STATS 504 Assignment 1
Derogatory Credit Reports

# 1. Introduction

Derogatory credit reports are reports with negative credit information which are generally used to deny loans. Often these derogatory credit reports are a result of missed payments, bankruptcy, repossessions, and foreclosures [1] which are influenced by external factors such as income and spending habits. Our client is interested in these derogatory reports and the factors that could contribute to the derogatory reports. However, she believes that there are some discrepancies between her data and the distribution of her model. Thus, this analysis aims to find a different model that is able to better represent the data as well as perform inference to analyze which predictors could be important and contribute to the derogatory reports. The models show that average monthly credit card expenditure, homeowner, and number of active credit accounts had the closest association with the number of derogatory reports.

# 2. Methods

## 2.1 Model Selection

Before the model fitting procedure, we first performed an exploratory study of the data set in order to obtain more information for our models. First looking at the response variable, we see the number of derogatory reports is count data and only takes on values greater than or equal to zero. Thus, we primarily focus on models in which the response variable is count data. This limits the family of models we examine to two main categories, the Poisson regression and the negative binomial regression, a less restrictive model that adds variation among observed counts and is also used to examine count data.

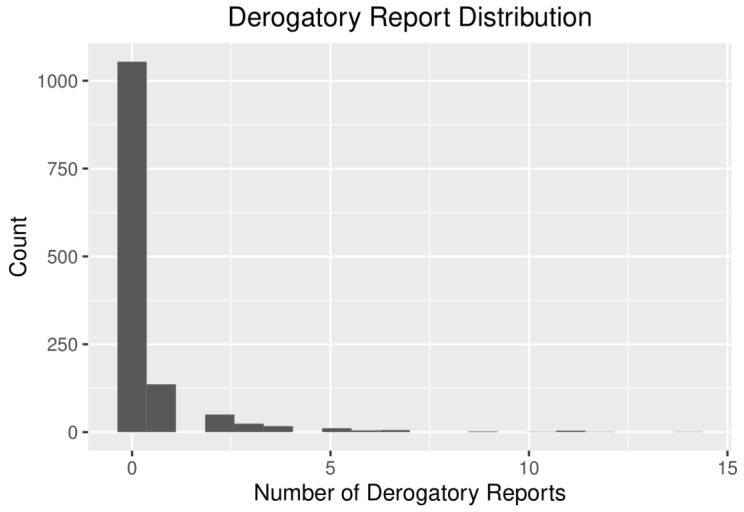



Figure 1: Histograms of the number of derogatory reports in the given data set.

From Figure 1 and our client's concerns, we know that the number of zeros in the report variable greatly outnumbers any other number of derogatory reports. Both the standard Poisson regression and the negative binomial regression are able to handle count data with zeros as long as the number of zeros is consistent with the respective distribution [2]. However, due to the large amount of zeros in the outcome, we also consider the zero-inflated Poisson regression model. The zero-inflated Poisson regression model is a model that handles large amounts of zeros by mixing two separate processes, one in which the response is zero and the other obtained from a Poisson distribution [2]. We also examine the Poisson regression allowing for overdispersion, a model which allows for greater variability in the data.

### 2.1 Variable Selection

Ultimately, the goal of the analysis is to understand which factors are associated with the outcome, the number of derogatory reports. However, variables that the outcome will affect should not be included in the model that is trying to predict this outcome. The variable card, representing whether or not the credit card application was accepted, is determined by all other variables in the data set. This means that the number of derogatory reports had an immediate impact on the variable card and thus, the variable card should not be included in the model.

Also, as provided, the variable share (ratio of monthly credit card expenditure to yearly income) is generated from the variables income and expenditure. Examining the correlation between the numerical variables in the data set, we found that the correlation between share and expenditure is around 0.84, which represents a very strong relation between the two variables.

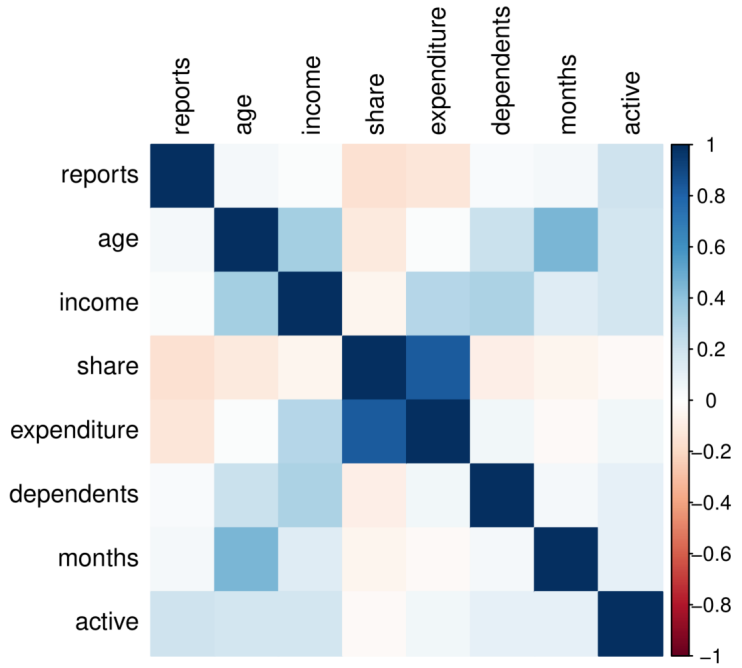



Figure 2: Correlation plot of the numerical variables in which values closer to 1 and -1 represent high correlation.

From Figure 2, it is evident that the correlation between the two variables share and expenditure is much higher than the correlation between other variables. This information along with the knowledge that the predictor share is generated from income and expenditure is evidence of multicollinearity, in which a predictor considered for the model can be predicted from the other variables considered with high accuracy [3]. Thus, one of the variables must be excluded from our model. In this case, we choose to exclude the variable share, because of its relation to both income and expenditure as well as the high correlation with expenditure. All other variables present in the data set do not seem to be influenced by the response variable or have high correlation with each other. In order to understand which factors are closely associated with the outcome, we include all the other explanatory variables available from the data set in the regression model. This is acceptable, because the number of variables in our data set is much less than the number of observations.

After performing variable selection, we compare the different models using a likelihood-based metric, AIC, to determine which model achieves the most optimal fit. AIC is a model fit metric that balances the model fit and number of parameters in the model to prevent overfitting, fitting any particular data set too closely. In order to ensure the model is interpretable, we do not perform any further transformations on the predictor or the response variables.

## 3. Results

The data set used for this analysis includes 1319 applications and 12 variables: card (credit card acceptance), number of derogatory reports, age, yearly income, share (ratio of monthly credit card expenditure to yearly income), average monthly credit card expenditure, home-owner, self-employed, number of dependents, months living at current address, other major credit cards, and number of active credit accounts. There are no missing values for any of the variables. After examining the data, we can see that there are seven applications in which the age was less than one year old. However, one must be at least 18 years of age to apply for a credit card [4]. It is not possible for an infant to apply for a credit card, thus, there must be some issues with these seven observations in the data set. Due to this, we decided to remove these seven observations from our later analysis. We choose not to re-impute the values, because it is difficult to evaluate the age from the data set and could lead to inaccurate results. Also, there are only seven observations that should be excluded, which is a small proportion in relation to the number of observations in the entire data set. After the removal of these entries, all entries had ages greater than 18 years old. Also, as mentioned in the Methods section, the variables card and share are dropped from the model. All other variables are included.

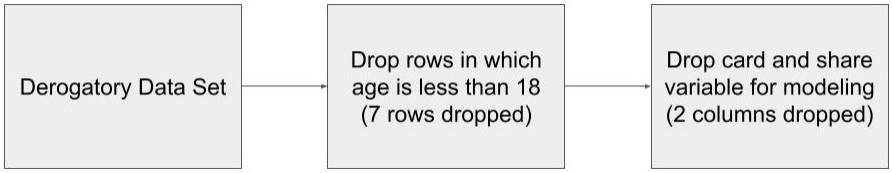



Figure 3: Data flow diagram of removed rows and columns in our data set before model fitting.

Table 1 displays the distribution metrics for the numeric variables in the data set after the removal of the seven rows in which the age was less than 18 years old as well as the two columns, card and share. The other three variables, homeowner, self-employed, and other major credit cards are all categorical variables taking on values of 'yes' and 'no'. The variable homeowner had 579 yes and 533 no values, self-employed had 91 yes and 1221 no values, and other major credit cards had 1073 yes and 239 no values. After evaluating the distribution of our variables, it seems that the values presented in Table 1 are quite reasonable and we thus proceed to model fitting.

| **Variable** | **Mean** | **Standard Deviation** |
|---|---|---|
| Number of Derogatory Reports | 0.458 | 1.348 |
| Age (years) | 33.390 | 9.884 |
| Yearly Income (in 10,000 USD) | 3.367 | 1.697 |
| Average Monthly Credit Card Expenditure | 184.970 | 272.715 |
| Number of Dependents | 0.994 | 1.247 |
| Months Living at Current Address | 55.180 | 66.254 |
| Number of Active Credit Accounts | 6.999 | 6.315 |

Table 1: Numeric variable distribution metrics providing both the variable mean and standard deviation, a measure of spread.

We examine several models, determining the most fitting model for our analysis. We choose to report the negative binomial regression model since this model provides the lowest AIC value, indicating a better fitting model than the standard Poisson regression model as well as the zero-inflated Poisson regression model, despite the larger amount of zeros in the data set. Another reason for choosing this model is that it is simple to interpret the results. The negative binomial regression model also allows for overdispersion, providing more leniency between the conditional variance and the conditional mean. Taking the overdispersion into account was vital in producing more accurate standard errors and significance values for the model at hand.

From the negative binomial regression model and Table 2, it can be seen that the variables average monthly credit card expenditure, homeowner, months living at current address, and number of active credit accounts are statistically significant (at the 0.05 p-value level) in determining the number of derogatory reports. However, the months variable appears to be less significant in comparison to the other three significant variables and is not statistically significant when fitting some of the Poisson models.

| Variable | Rate (95% CI) | p-value |
|---|---|---|
| Age (years) | 1.005 (0.986, 1.023) | 0.614 |
| Yearly Income (in 10,000 USD) | 1.086 (0.985, 1.202) | 0.096 |
| Average Monthly Credit Card Expenditure | 0.998 (0.997, 0.998) | <0.001 |
| Homeowner | 0.441 (0.312, 0.621) | <0.001 |
| Self-employed | 1.032 (0.592, 1.838) | 0.911 |
| Number of Dependents | 1.097 (0.967, 1.248) | 0.144 |
| Months Living at Current Address | 1.002 (1.000, 1.005) | 0.040 |
| Other Major Credit Cards | 1.018 (0.695, 1.481) | 0.930 |
| Number of Active Credit Accounts | 1.129 (1.098, 1.162) | <0.001 |

Table 2: Rate ratios and 95% confidence intervals of the number of derogatory reports for the variable assessed along with their p-values.

Based on the rates presented in Table 2, one can provide interpretations for each of the variables. In this case, we will focus on the variables that are statistically significant in the chosen model. Other variables must be held constant for each of these interpretations to hold. For a one unit increase in average monthly expenditure, the number of derogatory reports is expected to decrease by a factor of 0.998. Similarly, for a month increase in the number of months living at the current address, the number of derogatory reports is expected to increase by a factor of 1.002. Furthermore, for each additional active credit account, the number of derogatory reports is expected to increase by a factor of 1.129. The last statistically significant variable is a categorical variable and thus will have a bit of a different interpretation. The number of derogatory reports is expected to lower by a factor of 0.441 for homeowners in comparison to those who do not own their home.

## 4. Conclusion

The purpose of this analysis was to identify factors that may contribute to derogatory reports on a credit card applicant's credit history. From our analysis, the average monthly expenditure, whether the applicant is a homeowner, and the number of active credit accounts appear to have significant association with the number of derogatory reports and may even contribute to that number. The months an applicant has been living at their current address also seems to have some association, but this association is not as significant as the three factors previously mentioned. Contrarily, the variables such as age, self-employed, and other major credit cards are probably not influential to the number of derogatory reports. In the future, we hope to analyze other variables such as occupation, years in the workforce, and marital status to see if there are other factors with close relation to the number of derogatory reports. However, with this being said, the significant variables determined from this analysis are most definitely associated with the number of derogatory reports.

# References


[1] Lucas Downey. Derogatory Information, December 16, 2020. URL https://www.investopedia.com/terms/d/derogatory-information.asp.

[2] John Fox. Applied Regression Analysis and Generalized Linear Models, April 2015. URL https://www.sagepub.com/sites/default/files/upm-binaries/21121_Chapter_15.pdf.

[3] Adam Hayes. Multicollinearity, February 18, 2021. URL https://www.investopedia.com/terms/m/multicollinearity.asp.

[4] Alexandria White. How old do you have to be to get a credit card?, June 3, 2021. URL https://www.cnbc.com/select/how-old-do-you-have-to-be-to-get-a-credit-card/.

# Appendix

## Load Libraries

```
library(ggplot2)
library(tidyverse)
library(MASS)
library(corrplot)
library(GGally)
library(reshape2)
library(pscl)
library(regclass)
```

## Read Data

We change the character columns to factor to illustrate the two levels: yes and no more clearly.

```
derogatory_bf <- read.csv(file = "derogatory.csv")
derogatory_bf$card <- as.factor(derogatory_bf$card)
derogatory_bf$owner <- as.factor(derogatory_bf$owner)
derogatory_bf$selfemp <- as.factor(derogatory_bf$selfemp)
derogatory_bf$majorcards <- as.factor(derogatory_bf$majorcards)
```

## Data Preprocessing and Exploratory Data Analysis

A brief overview of what our data set looks like.

```
head(derogatory_bf)
```

```
##   card reports      age income       share expenditure owner selfemp dependents
## 1  yes       0 37.66667 4.5200 0.033269910  124.983300   yes      no          3
## 2  yes       0 33.25000 2.4200 0.005216942    9.854167    no      no          3
## 3  yes       0 33.66667 4.5000 0.004155556   15.000000   yes      no          4
## 4  yes       0 30.50000 2.5400 0.065213780  137.869200    no      no          0
## 5  yes       0 32.16667 9.7867 0.067050590  546.503300   yes      no          2
## 6  yes       0 23.25000 2.5000 0.044438400   91.996670    no      no          0
##   months majorcards active
## 1     54        yes     12
## 2     34        yes     13
## 3     58        yes      5
## 4     25        yes      7
## 5     64        yes      5
## 6     54        yes      1
```

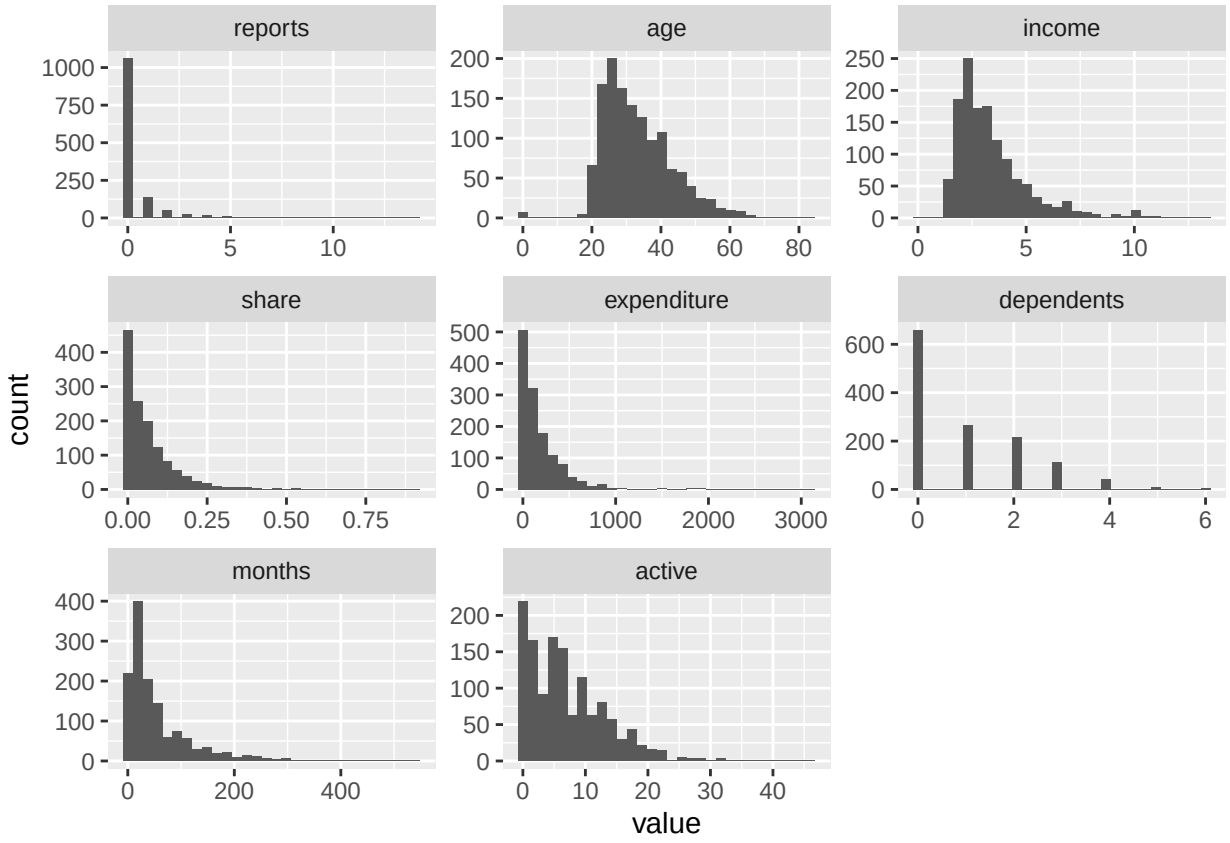
Data set dimensions:

```
dim(derogatory_bf)
```

```
## [1] 1319   12
```

Histograms of the distribution of numeric variables. From here we can see that there are a few age values that are very close to zero.

```
ggplot(melt(derogatory_bf), aes(x=value)) + geom_histogram() + facet_wrap(~variable, scales = "free")
```




After a quick examination of the data set and the histograms, we can see that the age variable has seven values less than 1. One must be at least 18 to apply for a credit card, thus, these rows will be dropped from our data set. All other analysis will be performed on the data set with these rows dropped.

```
derogatory <- derogatory_bf[which(derogatory_bf$age > 18),]
dim(derogatory)
```

```
## [1] 1312   12
```

Summary of the data and the general distribution for the variables. This report shows that our data set does not include any missing values.

```
summary(derogatory)
```

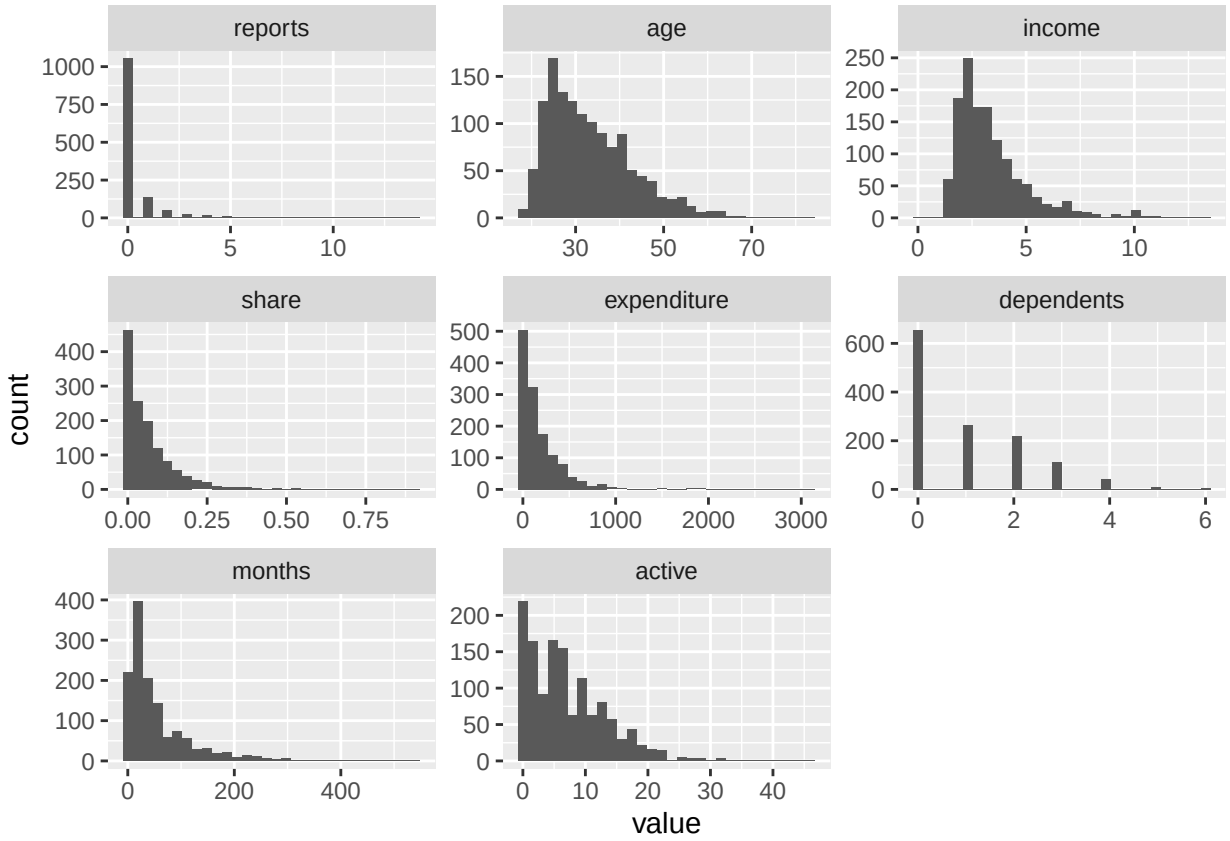
```
##   card         reports            age            income
##  no : 295   Min.   : 0.0000   Min.   :18.17   Min.   : 0.210
##  yes:1017   1st Qu.: 0.0000   1st Qu.:25.42   1st Qu.: 2.237
##             Median : 0.0000   Median :31.29   Median : 2.900
##             Mean   : 0.4581   Mean   :33.39   Mean   : 3.367
##             3rd Qu.: 0.0000   3rd Qu.:39.42   3rd Qu.: 4.000
##             Max.   :14.0000   Max.   :83.50   Max.   :13.500
##      share            expenditure       owner     selfemp      dependents
##  Min.   :0.0001091   Min.   :   0.000   no :733   no :1221   Min.   :0.0000
##  1st Qu.:0.0022080   1st Qu.:   4.583   yes:579   yes:  91   1st Qu.:0.0000
##  Median :0.0387754   Median : 101.232                       Median :1.0000
##  Mean   :0.0686361   Mean   : 184.970                       Mean   :0.9939
##  3rd Qu.:0.0935162   3rd Qu.: 248.971                       3rd Qu.:2.0000
##  Max.   :0.9063205   Max.   :3099.505                       Max.   :6.0000
##      months       majorcards     active
##  Min.   :  0.00   no : 239   Min.   : 0.000
##  1st Qu.: 12.00   yes:1073   1st Qu.: 2.000
##  Median : 30.00              Median : 6.000
##  Mean   : 55.18              Mean   : 6.999
##  3rd Qu.: 72.00              3rd Qu.:11.000
##  Max.   :540.00              Max.   :46.000
```

```
derogatory_numeric <- subset(derogatory, select = -c(card, owner, selfemp, majorcards))
sapply(derogatory_numeric, sd)
```

```
##      reports          age       income        share  expenditure   dependents
##   1.34841613   9.88420030   1.69737717   0.09478713 272.71474230   1.24740806
##       months       active
##  66.25413993   6.31471827
```

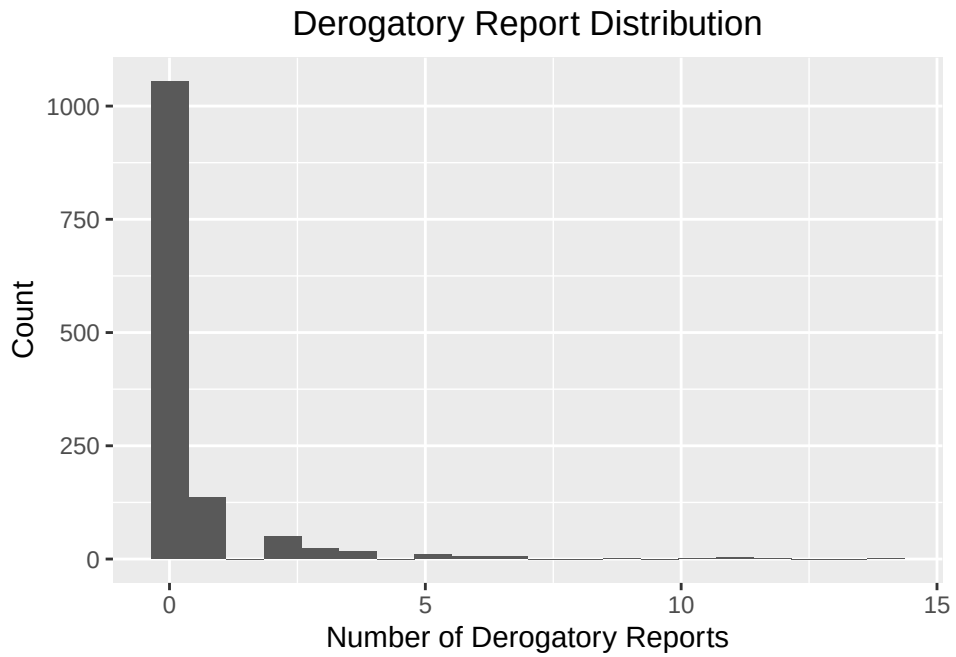
Histograms of the distribution of numeric variables after the 7 rows are removed.

```
ggplot(melt(derogatory), aes(x=value)) + geom_histogram() + facet_wrap(~variable, scales = "free")
```




Histogram of the response variable `reports`. Singling out this histogram to add to the report later on.

```
ggplot(derogatory, aes(x=reports)) + geom_histogram(bins=20) +
  ggtitle("Derogatory Report Distribution") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Number of Derogatory Reports") +
  ylab("Count")
```

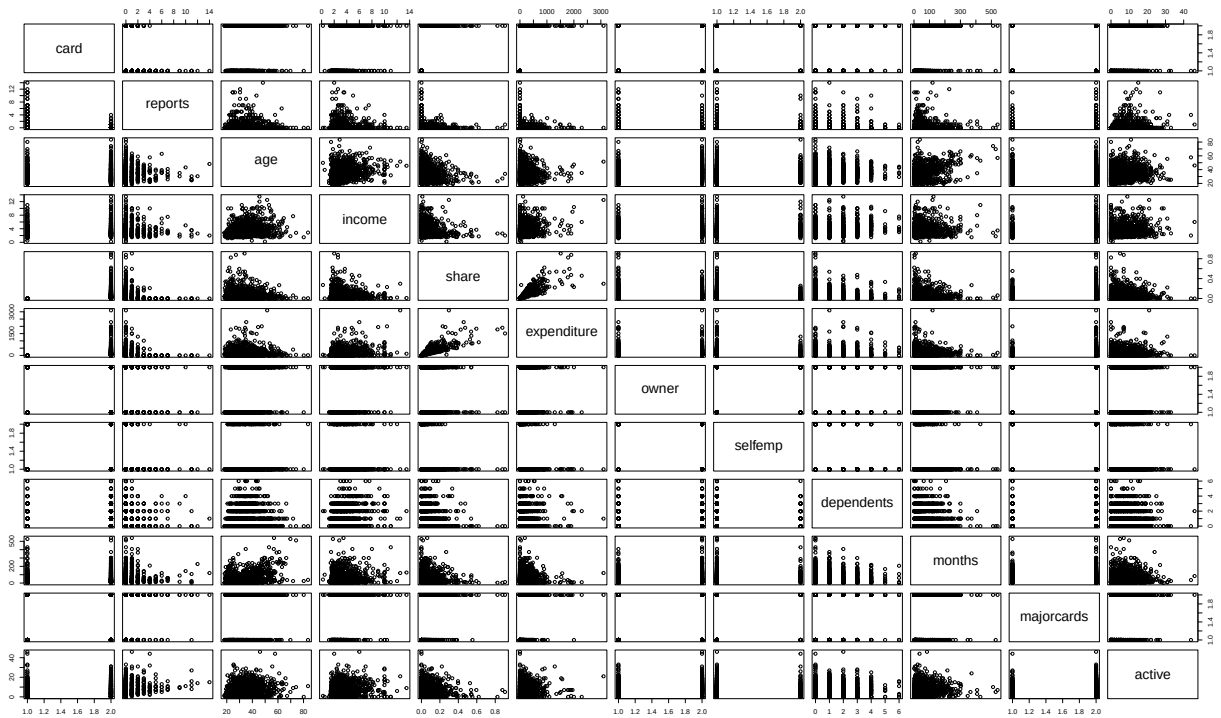



Scatter plots, density diagrams, and correlation values between different variables. This plot shows that expenditure and share seem to have a positive correlation. Looking at the row labeled reports, one can see the relation between reports and the other variables. The upper triangle of the visualization are the scatter plots, the diagonal are density distributions and the lower part consists of the correlations.

```
ggpairs(derogatory, axisLabels = 'none',
        upper = list(continuous = 'points', combo = 'dot'),
        lower = list(continuous = 'cor', combo = 'dot'),
        diag = list(continuous = 'densityDiag')) +
  theme_bw()
```

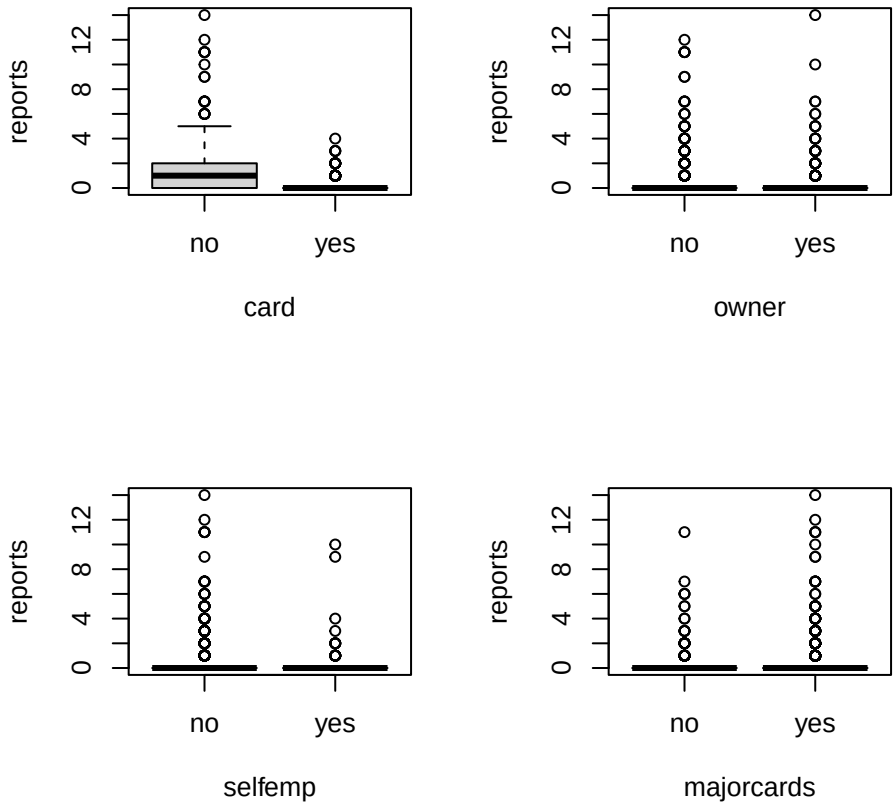

Also included the pairs plot for better visualization of the scatter plots.

```
pairs(derogatory)
```




Boxplots for each of the factor variables in relation to the response variable `reports`.

```
par(mfrow = c(2, 2))
names <- c('card', 'owner', 'selfemp', 'majorcards')
for (name in names) {
  df <- derogatory[, c('reports', name)]
  boxplot(df[,'reports']~df[,name], df, xlab = name, ylab = "reports")
}
```



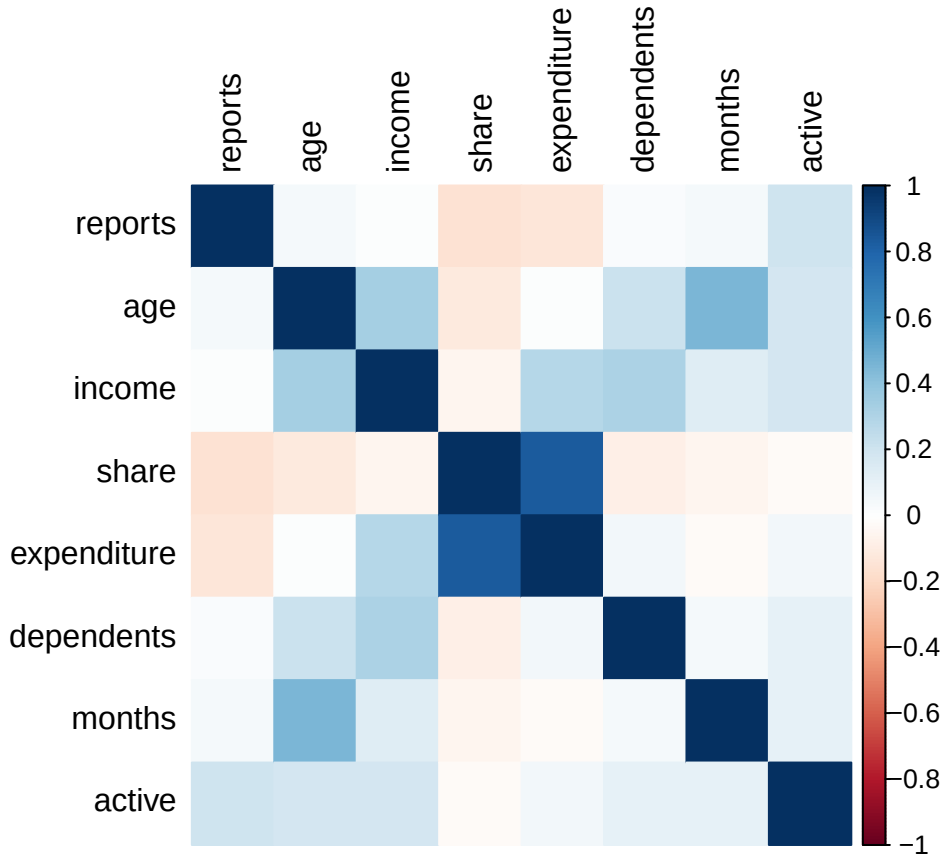

Correlation values and plot for the numerical variables.

```
df <- subset(derogatory, select = -c(card, owner, selfemp, majorcards))
corr <- cor(df)
corr
```

```
##                 reports         age      income       share expenditure
## reports      1.00000000  0.04125490  0.01054498 -0.15857393 -0.13627254
## age          0.04125490  1.00000000  0.33057310 -0.11579098  0.01649107
## income       0.01054498  0.33057310  1.00000000 -0.05367639  0.28157133
## share       -0.15857393 -0.11579098 -0.05367639  1.00000000  0.83873070
## expenditure -0.13627254  0.01649107  0.28157133  0.83873070  1.00000000
## dependents   0.02025483  0.21907121  0.31740910 -0.08468480  0.05046493
## months       0.04844560  0.45461945  0.13051897 -0.05428677 -0.02811653
## active       0.20813937  0.18538908  0.18109389 -0.02552547  0.05315996
```

```
##              dependents      months      active
## reports      0.02025483  0.04844560  0.20813937
## age          0.21907121  0.45461945  0.18538908
## income       0.31740910  0.13051897  0.18109389
## share       -0.08468480 -0.05428677 -0.02552547
## expenditure  0.05046493 -0.02811653  0.05315996
## dependents   1.00000000  0.04915109  0.10574382
## months       0.04915109  1.00000000  0.10159357
## active       0.10574382  0.10159357  1.00000000
```

```
corrplot(corr, method="color", tl.col="black")
```




## Model Fitting

```
# remove the two variables we will not be considering in our model
derogatory_rem <- subset(derogatory, select = -c(card, share))

# fit the Poisson model
mod_poisson <- glm(reports ~ ., family = poisson, data = derogatory_rem)
summary(mod_poisson)
```

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

```
## glm(formula = reports ~ ., family = poisson, data = derogatory_rem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8570  -0.9491  -0.7088  -0.3444   7.4064
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.1481978  0.1805244  -6.360 2.01e-10 ***
## age            0.0008230  0.0049259   0.167 0.867308
## income         0.0657931  0.0265197   2.481 0.013104 *
## expenditure   -0.0038057  0.0003669 -10.373  < 2e-16 ***
## owneryes      -0.7819979  0.1027541  -7.610 2.73e-14 ***
## selfempyes    -0.0236909  0.1502978  -0.158 0.874751
## dependents     0.0881746  0.0355811   2.478 0.013207 *
## months         0.0023639  0.0006192   3.818 0.000135 ***
## majorcardsyes -0.0308771  0.1056589  -0.292 0.770108
## active         0.0768453  0.0046422  16.554  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2341.5  on 1311  degrees of freedom
## Residual deviance: 1897.6  on 1302  degrees of freedom
## AIC: 2565.1
##
## Number of Fisher Scoring iterations: 6
```

```
# fit the Poisson model allowing for overdispersion
mod_quasipoisson <- glm(reports ~ ., family = quasipoisson, data = derogatory_rem)
summary(mod_quasipoisson)
```

```
##
## Call:
## glm(formula = reports ~ ., family = quasipoisson, data = derogatory_rem)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8570  -0.9491  -0.7088  -0.3444   7.4064
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.148198   0.412838  -2.781  0.00549 **
## age            0.000823   0.011265   0.073  0.94177
## income         0.065793   0.060647   1.085  0.27819
## expenditure   -0.003806   0.000839  -4.536 6.26e-06 ***
## owneryes      -0.781998   0.234987  -3.328  0.00090 ***
## selfempyes    -0.023691   0.343713  -0.069  0.94506
## dependents     0.088175   0.081370   1.084  0.27873
## months         0.002364   0.001416   1.669  0.09527 .
## majorcardsyes -0.030877   0.241629  -0.128  0.89834
## active         0.076845   0.010616   7.239 7.73e-13 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.229832)
##
##     Null deviance: 2341.5  on 1311  degrees of freedom
## Residual deviance: 1897.6  on 1302  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```r
# fit the negative binomial model
mod_nb <- glm.nb(reports ~ ., data = derogatory_rem)
summary(mod_nb)
```

```
##
## Call:
## glm.nb(formula = reports ~ ., data = derogatory_rem, init.theta = 0.2639500296,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4219  -0.6773  -0.5594  -0.3726   2.5302
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.9225923  0.3239929  -5.934 2.96e-09 ***
## age            0.0045524  0.0090345   0.504   0.6143
## income         0.0825333  0.0496417   1.663   0.0964 .
## expenditure   -0.0023705  0.0004364  -5.432 5.56e-08 ***
## owneryes      -0.8182972  0.1780034  -4.597 4.28e-06 ***
## selfempyes     0.0315603  0.2818887   0.112   0.9109
## dependents     0.0930299  0.0636163   1.462   0.1436
## months         0.0024388  0.0011892   2.051   0.0403 *
## majorcardsyes  0.0173520  0.1966613   0.088   0.9297
## active         0.1208934  0.0114956  10.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.264) family taken to be 1)
##
##     Null deviance: 838.54  on 1311  degrees of freedom
## Residual deviance: 680.00  on 1302  degrees of freedom
## AIC: 1990.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2640
##           Std. Err.:  0.0288
##
##  2 x log-likelihood:  -1968.8080
```

```r
# fit the zero-inflated Poisson model
mod_zeroinfl <- zeroinfl(reports ~ ., data = derogatory_rem)
summary(mod_zeroinfl)
```

```
##
## Call:
## zeroinfl(formula = reports ~ ., data = derogatory_rem)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.6829 -0.4260 -0.3384 -0.2344 13.5537
##
## Count model coefficients (poisson with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    7.038e-01  2.460e-01   2.861  0.00422 **
## age           -8.758e-03  6.841e-03  -1.280  0.20048
## income         7.832e-03  3.492e-02   0.224  0.82252
## expenditure   -2.259e-03  3.530e-04  -6.399 1.56e-10 ***
## owneryes      -4.728e-01  1.257e-01  -3.760  0.00017 ***
## selfempyes     6.619e-03  1.849e-01   0.036  0.97144
## dependents     7.650e-02  4.573e-02   1.673  0.09434 .
## months        -2.061e-05  7.619e-04  -0.027  0.97842
## majorcardsyes  1.510e-01  1.274e-01   1.185  0.23598
## active         3.820e-02  6.399e-03   5.969 2.38e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.2148826  0.3964404   5.587 2.31e-08 ***
## age           -0.0138532  0.0113434  -1.221  0.22199
## income        -0.0916060  0.0626910  -1.461  0.14395
## expenditure    0.0002395  0.0005006   0.478  0.63236
## owneryes       0.5149699  0.2075130   2.482  0.01308 *
## selfempyes    -0.0688650  0.3250118  -0.212  0.83220
## dependents    -0.0063217  0.0748275  -0.084  0.93267
## months        -0.0040930  0.0015418  -2.655  0.00794 **
## majorcardsyes  0.2758467  0.2247576   1.227  0.21971
## active        -0.0835282  0.0135779  -6.152 7.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 27
## Log-likelihood: -1029 on 20 Df
```

Check the multicollinearity between the variables within the negative binomial model after the removal of `card` and `share`, we see the variance inflation factor is low for all variables here, meaning there should not be much multicollinearity in our model. Also, we want to note that when we try to fit the negative binomial model with both `share` and `expenditure`, we are unable to obtain a fit (errors out). This is potentially due to the high correlation between the two variables, further verifying that one should be dropped.

```r
VIF(mod_nb)
```

```
##         age      income expenditure       owner     selfemp  dependents
```

```
##    1.526006    1.331585    1.085231    1.410547    1.020737    1.187199
##      months  majorcards      active
##    1.316622    1.027685    1.139831
```

Use AIC to compare the Poisson regression models to the negative binomial regression model. The lower value indicates a better fit. Thus, one can see that the negative binomial here fits better than the Poisson models.

```
print(paste("Poisson AIC: ", extractAIC(mod_poisson)[2]))
```

```
## [1] "Poisson AIC:  2565.0961948417"
```

```
print(paste("Negative Binomial AIC: ", extractAIC(mod_nb)[2]))
```

```
## [1] "Negative Binomial AIC:  1988.80781187038"
```

```
print(paste("Zero-Inflated Poisson AIC: ", extractAIC(mod_zeroinfl)[2]))
```

```
## [1] "Zero-Inflated Poisson AIC:  2098.73320693679"
```

Since the negative binomial regression model achieves a better fit, we choose to use that model. We calculate the 95% confidence intervals for the variables to display in the report.

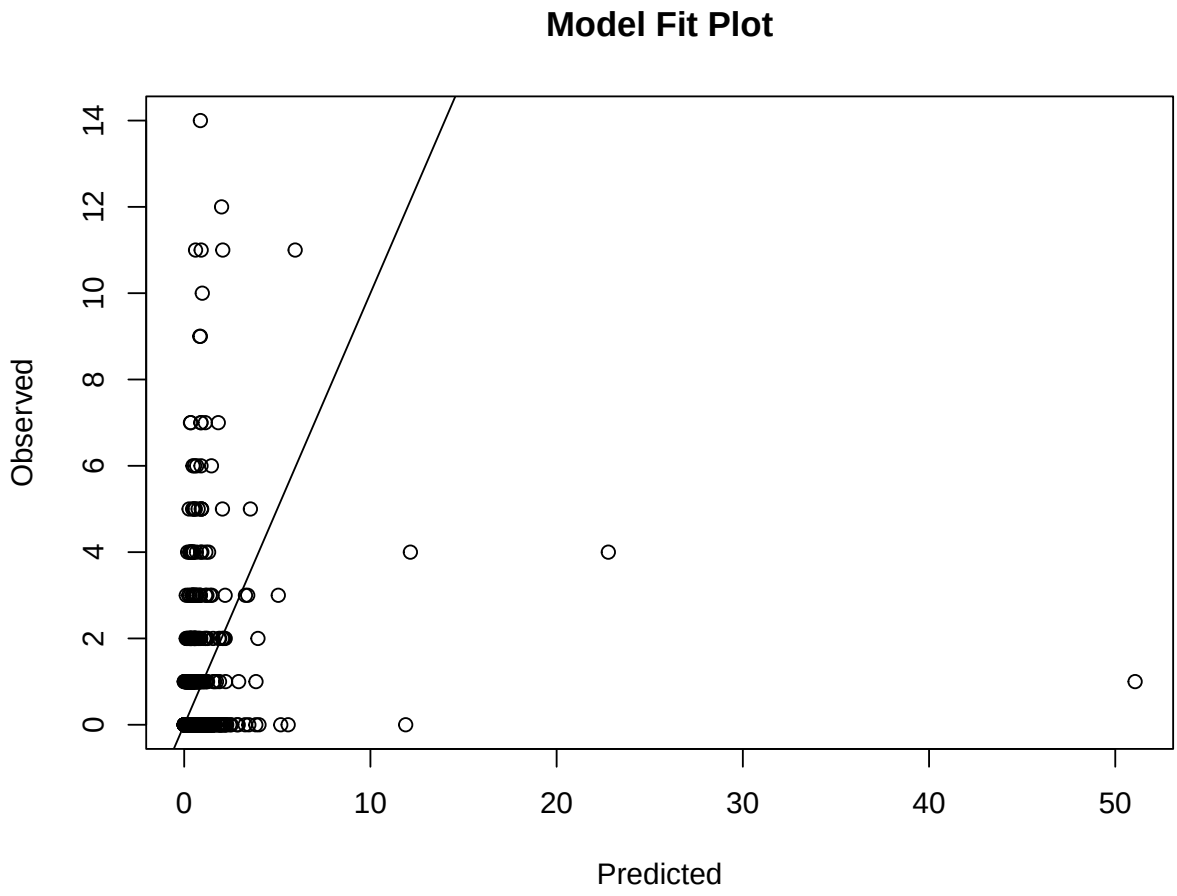
```
ci_table <- round(exp(confint(mod_nb, level=0.95)), digits = 3)
coef_table <- round(exp(coef(mod_nb)), digits = 3)
p_table <- round(coef(summary(mod_nb))[,4], digits = 3)
reports_table <- cbind.data.frame(coef_table, ci_table, p_table)
reports_table
```

```
##               coef_table 2.5 % 97.5 % p_table
## (Intercept)        0.146 0.074  0.286   0.000
## age                1.005 0.986  1.023   0.614
## income             1.086 0.985  1.202   0.096
## expenditure        0.998 0.997  0.998   0.000
## owneryes           0.441 0.312  0.621   0.000
## selfempyes         1.032 0.592  1.838   0.911
## dependents         1.097 0.967  1.248   0.144
## months             1.002 1.000  1.005   0.040
## majorcardsyes      1.018 0.695  1.481   0.930
## active             1.129 1.098  1.162   0.000
```

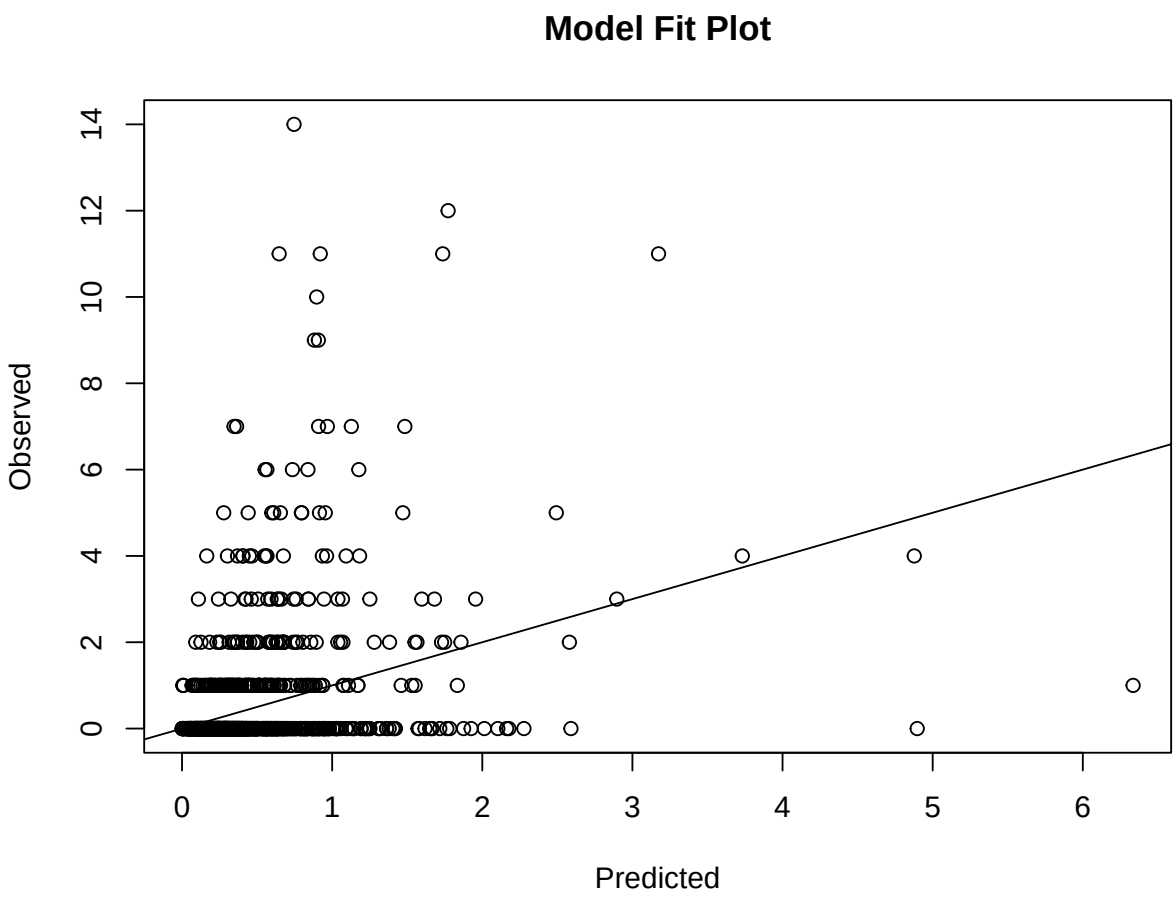
## Diagnostics

Plot the fitted values against the observed values to evaluate the fit of the best model, also present the zero inflated model for comparison. The fit does not seem to be great in either case. We see that there is a very large predicted value for the negative binomial regression model that seems to impact the overall fit; however, the rest of the model seems to fit better.

```r
plot(predict(mod_nb, newdata = derogatory_rem, type = "response"),
     derogatory_rem$reports,
     xlab = "Predicted", ylab = "Observed", main = "Model Fit Plot")
abline(a = 0, b = 1)
```

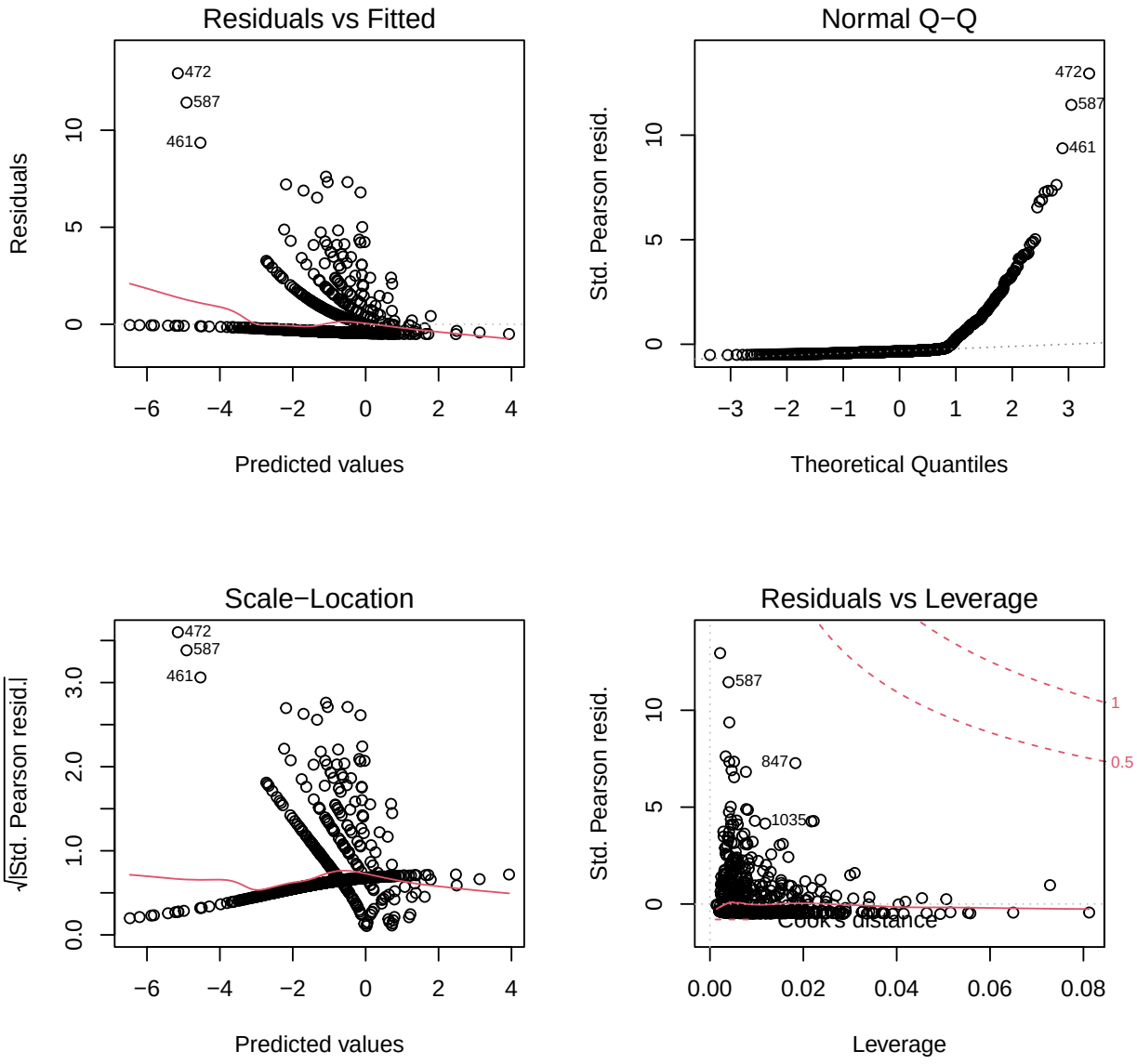



```r
plot(predict(mod_zeroinfl, newdata = derogatory_rem, type = "response"),
     derogatory_rem$reports,
     xlab = "Predicted", ylab = "Observed", main = "Model Fit Plot")
abline(a = 0, b = 1)
```

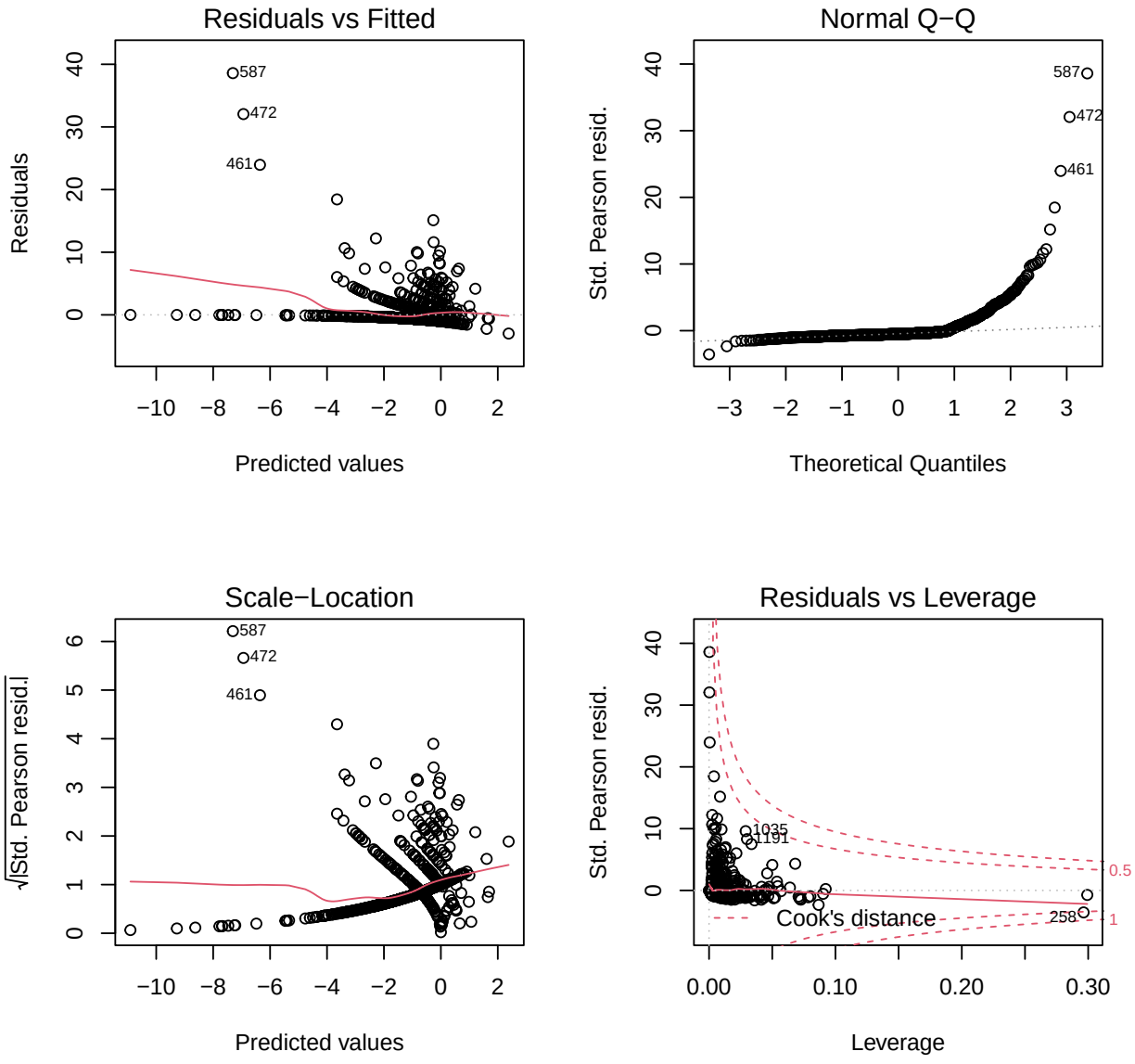



Other diagnostics plots for the negative binomial model and Poisson model.

```
par(mfrow = c(2,2))
# Negative binomial diagnostics
plot(mod_nb)
```




```
# Poisson diagnostics
plot(mod_poisson)
```


Residuals vs Fitted
Residuals
Predicted values
587
472
461
Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
Scale−Location
√|Std. Pearson resid.|
Residuals vs Leverage
Cook's distance
Leverage
1035
1191
258
0.5
1