9/3/2023

User Mode CPU Time Prediction – Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

Problem:

The comp-activ databases is a collection of a computer systems activity measures .
The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

Dataset for Problem 1: compactiv.xlsx

DATA DICTIONARY:

System measures used:

lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.
exec - Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls
wchar - Number of characters transfreed per second by system write calls
pgout - Number of page out requests per second
ppgout - Number of pages, paged out per second
pgfree - Number of pages per second placed on the free list.
pgscan - Number of pages checked if they can be freed per second
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second
pgin - Number of page-in requests per second
ppgin - Number of pages paged in per second
pflt - Number of page faults caused by protection errors (copy-on-writes).
vflt - Number of page faults caused by address translation .
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.
----
usr - Portion of time (%) that cpus run in user mode.

Linear Regression

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

The dataset that is given for analysis is comp-activ databases is a collection of a computer systems activity measures.

Shape:

```
The shape of the dataset is (8192, 22)
```

There are 8192 Rows and 22 columns in the dataset.

First Five (Head):

The First Five rows of the dataset (The rows and columns has been transposed for easier view). Refer jupyter workings for the output.

	0	1	2	3	4
Lread	1	0	15	0	5
lwrite	0	0	3	0	1
Scall	2147	170	2162	160	330
Sread	79	18	159	12	39
swrite	68	21	119	16	38
Fork	0.2	0.2	2	0.2	0.4
Exec	0.2	0.2	2.4	0.2	0.4
Rchar	40671	448			
wchar	53995	8385	31950	8670	12185
pgout	0	0	0	0	0
ppgout	0	0	0	0	0
pgfree	0	0	0	0	0
pgscan	0	0	0	0	0
Atch	0	0	1.2	0	0
Pgin	1.6	0	6	0.2	1
Ppgin	2.6	0	9.4	0.2	1.2
Pflt	16	15.63	150.2	15.6	37.8
Vflt	26.4	16.83	220.2	16.8	47.6
runqsz	CPU_Bound	Not_CPU_Bound	Not_CPU_Bound	Not_CPU_Bound	Not_CPU_Bound
freemem	4670	7278	702	7248	633
freeswap	1730946	1869002	1021237	1863704	1760253
usr	95	97	87	98	90

Table 1 First Five rows of dataset

Info:

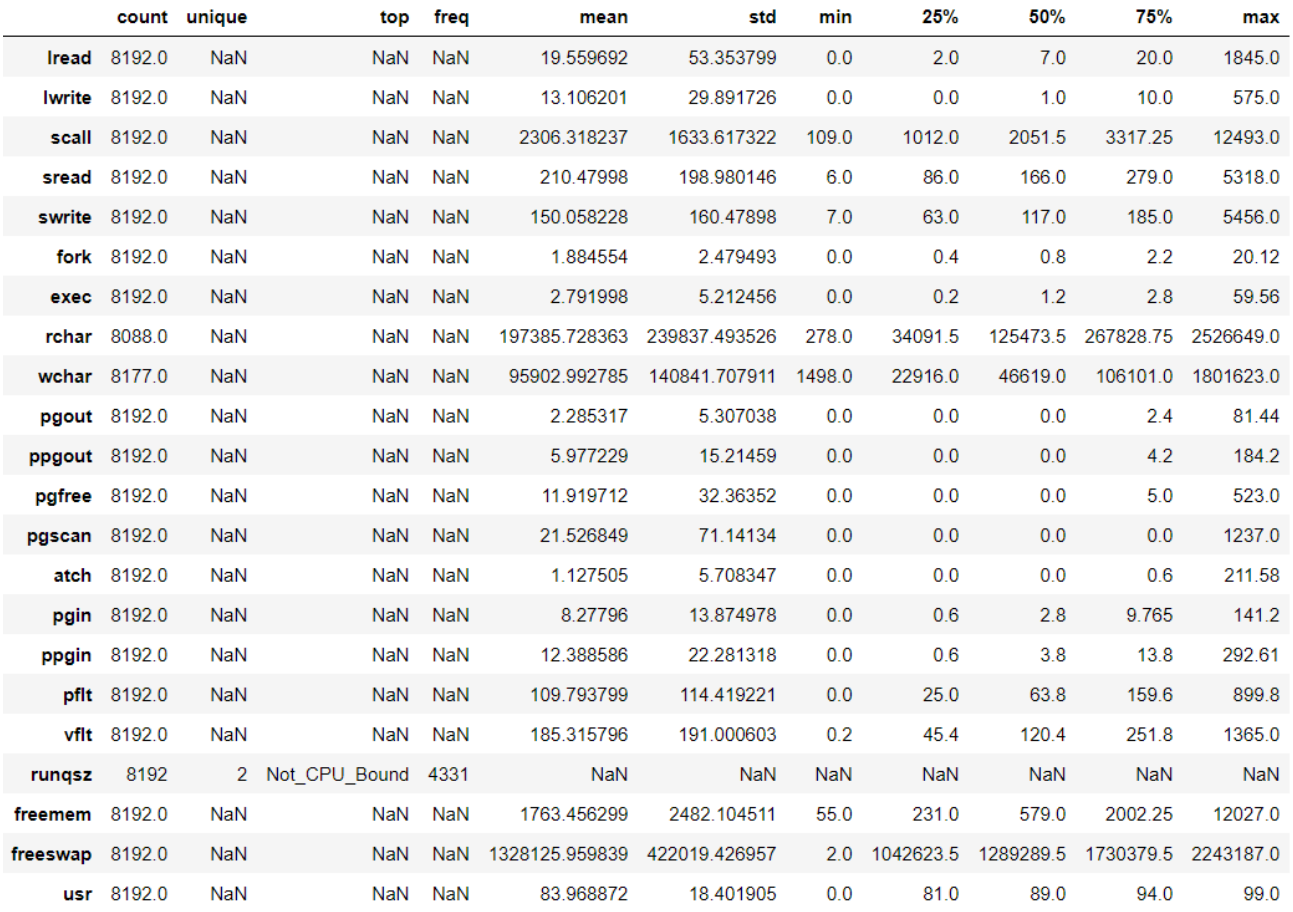
The Info of the dataset is

```
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   lread     8192 non-null   int64  
 1   lwrite    8192 non-null   int64  
 2   scall     8192 non-null   int64  
 3   sread     8192 non-null   int64  
 4   swrite    8192 non-null   int64  
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object 
 19  freemem   8192 non-null   int64  
 20  freeswap  8192 non-null   int64  
 21  usr       8192 non-null   int64  
dtypes: float64(13), int64(8), object(1)
memory usage: 1.4+ MB
```

There are null values in the 'rchar' and 'wchar' columns.

Other than ‘runqsz’, all the columns are numerical variables.

Five Point Summary:



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
lread	8192.0	NaN	NaN	NaN	19.559692	53.353799	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.106201	29.891726	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.318237	1633.617322	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.47998	198.980146	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.058228	160.47898	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.884554	2.479493	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.791998	5.212456	0.0	0.2	1.2	2.8	59.56
rchar	8088.0	NaN	NaN	NaN	197385.728363	239837.493526	278.0	34091.5	125473.5	267828.75	2526649.0
wchar	8177.0	NaN	NaN	NaN	95902.992785	140841.707911	1498.0	22916.0	46619.0	106101.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.285317	5.307038	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.977229	15.21459	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.919712	32.36352	0.0	0.0	0.0	5.0	523.0
pgscan	8192.0	NaN	NaN	NaN	21.526849	71.14134	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.127505	5.708347	0.0	0.0	0.0	0.6	211.58
pgin	8192.0	NaN	NaN	NaN	8.27796	13.874978	0.0	0.6	2.8	9.765	141.2
ppgin	8192.0	NaN	NaN	NaN	12.388586	22.281318	0.0	0.6	3.8	13.8	292.61
pflt	8192.0	NaN	NaN	NaN	109.793799	114.419221	0.0	25.0	63.8	159.6	899.8
vflt	8192.0	NaN	NaN	NaN	185.315796	191.000603	0.2	45.4	120.4	251.8	1365.0
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192.0	NaN	NaN	NaN	1763.456299	2482.104511	55.0	231.0	579.0	2002.25	12027.0
freeswap	8192.0	NaN	NaN	NaN	1328125.959839	422019.426957	2.0	1042623.5	1289289.5	1730379.5	2243187.0
usr	8192.0	NaN	NaN	NaN	83.968872	18.401905	0.0	81.0	89.0	94.0	99.0

Table 2 Five Point Summary

The Five point summary can be seen from the above table.

The portion of time that cpu runs in user mode is with a mean of 84 sec. The user mode time can’t be zero percent logically. It is only logical if the other independent variables are also zero. This is imputed in the later part.

runqsz has maximum of "Not_CPU_Bound" in the dataset.

EDA:

Univariate:

```
RUNQSZ
Not_CPU_Bound    4331
CPU_Bound        3861
Name: runqsz, dtype: int64
***********************************************************************
```

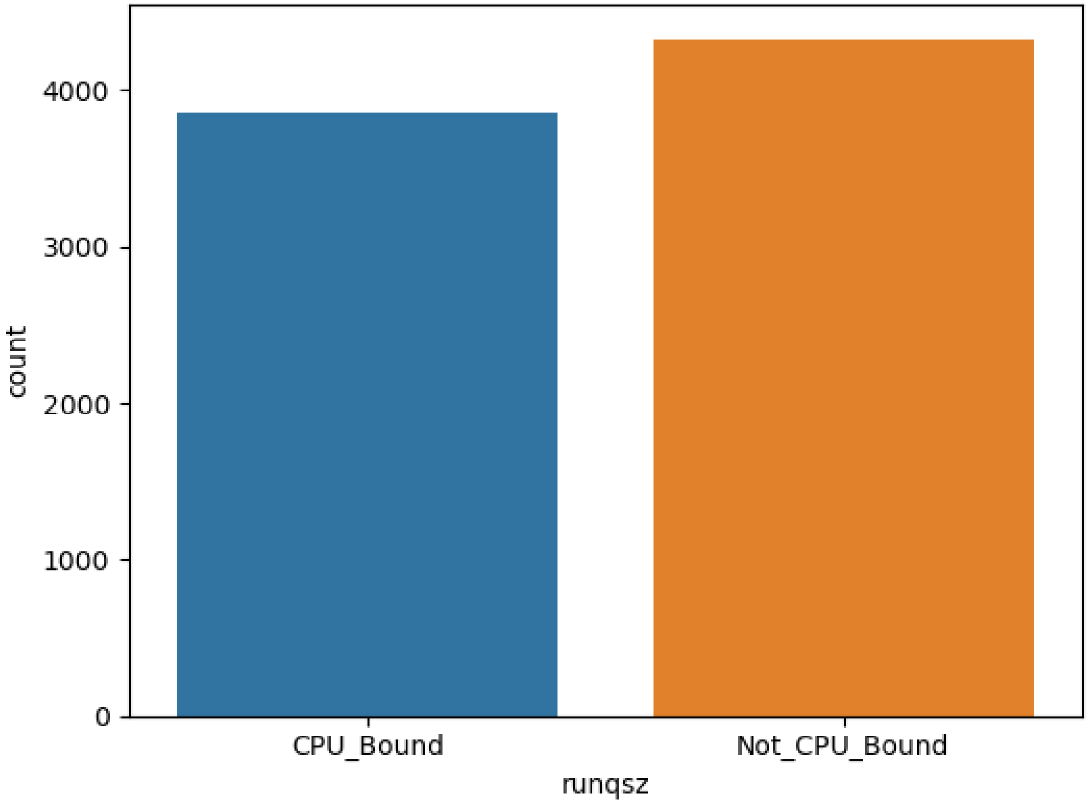


Figure A Count plot for runqsz

More than 50% of the dataset is Not CPU Bound - The number of kernel threads in memory that are waiting for a CPU to run is less than 2.

Histogram:

The below plot shows Histogram for lread,scall,freeswap and usr numerical variables. Refer jupyter notebook for histogram for all variables.

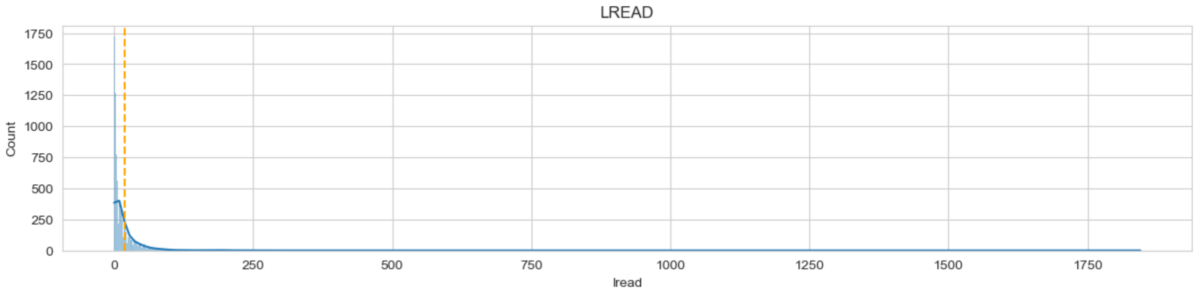


Figure B Histogram of lread

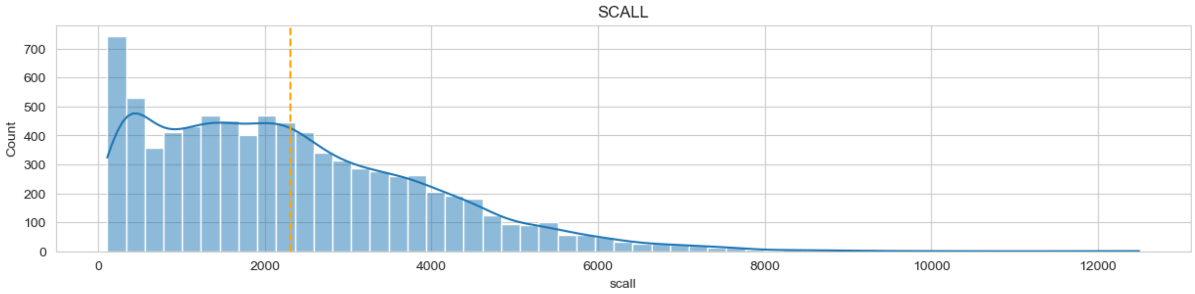


Figure C Histogram of scall

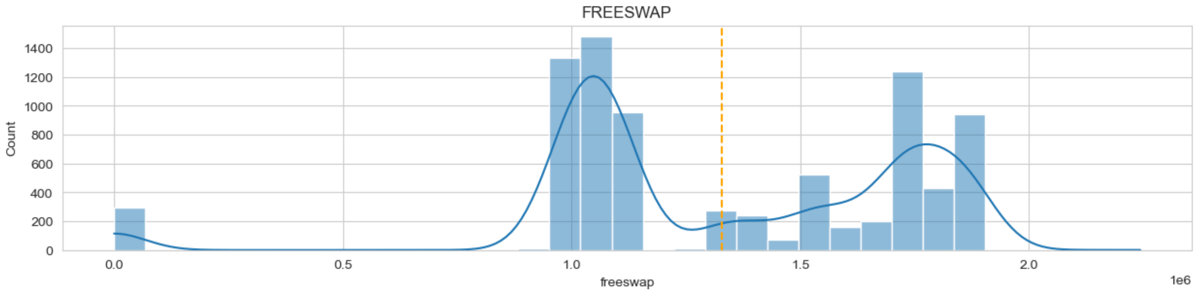


Figure D Histogram of freeswap

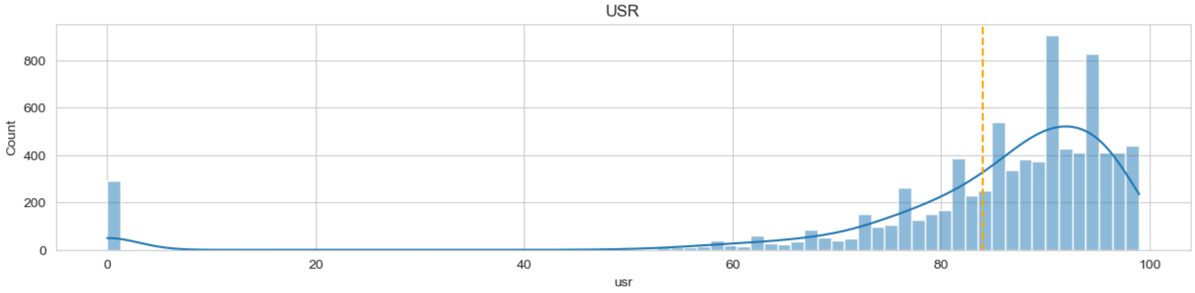


Figure E Histogram of usr

Most of the dataset is right skewed. We can confirm it through the skewness below also.

The usr is left skewed and skewness also confirms the outlier’s direction.

Freeswap is bit left skewed, and also the skewness is much closer to zero than usr.

Skewness:

```
Skewness in lread = 13.895307340580494

Skewness in lwrite = 5.276678111482097

Skewness in scall = 0.9023659545115966

Skewness in sread = 5.458466253528661

Skewness in swrite = 9.604084726213738

Skewness in fork = 2.249277187819112

Skewness in exec = 4.068492569973669

Skewness in rchar = nan

Skewness in wchar = nan

Skewness in pgout = 5.06605627889887

Skewness in ppgout = 4.679584596517809

Skewness in pgfree = 4.767318125798512

Skewness in pgscan = 5.812350621813612

Skewness in atch = 21.538075020993087

Skewness in pgin = 3.24181874259949

Skewness in ppgin = 3.902050260208112

Skewness in pflt = 1.71996910974841

Skewness in vflt = 1.7370084627910771

Skewness in freemem = 1.8072236633447214

Skewness in freeswap = -0.7915194783899336

Skewness in usr = -3.4161239456823633
```

Right Skewed Variables

- lread
- lwrite
- scall
- sread
- swrite

- fork
- exec
- rchar
- wchar
- pgout
- ppgout
- pgfree
- pgscan
- atch
- pgin
- ppgin
- pflt
- vflt
- freemem

Left Skewed

- freeswap
- usr

Note: rchar and wchar values are calculated once the null values are imputed. It shows nan due to the null values.

After imputing rchar and wchar null values with mean.

```
Skewness in rchar = 2.871802562692287

Skewness in wchar = 3.8504752401786484
```

Bivariate:

Heatmap(Correlation):

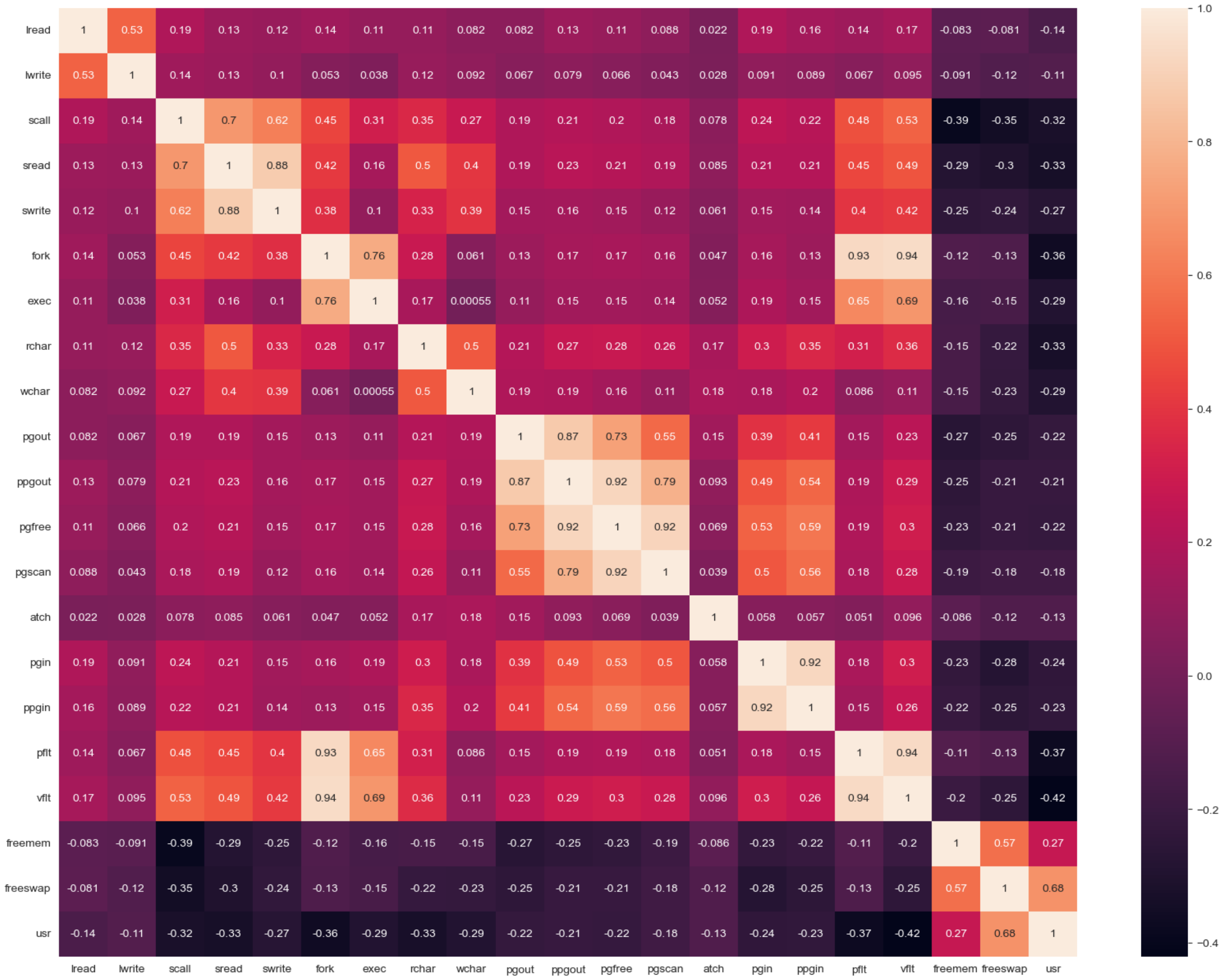


Figure F Correlation Heatmap

Most of the independent variables has correlation. There is multicollinearity between the variables. This has to be removed for the assumptions of linear regression. One of the assumptions of Linear regression is that the independent variables should not have any correlation between them.

The relationship of independent with dependent variables are also strong. The below heatmap has the correlation of dependent with independent variables.

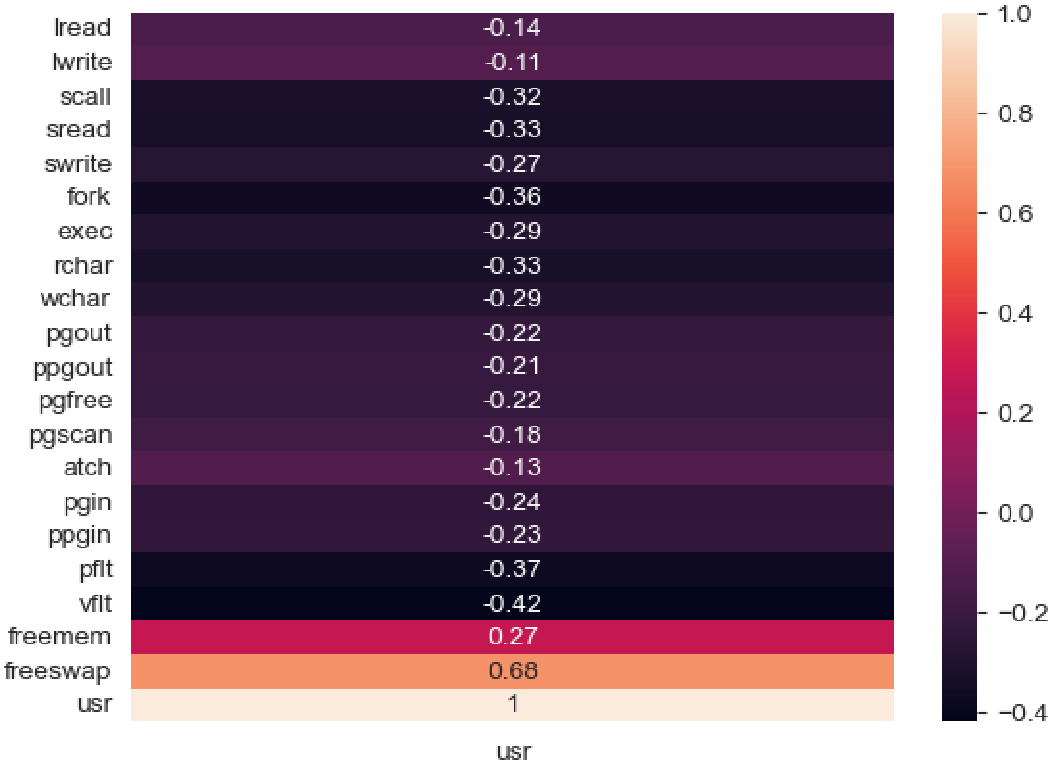


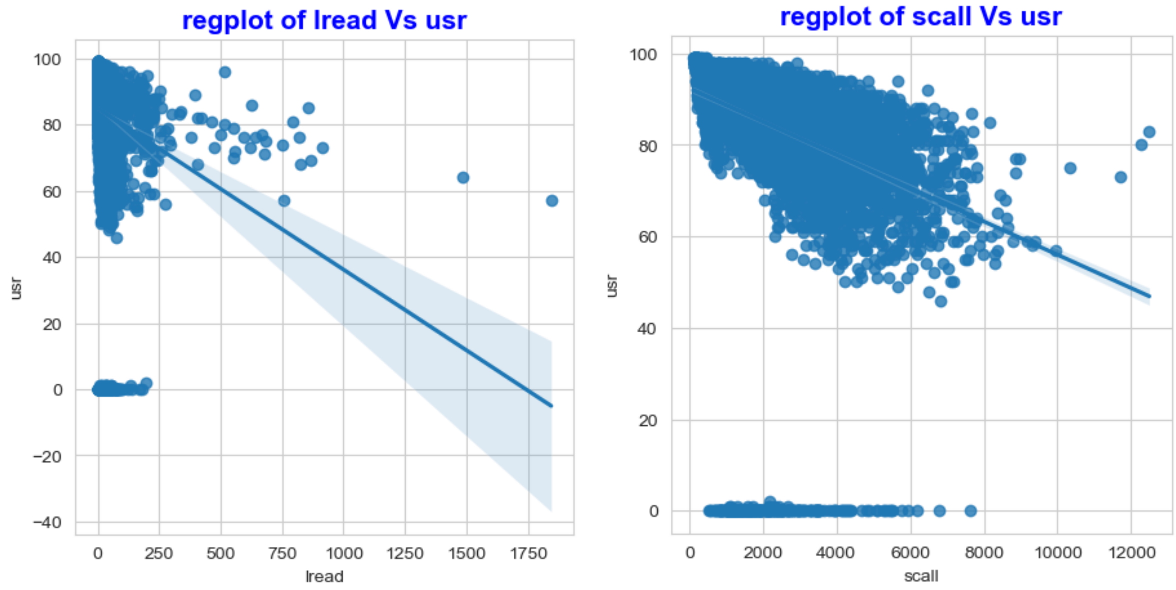
Figure G Correlation Heatmap (Independent vs dependent)

Freeswap and freemem has positive correlation with the usr.

Others are negatively correlated with usr.

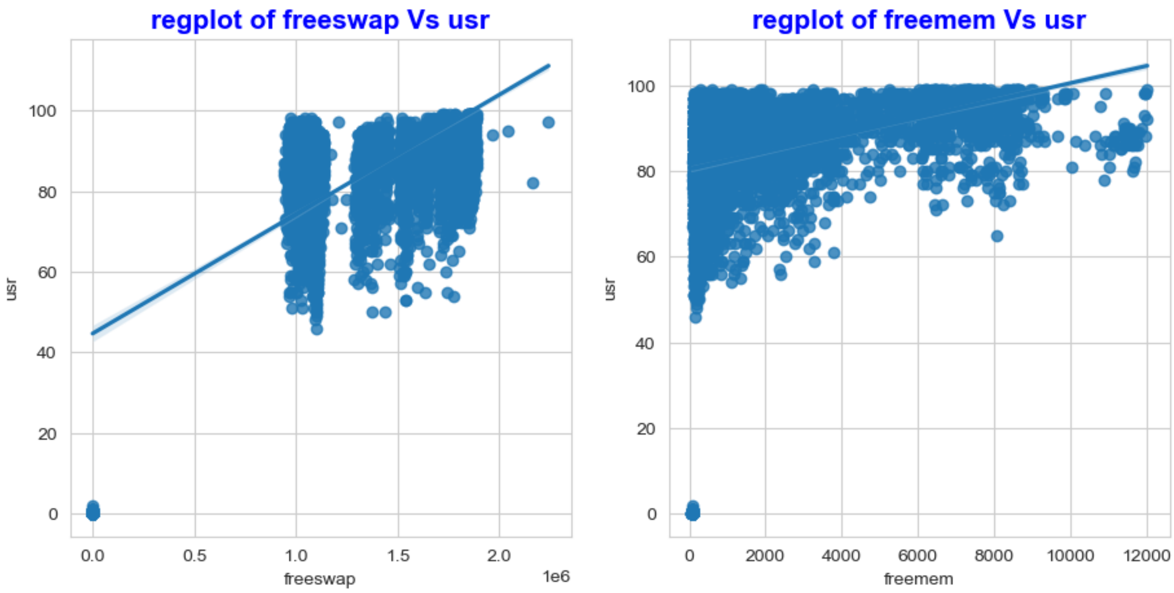
This correlation can also be seen in the regression plot with usr in the below plots. (Refer Juypter notebook for regression plot of all variables)

Figure H regplot of lread and scall Vs usr



Both lread and scall are negatively correlated with usr and the regression line is also downwards in the plots.

Figure I regplot of freeswap and freemem Vs usr



regplot of freeswap Vs usr

regplot of freemem Vs usr

Both freeswap and freemem is positively correlated with the usr, there is an upward slope in the regression plots.

Multivariate:

(Refer Juypter notebook for Scatter plot of all variables)

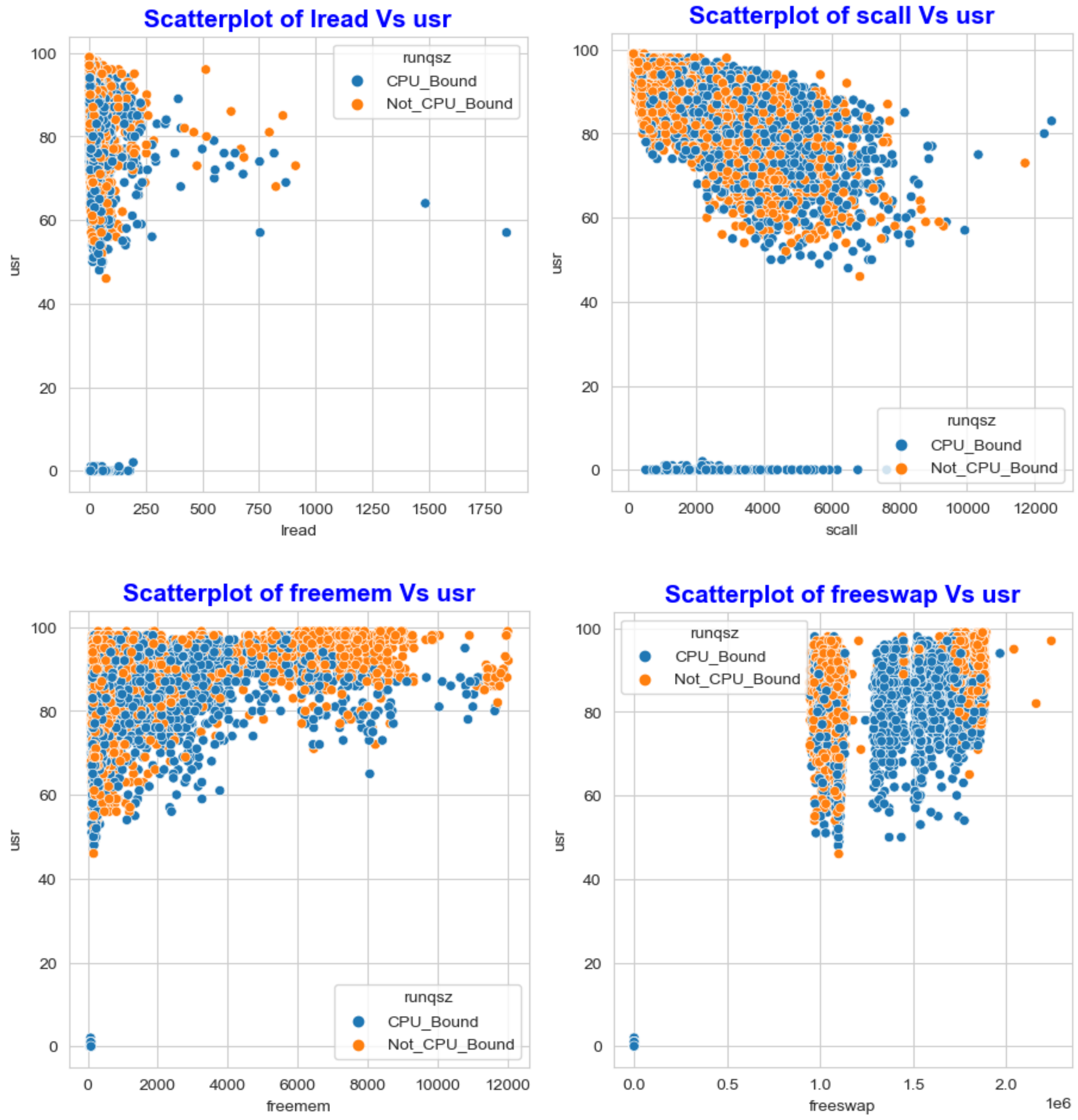


Figure J Scatterplot with usr

There is no clear visual difference with ‘CPU_Bound’ and ‘Not_CPU_Bound’. The dataset is intermingled with both category. There is not much segregation of ‘CPU_Bound’ and ‘Not_CPU_Bound’ with other variables.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.

Null values and Duplicates:

There are 104 and 15 null values with rchar and wchar respectively.

```
rchar       104
wchar        15
```

There are no duplicates with the dataset.

The Null values are imputed with the mean values of rchar and wchar respectively.

Outliers:

The boxplot visually shows the outliers which has been inferred from the skewness of the dataset as it confirms the outlier's direction.

We are keeping the outliers as they might be of some value to the analysis.

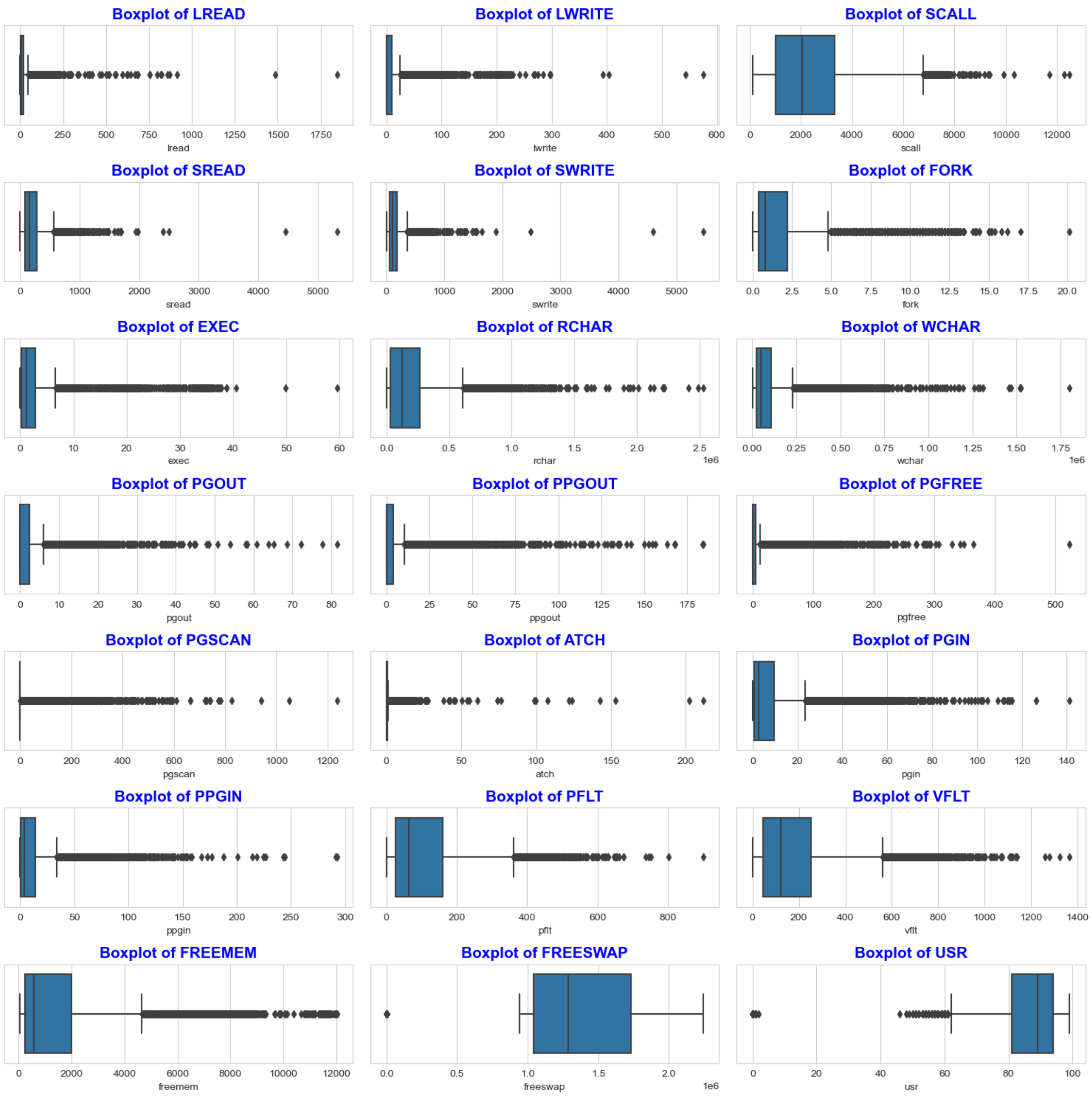


Figure K Boxplot

Zeros in USR:

usr - Portion of time (%) that cpus run in user mode can't be zero, imputing them with mean.

Other zeros are seeming valid as all of them are calls/pages/characters transferred per second

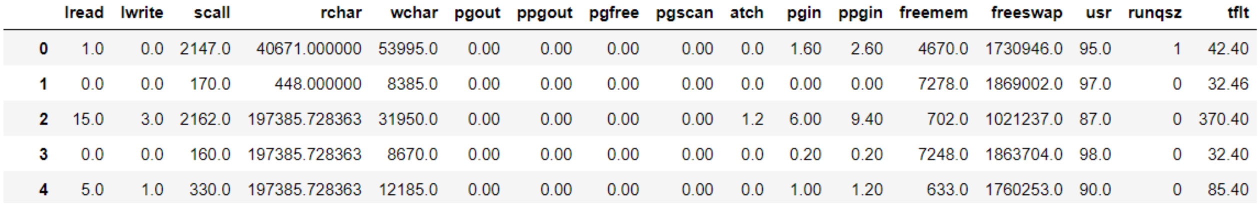
New Feature:

Dropping sread, swrite, fork, exec as they are covered in scall.

The pflt and vflt are combined to create a new feature of tflt.

tflt - Total number of page faults caused.

```
The shape of the dataset is (8192, 17)
```



	lread	lwrite	scall	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	freemem	freeswap	usr	runqsz	tflt
0	1.0	0.0	2147.0	40671.000000	53995.0	0.00	0.00	0.00	0.00	0.0	1.60	2.60	4670.0	1730946.0	95.0	1	42.40
1	0.0	0.0	170.0	448.000000	8385.0	0.00	0.00	0.00	0.00	0.0	0.00	0.00	7278.0	1869002.0	97.0	0	32.46
2	15.0	3.0	2162.0	197385.728363	31950.0	0.00	0.00	0.00	0.00	1.2	6.00	9.40	702.0	1021237.0	87.0	0	370.40
3	0.0	0.0	160.0	197385.728363	8670.0	0.00	0.00	0.00	0.00	0.0	0.20	0.20	7248.0	1863704.0	98.0	0	32.40
4	5.0	1.0	330.0	197385.728363	12185.0	0.00	0.00	0.00	0.00	0.0	1.00	1.20	633.0	1760253.0	90.0	0	85.40

Table 3 First five rows of dataset after new feature

Heatmap for new feature:

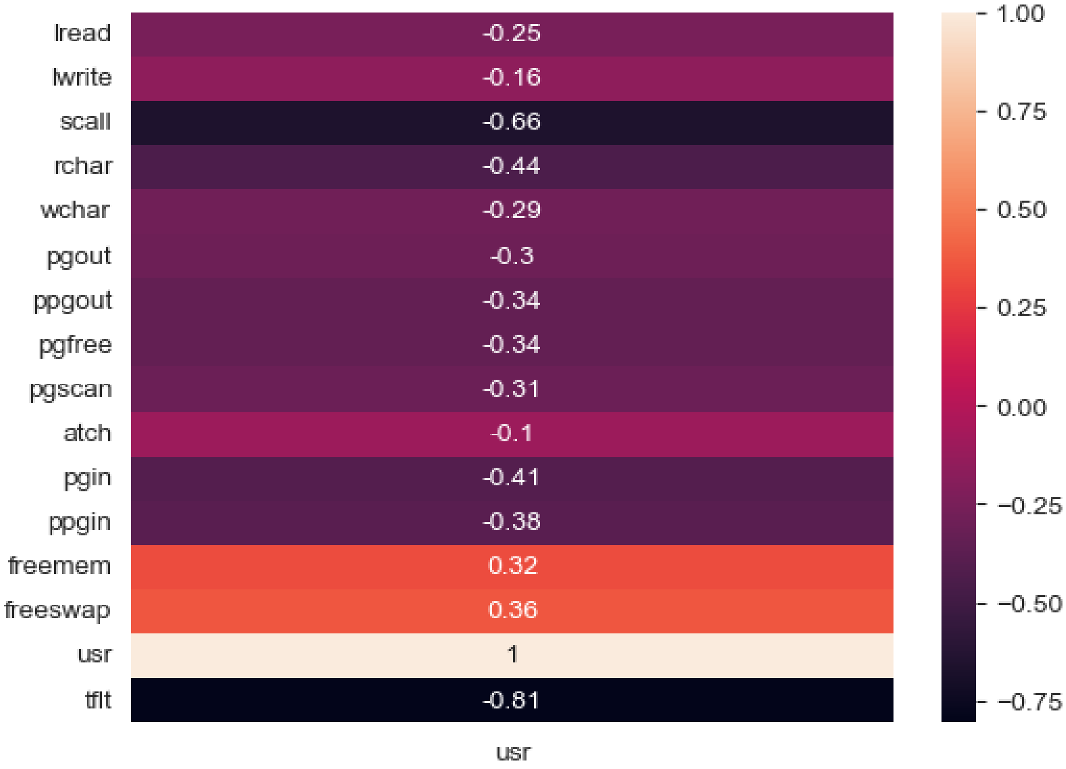


Figure L Correlation Heatmap(Independent vs dependent) with new feature

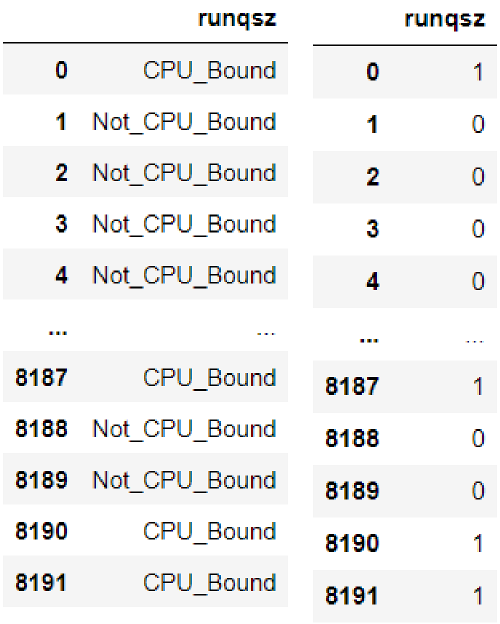
The new feature is tflt is negatively correlated like vflt and pflt.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

The runqsz has been labelled as 0 and 1 for 'Not_CPU_Bound' and 'CPU_Bound'.

Before and After Labelling:

Table 4 Before and After labelling runqsz



	runqsz
0	CPU_Bound
1	Not_CPU_Bound
2	Not_CPU_Bound
3	Not_CPU_Bound
4	Not_CPU_Bound
...	...
8187	CPU_Bound
8188	Not_CPU_Bound
8189	Not_CPU_Bound
8190	CPU_Bound
8191	CPU_Bound

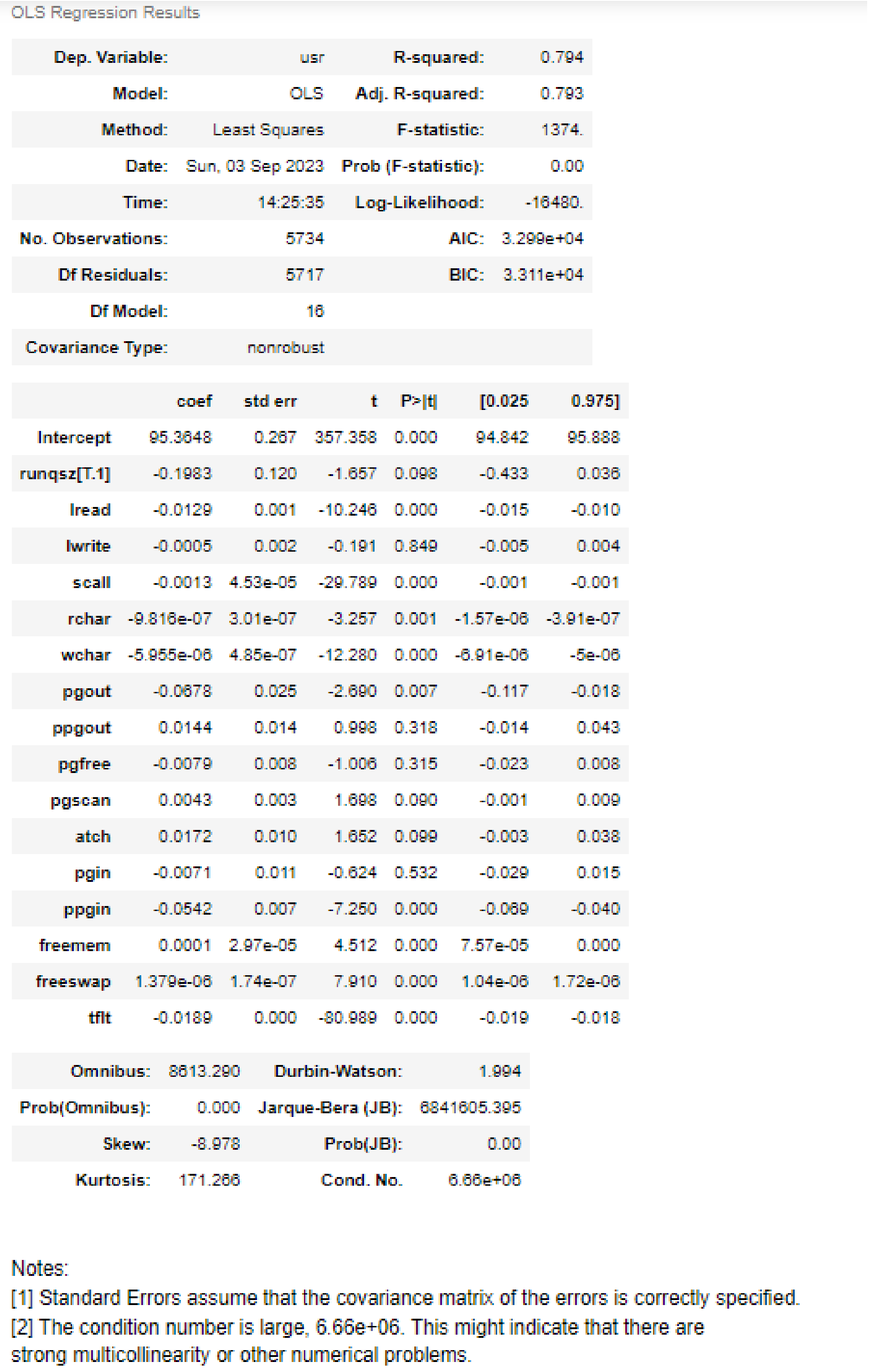
	runqsz
0	1
1	0
2	0
3	0
4	0
...	...
8187	1
8188	0
8189	0
8190	1
8191	1

The dataset is split into 70% Train and 30% Test data.

The training data is fitted into the model for linear regression. The model 1 has all the variables.

Model 1:

expr = 'usr~lread+lwrite+scall+rchar+wchar+pgout+ppgout+pgfree+pgscan+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	1374.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16480.
No. Observations:	5734	AIC:	3.299e+04
Df Residuals:	5717	BIC:	3.311e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3648	0.267	357.358	0.000	94.842	95.888
runqsz[T.1]	-0.1983	0.120	-1.657	0.098	-0.433	0.036
lread	-0.0129	0.001	-10.246	0.000	-0.015	-0.010
lwrite	-0.0005	0.002	-0.191	0.849	-0.005	0.004
scall	-0.0013	4.53e-05	-29.789	0.000	-0.001	-0.001
rchar	-9.816e-07	3.01e-07	-3.257	0.001	-1.57e-06	-3.91e-07
wchar	-5.955e-06	4.85e-07	-12.280	0.000	-6.91e-06	-5e-06
pgout	-0.0678	0.025	-2.690	0.007	-0.117	-0.018
ppgout	0.0144	0.014	0.998	0.318	-0.014	0.043
pgfree	-0.0079	0.008	-1.006	0.315	-0.023	0.008
pgscan	0.0043	0.003	1.698	0.090	-0.001	0.009
atch	0.0172	0.010	1.652	0.099	-0.003	0.038
pgin	-0.0071	0.011	-0.624	0.532	-0.029	0.015
ppgin	-0.0542	0.007	-7.250	0.000	-0.069	-0.040
freemem	0.0001	2.97e-05	4.512	0.000	7.57e-05	0.000
freeswap	1.379e-06	1.74e-07	7.910	0.000	1.04e-06	1.72e-06
tflt	-0.0189	0.000	-80.989	0.000	-0.019	-0.018

Omnibus:	8613.290	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6841605.395
Skew:	-8.978	Prob(JB):	0.00
Kurtosis:	171.266	Cond. No.	6.66e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.66e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 5 Model 1 Summary

There is multicollinearity in the data as the condition number is high.

The VIF for Model 1 are

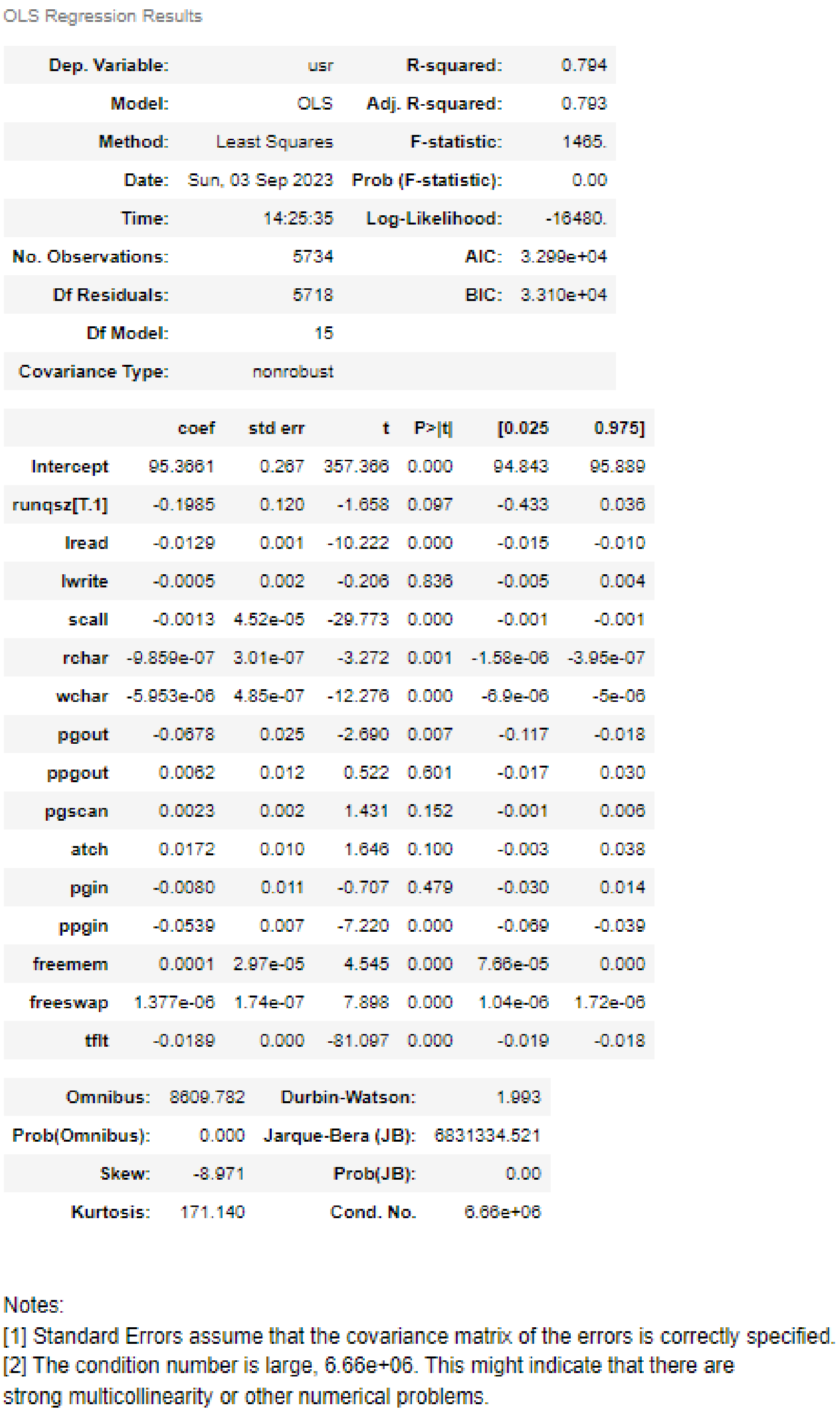
```
lread ---> 1.6414460735410656
lwrite ---> 1.652150073197707
scall ---> 4.543173080830532
rchar ---> 2.7684050909480993
wchar ---> 2.063967081976855
pgout ---> 6.596228402235933
ppgout ---> 17.906759870193884
pgfree ---> 22.827844329049146
pgscan ---> 10.096418470905832
atch ---> 1.1031849258374247
pgin ---> 10.241396699178875
ppgin ---> 11.166047520809958
freemem ---> 2.471116013825802
freeswap ---> 4.980633768190705
runqsz ---> 2.08971225768891
```

Five variables has VIF greater than 5. VIF > 5 means there exists multicollinearity.

Removing pgfree as it has high VIF.

Model 2:

expr2 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+ppgout+pgscan+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	1465.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16480.
No. Observations:	5734	AIC:	3.299e+04
Df Residuals:	5718	BIC:	3.310e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3661	0.267	357.366	0.000	94.843	95.889
runqsz[T.1]	-0.1985	0.120	-1.658	0.097	-0.433	0.036
lread	-0.0129	0.001	-10.222	0.000	-0.015	-0.010
lwrite	-0.0005	0.002	-0.206	0.836	-0.005	0.004
scall	-0.0013	4.52e-05	-29.773	0.000	-0.001	-0.001
rchar	-9.859e-07	3.01e-07	-3.272	0.001	-1.58e-06	-3.95e-07
wchar	-5.953e-06	4.85e-07	-12.276	0.000	-6.9e-06	-5e-06
pgout	-0.0678	0.025	-2.690	0.007	-0.117	-0.018
ppgout	0.0062	0.012	0.522	0.601	-0.017	0.030
pgscan	0.0023	0.002	1.431	0.152	-0.001	0.006
atch	0.0172	0.010	1.646	0.100	-0.003	0.038
pgin	-0.0080	0.011	-0.707	0.479	-0.030	0.014
ppgin	-0.0539	0.007	-7.220	0.000	-0.069	-0.039
freemem	0.0001	2.97e-05	4.545	0.000	7.66e-05	0.000
freeswap	1.377e-06	1.74e-07	7.898	0.000	1.04e-06	1.72e-06
tflt	-0.0189	0.000	-81.097	0.000	-0.019	-0.018

Omnibus:	8609.782	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6831334.521
Skew:	-8.971	Prob(JB):	0.00
Kurtosis:	171.140	Cond. No.	6.66e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.66e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 6 Model 2 Summary

The Adjusted R-Square is the same. The model strength has not dropped but the condition number is still high.

The VIF for Model 2 is

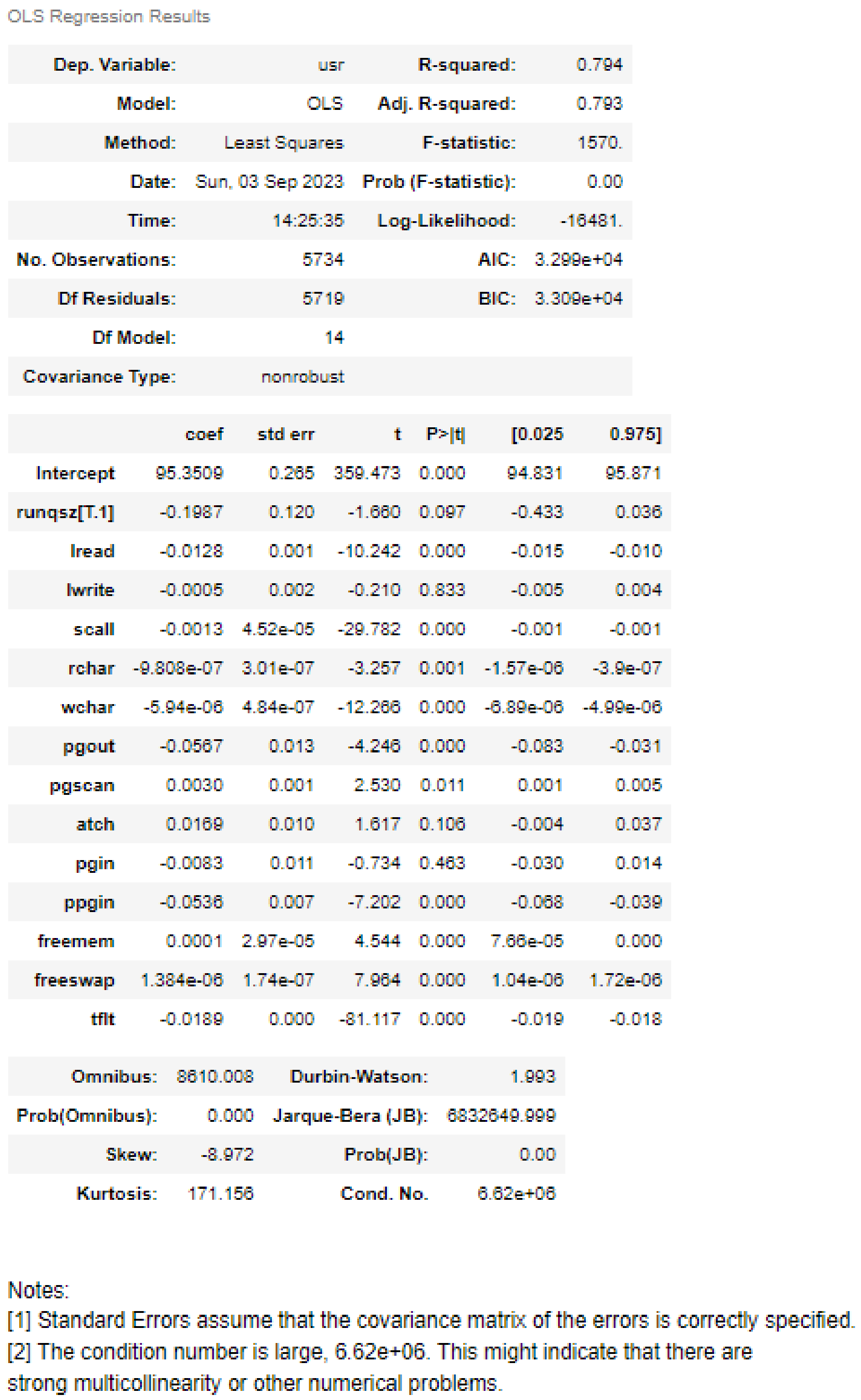
```
lread ---> 1.6401762039453445
lwrite ---> 1.651778117788454
scall ---> 4.539037799348833
rchar ---> 2.7678568889293516
wchar ---> 2.0639059185506707
pgout ---> 6.5962277130252
ppgout ---> 12.187296554356406
pgscan ---> 4.328482433619534
atch ---> 1.1031474690950958
pgin ---> 10.175697690473513
ppgin ---> 11.152956112580462
freemem ---> 2.4688787696016905
freeswap ---> 4.979405846676125
runqsz ---> 2.0897095734001585
tflt ---> 3.0084392653921337
```

The VIF of Pgscan has been decreased after removing pgfree. Removing ppgout as it has high VIF.

After repeating this process for other models.

Model 3:

expr3 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgscan+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.794
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	1570.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16481.
No. Observations:	5734	AIC:	3.299e+04
Df Residuals:	5719	BIC:	3.309e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3509	0.265	359.473	0.000	94.831	95.871
runqsz[T.1]	-0.1987	0.120	-1.660	0.097	-0.433	0.036
lread	-0.0128	0.001	-10.242	0.000	-0.015	-0.010
lwrite	-0.0005	0.002	-0.210	0.833	-0.005	0.004
scall	-0.0013	4.52e-05	-29.782	0.000	-0.001	-0.001
rchar	-9.808e-07	3.01e-07	-3.257	0.001	-1.57e-06	-3.9e-07
wchar	-5.94e-06	4.84e-07	-12.266	0.000	-6.89e-06	-4.99e-06
pgout	-0.0567	0.013	-4.246	0.000	-0.083	-0.031
pgscan	0.0030	0.001	2.530	0.011	0.001	0.005
atch	0.0169	0.010	1.617	0.106	-0.004	0.037
pgin	-0.0083	0.011	-0.734	0.463	-0.030	0.014
ppgin	-0.0536	0.007	-7.202	0.000	-0.068	-0.039
freemem	0.0001	2.97e-05	4.544	0.000	7.66e-05	0.000
freeswap	1.384e-06	1.74e-07	7.964	0.000	1.04e-06	1.72e-06
tflt	-0.0189	0.000	-81.117	0.000	-0.019	-0.018

Omnibus:	8610.008	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6832649.999
Skew:	-8.972	Prob(JB):	0.00
Kurtosis:	171.156	Cond. No.	6.62e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.62e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 7 Model 3 Summary

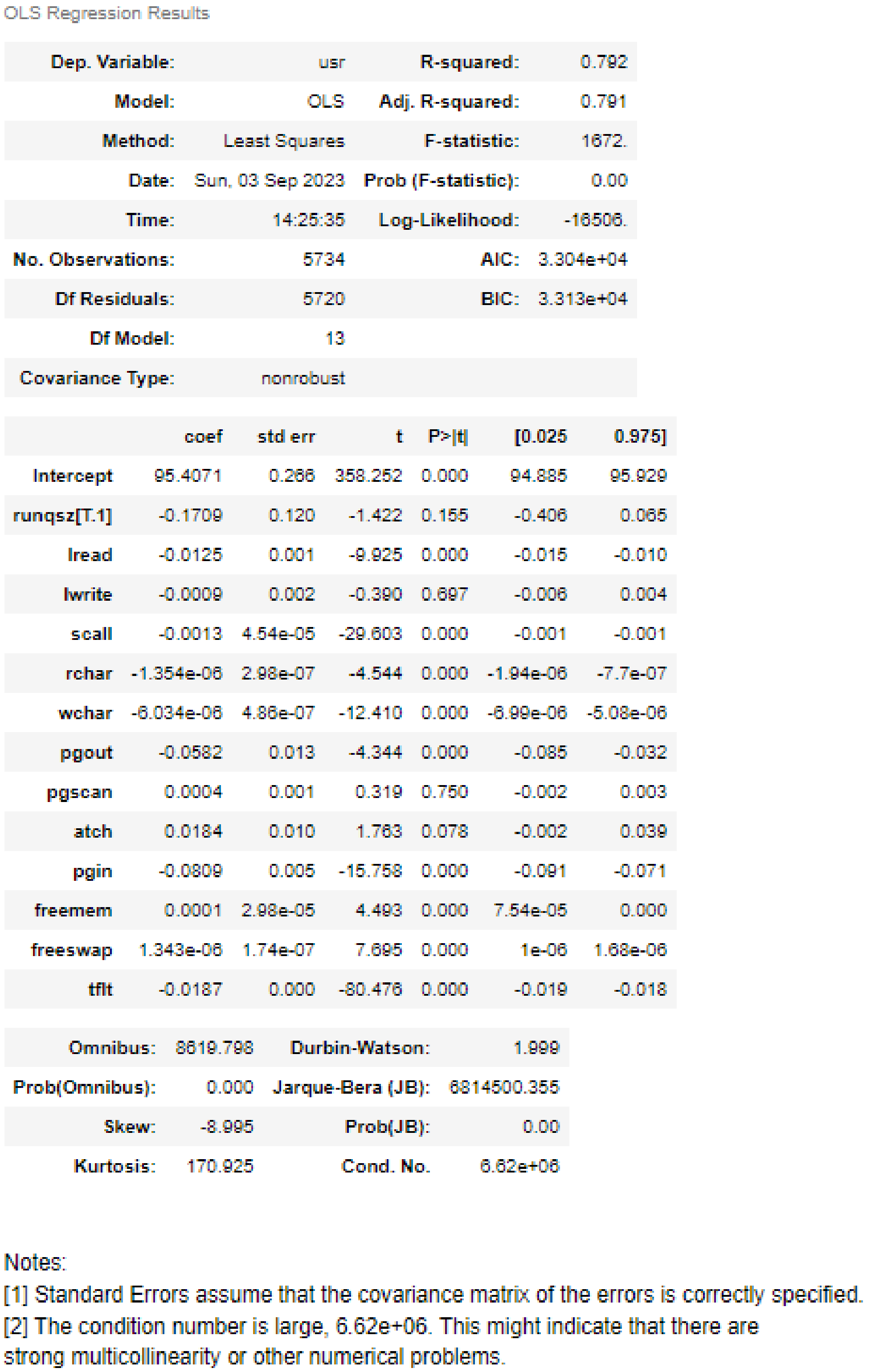
The VIF for Model 3 is

```
lread ---> 1.6094733227405498
lwrite ---> 1.6510609503307339
scall ---> 4.527606954291624
rchar ---> 2.7658143962039277
wchar ---> 2.0610149959790416
pgout ---> 1.8680552284400846
pgscan ---> 2.1865225389073593
atch ---> 1.0982395266971754
pgin ---> 10.13745770306169
ppgin ---> 11.080590224926018
freemem ---> 2.4684757670485364
freeswap ---> 4.975151662713798
runqsz ---> 2.0895275916481197
tflt ---> 3.0070807101363286
```

After Removing ppgout, VIF of pgout has decreased. removing ppgin of high VIF.

Model 4:

expr4 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgscan+atch+pgin+freemem+freeswap+runqsz+ tflt'



OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.792
Model:	OLS	**Adj. R-squared:**	0.791
Method:	Least Squares	**F-statistic:**	1672.
Date:	Sun, 03 Sep 2023	**Prob (F-statistic):**	0.00
Time:	14:25:35	**Log-Likelihood:**	-16506.
No. Observations:	5734	**AIC:**	3.304e+04
Df Residuals:	5720	**BIC:**	3.313e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.4071	0.266	358.252	0.000	94.885	95.929
runqsz[T.1]	-0.1709	0.120	-1.422	0.155	-0.406	0.065
lread	-0.0125	0.001	-9.925	0.000	-0.015	-0.010
lwrite	-0.0009	0.002	-0.390	0.697	-0.006	0.004
scall	-0.0013	4.54e-05	-29.603	0.000	-0.001	-0.001
rchar	-1.354e-06	2.98e-07	-4.544	0.000	-1.94e-06	-7.7e-07
wchar	-6.034e-06	4.86e-07	-12.410	0.000	-6.99e-06	-5.08e-06
pgout	-0.0582	0.013	-4.344	0.000	-0.085	-0.032
pgscan	0.0004	0.001	0.319	0.750	-0.002	0.003
atch	0.0184	0.010	1.763	0.078	-0.002	0.039
pgin	-0.0809	0.005	-15.758	0.000	-0.091	-0.071
freemem	0.0001	2.98e-05	4.493	0.000	7.54e-05	0.000
freeswap	1.343e-06	1.74e-07	7.695	0.000	1e-06	1.68e-06
tflt	-0.0187	0.000	-80.476	0.000	-0.019	-0.018

Omnibus:	8619.798	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6814500.355
Skew:	-8.995	**Prob(JB):**	0.00
Kurtosis:	170.925	**Cond. No.**	6.62e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.62e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 8 Model 4 Summary

The VIF for Model 4 is

```
lread ---> 1.6072507543377115
lwrite ---> 1.6502831381620369
scall ---> 4.52621349392138
rchar ---> 2.6850149121087816
wchar ---> 2.059880629500184
pgout ---> 1.8677203071374144
pgscan ---> 1.9749928866474429
atch ---> 1.0976615099948441
pgin ---> 2.030401354130604
freemem ---> 2.4683054653300642
freeswap ---> 4.973972430679899
runqsz ---> 2.0871179692928328
tflt ---> 2.9636567791454587
```

After Removing ppgin, VIF of pgin has decreased.

After removing the variables with VIF >5, the condition number is still high. Removing insignificant variables.

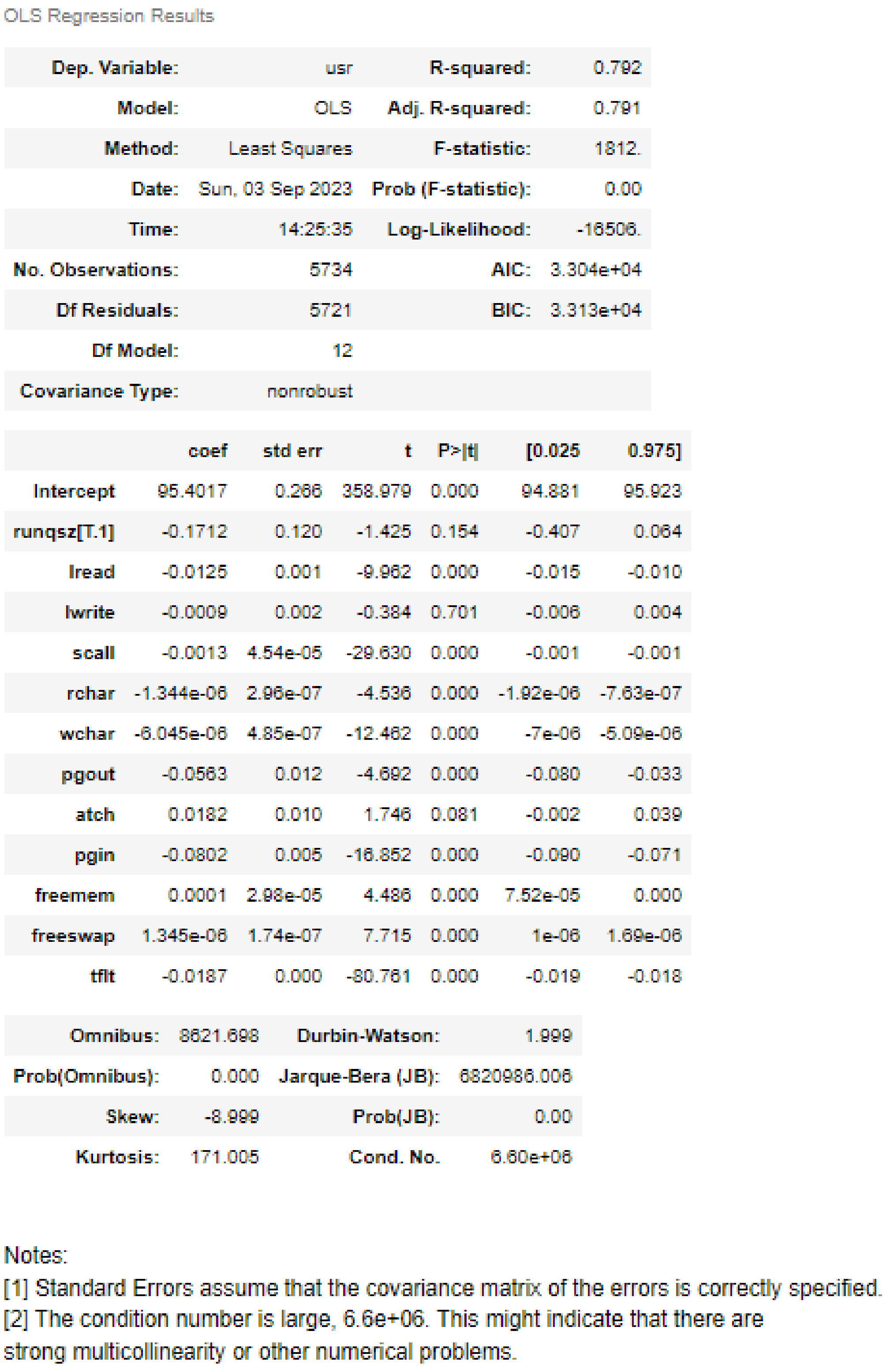
Null hypothesis : Predictor variable is not significant

Alternate hypothesis : Predictor variable is significant

P-values of pgscan is greater than 0.05, it is insignificant.

Model 5:

expr5 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+atch+pgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.791
Method:	Least Squares	F-statistic:	1812.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16506.
No. Observations:	5734	AIC:	3.304e+04
Df Residuals:	5721	BIC:	3.313e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.4017	0.266	358.979	0.000	94.881	95.923
runqsz[T.1]	-0.1712	0.120	-1.425	0.154	-0.407	0.064
lread	-0.0125	0.001	-9.962	0.000	-0.015	-0.010
lwrite	-0.0009	0.002	-0.384	0.701	-0.006	0.004
scall	-0.0013	4.54e-05	-29.630	0.000	-0.001	-0.001
rchar	-1.344e-06	2.96e-07	-4.536	0.000	-1.92e-06	-7.63e-07
wchar	-6.045e-06	4.85e-07	-12.462	0.000	-7e-06	-5.09e-06
pgout	-0.0563	0.012	-4.692	0.000	-0.080	-0.033
atch	0.0182	0.010	1.746	0.081	-0.002	0.039
pgin	-0.0802	0.005	-16.852	0.000	-0.090	-0.071
freemem	0.0001	2.98e-05	4.486	0.000	7.52e-05	0.000
freeswap	1.345e-06	1.74e-07	7.715	0.000	1e-06	1.69e-06
tflt	-0.0187	0.000	-80.761	0.000	-0.019	-0.018

Omnibus:	8621.698	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6820986.006
Skew:	-8.999	Prob(JB):	0.00
Kurtosis:	171.005	Cond. No.	6.60e+06

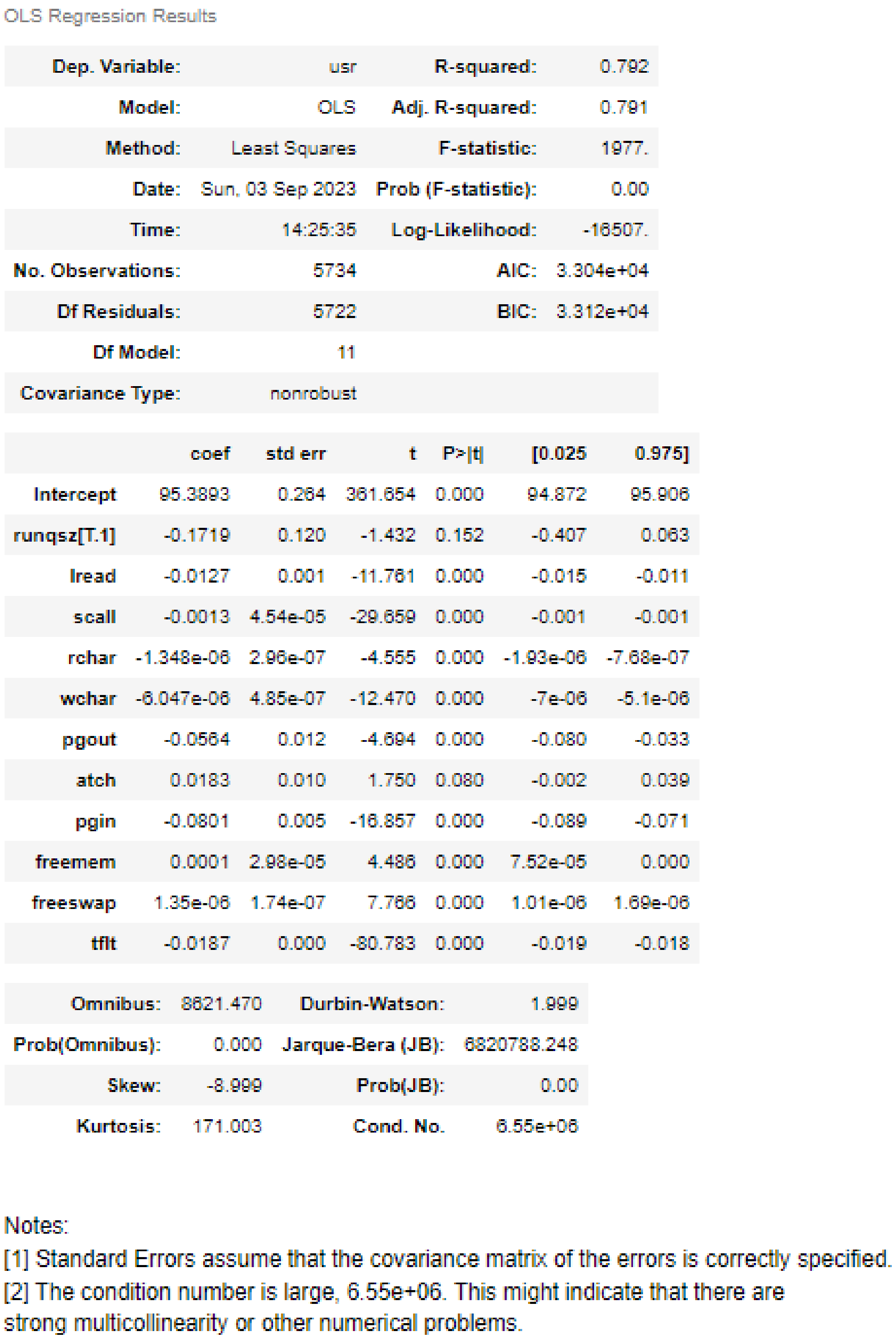
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.6e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 9 Model 5 Summary

Removing lwrite as it is insignificant. P-value is greater than 0.05.

Model 6:

expr6 = 'usr~lread+scall+rchar+wchar+pgout+atch+pgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.791
Method:	Least Squares	F-statistic:	1977.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16507.
No. Observations:	5734	AIC:	3.304e+04
Df Residuals:	5722	BIC:	3.312e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3893	0.264	361.654	0.000	94.872	95.906
runqsz[T.1]	-0.1719	0.120	-1.432	0.152	-0.407	0.063
lread	-0.0127	0.001	-11.761	0.000	-0.015	-0.011
scall	-0.0013	4.54e-05	-29.659	0.000	-0.001	-0.001
rchar	-1.348e-06	2.96e-07	-4.555	0.000	-1.93e-06	-7.68e-07
wchar	-6.047e-06	4.85e-07	-12.470	0.000	-7e-06	-5.1e-06
pgout	-0.0564	0.012	-4.694	0.000	-0.080	-0.033
atch	0.0183	0.010	1.750	0.080	-0.002	0.039
pgin	-0.0801	0.005	-16.857	0.000	-0.089	-0.071
freemem	0.0001	2.98e-05	4.486	0.000	7.52e-05	0.000
freeswap	1.35e-06	1.74e-07	7.766	0.000	1.01e-06	1.69e-06
tflt	-0.0187	0.000	-80.783	0.000	-0.019	-0.018

Omnibus:	8621.470	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6820788.248
Skew:	-8.999	Prob(JB):	0.00
Kurtosis:	171.003	Cond. No.	6.55e+06

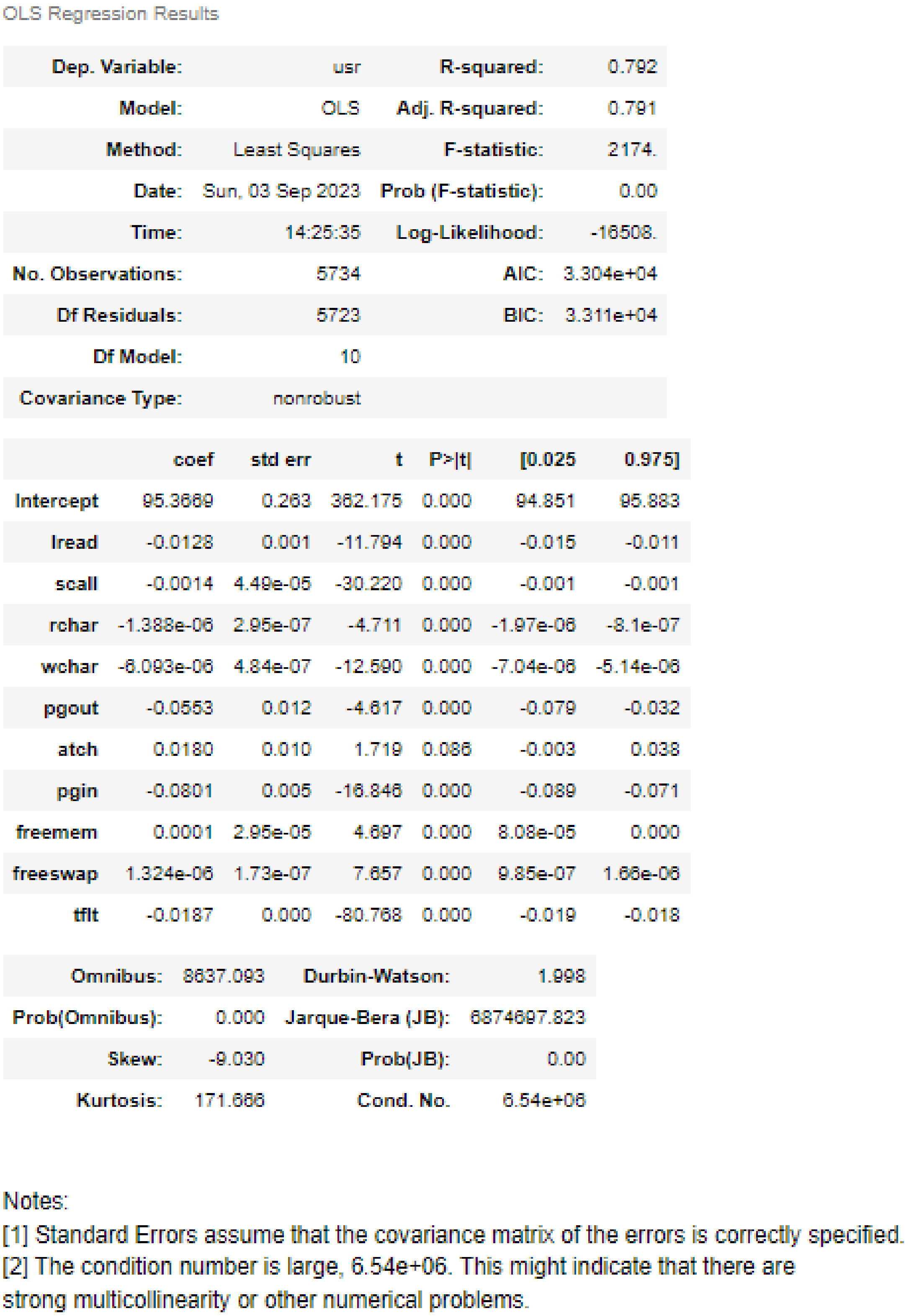
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.55e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 10 Model 6 Summary

Removing runqsz of insignificance.

Model 7:

expr7 = 'usr~lread+scall+rchar+wchar+pgout+atch+pgin+freemem+freeswap+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.791
Method:	Least Squares	F-statistic:	2174.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16508.
No. Observations:	5734	AIC:	3.304e+04
Df Residuals:	5723	BIC:	3.311e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3669	0.263	362.175	0.000	94.851	95.883
lread	-0.0128	0.001	-11.794	0.000	-0.015	-0.011
scall	-0.0014	4.49e-05	-30.220	0.000	-0.001	-0.001
rchar	-1.388e-06	2.95e-07	-4.711	0.000	-1.97e-06	-8.1e-07
wchar	-6.093e-06	4.84e-07	-12.590	0.000	-7.04e-06	-5.14e-06
pgout	-0.0553	0.012	-4.617	0.000	-0.079	-0.032
atch	0.0180	0.010	1.719	0.086	-0.003	0.038
pgin	-0.0801	0.005	-16.846	0.000	-0.089	-0.071
freemem	0.0001	2.95e-05	4.697	0.000	8.08e-05	0.000
freeswap	1.324e-06	1.73e-07	7.657	0.000	9.85e-07	1.66e-06
tflt	-0.0187	0.000	-80.768	0.000	-0.019	-0.018

Omnibus:	8637.093	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6874697.823
Skew:	-9.030	Prob(JB):	0.00
Kurtosis:	171.666	Cond. No.	6.54e+06

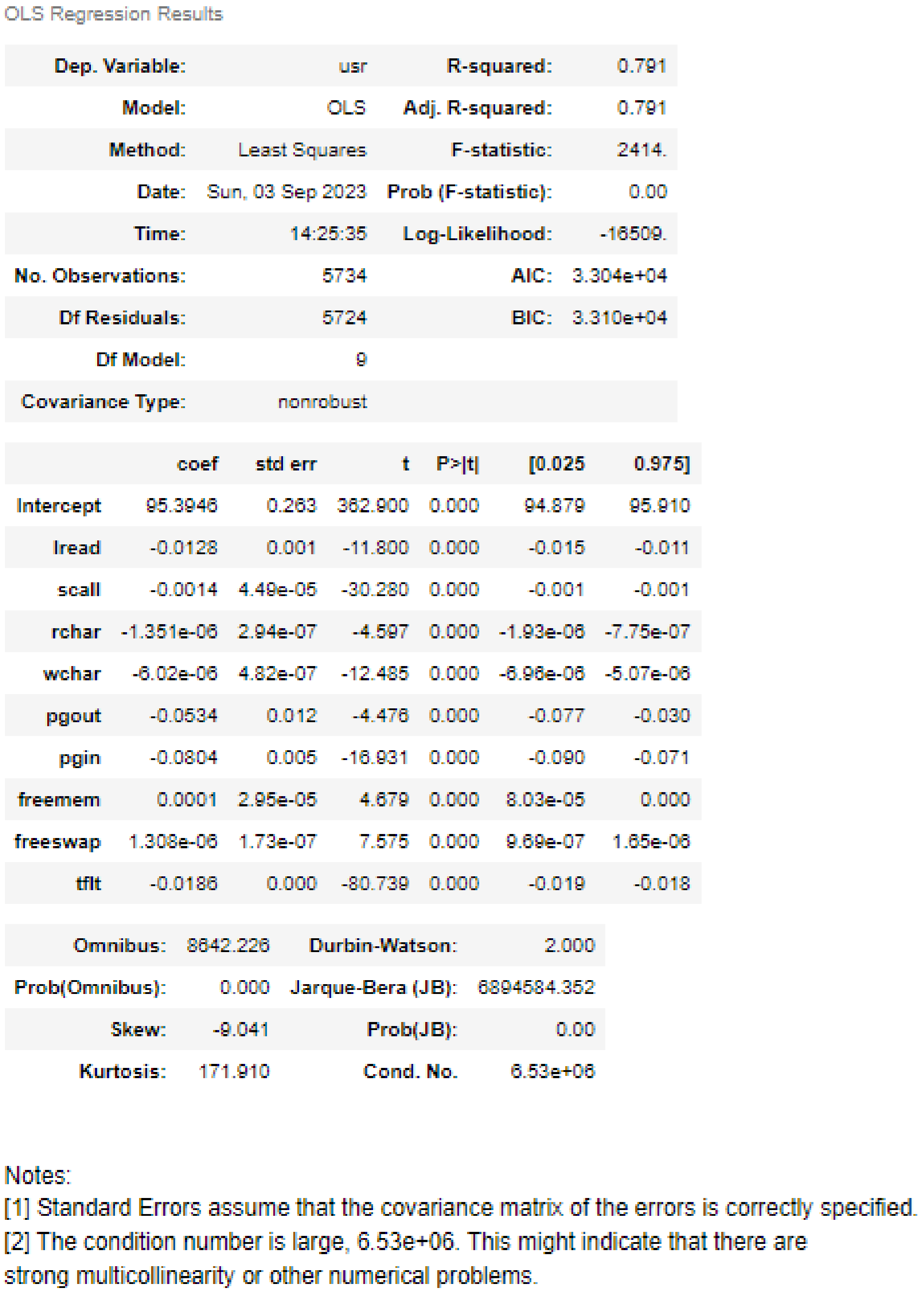
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.54e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Table 11 Model 7 Summary

Removing atch as its p-value is greater than 0.05

Model 8:

expr8 = 'usr~lread+scall+rchar+wchar+pgout+pgin+freemem+freeswap+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.791
Model:	OLS	Adj. R-squared:	0.791
Method:	Least Squares	F-statistic:	2414.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:35	Log-Likelihood:	-16509.
No. Observations:	5734	AIC:	3.304e+04
Df Residuals:	5724	BIC:	3.310e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	95.3946	0.263	362.900	0.000	94.879	95.910
lread	-0.0128	0.001	-11.800	0.000	-0.015	-0.011
scall	-0.0014	4.49e-05	-30.280	0.000	-0.001	-0.001
rchar	-1.351e-06	2.94e-07	-4.597	0.000	-1.93e-06	-7.75e-07
wchar	-6.02e-06	4.82e-07	-12.485	0.000	-6.96e-06	-5.07e-06
pgout	-0.0534	0.012	-4.476	0.000	-0.077	-0.030
pgin	-0.0804	0.005	-16.931	0.000	-0.090	-0.071
freemem	0.0001	2.95e-05	4.679	0.000	8.03e-05	0.000
freeswap	1.308e-06	1.73e-07	7.575	0.000	9.69e-07	1.65e-06
tflt	-0.0186	0.000	-80.739	0.000	-0.019	-0.018

Omnibus:	8842.226	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6894584.352
Skew:	-9.041	Prob(JB):	0.00
Kurtosis:	171.910	Cond. No.	6.53e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.53e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 12 Model 8 Summary

After removing all the insignificant variables also, the condition number is high. The same steps of removing variables VIF>5 and insignificant has been repeated after cleaning the outliers for the dataset.

Outlier’s Treatment:

The outliers are treated using the IQR lower and upper range for all variables. The pgscan has become zero after treatment. It wont have any significance in the dataset.

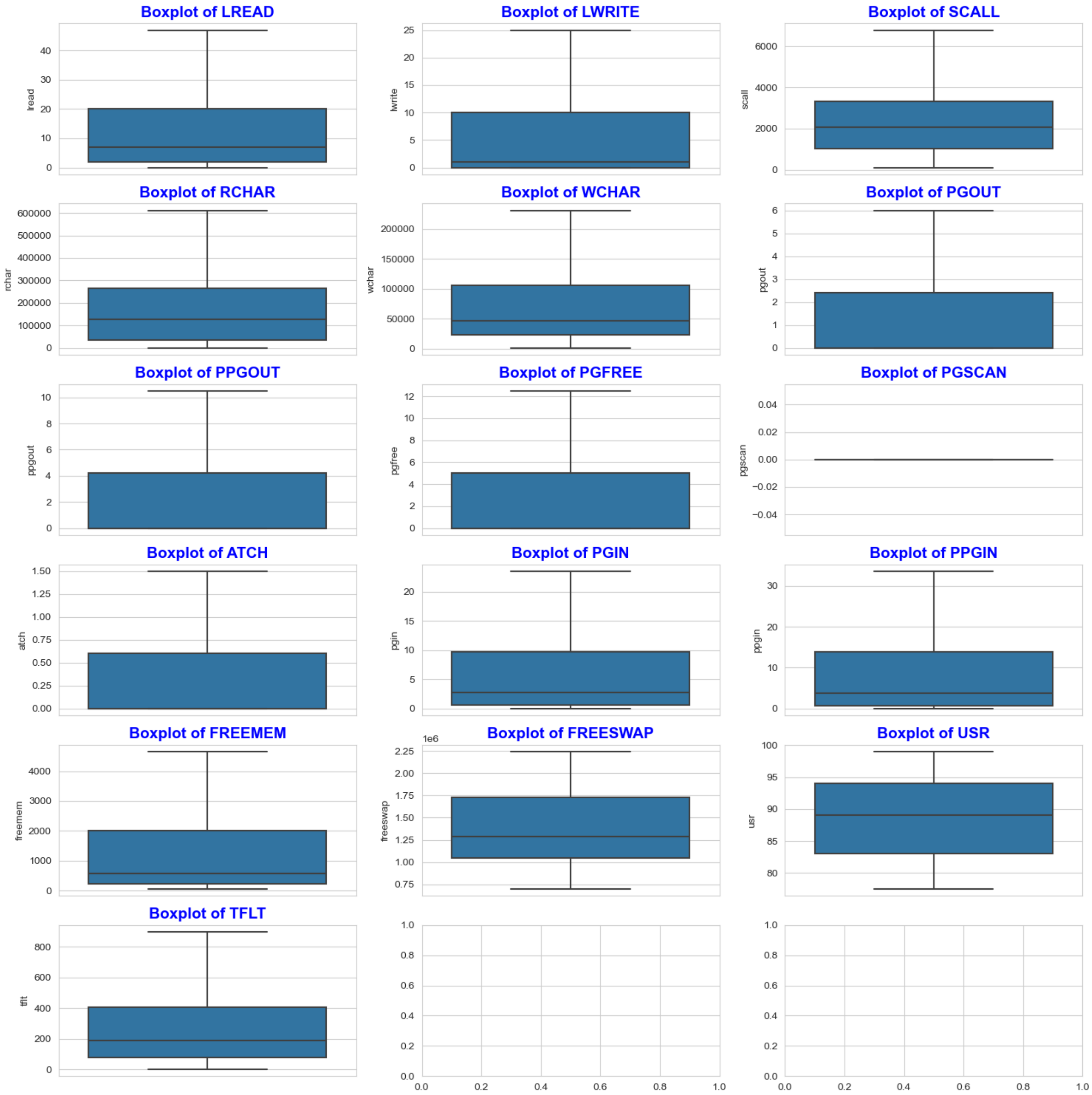


Figure M Boxplot (After outlier's treatment)

The model 9 to model 16 has the same steps as above models. Refer jupyter notebook for detailed summary of regression model.

Model 9:

expr = 'usr~lread+lwrite+scall+rchar+wchar+pgout+ppgout+pgfree+pgscan+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'

Model 10:

expr2 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+ppgout+pgfree+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'

Model 11:

expr3 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'

Model 12:

expr4 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'

Model 13:

expr5 = 'usr~lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'

Model 14:

expr6 = 'usr~lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'

Model 15:

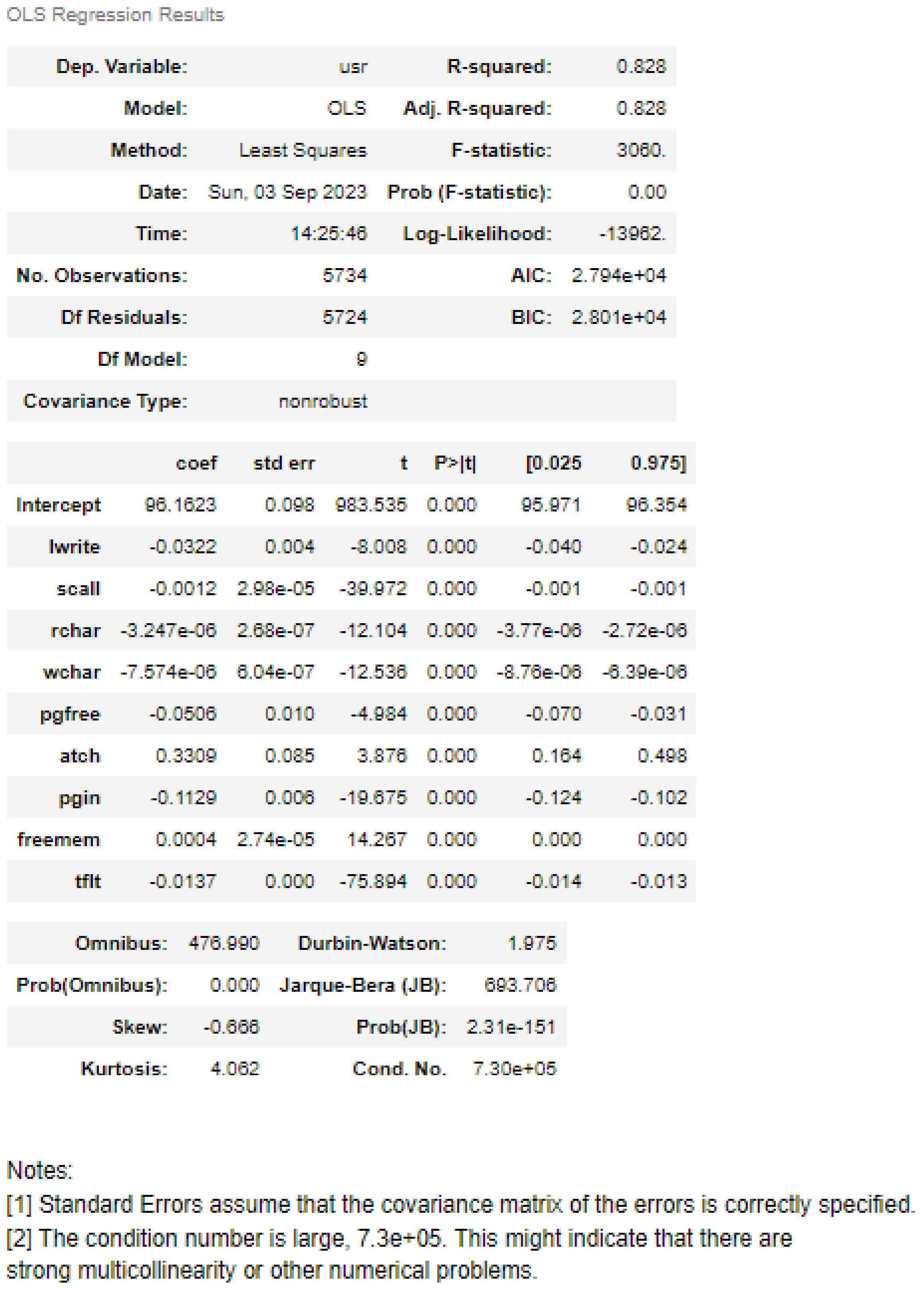
expr7 = 'usr~lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+runqsz+tflt'

Model 16:

expr8 = 'usr~lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+tflt'

The VIF for Model 16 is

```
lwrite ---> 1.5277087360677577
scall ---> 4.198293334168806
rchar ---> 3.4062842474391917
wchar ---> 2.8792222621418158
pgfree ---> 2.6267974367647313
atch ---> 2.4617757644858247
pgin ---> 2.3996498134201634
freemem ---> 1.3809322818970744
tflt ---> 3.5440971138222985
```



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.828
Model:	OLS	Adj. R-squared:	0.828
Method:	Least Squares	F-statistic:	3060.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:46	Log-Likelihood:	-13962.
No. Observations:	5734	AIC:	2.794e+04
Df Residuals:	5724	BIC:	2.801e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	96.1623	0.098	983.535	0.000	95.971	96.354
lwrite	-0.0322	0.004	-8.008	0.000	-0.040	-0.024
scall	-0.0012	2.98e-05	-39.972	0.000	-0.001	-0.001
rchar	-3.247e-06	2.68e-07	-12.104	0.000	-3.77e-06	-2.72e-06
wchar	-7.574e-06	6.04e-07	-12.536	0.000	-8.76e-06	-6.39e-06
pgfree	-0.0506	0.010	-4.984	0.000	-0.070	-0.031
atch	0.3309	0.085	3.876	0.000	0.164	0.498
pgin	-0.1129	0.006	-19.675	0.000	-0.124	-0.102
freemem	0.0004	2.74e-05	14.267	0.000	0.000	0.000
tflt	-0.0137	0.000	-75.894	0.000	-0.014	-0.013

Omnibus:	476.990	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	693.706
Skew:	-0.666	Prob(JB):	2.31e-151
Kurtosis:	4.062	Cond. No.	7.30e+05

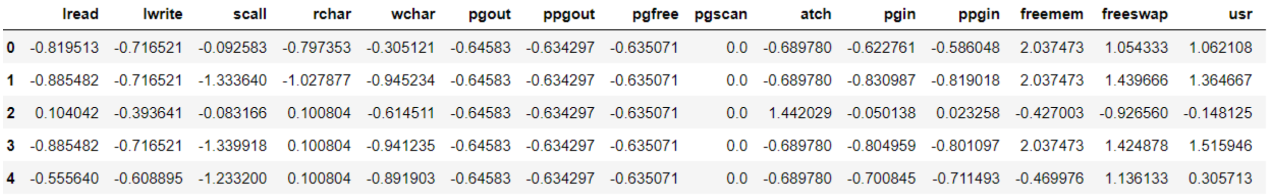
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.3e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 13 Model 16 Summary

After repeating the steps for removing multicollinearity variables and insignificant variables, the condition number is still high. The dataset has been scaled and the process is repeated.

Scaling:

The dataset has been scaled using the standard scaler.

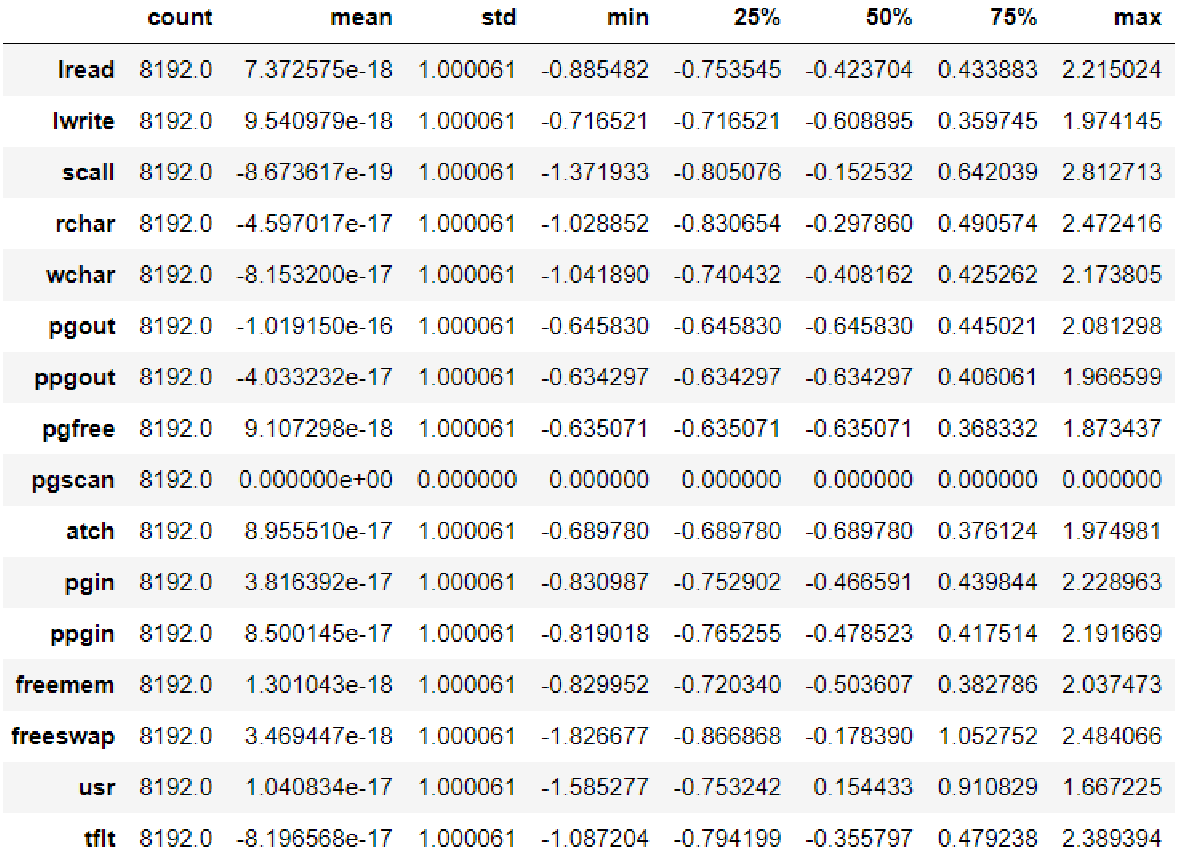


	lread	lwrite	scall	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	freemem	freeswap	usr
0	-0.819513	-0.716521	-0.092583	-0.797353	-0.305121	-0.64583	-0.634297	-0.635071	0.0	-0.689780	-0.622761	-0.586048	2.037473	1.054333	1.062108
1	-0.885482	-0.716521	-1.333640	-1.027877	-0.945234	-0.64583	-0.634297	-0.635071	0.0	-0.689780	-0.830987	-0.819018	2.037473	1.439666	1.364667
2	0.104042	-0.393641	-0.083166	0.100804	-0.614511	-0.64583	-0.634297	-0.635071	0.0	1.442029	-0.050138	0.023258	-0.427003	-0.926560	-0.148125
3	-0.885482	-0.716521	-1.339918	0.100804	-0.941235	-0.64583	-0.634297	-0.635071	0.0	-0.689780	-0.804959	-0.801097	2.037473	1.424878	1.515946
4	-0.555640	-0.608895	-1.233200	0.100804	-0.891903	-0.64583	-0.634297	-0.635071	0.0	-0.689780	-0.700845	-0.711493	-0.469976	1.136133	0.305713

Table 14 Scaled Dataset

The categorical variable is not scaled. All the numerical variables are standardised.

The below table confirms the data has been scaled with mean 0 and standard deviation of 1.

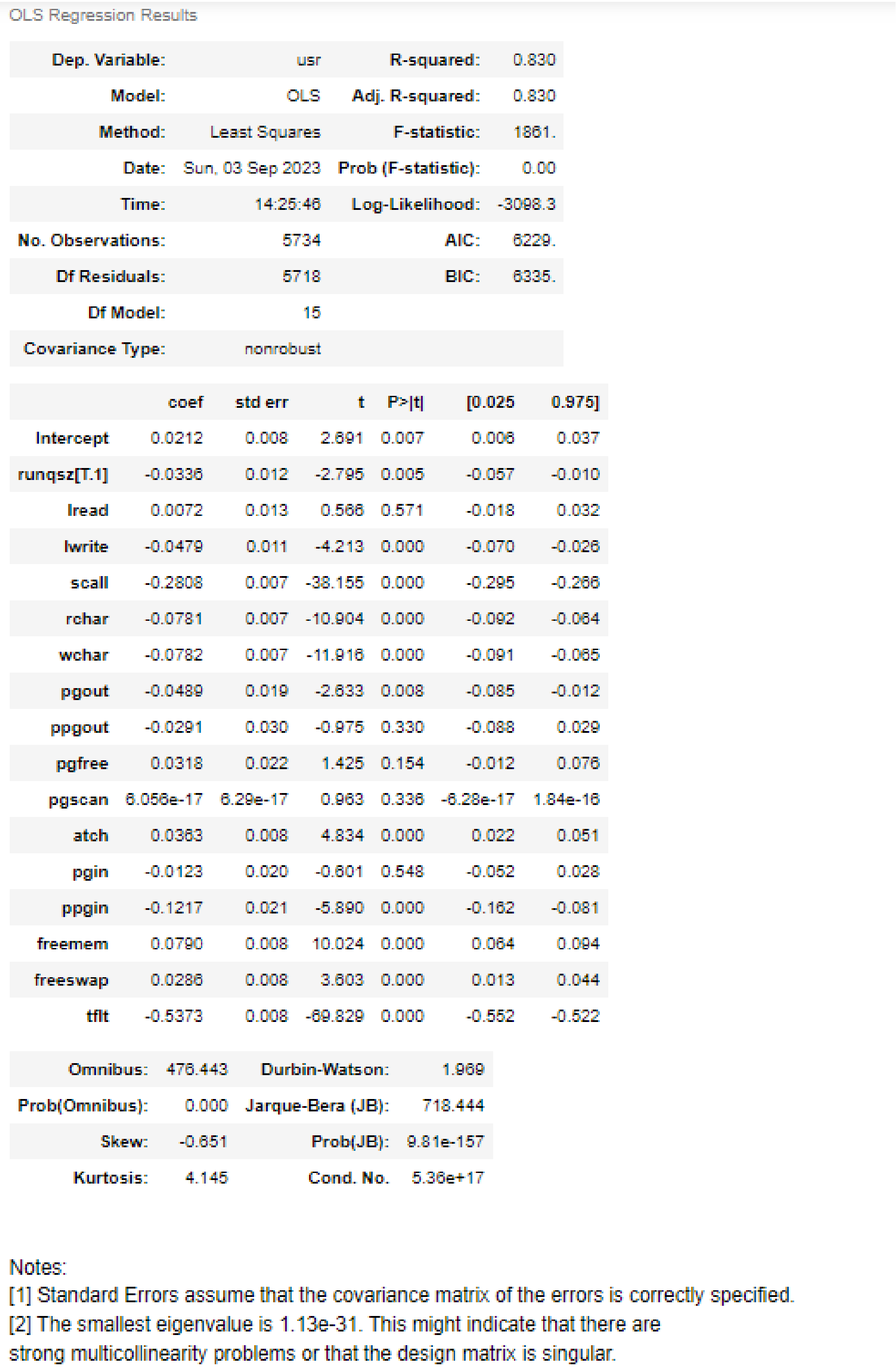


	count	mean	std	min	25%	50%	75%	max
lread	8192.0	7.372575e-18	1.000061	-0.885482	-0.753545	-0.423704	0.433883	2.215024
lwrite	8192.0	9.540979e-18	1.000061	-0.716521	-0.716521	-0.608895	0.359745	1.974145
scall	8192.0	-8.673617e-19	1.000061	-1.371933	-0.805076	-0.152532	0.642039	2.812713
rchar	8192.0	-4.597017e-17	1.000061	-1.028852	-0.830654	-0.297860	0.490574	2.472416
wchar	8192.0	-8.153200e-17	1.000061	-1.041890	-0.740432	-0.408162	0.425262	2.173805
pgout	8192.0	-1.019150e-16	1.000061	-0.645830	-0.645830	-0.645830	0.445021	2.081298
ppgout	8192.0	-4.033232e-17	1.000061	-0.634297	-0.634297	-0.634297	0.406061	1.966599
pgfree	8192.0	9.107298e-18	1.000061	-0.635071	-0.635071	-0.635071	0.368332	1.873437
pgscan	8192.0	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
atch	8192.0	8.955510e-17	1.000061	-0.689780	-0.689780	-0.689780	0.376124	1.974981
pgin	8192.0	3.816392e-17	1.000061	-0.830987	-0.752902	-0.466591	0.439844	2.228963
ppgin	8192.0	8.500145e-17	1.000061	-0.819018	-0.765255	-0.478523	0.417514	2.191669
freemem	8192.0	1.301043e-18	1.000061	-0.829952	-0.720340	-0.503607	0.382786	2.037473
freeswap	8192.0	3.469447e-18	1.000061	-1.826677	-0.866868	-0.178390	1.052752	2.484066
usr	8192.0	1.040834e-17	1.000061	-1.585277	-0.753242	0.154433	0.910829	1.667225
tflt	8192.0	-8.196568e-17	1.000061	-1.087204	-0.794199	-0.355797	0.479238	2.389394

Table 15 Five part summary for Scaled data

Model 17:

expr = 'usr~lread+lwrite+scall+rchar+wchar+pgout+ppgout+pgfree+pgscan+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.830
Method:	Least Squares	**F-statistic:**	1861.
Date:	Sun, 03 Sep 2023	**Prob (F-statistic):**	0.00
Time:	14:25:46	**Log-Likelihood:**	-3098.3
No. Observations:	5734	**AIC:**	6229.
Df Residuals:	5718	**BIC:**	6335.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.0212	0.008	2.691	0.007	0.006	0.037
runqsz[T.1]	-0.0336	0.012	-2.795	0.005	-0.057	-0.010
lread	0.0072	0.013	0.566	0.571	-0.018	0.032
lwrite	-0.0479	0.011	-4.213	0.000	-0.070	-0.026
scall	-0.2808	0.007	-38.155	0.000	-0.295	-0.266
rchar	-0.0781	0.007	-10.904	0.000	-0.092	-0.064
wchar	-0.0782	0.007	-11.916	0.000	-0.091	-0.065
pgout	-0.0489	0.019	-2.633	0.008	-0.085	-0.012
ppgout	-0.0291	0.030	-0.975	0.330	-0.088	0.029
pgfree	0.0318	0.022	1.425	0.154	-0.012	0.076
pgscan	6.056e-17	6.29e-17	0.963	0.336	-6.28e-17	1.84e-16
atch	0.0363	0.008	4.834	0.000	0.022	0.051
pgin	-0.0123	0.020	-0.601	0.548	-0.052	0.028
ppgin	-0.1217	0.021	-5.890	0.000	-0.162	-0.081
freemem	0.0790	0.008	10.024	0.000	0.064	0.094
freeswap	0.0286	0.008	3.603	0.000	0.013	0.044
tflt	-0.5373	0.008	-69.829	0.000	-0.552	-0.522

Omnibus:	476.443	**Durbin-Watson:**	1.969
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	718.444
Skew:	-0.651	**Prob(JB):**	9.81e-157
Kurtosis:	4.145	**Cond. No.**	5.36e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.13e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Table 16 Model 17 Summary

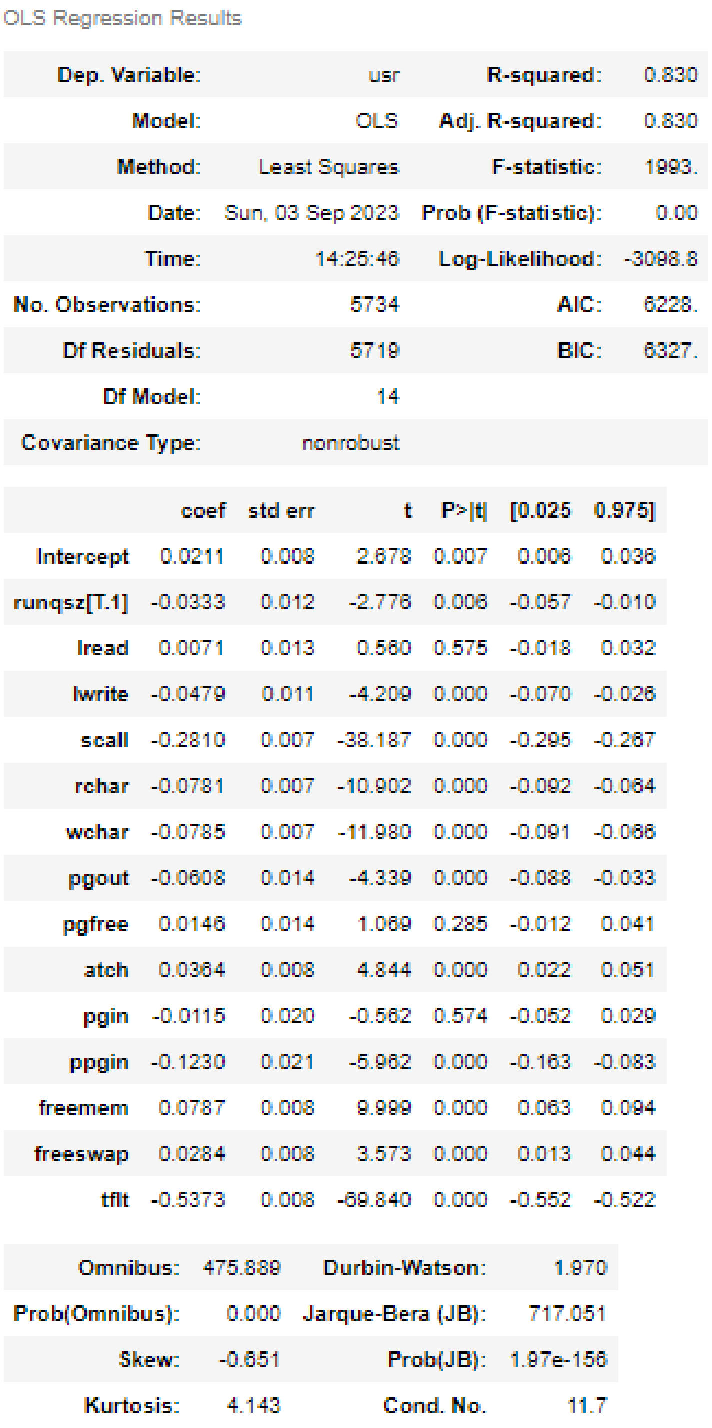
The VIF for Model 17 is

```
lread ---> 5.238811078987626
lwrite ---> 4.257513897235675
scall ---> 1.7908965731441515
rchar ---> 1.7023873961603637
wchar ---> 1.413056392631436
pgout ---> 11.320598700591798
ppgout ---> 29.35699339283489
pgfree ---> 16.477647289915833
pgscan ---> nan
atch ---> 1.8577244823340298
pgin ---> 13.716974279527667
ppgin ---> 13.944763096093599
freemem ---> 2.0585352824635015
freeswap ---> 2.0518495581993474
tflt ---> 1.9915155086862288
```

Removing pgscan as it is insignificant after outlier’s treatment.

Model 18:

expr3 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.830
Model:	OLS	Adj. R-squared:	0.830
Method:	Least Squares	F-statistic:	1993.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:46	Log-Likelihood:	-3098.8
No. Observations:	5734	AIC:	6228.
Df Residuals:	5719	BIC:	6327.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0211	0.008	2.678	0.007	0.006	0.036
runqsz[T.1]	-0.0333	0.012	-2.776	0.006	-0.057	-0.010
lread	0.0071	0.013	0.560	0.575	-0.018	0.032
lwrite	-0.0479	0.011	-4.209	0.000	-0.070	-0.026
scall	-0.2810	0.007	-38.187	0.000	-0.295	-0.267
rchar	-0.0781	0.007	-10.902	0.000	-0.092	-0.064
wchar	-0.0785	0.007	-11.980	0.000	-0.091	-0.066
pgout	-0.0608	0.014	-4.339	0.000	-0.088	-0.033
pgfree	0.0146	0.014	1.069	0.285	-0.012	0.041
atch	0.0364	0.008	4.844	0.000	0.022	0.051
pgin	-0.0115	0.020	-0.562	0.574	-0.052	0.029
ppgin	-0.1230	0.021	-5.962	0.000	-0.163	-0.083
freemem	0.0787	0.008	9.999	0.000	0.063	0.094
freeswap	0.0284	0.008	3.573	0.000	0.013	0.044
tflt	-0.5373	0.008	-69.840	0.000	-0.552	-0.522

Omnibus:	475.889	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	717.051
Skew:	-0.651	Prob(JB):	1.97e-156
Kurtosis:	4.143	Cond. No.	11.7

Table 17 Model 18 Summary

The condition number is drastically decreased. The model R-square is 83%. There are some insignificant variables and VIF>5.

The VIF for Model 18 is

```
lread ---> 5.238811078987626
lwrite ---> 4.257513897235675
scall ---> 1.790896573144152
rchar ---> 1.7023873961603637
wchar ---> 1.413056392631436
pgout ---> 11.320598700591798
ppgout ---> 29.35699339283489
pgfree ---> 16.477647289915833
atch ---> 1.8577244823340298
pgin ---> 13.716974279527667
ppgin ---> 13.944763096093599
freemem ---> 2.0585352824635015
freeswap ---> 2.0518495581993474
tflt ---> 1.9915155086862288
runqsz ---> 1.0961635239623044
```

After removing the variables with VIF>5. The final model is obtained.

Model 19:

expr3 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+ppgin+freemem+freeswap+runqsz+tflt'

Model 20:

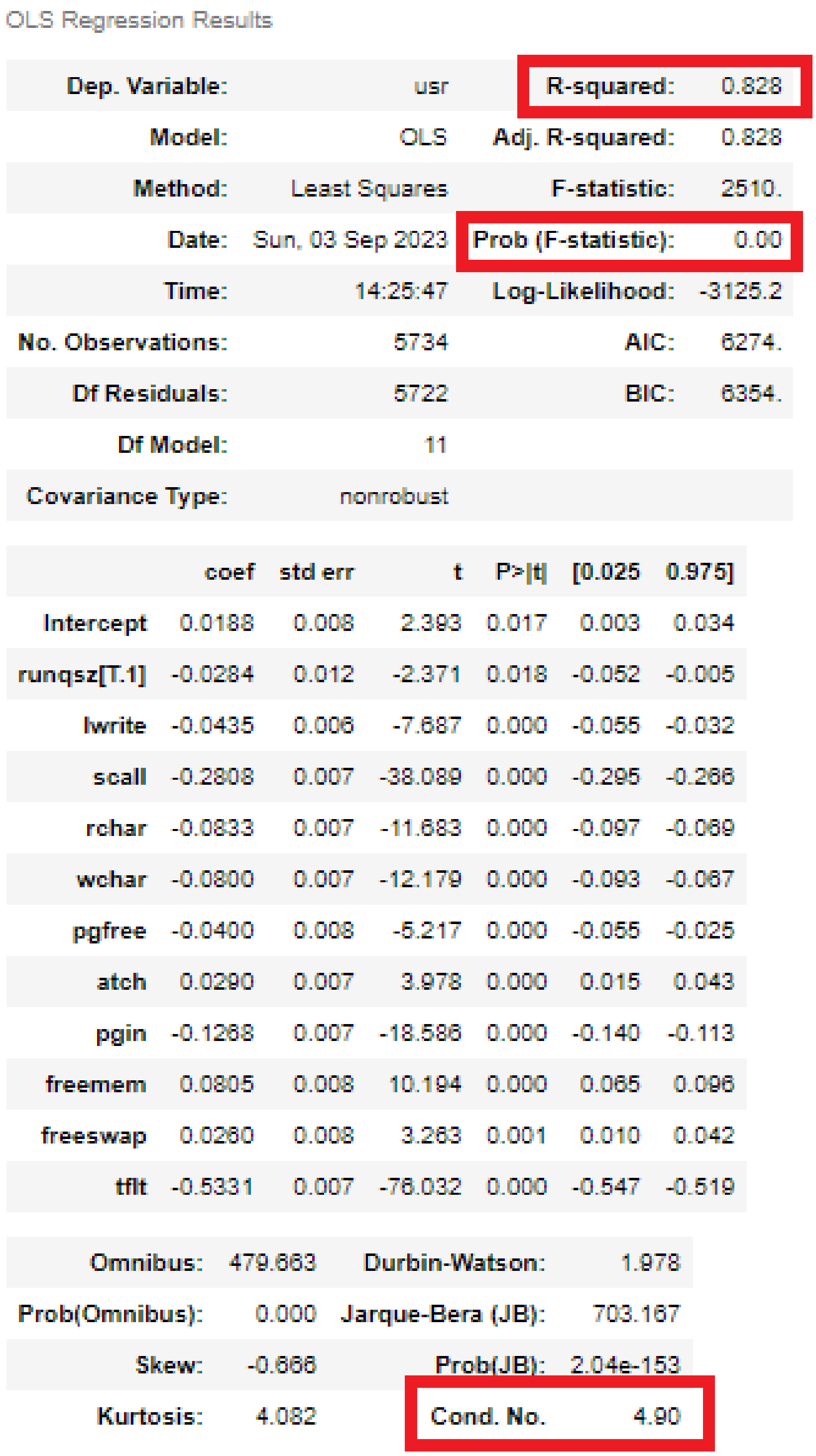
expr4 = 'usr~lread+lwrite+scall+rchar+wchar+pgout+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'

Model 21:

expr5 = 'usr~lread+lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'

Model 22 (Final Model):

expr6 = 'usr~lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.828
Model:	OLS	Adj. R-squared:	0.828
Method:	Least Squares	F-statistic:	2510.
Date:	Sun, 03 Sep 2023	Prob (F-statistic):	0.00
Time:	14:25:47	Log-Likelihood:	-3125.2
No. Observations:	5734	AIC:	6274.
Df Residuals:	5722	BIC:	6354.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0188	0.008	2.393	0.017	0.003	0.034
runqsz[T.1]	-0.0284	0.012	-2.371	0.018	-0.052	-0.005
lwrite	-0.0435	0.006	-7.687	0.000	-0.055	-0.032
scall	-0.2808	0.007	-38.089	0.000	-0.295	-0.266
rchar	-0.0833	0.007	-11.683	0.000	-0.097	-0.069
wchar	-0.0800	0.007	-12.179	0.000	-0.093	-0.067
pgfree	-0.0400	0.008	-5.217	0.000	-0.055	-0.025
atch	0.0290	0.007	3.978	0.000	0.015	0.043
pgin	-0.1268	0.007	-18.586	0.000	-0.140	-0.113
freemem	0.0805	0.008	10.194	0.000	0.065	0.096
freeswap	0.0260	0.008	3.263	0.001	0.010	0.042
tflt	-0.5331	0.007	-76.032	0.000	-0.547	-0.519

Omnibus:	479.663	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	703.167
Skew:	-0.666	Prob(JB):	2.04e-153
Kurtosis:	4.082	Cond. No.	4.90

Table 18 Model 22 Summary

The VIF for Final model is

```
lwrite ---> 1.048736618408507
scall ---> 1.7798803448653988
rchar ---> 1.6689470660454242
wchar ---> 1.4049055486474127
pgfree ---> 1.918674681379076
atch ---> 1.7277907462942845
pgin ---> 1.5107973541989992
freemem ---> 2.0496547838522545
freeswap ---> 2.0441506516662704
tflt ---> 1.6421425183590337
runqsz ---> 1.089444587133308
```

All the VIF are less than 5. There is no multicollinearity.

The model strength is 82% (R-Squared).

The hypotheses of the F test are as follows:

H0: β1=β2=…=βk=0

H1: At least one of β1,β2,…,βk≠0

Prob (F-statistic) is less than 0.05. Rejects the null Hypothesis.

All the predictor variables are significant.

The Final Equation for prediction of usr is

For Runqsz[CPU_Bound]

usr~0.018821-0.028395+(-0.043525)lwrite+(-0.280847)scall+(-0.083251)rchar+(-0.079990)wchar+(-0.039984)pgfree+(0.028972)atch+(-0.126763)pgin+(0.080485)freemem+(0.025975)freeswap+(-0.533106)tflt

For Runqsz[NOT_CPU_Bound]

usr~0.018821+(-0.043525)lwrite+(-0.280847)scall+(-0.083251)rchar+(-0.079990)wchar+(-0.039984)pgfree+(0.028972)atch+(-0.126763)pgin+(0.080485)freemem+(0.025975)freeswap+(-0.533106)tflt

If the runqsz is cpu bound then it has an additional change in its intercept, it is not, much difference. The same has been explored in EDA when the scatter plot doesn't show any clear differences between the two categories.

The correlation and coefficient are in the same direction. Only freeswap and freemem has positive coefficient.

Assumptions of Linear Regression:

Linearity:

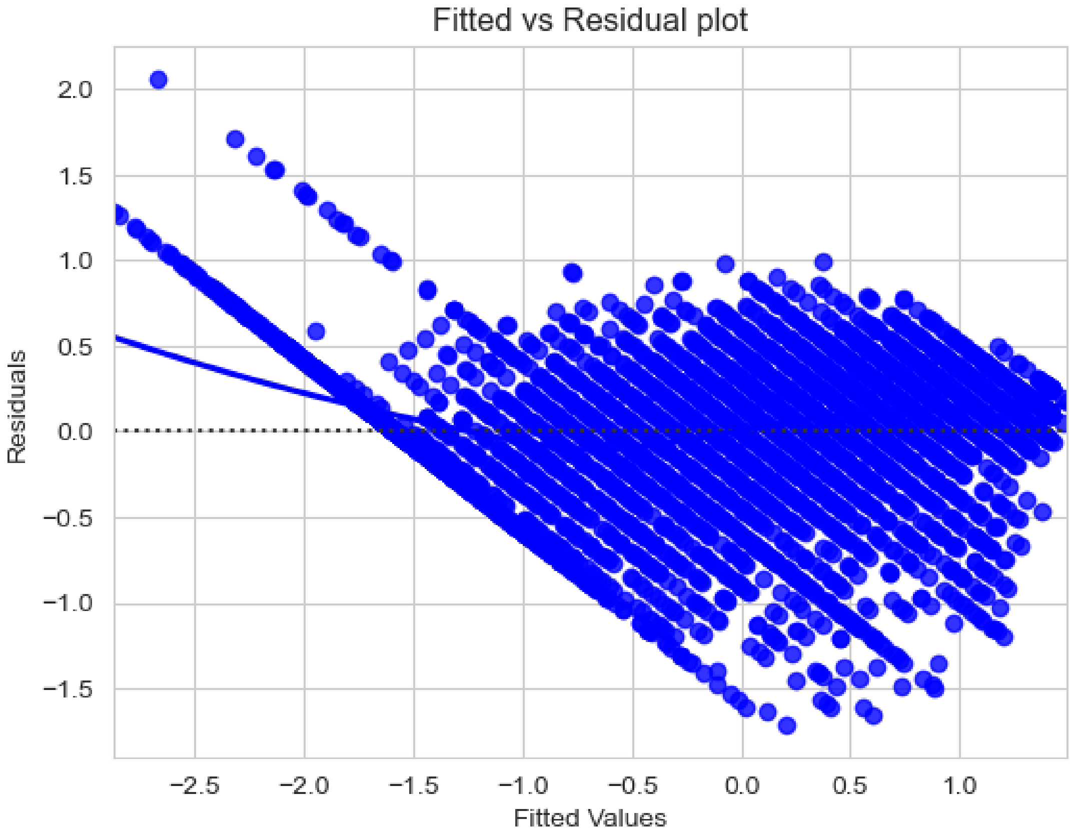


Figure N Fitted Vs Residual Plot

The Model is showing signs of non-linearity. The Regression plot in EDA showed all the variables has either upward or downward slope with linear relationship.

Normality:

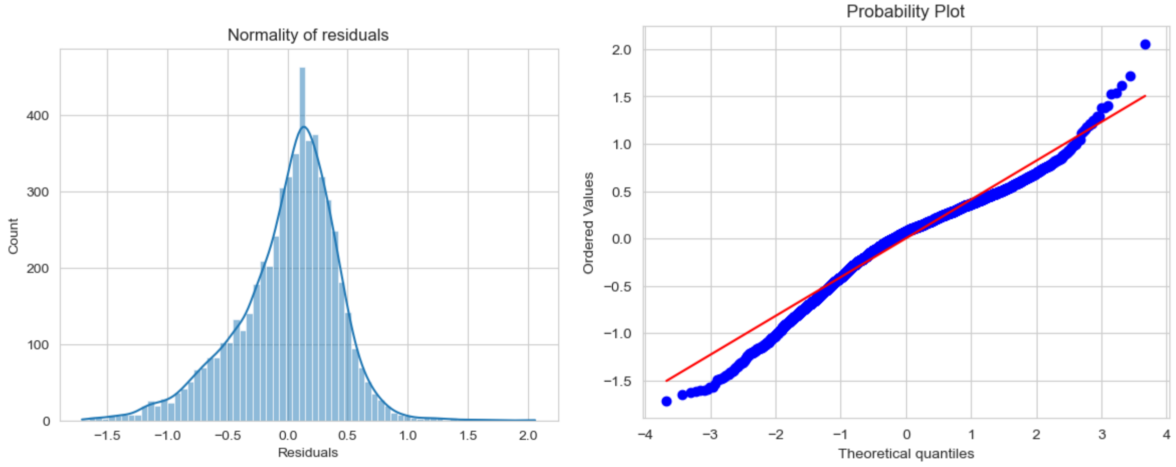


Figure O Normality and Q-Q plot for residuals

Shapiro Test:

Null hypothesis - Data is normally distributed.

Alternate hypothesis - Data is not normally distributed.

P-value is less than 0.05. Reject null hypothesis. The residuals are not normally distributed.

However, as an approximation from above plots, it can be accepted this distribution as close to being normal.

Homoscedasticity:

The null and alternate hypotheses of the goldfeldquandt test are as follows:

Null hypothesis : Residuals are homoscedastic

Alternate hypothesis : Residuals have hetroscedasticity

P-value is greater than 0.05. Do not reject null hypothesis.

Residuals are homoscedastic.

Train and Test comparision:

```
RMSE for Train - 0.4173182011920647
RMSE for Test - 0.43285977312490803
```

Lower the RMSE, better the model. The final model has low RMSE for both train and test.

1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

Step 1: Performing Descriptive Statistics and EDA.

From this, the insights from EDA are

- usr is positively correlated with freeswap and freemem. Negatively correlated with others.
- usr and freeswap are positively skewed.
- The outliers in usr is in the left side and 75% of the data takes more than 81% of the time.
- There is no difference with respect to CPU_Bound and Not_CPU_Bound.

Step 2: Null values and Duplicates treatment.

- The Null values are imputed with mean.
- There is no duplicates for this dataset.

Step 3: New feature Addition**.**

- vflt and pflt are combined to create tflt.
- Dropped variables which are covered in scall.

Step 4: Data Split 70:30

- The Data has been split into Train and Test.
- The Training Data is used for training the model and the test is used for validation.

Step 5: Building a Model

- The Train data is fitted into the model but the condition number is high.
- After removing the variables with multicollinearity and insignificant variables still it is huge.
- To remove any abnormality in data, the outliers are removed.

Step 6: Building a Model for Outlier treated data

- Step 5 is repeated for the new cleaned dataset.
- Still the condition number is high. The Data has been scaled to remove other numerical problems.

Step 7: Building a Model for Scaled data

- The condition number has been decreased and in acceptable range.
- The Model Strength is 83% and it has only the significant variables.
- Final Model - 'usr~lwrite+scall+rchar+wchar+pgfree+atch+pgin+freemem+freeswap+runqsz+tflt'.

Step 8: Checking the Assumptions of Linear Regression.

- Linearity – The residuals are non-linear.
- Normality – The residuals are close enough normally distributed.
- Multicollinearity – There is no multicollinearity exists. VIF<5 for all variables.
- Homoscedasticity – The residuals are homoscedastic.

Step 9: Performance of the Model.

- The RMSE is low for both Train and Test. The Model performs well with test data.

Step 10: Business Insights**.**

- Based on the model, the usr is positively correlated with freeswap and freemem.
- freemem - Number of memory pages available to user processes
  freeswap - Number of disk blocks available for page swapping.
- If these two are decreased, the usr will also decrease.
- If the performance of other variables increases i.e. if the other variables process more per second, it will also decrease the user time.
- Process run queue size doesn’t have any significant impact with the two category.