MATH 4044 – Statistics for Data Science

Practical Week 5

The data for this practical is stored in a SAS data file called `kbb.sas7bdat` located in `mydata` library on the SAS OnDemand server. Variables in that file are as follows:

Variable	Description
Mileage	Number of miles the car has driven
Make	Make of the car
Model	Specific model for each manufacturer
Trim	Specific type of car model such as SE Sedan 4D, Quad Coupe 2D
Type	Body type, e.g. sedan, coupe etc.
Cylinder	Number of cylinders in the engine
Liter	A more specific measure of engine size
Doors	Number of doors
Cruise	Indicator (binary) variable representing whether the car has cruise control (1 = cruise)
Sound	Indicator variable representing whether the car has upgraded speakers (1 = upgraded)
Leather	Indicator variable representing whether the car has leather seats (1 = leather)
Price	Suggested retail price of the used 2005 car in excellent condition.

The data was collected from Kelly Blue Book http://www.kbb.com for several hundred 2005 used General Motors (GM) cars. The goal is to develop a multivariate regression model to determine car values based on a variety of characteristics such as mileage, make, model, engine size, interior style, and cruise control. All cars in this data set were less than one year old when priced and considered to be in excellent condition.

(a) Fit a simple linear regression model with *Price* as the dependent variable and *Mileage* as the independent variable. Discuss the resulting model in terms of goodness of fit.

(b) Use model selection techniques available in SAS to identify a better model with multiple predictors. A good model should have a high R-squared and adjusted R-squared, and a C_p value that is close to the number of predictors contained in the model.

(c) Fit the model identified in part (b) and discuss goodness-of-fit. Also examine and discuss residual patterns. Are there any issues with collinearity?

(d) Create dummy variables based on the makes of cars in this data set (Buick, Cadillac, Chevrolet, Pontiac, SAAB, and Saturn). Also create a new variable *LPrice* = log(*Price*).

(e) Fit a multiple regression model with *LPrice* as the dependent variable and *Mileage*, *Cylinder* and *Make* dummy variables as explanatory variables. Examine residuals and comment on the goodness of fit. Try other models including additional explanatory variables and comment on the results.