The Effect of Food Access on Overweight and Obesity Levels in Colorado

Obesity levels in the US are alarmingly high and continue to rise. For 2015-2016, the Center for Disease Control and Prevention (CDC) found that 39.8% of American adults were obese, up from 33.7% in 2007-2008 ("Adult Obesity Facts"). The effects of obesity are at least three-fold. First, obesity is part of a general lack of physical health in the American population. Second, obesity-related conditions like heart disease, stroke, type-2 diabetes, and certain types of cancer can lead to shorter life span and keep individuals unhealthy. Third, the epidemic puts a strain on the US healthcare industry. The estimated annual medical cost of obesity in the US was $147 billion in 2008 (Finkelstein).

While obesity is increasing for Americans in general, researchers have found that the rise is disproportionately affecting low-income, Black, and Hispanic Americans. The CDC cites that Hispanics and non-Hispanic Blacks have the highest prevalence of obesity, and that men and women with college degrees had lower obesity prevalence than those with a college education ("Adult Obesity Facts"). Studies have shown that low-income communities have higher levels of childhood obesity ("Low-Income Communities More Likely to Face Childhood Obesity"), and that race and ethnicity is associated with higher levels of obesity (Crossow).

One suggested explanation for why low-income, Black, and Hispanic populations have higher prevalence of diabetes is that these groups might not have access to healthy food, geographically speaking. Regions with a lack of access to grocery stores and healthy food providers are known as food deserts. When people live in food deserts, it's possible that they rely on gas stations and fast food, which sell foods that contribute to the obesity epidemic ("USDA Defines Food Deserts").

Many studies have been done to investigate whether communities with low access to food has an effect on obesity, and results are mixed. Research by Ghosh-Dastidar et al. found that distance to store and higher food prices were positively associated with obesity in urban food deserts. But a comprehensive review of existing research by Cobb et al. found that there was not an association, although they suggested much of the research has not been productive, and more research needs to be done.

The goal of this project is to further research in this area by building a linear regression model to find whether low access to food is related to obesity levels in Colorado neighborhoods, and what other characteristics are associated with high obesity levels. In order to achieve this, we use data from two sources. The first is a dataset with estimates for the obesity rates amongst adults in Colorado neighborhoods, called "Obesity in Adults – Colorado BRFSS" from the ArcGIS Hub. The second dataset is the "Food Access Research Atlas" from the United States Department of Agriculture, which covers a wide range of census data like race and income, along with several variables related to food access, based on the distance from households in neighborhoods to the nearest grocery store.

Feature Generation, Exploratory Data Analysis

After joining the datasets described above, we have the following variables, for each neighborhood, or census tract, in Colorado.

Variable Name	Description
Urban	A flag indicating whether or not the census tract is urban
POP2010	The 2010 population
PovertyRate	The poverty rate
MedianFamilyIncome	The median family income
TractLOWI	The low income population
LA1and10	A flag for low food access at 1 mile for urban areas and 10 miles for rural areas

LAhalfand10	A flag for low food access at ½ mile for urban areas and 10 miles for rural areas
LA1and20	A flag for low food access at 1 mile for urban areas and 20 miles for rural areas
TractWhite	The number of residents who identify as White
TractBlack	The number of residents who identify as Black
TractAsian	The number of residents who identify as Asian
TractOMultir	The number of residents who identify as Multiracial
TractHispanic	The number of residents who identify as Hispanic
TractHUNV	The number of housing units with no vehicle
TractSNAP	The number of housing units receiving SNAP benefits
OHU2010	The total number of housing units
Obese_Census_Tract_Estimate	An estimate for the obese population

To start, we look at a summary of each variable to see if there are any implausible values which point to missing data or data entry errors, and find several instances where this occurs. We see that there are some observations where the population is zero, and we remove these observations from the dataset. There are also observations where the median family income is zero. For these observations, we set the median family income to be the median of the median family income for all other census tracts in the given county. There are several tracts where the number of housing units is zero, and we replace these values with the mean number of housing units per population size for the entire population, multiplied by the population in the given tract. Finally, there are two census tracts where the estimated number of obese adults is -1, which we assume implies that the data is missing and we remove those observations from the data.

Now that the data are clean, we modify the variables of interest to be proportions by dividing by the tract populations, so that comparing values across tracts is meaningful. The list of variables now under consideration is as follows:

Variable Name	**Description**
Urban	A flag indicating whether or not the census tract is urban
PovertyRate	The poverty rate
MedianFamilyIncome	The median family income
PercentLowIncome	The proportion of population that is low income

LA1and10	A flag for low food access at 1 mile for urban areas and 10 miles for rural areas
LAhalfand10	A flag for low food access at ½ mile for urban areas and 10 miles for rural areas
LA1and20	A flag for low food access at 1 mile for urban areas and 20 miles for rural areas
PercentWhite	The proportion of population that identifies as White
PercentBlack	The proportion of population that identifies as Black
PercentAsian	The proportion of population that identifies as Asian
PercentOMultir	The proportion of population that identifies as Multiracial
PercentHispanic	The proportion of population that identifies as Hispanic
PercentHUNV	The proportion of housing units with no vehicle
PercentSNAP	The proportion of housing units receiving SNAP benefits
PercentObese (response variable)	The estimated proportion of population that is obese

We continue to explore the data by making bivariate graphics. While not all variables appear to be related to the obesity level, there are some notable observations. First, the obesity level appears to be slightly higher in rural areas than in urban areas (Fig. 1). There also appears to be an association between the obesity level and income-related variables such as PovertyRate, MedianFamilyIncome, and PercentSNAP (Fig. 2, 3, and 8, respectively), which appears to confirm the existing research results, showing that obesity disproportionately affects low income individuals. For each of the three variables related to food access, it appears that low access to food is associated with higher obesity levels, but only slightly (Fig. 4 - 6). The only variable related to race which appears clearly associated with obesity levels is PercentHispanic (Fig. 7). This could be a result of genetic disposition, which we won't explore here, or it could be a result of the correlation with income-related variables. Either way, it appears to confirm what earlier research suggests – that obesity disproportionately affects Hispanic individuals. With these observations in mind, we begin the process of constructing a linear regression model that will predict obesity levels in Colorado neighborhoods.

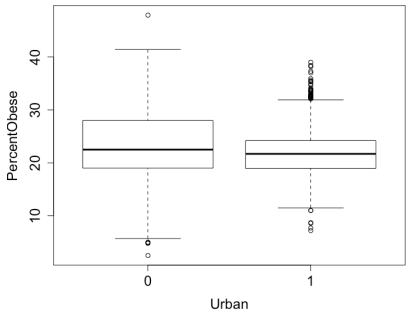

Figure 1. Relationship between neighborhoods in urban areas and obesity level

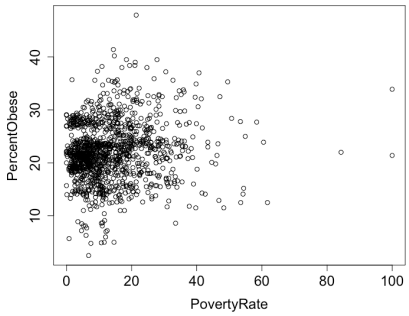

Figure 2. Relationship between poverty rate and obesity level

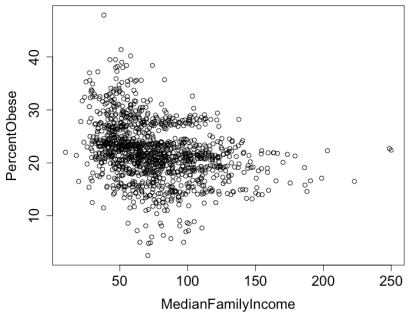

Figure 3. Relationship between median family income and obesity level

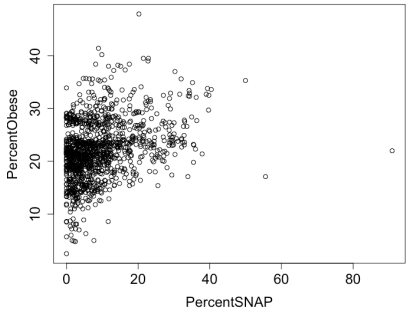

Figure 4. Relationship between proportion of population receiving SNAP benefits and obesity level

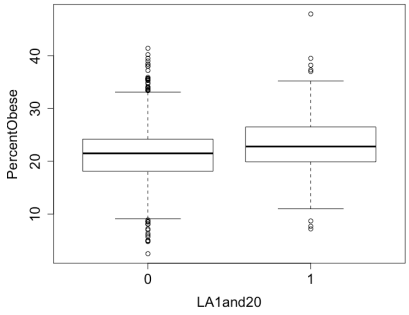

Figure 5. Relationship between low food access (at 1 mile for urban, 20 miles for non-urban areas) and obesity level

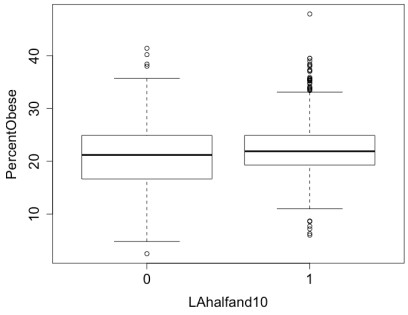

Figure 6. Relationship between low food access (at 1/2 mile for urban, 10 miles for non-urban areas) and obesity level

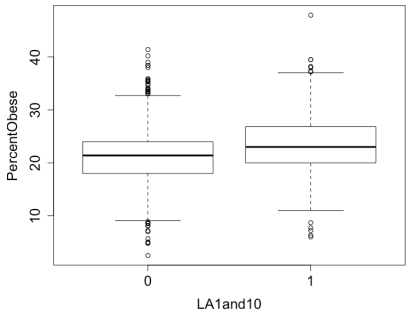

Figure 7. Relationship between low food access (at 1 mile for urban, 10 miles for non-urban areas) and obesity level

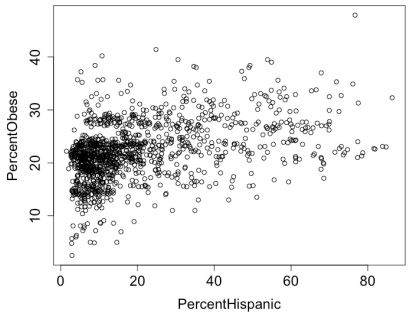

Figure 8. Relationship between Hispanic population and obesity level

Addressing Collinearity

We start by fitting a model with all variables in Fig. 2, and removing them as necessary based on tests for collinearity. Many of the variables are collinear, since they measure similar things (for example, PercentLowIncome and MedianFamilyIncome both measure income), so several of them will need to be removed.

To start, we look at the Variance Inflation Factors (VIF) of the predictors (Fig. 9). Several of the VIF's are greater than 5, which indicates collinearity between the predictors. It makes sense that the various food access indicators are collinear, so we'll remove two of them and keep one in the model. Since LA1and10 is the most highly correlated with the response variable, we'll keep that one and take out the others. The race proportion variables are all also very highly correlated. Since prior research indicated that lack of food access may be associated with Black and Hispanic populations, we'll take out all variables related to race except PercentBlack and PercentHispanic. After the removal of those variables, we look at the VIF's again and see they're all under 5. We also look at the condition numbers, see that they're all under 30, and conclude that they don't indicate any further evidence of colliearity. However, pairwise correlations show that the variables PercentLowIncome, PovertyRate, PercentSNAP, and MedianFamilyIncome are highly correlated, which we anticipate since all of these variables are related to income. We find that PercentSNAP and MedianFamilyIncome are the two variables from that group that have the highest correlation with the response, so we decide to leave those in the model and remove PercentLowIncome and PovertyRate.

Predictor	VIF	Predictor	VIF
PercentSNAP	3.7474	PercentAsian	23.4608
PercentOMultir	24.3428	PercentWhite	3.0750
MedianFamilyIncome	2.6974	LAhalfand10	3.7318
LA1and20	10.3008	Urban	13.7948
PovertyRate	3.0525	PercentHispanic	7.0193
PercentHUNV	1.9909	PercentBlack	5.4278
LA1and10	2.2655	PercentLowIncome	10.8557

Figure 9. Variance inflation factors (VIF) for the variables when they're all included in the model

Predictor	VIF	Predictor	VIF
PercentSNAP	3.5025	PercentHispanic	2.3606
MedianFamilyIncome	2.1342	PovertyRate	2.4099
PercentHUNV	1.8824	PercentBlack	1.1592
LA1and10	1.1073	Urban	1.2041

Figure 10. VIF's after removing some collinear variables from the model

Variable Name	Description
Urban	A flag indicating whether or not the census tract is urban
LA1and10	A flag for low food access at 1 mile for urban areas and 10 miles for rural areas
MedianFamilyIncome	The median family income
PercentBlack	The proportion of population that identifies as Black
PercentHispanic	The proportion of population that identifies as Hispanic
PercentHUNV	The proportion of housing units with no vehicle
PercentSNAP	The proportion of housing units receiving SNAP benefits
PercentObese (response variable)	The estimated proportion of population that is obese

Figure 11. Variables remaining in the model after removing all collinear variables

Model Selection

After addressing collinearity, our list of predictors is as listed in Figure 11. We proceed with trying to find a subset of these predictors which is optimal for predicting the response by using various model selection methods.

First, we perform a backward selection by removing variables from the fitted full model (Model 1 in Fig. 12) with the highest P-values, given that the P-values greater than or equal to 0.05, meaning that there is not sufficient evidence to conclude that the coefficients are not zero. The only non-significant predictor removed from the model by this method is PercentBlack, which results in Model 2 (Fig. 13).

We then compute the Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC) for the full model with each subset of predictors (Fig. 15 and Fig. 16). The AIC is lowest when all predictors are included in the model, which supports Model 1, and the BIC supports Model 3, which doesn't have PercentBlack or MedianFamilyIncome.

Finally, we compute Mallow's Cp for each subset (Fig. 17), which suggests that we leave all of the variables in the model (Model 1). The adjusted R-squared supports that model as well, as does stepwise selection with AIC. However, stepwise selection with BIC suggests Model 3.

Predictor	Est. Coef.	Pr(>|t|)
(Intercept)	21.60742	< 2e-16
PercentSNAP	0.10695	0.00014
PercentHUNV	-0.15892	1.0e-07
PercentHispanic	0.10661	< 2e-16
PercentBlack	-0.04334	0.05955
MedianFamilyIncome	-0.01557	0.01031
LA1and10	1.82550	7.4e-10
Urban	-1.44167	8.7e-05

Figure 12. Model 1, R-squared = 0.247

Predictor	Est. Coef.	Pr(>|t|)
(Intercept)	21.51809	< 2e-16
PercentSNAP	0.09888	0.00037
PercentHUNV	-0.16127	6.6e-08
PercentHispanic	0.10825	< 2e-16
MedianFamilyIncome	-0.01458	0.01596
LA1and10	1.83488	6.2e-10
Urban	-1.56657	1.5e-05

Figure 13. Model 2, R-squared = 0.2

Predictor	Est. Coef.	Pr(>|t|)
(Intercept)	20.1216	< 2e-16
PercentSNAP	0.1210	4.2e-06
PercentHUNV	-0.1549	2.0e-07
PercentHispanic	0.1155	< 2e-16
LA1and10	1.8689	3.1e-10
Urban	-1.6979	2.2e-06

Figure 14. Model 3, R-squared = 0.241

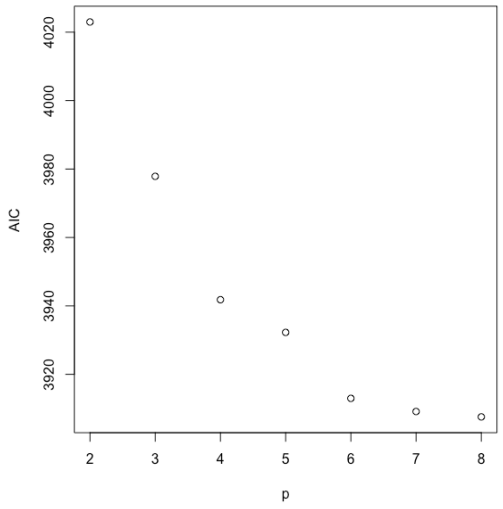

Figure 15. Plot of AIC for each subset of predictors

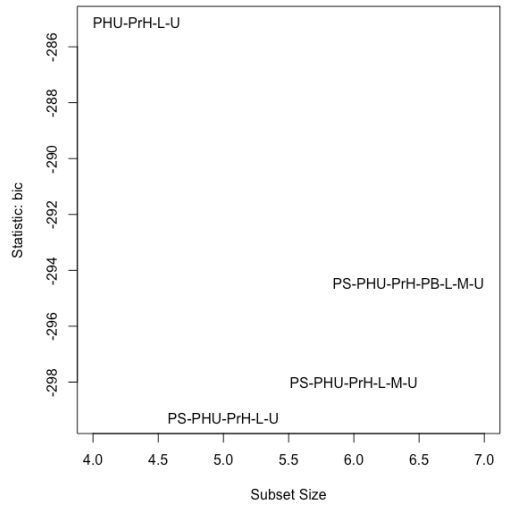

Figure 16. Plot of BIC for each subset of predictors

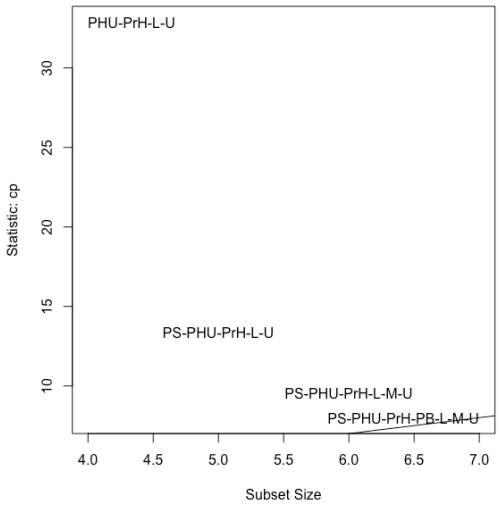

Figure 17. Mallow's CP statistic for each subset of predictors

Each of these models are valid, but we need to pick one to use moving forward. To see if one performs significantly better than the others, we use a 10-fold cross validation and compare the root mean square error (RMSE) and the mean absolute error (MAE) across the models (Fig. 18). Model 1 performs better than Model 2, which in turn performs better than Model 3, but only slightly so. Since the models all score about the same, we stick with Model 3, in an effort to keep the model parsimonious. This decision is supported by the fact that PercentBlack wasn't significant as a predictor in the full model, so there isn't evidence to conclude that it should be included in the model, and MedianFamilyIncome was found to be highly correlated with PercentSNAP, which will remain in the model.

Model	Root mean square error (RMSE)	Mean absolute error (MAE)
Model 1	4.8274	3.5825
Model 2	4.8326	3.5821
Model 3	4.8331	3.5922

Figure 18. Results of a 10-fold cross validation for the three models

Model Structure

Now that we've arrived at a model, we use several methods to check the structure of the model. We begin this process by looking at Component Plus Residual plots to see if any transformations are suggested (Fig 19). There is curvature in PercentHispanic, and some deviation in PercentSNAP. To address these, we try a log transformation of the PercentHispanic variable, refit the model, and observe that the curvature in residuals for that variable appears to be resolved (Fig. 20).

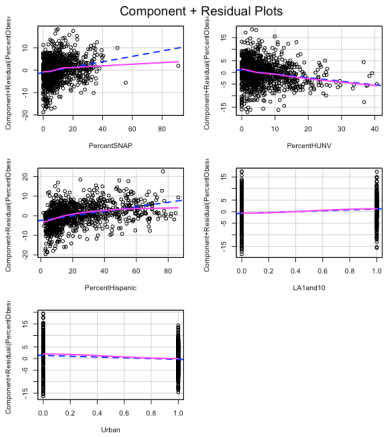

Figure 19. Component plus residual plots for the predictors in the model

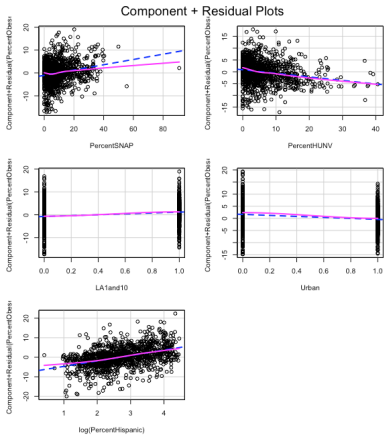

Figure 20. Component plus residual plots after adding a log transformation to PercentHispanic

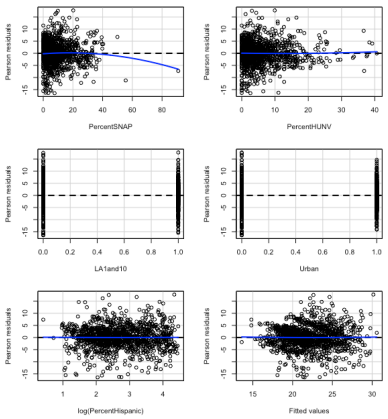

Figure 21. Residual plots for model predictors

To address the deviation in the plot for PercentSNAP, (although it's possible that it could be caused by leverage points or outliers), we perform a lack-of-fit test, which suggests to leave the model as it is. Next, we check the residual plots to make sure there's no systematic curves or patterns (Fig. 21). We see that the residuals all seem centered at zero and there are no systematic curves or patterns, except for PercentSNAP, which we think could be a result of leverage points or outliers. Finally, we use Tukey's test for non-additivity to see whether a polynomial term of the response should be added to the model and see that it should not.

We also consider whether interaction terms are necessary for the categorial and quantitative predictors by creating scatterplots where the data are split by the values of the two categorical predictors (Fig. 22 – 27). We see that there may be a difference in slope of the plotted points when PercentSNAP is plotted against the obesity level and the data is split by Urban. To try to address this, we add an interaction term between PercentSNAP and Urban, but find that it's not significant. Further, with the interaction term included, residual plots show patterns and Tukey's test is significant, meaning there's curvature in the data that is no longer explained by our model. Therefore, we move forward without the interaction term. None of the other plots indicate that other interaction terms are necessary.

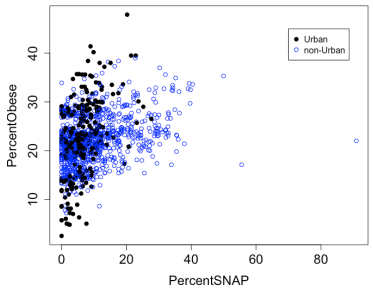

Figure 22. obesity level vs PercentSNAP, split by Urban

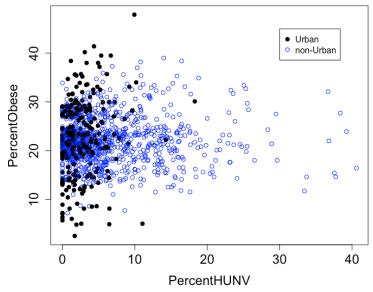

Figure 23. obesity level vs PercentHUNV, split by Urban

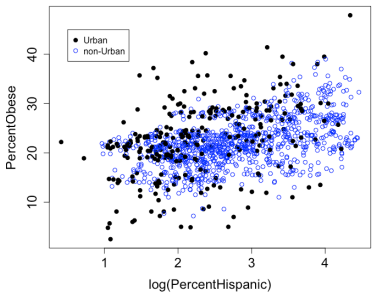

Figure 24. obesity level vs log(PercentHispanic), split by Urban

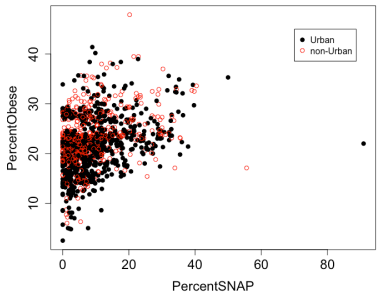

Figure 25. obesity level vs PercentSNAP, split by LA1and10

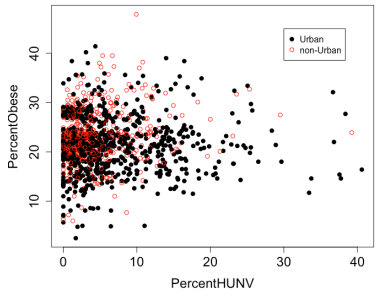

Figure 26. obesity level vs PercentHUNV, split by LA1and10

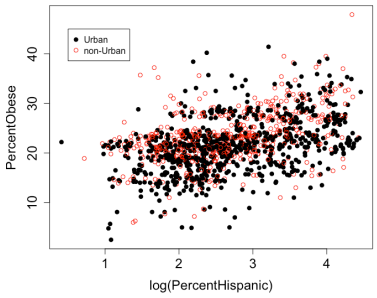

Figure 27. obesity level vs log(PercentHispanic), split by LA1and10

Now that the structure of the model has been verified, we check again for collinearity. The variance inflation factors, condition numbers, and pairwise correlations show no collinearity between variables that needs to be addressed. Repeating the model selection methods, including backward selection, AIC, BIC, Mallow's CP, and Adjusted R-squared, all recommend to keep our current model without removing variables. We feel confident in our model and move on to considering leverage points, outliers, and influential observations.

Leverage Values, Outliers, and Influential Observations

We look for leverage points by making a half-normal plot of the leverage values (Fig. 28) and see two leverage points. To make sure that these leverage points because of a data entry

error or missing data, we look revisit the data for those tracts, but nothing stands out. Next, we perform an outlier test, which shows that there are no studentized residuals that are significant with a Bonferroni $p < 0.05$, meaning that there don't appear to be any outliers in our model. Finally, we look for influential observations by creating a half-norm plot of the observations using Cook's distance and see three points that could be influential, two of which were points we considered might be leverage points (Fig. 29).

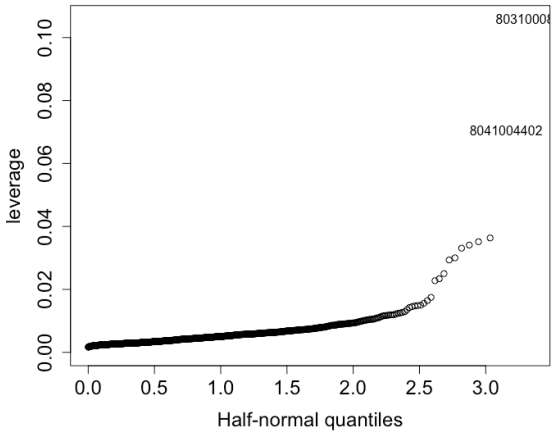

Figure 28. Half-normal plot of leverage values

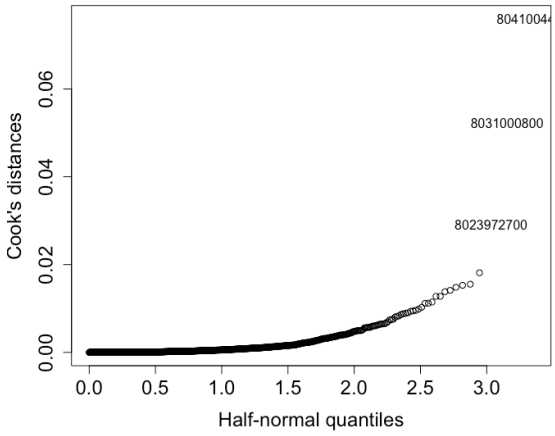

Figure 29. Half-normal plot of Cook's distances

To see if the potential influential points are, in fact, influencing our model, we remove them one at a time from the model to see if the model coefficients change in a way that is significant or would change the inferences we draw from the model. We see that removing these points doesn't change any of the coefficients or our interpretation in a significant way. Since it also isn't clear that these points stand out because of data entry errors or missing data, we decide to leave them in the model and move on to checking error assumptions.

Checking error assumptions

An assumption of linear regression is that the errors are normally distributed, with mean zero and constant variance. We need to check that our model adheres to these assumptions. We do this in two ways: looking at a plot of residuals vs. fitted values (Fig. 30) and by plotting residuals vs. the predictors (Fig. 31).

While it looks like the errors do have a mean of zero, it's not clear that the errors have constant variance, as indicated by the cone-shapes in Fig. 31. In order to address this, we attempt to weights to the least squared model. Unfortunately, weighting by population or the number of housing units doesn't appear to resolve the non-constant variance. Weighting by population density does fix the non-constant variance, but then residual plots and Tukey's test show that the weights introduced curvature in the data. Since this curvature wasn't resolved by adding quadratic terms to the predictors or to the response, we choose to move forward without the weights. As a last effort to address the non-constant variance, we take log and square root transformations of the response, but it still appears that the constant error assumption is violated. Thus, we move forward noting that the errors appear not to have constant variance, and note that this should be addressed in future work.

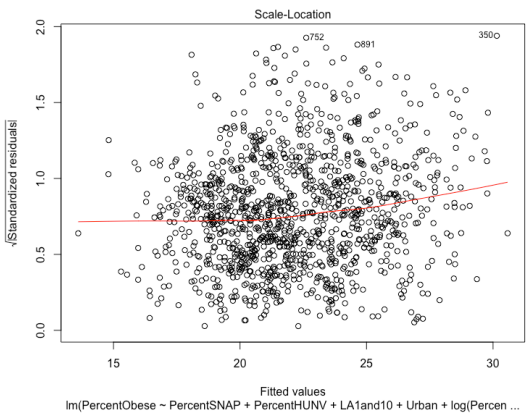

Figure 30. Square-root of standardized residuals vs. fitted values

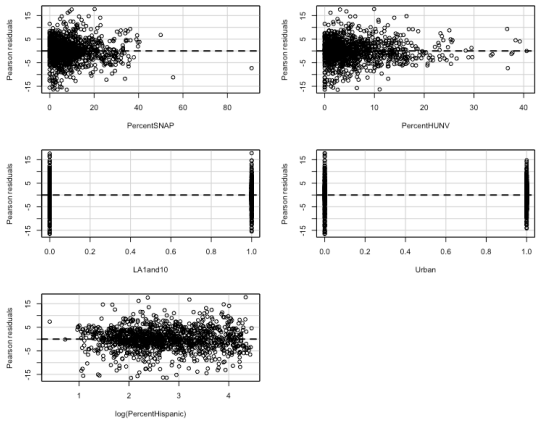

Figure 31. Plots of residuals vs. predictor values

Conclusions and Interpretation

Our final model is described in Figure 32.

Predictor	Est. Coef.	Standard error	Pr(>|t|)
(Intercept)	15.3308	0.5449	< 2e-16
PercentSNAP	0.1154	0.0242	4.2e-06
PercentHUNV	-0.1504	0.0291	2.0e-07
Log(PercentHispanic)	2.7827	0.2283	< 2e-16
LA1and10	1.902	0.2894	3.1e-10
Urban	-2.0790	0.3548	2.2e-06

Figure 32. Summary of the final model. R-squared = 0.268

Holding other things constant, we expect that obesity levels in Colorado neighborhoods will be higher in neighborhoods with more SNAP recipients. Specifically, if the proportion of SNAP recipients increases by 1, we expect that the level of obesity in the neighborhood will increase by 0.1154%. We also expect that, holding the other variables constant, the obesity level will be higher in neighborhoods with a greater Hispanic population. If the proportion of Hispanic population in a neighborhood increased by 1%, we would expect the obesity level to increase by 0.02743%. For neighborhoods considered as having low access to food, we expect that obesity levels will be 1.902% higher than in neighborhoods not considered as having low food access, holding other things constant. We also expect obesity levels to be lower in neighborhoods considered to be urban, and neighborhoods where there are more housing units without vehicles.

Two of these conclusions confirm the previous research – that we should expect obesity levels are higher in neighborhoods with higher low income and Hispanic populations, while holding the other predictors constant. Another conclusion confirms our suspicion – that we should expect obesity levels to be higher for neighborhoods with low access to food.

With these results in mind, there are several policy changes that we could implement in Colorado neighborhoods in an effort to combat obesity and its disproportionate effect on Black, Hispanic and low income communities. First, since we expect obesity levels to be higher in

neighborhoods that are considered as having low access to food, we can simply increase access to healthy foods. According to the definition used here, that can be done by simply reducing the distance between households in neighborhoods and grocery stores, especially in low-income and non-urban communities. Farmers markets and food co-ops can also be useful for increasing access to healthy foods. We can also incentivize corner stores and gas stations to carry a wider range of healthy fruits and vegetables. Second, since we expect obesity levels to be higher in neighborhoods with more SNAP recipients, we can focus on increasing enrollment in existing health-food education programs for SNAP recipients such as SNAP-Ed. Since there's a correlation between Hispanic population and the low income population, implementing these policies might not only reduce obesity levels in Colorado neighborhoods overall, but would also eliminate the relationship between obesity levels and the Hispanic population.

Summary

We should note that the R-squared value of our model is 0.268, which means that our model only explains about 27% of the observed variance. Thus, there's more to the picture than what our model explains. Further, the data used here were observational, meaning there are likely lurking and confounding variables affecting the results. And finally, since this is spatial data, we should consider spatial auto-correlation, which we did not do. With these things in mind, this project is still useful, because it gives us reason to believe that there may be a relationship between obesity levels and food access. It's certainly worthwhile to invest resources in continuing this research and implementing policies that address these factors to ultimately lower obesity levels in Colorado neighborhoods.

Works Cited

"Adult Obesity Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 Aug. 2018, www.cdc.gov/obesity/data/adult.html.

Cobb, Laura K., et al. "The Relationship of the Local Food Environment with Obesity: A Systematic Review of Methods, Study Quality, and Results." *Obesity*, vol. 23, no. 7, 2015, pp. 1331–1344., doi:10.1002/oby.21118.

Cossrow, Nicole, and Bonita Falkner. "Race/Ethnic Issues in Obesity and Obesity-Related Comorbidities." *The Journal of Clinical Endocrinology & Metabolism*, vol. 89, no. 6, 2004, pp. 2590–2594., doi:10.1210/jc.2004-0339.

Finkelstein, Eric A., et al. "Annual Medical Spending Attributable To Obesity: Payer-And Service-Specific Estimates." *Health Affairs*, vol. 28, no. Supplement 1, 2009, doi:10.1377/hlthaff.28.5.w822.

"Food Access Research Atlas." *USDA ERS - Food Access Research Atlas*, www.ers.usda.gov/data-products/food-access-research-atlas/.

Ghosh-Dastidar, Bonnie, et al. "Distance to Store, Food Prices, and Obesity in Urban Food Deserts." *American Journal of Preventive Medicine*, vol. 47, no. 5, 2014, pp. 587–595., doi:10.1016/j.amepre.2014.07.005.

Hales, Craig M., et al. "Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007-2008 to 2015-2016." *Jama*, vol. 319, no. 16, 2018, p. 1723., doi:10.1001/jama.2018.3060.

"Low-Income Communities More Likely to Face Childhood Obesity." *Low-Income Communities*

More Likely to Face Childhood Obesity | Michigan Medicine, 7 Jan. 2016, www.uofmhealth.org/news/archive/201601/low-income-communities-more-likely-face-childhood-obesity.

"Obesity in Adults - Colorado BRFSS 2014-2017 (County)." *Hub.arcgis.com*, hub.arcgis.com/datasets/CDPHE::obesity-in-adults-colorado-brfss-2014-2017-county/data.

"USDA Defines Food Deserts." *American Nutrition Association*, americannutritionassociation.org/newsletter/usda-defines-food-deserts.

Appendix

```
#####################################################
# Import data
#####################################################

setwd('/Users/Nick/Desktop/CU Classes Fall 2019/Applied
Regression/Project/Data')

food_df = read.csv('FoodAvailabilityData.csv')
overweight_df =
read.csv('Overweight_and_Obese_Adults__CDPHE_Community_Level_Estimates
_Census_Tracts.csv')
obesity_df =
read.csv('Obesity_in_Adults__CDPHE_Community_Level_Estimates_Census_Tr
acts.csv')

setwd('/Users/Nick/Downloads')
density_df = read.csv('Population_Density_Census_Tracts.csv')
density_df = density_df[,c('FIPS',
'Population_Density_PerLandSquareMile')]

food_df = food_df[food_df$State == 'Colorado',]

df = merge(food_df, obesity_df, by.x = 'CensusTract', by.y =
'Census_Tract_FIPS')

df = merge(df, density_df, by.x = 'CensusTract', by.y = 'FIPS')

df = df[,c('County', 'CensusTract', 'Urban', 'POP2010', 'PovertyRate',
'MedianFamilyIncome',
           'LA1and10', 'LAhalfand10', 'LA1and20', 'TractLOWI',
           'TractWhite', 'TractBlack', 'TractAsian', 'TractNHOPI',
           'TractAIAN', 'TractOMultir', 'TractHispanic', 'TractHUNV',
'TractSNAP',
           'OHU2010', 'Obese_Census_Tract_Estimate',
'Population_Density_PerLandSquareMile')]

#####################################################
# Feature generation and EDA
#####################################################

# Are there any unusual values?

summary(df)

# POP 2010:
# min = 0, which could be a problem, unless there are really
# tracts with no population

# MedianFamilyIncome:
# min = 0, which could be a problem, unless again there are
# tracts with no population and therefore no income

# Obese_Census_Tract_Estimate
# min is -1, so we should omit that row, since it's the predictor
variable
```

```
# For TractWhite, TractBlack, TractAsian, TractNHOPI,
# TractAIAN, TractOMultir, TractHispanic, TractHUNV, TractSNAP,
# OHU2010 all have min values of 0, so we should just make sure we're
# not missing data for those things

df[df['POP2010'] == 0, c('County', 'CensusTract')] # (NOTE: THIS ISN'T
A PROBLEM AFTER ADDING DENSITY DATAFRAME)

# This gives the census tracts with 0 population.
# Could replace with AdultPopulation_Age18_and_over,
# But that's also 0, so it's likely that there's just 0 population

# Are these the only rows with missing data?
# Let's drop those records and see
zero_pop_tracts = df[df['POP2010'] == 0, c('County',
'CensusTract')][['CensusTract']]

df = df[!(df$CensusTract %in% zero_pop_tracts), ]

summary(df)

# That got rid of a lot of problems.  We still have records
# where the MedianFamilyIncome is 0.
df[df$MedianFamilyIncome == 0,]

# There can't be a median family income of 0, so replace the missing
median family incomes with the median
# family income for the county
zero_medIncome_tracts = df[df$MedianFamilyIncome ==
0,][['CensusTract']]

for(tract in zero_medIncome_tracts){
  tract_county = df[df$CensusTract == tract, 'County']
  med_county_income = median(df[df$County == tract_county,
'MedianFamilyIncome'])
  df[(df$MedianFamilyIncome == 0 & df$CensusTract == tract),
'MedianFamilyIncome'] = med_county_income
}

summary(df)

# That seems to have fixed the problem for median family income

# We should look into the Poverty rate of 0
df[df$PovertyRate == 0,]

# The national poverty rate is $25,750 for a household of 4.
# Since all these tracts have a high median family income (>= 48,000),
# I think it's okay to leave it for now

summary(df)
# There are some tracts that also have 0 low income population
# These should at least be the same as the tracts with a 0 poverty
rate,
# and should have a reasonable median family income.

df[df$TractLOWI == 0,][,c('MedianFamilyIncome', 'CensusTract')]
```

```
df[(df$PovertyRate == 0  & df$TractLOWI == 0),]['CensusTract']

# They are the same, and the median family income id decent,
# so I think we can just leave it

summary(df)

# As far as races go, my concern would be that race just isn't
recorded for
# some tracts, so let's see if there are any rows where the counts for
ALL races is 0
nrow(df[(df$TractWhite == 0 & df$TractBlack == 0 &
       df$TractAsian == 0 & df$TractNHOPI == 0 &
       df$TractAIAN == 0 & df$TractOMultir == 0 &
       df$TractHispanic == 0),])

# There are no records, so I'll just really assume that if these are 0
# it's because there are 0 people of that race

summary(df)

# There are rows where TractHUNV is 0.  Are there really tracts
# where all tracts have a vehicle?
df[df$TractHUNV == 0,'PovertyRate']

# The poverty rate is high in some of these tracts, but it's plausible
that everyone has a vehicle
# (or at least there's no way to know for sure that it's an error), so
we'll leave it

summary(df)

# There are tracts where OHU2010 is 0.  That doesn't really make sense
df[df$OHU2010 == 0,]

# Let's find the mean number of households per population in the
dataframe
temp_df = df[df$OHU2010 != 0,]
mean(temp_df$OHU2010/temp_df$POP2010)
# output is  0.4004779

# We'll replace the 0 values with 0.4004779* pop2010
df[df$OHU2010 == 0,'OHU2010'] = 0.4004779*df[df$OHU2010 ==
0,'POP2010']

summary(df)

# The last thing to deal with is the Obese_Census_Tract_Estimate of -1
df[df$Obese_Census_Tract_Estimate == -1,]

# That's only the case for two counties, so we'll just drop them
obese_missing_tracts = df[df$Obese_Census_Tract_Estimate == -1,
'CensusTract']

df = df[!(df$CensusTract %in% obese_missing_tracts),]

summary(df)
```

```
######## I think by this point, the data is clean
######## Feature generation, then we'll look at univariate stats

# Percent population low income
df['PercentLowIncome'] = df$TractLOWI/df$POP2010*100

# Percent white
df['PercentWhite'] = df$TractWhite/df$POP2010*100

# Percent black
df['PercentBlack'] = df$TractBlack/df$POP2010*100

# Percent asian
df['PercentAsian'] = df$TractAsian/df$POP2010*100

# Percent native hawaiin or pacific islander
df['PercentNHOPI'] = df$TractNHOPI/df$POP2010*100

# Percent native american / alaskan native
df['PercentAIAN'] = df$TractAIAN/df$POP2010*100

# Percent multicultural or other
df['PercentOMultir'] = df$TractOMultir/df$POP2010*100

# Percent hispanic
df['PercentHispanic'] = df$TractHispanic/df$POP2010*100

# Percent housing units without a vehicle
df['PercentHUNV'] = df$TractHUNV/df$OHU2010*100

# Percent housing units with SNAP benefits
df['PercentSNAP'] = df$TractSNAP/df$OHU2010*100

# Percent obese
df['PercentObese'] = df$Obese_Census_Tract_Estimate

############ Now let's take a look at some bivariate statistics

boxplot(PercentObese~Urban, data = df, cex.lab = 1.5, cex.axis =
1.5)#, main = 'Proportion of Population that is Obese vs. Urban')
#****
# Nothing too outstanding.  Rural areas look like they're a bit more
obese

plot(PercentObese ~ PovertyRate, data = df, cex.lab = 1.5, cex.axis =
1.5)#main = 'Proportion of Population that is Obese vs. Poverty Rate')
#****
# Hard to tell, maybe a sight positive association

plot(PercentObese ~ MedianFamilyIncome, data = df, cex.lab = 1.5,
cex.axis = 1.5)# xlab = "Median Family Income (In Thousands of
Dollars)", ylab = "Obesity Level", main = 'Obesity Level vs. Median
Family Income') #****
# Looks like there's a negative association

boxplot(PercentObese ~ LA1and10, data = df,  cex.lab = 1.5, cex.axis =
1.5)#xlab = 'Lack of Food Access', ylab = 'Obesity Level', main =
"Obesity Level vs. Lack of Access to Food")
```

```
boxplot(PercentObese ~ LA1and20, data = df,  cex.lab = 1.5, cex.axis =
1.5)#main = 'Proportion of Population that is Obese vs. Limited Food
Access') #****
boxplot(PercentObese ~ LAhalfand10, data = df, cex.lab = 1.5, cex.axis
= 1.5)#
# It does look like low access has a higher percent of obesity

plot(PercentObese ~ PercentLowIncome, data = df, main = 'Proportion of
Population that is Obese vs. Proportion of Low Income Population')
#****
# Maybe a positive correlation

plot(df$PercentObese ~df$PercentWhite)
# Doesn't look like an association

plot(df$PercentObese ~df$PercentBlack)
# Doesn't look like an association

plot(df$PercentObese ~df$PercentAsian)
# Doesn't look like an association

plot(df$PercentObese ~df$PercentNHOPI)

plot(df$PercentObese ~df$PercentAIAN)

plot(df$PercentObese ~df$PercentOMultir)
# Maybe positive, not much

plot(PercentObese ~ PercentHispanic, data = df, cex.lab = 1.5,
cex.axis = 1.5)# main = 'Proportion of Population that is Obese vs.
\nProportion of Hispanic Population') #****
# Does look like a positive association

plot(df$PercentObese ~df$PercentHUNV)
# Maybe slightly positive

plot(PercentObese ~ PercentSNAP, data = df,  cex.lab = 1.5, cex.axis =
1.5)#xlab = 'Population SNAP benefits', ylab = 'Obesity Level', main =
'Obesity Levels vs. Proportion \nof Population with SNAP benefits')
#****
# Looks positive

##################################################
##Collinearity
##################################################

# Start creating a model with all variables
# But note that we don't need percent of all races
# since those are linearly dependent, so we'll take out
# PercentAIAN
lmod = lm(PercentObese ~ PercentSNAP +
          PercentHUNV +
          PercentHispanic +
          PercentOMultir +
          #PercentNHOPI +
          PercentAsian +
          PercentBlack +
          MedianFamilyIncome +
```

```
          PercentWhite +
          PercentLowIncome +
          LA1and20 +
          LAhalfand10 +
          LA1and10 +
          PovertyRate +
          Urban, data = df)

# Start by looking at variance inflation factors
library(car)
vif(lmod)

# The variables with VIF's greater than 5 are
# PercentHispanic
# PercentOMultir
# PercentBlack
# PercentWhite
# LA1and20
# LA1and10

# This makes sense because these things are related
# So we'll eliminate some of the races and some of the food access
flags

# For the food desert indicators, which one has the strongest
# association with PercentObese? We'll keep that one and throw
# out the others
lmod1 = lm(PercentObese ~ LA1and20, data = df)
summary(lmod1)
# R-squared is 0.0146

lmod1 = lm(PercentObese ~ LAhalfand10, data = df)
summary(lmod1)
# R-squared is 0.009568

lmod1 = lm(PercentObese ~ LA1and10, data = df)
summary(lmod1)
# R-squared is 0.02506

# So take out everything except hispanic, black, and white,
# and all the food desert indicators except LA1and20
lmod = lm(PercentObese ~ PercentSNAP + PercentHUNV +
            PercentHispanic +
            PercentBlack +
            PercentWhite + PercentLowIncome +
            MedianFamilyIncome +
            LA1and10 +
            PovertyRate + Urban, data = df)

vif(lmod)

# There's a couple that are higher than 5, so we should consider
taking out PercetWhite
# or PercentHispanic.  Since we think race has an impact let's take
PercentWhite out
lmod = lm(PercentObese ~ PercentSNAP + PercentHUNV +
            PercentHispanic +
            PercentBlack +
```

```
            PercentLowIncome +
            MedianFamilyIncome +
            LA1and10 +
            PovertyRate + Urban, data = df)

vif(lmod)

# PercentLowIncome is still high, so take it out
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            PercentBlack +
            #PercentLowIncome +
            MedianFamilyIncome +
            LA1and10 +
            PovertyRate +
            Urban, data = df)

vif(lmod)
# That looks good per VIF's

cor(lmod$model)

library(perturb)
colldiag(lmod)
# The colldiag method (condition numbers) also shows there is no more
collinearity

# Pairwise correlations
cor(lmod$model)
# There is high pairwise correlation between PercentLowIncome,
PovertyRate, PercentSNAP, and MedianFamilyIncome
# which isn't surprising, since they're all related to income.  And
even though the
# VIF's ane condition numbers are good, let's remove a couple of them

lmod = lm(df$PercentObese~df$PercentLowIncome)
summary(lmod) # R^2 = 0.0481
lmod = lm(df$PercentObese~df$PovertyRate)
summary(lmod) # R^2 = 0.0153
lmod = lm(df$PercentObese~df$PercentSNAP)
summary(lmod) # R^2 = 0.113
lmod = lm(df$PercentObese~df$MedianFamilyIncome)
summary(lmod) # R^2 = 0.0981

# Let's take out PercentLowIncome and PovertyRate, but keep
MedianFamilyIncome
lmod = lm(PercentObese ~ PercentSNAP + PercentHUNV +
            PercentHispanic +
            PercentBlack +
            #PercentLowIncome +
            MedianFamilyIncome +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)

vif(lmod)
colldiag(lmod)
```

```
cor(lmod$model)

#####################################################
# Model selection
#####################################################

df$MedianFamilyIncome = df$MedianFamilyIncome/1000
# Scale median income

############ 1. Backward elmination with a = 0.05
summary(lmod)

# Take out PercentBlack
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            PercentLowIncome +
            LA1and10 +
            #PovertyRate +
            MedianFamilyIncome +
            Urban, data = df)

summary(lmod)
# That's it, everything else is significant at the 0.05 level,
# but MedianFamilyIncome variable is very small

############ 2. AIC
b = regsubsets(PercentObese ~ PercentSNAP +
                 PercentHUNV +
                  PercentHispanic +
                  #PercentLowIncome +
                  PercentBlack +
                  LA1and10 +
                  MedianFamilyIncome +
                  #PovertyRate +
                 Urban, data = df)

rs = summary(b) # summarize model that minimizes RSS for each p

# calculate AIC of each model
n = nobs(lmod)
aic = n*log(rs$rss/n) + 2*(2:8)
aic2 = rs$bic + (2 - log(n)) * 2:8
# plot AIC vs p
plot(2:8, aic, xlab = "p", ylab = "AIC")

# It's hard to tell from the plot, so just look at the values
aic

# AIC recommends to keep the model with all variables

############ 3. BIC

library(car)
```

```
subsets(b, min.size = 4, statistic = "bic", legend = FALSE)

# BIC suggests the model without PercentBlack or
MedianFamilyIncome

############ 4. Mallow's CP
subsets(b, min.size = 4, statistic = "cp", legend = FALSE)
abline(1, 1)

# The one that has the lowest Cp is the model with all variables

############ 5. Adjusted R^2
subsets(b, statistic = "adjr2", min.size = 4, legend = FALSE)

# Says to use the model with all variables

############ 6. Stepwise selection
lmod = lm(PercentObese ~ PercentSNAP + PercentHUNV +
            PercentHispanic +
            PercentBlack +
            MedianFamilyIncome +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)

step(lmod) #with AIC

# Keeps everything

step(lmod, k = log(n)) # With BIC

# Takes out PercentBlack and MedianFamilyIncome

# So there are 3 models:
# The one with all the variables
# The one without PercentBlack
# The one without PercentBlack or MedianFamilyIncome

# Full model
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            PercentBlack +
            MedianFamilyIncome +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)
summary(lmod)

# Without PercentBlack
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            #PercentBlack +
```

```
            MedianFamilyIncome +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)
summary(lmod)

# Without PercentBlack or MedianFamilyIncome
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            #PercentBlack +
            #MedianFamilyIncome +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)
summary(lmod)

###### Cross validation on these models

install.packages('caret')
library(caret)

# All variables
ols_cv = train(PercentObese ~ PercentSNAP +
                 PercentHUNV +
                 PercentHispanic +
                 MedianFamilyIncome +
                 PercentBlack +
                 LA1and10 +
                 #PovertyRate +
                 Urban, data = df,
                  method = 'lm',
               trControl = trainControl(
                 method = "cv",
                 number = 10,
                 savePredictions = TRUE,
                 verboseIter = TRUE
               ))
ols_cv$results

# For the full model, the RMSE is 4.8274,
# MAE is 3.5825

# Without PercentBlack
ols_cv2 = train(PercentObese ~ PercentSNAP +
                  PercentHUNV +
                 PercentHispanic +
                 MedianFamilyIncome +
                 LA1and10 +
                 #PovertyRate +
                  Urban, data = df,
               method = 'lm',
               trControl = trainControl(
                 method = "cv",
```

```
                number = 10,
                savePredictions = TRUE,
                verboseIter = TRUE
              ))
ols_cv2$results

# For the model without PercentBlack,
# the RMSE is 4.8326,
# MAE is 3.5821

# Without PercentBlack or MedianFamilyIncome
ols_cv3 = train(PercentObese ~ PercentSNAP +
                  PercentHUNV +
                  PercentHispanic +
                  LA1and10 +
                  #PovertyRate +
                  Urban, data = df,
                method = 'lm',
                trControl = trainControl(
                  method = "cv",
                  number = 10,
                  savePredictions = TRUE,
                  verboseIter = TRUE
                ))
ols_cv3$results

# For the model without PercentBlack or
# PercentLowIncome, the RMSE is 4.8331,
# and the MAE is 3.5922

# They're all about the same, so in an effort to keep the model
parsimonious,
# we'll go without PercentBlack and MedianFamilyIncome (since
MedianFamilyIncome and PercentSNAP and PovertyRate
# had high correlation anyway)

#################################################
# Model structure
#################################################

# Our model so far...
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            PercentHispanic +
            LA1and10 +
            #PovertyRate +
            Urban, data = df)

# Component plus residual plots can suggest transformations
crPlots(lmod)

# That looks like we should add a log transformation for
PercentHispanic
lmod = lm(PercentObese ~ PercentSNAP + PercentHUNV +
```

```
            #PercentHispanic +
            LA1and10 +
            #PovertyRate +
            Urban +
            log(PercentHispanic), data = df)

summary(lmod)

crPlots(lmod)

# PercentHispanic looks much better after that transformation

# It's hard to tell if we should apply transformations to
PercentSNAP
#, since there are some influential observations that are
skewing the line.

# So let's try a lack of fit test
lmod = lm(PercentObese ~ PercentSNAP + I(PercentSNAP^2) +
            PercentHUNV +
            #PercentHispanic +
            LA1and10 +
            #PovertyRate +
            Urban +
            log(PercentHispanic), data = df)

summary(lmod)

# Adding PercentSNAP^2 isn't significant at 0.05 level, so leave
it out
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            #PercentHispanic +
            LA1and10 +
            #PovertyRate +
            Urban +
            log(PercentHispanic), data = df)

summary(lmod)

crPlots(lmod)

# Everything looks good on the crPlots, and everything is
significant

# The residual plots shouldn't have any systematic curves or
patterns
residualPlots(lmod)

# No curves or patterns (except for posible), so I think we're
good to go

# Tukey's test tells whether or not the response should be
squared
```

```
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

yhatall = predict(lmod, newdata = df)

lmod.tukeytest = lm(PercentObese ~ PercentSNAP +
                      PercentHUNV +
                      LA1and10 +
                      Urban +
                      log(PercentHispanic) +
                      I(yhatall^2), data = df)

summary(lmod.tukeytest)

# That test fails
# So this will be our model so far
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

summary(lmod)
# Note that the R^2 is 0.26 so the model explains about 1/3
# of the observed variation

#####################################################
# Check again for collinearity and model selection with
transformations
#####################################################

lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

######### Collinearity
vif(lmod)

colldiag(lmod)

cor(lmod$model)
# There is a bit of collinearity between PercentSNAP and
log(PercentHispanic),
# but everything else is low or certainly within reason. It
won't be possible
# to get no collinearity between these variables

######### Model selection
```

```
# Everything is significant at the 0.05 level, so no need to do
backward selection

# Try with AIC
b = regsubsets(PercentObese ~ PercentSNAP +
                 PercentHUNV +
                 LA1and10 +
                 Urban +
                 log(PercentHispanic), data = df)

rs = summary(b) # summarize model that minimizes RSS for each p

# calculate AIC of each model
n = nobs(lmod)
aic = n*log(rs$rss/n) + 2*(2:6)

# AIC says to keep the model with all the variables

# Try Mallow's CP
subsets(b, statistic = "cp", legend = FALSE)
abline(1, 1)

# Mallow's CP says that the modwl with all variables is good!

# Try Adjusted R^2
subsets(b, statistic = "adjr2", legend = FALSE)

# Says to keep the model with all variables.

# In conclusion, I think our model is good.
# We've checked model structure and applied transformations,
# addressed collinearity by amputating variables, and used
# model selection techniques to evaluate.

##################################################
# Influential observations
##################################################

# Leverage points
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

h = hatvalues(lmod)

row_names = df[,'CensusTract']
halfnorm(h, nlab = 2, labs = row_names, cex.lab = 1.5, cex.axis
= 1.5, ylab = "leverage")

# It looks like maybe there's two leverage points
sort(h, decreasing = TRUE)[1:2]
```

```
df[(df$CensusTract %in% c(8041004402, 8031000800)),]

# They're the points with indices 373 and 647, so let's take a
look
# at them and see if anything stands out (could be an entry
error, etc.)
df[(rownames(df) %in% c(373, 647)),]

# Nothing stands out or looks like a data entry error.

# Test for outliers
outlierTest(lmod)

# No outliers at the Bonferroni 0.05 level

# Make an inflence plot
influencePlot(lmod)
# This suggests that the rows with index 647 and 373 MIGHT be
influential, which we
# already found.

# Test for influential observations using Cook's distance
cook <- cooks.distance(lmod)
halfnorm(cook, n = 3, labs = row_names, cex.lab = 1.5, cex.axis
= 1.5, ylab = "Cook's distances")
sort(cook, decreasing = TRUE)[1:3]

# It looks like the possible influential points are the ones
with indexes 373, 647, and 350
# We already looked at 647 and 373, so let's look them all again
df[(rownames(df) %in% c(350, 373, 647)),]

# One thing that stands out to me about 350 is that nearly half
the population is obese.
# Maybe part of the policy is to see what's going on in that
tract specifically.

# Let's see what influence these points have by taking them out
1-by-1 and see what effect they have on
# the model

# Remove point 1084
lmod2 = lm(PercentObese ~ PercentSNAP +
             PercentHUNV +
             LA1and10 +
             Urban +
             log(PercentHispanic), data = df, subset =
(rownames(df) != 350))

compareCoefs(lmod, lmod2)

# Doesn't change the inference at all

# Remove point 377
```

```
lmod2 = lm(PercentObese ~ PercentSNAP +
             PercentHUNV +
             LA1and10 +
             Urban +
             log(PercentHispanic), data = df, subset =
(rownames(df) != 373))

compareCoefs(lmod, lmod2)

# Doesn't change inference at all

# Remove point 652
lmod2 = lm(PercentObese ~ PercentSNAP +
             PercentHUNV +
             LA1and10 +
             Urban +
             log(PercentHispanic), data = df, subset =
(rownames(df) != 647))

compareCoefs(lmod, lmod2)

# Again, doesn't change inference.

# In conclusion, I don't see any need to take any of these
points out of the model

#####################################################
# Checking error assumptions
#####################################################

# To check that the mean of the errors is 0, we look at
residuals vs. fitted values
# and residuals vs. values of the predictors to see that they're
centered around 0

lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

# plot of residuals versus fitted values
plot(lmod, which = 3)
# Looks good

# plot of residuals versus predictors
residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# We definitely see horns in the plots of PercentSNAP and
PercentHUNV,
# which means that our variance is not constant. We can try to
remedy that by applying
# transformations to the response or adding weights.
```

```
# Try weighing by the population
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), weights = df$POP2010, data =
df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# That didn't seem to fix it.  What about weighting by housing
units?
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), weights = df$OHU2010, data =
df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# That's better, but still not great.

# Try log transformation on the response
lmod = lm(log(PercentObese) ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# Not good.  What about a square root transformation of the
response?
lmod = lm(sqrt(PercentObese) ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# Better than log, but still not good.

# Try using the residuals from a model without weights as
weights (as recommended by online research)
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
```

```
            Urban +
            log(PercentHispanic), data = df)

lmod2 = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), weights =
abs(1/lmod$residuals), data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# That's about the same, still not great.

# After thinking about it a bit, it makes sense that we might
have less variability in more
# densely populated areas, so try weighting by density
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            Urban +
            log(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# That seems to fix PercentSNAP and PercentHUNV, but now Urban
doesn't have constant variance.
# So let's take Urban out.  The information we gain for that
variable is probably mostly
# accounted for by weighing for density

lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

# There's almost a wedge shape on log(PercentHispanic) now.  Try
a sqrt transormation instead
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            sqrt(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)
```

```
# I'd say those look pretty good

# So let's leave the model where it is, now that our error
assumptions are satisfied and re-check
# collinearity, model structure, and model selection methods

lmod = lm(PercentObese^2 ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            sqrt(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod, quadratic = FALSE, fitted = FALSE, tests =
FALSE)

######## Collinearity:
vif(lmod)
colldiag(lmod)
cor(lmod$model)

# Collinearity looks good

crPlots(lmod)

# Looks good aside from possible influential points on
PercentSNAP

residualPlots(lmod)

# Maybe curving on the fitted values? Try Tukey's test again
lmod = lm(PercentObese ~ PercentSNAP +
            Urban +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic),
          weights = df$Population_Density_PerLandSquareMile,
          data = df)

yhatall = predict(lmod, newdata = df)

lmod.tukeytest = lm(PercentObese ~ PercentSNAP +

                      PercentHUNV +
                      LA1and10 +
                      log(PercentHispanic) +
                      I(yhatall^2),
                    # weights =
df$Population_Density_PerLandSquareMile,
                     data = df)

summary(lmod.tukeytest)

# That is significant now.  So let's try undoing some of the
things we did
```

```
# Start by keeping log(PercentHispanic)
lmod = lm(PercentObese ~ PercentSNAP +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod)

# What about adding in the Urban variable again
lmod = lm(PercentObese ~ PercentSNAP +
            Urban +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), weights =
df$Population_Density_PerLandSquareMile, data = df)

residualPlots(lmod)

# No, so the weights are causing other problems, so remove the
weights
lmod = lm(PercentObese ~ PercentSNAP +
            Urban +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), data = df)

residualPlots(lmod)

# I think it's better to move ahead with this and note that some
of our error assumptions may
# be violated.  If I had more time, I would figure out how to
incorporate weights without
# adding unaccounted for curvature to the model

############ NOTE: IN paper, need to mention that we SHOULD
check for autocorrelation,
# but we're not going to do that (because we didn't learn it)

# Make sure to mention that WE ASSUME that the errors are
uncorrelated

#################################################
# Interpretation
#################################################
lmod = lm(PercentObese ~ PercentSNAP +
            Urban +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), data = df)


summary(lmod)
```

```
# The intercept is 15.8011
# PercentSNAP is .1209
# Urban is -2.3763
# PercentHUNV is -0.1563
# LA1and20 is 1.7211
# log(PercentHispanic) is 2.7430

# PercentSNAP
# When the percent of the population with SNAP benefits goes up,
we expect that the Percent of the population
# that is obese will go up by 0.1209

# Urban
# Locations that are urban have an expected percent of obesity
that's -2.3763 less than rural locations,
# while holding all other variables constant

# PercentHUNV
# When the percent of housing units without vehicles goes up by
1, we expect the percent of obesity
# to go down -0.1563, while holding other variables constant

# LA1and20
# For housing units with low access, we expect the percent of
obesity to be 1.7211 percent higher,
# while holding other things constant

# log(PercentHispanic)
# Increasing by 1%, while holding others constant, is associated
with
# APPROXIMATELY a 2.743/100 increase in the response variable,
on average

# Note that the R^2 value seems appropriate since a plot of y vs
yhat is approximately linear
plot(df$PercentObese~lmod$fitted)

####################
# Things to add
####################

# Effects plots (95% confidence bands of the effects of
regressor on response)
# This could be useful for a poster
library(effects) # for Effect function
plot(Effect("PercentHUNV", lmod))
plot(Effect("log(PercentHispanic)", lmod))

# Create a scatterplot of Urban vs. the other variables
# where urban or not are different colors to see if value,
slopes are the same
# If slopes change, consider including interactions.
# If not, use it as a justification why NOT to include
interactions
```

```
# You did this in a homework at some point, so just find the
code for that
sumary(lmod)

pch_type = c(19, 1)[factor(df$Urban)]
col_type = c(153, 4)[factor(df$Urban)]

pchs = c(19, 1)
coltypes = c(153, 4)

plot(PercentObese~PercentSNAP, data = df, col = col_type, pch =
pch_type, cex.lab = 1.5, cex.axis = 1.5)# main = "Obesity Level
vs. PercentSNAP")
legend(70, 45, legend = c('Urban', 'non-Urban'), pch = pchs, col
= coltypes)
# Maybe? Worth trying...
plot(PercentObese~PercentHUNV, data = df, col = col_type, pch =
pch_type, cex.lab = 1.5, cex.axis = 1.5)#, main = "Obesity Level
vs. PercentHUNV")
legend(30, 45, legend = c('Urban', 'non-Urban'), pch = pchs, col
= coltypes)
# Nothing there
plot(PercentObese~log(PercentHispanic), data = df, col =
col_type, pch = pch_type, cex.lab = 1.5, cex.axis = 1.5)#, main
= "Obesity Level vs. log(PercentHispanic")
legend(0.5, 45, legend = c('Urban', 'non-Urban'), pch = pchs,
col = coltypes)
# Those look the same, so good to go

# What about LA1and10?

pch_type = c(19, 1)[factor(df$LA1and10)]
col_type = c(153, 2)[factor(df$LA1and10)]

pchs = c(19, 1)
coltypes = c(153, 2)

plot(PercentObese~PercentSNAP, data = df, col = col_type, pch =
pch_type, cex.lab = 1.5, cex.axis = 1.5)#main = "Obesity Level
vs. PercentSNAP")
legend(70, 45, legend = c('Urban', 'non-Urban'), pch = pchs, col
= coltypes)
# No
plot(PercentObese~PercentHUNV, data = df, col = col_type, pch =
pch_type, cex.lab = 1.5, cex.axis = 1.5)#main = "Obesity Level
vs. PercentHUNV")
legend(30, 45, legend = c('Urban', 'non-Urban'), pch = pchs, col
= coltypes)
# No
plot(PercentObese~log(PercentHispanic), data = df, col =
col_type, pch = pch_type, cex.lab = 1.5, cex.axis = 1.5)#main =
"Obesity Level vs. log(PercentHispanic)")
legend(0.5, 45, legend = c('Urban', 'non-Urban'), pch = pchs,
col = coltypes)
```

```
# Those look the same, so good to go

# So let's try an interaction term for Urban and PercentSNAP
lmod = lm(PercentObese ~ PercentSNAP +
            Urban +
            PercentSNAP*Urban +
            PercentHUNV +
            LA1and10 +
            log(PercentHispanic), data = df)


summary(lmod)
# Makes the P-value for Urban really high

# What about residual plots?
residualPlots(lmod)

# The residual plots also look worse, and Tukey's test is
significant, so leave it out.

# Note that we chose not to scale the regressors to make it easy
to
# interpret the coefficients
```