ISyE 7401 Project Report

A Regression Analysis of Monthly Rent Paid by ISyE PhD Students

Christopher Shartrand

Professor Roshan Vengazhiyil

Motivation:

In the Fall of 2015, Dr. Alan Erera announced a framework for changes to the stipend system for ISyE PhD students funded under either teaching assistanceships. Under the new system, pay would be increased from a 1/3 time compensation to a 2/5 time compensation, with the additional money coming from each student’s advisor in addition to the original amount provided by the department. This increase, set to take effect for the 2016-2017 school year, was implemented in response to the suggestions submitted by ISyE GSAC, who argued that the current stipend system was no longer sufficient in covering the cost of living for the city of Atlanta and the surrounding suburbs. While an investigation into the estimated cost of living for the city was conducted before the implementation of the new plan, it lacked the breadth of understanding the student body for which it was directed to serve. As a result it was my belief that gathering data on the factors that influence the way the students spend the stipend was imperative for a thorough study on this issue. It was my hope that the department will use the data collected for making decisions on stipend changes in the future.

The Data:

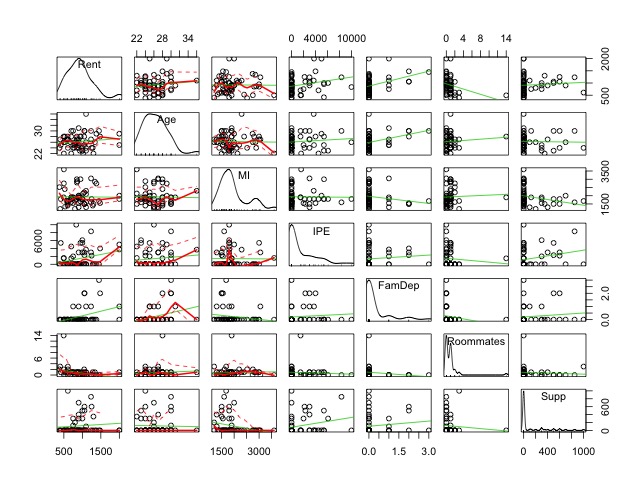
The data was collected over a two-week period in March of 2016 using the online data collection service, Survey Monkey. The survey was distributed using the ISyE PhD email blast on Microsoft Outlook. An initial email was sent on the start of the first week and a reminder email for those who had yet to participate was distributed at the beginning of the second week. For the purpose of this study, the variable of interest was determined to be monthly rent in US dollars. This was due to the fact that economically, rent contributes to a large majority of spent monthly income, especially at lower income brackets. The corresponding explanatory variables gathered were formulated with monthly rent in mind, but also focused to the desired respondents, the ISyE PhD students. In total, data was gathered for six separate explanatory variables which include; age in years, approximate monthly income from stipend or outside funding in US dollars, approximate monthly income from previous employment in the past 5 years in US dollars, number of family dependents, number of roommates who are able to split monthly rent, and approximate money spent to supplement the cost of rent in US dollars. A sample of the dataset can be seen in Table 1 below. The full dataset is attached in a .xlsx file with the digital file for this report.

Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rent($) | Age | MI($) | IPE($) | FamDep | Roommates | Supp($) |
| 1100 | 26 | 1300 | 0 | 1 | 0 | 100 |
| 608.34 | 27 | 1614 | 0 | 0 | 2 | 0 |
| 535 | 26 | 1340 | 0 | 0 | 0 | 0 |
| 1480 | 28 | 1552 | 4000 | 2 | 0 | 0 |
| 750 | 25 | 1550 | 2000 | 1 | 0 | 300 |
| 400 | 27 | 1200 | 0 | 0 | 3 | 0 |
| 1000 | 23 | 1800 | 5400 | 0 | 0 | 500 |
| 1057 | 30 | 1500 | 2300 | 2 | 0 | 500 |

Analysis:

All regression analysis conducted in this report was completed using R-Studio Version 3.0.3 (2014-03-06). The first step I took was to visually investigate the distribution of all variables from a scatterplot matrix, which can be seen in Figure 1. Attempts at identifying potential collinearity between predictors were found using the scatterplot in the cases of family dependents and age, supplemental rent and income from previous employment, supplemental rent and family dependents, income from previous employment and family dependents, and supplemental rent and monthly income. An enlarged plot of each of the 5 individual cases can be found in Appendix 1.

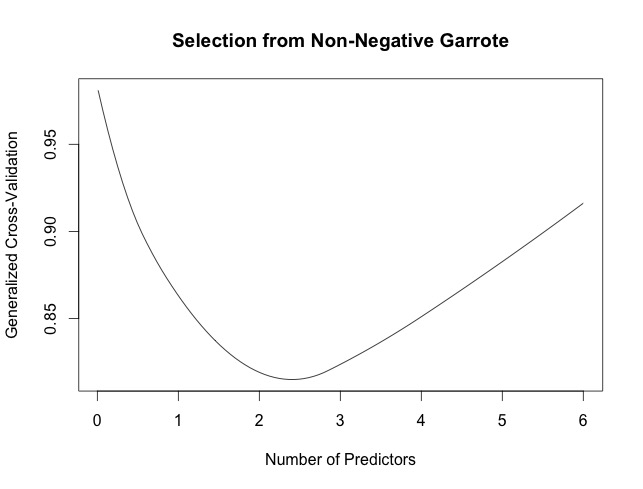
Figure 1.

Following the qualitative investigation from the scatterplots, a quantitative investigation was conducted through production on the correlation matrix seen in Table 2 below. The correlation matrix reveals that the largest connection between two predictors is that of age and family dependents with a value of 0.33. While there may be some evidence of correlation between these two predictors, the value is not large enough to worry for collinearity between predictors. To support this belief, analysis of the variance inflation factor reveals no values above 1.5, well below the threshold of 10. Hence it is concluded that collinearity between predictors is not an issue for this study.

Table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age | MI | IPE | FamDep | Roommates | Supp |
| Age | 1 | -0.02 | 0.1 | 0.33 | 0.08 | -0.03 |
| MI | -0.02 | 1 | -0.01 | -0.17 | 0.04 | -0.26 |
| IPE | 0.1 | -0.01 | 1 | 0.09 | -0.08 | 0.24 |
| FamDep | 0.33 | -0.17 | 0.09 | 1 | -0.18 | 0.11 |
| Roommates | 0.08 | 0.04 | -0.08 | -0.18 | 1 | -0.08 |
| Supp | -0.03 | -0.26 | 0.24 | 0.11 | -0.08 | 1 |

Following the analysis of collinearity, I looked to determining which main variables I should include in my model. In order to make the best decision of variables to use, I ran analysis for multiple variable selections methods. Beginning with subset regression and Mallow’s Cp as the indicator, I found that the model should be comprised of every variable with the exception of age based on a Cp of 5.006 for the five predictor variables. The plot for this can be seen in Appendix 2. Moving forward to stepwise regression, the process completed three iterations before a final AIC value of 677.81 was reported. In this case the three predictor variables suggested were income from previous employment, number of family dependents, and number of roommates. The selections of this method were further strengthened by the use of generalized cross-validation during a non-negative garrote calculation where the model coefficients of all variables except IPE, FamDep, and Roommates shrunk to zero. The plot of the M value, which minimizes GCV, can be viewed in Figure 2. Finally, a lasso regression was conducted against the six main explanatory variables and the results yielded a model with non-zero coefficients for only IPE, FamDep, and Roommates. Due to the support between stepwise, non-negative garrote and lasso methods, I determined that these were the variables to include in the model.

Figure 2.

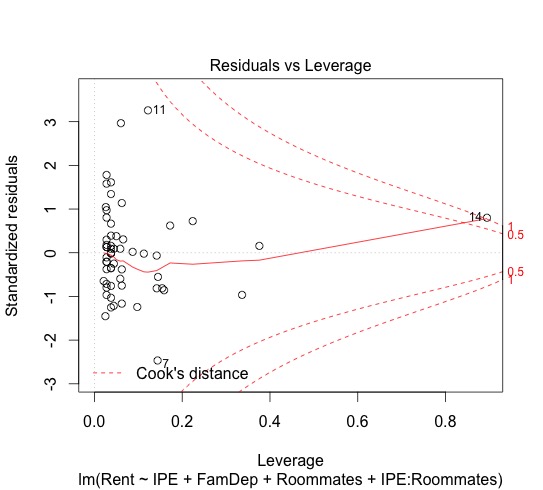
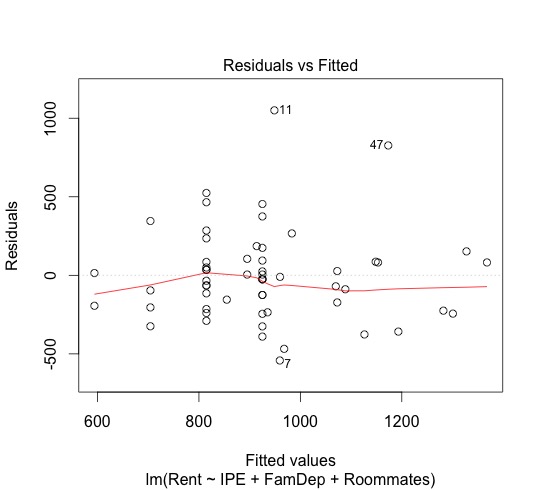
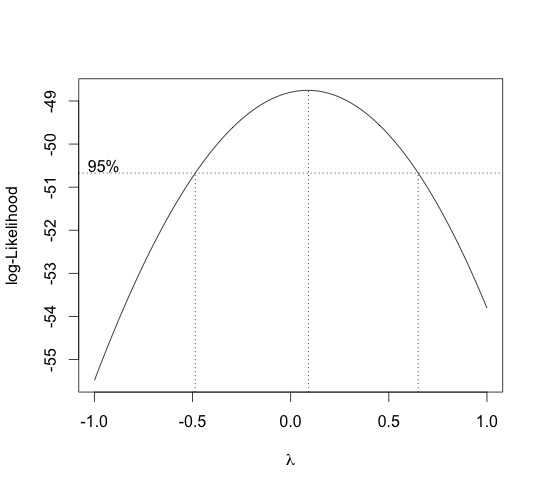
 Using these three variables, I tested for the presence of any significant interaction effects between the predictors. Using strong heredity as my determination factor and again using non-negative garrote to calculate the regression coefficients, I found that the along with the main predictors, only the interaction between IPE and Roommates was not shrunk to zero. Thus, I was left with a model containing the variables IPE, FamDep, Roommates and IPE:Roommates. Preliminary residual analysis of this model identified both heteroscedasticity and a significant influential observation. For the Cook’s Distance plot seen in Figure 3 below,

Figure 3.

observation 14 maintains a strong leverage on the fit of the model. Upon investigation of the given observation, it was found that the survey respondent had input that their housing situation included 14 distinct roommates who contribute monetarily to the monthly rent. Assuming that this was an erroneous response, the observation was removed from the dataset. Because of the updated dataset, variable selection for the model was rerun using the same methodology described above. All conclusions held with the exception of the selection of interaction terms using non-negative garrote. In this instance, the procedure shrunk all three interaction terms to zero, leaving only the three main predictors, IPE, FamDep, and Roommates. Regressing Rent against these three predictors under the updated dataset revealed an additional violation of model assumptions, heteroscedasticity, displayed by Figure 4.

Figure 4.

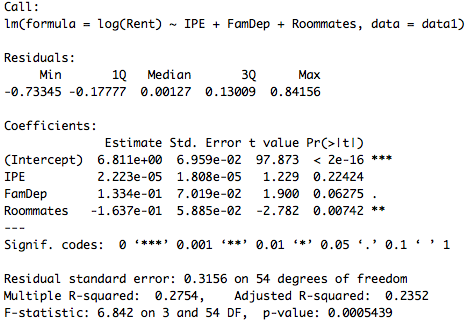
From the plot, it appears that a portion of the ‘fanning out’ of the variability is contributed by observations 11 and 47. Creation of the half-normal plot located in Appendix 3 details the two observations as significant outliers. Viewing the data of which the observations detail reveals rent values that are relatively large given the corresponding presence of roommates for two respondents. However, the rent values themselves are not outlying values when considered in the scope of the entire study. Based on this, coupled with fact that given a normal distribution of data, we would expect to observe approximately 5% of the 58 rent values (or 2-3 data points) to be more than 2 standard deviations away from the mean, the outliers were kept in the data set. Because data removal was decided against to deal with the non-constant variance, a transformation of the response was deemed necessary.

Figure 5.

Using the plot of the Box-Cox transformation in Figure 5, it was determined that the that maximized the log-Likelihood was . However because is included within the 95% confidence interval of the Box-Cox and is most closely located ‘simple’ transformation to the maximum , the log transformation was considered for this model.

Computing a final regression model with the log transformation produced the model:

for which the final residual plots can be found in Appendix 4. The regression output produced by R is seen in Table 3 below.

Table 3.

Results:

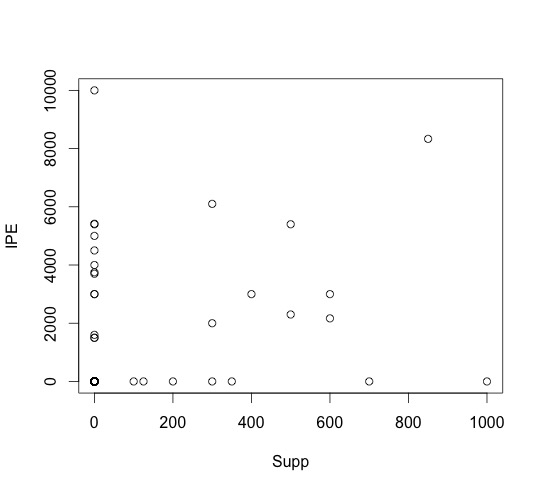
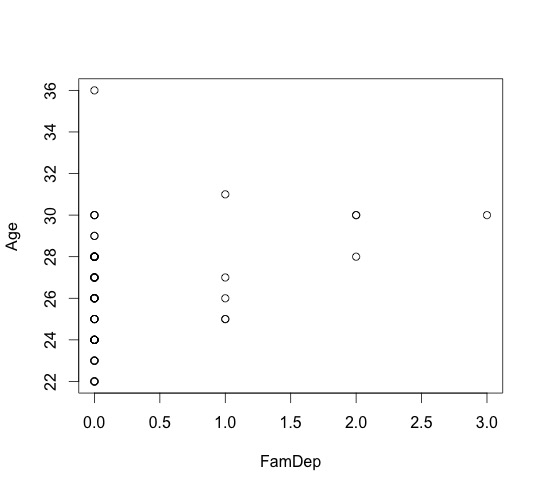
Described in the motivation section, the primary goal of this study was to gather data on the factors that influence the way the PhD students spend their stipends, specifically focusing on rent. Due to this, when attempting to model rent, much of the conventional predictor variables such as location, square footage, age of apartment, etc. were left out in favor of predictors specific to that of the students. Understandably, such a methodology resulted in a low , or 27.5% of the variance in rent is explained by the income from previous employment, number of family dependents, and number of roommates who split rent. Despite this, there are a number of important observations to gather from the model. First, it is found that a PhD student with no family dependents or roommates who does not have income from previous employment (such as the author), would expect to pay $907.77 a month on rent. With the inclusion of an additional roommate, monthly rent falls to approximately $770.70. Monthly rent for a student providing for one other person who cannot contribute as a roommate is approximately $1,037.32. Finally, monthly rent for students who previously held paid employment increases by approximately $21 dollars for every $1,000 of previous monthly employment.

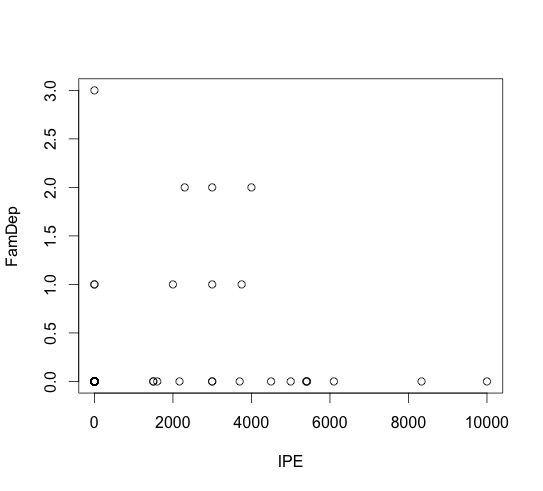
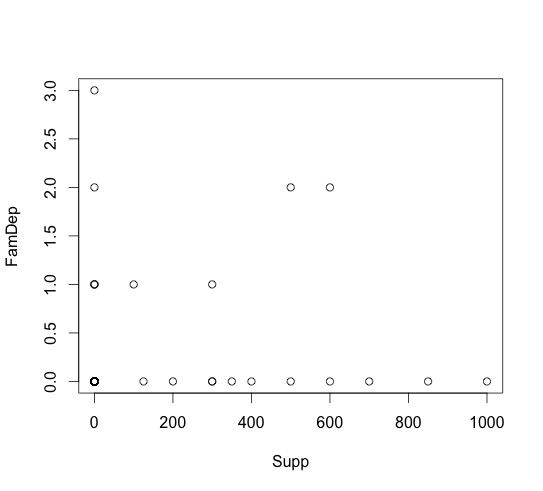
As previously stated, students with one family dependent expect to pay approximately $130 more a month on rent, or $1,560 extra over the course of the year. When considering the current guaranteed stipend provided by the department ($1,800 over 9 months), students with one family dependent would spend approximately one tenth more of their stipend to provide adequate housing than their department counterparts. Adding in the additional costs of necessity goods such as food, this percentage is certain to rise. In total, 9 of 58 observations in the data set contained students who support a family dependent. Under the assumption that this data set is reflective of the entire department, 15% of the PhD students use their stipend to at least partially support one or more family members. One important observation to make regarding the residual plots found in Appendix 4 is that of the normal probability plot. The residuals in the plot appear to display some evidence of a Cauchy distribution. The presence of ‘heavier tails’ to the distribution can most likely be attributed to the effect of the two predictors FamDep and Roommates. Those students with family dependents are adversely affected seeing their monthly rent rise, students who are able to find roommates to split rent can decrease their monthly rent and all others are bunched around the mean.

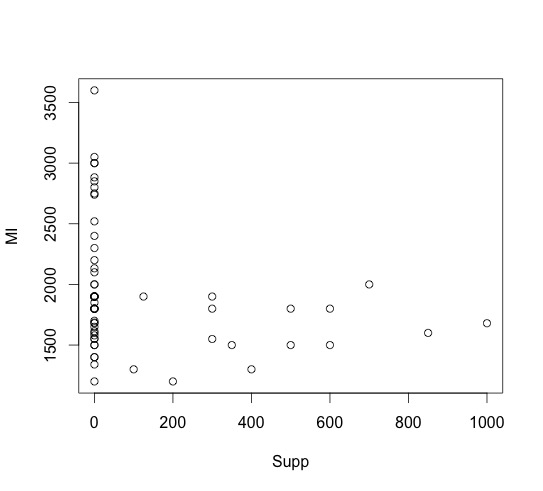
Future Work:

As I believe collecting relevant data regarding the student body is important for the implementation of sound objectives for both GSAC and the department, I would like to conduct a follow up investigation where I fix some of the issues described previously. First I would hope to gather data from every PhD student in order to make sure the data gathered is completely reflective of the department. Furthermore, I would look to include conventional predictor variables for apartments (location, square footage, age of apartment, etc) in order to capture the nature of the fluctuations in rent. Additionally, I believe that collection of such variables can further strengthen the significance of the family dependents and roommates predictors found in the initial study, as well as give more insight into the needs of the PhD students in regards to housing.

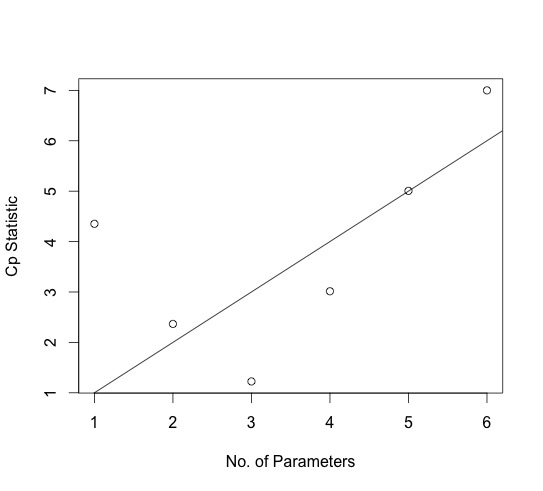
Appendix 1.



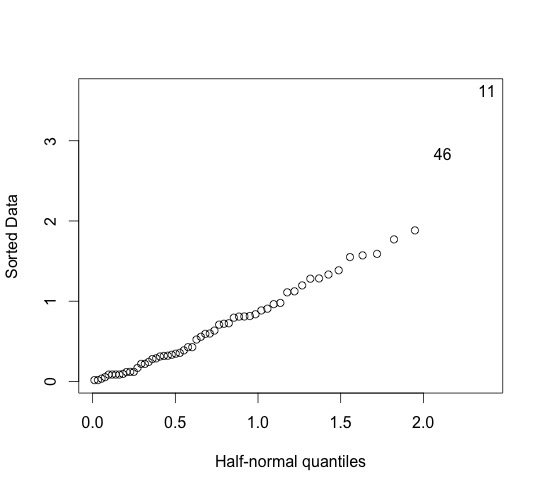




Appendix 2.



Appendix 3.



Appendix 4.

