Assignment #1 Appendix

1. Exploratory Data Analysis

Variables

Variable Name	Description
`card`	Was the application for a credit card accepted?
`reports`	Number of derogatory reports
`age`	Applicant age in years at time of application
`income`	Yearly income in 10,000 USD
`share`	Ratio of monthly credit card expenditure to yearly income (generated from `income` and `expenditure`)
`expenditure`	Average monthly credit card expenditure
`owner`	Does the applicant own their home?
`selfemp`	Is the individual self-employed?
`dependents`	Number of dependents
`months`	Number of months living at current address
`majorcards`	Does the applicant have other major credit cards?
`active`	Number of active credit accounts

Summary of EDA

- There are 7 observations with age of less than 18 years old. This is noticeable because people can only apply to credit card starting at the age of 18.
- The variable `reports` (the number of derogatory reports) contains many zeros. Need to keep this in mind when choosing a model.
- Correlations:
  - Not many variables correlated with `reports`. `expenditure` is negatively correlated (the more you spend using credit cards, the less deragatory reports you have) and `active` is positively correlated (the more active cards you have, the more derogatory reports)
- Almost all distributions for the numeric variables have a skewed distribution.

Structure of dataset

```
## 'data.frame':    1319 obs. of  12 variables:
##  $ card       : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ reports    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age        : num  37.7 33.2 33.7 30.5 32.2 ...
##  $ income     : num  4.52 2.42 4.5 2.54 9.79 ...
```

```
##  $ share      : num  0.03327 0.00522 0.00416 0.06521 0.06705 ...
##  $ expenditure: num  124.98 9.85 15 137.87 546.5 ...
##  $ owner      : Factor w/ 2 levels "no","yes": 2 1 2 1 2 1 1 2 2 1 ...
##  $ selfemp    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dependents : int  3 3 4 0 2 0 2 0 0 0 ...
##  $ months     : int  54 34 58 25 64 54 7 77 97 65 ...
##  $ majorcards : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ active     : int  12 13 5 7 5 1 5 3 6 18 ...
```

Summary of entire dataset

```
##   card         reports             age              income      
##  no : 296   Min.   : 0.0000   Min.   : 0.1667   Min.   : 0.210  
##  yes:1023   1st Qu.: 0.0000   1st Qu.:25.4167   1st Qu.: 2.244  
##             Median : 0.0000   Median :31.2500   Median : 2.900  
##             Mean   : 0.4564   Mean   :33.2131   Mean   : 3.365  
##             3rd Qu.: 0.0000   3rd Qu.:39.4167   3rd Qu.: 4.000  
##             Max.   :14.0000   Max.   :83.5000   Max.   :13.500  
##      share            expenditure       owner     selfemp      dependents    
##  Min.   :0.0001091   Min.   :   0.000   no :738   no :1228   Min.   :0.0000  
##  1st Qu.:0.0023159   1st Qu.:   4.583   yes:581   yes:  91   1st Qu.:0.0000  
##  Median :0.0388272   Median : 101.298                        Median :1.0000  
##  Mean   :0.0687322   Mean   : 185.057                        Mean   :0.9939  
##  3rd Qu.:0.0936168   3rd Qu.: 249.036                        3rd Qu.:2.0000  
##  Max.   :0.9063205   Max.   :3099.505                        Max.   :6.0000  
##      months        majorcards     active      
##  Min.   :  0.00   no : 241   Min.   : 0.000  
##  1st Qu.: 12.00   yes:1078   1st Qu.: 2.000  
##  Median : 30.00              Median : 6.000  
##  Mean   : 55.27              Mean   : 6.997  
##  3rd Qu.: 72.00              3rd Qu.:11.000  
##  Max.   :540.00              Max.   :46.000  
```

SD for Numeric Variables

```
## $age
## [1] 14
## 
## $reports
## [1] 0
## 
## $income
## [1] 1.75625
## 
## $share
## [1] 0.0913009
## 
## $expenditure
## [1] 244.4525
## 
## $dependents
## [1] 2
## 
```

```
## $months
## [1] 60
##
## $active
## [1] 9
```

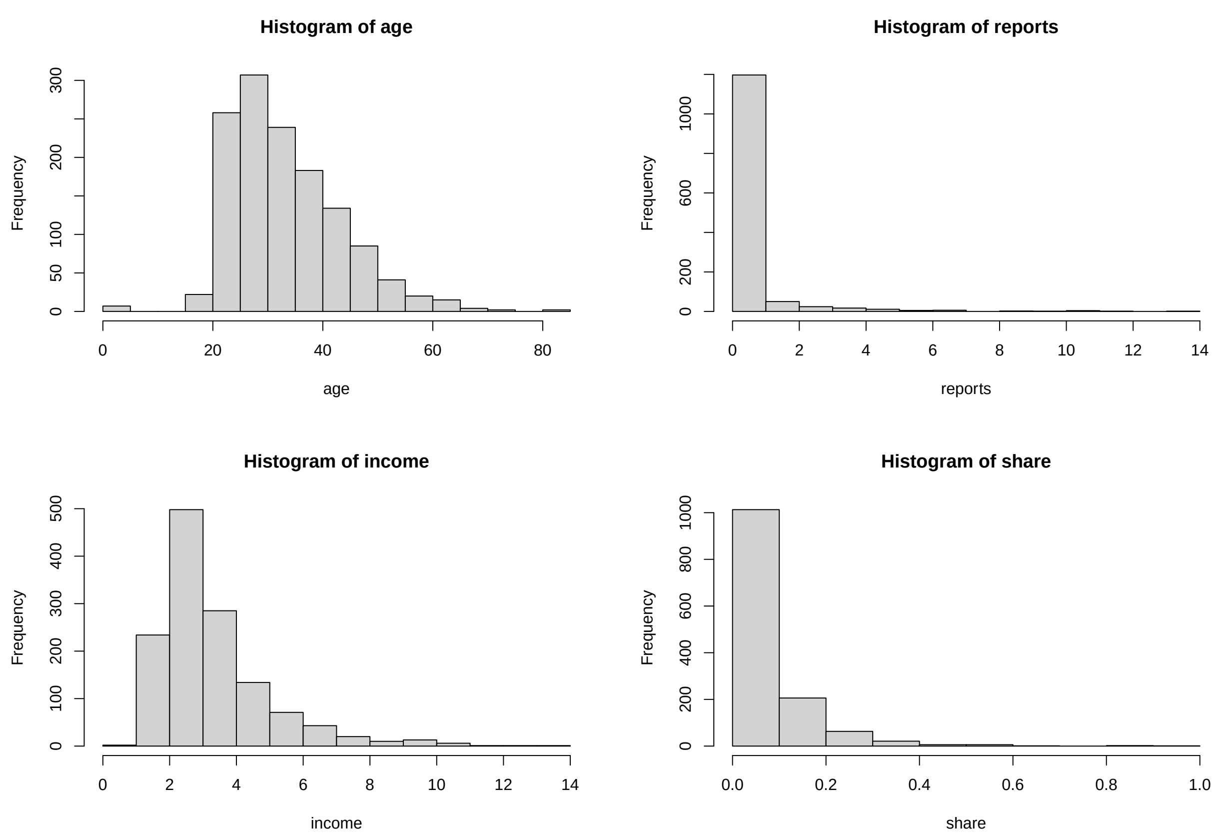
Number of applications with no deragotory reports

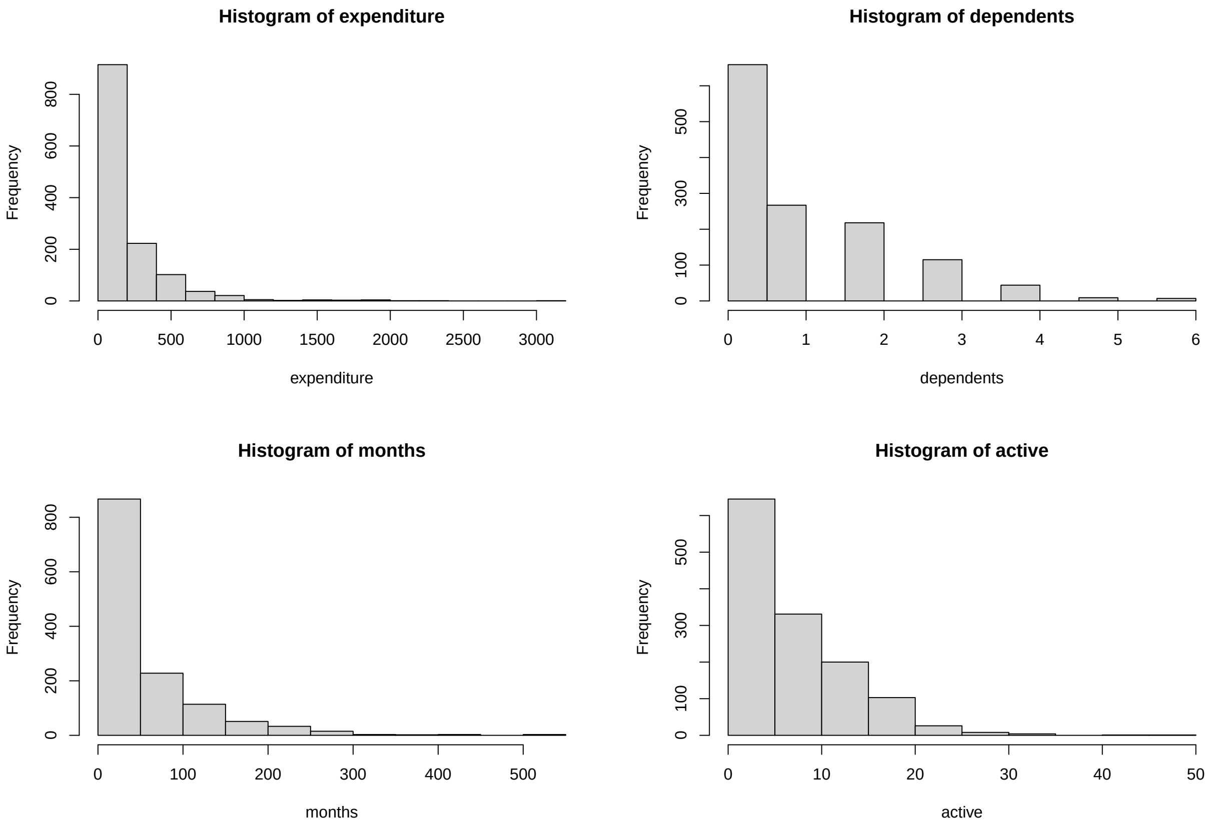
```r
# Identifies how many observations have zero derogatory reports
no_reports = (credit$reports == 0)

# Reports the proportion of observations with zero derogatory reports
sum(no_reports) / nrow(credit)
```

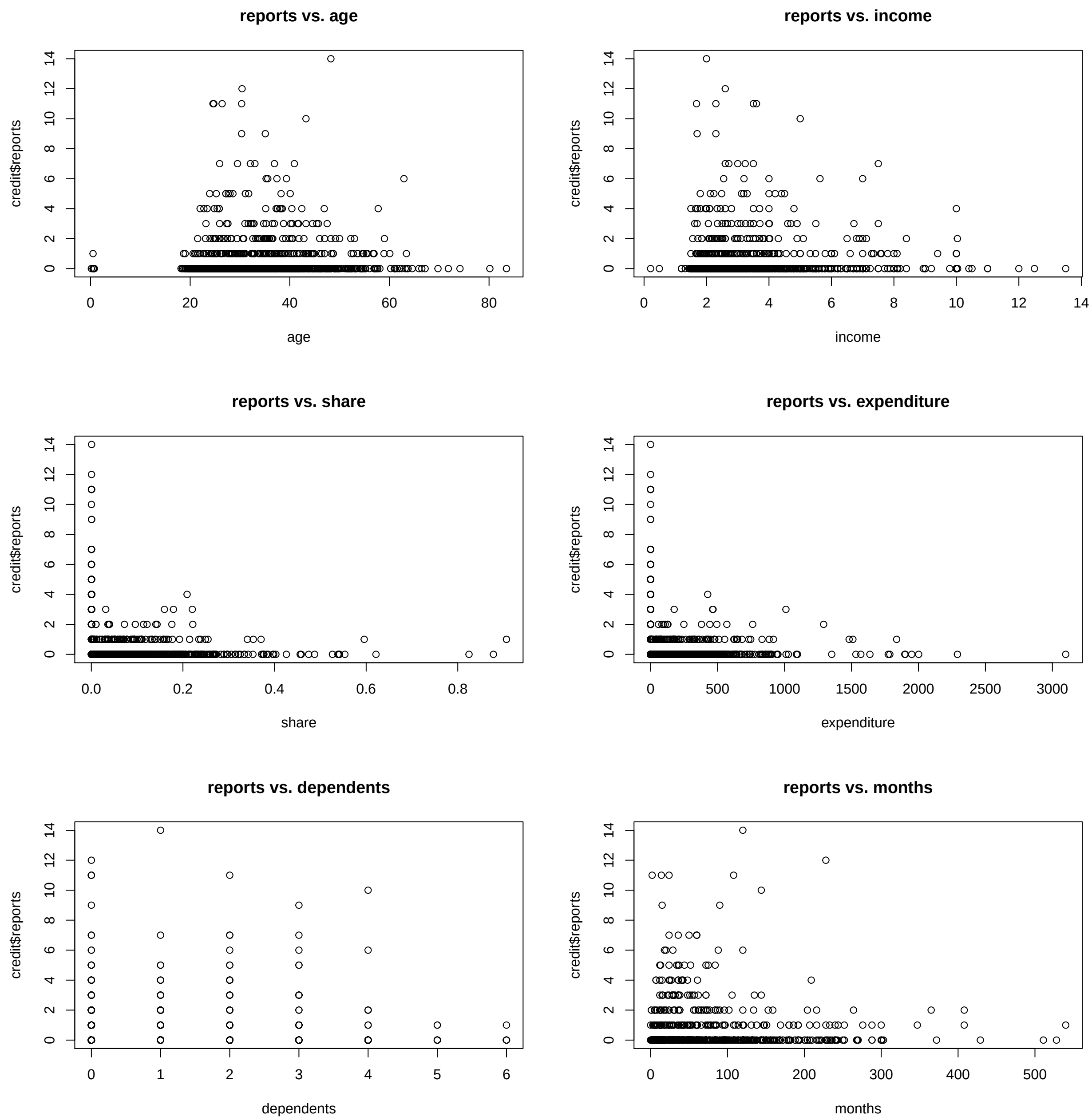
```
## [1] 0.8036391
```

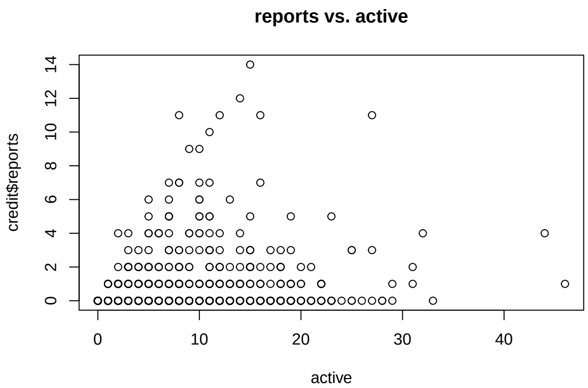
Histograms for all numeric variables




Histogram of expenditure
Frequency
0
200
400
600
800
0
500
1000
1500
2000
2500
3000
expenditure
Histogram of dependents
Frequency
0
100
300
500
0
1
2
3
4
5
6
dependents
Histogram of months
Frequency
0
200
400
600
800
0
100
200
300
400
500
months
Histogram of active
Frequency
0
100
300
500
0
10
20
30
40
50
active

Scatterplots of response vs. numeric variables





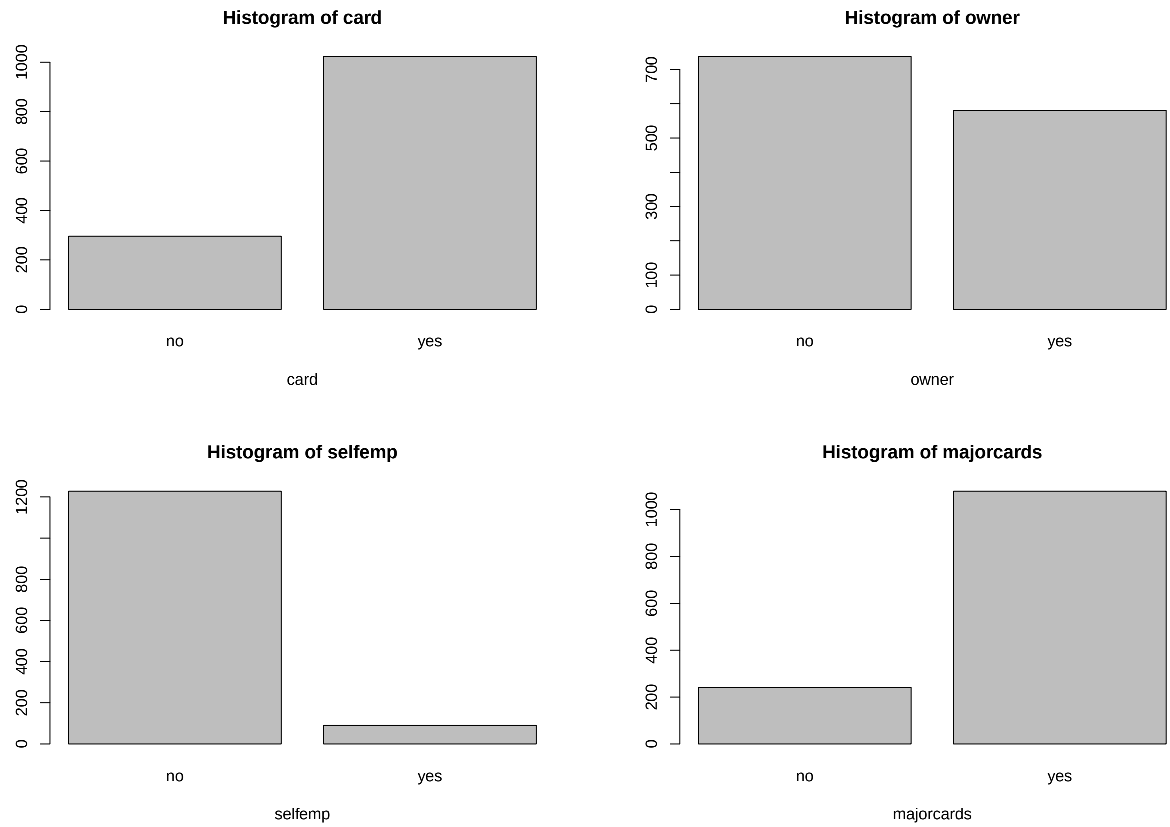
Observations with age of less than 18

```
##      card reports       age income      share expenditure
## 79    yes       0 0.5000000   3.05 0.10172430   258.54920
## 324   yes       0 0.1666667   3.24 0.18436640   497.70580
## 435   yes       0 0.5833333   2.50 0.08317120   173.02330
## 462    no       0 0.7500000   3.00 0.00040000     0.00000
## 656   yes       0 0.5833333   4.00 0.07266350   242.12830
## 659   yes       1 0.5000000   3.70 0.01063703    32.46416
## 1195  yes       0 0.7500000   1.60 0.15419060   205.25420
```

Correlation Matrix across all Numeric Variables

```
##                     age     reports      income       share expenditure
## age          1.00000000  0.04408851  0.32465320 -0.11569704  0.01494770
## reports      0.04408851  1.00000000  0.01102287 -0.15901079 -0.13653760
## income       0.32465320  0.01102287  1.00000000 -0.05442926  0.28110402
## share       -0.11569704 -0.15901079 -0.05442926  1.00000000  0.83877932
## expenditure  0.01494770 -0.13653760  0.28110402  0.83877932  1.00000000
## dependents   0.21214643  0.01973090  0.31760130 -0.08261776  0.05266406
## months       0.43642554  0.04896762  0.13034627 -0.05534756 -0.02900660
## active       0.18106971  0.20775502  0.18054026 -0.02347440  0.05472424
##              dependents      months      active
## age          0.21214643  0.43642554  0.18106971
## reports      0.01973090  0.04896762  0.20775502
## income       0.31760130  0.13034627  0.18054026
## share       -0.08261776 -0.05534756 -0.02347440
## expenditure  0.05266406 -0.02900660  0.05472424
## dependents   1.00000000  0.04651197  0.10713276
## months       0.04651197  1.00000000  0.10002764
## active       0.10713276  0.10002764  1.00000000
```

Bar Plots for All categorical varibles



2. Modeling and Diagnostics

Data Decisions

- We will be dropping the 7 observations that have an age of less than 18 years old

```
credit <- credit[!credit$age < 18, ]
```

Models

Modeling Decisions

- 7 observations with age less than 18 years old will be dropped.
- The variable `card` will not be included since this variable was created as a function of the other variables, and this will cause multicollinearity issues.
- The variable `ratio` will not be included in the model since this variable is created using `income` and `expenditure`, and since this information will already be available, we don't want redundancy in the variables of our model AND we don't want issues related to multicollinearity.
- We aim to choose the model that:
  - Handles excess amount of zeros in the `report` variable

    – Provides good interpertability of results
    – Is a good fit to the data

Discussion of each model

- Poisson Model:
    – Excess zeros in `report` will lead to problems
    – Overdispersion present
- Negative Binomial because:
    – Helps deal with overdispersion present in Poisson model
    – Helps deal with excess zeros
- Zero-Inflated Negative Binomial because:
    – Can help deal with excess zeros
    – Interpretation not clear

Poisson Regression Model

```
# Poisson model
poi.model = glm(reports ~ owner + selfemp + majorcards + age + income + expenditure + dependents + montl
summary(poi.model)
```

```
##
## Call:
## glm(formula = reports ~ owner + selfemp + majorcards + age +
##     income + expenditure + dependents + months + active, family = poisson,
##     data = credit)
##
## 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 ***
## owneryes      -0.7819979  0.1027541  -7.610 2.73e-14 ***
## selfempyes    -0.0236909  0.1502978  -0.158 0.874751
## majorcardsyes -0.0308771  0.1056589  -0.292 0.770108
## 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 ***
## dependents     0.0881746  0.0355811   2.478 0.013207 *
## months         0.0023639  0.0006192   3.818 0.000135 ***
## 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
```

```
# Overdispersion check
sigma2 = sum(residuals(poi.model, type="pearson")^2) / poi.model$df.residual
sigma2
```

```
## [1] 5.225628
```

Negative Binomial Model

```
# Negative Binomial
nb.model <- glm.nb(reports ~ owner + selfemp + majorcards + age +
                     income + expenditure + dependents + months + active, data=credit)
summary(nb.model)
```

```
##
## Call:
## glm.nb(formula = reports ~ owner + selfemp + majorcards + age +
##     income + expenditure + dependents + months + active, data = credit,
##     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 ***
## owneryes      -0.8182972  0.1780034  -4.597 4.28e-06 ***
## selfempyes     0.0315603  0.2818887   0.112   0.9109
## majorcardsyes  0.0173520  0.1966613   0.088   0.9297
## 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 ***
## dependents     0.0930299  0.0636163   1.462   0.1436
## months         0.0024388  0.0011892   2.051   0.0403 *
## 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
##
```

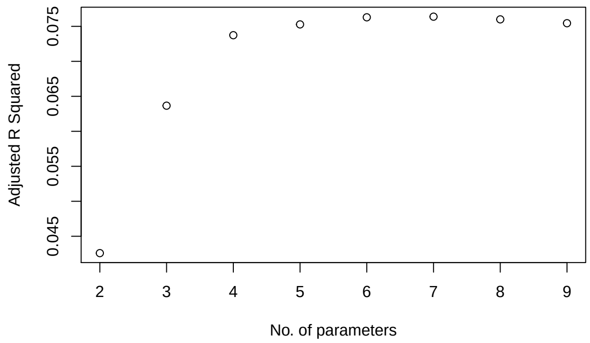
```
##
##              Theta:  0.2640
##          Std. Err.:  0.0288
##
##  2 x log-likelihood:  -1968.8080
```

Variable Selection

Adjusted R-Square Approach

```
## Subset selection object
## Call: regsubsets.formula(reports ~ owner + selfemp + majorcards + age +
##     income + expenditure + dependents + months + active, data = credit)
## 9 Variables  (and intercept)
##              Forced in Forced out
## owneryes         FALSE      FALSE
## selfempyes       FALSE      FALSE
## majorcardsyes    FALSE      FALSE
## age              FALSE      FALSE
## income           FALSE      FALSE
## expenditure      FALSE      FALSE
## dependents       FALSE      FALSE
## months           FALSE      FALSE
## active           FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          owneryes selfempyes majorcardsyes age income expenditure dependents
## 1  ( 1 ) " "      " "        " "           " " " "    " "         " "
## 2  ( 1 ) " "      " "        " "           " " " "    "*"         " "
## 3  ( 1 ) "*"      " "        " "           " " " "    "*"         " "
## 4  ( 1 ) "*"      " "        " "           " " " "    "*"         " "
## 5  ( 1 ) "*"      " "        " "           " " "*"    "*"         " "
## 6  ( 1 ) "*"      " "        " "           " " "*"    "*"         "*"
## 7  ( 1 ) "*"      " "        "*"           " " "*"    "*"         "*"
## 8  ( 1 ) "*"      " "        "*"           "*" "*"    "*"         "*"
##          months active
## 1  ( 1 ) " "    "*"
## 2  ( 1 ) " "    "*"
## 3  ( 1 ) " "    "*"
## 4  ( 1 ) "*"    "*"
## 5  ( 1 ) "*"    "*"
## 6  ( 1 ) "*"    "*"
## 7  ( 1 ) "*"    "*"
## 8  ( 1 ) "*"    "*"
```

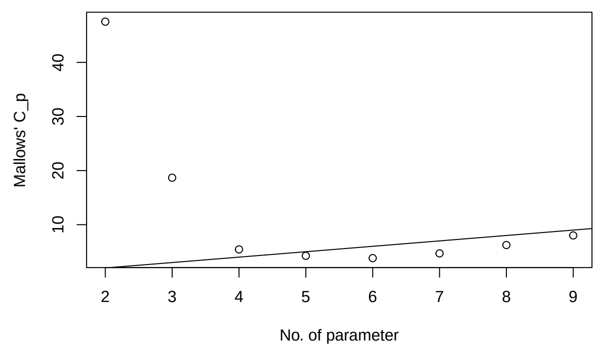
```
## [1] "selfemp"   "majorcard"
```




```
## [1] "The model with 6 predictors is the one that maximizes the adjusted R2"
```

Thus, our final model using Adjusted R^2 method would be:

```
##
## Call:
## glm.nb(formula = reports ~ owner + age + income + expenditure +
##     dependents + months + active, data = credit, init.theta = 0.2639276012,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4222  -0.6774  -0.5596  -0.3732   2.5283
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.9114389  0.2883022  -6.630 3.36e-11 ***
## owneryes    -0.8184571  0.1779509  -4.599 4.24e-06 ***
## age          0.0045943  0.0090182   0.509   0.6104
## income       0.0828643  0.0493452   1.679   0.0931 .
## expenditure -0.0023658  0.0004351  -5.437 5.41e-08 ***
## dependents   0.0937520  0.0635737   1.475   0.1403
## months       0.0024407  0.0011882   2.054   0.0400 *
## active       0.1210764  0.0114489  10.575  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2639) family taken to be 1)
##
##     Null deviance: 838.50  on 1311  degrees of freedom
## Residual deviance: 679.99  on 1304  degrees of freedom
## AIC: 1986.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2639
##           Std. Err.:  0.0288
##
##  2 x log-likelihood:  -1968.8290
```



Mallows' Cp Approach

```
## [1] "The model with 5 predictors is the one that minimizes the Mallows' C_p"
```

Thus our final model using Mallows' Cp will contain `owner`, `age`, `income`, `expenditure`, `months`, and `active`

```
##
## Call:
## glm.nb(formula = reports ~ owner + age + income + expenditure +
##     months + active, data = credit, init.theta = 0.262035756,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4263  -0.6774  -0.5625  -0.3704   2.6306
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.9089323  0.2871224  -6.648 2.96e-11 ***
## owneryes    -0.7556413  0.1732749  -4.361 1.30e-05 ***
## age          0.0059099  0.0089513   0.660   0.5091
## income       0.0970605  0.0480103   2.022   0.0432 *
## expenditure -0.0024177  0.0004382  -5.517 3.45e-08 ***
## months       0.0022920  0.0011816   1.940   0.0524 .
## active       0.1202970  0.0114524  10.504  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.262) family taken to be 1)
##
##     Null deviance: 835.50  on 1311  degrees of freedom
## Residual deviance: 679.82  on 1305  degrees of freedom
## AIC: 1987
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  0.2620
##          Std. Err.:  0.0285
##
##  2 x log-likelihood:  -1970.9640
```

NB Model Diagnostics

The Negative binomial models assume the conditional means are not equal to the conditional variances. This inequality is captured by estimating a dispersion parameter (not shown in the output) that is held constant in a Poisson model. From the values below, we conclude that the negative binomial model is more appropriate than the Poisson model.

```
chi_val <- 2 * (logLik(nb.model) - logLik(poi.model))
chi_val
```

```
## 'log Lik.' 576.2884 (df=11)
```

```
pchisq(chi_val, df = 1, lower.tail = FALSE)
```

```
## 'log Lik.' 2.406786e-127 (df=11)
```

NB Model Coefficents and CI's

```
estimates <- cbind(Estimate = coef(nb.model), confint(nb.model))
```

```
## Waiting for profiling to be done...
```

```
round(exp(estimates), 2)
```

```
##                Estimate 2.5 % 97.5 %
## (Intercept)        0.15  0.07   0.29
## owneryes           0.44  0.31   0.62
## selfempyes         1.03  0.59   1.84
## majorcardsyes      1.02  0.70   1.48
## age                1.00  0.99   1.02
## income             1.09  0.98   1.20
## expenditure        1.00  1.00   1.00
## dependents         1.10  0.97   1.25
## months             1.00  1.00   1.00
## active             1.13  1.10   1.16
```

Zero-Inflated Negative Binomial Regression

```
# A simple inflation model where all zero counts have the same probability of belonging to the zero com
nb.infl.model <- zeroinfl(reports ~ owner + selfemp + majorcards + age
                          + income + expenditure + dependents + months + active | 1, data = credit, dis
summary(nb.infl.model)
```

```
##
## Call:
## zeroinfl(formula = reports ~ owner + selfemp + majorcards + age + income +
##     expenditure + dependents + months + active | 1, data = credit, dist = "negbin")
```

```
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.5082 -0.3915 -0.3438 -0.2479 12.9350
##
## Count model coefficients (negbin with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.9226221  0.3444888  -5.581 2.39e-08 ***
## owneryes      -0.8183035  0.1751592  -4.672 2.99e-06 ***
## selfempyes     0.0315477  0.2878605   0.110   0.9127
## majorcardsyes  0.0173653  0.1926305   0.090   0.9282
## age            0.0045525  0.0093834   0.485   0.6276
## income         0.0825317  0.0506746   1.629   0.1034
## expenditure   -0.0023705  0.0003886  -6.100 1.06e-09 ***
## dependents     0.0930316  0.0648325   1.435   0.1513
## months         0.0024388  0.0012308   1.982   0.0475 *
## active         0.1208962  0.0143994   8.396  < 2e-16 ***
## Log(theta)    -1.3319950  0.1113856 -11.958  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -11.63     212.82  -0.055    0.956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.264
## Number of iterations in BFGS optimization: 51
## Log-likelihood: -984.4 on 12 Df
```

Vuong Test Among Three Models

```r
# Poisson vs. Negative Binomial
vuong(poi.model, nb.model)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   -6.482573 model2 > model1 4.5086e-11
## AIC-corrected         -6.482573 model2 > model1 4.5086e-11
## BIC-corrected         -6.482573 model2 > model1 4.5086e-11
```

```r
# Poisson vs. Zero-Inflated NB
vuong(poi.model, nb.infl.model)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
```

```
## Raw                    -6.482560 model2 > model1 4.5090e-11
## AIC-corrected          -6.460063 model2 > model1 5.2330e-11
## BIC-corrected          -6.401802 model2 > model1 7.6777e-11
```

```r
# NB vs. Zero-Inflated NB
vuong(nb.model, nb.infl.model)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A p-value
## Raw                9.651908e-02 model1 > model2 0.46155
## AIC-corrected      4.369047e+03 model1 > model2 < 2e-16
## BIC-corrected      1.568312e+04 model1 > model2 < 2e-16
```