CANADIAN HOUSEHOLD DEBT DETERMINANTS

Harpreet Kaur Shoker, Krishna Rajan

Part 1 - Introduction:

Canada is currently experiencing record-breaking levels of household debt. Consumer spending is central to the Canadian economy and therefore to financial stability. However, with the household debt ratio reaching 163% there is a growing concern that households are overextended. Household debt reporting often mentions low interest rates and rising real estate prices as the main drivers. I want to build a model that could accurately predict household debt levels for different types of households across regions, taking into account a variety of both quantitative and qualitative factors.

Part 2 - Data:

2.1 - Defination

Before proceeding, Here are few definition for "household". According Statistics Canada, a household "refers to a person or group of persons who occupy the same dwelling and do not have a usual place of residence elsewhere in Canada or abroad."
Moreover, debt is defined as "An amount of money borrowed by one party from another". More specific assumptions which will be detailed below.

2.2 - Data Sources

Our data collection process started with researching Statscan and other online resources for survey results related to household finances. The dataset was not available directly through Statistics Canada's online page, however, I was able to obtain the data via the University of Toronto online Database.

I found The 'Survey of Financial Security 2005', a comprehensive, Canada-wide survey that uses 5,276 households to represent the approximately 12.5 million household population of Canada at the time. Moreover, each household contained 86 predictor ("x"variables).

Part 3 - Exploratory data analysis:

3.1 - Data Cleanup

In order to cleanup the data, we are converting all the values into numeric fields and removing rows that contains 0 in wdtotal column.

```
financial_survey_converted = subset(financial_survey_raw, subset=(wdtotal != 0))
financial_survey_converted <- sapply(financial_survey_converted, as.numeric)
financial_survey_converted = as.data.frame(financial_survey_converted)
```

3.2 - Descriptive Statistics

There are 5 region included in the survey. There is a factor named "region" included in the financial_survey dataframe.

wdTotal	Canada
1	Atlantic
2	Quebec
3	Ontario
4	Prairies
5	British

– Need to add details regarding all the regions. The Ontario dataset contains 480 sample points (once filtering complete). The mean debt level is 161,122.18 with a range from 2,300 to 950,000. The dataset is skewed towards the right with a value of 2.10. This indicates there are some extreme debt values on the right side contributing to skewness. Because there are extreme values to the right, median debt (131,500) is less than the mean household debt (161,122). Mode debt value of 100,000 indicates there are quite a few number of households with debt level of 100,000, which is less than both mean and medium.

Standard deviation value is $120,164 which indicate some variability in the household debt levels. The coefficient of variance is 0.74, which mean there is good amount variability within dataset. This variability could be attributed to multiple household related factors.

3.2 - Linear Regression Analysis

With our cleaned up and simplified data set, we used regression analysis to construct three different model.

3.2.1 Select top 10 high correlation predictor.

```
cors <- sapply(financial_survey_converted, cor, y=financial_survey_converted$wdtotal)
mask <- (rank(-abs(cors)) <= 10 )
best10.pred <- financial_survey_converted[, mask]

best10.pred <- subset(best10.pred, select = c(-wdtotal, -wnetwpt, -wnetwpg) )
summary(best10.pred)
```

```
##     ecfexhmr         watotpt            watotpg            waprval
##  Min.   :     0   Min.   :     175   Min.   :     175   Min.   :      0
##  1st Qu.:  6000   1st Qu.:   97275   1st Qu.:   99225   1st Qu.:      0
##  Median : 10000   Median :  301500   Median :  305000   Median : 125000
##  Mean   : 13413   Mean   :  654468   Mean   :  658896   Mean   : 211485
##  3rd Qu.: 16500   3rd Qu.:  644238   3rd Qu.:  649625   3rd Qu.: 260000
##  Max.   :155000   Max.   :34927030   Max.   :34942030   Max.   :3800000
##     wastrest           wdprmor           wdstomor
##  Min.   :       0   Min.   :      0   Min.   :      0
##  1st Qu.:       0   1st Qu.:      0   1st Qu.:      0
##  Median :       0   Median :      0   Median :      0
##  Mean   :   72790   Mean   :  57417   Mean   :  15306
##  3rd Qu.:       0   3rd Qu.:  87500   3rd Qu.:      0
##  Max.   :11250000   Max.   :1450000   Max.   :1550000
```

3.2.2 Stepwise backword regression.

```
full.model.best10 <- lm (wdtotal ~ ecfexhmr + watotpt + watotpg + waprval + wastrest + wdprmor + wdstomo

reduced.model.best10<- step (full.model.best10, direction = "backward")
```

```
## Start:  AIC=76296.18
## wdtotal ~ ecfexhmr + watotpt + watotpg + waprval + wastrest +
##     wdprmor + wdstomor
##
##            Df  Sum of Sq        RSS   AIC
## - watotpt   1 8.9716e+06 5.8984e+12 76294
## - watotpg   1 1.0495e+07 5.8984e+12 76294
## <none>                   5.8984e+12 76296
## - ecfexhmr  1 4.1601e+09 5.9026e+12 76297
## - wastrest  1 6.1258e+10 5.9597e+12 76331
## - waprval   1 9.4037e+10 5.9925e+12 76351
## - wdprmor   1 1.8965e+13 2.4864e+13 81466
## - wdstomor  1 1.9244e+13 2.5142e+13 81506
##
## Step:  AIC=76294.18
## wdtotal ~ ecfexhmr + watotpg + waprval + wastrest + wdprmor +
##     wdstomor
##
##            Df  Sum of Sq        RSS   AIC
## - watotpg   1 1.4117e+08 5.8986e+12 76292
## <none>                   5.8984e+12 76294
## - ecfexhmr  1 4.2124e+09 5.9027e+12 76295
## - wastrest  1 6.1254e+10 5.9597e+12 76329
## - waprval   1 9.4028e+10 5.9925e+12 76349
## - wdprmor   1 1.8966e+13 2.4865e+13 81465
## - wdstomor  1 1.9244e+13 2.5142e+13 81504
##
## Step:  AIC=76292.27
## wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor
##
##            Df  Sum of Sq        RSS   AIC
## <none>                   5.8986e+12 76292
## - ecfexhmr  1 4.2474e+09 5.9028e+12 76293
## - wastrest  1 7.9589e+10 5.9782e+12 76338
## - waprval   1 1.5484e+11 6.0534e+12 76383
## - wdprmor   1 1.9079e+13 2.4977e+13 81479
## - wdstomor  1 2.0060e+13 2.5959e+13 81617
```

```
summary(reduced.model.best10)
```

```
##
## Call:
## lm(formula = wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor +
##     wdstomor, data = financial_survey_converted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124511  -17363   -9283    4850  465622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.681e+04  9.741e+02  17.261  < 2e-16 ***
## ecfexhmr    1.164e-01  7.241e-02   1.608    0.108
## waprval     2.622e-02  2.701e-03   9.706  < 2e-16 ***
## wastrest    1.226e-02  1.762e-03   6.959 4.06e-12 ***
```

```
## wdprmor     9.745e-01  9.045e-03 107.742  < 2e-16 ***
## wdstomor    1.011e+00  9.150e-03 110.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40540 on 3589 degrees of freedom
## Multiple R-squared:   0.93,  Adjusted R-squared:  0.9299
## F-statistic:  9535 on 5 and 3589 DF,  p-value: < 2.2e-16
```

3.2.3 Stepwise forward regression.

```
min.model.best10 <- lm (wdtotal ~ 1, data=financial_survey_converted)
forward.model.best10 <- step(min.model.best10, direction="forward", scope = ( ~ ecfexhmr + watotpt + wa
```

```
## Start:  AIC=85841.75
## wdtotal ~ 1
##
##            Df  Sum of Sq        RSS   AIC
## + wdprmor   1 4.8087e+13 3.6165e+13 82803
## + wdstomor  1 3.9459e+13 4.4794e+13 83573
## + ecfexhmr  1 2.7578e+13 5.6675e+13 84418
## + waprval   1 2.4808e+13 5.9444e+13 84590
## + watotpg   1 2.1871e+13 6.2381e+13 84763
## + watotpt   1 2.1826e+13 6.2426e+13 84766
## + wastrest  1 1.3460e+13 7.0793e+13 85218
## <none>                   8.4252e+13 85842
##
## Step:  AIC=82803.37
## wdtotal ~ wdprmor
##
##            Df  Sum of Sq        RSS   AIC
## + wdstomor  1 2.9911e+13 6.2537e+12 76496
## + wastrest  1 9.5211e+12 2.6644e+13 81707
## + watotpt   1 7.7510e+12 2.8414e+13 81938
## + watotpg   1 7.7506e+12 2.8414e+13 81938
## + waprval   1 2.7507e+12 3.3414e+13 82521
## + ecfexhmr  1 9.5371e+11 3.5211e+13 82709
## <none>                   3.6165e+13 82803
##
## Step:  AIC=76496.43
## wdtotal ~ wdprmor + wdstomor
##
##            Df  Sum of Sq        RSS   AIC
## + waprval   1 2.7171e+11 5.9820e+12 76339
## + watotpg   1 1.9595e+11 6.0577e+12 76384
## + watotpt   1 1.9584e+11 6.0579e+12 76384
## + wastrest  1 1.5869e+11 6.0950e+12 76406
## + ecfexhmr  1 5.3193e+10 6.2005e+12 76468
## <none>                   6.2537e+12 76496
##
## Step:  AIC=76338.74
## wdtotal ~ wdprmor + wdstomor + waprval
##
##            Df  Sum of Sq        RSS   AIC
```

```
## + wastrest  1 7.9154e+10 5.9028e+12 76293
## + watotpg   1 1.8704e+10 5.9633e+12 76329
## + watotpt   1 1.8687e+10 5.9633e+12 76329
## + ecfexhmr  1 3.8121e+09 5.9782e+12 76338
## <none>                   5.9820e+12 76339
##
## Step:  AIC=76292.86
## wdtotal ~ wdprmor + wdstomor + waprval + wastrest
##
##            Df  Sum of Sq        RSS   AIC
## + ecfexhmr  1 4247375794 5.8986e+12 76292
## <none>                   5.9028e+12 76293
## + watotpg   1  176099651 5.9027e+12 76295
## + watotpt   1  171756174 5.9027e+12 76295
##
## Step:  AIC=76292.27
## wdtotal ~ wdprmor + wdstomor + waprval + wastrest + ecfexhmr
##
##           Df Sum of Sq        RSS   AIC
## <none>                 5.8986e+12 76292
## + watotpg  1 141167925 5.8984e+12 76294
## + watotpt  1 139644344 5.8984e+12 76294
```

```
summary(forward.model.best10)
```

```
##
## Call:
## lm(formula = wdtotal ~ wdprmor + wdstomor + waprval + wastrest +
##     ecfexhmr, data = financial_survey_converted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124511  -17363   -9283    4850  465622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.681e+04  9.741e+02  17.261  < 2e-16 ***
## wdprmor     9.745e-01  9.045e-03 107.742  < 2e-16 ***
## wdstomor    1.011e+00  9.150e-03 110.479  < 2e-16 ***
## waprval     2.622e-02  2.701e-03   9.706  < 2e-16 ***
## wastrest    1.226e-02  1.762e-03   6.959 4.06e-12 ***
## ecfexhmr    1.164e-01  7.241e-02   1.608    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40540 on 3589 degrees of freedom
## Multiple R-squared:   0.93,  Adjusted R-squared:  0.9299
## F-statistic:  9535 on 5 and 3589 DF,  p-value: < 2.2e-16
```

3.2.4 Selection based on model and understanding of the data.

After analyzing the data,we identified following independent variables out of the 85 parameters could be used to generate our model.Following variables were kept:

Response Variable	Household Debt [wdtotal]
Quantitative Factor Variable	Family income after taxes [atinc27]
Quantitative Factor Variable	Number of family members [fmsz27]
Quantitative Factor Variable	Number of credit cards [dvfcrn]
Quantitative Factor Variable	Student debt [wdsloan]
Quantitative Factor Variable	Child related expense [ecfexchr]
Quantitative Factor Variable	Subtotal-credit card & instalment debt [wdstcred]
Quantitative Factor Variable	Assets (continue basis) [watotpt]
Quantitative Factor Variable	Mortgage on principle residence [wdprmor]
Quantitative Factor Variable	Mortgage on other residence [wdstomor]
Quantitative Factor Variable	Line of credit [wdstloc]
Quantitative Factor Variable	Net worth including pension [wnetwpg]
Quantitative Factor Variable	Age of major income earner [ecpage]
Qualitative Factor Variable	Number of earners in the family [nbear27]
Qualitative Factor Variable	Region [region]
Qualitative Factor Variable	Level of education [dvphlv2g]
Qualitative Factor Variable	Sex of major income earner [hcsex_r]

Once we generated the model we found from the above only below variables were statistically significant.

```
 reg_multi <- lm(wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc , data=financial_survey_converted)
summary(reg_multi)
```

```
##
## Call:
## lm(formula = wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc,
##     data = financial_survey_converted)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -192214  -24383  -13626   -2549 1497387
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.105e+03  3.198e+03   2.535 0.011302 *
## wdprmor     1.107e+00  1.525e-02  72.549  < 2e-16 ***
## nbear27     6.754e+03  1.800e+03   3.752 0.000178 ***
## wdsloan     5.788e-01  2.173e-01   2.664 0.007755 **
## wdstloc     1.307e+00  4.840e-02  26.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 90900 on 3590 degrees of freedom
## Multiple R-squared:  0.6479, Adjusted R-squared:  0.6475
## F-statistic:  1652 on 4 and 3590 DF,  p-value: < 2.2e-16
```

Part4 Model comparision

4.1 Comparing Models by using ANOVA

Comparing the both forward and backword step model using anova returns a p value of 0 , indicating that both models are same.

```
anova(reduced.model.best10, forward.model.best10)
```

```
## Analysis of Variance Table
##
## Model 1: wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor
## Model 2: wdtotal ~ wdprmor + wdstomor + waprval + wastrest + ecfexhmr
##   Res.Df        RSS Df Sum of Sq F Pr(>F)
## 1   3589 5.8986e+12
## 2   3589 5.8986e+12  0         0
```

Comparing Step model with custom model using anova retuns a p value less than 0.05, indicating that both models are different.

```
anova(reduced.model.best10, reg_multi)
```

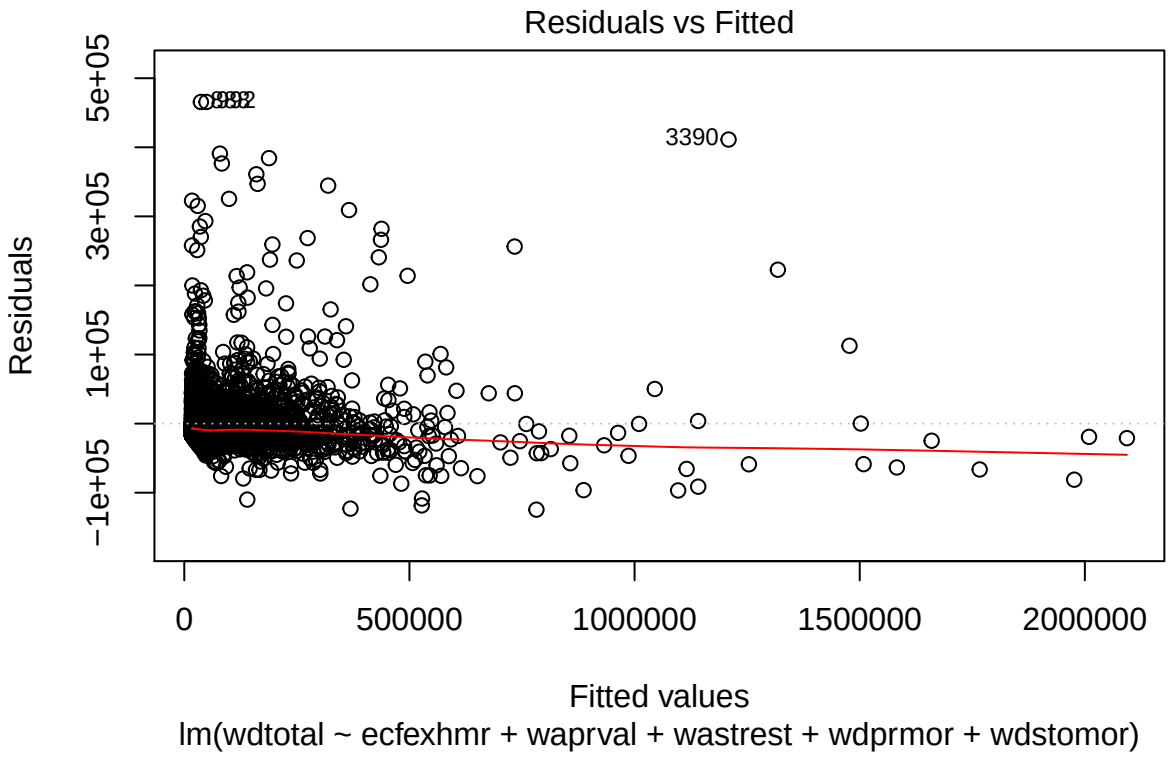
```
## Analysis of Variance Table
##
## Model 1: wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor
## Model 2: wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc
##   Res.Df        RSS Df   Sum of Sq     F    Pr(>F)
## 1   3589 5.8986e+12
## 2   3590 2.9664e+13 -1 -2.3766e+13 14460 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

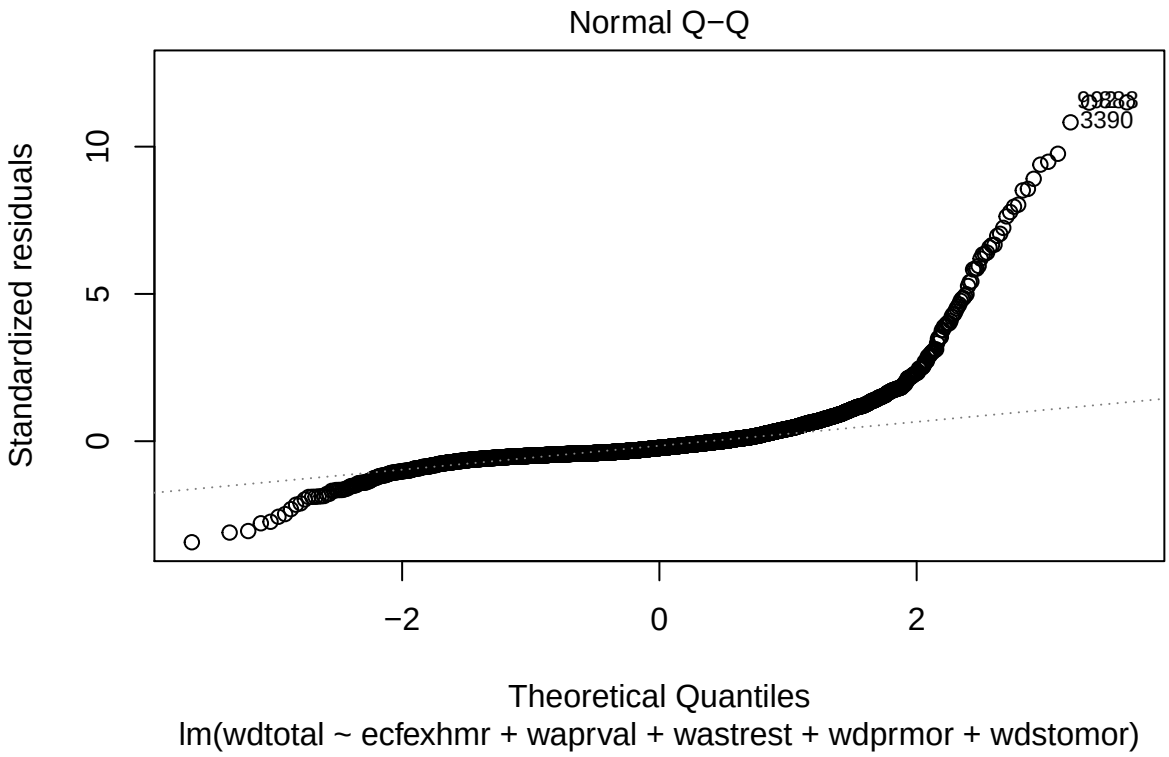
4.2 Diagnosing different linear models

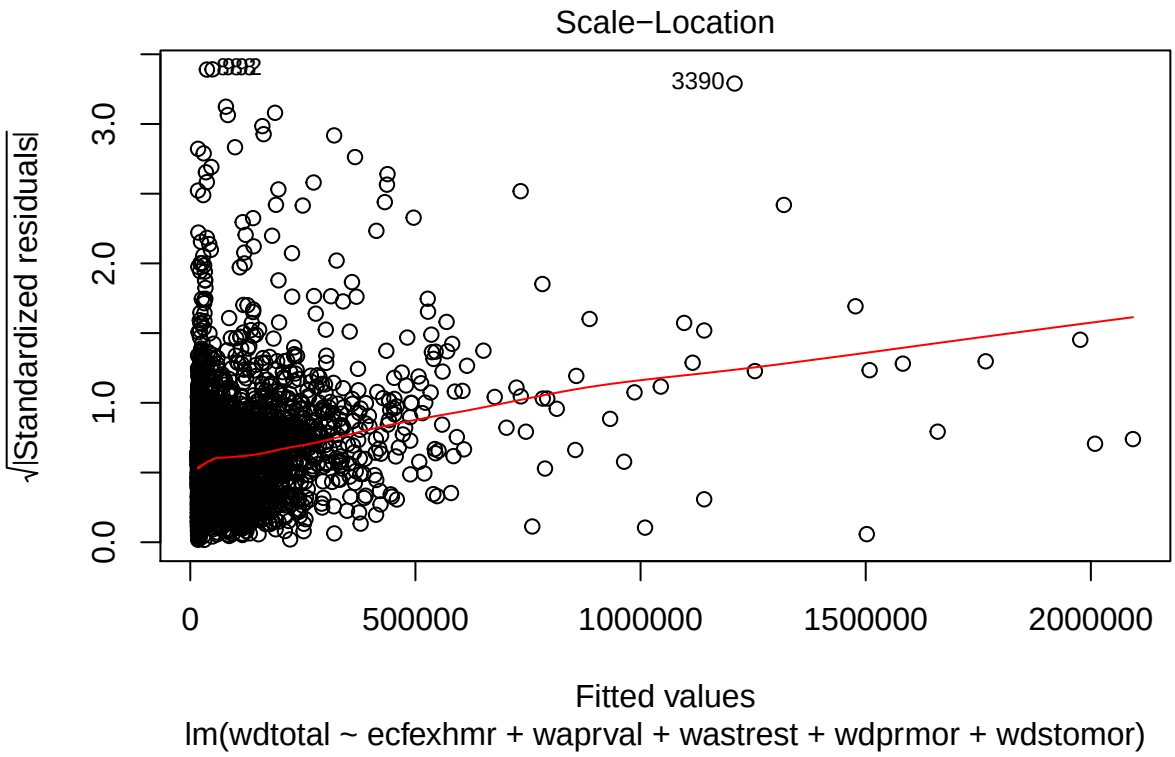
Diagostic plots for the both the models are show below. Based on the plots we can say these are pretty good regression models. The points in the Residuals Vs Fitted Plot are radomly scattered with no pattern for both model. The points in the Normal Q-Q plot are more or less on the line, indicating that residuals follow a normal distribution. In both Scale-Location plot and Residual Vs Leverage plots, the points are in the a group with none too far from center.

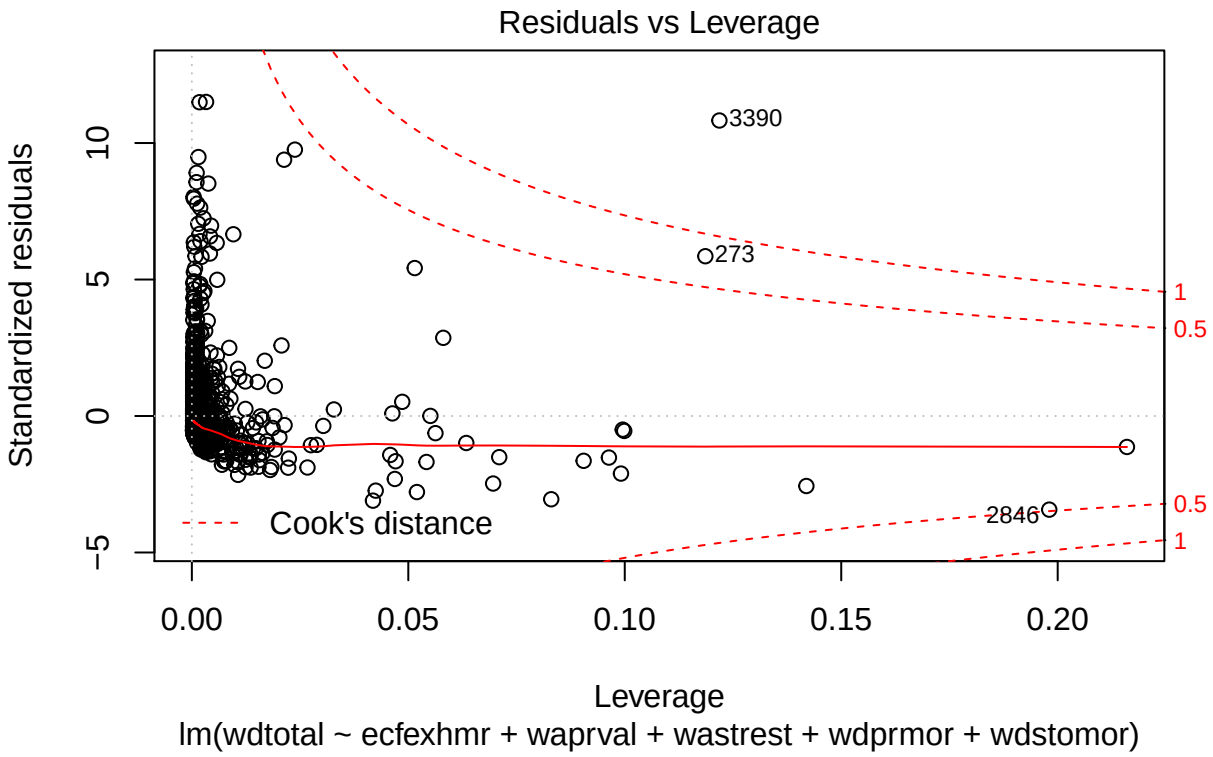
4.2.1 Step wise linear regression.

```
plot(reduced.model.best10)
```


Residuals vs Fitted
Residuals
5e+05
3e+05
1e+05
−1e+05
3390
0
500000
1000000
1500000
2000000
Fitted values
lm(wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor)

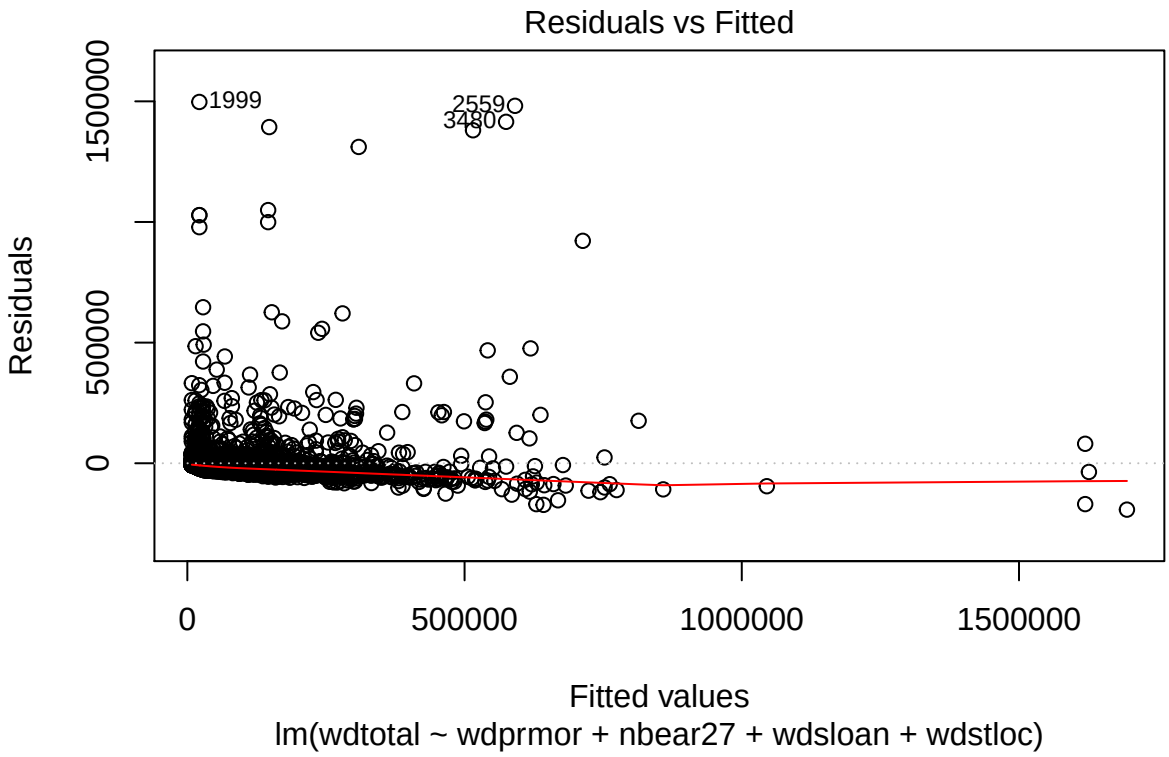

Normal Q−Q
Standardized residuals
0
5
10
−2
0
2
3390
Theoretical Quantiles
lm(wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor)

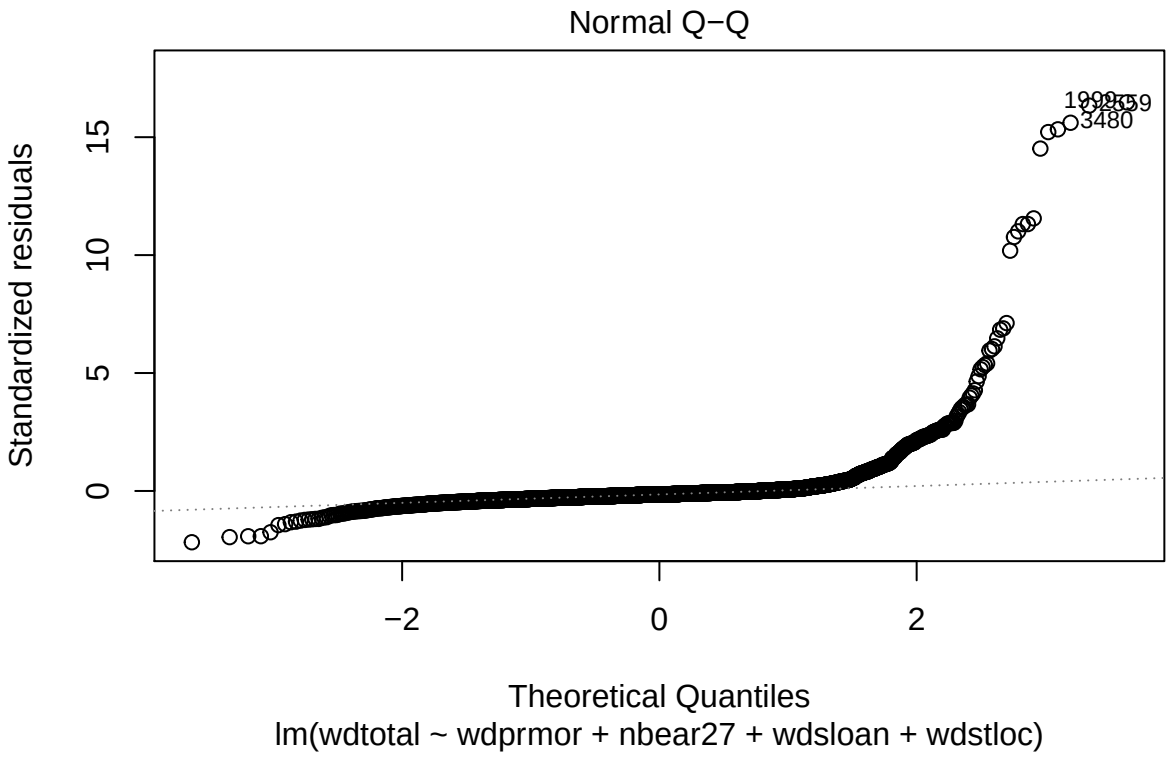

Scale−Location
3390
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
500000
1000000
1500000
2000000
Fitted values
lm(wdtotal ~ ecfexhmr + waprval + wastrest + wdprmor + wdstomor)

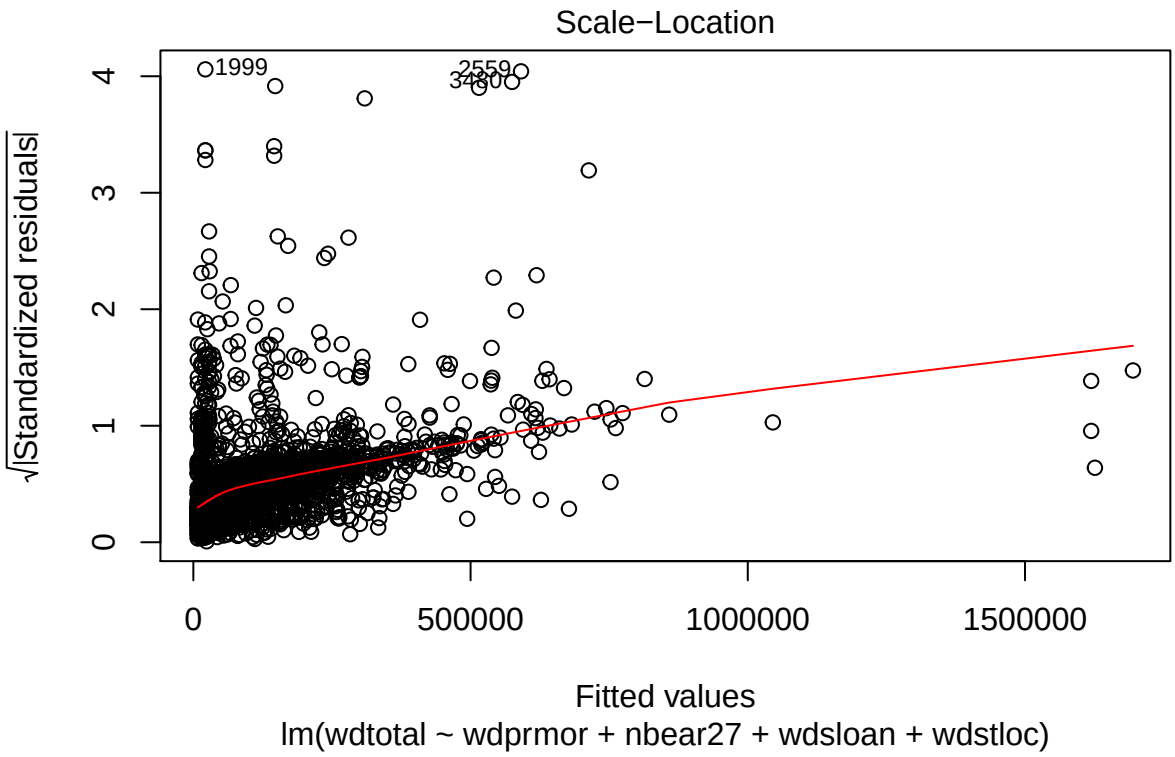


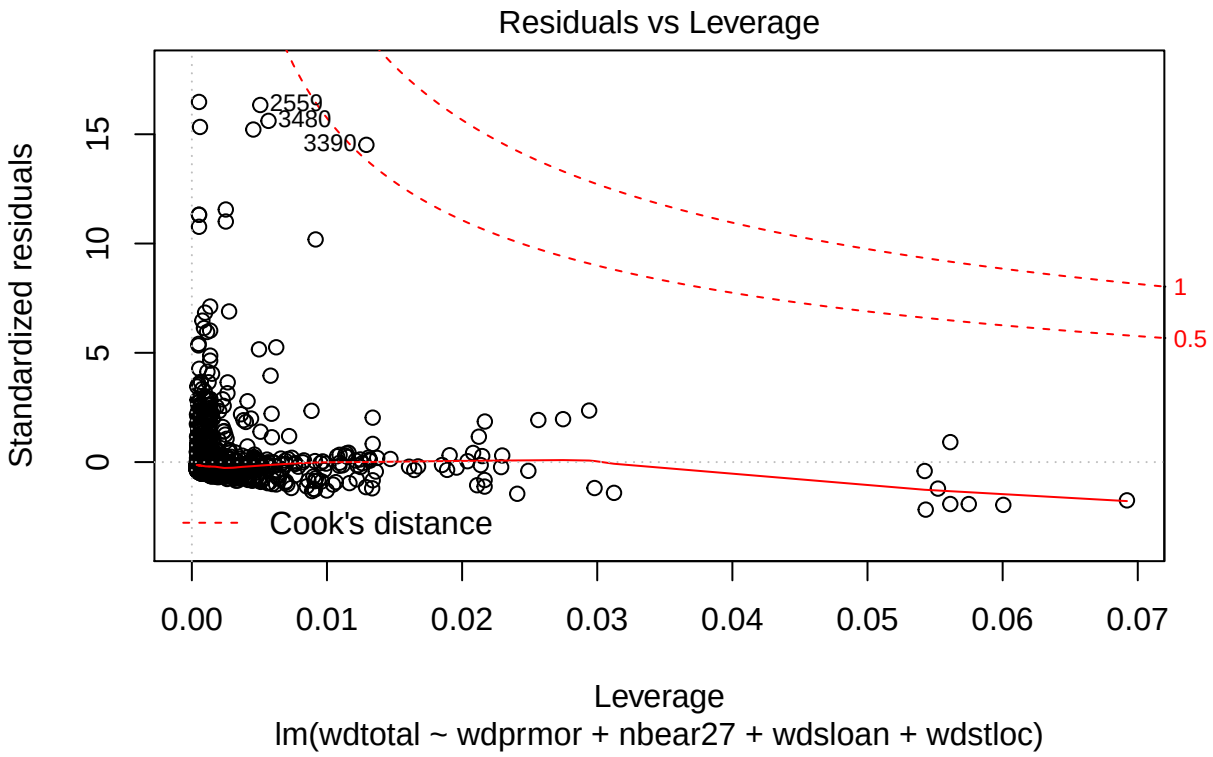
4.2.2 Custom Model linear regression.

```
plot(reg_multi)
```


Residuals vs Fitted
Residuals
1500000
500000
0
1999
2559
3480
0
500000
1000000
1500000
Fitted values
lm(wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc)


Normal Q−Q
Standardized residuals
0
5
10
15
−2
0
2
Theoretical Quantiles
lm(wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc)


Scale−Location
1999
2559
3480
√|Standardized residuals|
0
1
2
3
4
0
500000
1000000
1500000
Fitted values
lm(wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc)


Residuals vs Leverage
Standardized residuals
15
10
5
0
2559
3480
3390
1
0.5
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Leverage
lm(wdtotal ~ wdprmor + nbear27 + wdsloan + wdstloc)