STA6235 / Dr. Amin
Summer 2023
Group 4
Alzheimer's Project

Group Members:

Brad Lipson

Pamela Mishaw

Jorge Sanchez

Daniel Wilson

Jolie Wise

STA6235: Summer 2023 Group 4 Alzheimer's Project

As part of our linear regression class project, we used a dataset provided to us which was adapted from www.countyhealthrankings.org in order to learn more about Obesity within the Alzheimer's Disease data from across the US. We took an approach to the project by choosing five interesting states as a group and giving each student one of those states to study in depth. We carefully included certain variables:

- Age Adj Alzheimer's Rate
- Mental Distress
- Top 10
- Median Age
- 65+ Rate
- Smoking Rate
- Physical Inactivity Rate
- Diabetes Rate
- Heart Disease Rate
- Cancer Rate
- Glyphosate Rate
- NATA Cancer Rate
- Fine Particulate Matter
- Mercury
- Lead

We put all of our data into one shared Excel file by carefully collecting and organizing it for the five states. Using this common dataset, we wanted to use linear regression methods to investigate the variable of Obesity to find possible models and insights within the counties of our chosen states. This helped us get a better understanding of Obesity within each of the five states in order to find outliers, extreme values, multicollinearity, and the best overall model. This included simple and multiple linear regression analysis, lack of fit testing, matrix techniques, model adequacy, and influence diagnostics. We attempted to find correlations between the variables and determined the slope, intercept, and confidence intervals of each. This allowed us to pick candidate models for fitting that were also helpful for predicting Obesity.

PART I - MULTIPLE LINEAR REGRESSION ANALYSIS

1. **Run SAS to find any data points whose removal might improve the model:**

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Brad Lipson (NY)</td>
<td>To show data points that greatly affected my regression model's performance. To assess each point's effect using R Student, the hat matrix diagonals, and the covariance ratio among the 62 counties of the state of New York.</td>
<td>proc reg;
model obesity_age_adj = alz_ageadj_rate mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY /influence;
run;

(To remove observation 22 would be:)
z=_n_;
if z=22 then delete;</td>
<td>
<table>
<tr><th>Obs</th><th>Residual</th><th>RStudent</th><th>Hat Diag H</th><th>Cov Ratio</th></tr>
<tr><td>22</td><td>3.9919</td><td>3.2703</td><td>0.1882</td><td>0.0739</td></tr>
<tr><td>25</td><td>-2.7534</td><td>-2.1175</td><td>0.1796</td><td>0.417</td></tr>
</table>
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>162.44680</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>213.72110</td></tr>
</table>
<table>
<tr><th></th><th colspan="4">DFBETAS</th></tr>
<tr><th>DFFITS</th><th>Intercept</th><th>Mercury_TPY</th><th>Diabetes</th><th>Glyphosates</th></tr>
<tr><td>0.6394</td><td>0.1212</td><td>0.5445</td><td>-0.0937</td><td>-0.2164</td></tr>
<tr><td>0.0662</td><td>-0.0197</td><td>-0.0142</td><td>0.0138</td><td>-0.0311</td></tr>
<tr><td>0.7559</td><td>-0.5648</td><td>0.0825</td><td>0.1029</td><td>-0.0187</td></tr>
<tr><td>0.0186</td><td>-0.0033</td><td>-0.0048</td><td>0.0027</td><td>-0.0091</td></tr>
</table>
<table>
<tr><th colspan="6">Analysis of Variance</th></tr>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>5</td><td>302.92267</td><td>60.58453</td><td>20.89</td><td><.0001</td></tr>
<tr><td>Error</td><td>56</td><td>162.44680</td><td>2.90084</td><td></td><td></td></tr>
<tr><td>Corrected Total</td><td>61</td><td>465.36948</td><td></td><td></td><td></td></tr>
</table>
<table>
<tr><td>Root MSE</td><td>1.70318</td><td>R-Square</td><td>0.6509</td></tr>
<tr><td>Dependent Mean</td><td>26.89710</td><td>Adj R-Sq</td><td>0.6198</td></tr>
<tr><td>Coeff Var</td><td>6.33222</td><td></td><td></td></tr>
</table>
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>162.44680</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>213.72110</td></tr>
</table>
After deleting Observation 22:
<table>
<tr><th colspan="6">Analysis of Variance</th></tr>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>5</td><td>300.28355</td><td>60.05671</td><td>21.60</td><td><.0001</td></tr>
<tr><td>Error</td><td>55</td><td>152.89764</td><td>2.77996</td><td></td><td></td></tr>
<tr><td>Corrected Total</td><td>60</td><td>453.18119</td><td></td><td></td><td></td></tr>
</table>
<table>
<tr><td>Root MSE</td><td>1.66732</td><td>R-Square</td><td>0.6626</td></tr>
<tr><td>Dependent Mean</td><td>26.84033</td><td>Adj R-Sq</td><td>0.6319</td></tr>
<tr><td>Coeff Var</td><td>6.21200</td><td></td><td></td></tr>
</table>
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>152.89764</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>203.93229</td></tr>
</table>
</td>
<td>Observation 22: This is an extreme outlier since the R-student value of 3.2703 is higher than the cutoff value of 3. The observation should thus be considered to be removed from the model. The Cov Ratio value of 0.0739 is lower than the cutoff threshold of $1 - (3p/n) = 1 - ((3*16)/62) = 0.2259$ so this will not affect the variance of the regression coefficients

Observation 25: The Cov Ratio value of 0.417 is higher than the cutoff threshold of $1 - (3p/n) = 1 - ((3*16)/62) = 0.2259$ so this will affect the variance of the regression coefficients. It is only a mild outlier since under the cutoff of 2.5 for the R-student value.

The Hat Diagonal threshold is $(2p/n) = ((2*16)/62) = 0.5161$. So, both of these outliers have Hat Diagonal values that are less than the threshold and will not have a large effect on the regression model.

These two points do not have an absolute DFFITS value that is greater than the cutoff of $2((p/n)^{(½)}) = 2((16/62)^{(½)}) = 1.016$ which suggests that neither outlier has a significant influence on the model’s fitted values.

The cutoff value for DFBETAs is $2/sqrt(n)=2/sqrt(62)=0.254$, which was exceeded at observations 1,6,38,53, but not 22 or 25.

The PRESS statistic improves to give better model prediction when we remove Observation 22 from 213 to 203. There is also an increase in fit (with reduction of MSE to 1.66 and adjusted R-square to 0.6319).</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

Pamela Mishaw (IL)

To identify points of influence that may need to be removed from the data set of the counties in Illinois. The data points are evaluated based on the values of the R-student statistic, the hat diagonal, Cov Ratio, and DFFITS.

```
proc reg; by state;

model obesity = alz
mental_distress top10
Med_age sixtyfiveandup
Smoking_Rate
physical_inactivity
Diabetes Heart_Disease
Cancer Glyphosates
NATA_Cancer11
Fine_PM Mercury
Lead/influence;
run;
```

To remove counties 58 and 68:

```
if z = 183 then delete;
 if z = 193 then delete;
```

Outliers:

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS
58	2.8046	2.6373	0.1409	0.3987	1.0678	-0.2806
68	-2.8281	-2.7856	0.2101	0.3773	-1.4364	-0.1727

Before data point removal:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	114.35096	7.62340	5.42	<.0001
Error	86	121.03817	1.40742		
Corrected Total	101	235.38913			

Root MSE	1.18635	R-Square	0.4858
Dependent Mean	29.22353	Adj R-Sq	0.3961
Coeff Var	4.05956		

Sum of Residuals	0
Sum of Squared Residuals	121.03817
Predicted Residual SS (PRESS)	179.96536

After Removal of County 58 (z = 183):

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	119.39969	7.95998	6.05	<.0001
Error	85	111.88303	1.31627		
Corrected Total	100	231.28272			

Root MSE	1.14729	R-Square	0.5162
Dependent Mean	29.20356	Adj R-Sq	0.4309
Coeff Var	3.92859		

Sum of Residuals	0
Sum of Squared Residuals	111.88303
Predicted Residual SS (PRESS)	168.17358

After Removal of County 68 (z=193):

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	113.98191	7.59879	5.82	<.0001
Error	85	110.91319	1.30486		
Corrected Total	100	224.89510			

Root MSE	1.14231	R-Square	0.5068
Dependent Mean	29.25545	Adj R-Sq	0.4198
Coeff Var	3.90459		

Sum of Residuals	0
Sum of Squared Residuals	110.91319
Predicted Residual SS (PRESS)	166.12449

After Removing Both 58 and 68:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	118.24149	7.88277	6.45	<.0001
Error	84	102.67577	1.22233		
Corrected Total	99	220.91726			

Root MSE	1.10559	R-Square	0.5352
Dependent Mean	29.23560	Adj R-Sq	0.4522
Coeff Var	3.78166		

Of the 102 counties, two outliers (data points 58 and 68) were identified with absolute R-student values greater than 2.5. The Hat Diagonal threshold is (2p/n) = ((2*16)/102) = 0.314. Both of these outlier counties have Hat Diagonal values that exceed the threshold and thus are identified as having a large effect on the regression model. The Cov Ratio threshold is 1 - (3p/n) = 1 - ((3*16)/102) = 0.529. Point 68 is the only county of the two outliers which has a value less than the threshold which suggests that removing the point would reduce variability and thus improve the model. Neither of the points have an absolute DFFITS value that exceeds the threshold of 2((p/n)^(½)) = 2((16/102)^(½)) = 0.792 which suggests that neither outlier has a significant influence on the model's fitted values.

The removal of county 58 improved the model fit with a decrease in MSE from 1.407 to 1.316 and an increase in the adjusted R-square value from 0.3961 to 0.4309. This also improved the prediction capability of the model by decreasing the PRESS statistic from 179.965 to 168.174. The removal of county 68 showed a greater improvement in the fit with a decrease in MSE from 1.407 to 1.305 and increase in adjusted R-square from 0.3961 to 0.4198. The predictive ability of the model also is increased more than in the case of removing county 58 with a decrease in the PRESS statistic from 179.965 to 166.124. Removing both 58 and 68 shows the best increase in fit (MSE of 1.222 and adjusted R-square of 0.4522) and predictive ability (PRESS statistic of 155.596).

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr><td></td><td></td><td></td><td><table><tr><td>Sum of Residuals</td><td>0</td></tr><tr><td>Sum of Squared Residuals</td><td>102.67577</td></tr><tr><td>Predicted Residual SS (PRESS)</td><td>155.59553</td></tr></table></td><td></td></tr>
<tr><td>Jorge Sanchez (CA)</td><td>After conducting a thorough diagnostic analysis on my regression model with 58 counties in Texas, I focused on identifying potential, influential records that significantly impacted the model's performance. Employing various diagnostic tools, including R Student, the diagonals of the hat matrix, and the covariance ratio, I thoroughly evaluated each record's influence.</td><td>proc reg;
model obesity_age_adj = alz_ageadj_rate mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY /influence;
run;

//deleting records for /testing
if z = 38 then delete;
if z = 15 then delete;</td><td><table><tr><th>Obs</th><th>Residual</th><th>RStudent</th><th>Hat Diag H</th><th>Cov Ratio</th></tr><tr><td>15</td><td>-1.0420</td><td>-1.8110</td><td>0.8089</td><td>2.2456</td></tr><tr><td>38</td><td>-4.0671</td><td>-4.2519</td><td>0.2952</td><td>0.0060</td></tr></table></td><td>Record 38, exhibited a substantial negative studentized residual value of -4.2519, indicating its significant influence on the model as a potential outlier. I reviewed its hat diagonal (0.2952), but it did not exceed the threshold (2 (16/58) = 0.5517). In addition, I checked the covariance ratio (0.0060) and it is below the threshold of 0.1724 (1 – 3(16)/58)). Additionally, I assessed record 15, which had a hat diagonal value of 0.8089 (exceeded our threshold of 0.5517) , a covariance ratio of 2.24456, and a relatively low R student value of -1.811. To ascertain the model's sensitivity to these influential records, I performed several iterations of the regression. Firstly, I removed record 15 (Kern County) and observed a slight increase in the Adjusted R-Square to 0.8558. Subsequently, upon excluding record 38 (San Francisco County), the model showed further improvements, yielding an Adjusted R-Square of 0.8836, a lower MSE of 0.29814, and a reduced PRESS value of 118.83173. Considering the limited size of our dataset and the improved goodness-of-fit metrics, I concluded that removing additional records was unnecessary.</td></tr>
<tr><td>Daniel Wilson (TX)</td><td>In Texas, there are 254 counties. If there are any counties whose influence significantly disrupts the model, I want to remove those points. There are many diagnostic tools, including $R^2_{Studentiezed}$, the diagonals of the hat matrix, the covariance ratio, DFFITS, and DFBETAS.</td><td>proc reg;
model Obesity = Alz MentalDistress Top10 MedAge Sixtyfiveplus Smoking PhysInac Diabetes HeartDisease Cancer Glyphosates NATACancer Finepm Mercury Lead /influence;
run;

To remove county 129:
<pre>z=_n_;
if z=129 then delete;</pre></td><td><table><tr><th>Obs</th><th>Residual</th><th>RStudent</th><th>Hat Diag H</th><th>Cov Ratio</th><th>DFFITS</th></tr><tr><td>129</td><td>2.8295</td><td>3.4115</td><td>0.3000</td><td>0.7097</td><td>2.2335</td></tr></table></td><td>Though there are many approaches I could have taken, amongst 254 counties, I am choosing only to remove one. There were only six outlier counties with $\left|R^2_{stud}\right| > 2.5$. Of these six, only one was high-influence as determined by the Hat diagonal value. The threshold of $2\frac{p}{n} = 2\frac{16}{254} = 0.126$ was surpassed by county 129, with a staggering Hat diagonal value of 0.3. This was the only outlier that was high influence, and its removal increased the overall R^2_{adj} for the full model from .577 to .588. It also decreased the PRESS for the full model from 286 to 266. While overall that's not a large jump, considering it was one point out of 254, it does warrant removal in my analysis.</td></tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>First, I will evaluate the influence of each of the 67 counties in Florida by producing the RStudent values, Hat-Diagonal value, and covariance ratio. Based on these diagnostics, I will decide if any observations should be removed from the data.</td>
<td>proc reg;
where state='FL';
model obesity_age_adj = alz_ageadj_rate mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY/ influence;
run;</td>
<td><table>
<tr><th>Obs</th><th>Residual</th><th>RStudent</th><th>Hat Diag H</th><th>Cov Ratio</th><th>DFFITS</th></tr>
<tr><td>29</td><td>-3.3903</td><td>-3.6023</td><td>0.4176</td><td>0.0767</td><td>-3.0505</td></tr>
<tr><td>65</td><td>3.4734</td><td>2.9632</td><td>0.1558</td><td>0.1463</td><td>1.2731</td></tr>
</table></td>
<td>I picked out the 29th and 65th observation because their RStudent values were greater than the critical value of 2.5. We will use the Hat-Diagonal value and covariance ratio to decide if they should be removed.
Our critical value for the Hat-Diagonal value = 2(p/n) = 2(16/67) = 0.4776. The Hat-Diagonal values for observation 29 (0.4176) and observation 65 (0.1158) do not exceed this threshold.
Our critical value for the covariance ratio = 1 - (3p/n) = 1 - ((3)(16)/67) = 0.2836. Our covariance ratios for observation 29 (0.0767) and observation 65 (0.1463) do not exceed this threshold.
Since both of these observations did not exceed the critical value thresholds of their Hat-Diaginal and covariance ratio, I will not be removing them from the data. I’m also hesitant to remove any observations since we only have a sample size of 67.</td>
</tr>
</table>

2. Run SAS to check multicollinearity that may exist:

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Brad Lipson (NY)</td>
<td>To find variable inflation factors and assess predictor multicollinearity. To help us identify the variables affecting obesity rates and assess collinearity, which is key for our model's accuracy.</td>
<td>proc reg;
model
obesity_age_adj =
alz_ageadj_rate
mental_distress
top10
Med_age
sixtyfiveandup
Smoking_Rate
physical_inactivity
Diabetes
Heart_Disease
Cancer
Glyphosates
NATA_Cancer_11
Fine_PM_2_5
Mercury_TPY
Lead_TPY /vif collin;
run;
proc corr;
….run;</td>
<td><table>
<tr><th>Variance Inflation</th></tr>
<tr><td>0</td></tr>
<tr><td>1.86272</td></tr>
<tr><td>3.61361</td></tr>
<tr><td>1.42477</td></tr>
<tr><td>8.77557</td></tr>
<tr><td>7.11275</td></tr>
<tr><td>6.14169</td></tr>
<tr><td>2.56915</td></tr>
<tr><td>2.21804</td></tr>
<tr><td>2.52482</td></tr>
<tr><td>2.56405</td></tr>
<tr><td>2.36268</td></tr>
<tr><td>3.71149</td></tr>
<tr><td>2.40075</td></tr>
<tr><td>1.76399</td></tr>
</table></td>
<td>Lead and Mercury and Sixtyfiveandup is strongly correlated with Med_age. Also, Physical inactivity is correlated with sixtyfiveandup and smoking rate. The final model should eliminate these variables.

All of the variance inflation values are less than 20, showing that there is no multicollinearity in this full model.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>To examine the 16 variables for any multicollinearity that may exist. This allows for a reduction in the size of the model used to predict obesity in the state. This will be done by looking at the variance inflation values, eigenvalues, and condition indexes as well as the correlation values amongst the variables.</td>
<td>proc reg; by state;

model obesity = alz mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer11 Fine_PM Mercury Lead/vif collin;
run;

proc corr; by state;
var alz mental_distress top10 Med_age sixtyfiveandup obesity Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer11 Fine_PM Mercury Lead;
run;</td>
<td>
<table>
<tr><th>Variable</th><th>Variance Inflation</th></tr>
<tr><td>Intercept</td><td>0</td></tr>
<tr><td>alz</td><td>2.46888</td></tr>
<tr><td>mental_distress</td><td>2.32644</td></tr>
<tr><td>top10</td><td>2.45254</td></tr>
<tr><td>Med_age</td><td>7.86446</td></tr>
<tr><td>sixtyfiveandup</td><td>8.56391</td></tr>
<tr><td>Smoking_Rate</td><td>1.73554</td></tr>
<tr><td>physical_inactivity</td><td>2.31006</td></tr>
<tr><td>Diabetes</td><td>1.73625</td></tr>
<tr><td>Heart_Disease</td><td>1.49385</td></tr>
<tr><td>Cancer</td><td>1.54159</td></tr>
<tr><td>Glyphosates</td><td>1.37242</td></tr>
<tr><td>NATA_Cancer11</td><td>1.59634</td></tr>
<tr><td>Fine_PM</td><td>1.33099</td></tr>
<tr><td>Mercury</td><td>2.31451</td></tr>
<tr><td>Lead</td><td>2.41470</td></tr>
</table>
</td>
<td>Based on the variance inflation values alone, there appears to be no multicollinearity in the full model since none of the VIF values are greater than 20.

Using 100 as the threshold value for condition indices, there appears to be potential collinearity at numbers 15 and 16 (with values of about 112 and 156). According to the table, Fine_PM is the only variable that contributes most to the possible collinearity of number 15 with a proportion of variation of 0.71645. The variables med_age and sixtyfiveandup have the highest and only significant proportions of variation for number 16 (0.744 and 0.60771, respectively). This suggests that these variables should not be included together in the regression model. When sixtyfiveandup is removed, the only significant collinearity is between the intercept value and Fine_PM in number 15. I chose not to remove Fine_PM due to possible loss of valuable information.

The correlation coefficient table showed that, of the regressor variables, physical inactivity had the strongest correlation with the response variable obesity (r = 0.58926, p = <0.001). Smoking rate and diabetes showed weakly positive correlations with obesity (r = 0.41296 [p = <0.0001] and r = 0.43779 [p = <0.0001], respectively). This suggests that these variables should be included in the ideal regression model.

There is a strong correlation between sixtyfiveandup and Med_age (r = 0.87885, p = <0.0001 as well as between Lead and Mercury (r = 0.70259, p = <0.0001). There are moderate correlations between physical_inactivity and sixtyfiveandup (r = 0.54243, p = <0.0001) and between physical_inactivity and smoking rate (r = 0.52397, p = <0.0001). This indicates that these variable pairs should be avoided in the regression model.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr><td></td><td></td><td></td><td>
<table>
<tr><th>Number</th><th>Eigenvalue</th><th>Condition Index</th></tr>
<tr><td>1</td><td>12.91564</td><td>1.00000</td></tr>
<tr><td>2</td><td>1.49435</td><td>2.93989</td></tr>
<tr><td>3</td><td>0.87449</td><td>3.84310</td></tr>
<tr><td>4</td><td>0.27465</td><td>6.85760</td></tr>
<tr><td>5</td><td>0.24166</td><td>7.31063</td></tr>
<tr><td>6</td><td>0.10638</td><td>11.01858</td></tr>
<tr><td>7</td><td>0.03596</td><td>18.95238</td></tr>
<tr><td>8</td><td>0.02358</td><td>23.40596</td></tr>
<tr><td>9</td><td>0.00910</td><td>37.66578</td></tr>
<tr><td>10</td><td>0.00655</td><td>44.40984</td></tr>
<tr><td>11</td><td>0.00596</td><td>46.53741</td></tr>
<tr><td>12</td><td>0.00508</td><td>50.44032</td></tr>
<tr><td>13</td><td>0.00318</td><td>63.69822</td></tr>
<tr><td>14</td><td>0.00188</td><td>82.78495</td></tr>
<tr><td>15</td><td>0.00101</td><td>112.87072</td></tr>
<tr><td>16</td><td>0.00052403</td><td>156.99322</td></tr>
</table>
After removing the redundant variable sixtyfiveandup:
<table>
<tr><th>Number</th><th>Eigenvalue</th><th>Condition Index</th></tr>
<tr><td>1</td><td>11.95292</td><td>1.00000</td></tr>
<tr><td>2</td><td>1.48663</td><td>2.83554</td></tr>
<tr><td>3</td><td>0.86962</td><td>3.70743</td></tr>
<tr><td>4</td><td>0.27144</td><td>6.63589</td></tr>
<tr><td>5</td><td>0.24004</td><td>7.05664</td></tr>
<tr><td>6</td><td>0.10366</td><td>10.73835</td></tr>
<tr><td>7</td><td>0.03494</td><td>18.49578</td></tr>
<tr><td>8</td><td>0.01293</td><td>30.40513</td></tr>
<tr><td>9</td><td>0.00795</td><td>38.77564</td></tr>
<tr><td>10</td><td>0.00605</td><td>44.46332</td></tr>
<tr><td>11</td><td>0.00525</td><td>47.73636</td></tr>
<tr><td>12</td><td>0.00350</td><td>58.46725</td></tr>
<tr><td>13</td><td>0.00243</td><td>70.20669</td></tr>
<tr><td>14</td><td>0.00187</td><td>79.99114</td></tr>
<tr><td>15</td><td>0.00080344</td><td>121.97204</td></tr>
</table>
</td><td></td></tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>This section aims to create a multiple linear regression model to investigate the relationship between 'obesity_age_adj' and several predictor variables, including health, demographic, and environmental factors. The '/vif collin' option included in the model also allows for assessing multicollinearity among these predictors. This analysis will enable us to understand better the significant factors influencing obesity rates and evaluate the collinearity issues, which is crucial for the accuracy of our model</td>
<td>proc reg;
model obesity_age_adj =
alz_ageadj_rate
mental_distress top10
Med_age sixtyfiveandup
Smoking_Rate
physical_inactivity Diabetes

Heart_Disease Cancer
Glyphosates NATA_Cancer_11
Fine_PM_2_5 Mercury_TPY
Lead_TPY /vif collin;
run;

proc corr;
var obesity_age_adj
alz_ageadj_rate
mental_distress top10
Med_age sixtyfiveandup
Smoking_Rate
physical_inactivity Diabetes

Heart_Disease Cancer
Glyphosates NATA_Cancer_11
Fine_PM_2_5 Mercury_TPY
Lead_TPY;
run;</td>
<td><table>
<tr><th>Variance Inflation</th><th>Condition Index</th></tr>
<tr><td>0</td><td>1.00000</td></tr>
<tr><td>2.73424</td><td>3.04141</td></tr>
<tr><td>3.96857</td><td>3.56332</td></tr>
<tr><td>1.33478</td><td>4.08301</td></tr>
<tr><td>19.06908</td><td>6.61638</td></tr>
<tr><td>15.56891</td><td>9.95389</td></tr>
<tr><td>6.25686</td><td>14.86094</td></tr>
<tr><td>5.18918</td><td>20.01768</td></tr>
<tr><td>5.02714</td><td>24.22132</td></tr>
<tr><td>3.07961</td><td>29.78647</td></tr>
<tr><td>3.94116</td><td>40.84299</td></tr>
<tr><td>3.71619</td><td>44.08149</td></tr>
<tr><td>2.94464</td><td>58.92839</td></tr>
<tr><td>2.94615</td><td>86.31406</td></tr>
<tr><td>2.63065</td><td>106.20323</td></tr>
<tr><td>3.30435</td><td>147.78131</td></tr>
</table></td>
<td>We set a VIF threshold of > 20 based on Dr. Amin's recommendation to identify potential collinearity issues. Although most variables were within the limit, Med_age approached the threshold with a VIF value of 19.06908, assuring further examination. I reviewed the Condition index value and identified Mercury_TPY and Lead_TPY with values exceeding the 100 thresholds of 106.2032 and 147.7813, respectively. Additionally, while examining the correlation results, we identified pairs of independent variables with high correlation values: Med_age and sixtyfiveandup (0.94516), Cancer and Smoking_Rate (0.60510), physical_inactivity and Diabetes (0.76905), and physical_inactivity and Heart_Disease (0.70104). These correlations indicate possible multicollinearity concerns. We may consider dropping one variable from each correlated pair or exploring composite variables to address this.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>16 regressors is A LOT. I want to build a smaller, more efficient model that predicts obesity in TX. So, I must first Identify if any multicollinearity exists to ensure that the model we select to fit and predict obesity contains regressors that are not significantly dependent on one another..</td>
<td>proc reg;
model Obesity = Alz
MentalDistress Top10 MedAge
Sixtyfiveplus Smoking PhysInac
Diabetes HeartDisease Cancer
Glyphosates NATACancer
Finepm Mercury Lead
/vif collin;
run;

For the correlation matrix:
proc corr;
run;</td>
<td><table>
<tr><th>Variance Inflation</th><th>Condition Index</th></tr>
<tr><td>0</td><td>1.00000</td></tr>
<tr><td>2.00990</td><td>3.20238</td></tr>
<tr><td>1.43678</td><td>3.66885</td></tr>
<tr><td>1.82231</td><td>4.03765</td></tr>
<tr><td>6.21470</td><td>4.18698</td></tr>
<tr><td>5.32517</td><td>9.09047</td></tr>
<tr><td>2.57611</td><td>13.83283</td></tr>
<tr><td>2.14212</td><td>17.15505</td></tr>
<tr><td>1.92995</td><td>24.69708</td></tr>
<tr><td>1.36695</td><td>28.90774</td></tr>
<tr><td>1.48347</td><td>35.10927</td></tr>
<tr><td>1.42740</td><td>39.88819</td></tr>
<tr><td>1.87488</td><td>60.04024</td></tr>
<tr><td>1.17606</td><td>70.16688</td></tr>
<tr><td>1.10801</td><td>84.26609</td></tr>
<tr><td>1.56277</td><td>126.06104</td></tr>
</table></td>
<td>While the variance inflation does not indicate any multicollinearity, the condition index is potentially problematic (126>100). Scrolling over to the right, the greatest proportion of variation associated with this smallest eigenvalue is from the regressor, Finepm. This proportion of variation is 0.41 corresponding to $\lambda_{16} = 0.00076$. As a result, I will choose not to include Finepm.

It is worth noting that the correlation matrix revealed correlation coefficients > 0.5 for the following variable pairs: Top10 & Alz, MedAge & 65plus, Smoking & Physical Inactivity, Smoking & Diabetes, Smoking & Cancer, Physical inactivity & Diabetes. If possible, when I select a model, it will be valuable to include few or none of these pairs.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

Name	Objective	SAS Code	Notable SAS Output(s)	Conclusion
Jolie Wise (FL)	From this point on, the top 10 regressor will not be included in my model as every county in Florida had a value of 0. Since there is no variation in the regressor, there is no reason to include it in the model. I will identify any multicollinearity between models based on their VIF values. We will also observe any correlations between pairs of regressor variables using a correlation matrix.	proc reg; where state='FL'; model obesity_age_adj = alz_ageadj_rate mental_distress Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY/ vif collin; run; proc corr; where state='FL'; run;	Variance Inflation 0 1.14807 3.36028 17.52798 17.05880 6.23627 11.84865 8.20229 2.68433 2.38527 2.03043 3.40568 1.37468 1.69447 1.93367	We have a VIF threshold of 20, and all 14 of the regressor variables had VIF values below the threshold. The highest VIF value was 17.52798. Our confusion matrix showed that the following pairs of regressors had correlation coefficients of 0.70: 65+ & Med Age (.9357) Diabetes & Physical inactivity (.87999) Smoking Rate & Physical inactivity (.70515) When selecting models, I may want to consider avoiding models containing these pairs of regressors, especially 65+ and median age.

3. Run SAS in order to identify the best model:

Name	Objective	SAS Code	Notable SAS Output(s)	Conclusion
Brad Lipson (NY)	To find the five variables that determine the best model that maximized R2adj and minimized Mallow's C(p) and MSE for fitting. To find those models for prediction that maximized R2Pred and minimized PRESS. To choose a model that accurately predicts data variability without overfitting.	proc rsquare adjrsq mse cp; model obesity_age_adj= alz_ageadj_rate mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY; run;	R-Square | Adjusted R-Square | C(p) | MSE possible models: 5 | 0.7381 | 0.7148 | 5.0466 | 2.17613 5 | 0.7312 | 0.7072 | 6.4956 | 2.23341 5 | 0.7244 | 0.6998 | 7.9304 | 2.29014 5 | 0.7242 | 0.6996 | 7.9783 | 2.29203 full model: 14 | 0.7764 | 0.7098 | 15.0000 | 2.21382	Four models with high R-Square and Adjusted R-square (over 70%) values were chosen. The full model has the the second lowest MSE, equivalent R-square and Adjusted R-square values, and the best fitting. The first 5-variable model has the lowest C(p) value, suggesting the best fit of all models. All four were equally as short with 5 variables for the possible models given. Thus, obesity = physical_inactivity Heart_Disease Diabetes Fine_PM Mercury is selected as the optimal of the examined models (first one listed).

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>The models with the best fit abilities are selected based on adjusted R-squared, mallow's Cp, and mean squared error. The prediction capabilities are determined by Predicted R-square and PRESS values.</td>
<td>proc rsquare adjrsq mse cp; by state;

model obesity = alz mental_distress top10 Med_age Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer11 Fine_PM Mercury Lead;
run;

proc reg; by state;
model obesity = alz mental_distress top10 Med_age Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer11 Fine_PM Mercury Lead/ clm p;
run;

proc reg; by state;
model obesity = alz mental_distress physical_inactivity Diabetes Mercury Lead/ clm p;
run;

proc reg; by state;
model obesity = alz mental_distress physical_inactivity Diabetes Glyphosates Mercury Lead/ clm p;
run;

proc reg; by state;
model obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates Mercury Lead/ clm p;
run;

proc reg; by state;
model obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates NATA_Cancer11 Mercury Lead/ clm p;
run;</td>
<td>Selected "best" models:
<table>
<tr><th>Number in Model</th><th>R-Square</th><th>Adjusted R-Square</th><th>C(p)</th><th>MSE</th><th>Variables in Model</th></tr>
<tr><td>6</td><td>0.4683</td><td>0.4348</td><td>1.9538</td><td>1.31735</td><td>alz mental_distress physical_inactivity Diabetes Mercury Lead</td></tr>
<tr><td>7</td><td>0.4775</td><td>0.4386</td><td>2.4039</td><td>1.30842</td><td>alz mental_distress physical_inactivity Diabetes Glyphosates Mercury Lead</td></tr>
<tr><td>8</td><td>0.4803</td><td>0.4356</td><td>3.9226</td><td>1.31529</td><td>alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates Mercury Lead</td></tr>
<tr><td>9</td><td>0.4825</td><td>0.4319</td><td>5.5595</td><td>1.32410</td><td>alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates NATA_Cancer11 Mercury Lead</td></tr>
</table>
"Full" model with 14 variables (excludes sixtyfiveandup due to previous section's results):
<table>
<tr><td>14</td><td>0.4858</td><td>0.4030</td><td>15.0000</td><td>1.39125</td><td>alz mental_distress top10 Med_age Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer11 Fine_PM Mercury Lead</td></tr>
</table>
PRESS Statistics:

Full:
<table><tr><td>Predicted Residual SS (PRESS)</td><td>176.05778</td></tr></table>
obesity = alz mental_distress physical_inactivity Diabetes Mercury Lead:
<table><tr><td>Predicted Residual SS (PRESS)</td><td>151.73217</td></tr></table>
obesity = alz mental_distress physical_inactivity Diabetes Glyphosates Mercury Lead:
<table><tr><td>Predicted Residual SS (PRESS)</td><td>150.40559</td></tr></table>
obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates Mercury Lead:
<table><tr><td>Predicted Residual SS (PRESS)</td><td>155.04045</td></tr></table>
obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates NATA_Cancer11 Mercury Lead:
<table><tr><td>Predicted Residual SS (PRESS)</td><td>157.16320</td></tr></table>
<table>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th></tr>
<tr><td>Model</td><td>9</td><td>113.57230</td></tr>
<tr><td>Error</td><td>92</td><td>121.81683</td></tr>
<tr><td>Corrected Total</td><td>101</td><td>235.38913</td></tr>
</table>
</td>
<td>The four possible models were selected based on relatively large R-Square and Adjusted R-square values. The 6-variable model is identified as the "best" fitting due to comparable R-square and Adjusted R-square values as well as the lowest MSE. The C(p) value is also lowest for this model which suggests the best fit of the four. Additionally, it happens to be the shortest model which is typically ideal.

Looking at the PRESS statistics, the 6-variable model also has the lowest value which suggests it has the least impacted prediction capability compared to most of the other models. Only the seven-variable model has a lower value and is only marginally lower. This applies to the predicted R-squared values, as well, as seen below.

Calculated predicted R-squared values:

Full: 1 - (176.05778/235.38913) = 0.2521

obesity = alz mental_distress physical_inactivity Diabetes Mercury Lead: 1 - (151.73217/235.38913) = 0.3554

obesity = alz mental_distress physical_inactivity Diabetes Glyphosates Mercury Lead: 1 - (150.40559/235.38913) = 0.3610

obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates Mercury Lead: 1 - (155.04045/235.38913) = 0.3413

obesity = alz mental_distress Smoking_Rate physical_inactivity Diabetes Glyphosates NATA_Cancer11 Mercury Lead: 1 - (157.16320/235.38913) = 0.3323

Thus, obesity = alz mental_distress physical_inactivity Diabetes Mercury Lead is selected as the optimal of the examined models.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>To find the best model, we consider five crucial metrics. For fit, we maximize Adjusted R-squared (R2adj) and minimize Mallow's C(p) and Mean Squared Error (MSE). We maximize Predictive R-squared (R2Pred) and minimize Prediction Error Sum of Squares (PRESS) for Prediction. These metrics guide us in selecting a well-fitted model with accurate predictions, ensuring its effectiveness in capturing the data's variability without overfitting.</td>
<td>proc rsquare adjrsq mse cp;
model obesity_age_adj= alz_ageadj_rate mental_distress top10 Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY;
run;</td>
<td><table>
<tr><th>Number in Model</th><th>R-Square</th><th>Adjusted R-Square</th><th>C(p)</th><th>MSE</th></tr>
<tr><td colspan="5">Full model</td></tr>
<tr><td>15</td><td>0.9148</td><td>0.8836</td><td>16.0000</td><td>1.29814</td></tr>
<tr><td colspan="5">Possible models</td></tr>
<tr><td>5</td><td>0.8775</td><td>0.8655</td><td>13.9265</td><td>1.49990</td></tr>
<tr><td>6</td><td>0.8898</td><td>0.8766</td><td>10.0271</td><td>1.37674</td></tr>
<tr><td>7</td><td>0.8949</td><td>0.8799</td><td>9.5804</td><td>1.34001</td></tr>
<tr><td>8</td><td>0.9042</td><td>0.8883</td><td>7.0803</td><td>1.24623</td></tr>
</table></td>
<td>The model that includes Med_age, physical_inactivity, cancer, fine_PM_2_5, and Lead_TPY stands out as the most effective predictor for obesity age-adjusted. An impressive Adjusted R-Square of 0.8655 explains approximately 86.55% of the variability in the response variable. The high Predicted R-Square value of 0.8775 signifies its reliability in making accurate predictions on unseen data. The model's residuals display robustness, evident from the low PRESS value of 93.4783 and MSE of 1.499, indicating its ability to predict actual values closely. Even though Mallow's Cp value is 13.9265, showing a good balance between prediction error and model complexity, The Model outperforms other models in Adjusted R-Square, Predicted R-Square, PRESS, and MSE, affirming its superiority in predicting obesity age-adjusted. Moreover, after conducting a multicollinearity test, we carefully considered this model, ensuring that physical_inactivity is not paired with diabetes or heart_disease, addressing potential multicollinearity concerns.
obesity_age_adj = Med_age physical_inactivity cancer fine_PM_2_5 Lead_TPY</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>Five metrics are helpful in determining the best model. For fit, it is helpful to maximize the R^2_{adj} and to minimize Mallow’s C(p) and the Mean Squared Error. For prediction, it is helpful to maximize R^2_{Pred} and minimize the PRESS statistic.</td>
<td>proc rsquare adjrsq cp mse;
model Obesity = Alz MentalDistress Top10 MedAge Sixtyfiveplus Smoking PhysInac Diabetes HeartDisease Finepm Cancer Glyphosates NATACancer Mercury Lead;
run;

To find PRESS statistics:
proc reg;
model Obesity = [testing different models to get PRESS statistics]</td>
<td><table>
<tr><th>Number in Model</th><th>R-Square</th><th>Adjusted R-Square</th><th>C(p)</th><th>MSE</th></tr>
<tr><td colspan="5">Full Model:</td></tr>
<tr><td>15</td><td>0.6130</td><td>0.5885</td><td>16.0000</td><td>0.98282</td></tr>
<tr><td colspan="5">Possible models for selection:</td></tr>
<tr><td>11</td><td>0.6105</td><td>0.5927</td><td>9.4956</td><td>0.97260</td></tr>
<tr><td>9</td><td>0.6073</td><td>0.5927</td><td>7.4958</td><td>0.97269</td></tr>
<tr><td>6</td><td>0.5944</td><td>0.5845</td><td>9.3681</td><td>0.99228</td></tr>
<tr><td>4</td><td>0.5787</td><td>0.5719</td><td>14.9658</td><td>1.02231</td></tr>
</table>
The PRESS statistic in the full model was 266. For the subsequent models, the PRESS statistic never dipped below 259. they are all quite similar.</td>
<td>In consideration of all these factors, I eliminate the 4 regressor model due to low R^2_{adj} and higher C(p). I also eliminate the 11 regressor model because you can get the same R^2_{adj} with a lower C(p) with 9 regressors. According to my multicollinearity assessment, Finepm contributed most to the 126 condition index. And while the 9 regressor model has the lowest Cp and highest R^2_{adj}, it has Finepm as a regressor. The 6 regressor model does not. For a tradeoff of less than 0.01 in the R^2_{adj}, a C(p) within 2, and having 3 fewer regressors, this is a worthwhile trade. The model is…

Obesity = MentalDistress, Smoking, PhysInac, Diabetes, Mercury, Lead.

Unfortunately, it was not possible to separate the pairwise variables mentioned in the previous section. They are the primary variables related to obesity!</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>I will select the four models with the best chance of being the best model for fitting and predicting the obesity rate in Florida. I will use adjusted R-Square, MSE, and Mallow's Cp to evaluate fitting. I will use the PRESS Statistic to evaluate prediction.

I will also include these values for the full model for comparison.</td>
<td>proc rsquare adjrsq mse cp;
where state='FL';
model obesity_age_adj = alz_ageadj_rate mental_distress Med_age sixtyfiveandup Smoking_Rate physical_inactivity Diabetes Heart_Disease Cancer Glyphosates NATA_Cancer_11 Fine_PM_2_5 Mercury_TPY Lead_TPY;
run;

For PRESS & SSTotal:
proc reg;
where state='FL';
model obesity_age_adj = [each model]/CLM;</td>
<td>
<table>
<tr><th>Number in Model</th><th>R-Square</th><th>Adjusted R-Square</th><th>C(p)</th><th>MSE</th></tr>
<tr><td>5</td><td>0.9147</td><td>0.9077</td><td>12.4784</td><td>2.06998</td></tr>
<tr><td>6</td><td>0.9213</td><td>0.9134</td><td>9.2653</td><td>1.94190</td></tr>
<tr><td>7</td><td>0.9282</td><td>0.9197</td><td>5.8313</td><td>1.80247</td></tr>
<tr><td>8</td><td>0.9314</td><td>0.9219</td><td>5.3165</td><td>1.75241</td></tr>
<tr><td>14</td><td>0.9343</td><td>0.9166</td><td>15.0000</td><td>1.87125</td></tr>
</table>
5:
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>126.26876</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>160.42841</td></tr>
</table>
6:
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>116.51371</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>148.44733</td></tr>
</table>
7:
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>106.34544</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>145.46298</td></tr>
</table>
8:
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>101.63950</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>141.80600</td></tr>
</table>
14:
<table>
<tr><td>Sum of Residuals</td><td>0</td></tr>
<tr><td>Sum of Squared Residuals</td><td>97.30483</td></tr>
<tr><td>Predicted Residual SS (PRESS)</td><td>171.10874</td></tr>
</table>
</td>
<td>These are the four chosen models:

5:alz_ageadj_rate, sixtyfiveandup, physical_inactivity, Diabetes, Lead_TPY
6:alz_ageadj_rate, sixtyfiveandup, physical_inactivity, Diabetes, NATA_Cancer_11, Lead_TPY
7:alz_ageadj_rate, Smoking_Rate, physical_inactivity, Diabetes, Heart_Disease, NATA_Cancer_11, Lead_TPY
8:alz_ageadj_rate, Smoking_Rate, physical_inactivity, Diabetes, Heart_Disease, Glyphosates, NATA_Cancer_11, Lead_TPY

The model that appears to be the best for fitting and predicting while also having fewer regressors is the model with the following seven variables: alz_ageadj_rate, Smoking_Rate, physical_inactivity, Diabetes, Heart_Disease, NATA_Cancer_11, Lead_TPY. I was choosing between this model and the model with eight regressors as they performed very similarly. They both had the lowest Mallow's Cp (7: 5.8313 & 8: 5.3165), MSE (7: 1.80247 & 8: 1.75241), and PRESS statistics (7: 145.4630 & 8: 141.8060). They also had the highest adjusted R-Square values (7: 0.9197 & 8: 0.9219).

The model with eight regressors performed marginally better on the performance metrics; however, the improvements were not large enough to warrant including another regressor and risk overcomplicating the model.</td>
</tr>
</table>

4. Run PROC IML in SAS to perform a matrix application:

STA6235: Summer 2023 Group 4 Alzheimer's Project

Name	Objective	SAS Code	Notable SAS Output(s)	Conclusion
Brad Lipson (NY)	Since I am interested in looking for the influence of mercury on obesity, I used PROC IML. I started with two vectors, x and y, where x is the independent variable (mercury) with a fixed term and y is the dependent variable (obesity, adjusted for age). I found the regression coefficients (b) and the fitting values (yhat) by using matrix operations. This matrix operation gave me the sum of squared residuals (SSE), the sum of squared regression (ssr), and the total sum of squares (SSTO), which then gave the mean squared error (mse). Overall, the MSE showed how well the model worked.	proc iml; x={1 25.52,1 28.88, 1 28.1,1 27.32,1 28.14, 1 26.16,1 28,...}; y={0.085829488, 0.002864658, 0.038681403, 0.006773176, 0.003269837,..}; n={62}; Q=j(n,{1}, {1}); /* a vector of ones */ id=I(n); ….. sse=e`*e; SSR=y`*(h-q*q`/n)*y; SST=(y`*y)-(1/n)*(y`*q*q`*y); mse=sse/(n-2); yhat=h*y; hh=h*h ; print h, hh, b, yhat, e, sum_e , sse, SSR, sst, mse xpy xpx xpx_1; run;	**b** 0.147159 -0.004785 **yhat** 0.0250359 0.008957 0.0126896 0.0164222 0.0124981 **mse** 0.0006116 **ssto** 0.047351	The beta hat matrix shows that the expected obesity rate when the mercury level is zero in a county is 0.147159, and for every unit of lead level, the obesity rate decreases by 0.004785. The data fits the model well since the total of residuals is approaching zero. The regression model fits the data well, and the MSE is almost zero.

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>I will use proc iml to calculate various matrices/vectors needed for regression model building and analysis. I used obesity as my response variable and lead as my regressor variable as I was interested in the possible influence of lead on metabolism which may increase rates of obesity.</td>
<td>proc iml;
x = {1 0.097970701,1 0.032503659, 1 0.06448189,..., 1 0.105597407, 1 0.671436459, 1 0.005522918};

y = {31.24,31.2,28.62,...,30.76, 29, 29.22};

n={102};
Q=j(n,{1}, {1}); /* a vector of ones */

id=I(n);

xpx=x`*x;<br>ypy=(y`)*(y);

xpx_1=inv(xpx);

xpy=x`*y;<br><br>b=(xpx_1)*xpy;<br><br>h=x*xpx_1*x`;
ih=id-h;

yhat=h*y;

e=(id-h)*y;

sum_e=e`*Q;<br><br>sse=e`*e;

SSR=y`*(h-q*q`/n)*y;

SST=(y`*y)-(1/n)*(y`*q*q`*y);
mse=sse/(n-2);
yhat=h*y;

hh=h*h ; /* check if H is idempotent or not */
print h, hh, b, yhat, e, sum_e , sse, SSR, sst, mse xpy xpx xpx_1;
run;</td>
<td>b
29.3901
-0.740794

yhat
29.317524
29.366022
29.342332
29.246032
29.383315
29.382553
.
.
.
29.311874
28.892704
29.386009

e
1.9224759
1.8339783
-0.722332
.
.
.
1.4481257
0.1072958
-0.166009

sum_e
-2.23E-13

sse
219.82989

SSR
15.55924

SST
235.38913

<table>
<tr><th>mse</th><th>xpy</th><th>xpx</th><th></th><th>xpx_1</th><th></th></tr>
<tr><td>2.1982989</td><td>2980.8</td><td>102</td><td>22.935152</td><td>0.0115872</td><td>-0.007931</td></tr>
<tr><td></td><td>649.24261</td><td>22.935152</td><td>33.509727</td><td>-0.007931</td><td>0.0352701</td></tr>
</table></td>
<td>According to the beta hat matrix, the model can be written as y hat = 29.3901 - 0.740794x which indicates that the predicted obesity rate when the lead level equals zero in a county is 29.3901 and that for every increase in unit of lead level there is a decrease in obesity rate by 0.740794. The sum of residuals is very small value nearing zero which suggests that the data fits the model well. The MSE is also relatively small at a little more than 2 which also suggests a good fit of the data to the regression model.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>Using the PROC IML (Interactive Matrix Language) procedure, I performed a simple linear regression in SAS. The data is provided in two vectors, x, and y, where x represents the independent variable (physical inactivity) with a constant term, and y represents the dependent variable (obesity age adjusted). Using matrix operations, I calculated the regression coefficients (b) and fitted values (yhat). Additionally, it computes the sum of squared residuals (SSE), the sum of squared regression (ssr), and the total sum of squares (SSTO), which are later used to calculate the mean squared error (mse) for evaluating the model's performance.</td>
<td>proc iml;
x = {1 15.9, 1 17.9, 1 16.24, 1 16.46, 1 17.64, 1 16.5, 1 16.82, 1 17.92, 1 14.14, 1 20,
1 17.58…..}
y = {19.9, 23.56, 24.74, 24.58, 24.8, 23.34, 23.3, 26.64, 19.82, 28.74,
27.16,….}

n={57};
Q=j(n,{1}, {1});

Id=I(n);
….
ssr=y`*(h-q*q`*y/n);

ssto=(y`*y)-(y`*q*q`*y/n)
;
mse=sse/(n-2);

print h, b, yhat, e, sse, ssr, ssto, mse;
run;</td>
<td>
<table><tr><th>b</th></tr><tr><td>5.5298067</td></tr><tr><td>1.0700295</td></tr></table>
<table><tr><th>yhat</th></tr><tr><td>22.543276</td></tr><tr><td>24.683335</td></tr><tr><td>22.907086</td></tr><tr><td>23.142492</td></tr></table>
<table><tr><th>ssto</th></tr><tr><td>624.62396</td></tr></table>
<table><tr><th>mse</th></tr><tr><td>3.5134962</td></tr></table>
</td>
<td>The estimated intercept in the model is 5.52981, indicating the expected value of the obesity age-adjusted variable when physical inactivity is zero. Physical inactivity has a significant positive effect on obesity age-adjusted, with an estimated coefficient of 1.07003, suggesting that for each unit increase in physical inactivity, we expect an increase of approximately 1.07 units in obesity age-adjusted. SSTotal is 624.62396, representing the total variability in the dependent variable, obesity age-adjusted. MSE is 3.51350, reflecting the average squared difference between the predicted and actual values in the model, indicating the level of unexplained variability.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>Proc iml allows SAS to do matrix mathematics efficiently and effectively. It is another tool that can help us arrive at meaningful conclusions. Here, I use matrix mathematics to compute some summary statistics for the regression of obesity on Alzheimer’s. The code also calculates the predicted obesity at different Alzheimer's rates for counties in TX, the Hat matrix, and the errors or residuals. Those are not displayed for lack of space.</td>
<td>x = {1 2.81, 1 2.64, … }
* These are Alzheimer’s rates.
y = {32.68, 29.08, … }
* These are obesity rates.

n={254};
Q=j(n,{1}, {1});
id=I(n);
xpx=x`*x;<br>ypy=(y`)*(y);
xpx_1=inv(xpx);
xpy=x`*y;<br>b=(xpx_1)*xpy;<br>h=x*xpx_1*x`;
ih=id-h;
yhat=h*y;
e=(id-h)*y;
sum_e=e`*Q;<br>sse=e`*e;
SSR=y`*(h-q*q`/n)*y;
SST=(y`*y)-(1/n)*(y`*q*q`*y);
mse=sse/(n-2);
yhat=h*y;
hh=h*h ;
print b, sum_e , sse, SSR, sst, mse xpy xpx xpx_1;
run;</td>
<td>
<table><tr><th>b</th></tr><tr><td>28.794369</td></tr><tr><td>0.5899145</td></tr></table>
<table><tr><th>sum_e</th></tr><tr><td>-2.24E-12</td></tr></table>
<table><tr><th>sse</th></tr><tr><td>579.7115</td></tr></table>
<table><tr><th>SSR</th></tr><tr><td>35.155237</td></tr></table>
<table><tr><th>SST</th></tr><tr><td>614.86674</td></tr></table>
<table><tr><th>xpx_1</th><th></th></tr><tr><td>0.0171069</td><td>-0.011418</td></tr><tr><td>-0.011418</td><td>0.0098989</td></tr></table>
<table><tr><th>mse</th><th>xpy</th><th>xpx</th><th></th></tr><tr><td>2.3004425</td><td>7486.6</td><td>254</td><td>292.975</td></tr><tr><td></td><td>8694.9741</td><td>292.975</td><td>438.9515</td></tr></table>
</td>
<td>The regression line is:

$\widehat{y} = 0.59x + 28.79$

where $\widehat{y}$ is the predicted obesity rate based on x, the measured Alzheimer’s rate. For every one unit the Alzheimer’s rate increases, it’s predicted the obesity rate will climb by 0.59 units.

This relation is not very reliable. According to the calculated output, we can find the R^2 value using SSR and SST.

$R^2 = \frac{SSR}{SST} = \frac{35.15}{614.87} = 0.057$

About 6% of the variation in obesity rates in TX counties is due to the variation in Alzheimer’s rate changes.

Based on this proc iml analysis, obesity and Alzheimer’s are not related in TX.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

Name	Objective	SAS Code	Notable SAS Output(s)	Conclusion
Jolie Wise (FL)	I will also demonstrate how we can use proc iml to use the matrix approach to linear regression. In this demonstration, our regressor value was the smoking rate in Florida, and our response variable is still the obesity rate. We can produce the estimated y values for each of our x values. We’ll also calculate the SSE, SSReg, SSTotal, and MSE.	proc iml; x={1 18.1, 1 28.56, 1 26.4, 1 29.14, 1 24.58,..., 1 29.64}; y={25.28, 35.04, 27.84, 35.48, 27.64,..., 35.62}; n={67}; Q=j(n,{1}, {1}); id=I(n); xpx=x`*x; ypy=(y`)*(y); xpx_1=inv(xpx); xpy=x`*y; b=(xpx_1)*xpy; h=x*xpx_1*x`; ih=id-h; yhat=h*y; e=(id-h)*y; sum_e=e`*Q; sse=e`*e; SSR=y`*(h-q*q`/n)*y; SST=(y`*y)-(1/n)*(y`*q*q` *y); mse=sse/(n-2); yhat=h*y; hh=h*h ; print h, hh, b, yhat, e, sum_e , sse, SSR, sst, mse xpy xpx xpx_1; run;	**yhat**: 22.794496, 32.109842, 30.186214, 32.626372, 28.56538 **b**: 6.6752072, 0.8905685 **sse**: 771.7977 **sum_e**: -3.11E-12 **SSR**: 708.82544 **SST**: 1480.6231 mse / xpy / xpx / xpx_1: 11.873811 | 1974.92 | 67 | 1715.4 | 0.7483855 | -0.028647 | 51359.772 | 1715.4 | 44813.087 | -0.028647 | 0.0011189	The model create is y = 6.6752 + 0.8906x This indicates that if the smoking rate of the county increases by one unit, we expect the obesity rate to increase by 0.8906. The intercept indicates that when the smoking rate is held at zero, we expect the obesity rate to be 6.6752. If we want to take a peak at performance of the model, we can find the R-Square value by dividing the SSReg by the SSTotal, which gives us 0.4787. This tells me that smoking rate only accounts for 47.87% of the variability in the obesity rate, which is not desirable result. This is consistent the calculated Mean Square Error (MSE). The MSE was 11.8738. This value would ideally be much lower.

PART II - SIMPLE LINEAR REGRESSION ANALYSIS

5. Run SAS for a simple linear regression on obesity in your state:

Name	Objective	SAS Code	Notable SAS Output(s)	Conclusion
Brad Lipson (NY)	To predict the age-adjusted obesity levels in various counties in New York using mercury levels in a linear regression model. To examine R-Square, P value, and parameter estimations.	proc reg; model obesity_age_adj=Mercury_TPY; run;	Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| Intercept | 1 | 27.76463 | 0.37435 | 74.17 | <.0001 Mercury_TPY | 1 | -47.03123 | 11.26658 | -4.17 | <.0001 R-Square | 0.2251 Adj R-Sq | 0.2121	The model gives us y= 27.76463 -47.03123x. This shows that the obesity rate is 27.76463 when mercury levels are zero and 47.03123 for every unit decrease. The finding is highly significant since the p-value is <.0001 and the corrected R-squared value is 0.2121 for mercury as a significant predictor variable in this obesity model. The F-value is 17.43 with a p-value of <.0001 which is also highly significant.

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>I analyzed the regression of obesity on fine particulate matter using proc reg to see the generated model.</td>
<td>proc reg;

model obesity = Fine_PM;
run;</td>
<td>
<table>
<tr><th colspan="6">Analysis of Variance</th></tr>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>1</td><td>20.61435</td><td>20.61435</td><td>2.05</td><td>0.1529</td></tr>
<tr><td>Error</td><td>562</td><td>5655.01008</td><td>10.06230</td><td></td><td></td></tr>
<tr><td>Corrected Total</td><td>563</td><td>5675.62443</td><td></td><td></td><td></td></tr>
</table>
<table>
<tr><td>Root MSE</td><td>3.17211</td><td>R-Square</td><td>0.0036</td></tr>
<tr><td>Dependent Mean</td><td>28.41376</td><td>Adj R-Sq</td><td>0.0019</td></tr>
<tr><td>Coeff Var</td><td>11.16400</td><td></td><td></td></tr>
</table>
<table>
<tr><th colspan="6">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th></tr>
<tr><td>Intercept</td><td>1</td><td>26.81298</td><td>1.12634</td><td>23.81</td><td><.0001</td></tr>
<tr><td>Fine_PM</td><td>1</td><td>0.14948</td><td>0.10444</td><td>1.43</td><td>0.1529</td></tr>
</table>
</td>
<td>According to the output, the regression model can be described by y hat = 26.81298 +0.14948x. This suggests that when fine particulate matter levels are zero the predicted obesity rate is 29.34283 and that for every unit increase in fine particulate matter there is an increase in obesity rate by 0.14948.
The p-value for the F-statistic is high (0.1529) and the adjusted R-squared value is low (0.0019) which suggests that the finding is not significant.</td>
</tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>I created a linear regression model predicting obesity age-adjusted using the smoking rate as the independent variable for California. I will review the R-Square, P value, and parameter estimates.</td>
<td>proc reg;

model obesity_age_adj=Smoking_Rate;

run;</td>
<td>
<table>
<tr><td>R-Square</td><td>0.2637</td></tr>
</table>
<table>
<tr><th colspan="6">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th></tr>
<tr><td>Intercept</td><td>1</td><td>15.57488</td><td>1.96504</td><td>7.93</td><td><.0001</td></tr>
<tr><td>Smoking_Rate</td><td>1</td><td>0.46100</td><td>0.10388</td><td>4.44</td><td><.0001</td></tr>
</table>
</td>
<td>The analysis of variance (ANOVA) for predicting obesity using smoking rate indicates that the model is statistically significant, with a highly considerable F-value of 19.70 (p < 0.0001). However, the R-Square value of 0.2637 suggests that only about 26.37% of the variability in obesity can be explained by the smoking rate predictor. The parameter estimates reveal that the intercept is estimated to be 15.57488, indicating the expected obesity value when the smoking rate is zero. The estimated coefficient for the smoking rate is 0.46100, signifying that for each unit increase in the smoking rate, we expect an increase of approximately 0.461 units in obesity.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>I am curious if age and obesity are related. I wonder if an aging population is more likely to be obese due to a more sedentary lifestyle. Therefore, I will investigate if the median age of a county can be a good regressor for obesity in Texas.</td>
<td>proc reg;
model Obesity = MedAge;
run;</td>
<td>
Anova: Correlation:
<table>
<tr><th>Pr > F</th></tr>
<tr><td>0.7179</td></tr>
</table>
<table>
<tr><td>R-Square</td><td>0.0005</td></tr>
</table>
Model Estimates:
$\widehat{Obesity} = 29.7 - 0.0065(MedAge)$
<table>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th></tr>
<tr><td>Intercept</td><td>1</td><td>29.71163</td></tr>
<tr><td>MedAge</td><td>1</td><td>-0.00645</td></tr>
</table>
</td>
<td>Median age of a county in TX is not correlated with that county's obesity rate. The estimated slope is practically 0 and the ANOVA found no difference in means across different ages.</td>
</tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>Physical inactivity was a regressor in all four of the possible models I chose for multiple regression, so I decided to use it to create the simple linear model. I will also evaluate the R-Square value, F value, and p value.</td>
<td>proc reg;
where state='FL';
model obesity_age_adj=physical_inactivity;
run;</td>
<td>
<table>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th></tr>
<tr><td>Intercept</td><td>1</td><td>4.38240</td></tr>
<tr><td>physical_inactivity</td><td>1</td><td>0.96666</td></tr>
</table>
<table>
<tr><td>R-Square</td><td>0.8184</td></tr>
</table>
<table>
<tr><th colspan="6">Analysis of Variance</th></tr>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>1</td><td>1211.81290</td><td>1211.81290</td><td>293.02</td><td><.0001</td></tr>
<tr><td>Error</td><td>65</td><td>268.81024</td><td>4.13554</td><td></td><td></td></tr>
<tr><td>Corrected Total</td><td>66</td><td>1480.62314</td><td></td><td></td><td></td></tr>
</table>
</td>
<td>The simple linear regression model created was y = 4.38240 + 0.96666x.
This indicates that if the rate of physical inactivity increases by one unit, we expect the county's obesity rate to increase by 0.96666. The intercept indicates that when the rate of physical inactivity is held at 0, we expect the obesity rate of the county to be 4.38240.
The R-Square value tells us that our model explains 81.84% of variance in obesity.
The F-value (293.02) and p-value (<0.0001) tell us that our model is significantly significant.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

6. Run SAS for model adequacy - is there a lack of fit?:

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Brad Lipson (NY)</td>
<td>To study how mercury levels affect obesity by using linear regression to determine these variables' relationships. To use lackfit to evaluate if obesity on mercury is fitting a linear regression.</td>
<td>proc reg;
model obesity_age_adj=Mercury_TPY/lackfit;
run;</td>
<td><table>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>1</td><td>104.73738</td><td>104.73738</td><td>17.43</td><td><.0001</td></tr>
<tr><td>Error</td><td>60</td><td>360.63210</td><td>6.01054</td><td></td><td></td></tr>
<tr><td>Lack of Fit</td><td>60</td><td>360.63210</td><td>6.01054</td><td>.</td><td>.</td></tr>
</table></td>
<td>This model shows a significant lack of fit with an F-value of 17.43 (p = <.0001) in the ANOVA. The null hypothesis is rejected since the p-value is less than 0.05, indicating that the model has a lack of fit in the linear regression of mercury on obesity. So factors are impacting obesity that the model cannot explain. This model may be too simple since it only has mercury, but other variables also affect obesity.</td>
</tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>To determine if there is a significant lack of linearity in the regression of obesity on a selected regressor (Fine_PM).

Ho: There is no lack of fit in the linear regression.
Ha: There is a lack of fit in the linear regression.</td>
<td>proc reg;

model obesity = Fine_PM/lackfit;
run;</td>
<td><table>
<tr><td>Lack of Fit</td><td>521</td><td>5277.14882</td><td>10.12888</td><td>1.10</td><td>0.3669</td></tr>
</table></td>
<td>Since the p-value of the lack of fit test is greater than the standard significance level of 0.05, one fails to reject the null hypothesis and it is concluded that there is not enough evidence of a lack of fit.</td>
</tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>We are investigating the impact of the 'Smoking Rate' on 'Obesity Age-Adjusted.' As part of this task, we will execute a linear regression analysis to establish the relationship between these variables. Additionally, we will be utilizing the 'lackfit' option in our analysis to verify the suitability of a linear model for our data.</td>
<td>proc reg;
model obesity_age_adj=Smoking_Rate/lackfit;
run</td>
<td><table>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Lack of Fit</td><td>52</td><td>446.96994</td><td>8.59558</td><td>1.99</td><td>0.3177</td></tr>
</table></td>
<td>The lack of fit analysis in the ANOVA shows that the model has a lack of fit with an F-value of 1.99 (p = 0.3177) for the lack of fit term. Since the p-value is greater than the chosen significance level (0.05), we fail to reject the null hypothesis, suggesting that there is evidence that the model fits the data adequately.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>Ho: There is no lack of fit in the linear regression.
Ha: There is a lack of fit in the linear regression.

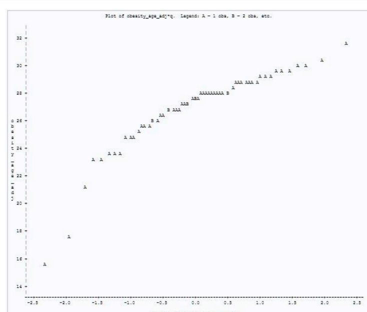
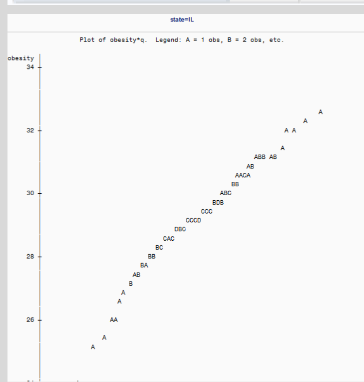
I am testing to see if the data appear to be in some function shape other than linear.</td>
<td>proc reg;
model Obesity = MedAge/lackfit;
run;</td>
<td><table>
<tr><td>Lack of Fit</td><td>144</td><td>330.10258</td><td>2.29238</td><td>0.90</td><td>0.7155</td></tr>
</table></td>
<td>Since the pvalue, 0.7155 > 0.05, the fit seems appropriate.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

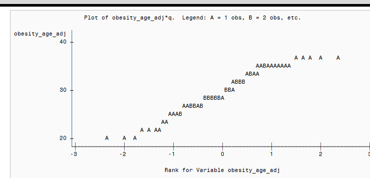
<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>The hypotheses to test for a lack of fit in our linear model are as follows:

H0: NO Lack of Fit
H1: Lack of Fit</td>
<td>proc reg;
where state='FL';
model obesity_age_adj=physical_inactivity/ lackfit;
run;</td>
<td>Analysis of Variance
<table>
<tr><th>Source</th><th>DF</th><th>Sum of Squares</th><th>Mean Square</th><th>F Value</th><th>Pr > F</th></tr>
<tr><td>Model</td><td>1</td><td>1211.81290</td><td>1211.81290</td><td>293.02</td><td><.0001</td></tr>
<tr><td>Error</td><td>65</td><td>268.81024</td><td>4.13554</td><td></td><td></td></tr>
<tr><td>Lack of Fit</td><td>64</td><td>268.74544</td><td>4.19915</td><td>64.80</td><td>0.0985</td></tr>
<tr><td>Pure Error</td><td>1</td><td>0.06480</td><td>0.06480</td><td></td><td></td></tr>
<tr><td>Corrected Total</td><td>66</td><td>1480.62314</td><td></td><td></td><td></td></tr>
</table></td>
<td>Since our p-value for our lack of fit test was 0.0985, which is larger than an alpha of 0.05, we fail to reject our null hypothesis. We do not have sufficient evidence to conclude that there is a lack of fit, meaning we will continue under the assumption that our model fits the data..</td>
</tr>
</table>

7. Run SAS for model adequacy to determine if normality assumption is appropriate?:

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Brad Lipson (NY)</td>
<td>To verify Obesity_age_adj normality using a Blom transformation. To construct rankings, display them alongside the original values then calculate their correlation to quantify the data's normalcy.</td>
<td>proc rank normal=blom out=normals;
var obesity_age_adj;
ranks q;
data normals;
set normals;
proc plot;
plot obesity_age_adj*q;
run;
proc corr;
var obesity_age_adj q;
run;</td>
<td>Pearson Correlation Coefficients, N = 62
Prob > |r| under H0: Rho=0
<table>
<tr><th></th><th>obesity_age_adj</th><th>q</th></tr>
<tr><td>obesity_age_adj
obesity_age_adj</td><td>1.00000</td><td>0.91606
<.0001</td></tr>
<tr><td>q
Rank for Variable obesity_age_adj</td><td>0.91606
<.0001</td><td>1.00000</td></tr>
</table>
</td>
<td>A positive 91.606% association between age-adjusted obesity and q was found. We need to have at least 0.91606 to assume that our data are normal based on the correlation coefficient test for normality with a sample size of 62 and a significance level less than 0.05. The data would be normally distributed because Obesity and q have a correlation of 0.91606. However, the QQ plot is not perfectly linear, so the normality is likely reduced, possibly due to outliers.</td>
</tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>The assumption of normality is applied to regression analysis. A Tukey transformation is used to verify the normality of the dataset.</td>
<td>proc rank normal = tukey out = normals; by state;
var obesity; ranks q;
data normals; set normals;
proc plot; by state; plot obesity*q; run;
proc corr; by state; var obesity q; run;</td>
<td>Pearson Correlation Coefficients, N = 102
Prob > |r| under H0: Rho=0
<table>
<tr><th></th><th>obesity</th><th>q</th></tr>
<tr><td>obesity</td><td>1.00000</td><td>0.98447
<.0001</td></tr>
<tr><td>q
Rank for Variable obesity</td><td>0.98447
<.0001</td><td>1.00000</td></tr>
</table>

</td>
<td>A minimum value of 0.9873 is needed for the correlation between q and obesity at the sample size of about 100 for the normality to be verified. The value calculated by SAS is slightly lower at 0.98447 so it cannot be concluded that the data is normal though is marginally insufficient. The QQ plot, which is nearly linear, suggests near normality, as well.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>Next, I am applying a Blom transformation to assess the normality of the 'obesity_age_adj' variable. We generate ranks, plot them against the original values for visual inspection, and then calculate their correlation to quantify the relationship, providing insight into the normality of our data.</td>
<td>proc rank normal=blom out=normals;
var obesity_age_adj;
ranks q;
data normals;
set normals;
proc plot;
plot obesity_age_adj*q;
run;
proc corr;
var obesity_age_adj q;
run;</td>
<td><table>
<tr><th colspan="3">Pearson Correlation Coefficients, N = 57
Prob > |r| under H0: Rho=0</th></tr>
<tr><th></th><th>obesity_age_adj</th><th>q</th></tr>
<tr><td>obesity_age_adj</td><td>1.00000</td><td>0.99351
<.0001</td></tr>
<tr><td>q
Rank for Variable obesity_age_adj</td><td>0.99351
<.0001</td><td>1.00000</td></tr>
</table></td>
<td>I observe a strong positive correlation of 99.351% between obesity age-adjusted and q. According to the Correlation Coefficient test for normality with a sample size of 57 and a significance level of 0.05, we should have at least 0.99351 to assume normality in our data. Based on a correlation value of 0.99351 between obesity and q, we can assume normality in our data.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>The test I conducted in #5 depends on an assumption of normal values in my response variable, Obesity. I will verify or refute whether or not that is the case.</td>
<td>proc rank normal=blom out=normals;
var Obesity;
ranks q;
data normals;
set normals;
proc plot;
plot Obesity*q;
run;
proc corr;
var Obesity q;
run;</td>
<td><table>
<tr><th colspan="3">Pearson Correlation Coefficients, N = 253
Prob > |r| under H0: Rho=0</th></tr>
<tr><th></th><th>Obesity</th><th>q</th></tr>
<tr><td>Obesity</td><td>1.00000</td><td>0.98054
<.0001</td></tr>
<tr><td>q
Rank for Variable Obesity</td><td>0.98054
<.0001</td><td>1.00000</td></tr>
</table></td>
<td>At a sample size of 200, the necessary value at the $\alpha = .01$ significance level to verify normality is 0.9905. This plot's correlation is 0.9805. Therefore, I ought not assume normality in the initial test I conducted. I should be wary of the results.</td>
</tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>Since simple linear regression has an assumption of normality, I will evaluate the normality of our response variable (obesity) using a Blom transformation to produce a QQ plot and the Pearson Correlation coefficient.</td>
<td>proc rank normal=blom out=normals;
where state='FL';
var obesity_age_adj;
ranks q;
data normals;
set normals;
proc plot;
where state='FL';
plot obesity_age_adj*q;
run;
proc corr;
where state='FL';
var obesity_age_adj q;
run;</td>
<td>

<table>
<tr><th colspan="3">Pearson Correlation Coefficients, N = 67
Prob > |r| under H0: Rho=0</th></tr>
<tr><th></th><th>obesity_age_adj</th><th>q</th></tr>
<tr><td>obesity_age_adj</td><td>1.00000</td><td>0.98115
<.0001</td></tr>
<tr><td>q
Rank for Variable obesity_age_adj</td><td>0.98115
<.0001</td><td>1.00000</td></tr>
</table></td>
<td>The critical values for a population of 60 and 75 are 0.9801 and 0.9838, respectively. Our population lies right around the middle of those at 67. The average of the critical values is 0.98195, which would be for a population of 67.5. The Pearson Correlation for obesity was 0.98115, which is close enough to the average critical value that I would assume normality. The QQ plot that was created backs up my assumption.</td>
</tr>
</table>

8. **Use the model to Run SAS to construct a confidence interval for the regression coefficient, a confidence interval for the intercept, a confidence interval for the mean obesity rate at a given regressor input, and/or a prediction interval for one obesity rate at a given regressor input:**

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Brad Lipson (NY)</td>
<td>To find confidence intervals for the intercept, slope, and mean response variable value for a certain mercury observation value were determined by the application of linear regression (PROC REG).</td>
<td>proc reg;
model
obesity_age_adj=Mercury_TPY/clb;
run;
proc reg;
model
obesity_age_adj=Mercury_TPY/cli clm;
run;</td>
<td>
<table>
<tr><th>Variable</th><th colspan="2">95% Confidence Limits</th></tr>
<tr><td>Intercept</td><td>27.01583</td><td>28.51343</td></tr>
<tr><td>Mercury_TPY</td><td>-69.56775</td><td>-24.49472</td></tr>
</table>
<table>
<tr><th>Obs</th><th>Dependent Variable</th><th>Predicted Value</th><th>Std Error Mean Predict</th></tr>
<tr><td>1</td><td>25.5</td><td>23.7280</td><td>0.8205</td></tr>
</table>
<table>
<tr><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th></tr>
<tr><td>22.0866</td><td>25.3693</td><td>18.5566</td><td>28.8994</td></tr>
</table>
</td>
<td>Here, the 95% confidence interval of the intercept is (27.01583, 28.51343) and the 95% confidence interval of the slope was found to be (-69.56775,-24.49472). This indicates that the model predicts that, when the mercury is zero, there is a 95% chance that the rate of obesity is between 27.01583 and 28.51343 and that there is a 95% chance that there is a change of -69.56775 and -24.49472 in obesity rate for every increase in unit of mercury. At mercury observation one with a value of 25.5, then there is a 95% change that the mean value of obesity is (22.08655,25.3693).</td>
</tr>
<tr>
<td>Pamela Mishaw (IL)</td>
<td>Linear regression was used to determine confidence intervals of the intercept, the slope, and the mean response variable value at a particular fine particulate matter observation value.</td>
<td>proc reg;
model Obesity = Fine_PM/cli clm clb;
run;</td>
<td>
<table>
<tr><th colspan="8">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th><th colspan="2">95% Confidence Limits</th></tr>
<tr><td>Intercept</td><td>1</td><td>26.81298</td><td>1.12634</td><td>23.81</td><td><.0001</td><td>24.60062</td><td>29.02534</td></tr>
<tr><td>Fine_PM</td><td>1</td><td>0.14948</td><td>0.10444</td><td>1.43</td><td>0.1529</td><td>-0.05565</td><td>0.35461</td></tr>
</table>
<table>
<tr><th>Obs</th><th>Dependent Variable</th><th>Predicted Value</th><th>Std Error Mean Predict</th><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th></tr>
<tr><td>1</td><td>19.9</td><td>28.1894</td><td>0.2059</td><td>27.7849</td><td>28.5939</td><td>21.9456</td><td>34.4332</td></tr>
</table>
</td>
<td>The 95% confidence interval of the intercept is (24.60062, 29.02534) and the 95% confidence interval of the slope was found to be (-0.05565, 0.34561). This indicates that the model predicts that, when the fine particulate matter is zero, there is a 95% chance that the rate of obesity is between 24.60062 and 29.02534 and that there is a 95% chance that there is a change of -0.05565 and 0.34561 in obesity rate for every increase in unit of fine particulate matter. At fine particulate matter observation one (where the value is 9.208) the 95% confidence interval of the mean value of obesity is (27.7849, 28.5939). Again, this indicates a 95% chance of the value of obesity rate being between 27.7849 and 28.5939 when the fine particulate matter value is 9.208.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer's Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jorge Sanchez (CA)</td>
<td>In this analysis, I use linear regression to examine the relationship between the 'obesity_age_adj' and the 'Smoking_Rate' variables. By incorporating confidence intervals, we estimate the likely relationship between these two variables and quantify the degree of certainty, or uncertainty, in our estimates. This provides us with a range of values that we can reasonably confidently contain the true parameter value.</td>
<td>proc reg;
model obesity_age_adj=Smoking_Rate/clb;
run;

proc reg;
model obesity_age_adj=Smoking_Rate/cli clm;
run;</td>
<td>
<table>
<tr><th>Variable</th><th colspan="2">95% Confidence Limits</th></tr>
<tr><td>Intercept</td><td>11.63685</td><td>19.51291</td></tr>
<tr><td>Smoking_Rate</td><td>0.25283</td><td>0.66917</td></tr>
</table>
<table>
<tr><th>Obs</th><th>Dependent Variable</th><th>Predicted Value</th><th>Std Error Mean Predict</th></tr>
<tr><td>1</td><td>19.9</td><td>21.9090</td><td>0.6299</td></tr>
</table>
<table>
<tr><th>Obs</th><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th></tr>
<tr><td>1</td><td>20.6466</td><td>23.1714</td><td>15.9779</td><td>27.8401</td></tr>
</table>
</td>
<td>We are 95% confident that the true value of the slope is between 0.25283 and 0.66917. This means that for every unit increase in the smoking rate, we can be 95% confident that the age-adjusted obesity rate increases between 0.25283 and 0.66917 units, assuming all other variables in the model are held constant.
We are 95% confident that the true value of the intercept is between 11.63685 and 19.51291. Assuming all other variables in the model are held constant if the smoking rate were zero, we would expect, with 95% confidence, that the age-adjusted obesity rate would be between 11.63685 and 19.51291.

Mean and single Prediction
For observation 1, with a smoking rate of 13.74, the model predicts an obesity rate 21.9090. The standard error of this Prediction is 0.6299, indicating some potential variability in this estimate. The 95% confidence interval for the mean response (20.6466 to 23.1714) suggests where the true mean response might lie. A wider 95% confidence interval for a single prediction (15.9779 to 27.8401) accounts for extra variability in predicting individual responses. The residual of -2.0090 suggests the model slightly overestimated the obesity rate for this observation.</td>
</tr>
<tr>
<td>Daniel Wilson (TX)</td>
<td>Median age does not correlate with obesity like I thought it would! Yet, I can still use my model to estimate values. The following is a confidence interval for the mean obesity rate for a Texas county with a median age of 27.8. This is appropriate since the 27.8 is within the domain (the youngest measured county is 24 years old). Then I will construct a prediction interval on the same input. I chose this value because it is closest to my actual age.</td>
<td>proc reg;
model Obesity = MedAge/cli clm;
run;</td>
<td>
<table>
<tr><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th></tr>
<tr><td>29.0123</td><td>30.0948</td><td>26.4570</td><td>32.6500</td></tr>
</table>
</td>
<td>Based on my (not very good) model, some TX county with a median age of 27.8 years old would have an obesity rate between 26.46 and 32.65. The mean obesity rate of all counties with median age 27.8 would fall in a narrower range of 29.0 to 30.1. I make both of these claims with 95% confidence.</td>
</tr>
</table>

STA6235: Summer 2023 Group 4 Alzheimer’s Project

<table>
<tr><th>Name</th><th>Objective</th><th>SAS Code</th><th>Notable SAS Output(s)</th><th>Conclusion</th></tr>
<tr>
<td>Jolie Wise (FL)</td>
<td>We will also find the confidence intervals for the true intercept and true slope of our model. Also, we will evaluate the confidence interval and prediction interval for a physical inactivity rate of 21.06.</td>
<td>proc reg;
where state='FL';
model obesity_age_adj=physical_inactivity/ clm clb cli;
run;</td>
<td>
<table>
<tr><th colspan="8">Parameter Estimates</th></tr>
<tr><th>Variable</th><th>DF</th><th>Parameter Estimate</th><th>Standard Error</th><th>t Value</th><th>Pr > |t|</th><th colspan="2">95% Confidence Limits</th></tr>
<tr><td>Intercept</td><td>1</td><td>4.38240</td><td>1.48685</td><td>2.95</td><td>0.0044</td><td>1.41295</td><td>7.35185</td></tr>
<tr><td>physical_inactivity</td><td>1</td><td>0.96666</td><td>0.05647</td><td>17.12</td><td><.0001</td><td>0.85388</td><td>1.07944</td></tr>
</table>
Physical Inactivity = 21.06
<table>
<tr><th>Obs</th><th>Dependent Variable</th><th>Predicted Value</th><th>Std Error Mean Predict</th></tr>
<tr><td>1</td><td>25.3</td><td>24.7403</td><td>0.3719</td></tr>
</table>
<table>
<tr><th colspan="2">95% CL Mean</th><th colspan="2">95% CL Predict</th></tr>
<tr><td>23.9977</td><td>25.4830</td><td>20.6116</td><td>28.8691</td></tr>
</table>
</td>
<td>Our 95% confidence interval for our intercept is: (1.41295, 7.35185), indicating that we can be 95% certain that the true intercept lies within this range. The 95% confidence interval for our slope of physical inactivity is: (0.85388, 1.07944), indicating that we can be 95% certain that the true slope lies within this range.

For a physical inactivity rate of 21.06, the predicted obesity value is 24.7403. The 95% confidence interval for the mean value is: (23.9977, 25.4830), indicating that we are 95% certain that the mean obesity rate for counties with a physical inactivity rate of 21.06 lies within this range. The 95% prediction interval is: (20.6116, 28.8691), indicating that we can predict with 95% certainty that the obesity value for a physical inactivity rate of 21.06 lies within this range.</td>
</tr>
</table>

Executive Summary:

Brad Lipson (NY):

The relationship between mercury exposure and age-adjusted obesity in 62 counties in New York was evaluated with linear regression methods. This model shows a significant lack of fit with an F-value of 17.43 (p = <.0001) in the ANOVA. Eliminating the outlier in observation 22 can improve this. This model may be too simple since it only has mercury. Also, the QQ plot is not perfectly linear, so the normality is likely reduced, possibly due to outliers. However, the model also demonstrated a significant lack of fit showing that many factors influencing obesity which cannot be explained by the model.

The correlation coefficient between obesity and q was 0.916006, indicating that the data had a normal distribution so parametric statistical methods can be used to analyze the data. The extreme outlier was Herkimer County, NY (observation 22) and it would be interesting to investigate the reasons such as mercury in the local lakes and rivers since it is a rural area. It is also one of the most obese counties in NY which can contribute to being an outlier.

This is an interesting finding with a p-value of <.0001 and corrected R-squared value of 0.2121 for mercury as a significant predictor variable in this obesity model. The best overall model was Obesity_age_adj=Mercury_TPY Diabetes Glyphosates Smoking_Rate mental_distress with an R squared of 65.09 for these 5 variables and Mallow C(p)=5.04. The PRESS statistic improves to give better model prediction when we remove Observation 22 from 213 to 203. There is also an increase in fit with the reduction of MSE to 1.66 and adjusted R-square to 0.6319 by removing observation 22. So this point should be removed from the model.

Pamela Mishaw (IL):

Two outlier data points (counties 58 and 68) were identified as influential points in the Illinois dataset. A notable improvement in the model was found after removing the points from the data set, based on the decrease in MSE, increase in adjusted R-squared, and decrease in the PRESS statistic. These indicate an improvement in the model's fit (lower MSE, higher R-squared) and the model's prediction ability (lower PRESS). These points were removed prior to further model analysis. Significant multicollinearity was found between the sixtyfiveandup and med_age variables, based on the condition indices and proportion of variation values found in the influence statistics table. The sixtyfiveandup variable was thus removed to improve the regression model. The possible regression models were then examined and the "best" of the selected models was determined to be obesity = alz mental_distress physical_inactivity Diabetes Mercury Lead based on the low C(p) value, high predicted R-squared value (both of which indicate good prediction capability), low MSE, high adjusted R-squared value, and low PRESS statistic value (the latter three indicating a good fit of the data to the model).

The simple linear regression model of obesity on fine particulate matter was found to be y-hat = 26.81298 +0.14948x. Lack of fit analysis showed a lack of evidence for a lack of fit in the model. The obesity dataset of Illinois was found to be marginally not normal based on the correlation between obesity and q.

The 95% confidence interval of the intercept was found to be(24.60062, 29.02534) and the 95% confidence interval of the slope was found to be (-0.05565, 0.34561) for the regression of obesity on fine particulate matter. At fine particulate matter observation one (where the value is 9.208) the 95% confidence interval of the mean value of obesity was determined to be (27.7849, 28.5939).

Jorge Sanchez (CA):

In conclusion, for California state, according to correlation coefficients, the most significant predictors for obesity age-adjusted in a county were diabetes (0.7765), physical inactivity (0.8310), and smoking rates (0.5135). After reviewing influence points, I detected San Francisco County as a possible outlier. Then, the model that includes Med_age, physical_inactivity, cancer, fine_PM_2_5, and Lead_TPY stands out as the most effective predictor for obesity age-adjusted for the California state with an adjusted R-Square of 0.8655, an MSE of 1.4999 and C(p) of 13.9265. I ran a linear regression on this model. I identified that the counties in California, for every one-unit increase in a county's median age (Med_age), the obesity rate tends to decrease by approximately 0.198 units. Conversely, for every unit increase in physical inactivity, obesity rates increase by about 0.611 units. Counties with a unit increase in cancer rates see an increase of roughly 0.095 units in obesity rates. When the levels of fine particulate matter (Fine_PM_2_5) go up by one unit, the obesity rate jumps by around 0.895 units. Interestingly, a one-unit increase in lead levels (Lead_TPY) is associated with a decrease in obesity rates by about 1.024 units. All these results are statistically significant, which means they're likely not due to random chance. However, it's essential to remember that these findings don't imply causation. While these factors are associated with obesity rates, it doesn't mean they directly cause changes in obesity. More research is necessary to delve deeper into these relationships.

Daniel Wilson (TX):

In Texas, the greatest predictors of obesity in a county were diabetes rates, physical inactivity rates, and smoking rates. These three variables had the highest pairwise correlation coefficients (0.68, 0.62, 0.55, respectively) with obesity, and were the only variables with correlation coefficients above 0.37. Coupled with mental distress levels, lead pollution, and mercury pollution, these six regressors formed a moderate-strong model for fitting and predicting obesity in Texas counties ($R^2_{adj} = 0.585$).

After generating this regression model, I personally wanted to explore a simple regression of age and obesity. I hypothesized that as people get older, they might tend to be more sedentary, and therefore more obese. Yet, my exploration yielded quite the opposite conclusion! Median age shared the second lowest correlation coefficient (after glyphosates) with obesity at 0.022 (R^2=.0005). Furthermore, the slope of the regression was -0.006, which is essentially a flat line. This number can be interpreted as follows: according to the regression, for every one year the median age increases, the obesity rate is predicted to decrease by 0.006 percent. Essentially, the median age of a county has absolutely no effect on the obesity rate of that county.

Jolie Wise (FL):

In Florida, the predictors with the greatest pairwise correlation coefficients were physical inactivity rates (0.90468), diabetes rates (0.91138), and smoking rates (0.69191). These three regressors were included in the final model to fit and predict the age-adjusted obesity rate in Florida counties. Age-adjusted Alzheimer's rate, rate of heart disease, lead pollution, and NATA Cancer rates were also included in the final model. This model was able to account for 91.97% of the variation (Adjusted R-Square) in obesity rates in Florida, as well as performing well on other metrics (MSE, Mallow's Cp, and PRESS Statistic).

Since we found that physical inactivity rates were one of the predators with the greatest correlation to our response variable, I wanted to see how this would be illustrated in a simple linear regression model. Our simple linear model shows that we expect obesity rates to increase by 0.96666 if physical inactivity rates were to increase by one unit. This model explains 81.84% of the variation in the obesity rates. This result backs the earlier claim that physical inactivity was one of the greatest predictors of obesity rates in Florida. We also failed to find that this linear model did not adequately fit our data.

Holmes County and Wakulla County were identified as potential outliers when initially evaluating the data. In the end, I decided to keep them in the data; however, it could be interesting to further investigate why these counties stood out. They are located relatively near each other, so this could be a clue.