Methods for Forecasting Freight in Uncertainty: Time Series Analysis of Multiple Factors

Final Report for:

Alabama Department of Transportation
Research Project 930-768

Project Advisory Committee
Mr. Robert Jilla – Chair
Mr. G. M. Harper
Mr. Bill Couch
Dr. Emmanuel C. Oranika
Ms. Michelle Owens – RAC Liaison
Mr. Rick Joki – FHWA Liaison

January 31, 2011

Principal Investigator:
Phillip A. Farrington, Ph.D.

Co-Investigator:
Gregory A. Harris, Ph.D., P.E.

Office for Freight, Logistics & Transportation
College of Engineering & College of Business Administration Research Centers
UAHuntsville

Methods for Forecasting Freight in Uncertainty:
Time Series Analysis of Multiple Factors

Table of Contents

Table of Tables

Table of Figures

Methods for Forecasting Freight in Uncertainty: Time Series Analysis of Multiple Factors

Executive Summary

The main goal of this research was to analyze and more accurately model freight movement in Alabama. Ultimately, the goal of this project was to provide an overall approach to the integration of accurate freight models into transportation plans and models in Alabama.
The first step in the process was to identify the dependent variable and collect the data necessary to develop the models. Initially, Truck Vehicle Miles Traveled (VMT) was the preferred dependent variable however, data collection revealed that the available VMT was not particularly accurate since it is derived from VMT data for all vehicles and there was no validated method for estimating the percentage of trucks in any one year. Therefore, the research team determined that annual Diesel Tax collections would be a good surrogate for Truck VMT. The Diesel Tax collections were used to estimate the Diesel Gallons Sold each year by dividing by the tax rate for that year. This variable, Diesel Gallons Sold (DGS), has the advantage that it could be used to estimate annual truck volumes based on estimate mileage performance for trucks. Thus, DGS was chosen as the dependent variable for this study.

The second step was to identify the set of candidate independent variables for the study. Based on the literature review a wide variety of variables were considered. Ultimately, nine variables were selected for further study because an appropriate amount of data was available with which to build and validate the models. The nine variables include: Alabama Population (ALPOP), Alabama Total Personal Income (ALTPI), Alabama Labor Force (ALLF), Alabama Employment (ALEMP), Alabama Unemployment (ALUEMP), Alabama Employment by Industry (Total using SIC and NAIC codes) (ATE), Alabama Value of Shipments (ALTVS), Alabama GDP (ALGDP), Southeast States GDP (SEGDP), and U.S. GDP (USGDP). At least 32 years of data was available for each of these variables.

The third step was to develop a set of models for predicting annual Diesel Gallons Sold. Over 600 potential models were investigated from which a reduced set of nine models were given extensive consideration with three final models, labeled Models E, H, and I respectively, were deemed to be the best candidates to use in predicting DGS annually. Model E used Alabama Employment and Alabama Gross Domestic Product as the predictors. Model H predicts DGS using Alabama Employment. Finally, Model I used Alabama Employment and U.S. Gross Domestic Product as the predictors of DGS annually. All three models predicted over 90% of the variability in the dependent variable (DGS). The validity of the models was supported by their high $R^2_{prediction}$ values which all exceeded 90% indicating good predictive capabilities. The UAHuntsville research team recommends that if a single model is used for prediction that Model H (the single variable model) be used because it is the simplest of the models with the prediction based on the Alabama Employment Levels (ALEMP). In addition, the prediction limits for this model are relatively easy to calculate. After identification and validation of the predictive models, a methodology was developed for estimating the number of trucks annually on Alabama roadways. This methodology involves the steps outlined below:

1. The number of truck trips for a reference year was taken from the Alabama State Freight Model.
2. The miles per trip were determined by calculating average miles per trip using DGS multiplied by the average fuel economy of over the road trucks, divided by 250 workdays per year, then divided by the number of trips found in step 1.
3. Average number of trucks on the road per day (i.e., truck trips per day) is then calculated based upon the projected DGS.

A spreadsheet was created which implements this methodology and generates estimates for the three models developed along with 95% prediction limits for these estimates for DGS and Average Truck Trips per Day.

1.0 Introduction

The main goal of this research was to develop a set of models that could be used to predict more accurately the movement of freight through the state of Alabama. The approach taken was to indentify a measure of truck traffic and a set of independent variables that could be used to predict the metric for truck traffic. These variables were then used to develop a set of models for estimating truck traffic in a given year utilizing available forecasts of the independent variables. A secondary goal of this study was to provide a framework for linking the models developed for this study and the Statewide Freight Flow model. This report will review the details of the study and provide the background explanation for the models and processes that were developed for predicting annual freight flow in Alabama.

2.0 Task 1: Literature Review on Factors applied for Freight Research

The movement of freight is vital to the economic growth of a state or region. An expanding population and significant growth in economic activity within the state of Alabama have brought about changes in the level of freight movement within the state. The increased volume of freight on the transportation network places significant demand on the existing infrastructure. These volumes are greatly affected by issues in the national and state economy and the ability to predict accurately the effect of national and state economic changes on future freight movements is essential to effective transportation planning activities (UAH Earmark, 2010).

Forecasting freight at the state level is more complex than modeling passenger volumes. Most freight forecasting to date involves analyzing truck traffic or using commodity flow models (Yang, Chow, & Regan, 2009). Some agencies have used time-series methods to predict future traffic volumes by extrapolating trends observed in historical data (Horowitz & Farmer, 1999). This literature review examines several types of freight forecasting studies to investigate and understand the factors that have been applied for freight in past research and the outcomes of those attempts.

I. Freight forecasting using commodity flow models

Commodity flow models attempt to represent freight movement using the flow of individual commodities from origin to destination. The data is usually derived from surveys of producing, consuming, or transportation-related entities. One public source for these data is the Commodity Flow Survey (CFS) produced by the US Census Bureau. Another is the Federal Highway Administration's Freight Analysis Framework (FAF). The FAF is based on the CFS plus some additional sources including Carload Waybill Samples and the US Army Corps of Engineers' Waterborne Commerce data. There are also private sources of commodity flow data. Global Insights' commercially available TRANSEARCH database is commonly used for these types of data.

Indiana

Freight flows in the state of Indiana were modeled on two occasions by the University of Indiana's Transportation Research Center. The first study used data from the 1977 Census of Transportation, the 1993 Commodity Flow Survey, County Business Patterns and a Carload Waybill Sample to determine base year, 1993, commodity flows (Black, 1997). In this effort, forecasts were developed for 2005 and 2015. To develop these flow forecasts, Black used forecasts of population and manufacturing from a private company, Woods and Poole. The data was available for Indiana counties as well as a national level forecast. Freight was allocated to the county level using the following factors:

- Manufactured goods: average annual growth in population and manufacturing employment in the county was used to forecast both production and attraction
- Farm products: farm earnings were used to forecast production and population growth used for attraction.
- Coal: forecasts of total earnings from mining were used to forecast both production and attraction
- Non-metallic minerals: total earnings from mining was used to forecast both production and attraction
- Waste & scrap: growth in employment in manufacturing was used to forecast both production and attraction

In the second study, the base year flows were derived from the 1997 Commodity Flow Survey from the US Department of Commerce (Black, 2006). Models were developed for each SCTG commodity by examining population and employment in various NAICS industries. From these models, future projections of the commodity flows were developed using gross (not commodity-specific) county-level employment projections. These were supplemented with labor productivity changes for several industry sectors derived from Indiana's REMI model. The final growth factors for future years 2015 and 2030 therefore included growth due to employment changes as well as growth in productivity of the employees in the various industries. The factors were applied to the base year commodity flows to generate production and attraction forecasts for 2015 and 2030, which were then input to a calibrated, fully-constrained gravity model. Traffic volumes were developed for both truck and rail modes.

Black does caution that the estimates could be improved by using county-level employment forecasts that are also industry-specific. An explicit assumption in this study is that all the industries in a county would grow at the same rate, which may be unrealistic.

Wisconsin

In a 2000 paper the method used by the Wisconsin DOT to model heavy truck trips at the state level using commodity flow data is described (Sorratini & Smith, 2000). Base year (1992) freight production was developed from two sources: the 1993 Commodity Flow Survey from the US Department of Commerce and the private TRANSEARCH database. Base year attraction was developed using state-level I-O direct coefficients from a software package. The state-level freight volumes were then disaggregated to county level using county employment data by sector from the County Business Patterns published by the US Census Bureau. These county-level flows were then used to model traffic counts and used as input into a trip generation step.

As a calibration procedure, 40 links were selected and model output compared to actual ground counts. To test forecasting ability of the final calibrated model, it was used to back-cast traffic flows for 1977. The 1977 estimate was developed from the base year using population and employment changes, as well as a productivity index based on a 4% annual increase in productivity. The model link volume to actual ground count ratio was 0.84 for the back-cast model. These results suggest that reasonable forecasts can be generated from this procedure if reliable employment, population and productivity data are available.

Montana

Montana used a private source for the commodity flow data for its freight forecasting procedure (Waliszewski, Ahanotu, & Fischer, 2004). The Reebie Transearch database was used to provide daily county level truck tonnage flows for the 2001 base year. Future year tonnage (2025) was developed using state-level growth factors from the FHWA FAF database. The FAF data included forecasted freight flows for 1998 and 2020. These flows were extrapolated to 2001 and 2025 to match the desired forecast period. A 2001 to 2025 growth factor was then calculated from these flows and applied to the 2001 Reebie Transearch data to obtain 2025 flows. The state-level forecasted flows were allocated to counties using county-level employment and population data from a private source, Woods and Poole Economics.

For internal and outbound trips, the Woods and Poole economic data was converted from 13 industry categories to Standard Transportation Commodity Codes (STCC). For inbound trips, the inbound flows were matched to consuming industries and personal consumption. The county-level change in employment from base to future year was calculated for each STCC commodity. This future distribution of employment was used to allocate the forecasted 2025 state-level flows for each commodity to the counties. The flows assigned to personal consumption were allocated to counties based on population.

This research effort sought primarily to determine if the method used produced statistically different results from the previous method used to forecast freight in Montana, which did not include sub-state economic demographic allocation. The new method was found to produce a statistically different forecast, but the authors also assert that the forecast is better than that produced by the previous method. How they came to this conclusion was not clear.

II. Freight forecasting using truck traffic analysis

Truck traffic analysis attempts to model freight traffic using the movement of commercial vehicles. The FHWA produces the Annual Highway Statistics Series which includes vehicle-miles of travel data. These statistics are generated from data submitted by the 50 states and the District of Columbia. The FHWA generates a stratification of vehicle types which gives information on the VMT for trucks. The data originates from the individual states and are thus available, in theory, for statewide and sub-state freight forecasting efforts. Individual freight companies do maintain freight flow data, but are usually reluctant to release them as they are considered proprietary.

South Dakota

In 2000, the South Dakota Department of Transportation (SDDOT) launched an investigation of their current freight forecasting methodology to identify potential improvements to the process and develop improved 20-year forecasting factors.

South Dakota collects and maintains truck traffic counts to develop its Vehicle Miles of Travel Report (Johnson, 2000). In the study, the historical VMT data for each county in South Dakota were distributed into appropriate categories by vehicle type and roadway classification. This differed from the previous method of apportioning statewide VMT to six traffic analysis regions. Linear regression analysis was used to determine 'Base Year' (1998) and 5-year VMT projections for each of the categories of traffic. These projections were converted to annual growth factors as a compound percentage annual growth, and then projected 20 years into the future.

A similar procedure was used to determine annual growth rates for county business data from the US Census Bureau. The business data were grouped by two categories and the growth factor for each was modified to reflect the appropriate passenger/commercial traffic split that the industry category is expected to generate. These business growth rates were then compared to the VMT growth rates, through a graphical analysis. VMT growth trends that differed significantly from business growth trends were investigated and adjusted through a simple algorithm. In situations where the application of the algorithm was inappropriate, the adjustment involved a review of population and building permit data or the professional judgment of the forecaster.

Final 20-year forecasting factors were generated from the base year and 20-year VMT forecasts. The factors were validated by applying them to historic VMT data and projecting a traffic forecast for a recent year with known VMT. The report indicated that the results were favorable, with forecasts being within a few percentage points of known VMT values.

Pennsylvania
In a Pennsylvania case study, several types of models were used to predict annual growth in VMT (Liu, Kaiser, Zekkos, & Allison, 2006). The Pennsylvania Department of Transportation (PENNDOT) provided an extensive traffic database for the project, which included VMT by functional classification, by county, for all public roads, annually between 1994 and 2003.

Several types of models were investigated, with the natural logarithm of VMT as the dependent variable, testing the effect of a variety of explanatory variables and indices. Ordinary least square (OLS) models were found to be most suitable for the needs of the state. The study recommended that PENNDOT use a range as its annual VMT growth rate with bounds defined by different models. The recommended model for the upper bound has households and mean household income as independent variables. The recommended model for the lower bound has household-based and county-group-level independent variables. A point estimate for the annual growth would be the average of the upper and lower boundaries.

The study recommends that future VMT forecasting should include both demand factors (socio-economic) and supply factors (lane-miles). In addition, it is proposed that the effect of VMT reclassification on model results should be considered, possibly through sensitivity analysis. This approach may be applicable to other areas, provided the data needed is easily available. In this case, the requesting organization was able to supply extensive data on the dependent variable, which proved invaluable to the analysis.

Truckload traffic forecasting
Another group of researchers analyzed truck traffic to investigate economic indices, which may have an impact on freight demand (Fite, Taylor, Usher, English, & Roberts, 2002). A slightly different approach was used in that truck movement for a single company was analyzed as opposed to state or citywide truck traffic.

Thirty-one months of trucking data were used as the dependent variable and 107 economic and industrial indices were considered as predictors for a stepwise multiple linear regression analysis. Producer commodities price index of construction materials and equipment (PCI-CM&E) was the only variable to enter the national model in a stepwise regression procedure using 10% to enter and showed a lead of 3 months. Months 32-36 were forecast by the model and showed an average error of 6.86%. Regional and industry models were developed, but the errors were much larger, diminishing their predictive suitability. The results from this analysis suggest that a fairly simple procedure can be used to model truckload trucking volumes with reasonable prediction errors. It would be useful to extend this type of analysis to wider data sets.

III. Time-series analysis

Time–series models consider the relationship between the response variable and time. These models look for trends in the growth or decline in the data. Time-series models are usually aggregate in nature and do not attempt to explicitly describe the explanatory factors in those trends. Time-series models ultimately try to predict the value of the response variable at some point in the future usually through the extrapolation of the identified historical trends in freight volumes. The benefit in using this modeling method is that it is relatively simple to implement and not particularly data intensive (Pendyala, Shankar, & McCullough, 2000).

In a 1986 paper, Julian Benjamin suggested that instead of a structural approach, under proper circumstances, only traffic trends need to be modeled (Benjamin, 1986). He proposed an equation modeling traffic volumes to time which results in an s-shaped curve, starting and ending with asymptotically constant rates. He tested this time-series forecasting approach in Greensboro, NC, selecting three streets with markedly different growth profiles. He used data available in 1975 to estimate 1981 traffic volumes and compared the estimates with actual traffic counts. He compared the results of the time-series model with estimates produced by the Urban Transportation Planning System (UTPS) computer package from the US Department of Transportation. The comparison showed that the estimates from Benjamin's model performed equivalently or better than the UTPS estimates for all three streets.

Benjamin contends that this time-series method is simpler, less costly and easier to understand than structural models, and still produce an acceptable level of accuracy. He does present some limitations to this method. He mentions its lack of sensitivity to large changes in the transportation system, which is a major weakness in this method. Situations involving new arterials, significant mode shifts, direction changes and other substantial events are not suitable for modeling with this approach. There is also an assumption that there is little variation in travel behavior during the modeling period. However, Benjamin asserts that although travel habits, changes in economic activity and gasoline prices cannot be forecast with this method, his results indicate that historically they have had little effect on average traffic. He ultimately suggests using structural models for areas with significant anticipated changes and the lower-cost time-series methods for the rest of the area under consideration.

Another researcher uses Benjamin's time-series approach to forecast average daily traffic volume on an Egyptian inter-city road (Sabry, Abd–El-Latif, Yousef, & Badra, 2007). Traffic data for 1990-2000 were used to fit the model, and then estimates were developed for 2001-2003. These estimates were compared with actual traffic volumes for January 2001 to December 2003.

In this application, three different time periods were examined for the traffic data: average annual daily traffic, average monthly daily and average weekly daily traffic. The results were then compared to actual traffic counts. The model performed best using the average annual daily traffic data. In both of these cases, total traffic is the variable being modeled, as opposed to freight traffic.

IV. Summary

The available literature presents a wide variety of methods and data used by state departments of transportation and other agencies in forecasting freight flows.

Techniques used
Researchers have successfully used growth factors, regression models and time series analysis methods to determine future freight flows [(Fite, Taylor, Usher, English, & Roberts, 2002), (Benjamin, 1986), and (Sabry, Abd–El-Latif, Yousef, & Badra, 2007)]. The growth factor method is simple, using historical data to calculate growth rates that can be applied to determine values at a future point. However, these require the assumption that the future will continue to behave like the past, which may be unrealistic. Time-series analysis exhibits a similar downside. In addition, it requires extensive historical information for the variable used, which may be difficult to obtain. Regression analysis offers insight into the major forces that drive freight growth and has been used effectively to describe the relationship between measures of freight flow and demand and supply related variables.

Data used
The types of data used to represent and explain freight movement also varied among research efforts. Commodity flow data at the state and/or county levels have been used as the response variable. Public sources for this type of information are not widely available at the level of detail required for state or sub-state level freight forecasting. Many have turned to private sources for their commodity flow data. Traffic counts are maintained by several city/state agencies but these generally rely on sampling methods. The sampling procedure used affects the quality of the estimates, and error is introduced when extrapolating from the sample to the larger network. There are some states that have not maintained extensive historic traffic data, which limits its use as a dependent variable in statewide freight forecasting.

A large variety of explanatory variables has been examined for use in freight forecasting procedures. Socio-economic variables like population, industry sector growth, and employment are some of the most popular. However, the classification and reporting schedule and format of these data usually do not match those of VMT, commodity flows or traffic counts. Additional procedures to match industries to commodities or traffic zones to county geography usually need to be undertaken.

Fuel-tax based methods of freight estimation have been undertaken (Fricker & Kumapley, 2000). However, fuel tax data do not necessarily represent freight flows very well over time. Fuel tax has to be converted into some measure of volume (gallons) and the conversion factor changes with policy changes with regard to fuel tax rates. In addition, changes in fuel economy also have an impact on the number of miles travelled per unit volume. Additional procedures to include these changes over time have to be included when using fuel tax in freight forecasting. Additional data sources have been explored including GIS, satellites and unmanned aerial vehicles (UAVs), but these are mostly as supplements to existing methods (Fricker & Kumapley, 2000).

V. Conclusion

A significant task in any freight forecasting effort is to determine what types of data and which techniques are best suited for the analysis. Part of that decision rests on where and how the results will be implemented. If freight traffic impact on roadway infrastructure is a primary concern, then truck traffic related data, like commercial VMT, is well suited as a dependent variable. If general freight flows are desired or additional modes needs to be considered, then commodity flow data may be fitting. However, the technique and data chosen for freight forecasting also depend heavily on the extent and quality of the data available to researchers. In modeling the freight movements, extensive historic traffic counts/VMT data lend very well to time series or regression analyses, but are not always readily available as many public sources have not maintained their traffic data adequately. Public and private sources exist for commodity flow data but the surveys, which generate these data, do not occur often enough to provide long-term historic trends for time-series analysis. Demand and supply explanatory variables seem to be more readily available, but the classification and reporting schedule and format of these data usually do not match the freight movements they are used to explain.

Several types of data and forecasting techniques have been used by state DOTs and other agencies to model and predict freight movements. This literature review suggests that a primary concern is quantity and quality of data available to forecast statewide freight. As a result, this availability (or lack thereof) becomes a driving force behind researchers' choices of type of data and technique for freight modeling and forecasting.

3.0 Task 2: Data Collection

After reviewing the literature the research team developed a set of candidate dependent and independent variables that were used for the study (Appendix A contains a more detailed description of each of the variables). There was significant discussion on the topic of the appropriate dependent variable. The previous research had utilized a variety of dependent variables to forecast freight flows including: vehicle miles travelled (VMT), traffic counts, and commodity flows. Of these, in initial discussions, the most promising of these for modeling freight flow in Alabama was thought to be highway truck counts and/or truck vehicle miles traveled, however, current data collection methods and data sources did not provide sufficiently accurate data to be usable for this investigation. A second variable that was considered was the annual Alabama Motor Vehicle Fuel Tax collections. Unfortunately, there was no way to effectively separate the car and truck data. Other variables that were investigated include Truck Registrations, and Diesel Tax Collections. Of these, the most promising was Diesel Tax Collection data, which was determined to be the best available dependent variable. Diesel Tax Collection data was available from 1970 through 2009, which provided sufficient data for model development and validation. The following list shows the variable, the label used in this study, and the source of the data.

Candidate Dependent Variables (Variable Label) - Source:

- **Alabama Truck Vehicle Miles Traveled (ALVMT) –** *Alabama Department of Transportation*
- **Alabama Diesel Tax Collections (DT) –** *Alabama Department of Revenue*
- **Alabama Motor Vehicle Fuel Taxes (VFT) –** *Alabama Department of Revenue*

To provide greater applicability the Diesel Tax Collection data was converted to Gallons of Diesel sold annually. This conversion was done by taking the Annual Diesel Tax Collections and dividing by the current tax rate for each year. For the remainder of this report the primary dependent variable used for the study was Diesel Gallons Sold (DGS). This was deemed to be a reasonable surrogate for freight traffic and can be used to estimate the number of trucks on Alabama roadways per day. The process for estimating trucks will be outlined under section 5 of this report.

The research team also identified several potential independent variables that could be good predictors of Freight volume in the state of Alabama. Ultimately, the 10 variables listed on the next page were identified as potential predictors of Freight traffic in the state of Alabama.

Independent Variables (Variable Label) – Source:

Alabama Population (ALPOP) – *US Census Bureau*
Alabama Total Personal Income (ALTPI) – *Bureau of Economic Analysis*
Alabama Labor Force (ALLF) – *Bureau of Labor Statistics*
Alabama Employment (ALEMP) – *Bureau of Labor Statistics*
Alabama Unemployment (ALUEMP) – *Bureau of Labor Statistics*
Alabama Employment by Industry (ATE) – *Bureau of Economic Analysis*
Alabama Value of Shipments (ALTVS) – *U.S. Census Bureau, Economic Census of Manufacturing, Geographic Area Series & State Statistical Abstracts*
Alabama GDP (ALGDP) – *Bureau of Economic Analysis*
Southeast States GDP (SEGDP) – *Bureau of Economic Analysis*
U.S. GDP (USGDP) – *Bureau of Economic Analysis*

Table 1 (on the next page) presents the variables with data available from 1970 through 2009.

Table 1: Variables and Data available from 1970 through 2009

	Diesel Gallons Sold	AL Total Personal Income	AL Population	AL labor force	AL employment	AL unemployment	Alabama Total Value of Shipments	AL Total Employment	AL Total GDP	SE Total GDP	US GDP
Year	**DGS**	**ALTPI**	**ALPOP**	**ALLF**	**ALEMP**	**ALUEMP**	**ALTVS**	**ATE**	**ALGDP**	**SEGDP**	**USGDP**
1970	113,776,961	10,218,849	3,444,354				9,455,965,000	1,412,928	12,455,000,000	155,825,000,000	1,038,300,000,000
1971	129,032,401	11,212,347	3,497,076				9,561,000,000	1,423,459	13,599,000,000	172,095,000,000	1,126,800,000,000
1972	184,171,630	12,483,468	3,539,400				11,195,000,000	1,470,523	15,336,000,000	194,650,000,000	1,237,900,000,000
1973	218,054,625	14,117,858	3,579,780				13,020,000,000	1,525,967	17,416,000,000	223,216,000,000	1,382,300,000,000
1974	232,882,598	15,731,053	3,626,499				14,126,700,000	1,552,266	19,438,000,000	245,909,000,000	1,499,500,000,000
1975	224,123,857	17,543,230	3,678,814				15,781,000,000	1,543,312	21,295,000,000	264,597,000,000	1,637,700,000,000
1976	253,223,332	19,856,618	3,735,139	1,491,367	1,396,193	95,174	17,988,000,000	1,593,952	24,206,000,000	296,895,000,000	1,824,600,000,000
1977	283,467,537	21,918,187	3,780,403	1,565,763	1,459,192	106,571	20,020,644,000	1,651,033	26,546,000,000	330,868,000,000	2,030,100,000,000
1978	319,647,225	24,782,560	3,831,836	1,621,071	1,524,618	96,453	22,623,327,720	1,712,582	30,377,000,000	378,030,000,000	2,293,800,000,000
1979	348,845,956	27,624,807	3,866,248	1,665,649	1,552,130	113,519	25,270,257,063	1,735,879	33,535,000,000	422,661,000,000	2,562,200,000,000
1980	241,922,058	30,521,535	3,893,888	1,680,780	1,538,910	141,870	28,552,000,000	1,731,866	36,006,000,000	467,887,000,000	2,788,100,000,000
1981	335,189,763	33,931,285	3,918,531	1,670,332	1,507,852	162,480	30,509,000,000	1,718,783	40,084,000,000	532,904,000,000	3,126,800,000,000
1982	340,236,292	35,925,679	3,925,266	1,702,879	1,478,260	224,619	29,939,000,000	1,687,466	41,478,000,000	557,473,000,000	3,253,200,000,000
1983	364,017,552	38,442,575	3,934,102	1,726,929	1,505,701	221,228	32,543,693,000	1,716,798	45,248,000,000	605,276,000,000	3,534,600,000,000
1984	411,754,534	42,488,038	3,951,820	1,755,660	1,576,853	178,807	36,302,000,000	1,779,584	49,713,000,000	682,138,000,000	3,930,900,000,000
1985	423,551,439	45,699,469	3,972,523	1,781,461	1,628,066	153,395	36,404,000,000	1,821,588	53,688,000,000	734,260,000,000	4,217,500,000,000
1986	444,103,887	48,218,721	3,991,569	1,838,361	1,671,375	166,986	38,515,432,000	1,858,269	56,046,000,000	772,343,000,000	4,460,100,000,000
1987	462,555,601	51,135,629	4,015,264	1,852,070	1,710,970	141,100	40,901,000,000	1,911,569	60,586,000,000	831,783,000,000	4,736,400,000,000
1988	505,209,862	54,881,335	4,023,844	1,871,682	1,742,480	129,202	44,050,377,000	1,969,768	65,435,000,000	902,110,000,000	5,100,400,000,000
1989	531,024,449	59,549,327	4,030,222	1,895,532	1,770,129	125,403	47,354,155,275	2,006,365	67,875,000,000	957,701,000,000	5,482,100,000,000
1990	558,858,572	63,254,104	4,048,508	1,903,248	1,782,700	120,548	48,748,000,000	2,047,865	71,085,000,000	1,006,364,000,000	5,800,500,000,000
1991	551,696,374	66,969,082	4,091,025	1,915,087	1,783,434	131,653	48,449,000,000	2,060,099	75,293,000,000	1,052,361,000,000	5,992,100,000,000
1992	440,370,827	71,714,011	4,139,269	1,943,033	1,809,337	133,696	52,856,900,000	2,097,425	80,450,000,000	1,118,157,000,000	6,342,300,000,000
1993	573,541,242	74,870,710	4,193,114	1,981,641	1,850,610	131,031	55,552,601,900	2,158,752	83,453,000,000	1,182,993,000,000	6,667,400,000,000
1994	578,186,306	79,480,755	4,232,965	2,018,524	1,909,881	108,643	59,052,415,820	2,180,001	88,581,000,000	1,273,469,000,000	7,085,200,000,000
1995	611,734,879	84,005,169	4,262,731	2,063,870	1,955,846	108,024	65,481,000,000	2,241,551	94,021,000,000	1,353,678,000,000	7,414,700,000,000
1996	757,621,954	87,682,073	4,290,403	2,086,493	1,992,652	93,841	66,257,000,000	2,275,108	97,941,000,000	1,430,349,000,000	7,838,500,000,000
1997	621,617,936	92,242,669	4,320,281	2,129,797	2,035,156	94,641	67,970,076,000	2,321,253	102,433,000,000	1,547,140,000,000	8,332,400,000,000
1998	654,570,182	97,858,395	4,351,037	2,142,512	2,059,310	83,202	71,708,430,180	2,370,943	106,656,000,000	1,635,430,000,000	8,793,500,000,000
1999	676,587,444	101,718,980	4,369,862	2,162,603	2,070,210	92,393	68,879,000,000	2,388,019	111,923,000,000	1,738,752,000,000	9,353,500,000,000
2000	711,022,159	107,150,846	4,451,849	2,154,545	2,067,147	87,398	70,290,000,000	2,399,989	114,576,000,000	1,812,329,000,000	9,951,500,000,000
2001	698,539,373	112,003,189	4,464,034	2,134,845	2,034,909	99,936	67,172,000,000	2,369,868	118,682,000,000	1,882,766,000,000	10,286,200,000,000
2002	693,086,316	115,396,846	4,472,420	2,107,858	1,994,748	113,110	66,686,220,000	2,369,236	123,805,000,000	1,958,127,000,000	10,642,300,000,000
2003	703,320,316	120,030,227	4,490,591	2,104,209	1,989,784	114,425	70,048,000,000	2,380,137	130,526,000,000	2,060,897,000,000	11,142,100,000,000
2004	758,316,052	128,009,032	4,512,190	2,113,781	2,007,153	106,628	76,096,000,000	2,440,586	141,366,000,000	2,212,168,000,000	11,867,800,000,000
2005	787,380,001	135,616,756	4,545,049	2,133,177	2,051,893	81,284	87,841,000,000	2,504,522	150,582,000,000	2,378,609,000,000	12,638,400,000,000
2006	840,948,686	144,436,849	4,597,688	2,176,529	2,100,558	75,971	101,862,000,000	2,564,654	158,858,000,000	2,535,414,000,000	13,398,900,000,000
2007	818,535,387	152,136,327	4,637,904	2,182,823	2,106,041	76,782	112,858,843,000	2,628,014	164,524,000,000	2,627,232,000,000	14,061,800,000,000
2008	714,747,434	157,421,997	4,677,464	2,155,941	2,044,406	111,535		2,640,717	170,014,000,000	2,689,360,000,000	14,369,100,000,000
2009	629,165,492	155,839,691	4,708,708	2,112,566	1,900,148	212,418					14,119,000,000,000
Average	493,665,962	69,103,007	4,102,342	1,936,548	1,812,019	124,528	46,892,658,894	2,006,736	74,746,948,718	1,108,313,230,769	6,334,012,500,000
Std. Dev	208,431,201	45,561,458	349,984	207,036	226,710	39,433	26,316,599,555	370,453	46,778,977,111	769,080,635,235	4,170,903,563,989
Minimum	113,776,961	10,218,849	3,444,354	1,491,367	1,396,193	75,971	9,455,965,000	1,412,928	12,455,000,000	155,825,000,000	1,038,300,000,000
Maximum	840,948,686	157,421,997	4,708,708	2,182,823	2,106,041	224,619	112,858,843,000	2,640,717	170,014,000,000	2,689,360,000,000	14,369,100,000,000

4.0 Task 3 – Statistical Evaluation of Data

A statistical analysis of the data set was performed to identify independent variables that were reasonable candidates for effectively predicting DGS each year and ultimately (after appropriate conversion) freight volumes in Alabama. The analysis focused on using regression analysis in order to develop a model for predicting DGS. The data plot shown in Figure 1 provides some idea about the dependent variable and how it varied in relation to candidate independent or predictor variables over time. The candidate variables have widely varying magnitudes, so standardized variables are used to make the plot. The standardized versions of the variables are computed by subtracting the average from the values and the result is divided by the standard deviation of the data series. The resultant standardized variables have an average of zero and standard deviation of one, with the majority of values between -2 and +2 in magnitude.

As shown in Figure 1 (following page) and Table 1 (preceding page), DGS is a broadly increasing function over time. In both 1980 and 1992, there are noticeable one-year drops in the values that recover in the following year. These drops coincide with the first year of an increase of the diesel tax rate. For instance, in 1980, the diesel tax rate increased by 50%, going from an 8% tax rate to a 12% tax rate, while in 1992, the tax rate increased by 41.67% from a tax rate of 12% to 17%. However, DGS was unaffected by a tax increase in 2005 that was only 11.8% from a rate of 17% to 19%, and there was no drop observed in DGS. On the other hand, in 1996 there was a noticeable one-year jump in DGS, which is unexplained. These three years are breaks in the relatively smooth increases in the DGS. These deviations are large in magnitude and consistently appear as outliers when compared to the fitted models.

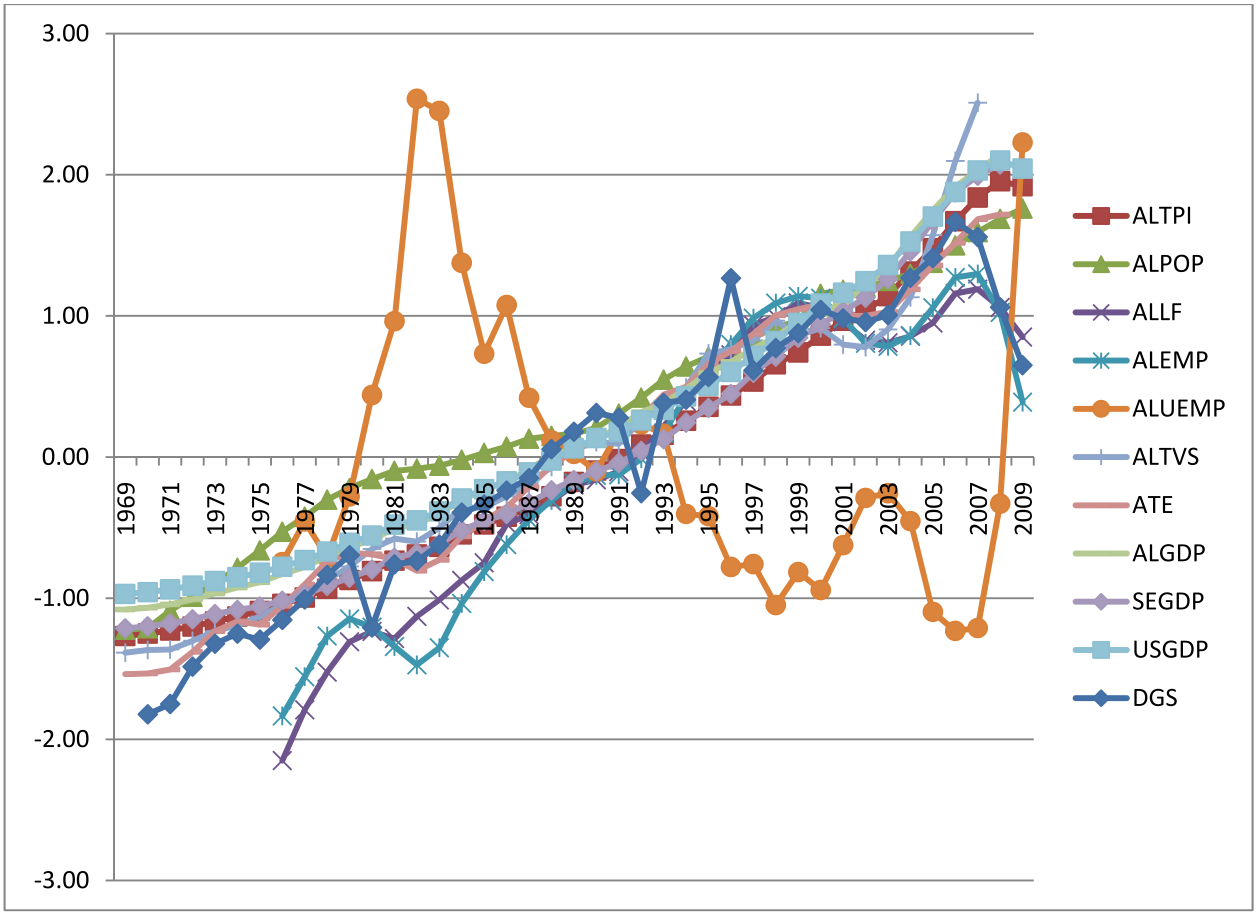


Figure 1: Diesel Gallons Sold versus Independent Variables (X-Axis Years - Y-Axis Standardized Variables)

The other exception to the generally smooth increase in DGS is the final period, from 2006 to 2009, in which DGS decreased by 25% from approximately $841M to $629M, or an average of 9% per year, reflecting the downturn in the economy. Not all variables are observed through 2009, but among those observed, both ALLF and ALEMP show similar though smaller decreases over the same time period. This is shown more clearly in Figure 2 (following page), which compares DGS to a selected set of variables.

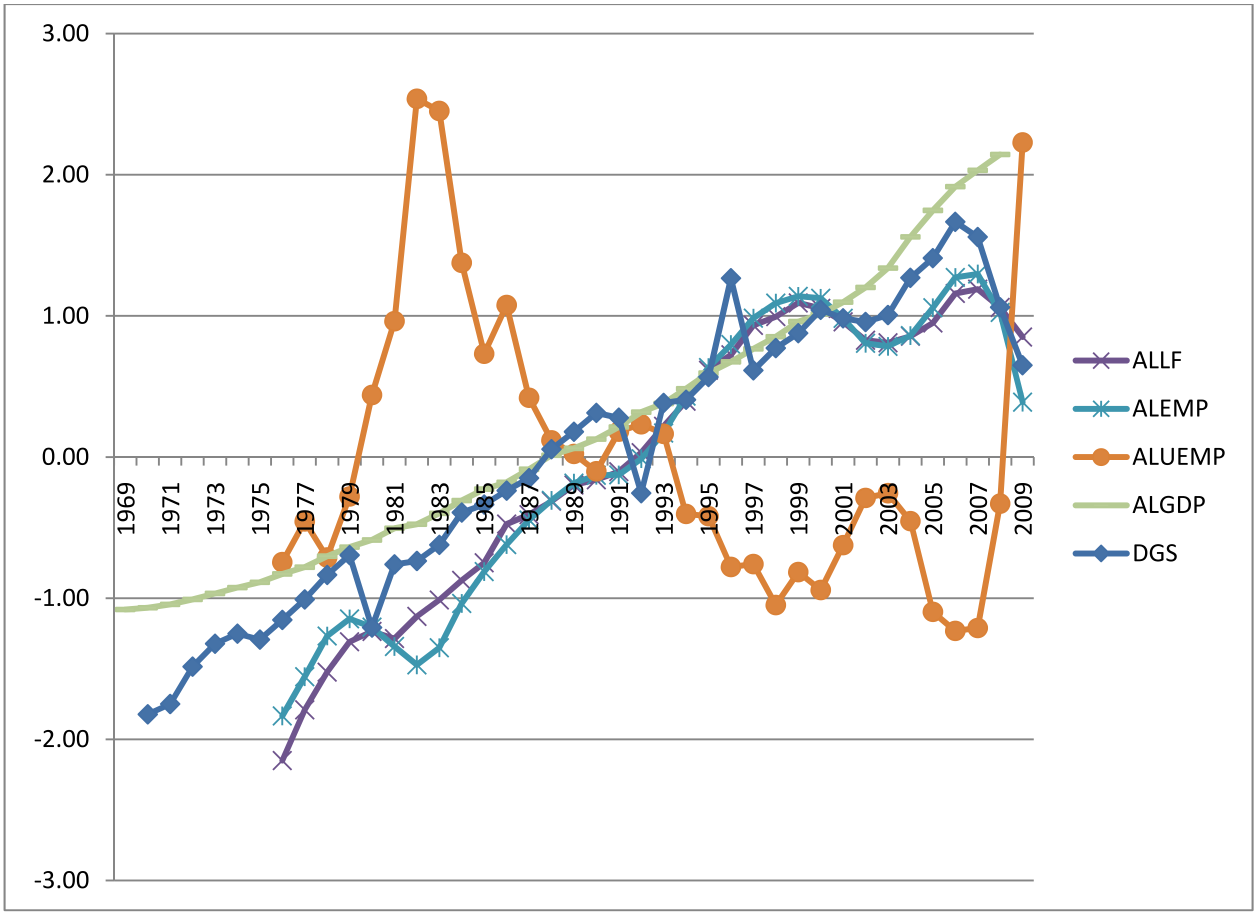


Figure 2: Diesel Gallons Sold versus Select Independent Variables (X-Axis: Years - Y-Axis: Standardized Variables)

As follow up to the graphical examination, a correlation analysis was conducted on the variables and summarized in Table 2 (following page). This preliminary analysis indicates a high correlation between most of the independent variables and the dependent variable DGS, suggesting that many of the variables would be suitable as candidate predictors. As observed from Figure 1, most of the variables are increasing over time, and therefore it is not surprising that many of the independent variables are also highly correlated with each other. This is consistent with the correlations that exceed 0.90 in magnitude. For example, the correlation between ALTPI and DGS is 0.935 indicating that as ALTPI increases (or decreases) a similar increase (or decrease) generally occurs in DGS. The only exception is ALUEMP which has a moderate negative correlation with the other variables (i.e., as ALEUMP increases the other variable tends to decrease).

Table 2: Correlation Analysis (Based on Data Shown in Table 1)

	DGS	ALTPI	ALPOP	ALLF	ALEMP	ALUEMP	ALTVS	ATE	ALGDP	SEGDP
ALTPI	0.935									
ALPOP	0.955	0.979								
ALLF	0.955	0.917	0.938							
ALEMP	0.961	0.901	0.919	0.988						
ALUEMP	-0.510	-0.365	-0.355	-0.427	-0.564					
ALTVS	0.965	0.983	0.969	0.929	0.935	-0.587				
ATE	0.978	0.982	0.987	0.971	0.980	-0.632	0.980			
ALGDP	0.957	0.999	0.979	0.928	0.930	-0.562	0.985	0.983		
SEGDP	0.947	0.999	0.971	0.918	0.921	-0.567	0.979	0.975	0.998	
USGDP	0.933	1.000	0.976	0.912	0.897	-0.368	0.980	0.978	0.999	1.000

The largest possible correlation is 1.00, which indicates perfect linear dependence. (Note, the two values of 1.000 in Table 2 are likely slightly less than 1.000 but "rounded" to that value because of the limited number of digits used.) The presence of so many highly inter-correlated variables will limit the regression models that can be constructed using these variables. The similarity of the variables is shown by their large correlations, so the variables are largely redundant in their information content. Regression models require variables that are not linearly dependent upon each other, and in an extreme case a subset of correlated variables may form a dependent set. This is known as the problem of multi-collinearity; when multi-collinearity is too large, a stable regression model cannot be fit. A variance inflation factor (VIF) is used in Minitab (the statistical analysis software used in this study) to detect this problem; values of the VIF in excess of 10 are considered a significant concern.

The next step was to generate a set of candidate models. There are a large number of models that can be developed, so variable selection methods were used. For instance, when there are 7 variables, a total of 127 models can be fit with all possible combinations of the variables, while with 9 variables, 511 models are possible. One approach to selecting models is Best Subset Selection where the best candidate models from among all possible models are presented using standard criteria such as largest squared correlation R^2, largest *Adjusted* $\bar{R}^2$ (or equivalently, smallest s^2), or according to Mallow's Cp criteria. The correlation measures are usually present as percentages, and measure the relative success in predication. For instance, an R^2 of 90% means that 90% of the variability in the dependent variable is explained by the fitted model and only 10% of the variation is unexplained. The Cp measure attempts to assess the point at which all relevant variables have been included in the model. The criteria are to pick the model that has the least Cp, which is "close" to p, the number of independent variables plus 1. When these criteria are satisfied, there is an indication that the significant predictors are included and that there are no superfluous variables included.

Table 3 (following page) summarizes the results for models of DGS with models selected from 7 candidate predictors ALTPI, ALPOP, ALTVS, ATE, ALGDP, SEGDP, or USGDP. There are 127 possible models that can be constructed. Each row represents one of the two "best" models with the number of predictors given by variables (first column). The second and third fourth columns provide the squared correlation R^2 (R-Sq), the Adjusted $\bar{R}^2$ (labeled R-Sq (adj)), both as percentages. Larger values are preferred, with 100 the largest possible. The fourth column is the Mallow's Cp. Smaller values close to the number of variables plus 1 are desired for these criteria. The fifth column is the standard deviation S of the residuals, which is the square root of the mean squared error (MSE) of the residuals. This is a measure of variation in the residuals, with smaller values preferred. The least value of S will correspond to the largest adjusted R^2 value. (For instance, see Model D.) The remaining columns are labeled for the predictor variables (e.g., ALTPI, in column six). The X's in each row indicate which predictors are included in the model (e.g., in model A.1, the X corresponds to the column for ATE).

Model A.1 is the best single variable model with ATE as the predictor. Its Cp measure is a bit high (4.7 compared to p=2), while its adjusted R^2 is almost comparable to the remaining models at 96.4. The best two variable model, Model B, consists of ATE and the Alabama Population (ALPOP), with a Cp of 5 (compared to p=3) and adjusted R^2 of 96.5. The best three variable model is Model C, with a Cp of 3.6 (compared to p=4) and adjusted R^2 of 96.7. Larger models will typically not be feasible because of multi-collinearity, but Model D with 5 variables is included as it has the minimum MSE among all candidate models.

Table 3: Results for Best Subsets Regression – (1970 to 2007) – Response Variable is DGS

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	ALTPI	ALPOP	ALTVS	ATE	ALGDP	SEGDP	USGDP	
1	96.5	96.4	4.7	39610444				X				(Model A.1)
1	95.0	94.9	21.4	47371398		X						
2	96.7	96.5	5.0	39298959		X		X				(Model B)
2	96.5	96.4	6.4	40015353			X	X				
3	97.0	96.7	3.6	37978220	X			X	X			(Model C)
3	96.9	96.6	4.8	38674177				X	X	X		
4	97.1	96.7	4.6	37993085	X	X		X	X			
4	97.1	96.7	4.7	38032360	X			X	X		X	
5	97.3	96.9	4.2	37113521	X			X	X	X	X	(Model D)
5	97.1	96.7	5.7	38042576	X	X		X	X		X	
6	97.3	96.8	6.0	37580738	X	X		X	X	X	X	
6	97.3	96.8	6.2	37706875	X		X	X	X	X	X	
7	97.3	96.7	8.0	38199859	X	X	X	X	X	X	X	

The preliminary analysis is repeated with nine variables (511 potential models), but this time restricted to 32 data points because this was all the data available for all nine variables, the results are summarized in Table 4 (next page). Model A.2 repeats the Model A.1, but in this specific model, the adjusted R^2 is lower (92.6) and the Cp is 2.5, somewhat smaller than before. For this smaller data set, the best two variable models consists of variables ALEMP and ATE (Adjusted R^2 of 95 and Cp of 0.3), the best three variables (Model C) consists of variables ALTPI, ALEMP, and ALGDP (Adjusted R^2 of 95.1 and Cp of 0.8). A potential three variable model is provided using variables ALPOP, ALLF, and USGDP (Adjusted R^2 of 95.1, Cp of 0.9).

Table 4: Results for Best Subsets Regression – (1976 to 2007) – Response Variable is DGS

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	ALTPI	ALPOP	ALLF	ALEMP	ALTVS	ATE	ALGDP	SEGDP	USGDP	
1	94.6	94.4	2.5	41864717						X				(Model A.2)
1	92.8	92.6	12.4	48145319							X			
2	95.3	95.0	0.3	39468013				X			X			(Model E)
2	95.2	94.8	1.1	40115525				X					X	
3	95.6	95.1	0.8	39059983	X			X			X			(Model F)
3	95.6	95.1	0.9	39080479		X	X						X	(Model G)
4	95.8	95.2	1.7	38864296		X		X		X	X			
4	95.7	95.1	2.2	39262112		X	X	X			X			
5	96.0	95.2	2.7	38756292	X			X			X	X	X	
5	95.8	95.0	3.5	39456168		X		X		X	X	X		
6	96.0	95.0	4.5	39385738	X			X		X	X	X	X	
6	96.0	95.0	4.5	39393273		X		X		X	X	X	X	
7	96.1	94.9	6.2	39866292	X	X		X		X	X	X	X	
7	96.0	94.9	6.3	40014819	X			X	X	X	X	X	X	
8	96.1	94.7	8.0	40594824	X	X		X	X	X	X	X	X	
8	96.1	94.7	8.2	40718279	X	X	X	X		X	X	X	X	
9	96.1	94.5	10.0	41483141	X	X	X	X	X	X	X	X	X	

The seven models, labeled A through I, were selected using a subset of the entire data set. To examine the models, the full data set was used. The total number of cases used in each model depended on the predictor variables used. Minitab uses all samples in which all variables had observations. In a few situations, this meant that that the "review" analysis had a different number of cases available compared to the selection analysis.

Once the models have been selected, they can be examined in more detail. Residual plots are used to assess whether the model assumptions are reasonable. For instance, residuals should be approximately normally distributed, and this can be assessed using a histogram and the normal probability plot. In the probability plot, the plotted points should be a straight line if the data are normally distributed. The residuals are also plotted against the fitted values (i.e., the values of DGS based upon the fitted regression model) and against the order that they were collected. The ideal is a random scatter plot with constant "spread." Appendix C contains the residual plots for each of the candidate models considered for a more detailed evaluation.

As noted in the preliminary analysis of the DGS, several values of DGS were observed in 1980, 1992, and 1996 that deviated from the relatively smooth progression of observations. These values also deviated from predicated values of DGS and resulted in large residual values. Residuals are considered to be potential outliers or values distinct from the other residuals if their "t-value" or "t-statistic" value exceeds 2 in magnitude. The t-statistic is obtained by dividing the residual by the standard error of the residuals, usually the square root of MSE. Such a value should occur only about 5% of the time.

Another method of assessing the model is in terms of their "deleted residuals" which are designed to assess how stable the model is. A deleted residual is obtained by fitting the model without a particular case and then comparing the fitted value obtained in this way to the actual observed value. A deleted residual is computed for each row of data. The PRESS statistic is obtained by squaring and summing all of the deleted residuals. The PRESS statistic is considered an alternative to the error sum of squares that is used to compute MSE. When the PRESS is divided by the total variation in the dependent variable, we have the Prediction R^2. This will be less than either of the model R^2 values. However, a good model will have only a small drop in the value of the Prediction R^2 compared to the other values.

Among the seven models, a number of similarities were observed. In most cases the models had adjusted R^2 values that were about 95%, sometimes a little more. The PRESS (or validation) R^2 values were a little lower, usually around 93%, but quite high overall. Another similarity was that, with the exception of Model G, the residuals for cases 11 (1980), 23 (1992), and 27 (1996) were consistently large enough to be tagged as potential outliers based upon the t-statistic. In model G, case 27 was not quite large enough to be identified as a potential outlier, though still fairly large. These three outlier values tended to make the normal probability plot appear to be heavy tailed and likely inflated the MSE.

In all evaluations in which case 39 (2008) was present, it was also a potential outlier. This reflects the large downturn in DGS values in 2007, 2008, and 2009, that was particularly marked in 2008 and 2009. Only Model G had data for 2009 and in that model, both the residuals for 2008 and 2009 were flagged as potential outliers. The presence of the other outliers may in fact have somewhat masked the size of this potential outlier. This is a potential concern because the three cases 11, 23, and 27 appear to be one-year anomalies; the downturn in 2007 through 2009 reflects a systematic change in the magnitude of DGS.

As noted in the preliminary analysis, many of the variables in analysis are highly correlated and the presence of multi-collinearity was a risk in building regression models from the variables. In fact, the variable for unemployment ALUEMP was eliminated from earlier analyses because its inclusion led to collinearity that was too large for numerical stability. Models B, C, D, F, and G all contained terms with Variance Inflation Factors (VIF) in excess of 30. For instance, Model B had a VIF of 37.84 for both variables, while in Model D the smallest VIF was 58.5 and the largest close to 4750. It was noted that Model D was expected to have this problem, as it contained five variables.

Overall, only Model A (with variable ATE) and Model E (with variables ALEMP and ALGDP) appear to be worth further examination. The first model contains only one variable, so collinearity is not an issue, and the VIF in the second model is a moderate value of only 7.36. To partially confirm this, both models are fit with the three outlier values omitted from the analysis (denoted Models A* and E* in the attached graphics). The overall analysis in both cases is the same. Without the other outlier values, the large residual in 2008 increases slightly in both cases and is now clearly tagged as the single outlier in each model. The overall residuals are improved, particularly in the case of Model E. In Model A, the residuals appear to be very non-independent and this is somewhat less evident in Model E. The adjusted R^2 and Predicted R^2 values increase to about 97%, reflecting the decrease in the error sum of squares in both cases.

In both Model E and E* (i.e., Model E with cases 11, 23, and 27 deleted – statistical analysis and residual plots shown in Appendix C) the last case is designated as potentially influential. This is because the large residual occurs in one of the more extreme cases in terms of the independent variables. The influence can be seen in the fitted model coefficients. For ALEMP changes its coefficient from 449 to 387 when the cases are deleted and from 0.00166 to 0.00184 for ALGDP. These differences do seem to be within their confidence intervals.

We may tentatively accept Model E as a reasonable model for predicting GDS. A concern remains that the model has a large, potential outlier with the most recent observed data point (year 2008).

To further, consider models that might accommodate the most recent data, we restricted attention to the variables to ALTPI, ALPOP, ALLF, ALEMP, ALUEMP, and USGDP for which both the years 2008 and 2009 are available. None of the models appear to be suitable according to Cp or MSE criteria until there are four variables. However, we will consider Model H using ALEMP (Adjusted R^2 is 92.2 and Cp of 16) and Model I with ALEMP and USGDP (Adjusted R^2 is 93.7 and Cp 7.8). Note that Model I is similar to Model E, except that USGDP is substituted for ALGDP for which there is data for 2009 as well as 2008 available.

Of the two models, Model H seems to be slightly better in predicting the most recent DGS values, while Model I is a slightly better overall predictor. In both cases, the undeleted DGS models are fairly similar to the deleted case DGS* models, so only the former will be reported.

Model H has a single predictor, ALEMP so collinearity is not an issue. The residuals for the final two years are relatively small, though cases 37 (2006) and 38 (2007) are over predicted. The adjusted R^2 is only 91.4 and the Prediction R^2 is 92.2. The chief advantage of this model is that the most recent cases seem to have fairly small residuals. Residuals do seem to be increasing over time with this model and there is some indication of this in the residuals vs. fitted plots.

Model I contains both ALEMP and USGDP and has a moderate VIF of 5.2. The adjusted R^2 is 93.7 and Prediction R^2 is 92.6 for a good overall fit. The last two residuals are fairly large, though not identified as outliers. When analyzed with the deleted DGS (i.e., Model I*), the overall residual pattern is good, though the last four residuals are large in magnitude (cases 37, 38 are large and positive, while cases 39 and 40 are large and negative). As with Model H, there is an indication of a funnel in the residuals vs. fitted plot, which means that the residuals seem to be increasing over time.

From the foregoing analysis, it appears that variable ALEMP is the best single variable predictor for GDS (Model H). The addition of either ALGDP (Model E) or USGDP (Model I) provides a slightly improved model in overall prediction. All three models seem to have increasing prediction errors over time, though Model H seems to have smaller errors for the final data cases. Table 5 shows all three regression models along with their Mean Square Error (MSE).

Table 5 – Best Predictive Models

Model	Regression Equation	MSE
Model E	DGS = -408,841,076.1 + 449.3521635*ALEMP + 0.001661122*ALGDP	1.79253E+15
Model H	DGS = -791,003,505.1 + 739.161045*ALEMP	2.38061E+15
Model I	DGS = -518,710,897.7 + 536.5062355*ALEMP + 0.0000131493*USGDP	1.91136E+15

This may be the best that can be done in a case where the most recent data (2006 and later) may represent a fundamentally different economic environment than historical data collected prior to that time.

In addition to the regression equations for predicting Diesel Gallons sold which can be converted to Average Truck Trips per Day as outlined above. It should be noted that the equations only provide point estimates for the dependent variable given the reality of variability in the predictions generated by application of the regression models. It is most appropriate to also provide prediction intervals which illustrate the range of potential outcomes. The research team developed prediction interval equations for calculating the prediction intervals for each of the candidate models (i.e., Models E, H, and I). The basic methodology is outlined below along with the elements of the equations necessary for their calculation.

Prediction Interval Computations

The method of least squares is used to estimate the parameters b_j in the linear regression model given by

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

The b_j represent the model coefficients, the x_j represent the predictor variables, and there are k predictors. For the models chosen in this work, k is either 1 or 2. In the method of least squares, the bj have been chosen to minimize the total sum of squares error, SSE, for the residuals, $e_i = y_i - \hat{y}_i$.

The first expression can be used to predict a new value of the response y as a function of the inputs. Such a prediction has two distinct sources of uncertainty: the uncertainty associated with the mean response at that set of predictors and the uncertainty associated with each response y. The former term consists of a combination of uncertainties from each of the estimated model coefficients. The computations are straightforward but somewhat complicated. We have provided a simple matrix formula which is convenient for computer calculation. The variances of these coefficients for $b^T = (b_0, b_1, b_2)$ can be obtained from the estimated variance-covariance matrix Var[b] given by:

$$Var[b] = MSE\ (X^T X)^{-1} = MSE\ \ V = MSE \begin{bmatrix} v_{00} & v_{01} & v_{02} \\ v_{10} & v_{11} & v_{12} \\ v_{20} & v_{21} & v_{22} \end{bmatrix}$$

The matrix V is symmetric, and therefore $v_{ij} = v_{ji}$. The variance of the estimator $\hat{y}$ can be expressed in the following formula (which takes advantage of the symmetrical terms):

$$Var[\hat{y}(x_0)] = MSE\ x_0^T V x_0 = MSE(v_{00} + x_{1,0}^2 v_{11} + x_{2,0}^2 v_{22} + 2x_{1,0}v_{01} + 2x_{2,0}v_{02} + 2x_{1,0}x_{2,0}v_{12})$$

Where the input vector is given by $x_0^T = (1, x_{1,0}, x_{2,0})$. The middle expression (using matrix expressions) is implemented by Excel using MMULT and TRANSPOSE commands. Model H has only one predictor (ALEMP), so there are fewer terms to compute. Table 5 (page 22) provided the details used in the equations, this table provides the model coefficients, identifies the variables that are used in each model, and the MSE for each model. Models E and I have two terms, while model H only has the single variable, ALEMP.

Tables 6 through 8 provide the terms for the V matrices for each model. The matrix V is 3x3 for models E and I, while the matrix V is only 2x2 for Model H.

Table 6 - Variance covariance matrix V for Model E

8.527012445	-6.04377E-06	2.85812E-11
-6.04377E-06	4.3622E-12	-2.16708E-17
2.85812E-11	-2.16708E-17	1.24574E-22

Table 7 - Variance covariance matrix V for Model H

1.965252998	-1.06833E-06
-1.06833E-06	5.89582E-13

Table 8 - Variance covariance matrix V for Model I

6.345332072	-4.32822E-06	2.11519E-13
-4.32822E-06	3.01577E-12	-1.57424E-19
2.11519E-13	-1.57424E-19	1.02145E-26

A spreadsheet was also developed which will facilitate calculation of the estimates and prediction intervals for Diesel Gallons sold and Average Truck Trips per Day for a given scenario. Section 6.0 provides examples which illustrate the application of these models and the spreadsheet.

Extrapolation Limits

A linear model fit over a particular range of data is usually considered a useful approximation to a more complicated, but unknown function. There is considerable mathematical justification for the usefulness of linear approximations. In addition, regression diagnostics are used to judge the suitability of the model so obtained. In particular, the plot of residuals versus fitted values is used to check for any indication of a systematic pattern of residuals. Such a pattern would suggest that the model is inappropriate over the range of the data.

While the usefulness of the model can be judged within the range of the data, no such test is available when extrapolating beyond the range of the available data. The absence of such a check means that there is no guarantee that the form of the underlying model is still adequate outside the range of the observable data.

The models proposed in this research relate predictor variables that have been increasing over several decades and whose values will likely continue to increase in the future so that predictions in the relatively near future will be extrapolation beyond the range of the available data. The suggested limits for these predictions are based upon an error bounding approach outlined below. While there is no guarantee that the observed relation will hold when used beyond the limits of the data, this approach provides guidance based upon one type of deviation from linearity.
To estimate an allowable extrapolation range, we suppose that the fitted first order model has a second order term whose magnitude is small enough that it is not detectable within the range of the data. For generality we conduct the analysis with scaled variables Z using:

$$Z = \frac{X - M}{H}$$
$$H = \frac{X_{max} - X_{min}}{2}$$
$$M = \frac{X_{max} + X_{min}}{2}$$

This coding will have the values +1 at the maximum observed value X_{max} and -1 at the observed value X_{min}. Thus we suppose that there may be a second order term that has been omitted from the analysis:

$$Y = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \epsilon$$

By our assumption the term $\beta_2 Z^2$ is too small over the range (-1, +1) in the coded variable Z. As Z increases, the magnitude of this omitted term increases rapidly. We can compute the value of Z at which the additional value of the missing term increases to a factor γ of the term omitted at the endpoint Z_0=+1:

$$\beta_2 Z^2 - \beta_2 Z_0^2 = \gamma \beta_2 Z_0^2$$
$$Z = Z_0 \sqrt{1 + \gamma}$$

This limit does not depend upon the value of the unknown second order coefficient, as it bounds the value that would be obtained relative to the maximum that would be obtained at the end of the data range. The limit to extrapolation would be given by the formula

$$X_U = M + H\sqrt{1 + \gamma}$$

That limits the magnitude of the omitted second order term. To illustrate, Table 9 summarizes the computation of the upper bound for each of the independent variables.

Table 9 – Computation of Variable Upper Bounds

Variable	X(min)	X(max)	M	H	γ	X_U	% Change from Maximum
ALEMP (Millions)	**1.396**	**2.106**	1.751	0.355	3	**2.461**	17%
ALGDP (Billions)	**12.455**	**170.014**	91.2345	78.7795	2	**227.685**	34%
USGDP (Trillions)	**1.038**	**14.369**	7.7035	6.6655	2	**19.248**	34%

The **γ** values were selected to provide upper bounds that would reflect an annual growth of approximately 5% for at least 5 years into the future. Based on these calculations and the logic outlined above, the UAHuntsville research team recommends that the boundaries for the input variables be within the values shown in Table 10 to ensure that the estimates generated by models (i.e., E, H, and I) do not involve unwarranted extrapolation. The lower boundaries were set at the minimum value observed for each of the input variables. These limits are provided in the spreadsheet accompanying this report (see Appendix D). This spreadsheet generates the estimated Diesel Gallons Sold and ultimately Average Truck Trips Traveled per Day along with their 95% Prediction Interval Limits based on the input variables (i.e., ALEMP, ALGDP, and USGDP).

Table 10 – Input Variable Boundaries

Variable	Lower Bound	Upper Bound
ALEMP	1,396,000	2,461,000
ALGDP	12,455,000,000	227,685,000,000
USGDP	1,038,000,000,000	19,248,000,000,000

It should also be noted that the UAHuntsville research team recommends that the models be revisited and updated within three to four years to ensure that the model coefficients reflect the most current trends in the independent variables (ALEMP, ALGDP, and USGDP) and the dependent variable (DGS).

5.0 Task 4: Validation of the Final Model

Given the limited size of the data set, the analysis team determined that the preferred method of validation was using the Prediction Error Sum of Squares (PRESS) statistic and R-squared (Predicted). These are deletion methods, which "drop" a single data point and use the remaining points to develop a regression model, which is then used to predict the "dropped" value. In the case of the PRESS statistic, the smaller the values are, the better. The $R^2_{prediction}$ is based on the PRESS and provides a clear indicator of performance by using a percentage scale. In this specific model, the higher the value of $R^2_{prediction}$ the better the model is for prediction. The PRESS and $R^2_{prediction}$ for the candidate models are shown in Table 11 (next page). As evidenced in the chart, all have very high R-squared Predicted values (i.e., greater than 0.90).

Table 11 – PRESS and $R^2_{prediction}$ for Best Predictive Models

Model	PRESS	$R^2_{prediction}$
Model E	6.872983E+16	93.10%
Model H	8.610266E+16	91.41%
Model I	7.473853E+16	92.55%

6.0 Task 5: Development of Implementation Plan

It is the recommendation of the UAHuntsville research team that ALDOT track the Diesel Tax Collections which can be easily converted to Diesel Gallons Sold per year. Then, what is needed to convert this data to the number of miles of truck travel on Alabama highways is an estimate of the miles per gallon rating for heavy trucks. The U.S. Department of Energy website *http://www1.eere.energy.gov/vehiclesandfuels/facts/2005/fcvt_fotw372.html* lists the average miles per gallon for different types of trucks as shown in Table 12 (next page). For the evaluations performed in this research project average fuel economy in miles per gallon was calculated from 5.5 to 6.5. This range encompasses most of the heavy trucks on the roadways. To provide a level of validation for this approach the total number of trucks from the Alabama Statewide Freight Model and Action Plan was used to calculate an average miles per gallon of 5.4 MPG. This validation approach is very close to the 5.5 MPG found in Table 11 (top of this page) for the heavy trucks and also provides evidence that the Alabama Statewide Freight Model and Action Plan provides a conservative, but accurate, estimate of trucks on Alabama highways.

It is felt that Alabama Diesel Tax Collections provides the best available source of data to represent Freight Traffic in the state each year. The Alabama Diesel Tax Collections can be converted to Diesel Gallons Sold by dividing by the Diesel Tax Collections for a given year by the Diesel Tax Rate for that year as shown in Appendix B. This data could then be used to estimate the number of trucks by multiplying the average miles per gallon and Gallons of Diesel Sold to obtain total miles driven per year.

Specifically, the method for calculating trucks on the road is accomplished in two steps:

1) Calculate the average miles per truck trip in the state.
    a. The number of truck trips is taken specifically from the Alabama Statewide Freight Model. The calculation uses the total gallons purchased (gallons/year) times fuel economy (mile per gallon), divided by 250 work days (days/year) divided by 285,000 trips (trips per day). The units are:
        i. Gallon/year * miles/gallon = miles per year
        ii. Miles per year / (days/year) = miles per day
        iii. Miles per day / (trips/day) = miles per trip

Utilizing this approach to obtain the total miles driven per year, the Alabama Statewide Freight Model developed for the Alabama Department of Transportation could be used to determine the number of trucks on the road. From the model, the total number of one-way truck trips expected per day was approximately 285,000 (UAHuntsville and JRWA, 2010). Using the gallons of diesel purchased, average fuel economy of the trucks, and a factor of 250 trucking days per year (Mobility.TAMU.edu, 2010), the average truck trip length in the state is calculated to be 53 miles, or 106 miles round trip. This average is intended to account for some truck movements, which are quite short, i.e. across metropolitan areas, versus other truck movements that cross the entire state, potentially up to 400 miles.

2) Calculate a projected number of trucks on the roadways.
    a. We make the assumption that the miles per trip will remain constant due to the fact that locations generally change infrequently.
    b. The projected DGS can be used to determine the number of trips per day, essentially trucks on the road.
    c. Trips per day, or trucks on the road, would be equal to:

$$\frac{\dfrac{\text{DGS (predicted)} * \text{MPG (5.5 to 6.5 mpg)}}{\text{250 days per year}}}{\text{53 miles per trip}}$$

Table 12 – Truck Fuel Economy by Size Class

Truck Fuel Economy by Size Class			
Manufacturer's GVW Class	**1992 TIUS**	**1997 TIUS**	**2002 TIUS**
1) 6,000 lb and less	17.2	17.1	17.6
2) 6,001–10,000 lb	13.0	13.6	14.3
Light truck subtotal	15.7	15.8	16.2
3) 10,000–14,000 lb	8.8	9.4	10.5
4) 14,001–16,000 lb	8.8	9.3	8.5
5) 16,001–19,500 lb	7.4	8.7	7.9
6) 19,501–26,000 lb	6.9	7.3	7.0
Medium truck subtotal	7.3	7.8	8.0
7) 26,001–33,000 lb	6.5	6.4	6.4
8) 33,001 lb and over	5.5	5.7	5.7
Large truck subtotal	5.6	5.7	5.8

Source: U.S. Bureau of the Census, "2002 Vehicle Inventory and Use Survey," Microdata file, January 2005
"1997 Vehicle Inventory and Use Survey," 2000
"1992 Vehicle Inventory and Use Survey," 1995

Data generated by Stacy Davis, Oak Ridge National Laboratory, March 2005

Assuming the spatial distribution of population and employment remain constant into the future, i.e. the locations of the major origin and destination locations for freight movement remain relatively constant and there is not a great change in the number of through trucks, this 53 mile average trip length can be used to determine the number of truck trips expected on Alabama roadways in future years.

These models can be used to estimate the average number of truck trips per day for a given year. The following example illustrates how the models can be used by following a four-step process:

1. Estimate the three independent variables (ALEMP, ALGDP, and USGDP) – these estimates can be based on forecasts provided by another state or federal agency or, more likely, they are based on a what-if analysis (i.e., What is the impact on the average trips per day if ALEMP, ALGDP, and USGDP each increases by 5%, 6%, and 8% from 2008 levels).
    a. For this example, the value for each of the variables after the increase from 2008 levels will be:
        i. ALEMP = 2,146,626
        ii. ALGDP = 180,214,840,000
        iii. USGDP = 15,518,628,000,000

2. Input the estimates for ALEMP, ALGDP, and USGDP into each of the models to develop an estimate for DGS.
    a. Model E: DGS = -408,841,076 + 449.35*ALEMP + 0.0016611*ALGDP
        i. Estimate for DGS = 855,108,879
    b. Model H: DGS = -791,003,505 + 739.16*ALEMP
        i. Estimate for DGS = 795,699,034
    c. Model I: DGS = -518,710,897 + 536.51*ALEMP + 0.00001315*USGDP
        i. Estimate for DGS = 837,026,787

3. Calculate the Average Truck Trips per Day for the Example using the miles per gallon, number of weekdays in a year and average trip length from the Statewide Freight Plan.
    a. Average Truck Trips per Day = DGS*mpg/250/53
        i. Estimate using 5.5 miles per gallon
            1. Model E: ATTD = 354,951
            2. Model H: ATTD = 330,290
            3. Model I: ATTD = 347,445
        ii. Estimate using 6.5 miles per gallon
            1. Model E: ATTD = 419,487
            2. Model H: ATTD = 390,343
            3. Model I: ATTD = 410,617

4. Determine the range of truck trips per day from the models.
    a. Average Truck Trips per Day is expected to range from 330,290 to 419,487 in the next year.

Under the conditions of this example, with the projected increases in ALEMP, ALGDP and USGDP the Average Truck Trips per Day are estimated to be between 330,290 (at 5.5 mpg) and 419,487 (at 6.5 mpg). This example illustrates the variability in the estimates among the three models (in addition to the truck miles per gallon value used). The proper use of the models would be to select a single model for estimating the DGS (and ultimately the Average Truck Trips per Day) and then calculate the 95% prediction limits for that model. Table 13 (next page) shows output from a spreadsheet tool that estimates the Average Number of Truck Trips per day based on ALEMP, ALGDP, and USGDP. The spreadsheet output generates estimates for Diesel Gallons Sold and the Average Truck Trips per Day along with 95% prediction limits for each. For this example, as shown in Table 13 if Model H is selected then the estimated Average Truck Trips per Day at 5.5 miles per gallon would be 330,290 with a lower prediction limit of 287,048 and an upper prediction limit of 373,533.

Table 13 - Predictions Generated by Candidate Models E, H, and I

Spreadsheet for Estimating Average Truck Trips per Day (ALDOT Research Project 930-768 by UAHuntsville)			
The input values for the ALEMP, ALGDP, and USDGP should fall within the ranges shown below. This will limit inappropriate extrapolation in the models which would reduce the accuracy of the predictions and prediction limits			
Description of Variables/Data to be entered in Next Column	Input Data Below	RECOMMENDED DATA BOUNDARIES	
Alabama Employment - ALEMP	2,146,626	1,396,193	2,461,000
Alabama Gross Domestic/State Product - ALGDP	180,214,840,000	12,455,000,000	227,685,000,000
U.S. Gross Domestic Product - USGDP	15,518,628,000,000	15,518,628,000,000	19,248,000,000,000
Estimate(s) and 95% Prediction Limit Calculations			
Models	Est. Diesel Gallons Sold	Lower Prediction Limit	Upper Prediction Limit
Model E (DGS = -408,841,076 + 449.35*ALEMP + 0.0016611*ALGDP)	855,108,879	758,115,206	952,102,551
Model H (DGS = -791,003,505 + 739.16*ALEMP)	795,699,034	691,523,538	899,874,530
Model I (DGS = -518,710,897 + 536.51*ALEMP + 0.00001315*USGDP)	837,026,787	739,477,035	934,576,539
Estimated Average Truck Trips Traveled per Day (5.5 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	354,951	314,689	395,212
Model H	330,290	287,048	373,533
Model I	347,445	306,953	387,937
Estimated Average Truck Trips Traveled per Day (6.0 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	387,219	343,297	431,141
Model H	360,317	313,143	407,490
Model I	379,031	334,858	423,204
Estimated Average Truck Trips Traveled per Day (6.5 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	419,487	371,906	467,069
Model H	390,343	339,238	441,448
Model I	410,617	362,762	458,472

7.0 Conclusions

This study developed a set of three regression models for estimating the gallons of diesel fuel sold annually. This value, Diesel Gallons Sold (DGS), was then used to estimate the average number of truck trips per day, which was validated against the Statewide Freight Flow model. The three regression models allowed the UAHuntsville research team to create a spreadsheet tool that can be used to generate an estimate for the Average Truck Trips per Day based on the estimates/projections of the three independent variables (ALEMP, ALGDP, and USGDP) along with their 95% Prediction Limits. It is recommended that the models and spreadsheet be updated periodically (i.e., at least every 3 or 4 years) based on additional values for the independent variables to ensure that the models capture new trends.

8.0 References

Benjamin, J. (1986). A time-series forecast of average daily traffic volume. *Transportation Research, Part A, 20* (1), 51-60.

Black, W. R. (2006). *Freight Flows of Indiana.* Springfield: National Technical Information Service.

Black, W. R. (1997). *Transport Flows in the State of Indiana: Commodity Database Development and Traffic Assignment, Phase 2.* Indiana University. Bloomington: Transportation Research Center, Indiana University.

Brown, C., Kennedy, N., Wright, D., & Zak, W. (2003). Methodology for Estimating Vehicle Miles Traveled for Commercial Motor Vehicles at the State Level. *Transportation Research Record , 1830*, 72-76.

Fite, J. T., Taylor, G. D., Usher, J. S., English, J. R., & Roberts, J. N. (2002). Forecasting Freight Demand Using Economic Indices. *International Journal of Physical Distribution and Logistics Management, 32* (4), 299-308.

Fricker, J., & Kumapley, R. (2000). *Updating Procedure to Estimate and Forecast Vehicle Miles Traveled.* West Lafayette, IN: Purdue University.

Horowitz, A. J., & Farmer, D. D. (1999). Statewide Travel Forecasting Practice: A Critical Review. *Transportation Research Record, 1685*, 13-20.

Johnson, D. L. (2000). *20-Year Traffic Forecasting Factors.* Pierre: South Dakota Department of Transportation.

Liu, F., Kaiser, R. G., Zekkos, M., & Allison, C. (2006). Growth Forecasting of Vehicle Miles of Travel at County and Statewide Levels. *Transportation Research Record, 1957*, 56-65.

Mobility.TAMU.edu (2010). http://tti.tamu.edu/documents/mobility_report_2009_wappx.pdf

Pendyala, R. M., Shankar, V. N., & McCullough, R. G. (2000). Freight Travel Demand Modeling: Synthesis of Approaches and Development of a Framework. *Transportation Research Record , 1725*, 9-16.

Polzin, S. E., Chu, X., & Toole-Holt, L. (2004). Forecasts of Future Vehicle Miles of Travel in the United States. *Transportation Research Record , 1895*, 147-155.

Sabry, M., Abd–El-Latif, H., Yousef, S., & Badra, N. (2007). A Time-Series Forecasting of Average Daily Traffic Volume. *Australian Journal of Basic and Applied Sciences , 1* (4), 386-394.

Sorratini, J. A., & Smith, R. L. (2000). Development of a Statewide Truck Trip Forecasting Model Based on Commodity Flows and Input-Output Coefficients. *Transportation Research Record , 1707*, 49-55.

UAHuntsville and JRWA (2010). Alabama Statewide Freight Study and Action Plan.

Waliszewski, J. M., Ahanotu, D. N., & Fischer, M. J. (2004). Comparison of Commodity Flow Forecasting Techniques in Montana. *Transportation Research Record , 1870*, 1-9.

Yang, C. H., Chow, J. Y., & Regan, A. C. (2009). State-of-the Art of Freight Forecasting Modeling: Lessons Learned and the Road Ahead. *Proceedings, 88th Annual Meeting of the Transportation Research Board.* Washington DC: Transportation Research Board.

9.0 Project Team

Phillip A. Farrington, Ph.D., – *Principal Investigator*
Gregory A. Harris, Ph.D., P.E. – *Co-Investigator*
Michael D. Anderson, Ph.D., P.E.
Niles C. Schoening, Ph.D.
James T. Simpson, Ph.D.
James J. Swain, Ph.D.
Jeff Thompson
Lauren Jennings Neppel
Lisa S. Blanchard
Karen E. Yarbrough

Appendix A – Description of Variables Used in Study

Description of Candidate Independent Variables:

Alabama Truck Vehicle Miles Traveled (ALVMT) – Annual miles travelled by trucks within the state of Alabama as determined from the Statewide Traffic Count Database.

Alabama Diesel Tax Collections (DT) – the total taxable gallons sold per year paid to the Alabama Department of Revenue.

Alabama Motor Vehicle Fuel Taxes (VFT) – revenues collected from taxes on the sales of motor fuels in the state of Alabama.

Description of Candidate Dependent Variables:

Alabama Population (ALPOP) – includes all persons living in a geographical area.

Alabama Total Personal Income (ALTPI) – Income received by persons from all sources. It includes income received from participation in production as well as from government and business transfer payments. It is the sum of compensation of employees (received), supplements to wages and salaries, proprietors' income with inventory valuation adjustment (IVA) and capital consumption adjustment (CCAdj), rental income of persons with CCAdj, personal income receipts on assets, and personal current transfer receipts, less contributions for government social insurance.

Alabama Labor Force (ALLF) – The labor force includes all persons classified as employed or unemployed.

Alabama Employment (ALEMP) – this variable includes all persons 16 years and over in the civilian non-institutional population who, during the reference week, (a) did any work at all (at least 1 hour) as paid employees; worked in their own business, profession, or on their own farm, or worked 15 hours or more as unpaid workers in an enterprise operated by a member of the family; and (b) all those who were not working but who had jobs or businesses from which they were temporarily absent because of vacation, illness, bad weather, childcare problems, maternity or paternity leave, labor-management dispute, job training, or other family or personal reasons, whether or not they were paid for the time off or were seeking other jobs. Each employed person is counted only once, even if he or she holds more than one job.

Alabama Unemployment (ALUEMP) – this variable includes all persons aged 16 years and older who had no employment during the reference week, were available for work, except for temporary illness, and had made specific efforts to find employment sometime during the 4-week period ending with the reference week. Persons who were waiting to be recalled to a job from which they had been laid off need not have been looking for work to be classified as unemployed.

Alabama Employment by Industry (ATE) – the number of people employed in a particular industry with industries presented by NAICS or SIC code.

Alabama Value of Shipments (ALTVS) – total value of shipments includes the received or receivable net selling values, “Free on Board” (FOB) plant (exclusive of freight and taxes), of all products shipped, both primary and secondary. Included are all items made by or for the establishments from material owned by it, whether sold, transferred to other plants of the same company, or shipped on consignment.
Total Value of Shipments data was obtained from the U.S. Census Bureau, Economic Census of Manufacturing for the years 1969 – 2007. However some years were unavailable (i.e., 1970, 1974, 1977-1979, 1983, 1986, 1988-1989, 1993-1994, and 1998). To estimate a value for the missing years, a calculation was made using Gross Domestic Product (GDP) and the Total Value of Shipments for the year prior to the missing year. For example, the value for 1970 was unavailable. Therefore, The Total Value of Shipments for 1969 was multiplied by the GDP for 1970. This value was then added to the Total Value of Shipments value for 1969 to obtain the Total Value of Shipments for 1970.

Alabama GDP (ALGDP) – A measurement of a state's output; it is the sum of value added from all industries in the state.

Southeast States GDP (SEGDP) – A measurement of a state's output; it is the sum of value added from all industries in the state.

U.S. GDP (USGDP) – This variable is the market value of goods and services produced by labor and property in the United States regardless of nationality.

Appendix B – Diesel Tax Collection Data

Table B1 - Diesel Tax Data & Conversion to Diesel Gallons Sold

*Beginning in 2008 IFTA taxes were no longer included with Diesel Fuel Collections

Year	Diesel Fuel Tax	Tax Rate	Diesel Gallons Sold
1970	9,102,156.88	0.08	113,776,961
1971	10,322,592.10	0.08	129,032,401
1972	14,733,730.42	0.08	184,171,630
1973	17,444,369.98	0.08	218,054,625
1974	18,630,607.83	0.08	232,882,598
1975	17,929,908.57	0.08	224,123,857
1976	20,257,866.53	0.08	253,223,332
1977	22,677,402.98	0.08	283,467,537
1978	25,571,778.01	0.08	319,647,225
1979	27,907,676.45	0.08	348,845,956
1980	29,030,646.96	0.12	241,922,058
1981	40,222,771.52	0.12	335,189,763
1982	40,828,355.04	0.12	340,236,292
1983	43,682,106.29	0.12	364,017,552
1984	49,410,544.07	0.12	411,754,534
1985	50,826,172.69	0.12	423,551,439
1986	53,292,466.46	0.12	444,103,887
1987	55,506,672.09	0.12	462,555,601
1988	60,625,183.45	0.12	505,209,862
1989	63,722,933.87	0.12	531,024,449
1990	67,063,028.63	0.12	558,858,572
1991	66,203,564.93	0.12	551,696,374
1992	74,863,040.60	0.17	440,370,827
1993	97,502,011.15	0.17	573,541,242
1994	98,291,671.96	0.17	578,186,306
1995	103,994,929.35	0.17	611,734,879
1996	128,795,732.19	0.17	757,621,954
1997	105,675,049.07	0.17	621,617,936
1998	111,276,930.94	0.17	654,570,182
1999	115,019,865.53	0.17	676,587,444
2000	120,873,767.04	0.17	711,022,159
2001	118,751,693.49	0.17	698,539,373
2002	117,824,673.80	0.17	693,086,316
2003	119,564,453.68	0.17	703,320,316
2004	128,913,728.92	0.17	758,316,052
2005	149,602,200.15	0.19	787,380,001
2006	159,780,250.35	0.19	840,948,686
2007	155,521,723.50	0.19	818,535,387
2008*	135,802,012.50	0.19	714,747,434
2009	119,541,443.57	0.19	629,165,492

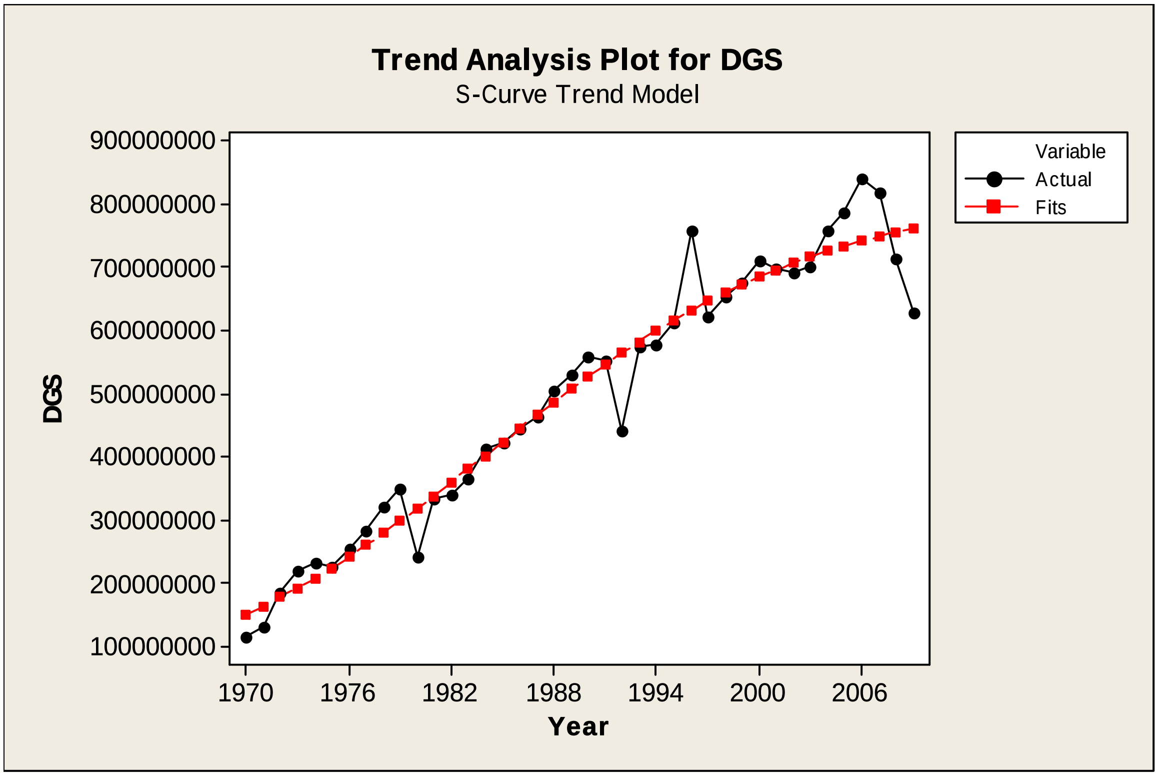
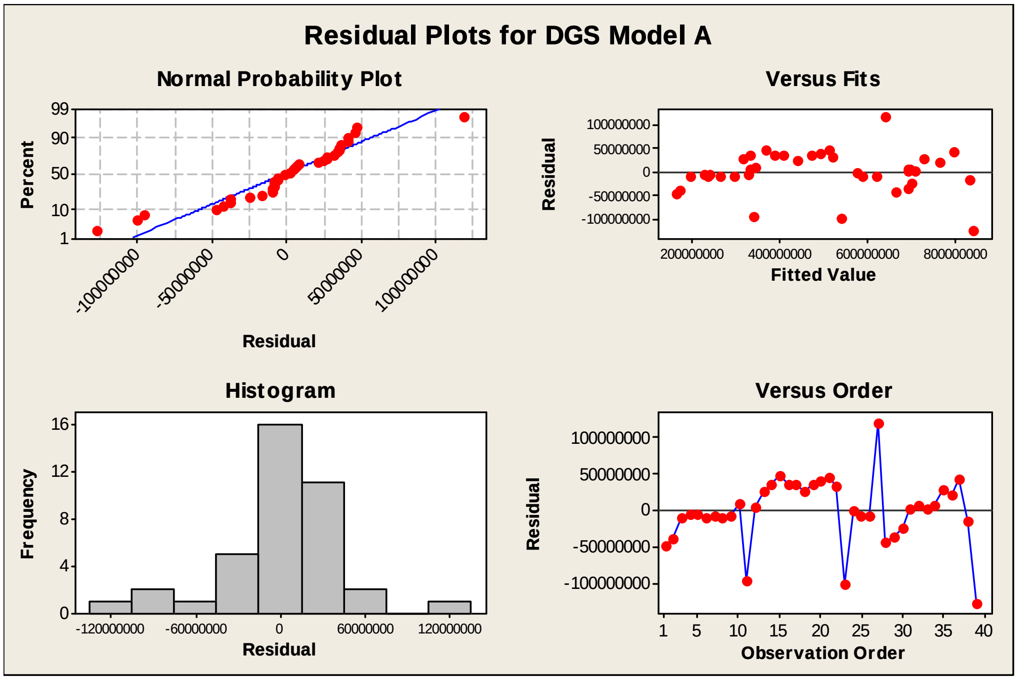


Figure B1 –Trend Analysis Plot of Annual Diesel Gallons Sold (1970 to 2009)

Appendix C – Statistical Analysis

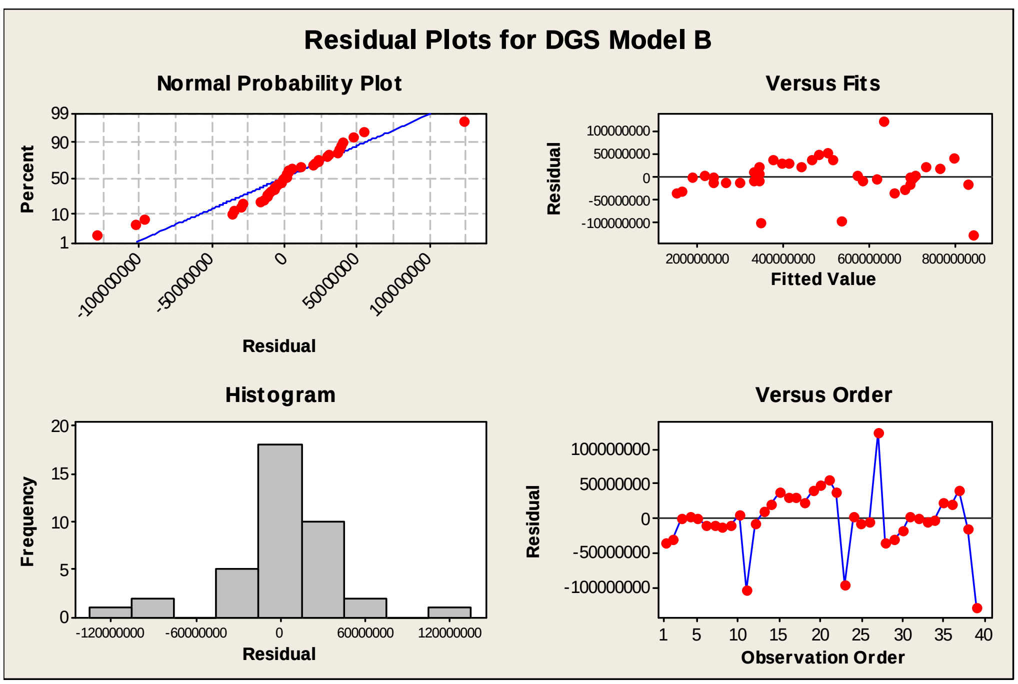
Model A: DGS versus ATE

```
The regression equation is
DGS = - 6.22E+08 + 554 ATE

39 cases used, 1 cases contain missing values

Predictor        Coef    SE Coef       T      P    VIF
Constant   -621724689   40048939  -15.52  0.000
ATE            554.09      19.63   28.22  0.000  1.000

S = 44836633   R-Sq = 95.6%   R-Sq(adj) = 95.4%
PRESS = 8.354039E+16   R-Sq(pred) = 95.01%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       1  1.60109E+18  1.60109E+18  796.43  0.000
Residual Error  37  7.43820E+16  2.01032E+15
Total           38  1.67547E+18

Unusual Observations
Obs      ATE        DGS        Fit    SE Fit    Residual  St Resid
 11  1731866  241922058  337888454   8981758   -95966396    -2.18R
 23  2097425  440370827  540441781   7397111  -100070954    -2.26R
 27  2275108  757621954  638894515   8905688   118727440     2.70R
 39  2640717  714747434  841475546  14369700  -126728112    -2.98R

R denotes an observation with a large standardized residual.
```



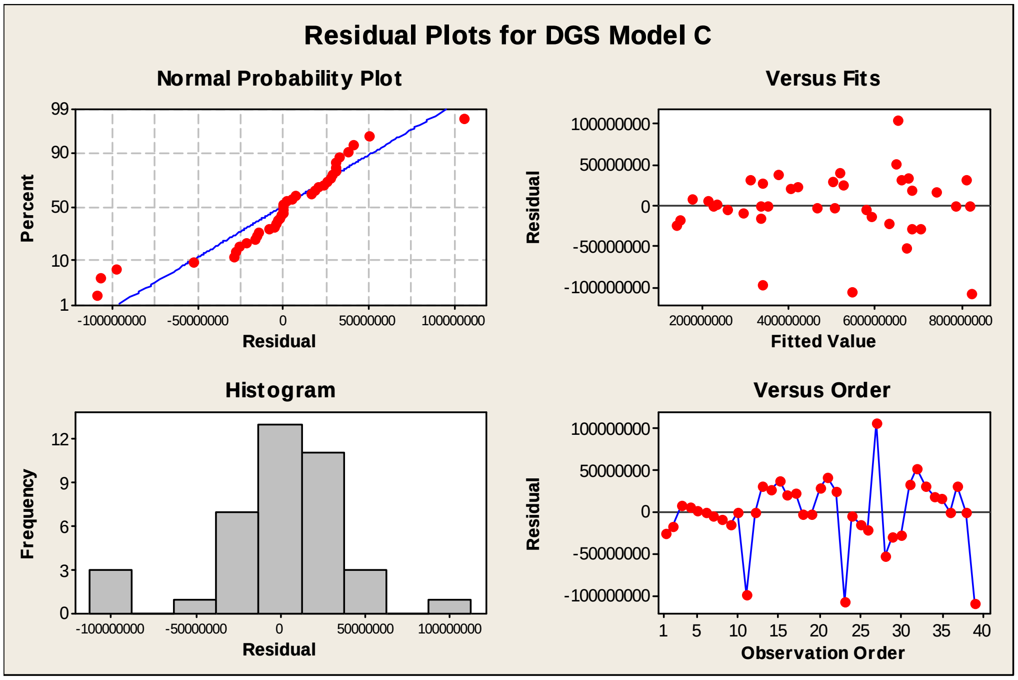
Model B: DGS versus ATE, ALPOP

```
The regression equation is
DGS = - 9.05E+08 + 441 ATE + 125 ALPOP

39 cases used, 1 cases contain missing values

Predictor         Coef    SE Coef      T      P     VIF
Constant    -904848128  301308089  -3.00  0.005
ATE              441.0      120.9   3.65  0.001  37.840
ALPOP            124.8      131.7   0.95  0.349  37.840

S = 44898039   R-Sq = 95.7%   R-Sq(adj) = 95.4%
PRESS = 8.438022E+16   R-Sq(pred) = 94.96%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       2  1.60290E+18  8.01449E+17  397.58  0.000
Residual Error  36  7.25700E+16  2.01583E+15
Total           38  1.67547E+18

Source  DF       Seq SS
ATE      1  1.60109E+18
ALPOP    1  1.81196E+15

Unusual Observations
Obs      ATE        DGS        Fit    SE Fit    Residual  St Resid
 11  1731866  241922058  344905865  11648091  -102983807    -2.38R
 23  2097425  440370827  536731955   8377267   -96361128    -2.18R
 27  2275108  757621954  633948325  10331805   123673629     2.83R
 39  2640717  714747434  843482608  14544271  -128735174    -3.03R

R denotes an observation with a large standardized residual.
```



MODEL C: DGS versus ATE, ALTPI, ALGDP

```
The regression equation is
DGS = - 7.03E+08 + 587 ATE - 11.1 ALTPI + 0.0102 ALGDP

39 cases used, 1 cases contain missing values

Predictor         Coef     SE Coef      T      P      VIF
Constant    -703464797   146467146  -4.80  0.000
ATE              587.1       103.1   5.69  0.000   30.331
ALTPI          -11.134       4.820  -2.31  0.027  931.251
ALGDP         0.010171    0.004686   2.17  0.037  999.288

S = 42749360   R-Sq = 96.2%   R-Sq(adj) = 95.9%
PRESS = 8.142654E+16   R-Sq(pred) = 95.14%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       3  1.61151E+18  5.37168E+17  293.93  0.000
Residual Error  35  6.39628E+16  1.82751E+15
Total           38  1.67547E+18

Source  DF       Seq SS
ATE      1  1.60109E+18
ALTPI    1  1.81087E+15
ALGDP    1  8.60834E+15

Unusual Observations
Obs      ATE        DGS        Fit    SE Fit    Residual  St Resid
 11  1731866  241922058  339613116   9349003   -97691058    -2.34R
 23  2097425  440370827  547601703   8586029  -107230876    -2.56R
 27  2275108  757621954  652017149  12326641   105604805     2.58R
 39  2640717  714747434  823190262  19681742  -108442827    -2.86R

R denotes an observation with a large standardized residual.
```



MODEL D: DGS versus ATE, ALTPI, ALGDP, SEGDP, USGDP

The regression equation is
DGS = - 8.72E+08 + 678 ATE - 25.7 ALTPI + 0.00987 ALGDP - 0.000594 SEGDP
+ 0.000268 USGDP

Predictor	Coef	SE Coef	T	P	VIF
Constant	-872087818	185656325	-4.70	0.000	
ATE	678.1	135.6	5.00	0.000	58.456
ALTPI	-25.702	8.253	-3.11	0.004	3040.071
ALGDP	0.009868	0.004497	2.19	0.035	1024.950
SEGDP	-0.0005941	0.0004126	-1.44	0.159	2331.800
USGDP	0.0002676	0.0001124	2.38	0.023	4746.250

S = 40506551   R-Sq = 96.8%   R-Sq(adj) = 96.3%
PRESS = 7.948012E+16   R-Sq(pred) = 95.26%

Analysis of Variance

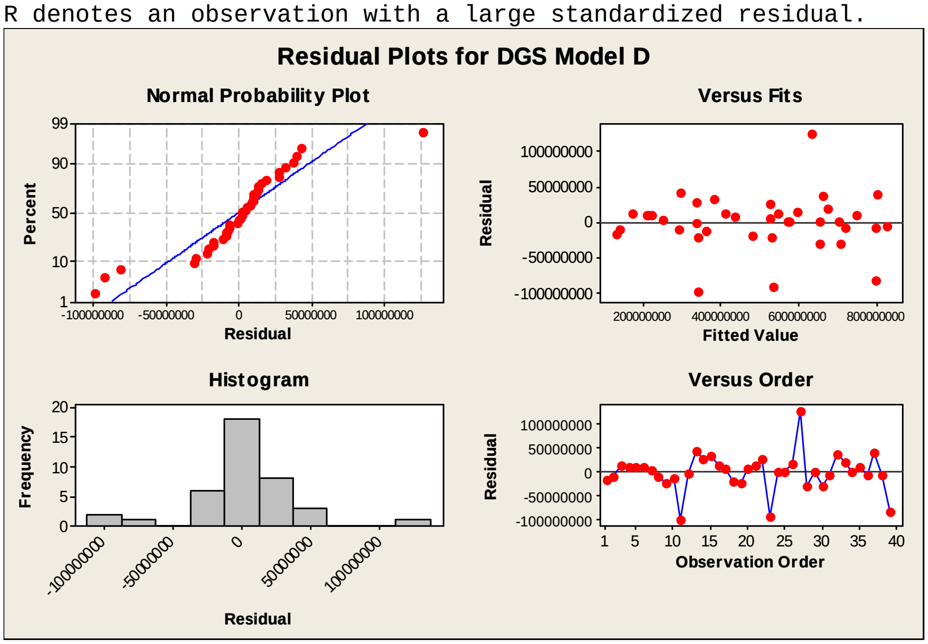
Source	DF	SS	MS	F	P
Regression	5	1.62132E+18	3.24264E+17	197.63	0.000
Residual Error	33	5.41458E+16	1.64078E+15		
Total	38	1.67547E+18			

Source	DF	Seq SS
ATE	1	1.60109E+18
ALTPI	1	1.81087E+15
ALGDP	1	8.60834E+15
SEGDP	1	5.21324E+14
USGDP	1	9.29568E+15

Unusual Observations

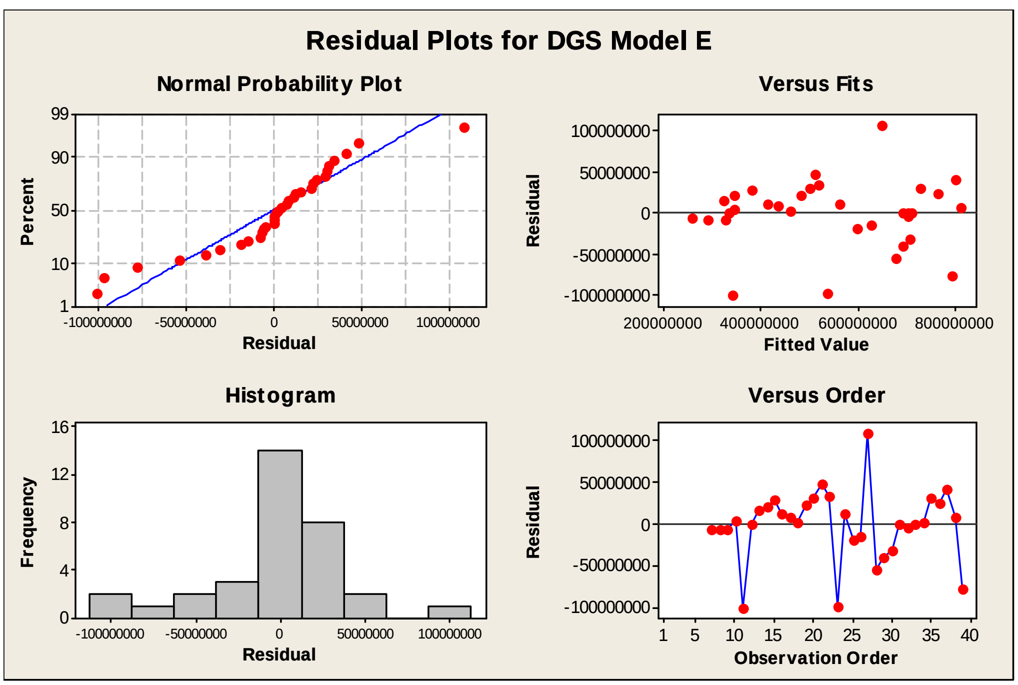
Obs	ATE	DGS	Fit	SE Fit	Residual	St Resid
11	1731866	241922058	341104592	8965809	-99182534	-2.51R
23	2097425	440370827	533471662	17894049	-93100835	-2.56R
27	2275108	757621954	630999041	14548632	126622913	3.35R
39	2640717	714747434	797039813	23177604	-82292379	-2.48R

R denotes an observation with a large standardized residual.



MODEL E: DGS versus ALEMP, ALGDP

```
The regression equation is
DGS = - 4.09E+08 + 449 ALEMP + 0.00166 ALGDP

33 cases used, 7 cases contain missing values

Predictor         Coef    SE Coef      T      P    VIF
Constant    -408841076  123632267  -3.31  0.002
ALEMP           449.35      88.43   5.08  0.000  7.364
ALGDP        0.0016611  0.0004725   3.52  0.001  7.364

S = 42338297   R-Sq = 94.6%   R-Sq(adj) = 94.2%
PRESS = 6.872983E+16   R-Sq(pred) = 93.10%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       2  9.42367E+17  4.71184E+17  262.86  0.000
Residual Error  30  5.37759E+16  1.79253E+15
Total           32  9.96143E+17

Source  DF       Seq SS
ALEMP    1  9.20217E+17
ALGDP    1  2.21501E+16

Unusual Observations
Obs    ALEMP        DGS        Fit    SE Fit    Residual  St Resid
 11  1538910  241922058  342481810  11538215  -100559752     -2.47R
 23  1809337  440370827  537825661   7720939   -97454834     -2.34R
 27  1992652  757621954  649253332  13148894   108368622      2.69R
 39  2044406  714747434  792231129  23268751   -77483695     -2.19RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```



Model F: DGS versus ALEMP, ALGDP, ALTPI

The regression equation is
DGS = - 4.54E+08 + 459 ALEMP + 0.00894 ALGDP - 7.73 ALTPI

33 cases used, 7 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	-454151834	125021204	-3.63	0.001	
ALEMP	458.64	86.94	5.28	0.000	7.402
ALGDP	0.008937	0.004927	1.81	0.080	832.631
ALTPI	-7.735	5.215	-1.48	0.149	836.773

S = 41516142   R-Sq = 95.0%   R-Sq(adj) = 94.5%
PRESS = 6.526980E+16   R-Sq(pred) = 93.45%

Analysis of Variance

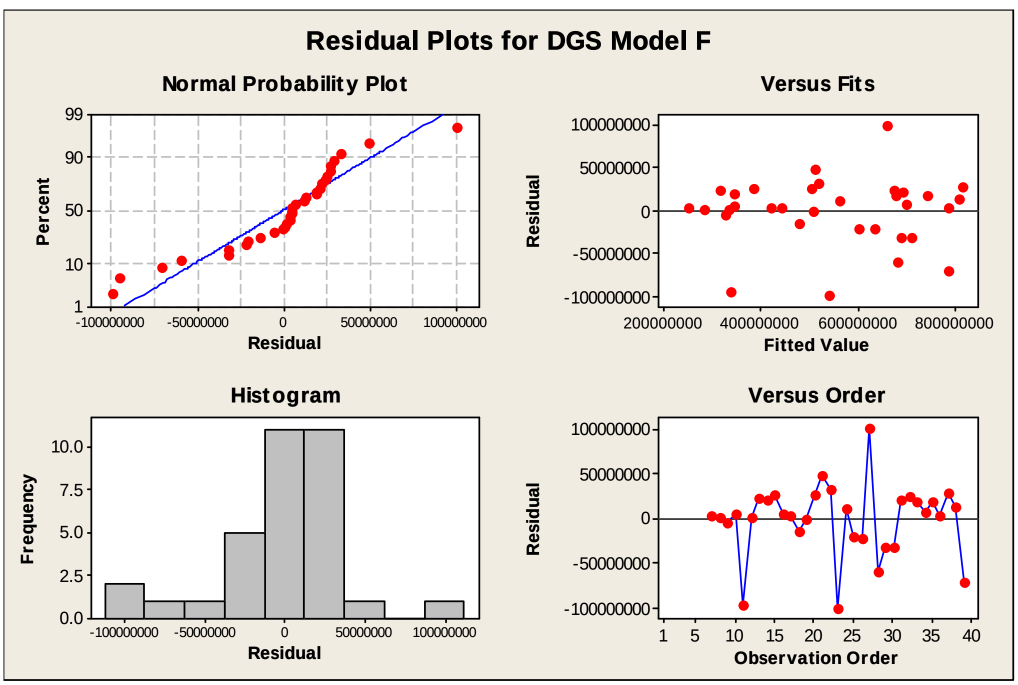
Source	DF	SS	MS	F	P
Regression	3	9.46159E+17	3.15386E+17	182.98	0.000
Residual Error	29	4.99841E+16	1.72359E+15		
Total	32	9.96143E+17			

Source	DF	Seq SS
ALEMP	1	9.20217E+17
ALGDP	1	2.21501E+16
ALTPI	1	3.79183E+15

Unusual Observations

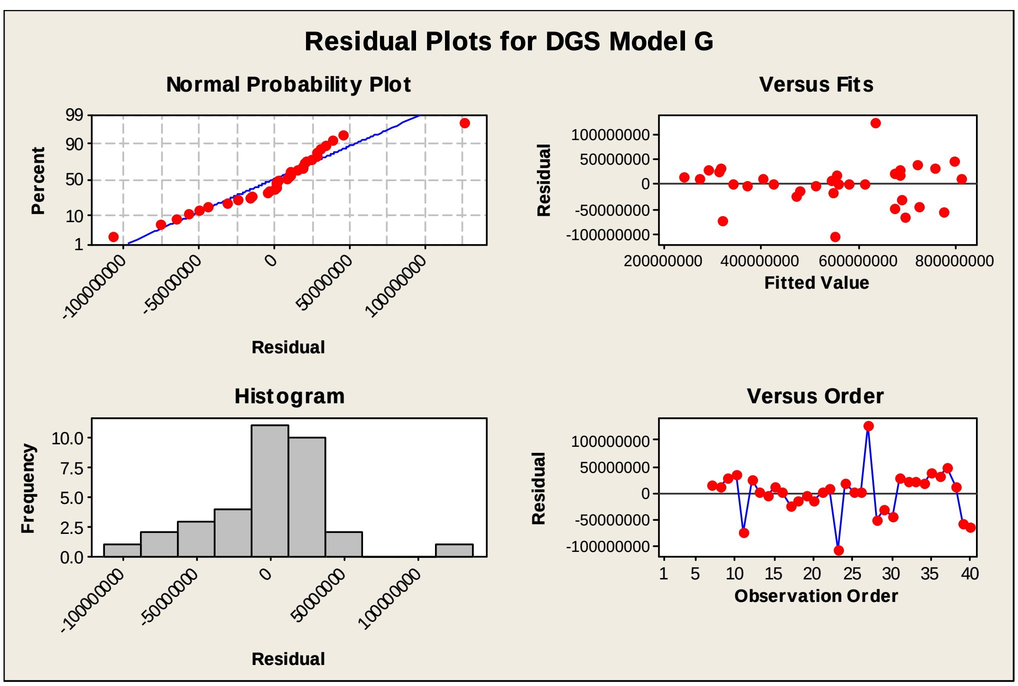
Obs	ALEMP	DGS	Fit	SE Fit	Residual	St Resid
11	1538910	241922058	337366937	11828024	-95444879	-2.40R
23	1809337	440370827	539982673	7709415	-99611846	-2.44R
27	1992652	757621954	656868162	13878093	100753792	2.57R
39	2044406	714747434	785307866	23289450	-70560432	-2.05R

R denotes an observation with a large standardized residual.



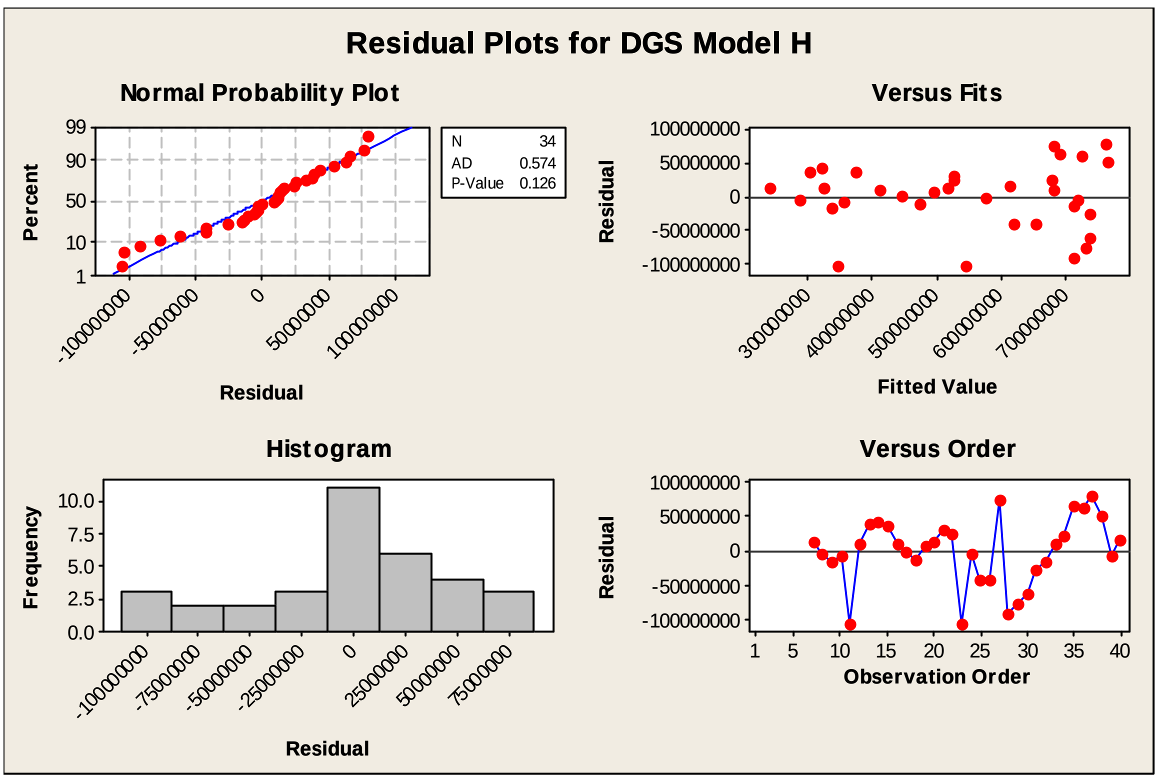
Model G: DGS versus ALLF, ALPOP, USGDP

```
The regression equation is
DGS = 2.05E+09 + 810 ALLF - 839 ALPOP + 0.000063 USGDP

34 cases used, 6 cases contain missing values

Predictor         Coef      SE Coef       T       P      VIF
Constant    2047980175    962636333    2.13   0.042
ALLF             810.4        119.8    6.76   0.000   10.545
ALPOP           -838.9        293.2   -2.86   0.008  119.227
USGDP       0.00006269   0.00001807    3.47   0.002   84.823

S = 43893819   R-Sq = 94.2%   R-Sq(adj) = 93.7%
PRESS = 7.706340E+16   R-Sq(pred) = 92.32%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       3  9.45069E+17  3.15023E+17  163.51  0.000
Residual Error  30  5.78000E+16  1.92667E+15
Total           33  1.00287E+18

Source  DF       Seq SS
ALLF     1  9.15381E+17
ALPOP    1  6.48457E+15
USGDP    1  2.32028E+16

Unusual Observations
Obs     ALLF        DGS         Fit    SE Fit    Residual  St Resid
 23  1943033  440370827   547841770   9216224  -107470943     -2.50R
 27  2086493  757621954   631112768  13945801   126509186      3.04R
 40  2112566  629165492   695071936  26228851   -65906444     -1.87 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```



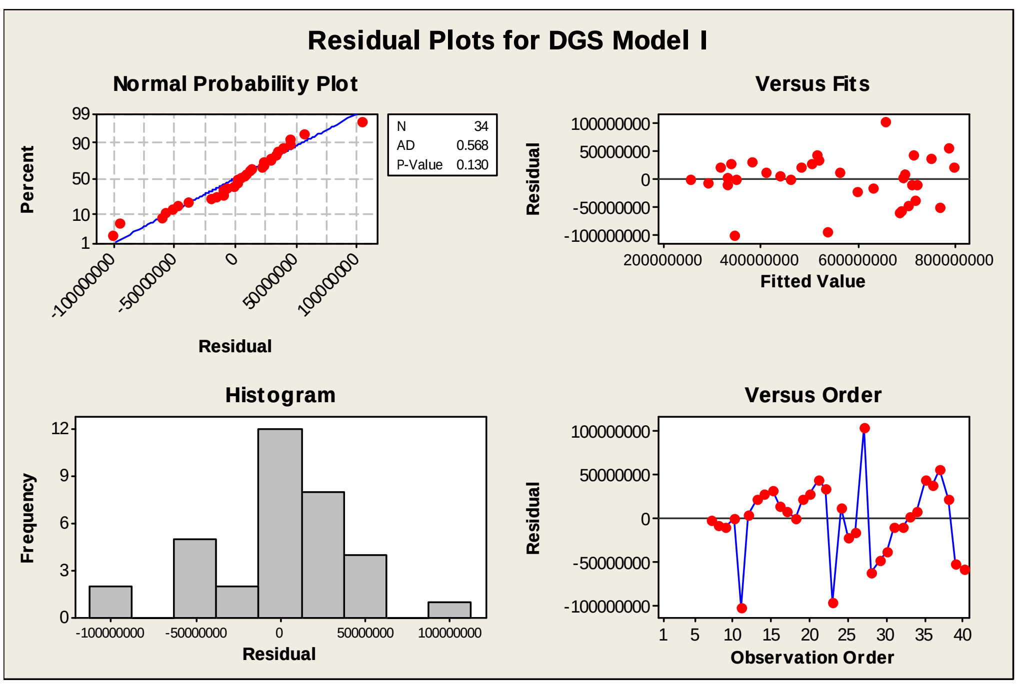
Model H: DGS versus ALEMP

```
The regression equation is
DGS = - 7.91E+08 + 739 ALEMP

34 cases used, 6 cases contain missing values

Predictor        Coef   SE Coef       T      P    VIF
Constant   -791003505  68399640  -11.56  0.000
ALEMP          739.16     37.46   19.73  0.000  1.000

S = 48791546   R-Sq = 92.4%   R-Sq(adj) = 92.2%
PRESS = 8.610266E+16   R-Sq(pred) = 91.41%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       1  9.26689E+17  9.26689E+17  389.26  0.000
Residual Error  32  7.61797E+16  2.38061E+15
Total           33  1.00287E+18

Unusual Observations
Obs    ALEMP        DGS        Fit    SE Fit    Residual  St Resid
 11  1538910  241922058  346498819  13217725  -104576761     -2.23R
 23  1809337  440370827  546387923   8368284  -106017096     -2.21R

R denotes an observation with a large standardized residual.
```



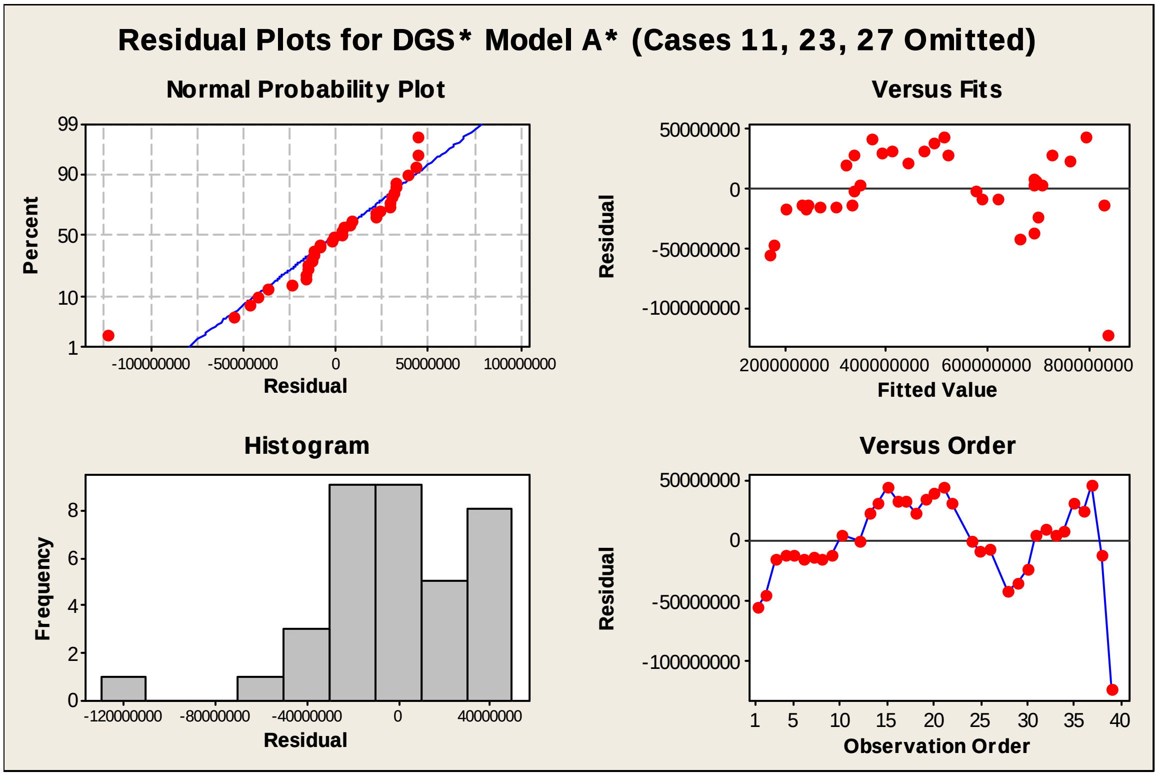
Model I: DGS versus ALEMP, USGDP

```
The regression equation is
DGS = - 5.19E+08 + 537 ALEMP + 0.000013 USGDP

34 cases used, 6 cases contain missing values

Predictor          Coef      SE Coef       T      P    VIF
Constant     -518710897    110128300   -4.71  0.000
ALEMP            536.51        75.92    7.07  0.000  5.115
USGDP        0.00001315   0.00000442    2.98  0.006  5.115

S = 43719155   R-Sq = 94.1%   R-Sq(adj) = 93.7%
PRESS = 7.473853E+16   R-Sq(pred) = 92.55%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       2  9.43616E+17  4.71808E+17  246.84  0.000
Residual Error  31  5.92523E+16  1.91136E+15
Total           33  1.00287E+18

Source  DF       Seq SS
ALEMP    1  9.26689E+17
USGDP    1  1.69274E+16

Unusual Observations
Obs    ALEMP        DGS        Fit    SE Fit    Residual  St Resid
 11  1538910  241922058  343585512  11883994  -101663454     -2.42R
 23  1809337  440370827  535406570   8357105   -95035743     -2.21R
 27  1992652  757621954  653430212  13580308   104191742      2.51R
 40  1900148  629165492  686385496  25780113   -57220004     -1.62 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```



Model A*: DGS* versus ATE (Cases 11, 23, and 27 deleted)

```
The regression equation is
DGS* = - 6.00E+08 + 544 ATE

36 cases used, 4 cases contain missing values

Predictor        Coef   SE Coef       T      P    VIF
Constant   -600168877  31259812  -19.20  0.000
ATE            544.41     15.33   35.51  0.000  1.000

S = 34480038   R-Sq = 97.4%   R-Sq(adj) = 97.3%
PRESS = 4.753035E+16   R-Sq(pred) = 96.91%

Analysis of Variance
Source          DF           SS           MS        F      P
Regression       1  1.49938E+18  1.49938E+18  1261.18  0.000
Residual Error  34  4.04217E+16  1.18887E+15
Total           35  1.53980E+18

Unusual Observations
Obs      ATE       DGS*        Fit    SE Fit    Residual  St Resid
 39  2640717  714747434  837461642  11321565  -122714208     -3.77R

R denotes an observation with a large standardized residual.
```



Model E*: DGS* versus ALEMP, ALGDP (Cases 11, 23, and 27 deleted)

The regression equation is
DGS* = - 3.09E+08 + 387 ALEMP + 0.00184 ALGDP

30 cases used, 10 cases contain missing values

Predictor	Coef	SE Coef	T	P	VIF
Constant	-309171104	83461586	-3.70	0.001	
ALEMP	387.24	59.70	6.49	0.000	7.427
ALGDP	0.0018442	0.0003160	5.84	0.000	7.427

S = 27543623   R-Sq = 97.6%   R-Sq(adj) = 97.4%
PRESS = 3.102508E+16   R-Sq(pred) = 96.33%

Analysis of Variance

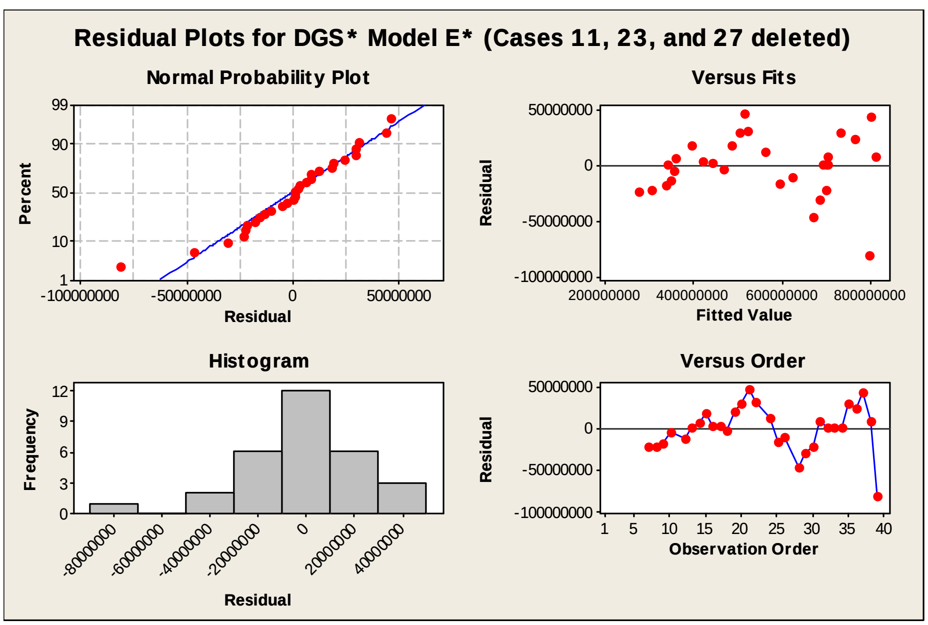
Source	DF	SS	MS	F	P
Regression	2	8.25981E+17	4.12990E+17	544.37	0.000
Residual Error	27	2.04836E+16	7.58651E+14		
Total	29	8.46464E+17			

Source	DF	Seq SS
ALEMP	1	8.00136E+17
ALGDP	1	2.58450E+16

Unusual Observations

Obs	ALEMP	DGS*	Fit	SE Fit	Residual	St Resid
39	2044406	714747434	796053886	15215894	-81306452	-3.54RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.



Correlations: DGS, DGS*, ALTPI, ALPOP, ALLF, ALEMP, ALUEMP, USGDP (DGS* has cases 11, 23, and 27 deleted)

	DGS	DGS*	ALTPI	ALPOP	ALLF	ALEMP	ALUEMP
DGS*	1.000						
	*						
ALTPI	0.935	0.944					
	0.000	0.000					
ALPOP	0.955	0.968	0.979				
	0.000	0.000	0.000				
ALLF	0.955	0.967	0.917	0.938			
	0.000	0.000	0.000	0.000			
ALEMP	0.961	0.972	0.901	0.919	0.988		
	0.000	0.000	0.000	0.000	0.000		
ALUEMP	-0.510	-0.500	-0.365	-0.355	-0.427	-0.564	
	0.002	0.004	0.034	0.039	0.012	0.001	
USGDP	0.933	0.944	1.000	0.976	0.912	0.897	-0.368
	0.000	0.000	0.000	0.000	0.000	0.000	0.032

Best Subsets Regression: DGS* versus ALTPI, ALPOP, ALLF, ALEMP, USGDP (Cases 11, 23, and 27 deleted)

Response is DGS*
31 cases used, 9 cases contain missing values

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	ALTPI	ALPOP	ALLF	ALEMP	USGDP
1	94.5	94.4	56.7	40051989				X	
1	93.5	93.3	72.0	43564567			X		
2	96.8	96.6	23.9	31161854				X	X
2	96.6	96.4	26.9	32120028	X			X	
3	97.8	97.5	11.2	26524122	X			X	X
3	97.6	97.3	14.2	27687644		X		X	X
4	98.0	97.7	9.4	25512648	X	X		X	X
4	98.0	97.7	9.6	25593770	X	X	X		X
5	98.4	98.0	6.0	23581295	X	X	X	X	X

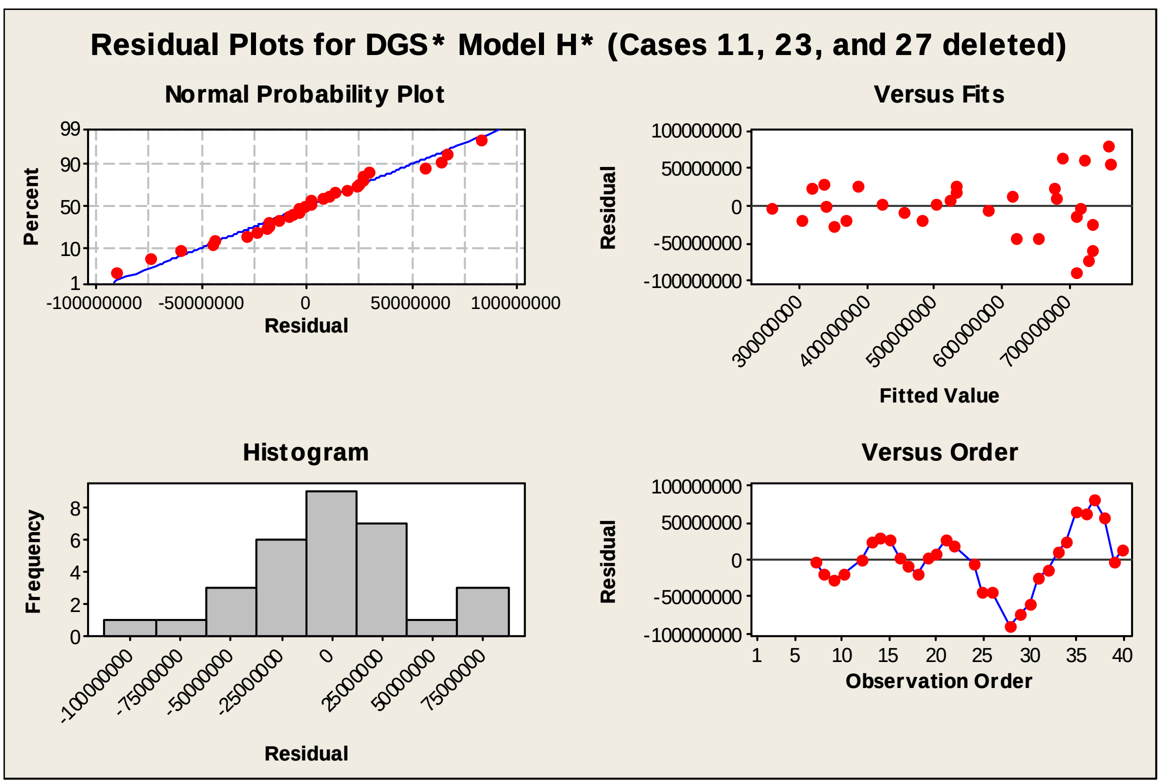
Best Subsets Regression: DGS versus ALTPI, ALPOP, ALLF, ALEMP, USGDP

Response is DGS
34 cases used, 6 cases contain missing values

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	ALTPI	ALPOP	ALLF	ALEMP	USGDP
1	92.4	92.2	16.0	48791546				X	
1	91.3	91.0	22.9	52287476			X		
2	94.1	93.7	7.8	43719155				X	X
2	93.9	93.5	8.8	44321860	X			X	
3	94.9	94.4	5.0	41338944	X			X	X
3	94.7	94.2	5.9	41945280		X		X	X
4	95.1	94.4	5.6	41121547	X	X		X	X
4	94.9	94.2	6.7	41835904	X		X	X	X
5	95.4	94.6	6.0	40685861	X	X	X	X	X

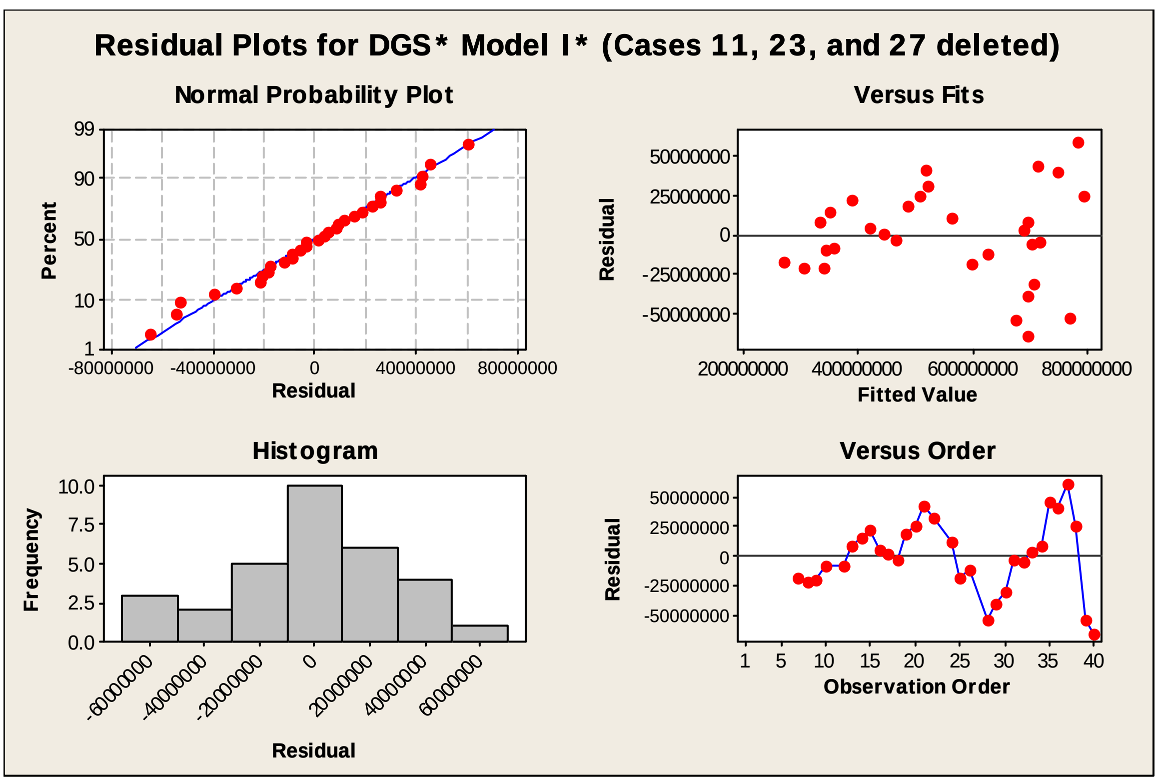
Model H*: DGS* versus ALEMP (Cases 11, 23, and 27 deleted)

```
The regression equation is
DGS* = - 7.38E+08 + 712 ALEMP

31 cases used, 9 cases contain missing values

Predictor         Coef   SE Coef       T      P    VIF
Constant    -737591944  58125430  -12.69  0.000
ALEMP           712.13     31.78   22.41  0.000  1.000

S = 40051989   R-Sq = 94.5%   R-Sq(adj) = 94.4%
PRESS = 5.358640E+16   R-Sq(pred) = 93.71%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       1  8.05629E+17  8.05629E+17  502.21  0.000
Residual Error  29  4.65207E+16  1.60416E+15
Total           30  8.52150E+17

Unusual Observations
Obs    ALEMP       DGS*        Fit    SE Fit   Residual  St Resid
 28  2035156  621617936  711706480  10032503  -90088544     -2.32R
 37  2100558  840948686  758281295  11577481   82667391      2.16R

R denotes an observation with a large standardized residual.
```



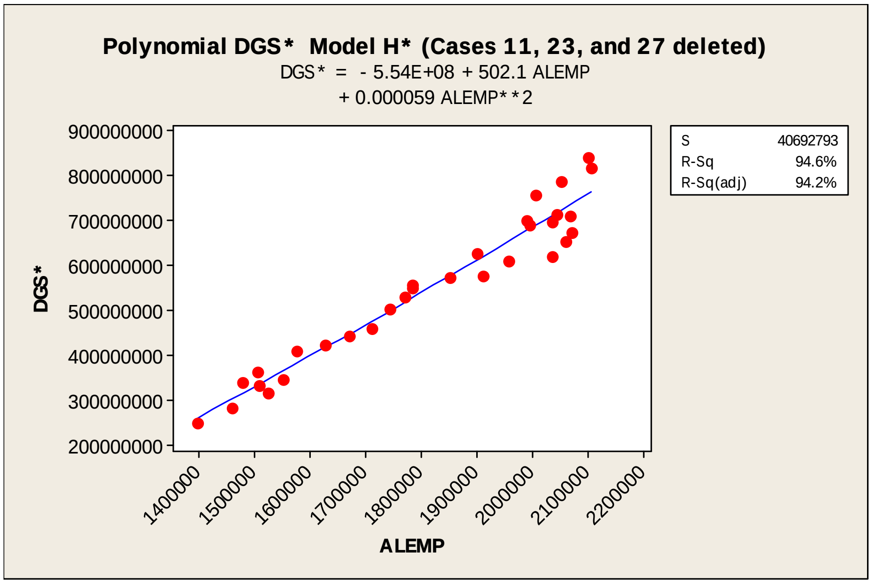
Model I*: DGS* versus ALEMP, USGDP (Cases 11, 23, and 27 deleted)

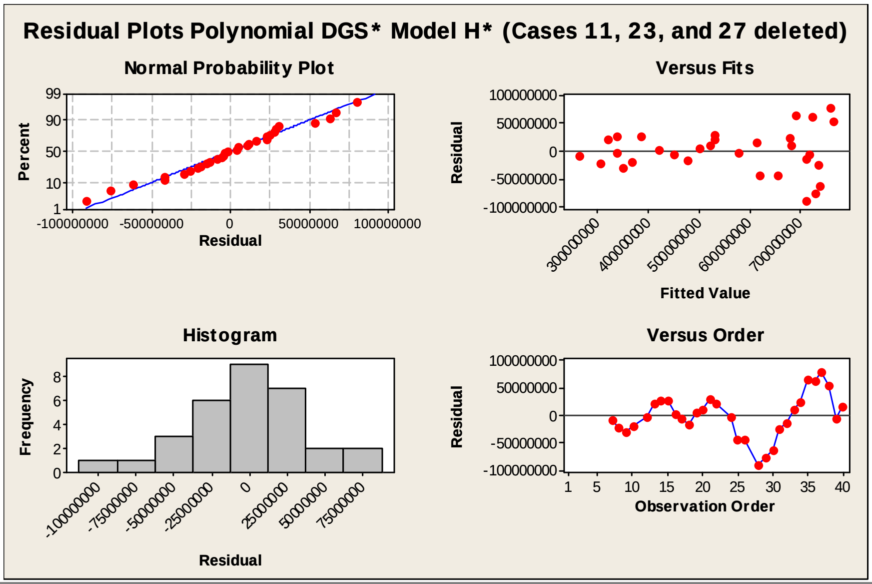
```
The regression equation is
DGS* = - 4.34E+08 + 486 ALEMP + 0.000014 USGDP

31 cases used, 9 cases contain missing values

Predictor          Coef      SE Coef      T      P    VIF
Constant    -433776066     81743044  -5.31  0.000
ALEMP           485.91        56.41   8.61  0.000  5.205
USGDP       0.00001449   0.00000325   4.46  0.000  5.205

S = 31161854   R-Sq = 96.8%   R-Sq(adj) = 96.6%
PRESS = 3.927766E+16   R-Sq(pred) = 95.39%

Analysis of Variance
Source          DF           SS           MS       F      P
Regression       2  8.24960E+17  4.12480E+17  424.77  0.000
Residual Error  28  2.71897E+16  9.71061E+14
Total           30  8.52150E+17

Source  DF       Seq SS
ALEMP    1  8.05629E+17
USGDP    1  1.93310E+16

Unusual Observations
Obs    ALEMP       DGS*        Fit    SE Fit   Residual  St Resid
 37  2100558  840948686  781061292  10354031   59887394     2.04R
 40  1900148  629165492  694113515  18592904  -64948022    -2.60RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```



Polynomial Regression Analysis: DGS* versus ALEMP (Cases 11, 23, and 27 deleted)

```
The regression equation is
DGS* = - 5.54E+08 + 502.1 ALEMP + 0.000059 ALEMP**2

S = 40692793   R-Sq = 94.6%   R-Sq(adj) = 94.2%

Analysis of Variance
Source      DF           SS           MS       F      P
Regression   2  8.05784E+17  4.02892E+17  243.31  0.000
Error       28  4.63653E+16  1.65590E+15
Total       30  8.52150E+17

Sequential Analysis of Variance
Source     DF           SS       F      P
Linear      1  8.05629E+17  502.21  0.000
Quadratic   1  1.55397E+14    0.09  0.762
```





Appendix D – Spreadsheet for Estimating Average Truck Trips per Day
(File name: 930-786 - Spreadsheet for Estimating Average Truck Trips per Day.xlsx)

This page provides screen shot of the Input worksheet. The next page provides a screen shot of the model data worksheet. The Excel File is provided separately. The spreadsheet file is password protected so that only the input variables can be entered so as to minimize the potential for inadvertently changing the model data.

Spreadsheet for Estimating Average Truck Trips per Day (ALDOT Research Project 930-768 by UAHuntsville)			
The input values for the ALEMP, ALGDP, and USDGP should fall within the ranges shown below. This will limit inappropriate extrapolation in the models which would reduce the accuracy of the predictions and prediction limits			
Description of Variables/Data to be entered in Next Column	Input Data Below	RECOMMENDED DATA BOUNDARIES	
Alabama Employment - ALEMP	2,146,626	1,396,193	2,461,000
Alabama Gross Domestic/State Product - ALGDP	180,214,840,000	12,455,000,000	227,685,000,000
U.S. Gross Domestic Product - USGDP	15,518,628,000,000	15,518,628,000,000	19,248,000,000,000
Estimate(s) and 95% Prediction Limit Calculations			
Models	Est. Diesel Gallons Sold	Lower Prediction Limit	Upper Prediction Limit
Model E (DGS = -408,841,076 + 449.35*ALEMP + 0.0016611*ALGDP)	855,108,879	758,115,206	952,102,551
Model H (DGS = -791,003,505 + 739.16*ALEMP)	795,699,034	691,523,538	899,874,530
Model I (DGS = -518,710,897 + 536.51*ALEMP + 0.00001315*USGDP)	837,026,787	739,477,035	934,576,539
Estimated Average Truck Trips Traveled per Day (5.5 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	354,951	314,689	395,212
Model H	330,290	287,048	373,533
Model I	347,445	306,953	387,937
Estimated Average Truck Trips Traveled per Day (6.0 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	387,219	343,297	431,141
Model H	360,317	313,143	407,490
Model I	379,031	334,858	423,204
Estimated Average Truck Trips Traveled per Day (6.5 mpg)	Estimated Average Truck Trips per Day	Lower Prediction Limit	Upper Prediction Limit
Model E	419,487	371,906	467,069
Model H	390,343	339,238	441,448
Model I	410,617	362,762	458,472

Appendix D - Continued

Model Data									
Model E				Model H			Model I		
Coefs		MSE		Coefs	MSE		Coefs		MSE
				-791003505.1	2.38061E+15				
-408841076.1		1.79253E+15		739.161045	ALEMP		-518710896.7		1.91136E+15
449.3521635	ALEMP						536.5062355	ALEMP	
0.001661122	ALGDP						1.31493E-05	USGDP	
XTXI Matrix				XTXI Matrix			XTXI Matrix		
8.527012445	-6.04377E-06	2.85812E-11		1.965252998	-1.06833E-06		6.345332072	-4.32822E-06	2.11519E-13
-6.04377E-06	4.3622E-12	-2.16708E-17		-1.06833E-06	5.89582E-13		-4.32822E-06	3.01577E-12	-1.57424E-19
2.85812E-11	-2.16708E-17	1.24574E-22					2.11519E-13	-1.57424E-19	1.02145E-26
1	2,146,626.30	1.80215E+11		1	2,146,626.30		1	2,146,626.30	1.55186E+13
Check Data		1			1				1
Predictions		855,108,878.55			795,699,034.04				837,026,786.92
SE(predictions)		47,545,918.10			51,066,419.51				47,818,505.97
PI Limits	758,115,205.64	952,102,551.47		691,523,538.23	899,874,529.85			739,477,034.74	934,576,539.09
If you have questions on this spreadsheet contact:									
Dr. Phillip A. Farrington									
The University of Alabama in Huntsville									
Phone: 256-824-6568									
e-mail: phillip.farrington@uah.edu									