Final Report

By Vipul Satone and Angadvir Paintal

OBJECTIVE

The aim of this project is to design an experiment to find out which file compression software (7zip, WinRAR and PeaZip) is fastest.

INTRODUCTION

Software like WinRAR, 7zip and PeaZip (henceforth referred to as zip softwares) are used to archive files. They compress large files into one ZIP/RAR file. These ZIP/RAR files are designed to compress data to save memory space make it easier to transfer, otherwise large files, through both online and offline platforms.

There are numerous free or paid versions of such software available. (Detailed list can be found in the references [1]).

Most of the users care little about the speed of file conversion and more about compression ratio. However, when file sizes exceed a certain amount (e.g. >1 GB), the time to compress becomes a more significant factor. Compression time becomes critical especially when large sets of data is needed to be archived and then uncompressed for database recovery and subsequent analysis.

Majority of the focus is weighed on the compression ratio for most compressive software applications, available online. Most of these softwares claim to provide high compressive speed along with a desirable compression ratio. However, not much analysis has been performed to quantify the "**fast compressive speed**" of these free and commercially available software, especially under certain file conditions. Subsequently, there are no conclusive results which help in making a decision for compression speed, while trying to make a decision of choosing one compression software over another. In light of this justification, we propose to design an experiment to analyze and identify the best software in terms of time taken to compress files varying sizes, types and numbers.

Design of the experiment

Various factors might affect the compression time. Following four factors are included in the design a they have a major effect on the speed of the compression:

1) Operating system (2 levels) – Windows 10 and Linux (Ubuntu 14.04 LTS).

2) Content of files to be compressed (3 levels) – 500 MB; 1.5 GB; 3 GB.

3) Number of files compressed in one instance (3 levels) – 1 file, 250 files, 500 files.

4) Zip software (3 levels) – 7zip, WinRar and PeaZip

The aim of this experiment is to find out fastest compression software for a range of file sizes and number of files compressed in a single instance. Therefore, these two factors along with compression software are classified as treatment variables. Windows and Linux Operating systems work differently. Hence, it was suspected that they might affect the compression

software speed. In this regard, operating system is considered as a block variable. The interaction of the operating system with other treatment effects is not taken into account since we assume that a user would either use one of the 2 OS at a time, while preforming any compression and un-compression tasks with files of various sizes, numbers and types.

Four factors, with one block and three treatments, are considered. Each treatment can be performed on each block level and number of levels are different in block factor and treatments. Considering all the above factors chosen for the experiment, a CRBD- Complete Randomized Block Design was selected as the design model. The response variable is taken as the time required to compress the various files. The unit of the response variable is seconds.

CRBD Experiment

Block factor has two levels and each level has 27 observations. We took 2 replications, so we had total 108 observations. Experiments were randomized for better results. Sequence of experiments was determined by sampling without replacement. All the possible combinations of block and treatment were written on paper chits and one chit was drawn from the bundle without replacement to determine the order in which experiments will be performed.

Most importance was given to keep all experimental conditions same throughout the different experiment. Therefore, laptop with partitioning between Linux and Windows was used as the hardware used in both the blocks levels will be same. All the experiments were performed on 6gb ram 250gb hard disk dell inspiron and i5 processor Ubuntu 14.04 windows 10. Furthermore, while running the experiments all other applications on the laptop were terminated.

Both Linux (Ubuntu) and Windows 10 record the time when the compression process begins (file created) and when it stops (file modified). This information is available in file properties (refer figure 1). By using this information, the human error in recording the time was alleviated or completely removed.

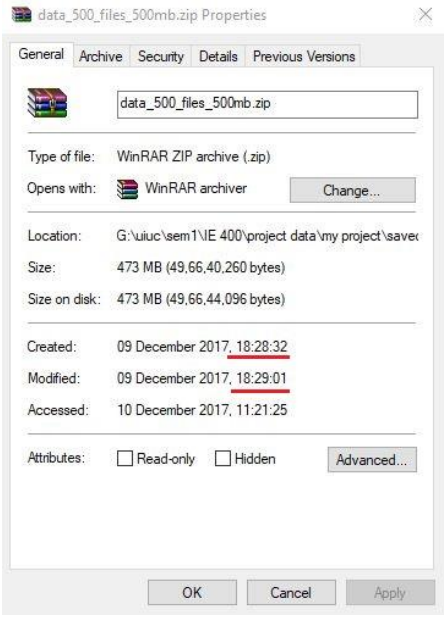

Figure 1

Observations

In the following table (Table 2)consist of first 30 observation of our experiment (Find the complete table in the appendix).

Abbreviation	Factor	Level '-1'	Level '0'	Level '+1'
Bl	Block - OS	Linux	-	Windows 10
Si	Size of files	0.5gb	1.5gb	3gb
So	Zip software	7zip	WinRAR	PeaZip
Nu	Number of files	1	250	50

Table 1: Explanation of treatment and block effects.

Obs	Bl	So	Si	Nu	Rep1	Rep2
1	-1	-1	-1	-1	110	107
2	-1	-1	-1	0	113	112
3	-1	-1	-1	1	58	60
4	-1	-1	0	-1	231	226
5	-1	-1	0	0	301	306
6	-1	-1	0	1	199	190
7	-1	-1	1	-1	878	858
8	-1	-1	1	0	694	688
9	-1	-1	1	1	676	671
10	-1	0	-1	-1	59	58
11	-1	0	-1	0	100	103
12	-1	0	-1	1	78	80
13	-1	0	0	-1	127	125
14	-1	0	0	0	235	237
15	-1	0	0	1	157	159
16	-1	0	1	-1	468	473
17	-1	0	1	0	479	476
18	-1	0	1	1	387	381
19	-1	1	-1	-1	69	71
20	-1	1	-1	0	92	90
21	-1	1	-1	1	38	39
22	-1	1	0	-1	180	186
23	-1	1	0	0	192	188
24	-1	1	0	1	124	127
25	-1	1	1	-1	496	493
26	-1	1	1	0	340	344
27	-1	1	1	1	327	322
28	1	-1	-1	-1	83	85
29	1	-1	-1	0	74	73
30	1	-1	-1	1	32	31

Table 2: A snippet of the 1st 30 runs of the complete experimental results

SAS output and Data Analysis:

Part-1: Model Selection

Since a CRBD ANOVA model was utilised for the design and analysis of the experiment, the following Factor effects model was utilized to perform statistical analysis of the data sets:

$$\boldsymbol{y_{ijklm} = \mu \ldots . + \rho_i + \tau_j + \gamma_k + \kappa_l + \epsilon_{ijklm}}$$

$$i = 1\ to\ 3\ , \quad j = 1\ to\ 3, \quad k = 1\ to\ 3, \quad l = 1\ to\ 2, \quad m = 1\ to\ 2$$

$y_{ijklm} = Time\ required\ to\ compress\ the\ file\ .(response\ )$

$such\ that\ y_{ijklm} are\ iid\ with\ constant\ variance$

$\epsilon_{ijklm} = Error\ factor\ for\ l^{th} block\ and\ i,j,k\ treatment\ and\ m^{th} repetition$

$such\ that\ \epsilon is\ normally\ distributed\ with\ constant\ variance\ \left(\epsilon \sim N(0,\sigma^2)\right).$

$\mu \ldots . = Constant\ representing\ overall\ mean$

$\rho_i = Constant\ representing\ treatment\ effect\ (\ file\ size)\ \ such\ that\ \sum \rho_i = 0$

$\tau_j = Constant\ representing\ treatment\ effect\ (number\ of\ files)\ \ such\ that \sum \tau_j = 0$

$\gamma_k\ = Constant\ representing\ treatment\ effect\ (zip\ software)\ \ such\ that \sum \tau_j = 0$

$\kappa_l = Constant\ representing\ block\ effect\ (operating\ system)\ \ such\ that \sum \kappa_l = 0$

The full model ANOVA was tested and the table shown below was obtained:

The ANOVA Procedure
Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	3586358.861	132828.106	29.69	<.0001
Error	80	357902.130	4473.777		
Corrected Total	107	3944260.991			

R-Square	Coeff Var	Root MSE	y Mean
0.909260	29.01675	66.88630	230.5093

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Nu	2	105805.352	52902.676	11.83	<.0001
Si	2	2598210.907	1299105.454	290.38	<.0001
Nu*Si	4	91312.815	22828.204	5.10	0.0010
So	2	381694.574	190847.287	42.66	<.0001
Nu*So	4	4424.815	1106.204	0.25	0.9105
Si*So	4	261807.593	65451.898	14.63	<.0001
Nu*Si*So	8	3925.685	490.711	0.11	0.9988
Bl	1	139177.120	139177.120	31.11	<.0001

Figure 1 – Anova Table

As seen from the above model(figure 1), the interaction between the number of flies archived or compressed, the compression software utilised and size of files is found to be insignificant for alpha equal to 0.05 significance level.

The diagnostics plot obtained from the above model is found to be as follows:

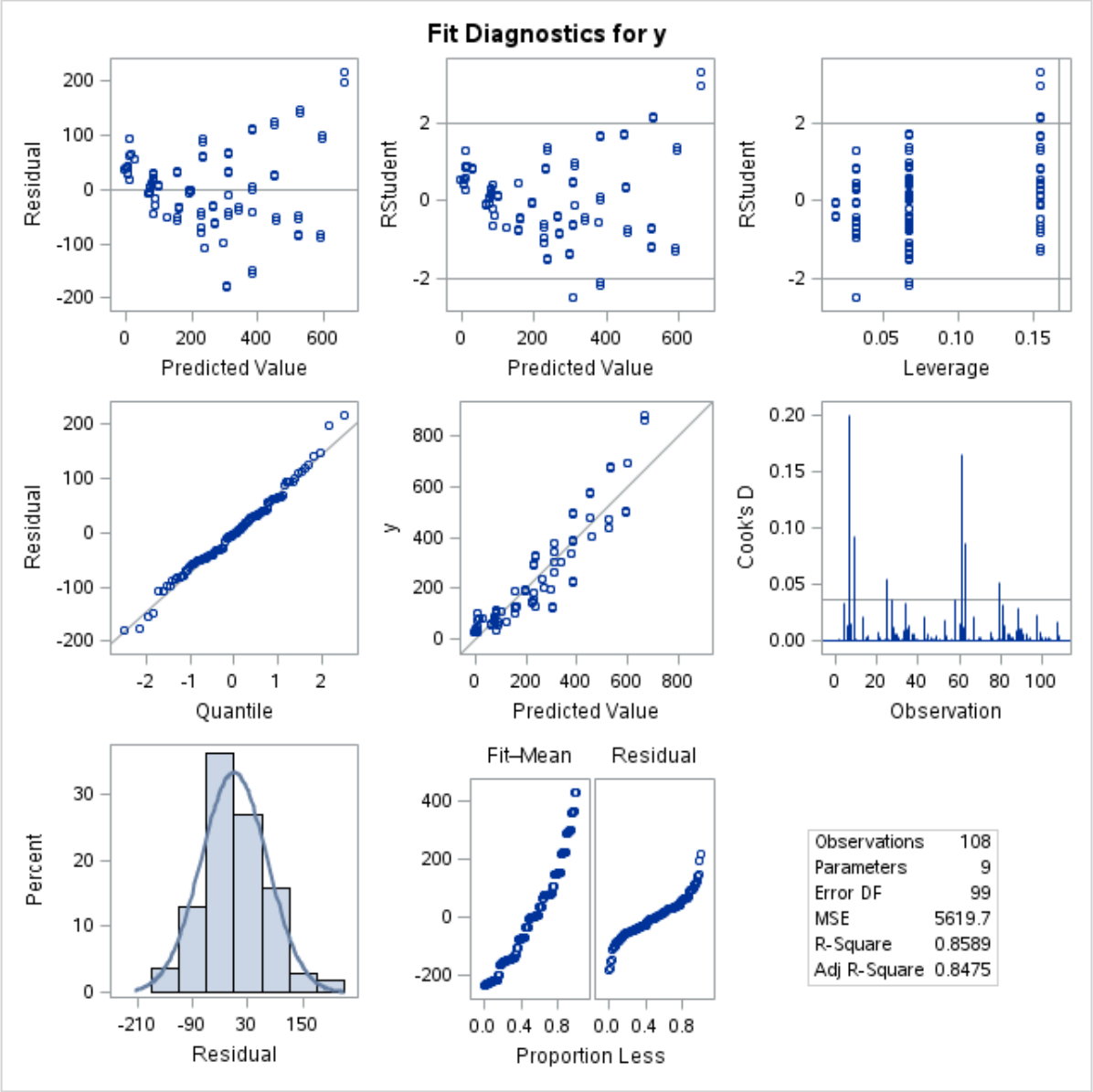

Figure 2 – Diagnostic plots of first model

The residual v.s. predicted (first graph in first row) values graph shows non-constant error variance, with an increase in the residuals, as the value of the predicted variable increases. This non-constant error variance needs to be address as it violates the ANOVA assumption in the model. However, from it can be concluded that normality and independence assumptions hold true.

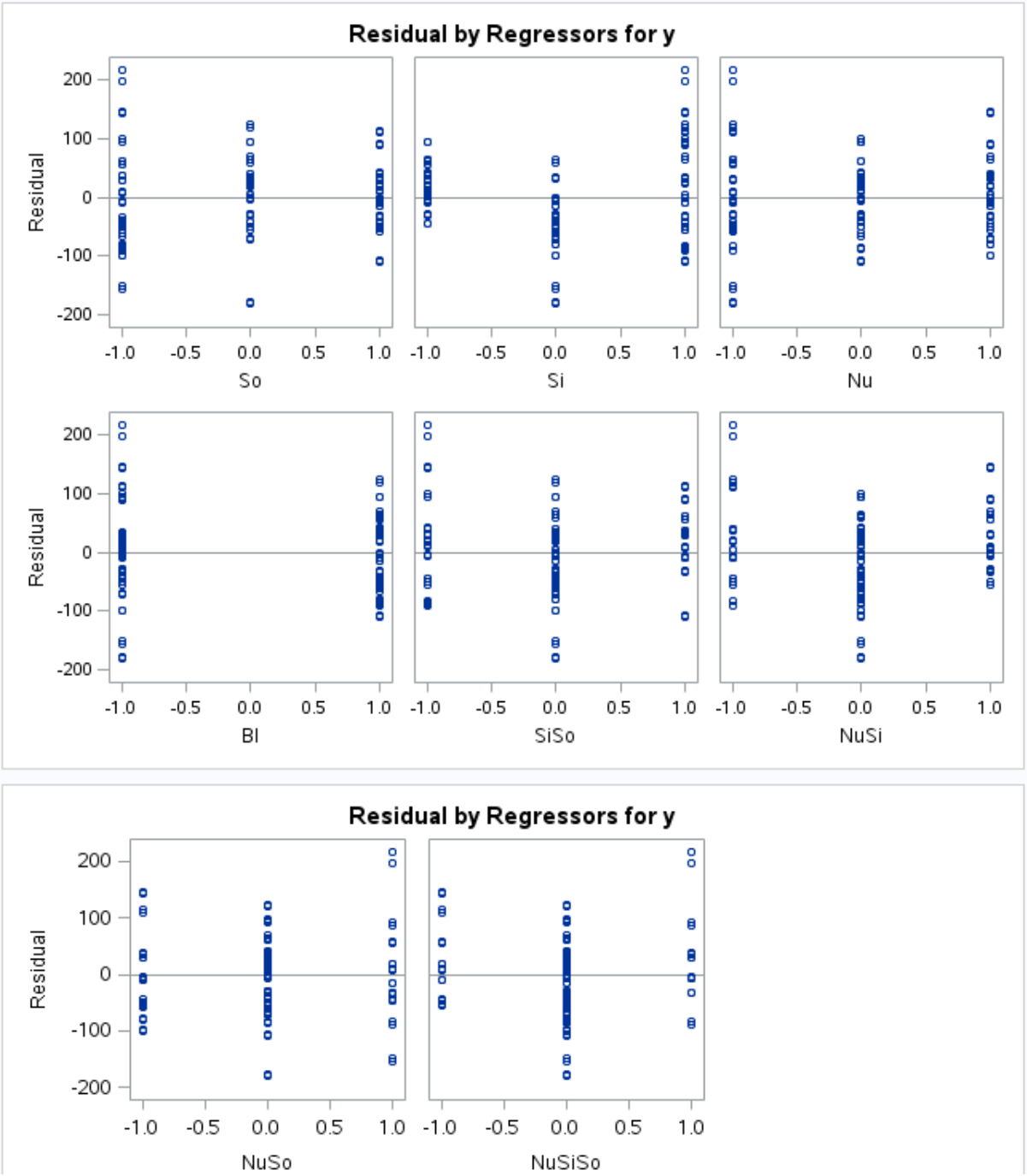

Figure 3 – Diagnostic plots of first model continued.

A similar result of non-constant error variance is seen in the residuals for the 3 different compression software as well as the residuals for Operating System or the block variable. The residuals for compression time for UBUNTU is much larger than that of the Windows software.

In order to address these ANOVA assumption violations, the response variable was transformed using the box-cox transformation. $Y \rightarrow y^{0.42}$.

The ANOVA model was re-run with the transformed response variable and the following ANOVA tables wert obtained.

The ANOVA Procedure
Dependent Variable: trans_y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	1070.378576	39.643651	44.83	<.0001
Error	80	70.740142	0.884252		
Corrected Total	107	1141.118719			

R-Square	Coeff Var	Root MSE	trans_y Mean
0.938008	10.39033	0.940347	9.050209

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Nu	2	34.0233523	17.0116761	19.24	<.0001
Si	2	833.5751394	416.7875697	471.34	<.0001
Nu*Si	4	17.6445640	4.4111410	4.99	0.0012
So	2	109.0811161	54.5405581	61.68	<.0001
Nu*So	4	2.9220758	0.7305190	0.83	0.5124
Si*So	4	28.6183066	7.1545767	8.09	<.0001
Nu*Si*So	8	4.3091974	0.5386497	0.61	0.7676
Bl	1	40.2048248	40.2048248	45.47	<.0001

Figure 4 –Anova model with transformed response.

After transforming the response variable to $y^{0.42}$, the interaction between the number of flies archived or compressed, the compression software utilised and size of files is still found to be insignificant for alpha equal to a 0.05 significance level. The interaction between the compression software and the number of files archived, is also found to be insignificant for a 0.05 significance level. The rest of the two way interaction and the main effects are found to be significant in the ANOVA model.

The diagnostics residual plots for the new model, with the transformed response variable is provided below:

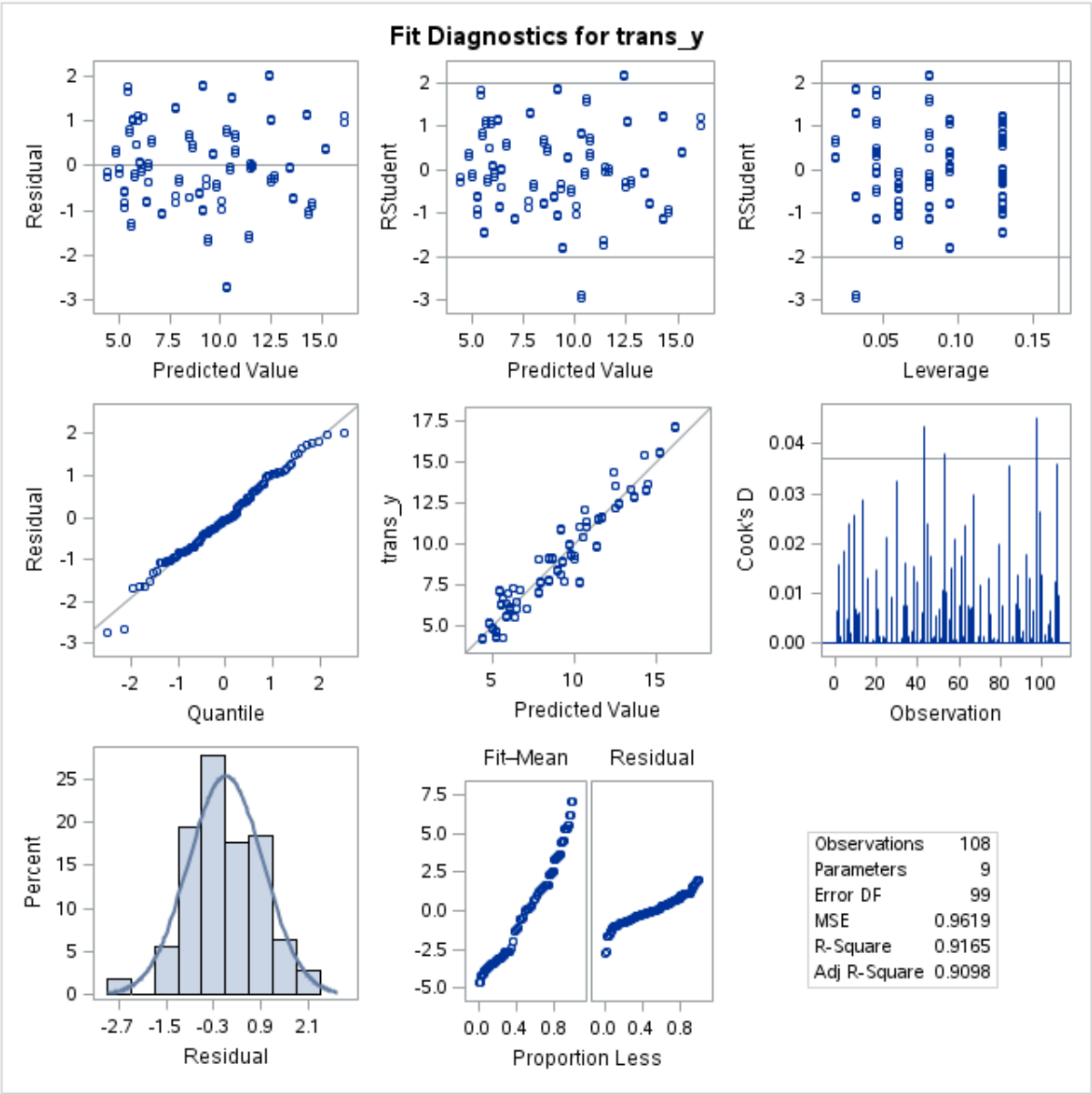

Figure 5 – Diagnostic plots for Anova model with transformed response.

The residual vs predicted values graph (see. first graph in first row) now shows constant error variance with no distinct pattern observed. The residuals follow a normal distribution, as seen from the histogram and the QQ plot (see first graph of second and third row), which supports the ANOVA assumption of the residuals for the response variable in the model to follow IID. The Cook's Distance is also reduced from a maximum value of 0.20 to approximately 0.04. The adjusted R^2 value is increased from 84.75% to 90.98%.

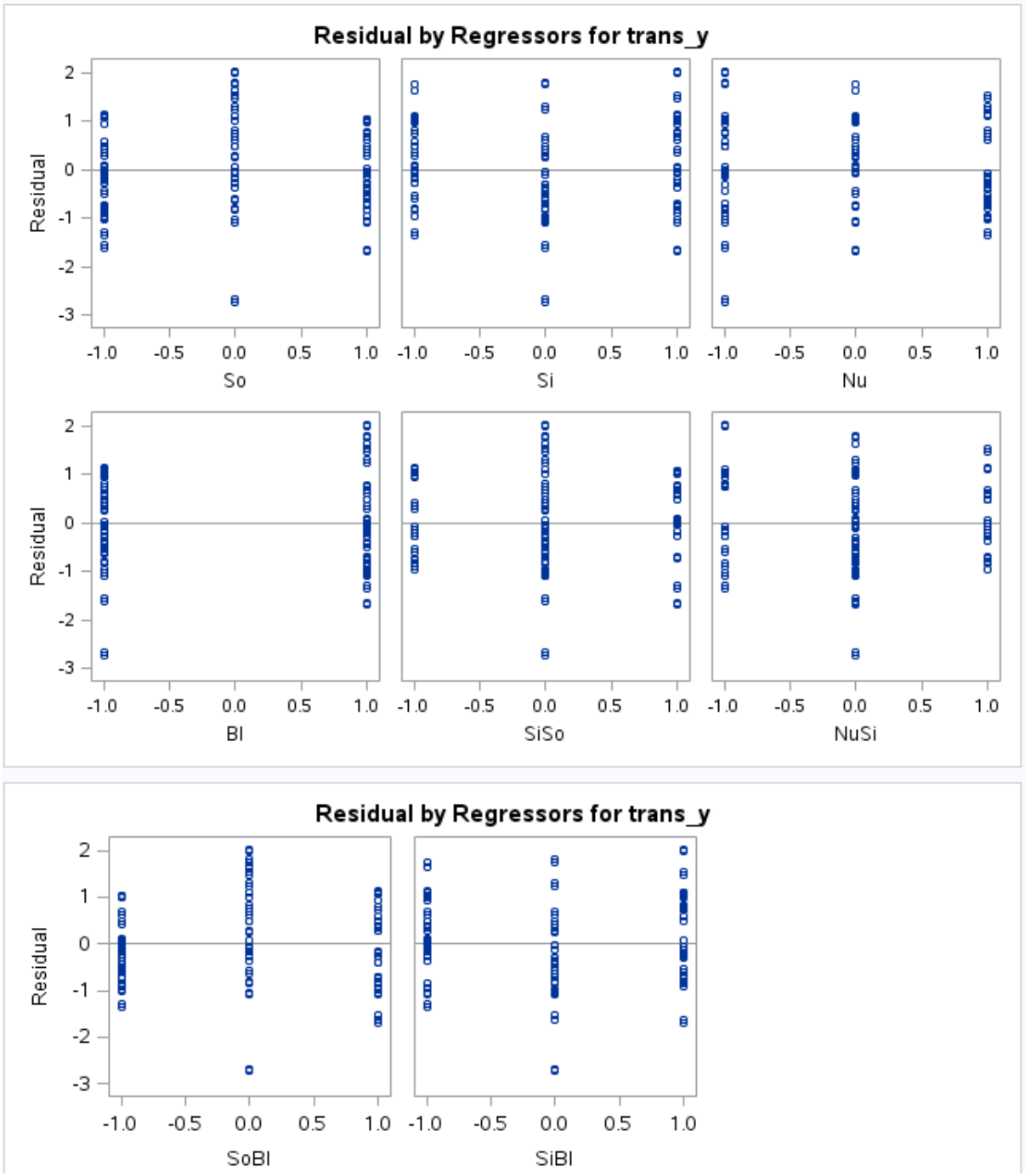

Figure 6 – Diagnostic plots for Anova model with transformed response continued.

The non-constant error variance which was seen in the residuals for the 3-different compression software as well as the residuals for Operating System or the block variable, is now resolved. The residuals for compression time for UBUNTU is much larger than that of the Windows software. Difference between residuals for different softwares was around 80 before transformation(see first graph in figure 3) now after transformation, the difference between residuals for different softwares is around 1 (see graph 1 in figure 6). Similarly, the difference in residuals between different softwares is reduced from 60 to 1.

After implementing the **backward elimination technique for the GLM** for the experiment, the significant terms left in the transformed model is as follows:

The GLM Procedure
Dependent Variable: trans_y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	1063.147303	70.876487	83.63	<.0001
Error	92	77.971415	0.847515		
Corrected Total	107	1141.118719			

R-Square	Coeff Var	Root MSE	trans_y Mean
0.931671	10.17221	0.920606	9.050209

Source	DF	Type I SS	Mean Square	F Value	Pr > F
So	2	109.0811161	54.5405581	64.35	<.0001
Si	2	833.5751394	416.7875697	491.78	<.0001
Si*So	4	28.6183066	7.1545767	8.44	<.0001
Nu	2	34.0233523	17.0116761	20.07	<.0001
Nu*Si	4	17.6445640	4.4111410	5.20	0.0008
BI	1	40.2048248	40.2048248	47.44	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
So	2	109.0811161	54.5405581	64.35	<.0001
Si	2	833.5751394	416.7875697	491.78	<.0001
Si*So	4	28.6183066	7.1545767	8.44	<.0001
Nu	2	34.0233523	17.0116761	20.07	<.0001
Nu*Si	4	17.6445640	4.4111410	5.20	0.0008
BI	1	40.2048248	40.2048248	47.44	<.0001

Figure 7 – Final Anova Model

The R^2 value for the fitted GLM model is 0.93167, this implies approximately 93.17% of the variation in the compression time of the file(s) is explained by the predictor variables chosen in this model.

The significant factors which statistically influence the compression time of the file(s) on the 2 operating systems, namely UBUNTU 14.04 and Windows 10, are the size of the file, the number of files compressed in one instance, for a particular total file size and the compression software used.

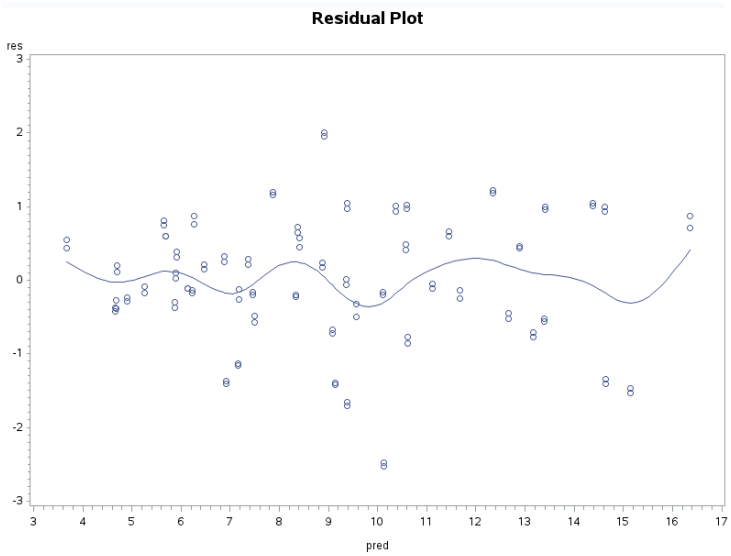

Figure 8 – Residual plot for Final Anova Model

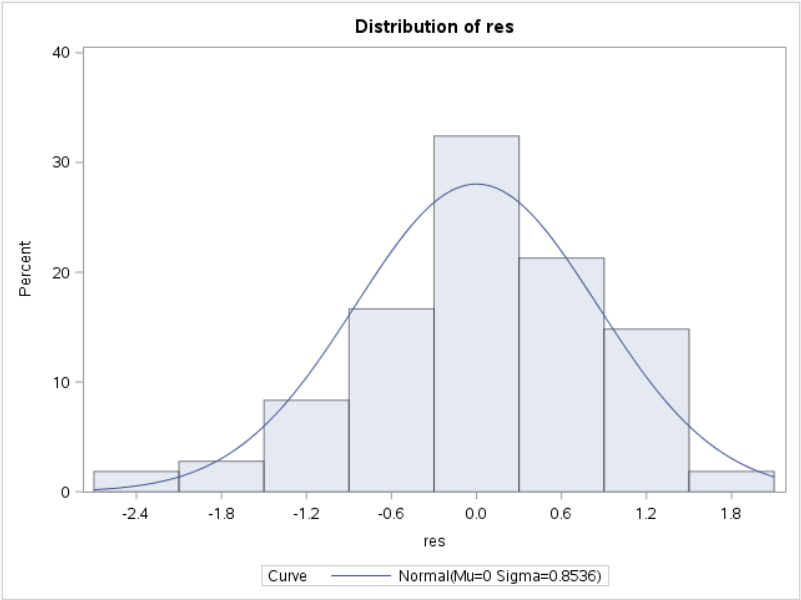

Figure 9 – Histogram for Final Anova Model

The graph above shows the residuals vs predicted values of the finalized GLM structure. As seen clearly, there's no distinct pattern observed in the residuals plot and the variance of the error from the 0 value is mostly constant. The residuals for the finalized GLM structure are found to be normally distributed, thereby supporting the ANOVA assumption of IID residuals.

Part-2: Inferences

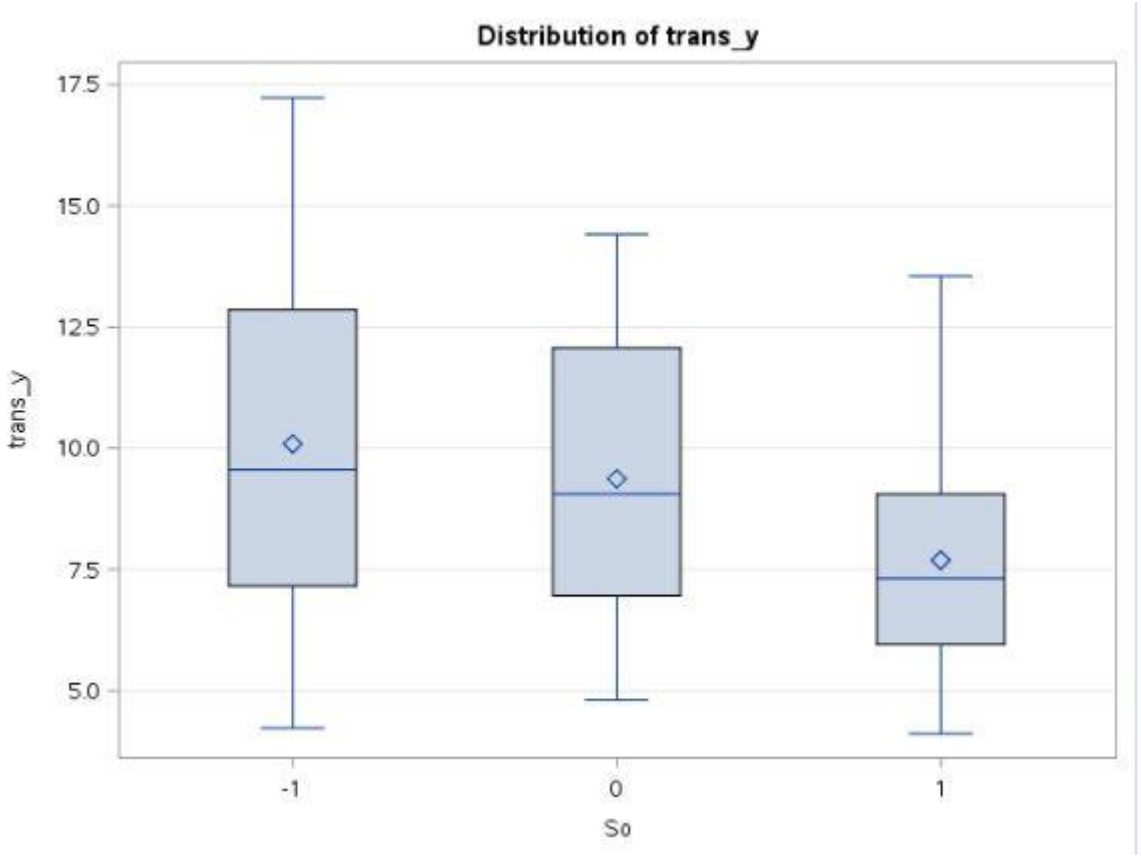

Figure 10 - Factor effects graph for the compression software.

The GLM Procedure
Tukey's Studentized Range (HSD) Test for trans_y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	92
Error Mean Square	0.847515
Critical Value of Studentized Range	3.36899
Minimum Significant Difference	0.5169

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	So
A	10.0897	36	-1
B	9.3698	36	0
C	7.6911	36	1

Figure 11 – Tuckey grouping table for the compression software.

From the Tukey groupings test and a significance level of 0.05, the 3 compression software have statistically significant difference in compression time for the same file sizes and

number of files. After un-transforming the response variable, mean compression times for the 3 software, 7zip, WinRAR and PeaZip, are found to be, **245.576 sec**, **205.894 sec** and **128.675 seconds** respectively.

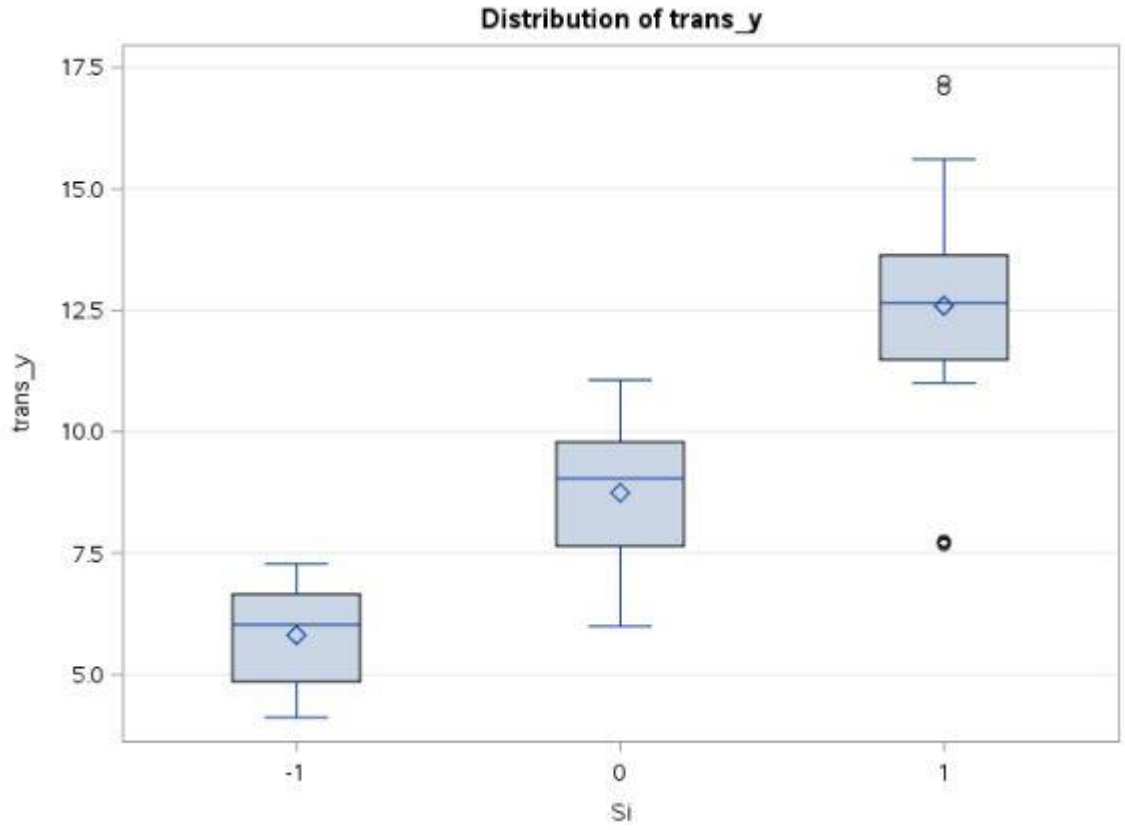

Figure 12 - Factor effects graph for the file size.

The GLM Procedure
Tukey's Studentized Range (HSD) Test for trans_y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	92
Error Mean Square	0.847515
Critical Value of Studentized Range	3.36899
Minimum Significant Difference	0.5169

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Si
A	12.5960	36	1
B	8.7429	36	0
C	5.8117	36	-1

Figure 13 – Tuckey grouping table for the file size.

From the Tukey groupings test and a significance level of 0.05, the 3 file sizes **(0.50 GB, 1.50 GB, and 3.00 GB)** have statistically significant difference in compression time for the same compression software, file types and number of files. After un-transforming the response variable, mean compression times for the 3 files sizes, 0.50GB, 1.50GB and 3.00GB, are found to be, **66.034 sec**, **174.597 sec** and **416.487 seconds** respectively.

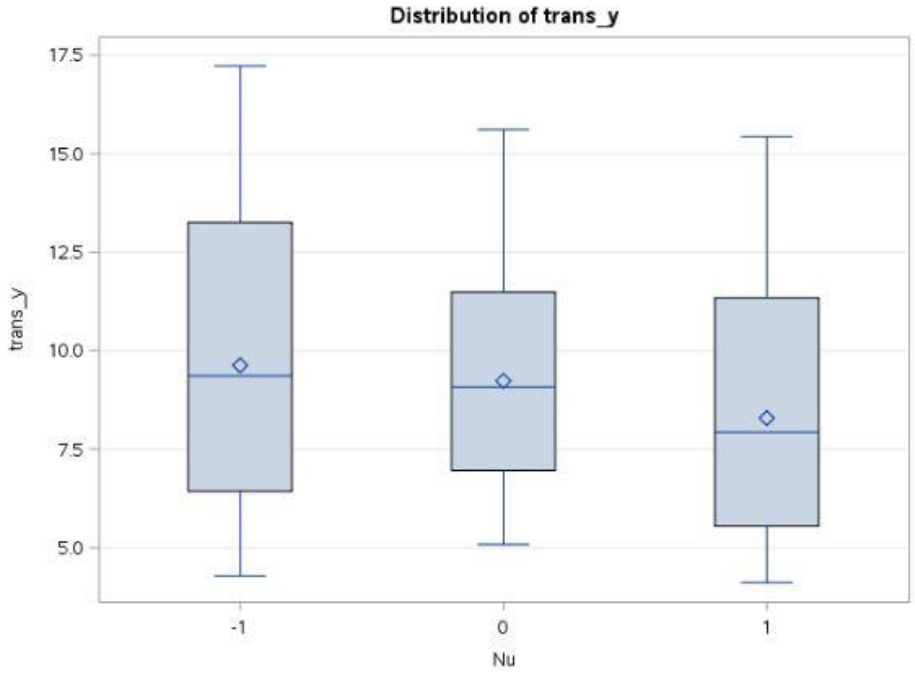

Figure 14 - Factor effects graph for number of files compressed at a time.

The GLM Procedure
Tukey's Studentized Range (HSD) Test for trans_y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	92
Error Mean Square	0.847515
Critical Value of Studentized Range	3.36899
Minimum Significant Difference	0.5169

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Nu
A	9.6286	36	-1
A			
A	9.2318	36	0
B	8.2902	36	1

Figure 15 – Tuckey grouping table for number of files compressed at a time.

From the Tukey groupings test and a significance level of 0.05, the time taken to compress files of various numbers **(single file, 250 files, and 500 files)** have statistically significant

difference in compression time for the same compression software and size of files. After un-transforming the response variable, mean compression times for folders of various file sizes, 1 file, 250 files and 500 files, are found to be, **219.693 sec**, **198.747 sec** and **153.837 seconds** respectively. According to the Tukey groupings test, there's no significant difference in compression times for compressing a single file or 250 files. However, the compression time for 500 files is found to be statistically smaller than compressing 250 files or smaller. This could imply that the software is otpimized for compressing larger file numbers than smaller file numbers.

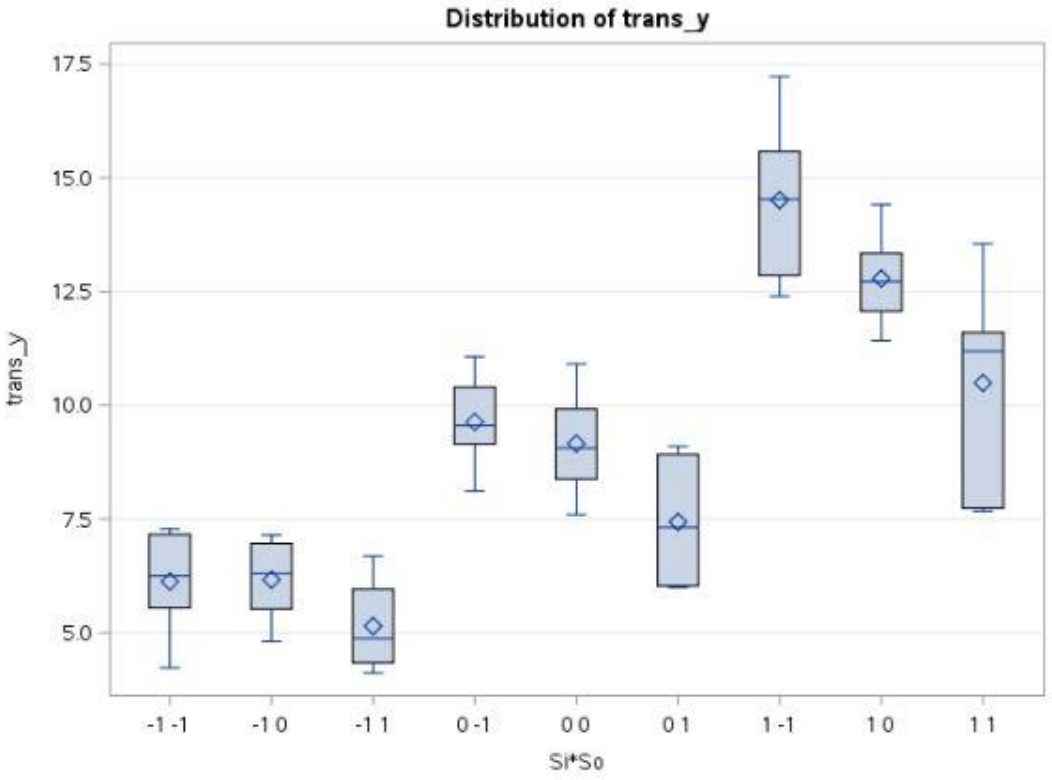

Figure 16 - Interactions effects graph for files sizes and compression software.

Level of Si	Level of So	N	trans_y	
			Mean	Std Dev
-1	-1	12	6.1231468	1.07036011
-1	0	12	6.1654844	0.81559387
-1	1	12	5.1466095	0.93277949
0	-1	12	9.6354087	0.96965722
0	0	12	9.1578285	1.09210662
0	1	12	7.4353457	1.28563339
1	-1	12	14.5106440	1.74475424
1	0	12	12.7861643	1.01825341
1	1	12	10.4912472	2.21193209

Figure 17 – Means Table for software and size interaction on transformed data.

Level of Si	Level of So	N	y
			Mean
-1	-1	12	74.772399
-1	0	12	76.0081458
-1	1	12	49.4420225
0	-1	12	220.062995
0	0	12	194.975005
0	1	12	118.719
1	-1	12	583.334347
1	0	12	431.609213
1	1	12	269.485939

Figure 18 – Means Table for software and size interaction on original un-transformed data.

From the Tukey groupings test and a significance level of 0.05, there is an interaction effect found between the type of compression software used and the size of the files compressed. Firstly, there's a significant difference found in the compression times between the 3 compression software, for the same file size and for each of the files sizes. Also, it is noticed that the **difference in compression times of the 3 software significantly increases, as the size (in GB) of the files to be compressed increases**. This relation can be clearly seen in the above graph, through **difference in the mean and median values for the box-plots** as the size of files to be compressed **increases**.

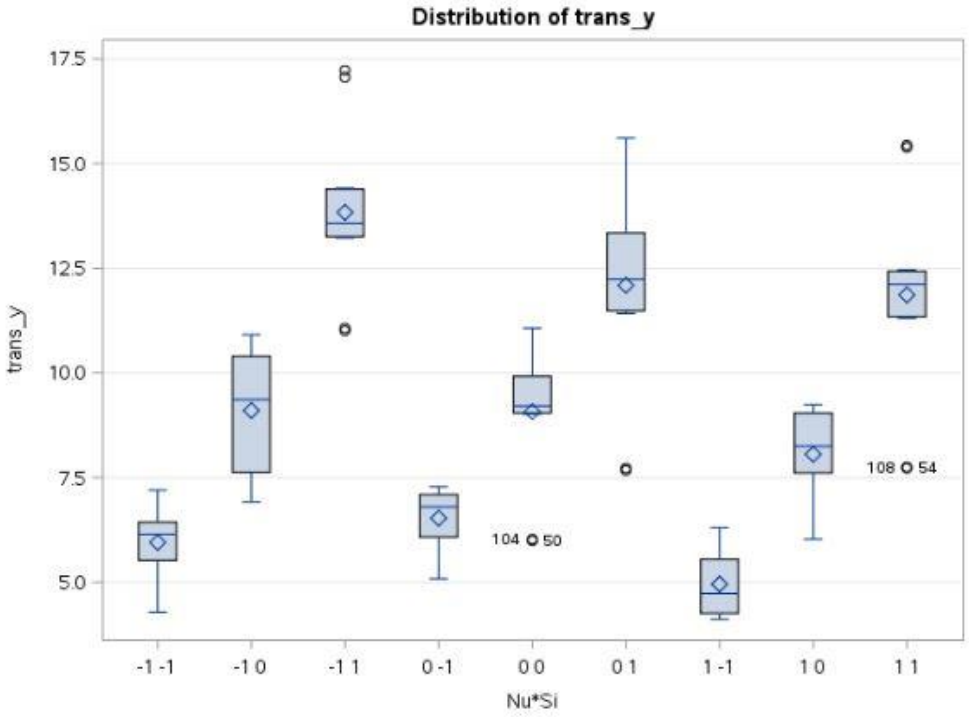

Figure 19 - Interactions effects graph for files sizes and number of files.

Level of Nu	Level of Si	N	trans_y	
			Mean	Std Dev
-1	-1	12	5.9516237	0.91339226
-1	0	12	9.1002120	1.47881351
-1	1	12	13.8340654	1.88689167
0	-1	12	6.5292741	0.76751902
0	0	12	9.0734584	1.59581428
0	1	12	12.0925367	2.49506775
1	-1	12	4.9543429	0.77299560
1	0	12	8.0549124	1.09515524
1	1	12	11.8614535	2.34987872

Figure 20 – Means Table for number of files and file size interaction on transformed data.

Level of Nu	Level of Si	N	y
			Mean
-1	-1	12	69.8373921
-1	0	12	192.068325
-1	1	12	520.645732
0	-1	12	87.1186457
0	0	12	190.703784
0	1	12	377.936882
1	-1	12	45.1564471
1	0	12	143.643533
1	1	12	360.966212

Figure 21 – Means Table for number of files and size interaction on original un- transformed data.

From the Tukey groupings test and a significance level of 0.05, there is an interaction effect found between the number of files to be compressed in a folder and the size of the files (in GB) compressed. Firstly, there's a significant difference found in the compression times between the 3 file sizes, for the same compression software and number of files compressed at once. Also, it is noticed that the **difference in compression times** for the 3 different file sizes of 0.50GB, 1.50GB and 3.00GB, **reduces** as the **number of files** to be compressed at once **increases** from a single file to 250 files to 500 files. This relation can be seen in the above graph, through the reduction in **difference of the mean and median values for the box-plots** for the 3 file number sizes compressed at once.

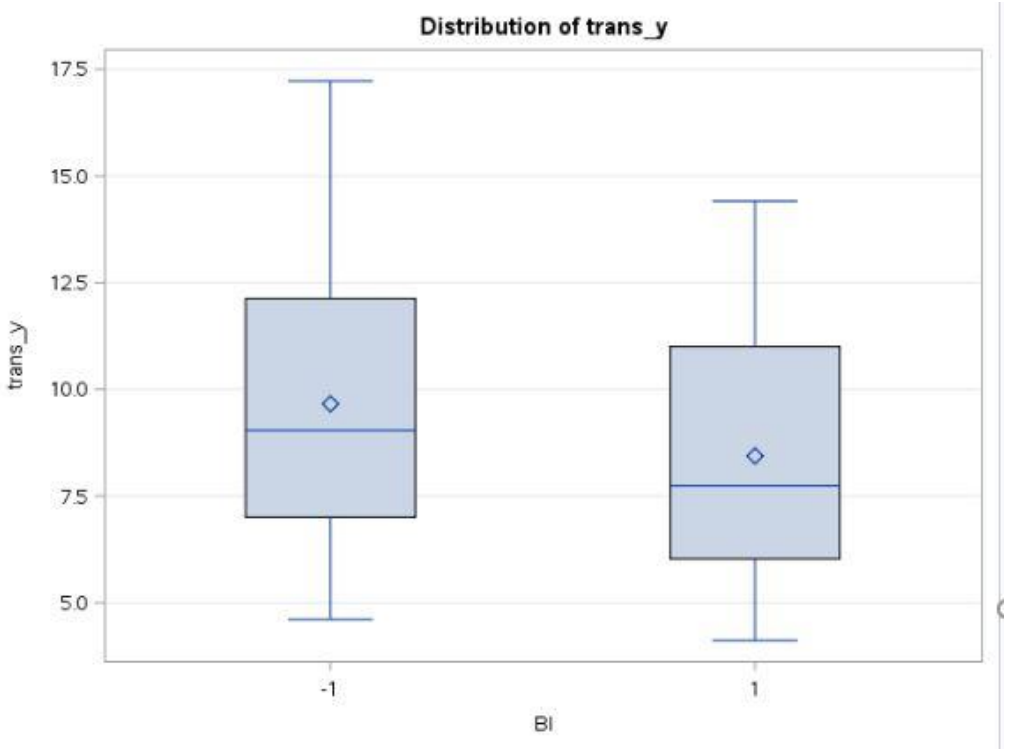

Figure 22- Factor effects graph for the Block variable (Operating System)

The GLM Procedure
Tukey's Studentized Range (HSD) Test for trans_y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	92
Error Mean Square	0.847515
Critical Value of Studentized Range	2.80875
Minimum Significant Difference	0.3519

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	Bl
A	9.6603	54	-1
B	8.4401	54	1

Figure 23 - Tuckey grouping table for Block factor (operating system).

From the Tukey groupings test and a significance level of 0.05, the 2 operating systems **(UBUNTU 14.04 and Windows 10)** have statistically significant difference in compression time for the same compression software, file types and number of files. After un-transforming the response variable, mean compression times for the 2 operating systems, Ubuntu 14.04 and Windows 10, are found to be, **221.419 sec** and **160.542 seconds** respectively.

Conclusion:

From the experiments performed it can be concluded that PeaZip is fastest followed by WinRAR and 7zip. Furthermore, these software work faster on Windows 10 than on Ubuntu 14.04 LTS.

PeaZip is 47.60% faster than 7zip, on average and is 37.5 % faster than WinRAR, on average, while taking into account the entire data set across the 2 operating systems. The compression time in Windows is 27.51 % faster, on average, than it is on Ubuntu.

It is also found that, while compressing files in larger numbers, the 3 compresion software are relatively quicker than compressing fewer file numbers. This interation graph between the number of files and the size of files reveals that for a fixed file size, as the number of files to be compressed increases, the difference in compression times between the 3 software becomes smaller. We can infer that all 3 compression software are optimized to compress larger file numbers

As the size of the files increases, the difference in performance of the 3 software becomes more apparent and subsequently, the difference in time taken for the 3 software increases.

In conclusion, choosing PeaZip to compress files on both Windows 10 and Ubuntu 14.04 LTS will be will be the quickest way (among the 3 compresion software) to compress any number of files of all file sizes.

References

[1] https://alternativeto.net/software/7-zip/

APPENDIX

SAS CODE:

```
/* Constructing the Data-Set for ANOVA and GLM Construction:
   Response Variable (y): Time taken to compress files in Seconds*/
data Time;
     do Bl=-1 to 1 by 2;
        do So=-1 to 1 by 1;
            do Si=-1 to 1 by 1;
                do Nu=-1 to 1 by 1;
                    input y @@;
                    output;
                end;
            end;
        end;
     end;
datalines;
110 113 58 231 301 199 878 694 676 59 100 78 127 235 157 468 479 387
69 92 38 180 192 124 496 340 327 83 74 32 261 206 147 500 436 401 80
104 42 293 190 188 570 338 374 32 48 29 100 71 72 302 128 130
107 112 60 226 306 190 858 688 671 58 103 80 125 237 159 473 476 381
71 90 39 186 188 127 493 344 322 85 73 31 266 202 146 506 439 406 80
108 44 296 193 190 574 330 379 34 50 31 105 72 72 306 130 131
;
/* Mapping Variable classifiactions to factor levels:
   Nu: [-1,0,1] =[1,250,500]; "Nu" - Number of files;
   Si: [-1,0,1] =[0.5,1.5,3.0]; "Si" - Size of the files in GB
   So: [-1,0,1] =[7Zip,WinRAR,PeaZIP]; "So" - Compression Sofwate
used
   Bl: [-1, 1 ] =[UBUNTU, Windows]; "Bl" - Operating System tested
on
*/

/* PRINTING OUT ORIGINAL Data-Set */
proc print data =Time;
run;

/* TRANSFORMING THE RESPONSE VARIABLE TO LOG10(Y),TO SATISFY ANOVA
ASSUMPTIONS
   AND REMOVE TREND IN MODEL RESIDUAL PLOT */
data two;
set Time;
trans_y = y**0.42;
run;

proc print data=two;
run;
```

```
/*Constructing the block model for the CRBD model with 4 treatment
effects   */

data inter;
    set two;
    NuSi=Nu*Si;
    NuSo=Nu*So;
    NuBl=Nu*Bl;
    SiSo=Si*So;
    SiBl=Si*Bl;
    SoBl=So*Bl;
    NuSiSo=NuSi*So;
    NuSiBl=NuSi*Bl;
    NuSoBl=NuSo*Bl;
    SiSoBl=SiSo*Bl;
    NuSiSoBl=NuSiSo*Bl;
run;

proc print data=inter;
run;

/*
 Obtaining the ANOVA table and checking for significant terms: *
 ONLY SIGNIFICANT TERMS INCLUDED:
 */
proc anova data=inter;
class Nu Si So Bl NuSi NuSo NuBl SiSo SiBl SoBl NuSiSo NuSiBl NuSoBl
SiSoBl NuSiSoBl;
model trans_y = Nu|Si|So Bl; /*Nu Si So Bl NuSi NuSo NuBl SiSo SiBl
SoBl SiBl; */
/*model log_y = So|Si Si|Nu Bl|So; NuSi NuSo NuBl SiSo SiBl SoBl
NuSiSo NuSiBl NuSoBl SiSoBl NuSiSoBl;*/
/*means So|Si Si|Nu Bl|So/lines tukey alpha=0.05; */
run;

/* Constructing a GLM model to identify and test for significant
predictor variables in the model.
   ONLY SIGNIFICANT TERMS INCLUDED [FINAL MODEL]:
*/
proc glm data=inter;
    class Nu Si So Bl NuSi NuSo NuBl SiSo SiBl SoBl NuSiSo NuSiBl
NuSoBl SiSoBl NuSiSoBl;
    model trans_y = So|Si Si|Nu Bl;
    means So|Si Si|Nu Bl/lines tukey alpha=0.05;
    output out=diag p=pred r=res;
run;
/* The reduced glm model for estimate and analysis of the predictor
variables in the model.
```

```
   Although the term NuSo is found to be insignificant, it is
included in the model since a higher order
   interaction term NuSiSo is found to be significant in the model.
*/

/* Residuals vs. Predicted values */
proc sort;
by pred;
symbol1 v=circle i=sm50;
title1 'Residual Plot';

proc gplot;
plot res*pred/frame;
run;

/* QQ-Plot and Histogram */
proc univariate data=diag noprint;
var res;
qqplot res / normal;
histogram res / normal;
run;

/*Constructing the diagnostics plots and checking for any violations
of ANOVA assumptions
  [Normality of model residuals, constant error variance, etc.] */
proc reg outest=effects data=inter;
    model trans_y = So Si Nu Bl SiSo NuSi SoBl SiBl; /*So*Si Si*Nu
Bl*So Bl*Si;/*A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD; */
run;

/*DIAGNOSTICS PLOTS OF TEH ORIGINAL (UN-tRANSFORMED) MODEL: */
proc reg outest=effects data=inter;
    model y = So Si Nu Bl SiSo NuSi NuSo; /* NuSiSoSo*Si Si*Nu Bl*So
Bl*Si;/*A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD; */
run;
```

All Experimental readings:

Obs	Bl	So	Si	Nu	Rep1	Rep2
1	-1	-1	-1	-1	110	107
2	-1	-1	-1	0	113	112
3	-1	-1	-1	1	58	60
4	-1	-1	0	-1	231	226
5	-1	-1	0	0	301	306
6	-1	-1	0	1	199	190
7	-1	-1	1	-1	878	858
8	-1	-1	1	0	694	688

9	-1	-1	1	1	676	671
10	-1	0	-1	-1	59	58
11	-1	0	-1	0	100	103
12	-1	0	-1	1	78	80
13	-1	0	0	-1	127	125
14	-1	0	0	0	235	237
15	-1	0	0	1	157	159
16	-1	0	1	-1	468	473
17	-1	0	1	0	479	476
18	-1	0	1	1	387	381
19	-1	1	-1	-1	69	71
20	-1	1	-1	0	92	90
21	-1	1	-1	1	38	39
22	-1	1	0	-1	180	186
23	-1	1	0	0	192	188
24	-1	1	0	1	124	127
25	-1	1	1	-1	496	493
26	-1	1	1	0	340	344
27	-1	1	1	1	327	322
28	1	-1	-1	-1	83	85
29	1	-1	-1	0	74	73
30	1	-1	-1	1	32	31
31	1	-1	0	-1	261	266
32	1	-1	0	0	206	202
33	1	-1	0	1	147	146
34	1	-1	1	-1	500	506
35	1	-1	1	0	436	439
36	1	-1	1	1	401	406
37	1	0	-1	-1	80	80
38	1	0	-1	0	104	108
39	1	0	-1	1	42	44
40	1	0	0	-1	293	296
41	1	0	0	0	190	193
42	1	0	0	1	188	190
43	1	0	1	-1	570	574
44	1	0	1	0	338	330
45	1	0	1	1	374	379
46	1	1	-1	-1	32	34
47	1	1	-1	0	48	50
48	1	1	-1	1	29	31
49	1	1	0	-1	100	105
50	1	1	0	0	71	72
51	1	1	0	1	72	72
52	1	1	1	-1	302	306
53	1	1	1	0	128	130

54	1	1	1	1	130	131