Assignments for Lesson 10

Note: You can find most of the answers by listening to the lectures.

1. What does the Pearson correlation coefficient describe and what is the range of values it can take on?

The Pearson correlation coefficient is basically a formula that describes a strength of a linear relationship between numeric variables. The Pearson correlation attempts to draw a the best fit lines between the data of two variables. The range of values are usually fall between +1 to -1 . The 0 usually indicates that there is no connection between two variables. The positive one indicates that as one values tends to increase the other also follows the same trend. However the value less than 0 indicates that value of one variable may increase while the value of another variable would decrease. For example if all the values are falling one the line it means that the the pearson correlation is equal to 1, they basically follow a straight line.

2. In linear regression, the equation has the form $Y=b_0 + b_1X_1 +b_2X_2 +...+b_kX_k +e$, what is e and what is its distribution?

e represents a random error term which by default is assumed to be normally distributed with a mean 0 constant variance which does not depend on the value of any other observation. For example in the linear regression of stock prices the error term may mean the difference between the actual price at a particular time versus the expected price of stock that was observed.

3. In this regression model what is Y and what is X?

h2s would be the X and taste is Y. Y is a dependant variable followed by a predictor x or in other terms independent variable.

```
/*Regression Line*/
proc reg data=cheddar;
 model taste=h2s;
run;
```

4. Given the following data:

X	Y	Z
1	3	15
7	13	7
8	12	5
3	4	14
4	7	10

(a) Write a SAS program to compute the Pearson correlation coefficient between X and Y; x and Z. What is the significance of each?

The correlation between x and y is 0.96509 which is very close to 1 which means they have a strong positive linear relationship . The relationship is positive means that as X increase the Y is also increases. The p value of 0.0078 value is also small. Therefore x and y has a significant positive relationship between each other. A low p value would allow us to reject the null hypothesis.

The correlation between X and Z is -0.97525 which means they have a strong negative linear relationship. Which means that as X increase the Z decreases. The p value which is 0.0047 is also small . Therefore x and z has a strong negative linear relationship. We can reject null hypothesis.

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
61
62         data correlationtest;
63         input @1 X 2. @3 Y 2. Z 2.;
64         datalines;

NOTE: The data set WORK.CORRELATIONTEST has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

70         ;

71         run;
72
73         proc corr data=correlationtest;
74         var X;
75          with Y Z;
76         run;

NOTE: PROCEDURE CORR used (Total process time):
      real time           0.08 seconds
      cpu time            0.07 seconds
```

The CORR Procedure

2 With Variables:	Y Z
1 Variables:	X

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	5	7.80000	4.54973	39.00000	3.00000	13.00000
Z	5	10.20000	4.32435	51.00000	5.00000	15.00000
X	5	4.60000	2.88097	23.00000	1.00000	8.00000

Pearson Correlation Coefficients, N = 5
Prob > |r| under H0: Rho=0

	X
Y	0.96509 0.0078
Z	-0.97525 0.0047

```
data correlationtest;
input @1 X 2. @3 Y 2. Z 2.;
datalines;
1    3    15
7    13   7
8    12   5
3    4    14
4    7    10
;
run;
```

```
proc corr data=correlationtest;
var X;
 with Y Z;
run;
```

(b) Change the correlation request to produce a correlation matrix; that is, the correlation coefficient between each variable versus every other variable.

The CORR Procedure

3 Variables:	X Y Z

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X	5	4.60000	2.88097	23.00000	1.00000	8.00000
Y	5	7.80000	4.54973	39.00000	3.00000	13.00000
Z	5	10.20000	4.32435	51.00000	5.00000	15.00000

Pearson Correlation Coefficients, N = 5 Prob > |r| under H0: Rho=0			
	X	Y	Z
X	1.00000	0.96509 0.0078	-0.97525 0.0047
Y	0.96509 0.0078	1.00000	-0.96317 0.0084
Z	-0.97525 0.0047	-0.96317 0.0084	1.00000

```
62          data correlationtest;
63          input @1 X 2. @3 Y 2. Z 2.;
64          datalines;

NOTE: The data set WORK.CORRELATIONTEST has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

70          ;

71          run;
72
73          proc corr data=correlationtest;
74          var x y z;
75          run;

NOTE: PROCEDURE CORR used (Total process time):
      real time           0.07 seconds
      cpu time            0.07 seconds
```

```
data correlationtest;
input @1 X 2. @3 Y 2. Z 2.;
datalines;
1    3    15
7    13   7
8    12   5
3    4    14
4    7    10
;
run;
```

```
proc corr data=correlationtest;
var x y z;
run;
```

5. 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.

```
NOTE: WORK.RECIPES data set was successfully created.
NOTE: The data set WORK.RECIPES has 11 observations and 11 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.06 seconds
      cpu time            0.06 seconds


116
117
118         proc reg data=recipes;
119         model flour = water oil sugar salt yeast;
120         run;

121
122

NOTE: PROCEDURE REG used (Total process time):
      real time           1.05 seconds
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.91244	0.56463	3.39	0.0195
water	1	0.26825	0.59587	0.45	0.6714
oil	1	0.00110	0.08216	0.01	0.9898
sugar	1	0.19465	0.14857	1.31	0.2471
salt	1	-0.07809	0.31857	-0.25	0.8161
yeast	1	0.31990	0.24479	1.31	0.2482

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

```
PROC IMPORT OUT=RECIPES
            DATAFILE="/folders/myshortcuts/sas/bread.txt"
            DBMS=CSV REPLACE;
            GETNAMES=YES;
            DATAROW=3;
RUN;
```

```
proc reg data=recipes;
model flour = water oil sugar salt yeast;
run;
```

```
data newrecipe;
input water oil sugar salt yeast;
flour = 1.91244 + 0.26825 * water + 0.00110 * oil + 0.19465 * sugar
- 0.07809 * salt + 0.31990 * yeast;
datalines;
1 2 2 1.5 2.25
;
run;
```

```
proc print data=newrecipe;
run;
```