Analysis on Consumer Expenditures in Holiday Season using Regression Methods

Terry Situ

San Jose State University

APPENDIX

INTRODUCTION

It is easy to understand the more money a household earns, the higher there expenditure will be. Beyond that, there are certainly other factors that would impact a family's overall level of expenditure. We investigated the relationship of household total expenditure with various predictors such as: age, income levels, regions, housing status, education background, marital status, vehicles owned and etc. Our dataset is from Consumer Expenditure (CE) Survey, which is run by the Bureau of Labor Statistics by the U.S. Census Bureau. The CE is essential since it is the only Federal survey that provides data on a complete range of information for consumer expenditure's and income, as well as other related consumer characteristics. Business and academic researchers use the CE to study the consumer spending habits and trends. We collected multiple independent variables that we think would possibly affecting people's expenditure. We are interested in seeing which characteristics are the ultimate underlying factors linearly predicting a household expenditure.

Consumer Expenditure can be affected by a variety of factors and choosing which variables to assign to our initial model up in our model is quite interesting and difficult. The CE Survey is an annual survey, which asks consumers throughout the nation, to record there spending habits based on 893 separate different parameters. The complete report consists of two separate surveys, diary and interview. The interview survey reports monthly personal expenditures such as: housing, apparel and services, transportation, health care, entertainment, personal care, reading, education, food, tobacco, cash contributions, and personal insurance and pensions. Seven total data sets make up the interview survey report. For the case of our analysis we will focus on the "fmli" set of data. While the diary survey reports weekly expenditure for: food at home, food away from home, alcoholic beverages, smoking supplies, personal care products and services, and nonprescription drugs. Model included 15 variables from a data set of nearly 893 variables. When we are choosing the variables, we try to cover aspects we've researched and thought would provide us the most indicative of household expenditure. On the other hand, what we can find is relationship with CE rather than causation, so the predictor chosen could not directly related to CE. The predictors we finally choose for our model are in the following table:

Table_1 Variables introduction

Name	Type	Interpretation
ETOTAPX4	NUM	Adjusted total major outlays such as food, housing, fuel, education etc, in K dollars, last quarter 2012; Y Response Variable
CUTENURE	CHAR	Housing tenure CODED: 1 Owned with mortgage 2 Owned without mortgage

		3 Owned mortgage not reported 4 Rented 5 Occupied without payment of cash rent 6 Student housing
EDUC_REF	CHAR	Education of reference person CODED 00 Never attended school 10 First through eighth grade 11 Ninth through twelfth grade (no H.S. diploma) 12 High school graduate 13 Some college, less than college graduate 14 Associate's degree (occupational/vocational or academic) 15 Bachelor's degree 16 Master's degree 17 Professional/Doctorate degree
BLS_URBN	CHAR	Urban/Rural CODED 1 Urban 2 Rural
INCLASS	CHAR	Income class of CU based on income before taxes. CODED: 01 Less than $5,000 02 $5,000 to $9,999 03 $10,000 to $14,999 04 $15,000 to $19,999 05 $20,000 to $29,999 06 $30,000 to $39,999 07 $40,000 to $49,999 08 $50,000 to $69,999 09 $70,000 and over
MARTIAL1	CHAR	Marital status of reference person CODED 1 Married 2 Widowed 3 Divorced 4 Separated 5 Never married
REGION	CHAR	Region CODED 1 Northeast 2 Midwest 3 South 4 West
VEHQ	NUM	Number of owned vehicles
FAM_SIZE	NUM	Number of members in Consumer Unit (CU)
FINCATAX	NUM	Amount of CU income after taxes in the past 12 months. In K dollars
AGE_REF	NUM	Age of reference person
NUM_AUTO	NUM	Number of owned automobiles

RENTEQVX	NUM	To rent out your home today, how much do you would charge rent for monthly, unfurnished and without utilities
INC_RANK	NUM	Weighted cumulative percent income ranking of CU to total population. Ranking based on income before taxes for complete reporters. Rank of incomplete income reporters is set to zero.
NO_EARNR	NUM	Number of earners

DATA CLEAN-UP

One of the around 6000 CE values are negative, which is worth mentioning. One element of CE is the expense (or return) of mortgage, and probably this observation is taking a reverse mortgage and gets paid from the bank more than he/she spent. Our dataset originally contained 6389 observations. However, though we chose variables with the least amount missing values as possible, we still have thousands of missing values. (44.7% of the data was removed in this process) Although, there may be potential dependence between the present observations and missing observations, due to the scope of our current knowledge, we decide to assume the missing observations have no dependence on the observations we choose to select. We decided to delete all the observations with any missing values. Thus, 3560 observations remained in our dataset.
After choosing variables and eliminating missing values, we examine variance inflation factors for all variables.

PRELIMINARY ANALYSIS

The conventional upper bound for an acceptable VIF (variance inflation factor) value is 10, and in this case all of our VIF values are smaller than 10. Therefore, we assume there is no issue of multicollinearity. So we attempted to fit the full regression model.

Table_2 VIF of variables

Variable	VIF	Variable	VIF
fam_size	1.773999	no_earnr	2.187233
vehq	1.515911	income.f	5.942470
fincatax	2.620526	bls.f	1.062047
age_ref	1.805348	cuten.f	1.349774
num_auto	1.334925	educ_ref.f	1.347094
renteqvx	1.386275	marital.f	1.368273
inc_rank	8.839525	region.f	1.016481

FULL MODEL

From the outputs, we can see an adjusted R^2 value of (0.3659). Several of our predictors have p-values larger than 0.1, indicating these variables are not significant to the model.

We then implemented three variable selection methods: forward, backward and stepwise method.

```
Residuals:
    Min      1Q  Median      3Q     Max
-35.408  -3.795  -0.659   2.601 143.824

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.294e+00  5.698e+00  -0.227 0.820344
vehq         5.357e-01  1.094e-01   4.897 1.02e-06 ***
fam_size     1.785e-01  1.204e-01   1.483 0.138144
fincatax     2.943e-05  2.938e-06  10.017  < 2e-16 ***
age_ref      1.329e-03  1.168e-02   0.114 0.909364
num_auto    -1.277e-01  1.735e-01  -0.736 0.461965
renteqvx     2.835e-03  1.689e-04  16.784  < 2e-16 ***
inc_rank     5.092e+00  1.488e+00   3.423 0.000627 ***
no_earnr     3.467e-01  2.053e-01   1.689 0.091390 .
income.f1    2.028e+00  1.421e+00   1.427 0.153588
income.f2    1.679e+00  1.409e+00   1.192 0.233436
income.f3    4.387e-01  1.170e+00   0.375 0.707642
income.f4    5.357e-01  1.069e+00   0.501 0.616295
income.f5    9.013e-01  8.763e-01   1.029 0.303772
income.f6   -2.734e-01  7.465e-01  -0.366 0.714241
income.f7    8.059e-02  6.716e-01   0.120 0.904489
income.f8   -2.436e-01  4.872e-01  -0.500 0.617131
bls.f1      -1.229e+00  5.645e-01  -2.178 0.029475 *
cuten.f1     1.150e+00  3.152e-01   3.647 0.000269 ***
educ_ref.f1 -1.615e+00  5.566e+00  -0.290 0.771654
educ_ref.f2 -1.634e+00  5.549e+00  -0.294 0.768455
educ_ref.f3 -1.190e+00  5.530e+00  -0.215 0.829640
educ_ref.f4 -5.649e-01  5.533e+00  -0.102 0.918680
educ_ref.f5 -9.586e-01  5.540e+00  -0.173 0.862633
educ_ref.f6  3.119e-01  5.535e+00   0.056 0.955066
educ_ref.f7 -1.491e-01  5.544e+00  -0.027 0.978543
educ_ref.f8  1.862e+00  5.575e+00   0.334 0.738427
marital.f1   7.780e-01  4.836e-01   1.609 0.107719
marital.f2   8.081e-01  6.337e-01   1.275 0.202312
marital.f3   8.366e-02  5.446e-01   0.154 0.877914
marital.f4  -5.371e-01  1.074e+00  -0.500 0.617124
region.f1    1.204e+00  4.222e-01   2.852 0.004373 **
region.f2    8.103e-01  4.139e-01   1.958 0.050330 .
region.f3    9.866e-01  3.761e-01   2.623 0.008754 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.781 on 3501 degrees of freedom
Multiple R-squared:  0.3718,    Adjusted R-squared:  0.3659
F-statistic: 62.79 on 33 and 3501 DF,  p-value: < 2.2e-16
```

VARIABLE SELECTION

When we manually conducted variable selection process, using R output, we noticed most predictors were very significant because almost all predictors have very small p-values. Hence, we also applied automated variable selection methods and compared models. Both results were nearly close.

After analyzing all above methods of variable selection, we saw that forward selection, backward selection and stepwise selection method obtained identical results with exact values of R^2 adjusted, AIC and BIC. Therefore, according to our results, one can conclude that any selection method is appropriate to prefer from forward selection, backward selection and stepwise selection since they have exact same results.

- The model obtained after applying regression subset selection is

$etotapx4 = \beta_0 + \beta_1 * vehq + \beta_2 * fam\_size + \beta_3 * fincatax + \beta_4 * renteqvx + \beta_5 * inc\_rank + \beta_6 * cuten.f + \beta_7 * educ\_ref.f.$

- The model obtained after applying forward selection process is

$etotapx4 = \beta_0 + \beta_1 * fincatax + \beta_2 * renteqvx + \beta_3 * inc\_rank + \beta_4 * vehq + \beta_5 * cuten.f + \beta_6 * educ\_ref.f + \beta_7 * fam\_size + \beta_8 * region.f + \beta_9 * bls.f + \beta_{10} * no\_earnr$

- The model obtained after applying backward selection process is

$etotapx4 = \beta_0 + \beta_1 * vehq + \beta_2 * fam\_size + \beta_3 * fincatax + \beta_4 * renteqvx + \beta_5 * inc\_rank + \beta_6 * no\_earnr + \beta_7 * bls.f + \beta_8 * cuten.f + \beta_9 * educ\_ref.f + \beta_{10} * region.f$

- The model obtained after applying stepwise selection process is

$etotapx4 = \beta_0 + \beta_1 * educ\_ref.f + \beta_2 * inc\_rank + \beta_3 * fincatax + \beta_4 * renteqvx + \beta_5 * vehq + \beta_6 * cuten.f + \beta_7 * fam\_size + \beta_8 * region.f + \beta_9 * bls.f$

FINAL MODEL

```
> summary(model.backward)

Call:
lm(formula = etotapx4 ~ vehq + fam_size + fincatax + renteqvx +
    inc_rank + no_earnr + bls.f + cuten.f + educ_ref.f + region.f,
    data = fmly)

Residuals:
   Min     1Q Median     3Q    Max
-36622  -3881   -624   2569 143974

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.689e+03  9.284e+02  -1.820 0.068898 .
vehq         5.466e+02  9.678e+01   5.648 1.75e-08 ***
fam_size     2.389e+02  1.083e+02   2.205 0.027497 *
fincatax     3.128e-02  2.773e-03  11.283  < 2e-16 ***
renteqvx     2.876e+00  1.672e-01  17.201  < 2e-16 ***
inc_rank     3.983e+03  7.990e+02   4.985 6.49e-07 ***
no_earnr     2.811e+02  1.933e+02   1.454 0.146015
bls.f1      -1.311e+03  5.610e+02  -2.336 0.019525 *
cuten.f1     1.014e+03  2.964e+02   3.420 0.000634 ***
educ_ref.f2 -2.141e+00  8.252e+02  -0.003 0.997930
educ_ref.f3  3.246e+02  7.211e+02   0.450 0.652602
educ_ref.f4  9.261e+02  7.410e+02   1.250 0.211469
educ_ref.f5  5.233e+02  7.912e+02   0.661 0.508362
educ_ref.f6  1.839e+03  7.455e+02   2.467 0.013681 *
educ_ref.f7  1.390e+03  8.100e+02   1.716 0.086247 .
educ_ref.f8  3.426e+03  9.901e+02   3.461 0.000546 ***
educ_ref.f9  7.990e+02  5.549e+03   0.144 0.885511
region.f1    1.247e+03  4.217e+02   2.957 0.003128 **
region.f2    8.551e+02  4.132e+02   2.070 0.038563 *
region.f3    1.060e+03  3.747e+02   2.828 0.004704 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7782 on 3515 degrees of freedom
Multiple R-squared:  0.3692,    Adjusted R-squared:  0.3657
F-statistic: 108.3 on 19 and 3515 DF,  p-value: < 2.2e-16
```

From the outputs, we can conclude that most predictors are insignificant because their p-values are higher that alpha level. There are very few predictors which are significant such as number of vehicles owned, amount of income after taxes, price of the rent, ranking based on income before taxes, are consumers urban/rural, housing tenured, and coded region. Moreover, mean square error is 60.543961, and R^2 is 37.18% which is moderately low amount. Furthermore, p-value is < 2.2e-16; hence, we then conducted variable selection methods in order to determine the most significant predictors. We applied regression subset, forward selection, backward selection and stepwise selection method.

We actually obtained identical results from all these three methods with exact same values of adjusted r-squared, AIC and BIC.

RESIDUALS ANALYSIS OF FINAL MODEL

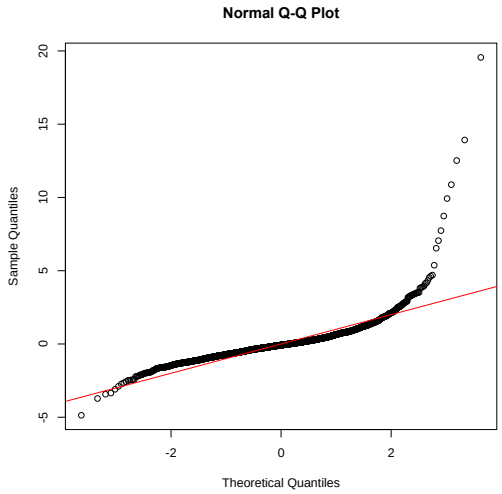


Figure_1 Q-Q plot of full model

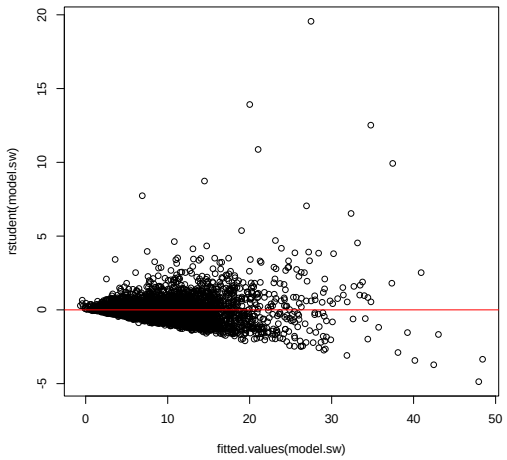


Figure_2 Residuals vs fitted values of full model

The normally probability plot of our final model (the stepwise regression model) exhibit patterns associated with positive skew, which might indicates the error terms are not normally distributed. In addition, the plot of externally studentized residuals against the fitted value shows the outward-opening funnel pattern, which implies that the variance is an increasing function of our response variable. As the result, we are convinced that the assumptions of constant and normally distributed disturbance terms are violated, and further transformation on our response should be applied.

TRANSFORMATIONS

Since there are enough evidences to convince that our response is not normally distributed nor it is constant, we wish to transform our response to correct non-normality and non-constant. A useful transformation we learned in our class is power transformation y^{λ}, where λ is a parameter to be determined. The method to obtain λ is called "Box-Cox" method.

The result we obtain by employing "Box-Cox" method is that the λ falls in to the range between -0.06 and -0.02, which is closed to 0. Hence, we decided to perform a logarithmic transformation on the response.

Now the residuals are normally distributed and became relative constant. In addition, the coefficients for the quantitative variables and the intercept are

significant. More important, as shown in the table below, the 5 possible criteria are better now, comparing to the non-transformed model.

Table_3 Analysis result of model with and without transformation

	R-Square	MSE	AIC	BIC	PRESS Statistic
Model w/ o Transformation	0.3692	60.56186	24560.3	24689.88	216708.7
Model w/ Transformation	0.4706049	0.2595201	5285.403	5414.983	922.8652

OUTLIER ANALYSIS

We find out more than 100 outliners and leverage points, respectively. However, we do not find out any influential point based on the criteria of Cook's Distance.

The transformed model is re-fitted after deleting 6 observations with largest absolute standard residuals and 6 observations with highest leverage points.

Talbe_6 Analysis result of model with and without outliers

	R-Square	MSE	AIC	BIC	PRESS Statistic
Model w/ outliners and leverage points	0.4706049	0.2595201	5285.403	5414.983	922.8652
Model w/o outliners and leverage points	0.4707473	0.2593874	5265.734	5395.242	918.2794

Based on the 5 criteria, our model seems only has minimal impacts after the outliners and leverage points are deleted. However, since we have a very large size of observations, there would be more impacts if more outliners and leverage points are deleted.

COMPARSION AND ANALYSIS OF RESIDUAL PLOTS

For the residual analysis we have compared the full model and final transformed model for comparison.

FULL MODEL

$$etotapx4=\beta_0+\beta_1 vehq + \beta_2 fam\_size + \beta_3 fincatax + \beta_4 age\_ref + \beta_5 num\_auto + \beta_6 renteqvx + \beta_7 inc\_rank + \beta_8 no\_earnr+\beta_9 income.f+\beta_{10} bls.f+\beta_{11} cuten.f+\beta_{12} educ\_ref.f+\beta_{13} marital.f+\beta_{14} region.f+\varepsilon$$

FINAL TRANSFORM MODEL

$$etotapx4=\beta_0+\beta_1 vehq + \beta_2 fam\_size + \beta_3 fincatax + \beta_4 renteqvx + \beta_5 inc\_rank + \beta_6 no\_earnr+\beta_7 income.f+\beta_8 cuten.f+\beta_9 educ\_ref.f+\beta_{10} marital.f+\beta_{11} region.f+\varepsilon$$

COMPARSION OF RESIDUALS PLOTS

Figure_3 to figure_6 are residual compared plots between full model and final model.

Upper ones are from full model, while lower ones are from our final model with log transformation. (Predictors are “renteqvx”, “inc_rank”, “vehq”, “fam_size”).

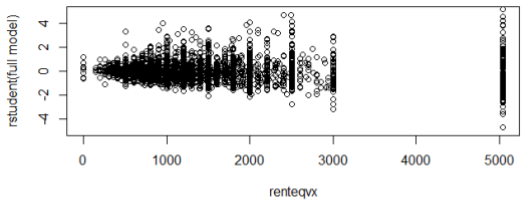
Figure_3: the plot of residuals against the renteqvx has a funnel pattern before transformation, which means nonconstant variance existed. After the log transformation, it looks more satisfy and normal.

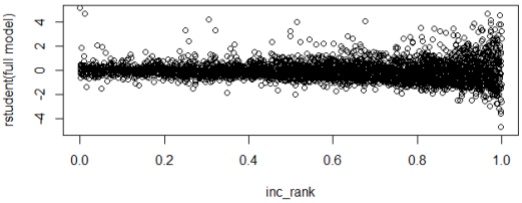
Figure_4: the plot of residuals against the inc_rank also has a funnel pattern in the full model. Thus, transformation should be considered. We took a log transformation, and the residual looks better, no pattern any more.

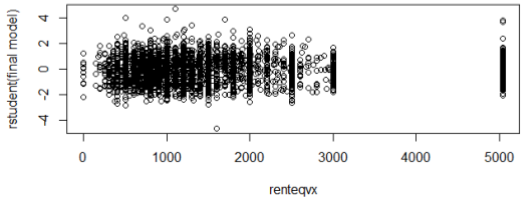
Figure_5 and figure_6: we can see that residual plot looks more random and satisfied.

Figure_7: is compared Q-Q plots between two models. Left is from full model and right is from our final model. No doubts, the model fits data very well after transforming. On the Q-Q plot, the resulting points lie approximately on a straight line, differently from full model which has light-tailed distribution.

Figure_8: two plots of residuals against the fitted values of two models. It’s obviously that the left plot has a funnel pattern. After the log transformation, the residuals are more random.







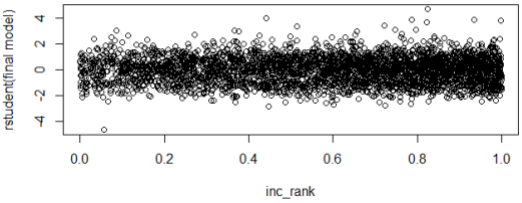
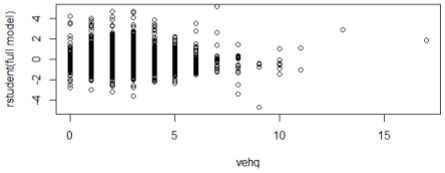
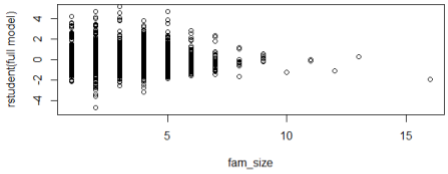
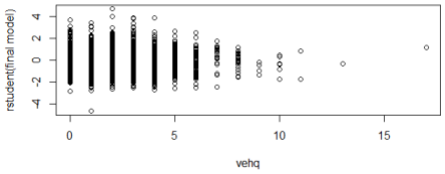


Figure_3 Residuals vs Renteqvx

Figure_4 Residuals vs inc_rank







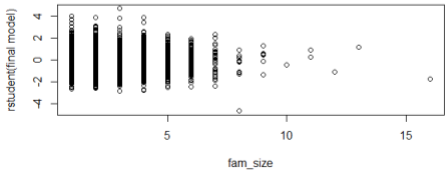


Figure_5 Residuals vs vehq

Figure_6 Residuals vs fam_size

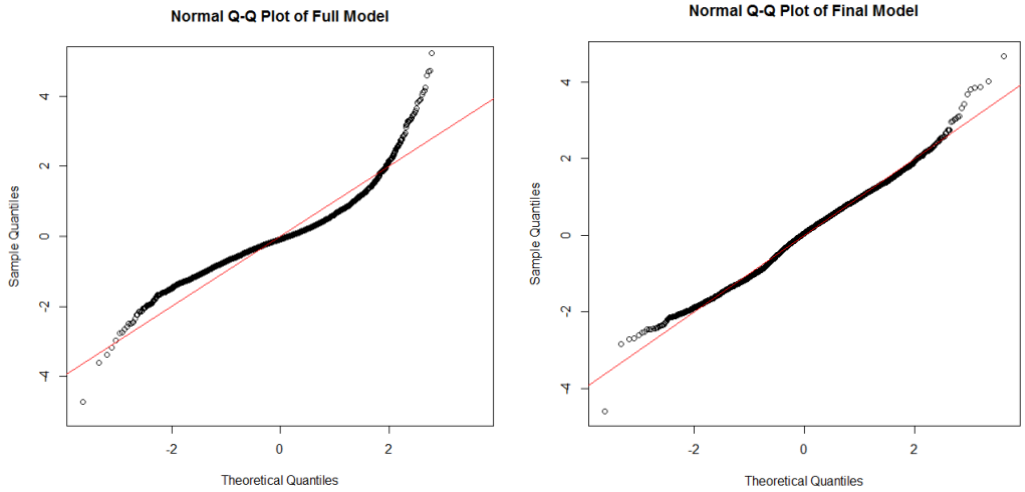


Figure_7 Q-Q plot

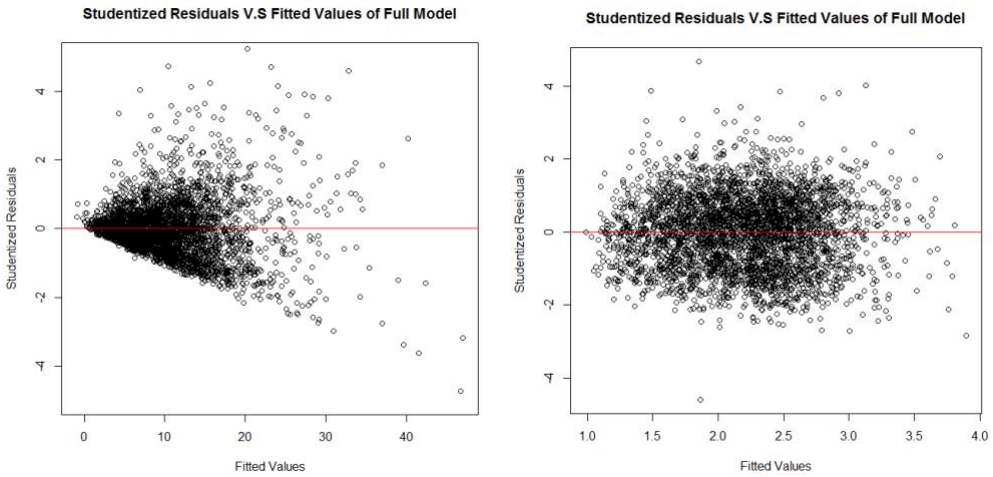


Figure_8 Residuals against fitted values plot

CONCLUSION

In conclusion, we encountered several limitations during our analysis. The first issue being, the selection criteria used for our initial set of predictors. The model consisted of over 800 variables and the variables selected were chosen due to our research, and the amount of available observations present. There may have been variables, which could have provided more explanatory power to our final model, due to errors in the data and time constraints we could not include every variables in the data set. Secondly, our model only took into account consumer expenditure information for a single quarter, while the reported income was from the entire year. It is safe to assume spending habits fluctuate during the year, especially in the fourth quarter. Although, a more in depth analysis based on income during all four quarters would require time-series analysis. Lastly, rows with missing values were omitted from our analysis. The issue of this method is, we are assuming independence between the omitted observations, and those included in our model. However, without further analysis we cannot conclude whether or not the omitted observations were independent.

REFERENCES

Laura Blow, Valérie Lechene, Peter Levell. (2011) Using the CE to Model Household Demand, National Bureau of Economic Research,1050 Massachusetts Ave., Cambridge, MA 02138. Retrieved from http://www.nber.org/chapters/c12676.pdf

Chang, Tsangyao and Fawson, Chris, "An Application of the Linear Expenditure Systems to the Pattern of Consumer Behavior in
Taiwan" (1994). Economic Research Institute Study Papers. Paper 37.
http://digitalcommons.usu.edu/eri/37

"Consumer Expenditure Survey" *Bureau of Labor and Statistics.* U.S. Bureau of Labor and Statistics, 10 September 2013 Web. 25 April 2014. http://www.bls.gov/cex/

Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to Linear Regression Analysis 5th edition. Wiley Books, 2012. Print.