Investigating the effects of race and gender on income using generalized linear models and penalized regression models

Christopher Odell[1]

Introduction

In recent years there have been many different social discussions over the impact of race, sex, education and other factors on income. These questions can be investigated with census data. In this analysis I used census data for the state of Oregon from 2013-2017 to examine four questions of interest: 1) Does race affect income, 2) how does gender affect total income, 3) does education or hours worked per week have a larger impact on total income, and 4) what is the relationship between hours of work and education attainment? I hypothesized that: 1) there is no difference between the different races and the affect that they have on income, 2) there is no difference in income between males and females, 3) there is no difference in the affect of education and hours worked per week on total income, and 4) there is no relation between hours of work and years of education. The alternative to these four hypotheses is that there is a difference. To make inference on these hypotheses I used a generalized linear model and model diagnostics that included penalized regression.

Data Description

The data is collected by Census Bureau, of the United States and readily available to the public. To answer my questions of interest I used a 5-year Public Use Microdata Sample (PUMS) Population dataset for the state of Oregon from 2013-2017. Public Use Microdata Sample (PUMS) files are a set of untabulated records about individual people and/or housing information. The 5-year PUMS is a combination of 1-year PUMS produced annually with appropriate adjustments to weights and inflation adjustment factors. The data for the state of Oregon from 2013-2017 came as comma separated values (csv) file with 286 variables, and 200,158 observations. I had to conduct extensive reading into the data collection and coding to understand what variables of interest would help answer my four hypotheses. After careful thought, I proceeded to filter the data down to the following explanatory variables: working hours per week ‘WKHP’, education level ‘SCHL’, gender ‘SEX’, age of the person ‘AGEP, what category of race ‘RAC1P’, marital status ‘MAR’, occupation ‘OCCP’, and class of worker ‘COW’. These variables would be used to estimate their causal impact onto the response variables of total personal income ‘PINCP’. A principal component analysis could have been conducted to help with variable reduction; however, my hypotheses were more geared towards specific questions of interest which I felt required careful selection of the explanatory variables.

Data Collection Analysis

The variables I had selected had a wide range of missing values, most likely due to survey response error. My initial analysis concerning missing data, in this case provided as NAs, showed that at least one of the variables had almost 50% of the observations as NA. The variable ‘WKHP’ made up the largest

[1] The initial analysis of this data was originally performed for a group project in ST_558, at Oregon State University in Spring 2019. The analysis was changed and updated to reflect statistical accuracy and new objectives in Winter 2020.

portion of the NAs in the dataset. Upon examination, it was clear that most of the NAs came from individuals below the working age of 16, or over the expected retirement age of 67. I made an informed decision after this point to limit the analysis to individuals between the ages of 16 and 67. After filtering out individuals under the age of 16 or over the age of 67 there were no longer NAs in any of the variables besides 'WKHP'. Now the data was represented by 912,000 observations with 'WKHP' still having 31,121 NAs. I had three options on how to proceed forward: I could either perform imputations to fill in the missing values with NAs, artificially change the NAs to 0, or remove the NAs altogether. Without a deeper understanding of why the NAs existed, it would not be appropriate for me to assume that these values were 0. Since I was conducting hypothesis testing, and I did not want to add bias by further exploration of the data, I decided not to perform imputations to fill in those NAs. I decided at this point to drop all observations in 'WKHP' that had NAs. I did this knowing that I had a large number of observations (N) which would still be representative of the population. During the analysis, it became clear that 'PINCP' was spread rather far apart with some individuals making less than $20,000 per year and some over $1 million per year. To help address this I proceeded forward with the log transformation of the 'PINCP' variable.

Next, I ran into a little bit of a problem with how some of the data was coded. For example, occupation 'OCCP' was coded into numerical variables. There is no quantitative difference in the numbers between one occupation and another (1500, 2500, 400 etc), so I created factors based on occupational groups. I used a break to separate them into 26 different discrete explanatory variables. This process was needed on other variables as well. I performed this also on: 'COW', 'MAR', and 'RAC1P'. Since 'COW' is class of worker I broke it down into eight discrete classifications. Marital status 'MAR' was converted to a discrete factor of five different levels. Race classification 'RAC1P', had eight different levels and was reclassified from continuous to discrete. Education 'SCHL' was given as a discrete variable that I transformed into years of education in order to use it as a continuous variable. I did this with the assumption that more years of education is a higher value than less years. After the completion of these transformations, I had five discrete explanatory variables and three continuous explanatory variables.

Since I was not familiar with the data and was not part of the data collection process, I produced some simple plots before doing model analysis. Figures 1 through 3 in Appendix A show race 'RAC1P', sex 'SEX', education 'SCHL' and hours worked per week 'WKHP' versus the log transformation of total income 'PINCP'. I observed a slight difference in the log of total income by race and gender, with some linearity observed in the relationship between the log of total income and hours worked per week. The distribution of education based on hours worked per week is shown in Figure 4 in Appendix A. There is no discernable trend between education and number of hours worked per week. During the exploratory process I saw an interesting impact of occupation, and gender versus that of the log transformed income which is shown in Figure 1 below:

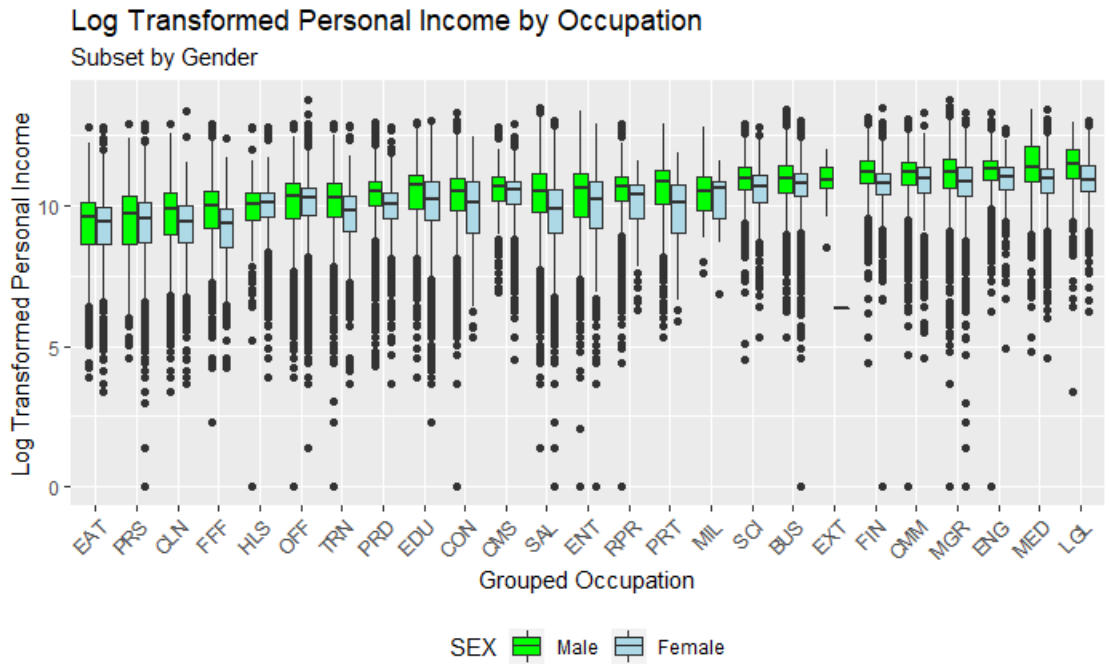

Figure 1. Box and whisker plot showing log transformed personal income by occupation

Model Selection:

To answer my questions of interest on this multivariate dataset I used a generalized linear model:

$Y_i = \beta_0 + \beta_1 x_i + e_i$

which I converted to:

$E(Y_i) = \beta_0 + \beta x_i$

For Questions 1 through 3 I used the following full model:

$\log(PINCP) \sim \beta_0 + \beta_1 WKHP + \beta_2 OCCP + \beta_3 AGEP + \beta_4 SCHL + \beta_5 SCHL2 + \beta_6 COW + \beta_7 MAR + \beta_8 SEX + \beta_9 RAC1P$

For Question 4 I used the following full model.

$WKHP \sim \beta_0 + \beta_1 \log(PINCP) + \beta_2 OCCP + \beta_3 SEX + \beta_4 AGEP + \beta_5 COW + \beta_6 MAR + \beta_7 RAC1P + \beta_8 SCHL$

The model assumptions were that response variables (log(PINCP) or WKHP) were normally distributed, errors were normally distributed as $e_i \sim N(0, \sigma^2)$, and independent, and that there was constant variance σ^2. On the continuous variables the β coefficients have a single coefficient estimate. On the discrete variables the β coefficient indicators have a k-1 estimate count for all factorial (discrete) variables. In the equations above the β coefficients are labeled as single elements to save space due to the high count of factors (50).

Diagnostics:

After running the models above, I performed model diagnostics before proceeding with the analysis. During my model diagnostics, I noted there were streaks in the residuals. After examining the residuals, I realized that these streaks were likely due to the categorical nature of the data. Another reason for the streaks besides categorical was that some of the data was collected in buckets but was used as continuous. For example, age was rounded to whole numbers as provided. These issues brought some linearity to the residuals. With this in mind, I focused on the distribution spread of the residuals.

To check the performance of the four models (one for each question) I used the residuals versus fitted values and the residuals for the variables of interest versus the explanatory variable. The figures illustrating these checks are shown in Appendix A and described below.

The figures for Question 1 are shown in Appendix A Figures 6 and 7. They have no concerning patterns. The Normal QQ plot for the model evaluated whether the assumption of normality was violated. The tails in the QQ plot showed a deviation from linearity, suggesting there were outliers in the dataset. I considered removing the outliers from the dataset, but since I had already subsetted the data and removed missing values, I made the decision to include the outliers and continue with the analysis assuming normality was a reasonable approximation.

The fitted versus residuals plot and residuals vs 'SEX' plot for Question 2 are shown in Figures 8 and 9 in Appendix A and both show adequate fits to the data. The Normal QQ plot for model 2 in Figure 19 showed an even larger deviation from the approximation of normality from that of Question 1, particularly near the lower tail. I found the plot to indicate a reasonable approximation of normality.

The fitted versus residuals, residuals versus 'SCHL' and residuals versus 'WKHP' plots for Question 3 are included in Appendix A Figures 10 and 11, and show no discernable pattern. The Normal QQ plot in Figure 23 again showed a deviation from linearity in the lower tail, but I deemed this as reasonable for approximation of normality.

The fitted versus residuals and residuals versus 'SCHL' plots for Question 4 shown in Figures 12 and 13 in Appendix A again return no identifiable pattern, suggesting an adequate fit. The Normal QQ plot for Question 4 in Figure 26 Appendix A deviates from linearity on the upper tail, which contrasts with the other three models. I found the plot to indicate a reasonable approximation of normality.

To check model fit with so many different variables a penalized regression model with forward selection was used. The penalized regression formula I used on the models was:

$$\hat{\beta}_{LS} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_p X_{ip})^2$$

To perform the penalized regression models, I divided the dataset at random between a training set of 75% and a test set of 25%. The initial model chosen to address Questions 1 through 3 was log(PINCP) ~ 1 and WKHP ~1 for Question 4. The stepwise variable selection was done on the training set and then model comparison was done on the test set. I did this to minimize some of the bias brought by variable selection.

For Questions 1 through 3 the penalized regression model with forward selection came out with the full model as the lowest AIC at 4000.71. For Question 4, the penalized regression model with forward

selection found that the model WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR + RAC1P to have the lowest AIC score at 348,879.9. To verify I used an ANOVA function to test the full model versus a simpler model and found that the p-value < 2.2e-16 at the α=0.05 gave strong evidence to support the fuller model over the simpler model. Even though the fuller model was supported, it is important to note that the original fuller model would still be necessary to make inference since variable selection weakens inference and brings bias to the model selection process.

To complete the full process of penalized regression I proceeded with the following the model analysis. The elastic net Mean-Squared Error, Cumulative Distribution, Residuals versus Fitted, and Normal QQ plots for the full model with log(PINCP) as an output are shown in Appendix A Figure 27. The 'WKHP' results are shown in Figure 28. The estimates for the coefficients based on the elastic net returned only one coefficient estimate for the factored variables, which did not allow for any comparison to the coefficients found in the linear regression models. The adaptive lasso was also done, with the same handling of categorical variables. The estimates for the coefficients based on the adaptive lasso returned only one coefficient for the factored variables as well. These penalized regression models are viable for prediction but cannot be used to compare the coefficients for explanatory variables as they introduce bias, and do not handle categorical variables adequately. Based on the conclusions of the penalized linear models and wanting to test the hypotheses for Questions 1 through 4, I went back to my original models and proceeded forward with inference.

Results

Question 1 considered whether race affects total income. Based on the output of model 1, I failed to reject the null hypothesis that 'RAC1P' is equal to zero for one of the indicator variables (Other), while I reject the null hypothesis for Alaska Native, Native American, African American, American Indian, Asian, Native Hawaiian, and Two or More Races. The t-test for the indicator variables returned p<0.05 for the seven significant races. This suggests that race does in fact affect total income. When compared to the base value of White, all seven of the significant race indicator variable coefficients were negative, which also suggests they have a negative effect on total income.

Question 2 considered if sex had a significant effect on total income. Based on the output of the model, I rejected the null hypothesis that sex does not have a significant effect on total income. The coefficient for females (SEX2) is -0.1726 with a p-value of p<2.2e-16. This outcome of a t-test suggests that sex is significant in determining the total income of an individual, and that women tend to make less money than men.

Question 3 asked whether education 'SCHL' or hours worked per week 'WKHP' had a greater effect on total income. To compare these two variables both the estimated coefficient and t-tests were compared. The coefficient for 'SCHL' is 0.09169 while the coefficient for 'WKHP' is 0.0726. This suggests that education has a greater effect on total income compared to hours worked per week. Based on the t-test for the variables, both 'SCHL' and 'WKHP' had p<2e-16. This suggests that at the α=0.05 level both variables are significant, and therefore 'SCHL' has a greater effect on total income than 'WKHP', though both are significant to the model. A further test could be completed to find out if they are statistically different.

Below is the output results from Questions 1 through 3:

```
Call:
lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
    MAR + SEX + SCHL + RAC1P, data = q_data_fitlered)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.4094  -0.3591   0.1174   0.5115   4.2052

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.4313082  0.0386503 218.144  < 2e-16 ***
WKHP                         0.0372583  0.0002294 162.426  < 2e-16 ***
OCCPBUS                     -0.0076559  0.0196003  -0.391 0.696091
OCCPFIN                      0.0602947  0.0225602   2.673 0.007527 **
OCCPCMM                      0.1845301  0.0186890   9.874  < 2e-16 ***
OCCPENG                      0.1442320  0.0208234   6.926 4.34e-12 ***
OCCPSCI                     -0.1276777  0.0285955  -4.465 8.02e-06 ***
OCCPCMS                     -0.2783533  0.0223363 -12.462  < 2e-16 ***
OCCPLGL                      0.0842471  0.0294628   2.859 0.004245 **
OCCPEDU                     -0.4845826  0.0154037 -31.459  < 2e-16 ***
OCCPENT                     -0.2549069  0.0199911 -12.751  < 2e-16 ***
OCCPMED                      0.2429154  0.0153177  15.858  < 2e-16 ***
OCCPHLS                     -0.2816258  0.0216953 -12.981  < 2e-16 ***
OCCPPRT                     -0.2508554  0.0239307 -10.483  < 2e-16 ***
OCCPEAT                     -0.5688005  0.0154383 -36.843  < 2e-16 ***
OCCPCLN                     -0.5993640  0.0180379 -33.228  < 2e-16 ***
OCCPPRS                     -0.6626393  0.0165102 -40.135  < 2e-16 ***
OCCPSAL                     -0.3126988  0.0129253 -24.193  < 2e-16 ***
OCCPOFF                     -0.3295772  0.0122372 -26.932  < 2e-16 ***
OCCPFFF                     -0.7791762  0.0227185 -34.297  < 2e-16 ***
OCCPCON                     -0.2185854  0.0166600 -13.120  < 2e-16 ***
OCCPEXT                     -0.3202563  0.1498735  -2.137 0.032613 *
OCCPRPR                     -0.1987500  0.0197436 -10.067  < 2e-16 ***
OCCPPRD                     -0.3231783  0.0155414 -20.795  < 2e-16 ***
OCCPTRN                     -0.4800574  0.0152098 -31.562  < 2e-16 ***
OCCPMIL                     -0.5590036  0.1017082  -5.496 3.89e-08 ***
AGEP                         0.0197456  0.0002590  76.239  < 2e-16 ***
SCHL2                        0.0049126  0.0001303  37.702  < 2e-16 ***
COW2                        -0.0684248  0.0103325  -6.622 3.56e-11 ***
COW3                         0.0382712  0.0122138   3.133 0.001728 **
COW4                        -0.0033437  0.0132916  -0.252 0.801376
COW5                         0.1055012  0.0197836   5.333 9.69e-08 ***
COW6                        -0.4280250  0.0115107 -37.185  < 2e-16 ***
COW7                        -0.0666941  0.0149839  -4.451 8.55e-06 ***
COW8                        -0.8857681  0.0482307 -18.365  < 2e-16 ***
MAR2                         0.0322407  0.0253088   1.274 0.202704
MAR3                        -0.1031376  0.0091424 -11.281  < 2e-16 ***
MAR4                        -0.1613203  0.0230439  -7.001 2.57e-12 ***
MAR5                        -0.2930468  0.0077585 -37.771  < 2e-16 ***
SEXFemale                   -0.1725801  0.0065376 -26.398  < 2e-16 ***
SCHL                        -0.0916924  0.0040750 -22.501  < 2e-16 ***
RAC1PAfrican American alone -0.1587586  0.0239979  -6.616 3.72e-11 ***
RAC1PAmerican Indian alone  -0.1412517  0.0275943  -5.119 3.08e-07 ***
RAC1PAlaska Native alone    -0.2423427  0.1162461  -2.085 0.037096 *
RAC1PNative American        -0.1633178  0.0633028  -2.580 0.009883 **
RAC1PAsian alone            -0.0655797  0.0144246  -4.546 5.46e-06 ***
RAC1PNative Hawaiian alone  -0.1729751  0.0509590  -3.394 0.000688 ***
RAC1POther                  -0.0206986  0.0187939  -1.101 0.270746
RAC1PTwo or More Races      -0.1123483  0.0153910  -7.300 2.91e-13 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8845 on 97255 degrees of freedom
Multiple R-squared:  0.5015, Adjusted R-squared:  0.5013
F-statistic:  2039 on 48 and 97255 DF,  p-value: < 2.2e-16
```

Question 4 asked what the effect of education was on hours worked per week. Based on the output of its model, I rejected the null hypothesis that the coefficient for the ‘SCHL’ variable in the model was equal to zero based on a t-test statistic of t= -8.993 with a p-value < 2e-16. This suggests that there is a correlation between hours worked per week and education. The coefficient for ‘SCHL’ in the results was -0.1385, which suggests the higher the level of education one has, the lower their average reported hours worked per week. The data could further be put into a matrix, standardized and then used with canonical correlation analysis to determine the significance of correlation.

Below is the output for Question 4:

```
Call:
lm(formula = WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR +
    RAC1P + SCHL, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-53.025  -5.313  -0.043   4.834 112.405

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -12.538144   0.514085 -24.389  < 2e-16 ***
log(PINCP)                   5.816551   0.040118 144.988  < 2e-16 ***
OCCPBUS                     -3.018009   0.277953 -10.858  < 2e-16 ***
OCCPFIN                     -2.878983   0.321436  -8.957  < 2e-16 ***
OCCPCMM                     -4.807460   0.266079 -18.068  < 2e-16 ***
OCCPENG                     -3.518506   0.297946 -11.809  < 2e-16 ***
OCCPSCI                     -3.612686   0.402527  -8.975  < 2e-16 ***
OCCPCMS                     -2.963511   0.317985  -9.320  < 2e-16 ***
OCCPLGL                     -3.059615   0.417074  -7.336 2.23e-13 ***
OCCPEDU                     -4.021827   0.219066 -18.359  < 2e-16 ***
OCCPENT                     -5.249485   0.284418 -18.457  < 2e-16 ***
OCCPMED                     -5.593347   0.216921 -25.785  < 2e-16 ***
OCCPHLS                     -3.167979   0.310277 -10.210  < 2e-16 ***
OCCPPRT                     -1.647446   0.337905  -4.875 1.09e-06 ***
OCCPEAT                     -4.827093   0.220656 -21.876  < 2e-16 ***
OCCPCLN                     -5.196828   0.257016 -20.220  < 2e-16 ***
OCCPPRS                     -2.057685   0.235734  -8.729  < 2e-16 ***
OCCPSAL                     -2.889915   0.183776 -15.725  < 2e-16 ***
OCCPOFF                     -3.234222   0.173831 -18.606  < 2e-16 ***
OCCPFFF                      3.977380   0.327290  12.152  < 2e-16 ***
OCCPCON                     -2.563473   0.236856 -10.823  < 2e-16 ***
OCCPEXT                     12.123384   1.996227   6.073 1.26e-09 ***
OCCPRPR                     -1.749656   0.282845  -6.186 6.21e-10 ***
OCCPPRD                     -0.983336   0.220494  -4.460 8.22e-06 ***
OCCPTRN                     -0.967860   0.217084  -4.458 8.27e-06 ***
OCCPMIL                      5.687701   1.407137   4.042 5.30e-05 ***
SEXFemale                   -2.505476   0.092934 -26.960  < 2e-16 ***
AGEP                        -0.060578   0.003792 -15.977  < 2e-16 ***
```

```
COW2                           -1.196330   0.147135  -8.131 4.33e-16 ***
COW3                           -0.233781   0.173832  -1.345 0.178672
COW4                            0.158742   0.188629   0.842 0.400039
COW5                            0.491322   0.282416   1.740 0.081915 .
COW6                           -1.309132   0.165282  -7.921 2.40e-15 ***
COW7                            0.855102   0.214111   3.994 6.51e-05 ***
COW8                           -2.749235   0.693698  -3.963 7.40e-05 ***
MAR2                           -1.554162   0.359256  -4.326 1.52e-05 ***
MAR3                            0.734924   0.130176   5.646 1.65e-08 ***
MAR4                            1.212819   0.328445   3.693 0.000222 ***
MAR5                           -0.884206   0.111280  -7.946 1.96e-15 ***
RAC1PAfrican American alone     0.720972   0.339629   2.123 0.033772 *
RAC1PAmerican Indian alone      1.453307   0.390698   3.720 0.000200 ***
RAC1PAlaska Native alone        0.383846   1.559779   0.246 0.805613
RAC1PNative American            1.824404   0.889464   2.051 0.040258 *
RAC1PAsian alone                0.063880   0.205457   0.311 0.755864
RAC1PNative Hawaiian alone      3.113337   0.718217   4.335 1.46e-05 ***
RAC1POther                      1.180797   0.265542   4.447 8.73e-06 ***
RAC1PTwo or More Races          0.316507   0.219601   1.441 0.149509
SCHL                           -0.138533   0.015404  -8.993  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.91 on 72930 degrees of freedom
Multiple R-squared:  0.3401, Adjusted R-squared:  0.3397
F-statistic: 799.9 on 47 and 72930 DF,  p-value: < 2.2e-16
```

Obstacles

During the model fitting process, the Normal QQ plots for models 1 through 3 with log(Income) as the response variable all returned lower tails that deviated from linearity. This could be concerning as the approximation to normality may not be reasonable for the analysis if a small N was provided. Bet I felt with a large enough N and the concept of central limit theorem that this was not an issue. Based on this I tried many different transformations of the response and explanatory variables to improve the Normal QQ plots. The best result was found when the quadratic form of ‘SCHL’ was used, but the deviation from linearity was still present.

When variable selection was utilized via forward selection to identify the best models to answer the questions of interest, the first three models with the log(Income) as the response variable returned the full model as the best fit. This suggests that the training dataset was too large or that the full model was truly the best performing model.

Incorporating the penalized linear regression proved to be difficult. Because the penalized regression models are estimated using different packages in R such as ‘glmnet’ or ‘adalasso’, after the model solved for the coefficients the results differed from that of the generalized linear models. The residuals versus fitted values for the overall model and the Normal QQ plot were created, though the interpretation of individual coefficients could not be made. The ability to handle categorical variables in the penalized regression would entail creating dummy variables for each of the factored categorical variables.

Conclusion

In the state of Oregon, using the 5-year Public Use Microdata Sample (PUMS) Population dataset for 2013-2017, there is statistical difference between income and explanatory variables for individuals between the ages of 16 and 67. Race as an explanatory variable had a significant difference in log of income for Alaska Native, Native American, African American, American Indian, Asian, Native Hawaiian, and Two or More Races compared to the baseline of White with all having a p-value less than 0.05. Females were found to have a significantly lower log transformed income than that of males with a p-value of $p<2.2e-16$. Level of education and hours worked per week both played a statistically significant role in explaining the log income, with both having a p-value less than 0.05. However, to answer the original question it should be noted that education had a larger impact on log income with a coefficient of -0.09117, while the coefficient for hours worked was 0.03725. A t-test between the two explanatory variables shows a statistically significant difference with a p-value less than 0.05. Lastly, the impact of education on hours worked per week resulted in a correlation between the two variables. The model showed a t-test statistic of $t=-8.993$ with a p-value $< 2e-16$ with the coefficient of -0.1386 which indicates the higher the level of education one has, the lower their average reported hours worked per week is.

While the analysis was conducted on interesting social discussions, penalized linear models and non-linear decision tree models were also explored, but were unable to assist with answering the questions of interest posed. Going forward, those models could be utilized to assist with predictions. Now that I completed the analysis to answer the questions of interest, there are many great places that this analysis could continue going forward. Directions in which this analysis could continue with: Which of the explanatory variables is most impactful? Do combinations of these variables play more of a role over others? How does 2013-2017 compare against years past, or to other states? With such a large and complex dataset, the possibilities are nearly endless.

Appendix A

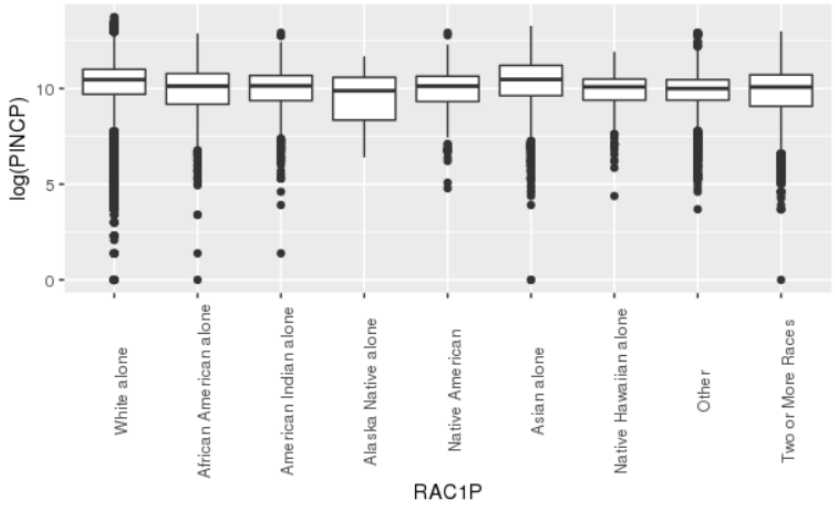

Figure 1: Log of total income versus race.

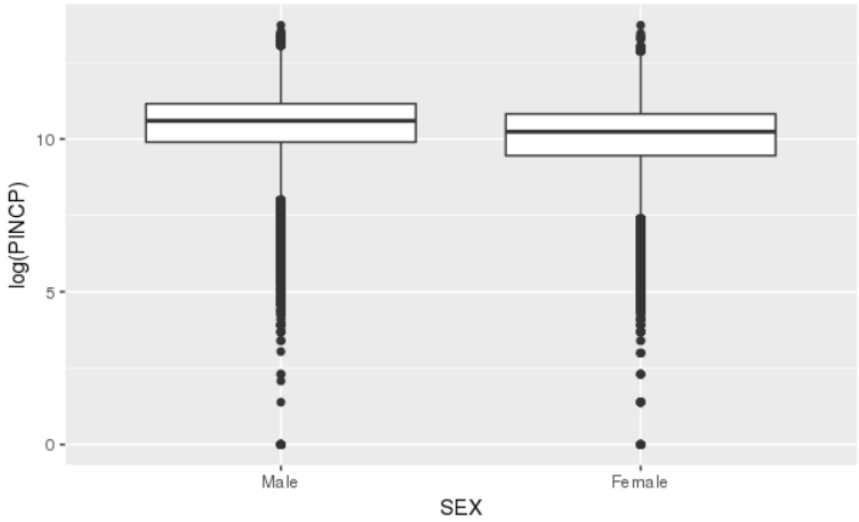

Figure 2: Log of total income versus sex.

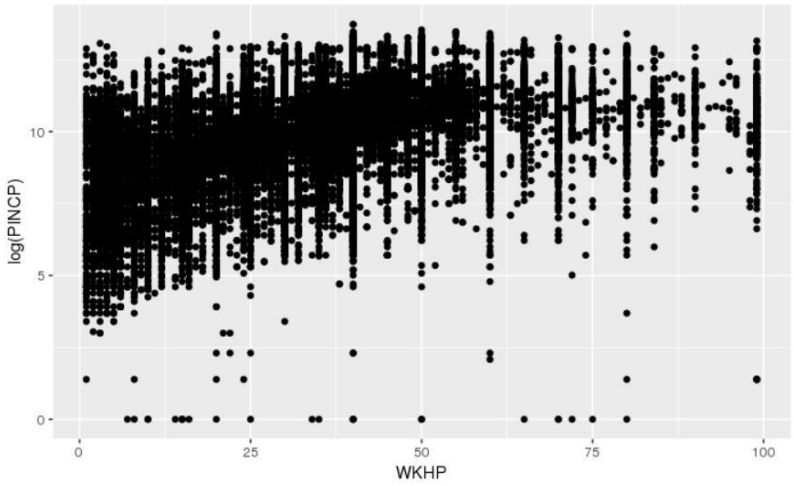

Figure 3: Log of total income versus hours worked per week.

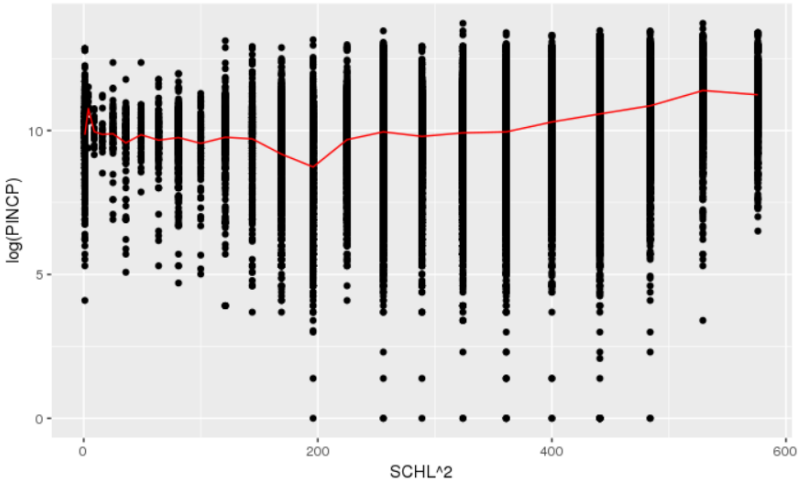

Figure 4: Log of total income versus quadratic form of education.

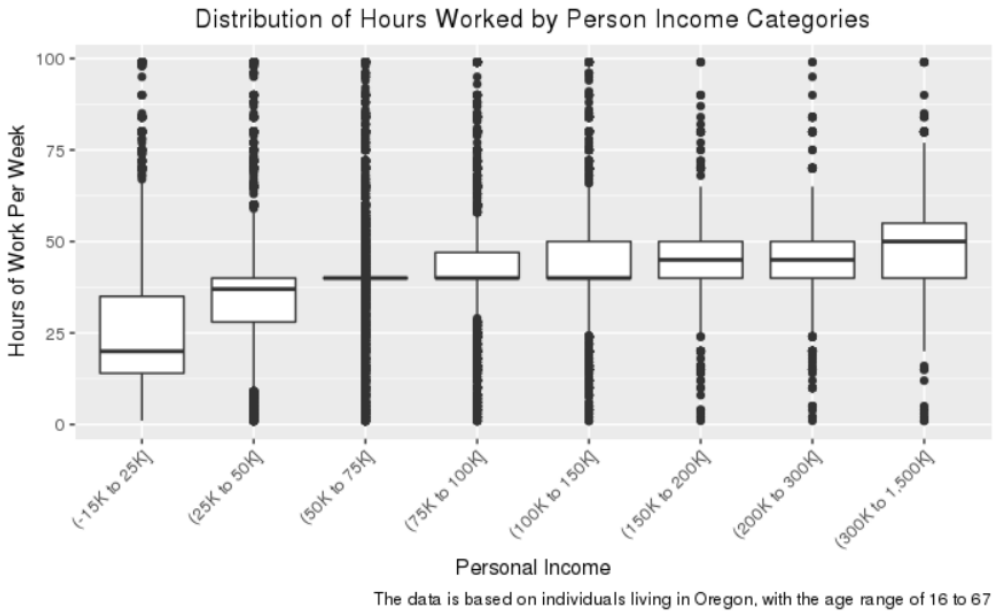

Figure 5: Hours worked per week versus education.

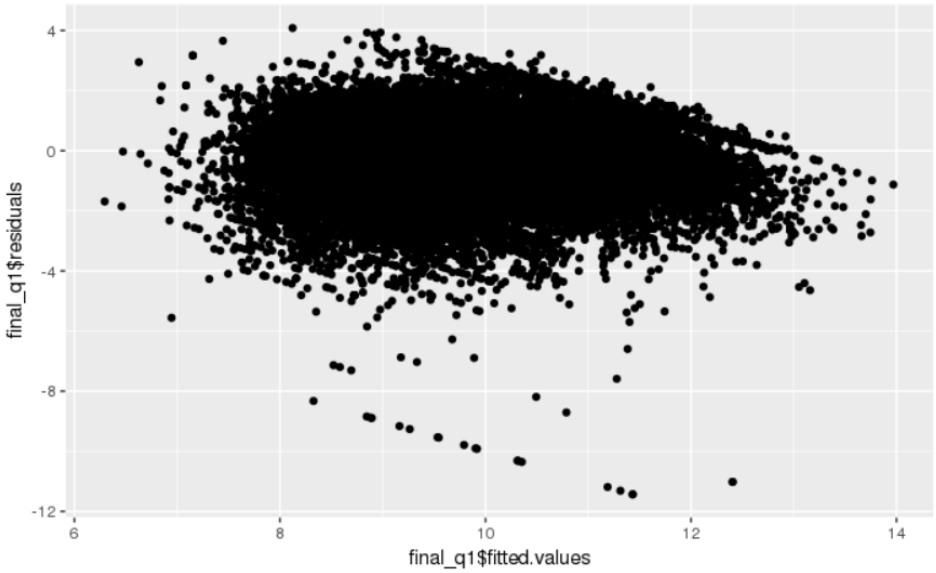

Figure 6: Fitted versus Residuals for model 1.

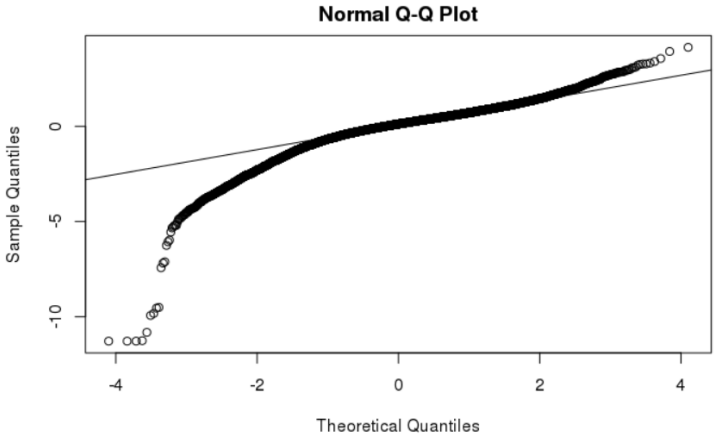

Figure 7: Normal QQ plot for model 1.

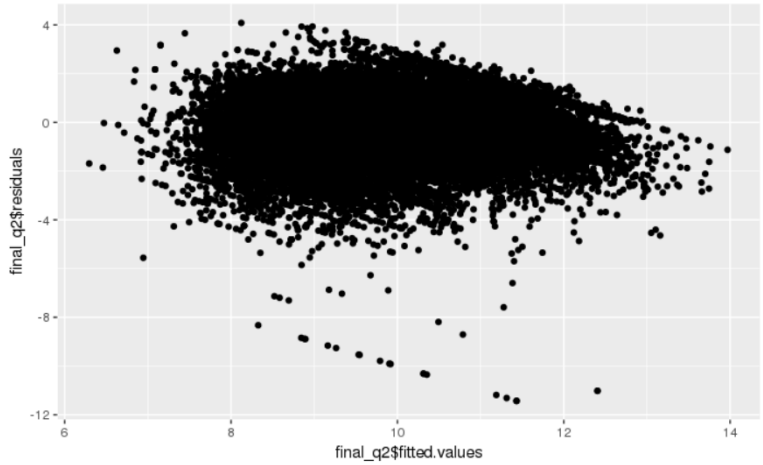

Figure 8: Fitted versus Residuals for model 2.

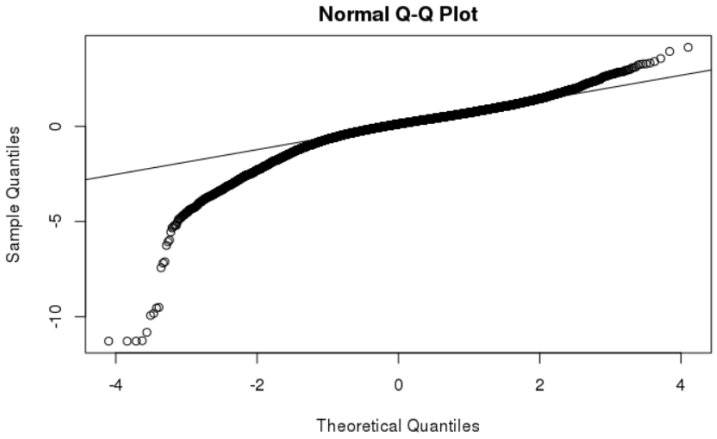

Figure 9: Normal QQ plot for training dataset model 2.

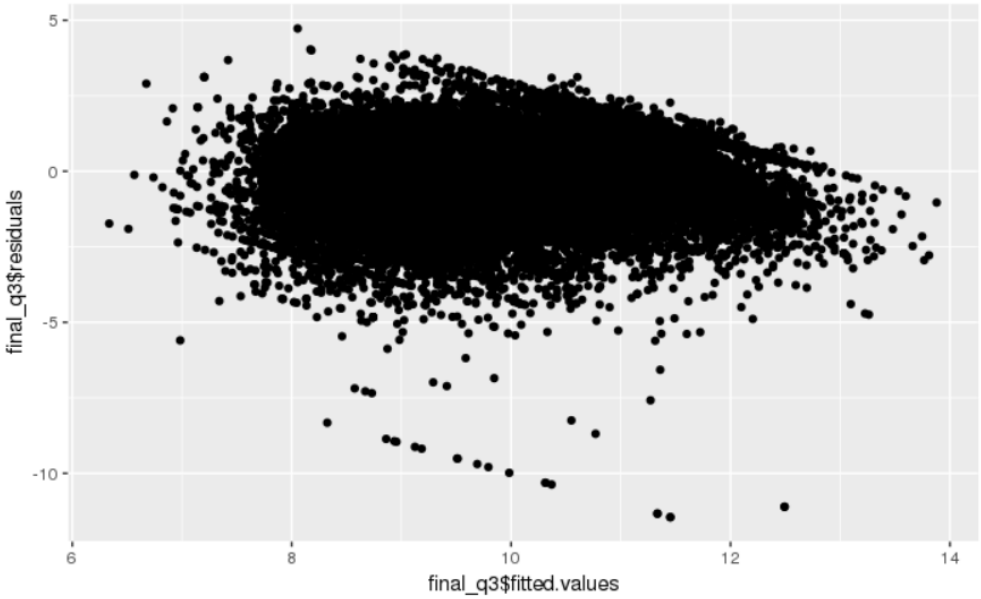

Figure 10: Fitted versus residual values for training dataset model 3.

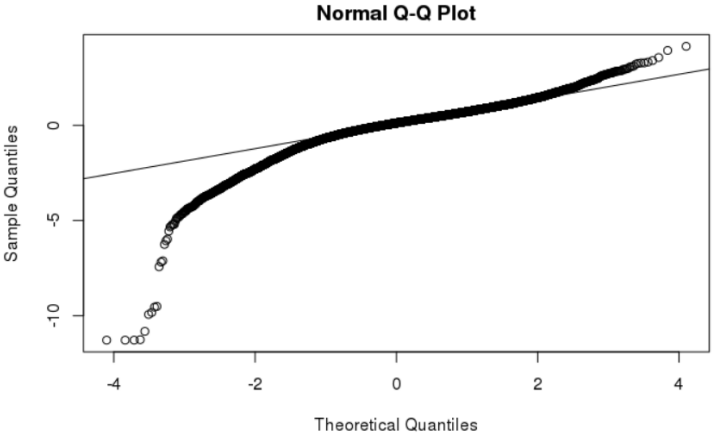

Figure 11: Normal QQ plot for training dataset model 3.

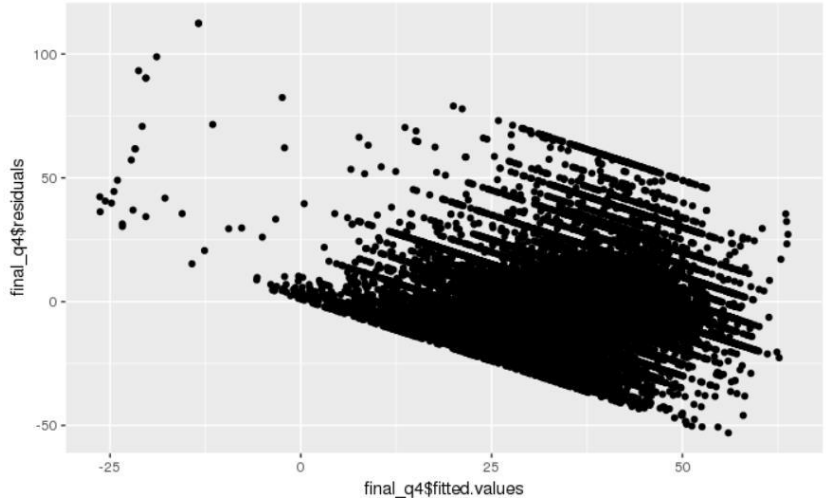

Figure 12: Fitted versus Residual Values for model 4.

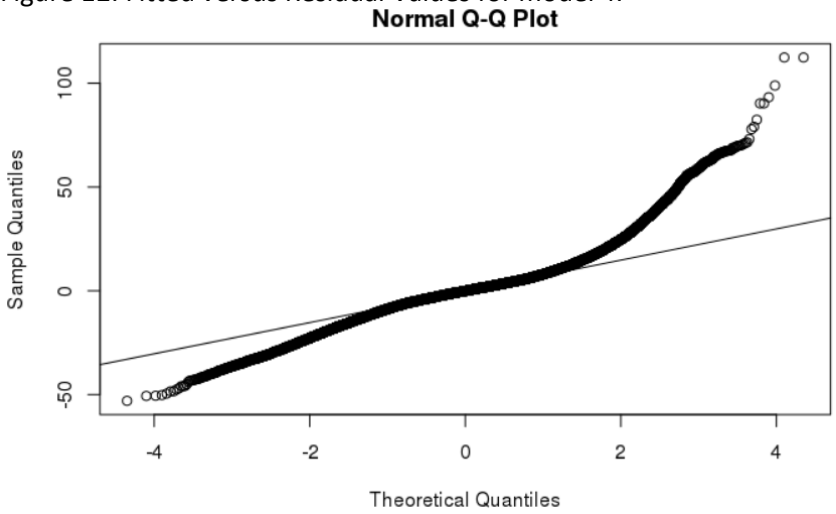

Figure 13: Normal QQ plot for model 4 using training data.

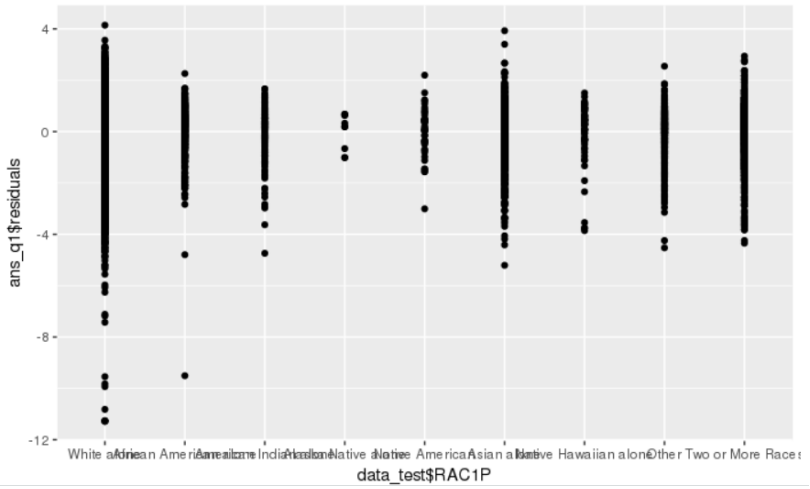

Figure 15: Residuals for RAC1P versus observed values model 1.

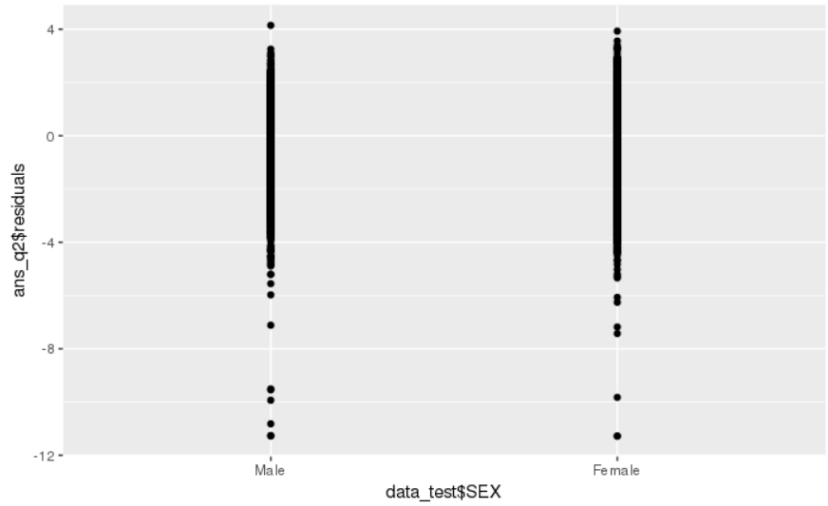

Figure 18: Residuals for SEX versus observed values model 2.

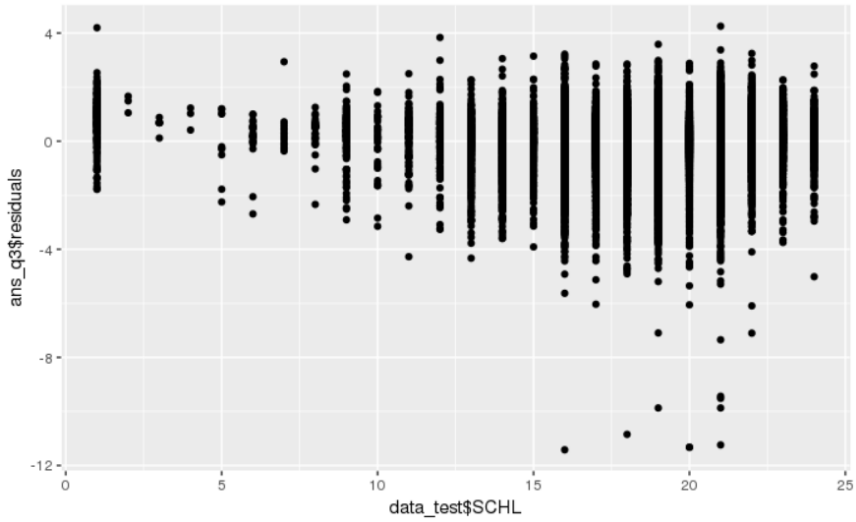

Figure 21: Residuals for SCHL versus observed values model 3.

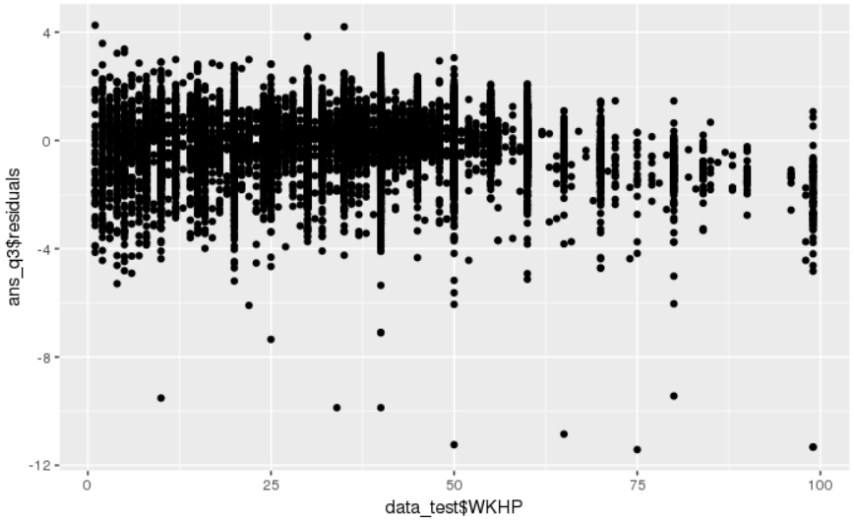

Figure 22: Residuals for WHKP versus observed values model 3.

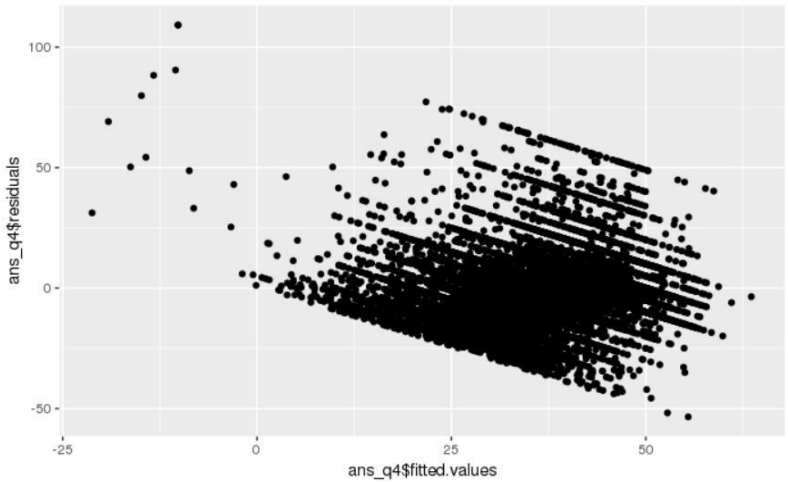

Figure 24: Residuals vs Fitted values for model 4 using test data.

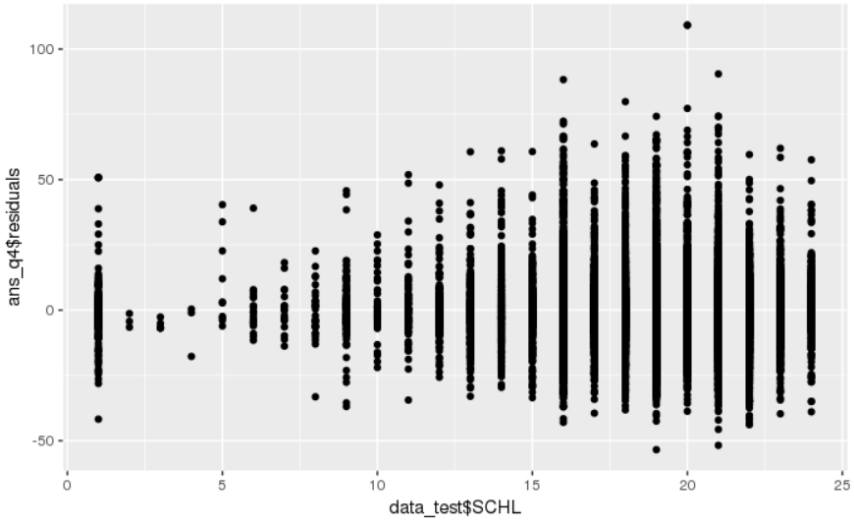

Figure 25: Residuals for SCHL versus observed values model 4.

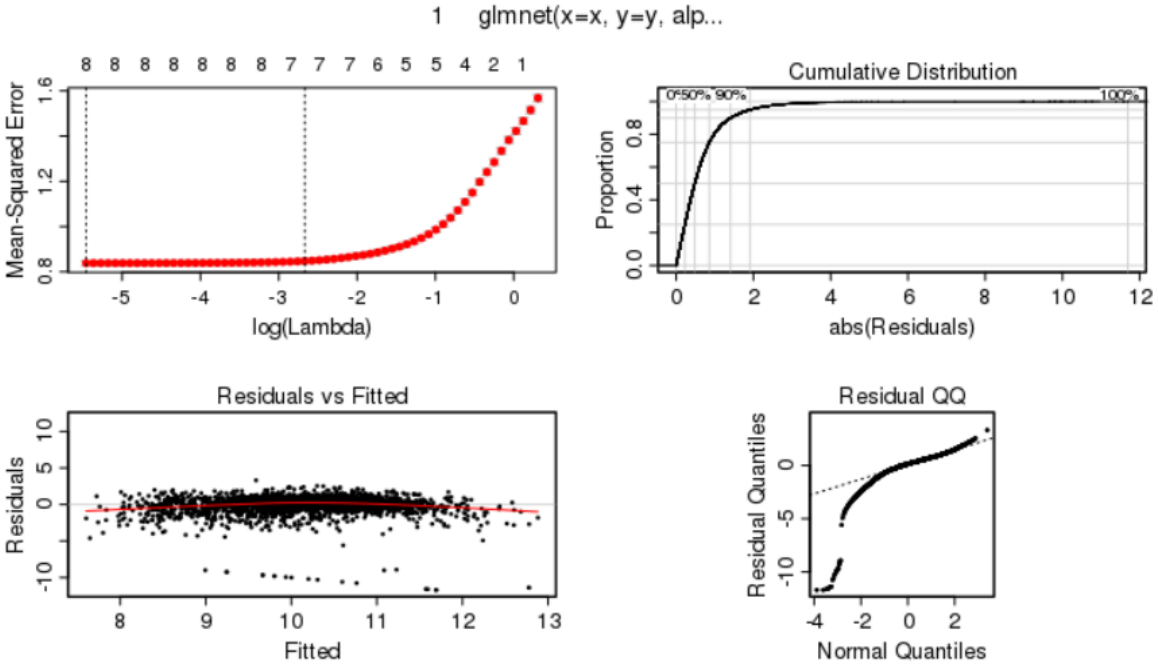

Figure 27: Elastic Net Mean-Squared Error, Cumulative Distribution, Residuals versus Fitted and Normal QQ plots for the full model with log(PINCP).

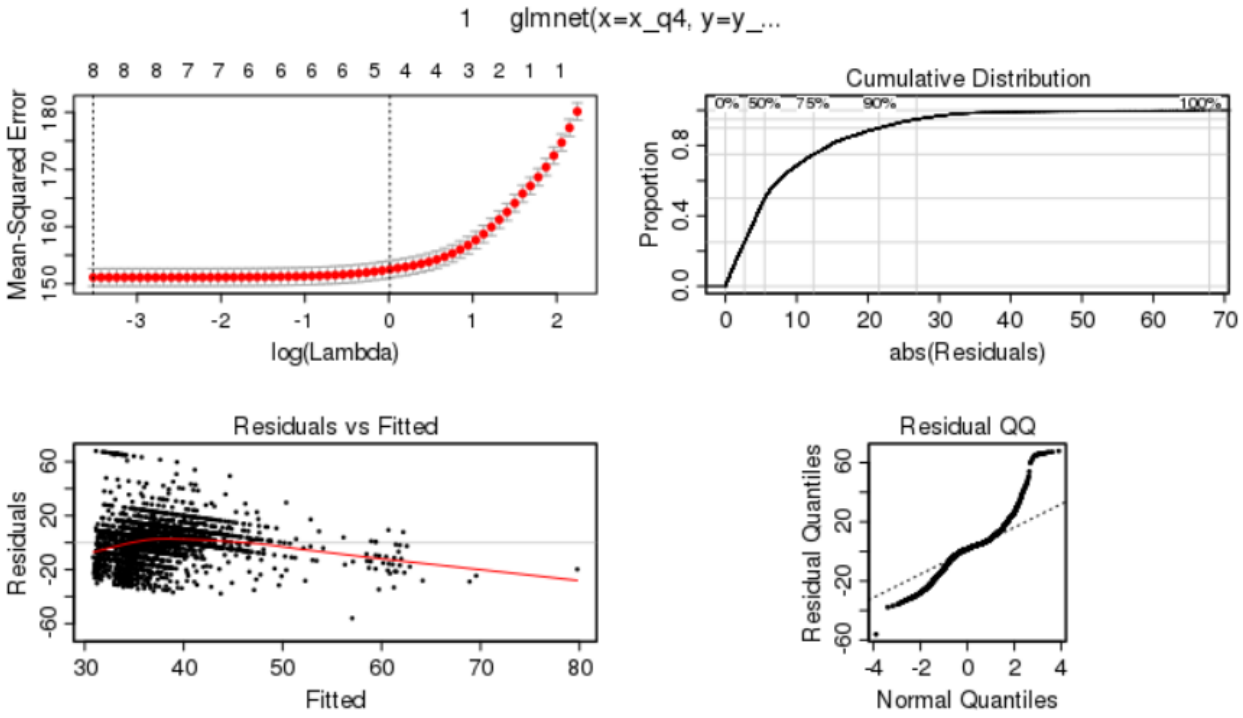

Figure 28: Elastic Net Mean-Squared Error, Cumulative Distribution, Residuals versus Fitted and Normal QQ plots for the full model with WKHP.

Appendix B - Source Code

```
data <- read.table("psam_p41.csv", header=T, quote="\"", sep=",")

q_data<- data %>% dplyr::select("SEX", "AGEP", "SCHL", "RAC1P", "MAR", "WKHP", "PINCP", "OCCP",
"COW")
q_data <- as_tibble(q_data)
summary(q_data)
q_data_fitlered <- q_data %>% filter(AGEP > 16 , AGEP < 67, !is.na(SCHL),!is.na(WKHP), PINCP >0)
summary(q_data %>% filter(AGEP > 16 , AGEP < 67))
q_data_fitlered$RAC1P <- as.factor(q_data_fitlered$RAC1P)
q_data_fitlered$SEX <- as.factor(q_data_fitlered$SEX)
q_data_fitlered$MAR <- as.factor(q_data_fitlered$MAR)
# q_data$OCCP <- as.factor(q_data$OCCP)
q_data_fitlered$COW <- as.factor(q_data_fitlered$COW)
q_data_fitlered$OCCP
q_data_fitlered$OCCP <- cut(q_data_fitlered$OCCP,
                    breaks=c(0, 430, 740, 950, 1240,1560,
                          1965,2060,2160,2550,2920,3540,
                          3655,3955,4150,4250,4650, 4965,
                          5940,6130,6765,6940,7630,8965,
                          9750,9830,Inf),
                    labels=c("MGR", "BUS","FIN","CMM","ENG","SCI",
                             "CMS", "LGL", "EDU", "ENT", "MED", "HLS",
```

```
                              "PRT","EAT", "CLN", "PRS","SAL", "OFF",
                              "FFF","CON", "EXT", "RPR", "PRD","TRN",
                              "MIL","OTHER"))
#class(q_data_fitlered$OCCP)
ethnicity <- c("White alone", "African American alone", "American Indian alone", "Alaska Native alone",
"Native American", "Asian alone", "Native Hawaiian alone", "Other", "Two or More Races")

levels(q_data_fitlered$RAC1P) <- ethnicity
levels(q_data_fitlered$SEX) <- c("Male", "Female")

#ggplot(q_data_fitlered,aes(RAC1P,PINCP))+geom_boxplot()
ggplot(q_data_fitlered,aes(RAC1P,log(PINCP)))+
  theme(axis.text.x = element_text(angle = 90)) +geom_boxplot()

#ggplot(q_data_fitlered,aes(SEX,PINCP))+geom_boxplot()
ggplot(q_data_fitlered,aes(SEX,log(PINCP)))+geom_boxplot()

#qplot(WKHP, PINCP, data = q_data_fitlered)
qplot(WKHP, log(PINCP), data = q_data_fitlered)

#qplot(SCHL, PINCP, data = q_data_fitlered)
#qplot(SCHL^2, PINCP, data = q_data_fitlered)
#ggplot(q_data_fitlered,aes(SCHL,log(PINCP)))+ geom_point()+
 # stat_summary(aes(y=log(PINCP),group=1), fun.y=mean,geom='line',color="red",group=1)
ggplot(q_data_fitlered,aes(SCHL^2,log(PINCP)))+ geom_point()+
  stat_summary(aes(y=log(PINCP),group=1), fun.y=mean,geom='line',color="red",group=1)
q_data_fitlered$SCHL2 <- (q_data_fitlered$SCHL)^2

ggplot(q_data_fitlered,aes(x=reorder(q_data_fitlered$OCCP, q_data_fitlered$PINCP, mean),log(PINCP),
fill=SEX))+
geom_boxplot()+
scale_fill_manual(labels=c("Male", "Female"),values=c("green", "light blue"))+
theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position = "bottom" )+
labs(title = "Log Transformed Personal Income by Occupation", subtitle = "Subset by Gender", y ="Log
Transformed Personal Income", x="Grouped Occupation")

ggplot(q_data_fitlered,aes(COW,log(PINCP)))+geom_boxplot()

ggplot(q_data_fitlered,aes(MAR,log(PINCP)))+geom_boxplot()

qplot(AGEP, log(PINCP), data = q_data_fitlered)

```{r}

q4_plot_data <- q_data_fitlered

#Verify that it filtered
print(q4_plot_data)
```
```

```
#Setting the boundaries for the categories
max(q4_plot_data$WKHP)
min(q4_plot_data$WKHP)
max(q4_plot_data$PINCP)
min(q4_plot_data$PINCP)

#q4_plot_data$WKHP<-cut(q4_plot_data$WKHP, c(0, 10, 20, 30, 40, 50, 60, 70, 99))
q4_plot_data$PINCP<-cut(q4_plot_data$PINCP, c(-15000, 25000, 50000, 75000, 10000, 150000,
200000, 300000, 1500000))

#print(q4_plot_data)

#q4_plot_data %>%
#  ggplot()+
#  geom_bar(aes(WKHP ,..count..))

q4_plot_data %>%
  ggplot()+
  geom_boxplot(aes(PINCP ,WKHP))+
  scale_x_discrete(labels=c("(-15K to 25K]", "(25K to 50K]", "(50K to 75K]", "(75K to 100K]", "(100K to
150K]", "(150K to 200K]", "(200K to 300K]", "(300K to 1,500K]")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),plot.title = element_text(hjust = 0.5))+
  labs(x="Personal Income", y="Hours of Work Per Week",title="Distribution of Hours Worked by Person
Income Categories",
       caption="The data is based on individuals living in Oregon, with the age range of 16 to 67")
```

```{r}
set.seed(2020)

(n <- dim(q_data_fitlered)[1])
(r <- round(n * .75))
idx <- 1:n
nidx <- sample(idx, r, replace = FALSE)
data_train <- q_data_fitlered[nidx, ]
data_test <- q_data_fitlered[-nidx, ]

```

```{r}

final_q1_q3 <- lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
    MAR + SEX + SCHL + RAC1P, data = q_data_fitlered)

final_q4_a <- lm(formula = WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR + RAC1P + SCHL,
            data = data_train)
```

```
summary(final_q1_q3)
summary(final_q4_a)

```

Question 1
```{r}
q1_lower <- lm(log(PINCP) ~ 1, data=data_train)
q1_upper <- lm(log(PINCP) ~ SEX + AGEP + SCHL + SCHL2 + MAR + WKHP + COW + OCCP,
data=data_train)
step(q1_lower, scope=list(upper=q1_upper, lower=q1_lower),direction="forward",test="F")

q1_tree <- data_train

q1_tree$PINCP2<-cut(q1_tree$PINCP, c(0, 25000, 50000, 75000, 10000, 150000, 200000, 300000,
1500000))
tr <- tree(PINCP2 ~ COW  + WKHP + SCHL + AGEP + SEX + MAR + RAC1P, data = q1_tree)
summary(tr)

plot(tr); text(tr)

```

```{R}
# Final model
# + RAC1P
final_q1 <- lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
   MAR + SEX + SCHL + RAC1P, data = data_train)

# Residual vs Fitted
qplot(final_q1 $fitted.values, final_q1 $residuals)

# Residual VS Explanatory
qplot(data_train$WKHP, final_q1$residuals)
qplot(data_train$AGEP, final_q1$residuals)
qplot(data_train$SCHL, final_q1$residuals)

# QQ Plot
par(mfrow = c(1, 1))
qqnorm(final_q1$residuals)
qqline(final_q1$residuals)
```
The plot of Residual vs Fitted looks good, but normality is violated.

```{R}
# To answer Q1
ans_q1 <- lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
   MAR + SEX + SCHL + RAC1P, data = data_test)
```

```
summary(ans_q1)
# Residual vs Fitted
qplot(ans_q1$fitted.values, ans_q1$residuals)

# Residual VS Explanatory
qplot(data_test$WKHP, ans_q1$residuals)
qplot(data_test$AGEP, ans_q1$residuals)
qplot(data_test$SCHL, ans_q1$residuals)
qplot(data_test$RAC1P, ans_q1$residuals)


# QQ Plot
par(mfrow = c(1, 1))
qqnorm(ans_q1$residuals)
qqline(ans_q1$residuals)
```

Question 2
```{r}
q2_lower <- lm(log(PINCP) ~ 1, data=data_train)
q2_upper <- lm(log(PINCP) ~ RAC1P + AGEP + SCHL + SCHL2 + MAR + WKHP + COW + OCCP,
data=data_train)
step(q2_lower, scope=list(upper=q2_upper, lower=q2_lower),direction="forward", test="F")
```

```{R}
# Final model
# + SEX
final_q2 <- lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
   MAR + SCHL + RAC1P + SEX, data = data_train)


test <- lm(formula = log(PINCP) ~WKHP + OCCP, data = data_train)
anova(test, final_q2)

# Residual vs Fitted
qplot(final_q2$fitted.values, final_q2$residuals)

# Residual VS Explanatory
qplot(data_train$WKHP, final_q2$residuals)
qplot(data_train$AGEP, final_q2$residuals)
qplot(data_train$SCHL, final_q2$residuals)
qplot(data_train$SCHL2, final_q2$residuals)
# QQ Plot
par(mfrow = c(1, 1))
qqnorm(final_q2$residuals)
qqline(final_q2$residuals)
```

Fitted resisual plot looks good, normality is violated.

```{R}
# To answer Q2
ans_q2 <- lm(formula = log(PINCP) ~ WKHP + OCCP + AGEP + SCHL2 + COW +
   MAR + SCHL + RAC1P + SEX, data = data_test)
summary(ans_q2)
# Residual vs Fitted
qplot(ans_q2$fitted.values, ans_q2$residuals)

# Residual VS Explanatory
qplot(data_test$WKHP, ans_q2$residuals)
qplot(data_test$AGEP, ans_q2$residuals)
qplot(data_test$SCHL, ans_q2$residuals)
qplot(data_test$SEX, ans_q2$residuals)


# QQ Plot
par(mfrow = c(1, 1))
qqnorm(ans_q2$residuals)
qqline(ans_q2$residuals)
```

Question 3
```{r}
q3_lower <- lm(log(PINCP) ~ 1, data=data_train)
q3_upper <- lm(log(PINCP) ~ SEX + AGEP + MAR + COW + OCCP + RAC1P, data=data_train)
step(q3_lower, scope=list(upper=q3_upper, lower=q3_lower),direction="forward", test="F")

```



```{R}
# Final model
# + SCHL, +WKHP
final_q3 <- lm(formula = log(PINCP) ~ OCCP + AGEP + COW + MAR + SEX + RAC1P + SCHL+ SCHL2 +
WKHP, data = data_train)

# Residual vs Fitted
qplot(final_q3$fitted.values, final_q3$residuals)

# Residual VS Explanatory
qplot(data_train$WKHP, final_q3$residuals)
qplot(data_train$AGEP, final_q3$residuals)
qplot(data_train$SCHL, final_q3$residuals)
qplot(data_train$SCHL2, final_q3$residuals)
# QQ Plot
par(mfrow = c(1, 1))
```

```
qqnorm(final_q3$residuals)
qqline(final_q3$residuals)
```

Again, fitted resisual plot looks good, normality is violated.

```{R}
# To answer Q3
ans_q3 <- lm(formula = log(PINCP) ~ OCCP + AGEP + COW + MAR + SEX + RAC1P + SCHL + SCHL2 + WKHP, data = data_test)

test <- lm(formula = log(PINCP) ~ OCCP + AGEP + COW + MAR + SEX + RAC1P, data = data_test)

summary(ans_q3)

anova(test, ans_q3)

# Residual vs Fitted
qplot(ans_q3$fitted.values, ans_q3$residuals)

# Residual VS Explanatory
qplot(data_test$AGEP, ans_q3$residuals)
qplot(data_test$SCHL, ans_q3$residuals)
qplot(data_test$WKHP, ans_q3$residuals)

# QQ Plot
par(mfrow = c(1, 1))
qqnorm(ans_q3$residuals)
qqline(ans_q3$residuals)
```

Issues for 1-3:
Normality is violated;
Constant vairance may be violated;
All three questions are answered by the full model, train set is too large?

Question 4

To find potential relationship between varaibles and proper forms of variables included in linear models, based on following plots for Q4:

```{r}
ggplot(q_data_fitlered,aes(SEX,WKHP))+geom_boxplot()

ggplot(q_data_fitlered,aes(OCCP,WKHP))+geom_boxplot()+ theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggplot(q_data_fitlered,aes(COW,WKHP))+geom_boxplot()
```

```
ggplot(q_data_fitlered,aes(RAC1P,WKHP))+geom_boxplot()

ggplot(q_data_fitlered,aes(MAR,WKHP))+geom_boxplot()

ggplot(q_data_fitlered,aes(AGEP, WKHP))+ geom_point()+
  stat_summary(aes(y=log(PINCP),group=1), fun.y=mean,geom='line',color="red",group=1)

ggplot(q_data_fitlered,aes(SCHL, WKHP))+ geom_point()+
  stat_summary(aes(y=SCHL,group=1), fun.y=mean,geom='line',color="red",group=1)

ggplot(q_data_fitlered,aes(log(PINCP), WKHP))+ geom_point()+
  stat_summary(aes(y=log(PINCP),group=1), fun.y=mean,geom='line',color="red",group=1)
```

```
# Question 4
```{r}
q4_lower <- lm(WKHP ~ 1, data=data_train)
q4_upper <- lm(WKHP ~ SEX + AGEP + MAR + COW + log(PINCP) + RAC1P + OCCP, data=data_train)
step(q4_lower, scope=list(upper=q4_upper, lower=q4_lower),direction="forward")

```

```{R}
# Final model
# + SCHL
final_q4 <- lm(formula = WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR + RAC1P + SCHL,
            data = data_train)
test <- lm(formula = WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR +  RAC1P + SCHL,
            data = data_train)
summary(final_q4)
anova(test, final_q4)

# Residual vs Fitted
qplot(final_q4$fitted.values, final_q4$residuals)

# Residual VS Explanatory
log_PINCP <- log(data_train$PINCP)
qplot(log_PINCP, final_q4$residuals)
qplot(data_train$AGEP, final_q4$residuals)
qplot(data_train$SCHL, final_q4$residuals)

# QQ Plot
par(mfrow = c(1, 1))
qqnorm(final_q4$residuals)
qqline(final_q4$residuals)
```
```

patterns of residual is not explained by current variables.

```{R}
# To answer Q4
ans_q4 <- lm(formula = WKHP ~ log(PINCP) + OCCP + SEX + AGEP + COW + MAR + RAC1P + SCHL, data =
data_test)
summary(ans_q4)
# Residual vs Fitted
qplot(ans_q4$fitted.values, ans_q4$residuals)

# Residual VS Explanatory
qplot(data_test$WKHP, ans_q4$residuals)
qplot(data_test$AGEP, ans_q4$residuals)
qplot(data_test$SCHL, ans_q4$residuals)

# QQ Plot
par(mfrow = c(1, 1))
qqnorm(ans_q4$residuals)
qqline(ans_q4$residuals)
```

```{R}
# Try penalized models

#install.packages("plotmo")
library(glmnet)
library(ncvreg)
library(parcor)
library(mht)
library(plotmo)

# Q1-Q3
x <- data.matrix(data_train[,1:6])
z <- data.matrix(data_train[,8:9])
x <- cbind(x, z)
y <- log(as.matrix(data_train[,7]))

## issue: factors are suspected to treat as numerical???
elnet <- glmnet(x, y, alpha = 0.5)
elnet.cv <- cv.glmnet(x, y, alpha = 0.5)
# Finding the Elastic net estimates for lambda with minimum cross-validation error
elnet1 <- glmnet(x, y, lambda = elnet.cv$lambda.min, alpha = 0.5)
# Elastic net estimates of coefficients
elnet1$beta

plotres(elnet.cv)

##### Adaptive Lasso Estimates
```

```
adlasso <- adalasso(x, y)
adlasso$coefficients.adalasso
adlasso


# This part is not quite right, trying to plot the residuals
#x_test <- data.matrix(data_test[,1:6])
#z_test <- data.matrix(data_test[,8:9])
#x_test <- cbind(x_test, z_test)

#preds <- predict(elnet1, newx=x_test)
#test1 <- data_test
#test1
#test1$preds <- preds
#test1$residuals <- test1$preds-test1$PINCP

#qplot(log(test1$PINCP), test1$residuals)


# Q4
x_q4 <- data.matrix(data_train[,1:5])
z_q4 <- data.matrix(data_train[,7:9])
x_q4 <- cbind(x_q4, z_q4)
y_q4 <- as.matrix(data_train[,6])

## issue: factors are suspected to treat as numerical???
elnet_q4 <- glmnet(x_q4, y_q4, alpha = 0.5)
elnet.cv_q4 <- cv.glmnet(x_q4, y_q4, alpha = 0.5)
# Finding the Elastic net estimates for lambda with minimum cross-validation error
elnet1_q4 <- glmnet(x_q4, y_q4, lambda = elnet.cv$lambda.min, alpha = 0.5)
# Elastic net estimates of coefficients
elnet1_q4$beta

plotres(elnet.cv_q4)

##### Adaptive Lasso Estimates
adlasso_q4 <- adalasso(x, y)
adlasso_q4$coefficients.adalasso
adlasso_q4
```


#Appendix A - Source Code
```{r code=readLines(knitr::purl('./Project1_report_Code.Rmd', documentation = 0)), eval = FALSE,
echo=TRUE}

```