STATISTICAL PROGRAMMING FOR BUSINESS ANALYTICS

Assignment 4

Tanay Bhalerao

U47707491

MARCH 13, 2015

UNIVERSITY OF SOUTH FLORIDA
Management Information Systems

Homework 4

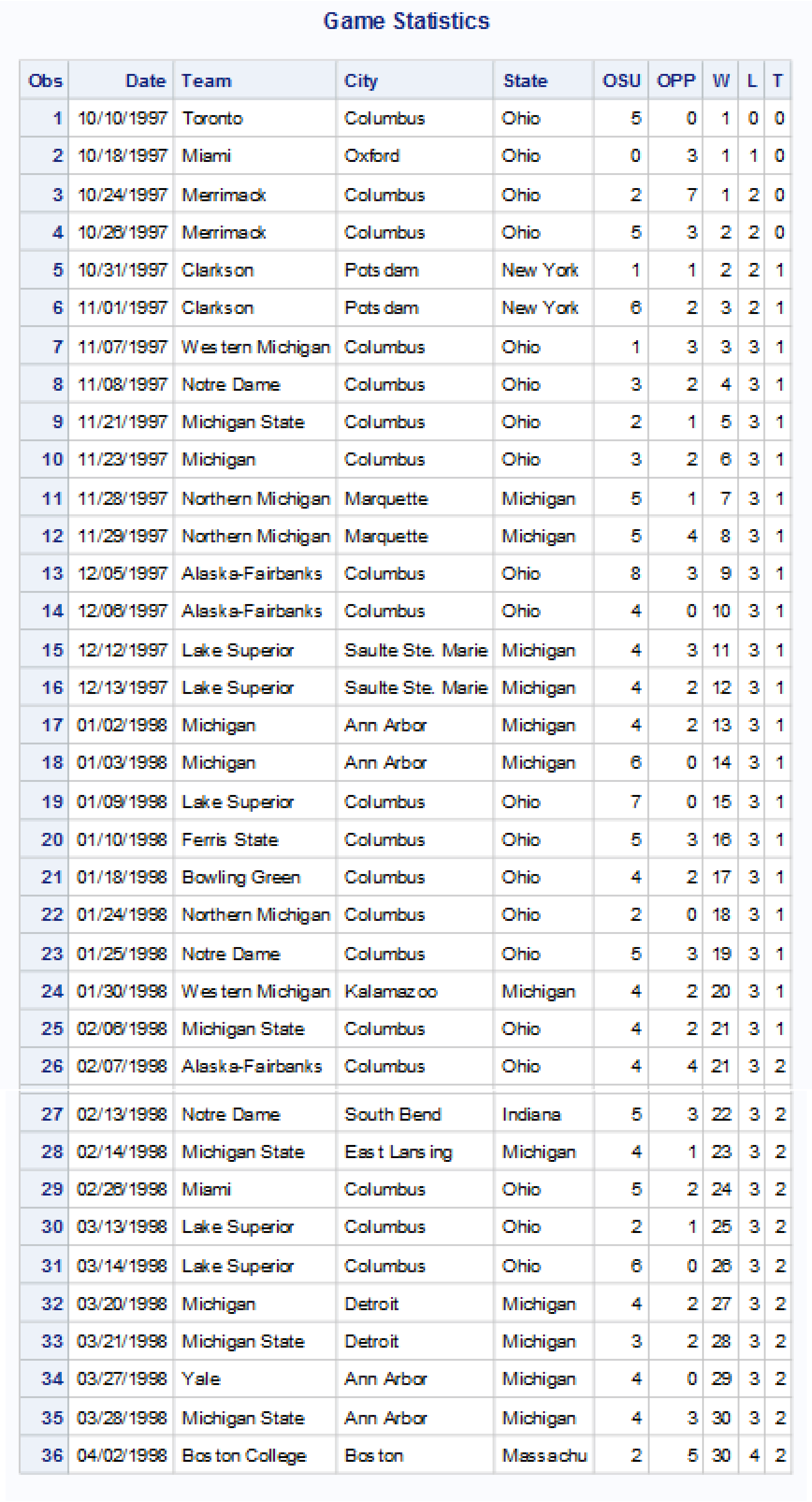
1. Refer to HOCKEY data. Write a SAS program which calculates the number of games won, lost, and tied up to and including the current observation. Print the dataset with an appropriate format for the date. Don't forget to change the score of the final game to Boston College 5, Ohio State 2 (do this in your code, don't change the original file). The first few lines of output should be similar to this:

DATE	TEAM	CITY	STATE	OSU	OPP	W	L	T
10/10/97	Toronto	Columbus	Ohio	5	0	1	0	0
10/18/97	Miami	Oxford	Ohio	0	3	1	1	0
10/24/97	Merrimack	Columbus	Ohio	2	7	1	2	0
10/26/97	Merrimack	Columbus	Ohio	5	3	2	2	0
10/31/97	Clarkson	Potsdam	New York	1	1	2	2	1

Use this code to import the data:

```
PROC IMPORT OUT= WORK.HOCKEY
            DATAFILE= "C:\Users\....\hockey.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

libname tan "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";

PROC IMPORT OUT= tan.hockey
            DATAFILE=
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\hockey.csv"
      DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

DATA tan.dathockey;
      set tan.hockey end=M_last;
      by Date;
      format      Date  mmddyy10.;
      retain      W     0
                  L     0
                  T     0;
      if M_last then do;
            OSU = 2;
            OPP = 5;
      end;
      do;
            if OSU > OPP then W = W + 1;
            else if OSU < OPP then  L = L + 1;
            else T = T + 1;
      end;

run;
proc print data=tan.dathockey;
title "Game Statistics";
```

```
run;
```

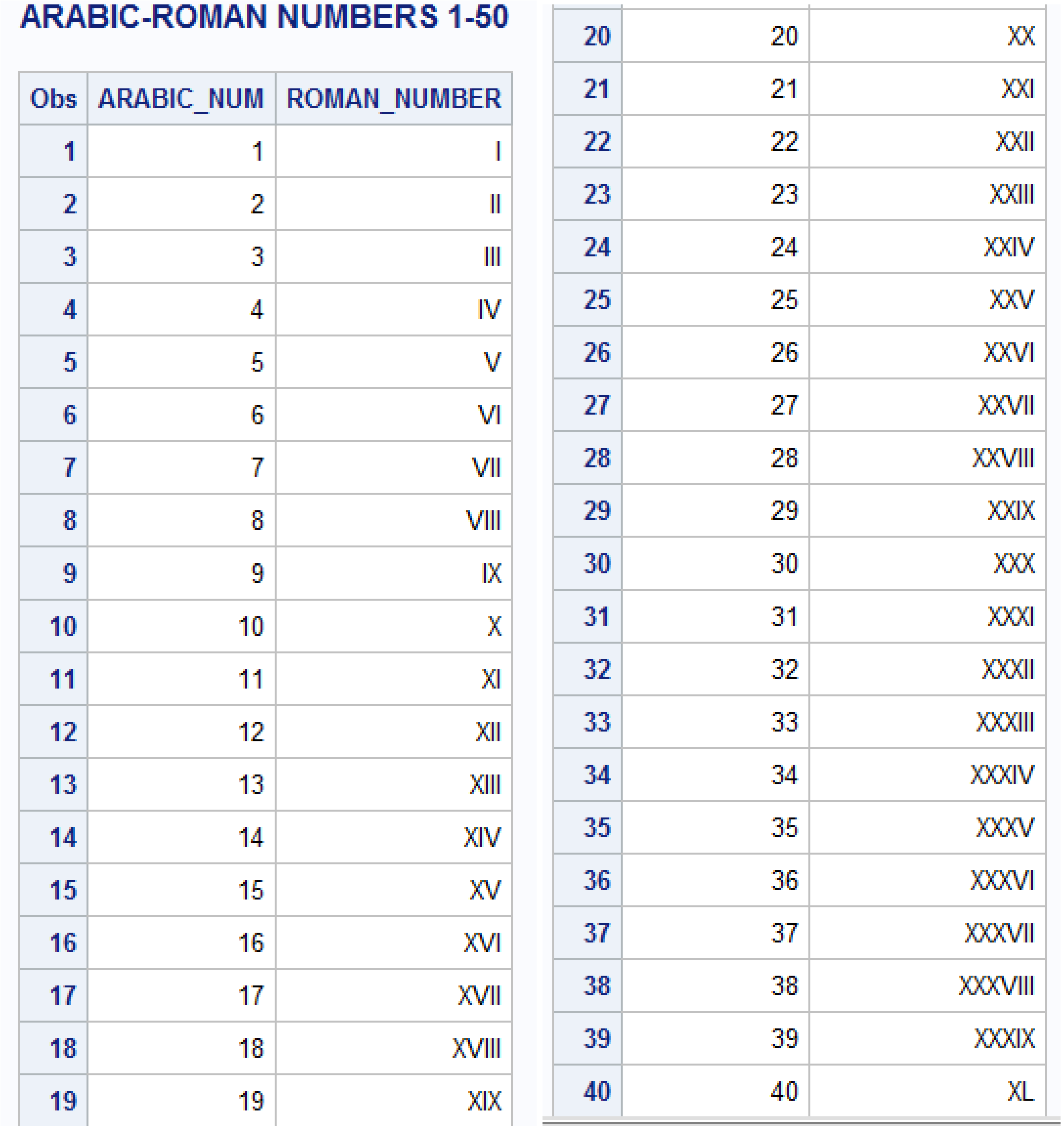


Game Statistics

Obs	Date	Team	City	State	OSU	OPP	W	L	T
1	10/10/1997	Toronto	Columbus	Ohio	5	0	1	0	0
2	10/18/1997	Miami	Oxford	Ohio	0	3	1	1	0
3	10/24/1997	Merrimack	Columbus	Ohio	2	7	1	2	0
4	10/26/1997	Merrimack	Columbus	Ohio	5	3	2	2	0
5	10/31/1997	Clarkson	Potsdam	New York	1	1	2	2	1
6	11/01/1997	Clarkson	Potsdam	New York	6	2	3	2	1
7	11/07/1997	Western Michigan	Columbus	Ohio	1	3	3	3	1
8	11/08/1997	Notre Dame	Columbus	Ohio	3	2	4	3	1
9	11/21/1997	Michigan State	Columbus	Ohio	2	1	5	3	1
10	11/23/1997	Michigan	Columbus	Ohio	3	2	6	3	1
11	11/28/1997	Northern Michigan	Marquette	Michigan	5	1	7	3	1
12	11/29/1997	Northern Michigan	Marquette	Michigan	5	4	8	3	1
13	12/05/1997	Alaska-Fairbanks	Columbus	Ohio	8	3	9	3	1
14	12/06/1997	Alaska-Fairbanks	Columbus	Ohio	4	0	10	3	1
15	12/12/1997	Lake Superior	Saulte Ste. Marie	Michigan	4	3	11	3	1
16	12/13/1997	Lake Superior	Saulte Ste. Marie	Michigan	4	2	12	3	1
17	01/02/1998	Michigan	Ann Arbor	Michigan	4	2	13	3	1
18	01/03/1998	Michigan	Ann Arbor	Michigan	6	0	14	3	1
19	01/09/1998	Lake Superior	Columbus	Ohio	7	0	15	3	1
20	01/10/1998	Ferris State	Columbus	Ohio	5	3	16	3	1
21	01/18/1998	Bowling Green	Columbus	Ohio	4	2	17	3	1
22	01/24/1998	Northern Michigan	Columbus	Ohio	2	0	18	3	1
23	01/25/1998	Notre Dame	Columbus	Ohio	5	3	19	3	1
24	01/30/1998	Western Michigan	Kalamazoo	Michigan	4	2	20	3	1
25	02/06/1998	Michigan State	Columbus	Ohio	4	2	21	3	1
26	02/07/1998	Alaska-Fairbanks	Columbus	Ohio	4	4	21	3	2
27	02/13/1998	Notre Dame	South Bend	Indiana	5	3	22	3	2
28	02/14/1998	Michigan State	East Lansing	Michigan	4	1	23	3	2
29	02/26/1998	Miami	Columbus	Ohio	5	2	24	3	2
30	03/13/1998	Lake Superior	Columbus	Ohio	2	1	25	3	2
31	03/14/1998	Lake Superior	Columbus	Ohio	6	0	26	3	2
32	03/20/1998	Michigan	Detroit	Michigan	4	2	27	3	2
33	03/21/1998	Michigan State	Detroit	Michigan	3	2	28	3	2
34	03/27/1998	Yale	Ann Arbor	Michigan	4	0	29	3	2
35	03/28/1998	Michigan State	Ann Arbor	Michigan	4	3	30	3	2
36	04/02/1998	Boston College	Boston	Massachu	2	5	30	4	2

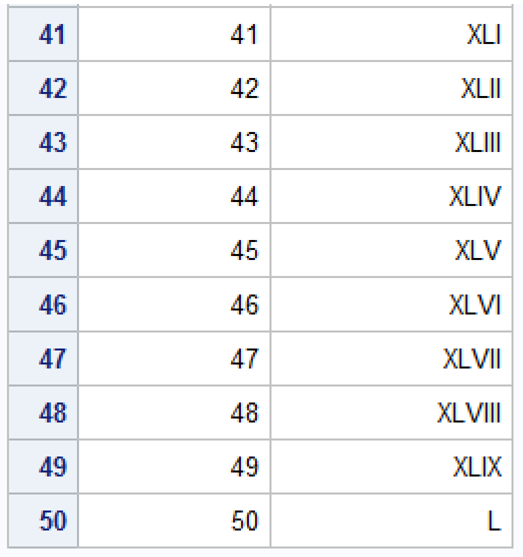
2. Suppose that your 5th grader is learning how to write Roman numerals, and you want to help her or him by preparing a study guide. Write a SAS program which uses DO loop to print the Arabic numbers 1, 2, 3, …,49, 50 AND their Roman equivalents. The ROMAN7. format in SAS will be helpful.

```
DATA ARAB_ROMAN;
DO ARABIC_NUM = 1 TO 50;
ROMAN_NUMBER = ARABIC_NUM;
FORMAT ROMAN_NUMBER ROMAN7.;
OUTPUT;
END;
RUN;
PROC PRINT DATA = ARAB_ROMAN;
TITLE "ARABIC-ROMAN NUMBERS 1-50";
RUN;
```



ARABIC-ROMAN NUMBERS 1-50

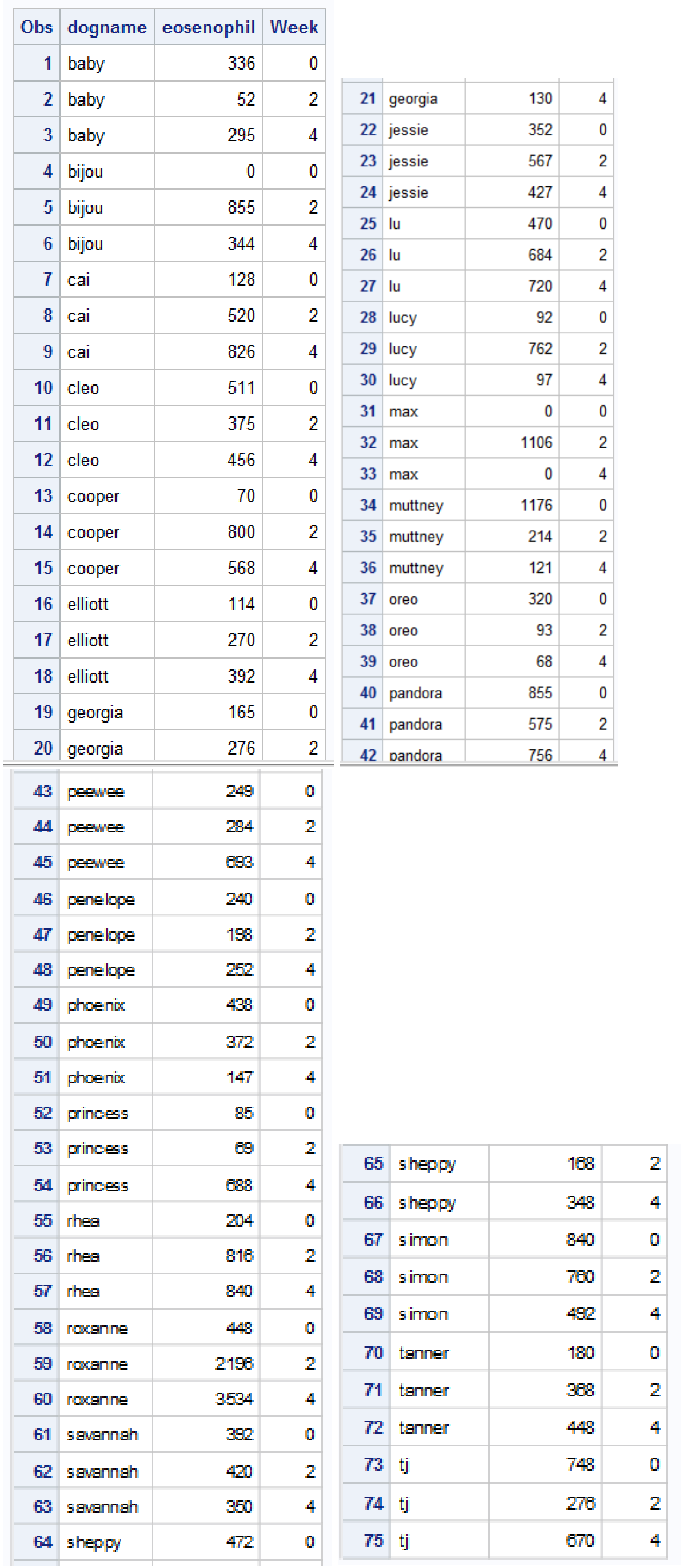
Obs	ARABIC_NUM	ROMAN_NUMBER
1	1	I
2	2	II
3	3	III
4	4	IV
5	5	V
6	6	VI
7	7	VII
8	8	VIII
9	9	IX
10	10	X
11	11	XI
12	12	XII
13	13	XIII
14	14	XIV
15	15	XV
16	16	XVI
17	17	XVII
18	18	XVIII
19	19	XIX
20	20	XX
21	21	XXI
22	22	XXII
23	23	XXIII
24	24	XXIV
25	25	XXV
26	26	XXVI
27	27	XXVII
28	28	XXVIII
29	29	XXIX
30	30	XXX
31	31	XXXI
32	32	XXXII
33	33	XXXIII
34	34	XXXIV
35	35	XXXV
36	36	XXXVI
37	37	XXXVII
38	38	XXXVIII
39	39	XXXIX
40	40	XL



41	41	XLI
42	42	XLII
43	43	XLIII
44	44	XLIV
45	45	XLV
46	46	XLVI
47	47	XLVII
48	48	XLVIII
49	49	XLIX
50	50	L

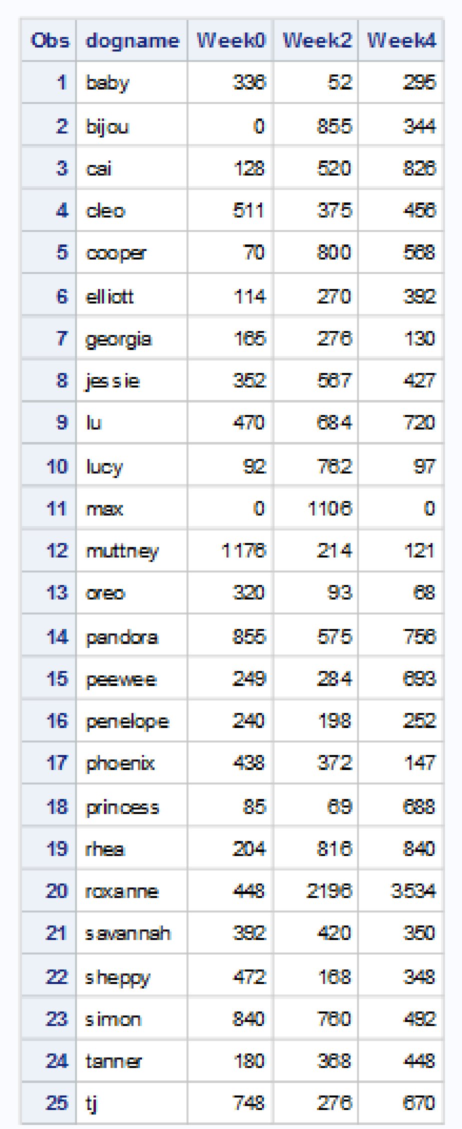
3. Refer to the DPGS3 data. Write a SAS program which creates a dataset using the INFILE statement. Then, create a new dataset which contains three variables: the name of the dog, the week of the measurement, and the eosinophil count in that week. There should be 75 observations in the new dataset. Print both datasets.

```
data tan.dogs3;
infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\dogs3.txt"
firstobs = 3 LRECL= 200 ;
input dogname     $     1-12
            Week0       13-16
            Week2       21-24
            Week4       29-32;
run;

PROC SORT DATA=tan.dogs3;
by dogname;
run;

proc transpose data = tan.dogs3 out = tan.trans_dogs3;
by dogname;
run;

data tan.trans_dogs3;
set tan.trans_dogs3
(rename=(col1=eosenophil _name_=Temp_Week));
      if Temp_Week = 'Week0' then Week = 0;
      else if Temp_Week = 'Week2' then Week = 2;
      else Week = 4;
      drop Temp_Week;
run;
proc print data=tan.trans_dogs3;
      run;
```



Obs	dogname	eosenophil	Week
1	baby	336	0
2	baby	52	2
3	baby	295	4
4	bijou	0	0
5	bijou	855	2
6	bijou	344	4
7	cai	128	0
8	cai	520	2
9	cai	826	4
10	cleo	511	0
11	cleo	375	2
12	cleo	456	4
13	cooper	70	0
14	cooper	800	2
15	cooper	568	4
16	elliott	114	0
17	elliott	270	2
18	elliott	392	4
19	georgia	165	0
20	georgia	276	2
21	georgia	130	4
22	jessie	352	0
23	jessie	567	2
24	jessie	427	4
25	lu	470	0
26	lu	684	2
27	lu	720	4
28	lucy	92	0
29	lucy	762	2
30	lucy	97	4
31	max	0	0
32	max	1106	2
33	max	0	4
34	muttney	1176	0
35	muttney	214	2
36	muttney	121	4
37	oreo	320	0
38	oreo	93	2
39	oreo	68	4
40	pandora	855	0
41	pandora	575	2
42	pandora	756	4
43	peewee	249	0
44	peewee	284	2
45	peewee	693	4
46	penelope	240	0
47	penelope	198	2
48	penelope	252	4
49	phoenix	438	0
50	phoenix	372	2
51	phoenix	147	4
52	princess	85	0
53	princess	69	2
54	princess	688	4
55	rhea	204	0
56	rhea	816	2
57	rhea	840	4
58	roxanne	448	0
59	roxanne	2196	2
60	roxanne	3534	4
61	savannah	392	0
62	savannah	420	2
63	savannah	350	4
64	sheppy	472	0
65	sheppy	168	2
66	sheppy	348	4
67	simon	840	0
68	simon	760	2
69	simon	492	4
70	tanner	180	0
71	tanner	368	2
72	tanner	448	4
73	tj	748	0
74	tj	276	2
75	tj	670	4

```
proc print data=tan.dogs3;
       run;
```

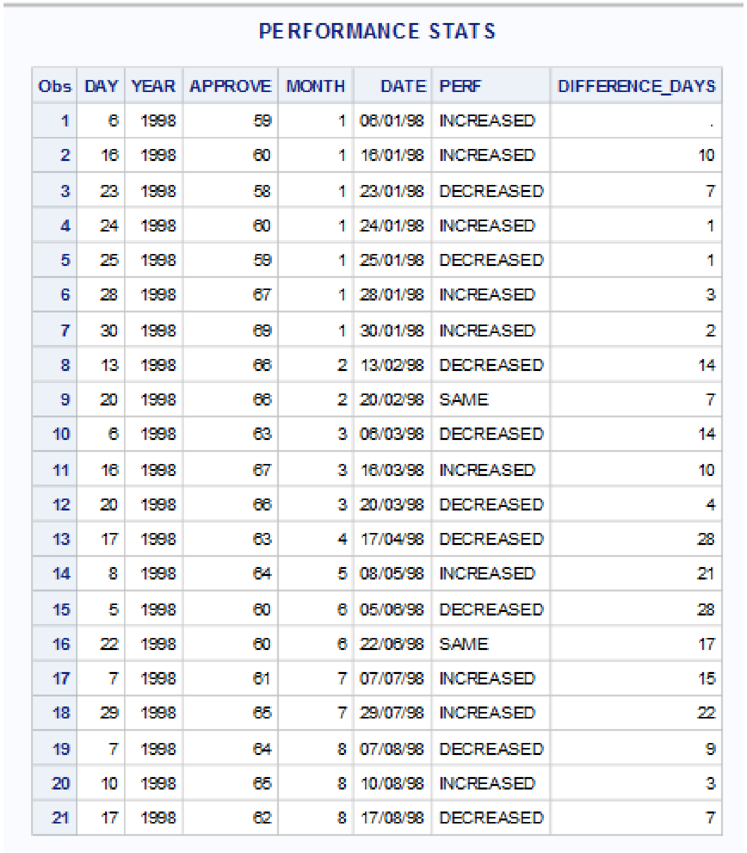


Obs	dogname	Week0	Week2	Week4
1	baby	336	52	295
2	bijou	0	855	344
3	cai	128	520	826
4	cleo	511	375	456
5	cooper	70	800	568
6	elliott	114	270	392
7	georgia	165	276	130
8	jessie	352	567	427
9	lu	470	684	720
10	lucy	92	762	97
11	max	0	1106	0
12	muttney	1176	214	121
13	oreo	320	93	68
14	pandora	855	575	756
15	peewee	249	284	693
16	penelope	240	198	252
17	phoenix	438	372	147
18	princess	85	69	688
19	rhea	204	816	840
20	roxanne	448	2196	3534
21	savannah	392	420	350
22	sheppy	472	168	348
23	simon	840	760	492
24	tanner	180	368	448
25	tj	748	276	670

4. Refer to the CLINTON data. Write a SAS program which reads the data. Using only the polls taken in the year 1998, create a new variable which indicates whether the percentage of people approving of the President’s performance increased, decreased, or stayed the same from the time the last poll was taken. Also, create another variable which indicates the number of days elapsed from the previous poll to the current one. Print the dataset with the new varaibles.

```
DATA clinton;
INFILE "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\clinton.txt"
FIRSTOBS = 2 LRECL=200;
INPUT @5 DAY @9 MO $ @13 YEAR @18 APPROVE @26 DISAPPROVE @37 NO_OPINION;
MONTH = INT(MONTH(INPUT("01"||SUBSTR(MO,1,3)||"2001",DATE9.)));
DROP MO;
RUN;
```

```
DATA CLINTON_PERF;
SET CLINTON;
IF YEAR = 1998;
RUN;
DATA CLINTON_PERF;
SET CLINTON_PERF;
DATE = MDY(MONTH, DAY, YEAR);
FORMAT DATE DDMMYY9.;
DROP DISAPPROVE NO_OPINION;
RUN;
PROC SORT DATA = CLINTON_PERF;
BY DATE;
RUN;
DATA CLINTON_PERF;
SET CLINTON_PERF;
IF LAG(APPROVE) = . THEN PERF = "NO PREV DATA";
IF APPROVE < LAG(APPROVE) THEN PERF = "DECREASED";
IF APPROVE > LAG(APPROVE) THEN PERF = "INCREASED";
IF APPROVE = LAG(APPROVE) THEN PERF = "SAME";
RUN;
DATA CLINTON_PERF;
SET CLINTON_PERF;
DIFFERENCE_DAYS = DATE - LAG(DATE);
RUN;
PROC PRINT DATA = CLINTON_PERF;
TITLE "PERFORMANCE STATS";
      RUN;
```



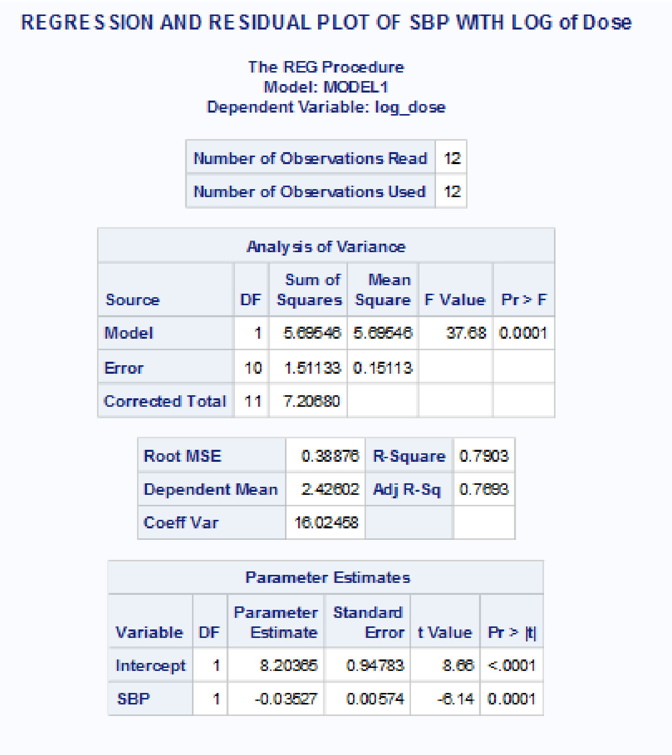
PERFORMANCE STATS

Obs	DAY	YEAR	APPROVE	MONTH	DATE	PERF	DIFFERENCE_DAYS
1	6	1998	59	1	06/01/98	INCREASED	.
2	16	1998	60	1	16/01/98	INCREASED	10
3	23	1998	58	1	23/01/98	DECREASED	7
4	24	1998	60	1	24/01/98	INCREASED	1
5	25	1998	59	1	25/01/98	DECREASED	1
6	28	1998	67	1	28/01/98	INCREASED	3
7	30	1998	69	1	30/01/98	INCREASED	2
8	13	1998	66	2	13/02/98	DECREASED	14
9	20	1998	66	2	20/02/98	SAME	7
10	6	1998	63	3	06/03/98	DECREASED	14
11	16	1998	67	3	16/03/98	INCREASED	10
12	20	1998	66	3	20/03/98	DECREASED	4
13	17	1998	63	4	17/04/98	DECREASED	28
14	8	1998	64	5	08/05/98	INCREASED	21
15	5	1998	60	6	05/06/98	DECREASED	28
16	22	1998	60	6	22/06/98	SAME	17
17	7	1998	61	7	07/07/98	INCREASED	15
18	29	1998	65	7	29/07/98	INCREASED	22
19	7	1998	64	8	07/08/98	DECREASED	9
20	10	1998	65	8	10/08/98	INCREASED	3
21	17	1998	62	8	17/08/98	DECREASED	7

5. Book Chapter 5 numbers 5.10 and 5.12

5.10

```
data tan.dose;
       infile
"\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\dose.txt"
firstobs=2;
       input Dose
                    SBP
                    DBP;
                    log_dose = LOG(Dose);
run;
PROC PRINT data=tan.dose;
RUN;

symbol1 value = dot color = red;

proc reg data = tan.dose;
       title "REGRESSION AND RESIDUAL PLOT OF SBP WITH LOG of Dose";
       model log_dose = SBP;
       plot  log_dose*SBP
                    residual. * SBP;
run;
proc reg data = tan.dose;
       title "REGRESSION AND RESIDUAL PLOT OF DBP WITH LOG of Dose";
       model log_dose = DBP;
       plot  log_dose*DBP
                    residual. * DBP;
RUN;
```



REGRESSION AND RESIDUAL PLOT OF SBP WITH LOG of Dose

The REG Procedure
Model: MODEL1
Dependent Variable: log_dose

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.69546	5.69546	37.68	0.0001
Error	10	1.51133	0.15113		
Corrected Total	11	7.20680			

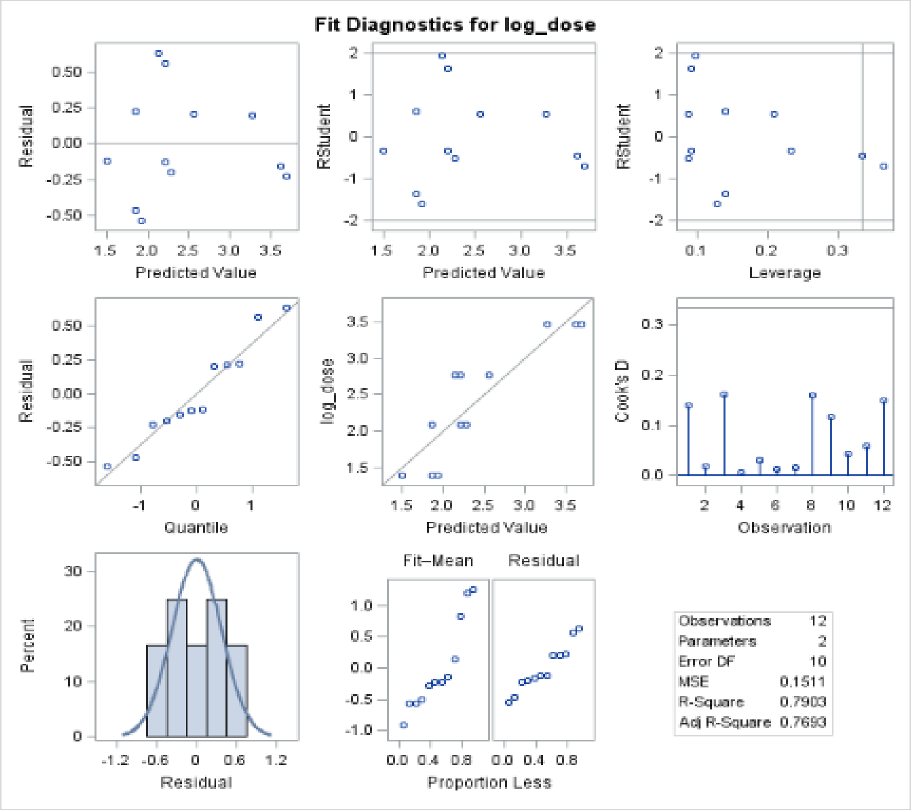
Root MSE	0.38876	R-Square	0.7903
Dependent Mean	2.42602	Adj R-Sq	0.7693
Coeff Var	16.02458		

Parameter Estimates

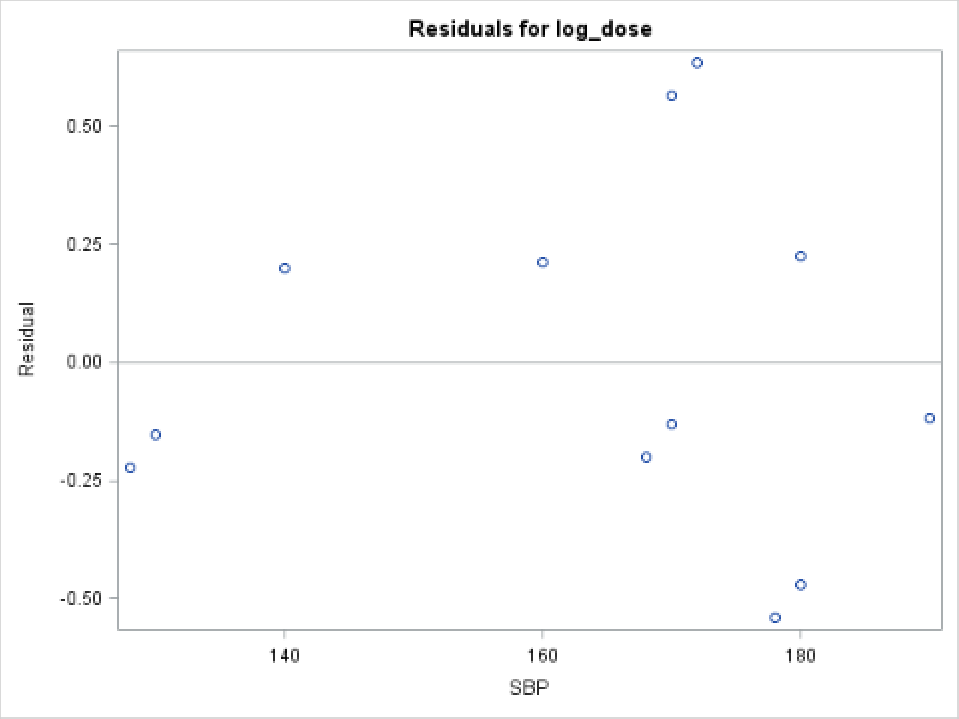
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.20365	0.94783	8.66	<.0001
SBP	1	-0.03527	0.00574	-6.14	0.0001

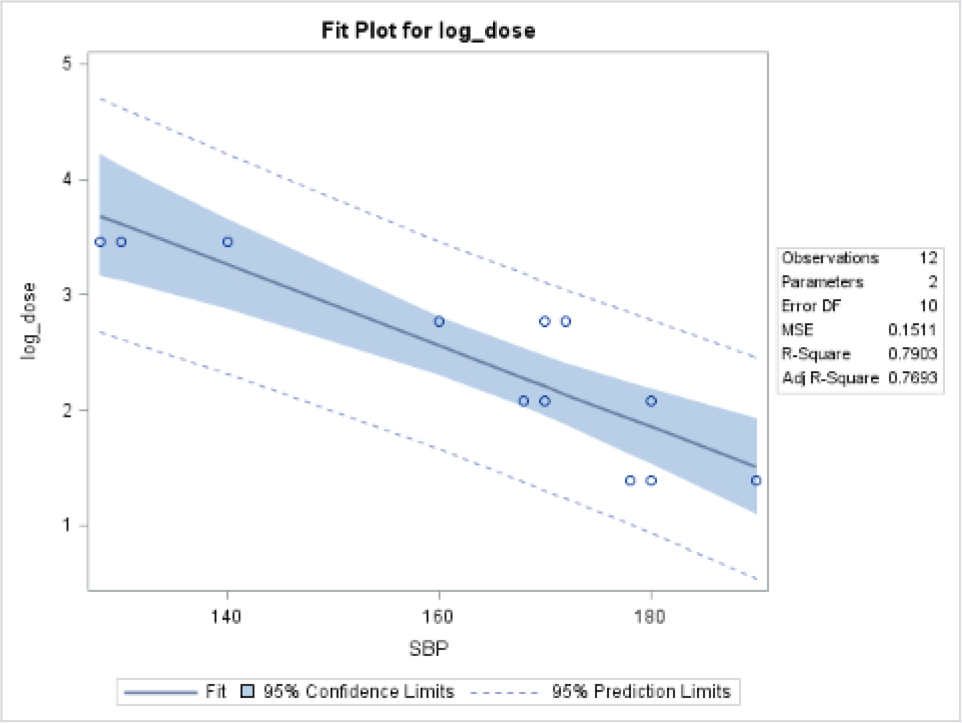
REGRESSION AND RESIDUAL PLOT OF SBP WITH LOG of Dose

The REG Procedure
Model: MODEL1
Dependent Variable: log_dose

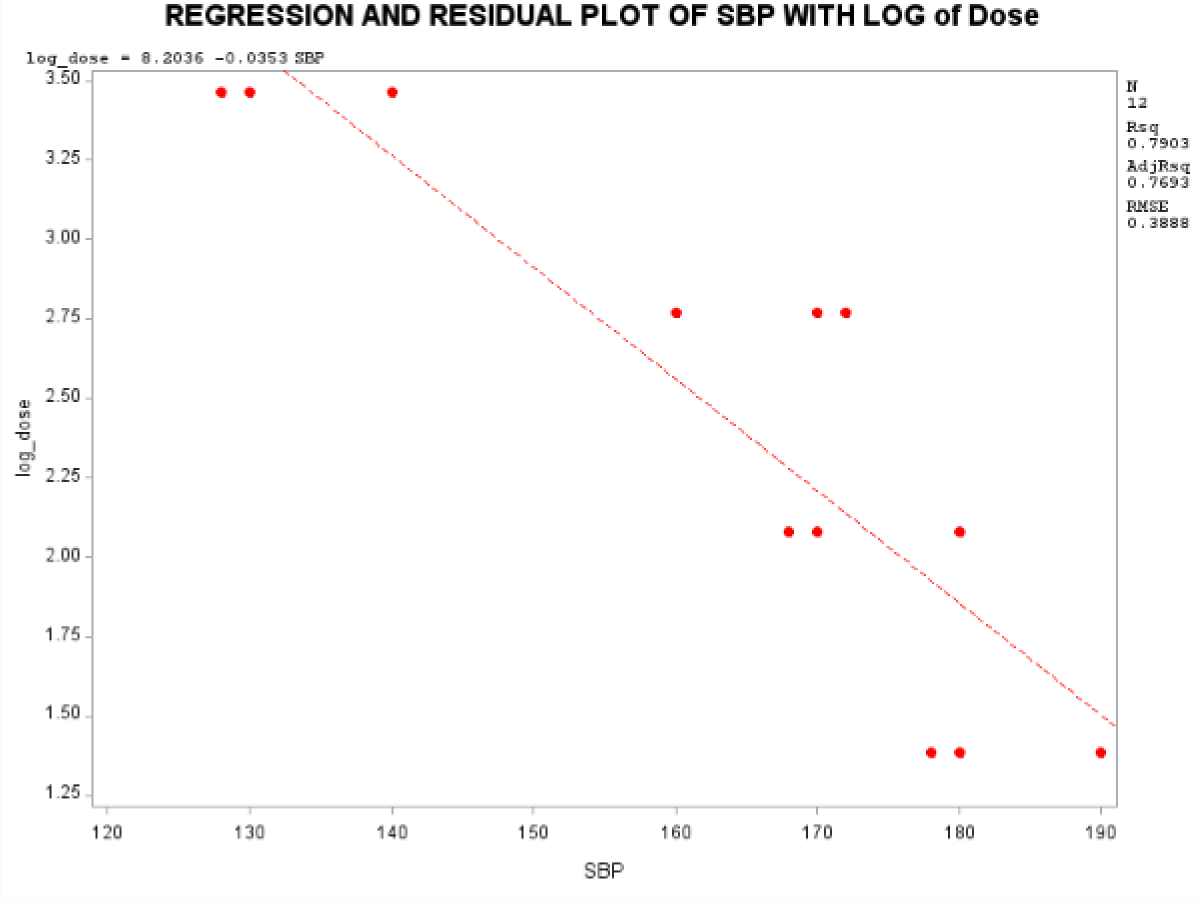


```
Observations        12
Parameters           2
Error DF            10
MSE             0.1511
R-Square        0.7903
Adj R-Square    0.7693
```



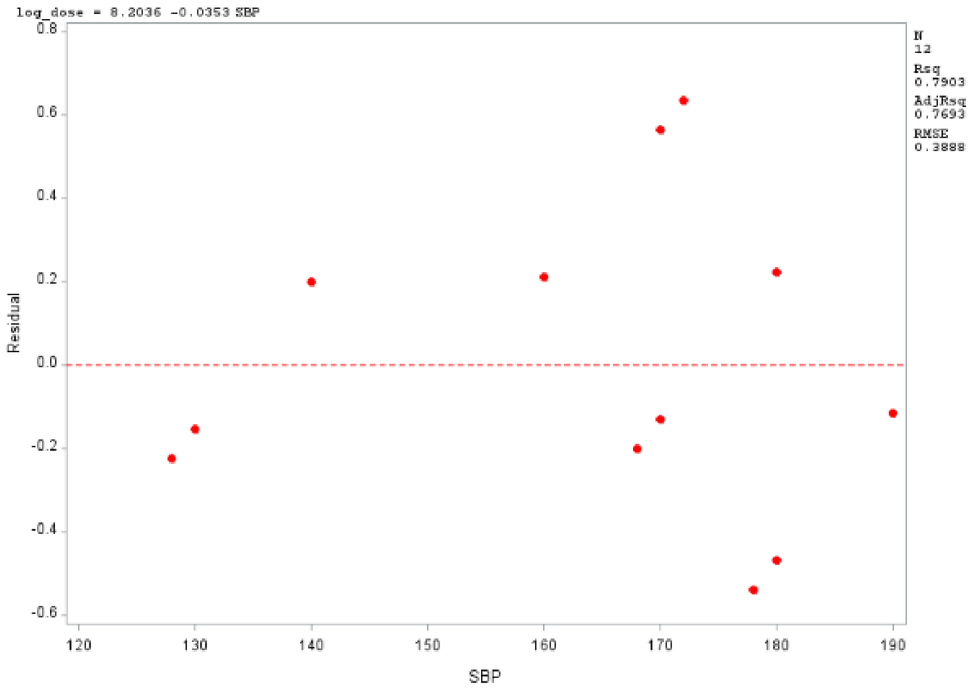


The REG Procedure



The REG Procedure

REGRESSION AND RESIDUAL PLOT OF SBP WITH LOG of Dose



REGRESSION AND RESIDUAL PLOT OF DBP WITH LOG of Dose

The REG Procedure
Model: MODEL1
Dependent Variable: log_dose

Number of Observations Read	12
Number of Observations Used	12

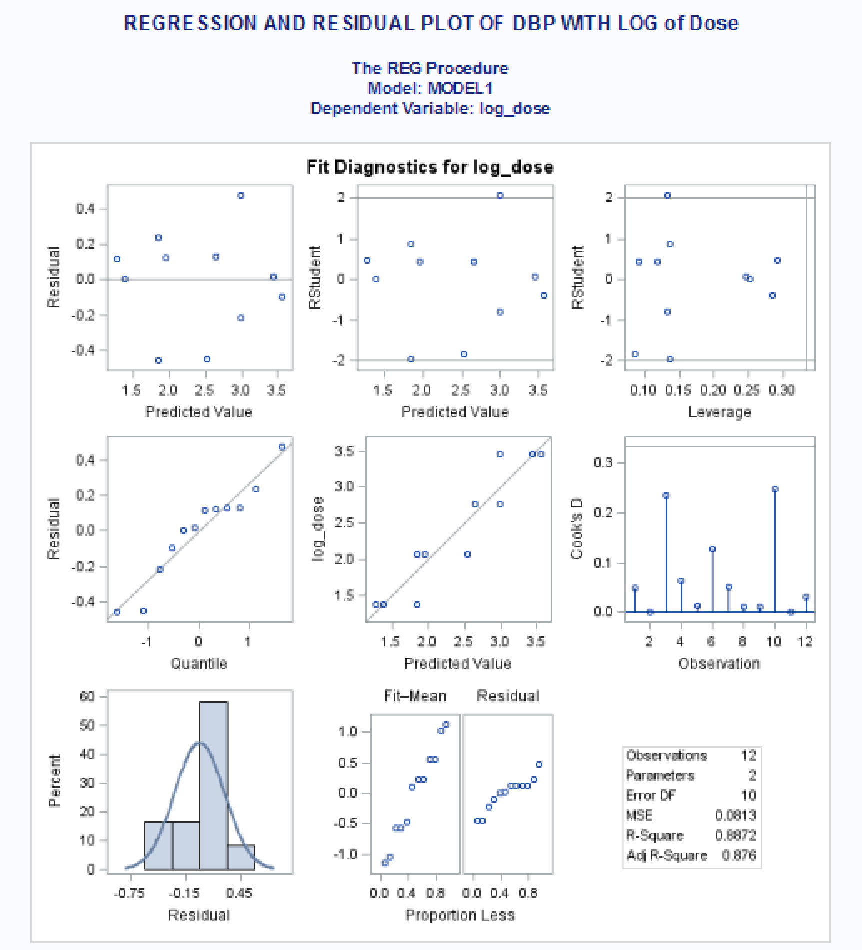
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.39423	6.39423	78.69	<.0001
Error	10	0.81257	0.08126		
Corrected Total	11	7.20680			

Root MSE	0.28506	R-Square	0.8872
Dependent Mean	2.42602	Adj R-Sq	0.8760
Coeff Var	11.74996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.57325	0.58605	12.92	<.0001
DBP	1	-0.05730	0.00646	-8.87	<.0001

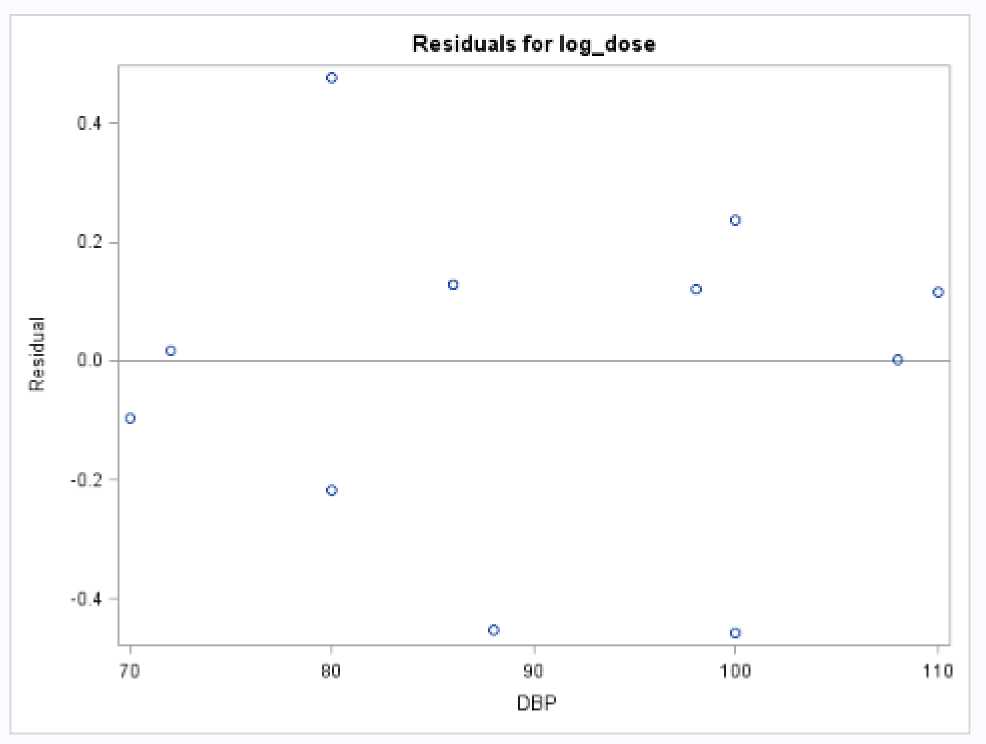


REGRESSION AND RESIDUAL PLOT OF DBP WITH LOG of Dose

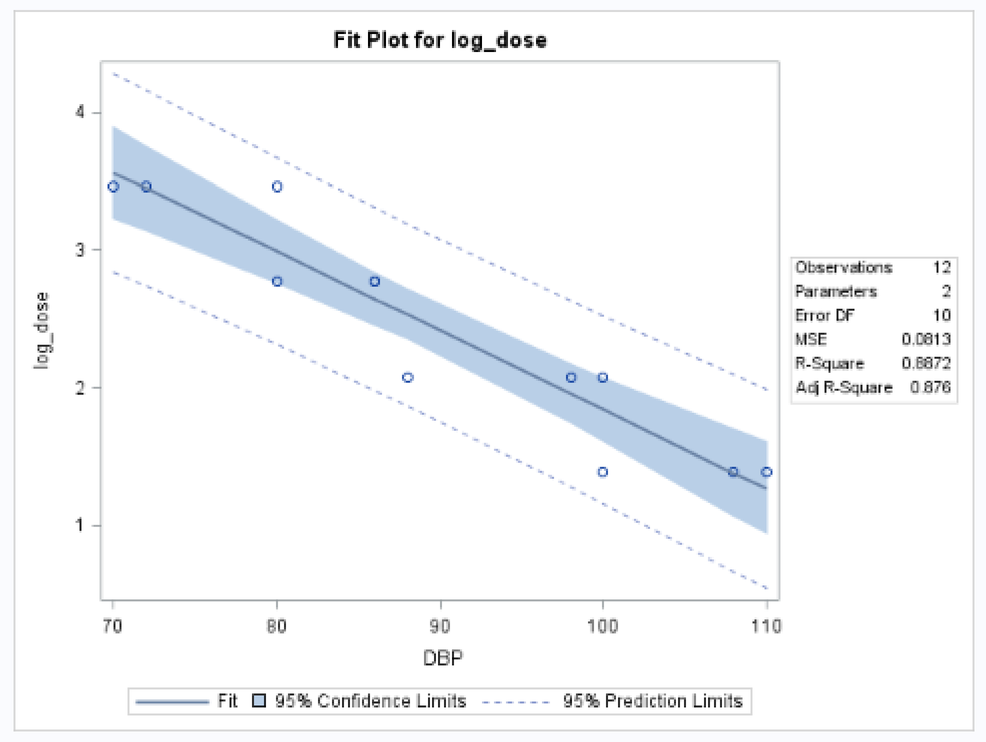
The REG Procedure
Model: MODEL1
Dependent Variable: log_dose

Fit Diagnostics for log_dose

```
Observations      12
Parameters         2
Error DF          10
MSE           0.0813
R-Square      0.8872
Adj R-Square   0.876
```

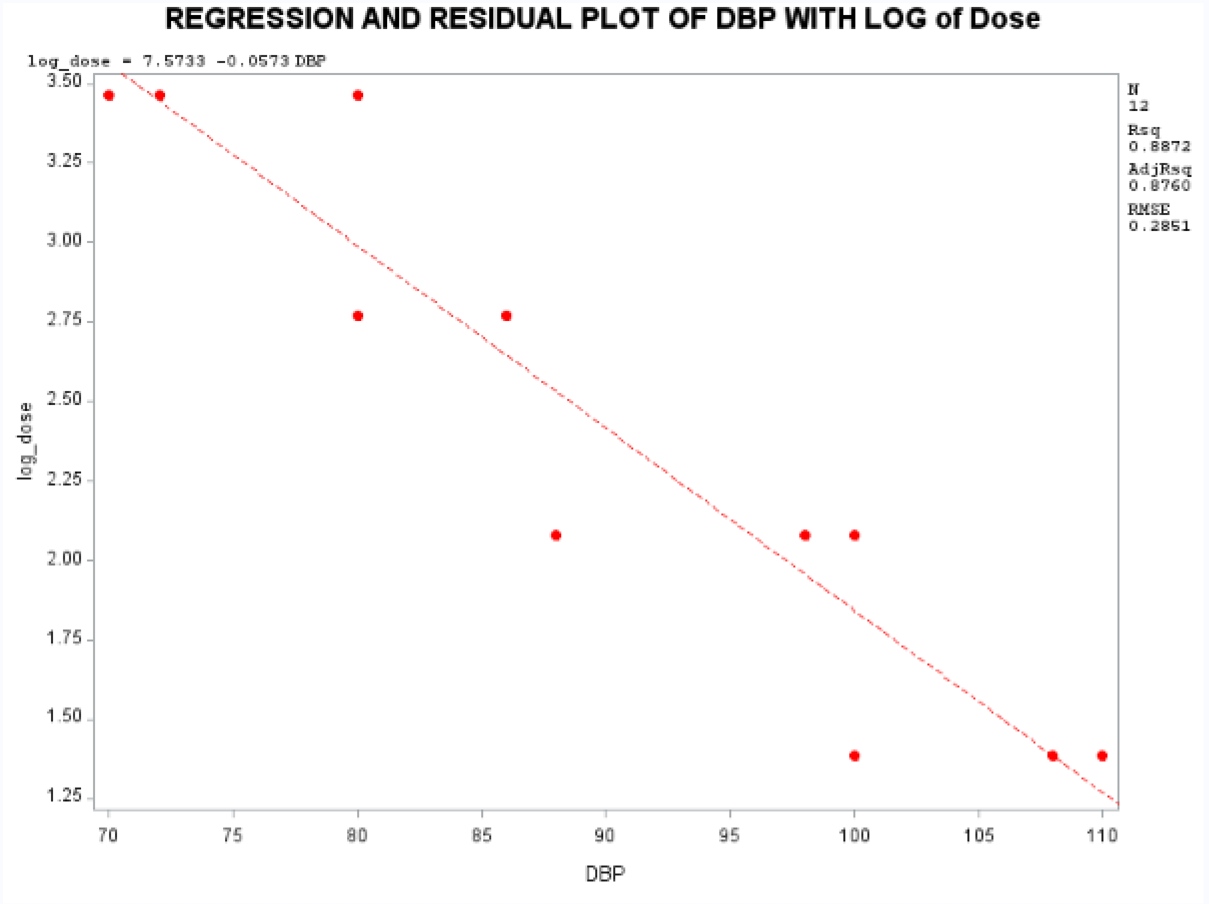


Residuals for log_dose

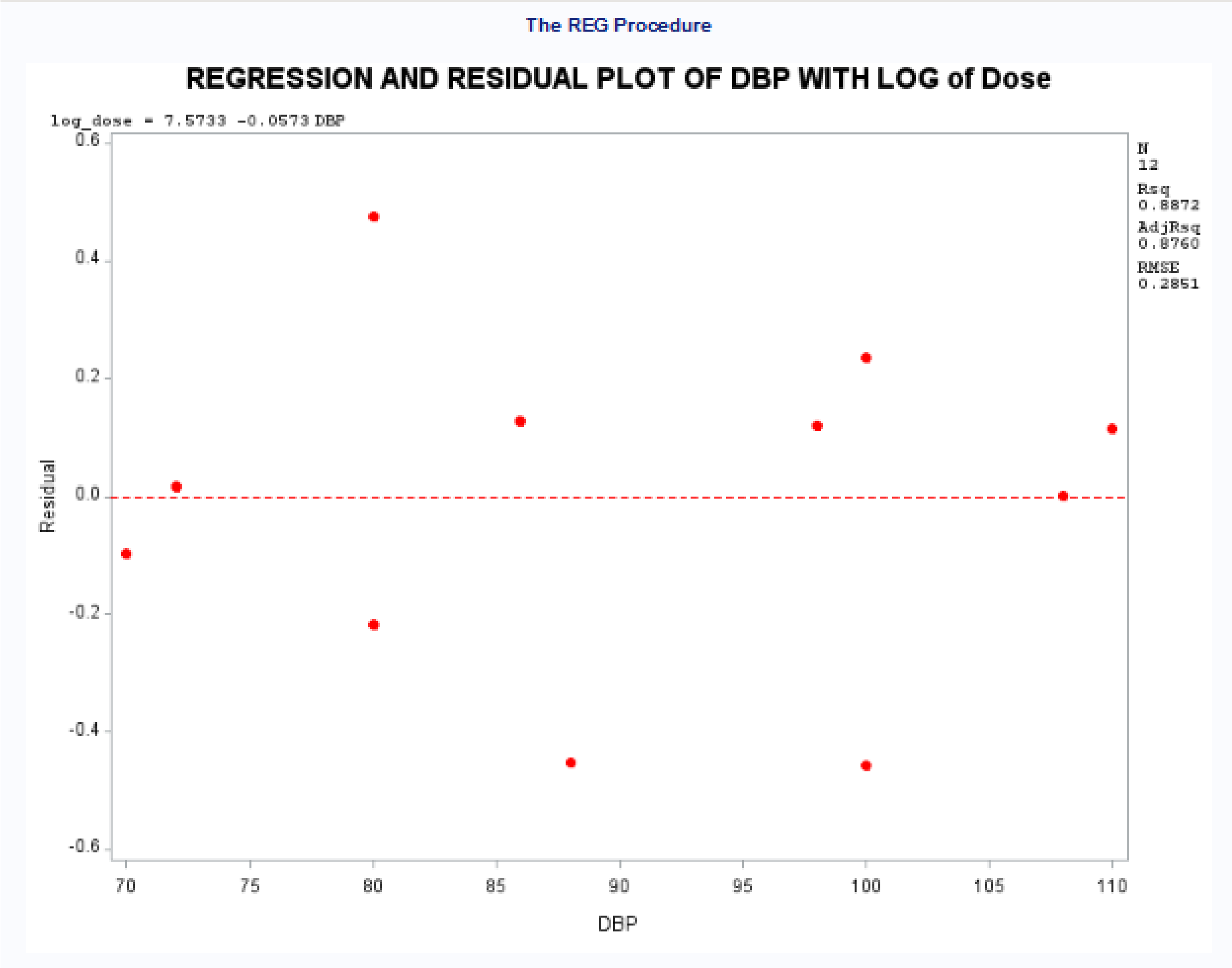


Observations	12
Parameters	2
Error DF	10
MSE	0.0813
R-Square	0.8872
Adj R-Square	0.876

The REG Procedure



N	12
Rsq	0.8872
AdjRsq	0.8760
RMSE	0.2851



The REG Procedure

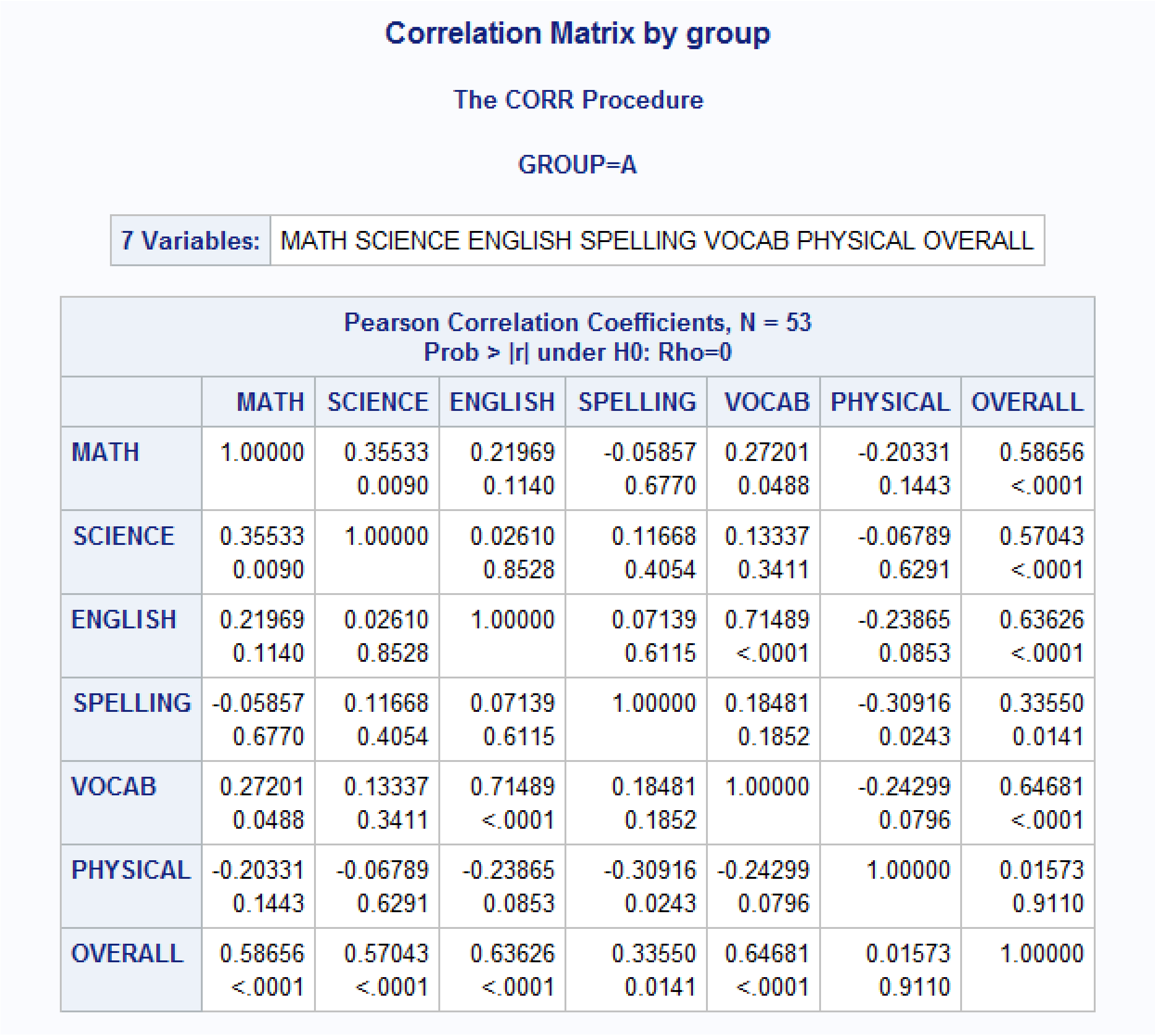
REGRESSION AND RESIDUAL PLOT OF DBP WITH LOG of Dose

Comparing the fit statistics using the plots for SBP and DBP we find that the number of points in the residual plot are dispersed more in DBP as compared to SBP. Also looking at the fit plot for DBP we find the R-squared value to be higher(0.8872) than the R-squared value for SBP(.7903). It is stated that higher the R-Square better the model fits the data. Also the MSE for SBP (0.15) is high compared to MSE for DBP(0.08) which helps us in the statistical estimation.

We can conclude that log of dose is a better fit for DBP than SBP.

5.12

```
DATA tan.SCORES;
   DO SUBJECT = 1 TO 100;
      IF RANUNI(1357) LT .5 THEN GROUP = 'A';
      ELSE GROUP = 'B';
      MATH = ROUND(RANNOR(1357)*20 + 550 + 10*(GROUP EQ 'A'));
      SCIENCE = ROUND(RANNOR(1357)*15 + .4*MATH + 300);
      ENGLISH = ROUND(RANNOR(1357)*20 + 500 + .05*SCIENCE +
               .05*MATH);
      SPELLING = ROUND(RANNOR(1357)*15 + 500 + .1*ENGLISH);
      VOCAB = ROUND(RANNOR(1357)*5 + 400 + .1*SPELLING +
             .2*ENGLISH);
      PHYSICAL = ROUND(RANNOR(1357)*20 + 550);
      OVERALL = ROUND(MEAN(MATH, SCIENCE, ENGLISH, SPELLING, VOCAB,
                     PHYSICAL));
      OUTPUT;
   END;
RUN;

proc sort data=tan.SCORES;
by GROUP;
run;

proc corr data=tan.SCORES nosimple;
      title "Correlation Matrix by group";
      by GROUP;
      var MATH SCIENCE ENGLISH SPELLING VOCAB PHYSICAL OVERALL;
      run;
```



Correlation Matrix by group

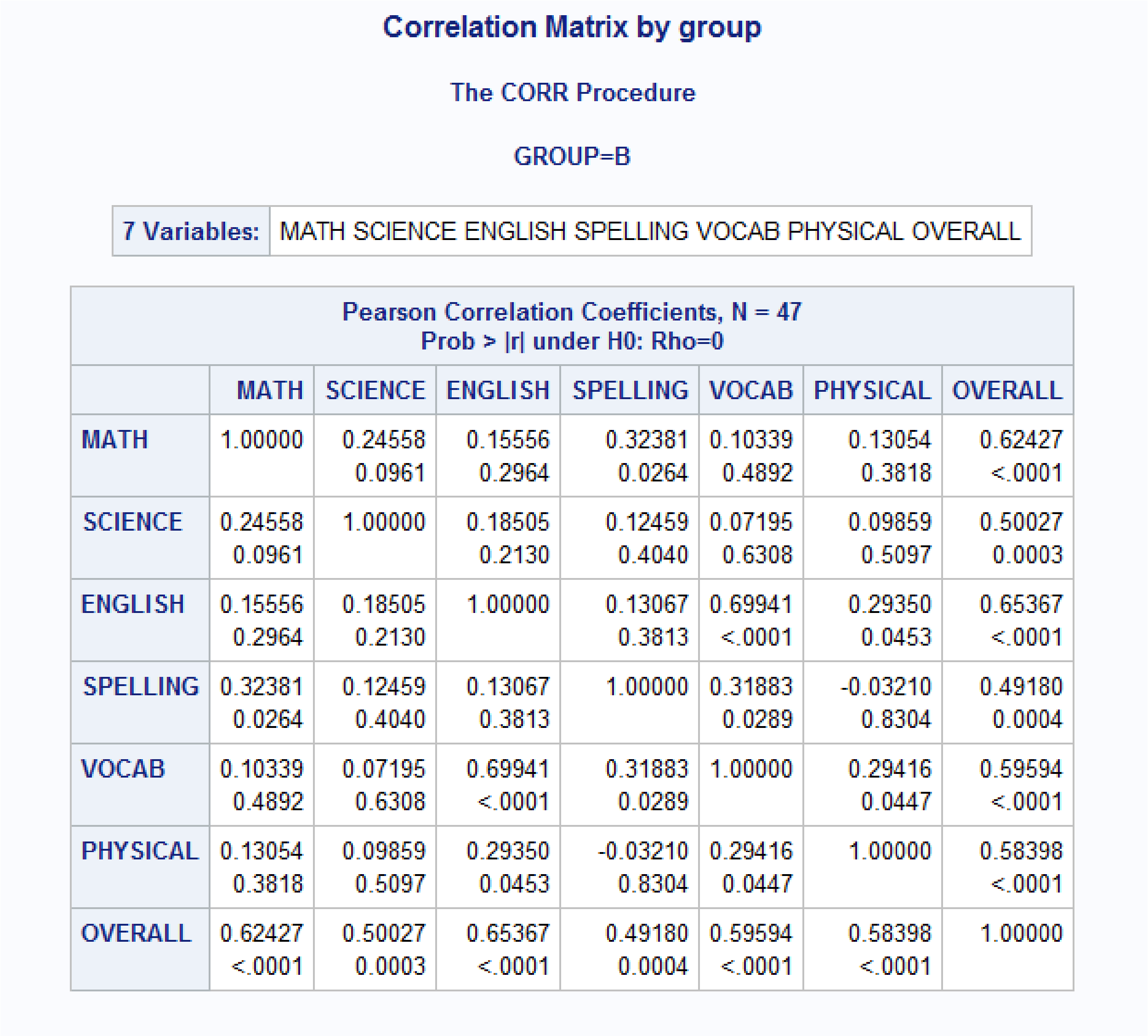
The CORR Procedure

GROUP=A

7 Variables:	MATH SCIENCE ENGLISH SPELLING VOCAB PHYSICAL OVERALL

Pearson Correlation Coefficients, N = 53
Prob > |r| under H0: Rho=0

	MATH	SCIENCE	ENGLISH	SPELLING	VOCAB	PHYSICAL	OVERALL
MATH	1.00000	0.35533 0.0090	0.21969 0.1140	-0.05857 0.6770	0.27201 0.0488	-0.20331 0.1443	0.58656 <.0001
SCIENCE	0.35533 0.0090	1.00000	0.02610 0.8528	0.11668 0.4054	0.13337 0.3411	-0.06789 0.6291	0.57043 <.0001
ENGLISH	0.21969 0.1140	0.02610 0.8528	1.00000	0.07139 0.6115	0.71489 <.0001	-0.23865 0.0853	0.63626 <.0001
SPELLING	-0.05857 0.6770	0.11668 0.4054	0.07139 0.6115	1.00000	0.18481 0.1852	-0.30916 0.0243	0.33550 0.0141
VOCAB	0.27201 0.0488	0.13337 0.3411	0.71489 <.0001	0.18481 0.1852	1.00000	-0.24299 0.0796	0.64681 <.0001
PHYSICAL	-0.20331 0.1443	-0.06789 0.6291	-0.23865 0.0853	-0.30916 0.0243	-0.24299 0.0796	1.00000	0.01573 0.9110
OVERALL	0.58656 <.0001	0.57043 <.0001	0.63626 <.0001	0.33550 0.0141	0.64681 <.0001	0.01573 0.9110	1.00000



Correlation Matrix by group

The CORR Procedure

GROUP=B

7 Variables:	MATH SCIENCE ENGLISH SPELLING VOCAB PHYSICAL OVERALL

Pearson Correlation Coefficients, N = 47 Prob > |r| under H0: Rho=0							
	MATH	SCIENCE	ENGLISH	SPELLING	VOCAB	PHYSICAL	OVERALL
MATH	1.00000	0.24558 0.0961	0.15556 0.2964	0.32381 0.0264	0.10339 0.4892	0.13054 0.3818	0.62427 <.0001
SCIENCE	0.24558 0.0961	1.00000	0.18505 0.2130	0.12459 0.4040	0.07195 0.6308	0.09859 0.5097	0.50027 0.0003
ENGLISH	0.15556 0.2964	0.18505 0.2130	1.00000	0.13067 0.3813	0.69941 <.0001	0.29350 0.0453	0.65367 <.0001
SPELLING	0.32381 0.0264	0.12459 0.4040	0.13067 0.3813	1.00000	0.31883 0.0289	-0.03210 0.8304	0.49180 0.0004
VOCAB	0.10339 0.4892	0.07195 0.6308	0.69941 <.0001	0.31883 0.0289	1.00000	0.29416 0.0447	0.59594 <.0001
PHYSICAL	0.13054 0.3818	0.09859 0.5097	0.29350 0.0453	-0.03210 0.8304	0.29416 0.0447	1.00000	0.58398 <.0001
OVERALL	0.62427 <.0001	0.50027 0.0003	0.65367 <.0001	0.49180 0.0004	0.59594 <.0001	0.58398 <.0001	1.00000

Looking at the Correlation coefficients and the p values from the two matrices A and B. Small values for correlation coefficients indicate weak correlation and negative values indicates an inverse relation.

Observing the values from both the groups we observe the correlation instances to be weaker and more inverse in group A as compared to group B.

For ex. If we see the group B matrix for ENGLISH vs OVERALL we find that they share a strong correlation (0.65367) between them and the p value is <.0001 which makes it significant.