Project 7 (Part 1 of Final): Regression and Correlation

Name:

The following SAS code may be helpful for this assignment:

Simple Linear Regression:

```
proc glm data=dataset;
model y=x;
run;
```

Simple Linear Regression with Confidence Intervals:

```
proc glm data=dataset;
model y=x/clparm;
run;
```

Pearson’s Correlation:

```
proc corr data=dataset;
var var1 var2;
run;
```

Scatterplot:

```
proc sgplot data=dataset noautolegend;
scatter y=var1 x=var2; /* just a scatterplot */
reg y=var1 x=var2; /* scatterplot with reg line */
xaxis label=’labelx’;
yaxis label=’labely’;
run;
```

Correlation Confidence Interval:

```
proc corr data=dataset fishers;
var var1 var2;
run;
```

Shapiro-Wilk’s Test:

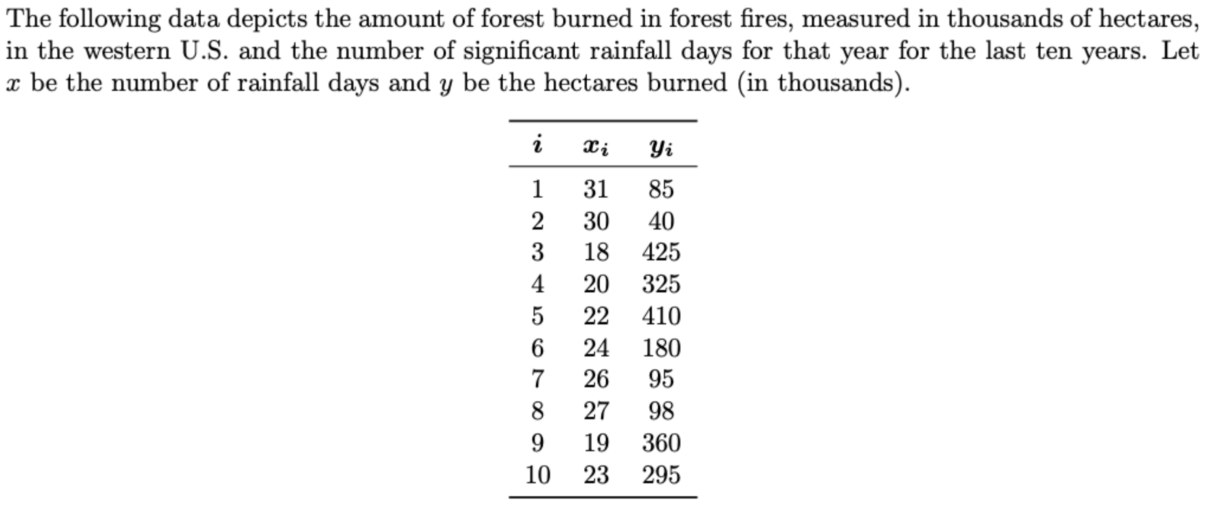
```
proc univariate normal;
var eggs weight;
run;
```

Spearman’s Correlation:

```
proc corr data=dataset spearman;
var var1 var2;
run;
```

Multiple Regression:

```
proc glm data=dataset;
model y=var1 var2 var3;
run; quit;
```



The following data depicts the amount of forest burned in forest fires, measured in thousands of hectares, in the western U.S. and the number of significant rainfall days for that year for the last ten years. Let x be the number of rainfall days and y be the hectares burned (in thousands).

i	x_i	y_i
1	31	85
2	30	40
3	18	425
4	20	325
5	22	410
6	24	180
7	26	95
8	27	98
9	19	360
10	23	295

1. Create a scatter plot of the data. Do you think the slope will be positive or negative?

   I think it will be negative because of the y-values with the x-values of 18 and 19 decrease.

2. Determine whether the regression is significant. Include your SAS output, hypothesis, and conclusion.

   H_0: $\beta_1 = 0$

   H_1: $\beta_1 \neq 0$

   We reject H_0 if p-val < 0.05

   p-val = 0.0002

   Since p-val is less than 0.05, we reject the H_0. This means that the days of rainfall are a significant indicator of hectacres burned.

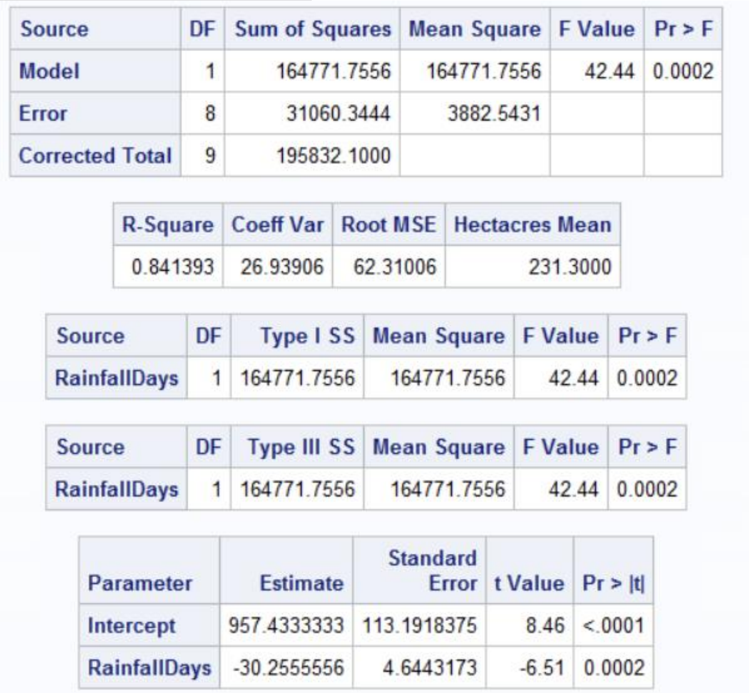
   ```
   Code: data forest;
   input RainfallDays Hectacres @@;
   cards;
   31 85 30 40 18 425 20 325 22 410 24 180 26 95 27 98 19 360 23 295
   ;
   run;

   proc glm data=forest;
   model Hectacres=RainfallDays;
   run;
   ```

3. If the regression is significant, fit the linear regression and write an interpretation of the line. Include your SAS output and code.

$$\hat{y} = 957.43 - 30.25x$$

This means that if there are no days of rainfall, 957.43 hectacres of forest are burned. But, for each day of rainfall, 30.25 less hectacres are burned.

Same Code as before



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	164771.7556	164771.7556	42.44	0.0002
Error	8	31060.3444	3882.5431		
Corrected Total	9	195832.1000			

R-Square	Coeff Var	Root MSE	Hectacres Mean
0.841393	26.93906	62.31006	231.3000

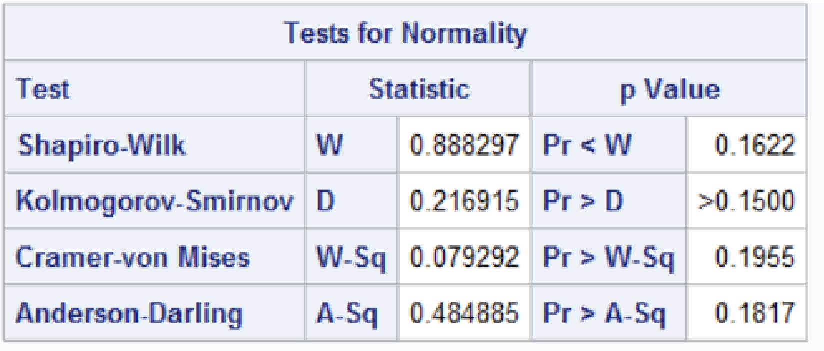
Source	DF	Type I SS	Mean Square	F Value	Pr > F
RainfallDays	1	164771.7556	164771.7556	42.44	0.0002

Source	DF	Type III SS	Mean Square	F Value	Pr > F
RainfallDays	1	164771.7556	164771.7556	42.44	0.0002

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	957.4333333	113.1918375	8.46	<.0001
RainfallDays	-30.2555556	4.6443173	-6.51	0.0002

4. Determine what percentage of the variability in y is explained by the regression. Include your SAS output and code.
   As determined by the R-Square table above, 84% of variability in hectacres burned is explained by the regression
5. Determine which correlation coefficient is appropriate. Justify your answer with SAS output and code.
   I tested for normality and got a p-val of 0.1622 which is more than 0.05 which means that the data is normally distributed. So we run Pearson Correlation

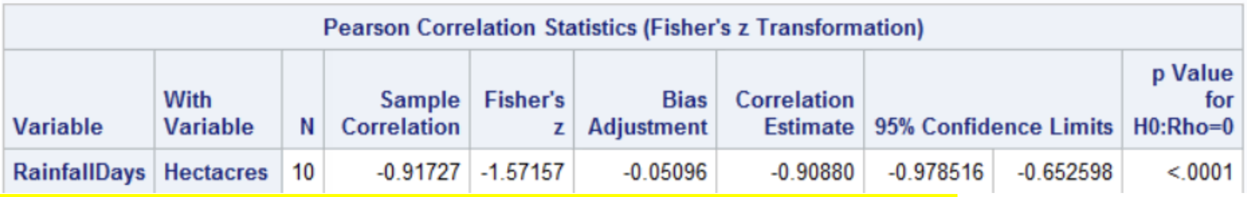
   Code: proc univariate normal data=forest;
   var RainfallDays Hectacres;
   run;



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.888297	Pr < W	0.1622
Kolmogorov-Smirnov	D	0.216915	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.079292	Pr > W-Sq	0.1955
Anderson-Darling	A-Sq	0.484885	Pr > A-Sq	0.1817

6. Calculate the correlation coefficient and determine whether the correlation is significant. Justify your answer with SAS output and code.
   After running Pearson's Correlation, we get a p-val of <0.0001, so the correlation between days of rainfall and hectacres of forest burned is significant.

```
Code: proc corr data=forest fisher;
var RainfallDays Hectacres;
run;
```



Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
RainfallDays	Hectacres	10	-0.91727	-1.57157	-0.05096	-0.90880	-0.978516	-0.652598	<.0001

7. Calculate the 95% confidence interval for the correlation coefficient.

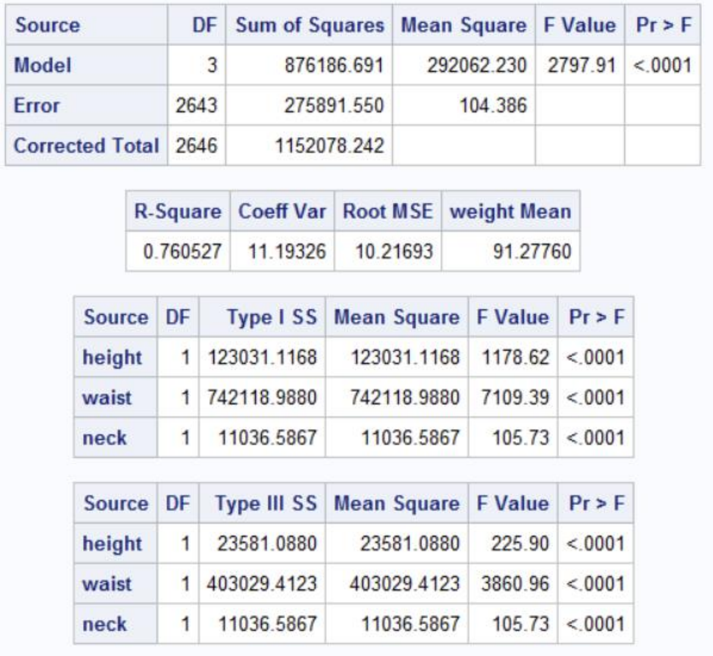
With 95% confidence, the correlation of rainfall days and hectacres of forest burned are between -0.978516 and -0.652598

Download the analysis1.csv file from eLearning and create a SAS data set.

We would like to determine if weight can be modeled from height, waist, and neck.

1. Determine whether the regression is significant. Include your SAS output, hypothesis, and conclusion.
   H_0: β_1= β_2= β_3= β_4=0
   H_1: at least one $\beta \neq 0$
   We reject H_0 if p-val < 0.05
   p-val < 0.0001
   Since p-val is less than 0.05, we can conclude that the regression is significant.

```
Code: pROC IMPORT OUT= WORK.analysis1
DATAFILE= "G:\My Drive\STA5990Data\analysis1.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

proc glm data=analysis1;
model weight=height waist neck;
run;
```



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	876186.691	292062.230	2797.91	<.0001
Error	2643	275891.550	104.386		
Corrected Total	2646	1152078.242			

R-Square	Coeff Var	Root MSE	weight Mean
0.760527	11.19326	10.21693	91.27760

Source	DF	Type I SS	Mean Square	F Value	Pr > F
height	1	123031.1168	123031.1168	1178.62	<.0001
waist	1	742118.9880	742118.9880	7109.39	<.0001
neck	1	11036.5867	11036.5867	105.73	<.0001

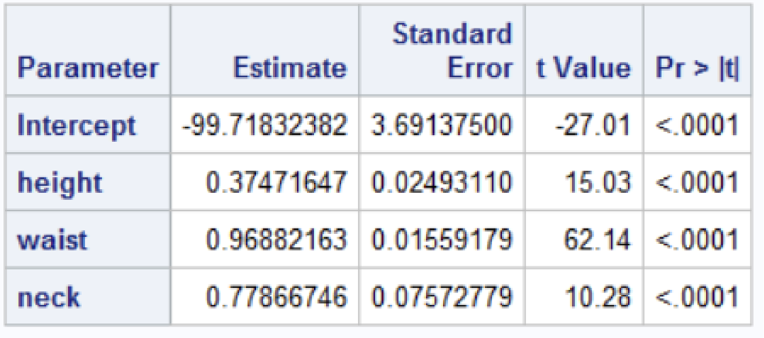
Source	DF	Type III SS	Mean Square	F Value	Pr > F
height	1	23581.0880	23581.0880	225.90	<.0001
waist	1	403029.4123	403029.4123	3860.96	<.0001
neck	1	11036.5867	11036.5867	105.73	<.0001

2. If the regression is significant, fit the linear regression and write an interpretation of the line. Include your SAS output and code.

$$\hat{y} = -99.72 + 0.37x_{height} + 0.97x_{waist} + 0.78x_{neck}$$

Weight increases .37 (pounds?) as height increases by 1 (inch?). Weight increases by .97 pounds as waist increases by 1 inch, and weight also increases by .78 pounds as neck increases by 1 inch.

Same Code



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-99.71832382	3.69137500	-27.01	<.0001
height	0.37471647	0.02493110	15.03	<.0001
waist	0.96882163	0.01559179	62.14	<.0001
neck	0.77866746	0.07572779	10.28	<.0001

3. Determine what percentage of the variability in y is explained by the regression. Include your SAS output and code.

As determined by the R-Square table above, 76% of variability in weight is explained by the regression

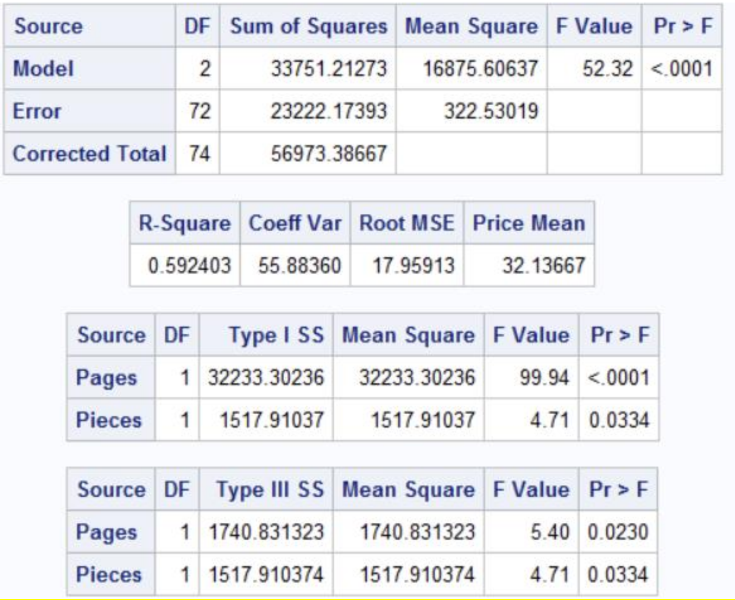
---

Download the lego.sample.csv file from eLearning and create a SAS data set.

We would like to determine whether price can be modeled from number of pieces and pages in the manual.

1. Determine whether the regression is significant. Include your SAS output, hypothesis, and conclusion.
   H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
   H_1: at least one $\beta \neq 0$
   We reject H_0 if p-val < 0.05
   p-val < 0.0001
   Since p-val is less than 0.05, we can conclude that the regression is significant.

   ```
   Code: pROC IMPORT OUT= WORK.lego
   DATAFILE= "G:\My Drive\STA5990Data\lego.sample.csv"
   DBMS=CSV REPLACE;
   GETNAMES=YES;
   DATAROW=2;
   RUN;

   proc glm data=lego;
   model price=pages pieces;
   run;
   ```



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	33751.21273	16875.60637	52.32	<.0001
Error	72	23222.17393	322.53019		
Corrected Total	74	56973.38667			

R-Square	Coeff Var	Root MSE	Price Mean
0.592403	55.88360	17.95913	32.13667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Pages	1	32233.30236	32233.30236	99.94	<.0001
Pieces	1	1517.91037	1517.91037	4.71	0.0334

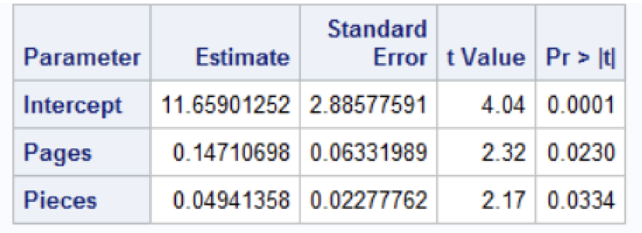
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Pages	1	1740.831323	1740.831323	5.40	0.0230
Pieces	1	1517.910374	1517.910374	4.71	0.0334

2. If the regression is significant, fit the linear regression and write an interpretation of the line. Include your SAS output and code.

$$\hat{y} = 11.66 + 0.05x_{pieces} + 0.15x_{pages}$$

   The legos start off at $11.66 and increase by 5 cents for each piece and 15 cents for each page in the manual.

   Same Code



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.65901252	2.88577591	4.04	0.0001
Pages	0.14710698	0.06331989	2.32	0.0230
Pieces	0.04941358	0.02277762	2.17	0.0334

3. Determine what percentage of the variability in y is explained by the regression. Include your SAS output and code.
   As determined by the R-Square table above, 59% of variability in price is explained by the regression