Drive Into the Future

Students: Richard Qian, Ryan Ho, Colin Menz

Faculty Advisors: Janet Duncan, Roberto Molinari

Abstract

The goal of this project was to forecast future miles driven over the next year to help the California State Auto Association estimate their future risks. Various economic indices were compared and selected to build a linear regression model to describe how said variables correlate to miles driven. The final model included US unemployment rate, US total vehicle sales and time of year and was able to predict mileage one year into the future.

Data

Initial mileage data came from US Department of Transportations Federal Highway Administration and appeared to exhibit seasonality. The patterns for US and California mileage data were similar enough that it was appropriate to consider indices from both the US and California. However, the two sources of data were not mixed when building models.

California indices considered were:

- Gross Domestic Product (GDP)
- Average Retail Gas Price per Gallon
- Consumer Price Index (CPI)
- Number of Licensed Drivers
- Number of Registered Vehicles
- Unemployment Rate
- Per Capita Real GDP

National indices considered were:

- Unemployment Rate
- Crude Oil (WTI) Price per Barrel
- Real Per Capita Disposable Income
- Total Vehicle Sales
- Consumer Sentiment Index

A binary variable called "seasonal indicator" was created to separate the data into two normal distributions (Figure 1, bottom right).

Linear Regression

This project used the most basic form of regression, linear regression. It is assumed that the data is approximately normally distributed with normal residuals. The equation to regress an explanatory variable on p predictors is:

$$y = \beta + \beta_0 x_1 + \beta_1 x_2 + ... + \beta_{p-1} x_p$$

The data appeared to have two overlapping normal distributions (Figure 1, top right), indicating linear regression would be an appropriate approach once seasonality was accounted for.

References

[1] Economic indices from Trading Economics website.

[2] Mileage from Federal Highway Administration website.

Acknowledgements

We would like to thank our faculty advisors, Janet Duncan and Roberto Molinari, for their insight, passion and motivation for this research. We would also like to thank our sponsor CSAA for their guidance, as well as the SOA Center of Actuarial Excellence 2015 Education Grant.

Graphics

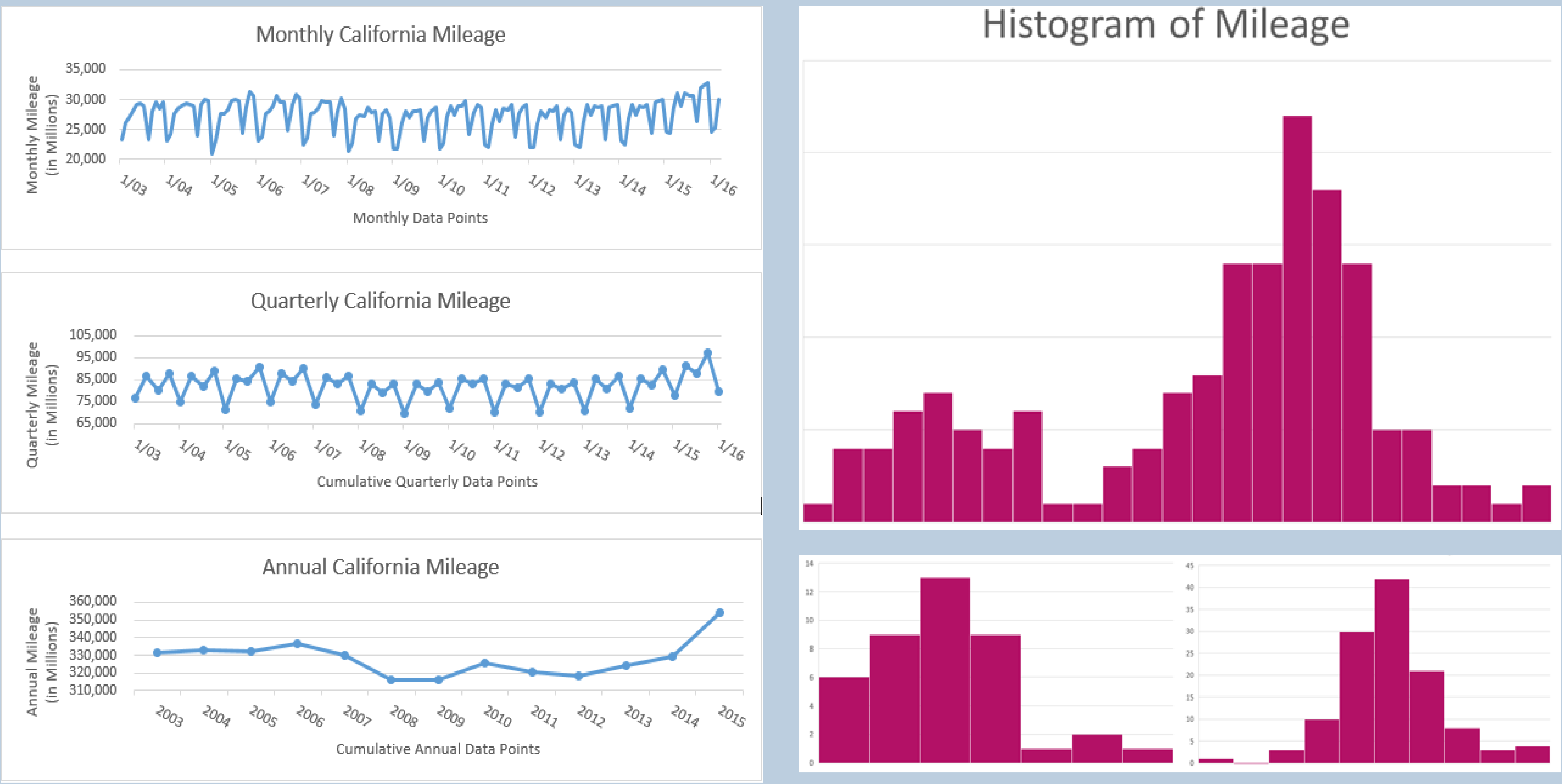

Figure 1: Graph of monthly mileage (left), histogram of data (top right), histogram of overlapping normal distributions (bottom right). The seasonal indicator was used to account for the bimodal distribution.

Model Selection

Once the variables were paired down using R^2, forward selection was used to determine the optimal combination of indices by adding one variable at a time to the model and testing for significance. Four models were eventually selected for testing and training. Then, the data was split into a testing set consisting of twelve random observations and a training set consisting of the rest of the data. Another training set consisting of the last twelve observations was created. Each model had its predictive power assessed using mean squared error (MSE). The combination that gives the lowest MSE was determined to be model 3.

Model	Random MSE	Last 12 MSE
Model 1	1,321,008	4,099,134
Model 2	800,831	1,644,963
Model 3	744,013	1,615,029
Model 4	10,722,138	5,980,309

Table 1: Table of Mean Squared Error

Forecasting

Model 3, using the predict function in R, used forecasted indices to generate mileage from April 2016 to March 2017. The forecasts were relatively flat due to the low volatility in the projections.

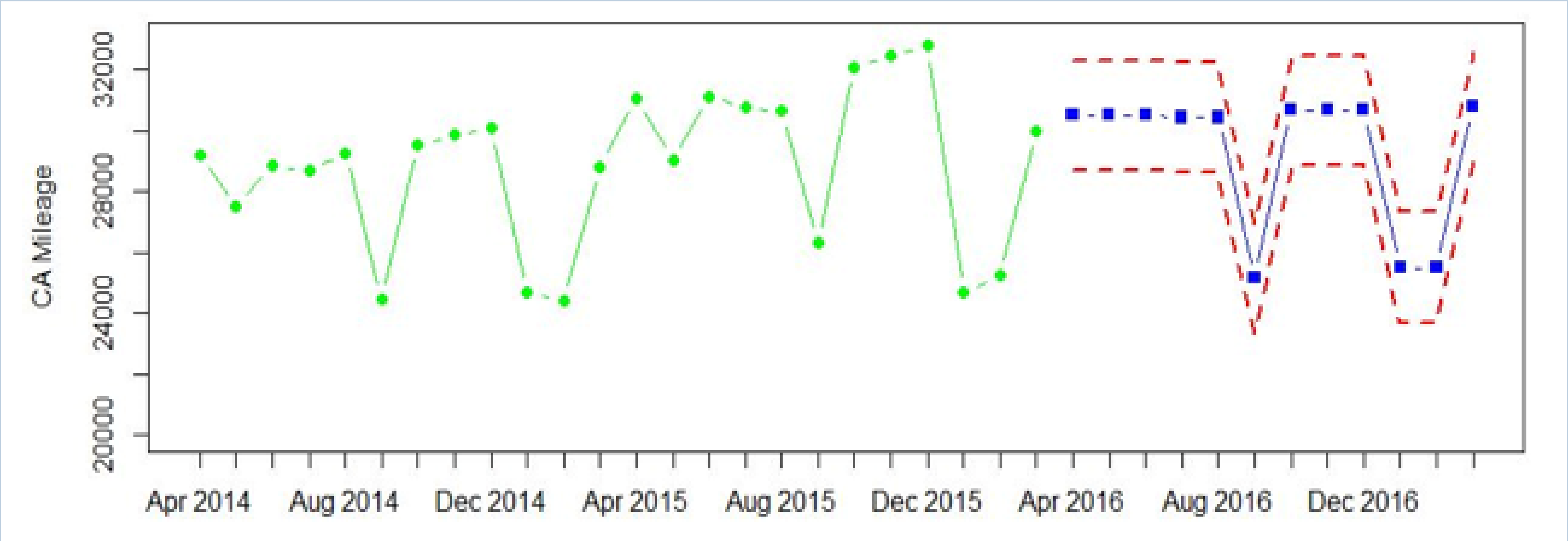

Figure 2: Forecasted mileage for the next 12 months using model 3. The green dots represent actual data, the blue squares are the forecasted mileage, with the dashed red line representing the 95% confidence interval.

Final Equation

The final formula is shown below. The seasonal indicator separates the two histograms.

$$\text{Mileage} = 12{,}046 - (5583 * \text{Seasonal Indicator}) + (1756 * \text{US Unemployment Rate})$$
$$+ (1127 * \text{US Total Vehicle Sales}) - (124 * \text{US Unemployment Rate} * \text{US Total Vehicle Sales})$$