Food Insecurity in Texas Counties

Jaehwan Han

INTRODUCTION

Food insecurity is the limited or uncertain availability of enough food to support an active, healthy life. According to the USDA, 11.8 percent of U.S. households were food insecure at some time during 2017.[1] Food insecurity is a national problem and an underrecognized social determinant of health. In the literature, food insecurity has been linked to poor health outcomes, increased health care costs, and impaired cognitive and psychological functioning.[2-6] Understanding the underlying factors of food insecurity is therefore important towards addressing the issue.

Due to lack of easy access to proper grocery stores and community services, individuals living in rural communities have reported higher rates of food insecurity compared to individuals living in urban communities.[3] Like other economic and social indicators, food insecurity disproportionally affects different demographic groups; food insecurity levels are higher for households with children headed by a single woman (30.3 percent), by Black non-Hispanics, (21.8 percent) and by Hispanics (18 percent).[1] The prevalence of food insecurity is also affected by household-level characteristics such as income, employment, average wages, cost of food and housing, household structure, and access to/participating in food assistance programs.

This project uses data from Feeding America's Map the Meal Gap study, the County Health Rankings, and the US Census Bureau to assess potential factors influencing food insecurity across Texas counties.[9] The average of food insecurity rate in Texas is 14 percent, higher than the national average. A more detailed examination shows that food insecurity levels vary between 5 - 26 percent among Texas counties (see **Figure 1** in Appendix). Although a previous study found ethnicity to be an important predictor for food insecurity, our study focuses on predictors influenced by social economic policies.[1] The initial set of variables include rural/urban status, median household income, poverty level (a measure of poverty among those who are poor), county population, cost per meal, percent of county population who are subject to diabetes monitoring, and percent of county population who are high school graduates. We performed multiple regression analysis to identify important predictors of food insecurity and potential interactions among selected predictors.

METHODOLOGY

Data

Our sample comprised of all 254 Texas counties. Our outcome variable, county-level estimate of food insecurity rate, comes from the annual Map the Meal Gap study. Our continuous variables, county population, median household income, cost per meal, diabetes prevalence, and high school graduation rates, come from all three sources. Our categorical variables, poverty-level and urban/rural status, come from Map the Meal Gap and the US Census Bureau.

Data Cleaning and Software

Since poverty category was a three-level dummy variable, this was dummy-coded by setting the “poverty category 1” as the reference category. This variable is a proxy for need beyond the scope of poverty level within the estimated food insecure population.This reference category represents the individuals who are less than 165% below the poverty level whereas level 3 consisted of individuals who are more than 185% below poverty level and level 2 consisted of individuals in between the two percentages. Since we were going to include interactions in the model, we immediately centered and scaled the variables so they would have a mean of 0 and a standard deviation of 1. No observations were removed.

All analyses were run using SAS 9.4, and some graphs were created using R and Tableau. Significance was based on a p-value less than 0.05.

Initial Variable Investigation

We first created a scatterplot matrix of all the variables to understand the relationships between the variables and their distributions; we also created a correlation ellipse plot in R to quickly assess multicollinearity (or relationships) among predictor variables.

A histogram and QQ-plot of food insecurity was also checked to confirm the assumption of a normally distributed outcome variable was met.

Model Selection

There were four model selection methods employed in this project: stepwise, backward, best subset (using Akaike's criterion, AIC), and F-tests for polynomial terms. For polynomial terms, such as the squared of some variables expected to have a curvilinear rather than linear relationship with food insecurity, we used the hierarchical approach to fitting. This means we tested whether it was necessary to use a higher order model or if a first-order model would suffice.

Diagnostics

Diagnostics were performed on all models fit. This included: (1) analyzing linearity based on scatterplot matrices, (2) normality based on the histograms, QQ-plots, and Shapiro-Wilk test of normality of the residuals, (3) homoscedasticity of the residuals based on the residuals v. fitted plot and the Brausch-Pagan test, and (4) independence based on the study design and residual vs. fitted plot.

Model Validation

Since this model focused on inference rather than prediction, all validation came from literature review. We also used LASSO to determine whether our selection methods selected the best variables.

RESULTS

Descriptive Statistics

Descriptive statistics for all continuous variables used in the analysis is presented in **Table 1** below. The high variance of predictor ranges underscores the need to scale the variables when using techniques such as LASSO or when including interaction terms. It is important to note that our class variable poverty category is not evenly distributed. Almost all counties have a majority of those in poverty in category 3, and only one county belongs to category 2. This may indicate that this is not the best variable to use and, rather, a numeric variable that captured the percentage in category 3 would have been more useful.

Table 1: Descriptive summary of continuous variables

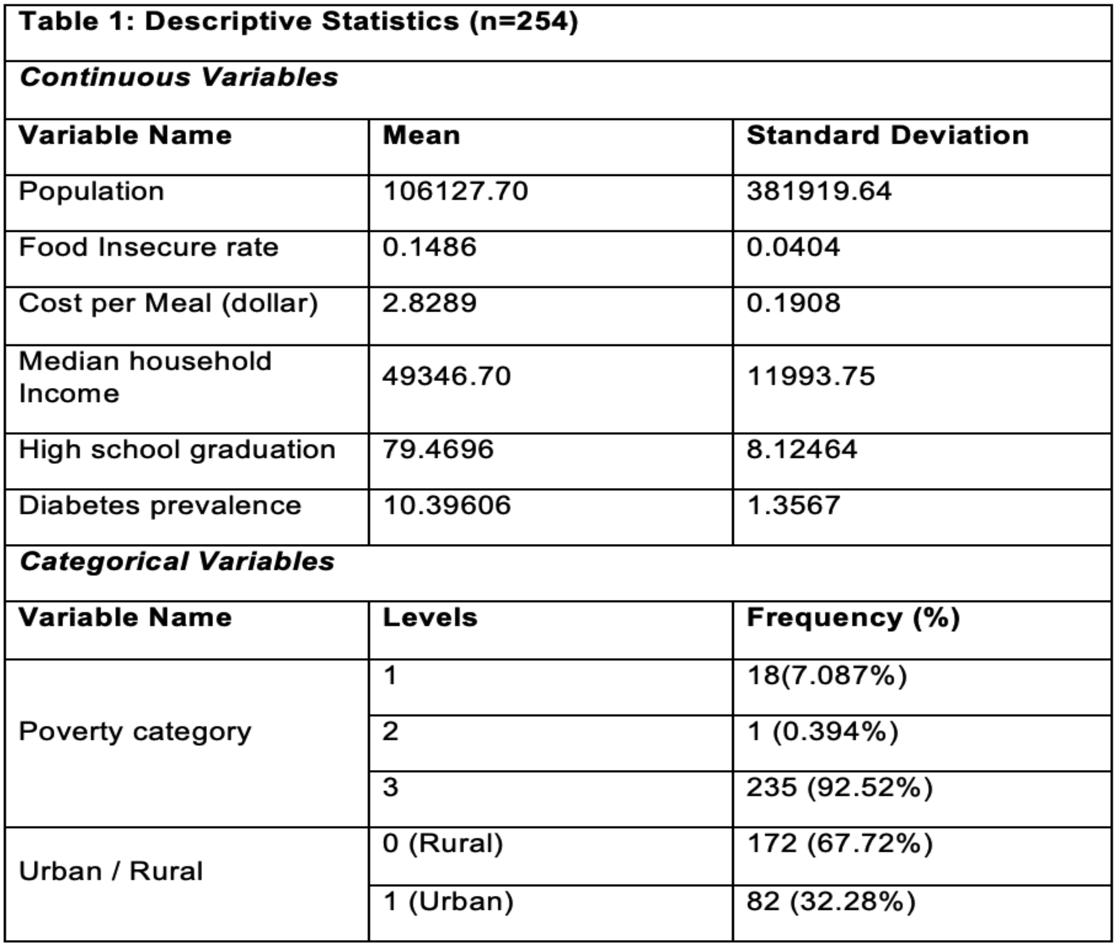


Table 1: Descriptive Statistics (n=254)		
Continuous Variables		
Variable Name	**Mean**	**Standard Deviation**
Population	106127.70	381919.64
Food Insecure rate	0.1486	0.0404
Cost per Meal (dollar)	2.8289	0.1908
Median household Income	49346.70	11993.75
High school graduation	79.4696	8.12464
Diabetes prevalence	10.39606	1.3567
Categorical Variables		
Variable Name	**Levels**	**Frequency (%)**
Poverty category	1	18(7.087%)
	2	1 (0.394%)
	3	235 (92.52%)
Urban / Rural	0 (Rural)	172 (67.72%)
	1 (Urban)	82 (32.28%)

Correlation and Scatterplot Matrix

Figure 2 shows that there are few strong, linear relationships between food insecurity and the predictor variables. Diabetes prevalence appears to have a medium, linear relationship with food insecurity while high school graduation rate and median house income indicate a possible polynomial relationship. Population appears to have many outliers and could benefit from a log transformation. The correlation plot shown below in **Figure 3** confirms that only median household income has a stronger relationship with food insecurity. However, it also shows that there is little multicollinearity present in the model as even the ones that appear more strongly correlated, such as median household income and poverty category, have a correlation r-value < 0.5, indicating weak correlation (**Figure 4).**

Preliminary Model

To further explore the data, a preliminary regression model was fit including all predictor variables. Results of this preliminary model are reported in **Figure 6.**

Due to the nonlinear relationship high school graduation rate and median household income appeared to have with food insecurity, we used the hierarchical approach to fitting. We used an F-test to determine whether to include the squared terms of these two variables in the model. The results of this F-test showed that a first-order model was sufficient ($F = 1.22$, $p > 0.05$). Therefore, these two squared terms were not included in the final model (**Figure 5**).

Results of the preliminary model showed that while the overall ANOVA F-test was significant ($p<0.0001$), the adjusted R-squared was fairly low (0.3861) and several variables did not appear to contribute very much to the model: population, the two dummy variables (poverty category 2, poverty category 3) and the two interaction variable (urban/rural * diabetes prevalence, median house income * diabetes prevalence). However in the initial model, cost per meal, median income, urban rural, diabetes prevalence and high school graduation were significant to the model ($p<0.05$).

Variance Inflation Factor (VIF) values from the initial model were all less than 3, confirming that multicollinearity was not a major concern for any of the included (untransformed) predictors. Additionally, QQ-plot, histogram, and Shapiro-Wilk test for normality of the residuals all indicated that variables were normally distributed (**Figure 7**).

While plots of residual against the predictors raised concerns about potential outliers and non-constant variance of the residuals with respect to some predictors (population, cost per meal, median household income, high school graduation), conducting a Breusch-Pagan test for constant variance of the residuals with respect to the predictors indicated that heteroscedasticity was not significant overall (Breusch-Pagan $p=0.0534$). Outputs of potentially problematic variance tests are reported in **Figure 7.**

Dimension Reduction (*Stepwise Selection Results)*

First, stepwise selection was used to identify the best model based on the inclusion criterion of α enter=0.05 and retention criterion of α stay=0.1. The resulting model for food insecurity rate (**Figure 8**) included 5 predictors: cost per meal, median household income, urban rural, diabetes prevalence, and high school graduation rate. Population, two dummy categorical variables (poverty category), and two interaction variables (urban rural * diabetes prevalence, median house income * diabetes prevalence) were dropped from the model.

Backward selection results

Backward selection shown in **Figure 9** yielded the same results as stepwise selection. The resulting model again included cost per meal, median household income, urban/rural status, diabetes prevalence, and high school graduation rate.

Best subset selection results

Best subset selection using AIC as selection criterion and with a best parameter value of 10 again yielded the same result as the selection methods discussed above (**Figure 10)**. Thus, our final only included the five variables cost per meal, median household income, urban/rural status, diabetes prevalence, and high school graduation. Due to previous literature indicating diabetes prevalence was related to median household income and urban/rural status, we chose to keep those interactions despite their insignificance based on the first three selection methods. [3-4]

LASSO

LASSO selection resulted in a different set of final predictors than the other three selection methods. The variables selected by LASSO were: urban/rural status, median household income, cost per meal, diabetes prevalence, high school graduation rate, and the interaction between median household income and diabetes prevalence. This is the only model that selected one of the interaction terms in the model (**Figure 11**).

Final Model

Due to the consistency of the first three selection methods as well as the confirmation that at least one of the interactions was deemed important by LASSO, our final model included the five variables previously discussed as well as the two interaction terms.

All diagnostics were rechecked. All residuals were normally distributed, but this time the Breusch-Pagan test was significant (16.12, $p < 0.05$). We therefore employed weighted least squares regression to fix the violation of this assumption. The assumption of independence was assumed to be met, and there was not any clustering in the residual v. fitted plot to indicate otherwise. Lastly, the linearity assumption was checked once more to confirm that the squared terms of median household income and high school graduation rate were not necessary to include in the model. Again, the F-test was not significant. After using weighted least squares regression, assumptions of normality were checked one last time (**Figure 12)**. This assumption was not violated. Outliers and leverage points were checked on this final model using DFFITS, DFBETAS, and Cook's distance (**Figure 13)**. There were no outliers according to Cook's distance, as all values were below the cutoff point of greater than 50%. As no DFFITS values were greater than 1, there were no points that had strong influence. Lastly, there appears to be some points that have high leverage, but these values are still well below 0.2, so they can be ignored.

The final model displayed in **Table 2** below shows median household income, cost per meal, prevalence of diabetes, and high school graduation rate were significant at the 0.05 level. The overall F-test was significant (F = 23.65, $p < 0.05$) as shown in **Figure 14**. Though the R-square value improved from the preliminary model, this criteria of model fit is no longer reliable because weighted regression was used.

Variable	**Parameter Estimate**	**Standard Error**	**t-value**	**Pr>|t|**
Urban/Rural status**	0.00983	0.00532	1.85	0.0659
Median Household Income	-0.0147	0.00312	-4.71	<0.0001
Cost Per Meal	0.00469	0.00209	2.25	0.0255
Prevalence of Diabetes	0.00974	0.00315	3.09	0.0022
High School Graduation Rate	0.01745	0.00249	7.00	<0.0001
Rural Status*Diabetes	-0.00408	0.0604	-0.67	0.5006
Median Household Income*Diabetes	0.00131	0.0249	0.53	0.5999

** Rural is the reference group

Validation

As our model was focused on inference rather than prediction, our validation was completed using literature review rather than comparing the accuracy of a model on a test set based on a training set. The literature review indicated that our model may have some problems. One of these is that our model gives the opposite of our expected relationship between high school graduation rate and food insecurity.[1] Typically food insecurity is linked to lower high school graduation rates, but our model indicates that a higher high school graduation rate leads to higher food insecurity. We attempted LASSO as another check of our validation methods. Although one extra variable was selected, the actual relationships between the predictors and food insecurity did not change (**Figure 11**).

DISCUSSION

The overall model predicting food insecurity was significant (F = 24.03, $p < 0.05$). Median household income, cost per meal, prevalence of diabetes, and high school graduation rate were significant at the 0.05 level (**Table 2 and Figure 14)** . Median household income and high school graduation rate have the most significant effects on food insecurity. The two interaction terms, however, are not significant despite literature supporting an interaction.

There were several surprising results regarding the coefficients of the model. Literature indicates that higher graduation rates lead to lower rates of food insecurity. However, our model shows that for every one standard deviation increase in high school graduation rate, food insecurity rate increased by 0.01745. Though literature found rural communities to have a higher prevalence of food insecurity, this model found that the mean of food insecurity for urban counties is 0.00983 standard deviations higher than for rural communities. Unlike previous literature, we did not find that diabetes, rural/urban status, or median income had interacted with one another to have a significant effect on food insecurity. However, the interaction between diabetes and rural/urban status was in line with expectation: for urban communities, for every one standard deviation increase in diabetes prevalence, food insecurity decreases by 0.00408 when controlling other factors. Median household income and cost per meal did fall in line with the current literature. For every one standard deviation increase in median household income, food insecurity decreased by 0.0147; for every one unit increase in cost per meal, food insecurity increased by 0.00469.

Due to multiple unexpected results and a violation of the homoscedasticity assumption, we believe that this is not the best model to understand the variance of food insecurity among Texas counties. This may be due to missing factors in the model or confounding variables. There are several factors not included in this model that have historically been included, such as race. However, this was not included because we only wanted to include variables that can be affected by actionable change via social economic policies. Racial factors cannot be changed. Variables that have been included in the past are unemployment rate and a measure of available resources in the area, so these may be worthwhile additions. **Figure 1** also indicates that a geographic analysis may be warranted, as a disproportionate number of the counties in East Texas are largely impacted by food insecurity. The poverty category may also be better represented as a numeric value as a percentage who fall in the “worst” category of 3 - 185% below the poverty line. This is because the large majority (92.54%) of counties have this group as the largest category. Lastly, including diabetes as a predictor of food insecurity may be misleading as diabetes may be more a result of food insecurity or the other socio-economic factors that lead to both outcomes. However, its addition still proves that there is a linear relationship between these two variables.

REFERENCES

1. Alisha Coleman-Jensen, Matthew P. Rabbitt, Christian A. Gregory, and Anita Singh. 2018. Household Food Security in the United States in 2017, ERR-256, U.S. Department of Agriculture, Economic Research Service

2. Berkowitz SA, Seligman HK, Meigs JB, Basu S. Food insecurity, healthcare utilization, and high cost: A longitudinal cohort study. Am J Manag Care. 2018;24(9):399-404. PMID: 30222918

3. Conway BN, Han X, Munro HM, Gross AL, Shu X-O, Hargreaves MK, et al. (2018) The obesity epidemic and rising diabetes incidence in a low-income racially diverse southern US cohort. PLoS ONE 13(1): e0190993. https://doi.org/10.1371/journal.pone.0190993

4. Heerman WJ, Wallston KA, Osborn CY, et al. Food insecurity is associated with diabetes self-care behaviours and glycaemic control. Diabet Med. 2016;33:844-850

5. O’Connor, A. Wellenius, G. Rural–urban disparities in the prevalence of diabetes and coronary heart disease. Public Health, 2012. Vol.126(10), p.813-820.

6. Olson CM. Nutrition and health outcomes associated with food insecurity and hunger. J Nutr Feb; 1999 129(2S Suppl):521S–524S. [PubMed: 10064322]

7. Seligman HK, Laraia BA, Kushel MB. Food insecurity is associated with chronic disease among low-income NHANES participants. J Nutr Feb;2010 140(2):304–10. Epub 2009 Dec 23. [PubMed: 20032485]

8. Shankar, A., Priya ; Chung, A., Rainjade ; Frank, A., Deborah. Journal of Developmental & Behavioral Pediatrics, 2017, Vol.38(2), p.135-150

9. University of Wisconsin Population Health Institute. County Health Rankings 2016.

APPENDIX

Figure 1: Estimates of Food Insecurity in Texas Counties, 2016

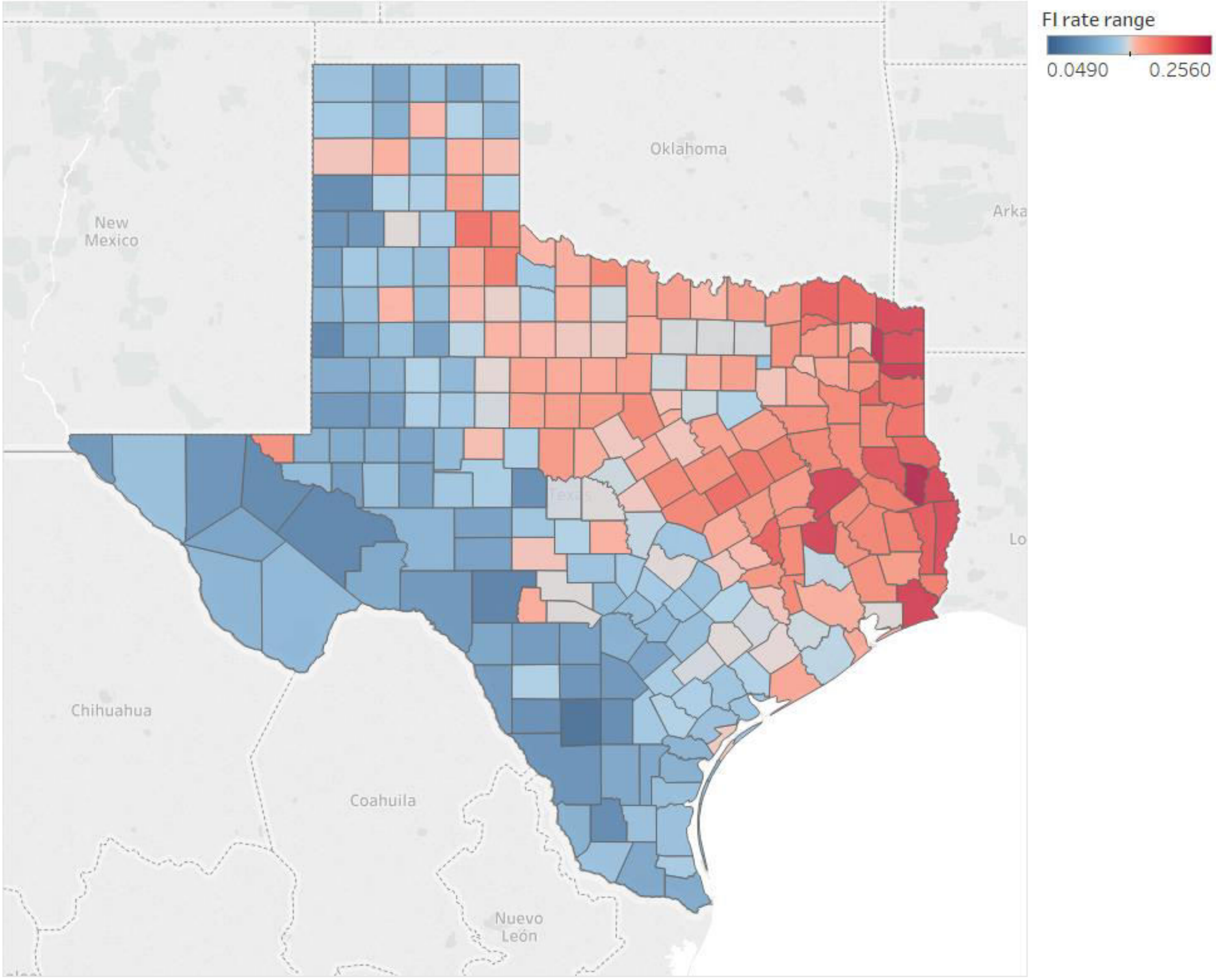


Figure 2: Scatterplot Matrix

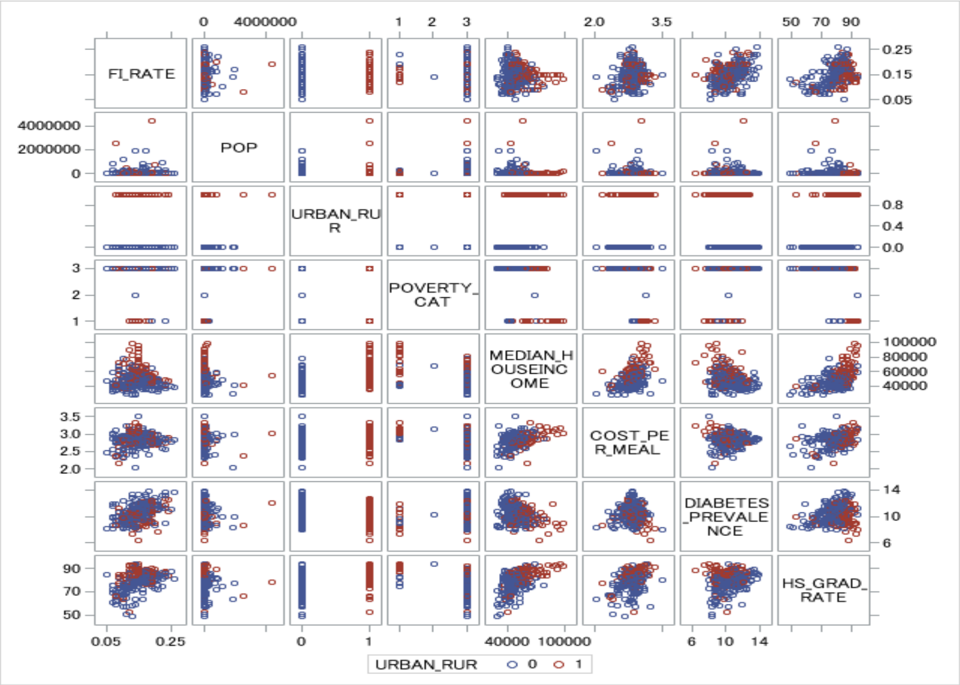


Figure 3: Ellipse Correlation Matrix

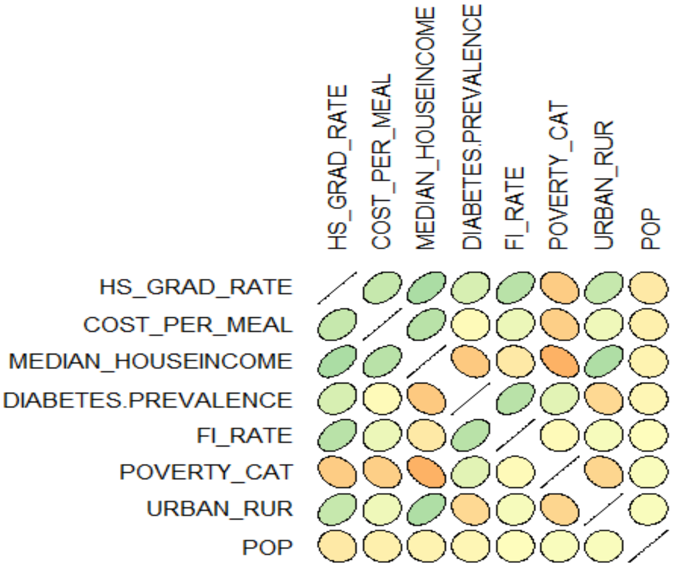
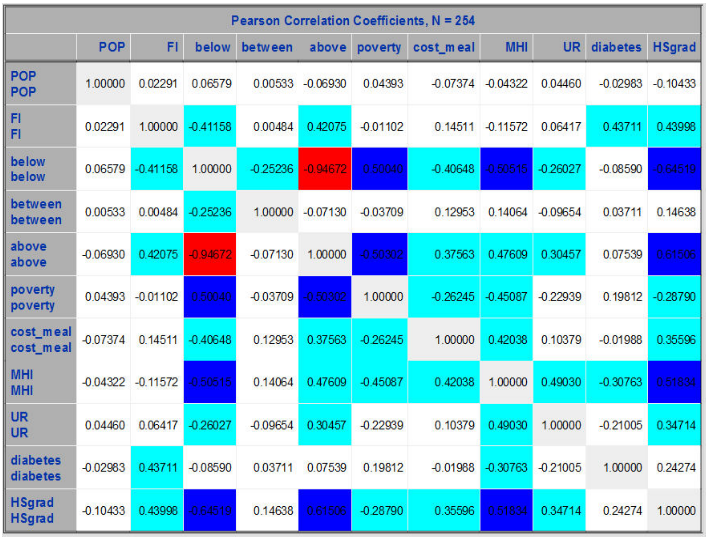


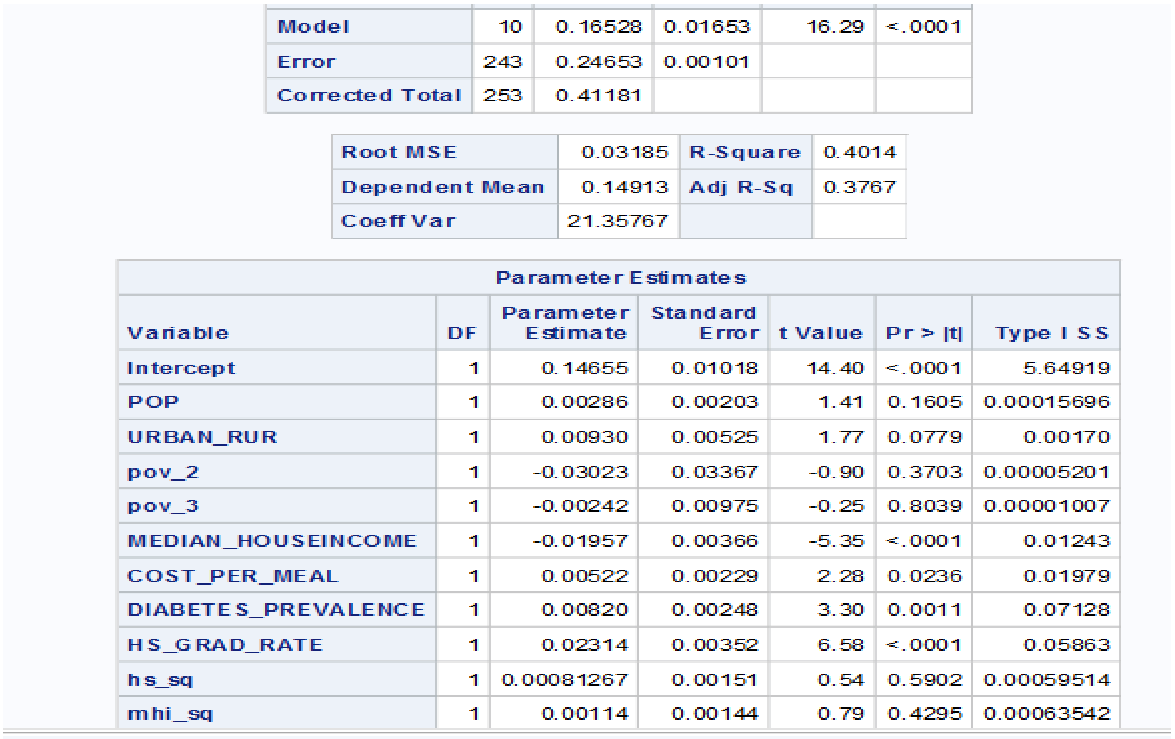
Figure 4: Correlation Matrix



Pearson Correlation Coefficients, N = 254

	POP	FI	below	between	above	poverty	cost_meal	MHI	UR	diabetes	HSgrad
POP POP	1.00000	0.02291	0.06579	0.00533	-0.06930	0.04393	-0.07374	-0.04322	0.04460	-0.02983	-0.10433
FI FI	0.02291	1.00000	-0.41158	0.00484	0.42075	-0.01102	0.14511	-0.11572	0.06417	0.43711	0.43998
below below	0.06579	-0.41158	1.00000	-0.25236	-0.94672	0.50040	-0.40648	-0.50515	-0.26027	-0.08590	-0.64519
between between	0.00533	0.00484	-0.25236	1.00000	-0.07130	-0.03709	0.12953	0.14064	-0.09654	0.03711	0.14638
above above	-0.06930	0.42075	-0.94672	-0.07130	1.00000	-0.50302	0.37563	0.47609	0.30457	0.07539	0.61506
poverty poverty	0.04393	-0.01102	0.50040	-0.03709	-0.50302	1.00000	-0.26245	-0.45087	-0.22939	0.19812	-0.28790
cost_meal cost_meal	-0.07374	0.14511	-0.40648	0.12953	0.37563	-0.26245	1.00000	0.42038	0.10379	-0.01988	0.35596
MHI MHI	-0.04322	-0.11572	-0.50515	0.14064	0.47609	-0.45087	0.42038	1.00000	0.49030	-0.30763	0.51834
UR UR	0.04460	0.06417	-0.26027	-0.09654	0.30457	-0.22939	0.10379	0.49030	1.00000	-0.21005	0.34714
diabetes diabetes	-0.02983	0.43711	-0.08590	0.03711	0.07539	0.19812	-0.01988	-0.30763	-0.21005	1.00000	0.24274
HSgrad HSgrad	-0.10433	0.43998	-0.64519	0.14638	0.61506	-0.28790	0.35596	0.51834	0.34714	0.24274	1.00000

Figure 5: Testing squared terms



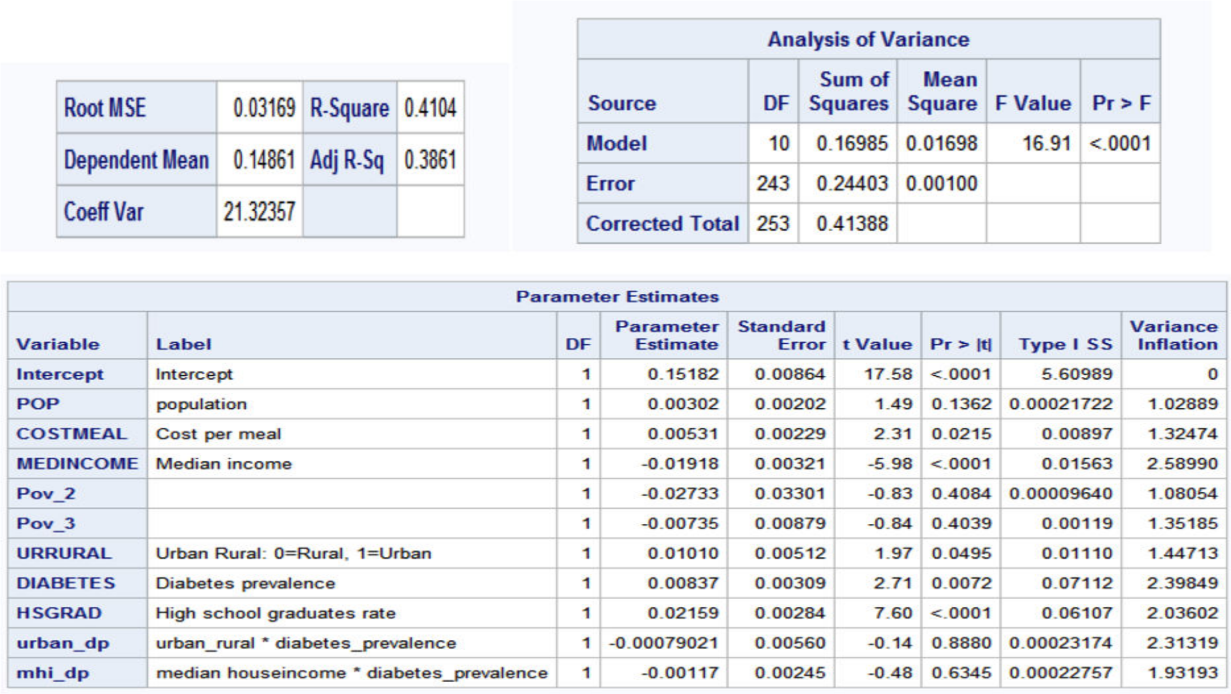
Model	10	0.16528	0.01653	16.29	<.0001
Error	243	0.24653	0.00101		
Corrected Total	253	0.41181			

Root MSE	0.03185	R-Square	0.4014
Dependent Mean	0.14913	Adj R-Sq	0.3767
Coeff Var	21.35767		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	0.14655	0.01018	14.40	<.0001	5.64919
POP	1	0.00286	0.00203	1.41	0.1605	0.00015696
URBAN_RUR	1	0.00930	0.00525	1.77	0.0779	0.00170
pov_2	1	-0.03023	0.03367	-0.90	0.3703	0.00005201
pov_3	1	-0.00242	0.00975	-0.25	0.8039	0.00001007
MEDIAN_HOUSEINCOME	1	-0.01957	0.00366	-5.35	<.0001	0.01243
COST_PER_MEAL	1	0.00522	0.00229	2.28	0.0236	0.01979
DIABETES_PREVALENCE	1	0.00820	0.00248	3.30	0.0011	0.07128
HS_GRAD_RATE	1	0.02314	0.00352	6.58	<.0001	0.05863
hs_sq	1	0.00081267	0.00151	0.54	0.5902	0.00059514
mhi_sq	1	0.00114	0.00144	0.79	0.4295	0.00063542

Figure 6: Preliminary Model Results

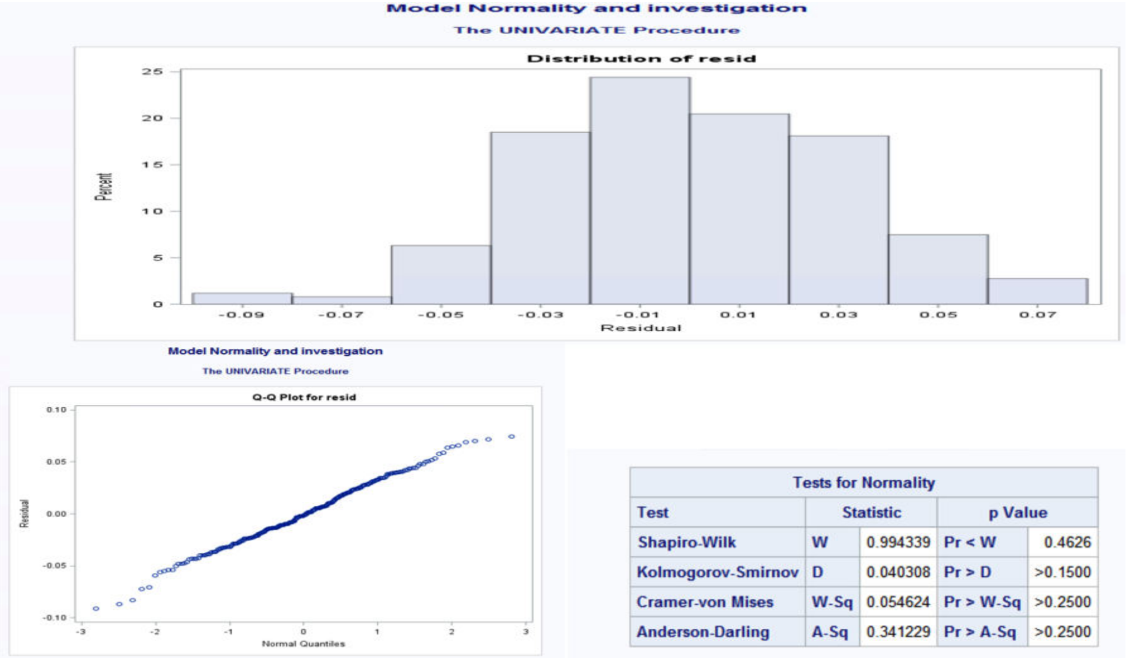


Root MSE	0.03169	R-Square	0.4104
Dependent Mean	0.14861	Adj R-Sq	0.3861
Coeff Var	21.32357		

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	0.16985	0.01698	16.91	<.0001
Error	243	0.24403	0.00100		
Corrected Total	253	0.41388			

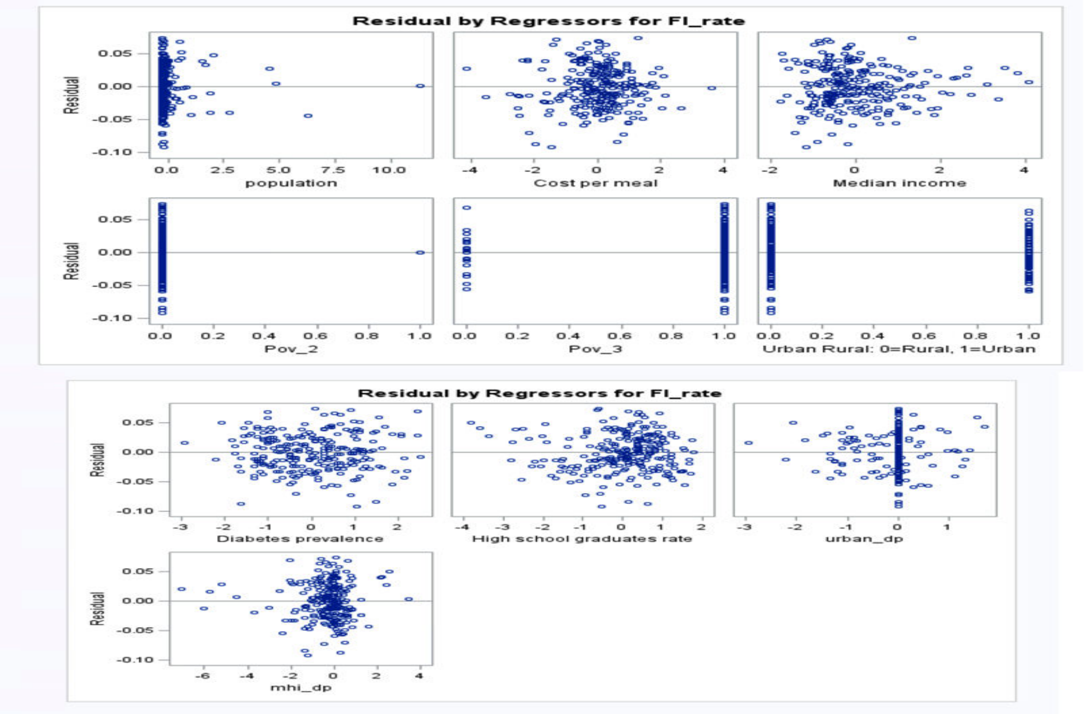
Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	Intercept	1	0.15182	0.00864	17.58	<.0001	5.60989	0
POP	population	1	0.00302	0.00202	1.49	0.1362	0.00021722	1.02889
COSTMEAL	Cost per meal	1	0.00531	0.00229	2.31	0.0215	0.00897	1.32474
MEDINCOME	Median income	1	-0.01918	0.00321	-5.98	<.0001	0.01563	2.58990
Pov_2		1	-0.02733	0.03301	-0.83	0.4084	0.00009640	1.08054
Pov_3		1	-0.00735	0.00879	-0.84	0.4039	0.00119	1.35185
URRURAL	Urban Rural: 0=Rural, 1=Urban	1	0.01010	0.00512	1.97	0.0495	0.01110	1.44713
DIABETES	Diabetes prevalence	1	0.00837	0.00309	2.71	0.0072	0.07112	2.39849
HSGRAD	High school graduates rate	1	0.02159	0.00284	7.60	<.0001	0.06107	2.03602
urban_dp	urban_rural * diabetes_prevalence	1	-0.00079021	0.00560	-0.14	0.8880	0.00023174	2.31319
mhi_dp	median houseincome * diabetes_prevalence	1	-0.00117	0.00245	-0.48	0.6345	0.00022757	1.93193

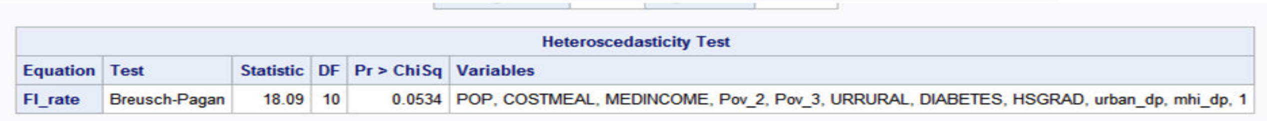
Figure 7: Model Normality and Investigation



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994339	Pr < W	0.4626
Kolmogorov-Smirnov	D	0.040308	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.054624	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.341229	Pr > A-Sq	>0.2500

Figure 8: Selected Variance Output for Preliminary Model

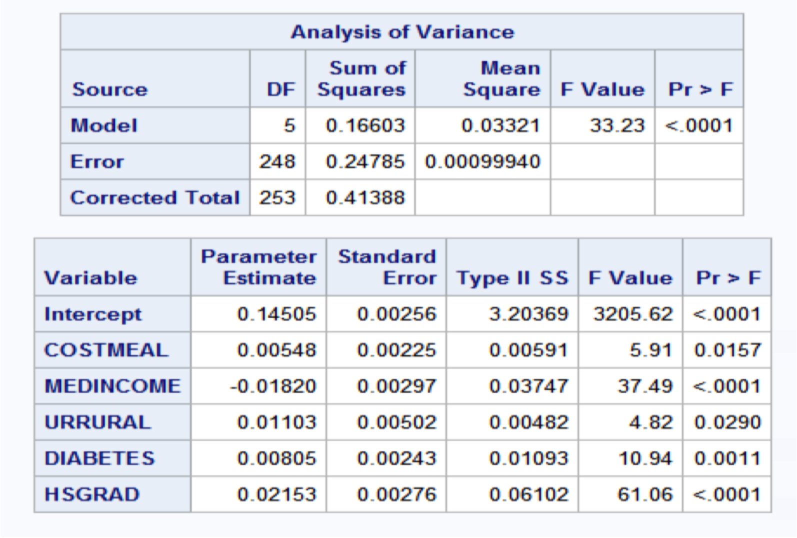




Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
FI_rate	Breusch-Pagan	18.09	10	0.0534	POP, COSTMEAL, MEDINCOME, Pov_2, Pov_3, URRURAL, DIABETES, HSGRAD, urban_dp, mhi_dp, 1

Figure 9: Stepwise Selection Results

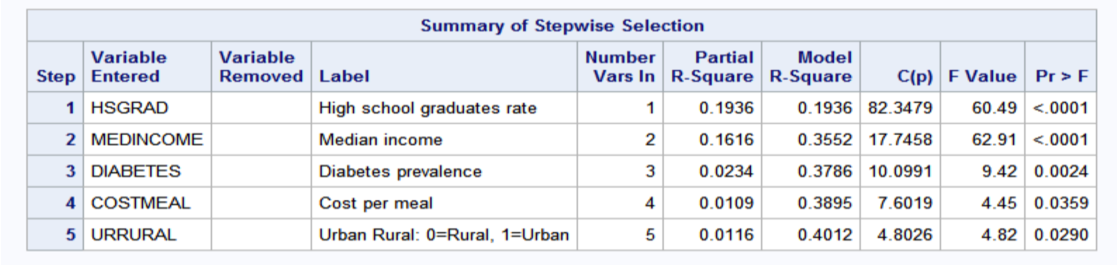


Analysis of Variance

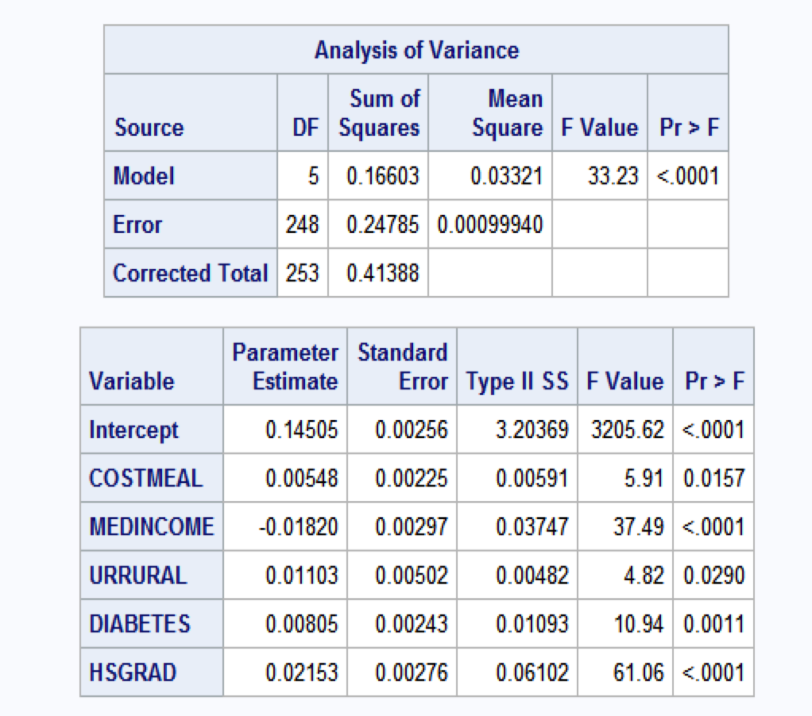
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.16603	0.03321	33.23	<.0001
Error	248	0.24785	0.00099940		
Corrected Total	253	0.41388			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.14505	0.00256	3.20369	3205.62	<.0001
COSTMEAL	0.00548	0.00225	0.00591	5.91	0.0157
MEDINCOME	-0.01820	0.00297	0.03747	37.49	<.0001
URRURAL	0.01103	0.00502	0.00482	4.82	0.0290
DIABETES	0.00805	0.00243	0.01093	10.94	0.0011
HSGRAD	0.02153	0.00276	0.06102	61.06	<.0001

Figure 10: Backward Selection Results



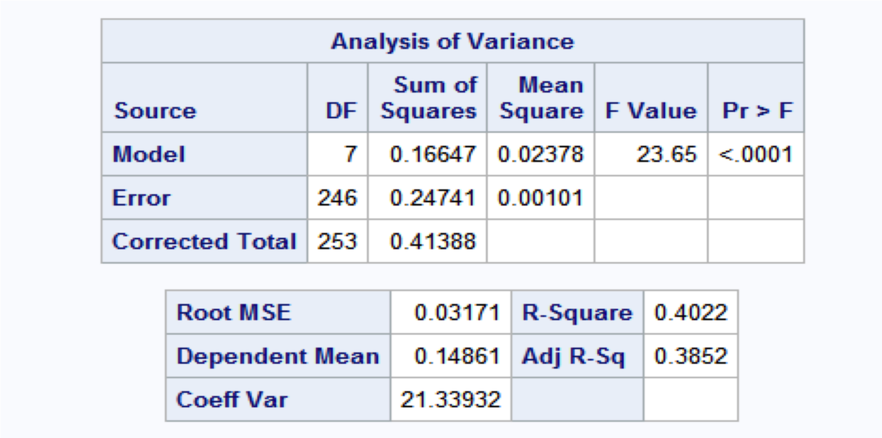
Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	HSGRAD		High school graduates rate	1	0.1936	0.1936	82.3479	60.49	<.0001
2	MEDINCOME		Median income	2	0.1616	0.3552	17.7458	62.91	<.0001
3	DIABETES		Diabetes prevalence	3	0.0234	0.3786	10.0991	9.42	0.0024
4	COSTMEAL		Cost per meal	4	0.0109	0.3895	7.6019	4.45	0.0359
5	URRURAL		Urban Rural: 0=Rural, 1=Urban	5	0.0116	0.4012	4.8026	4.82	0.0290



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.16603	0.03321	33.23	<.0001
Error	248	0.24785	0.00099940		
Corrected Total	253	0.41388			

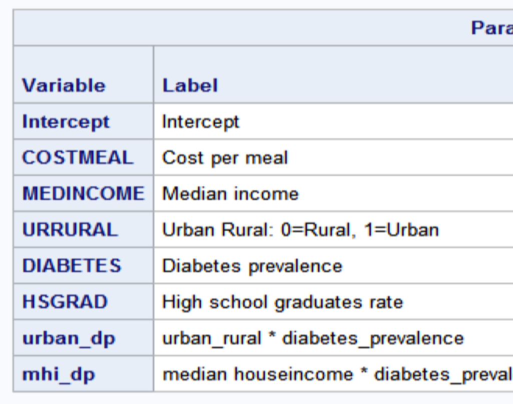
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.14505	0.00256	3.20369	3205.62	<.0001
COSTMEAL	0.00548	0.00225	0.00591	5.91	0.0157
MEDINCOME	-0.01820	0.00297	0.03747	37.49	<.0001
URRURAL	0.01103	0.00502	0.00482	4.82	0.0290
DIABETES	0.00805	0.00243	0.01093	10.94	0.0011
HSGRAD	0.02153	0.00276	0.06102	61.06	<.0001

Figure 11: LASSO model results



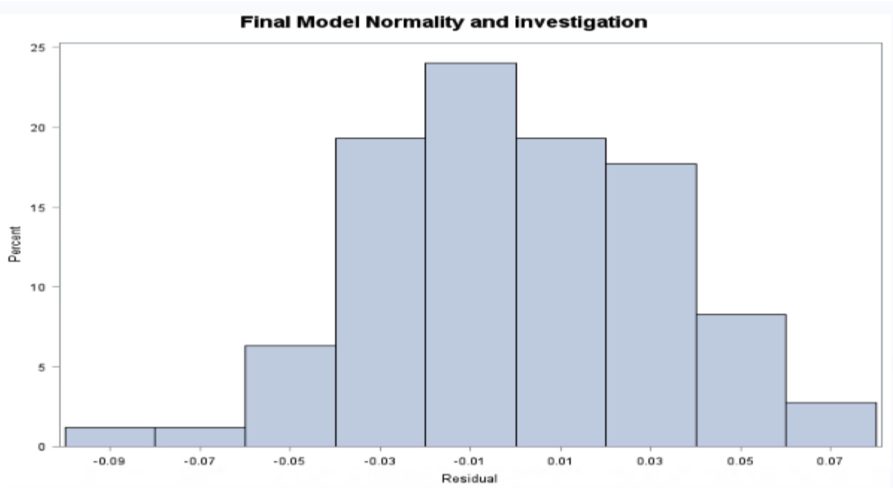
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.16647	0.02378	23.65	<.0001
Error	246	0.24741	0.00101		
Corrected Total	253	0.41388			

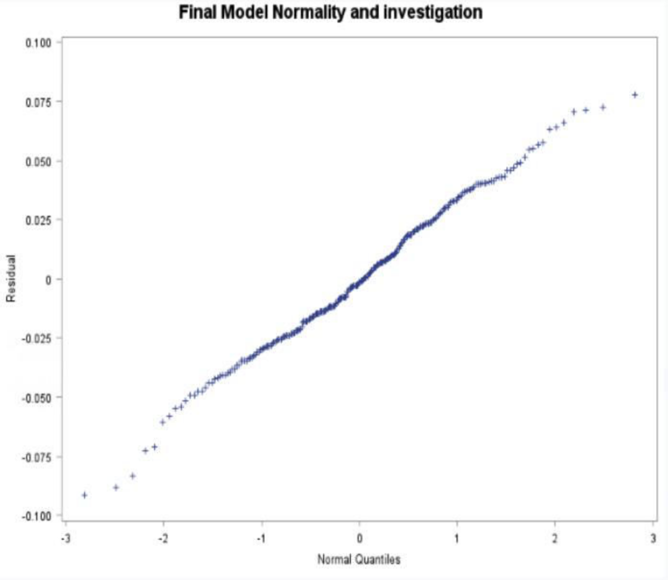
Root MSE	0.03171	R-Square	0.4022
Dependent Mean	0.14861	Adj R-Sq	0.3852
Coeff Var	21.33932		

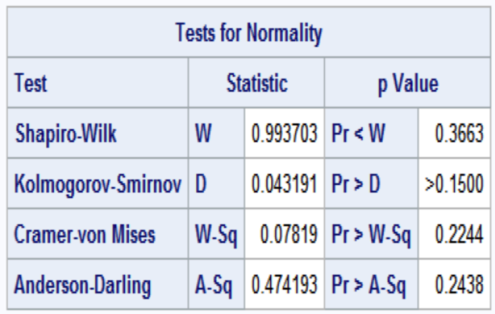


Para	
Variable	Label
Intercept	Intercept
COSTMEAL	Cost per meal
MEDINCOME	Median income
URRURAL	Urban Rural: 0=Rural, 1=Urban
DIABETES	Diabetes prevalence
HSGRAD	High school graduates rate
urban_dp	urban_rural * diabetes_prevalence
mhi_dp	median houseincome * diabetes_preval

Figure 12: Normality plot for Final Model

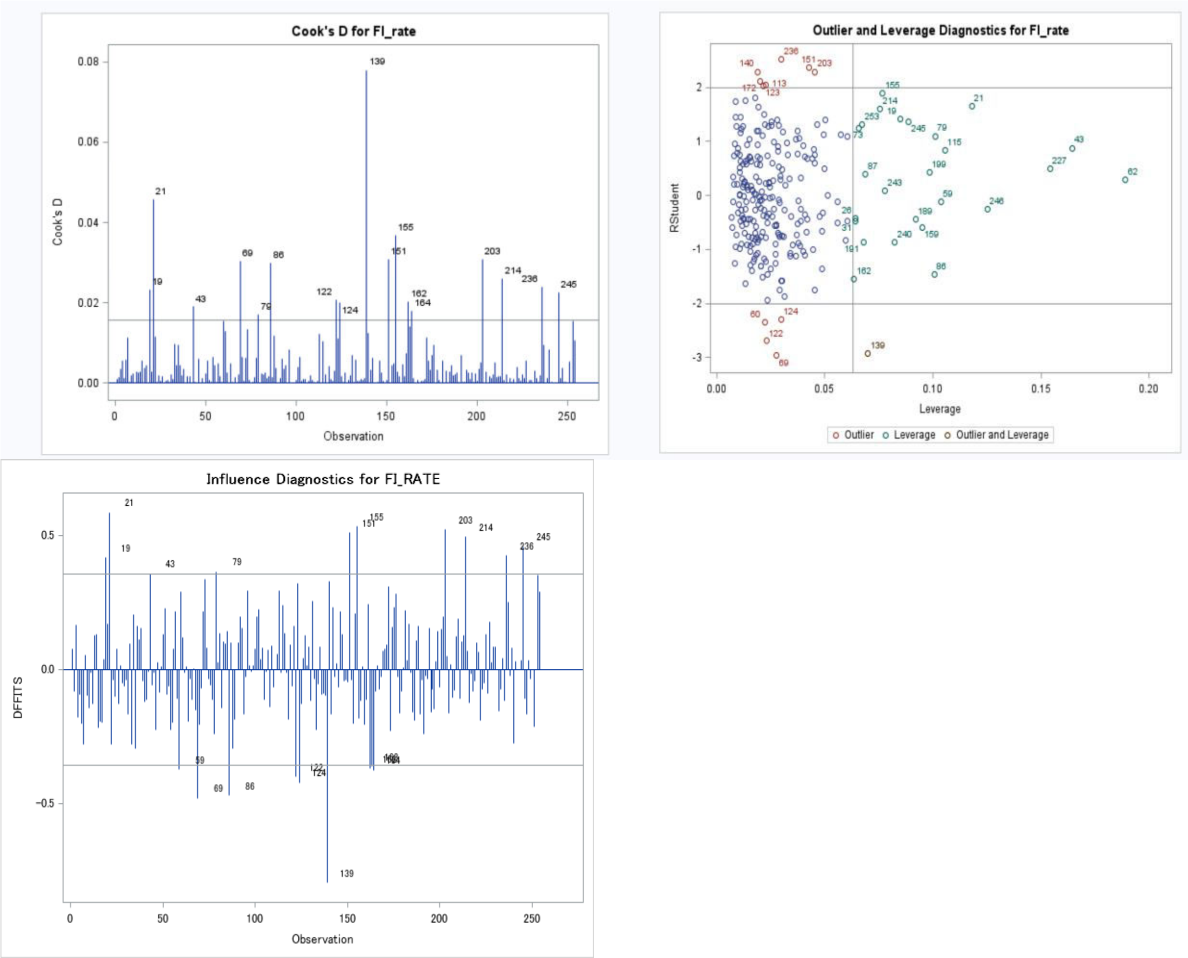




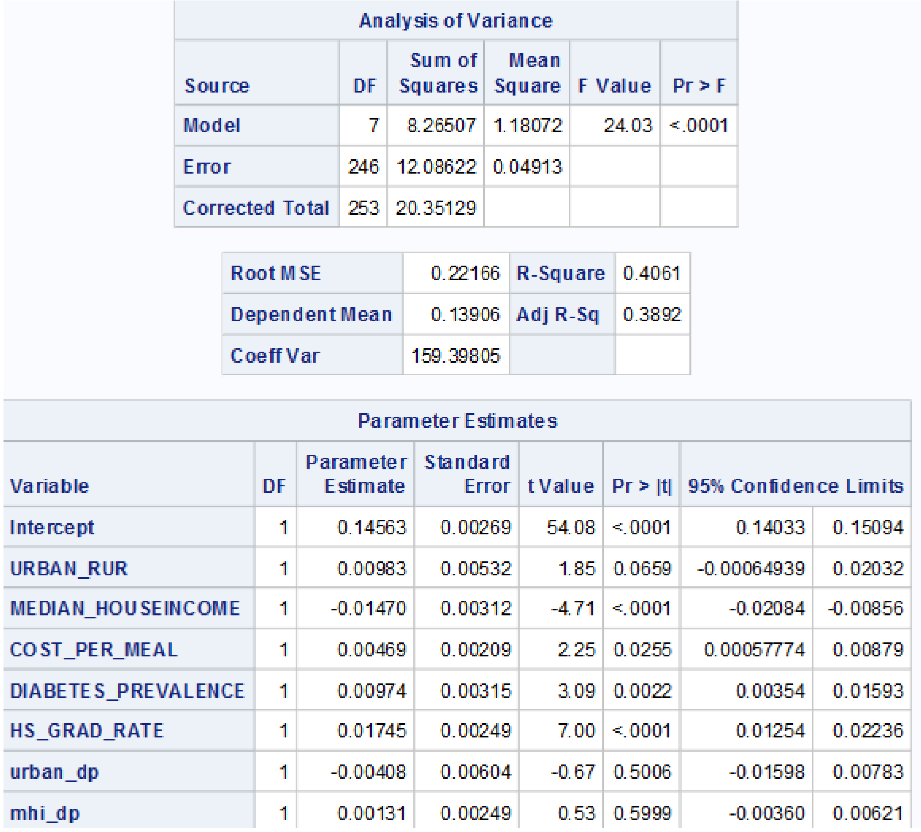


Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.993703	Pr < W	0.3663
Kolmogorov-Smirnov	D	0.043191	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.07819	Pr > W-Sq	0.2244
Anderson-Darling	A-Sq	0.474193	Pr > A-Sq	0.2438

Figure 13: Leverage, DFFITS and Cook’s D Test for Final Model



\
Figure 14: Final Model Results



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8.26507	1.18072	24.03	<.0001
Error	246	12.08622	0.04913		
Corrected Total	253	20.35129			

Root MSE	0.22166	R-Square	0.4061
Dependent Mean	0.13906	Adj R-Sq	0.3892
Coeff Var	159.39805		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.14563	0.00269	54.08	<.0001	0.14033	0.15094
URBAN_RUR	1	0.00983	0.00532	1.85	0.0659	-0.00064939	0.02032
MEDIAN_HOUSEINCOME	1	-0.01470	0.00312	-4.71	<.0001	-0.02084	-0.00856
COST_PER_MEAL	1	0.00469	0.00209	2.25	0.0255	0.00057774	0.00879
DIABETES_PREVALENCE	1	0.00974	0.00315	3.09	0.0022	0.00354	0.01593
HS_GRAD_RATE	1	0.01745	0.00249	7.00	<.0001	0.01254	0.02236
urban_dp	1	-0.00408	0.00604	-0.67	0.5006	-0.01598	0.00783
mhi_dp	1	0.00131	0.00249	0.53	0.5999	-0.00360	0.00621

(SAS code)

```
/* TX_food 2016 */
data TX_food;
infile "H:\food.csv" DSD firstobs=2;
input COUNTY $ POP:comma. FI_rate:percent. BELOW:percent. BETWEEN:percent.
ABOVE:percent.  POVERCAT  COSTMEAL:dollar.  MEDINCOME:dollar. URRURAL
DIABETES  HSGRAD;
label COUNTY = "County"
      POP = "population"
      FI_rate = "Food Insecure rate"
      BELOW = "Below 165"
      BETWEEN = "Between 165-185"
      ABOVE = "Above 185"
      POVERCAT = "Poverty category: 1= 165% below poverty level, 3= more than
185% poverty level"
      COSTMEAL = "Cost per meal"
      MEDINCOME = "Median income"
      URRURAL = "Urban Rural: 0=Rural, 1=Urban"
      DIABETES = "Diabetes prevalence"
      HSGRAD = "High school graduates rate";
run;
```

```
/* Scatterplot matrix */
proc sgscatter data=TX_food;
matrix FI_rate POP URRURAL POVERCAT MEDINCOME COSTMEAL DIABETES HSGRAD /
group=URRURAL;
run;

/* Ellipse Correlation Matrix */
names(food)

food <- TX_food %>% select(-c(BETWEEN165_185, ABOVE_185, BELOW_165))
data=cor(food[,-1])
# Build a Pannel of 100 colors with Rcolor Brewer
my_colors <- brewer.pal(5, "Spectral")
my_colors=colorRampPalette(my_colors)(100)

# Order the correlation matrix
ord <- order(data[1, ])
data_ord = data[ord, ord]
plotcorr(data_ord , col=my_colors[data_ord*50+50] , mar=c(1,1,1,1)  )

/* standardized var  */
PROC STANDARD DATA=TX_food MEAN=0 STD=1 OUT=food_scaled;
  VAR POP MEDINCOME COSTMEAL DIABETES HSGRAD;
RUN;

/* dummy variable */
data food2;
set food_scaled;
if POVERCAT = 2 then Pov_2 =1; else Pov_2 =0;
if POVERCAT = 3 then Pov_3 =1; else Pov_3 =0;
urban_dp = URRURAL*DIABETES;
mhi_dp = MEDINCOME*DIABETES;
hs_sq = HSGRAD**2;
mhi_sq = MEDINCOME**2;
label urban_dp ="urban_rural * diabetes_prevalence"
      mhi_dp = "median houseincome * diabetes_prevalence"
      hs_sq = "high school graduation squared"
      mhi_Sq = "median house income squared";
Run;

/*Testing squared terms*/
proc reg data = food2 outest=est;
      model FI_rate = POP URRURAL Pov_2 Pov_3 MEDINCOME COSTMEAL DIABETES
HSGRAD hs_sq mhi_sq/ss1 r p;
      plot r.*p.;
      plot npp.*r.;
      output out = outfood p=yhat r=resid stdr = stdresid ;
run;
```

```
/*Drop varaibles (sqaured terms, below, median, above, county, Povertcat)*/
data Total;
set food2 (KEEP = POP FI_rate Pov_2 Pov_3 COSTMEAL MEDINCOME URRURAL DIABETES
HSGRAD urban_dp mhi_dp);
run;

/* Premeliminary Model Diagnostics */
proc reg data = Total outest=est;
      model FI_rate = POP COSTMEAL MEDINCOME Pov_2 Pov_3 URRURAL DIABETES
HSGRAD urban_dp mhi_dp/ss1 vif r p;
      plot r.*p.;
      plot npp.*r.;
      output out = outfood p=yhat r=resid stdr = stdresid ;
run;

/* Check normality */
title "Model Normality and investigation";
proc univariate data= outfood normal;
var resid;
histogram;
qqplot;
run;

/* BP test to detect homoscedasticity */
proc model data= Total;
FI_rate = a0+ b1*POP + b2*COSTMEAL + b3*MEDINCOME + b4*Pov_2 + b5*Pov_3 +
b6*URRURAL + b7*DIABETES + b8*HSGRAD + b9*urban_dp + b10*mhi_dp;
fit FI_rate PARMS= (a0 b1 b2 b3 b4 b5 b6 b7 b8 b9 b10) / breusch=(POP
COSTMEAL MEDINCOME Pov_2 Pov_3 URRURAL DIABETES HSGRAD urban_dp mhi_dp);
run;

/*Perform stepwise selection */
title "Stepwise Model selection";
proc reg data = Total outest = est1;
      model FI_rate = POP COSTMEAL MEDINCOME Pov_2 Pov_3 URRURAL DIABETES
HSGRAD urban_dp mhi_dp / r p selection = stepwise sle = 0.05 sls=0.1 aic;
      plot r.*p.;
      plot npp.*r.;
       output out = outfood p=yhat r=resid stdr = stdresid ;
run;

/*Perform backward selection */
title "backward Model selection";
proc reg data = Total outest = est;
      model FI_rate = POP COSTMEAL MEDINCOME Pov_2 Pov_3 URRURAL DIABETES
HSGRAD urban_dp mhi_dp / r p selection = backward aic;
      plot r.*p.;
      plot npp.*r.;
      output out = outfood p=yhat r=resid stdr = stdresid ;
```

```
run;

/*Lasso model*/
proc glmselect data = food2 plots(stepaxis=normb)=all;
model FI_rate = POP COSTMEAL MEDINCOME Pov_2 Pov_3 URRURAL DIABETES HSGRAD
urban_dp mhi_dp / selection=lasso(stop=none choose=AIC);
run;

/*Final Model */
data Final;
set Total (KEEP = FI_rate COSTMEAL MEDINCOME URRURAL DIABETES HSGRAD urban_dp
mhi_dp);
run;
proc sgscatter data=Final;
matrix FI_rate COSTMEAL MEDINCOME URRURAL DIABETES HSGRAD urban_dp mhi_dp /
group=URRURAL;
run;

/* Final Model Diagnostics */
proc reg data= Final;
model FI_rate = COSTMEAL MEDINCOME URRURAL DIABETES HSGRAD urban_dp mhi_dp /
ss1;
plot nap.*r.;
plot r.*p.;
output out =outfoodfinal p=yhat r=resid stdr=stdresid;
run;

/* Final Model Check normality */
title "Final Model Normality and investigation";
proc univariate data= outfoodfinal normal;
var resid;
histogram;
qqplot;
run;

/*Cook, DFFITS, leverage */
ods graphics on;
proc reg data= Total
plots(label) = (CooksD RstudentByLeverage DFFITS DFBETAS);
model FI_rate = COSTMEAL MEDINCOME URRURAL DIABETES HSGRAD urban_dp mhi_dp;
run;
ods graphics off;

/* Final model BP test to detect homoscedasticity */
proc model data= Total;
FI_rate = a0+ b1*COSTMEAl + b2*MEDINCOME + b3*URRURAL + b4*DIABETES +
b5*HSGRAD + b6*urban_dp + b7*mhi_dp;
fit FI_rate PARMS= (a0 b1 b2 b3 b4 b5 b6 b7) / breusch=(COSTMEAL MEDINCOME
URRURAL DIABETES HSGRAD urban_dp mhi_dp);
```

```
run;

/*Weighted Final model*/
data outfoodfinal;
set outfoodfinal;
absr=abs(resid); sqrr=resid*resid;
proc reg data=outfoodfinal;
model FI_rate = COSTMEAL MEDINCOME URRURAL DIABETES HSGRAD urban_dp mhi_dp;
output out=Weightmodel p=shat;
data Weightmodel;set Weightmodel;
wt=1/(shat*shat);


/* Weighted model BP test to detect homoscedasticity */
proc model data= Weightmodel;
FI_rate = a0+ b1*COSTMEAl + b2*MEDINCOME + b3*URRURAL + b4*DIABETES +
b5*HSGRAD + b6*urban_dp + b7*mhi_dp;
fit FI_rate PARMS= (a0 b1 b2 b3 b4 b5 b6 b7) / breusch=(COSTMEAL MEDINCOME
URRURAL DIABETES HSGRAD urban_dp mhi_dp);
run;
```