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

The command in SAS to calculate TINV(p,df) where p = 1 - $\frac{\alpha}{2}$ and df = the degrees of freedom. The command is called TINV (t inverse) because this is calculating the percentile (or inverse) of the t distribution (function PROBT).

To calculate the P-value in SAS, you use the function PROBT which is the probability that we are less than or equal to a certain value of the appropriate t distribution.

For a one-tailed alternative hypothesis (directional), the formula is

Pvalue1 = 1-PROBT(abs(ts),df).

Note that the value of ts has to be positive (abs = absolute value) in the formula. This is just the probability using the appropriate t distribution that we are beyond the values of ts. Since we are using the absolute value, you only use this formula when the direction is 'correct'.

For a two-tailed alternative hypothesis (non-directional) the formula Pvalue2 = (1-PROBT(abs(ts),df))*2. This is two times the probability that we are past the value of ts.

In the following learning code, I am calculating $t(20)_{0.025}$, the directional and non-directional P-values when df = 20 and ts = 2.758.

SAS Learning code: (b1.sas)

```
*t critival value;
data tcritval;
*p = 1 - alpha/2, df = degrees of freedom;
alpha = 0.05;
p = 1 - alpha/2;
df = 20;
CritVal = TINV(p,df);
proc print data=tcritval; run;

*P values;
data Pvalue;
*ts = the test statistic, df = degrees of freedom;
ts = 2.758; df = 20;
Pvalue2 = (1-PROBT(abs(ts),df))*2; /*nondirectional P-value */
Pvalue1 = 1-PROBT(abs(ts),df); /*directional P-value */
proc print data=Pvalue; run;

quit;
```

SAS Learning output:

Obs	alpha	p	df	CritVal
1	0.05	0.975	20	2.08596

Obs	ts	df	Pvalue2	Pvalue1
1	2.758	20	0.012131	.006065423

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.

1.2. Inference for 1 - sample for the mean:

To have SAS calculate the confidence intervals and hypothesis testing, we use the proc ttest. We need to specify the α not the confidence level which is 1 - α. Therefore, the confidence level is the same for a 95% confidence interval and a hypothesis test with $\alpha = 0.5$. For the hypothesis test, the value of μ_0 is given by H0 = value in the procedure line.

This example in the learning code is taken from the textbook (Example 6.7.3) concerning the thorax weight from male and female Monarch butterflies. I will be using the same example for the next section (1.3) when we will be looking a 2 – sample inference for the mean. In this example, we are a) determining the 95% confidence interval and b) performing the hypothesis test to see if the weight of the male thorax is the same as the average weight of the female thorax (63.4)

SAS Learning code: (b2.sas)

```
data thorax;
infile 'H:\b2.dat';
input male;
run;

proc print data=thorax; run;

Title 'One Sample Inference';
proc ttest data=thorax alpha=0.05 H0 = 63.4;
   *alpha: the default value is 0.05, it is required for
         both CI and hypothesis tests.;
   *H0: The default value is 0, it is only required for hypothesis tests.;
var male;
run;

quit;
```

SAS Learning output:

The TTEST Procedure

Variable: male

N	Mean	Std Dev	Std Err	Minimum	Maximum
7	75.7143	8.4007	3.1752	63.0000	85.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
75.7143	67.9450	83.4836	8.4007	5.4133	18.4989

DF	t Value	Pr > |t|
6	3.88	0.0082

The 95% Confidence interval is (67.9450, 83.4836).

For the hypothesis test with $\alpha = 0.05$ (95% confidence interval),

df = 6, t value = 3.88 and the p-value = 0.0082

The rest of the output is for diagnostics and is not covered in this class.

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 ($\alpha = 0.05$) 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.

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

For two means, we will again use “proc ttest”. However, like in the boxplot, we need to indicate which value is for which of the two cases. In example 6.7.3:

weight	gender
67	Male
73	Male
…	
54	Female
61	Female
…	

Be aware that you must sort the data by the group variable first (gender in this example), in order for the proc ttest to work correctly. Because you are sorting the data, be careful to determine what is ‘1’ and what is ‘2’. SAS always does the inference as ‘1’ – ‘2’. H0 is NOT required in the procedure statement.

For pooled variance, use the ‘Pooled’ rows, for unpooled variance, use the ‘Satterthwaite’ rows. The section titled ‘Equality of Variances’ is a hypothesis test whether the variances are the same.

SAS Learning code: (b3.sas)

```
data thorax;
infile 'H:\b3.dat';
*note: gender is a categorical variable and is the group variable;
input weight gender$;
run;

proc print data=thorax; run;

proc sort data=thorax;
by gender;
run;

Title 'Two Sample Inference';
proc ttest data=thorax alpha=0.05 ;
class gender; *class statements are used for categorical variables;
var weight;
run;

quit;
```

SAS Learning output:

The TTEST Procedure
Variable: weight

gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female	8	63.3750	7.5392	2.6655	54.0000	75.0000
Male	7	75.7143	8.4007	3.1752	63.0000	85.0000
Diff (1-2)		-12.3393	7.9484	4.1137		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		63.3750	57.0721	69.6779	7.5392	4.9847	15.3443
Male		75.7143	67.9450	83.4836	8.4007	5.4133	18.4989

Diff (1-2)	**Pooled**	-12.3393	**-21.2264**	**-3.4522**	7.9484	5.7622	12.8052
Diff (1-2)	**Satterthwaite**	-12.3393	**-21.3531**	**-3.3255**			

Method	**Variances**	**DF**	**t Value**	**Pr > |t|**
Pooled	Equal	**13**	**-3.00**	**0.0102**
Satterthwaite	Unequal	**12.23**	**-2.98**	**0.0114**

Equality of Variances

Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	6	7	1.24	0.7751

The 95% Confidence interval (pooled) is (-21.2264, -3.4522), df = 13
The 95% Confidence interval (unpooled) is (-21.3531, -3.3255), df = 12.23

For the hypothesis test with α = 0.05 (again from the 95% CI),
pooled: df = 13, t value = -3.00 and the p-value = 0.0102
unpooled: df = 12.23, t-value = -2.98, p-value = 0.0114.

The rest of the output is for diagnostics and is not covered in this class.

Note: the numbers are opposite those in the example in the book because the book performs the test male – female, and SAS performs the test female – male because female is first alphabetically.

Notice also that the p-value is different in the parts 1.2 and 1.3. The latter is better because it includes more information, the variability of the data for the female butterflies in addition to the average value.

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).

1.4 Paired - Design

For paired data, we will again use “proc ttest” like we did in 1.2 and 1.3; however, we just use a “paired” command for both variables in place of the “var” command for a single variable. The difference will be the first variable in the paired command minus the second variable. In addition, we now have a separate column for each of the two variables instead of using a grouping variable. The file b4.dat was taken from Example 8.2.4 in the book for the following learning code. In addition, this example uses a directional alternative hypothesis. When you are performing directional hypothesis, be sure to know which variable is which so the direction is appropriate. In the example in the book (which was nondirectional), I am choosing the direction of the drug decreases how hungry the women are.

SAS Learning code: (b4.sas)

```
data hunger;
infile 'H:\b4.dat';
input subject drug placebo;
run;

proc print data=hunger; run;

Title 'Two Sample Paired Inference';
proc ttest data=hunger alpha=0.01 SIDE=L;
paired drug*placebo;
/*generates the hypothesis test with alpha = 0.01 (99% CI) for drug - placebo
  with Ha = mu1 < mu2 (SIDE = L). If SIDE=U, it would be Ha: mu1 > mu2*/
run;
```

SAS Learning output:

The TTEST Procedure
Difference: drug - placebo

N	Mean	Std Dev	Std Err	Minimum	Maximum
9	-29.5556	32.8219	10.9406	-90.0000	8.0000

Mean	99% CL Mean		Std Dev	99% CL Std Dev	
-29.5556	-Infty	2.1336	32.8219	19.8127	80.0650

Note: -Infty means – infinity.

DF	t Value	Pr < t
8	-2.70	0.0135

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.

2. Categorical Data:

2.1 χ^2 Distribution

Similar to the t distribution, the χ^2 distribution depends on the "degrees of freedom". Unlike the t distribution, the χ^2 distribution is skewed and only positive values can occur.

When performing a test that involves a chi-square statistic, large values suggest a departure from the Null hypothesis. As a result, p-values of hypothesis tests involve determining an upper tail area.

SAS provides two functions involving the chi-square distribution. PROBCHI(χ^2_s, df) returns the upper tail area for a specific test statistic which is the P-value. The Quantile function which we have seen before will calculate x for a specific quantile p, or QUANTILE('CHISQ',p,df) = $Pr(\chi^2>x) = p$ for df =df. This function can be used to determine the critical value of a test. The sample code is based on Example 9.4.5 where $\chi^2_s = 43.2$, df = 3 using an alpha of 0.05.

SAS Learning code: (b5.sas)

```
data chisquared;
*chis is the calculated test statistic, df is the degrees of freedom for
  the test, alpha is the significance level, for the critical value, we want
  the percentile to be alpha;
alpha=0.05;
df = 3;
CritVal = QUANTILE('CHISQ',alpha,df);
chis = 43.2;
Pvalue = 1 - PROBCHI(chis,df);

proc print data=chisquared; run;

quit;
```

SAS Learning output:

Obs	alpha	df	CritVal	chis	Pvalue
1	0.05	3	0.35185	43.2	2.2318E-9

P-value = 2.2318×10^{-9}, $\chi^2_c = 0.35185$

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.

2.2. Goodness of Fit Test

The Goodness-of-Fit test uses proc freq. The difficulty is the output does not include the expected value (E_i) only the percent. Therefore, to convert this back to what we use in class, you will have to manually compute the E_i's for each of the observations. However, the program does calculate χ^2_s and the P-value for you.

The learning code is based on Example 9.4.6 (p. 353). SAS does not like 0's, it considers that missing data, so I inputted a value of 0.001 for the 0 for Variegated Low.

SAS Learning code: (b6.sas)

```
data flaxseed;
infile 'H:\b6.dat';
input coloracid $ count;
*I converted the two inputs Color and Acid Level into one variable;
run;

proc print data=flaxseed; run;

Title 'Goodness of Fit';
proc freq data=flaxseed order=data;
tables coloracid/nocum chisq
   testp = (0.1875,0.375,0.1875,0.0625,0.125,0.0625);
*the qualitative variable is listed in the Table statement;
*nocum: prevents printout of the cumulative percentages;
*testp: the percentages for the expected, these need to be calculated out;
weight count; *the number of observations or 'count' is listed as the weight;
run;

quit;
```

SAS Learning output:

Goodness of Fit

The FREQ Procedure

coloracid	Frequency	Percent	Test Percent
BrownLow	15	20.83	18.75
BrownInt	26	36.11	37.50
BrownHig	15	20.83	18.75
VarLow	0.001	0.00	6.25
VarInter	8	11.11	12.50
VarHigh	8	11.11	6.25

Chi-Square Test for Specified Proportions

Chi-Square	7.7016
DF	5
Pr > ChiSq	0.1735

WARNING: 33% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 72.001

The information that is required from this output is: test statistic is in red, the df is in **green** and the P-value is in blue.

If you look in the book, the expected values of 2/6 = 1/3 of the entries are less than 5. This is what the warning is for.

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?

2.3. Test of Independence:

The test of independence (2 x 2 contingency table) also uses the proc freq. The example I am using is Migraine Headaches in section 10.2 (p. 365 – 369) except that that the H_a is nondirectional instead of directional.

SAS Learning code: (b7.sas)

```
data migraine;
infile 'H:\b7.dat';
input success $ sugery $ count;
run;

proc print data=migraine; run;

Title 'Test of Independence';
proc freq data=migraine order=data;
tables success*sugery/nocum chisq expected nopercent;
*expected: provides the expected values of each cell;
*nopercent: doesn't print out the percentages, just the frequencies;
weight count;
run;

quit;
```

SAS Learning output:

The FREQ Procedure

Frequency Expected Row Pct Col Pct	success	Table of success by sugery sugery real	sham	Total
	success	41 36.587 73.21 83.67	15 19.413 26.79 57.69	56
	no	8 12.413 42.11 16.33	11 6.5867 57.89 42.31	19
	Total	49	26	75

Statistics for Table of success by sugery

Statistic	DF	Value	Prob
Chi-Square	1	6.0619	0.0138
Likelihood Ratio Chi-Square	1	5.8549	0.0155
Continuity Adj. Chi-Square	1	4.7661	0.0290
Mantel-Haenszel Chi-Square	1	5.9810	0.0145
Phi Coefficient		0.2843	
Contingency Coefficient		0.2735	
Cramer's V		0.2843	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	41
Left-sided Pr <= F	0.9965
Right-sided Pr >= F	0.0156
Table Probability (P)	0.0121
Two-sided Pr <= P	0.0241

Sample Size = 75

Note: We are not covering the Fisher's Exact Test in the class though it is discussed in section 10.4

The contingency table that you would report in your homework is in red. The test statistic, df and P-value are in blue. We will not be using anything else in this output.

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?

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.

3. ANOVA:

3.1 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.

SAS Learning code: (b8.sas)

```
data lambs;
infile 'H:\b8.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

Source	DF	Sum of 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 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).

3.2 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: (b9.sas)

```
data iron;
infile 'H:\b9.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

Source	DF	Sum of 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.

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).

4. 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 **b10.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).

4.1. Scatterplot

The scatterplot is a visual presentation of the data. Though there are more ways to create a scatterplot than I am showing you, the best way to create it in SAS is using the plotting process, proc gplot.

SAS Learning code: (b10.sas)

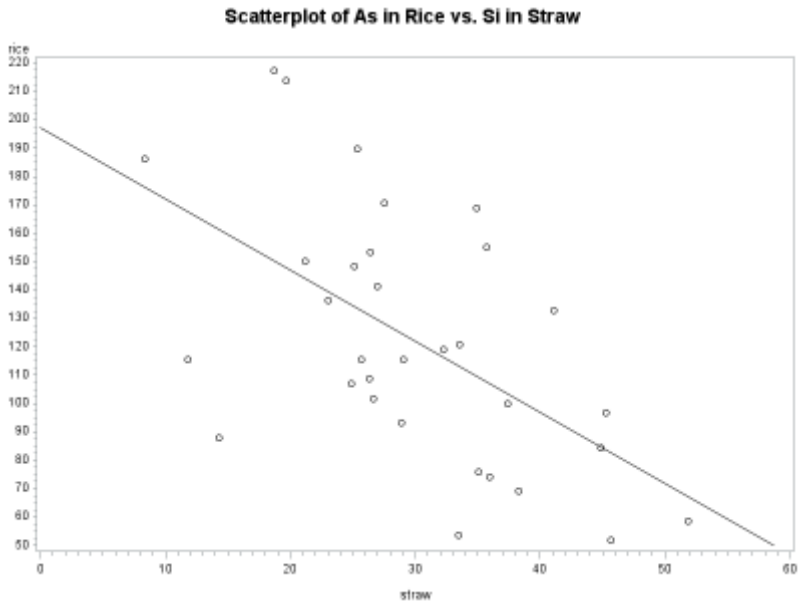
```
data pollutant;
infile 'H:\b10.dat';
input straw rice;
run;

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 c=black;
*the plot statement is y*x.;
proc gplot data=pollutant;
  plot rice*straw;
run;

quit;
```

SAS Graph:



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.

4.2. 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: (b11.sas)

```
data pollutant;
infile 'H:\b10.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 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.

4.3. 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.

SAS Learning code: (b12.sas)

```
data pollutant;
infile 'H:\b10.dat';
input straw rice;
run;

proc print data=pollutant; run;

*objective 4.3 and 4.4;

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*/
*proc reg automatically creates the residual plot and QQplot;
run;
```

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 test statistics and p-value of β_1 are -3.76 and 0.0007, the p-value of β_0 **< 0.0001**. The df of the t-tests is df for the error (within) = **30**.

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?)

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.

4.4 Residual Plots and QQplots

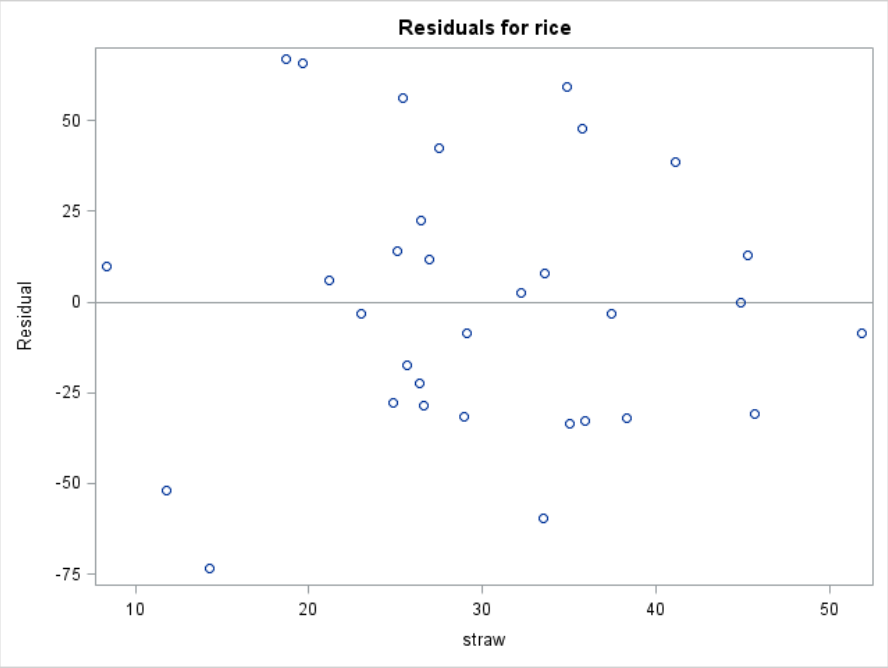
It is always important to check assumptions. In linear regression, these assumptions can be assessed by looking at the residuals.

In linear regression, one assumes that the residuals are independent and normally distributed with constant variance and randomly above and below the x axis (linearity). 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 and linearity, a residual plot is very helpful.

The residual plot, QQplot and histogram are generated automatically in proc reg which we ran in the previous objective.

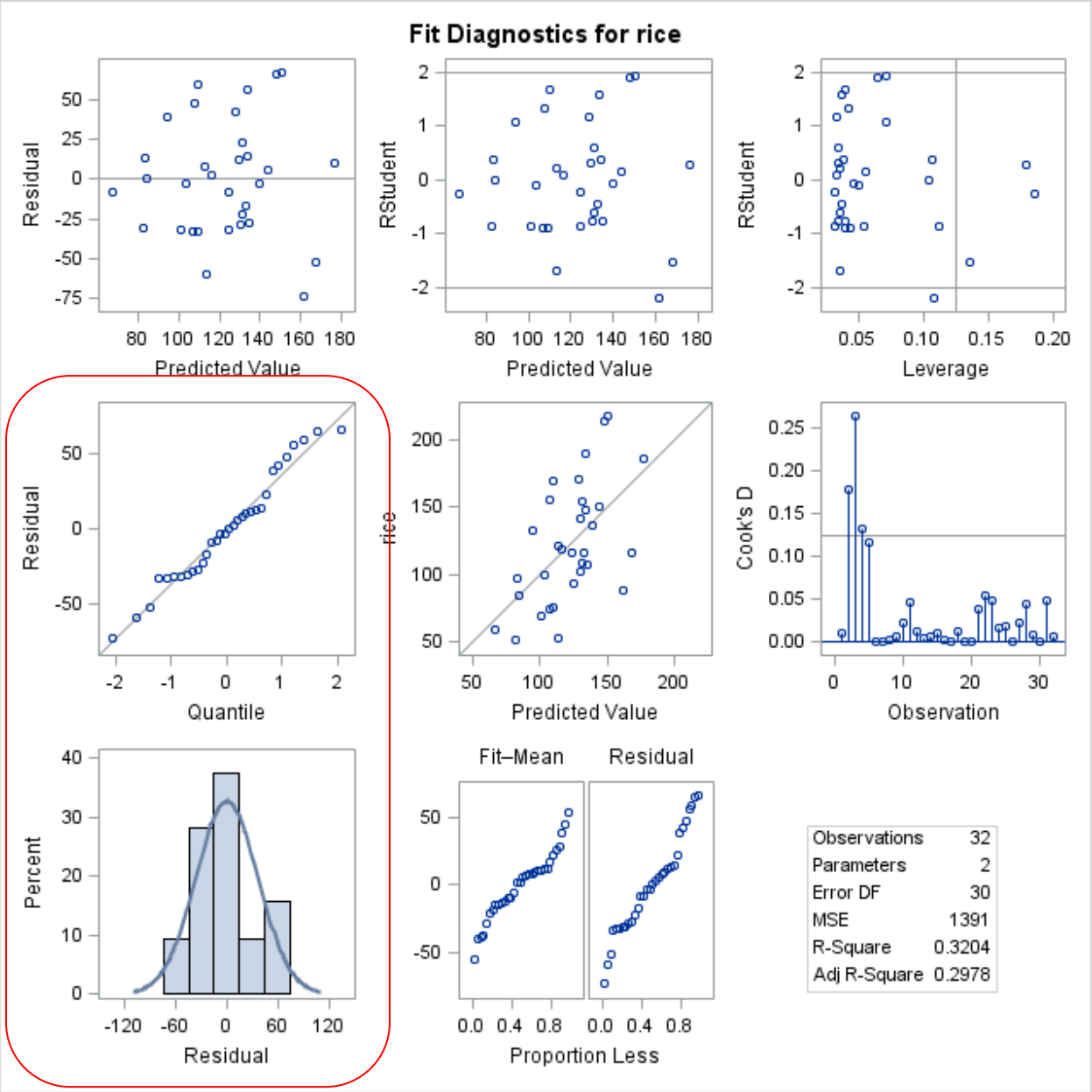
SAS Learning code: (output is generated in Objective 4.3)

SAS Learning graphics:



I see no pattern (linearity) and the variance looks constant.

Normality plots



It looks like the residuals are normal here.

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.