In this project, we want to justify whether the Generic Substitute (GS) and Restriction Index (RI) are efficient in controlling drugs costs. We analyzed collected data of “COST” (response) and related seven potential explanatory variables (including GS and RI) with several quantitative steps with different methods. And we drew a conclusion that …. The specific analysis process is as follows.

1. EDA

First, we explored the relationships between response and each variable. By analyzing correlation matrix, histogram and plots between each two variables, we found some potential problems:

1. Variable RI and COPAY may have co-linearity (correlation = 0.72)
2. Most variables are not normally distributed (from histogram)
3. There may exists influential outliers (point 19’s cook’s distance> 1)

2. Data transformation and cleaning

With problems discovered above, we did data transformation and cleaning respectively to each problem:

1. Using unconditional box-cox transformations to each variable. The rounded lambda for each variable that needs transformation is (GS=2, RI=0.5, COPAY=0,MM=0.5)
2. Testing VIF of the full model with transformed data.
3. Detecting outliers with transformed data.

We get the following results:

1. After using unconditional box-cox transformations for each independent variable, all variables are normally distributed.
2. After variable transformation, all VIFs are less than 5, so there doesn’t exist co-linearity.
3. After transformation, no point has cook’s distance greater than 0.5, so there doesn’t exist influential point.

Thus, we will use transformed data to fit models.

3. Fitting Models

First, we fit linear model using full subsets

1. 1