ST466 Assignment 3

Nishant Kumar

10 April, 2020

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

```
library(tidyverse)
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.3
## 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
```

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

```
library(MASS)
```

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

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

```
library(readr)
attendance <- read_csv("attendance.csv")
```

```
## Parsed with column specification:
## cols(
##   gender = col_character(),
##   math = col_double(),
##   daysabs = col_double()
## )
```

(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.

```
fit_pois<-glm(daysabs~.,family=poisson,data=attendance)
summary(fit_pois)
```

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

U^= exp(2.4031 -0.2548 gendermale-0.0112 math).

Interpretation to the coefficients :

Coef. –for a one unit change in the predictor variable, 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-If a student were to increase her maths test score by one point, 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.

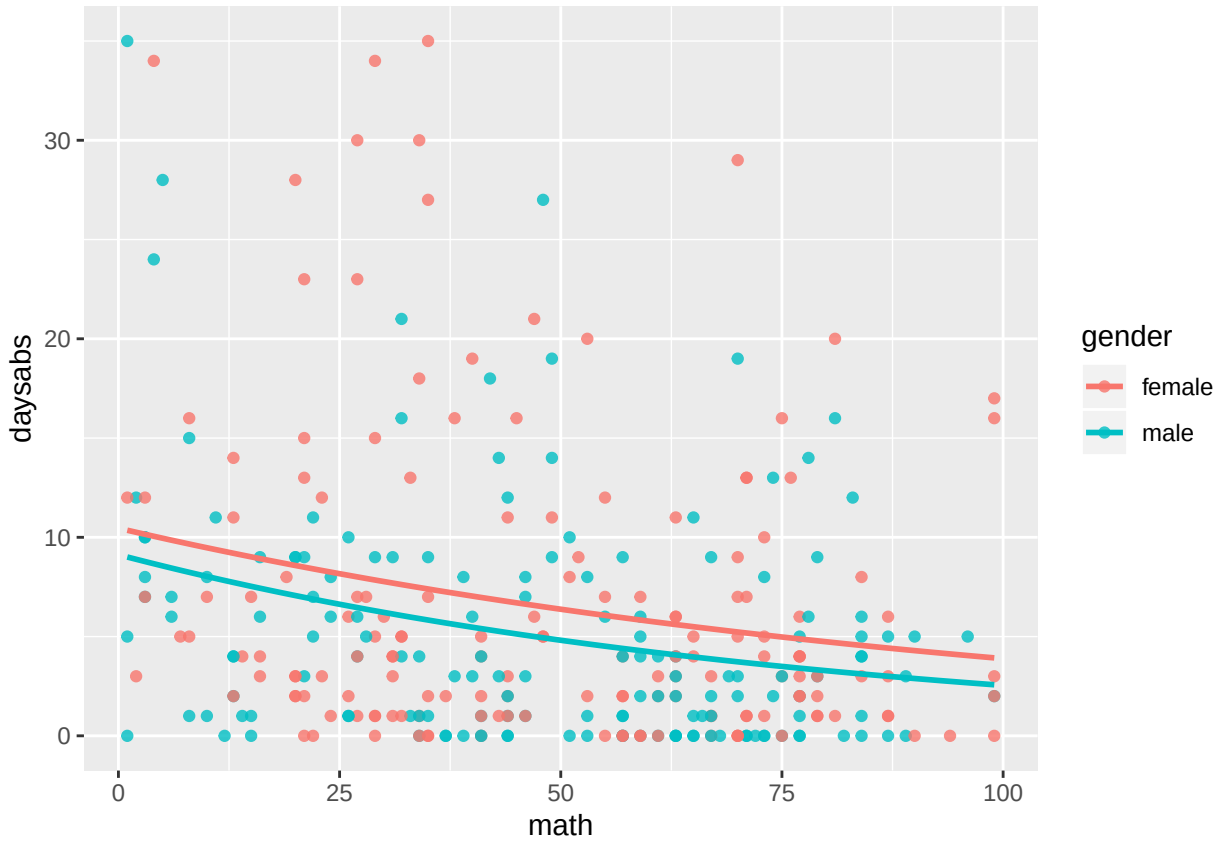
gendermale – This is the estimated Poisson regression coefficient comparing male to female, given the other variables are held constant in the model. The difference in the logs of expected counts is expected to decreease by 0.2548 unit for males compared to females, while holding the other variables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

daysabs – This is the response variable in the Poisson regression.Underneath daysabs are the predictor variables and the intercept (_cons).

(b) Plot 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.

```
attendance_res <- attendance %>%
  mutate(fit_p = predict(fit_pois, type = "response"),
res_p = residuals(fit_pois))

ggplot(data=attendance_res , aes(x=math, y=daysabs,col=gender)) +
geom_point(alpha=.8)+geom_smooth(method=glm,method.args=list(family="poisson"), se = FALSE,aes(fill=fit
```



#comment on the model fit

1>Dispersion value obtained from the residual diviance and degree of freedom is 6.567 which proves that the model has overdispresion in the model.So, the model fit is not good. 1>Female have higher abesent days in class compared to the males in the class. 2>Those student who attend the classes have higher scores from the data given though there are outlier in the data but the trend of prediction follows this trend.

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

###daysabs = exp(Intercept+b1(genderfemale) + b2(gendermale)+ b3 math)

###random component=refers to the probability distribution of the response variable (Y) ###systematic component=Intercept+b1(genderfemale) + b2(gendermale)+ b3 math

###link function=log(u)=log(r(1-pie)/pie)

(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?

```
fit_nb<-glm.nb(daysabs~.,data=attendance)
summary(fit_nb)
```

```
##
## Call:
## glm.nb(formula = daysabs ~ ., data = attendance, 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
```

yes. The coefficient of intrepretation have not changed much. yes. The standard errors have changed when compared to the poisson distribution. There is underestimation of standard error in poission distribution this may be due to dispersion.

(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?

Mean=Variance By definition, the mean of a Poisson random variable must be equal to its variance. Negative binomial model assumes that the conditional variance is not similar than the conditional mean.

Dispersion parameter for Negative Binomial=0.8706

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

```
q<- (-2*logLik(fit_pois)+2*3)
q
```

```
## 'log Lik.' 2931.868 (df=3)
```

```
q1<- (-2*logLik((fit_nb))+2*4)
q1
```

```
## 'log Lik.' 1780.074 (df=4)
```

The AIC value for the poisson is higher while the AIC value of negative binomial regression is lower.so , we choose the negative binomial regression model better the two model given.