Computational Assigment #1

Brandon Moretz

### 1.) Given the variables in this dataset, which variables can be considered explanatory (X) and which considered response (Y)? Can any variables take on both roles? What is the population of interest for this problem (yes – this is a trick question!)?

Explanatory variables:

* High School
* Insured
* College
* Smokers
* Obese
* Heavy Drink

Response variables:

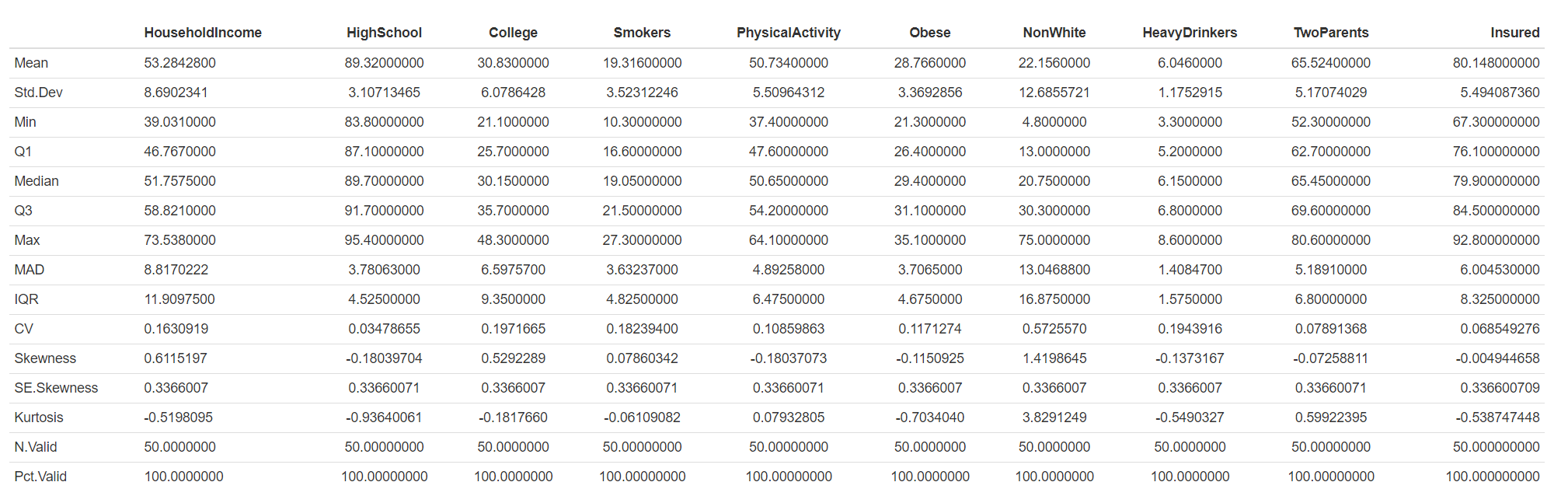
Both:

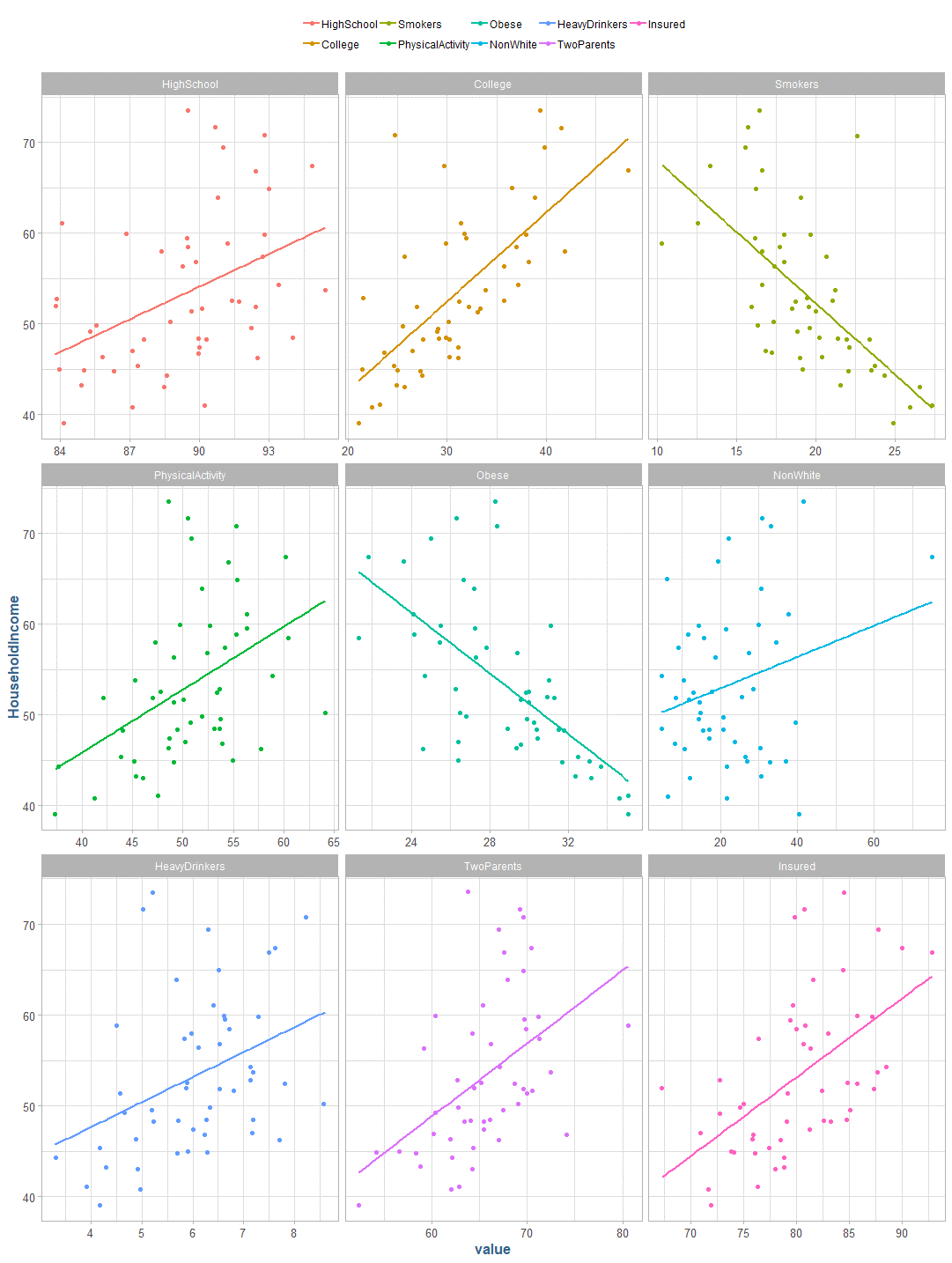
* Two Parents
* Heavy Drinkers
* Household Income
* Physical Activity

This is a census, because the data is summary data for each state. The population is the set of states.

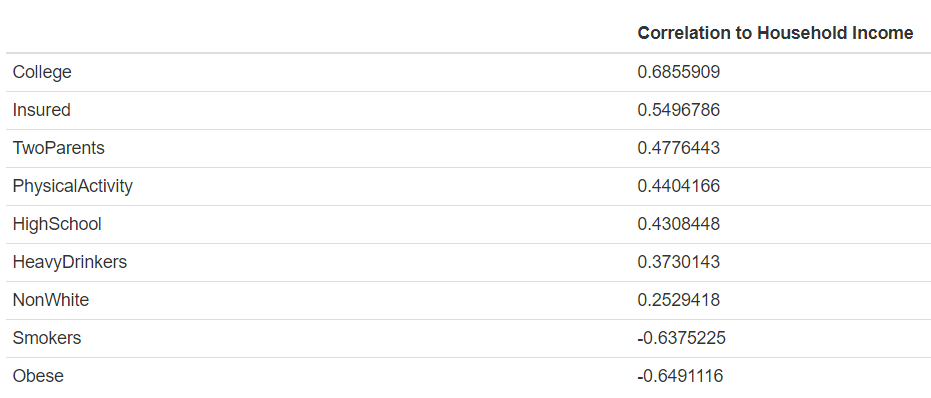
Most all the variables can be considered both explanatory and response variables. The state name is just an indicator, and should not be used in analysis, except to identify records. Region is a demographic variable of the population and would be explanatory.

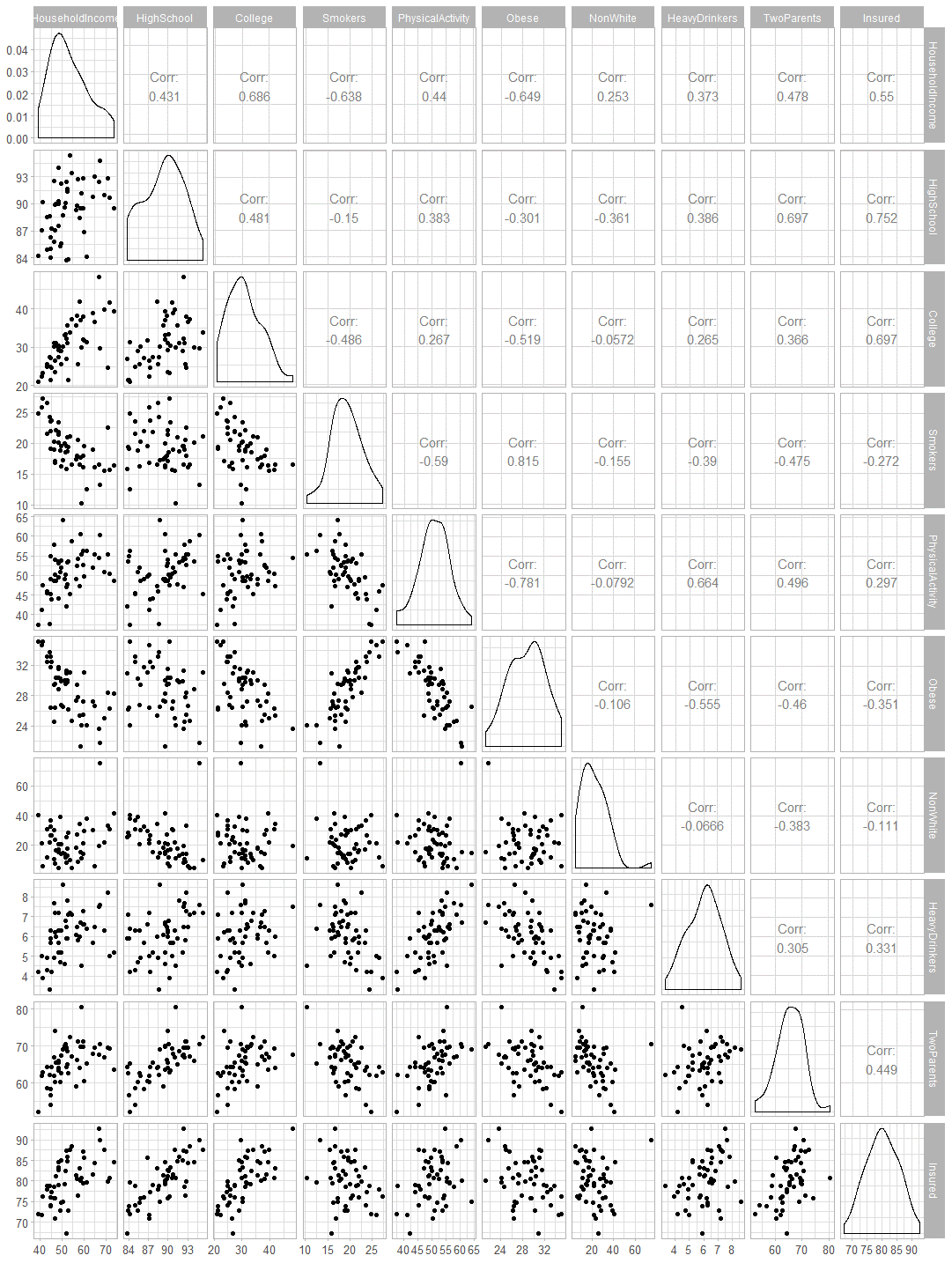
### 2.) For the duration of this assignment, let’s have HOUSEHOLDINCOME be the response variable (Y). Also, please consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all possible scatterplots relating the non-demographic explanatory variables to the response variable (Y).





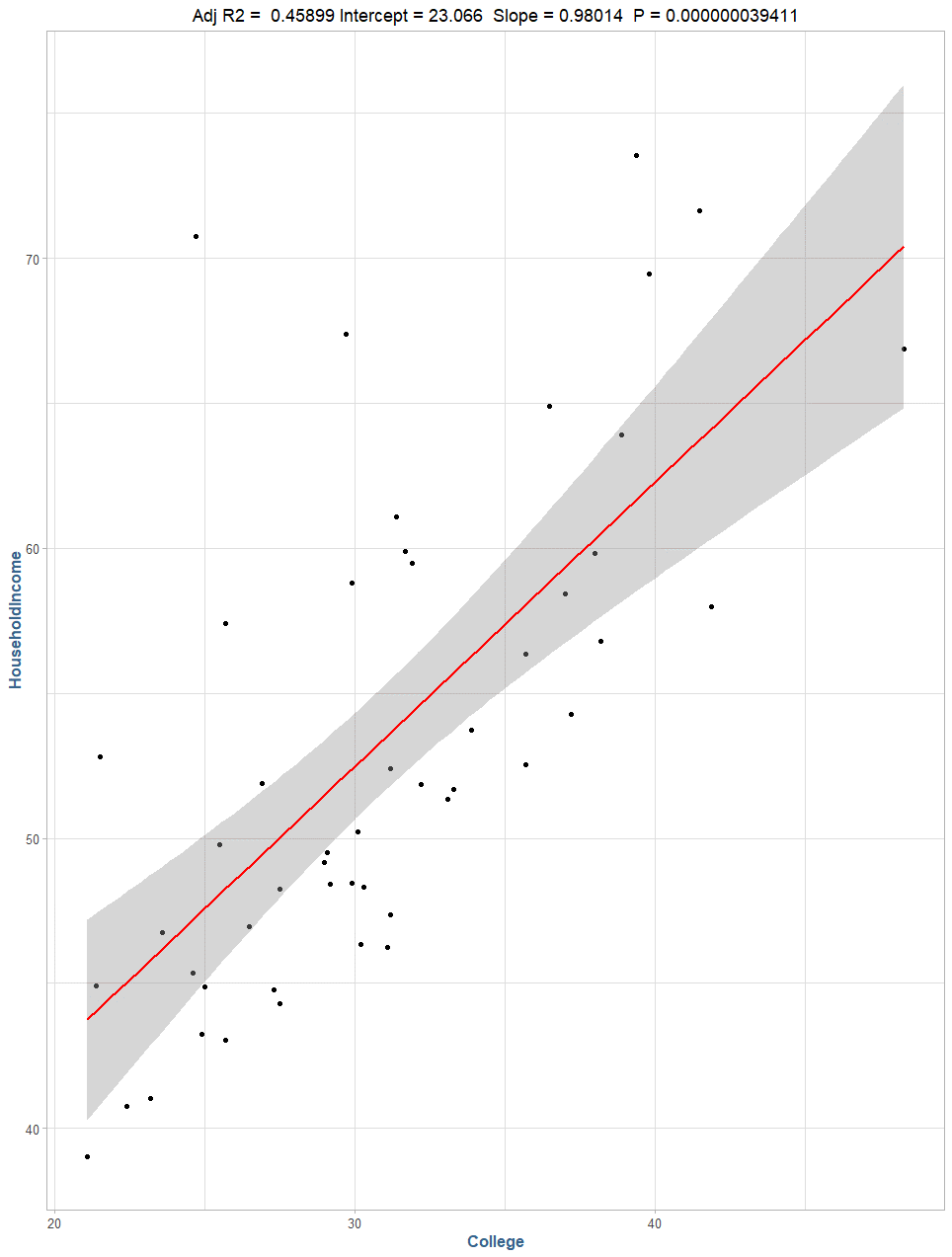
### 3.) Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?





Based upon the correlation to household income, there appears to be four variables which we could fit a linear model to with some success. These variables college, insured, smokers and obese have a semi-colinear relationship to the household income response.

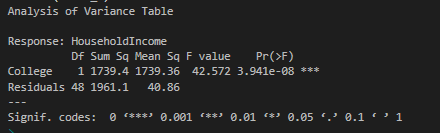
### 4.) Fit a simple linear regression model to predict Y using the COLLEGE explanatory variable. Use the base STAT lm(Y~X) function. Why would you want to start with this explanatory variable? Call this Model 1. Report the results of Model 1 in equation form and interpret each coefficient of the model in the context of this problem. Report the ANOVA table and model fit statistic, R-squared.

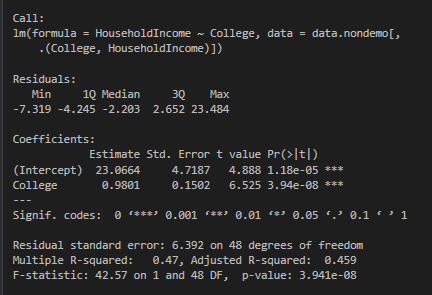


Here, we can see a simple linear model fitted to the college explanatory variable for the household income response variable. We start with the college variable as it has the highest colinearly relationship to the target response variable, household income.

 ŷ = 23.066 + 0.98X1,

where X1 is the percent of resident’s report to have a college education per state.





We can verify the slope and y-intercept by the following long-hand calculations:

slope <- cor(m1$College, m1$HouseholdIncome) \* (sd(m1$HouseholdIncome) / sd(m1$College))

**Output**: 0.9801

intercept <- mean(m1$HouseholdIncome) - (slope \* mean(m1$College))

**Output:** 23.0664

### 5.) Write R-code to calculate and create a variable of predicted values based on Model 1. Use the predicted values and the original response variable Y to calculate and create a variable of residuals (i.e. residual = Y – Y\_hat = observed minus predicted) for Model 1. Using the original Y variable, the predicted, and/or residual variables, write R-code to:

### Square each of the residuals and then add them up. This is called sum of squared residuals, or sums of squared errors.

m1$Y\_Hat <- predict(model\_1)

m1$residual <- m1$HouseholdIncome - m1$Y\_Hat

sum(m1$residual \*\* 2)

**Output**: 1961.13

### Deviate the mean of the Y’s from the value of Y for each record (i.e. Y – Y\_bar). Square each of the deviations and then add them up. This is called sum of squares total.

y\_bar <- mean(m1$HouseholdIncome)

sum((m1$HouseholdIncome - y\_bar) \*\* 2)

**Output**: 3700.488

### Deviate the mean of the Y’s from the value of predicted (Y\_hat) for each record (i.e. Y\_hat – Y\_bar). Square each of these deviations and then add them up. This is called the sum of squares due to regression.

sum((m1$Y\_Hat - y\_bar) \*\* 2)

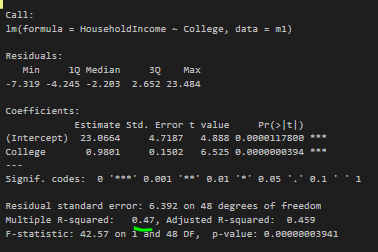
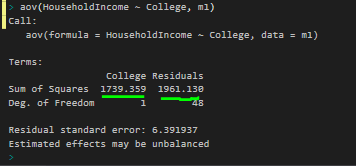
**Output**: 1739.359

### Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total)

(ssr / sst)

**Output**: 0.4700

### Verify and note the accuracy of the ANOVA table and R-squared values from the regression printout from part 4), relative to your computations here.

### 6.) Fit a multiple linear regression model to predict Y using COLLEGE and INSURED as the explanatory variables. Use the base lm(Y~X) function. Call this Model 2. Report the results of Model 2 in equation form, interpret each coefficient of the model in the context of this problem, and report the model fit statistic, R-squared. How have the coefficients and their interpretations changed? Calculate the change in R-squared from Model 1 to Model 2 and interpret this value. For this specific problem, is it OK to use the hypothesis testing results to determine if the additional explanatory variable should be retained or not? Think statistically using first principals. Discuss. NOTE: The topic of hypothesis testing in regression is the focus of Module 2 – you should NOT need to read anything about hypothesis testing to answer this.

Model 2:

9.6725 + 0.8411X1 + 0.2206X2

Where,

Y-Intercept: 9.6728

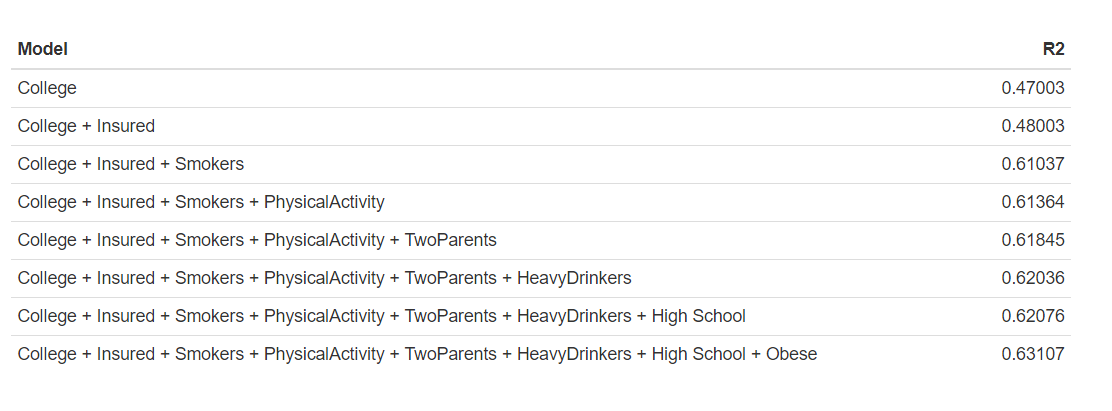
X1: Percent of respondents with a college degree

X2: Percent of respondents that have insurance

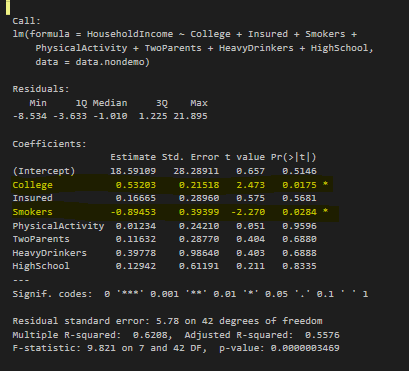
R2: 0.48

When both X1 and X2 are equal to zero, or the percentage of the population with no college degree and having insurance is zero, the average household income is $9,673. For both variables, 0 is outside the rage of the data, so the constant term is more of a place holder than an interpretable value. For a 1 unit change in College percentage, the average household income increases by $841, all other things held constant. Similarly, a 1 unit change increase in the percentage of the population with insurance increases household income by $220.60.

### 7.) In a sequential fashion, continue to add in the non-demographic variables into the prediction model, one variable at a time. Make a table summarizing the change in R-squared that is associated with each variable added. Based on this information, what variables should be retained for a “best” predictive model? What criteria seems appropriate to you?



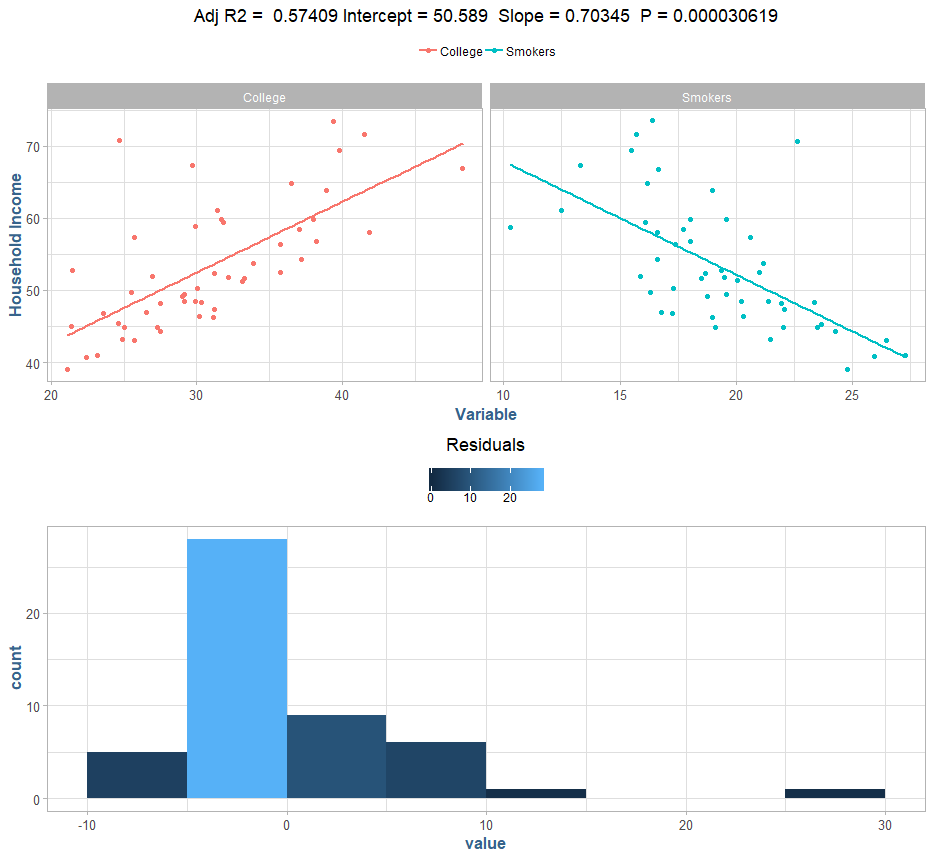
Based upon the information above relating to the R2, Smokers appears to explain the most variance in the data after College. Looking closer at the final model with all the variables, we can see that both the variables “College” and ”Smokers” have p-values at the 0.01 significance level:



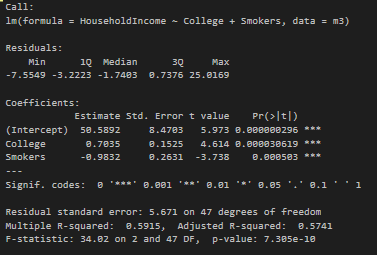
However, they appear to be somewhat colinear with a correlation value of -0.49.

### 8.) Now that you have a sense of which explanatory variables contribute to explaining HOUSEHOLDINCOME, refit a model using only the set of variables you consider to be appropriate to model Y. Report this model, interpret the coefficients, and interpret R-squared in the context of this problem. Discuss why is it necessary to refit this model.

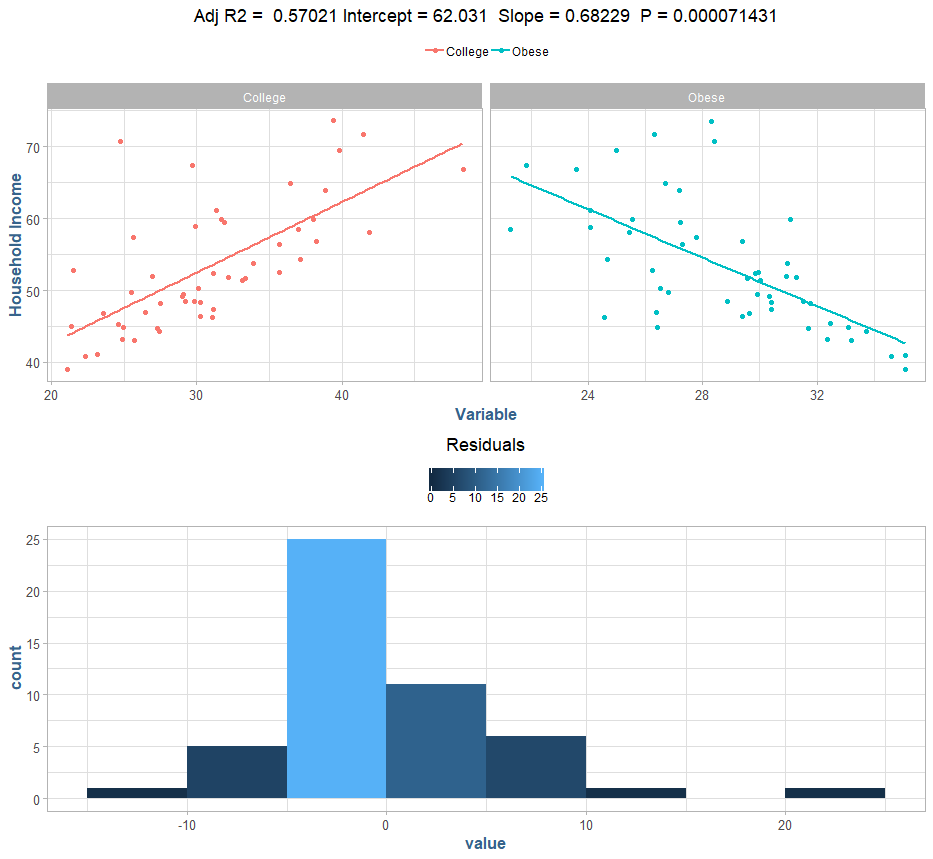
Refitting a multiple linear regression model on the variables “College” and “Smokers”, we can see the strong positive association between household income and residents with a college education and a strong negative association with the number of residents who smoke:



The R2of the new model is 0.57, with both variables having significance at the 0.001 level.

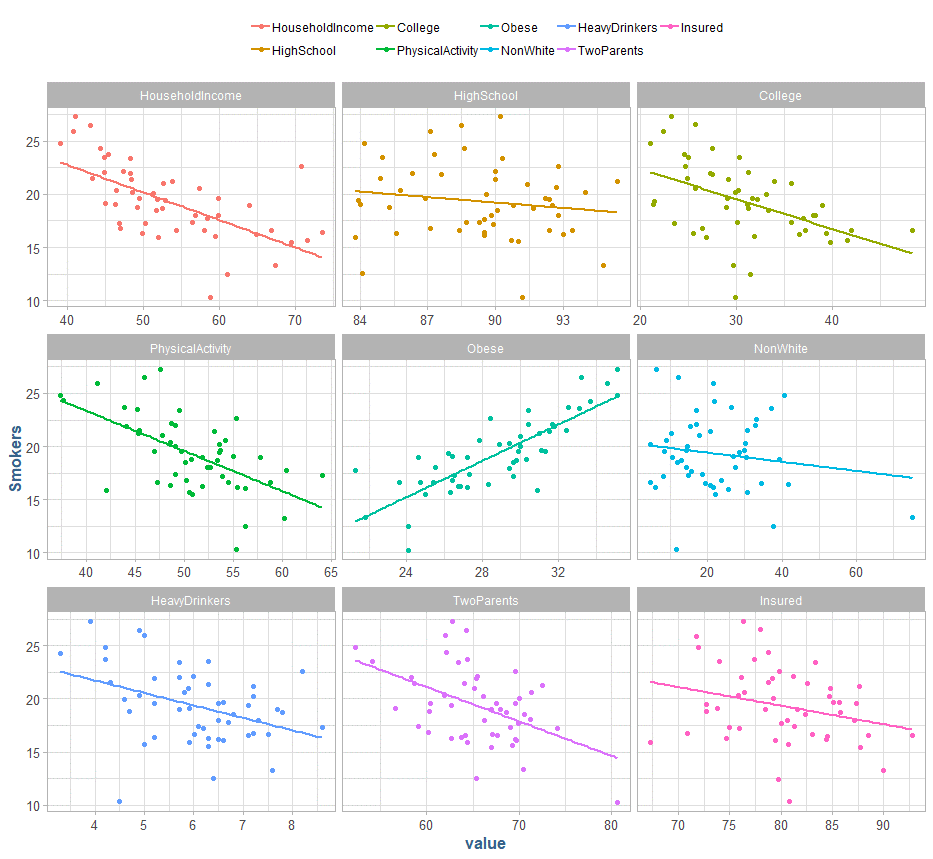


In the correlation matrix above, we noted that “Obese” had the strongest negative correlation to the target response, household income. However, a multiple linear regression model using obese over smoking yielded a slightly weaker R2 value:

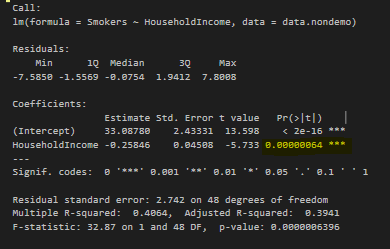


### 9.) You are welcome to conduct any other analyses you wish to embellish your understanding of this dataset.

I find it interesting that the “Smokers” variable has such an influence on the household income. Let’s look at smokers as a response to the other variables.



Physical activity is negatively correlated and obese is positively correlated as we might intuitively expect, and we can clearly see the strong association to household income as explored above. There appears to be a genuine association between smokers and household income as opposed to it being a confounding variable given its p-value significance.



### 10.) Given what you’ve learned from this modeling endeavor, what overall conclusions do you draw? What is the “Story” contained in this data? What have you learned? What are your Prescriptive Recommendations for action based on this evidence? Finally, feel free to reflect on what you’ve learned from a modeling perspective.

In this lab I learned how to interpret census data, and about the relationships around various non-demographic variables have on the expected household income in a given state. We looked at multiple non-demographic variables as possible explanatory variables for household income, starting with the variables that have the highest degree of collinearity, namely college educated, insurance, smokers and obese.

We formed a simple linear regression model using the strongest positive relationship, college, which lead to an overall weakly explanatory model of the data. The R2 value is not the only indicator of “goodness of fit”, however, it is a strong indicator of how well the model explains the overall variance in the data. In this instance, the resulting R2 was examined using the summary of the linear model, as well as calculated “long-hand” by looking at the sum of squared due to regression ((y - ŷ)2) as a proportion of the total sum of squares (total variation in the data, (y - ȳ )2). This indicator leads us to a relatively weak explanation of the household income due to the residents who attended college.

The simple linear model we developed above was not indicative enough of the response variable alone, so we chose to expand our analysis using multiple linear regression. We built several combinations of the multiple linear regression model adding explanatory variables in a serial fashion and noting the R2, or portion of explained variance, with each iteration. We noted that with every increased variable in the model, the resulting R2 value did in fact increase, most of these values did not display statistical significance to the explanation of the variance in the data. In our final model, we noted two variables had statistical significance in their p-values at the 0.01 level, college and smoking. We also note here that simply because there is p-value significance, it does not mean that the variable is a definitive explanatory variable for household income.

We explored this relationship in further detail, noting that there is a weak colinear relationship between college and smoking as to provide some evidence that smoking is not simply a colinear explanatory variable with college. We adjusted our simple model to include smoking, and we noted the increase of our R2 to 0.57, which is better than our original, however ultimately not ideal for building a predictive model. Having examined all the variables in the available data, we should either seek more data points to explain the household income variance or look to build a non-linear model to seek greater accuracy in our predictability.