A Statistical Investigation into the Causes of Adult Obesity

Yuzhen Ke, Quintin Pittendrigh, Qihui Feng, Xiuxiu Tang

1. Introduction

Adult obesity represents one of the most pervasive issues in the developed world, such that, resulting from drastic changes over the last century, access to unhealthy food has expanded whereas time spent exercising has decreased, which, in combination with other factors, has negatively impacted health outcomes (Wright & Aronne, 2012). Notably, obesity leads to increased levels of heart disease and cancer rates, lowered lifespans as well as a worsened quality of life (Wright & Aronne, 2012). Yet, since the exact causes of increased adult obesity levels are diverse, we wish to address what factors contribute most significantly in augmenting obesity levels, including whether other unhealthy habits, poverty and education factor into predicting and explaining adult obesity rates (Wright & Aronne, 2012). Hence, the scientific question that we address is precisely what factors, excluding genetics, impact a population's adult obesity rate in the sense of both constructing a predictive model, while also maintaining a model to be used for inference. In doing so, we used the 2019 County Health Rankings dataset from California, where our results, which provide us with information on the impact of influential factors on adult obesity rates, are used towards providing recommendations regarding lifestyle changes to those suffering from obesity. Therefore, the importance of our study is of practical use, as we aim to provide guidelines to the public as to how to improve their health outcomes via adjusting to their lifestyles in order to prevent obesity. Our expectations regarding the most influential factors, purely based on common sense, are that physical health status, physical activity, access to healthy food, household income level, and food environment represent the factors of utmost importance in predicting adult obesity (Wright & Aronne, 2012).

2. Methods

2.1 Description of Data

The data used in this study originates from County Health Rankings, where we decided to use the 2019 health rankings for the state of California, a file that includes data on 58 California counties across several variables (County Health Rankings, 2019). An issue arises in that a small number of counties have missing data, ergo, if a county was missing many of the variables used in our study, we omitted that county from the study. We do not expect any outliers in this dataset initially, yet, we run multiple diagnostics, such as applying Cook's Distance, DFFITS, and others, to analyze the leverage that any potential outlier may have on the dataset. Provided that an observation is indeed an outlier, we may also omit the county from use in our study.

2.2 Methodology

Our study runs a multiple regression analysis with Adult Obesity Rate as the response variable, and Poor or Fair Health, Poor Physical Health Days, Poor Mental Health Days, Adult Smoking, Food Environment Index, Physical Inactivity, Access to Exercise Opportunities, Excessive Drinking, Primary Care Physician, Dentists, Mental Health Providers, Some College, Unemployment, Low Income, Long Commute & Driving Alone, Diabetes Prevalence, Food Insecure, Insufficient Sleep, Median Household Income, and Limited Access to Healthy Foods as explanatory variables. In translating the scientific question previously posed into a statistical one, we wish to verify which factors explain a particular county's obesity rate for both prediction and inference. As such, the goal is to initially provide a predictive model, which could be easily done by including as many regressors as the dataset provides, however, we would also like to construct a model that can be used for inference in further investigations into the causes of the obesity epidemic impacting California. Furthermore, we begin our study by creating a predictive model aimed at estimating future obesity rates, while then

continuing with a model to be used for inference in interpreting obesity trends, all of which is based on multiple regression analysis.

As mentioned, in constructing a multiple regression model for inference, we must initially address the issue of multicollinearity, since several regressors in the model are likely to be highly correlated with other regressors. Therefore, we will initially investigate the correlation matrix of variables in order to check whether multicollinearity occurs. If high correlations among the selected variables exist, such an issue may be addressed by excluding certain regressors, which are correlated with other regressors, from our model. We use the best model selection procedure to choose a model that maximizes the Adjusted R-Square and minimizes AIC/BIC in order to resolve which regressors to exclude, while also considering the Cp criterion. Provided the selected model is correct, bias will be nonexistent, such that the expected value of Cp is approximately equal to (or less than) p, the number of predictors plus the intercept. Based on each of these criterions, we may have up to four "best" models, however, if all methods coincide in determining the foremost model, we will select such a model. Otherwise, we will use BIC as our criterion to determine which model is most useful for inference, such that BIC uses a maximum likelihood procedure to penalize additional factors being included in the model. In contrast to AIC, BIC penalizes the inclusion of extra regressors in a logarithmic manner as opposed to a linear manner, and as such, we expect to obtain a simpler model to be used for inference.

3. Results

3.1 Descriptive Analysis of Variables

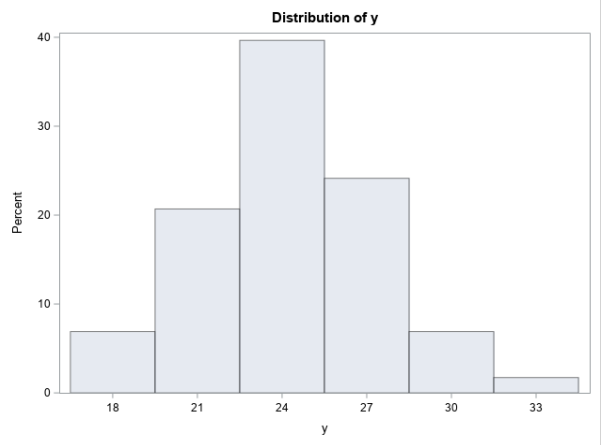
Our study involves 21 continuous variables, including the response variable of "adult obesity" and 20 explanatory variables, as well as a sample size of 58 (County Health Rankings, 2019), where detailed information about each variable is provided in Table 2 of the Appendix. Table 1 presents the basic statistical measurements of the response variable (obesity), where obesity is measured as a

percentage of adults that report BMI >= 30, with a mean value of 24.19 and a standard deviation of 3.43. The distribution of obesity is normal based on Figure 1.

Table 1. Basic statistical measures of Y(obesity)

Basic Statistical Measures			
Location		Variability	
Mean	24.18966	Std Deviation	3.43080
Median	24.50000	Variance	11.77042
Mode	25.00000	Range	17.00000
		Interquartile Range	5.00000

Figure 1. Distribution of y



3.2 Correlation

Following an initial inquiry into our response variable, we analyze the correlations of the 20 predictor variables, where many are intrinsically related, such as Poor or Fair Health (PH) with Poor Physical Health Days (poorPH), as well as Low Income (low_income) with Median Household Income (median_income). If correlation between regressors occurs, our interpretation of how changes in a given explanatory variable impact the response variable, while keeping all else constant, may be affected. Furthermore, highly correlated predictor variables contribute to large standard errors, and as such, the test for multicollinearity, specifically for pairwise correlation between highly correlated variables, is essential in our data analysis. The results of pairwise correlations are presented in Table 3 of the Appendix, such that this table provides the correlation of explanatory variables. Specifically, in analyzing the table, highly correlated explanatory variables, defined as having an absolute value of

correlation $r(x_i, x_j)$ $(i \neq j)$ greater than 0.95, did not occur; however, we decided to further investigate potential explanatory variables having an the absolute value of correlation greater than 0.80. Therefore, we have following correlations that require attention: r(PH, poorPH) = 0.83695, r(poorPH, poorMH) = 0.85463, r(poorPH, smoking) = 0.87227, r(poorPH, low_income) = -0.80237, r(poorPH, median_income) = -0.83301, r(poorMH, smoking) = 0.86934, r(poorMH, low_income) = -0.84879, r(poorMH, median_income) = -0.87728, r(smoking, low_income) = -0.80583, r(smoking, median_income) = -0.83710, r(FEI, food_insecure) = -0.84931, r(FEI, lmt_HF) = -0.80178, r(low_income, median_income) = 0.90703. Since some of these regressors are sufficiently correlated, we must consider only certain explanatory variables that are more representative or significant in the model, as well as in selecting simpler or "best" models in SAS. As such, we retain the models which have better representations of differing regressors and ignore those containing highly correlated regressors. Ergo, such an analysis of correlation will provide supporting evidence for our best-model selection step.

3.3 Best-Model Selection

Following our analysis of correlations between the regressors, we shift our focus to the best model selection method in order to find an appropriate model for inference. The selection criterions included in our study are Adjusted R-square, C(p), AIC and BIC, where we wish to maximize Adjusted R-square, and to minimize AIC and BIC. With regards to the C(p) criterion, a reasonable model should have a C(p) value that is equal to or below p, the number of predictors in the model plus one (which accounts for the intercept), which would suggest that the selected model has a small bias. Results are presented in Table 4 (below).

Table 4. Best Model Selection Results

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	BIC	Variables in Model
1	0.6160	0.6092	26.3305	90.4769	91.2104	inactivity
1	0.4630	0.4534	58.3522	109.9352	109.5321	PH
2	0.7268	0.7169	5.1491	72.7287	74.8176	poorPH inactivity
2	0.7091	0.6985	8.8629	76.3804	78.0937	PH inactivity
3	0.7422	0.7279	3.9354	71.3718	73.9633	poorPH poorMH inactivity
3	0.7412	0.7268	4.1449	71.5967	74.1562	poorPH inactivity ins_sleep
4	0.7563	0.7380	2.9776	70.1006	73.4402	poorPH poorMH smoking inactivity
4	0.7533	0.7347	3.6099	70.8155	74.0230	poorPH inactivity long_commute ins_sleep
5	0.7671	0.7447	2.7192	69.4725	73.6859	poorPH inactivity long_commute ins_sleep lmt_HF
5	0.7664	0.7439	2.8706	69.6524	73.8243	poorPH poorMH inactivity long_commute ins_sleep
6	0.7747	0.7482	3.1424	69.5642	74.6825	poorPH poorMH inactivity long_commute ins_sleep lmt_HF
6	0.7743	0.7477	3.2213	69.6613	74.7527	poorPH inactivity low_income long_commute ins_sleep lmt_HF
7	0.7848	0.7547	3.0164	68.8877	75.3361	poorPH FEI inactivity low_income long_commute food_insecure ins_sleep
7	0.7824	0.7520	3.5189	69.5317	75.7684	poorPH poorMH inactivity low_income long_commute food_insecure ins_sleep
8	0.7913	0.7573	3.6542	69.1055	76.8048	poorPH FEI inactivity low_income long_commute food_insecure ins_sleep lmt_HF
8	0.7897	0.7553	4.0058	69.5708	77.0927	poorPH poorMH FEI inactivity low_income long_commute food_insecure ins_sleep
9	0.7953	0.7569	4.8365	70.0089	78.8380	poorPH FEI inactivity exercise_opp low_income long_commute food_insecure ins_sleep lmt_HF
9	0.7952	0.7568	4.8474	70.0236	78.8463	poorPH poorMH FEI inactivity low_income long_commute food_insecure ins_sleep median_income
10	0.8015	0.7592	5.5376	70.2229	80.6279	poorPH poorMH FEI inactivity exercise_opp dentists long_commute food_insecure ins_sleep lmt_HF
10	0.8006	0.7582	5.7207	70.4780	80.7579	poorPH poorMH FEI inactivity exercise_opp physicians low_income food_insecure median_income lmt_HF
11	0.8084	0.7626	6.0816	70.1533	82.4828	poorPH poorMH FEI inactivity exercise_opp physicians dentists long_commute food_insecure ins_sleep lmt_HF
11	0.8072	0.7611	6.3317	70.5142	82.6442	poorPH poorMH FEI inactivity exercise_opp low_income long_commute food_insecure ins_sleep median_income lmt_HF
12	0.8122	0.7621	7.2997	71.0108	84.9297	poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists long_commute food_insecure ins_sleep lmt_HF
12	0.8118	0.7616	7.3807	71.1302	84.9760	poorPH poorMH FEI inactivity exercise_opp physicians low_income long_commute food_insecure ins_sleep median_income lmt_HF
13	0.8189	0.7654	7.8900	70.8920	87.1896	poorPH poorMH FEI inactivity exercise_opp physicians dentists low_income long_commute food_insecure ins_sleep median_income lmt_HF
13	0.8154	0.7609	8.6133	71.9888	87.5427	poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists low_income food_insecure ins_sleep median_income lmt_HF
14	0.8208	0.7625	9.4835	72.2664	90.0877	poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists low_income long_commute food_insecure ins_sleep median_income lmt_HF
14	0.8192	0.7603	9.8241	72.7911	90.2244	poorPH poorMH FEI inactivity exercise_opp physicians dentists college low_income long_commute food_insecure ins_sleep median_income lmt_HF
15	0.8213	0.7575	11.3892	74.1203	93.1526	poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists low_income long_commute food_insecure ins_sleep median_income lmt_HF
15	0.8211	0.7572	11.4334	74.1888	93.1665	poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists unemployment low_income long_commute food_insecure ins_sleep median_income lmt_HF
16	0.8220	0.7525	13.2432	75.8034	96.2214	PH poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists unemployment low_income long_commute food_insecure ins_sleep median_income lmt_HF
16	0.8217	0.7521	13.3123	76.0010	96.2373	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists low_income long_commute food_insecure ins_sleep median_income lmt_HF
17	0.8228	0.7475	15.0687	77.6210	99.3112	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists unemployment low_income long_commute food_insecure ins_sleep median_income lmt_HF
17	0.8221	0.7465	15.2182	77.8545	99.3325	PH poorPH poorMH smoking FEI inactivity exercise_opp physicians dentists college unemployment low_income long_commute food_insecure ins_sleep median_income lmt_HF
18	0.8231	0.7415	17.0096	79.5285	102.4373	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists unemployment low_income long_commute diabetes food_insecure ins_sleep median_income lmt_HF
18	0.8229	0.7411	17.0619	79.6103	102.4401	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists college unemployment low_income long_commute food_insecure ins_sleep median_income lmt_HF
19	0.8231	0.7347	19.0037	81.5192	105.5718	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists college unemployment low_income long_commute diabetes food_insecure ins_sleep median_income lmt_HF
19	0.8231	0.7347	19.0064	81.5234	105.5717	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists MHprovider unemployment low_income long_commute diabetes food_insecure ins_sleep median_income lmt_HF
20	0.8231	0.7276	21.0000	83.5134	108.7070	PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking physicians dentists MHprovider college unemployment low_income long_commute diabetes food_insecure ins_sleep median_income lmt_HF

In accordance with the above table, our initial model, Model 1 (Obesity = poorPH + poorMH + FEI + inactivity + exercise_opp + physicians + dentists + low_income + long_commute + food_insecure + ins_sleep + median_income + lmt_HF), was selected based on the criterion of adjusted R-square. This model contains 13 predictors and maximizes Adjusted R-square value, equal to 0.7654, in comparison to all other models.

The second model, Model 2 (Obesity = poorPH + poorMH + inactivity), containing 3 predictor variables and a C(p) value equal 3.9354, which is closest to p, was selected based on the C(p) criterion. The model includes an adjusted R-square equal to 0.7279, an AIC value equal to 71.3718, and a BIC value equal to 73.9633.

Furthermore, Model 3 (Obesity = poorPH FEI + inactivity + low_income + long_commute + food_insecure + ins_sleep) was selected based on the criterion of minimizing AIC. The model contains 6 predictor variables, and has the smallest AIC value of all models, namely, AIC equals 68.8878. Also, the model has an adjusted R-square equal to 0.7547, a C(p) value equal to 3.0164, and a BIC value equal to 75.3361.

Finally, Model 4 (Obesity = poorPH + poorMH + smoking + inactivity) was selected based on the criterion of minimizing BIC, where the model contains 4 predictor variables and has the smallest BIC value, namely, 73.4402. Furthermore, the model has an adjusted R-square equal to 0.7380, a C(p) value equal to 2.9776, and an AIC value equal to 70.1006.

Following the selection of the above models, we compared the results of each model, as is summarized in Table 9. Generally, Model 2 exhibits a better performance with respect to the four selection criterions, and as such, this was selected as our model to use in inference. While its adjusted R-square is the smallest among the four models, the difference between the value achieved by this model and the largest adjusted R-square (0.7654) is only 0.0375, which is an insignificant difference that can be ignored. In terms of the C(p) criterion, Model 2 is preferred as its C(p) value is closest to its p value, while the C(p) values of other models differ significantly from their respective p values. The difference in AIC values between Model 2 and Model 3, the model with the smallest AIC value of 68.8877, is the negligible amount of 2.4841. Furthemore, with respect to BIC, Model 2 has the second smallest BIC value among the four models, where the difference between Model 4's BIC value and Model 2's BIC value is a mere 0.5231. Therefore, when contrasting the four models, we found that Model 2 performed most satisfactorily towards inference in analyzing the causes of adult obesity.

Table 9. Comparisons of Models

	adjusted R-square	C(p)	AIC	BIC
Model 1	0.7654	7.8900	70.8920	87.1896
Model 2	0.7279	3.9354	71.3718	73.9633
Model 3	0.7547	3.0164	68.8877	75.3361

Model 4	0.7380	2.9776	70.1006	73.4402

After conducting regression analysis on Model 2 (Obesity = poorPH + poorMH + inactivity), we found that poorMH was not a significant predictor variable in this model (P-value = 0.0785 >0.05). As a result, we decided to delete it from Model 2, such that the final model only contains poorPH and inactivity.

3.4 Final Model

Following the best-model selection step, the final model selected includes 2 explanatory variables: poorPH and inactivity. Based on our output, as provided in Figure 2 below. Analysis of Final Model (before deleting outliers), we can see that the R-Square and adjusted R-square values are 0.7268 and 0.7169, respectively, which suggests that the 2 predictor variables selected, poorPH and inactivity, explain most of the total variation in obesity rates. Also, from the parameter estimates table in Figure 2, the p-values of poorPH and inactivity are both smaller than 0.0001; the standard errors of poorPH and inactivity are 0.66510 and 0.09414, respectively. Furthermore, the VIF (variance inflation factor) value is 1.32567, therefore, our final model is appropriate for future inference and prediction. According to Figure 3. Fit Diagnostic for y (final model before deleting outliers), found in the Appendix, the residual plot shows that these points are scattered randomly, however, there could be potential outliers; in the Leverage vs. RStudent plot, there are 4 points which are on the right of the cut-off and possibly have large influence in the model. The Cook's Distance plot shows us that there are 7 cases that have greater influence than other cases and thus need further investigation. Both the QQ-plot and the distribution of residual plot suggests that the residuals are reasonably normally distributed. As such, the investigation for x and y outliers and influential cases are applied later for further analysis.

Figure 2. Analysis of Final Model (before deleting outliers)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	487.64803	243.82402	73.17	<.0001
Error	55	183.26576	3.33210		
Corrected Total	57	670.91379			

Root MSE	1.82541	R-Square	0.7268
Dependent Mean	24.18966	Adj R-Sq	0.7169
Coeff Var	7.54622		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	1	-0.20630	2.21653	-0.09	0.9262	33938	0.02886	0
poorPH	1	3.14148	0.66510	4.72	<.0001	308.52392	74.33824	1.32567
inactivity	1	0.69023	0.09414	7.33	<.0001	179.12411	179.12411	1.32567

Table 10. Output for Identifying x and y Outliers and Influential Cases in the Appendix provides us the values calculated based on different tests for each case. Henceforth, we use Hat Matrix Diagonals to identify outlying x observations in the final model and look for large h_{ii} values which are greater than double the mean leverage value (2p/n = 0.1034). The cases for possible x outliers are as follows: # 21, 41, 44, and 39. Based on the studentized deleted residual values (Student Residual column in Table 6), we identify significant cases if their absolute values are greater than 2, namely, these are # 40, 52, 54. Furthermore, there are three different rules to identifying influential cases: DFBETAS, DFFITS, and Cook's Distance. For DFBETAS, if the absolute value is greater than 2/sqrt(n), or 0.2626, it may be considered significant, such that the significant cases include: #7, 12, 23, 41, 44, 52, 54. With regards to DFFITS, if the absolute value is greater than 2sqrt(p/n), or 0.4548, then it can be considered to have greater influence in contrast to other cases. The influential cases are as follows: #12, 23, 40, 41, 44, 52, 54. Figure 4. Studentized Residuals and Cook's D for y in the Appendix provides us a visual plot for identifying influential cases under Cook's Distance measure, and these include #12, 23, 40, 41, 44, 52, 54 according to the plot.

Therefore, given all influential cases, x outliers, and y outliers, we decided to delete cases # 40, 41, 44, 52, 54, since they represent both outliers and influential cases. After rerunning the basic

analysis for the final model following the deletion of outliers and influential cases, we obtain Figure 5. Analysis of Final Model (after deleting outliers) as shown below. Compared to Figure 2. Analysis of Final Model (before deleting outliers), the R-square value increases from 0.7268 to 0.7583, while the adjusted R-square is augmented from 0.7179 to 0.7487. The VIF values are decreased from 1.32567 to 1.22991, and both standard error values for poorPH and inactivity decrease. In Figure 6. Fit Diagnostic for y (final model after deleting outliers), the residual plot shows that the residual points are randomly distributed, and that the QQ-plot and distribution of residuals are both normal. Ergo, the final model, after deleting outliers, is more precise as it exhibits normally distributed residuals.

Figure 5. Analysis of Final Model (after deleting outliers)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	388.35474	194.17737	78.45	<.0001
Error	50	123.75847	2.47517		
Corrected Total	52	512.11321			

Root MSE	1.57327	R-Square	0.7583
Dependent Mean	24.18868	Adj R-Sq	0.7487
Coeff Var	6.50415		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	1	-1.56367	2.18518	-0.72	0.4776	31010	1.26743	0
poorPH	1	3.41049	0.61469	5.55	<.0001	240.55649	76.19623	1.22991
inactivity	1	0.70745	0.09155	7.73	<.0001	147.79825	147.79825	1.22991

4. Discussion

Our final model includes only two predictor variables, suggesting that most of the total variation in the response variable, obesity rates, is explained by poor physical health and smoking. Hence, we conclude that the two main causes of adult obesity are related to poor physical health and smoking, results which are intuitive as both factors directly influence physical status, a direct cause of obesity (Wright & Aronne, 2012).

In the process of data collection, it would be recommended to also collect information regarding other explanatory variables, such as genetics, lifestyle choices, dietary habits, daily calorie intake, as well as decreased energy expenditure. Such variables are likely to be significant factors that are strongly associated with adult obesity. Another weakness of this study is that the sample size is relatively small; we only obtained 58 observations originating from California. In the future, we may increase the sample size by using data from the entirety of the US or several countries to get a more precise model for inference and prediction.

References

County Health Rankings (2019) *2019 County Health Rankings Report:California*. [2019 California Data]. https://www.countyhealthrankings.org/app/california/2020/downloads

Wright, S. M., & Aronne, L. J. (2012). *Causes of obesity. Abdominal Radiology*. 37(5), 730-732.

Appendix

Table 2. Variables and their descriptions

Measure	SAS code	Description
Adult obesity	y	Percentage of adults that report BMI >= 30

Poor or fair health	PH	Percentage of adults that report fair or poor health
Poor physical health days	poorPH	Average number of reported physically unhealthy days per month
Poor mental health days	poorMH	Average number of reported mentally unhealthy days per month
Adult smoking	smoking	Percentage of adults that reported currently smoking
Food environment index	FEI	Indicator of access to healthy foods - 0 is worst, 10 is best
Physical inactivity	inactivity	Percentage of adults that report no leisure-time physical activity
Access to exercise opportunities	exercise_opp	Percentage of the population with access to places for physical activity
Excessive drinking	ex_driking	Percentage of adults that report excessive drinking
Primary care physicians	physicians	Primary Care Physicians per 100,000 population
Dentists	dentists	Dentists per 100,000 population
Mental health providers	Mhprovider	Mental Health Providers per 100,000 population
Some college	college	Percentage of adults age 25-44 with some post-secondary education
Unemployment	unemployment	Percentage of population ages 16+ unemployed and looking for work
Low income	low_income	20th percentile of median household income

Long commute & driving alone	long_commute	The percentage that workers commute in their car alone and drive more than 30 minutes
Diabetes prevalence	diabetes	Crude Percentage of Adults aged 20+ years that have diabetes in 2015
Food Insecure	food_insecure	Percentage of population with food insecurity in 2016
Insufficient sleep	ins_sleep	Percentage of population that report insufficient sleep in 2016
Median household income	median_income	Median household income
Limited access to healthy foods	lmt_HF	Percentage of population with limited access to healthy foods in 2015

Table 3. Correlation of Explanatory Variables

	PH	poorPH	poorMH	smoking	FEI	inactivity	exercise_opp	ex_driking	physicians	dentists	MHprovider	college	unemployment	low_income	long_commute	diabetes	food_insecure	ins_sleep	median_income	lmt_HF
PH	1.00000	0.83695 <.0001	0.53176 <.0001	0.63130 <.0001	-0.16104 0.2272	0.53975 <.0001	-0.37405 0.0038	-0.38736 0.0027	-0.58052 <.0001	-0.57548 <.0001	-0.45445 0.0003	-0.72409 <.0001	0.74741 <.0001	-0.52348 <.0001	-0.33735 0.0096	0.03847 0.7744	0.09623 0.4724	-0.56560 <.0001	0.15126 0.2571
poorPH	0.83695 <.0001	1.00000	0.85463 <.0001	0.87227 <.0001	-0.50749 <.0001	0.49564 <.0001	-0.46794 0.0002	-0.34226 0.0085	-0.66109 <.0001	-0.66086 <.0001	-0.35246 0.0067	-0.73207 <.0001	0.64838 <.0001	-0.80237 <.0001	-0.54206 <.0001	0.18881 0.1558	0.45425 0.0003	-0.83301 <.0001	0.36420 0.0049
poorMH	0.53176 <.0001	0.85463 <.0001	1.00000	0.86934 <.0001	-0.68187 <.0001	-0.41936 0.0011	-0.21995 0.0971	-0.21995 0.0971	-0.59181 <.0001	-0.63594 <.0001	-0.21797 0.1002	-0.58617 <.0001	0.48294 0.0001	-0.84879 <.0001	-0.57853 <.0001	0.36225 0.0052	0.62152 <.0001	-0.87728 <.0001	0.48257 0.0001
smoking	0.63130 <.0001	0.87227 <.0001	0.86934 <.0001	1.00000	-0.69109 <.0001	0.47147 0.0002	-0.48611 0.0001	-0.16840 0.2064	-0.61230 <.0001	-0.60673 <.0001	-0.23242 0.0792	-0.67036 <.0001	0.47203 0.0002	-0.80583 <.0001	-0.63932 <.0001	0.27348 0.0378	0.64427 <.0001	-0.83710 <.0001	0.48092 0.0001
FEI	-0.16104 0.2272	-0.50749 <.0001	-0.68187 <.0001	-0.69109 <.0001	1.00000	-0.29191 0.0262	0.37689 0.0035	0.11796 0.3778	0.34291 0.0084	0.37406 0.0050	0.55395 <.0001	0.55395 <.0001	-0.56944 <.0001	0.56956 <.0001	0.43685 0.0006	-0.21825 0.2905	-0.76380 <.0001	0.71830 <.0001	-0.80178 <.0001	0.30328 0.0207
inactivity	0.53975 <.0001	0.49564 <.0001	0.40587 0.0016	0.47147 0.0002	-0.29191 0.0262	1.00000	-0.48679 0.0001	-0.32357 0.0132	-0.61607 <.0001	-0.50080 <.0001	-0.65025 <.0001	-0.63598 <.0001	0.51012 <.0001	-0.46021 0.0003	-0.09339 0.4856	0.48958 <.0001	0.17620 0.1858	-0.54450 <.0001	0.30328 0.0207
exercise_opp	-0.37405 0.0038	-0.46794 0.0002	-0.41936 0.0011	-0.48611 0.0001	0.37689 0.0035	-0.48679 0.0001	1.00000	0.15184 0.2552	0.43557 0.0006	0.36371 0.0050	0.36710 0.0046	0.60104 <.0001	-0.43238 0.0007	0.52033 <.0001	0.38213 0.0031	-0.27699 0.0353	-0.26680 0.0429	0.59557 <.0001	-0.39926 0.0019
ex_driking	-0.38736 0.0027	-0.34226 0.0085	-0.21995 0.0971	-0.16840 0.2064	0.11796 0.3778	-0.32357 0.0132	0.15184 0.2552	1.00000	0.33102 0.0111	0.37406 0.0038	0.34368 0.0083	0.29119 0.0266	-0.34771 0.0075	0.21563 0.1040	-0.04177 0.7555	-0.49204 <.0001	-0.07338 0.5841	0.25303 0.0553	-0.10520 0.4319
physicians	-0.58052 <.0001	-0.66109 <.0001	-0.59181 <.0001	-0.61230 <.0001	0.37451 0.0038	-0.61607 <.0001	0.43557 0.0006	0.33102 0.0111	1.00000	0.75647 <.0001	0.60338 <.0001	0.69501 <.0001	-0.56300 <.0001	0.52929 <.0001	0.22517 0.0892	-0.30524 0.0198	-0.20810 0.1170	0.66516 <.0001	-0.40963 0.0014
dentists	-0.57548 <.0001	-0.66086 <.0001	-0.65394 <.0001	-0.60673 <.0001	0.34291 0.0084	-0.50080 <.0001	0.36371 0.0050	0.37406 0.0038	0.75647 <.0001	1.00000	0.45736 0.0003	0.55395 <.0001	-0.52418 <.0001	0.46503 0.0002	0.20237 0.1277	-0.28341 0.0311	-0.16750 0.2088	0.62800 <.0001	-0.39105 0.0024
MHprovider	-0.45445 0.0003	-0.35246 0.0067	-0.21797 0.1002	-0.23242 0.0792	0.07115 0.5956	-0.65025 <.0001	0.36710 0.0046	0.34368 0.0083	0.60338 <.0001	0.45736 0.0003	1.00000	0.55393 <.0001	-0.46745 0.0002	0.20653 0.1198	-0.03216 0.8105	-0.20356 0.1254	0.11856 0.3754	0.40689 0.0015	-0.24912 0.0593
college	-0.72409 <.0001	-0.73207 <.0001	-0.58617 <.0001	-0.67036 <.0001	0.33288 0.0107	-0.63598 <.0001	0.60104 <.0001	0.29119 0.0266	0.69501 <.0001	0.55395 <.0001	0.55393 <.0001	1.00000	-0.56944 <.0001	0.56956 <.0001	0.43685 0.0006	-0.21825 0.0998	-0.14116 0.2905	0.69173 <.0001	-0.42848 0.0008
unemployment	0.74741 <.0001	0.64838 <.0001	0.48294 0.0001	0.47203 0.0002	-0.30113 0.0216	0.51012 <.0001	-0.43238 0.0007	-0.34771 0.0075	-0.56300 <.0001	-0.52418 <.0001	-0.46745 0.0002	-0.56944 <.0001	1.00000	-0.51881 <.0001	-0.38629 0.0027	0.16176 0.2251	0.23995 0.0696	-0.57139 <.0001	0.23866 0.0712
low_income	-0.52348 <.0001	-0.80237 <.0001	-0.84879 <.0001	-0.80583 <.0001	0.77828 <.0001	-0.46021 0.0003	0.52033 <.0001	0.21563 0.1040	0.52929 <.0001	0.46503 0.0002	0.20653 0.1198	0.56956 <.0001	-0.51881 <.0001	1.00000	0.62891 <.0001	-0.41622 0.0012	-0.76380 <.0001	0.90703 <.0001	-0.50020 <.0001
long_commute	-0.33735 0.0096	-0.54206 <.0001	-0.57853 <.0001	-0.63932 <.0001	0.64372 <.0001	-0.09339 0.4856	0.38213 0.0031	-0.04177 0.7555	0.22517 0.0892	0.20237 0.1277	-0.03218 0.8105	0.43685 0.0006	-0.38629 0.0027	0.62891 <.0001	1.00000	-0.10101 0.4506	-0.59366 <.0001	0.64828 <.0001	-0.47843 0.0001
diabetes	0.03847 0.7744	0.18881 0.1558	0.36225 0.0052	0.27348 0.0378	-0.41604 0.0012	0.48958 <.0001	-0.27699 0.0353	-0.49204 <.0001	-0.30524 0.0198	-0.28341 0.0311	-0.20356 0.1254	-0.21825 0.0998	0.16176 0.2251	-0.41622 0.0012	-0.10101 0.4506	1.00000	0.39228 0.0023	0.90703 <.0001	-0.50020 <.0001
food_insecure	0.09623 0.4724	0.45425 0.0003	0.62152 <.0001	0.64427 <.0001	-0.84931 0.1858	0.17620 0.1858	-0.26680 0.0429	-0.07338 0.5841	-0.20810 0.1170	-0.16750 0.2088	0.11856 0.3754	-0.14116 0.2905	0.23995 0.0696	-0.76380 <.0001	-0.59366 <.0001	0.39228 0.0023	1.00000	0.64828 <.0001	-0.42814 0.0008
ins_sleep	0.63396 <.0001	0.49888 <.0001	0.29576 0.0242	0.43144 0.0007	-0.05973 0.6561	0.57755 <.0001	-0.14686 0.2713	-0.26480 0.0446	-0.40290 0.0017	-0.35992 0.0055	-0.53486 <.0001	-0.52875 <.0001	0.39485 0.0022	-0.28416 0.0306	0.00126 0.9925	0.15763 0.2373	0.03344 0.8032	-0.42814 0.0008	0.37414 0.0038
median_income	-0.56560 <.0001	-0.83301 <.0001	-0.87728 <.0001	-0.83710 <.0001	0.71830 <.0001	-0.54450 <.0001	0.59557 <.0001	0.25303 0.0553	0.66516 <.0001	0.62800 <.0001	0.40689 0.0015	0.69173 <.0001	-0.57139 <.0001	0.90703 <.0001	0.64828 <.0001	-0.42814 0.0008	-0.59842 <.0001	1.00000	-0.47843 0.0001
lmt_HF	0.15126 0.2571	0.36420 0.0049	0.48257 0.0001	0.48092 0.0001	-0.80178 <.0001	0.30328 0.0207	-0.39926 0.0019	-0.10520 0.4319	-0.40963 0.0014	-0.39105 0.0024	-0.24912 0.0593	-0.42848 0.0008	0.23866 0.0712	-0.50020 <.0001	-0.47843 0.0001	0.26754 0.0423	0.37414 0.0038	-0.47843 0.0001	1.00000

Figure 3. Fit Diagnostic for y (final model before deleting outliers)

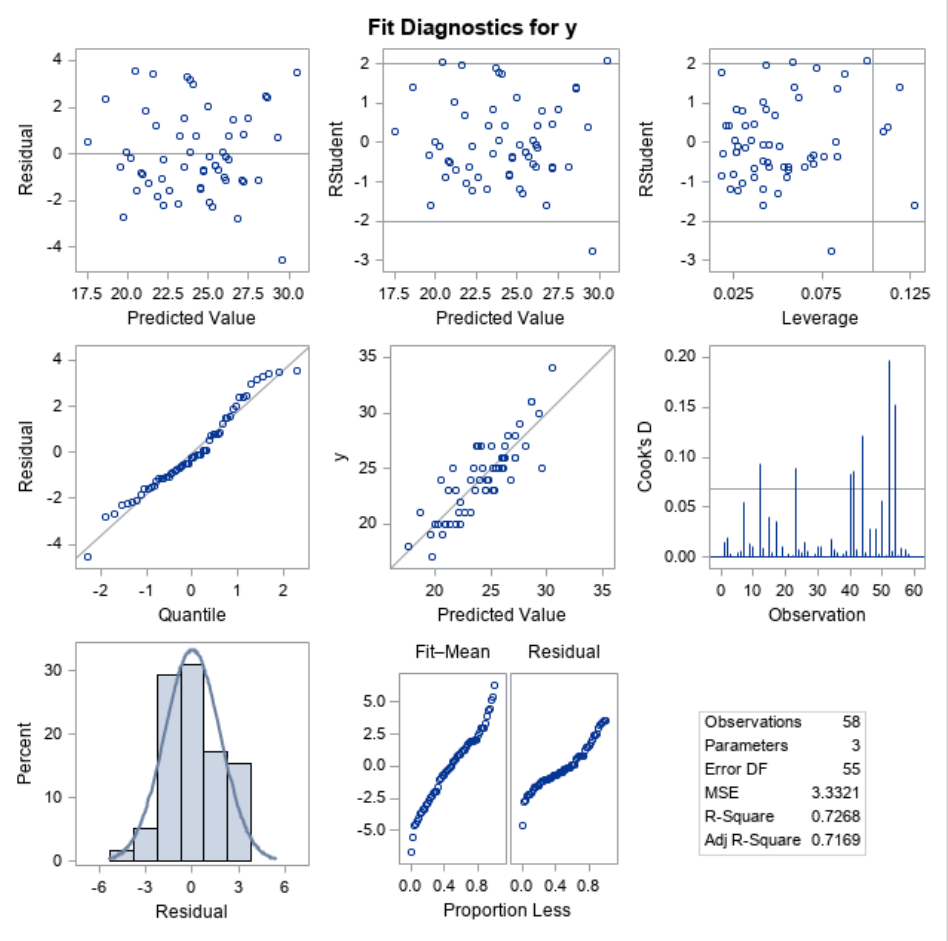


Table 10. Output for Identifying x and y Outliers and Influential Cases

Output Statistics														
												DFBETAS		
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	poorPH	inactivity
1	19	20.5759	0.4267	-1.5759	1.775	-0.888	0.015	-0.8862	0.0546	1.0703	-0.2130	-0.1863	0.1447	0.0157
2	23	25.0979	0.3721	-2.0979	1.787	-1.174	0.020	-1.1780	0.0415	1.0216	-0.2453	0.1004	-0.1859	0.1139
3	24	24.6553	0.5025	-0.6553	1.755	-0.373	0.004	-0.3705	0.0758	1.1346	-0.1061	-0.0353	0.0823	-0.0788
4	25	25.4739	0.2964	-0.4739	1.801	-0.263	0.001	-0.2609	0.0264	1.0811	-0.0429	0.0183	-0.0245	0.0066
5	25	25.6597	0.5273	-0.6597	1.748	-0.378	0.004	-0.3745	0.0834	1.1438	-0.1130	-0.0208	0.0793	-0.0931
6	28	26.5402	0.3170	1.4598	1.798	0.812	0.007	0.8095	0.0302	1.0507	0.1427	-0.0548	0.0041	0.0790
7	25	21.5803	0.3803	3.4197	1.785	1.915	0.056	1.9646	0.0434	0.8979	0.4185	0.3334	-0.2970	0.0327
8	25	25.4739	0.2964	-0.4739	1.801	-0.263	0.001	-0.2609	0.0264	1.0811	-0.0429	0.0183	-0.0245	0.0066
9	20	22.2086	0.3009	-2.2086	1.800	-1.227	0.014	-1.2325	0.0272	0.9994	-0.2060	-0.1415	0.0937	0.0248
10	29	27.4827	0.3793	1.5173	1.786	0.850	0.011	0.8476	0.0432	1.0613	0.1801	-0.1264	0.0990	0.0364
11	26	26.1022	0.3879	-0.1022	1.784	-0.057	0.000	-0.0568	0.0452	1.1064	-0.0124	0.0071	-0.0096	0.0036
12	27	24.0315	0.5416	2.9685	1.743	1.703	0.093	1.7336	0.0880	0.9848	0.5386	-0.1559	0.4394	-0.3919
13	27	28.1110	0.4659	-1.1110	1.765	-0.629	0.009	-0.6260	0.0651	1.1060	-0.1652	0.1286	-0.1211	-0.0038
14	24	23.9032	0.2935	0.0968	1.802	0.054	0.000	0.0532	0.0259	1.0845	0.0087	0.0034	-0.0048	0.0035
15	31	28.5490	0.4442	2.4510	1.771	1.384	0.040	1.3962	0.0592	1.0097	0.3503	-0.2065	0.0384	0.2349
16	24	24.7218	0.4775	-0.7218	1.762	-0.410	0.004	-0.4065	0.0684	1.1238	-0.1102	0.0412	-0.0912	0.0692
17	24	26.7925	0.3719	-2.7925	1.787	-1.563	0.035	-1.5838	0.0415	0.9620	-0.3296	0.2166	-0.2262	0.0155
18	26	25.9119	0.3402	0.0881	1.793	0.049	0.000	0.0487	0.0347	1.0945	0.0092	-0.0007	-0.0034	0.0066
19	21	23.1510	0.2752	-2.1510	1.805	-1.192	0.011	-1.1967	0.0227	0.9995	-0.1825	-0.0412	-0.0415	0.0896
20	26	26.1022	0.3879	-0.1022	1.784	-0.057	0.000	-0.0568	0.0452	1.1064	-0.0124	0.0071	-0.0096	0.0036
21	18	17.5009	0.6047	0.4991	1.722	0.290	0.003	0.2874	0.1097	1.1814	0.1009	0.0849	-0.0283	-0.0626
22	25	24.2173	0.2623	0.7827	1.806	0.433	0.001	0.4300	0.0206	1.0679	0.0624	0.0158	-0.0231	0.0205
23	27	23.7174	0.4888	3.2826	1.759	1.866	0.090	1.9109	0.0717	0.9353	0.5312	-0.1180	0.4078	-0.3924
24	26	27.1066	0.4204	-1.1066	1.776	-0.623	0.007	-0.6195	0.0530	1.0923	-0.1466	0.1020	-0.1122	0.0176
25	23	24.4696	0.2848	-1.4696	1.803	-0.815	0.006	-0.8125	0.0243	1.0442	-0.1283	0.0253	-0.0671	0.0482
26	23	21.1423	0.3718	1.8577	1.787	1.039	0.016	1.0403	0.0415	1.0386	0.2164	0.1178	0.0134	-0.1499
27	25	23.4652	0.2975	1.5348	1.801	0.852	0.007	0.8500	0.0266	1.0430	0.1404	0.0066	0.0610	-0.0793
28	22	22.2086	0.3009	-0.2086	1.800	-0.116	0.000	-0.1148	0.0272	1.0853	-0.0192	-0.0132	0.0087	0.0023
29	20	20.8901	0.3846	-0.8901	1.784	-0.499	0.004	-0.4954	0.0444	1.0907	-0.1068	-0.0902	0.0620	0.0179
30	20	21.8325	0.3141	-1.8325	1.798	-1.019	0.011	-1.0195	0.0296	1.0283	-0.1781	-0.1023	0.0134	0.0926
31	20	21.2661	0.4301	-1.2661	1.774	-0.714	0.010	-0.7105	0.0555	1.0879	-0.1723	-0.1406	0.1347	-0.0251
32	23	23.5271	0.2479	-0.5271	1.808	-0.291	0.001	-0.2890	0.0184	1.0715	-0.0396	-0.0141	0.0081	0.0013
33	26	26.2261	0.3220	-0.2261	1.797	-0.126	0.000	-0.1247	0.0311	1.0896	-0.0223	0.0051	0.0040	-0.0145
34	27	23.8413	0.2432	3.1587	1.809	1.746	0.018	1.7801	0.0178	0.9066	0.2393	0.0396	0.0157	-0.0403
35	25	25.9739	0.4829	-0.9739	1.760	-0.553	0.008	-0.5497	0.0700	1.1172	-0.1508	-0.0153	0.0948	-0.1254
36	26	27.1685	0.3472	-1.1685	1.792	-0.652	0.005	-0.6486	0.0362	1.0710	-0.1257	0.0789	-0.0514	-0.0397
37	20	20.1998	0.4110	-0.1998	1.779	-0.112	0.000	-0.1113	0.0507	1.1122	-0.0257	-0.0217	0.0111	0.0099
38	19	19.5715	0.4833	-0.5715	1.760	-0.325	0.003	-0.3220	0.0701	1.1297	-0.0884	-0.0811	0.0566	0.0169
39	30	29.3011	0.6119	0.6989	1.720	0.406	0.007	0.4033	0.1124	1.1797	0.1435	-0.0491	-0.0416	0.1294
40	24	20.4521	0.4423	3.5479	1.771	2.003	0.083	2.0617	0.0587	0.8943	0.5149	0.2610	0.0810	-0.4093
41	21	18.6291	0.6300	2.3709	1.713	1.384	0.086	1.3958	0.1191	1.0785	0.5133	0.4787	-0.4148	0.0051
42	21	22.0847	0.4327	-1.0847	1.773	-0.612	0.007	-0.6082	0.0562	1.0968	-0.1484	-0.0198	-0.0763	0.1222
43	20	19.9476	0.5261	0.0524	1.748	0.030	0.000	0.0297	0.0831	1.1523	0.0089	0.0080	-0.0072	0.0006
44	17	19.6999	0.6515	-2.6999	1.705	-1.583	0.122	-1.6059	0.1274	1.0527	-0.6136	-0.1617	-0.2787	0.5706
45	23	24.5315	0.2456	-1.5315	1.809	-0.847	0.004	-0.8445	0.0181	1.0345	-0.1146	-0.0086	0.0143	-0.0248
46	23	25.2836	0.4080	-2.2836	1.779	-1.284	0.029	-1.2913	0.0500	1.0152	-0.2961	-0.0514	0.1865	-0.2231
47	25	25.0979	0.3721	-0.0979	1.787	-0.055	0.000	-0.0543	0.0415	1.1022	-0.0113	0.0046	-0.0086	0.0052
48	27	24.9695	0.4528	2.0305	1.768	1.148	0.029	1.1517	0.0615	1.0468	0.2949	0.0775	-0.2107	0.2216
49	20	20.8281	0.3737	-0.8281	1.787	-0.463	0.003	-0.4602	0.0419	1.0899	-0.0962	-0.0645	0.0112	0.0578
50	31	28.6109	0.5286	2.3891	1.747	1.367	0.057	1.3785	0.0838	1.0396	0.4170	-0.1479	-0.1033	0.3613
51	27	26.2261	0.3220	0.7739	1.797	0.431	0.002	0.4275	0.0311	1.0795	0.0766	-0.0175	-0.0137	0.0496
52	25	29.5534	0.5166	-4.5534	1.751	-2.601	0.196	-2.7518	0.0801	0.7748	-0.8120	0.5218	-0.1078	-0.5643
53	25	26.1022	0.3879	-1.1022	1.784	-0.618	0.006	-0.6144	0.0452	1.0837	-0.1336	0.0766	-0.1040	0.0385
54	34	30.4958	0.5780	3.5042	1.731	2.024	0.152	2.0845	0.1003	0.9310	0.6959	-0.5736	0.3228	0.3132

55	24	23.2130	0.2694	0.7870	1.805	0.436	0.001	0.4327	0.0218	1.0690	0.0646	0.0327	-0.0280	0.0060
56	21	22.5847	0.3485	-1.5847	1.792	-0.884	0.010	-0.8826	0.0364	1.0504	-0.1716	-0.1147	0.1232	-0.0448
57	23	21.7706	0.4026	1.2294	1.780	0.691	0.008	0.6872	0.0486	1.0820	0.1554	0.0409	0.0602	-0.1248
58	28	27.1685	0.3472	0.8315	1.792	0.464	0.003	0.4606	0.0362	1.0834	0.0893	-0.0561	0.0365	0.0282

Figure 4. Studentized Residuals and Cook's D for y

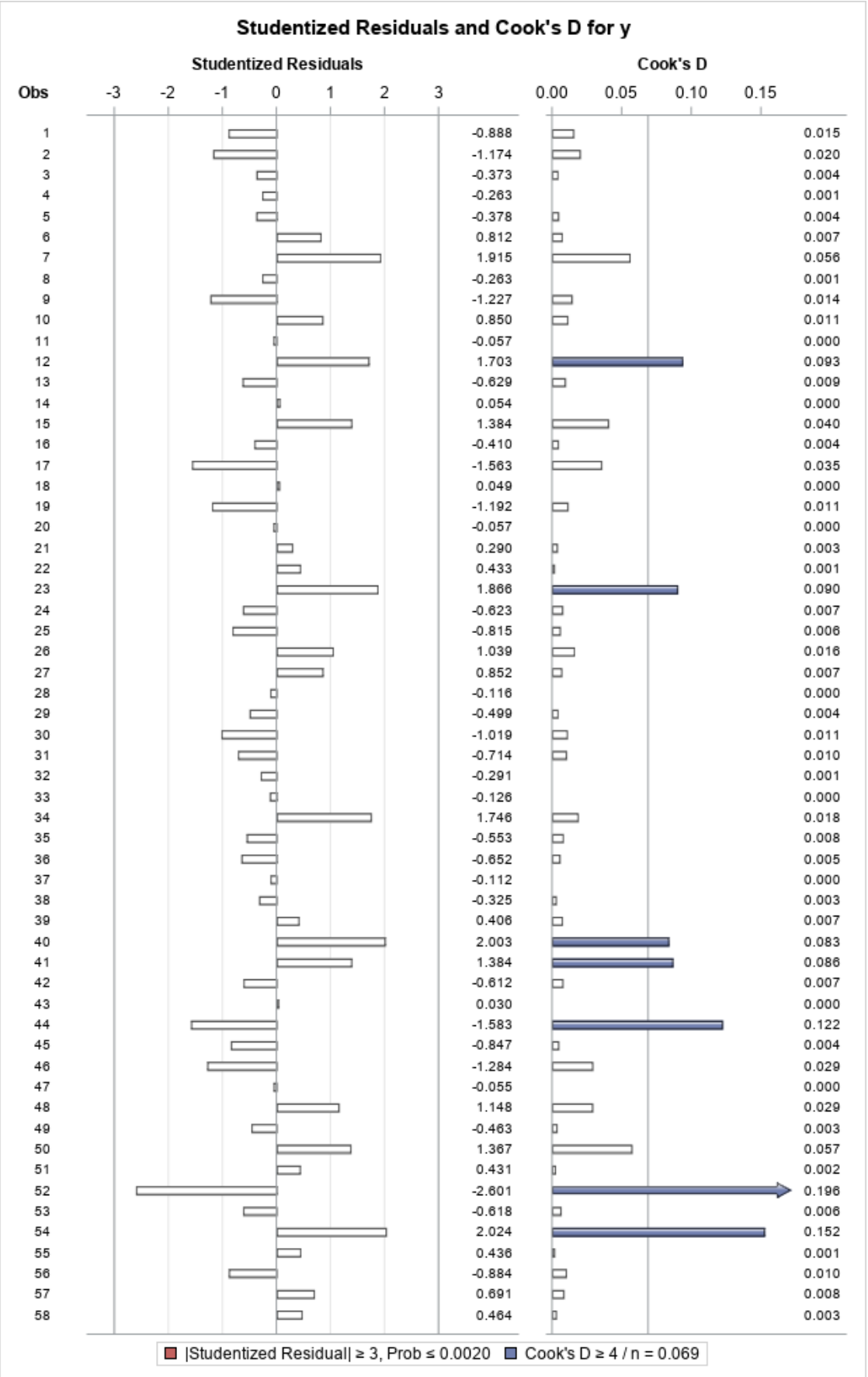
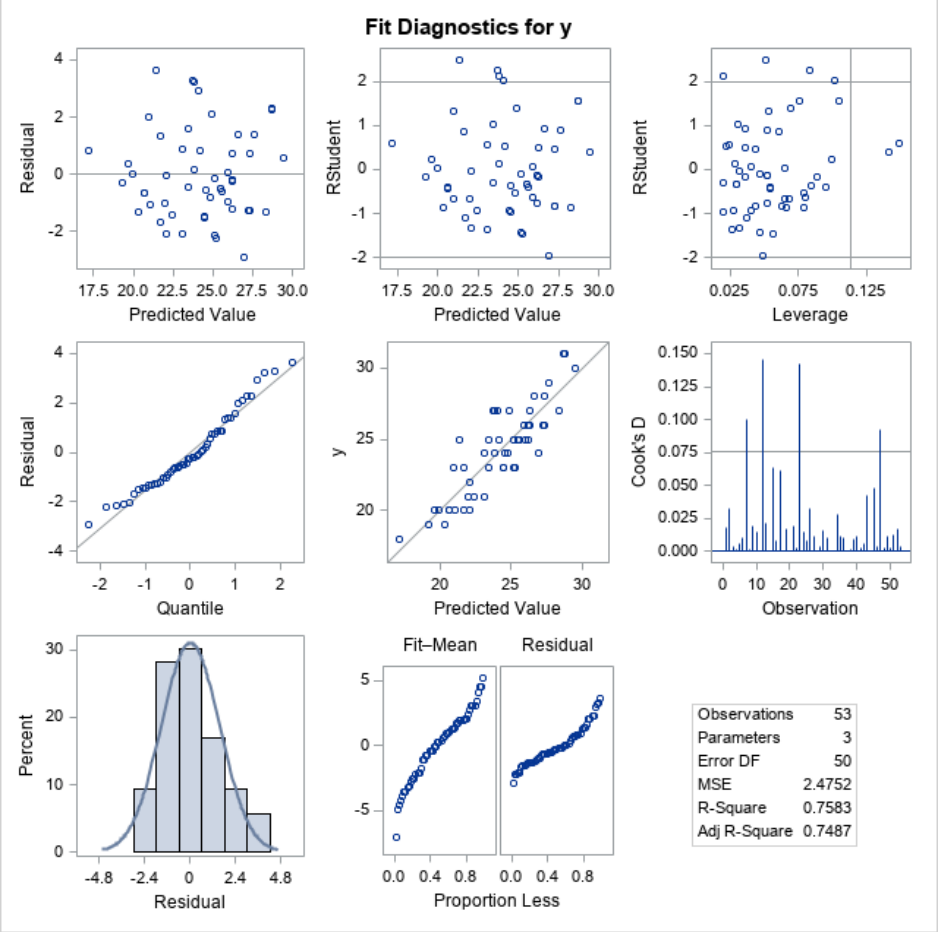


Figure 6. Fit Diagnostic for y (final model after deleting outliers)



7. SAS Codes

```
**read data;

data a1;

infile 'C:\Users\pmy\Desktop\2021Spring\STAT 525\Project\datanew2.txt' dlm='09'x;

input y PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking

physicians dentists MHprovider college unemployment low_income

long_commute diabetes food_insecure ins_sleep median_income lmt_HF;

**histogram of y;
```

```
proc univariate data=a1;

var y;

histogram y;

run;

**correlation btw predictor variables;

proc corr;

var PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking

physicians dentists MHprovider college unemployment low_income

long_commute diabetes food_insecure ins_sleep median_income lmt_HF;

run;

**model selection;

proc reg data=a1;

model y = PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking

physicians dentists MHprovider college unemployment low_income

long_commute diabetes food_insecure ins_sleep median_income lmt_HF /

selection=rsquare adjrsq cp aic bic best=2 vif;

run;

**outliers/;

**cp;

proc reg data=a1;
```

```
model y = poorPH poorMH inactivity/

r partial influence vif ss1 ss2;

run;

**2 variables (selected final model);

proc reg data=a1;

model y = poorPH inactivity /

r partial influence vif ss1 ss2;

run;

**remove outliers that are influential;

**read data;

data a2;

infile 'C:\Users\pmy\Desktop\2021Spring\STAT 525\Project\datanew3_nooutlier.txt' dlm='09'x;

input y PH poorPH poorMH smoking FEI inactivity exercise_opp ex_driking

physicians dentists MHprovider college unemployment low_income

long_commute diabetes food_insecure ins_sleep median_income lmt_HF;

**histogram of y;

proc univariate data=a2;

var y;

histogram y;

run;
```

```
**2 variables;

proc reg data=a2;

model y = poorPH inactivity /

r partial influence vif ss1 ss2;

run;
```