Poisson Regression_Scholarship Prediction

Thien Dinh Van

14/02/2023

1. Introduction

Scholarship is one of the most important sources of financial aid for any students. In fact, it can provide a significant amount of assistant towards tuition fee and other expenses. There are a number of factors which affect to how many scholarship offers for a student.

In this report, our objective is to use Poisson Regression for looking at whether there are different number of offers in various factors at Okanagan College.

Using Poisson Regression are required some of conditions as below:

- `Poisson Response`: The response variable is a count per unit of time or space, described by a Poisson distribution.
- `Independence`: The observations must be independent of one another.
- `Mean = Variance`: By definition, the mean of a Poisson random variable must be equal to its variance.
- `Linearity`: The log of the mean rate, log (y), must be a linear function of x.

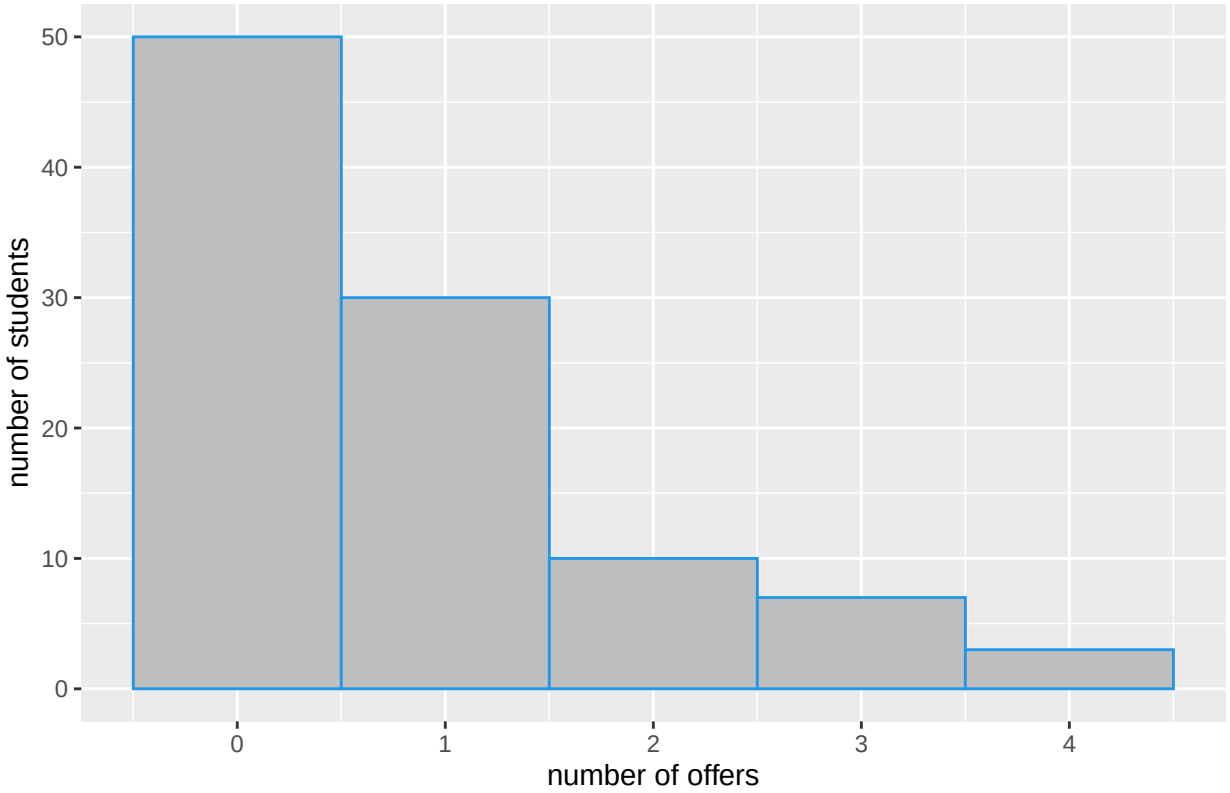
2. Data

The data set has 100 observations and 4 variables, including integer, numeric, character variables. The variables as the detail below:

- `offers` = number of scholarship offers of a student applying at Okanagan College received.
- `division`: application address, `A` - British Columbia, `B` - Canada (outside of BC), `C` - Outside of Canada.
- `exam`: college entrance exam score (measured from 0 to 1000) -`sex`: gender of student, `M` - male, and `F` - female.

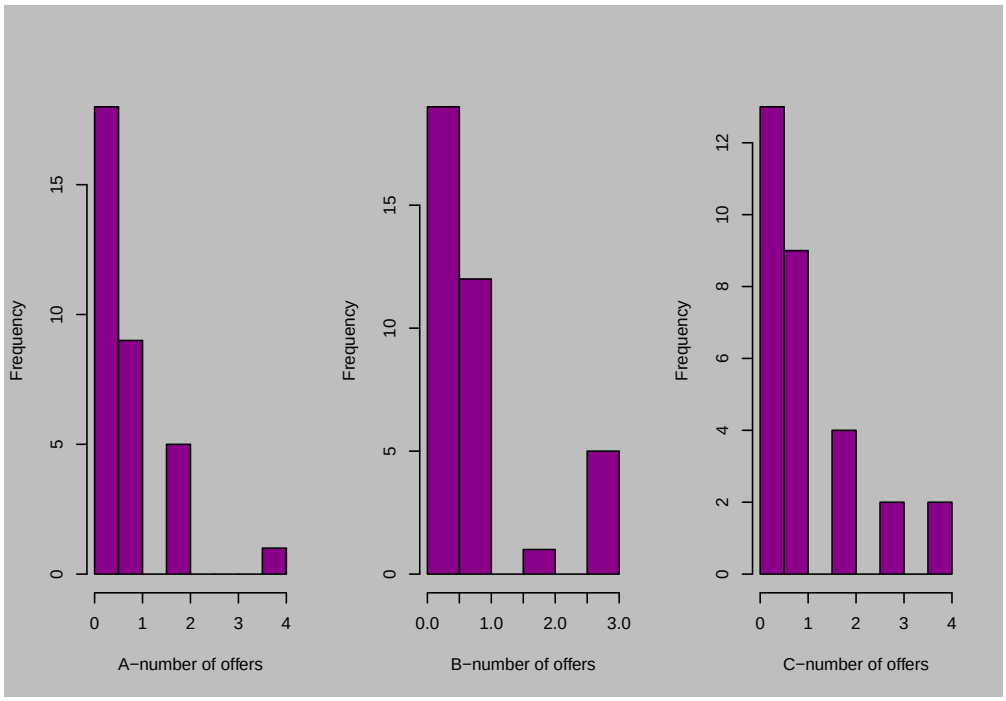
2.1. Exploratory Data Analysis

Figure 2.1: Histogram of number of scholarship offers at Okanagan College.



A histogram of number of scholarships offers of student at OC, figures 2.1, illustrate the difference in the distribution of offer's number. A larger number of students who has no offers and there is a significant difference between the number of students who has 1 offer and 4 offers. Therefore, Poisson Regression is make sense for using to model in this data set. Furthermore, response variable is count as number of offers, so we will consider λ , the average number of offers and $\lambda = 0.83\$$.

`Figure 2.2: Histogram the number of offers for student in BC, outside BC (Canada), and outside Canada.`



Looking at figure 2.2, the graphs show right skewed in the distribution, and there is the difference in the offer's number in the various application address. Hence, these plots suggest for condition 1 of Poisson Regression which is Poisson Response.

For Poisson Regression, the mean of a Poisson random variable must be equal to its variance. Therefore, we will discover the changing of offer's number and gender, as well as location. As regards to table 2.1 and 2.2, we can see the approximately similar between mean and variance;however, there is a tiny difference between mean and variance in number of scholarship offers of student outside Canada. Although, this is not really significant.

`Table 2.1: Compare mean and variance of number of scholarship offers of OC students by gender.`

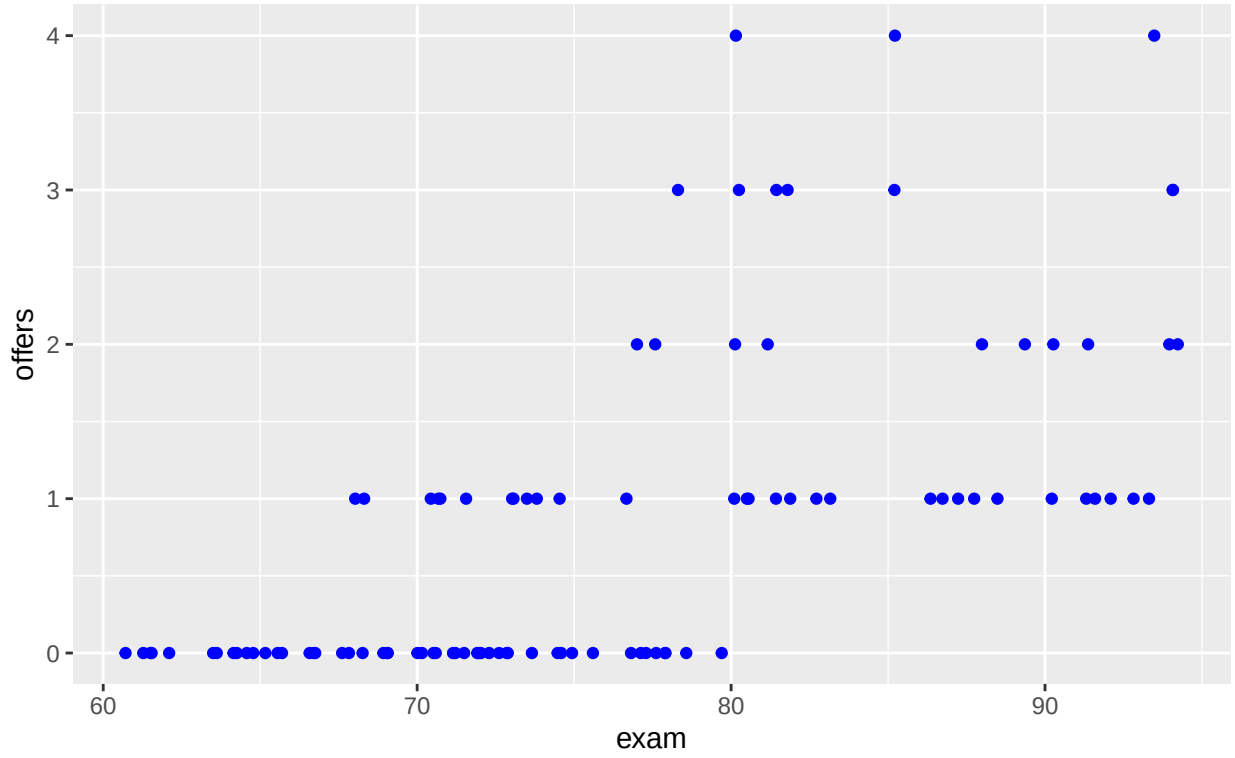
Sex	Mean	Variance	n
M	0.84	1.1167	50
F	0.82	1.1710	50

`Table 2.2: Compare mean and variance of number of scholarship offers of OC students by application address.`

Location	Mean	Variance	n
A	0.70	0.91	33
B	0.78	1.06	37
C	1.03	1.48	30

For checking linearity assumption, we will get the relationship between exam score and number of offers as the plot from figure 2.3 for explaining. In fact, Poisson Regression model implied that $\log(\lambda)$, not the mean of offers number λ. So the linear function as example: $\log(\lambda) = \beta_0 + beta_1\text{exam}$. However, we can not calculate the $\log(\lambda)$ now, but we can use the plot below and guess that linearity assumption is not satisfy.

Figure 2.3: The retionship between number of scholarship offers and exam score at Okanagan College



2.2. Fit a Poisson model to the data

Table 2.3: Fit measures from full model

Coefficients	Estimate	P-value
Intercept	-7.34	1.31e-11
divisionB	0.07	0.805
divisionC	0.28	0.308
exam	0.09	4.12e-12
sexM	0.11	0.610

Using log link function and we can get the full model of Poisson Regression in this data set is:

$log(offers) = -7.34 + 0.07 divisionB + 0.28 divisionC + 0.09 exam + 0.11 sexM$

3. Explain the Poisson model

- Looking first at coefficient of exam which is $\beta_{exam} = 0.09$. Because $0.09 > 0$, then $e^{0.09}$ and is referred to as a ratio. For this case, it explain that the probability of receiving an offer based on the entrance exam will increase around 9% by every unit increase in exam score.
- As regards to address application, we can see that, the probability of receiving an offer of an international students who outside BC (in Canada) and outside Canada will going up 7% and approximately 32%, respectively.

- Interestingly, if a student is male, the probability if receiving an offer is increase 12% over a student who is female,

4. Improvement and testing the model

First of all, we can see that, the p-value of `exam` = variable shows that is significant. Therefore, we will keep this variable in the final model.

Next, we will consider with `sex` and `division`, let's make a guess that `sex` is not significant in the model. Then, we will use the ANOVA test between 2 models (full model and reduce model(remove `sex`)). As a result, the p_value in the ANOVA test above = 0.61 » 0.05, so we can reject the `sex` variable in the model (null hypothesis is the coefficient of `sex` = 0).

Similarity, we will get the final model which is only `exam` is predictor variable. And, we will make the ANOVA test, as the same above. Hence, we get the p_value = 0.58 » 0.05, so we will reject the `division` variable in the final model (null hypothesis is the coefficient of `division` = 0).

After analysis, we will get the final model as below:

$log(offers) = -7.16 + 0.087exam$

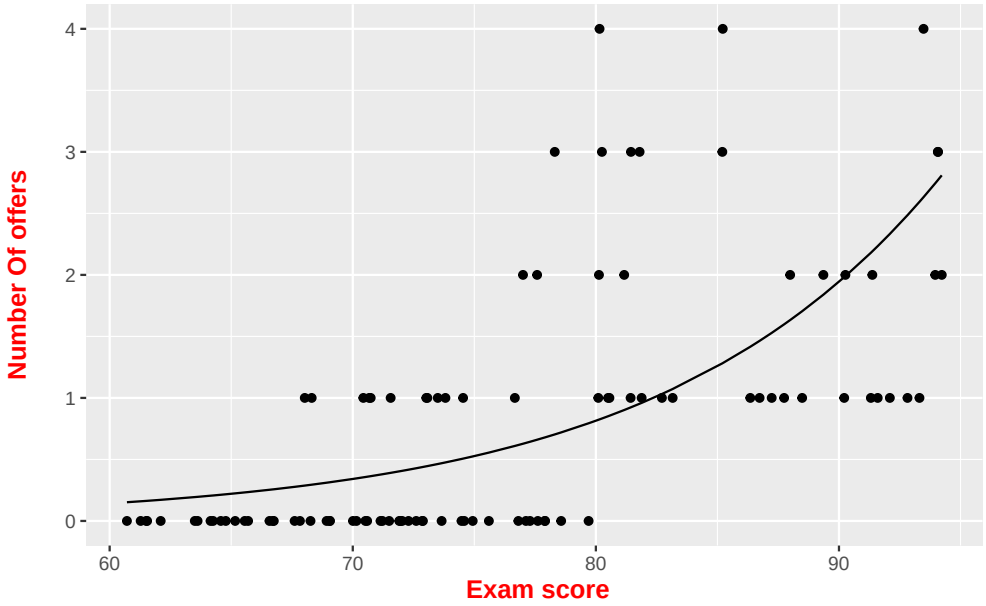
Moreover, we will using likelihood ratio test to conclude any significance:

- H_0: Only `exam` is significant in the model
- H_A: Exam is not significant in the model

The p-value from lrtest() functions for Likelihood ratio test is p-value = 0.6101 » 0.05, so, we can conclude that only `exam` is significant in Poisson Regression model.

5. Plot the data, final model, and conclusion.

Figure 5.1: Poisson Regression: Comparing Number of Offers To Exam Score



From the graph, we can conclude some primary points that:

- There is a positive relationship between exam score and number of offers.
- The number of scholarship offers will increase significantly if the exam score between over 80.

In conclusion, exam score is one of the most important factor which affect to success in applying scholarship offers. Therefore, a great idea for student is that concentrate in studying and try to get higher score in exams, as well as other activity.

6. Appendix

6.1. Data set

```
library(readxl)
oc <- read.csv("oc.csv")
```

6.2. EDA

```
par(mfrow=c(1,3), bg ="grey")
hist(oc[oc$division=='A',]$offers, main= "",xlab = "A-number of offers", col = "darkmagenta")
hist(oc[oc$division=='B',]$offers, main= "",xlab = "B-number of offers", col = "darkmagenta")
hist(oc[oc$division=='C',]$offers, main= "",xlab = "C-number of offers", col = "darkmagenta")
```

```
ggplot(oc, aes(x=exam, y = offers)) +
  geom_point(color ="blue") +
  ggtitle("Figure 2.3: The retionship between number of scholarship offers and exam score at Okanagan C
```

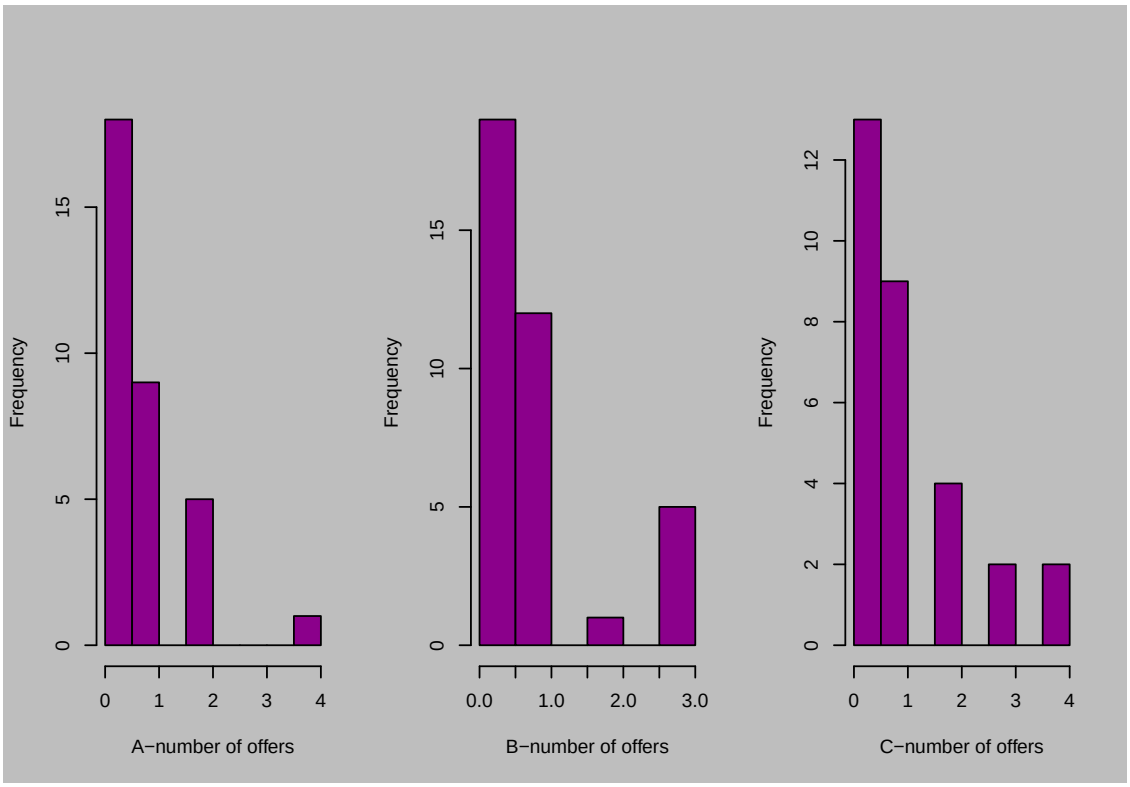
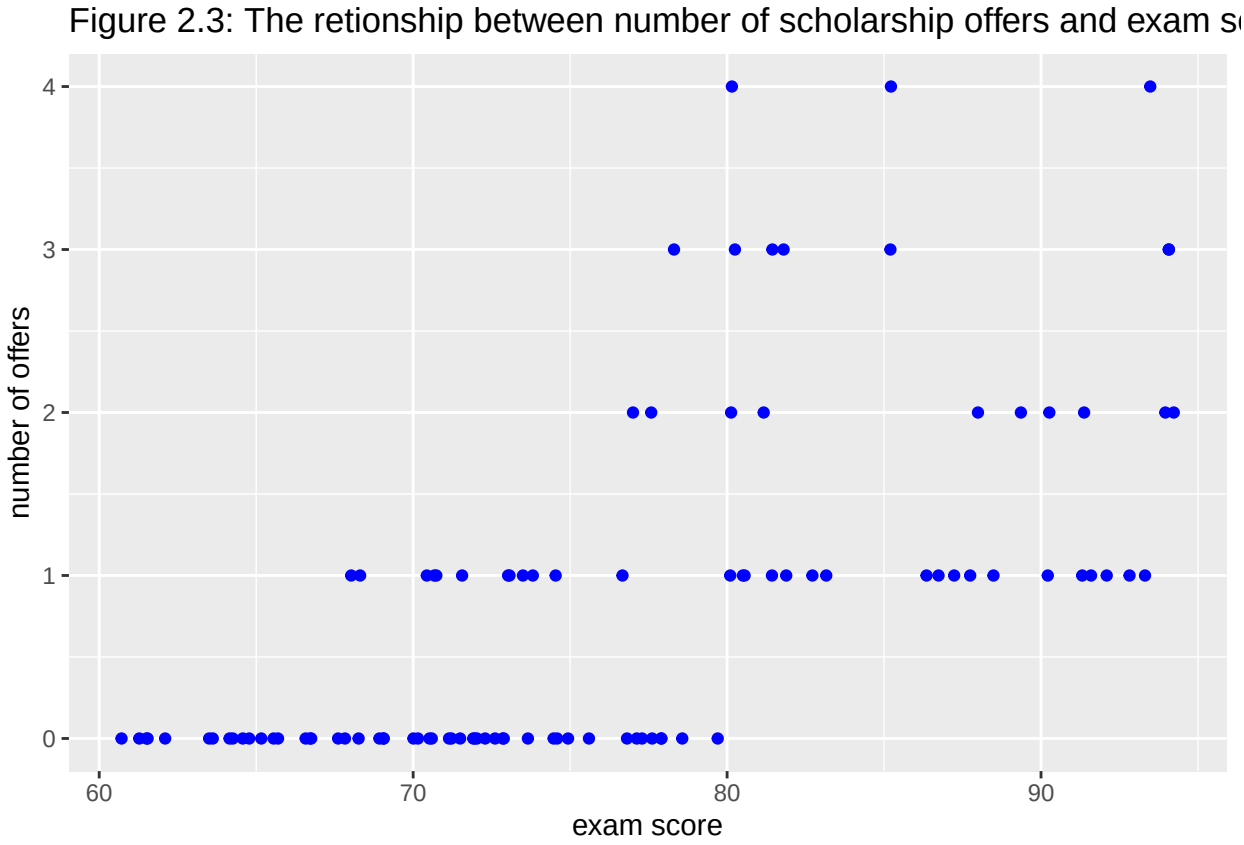


Figure 1: Histogram the number of offers for student in BC, outside BC (Canada), and outside Canada.



6.3 Poisson Regression models

```
# full model
oc_model = glm(offers ~., family = poisson(link ="log"), data = oc[,-1])
summary(oc_model)
```

```
##
## Call:
## glm(formula = offers ~ ., family = poisson(link = "log"), data = oc[,
##     -1])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4505  -0.8732  -0.5942   0.3194   2.5868
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.33864    1.08437  -6.768 1.31e-11 ***
## divisionB    0.06888    0.27948   0.246    0.805
## divisionC    0.28276    0.27719   1.020    0.308
## exam         0.08693    0.01254   6.933 4.12e-12 ***
## sexM         0.11392    0.22343   0.510    0.610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 138.069  on 99  degrees of freedom
## Residual deviance:  82.481  on 95  degrees of freedom
## AIC: 209.36
##
## Number of Fisher Scoring iterations: 5
```

```
# reduce model
reduce_model =  glm(offers ~ division + exam, family = poisson(link ="log"), data = oc[,-1])
summary(reduce_model)
```

```
##
## Call:
## glm(formula = offers ~ division + exam, family = poisson(link = "log"),
##     data = oc[, -1])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3376  -0.8612  -0.6167   0.2442   2.6496
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.21183    1.04877  -6.876 6.14e-12 ***
## divisionB    0.07156    0.27935   0.256    0.798
## divisionC    0.26906    0.27585   0.975    0.329
## exam         0.08614    0.01236   6.969 3.20e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 138.069  on 99  degrees of freedom
## Residual deviance:  82.741  on 96  degrees of freedom
## AIC: 207.62
##
## Number of Fisher Scoring iterations: 5
```

```
# ANOVA test
anova_1 = anova(reduce_model,oc_model,test ="Chisq")
library(knitr)
kable(anova_1)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
96	82.74063	NA	NA	NA
95	82.48060	1	0.2600349	0.6100962

```
# final model
final_model =  glm(offers ~ exam, family = poisson(link ="log"), data = oc[,-1])
summary(final_model)
```

```
##
## Call:
## glm(formula = offers ~ exam, family = poisson(link = "log"),
##     data = oc[, -1])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2601  -0.8852  -0.6244   0.2118   2.5049
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.16456    1.04220  -6.874 6.22e-12 ***
## exam         0.08700    0.01234   7.048 1.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 138.069  on 99  degrees of freedom
## Residual deviance:  83.806  on 98  degrees of freedom
## AIC: 204.68
##
## Number of Fisher Scoring iterations: 5
```

```
anova_2 = anova(final_model,reduce_model,test ="Chisq")
kable(anova_2)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
98	83.80559	NA	NA	NA
96	82.74063	2	1.064955	0.5871486

```
# likelihood ratio test
library(lmtest)
lrtest(oc_model, final_model)
```

```
## Likelihood ratio test
##
## Model 1: offers ~ division + exam + sex
## Model 2: offers ~ exam
##   #Df   LogLik Df Chisq Pr(>Chisq)
## 1   5  -99.679
## 2   2 -100.341 -3 1.325     0.7232
```