Modeling Assignment 2: Fitting and Interpreting Simple Linear Regression Models

The Data

"The US State data set (USStates.xlsx) is a 12 variable dataset with n=50 records. The data, calculated from census data, consists of state-wide average or proportion scores. As such, higher scores for these variables translate into having more of that quality."

Note that valid values for these variables will range from 0-100.

Assignment Tasks

1. Which variables can be considered explanatory (X) and which considered response (Y)? Can any variables take on both roles? Make a table that summarizes your conclusions.

State is more of an identifier and Population is more of a weight variable which is why they are not included in the tables below.

General Variables

Possible Response (Y)	Possible Both (X and Y)	Possible Explanatory (X)
		Region NonWhite TwoParents

Region Nonwhite and TwoParents are all demographics variables that can fit into any story.

Possible "Stories" to Investigate

Story #	Possible Response (Y)	Possible Both (X and Y)	Possible Explanatory (X)
1.) Education vars	College		HighSchool
2.) Health vars	Obese		Smokers PhysicalActivity HeavyDrinkers
3.) Money vars		HouseholdIncome Insured	

Based off looking at the data College could be predicted from HighSchool because they are education-related variables, and HighSchool rates may influence College rates. Obese could be predicted from Smokers, PhysicalActivity, and HeavyDrinkers (if this applies to the individual who answered) because they are health-related variables, and Obesity could be a result of the other variables. HouseholdIncome and Insured may be related as well because they are both money-related variables, either of which may be a predictor or response.

2. What is the population of interest for this problem (yes – this is a trick question!)? Be sure your answer is clear and complete.

The population of *data* is not clearly defined. All we know is that the dataset reflects US Census data for 50 states. The most we can assume is that the population of interest is a subset of households that took the Census in one of the 50 states of the dataset. The time period is unknown.

The population of *interest* is determined by what "story" we are trying to tell. For example, predicting Statewide rates of Highschool would indicate that our population of interest may be States with a high school graduation rate (if that is what the variable represents) over 50%. As there are several different stories that can be told (about a vaguely defined dataset), the population of interest will change depending on the population we would like to draw conclusions about.

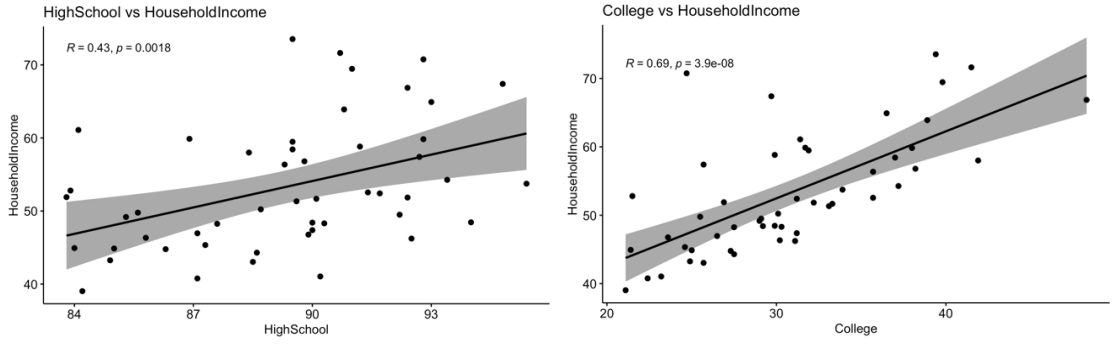
3. For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all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

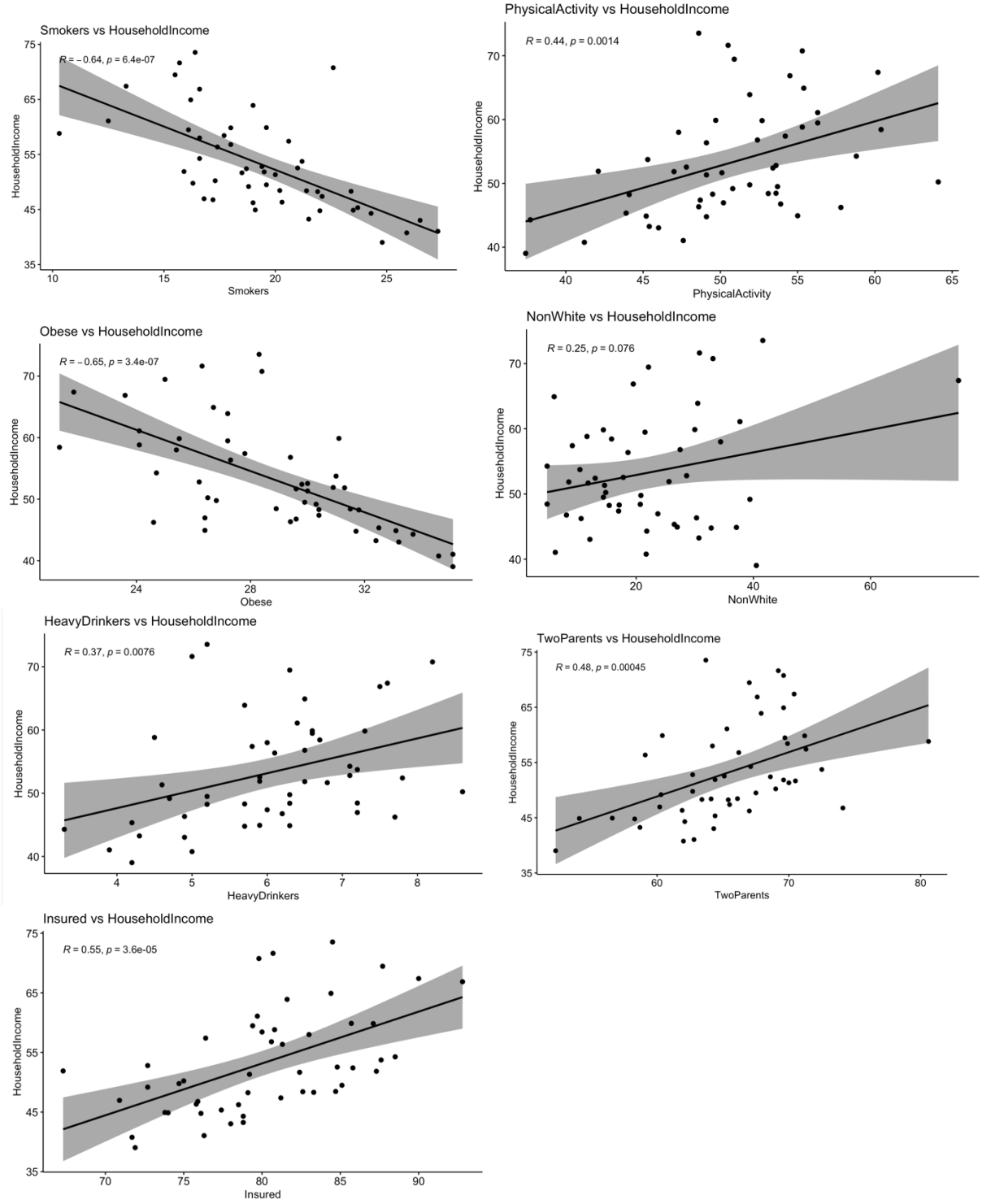
Basic Summary Statistics of Each Variable

	State	Region	Population	HouseholdIncome	HighSchool	College	Smokers	PhysicalActivity	Obese	NonWhite	HeavyDrinkers	TwoParents	Insured
nbr.val	*NA*	*NA*	50	50	50	50	50	50	50	50	50	50	50
nbr.null	*NA*	*NA*	0	0	0	0	0	0	0	0	0	0	0
nbr.na	*NA*	*NA*	0	0	0	0	0	0	0	0	0	0	0
min	*NA*	*NA*	0.58	39.03	83.8	21.1	10.3	37.4	21.3	4.8	3.3	52.3	67.3
max	*NA*	*NA*	38.8	73.54	95.4	48.3	27.3	64.1	35.1	75	8.6	80.6	92.8
range	*NA*	*NA*	38.22	34.51	11.6	27.2	17	26.7	13.8	70.2	5.3	28.3	25.5
sum	*NA*	*NA*	318.2	2664.21	4466	1541.5	965.8	2536.7	1438.3	1107.8	302.3	3276.2	4007.4
median	*NA*	*NA*	4.53	51.76	89.7	30.15	19.05	50.65	29.4	20.75	6.15	65.45	79.9
mean	*NA*	*NA*	6.36	53.28	89.32	30.83	19.32	50.73	28.77	22.16	6.05	65.52	80.15
SE.mean	*NA*	*NA*	1.01	1.23	0.44	0.86	0.5	0.78	0.48	1.79	0.17	0.73	0.78
CI.mean	*NA*	*NA*	2.03	2.47	0.88	1.73	1	1.57	0.96	3.61	0.33	1.47	1.56
var	*NA*	*NA*	51.14	75.52	9.65	36.95	12.41	30.36	11.35	160.92	1.38	26.74	30.18
std.dev	*NA*	*NA*	7.15	8.69	3.11	6.08	3.52	5.51	3.37	12.69	1.18	5.17	5.49
coef.var	*NA*	*NA*	1.12	0.16	0.03	0.2	0.18	0.11	0.12	0.57	0.19	0.08	0.07

The average HouseholdIncome appears to be 53.28%. the average HighSchool rate (89.32%) is much higher than the average College rate (30.15%). The variable with the largest standard deviation (and widest distribution of data) is NonWhite (12.69%). Values range from 0.58-95.4 across all variables which is within the valid region. There are no missing values in the data.

Scatterplots of Non-demographics Explanatory Variables vs. HouseholdIncome





Here are a few observations from the scatterplots above: College seems to have the strongest relationship to HouseholdIncome (R = 0.69) where higher rates of college are related to higher household incomes. Rates of Smokers (R = -0.64) and Obese (R = -0.65) however seem to indicate a negative relationship with Household Income. All other non-demographic explanatory variables have weaker relationships with HouseholdIncome based off the scatterplots. It also appears that NonWhite may have an outlier of a predominantly POC state with a high Income.

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

Pearson Product Moment correlations + Linear Regression Assumptions

Variable	PPM Corr	Linear Relationship Between Variables	Independent Residuals (Durbin-Watson)	Normality of Residuals (Shapiro-Wilk)	Constant Variance of Residuals (ncvTest)
HighSchool	0.4308448	Yes, the scatterplots from step 3 all indicate a *linear* relationship with Household Income.	0.2037	0.1438	0.41645
College	0.6855909		0.2483	0.1804	0.37503
Smokers	-0.6375225		0.7051	0.8029	0.24569
PhysicalActivity	0.4404166		0.7351	0.9018	0.27302
Obese	-0.6491116		0.8409	0.7077	0.028969
NonWhite	0.2529418		0.2741	0.0003358	0.068237
HeavyDrinkers	0.3730143		0.3907	0.926	0.90145
TwoParents	0.4776443		0.1665	0.5808	0.29188
Insured	0.5496786		0.2653	0.9957	0.5744

The linear relationship between explanatory variables and HouseholdIncome is confirmed through the scatterplots in step 3. However, we ran tests on the 3 assumptions for linear regressions and found the following:

1) Via the Durbin-Watson test, all variables have independent residuals.
2) Via the Shapiro-Wilk test, all except NonWhite variables have a normal distribution of errors. This can be fixed by removing the outlier from NonWhite before using it in a linear regression.
3) Via the ncvTest, all except Obese variables have a constant variance of residuals. It is apparent in the scatterplot that the spread of residuals becomes smaller as Obese % becomes larger. We can exclude Obese from any linear regressions built with Income.

5. 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 prediction equation for Model 1 and interpret each coefficient of the model in the context of this problem. In addition, report and interpret the R-squared statistic for Model 1.

We use College as the first predictor of HouseholdIncome in Model 1 because it has the strongest correlation with each other. In other words, compared to any other 1-variable model, College will explain the most variation in HouseholdIncome.

Model 1: lm(HouseholdIncome ~ College)
predicted(HouseholdIncome) = 23.0664 + 0.9801*College

B_0 is 23.0664 which means that a state with 0% College would be predicted to have a 23.0664% value of HouseholdIncome. This tells us that all predicted values of HouseholdIncome will have a mimimum average rate of 23.0664%.

B_1 is 0.9801 which means that for every additional 1% of College for a given state, that same state's average predicted HouseholdIncome will go up by 0.9801%. This tells us that states with higher average rates of college will also have higher average rates of HouseholdIncome.

The R2 of this model is 0.458994 which means that this model (College in particular) explains 45.9% of the variation in HouseholdIncome.

6. From your Model 1 results for task 5) – Specify the null and alternative hypothesis separately for each of the two parameters in the model. Report and interpret the results of the T-tests for these hypotheses. In addition, state the null and alternative hypotheses for the omnibus (i.e. overall) model. Report the ANOVA table and interpret the results of the F-test.

Test	Null hyp	Alt hyp	Stat	Conclusion
B_0	B_0 = 0	B_0 =! 0	t = 4.888 p = 1.18e-05	Reject null. B_0 is not 0
B_1	B_1 = 0	B_1 =! 0	t = 6.525 p = 3.94e-08	Reject null. B_1 is not 0
Omnibus	B_1 = 0	B_1 =! 0	p = 42.57 on 1 and 48 DF p = 3.941e-08	Reject null. B_1 is not 0

ANOVA table for Model 1

```
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
```

From the tables above, we can note that the constant (B_0) and coefficient (B_1) of Model 1 are statistically significant (and not 0). The F test tells us if any coefficients are non-0, but because Model 1 only has 1 coefficient, we conclude that B_1 is not 0.

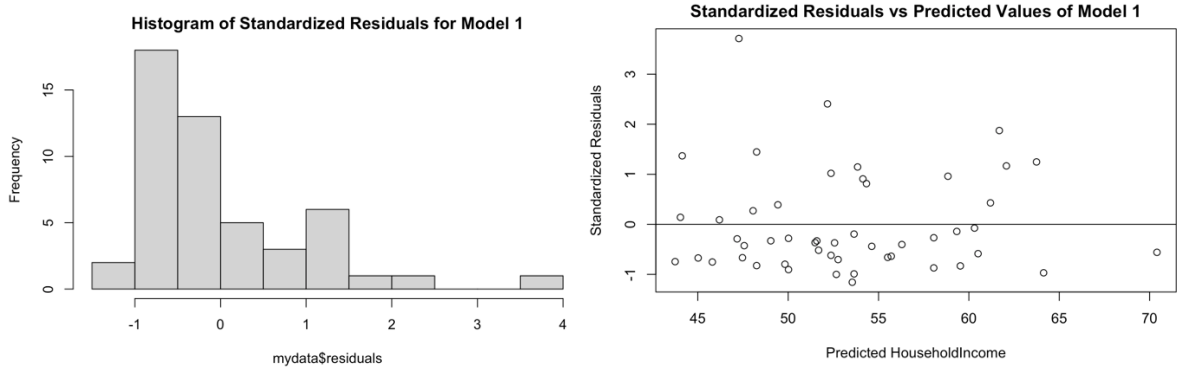
7. For Model 1, write R-code to
   - Y – Y_hat. Square the deviations and add them up. This is sum of squared errors.
   - Y – Y_bar. Square the deviations and add them up. This is sum of squares total.
   - Y_hat – Y_bar). Square the deviations and add them up. This is sum of squares due to regression.
   - R2: (Sum of Squares due to Regression) / (Sum of squares Total)

   Verify and note the accuracy of the ANOVA table and R-squared values from pt 4.

	R Code	R Output	ANOVA Output
SSE	SSE = sum((mydata$HouseholdIncome - mydata$HouseholdIncome_hat) ^ 2) #SSE = sum((Y - Y_hat) ^ 2)	1961.13	1961.1
SST	SST = sum((mydata$HouseholdIncome - mean(mydata$HouseholdIncome)) ^ 2) # SST = sum((Y - Y-bar) ^ 2)	3700.488	3700.46
SSR	SSR = sum((mydata$HouseholdIncome_hat - mean(mydata$HouseholdIncome)) ^ 2) # SSR = sum((Y_hat - Y-bar) ^ 2)	1739.202	1739.36
R2	SSR / SST	0.4699927	0.47

The calculations using manual R Code confirm the ANOVA outputs for all summary statistincs including R2. They are accurate within 0.05 for SSE, SST, and R2, and within 0.2 for SSR which are relatively close.

8. From task 7 you created a variable of residuals for Model 1. Write R-code to standardize the residuals. Do not use residuals from the lm(). Plot the standardized residuals using a histogram. Also, plot the standardized residuals in a scatterplot with the predicted values. Discuss what you see in these two graphs.



The histogram shows that the distribution of standardized residuals is skewed right with a median value of about -0.75. The scatterplot shows the residuals vs fitted plot also indicate the variance of residuals are not evenly distributed above and below the 0 residual line – the positive residuals are typically greater in magnitude than those below, whereas ideally we would like to see more of a balance between the two in a simple linear regression.

9. Select a different explanatory variable and use that variable in a Simple Linear Regression model to predict Y, HOUSEHOLDINCOME. Call this Model 2. Report and interpret the results of Model 2. Which is the better model, Model 1 or Model 2? Give evidence to justify your answer.

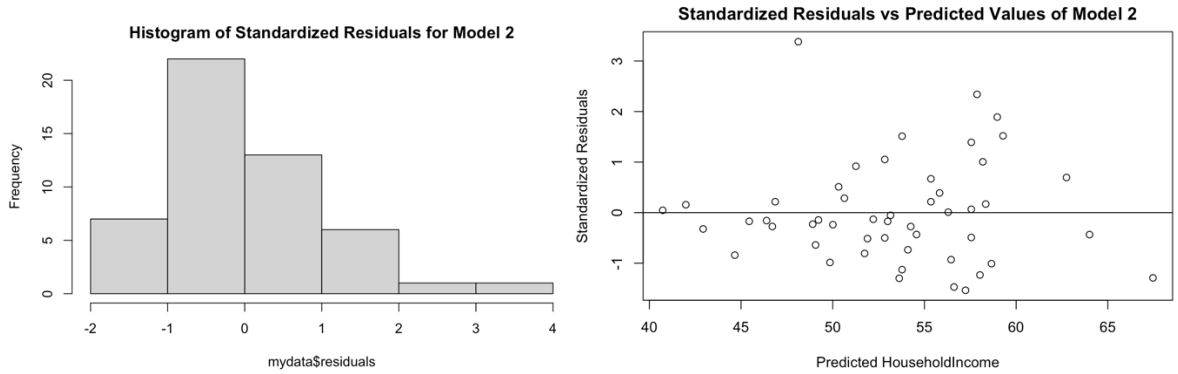
Although Obese has the next strongest correlation, it failed the Constant Variance test, so model 2 will use Smokers.

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

Residuals:
    Min      1Q  Median      3Q     Max
-10.285  -4.074  -1.100   2.434  22.640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.6593     5.3840  15.539  < 2e-16 ***
Smokers      -1.5725     0.2743  -5.733  6.4e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.765 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: 6.396e-07
```

Model 2: lm(HouseholdIncome ~ Smokers)
predicted(HouseholdIncome) = 23.0664 + 0.9801*Smokers

Similar to Model 1, Model 2 has statistically significant B_0, B_1, and Omnibus values with p-values less than 0.05. However, Model 2 has an adjusted R2 value of 0.394069 which is less

than the R2 value of Model 1 (0.458994) meaning that the model with Smokers explain 6-7% less variation than the model with College.



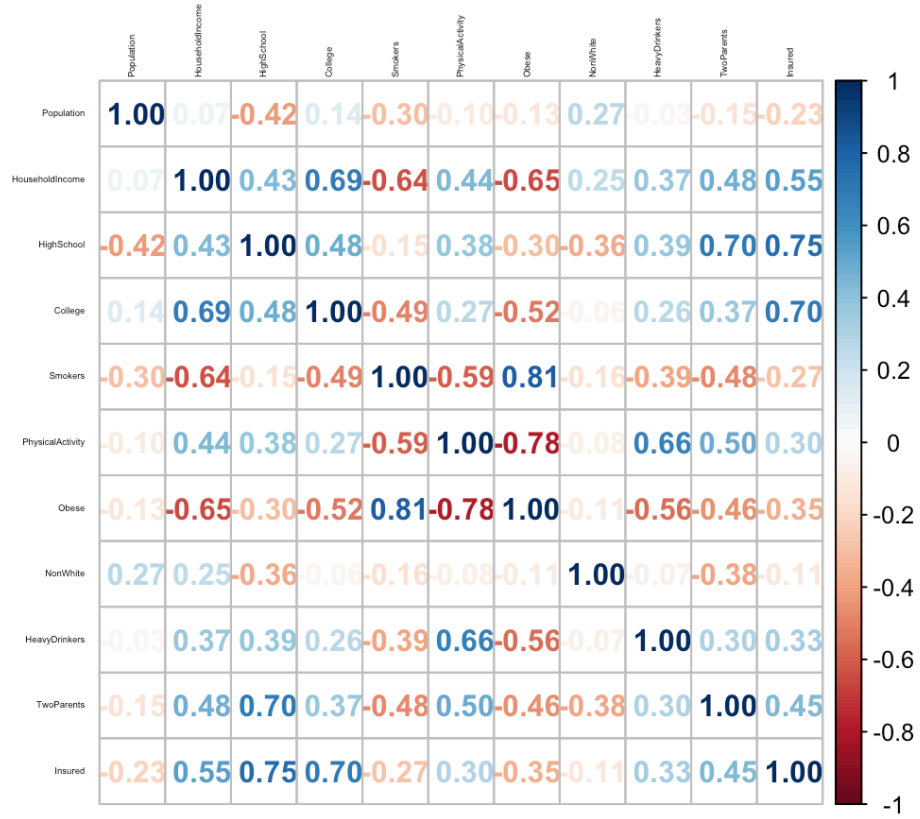
A quick look at the histogram and scatterplot of standardized residuals of Model 2 will also indicate that they are skewed right with a typically negative residual just like Model 1.

With similar residuals and significant model coefficients, Model 1 is better due to its higher R2 value.

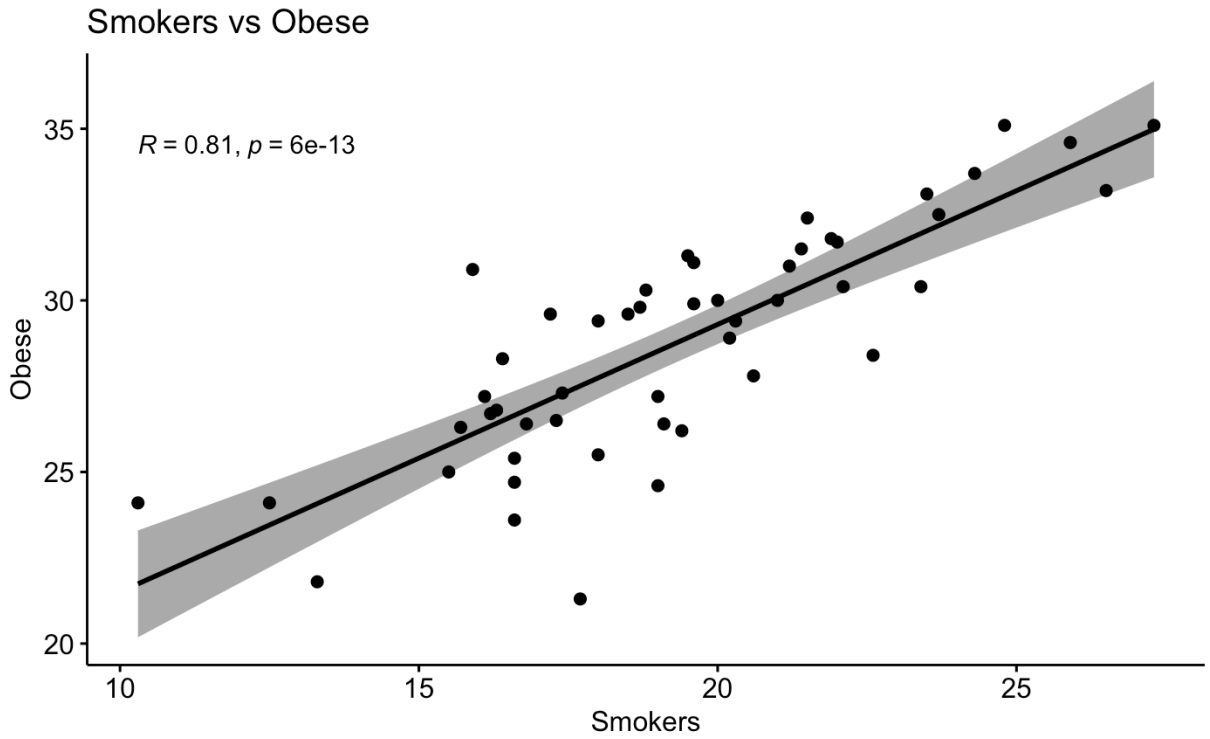
10. For this last task, you are welcome to fit any Simple Linear Regression model that you wish on the US States data. You'll need to decide on the response variable as well as the explanatory variable. Call this Model 3. Report and interpret the results of Model 3.

A quick glance at the correlation plot below shows the highest correlation (0.81) between Obese and Smokers. We will investigate this relationship in Model 3

.



Check the model assumptions:



The scatter plot confirms a linear relationship between variables.

Independent Residuals (Durbin-Watson)	Normality of Residuals (Shapiro-Wilk)	Constant Variance of Residuals (ncvTest)
0.7935	P(Smokers) = 0.8029 P(Obese) = 0.7077	0.11371

This model does not violate any of the 3 assumptions needed to perform a simple linear regression.

Model 3 Summary:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.71332    1.57050   8.732 1.77e-11 ***
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
```

Predicted(Obese) = 13.71332 + 0.77929 * Smokers

B_0

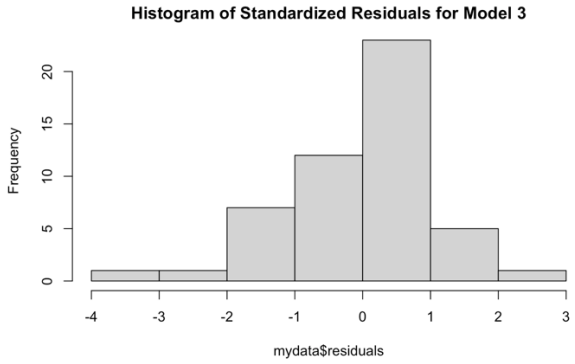
P-value < .05, can conclude that a non-0 B_0 is statistically significant. In this example, Model 3 predicts a data point with 0% Smokers to have a 13.7% Obese. Although there are no datapoints with 0-value Smokers, this constant shows that all data points are predicted to have Obese rates above 13.7% despite the value of Smokers.

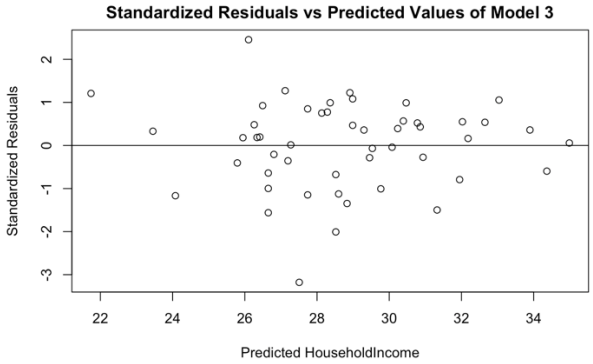
B_1

P-value < .05, can conclude that a non-0 B_1 is statistically significant. In this example, Model 3 predicts an average increase of 0.78% Obese for every 1% increase in Smokers. This shows that states with higher average rates of smoking typically also have higher average rates of Obesity. However, we cannot jump to causation of one due to the other.

Additionally, the F-stat p-value indicates the entire model is statistically significant (at least one of the coefficients are non-0) and the adjusted R2 value tells us that 65.7% of the variation in Obese can be explained by Smokers.

Standardized Residuals





Both the histogram and scatterplot show a fairly normal distribution of residuals with an average residual value close to 0 which is what we would like to see. Model 3 provides a strong relationship between explanatory and response without violating any assumptions and producing output with proper residual distribution.

Conclusion

Here are some key takeaways from Assignment 2.

Defining the population of interest is important before jumping into creating models. Doing so can require looking into data documentation and metadata available on your dataset. Caution must be taken when proceeding with a dataset will little to no documentation.

Always check assumptions of modeling and make sure to have either visual or statistical evidence that backs a claim.

Model interpretation can be very difficult when variables are not clearly defined. Although difficult, it is still possible to interpret model coefficients and summary statistics in a meaningful and contextual way.

Not all models are valid, it is important to check model outputs after to make sure that errors are independent, normally distributed, with a constant variance.

Always check and double check results, and always make sure to explain results in context.