Assignment: Homework 3 Report

Name: Horace Fung

Instructor: Professor Simonoff

Course: Regression and Multivariate Data Analysis

Modelling the relationship between mental health and other health indicators in U.S. cities

This report builds on the first report's analysis of sleep deprivation as a predictor of poor mental health in American cities, where the best simple regression model suggested that sleep deprivation has a significant positive association with poor mental health. Both reports are motivating by the large number of people affected by mental health conditions in the United States—estimated to be around 20%— and the potential benefits to policy-making from understanding how various health indicators may improve mental health across the country.

The data is from 500 Cities Project, a collaboration between Centers for Disease Control and Prevention (CDC) and The Robert Wood Johnson Foundation. Additional information about the region of each city is from the U.S. Census Bureau. Lastly, information about the status of Medicaid expansion for each state is from the Kaiser Family Foundation, a non-profit that provides data on U.S. healthcare. The observations in the data are on an aggregate city level. For example, they show the percentage of total respondents with poor mental health and the percentage of total respondents with sleep deprivation. Even if a respondent answers yes to poor mental health and sleep deprivation, there is no way to link both variables to one individual. Hence, this is not a 1-to-1 analysis Therefore, the analysis can only examine the general association between the pervasiveness of poor mental health and the pervasiveness of other health indicators.

1. Definition of Variables:

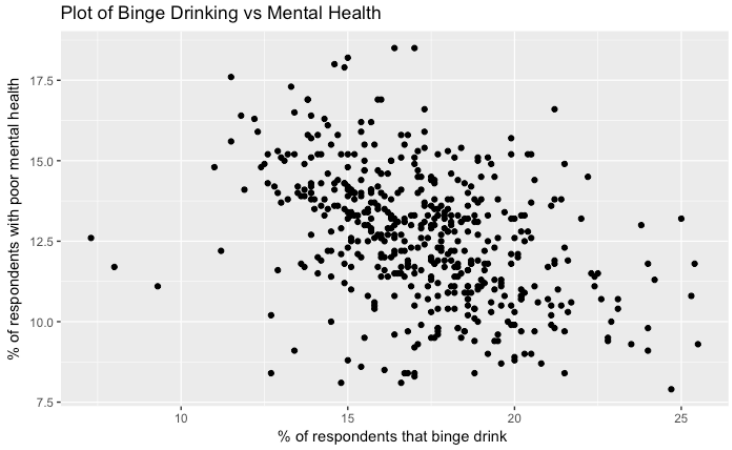
The initial data contains nine predictors and one response. An addition predictor on whether a city has implemented Medicaid expansion or not was added. The rationale will be explained in the next section. The definitions of all the variables are listed below:

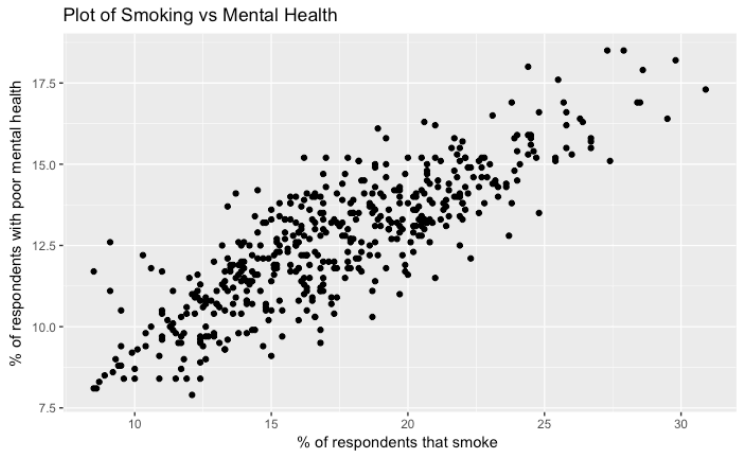
1. *Mental_Health:* Percentage of respondents who report 14 or more days during the past 30 days in which their mental health was not good.
2. *Binge_Drinking:* Percentage of respondents who report having five or more drinks (men) or four or more drinks (women) on an occasion in the past 30 days
3. *Smoking*: Percentage of respondents who report having smoked ≥ 100 cigarettes in their lifetime and currently smoke every day or some days.

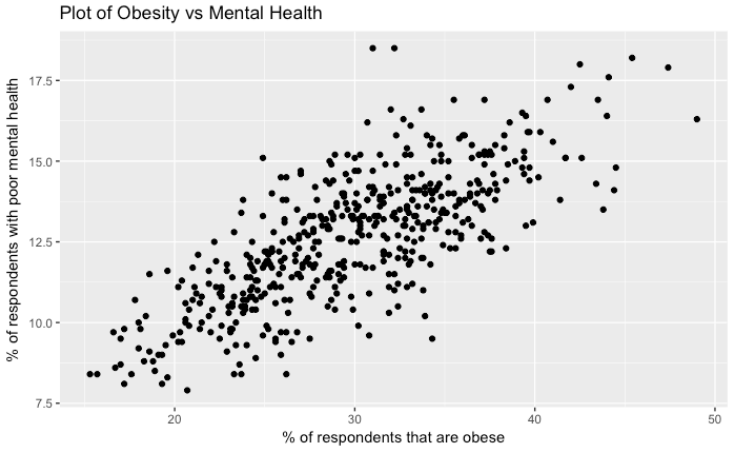
4. *Obesity*: Percentage of respondents who have a body mass index (BMI) ≥ 30.0 kg/m².
5. *Lack_of_Sleep*: Percentage of respondents aged who report getting less than 7 hours of sleep on average.
6. *Checkup*: Percentage of respondents who report having been to a doctor for a routine checkup in the previous year.
7. *Health_Insurance*: Percentage of respondents who report having no current health insurance coverage.
8. *Region*: The region of a city (Northeast, South, Midwest, West). This is one-hot encoded into three binary variables. The fourth variable is unnecessary as Northeast = South = Midwest = 0 implies the observation is in the West region.
9. *Medicaid*: Binary variable where 1 means a state has implemented Medicaid expansion before 2016, and 0 means a state has not.

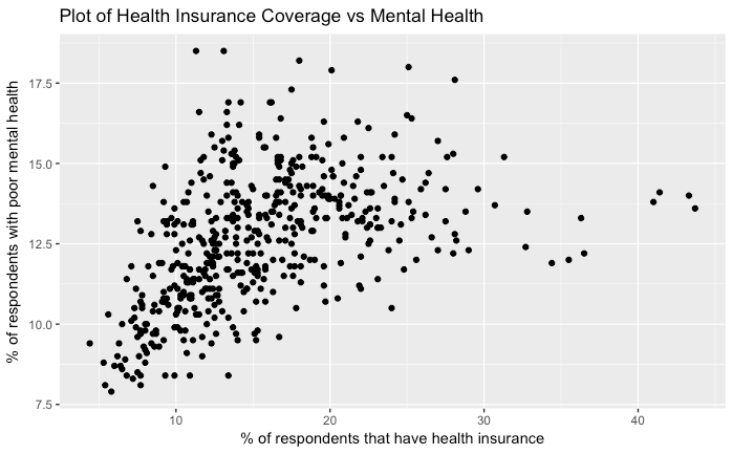
2. Identifying outliers and unusual observations:

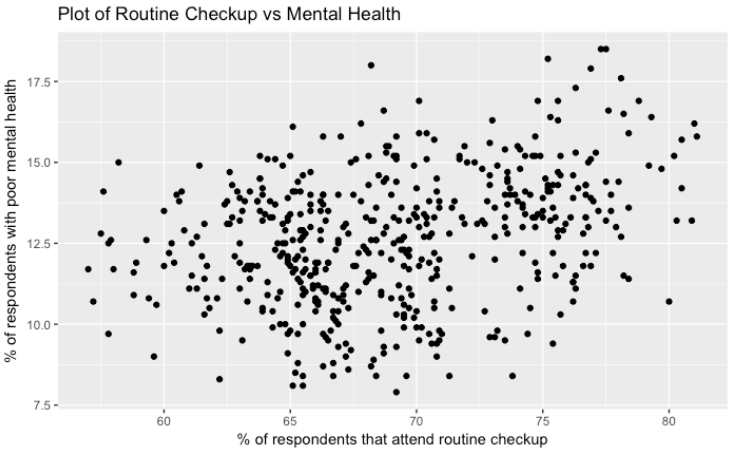
The plots of the initial predictors vs the response are shown below:

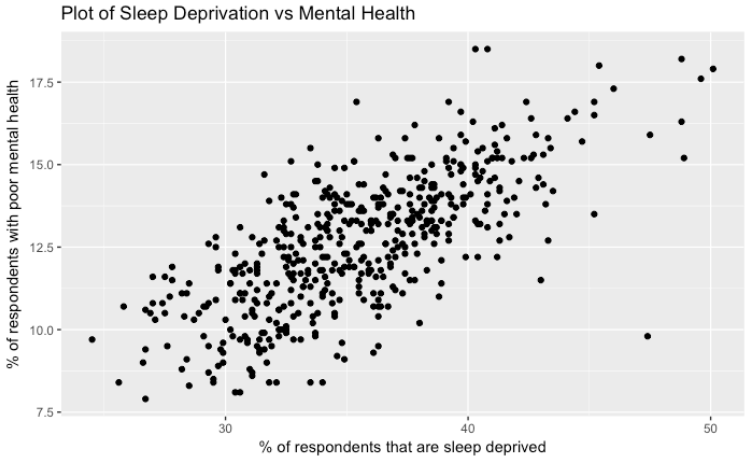


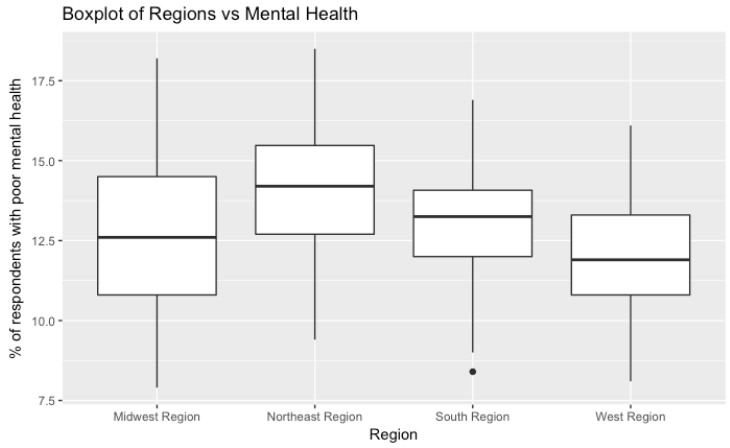




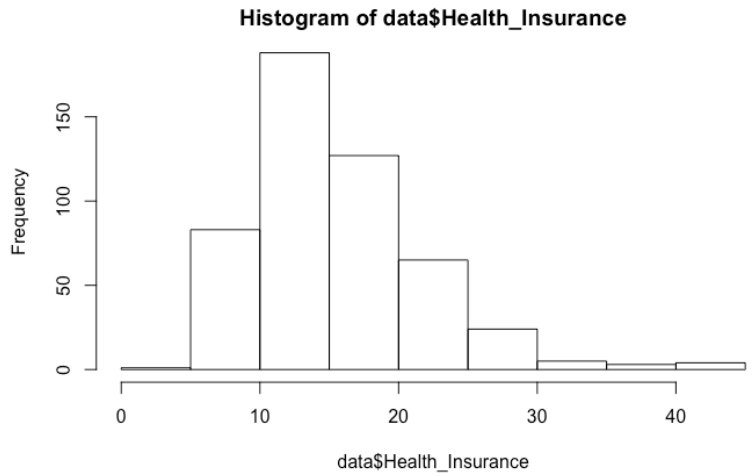




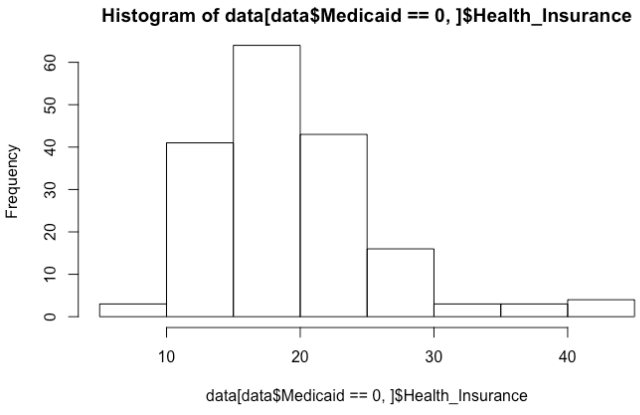


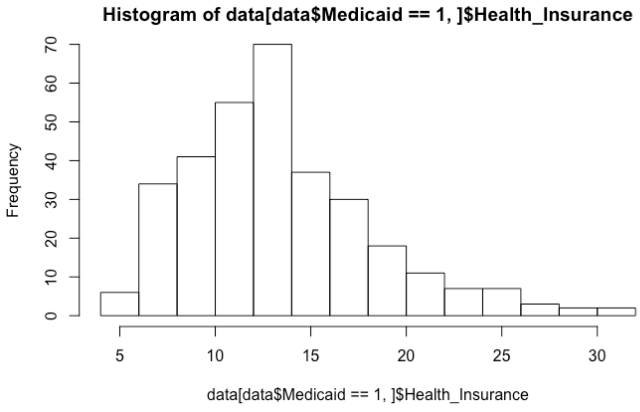


The most notable plots are binge drinking, routine checkups and health insurance coverage. The negative relationship in binge drinking is surprising since heavy drinking is often associated with mental illness, either as a method of coping or as a contributor to worse physical and mental health. The data, however, is self-reported, so perhaps, many respondents temporarily alleviated some symptoms of poor mental health through drinking but are not actually in better conditions. The opportunity to drink excessively once in a while may also be in fact associated with better mental wellbeing. There seems to be some potential leverage points to the left, and outliers when the response is less than 10% and the predictor is between 10-20%. In the routine checkups plot, the variance is high and there seems to be a weak positive relationship. This is reasonable as the typical medical checkup does not provide much for mental wellbeing and may have no real relationship with the response variable at all. In the health insurance plot, there is clearly non-constant variance. The observations with over 30% of respondents without health insurance appears to be leverage points. But even ignoring those points, the plot is still cone shaped. A histogram of health insurance coverage below shows a long right tail:



A semi-log model with y = log(Health_Insurance) can reduce heteroscedasticity, but there is no evidence to support that type of relationship. A semi-log model implies a multiplicative/additive relationship. In other words, in cities where the lack of health insurance is lower, an equal percentage point decrease in the lack of health insurance is associated with a higher percentage point decrease in poor mental health. Yet, one would suspect that cities which already have good health insurance coverage will benefit less from extra coverage. Another hypothesis is that this data contains two subgroups. Scanning through the data, it seems that many of the cities with lower coverage are in states that did not have Medicaid expansion implemented prior to 2016, which means the cost of insurance for low-income individuals tended to be higher. A Medicaid binary variable is introduced; 1 means the state is has implemented expansion and 0 means it has not. The histograms for the subgroups are shown below:





Both subgroups still have long right tails. This could, however, simply be the effect of unusual observations. The Medicaid variable may still be informative and could help reduce heteroscedasticity in the health insurance variable. Therefore, it will continue to be included in the analysis. For the remaining predictors, the plots are promising. In smoking, obesity and sleep deprivation, the plots suggest strong positive relationships with poor mental health, which is consistent with common knowledge that these predictors tend to be bad for mental health. One note to make, however, is that these three plots look very similar, which may indicate high correlation and multicollinearity in models containing these variables.

To identify unusual observations, the analysis will look for absolute standardized residuals > 2.5, hat values > 2.5*(p+1/n) where p is the number of predictors and n is the sample size, and Cook's distance > 1. However, these are guidelines and any observations that seem unusual despite having diagnostic values below the thresholds will still be removed. To account for outliers and unusual observations with respect to all the predictors, the model fitted in this section will be the full model. In deciding between a full model with an interaction term Medicaid * Health_Insurance versus a constant shift model, a partial-F test was completed. The output is shown below:

```
Analysis of Variance Table


Model 1: Mental_Health ~ Drinking + Smoking + Medicaid + Health_Insurance +
    Obesity + Lack_of_Sleep + Checkup + Midwest + Northeast + South
Model 2: Mental_Health ~ Drinking + Smoking + Medicaid + Health_Insurance+ Ob
esity + Lack_of_Sleep + Checkup + Midwest + Northeast + South + Medicaid*Heal
th_Insurance
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    489 251.54
2    488 248.69  1      2.85 5.5926 0.01843 *
```

The adjusted R^2 improvement was low (about 1%) but this could be a result of leverage points obscuring interaction between Medicaid and health insurance. Since p-value of the partial F-test

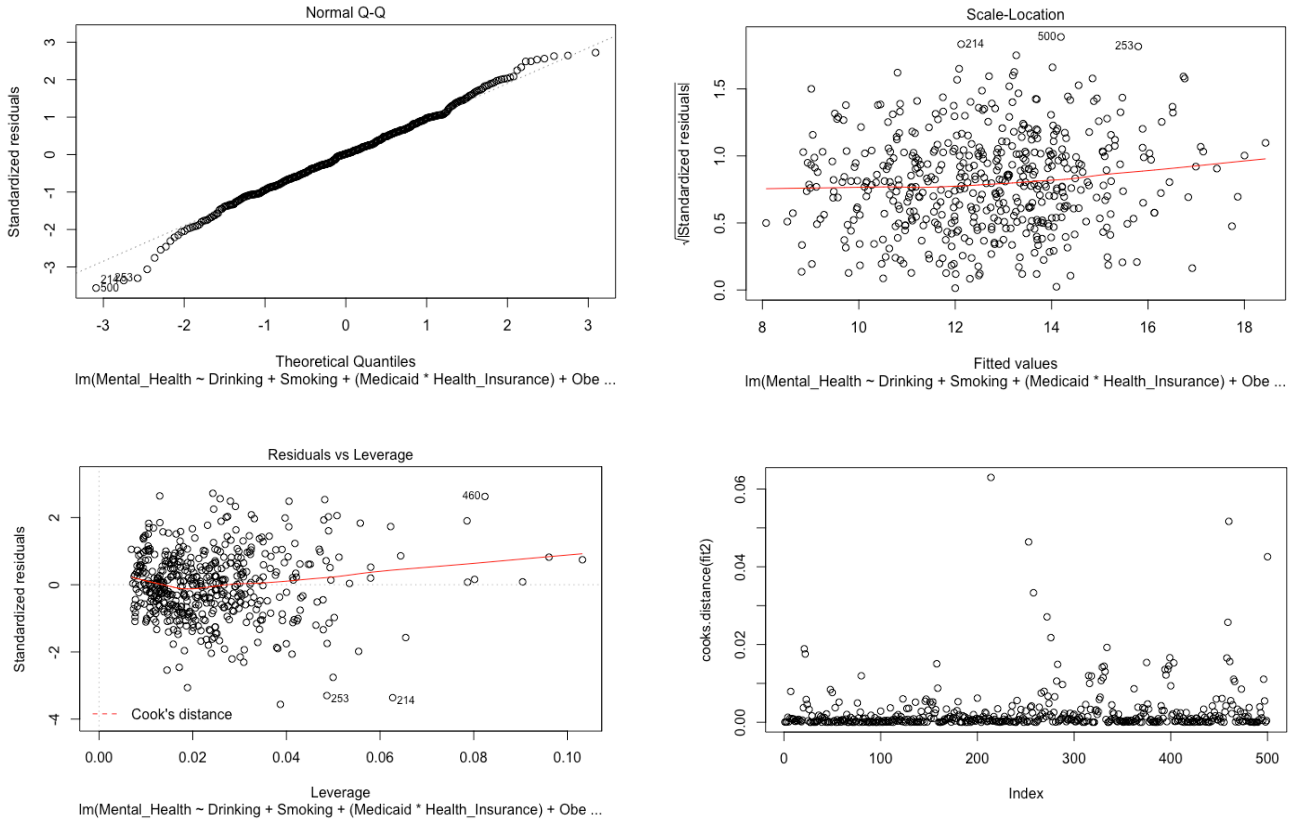
is statistically significant, the analysis will continue with the full model for identifying unusual observations. Below is the multiple regression model fitted and its summary output:

```
lm(formula = Mental_Health ~ Drinking + Smoking + Medicaid + Health_Insurance
+ Obesity + Lack_of_Sleep + Checkup + Midwest + Northeast + South + Medicaid*
Health_Insurance, data = data)


Coefficients:              Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.776263   0.859256   5.559 4.48e-08 ***
Drinking                  -0.053238   0.016203  -3.286 0.001091 **
Smoking                    0.350145   0.014599  23.984  < 2e-16 ***
Medicaid                  -0.534670   0.212969  -2.511 0.012377 *
Health_Insurance           0.042021   0.009133   4.601 5.37e-06 ***
Obesity                    0.040854   0.011433   3.574 0.000387 ***
Lack_of_Sleep              0.055046   0.014615   3.766 0.000186 ***
Checkup                   -0.005355   0.013672  -0.392 0.695502
Midwest                   -1.450280   0.158070  -9.175  < 2e-16 ***
Northeast                 -0.255518   0.169012  -1.512 0.131224
South                     -1.227448   0.182835  -6.713 5.30e-11 ***
Medicaid:Health_Insurance  0.029978   0.012676   2.365 0.018427 *


Residual standard error: 0.7139 on 488 degrees of freedom
Multiple R-squared:  0.8778,    Adjusted R-squared:  0.875
F-statistic: 318.6 on 11 and 488 DF,  p-value: < 2.2e-16
```

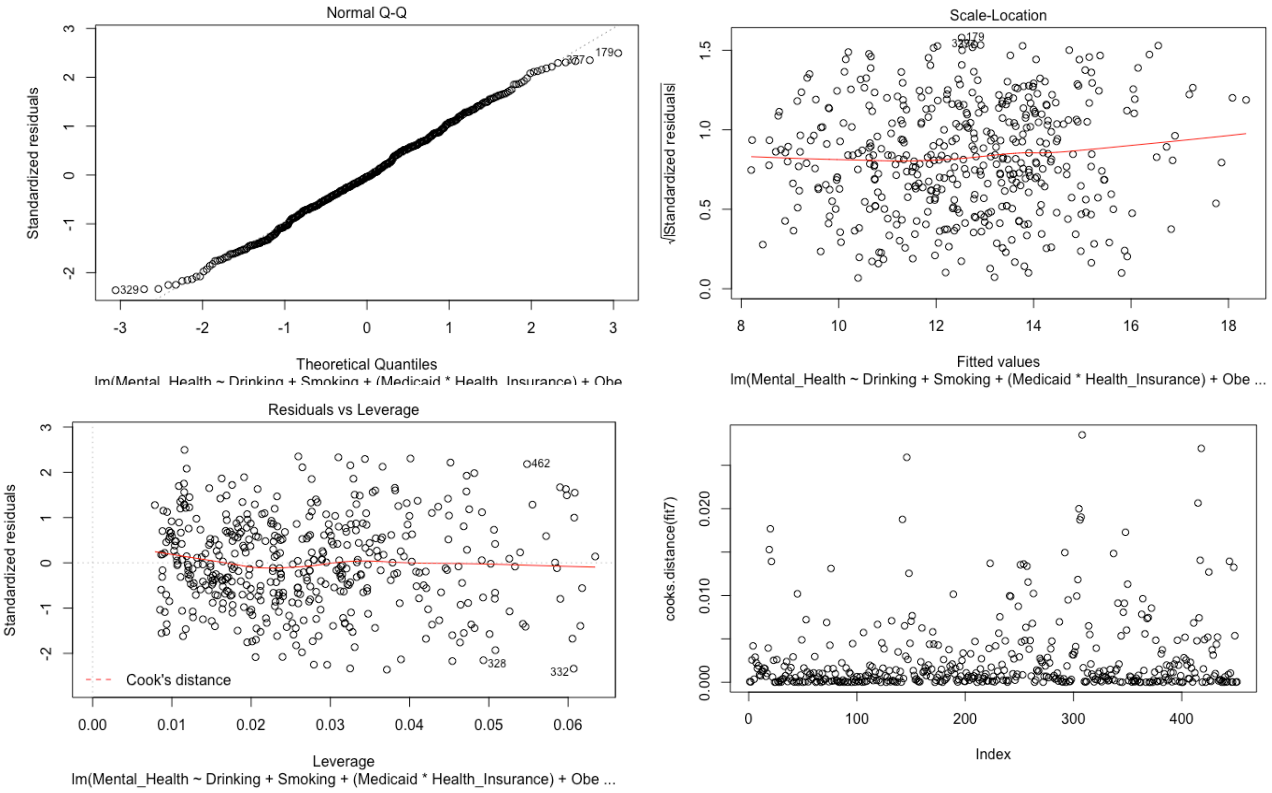
Unusual observations and outliers are removed and a model with the same predictors is fitted again on the new subset of data. This process was repeated six times and a total of 52 observations were removed (10.4% of the dataset). The process was repeated due to masking

effect, where outliers or leverage points can be hidden before the removal of another usually more extreme outlier or leverage point. The full list of observations removed can be found in Appendix 1. Many of the observations were unusual. For example, the outlier Fall River, Massachusetts has a significant drug abuse problem, especially among their homeless population, which lead to abnormally high percentage of the population with poor mental health. Honolulu was both an outlier and a leverage point; the city had an abnormally low value for poor mental health, perhaps due to the lifestyle and environment there. It also had a very high rate of sleep deprivation due to the work culture where many people take on multiple jobs and work extremely long hours, contributing to its high hat value. Some observations like Thornton, Colorado was flagged as an outlier, but there is no evidence to explain why. These observations were still left out of the analysis due to their influence on the regression, but that simply reflected the limitation of the model and not a problem with the data. The diagnostic plots for the initial fit and the final fit are shown below. The diagnostic plots and regression outputs for the rounds in between are in Appendix 2:

Initial Sample (n=500):



Final Sample (n=448):




```
Coefficients:              Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.496101   0.795408   3.138 0.001815 **
Drinking                  -0.004225   0.014766  -0.286 0.774908
Smoking                    0.382987   0.013308  28.779  < 2e-16 ***
Medicaid                  -0.742406   0.213733  -3.474 0.000565 ***
Health_Insurance           0.051315   0.010786   4.757 2.67e-06 ***
Obesity                    0.034300   0.009733   3.524 0.000470 ***
Lack_of_Sleep              0.032662   0.013002   2.512 0.012358 *
Checkup                    0.023025   0.012485   1.844 0.065815 .
Midwest                   -1.825477   0.140439 -12.998  < 2e-16 ***
Northeast                 -0.674643   0.146165  -4.616 5.15e-06 ***
South                     -1.687157   0.163301 -10.332  < 2e-16 ***
Medicaid:Health_Insurance  0.043556   0.012995   3.352 0.000873 ***
---
```

```
Residual standard error: 0.5669 on 439 degrees of freedom
Multiple R-squared:  0.9215,    Adjusted R-squared:  0.9196
F-statistic: 468.7 on 11 and 439 DF,  p-value: < 2.2e-16
```

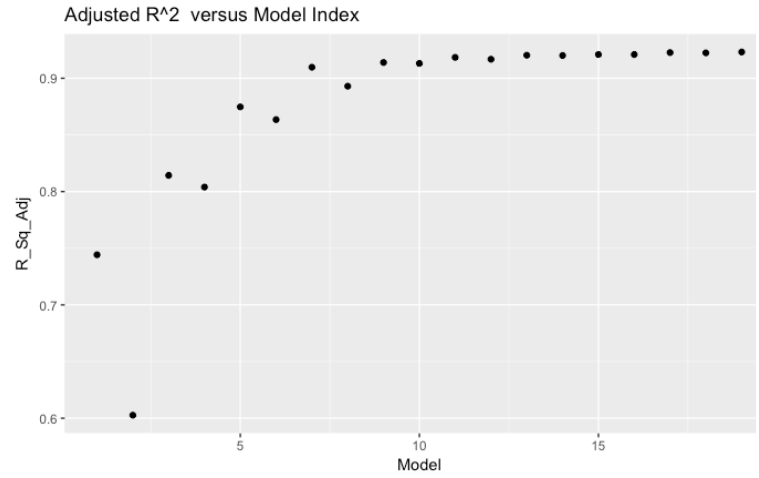
In the initial standardized residuals vs fitted values plot, the residuals look random and spread out which implies the variance is close to constant. There are outliers where the standardized residuals are above 2.5. In the normality plot, the residuals look mostly normal with some outliers, mostly concentrated in values with surprisingly low percentage of poor mental health. Lastly, there are clear leverage points in the standardized residuals vs leverage plot. After cleaning, the final dataset no longer has any clear outliers or leverage points. Furthermore, while the initial Cook's D values were far below 1, some of them were vastly larger than the rest. Those observations were also removed. On the normal probability plot, the residuals still show signs of non-normality which indicates that a linear model may not fully capture the relationships in this data. The adjusted R^2 has improved from 87.5% to 91.96%. The precision of this model on the cleaned dataset, however, is mediocre. The 95% confidence interval is (12.27, 12.67) and the prediction interval for New York City is (11.34, 13.60). The prediction interval tells us there is a 95% probability that the percentage of the population with poor mental health in a city like New York City is between 11.34% and 13.60%. That is not precise enough to tell us how well the city is doing since the lower bound is considered acceptable but the upper bound is concerning (the full range in this subset is 7.9-18.2%). The most significant impact of removing the unusual observations is that the predictors Northeast and Checkup are now significant, while Drinking is no longer significant. That is important because the scatterplot showed a negative relationship between binge drinking and poor mental health, which contradicts many established links between the two. The other coefficients are similar in magnitude and the same in direction.

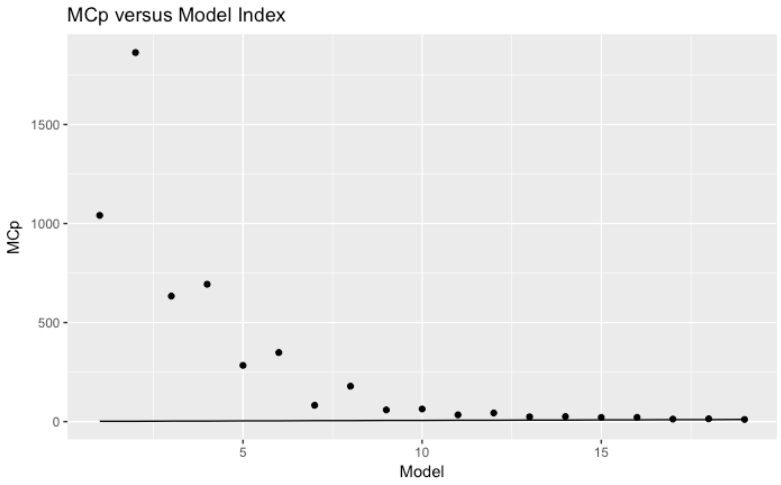
3. Model Selection:

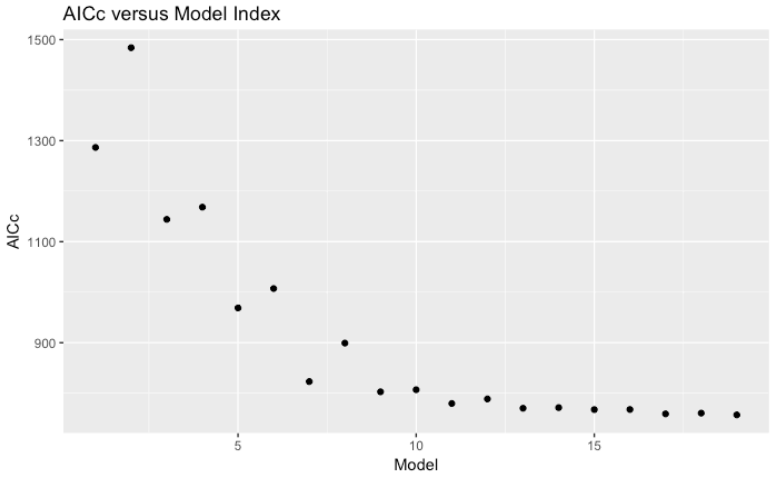
The full model with all the predictors is not the best model. In the previous section, the predictor binge drinking is clearly not significant and can be dropped from model selection. The predictor routine checkup is close to significant with p-value = 0.0658 and can be further examined. To perform model selection, the best subset selection algorithm identifies the best models given p predictors. The output below are 19 models identified by the algorithm, two models for each p from p=1 to p=9, along with one full model p=10:

	Smoking	Medicaid	Health_Insurance	Medicaid*Health	Obesity	Lack_of_Sleep	Checkup	Midwest	Northeast	South
1 (1)	*									
1 (2)						*				
2 (1)	*								*	
2 (2)	*				*					
3 (1)	*				*				*	
3 (2)	*								*	*
4 (1)	*			*					*	*
4 (2)	*			*	*				*	
5 (1)	*			*			*		*	*
5 (2)	*			*					*	*	*
6 (1)	*			*			*		*	*	*
6 (2)	*			*				*	*	*	*
7 (1)	*			*		*	*		*	*	*
7 (2)	*			*		*		*	*	*	*
8 (1)	*	*		*	*		*		*	*	*
8 (2)	*			*		*	*	*	*	*	*
9 (1)	*	*		*	*	*		*	*	*	*
9 (2)	*	*		*	*	*	*		*	*	*
10 (1)	*	*		*	*	*	*	*	*	*	*

In order to narrow down the range of models to choose from, the adjusted R^2, MCp and AICc of each model are compared. The ideal range of models would be where the measures peak and begin to plateau, which means additional predictors are not improving model performance. Moreover, a good model for the MCp measure should have MCp ≈ 1 + p. Below are the plots of each measure against the index of models. The horizontal line in the MCp plot is 1 + p:







The adjusted R^2 begins to plateau around model 11, while both MCp and AICc flattens around model 13 and 14. A good range to test is models 9-14, which gives us the option to choose simpler models compare to the optimal models identified in the graphs. This range includes p = 5, 6, 7. The better of each pair generated by best subset algorithm is fitted. The regression outputs are shown below:

p = 5:

```
lm(formula = Mental_Health ~ Smoking + Health_Insurance + Lack_of_Sleep +
    Midwest + South, data = data)

Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.325404   0.244439  13.604  < 2e-16 ***
Smoking           0.398435   0.010423  38.225  < 2e-16 ***
Health_Insurance  0.093778   0.006812  13.768  < 2e-16 ***
Lack_of_Sleep     0.045433   0.009554   4.756 2.68e-06 ***
Midwest          -1.451689   0.091266 -15.906  < 2e-16 ***
South            -1.367703   0.073574 -18.590  < 2e-16 ***

Residual standard error: 0.5874 on 442 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9139
F-statistic: 950.2 on 5 and 442 DF,  p-value: < 2.2e-16
VIF:
Smoking   Health_Insurance    Lack_of_Sleep     Midwest            South
2.598759      1.715631         2.221754         1.662797         1.504229
```

p = 6

```
lm(formula = Mental_Health ~ Smoking + Health_Insurance + Lack_of_Sleep +
    Midwest + Northeast + South, data = data)

Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.100289   0.242058  12.808  < 2e-16 ***
Smoking           0.411736   0.010481  39.282  < 2e-16 ***
Health_Insurance  0.088326   0.006717  13.149  < 2e-16 ***
```

```
Lack_of_Sleep     0.051196   0.009369   5.465 7.77e-08 ***
Midwest          -1.626906   0.095360 -17.061  < 2e-16 ***
Northeast        -0.521862   0.103268  -5.053 6.37e-07 ***
South            -1.500712   0.076296 -19.670  < 2e-16 ***

Residual standard error: 0.5717 on 441 degrees of freedom
Multiple R-squared:  0.9195,    Adjusted R-squared:  0.9185
F-statistic: 840.1 on 6 and 441 DF,  p-value: < 2.2e-16
VIF:
Smoking   Health_Insurance  Lack_of_Sleep  Midwest  Northeast   South
2.773654      1.761057     2.255176       1.916111   1.372570  1.707424
```

$p = 7$

```
lm(formula = Mental_Health ~ Smoking + Health_Insurance + Obesity +
    Lack_of_Sleep + Midwest + Northeast + South, data = data)

Coefficients:     Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.947189   0.243480  12.104  < 2e-16 ***
Smoking           0.383576   0.013288  28.867  < 2e-16 ***
Health_Insurance  0.079498   0.007133  11.145  < 2e-16 ***
Obesity           0.031273   0.009241   3.384 0.000778 ***
Lack_of_Sleep     0.047632   0.009319   5.111 4.79e-07 ***
Midwest          -1.698746   0.096611 -17.583  < 2e-16 ***
Northeast        -0.535627   0.102146  -5.244 2.45e-07 ***
South            -1.530811   0.075931 -20.161  < 2e-16 ***

Residual standard error: 0.5651 on 440 degrees of freedom
Multiple R-squared:  0.9216,    Adjusted R-squared:  0.9203
F-statistic: 738.8 on 7 and 440 DF,  p-value: < 2.2e-16
VIF:
Smoking  Health_Insurance   Obesity    Lack_of_Sleep   Midwest   Northeast
4.563291   2.032955         4.481496     2.284360      2.013319    1.374750
South
1.731177
```

Examining the three models, the adjusted R^2 are all very similar which implies there is limited improvement in performance from adding additional predictors to the p = 5 model. The F-statistics are all very significant, which means the slope coefficients are not zero and the models contribute predictive power. The predictors for all three models are also very significant. The t-statistics, and consequently p-values, have to be adjusted to roughly account for the problem of post selection inference. However, the p-values are extremely small and the sample size (n=448) is large which means all predictors will clearly be significant after adjustment. The VIFs for the predictors of all three models are below 10, in fact, the highest VIF is only 4.56 for the smoking predictor in p = 7 model. Hence, none of the models suffer from multicollinearity which could result in an unstable model with poor generalization. Given the similarity in performance and stability of the models, the best model is the simplest one— the p = 5 model.

4. Interpreting the best model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

$$y = 3.3254 + 0.3984 x_1 + 0.0938 x_2 + 0.0454 x_3 - 1.4517 x_4 - 1.368 x_5$$

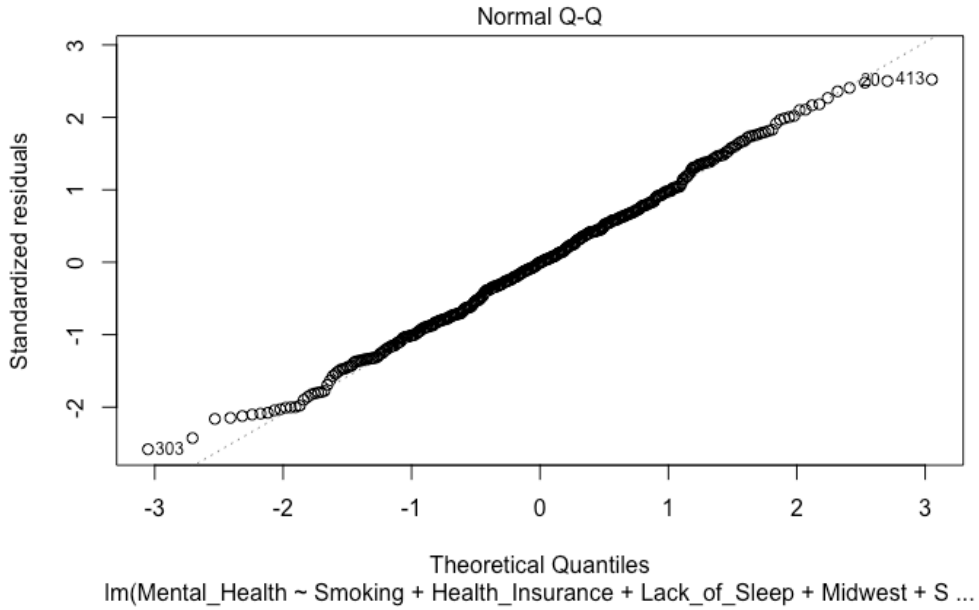
x_1: Percentage of respondents that smoke frequently
x_2: Percentage of respondents that lack health insurance
x_3: Percentage of respondents that are sleep deprived
x_4: Whether the city is located in the Midwest (1/0)
x_5: Whether the city is located in the South (1/0)

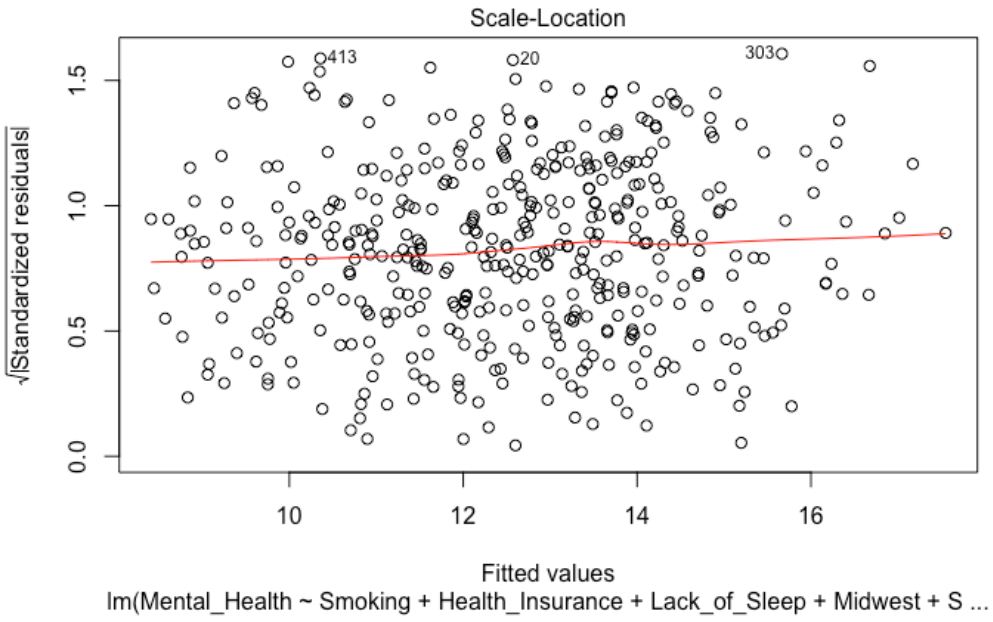
The intercept β_0 has no practical meaning. It represents the response when a city is located in either Northeast region or West region, and none of its population smoke frequently, lack health insurance or lack sleep. Unfortunately, such a city is far from existing in America. The slope coefficients $\beta_1, \beta_2, \beta_3$ are all additive/additive relationships and can be interpreted in the same manner. For example, $\beta_1 = 0.3984$ means that a one percentage point change in the percentage of people that smoke frequently is associated with a 0.3984 percentage point change in the percentage of people with poor mental health in U.S. cities, given that all other variables are held constant. The coefficients β_4, β_5 are constant shifts which results in three parallel

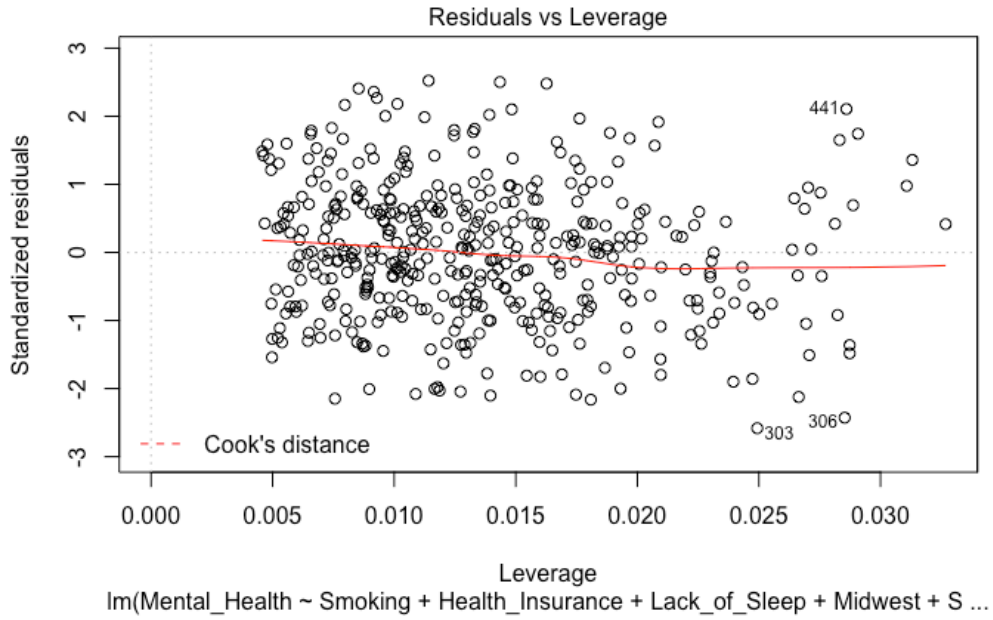
hyperplanes. For example, β_4 = -1.4517 means that compared to the reference group, which are cities in Northeast region or West region, cities located in the Midwest experience a -1.4517 percentage point change in the percentage of people with poor mental health. Looking at the size of the coefficients, it seems like location has a large impact on a city's mental health. In particular, cities in the Midwest or South have much better mental health. Of the health indicators, smoking has the largest positive coefficient, which suggests smoking greatly deteriorates a city's mental health. The approximate 95% prediction interval for New York City is (11.18, 13.53). However, this interval needs to be adjusted for post selection inference:

$$\tilde{s} = s\sqrt{\frac{n-p-1}{n-p*-1}} = 0.5874\sqrt{\frac{448-5-1}{448-10-1}} = 0.5908$$

The new 95% prediction interval for New York City is approximately ±(2)(0.5908) = ±1.1816, which means the model predicts with 95% probability that the response value lies between (11.17, 13.54). The interval is only about 0.85% wider. However, as mentioned previously, a prediction interval of (11.17, 13.54) is quite wide in this context, given that a city in the lower bound is considered to have decent mental health while a city in the upper bound is facing a slight mental health problem— for observations like New York City, the model does not precisely predict a populations' mental wellbeing. The diagnostic plots for the model are shown below:







There are no clear outliers or leverage points, although some observations do have high standardized residuals and hat values. The cook's distance for all the observations are very small and close together, the largest value was 0.029. On the standardized residuals vs fitted plot, the residuals look very randomly distributed and even across the zero line. On the normal probability plot, the data is mostly normally distributed, but there are a few points that were surprisingly large or small. These observations possibly reflect nonlinear relationships the model failed to capture. But overall, the data seems fairly consistent with the assumptions of linear regression.

6. Conclusion:

The best model based on this dataset suggests that mental health in American cities can be modelled with a surprisingly small number of predictors. Moreover, it seems that location contributes the most to mental health, as opposed to large scale health indicators like obesity. In addition, health insurance coverage has a small coefficient compare to specific unhealthy behaviours like smoking and sleep deprivation. Therefore, policies targeting mental health may be more effective if it also targets smoking and sleep deprivation, as opposed to just improving healthcare coverage. Overall, this model has many limitations. The wide prediction interval severely limits the usefulness of predictions from this model. In addition, there is no new data available to validate the model. Lastly, the region predictors very likely substitutes for a variable or several variables that were not present in the dataset but has a strong relationship with mental health, such as a region's attitude and stigma towards mental health. A stronger model would be one which contains these variables and captures their relationships with mental health.

Data:

1) 500 Cities: Local Data for Better Health, 2018 Release | Chronic Disease and Health Promotion Data & Indicators." *Centers for Disease Control and Prevention, Centers for Disease Control and Prevention*, chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2018-relea/6vp6-wxuq.

2) *United States Census Bureau*, www2.census.gov/programs-surveys/popest/geographies/2017/.

3) The Status of State Action on Medicaid Expansion Decision." *Henry J. Kaiser Family Foundation*, https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-careact/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.

Appendix 1:

Round 1

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Massachusetts	Fall River	17.0	27.9	31.0	40.3	77.3	11.3	1	11.3	Northeast	18.5	2.535	-	-
New Jersey	Passaic	13.2	21.0	37.4	41.2	74.9	31.3	1	31.3	Northeast	15.2	-	0.065	-
California	El Cajon	19.2	18.2	24.9	35.3	64.1	14.1	1	14.1	West	15.1	2.645	-	-
South Carolina	Columbia	16.3	18.8	35.2	38.5	67.4	17.5	0	0.0	South	15.0	2.561	-	-
Texas	Brownsville	13.7	17.4	44.4	35.0	65.4	41.4	0	0.0	South	14.1	-	0.090	-
Oklahoma	Norman	13.5	16.8	32.0	34.0	69.0	13.3	0	0.0	South	14.0	2.722	-	-
Texas	Pharr	12.9	16.5	35.2	34.3	63.9	43.3	0	0.0	South	14.0	-	0.096	-
Texas	Laredo	14.3	16.9	41.4	35.0	66.2	41.0	0	0.0	South	13.8	-	0.080	-
Illinois	Cicero	18.9	18.7	36.7	35.6	66.3	30.7	1	30.7	Midwest	13.7	-	0.079	-
Florida	Hialeah	14.7	16.4	30.2	34.7	73.1	43.7	0	0.0	South	13.6	-	0.103	-
Kansas	Kansas City	14.0	24.8	43.8	36.9	66.2	25.4	0	0.0	Midwest	13.5	3.300	-	0.046
Utah	Provo	7.3	9.1	29.3	31.4	57.9	14.6	0	0.0	West	12.6	2.628	0.082	0.052
Kansas	Wichita	14.8	22.3	31.6	35.6	64.9	18.3	0	0.0	Midwest	12.1	2.757	-	-
Wyoming	Cheyenne	16.8	19.9	29.4	33.4	63.4	15.3	0	0.0	West	11.7	3.562	-	0.043
Utah	Orem	8.0	8.5	27.7	30.7	58.0	13.3	0	0.0	West	11.7	-	0.079	-
Colorado	Thornton	18.6	18.7	25.9	30.6	61.0	14.6	1	14.6	West	11.1	3.063	-	-
Utah	Layton	9.3	9.1	28.9	30.8	63.0	11.9	0	0.0	West	11.1	-	0.064	-
Alaska	Anchorage	18.2	17.2	27.5	30.7	58.8	12.6	1	12.6	West	10.9	2.541	-	-
South Carolina	Mount Pleasant	23.1	13.0	24.0	29.2	63.3	9.1	0	0.0	South	10.7	-	0.062	-
Hawaii	Honolulu	18.3	14.1	23.2	47.4	70.9	10.3	1	10.3	West	9.8	3.362	0.063	0.063

Round 2

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Massachusetts	New Bedford	16.4	27.3	32.2	40.8	77.5	13.1	1	13.1	Northeast	18.5	2.531	-	-
Michigan	Dearborn	21.2	25.8	33.7	39.7	68.7	13.3	1	13.3	Midwest	16.6	2.567	-	-
Rhode Island	Pawtucket	15.7	21.0	30.7	41.4	81.0	13.3	1	13.3	Northeast	16.2	2.823	-	-
Utah	Ogden	11.9	13.7	37.5	34.2	62.9	19.3	0	0.0	West	14.1	-	0.083	-
New Jersey	Union City	14.4	17.9	34.4	42.0	75.0	28.8	1	28.8	Northeast	13.5	-	0.065	-
Florida	Miami	15.2	16.9	29.8	35.2	74.7	36.3	0	0.0	South Region	13.3	-	0.062	-
Texas	Mission	14.4	14.3	31.6	32.6	64.7	36.5	0	0.0	South	12.2	-	0.079	-
Utah	St. George	11.2	10.3	25.4	32.4	60.2	13.6	0	0.0	West	12.2	-	0.086	-
Texas	Edinburg	14.1	14.0	32.4	32.8	64.9	35.5	0	0.0	South	12.0	-	0.071	-
Texas	McAllen	14.2	13.8	31.6	32.5	64.9	34.4	0	0.0	South	11.9	-	0.072	-
Montana	Billings	21.2	19.2	30.0	31.0	65.0	12.4	1	12.4	West	11.8	2.573	-	-
Nebraska	Omaha	21.3	19.7	33.8	31.8	65.6	16.3	0	0.0	Midwest	11.0	2.574	-	-
Michigan	Ann Arbor	20.7	12.5	25.6	31.1	69.5	7.8	1	7.8	Midwest	10.6	2.560	-	-
Colorado	Westminster	20.0	16.6	23.8	28.7	61.6	11.5	1	11.5	West	10.3	2.509	-	-

Round 3

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Rhode Island	Providence	13.8	19.2	32.3	41.6	81.1	16.5	1	16.5	Northeast	15.8	2.603	-	-
Arkansas	Jonesboro	13.9	21.9	36.5	32.7	70.4	11.6	1	11.6	South	15.1	2.514	-	-
Texas	Pasadena	15.9	16.9	32.5	34.2	66.5	32.8	0	0.0	South	13.5	-	0.076	-
Texas	El Paso	17.3	15.6	34.9	36.9	67.9	32.7	0	0.0	South	12.4	-	0.073	-
Montana	Missoula	22.4	18.8	26.5	28.5	61.6	11.2	1	11.2	West	11.4	2.542	-	-
South Dakota	Sioux Falls	19.3	18.7	31.9	27.1	69.5	11.3	0	0.0	Midwest	10.3	2.674	-	-

Round 4

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Arkansas	Fort Smith	13.8	25.7	35.5	35.4	70.1	16.2	1	16.2	South	16.9	2.505	-	-
South Carolina	North Charleston	16.8	21.7	36.1	38.8	66.3	20.9	0	0.0	South	15.8	2.555	-	-
Michigan	Kalamazoo	18.2	21.2	41.7	39.4	71.7	13.0	1	13.0	Midwest	15.1	2.531	-	-
Utah	West Valley City	13.0	13.4	29.5	32.8	62.4	20.6	0	0.0	West	15.1	2.616	-	-
Colorado	Aurora	16.8	17.9	25.3	31.3	60.0	15.8	1	15.8	West	11.8	2.520	-	-

Round 5

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Oklahoma	Oklahoma City	12.6	18.8	32.9	34.3	68.1	18.9	0	0.0	South	14.3	2.599	-	-
South Carolin	Charleston	20.5	15.9	29.1	33.2	65.1	12.3	0	0.0	South	12.6	2.617	-	-
Colorado	Lakewood	20.2	16.5	24.3	27.4	61.7	12.1	1	12.1	West	10.8	2.506	-	-
District of C	Washington	22.6	17.2	26.4	36.7	80.0	7.7	1	7.7	South	10.7	-	0.067	-

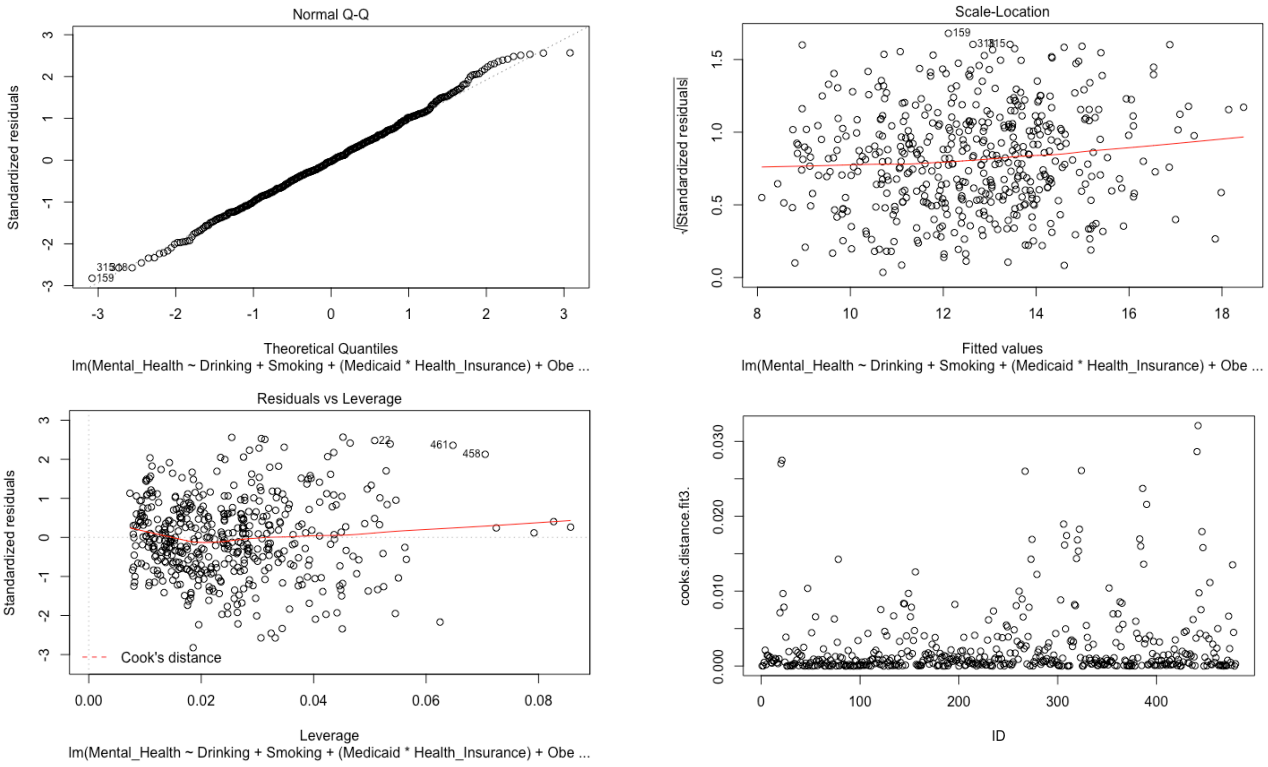
Round 6

State	CityName	Drinking	Smoking	Obesity	Lack_of_Sleep	Checkup	Health_Insurance	Medicaid	Medi_Health	Region	Mental_Health	Std.Residuals	Hat Values	Cook's D
Colorado	Arvada	21.4	16.3	23.2	26.9	61.8	10.2	1	10.2	West	10.5	2.636	-	-

Appendix 2:

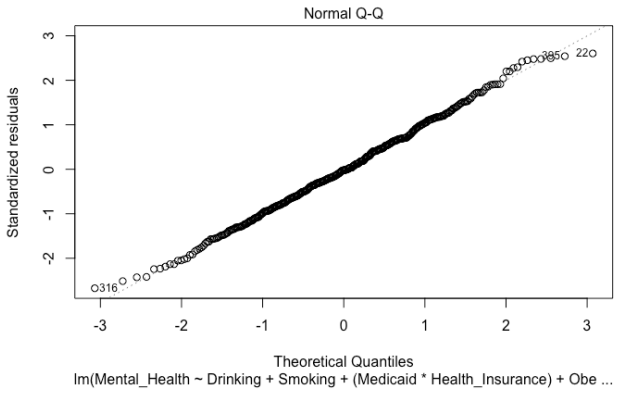
Round 2:

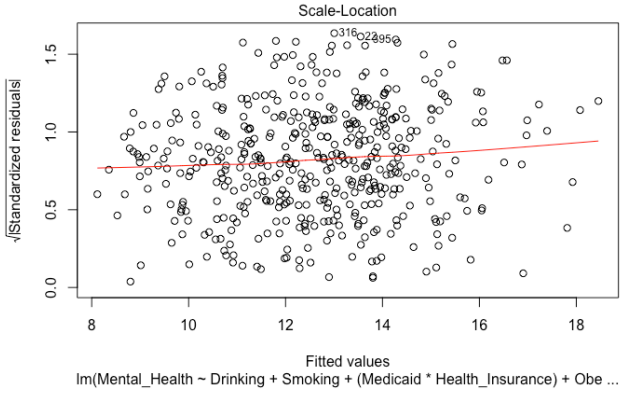
```
Coefficients:                Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.876640   0.839833   4.616 5.06e-06 ***
Drinking                    -0.039408   0.015819  -2.491 0.013076 *
Smoking                      0.366005   0.014032  26.084  < 2e-16 ***
Medicaid                    -0.498048   0.216156  -2.304 0.021653 *
Health_Insurance             0.051807   0.010053   5.153 3.78e-07 ***
Obesity                      0.035677   0.010702   3.334 0.000925 ***
Lack_of_Sleep                0.054119   0.013899   3.894 0.000113 ***
Checkup                      0.002571   0.013055   0.197 0.843963
Midwest                     -1.548206   0.148642 -10.416  < 2e-16 ***
Northeast                   -0.409956   0.157695  -2.600 0.009626 **
South                       -1.454420   0.172716  -8.421 4.61e-16 ***
Medicaid:Health_Insurance    0.025018   0.013078   1.913 0.056367 .

Residual standard error: 0.6472 on 468 degrees of freedom
Multiple R-squared:  0.8995,    Adjusted R-squared:  0.8971
F-statistic: 380.7 on 11 and 468 DF,  p-value: < 2.2e-16
```

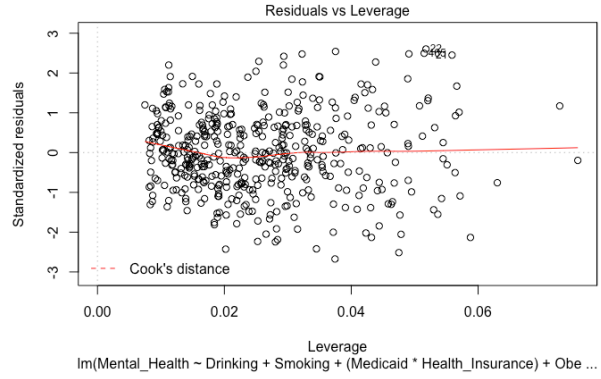


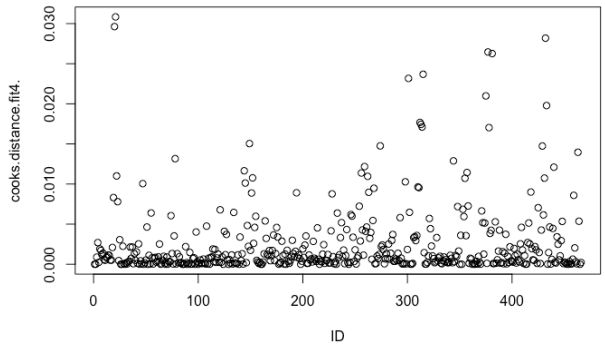
Round 3:

```
Coefficients                Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.461441   0.821042   4.216 3.00e-05 ***
Drinking                   -0.022155   0.015384  -1.440 0.150522
Smoking                     0.369998   0.013834  26.745  < 2e-16 ***
Medicaid                   -0.619836   0.220477  -2.811 0.005147 **
Health_Insurance            0.050465   0.010860   4.647 4.42e-06 ***
Obesity                     0.037407   0.010350   3.614 0.000335 ***
Lack_of_Sleep               0.045870   0.013400   3.423 0.000675 ***
Checkup                     0.006946   0.012769   0.544 0.586724
Midwest                    -1.640813   0.147180 -11.148  < 2e-16 ***
Northeast                  -0.485498   0.153335  -3.166 0.001648 **
South                      -1.454539   0.168205  -8.647  < 2e-16 ***
Medicaid:Health_Insurance   0.036813   0.013343   2.759 0.006031 **

Residual standard error: 0.6119 on 454 degrees of freedom
Multiple R-squared:  0.909, Adjusted R-squared:  0.9068
F-statistic: 412.5 on 11 and 454 DF,  p-value: < 2.2e-16
```



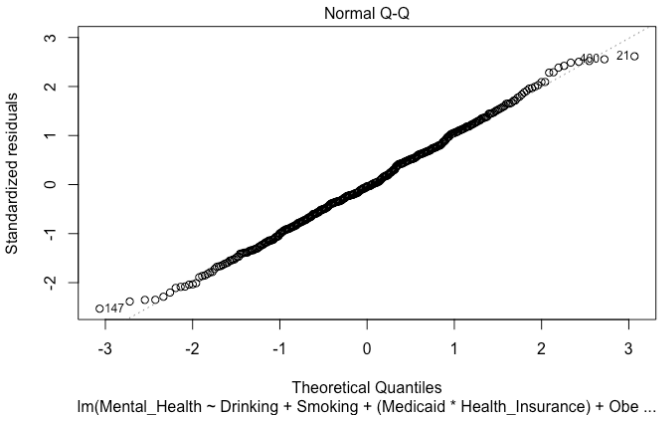


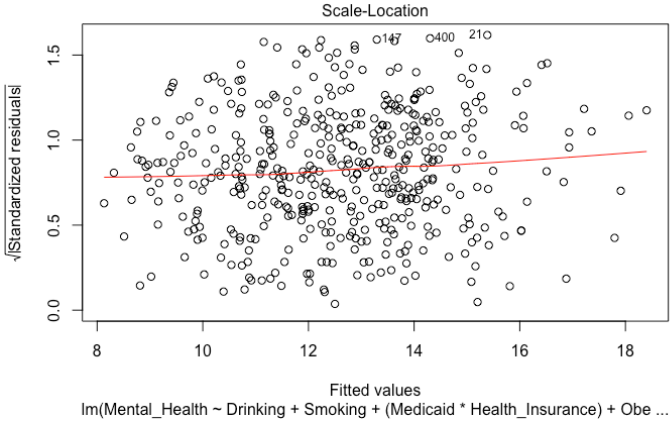


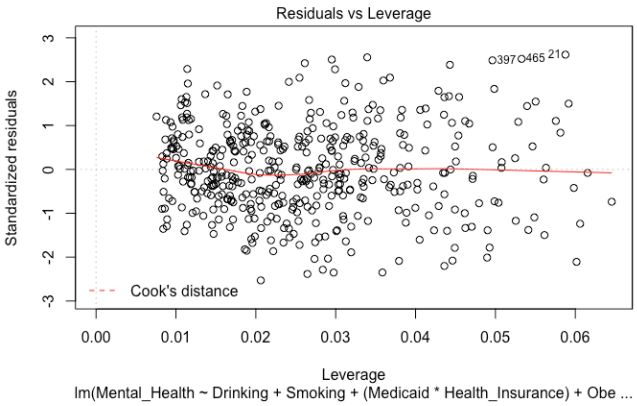


Round 4:

```
Coefficients:                Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.371586   0.804959   4.189 3.38e-05 ***
Drinking                    -0.013343   0.015150  -0.881 0.378936
Smoking                      0.374733   0.013675  27.403  < 2e-16 ***
Medicaid                    -0.754099   0.223325  -3.377 0.000798 ***
Health_Insurance             0.047500   0.011209   4.238 2.74e-05 ***
Obesity                      0.037626   0.010115   3.720 0.000225 ***
Lack_of_Sleep                0.043538   0.013355   3.260 0.001199 **
Checkup                      0.008031   0.012677   0.633 0.526741
Midwest                     -1.700219   0.145756 -11.665  < 2e-16 ***
Northeast                   -0.548314   0.151091  -3.629 0.000317 ***
South                       -1.549994   0.166618  -9.303  < 2e-16 ***
Medicaid:Health_Insurance    0.040888   0.013478   3.034 0.002556 **

Residual standard error: 0.5966 on 448 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.9115
F-statistic: 430.6 on 11 and 448 DF,  p-value: < 2.2e-16
```







Round 5:

```
Coefficients:                Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.150371   0.789098   3.992 7.66e-05 ***
Drinking                    -0.009107   0.014791  -0.616 0.538416
Smoking                      0.378130   0.013491  28.028  < 2e-16 ***
Medicaid                    -0.803889   0.216639  -3.711 0.000233 ***
Health_Insurance             0.047377   0.010900   4.347 1.72e-05 ***
Obesity                      0.033811   0.009921   3.408 0.000714 ***
Lack_of_Sleep                0.040718   0.013096   3.109 0.001997 **
Checkup                      0.012514   0.012443   1.006 0.315098
Midwest                     -1.735442   0.141633 -12.253  < 2e-16 ***
Northeast                   -0.575299   0.147073  -3.912 0.000106 ***
South                       -1.583650   0.164265  -9.641  < 2e-16 ***
Medicaid:Health_Insurance    0.045026   0.013176   3.417 0.000691 ***

Residual standard error: 0.5782 on 443 degrees of freedom
Multiple R-squared:  0.9181,    Adjusted R-squared:  0.916
F-statistic: 451.4 on 11 and 443 DF,  p-value: < 2.2e-16
```