Adam Chu and Dennis Lu (TF: Bambo Sosina)

Statistics 104

December 5, 2011

**Influences on State Obesity Rates**

**Abstract**

This project attempted to determine the relationship between state obesity rates and variables including but not limited to fast food prevalence, level of education, poverty rates, elevation of geographical residence, and political affiliation. By comparing differences in obesity rates among the 50 U.S. states with variations in these factors, inferences could be drawn about possible predictors of one of the fastest growing causes of death and disease in America. The analysis was split into several stages, and a final model for obesity rate was obtained through multivariate linear regression analysis. According to the final regression model, number of fast food restaurants per 100,000 people and percent below the poverty line were positively correlated with obesity rates, while percentage with a bachelor’s degree and average elevation were negatively correlated with obesity rates. One categorical variable, political affiliation, revealed that those states who voted for Obama in the 2008 presidential election had on average 1.32% lower obesity rates than those who voted against him, implicitly suggesting that on average, obesity rates are lower among Democrats. This analysis may ultimately expose insight on the consequences of government policies and the best way to improve the well-being of the public. For example, policy affecting the level of education, as suggested by the model, may have a potentially positive impact by reducing state obesity rates. The analysis demonstrates the surprising correlation between level of education and an issue as seemingly unrelated as the medical condition of obesity.

**Introduction**

About one-third of U.S. adults (33.8%) are obese. Approximately 17% (or 12.5 million) of children and adolescents aged 2-19 years are obese (CDC Obesity). During the past 20 years, there has been a dramatic increase in obesity in the United States and rates remain high. In 1996, no state had an obesity rate greater than 20%; as recently as 2010, however, all states had an obesity rate greater than 20%, with many exceeding even 30%. Clearly, this growing problem in America is reason for concern. Aside from the negative health consequences, obesity may even have economic implications, such as loss of worker productivity.

With this information in mind, the current analysis seeks to identify possible influences on state obesity rates. Starting with a diverse array of possible explanatory variables, the analysis attempts to determine the extent of associations between the explanatory variables and state obesity rates. Furthermore, a final goal of the analysis is the creation of a model to reliably predict state obesity rates given certain state statistics.

**Methods**

*Variable sources and definitions.* All data was obtained from reputable Internet sources (see References for sites). The most recent data was found for the variables, and the only data not within the last decade or so was for the explanatory variable average annual temperature in degrees Fahrenheit (“temperature,” 1971-2000), which is expected not to have changed significantly since then. The other variables are the primary variable of interest, obesity rate (“obesity,” 2010), and the other explanatory variables, population density in people per square mile (“density,” 2010), average income in dollars (“income,” 2010), average age (“age,” 2009), number of fast food restaurants per 100,000 people (“fastfood,” 2005), region (“south,” “midwest,” “west,” South, Midwest, West, or Northeast), geography (“coastal,” coastal or non-coastal), violent crimes per 100,000 people (“crimes,” 2006), percent of population who have completed a bachelor’s degree (“bachelorsdegree,” 2004), vote for 2008 presidential election (“obama,” Republican or Democrat), dollars of health care spending per capita (“healthcare,” 2007), average household size (“household,” 2004), percent of population below poverty level (“poverty,” 2004), average elevation in meters (“elevation”), urban-rural population ratio (“urbanrural,” 2000), percent of employed who used public transit (“publictransit,” 2010), percent of employed who walked to work (“walk,” 2010), BBtu of energy consumption per 10 people (“energy,” 2001), and unemployment rate (“unemployment,” 2010).

*Choice of variables.* Some of the variables, such as number of fast food restaurants and average income, were chosen because it was intuitively believed that such variables would naturally be correlated with obesity rates (e.g., greater presence of fast food restaurants seems to be associated with greater prevalence of unhealthy dining habits, larger average income seems to be associated with greater food quality and diversity). Many of the other variables, however, were selected because of curiosity about possible unforeseen correlations between obesity rates and a wide array of factors (e.g., population density, vote for the 2008 election, etc.).

*Initial inspection for dramatic outliers.* After initially obtaining data for the 50 states and the District of Columbia, it was decided that because the population densities for Alaska and the District of Columbia were exceptionally low and high, respectively, they would be removed from the data analysis.

*Analytical methods.* Since there were many variables initially, single-variable linear regressions were performed to determine which individual variables were significant in explaining obesity rates. With regard to the categorical variables (region, geography, and vote during 2008 election), dummy variables were established (three for region, one for geography, and one for 2008 election). Upon obtaining a list of significant individual variables, a multivariate linear regression was performed on them, and then backwards regression analysis was employed to determine the significant variables for the overall model. Tests for omitted variables, heteroskedasticity, and normal residuals were performed. A complete residual analysis, examining residuals and standardized residuals for patterns and outliers, was also performed. After removing two outliers, multivariate linear regression was run again to obtain the final model. As a final check, tests for omitted variables, heteroskedasticity, and residual normality were performed a second time, and another residual analysis was completed.

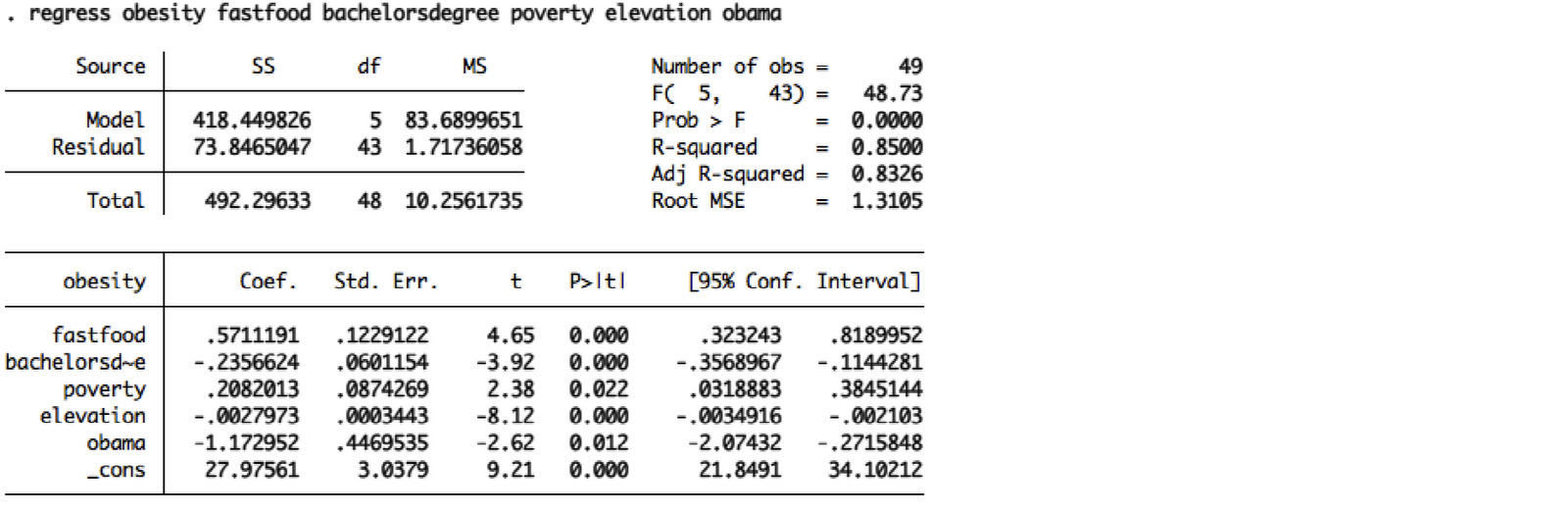
**Results**

Initial single-variable linear regressions determined that the following explanatory variables are significant:

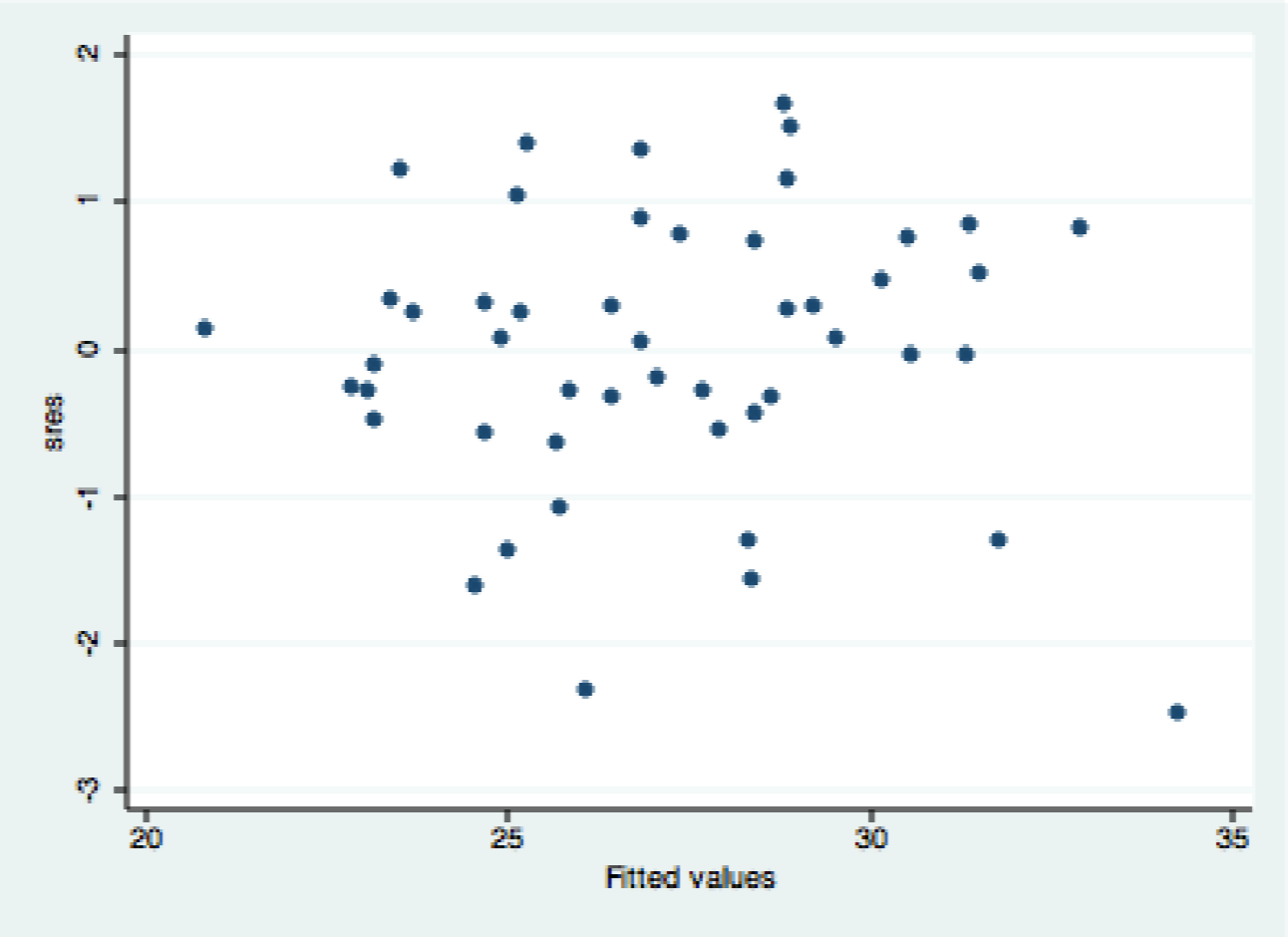
*Quantitative*: income, temperature, fastfood, bachelorsdegree, poverty, elevation, urbanrural, publictransit, walk, energy

*Categorical*: obama, south, midwest

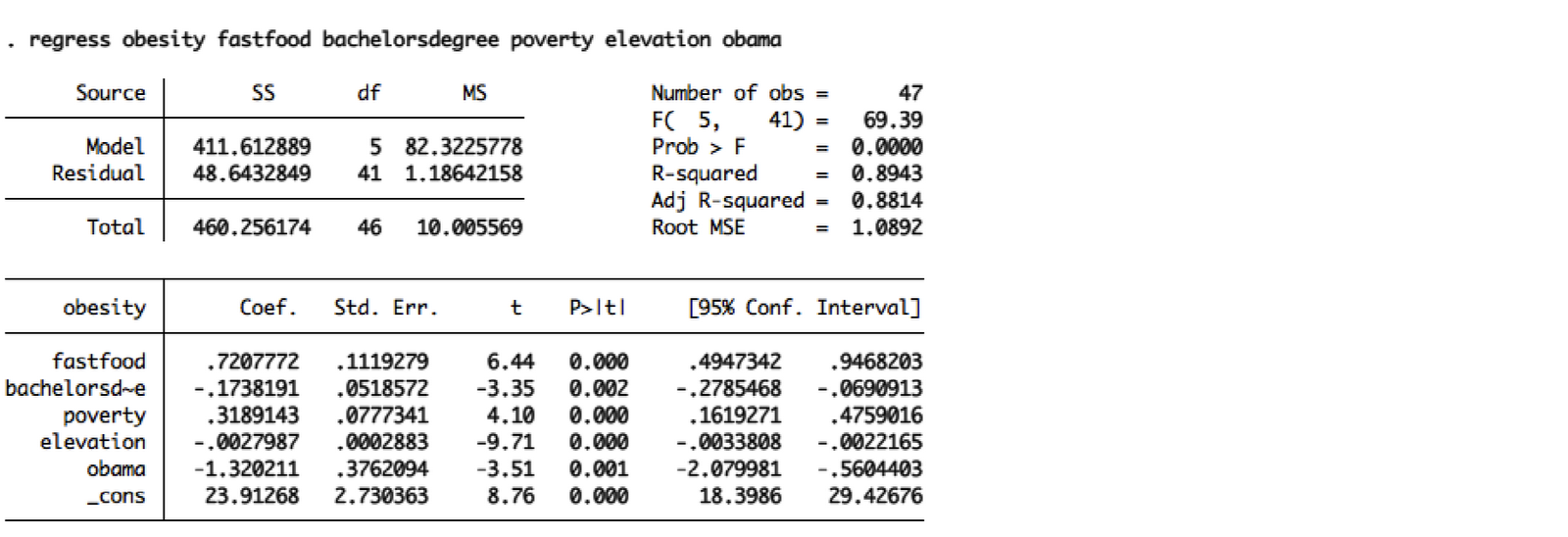
Multivariate linear regression and backwards regression analysis resulted in following model:



Tests for heteroskedasticity, normal residuals, and omitted variables indicate that although the variance is constant and the residuals are normal, there seems to be omitted variables. A complete residual analysis shows us that though there is no apparent pattern, there are two outliers (standardized residuals < -2).

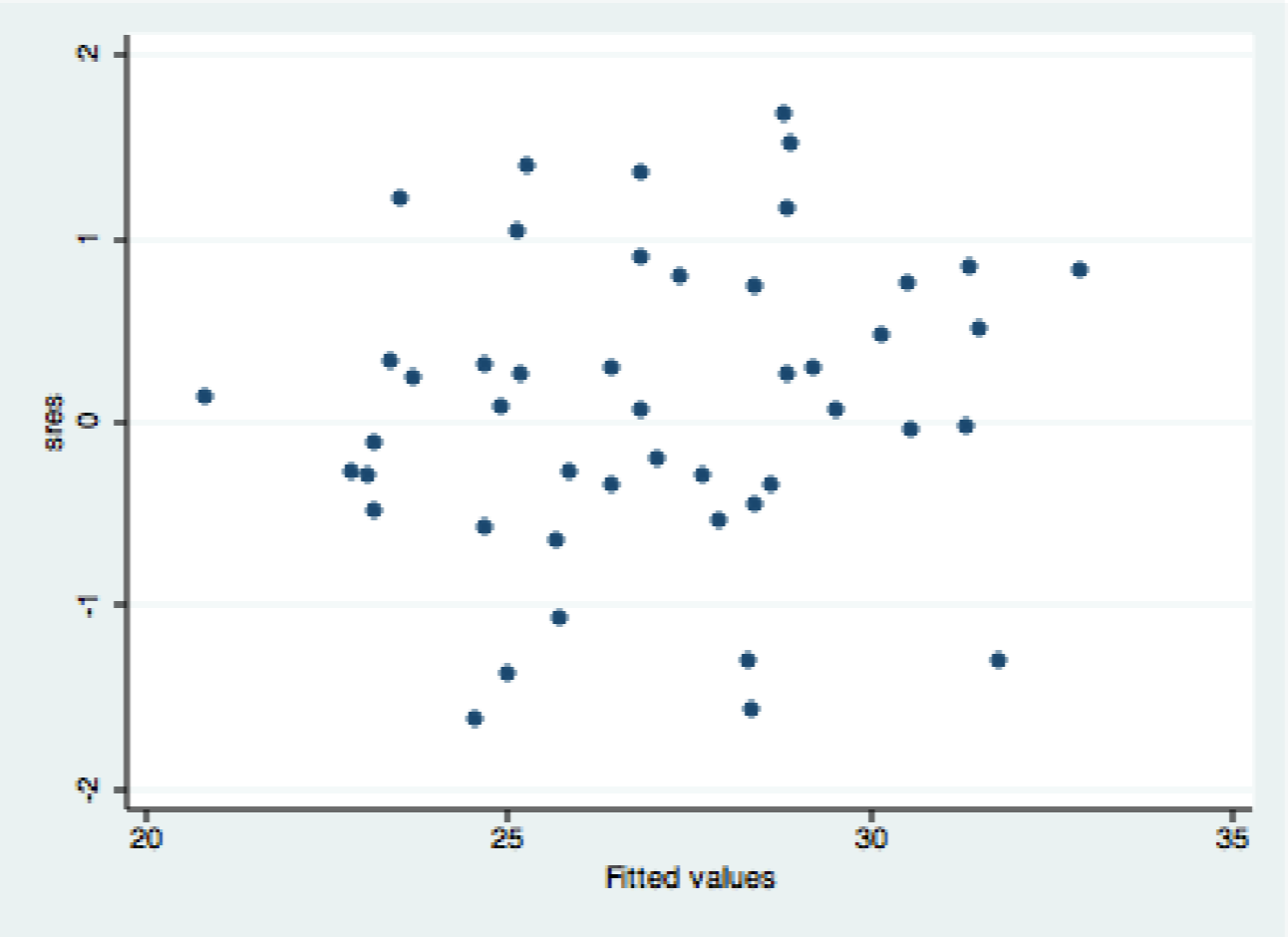
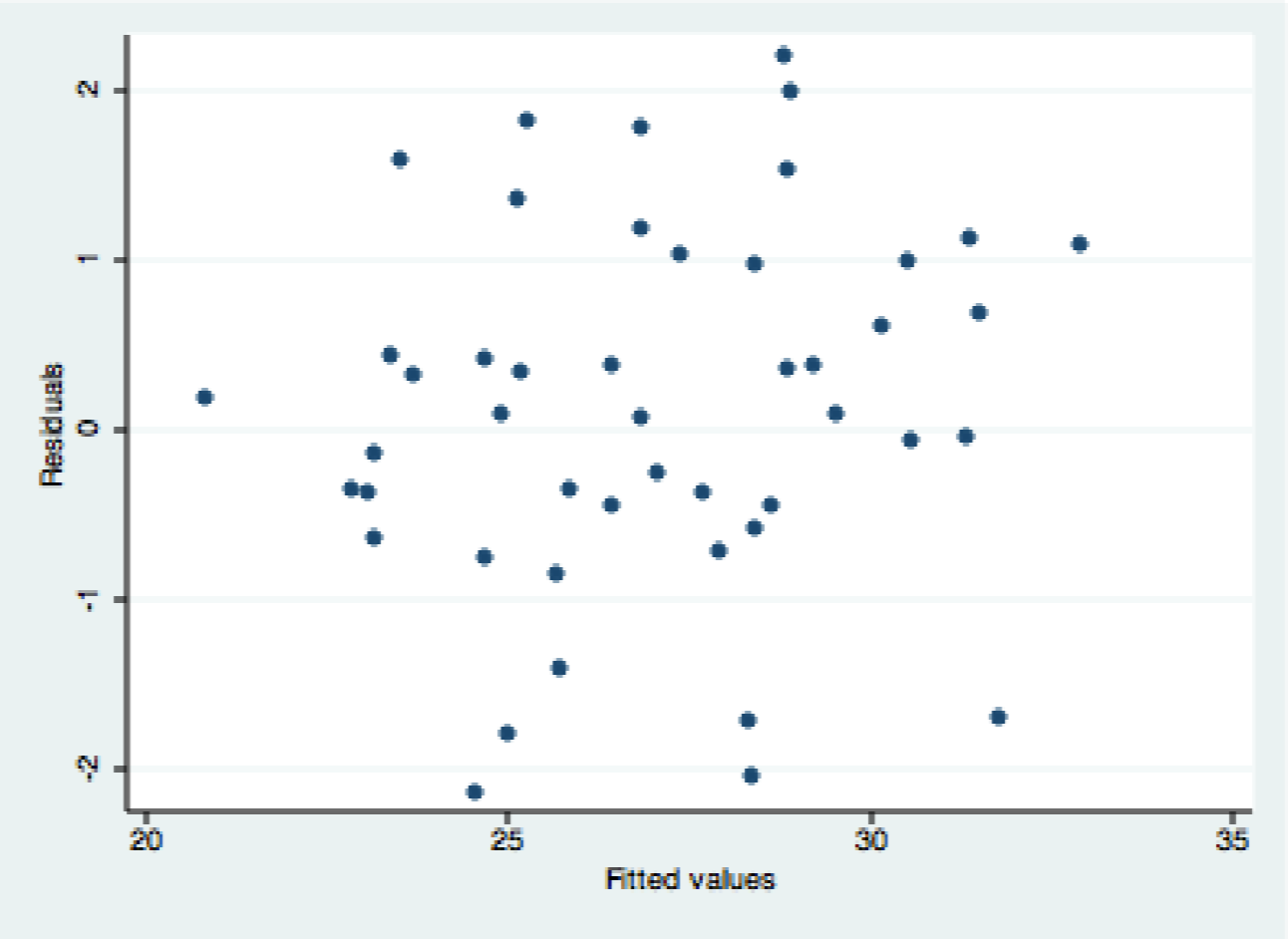


Upon removal of the outliers (Louisiana and Montana), a final regression model was obtained:



*obesity = .721\*fastfood – .174\*bachelorsdegree + .319\*poverty – .00280\*elevation – 1.32\*obama + 23.9*

Tests for heteroskedasticity, residual normality, and omitted variables show that the variance is constant, the residuals are normal, and there are no omitted variables. A complete residual analysis shows no pattern and no outliers:



**Conclusions and Discussion**

The final regression model was:

*obesity = .721\*fastfood – .174\*bachelorsdegree + .319\*poverty – .00280\*elevation – 1.32\*obama + 23.9*

Performing the F test on this model revealed Prob > F ≈ 0.000 < 0.05, which is a significant test statistic. This indicated that at least one of the variables in the model was statistically significant. Upon further examination, the variables were all determined to be significant, as the individual p-values for each of the predictors were less than .05:

|  |  |
| --- | --- |
| Model component/variable | P-value |
| fastfood | 0.000 |
| bachelorsdegree | 0.002 |
| poverty | 0.000 |
| elevation | 0.000 |
| obama | 0.001 |
| constant | 0.000 |

The model indicated the following about the impact of the explanatory variables on the outcome variable obesity (holding all other variables constant):

* Number of fast food restaurants per 100,000 people was positively correlated with obesity rates; for every unit increase of fast food restaurants per 100,000 people, obesity rates increased by .721% on average.
* Percent who have completed a bachelor’s degree was negatively correlated with obesity rates; for every unit percent increase in people with bachelor’s degrees, obesity rates decreased by .174% on average.
* Percent below poverty level was positively correlated with obesity rates; for every unit percent increase in people below poverty level, obesity rates increased by .319% on average.
* Average elevation was negatively correlated with obesity rates; for every unit increase (m) in elevation, obesity rates decreased by .00280% on average.
* Political affiliation was also significantly correlated with obesity rates. Statistical output for our dummy variable “obama” (1 - voted for, 0 - voted against) showed that states who voted for Obama in the 2008 presidential election (likely assumed Democrats) had on average 1.32% lower obesity rates than those who voted against him (likely assumed Republicans).

Overall, 89.4% of the variation in state obesity rates could be explained by the model.

Some of the initial variables that seemed to be intuitively tied to obesity rates wound up in the final regression model, while some variables that seemed intuitive did not. Likewise, some of the variables that were initially chosen for pure curiosity’s sake wound up in the final regression model, while others did not. This allowed for an interesting interpretation of the variety of factors that seem to be related to obesity rates.

Variables such as the number of fast food restaurants per 100,000 people were expected to correlate with obesity rates, and the variable indeed remained in the model. Variables such as average income were also expected to correlate with obesity rates, but surprisingly, it was not found in the end model. It was initially thought that higher income would lead to greater food quality and diversity. After the regressions were performed, however, it was realized that income would likely be related to too many other variables for to have a distinct impact on obesity rates. For instance, higher income might correlate with confounding variables, including better access to technology and increased comfort as well as greater access to health resources, which may make the effect of income on obesity rates ambiguous.

The variables that were chosen and regressed for pure curiosity showed interesting results as well. Variables such as population density did not end up in the final model, while surprisingly, vote for the 2008 election did. The dummy variable for political affiliation suggested that on average, states who voted for Obama had obesity rates 1.32% lower the obesity rates for states that voted against him, with the implication that the prevalence of obesity rates is 1.32% lower among Democratic states than Republican states.

The above results suggest that while conventional thoughts regarding the predictors of obesity (unhealthiness of eating fast food) generally prevailed, obesity as a disease is more complicated and intricately linked with other variables than previously thought.

Obesity is a sizable problem for society; in addition to the health impacts, obesity may lead to increased business costs and disadvantages in employment. These affect society at all levels, from governments to corporations to individuals. The United States government has tried to address this issue with funding for obesity prevention programs (“Obesity”). One effect of these programs, an increase in lifespan, may actually increase medical costs over the individuals’ lifetime, so economically, the positive effects of such programs may be diminished.

With regard to public policy, our model offers some further insight as well.

Obesity rates were shown to be negatively correlated with level of education, i.e., lower obesity rates were associated with increased prevalence of bachelor degrees. This proves interesting for Washington policymakers because they have been making a consistent push for academic competitiveness and education for innovation. This lends support to the idea that increased education among the population might have spillover benefits - higher education leads to better health in general. Based on the results of this study, it would be reasonable for policymakers to investigate further, if they have not done so already, the other spillover benefits that education might have in store.

Number of fast food restaurants per 100,000 people was shown to be positively correlated with obesity rates. Again, this proves interesting for policymakers. Because these restaurants create externalities that negatively impact society’s health in certain aspects (obesity and weight being an important component), the government might find it prudent to show preference to applicants of certain funding programs. Knowing that decreasing the overall presence of fast food restaurants will likely decrease obesity rates over time might be a minute but significant factor in some of the decisions that the government makes with regard to building and residential development.

Likewise, percent below poverty was shown to be positively correlated obesity rates. Like fast food restaurants, the tangible presence of poverty can be viewed as a negative externality on the rest of society. This indicates that facing the issue of poverty through whichever means (funding social welfare, creating jobs, etc.) has the benefit of increasing society’s utility in addition to potentially enhancing the lives of those who were below the poverty line.

Elevation and political affiliation seem to be variables that the government could not easily (nor rationally) change. Regardless, speculative inferences can be drawn regarding the possible connections of elevation and political affiliation to obesity rates.

The model demonstrated that an increase in average elevation is associated with a decrease in state obesity rate. Experimental studies have shown that altitude at which exercise is carried out can have a positive correlation to the number of calories burned (“Altitude”). At higher altitudes with thinner air, the body responds by increasing breathing rate, increasing blood and muscle pH, and using more carbohydrates as sources of energy. These contribute to an overall increase in the BMR (Basal Metabolic Rate) and rate at which calories are burned. Because human body weight is at its most basic level a matter of energy intake and expenditure, it is natural to speculate that higher altitudes could play a significant role in the body’s energy cycle, with the resulting influence on weight and obesity rates.

The variable political affiliation revealed that those states who voted for Obama in the 2008 presidential election had on average 1.32% lower obesity rates than those who voted against him, implicitly suggesting that on average, obesity rates are lower among Democrats. Though the experimenters attempted to draw meaningful inferences as to the possible relation, political affiliation was thought to be somewhat correlated with too many other variables relating to lifestyle (income, family attitude, etc.) to pinpoint a strong potential relation. This was also the reason why interaction terms involving the variabl

Despite these potential outlets for insight, there are certain limitations of the model. The model merely indicates that there is a strong association between the explanatory variables and state obesity rates. This does not imply causation, i.e., changes in the explanatory variables does not directly cause a change in state obesity rate. As a result, there may be possible unidentified confounding variables that may influence both obesity rate and the explanatory variables.

**References**

“Does Altitude Affect Calorie Burning?” *Go Ask Alice*. Columbia University.

Web. 5 December 2011.

“Income Statistics.” U.S. Census Bureau. 1 Nov 2011. Web. 3 Dec. 2011.

“Median Age, by state.” USA Today, 10 June 2010. Web. 3 Dec. 2011.

“Obesity.” *Wikipedia*. Wikipedia, the Free Encyclopedia. Web. 1 Dec. 2011.

“Population.” *The 2012 Statistical Abstract*. U.S. Census Bureau. Web. 3

StataCorp (2010). StataSE (Version 11) [Computer software]. College Station, TX: StataCorp

LP.

*StateMaster*. Rapid Intelligence Web Publishing Company. Web. 4 December 2011.

“State Transportation Statistics.” *Bureau of Transportation Statistics*. Research and Innovative

Technology Administration. Web. 3 December 2011.

“Unemployment Rates for States.” *Bureau for Labor Statistics*. United States Department of

Labor.” Web. 3 December 2011.

“United States Presidential Election, 2008.” *Wikipedia*. Wikipedia, the Free Encyclopedia. Web. 1 Dec. 2011.

“U.S. Obesity Trends.” *Center for Disease Control and Prevention*. CDC, 2011.

Web. 3 December 2011.

**Appendix**

**Figures Group 1: Regressing Individually with Quantitative Variables**

Figure 1.1

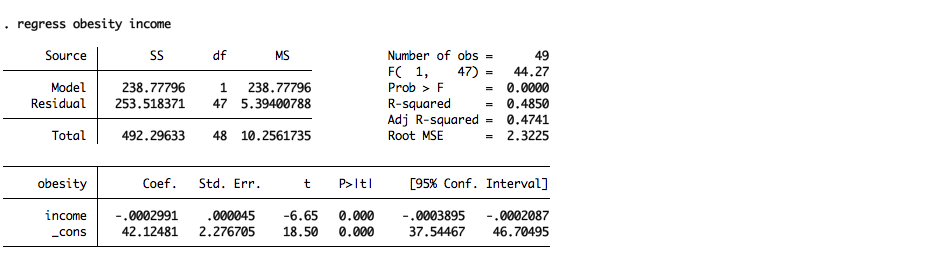


Figure 1.2

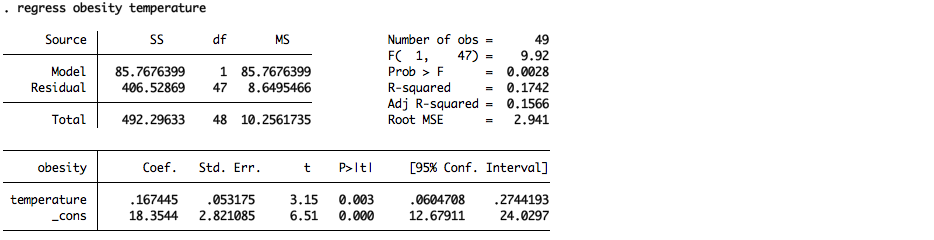


Figure 1.3

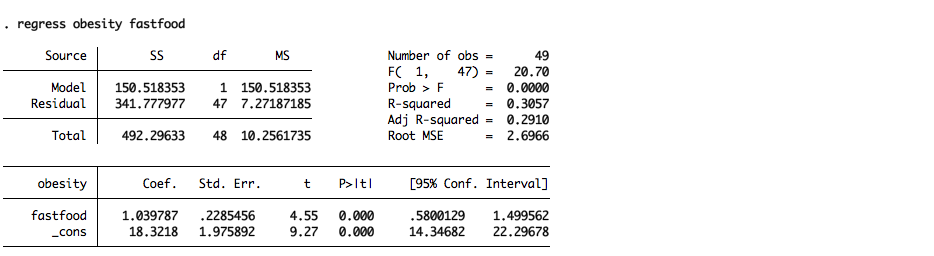


Figure 1.4

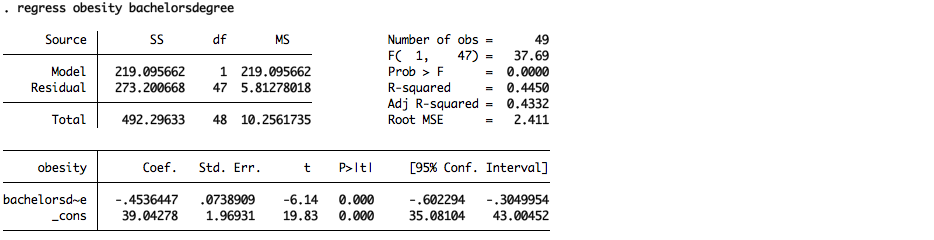


Figure 1.5

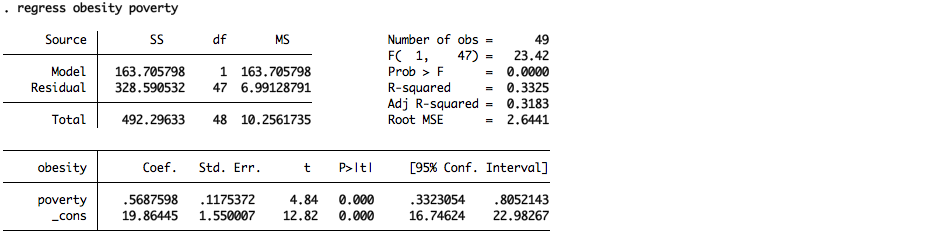
\

Figure 1.6

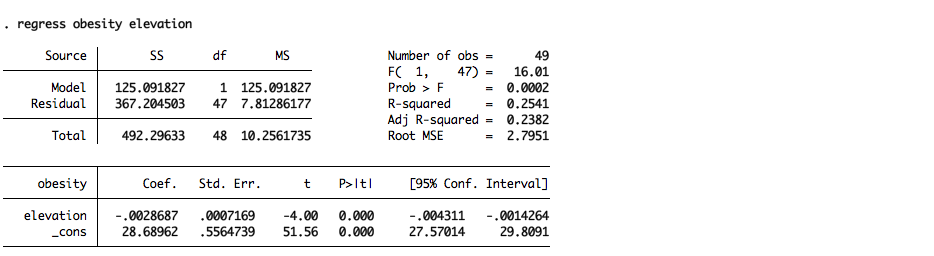


Figure 1.7

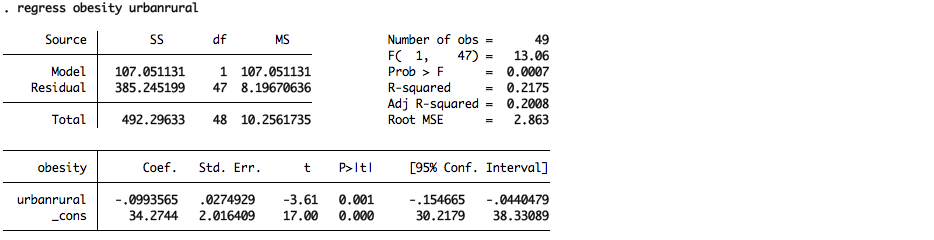


Figure 1.8

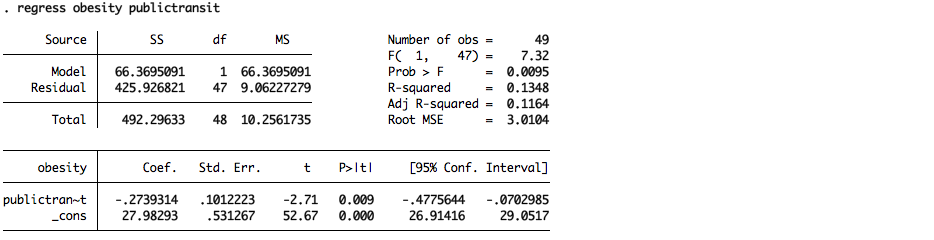


Figure 1.9

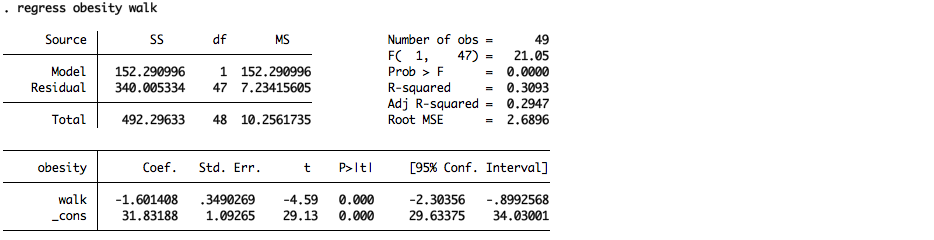
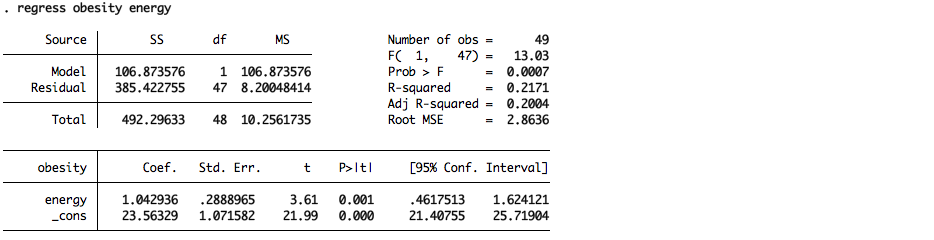


Figure 1.10



**Figures Group 2: Regressing Obesity Individually with Categorical Variables**

Figure 2.1

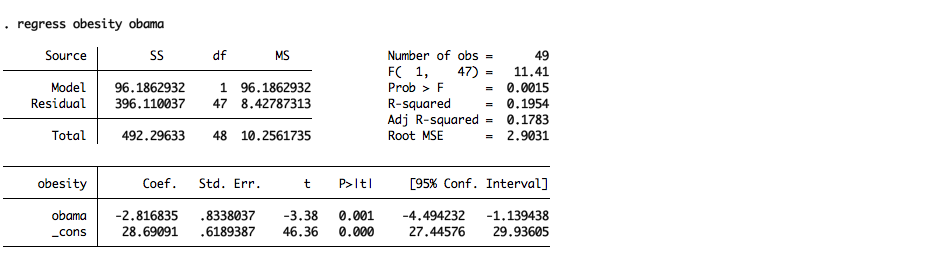


Figure 2.2

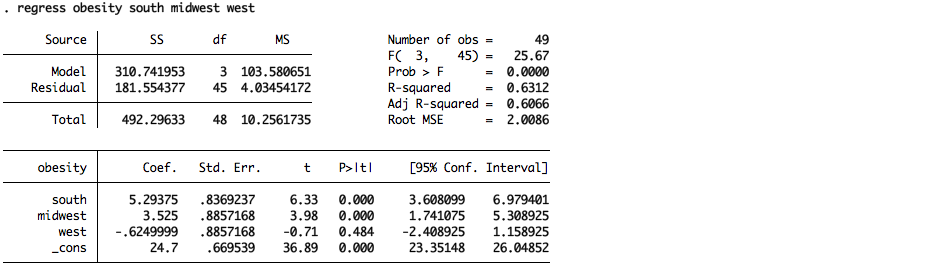
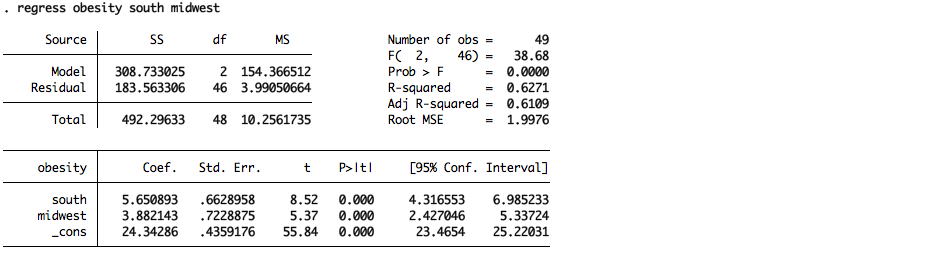


Figure 2.3



**Figure 3: Backwards Regression of Significant Variables**

Figure 3.1 Regression of Significant Variables

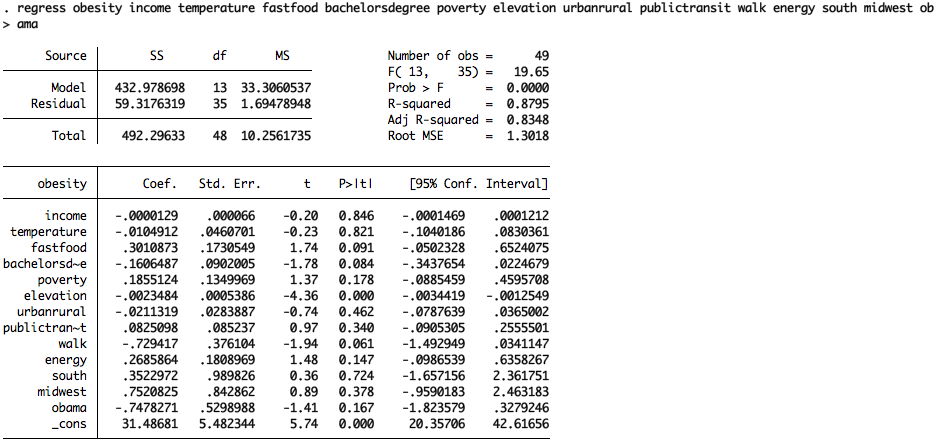


Figure 3.2 Backwards Regression Analysis gives a final model

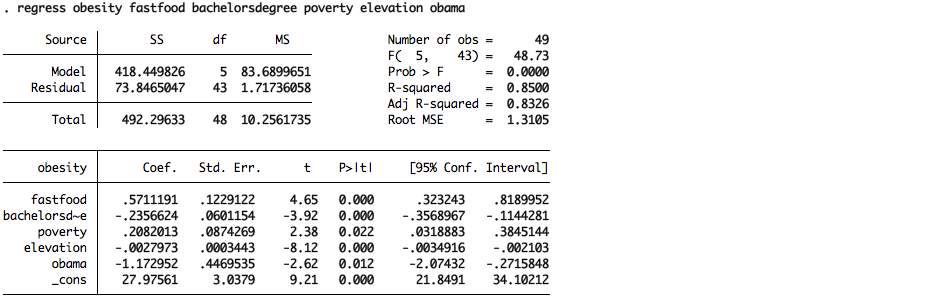
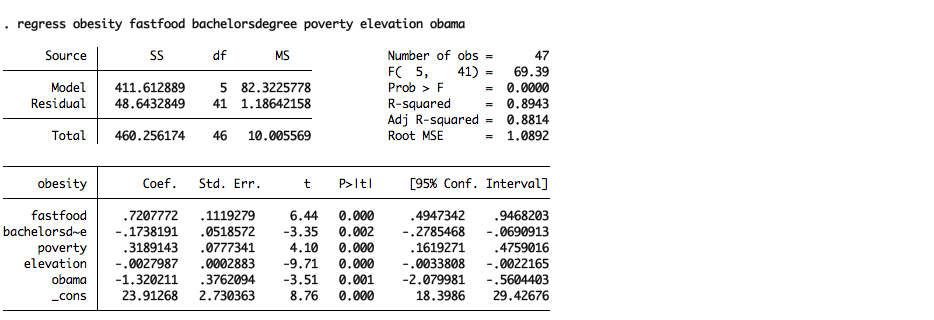
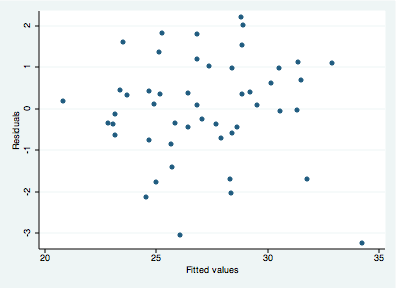


Figure 3.3 New regression model (after dropping Louisiana and Montana):



**Diagnostic (before dropping Montana and Louisiana)**

Graph 1: Scatter Plot of Residuals



Graph 2: Scatter Plot of Standardized Residuals

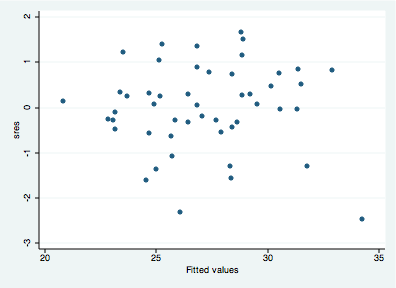


Table 1: Test for heteroskedasticity:



Table 2: Test for normal residuals:

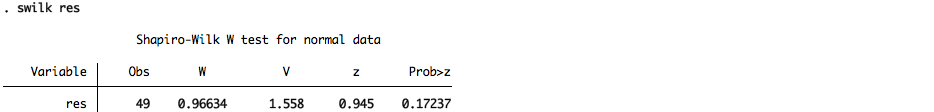
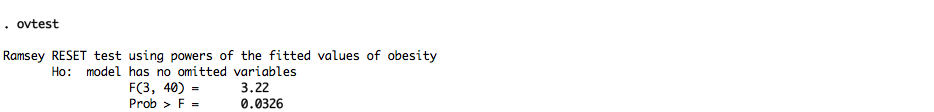


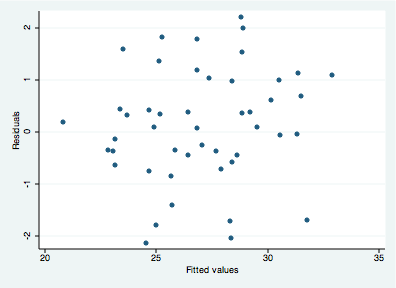
Table 3: Test for omitted variables:



**Diagnostic (before dropping Montana and Louisiana)**

Graph 1: Scatter Plot of Residuals after removing two outliers (Montana and Louisiana)





Graph 2: Scatter Plot of Standardized Residuals after removing two outliers (Montana and Louisiana)3

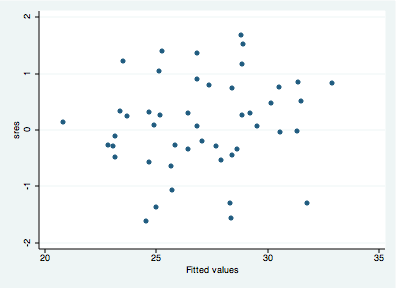


Table 1: New test for heteroskedasticity:



Table 2: New test for normal residuals:

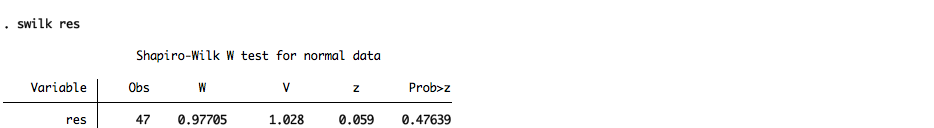


Table 3: New test for omitted variables:

