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:

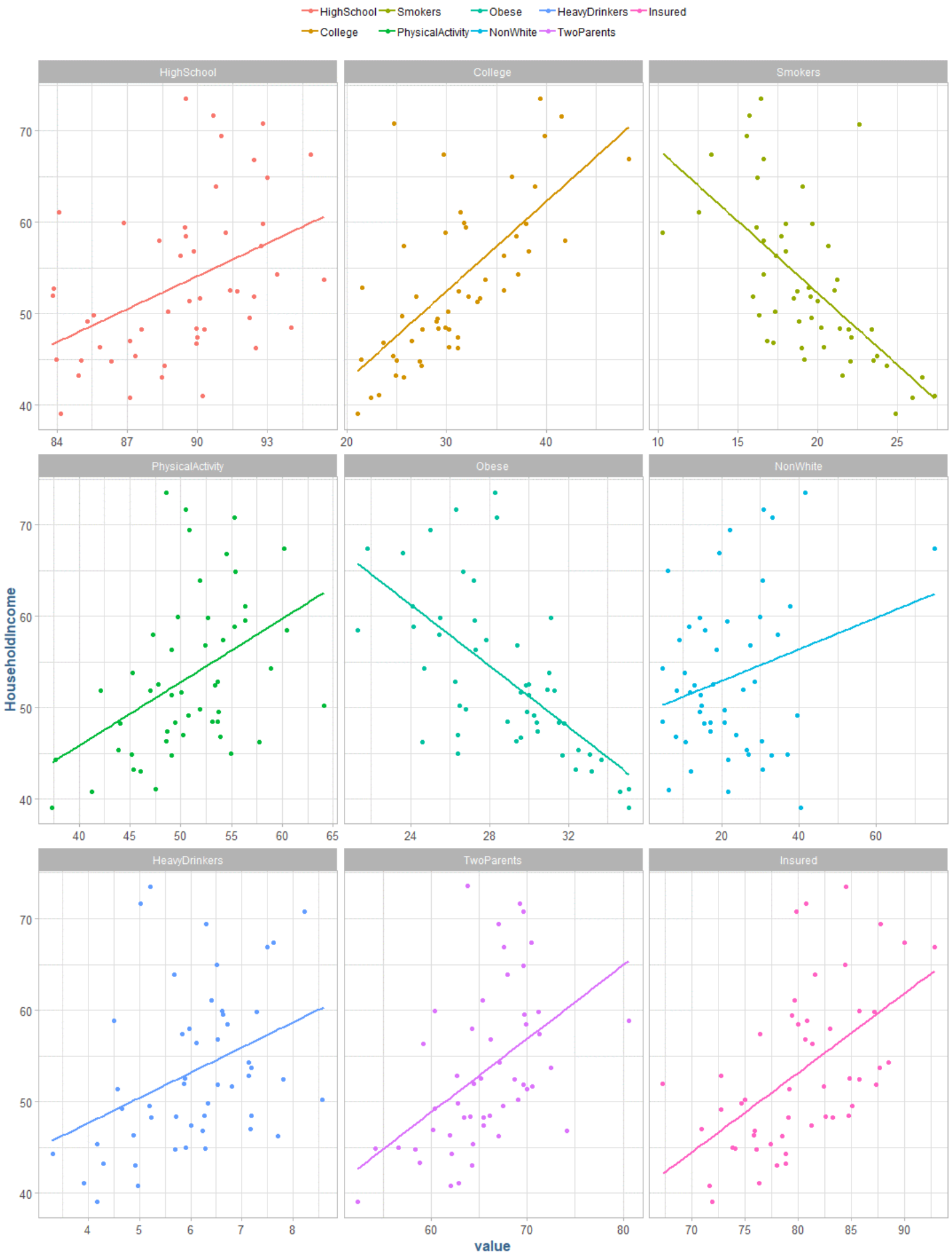
Both:

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

The population of interest for this dataset is the US population.

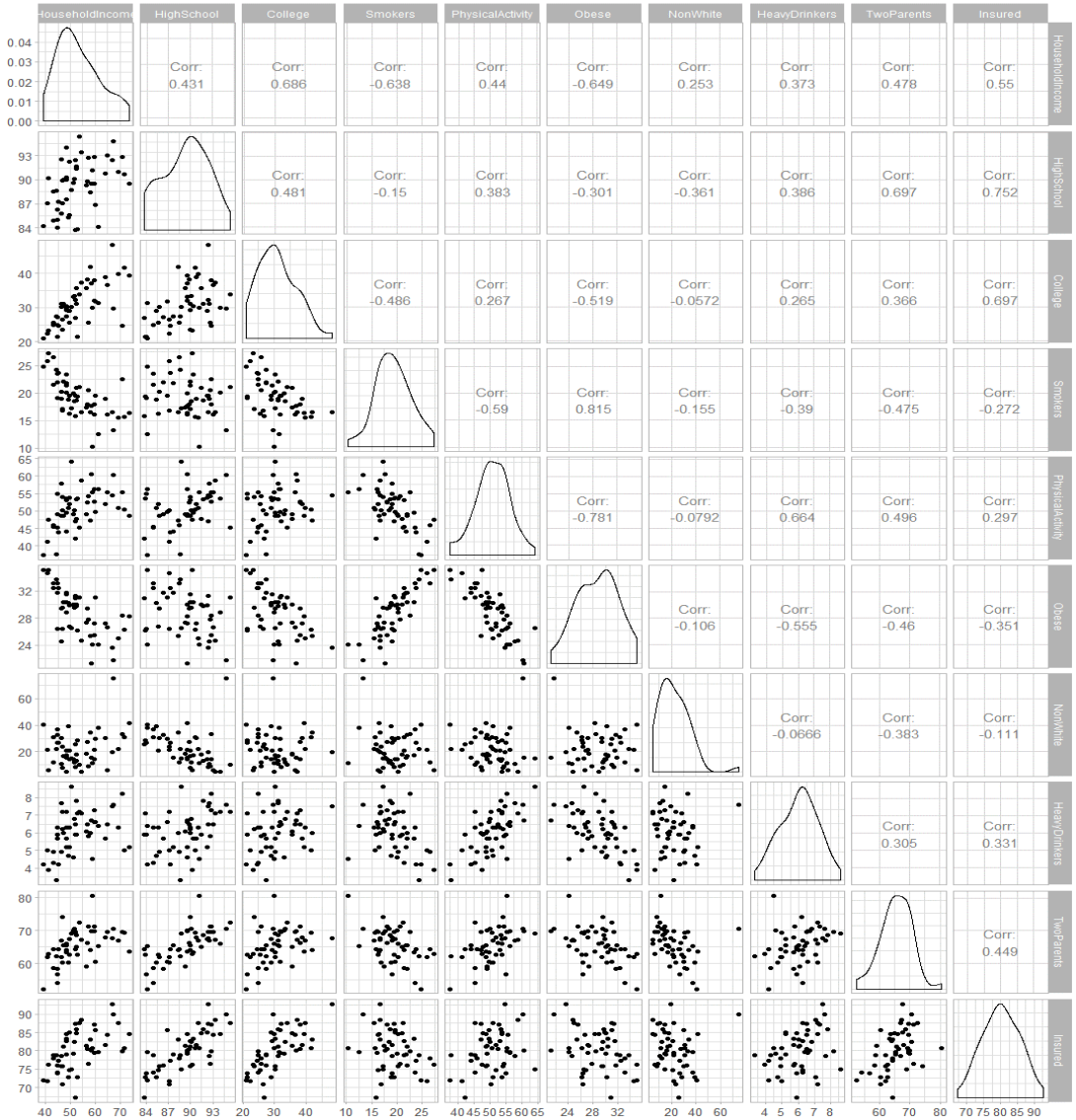
2.) For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, please consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain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

	HouseholdIncome	HighSchool	College	Smokers	PhysicalActivity	Obese	NonWhite	HeavyDrinkers	TwoParents	Insured
Mean	53.2842800	89.32000000	30.8300000	19.31600000	50.73400000	28.7660000	22.1560000	6.0460000	65.52400000	80.148000000
Std.Dev	8.6902341	3.10713465	6.0786428	3.52312246	5.50964312	3.3692856	12.6855721	1.1752915	5.17074029	5.494087360
Min	39.0310000	83.80000000	21.1000000	10.30000000	37.40000000	21.3000000	4.8000000	3.3000000	52.30000000	67.300000000
Q1	46.7670000	87.10000000	25.7000000	16.60000000	47.60000000	26.4000000	13.0000000	5.2000000	62.70000000	76.100000000
Median	51.7575000	89.70000000	30.1500000	19.05000000	50.65000000	29.4000000	20.7500000	6.1500000	65.45000000	79.900000000
Q3	58.8210000	91.70000000	35.7000000	21.50000000	54.20000000	31.1000000	30.3000000	6.8000000	69.60000000	84.500000000
Max	73.5380000	95.40000000	48.3000000	27.30000000	64.10000000	35.1000000	75.0000000	8.6000000	80.60000000	92.800000000
MAD	8.8170222	3.78063000	6.5975700	3.63237000	4.89258000	3.7065000	13.0468800	1.4084700	5.18910000	6.004530000
IQR	11.9097500	4.52500000	9.3500000	4.82500000	6.47500000	4.6750000	16.8750000	1.5750000	6.80000000	8.325000000
CV	0.1630919	0.03478655	0.1971665	0.18239400	0.10859863	0.1171274	0.5725570	0.1943916	0.07891368	0.068549276
Skewness	0.6115197	-0.18039704	0.5292289	0.07860342	-0.18037073	-0.1150925	1.4198645	-0.1373167	-0.07258811	-0.004944658
SE.Skewness	0.3366007	0.33660071	0.3366007	0.33660071	0.33660071	0.3366007	0.3366007	0.3366007	0.33660071	0.336600709
Kurtosis	-0.5198095	-0.93640061	-0.1817660	-0.06109082	0.07932805	-0.7034040	3.8291249	-0.5490327	0.59922395	-0.538747448
N.Valid	50.0000000	50.00000000	50.0000000	50.00000000	50.00000000	50.0000000	50.0000000	50.0000000	50.00000000	50.000000000
Pct.Valid	100.0000000	100.00000000	100.0000000	100.00000000	100.00000000	100.0000000	100.0000000	100.0000000	100.00000000	100.000000000


HighSchool
Smokers
Obese
HeavyDrinkers
Insured
College
PhysicalActivity
NonWhite
TwoParents
HighSchool
College
Smokers
PhysicalActivity
Obese
NonWhite
HeavyDrinkers
TwoParents
Insured
HouseholdIncome
value

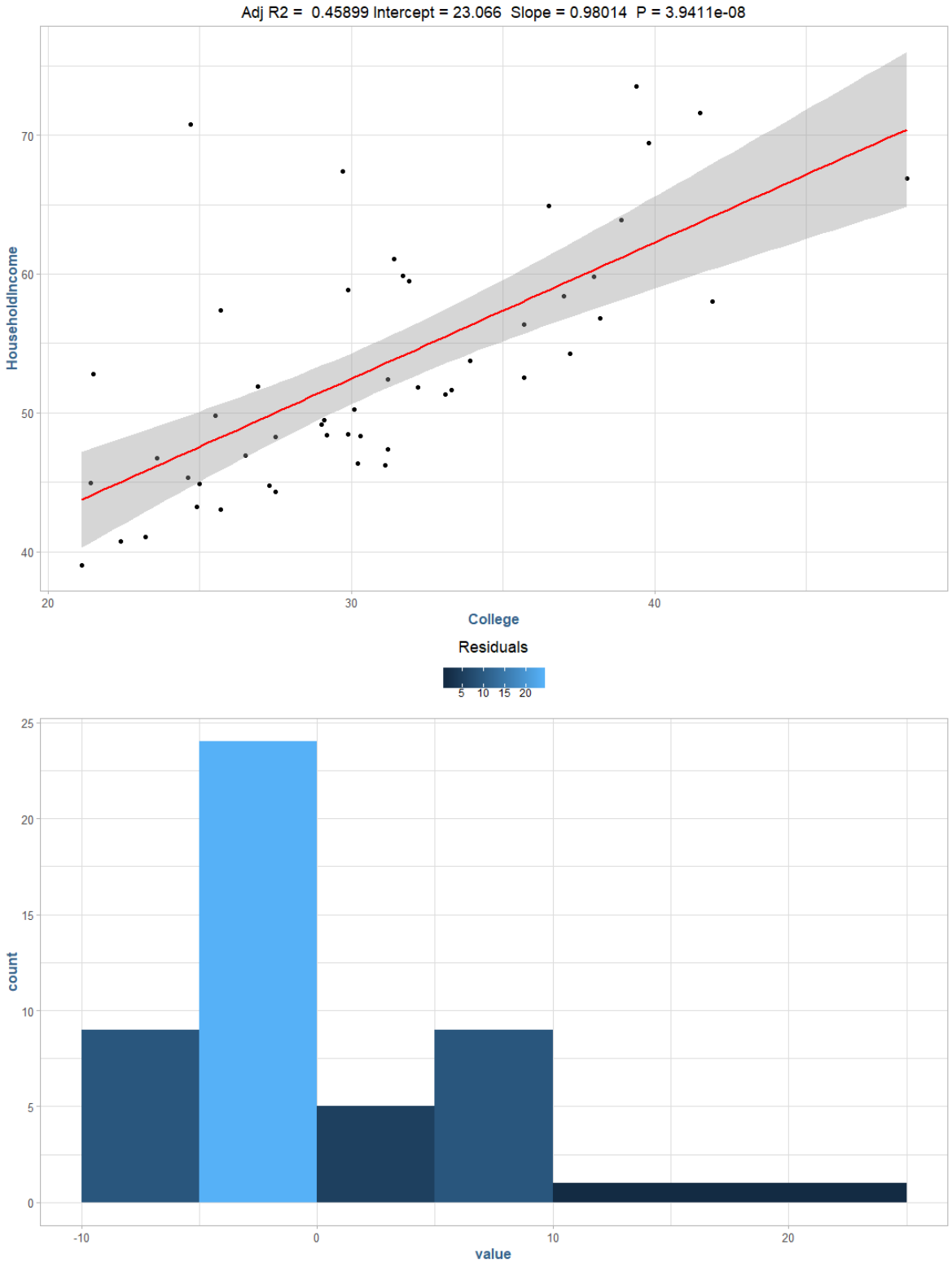
3.) Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or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

	Correlation to Household Income
College	0.6855909
Insured	0.5496786
TwoParents	0.4776443
PhysicalActivity	0.4404166
HighSchool	0.4308448
HeavyDrinkers	0.3730143
NonWhite	0.2529418
Smokers	-0.6375225
Obese	-0.6491116



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.

$\hat{y} = 23.066 + 0.98X_1$,

where X_1 is the percent of resident's report to have a college education per state.

```
Analysis of Variance Table

Response: HouseholdIncome
          Df Sum Sq Mean Sq F value    Pr(>F)
College    1 1739.4 1739.36  42.572 3.941e-08 ***
Residuals 48 1961.1   40.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = HouseholdIncome ~ College, data = data.nondemo[,
    .(College, HouseholdIncome)])

Residuals:
   Min     1Q Median     3Q    Max
-7.319 -4.245 -2.203  2.652 23.484

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.0664     4.7187   4.888 1.18e-05 ***
College       0.9801     0.1502   6.525 3.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.392 on 48 degrees of freedom
Multiple R-squared:   0.47, Adjusted R-squared:  0.459
F-statistic: 42.57 on 1 and 48 DF,  p-value: 3.941e-08
```

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: 0.9801441

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

Output: 23.06644

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: 0.4700349

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.

```
Call:
lm(formula = HouseholdIncome ~ College, data = m1)

Residuals:
   Min     1Q Median     3Q    Max
-7.319 -4.245 -2.203  2.652 23.484

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  23.0664     4.7187   4.888 0.0000117800 ***
College       0.9801     0.1502   6.525 0.0000000394 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.392 on 48 degrees of freedom
Multiple R-squared:   0.47, Adjusted R-squared:  0.459
F-statistic: 42.57 on 1 and 48 DF,  p-value: 0.00000003941
```

```
> aov(HouseholdIncome ~ College, m1)
Call:
   aov(formula = HouseholdIncome ~ College, data = m1)

Terms:
                College Residuals
Sum of Squares  1739.359  1961.130
Deg. of Freedom       1        48

Residual standard error: 6.391937
Estimated effects may be unbalanced
>
```

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:

$$9.6725 + 0.8411X_1 + 0.2206X_2$$

Where,

Y-Intercept: 9.6728

X_1: Percent of respondents with a college degree

X_2: Percent of respondents that have insurance

R^2: 0.48

In this multiple linear regression model, we can see that the additional feature of "Insured" does not contribute significantly to the overall variance explained by the simple linear model using only the "College" variable, having only a 0.01 total delta in R^2. Additionally, we can see that the p-value for insured is large a .3468 indicating that the null hypothesis, that this value adds no additional information to the model, cannot be rejected.

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?

Model	R2
College	0.47003
College + Insured	0.48003
College + Insured + Smokers	0.61037
College + Insured + Smokers + PhysicalActivity	0.61364
College + Insured + Smokers + PhysicalActivity + TwoParents	0.61845
College + Insured + Smokers + PhysicalActivity + TwoParents + HeavyDrinkers	0.62036
College + Insured + Smokers + PhysicalActivity + TwoParents + HeavyDrinkers + High School	0.62076
College + Insured + Smokers + PhysicalActivity + TwoParents + HeavyDrinkers + High School + Obese	0.63107

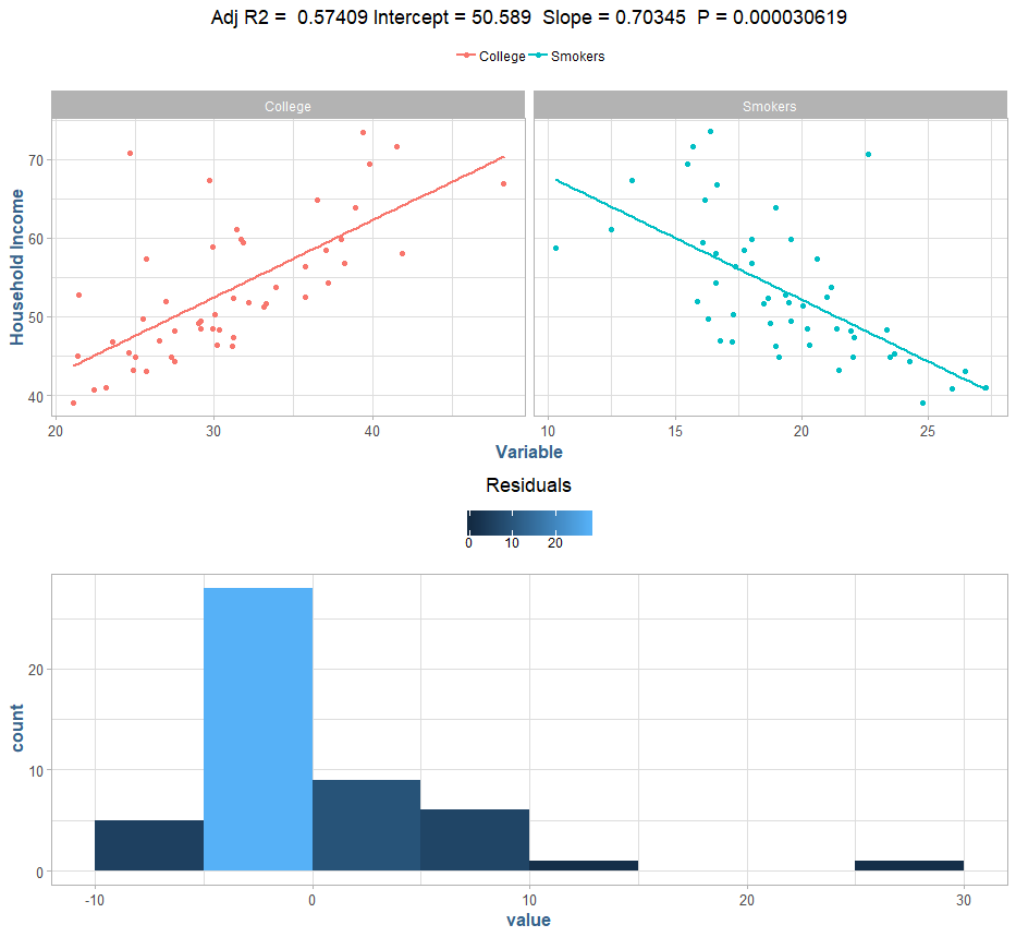
Based upon the information above relating to the R^2, Smokers appears to explain the most variance in the data after College. Looking closer at the final model with all the variables, we can see that both the variables "College" and "Smokers" have p-values at the 0.01 significance level:

```
Call:
lm(formula = HouseholdIncome ~ College + Insured + Smokers +
    PhysicalActivity + TwoParents + HeavyDrinkers + HighSchool,
    data = data.nondemo)

Residuals:
   Min     1Q Median     3Q    Max
-8.534 -3.633 -1.010  1.225 21.895

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     18.59109   28.28911   0.657   0.5146
College          0.53203    0.21518   2.473   0.0175 *
Insured          0.16665    0.28960   0.575   0.5681
Smokers         -0.89453    0.39399  -2.270   0.0284 *
PhysicalActivity  0.01234    0.24210   0.051   0.9596
TwoParents       0.11632    0.28770   0.404   0.6880
HeavyDrinkers    0.39778    0.98640   0.403   0.6888
HighSchool       0.12942    0.61191   0.211   0.8335
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.78 on 42 degrees of freedom
Multiple R-squared:  0.6208,  Adjusted R-squared:  0.5576
F-statistic: 9.821 on 7 and 42 DF,  p-value: 0.0000003469
```

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.

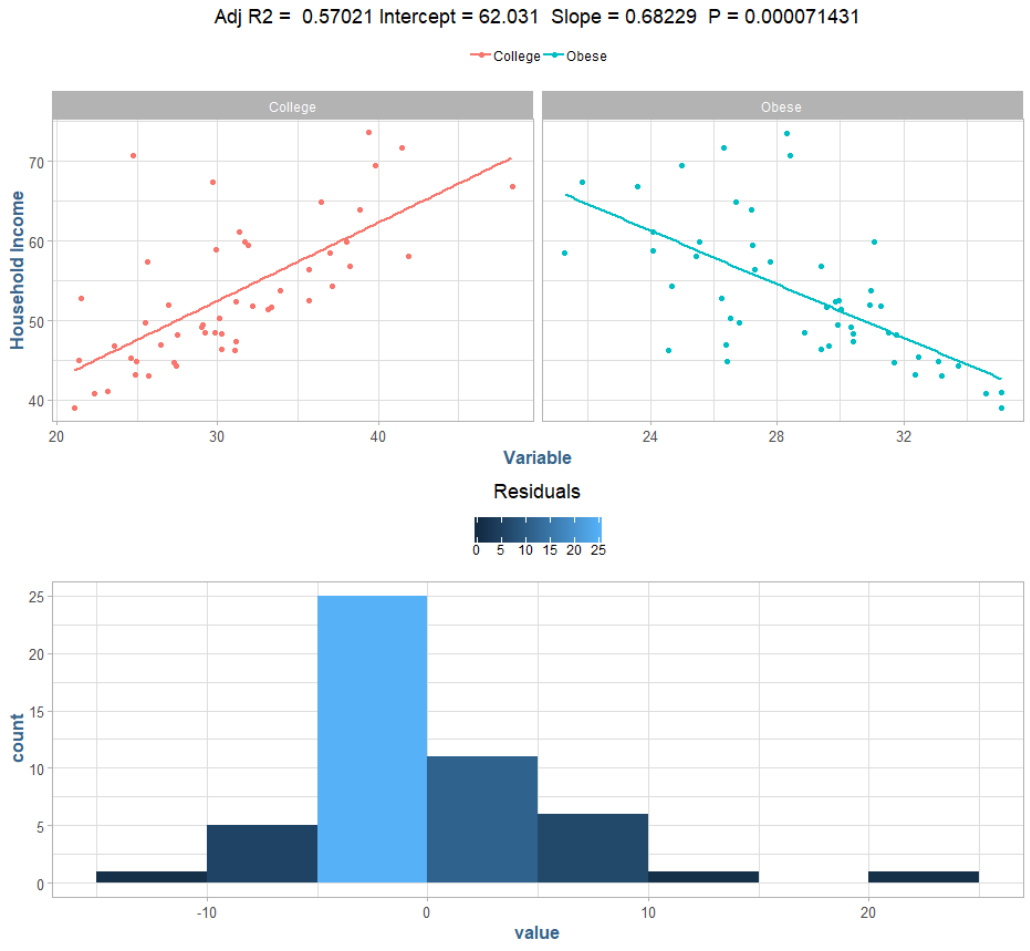
Refitting a multiple linear regression model on the variables "College" and "Smokers", we can see the strong positive association between household income and residents with a college education and a strong negative association with the number of residents who smoke:



The R^2 of the new model is 0.57, with both variables having significance at the 0.001 level.

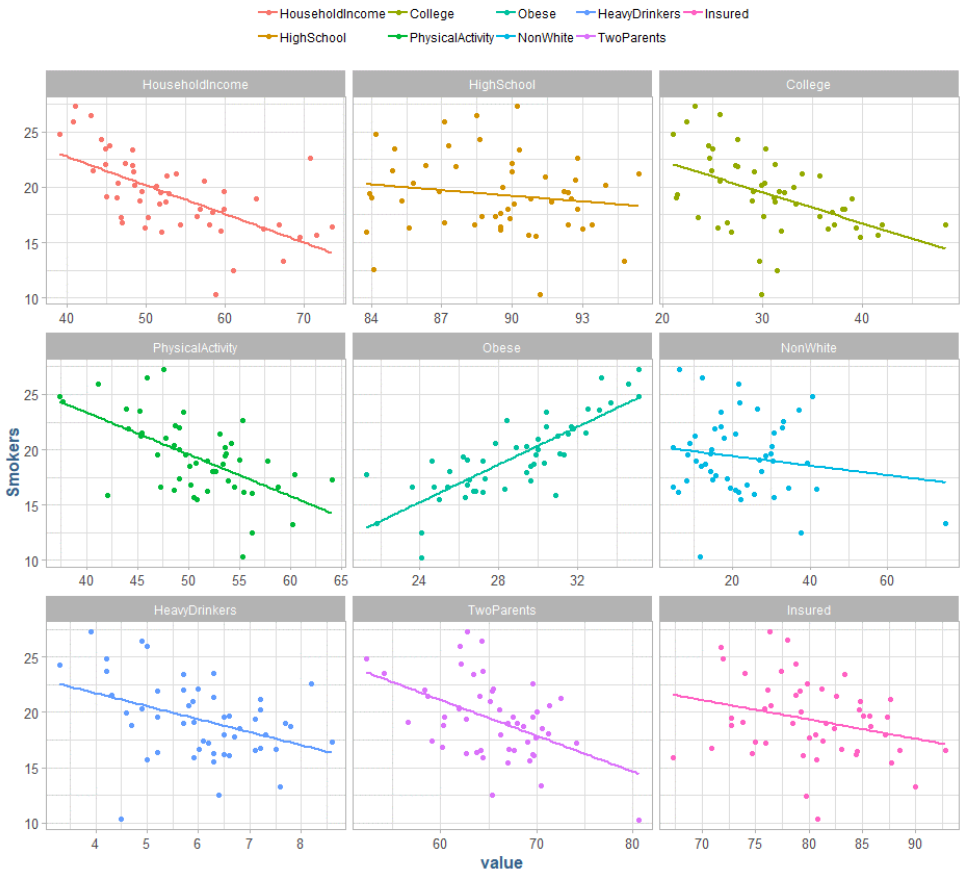
```
Call:
lm(formula = HouseholdIncome ~ College + Smokers, data = m3)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5549 -3.2223 -1.7403  0.7376 25.0169

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)  50.5892     8.4703   5.973 0.000000296 ***
College       0.7035     0.1525   4.614 0.000030619 ***
Smokers      -0.9832     0.2631  -3.738    0.000503 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.671 on 47 degrees of freedom
Multiple R-squared:  0.5915,  Adjusted R-squared:  0.5741
F-statistic: 34.02 on 2 and 47 DF,  p-value: 7.305e-10
```

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 R^2 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 apposed to it being a confounding variable given its p-value significance.

```
Call:
lm(formula = Smokers ~ HouseholdIncome, data = data.nondemo)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5850 -1.5569 -0.0754  1.9412  7.8008

Coefficients:
               Estimate Std. Error t value   Pr(>|t|)
(Intercept)     33.08780    2.43331  13.598    < 2e-16 ***
HouseholdIncome -0.25846    0.04508  -5.733 0.00000064 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.742 on 48 degrees of freedom
Multiple R-squared:  0.4064,  Adjusted R-squared:  0.3941
F-statistic: 32.87 on 1 and 48 DF,  p-value: 0.0000006396
```

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 R^2 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 R^2 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 $((y - \hat{y})^2)$ as a proportion of the total sum of squares (total variation in the data, $(y - \bar{y})^2$). 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 R^2, or portion of explained variance, with each iteration. We noted that with every increased variable in the model, the resulting R^2 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 R^2 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.