Lab 5: ANOVA, Linear Regression (10 pts. + 3 pts. Bonus)

Objectives

- **Part 1: ANOVA**
  - **1.1) One-way ANOVA**
  - **1.2) Residual Plot**
  - **1.3) Two-Way ANOVA (BONUS)**
- **Part 2: Linear Regression**
  - **2.1) Scatterplot**
  - **2.2) Correlation Coefficient and a Scatterplot**
  - **2.3) Linear Regression**

Remember:
a) Please put your name, STAT 503 and the lab # on the front of the lab
b) Label each part and put them in logical order.
c) ALWAYS include your SAS code for each problem.

1. ANOVA:

1.1 One-Way ANOVA

For ANOVA, we use the process 'glm'. Again, the data will have the quantity in one column and which categorical variable it refers to in the other. For the learning code, I am using the Weight Gain of Lambs data in several examples in chapter 11.

```
data lambs;
infile 'H:\e1.dat';
input diet gain;
*note that diet = 1 or 2 or 3;
run;

proc print data=lambs; run;

Title 'One-Way ANOVA';
*The categorical variable is in the class statement;
*the model statement is numeric variable = categorical variable;
proc glm data=lambs alpha=0.05;
class diet;
model gain = diet;
run;

quit;
```

SAS Learning output:

```
Dependent Variable: gain

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        2      36.0000000      18.0000000       0.77    0.4907
Error                        9     210.0000000      23.3333333
Corrected Total             11     246.0000000

                  R-Square     Coeff Var      Root MSE     gain Mean
                  0.146341      37.15738      4.830459      13.00000
```

To compare this output to the tables in the book Table 11.2.3 (p.426) and example 11.4.1 (p.430); Between diets is the same as Model and Within diets is the same as Error.

Problem 1 (2 pts.)

A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation. Patient records in the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful completion of physical therapy and the prior physical fitness status (below average, average, above average) for each patient are in the data file fitness.dat (the first column has the fitness status and the second column has the days required). Perform a one-way ANOVA analysis (hypothesis test) in this situation.

Your submission should consist of the code and relevant parts of the output in addition to the ANOVA table and all of the steps of the hypothesis test. No work is required, just provide the information from the SAS output.

1.2 Residual Plot

It is always important to check assumptions. In ANOVA (and linear regression), these assumptions can be assessed by looking at the residuals.

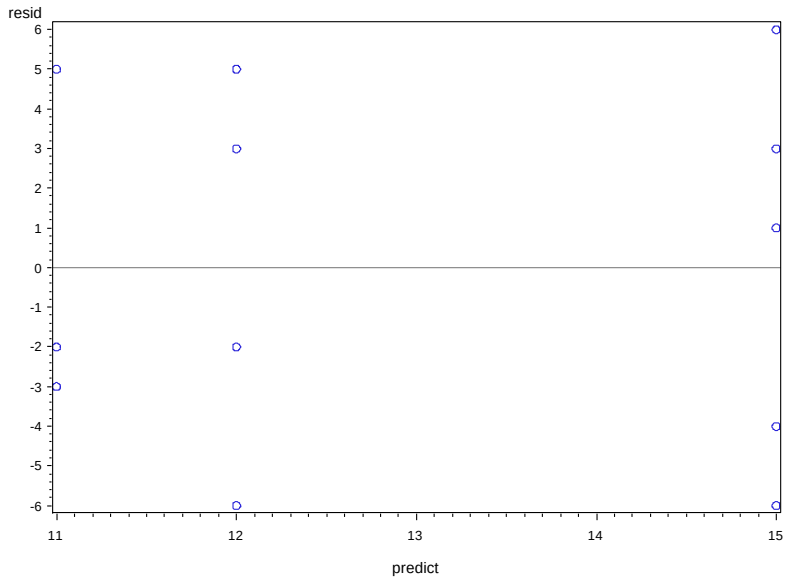
In ANOVA, one assumes that the residuals are independent and normally distributed with constant variance. Independence is often difficult to assess but if the experiment is performed through a sequence of individual runs, one could plot the residuals versus the run order to make sure there is no pattern of positive and negative residuals. Also, if the experiment was done in a field, one could plot the residuals based on location to see if there is any sort of clustering of negative or positive residuals. One usually hopes this assumption is satisfied through proper randomization. To check normality, one can generate histograms, boxplots or Normal QQ plots of the residuals. To check constant variance, a residual plot is very helpful.

In one-factor ANOVA, the predicted value is the sample mean so the residual is $x - \bar{x}$. Using the same data as in 1.1, the following code shows how to generate this type of plot in SAS.

SAS Learning code: (e2.sas)

```
data lambs;
infile 'H:\e1.dat';
input diet gain;
run;

proc print data=lambs; run;

Title 'Residual Plot';
proc glm data=lambs alpha=0.05;
class diet;
model gain = diet;
*the output statement creates a new data file which contains the residuals
  using the keyword r. I am calling this resid. The predicted values are
  printed out using the keyword p. I am calling them predict. Note: the
  output statement is identical in proc glm and proc reg (used for linear
  regression;
output out=resid r=resid p=predict;
run;

symbol1 v=circle;
*the plot statement is y*x, I want a reference line at y = 0.;
proc gplot data=resid;
  plot resid*predict/vref=0;
run;

quit;
```

SAS Learning graphics:



Problem 2 (1 pts.)

For the same problem as described in Problem 1, please create the residual plot. We are trying to determine if the variances are approximately the same for the three physical fitness statuses and if there any outliers.

Your submission should consist of the code and the residual plot. In addition, please indicate if the variances look approximately the same and if there are any outliers.

1.3 Two-Way ANOVA (BONUS)

The only difference between one-way ANOVA and two-way ANOVA is how you state the variables in proc glm. The term with the '*' is the interaction term. This would be included for two-way ANOVA (11.7) but excluded when using the randomized blocks design (11.6). For the learning code, I am using the Iron Supplements example (11.7.3, 11.7.4) in Chapter 11.

SAS Learning code: (e1a.sas)

```
data iron;
infile 'H:\e1a.dat';
input retention iron$ zinc$;
run;

proc print data=iron; run;

Title 'Two-Way ANOVA';
*The categorical variables are in the class statement;
*the iron*zinc term is the interaction term;
proc glm data=iron alpha=0.05;
class iron zinc;
model retention=iron zinc iron*zinc;
run;

quit;
```

SAS Learning output:

```
                                The GLM Procedure
Dependent Variable: retention

                                           Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3      6.06863649      2.02287883    1082.46    <.0001
Error                       28      0.05232581      0.00186878
Corrected Total             31      6.12096230

                  R-Square     Coeff Var      Root MSE    retention Mean
                  0.991451      5.198064      0.043229          0.831644

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
iron                         1      4.40228628      4.40228628    2355.70    <.0001
zinc                         1      0.01087813      0.01087813       5.82    0.0226
iron*zinc                    1      1.65547208      1.65547208     885.86    <.0001
```

Look in the lower table (red) to compare the values from the SAS output to Table 11.7.4 in the book where Between Fe levels is iron, Between Zn levels is zinc and Interaction is iron*zinc. The within groups is the Error in blue. The total is in green. We do not use the Model row in the top table.

Problem B1 (2 pts.)

This is based from exercises 11.7.1 and 11.7.2 in the book. The data is in birch.dat where the concentration of ATP is the first column, whether it is River Birch or European Birch is in the second column and whether it is Flooded or Control is in the third column. Perform a two-way ANOVA analysis (hypothesis test) in this situation.

Your submission should consist of the code and relevant parts of the output in addition to the relevant parts of the ANOVA table (see Table 11.7.7). Please indicate whether the interaction term is significant and whether each of the main effects (type of Birch, flooding) are significant and why (appropriate P-values).

2. Linear Regression:

The learning code for the linear regression is based on the Arsenic in Rice example which starts on p.481. In data file e3.dat, the concentration of silicon in straw is the first variable (X) and the concentration of arsenic in the polished rice is the second variable (Y).

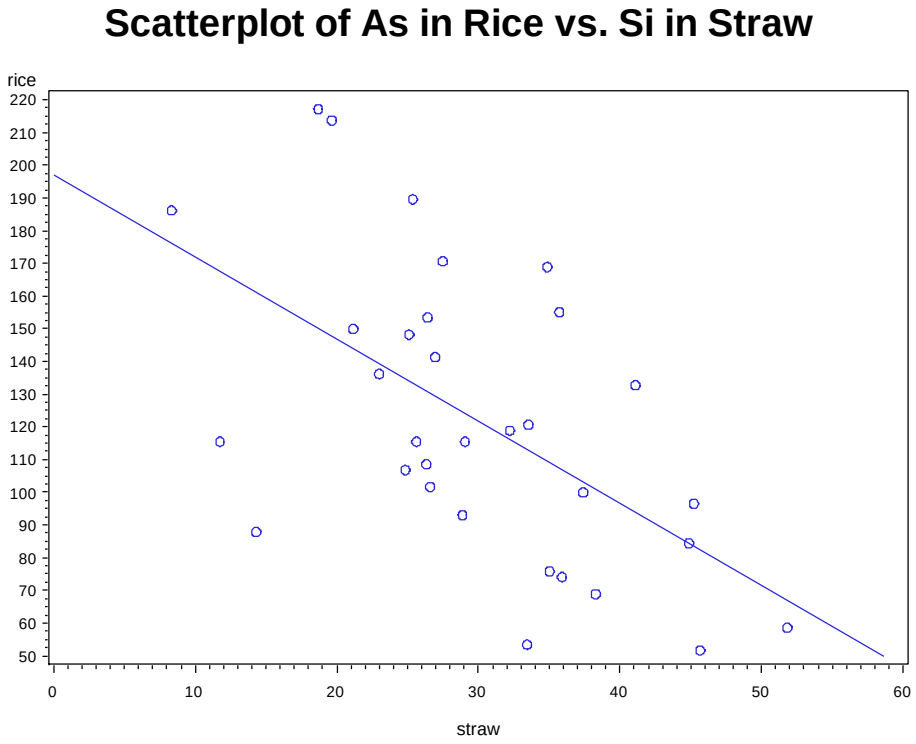
2.1. Scatterplot

The scatterplot is a visual presentation of the data. There are two ways to make a scatterplot is SAS, using the plotting process, proc gplot and when we perfrom the linear regression in objective 2.3 using proc reg.

SAS Learning code: (e3.sas)

```
data pollutant;
infile 'H:\e3.dat';
input straw rice;
run;

proc print data=pollutant; run;

Title 'Scatterplot of As in Rice vs. Si in Straw';
*i = rl means that the regression line will be drawn;
symbol1 v=circle i=rl;
*the plot statement is y*x.;
proc gplot data=pollutant;
  plot rice*straw;
run;

quit;
```

SAS Graph:



Problem 3 (2 pts.)

Laetisaric acid is a newly discovered compound that holds promise for control of fungus in crop plants. The data file fungus.dat shows the results of growing the fungus Pythium ultimum in various concentrations of laetisaric acid. Each growth value is the average of four radial measurements of a Pythium ultimum colony grown in a petri dish for 24 hours; there were two petri dishes at each concentration (Problem 12.2.6). In the data file, the concentration of acid is the first variable and the growth rate is the second variable. Construct the scatterplot with the regression line. Does there seem to be an association between concentration of laetisaric acid (X) and the amount of growth of the fungus (Y)?

Your submission should consist of the code, the graph and the answer to whether there seems to be an association between the variables.

2.2. Correlation Coefficient and Scatterplot:

The correlation is calculated in “proc corr”. It simply gives the pairwise Pearson’s correlation coefficients for each pair of quantitative variables in your dataset. (Note: it will ignore any categorical variables if you have any.) It also gives some useful summary statistics for each variable like mean, standard deviation, etc. You might like to use this in the future in place of proc univariate or proc means.

SAS Learning code: (e4.sas)

```
data pollutant;
infile 'H:\e3.dat';
input straw rice;
run;

proc print data=pollutant; run;

*in the matrix, the correlation coefficient is in the first row and the
  P value with the null hypothesis of rho = 0 is in the second row.;
Title 'Correlation';
proc corr data=pollutant;
run;

quit;
```

SAS Learning output:

```
                          The CORR Procedure
                       2  Variables:    straw    rice

                          Simple Statistics

Variable           N          Mean       Std Dev           Sum       Minimum       Maximum
straw             32      29.85156      10.03738     955.25000       8.35000      51.83000
rice              32     122.25281      44.50685          3912      51.72000     217.24000

                  Pearson Correlation Coefficients, N = 32
                        Prob > |r| under H0: Rho=0

                                    straw          rice

                    straw        1.00000       -0.56607
                                               0.0007

                    rice        -0.56607        1.00000
                                 0.0007
```

r = -0.56607 with a P value = 0.0007 (H_0: $\rho = 0$, H_a: $\rho \neq 0$)

Problem 4 (1 pts.)

Calculate the correlation coefficient between laetisaric acid and fungus growth (same data set as before). Do you think there is an association between the two variables?

Your submission should consist of the code, relevant output, the correlation coefficient and the answer to whether there seems to be an association between the variables.

The second part of this objective is to get a visual idea of what the correlation coefficient means. You will generate a scatterplot for various different correlation coefficients and then answer a question about the relative association of the variables.

SAS Learning code: (e5.sas)

```
%LET rho=0; *Choose a value between -1 and 1 for the correlation coefficient;
title1 'Correlation for Rho = ' &rho;
data correlation;
do j = 1 to 50 by 1;
   r1 = rand('uniform');
   r2 = rand('uniform');
   x = quantile('normal',r1,0,1);
   if abs(&rho) < 1
      then y=x*&rho + quantile('normal',r2,0,1)*(1-abs(&rho)**2);
      else y=&rho*x;
   output;
 end;
 run;

 proc print data=correlation; run;

 symbol1 v=circle i = none;
 proc gplot data = correlation;
   plot y*x;
 run;
 quit;
```

Problem 5 (2 pts.)

Using learning code e5.sas, choose the following values of ρ: -0.9, -0.4, 0, 0.4, 0.8 and determine which correlation coefficient produces the least association and which produces the most association between x and y?

Your submission should consist of the relevant parts of the output, the correlation coefficient and the graph ONLY for the least association and most association. What is the difference in the graphs when you have positive correlations or negative correlations? No code is required because you are just modifying the learning code.

2.3. Linear Regression:

"proc reg" is the regression procedure, it is quite powerful and has lots of options but we will just do simple linear regression with it. In addition, I will show you how to create the scatterplot with regression line using this method.

SAS Learning code: (e6.sas)

```
data pollutant;
infile 'H:\e3.dat';
input straw rice;
run;

proc print data=pollutant; run;

Title 'Linear Regression for Y=As (rice) and X=Si (straw)';
symbol1 v=circle i=none;
proc reg data=pollutant;
model rice=straw;/*y = x*/
plot rice*straw;
/*proc reg can create a scatterplots with the equation for the regression
line */
run;
quit;
```

SAS Learning output:

The REG Procedure
Model: MODEL1
Dependent Variable: rice

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	19677	19677	14.15	0.0007
Error	30	41730	1390.99380		
Corrected Total	31	61407			

Root MSE	37.29603	R-Square	0.3204
Dependent Mean	122.25281	Adj R-Sq	0.2978
Coeff Var	30.50730		

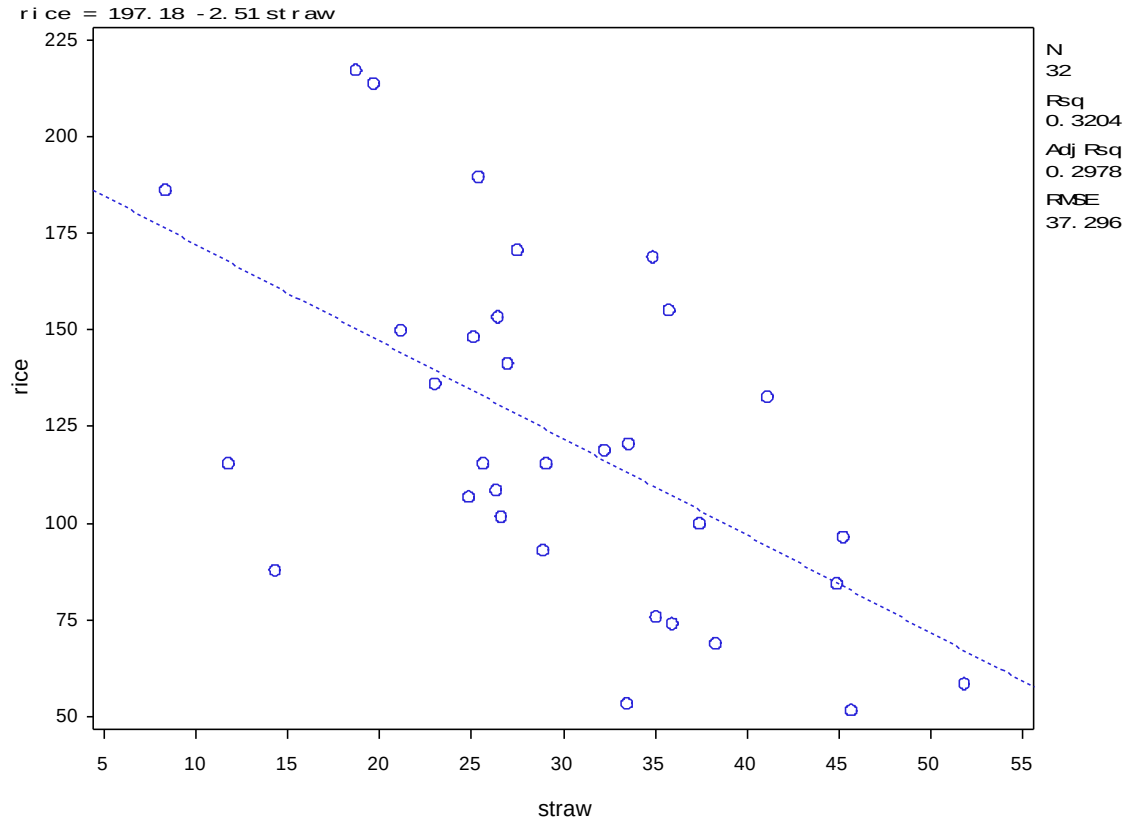
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	197.18070	20.98443	9.40	<.0001
straw	1	-2.51002	0.66736	-3.76	0.0007

The equation of the line is rice = 197.18070 – 2.51002 straw with a R^2 = 0.3204. The p-value of β_1 = 0.0007, the p-value of β_0 < 0.0001.

SAS Graph:

Linear Regression for Y=As (rice) and X=Si (straw)



Problem 6 (2 pts.)

Perform the linear regression analysis between laetisaric acid and fungus growth (same data set as before). Do you think there is an association between the two variables?

Your submission should consist of the code, relevant output, the equation of the line, R^2, and the P-value for the association between the two variables. (Hint: what is H_0?) Is there an association between the variables? No scatterplot is required.

BONUS B2 (1 pt.)

For questions 3, 4 and 6, you answered the question whether there was an association between laetisaric acid and fungus growth. Are your answers the same? If not, which answer would you say is correct and why. If the answers are the same, which of the methods would provide the most complete information and why.