HW1 Proposal

RJ Cass

1. Understanding the Problem

Background:

Kelley Blue Book (KBB) has extensive car selling data: car conditions, mileage, packages, etc., associated with the selling price. We want to use this data to accurately guide consumers on what their car is worth, and how much they could expect to sell their car for.

Goals:

In this analysis, we want to address several key questions:

i. What factors lead to higher/lower resale values?

- We will perform variable selection to determine which factors or impactful to price. Given scale of the data and variables, we will use hybrid stepwise selection (both forward and backwards) to select the model with the appropriate covariates.

ii. Are there other factors not included in this dataset that likely explain how much a car is worth? If so, what other factors do you think explain resale value?

- We will determine what other factors may be impactful by consulting expertise (our own experience, search online) as well as viewing the R^2 value from our model (if R^2 is low, it indicates there are variables missing that would help better explain the relationship to price).

iii. Generally, as mileage increases, the price should decrease. But, does the amount of decrease in value from additional mileage differ depending upon the make of the car? If so, which makes hold the value better with more miles?

- To determine how the make of the car impacts the effect of mileage on the price of the car, we will need to see if the interaction of Make and Mileage is significant. If it is significant we need to ensure it appears in any models we create.

iv. Which car (and with what characteristics) has the highest resale value at 15000 miles?

- To see which car has the highest resale value at 15k miles, we will build a predictive model, plug in a value of 15k miles, and see what values for the other variables give the highest price.

v. What is a reasonable resale value for the following vehicle: Cadillac CTS 4D Sedan with 17,000 miles, 6 cylinder, 2.8 liter engine, cruise control, upgraded speakers and leather seats?

- To identify what is a reasonable price for a car with the above values, we will create a predictive model in which we can enter the relevant values, and it will predict a price range that the car is worth. This will require that our model by able to accurately predict price given the variable inputs.

2. Exploratory Data Analysis

In our initial exploration, the first things that stick out about the data is that price is pretty heavily skewed with a long right tail. This violates the assumption of normality so we may need to perform a transformation. We may also need to address colinearity between variables. We did not identify any other concerning factors (all fields have complete data - no missing values, etc.)

3. Desired Attributes

Model attributes required from the research questions

The research goals require an interpretable model that gives predictions, R^2, and that can include Mileage*Make interactions

Model attributes required from the data

The data show we may need to address collinearity and non-normality.

Any other anticipated problems

I do not anticipate any other issues in the data.

What goes wrong if the above are not accounted for?

Without meeting the requirements of the research questions, I will not be able to answer those questions. If I don't address collinearity, my coefficient estimates won't be true. If I don't address non-normality, the standard models' likelihoods won't be appropriate.

4. Proposed Method

Appropriate models

In this analysis, a linear regression model is appropriate.

Specific model proposal

I will use Oridnary Least Squares (OLS) regression using the format:

$$Y = \sum X_i \beta_i + \epsilon : \epsilon \sim N(0, \sigma^2)$$

Method strengts/weakness

i. How this method accounts for issues in the dataset

We can perform transformations in OLS to account for non-normality. We can use stepwise selection to help with collinearity.

ii. How this method accomplishes research/analysis goals and yields appropriate estimators

For the goals of the analysis, regression can have interactions and predict, meeting the requirements of the research questions.

iii. How this method will answer the research questions

This method will answer the research questions by (for each question) 1) stating the included variables, 2) state R^2 value and explain ideas for non-included variables that may be significant, 3) report significance of interactions, 4) predict appropriately 5) predict appropriately

iv. What assumptions are needed to use the model adequately? Are they reasonable to assume and explained well?

To properly use this model we muse assume linearity, independence, normality, and equal variance. Linearity appears safe to assume, and we have toolks to address the other assumptions should they turn out not to be true.