

USStates Report

1. Dataset overview

The *USStates* dataset contains 12 variables with state-wide average scores from census data, with each state contributing one record for a total of fifty observations. A higher score indicates a greater number of that observation in that state. Of the 12 variables, 10 are continuous and 1 is character data types. Moreover, the 10 continuous variables span a variety of different areas in state demographics, such as population, household income, education level, and health behaviors.

Whether a variable is considered explanatory (X) or response (Y) depends on the overarching goals of the model. Some variables align more naturally with the role of a response variable because they represent outcomes of interest. For instance, variables such as *Population* or *Household Income* could be modeled as response variables when the goal is understanding which factors predict differences across states. Other variables, such as *PhysicalActivity*, *Smokers*, or *NonWhite*, describe characteristics of state populations that may help explain variance seen in response variables, therefore, are more often used as explanatory variables.

Some variables can more reasonably take both the response variable and explanatory variable role. For example, *Household Income*, is commonly used as a response variable in OLS regression models to understand demographic factors that predict income. However, *HouseholdIncome* could also be used as an explanatory variable in an OLS regression model where the outcome is obesity or insurance coverage. Thus, the role of a variable is dependent on the research question being asked. Table 1 summarizes the variables in the dataset according to whether they are more commonly used as response variables, explanatory variables, or both.

Table 1. Roles of Variables in the USStates Dataset

Variables	Response Variable	Explanatory Variable
State	-	-
Region	-	X
Population	X	X
HouseholdIncome	X	X
HighSchool	-	X
College	-	X
Smokers	X	X
PhysicalActivity	-	X
Obese	X	X
NonWhite	-	X
HeavyDrinkers	X	X
TwoParents	-	X
Insured	X	X

2. Population of interest

The population of interest is the collection of U.S. states, as state-level observational units are characterized by average demographic, socioeconomic, and health-related measures derived from census data. Each record in the dataset represents a summary of individuals living in each state, rather than the individual people themselves. Since the dataset contains all 50 states, it is a census of the states, not a sample drawn from a larger population. As such, any statistical analysis pertains to the differences across states and does not represent individual-level outcomes.

3. *HouseholdIncome* Summary Statistics & Scatterplots

Table 2 below illustrates the basic summary statistics for the response variable (*HouseholdIncome*) and the demographic variables that we determine (*State*, *Region*, and *Population*). Since there are only 50 records for each state, the *State* variable contains one record for each state. Additionally, there are only four distinct regions, South (S), West (W), Midwest (MW), and Northeast (NE). All regions except for the Northeast contained 13 states, which had 11 states. This suggests that there is an even class distribution between the regions. The *HouseholdIncome* mean value was 53.28 and the standard deviation was 8.69. This indicates a widespread of data points and some slightly right skewness. Similarly, the *Population* mean value was 6.34 and the standard deviation was 7.15, again, pointing to a greater distribution of data points and some large data point in the right tail pulling the distribution to the right.

Table 2. Summary Statistics of HouseholdIncome with demographic variables

Summary Statistics

HouseholdIncome (Y) and Demographic Variables				
Variable	N	Mean	SD	Unique
HouseholdIncome	50	53.284	8.690	NA
Population	50	6.364	7.151	NA
State	50	NA	NA	NA
Region	50	NA	NA	4
Region_MW	13	NA	NA	NA
Region_NE	11	NA	NA	NA
Region_S	13	NA	NA	NA
Region_W	13	NA	NA	NA

Scatterplots were created between the response variable, *HouseholdIncome*, and each non-demographic explanatory variable, where most variables appeared generally linear and positive. Variables such as higher education level (*College*), physical activity, and insurance illustrated

strong associations with *HouseholdIncome*, suggesting that higher values of these predictors are associated with higher average household income across states. *College* specifically indicated the strongest positive linear association with *HouseholdIncome*. Both *Obese* and *Smokers* showed a strong negative relationship with household income, indicating that states with higher obesity rates and smokers had lower average household incomes. Overall, the strengths of these relationships vary, the scatterplots do suggest that linear models would be appropriate for most predictors, with *Obese* and *Smokers* standing out as variables with an inverse relationship with household income. Figures A1-5 in the appendix display the scatterplots with the relationships described above.

4. Pearson Product Moment Correlation

Table 3 below reports the Pearson Product Moment correlation coefficients between household income (Y) and each non-demographic explanatory variable (X). The correlations indicate moderate to strong linear relationships between *HoseholdIncome* and some of the predictors. *College* shows the strongest positive correlation ($r = 0.69$), while *Obese* ($r = -0.65$) and *Smokers* ($r = -0.64$) exhibit strong negative correlations with *HoseholdIncome*. Other variables such as *Insured*, *PhysicalActivity*, and *TwoParents* demonstrated moderate positive associations, whereas *NonWhite* shows only a weak positive correlation.

Table 3. Pairwise Pearson Product Moment Correlation Coefficients

Pearson Correlations with Household Income	
Variable	Correlation
HighSchool	0.43
College	0.69
Smokers	-0.64
PhysicalActivity	0.44
Obese	-0.65
NonWhite	0.25
HeavyDrinkers	0.37
TwoParents	0.48
Insured	0.55

Based on the scatterplots developed in section 3 and the correlation coefficients in Table 2, simple linear regression models would be an appropriate analytic method for examining the relationship between *HouseholdIncome* and individual non-demographic explanatory variables. The scatterplots indicate a generally linear relationship without strong curvature in most variables, and the Pearson's correlation coefficients confirm meaningful linear relationships for

most variables. However, while a simple linear regression is appropriate for individual predictors, the presence of strong correlations among explanatory variables suggests that multiple linear regression may be more suitable for capturing the joint effects of these predictors and address any potentially confounding in a more comprehensive model.

5. Model 1 Fit & Interpretation

A simple linear regression model was fit to predict *HouseholdIncome* (Y) using *College* (X) as the explanatory variable. *College* was chosen as the initial predictor as it had the strongest positive Pearson correlation coefficient with *HouseholdIncome* ($r = 0.69$) among the individual non-demographic explanatory variables and demonstrated a clear linear relationship in the scatterplots (Figure A1). This makes it a natural starting point for a simple linear regression model.

5.1 Prediction Equation for Model 1

The fitted prediction equation is:

$$\widehat{\text{HouseholdIncome}} = 23.07 + 0.98(\text{College})$$

The intercept (23.07) represents the predicted average *HouseholdIncome* for a state in which 0% of the population has a college degree. While this value is not realistic in practice, it serves as a baseline for regression line. The slope for *College* (0.98) indicates that for each 1% increase in the proportion of residents that get a college degree, the average household income is expected to increase on average by \$980, holding all else constant

5.2 R² Interpretation

The R² value of Model 1 was 0.47, meaning that 47% of the variation in *HouseholdIncome* of the model is explained by the differences in college level education alone. This indicates a moderately strong positive relationship and suggests that college education is an important predictor of household income across states.

6. Model 1 Hypothesis Testing

Model 1 can be written as $\widehat{\text{HouseholdIncome}} = \beta_0 + \beta_1(\text{College})$. The model has two parameters estimated in the model, the intercept (β_0) and the slope associated with *College* (β_1). Separate hypothesis tests are conducted for each parameter to assess the statistical significance. For the intercept, the null hypothesis states that $\beta_0 = 0$, while the alternative hypothesis states that $\beta_0 \neq 0$. This test evaluates whether the expected household income is zero when the *College* value is zero. While the intercept estimate (23.07) is statistically significant different from zero ($t = 4.89, p < 0.001$), this result is not meaningful because a state with zero college level education is unrealistic. For the slope, the null hypothesis states that $\beta_1 = 0$, meaning that there is no linear relationship between *College* and *HouseholdIncome*, while the alternative hypothesis states that $\beta_1 \neq 0$. The estimated slope was 0.98, which is statistically significant different than 0 ($t = 6.53, p < 0.001$). This results in strong evidence of a positive linear association between the college-

level educated residents and household income, suggesting that states with higher levels of college education tend to have significantly higher average household income.

In addition to the individual parameters, an omnibus test is used to evaluate the overall significance of the regression model. The null hypothesis for the omnibus model test states that the model with *College* as the predictor variable does not explain more variation in household income than a model with no predictors. The alternative hypothesis for the omnibus model test states that the model explains a significant amount of variation. The ANOVA table reports an F-statistic of 42.57 with 1 and 48 degrees of freedom was highly significant ($p < 0.001$), which leads to rejection of the null hypothesis.

Overall, the results of both the individual t-tests for the College coefficient and the omnibus F-test suggest that college education is statistically significant predictor of household income. The model explains a meaningful amount of variation in household income across the states.

7. Manual Calculation of ANOVA table

For Model 1, predicted values of household income were calculated using the fitted linear regression equation from section 5. Residuals were then computed as the difference between the observed household income and the predicted values for each state. Using these residuals, several important sums of squares were calculated to verify the variance decomposition underlying the regression model.

First, the residuals were summed and squared to obtain the sum of squared errors (SSE), which measures the proportion of variability in *HouseholdIncome* not explained by the model. Next, the sum of squares total (SST) was calculated the deviations, subtracting the mean *HouseholdIncome* from each observed value, squaring and summing them. This value represents the total variability in household income across the states. Lastly, the sum of squares due to the regression (SSR) was calculated by subtracting the mean HouseholdIncome from each predicted value, squaring these deviations and then summing them. This value represents the amount of variability explained by Model 1.

The ratio of SSR to SST was calculated to obtain the manual computed R^2 statistic. When the code was executed, this R^2 statistic was identical to the R^2 statistic reported by the ANOVA table output for Model 1. This confirms the ANOVA results and the R^2 statistic produced by the base stats function are consistent with the fundamental definitions of ANOVA in the simple linear regression and verifies the accuracy of the Model 1 output.

8. Standardized Residual Plots

Standardized residuals were manually calculated for Model 1 by dividing the residuals calculated from section 7 and the standard deviation of the residuals. Figure A6 illustrates the histogram of the standardized residuals for Model 1. The histogram shows the residuals are generally centered around zero, which is consistent with the assumptions of linear regressions. Most of the standardized residuals are between -2 and 2 standard deviations, however, the distribution is slightly right skewed, with a small number of positive residuals exceeding 3

standard deviations from the mean. This indicates the presence of one or two potential outlier states where *HouseholdIncome* is higher than the predicted value. The distribution of the residuals does not appear to be perfectly normal, but the overall shape is reasonably close to normal given the small sample size of 50 states and no extreme outliers present.

The scatterplots of the standardized residuals and predicted values display a larger random scatter of data points around the horizontal zero line, suggesting a strong assumption of linearity. Figure A7 reports the scatterplot of the standardized residuals versus the predicted values. In the scatterplot, there is no evidence of curvature, suggesting that variance of the residuals is approximately uniform across the range of the predicted household income values. There are a few observations that exhibit larger positive or negative residuals; however, these points do not seem to form a pattern and do not suggest any violation of homoscedasticity.

These diagnostic plots indicate that the assumptions of normality, linearity, and constant variance are reasonably satisfied with Model 1. While there are a small number of potential outliers, the residual behavior does not suggest serious model specification error, signaling the appropriateness of a simple linear regression of household income on college education.

9. Model 2 Fit & Interpretation

Model 2 uses *Obese* as the explanatory variable to predict *HouseholdIncome*. This variable was chosen as it had the second highest correlation coefficient with *HouseholdIncome* ($r = -0.65$) and was one of two variables that had a negative association with *HouseholdIncome*. The fitted regression equation for Model 2 is:

$$\text{HouseholdIncome} = 101.44 - 1.67(\text{Obese})$$

The estimate slope for *Obese* is -1.67 and is statistically significant ($t = -5.92, p < 0.001$). This indicates a negative linear relationship between *Obese* with *HouseholdIncome*, for each one-percentage point increase in the *Obese* variable, the *HouseholdIncome* decreases by an average of 1.67 . The intercept suggests that when *Obese* is zero, the predicted value of *HouseholdIncome* is an average of 101.44 , which is unrealistic that the average obesity would be zero. The overall model is statistically significant, as reported by the ANOVA table with an F-test ($F = 34.95, p < 0.001$). The R^2 statistic for Model 2 was 0.421 , meaning that 42.1% of the variation in *HouseholdIncome* is explained by the *Obese* value alone. This indicates a moderately strong relationship for a simple linear regression.

9.1 Comparison of Model 1 & Model 2

When comparing Model 1 and Model 2, Model 1 is a better model. Model 1 has a higher R^2 value than Model 2, suggesting that a larger proportion of the variability in *HouseholdIncome* is explained by Model 1. Additionally, Model 1 had a lower residual standard error, indicating that more accurate predictions are made on average over Model 2. *Obese* is a statistically significant predictor of *HouseholdIncome* and shows a negative linear relationship with *HouseholdIncome*, however, it does not perform as well as an explanatory variable used in Model 1 (*College*) in terms of overall fit.

10. Model 3 Fit & Interpretation

Model 3 uses *Insured* as the explanatory variable to predict *HouseholdIncome*. This variable was chosen because it had a strong correlation to *HouseholdIncome* ($r = 0.55$). The fitted regression equation for Model 3 is:

$$\widehat{\text{HouseholdIncome}} = -16.40 + 0.87(\text{Insured})$$

The slope coefficient for *Insured* is positive and statistically significant ($t = 4.56, p < 0.001$), suggesting that states with higher insurance coverage tend to have higher household income. Specifically, for every one-percentage point increase in insured rate, the predicted household income increases by approximately 0.87, on average. The intercept is not statistically significant ($t = -1.07, p = 0.29$), which suggests that the predicted income when insured is zero is not meaningfully different than zero. Like Model 1 and Model 2, this value has little practical interpretation since this scenario is out of the range of data. The ANOVA results confirm that the overall model was statistically significant ($F = 20.78, p < 0.001$), meaning that *Insured* explains a significant amount proportion of the variability in *HouseholdIncome*. The R^2 value for Model 3 is 0.302, indicating that 30.2% of the variation in *HouseholdIncome* is explained by *Insured* alone. While the model was statistically significant, it demonstrated the weakest explanatory power than Model 1 and Model 2, making it the least strong of the three models.

A. Appendix

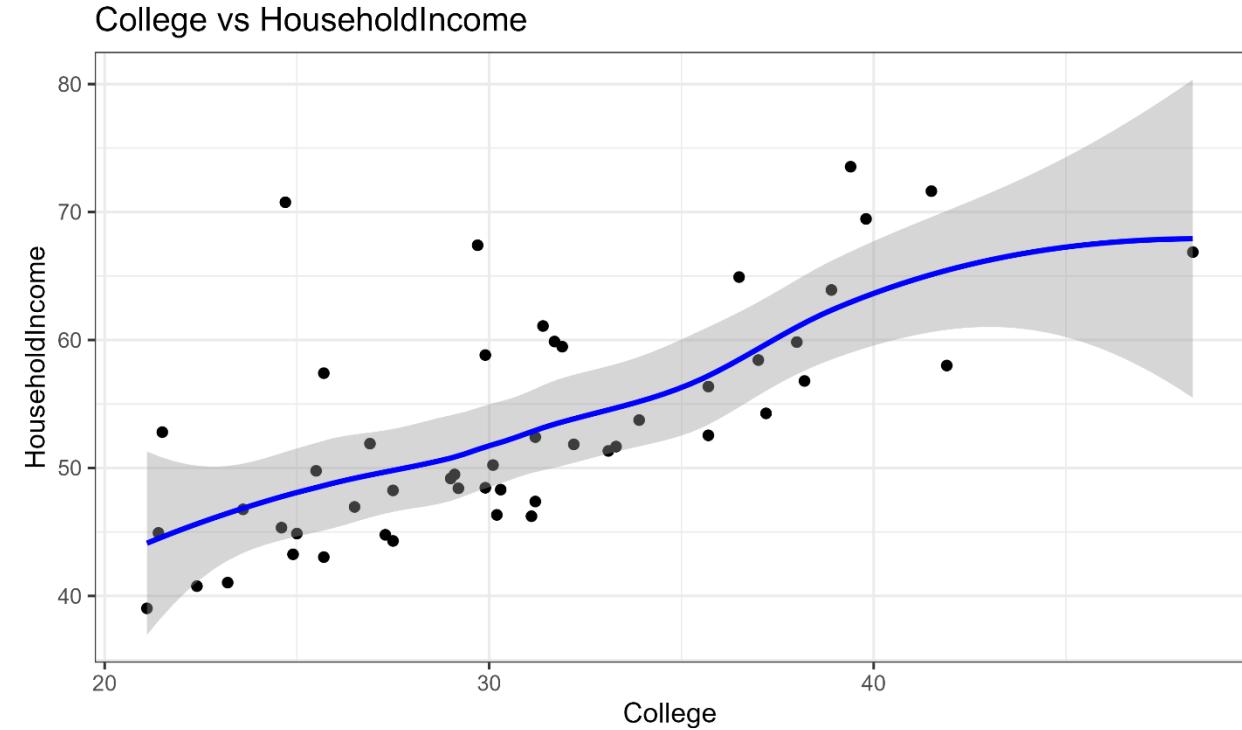
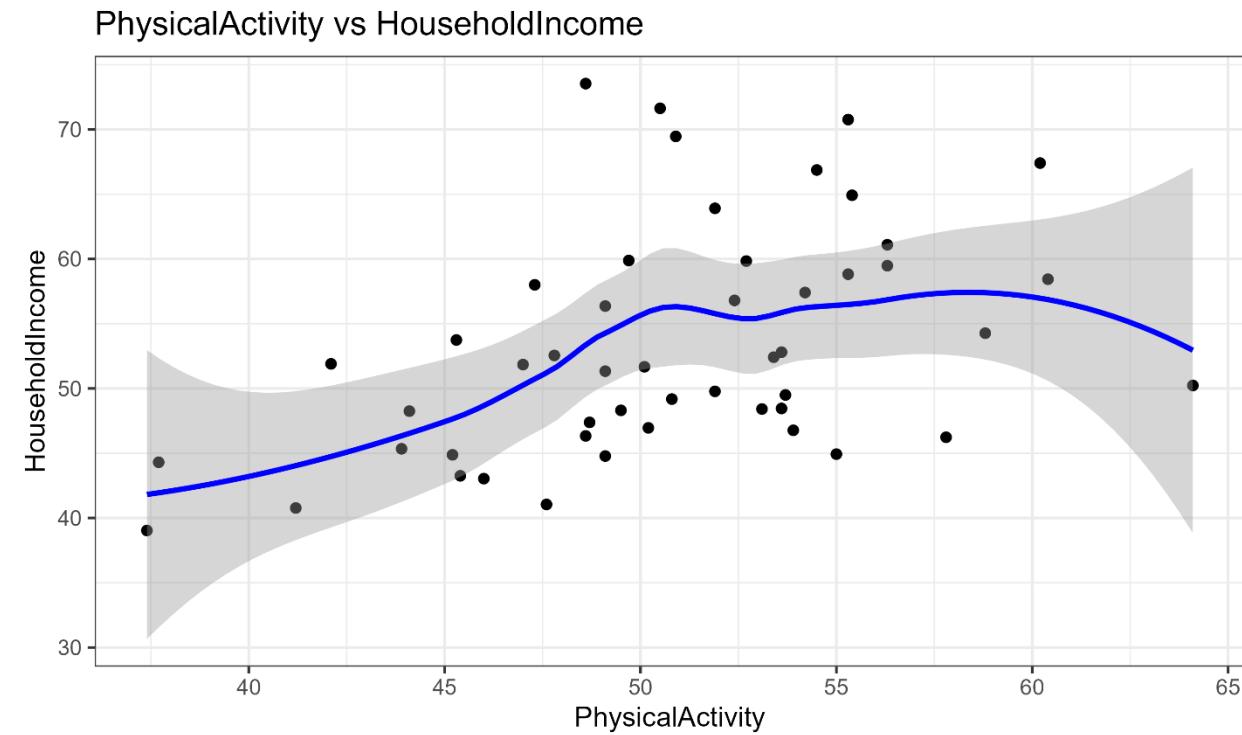
Figure A1. Scatterplot of College vs HouseholdIncome*Figure A2. Scatterplot of PhysicalActivity vs HouseholdIncome*

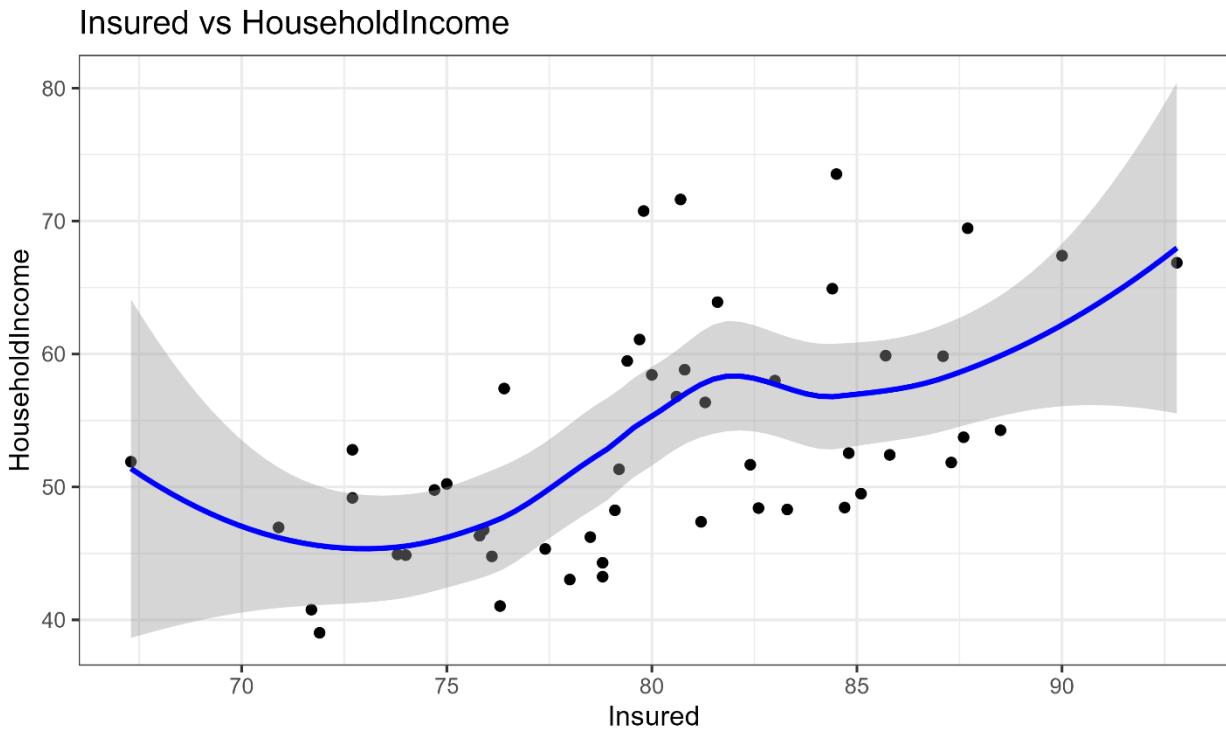
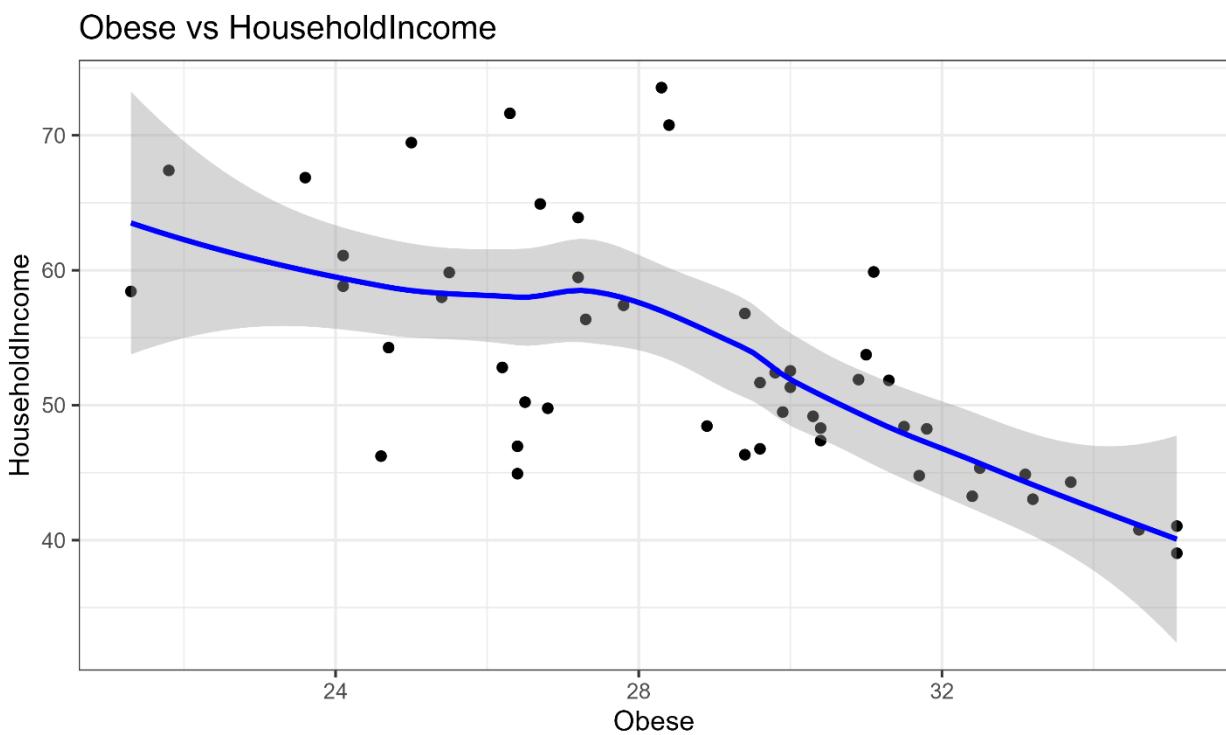
Figure A3. Scatterplot of Insured vs HouseholdIncome*Figure A4. Scatterplot of Obese vs HouseholdIncome*

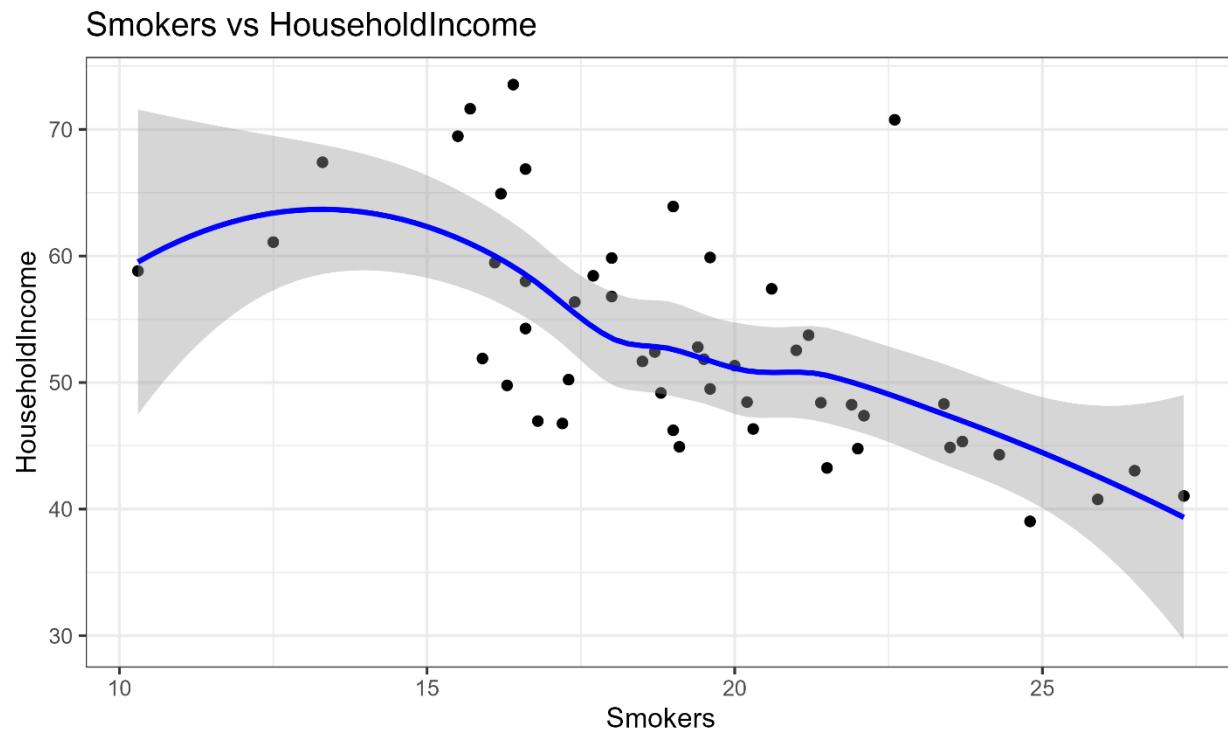
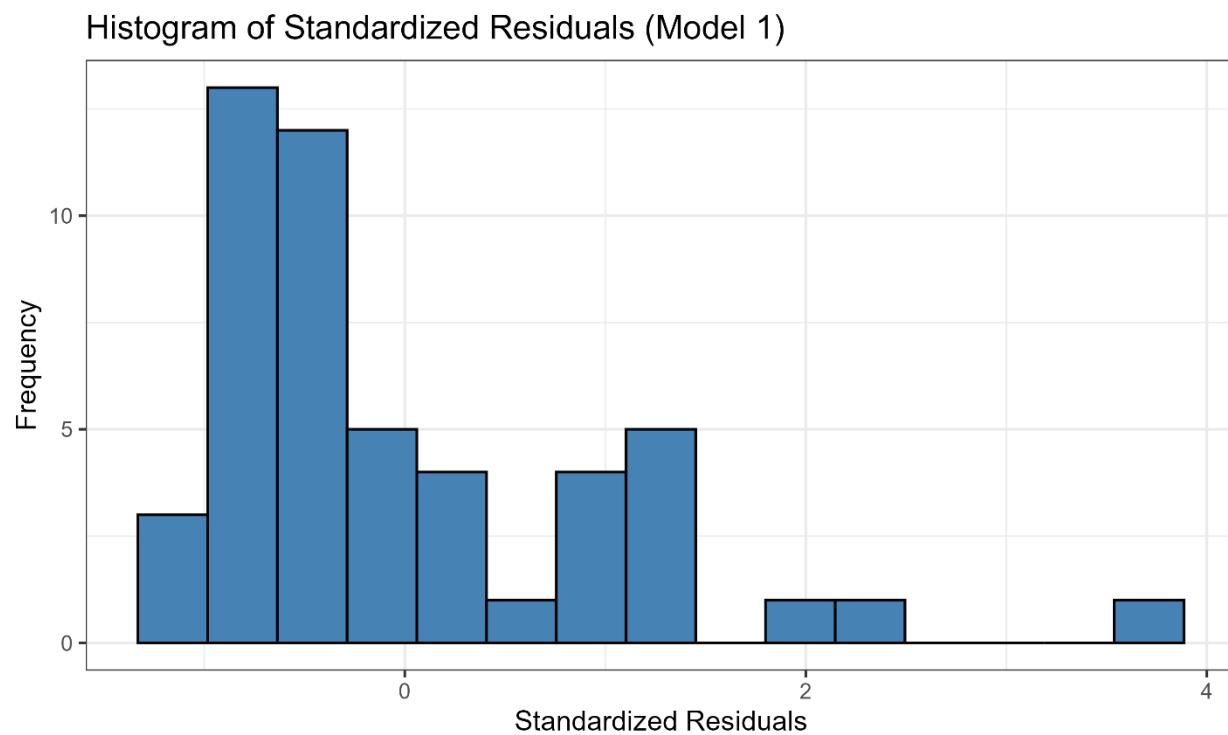
Figure A5. Scatterplot of Smokers vs HouseholdIncome*Figure A6. Histogram of Standardized Residuals*

Figure A7. Scatterplot of Standardized Residuals and Predicted Values

Standardized Residuals vs Predicted Values (Model 1)

