Project 1 Proposal

Analyzing Vehicle Recalls, 1967-Present

Math 519

William Gillespie

2017/02/16

My project proposal is based on the Kaggle dataset of vehicle recalls submitted by the Department of Transportation’s National Highway Traffic Safety Administration (NHTSA). The dataset is record-based; each row is a unique recall event (although sometimes several rows are variations on the same recall event, such as when a component that is used across several models or years is subject to recall), and each column is a field such as vehicle make, model, model year, recall component (structural or electrical), estimated units recalled, and the recall notification date. The link to the data is here: <https://www.kaggle.com/nhtsa/safety-recalls>.

I chose this dataset because I thought it was interesting and had lots of possible questions of interest. I plan to concentrate on the time series aspect of the data, i.e., patterns in the recall data over the years. I had originally wanted to do a project using an AR model, but am unsure whether this dataset would fit this model. I do have some concern that recall data may not have much autocorrelation, especially after removing trends over time. This will be interesting to explore. I plan to first detrend the data to remove any general patterns over time, such as an anticipated exponential increase. I will also check the autocorrelation function for the number of lags that still have some correlation, positive or negative (compared to having non-significant correlation, essentially noise). Several possible models of the remaining correlation will be investigated, including ARMA and ARIMA.

I also plan to look for time series across the same years as the recall data with data on car ownership in the U.S. I will use the data from the energy.gov at <https://energy.gov/eere/vehicles/vehicle-technologies-office>. These data cover the year ranges I am interested in. I intend to calculate the cross correlation between the number of vehicles affected by recalls and the number of vehicles per 1000 people.

One possible predictor of recalls is economic downturns, which may prompt automakers to skimp on quality to make cheaper cars. I could detect such an effect from a cross correlation function of recall data with economic indicators over time. I will look for datasets of various economic indicators to test this hypothesis.

Another topic I would like to explore is trends in recall data for each recall component (structural or electrical). My hypothesis is that structural components will have a decreasing trend in recall events over time, while electrical components will tend to have increased recalls over time (perhaps due to more features which means that more things can go wrong). I will test for decreasing mechanical recalls over time and increasing computer component recalls over time. Another trend I would like to explore is whether there is a difference between domestic and foreign manufacturer recalls.

There have been historic recalls involving huge numbers of cars. I could identify some of these historic recalls (anomalies) and model how often they happen.