An Econometric Analysis of Obesity

Debarshi Dutta

Adult BMI and Obesity

- BMI* – ratio of the weight (in Kg) and the square of height (in m)
  - Unit: kg/m^2

- An adult is obese* if their calculated BMI is 30.0 or above.

BMI range	Category
<18.5	Underweight
18.5 - <25.0	Normal
25.0 - <30.0	Overweight
>=30.0	Obese

*All definitions are from the CDC

Obesity- an epidemic

- Obesity is recognized as a standalone disease by the CDC
- Leads to co-morbidities like
  - Type 2 diabetes and prediabetes
  - Cancer
  - Coronary Artery Disease (CAD)

- Obesity in children is rising as well
  - causes dyslipidemia, hypertension, and hyperinsulinemia later in life

Obesity- in numbers

- Prevalence in Adults*
  - 42.4% in 2017-2018.
  - Spike of ~12% from 1999–2000
  - Severe obesity increased from 4.7% to 9.2%
- Prevalence in Children*
  - 18.5% = ~13.7 million children and adolescents (aged 2-19)
- Obesity is expensive to treat ($$$$)
  - Estimated annual medical cost - $147 billion (2008)
  - +$1,429 in average more than the non obese people/year

All data collected from the CDC website

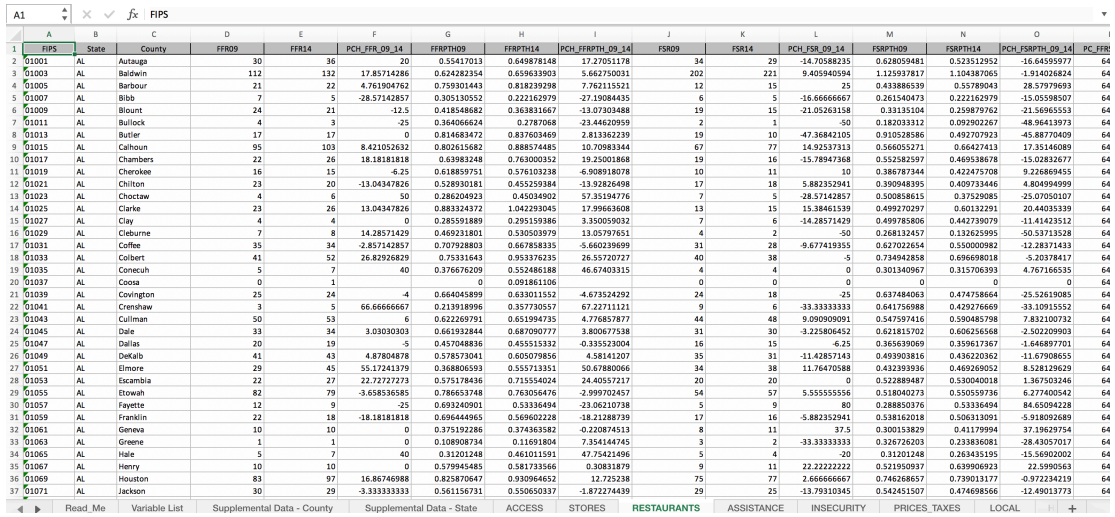
Motivation: What is the working hypothesis

- Who is most susceptible-
  - Low income households
    - Households with no cars
    - Low access to grocery stores
    - Families with children eligible for free lunch
    - Participants of the SNAP program
    - WIC participants

The Data

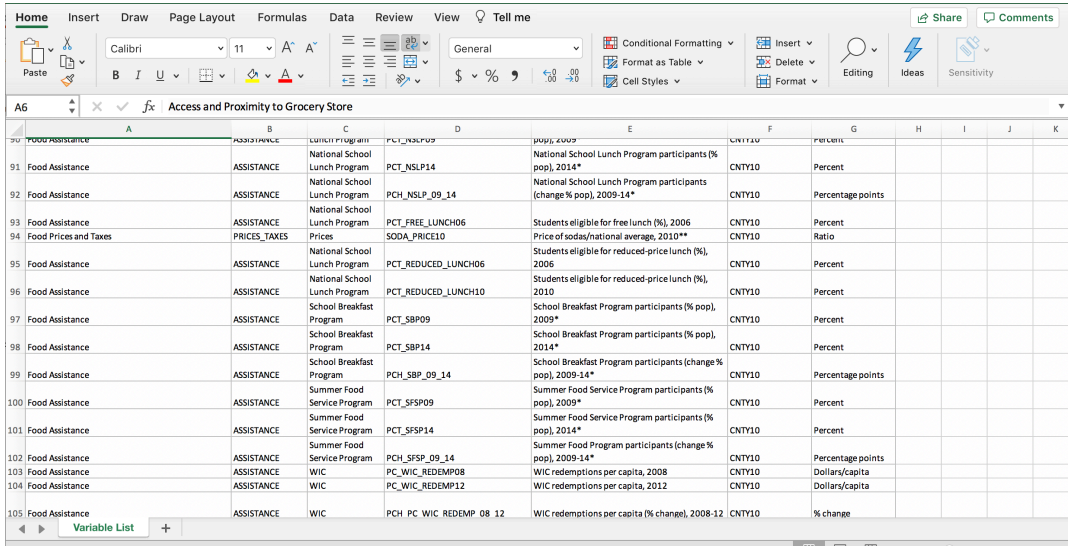
- The data is from the USDA ATLAS
- It is collected from multiple sources and cover a range of years and geographic levels.
- Obesity rates in the data set used in the paper are from 2010 released in 2011 by the USDA.
- The data comprises of two hundred and eleven variables in three categories-
  - Food Choices- the proximity of the population divided by counties to healthy, affordable food
  - Health and Well being- describes diabetes, and obesity rates
  - Community Characteristics- demography divided by age and race, income and poverty, metro and non metro status of a county, availability of recreational facilities.

The Data



FIPS	State	County	FFR09	FFR14	PCH_FFR_09_14	FFRPTH09	FFRPTH14	PCH_FFRPTH_09_14	FSR09	FSR14	PCH_FSR_09_14	FSRPTH09	FSRPTH14	PCH_FSRPTH_09_14
01001	AL	Autauga	30	36	20	0.55417013	0.649878148	17.27051178	34	29	-14.70588235	0.628059481	0.523512952	-16.64595977
01003	AL	Baldwin	112	132	17.85714286	0.624282354	0.659633903	5.662750031	202	221	9.405940594	1.125937817	1.104387065	-1.914026824
01005	AL	Barbour	21	22	4.761904762	0.759301443	0.818239298	7.762115521	12	15	25	0.433886539	0.55789043	28.57979693
01007	AL	Bibb	7	5	-28.57142857	0.305130552	0.222162979	-27.19084435	6	5	-16.66666667	0.261540473	0.222162979	-15.05598507
01009	AL	Blount	24	21	-12.5	0.418548682	0.363831667	-13.07303488	19	15	-21.05263158	0.33135104	0.259879762	-21.56965553
01011	AL	Bullock	4	3	-25	0.364066624	0.2787068	-23.44620959	2	1	-50	0.182033312	0.092902267	-48.96413973
01013	AL	Butler	17	17	0	0.814683472	0.837603469	2.813362239	19	10	-47.36842105	0.910528586	0.492707923	-45.88770409
01015	AL	Calhoun	95	103	8.421052632	0.802615682	0.888574485	10.70983344	67	77	14.92537313	0.566055271	0.66427413	17.35146089
01017	AL	Chambers	22	26	18.18181818	0.63983248	0.763000352	19.25001868	19	16	-15.78947368	0.552582597	0.469538678	-15.02832677
01019	AL	Cherokee	16	15	-6.25	0.618859751	0.576103238	-6.908918078	10	11	10	0.386787344	0.422475708	9.226869455
01021	AL	Chilton	23	20	-13.04347826	0.528930181	0.455259384	-13.92826498	17	18	5.882352941	0.390948395	0.409733446	4.804994999
01023	AL	Choctaw	4	6	50	0.286204923	0.45034902	57.35194776	7	5	-28.57142857	0.500858615	0.37529085	-25.07050107
01025	AL	Clarke	23	26	13.04347826	0.883324372	1.042293045	17.99663608	13	15	15.38461539	0.499270297	0.60132291	20.44035339
01027	AL	Clay	4	4	0	0.285591889	0.295159386	3.350059032	7	6	-14.28571429	0.499785806	0.442739079	-11.41423512
01029	AL	Cleburne	7	8	14.28571429	0.469231801	0.530503979	13.05797651	4	2	-50	0.268132457	0.132625995	-50.53713528
01031	AL	Coffee	35	34	-2.857142857	0.707928803	0.667858335	-5.660239699	31	28	-9.677419355	0.627022654	0.550000982	-12.28371433
01033	AL	Colbert	41	52	26.82926829	0.75331643	0.953376235	26.55720727	40	38	-5	0.734942858	0.696698018	-5.20378417
01035	AL	Conecuh	5	7	40	0.376676209	0.552486188	46.67403315	4	4	0	0.301340967	0.315706393	4.767166535
01037	AL	Coosa	0	1		0	0.091861106		0	0	0	0	0	
01039	AL	Covington	25	24	-4	0.664045899	0.633011552	-4.673524292	24	18	-25	0.637484063	0.474758664	-25.52619085
01041	AL	Crenshaw	3	5	66.66666667	0.213918996	0.357730557	67.22711121	9	6	-33.33333333	0.641756988	0.429276669	-33.10915552
01043	AL	Cullman	50	53	6	0.622269791	0.651994735	4.776857877	44	48	9.090909091	0.547597416	0.590485798	7.832100732
01045	AL	Dale	33	34	3.03030303	0.661932844	0.687090777	3.800677538	31	30	-3.225806452	0.621815702	0.606256568	-2.502209903
01047	AL	Dallas	20	19	-5	0.457048836	0.455515332	-0.335523004	16	15	-6.25	0.365639069	0.359617367	-1.646897701
01049	AL	DeKalb	41	43	4.87804878	0.578573041	0.605079856	4.58141207	35	31	-11.42857143	0.493903816	0.436220362	-11.67908655
01051	AL	Elmore	29	45	55.17241379	0.368806593	0.555713351	50.67880066	34	38	11.76470588	0.432393936	0.469269052	8.528129629
01053	AL	Escambia	22	27	22.72727273	0.575178436	0.715554024	24.40557217	20	20	0	0.522889487	0.530040018	1.367503246
01055	AL	Etowah	82	79	-3.658536585	0.786653748	0.763056476	-2.999702457	54	57	5.555555556	0.518040273	0.550559736	6.277400542
01057	AL	Fayette	12	9	-25	0.693240901	0.53336494	-23.06210738	5	9	80	0.288850376	0.53336494	84.65094228
01059	AL	Franklin	22	18	-18.18181818	0.696444965	0.569602228	-18.21288739	17	16	-5.882352941	0.538162018	0.506313091	-5.918092689
01061	AL	Geneva	10	10	0	0.375192286	0.374363582	-0.220874513	8	11	37.5	0.300153829	0.41179994	37.19629754
01063	AL	Greene	1	1	0	0.108908734	0.11691804	7.354144745	3	2	-33.33333333	0.326726203	0.233836081	-28.43057017
01065	AL	Hale	5	7	40	0.31201248	0.461011591	47.75421496	5	4	-20	0.31201248	0.263435195	-15.56902002
01067	AL	Henry	10	10	0	0.579945485	0.581733566	0.30831879	9	11	22.22222222	0.521950937	0.639906923	22.5990563
01069	AL	Houston	83	97	16.86746988	0.825870647	0.930964652	12.725238	75	77	2.666666667	0.746268657	0.739013177	-0.972234219
01071	AL	Jackson	30	29	-3.333333333	0.561156731	0.550650337	-1.872274439	29	25	-13.79310345	0.542451507	0.474698566	-12.49013773

Read_Me | Variable List | Supplemental Data - County | Supplemental Data - State | ACCESS | STORES | RESTAURANTS | ASSISTANCE | INSECURITY | PRICES_TAXES | LOCAL



91	Food Assistance	ASSISTANCE	National School Lunch Program	PCT_NSLP14	National School Lunch Program participants (% pop), 2014*	CNTY10	Percent
92	Food Assistance	ASSISTANCE	National School Lunch Program	PCH_NSLP_09_14	National School Lunch Program participants (change % pop), 2009-14*	CNTY10	Percentage points
93	Food Assistance	ASSISTANCE	National School Lunch Program	PCT_FREE_LUNCH06	Students eligible for free lunch (%), 2006	CNTY10	Percent
94	Food Prices and Taxes	PRICES_TAXES	Prices	SODA_PRICE10	Price of sodas/national average, 2010**	CNTY10	Ratio
95	Food Assistance	ASSISTANCE	National School Lunch Program	PCT_REDUCED_LUNCH06	Students eligible for reduced-price lunch (%), 2006	CNTY10	Percent
96	Food Assistance	ASSISTANCE	National School Lunch Program	PCT_REDUCED_LUNCH10	Students eligible for reduced-price lunch (%), 2010	CNTY10	Percent
97	Food Assistance	ASSISTANCE	School Breakfast Program	PCT_SBP09	School Breakfast Program participants (% pop), 2009*	CNTY10	Percent
98	Food Assistance	ASSISTANCE	School Breakfast Program	PCT_SBP14	School Breakfast Program participants (% pop), 2014*	CNTY10	Percent
99	Food Assistance	ASSISTANCE	School Breakfast Program	PCH_SBP_09_14	School Breakfast Program participants (change % pop), 2009-14*	CNTY10	Percentage points
100	Food Assistance	ASSISTANCE	Summer Food Service Program	PCT_SFSP09	Summer Food Service Program participants (% pop), 2009*	CNTY10	Percent
101	Food Assistance	ASSISTANCE	Summer Food Service Program	PCT_SFSP14	Summer Food Service Program participants (% pop), 2014*	CNTY10	Percent
102	Food Assistance	ASSISTANCE	Summer Food Service Program	PCH_SFSP_09_14	Summer Food Program participants (change % pop), 2009-14*	CNTY10	Percentage points
103	Food Assistance	ASSISTANCE	WIC	PC_WIC_REDEMP08	WIC redemptions per capita, 2008	CNTY10	Dollars/capita
104	Food Assistance	ASSISTANCE	WIC	PC_WIC_REDEMP12	WIC redemptions per capita, 2012	CNTY10	Dollars/capita
105	Food Assistance	ASSISTANCE	WIC	PCH_PC_WIC_REDEMP_08_12	WIC redemptions per capita (% change), 2008-12	CNTY10	% change

Variable List

The link for the data download available at https://www.ers.usda.gov/data- products/food- environment- atlas/data- access- and- documentation- downloads/

Pre-Processing

- The data is available across multiple excel sheets.
- Was combined into one comma separated value file containing sixty six relevant columns
  - The general principle of operation for combining the data was, that after importing a spreadsheet, the feature correlation was collected.
  - The features with low correlation were dropped.
- Dependent variable- adult obesity rate*, is for the year 2010.
  - Any data after that year was dropped
  - For data missing for the year 2010, the available data closest to 2010 was used

*available from the health and physical activity data

Training and Testing

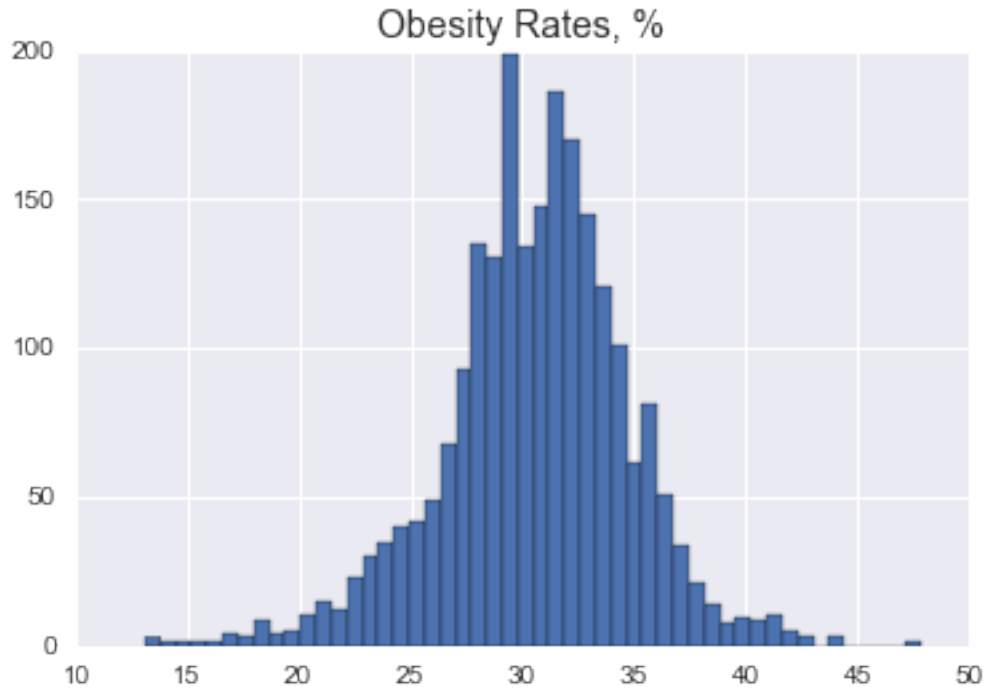
- EDA of Training Data
- Modeling
- Testing

Training and Testing

- **EDA of Training Data**
- Modeling
- Testing

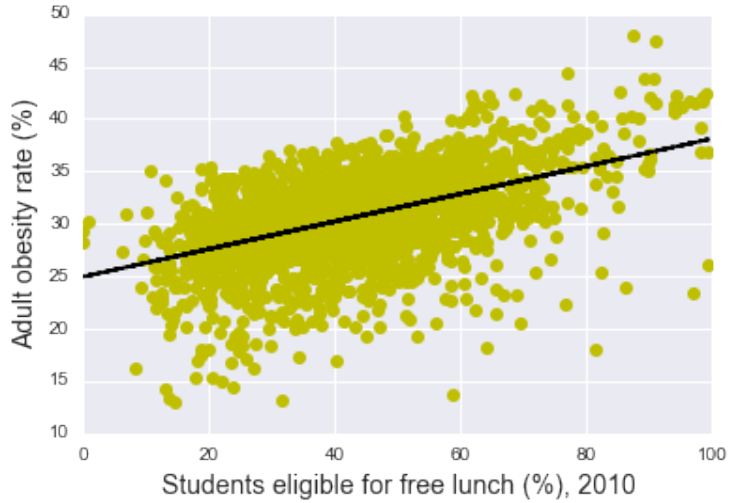
EDA of Training Data

- Outliers- using Histogram analysis



EDA of Training Data

- Correlation of the Adult Obesity Percent to Low income households
- All 'Access and Proximity' predictor variables with the exception of *PCT_LACCESS_HHNV10* are highly correlated



EDA of Training Data

average of obesity rate for Persistent-poverty counties, 2010

PERPOV10	PCT_OBESE_ADULTS10
0	30.072933
1	34.618504

average of obesity rate for Persistent-child-poverty counties, 2010

PERCHLDPOV10	PCT_OBESE_ADULTS10
0	29.774134
1	33.377470

EDA of Training Data

Test for Multicollinearity

Eigen values and Eigen Vectors were used

- The Eigen vectors were found for the features who Eigen values were close to zero

	epigen_value	feature
27	0.000595	PCT_WIC09
28	0.001541	PCT_CACFP09
29	0.007749	FOODINSEC_07_09

WIC participants (% pop), 2009

Child & Adult Care (% pop), 2009

Household food insecurity (%, three-year average), 2007-09

- Features with eigen vectors that are NOT close to zero i.e strong correlation
- The best feature was kept for testing and the rest were dropped

Training and Testing

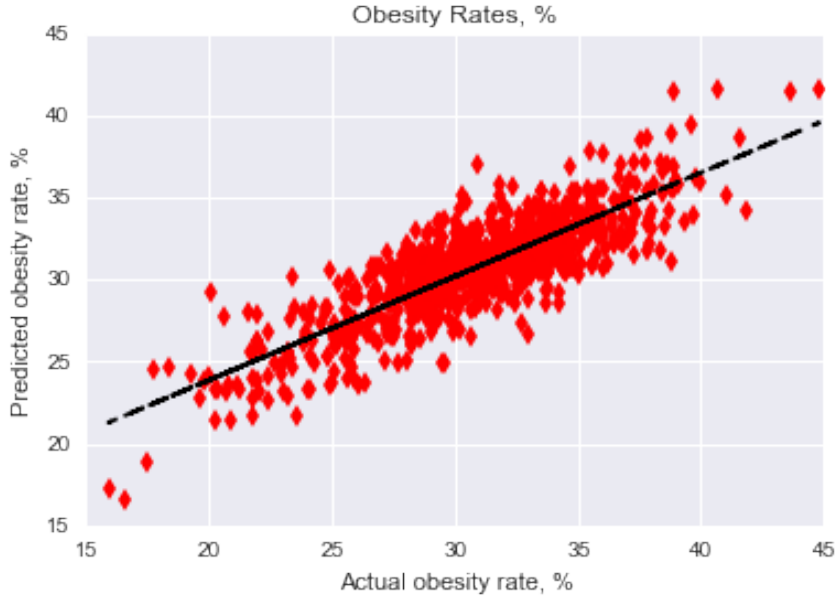
- EDA of Training Data
- **Modeling**
- Testing

Modeling

- Two modeling techniques used
  - Random Forest
  - Linear Regression

Modeling

- Random Forest*
  - Used to find feature importance
  - Estimators (n) =100



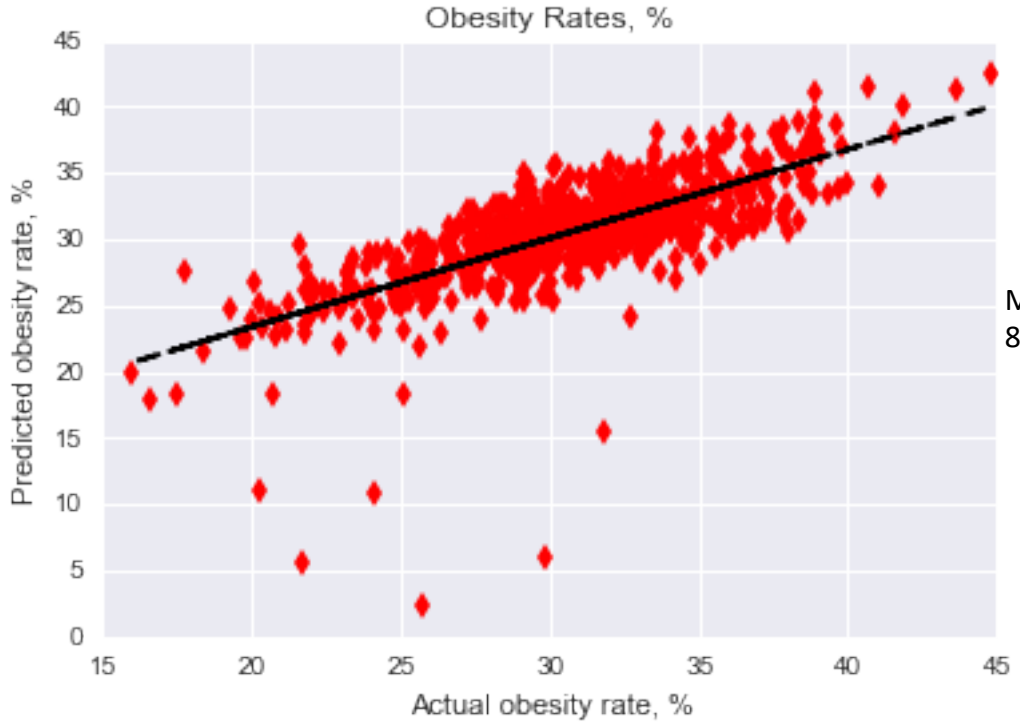
MSE for training set = 5.39
R-squared = 0.685776

	Features	Importance Score
23	PCT_FREE_LUNCH10	0.1057
47	NATAMEN	0.0962
15	PCT_SNAP09	0.0948
49	PCT_NHBLACK10	0.0788
14	PC_FSRSALES07	0.0614
51	PCT_NHASIAN10	0.0470
22	PCT_NSLP09	0.0418
12	FSRPTH07	0.0309
9	SNAPSPTH08	0.0305
52	PCT_NHNA10	0.0210

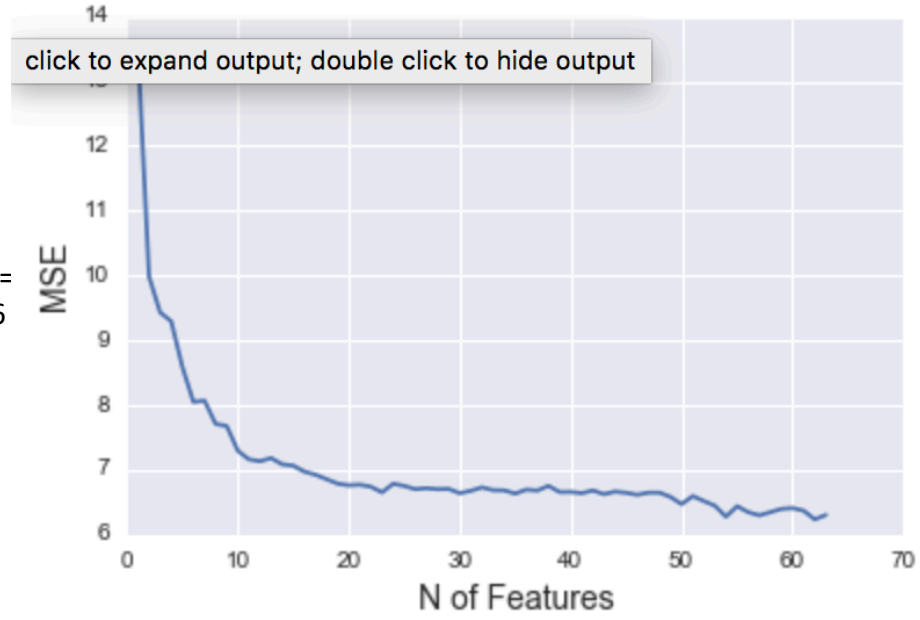
The scikit learn Random forest Regressor was used

Modeling

- Linear Regression Model*



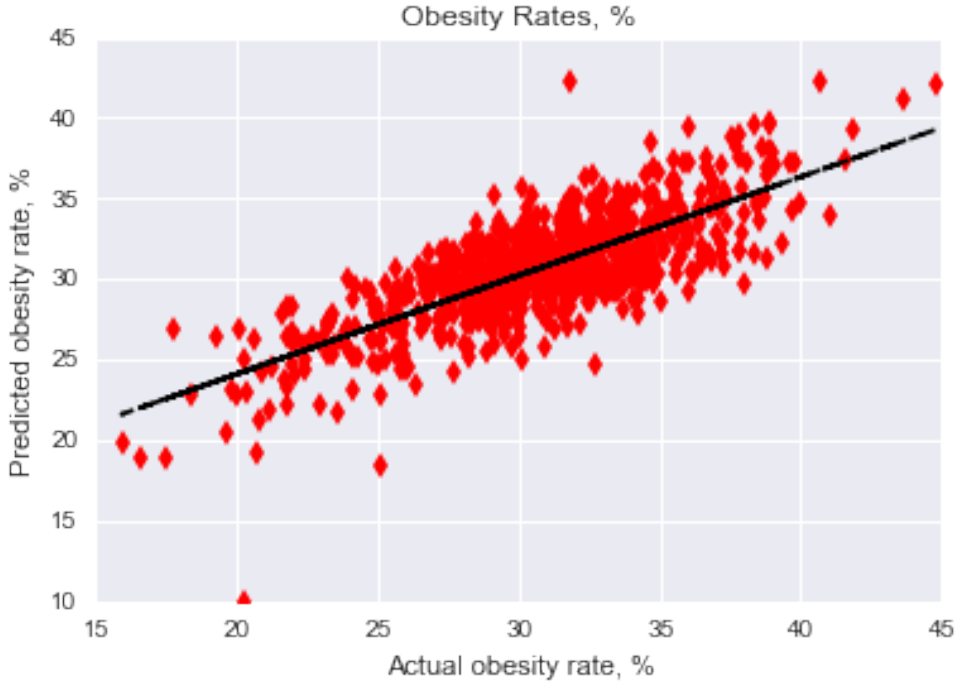
MSE for training set = 8.619311726714356



The LinearRegression module from sklearn.linear_model was used

Modeling

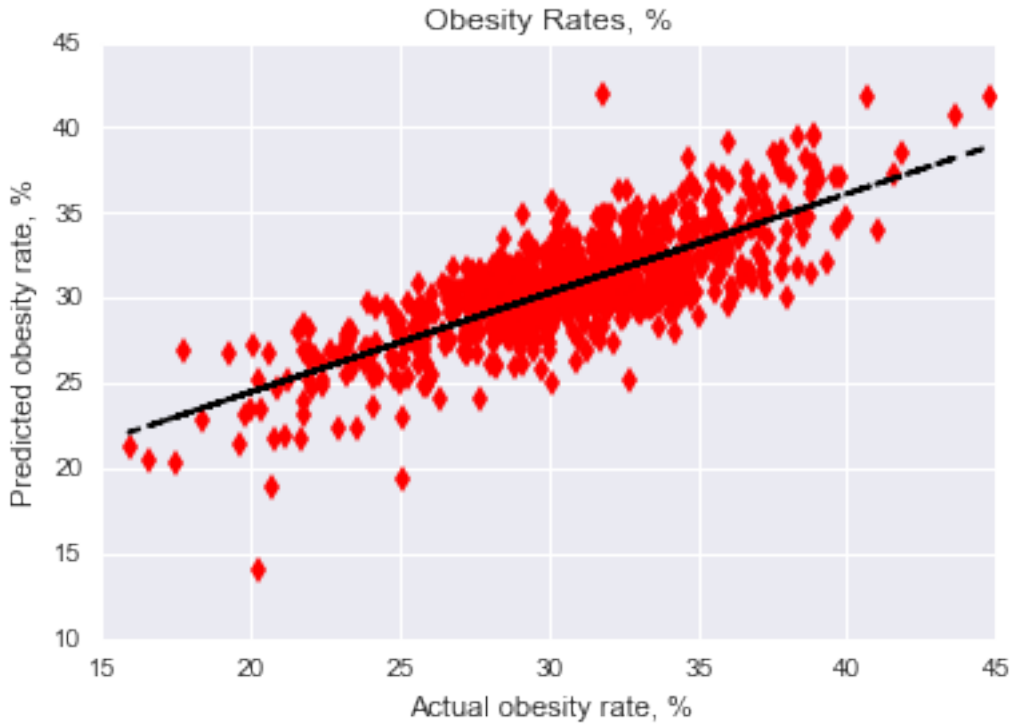
- Linear Regression Model with top 80% features from Random Forest
  - Top 80% of features by importance score
  - 22 features



MSE for training set = 7.660215
R-squared = 0.588249

Analysis

- Linear Regression Model without outliers



Training

- Training model performance comparison

	Model	n_features	MSE
5	Random Forest	63	5.418264
3	Linear Regression _all_features	63	8.619312
2	Linear Regression	22	7.660215
1	Linear Regression_no outliers	22	6.348386

Training and Testing

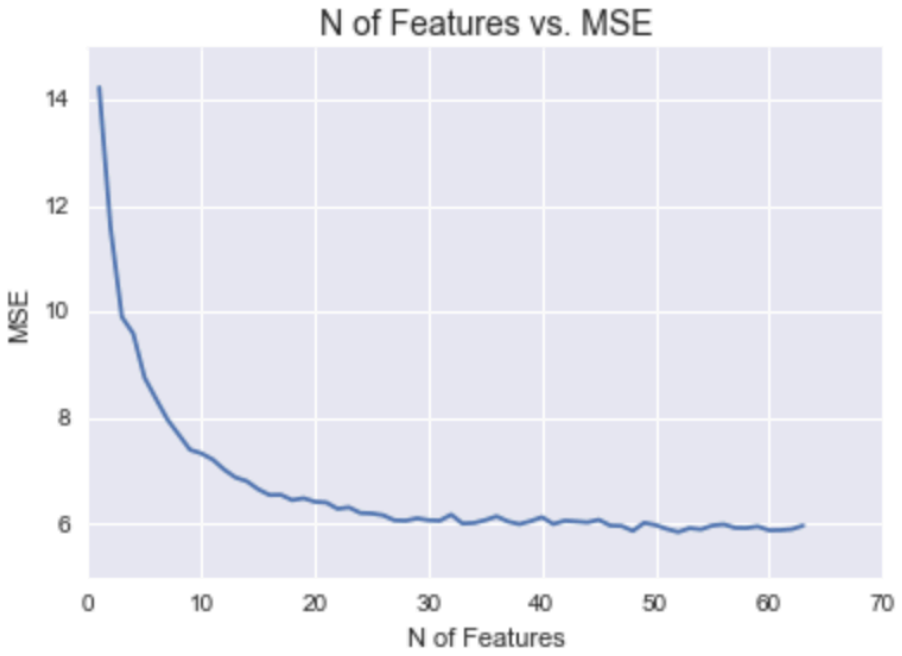
- EDA of Training Data
- Modeling
- **Testing**

Testing

- The trained model was tested with the all non-collinear features and best feature out of collinear feature set.
- MSE calculated - **9.3571769776865317**

Feature Selection

- Technique used - Recursive Feature Elimination*




```
31     FOODINSEC_CHILD_01_07
27                 PCT_WIC09
29           FOODINSEC_07_09
17               SNAP_OAPP10
15                PCT_SNAP09
35             PCT_LOCLFARM07
37                VEG_FARMS07
12                   FSRPTH07
22                PCT_NSLP09
47                     NATAMEN
38           FRESHVEG_FARMS07
49              PCT_NHBLACK10
7                  CONVSPTH07
50                PCT_HISP10
51              PCT_NHASIAN10
52                 PCT_NHNA10
55            PCT_18YOUNGER10
56                 MEDHHINC10
48              PCT_NHWHITE10
23           PCT_FREE_LUNCH10
Name: Feature, dtype: object
```

RFE available from the sklearn.feature_selection library

Conclusion

- Counties with a high poverty rate have around 4-5% higher obesity rate than others.
- Percentage of students in a county who are eligible for free lunch program have the strongest correlation with a predictor variable, that is they are at a higher risk than others to be obese- which is the other hypothesis.
- The feature selection from the Linear Regression model shows that
  - child food insecurity percentage in households aggregated over the years 2001-07 is the highest ranked feature. This is significant because the response variable adult obesity rateis for the year 2010.
  - Inferences made from the other top features such as- WIC participants from the year 2009, state widefood insecurity aggregated over 2007-09 and SNAP participants from year 2009 also verify the initial hypothesis

THANK YOU!