261 Project - Consumer Expenditure Analysis

Junyang Liu

12/7/2016

Math 261A Class Project - Consumer Expenditure Analysis

Background Introduction & Motivation

U.S. Census Bureau official site gives much resources of quality data about the people or economy. Our Consumer Expenditure Survey data provides the information about the consumer's expenditures, buying habits, and the characteristics of those consumers.The reason why we are particularly interested in this topic is that the result of consumer expenditure performers an important role in the our real life.

For individuals, the statistics of total expenditure influence the consumers' buying habits. Of course, it also leads the tendency of consume purpose.

For companies, the result of the total expenditure statistics has a great effect on almost all of company's' investments. Sometimes, it decides whether companies will survive or not.

For Government, the result of the total expenditure statistics provides a crucial factor when the government do any decision because it is important to determine the economic performance.

Questions

In this analysis, we are interested in tackling down a combination of problems that contain both interesting trivial question and noteworthy non-trivial question. Our set of questions includes:

- What is the average American consumption level for current quarter (4th quarter 2015)?
- Is it true that if a household has more women than men, then this household is tend to spend more? If true, how much does an addition women spend compared to additional men?
- With given parameters, what is the best model to predict our expenditure? What does it tell us?
- Are these predictors a valid predictor in our best model? Does it violate any assumptions? How can we make sure our best model is valid model?
- Is there any outliers? Are they reasonable? What do they tell us?

The Process

Getting data

This data set is from Consumer Expenditure Survey, collected from U.S census bureau using Data Ferret.link. It is the Consumer Expenditure data in the fourth quarter of 2015. We picked 11 variables that contains both individual features such as age and gender, as well as household expenditure features such as home operating cost and education cost. The full list of variables and explanation is listed down below:

Variable Names	Variable Explanation
age2	Age of spouse
age_ref	Age of ref. person
as_comp1	Number of Males Over 16 In A Household
as_comp2	Number of Females Over 16 In A Household

Variable Names	Variable Explanation
bls_urbn	Rural/Urban
cshcntbx	Cash Contribution
foodcq	Food Current Quarter
housopcq	Household Operation Current Quarters
totexpcq	Total Expenditure This Quarter
educacq	Education Current Quarters
vehq	Number of Vehicle

Cleaning data & deleting unnecessary variables

The first three columns of the data are id, and system automated variables. These variables are not quite useful in our analysis, so we have decided to delete these three unnecessary variables.

```
##      NEWID           FINLWT21         QINTRVMO          AGE2
##  Min.   :512675   Min.   :   649   Min.   :10.00   Min.   :-1.00
##  1st Qu.:532589   1st Qu.: 12283   1st Qu.:10.00   1st Qu.:-1.00
##  Median :549809   Median : 18008   Median :11.00   Median :25.00
##  Mean   :548361   Mean   : 19236   Mean   :11.01   Mean   :23.37
##  3rd Qu.:564920   3rd Qu.: 24254   3rd Qu.:12.00   3rd Qu.:44.00
##  Max.   :578102   Max.   :111427   Max.   :12.00   Max.   :90.00
##     AGE_REF         AS_COMP1         AS_COMP2        BLS_URBN
##  Min.   :17.00   Min.   :0.0000   Min.   :0.000   Min.   :1.000
##  1st Qu.:33.00   1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:1.000
##  Median :44.00   Median :1.0000   Median :1.000   Median :1.000
##  Mean   :47.21   Mean   :0.8963   Mean   :1.014   Mean   :1.112
##  3rd Qu.:60.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :90.00   Max.   :6.0000   Max.   :5.000   Max.   :2.000
##     CSHCNTBX           FOODCQ          HOUSOPCQ          TOTEXPCQ
##  Min.   :   -1.00   Min.   :   0.0   Min.   :   0.00   Min.   :-2214
##  1st Qu.:   -1.00   1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.:    0
##  Median :   -1.00   Median : 260.0   Median :   0.00   Median : 1508
##  Mean   :   77.45   Mean   : 379.4   Mean   :  44.76   Mean   : 2534
##  3rd Qu.:   -1.00   3rd Qu.: 553.0   3rd Qu.:   0.00   3rd Qu.: 3444
##  Max.   :30000.00   Max.   :9244.0   Max.   :8720.00   Max.   :56644
##     EDUCACQ             VEHQ
##  Min.   :    0.00   Min.   : 0.000
##  1st Qu.:    0.00   1st Qu.: 1.000
##  Median :    0.00   Median : 2.000
##  Mean   :   22.75   Mean   : 1.939
##  3rd Qu.:    0.00   3rd Qu.: 3.000
##  Max.   :15554.00   Max.   :17.000
```

We believe that a non-positive value in our response `TOTEXPCQ` can not be applied to the real world. It cannot help us to identify the features of consumer expenditure either. Therefore,we eliminated all the responses that have a non-positive value.

We also deleted all the observations where the spouse's age `AGE2` is `-1`. The reason is that `-1` represents those people who don't have spouse. However, Since this survey is household based, we want to make sure that the household expenditure is a family effort, not a personal effort.

We also deleted cash contribution predictor `CSHCNTBX`. The reason is that most of observations of this variable are either `0` or `-1`. We believe a numeric variable contain such values for most of its observations is not good for analysis.

Lastly, we transformed variables that we believe should be factor variables into the right class. In our data analysis, the numeric variable rural/urban `BLS_URBN` is transformed into factor variable.

Feature Analysis

In this feature analysis, we are trying to answer questions:

- What is the average American consumption level for current quarter (4th quarter 2015)?
- Is it true that if a household has more women than men, then this household is tend to spend more? If true, how much does an additional women spend compared to additional men?

First question

we believe the median total expenditure is the best representation of the average U.S consumer expenditure level.

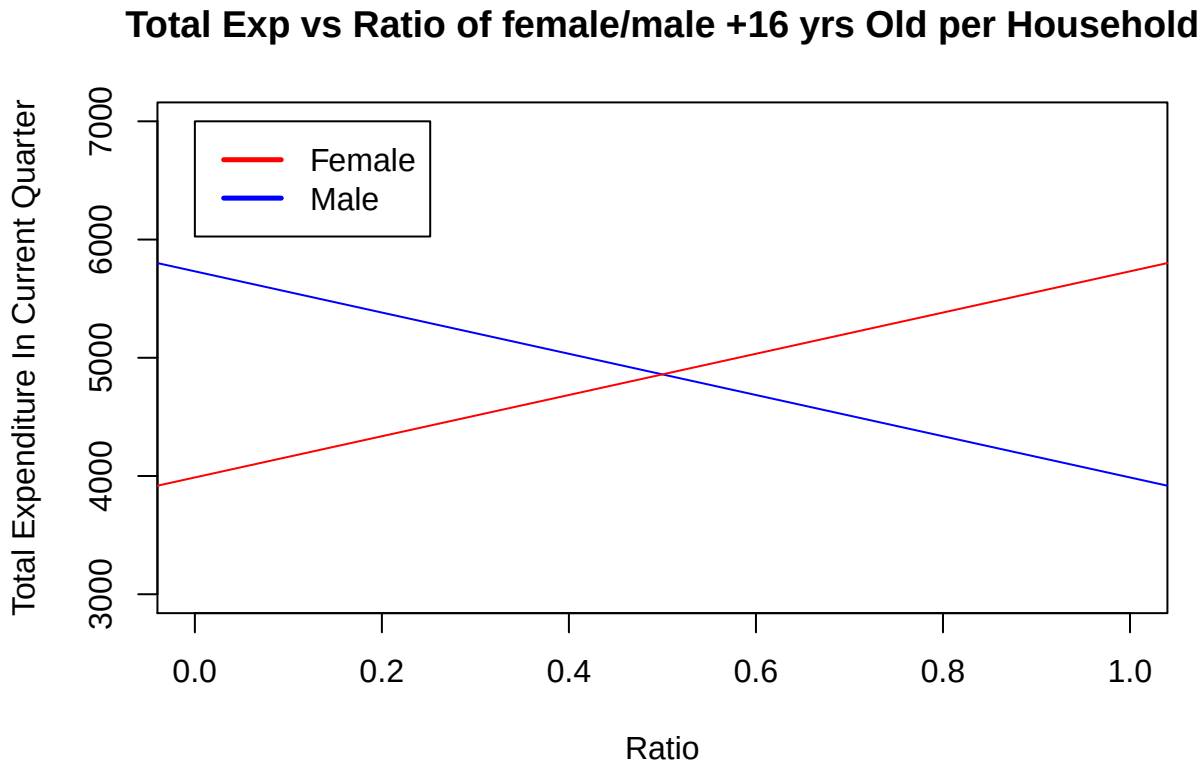
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250    2188    3582    4855    5844   56640
```

From the above results we can tell that the average American consumption level is around $3582.

Second question

Since we only have number of male over 16 in a household `AS_COMP1` and number of female over 16 in a household `AS_COMP2` two variables, we are going to create two new ratio variables `male_ratio` and `female_ratio`. In order to do that, we created a variable called `AS_COMP3`, which is the total number of people over 16 in a household to help us calculate ratio for men and women.

After creating necessary variables, we combine them to our data and fit them with Single Linear Model and the plotted results shows below.



From the above graph, we can clearly see the upward trending red line that indicates that as the ratio of female in a family goes up, the corresponding total expenditure goes up as well. On the other hand, the downward trending blue line, which indicates the male ratio in a family, decreases as the ratio of female increases. This also shows that the when male ratio in a family is low(female ratio is high), then we have higher total expenditure.

For the second part of the feature analysis question, we need to take a look at the slope of each variable. At first, we build a model called `fullmodel` with every variable included.

```
##
## Call:
## lm(formula = TOTEXPCQ ~ ., data = deleted_good_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18307  -1625   -774    499  45673
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1931.9810   396.5612   4.872 1.20e-06 ***
## AGE2          18.2509    16.9321   1.078   0.2812
## AGE_REF      -34.6026    16.5782  -2.087   0.0370 *
## AS_COMP1    -105.1484   183.9554  -0.572   0.5677
## AS_COMP2     266.7162   186.3358   1.431   0.1525
## BLS_URBN2   -551.7257   259.9479  -2.122   0.0339 *
## FOODCQ         3.5028     0.1578  22.198  < 2e-16 ***
## HOUSOPCQ       3.5534     0.3304  10.755  < 2e-16 ***
## EDUCACQ        2.4967     0.3382   7.382 2.35e-13 ***
## VEHQ         271.0673    52.2474   5.188 2.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3647 on 1834 degrees of freedom
## Multiple R-squared:  0.3477, Adjusted R-squared:  0.3445
## F-statistic: 108.6 on 9 and 1834 DF,  p-value: < 2.2e-16
```

From the full model we can interpret our variable `AS_COMP1`and `AS_COMP2` as below:

- `AS_COMP1`: With all other predictors fixed, on average, the total expenditure of a household will decrease by $105.1481 for every new male introduced in the household.
- `AS_COMP2`: With all other predictors fixed, on average, the total expenditure of a household will increase by $266.7162 for every new female introduced in the household.

Therefore, we can conclude that when introducing a female into family, the total expenditure is 371.8656 higher than introducing a male.

Exploratory Data Analysis & Find Our Best Model

```
##    p (Intercept) AGE2 AGE_REF AS_COMP1 AS_COMP2 BLS_URBN2 FOODCQ HOUSOPCQ
## 1  2           1    0       0        0        0         0      1        0
## 2  3           1    0       0        0        0         0      1        1
## 3  4           1    0       0        0        0         0      1        1
```

```
## 4  5           1    0       0        0        0         0      1        1
## 5  6           1    0       1        0        0         0      1        1
## 6  7           1    0       1        0        0         1      1        1
## 7  8           1    0       1        0        1         1      1        1
## 8  9           1    1       1        0        1         1      1        1
## 9 10           1    1       1        1        1         1      1        1
##   EDUCACQ VEHQ       SSRes        R2     AdjR2      MSE         Cp
## 1       0    0 27662681474 0.2602087 0.2598071 15017742 240.124790
## 2       0    0 25789781634 0.3102962 0.3095470 14008572 101.290092
## 3       1    0 25011761301 0.3311031 0.3300125 13593349  44.786025
## 4       1    1 24633332554 0.3412235 0.3397906 13394961  18.329672
## 5       1    1 24495445132 0.3449111 0.3431290 13327228   9.961081
## 6       1    1 24433719650 0.3465618 0.3444276 13300882   7.319567
## 7       1    1 24409127379 0.3472195 0.3447307 13294732   7.470325
## 8       1    1 24393919111 0.3476262 0.3447821 13293689   8.326724
## 9       1    1 24389574155 0.3477424 0.3445416 13298568  10.000000
```

From the output above, we notice that the best models we want to choose are #6, #7, and #8. They have smallest Cp, small SSres, and good R^2 and AdjR^2.

Assumptions

- we assume that our data is still valid after deleting all unnecessary variables and observations
- we assume the response and predictors have linear relationship
- we assume our variables are independent.
- we assume the variables we choose is to our best knowledge. they are unbiased and only variables available.

Execution

Now we have all the key parameters for our best model, which are `AGE_REF`, `BLS_URBN` ,`FOODCQ`,`HOUSOPCQ`,`EDUCACQ`, and `VEHQ`.

```
##
## Call:
## lm(formula = TOTEXPCQ ~ AGE_REF + BLS_URBN + FOODCQ + HOUSOPCQ +
##     EDUCACQ + VEHQ, data = deleted_good_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17902  -1605   -781    465  45713
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2107.6317   328.6080   6.414 1.80e-10 ***
## AGE_REF      -17.6295     5.6273  -3.133  0.00176 **
## BLS_URBN2   -559.4175   259.6833  -2.154  0.03135 *
## FOODCQ         3.5193     0.1569  22.430  < 2e-16 ***
## HOUSOPCQ       3.5386     0.3287  10.767  < 2e-16 ***
## EDUCACQ        2.4981     0.3382   7.387 2.26e-13 ***
```

```
## VEHQ         281.5805    50.3586   5.592 2.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3647 on 1837 degrees of freedom
## Multiple R-squared:  0.3466, Adjusted R-squared:  0.3444
## F-statistic: 162.4 on 6 and 1837 DF,  p-value: < 2.2e-16
```

From above we can conclude that the best model is:

- TOTEXPCQ = -17.6295 AGE_REF - 559.4175 BLS_URBN2 + 3.5193 FOODCQ + 3.5386 HOUSOPCQ + 2.4981 EDUCACQ + 281.5805 VEHQ

From this model, we can interpret our coefficient as follows:

- With other predictors fixed, on average, the total expenditure of a household will decrease by $17.6295 for every year a person gets older.
- With other predictors stay fixed, on average,the total expenditure of a household will decrease by $559.4175 when this household live in rural.
- With other predictors stay fixed, on average, the total expenditure of a household will increase by $3.5193 when the household spend one more dollar on food in current quarter.
- With other predictors stay fixed, on average, the total expenditure of a household will increase by $3.5386 when the household spend one more dollar on household operation in current quarter.
- With other predictors stay fixed, on average, the total expenditure of a household will increase by $2.4981 when the household spend one more dollar on education.
- With other predictors stay fixed, on average, the total expenditure of a household will increase by $281.5805 when the household has one more vehicle.

Multicollinearity

```
##             [,1]        [,2]        [,3]        [,4]        [,5]
## [1,]  1.00000000  0.04326091 -0.04610528 -0.03171728 -0.01993570
## [2,]  0.04326091  1.00000000 -0.09282363 -0.03968033  0.17242872
## [3,] -0.04610528 -0.09282363  1.00000000  0.12376593  0.08111350
## [4,] -0.03171728 -0.03968033  0.12376593  1.00000000  0.03472454
## [5,] -0.01993570  0.17242872  0.08111350  0.03472454  1.00000000
```

From above we can conclude that in our best model, multicollinearity among predictors does not appear to be a problem, because we have very low correlation between predictors.

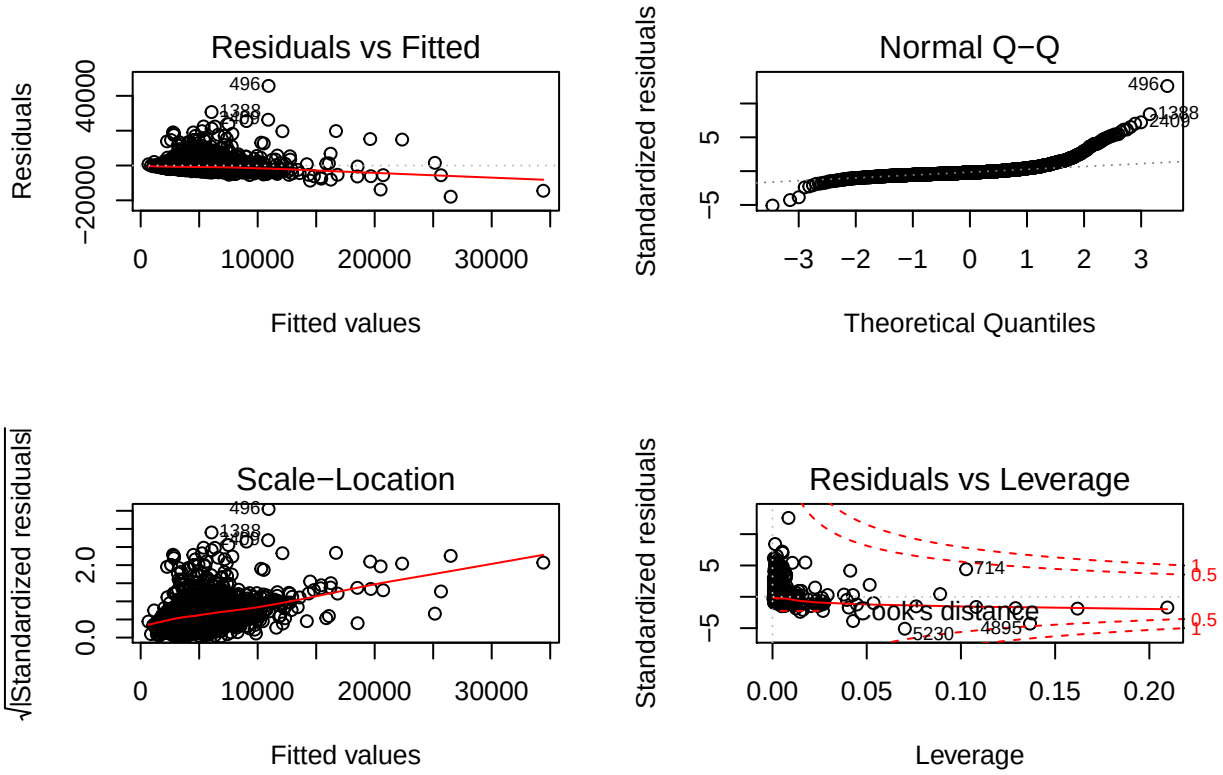
Residual Control

Residual control is an essential part to ensure our best model is valid. In a perfect world, we want to see our residual to have:

- Constant variance
- Normally distributed
- Independent

- Zero mean

When we look at residual plot, we discover that the residuals for this model do not have constant variance because it formed a triangle pattern. Ideally we want to see a cloud of residuals that does not appear to be any pattern. Also looking at QQ plot, we see the heavy tail on both sides of the qqline, which indicates that the residual does not appear to be normal. Ideally, we want to see a straight line to show that it is part of the normal distribution family.Therefore, Response transformation is needed to overcome this issue.

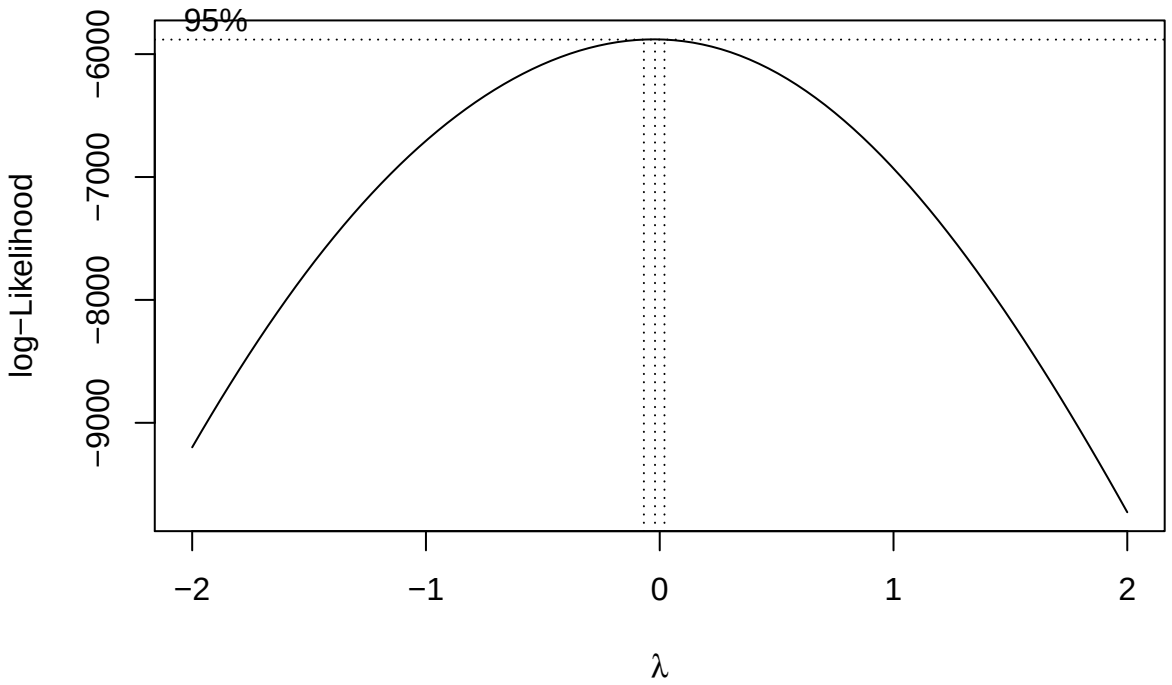


Transformation

According to Boxcox method introduced in class, the transformation method guideline follows the following table:

Lambda	Transformation needed
-2	1/yˆ2
-1	1/y
-0.5	1/sqrt(y)
0	1/ln(y)
0.5	sqrt(y)
1	y
2	yˆ2

Now let's look at our Boxcox Lambda value in our model.

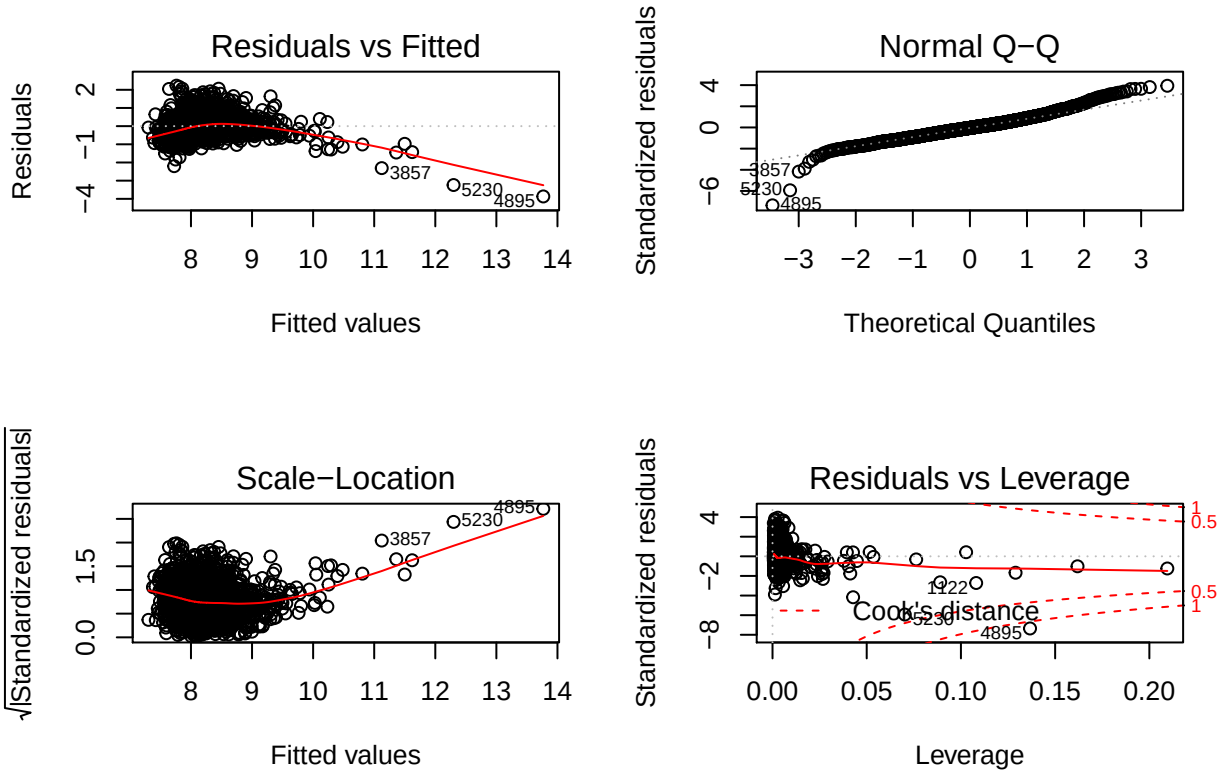


From the result above, we have found that the best transformation method is log transformation since our lambda value is very close to `0`.

Therefore, our transformed model becomes:

log(TOTEXPCQ) = 7.810 - 5.943e-03*AGE_REF - 1.295e-01*BLS_URBN2 + 6.601e-04*FOODCQ + 5.213e-04*HOUSOPCQ + 2.772e-04*EDUCACQ + 6.198e-02*VEHQ.

Then we need to verify that our transformation is indeed effective by looking at different plots for our new best model. We will examine our new residuals and outlier to confirm the validity of our new best model.



In the transformed model, we can clear see the improvements of our residual from "Residual vs Fitted" plot, as the triangle pattern disappeared. QQ plot also have better looks as the previous one has heavy tails and now mass majority of the data follow the normal distributions. Although our residuals do not appear to be perfect, we can still conclude that our residual control is good enough to make sure our best model is valid.

Validation

Since we have confirm that our best model is a valid model, we want to know can our model well predict outcomes? If our best model fail to predict outcomes, even if it is a valid model, it is not a good regression model we want to use in our real life. Therefore, we have used several methods to verify the accuracy of our model.

```
##
## Call:
## lm(formula = log(TOTEXPCQ) ~ AGE_REF + BLS_URBN + FOODCQ + HOUSOPCQ +
##     EDUCACQ + VEHQ, data = deleted_good_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8692 -0.3560 -0.0239  0.3034  2.2262
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.810e+00  5.100e-02 153.121  < 2e-16 ***
## AGE_REF     -5.943e-03  8.734e-04  -6.804 1.37e-11 ***
## BLS_URBN2   -1.295e-01  4.031e-02  -3.213  0.00134 **
## FOODCQ       6.601e-04  2.435e-05  27.105  < 2e-16 ***
## HOUSOPCQ     5.213e-04  5.101e-05  10.219  < 2e-16 ***
```

```
## EDUCACQ      2.772e-04  5.249e-05   5.282 1.43e-07 ***
## VEHQ         6.198e-02  7.816e-03   7.929 3.80e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5661 on 1837 degrees of freedom
## Multiple R-squared:  0.415,  Adjusted R-squared:  0.413
## F-statistic: 217.2 on 6 and 1837 DF,  p-value: < 2.2e-16
```

```
##
## Attaching package: 'MPV'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     stackloss
```

```
## [1] 601.1452
```

PRESS residuals (deleted residuals) are obtained from fitting a regression model with a single observation deleted. PRESS statistics is calculated as the sums of squares of the prediction residuals for those observations. In our analysis, our PRESS statistics is 601.1452. Considering we have around 2000 observation, our PRESS statistics is good for our model.

Another validation method is data-splitting method. We use half of our data from form the same model. And use this model to predict the other half of the data. We then compare our predicted value against our real observation values to see the accuracy of our model. If our model is indeed a good predicting model, we should see a small difference in the errors. In our analysis, we will use MSP vs MSE to help us make a decision.

```
## Analysis of Variance Table
##
## Response: log(TOTEXPCQ)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## AGE_REF     1  11.692  11.692  39.127 6.088e-10 ***
## BLS_URBN    1   4.657   4.657  15.583 8.501e-05 ***
## FOODCQ      1 142.511 142.511 476.900 < 2.2e-16 ***
## HOUSOPCQ    1  22.509  22.509  75.324 < 2.2e-16 ***
## EDUCACQ     1   7.355   7.355  24.612 8.359e-07 ***
## VEHQ        1   5.612   5.612  18.781 1.629e-05 ***
## Residuals 915 273.428   0.299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above table, we have found the Mean Square Error(MSE) is 0.290

```
## [1] 0.3466156
```

Compare to MSE = 0.290, our model MSP = 0.3575745, while they are not equal to each other, we can tell that they are pretty close. This means our predicting model is accurate enough to do the job. Hence we conclude that our model is a valid regression model.

Outlier Control

Find Influential Points

In our outlier analysis, we are going to use cook's distance to determine influence, and Hii matrix to determine leverage points.

Cook's distance is a measure for the influence each point in linear regression. It is computed by comparing the parameter estimates obtained when using all points and the parameter estimates obtained when deleting the i th observation. In general, a cook's distance, Di > 1 is good indicator that our observation has large influence to the model.

```
##       4783       1192       3857       1122       5230       4895
## 0.05926997 0.09774817 0.11113228 0.12798079 0.38026987 1.22386494
```

We find that position 4895 is 1.22386494. This indicate that point at position 4895 is an influential point.

Interpret Influential Point

```
##      AGE2 AGE_REF AS_COMP1 AS_COMP2 BLS_URBN FOODCQ HOUSOPCQ TOTEXPCQ
## 4895   44      45        1        1        1   9244        0    19911
##      EDUCACQ VEHQ
## 4895       0    2
```

We see that this house only have 2 people lived in and their age are 44 and 45. but their food expenditure in current quarter is the highest in the whole data set. Thus we assume this might be the reason why it is such an influential outlier.

Find Leverage Points

A leverage point is an observation, that has an unusual predictor value(very different from the bulk of the observations), but that lies on or at least very close to the regression surface determined by the rest of the data. In our analysis, we use hat matrix to identifying leverage points. The elements hii if the hat matrix may be interpreted as the leverage that the ith observation yi exerts on the jth fitted value yi_hat. The diagonal entries hii of H can be seen as a measure for how far the ith observation lies from the center of the x-space. The rule of thumb is to consider any point for which hii exceeds 2(k+1)/n a leverage point.

```
## [1] 102
```

hii > 2(k+1)/n is the formula we used to find the cutoff point. In our case, hii > 2*(6+1)/1844 = 0.007592191. We see 102 leverage points but this is only 5.5% of the total data points.

Conclusion

From all the analyses above, we have known the average U.S household expenditure level ($3582), we have confirmed that the women tend to spend more than men by analyzing the effect of increasing ratio of female in a household on total expenditure. We have also found our best model and we have gone through a chain of processes to make sure we have a valid model. Our best model contains 6 variables. Among those 6 variables, 5 of those variables are household features and only 1 variable is individual features. This finding leads us to think that the deciding factors of household expenditure might lean more towards the family financial healthiness rather than individual influence such as age or gender. Because the driver of our best models are mostly household expenditures variables. However, we need more data and evidence to perform more detailed analysis to confirm this theory.