Poisson Regression

Vazgen Tadevosyan

March 7, 2019

For this Homework, you are required to submit both Markdown and HTML files with your answers and code in it. Be sure that the .Rmd file is working, so when I run it, there would be no errors and represent the same information as HTML. Write your code and interpretations under each question. The interpretations of the results need to be written below or above all the charts, summaries, or tables. Do not remove problems from your Markdown file.

Use awards.csv dataset uploaded on Moodle to analyze the relationship between the target variable and different factors. The description of the variables is given in a separate file. Pay close attention to the names of axes, titles, and labels.

```
library(ggplot2)
library(tidyr)
library(data.table)
library(gridExtra)
library(MASS)
library(AER)
library(dplyr)
```

Problem 1. 2 pt.

- Load the file.
- Use the function str() to understand the structure of your data.
- Get rid of variables that are irrelevant for Poisson regression analysis using function select().
- Pay attention to the last column of your data. Use the separate() function to solve the problem based on data description.
- Check whether the data types are correct, if not, make appropriate corrections, assigning labels to each level according to the data description.
- Use the glimpse() function to see the structure of the final data.

```
df<-read.csv("awards.csv")
str(df)
```

```
## 'data.frame':    2500 obs. of  10 variables:
##  $ X          : int  27 47 116 281 409 463 488 491 513 532 ...
##  $ id_num     : int  56263 47181 97924 32940 89033 85747 80297 51268 20093 12830 ...
##  $ date       : Factor w/ 34 levels "2019-01-01","2019-01-02",..: 20 17 34 12 31 30 28 18 7 5 ...
##  $ awards     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ math       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ physics    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ hpw        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gender     : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ imp        : int  2 1 2 1 1 1 1 1 1 1 ...
##  $ school.prog: Factor w/ 8 levels "Private/0","Private/1",..: 1 2 2 3 1 3 3 2 2 1 ...
```

```
df<-df %>% dplyr::select(-c(X,id_num,date))
#sapply(df, function(x) sum(is.na(x)))
#summary(df)
df<-separate(df,school.prog,c("School", "Program"),sep="/")
df$imp<-ordered(df$imp,levels=c(1,2,3,4),labels=c('Not important','Neutral','Important','Very important'))
df$School<-factor(df$School)
df$Program<-factor(df$Program,levels = c(0,1,2,3),labels = c("General","Pre-Academic","Academic","Vocational"))
df<-df[!df$hpw<0,]
glimpse(df)
```

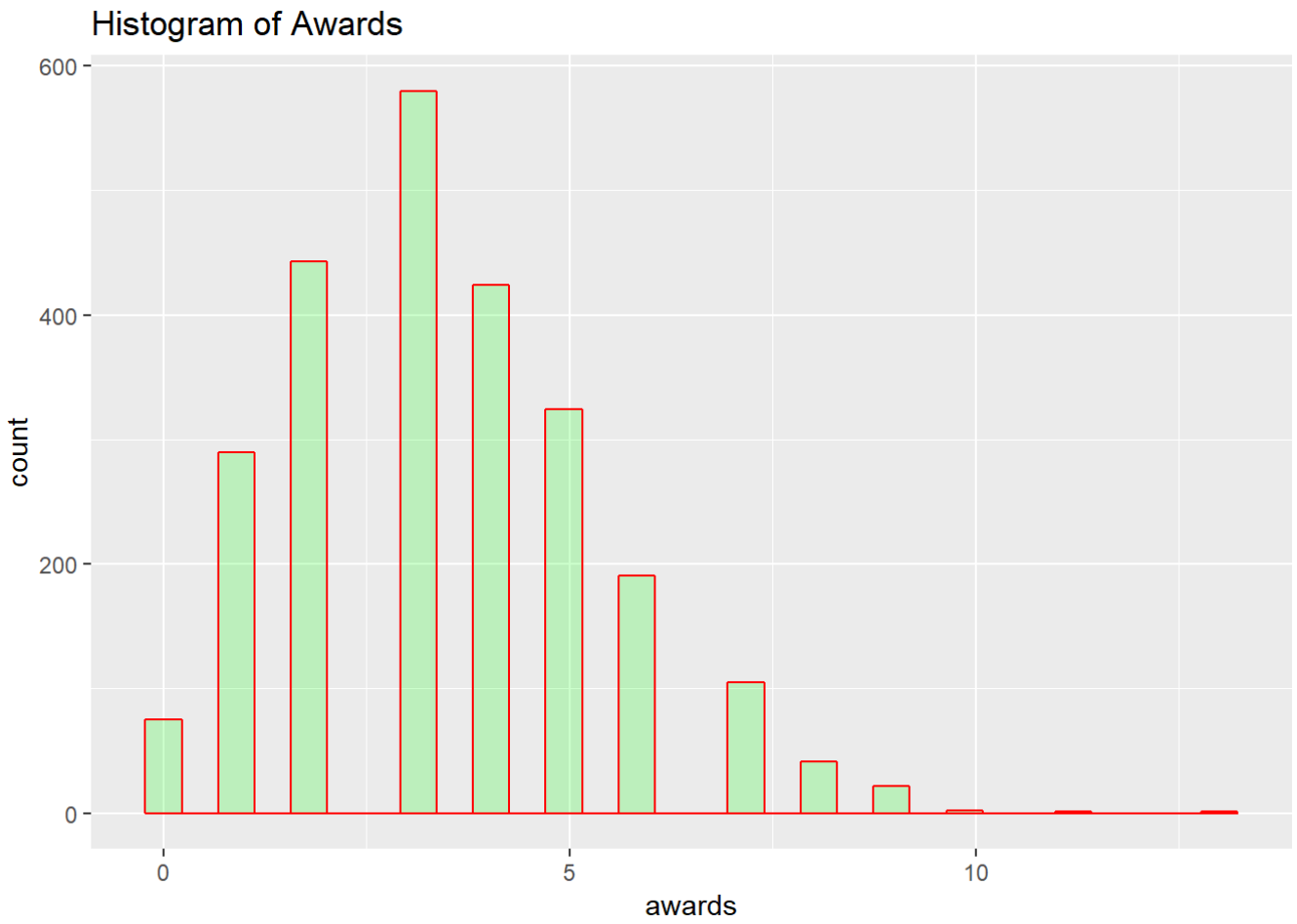
```
## Observations: 2,499
## Variables: 8
## $ awards  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ math    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ physics <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ hpw     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ gender  <fct> female, female, female, female, female, female, female...
## $ imp     <ord> Neutral, Not important, Neutral, Not important, Not im...
```

```
## $ School  <fct> Private, Private, Private, Private, Private, Private, ...
## $ Program <fct> General, Pre-Academic, Pre-Academic, Academic, General...
```

Problem 2. 4 pt.

a. Find your dependent variable for Poisson regression analysis. Plot the histogram of your target variable. Calculate the unconditional mean and variance of your target variable. What can you notice?

```
#summary(df[which(unlist(lapply(df,is.numeric)))])
ggplot(df,aes(awards))+geom_histogram( col="red", fill="green", alpha=.2)+ggtitle("Histogram of Awards")
```




```
var(df$awards)
```

```
## [1] 3.568109
```

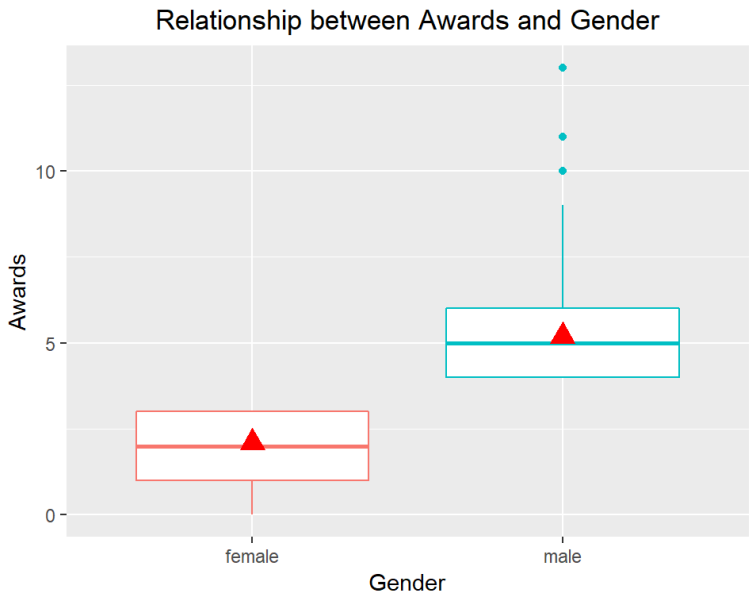
```
mean(df$awards)
```

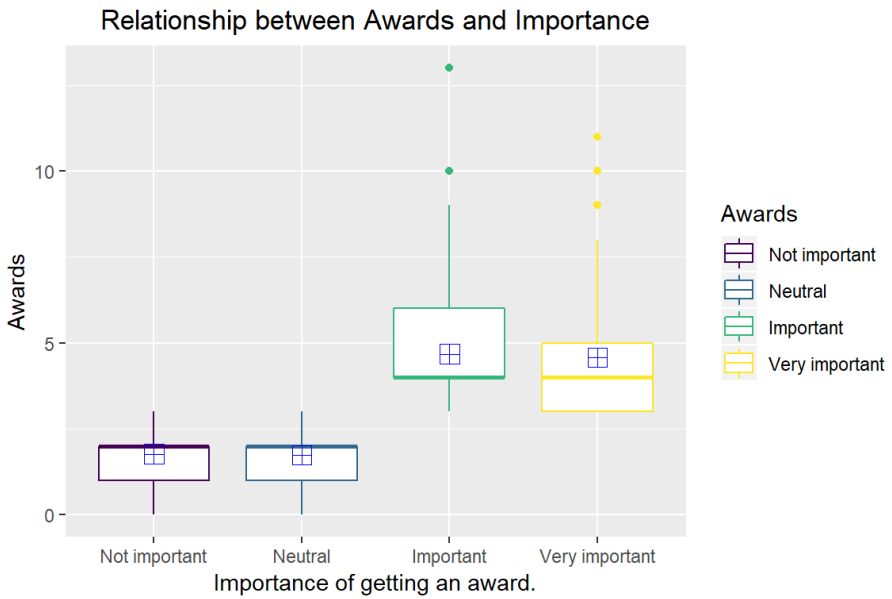
```
## [1] 3.47459
```

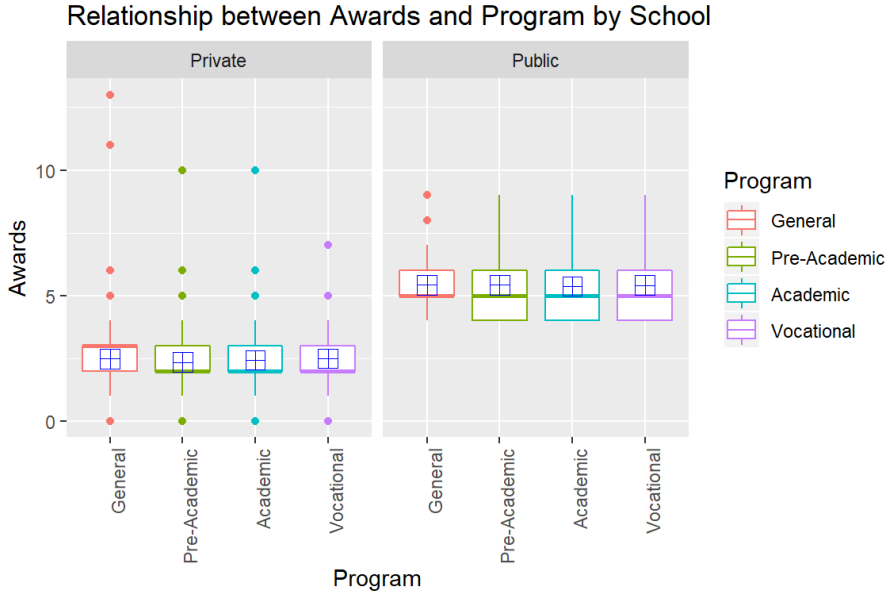
Our dependent variable is Awards as it is only count variable in our dataset.
Variance and mean of variable awards are close to each other.

b. Find the **categorical** variables which affect your target variable using boxplots. Comment on it.

```
g1=ggplot(data = df, aes(x = gender, y = awards,color=gender))+
  geom_boxplot()+
  scale_x_discrete(labels= levels(df$gender))+
  stat_summary(fun.y = "mean",geom = 'point',col='red', shape=17, size=4)+
  labs(title ="Relationship between Awards and Gender" ,y='Awards',x="Gender",color="Gebder")+theme(plot.title =
 element_text(hjust = 0.5))
g2 = ggplot(data = df, aes(x = imp, y = awards,color = imp))+
  geom_boxplot()+
  scale_x_discrete(labels= levels(df$imp) )+
  stat_summary(fun.y = "mean",geom = 'point',col='blue', shape=12, size=4)+
  labs(title ="Relationship between Awards and Importance" ,color =  "Awards",y='Awards',x="Importance of getting
 an award.")+theme(plot.title = element_text(hjust = 0.5))
g3<-ggplot(data = df, aes(x = Program, y = awards,color = Program))+
  geom_boxplot()+facet_grid(.~School)+
  scale_x_discrete(labels= levels(df$Program) )+
  stat_summary(fun.y = "mean",geom = 'point',col='blue', shape=12, size=4)+
  labs(title ="Relationship between Awards and Program by School" ,color =  "Program",y='Awards',x="Program")+the
me(axis.text.x = element_text(angle = 90, hjust = 1))
grid.arrange(g1, g2,g3,nrow=2)
```







It seems all categorical variables except Program affect our dependent variable.From the plot we can conclude that levels male(gender),important and very important(imp), Public(School) affect variable Awards positively meaning it is more likely to have more awards when a person belongs to one of this levels than when he/she does not.

c. Use group_by() and summarise() functions to conclude about conditional variances and the means of your target variable grouped by categorical variables. Comment on it: do you have the problem of overdispersion?

```
df%>%group_by(gender)%>%
  summarise(var=var(awards),mean=mean(awards))
```

```
## # A tibble: 2 x 3
##   gender   var  mean
##   <fct>  <dbl> <dbl>
## 1 female 0.833  2.10
## 2 male   1.68   5.19
```

```
df%>%group_by(imp)%>%
  summarise(var=var(awards),mean=mean(awards))
```

```
## # A tibble: 4 x 3
##   imp              var  mean
##   <ord>          <dbl> <dbl>
## 1 Not important  0.753  1.77
## 2 Neutral        0.691  1.74
## 3 Important      2.22   4.68
## 4 Very important 2.12   4.57
```

```
df%>%group_by(School)%>%
  summarise(var=var(awards),mean=mean(awards))
```

```
## # A tibble: 2 x 3
##   School    var  mean
##   <fct>   <dbl> <dbl>
## 1 Private  1.45  2.41
## 2 Public   1.59  5.41
```

```
df%>%group_by(Program)%>%
  summarise(var=var(awards),mean=mean(awards))
```

```
## # A tibble: 4 x 3
##   Program        var  mean
##   <fct>        <dbl> <dbl>
## 1 General       3.50  3.5
## 2 Pre-Academic  3.66  3.42
## 3 Academic      3.48  3.54
## 4 Vocational    3.63  3.54
```

It seems there is no overdispersion as most of cases observed variation is less than the expected variance(mean in this case).

d. Why Poisson regression is called log-linear?
A log-linear model is a mathematical model that takes the form of a function whose logarithm equals a linear combination of the parameters of the model, which makes it possible to apply (possibly multivariate) linear regression.Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. Here is the formula.

Problem 3. 7 pt.

a. Use the glm() function to perform an intercept-only Poisson regression model with your chosen (see Problem 2) target variable as the response. Use the output of your model to calculate the mean of your target variable.

```
intercept_only<-glm(awards~1,data = df,family = poisson(link = log))
summary(intercept_only)
```

```
##
## Call:
## glm(formula = awards ~ 1, family = poisson(link = log), data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
```

```
## -2.6361  -0.8602  -0.2608   0.7673   3.9058
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.24548    0.01073   116.1   <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: 2752.3  on 2498  degrees of freedom
## Residual deviance: 2752.3  on 2498  degrees of freedom
## AIC: 10106
##
## Number of Fisher Scoring iterations: 4
```

```
mean(df$awards)
```

```
## [1] 3.47459
```

```
exp(intercept_only$coefficients)
```

```
## (Intercept)
##     3.47459
```

b. Exclude from full model variables with insignificant coefficients. Show the result. Explain the meanings of coefficients of your model (at least one numeric and one categorical).

```
final_model<-glm(awards~.-Program-School,data = df,family = poisson(link = log))
summary(final_model)
```

```
##
## Call
```

```
## glm(formula = awards ~ . - Program - School, family = poisson(link = log),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.69596  -0.13081   0.03392   0.15265   1.11255
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.390263   0.027009  14.450  < 2e-16 ***
## math         0.015294   0.002048   7.468 8.16e-14 ***
## physics      0.011168   0.001812   6.163 7.15e-10 ***
## hpw          0.012509   0.005513   2.269  0.02326 *
## gendermale   0.113881   0.038033   2.994  0.00275 **
## imp.L        0.288284   0.036571   7.883 3.20e-15 ***
## imp.Q        0.009075   0.026751   0.339  0.73442
## imp.C       -0.145817   0.029788  -4.895 9.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2752.33  on 2498  degrees of freedom
## Residual deviance:  375.42  on 2491  degrees of freedom
## AIC: 7742.9
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(final_model))
```

```
## (Intercept)        math     physics         hpw  gendermale       imp.L
##   1.4773700   1.0154114   1.0112302   1.0125872   1.1206188   1.3341359
##       imp.Q       imp.C
##   1.0091164   0.8643158
```

1 unit increase in math test would increase expected number of awards by a 1.5%, while holding all other variables in the model constant. Males compared to females, are expected to have 1,12 times more awards while holding the other variable constant in the model.

$$\log(\lambda) = \beta_0 + \beta_1 X$$

c. Pick your own new observation and predict the lambda. Comment on it.

```
options(scipen=999)
predict(final_model,newdata=data.frame(math=mean(df$math),physics=mean(df$physics),hpw=mean(df$hpw),gender="male"
,imp="Very important",School="Private",Program="General"),type = "response")
```

```
##        1
## 3.679907
```

An expected number of awards is 3.679907 for a male who got average scores of tests and it is very important for him to get an award.

d. Calculate the probability of having more than 15 awards using your predicted lambda from Problem 3 c.

```
ppois(15,lambda = 3.679907,lower.tail = F)
```

```
## [1] 0.000001733085
```

Probability of getting more than 15 awards for such a person is very small(0.000001).

e. Formulate Null and Alternative hypotheses for chi-squared and deviance test. Do it both mathematically and with explanation. Conclude about goodness of fit for the full model (with significant coefficients) using chi-squared and deviance tests.

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i} = \sum_{i=1}^{n} \left[\frac{(y_i - \lambda_i)}{\sqrt{\lambda_i}} \right]^2$$

$\left[\frac{(y_i - \lambda_i)}{\sqrt{\lambda_i}} \right]$ *is called Pearson residual.*

- *Where y is actual number for the dependent variable.*
- $\lambda_i = e^{x_1 \hat{\beta}}$ *is the predicted value using Poission regression. So, smaller number of* χ^2 *indicates that predicted and actual values are close to each other, thus better the model fits the data.*

H_0 *: The model fits the data well* $y_i = \lambda_i$
H_1 *: The model doesn't fits the data well* $y_i \neq \lambda_i$
if $\chi^2 > \chi^2$*critcal then we reject* H_0

```
degree_of_f<-df.residual(final_model)
chi_sq<-sum(resid(final_model,type="pearson")^2)
pchisq(chi_sq,df=degree_of_f,lower.tail = F)
```

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

We fail to reject H_0 *so model fits well.*

Another way to make assumption for goodnes of fit is to look Deviance statistics.

$$D = 2 * \sum_{i=1}^{n} \left[y_i log\left(\frac{y_i}{\lambda_i}\right) - (y_i - \lambda_i) \right]$$

If the model fits well the observed values y_i *will be close to their predicted means* λ_i*. Thus deviance will be small.* H_0*The model fits the data well* $D = 0$
H_1*The model doesn't fits the data well* $D \neq 0$ *In this case we compare our specified model with saturated model where all predicted values exactily match actual values.*

```
pchisq(q=375.42,df=2491,lower.tail = F)#numbers are from summary.
```

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

We fail to reject H_0 as p-values is less than α(0.05).

Problem 4. 7 pt.

a. What is the equidispersion in Poisson regression? Why do we need to avoid overdispersion?
The only parameter the mean occurence rate λ_I describes mean and variance of the distribution at the same time:

$$E(Y_i) = VAR(Y_i) = \tilde{A} \times \hat{A}\text{»}_i$$

This phenomenon is called equidispersion.` When we have overdispersion we cannot use anymore Poisson regression as it will not provide proper values for significant tests. An alternative model with additional free parameters may provide a better fit.

b. Add to your data a new (created) variable with the problem of unconditional overdispersion. Show the problem by computing the average and variance of your variable. (Your variable needs to have a similar meaning to your target variable.).

```
set.seed(1)
df$articles<-rnbinom(nrow(df),mu=5,size = 1)
var(df$articles)
```

```
## [1] 30.28106
```

```
mean(df$articles)
```

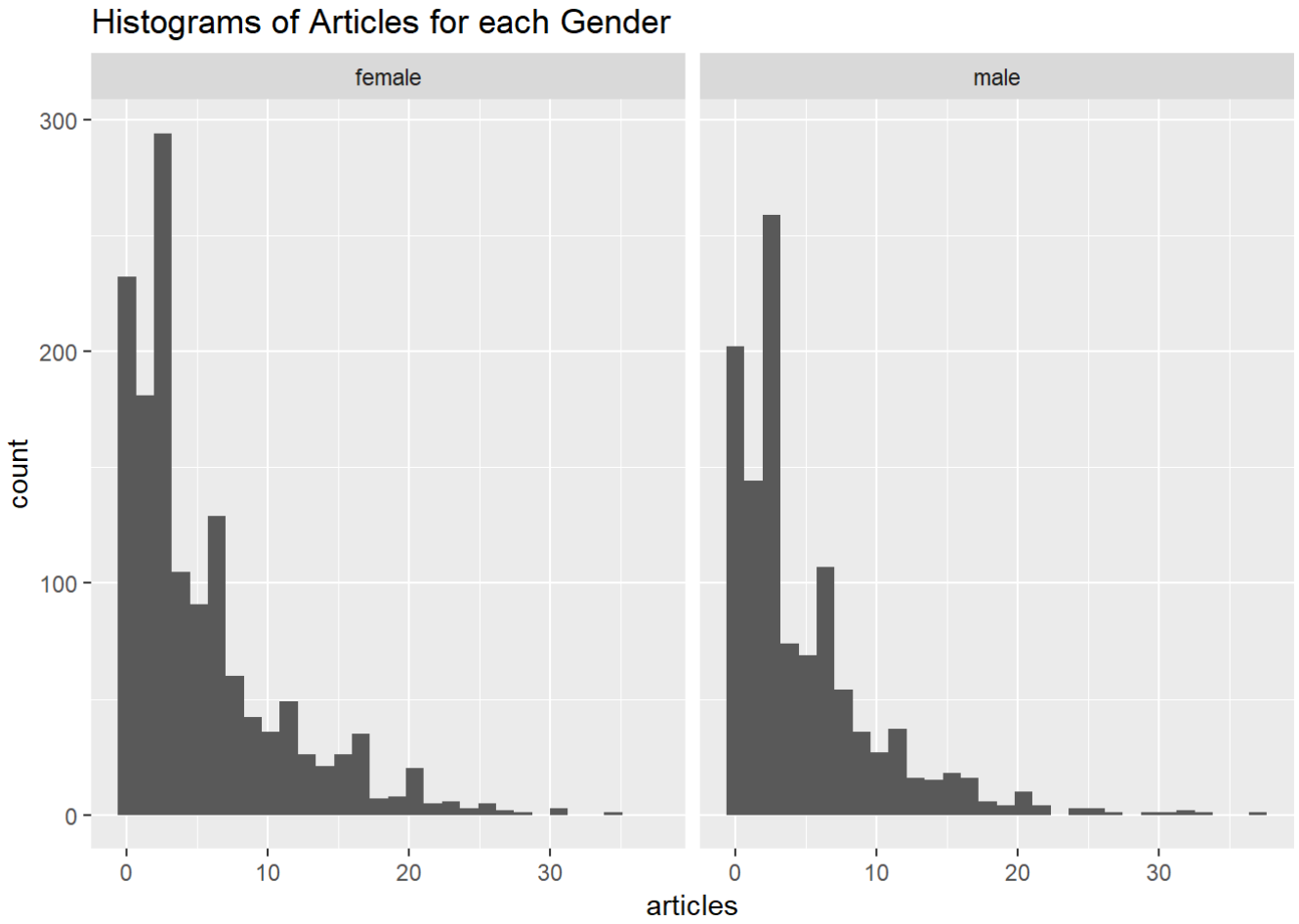
```
## [1] 5.077231
```

Now there is a new generated count variable in the dataframe representing number of articles written by students, where variance of number of atricles is bigger than average of it.

c. Plot the histogram of your created variable grouped by a nominal variable. Does your variable have conditional overdispersion with the nominal variable in your data?

```
ggplot(df)+geom_histogram(aes(articles))+facet_grid(.~gender)+ggtitle("Histograms of Articles for each Gender")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It seems that articles is distributed similarly for each level, so there is not conditional overdispersion

d. Run the model with the new variable as a response. Your model must contain only significant coefficients.

```
new_model<-glm(articles~.-Program-physics-gender-School-hpw,data = df,family = poisson(link = log))
summary(new_model)
```

```
##
## Call:
## glm(formula = articles ~ . - Program - physics - gender - School -
##     hpw, family = poisson(link = log), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3243  -2.1921  -0.9267   0.9244   9.1352
##
## Coefficients:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  1.604510   0.024361  65.863 < 0.0000000000000002 ***
## awards      -0.032504   0.019171  -1.696              0.08998 .
## math         0.006656   0.002931   2.271              0.02314 *
## imp.L       -0.072775   0.025479  -2.856              0.00429 **
## imp.Q       -0.007656   0.018020  -0.425              0.67094
## imp.C        0.063349   0.020404   3.105              0.00190 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13495  on 2498  degrees of freedom
## Residual deviance: 13472  on 2493  degrees of freedom
## AIC: 20359
##
## Number of Fisher Scoring iterations: 5
```

e. Use the function dispersiontest to find out overdispersion. Formulate Null and Alternative hypotheses for trafo=Null (mathematically and explain it). Do you have an overdispersion?

```
dispersiontest(new_model,trafo = NULL,alternative = "greater")
```

```
##
##  Overdispersion test
##
## data:  new_model
## z = 19.038, p-value < 0.00000000000000022
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##    5.93939
```

$$Var(y_i|x_I) = (1+a) * \lambda_i$$

$$H_0 : There\ is\ notoverdispersion\ a \leq 0 H_1 : There\ is\ overdispersion\ a > 0$$

P-values is a very small number so we reject H_0 and we can cliam that there is overdispersion.

f. Run the negative binomial and quasi-Poisson model. Show only coefficients. Find the best model based on deviance and AIC. Which is the best model?

```
mod_q<-glm(articles~.-Program-physics-gender-School-hpw, data = df,family = quasipoisson(link = log))
mod_n<-glm.nb(articles~.-Program-physics-gender-School-hpw, data = df)
data.frame(coef(mod_q),coef(mod_n))
```

```
##              coef.mod_q.  coef.mod_n.
## (Intercept)  1.604510330  1.605245910
## awards      -0.032504228 -0.034248730
## math         0.006656038  0.006903808
## imp.L       -0.072774867 -0.072003301
## imp.Q       -0.007656143 -0.007580207
## imp.C        0.063348617  0.063987981
```

```
##  (Intercept)       awards         math        imp.L        imp.Q
##  1.604510330 -0.032504228  0.006656038 -0.072774867 -0.007656143
##        imp.C
##  0.063348617
```

```
coef(mod_n)
```

```
##  (Intercept)       awards         math        imp.L        imp.Q
##  1.605245910 -0.034248730  0.006903808 -0.072003301 -0.007580207
##        imp.C
##  0.063987981
```

```
data.frame(deviance(new_model),deviance(mod_q),deviance(mod_n))
```

```
##   deviance.new_model. deviance.mod_q. deviance.mod_n.
## 1            13471.71        13471.71        2841.559
```

```
data.frame(deviance(new_model),deviance(mod_q),deviance(mod_n))
```

```
##   deviance.new_model. deviance.mod_q. deviance.mod_n.
## 1            13471.71        13471.71        2841.559
```

```
mod_n$aic
```

```
## [1] 13590.44
```

```
new_model$aic
```

```
## [1] 20358.69
```

The best model is Negative binomial model as values of AIC and Deviance were the least ones.

g. Why does not quasi-Poisson model have AIC?

$$AIC = 2k - 2ln(\hat{L})$$

k is number of predictor and L is $\hat{L}$ be the maximum value of the likelihood function for the model. *AIC is calculated using log-likelihood function ,however quasi-Poisson model does not use log likelihood estimation it use quasi-likelihood , so it does not provide AIC score.* The quasi-likelihood approach is based on this fact, requiring that only the mean and variance of the distribution be specified. And then the quasi-likelihood estimates are obtained through the solution of the likelihood equations for GLMs. As focusing in the quasi-Poisson model, a dispersion parameter is included, giving us: $V(\mu) = \phi\mu$ This new parameter can be estimated with: $\hat{\phi} = \frac{X^2}{n-p}$ where $X^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$ This means that que quasi-Poisson model is equivalent to a Poisson model, with the $\hat{\beta}$ standard errors multiplied by $\sqrt{(\hat{\phi})}$But it's not exactly Poisson because we do not have the property of mean = variance. This kind of model is usually considered when one wants to account for overdispersion in count data.

So I think the main difference is that $\hat{\phi})$ which is included in quasilikelihood but not in log-likelihood estimation from which we get AIC.

Please, make brief and meaningful conclusions.