Machine Learning:

Learning from Data Using Linear Regression

Wright Analytics

Introduction

- The linear regression model is a parametric learning method that can be used to model the relationship between a ***continuous*** dependent variable and one or more ***continuous*** independent variables

- What do we exactly mean by “parametric learning method” and continuous independent variables? We will explain shortly

Introduction to Simple Linear Regression

- Linear regression is a ***parametric learning method*** because the function used to "describe" or "model" the dependency between ***one*** dependent variable and ***one or more*** independent variables has a ***specific form*** that we assume is appropriate for the application

- The form of the ***simple linear regression model*** is

$$E[Y \mid X] = \beta_0 + \beta_1 * X + \epsilon$$

Introduction to Simple Linear Regression

- So the ***simple linear regression model*** is a ***line***. And $E[Y \mid X]$ is the ***conditional expectation*** of Y given X.

- β_0, β_1 and ϵ are ***model parameters*** which are ***unknown***

- In linear regression, we estimate these parameters with ***continuous information***

Introduction to Simple Linear Regression

- Furthermore, we previously stated that $E[Y \mid X]$ is ***equal to*** a line. This assumption makes linear regression a parametric learning method

- Note, this model is appropriate to apply when an ***obvious*** or ***approximate*** linear relationship is observed in a ***scatterplot*** of variables

Introduction to Multiple Linear Regression

- Multiple linear regression is an ***extension*** of simple linear regression
- Multiple linear regression can be considered when ***two or more*** independent variables are available
- Multiple linear regression attempts to explain the ***impact*** each independent variable has on Y

Introduction to Multiple Linear Regression

- The ***form*** of the multiple linear regression model is

$$E[Y \mid \vec{X}] = \beta_0 + \beta_1 * X_1 + ... + \beta_n * X_n + \epsilon$$

  and contains *n* independent variables where *n* is just a number

- Again, the parameters are unknown and need to be estimated

A Note about Parameters and Continuity

- In either the simple or multiple linear regression model, the parameters are ***real numbers*** as well as the ***continuous*** information contained in the dependent and independent variable(s)

- The parameter ***estimates*** minimize a quantity called the ***residual sum of squares*** also known as the ***RSS***

- Once the parameters are known, we can make ***predictions***

A Note about Parameters and Continuity

- The idea of ***continuity***, or ***continuous*** independent variables, is somewhat abstract. Since this is the case, we explain the concept of time as a continuous quantity

- Many units can be used to ***measure*** time. Time can be expressed in years, months, weeks, days, hours, minutes, seconds, milliseconds, etc.

- So time be ***continuously divided*** into smaller and smaller intervals/ units of measurement. This the concept of continuity

Project Results: Overview

- In the project ***Machine Learning: Learning from Data Using Linear Regression***, we describe how we search for a robust linear regression model we can use to predict a vehicle's ***miles per gallon*** rating based on vehicle ***features***

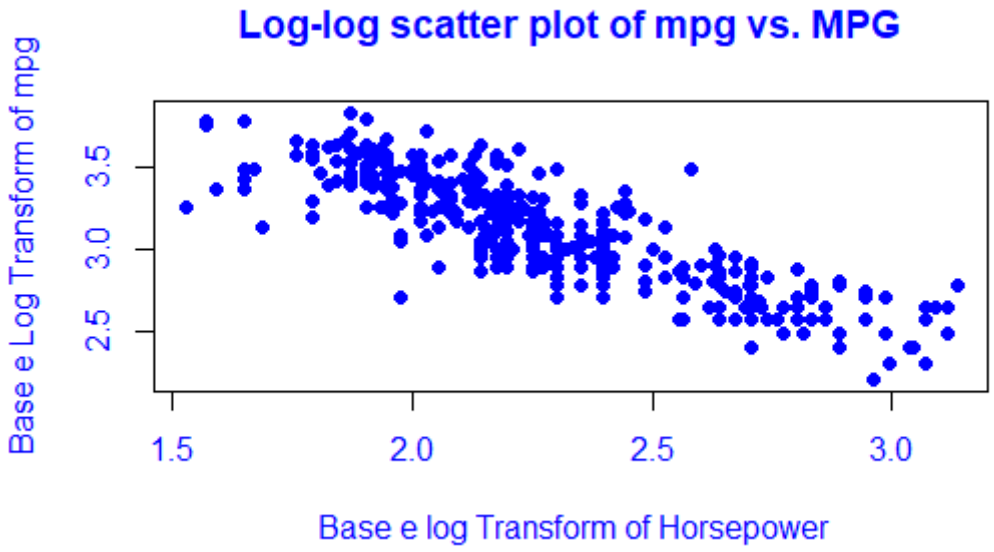
- In the project, we present our modeling process and search for robust linear regression models

Project Results: Overview

- The “best” linear regression models we were able to find contain ***transformed*** variables
- Transformations of variables are often used to obtain ***linearity*** or capture ***non-linear*** relationships that appear to be present
- During our search, we were able to find a robust simple and multiple linear regression model. We next discuss them

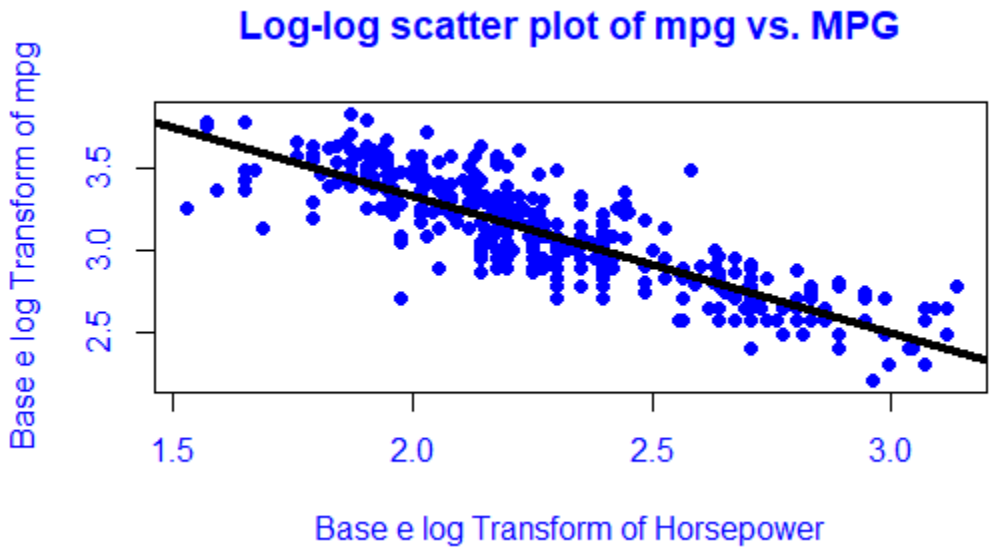
Project Results: Simple Linear Regression

- We applied natural logarithm transforms to the data contained in the dependent ***mpg*** and independent variable ***horsepower*** to obtain linearity. The scatterplot found below shows this



Project Results: Simple Linear Regression

- A scatterplot with the estimated model is found below. The solid black line is the estimated model



Project Results: Simple Linear Regression

- The estimated model written down is

$$\ln(\hat{E}[Y \mid X]) = 5.02 - 0.842 * \ln(Horsepower)$$

where the ^ symbol indicates an ***estimation***

Project Results: Simple Linear Regression

- So for a car with that has a 150 horsepower engine, we expect that vehicles mpg to be

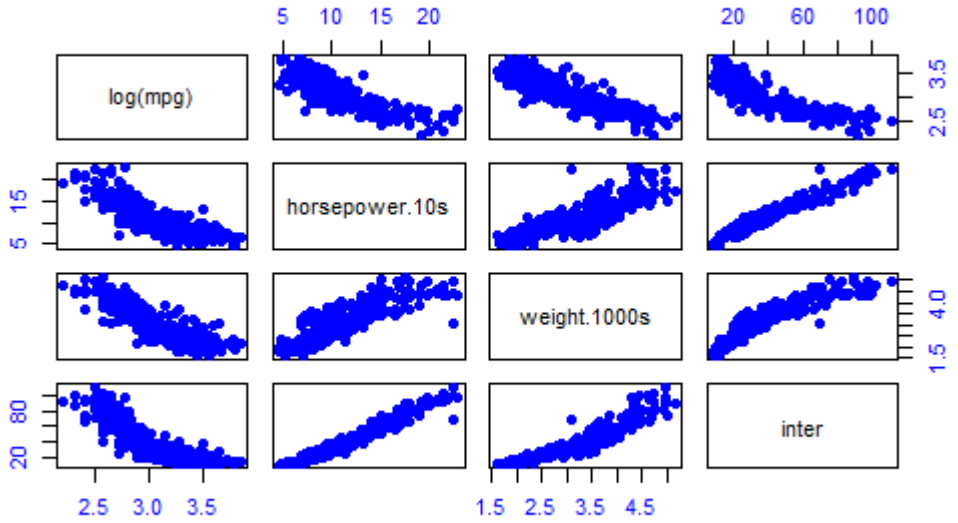
$$\ln(\hat{E}[Y \mid X = 15]) = 5.02 - 0.842 * \ln(15)$$

- ***Note, the original values of the variable horsepower have been divided by 10***

- So the predicted value in the transformed scale is 2.74. Exponentiating the result yields 15.49 miles per gallon

Project Results: Multiple Linear Regression

- To assess linearity in multiple linear regression, we produce ***scatter plot matrices***



Project Results: Multiple Linear Regression

- Looking at the top row of the plot, we see that the relationship between the natural log of mpg and independent variables horsepower, weight, and the interaction of horsepower and weight is linear

- The other variables we include in the model are American.Dummy, European.Dummy, and X4.cylinders.dummy which are ***indicator variables*** that represent different categories.

Project Results: Multiple Linear Regression

- American.Dummy takes on a value of 1 if the vehicle is manufactured in America. It takes on a value of 0 if the vehicle is manufactured in Japan or Europe

- European.Dummy takes takes on a value of 1 if the vehicle is manufactured in Europe. It takes on a value of 0 if the vehicle is manufactured in Japan or America

Project Results: Multiple Linear Regression

- X4.cylinders.dummy takes on a value of 1 if the vehicle's engine has four cylinders. It takes on a value of 0 if it does not.

- The idea of ***dummy/indicator variables*** is that they can be used to quantify how certain ***qualities*** of an observation ***impact*** the dependent variable. Qualities are ***not*** numbers. However, they can be ***represented*** with numbers

Project Results: Multiple Linear Regression

- The estimated multiple linear regression we obtained is

$$\ln(\hat{E}[Y \mid \vec{X}]) = 4.112 - 0.060 * X.1 - 0.067 * X.2 + 0.121 * X.3 - 0.053 * X.4 - 0.235 * X.5 + 0.007 * X.4 * X.5$$

- ***Where X.1 is American.Dummy, X.2 is European.Dummy, X.3 is X4.cylinders.dummy, X.4 is horsepower, and X.5 is weight***

- Note, the original values in horsepower have been divided by 10. Likewise the original values of weight have been divided by 1000

Project Results: Multiple Linear Regression

- So if a vehicle manufactured in America, weighs 3500 pounds, has a 155-horsepower engine, and is not a four cylinder we expect that vehicle to get

$$\ln(\hat{E}[Y \mid \vec{X}]) = 4.112 - 0.060 * 1 - 0.053 * 15.5 - 0.235 * 3.5 + 0.007 * 15.5 * 3.5$$

- The predicted value in the transformed scale is 2.788. Exponentiating the result yields 16.28 mpg