For office use only	For office use only
T1 ________________	F1 ________________
T2 ________________	F2 ________________
T3 ________________	F3 ________________
T4 ________________	F4 ________________

2012

15th Annual High School Mathematical Contest in Modeling (HiMCM) Summary Sheet

(Please attach a copy of this page to each copy of your Solution Paper.)

Team Control Number: 3874

Problem Chosen: B

The cost of gasoline plays a major role in the everyday life of middle-class families. However, gasoline prices fluctuate significantly, increasing the need for wise financial decision-making. Our model both predicts gasoline price

The first part of our model identified possible predictors of gasoline price. We took into account both economic factors, such as gasoline company stock prices and crude oil price, as well as natural factors, including the weather. We constructed a multiple regression model from 2011 data and reduced it through AIC-selection, until only three significant predictor variables remained: the change in gasoline and crude oil price from the past week and the value of Exxon-Mobil stock. Because we found that prices from two weeks ago were only weakly correlated with current prices, we decided not to model gasoline price more than one week into the future.

The second part of our model used these variables to generate a probabilistic prediction of the change in gasoline price in the following week. We split the possible percent changes in gasoline price into five quintiles, then used the 2011 training data to fit the parameters of each quintile to a multivariate normal distribution. Testing the model on 2012 data, we obtained, for each week, the probabilities that the change in gas would be in each of the five quintiles. We then compared our predictions with actual 2012 gasoline prices and showed that our model was 10^{20} times more likely than the null hypothesis.

The final part of our model used the probabilities generated to decide what actions to take. Because it is more advantageous to buy gas right before a price increase, we used the idea of thresholds: if the predicted change in price is above a certain threshold, we advise the consumer to buy more gas (a full tank instead of a half tank, or a half tank instead of nothing). We found the optimal thresholds by training on 2011 data and then applied them to 2012 data using the second part's price predictions. We calculated the efficiency of our strategy by comparing the resulting savings to the best possible savings.

Applying this model to gasoline prices from New York City gave an efficiencies of 35% and 53% for cases 1 and 2, respectively. These correspond to savings of \$3.84 and \$5.75 per year, respectively.

For office use only	For office use only
T1 ________________	F1 ________________
T2 ________________	F2 ________________
T3 ________________	F3 ________________
T4 ________________	F4 ________________

2012

15th Annual High School Mathematical Contest in Modeling (HiMCM) Summary Sheet

(Please attach a copy of this page to each copy of your Solution Paper.)

Team Control Number: 3874

Problem Chosen: B

The cost of gasoline plays a major role in the everyday life of middle-class families. However, gasoline prices fluctuate significantly, increasing the need for wise financial decision-making. Our model both predicts gasoline price

The first part of our model identified possible predictors of gasoline price. We took into account both economic factors, such as gasoline company stock prices and crude oil price, as well as natural factors, including the weather. We constructed a multiple regression model from 2011 data and reduced it through AIC-selection, until only three significant predictor variables remained: the change in gasoline and crude oil price from the past week and the value of Exxon-Mobil stock. Because we found that prices from two weeks ago were only weakly correlated with current prices, we decided not to model gasoline price more than one week into the future.

The second part of our model used these variables to generate a probabilistic prediction of the change in gasoline price in the following week. We split the possible percent changes in gasoline price into five quintiles, then used the 2011 training data to fit the parameters of each quintile to a multivariate normal distribution. Testing the model on 2012 data, we obtained, for each week, the probabilities that the change in gas would be in each of the five quintiles. We then compared our predictions with actual 2012 gasoline prices and showed that our model was 10^{20} times more likely than the null hypothesis.

The final part of our model used the probabilities generated to decide what actions to take. Because it is more advantageous to buy gas right before a price increase, we used the idea of thresholds: if the predicted change in price is above a certain threshold, we advise the consumer to buy more gas (a full tank instead of a half tank, or a half tank instead of nothing). We found the optimal thresholds by training on 2011 data and then applied them to 2012 data using the second part's price predictions. We calculated the efficiency of our strategy by comparing the resulting savings to the best possible savings.

Applying this model to gasoline prices from New York City gave an efficiencies of 35% and 53% for cases 1 and 2, respectively. These correspond to savings of \$3.84 and \$5.75 per year, respectively.

The Weekly Drill: Gas or Pass?

Team #3874

November 5, 2012

Contents

1 Letter to the Local Newspaper

Dear Editor of the New York Times,

As avid readers of your newspaper, our team could not help but notice the number of residents concerned with the rising trend of gasoline prices. Being math enthusiasts, we took on the task of developing a model to help consumers predict whether to purchase a full tank or half tank of gasoline for their cars on a weekly basis. This decision takes into account several factors, including average weekly mileage driven, amount of gasoline already in the consumer's tank, and change in gasoline price based on data from 2011.

We considered many factors that might be correlated to change in gasoline price. It was determined that previous change in gasoline price, previous change in crude oil price, and Exxon-Mobile stock price per share all significantly affect the predicted change in gasoline price at the pump over the next week.

After simulating buying gas for the years 2011 and 2012, we found that, at best, a family may save up to 1.5% on their yearly gas expenses by selectively choosing when to fill up on gas. Using our model, which only requires three variables, we can achieve up to 53% of the ideal savings. This results in savings of about $5.75 per year. Over the lifespan of a car, this adds up to savings of over $100. When our model is applied to New York City, the citizens would save over $14 million in total each year.

Our model is not only effective but also easy to use. As detailed in the technical report, the distinct components of our model afford great versatility in simulating different situations. It can also be adjusted for other regions by substituting local gas price data for the New York City data. Results may be obtained by inserting the local data into the source code provided in the full paper. Currently, our team is developing a smartphone and web application that only requires an input of residential location. If historical gas prices cannot be found for a given city or town, our application will use information from the nearest location with available data.

Through the creation of this novel algorithm, we hope that we will be able to help consumers throughout the country save money on gasoline during the ongoing financial recession. Our team believes that our model will allow citizens to make an accurate prediction on a weekly basis of how much gasoline to purchase.

Sincerely,
Team #3874

2 Introduction

2.1 Background

Since the advent of the gasoline-powered automobile, the world has grown close to the pump (Fig. 1). Patented in 1886 by Karl Benz, the gasoline-fueled car has not only revolutionized personal transportation but has also given rise to the popular usage of gasoline [1]. In fact, the commercialization and standardization of internal combustion vehicles account for the dramatic growth in demand for gasoline beginning in the 19th century: consumption soared from less than three billion gallons in 1919 to more than 135 in 2002 [2].

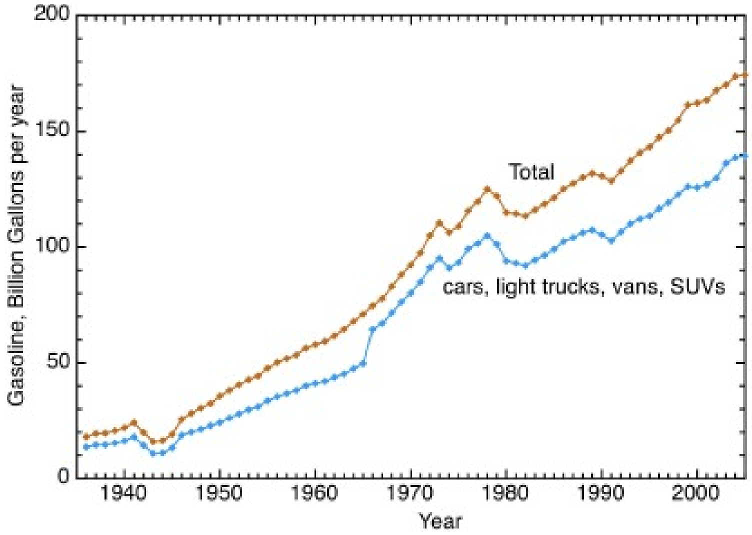

Figure 1: Trend of gasoline consumption in the US from 1935 to 2005

Currently, the United States consumes about 367.08 million gallons of gasoline daily [3]. Because the modern car has become such a commonly used mode of transportation, gasoline costs are an inevitable expense for middle class American families. The Associated Press reports that the typical U.S. household spent $4,155 last year, and this annual amount is likely to increase as demand for gasoline grows and existing supplies diminish [4].

As the market value of gasoline increases, the price automobile owners pay at the pump increases as well. Though the overall demand for gasoline tends to rise with time, historical data shows that the week-to-week price of gasoline can change quite sharply in both positive and negative directions [5]. Furthermore, domestic prices can vary drastically from region to region: in October, the average price for gasoline in Chico, California was $4.46 per gallon while the average price in San Luis Obispo, California was $4.71 [6]. As the present-day financial recession continues, these seemingly fickle price fluctuations pose great concerns for citizens looking to economize spending.

Gasoline prices have long been under scrutiny, leading to the development of a vast, global gasoline futures market. Past analyses have shown that gasoline prices respond more quickly to increases than to decreases in crude oil price [7, 8]. Although there is an asymmetrical response, perhaps due to inventory adjustment effects or independent retail outlets, changes in gasoline price tend to follow changes in crude oil price [9]. By considering factors that significantly affect gasoline price changes, a beneficial model may be constructed to help consumers make decisions regarding gasoline consumption.

2.2 Problem Restatement

In our paper, we propose solutions to the following objectives:

- Develop a model that helps consumers predict how much gasoline to purchase each week (none, half tank, or full tank).
- Determine the amount of gasoline to be purchased by consumers assumed to drive 100 or 200 miles per week.
- Identify a threshold value for weekly mileage, if applicable, that changes the amount of gasoline a consumer would buy.
- Apply the model to data for a large, metropolitan area in the United States.

2.3 Global Assumptions

1. Gasoline and crude oil prices are not significantly autocorrelated for time lags greater than several months. This is apparent from a graph of the prices over the past few years (see Fig. 2): fluctuations follow no pattern in length or frequency, and the prices do not stay around a fixed mean.
2. No natural disasters occur. Such events are outside the scope of our model, since major natural disasters vary significantly in effect and occur relatively scarcely.
3. We will ignore long-term gas price trends because they are unpredictable (by global assumption (1)) and much smaller in magnitude than weekly fluctuations, and because customers cannot avoid buying gas for longer than a month given the parameters for the cases studied. As such, we will only consider differences in prices from week to week.
4. A time difference of one year will not significantly change the general behavior of gasoline prices; the prices in 2012 are similar to those of 2011.
5. A gas tank holds 16 gallons, and average car mileage is 25 miles per gallon.

3 Model

3.1 Identification of Significant Variables

The first part of the task is to determine which factors significantly affect the price of gasoline. There are two types of factors: economic factors, such as the price of crude oil and the demand for gasoline, and natural factors, including the weather and natural disasters. We will first consider a wide variety of explanatory variables. Then, we will conduct stepwise regression on the 2011 data using the Akaike Information Criterion to identify the most significant predictor variables.

3.1.1 Assumptions

1. By global assumption (1), we will not consider gasoline prices (nor crude oil prices) from more than a month in the past.

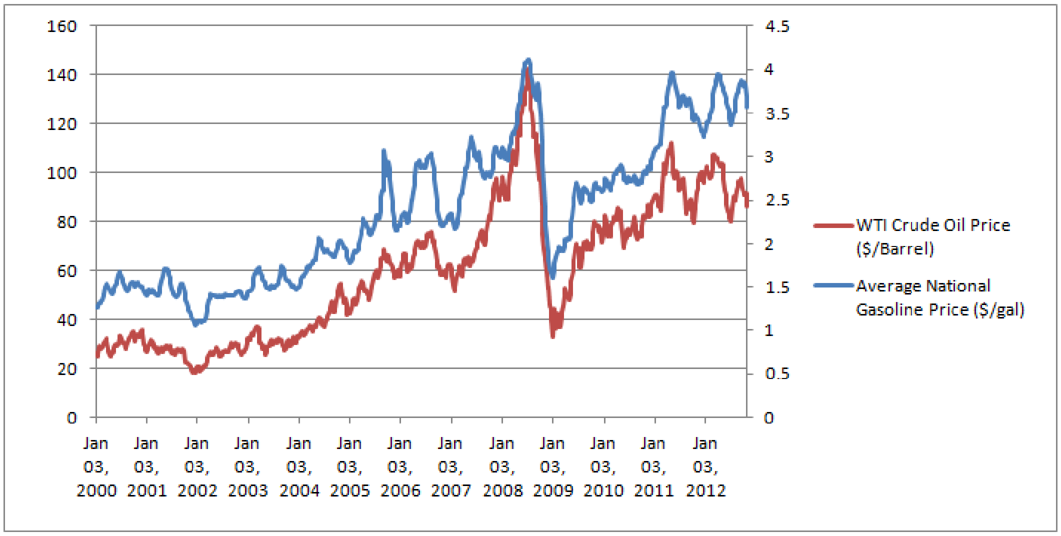

Figure 2: Gasoline and crude oil prices over the past 10 years. The prices are very closely correlated with each other, but not significantly autocorrelated over the long term. In the short term, however, they are autocorrelated.

2. The weather (not including natural disasters as stated in global assumption (2)) only affects the demand for gasoline, which is closely correlated with the amount of gasoline sold [10]. This is reasonable since extraction of crude oil acts independently of weather, and weather affects the number of trips and vacations.
3. Overall indicators of economic health (e.g. GDP) do not change significantly.
4. Refinery capacity is consistent and sufficient, and refineries do not malfunction.
5. The distribution of percent change in gas price is normal.

3.1.2 Variables Considered

Let $g(t)$ denote the weekly average national gas price, in dollars per gallon, where t is in weeks. Let $c(t)$ denote the weekly average WTI (West Texas Intermediate) crude oil price, in dollars per barrel, where t is in weeks. Weekly average data is provided for Mondays for $g(t)$ (which will be the standard for our other variables), but only provided for Fridays for crude oil prices. We accept this three day lag in crude oil prices as a source of error. Let $\Delta g(t) = g(t) - g(t-1)$ and $\Delta c(t) = c(t) - c(t-1)$. We wish to predict the change in gas price next week, $\Delta g(t+1)$. Since there is no well-defined mean, by global assumption (1), we will only consider differences in prices, and not use any data involving prices of means.

Though gasoline prices are not long-term autocorrelated, they are short-term autocorrelated, since most fluctuations last at least 2 to 3 weeks. The same is true for crude oil prices. This is supported by the results of Hughes *et al.*, a study that confirmed the existence of short-term autocorrelation in gasoline and crude oil prices [11]. As such, we will use the differences in gas and crude oil prices from those of one week ago, as well as the weekly changes 1 and 2 weeks ago:

$$\Delta g(t),\ \Delta g(t-1),\ \Delta g(t-2),\ \Delta c(t),\ \Delta c(t-1),\ \Delta c(t-2)$$

It is assumed as an immediate result of assumption (4) that any variations in the supply of crude oil (or gasoline in its unprocessed state) will be reflected in the price trends of crude oil, which in turn influence the price trend of gasoline. Therefore, we will not add the supply of crude oil as another predictor variable.

To account for short-term economic fluctuations, we will use high stock prices on each Monday for the three largest oil companies that do business in the United States: Exxon-Mobil (XOM), British Petroleum (BP), and Chevron (CVX). As per assumption (3), we will disregard long-term indicators such as GDP or unemployment rate. The fluctuations of these economic measures are affected by many factors other than demand for gasoline and do not change fast enough to provide significant predictive capacity over the span of a few weeks.

A final factor that can affect gasoline price is weather. There is notable seasonality in gasoline demand, which good weather conditions account for. However, by assumption (2), we can replace the quality of weather, which is difficult to quantify, with the amount of gasoline sold per week.

Given the aforementioned premises, we have ten predictor variables and one response variable. The predictor variables are weekly change in gas price from 1, 2, and 3 weeks ago, weekly change in crude oil price from 1, 2, and 3 weeks ago, quantity of gas sold in the previous week, and stock prices of Exxon-Mobil, BP, and Chevron. The response variable is the change in gas price in the next week. Although the data provided for change in gas prices and crude oil prices differs by three days in initial lag, we assume that the resulting change in prices is negligible.

3.1.3 Significance Testing

To identify which of these hypothesized predictor variables do indeed have significant predictive capacity, we construct a multiple regression model incorporating all ten variables. Although we will not use this model to predict future changes in gas prices, it can provide valuable information about the relationships between the response and predictor variables. The Student's t-test p-value of each predictor is computed from the model, providing a measure of individual predictive capacities. The null hypothesis of these tests is that the hypothesized predictor variables do not significantly affect the variation of gas prices. Furthermore, our significance level of $\alpha = 0.05$ indicates that predictions of predictor variables with a p-value of less than the α will only occur 5% of the time in a random simulation.

In anticipation of the non-significance of some hypothesized predictor variables, we propose simplifying the regression model in a backwards stepwise fashion. This approach is conducted as follows: the Akaike Information Criterion (AIC) of the original model is computed, the variable with the highest p-value is removed, the AIC is re-computed, and the p-values are recomputed. If the new AIC is not lower than the original AIC, then the variable is replaced and the variable with the next highest p-value in the original model is removed. This continues until removal of one variable results in a lower AIC. That variable is then permanently removed, and the entire process iterates. Model reduction is complete once all remaining predictor variables have significant p-values. Using the AIC to compare models is beneficial in this case because it promotes the lossless model but penalizes its overfitted counterpart [12].

As indicated in Table 1, the only variables remaining after the stepwise AIC-selection of the best multiple regression model were change in gas prices from a week ago, change in crude oil prices from a week ago, and Exxon Mobil stock prices (Fig. 3). (A full table listing the AIC value of each step can be found in the Appendix, Part 5.1.) Therefore, these variables will be used to predict future changes in prices of gas. Since gasoline prices are local, while the other explanatory variables are the same nationally, the next section will use gasoline prices from New York City.

Predictor Variable	Estimate	Standard Error	t-value	p-value
Intercept	-0.252	N/A	N/A	N/A
$\Delta g(t-1)$	0.411	0.089	4.631	2.79×10^{-05}
$\Delta c(t-1)$	0.007	0.001	5.289	3.00×10^{-06}
XOM Price	0.003	0.001	2.586	0.013

Table 1: Estimates, standard errors, t-values, and p-values of the significant predictor variables of the stepwise AIC-selected multiple regression model. Analysis was performed with the use of R software, version 2.15.1.

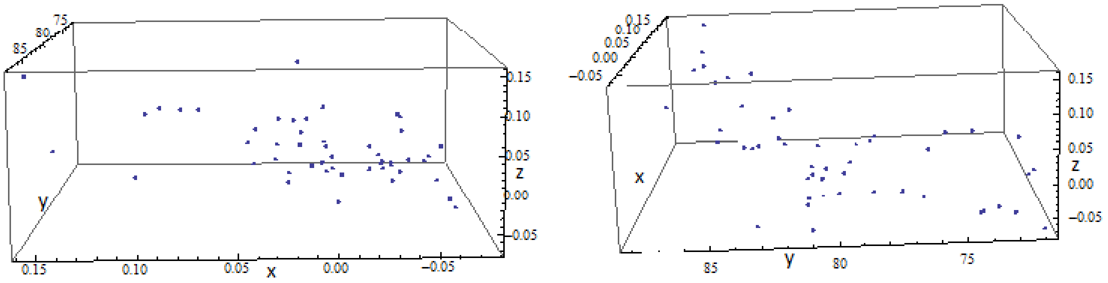

Figure 3: Two different views of the 3-dimensional plot of the predictor variables versus the response variable. On the x-axis is the percent change in gas prices from a week ago, on the y-axis is the percent change in crude oil prices from a week ago, and on the z-axis is the response variable, percent change in gas prices from the present week to the next.

3.2 Price Prediction

3.2.1 Overview

Since the decision between buying a full tank, half tank, or no gas at all depends on the change in gas prices, we propose here a probabilistic model that predicts a range of percent change in gasoline prices. It provides a general estimate of the direction and magnitude of price change the next week, rather than an exact number. The advantage of this technique lies in the fact that most deterministic models have a root mean square error (RMSE) of \$0.60 or more, a margin of error which is considerably large compared to gas prices. We avoid this error by stratifying the range of percent change in price into five classes, or quintiles.

The model is first trained on 2011 data using local gasoline prices from New York City, with quintile boundaries found from 2011 $\Delta g(t)$ values. The model is then tested on data from 2012. By global assumption (4), the gasoline market in 2012 is similar to that of 2011, so we use the same quintile boundaries. The model then provides five values per experimental data point: the probabilities, given the values of the three predictor variables, that $\Delta g(t+1)$ lies in each class.

We make no attempt to predict change in gas price more than one week in the future. This is because, as shown in 3.1.3, gas price changes are not significantly correlated with crude oil or gas price changes two or three weeks ago.

Mathematically speaking, let the bins be $\omega_1, \omega_2, ..., \omega_5$ and $\vec{x}$ be a 1×3 vector of data from the current week (change in gas prices from a week ago, change in crude oil prices from a week ago, and Exxon-Mobil stock price). $\vec{x}$ will be referred to as the *predictive parameter set.* Then according to Bayes' Theorem, the posterior probability that the percent price change

Boundary	% Change in Gas Prices
Between ω_1 and ω_2	-1.2
Between ω_2 and ω_3	-0.7
Between ω_3 and ω_4	0.2
Between ω_4 and ω_5	0.8

Table 2: Boundaries between the quintiles of weekly percent change in New York City gas prices from 2011. The upper and lower bounds were assumed to be ∞ and $-\infty$, respectively.

Class	ω_1	ω_2	ω_3	ω_4	ω_5
Mean	0.071	0.078	0.141	0.446	0.264
SD	0.053	0.110	0.185	0.284	0.302

Table 3: Mean and standard deviation of predicted quintile probabilities for 2012 New York City data.

next week will lie within bin ω_i given the predictive parameter set $\vec{x}$ is:

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)} \tag{1}$$

In this equation, $P(x|\omega_i)$ is the likelihood, or conditional probability, of predictive parameter set x being in class ω_i, and $P(\omega_i)$ is the prior probability. The conditional probability can be calculated from the distribution of the percent gas price changes from 2011 with respect to the $n \times 3$ training data set, where n is the number of weeks, or observations. Each row contains the predictive parameter set of each week. Since the aforementioned distribution was not sampled frequently enough to adequately approximate a continuous distribution, we fit the training data to a multivariate normal probability distribution [13] instead:

$$P(\vec{x}) = \frac{e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\sum^{-1}(\vec{x}-\vec{\mu})}}{(2\pi)^{d/2}|\sum|^{1/2}} \tag{2}$$

where $\vec{\mu}$ is the 3-dimensional row vector of the means of each set of predictor variables, $\sum$ is the 3×3 covariance matrix, $|\sum|$ is the determinant of the covariance matrix, and $\sum^{-1}$ is the inverse of the covariance matrix. The mean vector and the covariance matrix can be expressed as follows:

$$\begin{aligned} \vec{\mu} &= \int \vec{x}P(\vec{x})\,\mathrm{d}\vec{x} \\ \sum &= \int (\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T P(\vec{x})\,\mathrm{d}\vec{x} \end{aligned} \tag{3}$$

We use the above equations to calculate the posterior probabilities $P(\omega_1|x)$, $P(\omega_2|x)$, $P(\omega_3|x)$, $P(\omega_4|x)$, and $P(\omega_5|x)$ for $\vec{x}$. The MATLAB (version R2011b) source code for these computations can be found in 5.2.

As shown in Table 2, the quintile widths were generally no larger than %1.0 weekly change in gasoline price. Furthermore, Table 3 provides some descriptive statistics about the predicted class probabilities.

3.2.2 Model Validation

In order to validate our model on the 2012 data, we used Bayesian inference, an alternative method for comparing the likelihoods of two hypotheses [14]. Given two hypotheses, H_0 and

H_1, an application of Bayes' Theorem on the given data, D, gives:

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)}{P(D|H_0)} \cdot \frac{P(H_1)}{P(H_0)} \tag{4}$$

Thus, the effect of the data on the ratios of the likelihoods can be measured as $\frac{P(D|H_1)}{P(D|H_0)}$. Let the null hypothesis, H_0, be that the future gas price is randomly distributed among the quintiles. Let the alternative hypothesis, H_1, be that the future gas price has probabilities as determined by our model. Then, the effect of the data on the ratios of the likelihoods is:

$$\frac{P(D|H_1)}{P(D|H_0)} = \prod \frac{P_i}{0.2} \tag{5}$$

where P_i is the probability that the i^{th} trial's gas price is in the class that it is predicted to be in.

Our computation gave that this value is 9.524×10^{19}, so the results predicted by our model are significantly more likely than those predicted by the null hypothesis. Thus, for any reasonable priors, the likelihood of the alternative hypothesis is significantly greater than that of the null. For example, if our priors are that the null and alternative hypotheses are equally likely, each having a 50% probability, then after incorporating the data, the probability of the null hypothesis reduces to $\frac{1}{1+9.524\times10^{19}} \times 100\% = 1.05 \times 10^{-18}\%$.

3.3 Decision Model

Now, we will determine the optimal action for consumers to take given the estimate for the price change. In our model, we consider the consumer's possible options every week. A diagram showing the decisions that consumers can make each week (in the two cases, 100 and 200 miles per week) is below. The consumer can look at how much gasoline is in their gas tank and then determine whether to buy by comparing the expected change in gasoline price (as determined using 3.2) to the optimal threshold values as determined by training on the 2011 data.

3.3.1 Assumptions

1. The most significant factor in deciding how much gas to buy is the change in gas price over the next week, as our model from 3.2 does not predict gas prices further in the future. As discussed in 3.2, gasoline price is too unpredictable to model further than one week in the future. Also, since the major impact on the cost by the decision of whether to buy more or less gas this week is whether the consumer needs to buy next week, the difference in price between consecutive weeks is the most important.

2. As the expected gas price next week increases, the optimal action is to buy either the same quantity or more of gas. This is because, if gas prices are forecasted to go up significantly, it is advantageous to buy more gas earlier, rather than later.

3. The gas estimate based on 3.2 is a reasonable predictor of the change in gas price over the next week. This is reasonable since as shown in 3.2, our model is significantly better than randomly guessing that it is one of the quintiles.

4. Consumers cannot store gas: they cannot exceed 100% of their tank's capacity.

5. As mentioned in the problem statement, the consumer can only buy a half gallon or full gallon of gas.

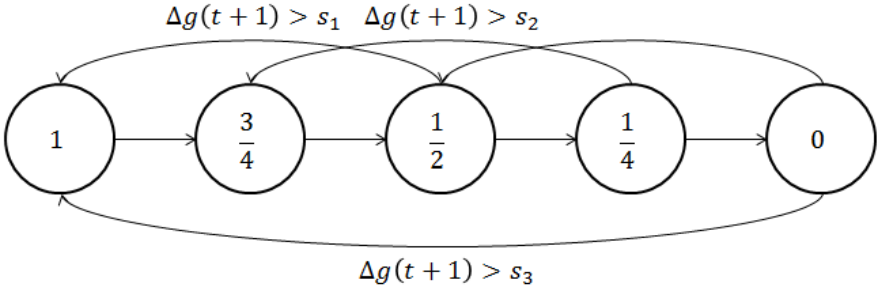

Figure 4: Decision Procedure for Case 1.

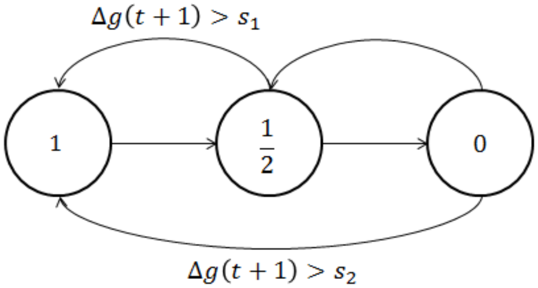

Figure 5: Decision Procedure for Case 2.

3.3.2 Decision Procedure

Now, we will show how our model comes out of the above assumptions. At the three gas levels in Case 1, and at two gas levels in Case 2, the customer has the choice to buy either less gas (none or a half tank) or more gas (a half tank or a full tank). By assumption (2), the optimal quantity of gas to buy in the current week must be a increasing function of the expected rise in cost of gas next week. Therefore, since there are only two options, there must be some threshold on the expected rise in cost of gas next week which determines the consumer's choice. By assumption (1), the expected rise in gas cost is the most significant factor and so the only factor considered at any point in time.

Denote the thresholds as s_1 through s_3 for Case 1, and s_1 and s_2 for Case 2. We assume that these threshold values are constant over time (by global assumption (4)) and independent of gas price. This is reasonable since the ideal threshold values are parameters of the decision process, which is independent of the gas price. We observe that, for Case 2, the thresholds should be fairly close to \$0, because even filling a full tank will only delay refilling for two weeks, but for Case 1, a full tank will last 4 weeks. In the latter case, the thresholds should be higher, in order to justify getting 4 weeks' worth of gas. The procedure is summarized in Figures 4 and 5.

3.3.3 Training the Model

We first trained our model on 2011 New York City gasoline price data. We used the actual change in gas price as the predicted change, and used an Excel spreadsheet to simulate Case 1 and Case 2 customers over one year. The threshold values were then adjusted in increments of a tenth of a cent.

Because there are only 52 weeks in our training data, there were intervals for each threshold

	Case 1	Case 2
s_1	0.010	-0.0015
s_2	-0.005	-0.0005
s_3	0.0035	N/A

Table 4: Optimal threshold values for New York City in 2011 ($/gal)

value which all gave the same lowest value for price per gallon. The optimal threshold value was taken to be the midpoint of the corresponding interval. Table 4 summarizes the optimal threshold values found.

These threshold values yielded an average price per gallon of $3.563 in Case 1 and $3.587 in Case 2. As predicted, the Case 2 thresholds are very close to zero, and the Case 1 price is slightly lower, since the consumer has more freedom.

3.3.4 Testing the Model

The model was tested on 2012 New York City price data, using the probabilities generated by 3.2. The predicted change in gas price, $\Delta\hat{g}(t)$, is found by weighting the probabilities by the median 2011 value within each quintile. Denoting these values m_i, we have that the expected value of the change in gas price is:

$$\Delta\hat{g}(t) = \sum_{i=1}^{5} P_i \times m_i$$

We then simulated applying our model during 2012, up to the beginning of November, by using the thresholds previously found and these estimates with our decision model to determine the appropriate action. The results are shown in Tables 5 and 6.

In order to give a reference point to analyze the cost to the consumer, this strategy was compared to two other strategies: the random strategy and the true optimal strategy. In the random strategy, the consumer buys just as much gas as is needed, without considering prices. Thus, the cost per gallon is simply the average cost of gasoline over the year. In the optimal strategy, the consumer is able to perfectly predict gas prices for the entire year. A dynamic programming algorithm was written to find the best possible cost per gallon assuming this perfect knowledge.

The procedure works on the following principle: there are only a few states that are possible in each week (no gas, 3/4 gas, etc.). Let $f(i,j)$ denote the cheapest possible way to, starting with an empty tank on week 1, end up with $j/4$ of a full tank on week i. If we know $f(i,0)$ through $f(i,4)$, it is simple to find $f(i+1,0)$ through $f(i+1,4)$. We simply take all possible actions on week i and find the minimum for a given fuel level on week $i+1$. For example, in Case 1 $f(5,2) = \min(f(4,3), f(4,1) + g(4))$. To guarantee that we start with an empty tank, we set $f(0,0) = 0$ and $f(0,1)$ through $f(0,4)$ to ∞, and we continue until we reach the end of the year.

We define the efficiency ϵ of a strategy as follows: let the average prices per gallon in the model, random, and optimal strategy be P_m, P_r, and P_o, respectively. Then the efficiency is

$$\epsilon = \frac{P_r - P_m}{P_r - P_o}$$

This determines how close our model gets to the best possible result.

	2011 (Training)	2012 (Testing)
P_m(\$/gal)	3.563	3.729
P_r(\$/gal)	3.608	3.747
P_o(\$/gal)	3.554	3.695
ϵ	0.82	0.35
Total Savings (\$)	9.38	3.84

Table 5: Results for Case 1.

	2011 (Training)	2012 (Testing)
P_m(\$/gal)	3.587	3.733
P_r(\$/gal)	3.608	3.747
P_o(\$/gal)	3.582	3.721
ϵ	0.82	0.53
Total Savings (\$)	8.91	5.75

Table 6: Results for Case 2.

3.4 Final Results

The results of our model on the 2012 New York City data are shown in Tables 5 and 6. To give an estimate of the accuracy of 3.2, we also include the results when the model is run on the 2011 training data using the actual gas price differences. We note that the maximum possible improvement amounts to less than 1.5% in both cases, which corresponds to less than \$10 in savings throughout the entire year. As a rough estimate, suppose there are 3 million cars in New York City, half of which correspond to each case. Then our plan, implemented through 2012, saves the citizens of the city over 14 million dollars.

3.4.1 Changing the Mileage

The models in our paper are specifically designed for consumers who drive 100 or 200 miles per week (using 1/4 or 1/2 of their tank per week). However, given the same assumptions, the model can be extended to any mileage. Our model requires a separate threshold value for each possible amount of gas in the tank less than one-half. If a consumer uses a rational fraction of their gas tank per week (p/q), the set of possible amounts of gas in the tank is finite, and a similar training procedure can be used to find optimal threshold values (at $0/q$, $1/q$, ..., $1/2$ if q is even or $0/2q$, ..., $(q-1)/2q$ if q is odd). In the case where an irrational fraction of the tank is used per week, we can interpolate values between those of existing threshold values to get a continuous function. The small errors introduced in this approximation would be overshadowed by the assumption that only exactly a half tank or full tank of gas can be bought. Thus, this model can be extended to take into account arbitrary miles per week.

Now we address what happens to the thresholds as the mileage driven increases. As gas usage per week increases, customers cannot afford to wait as long, so the threshold values should decrease. This is supported by our model, since the threshold values for Case 2 are less than those of Case 1, which corresponds to half the mileage ($.01 < -.0015$ and $.0035 > -.0005$). Also, threshold values should decrease with the amount of gas in the tank, again because customers cannot afford to wait as long. This is also supported by our data since $s_2 = -0.005 < 0.0035 = s_3$. Thus, as gas mileage increases, consumers are more likely to buy a larger quantity of gas than a smaller quantity when they have the same amount in the tank. There will be some threshold mileage where the consumer should switch from waiting another week to filling up their tank.

4 Discussion

The three parts of our model, though interdependent, are mostly independent in terms of internal structure. As a result, it may be easily adapted or extended. Extra variables can be added to the first part (Significant Variables), and, if significant, added to the second (Price Prediction). The second part can be specialized to any city or region by using values for the significant variables specific to that area. It also accommodates an arbitrarily long training time, and can be extended to future years. Since the model is entirely probabilistic, not relying on regression, the timescale can also be changed to predict gas prices within days or months. If gas prices change arbitrarily in nature, such as in the overall strength or frequency of fluctuations, the thresholds in the third part of the model (Decision Model) can be automatically adjusted to find the new optimal strategy. Finally, as discussed in 3.4.1, the decision model can be adjusted for arbitrary car mileages.

However, there are aspects to the problem that were simplified. For example, it has been found that gasoline price responds asymmetrically to crude oil price changes, tending to take longer to respond to decreases than increases [7, 8]. Our model does not take into account differences in the nature of gasoline prices between 2011 (the training set) and 2012 (the testing set), since it is difficult to quantify with any one factor. It also does not predict changes in price due to natural disasters, since there are only a handful of major natural disasters in our training set to work with. Finally, the Bayesian price model assumes that all variables are normally distributed, but graphing changes in gasoline shows that it is slightly heavy-tailed.

Possible extensions of the model which could be implemented in future iterations include the following:

- Implement the model for different cities. In particular, gasoline costs in some regions of the United States, such as California, are significantly higher than those in New York.
- Explicitly take into account the change of gasoline demand with the seasons, using gasoline's price elasticity of demand.
- Find threshold cost changes for arbitrary mileages, and, by using statistics on the distribution of mileage driven per week, obtain a better estimate of the average amount of money saved.

Despite these shortcomings, our model still achieves efficiencies of 35% and 53%, predicting a highly volatile price with only one year of training data. It retains great potential for aiding the average consumer in saving immense amounts of money without much effort. All computation is performed by computer programs using data publicly available online, and no work is required on the consumer's part. Moreover, we propose that the code be packaged as a smartphone application as well as an online applet, further facilitating the decision-making process. In the current, highly gasoline-dependent era, aiding consumers in making wise financial choices will not only benefit the individual consumer, but also the economy as a whole.

5 Appendices

5.1 Intermediate Results

Iteration	Predictor Variables	AIC
1	$\Delta g(t), \Delta g(t-1), \Delta g(t-2), \Delta c(t), \Delta c(t-1), \Delta c(t-2), G, P_{XOM}, P_{BP}, P_{CVX}$	-338.88
2	$\Delta g(t), \Delta g(t-2), \Delta c(t), \Delta c(t-1), \Delta c(t-2), G, P_{XOM}, P_{BP}, P_{CVX}$	-340.87
3	$\Delta g(t), \Delta g(t-2), \Delta c(t), \Delta c(t-1), \Delta c(t-2), G, P_{XOM}, P_{CVX}$	-342.86
4	$\Delta g(t), \Delta c(t), \Delta c(t-1), \Delta c(t-2), G, P_{XOM}, P_{CVX}$	-344.5
5	$\Delta g(t), \Delta c(t), \Delta c(t-1), \Delta c(t-2), G, P_{XOM}$	-346
6	$\Delta g(t), \Delta c(t), \Delta c(t-1), \Delta c(t-2), P_{XOM}$	-347.02
7	$\Delta g(t), \Delta c(t), \Delta c(t-2), P_{XOM}$	-347.51
8	$\Delta g(t), \Delta c(t), P_{XOM}$	-350

Table 7: AIC values of the selection of the best multiple regression model, where G is the amount of gas sold

5.2 Source Code

5.2.1 Model Training and Testing Script

```
% Team #3874 - Model Training and Testing
[data,~,~]=xlsread(’Training_Data.xlsx’);
shiftedData=data(2:size(data,1),:);
shiftedData=[shiftedData; zeros(1,size(shiftedData,2))];
change=(shiftedData-data)./data;
change=change(1:(size(change,1)-1))’; % percent change of gas prices
changeOrdered=sort(change);
n=5; % number of classes
classIntervalWidth=round(size(change,1)/n);
class1Bound=changeOrdered(classIntervalWidth); % boundary between classes 1 & 2
class2Bound=changeOrdered(2*classIntervalWidth); % boundary between classes 2 & 3
class3Bound=changeOrdered(3*classIntervalWidth); % boundary between classes 3 & 4
class4Bound=changeOrdered(4*classIntervalWidth); % boundary between classes 4 & 5
class=zeros(size(change,1),1);
class(change<=class1Bound)=1;
class(change>class1Bound & change<=class2Bound)=2;
class(change>class2Bound & change<=class3Bound)=3;
class(change>class3Bound & change<=class4Bound)=4;
class(change>class4Bound)=5;
data=data(2:size(data,1),2:size(data,2));
data=[data class]; % training data matrix from 2011
[eD,~,~]=xlsread(’Experimental_Data.xlsx’);
response=eD(:,1);
shiftedResponse=response(2:size(response,1),:);
shiftedResponse=[shiftedResponse;zeros(1,size(shiftedResponse,2))];
expChange=(shiftedResponse-response)./response;
expChange=expChange(1:(size(expChange,1)-1));
classExp=zeros(size(expChange,1),1); % matrix of classes of experimental data
classExp(expChange<=class1Bound)=1;
classExp(expChange>class1Bound & expChange<=class2Bound)=2;
classExp(expChange>class2Bound & expChange<=class3Bound)=3;
classExp(expChange>class3Bound & expChange<=class4Bound)=4;
classExp(expChange>class4Bound)=5;
eD=eD(2:size(eD,1),2:size(eD,2));
modelResults=zeros(size(eD,1),n);
predictedClass=zeros(size(eD,1),1);
for i=1:size(eD,1)
    modelResults(i,:)=Bayes(eD(i,:),data);
    predictedClass(i)=find(modelResults(i,:)==max(modelResults(i,:)));
end
```

5.2.2 Probabilistic Model

```
% Team #3874 - Prediction of Change in Gas Prices
function predictedProbChange = Bayes(x,data1)
% last column of data1 is class
n=size(data1,1); % number of observations
C=data1(:,size(data1,2)); % classes labeled 1 through 5
A=data1(:,(1:(size(data1,2)-1))); % data matrix of past predictors
indClass1=find(C==1);
indClass2=find(C==2);
indClass3=find(C==3);
indClass4=find(C==4);
indClass5=find(C==5);
priorProbClass1=size(indClass1,1)/n;
priorProbClass2=size(indClass2,1)/n;
priorProbClass3=size(indClass3,1)/n;
priorProbClass4=size(indClass4,1)/n;
priorProbClass5=size(indClass5,1)/n;
AClass1=A(indClass1,:);
AClass2=A(indClass2,:);
AClass3=A(indClass3,:);
AClass4=A(indClass4,:);
AClass5=A(indClass5,:);
mu1=mean(AClass1);
mu2=mean(AClass2);
mu3=mean(AClass3);
mu4=mean(AClass4);
mu5=mean(AClass5);
sigma1=cov(AClass1);
sigma2=cov(AClass2);
sigma3=cov(AClass3);
sigma4=cov(AClass4);
sigma5=cov(AClass5);
predictedProbChange=zeros(1,5);
predictedProbChange(1)=priorProbClass1*mvnpdf(x,mu1,sigma1)/
    (priorProbClass1*mvnpdf(x,mu1,sigma1)+priorProbClass2*
    mvnpdf(x,mu2,sigma2)+priorProbClass3*mvnpdf(x,mu3,sigma3)+
    priorProbClass4*mvnpdf(x,mu4,sigma4)+priorProbClass5*
    mvnpdf(x,mu5,sigma5));
predictedProbChange(2)=priorProbClass2*mvnpdf(x,mu2,sigma2)/
    (priorProbClass1*mvnpdf(x,mu1,sigma1)+priorProbClass2*
    mvnpdf(x,mu2,sigma2)+priorProbClass3*mvnpdf(x,mu3,sigma3)+
    priorProbClass4*mvnpdf(x,mu4,sigma4)+priorProbClass5*
    mvnpdf(x,mu5,sigma5));
predictedProbChange(3)=priorProbClass3*mvnpdf(x,mu3,sigma3)/
    (priorProbClass1*mvnpdf(x,mu1,sigma1)+priorProbClass2*
    mvnpdf(x,mu2,sigma2)+priorProbClass3*mvnpdf(x,mu3,sigma3)+
    priorProbClass4*mvnpdf(x,mu4,sigma4)+priorProbClass5*
    mvnpdf(x,mu5,sigma5));
predictedProbChange(4)=priorProbClass4*mvnpdf(x,mu4,sigma4)/
    (priorProbClass1*mvnpdf(x,mu1,sigma1)+priorProbClass2*
    mvnpdf(x,mu2,sigma2)+priorProbClass3*mvnpdf(x,mu3,sigma3)+
    priorProbClass4*mvnpdf(x,mu4,sigma4)+priorProbClass5*
    mvnpdf(x,mu5,sigma5));
predictedProbChange(5)=priorProbClass5*mvnpdf(x,mu5,sigma5)/
    (priorProbClass1*mvnpdf(x,mu1,sigma1)+priorProbClass2*
    mvnpdf(x,mu2,sigma2)+priorProbClass3*mvnpdf(x,mu3,sigma3)+
    priorProbClass4*mvnpdf(x,mu4,sigma4)+priorProbClass5*
    mvnpdf(x,mu5,sigma5));
% output is a row vector with conditional probability for each class
```

6 Bibliography

[1] Deffree, Suzanne. “Karl Benz drives the first automobile, July 3, 1886.” *EDN Network.* UBM Tech, 3 July 2012. Web. 4 Nov. 2012.

[2] Melosi, Martin V. “Energy Use and the Internal Combustion Engine.” *Automobile in American Life and Society.* The Henry Ford, n.d. Web. 4 Nov. 2012.

[3] “How much gasoline does the United States consume?” *Energy Information Administration.* U.S. Department of Energy, n.d. Web. 4 Nov. 2012.

[4] Tuttle, Brad. “2011 Is Priciest Year Ever for Gasoline: $3.53 Per Gallon, Over $4K Spent Per Household.” *TIME Business & Money.* TIME, 20 Dec. 2011. Web. 4 Nov. 2012.

[5] “Independent Statistics and Analysis.” *Petroleum & Other Liquids.* U.S. Energy Information Administration, Nov. 2012. Web. 05 Nov. 212.

[6] “California gas prices hit all-time high as average soars to $4.61 a gallon.” *Fox News.* Fox News Network, 6 Oct. 2012. Web. 5 Nov. 2012.

[7] Balke, Nathan S., Stephen P. A. Brown, and Mine K. Yücel.“Crude oil and gasoline prices: an asymmetric relationship?” *Economic and Financial Policy Review* Q1 (1998): 2-11. Print.

[8] Borenstein, Severin, A. Colin Cameron, and Richard Gilbert. “Do Gasoline Prices Respond Asymmetrically to Crude Oil Price Changes?” *The Quarterly Journal of Economics* 112.1 (1997): 305-39. Print.

[9] Karrenbrock, Jeffrey D. “The Behavior of Retail Gasoline Prices: Symmetric or Not?” *Federal Reserve Bank of St. Louis Review* 73.4 (1991): 19-29. Print.

[10] Villar, Jose A., and Frederick L. Joutz. “The relationship between crude oil and natural gas prices.” *EIA manuscript*, October(2006).

[11] Hughes, Jonathan E., Christopher R. Knittel, and Daniel Sperling. “Evidence of a Shift in the Short-Run Price Elasticity of Gasoline Demand.” *The Energy Journal* 29.1 (2008): 113-34. *The National Bureau of Economic Research.* International Association for Energy Economics. Web. 5 Nov. 2012.

[12] “Information Theory and an Extension of the Maximum Likelihood Principle.” *International Symposium on Information Theory* 2nd (1973): 267-81. Web.

[13] Meerschaert, Mark M. *Mathematical Modeling.* Amsterdam: Elsevier Academic, 2007. Print.

[14] Box, George, and George Tiao. *Bayesian Inference in Statistical Analysis.* N.p.: Wiley Classics Library, 1973. Print.