SAI-assignment-5

Daniel Verdejo 22240224

2022-12-10

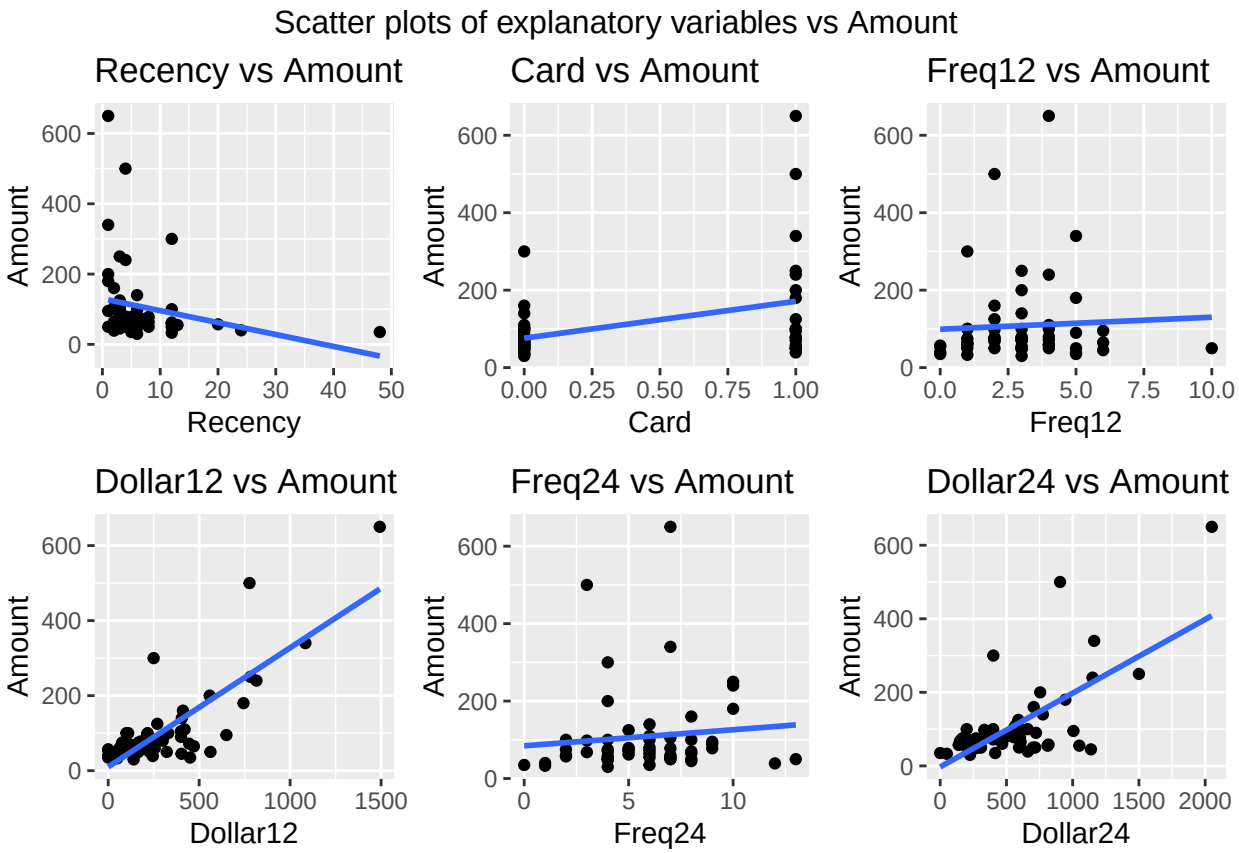
1. Question of Interest
2. Subjective Impressions or Exploratory Analysis
3. Formal Analysis
4. Conclusion and Translation

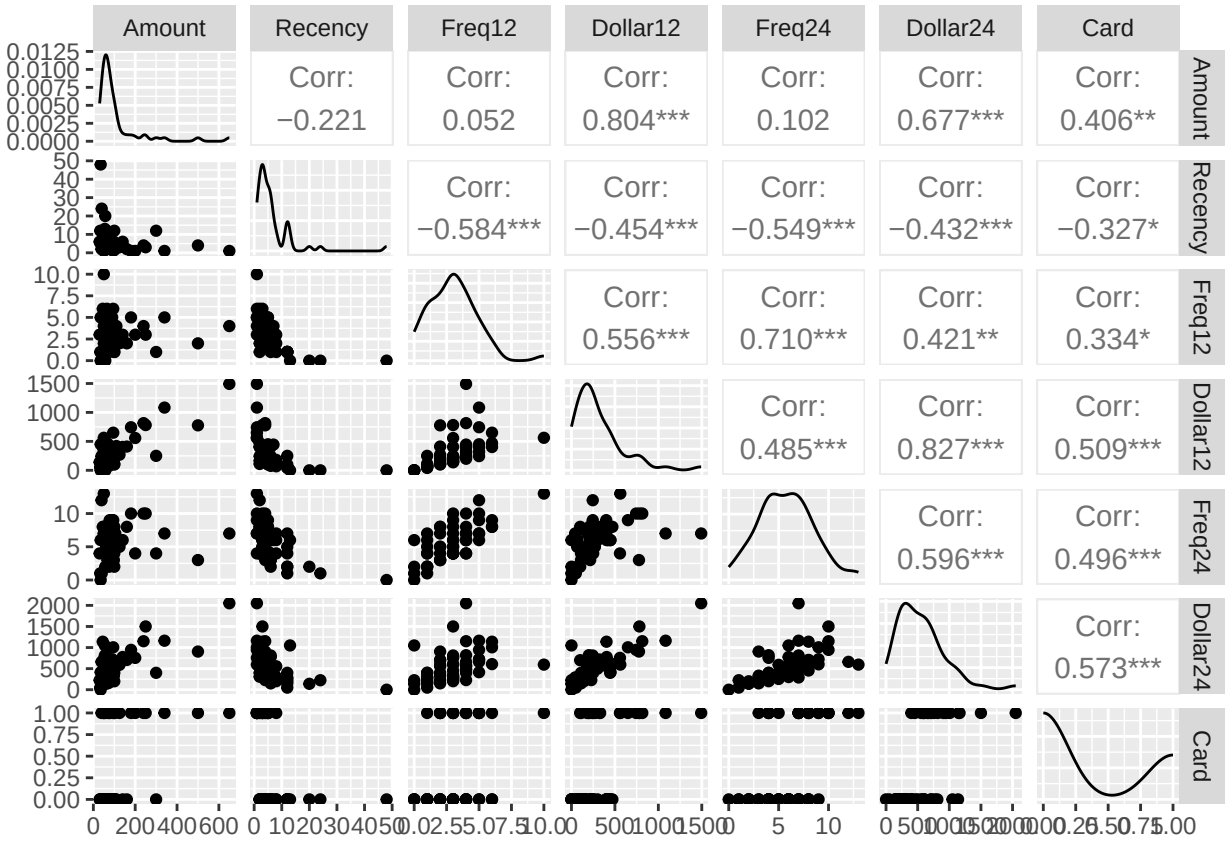
```
##   Amount Recency Freq12 Dollar12 Freq24 Dollar24 Card
## 1      0      22      0        0      3      400    0
## 2      0      30      0        0      0        0    0
## 3      0      24      0        0      1      250    0
## 4     30       6      3      140      4      225    0
## 5     33      12      1       50      1       50    0
## 6     35      48      0        0      0        0    0
```

```
## [1] "Amount"   "Recency"  "Freq12"   "Dollar12" "Freq24"   "Dollar24" "Card"
```

```
##      Amount           Recency          Freq12          Dollar12
##  Min.   :      0   Min.   : 1.000   Min.   : 0.000   Min.   :   0.0
##  1st Qu.:     50   1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 107.5
##  Median :     70   Median : 4.500   Median : 3.000   Median : 223.5
##  Mean   :  25201   Mean   : 7.217   Mean   : 2.883   Mean   : 372.0
##  3rd Qu.:    100   3rd Qu.: 8.000   3rd Qu.: 4.000   3rd Qu.: 406.5
##  Max.   :1506000   Max.   :48.000   Max.   :10.000   Max.   :5000.0
##      Freq24          Dollar24          Card
##  Min.   : 0.000   Min.   :   0.0   Min.   :0.0000
##  1st Qu.: 4.000   1st Qu.: 260.2   1st Qu.:0.0000
##  Median : 6.000   Median : 461.5   Median :0.0000
##  Mean   : 5.617   Mean   : 660.8   Mean   :0.3333
##  3rd Qu.: 7.250   3rd Qu.: 718.5   3rd Qu.:1.0000
##  Max.   :13.000   Max.   :8000.0   Max.   :1.0000
```

Lets first visualise the correlation and relationship between the explanatory variables and the target variable:


Scatter plots of explanatory variables vs Amount
Recency vs Amount
Amount
Recency
Card vs Amount
Card
Freq12 vs Amount
Freq12
Dollar12 vs Amount
Dollar12
Freq24 vs Amount
Freq24
Dollar24 vs Amount
Dollar24



As we can see from the scatter plot of explanatory variables vs the target variable Amount there are linear relationships between the explanatory variables and the target variable.

Recency has a negative linear relationship while the Dollar12, and Dollar24 have positive linear relationships with the target variable, and the rest have weak relationships with the target variable. The Dollar12 and Dollar24 Have highly positive relationships while Amount vs Card, Freq12, and Freq24 all have weak relationships: In this case, the line suggests that the explanatory variables have little to no effect on the target variable. This is backed up by the weak correlation values we see between the Freq12 / Freq24 variables and Amount. They appear to have moderate correlation to Dollar12 / Dollar24: - Freq12 and Dollar12: 0.556 - Freq24 and Dollar24: 0.596

The Dollar 12, and Dollar24 both have strong correlations to the target variable Amount, with the Dollar12 being slightly stronger. The Dollar12 and Dollar24 features also have a strong correlation between each other, which could indicate that they could both be measuring the same underlying concept. The Freq12 and Freq24 also have a strong correlation between each other, and a weak correlation to the target variable, again indicating that they could both be measuring the same underlying concept. The Recency is the only variable which has a negative correlation with all variables including the target variable.

We should first see what we can find out when we apply all explanatory variables against the multiple linear regresssion model (MLR), this should indicate where improvements can be made.

```
model <- lm(Amount ~ Recency + Freq12 + Dollar12 + Freq24 + Dollar24 + Card, data = clothing)

summary(model)
```

```
##
## Call:
## lm(formula = Amount ~ Recency + Freq12 + Dollar12 + Freq24 +
##     Dollar24 + Card, data = clothing)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.799 -12.218  -3.334   7.299 156.822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.251935  19.834341   5.256 3.20e-06 ***
## Recency      -1.345963   0.971053  -1.386    0.172
## Freq12      -32.353539   5.187870  -6.236 1.01e-07 ***
## Dollar12      0.429683   0.041325  10.398 5.43e-14 ***
## Freq24       -5.173593   3.619661  -1.429    0.159
## Dollar24      0.001756   0.031850   0.055    0.956
## Card         14.624409  14.575770   1.003    0.321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.83 on 49 degrees of freedom
## Multiple R-squared:  0.882,  Adjusted R-squared:  0.8675
## F-statistic: 61.02 on 6 and 49 DF,  p-value: < 2.2e-16
```

```
regression_table <- get_regression_table(model = model)

regression_table
```

```
## # A tibble: 7 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept  104.       19.8       5.26    0       64.4    144.
## 2 Recency     -1.35      0.971    -1.39    0.172   -3.30     0.605
## 3 Freq12     -32.4       5.19     -6.24    0      -42.8    -21.9
## 4 Dollar12     0.43      0.041    10.4     0        0.347    0.513
## 5 Freq24      -5.17      3.62     -1.43    0.159  -12.4      2.1
## 6 Dollar24     0.002     0.032     0.055   0.956   -0.062    0.066
## 7 Card        14.6      14.6       1.00    0.321  -14.7     43.9
```

Observations of the model

- Freq12 and Freq24 have extremely different coefficients, which given their strong positive association of 0.710 we may have expected them to be similar.
- Dollar12 and Dollar24 also have very differing coefficients, again given their strong positive association of 0.827 we also would have expected them to be similar.
- We can see that Dollar24, and Card have quite large p values despite them having moderate correlation values to the target variable this is somewhat contradictory of what we would expect.
- Furthermore, if we inspect the confidence intervals for these two in particular, we see that they both include zero.

If we look at the F-statistic we find that we have a large value for the F test statistic of 61.02 and a p value of 2.2e-16 which is lower than the 0.05 cutoff so its considered significant and we can reject the null hypothesis. All this is evidence to the explanatory variables and model being useful for explaining variation, so there is a contradiction from what discovered before.

The reason we are seeing these contradictions is multicollinearity. Multicollinearity is where two explanatory variables are measuring the same underlying concept or linearly related. Its Generally good practice to exclude one of the two explanatory variables that have a strong correlation between each other from the model to improve the accuracy of our predictions but more importantly reduce or eradicate the contradictions

we have just discovered. first we will experiment and see what the 3 different scenarios produce (i.e how accurate is our prediction?).

Some findings to reinforce the decision to exclude expanatory variables are as follows: - Given that Dollar12 is in the model the Dollar24 feature does not appear to add any more useful information for predicting the Amount, this is shown by Dollar24 having a high p-value of 0.956 and 95% confidence interval (-14.667, 43.916) which includes zero. - We can also see that there is a strong correlation between Dollar12, and Dollar24 of ~ 0.827 from the previous correlation plot. Therefore if Dollar12 is included in the model, then Dollar24 is not needed (Dollar12 is preferable as it has a stronger correlation to the target variable than Dollar24)(add note about overfitting). - We see a similar observation can be found for the Freq12 and Freq24 features, there is a strong correlation between the two features of ~ 0.710 again overfitting and same concept measured

```
##    Amount Recency Freq12 Dollar12 Freq24 Dollar24 Card
## 12     50       1     10      562     13      595    1
```

lets do some predictions from our fitted model for row number 12 from our dataset: The actual value is 50, the predicted amount if we calculate using the estimate from above:

On both plots we can see a positive linear correlation between the frequency and amount spent

```
paste("Predicted value: ", regression_table$estimate[1] + prod(regression_table$estimate[2:7]),
    "  vs Actual value: ", row12$Amount)
```

```
## [1] "Predicted value:  101.41823328352   vs Actual value:  50"
```

The prediction is quite poor. This is likely due to some of the points we discussed earlier. We will carry out feature subset selection techniques in order to choose the best model.

First we will do the best subset regression

```
best <- regsubsets(Amount ~ ., data = clothing)
summary(best)
```

```
## Subset selection object
## Call: regsubsets.formula(Amount ~ ., data = clothing)
## 6 Variables  (and intercept)
##          Forced in Forced out
## Recency      FALSE      FALSE
## Freq12       FALSE      FALSE
## Dollar12     FALSE      FALSE
## Freq24       FALSE      FALSE
## Dollar24     FALSE      FALSE
## Card         FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          Recency Freq12 Dollar12 Freq24 Dollar24 Card
## 1  ( 1 ) " "     " "    "*"      " "    " "      " "
## 2  ( 1 ) " "     "*"    "*"      " "    " "      " "
## 3  ( 1 ) "*"     "*"    "*"      " "    " "      " "
## 4  ( 1 ) "*"     "*"    "*"      "*"    " "      " "
## 5  ( 1 ) "*"     "*"    "*"      "*"    " "      "*"
## 6  ( 1 ) "*"     "*"    "*"      "*"    "*"      "*"
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```


```
step(model, direction = "backward")    # backward selection
```

```
## Start:  AIC=421.99
## Amount ~ Recency + Freq12 + Dollar12 + Freq24 + Dollar24 + Card
##
##            Df Sum of Sq    RSS    AIC
## - Dollar24  1         5  81712 419.99
## - Card      1      1679  83385 421.13
## <none>                   81707 421.99
## - Recency   1      3204  84910 422.14
## - Freq24    1      3407  85113 422.28
## - Freq12    1     64853 146559 452.71
## - Dollar12  1    180274 261981 485.24
##
## Step:  AIC=419.99
## Amount ~ Recency + Freq12 + Dollar12 + Freq24 + Card
##
##            Df Sum of Sq    RSS    AIC
## - Card      1      1747  83459 419.18
## <none>                   81712 419.99
## - Recency   1      3229  84940 420.16
## - Freq24    1      4332  86044 420.89
## - Freq12    1     79391 161102 456.01
## - Dollar12  1    464275 545987 524.36
##
## Step:  AIC=419.18
```

```
## Amount ~ Recency + Freq12 + Dollar12 + Freq24
##
##            Df Sum of Sq    RSS    AIC
## - Freq24    1      2987  86446 419.15
## <none>                   83459 419.18
## - Recency   1      3382  86841 419.40
## - Freq12    1     86150 169609 456.89
## - Dollar12  1    567743 651202 532.23
##
## Step:  AIC=419.15
## Amount ~ Recency + Freq12 + Dollar12
##
##            Df Sum of Sq    RSS    AIC
## - Recency   1      2270  88716 418.60
## <none>                   86446 419.15
## - Freq12    1    140610 227056 471.23
## - Dollar12  1    565766 652212 530.32
##
## Step:  AIC=418.6
## Amount ~ Freq12 + Dollar12
##
##            Df Sum of Sq    RSS    AIC
## <none>                   88716 418.60
## - Freq12    1    156406 245122 473.51
## - Dollar12  1    601685 690400 531.50
##
## Call:
## lm(formula = Amount ~ Freq12 + Dollar12, data = clothing)
##
## Coefficients:
## (Intercept)       Freq12     Dollar12
##     73.8976     -34.4259       0.4431
```

```
step(model, direction = "both")       # stepwise backwards
```

```
## Start:  AIC=421.99
## Amount ~ Recency + Freq12 + Dollar12 + Freq24 + Dollar24 + Card
##
##            Df Sum of Sq    RSS    AIC
## - Dollar24  1         5  81712 419.99
## - Card      1      1679  83385 421.13
## <none>                   81707 421.99
## - Recency   1      3204  84910 422.14
## - Freq24    1      3407  85113 422.28
## - Freq12    1     64853 146559 452.71
## - Dollar12  1    180274 261981 485.24
##
## Step:  AIC=419.99
## Amount ~ Recency + Freq12 + Dollar12 + Freq24 + Card
##
##            Df Sum of Sq    RSS    AIC
## - Card      1      1747  83459 419.18
## <none>                   81712 419.99
## - Recency   1      3229  84940 420.16
```

```
## - Freq24    1      4332  86044 420.89
## + Dollar24  1         5  81707 421.99
## - Freq12    1     79391 161102 456.01
## - Dollar12  1    464275 545987 524.36
##
## Step:  AIC=419.18
## Amount ~ Recency + Freq12 + Dollar12 + Freq24
##
##            Df Sum of Sq    RSS    AIC
## - Freq24    1      2987  86446 419.15
## <none>                    83459 419.18
## - Recency   1      3382  86841 419.40
## + Card      1      1747  81712 419.99
## + Dollar24  1        74  83385 421.13
## - Freq12    1     86150 169609 456.89
## - Dollar12  1    567743 651202 532.23
##
## Step:  AIC=419.15
## Amount ~ Recency + Freq12 + Dollar12
##
##            Df Sum of Sq    RSS    AIC
## - Recency   1      2270  88716 418.60
## <none>                    86446 419.15
## + Freq24    1      2987  83459 419.18
## + Dollar24  1       519  85927 420.81
## + Card      1       402  86044 420.89
## - Freq12    1    140610 227056 471.23
## - Dollar12  1    565766 652212 530.32
##
## Step:  AIC=418.6
## Amount ~ Freq12 + Dollar12
##
##            Df Sum of Sq    RSS    AIC
## <none>                    88716 418.60
## + Recency   1      2270  86446 419.15
## + Freq24    1      1874  86841 419.40
## + Card      1       623  88093 420.20
## + Dollar24  1       207  88508 420.47
## - Freq12    1    156406 245122 473.51
## - Dollar12  1    601685 690400 531.50
##
## Call:
## lm(formula = Amount ~ Freq12 + Dollar12, data = clothing)
##
## Coefficients:
## (Intercept)       Freq12     Dollar12
##     73.8976     -34.4259       0.4431
```

```r
min.model <- lm(Amount ~ 1, data = clothing) # smallest model to consider

step(min.model, direction = "forward",   # forward selection
     scope = list(lower = ~ 1,
                  upper = ~ Recency + Freq12 + Dollar12 + Freq24 + Dollar24 + Card))
```

```
## Start:  AIC=529.65
## Amount ~ 1
##
##            Df Sum of Sq    RSS    AIC
## + Dollar12  1    447122 245122 473.51
## + Dollar24  1    317577 374666 497.27
## + Card      1    113863 578380 521.59
## + Recency   1     33753 658490 528.85
## <none>                  692243 529.65
## + Freq24    1      7162 685081 531.07
## + Freq12    1      1843 690400 531.50
##
## Step:  AIC=473.51
## Amount ~ Dollar12
##
##            Df Sum of Sq    RSS    AIC
## + Freq12    1    156406  88716 418.60
## + Freq24    1     75087 170035 455.03
## + Recency   1     18066 227056 471.23
## <none>                  245122 473.51
## + Dollar24  1       333 244789 475.44
## + Card      1        10 245112 475.51
##
## Step:  AIC=418.6
## Amount ~ Dollar12 + Freq12
##
##            Df Sum of Sq   RSS    AIC
## <none>                  88716 418.60
## + Recency   1   2269.52 86446 419.15
## + Freq24    1   1874.47 86841 419.40
## + Card      1    622.68 88093 420.20
## + Dollar24  1    207.39 88508 420.47
```

```
##
## Call:
## lm(formula = Amount ~ Dollar12 + Freq12, data = clothing)
##
## Coefficients:
## (Intercept)     Dollar12       Freq12
##     73.8976       0.4431     -34.4259
```

```
step(min.model, direction = "both",    # stepwise forward
     scope = list(lower = ~ 1,
                  upper = ~ Recency + Freq12 + Dollar12 + Freq24 + Dollar24 + Card))
```

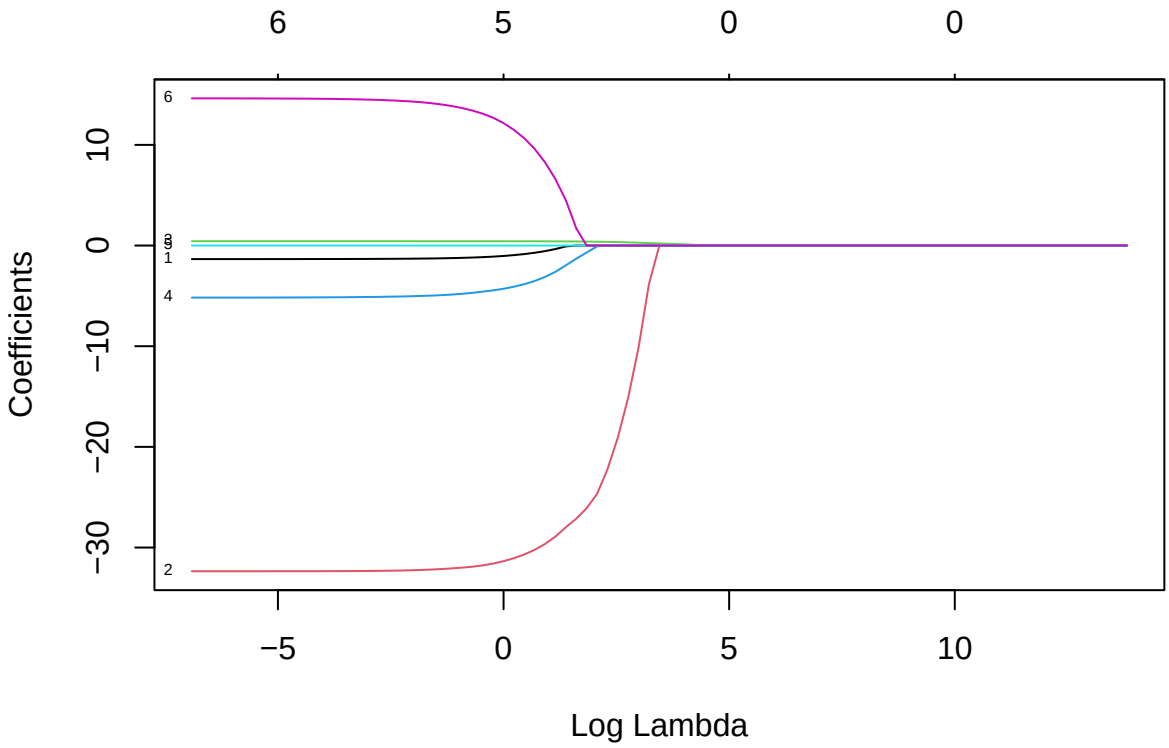
```
## Start:  AIC=529.65
## Amount ~ 1
##
##            Df Sum of Sq    RSS    AIC
## + Dollar12  1    447122 245122 473.51
## + Dollar24  1    317577 374666 497.27
## + Card      1    113863 578380 521.59
## + Recency   1     33753 658490 528.85
## <none>                  692243 529.65
## + Freq24    1      7162 685081 531.07
```

```
## + Freq12    1      1843 690400 531.50
##
## Step:  AIC=473.51
## Amount ~ Dollar12
##
##            Df Sum of Sq    RSS    AIC
## + Freq12    1    156406  88716 418.60
## + Freq24    1     75087 170035 455.03
## + Recency   1     18066 227056 471.23
## <none>                  245122 473.51
## + Dollar24  1       333 244789 475.44
## + Card      1        10 245112 475.51
## - Dollar12  1    447122 692243 529.65
##
## Step:  AIC=418.6
## Amount ~ Dollar12 + Freq12
##
##            Df Sum of Sq    RSS    AIC
## <none>                   88716 418.60
## + Recency   1      2270  86446 419.15
## + Freq24    1      1874  86841 419.40
## + Card      1       623  88093 420.20
## + Dollar24  1       207  88508 420.47
## - Freq12    1    156406 245122 473.51
## - Dollar12  1    601685 690400 531.50
```

```
##
## Call:
## lm(formula = Amount ~ Dollar12 + Freq12, data = clothing)
##
## Coefficients:
## (Intercept)     Dollar12       Freq12
##     73.8976       0.4431     -34.4259
```

```r
# Fit the model using glmnet() with the lasso method
lasso <- glmnet(x = as.matrix(clothing[,-1]), y = clothing$Amount, alpha = 1, lambda = 10^seq(-3, 6, 0.

plot(lasso, xvar = "lambda", label = TRUE)
```



From the above model selection process we can see a clear indication that the lowest AIC of **418.6** consistently uses the model `Amount ~ Dollar12 + Freq12` in all selection processes `backward, stepwise backward, forward, stepwise forward`. We also see similiar information from the best regression and exhaustive

Parsimony - It is desirable to find the simplest model required for the application, which is captured by the concept of parsimony: "If two competing models have statistically the same predictive ability then the parsimonious model is the one with the smaller number of parameters - Similar to the concept of Occam's Razor:"the simplest explanation is usually the right one"

Boot strapping 10000 samples

```
#set the seed and then run the bootstrap samples
set.seed(22240224)
modelFn <- function(data, i) {
  m <- lm(Amount ~ Freq12 +
            Dollar12 +
            Freq12 * Dollar12 +
            Card, data = data[i, ])

  return(coef(m))
}

results <- boot(clothing, modelFn, R = 10000)

results
```

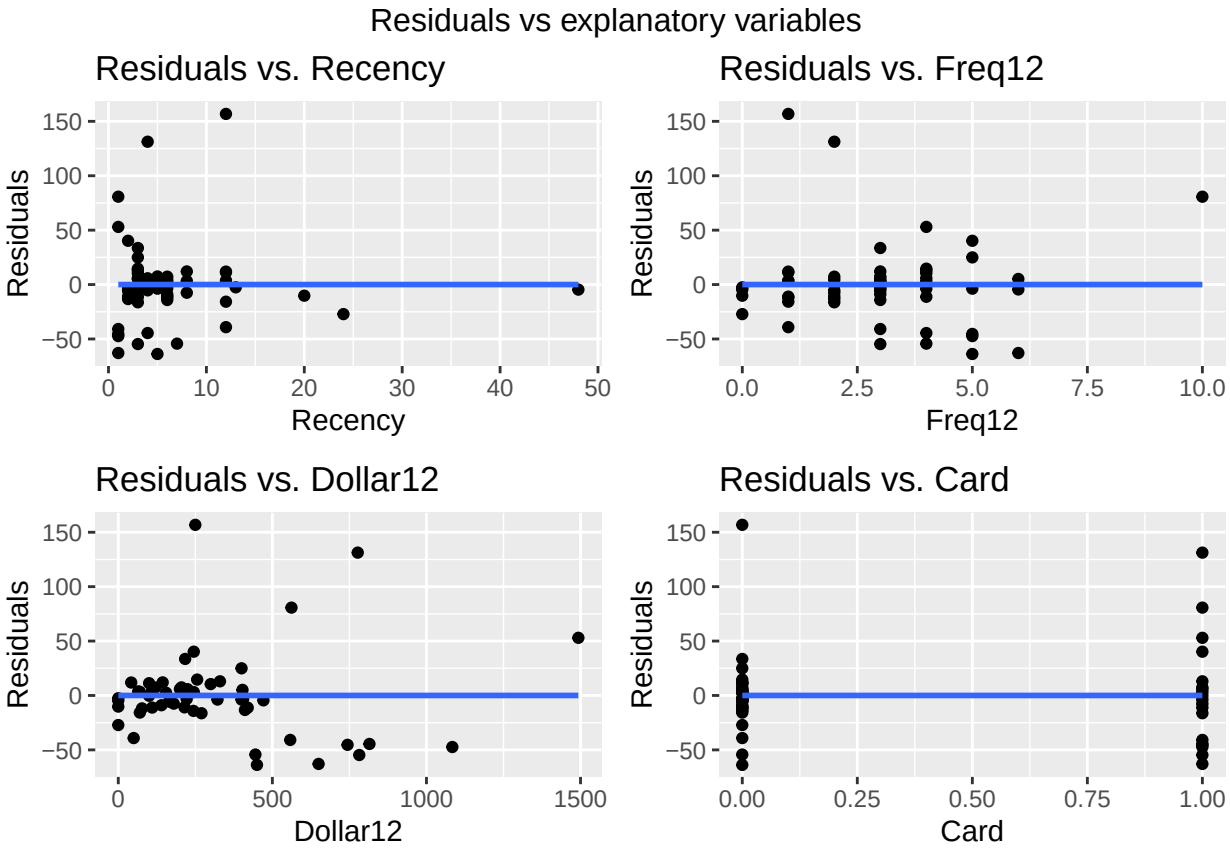
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
```

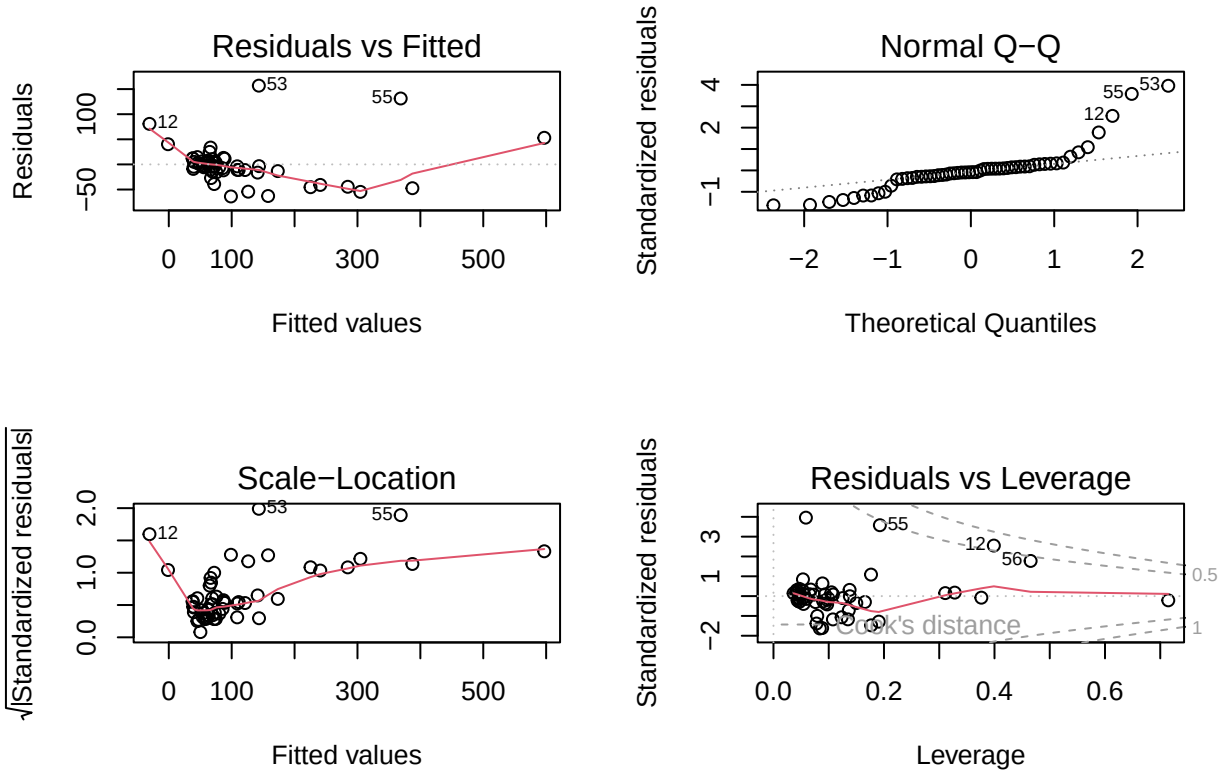
```
##
##
## Call:
## boot(data = clothing, statistic = modelFn, R = 10000)
##
##
## Bootstrap Statistics :
##         original       bias    std. error
## t1*  56.51758985 -0.079598991 11.11779494
## t2* -27.71103941  0.551596246  4.48545672
## t3*   0.51385290  0.016612582  0.11458607
## t4*   8.17801165 -1.420797121 11.21121501
## t5*  -0.02252985 -0.006427376  0.02416982
```

```
conf_intervals <- boot.ci(boot.out = results, type =  c("norm", "basic", "perc", "bca"))

conf_intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = c("norm", "basic", "perc",
##     "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   (34.81, 78.39 )   (31.67, 75.46 )
##
## Level     Percentile            BCa
## 95%   (37.57, 81.37 )   (38.87, 83.25 )
## Calculations and Intervals on Original Scale
```


Residuals vs explanatory variables
Residuals vs. Recency
Residuals
Recency
Residuals vs. Freq12
Freq12
Residuals vs. Dollar12
Dollar12
Residuals vs. Card
Card



1. Using a subset selection procedure on the variables above, construct a model that best predict the Amount of money spent by a customer. Summarise the key features of the model fit and performance, and check the underlying assumptions.

When evaluating the fit and performance of a statistical model, there are several key features that you should summarize and assess. These include: 1. The model's overall fit to the data, which can be assessed using measures such as the R-squared value, the adjusted R-squared value, and the residual standard error. 2. The statistical significance of the model's coefficients, which can be assessed using hypothesis tests or confidence intervals. 3. The individual effects of the predictor variables on the response variable, which can be assessed using measures such as t-tests, p-values, and confidence intervals. 4. The model's predictive accuracy, which can be assessed using measures such as the mean squared error (MSE), the root mean squared error (RMSE), and the mean absolute error (MAE).

Additionally, when building and evaluating a statistical model, it is important to check the underlying assumptions of the model. These assumptions include:

1. Linearity: The relationship between the predictor and response variables is linear.
2. Normality: The residuals are normally distributed.
3. Independence: The observations are independent of each other.
4. Equal variances: The variances of the residuals are equal for all values of the predictor variable.
5. Outlier detection: The data does not contain any extreme values or outliers.

By summarizing the key features of the model fit and performance, and checking the underlying assumptions, you can gain a better understanding of the model's strengths and weaknesses, and make informed decisions about how to improve it.

Assumptions Underlying the Model: LINE 1.Population relationship between the mean response and the features is linear (Linearity assumption); 2.Sample is representative of the population and the subjects are independent (Independence assumption); 3.Errors follow a normal distribution, centred about the regression

line (Normality assumption); 4.Variance of the errors is the same for any value of the explanatory variable (Equal Spreads assumption).