Assignment 2: (HET) Heteroscedasticity – Credit Card Data

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

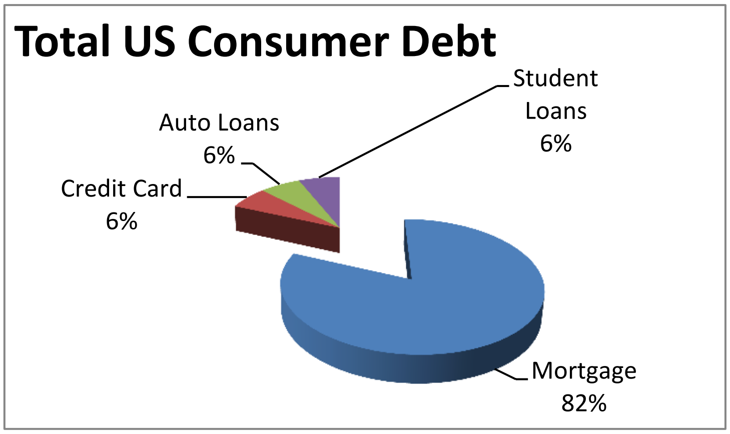
Eden Prairie, MN 55344

Executive Summary

The current credit card industry is worth 2.7 trillion dollars (federalreserve.gov). Maximizing marketing efforts to reach the most lucrative segment is a great strategy to grow revenue. Through the initial exploratory data analysis (EDA), it was concluded that the credit card data suffered from heteroscedasticity. This finding was validated through using specific heteroscedasticity tests. An ordinary least squares (OLS) regression method should not be used due to the violation of the Exogeneity for the Explanatory Variables assumption. In order to mine the data set with efficacy, an additional EDA should be conducted utilizing a different regression technique such as Weighted Least Squares or Logistic Regression.

Introduction

Debt is big business in the US; such that as of 2011 USA consumer debt was 11.4 trillion dollars (saveup.com). The pie chart below displays the percentage of total debt per category for an American household (census et. al). Of the 11.4 trillion dollars, 2.7 trillion is consumer credit debt (http://www.federalreserve.gov/releases/g19/Current/). Management has requested an exploratory data analysis (EDA) on a specific credit card data set of which the following variables will be explored:



- Age: indicator variable (IV), expressed in years.
- Income: yearly income divided by 10,000, IV.
- OwnRent: denotes whether the individual owns (1) or rents (0), notice this is binary, IV.

- AvgExp: the average monthly credit card expenditure expressed in dollars (USD). This is the dependent variable for the model.

At first glance, I would expect a strong correlation between Income and AvgExp based on the logic that as one has more income their average monthly credit expenditure would increase. Management has requested another variable called Income Squared which is Income squared. I assume this variable will behave very similar to Income. In this initial analysis, if customer "A" made 10,000 dollars a year and 10 percent of their income was credit card debt, their monthly payment would be 1000/12= 83 dollars. On the flip side, if customer "B" made 100,000 dollars a year and 10 percent of their income was credit card debt, their monthly payment would be 830 dollars a month. If percent per household remains the same, there will be a very strong correlation, such that as Income increases so too will AvgExp, holding all other variables constant.

The variable Age is initially more dynamic than Income. I would expect at around age 18 to 28 there would be the highest AvgExp correlation, and as age increases the AvgExp would begin to drop as customers become more financially independent, given that all other variables are held constant.

OwnRent is a binary variable, and given that we are using OLS I expect to see further issues in regard to satisfying the assumptions with this variable. I would expect customers that rent to have a higher correlation with AvgExp based on the assumption that renters are not as financially independent as individuals that own a home, holding all other variables constant.

For this EDA, AvgExp is the dependent variable. This variable is structured solely on amount. In my opinion, this is a rather myopic variable because it does not take into account the percentage of debt compared to total net pay. As a result, individuals that make large sums of money will be disproportionately represented in the model.

In addition to the variables, management has assigned the task of specifically exploring the possibility of heteroscedasticity. If the EDA reveals the presence heteroscedasticity, it is expected that sound statistical practice will be deployed to use the correct model for analysis. The relationships between the four main variables hold insights into exploring and eventually targeting the most financially lucrative demographic.

Analysis

In order to meet the objective of exploring the relationship between the dependent variable and independent variables, an exploratory data analysis must be conducted. This EDA will use specific techniques intended to work with data that does not conform to OLS regression. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

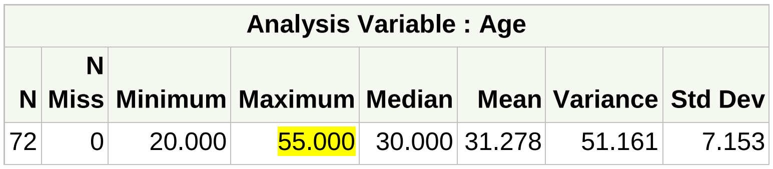
> Problem/Objective: Explore the relationship between Average Expense (Y), and the independent variables Age, Income, Income Squared, and Own versus Rent. While exploring the relationship, utilize a statistical technique that is valid when the data suffers from heteroscedasticity.
>
> Data: The data has been aggregated and has been supplied from management.
>
> Analysis: Scatter plots and correlation coefficients will be used to study the nature of the relationships between the independent variables and their relation to the dependent variable.
>
> Model: After assessing the data, a model will be used. Management has recommended using a regression model, but the standard OLS assumptions will need to be validated. If/when the assumptions do not hold another model such as Weighted Least Squares will be used to explore the relationship.
>
> Results/Interpretation: Once the model has been validated and iterations complete, a recommendation will be written to management in regard to the relational dynamics amongst the variables listed above.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives which model is used, and the analyst's personal bias is mitigated.

Data

There are a total of 100 observations with 0 completely missing values per row. Management has requested to remove attributes or rows that contain an AvgExp of 0, expressed as "y" in this EDA. Removing these rows focuses the experiment on actual customers that have an average monthly expense. After removing the values with 0 for "y", there are 72 total observations. At this point, none of the variables require a further transformation. Each variable has its own descriptive breakdown explained in a subsection below.

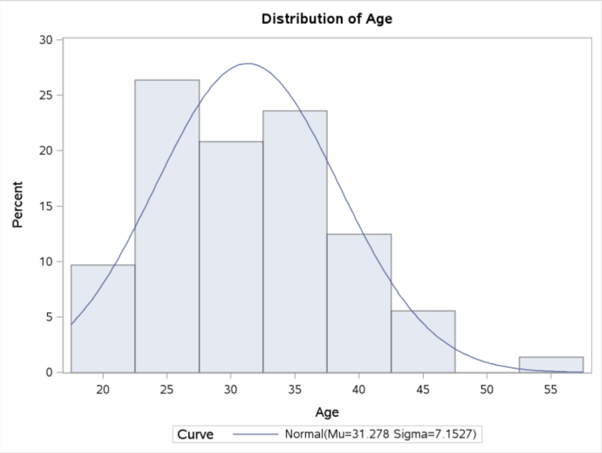
Age: This variable did not require a data transformation and represents the age in years for each individual in a row.



Analysis Variable : Age

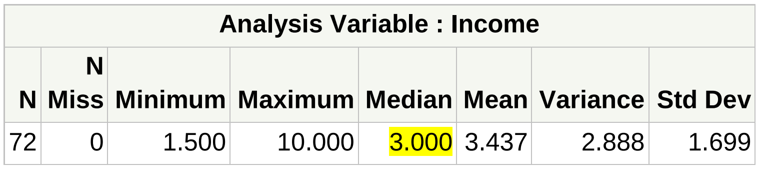
N	N Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
72	0	20.000	55.000	30.000	31.278	51.161	7.153

Age is the easiest variable to understand since the unit is easily comprehended. What initially concerns me with this variable is the maximum age is 55 years old. I would have expected the maximum age to be higher. In fact, between 2005 – 2008 those 65 and older accumulated the most credit card debt out of any demographic in the US (Public Policy Group Demos – ehow.com). Not having data that covers this age group could lead to results that are skewed towards a younger demographic, and the company missing out on a financially attractive demographic all together. The mean and median are relatively close to one another which would lead me to believe there is a relatively normal distribution. The visual demonstration, via the

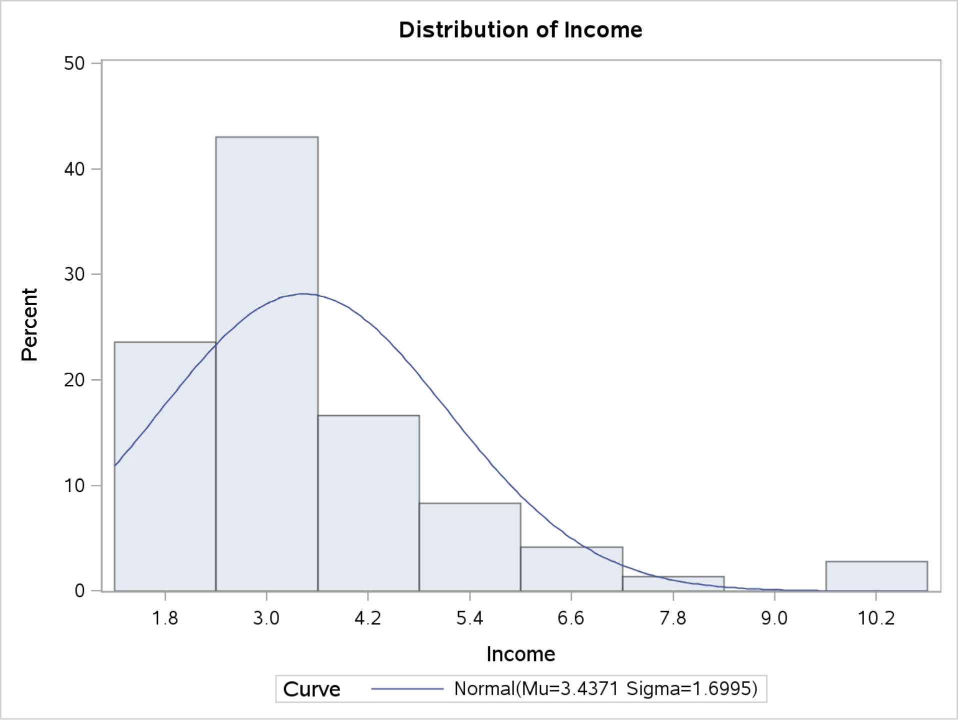


histogram to the left, reveals exactly what would be expected from the table above and my preliminary thoughts. Young people have a higher average monthly credit card expense than those of older people. This variable is positively skewed by .674, but I would question the age range of this data before drawing any concrete conclusions.

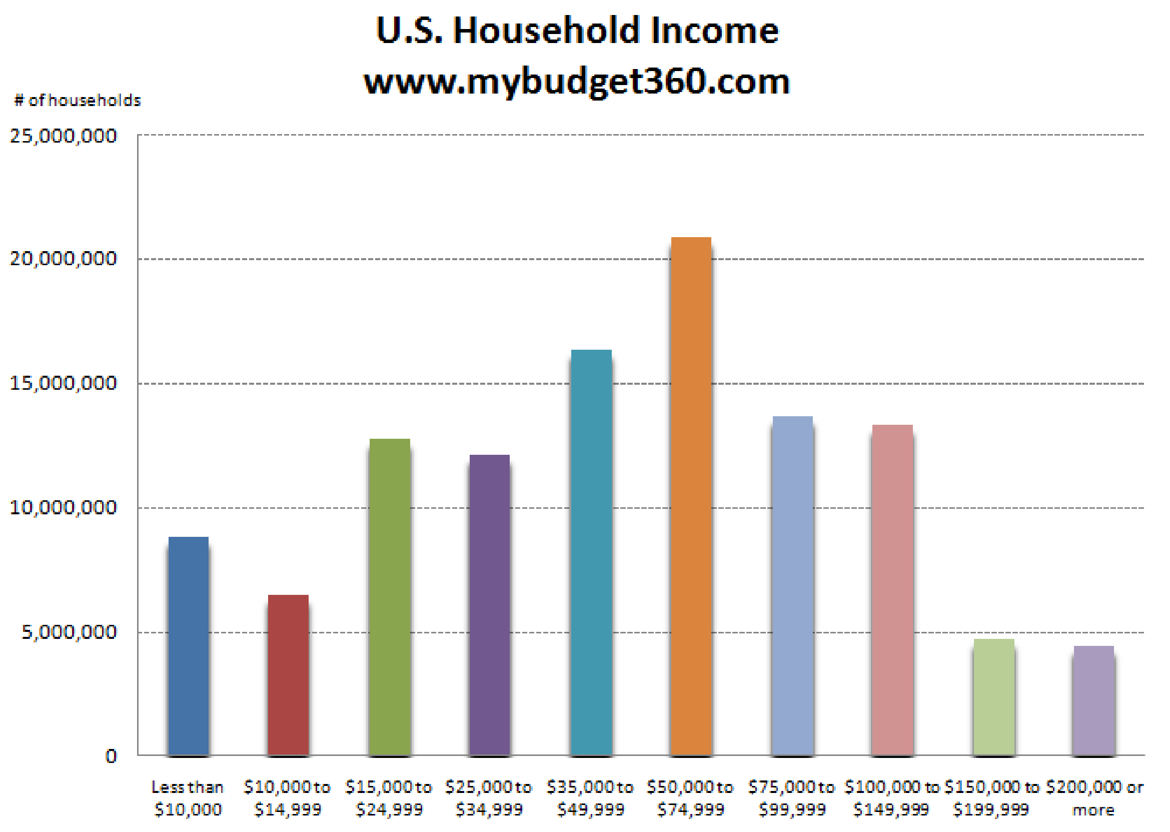
Income: This variable is yearly income divided by 10,000. The minimum value is $15,000 a year, and the maximum is $100,000 a year. According to the 2010 US Census, the median income for Americans was $50,221 a year (mybudget360.com). The conclusion I draw from this data is the target market for this company is not in line with the median income for Americans. This could be a result of purposeful demographic segmentation or a new insight for management.


Analysis Variable : Income

N	N Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
72	0	1.500	10.000	3.000	3.437	2.888	1.699



The histogram visually demonstrates how the data is poistivley skewed. In addition, the mean is roughly $34,000 a year with one standard deviation being $16,990 a year. This is a rather large standard deviation. Looking below, the US household income follows a more normal distribution, and is significnatly greater than the data set.

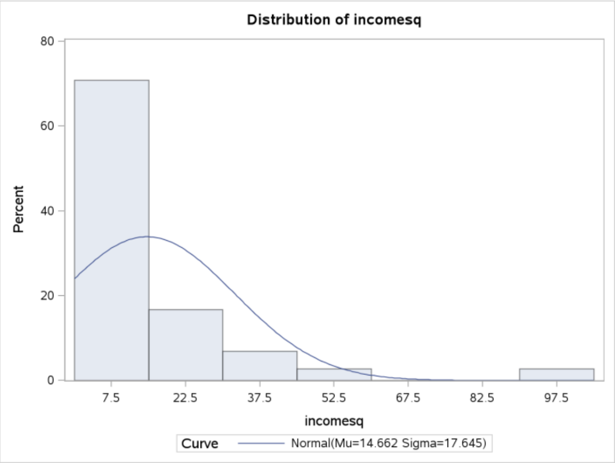


Income Squared: Similar to the variable Income, this variable is simply the square of Income. As the analysis progresses, I am surmising it will be easier to understand the results from using the square of income rather than the variable income.

Analysis Variable : Income Squared

N	N Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
72	0	2.250	100.000	9.000	14.662	311.344	17.645

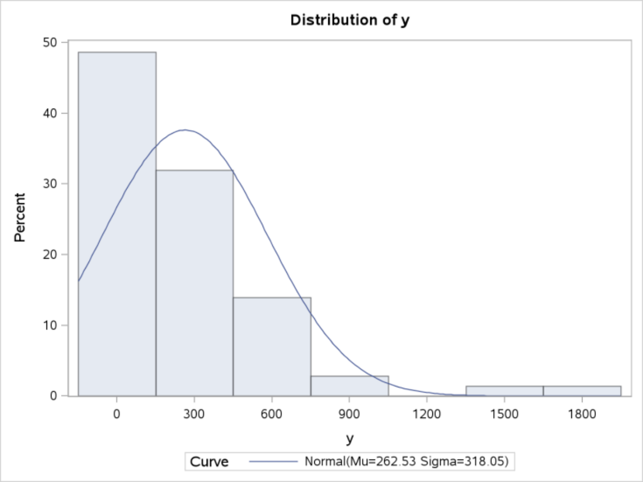
The results are very similar to Income, and please refer to the above section for further insights into this variable. The skew of the data has been exacerbated by the square function. In a prior report, I used the natural log to transform positively skewed data such that after the transformation the bulk of the data was in a manner that made it more normal. In a further analysis, I would suggest applying the natural log to this variable with the goal of making it follow a more normal distribution.



<u>Average Expense:</u> This variable represents the average monthly credit card expenditure expressed in dollars (USD), and is the dependent variable for the model.

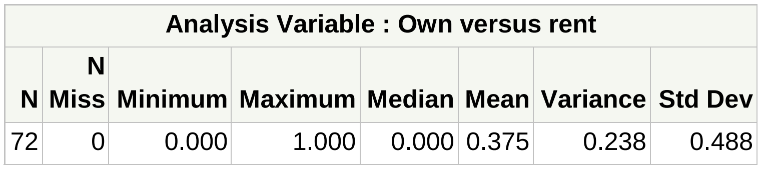
Analysis Variable : Y							
N	N Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
72	0	9.580	1898.030	158.320	262.532	101153.787	318.047

The range for this variable, 1,888, is quite large and represents the different size of payments customers pay each moths. An additional helpful insight would be to analyze the percentage of revolving debt verses stationary debt. The mean and median have quite a gap, which leads me to believe the distribution is positively skewed. Similar to the Income variable, this variable is positively skewed 3.02. The vast majority of monthly payments are under $300 a month. The natural log could also be applied to this



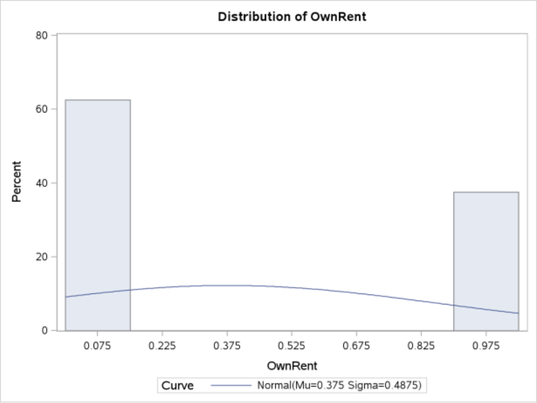
variable such that is would follow a more normal distribution.

Own versus Rent: A binary variable has the characteristic of having only two possible outcomes. It is hard to establish that this variable is a continuous variable given its binary outcome.



Analysis Variable : Own versus rent							
N	N Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
72	0	0.000	1.000	0.000	0.375	0.238	0.488

Own versus Rent is broken down such that 1 equals a row that owns a home, and 0 equals a row that rents. From the table above it can be seen that .375 or 38 percent of the rows own a home.



The histogram visually demonstrates that more rows rent versus owning a home. As of 2012, 66 percent of Americans owned their own home (US Census & usatoday.com). This data set does not reflect the US homeownership rate, but it could reflect the homeownership rate for credit card holders. I would note this finding for further study because targeting homeowners could be a marketing objective in the future. The issue I have with a binary variable is I am coding it as a dummy variable and this is a data transformation. I am allowing the regression model to capture a nonlinear relationship between the predictor and the response variable. As a result, I can expect to run into the following problems that I learned from the NYU PDF "Binary Response Models" (nyu.edu) : "

- The errors can only take on two values, 1 - xiB or -xiB. As a result the errors can never be normally distributed, therefore causing problems for hypothesis testing.
- Unbounded Predicted Values: xi¯ can take on values greater than 1 and less than 0.

- Conditional Heteroskedasticity: The variance of the residual is related to the value of x. Specifically,

var(y) = E[y] (1 -E[y]
= xiB(1 - xiB) (4) B=Coefficient, which is the symbol that looks like a cursive B.

As this illustrates, the variance of y depends on the values of X and B and is, therefore, heteroskedastic by construction.”

The data has been analyzed in its original form, and one categorical variable has been transformed for better analysis. The results from building an OLS Least Weighted Squares (LWS) should be straight forward in their interpretation.

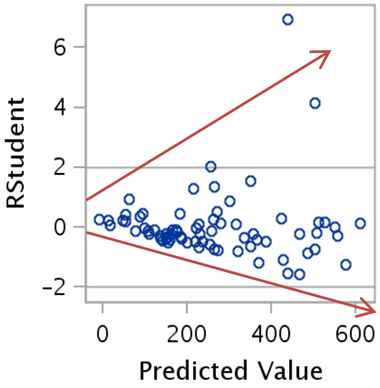
Results

From the data analysis, management has encouraged to start with an Ordinary Least Squares (OLS) regression model. Through using this model, the following analysis will be a brief overview of the data.

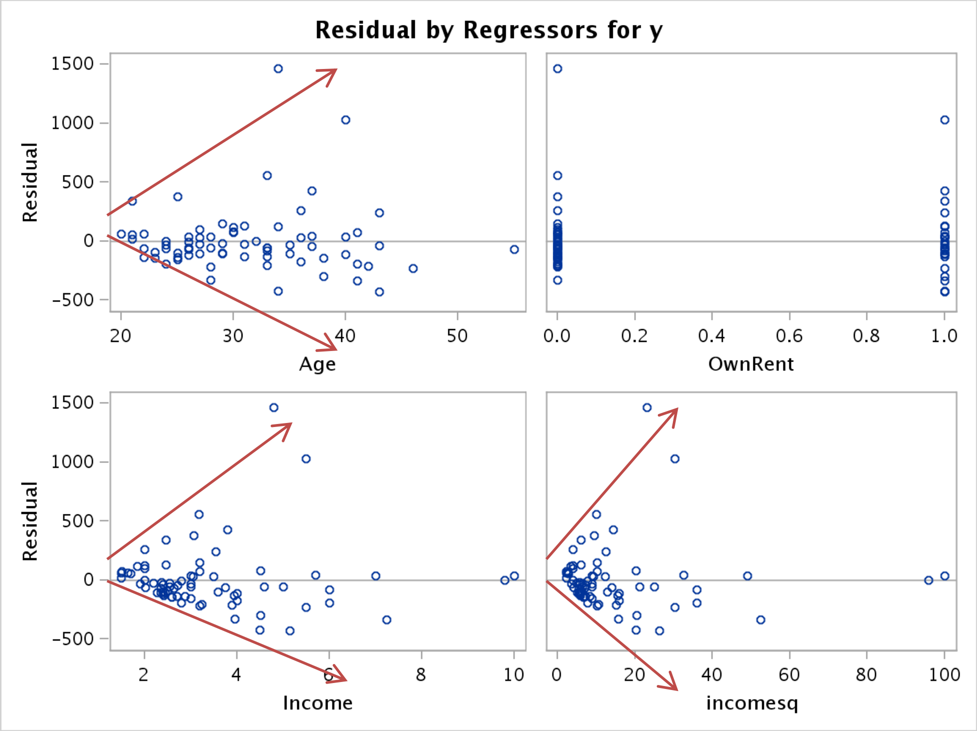
The model has a statistically strong P-value for the F-value. For this model, the critical F-value at .95 significance is the intersection of 4 for the numerator and 67 for the denominator, which is greater than or equal to 2.53. This can be interpreted as at least one variable is explanative of the dependent variable in the model. The adjusted R-squared is low, and from a brief synopsis standpoint I would conclude that this model is not very predictive.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1749357	437339	5.39	0.0008
Error	67	5432562	81083		
Corrected Total	71	7181919			
R-Square	0. 2436	**Adj R-Sq**		0. 1984	

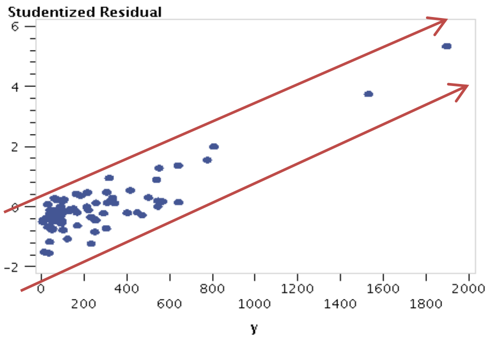
At this point in the EDA, checking the models diagnostics is done to validate the OLS assumptions for further assessment of the model adequacy. Specifically honing in on the residuals is vital for assessing the presence of heteroscedasticity. Utilizing scatterplots of the variables versus the standardized residuals is a great initial process for detecting heteroscedasticity. The first scatter plot captures the relationship of the standardized residuals with the predicted variables. Notice the funnel/cone like scatter of the plots. Scatter plots can be interpreted subjectively, but it is rather apparent that this plot suffers from heteroscedasticity. The next step is to assess scatter plots of each individual variable against their corresponding residuals to narrow down the variable/variables causing the heteroscedasticity.



Below, are the individual scatter plots of the residuals versus the corresponding variables. It can be seen that Age has a mild cone shape as age increases, which leads one to suspect that a function of Age is correlated with the residuals. Surprisingly, OwnRent does not appear to suffer from heteroscedasticity. Income and

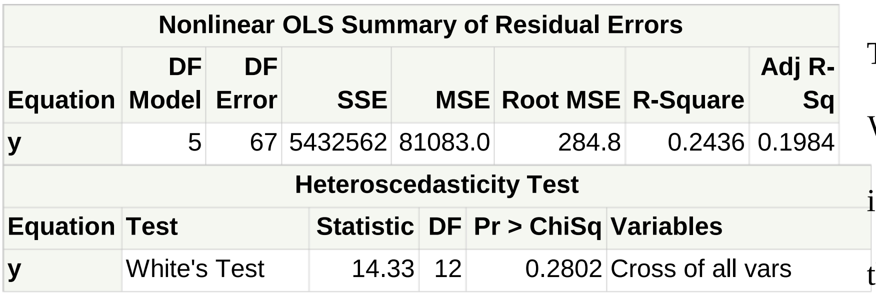


Incomesq both look to suffer from heteroscedasticity. Like Age, it could be specified that the residuals are correlated to a function of Income and Incomesq. An assessment of the response variable is needed to verify if it also suffer from heteroscedasticity. From looking at the scatter plot to the right, it can be seen that there is no cone-like scatter from the plots. But, there appears to be a linear relationship amongst the response variable and the studentized residuals. While this is not a violation of heteroscedasticity, it should be further investigated. The ideal scatter plot of the residuals would be a random scatter with no discernible pattern.



At this point in the EDA, it can be conferred that the credit card data initially suffers from heteroscedasticity. The next step is to validate the visual inspections with formal tests that identify the presence of heteroscedasticity (Ajmani 2009).

White's general test assesses the relationship between the variance of the disturbances and regressor variables. Essentially, this test squares the residuals and regresses them on all the variables, their squares, and cross products (Gupta 2000). This test assumes no form of a relationship between either variable, and is validated through a hypothesis test. It assumes an asymptotically chi-squared distribution with p-1 degrees of freedom.



Nonlinear OLS Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
y	5	67	5432562	81083.0	284.8	0.2436	0.1984

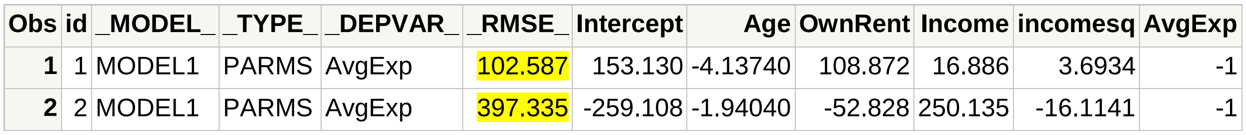
Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
y	White's Test	14.33	12	0.2802	Cross of all vars

The output from running White's general is interesting and points out the weakness of using this test. What I find amazing about using this regression technique is the R-square and Adjusted R-squared are the same as the OLS model. White’s heteroscedasticity test generated a test statistic of 14.33 with 12 degrees of freedom. In order to diagnose a model with heteroscedasticity at the 95 percent confidence level, it must have a test statistic of 21.02 for 12 degrees of freedom (http://www.itl.nist.gov). One cannot reject the null hypothesis, and this test states that the model does not suffer from heteroscedasticity. From the scatter plots, it was clear that heteroscedasticity was present in the model, but this test failed to validate this finding. This exemplifies how White’s test is not a solid test for detecting heteroscedasticity. When conducting an EDA, it is important to use multiple tools and techniques to validate a model because weaknesses exist for specific statistical tests and techniques. The vast majority of this content was referenced from Vivek Ajmani’s book *Applied Econometrics*.

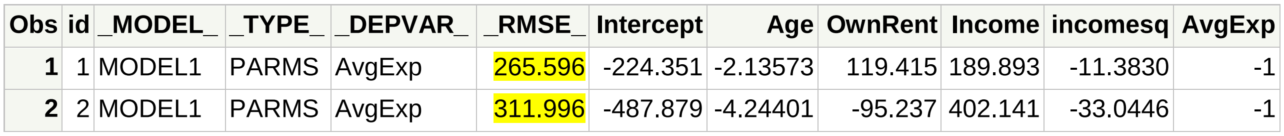
Other tests for detecting heteroscedasticity involve a more myopic approach than White’s test, but are more accurate given that they focus on individual variables. Recall from the initial assessment that variables Income, Income Square, and Age all visually showed signs of heteroscedasticity. The two tests I am going to deploy, Goldfeld-Quandt and Breusch-Pagan, focus on the variables that are presumed to be causing the heteroscedasticity.

Using the Goldfeld-Quandt test, the data is ranked/split into two groups. The first group is based on the variance of the problematic variables, and the second group is the variance of the variable(s) that are not suspected of heteroscedasticity. After splitting the data, the F-test is used to compare the two variances. A hypothesis test is used to determine heteroscedasticity with an F-distribution of n1-k-1 and n2-k-1 degrees of freedom with the subsequent threshold. In order to demonstrate how this test was conducted on the credit card data, I will break down the steps for splitting the data. Recall from the residual scatter plots for variables income and income squared, that as income increased the variance increased as well. Thus, ranking the data based on income is a way to linearly sort the data from least variance too greatest variance. After ranking the data, simply splitting it in half satisfies splitting the variances into two groups that differ greatest.



Obs	id	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	OwnRent	Income	incomesq	AvgExp
1	1	MODEL1	PARMS	AvgExp	102.587	153.130	-4.13740	108.872	16.886	3.6934	-1
2	2	MODEL1	PARMS	AvgExp	397.335	-259.108	-1.94040	-52.828	250.135	-16.1141	-1

The table above shows the results of splitting the data by greatest variance from the Income/Income Squared variables. To conduct the F-test, divide the model with the suspected greater variance by the model with less variance. Given the parameters, the threshold for the F-test needs to be greater than 1.822 for 95 percent confidence. 397.335/102.587 equals 3.873, and I can reject the null hypothesis of homoscedasticity and accept the hypothesis of heteroscedasticity. Now, let’s analyze the F-test if I did not sort the data based on the variable(s) Income’s variance. Here it can be seen that the arbitrary ranking by observation number yields a F-test of 311.996/265.596 equals 1.175, which is below the threshold to reject the null hypothesis of homoscedasticity. While the Goldfeld-Quandt test identified one variable that suffered from heteroscedasticity, its limitations lie in its ability to delineate multiple variables that suffer from heteroscedasticity.



Obs	id	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	OwnRent	Income	incomesq	AvgExp
1	1	MODEL1	PARMS	AvgExp	265.596	-224.351	-2.13573	119.415	189.893	-11.3830	-1
2	2	MODEL1	PARMS	AvgExp	311.996	-487.879	-4.24401	-95.237	402.141	-33.0446	-1

When multiple variables are suspected of heteroscedasticity, Breusch and Pagan developed the Lagrange Multiplier test, which assesses multiple regressors. In my opinion, this test has a similar framework to White’s test in that it regresses the squared residuals against independent variables testing for significant variations (MTSU.edu). This test is validated through a hypothesis test, and assumes ac hi-squared distribution with *k* degrees of freedom based on the observations in the model. For the credit card data, all the variables form one vector, and the variables suspected of heteroscedasticity, Income and Income Squared, are in another vector. The residuals are then calculated from OLS and the sum of squares is regressed on the vector with the suspected heteroscedastic variables. The test statistic generated needs to cross the appropriate threshold given the degrees of freedom. The threshold for the chi-squared distribution is 5.991 and has 2 degrees of freedom based on the fact that there are two variables in the suspected heteroscedastic vector. After conducting the Breusch Pagan Test Statistic the value is 41.9203 with a p-value of 7.891E-10, which means that the null hypothesis of homoscedasticity is rejected. Using this technique, I went back into the SAS program and put other variables into the Z vector. Below are my findings.

Age	LM
The Breusch Pagan Test Statistic Value is	10.001002
The p value associated with this is	0.0015646
The null hypothesis of homoscedasticity is rejected for the variable Age, I expected this based on the scatter plot for Residuals versus Age.	

Rent/Own	LM
The Breusch Pagan Test Statistic Value is	0.0720366
The p value associated with this is	0.7883942
The null hypothesis of homoscedasticity is not rejected, and this aligns with the initial scatter plot assessment.	

Income	LM
The Breusch Pagan Test Statistic Value is	14.942254
The p value associated with this is	0.0001109
The null hypothesis of homoscedasticity is rejected, although note how these numbers are different than the BP test when conducting White’s test.	

Income Squared	LM
The Breusch Pagan Test Statistic Value is	5.5760606
The p value associated with this is	0.0182076
The null hypothesis of homoscedasticity is rejected, although I would have expected this to be the same as Income.	

In addition to these findings, seeing that we are using nested models one could conduct the F-test for nested models to verify that added variables are contributing to the increase in heteroscedasticity. For example, if I wanted to test the variables Income and Own/Rent to see if heteroscedasticity is present I would run the following test statistic:

F = [(SSE(RM) – SSE(FM)) / (p-k)] / [SSE(FM) / (n-p-1)]

The findings from the two of three formal tests, White's - Goldfeld Quandt - and Breusch-Pagan, validated the findings in the scatter plots of heteroscedasticity. White’s test did not validate heteroscedasticity, but it assumes no form of a relationship between the variables being studies, which is a plus in some circumstances. The Goldfeld-Quandt and Breusch-Pagan tests validated heteroscedasticity and focus on specific variables that are presumed to be the culprit. Each test utilizes different distributions and parameters for establishing degrees of freedom, which need to be understood in order to come to an accurate conclusion.

After conducting my initial EDA, I would conclude that the credit card data suffers from heteroscedasticity with at least three of the four variables. OLS should not be used as the regression model for the following reasons: the variance is biased for the estimated parameters, and the t-values are invalid for the estimated coefficients (MTSU.edu). My recommendation to management is to use a different regression technique that does not require the assumption of Exogeneity for the Explanatory Variables. Other techniques to consider would be Weighted Least Squares or Logistic Regression.

Future Work

Further recommendations on how this study can be improved upon are the following:

- Utilize a data set that reflects the desired target market(s).
  - Include data that has older age groups represented.
  - Analyze data that has a higher median income.
- Use Weighted Least Square or Logistic Regression to further conclude the EDA.
- Explore the Goldfeld-Quandt test for multiple variables.

Through this initial EDA, coupled with the future work recommendations, delineating the demographic that has the highest average monthly credit card expense can be utilized to capture additional revenue.

References

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.

FRB: G.19 Release-- Consumer Credit. (n.d.). *Board of Governors of the Federal Reserve System*. Retrieved January 18, 2013, from http://www.federalreserve.gov/releases/g19/Current/

Forbes. (2012, November 19). 2.7 Trillion. *Forbes*, *190*, 28.

Gupta, V. (2000). Regression explained in simple terms. *Vijay Gupta Publication*, *1*, 5. Retrieved January 18, 2013, from https://mywebspace.wisc.edu/rlbrown3/web/library/regression_explained.pdf

How Much Debt Does the Average American Have?. (n.d.). *USA Consumer Debt*. Retrieved January 17, 2013, from https://www.saveup.com/blog/us-debt-statistics/

How much does the average American make in 2010? . (n.d.). *My Budget 360*. Retrieved January 18, 2013, from http://www.mybudget360.com/how-much-does-the-average-american-make-in-2010-household-income-new-data-100-million-make-less-than-40000/

Huebsch, R. (n.d.). What to Do About Credit Card Debt for Older People | eHow.com. *eHow | How to Videos, Articles & More - Discover the expert in you. | eHow.com*. Retrieved January 18, 2013, from http://www.ehow.com/info_7754881_do-card-debt-older-people.html

NYU. (n.d.). Binary Response Models. *NYU*. Retrieved January 18, 2013, from https://files.nyu.edu/mrg217/public/binaryresponse.pdf

University, M. T. (0). What is heteroscedasticity and why is it a problem?. *Middle Tenessee State University*, *1*, 1.