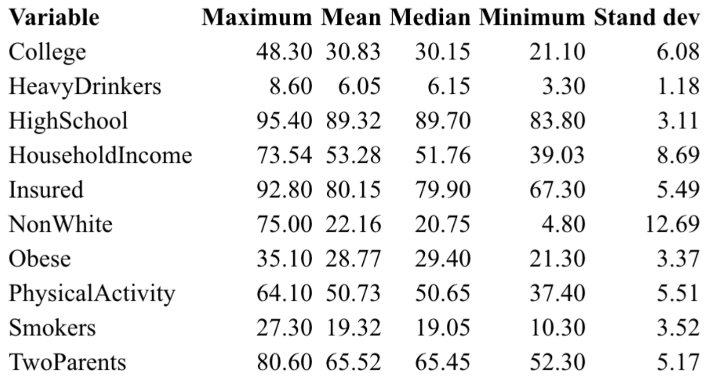
Michael Venit

MSDS 410

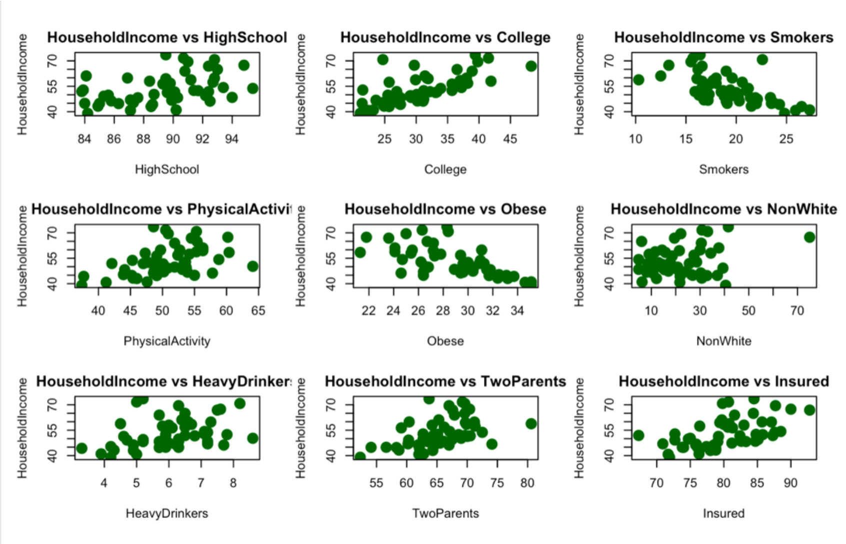
Computational Assignment #1

1. After reading the US States data for this assignment, I was able to conclude that household income is a response variable that one would choose to predict in this dataset. Other features in the data would seem to serve as explanatory variables for household income. The population of interest as it pertains to the purpose of the assignment would be considered household income even though a case can be made for the state variable.

2. I decided to remove state, region and population as features that would not contribute to the prediction of the response variable since these are demographic fields. Hence, 10 features were left to explore as can be seen by the summary statistics below.

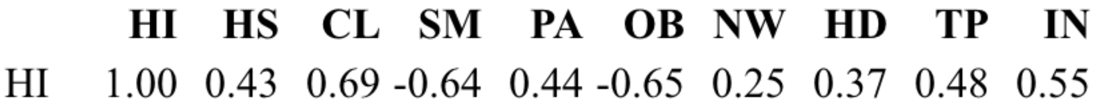


Variable	Maximum	Mean	Median	Minimum	Stand dev
College	48.30	30.83	30.15	21.10	6.08
HeavyDrinkers	8.60	6.05	6.15	3.30	1.18
HighSchool	95.40	89.32	89.70	83.80	3.11
HouseholdIncome	73.54	53.28	51.76	39.03	8.69
Insured	92.80	80.15	79.90	67.30	5.49
NonWhite	75.00	22.16	20.75	4.80	12.69
Obese	35.10	28.77	29.40	21.30	3.37
PhysicalActivity	64.10	50.73	50.65	37.40	5.51
Smokers	27.30	19.32	19.05	10.30	3.52
TwoParents	80.60	65.52	65.45	52.30	5.17



Each of the 9 continuous explanatory variables were plotted against the response variable (HouseholdIncome) as a first step in determining if there was any noticeable relationship. Several of the explanatory variables have a linear relationship with HouseholdIncome. Smokers seems to be negatively correlated with HouseholdIncome while College has a strong positive linear correlation.

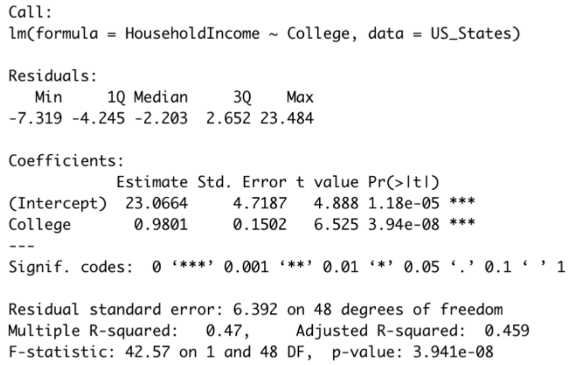
3. The next step in the process of exploring these non-demographic features is to determine their respective correlation to household income via Pearson Product Moment correlation. The table below gives the correlation between our response variable and explanatory variables.



	HI	HS	CL	SM	PA	OB	NW	HD	TP	IN
HI	1.00	0.43	0.69	-0.64	0.44	-0.65	0.25	0.37	0.48	0.55

After creating our scatterplots and determining our correlation coefficients, no single variable has a particularly strong relationship with household income. The variable with the strongest positive correlation coefficient of 0.69 is college, while the feature with the strongest negative relationship is that of the obese variable. From the table, the use of the college, insured and smokers variables can be used for multiple linear regression to predict household income.

4. I first used college, as the explanatory variable, to create a simple linear regression model to predict the response variable. This feature appears to have the strongest positive correlation with household income, while also showing a strong, positive linear relationship as can be seen in the scatterplot. Below, the summary statistics for Model 1 can be found as well as the linear equation for prediction. The value of 23.0664 represents the Y-intercept, or the value when X = 0. Hence, household income would be 23.0664 when the college variable is 0. The intercept coefficient also has a standard error of 4.7187 with a corresponding t-value of 4.888. The t-value allows for the rejection of the null hypothesis that our intercept is equal to 0. On the other hand, the regression coefficient for the college variable is 0.9801, meaning that for every 1 unit increase in the value of college, we can expect household income to also increase by an increment of 0.9801. The ANOVA table shows an F-value of 42.572 that allows us to reject the null hypothesis while the r-squared value of 0.47 demonstrates that approximately 47% of the variance in household income can be explained with college.

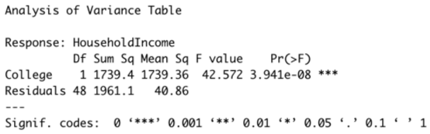


```
Call:
lm(formula = HouseholdIncome ~ College, data = US_States)

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
```



```
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
```

$$\hat{Y} = 23.0664 + 0.9801X$$

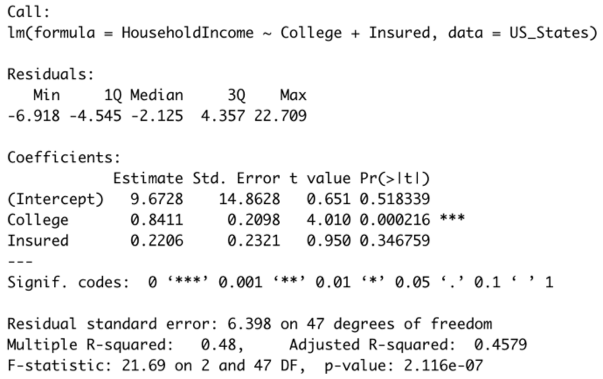
5. The following statistics are required items for this particular step

- Sum of Squared Residuals: 1961.1
- Sum of Squares Total: 3700.5
- Sum of Squares due to Regression: 1739.36

- SSR/SST: 0.47003

It appears that the SSR/SST value matches up with the multiple R-squared value 0.47. Manual computations are aligned with the statistics generated from the program. This means that approximately 47% of the variance in the variable household income is explained with the feature college.

6. A multiple linear regression model was crafted using college and insured as the explanatory variables, while household income remains the response variable. The intercept coefficient is now 9.6728, meaning that if the value of all explanatory variables is equal to 0, then the value of household income would be the value of the intercept coefficient. The college regression coefficient is equal to 0.8411 which indicates that a 1 unit increase in the college proportion will result in a 0.8411 increase in average household income when all other variables are held constant. The coefficient for the college variable has decreased, but it is still considered statistically significant to the model with a p-value of 0.000216. The insured coefficient is 0.2206. This means that for a 1 unit increase in the proportion of people insured, the average household income will increase 0.2206. The r-squared value for Model 2 is 0.48, which is an increase of 0.01 from Model 1 (0.47). This means that by adding the insured variable, our model was only able to account for an additional 1% of the variance in household income. Seeing as how adding variables inherently increases the r-squared value, the addition of insured as an explanatory variable has not produced anything meaningful. The standard error of the model is greater than the coefficient value, which indicates that the coefficient could in fact be of the opposite sign or zero. Due to the t-value of the coefficient, we cannot reject the null hypothesis that it is zero and the variable should not be used again in further models.

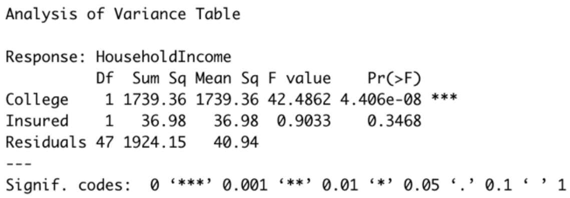


```
Call:
lm(formula = HouseholdIncome ~ College + Insured, data = US_States)

Residuals:
   Min     1Q Median     3Q    Max
-6.918 -4.545 -2.125  4.357 22.709

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6728    14.8628   0.651 0.518339
College       0.8411     0.2098   4.010 0.000216 ***
Insured       0.2206     0.2321   0.950 0.346759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.398 on 47 degrees of freedom
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4579
F-statistic: 21.69 on 2 and 47 DF,  p-value: 2.116e-07
```



```
Analysis of Variance Table

Response: HouseholdIncome
          Df  Sum Sq Mean Sq F value    Pr(>F)
College    1 1739.36 1739.36 42.4862 4.406e-08 ***
Insured    1   36.98   36.98  0.9033    0.3468
Residuals 47 1924.15   40.94
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

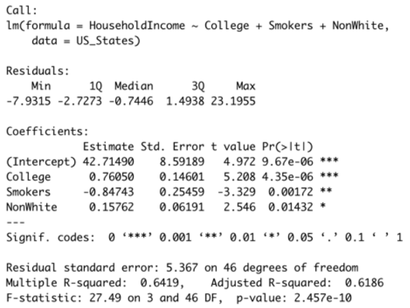
$$\hat{Y} = 9.6728 + 0.8411X_1 + 0.2206X_2$$

7. Following the creation and analysis of Model 2, the next task was to continue adding in non-demographic features into the prediction model, one variable at a time. A table was then created with each model and its generated r-squared value. We can see the effect that additional explanatory variables have on the r-squared value. There is a drastic increase in the explained variability of household income when incorporating smokers and non-white as explanatory variables. I believe that the variables college, smokers and non-white should be retained for the final model. Their additions to the model yielded the biggest increase in the r-squared value. Intuitively, common logic would state that people who have gone to college, on average, make more money than those who do not.

Variables	**R-Squared**
CL + IN + HS	0.4843539
CL + IN + HS + SM	0.6175388
CL + IN + HS + SM + PA	0.6183680
CL + IN + HS + SM + PA + OB	0.6279742
CL + IN + HS + SM + PA + OB + NW	0.7112100
CL + IN + HS + SM + PA + OB + NW + HD	0.7114523
CL + IN + HS + SM + PA + OB + NW + HD + TP	0.7354778

8. For the final model, college, smokers and non-white features were used as predictors for our response variable. The intercept coefficient is 42.71490, meaning that if all of the explanatory

variables were to equal 0, then the average household income is 42.71490. This coefficient has a corresponding t-value of 4.972 which allows us to reject the null hypothesis that it's zero.. The college coefficient equals 0.76050. As the proportion of the population who've attended college increases by 1%, the average household income will increase 0.76050. The smokers regression coefficient equals -0.84743 and reflects the proportion of the population who smoke. As this proportion increases 1%, the average household income will decrease by 0.84743 units. The non-white regression coefficient equals 0.15762. This was interesting as the proportion of minorities increases by 1%, household income will increase by 0.15762. The multiple r-squared value of this model was 0.6419 as the model has accounted for 64% of the variance in household income.

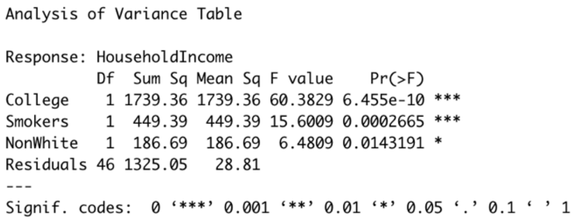


```
Call:
lm(formula = HouseholdIncome ~ College + Smokers + NonWhite,
    data = US_States)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9315 -2.7273 -0.7446  1.4938 23.1955 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.71490    8.59189   4.972 9.67e-06 ***
College      0.76050    0.14601   5.208 4.35e-06 ***
Smokers     -0.84743    0.25459  -3.329  0.00172 ** 
NonWhite     0.15762    0.06191   2.546  0.01432 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.367 on 46 degrees of freedom
Multiple R-squared:  0.6419,	Adjusted R-squared:  0.6186 
F-statistic: 27.49 on 3 and 46 DF,  p-value: 2.457e-10
```



```
Analysis of Variance Table

Response: HouseholdIncome
          Df  Sum Sq Mean Sq F value    Pr(>F)    
College    1 1739.36 1739.36 60.3829 6.455e-10 ***
Smokers    1  449.39  449.39 15.6009 0.0002665 ***
NonWhite   1  186.69  186.69  6.4809 0.0143191 *  
Residuals 46 1325.05   28.81                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\hat{Y} = 42.71490 + 0.76050X_1 - 0.84743X_2 + 0.15762X_3$$

9. Overall, there are some promising features in this dataset that can help in the understanding of the average household income in the United States. From the analysis performed, one can conclude that a college educated and healthy population has a higher average household income. Populations that show higher rates of unhealthy attributes, like smoking and obesity, tend to display lower average household income. I would recommend stronger education for those areas of the country that show less college educated individuals as well as a health program to discuss the impact of smoking on the health and well-being of individuals. If a population is generally unhealthy and has lower average household income, then there are typically higher associated healthcare costs. For this project, I feel like I was able to learn a lot about modeling. I havent

had much experience with modeling, let alone have I approached modeling from the perspective of conducting a hypothesis test on the actual regression coefficients. Something that I found to be incorrect about this analysis was the fact that we performed the model fits on the entire dataset. I suspect that several of the models would not generalize well to unseen data due to the high standard error in some of the explanatory variable's coefficients. Overall, this assignment was very helpful to me.