MSDS660_Week6_Assignment_Apeetz

2022-11-13

Adam Peetz

MSDS660 Week 6 Assignment

Regis University

Dr. Siripun Sanguansintukul

November 27th 2022

```
# load libraries
library(tidyverse)
library(data.table)
library(dplyr)
library(car)
library(corrplot)
library(MASS)
library('fastDummies')

#heatmap and custom colors
#install.packages("reshape2")
library(reshape2)
#install.packages("viridis")
library("viridis")

# load data
data <- read_csv("marketing.csv",show_col_types = FALSE)
# convert data to table
df<-as.data.table(data)
```

Multiple Linear Regression on Marketing Data

Hypothesis: The total number of purchases made by a customer is significantly impacted by demographic variables such as their education, geographic location, and marital status.

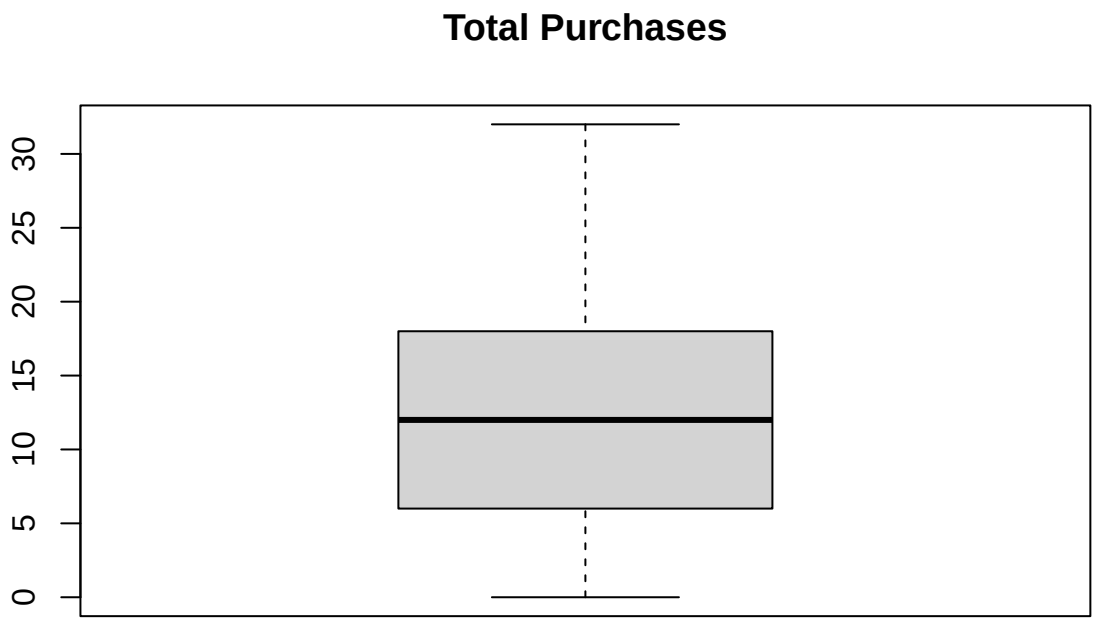
Data: The data used in this notebook is marketing data provided by the Hult International School of Business.

Total Purchases Variable:

Total purchases can be created by adding together the number of purchases made in stores, on the web, and through the company's catalog. The newly created total purchases variable has an almost normal looking distribution with a spike in customers who have made fewer purchases. A boxplot shows that there are no outliers in the total purchases variable.

```r
#create totalpsum variable
df$totalpsum <- df$NumStorePurchases +
                df$NumWebPurchases +
                df$NumCatalogPurchases

# histogram total sales variable
hist(df$totalpsum,
     main="Purchases at Store",
     xlab="Total Purchases",
     ylab="Customers",
     col="blue",
     breaks=100)
```


```r
# boxplot total sales variable
boxplot(df$totalpsum, main="Total Purchases")
```



Cleaning the data set

Removing NAs from Income

Several customers have not supplied an Income in the data set. Rows for these customers will be removed from the data set. Income also needs to have the $ sign removed so it can be treated as a numerical variable.

```
#remove NA from column
df<- df[-which(is.na(df$Income)), ]

#remove $ signs
df$Income <- parse_number(df$Income)
```

Education Feature Engineering

2nd Cycle refers to people who have completed their 2nd cycle of college. This is the same education level as Masters degree. Rows for 2n Cycle and Master will have labels applied to combine them in the dataset.

```
# feature engineer education
# combine 2nd Cycle and Masters
df<-df %>% mutate(Education_1 = case_when(Education == "PhD" ~ "Doctors",
                                          Education == "Master" ~ "Masters",
                                          Education == "2n Cycle" ~ "Masters",
                                          Education == "Graduation" ~ "Bachelors",
                                          Education == "Basic" ~ "High School") %>%
           fct_relevel("High School",
```

```
                        "Bachelors",
                        "Masters",
                        "Doctors"))
```

Feature Selection

All variables will be kept and tested except for those that identify unique customer details and the variables for purchase counts. Unique customer details like customer ID will not meaningfully correlate to the purchase count. Variables representing purchase counts from different sources were used to construct the dependent variable and have too much influence on that variable to be left in the model.

```
# subset dataframe
df_1 <- df %>% dplyr::select(totalpsum, Education_1, Income, Kidhome, MntWines, MntFruits, MntMeatProdu
```

Dummies Variables

Education, country, and marital status are all categorical features that contain text labels. These labels must be converted into a sparse binary matrix prior to analysis. This operation can be performed with the dummies function which applies one hot encoding to the categorical columns in the dataset.

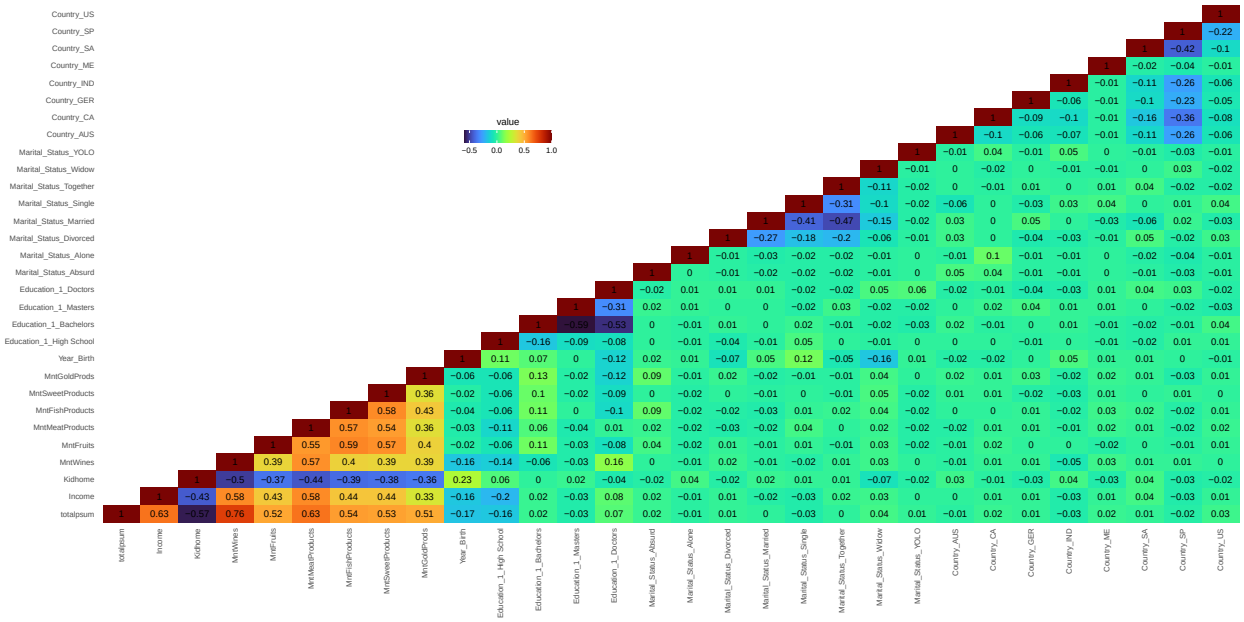
```
# One hot encoding categorical variables
dum_df_1 <- dummy_cols(df_1,
                        select_columns=c('Education_1','Marital_Status', 'Country'),
                        remove_selected_columns = TRUE)
```

Correlation Plot

A correlation heatmap can be created to check for multicollinearity in the dataset. Any features that have correlation coefficients above 0.8 or below -0.8 have multicollinearity issues and one should be selected and removed from the model. The heatmap below does not reveal any collinearity issues in the selected features.

```
# correlation plot
# define lower triangle function
get_lower_tri<-function(cormat){
    cormat[lower.tri(cormat)] <- NA
    return(cormat)}

# define upper triangle function
get_upper_tri <- function(cormat){
    cormat[upper.tri(cormat)]<- NA
    return(cormat)}

# translate dataframe to correlation dataframe
cormap <- round(cor(dum_df_1),2)

# get lower triangle
tri <- get_lower_tri(cormap)

# melt the corrleation dataframe
melted_cormap <- melt(tri, na.rm=TRUE)

# apply gg plotting function
ggplot(data = melted_cormap, aes(x=Var2, y=Var1, fill=value)) +
        geom_tile() +
```

```
        geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
        scale_fill_viridis(discrete = FALSE, option="H") +
        theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.justification = c(1, 0),
        legend.position = c(0.4, 0.7),
        legend.direction = "horizontal")+
        guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                     title.position = "top", title.hjust = 0.5))
```



Model: Fit_1

A linear regression model can now be created for the features in the data set. Features for -Marital_Status_YOLO -Country_US and -Education_1_Doctors were shown to be multicollinear by the regression model and have been removed.

Summary

Most of the demographic features created for education, marital status, and country do not pass the test for significance. The block of features that pass the test for significance are the amounts spent on different product categories. There are a few demographic features that are considered significant, such as being from the country Spain, which reduces purchase count by -1.05, or having a high school education which reduces purchase count by -1.896.

AIC Score: This model achieves and AIC score of 12231

Plots:

Residuals vs Fitted: The residual plots show an increase in variance as purchase counts increase.

QQ Plots: The QQ Plot is approximately normal. There are a few outlying datapoints on either end of the model that sway the distribution.

Outliers: Several outlying datapoints are revealed in the diagnostic plots of the model. These rows were investigated and found to contain examples of extreme outliers or perhaps even forgery and will be removed from the dataset

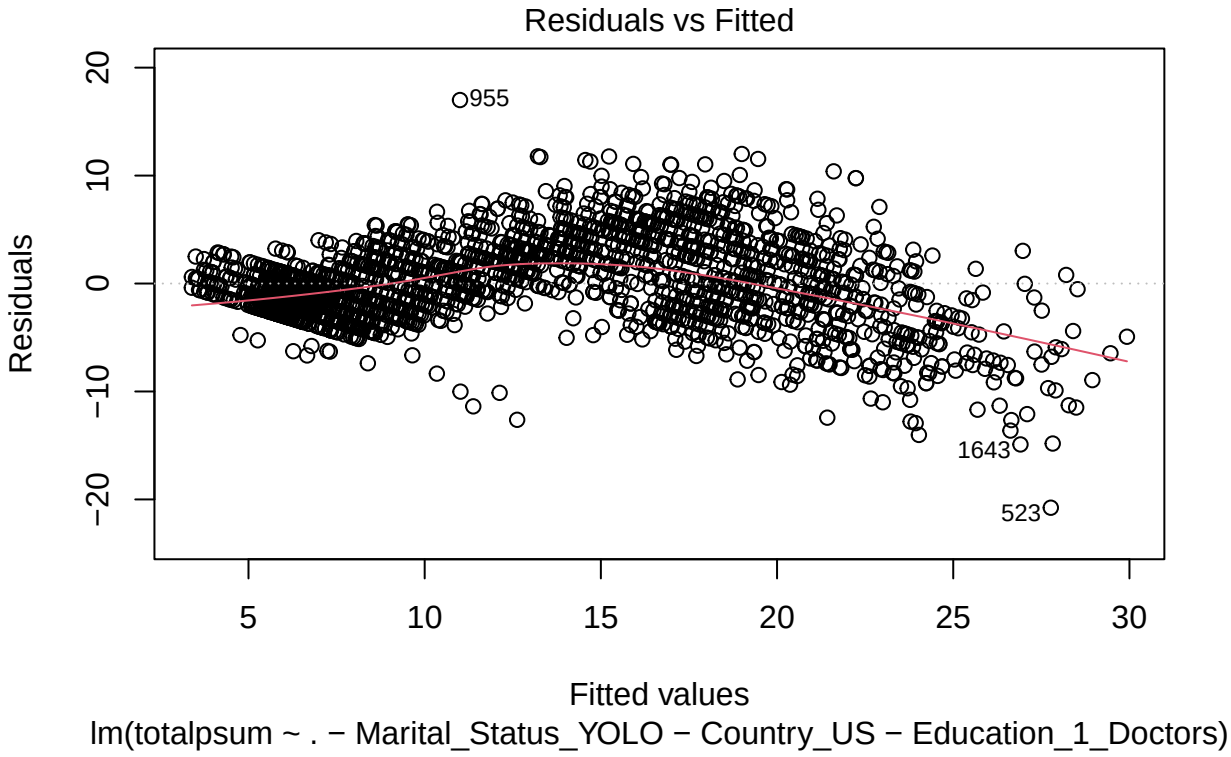
- Row 523, income is 666,666 - removing because this doesn't seem like a real salary.
- Row 955, income is 2447, and is an outlier for meat purchases (1750) and catalog purchases (28). This customer spends more than half of their total income on catalog meat purchases. They are an oddity and will be removed.

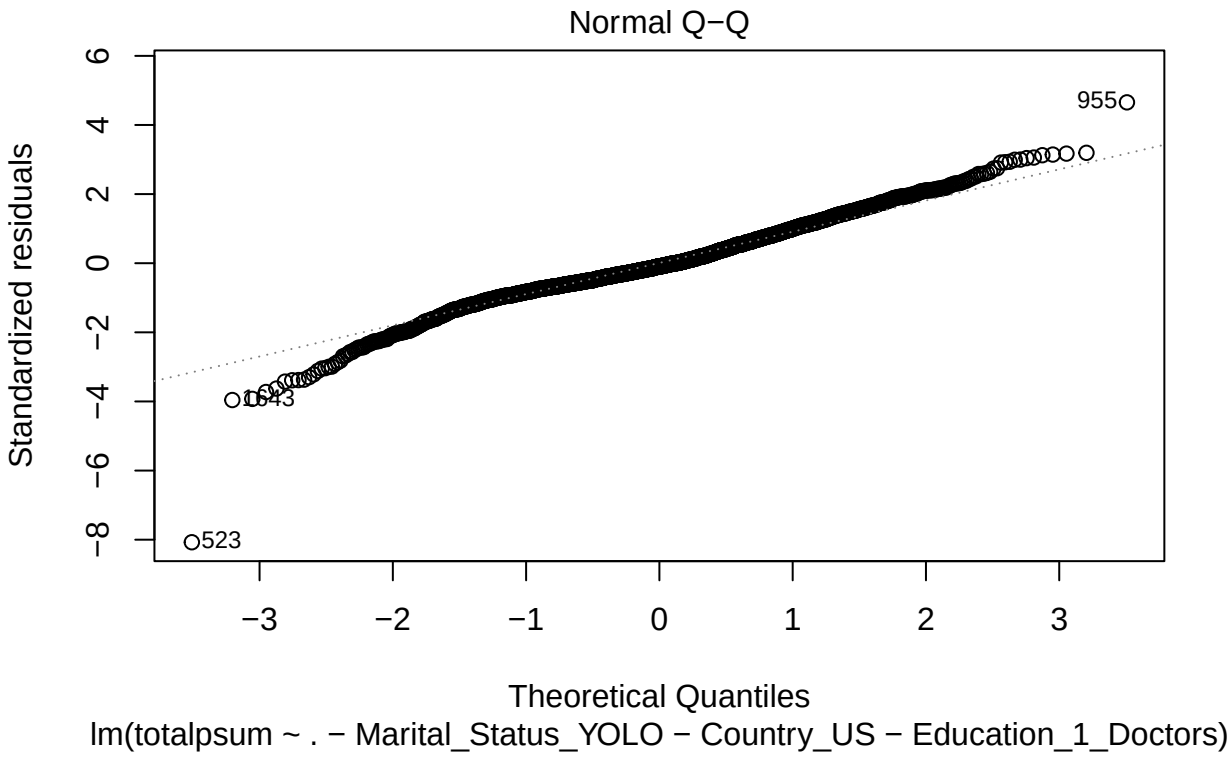
```
# initial MLR w/o stepwise AIC
# fit model
fit_1 <- lm(totalpsum ~ . -Marital_Status_YOLO -Country_US -Education_1_Doctors, data = dum_df_1)

# show model statistics
summary(fit_1)
```

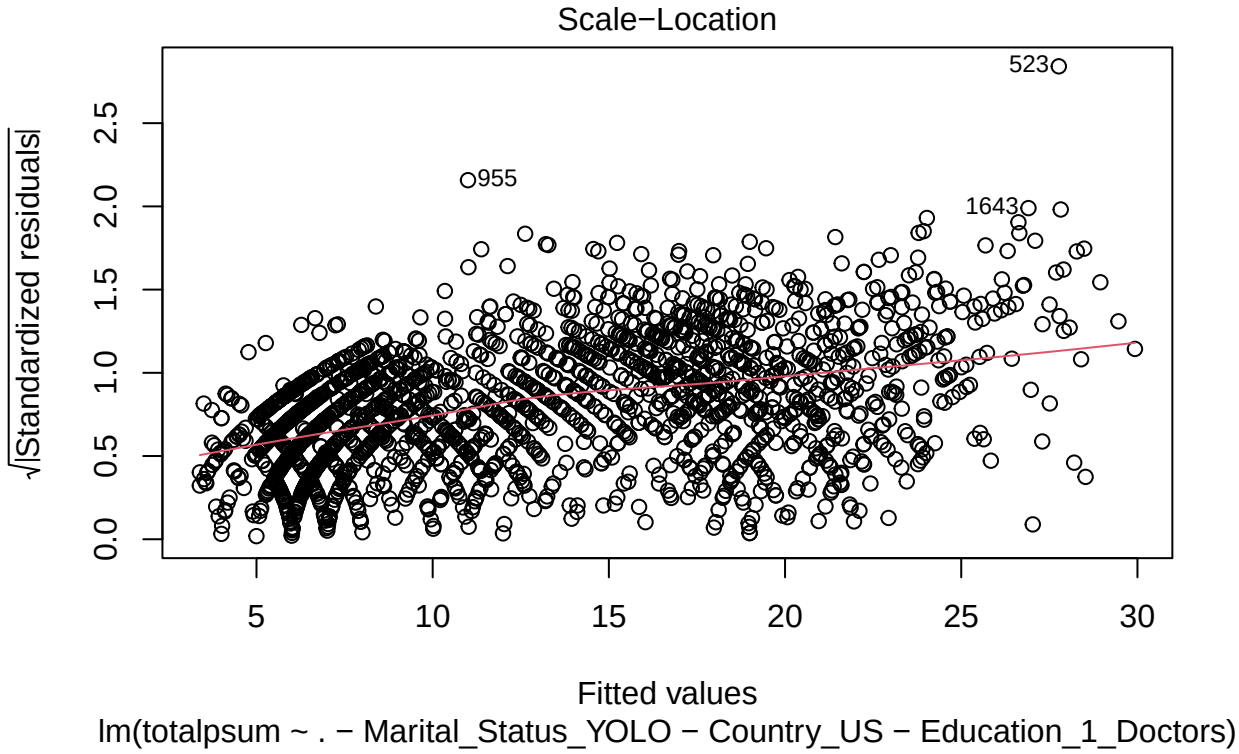
```
##
## Call:
## lm(formula = totalpsum ~ . - Marital_Status_YOLO - Country_US -
##     Education_1_Doctors, data = dum_df_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.7685  -2.2715  -0.3533   2.3360  16.9973
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.897e+01  1.455e+01   2.678 0.007455 **
## Income                      3.400e-05  4.428e-06   7.678 2.43e-14 ***
## Kidhome                    -1.776e+00  1.882e-01  -9.434  < 2e-16 ***
## MntWines                    8.920e-03  3.376e-04  26.419  < 2e-16 ***
## MntFruits                   9.969e-03  2.808e-03   3.550 0.000393 ***
## MntMeatProducts             3.074e-03  5.398e-04   5.695 1.40e-08 ***
## MntFishProducts             8.108e-03  2.116e-03   3.832 0.000131 ***
## MntSweetProducts            1.730e-02  2.688e-03   6.438 1.48e-10 ***
## MntGoldProds                1.971e-02  1.853e-03  10.641  < 2e-16 ***
## Year_Birth                 -1.512e-02  7.237e-03  -2.089 0.036796 *
## `Education_1_High School` -1.896e+00  5.674e-01  -3.341 0.000848 ***
## Education_1_Bachelors      -2.885e-01  2.191e-01  -1.317 0.187963
## Education_1_Masters        -2.557e-01  2.425e-01  -1.054 0.291811
## Marital_Status_Absurd      -4.109e+00  3.829e+00  -1.073 0.283351
## Marital_Status_Alone       -6.234e-01  3.480e+00  -0.179 0.857830
## Marital_Status_Divorced    -1.689e+00  2.711e+00  -0.623 0.533481
## Marital_Status_Married     -1.428e+00  2.702e+00  -0.528 0.597296
## Marital_Status_Single      -1.828e+00  2.705e+00  -0.676 0.499333
## Marital_Status_Together    -1.661e+00  2.705e+00  -0.614 0.539253
## Marital_Status_Widow       -1.748e+00  2.736e+00  -0.639 0.523019
## Country_AUS                -8.885e-01  4.849e-01  -1.832 0.067054 .
```

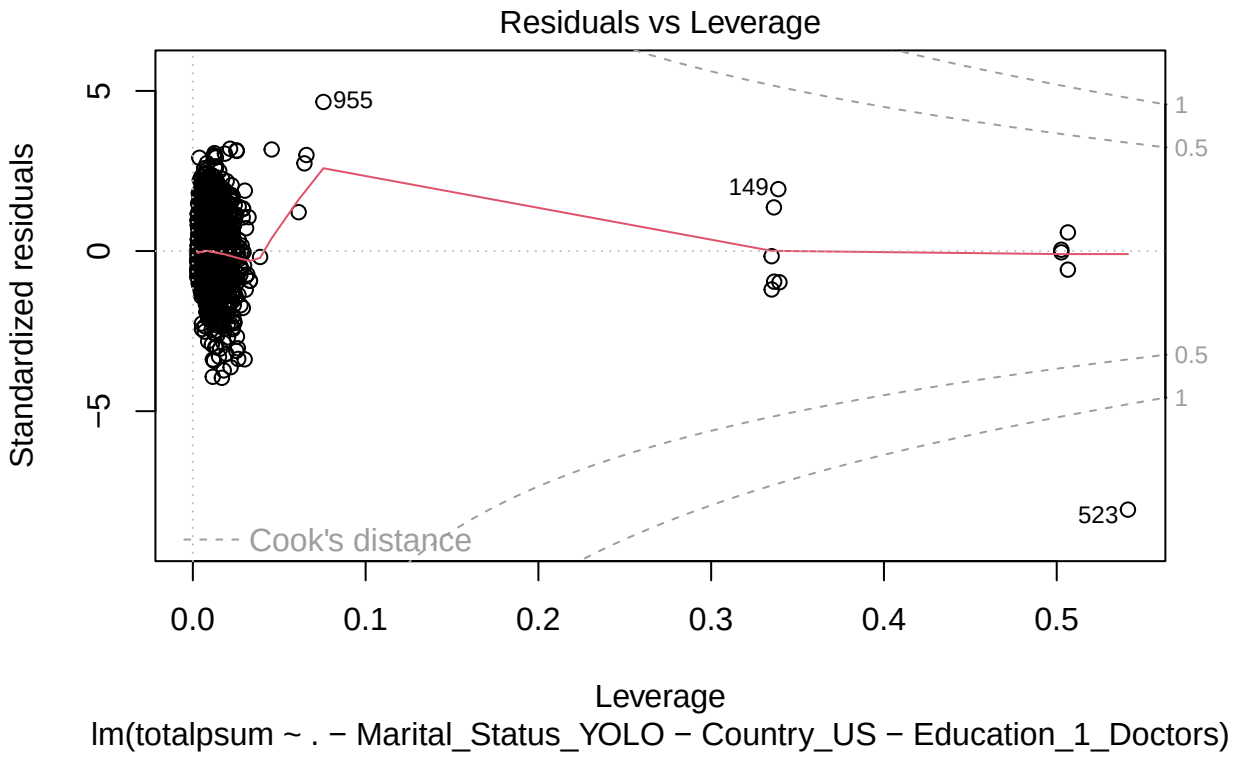
```
## Country_CA                -8.606e-01  4.367e-01  -1.971 0.048886 *
## Country_GER               -9.817e-01  5.115e-01  -1.919 0.055084 .
## Country_IND               -6.401e-01  4.845e-01  -1.321 0.186576
## Country_ME                -5.847e-01  2.231e+00  -0.262 0.793248
## Country_SA                -8.401e-01  4.230e-01  -1.986 0.047179 *
## Country_SP                -1.050e+00  3.860e-01  -2.719 0.006600 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.798 on 2189 degrees of freedom
## Multiple R-squared:  0.7254, Adjusted R-squared:  0.7221
## F-statistic: 222.4 on 26 and 2189 DF,  p-value: < 2.2e-16
```

```
plot(fit_1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(totalpsum ~ . − Marital_Status_YOLO − Country_US − Education_1_Doctors)
955
643
523
6
4
2
0
−2
−4
−6
−8
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
5
10
15
20
25
30
523
955
1643
Fitted values
lm(totalpsum ~ . − Marital_Status_YOLO − Country_US − Education_1_Doctors)




```
AIC(fit_1)
```

```
## [1] 12231.61
```

Fit_1 VIF

Variance-Inflation-Factor (VIF) is a check for multicollinearity issues in the model. The best possible VIF score is 1, which indicates the absence of multicollinearity. VIF values that exceed 5 are problematic and should be addressed in the model.

Dummy features created for Marital_Status have high VIF scores, with "Married", "Single", "Divorced", and "Together", scoring in the 100's. These features should be removed from the model due to high collinearity.

```
# initial MLR w/o stepwise AIC VIF
vif(fit_1)
```

```
##                   Income                  Kidhome                 MntWines
##                 1.908268                 1.568622                 1.992020
##                MntFruits          MntMeatProducts          MntFishProducts
##                 1.917795                 2.250755                 2.060753
##         MntSweetProducts             MntGoldProds               Year_Birth
##                 1.871645                 1.415292                 1.155390
## `Education_1_High School`     Education_1_Bachelors       Education_1_Masters
##                 1.176031                 1.843153                 1.716754
##     Marital_Status_Absurd      Marital_Status_Alone   Marital_Status_Divorced
##                 2.031023                 2.515217               105.882018
##    Marital_Status_Married     Marital_Status_Single   Marital_Status_Together
```

```
##                266.103587                188.148424                215.460574
##      Marital_Status_Widow               Country_AUS               Country_CA
##                38.084366                  2.237691                  3.094931
##               Country_GER               Country_IND               Country_ME
##                 1.994125                  2.233679                  1.033679
##               Country_SA               Country_SP
##                 3.545559                  5.722573
```

Perform feature selection via stepwise AIC

An optimum combination of features for regression can be selected using a stepwise process that tests different combinations until an optimum set is found. This operation is performed for the fitted model below. Stepwise selection should solve the multicollinearity issues revealed by the VIF.

The outlier rows identified in the diagnostic plots will also be removed here and the original model refitted without those outliers.

```
# remove outlier rows
dum_df_1 <- dum_df_1[-c(955, 523), ]

# refit due to removed rows.
fit_1 <- lm(totalpsum ~ . -Marital_Status_YOLO -Country_US -Education_1_Doctors, data = dum_df_1)

# stepwise AIC
stepAIC(fit_1, direction="both")
```

```
## Start:  AIC=5837.97
## totalpsum ~ (Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Education_1_Masters +
##     Education_1_Doctors + Marital_Status_Absurd + Marital_Status_Alone +
##     Marital_Status_Divorced + Marital_Status_Married + Marital_Status_Single +
##     Marital_Status_Together + Marital_Status_Widow + Marital_Status_YOLO +
##     Country_AUS + Country_CA + Country_GER + Country_IND + Country_ME +
##     Country_SA + Country_SP + Country_US) - Marital_Status_YOLO -
##     Country_US - Education_1_Doctors
##
##                             Df Sum of Sq   RSS    AIC
## - Country_ME                 1       0.0 30182 5836.0
## - Marital_Status_Alone       1       0.7 30183 5836.0
## - Marital_Status_Married     1       4.1 30186 5836.3
## - Marital_Status_Together    1       5.1 30187 5836.3
## - Marital_Status_Widow       1       5.5 30188 5836.4
## - Marital_Status_Divorced    1       5.5 30188 5836.4
## - Marital_Status_Single      1       6.1 30188 5836.4
## - Education_1_Masters        1       6.7 30189 5836.5
## - Education_1_Bachelors      1       8.9 30191 5836.6
## - Country_IND                1       9.5 30192 5836.7
## - Year_Birth                 1      15.7 30198 5837.1
## - Marital_Status_Absurd      1      19.6 30202 5837.4
## <none>                                   30182 5838.0
## - Country_SA                 1      27.4 30210 5838.0
## - Country_AUS                1      29.6 30212 5838.1
## - Country_CA                 1      32.1 30214 5838.3
## - Country_GER                1      32.7 30215 5838.4
```

```
## - MntMeatProducts            1      43.6 30226 5839.2
## - `Education_1_High School`  1      52.6 30235 5839.8
## - Country_SP                 1      63.7 30246 5840.6
## - MntFruits                  1     156.6 30339 5847.4
## - MntFishProducts            1     194.5 30377 5850.2
## - MntSweetProducts           1     456.7 30639 5869.2
## - Kidhome                    1    1089.5 31272 5914.5
## - MntGoldProds               1    1581.3 31764 5949.0
## - Income                     1    2024.4 32207 5979.7
## - MntWines                   1    7704.2 37886 6339.3
##
## Step:  AIC=5835.97
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Education_1_Masters +
##     Marital_Status_Absurd + Marital_Status_Alone + Marital_Status_Divorced +
##     Marital_Status_Married + Marital_Status_Single + Marital_Status_Together +
##     Marital_Status_Widow + Country_AUS + Country_CA + Country_GER +
##     Country_IND + Country_SA + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Marital_Status_Alone       1       0.7 30183 5834.0
## - Marital_Status_Married     1       4.1 30186 5834.3
## - Marital_Status_Together    1       5.1 30187 5834.3
## - Marital_Status_Widow       1       5.5 30188 5834.4
## - Marital_Status_Divorced    1       5.5 30188 5834.4
## - Marital_Status_Single      1       6.1 30188 5834.4
## - Education_1_Masters        1       6.7 30189 5834.5
## - Education_1_Bachelors      1       8.9 30191 5834.6
## - Country_IND                1       9.5 30192 5834.7
## - Year_Birth                 1      15.8 30198 5835.1
## - Marital_Status_Absurd      1      19.6 30202 5835.4
## <none>                                   30182 5836.0
## - Country_SA                 1      27.6 30210 5836.0
## - Country_AUS                1      29.8 30212 5836.2
## - Country_CA                 1      32.4 30215 5836.3
## - Country_GER                1      32.9 30215 5836.4
## - MntMeatProducts            1      43.6 30226 5837.2
## - `Education_1_High School`  1      52.6 30235 5837.8
## + Country_ME                 1       0.0 30182 5838.0
## - Country_SP                 1      64.8 30247 5838.7
## - MntFruits                  1     157.3 30340 5845.5
## - MntFishProducts            1     194.5 30377 5848.2
## - MntSweetProducts           1     456.6 30639 5867.2
## - Kidhome                    1    1089.7 31272 5912.5
## - MntGoldProds               1    1581.4 31764 5947.0
## - Income                     1    2025.7 32208 5977.8
## - MntWines                   1    7707.4 37890 6337.5
##
## Step:  AIC=5834.02
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Education_1_Masters +
##     Marital_Status_Absurd + Marital_Status_Divorced + Marital_Status_Married +
```

```
##     Marital_Status_Single + Marital_Status_Together + Marital_Status_Widow +
##     Country_AUS + Country_CA + Country_GER + Country_IND + Country_SA +
##     Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Marital_Status_Married      1       4.8 30188 5832.4
## - Marital_Status_Together     1       6.5 30189 5832.5
## - Education_1_Masters         1       6.9 30190 5832.5
## - Marital_Status_Widow        1       7.0 30190 5832.5
## - Marital_Status_Divorced     1       7.2 30190 5832.6
## - Marital_Status_Single       1       8.4 30191 5832.6
## - Education_1_Bachelors       1       9.1 30192 5832.7
## - Country_IND                 1       9.4 30192 5832.7
## - Year_Birth                  1      15.7 30199 5833.2
## - Marital_Status_Absurd       1      22.5 30205 5833.7
## <none>                                    30183 5834.0
## - Country_SA                  1      27.6 30211 5834.0
## - Country_AUS                 1      29.8 30213 5834.2
## - Country_CA                  1      32.6 30216 5834.4
## - Country_GER                 1      32.9 30216 5834.4
## - MntMeatProducts             1      43.5 30227 5835.2
## - `Education_1_High School`   1      52.8 30236 5835.9
## + Marital_Status_Alone        1       0.7 30182 5836.0
## + Country_ME                  1       0.0 30183 5836.0
## - Country_SP                  1      64.8 30248 5836.8
## - MntFruits                   1     157.3 30340 5843.5
## - MntFishProducts             1     194.4 30377 5846.2
## - MntSweetProducts            1     456.5 30639 5865.3
## - Kidhome                     1    1095.3 31278 5910.9
## - MntGoldProds                1    1581.7 31765 5945.1
## - Income                      1    2025.3 32208 5975.8
## - MntWines                    1    7706.8 37890 6335.5
##
## Step:  AIC=5832.38
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Education_1_Masters +
##     Marital_Status_Absurd + Marital_Status_Divorced + Marital_Status_Single +
##     Marital_Status_Together + Marital_Status_Widow + Country_AUS +
##     Country_CA + Country_GER + Country_IND + Country_SA + Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Marital_Status_Widow        1       4.2 30192 5830.7
## - Education_1_Masters         1       7.3 30195 5830.9
## - Country_IND                 1       9.1 30197 5831.0
## - Marital_Status_Together     1       9.5 30197 5831.1
## - Education_1_Bachelors       1       9.7 30197 5831.1
## - Marital_Status_Divorced     1      10.5 30198 5831.1
## - Year_Birth                  1      15.3 30203 5831.5
## - Marital_Status_Absurd       1      17.9 30206 5831.7
## <none>                                    30188 5832.4
## - Country_SA                  1      27.7 30215 5832.4
## - Country_AUS                 1      29.9 30218 5832.6
## - Country_CA                  1      31.3 30219 5832.7
```

```
## - Marital_Status_Single      1      32.2 30220 5832.7
## - Country_GER                1      33.0 30221 5832.8
## - MntMeatProducts            1      43.1 30231 5833.5
## + Marital_Status_Married     1       4.8 30183 5834.0
## + Marital_Status_Alone       1       1.4 30186 5834.3
## - `Education_1_High School`  1      53.5 30241 5834.3
## + Country_ME                 1       0.0 30188 5834.4
## - Country_SP                 1      65.0 30253 5835.1
## - MntFruits                  1     156.7 30344 5841.8
## - MntFishProducts            1     194.4 30382 5844.6
## - MntSweetProducts           1     456.0 30644 5863.6
## - Kidhome                    1    1095.5 31283 5909.3
## - MntGoldProds               1    1583.6 31771 5943.6
## - Income                     1    2026.3 32214 5974.2
## - MntWines                   1    7706.6 37894 6333.8
##
## Step:  AIC=5830.68
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Education_1_Masters +
##     Marital_Status_Absurd + Marital_Status_Divorced + Marital_Status_Single +
##     Marital_Status_Together + Country_AUS + Country_CA + Country_GER +
##     Country_IND + Country_SA + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Education_1_Masters        1       7.0 30199 5829.2
## - Marital_Status_Together    1       7.6 30200 5829.2
## - Marital_Status_Divorced    1       8.9 30201 5829.3
## - Country_IND                1       9.2 30201 5829.4
## - Education_1_Bachelors      1       9.3 30201 5829.4
## - Year_Birth                 1      13.3 30205 5829.7
## - Marital_Status_Absurd      1      17.6 30210 5830.0
## <none>                                   30192 5830.7
## - Country_SA                 1      28.1 30220 5830.7
## - Marital_Status_Single      1      29.4 30221 5830.8
## - Country_AUS                1      29.9 30222 5830.9
## - Country_CA                 1      31.2 30223 5831.0
## - Country_GER                1      33.0 30225 5831.1
## - MntMeatProducts            1      43.3 30235 5831.9
## + Marital_Status_Widow       1       4.2 30188 5832.4
## + Marital_Status_Married     1       2.0 30190 5832.5
## + Marital_Status_Alone       1       1.4 30191 5832.6
## - `Education_1_High School`  1      53.3 30245 5832.6
## + Country_ME                 1       0.0 30192 5832.7
## - Country_SP                 1      65.5 30257 5833.5
## - MntFruits                  1     156.9 30349 5840.2
## - MntFishProducts            1     193.2 30385 5842.8
## - MntSweetProducts           1     453.7 30646 5861.7
## - Kidhome                    1    1092.8 31285 5907.4
## - MntGoldProds               1    1580.7 31773 5941.7
## - Income                     1    2031.4 32223 5972.9
## - MntWines                   1    7710.6 37903 6332.2
##
## Step:  AIC=5829.19
```

```
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Education_1_Bachelors + Marital_Status_Absurd +
##     Marital_Status_Divorced + Marital_Status_Single + Marital_Status_Together +
##     Country_AUS + Country_CA + Country_GER + Country_IND + Country_SA +
##     Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Education_1_Bachelors      1       3.2 30202 5827.4
## - Marital_Status_Together    1       8.1 30207 5827.8
## - Marital_Status_Divorced    1       9.1 30208 5827.9
## - Country_IND                1       9.4 30208 5827.9
## - Year_Birth                 1      14.3 30213 5828.2
## - Marital_Status_Absurd      1      17.7 30217 5828.5
## <none>                                   30199 5829.2
## - Country_SA                 1      27.8 30227 5829.2
## - Marital_Status_Single      1      29.5 30228 5829.4
## - Country_AUS                1      30.1 30229 5829.4
## - Country_CA                 1      31.5 30230 5829.5
## - Country_GER                1      34.0 30233 5829.7
## - MntMeatProducts            1      43.2 30242 5830.4
## - `Education_1_High School`  1      46.8 30246 5830.6
## + Education_1_Masters        1       7.0 30192 5830.7
## + Marital_Status_Widow       1       3.9 30195 5830.9
## + Marital_Status_Married     1       1.7 30197 5831.1
## + Marital_Status_Alone       1       1.5 30197 5831.1
## + Country_ME                 1       0.0 30199 5831.2
## - Country_SP                 1      65.2 30264 5832.0
## - MntFruits                  1     155.9 30355 5838.6
## - MntFishProducts            1     189.2 30388 5841.0
## - MntSweetProducts           1     449.4 30648 5859.9
## - Kidhome                    1    1093.9 31293 5906.0
## - MntGoldProds               1    1574.9 31774 5939.7
## - Income                     1    2052.5 32251 5972.8
## - MntWines                   1    7869.2 38068 6339.9
##
## Step:  AIC=5827.43
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Marital_Status_Absurd + Marital_Status_Divorced +
##     Marital_Status_Single + Marital_Status_Together + Country_AUS +
##     Country_CA + Country_GER + Country_IND + Country_SA + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Marital_Status_Together    1       8.2 30210 5826.0
## - Country_IND                1       9.1 30211 5826.1
## - Marital_Status_Divorced    1       9.2 30211 5826.1
## - Year_Birth                 1      15.2 30217 5826.5
## - Marital_Status_Absurd      1      17.4 30219 5826.7
## - Country_SA                 1      27.0 30229 5827.4
## <none>                                   30202 5827.4
## - Marital_Status_Single      1      29.8 30232 5827.6
## - Country_AUS                1      29.8 30232 5827.6
## - Country_CA                 1      30.9 30233 5827.7
```

```
## - Country_GER                1      33.5 30236 5827.9
## - MntMeatProducts            1      42.5 30245 5828.5
## - `Education_1_High School`  1      44.0 30246 5828.6
## + Marital_Status_Widow       1       3.8 30198 5829.1
## + Education_1_Bachelors      1       3.2 30199 5829.2
## + Marital_Status_Married     1       1.6 30200 5829.3
## + Marital_Status_Alone       1       1.5 30201 5829.3
## + Education_1_Masters        1       0.8 30201 5829.4
## + Country_ME                 1       0.0 30202 5829.4
## - Country_SP                 1      64.4 30266 5830.1
## - MntFruits                  1     154.1 30356 5836.7
## - MntFishProducts            1     187.7 30390 5839.1
## - MntSweetProducts           1     446.7 30649 5857.9
## - Kidhome                    1    1094.7 31297 5904.3
## - MntGoldProds               1    1580.8 31783 5938.4
## - Income                     1    2065.1 32267 5971.9
## - MntWines                   1    8020.3 38222 6346.8
##
## Step:  AIC=5826.03
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Marital_Status_Absurd + Marital_Status_Divorced +
##     Marital_Status_Single + Country_AUS + Country_CA + Country_GER +
##     Country_IND + Country_SA + Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Marital_Status_Divorced     1       5.5 30216 5824.4
## - Country_IND                 1       9.0 30219 5824.7
## - Year_Birth                  1      14.4 30225 5825.1
## - Marital_Status_Absurd       1      16.7 30227 5825.3
## - Marital_Status_Single       1      22.6 30233 5825.7
## <none>                                    30210 5826.0
## - Country_SA                  1      27.6 30238 5826.1
## - Country_AUS                 1      29.4 30240 5826.2
## - Country_CA                  1      30.6 30241 5826.3
## - Country_GER                 1      33.0 30243 5826.4
## - MntMeatProducts             1      41.9 30252 5827.1
## - `Education_1_High School`   1      44.3 30255 5827.3
## + Marital_Status_Married      1       9.8 30201 5827.3
## + Marital_Status_Together     1       8.2 30202 5827.4
## + Education_1_Bachelors       1       3.3 30207 5827.8
## + Marital_Status_Widow        1       1.9 30208 5827.9
## + Marital_Status_Alone        1       1.7 30209 5827.9
## + Education_1_Masters         1       0.9 30209 5828.0
## + Country_ME                  1       0.1 30210 5828.0
## - Country_SP                  1      63.9 30274 5828.7
## - MntFruits                   1     154.8 30365 5835.3
## - MntFishProducts             1     185.5 30396 5837.6
## - MntSweetProducts            1     449.3 30660 5856.7
## - Kidhome                     1    1098.4 31309 5903.1
## - MntGoldProds                1    1582.6 31793 5937.1
## - Income                      1    2068.0 32278 5970.6
## - MntWines                    1    8017.7 38228 6345.2
##
```

```
## Step:  AIC=5824.43
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Marital_Status_Absurd + Marital_Status_Single +
##     Country_AUS + Country_CA + Country_GER + Country_IND + Country_SA +
##     Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Country_IND                 1       8.4 30224 5823.0
## - Year_Birth                  1      13.7 30229 5823.4
## - Marital_Status_Absurd       1      16.5 30232 5823.6
## - Marital_Status_Single       1      19.5 30235 5823.9
## <none>                                    30216 5824.4
## - Country_SA                  1      27.3 30243 5824.4
## - Country_AUS                 1      28.9 30245 5824.5
## - Country_CA                  1      29.7 30246 5824.6
## - Country_GER                 1      31.5 30247 5824.7
## + Marital_Status_Married      1      14.3 30201 5825.4
## - MntMeatProducts             1      42.7 30259 5825.6
## - `Education_1_High School`   1      43.3 30259 5825.6
## + Marital_Status_Divorced     1       5.5 30210 5826.0
## + Marital_Status_Together     1       4.4 30211 5826.1
## + Education_1_Bachelors       1       3.3 30212 5826.2
## + Marital_Status_Alone        1       1.8 30214 5826.3
## + Marital_Status_Widow        1       1.3 30214 5826.3
## + Education_1_Masters         1       0.9 30215 5826.4
## + Country_ME                  1       0.0 30216 5826.4
## - Country_SP                  1      62.4 30278 5827.0
## - MntFruits                   1     153.4 30369 5833.6
## - MntFishProducts             1     187.1 30403 5836.1
## - MntSweetProducts            1     449.2 30665 5855.1
## - Kidhome                     1    1097.1 31313 5901.4
## - MntGoldProds                1    1580.4 31796 5935.3
## - Income                      1    2068.6 32284 5969.0
## - MntWines                    1    8013.5 38229 6343.2
##
## Step:  AIC=5823.05
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth +
##     `Education_1_High School` + Marital_Status_Absurd + Marital_Status_Single +
##     Country_AUS + Country_CA + Country_GER + Country_SA + Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Year_Birth                  1      14.2 30238 5822.1
## - Marital_Status_Absurd       1      16.5 30241 5822.3
## - Marital_Status_Single       1      18.9 30243 5822.4
## - Country_SA                  1      19.4 30244 5822.5
## - Country_AUS                 1      20.5 30245 5822.5
## - Country_CA                  1      21.9 30246 5822.6
## - Country_GER                 1      23.2 30247 5822.7
## <none>                                    30224 5823.0
## + Marital_Status_Married      1      14.0 30210 5824.0
## - MntMeatProducts             1      41.8 30266 5824.1
## - `Education_1_High School`   1      43.0 30267 5824.2
```

```
## + Country_IND                1       8.4 30216 5824.4
## + Marital_Status_Divorced    1       4.9 30219 5824.7
## + Marital_Status_Together    1       4.5 30220 5824.7
## + Education_1_Bachelors      1       3.1 30221 5824.8
## + Marital_Status_Alone       1       1.9 30222 5824.9
## + Marital_Status_Widow       1       1.4 30223 5824.9
## + Education_1_Masters        1       1.1 30223 5825.0
## + Country_ME                 1       0.0 30224 5825.0
## - Country_SP                 1      70.4 30295 5826.2
## - MntFruits                  1     152.3 30376 5832.2
## - MntFishProducts            1     188.8 30413 5834.8
## - MntSweetProducts           1     450.0 30674 5853.8
## - Kidhome                    1    1101.9 31326 5900.3
## - MntGoldProds               1    1580.0 31804 5933.9
## - Income                     1    2074.8 32299 5968.0
## - MntWines                   1    8017.6 38242 6342.0
##
## Step:  AIC=5822.09
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Absurd + Marital_Status_Single + Country_AUS +
##     Country_CA + Country_GER + Country_SA + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Marital_Status_Absurd      1      17.4 30256 5821.4
## - Country_SA                 1      19.2 30258 5821.5
## - Country_AUS                1      19.5 30258 5821.5
## - Country_CA                 1      21.0 30259 5821.6
## - Marital_Status_Single      1      22.6 30261 5821.7
## - Country_GER                1      23.2 30262 5821.8
## <none>                                   30238 5822.1
## - MntMeatProducts            1      36.7 30275 5822.8
## + Year_Birth                 1      14.2 30224 5823.0
## + Marital_Status_Married     1      11.0 30227 5823.3
## + Country_IND                1       9.0 30229 5823.4
## - `Education_1_High School`  1      46.5 30285 5823.5
## + Marital_Status_Divorced    1       4.2 30234 5823.8
## + Marital_Status_Together    1       4.1 30234 5823.8
## + Education_1_Bachelors      1       3.9 30235 5823.8
## + Marital_Status_Alone       1       1.7 30237 5824.0
## + Education_1_Masters        1       1.1 30237 5824.0
## + Marital_Status_Widow       1       0.4 30238 5824.1
## + Country_ME                 1       0.0 30238 5824.1
## - Country_SP                 1      69.5 30308 5825.2
## - MntFruits                  1     148.6 30387 5830.9
## - MntFishProducts            1     188.8 30427 5833.9
## - MntSweetProducts           1     442.1 30681 5852.2
## - Kidhome                    1    1191.6 31430 5905.7
## - MntGoldProds               1    1577.6 31816 5932.7
## - Income                     1    2175.9 32414 5973.9
## - MntWines                   1    8033.0 38271 6341.7
##
## Step:  AIC=5821.36
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
```

```
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Single + Country_AUS + Country_CA + Country_GER +
##     Country_SA + Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Country_SA                  1      19.2 30275 5820.8
## - Country_AUS                 1      21.3 30277 5820.9
## - Marital_Status_Single       1      22.1 30278 5821.0
## - Country_CA                  1      22.3 30278 5821.0
## - Country_GER                 1      23.1 30279 5821.1
## <none>                                    30256 5821.4
## + Marital_Status_Absurd       1      17.4 30238 5822.1
## - MntMeatProducts             1      37.7 30294 5822.1
## + Year_Birth                  1      15.1 30241 5822.3
## + Marital_Status_Married      1      11.8 30244 5822.5
## + Country_IND                 1       9.1 30247 5822.7
## - `Education_1_High School`   1      46.5 30302 5822.8
## + Marital_Status_Divorced     1       3.9 30252 5823.1
## + Marital_Status_Together     1       3.7 30252 5823.1
## + Education_1_Bachelors       1       3.5 30252 5823.1
## + Marital_Status_Alone        1       1.8 30254 5823.2
## + Education_1_Masters         1       1.2 30255 5823.3
## + Marital_Status_Widow        1       0.4 30255 5823.3
## + Country_ME                  1       0.0 30256 5823.4
## - Country_SP                  1      69.8 30326 5824.5
## - MntFruits                   1     148.1 30404 5830.2
## - MntFishProducts             1     180.2 30436 5832.5
## - MntSweetProducts            1     454.7 30710 5852.4
## - Kidhome                     1    1191.4 31447 5904.9
## - MntGoldProds                1    1561.7 31818 5930.8
## - Income                      1    2170.7 32427 5972.8
## - MntWines                    1    8067.3 38323 6342.7
##
## Step:  AIC=5820.76
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Single + Country_AUS + Country_CA + Country_GER +
##     Country_SP
##
##                              Df Sum of Sq   RSS    AIC
## - Country_CA                  1       7.9 30283 5819.3
## - Country_AUS                 1       8.6 30284 5819.4
## - Country_GER                 1      10.7 30286 5819.5
## - Marital_Status_Single       1      20.6 30296 5820.3
## <none>                                    30275 5820.8
## + Country_SA                  1      19.2 30256 5821.4
## + Marital_Status_Absurd       1      17.3 30258 5821.5
## - MntMeatProducts             1      38.5 30313 5821.6
## + Year_Birth                  1      14.8 30260 5821.7
## + Marital_Status_Married      1      13.1 30262 5821.8
## - `Education_1_High School`   1      46.1 30321 5822.1
## + Marital_Status_Divorced     1       4.4 30271 5822.4
## + Marital_Status_Together     1       4.1 30271 5822.5
## + Education_1_Bachelors       1       3.0 30272 5822.5
```

```
## + Marital_Status_Alone       1       1.8 30273 5822.6
## + Education_1_Masters        1       1.2 30274 5822.7
## - Country_SP                 1      53.6 30329 5822.7
## + Marital_Status_Widow       1       0.5 30275 5822.7
## + Country_IND                1       0.4 30275 5822.7
## + Country_ME                 1       0.2 30275 5822.7
## - MntFruits                  1     149.9 30425 5829.7
## - MntFishProducts            1     178.4 30453 5831.8
## - MntSweetProducts           1     452.6 30728 5851.6
## - Kidhome                    1    1201.7 31477 5904.9
## - MntGoldProds               1    1562.0 31837 5930.1
## - Income                     1    2166.7 32442 5971.8
## - MntWines                   1    8051.3 38326 6340.9
##
## Step:  AIC=5819.34
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Single + Country_AUS + Country_GER + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Country_AUS                1       5.3 30288 5817.7
## - Country_GER                1       7.3 30290 5817.9
## - Marital_Status_Single      1      20.3 30303 5818.8
## <none>                                   30283 5819.3
## + Marital_Status_Absurd      1      18.2 30265 5820.0
## - MntMeatProducts            1      39.0 30322 5820.2
## + Year_Birth                 1      14.3 30269 5820.3
## + Marital_Status_Married     1      12.4 30270 5820.4
## - Country_SP                 1      46.1 30329 5820.7
## - `Education_1_High School`  1      46.5 30329 5820.7
## + Country_CA                 1       7.9 30275 5820.8
## + Country_SA                 1       4.8 30278 5821.0
## + Marital_Status_Divorced    1       4.1 30279 5821.0
## + Marital_Status_Together    1       3.9 30279 5821.1
## + Education_1_Bachelors      1       3.0 30280 5821.1
## + Country_IND                1       2.1 30281 5821.2
## + Education_1_Masters        1       1.3 30282 5821.2
## + Marital_Status_Alone       1       1.2 30282 5821.3
## + Marital_Status_Widow       1       0.4 30282 5821.3
## + Country_ME                 1       0.3 30283 5821.3
## - MntFruits                  1     148.9 30432 5828.2
## - MntFishProducts            1     180.7 30464 5830.5
## - MntSweetProducts           1     450.4 30733 5850.0
## - Kidhome                    1    1197.5 31480 5903.2
## - MntGoldProds               1    1562.1 31845 5928.7
## - Income                     1    2166.4 32449 5970.3
## - MntWines                   1    8048.4 38331 6339.1
##
## Step:  AIC=5817.73
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Single + Country_GER + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
```

```
## - Country_GER               1       5.8 30294 5816.2
## - Marital_Status_Single     1      19.1 30307 5817.1
## <none>                                  30288 5817.7
## + Marital_Status_Absurd     1      19.1 30269 5818.3
## - MntMeatProducts           1      39.5 30328 5818.6
## + Year_Birth                1      13.9 30274 5818.7
## - Country_SP                1      40.9 30329 5818.7
## + Marital_Status_Married    1      12.0 30276 5818.8
## - `Education_1_High School` 1      46.7 30335 5819.1
## + Country_AUS               1       5.3 30283 5819.3
## + Country_CA                1       4.6 30284 5819.4
## + Marital_Status_Divorced   1       4.1 30284 5819.4
## + Marital_Status_Together   1       3.6 30285 5819.5
## + Country_IND               1       3.3 30285 5819.5
## + Education_1_Bachelors     1       3.1 30285 5819.5
## + Country_SA                1       2.0 30286 5819.6
## + Marital_Status_Alone      1       1.3 30287 5819.6
## + Education_1_Masters       1       1.3 30287 5819.6
## + Country_ME                1       0.4 30288 5819.7
## + Marital_Status_Widow      1       0.4 30288 5819.7
## - MntFruits                 1     149.8 30438 5826.7
## - MntFishProducts           1     180.6 30469 5828.9
## - MntSweetProducts          1     448.5 30737 5848.3
## - Kidhome                   1    1200.8 31489 5901.8
## - MntGoldProds              1    1558.3 31846 5926.8
## - Income                    1    2165.0 32453 5968.6
## - MntWines                  1    8049.9 38338 6337.5
##
## Step:  AIC=5816.16
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Marital_Status_Single + Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## - Marital_Status_Single      1      18.5 30313 5815.5
## <none>                                  30294 5816.2
## - Country_SP                 1      35.9 30330 5816.8
## + Marital_Status_Absurd      1      18.7 30275 5816.8
## - MntMeatProducts            1      39.3 30333 5817.0
## + Year_Birth                 1      14.2 30280 5817.1
## + Marital_Status_Married     1      11.3 30283 5817.3
## - `Education_1_High School` 1      46.5 30341 5817.5
## + Country_GER                1       5.8 30288 5817.7
## + Country_IND                1       4.5 30290 5817.8
## + Country_AUS                1       3.9 30290 5817.9
## + Marital_Status_Divorced    1       3.6 30290 5817.9
## + Marital_Status_Together    1       3.6 30290 5817.9
## + Education_1_Bachelors      1       3.1 30291 5817.9
## + Country_CA                 1       2.7 30291 5818.0
## + Education_1_Masters        1       1.5 30293 5818.0
## + Marital_Status_Alone       1       1.4 30293 5818.1
## + Country_SA                 1       0.7 30293 5818.1
## + Country_ME                 1       0.5 30294 5818.1
## + Marital_Status_Widow       1       0.4 30294 5818.1
```

```
## - MntFruits                  1     150.6 30445 5825.1
## - MntFishProducts            1     180.2 30474 5827.3
## - MntSweetProducts           1     451.2 30745 5846.9
## - Kidhome                    1    1196.3 31490 5899.9
## - MntGoldProds               1    1555.7 31850 5925.0
## - Income                     1    2166.5 32461 5967.1
## - MntWines                   1    8051.5 38346 6336.0
##
## Step:  AIC=5815.51
## totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts +
##     MntFishProducts + MntSweetProducts + MntGoldProds + `Education_1_High School` +
##     Country_SP
##
##                             Df Sum of Sq   RSS    AIC
## <none>                                   30313 5815.5
## + Marital_Status_Married     1      23.4 30289 5815.8
## - MntMeatProducts            1      35.3 30348 5816.1
## + Marital_Status_Single      1      18.5 30294 5816.2
## - Country_SP                 1      36.5 30349 5816.2
## + Marital_Status_Absurd      1      18.1 30294 5816.2
## + Year_Birth                 1      17.5 30295 5816.2
## - `Education_1_High School`  1      48.7 30361 5817.1
## + Country_GER                1       5.2 30307 5817.1
## + Country_IND                1       4.0 30308 5817.2
## + Education_1_Bachelors      1       3.4 30309 5817.3
## + Country_AUS                1       3.0 30310 5817.3
## + Country_CA                 1       2.8 30310 5817.3
## + Marital_Status_Alone       1       1.6 30311 5817.4
## + Education_1_Masters        1       1.3 30311 5817.4
## + Marital_Status_Divorced    1       1.3 30311 5817.4
## + Country_SA                 1       0.7 30312 5817.5
## + Country_ME                 1       0.2 30312 5817.5
## + Marital_Status_Together    1       0.2 30312 5817.5
## + Marital_Status_Widow       1       0.0 30312 5817.5
## - MntFruits                  1     150.1 30463 5824.4
## - MntFishProducts            1     180.2 30493 5826.6
## - MntSweetProducts           1     453.2 30766 5846.4
## - Kidhome                    1    1199.8 31512 5899.4
## - MntGoldProds               1    1556.7 31869 5924.4
## - Income                     1    2187.3 32500 5967.8
## - MntWines                   1    8069.9 38382 6336.1
##
## Call:
## lm(formula = totalpsum ~ Income + Kidhome + MntWines + MntFruits +
##     MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds +
##     `Education_1_High School` + Country_SP, data = dum_df_1)
##
## Coefficients:
##               (Intercept)                     Income
##                 4.8801819                  0.0000799
##                   Kidhome                   MntWines
##                -1.6877615                  0.0082067
##                 MntFruits            MntMeatProducts
```

```
##                0.0090354                 0.0008972
##          MntFishProducts          MntSweetProducts
##                0.0074321                 0.0150438
##             MntGoldProds `Education_1_High School`
##                0.0189579                -0.9961723
##               Country_SP
##               -0.2575305
```

Model: Fit_2 from stepwiseAIC

A linear regression model can now be created for the features selected by stepwise AIC. Income + Kidhome + MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds + Year_Birth + Education_1_High School + Country_SP + will be left in the model.

AIC Score: This model improves on the previous score of 12231 with an AIC score of 12101.

Plots:

Residuals vs Fitted: The residual plots show an increase in variance as purchase counts increase. There is a small improvement from removing the outlying datapoints shown in the previous model.

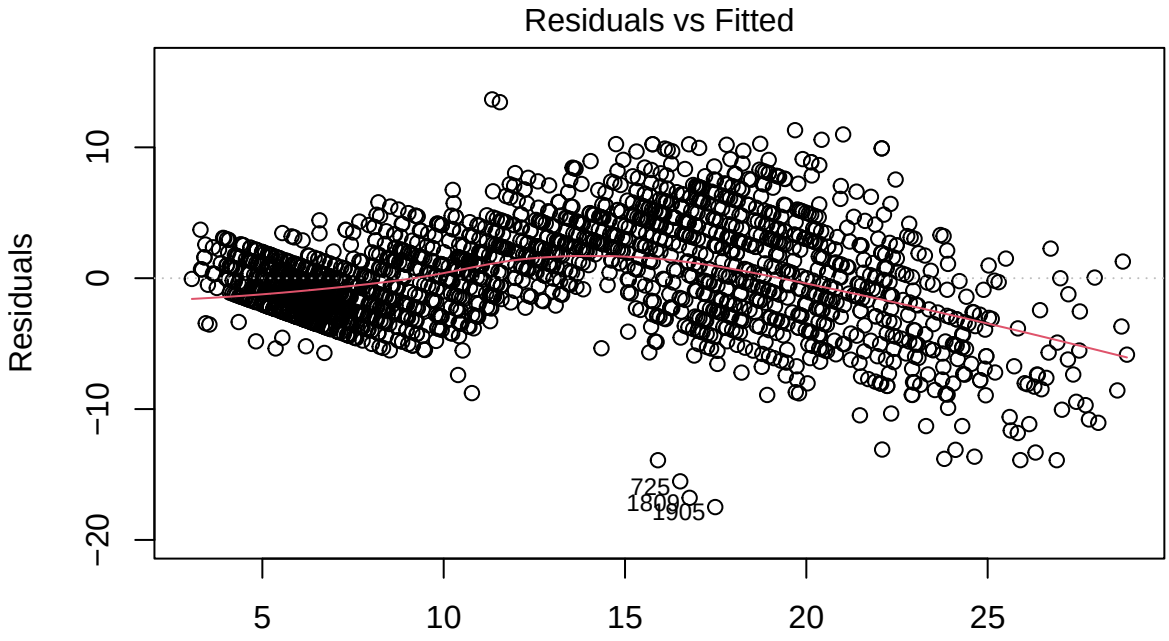
QQ Plots: The QQ Plot is approximately normal.

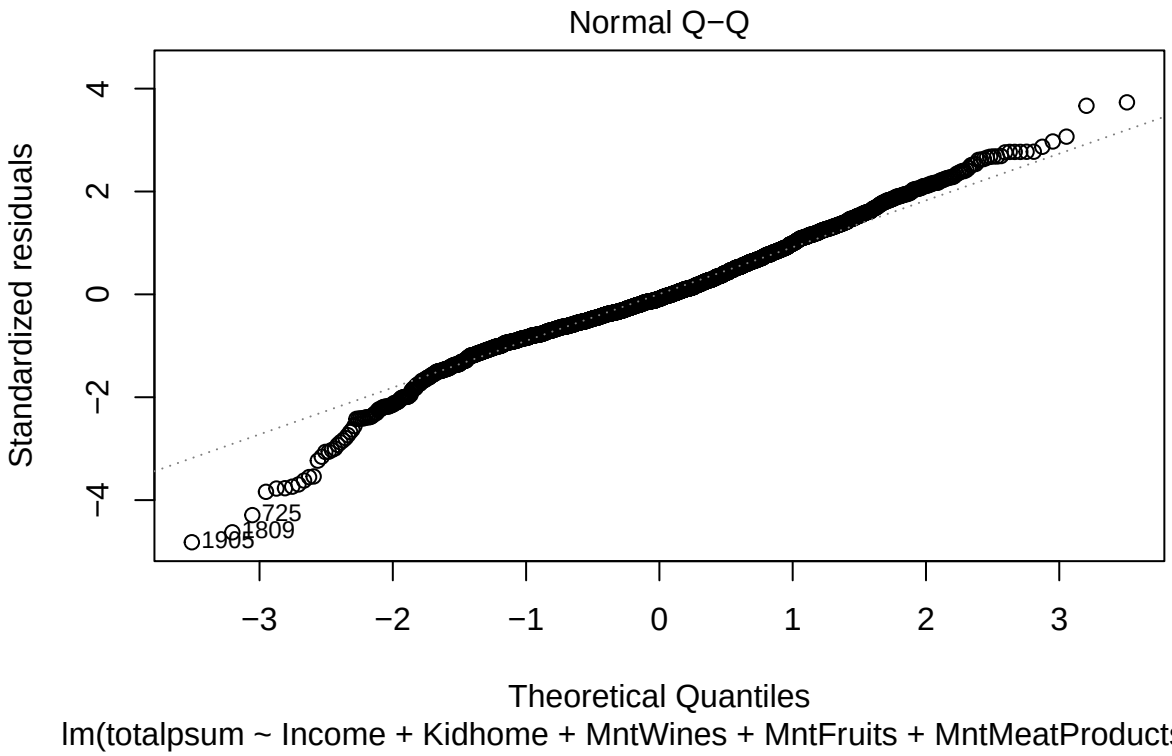
```
# stepwise model
# fit model
fit_2 <- lm(formula = totalpsum ~ Income + Kidhome + MntWines + MntFruits +
                      MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds +
                      Year_Birth + `Education_1_High School` + Country_SP,
            data = dum_df_1)

# show model statistics
summary(fit_2)
```

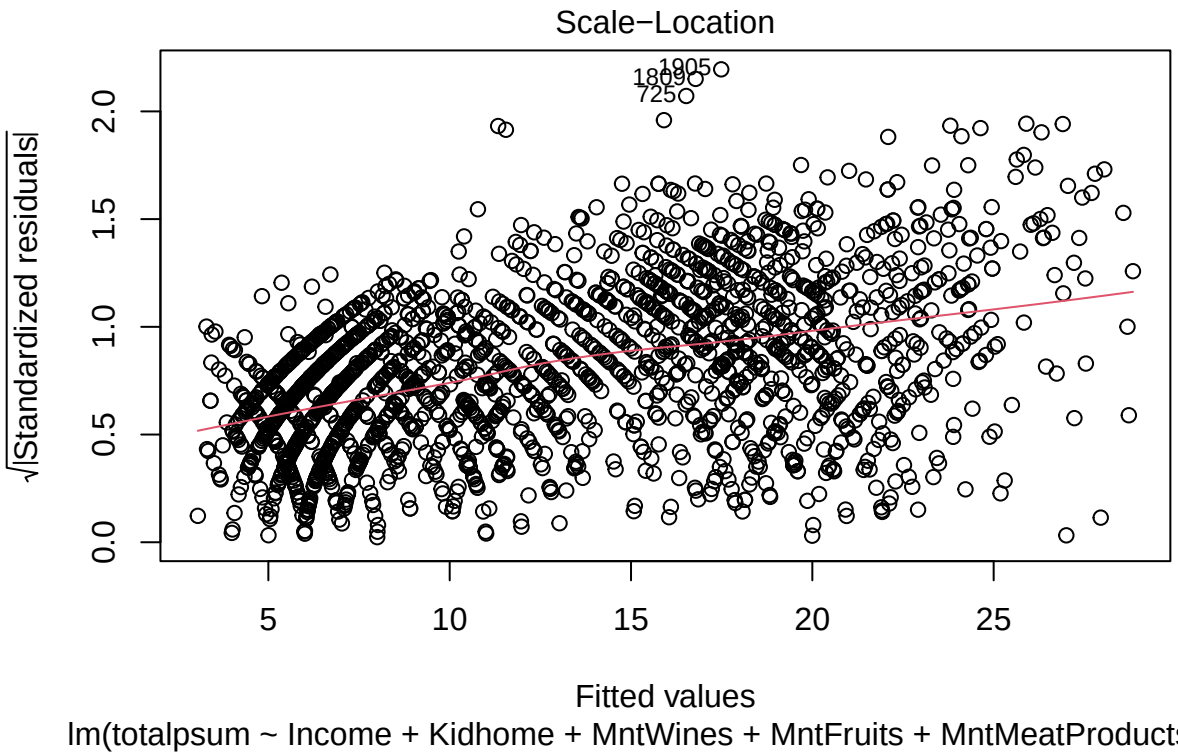
```
##
## Call:
## lm(formula = totalpsum ~ Income + Kidhome + MntWines + MntFruits +
##     MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds +
##     Year_Birth + `Education_1_High School` + Country_SP, data = dum_df_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4882  -2.2474  -0.3104   2.2971  13.6631
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.035e+01  1.370e+01   1.485 0.137676
## Income                    7.879e-05  6.413e-06  12.286  < 2e-16 ***
## Kidhome                  -1.649e+00  1.840e-01  -8.961  < 2e-16 ***
## MntWines                  8.198e-03  3.389e-04  24.189  < 2e-16 ***
## MntFruits                 9.171e-03  2.738e-03   3.349 0.000823 ***
## MntMeatProducts           9.787e-04  5.651e-04   1.732 0.083452 .
## MntFishProducts           7.438e-03  2.053e-03   3.622 0.000299 ***
## MntSweetProducts          1.520e-02  2.625e-03   5.792 7.94e-09 ***
## MntGoldProds              1.898e-02  1.782e-03  10.648  < 2e-16 ***
```

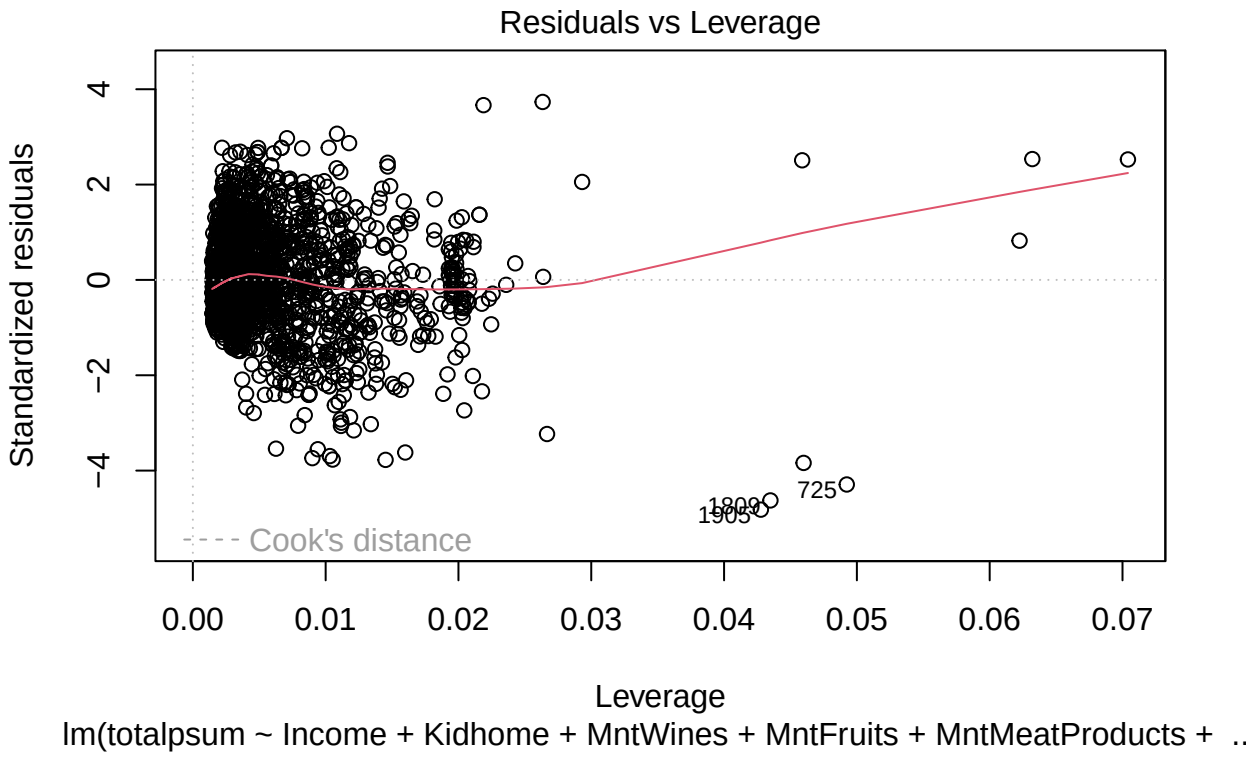
```
## Year_Birth                -7.848e-03  6.950e-03  -1.129 0.258918
## `Education_1_High School` -9.547e-01  5.306e-01  -1.800 0.072075 .
## Country_SP                -2.572e-01  1.581e-01  -1.627 0.103876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.709 on 2202 degrees of freedom
## Multiple R-squared:  0.7359, Adjusted R-squared:  0.7346
## F-statistic: 557.8 on 11 and 2202 DF,  p-value: < 2.2e-16
```

```
plot(fit_2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts + ..
1905
1809
725


Scale−Location
1905
1809
725
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
Fitted values
lm(totalpsum ~ Income + Kidhome + MntWines + MntFruits + MntMeatProducts + ..




```
AIC(fit_2)
```

```
## [1] 12101.28
```

Fit_2 VIF

Removing many of the variables through stepwise AIC has taken care of the high VIF scores shown in the last model. All variables now have a VIF score of 1 or 2.

```
# stepwise model VIF
vif(fit_2)
```

```
##                    Income                   Kidhome                 MntWines
##                  3.059510                  1.569709                  2.103023
##                 MntFruits           MntMeatProducts          MntFishProducts
##                  1.910993                  2.529464                  2.034266
##          MntSweetProducts              MntGoldProds                Year_Birth
##                  1.870662                  1.372435                  1.116525
## `Education_1_High School`                Country_SP
##                  1.077916                  1.005489
```

AIC summary scores

The original model scores an AIC of 12123, the stepwise model improves upon this, achieving an AIC of 12101.

```
# score all models
# score fit_1
AIC(fit_1)
```

```
## [1] 12123.03
```

```
# score fit_2
AIC(fit_2)
```

```
## [1] 12101.28
```

Summary

The R square score of this model indicates that 73% of the variance in the dependent variable can be explained by the model. An F-test statistic less than 0.05 indicates that at least one variable in the model significantly impacts the dependent variable.

The highest positively correlated features in the model are the amounts spent on fruits and meats. Customers who spend more on fruits and meats will have a higher purchase count. This is not a surprising outcome as customers with the highest spend probably make more purchases.

The feature with the greatest negative correlation was the binary column for a high school education level. Customers with a high school education have around 9 fewer purchases than customers with a higher level of education.

It would be helpful to have additional demographic variables in the survey to test. It would allow analysis of how things like religion or political affiliation effect the count of purchases.

The original hypothesis does not hold out through the regression. Very few demographic features were chosen in the final model; only high_school_education and country_sp were included in the highest performing model. The variables that had the biggest impact were numerical values such as income or those corresponding to purchase amounts.

References

MSDS660. (2022). Statistical Methods and Experimental Design. Taught by Dr. Siripun Sanguansintukul.

Hult International Business School. (n.d.). marketing data . dataset. retrieved 10/22/22 from https://worldclass.regis.edu/d2l/le/content/297311/Home