Assignment 2

April 2021

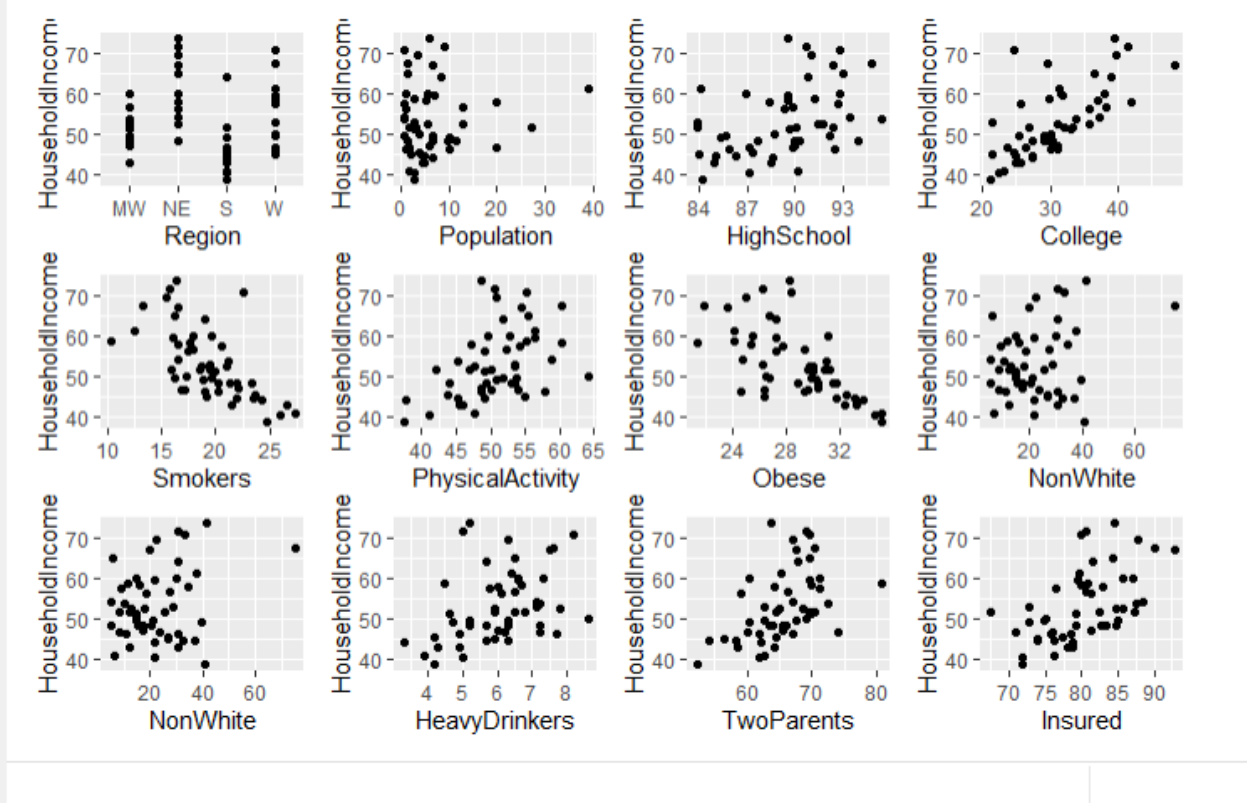
1) In determining the response versus explanatory variables it is important to know what question is being asked of the data and what is trying to be solved. Response variables are those that could be predicted or the focus of a question. Explanatory variables explain changes. Here, response variables could be: Population, Household income, Insured, or Two Parents. Explanatory variables could be: Region, Highschool, college, smokers, physical activity, non white, heavy drinkers. Insure, two parents, and more could take on both roles. The population of interest is that of the United states broken down into states and regions.

2)

-- Variable type: numeric ----------

A tibble: 11 x 11

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Population	0	1	6.36	7.15	0.584	1.86	4.53	6.98	38.8	
HouseholdIncome	0	1	53.3	8.69	39.0	46.8	51.8	58.7	73.5	
HighSchool	0	1	89.3	3.11	83.8	87.1	89.7	91.6	95.4	
College	0	1	30.8	6.08	21.1	25.9	30.2	35.2	48.3	
Smokers	0	1	19.3	3.52	10.3	16.7	19.0	21.5	27.3	
PhysicalActivity	0	1	50.7	5.51	37.4	47.7	50.6	54.1	64.1	
Obese	0	1	28.8	3.37	21.3	26.4	29.4	31.1	35.1	
NonWhite	0	1	22.2	12.7	4.8	13.4	20.8	30.2	75	
HeavyDrinkers	0	1	6.05	1.18	3.3	5.2	6.15	6.78	8.6	
TwoParents	0	1	65.5	5.17	52.3	62.7	65.4	69.5	80.6	
Insured	0	1	80.1	5.49	67.3	76.1	79.9	84.5	92.8	



3)

	HouseholdIncome
HighSchool	0.4308448
College	0.6855909
Smokers	-0.6375225
PhysicalActivity	0.4404166
Obese	-0.6491116
NonWhite	0.2529418
HeavyDrinkers	0.3730143
TwoParents	0.4776443
Insured	0.5496786

After reviewing the data of correlation and scatterplots, it does not seem that a linear regression is appropriate for analysis. All of the variables are too weak in relation against household income to create a strong linear regression model. However, a multiple regression model with the top variables might be more appropriate.

4) You would want to start with the college variable because it has the highest correlation to household income with .68 correlation.

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

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
```

Equation: y_hat = 23.0664 + 0.9801 * X

```
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


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
```

5) Sum of Squared Residuals is 1981.13, Sum of Squares Total is 3700.49, Sum of Squares due to regression is 1739.20, SSR/SST is 0.4700349. The anova table as put above matches these values.

6)

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

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
```

Equation = y_hat = 9.6728 + .8411 * X1 = .2206 * X2

```
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
```

R-Squared value is .48. The coefficient decreases from model 1 to model 2 for college. The r squared value only increased by .01. There is not enough evidence to reject the null hypothesis of zero and Insured doesn’t/t explain much of the variability in the models so it can be rejected.

7)

```
Call:
lm(formula = HouseholdIncome ~ ., data = x2)

Residuals:
   Min     1Q Median     3Q    Max
-7.541 -2.543 -1.260  1.515 15.204

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -15.52228   33.84632  -0.459 0.648996
HighSchool         0.22500    0.52624   0.428 0.671257
College            0.61379    0.19794   3.101 0.003528 **
Smokers           -0.26301    0.42024  -0.626 0.534959
PhysicalActivity  -0.02829    0.25515  -0.111 0.912257
Obese             -0.27036    0.51896  -0.521 0.605257
NonWhite           0.27281    0.06866   3.973 0.000288 ***
HeavyDrinkers      0.52234    0.84689   0.617 0.540883
TwoParents         0.50137    0.26304   1.906 0.063847 .
Insured            0.02526    0.25319   0.100 0.921014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.947 on 40 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.676
F-statistic: 12.36 on 9 and 40 DF,  p-value: 4.541e-09
```

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	0.62076

It seems that the best variables for a predictive model are College and Smokers. The r2s seem to have very similar values the more variables added to it with twoparents and PhysicalActivty being more minimal in terms of effect. The criteria that seems appropriate are those which would have a significant impact on the model and since most of the variables do not they can be discarded. The interpretations do not become more counterintuitive but they do show which variables have minimal or no impact. This means that it would be important to understand all the r squared or other coeffecients regarding impact otherwise you could end up with a bloated model which includes variables that have little to no impact.

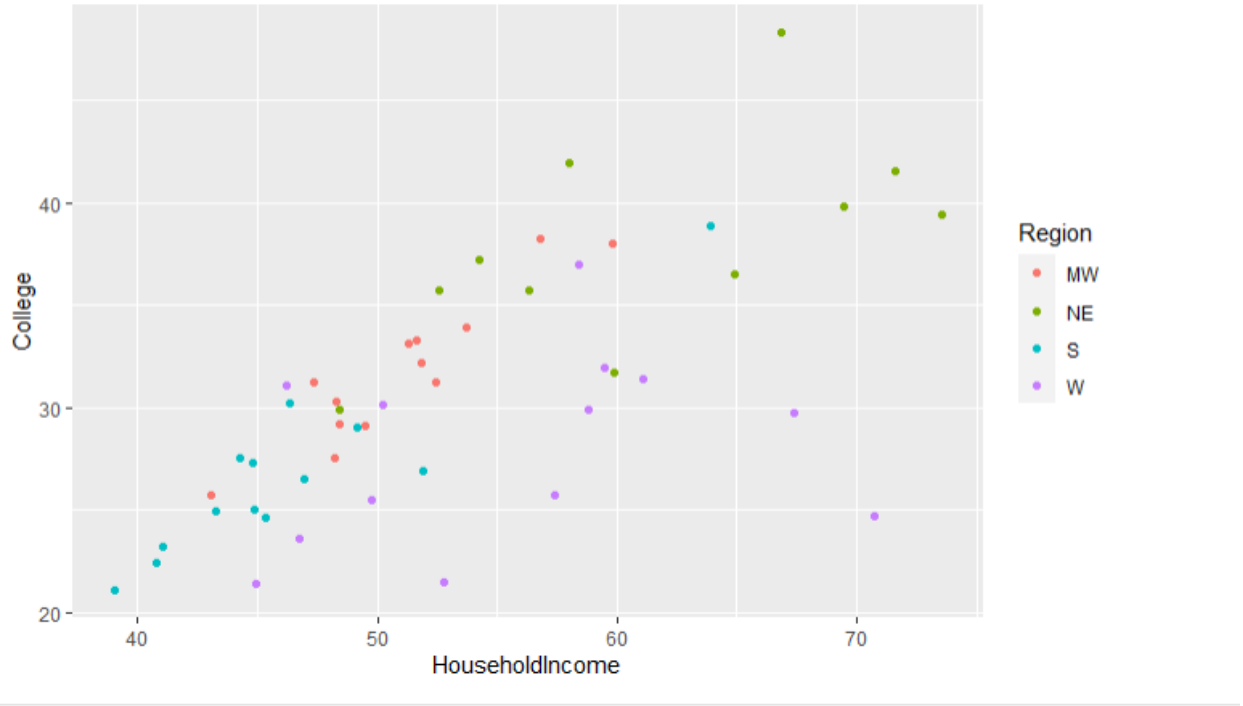
8)

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

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 2.96e-07 ***
College       0.7035     0.1525   4.614 3.06e-05 ***
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
```

```
Response: HouseholdIncome
          Df  Sum Sq Mean Sq F value    Pr(>F)
College    1 1739.36 1739.36  54.077 2.382e-09 ***
Smokers    1  449.39  449.39  13.972 0.0005027 ***
Residuals 47 1511.74   32.16
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This model which is household income against college and smokers has an r squared value of .59, much higher than our previous models. It was necessary to refit this model because those are the best fitting variables which show the most direct impact on household income.

9)



I wanted to see if the variables that we didn't look at like region would have an effect on the data. From the chart it looks like there is as those in the West with low college percentages are still making quite a lot of income, whilst in New England it is those who are highly college educated who make the most income.

10)

The conclusions I can draw are that very few factors affect the r squared or fit for regression models. It seems only College or smokers affects household income in a great amount whilst the other variables are marginal. It was interesting to look at how region affects household income and college, with those

in the west not being very educated but still making quite a lot of money. The story in this data is that region, college education and smoking have a great impact upon the income of the household. Despite their low affects on r squared it would still be a good idea to keep the other variables in the model if one wants to get a more nuanced view of the model. This could especially help with region and state as there might be different regions or states that have higher or lower affects which would be interesting to note.