Assignment 1 Appendix

```
library(stats)
library(ggplot2)
library(MASS)
```

#1 Data skimming

```
dero = read.csv(file = 'C:/Users/krss9/Desktop/FS21/Stats 504/derogatory.csv')
head(dero)
```

```
##   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
```

#2 Converting categorical variables into numeric

```
dero$card2 <- ifelse(dero$card == 'yes', 1,0)
dero$owner2 <- ifelse(dero$owner == 'yes', 1, 0)
dero$selfemp2 <- ifelse(dero$selfemp == 'yes', 1, 0)
dero$majorcards2 <- ifelse(dero$majorcards == 'yes', 1, 0)
```

#3 glm with Poisson assumption

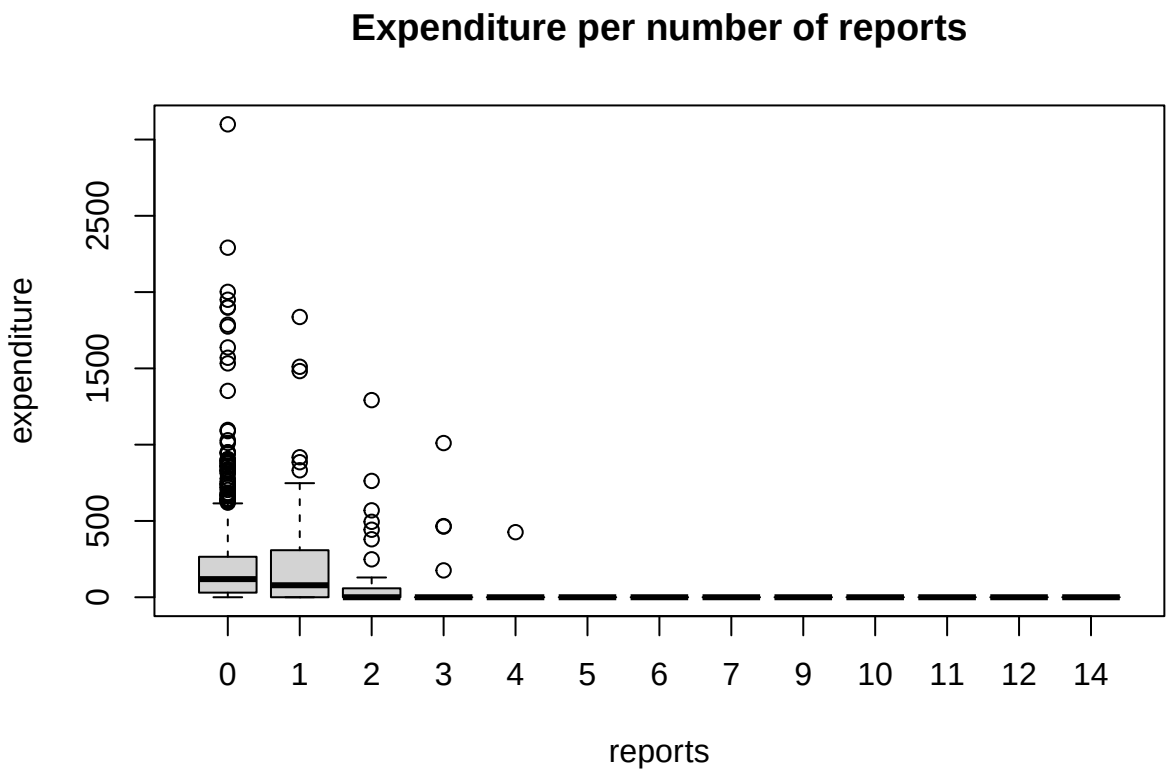
```
expr1 = 'reports ~ age + income + expenditure + owner2 + selfemp2 + dependents + months + majorcards2 +
model1_GLM = glm(expr1, family=poisson(), data=dero)
summary(model1_GLM)
```

```
##
## Call:
## glm(formula = expr1, family = poisson(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

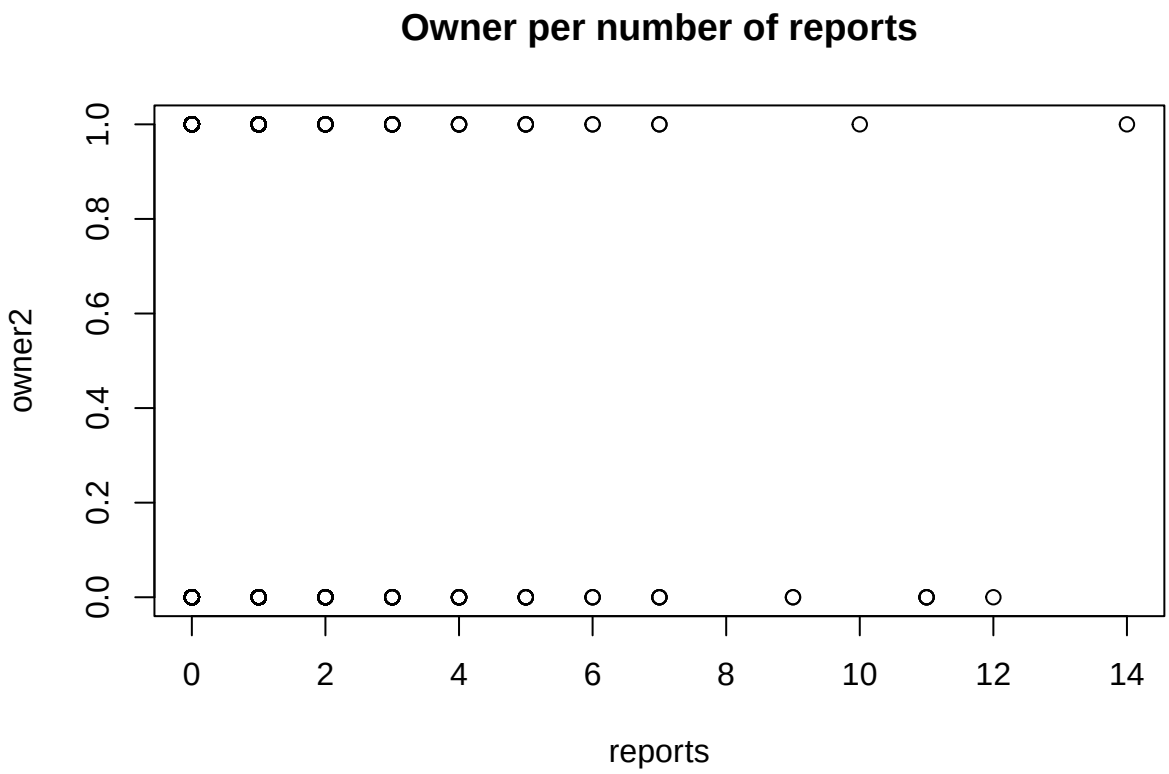
```
## -3.8691  -0.9467  -0.7081  -0.3476   7.3921
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1794113  0.1763576  -6.688 2.27e-11 ***
## age          0.0018484  0.0047569   0.389 0.697598
## income       0.0655500  0.0264859   2.475 0.013327 *
## expenditure -0.0038243  0.0003674 -10.409  < 2e-16 ***
## owner2      -0.7866639  0.1027559  -7.656 1.92e-14 ***
## selfemp2    -0.0252848  0.1503499  -0.168 0.866447
## dependents   0.0881904  0.0355773   2.479 0.013181 *
## months       0.0023190  0.0006124   3.787 0.000153 ***
## majorcards2 -0.0298881  0.1052483  -0.284 0.776428
## active       0.0767950  0.0046391  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: 2347.4  on 1318  degrees of freedom
## Residual deviance: 1901.0  on 1309  degrees of freedom
## AIC: 2570.5
##
## Number of Fisher Scoring iterations: 6
```

#4 Graphs of variables and reports

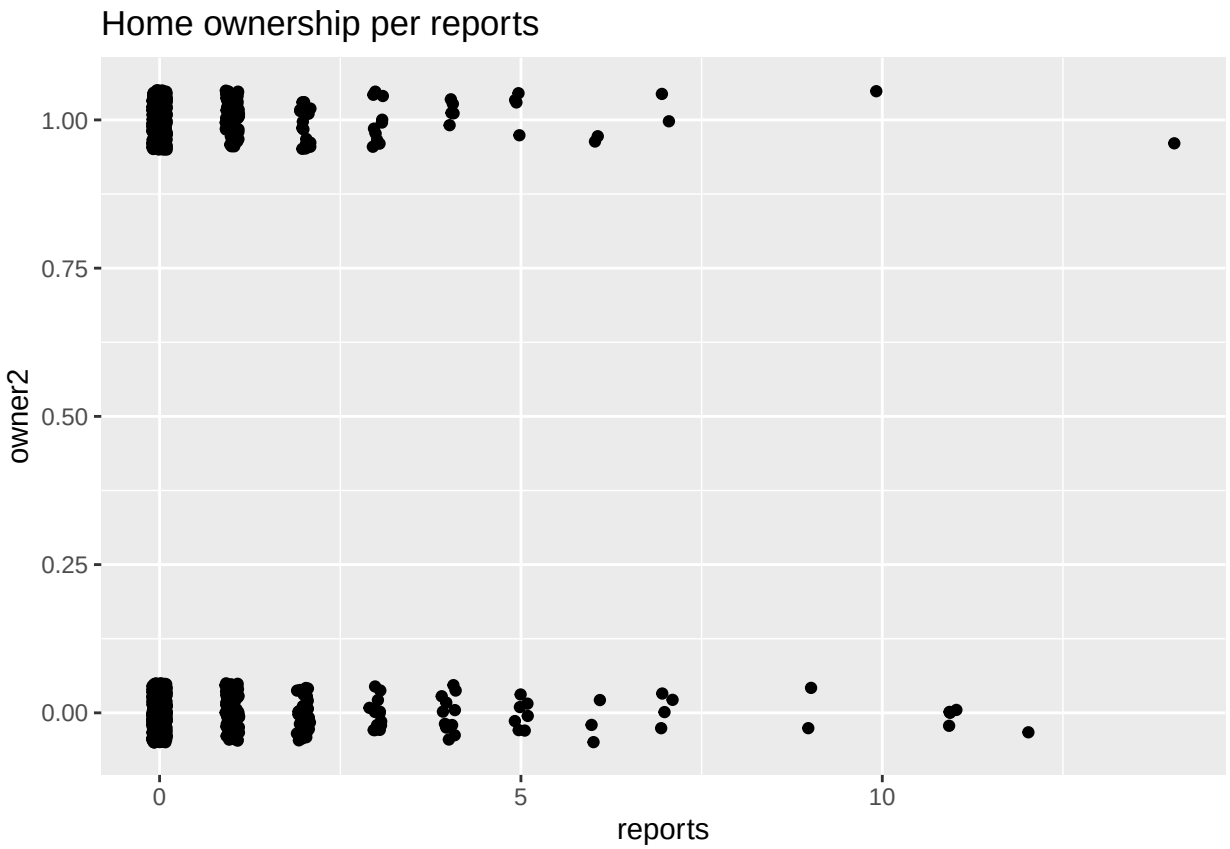
```
boxplot(expenditure ~ reports, data=dero, main ="Expenditure per number of reports")
```



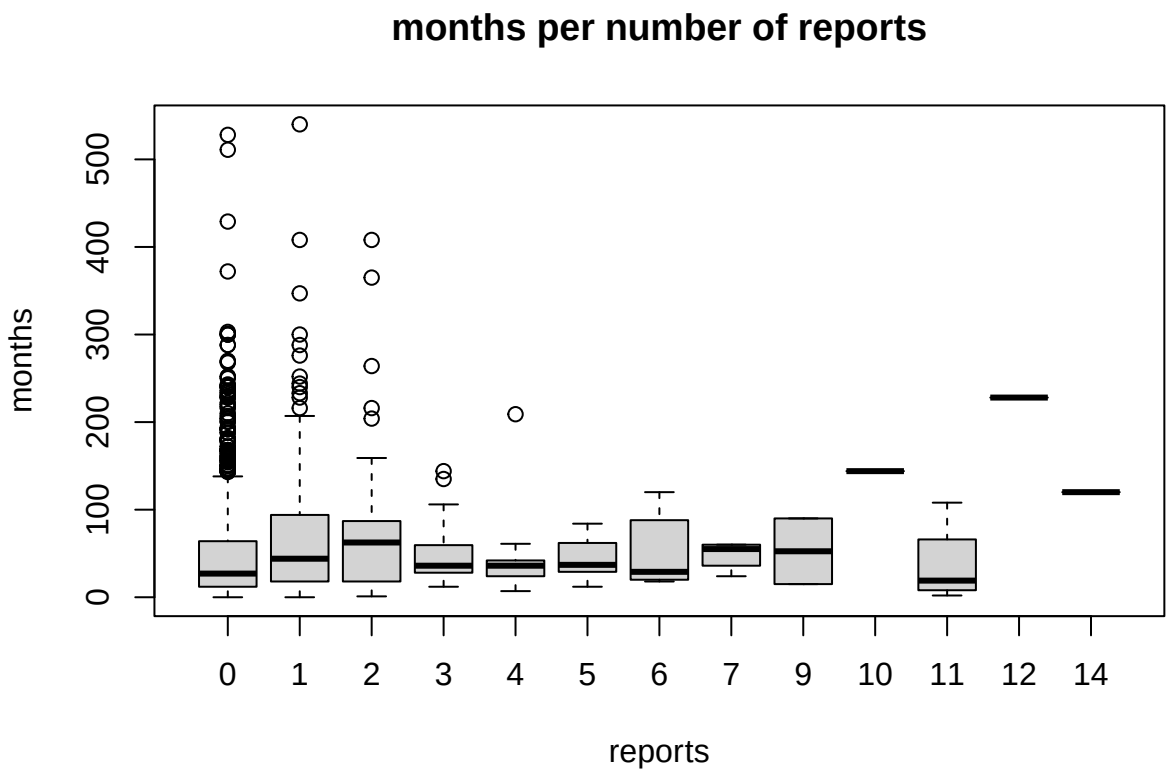

```
plot(owner2 ~ reports, data=dero, main ="Owner per number of reports")
```



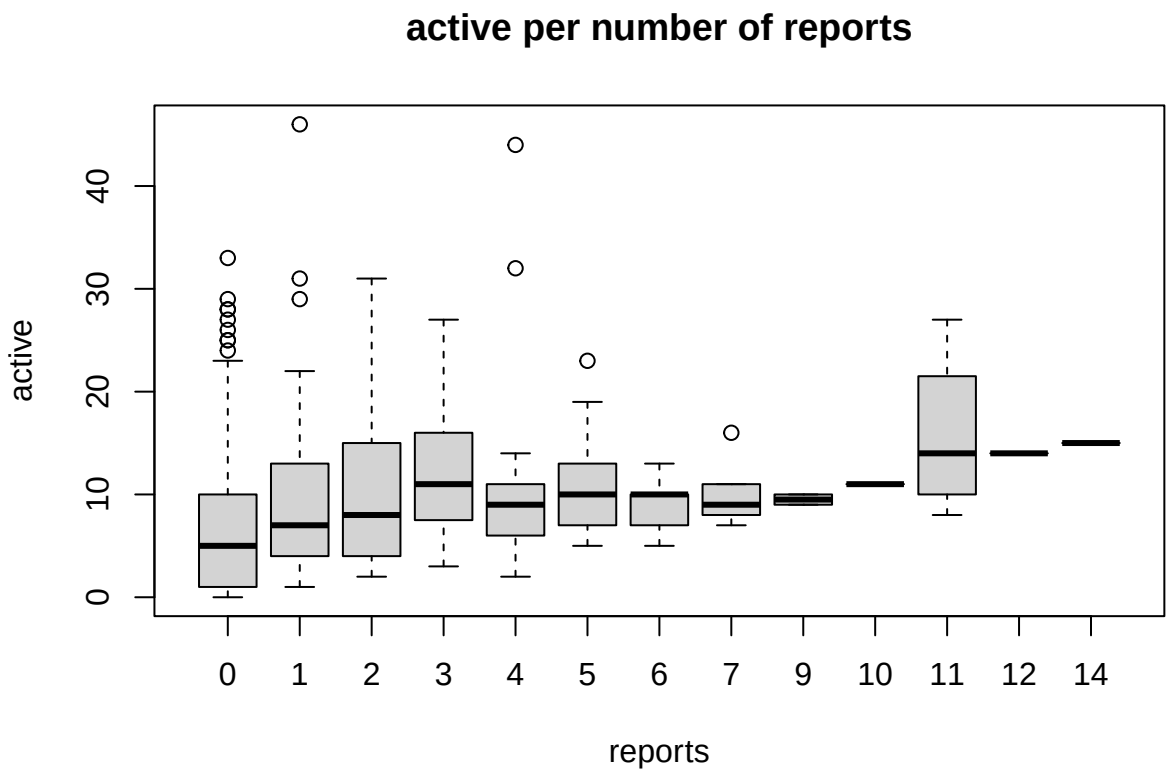

```
ggplot(data = dero, aes(x = reports, y = owner2))+
  geom_jitter(width = 0.1, height = 0.05) +
  ggtitle("Home ownership per reports")
```




```
boxplot(months ~ reports, data=dero, main ="months per number of reports")
```




```
boxplot(active ~ reports, data=dero, main ="active per number of reports")
```



#5 Distribution of reports and missing values check

```
dero_rep0 = dero[which(dero$reports == 0),]
dim(dero_rep0)
```

```
## [1] 1060   16
```

```
dero_rep1 = dero[which(dero$reports == 1),]
dim(dero_rep1)
```

```
## [1] 137  16
```

```
dero_rep2 = dero[which(dero$reports > 1),]
dim(dero_rep2)
```

```
## [1] 122  16
```

```
which(is.na(dero))
```

```
## integer(0)
```

#6 glm with four assumptions _ Guassian, Poisson, NB, Quasi

```r
expr1 = 'reports ~ age + income + expenditure + owner2 + selfemp2 + dependents + months + majorcards2 +
model1_GLM = glm(expr1, family=poisson(), data=dero)
summary(model1_GLM)
```

```
##
## Call:
## glm(formula = expr1, family = poisson(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8691  -0.9467  -0.7081  -0.3476   7.3921
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1794113  0.1763576  -6.688 2.27e-11 ***
## age          0.0018484  0.0047569   0.389 0.697598
## income       0.0655500  0.0264859   2.475 0.013327 *
## expenditure -0.0038243  0.0003674 -10.409  < 2e-16 ***
## owner2      -0.7866639  0.1027559  -7.656 1.92e-14 ***
## selfemp2    -0.0252848  0.1503499  -0.168 0.866447
## dependents   0.0881904  0.0355773   2.479 0.013181 *
## months       0.0023190  0.0006124   3.787 0.000153 ***
## majorcards2 -0.0298881  0.1052483  -0.284 0.776428
## active       0.0767950  0.0046391  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: 2347.4  on 1318  degrees of freedom
## Residual deviance: 1901.0  on 1309  degrees of freedom
## AIC: 2570.5
##
## Number of Fisher Scoring iterations: 6
```

```r
expr2 = 'reports ~ age + income + expenditure + owner2 + selfemp2 + dependents + months + majorcards2 +
model2_GLM = glm.nb(expr2, data=dero)
summary(model2_GLM)
```

```
##
## Call:
## glm.nb(formula = expr2, data = dero, init.theta = 0.2648349349,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4214  -0.6764  -0.5587  -0.3723   2.5341
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.9646475  0.3155450  -6.226 4.78e-10 ***
## age          0.0060381  0.0086661   0.697   0.4860
## income       0.0830533  0.0494830   1.678   0.0933 .
```

```
## expenditure -0.0023861  0.0004366  -5.465 4.63e-08 ***
## owner2      -0.8233963  0.1776964  -4.634 3.59e-06 ***
## selfemp2     0.0322711  0.2815948   0.115   0.9088
## dependents   0.0909177  0.0634911   1.432   0.1522
## months       0.0023967  0.0011723   2.044   0.0409 *
## majorcards2  0.0132247  0.1957436   0.068   0.9461
## active       0.1207594  0.0114787  10.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2648) family taken to be 1)
##
##     Null deviance: 843.27  on 1318  degrees of freedom
## Residual deviance: 683.24  on 1309  degrees of freedom
## AIC: 1996.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2648
##           Std. Err.:  0.0289
##
##  2 x log-likelihood:  -1974.9280
```

```
expr3 = 'reports ~ age + income + expenditure + owner2 + selfemp2 + dependents + months + majorcards2 +
model3_GLM = glm(expr3, family=gaussian(), data=dero)
summary(model3_GLM)
```

```
##
## Call:
## glm(formula = expr3, family = gaussian(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9799  -0.5228  -0.3066   0.0473  13.1757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1928631  0.1490863   1.294    0.196
## age          0.0027089  0.0042444   0.638    0.523
## income       0.0284319  0.0247086   1.151    0.250
## expenditure -0.0007197  0.0001376  -5.231 1.96e-07 ***
## owner2      -0.3845987  0.0833708  -4.613 4.36e-06 ***
## selfemp2     0.0145824  0.1421554   0.103    0.918
## dependents   0.0303756  0.0311287   0.976    0.329
## months       0.0007878  0.0006040   1.304    0.192
## majorcards2 -0.0638952  0.0935942  -0.683    0.495
## active       0.0511610  0.0059530   8.594  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.674134)
##
##     Null deviance: 2385.2  on 1318  degrees of freedom
```

```
## Residual deviance: 2191.4  on 1309  degrees of freedom
## AIC: 4434.8
##
## Number of Fisher Scoring iterations: 2
```

```
expr4 = 'reports ~ age + income + expenditure + owner2 + selfemp2 + dependents + months + majorcards2 +
model4_GLM = glm(expr4, family=quasi(), data=dero)
summary(model4_GLM)
```

```
##
## Call:
## glm(formula = expr4, family = quasi(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9799  -0.5228  -0.3066   0.0473  13.1757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1928631  0.1490863   1.294    0.196
## age          0.0027089  0.0042444   0.638    0.523
## income       0.0284319  0.0247086   1.151    0.250
## expenditure -0.0007197  0.0001376  -5.231 1.96e-07 ***
## owner2      -0.3845987  0.0833708  -4.613 4.36e-06 ***
## selfemp2     0.0145824  0.1421554   0.103    0.918
## dependents   0.0303756  0.0311287   0.976    0.329
## months       0.0007878  0.0006040   1.304    0.192
## majorcards2 -0.0638952  0.0935942  -0.683    0.495
## active       0.0511610  0.0059530   8.594  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 1.674134)
##
##     Null deviance: 2385.2  on 1318  degrees of freedom
## Residual deviance: 2191.4  on 1309  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 2
```

#7 Correlation table

```
res <- cor(dero[,c(2,3,4,5,6,9,10,12,14,15,16)])
as.data.frame(round(res, 3))
```

```
##             reports    age income  share expenditure dependents months active
## reports       1.000  0.044  0.011 -0.159      -0.137      0.020  0.049  0.208
## age           0.044  1.000  0.325 -0.116       0.015      0.212  0.436  0.181
## income        0.011  0.325  1.000 -0.054       0.281      0.318  0.130  0.181
## share        -0.159 -0.116 -0.054  1.000       0.839     -0.083 -0.055 -0.023
## expenditure  -0.137  0.015  0.281  0.839       1.000      0.053 -0.029  0.055
## dependents    0.020  0.212  0.318 -0.083       0.053      1.000  0.047  0.107
## months        0.049  0.436  0.130 -0.055      -0.029      0.047  1.000  0.100
```

```
## active        0.208  0.181  0.181 -0.023       0.055      0.107  0.100  1.000
## owner2       -0.054  0.368  0.325 -0.016       0.093      0.309  0.239  0.275
## selfemp2      0.019  0.100  0.112 -0.079      -0.036      0.042  0.066  0.030
## majorcards2  -0.007  0.010  0.107  0.051       0.078      0.010 -0.041  0.120
##             owner2 selfemp2 majorcards2
## reports     -0.054    0.019      -0.007
## age          0.368    0.100       0.010
## income       0.325    0.112       0.107
## share       -0.016   -0.079       0.051
## expenditure  0.093   -0.036       0.078
## dependents   0.309    0.042       0.010
## months       0.239    0.066      -0.041
## active       0.275    0.030       0.120
## owner2       1.000    0.042       0.064
## selfemp2     0.042    1.000       0.005
## majorcards2  0.064    0.005       1.000
```

#8 glm with fewer variables

```
expr_fin = 'reports ~ expenditure + owner2 + months + active'
model5_GLM = glm(expr_fin, family = poisson(), data = dero)
model6_GLM = glm.nb(expr_fin, data = dero)
model7_GLM = glm(expr_fin, family = gaussian(), data = dero)
model8_GLM = glm(expr_fin, family = quasi(), data = dero)
```

```
summary(model5_GLM)
```

```
##
## Call:
## glm(formula = expr_fin, family = poisson(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4627  -0.9392  -0.7245  -0.3446   7.0969
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.9127446  0.0769756 -11.858  < 2e-16 ***
## expenditure -0.0037563  0.0003662 -10.258  < 2e-16 ***
## owner2      -0.6409897  0.0922590  -6.948 3.71e-12 ***
## months       0.0024628  0.0005512   4.468 7.89e-06 ***
## active       0.0775611  0.0045252  17.140  < 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: 2347.4  on 1318  degrees of freedom
## Residual deviance: 1917.4  on 1314  degrees of freedom
## AIC: 2576.9
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(model6_GLM)
```

```
##
## Call:
## glm.nb(formula = expr_fin, data = dero, init.theta = 0.2584495758,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4419  -0.6779  -0.5687  -0.3689   2.5203
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.5035756  0.1407527 -10.682  < 2e-16 ***
## expenditure -0.0022398  0.0004248  -5.273 1.34e-07 ***
## owner2      -0.6428863  0.1626833  -3.952 7.76e-05 ***
## months       0.0026630  0.0010710   2.487   0.0129 *
## active       0.1239529  0.0114320  10.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2584) family taken to be 1)
##
##     Null deviance: 833.08  on 1318  degrees of freedom
## Residual deviance: 682.65  on 1314  degrees of freedom
## AIC: 1994.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2584
##           Std. Err.:  0.0281
##
##  2 x log-likelihood:  -1982.2150
```

```r
summary(model7_GLM)
```

```
##
## Call:
## glm(formula = expr_fin, family = gaussian(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0421  -0.5351  -0.3142   0.0526  13.1192
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3062669  0.0650670   4.707 2.78e-06 ***
## expenditure -0.0006785  0.0001317  -5.151 2.99e-07 ***
## owner2      -0.3233147  0.0768345  -4.208 2.75e-05 ***
## months       0.0009976  0.0005549   1.798   0.0724 .
## active       0.0518767  0.0058845   8.816  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.673544)
##
##     Null deviance: 2385.2  on 1318  degrees of freedom
## Residual deviance: 2199.0  on 1314  degrees of freedom
## AIC: 4429.4
##
## Number of Fisher Scoring iterations: 2
```

```
summary(model8_GLM)
```

```
##
## Call:
## glm(formula = expr_fin, family = quasi(), data = dero)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0421  -0.5351  -0.3142   0.0526  13.1192
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3062669  0.0650670   4.707 2.78e-06 ***
## expenditure -0.0006785  0.0001317  -5.151 2.99e-07 ***
## owner2      -0.3233147  0.0768345  -4.208 2.75e-05 ***
## months       0.0009976  0.0005549   1.798   0.0724 .
## active       0.0518767  0.0058845   8.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 1.673544)
##
##     Null deviance: 2385.2  on 1318  degrees of freedom
## Residual deviance: 2199.0  on 1314  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 2
```

#9 AIC comparison

```
AIC(model5_GLM)
```

```
## [1] 2576.902
```

```
AIC(model6_GLM)
```

```
## [1] 1994.215
```

```
AIC(model7_GLM)
```

```
## [1] 4429.361
```

#10 Summary of the selected model

```
summary(model6_GLM)
```

```
##
## Call:
## glm.nb(formula = expr_fin, data = dero, init.theta = 0.2584495758,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4419  -0.6779  -0.5687  -0.3689   2.5203
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.5035756  0.1407527 -10.682  < 2e-16 ***
## expenditure -0.0022398  0.0004248  -5.273 1.34e-07 ***
## owner2      -0.6428863  0.1626833  -3.952 7.76e-05 ***
## months       0.0026630  0.0010710   2.487   0.0129 *
## active       0.1239529  0.0114320  10.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2584) family taken to be 1)
##
##     Null deviance: 833.08  on 1318  degrees of freedom
## Residual deviance: 682.65  on 1314  degrees of freedom
## AIC: 1994.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2584
##           Std. Err.:  0.0281
##
##  2 x log-likelihood:  -1982.2150
```

```
summary(model6_GLM)$coefficients[,c(1,2,4)]
```

```
##                 Estimate   Std. Error     Pr(>|z|)
## (Intercept) -1.503575613 0.1407527227 1.230606e-26
## expenditure -0.002239835 0.0004247969 1.344162e-07
## owner2      -0.642886284 0.1626833416 7.757701e-05
## months       0.002663039 0.0010709985 1.290063e-02
## active       0.123952861 0.0114319985 2.161801e-27
```