**Predicting Performance of Integrated Circuits using Regression Analysis**

**Part Two**

**Joanna Duran, Nikhil Gupta, Max Moro**

# Introduction

This project is a continuation of Project 1 for course MSDS 6372 “Applied Statistics”. From project 1, we identified a couple of possible improvements that can be made to the model fit. These improvements are performing intelligent feature engineering and selection and cluster high Cook’s D points. In this project, we will act upon these ideas using the techniques learned in the second half of the course to see if we can improve the model fit.

**Synopsis of Project 1**

Semiconductor manufacturing is a variable process and outcomes depend on several factors. In order to meet target specifications, some parameters are controlled by engineers. However, some parameters are beyond human control (e.g. process variation). Variation in the process leads to issues if the outputs are outside target specifications.

Data for this project has been sponsored and approved for use by Texas Instruments Inc. (TI) who provided two files. Due to proprietary nature of the information, the variables have been anonymized.

The original data set consists of 10,000 rows capturing the performance of an integrated circuit under various conditions. There are 240 features consisting of:

* Engineer-controlled variables (x1 – x23). Values differ, some are between 1 to 100 while others are in Nano or Micro range.
* Process variation variables (stat1 – stat217). These parameters are beyond human control. They represent various statistical manufacturing parameters whose values represent the sigma variation around the mean. Range is from -3 (sigma) to 3 (sigma).
* Output Variables (y1 - y19) which represent various output variables.

Like most data, this dataset also needed some cleaning. Basic descriptive statistics revealed that 3020 NA values were present. After consulting with the expert from TI, we found that the predictors (features) for these data points were not practical in combination with each other. Hence, these points are not valid and can be removed without impacting the predicting power of the model being developed. We cleaned up labels and removed NA values. Once the data was cleaned, the datasets were merged.

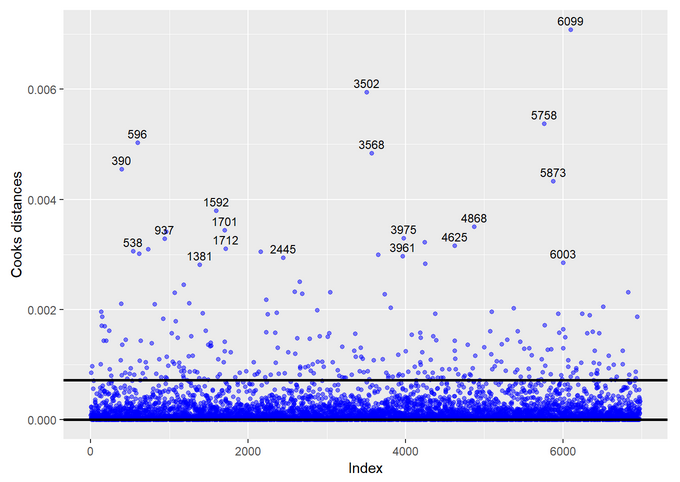
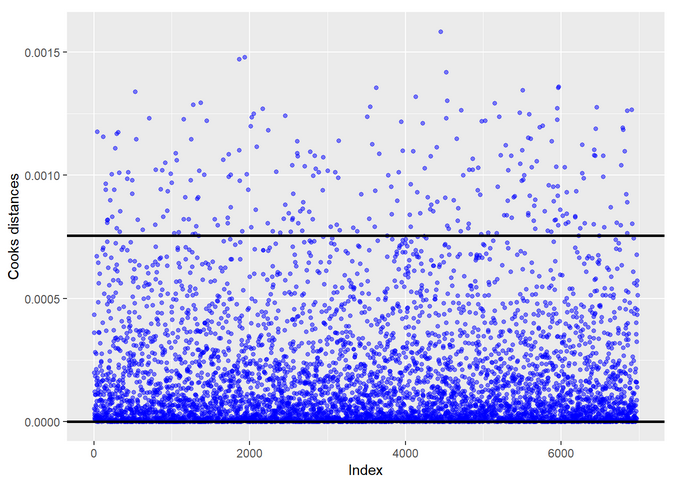
We began exploratory data analysis on the target variable y3 and notice right skewness, therefore we perform a log transformation. Log transformation makes the data a little more normal, therefore we proceeded with the log transformed variable “y3.log”

Since inputs were randomly selected, we did not expect there to be multicollinearity. After running the analysis, the correlation table and VIF values confirmed that there is no multicollinearity.

One of the highly correlated features, x18, shows a slight curvature in the scatter plot therefore a square root transformation was performed. After the transformation, the scatter plot showed slightly better linearity. Based on our observations we proceed to modeling with the target variable y3 as log transformed and the predictor x18 transformed with square root.

The full model analysis was performed by analyzing the full model consisting of all 240 features with the transformations. The adjusted is 0.2275 which is low and expected based on low correlation of features to outputs.

The influence statistics were also analyzed and from the Cook’s D plot we notice that there are 288 points beyond the 4/n line (Fig 1). It was tempting to remove these points because it would improve the model fit drastically (Fig 1) but after consulting with domain experts, we found that these are valid points and there is no justification to remove them.

*Fig 1: Cook’s D before and after*

It was determined that variable selection may be appropriate at this step to remove unnecessary predictor variables. The variable selection techniques that were used were Forward Selection, Backward Elimination, Stepwise Selection, LASSO, and LARS.

We found that all techniques give essentially the same fit statistics. The best model is Backward Elimination. It was chosen with one of the reasons being that it has the least number of predictors. RMSE was used to choose the final model during backward selection. The lowest value of RMSE obtained for the model with 13 variables. We also note that the number of predictors selected would have been the same even if we would have chosen metric MAE or instead as RMSE for final model selection.

In analyzing the residuals, we noticed the histogram of the residuals shows to be right skewed and the equality of variance at lower values of prediction are questionable as before. We conclude that the final model suffers from same assumption violations as the full model.

# Goal for project 2

The goal of the second project was to Improve the model fit from Project 1 utilizing Principal Component Regression, Linear/Quadratic Discriminate Analysis and Logistic Regression. Target accuracy of ±10% is desired, but ±15% would be acceptable.

# Approach 1: Principal Component Regression with intelligent feature selection

## Intro

In project 1 we realized that doing a brute force interaction of all the predictors would give us more than 28,000 predictors in the model which would not be practical to use. In this project, we will perform feature engineering to create new features (using the ones already available) and do intelligent interactions with a subset of the available predictors. Both feature engineering and intelligent feature interaction will be done by taking domain expertise into account. We have about 6,000 observations in our dataset and we need at least 10 observations per feature to avoid overfitting. With interactions included, we can only have about 600 features. However, our dataset already consists of over 240 features and after doing intelligent feature engineering and interactions, we can easily end up with more than 600 features. Hence we will finally perform PCR to reduce the dimensionality of the predictors and build a linear model using our dataset.

# Approach 2: Clustering based on High Cook’s D and using the cluster as a predictor in the regression model

## Intro

In project 1 we found that the distribution of the predictors for the High Cook’s D points were very different from those of the low Cook’s D points. If we can classify the observations as belonging to the high Cook’s D group or the low Cook’s D group, we can possible add that as a predictor variable to improve the fit of the model. Since we don’t know which points belong to high Cook’s D group and the low Cook’s D group, we will run a basic linear regression model to determine and label. Using this label, we will perform LDA/QDA/Logistic Regression to train a model to predict the right group for a new observation. We will add the predicted group (from LDA/QDA/Logistic Regression) as a feature into the linear model for predicting the value of the output variable. In a way this is a 2-step machine learning model - first cluster the incoming observation and then use the cluster information as a predictor in the final model to improve its fit.

# Appendix

*Project 1 paper*

Refer to [this link](https://github.com/ngupta23/MSDS-6372-Project1) for detailed and reproducible analysis and [this link](https://htmlpreview.github.io/?https://github.com/ngupta23/MSDS-6372-Project1/blob/master/Analysis/Master_Submitted/Master.html) for the final HTML file of the analysis.