Predicting the Number of Patients’ Complaints against Emergency Department Doctors Based on Doctor’s Demographic Characteristics

Fanchao(Franco) Meng 23370209

05 November, 2022

Abstract

Health care system, has always been one of the key sectors in the societies. With the current global pandemic environment, the emergency department has been under enormous pressure to keep up with the demand. It is inevitable that complaints from patients may often be made against doctors. The dramatic increase in the size and frequency of medical complaints over the last decade, has made the development of modeling on related issues, increasingly important (Burce Cooil (1991)). The background for complaints is related both to the doctor and medical practice in general, as well as to the patients (Bianca Hanganu (2022)). The aim for this paper is to identify a model, that provides predictions and explanations of the number of claims for a certain doctor, based on doctor’s social-demographic and professional characteristics. The data file given by the unit coordinator, contains 94 doctors records who worked in an emergency service at a hospital. We analysis the number of the complaints each doctor received in one year of time, based on doctors gender, residency training status, hourly income and the total number of hours the doctor worked in a year. The analysis will help us explore and discover which measurement may affect the changes in the number of complaints.

A zero inflated poisson model was fitted to the data. The model has revealed that, the doctors who had more numbers of total patient visits in one year, are more likely to receive an increasing number of total complaints during that time frame, a factor change of 1.15 per 100 increase in the total number of patient visits. Doctors with higher hourly income rate or more accumulated working hours are more likely to receive less number of complaints, a factor change of 0.991 per 1 dollar increase in hourly rate, and a factor change of 0.92 per 100 increase in the number of hours worked in a year respectively.

Moreover, being female without any residency training, or being male with residency training , increase the odds of receiving no complaint significantly, comparing with the group of being female doctors with residency training, or male without any residency training.

1 Introduction

With the current ongoing pandemic crisis, health care industry, has become the most important sector globally. A better health care system, will provide a solid foundation to battle with contagious diseases, and to maintain a well functioning society.

In order to improve the health care system, patients’ complaint, is one of the crucial areas to be investigated. The better understanding on why the complaints were made, the more efficient to address and resolve the issues. The complaints may concerned different domains, from clinical care, management to patient or caregiver relationship. Patient complaints can provide useful insights into quality of care and patient dissatisfaction (David M. Studder Matthew J Spittal Marie M Bismark (2011)). This paper is focused on investigating the relationship between the complaints and the doctors, who worked in an emergency service at a hospital. The data contained 94 doctors with some of their demographic information. Various literature have also inspected the similar matters.

In the book 'Applied Categorical Data Analysis and Translational Research', Chap T.Le (2010) investigated a very similar data with 44 observations, which including same variables. A simple poisson regression model has been fit in the data. The result indicates that the common perception is almost true, that those without previous training are more likely to receive the same number of complaints as those who were trained in the specialty, the model showed the likelihood will be increased by 36%.

To identify characteristics of doctors who are repeated subjects of complaints made by patients, David M. Studder Matthew J Spittal Marie M Bismark (2011) designed a case-control study of doctors whom patients had complained to the Victoria Health Commissioner between 1 January 2000 and 31 December 2009. The case comprised 96 doctors who were the subject of four or more separate complaints, and the control group comprised 288 doctors who were subjects of one single complaint over the study period.

The multivariate analyses showed that certain distinctive characteristics of complain-prone doctors were: being male, or the training was completed within Australia. Surgeons and psychiatrists also had higher odds of being in the complaint-prone group than general practitioners.(David M. Studder Matthew J Spittal Marie M Bismark (2011))

In another article produced by the same group of researchers, David M. Studdert Matthew J Spittal Marie M Bismark (2015) aimed to develop a reliable scoring system and algorithm for identifying, and predicting Australia doctors' risk of becoming the subject of patient complaints. 2011-2016 administrative data collected by the national regulator of healthy practitioners in Australia (AHPRA) has been used , a retrospective cohort study of 14 registered health professions has been conducted. A total of 715,415 registered health practitioners were participated, the modelling incorporated multivariate predictors including: gender, age, profession and specialty, number of prior complaints and complaint issue.

The research found out that male practitioners' compliant risk was 1.5 times that of female practitioners, there was no significant difference in complaint risk for practitioners 26-35 years and those ages less than 25 years, but practitioners in the older age group had 1.5 to 2.1 times higher risk and it increased with age. Practitioners working based in regional Australia also had 1.1 times higher complaint risk comparing with working in major cities, and those in remote area had 1.3 times the risk with practitioners in the cities. The finding also thoroughly explored the differences between different profession and medical specialty, however this information was not presented in our dataset.

Bianca Hanganu (2022) utilized a binomial logistic regression model to identify independent predictors for Romanian doctors who are more prone to receiving complaints . The analysis was based on a data set consisted of 1684 doctors records, which detailed each doctor's social-demographic (gender, age, marital status etc), professional (medical specialty, area of activity), and institutional characteristics (type of medical institution, type of patients).

The result showed male doctors face complaints more often than female doctors, doctors with children declared to a higher extend that they experienced complaints than doctors without children. There are also significant differences between the doctors with different professional degrees. Senior doctors, and specialist are more likely to receive complaints as opposed to resident doctors. While surgical specialties have the highest risk receiving complaints.

Sara C. Charles (1992)'s article on predicting risk for medical malpractice claims using quality of care characteristics, has explored various characteristics which may contribute to higher risk of doctor receiving complaints. The "high risk" and "low risk" categories were distinguished based on the payouts amount. The chi-squared statistic was used for categorical variables, one-way analysis of variance was used for continuous variables, and individual pairwise group differences were computed using Duncan's multiple range test. Variables found significant in individuals logistic regressions within each Donabedian component area were combined into a single logistic regression analysis.

The model shows significant predictors of the high-risk group include increasing age, surgical specialty, emergency department coverage, increased number of days away from practice, and feeling that the current climate of litigation is unfair.

This report is constructed as follows: the methodology section, will detail the computer tools, and statistical methods used for the modelling process, the steps to be undertaken and the reason to use certain methods. Followed by a result section, where all the detailed data summary will be given, and the final model building, validation process with the interpretation of the final model. This is followed by a discussion of the findings, explaining reasons for the model validation and selection, comparing with other similar modelling results from literature review, and listing certain approaches on how the model can be improved.

2 Methodology

R statistical tool environment was used for the modelling process, with multiple packages including MASS, AER, Lattice, KableExtra, Countreg, Lmtest etc. Statistical significance was determined at the 5% (0.05) level.

In order to build a model to fit the data, all the variables in the dataset were firstly transformed into correct data type, namely categorical and numeric variables. The dataset was then explored by numerical and graphical summaries. The numerical summary included results of minimum, maximum, median, mean, counts, range, standard deviation for each variables, to ensure all the data entries were proper and there is no invalid, missing entries in the dataset. The graphic summaries, including scatter plots, box plots, histogram were produced, aiming to explore the relationship between two or three different variables. These plots will facilitate us to visually examine relationships between variables. A correlation map with correlation coefficients, has also been produced to further explore the degree of the correlation between each pair of variables.

The dataset includes the number of complaints against each doctor, which will be our response variables for the model, along with demographic characteristics of the doctors as explanatory variables. As the number of complaints are discreet, non-negative count data, and the events were independently occurred. The data were firstly fit into a log-linear regression for poisson counts, with all the variables and two way interactions between categorical variables considered.

As Alain F. Zuur (2015) mentioned in the book 'Mixed Effects Models and Extensions in Ecology with R', the over/under dispersion, and zero inflated/truncated data are common issues to properly fit in a log-linear poisson regression. It is safest to assume these issues are present in our modelling process. Ignoring zero inflation can cause the estimated parameters and standard errors biased, and the extra zeros can also cause over-dispersion. A histogram of complaints counts grouped by gender and residency was produced, as an evidence that zero-inflated issue existed in our dataset. In order to examine the extra-poisson variation, we here used both numerical and graphical approach, by observing plots of number of complaints against each variables, conducting dispersion tests, and visually checking the Pearson residue plots after fitting the model.

The data then has been fitted in the following model selection process, with interactions between categorical variables considered. In the order of : poisson regression, quasi-Poisson regression, negative binomial regression, zero-inflated poisson regression, and zero-inflated negative binomial. The below table has demonstrated the reason for this approach. The table has listed the model variance parameters for these 5 models, how the different models gradually allow for extra over dispersion from the data, and dealing with extra zeroes issues : Fred L.Ramsey (2013)

During the model selection process, the dispersion test, residual deviance, rootogram, and Pearson residual plots were used to examine if the model is improving, or the issues like under fitting zero counts, and whether the over dispersion issue is still existed.

Once the most suitable model has been decided, the model was reduced to the final model,by using likelihood ratio tests, to examine the significance of each term, accompanied by Wald test P value , AIC (Akaike information criterion) , chi-square statistic of the goodness of fit. The MLE (maximum likelihood estimation)

Table 1: Models allow for extra dispersion

	Models	Variance
1	Poisson	μ
2	Quasi Poisson	$\Psi\mu$
3	Negative Binomial	μ (1 + $\Phi\mu$)
4	Zero Inflated Poisson	(1-π) * (μ + π x μ^2)
5	Zeor Inflated Negative Binomial	(1-π) * (μ + $\frac{\mu^2}{k}$) + μ^2 * (π^2 + π)

was used to estimate the parameters of the model. The necessity of the interaction terms between categorical variables will be explored, Pearson residue plot, rootogram will also facilitate to monitor the model improvement.

3 Results

Summary Tables

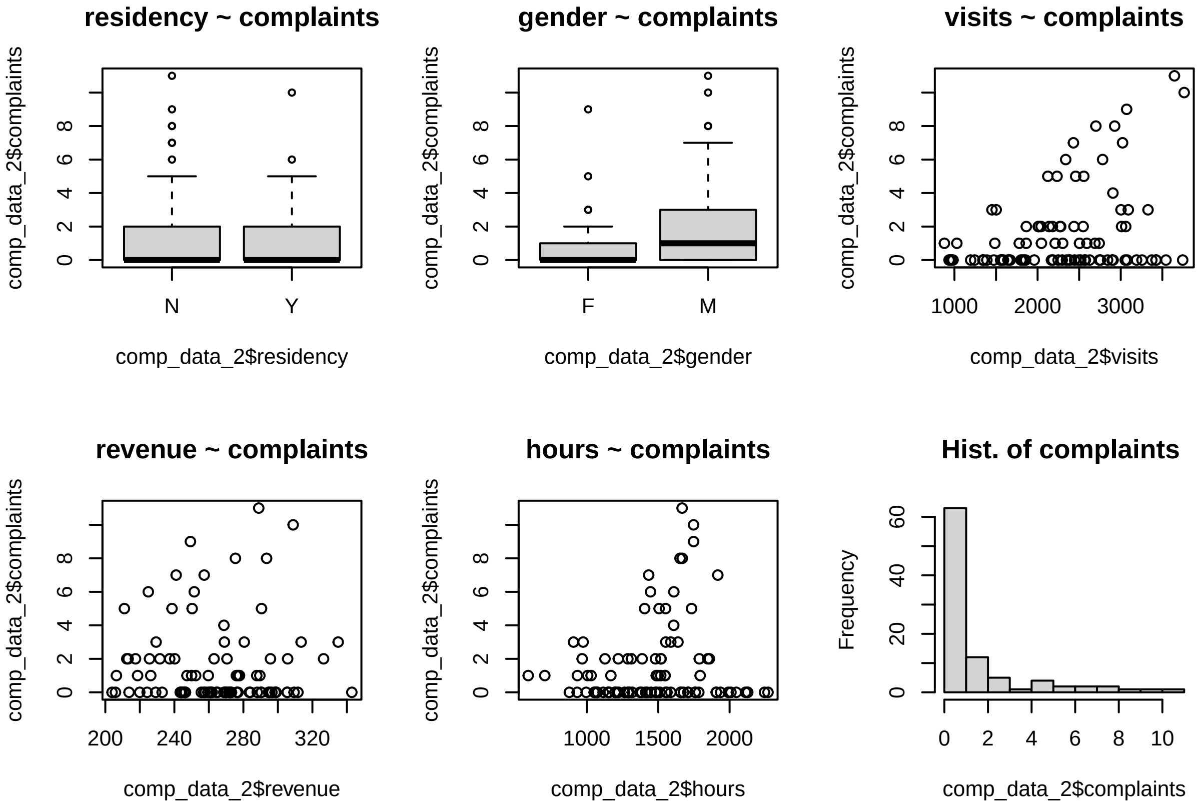
The data contains information for 94 doctors in emergency department. A description and summary statistics are presented in table below, divided by numeric variables and categorical variables. Out of 94 doctors, 37 were female and 57 males, 49 were not in residency training, while the rest 45 were. There was no obvious issue with the data, the numeric data were fairly normally distributed.

	Residency Training N	Residency Training Y	Row Total
Female	30	7	37
Male	19	38	57
Column Total	49	45	94 (obs. in data)

	vars	n	mean	sd	min	max	range	se
visits	1	94	2270.59574	723.91900	879.0000	3763.0000	2884.0000	74.6665436
complaints	2	94	1.56383	2.52113	0.0000	11.0000	11.0000	0.2600347
revenue	3	94	263.76358	30.90041	203.8928	342.8607	138.9678	3.1871335
hours	4	94	1468.91014	351.35986	589.0000	2269.0726	1680.0726	36.2400027

residency : is the doctor in residency training (Y = Yes, N = No)
gender: gender of the doctor (M = male, F = female)
visits : the number of patient visits
complaints : the number of complaints against the doctor in previous year
revenue: doctor's hourly income (dollars)
hours: total number of hours the doctor worked in a year

Graphic summaries (boxplots, scatter plots, histogram) has also been produced as below. The median number of complaints is nearly zero for the both group with and without residency training, and for the female doctor group too, while the male group has significant higher median and bigger range in terms of number of complaints. The number of visits appear to positively correlated with number of complaints in none zero part of the data, and lastly the histogram shows our data is zero inflated .

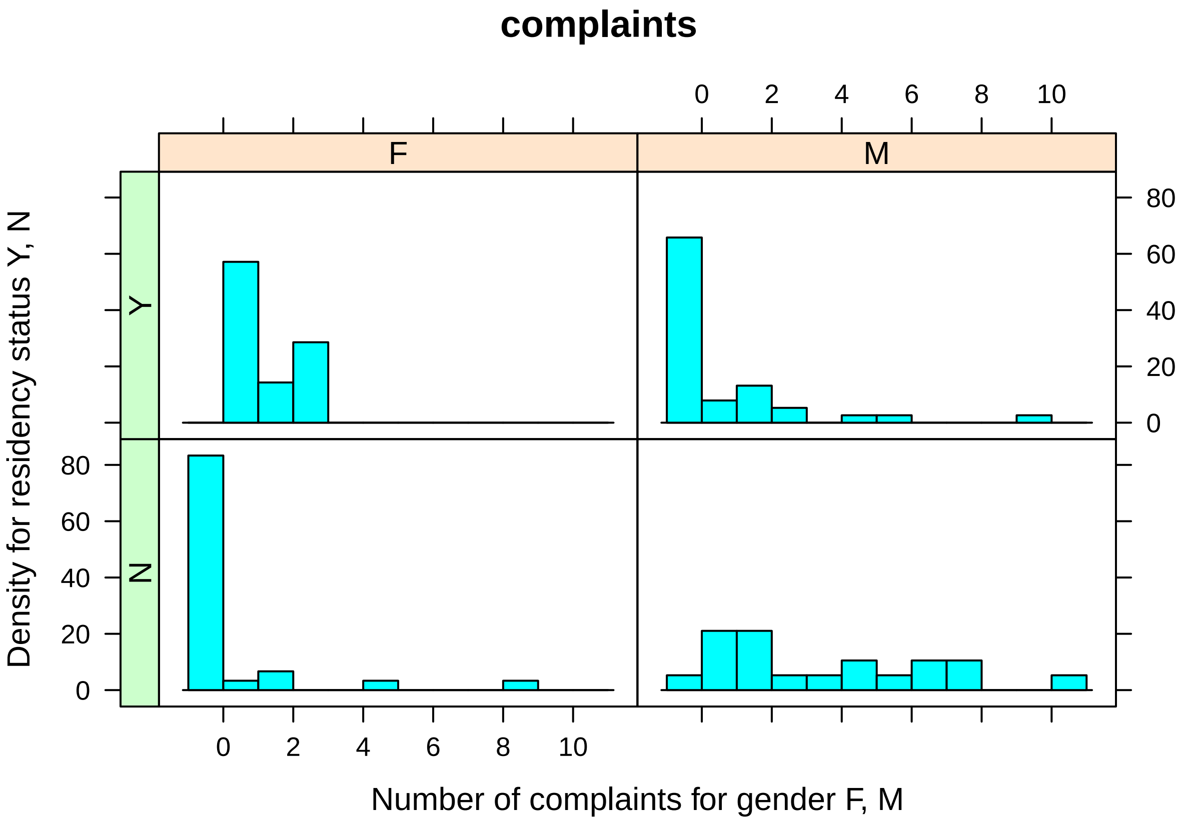


Correlation Map (Appendix A.1)

The categorical data residency and gender, were factored into numeric value in order to produce correlation map, (female = 1, male = 2, no residency training = 1, residency training = 2). The positive and negative correlation are indicated by blue and red colors, with the correlation coefficient listed. 95 percent of significance level is used for the P value, the × indicates the correlation is not significant. The map presented there is no strong correlated relationship (no correlation coefficient lies between ± 0.50 and ± 1) between any sets of variables.

Histogram

The below histogram plotted the number of complaints, split by gender and residency status, shows that the zero inflation is present in our data, and specifically in the groups of : Female with no residency training, and Male with residency training. This information is valuable for us to deal with zero inflation issue later on.



A poisson regression has been initially fit to our data, the formula below has been used for model comparison and selection.Interaction between residency and gender was included.

$$formula = complaints \sim visits + residency * gender + revenue + hours$$

The detailed step by step model fitting and selection process with R code is listed in appendix A.2, below table listed AIC values for each models, and dispersion parameter, by using dispersion test on the poisson model, or the ratio of residual deviance and residual degrees of freedom.

The most suitable model, zero-inflated poisson, has been selected based on checking the AIC, dispersion parameters, rootogram, and pearson residuals plots. The likelihood ratio test has been conducted between nested models, eg: poisson is nested in zero inflated poisson, ZIP is nested in ZINB. Below table listed dispersion parameters, degrees of freedom, and AIC for the model validation process.

	Poisson	quasi-Poisson	Negative Binomial	Zero-Inflated Poisson	Zero-inflated NB
dispersion	2.776	2.249	0.991	-	-
df	7	7	8	13	14
AIC	332.39	-	293.95	266.536	268.206

In order to reduce the model , the likelihood ratio test has been conducted, along with rootogram and pearson residue plots. Below table shows the significance to keep the interactions: between gender and residency in the zero-inflation part of the model, the p value was given by likelihood ratio test. The AIC value was increased from 257 to 288 in the interaction dropped from count model.

Here are the reported results

for the count model :

$$Gender \times\ Residency\ Interaction : \mathbf{Pr(\chi^2_1 > 32.59) = 1.136^{-8}}$$

for the logistic model:

$$Visits : \mathbf{Pr(\chi^2_1 > 15.478) = 8.346^{-5}}$$

$$Revenue : \mathbf{Pr(\chi^2_1 > 5.3696) = 0.02049}$$

$$Hours : \mathbf{P = 0.044988}$$

* Wald test P value reported for the hours, dropping hours causing misfit NAs.

The final model variables and coefficients are given below, along with factor change in number of complaints, for the indicated change in the variable. Due to the large range for visits : 879 to 3763 , and large range for hours: 589 to 2269.0726, both variable has been interpreted in the change of every 100 rather than 1.

Count model coefficients (poisson with log link):	Coefficient	factor Change in number of complaints
visits	0.001433	1.154076 (per 100 increase)
revenue	-0.009228	0.9908143 (per $1 increase)
hours	-0.000837	0.9197071 (per 100 increase)
Zero-inflation model coefficients (binomial with logit link):	Coefficient	odds ratio (change) of having zero complaints
(Intercept)	1.265	3.543235
Residency Yes	-14.2266	6.629167e-07
Gender Male	-14.516	4.9632e-07
Residency Yes : Gender Male	27.6935	1.064516e+12
AIC	257.8263	

The followings are the findings from the model:

To predict none zero numbers of complaints:

1: For every 100 increase in the number of visits, the estimated mean number of complains increase by factor of 1.15 on average.

2: For every $1 increase in the revenue, the estimated mean number of complains decrease by factor of 0.991 on average.

3: For every 100 increase in the number of hours, the estimated mean number of complains decrease by factor of 0.92 on average.

To predict the odds of having zero complaint:

Male doctors, with residency training, and female doctors, without residency training, are significantly more likely to receive no complaints, than female doctors with residency training, or male doctors without residency training.

Discussion

There are several researchers using various models to predict number of complaints doctor received, some of them covered in the introduction literature review. We here used a zero inflated poisson model for the number of complaints doctor received in previous year, based on 5 characteristics.

1. the modeling initial visual plots discovered a strong interaction relationship between gender and residency, that male doctors with residency training, and female doctors without residency training are a lot more likely to receive no complaints. The Wald test P value shows the interaction was not significant in the zero count part of the model, the coefficients are very large and the standard errors are even larger, an indication that the model might have some issue with the interaction term.

The below table has quantified our data:

	Number of Obs with at least 1 complaint	Number of Obs with 0 complaint
Female Not in Residency Training	5	25
Female In Residency Training	7	0
Male Not in Residency Training	18	1
Male In Residency Training	13	25

The complete separation / quasi-separation issues are existed. That based on our data, the prediction can be made that if you are female in residency training, you would definitely get at lease 1 complaints, and similar situation goes to female not in residency training, where only 1 exception presented in our data.

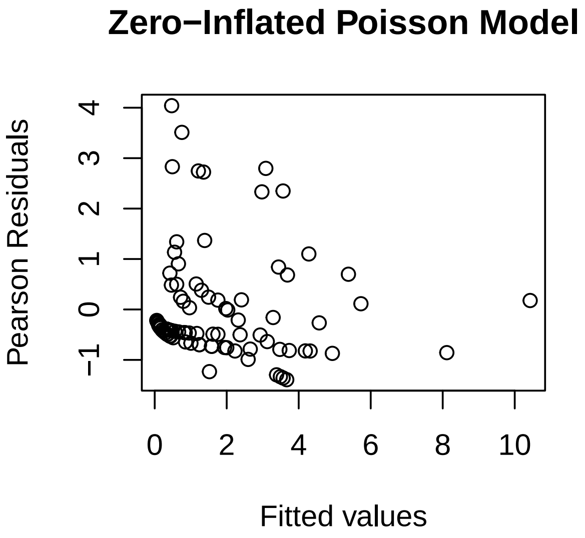
We can try to omit the interaction terms, however it would leads bias estimates for the other predictor variables in the model. The likelihood ratio test has provided an overwhelming evidence (p = 1.136e-08) to include interaction terms, the significant decrease in AIC also shows the significance of the interactions.

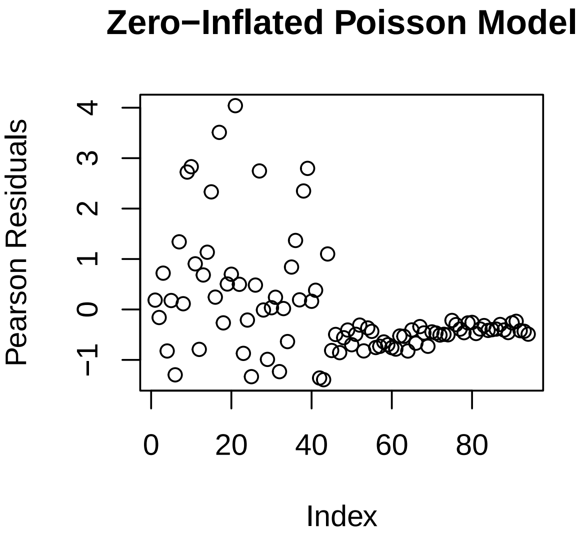
2. Separation occurs when the dataset is too small to observe events with low probabilities. The female in residency training group only has 7 observations in our dataset. The issue may be improved if more observation presented. For the interaction terms between gender and residency, we can try remove the predictor, collapse predictor categories, or re-expressing the predictors. However these are not very practical in our case.

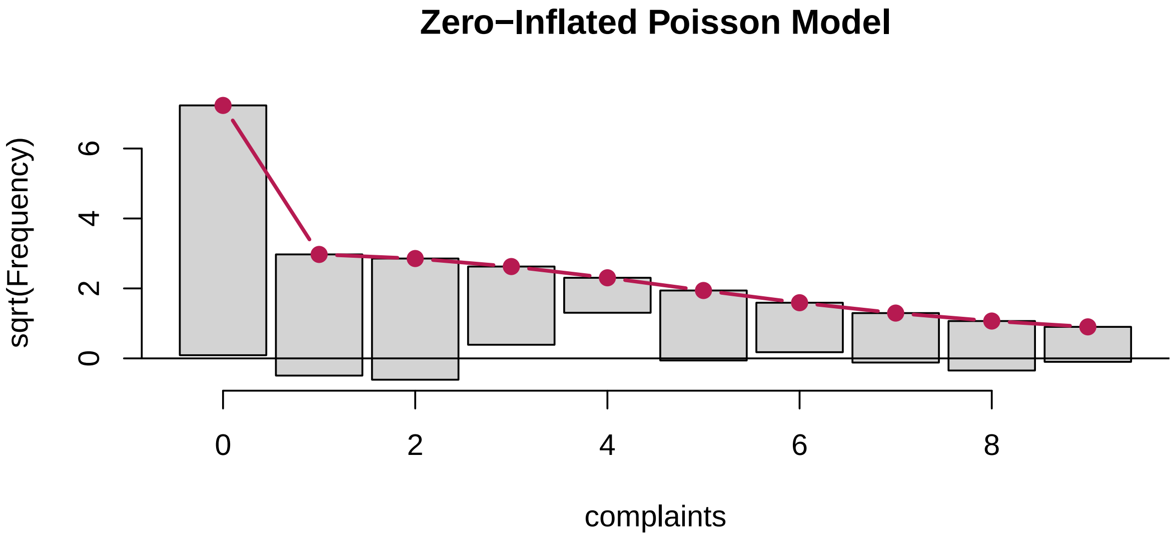
3. Bayesian method can be used when some additional information given. The model may be improved if regularization or Bayesian priors are used.

4. The negative binomial model also has dealt with over dispersion issue well. However the model includes more interactions: between residency and hours, gender and revenue. The AIC and residue plot comparison shows the simpler ZIP model perform better.

5. Plots of the Pearson residuals below indicates the presence of two outliers (>3 standardized distance). Index 17 & 21 in the data. The rootogram shows the zero counts fitted well, some other counts are slightly under or over fit but overall it was satisfactory.







5. some other characteristics, which were not included in our dataset, may be valuable to better predict the number of complaints. for example, doctor's age and hospital location (city/rural) are significant in David M. Studdert Matthew J Spittal Marie M Bismark (2015)'s finding. Certain doctor's occupation specialties, and the levels of their professional degrees, may also affect the number of received complaints (Bianca Hanganu (2022))

6. When swapping two parts of our model, move the interaction term to the poisson part of the model, and move revenue,hours,and visits to zero count part of the model, all the terms were statistically significant. However AIC was increased significantly, and the residue appears to have much larger variance (-2 to 8) .

References

Alain F. Zuur, Neil Walker, Elena N. Ieno. 2015. *Mixed Effects Models and Extensions in Ecology with r.* Springer. https://www.springer.com/series/2848.

Bianca Hanganu, Lavinia Maria Pop, Magdalena Iorga. 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.

Burce Cooil. 1991. "Using Medical Malpractice Data to Predict the Frequency of Claims: A Study of Poisson Process Models with Random Effects." *Journal of the American Statistical Association* 86, No 414: 285–95. https://doi.org/10.2307/2290560.

Chap T.Le. 2010. *Applied Categorical Data Analysis and Translational Research.* 1st ed. Hoboken, New Jersey, USA: John Wiley & Sons.

Fred L.Ramsey, Daniel W. Schafer. 2013. *The Statistical Sleuth: A Course in Methods of Data Analysis.* 3rd ed. 20 Channel Centre Street, Boston, USA: CENGAGE Learning.

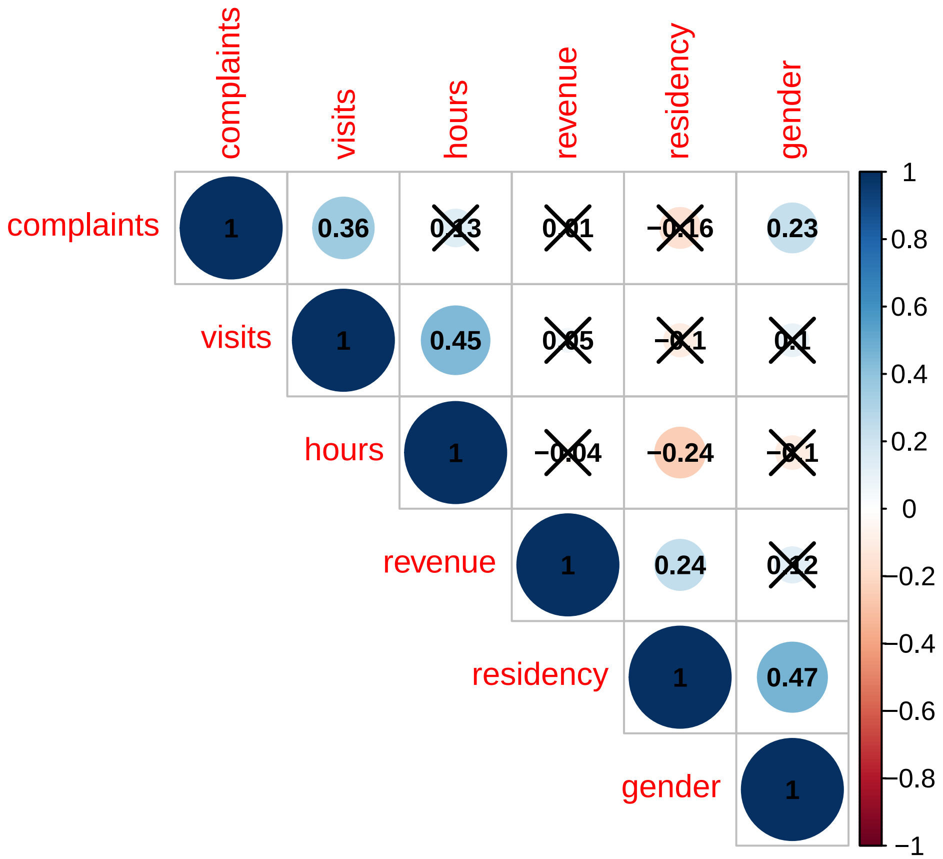
Matthew J Spittal, David M Studder, Marie M Bismark. 2011. "Prevalence and Characteristics of Complaint-Prone Doctors in Private Practice in Victoria." *Med J*, 195. https://doi.org/10.5694/j.1326-5377.2011.tb03183.x.

Matthew J Spittal, David M Studdert, Marie M Bismark. 2015. "The PRONE Score: An Algorithm for Predicting Doctors' Risks of Formal Patient Complaints Using Routinely Collected Administrative Data." *BMJ Quality & Safety, 24(6)*, 360–68. https://doi.org/10.1136/bmjqs-2014-003834.

Sara C. Charles, Paul R. Frisch, Robert D. Gibbons. 1992. "Predicting Risk for Medical Malpractice Claims Using Quality-of-Care Characteristics." *West J Med* 157(4): 433–39. https://doi.org/10.3390/medicina58020287.

Appendix

Appendix A.1



*The positive and negative correlation are indicated by blue and red colors, with the correlation coefficient listed. 95 percent of significance level is used for the P value, the × indicates the correlation is not significant.

Appendix A.2

R Code

load Packages

library(corrplot) library(knitr) library(MASS) library(AER) library(lattice) library(papeR) library(kableExtra) library(latticeExtra) library(countreg) library(lmtest) library(psych) library(equatiomatic) library(pander)

load data in

```
comp_data_2 <-read.table("compdat.txt", header=T, stringsAsFactors = T, sep = "\t")
summary(comp_data_2)
```

```
##      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
```

```
str(comp_data_2)
```

```
## 'data.frame':    94 obs. of  6 variables:
##  $ visits    : int  2014 3091 879 1780 3646 2690 1864 2782 3071 1502 ...
##  $ complaints: int  2 3 1 1 11 1 2 6 9 3 ...
##  $ residency : Factor w/ 2 levels "N","Y": 2 1 2 1 1 1 2 1 1 2 ...
##  $ gender    : Factor w/ 2 levels "F","M": 1 2 2 2 2 2 2 2 1 2 ...
##  $ revenue   : num  263 335 206 226 289 ...
##  $ hours     : num  1287 1588 705 1006 1667 ...
```

producing numeric and graphic summary tables

```
tab <- matrix(c(30, 7, 37, 19, 38, 57, 49, 45, "94 (obs. in data)"), ncol=3, byrow=TRUE)
colnames(tab) <- c('Residency Training N','Residency Training Y','Row Total')
rownames(tab) <- c('Female','Male','Column Total')
tab <- as.table(tab)
y = kbl(tab,longtable = T, booktabs = TRUE, align="c", linesep = "")%>%
  kable_styling() %>%
  row_spec(0,background="yellow")
y
```

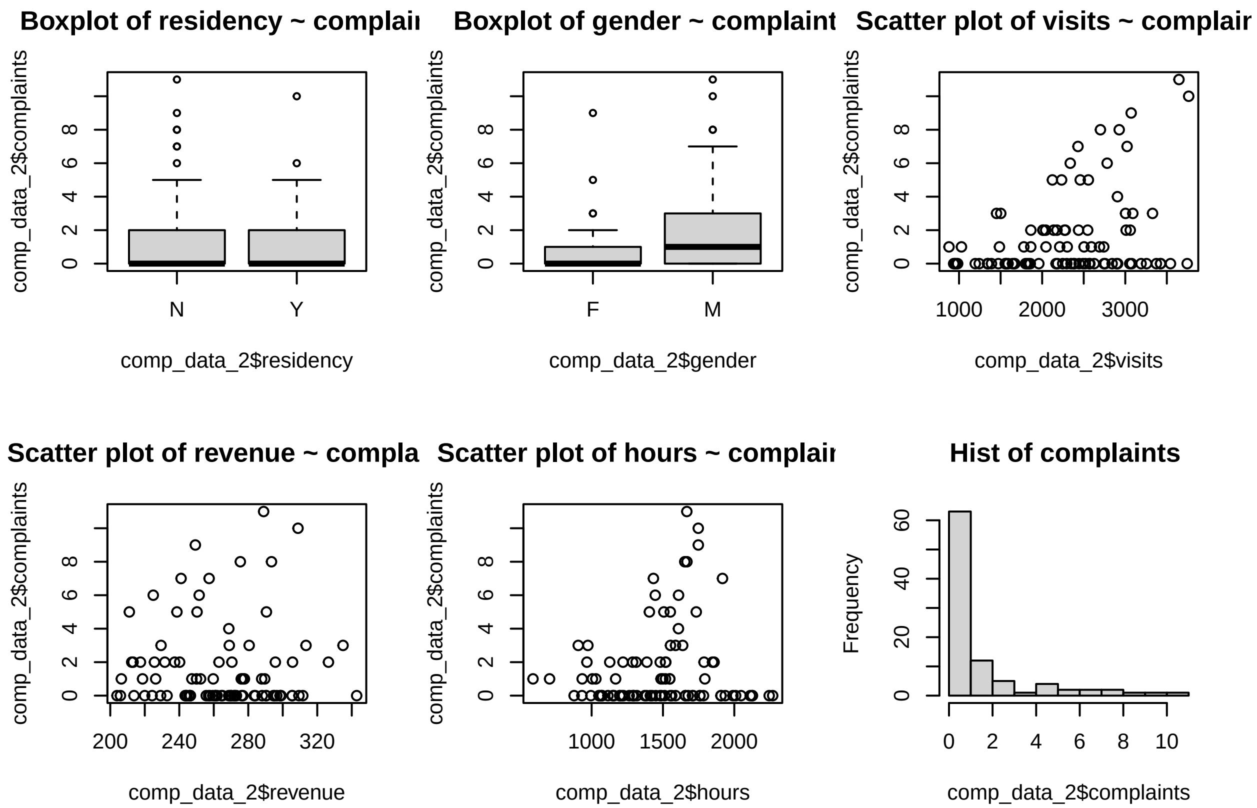
	Residency Training N	Residency Training Y	Row Total
Female	30	7	37
Male	19	38	57
Column Total	49	45	94 (obs. in data)

```
x = kbl(describe(comp_data_2[, c("visits", "complaints", "revenue", "hours")], fast=TRUE),longtable = F
  kable_styling(latex_options = c( "hold_position","striped","scale_down"))%>%
  row_spec(0,background="yellow")
add_footnote(x, c("  "," ", "residency : is the doctor in residency training (Y = Yes, N = No)", "gende
```

	vars	n	mean	sd	min	max	range	se
visits	1	94	2270.59574	723.91900	879.0000	3763.0000	2884.0000	74.6665436
complaints	2	94	1.56383	2.52113	0.0000	11.0000	11.0000	0.2600347
revenue	3	94	263.76358	30.90041	203.8928	342.8607	138.9678	3.1871335
hours	4	94	1468.91014	351.35986	589.0000	2269.0726	1680.0726	36.2400027

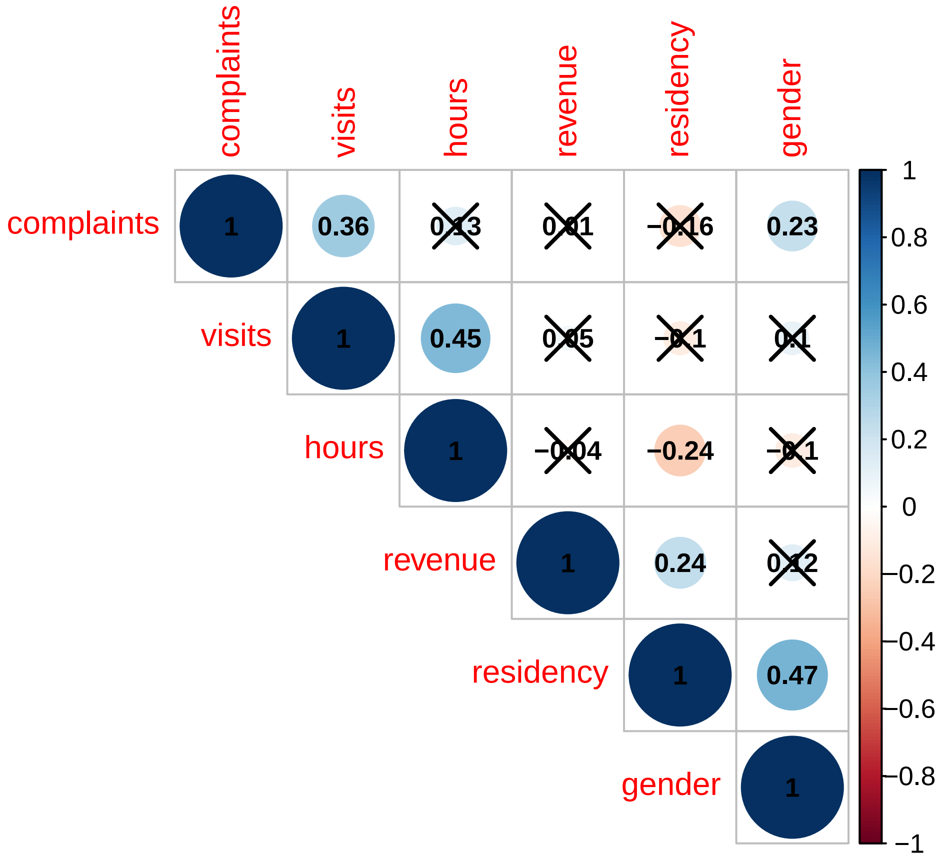
residency : is the doctor in residency training (Y = Yes, N = No)
gender: gender of the doctor (M = male, F = female)
visits : the number of patient visits
complaints : the number of complaints against the doctor in previous year
revenue: doctor's hourly income (dollars)
hours: total number of hours the doctor worked in a year

```
par(mfrow=c(2,3))
plot(comp_data_2$complaints ~ comp_data_2$residency, main = "Boxplot of residency ~ complaints")
plot(comp_data_2$complaints ~ comp_data_2$gender,main = "Boxplot of gender ~ complaints")
plot(comp_data_2$complaints ~ comp_data_2$visits,main = "Scatter plot of visits ~ complaints")
plot(comp_data_2$complaints ~ comp_data_2$revenue,main = "Scatter plot of revenue ~ complaints")
plot(comp_data_2$complaints ~ comp_data_2$hours,main = "Scatter plot of hours ~ complaints")
hist(comp_data_2$complaints,main = "Hist of complaints ")
```



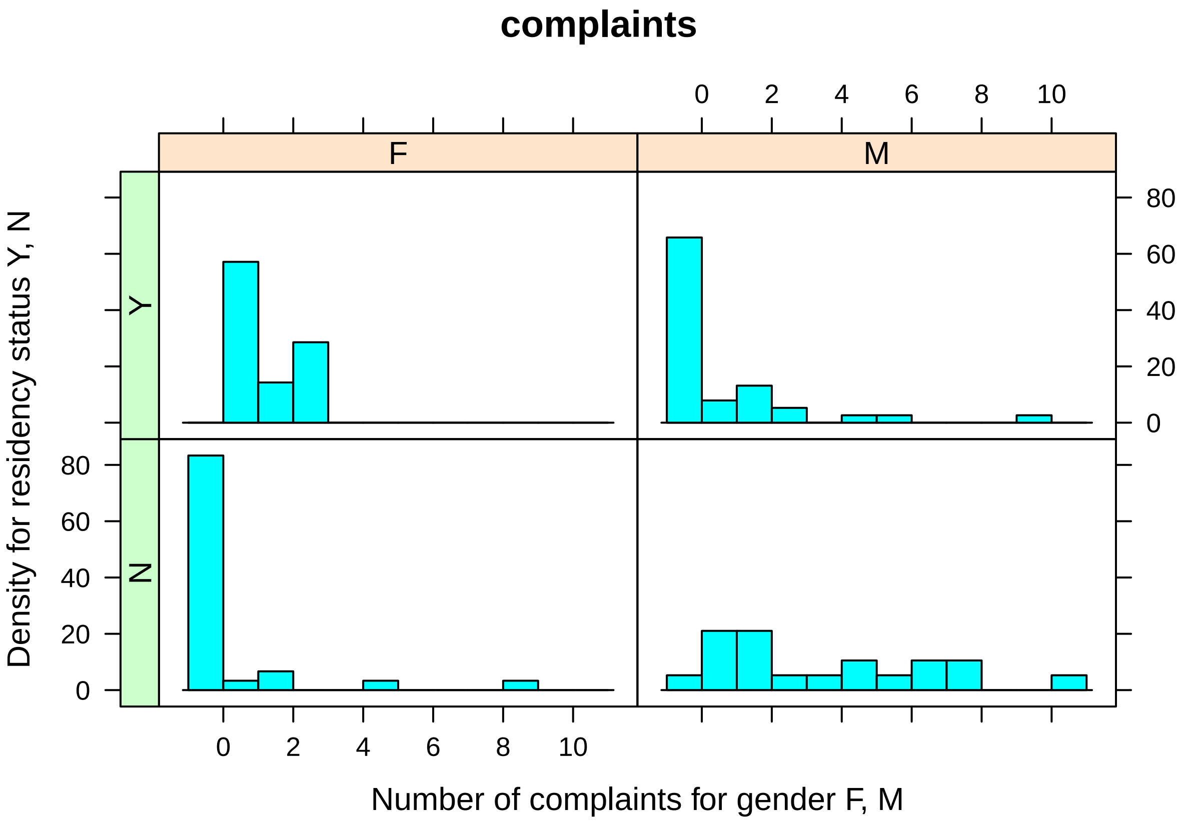
Producing correlation heat map

```
comp_data_3 <- within(comp_data_2, {residency <- as.numeric(residency)
gender <- as.numeric(gender)})



M_1 <- cor(comp_data_3)
test_Res_1 = cor.mtest(comp_data_3, conf.level = 0.95)

corrplot( M_1, p.mat = test_Res_1$p, method = 'circle',  type = 'upper', insig='pch',addCoef.col ='blac
```



Producing histogram

```
histcomp <- histogram (~complaints | gender * residency, comp_data_2, breaks = 0:12 - 1, xlab = "Number
useOuterStrips(histcomp)
```



building initial formula for data selection

1. poisson model with summary and diagnostics, dispersion test etc

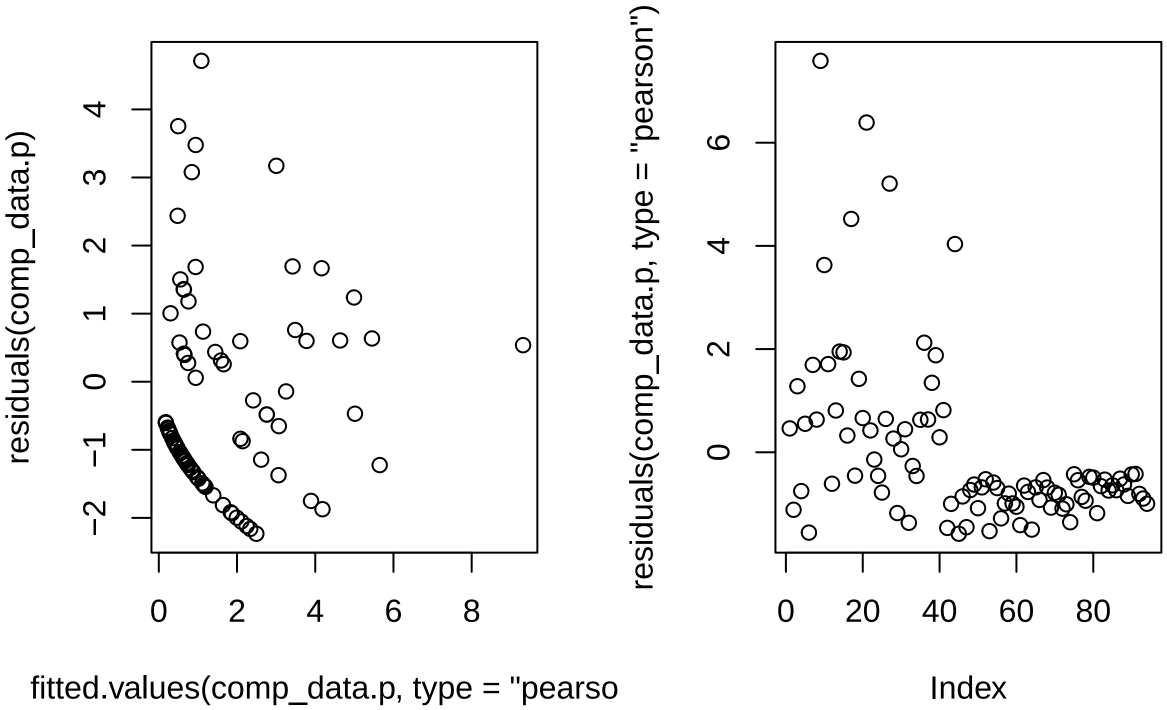
```
f <- formula(complaints ~ visits + residency * gender + revenue + hours)
comp_data.p <- glm(f, data = comp_data_2, family = poisson)
summary(comp_data.p)
```

```
##
## Call:
## glm(formula = f, family = poisson, data = comp_data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2330  -1.2243  -0.8338   0.4307   4.7130
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.2075577  0.8407904  -2.626  0.00865 **
## visits              0.0008632  0.0001757   4.914 8.92e-07 ***
## residencyY          1.1528999  0.3767804   3.060  0.00221 **
## genderM             1.6677457  0.2597078   6.422 1.35e-10 ***
## revenue            -0.0006513  0.0028746  -0.227  0.82075
## hours              -0.0001127  0.0003616  -0.312  0.75527
## residencyY:genderM -2.3631371  0.4207960  -5.616 1.96e-08 ***
## ---
## 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: 195.61  on 87  degrees of freedom
## AIC: 332.39
##
## Number of Fisher Scoring iterations: 6
```

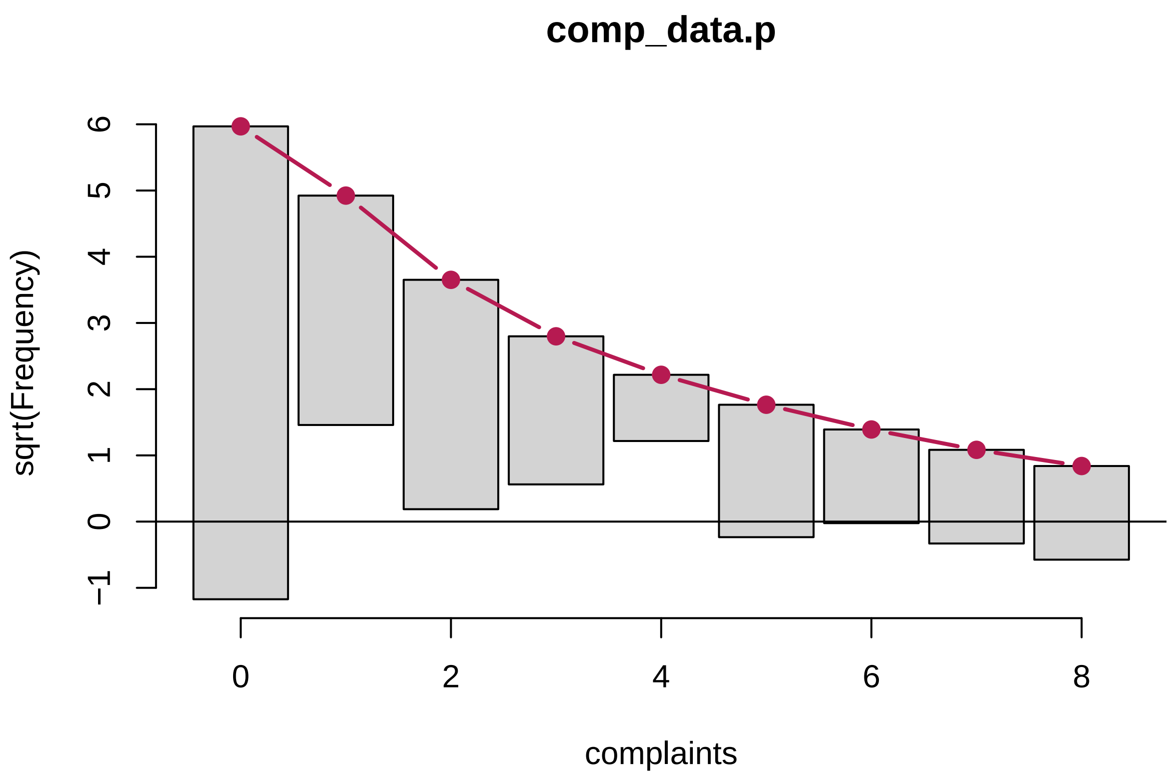
```
dispersiontest(comp_data.p)
```

```
##
##  Overdispersion test
##
## data:  comp_data.p
## z = 2.6148, p-value = 0.004463
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##    2.77628
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.p) ~ fitted.values(comp_data.p, type = 'pearson'))
plot(residuals(comp_data.p, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.p)
```

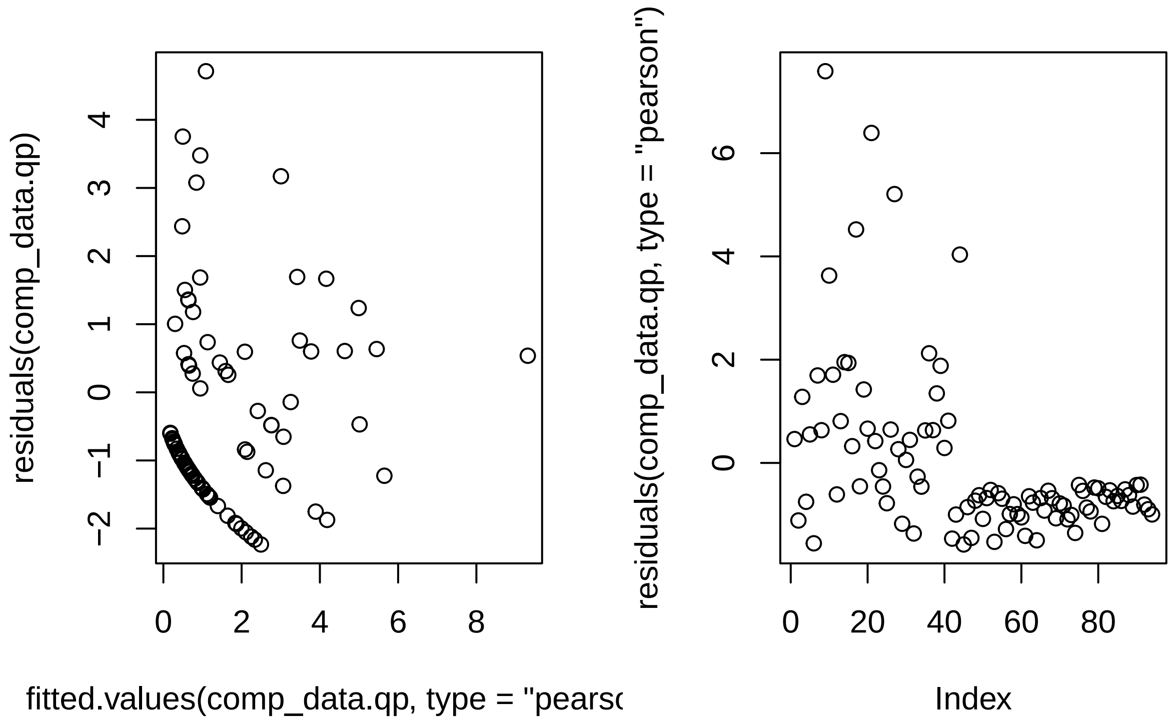


2. quasi poisson model with summary and diagnostics

```
comp_data.qp <- glm(f, data = comp_data_2, family = quasipoisson)
summary(comp_data.qp)
```

```
##
## Call:
## glm(formula = f, family = quasipoisson, data = comp_data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2330  -1.2243  -0.8338   0.4307   4.7130
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.2075577  1.4536952  -1.519 0.132493
## visits              0.0008632  0.0003037   2.842 0.005581 **
## residencyY          1.1528999  0.6514394   1.770 0.080269 .
## genderM             1.6677457  0.4490251   3.714 0.000359 ***
## revenue            -0.0006513  0.0049701  -0.131 0.896038
## hours              -0.0001127  0.0006252  -0.180 0.857352
## residencyY:genderM -2.3631371  0.7275406  -3.248 0.001651 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.989312)
##
##     Null deviance: 312.62  on 93  degrees of freedom
## Residual deviance: 195.61  on 87  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.qp) ~ fitted.values(comp_data.qp, type = 'pearson'))
plot(residuals(comp_data.qp, type = 'pearson'))
```



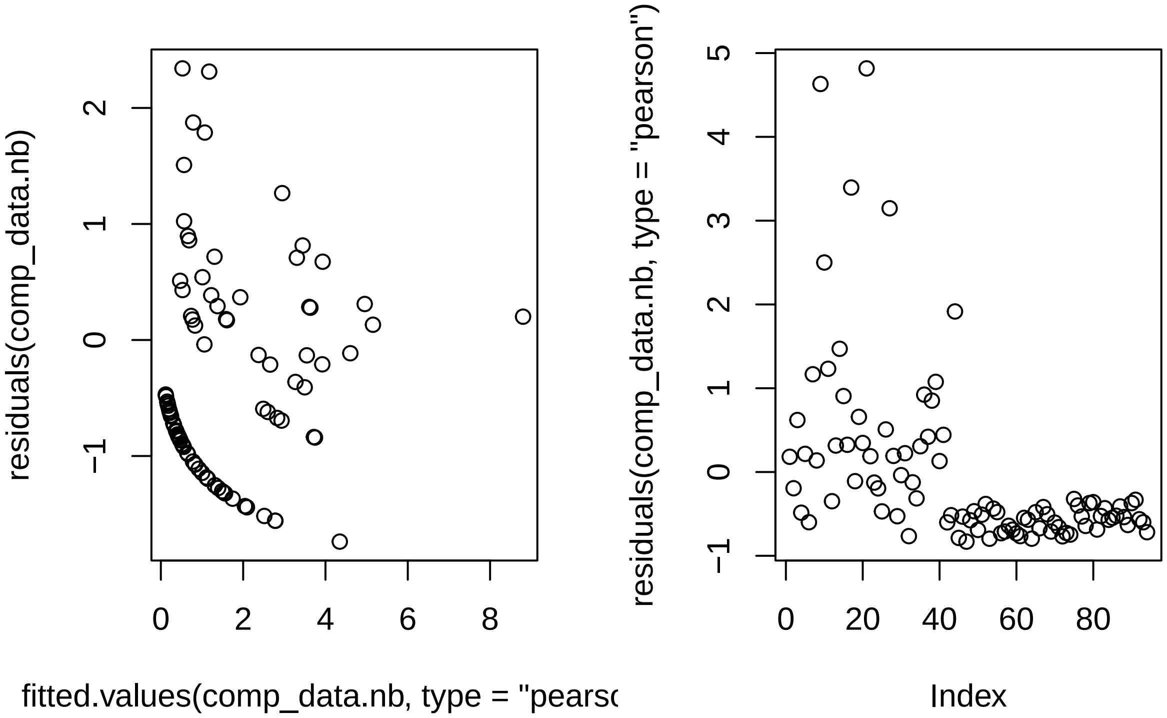
3. negative binomial with summary and diagnostics, with step AIC used

```
comp_data.nb <- glm.nb(f , data = comp_data_2)
summary(comp_data.nb)
```

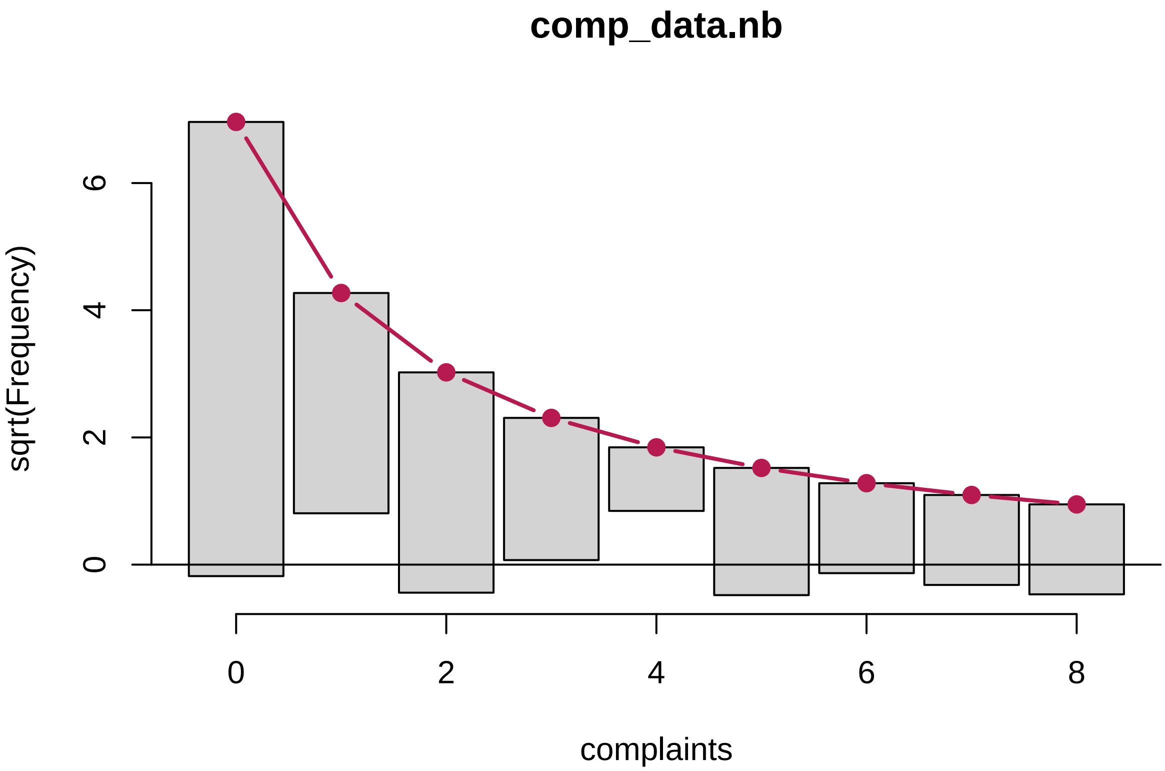
```
##
## Call:
```

```
## glm.nb(formula = f, data = comp_data_2, init.theta = 0.8200922261,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7376  -0.9598  -0.6240   0.1958   2.3408
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.5739755  1.5181086  -0.378 0.705367
## visits              0.0010838  0.0002841   3.815 0.000136 ***
## residencyY          1.2235477  0.6229425   1.964 0.049514 *
## genderM             1.5605405  0.4307888   3.623 0.000292 ***
## revenue            -0.0056565  0.0053865  -1.050 0.293663
## hours              -0.0006772  0.0005634  -1.202 0.229306
## residencyY:genderM -2.2770372  0.7207557  -3.159 0.001582 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(0.8201) family taken to be 1)
##
##     Null deviance: 125.563  on 93  degrees of freedom
## Residual deviance:  86.176  on 87  degrees of freedom
## AIC: 293.95
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.820
##           Std. Err.:  0.256
##
##  2 x log-likelihood:  -277.945
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.nb) ~ fitted.values(comp_data.nb, type = 'pearson'))
plot(residuals(comp_data.nb, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.nb)
```



```
stepAIC(comp_data.nb)
```

```
## Start:  AIC=291.95
## complaints ~ visits + residency * gender + revenue + hours
##
##                     Df    AIC
## - revenue            1 290.85
## - hours              1 290.85
## <none>                 291.94
## - residency:gender  1 299.04
## - visits             1 301.01
##
## Step:  AIC=290.85
## complaints ~ visits + residency + gender + hours + residency:gender
##
##                     Df    AIC
## - hours              1 289.51
## <none>                 290.85
## - residency:gender  1 298.96
## - visits             1 299.05
##
## Step:  AIC=289.51
## complaints ~ visits + residency + gender + residency:gender
##
##                     Df    AIC
## <none>                 289.51
## - visits             1 299.17
## - residency:gender   1 299.19
```

```
##
## Call:  glm.nb(formula = complaints ~ visits + residency + gender + residency:gender,
##     data = comp_data_2, init.theta = 0.8544720303, link = log)
##
## Coefficients:
##       (Intercept)              visits          residencyY             genderM
##        -2.4123731           0.0008078           1.3282257           1.6213312
## residencyY:genderM
##        -2.4610397
##
## Degrees of Freedom: 93 Total (i.e. Null);  89 Residual
## Null Deviance:       128.1
## Residual Deviance: 89.31     AIC: 291.5
```

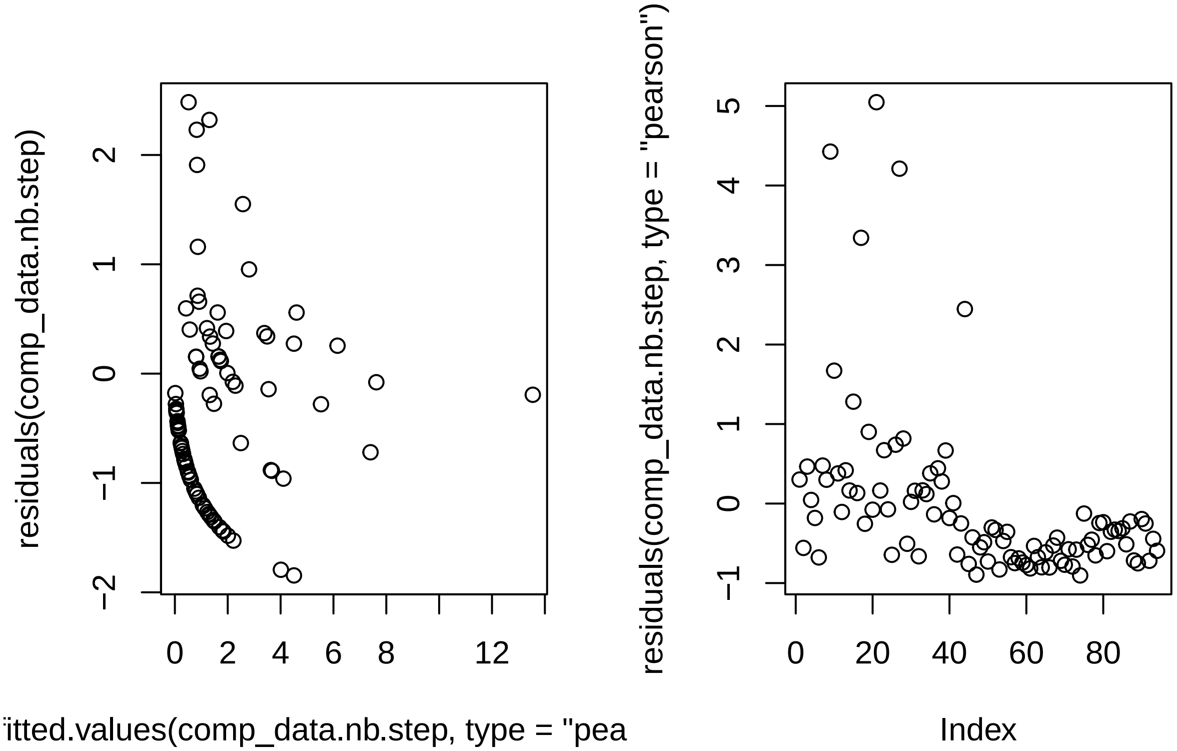
3.1 negative binomial final model after backwards stepwise selection

```
comp_data.nb.step <- glm.nb(formula = complaints ~ visits + residency + gender + revenue +
    hours + residency:gender +
    residency:hours + gender:revenue, data = comp_data_2)
summary(comp_data.nb.step)
```

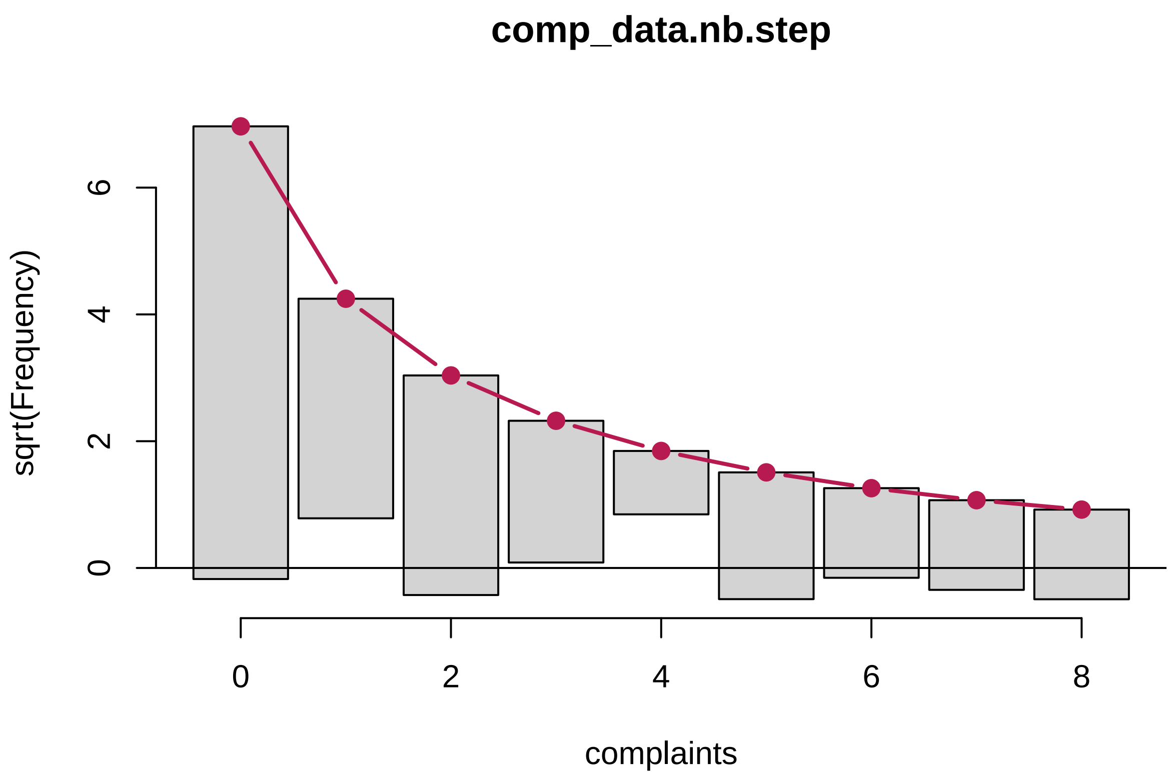
```
##
## Call:
## glm.nb(formula = complaints ~ visits + residency + gender + revenue +
##     hours + residency:gender + residency:hours + gender:revenue,
##     data = comp_data_2, init.theta = 0.9968612672, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8455  -0.9559  -0.4796   0.1545   2.4827
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.7762319  2.8508020   0.974 0.330135
## visits              0.0011079  0.0002974   3.725 0.000195 ***
## residencyY          5.2621274  1.6642173   3.162 0.001567 **
## genderM            -5.9295699  3.0453947  -1.947 0.051527 .
## revenue            -0.0290563  0.0108109  -2.688 0.007195 **
## hours               0.0007630  0.0007822   0.975 0.329374
## residencyY:genderM -2.7350420  0.7147973  -3.826 0.000130 ***
## residencyY:hours   -0.0025233  0.0010064  -2.507 0.012164 *
## genderM:revenue     0.0306058  0.0121969   2.509 0.012097 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9969) family taken to be 1)
##
```

```
##     Null deviance: 137.736  on 93  degrees of freedom
## Residual deviance:  83.606  on 85  degrees of freedom
## AIC: 288.13
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.997
##           Std. Err.:  0.328
##
##  2 x log-likelihood:  -268.128
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.nb.step) ~ fitted.values(comp_data.nb.step, type = 'pearson'))
plot(residuals(comp_data.nb.step, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.nb.step)
```



4.ZIP model with summary and diagnostics

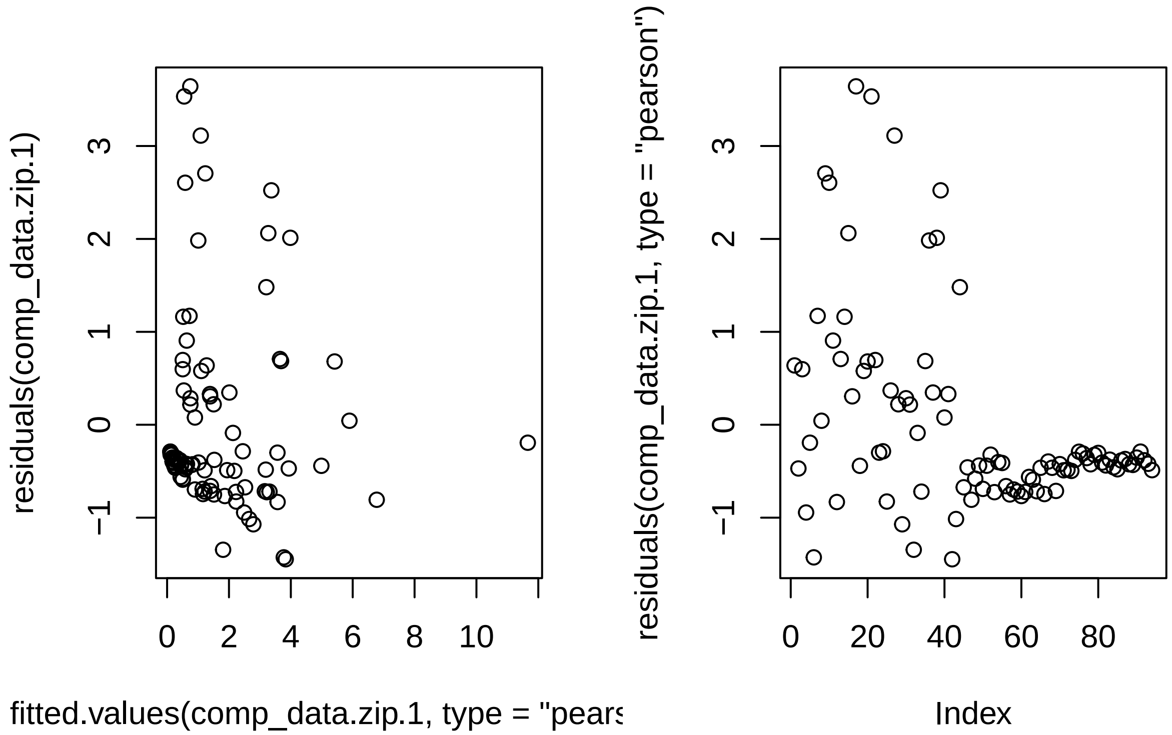
```
comp_data.zip.1 <- zeroinfl(complaints ~ visits + residency * gender + revenue + hours | visits + reside
, data = comp_data_2, dist = "poisson")
summary(comp_data.zip.1)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ visits + residency * gender + revenue +
##     hours | visits + residency * gender + revenue + hours, data = comp_data_2,
##     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 ***
## residencyY            -5.074e-01  4.370e-01  -1.161  0.24565
## genderM               -7.107e-02  3.219e-01  -0.221  0.82524
## revenue               -7.918e-03  2.860e-03  -2.769  0.00563 **
## hours              -9.846e-04  5.355e-04  -1.839  0.06597 .
## 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
## residencyY          -1.370e+01  2.018e+02  -0.068    0.946
## genderM             -1.536e+01  2.881e+02  -0.053    0.957
## revenue             -1.255e-03  1.212e-02  -0.104    0.918
## hours                2.810e-04  1.840e-03   0.153    0.879
## 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
```

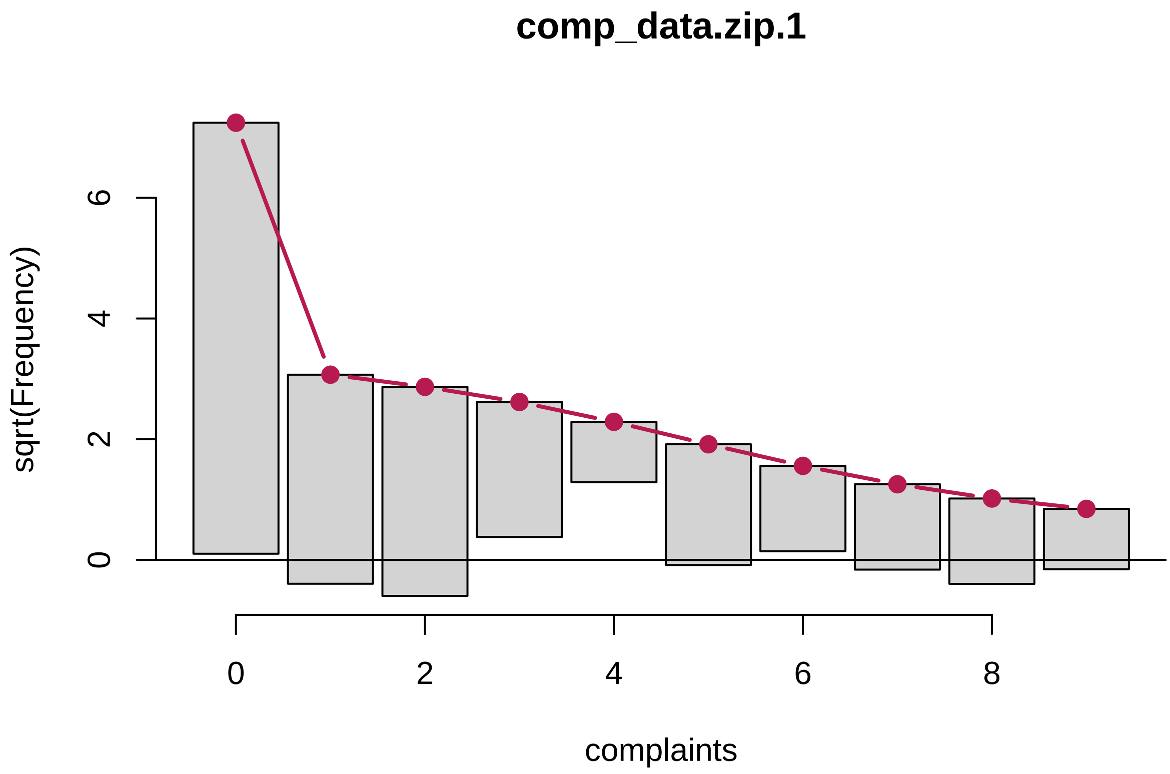
```
AIC(comp_data.zip.1)
```

```
## [1] 266.5369
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.zip.1) ~ fitted.values(comp_data.zip.1, type = 'pearson'))
plot(residuals(comp_data.zip.1, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.zip.1)
```



5. ZINB model with summary and diagnostics

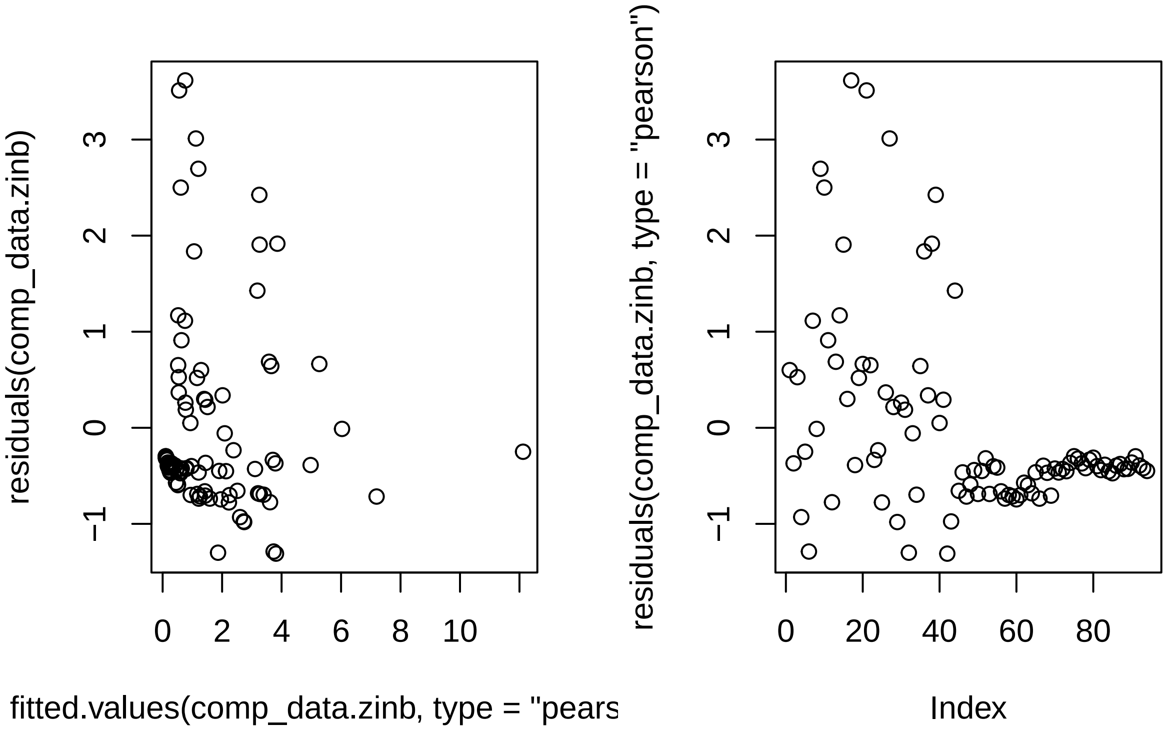
```
comp_data.zinb <- zeroinfl(complaints ~ visits + residency * gender + revenue + hours | visits + residen
summary(comp_data.zinb)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ visits + residency * gender + revenue +
##     hours | visits + residency * gender + revenue + hours, data = comp_data_2,
##     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  2.9460530   0.490   0.6243
## visits               0.0015351  0.0003813   4.026 5.67e-05 ***
## residencyY         -0.4587666  0.4808625  -0.954   0.3401
## genderM            -0.0527055  0.3312755  -0.159   0.8736
## revenue            -0.0087563  0.0047726  -1.835   0.0665 .
## hours              -0.0011771  0.0018345  -0.642   0.5211
## 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
## residencyY         -1.370e+01  2.069e+02  -0.066    0.947
## genderM            -1.536e+01  2.915e+02  -0.053    0.958
## revenue            -2.171e-03  3.342e-02  -0.065    0.948
## hours               4.494e-05  1.042e-02   0.004    0.997
## 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
```

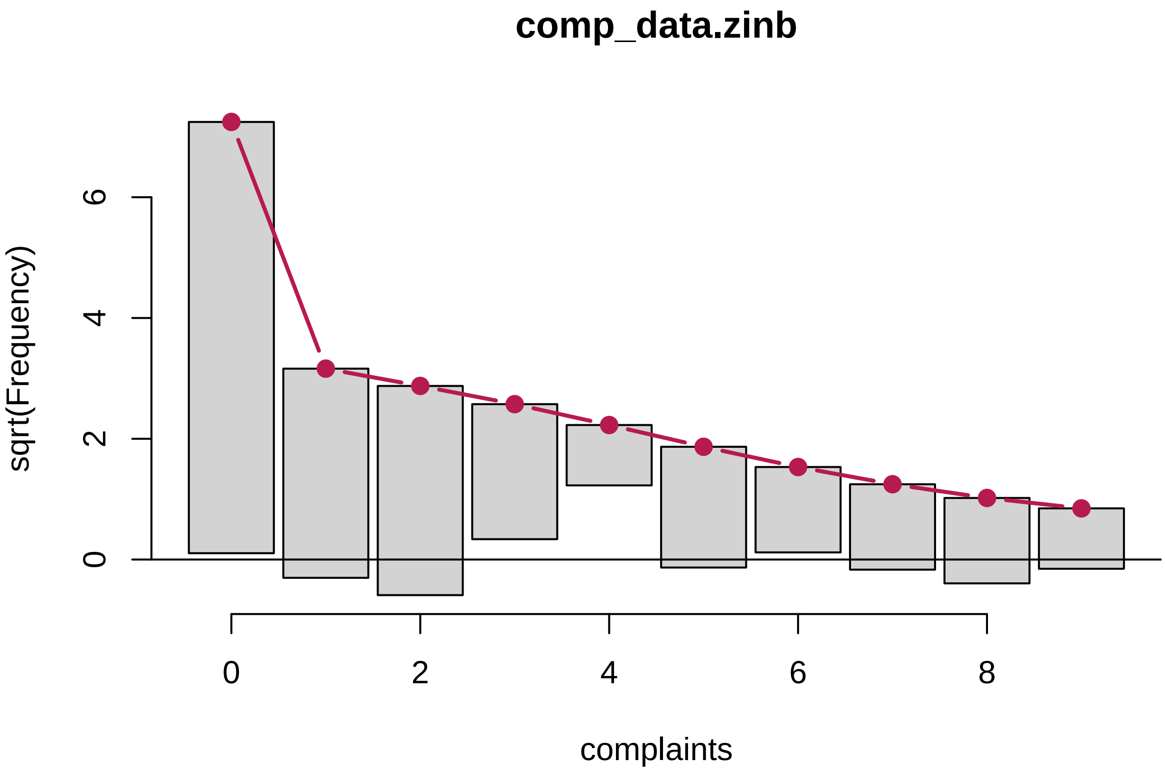
```
AIC(comp_data.zinb)
```

```
## [1] 268.2066
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.zinb) ~ fitted.values(comp_data.zinb, type = 'pearson'))
plot(residuals(comp_data.zinb, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.zinb)
```



5.1 likelihood ratio test to compare ZINB and ZIP, prove that ZINB is not improving the model significantly

```
lrtest(comp_data.zip.1, comp_data.zinb)
```

```
## Likelihood ratio test
##
## Model 1: complaints ~ visits + residency * gender + revenue + hours |
##     visits + residency * gender + revenue + hours
## Model 2: complaints ~ visits + residency * gender + revenue + hours |
##     visits + residency * gender + revenue + hours
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  14 -119.27
## 2  15 -119.10  1 0.3302     0.5655
```

```
AIC(comp_data.zip.1, comp_data.zinb)
```

```
##                df      AIC
## comp_data.zip.1 14 266.5369
## comp_data.zinb  15 268.2066
```

6.1 Dropping interaction terms between residency x gender from count part of ZIP, showing the compelling evidence of the interaction term significance

```
comp_data.zip.1.drop <- zeroinfl(complaints ~ visits + residency * gender + revenue + hours | visits + r
, data = comp_data_2, dist = "poisson")
summary(comp_data.zip.1)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ visits + residency * gender + revenue +
##     hours | visits + residency * gender + revenue + hours, data = comp_data_2,
##     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 ***
## residencyY         -5.074e-01  4.370e-01  -1.161  0.24565
## genderM            -7.107e-02  3.219e-01  -0.221  0.82524
## revenue            -7.918e-03  2.860e-03  -2.769  0.00563 **
## hours              -9.846e-04  5.355e-04  -1.839  0.06597 .
## 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
## residencyY         -1.370e+01  2.018e+02  -0.068    0.946
## genderM            -1.536e+01  2.881e+02  -0.053    0.957
## revenue            -1.255e-03  1.212e-02  -0.104    0.918
## hours               2.810e-04  1.840e-03   0.153    0.879
## 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
```

```
lrtest(comp_data.zip.1, comp_data.zip.1.drop)
```

```
## Likelihood ratio test
##
## Model 1: complaints ~ visits + residency * gender + revenue + hours |
##     visits + residency * gender + revenue + hours
## Model 2: complaints ~ visits + residency * gender + revenue + hours |
##     visits + residency + gender + revenue + hours
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  14 -119.27
## 2  13 -135.11 -1 31.675  1.823e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.2 Final model after using likelihood ratio test to drop each non-significant term

```
comp_data.zip.2 <- zeroinfl(complaints ~ visits + revenue + hours | residency * gender, data = comp_data
summary(comp_data.zip.2)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ visits + revenue + hours | residency *
##     gender, data = comp_data_2, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3927 -0.5535 -0.3901  0.1891  4.0408
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.181e+00  1.231e+00   0.959 0.337668
## visits       1.433e-03  1.287e-05 111.367  < 2e-16 ***
## revenue     -9.228e-03  2.645e-03  -3.489 0.000485 ***
## hours       -8.370e-04  4.167e-04  -2.009 0.044589 *
##
## Zero-inflation model coefficients (binomial with logit link):
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.2650     0.5162   2.451   0.0143 *
## residencyY         -14.2266   246.6172  -0.058   0.9540
## genderM            -14.5160   198.5522  -0.073   0.9417
## residencyY:genderM  27.6935   316.6113   0.087   0.9303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 42
## Log-likelihood: -120.9 on 8 Df
```

```
AIC(comp_data.zip.2)
```

```
## [1] 257.8263
```

7. Another alternative model for ZIP, place variable in the opposite part of the model

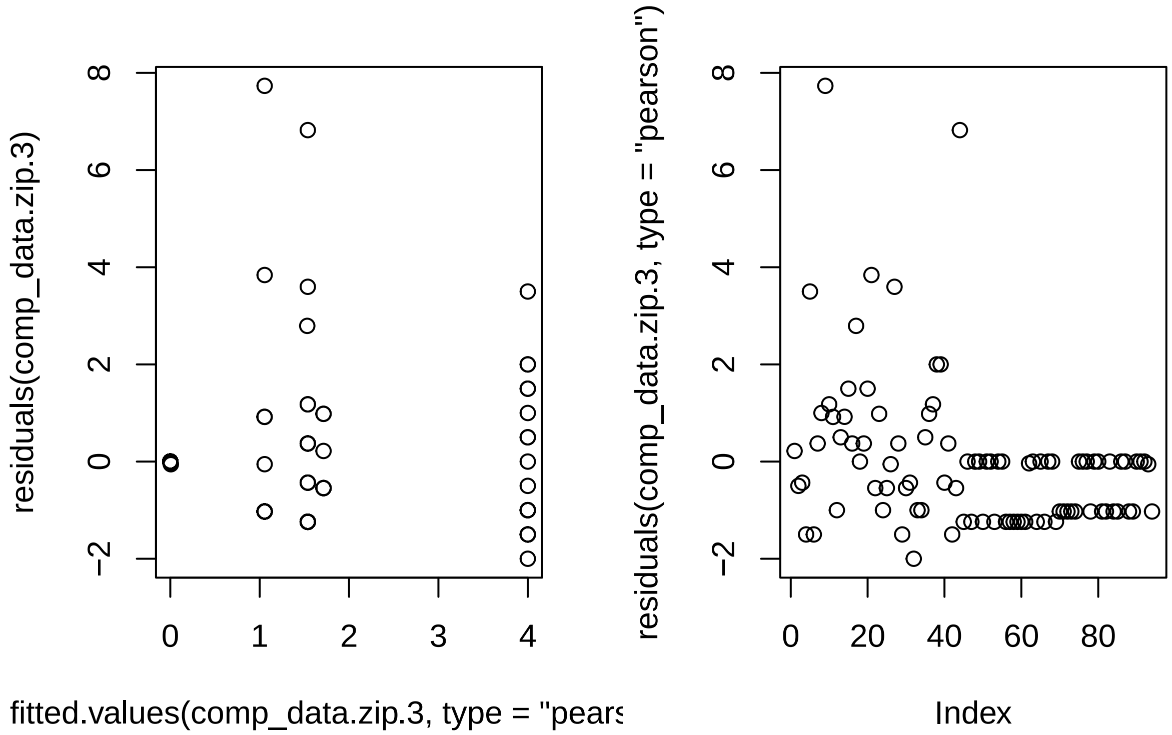
```
comp_data.zip.3 <- zeroinfl(complaints ~ residency * gender | visits   + revenue + hours , data = comp_
summary(comp_data.zip.3)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ residency * gender | visits + revenue +
##     hours, data = comp_data_2, dist = "poisson")
##
## Pearson residuals:
##        Min         1Q     Median         3Q        Max
## -1.9999887 -1.0273637 -0.0008131  0.3336592  7.7329224
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.05399    0.22943   0.235 0.813954
## residencyY         0.48497    0.36875   1.315 0.188452
## genderM            1.33229    0.25651   5.194 2.06e-07 ***
## residencyY:genderM -1.44048    0.41729  -3.452 0.000557 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.355e+02  1.888e+02  -1.777  0.07558 .
## visits      -1.293e-01  7.705e-03 -16.776  < 2e-16 ***
## revenue      1.133e+00  3.862e-01   2.934  0.00335 **
## hours        2.062e-01  6.774e-02   3.044  0.00233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 473
## Log-likelihood: -153.4 on 8 Df
```

```
AIC(comp_data.zip.3)
```

```
## [1] 322.715
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.zip.3) ~ fitted.values(comp_data.zip.3, type = 'pearson'))
plot(residuals(comp_data.zip.3, type = 'pearson'))
```



8. Final model diagnostics

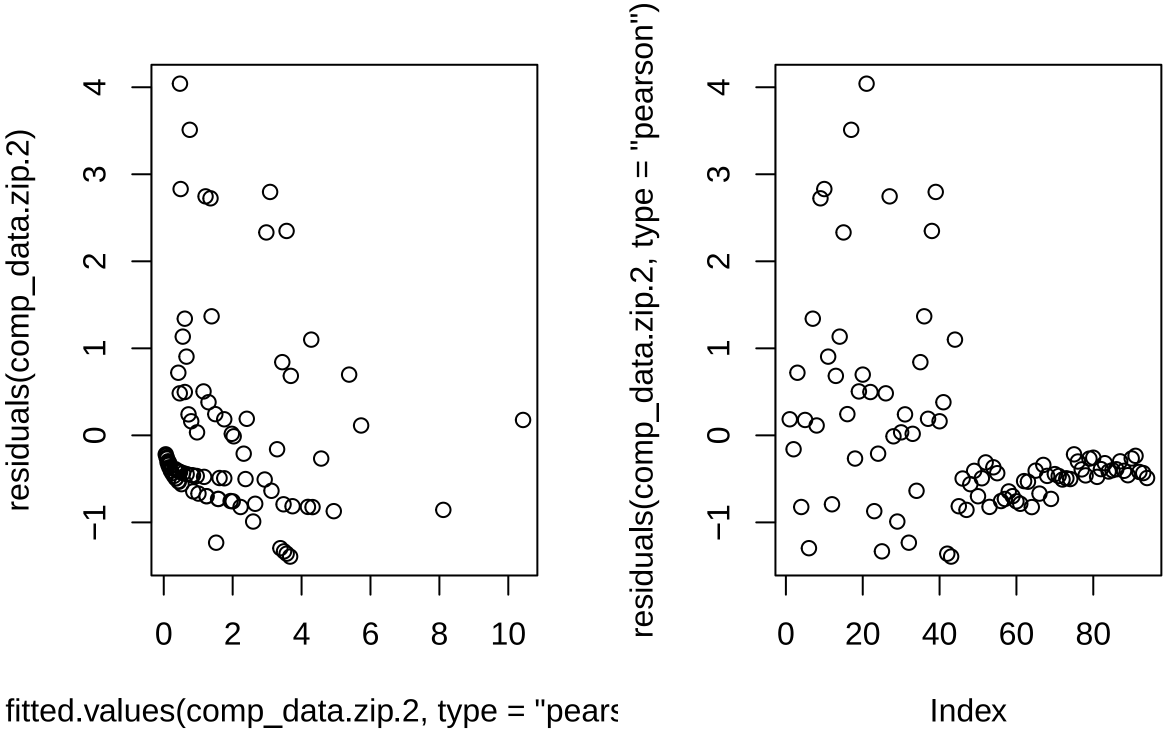
```
summary(comp_data.zip.2)
```

```
##
## Call:
## zeroinfl(formula = complaints ~ visits + revenue + hours | residency *
##     gender, data = comp_data_2, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3927 -0.5535 -0.3901  0.1891  4.0408
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.181e+00  1.231e+00   0.959 0.337668
## visits       1.433e-03  1.287e-05 111.367  < 2e-16 ***
## revenue     -9.228e-03  2.645e-03  -3.489 0.000485 ***
## hours       -8.370e-04  4.167e-04  -2.009 0.044589 *
##
## Zero-inflation model coefficients (binomial with logit link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.2650     0.5162   2.451   0.0143 *
## residencyY          -14.2266   246.6172  -0.058   0.9540
## genderM             -14.5160   198.5522  -0.073   0.9417
## residencyY:genderM  27.6935   316.6113   0.087   0.9303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 42
## Log-likelihood: -120.9 on 8 Df
```

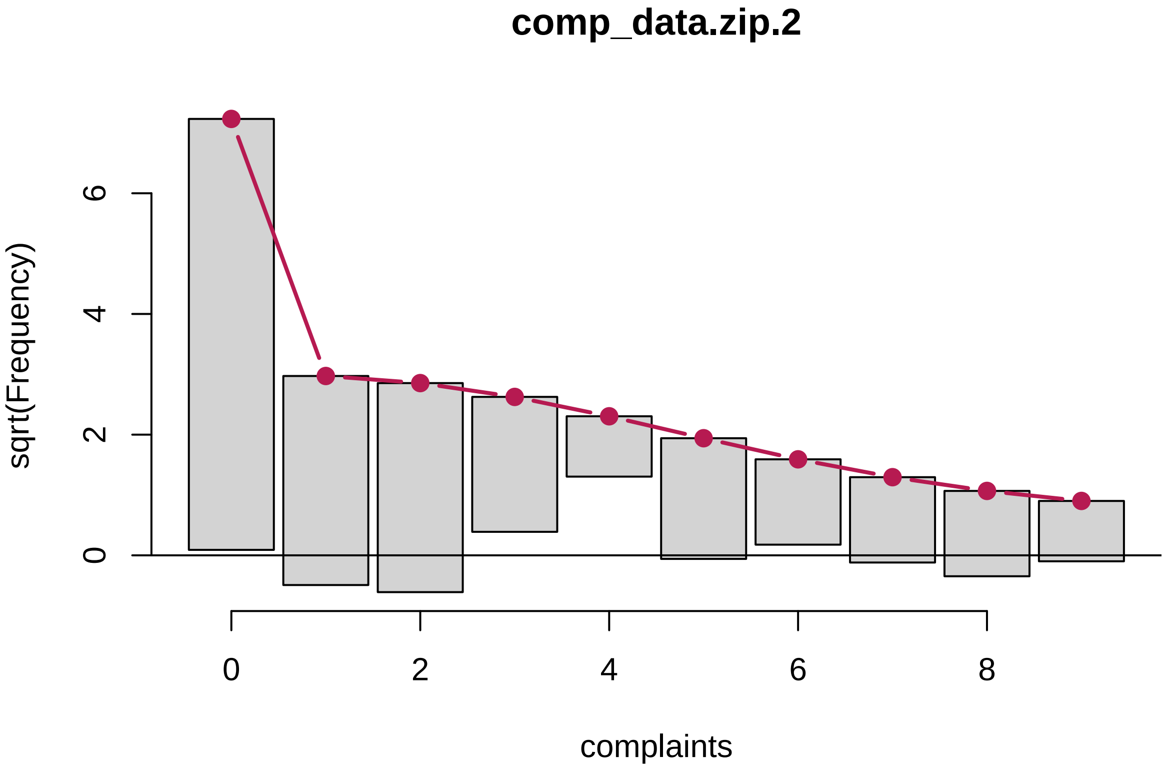
```
AIC(comp_data.zip.2)
```

```
## [1] 257.8263
```

```
par(mfrow=c(1,2))
plot(residuals(comp_data.zip.2) ~ fitted.values(comp_data.zip.2, type = 'pearson'))
plot(residuals(comp_data.zip.2, type = 'pearson'))
```



```
par(mfrow=c(1,1))
rootogram(comp_data.zip.2)
```


comp_data.zip.2
sqrt(Frequency)
0
2
4
6
0
2
4
6
8
complaints