SAS Project 2 (31.3 pts. + 4.2 Bonus)
Due Dec. 8

Objectives

- **Part 1: Confidence Intervals and Hypothesis tests**
  - **1.1) Calculation of t critical values and P-values**
  - **1.2) 1 – sample t-test**
  - **1.3) 2 – sample t-test (independent)**
  - **1.4) Paired Design**
- **Part 2: Inference with Categorical Variables**
  - **2.1) Chi-Square Distribution**
  - **2.2) Goodness of Fit Test**
  - **2.3) Test of Independence**
- **Part 3: ANOVA**
  - **3.1) One-way ANOVA**
  - **3.2) Two-Way ANOVA (BONUS)**
- **Part 4: Linear Regression**
  - **4.1) Scatterplot**
  - **4.2) Correlation Coefficient and a Scatterplot**
  - **4.3) Linear Regression**
  - **4.4) Residual Plots and QQplots**

Remember:
a) Please only turn in one paper per group.
b) Please put all of your names, STAT 503 and the project # on the front of the project.
c) Label each part and put them in logical order.
d) ALWAYS include your SAS code at the beginning of each problem.
e) Do NOT include more SAS output then is required to answer the question.

1. Confidence Intervals/Hypothesis tests:

1.1. Calculation of t critical values and P-values

Problem 1 (3 pt.)
Calculate the following:

- **a) t critical value when $\alpha = 0.05$, df = 21.9**
- **b) t critical value when $\alpha = 0.01$, df = 37.8**
- **c) with H_a: $\mu_1 \neq \mu_2$, df = 10.5, $t_s = 0.7757$**
- **d) with H_a: $\mu_1 > \mu_2$, df = 63.8, $t_s = 1.218$**
- **e) with H_a: $\mu_1 < \mu_2$, df = 26.4, $t_s = 2.194$**

Your submission should consist of the answers to all of the parts, your code (clearly indicating which part is coming from which line) and the appropriate parts of the output file.

SAS Code:

```
*in the code, each calculation has the part number after the variabel name;

*t critival value;
data tcritval;
*p = 1 - alpha/2, df = degrees of freedom;
alphaa = 0.05;
pa = 1 - alphaa/2;
dfa = 21.9;
CritVala = TINV(pa,dfa);
alphab = 0.01;
pb = 1 - alphab/2;
dfb = 37.8;
CritValb = TINV(pb,dfb);

proc print data=tcritval; run;

*P values;
data Pvalue;
*ts = the test statistic, df = degrees of freedom;
tsc = 0.7757; dfc = 10.5;
Pvalue2c = (1-PROBT(abs(tsc),dfc))*2; /*nondirectional P-value */
tsd = 1.218; dfd = 63.8;
Pvalued = 1-PROBT(abs(tsd),dfd); /*directional P-value */
tse = 2.194; dfe = 26.4;
Pvaluee = 1-PROBT(abs(tse),dfe); /*directional P-value */
proc print data=Pvalue; run;

quit;
```

Note: it is acceptable to include one code per part and hence, you would have one output statement per part.

SAS Output:

Obs	alphaa	pa	dfa	CritVala	alphab	pb	dfb	CritValb
1	0.05	0.975	21.9	2.07442	0.01	0.995	37.8	2.71231

Obs	tsc	dfc	Pvalue2c	tsd	dfd	Pvalued	tse	dfe	Pvaluee
1	0.7757	10.5	0.45504	1.218	63.8	0.11385	2.194	26.4	0.018614

Answer:

a) t_c = 2.07442, b)) t_c = 2.71234, c) P = 0.45504, d) P = 0.11385, e) P = 0.018614

If these values are not explicitly stated, it is required that they are clearly indicated which value is for which part.

1.2. Inference for 1 - sample for the mean:

Problem 2 (3.7 pts.)
This problem is based from exercise 7.2.17 on p. 234 in the textbook. As is stated in the textbook, the mean height of the control group is 2.58 cm. The data file is radishf.dat.

a) Calculate the 95% confidence interval for mean height of the fertilized radishes (α = 0.05) including the interpretation of the interval in the context of the example. Is the mean height of the control group in the confidence interval?

b) Perform a hypothesis test (significance level = 0.05) to determine if the mean height of the fertilized radishes is different from the mean height of the control group. Please include the 9 (8) steps for the hypothesis tests. In Step 6, you only need to include the P-value. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output.

c) Are parts a) and b) consistent with each other? Why or why not?

Your submission should consist of one code for both parts and the relevant parts of the output in addition to the confidence interval and interpretation, whether the mean value is in the confidence interval and the steps of the hypothesis test.

SAS Code:

```
data radish;
infile 'I:\My Documents\Stat 503\SAS\radishf.dat';
input height;
run;

proc print data=radish; run;

Title 'One Sample Inference';
proc ttest data=radish alpha=0.05 H0 = 2.58;
var height;
run;

quit;
```

SAS output:

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.1821	-0.2808	8.6451	11.5096	9.0998	15.6662

DF	t Value	Pr > |t|
27	0.74	0.4677

Answer:

a) CI: (-0.2808, 8.6451).

We are 95% confidence that the mean height of radishes with fertilizer sticks after 2 weeks is between -0.2808 cm (0 cm) and 8.6451 cm.

The value of 2.58 cm is in this range.

b)
Step 1: Does the fertilizer stick affect the mean height of radish plants.
Step 2: μ = mean height of the radish plant using the fertilizer stick.
Step 3: H_0:μ = 2.58: the mean height is 2.58 cm
H_A:$\mu \neq$ 2.58; the mean height is not 2.58 cm.
Step 4: α = 0.05
Step 5: t_s = 0.74, df = 27
Step 6: P-value = 0.4677
Step 7: 0.4677 > 0.05
Step 8: fail to reject H_0
Step 9: This study does not provide evidence (P = 0.4677) at the 0.05 significance level that the fertilizer sticks changes the height from 2.58 cm.

c) Yes the two parts are consistent because they both state that the 2.58 is a possible height of the radishes when using the fertilizer sticks.

1.3. Inference for 2-sample for two means (independent):

Problem 3 (3.7 pts.)
Now suppose that we do not know the height of the control radishes (we are using the full data set in question 7.2.17). This is in file radishall.dat. Let μ_1 refer to the mean height for the control radishes and μ_2 refer to the mean height of the fertilized radishes. We will be assuming that the variances are NOT the same.
a) Calculate the 95% confidence interval for this example including the interpretation of the interval in the context of the example. Are the two means the same?
b) Perform the hypothesis test (significance level = 0.05) to determine if the two heights are different. Please include the 9 (8) steps for the hypothesis tests. In Step 6, you only need to include the P-value. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output.
c) Are parts a) and b) consistent with each other? Why or why not?
Your submission should consist of one code for both parts and the relevant parts of the output in addition to the confidence interval and interpretation, whether the two means are the same or not (using the confidence interval) and the steps of the hypothesis test. Be sure to use the appropriate method to calculate the variance (pooled or unpooled).

SAS Code:

```
data radish;
infile 'I:\My Documents\Stat 503\SAS\radishall.dat';
input height type$;
run;

proc print data=radish; run;

proc sort data=radish;
by type;
run;

Title 'Two Sample Inference';
proc ttest data=radish alpha=0.05 ;
class type;
var height;
run;

quit;
```

SAS Output:

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Control		2.5821	2.3284	2.8359	0.6544	0.5174	0.8907
Fertiliz		2.0393	1.7598	2.3188	0.7208	0.5699	0.9811
Diff (1-2)	Pooled	0.5429	0.1740	0.9117	0.6884	0.5795	0.8480
Diff (1-2)	Satterthwaite	0.5429	0.1739	0.9118			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	54	2.95	0.0047
Satterthwaite	Unequal	53.503	2.95	0.0047

Answer:

a) CI: (0.1739, 0.9118).

We are 95% confident that the mean different in heights of control vs. fertilized is between 0.1739 and 0.9118 cm larger.

The two means are not the same because 0 is not in the interval.

b)
Step 1: Does the fertilizer stick affect the mean height of radish plants.
Step 2: μ_1 = mean height of the radish plant without using the fertilizer stick.
μ_2 = mean height of the radish plant using the fertilizer stick.
Step 3: H_0:$\mu_1 = \mu_2$: the mean heights are the same
H_A:$\mu_1 \neq \mu_2$; the mean heights are different
Step 4: $\alpha = 0.05$
Step 5: $t_s = 0.2.95$, df = 53.503
Step 6: P-value = 0.0047
Step 7: $0.0047 \leq 0.05$
Step 8: reject H_0
Step 9: This study does provide evidence (P = 0.0047) at the 0.05 significance level that the heights are different with and not using the fertilizer sticks.

c) Yes they are the same, in a) they were not the same and in b) we reject H_0 which is that they are not the same.

1.4 Paired - Design

Problem 4 (3.2 pts.)

This problem is based from exercise 8.2.4 on p. 308 in the textbook. The data file is wloss.dat. (Note: the order of the variables is the same as in the book.)
a) Calculate the 90% confidence interval for the difference in weight loss with and without the drug including the interpretation of the interval in the context of the example.
b) Perform a hypothesis test (significance level = 0.10) to determine if the mean weight loss for the people taking MCPP is greater than those who did not take the drug. Please include the 9 (8) steps for the hypothesis tests. In Step 6, you only need to include the P-value. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output.
c) Are parts a) and b) consistent with each other? Why or why not?
Your submission should consist of one code for both parts and the relevant parts of the output in addition to the confidence interval and interpretation, whether the mean value is in the confidence interval and the steps of the hypothesis test.

SAS Code:

```
data weightloss;
infile 'I:\My Documents\Stat 503\SAS\wloss.dat';
input subject MCPP placebo;
run;

proc print data=hunger; run;

Title 'Two Sample Paired Inference';
proc ttest data=weightloss alpha=0.1;
paired MCPP*placebo;
run;

quit;
```

SAS Output:

Mean	90% CL Mean		Std Dev	90% CL Std Dev	
1.0000	0.5541	1.4459	0.7194	0.5167	1.2309

DF	t Value	Pr > |t|
8	4.17	0.0031

Answer:

a) CI: (0.5541, 1.4459).
We are 90% confident that the mean difference in weight loss of drug vs. placebo is between 0.5541 and 1.4459 kg larger.
The mean value of the difference (1.00) is in this interval.

b)
Step 1: Does the drug affect weight loss?
Step 2: μ_1 = mean weight loss using the drug.
μ_2 = mean weight loss when drug is not used.
$d = \mu_1 - \mu_2$
Step 3: H_0:$\mu_1 = \mu_2$ or H_0: $\mu_d = 0$; the mean weight losses are the same
H_A:$\mu_1 \neq \mu_2$ H_0: $\mu_2 \neq 0$: the mean weight losses are different
Step 4: $\alpha = 0.10$
Step 5: $t_s = 4.17$, df = 8
Step 6: P-value = 0.0031
Step 7: $0.0031 \leq 0.10$
Step 8: reject H_0
Step 9: This study does provide evidence (P = 0.0031) at the 0.10 significance level that the mean weight losses are different with and without the drug.

c) Yes they are the same, in a) they were not the same because 0 was not in the interval and in b) we reject H_0 which is that they are not the same.

2. Categorical Data:

2.1 χ^2 Distribution

Problem 5 (1.5 pt.)
Calculate the following which refers to Example 9.4.6 (p.353).
a) The critical value for $\alpha = 0.005$.
b) The P-value ($\chi^2_s = 7.71$, df = 5).
Your submission should consist of the answers to parts a and b, your code (clearly indicating which part is coming from which line) and the appropriate parts of the output file.

SAS Code:

```
data chisquared;
alpha=0.0005;
df = 5;
CritVal = QUANTILE('CHISQ',alpha,df);
chis = 7.71;
Pvalue = 1 - PROBCHI(chis,df);

proc print data=chisquared; run;

quit;
```

SAS Output:

Obs	alpha	df	CritVal	chis	Pvalue
1	.0005	5	0.15814	7.71	0.17296

Answer:

a) $\chi^2_c = 0.15814$, b) P = 0.17296

2.2. Goodness of Fit Test

Problem 6 (3 pts.)

Suppose you do not feel comfortable with SAS's random number generator and decide to develop your own discrete integer algorithm. This algorithm is to generate the digits 0-9 with equal probability. After writing the macro, you decide to test it out by generating 1000 digits and checking to see if there is any evidence that the algorithm is not performing properly. That number of times that you generated each digit is in rannum.dat. The infile contains the digit in the first column and the number of times that it appears in the second. Perform the goodness of fit test for this data. Remember, you need to decide what percentage to use for each of the digits.

Your submission should consist of the code, the relevant parts of the output, the 9 (8) steps of the hypothesis test. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output. In addition, explain why you chose the expected percentages that you did. Does your macro work (that is, are each of the digits from 0 – 9 random?)? Why or why not? Hint: What is H_0?

SAS Code:

```
data RandomNumber;
infile 'I:\My Documents\Stat 503\SAS\rannum.dat';
input number $ count;
run;

proc print data=RandomNumber; run;

Title 'Goodness of Fit';
proc freq data=RandomNumber order=data;
tables number/nocum chisq
   testp = (0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1);
weight count;
run;

quit;
```

SAS Output:

Chi-Square Test for Specified Proportions	
Chi-Square	4.9800
DF	9
Pr > ChiSq	0.8360

Answer:

Step 1: Are the digits random? OR Do each of the digits occur with the same frequency?
Step 2: p_i = the probability that digit i occurs where i = 0, …, 9
Step 3: H_0: all of the p_i's are the same
H_A: at least on p_i is different
Step 4: $\alpha = 0.05$
Step 5: $\chi^2_s = 4.98$, df = 9
Step 6: P-value = 0.8360
Step 7: 0.8360 > 0.05
Step 8: fail to reject H_0
Step 9: This study does not provide evidence (P = 0.8360) at the 0.05 significance level that the percentage of the time that the digits occur is different.

I chose a percentage of 10% because there are 10 digits and each has an equally likely chance of occurring.

The macro worked because the χ^2 test says that the number of times each digit occurred is the same (fail to reject H_0).

2.3. Test of Independence:

Problem 7 (2.5 pts.)
The file plover.dat includes the data from example 10.5.1 (p.385). The first column is the location, the second is the year and then the appropriate counts. Perform a chi-squared test of independence to determine whether nesting location is independent of year.
Your submission should consist of the code and relevant parts of the output, the 9 (8) steps of the hypothesis test. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output. In addition, please state whether nest location is independent of year. Why or why not? Hint: What is H_0?

SAS Code:

```
data plover;
infile 'I:\My Documents\Stat 503\SAS\plover.dat';
input location $ year count;
run;

proc print data=plover; run;

Title 'Test of Independence';
proc freq data=plover order=data;
tables location*year/nocum chisq expected nopercent;
weight count;
run;

quit;
```

SAS Output:

Table of location by year

location	year			
Frequency / Expected / Row Pct / Col Pct	2004	2005	2006	Total
AF	21	19	26	66
	18.549	27.176	20.275	
	31.82	28.79	39.39	
	48.84	30.16	55.32	
PD	17	38	12	67
	18.83	27.588	20.582	
	25.37	56.72	17.91	
	39.53	60.32	25.53	
G	5	6	9	20

	5.6209	8.2353	6.1438	
	25.00	30.00	45.00	
	11.63	9.52	19.15	
Total	43	63	47	153

Statistics for Table of location by year

Statistic	**DF**	**Value**	**Prob**
Chi-Square	4	14.0894	0.0070
Likelihood Ratio Chi-Square	4	14.3582	0.0062
Mantel-Haenszel Chi-Square	1	0.0010	0.9753
Phi Coefficient		0.3035	
Contingency Coefficient		0.2904	
Cramer's V		0.2146	

Answer:

Step 1: Are population distribution in the nests are the same for the three years?
Step 2: not required
Step 3: H_0: the population distributions of the nest locations are the same.
H_A: the population distributions are not all the same, at least one is different.
Step 4: $\alpha = 0.05$
Step 5: $\chi^2_s = 14.0894$, df = 4
Step 6: P-value = 0.0070
Step 7: $0.0070 \leq 0.05$
Step 8: reject H_0
Step 9: This study does provide evidence (P = 0.0070) at the 0.05 significance level that the population distribution of the plovers is different in the different years.

Therefore the nest location is NOT independent of year.

Bonus B1 (1.7 pt.):
Repeat Problem 7 assuming that the count of Prairie dog habitat in 2005 was 28 instead of 38.
Your submission should consist of the infile, code and relevant parts of the output, the test statistic and the P-value. In addition, please state whether nest location is now independent of year.

SAS Infile:

AF 2004 21
AF 2005 19
AF 2006 26
PD 2004 17
PD 2005 28
PD 2006 12
G 2004 5
G 2005 6
G 2006 9

SAS Code:

```
data plover;
infile 'I:\My Documents\Stat 503\SAS\ploverB1.dat';
input location $ year count;
run;

proc print data=plover; run;

Title 'Test of Independence';
proc freq data=plover order=data;
tables location*year/nocum chisq expected nopercent;
weight count;
run;

quit;
```

SAS Output:

Frequency Expected Row Pct Col Pct	Table of location by year			
location	**year**			
	2004	**2005**	**2006**	**Total**
AF	21	19	26	66
	19.846	24.462	21.692	
	31.82	28.79	39.39	
	48.84	35.85	55.32	
PD	17	28	12	57
	17.14	21.126	18.734	
	29.82	49.12	21.05	
	39.53	52.83	25.53	
G	5	6	9	20
	6.014	7.4126	6.5734	
	25.00	30.00	45.00	
	11.63	11.32	19.15	
Total	43	53	47	143

Statistics for Table of location by year

Statistic	DF	Value	Prob
Chi-Square	4	8.1365	0.0867
Likelihood Ratio Chi-Square	4	8.2696	0.0822
Mantel-Haenszel Chi-Square	1	0.0018	0.9659
Phi Coefficient		0.2385	
Contingency Coefficient		0.2320	
Cramer's V		0.1687	

Answer.

$\chi^2_s = 8.1365$, df = 4
P-value = 0.0867
Therefore the nest location IS independent of year.

3. ANOVA:

3.1 One-Way ANOVA

Problem 8 (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. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output.

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 (1 is optional).

SAS Code:

```
data fitness;
infile 'I:\My Documents\Stat 503\SAS\fitness.dat';
input status $ days;
run;

proc print data=fitness; run;

Title 'One-Way ANOVA';
proc glm data=fitness alpha=0.05;
class status;
model days = status;
run;

quit;
```

SAS Output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	672.000000	336.000000	16.96	<.0001
Error	21	416.000000	19.809524		
Corrected Total	23	1088.000000			

Answer:

Step 1: Does the number of days for recovery depend on the fitness status?
Step 2: μ_1: mean number of days for recovery for people with below average fitness status.
μ_2: mean number of days for recovery for people with average fitness status.
μ_3: mean number of days for recovery for people with above average fitness status.
Step 3: H_0: $\mu_1 = \mu_2 = \mu_3$: the average recovery time is the same for all fitness statuses
H_A: at least one μ_i is different.
Step 4: $\alpha = 0.05$
Step 5: F = 16.96, dfb = 2, dfw = 21
Step 6: P-value < 0.0001
Step 7: 0.0001 ≤ 0.05
Step 8: reject H_0
Step 9: This study does provide evidence (P < 0.0001) at the 0.05 significance level that the mean days for recovery is different for at least one of the fitness levels.

3.2 Two-Way ANOVA (BONUS)

Bonus B2 (2.5 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 (Hint: use the appropriate P-values).

SAS Code:

```
data birch;
infile 'I:\My Documents\Stat 503\SAS\birch.dat';
input ATP type$ flood$;
run;

proc print data=birch; run;

Title 'Two-Way ANOVA';
proc glm data=birch alpha=0.05;
class type flood;
model ATP=type flood type*flood;
run;

quit;
```

SAS Output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.55296875	1.51765625	38.39	<.0001
Error	12	0.47437500	0.03953125		

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Corrected Total	15	5.02734375			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	1	2.19780625	2.19780625	55.60	<.0001
flood	1	2.25750625	2.25750625	57.11	<.0001
type*flood	1	0.09765625	0.09765625	2.47	0.1420

Answer:

Interaction term is NOT significant because P = 0.1420 > 0.05
the type of tree IS significant because P < 0.001 ≤ 0.05
whether the tree is flooded or not IS significant because P < 0.001 ≤ 0.05

4. Linear Regression:

4.1. Scatterplot

Problem 9 (1.5 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.

SAS Code:

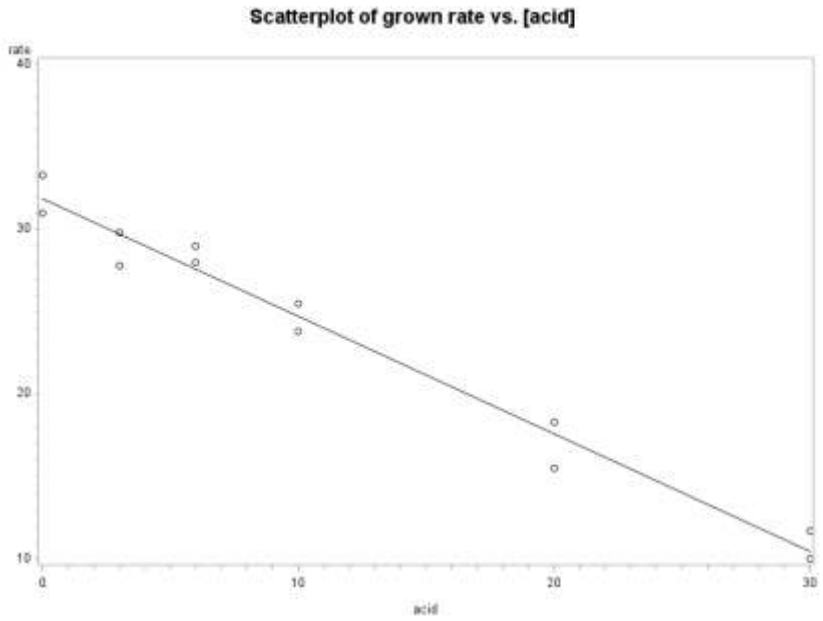
```
data LA;
infile 'I:\My Documents\Stat 503\SAS\fungus.dat';
input acid rate;
run;

proc print data=LA; run;

Title 'Scatterplot of grown rate vs. [acid]';
symbol1 v=circle i=rl c=black;
proc gplot data=LA;
  plot rate*acid;
run;

quit;
```

SAS Graph:



Answer:

It looks like there is a non-zero (negative) slope, so there is an association between concentration of acid and growth rate of the fungus

4.2. Correlation Coefficient and Scatterplot:

Problem 10 (1.7 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.

SAS Code:

```
data LA;
infile 'I:\My Documents\Stat 503\SAS\fungus.dat';
input acid rate;
run;

proc print data=LA; run;

Title 'Correlation';
proc corr data=LA;
run;

quit;
```

SAS Output:

Pearson Correlation Coefficients, N = 12
Prob > |r| under H0: Rho=0

	acid	rate
acid	1.00000	-0.98753
		<.0001
rate	-0.98753	1.00000
	<.0001	

Answer:

r = -0.98753. Because the r value is so close to -1, I would say that there is an association between the variables.

4.3. Linear Regression:

Problem 11 (2.5 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, the 9 (8) steps of the hypothesis test for the association between the two variables.. No calculations by hand are required; that is, all of the values for the steps are to be taken from the SAS output. Is there an association between the variables? (Hint: what is H_0?)

SAS Code:

```
data LA;
infile 'I:\My Documents\Stat 503\SAS\fungus.dat';
input acid rate;
run;

proc print data=LA; run;

Title 'Linear Regression for Y=growth rate and X=[acid]';
symbol1 v=circle i=none;
proc reg data=LA;
model rate=acid;
run;

quit;
```

SAS Output:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	660.56797	660.56797	393.64	<.0001
Error	10	16.78120	1.67812		
Corrected Total	11	677.34917			

Root MSE	1.29542	**R-Square**	0.9752
Dependent Mean	23.64167	**Adj R-Sq**	0.9727
Coeff Var	5.47940		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	31.82979	0.55693	57.15	<.0001
acid	**1**	-0.71201	0.03589	-19.84	<.0001

Answer:

Equation of the line: $Y = 31.82979 - 0.71201\ X$, $R^2 = 0.9752$

Step 1: Is there an association between acid concentration and growth rate of the fungus?
Step 2: β_1: population slope of growth rate vs. acid concentration.
Step 3: H_0: $\beta_1 = 0$: there is no association between acid concentration and growth rate.
H_A: $\beta_1 \neq 0$: there is an association between acid concentration and growth rate.
Step 4: $\alpha = 0.05$
Step 5: t = -19.84, df = 10
Step 6: P-value < 0.0001
Step 7: $0.0001 \leq 0.05$
Step 8: reject H_0
Step 9: This study does provide evidence (P < 0.0001) at the 0.05 significance level that there is an association between growth rate and acid concentration.

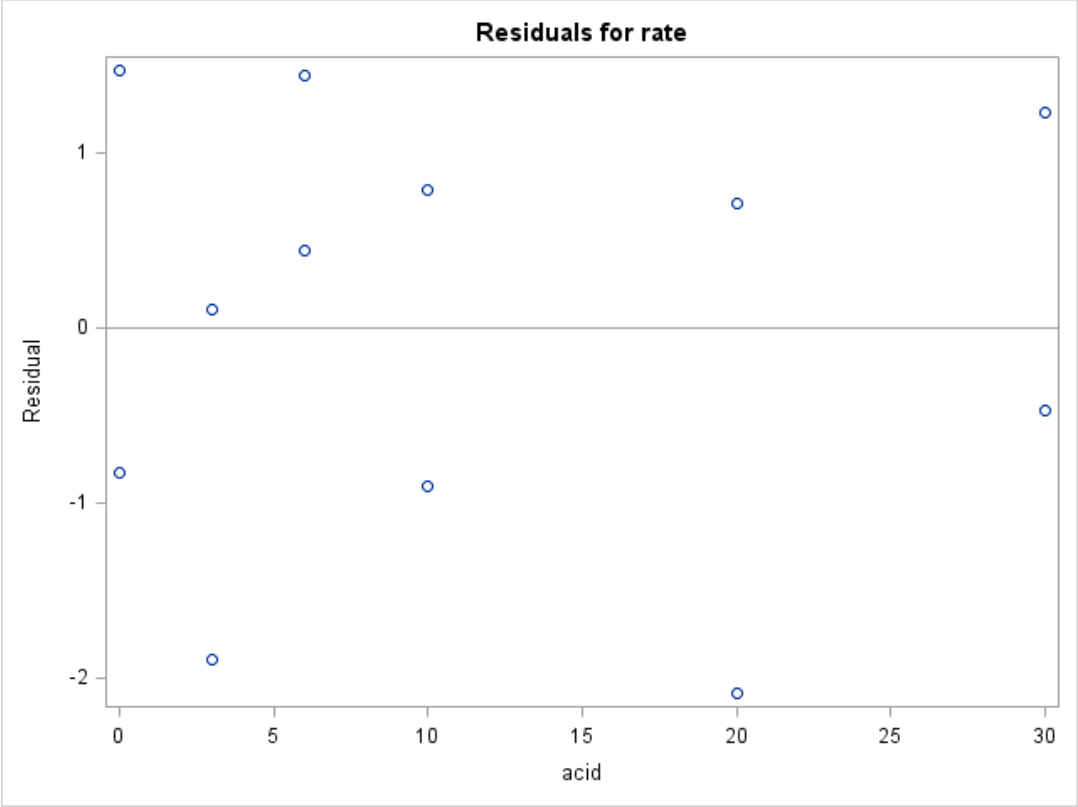
Problem 12 (1 pt.)
For questions 9, 10 and 11, you answered the question whether there was an association between laetisaric acid and fungus growth. Are your answers the same? If not, which answer would you say is correct and why. If the answers are the same, which of the methods would provide the most complete information and why.

All of the answers are the same in questions 9, 10 and 11. The last method using the linear regression would provide the most complete information because this method takes into account all of the information in the previous sections. However, that does not mean that you should not create the scatterplot first to visually look at your data.

4.4 Residual Plots and QQplots

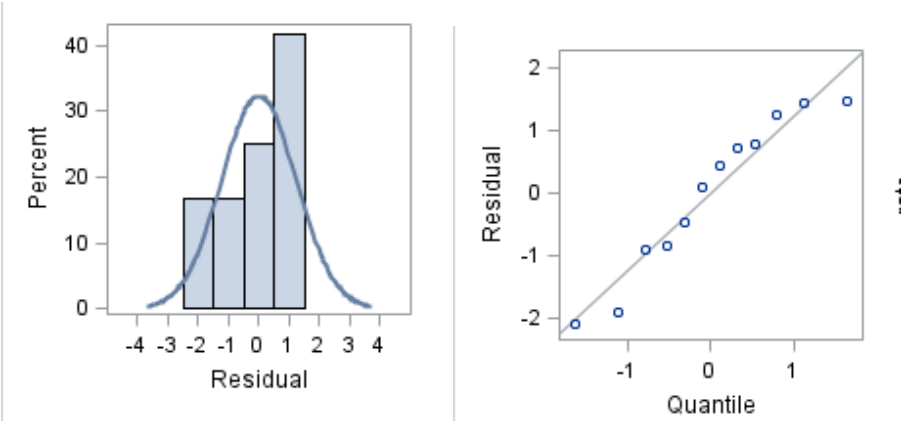
Problem 12 (2 pts.)
We are going to continue our analysis of the linear regression analysis between laetisaric acid and fungus growth (same data set as before).
a) Using the same output as in Objective 4.3, determine using the residual plot if the data is linear and the variances are approximately constant.
b) Using the same output as in Objective 4.3, determine using the histogram and QQplot if the normality assumption is met.
Your submission should consist of the residual plot, histogram, QQplot and the answers to the questions about linearity, constant variance and normality.

a) Residual plot



I do not see any patterns here so I would say that the data is linear. It looks like the variance is approximately constant no matter what the concentration of acid is.

b) Normality assumptions



From the histogram (left) and the QQplot (right), it looks like the normality assumption is met.