Homework 4

Elizabeth Chung, Shay Diamond, Flaka Bajraktari, Ekaterina Marbot, and Om0lara Adelaja

10/9/2018

1. Money spent on entertainment

Descriptive Statistics

We'll first start with the descriptive statistics for the entertainment/total expenditure ratio (ENTERTPQ/TOTEXPPQ) in it's entirety:

```
entertainment_prop <- clean_data$ent_prop
meanProp <- mean(entertainment_prop)
# [1] 0.05189267

ent_sd <- sd(entertainment_prop)
# [1] 0.05735944

n <- length(entertainment_prop)
stdError <- qt(.025, df = n -1, lower.tail = FALSE) * sd(entertainment_prop) / sqrt(n)
# [1] 0.00135977

upperBound = meanProp + stdError
# [1] 0.05325244

lowerBound = meanProp - stdError
# [1] 0.0505329
```

At a 95% confidence interval, this gives us range of (5.053%, 5.325%) with a mean of 5.189% and standard deviation of 0.0013 which at first glance seems like a relatively small range. So to understand how other factors play into the variation, we will first compare a sub-category that looks at how being a male or female factors into the proportion:

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Wed, Oct 17, 2018 - 08:09:56
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multic
## \hline \\[-1.8ex]
## female & 2 & 0.500 & 0.707 & 0 & 0.2 & 0.8 & 1 \\
## n & 2 & 3,419.000 & 321.026 & 3,192 & 3,305.5 & 3,532.5 & 3,646 \\
## meanProp & 2 & 0.052 & 0.002 & 0.051 & 0.051 & 0.052 & 0.053 \\
## sdProp & 2 & 0.057 & 0.003 & 0.055 & 0.056 & 0.059 & 0.060 \\
## stdError & 2 & 0.002 & 0.0002 & 0.002 & 0.002 & 0.002 & 0.002 \\
## lowerBound & 2 & 0.050 & 0.002 & 0.048 & 0.049 & 0.051 & 0.051 \\
## upperBound & 2 & 0.054 & 0.002 & 0.053 & 0.053 & 0.054 & 0.055 \\
## \hline \\[-1.8ex]
```

```
## \end{tabular}
## \end{table}
```

Based on these results, the means slightly vary between the two; where females appear to have a slightly higher percentage spent on entertainment compared to males (5.31% to 5.06%), which leads to interest in running statistical significance tests later.

The next sub-group we looked at was if the individual was from an urban or rural area:

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Wed, Oct 17, 2018 - 08:09:56
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multi
## \hline \\[-1.8ex]
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

In this case, being from an urban area gives a factor of 1, rural being 2. At first glace we can see that the data set has significantly more observations from urban than rural with the mean from urban being lower than those from rural at 5.18% vs. 5.44%. This logically makes sense as those in rural areas have less access to affordable entertainment so they would have to spend more at the few entertainment options vs. the people in a urban area which have many options at varying levels of prices to pick from. Definitely a factor to consider for future statistical significance testing.

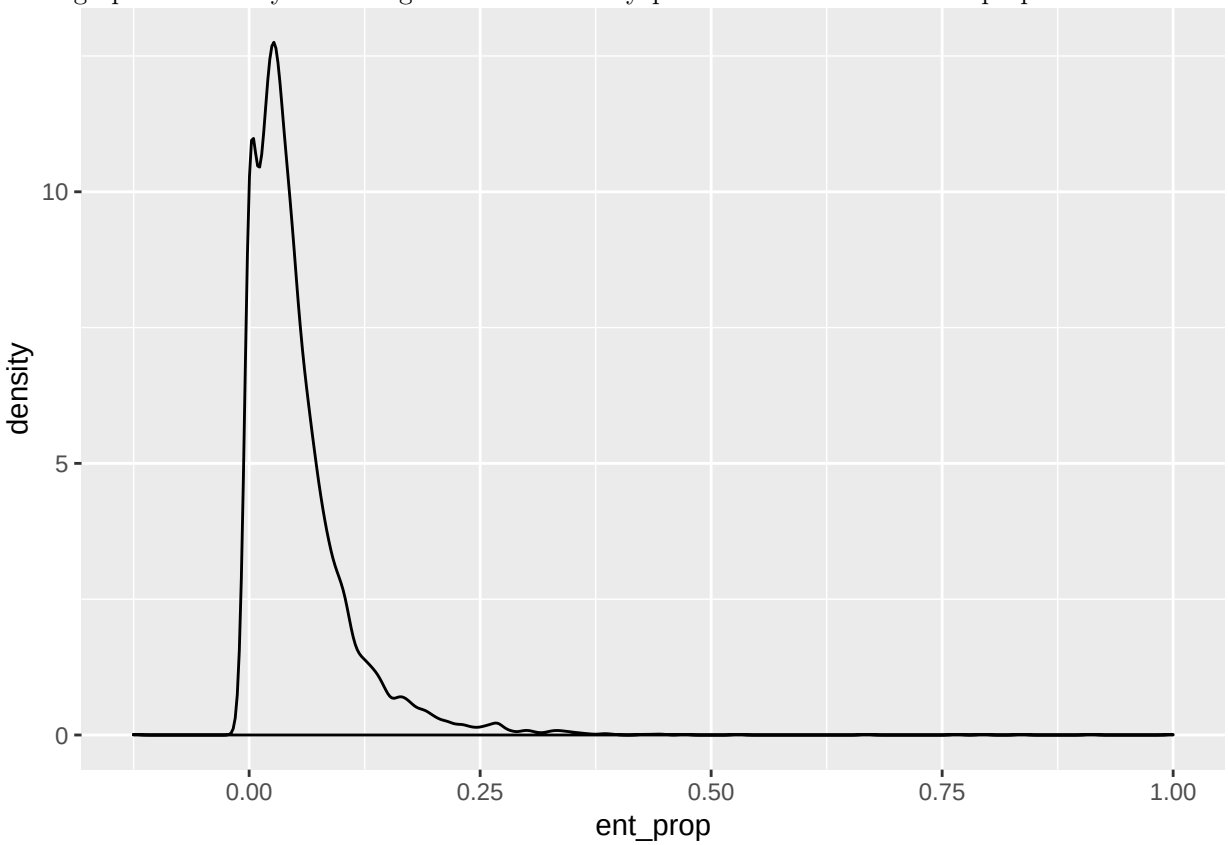
The next sub-group tested was to look at the difference of the entertainment proportion based on education level:

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Wed, Oct 17, 2018 - 08:09:56
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multi
## \hline \\[-1.8ex]
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

Based on these results, we can see that there were the most observations from those who graduated high school (factor 12), have some college (factor 13), and have a bachelors degree (factor 15), and that these groups also have higher means than most of the other categories. There is likely other factors at play that would explain the proportion spent on entertainment such as age or income level (and how that relates to education level) that could explain these variations.

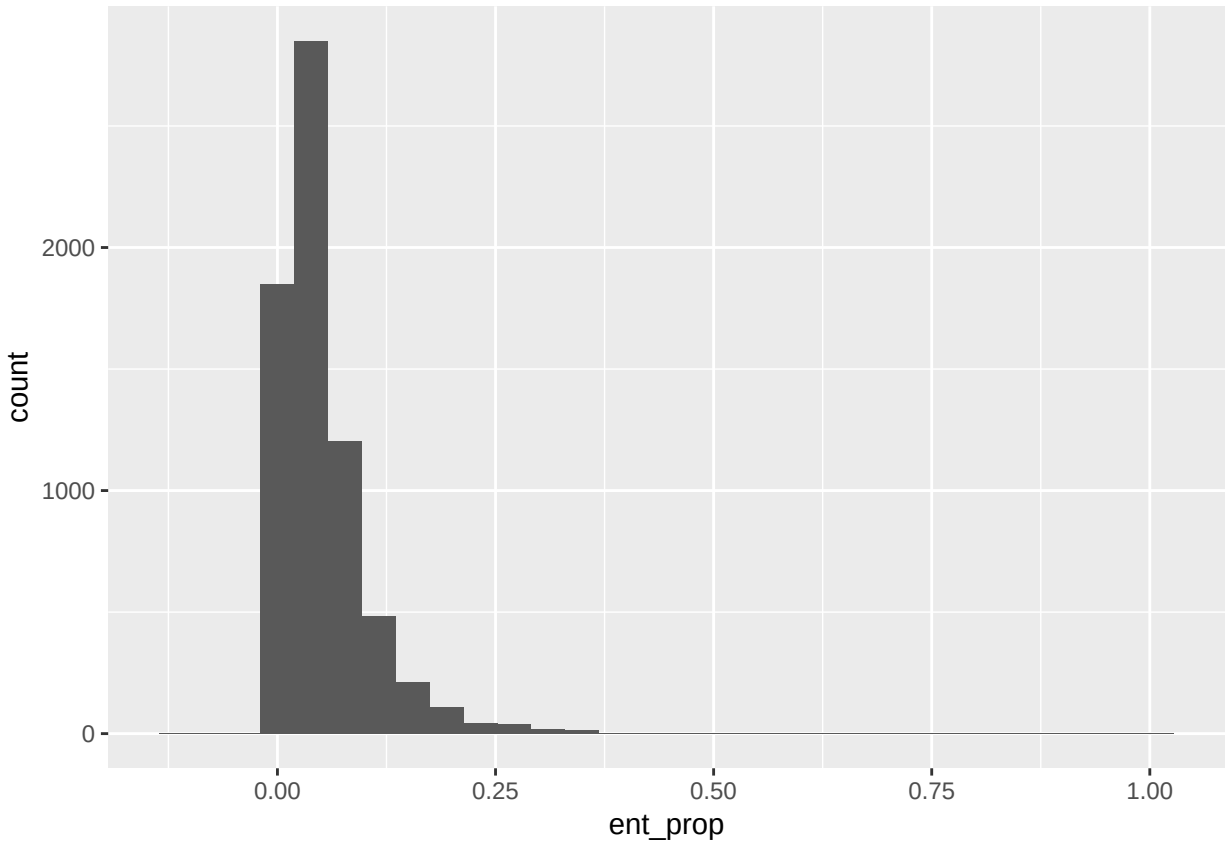
Plots

First graph we will try is looking at a basic density plot for the entertainment proportion as a whole:



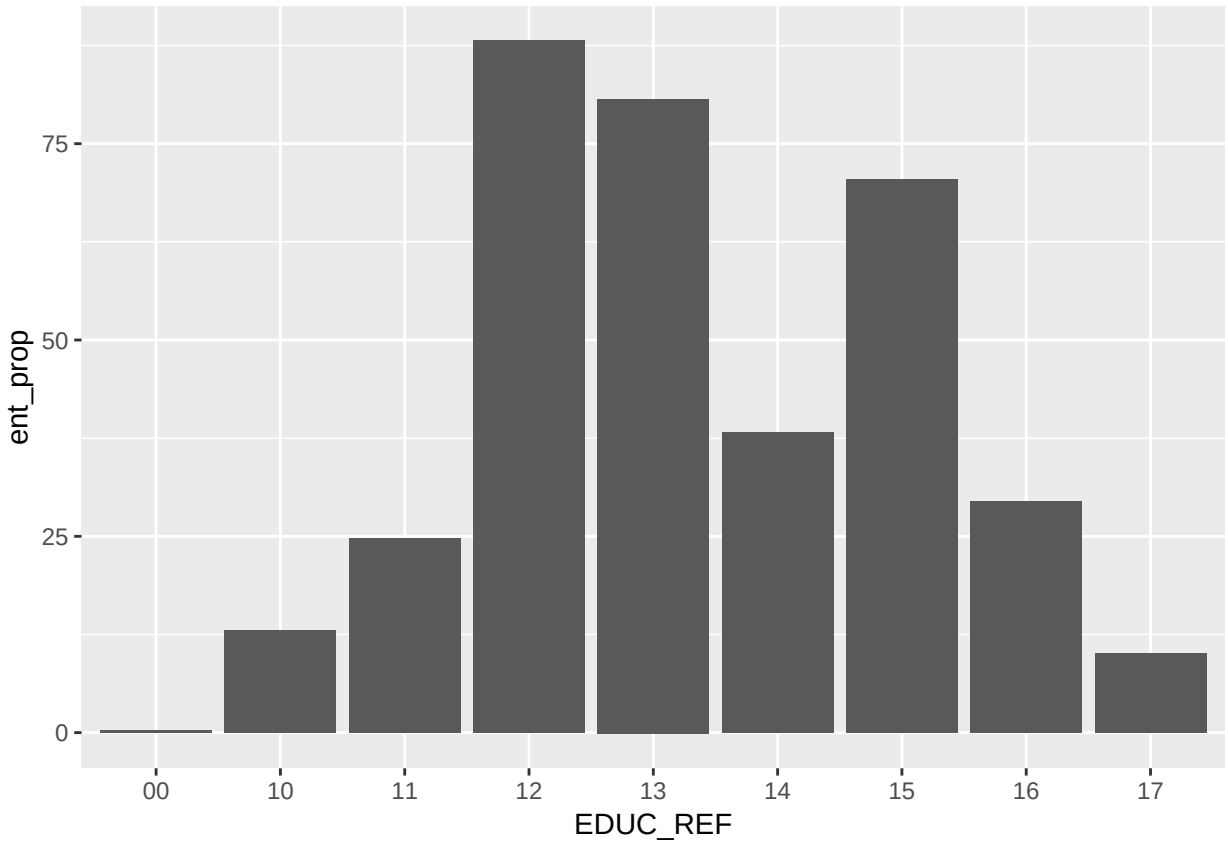
This graph is not surprising based with the sharp slope around the mean based on the results we obtained previously with the limited variation and confidence interval we found previously

Next we will try doing a histogram with the same entertainment proportion:



Again, the high levels indicated in the baskets immediately surrounding the mean are not surprising as it is reflective of the density graph previously used, but still a decent representation of the results found earlier.

Based on our subgroups we chose earlier, we will now look at the different levels of entertainment proportion with respect to education level:



This is a great way to visualize a category that has multiple factors in it such as education level because you can potentially identify likely trends or factors of interest to do further statistical analysis on to determine their statistical significance.

OLS

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Wed, Oct 17, 2018 - 08:09:58
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & ent\_prop \\
## \hline \\[-1.8ex]
##  BLS\_URBN2 & 0.004 \\
##   & (0.003) \\
##   & \\
##  educ\_nohs & 0.0001 \\
##   & (0.003) \\
##   & \\
##  educ\_hs & 0.009$^{***}$ \\
```

```
##   & (0.003) \\
##   & \\
##  educ\_smcoll & 0.014$^{***}$ \\
##   & (0.003) \\
##   & \\
##  educ\_as & 0.011$^{***}$ \\
##   & (0.003) \\
##   & \\
##  educ\_bach & 0.008$^{***}$ \\
##   & (0.003) \\
##   & \\
##  educ\_adv & 0.006$^{*}$ \\
##   & (0.003) \\
##   & \\
##  female & 0.003$^{**}$ \\
##   & (0.001) \\
##   & \\
##  ALCBEVPQ & 0.00001$^{***}$ \\
##   & (0.00000) \\
##   & \\
##  AGE\_REF & $-$0.0001$^{***}$ \\
##   & (0.00004) \\
##   & \\
##  Constant & 0.048$^{***}$ \\
##   & (0.003) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 6,838 \\
## R$^{2}$ & 0.011 \\
## Adjusted R$^{2}$ & 0.010 \\
## Residual Std. Error & 0.057 (df = 6827) \\
## F Statistic & 7.855$^{***}$ (df = 10; 6827) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

The model we came up with measures many of the basic variables you'd expect. Education, gender, age. We added in the urban/rural variable and alcoholic beverage spending. Not totally sure that's kosher since we woul dbe comparing a proportion to a proportion with the same denominator. I was really surprised to see that rural areas had a higher coefficient than urban areas. There's so much to do in cities. I'm interested in exploring that further.

kNN model for bottom quartile

```
ent_prop_25 <- quantile(clean_df$ent_prop, 0.25, na.rm = TRUE)

knn_data <- clean_df %>%
  select(ent_prop,
           BLS_URBN,
           female,
           AGE_REF
```

```
        ) %>%
  mutate(bottom25 = ent_prop < ent_prop_25) %>%
  select( -ent_prop)

good.obs <- complete.cases(knn_data)
knn_data  <- subset(knn_data, good.obs)

y.data <- knn_data

set.seed(1485)
NN_obs <- sum(good.obs)
train.obs <- (runif(NN_obs) < 0.8)

train.data <- subset(y.data, train.obs)
test.data  <- subset(y.data, !train.obs)
cl.data    <- y.data$bottom25[train.obs]
true.data  <- y.data$bottom25[!train.obs]

predicted.ent <- knn(train = train.data[-1],
                     test = test.data[-1],
                     cl = cl.data,
                     k = 3)
n.correctly.predicted <- sum(predicted.ent == true.data)
correct.rate <- n.correctly.predicted / length(predicted.ent)
print(correct.rate)
```

```
## [1] 0.9971119
```

Something must be wrong with this. The prediction rate is way too close to perfect. We tried several different variable combinations, and this was the strongest. Both education and alcohol expenditure reduced our rate of success in prediction, so we chose not to use them. We even played with values of k. Things got murkier as k got larger. Very open to your opinion on this! Are we missing something?

2. Expenditure on Apparel

Models for APPARPQ including education

1. Again with the CEX data; estimate some models to explain APPARPQ, expenditure on apparel in the previous quarter (includes MENBOYPQ, WOMGRLPQ, and FOOTWRPQ - expenditure on apparel for Men and boys; Women and girl; footwear). How important is educational attainment on this expenditure category?

```
model1 <- lm(APPARPQ ~ educ_adv + educ_as + educ_smcoll + educ_nohs + educ_hs)
summary(model1)
```

```
##
## Call:
## lm(formula = APPARPQ ~ educ_adv + educ_as + educ_smcoll + educ_nohs +
##     educ_hs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -424   -233   -148     48  52717
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   382.91      19.43  19.712  < 2e-16 ***
## educ_adv       41.10      39.85   1.031   0.3023
## educ_as       -80.44      37.40  -2.151   0.0315 *
## educ_smcoll  -150.27      29.82  -5.039 4.81e-07 ***
## educ_nohs    -237.04      40.11  -5.910 3.59e-09 ***
## educ_hs      -181.20      28.22  -6.421 1.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 843.6 on 6832 degrees of freedom
## Multiple R-squared:  0.01147,    Adjusted R-squared:  0.01074
## F-statistic: 15.85 on 5 and 6832 DF,  p-value: 1.488e-15
```

```
educ_indx <-  educ_nohs +
              2*educ_hs +
              3*educ_smcoll +
              4*educ_as +
              5*educ_bach +
              6*educ_adv
model2 <- lm(APPARPQ ~ educ_indx)
summary(model2)
```

```
##
## Call:
## lm(formula = APPARPQ ~ educ_indx)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -421   -251   -161     47  52727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  135.377     21.012   6.443 1.25e-10 ***
## educ_indx     47.609      5.937   8.019 1.24e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 844.3 on 6836 degrees of freedom
## Multiple R-squared:  0.00932,    Adjusted R-squared:  0.009175
## F-statistic: 64.31 on 1 and 6836 DF,  p-value: 1.242e-15
```

```
model3 <- lm(APPARPQ ~ FAM_SIZE + FINCATAX + HOUSPQ + Married + unmarried )
summary(model3)
```

```
##
## Call:
## lm(formula = APPARPQ ~ FAM_SIZE + FINCATAX + HOUSPQ + Married +
##     unmarried)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3574   -195    -74     53  51301
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.660e+01  2.351e+01  -1.982  0.04752 *
## FAM_SIZE     1.148e+01  7.378e+00   1.557  0.11963
## FINCATAX     1.186e-03  1.702e-04   6.968  3.5e-12 ***
## HOUSPQ       7.238e-02  3.874e-03  18.684  < 2e-16 ***
## Married      6.764e+01  2.593e+01   2.608  0.00912 **
## unmarried    2.586e+01  2.810e+01   0.920  0.35746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 809.3 on 6832 degrees of freedom
## Multiple R-squared:  0.0903, Adjusted R-squared:  0.08963
## F-statistic: 135.6 on 5 and 6832 DF,  p-value: < 2.2e-16
```

Conditional Means by Education level

2. What are conditional mean expenditure on apparel for different educational levels? What about conditional means for those who spent a non-zero amount?

```
# &gt; mean(APPARPQ[as.logical(educ_adv)])
# [1] 424.0147
# &gt; mean(APPARPQ[as.logical(!educ_adv)])
# [1] 269.3451
# &gt; mean(APPARPQ[as.logical(educ_nohs)])
# [1] 145.876
# &gt; mean(APPARPQ[as.logical(!educ_nohs)])
# [1] 295.2734
# &gt; mean(APPARPQ[as.logical(educ_bach)])
# [1] 427.8435
# &gt; mean(APPARPQ[as.logical(!educ_bach)])
# [1] 247.3224
# &gt; mean(APPARPQ[as.logical(educ_as)])
# [1] 302.4739
# &gt; mean(APPARPQ[as.logical(!educ_as)])
# [1] 280.3946
# &gt; mean(APPARPQ[as.logical(educ_smcoll)])
# [1] 232.6468
# &gt; mean(APPARPQ[as.logical(!educ_smcoll)])
# [1] 295.4017
# &gt; mean(APPARPQ[as.logical(!educ_hs)])
# [1] 309.4018
# &gt; mean(APPARPQ[as.logical(educ_hs)])
# [1] 201.7139
```

People with college degrees are more likely to spend on foods, footwear, than those who do not have college degrees.

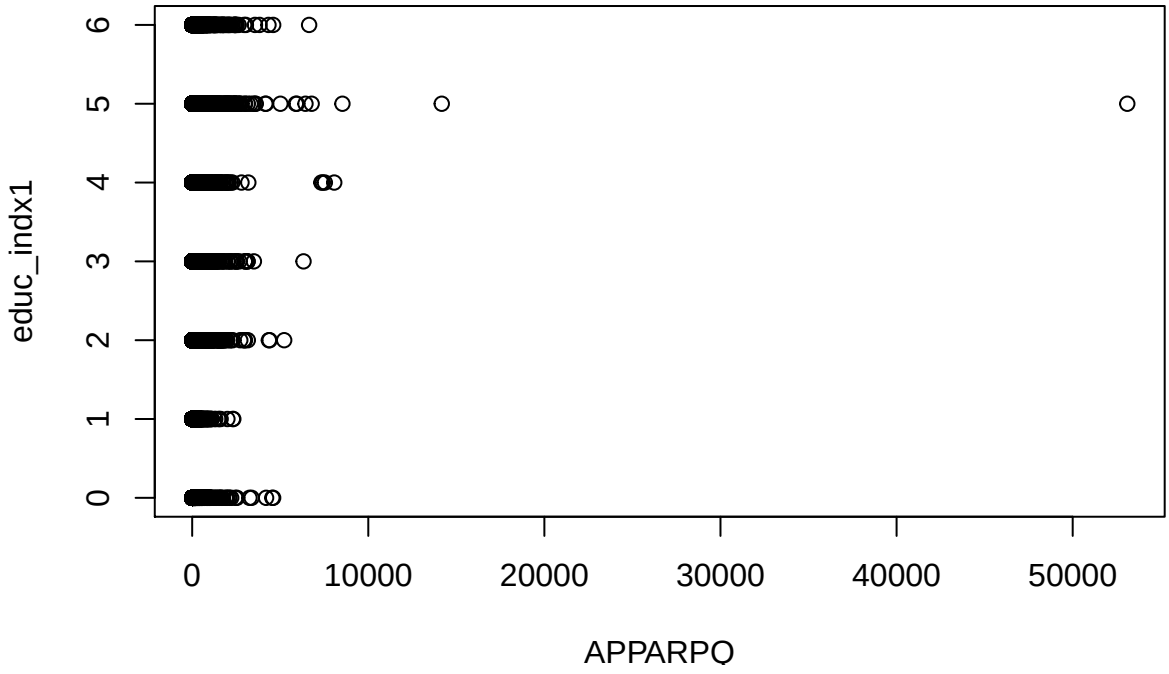
OLS models

As mentioned below we can analyse the relation between expenditure on apparel and educational attainment:

```
ols1 <- lm(APPARCQ ~ educ_nohs + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv)
summary(ols1)
```

```
##
## Call:
## lm(formula = APPARCQ ~ educ_nohs + educ_hs + educ_smcoll + educ_as +
##     educ_bach + educ_adv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -115.6  -62.5  -46.7  -20.5 5405.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   64.541      9.270   6.962 3.66e-12 ***
## educ_nohs    -26.844     12.939  -2.075 0.038049 *
## educ_hs      -17.859     10.661  -1.675 0.093946 .
## educ_smcoll   -5.972     10.946  -0.546 0.585339
## educ_as       -2.088     12.389  -0.169 0.866179
## educ_bach     36.403     11.006   3.308 0.000946 ***
## educ_adv      51.043     12.885   3.961 7.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 217 on 6831 degrees of freedom
## Multiple R-squared:  0.013,  Adjusted R-squared:  0.01214
## F-statistic:    15 on 6 and 6831 DF,  p-value: < 2.2e-16
```

```r
educ_indx1 <- educ_nohs +
              2*educ_hs +
              3*educ_smcoll +
              4*educ_as +
              5*educ_bach +
              6*educ_adv
plot(APPARPQ, educ_indx1)
```



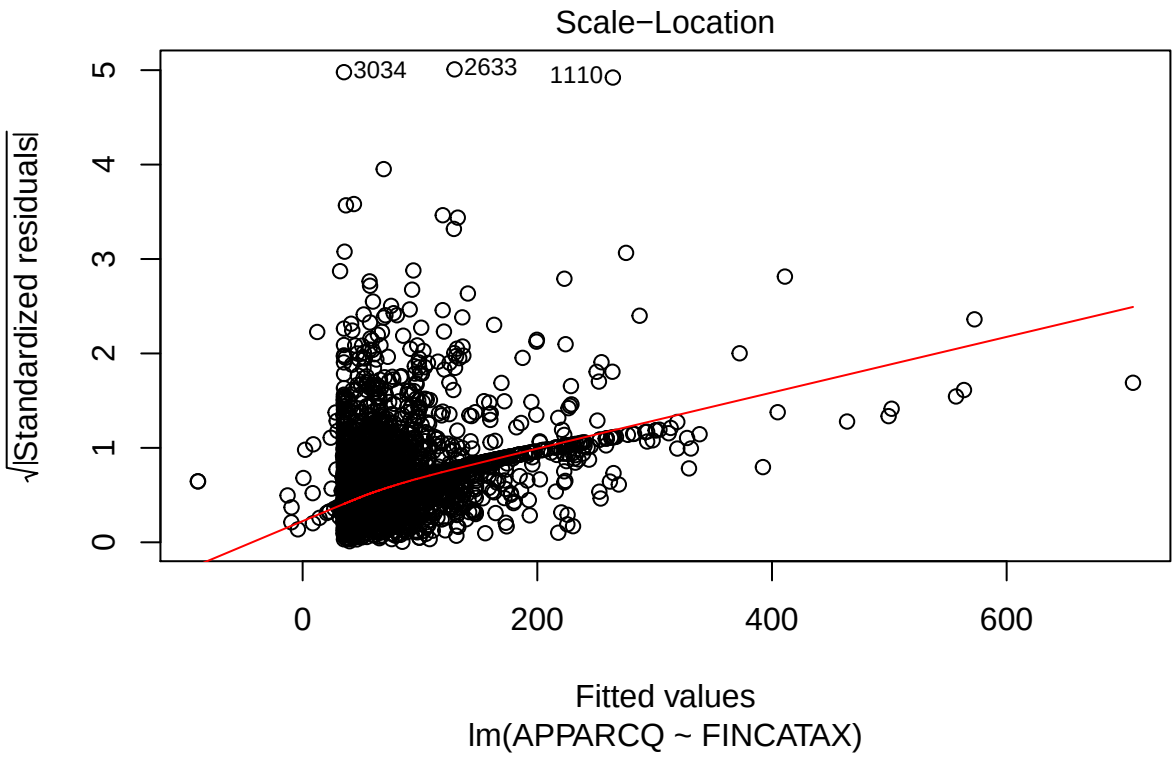
We can see that people with Bachelor's degree have highest expenditure on apparel and their expenditure is more spread than the other.

However, from our data we can observe the obvious thing - how income affects expenditure on apparel:

```
ols_INC_AP <- lm(APPARCQ ~ FINCATAX, data_cex)
summary (ols_INC_AP)
```

```
##
## Call:
## lm(formula = APPARCQ ~ FINCATAX, data = data_cex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -601.0  -65.7  -43.5  -15.4 5377.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.522e+01  3.315e+00   10.62   <2e-16 ***
## FINCATAX    6.423e-04  4.055e-05   15.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 214.5 on 6836 degrees of freedom
## Multiple R-squared:  0.0354, Adjusted R-squared:  0.03526
## F-statistic: 250.9 on 1 and 6836 DF,  p-value: < 2.2e-16
```

```
plot (ols_INC_AP, which = c(3))
```



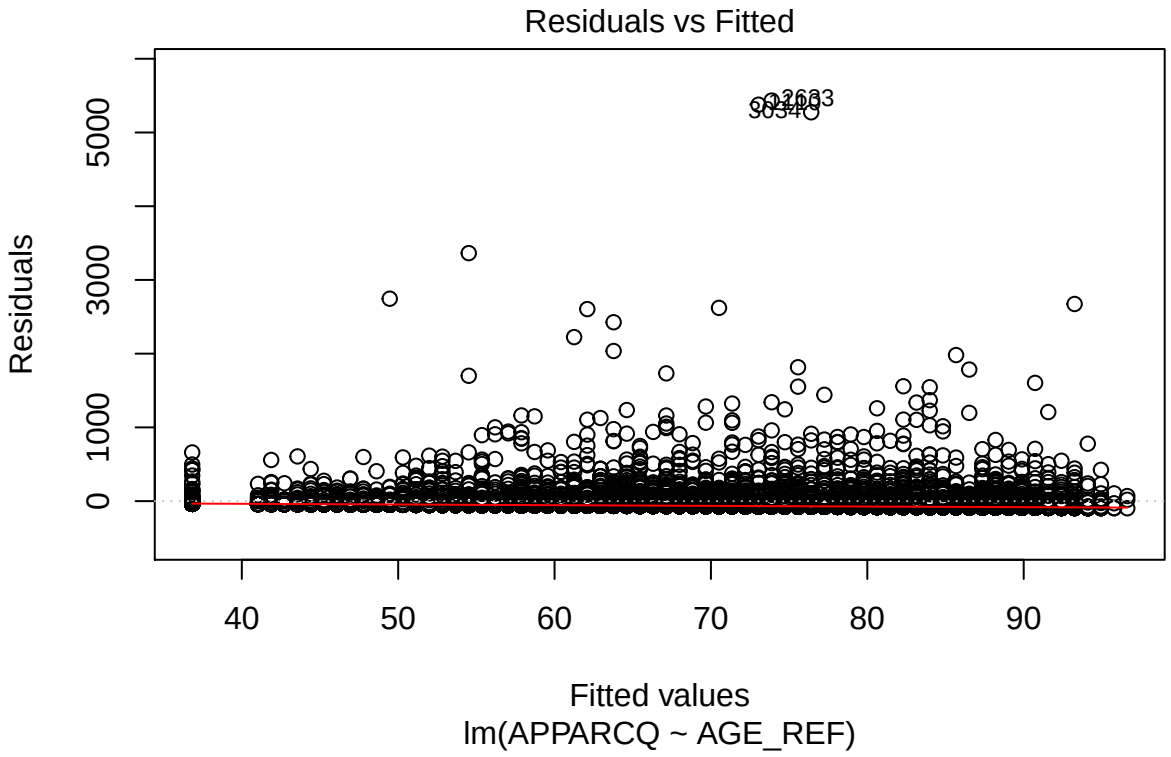
More Models

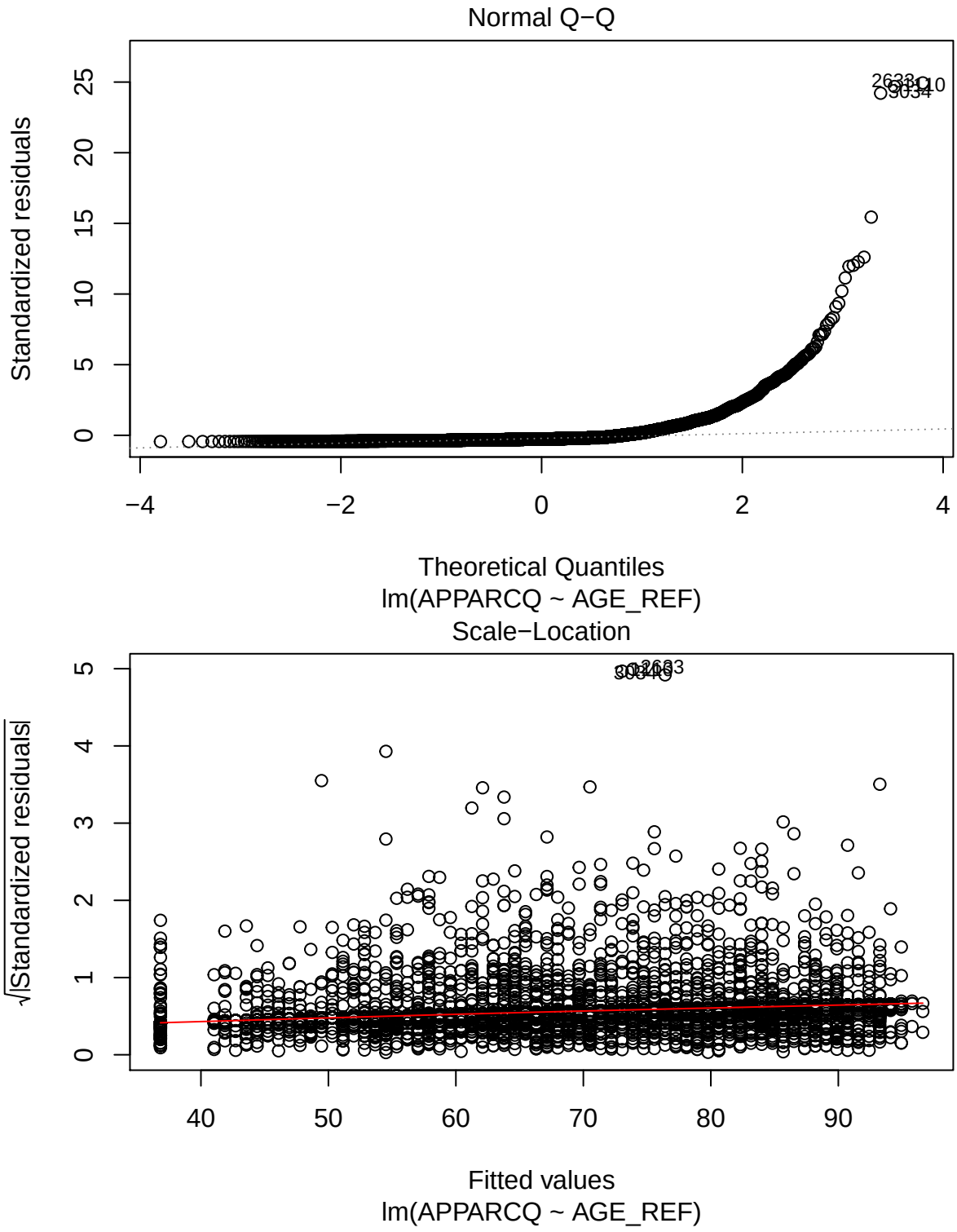
Let's estimate the relation between age and expenduture on apparel.

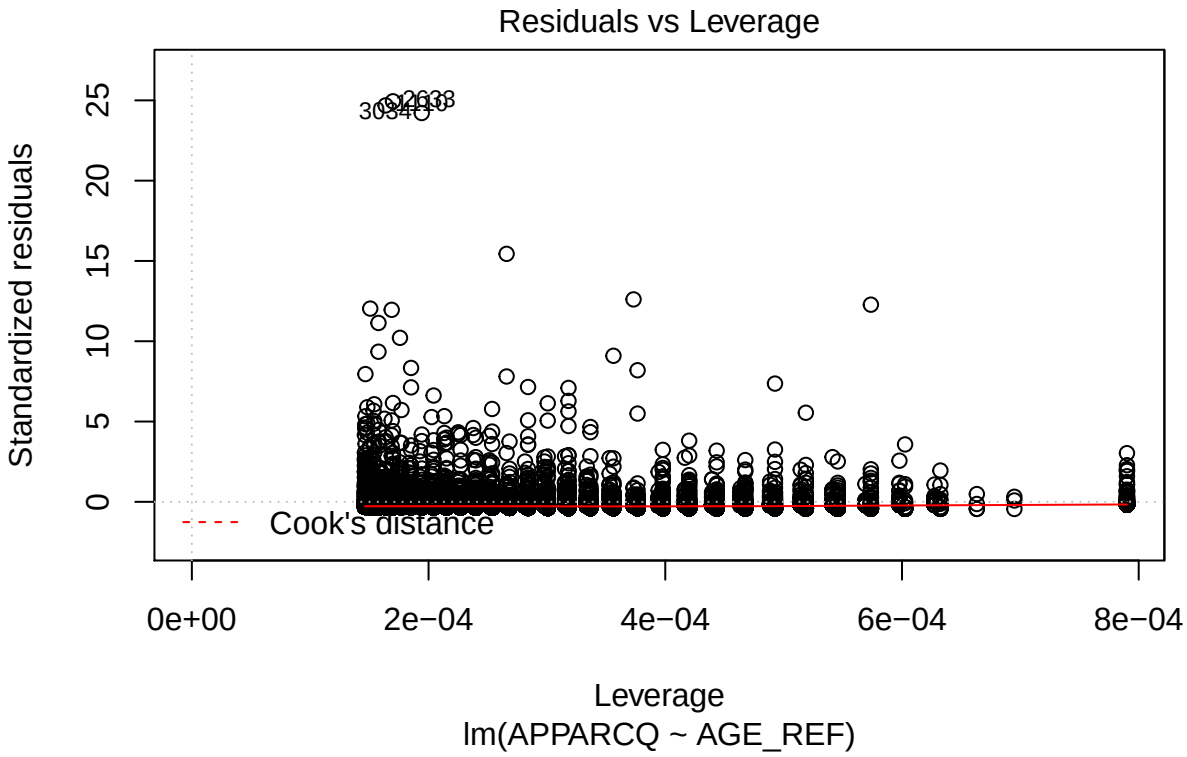
```
model_Age_Ap <- lm(APPARCQ ~ AGE_REF, data_cex)
summary(model_Age_Ap)
```

```
##
## Call:
## lm(formula = APPARCQ ~ AGE_REF, data = data_cex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -96.6  -72.2  -57.0  -23.1 5432.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 110.1001     7.9480  13.853  < 2e-16 ***
## AGE_REF      -0.8422     0.1497  -5.625 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 217.8 on 6836 degrees of freedom
## Multiple R-squared:  0.004607,   Adjusted R-squared:  0.004462
## F-statistic: 31.64 on 1 and 6836 DF,  p-value: 1.928e-08
```

```
plot(model_Age_Ap)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(APPARCQ ~ AGE_REF)
Scale−Location
√|Standardized residuals|
Fitted values
lm(APPARCQ ~ AGE_REF)



Then we can also find, for example, relation between franction of income women spent on apparel and their age:

```
woman_ap_franction <- WOMGRLCQ/FINCATAX
woman_ap_franction [is.infinite(woman_ap_franction)] <- NA
model_W_Ap_Age <- lm(woman_ap_franction ~ AGE_REF)
summary (model_W_Ap_Age)
```

```
##
## Call:
## lm(formula = woman_ap_franction ~ AGE_REF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2712 -0.0060 -0.0055 -0.0050 12.8442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.928e-03  7.027e-03   0.559    0.576
## AGE_REF     3.413e-05  1.316e-04   0.259    0.795
##
## Residual standard error: 0.1828 on 6158 degrees of freedom
##   (678 observations deleted due to missingness)
## Multiple R-squared:  1.092e-05,  Adjusted R-squared:  -0.0001515
## F-statistic: 0.06724 on 1 and 6158 DF,  p-value: 0.7954
```

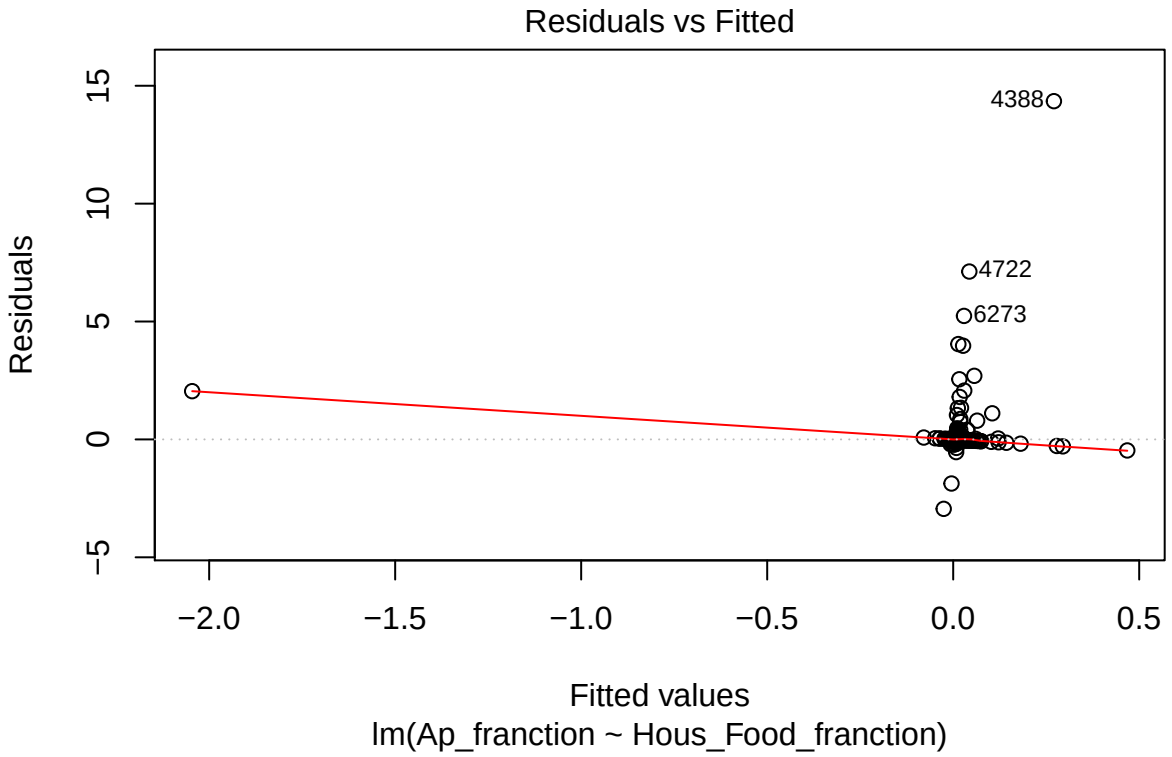
It is also interesting to see how expenditure on apparel are related with expenditure on housing and food:

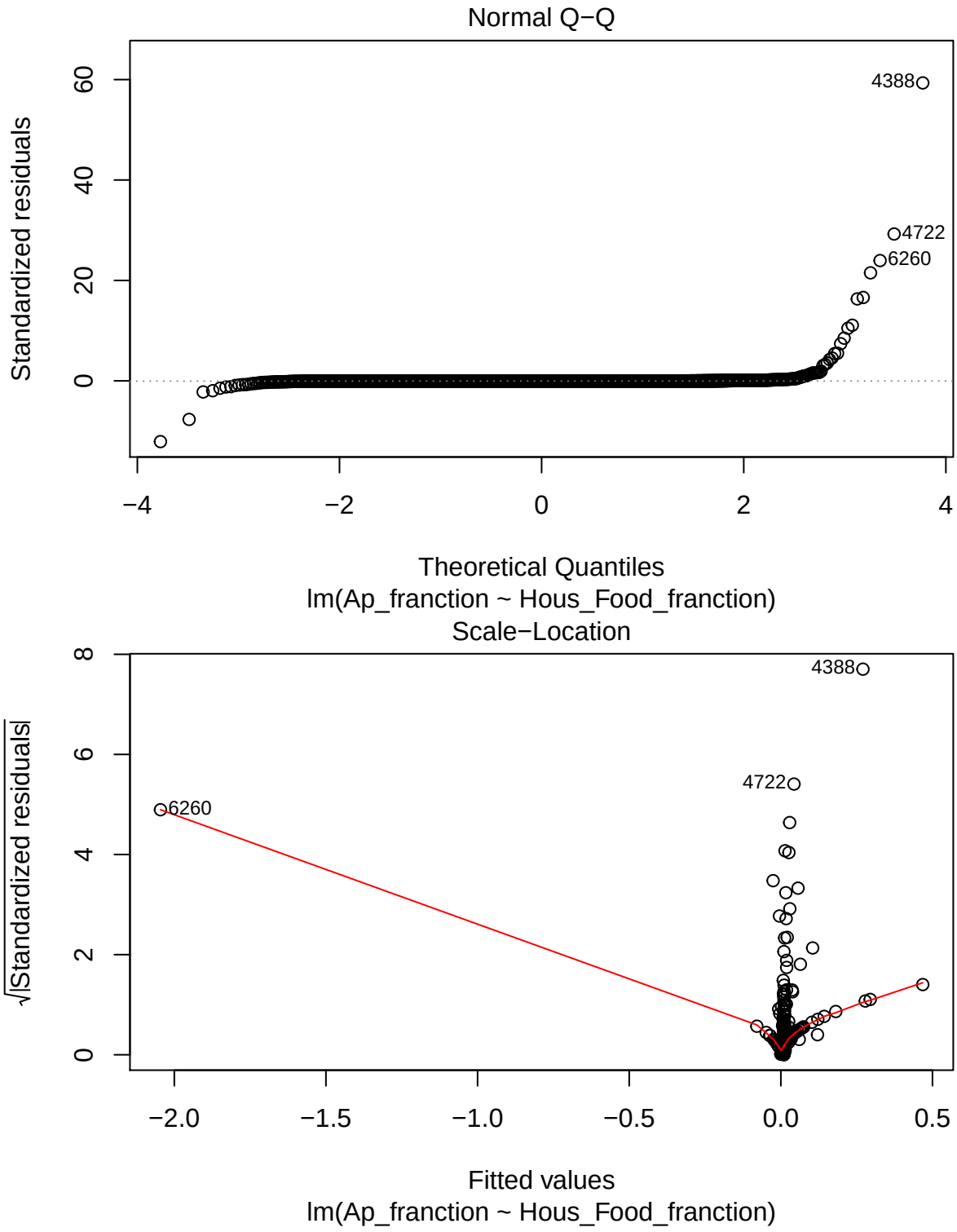
```
Ap_franction <- APPARCQ/FINCATAX
Hous_Food_franction <- (HOUSPQ + FOODPQ)/FINCATAX
Ap_franction[is.infinite(Ap_franction)] <- NA
```

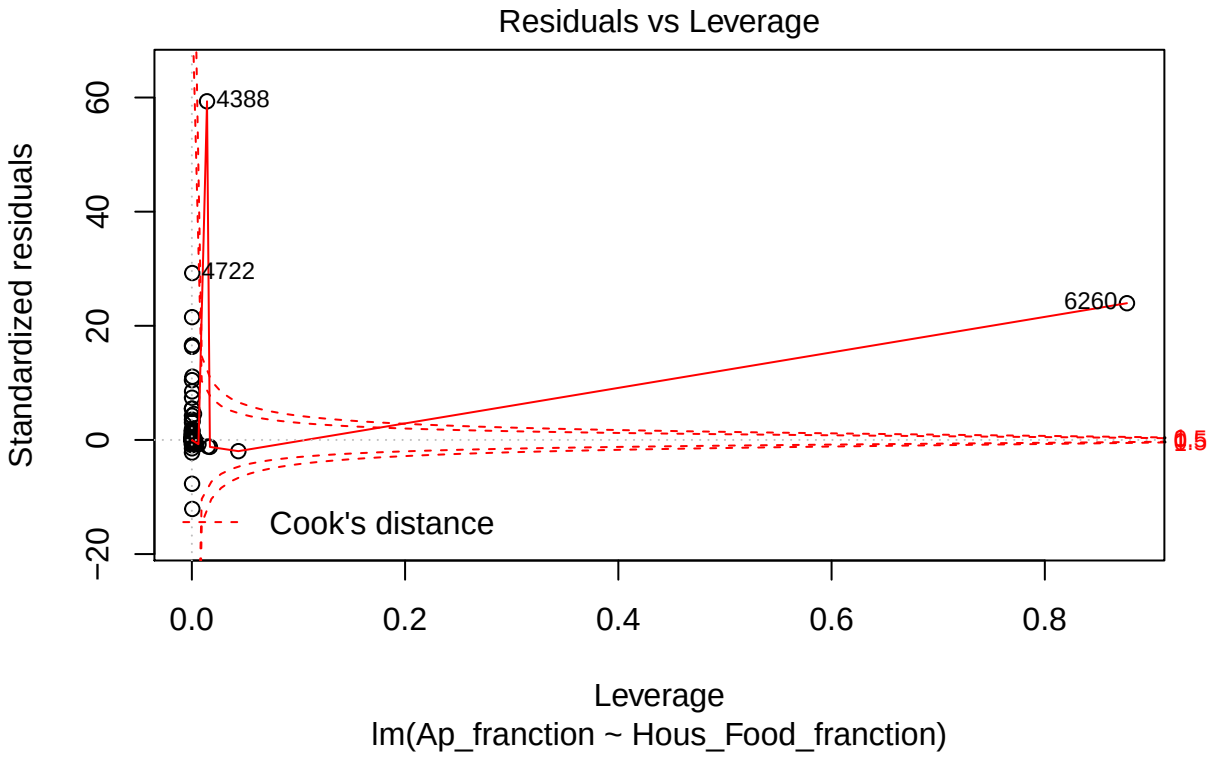
```r
Hous_Food_franction[is.infinite(Hous_Food_franction)] <- NA
model_A_HF <- lm(Ap_franction ~ Hous_Food_franction)
summary (model_A_HF)
```

```
##
## Call:
## lm(formula = Ap_franction ~ Hous_Food_franction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9448 -0.0101 -0.0101 -0.0088 14.3459
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.0100076  0.0031027   3.225  0.00126 **
## Hous_Food_franction 0.0009970  0.0001106   9.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2435 on 6158 degrees of freedom
##   (678 observations deleted due to missingness)
## Multiple R-squared:  0.01303,    Adjusted R-squared:  0.01287
## F-statistic:  81.3 on 1 and 6158 DF,  p-value: < 2.2e-16
```

```r
plot(model_A_HF)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Ap_franction ~ Hous_Food_franction)
4388
4722
6260
Scale−Location
√|Standardized residuals|
Fitted values
lm(Ap_franction ~ Hous_Food_franction)


Residuals vs Leverage
Standardized residuals
60
40
20
0
−20
4388
4722
6260
0.5
0.5
Cook's distance
0.0
0.2
0.4
0.6
0.8
Leverage
lm(Ap_franction ~ Hous_Food_franction)