispersiontest(fit)
```

```
##
##  Overdispersion test
##
## data:  fit
## z = 7.6139, p-value = 1.329e-14
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   7.333203
```

We can see that the model is not perfectly fit to this dataset and also we can see from the dispersion test that alternative hypothesis is true which suggest that the dispersion is greater than 1 which suggest that there is overdispersion of data. So there is further investigation is required for this and we can also check for quassi poisson or negative disperssion methods to overcome this overdispersion.

c). Using equations, specify a negative binomial regression model for these data. Identify the random component, the systematic component and the link function.

Answer:-

random component = Y ~ Neg.binom(r, pi)
systematic component = BetaNot + BetaOneX1 + BetaTwoX2
link function:-
log(mu) = BetaNot + BetaOneX1 + BetaTwoX2 Where log(mu) is called as link function.

d). Fit the negative binomial model to these data. Has your interpretation of the coefficients changed compared to the fitted Poisson model? How have the standard errors been impacted?

```
fit1 <- glm.nb(formula = daysabs ~ ., data = attend)
summary(fit1)
```

```
##
## Call:
## glm.nb(formula = daysabs ~ ., data = attend, init.theta = 0.8705938674,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0413  -1.0659  -0.3533   0.2983   2.1304
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.380579   0.152939  15.566  < 2e-16 ***
## gendermale  -0.269599   0.130290  -2.069   0.0385 *
## math        -0.010561   0.002575  -4.102  4.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8706) family taken to be 1)
##
##     Null deviance: 379.60  on 313  degrees of freedom
## Residual deviance: 357.93  on 311  degrees of freedom
## AIC: 1780.1
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##                Theta:  0.8706
##            Std. Err.:  0.0848
##
##  2 x log-likelihood:  -1772.0740
```

Answer:-

Yes, The interpretation would slightly change for negative binomial from poission model as there will be slight change in coefficient of the parameters.

standard errors have been scaled in negative binomial model and it is biased in poisson distribution. It is very high in Negative binomial then poisson distribution.

e).Provide a brief description of how the variance assumptions underlying the models specified in (a) and (c) differ from each other. What is the estimated dispersion parameter for the Negative Binomial model?

Answer:-

The Poisson distribution assumes that the mean and variance are the same. When data shows extra variation that is greater than the mean. This situation is called overdispersion and negative binomial regression is more flexible in that regard than Poisson regression. The negative binomial distribution has one parameter more than the Poisson regression that adjusts the variance independently from the mean.

Estimated dispersion parameter for the negative binomial model is: 0.8706

f).Using equations, describe how you would calculate AIC for the fitted models. Use AIC to choose between the models fitted above.

```
#Answer:-

#AIC = -2l + 2p
#where l is log likelihood of the model and p is the number of parameters

#for poisson binomial model :-
 AIC <- -2*as.numeric(logLik(fit))+2*(length(fit$coefficients))
 AIC
```

```
## [1] 2931.868
```

```
#for negative binomial model :-
 # We are adding one to the number of parameters for variance
 AIC_NB <- -2*as.numeric(logLik(fit1))+2*(length(fit1$coefficients)+1)
 AIC_NB
```

```
## [1] 1780.074
```