Abstract

This report attempts to analyze the performance of sixty-five Military Entrance Processing Stations (MEPS) in regards to the aptitude, medical, and miscellaneous accession processing functions. Linear regression is utilized to model the percentage of points a MEPS earned in a year, using quarterly predictor variables from the previous year. Multiple models are considered with the best predictive model chosen as the final model. Finally, a logistic regression model is used to predict whether a MEPS will win at least one Unit Pennant award in the next year.

Table of Contents

Pages

List of Tables

List of Figures

I. Introduction

Background

As noted in the project instructions, this project is motived by concerns of a systemic advantages within the MEPS scoring system for certain awards and other applications. We have been tasked to conduct an analysts and quantify the relationship between MoE metric scores for Year n and MoE aggregate scores for Year n+1.

II. Methodology

Exploratory Data Analysis

Before we discuss any linear regression model formulation, we perform some initial exploratory analysis of the data. Immediately we remove some predictor variables from consideration of including in our model due to their lack of information provided. These include variables such as Total Students Tested and Supervisor Training since these variables only contained NA values as well Timeliness of Awards and Timeliness of Evaluations since they only contained 1 factor level and NA values, thus providing no meaningful information. We also only keep CLIP_Q3[1] since all other quarters only have NA values. By centering and scaling the relevant numeric data, we can consider the correlation matrix between the predictor variables to note that no variables have a correlation greater than 0.5 which indicates multicollinearity should not be a major issue. Finally, we establish the variables corresponding to TLC, CBA, and IBA to be factor variables.

Model Formulation

Our model development process utilizes a number of performance measures and diagnostic checks in order to determine the best combination of variables and transformations so that our model's predictive power is highest. For each model under consideration, we use diagnostic checks to verify the linear regression assumptions of linearity, homoscedasticity, independence between error terms, and normality. We then compare the models with the Adjusted R^2 value, the root mean squared error (RMSE), the

[1]See Appendix for variable names and acronyms

Akaike's Information Criterion (AIC), and Leave-One-Out Cross Validation prediction error (LoOPE) to determine the optimal model for the given data. The final model includes the following predictor variables to model the FY19 Percentage:

> FY18 Percent, CBA_Q1, log(CICO_Q1), HIV_Q1, TNG_Q1, log(CICO_Q1):HIV_Q1, CBA_Q2, DSP_Q2, CLIP_Q3, log(CICO_Q3), TNG_Q3, DSP_Q3, CLIP_Q3:DSP_Q3, log(CICO_Q3):DSP_Q3, log(CICO_Q4), HIV_Q4, FBP_Q4, log(CICO_Q4):HIV_Q4,

where the notation "x:y" indicates an interaction term where the variable x is multiplied by the variable y and is considered a single variable. We will continue by explaining our process of how we came to our final model by reviewing some choice models that we considered. Table 1 summarizes the performance measures for the models discussed below.

We begin our model selection process by considering the model that includes all available predictor variables with no interaction terms (Model 1). As we do for all models considered, we first conduct the diagnostic checks by looking at the Fitted vs Residuals plot, the Scale-Location plot, and the Normal Q-Q plot to check the linearity, homoscedasticity, and normality assumptions, respectively. We follow these checks with statistical tests to confirm any diagnostic check that is not clear from the respective plot. We use the Breusch-Pagan test and the Shapiro-Wilk test to verify the homoscedasticity and normality assumptions while we use the Durbin-Watson test to check the independence of the error terms. For each model discussed in this report, the model diagnostics are satisfied so these are never used to dismiss a model from consideration.

As such, we only discuss in more detail the diagnostic checks for the final model in the next section of this report. Of note for Model 1 is that since CBA_Q3 only has one instance of level 0, the cross-validation prediction fails for that instance.

In searching for an optimal model, we utilize many tools and tests to identify the key variables to include in the model. The Variance Inflation Factor (VIF), Residual vs Covariate plots, and Partial Residual vs Covariate plots are some of the methods used to select our final model. For brevity, additional insights into formulation can be found in the R code comments regarding the model formulation process. From the aforementioned methods, we discover some variables are heavily correlated with other variables and can be removed while others require transformations. This brings us to Model 2 which includes logarithmic transformations to the CICO quarterly variables while accounting for interaction between the HIV quarterly variables and DSP quarterly variables with the CICO variables as well as the CLIP_Q3 variable with DSP_Q3 and HIV_Q3. Logically, these variables potentially have the most interaction between each other.

Table 1. Model Performance Measures

Model	Adj. R^2	RMSE	LoO PE	AIC
Model 1	0.9004	0.04341	NA	-149.35
Model 2	0.2319	0.04057	0.0223	-160.16
Model 3 (log-odds)	0.2151	1.04272	1.8365	126.46
Model 4	0.4147	0.04288	0.0052	-180.93

These changes to the model seem to have made some improvement to the model fit as seen with the increased adjusted R^2 but at a cost of increased prediction error.

In a separate vein of analysis, we also consider modeling the log-odds of FY19 Percentage instead to map the dependent variable onto the set of real numbers, $\mathbb{R}$, as a Gaussian variable would be. Testing this against multiple combinations of predictor variables, Model 3 was the best performing of them which used the same predictors from Model 2. Regardless, we see a significant increase in the error performance measures as well as the AIC value compared to the previous models considered, and so we return to modeling the FY19 Percentage.

Model 2 seems to be the best performing model so far, but we want to reduce the possibility of over-fitting our model, noticing some variables with high VIF, as that could negatively impact the prediction capabilities of the model. To this point, we use a backwards step-wise model selection process with the minimum AIC as the selection criteria. This model yields the final model described above and is reported in Table 1 as Model 4. We see a marked improvement in all the performance measures compared to Model 2, and so we finish our model selection process with this as our final model. In the next section we will discuss the diagnostic checks associated with this model to verify the model assumptions are satisfied.

Model Validation

To verify the assumptions of linear regression modeling, we look at certain plots and conduct statistical tests to confirm that the model is linear, with independent error terms that follow a normal distribution with constant variance. Breaks in some of these

assumptions can be seen from the plots in Figure 1. The lack of any pattern in the Residuals vs. Fitted plot confirms the model is linear with additive residuals. Meanwhile the Normal Q-Q plot appears to also confirm the residuals mostly follow a normal distribution, but we verify this statistically with a Shapiro-Wilk test, seen in Table 2, as the residuals do deviate from the standard Normal quantiles on the right tail. The lack of pattern in the Scale-Location plot and flat locally fit line indicates the residuals have a constant variance and are independent of each other. These are also verified with a Breusch-Pagan test and Durbin-Watson test, respectively. The results of the statistical tests seen in Table 2 allow us to conclude that the model assumptions are satisfied.

Table 2. Statistical Tests for Model Assumptions

Statistical Test	Test Statistic	P-value
Shapiro-Wilk	$W = 0.9768$	0.2606
Breusch-Pagan	$\chi^2 = 17.379$	0.6283
Durbin-Watson	$d = 1.8928$	0.626

The final graph in Figure 1 plots the Residuals vs. Leverage with Cook's Distance overlaid to give an indication to observations that have high leverage and influence. These observations and others will be investigated further in the following section.

Analysis of influential observations

There are some observations that appear to have high leverage, and we want to investigate whether these observations have significant influence on the overall fit of our model. To start this investigation, we consider the DFBETA plots in Figures 2, 3, and 4 and the Influence Plot in Figure 5 while considering the observations with the largest studentized residuals. This analysis leads us to consider observation number 39 that

corresponds to the Pittsburgh MEPS to be a potential outlier with a reasonable amount of influence on the overall fit of the model. However, there does not appear to be a value for any single variable that seems a data entry error was committed, nor does it appear that any values are an outlier for any single variable. With that in consideration, we have no reason to remove the observation.

Logistic Model

Finally we develop a logistic regression model to make predictions on a MEPS winning a Unit Pennant award in the following year. Since there are four quarters in which a MEPS could win the award, we assume our response variable to be a Binomial random variable which represents the number of Unit Pennant awards a MEPS wins in a year with parameters n equal to four and unknown p. To make this model we create a variable in the FY19 data that counts the number of Unit Pennant awards each MEPS won according to their percentage each quarter and regress this variable against the quarterly predictor variables in the FY18 data.

III. Present the model, results, and analysis

Model

The final linear model can be expressed as

$$Y = \beta_0 + \sum_i \beta_i X_i \text{ , for } i=1,\ldots,21$$

where Y is FY19 Percent and the X_i's are the predictor variables listed in the Model Formulation section. Table 3 in the Appendix displays the β parameter estimates.

Following a similar model formulation process as described above though with additional steps to verify the logistic regression assumptions, we develop a model that has small LoO prediction error equal to 0.0869 and RMSE equal to 0.8331 as well as a decent fit to the data, with McFadden's pseudo R^2 equal to 0.2810 to measure model fit. The final model chosen is defined through the following:

Y_i: Number of Unit Pennant awards i^{th} MEPS wins in a year
$Y_i/4 \sim \text{BinProp}(p_i)$, for i=1,…,65
$\eta_i = logit(p_i) = log(p_i/(1-p_i))$
$\eta_i = \beta_0 + \sum_j \beta_j X_{ij}$, for j=1,…,15 and i=1,…,65

where the selected predictor variables X_j are Per_Q1, CBA_Q2==5, CBA_Q2==10, Per_Q2, CLIP_Q3, DSP_Q3, log(CICO_Q3), Per_Q3, log(CICO_Q4), HIV_Q4, Per_Q4, CLIP_Q3:DSP_Q3, DSP_Q3:log(CICO_Q3), log(CICO_Q4):HIV_Q4. The parameter estimates are presented in Table 4 in the Appendix.

Results

The predictions for FY2020 are presented in Table 3 and Table 4. For each MEPS, we present the FY2020 Percent prediction and whether they would win the MOY

award, an MOQ award for each quarter all according to our linear regression model, and whether they will win at least one Unit Pennant in 2020 per our logistic regression model.

Analysis

In the linear regression model predicting FY19 Percentage, we find a number of our parameters to be statistically significant, meaning their corresponding predictors have a non-zero influence on the future Percentage. We find the most influential predictor to be obvious, with the FY18 Percentage having a parameter estimate of just under one, and HIV_Q1 is the second most influential predictor. Meanwhile, in the logistic model the most influential predictor is CICO_Q3 with DSP_Q3 or CBA being the next most influential, depending on the value of CICO_Q3 due to the interaction effect it has with DSP_Q3.

IV. Discuss the results and provide insights and recommendations

Insights and Recommendations

Regarding the Gaussian model, our predicted results seem to be consistent with the past results; however, the somewhat small sample size does raise concern for the accuracy of our inference. One possible consideration is to utilize bootstrap sampling to replicate more observations using the variables' empirical distributions. Or we could expand our model by using a time series approach and utilizing the additional historical data. Additionally, one could consider using a different set of performance measures to determine the best model such as the Mean Absolute Error or use k-fold Cross Validation for k greater than one.

Regarding the logistic model, one concern is the potentially large dispersion parameter. For the final model chosen, the estimated parameter was around thirteen which is larger than the theoretical value of one. To account for this, we could attempt to model this over dispersion by using a Beta-Binomial distribution.

Discussion

A surprise in both our linear regression and logistic models is the predictor influence of CLIP_Q3 and DSP. We would expect a positive parameter value from these predictors as we expect a larger value in the measures to improve the MEPS's performance, but our model indicates the opposite effect. But these models were both constructed for predictive power and not inferential power so we did not tailor our model parameter estimates for interpretability.

Appendix A. Variable Name Abbreviations

Variable Name Abbreviations
CICO: Check-in/Check-out
TLC: Test Loss Compromise
DSP: Drug Specimen Processing
HIV: HIV Sample Processing
CLIP: CLIP
FBP: Accuracy of Fee-Basis Provider Work Hour Data
IRP: Accuracy of Invoice Reconciliation Program
CBA: Citibank CBA
IBA: Citibank IBA
ToA: Timeliness of Awards
ToE: Timeliness of Evals
TNG: Training
Per: Percentage
*_Q#: denotes the quarter

Appendix B. Additional Tables

Table 3: Linear Regression Model Estimates

Predictor variable	β parameter estimate
Intercept	4.197283122 *
FY18_Percent	0.954518503 ***
CBA_Q1==5	-0.051526
CBA_Q1==10	-0.087661 .
log(CICO_Q1)	-0.028831
HIV_Q1	-0.214162
TNG_Q1	-0.014324
CBA_Q2==5	0.142853 **
CBA_Q2==10	0.078023
DSP_Q2	-0.007237 *
CLIP_Q3	-0.204592 **
log(CICO_Q3)	-1.735521 *
TNG_Q3	-0.013613
DSP_Q3	-0.385254 *
log(CICO_Q4)	0.042757
HIV_Q4	0.307731 **
FBP_Q4	0.009931
log(CICO_Q1):HIV_Q1	0.081969
CLIP_Q3:DSP_Q3	0.021069 **
log(CICO_Q3):DSP_Q3	0.162224 *
log(CICO_Q4):HIV_Q4	-0.133737 **

Table 4: Logistic Regression Model Estimates

Predictor variable	β parameter estimate
Intercept	255.42637
Per_Q1	0.0.09613
CBA_Q2==5	18.85170
CBA_Q2==10	17.28383
Per_Q2	3.40096 .
CLIP_Q3	-9.00653 **
DSP_Q3	-28.94637 *
log(CICO_Q3)	-122.64871 *
Per_Q3	3.37885 .
log(CICO_Q4)	1.68667
HIV_Q4	12.89337 *
Per_Q4	7.71460 **
CLIP_Q3:DSP_Q3	0.92790 **
DSP_Q3:log(CICO_Q3)	12.23197 *
log(CICO_Q4):HIV_Q4	-5.48043 *

Signif. codes: 0 ' *' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Table 5: FY2020 Predictions

MEPS	FY2020 Prediction	MOY	MOQ1	MOQ2	MOQ3	MOQ4	Win Unit in 2020
Atlanta	0.9018909			Winner	Winner	Winner	Yes
Baltimore	0.8370317		Winner				
Chicago	0.82281						Yes
Columbus	0.8313931						Yes
Dallas	0.7660713						
Fort Jackson	0.9129501						Yes
Fort Lee	0.8723752						Yes
Houston	0.911784						
Jacksonville	0.915199		Winner	Winner	Winner	Winner	Yes
Los Angeles	0.832069						Yes
Miami	0.8542421						
Montgomery	0.8424609						Yes
New York	0.8198566						Yes
Phoenix	0.8294936						Yes
Raleigh	0.930949						
Sacramento	0.8703183				Winner		
San Antonio	0.7524866						Yes
San Diego	0.8031394						Yes
San Jose	0.8553434						Yes

St Louis	0.8031368					Winner	
Tampa	0.9321654	Winner	Winner	Winner			Yes
Boston	0.8716037						Yes
Charlotte	0.9091162						
Cleveland	0.7938987						
Denver	0.751855						Yes
Detroit	0.8231315						Yes
Fort Dix	0.9591863			Winner		Winner	Yes
Harrisburg	0.8579761				Winner		Yes
Indianapolis	0.8556916						
Kansas City	0.8680807						
Knoxville	0.9460346		Winner	Winner	Winner		Yes
Lansing	0.8098692						
Louisville	0.9271158		Winner			Winner	Yes
Milwaukee	0.6926585						Yes
Minneapolis	0.890524						Yes
Nashville	0.9039894		Winner	Winner			Yes
New Orleans	0.8440103						
Oklahoma City	0.5482789						
Pittsburgh	1.0061532					Winner	Yes
Portland OR	1.8769807	Winner			Winner		
Salt Lake City	0.8365917						
Seattle	0.9142306						
Springfield	0.8488509						Yes
Albany	0.8576016		Winner				Yes
Albuquerque	0.7970988						
Amarillo	0.8833072						Yes
Anchorage	0.7784369			Winner		Winner	Yes
Beckley	0.9676862	Winner		Winner			
Boise	0.7979164						
Buffalo	0.8328573						Yes
Butte	0.8493384						Yes
Des Moines	0.8775016						Yes
El Paso	0.9004641			Winner	Winner		Yes
Fargo	0.9250971						Yes
Honolulu	0.7982182						Yes
Jackson	0.9239914		Winner		Winner	Winner	Yes
Little Rock	0.7099517						
Memphis	0.894485				Winner	Winner	Yes
Omaha	0.8773086						
Portland ME	0.9530189		Winner				Yes

San Juan	0.712587						Yes
Shreveport	0.9139973						Yes
Sioux Falls	0.9082461						
Spokane	0.8972412						Yes
Syracuse	0.7839002						Yes

Table 6: FY2020 Quarterly UNIT Pennant Predictions

MEPS	Unit Q1	Unit Q2	Unit Q3	Unit Q4
Atlanta	Yes	Yes	Yes	Yes
Baltimore	Yes	Yes		Yes
Chicago		Yes		
Columbus	Yes			Yes
Dallas				
Fort Jackson		Yes		
Fort Lee				Yes
Houston				
Jacksonville	Yes	Yes	Yes	Yes
Los Angeles				
Miami		Yes		Yes
Montgomery		Yes		Yes
New York				Yes
Phoenix				
Raleigh		Yes		
Sacramento			Yes	
San Antonio				
San Diego		Yes		
San Jose		Yes		
St Louis				Yes
Tampa	Yes	Yes		
Boston	Yes			Yes
Charlotte				Yes
Cleveland				
Denver				
Detroit				Yes
Fort Dix		Yes		Yes
Harrisburg			Yes	
Indianapolis				
Kansas City	Yes			
Knoxville	Yes	Yes	Yes	Yes
Lansing				
Louisville	Yes	Yes	Yes	Yes
Milwaukee	Yes			

Minneapolis	Yes			Yes
Nashville	Yes	Yes	Yes	Yes
New Orleans				
Oklahoma City	Yes			Yes
Pittsburgh	Yes	Yes		Yes
Portland OR			Yes	
Salt Lake City				
Seattle				
Springfield				
Albany	Yes	Yes		Yes
Albuquerque				
Amarillo		Yes	Yes	
Anchorage		Yes		Yes
Beckley	Yes	Yes		Yes
Boise	Yes			Yes
Buffalo		Yes		Yes
Butte		Yes		
Des Moines	Yes			
El Paso		Yes	Yes	
Fargo				Yes
Honolulu				
Jackson	Yes	Yes	Yes	Yes
Little Rock				
Memphis	Yes		Yes	Yes
Omaha				
Portland ME	Yes	Yes		Yes
San Juan				
Shreveport				
Sioux Falls			Yes	Yes
Spokane	Yes	Yes		Yes
Syracuse	Yes			

Appendix C. Figures

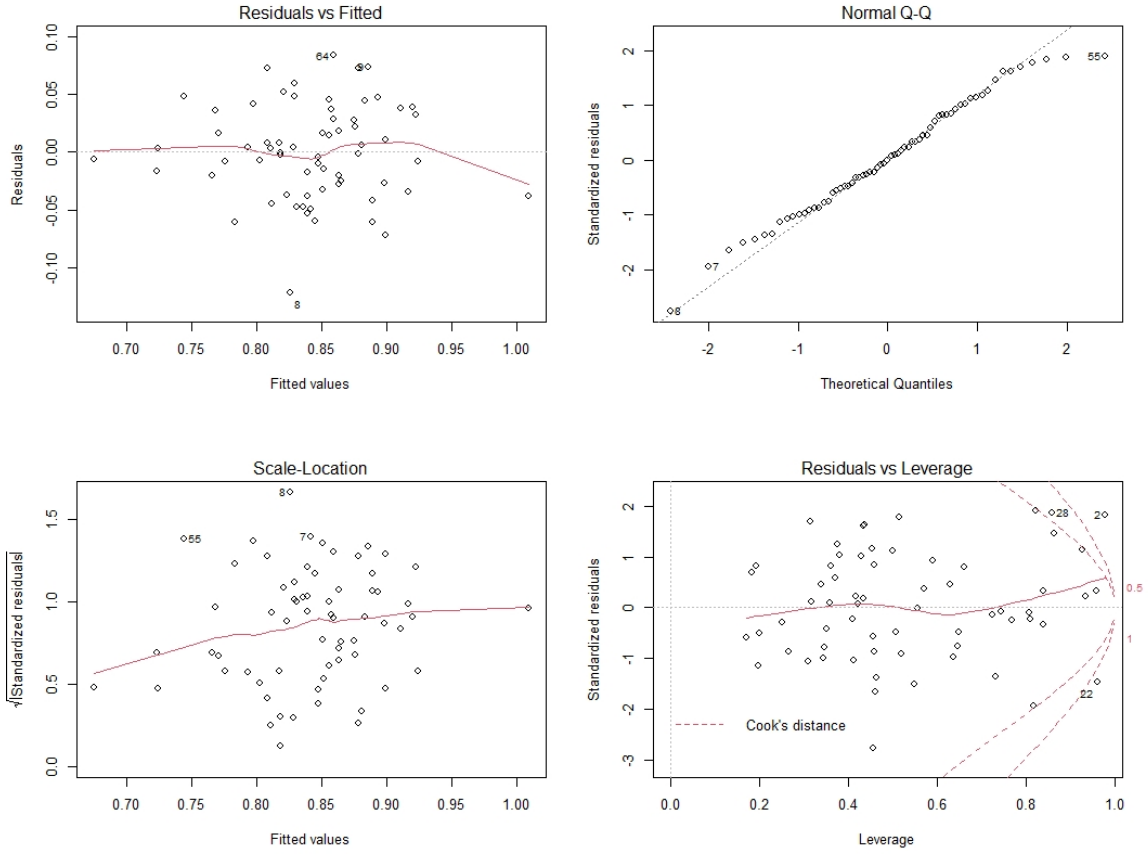

Figure 1. Diagnostic Plots for Final Model

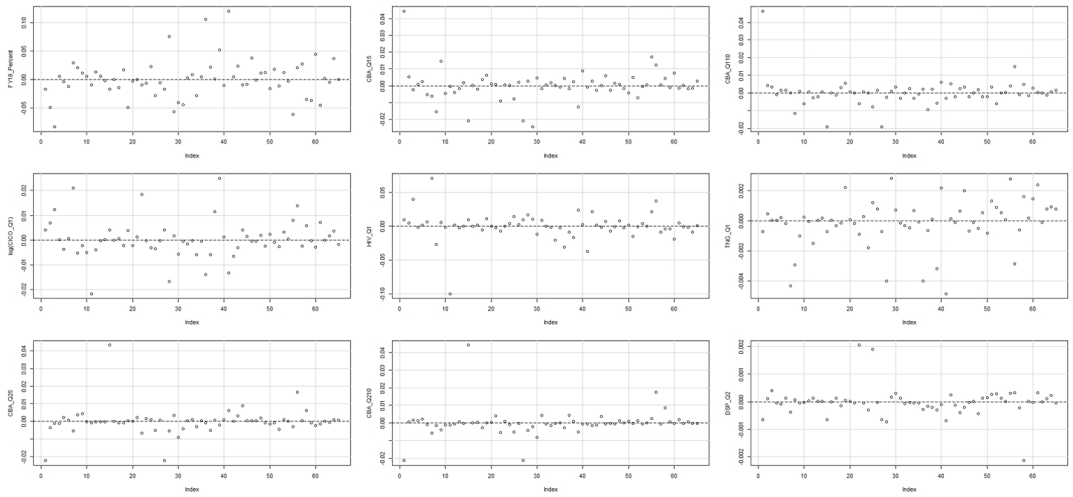

Figure 2: DFBETA Plot 1

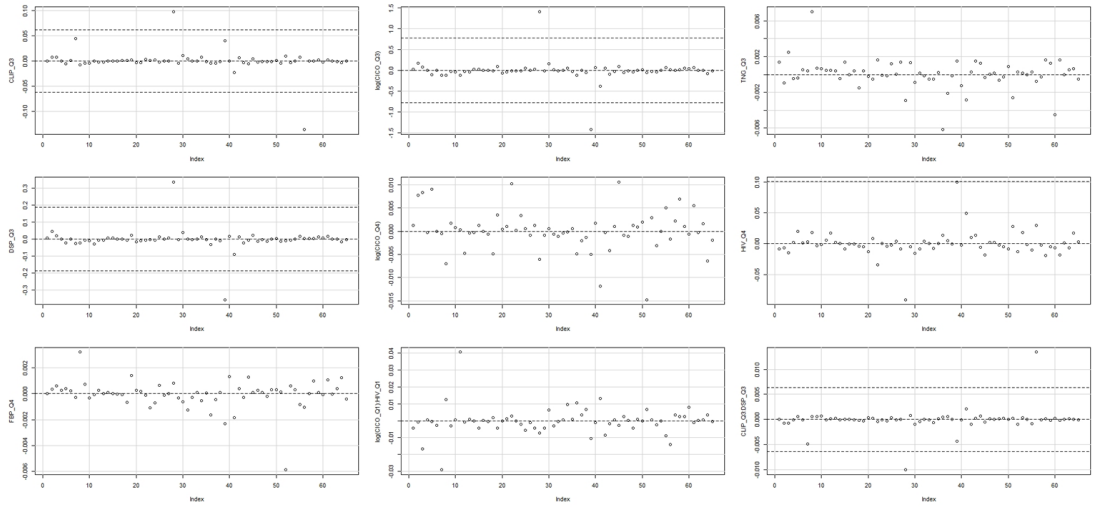

Figure 3: DFBETA Plot 2

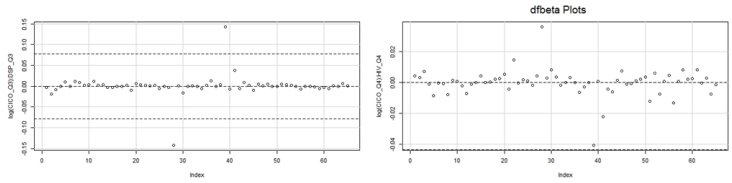

Figure 4: DFBETA Plot 3

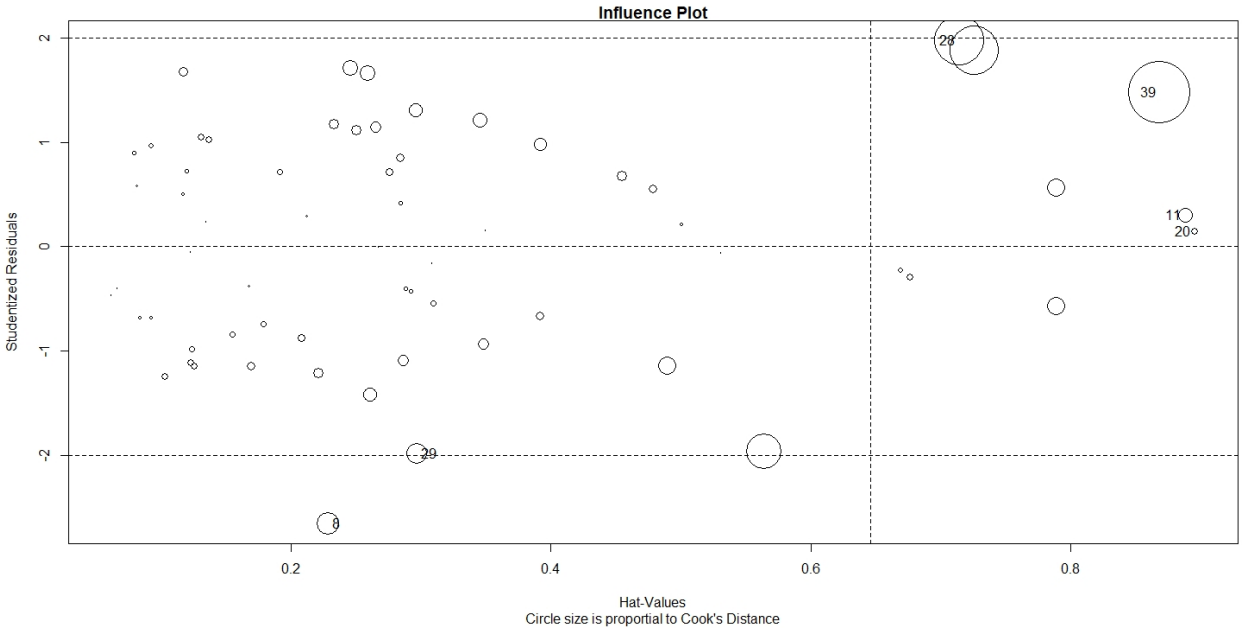

Figure 5: Influence Plot