Scott Jue
MSDS 410
1/15/2023

Assignment #2: Fitting and Interpreting Simple Linear Regression Models

Introduction

The dataset provided for this assignment includes state-wide average or proportion scores data on all 50 US states calculated from census data. There are 13 variables and 50 observations, with each observation representing one of the 50 states. There are 2 nominal categorical variables (STATE and REGION) and the remaining 11 variables are continuous variables. The objective of this analysis is to develop various linear regression models for the given dataset to determine if the variables included can be used to predict Household Income and/or any additional response variables of interest.

Results

1. The dataset contains 3 demographic variables (State, Region, and Population). As such, these variables will not be considered explanatory variables for the linear regression model. The remaining variables, however, can be used as either explanatory or response variables. The below table summarizes if these variables can be considered explanatory or response.

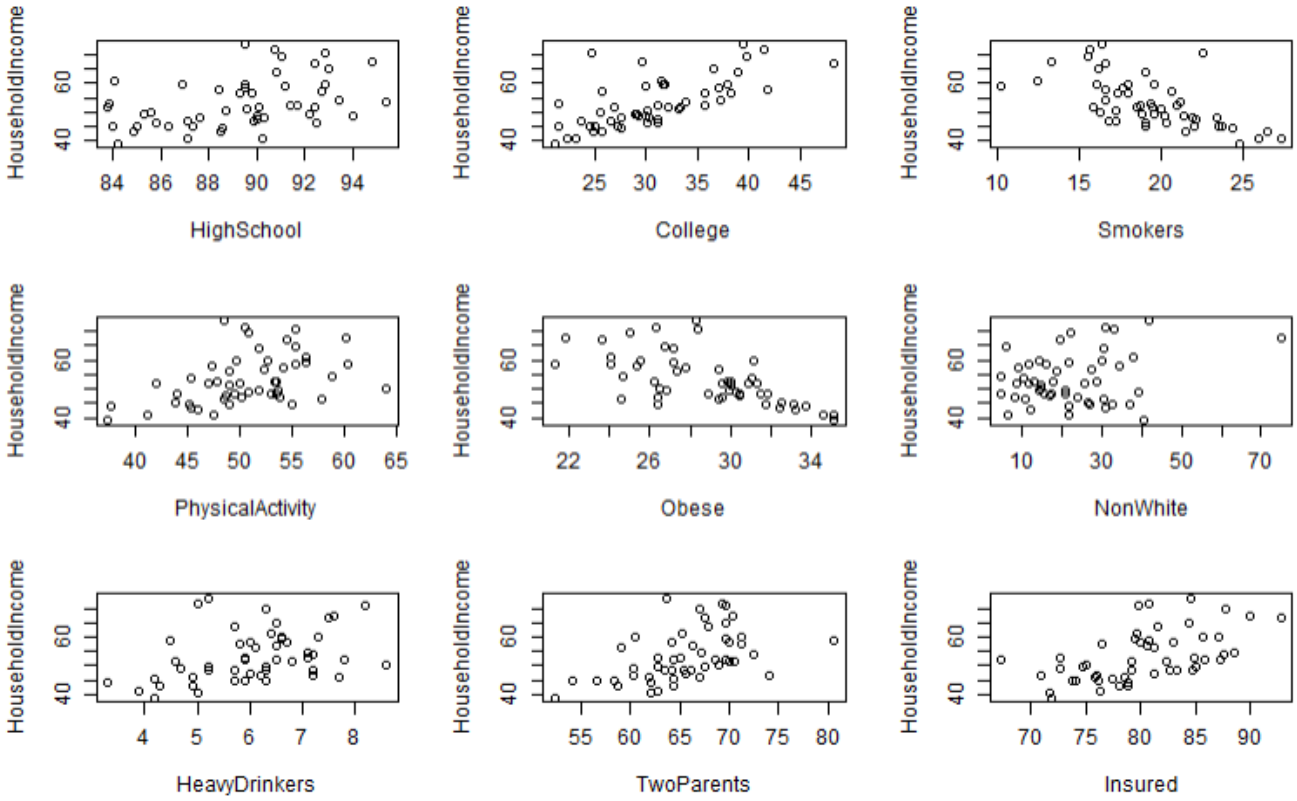
Variable	Can be explanatory (x)?	Can be response (y)?
State	No	No
Region	No	No
Population	No	No
HouseholdIncome	Yes	Yes
HighSchool	Yes	Yes
College	Yes	Yes
Smokers	Yes	Yes
PhysicalActivity	Yes	Yes
Obese	Yes	Yes
NonWhite	Yes	Yes
HeaveDrinkers	Yes	Yes
TwoParents	Yes	Yes
Insured	Yes	Yes

2. The population that we will be modeling is the US population as presented through state-wide average metrics. However, technically there is no "population of interest" in this analysis since the data provided is total US population that is aggregated at the state level. Additionally, since this dataset is census data for all 50 US States, no sampling is needed since we will not be performing inferential statistics. Instead, we will be looking at population parameters and will include all observations (States) in our analysis and modeling.

3. For the initial linear regression model, we will consider HOUSEHOLDINCOME as the response variable. Also, since the STATE, REGION and POPULATION variables are considered demographic variables, we will remove these variables from the dataset since they will not be used for the linear model. The table below displays the summary statistics for all explanatory variables.

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

The scatterplots below were created to visualize any linear relationships between each of the 9 continuous explanatory variables with the response variable HouseholdIncome. It appears that several of the explanatory variables have a linear relationship with HouseholdIncome. We can see that some explanatory variables have a stronger linear relationship to HouseholdIncome than other variables. We can see both positive and negative linear relationships in these plots. For example, COLLEGE and INSURED both show a positive linear relationship, while SMOKERS and OBESE seem to have a negative linear relationship with HouseholdIncome.



4. To continue exploring the linear relationships, the Pearson Product Moment correlation coefficients were calculated for each explanatory variable and shown in the below table. We can see that the College has the highest positive correlation (0.69) and OBESE has the highest negative correlation (-0.65) to HouseholdIncome.

	Pearson Correlation Coefficient
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

Given the scatterplots from step (3) and the correlation coefficients from the above table a simple linear regression model is an appropriate analytical method for this data. The scatterplots show several variables with a linear relationship to HouseholdIncome. Furthermore, after calculating the Pearson correlation coefficients we can see that none of the variables have a coefficient near or at 0 which would suggest no linear relationship. However, some variables are shown to have a stronger relationship than others. Therefore, for optimal modeling we should select the explanatory variables with the strongest relationships as denoted by Pearson correlation coefficients closest to 1 or -1.

5. For the first linear regression model (Model 1), we will select COLLEGE as the explanatory variable (X). Since COLLEGE has the highest Pearson Correlation Coefficient (0.69) with HOUSEHOLDINCOME, we will want to start with this variable as there is a likelihood it will produce the best fitting simple linear regression model from the explanatory variables available to us. As mentioned previously, we will consider HouseholdIncome as the response variable (Y). Using base R STAT package, it has been determined that the linear regression equation for Model 1 is *Y_hat = 23.0664 + 0.9801*COLLEGE.* Below is the R summary output for Model 1.

   Model 1 Summary Output:

```
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 ***
subdat$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
```

   The y-intercept for this equation is 23.0664 which can be interpreted as if COLLEGE=0 then the predicted HouseholdIncomeis 23.0664. However, the minimum value for HouseholdIncome is 39.03. So, the y-intercept for in this equation is only relevant to this dataset and cannot be interpreted outside of this dataset. The coefficient for the explanatory variable COLLEGE can be interpreted as for every unit increase in COLLEGE, HouseholdIncome will increase by .9801. The Multiple R-squared is 0.47. This statistic can be interpreted as 47% of the variability in the response variable can be accounted for or explained by this explanatory variable in this model.

6. For the two parameters in the model, the null hypotheses are that the coefficients are equal to 0. This means that we have a null hypothesis that the y-intercept is 0 and an alternative hypothesis that the y-intercept is not equal to 0. The t-statistic value of 4.88 allows us to reject this null hypothesis that this value is 0. This makes sense as it would be unrealistic for HouseholdIncome to be completely dependent on COLLEGE as this would imply that if there was no college degree, then there would be no household income which is not true. For the second parameter, we have a null hypothesis that the COLLEGE coefficient is equal to 0 and a null hypothesis that the coefficient is not equal to 0. However, the t-value of 6.525 suggests that we should reject the null hypotheses and accept the alternative hypothesis.

Next, we will perform a hypothesis test of the omnibus. That is, we will look at the overall model to see if the model coefficients are equal to 0, which will be our null hypothesis. The alternative hypothesis is that the model coefficients are not equal to 0. The below ANOVA summary output shows an F-static value of 42.572 which suggests that we should reject the null hypothesis that our model coefficients are both equal to 0.

Model 1 ANOVA Table:

```
Analysis of Variance Table

Response: subdat$HouseholdIncome
                Df Sum Sq Mean Sq F value    Pr(>F)
subdat$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
```

7. To ensure accuracy of the ANOVA table and R-squared values from the regression summary output, I have manually calculated Sum of Squared Residuals, Sum of Squares Total, Sum of Squares due to Regression, and R Squared using the below R-code. We can see that these values match the ANOVA table and regression model summary output; therefore, we can confirm their accuracy.

Sum of Squared Residuals:

```
subdat2$resid_sq <- subdat2$residual^2
sumsquare_resid <- sum(subdat2$resid_sq)

sumsquare_resid
```

```
[1] 1961.13
```

Sum of Squares Total:

```
subdat2$y_dev <- (subdat2$HouseholdIncome - y_bar)
subdat2$square_y_dev <- (subdat2$y_dev)^2

sumsquare_total <- sum(subdat2$square_y_dev)

sumsquare_total
```

```
[1] 3700.488
```

Sum of Squares due to Regression:

```
subdat2$yhat_dev <- subdat2$y_hat - y_bar

subdat2$square_yhat_dev <- subdat2$yhat_dev^2

sumsquare_reg <- sum(subdat2$square_yhat_dev)

sumsquare_reg
```

```
[1] 1739.202
```

R Squared:

```
sumsquare_reg/sumsquare_total
```

```
[1] 0.4699927
```

8. Using the below R-code, I calculated the standardized residuals for Model.

Standardized Residuals:

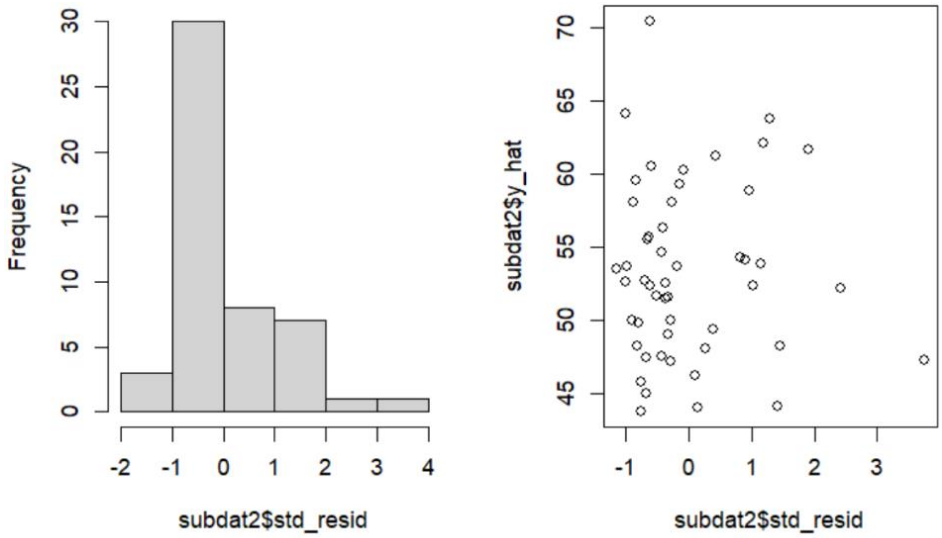
```
n <- length(subdat2$residual)
x_mean <- mean(subdat2$College)
x_ss <- sum((subdat2$College-x_mean)^2)

subdat2$leverage <- 1/n + (subdat2$College-x_mean)^2/x_ss

subdat2$std_resid <- subdat2$residual/(6.392*sqrt(1-subdat2$leverage))
```

Next, I created a histogram of the standardized residuals and a scatterplot with the predict values for HouseholdIncome. The histogram shows a right-skewed distribution with the peak occurring between -1 and 0. Additionally, the majority of the standardized residuals fall within -1 and 2. According to the scatterplot, there does not appear to be any significant linear relationship between the standardized residuals and the predicted values. We can see that most of the points are grouped in between -1 and 0 which is consistent with the histogram. Additionally, I noticed that the 2 points that have a standardized residual greater than 2 have a predicted value below 55. These could be outliers or perhaps there is more variance in the lower HouseholdIncome range.

Standardized Residuals Plots:



9. For Model 2, I will select OBESE as the explanatory variable for the linear regression model, as this variable had the second strongest correlation coefficient (-0.65) with HouseholdIncome. Below is the summary output and ANOVA table for Model 2.

Model 2 Summary Output:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.0291  -3.6348  -0.8212   2.4921  19.4735

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   101.4449     8.2009  12.370  < 2e-16 ***
subdat2$Obese  -1.6742     0.2832  -5.912 3.42e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.679 on 48 degrees of freedom
Multiple R-squared:  0.4213,    Adjusted R-squared:  0.4093
F-statistic: 34.95 on 1 and 48 DF,  p-value: 3.416e-07
```

Model 2 ANOVA Table:

```
Analysis of Variance Table

Response: subdat2$HouseholdIncome
              Df Sum Sq Mean Sq F value    Pr(>F)
subdat2$Obese  1 1559.2 1559.19  34.951 3.416e-07 ***
Residuals     48 2141.3   44.61
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear equation for Model 2 is *Y_hat = 101.449 – 1.6742*OBESE*. The interpretation of this equation is that if OBESE is 0, then the predicted HouseholdIncome is 101.449. Again, this y-intercept value is outside the range of the dataset as the max HouseholdIncome is 73.54. So, this y-intercept is not able to be interpreted outside of this dataset. The coefficient for OBESE can be interpreted that for every unit increase in OBESE the predicted HouseholdIncome decreases by 1.6742 since OBESE is negatively correlated to HouseHoldIncome. For hypothesis testing, the y-intercept has a t-statistic value of 12.37. As such, we should reject the null hypthesis that the coefficient is equal to 0. The OBESE coefficient t-statistuc value is -5.912; therefore, we should also reject the null hypthosis that this coefficient is equal to 0. Furthermore, we can also conclude that we should reject the null hypothesis for the omnibus test since the F-statistic value of this model is 34.951. Model 2 has a R-squared value of 0.4213 which means that about 42% of the variability in the response variable, Household Income, can be accounted for or explained by this explanatory variable in this model.

When determining best fit, we can conclude that Model 1 is a better fit than Model 2 as it has a higher R-squared value (0.47 vs 0.42) meaning that a greater portion of the response variable can be explained by using COLLEGE as the explanatory variable as opposed to OBESE. We can also see that Model 1 also produces a slightly lower residual standard error

10. In selecting the explanatory and response variables for Model 3, I used a heat map to display the correlation coefficients between each variable which can be seen in Appendix A. From this correlation heat map, we can see that the two variables with the highest correlation coefficient (0.81) are SMOKERS and OBESE. Therefore, I will use OBESE as the response variable and SMOKERS as the explanatory variable for this linear model. The summary output and ANOVA table for Model 3 can be seen below.

Model 3 Summary Output:

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.2067 -1.2282  0.3571  1.1008  4.7960

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.71332    1.57050   8.732 1.77e-11 ***
subdat3$Smokers  0.77929    0.08001   9.740 5.96e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.973 on 48 degrees of freedom
Multiple R-squared:  0.664,     Adjusted R-squared:  0.657
F-statistic: 94.86 on 1 and 48 DF,  p-value: 5.964e-13
```

Model 3 ANOVA Table:

```
Analysis of Variance Table

Response: subdat3$Obese
                Df Sum Sq Mean Sq F value    Pr(>F)
subdat3$Smokers  1 369.36  369.36  94.861 5.964e-13 ***
Residuals       48 186.90    3.89
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The linear equation for Model 3 is *Y-hat = 13.71332 + 0.77929*SMOKERS*. The y-intercept tells us that if SMOKERS equals 0, then OBESE is equal to 13.71332. The minimum value in the dataset for OBESE is 21.3; therefore, the y-intercept is outside of the range of the given dataset. For this reason, the y-intercept is not feasible outside of this dataset and cannot be interpreted outside of this context. The coefficient for SMOKERS can be interpreted as for every unit increase in SMOKERS the predicted value for OBESE will increase by 0.77929. For hypothesis testing, we should reject the null hypothesis that the intercept is equal 0 and reject the null hypotheses that the SMOKERS coefficient is equal to 0, since the t-statistic values for the intercept and SMOKERS coefficient are 8.732, and 9.74, respectively. The F-statistic value is 94.861, so we can also reject the null hypothesis for the omnibus test that the overall model coefficients are equal to 0. Lastly, the R-squared value is 0.664 which means that about 66% of the variability in the response variable, OBESE, can be accounted for or explained by the explanatory variable, SMOKER, in this linear model.

Conclusion

From the exploratory data analysis, we can see that there are several variables in the dataset that display a linear relationship to HouseholdIncome. These variables can be used as explanatory variables to help us predict the average household income using a linear regression model. From the analysis performed, we can see that COLLEGE has the strongest positive linear relationship to HouseholdIncome, while OBESE has the strongest negative linear relationship. Looking at summary statistics for Model 1 and Model 2, I can conclude that a college education generally is correlated to a higher average household income and that populations that have higher rates of obesity tend to have lower average household income. However, if were to select one model, Model 1 provides a better fit to the dataset provided as shown through the higher R-squared value. Nonetheless, households that have

a higher household income generally obtain a college education and remain physically healthy such that they do not fall into the obese category. Obesity is also highly correlated with smoking as there is a high correlation coefficient between these variables. Model 3 is a regression model that represents the linear relationship between these variables. This model suggests the smoking can account for roughly two-thirds of the variability in obesity. So, we can further conclude that since obesity and smoking are highly correlated that households with higher income generally have a college education, are not obese, are non-smoking. It is important to note that we are not able to draw causation between these variables as part of this analysis.

From this assignment, I was able gain additional proficiency in creating OLS regression models. Although, I have experience in creating these types of regression models, I have never really conducted hypothesis testing on the regression coefficients. So, this was a new learning area for me. One thing that I found to be a limiting factor of this analysis was the lack of a data dictionary and detail surrounding the dataset. However, I suspect this happens often in the real world. So, taking that into account, I appreciated the challenge in working with data that is not spoon-fed to us in the most clean and detailed format.

Appendix A: Correlation Heat Map