STAT2402 Exam: Analysis on Doctor's Complaints

Aaminah Irfan: 23642166

Abstract:

The primary aim of this report is to develop a simplified model to understand the relationship between number of complaints against doctors at the emergency department and their demographic factors. The data set used includes information on 94 doctors including: number of complaints, number of patient visits, whether the doctor is in residency training, gender, hourly income, and total hours worked. To achieve the objective, five distinct Poisson models were employed including: Poisson Model, Quasi-Poisson, Negative Binomial Poisson (NBP), Zero-Inflated Poisson Regression (ZIPR), and Zero-Inflated Negative Binomial (ZINB) models, each addressing different aspects of the data's distribution and characteristics. Best model for each 5 models were developed and compared, using criterion such as dispersion test, `AIC, rootograms` and Pearson's Residual Plot. The Zero-Inflated Negative Binomial (ZINB) Model was found to be the most suitable, addressing over-dispersion and excess zero counts in the data. Results showed that an increase in patient visits significantly increased complaints, while more working hours and higher revenue reduced the number of complaints. Gender and revenue were not statistically significant factors, indicating limited influence on complaints. However, they exhibited a high probability of receiving zero complaints. This research offers valuable insights into the complex interplay between demographic factors and patient complaints, informing healthcare providers and policymakers for better service quality.

Introduction:

Medico-legal matters in healthcare are far from homogeneous and can vary significantly from one country to another. Past research has highlighted a crucial finding: malpractice lawsuits, complaints, disciplinary proceedings, and other legal matters tend to cluster among specific and small groups of doctors with identifiable characteristics [1]. This clustering raises significant questions about the factors contributing to these complaints and the demographic characteristics of the doctors involved. However, limited research has been conducted to explore complaints lodged against doctors. Such complaints are vital in shedding light on the quality of patient care and the levels of patient dissatisfaction within the healthcare system [1]. Within the intricate landscape of healthcare, identifying doctors at risk of receiving complaints before they accumulate troubling records is an ongoing challenge. Clinical leaders, risk managers, and regulators often lack reliable methods for systematically determining which doctors should be targeted for assistance and preventive action. Consequently, the medico-legal enterprise remains reactive, primarily dealing with the aftermath of adverse events and behaviors that lead to costly disputes [2].

To address this gap, recent research has delved into the factors contributing to complaints against doctors and doctors' demographics. Notably, Bratland et al. (2020) aimed to identify predictors of complaint-prone doctors in Australia and develop a robust method for forecasting medico-legal risk [2]. Their findings revealed that male doctors had a 40% higher risk of recurrence of complaints than their female counterparts. Additionally, physicians without a specialty in general practice were at a significantly higher risk of receiving a complaint. On the other hand, doctors with medium-low and medium-high workloads were at a lower risk of complaints than those without duty during the preceding fourteen days [2]. In essence, continuous medical training and achieving specialization in general practice were decisively associated with a reduced risk of complaints in primary care emergency services [2]. Furthermore, research conducted by Walton et al. (2020) aimed to profile the most common complaints against doctors and explore the demographic factors associated with receiving complaints [3]. Their work found that males were more likely to receive complaints than females, and doctors under 35 were less likely to receive complaints. Interestingly, complaints regarding communication, documentation, and medico-legal conduct became more common as doctors' age increased. On the other hand, complaints related to health impairment and offense decreased with age [3].

International research has also provided valuable insights into complaints against doctors. For instance, a study conducted in Singapore by Wong et al. (2007) determined specific rates and rate ratios for various demographic groups among self-complainants. This study revealed a complaint case rate of 1.17 per 1,000 visits in the emergency department [4]. Meanwhile, a study in Romania by Hanganu et al. (2022) applied a logistic regression model to identify independent predictors for receiving complaints. Their findings demonstrated that men, senior doctors from surgical specialties, those performing a greater number of on-call shifts, doctors working in regional or county hospitals, those with a heightened fear of receiving complaints, and those whose life partners were doctors in the same specialty were more prone to receiving complaints. Furthermore, research in Canada conducted by Vogel et.al (2019) revealed that during the period from 2008 to 2017, there were 854 regulatory complaints that involved medical trainees, demonstrating an annual increase of 94% in complaints. This reveals that amount of training and whether the doctor is in residency can significantly affect the number of complaints received.

The primary aim of this analysis is to determine the effect of demographic factors on the number of complaints received by doctors. By doing so, this paper aspires to create a model that can predict the number of complaints received based on demographic factors. This report is organised is as follows. In the following section statistical methodology is described in depth, followed by Results section which presents the results of the analysis. This is followed by Discussion section where it discuses the results in detail and compares the results with research papers discussed in the Introduction. Finally, paper is concluded by summarizing results and provide implications that can potentially guide healthcare institution in proactively addressing patient complaints and improving patient care.

Methodology:

Statistical Methods were employed to predict the number of complaints based on demographic factors. Response variable (y) considered was Number of complaints, and Explanatory variables (x) were Visits, Residency, Gender, Revenue and Hours. The data was initially explored through graphical and numerical summaries. To facilitate interpretation, categorical variables such as Gender and Residency were re-leveled to 'Male' and 'No,' respectively. To achieve the research objectives, various types of Poisson models were fitted, given that the data set is comprised of count data. These models included the Poisson Model, Quasi-Poisson Model, Negative Binomial Poisson Model, Zero-Inflated Poisson Regression Model, and Zero-Inflated Negative Binomial Model. Interaction variables were considered where appropriate and applicable. For all statistical tests, a significance level of 5% was used. To assess the significance of the variables or the models, the following null and alternate hypothesis were employed:

- Null Hypothesis: There is no relationship between explanatory variable(s) and the Number of Complaints.
- Alternate Hypothesis: There is a relationship between explanatory variable(s) and the Number of Complaints.

Poisson Model: To fit a Poisson Model, step-wise approach was used, initially a full model (Model 1) with all explanatory variables was fitted. Model 1 was expanded to include two-way interactions between the explanatory variables. To identify the most relevant variables and eliminate irrelevant ones, Stepwise Akaike Information Criterion `(stepAIC)` method was employed, leading to the selection of Model 2. Model 2 was furthered refined by removing insignificant variables using method of Backward selection, resulting in Model 3. All three models were compared based on AIC values and dispersion test and final and best Poisson Model was chosen.

Quassi-Poisson Model: For the Quasi-Poisson model, an initial full model was fitted, including all explanatory variables.

Negative Binomial Poisson Mode: To fit Negative Binomial Poisson Model, initially a full model with all explanatory variables was fitted. "`drop-term`" method was utilized to optimize the model by removing variables based on AIC and p-values. The best model was chosen based on AIC and rootogram analysis.

Zero-Inflated Models: For the Zero-Inflated Poisson Regression and Zero-Inflated Negative Binomial models, full models with all explanatory variables were initially fitted. Multiple models were iteratively constructed by excluding highly insignificant variables, employing criteria such as AIC, dispersion tests, and variable significance. The best models were chosen considering AIC, dispersion tests, and rootogram analysis.

All final models from each type were compared using dispersion tests, `rootograms`, and AIC. Based on these criteria, the final and best model was selected.

Results:

Exploratory Data Analysis (EDA): Data set used contained information 94 doctors and their demographics. Table 1 presents summary statistics for the data set, revealing over-dispersion due to variances exceeding means. Figure 1 further confirms over-dispersion as the frequency of zero complaints is notably high. It also illustrates that males tend to receive more complaints than females, and doctors not in residency training receive more complaints than those in residency. Figure 2 displays scatter plots for numerical variables with trend lines for each gender, showing a concentration of complaints in the 0-2 range, indicating an excess of lower complaint numbers.

Table 1: Data Summary Table

Variable	Number of Observations	Value Range	Mean	Variance
Visits: the number of patient visits	94	879-3763	2271	524058.7
Complaints: the number of complaints against the doctor in the previous year	94	0-11	1.56	6.36
Residency: is the doctor in residency training (Y = Yes, N= No)				
	Yes (Y) = 45 No (N) = 49	Range of complaints by Residency N: 0-11 Y: 0-10	Mean number of complaints: Y = 1.16 N= 1.94	Variance of number of complaints: Y = 3.82 N= 8.52
Gender: gender of the doctor (M = Male, F= Female)				
	M = 57 F = 37	Range of complaints by gender F: 0-9 M: 0-11	Mean number of complaints: F = 0.83 M = 2.03	Variance of number of complaints: F = 3.19 M = 7.92
Revenue: doctor's hourly income (dollars $)	94	203.9-342.9	263.9	954.84
Hours: total number of hours the doctor worked in a year.	94	589-2269	1469	123453.80

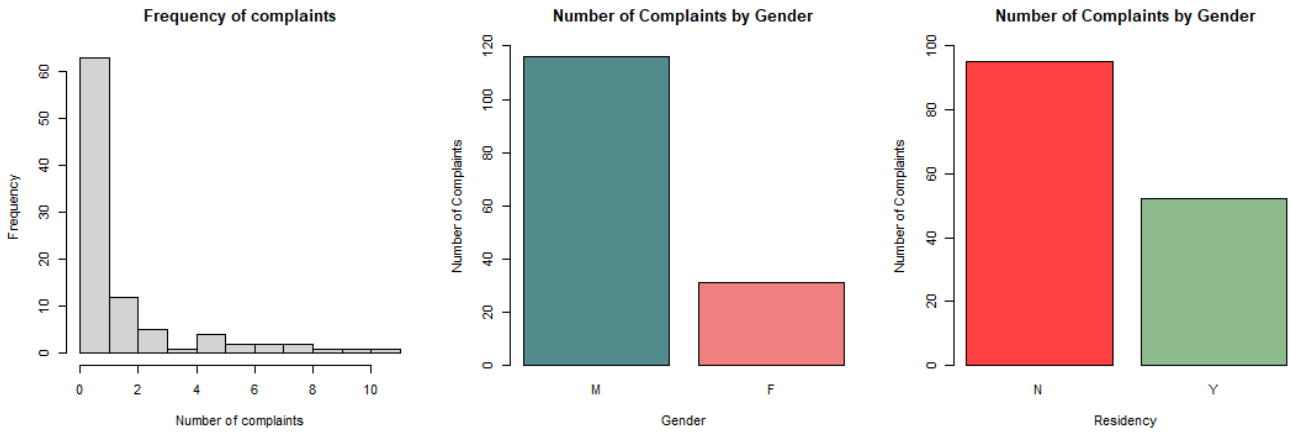

Figure 1: Explanatory Data Analysis Plots

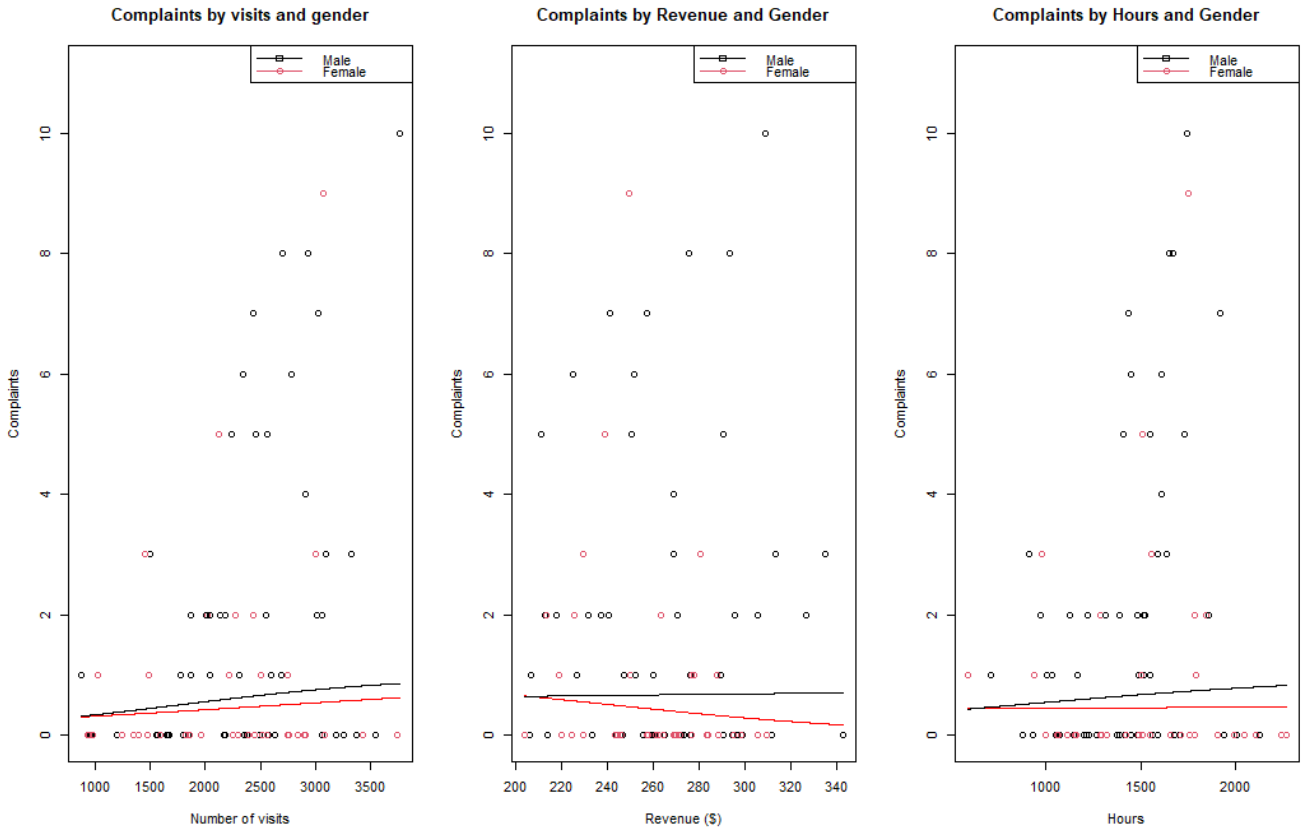

Figure 2: Scatter plots for numerical variables

Model Assessment:

1. **Poisson Model:** Initially, a full Poisson model (Model 1) with all explanatory variables was fitted. Subsequently, two-way interactions between the variables were introduced and `stepAIC (from R)` was employed to reduce the model to only relevant and significant terms, resulting in Model 2. Model 2 was further reduced by removing insignificant variables using Backward Selection Method, resulting in Model 3. Table 2 presents AIC and degrees of freedom `(df)` for the models obtained from summary results. Model 2, with the lowest AIC and better dispersion (166/83 =2), was chosen as the best Poisson Model. Figure 3 displays rootogram and Pearson residual for best Poisson Model.

Table 2: Poisson Model Comparison

Model	Residual Deviance	Model df	AIC
1	221.66	88	356.44
2	166	83	310.78
3	218	89	351.63

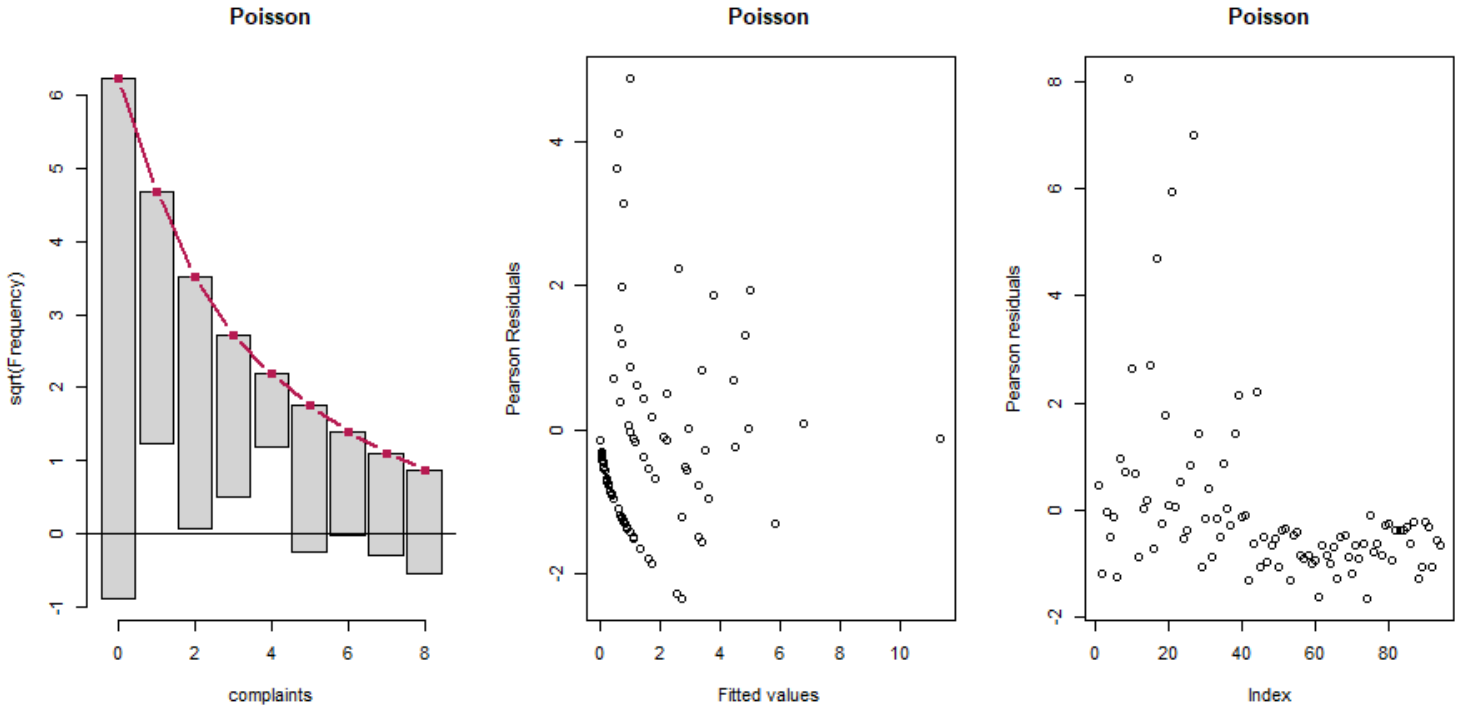

Figure 3: Poisson Model Diagnostics

Equation for best Poisson Model is as follows:

$$\ln(\lambda) = (7.481) - (0.00255) \cdot \text{visits} - (2.139) \cdot \text{residency} + (2.939) \cdot \text{genderF} \quad (1)$$
$$- (3.847) \cdot \text{revenue} + (0.001) \cdot \text{hours} + (0.000013) \cdot \text{visits} \cdot \text{revenue} \quad (2)$$
$$+ (2.868) \cdot \text{residency} \cdot \text{genderF} + (0.017) \cdot \text{residency} \cdot \text{revenue} \quad (3)$$
$$- (0.00256) \cdot \text{residency} \cdot \text{hours} - (0.0189) \cdot \text{genderF} \cdot \text{revenue} \quad (4)$$

However, Model 2 fails the dispersion test where p-value $(0.013) < 0.05$, leading to rejection of null hypothesis that dispersion is less than 1. This indicated that Model 2 is over-dispersed.

2. **Quasi-Poisson Model:** Quassi-Poisson Model was fitted to account for over-dispersion and unequal variance and mean. Residual Deviance for this model was 221.66 on 88 degrees of freedom. This means that model is still over-fitted i.e., (Residual Deviance (221.6) / 88) = 2.52. Figure 4 displays Pearson residual for best QP Model.

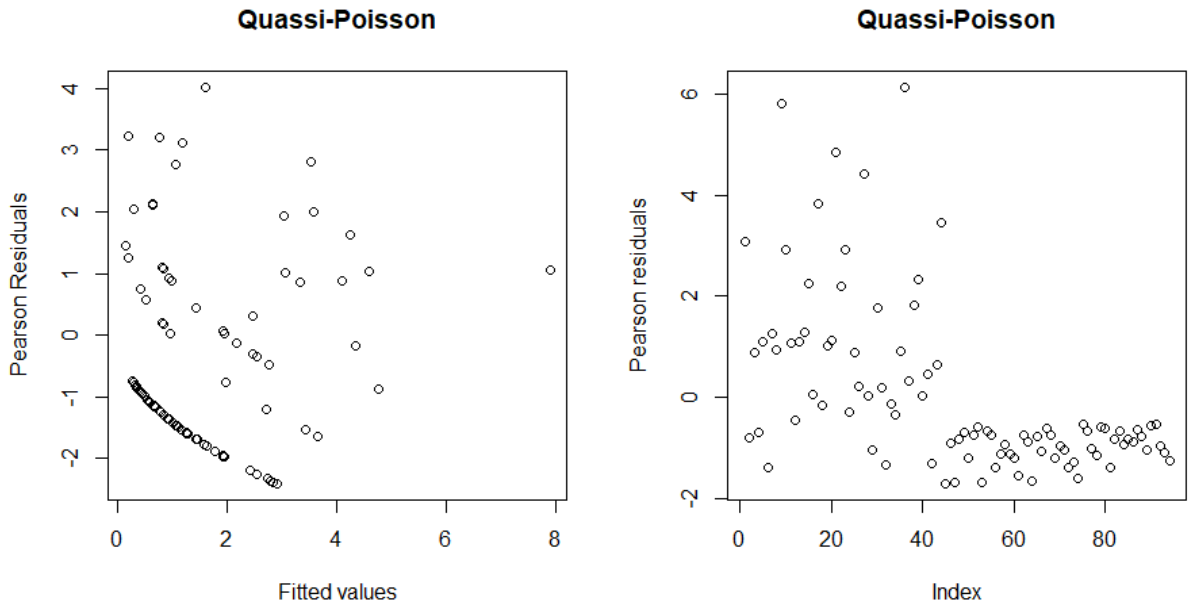

Figure 4: Quassi-Poisson Model Diagnostics

Equation for Quassi-Poisson Model is as follows:

$$\ln(\lambda) = -0.3665 + (0.0009 \cdot \text{visits}) - (0.8648 \cdot \text{residencyY}) - (1.1203 \cdot \text{genderF}) - (0.0010 \cdot \text{revenue}) - (0.0002 \cdot \text{hours})$$

3. **Negative Binomial Poisson (NBP) Model:** Initially, Full model including all explanatory variables was fitted (Model 1). Using `drop-term` method with `Chi-squared` test p-values, variables were dropped on the basis of highest p-value until all variables were significant. Variables were dropped in the order of residency `(p-value = 0.295)` (Model 2) , gender `(p-value = 0.068)` (Model 3), revenue `(p-value = 0.0519)` (Model 4), hours `(p-value = 0.0841)` (Model 5). `AIC` was used to compare all 5 models. Table 3 displays results for all NBP Models.

Table 3: NBP Models Comparison

Model	Residual Deviance	df	AIC
1	83.535	88	301.04
2	82.376	89	300.14
3	81.021	90	301.13
4	83.394	91	301.79
5	95.478	92	301.59

Since Model 2 has the lowest AIC, Model 2 was chosen as the best NBP Model. Figure 5 displays rootogram and Pearson residual for best ZIPR Model.

Equation for best NBP Model is as follows:

$$\ln(\lambda) = 1.8920 + (0.0014 \cdot \text{visits}) - (0.6744 \cdot \text{genderF}) - (0.0111 \cdot \text{revenue}) - (0.0012 \cdot \text{hours})$$

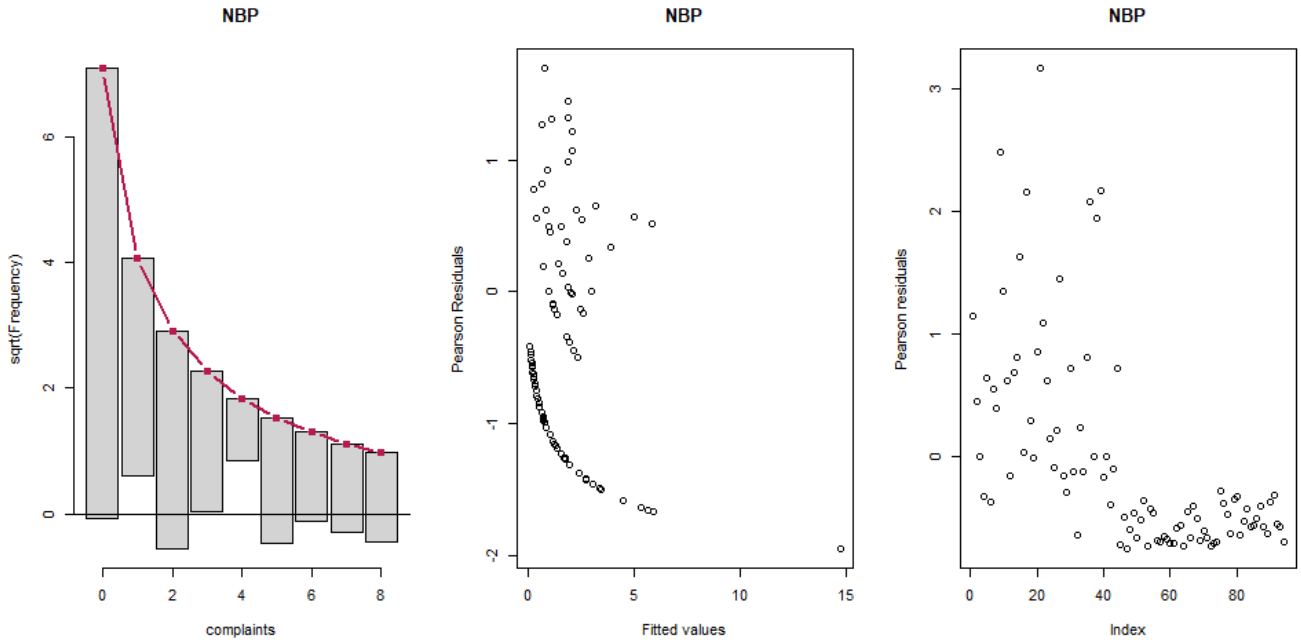

Figure 5: NBP Model Diagnostics

4. **Zero Inflated Poisson Regression (ZIPR) Model:** Initially, a full model incorporating all explanatory variables was applied. Due to the generation of `Not-a-Number (NAN)` values in the summary results, the "visits" variable was excluded from the Bernoulli component of the ZIPR model. Subsequently, a series of models was systematically generated by removing highly insignificant variables. These models were continuously compared using evaluation criteria such as `rootograms`, `AIC`, dispersion tests, and variable significance. The final model achieved the lowest `AIC` of `282.93` with `8` degrees of freedom. Figure 6 presents the `rootogram` and Pearson residual for the optimal ZIPR Model.

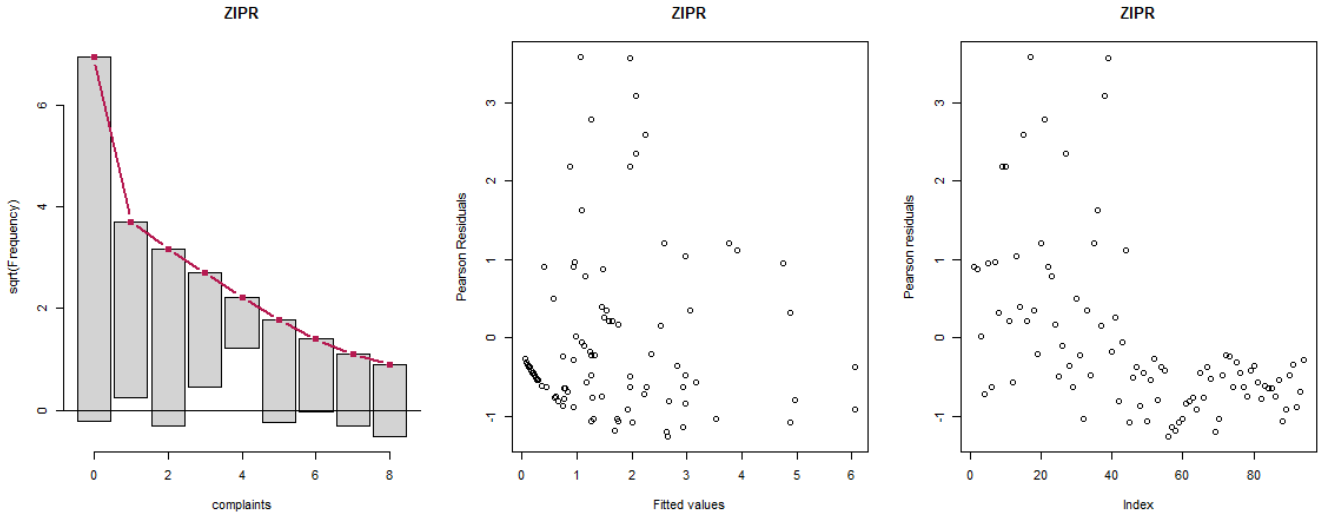

Figure 6: ZIPR Model Diagnostics

Equation for best ZIPR Model is as follows:

$$\text{Poisson} : \ln(\lambda) = 2.509 \times \text{Intercept} + 0.0020 \times \text{visits} - 0.0145 \times \text{revenue} - 0.0018 \times \text{hours}$$

$$\text{Bernoulli} : logit(\pi) = -18.8570 + 1.6476 \times \text{genderF} + 6.9085 \times \log(\text{visits}) - 5.0227 \times \log(\text{hours})$$

5. **Zero Inflated Negative Binomial (ZINB) Model:** Initially, the full model encompassing all explanatory variables was implemented. However, it was observed that the "visits" and "hours" variables generated `Not-a-Number (NAN)` values in the summary results. To address this, these variables were log-transformed in both the Bernoulli and Poisson components of the ZINB model. Subsequently, a series of models were iteratively constructed by excluding highly insignificant variables, and these models were assessed using various evaluation criteria, including `rootograms, AIC,` dispersion tests, and variable significance. This process resulted in the identification of two final best models. `Model 1` achieved the lowest `AIC of 273.3` in comparison to all other models, suggesting it as the best-performing model among the candidates. However, `Model 2`, with an `AIC` of `277.82,` exhibited better `rootogram` and Pearson Residual plot compared to `Model 1` and all other models. Furthermore, `Model 2` resulted in a higher number of significant variables. Consequently, `Model 2` was selected as the optimal ZINB model. To gain a better understanding of the model's fit, Figure 7 illustrates the `rootogram` and Pearson residual for this chosen ZINB Model.

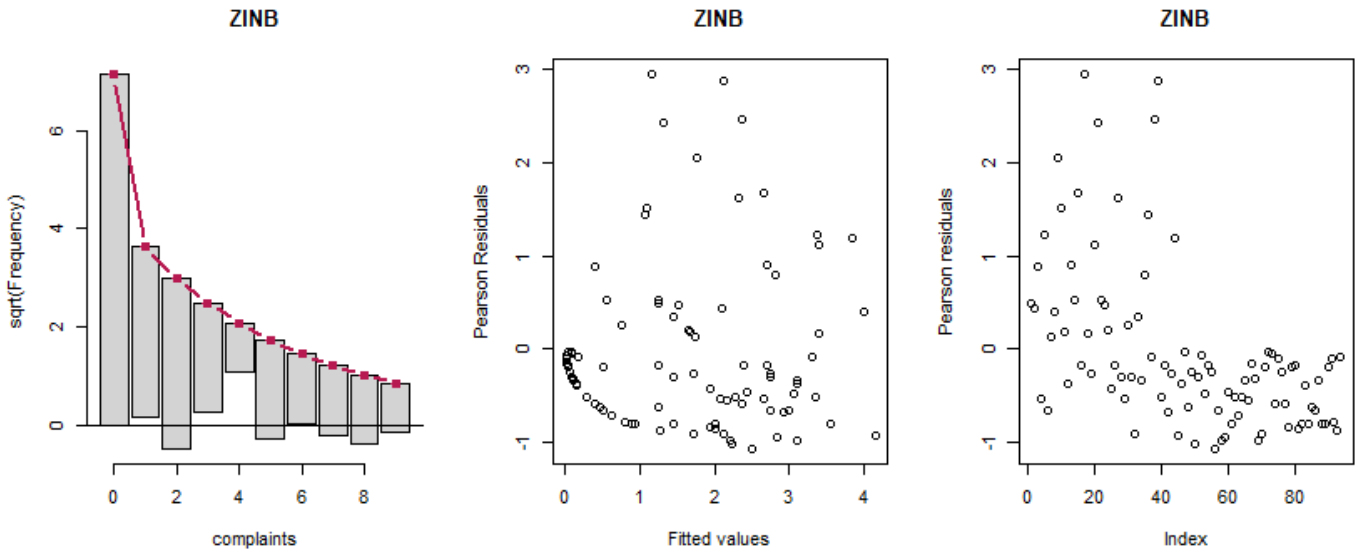

Figure 7: ZINB Model Diagnostics

Equation for best ZINB model (Model 2) is as follows:

$$\text{Poisson} : ln(\lambda) = -8.3530 + 5.8737 \cdot \log(visits) - 4.4292 \cdot \log(hours) - 0.0161 \cdot \text{revenue}$$

$$\text{Bernoulli} : log(\pi) = -11.4292 + 13.9128 \cdot \log(\text{visits}) - 12.3762 \cdot \log(\text{hours}) + 1.6761 \cdot \text{genderF} - 0.0352 \cdot \text{revenue}$$

Comparison of all 5 Models: All five models were evaluated using a combination of `AIC` and `rootograms` for model comparison. Table 4 presents the `AIC` values for the best models from each type.

Table 4: Final Models Comparison

Model	AIC
Poisson	310.78
Quassi-Poisson	NA
Negative Binomial Poisson	300.14
Zero Inflated Poisson Regression	282.93
Zero Inflated Negative Binomial	277.82

The ZINB model emerged as the best model due to its lowest AIC value, superior rootogram, and Pearson's Residuals.

Discussion:

In this analysis, we aimed to explore the impact of demographic factors on the number of complaints against doctors, employing five different Poisson Models for modelling the data. The first model, the Poisson Model, was chosen due to the count nature of the data. While it included significant variables and interaction terms, it assumed equal variance and mean, which did not hold true for the data set, indicating over-dispersion. Moreover, the model did not account for excess zero counts, indicating that it is a least good-fit model. The Quasi-Poisson Model was chosen to accommodate over-dispersion. However, it still exhibited over-dispersion as indicated by dispersion tests and Pearson Residuals Plot. It also did not address the issue of excess zeros. To address both over-dispersion and excess zero counts, the Negative Binomial Poisson (NBP) Model was introduced. The best NBP model demonstrated a lower AIC value than the previous two models, indicating a better fit. While the Pearson Residuals plot showed improvements, the residuals still clustered to the left, suggesting it might not be the ideal choice. The Zero-Inflated Poisson Regression (ZIPR) Model was considered to tackle excess zeros and over-dispersion, and it showed promise with a lower AIC and improved Pearson Residuals plots. However, it had certain limitations in handling high over-dispersion, which led to the consideration of the Zero-Inflated Negative Binomial (ZINB) Model. The ZINB Model, with the lowest AIC among all the models, emerged as the most suitable choice. It effectively addressed over-dispersion and excess zero counts as evident from improved residual plots and rootograms. However, it should be noted that the model's performance is not perfect for all counts, but it represents the best option among all other models.

Interpretation of the model chosen is as follows:

- For one unit increase in log(visits), mean number of complaints increases by a factor of $exp(5.874) = 325.06$, holding all other variables fixed. As number of visits increases, mean number of complaints increases.
- For one unit increase in log(hours), mean number of complaints decreases by a factor $exp(4.429) = 83.85$, holding all other variables fixed. As number of hours increases, mean number complaints decreases.
- For one unit increase in revenue, mean number of complaints decrease by a factor of $exp(0.016) = 1.02$, holding all other variables fixed. As revenue increases, mean number of complaints decreases.
- The probability of zero complaints, increases with more visits and if the doctor is female, and decreases as number of hours and revenue increases.

Note that gender variable is not significant at 5% level (p-value =0.09), and revenue is also not significant (p-value = 0.123).All other variables are statistically significant. This suggests that with more visits to doctors, number of complaints increase, however there is also a high probability of zero complaints. Furthermore, as number of hours worked increases which also indicate more revenue, means that average number of complaints decrease with a lower probability of receiving zero complains. Model also suggest that female doctors tend to have a higher probability of zero complaints than males, however this is insignificant. The results align with research conducted by Wong et al. (2007) where number of complaints increase with number of visits. According to research by Bratland et al (2020) and Walton et al.(2020) , male doctors tend to receive more complaints. However, gender is not incorporated in our final model chosen i.e., it is not included in the Poisson Model but in Bernoulli

Model which models variables with excess zeros. ZINB model consists of two components: the count model and the zero-inflation model. Whether a predictor variable appears in the count model or the zero-inflation model is based on the nature of the data. Since, there is zero-inflation in data, some predictors were likely to have high probability of having zero complaints, while others do not. For example, `genderF` i.e., Female appeared in the Bernoulli model of the ZINB because female doctors tend to have a higher probability of receiving zero complaints than Males. This is also depicted in the scatter plots and bar chart where female doctors tend to receive less complaints than males.

While this research aimed to devise a more simple model to predict complaints against doctors, there are some limitations. Firstly, the selected final model excludes certain demographic factors from the data set, primarily because they were found to be statistically insignificant. However, the insignificance of these factors might result from limitations in the modeling approach. Secondly, the relatively small sample size of 94 observations may not comprehensively represent all doctors. Future research should focus on incorporating additional information, such as the reasons for complaints, as well as considering the demographics of the patients filing the complaints to develop a more robust model

References:

[1] Marie M Bismark, Matthew J Spittal and David M Studdert Med J Aust 2011; 195 (1): 25-28. || doi: 10.5694/j.1326-5377.2011.tb03183.x Published online: 4 July 2011

[2] Bratland, S.Z., Baste, V., Steen, K. et al. Physician factors associated with increased risk for complaints in primary care emergency services: a case – control study. BMC Fam Pract 21, 201 (2020). https://doi.org/10.1186/s12875-020-01272-0

[3] Walton, M., Kelly, P. J., Chiarella, E. M., Carney, T., Bennett, B., Nagy, M., & Pierce, S. (2020). Profile of the most common complaints for five health professions in Australia. Australian Health Review, 44(1), 15. https://doi.org/10.1071/ah18074

[4] Wong, L. L., Ooi, S. B., & Goh, L. G. (2007). Patients' complaints in a hospital emergency department in Singapore. *Singapore medical journal*, *48*(11), 990–995.

[5] Hanganu, B., Iorga, M., Pop, L. M., & Ioan, B. G. (2022). Socio-Demographic, Professional and Institutional Characteristics That Make Romanian Doctors More Prone to Malpractice Complaints. *Medicina*, *58*(2), 287. https://doi.org/10.3390/medicina58020287

[6] Vogel L. (2019). Growing number of medical trainees named in complaints. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 191(25), E717–E718. https://doi.org/10.1503/cmaj.109-5762

Appendix:

Relevant libraries were imported and data was read. Structure and summary results were obtained and data was visualized using various plots.

```
library(pscl)
library(AER)
library(MASS)
library(vcd)
library(latexpdf)
library(tinytex)
library(rmarkdown)
#install.packages("countreg", repos="http://R-Forge.R-project.org")
library(countreg)


comp <- read.table('compdat.txt', header = T, stringsAsFactors = T)
```

```r
str(comp)
summary(comp)
head(comp)
comp$gender <- relevel(comp$gender, ref='M')
comp$residency <- relevel(comp$residency, ref='N')
```

Code for Figure 1:

```r
par(mfrow = c(1,3))
hist(comp$complaints, xlab = "Number of complaints", main = "Frequency of complaints")
barplot(height = tapply(comp$complaints, comp$gender, sum),
        col = c("darkslategray4", "lightcoral"),
        main = "Number of Complaints by Gender",
        xlab = "Gender",
        ylim = c(0,120),
        ylab = "Number of Complaints")

barplot(height = tapply(comp$complaints, comp$residency, sum),
        col = c("brown1", "darkseagreen"),
        main = "Number of Complaints by Gender",
        xlab = "Residency",
        ylim = c(0,100),
        ylab = "Number of Complaints")
```

Code for Figure 2:

```r
## INDIVIDUAL PLOTS BASED ON MODELS BY GENDER:
# VISITS
summary(visits.lm <- glm(complaints ~ visits + gender + visits:gender,
                         data = comp, family = poisson)) # significant
with(comp, plot(jitter(visits), complaints, col = gender,
                ylab = "Complaints", xlab = "Number of visits",
                main = "Complaints by visits and gender"))
grid <- seq(min(comp$visits), max(comp$visits), by = 0.1)

coefs <- coef(visits.lm)
female.int <- coefs[1] + coefs[3]
female.slope <- coefs[2] + coefs[4]
fem.pi <- exp(female.int + female.slope * grid)/(1 + exp(female.int + female.slope *
                                                          grid))
male.pi <- exp(coefs[1] + coefs[2] * grid)/(1 + exp(coefs[1] + coefs[2] * grid))
lines(grid, fem.pi, type = "l", col = "red")
lines(grid, male.pi, type = "l", col = "black")
legend("topright", col = c(1, 2), pch = c(0, 1), lty = 1, legend = c("Male", "Female"))


# REVENEUE
summary(revenue.lm <- glm(complaints ~ revenue + gender + revenue:gender,
                          data = comp, family = poisson))
with(comp, plot(jitter(revenue), complaints,
                col = gender, ylab = "Complaints",
                xlab = "Revenue ($)", main = "Complaints by Revenue and Gender"))
grid <- seq(min(comp$revenue), max(comp$revenue), by = 0.1)

coefs <- coef(revenue.lm)
```

```
female.int <- coefs[1] + coefs[3]
female.slope <- coefs[2] + coefs[4]
fem.pi <- exp(female.int + female.slope * grid)/(1 + exp(female.int + female.slope *
                                                          grid))
male.pi <- exp(coefs[1] + coefs[2] * grid)/(1 + exp(coefs[1] + coefs[2] * grid))
lines(grid, fem.pi, type = "l", col = "red")
lines(grid, male.pi, type = "l", col = "black")
legend("topright", col = c(1, 2), pch = c(0, 1), lty = 1, legend = c("Male", "Female"))

# HOURS
summary(hours.lm <- glm(complaints ~ hours + gender + hours:gender,
                        data = comp, family = poisson))
with(comp, plot(jitter(hours), complaints, col = gender,
                xlab = "Hours", main = "Complaints by Hours and Gender",
                ylab = "Complaints"))
grid <- seq(min(comp$hours), max(comp$hours), by = 0.1)
coefs <- coef(hours.lm)

female.int <- coefs[1] + coefs[3]
female.slope <- coefs[2] + coefs[4]
fem.pi <- exp(female.int + female.slope * grid)/(1 + exp(female.int + female.slope *
                                                          grid))
male.pi <- exp(coefs[1] + coefs[2] * grid)/(1 + exp(coefs[1] + coefs[2] * grid))
lines(grid, fem.pi, type = "l", col = "red")
lines(grid, male.pi, type = "l", col = "black")
legend("topright", col = c(1, 2), pch = c(0, 1), lty = 1, legend = c("Male", "Female"))
```

5 Models were fitted.

1. Poisson Model

```
# POISSON MODEL
summary(p.model <- glm(complaints ~ visits + residency + gender + revenue + hours,
                       data = comp, family = poisson)) # MODEL 1
summary(p.model2 <- glm(complaints~ .^2, data = comp, family = poisson))
stepAIC(p.model2)
summary(final.p<- glm(formula = complaints ~ visits + residency + gender + revenue +    # MODEL 2
                hours + visits:revenue + residency:gender + residency:revenue +
                residency:hours + gender:revenue, family = poisson, data = comp))
summary(lm1 <- update(final.p, . ~ . - residency ))
summary(lm2 <- update(lm1, . ~ . - gender:residency ))
summary(lm3 <- update(lm2, . ~ . - gender ))
summary(lm4 <- update(lm3, . ~ . - hours ))
summary(lm5 <- update(lm4, . ~ . - residency:hours ))
summary(lm6 <- update(lm5, . ~ . - visits ))
summary(lm6<- glm(formula = complaints ~ revenue + visits:revenue + revenue:residency +  # MODEL 3
                    revenue:gender, family = poisson, data = comp))

AIC(p.model, final.p, lm6) # final.p is better i.e, MODEL


# DISPERSION TEST ON MODEL 1 , 2, 3
dispersiontest(p.model)
dispersiontest(final.p)
dispersiontest(lm6)
```

```r
# ROOTOGRAM, PEARSON RESIDUAL'S PLOT FOR POISSON MODEL:
par(mfrow = c(1,3))
rootogram(final.p, main = "Poisson")
plot(residuals(final.p) ~ fitted.values(final.p, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "Poisson")
plot(residuals(final.p, type = "pearson"), ylab = "Pearson residuals",  main = "Poisson")
```

2. Quassi-Poisson Model

```r
# QUASSI POISSON
summary(qp <- glm(complaints ~ visits + residency + gender + revenue + hours,
                  data = comp, family = quasipoisson(link=log)))

# PLOT PEARSON'S RESIDUALS PLOT
par(mfrow = c(1,2))
plot(residuals(qp) ~ fitted.values(qp, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "Quassi-Poisson")
plot(residuals(qp, type = "pearson"), ylab = "Pearson residuals",  main = "Quassi-Poisson")
```

3. Negative Binomial Poisson (NBP) Model:

```r
# NEGATIVE BINOMIAL POISSON
# MODEL 1
summary(nbp<- glm.nb(complaints ~ visits + residency + gender + revenue + hours, data = comp))
dropterm(nbp, test="Chisq")
# its better without removing any term, also visits is highly significant
# however: remove least significant term (highest p-value)
summary(nbp.1 <- update(nbp, . ~ . - residency)) # MODEL 2
summary(nbp.2 <- update(nbp.1, . ~ . - gender)) # MODEL 3
summary(nbp.3 <- update(nbp.2, . ~ . - revenue)) # MODEL 4
summary(nbp.4 <- update(nbp.3, . ~ . - hours)) # MODEL 5

AIC(nbp, nbp.1, nbp.2, nbp.3, nbp.4) # nbp.1 (MODEL 2) is better (without residency)
summary(nbp.1) # better than qp and poisson, less over dispersed (dispersion is less 1)
# final model : nbp.1 (MODEL 2)

# ROOTOGRAM, PEARSON'S RESIDUAL PLOT
par(mfrow = c(1,3))
rootogram(nbp.1, main = "NBP")
plot(residuals(nbp.1) ~ fitted.values(nbp.1, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "NBP")
plot(residuals(nbp.1, type = "pearson"), ylab = "Pearson residuals",  main = "NBP") # best one
```

4. Zero Inflated Poisson Regression (ZIPR) Model:

```r
# ZERO INFLATED POISSON REGRESSION
summary(zip <- zeroinfl(complaints ~ visits + residency + gender + revenue + hours |     # full model
                          residency + gender + revenue + hours, data = comp, dist = "poisson"))
# remove gender from poisson part - highest p-value
summary(zip.1 <- zeroinfl(complaints ~ visits + residency + revenue + hours |           # MODEL 2
                            residency + gender + revenue + hours, data = comp, dist = "poisson"))
# remove revenue from bern part - highest p-value
summary(zip.2 <- zeroinfl(complaints ~ visits + residency + revenue + hours |           # MODEL 3
                            residency + gender + hours, data = comp, dist = "poisson"))
# remove hours from bern part - highest p-value
```

```r
summary(zip.3 <- zeroinfl(complaints ~ visits + residency + revenue + hours |          # MODEL 4
                          residency + gender, data = comp, dist = "poisson"))
# remove residency from poisson and bern part
summary(zip.4 <- zeroinfl(complaints ~ visits + revenue + hours |                       # MODEL 5
                          gender, data = comp, dist = "poisson"))
# remove residency from poisson part
summary(zip.5 <- zeroinfl(complaints ~ visits +  revenue + hours |                      # MODEL 6
                          gender + log(visits)+ log(hours), data = comp, dist = "poisson"))


AIC(zip, zip.1, zip.2, zip.3, zip.4, zip.5) # compare all models
rootogram(zip.1)
rootogram(zip.2)
rootogram(zip.3)
rootogram(zip.4)
rootogram(zip.5) # BEST ONE

# ROOTOGRAM AND PEARSON'S RESIDUAL PLOT
par(mfrow = c(1,3))
plot(residuals(zip.5) ~ fitted.values(zip.5, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "ZIPR")
plot(residuals(zip.5, type = "pearson"), ylab = "Pearson residuals", main = "ZIPR")
```

5. Zero Inflated Negative Binomial (ZINB) Model:

```r
```r
# ZERO INFLATED NEGATIVE BINOMIAL
summary(zinb <- zeroinfl(complaints ~ log(visits) + residency + gender + revenue +log(hours) |
                           log(visits) + residency + gender + revenue + log(hours),
                         data = comp, dist = "negbin"))

summary(zinb.1<- zeroinfl(complaints ~ log(visits) + residency  + revenue + log(hours) |
                  log(visits)  + gender + revenue , data = comp, dist = "negbin"))

summary(zinb.2<- zeroinfl(complaints ~ log(visits) + residency  + revenue + log(hours) |
                            log(visits)  + gender , data = comp, dist = "negbin"))

summary(zinb.3<- zeroinfl(complaints ~ log(visits) + revenue + log(hours) |
                            log(visits)  + gender , data = comp, dist = "negbin"))




summary(zinb.4 <- zeroinfl(complaints ~ log(visits) +log(hours)+ revenue+gender + residency|
                             log(visits)+log(hours) + gender + revenue, data = comp, dist = "negbin"))

summary(zinb.5 <- zeroinfl(complaints ~ log(visits) +log(hours) + revenue|
                             log(visits) +log(hours) + gender +revenue , data = comp, dist = "negbin"))

# COMPARE ALL MODELS
AIC(zinb, zinb.1, zinb.2, zinb.3, zinb.4, zinb.5) # MODEL 4 has lower AIC
par(mfrow = c(2,2))
rootogram(zinb.4)
rootogram(zinb.5) #  but MODEL 5 has a better rootogram
```
```

```
# COMPARE RESIDUAL PLOT FOR MODEL 4 AND 5:

# MODEL 4
plot(residuals(zinb.4) ~ fitted.values(zinb.4, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "ZINB")
plot(residuals(zinb.4, type = "pearson"), ylab = "Pearson residuals",  main = "ZINB")
summary(zinb.4)
# MODEL 5
plot(residuals(zinb.5) ~ fitted.values(zinb.5, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "ZINB")
plot(residuals(zinb.5, type = "pearson"), ylab = "Pearson residuals",  main = "ZINB")
summary(zinb.5)

# MODEL 5 HAS A BETTER REISDUALS PLOT AND ROOTOGRAM, HENCE THE BEST MODEL
# FIGURE 7
par(mfrow = c(1,3))
rootogram(zinb.5, main = "ZINB")
plot(residuals(zinb.5) ~ fitted.values(zinb.5, type = "pearson"), xlab = "Fitted values",
     ylab = "Pearson Residuals" , main = "ZINB")
plot(residuals(zinb.5, type = "pearson"), ylab = "Pearson residuals",  main = "ZINB")
summary(zinb.5)
```