University of Nevada, Reno

Applications of Linear Regression and Other Statistical Tools in the Study of Alcohol Sales

A thesis submitted in partial fulfillment
of the requirements for the degree of

BACHELOR OF ARTS, MATHEMATICS

by

NICHOLAS BERTRANDO

Birant Ramazan, Ph.D., Thesis Advisor

May, 2013

UNIVERSITY OF NEVADA RENO

THE HONORS PROGRAM

We recommend that the thesis
prepared under our supervision by

NICHOLAS BERTRANDO

entitled

Applications of Linear Regression and Other Statistical Tools in the Study of Alcohol Sales

be accepted in partial fulfillment of the
requirements for the degree of

BACHELOR OF ARTS, MATHEMATICS

_____

Birant Ramazan, Ph.D., Thesis Advisor

_____

Tamara Valentine, Ph.D., Director, Honors Program

May, 2013

Abstract

This study uses linear regression to model the relationship of alcohol sales as a function of economic and social variables. Data were obtained on the whole U.S. for this study's dependent variable of alcohol sales and independent variables of unemployment rate, personal consumption expenditures, consumer price index, population, and high school graduation rate. For the purposes of developing a reliable regression model, this study focuses on satisfying the seven classical assumptions of ordinary least squares regression. The results of this study show a statistically significant positive relationship between alcohol sales and the variables of unemployment rate, personal consumption expenditures, and high school graduation rate.

Table of Contents

List of Figures

1. Introduction/ Literature Review

Studies in the past, which will be discussed in this section, have attempted to create different mathematical models to show a relationship between alcohol related variables and other economic and social variables, and not all have been successful. Furthermore, past studies have not analyzed alcohol sales on a nationwide scale or used data that encompass the entire U.S.

A study conducted in Finland by Luoto, Poikolainen, and Uutela analyzes alcohol use in relation to unemployment, education, marital status, and sex among individuals in Finland during two time periods, one of high unemployment and one of low unemployment. The study uses a total of 44,391 respondents aged 18 to 64 years old from 1982 to 1995. Using univariate analysis, the study finds that unemployment is related to alcohol use, but when the study uses logistic regression to analyze the relation between alcohol consumption, unemployment, education, and marital status, the previous results change. During the times of low unemployment, unemployment is not found to be associated with high levels of alcohol consumption. In times of high unemployment, single people are the only group that shows a relationship between being unemployed and consuming high levels of alcohol. More specifically, poorly-educated, single, unemployed men and highly-educated, single, unemployed women are more likely to consume high levels of alcohol as compared to other groups within their sex. The conclusion is that unemployment is weakly but significantly associated with a higher consumption of alcohol among single people during economic recession (time of high

unemployment), but not during economic expansion or time of low unemployment (Luoto, Poikolainen, and Uutela 623-29).

Another study by Waller, Zhu, Gotway, Gorman, and Gruenewald analyzes the relationship between alcohol sales and violent crime. The study reviews and contrasts two methods for modeling associative factors whose impacts on the dependent variable vary throughout geographic space. One of these methods uses Poisson "geographically weighted regression" (GWR) models, which allow covariate effects to vary in space. The other method uses "variable coefficient" models, which allow varying effects through spatial random fields, but they are more computationally involved. The study compares the two methods with respect to conceptual structures, computational implementation, and inferential output. This study analyses violent crime, illegal drug arrest, and alcohol distribution data in Houston, Texas and compares the results obtained by using the two methods described. Local rates of violent crime are used as the dependent variable or outcome, and local alcohol sales and local illegal drug activities are the two covariates of interest. The study concludes that GWR provides a quicker descriptive result, and it provides maps of smoothed general differences in association. The GWR model is found to be somewhat limited with respect to statistical inference about the amount and extent of spatial pattern. The random effects spatially varying coefficient models provide a wider inferential basis for statistical analysis of spatial pattern, but these models are more computationally intensive. The conclusion is that neither model is necessarily better for estimating the relationship between alcohol and violence than the other, but each model shows that there is significant importance in using spatial analysis. The results obtained by this study show that spatial analysis can vastly improve regression models in the study

of the relationship between alcohol and violence (Waller, Zhu, Gotway, Gorman, and Gruenewald 573-88).

Alcohol abuse is found to have a statistically significant positive effect on the likelihood of being unemployed according to a study published in 2002 by Joseph V. Terza. The study uses data from the 1988 Alcohol Supplement of the National Health Interview Survey. Terza performs regression analysis using several variables to account for endogeneity and nonlinearity in the variables. The method of regression that is used also allows for the likely possibility that alcohol abuse effects are heterogeneous concerning the observed and unobserved characteristics of people in the population. The study accomplishes this by computing alcohol abuse effects for two very different subgroups within the population. The results show a large difference in the two subgroups. More specifically, the effect of alcohol abuse on one subgroup is over three times more on the other subgroup. This difference illustrates the potential importance for recognizing heterogeneity (Terza 393-404).

Another study by Susan L. Ettner uses data from the same 1988 Alcohol Supplement of the National Health Interview Survey and looks at whether unemployment affects alcohol abuse. The study uses two-stage instrumental variables methods and looks at the effects of both non-employment and involuntary unemployment. Non-employment is found to significantly reduce both alcohol consumption and alcohol dependence symptoms. According to Ettner, the effects of non-employment are most likely because of an income effect. Involuntary unemployment is found to increase alcohol consumption, but it is also found to decrease alcohol dependence symptoms among single respondents (Ettner 251-60).

I attempt to determine if alcohol sales on a nationwide scale can be effectively modeled using linear regression techniques. I look at the economic and social variables of the national unemployment rate, high school graduation rate, personal consumption expenditures, population, and consumer price index. These data are chosen based on availability. I use the program STATA with these data to develop a linear regression model, which models alcohol sales in the U.S.

I expect to find a statistically significant relationship between alcohol sales and the economic and social variables and be able to model that relationship by linear regression. The intent is to create a unique mathematical model, which relates alcohol sales to other variables on a nationwide scale. My study will have significance from a mathematical as well as an economic perspective.

2. Linear Regression and Statistical Theory

To develop the linear regression model for alcohol sales, the program Stata 10 was used. Stata is a general purpose statistical software program developed by StataCorp. Stata's capabilities of data management and statistical analysis provide the necessary tools for developing a linear regression model for alcohol sales as a function of multiple variables. In addition to being able to create regression models, Stata contains built-in tests to check for serial correlation, stationarity, and multicollinearity. The tests used in this regression analysis include the Breusch-Godfrey test for serial correlation, the Dickey-Fuller test for stationarity, and the variance-inflating factor test for multicollinearity. Data can be easily imported from Microsoft Excel into Stata, which is what is done in this project.

Ordinary least squares (OLS) regression is used in this study as the basis of forming the model for alcohol consumption. This method predicts a linear approximation of the relationship among the data by minimizing the sum of squared vertical distances between the observed values in the datasets and the values predicted by the linear approximation. The resulting equation is expressed as a simple linear formula in n dimensions according to the number n of independent variables, as shown in Eqn. 1:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon$$

Eqn.1: Multivariable regression formula.

In this example, Y is the dependent variable, X_1, X_2,…, X_n are the independent variables, β_0, β_1,…, β_n are constants, and ε is the error term. β_1 through β_n represent the slope of the regression line with respect to each independent variable X_1, X_2,…, X_n. The error term ε

accounts for other independent variables not in the regression, mis-measurement of variables, random noise, and incorrect functional form (Gujarati and Porter 55-61). In Stata, OLS is the default regression, which makes OLS regression a good starting point for developing the final regression model.

To better understand the significance or reliability of a regression, it is important to be aware of certain components of the regression. The overall goodness of fit of the regression is measured by the coefficient of determination R^2. The coefficient of determination is well known in the statistical community, so for the purposes of this project, the mathematics will not be described. It suffices to know that R^2 represents what proportion of the variation in the dependent variable is explained by the independent variable(s). The value of R^2 is always between 0 and 1. A value closer to 1 indicates a better fit (Gujarati and Porter 73-76). Standard errors, a term that is important to regression analysis, measure the precision of the estimated β coefficients (Gujarati and Porter 84). T-statistics are derived from standard errors based on the null hypotheses that all the coefficients are equal to zero. The t-stats are calculated by the formula shown in Eqn. 2, where $\beta_{true}=0$ and $se(\beta_{est.})$ is the standard error of the estimated coefficient:

$$t = \frac{\beta_{est.} - \beta_{true}}{se(\beta_{est.})}$$

Eqn. 2: t-stat formula.

The t-stats are then used in the t test to determine whether the null hypothesis H_0: $\beta_{true}=0$ is accepted or rejected. From the t test, a p value is obtained, which determines whether to accept or reject H_0. P values will be between 0 an1, with lower values indicating greater statistical significance. A p value that is less than 0.05 indicates the null

hypothesis is rejected at the 5% level, or that the estimated coefficient is significant at the 5% level. It is standard practice to consider statistical significance at the 5% level; however, statistical significance can easily be set at different levels. For example, a p value of less than 0.01 indicates significance at the 1% level (Gujarati and Porter 115-116).

In order for a regression to be considered reliable, seven classical assumptions must hold. The first assumption is that the model is linear in the parameters, but the model does not necessarily have to be linear in the variables. The second assumption is that independent variables are not correlated with the error term (Gujarati and Porter 61-69). The violation of this assumption is known as endogeneity, which can cause biased coefficients. The common reason for this bias is the omission of a relevant variable from the regression. For example, consider the following equations:

True Model **(Eqn. 3)** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon^*$

Estimated model **(Eqn. 4)** $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Then **(Eqn. 5)** $\varepsilon = \beta_2 X_2 + \varepsilon^*$

If x_1 and x_2 are correlated, then ε and x_1 are correlated, which violates the second assumption (Nichols). The third assumption is that the error term or residuals have a zero mean. This assumption is essentially another way of saying there is no specification error in the chosen regression model. A violation of the third assumption includes modeling a dataset as linear when in reality the dataset is something else such as logarithmic or quadratic. The fourth assumption is that there is homoscedasticity or constant variance in the residuals. A violation of the fourth assumption is called heteroscedasticity, and it occurs when the variance of the data values around a regression line change as x varies.

Heteroscedasticity is most common in cross-sectional data and causes inefficient estimates. The fifth assumption is the absence of auto- or serial correlation, which means that the residuals are not correlated with each other. Serial correlation, to be discussed in more detail, is common in time series data, and it is one of the focuses in the regression for my study. Serial correlation generally produces inefficient, but unbiased estimates. The sixth assumption is the number of observations must be greater than the number of parameters to be estimated, which is somewhat obvious. Finally, the seventh assumption is the existence of variability in the independent variables, or the values for a given variable must not all be the same. Furthermore, there must not be significant outliers in each variable, as these outliers will overly influence the regression. It is very difficult to completely satisfy each assumption, but the seven assumptions serve as an important guideline when conducting regression analysis (Gujarati and Porter 61-69).

All of the data used are time series data. The use of time series creates several issues in creating a reliable linear regression model, one of which is serial correlation. Serial correlation, also known as autocorrelation, is a correlation of the error terms in a regression. For the purposes of this regression analysis, the term is defined as correlation between members of series of observations ordered in time. Serial correlation is prevalent among many time series data, and especially among economic data, which were used in this project. An explanation for serial correlation in economic data is inertia. Economic time series data, such as GDP, unemployment, and price indexes, exhibit business cycles. Starting from the trough of a recession, as recovery begins, these series start to move in a positive direction. As recovery grows, the values in the series increase by more and more until something, such as higher taxes, slows them down. In this way, the time series have

a certain inertia built into them. This inertia dramatically increases the likelihood of each successive observation to be interdependent. Other factors like the omission of a relevant variable (endogeneity) and the use of an incorrect functional form may also cause serial correlation. Nonstationarity in the variables, to be discussed later, is another possible contributor to this problem (Gujarati and Porter 413-418).

To check for the presence of serial correlation, there are several tests that can be used. For this study, the Breusch-Godfrey test was implemented. To illustrate how this test works, consider the following regression, Eqn. 6:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

Eqn. 6: Simple two-variable regression.

Assume the error term ε_t follows the pth order autoregressive scheme, as shown in Eqn. 7:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \cdots + \rho_p \varepsilon_{t-p}$$

Eqn. 7: pth order autoregressive scheme.

The null hypothesis H_0 to be tested is shown in Eqn. 8:

$$H_0: \rho_1 = \rho_2 = \ldots = \rho_p = 0$$

Eqn. 8: Null hypothesis of the Breusch-Godfrey test

Eqn. 8 means that there is no serial correlation of any order. If we reject the null hypothesis H_0, at least one ρ is statistically significantly different from zero, and there exists serial correlation (Gujarati and Porter 438-439).

The method used to correct for serial correlation in this study involves using Newey-West standard errors, also known as HAC (heteroscedasticity- and autocorrelation-consistent) standard errors. The Newey-West method uses OLS

regression, but it corrects the standard errors to account for serial correlation. Newey-West regression is only valid for use in large samples, and it is not appropriate for small ones. One very significant advantage to using Newey-West standard errors is that they not only corrects serial correlation, but they also correct heteroscedasticity. The mathematics are very involved and not necessarily that important to understanding this method, so the math behind Newey-West regression will not be discussed in detail. What is important to know, however, is that Newey-West regression will produce the same estimated coefficients and R^2 value as normal OLS, but Newey-West regression will produce much greater standard errors than OLS, which makes statistical significance harder to achieve (Gujarati and Porter 447-448).

Another important factor in making a reliable regression model is checking for stationarity. Stationary time series have constant mean and variance, which is needed to create a reliable regression model. Nonstationary variables, also known as random walk or unit root variables, can cause spurious regression results and t-stats not to follow t-distributions. An example of a pure random walk is shown in Eqn. 9:

$$Y_t = Y_{t-1} + u_t \text{ where } u_t \text{ is a white noise error term}$$

Eqn. 9: Pure random walk.

There are a few different versions of this random walk, which expands upon the pure random walk. We can have a random walk with drift, as shown in Eqn. 10:

$$Y_t = \beta_1 + Y_{t-1} + u_t$$

Eqn. 10: Random walk with drift.

A random walk with a deterministic trend can also exist, as shown in Eqn. 11:

$$Y_t = \beta_1 + \beta_2 t + Y_{t-1} + u_t$$

Eqn. 11: Random walk with a deterministic trend.

Equations 9, 10, and 11 are all examples of nonstationary variables. The phenomenon of spurious regression causes an apparently strong statistical relationship between variables, but in reality there may exist no relationship at all. Therefore, it is very important to test each variable to ensure the variable is stationary and correct for nonstationarity as needed (Gujarati and Porter 740-748).

There are many tests that check for stationarity, but this study uses the Dickey-Fuller test. The Dickey-Fuller test involves several decisions to determine whether a series is stationary. The null hypothesis in this test is that the time series is nonstationary. To account for the three forms of random walks described earlier, the Dickey-Fuller test is estimated in three different forms. If the null hypothesis is rejected, the time series is stationary (Gujarati and Porter 755-756).

If a time series variable is found to be nonstationary, it still may be possible to create a regression without spurious results. Two or more variables may share the same common trend, which avoids the problem of spurious results. Although two or more variables may be nonstationary on their own, their linear combination may still be stationary. If this characteristic is true, the variables are said to be cointegrated and will not produce spurious results when regressed together. From an economic perspective, two variables will likely be cointegrated if they have a long term relationship between them. Regressing cointegrated variables will still produce a model that is considered reliable (Gujarati and Porter 762).

To check for cointegration, it is common to use the simple method of applying the Dickey-Fuller test to the residuals of the regression between the nonstationary variables. Once the residuals are obtained, the process is the same as checking a single time series for stationarity (Gujarati and Porter 763).

When dealing with time series regressions, problems like serial correlation and stationarity are among the most important. Multicollinearity is a smaller problem, but it may still affect the results of the regression, so it should not be ignored. Multicollinearity is a high correlation among the variables, which causes an increase in standard errors that may reduce the statistical significance of a variable's coefficient. In perfect multicollinearity, the variables are perfect linear combinations of each other. We can express multicollinearity in mathematical terms, as shown in Eqn. 12:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + v_i = 0$$

Eqn. 12: Multicollinearity.

where v_i is a stochastic error term (Gujarati and Porter 321).

To detect the presence of multicollinearity, a few different tests can be performed. A very simple technique is to generate a correlation matrix. This matrix will give correlations from -1 to 1 between each variable, -1 being perfect negative correlation and 1 being perfect positive correlation. A correlation with an absolute value above 0.8 is considered cause for concern. A correlation from -1 to -0.8 indicates strong negative correlation and from 0.8 to 1 indicates strong positive correlation (Nichols). Another way to test for multicollinearity is to calculate the variance-inflating factor (VIF). The VIF shows how the variance of an estimator is inflated as a result of multicollinearity. To understand the calculation of the VIF, suppose our regression is (Eqn. 13):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon$$

Eqn. 13: Sample regression.

Then the first step would be to regress each independent variable X_i as a function of all other independent variables. Then the VIF can be calculated for each X_i. For example, if i=1 (Eqn. 14):

$$X_1 = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + \cdots + \alpha_{n-1} X_n$$

Eqn. 14: Regression of X_1 on all other indep. variables.

The VIF for X_1 is then calculated as follows in Eqn. 15:

$$VIF_{X_1} = \frac{1}{(1 - {R_1}^2)}$$

Eqn. 15: VIF calculation.

where ${R_1}^2$ is the coefficient of determination from the regression of X_1 (Gujarati and Porter 328). As a general rule of thumb, a VIF greater than 5 indicates there exists significant signs multicollinearity in the original regression. Multicollinearity can be remedied by removing one or more of the variables, which are strongly correlated with other variables in the regression (Nichols).

3. Methods Used in the Study of Alcohol Sales

I have collected data on alcohol sales, unemployment rates, personal consumption expenditures, consumer price indexes, population, and high school graduation rates. The data on alcohol sales are seasonally adjusted, nationwide (total U.S. sales), monthly (data value for each month) from January 1995 to December 2012, and in units of millions of dollars. More specifically, these data are the total U.S. sales for "beer, wine, and liquor stores" from the U.S. Census Bureau website. The data on unemployment rates, personal consumption expenditures, and consumer price indexes are also seasonally adjusted, nationwide, and monthly from January 1995 to December 2012. The data are in units of percentage, billions of dollars, and index (with 1982-84=100) respectively and collected from the U.S. Federal Reserve Bank of St. Louis website. The data on population are monthly from January 1995 to December 2012 and in units of thousands of people. These population data were collected from the U.S. Bureau of Economic Analysis website (United States). The data for high school graduation rates are nationwide and yearly from 1990 to 2009 with missing values for the years 1991, 1992,1993,1994, and 2007 (Preparation for College). Each dataset used, shown in the appendix, yields a total of 216 observations, which account for every month over a total of 18 years.

The selection of these data is based on previous studies discussed in the literature review and the availability of data. Monthly data are used to allow for the most data points possible, so that the accuracy of the regression could be maximized. I hypothesize that the graduation rate from a certain year affects the sale of alcohol three years later. High school students are most commonly 18 years old when they graduate and are 21

years old, which is the legal U.S. drinking age, three years later. To account for this lag, values for graduation rates are taken from 1992 to 2009 (all other data are from 1995 to 2012). Missing data for 1992, 1993, and 1994 were created using the linear trend between 1990 and 1995. The missing value for 2007 was created by taking the average value between 2006 and 2008. To allow the high school graduation data to be regressed with the other data, it was necessary to generalize each year's graduation rate to every month of that year.

Using the computer program Stata 10, several different regressions are analyzed. A preliminary linear regression is performed using alcohol sales as the dependent variable and unemployment rate, personal consumption expenditures, consumer price index, population, and high school graduation rate as the independent variables. The results are shown in Fig. 1:

```
. regress AlcSales Urate pce cpi Population GradRate
```

Source	SS	df	MS		
Model	74288817	5	14857763.4	Number of obs =	216
Residual	484363.588	210	2306.49327	F(5, 210) =	6441.71
				Prob > F =	0.0000
Total	74773180.6	215	347782.235	R-squared =	0.9935
				Adj R-squared =	0.9934
				Root MSE =	48.026

AlcSales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Urate	-8.125275	4.599425	-1.77	0.079	-17.19224	.9416866
pce	-.1357915	.0390592	-3.48	0.001	-.2127899	-.0587931
cpi	18.18213	2.454694	7.41	0.000	13.34313	23.02113
Population	.0275893	.0041633	6.63	0.000	.019382	.0357966
GradRate	36.55653	5.610898	6.52	0.000	25.49562	47.61743
_cons	-10127	1211.399	-8.36	0.000	-12515.06	-7738.938

Fig. 1: Alcohol sales as a function of unemployment rate, personal consumption expenditures, consumer price index, population, and high school graduation rate.

The results of this preliminary regression show a strong relationship between alcohol sales and the previously mentioned variables, as evidenced by an extremely high R^2 value of .9935. In addition, all of the coefficients are statistically significant at the 5% level, with the exception of the unemployment rate coefficient. This significance is shown by the values which fall under "P>|t|", which are the p values. All of these values are well below 0.05, with the exception of the unemployment coefficient, which is 0.079. There are a few apparent problems with this regression. The coefficients of unemployment and personal consumption expenditures are both negative; as unemployment and personal consumption go down, alcohol sales go up. These results are both counterintuitive and opposite to the findings of previous studies in the case of unemployment (Ettner 251-60, Luoto, Poikolainen, and Uutela 623-29, Terza 393-404). The negative coefficient for personal consumption is especially counterintuitive since one would expect the sales of most goods, alcohol included, to increase as people across the country spend more money. In the case of graduation rates, a positive value in the coefficient is counterintuitive since one would guess more people graduating from high school would lead to lower levels of alcohol abuse and consumption; however, the study conducted by Luoto, Poikolainen, and Uutela in Finland did find mixed effects of education on alcohol consumption (Luoto, Poikolainen, and Uutela 623-29). The cause of these problems is unclear after looking at the regression, and the cause is to be diagnosed after running further tests. On the issue of the statistically insignificant coefficient value for unemployment, multicolinearity is a definite possibility. Since the data are time series, there is a strong chance of serial correlation or autocorrelation. This problem causes inaccurate t-statistics, which changes the statistical significance of coefficients. To check

for this, the Breusch-Godfrey Test for first order serial correlation is used (Nichols). The results of this test are shown in Fig.2:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

lags(*p*)	chi2	df	Prob > chi2
1	**98.601**	**1**	**0.0000**

H0: no serial correlation

Fig. 2: Results of Breusch-Godfrey test on preliminary regression.

The results of this test indicate that there is, in fact, a strong presence of serial correlation. The presence of serial correlation is indicated by "Prob > chi^2 = 0.0000," which is a p value. Because the p value is less than 0.05, the null hypothesis that there is no serial correlation is rejected; there exists serial correlation (Nichols). Stationarity is also a possible problem. Because this problem is especially common in economic time series variables, as are used in this regression, there may be some spurious regression results (Gujarati and Porter 740-748). All problems mentioned are to be addressed to create a reliable regression.

In trying to create a better regression model, a trend variable "time" was included. In some cases, a trend variable can remedy the problem of serial correlation, and it can affect the estimated coefficients for the other independent variables. The time variable is the data set 1,2,3,…,216 since there are 216 data values for each of the other variables. This time variable is designated in Stata as the trend variable (Nichols). The regression with "time" included is shown in Fig. 3:

```
. regress AlcSales Urate pce cpi Population GradRate time
```

Source	SS	df	MS		
Model	74289185.9	6	12381531	Number of obs	216
Residual	483994.759	209	2315.7644	F(6, 209)	5346.63
				Prob > F	0.0000
Total	74773180.6	215	347782.235	R-squared	0.9935
				Adj R-squared	0.9933
				Root MSE	48.122

AlcSales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Urate	-6.770186	5.724431	-1.18	0.238	-18.05521	4.51484
pce	-.1314504	.0406211	-3.24	0.001	-.2115301	-.0513708
cpi	18.16655	2.459933	7.38	0.000	13.31709	23.01601
Population	.0321173	.0120885	2.66	0.008	.0082863	.0559482
GradRate	38.33979	7.181581	5.34	0.000	24.18217	52.49741
time	-1.224915	3.06931	-0.40	0.690	-7.27569	4.825861
_cons	-11476.87	3593.614	-3.19	0.002	-18561.24	-4392.488

Fig. 3: Regression with trend variable "time" included.

This regression yields coefficients for the unemployment rate and time that are statistically insignificant at the 5% level, which is indicated by P values of 0.238 and 0.69 respectively. Also, despite having included a trend variable, the regression has strong signs of serial correlation when using another Breusch-Godfrey test, as shown in Fig.4:

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

lags(*p*)	chi2	df	Prob > chi2
1	98.565	1	0.0000

H0: no serial correlation

Fig.4: Results of Breusch-Godfrey test on regression with trend variable included.

The p value of 0.0000, which is the same as the Breusch-Godfrey test on the preliminary regression, indicates there is still serial correlation present. Since the regression actually gets worse by including the trend variable, it may be best to leave the trend variable out.

One possible reason for the non-statistically significant coefficients, in particular on unemployment, could be from the effects of multicollinearity. Multicollinearity means there is a high correlation between independent variables that causes t-statistics to be smaller and the coefficients to not be statistically significant (Nichols). Two methods are used to check for multicollinearity. One method is to calculate the variance-inflating factor (VIF), as shown in Fig. 5:

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        time |   3416.32    0.000293
  Population |   2910.80    0.000344
         pce |    575.15    0.001739
         cpi |    332.52    0.003007
       Urate |      9.52    0.105051
    GradRate |      4.94    0.202545
-------------+----------------------
    Mean VIF |   1208.21
```

Fig. 5: Variance-inflating factor (VIF) for all independent variables.

The general rule of thumb is that a VIF greater than five, for any of the independent variables, indicates possible multicollinearity (Nichols). Since all of the VIF values are greater than five, except for the VIF of high school graduation, there is likely strong multicollinearity. To confirm and diagnose the problem, the method of analyzing a correlation matrix is used. This correlation matrix is shown in Fig. 6:

```
. correlate Urate pce cpi Population GradRate time
(obs=216)
```

	Urate	pce	cpi	Popula~n	GradRate	time
Urate	1.0000					
pce	0.6202	1.0000				
cpi	0.6743	0.9952	1.0000			
Population	0.6651	0.9952	0.9934	1.0000		
GradRate	0.3260	0.3670	0.4057	0.3227	1.0000	
time	0.6800	0.9952	0.9957	0.9992	0.3517	1.0000

Fig 6: Correlation matrix of all independent variables.

A correlation of above 0.8 is considered a cause for concern (Nichols). This matrix shows that personal consumption expenditures, consumer price index, population, and time are all heavily correlated as indicated by correlations all above .99. Because of this high multicollinearity, the decision is made to use only personal consumption expenditures and to drop population, consumer price index, and time as variables. Personal consumption is chosen over the other variables, since it intuitively accounts the most for alcohol sales because personal consumption is directly related to the sale of all goods, including alcohol. The dropping of the variables population, consumer price index, and time leaves unemployment, personal consumption, and high school graduation as the three remaining independent variables. Another regression, as seen in Fig. 7, as well as a new VIF table and correlation matrix, Fig. 8 and 9 respectively, show these changes:

```
. regress AlcSales Urate pce GradRate
```

Source	SS	df	MS		
Model	74028889.7	3	24676296.6	Number of obs	216
Residual	744290.91	212	3510.80618	F(3, 212)	7028.67
				Prob > F	0.0000
				R-squared	0.9900
				Adj R-squared	0.9899
Total	74773180.6	215	347782.235	Root MSE	59.252

AlcSales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Urate	33.55717	2.939019	11.42	0.000	27.76373	39.35062
pce	.276802	.0027268	101.51	0.000	.2714269	.282177
GradRate	29.10476	4.31752	6.74	0.000	20.59399	37.61553
_cons	-1708.775	289.4566	-5.90	0.000	-2279.357	-1138.193

Fig. 7: Regression with remaining independent variables.

```
. estat vif
```

Variable	VIF	1/VIF
pce	1.71	0.584975
Urate	1.66	0.604188
GradRate	1.18	0.849582
Mean VIF	1.51	

Fig. 8: VIF with remaining independent variables.

```
. correlate Urate pce GradRate
(obs=216)
```

	Urate	pce	GradRate
Urate	1.0000		
pce	0.6202	1.0000	
GradRate	0.3260	0.3670	1.0000

Fig. 9: Correlation matrix with remaining independent variables.

Fig. 8 and Fig. 9 show how the problem of multicolinearity has been remedied. The elimination of multicollinearity is shown by the VIF values of less than 5 in Fig. 8 and by the correlations of less than 0.8 in Fig. 9. There are a few very significant changes to the regression as well, as seen in Fig. 7. All of the coefficients are now statistically significant at the 5% level, and the coefficients for unemployment and personal consumption went from negative to positive, which agrees with my original expectations. The R^2 value is still extremely high at 0.9900, indicating a strong goodness of fit (Gujarati and Porter 73-76). Despite solving some of the previous issues, there remain some issues that have to be addressed.

Since all of the data are time series, it is important to check for stationarity. Stationary time series have constant mean and variance. Non-stationary variables cause spurious regression results (cause false statistical significance) and t-stats not to follow t-distributions. A non-stationary process is known as a random walk or unit root. To check variables for stationarity, the Dickey-Fuller test is used. The results of this test on the variables of alcohol sales, high school graduation rate, unemployment rate, and personal consumption expenditures are shown in Fig. 10, 11, 12, and 13 respectively:

```
. dfuller AlcSales
Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ———— Interpolated Dickey-Fuller ————
                    Test         1% Critical       5% Critical      10% Critical
                Statistic             Value             Value             Value
——————————————————————————————————————————————————————————————————————————————
 Z(t)              0.584            -3.472            -2.882            -2.572
——————————————————————————————————————————————————————————————————————————————
MacKinnon approximate p-value for Z(t) = 0.9872
```

Fig. 10: Dickey-Fuller test for stationarity in alcohol sales.

```
. dfuller  GradRate

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ———— Interpolated Dickey-Fuller ————
                    Test         1% Critical       5% Critical      10% Critical
                Statistic           Value             Value            Value
———————————————————————————————————————————————————————————————————————————
 Z(t)             -0.943            -3.472            -2.882           -2.572
———————————————————————————————————————————————————————————————————————————
MacKinnon approximate p-value for Z(t) = 0.7736
```

Fig. 11: Dickey-Fuller test for stationarity in high school graduation rates.

```
. dfuller  Urate

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ———— Interpolated Dickey-Fuller ————
                    Test         1% Critical       5% Critical      10% Critical
                Statistic           Value             Value            Value
———————————————————————————————————————————————————————————————————————————
 Z(t)             -0.247            -3.472            -2.882           -2.572
———————————————————————————————————————————————————————————————————————————
MacKinnon approximate p-value for Z(t) = 0.9327
```

Fig. 12: Dickey-Fuller test for stationarity in unemployment rates.

```
. dfuller  pce

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ———— Interpolated Dickey-Fuller ————
                    Test         1% Critical       5% Critical      10% Critical
                Statistic           Value             Value            Value
———————————————————————————————————————————————————————————————————————————
 Z(t)             -0.139            -3.472            -2.882           -2.572
———————————————————————————————————————————————————————————————————————————
MacKinnon approximate p-value for Z(t) = 0.9454
```

Fig. 13: Dickey-Fuller test for stationarity in personal consumption expenditures.

The null hypothesis in this test is that the variable is non-stationary, and since all the variables show p-values greater than 0.05, they are all shown to be non-stationary (Gujarati and Porter 755-756). Although each variable by itself may be non-stationary,

the variables are found to be cointegrated, or collectively stationary (Nichols). Cointegration is shown by using the Dickey-Fuller Test on the residuals of the regression in Fig. 7. These results are shown in Fig. 14:

```
. dfuller  resids

Dickey-Fuller test for unit root                    Number of obs   =       215

                                 ———— Interpolated Dickey-Fuller ————
                 Test         1% Critical       5% Critical      10% Critical
               Statistic         Value             Value            Value
———————————————————————————————————————————————————————————————————————————
 Z(t)            -4.023           -3.472            -2.882           -2.572
———————————————————————————————————————————————————————————————————————————
MacKinnon approximate p-value for Z(t) = 0.0013
```

Fig. 14: Dickey-Fuller test for stationarity in residuals of regression in Fig.7.

The p-value of 0.0013 (<0.05) indicates that the null hypothesis of non-stationarity is to be rejected and the alternative hypothesis of stationarity is accepted. This means the regression is safe from the negative effects of non-stationary variables (Nichols). This leaves serial correlation as the only remaining problem to making the regression reliable.

To resolve the issue of serial correlation, Newey-West standard errors are used in the regression of alcohol sales as a function of the unemployment rate, high school graduation rate, and personal consumption expenditures. The results of this regression are shown in Fig. 15:

```
. newey AlcSales Urate pce GradRate, lag(5)

Regression with Newey-West standard errors          Number of obs  =       216
maximum lag: 5                                      F(  3,   212)  =   2978.50
                                                    Prob > F       =    0.0000
```

AlcSales	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
Urate	33.55717	4.69971	7.14	0.000	24.29302	42.82132
pce	.276802	.0046875	59.05	0.000	.2675618	.2860422
GradRate	29.10476	9.555998	3.05	0.003	10.26781	47.94171
_cons	-1708.775	653.363	-2.62	0.010	-2996.695	-420.8544

Fig. 15: Regression using Newey-West standard errors.

The Newey-West regression gives standard errors that are robust to serial correlation, which fixes the problems associated with serial correlation. This Newey-West regression is exactly the same as the one in Fig. 7 with the exception of the standard errors, t-stats, and p values, which are all directly related. Newey-west standard errors are greater than regular OLS standard errors, making t-stats smaller, and p values larger. These standard errors make coefficients less statistically significant (Gujarati and Porter 447-448). Despite having used these Newey-West standard errors, this regression still achieves statistical significance at the 5% level. This is determined to be the best regression model for alcohol sales in the U.S. since it corrects and accounts for all of the problems previously mentioned in this section.

4. Results

This regression analysis yields several interesting results. The final regression, shown in Fig.15, models the data effectively, as evidenced by the high R^2 value of 0.9900 (same as R^2 in Fig. 7). In addition, statistical significance is achieved in all of the variable coefficients while satisfying the seven classical assumptions of OLS regression to the best of my ability.

Through analyzing this regression, several conclusions with respect to the nature of alcohol sales can be inferred. The coefficient on the unemployment rate means that for every percentage increase in unemployment, the sale of alcohol increases by 33.56 million dollars in the U.S. The coefficient on personal consumption expenditures means that for every billion dollars spent in the U.S. by consumers, alcohol sales increase by 0.2768 million dollars. And finally, the coefficient on the high school graduation rate means that for every percentage increase in the U.S. high school graduation rate, the sale of alcohol increases by 29.10 million dollars three years later (graduation rates are the only variable in which observations are not compared to the observations for alcohol sales in the same instance in time). It is also important to note that this coefficient on graduation rates is positive, rather than negative as I would expect, but this will be discussed in the next section. Not surprisingly, the variable of personal consumption expenditures has the strongest statistical relationship with alcohol sales. The t-stat of 59.05 is astronomical, which shows just how strong this relationship is. The strong relationship makes sense since if nationwide consumer spending goes up, one would expect nationwide spending on alcohol to go up as well. Also as I expected, high school

graduation rates have the weakest statistical relationship with alcohol sales, as evidenced by the t-stat of 3.05. This variable is much more experimental because the data behind it requires so much manipulation. Just achieving statistical significance with this particular variable is somewhat peculiar, and it raises some interesting questions to be discussed.

5. Significance

This regression shows that alcohol sales are effectively modeled on a nationwide scale as a function of economic and social variables. The significance of this regression model is less associated with the physical results and more associated with the statistical processes and theory that go in to obtaining those results. The mathematical theory behind many of the processes used in this study is in many cases much denser than what is presented, but it is the intention of this study to show a balance between the application and theory of these processes. This study is successful from the standpoint of showing the applications of linear regression and other statistical tools in the study of alcohol sales, but there are still some minor problems.

The results of this regression show relationships that are consistent with those found in the previous studies discussed in the literature review. The idea of unemployment having a positive relationship with alcohol sales or consumption is supported by this study, as well as in the studies of Ettner; Luoto, Poikolainen, and Uutela; and Terza. The results of the high school graduate rate coefficient is also somewhat supported by the study of Luoto, Poikolainen, and Uutela. This study finds a positive relationship between high school graduation and alcohol sales, and their study finds a positive relationship between highly-educated, single women and alcohol consumption (Luoto, Poikolainen, and Uutela 623-29). These results are interesting, since one would expect alcohol sales or consumption to decline among an educated population. All of the studies mentioned in the literature review, as well as this study, show the many different ways to use alcohol sales or consumption in mathematical modeling.

The largest problem, or obstacle, when creating a reliable model for alcohol sales in the U.S. relates to the availability of data. General economic data like those on unemployment, consumer spending, price index, and population is relatively easy to come by; however, social variables are much harder to find. A good example of the unavailability of data is shown in the high school graduation rates used in this study. I was unable to find high school graduation rates from every year that is used in this study, so I created a few artificial values based on apparent trends. This certainly dissolves some significance from the results of the regression model, even though the mathematical theory behind the regression is the same had all the values been available. The legitimacy of taking yearly data and using the same value over each month in a certain year to create monthly data is questionable. But, as in the case of the graduation rates, it was necessary to be able to include the variable in the regression. It also would be interesting to use other social variables such as divorce rates and crime rates in the study, but I was unable to find reliable data for these variables. Being able to use other social variables would offer more interesting comparisons between some of the previous studies discussed in the literature review. Nevertheless, the study still presents linear regression in an interesting way.

Studying regional data, rather than just data of the whole U.S., would have also given this study an interesting new dimension. There are several ways to study regional data, so there are many options to go about incorporating this into a similar study. One such study by Waller, Zhu, Gotway, Gorman, and Gruenewald discussed in the literature review shows clearly how geographic location in regards to alcohol consumption can be analyzed. The relationship from state-to-state or even country-to-country, in regards to

alcohol consumption, could be examined. Many states have different laws with regards to alcohol, so one could study the effect these laws have across the U.S. The major set back to studying regional data is obtaining it; nevertheless, if the data were made more available, then it would certainly lead to new and exciting studies.

This study on alcohol sales not only uses many mathematical and statistical methods and concepts, but it is also uses concepts very closely related to economics. More specifically, this study falls under the subject of econometrics, which is a combination of economics, statistics, and applied mathematics. Although this study has some weaknesses, as previously described, it can be used as a reference point for future econometric studies, which incorporate more social variables or even regional variables. This study presents many interesting and useful applications of linear regression and other statistical methods.

Works Cited

Ettner, Susan L. "Measuring the Human Cost of a Weak Economy: Does Unemployment Lead to Alcohol Abuse?" *Social Science & Medicine* 44.2 (1997): 251-60. Print.

Gujarati, Damodar N., and Dawn C. Porter. *Basic Econometrics*. Boston: McGraw-Hill Irwin, 2009. Print.

Luoto, Riitta, Kari Poikolainen, and Antti Uutela. "Unemployment, Sociodemographic Background and Consumption of Alcohol before and during the Economic Recession of the 1990s in Finland." *International Journal of Epidemiology* 27 (1998): 623-29. Print.

Nichols, Mark. Econ 441: Introduction to Econometrics. University of Nevada, Reno. July 2012. Lecture.

"Preparation for College: Public High School Graduation Rates." *The National Center for Higher Education Management Systems*. N.p., n.d. Web. 12 Feb. 2013. *Stata.com*. N.p., n.d. Web. 06 Apr.2013.

Terza, Joseph V. "Alcohol Abuse and Employment: A Second Look." *Journal of Applied Econometrics* 17.4 (2002): 393-404. Print.

United States. *Bureau of Economic Analysis (BEA)*. N.p., n.d. Web. 06 Apr. 2013. <http://www.bea.gov/>.

United States. Census Bureau. *Census.gov*, n.d. Web. 10 Feb. 2013.

United States. "Federal Reserve Economic Data." *Federal Reserve Bank of St. Louis*. N.p., n.d. Web. 10 Feb. 2013.

Waller, Lance A., Li Zhu, Carol A. Gotway, Dennis M. Gorman, and Paul J. Gruenewald."Quantifying Geographic Variations in Associations between Alcohol Distribution and Violence: A Comparison of Geographically Weighted Regression and Spatially Varying Coefficient Models." *Stochastic Environmental Research and Risk Assessment* 21.5 (2007): 573-88. Print.

Appendix

	Alcohol Sales	PCE	Population	CPI	Unemployment Rate	High School Grad Rate (Modified)
Units	$ millions	$ billions	thousands	Index 1982-84=100	Percent	Percent
Source	(United States. Census)	(United States. Federal)	(United States. Bureau)	(United States. Federal)	(United States. Federal)	(Preparation for College)
Date						
Jan-1995	1832	4878.4	265157	150.5	5.6	69.47
Feb-1995	1802	4877.0	265383	150.9	5.4	69.47
Mar-1995	1810	4910.6	265625	151.2	5.4	69.47
Apr-1995	1817	4914.5	265877	151.8	5.8	69.47
May-1995	1809	4956.8	266134	152.1	5.6	69.47
Jun-1995	1818	5001.2	266414	152.4	5.6	69.47
Jul-1995	1808	4994.2	266700	152.6	5.7	69.47
Aug-1995	1830	5029.6	266998	152.9	5.7	69.47
Sep-1995	1877	5044.9	267304	153.1	5.6	69.47
Oct-1995	1866	5038.1	267585	153.5	5.5	69.47
Nov-1995	1859	5079.9	267829	153.7	5.6	69.47
Dec-1995	1888	5122.2	268047	153.9	5.6	69.47
Jan-1996	1950	5111.1	268258	154.7	5.6	69.12
Feb-1996	1908	5158.7	268480	155.0	5.5	69.12
Mar-1996	1930	5199.7	268724	155.5	5.5	69.12
Apr-1996	1931	5235.6	268980	156.1	5.6	69.12
May-1996	1907	5251.9	269247	156.4	5.6	69.12
Jun-1996	1954	5258.9	269527	156.7	5.3	69.12
Jul-1996	1960	5281.4	269822	157.0	5.5	69.12
Aug-1996	1955	5305.1	270130	157.2	5.1	69.12
Sep-1996	1917	5326.8	270433	157.7	5.2	69.12
Oct-1996	1902	5359.2	270730	158.2	5.2	69.12
Nov-1996	1914	5383.6	271002	158.7	5.4	69.12
Dec-1996	1904	5411.3	271243	159.1	5.4	69.12
Jan-1997	1939	5445.2	271472	159.4	5.3	68.77
Feb-1997	1939	5468.3	271703	159.7	5.2	68.77
Mar-1997	1973	5487.7	271952	159.8	5.2	68.77
Apr-1997	1962	5491.9	272213	159.9	5.1	68.77
May-1997	1977	5493.0	272482	159.9	4.9	68.77
Jun-1997	2039	5527.0	272767	160.2	5.0	68.77
Jul-1997	2017	5582.5	273074	160.4	4.9	68.77
Aug-1997	2009	5621.7	273395	160.8	4.8	68.77
Sep-1997	2024	5635.8	273703	161.2	4.9	68.77
Oct-1997	2056	5671.5	273989	161.5	4.7	68.77

Nov-1997	2055	5695.7	274249	161.7	4.6	68.77
Dec-1997	2055	5727.2	274499	161.8	4.7	68.77
Jan-1998	2072	5725.6	274732	162.0	4.6	68.62
Feb-1998	2112	5760.0	274943	162.0	4.6	68.62
Mar-1998	2074	5786.9	275175	162.0	4.7	68.62
Apr-1998	2095	5825.1	275434	162.2	4.3	68.62
May-1998	2092	5875.7	275700	162.6	4.4	68.62
Jun-1998	2076	5909.9	275976	162.8	4.5	68.62
Jul-1998	2094	5928.9	276266	163.2	4.5	68.62
Aug-1998	2135	5965.9	276566	163.4	4.5	68.62
Sep-1998	2133	6009.3	276859	163.5	4.6	68.62
Oct-1998	2137	6046.9	277140	163.9	4.5	68.62
Nov-1998	2152	6065.9	277402	164.1	4.4	68.62
Dec-1998	2169	6121.9	277658	164.4	4.4	68.62
Jan-1999	2145	6125.5	277891	164.7	4.3	68.06
Feb-1999	2158	6155.6	278095	164.7	4.4	68.06
Mar-1999	2145	6191.2	278324	164.8	4.2	68.06
Apr-1999	2219	6258.0	278584	165.9	4.3	68.06
May-1999	2201	6290.8	278859	166.0	4.2	68.06
Jun-1999	2167	6321.0	279148	166.0	4.3	68.06
Jul-1999	2190	6351.8	279448	166.7	4.3	68.06
Aug-1999	2193	6396.9	279752	167.1	4.2	68.06
Sep-1999	2229	6447.9	280053	167.8	4.2	68.06
Oct-1999	2235	6467.7	280337	168.1	4.1	68.06
Nov-1999	2231	6503.7	280594	168.4	4.1	68.06
Dec-1999	2267	6603.1	280846	168.8	4.0	68.06
Jan-2000	2259	6602.8	281083	169.3	4.0	67.15
Feb-2000	2290	6689.2	281299	170.0	4.1	67.15
Mar-2000	2327	6757.0	281531	171.0	4.0	67.15
Apr-2000	2297	6740.1	281763	170.9	3.8	67.15
May-2000	2355	6775.6	281996	171.2	4.0	67.15
Jun-2000	2384	6811.2	282247	172.2	4.0	67.15
Jul-2000	2401	6832.2	282504	172.7	4.0	67.15
Aug-2000	2417	6864.7	282769	172.7	4.1	67.15
Sep-2000	2416	6948.3	283033	173.6	3.9	67.15
Oct-2000	2450	6955.4	283285	173.9	3.9	67.15
Nov-2000	2475	6970.7	283523	174.2	3.9	67.15
Dec-2000	2344	7017.2	283748	174.6	3.9	67.15
Jan-2001	2496	7047.7	283960	175.6	4.2	67.76
Feb-2001	2457	7066.5	284166	176.0	4.2	67.76
Mar-2001	2444	7060.1	284380	176.1	4.3	67.76

Apr-2001	2444	7079.9	284602	176.4	4.4	67.76
May-2001	2455	7130.8	284834	177.3	4.3	67.76
Jun-2001	2474	7145.3	285076	177.7	4.5	67.76
Jul-2001	2458	7156.5	285324	177.4	4.6	67.76
Aug-2001	2454	7195.8	285584	177.4	4.9	67.76
Sep-2001	2451	7101.4	285842	178.1	5.0	67.76
Oct-2001	2468	7298.6	286086	177.6	5.3	67.76
Nov-2001	2512	7259.9	286315	177.5	5.5	67.76
Dec-2001	2510	7243.0	286533	177.4	5.7	67.76
Jan-2002	2493	7275.0	286739	177.7	5.7	67.18
Feb-2002	2532	7316.7	286935	178.0	5.7	67.18
Mar-2002	2517	7335.2	287131	178.5	5.7	67.18
Apr-2002	2503	7402.9	287343	179.3	5.9	67.18
May-2002	2511	7380.6	287571	179.5	5.8	67.18
Jun-2002	2511	7426.7	287808	179.6	5.8	67.18
Jul-2002	2505	7484.4	288051	180.0	5.8	67.18
Aug-2002	2481	7507.3	288303	180.5	5.7	67.18
Sep-2002	2467	7481.9	288554	180.8	5.7	67.18
Oct-2002	2438	7517.6	288794	181.2	5.7	67.18
Nov-2002	2454	7543.9	289012	181.5	5.9	67.18
Dec-2002	2526	7598.1	289214	181.8	6.0	67.18
Jan-2003	2461	7629.7	289412	182.6	5.8	67.10
Feb-2003	2459	7629.8	289606	183.6	5.9	67.10
Mar-2003	2487	7678.2	289809	183.9	5.9	67.10
Apr-2003	2506	7705.6	290024	183.2	6.0	67.10
May-2003	2479	7717.1	290250	182.9	6.1	67.10
Jun-2003	2493	7759.4	290484	183.1	6.3	67.10
Jul-2003	2525	7821.2	290726	183.7	6.2	67.10
Aug-2003	2580	7915.5	290974	184.5	6.1	67.10
Sep-2003	2607	7909.2	291222	185.1	6.1	67.10
Oct-2003	2628	7921.3	291463	184.9	6.0	67.10
Nov-2003	2593	7968.1	291677	185.0	5.8	67.10
Dec-2003	2629	7994.3	291868	185.5	5.7	67.10
Jan-2004	2627	8065.9	292046	186.3	5.7	67.30
Feb-2004	2623	8096.5	292230	186.7	5.6	67.30
Mar-2004	2632	8131.7	292434	187.1	5.8	67.30
Apr-2004	2668	8147.1	292651	187.4	5.6	67.30
May-2004	2680	8221.7	292872	188.2	5.6	67.30
Jun-2004	2664	8212.9	293103	188.9	5.6	67.30
Jul-2004	2672	8277.3	293350	189.1	5.5	67.30
Aug-2004	2676	8298.4	293603	189.2	5.4	67.30

Sep-2004	2696	8373.7	293857	189.8	5.4	67.30
Oct-2004	2722	8424.1	294104	190.8	5.5	67.30
Nov-2004	2709	8468.6	294337	191.7	5.4	67.30
Dec-2004	2690	8528.7	294561	191.7	5.4	67.30
Jan-2005	2648	8542.4	294768	191.6	5.3	68.20
Feb-2005	2787	8591.2	294955	192.4	5.4	68.20
Mar-2005	2738	8642.2	295149	193.1	5.2	68.20
Apr-2005	2750	8726.2	295359	193.7	5.2	68.20
May-2005	2739	8685.3	295582	193.6	5.1	68.20
Jun-2005	2777	8779.5	295824	193.7	5.0	68.20
Jul-2005	2766	8871.0	296077	194.9	5.0	68.20
Aug-2005	2797	8879.6	296338	196.1	4.9	68.20
Sep-2005	2824	8936.8	296606	198.8	5.0	68.20
Oct-2005	2852	8970.5	296857	199.1	5.0	68.20
Nov-2005	2878	8992.5	297089	198.1	5.0	68.20
Dec-2005	2901	9025.2	297311	198.1	4.9	68.20
Jan-2006	2945	9098.1	297526	199.3	4.7	69.70
Feb-2006	3031	9123.0	297734	199.4	4.8	69.70
Mar-2006	2944	9157.3	297950	199.7	4.7	69.70
Apr-2006	2979	9220.6	298170	200.7	4.7	69.70
May-2006	2996	9248.3	298401	201.3	4.6	69.70
Jun-2006	2986	9278.9	298653	201.8	4.6	69.70
Jul-2006	2991	9357.4	298910	202.9	4.7	69.70
Aug-2006	2994	9368.2	299178	203.8	4.7	69.70
Sep-2006	3021	9389.6	299452	202.8	4.5	69.70
Oct-2006	3022	9412.4	299710	201.9	4.4	69.70
Nov-2006	3069	9433.3	299950	202.0	4.5	69.70
Dec-2006	3069	9524.8	300178	203.1	4.4	69.70
Jan-2007	3080	9561.4	300398	203.4	4.6	69.70
Feb-2007	3129	9600.5	300608	204.2	4.5	69.70
Mar-2007	3157	9643.3	300823	205.3	4.4	69.70
Apr-2007	3116	9688.8	301045	205.9	4.5	69.70
May-2007	3214	9730.8	301278	206.8	4.4	69.70
Jun-2007	3250	9743.2	301528	207.2	4.6	69.70
Jul-2007	3232	9775.4	301790	207.6	4.7	69.70
Aug-2007	3178	9815.5	302064	207.7	4.6	69.70
Sep-2007	3194	9862.2	302334	208.5	4.7	69.70
Oct-2007	3163	9885.0	302590	209.2	4.7	69.70
Nov-2007	3176	9957.9	302834	210.8	4.7	69.70
Dec-2007	3238	10003.2	303062	211.4	5.0	69.70
Jan-2008	3193	10014.5	303280	212.2	5.0	68.80

Feb-2008	3202	9997.2	303494	212.6	4.9	68.80
Mar-2008	3230	10043.8	303707	213.4	5.1	68.80
Apr-2008	3239	10081.5	303926	214.0	5.0	68.80
May-2008	3257	10121.7	304157	215.2	5.4	68.80
Jun-2008	3329	10176.2	304396	217.5	5.6	68.80
Jul-2008	3349	10171.0	304646	219.1	5.8	68.80
Aug-2008	3335	10143.8	304903	218.7	6.1	68.80
Sep-2008	3325	10092.5	305158	218.9	6.1	68.80
Oct-2008	3349	9992.2	305403	217.0	6.5	68.80
Nov-2008	3339	9847.0	305620	213.1	6.8	68.80
Dec-2008	3312	9744.8	305827	211.4	7.3	68.80
Jan-2009	3356	9790.3	306035	212.0	7.8	68.60
Feb-2009	3315	9780.5	306237	212.8	8.3	68.60
Mar-2009	3329	9734.5	306438	212.6	8.7	68.60
Apr-2009	3323	9730.3	306645	212.7	9.0	68.60
May-2009	3359	9753.1	306863	213.0	9.4	68.60
Jun-2009	3301	9808.3	307090	214.7	9.5	68.60
Jul-2009	3314	9834.5	307322	214.7	9.5	68.60
Aug-2009	3369	9960.2	307570	215.5	9.6	68.60
Sep-2009	3357	9871.7	307826	215.9	9.8	68.60
Oct-2009	3353	9925.2	308071	216.5	10.0	68.60
Nov-2009	3350	9951.3	308289	217.1	9.9	68.60
Dec-2009	3427	10011.0	308495	217.3	9.9	68.60
Jan-2010	3367	10024.7	308706	217.5	9.8	68.60
Feb-2010	3447	10058.5	308904	217.4	9.8	68.60
Mar-2010	3447	10124.2	309089	217.4	9.9	68.60
Apr-2010	3466	10131.8	309268	217.4	9.9	68.60
May-2010	3420	10155.6	309453	217.2	9.6	68.60
Jun-2010	3429	10157.3	309649	217.2	9.4	68.60
Jul-2010	3368	10187.9	309858	217.6	9.5	68.60
Aug-2010	3471	10260.6	310071	218.1	9.5	68.60
Sep-2010	3465	10282.3	310283	218.4	9.5	68.60
Oct-2010	3505	10350.2	310488	219.0	9.5	68.60
Nov-2010	3501	10405.3	310673	219.4	9.8	68.60
Dec-2010	3474	10450.3	310863	220.4	9.3	68.60
Jan-2011	3499	10496.6	311031	221.0	9.1	70.06
Feb-2011	3586	10561.5	311189	222.0	9.0	70.06
Mar-2011	3546	10640.8	311356	223.2	8.9	70.06
Apr-2011	3542	10680.0	311534	224.0	9.0	70.06
May-2011	3566	10692.1	311715	224.6	9.0	70.06
Jun-2011	3611	10682.5	311905	224.8	9.1	70.06

Jul-2011	3619	10758.6	312108	225.5	9.0	70.06
Aug-2011	3629	10778.5	312317	226.3	9.0	70.06
Sep-2011	3652	10836.4	312531	226.9	9.0	70.06
Oct-2011	3651	10861.1	312735	226.8	8.9	70.06
Nov-2011	3670	10874.0	312919	227.0	8.6	70.06
Dec-2011	3611	10886.3	313095	227.0	8.5	70.06
Jan-2012	3692	10941.8	313261	227.5	8.3	70.50
Feb-2012	3712	11025.8	313422	228.4	8.3	70.50
Mar-2012	3720	11054.1	313593	229.1	8.2	70.50
Apr-2012	3744	11080.3	313773	229.2	8.1	70.50
May-2012	3763	11061.7	313957	228.5	8.2	70.50
Jun-2012	3734	11059.5	314150	228.6	8.2	70.50
Jul-2012	3700	11102.6	314353	228.7	8.2	70.50
Aug-2012	3656	11137.2	314562	230.1	8.1	70.50
Sep-2012	3711	11223.4	314777	231.4	7.8	70.50
Oct-2012	3747	11213.8	314981	231.8	7.9	70.50
Nov-2012	3803	11255.4	315165	231.0	7.8	70.50
Dec-2012	3924	11278.0	315342	231.0	7.8	70.50

	High School Grad Rate (Original) **Source:** (Preparation for College)
1990	71.18%
1991	n/a
1992	n/a
1993	n/a
1994	n/a
1995	68.62%
1996	68.06%
1997	67.15%
1998	67.76%
1999	67.18%
2000	67.10%
2001	67.30%
2002	68.20%
2003	69.70%
2004	69.70%
2005	68.80%
2006	68.60%
2007	n/a
2008	70.06%
2009	70.50%