Pooja Deshpande

Computational Assignment 1
MSDS – 410 Data Modeling for Supervised Learning, Summer 2020
Northwestern University

Contents:

Q.1

US States data consists of demographic and non-demographic variables. The shape of the dataset is (50,13). It is calculated from census data. Given the variables in the dataset, following combinations of response variable and explanatory variables can be considered:

Response variable	**Explanatory variable**
HouseholdIncome	HighSchool, College, Insured,
Obese	HeavyDrinkers, PhysicalActivity, Smokers
PhysicalActivity	Obese, Smokers, HeavyDrinkers
Insured	HighSchool, College, HouseholdIncome

Table 1: Explanatory and response variables

All the above-mentioned response variables can be used interchangeably as explanatory and response variables. The population of interest for this problem is US states.

Q.2

Basic Summary statistics of the continuous variables

Variable Name	**Min**	**Max**	**Mean**	**Median**	**Std.**	**Skewness**	**Kurtosis**
Population	0.584	38.803	6.364	4.532	7.1509	2.603	10.9078
HouseholdIncome	39.03	73.54	53.28	51.76	8.6902	0.6303	2.5824

HighSchool	83.8	95.4	89.32	89.7	3.1071	-0.1859	2.14868
College	21.2	48.3	30.83	30.15	6.0786	0.5455	2.9344
Smokers	10.3	27.3	19.32	19.05	3.5231	0.0810	3.06
PhysicalActivity	37.4	64.1	50.73	50.65	5.509	-0.1859	3.206
Obese	21.3	35.1	28.77	29.4	3.3692	-0.1186	2.3912
NonWhite	4.8	75	22.16	20.75	12.68557	1.4635	7.110
HeavyDrinkers	3.3	8.6	6.046	6.15	1.175	-0.1415	2.5520
TwoParents	52.3	80.6	65.52	65.45	5.1707	-0.0748	3.7476
Insured	67.3	92.8	80.15	79.9	5.4940	-0.0050	2.5627

Table 2: Basic summary statistics

Plots

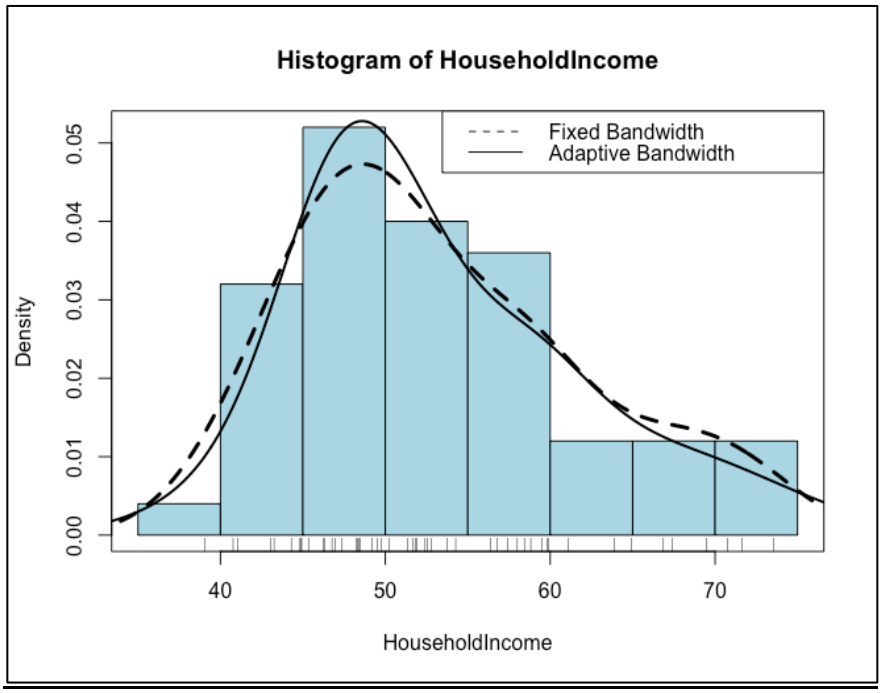

Fig 1: Histogram of HouseholdIncome

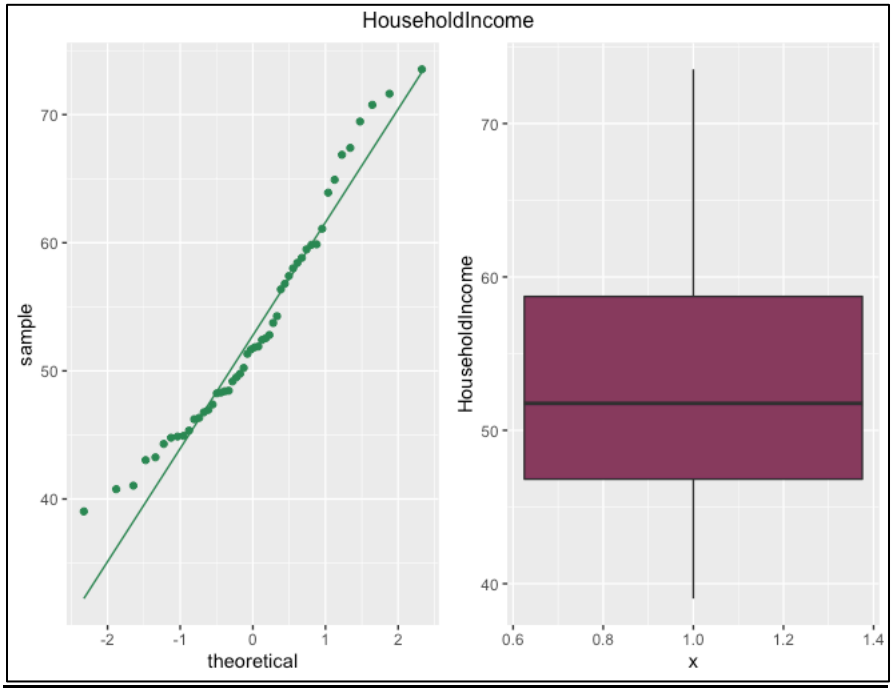

Fig 2: QQ plot and boxplot of HouseholdIncome

HouseholdIncome is slightly skewed to the right and also slightly platykurtic. As observed in the boxplot, there are no outliers.

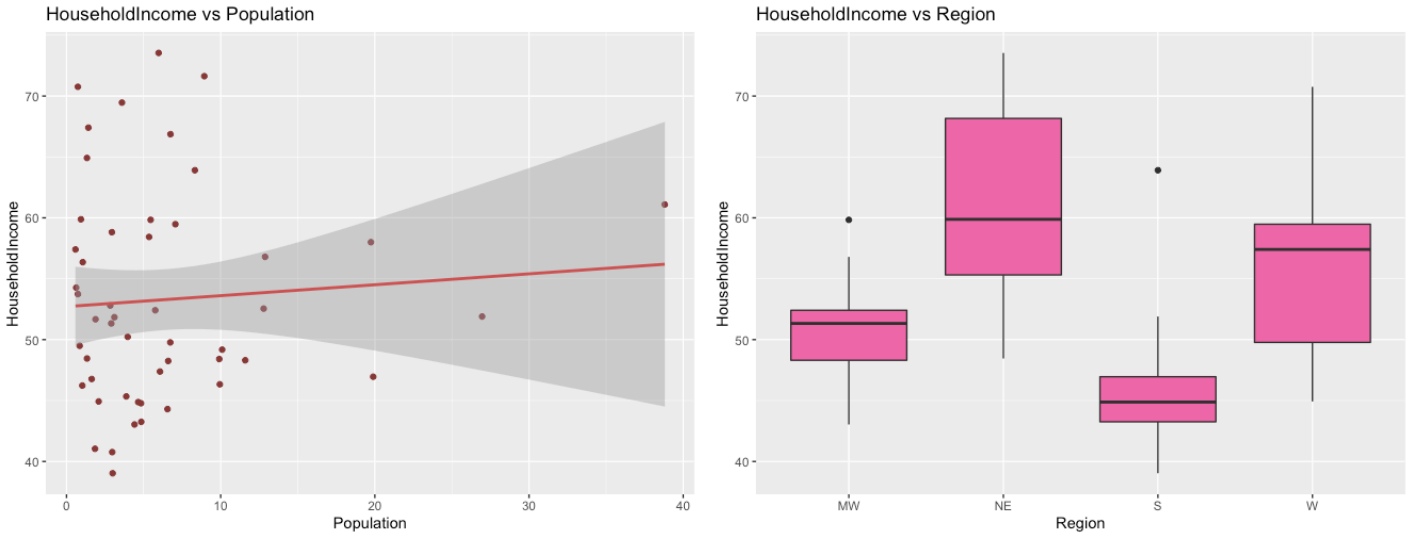

Fig 3,4: HouseholdIncome vs population and HouseholdIncome vs Region

HouseholdIncome is not correlated with population. Population has extreme outliers and the regression line is affected by it. There is a clear trend in the relationship between HouseholdIncome and region. North East and West regions have higher household incomes than the south and north west.

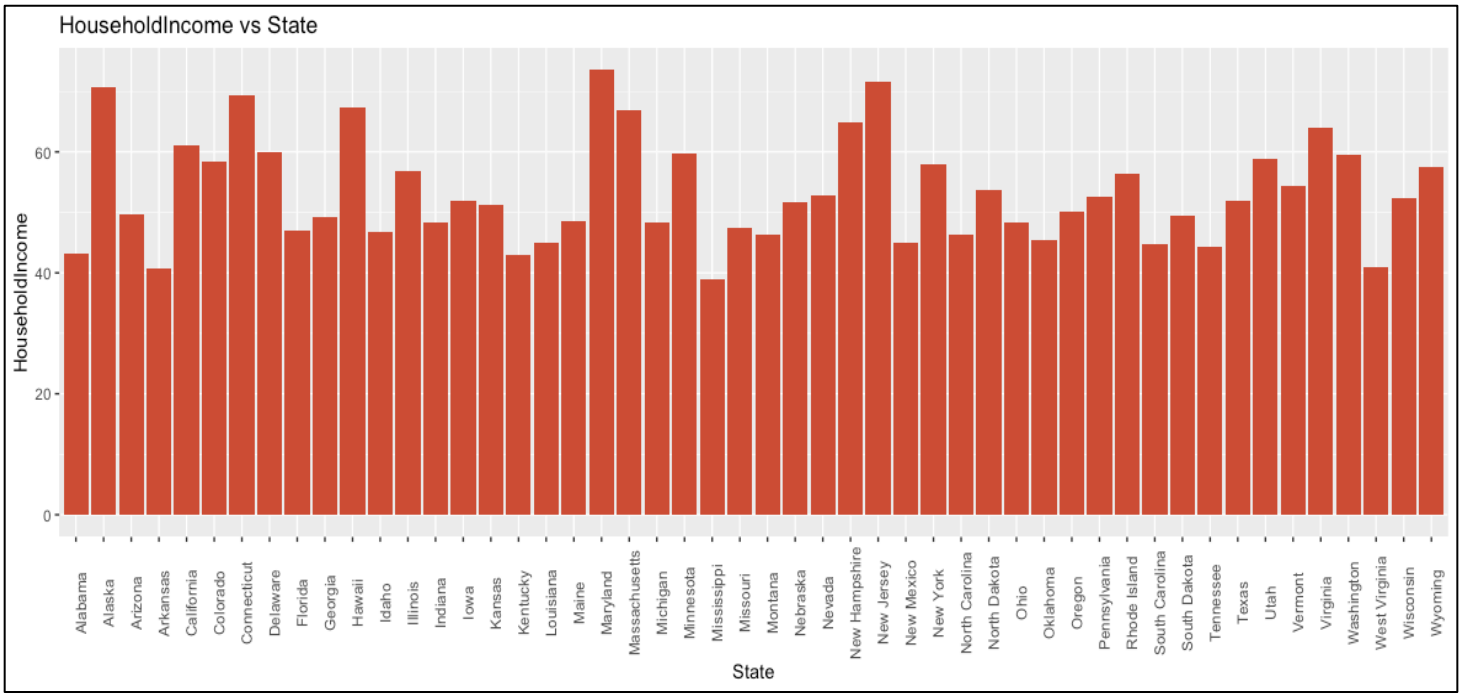

Fig 5: HouseholdIncome vs State

Maryland, New Jersey and Alaska have the highest household incomes whereas Mississippi, West Virginia and Arkansas have the lowest household incomes.

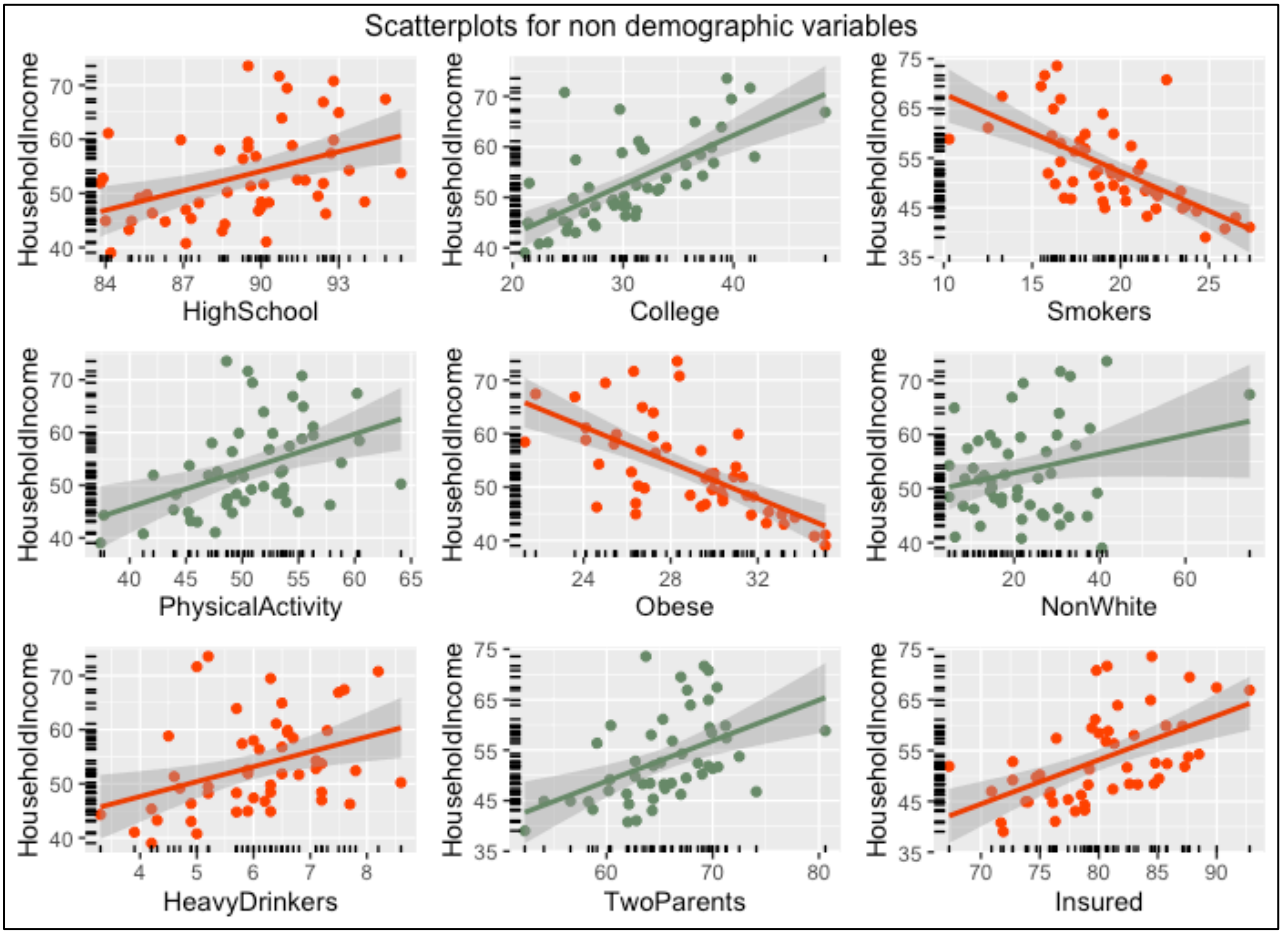

Fig 6: Non-demographic variables

Q.3

Pearson's correlation coefficients

	HouseholdIncome	HighSchool	College	Smokers	PhysicalActivity	Obese	NonWhite	HeavyDrinkers	TwoParents	Insured
HouseholdIncome	1	0.430844784	0.68559094	-0.637522473	0.440416649	-0.649111609	0.252941779	0.373014276	0.477644344	0.549679
HighSchool	0.430844784	1	0.480685013	-0.150124728	0.382558816	-0.301333961	-0.361481657	0.386469782	0.69702228	0.751874
College	0.68559094	0.480685013	1	-0.485521687	0.267362114	-0.519044572	-0.057217763	0.264925317	0.365849984	0.697235
Smokers	-0.637522473	-0.150124728	-0.485521687	1	-0.59024279	0.814866828	-0.155120004	-0.390434212	-0.475431311	-0.27249
PhysicalActivity	0.440416649	0.382558816	0.267362114	-0.59024279	1	-0.780684775	-0.079207166	0.664022009	0.495816725	0.297089
Obese	-0.649111609	-0.301333961	-0.519044572	0.814866828	-0.780684775	1	-0.106160627	-0.555064485	-0.459993042	-0.35112
NonWhite	0.252941779	-0.361481657	-0.057217763	-0.155120004	-0.079207166	-0.106160627	1	-0.066591743	-0.383024387	-0.11074
HeavyDrinkers	0.373014276	0.386469782	0.264925317	-0.390434212	0.664022009	-0.555064485	-0.066591743	1	0.304604178	0.331003
TwoParents	0.477644344	0.69702228	0.365849984	-0.475431311	0.495816725	-0.459993042	-0.383024387	0.304604178	1	0.449379
Insured	0.549678617	0.751874009	0.697234724	-0.272492447	0.297089254	-0.351115912	-0.110736407	0.331003201	0.4493786	1

Table 3: Correlation coefficients of non-demographic variables and response variable

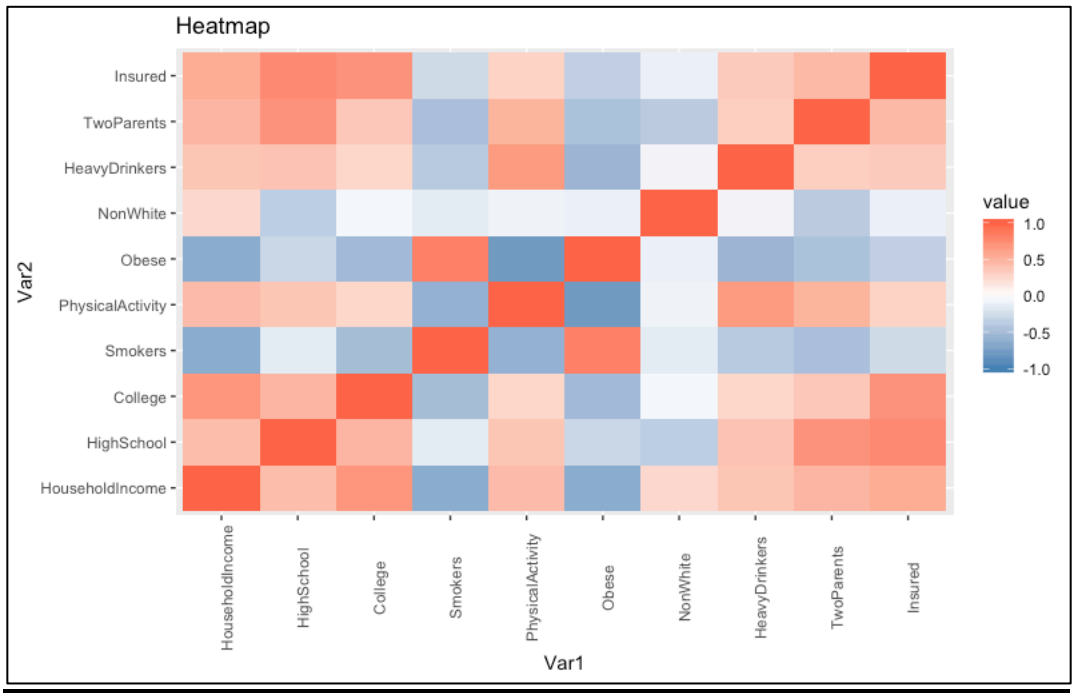

Fig 7: Correlation heatmap

Given the scatterplots, correlation coefficients and the correlation heatmap, it can be concluded that the response variable is linearly related to most of the non-demographic variables. HouseholdIncome is most **strongly related to College** and **weakly related to NonWhite**. It is **negatively correlated to Smokers and Obese**. In this case, linear regression can be used to represent the functional relationship between HouseholdIncome and non-demographic variables. However, simple linear regression may not be the most appropriate since HouseholdIncome has strong correlations with many variables. No single variable will be able to account for the entire variance in HouseholdIncome. Hence, multiple linear regression should be used for this problem.

Q. IV

Model 1 using only College as the explanatory variable:

We can start with College as the explanatory variable since it has the highest correlation coefficient (0.6855). It is positively correlated with HouseholdIncome.

```
> #Question 4: Model with only COLLEGE variable
> model1 <- lm(HouseholdIncome~College,data = data)
> summary(model1)

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

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
```

```
> coef(model1)
(Intercept)     College
 23.0664377   0.9801441
> confint(model1)
                 2.5 %    97.5 %
(Intercept) 13.5788990 32.553976
College      0.6781065  1.282182
> anova(model1)
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
```

Fig 7,8: Model1

Equation form of Model1:

HouseholdIncome = 23.0664 + 0.9801 * College

Model interpretation:

For every unit increase in the proportion of people who attend college, the household income proportion increases by 0.9801. When the proportion of college equals zero, the proportion of household income equals 23.0664

R – Squared = 0.47

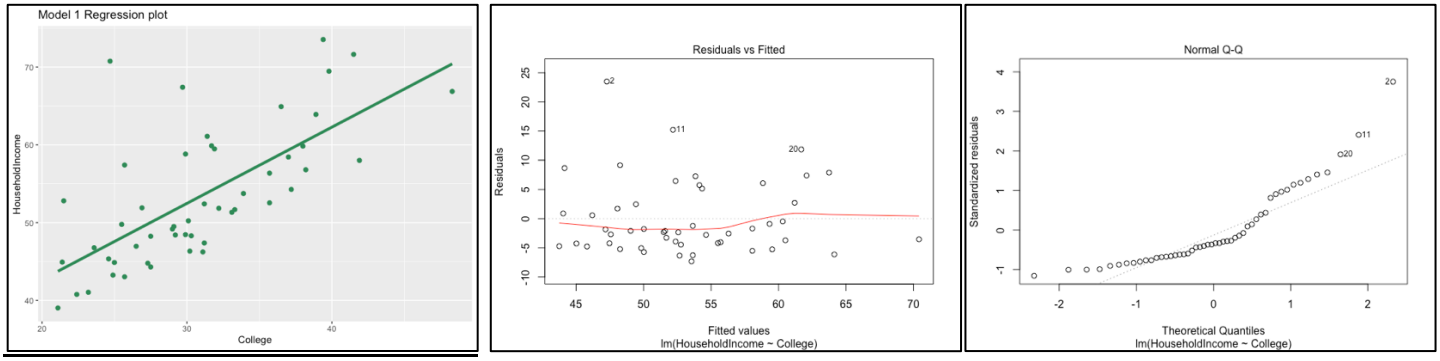

Fig 9,10,11: Regression plot, Residual vs fitted plot, QQ plot of residuals

Verification of coefficient and intercept by hand:

```
> #Verification of coefficient and intercept by hand
> y <- data$HouseholdIncome
> y_bar <- mean(y)
> x <- data$College
> x_bar <- mean(x)
> numerator <- sum((y-y_bar)*(x-x_bar))
> denominator <- sum((x-x_bar)^2)
> b1 <- numerator/denominator
> b1
[1] 0.9801441
> b0 <- y_bar - (b1*x_bar)
> b0
[1] 23.06644
```

Fig 12: Coefficient estimate verification

Q. V

Verifications of R^2 square and ANOVA table:

```
> #Question 5 - Comparing manual calculations with ANOVA table
> y <- data$HouseholdIncome
> y_fitted <- 23.0664 + 0.9801*data$College
> residuals <- y - y_fitted
> squared_residuals <- residuals^2
> sum_squared_residuals <- sum(squared_residuals)
> y_bar <-mean(y)
> mean_deviations <- y - y_bar
> sum_squares_total <- sum(mean_deviations^2)
> fitted_deviations <- y_fitted - y_bar
> sum_squares_regression <- sum(fitted_deviations^2)
> r_squared <- sum_squares_regression/sum_squares_total
> cat("r_squared : ",r_squared)
r_squared :  0.4699927
> cat("sum_squares_regression : ",sum_squares_regression)
sum_squares_regression :  1739.202
> cat("sum_squared_residuals : ",sum_squared_residuals)
sum_squared_residuals :  1961.13
```

Fig 13: Comparing manual calculations of R^2

ANOVA table computations performed manually result in the same values as the ones from the linear model in Q. IV.

Q.VI

Model 2 using College and Insured as explanatory variables:

```
> #Question 6  - Model2
> model2 <- lm(HouseholdIncome~College+Insured,data = data)
> summary(model2)

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

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

> coef(model2)
(Intercept)     College     Insured
  9.6728439   0.8411368   0.2205818
> confint(model2)
                 2.5 %     97.5 %
(Intercept) -20.2272207 39.5729086
College       0.4191356  1.2631380
Insured      -0.2463192  0.6874828
```

```
> anova(model2)
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
```

Fig 14: Model 2 results

Equation form of Model1:

HouseholdIncome = 9.6728 + 0.8411 * College + 0.2206 * Insured

Model interpretation:

For every unit increase in the proportion of people who attend college, the household income proportion increases by 0.8411 when Insured is fixed. For every unit increase in the proportion of people that are insured, the proportion of HouseholdIncome increases by 0.2206 when College is fixed.

R – Squared = 0.48

The partial coefficient for College and the intercept have decreased in magnitude. This is due to the addition of another explanatory variable.

Difference in R-Squared: 0.48-0.47 = 0.01

Addition of another explanatory variable increases the r-squared by 1%. Overall, 48% of the variance in HouseholdIncome is captured by model 2.

For this problem, hypothesis testing results can be used to determine whether Insured should be included in the model or not. The p-value for Insured is 0.34 which means that it doesn't have a statistically significant relationship with HouseholdIncome. It's possible that the correlation has occurred by chance. R-squared always increases with the addition of new explanatory variables. It may be that the variable is actually capturing noise in the data rather than the true pattern. It is always better to check the adjusted r-squared value which penalizes the model for

every new explanatory variable that is added to it. In this case, the adjusted r-squared value has decreased from 0.459 to 0.4579.

Q. VII

Summary of the change in R-squared with each added variable:

Model name	Explanatory variables	R-square	Adjusted R-square
Model 1	College	0.47	0.459
Model 2	College+Insured	0.48	0.4579
Model 3	College+Insured+Smokers	0.6104	0.585
Model 4	College+Insured+Smokers+Obese	0.6225	0.5889
Model 5	College+Insured+Smokers+Obese+TwoParents	0.6278	0.5855
Model 6	College+Insured+Smokers+Obese+TwoParents+PhysicalActivity	0.6296	0.5779
Model 7	College+Insured+Smokers+Obese+TwoParents+PhysicalActivity+ HighSchool	0.63	0.5684
Model 8	College+Insured+Smokers+Obese+TwoParents+PhysicalActivity+ HighSchool+HeavyDrinkers	0.6311	0.5591
Model 9	College+Insured+Smokers+Obese+TwoParents+PhysicalActivity+ HighSchool+HeavyDrinkers+NonWhite	0.7355	0.676

Table 4: Summarization of model results

It can be observed from the above table that the r-squared value always increases with the addition of new explanatory variables. It doesn't consider if the explanatory variable truly captures the pattern in the data. Adjusted R-squared is a better metric to evaluate the model and to choose the explanatory variables. It only increases when the new explanatory variable improves the model fit more than expected by chance. In this case, College, Smokers, Obese and NonWhite variables increase the Adjusted R-squared value.

In models 6, 7, 8 and 9, the partial coefficient of PhysicalActivity has a negative sign. This result is rather counter intuitive. As the model gets more complex, many of the explanatory variables start to become statistically insignificant. When all the variables are used in the model, only College and NonWhite are statistically significant. This could be due to multicollinearity amongst the explanatory variables.

Q. VIII

It can be observed from the previous models that there is a need to refit the model with fewer explanatory variables. Complex models are usually not the best ones. A simpler model is able to generalize to the data better than a complex model. Based on the adjusted r-squared values, **College, Smokers, Obese, and NonWhite** were initially chosen as explanatory variables for the final model. **HighSchool** was also added to the model as it makes intuitive sense that a state with higher proportion of people who have studied high school also has a higher household income. However, **Obese** was then removed from the model as it is very highly correlated with **Smokers (correlation coefficient = 0.81).**

In the final model, all the variables are statistically significant.

Equation form:

HouseholdIncome = -30.48025 + 0.54596 *College -0.87079*Smokers + 0.88093 *HighSchool +

0.22873*NonWhite

```
> #Question 8
> model <- lm(HouseholdIncome~College+Smokers+HighSchool+NonWhite,data=data)
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max
-7.635 -3.064 -1.081  1.854 18.113

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.48025   24.44106  -1.247 0.218816
College       0.54596    0.14977   3.645 0.000689 ***
Smokers      -0.87079    0.23294  -3.738 0.000521 ***
HighSchool    0.88093    0.27854   3.163 0.002799 **
NonWhite      0.22873    0.06092   3.754 0.000496 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.908 on 45 degrees of freedom
Multiple R-squared:  0.707,     Adjusted R-squared:  0.681
F-statistic: 27.15 on 4 and 45 DF,  p-value: 1.703e-11
```

Fig 15: Final model

Model interpretation:

For every unit increase in the proportion of people who attend college, the household income proportion increases by 0.54596 when adjusted for all other explanatory variables. For every unit increase in the proportion of Smokers, the proportion of HouseholdIncome decreases by 0.87079 when adjusted for all other explanatory variables. For every unit increase in the proportion of people who attend HighSchool, the household income proportion increases by 0.88093 when adjusted for all other explanatory variables. For every unit increase in the proportion of NonWhite, the household income proportion increases by 0.22873 when adjusted for all other explanatory variables.

R – Squared = 0.7

Adjusted R-Squared = 0.68

Q. IX

This section looks at the **backward selection process** to finalize the explanatory variables in the multiple linear regression model.

In the backward selection method, a model is built using ALL the explanatory variables. The variable with the highest p-value is removed and a new model is built with the remaining regressors. This process continues until a stopping criterion is met. In this case, the stopping criteria is when the number of predictors reaches a value of 4 (since the final model has 4 predictors).

A complete illustration of this process along with the code can be found in the Appendix section. The final model that was built using this method resulted in a higher R-square and adjusted r-squared scores.

Final model with Backward Selection:

```
> #Build a new model without Smokers
> mod6 <- lm(HouseholdIncome~College+TwoParents+HeavyDrinkers+NonWhite,data = data)
> summary(mod6)

Call:
lm(formula = HouseholdIncome ~ College + TwoParents + HeavyDrinkers +
    NonWhite, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-6.407 -2.854 -1.210  2.302 13.610

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -28.75489    9.96237  -2.886  0.00597 **
College         0.74648    0.12280   6.079 2.39e-07 ***
TwoParents      0.70230    0.15799   4.445 5.68e-05 ***
HeavyDrinkers   1.01669    0.61870   1.643  0.10730
NonWhite        0.30966    0.05848   5.295 3.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.771 on 45 degrees of freedom
Multiple R-squared:  0.7232,    Adjusted R-squared:  0.6986
F-statistic: 29.39 on 4 and 45 DF,  p-value: 4.853e-12
```

Fig 16: Final model with backward selection

The predictors in this model are College, TwoParents, HeavyDrinkers and NonWhite. Two of the predictors in this model are not a part of the final model that was illustrated in **Q.VIII.**

R-Square: 0.7232

Adjusted-R square: 0.6986

This was surprising because TwoParents and HeavyDrinkers did not have high correlation coefficients. It was also found that a model with only three predictors, College, TwoParents and NonWhite performed better than the model that was chosen in **Q.VIII.**

Q. X

The objective of this project was to analyze US States data. The dataset consisted of demographic and non-demographic variables. Exploratory data analysis was performed, and multiple linear regression models were built to predict the HouseholdIncome. Correlation coefficients, R-Square and adjusted r-square were examined. A final model was built using College, HighSchool, NonWhite and Smokers as the explanatory variables. Backward selection process revealed that College, NonWhite, TwoParents, and HeavyDrinkers may be better predictors for HouseholdIncome. Although some of the above variables do not have a high correlation coefficient, they have turned out to be better predictors of HouseholdIncome. This may be due to the fact that there is multi collinearity amongst the explanatory variables. NonWhite is not correlated to many of the predictors, hence most models that include

NonWhite have a higher adjusted r squared. The statistical significance of the variables also depends on multi collinearity. In the future, variance inflation factor could be considered to further analyze correlated explanatory variables.

Appendix

```
> #Question 9
> #Backward selection
> mod1 <- lm(HouseholdIncome~College+Insured+Smokers+Obese+TwoParents+PhysicalActivity+HighSchool+HeavyDrinke
rs+NonWhite,data = data)
> summary(mod1)

Call:
lm(formula = HouseholdIncome ~ College + Insured + Smokers +
    Obese + TwoParents + PhysicalActivity + HighSchool + HeavyDrinkers +
    NonWhite, data = data)

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
College            0.61379    0.19794   3.101 0.003528 **
Insured            0.02526    0.25319   0.100 0.921014
Smokers           -0.26301    0.42024  -0.626 0.534959
Obese             -0.27036    0.51896  -0.521 0.605257
TwoParents         0.50137    0.26304   1.906 0.063847 .
PhysicalActivity  -0.02829    0.25515  -0.111 0.912257
HighSchool         0.22500    0.52624   0.428 0.671257
HeavyDrinkers      0.52234    0.84689   0.617 0.540883
NonWhite           0.27281    0.06866   3.973 0.000288 ***
---
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
```

```
> #Build a new model without Insured as it has the highest p-value
> mod2 <- lm(HouseholdIncome~College+Smokers+Obese+TwoParents+PhysicalActivity+HighSchool+HeavyDrinkers+NonWh
ite,data = data)
> summary(mod2)

Call:
lm(formula = HouseholdIncome ~ College + Smokers + Obese + TwoParents +
    PhysicalActivity + HighSchool + HeavyDrinkers + NonWhite,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
-7.490 -2.656 -1.235  1.544 15.161

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -16.55121   31.84575  -0.520 0.606046
College            0.62382    0.16844   3.704 0.000628 ***
Smokers           -0.26769    0.41255  -0.649 0.520041
Obese             -0.26329    0.50785  -0.518 0.606937
TwoParents         0.49670    0.25570   1.942 0.058968 .
PhysicalActivity  -0.02675    0.25159  -0.106 0.915844
HighSchool         0.25664    0.41491   0.619 0.539637
HeavyDrinkers      0.52345    0.83653   0.626 0.534957
NonWhite           0.27400    0.06679   4.103 0.000189 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.887 on 41 degrees of freedom
Multiple R-squared:  0.7354,    Adjusted R-squared:  0.6838
F-statistic: 14.24 on 8 and 41 DF,  p-value: 1.153e-09
```

```
> #Build a new model without PhysicalActivity as it has the highest p-value
> mod3 <- lm(HouseholdIncome~College+Smokers+Obese+TwoParents+HighSchool+HeavyDrinkers+NonWhite,data = data)
> summary(mod3)

Call:
lm(formula = HouseholdIncome ~ College + Smokers + Obese + TwoParents +
    HighSchool + HeavyDrinkers + NonWhite, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-7.484 -2.571 -1.219  1.505 15.180

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.23176   27.31906  -0.667 0.508190
College         0.62979    0.15693   4.013 0.000242 ***
Smokers        -0.26878    0.40754  -0.660 0.513158
Obese          -0.23193    0.40852  -0.568 0.573240
TwoParents      0.49409    0.25150   1.965 0.056108 .
HighSchool      0.25235    0.40805   0.618 0.539636
HeavyDrinkers   0.48918    0.76282   0.641 0.524825
NonWhite        0.27493    0.06544   4.201 0.000135 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.829 on 42 degrees of freedom
Multiple R-squared:  0.7353,    Adjusted R-squared:  0.6912
F-statistic: 16.67 on 7 and 42 DF,  p-value: 2.697e-10
```

```
> #Build a new model without Obese as it has the highest p-value
> mod4 <- lm(HouseholdIncome~College+Smokers+TwoParents+HighSchool+HeavyDrinkers+NonWhite,data = data)
> summary(mod4)

Call:
lm(formula = HouseholdIncome ~ College + Smokers + TwoParents +
    HighSchool + HeavyDrinkers + NonWhite, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-7.264 -2.481 -1.133  1.615 15.342

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -25.36291   24.06828  -1.054 0.297865
College         0.64385    0.15374   4.188 0.000137 ***
Smokers        -0.40955    0.32088  -1.276 0.208689
TwoParents      0.49718    0.24946   1.993 0.052625 .
HighSchool      0.26856    0.40383   0.665 0.509587
HeavyDrinkers   0.65625    0.69821   0.940 0.352518
NonWhite        0.27873    0.06458   4.316 9.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.791 on 43 degrees of freedom
Multiple R-squared:  0.7333,    Adjusted R-squared:  0.6961
F-statistic: 19.71 on 6 and 43 DF,  p-value: 6.699e-11
```

```
> #Build a new model without HighSchool
> mod5 <- lm(HouseholdIncome~College+Smokers+TwoParents+HeavyDrinkers+NonWhite,data = data)
> summary(mod5)

Call:
lm(formula = HouseholdIncome ~ College + Smokers + TwoParents +
    HeavyDrinkers + NonWhite, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-6.384 -2.606 -1.164  1.352 15.343

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13.94903   16.76602  -0.832  0.40991
College         0.69688    0.13061   5.336 3.16e-06 ***
Smokers        -0.29517    0.26917  -1.097  0.27878
TwoParents      0.61303    0.17741   3.455  0.00123 **
HeavyDrinkers   0.83788    0.63849   1.312  0.19622
NonWhite        0.28055    0.06411   4.376 7.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.76 on 44 degrees of freedom
Multiple R-squared:  0.7306,    Adjusted R-squared:  0.6999
F-statistic: 23.86 on 5 and 44 DF,  p-value: 1.592e-11
```

```
> #Build a new model without Smokers
> mod6 <- lm(HouseholdIncome~College+TwoParents+HeavyDrinkers+NonWhite,data = data)
> summary(mod6)

Call:
lm(formula = HouseholdIncome ~ College + TwoParents + HeavyDrinkers +
    NonWhite, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-6.407 -2.854 -1.210  2.302 13.610

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -28.75489    9.96237  -2.886  0.00597 **
College         0.74648    0.12280   6.079 2.39e-07 ***
TwoParents      0.70230    0.15799   4.445 5.68e-05 ***
HeavyDrinkers   1.01669    0.61870   1.643  0.10730
NonWhite        0.30966    0.05848   5.295 3.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.771 on 45 degrees of freedom
Multiple R-squared:  0.7232,    Adjusted R-squared:  0.6986
F-statistic: 29.39 on 4 and 45 DF,  p-value: 4.853e-12
```