ST-512

Final Project Report

Power Consumption Analysis

Oregon State University

Group 10

Presented by:
Praveen Ilango (ilangop@oregonstate.edu)
Fransiskus Derian (derianf@oregonstate.edu)
You-Jen Lin (linyou@oregonstate.edu)

1. Introduction

1.1. Project and Data Description

In this project, our group decided to conduct a data analysis approach on datasets related to power consumption of virtual machine activities. The input data was retrieved by recording some results of cloud computing experiments that was conducted by the Computer Science department at North Carolina State University. This experiments measures the power consumption (Response Variable) of a virtual machine based on their CPU, Memory, and Disk usages (Numerical Explanatory Variables) for different hadoop workload sizes (Categorical Explanatory Variables - two levels). As power consumption has become an alarming issue in the field of cloud computing, our group decided to answer several questions that is mentioned in the next section.

1.2. Questions of interest

In this project, our group is trying to answer the following questions:

1. Does the workload size have a significant effect on power consumption?
2. What is the effect of CPU, Memory and Disk usage on the host's power consumption?

2. Methodologies

2.1. Assumptions

In order to perform the analysis for this report, the following assumptions have to be met:

a. Non-Collinearity within explanatory variables (See Correlation Matrix in Appendix A)
b. Data Independence
c. Constant Variance (See Residual Plot in Appendix A)
d. Normality (See Normal Q-Q plot in Appendix A)

2.2. Approach

a. Ensuring that our assumptions have been met (Residuals Plot)
b. Do necessary transformation when needed (Boxcox plot)
c. Check and remove any influential data points (Case influential statistics plot)
d. Make a statistical inference (ESS F-statistics)

3. Summary

3.1. Challenges

This section describes several issues that our group encountered in the process of conducting data analysis and answering the questions of interest. The issues are mentioned below:

a. **Formatting Issue:** The input data that our group is using come with some formatting issues. Similar data type with different workloads (small, big) are represented as multiple different columns.
b. **Missing Values:** While checking the input data, we found some missing values in certain variables of the input data.

3.2. Solution

In order to solve the issues mentioned in the previous section, our group performed some additional works as follow:

a. **Formatting Issue:** To solve the formatting issue in the input data, we created two data frames containing the response variable and the explanatory variables which are differentiated based on the workload indicator variable (small, big). At the end, we merged these two dataframes to create one big data frame that is used as the primary data of the analysis.
b. **Missing Values:** To avoid potential issues with our analysis, we decided to remove the rows containing missing values in the input data.

3.3. Findings

Assumptions Checking

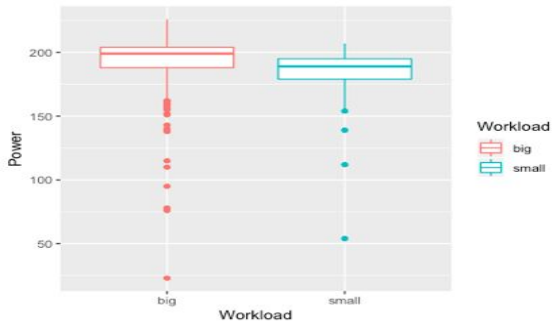
a. By looking at the Correlation Matrix, the Non-Collinearity assumption is satisfied (See the Correlation Matrix in Appendix B)
b. Data Independence is assumed to be satisfied
c. There seems to be a Non-Constant variance in the original datasets. Hence, we decided to use Cube of Power Consumption (as suggested in the Box-Cox Plot) as the response variable due to non-constant variance in the initial residual plot. (See the before transformation Box-Cox Plot in Appendix A)
d. According to the Normal Q-Q plot, there is no violation on the Normality assumption (See Normality Q-Q plot in Appendix A)

Case Influential Analysis

Case Influence statistics show that some data points have high leverage and high studentized residuals, however none of the data points are influential. (See the Case Influence Statistics Plot in Appendix B)

Q1. Does the workload size have a significant effect on power?

Box Plot



From the above boxplot, we can observe that the power consumption for a big workload is slightly higher when compared to the power consumption for a small workload. We fit the following model to test whether workload has a significant effect on power consumption.

Model:

Response variable : Power³

Explanatory variables : CPU, Memory, Disk, Workloadsmall

Base Workload Category: Workloadbig

$\mu(\text{Power}^3 \mid \text{CPU, Memory, Disk, Workloadsmall}) = \beta_0 + \beta_1 CPU + \beta_2 Memory + \beta_3 Disk + \beta_4 Workloadsmall$

Null Hypothesis : $\beta_4 = 0$ (Workload does not have an effect on Cube of Power Consumption)
Alternative Hypothesis: $\beta_4 != 0$ (Workload has a significant effect on Cube of Power Consumption)

Summary:

Coefficients	Estimate	Std.Error	T-statistic	P-value
Intercept	-3,445,050	670,520	-5.138	0.00000049
CPU	117,849	6,121	19.254	<2e-16
Memory	-2,634	5,089	-0.518	0.605
Disk	-43,399	9,884	-4.391	0.0000155
WorkloadSmall	-885,107	189,760	-4.664	0.00000459

Conclusion: We reject the null hypothesis that workload does not have an effect on Cube of Power Consumption at significance level alpha = 0.05. Since our p-value = < 0.05, we have strong evidence that workload has a significant effect on Cube of Power Consumption.

Effect of Workload on Power Consumption

With 95% confidence the mean Cube of Power Consumption spent by Small Workload is between 1,258,473.81 and 511,739.469 Watts less than the mean Cube of Power Consumption spent by Big Workload with other explanatory variables remain the same (See the table in Appendix C).

Q2. What is the effect of CPU, Memory and Disk usage on the host's power consumption?

Effect of CPU Usage on Power Consumption

With 95% confidence the mean Cube of Power Consumption spent is between 105,805.75 and 129,892.271 Watts more for 1 unit increase in the CPU usage with other explanatory variables remain the same (See the table in Appendix C).

Effect of Memory Usage on Power Consumption

With 95% confidence the mean Cube of Power Consumption spent is between 12,646.25 Watts less and 7,378.613 Watts more for 1 unit increase in the CPU usage with other explanatory variables remain the same (See the table in Appendix C).

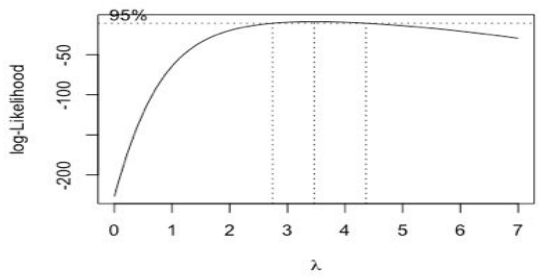
Effect of Disk Usage on Power Consumption

With 95% confidence the mean Cube of Power Consumption spent is between 62,845.40 and 23,951.707 Watts less for 1 unit increase in the CPU usage with other explanatory variables remain the same (See the table in Appendix C).

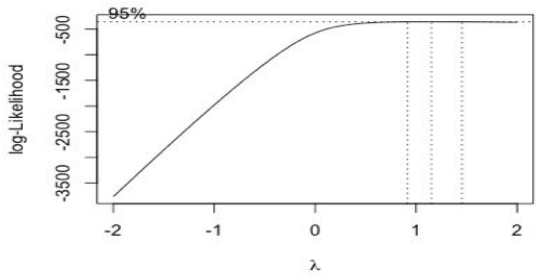
Appendix A

Boxcox Plot before and after transformation

Before Transformation

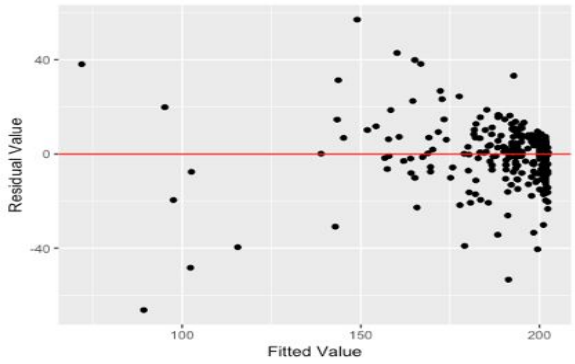


After Transformation

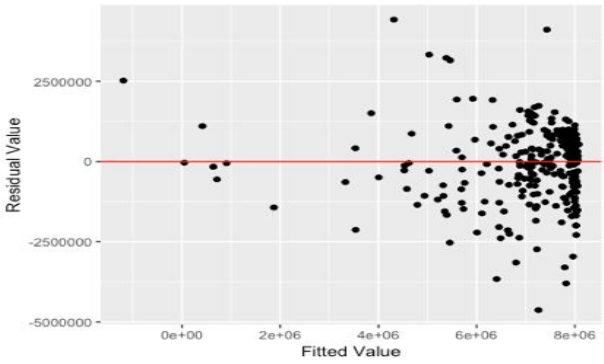


Residuals Plot before and after transformation

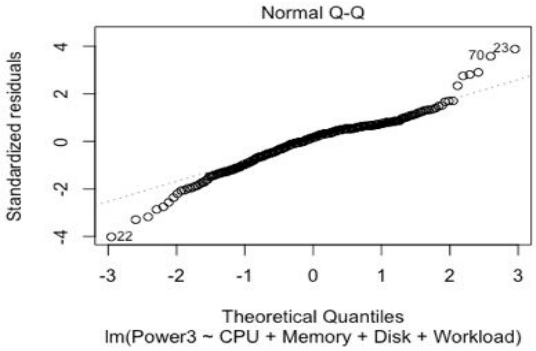
Before Transformation



After Transformation

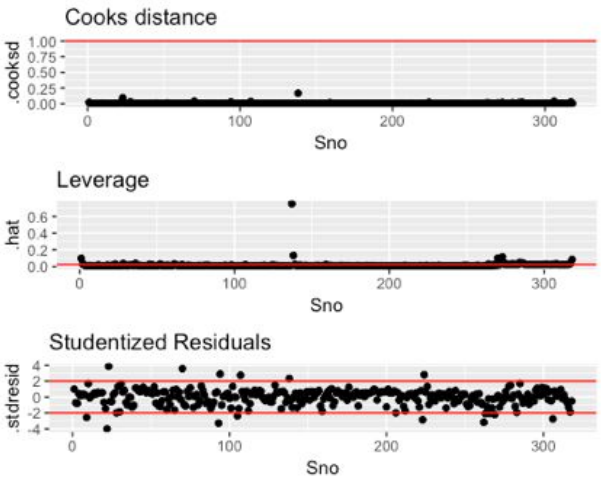


Normal QQ-Plot

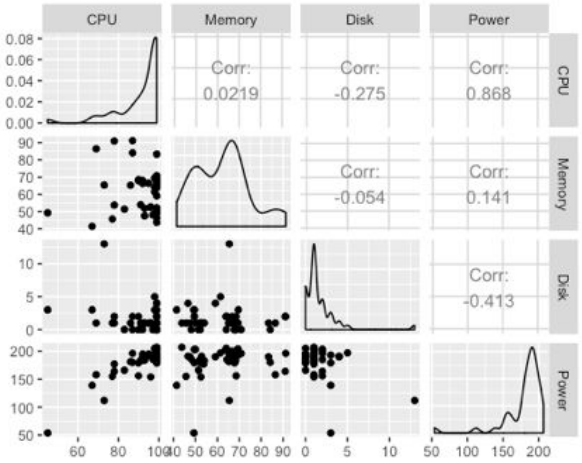


Appendix B

Case Influence Statistics



Correlation Matrix



Summary Output

```
##
## Call:
## lm(formula = Power3 ~ CPU + Memory + Disk + Workload, data = Data_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4630748  -629803   187600   701862  4425476
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3445050     670520  -5.138 4.90e-07 ***
## CPU             117849       6121  19.254  < 2e-16 ***
## Memory           -2634       5089  -0.518    0.605
## Disk             -43399       9884  -4.391 1.55e-05 ***
## Workloadsmall  -885107     189760  -4.664 4.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1158000 on 313 degrees of freedom
## Multiple R-squared:  0.5903, Adjusted R-squared:  0.5851
## F-statistic: 112.7 on 4 and 313 DF,  p-value: < 2.2e-16
```

Appendix C

Coefficient	Confidence Interval	
	Lower Bound	**Upper Bound**
Intercept	-4,764,345.90	-2,125,754.054
CPU	105,805.75	129,892.271
Memory	-12,646.25	7,378.613
Disk	-62,845.40	-23,951.707
WorkloadSmall	-1,258,473.81	-511,739.469