STAT2402 ANALYSIS OF OBSERVATIONS
Journal Article

An Observational Analysis of Patient Complaints Among Emergency Department Doctors

Cedrus Dang

2024-10-23

Abstract: This study aims to improve a hospital's management system by examining how specific demographic factors of doctors of a hospital emergency department correlate with the number of patient complaints received in a year. Specifically, patient complaints are critical to healthcare service quality and present patient satisfaction.

By using the records of 94 doctors over a year, this study investigates the impact of variables, including the number of patient visits, residency training status, gender, hourly income, and total working hours on complaint counts. As the target variable consists of count values, with a high proportion of zero cases and over-dispersion can present in the dataset, this study employed four regression models, specifically Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP), and Zero-Inflated Negative Binomial (ZINB) models. Model selection was conducted based on Akaike Information Criterion (AIC) values, while Vuong's test was used to compare non-nested models, aiming to identify the optimal model that balances accuracy and simplicity. Additional performance valuation has also been done to ensure the accuracy of the selection.

The results show that the ZIP model is optimal for the data, addressing over-dispersion and zero inflation more effectively than other models. The model has the lowest AIC value among models and is statistically significant in Vuong's test results compared to Poisson and NB models. The ZIP model indicated that the number of patient visits positively predicts complaint counts, with the complaint case increasing by 0.14% for each additional patient's visit, suggesting that doctors with a higher patient load may receive more complaints than others. Meanwhile, a higher doctor's hourly income will reduce complaints, with the expected number of complaints decreasing by 0.79% for each additional dollar of hourly income. In contrast, residency training, gender, and total working hours were statistically insignificant in predicting the number of complaints in this model.

There are also limitations in the study. The ZIP model's zero cases probability estimation part has no significant variables, and the intercept is also insignificant. This indicates that this part can not capture any factors that impact the probability of not having any complaints. This suggests other factors not included in this dataset affect complaint rates. Another potential reason is the small sample size and limited timeline of one year for this data. Those findings suggest that hospitals can improve patient satisfaction by effectively distributing the number of patient visits to each doctor and their payment. Furthermore, future studies could also be conducted to enhance this analysis by extending the data collection time frame with additional factors in order to have a better analysis.

I. INTRODUCTION

In medical facilities, patient complaints are widely recognized as a primary indicator of service quality and patient satisfaction. Therefore, they can be used to identify system flaws and improve service and staff quality [1] [2] [3].

Specifically, a high number of complaints often signals that the quality of service or the attitude of doctors may not meet the standards of an effective healthcare system [2]. In the emergency department, besides the quality of facilities and service, communication between medical staff and patients remains a significant factor in patient dissatisfaction, indicating the demand for analysis of the relationship between complaints and doctor's demographic factors [3].

Based on the significant role of this figure, this study seeks to determine how the demographic factors of the doctors in emergency service influenced the number of complaints they received. This will help determine what can be improved to increase patients' satisfaction in the emergency department of hospitals. This objective will be conducted by determining the effect of several demographic variables, such as the number of visits, residency training status, gender, revenue, and working hours, on the number of complaints received by each doctor, with this expressed as counts of complaints in a year. The utilized data is the records of 94 doctors who worked in an emergency service at a hospital in a year.

With medical outcomes expressed as counts, Poisson, Negative Binomial (NB), Zero Inflated Poisson (ZIP), and Zero Inflated Negative Binomial (ZINB) regression are popular methods for an accurate analysis. This expectation is based on the successful application of those models in multiple studies in medical systems that have count outcomes in their models [4] [5] [6][7].

In one study of those studies [4], the authors applied the Poisson regression model to analyze hospitalization data when outcomes are expressed as counts, in this case, the number of days the patients were hospitalized. Then, they conducted over-dispersion testing (where the variance exceeds the mean) and zero inflation (datasets with excess zeros) in the models. When the authors saw over-dispersion and zero-inflation were confirmed, they deployed other models as potential solutions for better fitting. They then compared them and concluded that NB and ZINB models best fit their counts data, which had over-dispersion and zero inflation.

Another study [5] applied additional Poisson and Negative Binomial Hurdle Models after Poisson, NB, ZIP, and ZINB. Those additional models are similar to the ZIP model but assume that all zeros are of the same origin (structural zeros) and use a truncated count model for non-zero counts. However, as expected, the additional models did not perform as well as the ZINB model, and ZINB was still the optimal choice. Meanwhile, Generalized Estimation Equations models and Zero-Inflated Generalized Poisson were used in a different study [6] to handle over and under-dispersion in the data set.

In this article, we will conduct the study using Poisson, NB, ZIP, and ZINB models due to their high potential in solving this analysis. The next section will describe the statistical methodology used, followed by the results section and the discussion section for the findings. The final part is the appendix, which will show the R codes used for this study.

II. METHODOLOGY

In this study starts with exploratory data analysis (EDA) to understand each variable's attributes and handle any special phenomena, such as high correlation and zero inflation. Only valid variables will be used for modelling.

After that, four regression models that include all variables and their interactions will be fitted and the n simplified using the Akaike Information Criterion (AIC) values following this order:

1. **Poisson model:** This model will also be used for zero inflation and over-dispersion tests.
2. **Negative Binomial model (NB):** If over-dispersion is discovered, this model will be applied with the expectation that it will fit better with the data set.
3. **Zero Inflated Poisson (ZIP):** This model will be used if zero inflation is detected.
4. **Zero Inflated Negative Binomial (ZINB):** If over-dispersion and zero inflation are both approaches, this model will be applied to get a better fit.

Model selection is based on AIC for model comparison, while Wald's test was used to assess variable significance. Pearson's Test for over-dispersion and Residual Analysis with diagnostic plots will be used to diagnose the models. Additionally, Vuong's test has also been used to compare two groups of non-nested models, in this case, Poisson and NB against ZIP and ZINB, to have a better valuation on the ability to provide a significantly better fit to the data than the other models. The optimal model will then be interpreted to determine the variables' effect on the number of complaints received.

All statistical analysis will be conducted in the R statistical environment, and statistical significance will be taken at α = 0.05 (5%). AIC stepwise selection will only be used to reduce the Poisson and NB models. The library used is "MASS," with the method being backward selection. For ZIP and ZINB, the AIC stepwise selection is unsuitable since there are two parts of the model: the zero inflation prediction for zero outcomes and the count's prediction. Therefore, manual selection using AIC values, focusing on interactions of variables that have a high chance of correlation with each other or the outcomes, such as visits, revenue, and hours, will be conducted. For other analyses, the R code will be shown in the Appendix section of this article.

III. RESULTS

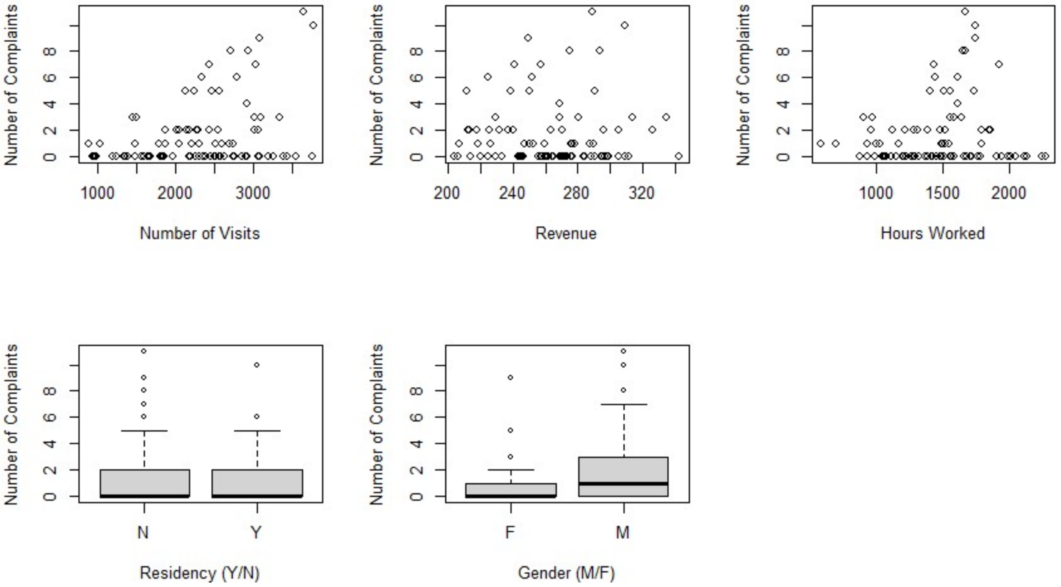
The description and summary of the variables in the data set are given in Table 1.

Table 1: Variables Summary

Variable	Description	Summary
complaints	Number of complaints against the doctor in the previous year	Min: 0, Mean: 1.6, Max: 11, SD: 2.5
visits	Number of patient visits	Min: 879, Mean: 2271, Max: 3763, SD: 723.9
residency	Doctor in residency training	Yes (Y) = 49, No (N) = 45
gender	Gender of the doctor	Male (M) = 37, Female (F) = 57
revenue	Doctor’s hourly income (dollars)	Min: 203.9, Mean: 263.7, Max: 342.9, SD: 30.9
hours	Total hours worked by the doctor in a year	Min: 589, Mean: 1469, Max: 2269, SD: 351.4

The data set summary in Table 1 shows the mean of complaints is only 1.6, while the max value is 11, with SD 2.5. This indicates the potential appearance of zero inflation in the data. A further investigation using scatterplots is shown in Figure 1:

Figure 1: Scatterplots of Complaints against Other Variables



The plots in Figure 1 show a high number of zero complaints cases across all factors and are not concentrated. This indicates a high chance of zero inflation in the data.

After fitting the Poisson model and reducing it using the AIC value, the Pearson test for over-dispersion showed a statistics value of 2.74, with a P-value of 0.013, lower than the significance level of 5%, indicating a dispersion phenomenon in the data set. Since both overdispersion and zero inflation were detected, the NB, ZIP, and ZINB models were developed with the expectation of a better fit.

Using the AIC values comparison method to find optimal models and additional Vuong’s test to compare two non-nested models, in this case, Poisson - NB against ZIP - ZINB, to have better valuation on the ability to provide a significantly better fit to the data than the other models. The comparison results are shown in the follow table:

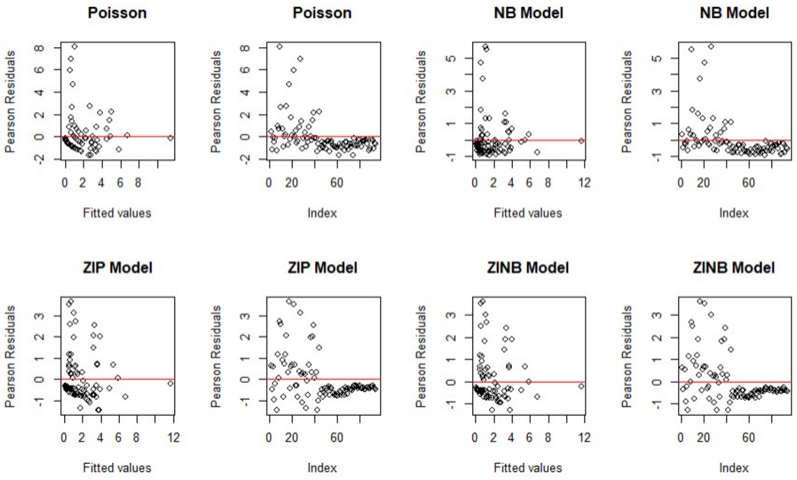
Table 2: Comparison Tests's Results

Model	Residual deviance	Log-likelihood	AIC	Ranking
Poisson	166.00 on 83 Df	N/A	310.78	4
NB	85.431 on 83 Df	N/A	287.79	3
ZIP	N/A	-119.3 on 14 Df	266.54	1
ZINB	N/A	-119.1 on 15 Df	268.21	2

The test results in Table 2 show that the ZIP model has the best AIC value and is close to the ZINB model. The variables of the ZIP and ZINB models are similar, and their log-likelihood values are also close to each other. This indicates that both models have similar performance, showing that without the Negative Binomial part in the Zero Inflated model, this model still can handle over-dispersion well. Therefore, ZIP would be the optimal model with a priority on simplicity and fitting.

Vuong's tests on the two groups show that when doing raw comparing, the ZIP and ZINB models fit better than the Poisson and NB models, proving zero inflation in the data set. Specifically, with raw comparison, all tests have P-values<0.05. With AIC-corrected methods in Vuong's test to include the complexity and fitting trade-off, the ZINB model is statistically better (P-value < 0.05) than the NB model in AIC corrected while in other comparisons, the AIC-corrected P-values are all marginally significant (P-values < 0.06), suggesting t the ZINB and ZIP models are close to being statistically significantly better than Poisson and NB models when complexity is accounted.

Figure 2: Pearson Residuals against Fitted Values and Index



Diagnostics plots in Figure 2 also show that ZIP and ZINB are the best models, with Pearson residuals improved and now closer to 0 than other models, although there is still high randomness in observations numbers 0 to 40, suggesting that the models may not fully capture the variability in the data. This indicated a better fit than Poisson and NB. However, it still can not fit perfectly with the data.

The equations for the ZIP model have two parts: the Poisson count model and the logistic model for the excess zeros. After excluding all insignificant variables (P-value of Wald's test lower than 5%), the equations of the model after excluding insignificant variables for simplicity in relationship analysis (without refitting the model) are:

The estimated equation for the count ($\hat{y}$)

$$\log(\hat{y}) = 1.199 + 0.001442 \text{ x visits} - 0.007918 \text{ x revenue}$$

The estimated equation for the probability of being in the zero-inflated group (p):

$$\log\left(\frac{p}{1-p}\right) = 0.5588$$

IV. DISCUSSION

When analyse the relationship between various demographic factors of doctors in the target emergency department and the number of complaints they received in a year, with the knowledge from previous studies, the study has detect zero inflation and over-dispersion phenomena in the data. This leading to the deployment of Negative Binomial(NB), Zero-Inflated Poisson (ZIP), and Zero-Inflated Negative Binomial (ZINB) models besides the Poisson model to fit the data better.

The study's results revealed that the ZIP model is the optimal model. The ZIP model performance is outweighs other models, evaluated by its lower AIC value and a better log-likelihood, although this difference is not much between ZIP and ZINB models. In Vuong's test, the ZIP and ZINB models were also better than the Poisson and NB models, strengthening the evidence that ZIP is the optimal model.

This highlighted that it could handle zero inflation better than Poisson and NB models while working with over-dispersion effectively, even without the additional complexity of the Negative Binomial structure in the ZINB model. This matches the findings of one previous study [4] when their best models are both NB and ZINB models, which indicates that ZINB is not the only option to handle cases with zero inflation and over-dispersion.

From the ZIP model, no explanatory variables in the data can explain the estimated probability of being in the zero-inflated group. The intercept is also highly statistically insignificant, meaning the model can not provide a reliable base probability. This indicates that the probability of being in the zero-inflated group is unaffected by any given factors, and the probability is not fixed as the intercept is statistically insignificant.

Meanwhile, if the case is not in the zero-inflated group, the key predictors of complaint counts are the number of patient visits and the Doctor’s hourly income in dollars. The model indicates that the complaint case will increase by 0.14% for each additional patient’s visit. The expected number of complaints decreases by 0.79% for each additional dollar of hourly income. Meanwhile, residency training, working hours, and gender are insignificant in this mode. The intercept of this model is also highly statistically insignificant.

Those are interesting findings as they show that the more visits the doctors have, the more chance they can get complaints. Furthermore, if their income increases, the number of complaints also decreases. As income presents doctors' experience, attitude, role and skills, this also suggests that further study on those related factors can receive valuable in sights. These findings suggest that hospitals can improve patient satisfaction by increasing the number of doctors to reduce the patient load per doctor.

However, there is a significant drawback, as the ZIP model can not capture all factors that affect the probability of being in the zero-inflated group, and the number of complaints even has high performance in AIC values and Vuong’s tests. This can come from the data being only one year, with a small sample size. It also suggests that there can be missing factors that can significantly affect the number of complaints. Further study using a more comprehensive range of factors and data sets may give better results.

REFERENCES

[1] T. Mirzoev and S. Kane, “Key strategies to improve systems for managing patient complaints within health facilities – what can we learn from the existing literature?,” *Global Health Action*, vol. 11, no. 1, p. 1458938, Jan. 2018, doi: 10.1080/16549716.2018.1458938.

[2] S. Zengin *et al.*, “Analysis of complaints lodged by patients attending a university hospital: A 4-year analysis,” *Journal of Forensic and Legal Medicine*, vol. 22, pp. 121–124, Dec. 2013, doi: 10.1016/j.jflm.2013.12.008.

[3] D. M. Taylor, R. Wolfe, and P. A. Cameron, “Complaints from emergency department patients largely result from treatment and communication problems,” *Emergency Medicine*, vol. 14, no. 1, pp. 43–49, Mar. 2002, doi: 10.1046/j.1442-2026.2002.00284.x.

[4] C. G. Weaver, P. Ravani, M. J. Oliver, P. C. Austin, and R. R. Quinn, “Analyzing hospitalization data: potential limitations of Poisson regression,” *Nephrology Dialysis Transplantation*, vol. 30, no. 8, pp. 1244–1249, Mar. 2015, doi: 10.1093/ndt/gfv071.

[5] M.-C. Hu, M. Pavlicova, and E. V. Nunes, “Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial,” The American Journal of Drug and Alcohol Abuse, vol. 37, no. 5, pp. 367–375, Aug. 2011, doi: 10.3109/00952990.2011.597280.

[6] F. Sarvi, A. Moghimbeigi, and H. Mahjub, “GEE-based zero-inflated generalized Poisson model for clustered over or under-dispersed count data,” Journal of Statistical Computation and Simulation, vol. 89, no. 14, pp. 2711–2732, Jun. 2019, doi: 10.1080/00949655.2019.1632857.

[7] G. A. Fernandez and K. P. Vatcheva, “A comparison of statistical methods for modeling count data with an application to hospital length of stay,” BMC Medical Research Methodology, vol. 22, no. 1, Aug. 2022, doi: 10.1186/s12874-022-01685-8.

APENDIX: R CODE AND TECHNICAL ANALYSIS

1. Explanatory Data Analysis (EDA)

First, we use a summary to analyse the data:

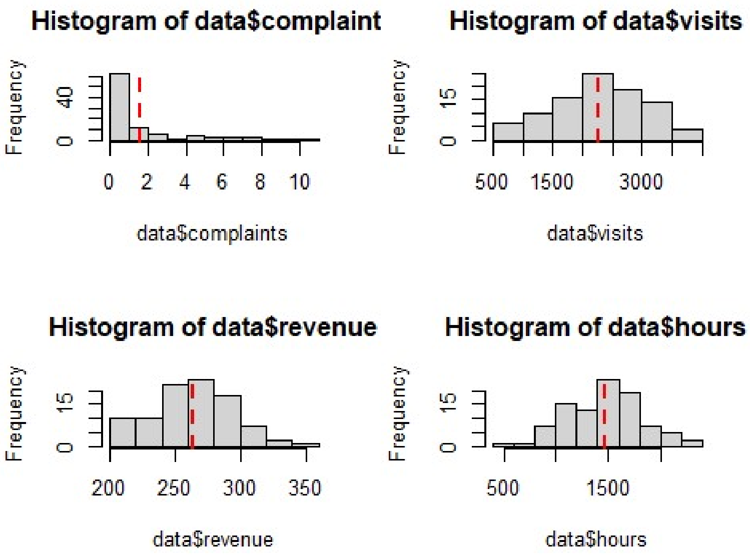
```
data <- read.csv("compdat.txt", sep = "\t")
data$residency <- as.factor(data$residency)
data$gender <- as.factor(data$gender)
data2 <- data
summary(data)
##      visits       complaints     residency gender    revenue
##  Min.   : 879   Min.   : 0.000   N:49      F:37   Min.   :203.9
##  1st Qu.:1698   1st Qu.: 0.000   Y:45      M:57   1st Qu.:243.8
##  Median :2299   Median : 0.000                    Median :263.7
##  Mean   :2271   Mean   : 1.564                    Mean   :263.8
##  3rd Qu.:2776   3rd Qu.: 2.000                    3rd Qu.:288.0
##  Max.   :3763   Max.   :11.000                    Max.   :342.9
##      hours
##  Min.   : 589
##  1st Qu.:1201
##  Median :1494
##  Mean   :1469
##  3rd Qu.:1700
##  Max.   :2269
```

We then calculated the standard deviation of each numerical variable.

```
round(sd(data$visits),2)
## [1] 723.92

round(sd(data$complaints),2)
## [1] 2.52

round(sd(data$revenue),2)
## [1] 30.9

round(sd(data$hours),2)
## [1] 351.36
```

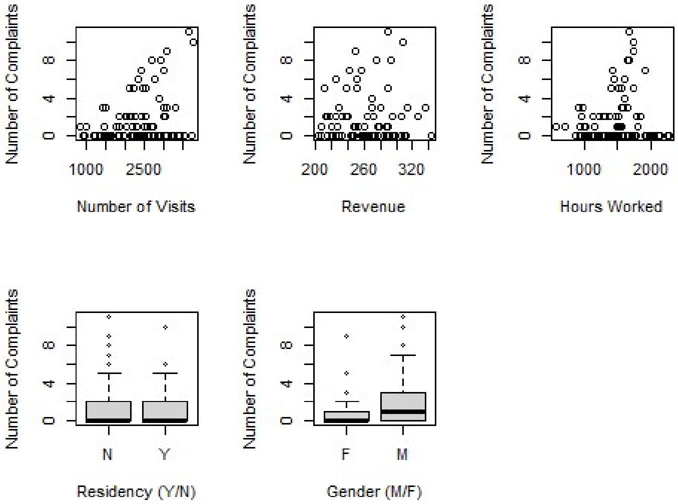
We then use histograms to analyse the distribution of the numerical variables.

```
par(mfrow = c(2, 2))
hist(data$complaints)
abline(v = mean(data$complaints), col = "red", lwd = 2, lty = 2,)
hist(data$visits)
abline(v = mean(data$visits), col = "red", lwd = 2, lty = 2)
hist(data$revenue)
abline(v = mean(data$revenue), col = "red", lwd = 2, lty = 2)
hist(data$hours)
abline(v = mean(data$hours), col = "red", lwd = 2, lty = 2)
```



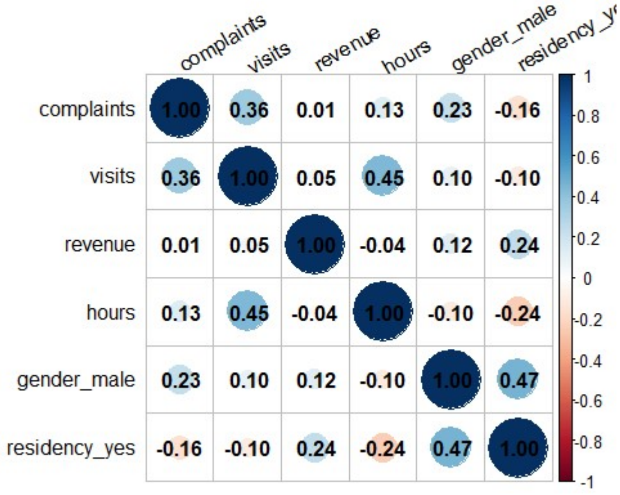
The complaint's plot shows a high volume of zero cases. We then used the scatter plots to analyse the relationship between complaints and other factors.

```
par(mfrow = c(2, 3))
plot(x = data$visits, y = data$complaints, xlab = "Number of Visits", ylab = "Number of Co
mplaints")
plot(x = data$revenue, y = data$complaints, xlab = "Revenue", ylab = "Number of Complaints
")
plot(x = data$hours, y = data$complaints, xlab = "Hours Worked", ylab = "Number of Complai
nts")
plot(x = data$residency, y = data$complaints, xlab = "Residency (Y/N)", ylab = "Number of
Complaints")
plot(x = data$gender, y = data$complaints, xlab = "Gender (M/F)", ylab = "Number of Compla
ints")
```



The sign of zero inflation is clear in this plot. We then use a correlation matrix plot to figure out if there is a high correlation between variables.

```
library(ggplot2)
library(corrplot)
library(dplyr)
data2$gender_male <- ifelse(data2$gender == "M", 1, 0)  # M/F becomes 1/2
data2$residency_yes <- ifelse(data2$residency == "Y", 1, 0)  # Y = 1, N = 0
continuous_vars <- data2[, c("complaints", "visits", "revenue", "hours", "gender_male","re
sidency_yes")]
# Calculate the correlation matrix
cor_matrix <- cor(continuous_vars)
# Visualize the correlation matrix
corrplot(cor_matrix, method = "circle",
         addCoef.col = "black",
         tl.col = "black", tl.srt = 30)
```



	complaints	visits	revenue	hours	gender_male	residency_y
complaints	1.00	0.36	0.01	0.13	0.23	-0.16
visits	0.36	1.00	0.05	0.45	0.10	-0.10
revenue	0.01	0.05	1.00	-0.04	0.12	0.24
hours	0.13	0.45	-0.04	1.00	-0.10	-0.24
gender_male	0.23	0.10	0.12	-0.10	1.00	0.47
residency_yes	-0.16	-0.10	0.24	-0.24	0.47	1.00

The plot shows that there is no high correlation in the matrix. However, the low correlation with complaints suggests there will be a challenge in fitting models using this data.

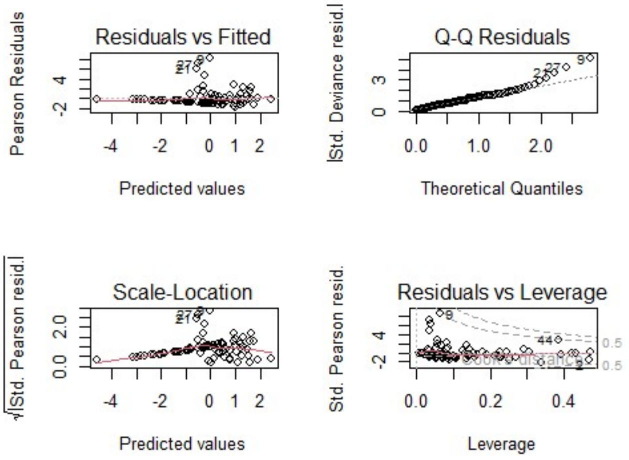
2. Model Fitting: Poisson Model

We begin to fit the Poisson model with all two-way interactions and then use the AIC step-wide backward method to reduce it.

```
library(MASS)
# Full model with all two-way interactions
model_Poisson_full <- glm(complaints ~
                            (visits + residency + gender + revenue + hours)^2,
                              data = data, family = poisson(link = "log"))
# Stepwise backward selection based on AIC
model_Poisson <- stepAIC(object = model_Poisson_full, direction = "backward", trace = F)
summary(model_Poisson)
##
## Call:
## glm(formula = complaints ~ visits + residency + gender + revenue +
##     hours + visits:revenue + residency:gender + residency:revenue +
##     residency:hours + gender:revenue, family = poisson(link = "log"),
##     data = data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.042e+01  3.275e+00   3.182 0.001465 **
## visits             -2.554e-03  1.227e-03  -2.081 0.037411 *
## residencyY          7.290e-01  1.670e+00   0.437 0.662380
## genderM            -2.939e+00  1.955e+00  -1.503 0.132728
## revenue            -5.744e-02  1.301e-02  -4.414 1.01e-05 ***
## hours               1.035e-03  5.454e-04   1.898 0.057673 .
## visits:revenue      1.296e-05  4.489e-06   2.888 0.003883 **
## residencyY:genderM -2.868e+00  4.776e-01  -6.005 1.91e-09 ***
## residencyY:revenue  1.738e-02  5.966e-03   2.913 0.003584 **
## residencyY:hours   -2.558e-03  6.845e-04  -3.736 0.000187 ***
## genderM:revenue     1.896e-02  7.957e-03   2.384 0.017147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 312.62  on 93  degrees of freedom
## Residual deviance: 166.00  on 83  degrees of freedom
## AIC: 310.78
##
## Number of Fisher Scoring iterations: 7
```

We then used diagnostic plots on the model.

```
par(mfrow=c(2, 2))
plot(model_Poisson)
```



In the Q-Q residual plot, it seems the model is not working well at the end of the tail. Meanwhile, other plots show an acceptable fit with the data. We then conduct the over-dispersion test using this model.

```
library(AER)
# Overdispersion test
dispersiontest(model_Poisson)
##  Overdispersion test
##
## data:  model_Poisson
## z = 2.2234, p-value = 0.01309
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.746316
# Mean and variance check for numerical variables
cat("Visits: mean =", mean(data$visits),"- variance =", var(data$visits), "\n")
## Visits: mean = 2270.596 - variance = 524058.7
cat("Revenue: mean =", mean(data$revenue),"- variance =", var(data$revenue), "\n")
## Revenue: mean = 263.7636 - variance = 954.8351
cat("Hours: mean =", mean(data$hours),"- variance =", var(data$hours), "\n")
## Hours: mean = 1468.91 - variance = 123453.8
```

The result indicates a high chance that there is over-dispersion in the data.

3. Model Fitting: Negative Binomial Model (NB)

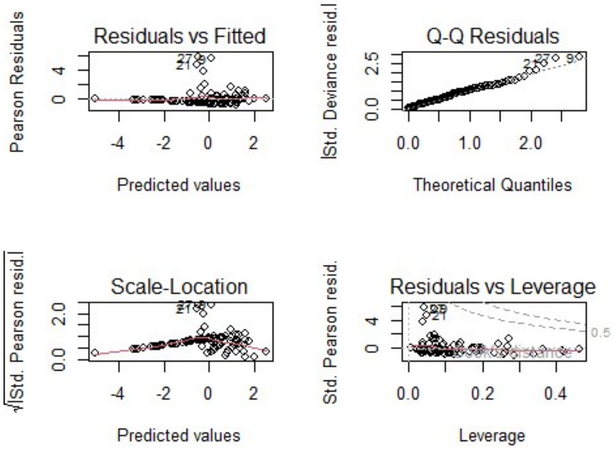
We begin to fit the Negative Binomial model with all two-way interactions and then use the AIC step-wide backward method to reduce it.

```
library(MASS)
# Negative Binomial full model with all two-way interactions
model_NB_full <- glm.nb(complaints ~
                          (visits + residency + gender + revenue + hours)^2,
                              data = data)
# Step wise backward selection based on AIC
```

```
model_NB <- stepAIC(object = model_NB_full,direction = "backward", trace = F)
summary(model_NB)
## Call:
## glm.nb(formula = complaints ~ visits + residency + gender + revenue +
##     hours + visits:revenue + residency:gender + residency:revenue +
##     residency:hours + gender:revenue, data = data, init.theta = 1.191894522,
##     link = log)
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.146e+01  5.103e+00   2.246  0.02470 *
## visits             -2.423e-03  1.832e-03  -1.322  0.18604
## residencyY          1.510e+00  2.815e+00   0.536  0.59181
## genderM            -4.251e+00  2.946e+00  -1.443  0.14902
## revenue            -6.260e-02  1.983e-02  -3.157  0.00160 **
## hours               1.068e-03  7.767e-04   1.376  0.16895
## visits:revenue      1.280e-05  6.816e-06   1.878  0.06039 .
## residencyY:genderM -2.814e+00  7.059e-01  -3.987 6.69e-05 ***
## residencyY:revenue  1.666e-02  1.010e-02   1.649  0.09911 .
## residencyY:hours   -2.972e-03  1.011e-03  -2.939  0.00329 **
## genderM:revenue     2.398e-02  1.184e-02   2.026  0.04279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.1919) family taken to be 1)
##     Null deviance: 149.163  on 93  degrees of freedom
## Residual deviance:  85.431  on 83  degrees of freedom
## AIC: 287.79
##
## Number of Fisher Scoring iterations: 1
##               Theta:  1.192
##           Std. Err.:  0.436
##  2 x log-likelihood:  -263.793
```

We then used diagnostic plots on the model.

```
par(mfrow=c(2, 2))
plot(model_NB)
```



Diagnostic plots show a better fit in this model than the Poisson model.

4. Model Fitting: Zero-Inflated Poisson Model (ZIP)

First, we try to fit the Zero-Inflated Poisson Model with all variables and suitable two-way interactions to avoid issues with model convergence or multicollinearity we have when using all interactions.

```
library(pscl)
model_ZIP <- zeroinfl(complaints ~
                          visits+ hours+ revenue + (residency + gender)^2|
                          visits+ hours+ revenue + (residency + gender)^2,
                           data = data, dist = 'poisson')
summary(model_ZIP)
## 
## Call:
## zeroinfl(formula = complaints ~ visits + hours + revenue + (residency + 
##     gender)^2 | visits + hours + revenue + (residency + gender)^2, data = data, 
##     dist = "poisson")
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -1.4474 -0.5755 -0.3986  0.3008  3.6420 
## 
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         1.199e+00  1.473e+00   0.814  0.41572    
## visits              1.442e-03  3.238e-05  44.536  < 2e-16 ***
## hours              -9.846e-04  5.355e-04  -1.839  0.06597 .  
## revenue            -7.918e-03  2.860e-03  -2.769  0.00563 ** 
## residencyY         -5.074e-01  4.370e-01  -1.161  0.24565    
## genderM            -7.107e-02  3.219e-01  -0.221  0.82524    
## residencyY:genderM  2.904e-01  4.534e-01   0.641  0.52181    
## 
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         5.588e-01  3.285e+00   0.170    0.865
## visits              2.561e-04  1.793e-03   0.143    0.886
## hours               2.810e-04  1.840e-03   0.153    0.879
## revenue            -1.255e-03  1.212e-02  -0.104    0.918
## residencyY         -1.370e+01  2.018e+02  -0.068    0.946
## genderM            -1.536e+01  2.881e+02  -0.053    0.957
## residencyY:genderM  2.798e+01  3.518e+02   0.080    0.937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 59 
## Log-likelihood: -119.3 on 14 Df

AIC(model_ZIP)
## [1] 266.5369
```

The AIC of the model is better than the previous models, and it is reasonable; further reducing or increasing the complexity did not give a better result. We then used diagnostic plots on the model.

5. Model Fitting: Zero-Inflated Negative Binomial Model (ZINB)

We begin to fit the Zero-Inflated Negative Binomial model with all variables and suitable two-way interactions to avoid issues with model convergence or multicollinearity we have when using all interactions.

```
model_ZINB <- zeroinfl(complaints ~
                          visits+ hours+ revenue + (residency + gender)^2|
                          visits+ hours+ revenue + (residency + gender)^2,
                           data = data, dist = 'negbin')
summary(model_ZINB)
## Call:
## zeroinfl(formula = complaints ~ visits + hours + revenue + (residency +
##     gender)^2 | visits + hours + revenue + (residency + gender)^2,
##                                            data = data,dist = "negbin")
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3099 -0.5856 -0.3937  0.2842  3.6156
##
## Count model coefficients (negbin with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.4428156  2.9460534   0.490   0.6243
## visits              0.0015351  0.0003813   4.026 5.67e-05 ***
## hours              -0.0011771  0.0018345  -0.642   0.5211
## revenue            -0.0087563  0.0047726  -1.835   0.0665 .
## residencyY         -0.4587666  0.4808625  -0.954   0.3401
## genderM            -0.0527055  0.3312755  -0.159   0.8736
## residencyY:genderM  0.2341677  0.4897740   0.478   0.6326
## Log(theta)          2.9028077  0.5483932   5.293 1.20e-07 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         7.112e-01  7.464e+00   0.095    0.924
## visits              4.205e-04  1.363e-02   0.031    0.975
## hours               4.494e-05  1.042e-02   0.004    0.997
## revenue            -2.171e-03  3.342e-02  -0.065    0.948
## residencyY         -1.370e+01  2.069e+02  -0.066    0.947
## genderM            -1.536e+01  2.915e+02  -0.053    0.958
## residencyY:genderM  2.794e+01  3.595e+02   0.078    0.938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 18.2252
## Number of iterations in BFGS optimization: 63
## Log-likelihood: -119.1 on 15 Df

AIC(model_ZINB)
## [1] 268.2066
```

The AIC of the model is better than the previous models, and it is reasonable but not as good as the ZIP model; further reducing or increasing the complexity did not give better results. We then used diagnostic plots on the model.

6. Model Comparison

We first conduct an AIC comparison between models.

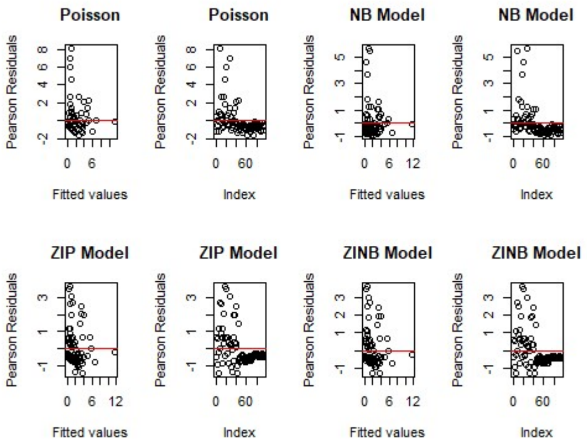
```
AIC(model_Poisson)
## [1] 310.7804

AIC(model_NB)
## [1] 287.7934

AIC(model_ZIP)
## [1] 266.5369

AIC(model_ZINB)
## [1] 268.2066
```

ZIP model give the Lowest AIC value, this indicate it is optimal model in this comparison. We then conduct Vuong’s tests to test if ZIP and ZINB models are better than Poisson and NB models.

```
vuong(model_Poisson, model_ZIP)
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A  p-value
## Raw                    -1.835148 model2 > model1 0.033242
## AIC-corrected          -1.615998 model2 > model1 0.053047
## BIC-corrected          -1.337316 model2 > model1 0.090560

vuong(model_Poisson, model_ZINB)
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A  p-value
## Raw                    -1.864772 model2 > model1 0.031107
## AIC-corrected          -1.643538 model2 > model1 0.050136
## BIC-corrected          -1.362207 model2 > model1 0.086566

vuong(model_NB, model_ZIP)
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A  p-value
## Raw                   -2.0435965 model2 > model1 0.020497
## AIC-corrected         -1.5581147 model2 > model1 0.059603
## BIC-corrected         -0.9407532 model2 > model1 0.173416

vuong(model_NB, model_ZINB)
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A  p-value
## Raw                        -2.196740 model2 > model1 0.014020
## AIC-corrected              -1.681612 model2 > model1 0.046322
## BIC-corrected              -1.026551 model2 > model1 0.152316
```

We then conduct diagnostic plots on each model to ensure the ZIP model fits well with the data.

```r
par(mfrow=c(2, 4))
residuals_Poisson <- residuals(model_Poisson, type = "pearson")
plot(fitted(model_Poisson), residuals_Poisson,
     xlab = "Fitted values", ylab = "Pearson Residuals", main = "Poisson")
abline(h = 0, col = "red")
plot(seq_along(residuals_Poisson), residuals_Poisson,
     xlab = "Index", ylab = "Pearson Residuals", main = "Poisson")
abline(h = 0, col = "red")
residuals_NB <- residuals(model_NB, type = "pearson")
plot(fitted(model_ZIP), residuals_NB,
     xlab = "Fitted values", ylab = "Pearson Residuals", main = "NB Model")
abline(h = 0, col = "red")
plot(seq_along(residuals_NB), residuals_NB,
     xlab = "Index", ylab = "Pearson Residuals", main = "NB Model")
abline(h = 0, col = "red")
residuals_zip <- residuals(model_ZIP, type = "pearson")
plot(fitted(model_ZIP), residuals_zip,
     xlab = "Fitted values", ylab = "Pearson Residuals", main = "ZIP Model")
abline(h = 0, col = "red")
plot(seq_along(residuals_zip), residuals_zip,
     xlab = "Index", ylab = "Pearson Residuals", main = "ZIP Model")
abline(h = 0, col = "red")
residuals_ZINB <- residuals(model_ZINB, type = "pearson")
plot(fitted(model_ZIP), residuals_ZINB,
     xlab = "Fitted values", ylab = "Pearson Residuals", main = "ZINB Model")
abline(h = 0, col = "red")
plot(seq_along(residuals_ZINB), residuals_ZINB,
     xlab = "Index", ylab = "Pearson Residuals", main = "ZINB Model")
abline(h = 0, col = "red")
```



Combining all the comparisons, the best model is the ZIP model, when AIC, Vuong's test and diagnostics plots all show good performance that overshadows the other models.