STATISTICAL PROGRAMMING FOR BUSINESS ANALYTICS

ASSIGNMENT NO.7

SUBMITTED BY:

TANAY BHALERAO

U47707491

APRIL 3, 2015

UNIVERSITY OF SOUTH FLORIDA
Management Information Systems

Homework for Chapter 9

1. Refer to the GRADES data. Suppose that the instructor of the class wants to see if students performed at consistent levels during the semester. There would be a problem with the grading procedure if, for example, students who earned high grades at the beginning of the semester tended to have lower grades toward the end of the semester, or if students who performed well in one week performed poorly the next week. One way to evaluate this consistency numerically is to use a split-half reliability coefficient. Choose one way to divide the 13 homework grades into two groups: one with seven assignments, one with six assignments. For example, you may choose to divide the assignment into early and late assignments, odd-number and even-number assignments, or a randomly chosen group of seven and the remaining six assignments. Then calculate the correlation between the total of the first group of assignments and the total of the second group of assignments. Typically, for a grading procedure to be considered 'reliable', this correlation should be 0.7 or higher. Would you conclude that the grading policy is reliable from your calculations?

```
LIBNAME tan "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";
DATA tan.GRADES;
INFILE '\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\grades.txt';
INPUT ID 1-4 G1 6 G2 7 G3 8 G4 9 G5 10 G6 11 G7 12 G8 13 G9 14 G10 15 G11 16
G12 17 G13 18;
      Group1 = G1+G2+G3+G4+G5+G6+G7;
      Group2 = G8+G9+G10+G11+G12+G13;
RUN;

PROC CORR DATA=tan.GRADES;
      Title "Consistency in Grades";
      Var Group1 Group2;
RUN;
```

Consistency in Grades

The CORR Procedure

2 Variables:	Group1 Group2

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Group1	23	49.47826	4.82295	1138	36.00000	55.00000
Group2	23	38.17391	8.32112	878.00000	10.00000	48.00000

Pearson Correlation Coefficients, N = 23 Prob > |r| under H0: Rho=0		
	Group1	Group2
Group1	1.00000	0.17112 0.4350
Group2	0.17112 0.4350	1.00000

The correlation between group1 and group 2 is 0.17112 which is less than 0.7. Hence the grading policies <u>cannot</u> be considered reliable.

2. I was trying out a new bread recipe the other day. I spilled something on the recipe booklet, and I can't read how much flour I'm supposed to use in the recipe. I do know that I need to use 1 cup of water, 2 tablespoons of oil, 2 tablespoons of sugar, 1 ½ teaspoons of salt, and 2 ¼ teaspoons of yeast.

   Help me out. Refer to BREAD data. Find the least-squares regression equation to predict flour amounts from water, oil, sugar, salt, and yeast, and use that equation to estimate how much flour I need in my recipe. Make sure that SAS prints the estimated amount of flour needed.

```
DATA tan.Bread;
INFILE '\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\bread.txt'
LRECL=200 dlm = ',' firstobs = 3;
INPUT dough $ water oil sugar salt dry_milk flour yeast wheat oregano eggs;
RUN;

PROC REG DATA=tan.Bread OUTEST = Flour_score;
TITLE "Regression Scores";
flour_m : MODEL flour = water oil sugar salt yeast;
RUN;
DATA new_ingre;
INPUT water oil sugar salt yeast;
DATALINES;
1 2 2 1.5 2.25
;
RUN;
PROC SCORE DATA=new_ingre SCORE=Flour_score TYPE=parms NOSTD PREDICT
OUT=Flour_predict;
VAR water oil sugar salt yeast;
RUN;
PROC PRINT DATA=Flour_predict;
TITLE "PREDICTION FOR FLOUR QUANTITY";
      RUN;
```

Obs	water	oil	sugar	salt	yeast	flour_m
1	1	2	2	1.5	2.25	3.17483

3. Refer to the USEDCARS data. Calculate the regression line to predict the price of a used car based on the year in which it was manufactured. Obtain the residuals from this regression model, and use the PROC UNIVARIATE to examine their distribution. In the PROC UNIVARIATE output, identify the largest five and smallest five residuals by the name of the used car dealer. Do the residuals appear to be normally distributed, as we assume when conducting the t- and F-tests?

```
DATA tan.USEDCARS;
      INFILE
'\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\usedcars.txt'
FIRSTOBS=2 OBS=51;
      INPUT Year 1-2 Manufacturer $ 9-23 Model $ 24-37 Miles $ 38-48 Price $
49-60 Dealer $ 61-86;
      Price_d = INPUT(Price,comma9.);
RUN;
PROC REG DATA=tan.USEDCARS;
      TITLE "Car Price Prediction";
      MODEL Price_d=Year;
      OUTPUT OUT= USEDCARS_R residual=_N_;
RUN;

PROC UNIVARIATE DATA=USEDCARS_R PLOT NORMAL;
ID Dealer;
VAR _N_;
RUN;
```

Car Price Prediction

The REG Procedure
Model: MODEL1
Dependent Variable: Price_d

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	565089782	565089782	23.17	<.0001
Error	48	1170485556	24385116		
Corrected Total	49	1735575338			

Root MSE	4938.12877	R-Square	0.3256
Dependent Mean	9182.54000	Adj R-Sq	0.3115
Coeff Var	53.77737		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-78803	18291	-4.31	<.0001
Year	1	939.81722	195.23024	4.81	<.0001

The UNIVARIATE Procedure
Variable: _N_ (Residual)

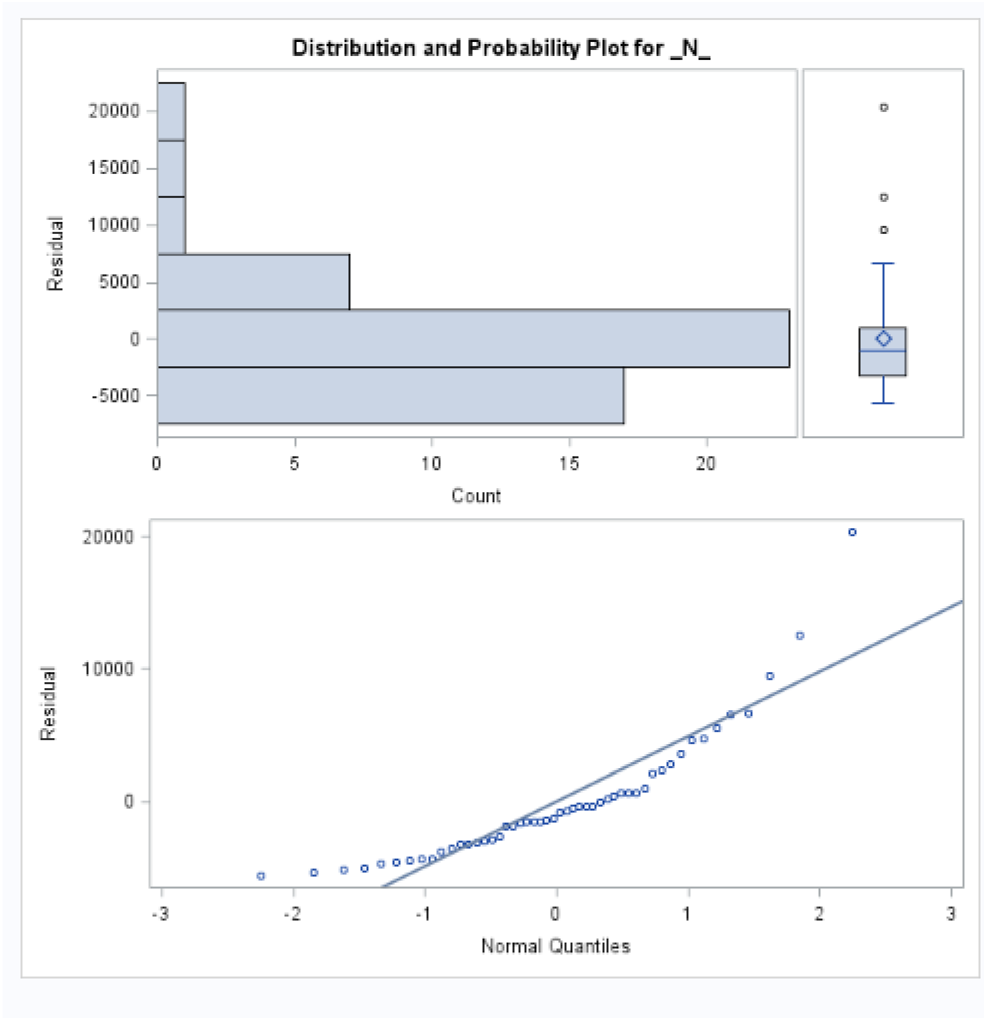
Moments			
N	50	Sum Weights	50
Mean	0	Sum Observations	0
Std Deviation	4887.47996	Variance	23887460.3
Skewness	1.99219857	Kurtosis	5.54907769
Uncorrected SS	1170485556	Corrected SS	1170485556
Coeff Variation	.	Std Error Mean	691.194044

Basic Statistical Measures			
Location		Variability	
Mean	0.00	Std Deviation	4887
Median	-1036.04	Variance	23887460
Mode	-424.30	Range	25955
		Interquartile Range	4209

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-7	Pr >= |M|	0.0649
Signed Rank	S	-125.5	Pr >= |S|	0.2293

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.830673	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.185588	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.351074	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.072583	Pr > A-Sq	<0.0050

Extreme Observations					
Lowest			Highest		
Value	Dealer	Obs	Value	Dealer	Obs
-5604.85	Magic Imports	41	6579.70	Santa Fe Ford	46
-5371.12	Budget Car Sales	27	6601.06	Tomlinson Motor Company	42
-5144.67	Bush Gator	22	9515.51	Taylor Volvo	39
-4984.49	Bush Gator	37	12515.51	Gatorland Toyota	14
-4665.04	Magic Imports	47	20350.15	Kraft Motorcar	10



The residuals are not normally distributed as seen from the plots. Also the difference between mean and median is large which again conforms that the residuals are not normally distributed.

4. Refer to the HANKS data. During Tom Hanks's career, he has played both humorous and dramatic roles. Over time, has he increasingly accepted serious roles over lighter ones? To see if this is true, find the correlations of the length of the movie, the drama rating, and the humor rating with year. (The ratings are ordinal, so the correlations of year with humor and drama must not be interpreted too rigidly. Instead, they give a rough indication of positive or negative trends with time.)

```
DATA tan.Hanks;
      INFILE
'\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\Hanks.txt' FIRSTOBS
= 2;
      INPUT Title $ 1-25 Year 26-30 Length 34-37 MPAA $ 42-47 Action 50-52
Drama 58-60 Humor 66-68 Sex 74 Violence 81-83 Suspense 90 Offbeat 98;
RUN;
PROC CORR DATA = tan.Hanks;
      TITLE "TRENDS WITH TIME";
      VAR Length Drama Humor;
```

```
    WITH year;
RUN;
```

TRENDS WITH TIME

The CORR Procedure

1 With Variables	Year
3 Variables	Length Drama Humor

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Year	22	1990	4.22449	43773	1984	1998
Length	22	110.77273	18.30153	2437	80.00000	160.00000
Drama	21	4.85714	2.66994	102.00000	1.00000	10.00000
Humor	21	5.76190	1.84132	121.00000	2.00000	10.00000

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	Length	Drama	Humor
Year	0.48867 0.0210 22	0.55815 0.0086 21	-0.10456 0.6519 21

5. Chapter 9: 9.4, 9.10

9.4

```
DATA LIBRARY;
INPUT Books Student_Enrollment Highest_Degree Area;
Log_Area = Log(Area);
If Highest_Degree in (1 2 3) then do;
MA = (Highest_Degree EQ 2);
PhD = (Highest_Degree EQ 3);
END;

DATALINES;
4   5   3   20
5   8   3   40
10  40  3   100
1   4   2   50
.5  2   1   300
2   8   1   400
7   30  3   40
4   20  2   200
1   10  2   5
1   12  1   100
;
PROC REG DATA = LIBRARY;
TITLE "Estimate number of books";
MODEL Books = Student_Enrollment MA PhD Log_Area / SELECTION = Forward;
```

```
RUN;
```

Variable MA Entered: R-Square = 0.9864 and C(p) = 5.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	86.03915	21.50979	90.69	<.0001
Error	5	1.18585	0.23717		
Corrected Total	9	87.22500			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2.91829	0.94453	2.26404	9.55	0.0272
Student_Enrollment	0.13128	0.01635	15.29338	64.48	0.0005
MA	1.35800	0.52811	1.56823	6.61	0.0499
PhD	4.54244	0.56785	15.17652	63.99	0.0005
Log_Area	0.57463	0.17335	2.60614	10.99	0.0211

Bounds on condition number: 3.263, 37.165

All variables have been entered into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Student_Enrollment	1	0.7428	0.7428	88.5868	23.11	0.0013
2	PhD	2	0.2129	0.9557	12.2968	33.63	0.0007
3	Log_Area	3	0.0127	0.9684	9.6123	2.42	0.1708
4	MA	4	0.0180	0.9864	5.0000	6.61	0.0499

9.10

```
PROC FORMAT;
      VALUE YESNO 1='YES' 0='NO';
      VALUE OUTCOME 1='Case' 0='Control';
RUN;
DATA SMOKING;
      DO SUBJECT = 1 TO 1000;
      DO OUTCOME = 0,1;
            IF RANUNI(567) LT .1 OR RANUNI(0)*OUTCOME GT .5 THEN
                  SMOKING = 1;
            ELSE SMOKING = 0;
            IF RANUNI(0) LT .05 OR
                  (RANUNI(0)*OUTCOME + .1*SMOKING) GT .6 THEN ASBESTOS = 1;
            ELSE ASBESTOS = 0;
```

```
            IF RANUNI(0) LT .3 OR OUTCOME*RANUNI(0) GT .9 THEN
                  SES = '1-Low    ';
            ELSE IF RANUNI(0) LT .3 OR OUTCOME*RANUNI(0) GT .8 THEN
                  SES = '2-Medium';
            ELSE SES = '3-High';
            OUTPUT;
      END;
END;
FORMAT SMOKING ASBESTOS YESNO. OUTCOME OUTCOME.;
RUN;
PROC LOGISTIC DATA = SMOKING;
      TITLE "Logistic Regression";
      CLASS SES (PARAM = Ref REF ='2-Medium');
      MODEL OUTCOME = SMOKING ASBESTOS SES;
RUN;
```

Logistic Regression

The LOGISTIC Procedure

Model Information	
Data Set	WORK.SMOKING
Response Variable	OUTCOME
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	2000
Number of Observations Used	2000

Response Profile

Ordered Value	OUTCOME	Total Frequency
1	Case	1000
2	Control	1000

Probability modeled is OUTCOME='Case'.

Class Level Information

Class	Value	Design Variables	
SES	1-Low	1	0
	2-Medium	0	0
	3-High	0	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2774.589	1946.815
SC	2780.190	1974.819
-2 Log L	2772.589	1936.815

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	835.7741	4	<.0001
Score	696.0566	4	<.0001
Wald	490.5215	4	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
SMOKING	1	277.2260	<.0001
ASBESTOS	1	232.7040	<.0001
SES	2	27.5915	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.9494	0.1179	64.7902	<.0001
SMOKING		1	2.2050	0.1324	277.2260	<.0001
ASBESTOS		1	2.5606	0.1679	232.7040	<.0001
SES	1-Low	1	0.1275	0.1484	0.7381	0.3903
SES	3-High	1	-0.5203	0.1431	13.2191	0.0003

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SMOKING	9.070	6.997	11.758
ASBESTOS	12.944	9.315	17.986
SES 1-Low vs 2-Medium	1.136	0.849	1.519
SES 3-High vs 2-Medium	0.594	0.449	0.787

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.7	Somers' D	0.665
Percent Discordant	12.1	Gamma	0.733
Percent Tied	9.2	Tau-a	0.333
Pairs	1000000	c	0.833