Siddharth Potti

Econometrics Final Project

Introduction:

Our topic involves finding a relationship between socioeconomic status and obesity. This topic was of interest to us, because obesity is a factor that increases each year. There are many factors influencing it, so as researchers, we wanted to test whether socioeconomic status, a factor which we believed was important, would have a quantitative effect on obesity.

Literature Review:

Our research primarily focused on different research papers correlated to the cause of obesity, as well as the effect obesity has on socioeconomic status. In our findings from BMC Public Health and Science Direct, we found that the primary causes of rising obesity with certain socioeconomic factors included three main categories: poverty, unemployment, income level, and receival of SNAP. In addition from Nature Magazine and medicalexpress, we found that obesity can also strongly be correlated with environmental factors, such as quantity/quality of food consumption for an individual. Many of these studies focused on measuring adults, as the results were more straightforward when measuring obesity and socioeconomic status. One study from Plos Medicine has revealed that as unemployment rates rose, so did the prevalence of obesity. While many studies directly hypothesize that lower socioeconomic status does have a correlation with obesity, psychological mechanisms are still unclear. Similar to our hypothesis, the results of our research did reveal an increase in obesity rates due to specific socioeconomic and environmental factors. Thus, we used our research and decided that we would use features that represented socioeconomic status for each state and we decided to measure average obesity for each state as our dependent variable, as much detail was available for that.

Below are the sources we have used to conduct research:

https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-8322-8
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101796/
https://link.springer.com/article/10.1007/s13679-020-00398-7
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003243
https://www.frontiersin.org/articles/10.3389/fnut.2021.585318/full
https://www.euro.who.int/__data/assets/pdf_file/0003/247638/obesity-090514.pdf
https://www.sciencedirect.com/science/article/pii/S2352827316301896
https://www.nature.com/articles/ijo2016109
https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(19)30421-8/fulltext
https://medicalxpress.com/news/2017-05-socioeconomic-status-linked-obesity-distress.html

Hypothesis:

We believe that lower socioeconomic status and poorer conditions would lead to a higher chance of obesity in a community. We also believe that average obesity rates have increased in each state over the past few years. We aim to measure socioeconomic status by using the average income level in each state which would be one of our independent variables. Another independent variable would be education attainment by state based on receiving a bachelors or higher, as it would be a predictor for a higher average socioeconomic status. Our third variable to measure socioeconomic status would be median home value in each state. We believe that a higher median home value would signify a higher average socioeconomic status for each state. Our dependent variable would be the rate of obesity for each state.

Experimental Design:

Our theoretical model is: $y = b0 + b1x1n + b2x2n + b3x3n + un$. Where y is average obesity in each state, and x1n is average income level in each state. x2n is education attainment by state based on the percentage of people getting a bachelors or high. x3n is the median home value in each state in US dollars.

Analysis:

For this project, We used Google Colab and Python modules to visualize our data and create our models. Please look at the text within our screenshots as well as the graphs, as they also contain explanation/ analysis of our results.

As our first step, we parsed my data which included creating a csv with each feature and dependent variable labeled as a column header.

	MedianHomeValue	State	AverageIncome	BachelorsOrHigher	ObesityRate
0	141302	Alabama	80098	26.0	39.0
1	326000	Alaska	95705	30.0	31.9
2	257600	Arizona	93563	30.0	30.9
3	129500	Arkansas	73595	23.0	36.4
4	550800	California	109260	34.0	30.3
5	381000	Colorado	11730	41.0	24.2
6	244800	Connecticut	112717	39.0	29.2
7	236000	Delaware	88015	32.0	36.5
9	237900	Florida	80986	30.0	28.4
10	193500	Georgia	89679	31.0	34.3
11	345963	Hawaii	100865	33.0	24.5
12	274200	Idaho	86823	28.0	31.1
13	183500	Illinois	105406	35.0	32.4
14	148700	Indiana	93692	27.0	36.8
15	146500	Iowa	89448	29.0	36.5
16	141500	Kansas	92077	33.0	35.3
17	148400	Kentucky	79926	24.0	36.6
18	147600	Louisiana	75323	24.0	38.1
19	237800	Maine	87585	32.0	31.0
20	290500	Maryland	130850	40.0	31.0
21	408100	Massachusetts	127044	44.0	24.4
22	154500	Michigan	98869	29.0	35.2
23	239900	Minnesota	112222	36.0	30.7
24	130200	Mississippi	66127	22.0	39.7
25	163700	Missouri	89400	29.0	34.0
26	242100	Montana	81696	32.0	28.5
27	169900	Nebraska	98805	32.0	34.0
28	291800	Nevada	83588	25.0	28.7
29	280400	New Hampshire	111787	37.0	29.9
30	329000	New Jersey	119883	40.0	27.7
31	197400	New Mexico	72198	27.0	30.9

Next, we removed outliers - states with N/A data such as Vermont and the District of Columbia as those would disrupt our data visualization and our regression models.

Then, I normalized the columns of Average Income and Median Home Value, so that I could eventually create a plot with all of those features on the same x-y scale.

```
#normalizing
#Normalizing the dataframe
establisheddata['AverageIncome'] = establisheddata['AverageIncome'].div(3249)
establisheddata['MedianHomeValue'] = establisheddata['MedianHomeValue'].div(8966)
establisheddata
```

	MedianHomeValue	State	AverageIncome	BachelorsOrHigher	ObesityRate
0	15.759759	Alabama	24.653124	26.0	39.0
1	36.359581	Alaska	29.456756	30.0	31.9
2	28.730761	Arizona	28.797476	30.0	30.9
3	14.443453	Arkansas	22.651585	23.0	36.4
4	61.432077	California	33.628809	34.0	30.3
5	42.493866	Colorado	3.610342	41.0	24.2
6	27.303145	Connecticut	34.692829	39.0	29.2
7	26.321660	Delaware	27.089874	32.0	36.5
9	26.533571	Florida	24.926439	30.0	28.4
10	21.581530	Georgia	27.602031	31.0	34.3
11	38.586103	Hawaii	31.044937	33.0	24.5
12	30.582199	Idaho	26.722992	28.0	31.1
13	20.466206	Illinois	32.442598	35.0	32.4
14	16.584876	Indiana	28.837181	27.0	36.8
15	16.339505	Iowa	27.530933	29.0	36.5
16	15.781843	Kansas	28.340105	33.0	35.3
17	16.551416	Kentucky	24.600185	24.0	36.6
18	16.462190	Louisiana	23.183441	24.0	38.1

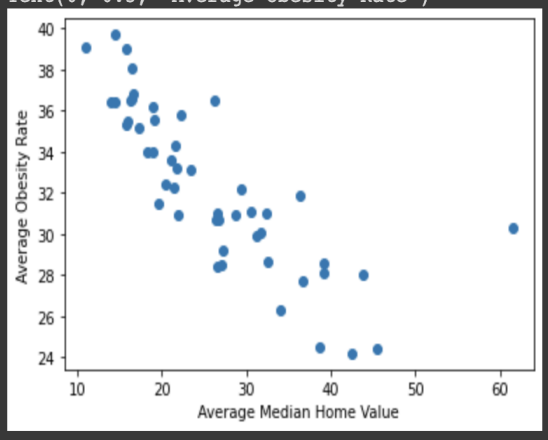
As you can see, our data is now much more manageable.

We then decided to individually plot each of our features with Average Obesity Rate:

Here are our visualizations with our analysis:

```
[ ] plt.plot(establisheddata.MedianHomeValue,establisheddata.ObesityRate,'o')
    plt.xlabel('Average Median Home Value')
    plt.ylabel('Average Obesity Rate')
```

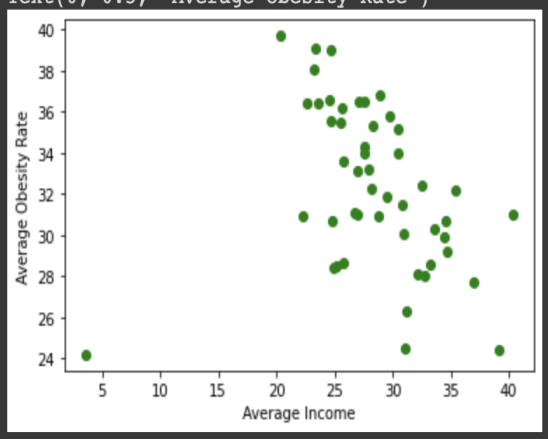
```
Text(0, 0.5, 'Average Obesity Rate')
```



Occurent negative correlation between Average Median Home Value and Average Obesity Rate

```
plt.plot(establisheddata.AverageIncome,establisheddata.ObesityRate,'o', color = 'green')
plt.xlabel('Average Income')
plt.ylabel('Average Obesity Rate')
```

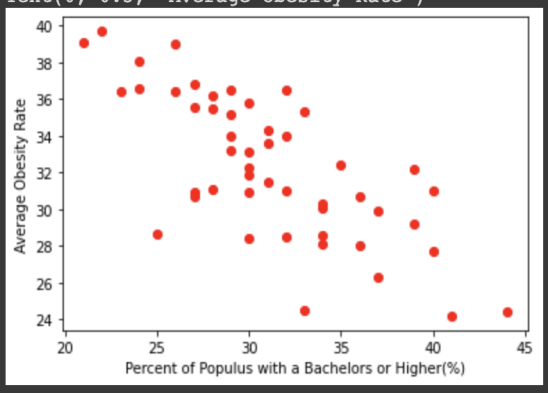
```
Text(0, 0.5, 'Average Obesity Rate')
```



This is Good! It makes sense that with average income increasing, average obesity would decrease.

```
[ ]  #Lets do the same thing with our last feature - education level
     plt.plot(establisheddata.BachelorsOrHigher,establisheddata.ObesityRate,'o', color = 'red')
     plt.xlabel('Percent of Populus with a Bachelors or Higher(%)')
     plt.ylabel('Average Obesity Rate')
```

```
Text(0, 0.5, 'Average Obesity Rate')
```

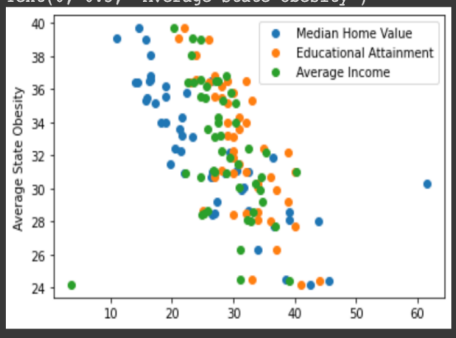


A little more varied, but still an apparent negative relationship between the two - which is expected.

Then after creating our initial visualizations, we created a plot with all of our features.

```
[ ]  #Now lets compare them together in one plot
     #First we have to normalize two of our features
     plt.plot(establisheddata.MedianHomeValue,establisheddata.ObesityRate,'o')
     plt.plot(establisheddata.BachelorsOrHigher,establisheddata.ObesityRate,'o')
     plt.plot(establisheddata.AverageIncome,establisheddata.ObesityRate,'o')
     plt.legend(['Median Home Value','Educational Attainment','Average Income'])
     plt.ylabel('Average State Obesity')
     #divide by specific numbers to fit the X axis in equal proportion
```

```
Text(0, 0.5, 'Average State Obesity')
```



This also makes sense as with the increase of all three of these categories, there should be an expected decrease in average state obesity

Now that we finished all of our visualizations, we moved on to creating a simple linear regression to test the relationship between Average Income and Average Obesity Rate.

```python
#Simple Linear Regression
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.formula.api import ols

#Simple Linear Regression Model
import pandas as pd
from sklearn import linear_model
import statsmodels.api as sm
X = establisheddata[['AverageIncome']]
Y = establisheddata['ObesityRate']

# with sklearn
regr = linear_model.LinearRegression()
regr.fit(X, Y)

print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)

# prediction with sklearn
New_AverageIncome = [33.628809];
print ('Predicted Obesity Rate: \n', regr.predict([New_AverageIncome]))
print('Real Obesity Rate: 30.3')

# with statsmodels
X = sm.add_constant(X) # adding a constant

model = sm.OLS(Y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

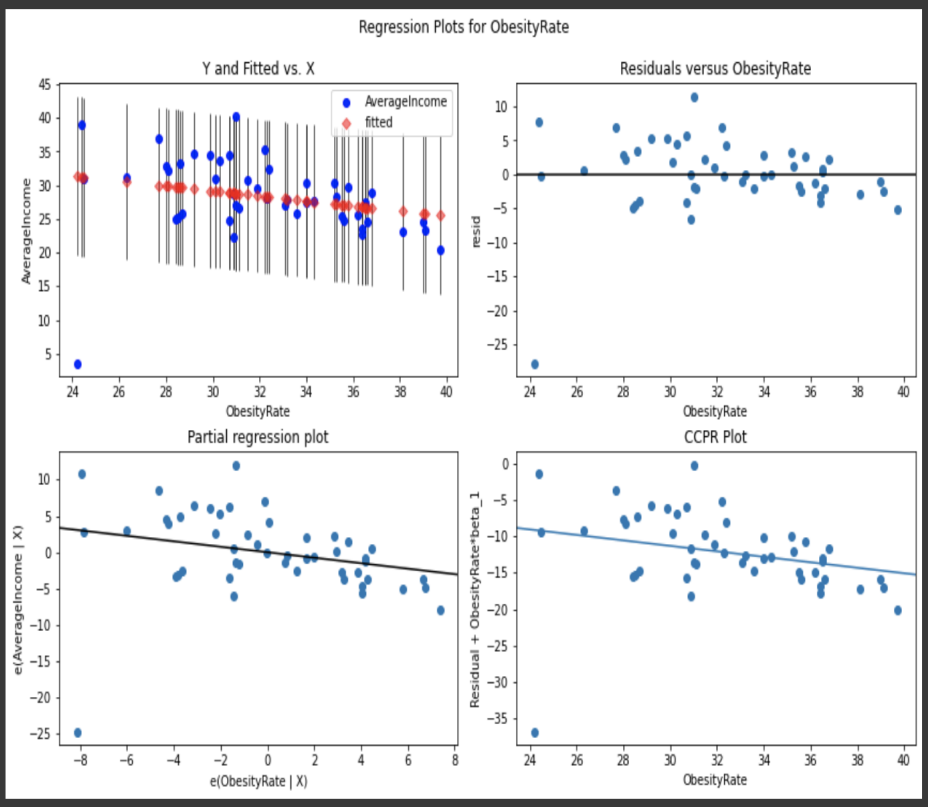
Output of our simple linear regression model:

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the functions in the public API at panda
  import pandas.util.testing as tm
Intercept:
 37.36913586175232
Coefficients:
 [-0.17795671]
Predicted Obesity Rate:
 [31.38466361]
Real Obesity Rate: 30.3
                            OLS Regression Results
==============================================================================
Dep. Variable:            ObesityRate   R-squared:                       0.067
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     3.376
Date:                Wed, 19 Jan 2022   Prob (F-statistic):            0.0725
Time:                        22:14:13   Log-Likelihood:                -134.45
No. Observations:                  49   AIC:                             272.9
Df Residuals:                      47   BIC:                             276.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         37.3691      2.795     13.372      0.000      31.747      42.991
AverageIncome -0.1780      0.097     -1.837      0.072      -0.373       0.017
==============================================================================
Omnibus:                        6.876   Durbin-Watson:                   1.942
Prob(Omnibus):                  0.032   Jarque-Bera (JB):                5.778
Skew:                          -0.729   Prob(JB):                       0.0556
Kurtosis:                       3.841   Cond. No.                         147.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but LinearRegression was fitted with feature names
  "X does not have valid feature names, but"
```

As you can see, our R^2 value of 0.067 was quite low. But since this Average Income is only one of the features in our multiple linear regression, a low R^2 value for a simple linear regression model was expected.

In order to test whether we could create a multiple linear regression model, I had to first create a residual plot.

```
#fit simple linear regression model
model = ols('AverageIncome ~ ObesityRate', data=establisheddata).fit()
#Regression Plot
fig = plt.figure(figsize=(12,8))

#produce regression plots
fig = sm.graphics.plot_regress_exog(model, 'ObesityRate', fig=fig)
```



Since the residuals appear to be randomly scattered around zero, this is an indication that heteroscedasticity is not a problem with the predictor variable.

Straight Enough Condition is met so we can proceed and create our multiple linear regression model

Then, we created our multiple linear regression model.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
x = establisheddata[['AverageIncome', 'MedianHomeValue', 'BachelorsOrHigher']]
y = establisheddata['ObesityRate']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 100)
#Fitting the Multiple Linear Regression model
mlr = LinearRegression()
mlr.fit(x_train, y_train)
print("Intercept: ", mlr.intercept_)
print("Coefficients:")
print(list(zip(x, mlr.coef_)))
#Prediction of test set
y_pred_mlr= mlr.predict(x_test)
#Predicted values
print("Prediction for test set: {}".format(y_pred_mlr))
mlr_diff = pd.DataFrame({'Actual value': y_test, 'Predicted value': y_pred_mlr})
mlr_diff.head()
```

```
Intercept:  43.398484027707404
Coefficients:
[('AverageIncome', 0.3365777053754754), ('MedianHomeValue', -0.21833452315802304), ('BachelorsOrHigher', -0.4820175080551816)]
Prediction for test set: [30.31541465 25.41294061 32.90047409 30.31619769 34.06364653 32.96248702
 15.5730488  32.505145   34.53149927 28.63402195 35.72282125 32.2192071
 34.5231476  31.25646632 35.76166954]
```

	Actual value	Predicted value
6	29.2	30.315415
21	24.4	25.412941
34	33.1	32.900474
29	29.9	30.316198
43	35.8	34.063647

As you can see, the predicted value is very close to the actual value.

Here in this screenshot, you can see the results of our multiple linear regression model which we trained using machine learning. It gave us our intercept, as well as our coefficients for each of our variables. Just through this data, I expected our model to have a high accuracy as the predicted values were very close to the actual ones.

Next, we proceeded with evaluating our model.

```
#Model Evaluation
from sklearn import metrics
meanAbErr = metrics.mean_absolute_error(y_test, y_pred_mlr)
meanSqErr = metrics.mean_squared_error(y_test, y_pred_mlr)
rootMeanSqErr = np.sqrt(metrics.mean_squared_error(y_test, y_pred_mlr))
print('R squared: {:.2f}'.format(mlr.score(x,y)*100))
print('Mean Absolute Error:', meanAbErr)
print('Mean Square Error:', meanSqErr)
print('Root Mean Square Error:', rootMeanSqErr)
```

```
R squared: 63.21
Mean Absolute Error: 1.8787718978726489
Mean Square Error: 7.628968972381764
Root Mean Square Error: 2.762058828551949
```

** Analysis: We have a R squared value of 63.21, which means that 63.21% of variation in the dependent variable is accounted for by the multiple linear regression model. Our mean absolute error, our mean square error, and our root mean square error are all relatively low, which is good for this model**

Our analysis above our model evaluation is in the screenshot above.

Our last step was completing the Z-Test, to test our hypothesis of whether the average obesity rate of most states increased recently or not.

Final Step: Completing the Z-Test

Hypothesis: According to the most recent Behavioral Risk Factor Surveillance System (BRFSS) data, adult obesity rates exceed 35% in 16 states. I predict that obesity rate will increase In my data, out of 52 states, there are 15 states with obesity rates above 35%. Is this evidence of a change in the amount of states with obesity rates above 35%?

Null Hypothesis: Po = 0.35; Ha; P != 0.35

Qo = 0.65; $SD\hat{p}$ = 0.066143

Z = (0.288-0.35)/0.066143 = -0.93736298625

Using the Z-Table: P Value = 0.1762

Since the P value > α of 0.05, we reject the null hypothesis, so there is no evidence of a change in the amount of states with obesity rates above 35%

Conclusion:

I believe that our model did give results that do make sense, as there is an expected negative relationship between all three of these variables and average obesity. However, since the independent assumption might not work, as average income could affect the median home value of a state, our data does not have complete statistical significance, and it is hard to generalize our results. Therefore we cannot truly state that every state might have a lower obesity rate if their average income, median home value, and rate of education attainment are high. If we had more time and resources, I would like to have measured features that are completely independent from each other, so that our results could potentially be generalizable. Since I would have access to more resources, I could measure data for smaller features that would normally take a long time to gather. However even with the independence assumption not holding 100% true, our results do show that these factors are in fact related to average obesity rate in each state. This could mean that even if it isn't in the exact proportion predicted by our model, education attainment, median home value, and average income are influencers of average obesity among the United States.