APPLICATION OF MULTIPLE REGRESSION ANALYSIS OF BASE BALL GAME

TESTING FOR WHICH VARIABLE CORRELATE WITH RUN_SCORE

PERFORM MULTIPLE REGRESSION BETWEEN RUNS-SCORE IN A BASE BALL GAME AND OTHER EXPLANATORY VARIABLES AS EXPLANINED IN THIS REPORT

TAIWO FAMUYIWA
T00589082
ADVANCE MATHS PROJECT

INTRODUCTION

- MOTIVATION

The main motive behind this report is to analyze and predict the RUN-SCORE of 30 MLB teams of a base game within certain period using explanatory variables. Additionally, I want to know which explanatory variable has strong correlation with the RUN-SCORE variable and how relevant is this variable to the prediction of the response variable. Also, other statistical test will be carried out to validate the result.

- EXPLANATION OF THE PROJECT

This project consists of data from baseball games of 30 MLB teams. One of the tests is to predict which variable correlate strongly with the response variable. The data shows the runs score bats, batting-average etc. To start with, descriptive statistic like DATA COLLECTION, EXPLAINATION OF VARIABLES will be explained etc. Followed by STATISTICAL ANALYSIS such as statistical test, analysis etc. The project will be rounded up with a result summarizing all the analytical work.

DESCRIPTIVE STATISTICS

- DATA COLLECTION

Data was collected from www.statcrunch.com. The data consist of RUN-SCORE as a response variable (dependent variable) and four independent variable; BATTING-AVG, OBP, SLG, OPS. Each variable will be discussed in the subsequent heading.

- EXPLANATION OF A VARIABLE

As mentioned above, the following are the variables used in this project with their respective meaning:

  - RUNS-SCORE
    This is the total number of all runs the baseball team scored by the end of the season (dependent variable) $represent\ as\ y$
  - BATTING-AVG
    This is equal to the number of hits divided by bats $represent\ as\ x_1$
  - OBP
    On base percentage $represent\ as\ x_2$
  - SLG (SLUGGING)
    Weights hits to first base as 1 point $represent\ as\ x_3$
  - OPS
    On base plus slugging. $represent\ as\ x_4$

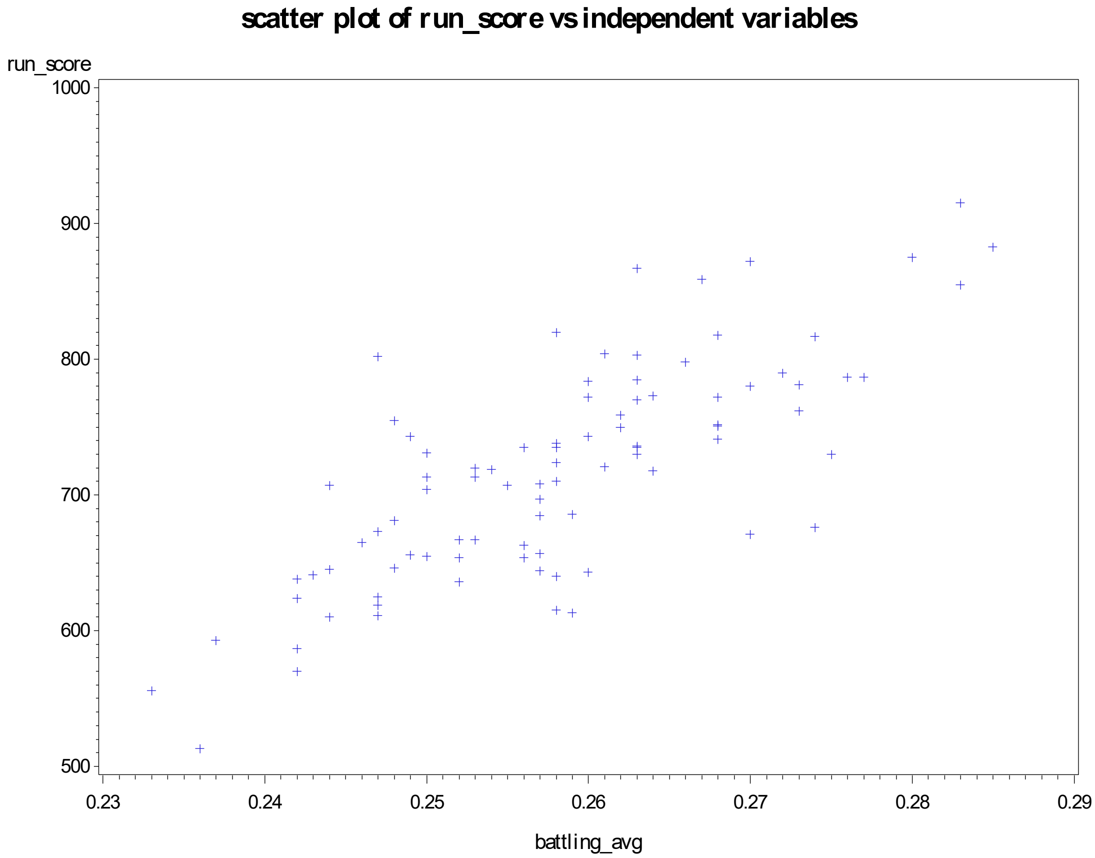
- DISCRIPTIVE STATISTICS

  - GETTING THE MEAN, NUMBER, MINIMUM AND MAXIMUM VALUE FOR EACH VARIABLE
    Conducting a descriptive test on each variable, starting with total number for each variable using proc mean.

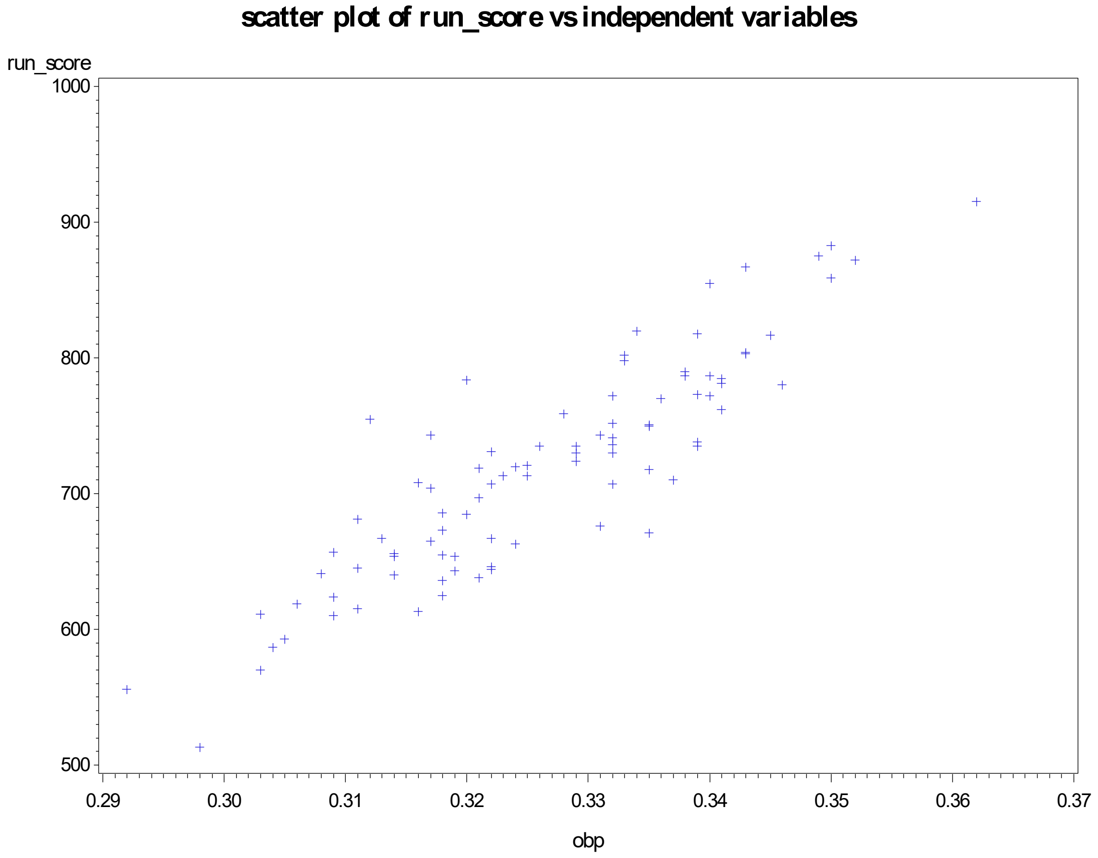
Variable	N	Mean	Median	Minimum	Maximum
run_score	90	717.06	719.50	513.00	915.00
battling_avg	90	0.26	0.26	0.23	0.29
obp	90	0.33	0.33	0.29	0.36
slg	90	0.41	0.41	0.34	0.48
ops	90	0.73	0.73	0.64	0.84

From the data above, it is shown that all explanatory variables have equal mean and median except for the response variable whose mean differs from the median.
This means that the distribution of each explanatory variable is symmetrical and that their respective distribution will have a zero skewness.

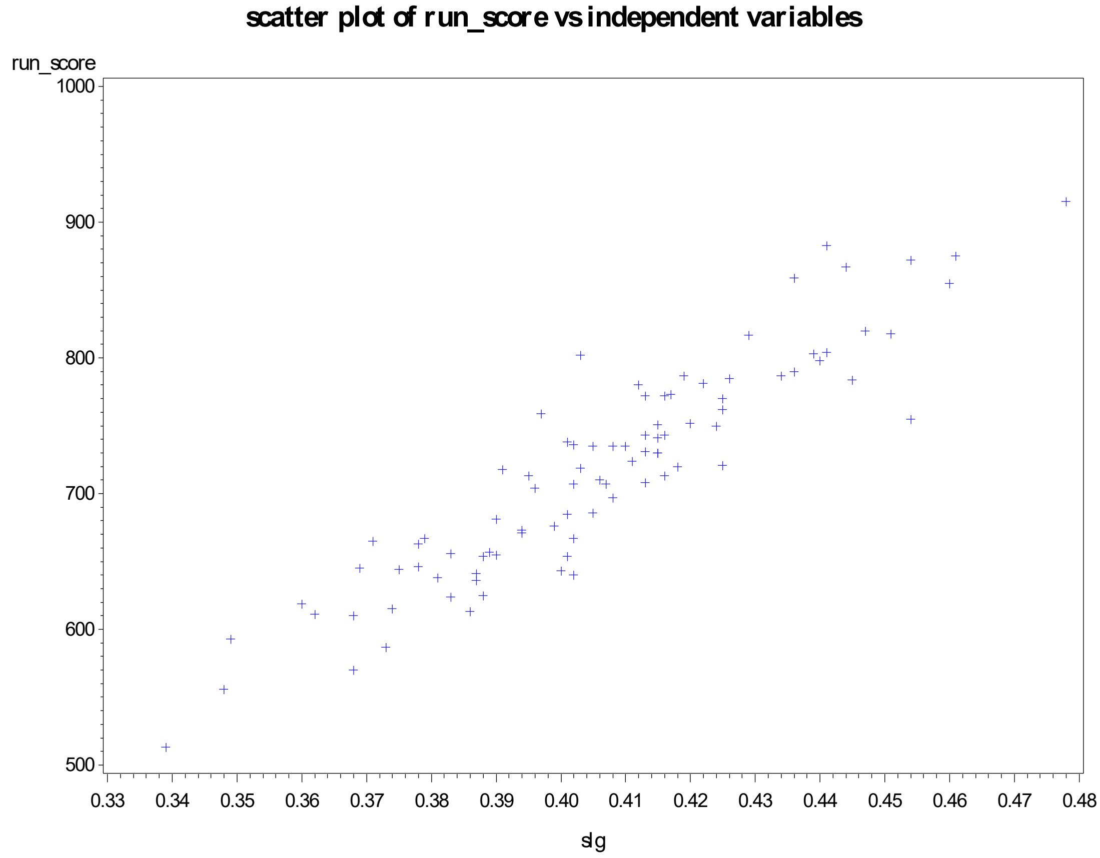
- SCATTER DIAGRAM BETWEEN RUN-SCORE AND THE EXPLANATORY VARIABLES



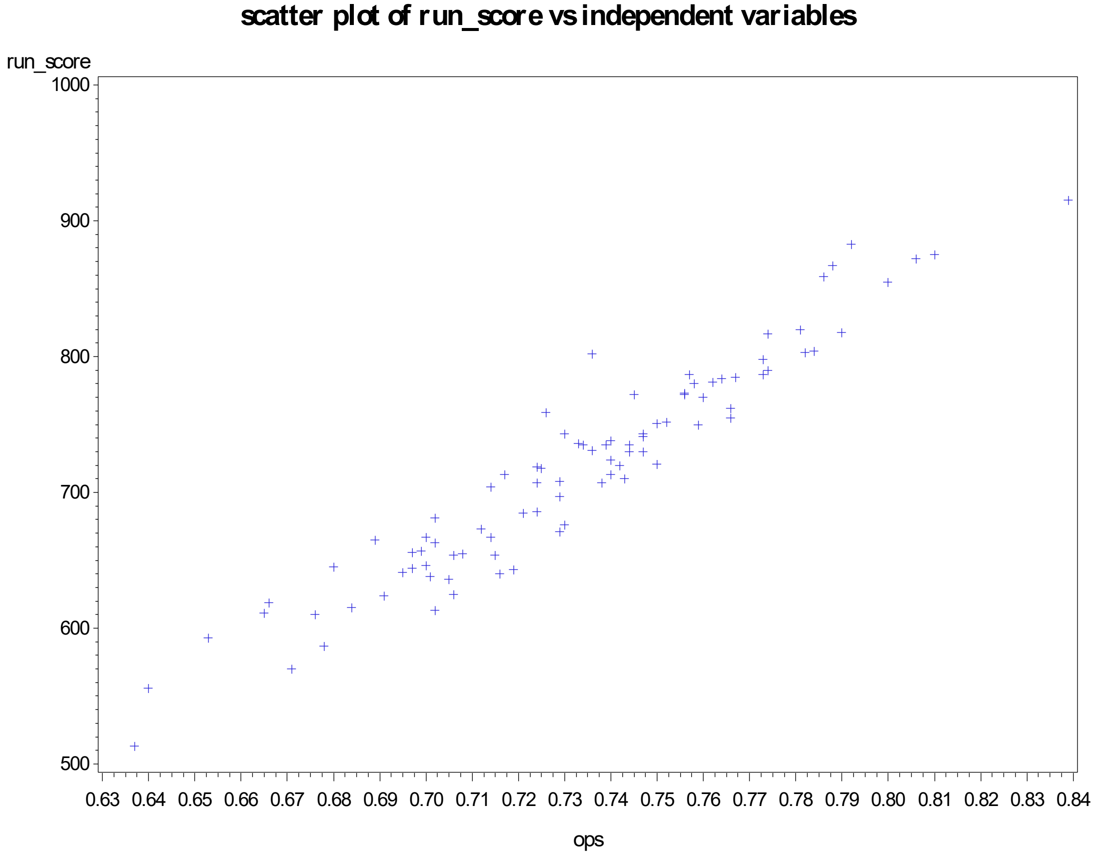
For scatter plot of RUN_SCORE Vs BATTLING_AVG, a moderate positive correlation exist between the two variables



As shown above, their appears to be a moderate positive correlation between the RUN_SCORE Vs OBP



The scatter plot of RUN_SCORE Vs SLG appear to be positively correlated.



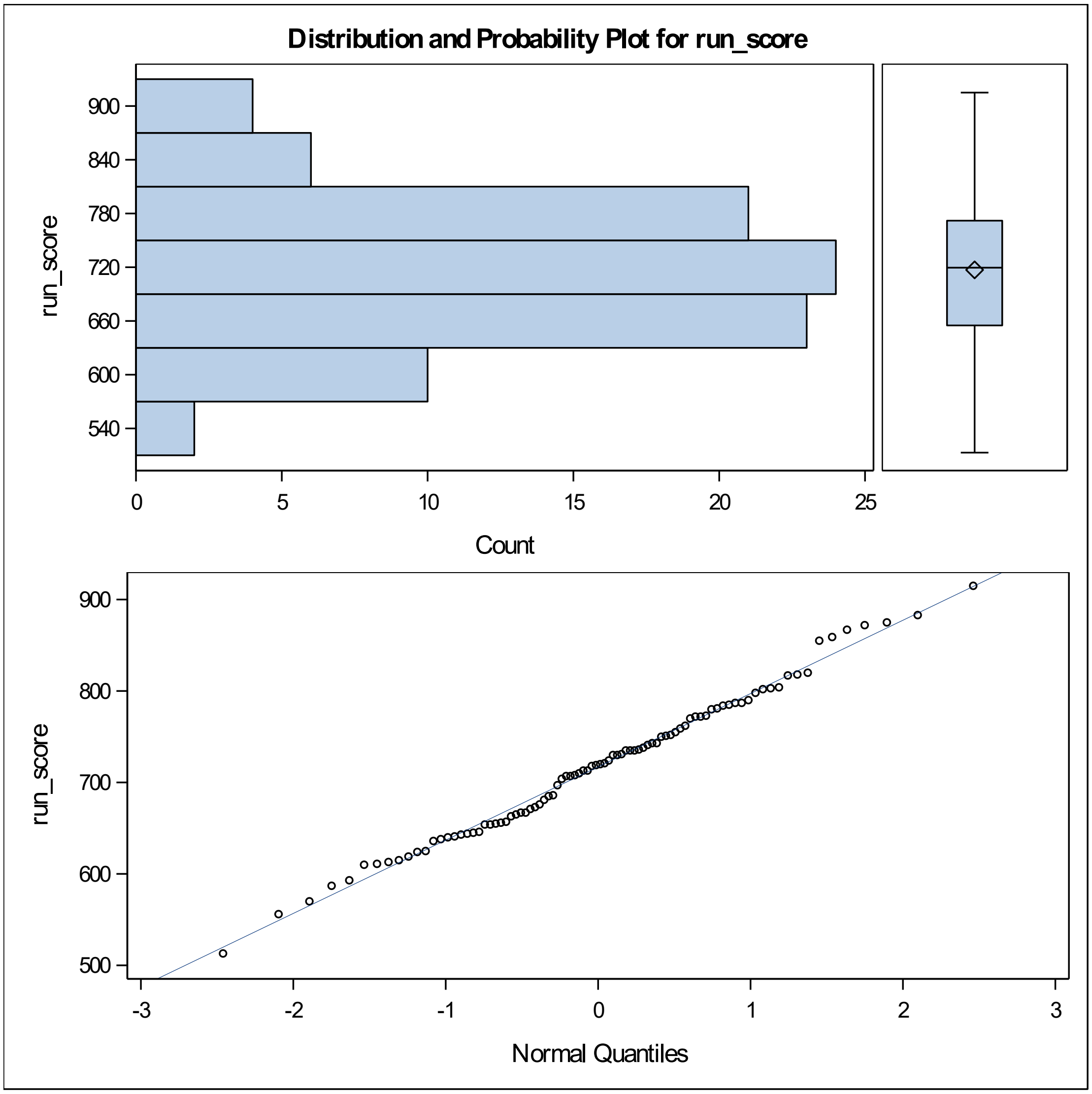
Likewise, the scatter plot between RUN_SCORE and OPS appears to be positively correlated.

From all the scatter plots above, relationship between RUNS_SCORE and all independent variables appear to be linear, so there is no need for transformation of data during multiple regression analysis.

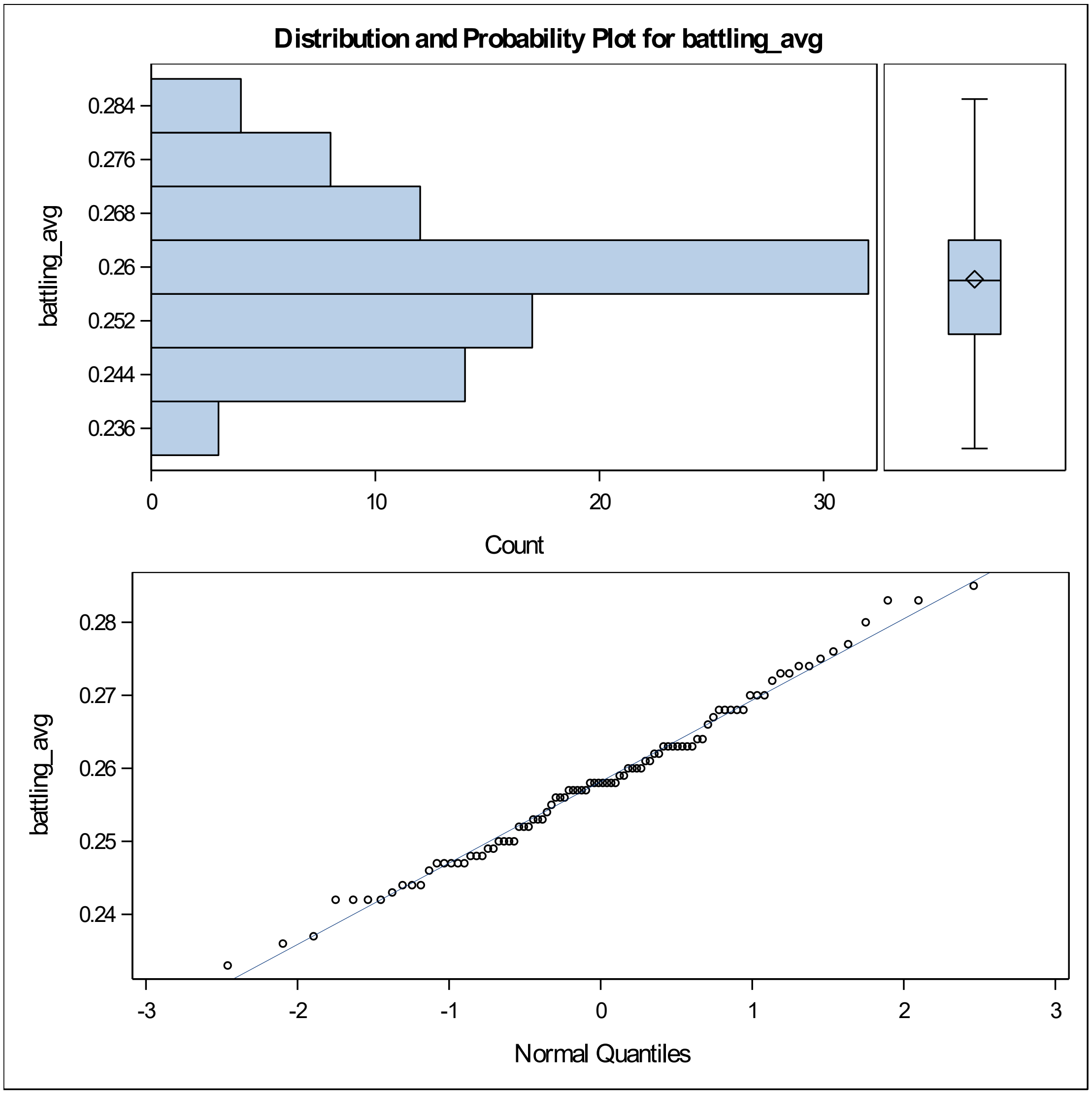
- GRAPHS

NORMAL PLOT

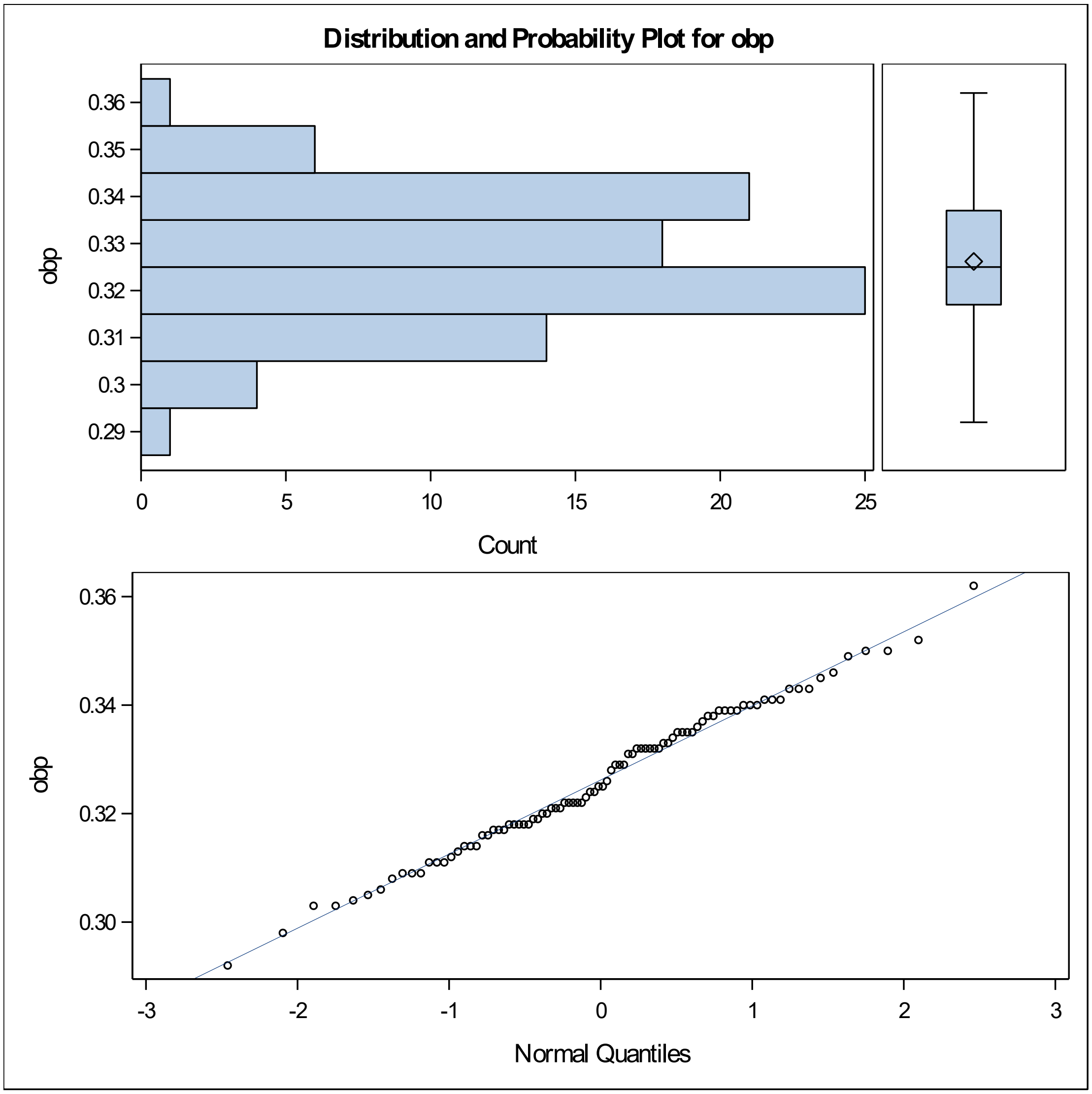
Testing for normality of all variables, we have:



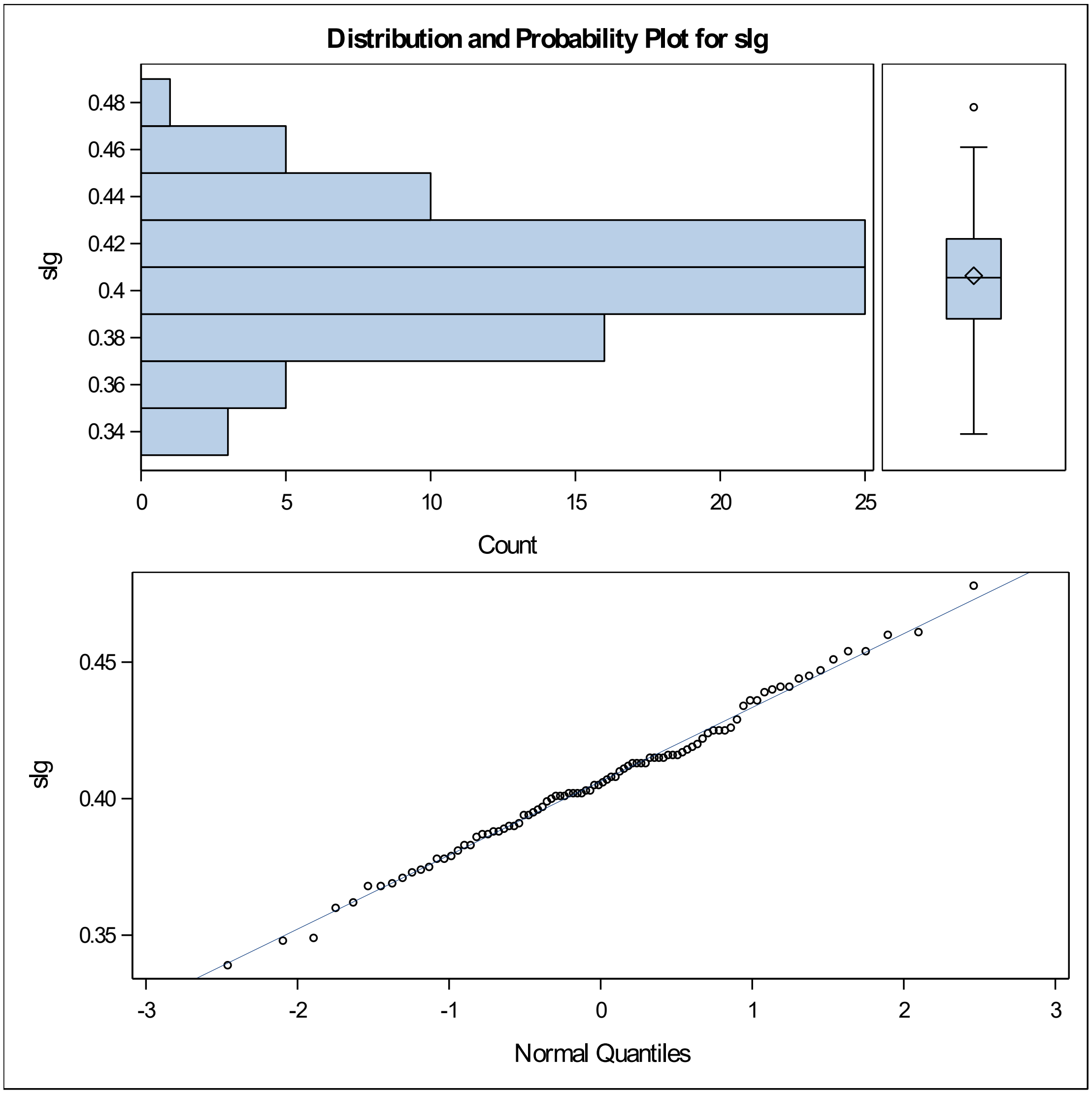
The points in this plot form a linear pattern. Its shows that the normal distribution is a good model for this data set



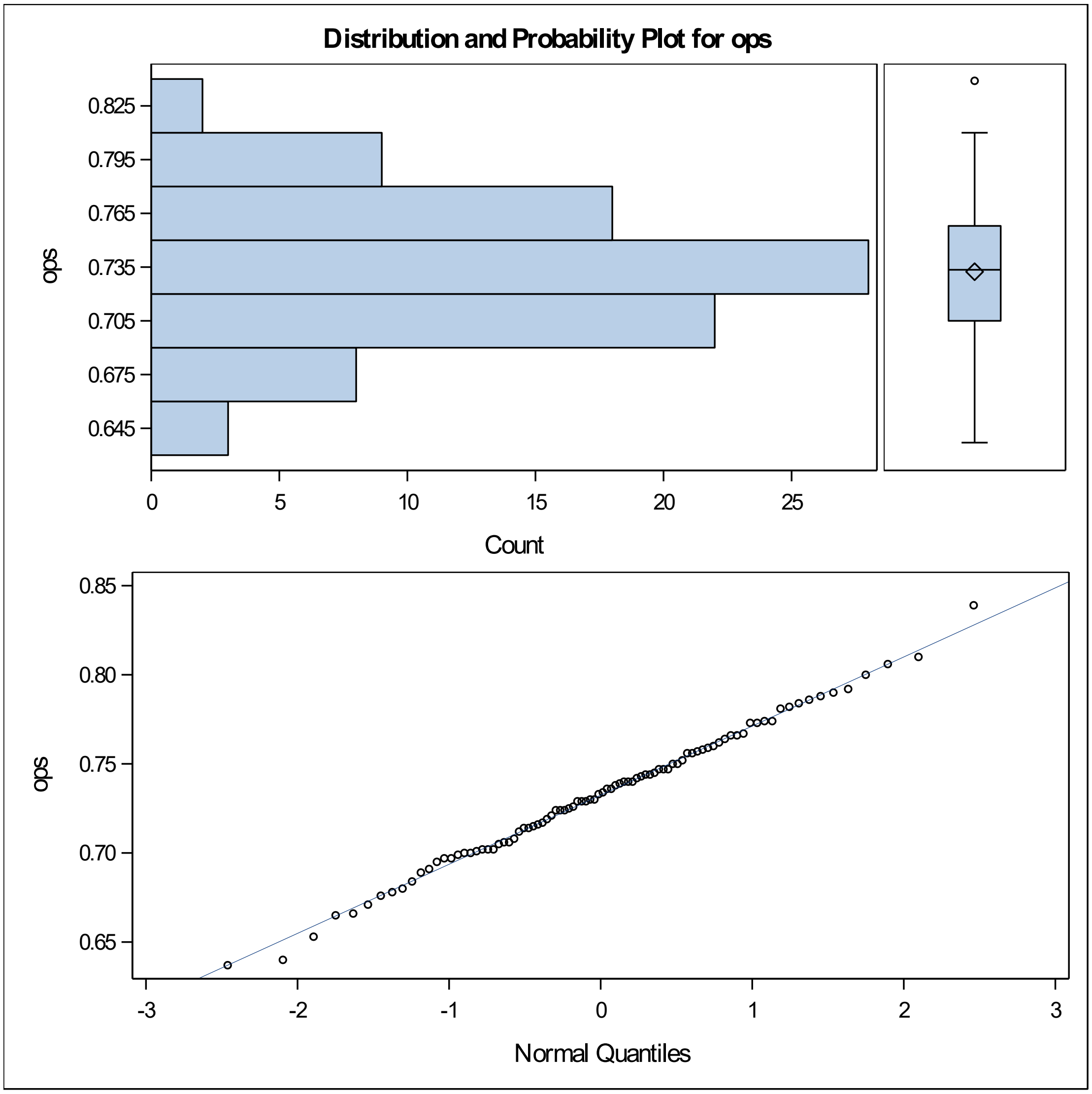
Likewise, BATTLING_AVG data form a linear pattern. It shows that the data is normally distributed.



For the OBP data, all points form a nearly pattern. It shows that the normal distribution is good model for this data set.



This data is normally distributed and all points form a moderately linear pattern.



For OPS variables, all values are normally distributed. The points on this plot form a nearly linear pattern.

STATISTICAL TEST

The following test will be carried out for analysis

- FITTING MULTIPLE REGRESSION MODEL AND TEST FOR VARIABILITY
- STATISTICAL INFERENCE ON MULTIPLE REGRESSION
- REGRESSION DIAGNOSIS
- RESIDUALS ANALYSIS
- MULTICOLLINEARITY
- VARIABLE SELECTION METHODS

JUSTIFICATION

The reason for using all this test for analysis is to test which of the predictive variable will have high correlation with the independent variable (RUN_SCORE).

STATISTICAL ANALYSIS

In this section, statistical analysis will be carried out on the response variable vs the explanatory variables using multiple regression analysis. Also, a test to know which variable has a greater correlation with the response variable will be conducted. Testing which explanatory variable fits the model will too and several other test.

- FITTING MULTIPLE REGRESSION MODEL AND TEST FOR VARIABILITY

_To predict the RUNS-SCORE, a model is developed. The model is

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$. A PROC REG of SAS software will be used to fit the model as follows:

$y = -786.01 - 789.47x_1 - 7839.50x_2 - 9173.3x_3 + 10911x_4$

Where $\beta_0 = -786.01, \beta_1 = -789.4, \beta_2 = -7839.50, \beta_3 = -9173.3, \beta_4 = 10911$

$All\ these\ are\ all\ cofficients\ of\ multiple\ regression.$

Root MSE	21.99046	**R-Square**	0.9281
Dependent Mean	717.05556	**Adj R-Sq**	0.9247
Coeff Var	3.06677		

Also from the table above, about 92.8% in RUN_SCORE IS accounted for by BATLLING_AVG, OBP, SLG and OPS.

- STATISTICAL INFERENCE FOR MULTIPLE REGRESSION

After fitting a multiple regression, the next is to determine which explanatory variables (BATLLING_AVG, OBP, SLG and OPS.) have a statistically significant effect on the response variable (RUN_SCORE).

This can be done by testing the hypothesis $H_{0j} : \beta_j = 0\ Vs\ H_{1j} : \beta_j \neq 0\ for\ each\ \beta_j.$

$That\ is, test\ for\ H_0 := \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0\ Vs$

$H_1 :\neq \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$

Using PROC REG in SAS, derive analysis of variance (ANOVA) for this test as follows:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	530482	132621	274.25	<.0001
Error	85	41104	483.58041		
Corrected Total	89	571587			

From SAS output, we can see that the F-statistic for testing $at\ \alpha = 0.05\ significant\ level.$
$H_0:\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ equals $274.25 > f_{4,85,0.05} = 2.479$
This shows that one or all of $\beta_1, \beta_2, \beta_3, \beta_4$ non zero

$Next\ we\ test\ the\ significance\ of\ individual\ terms$
That is $H_{01}:\ \beta_1 = 0, H_{02} = \beta_2 = 0, H_{02} = \beta_3 = 0, H_{04} = \beta_4 = 0$

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-786.01473	60.31856	-13.03	<.0001
battling_avg	1	-789.46564	397.51178	-1.99	0.0503
obp	1	-7839.49410	4212.63339	-1.86	0.0662
slg	1	-9173.65474	4250.71303	-2.16	0.0337
ops	1	10911	4237.86669	2.57	0.0118

$testing\ for\ t_{85,0.025} = 1.99\ at\ \alpha = 0.05. If\ we\ use\ the\ T - statistic\ of\ each\ variable$
$as\ shown\ in\ the\ table\ above, we\ have\ the\ following$

$for\ \beta_1; |t| = 1.99 = 1.99, accept\ H_{01}\ no\ significant\ different\ and\ reject\ x_1$

$for\ \beta_2; |t| = 1.86 < 1.99, accept\ H_{02}\ No\ significant\ different\ and\ reject\ x_2$

$for\ \beta_3; |t| = 2.16 > 1.99, reject\ H_{03}\ There\ is\ significant\ different\ and\ accept\ x_3$

$for\ \beta_4; |t| = 2.57 > 1.99, reject\ H_{04}. There\ is\ significant\ different\ and\ accept\ x_4$

$Also, the\ P - Value\ shows\ that\ \ x_1\ and\ x_2\ should\ be\ removed\ from\ the\ model,$
$leaving\ behind\ \ x_3\ and\ x_4$

$We\ can\ now\ compute$ 95% $CI\ for\ \beta_3\ and\ \beta_{4\ .}\ that\ is$

$$CI\ for\ \beta_3:\ \left[\hat{\beta}_3 \pm t_{85,0.025} SE\left(\hat{\beta}_3\right)\right] = [-9173.65 \pm 1.99 \times 4250.71] = [17632.56, -714.74]$$

$$CI\ for\ \beta_4:\ \left[\hat{\beta}_4 \pm t_{85,0.025} SE\left(\hat{\beta}_4\right)\right] = [10911 \pm 1.99 \times 4237.84] = [2477.64, 19344.36]$$

$We\ may\ now\ refit\ the\ rectify\ \ model.$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-820.35086	59.49940	-13.79	<.0001
slg	1	-947.95439	414.65213	-2.29	0.0247
ops	1	2624.96315	289.22612	9.08	<.0001

$Thus\ we\ have\ y = \ \beta_0 + \beta_3 x_3 + \beta_4 x_4 = \ -820.35 - 947.95 x_3 + 2624.96 x_4$

- REGRESSION DIAGNOSTICS OF THE NEW MODEL

RESIDUAL ANALYSIS

With the new fitted model, I want to test if there is an influential and an outliers' observations.

$Using\ INFLUENCE\ option\ of\ PROC\ REG, we\ derive\ the\ following\ table.$

Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	slg	ops
1	10.8831	0.4825	0.0172	1.0449	0.0639	0.0228	0.0378	-0.0359
2	3.8598	0.1724	0.0343	1.0710	0.0325	0.0272	0.0198	-0.0232
3	6.2579	0.2802	0.0384	1.0737	0.0560	0.0337	0.0471	-0.0466
4	6.1377	0.2776	0.0572	1.0951	0.0684	-0.0388	0.0095	0.0037
5	-22.3681	-1.0014	0.0277	1.0284	-0.1692	-0.1198	-0.1217	0.1293
6	-11.0668	-0.4904	0.0164	1.0437	-0.0632	-0.0274	0.0029	0.0048
7	15.3933	0.6813	0.0113	1.0303	0.0727	0.0055	0.0076	-0.0068
8	25.5171	1.1370	0.0153	1.0054	0.1418	0.0790	0.0468	-0.0580
9	4.1644	0.1839	0.0117	1.0463	0.0200	-0.0035	-0.0027	0.0034
10	-10.3335	-0.4595	0.0234	1.0523	-0.0711	0.0372	-0.0003	-0.0108
11	-40.0668	-1.8064	0.0164	0.9413	-0.2330	-0.1011	0.0107	0.0178
12	-5.5890	-0.2490	0.0288	1.0637	-0.0429	-0.0286	-0.0031	0.0102
13	-9.2207	-0.4076	0.0123	1.0422	-0.0455	0.0043	-0.0059	0.0030
14	-5.7952	-0.2592	0.0363	1.0718	-0.0503	-0.0369	-0.0398	0.0417
15	-9.7656	-0.4374	0.0378	1.0688	-0.0867	0.0031	0.0587	-0.0481
16	-24.4909	-1.0981	0.0285	1.0220	-0.1880	-0.0522	-0.1367	0.1220
17	32.3890	1.4724	0.0445	1.0055	0.3176	0.1958	-0.0294	-0.0298
18	5.9027	0.2655	0.0468	1.0835	0.0588	-0.0277	-0.0511	0.0487
19	39.7716	1.8097	0.0342	0.9583	0.3405	-0.2096	-0.0064	0.0661
20	30.1711	1.3607	0.0325	1.0038	0.2494	0.1307	-0.0406	-0.0027
21	25.6943	1.1438	0.0131	1.0026	0.1318	0.0308	-0.0172	0.0064
22	4.7230	0.2111	0.0349	1.0711	0.0401	0.0265	0.0005	-0.0076
23	30.0860	1.3781	0.0617	1.0334	0.3534	0.1874	-0.0836	0.0155
24	26.2626	1.2134	0.0822	1.0720	0.3631	0.3189	0.1078	-0.1738
25	-22.1522	-0.9998	0.0435	1.0455	-0.2131	-0.1778	-0.0660	0.1012
26	-25.4903	-1.1418	0.0254	1.0154	-0.1845	0.1335	0.0889	-0.1096
27	7.9552	0.3516	0.0123	1.0437	0.0393	0.0136	0.0092	-0.0107

Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	slg	ops
28	11.4394	0.5200	0.0648	1.0966	0.1368	-0.0249	0.0751	-0.0521
29	38.6329	1.7562	0.0344	0.9646	0.3314	0.1926	0.2724	-0.2687
30	-6.4322	-0.2871	0.0320	1.0664	-0.0522	-0.0433	-0.0262	0.0326
31	-14.7729	-0.6560	0.0184	1.0390	-0.0897	-0.0277	-0.0547	0.0505
32	-3.9922	-0.1796	0.0484	1.0867	-0.0405	0.0250	0.0353	-0.0353
33	-43.4629	-1.9679	0.0181	0.9238	-0.2670	-0.1438	-0.0147	0.0498
34	-7.8426	-0.3526	0.0458	1.0803	-0.0773	0.0185	-0.0360	0.0230
35	-7.1179	-0.3148	0.0141	1.0465	-0.0377	-0.0181	-0.0143	0.0161
36	-3.4806	-0.1539	0.0140	1.0491	-0.0183	0.0050	-0.0010	-0.0008
37	-8.0625	-0.3586	0.0246	1.0567	-0.0570	0.0241	-0.0090	-0.0001
38	-12.7966	-0.5719	0.0320	1.0574	-0.1039	0.0011	0.0663	-0.0537
39	-1.7405	-0.0770	0.0168	1.0528	-0.0101	0.0042	0.0002	-0.0014
40	-3.9704	-0.1758	0.0171	1.0521	-0.0232	0.0127	0.0107	-0.0123
41	20.9032	0.9283	0.0136	1.0186	0.1090	0.0471	0.0413	-0.0450
42	28.9098	1.3120	0.0459	1.0225	0.2879	0.2200	0.0397	-0.0918
43	-41.6384	-1.8857	0.0219	0.9374	-0.2824	0.1102	0.1980	-0.1907
44	28.3289	1.2706	0.0245	1.0037	0.2015	0.1540	0.1037	-0.1242
45	9.1514	0.4078	0.0281	1.0591	0.0694	0.0041	-0.0396	0.0308
46	-20.0635	-0.8920	0.0165	1.0239	-0.1154	0.0506	0.0071	-0.0209
47	1.1657	0.0518	0.0244	1.0610	0.0082	-0.0058	-0.0043	0.0051
48	9.8181	0.4359	0.0207	1.0502	0.0634	0.0386	0.0066	-0.0155
49	29.4379	1.3380	0.0482	1.0225	0.3010	-0.2580	-0.1714	0.2104
50	-1.0465	-0.0468	0.0373	1.0754	-0.0092	0.0011	0.0066	-0.0056
51	28.2585	1.2602	0.0135	0.9934	0.1474	-0.0563	-0.0439	0.0523
52	-18.7872	-0.8441	0.0380	1.0498	-0.1677	-0.1395	-0.0592	0.0852
53	28.4423	1.2837	0.0362	1.0147	0.2489	0.0351	-0.1373	0.1009
54	-17.3941	-0.7993	0.0809	1.1017	-0.2372	-0.1465	0.0352	0.0122

Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	slg	ops
55	-9.4819	-0.4201	0.0166	1.0463	-0.0546	-0.0249	-0.0311	0.0312
56	13.3305	0.5918	0.0186	1.0421	0.0814	-0.0315	-0.0516	0.0506
57	72.4036	3.4261	0.0223	0.7214	0.5177	-0.2472	-0.3653	0.3641
58	17.4466	0.7767	0.0212	1.0357	0.1142	-0.0746	-0.0585	0.0684
59	-4.9996	-0.2494	0.2256	1.3340	-0.1346	-0.0645	-0.1287	0.1201
60	-13.4208	-0.5949	0.0156	1.0388	-0.0749	-0.0353	-0.0035	0.0120
61	-11.1269	-0.4942	0.0207	1.0482	-0.0718	-0.0226	-0.0471	0.0432
62	-13.7002	-0.6152	0.0405	1.0649	-0.1265	0.0809	0.1059	-0.1077
63	-6.0955	-0.2694	0.0132	1.0465	-0.0312	0.0111	0.0062	-0.0084
64	7.0018	0.3161	0.0536	1.0901	0.0752	-0.0579	-0.0146	0.0282
65	-24.0545	-1.0706	0.0146	1.0097	-0.1304	0.0478	0.0616	-0.0637
66	-8.5126	-0.3761	0.0116	1.0423	-0.0408	0.0065	0.0034	-0.0050
67	-2.1289	-0.0942	0.0151	1.0508	-0.0116	-0.0062	-0.0028	0.0038
68	4.1757	0.1856	0.0250	1.0605	0.0297	-0.0204	-0.0185	0.0207
69	-15.5724	-0.6963	0.0311	1.0506	-0.1247	0.0769	0.0060	-0.0272
70	-3.1476	-0.1390	0.0127	1.0479	-0.0158	0.0041	0.0007	-0.0018
71	2.2277	0.0992	0.0289	1.0658	0.0171	-0.0121	-0.0118	0.0129
72	-44.8159	-2.0289	0.0151	0.9136	-0.2514	-0.1341	-0.1056	0.1192
73	-10.2009	-0.4524	0.0182	1.0470	-0.0615	-0.0338	-0.0369	0.0383
74	42.4279	1.9517	0.0494	0.9563	0.4447	-0.3800	-0.2167	0.2815
75	1.1860	0.0538	0.0623	1.1039	0.0138	-0.0104	-0.0119	0.0125
76	-4.1673	-0.1853	0.0251	1.0607	-0.0297	0.0213	0.0133	-0.0168
77	12.3018	0.5505	0.0347	1.0613	0.1044	-0.0836	-0.0580	0.0702
78	-48.7533	-2.2528	0.0446	0.9124	-0.4870	0.2461	0.4215	-0.4071
79	-13.8710	-0.6419	0.0963	1.1293	-0.2095	0.1546	0.0116	-0.0534
80	49.9655	2.2808	0.0196	0.8854	0.3226	-0.0942	-0.2086	0.1951
81	13.9902	0.6308	0.0480	1.0725	0.1416	-0.0047	0.0897	-0.0692

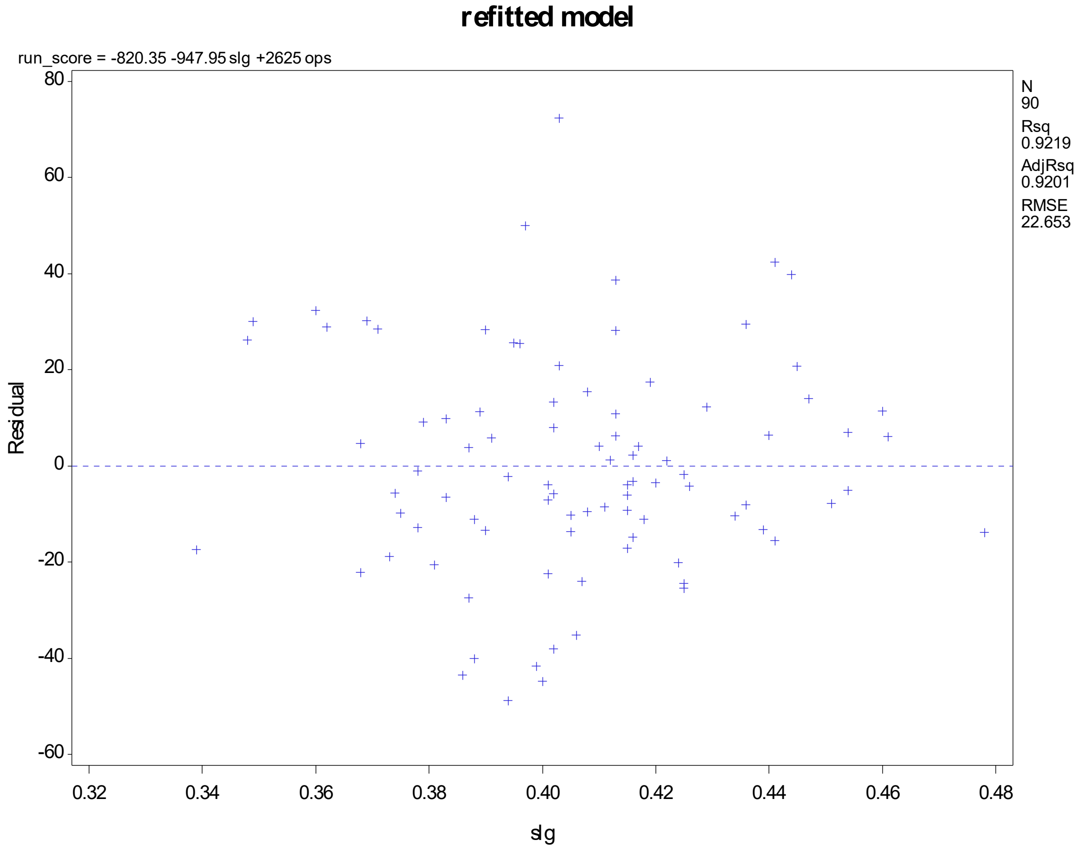
Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	slg	ops
82	-27.3898	-1.2229	0.0169	1.0000	-0.1602	-0.0664	0.0144	0.0057
83	-20.5777	-0.9185	0.0236	1.0297	-0.1428	-0.0218	0.0666	-0.0479
84	-38.0451	-1.7235	0.0289	0.9629	-0.2974	-0.2093	-0.2198	0.2315
85	11.2559	0.5022	0.0294	1.0573	0.0874	0.0705	0.0508	-0.0596
86	-17.0955	-0.7578	0.0132	1.0284	-0.0877	0.0313	0.0175	-0.0236
87	-13.2184	-0.5902	0.0299	1.0543	-0.1036	0.0672	0.0121	-0.0292
88	20.7187	0.9731	0.1172	1.1349	0.3547	0.1465	0.3254	-0.2987
89	6.3543	0.2841	0.0356	1.0705	0.0546	-0.0026	0.0319	-0.0243
90	-35.1273	-1.5889	0.0309	0.9795	-0.2839	0.1725	0.2225	-0.2273

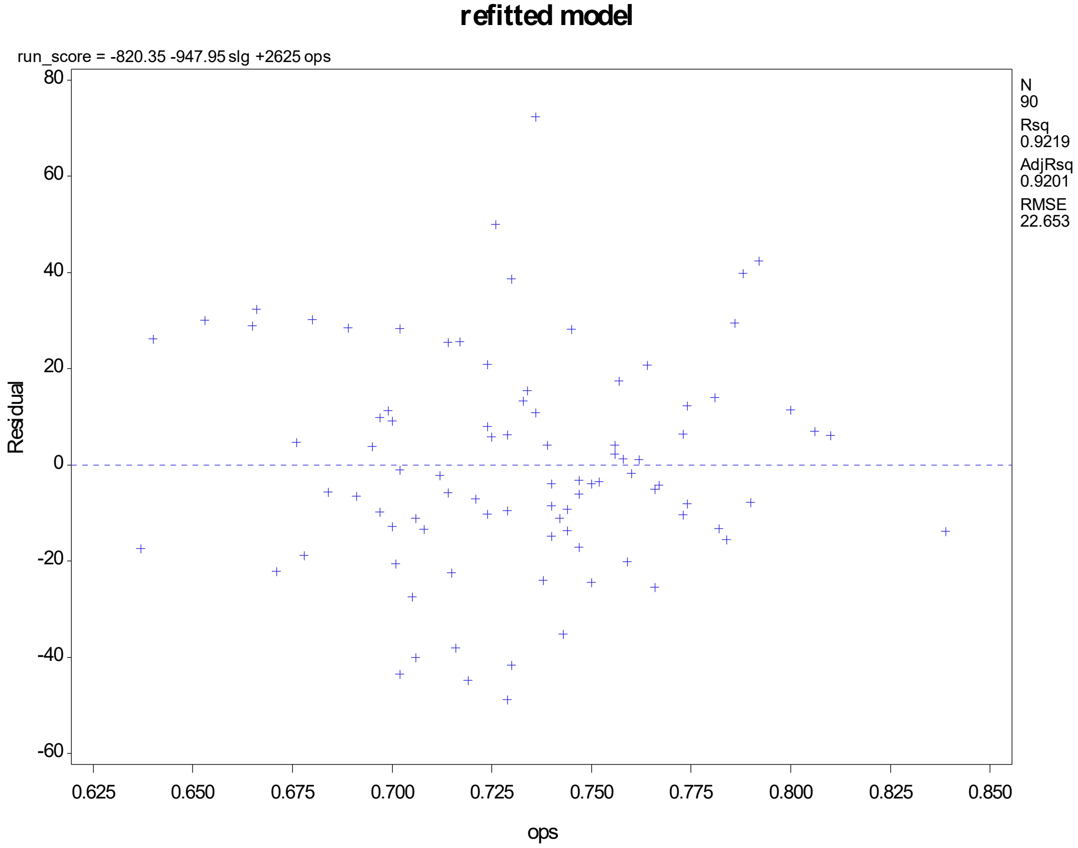
From the above table, we can test if there is an influential observations using the following formula;

$$if\ h_{ii} > \frac{2(k+1)}{n}\ then\ we\ have\ an\ influential\ observation.$$

$$For\ influential\ obervation, we\ have\ observation\ 24, 54\ and\ 79.$$

PLOTS OF RESIDUALS AGAINST INDIVIDUAL PREDICTORS VARIABLES FOR THE NEW FITTED MODEL

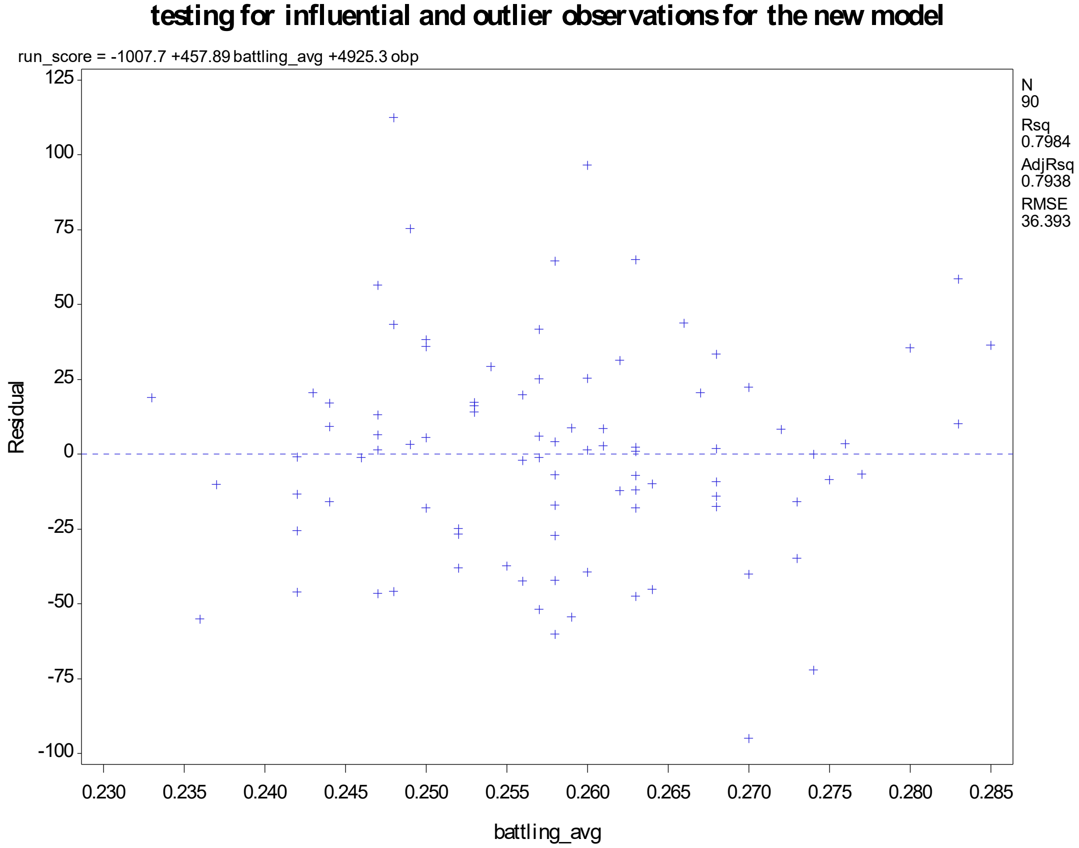


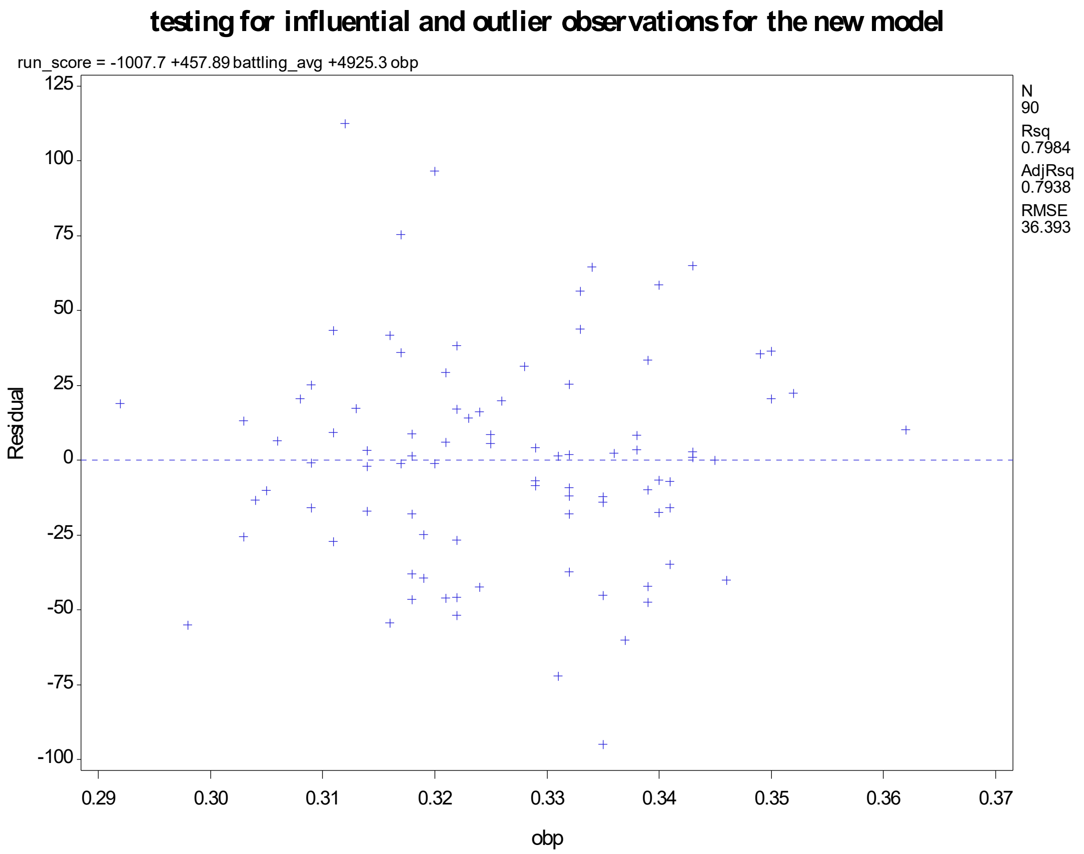


For each predictor variable, each plots are randomly dispersed.

PLOTS OF RESIDUALS AGAINST ANY OMITTED PREDICTOR VAIRABLES

This test is done to show if the excluded predictor variables BATTLING_AVG and OBP is worth included in the data. The two plots show points to be randomly scatter though out both plots and also a systematic pattern.





The two omitted variables look to be included in the model. But to test this further, we shall test for ru multicollinearity.

MULTICOLLINEARITY

Multicollinearity occurs when predictor variables are linearly related.
Going back to the first model, we can also use multicollinearity to test which variables should be excluded. This will also be used to test the first assumption of omitting BATTLING_AVG and OBP variables in the first place.

MEASURES OF MULTICOLLINEARITY

Use VIF option of PROC REG to measure MULTICOLLINEARITY.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-786.01473	60.31856	-13.03	<.0001	0
battling_avg	1	-789.46564	397.51178	-1.99	0.0503	3.61502
obp	1	-7839.49410	4212.63339	-1.86	0.0662	609.24717
slg	1	-9173.65474	4250.71303	-2.16	0.0337	2433.21401
ops	1	10911	4237.86669	2.57	0.0118	4971.00404

From the table above, there are large VIF values, we can assume there are serious COLLINEARITY problems.

STANDARDIZED REGRESSION COEFFICIENTS

Comparing predictors variables magnitudes of their effects on the response variable. In order to do this effectively, I used STB option of PROC REG in SAS. The output is given as follows

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**
Intercept	1	-786.01473	60.31856	-13.03	<.0001	0
battling_avg	1	-789.46564	397.51178	-1.99	0.0503	-0.10983
obp	1	-7839.49410	4212.63339	-1.86	0.0662	-1.33606
slg	1	-9173.65474	4250.71303	-2.16	0.0337	-3.09645
ops	1	10911	4237.86669	2.57	0.0118	5.28002

From the table above, variable OPS seems to have larger effect on RUN_SCORE than any other variables.

VARIABLE SELECTION METHOD

STEPWISE REGRESSION

This is used for variable selection into the model. To know which variable deserves to be in the model. So I use stepwise regression of SAS solve the problem.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ops		1	0.9172	0.9172	11.8708	974.78	<.0001
2	slg		2	0.0047	0.9219	8.3245	5.23	0.0247
3	battling_avg		3	0.0033	0.9252	6.4631	3.75	0.0560
4	obp		4	0.0029	0.9281	5.0000	3.46	0.0662

Its shows that all variables seem to have entered the model. Known was removed at 0.015 significant level.

- ASSUMPTIONS

The following are the assumptions from this analysis:

1. All predictive variables are normally distributed. This shown from the normal plot above.
2. At the initial part of this variable, we tested for hypothesis to know whether each coefficient is significant. Some variables were rejected as a result of this.
3. It is possible to accept the rejected explanatory variables earlier rejected if a residual plot between the predictive variables show a regular pattern
4. Using VIF, we see that variable OBP has a high magnitude influence on RUN_SCORE than any other variables.
5. Using STEPWISE REGRESSION, we see that all variables are included in the model. .

RESULTS

From the above analysis, It is clear that variable OPS has a strong influence to predict the RUN_SCORE of the BASE BALL in the future. Also, there is a strong correlation between RUN_SCORE and all the four variables.

In conclusion, the RUN_SCORE for the MLS players can be predicted by all the variables but OBP has a strongest CORRELATION between it and RUN_SCORE. Variable OPS can influence the prediction of RUN_SCORE in the future.