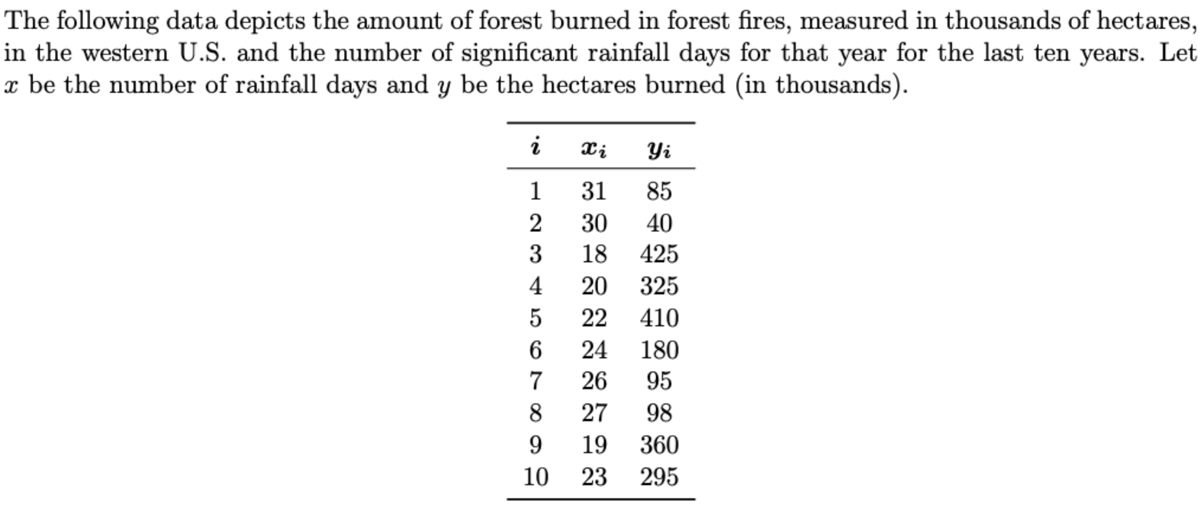
Project 7 (Part 1 of Final): Regression and Correlation

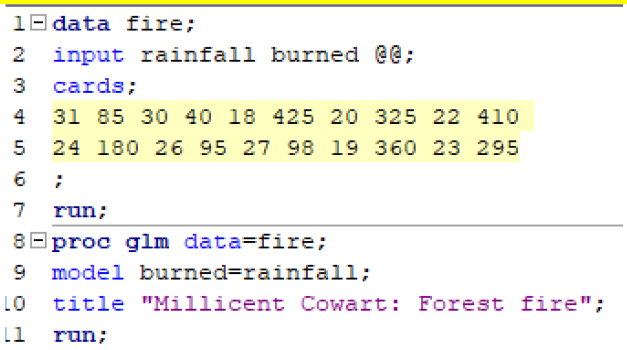
Name: Millicent Cowart



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).

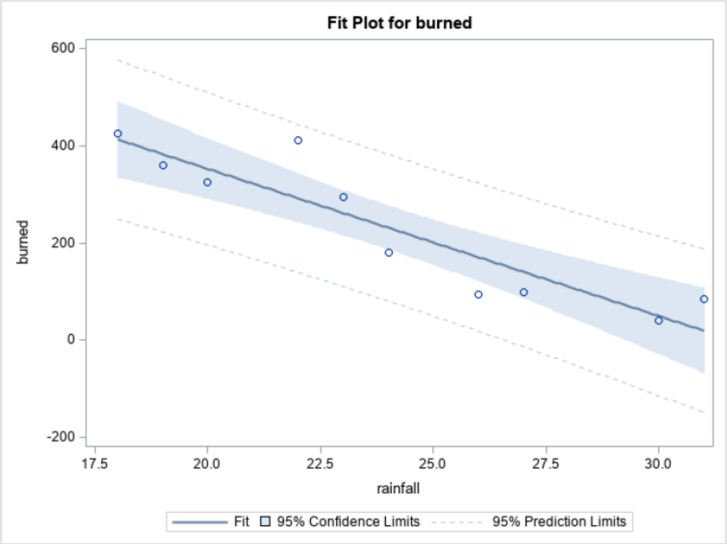
i	x_i	y_i
1	31	85
2	30	40
3	18	425
4	20	325
5	22	410
6	24	180
7	26	95
8	27	98
9	19	360
10	23	295

1. Consider the data above and do the following:

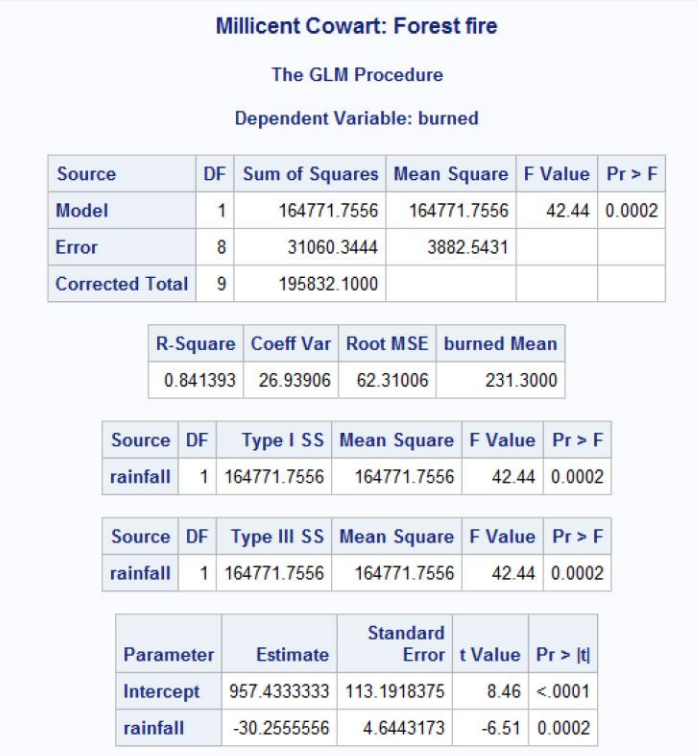


```
data fire;
input rainfall burned @@;
cards;
31 85 30 40 18 425 20 325 22 410
24 180 26 95 27 98 19 360 23 295
;
run;
proc glm data=fire;
model burned=rainfall;
title "Millicent Cowart: Forest fire";
run;
```

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



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



Millicent Cowart: Forest fire

The GLM Procedure

Dependent Variable: burned

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	burned Mean
0.841393	26.93906	62.31006	231.3000

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

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rainfall	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
rainfall	-30.2555556	4.6443173	-6.51	0.0002

H0:$\beta_1 = 0$
H1:$\beta_1 \neq 0$
P-value: 0.0002
Conclusion: Reject the null hypothesis. The slope is not equal to zero, so rainfall is a significant predictor of burning.

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

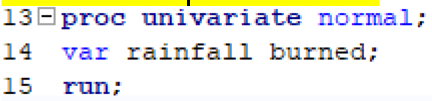
$$\hat{y} = 957.433 - 30.256x$$

For every day increase of rainfall there is an expected decrease of 30.256 hectares burned.

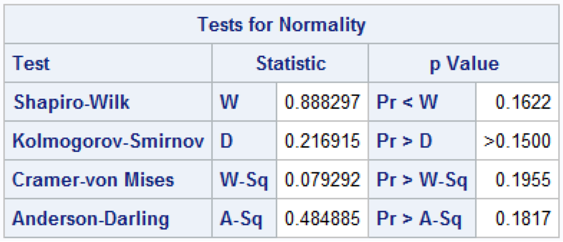
- Determine what percentage of the variability in y is explained by the regression. Include your SAS output and code.

84% of the variability in hectares burned is explained by the regression.

- Determine which correlation coefficient is appropriate. Justify your answer with SAS output and code.



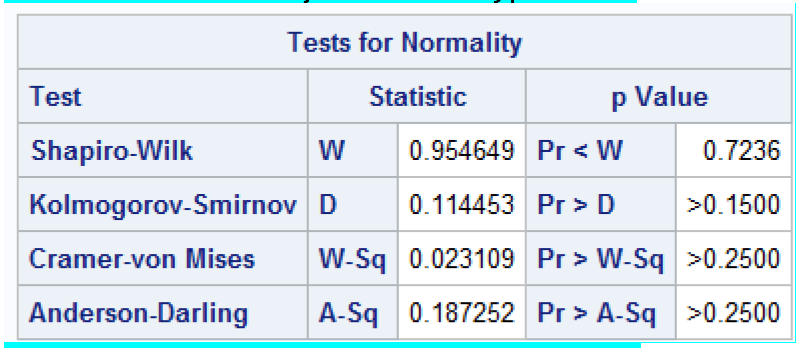
```
proc univariate normal;
 var rainfall burned;
 run;
```



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

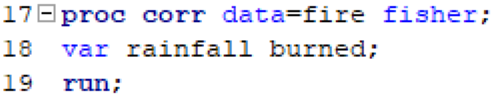
H0: The data (forests burned) are normally distributed.
H1: The data (forests burned) is not normally distributed.
P-Value: 0.1622
Conclusion: Fail to reject the null hypothesis.



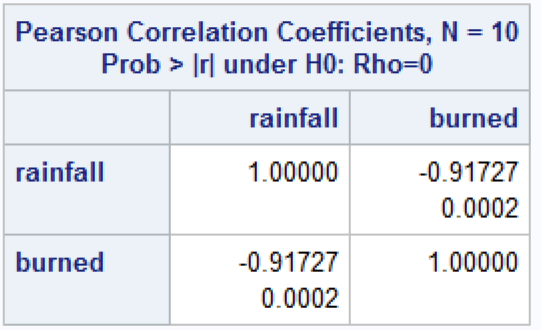
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.954649	Pr < W	0.7236
Kolmogorov-Smirnov	D	0.114453	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.023109	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.187252	Pr > A-Sq	>0.2500

H0: The data (rainfall) is normally distributed.
H1: The data (rainfall) is not normally distributed.
P-value: 0.7236
Conclusion. Fail to reject the null hypothesis.
Since both variables are normally distributed, use Pearson's correlation coefficient.

- Calculate the correlation coefficient and determine whether the correlation is significant. Justify your answer with SAS output and code.



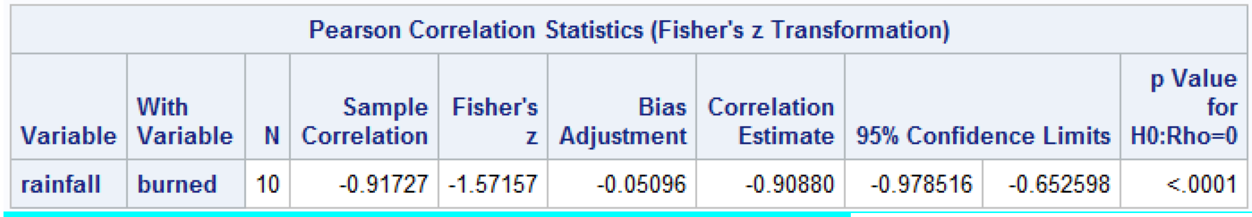
```
proc corr data=fire fisher;
var rainfall burned;
run;
```



Pearson Correlation Coefficients, N = 10 Prob > |r| under H0: Rho=0		
	rainfall	burned
rainfall	1.00000	-0.91727 0.0002
burned	-0.91727 0.0002	1.00000

The correlation coefficient is -0.91727 and it is significant.

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



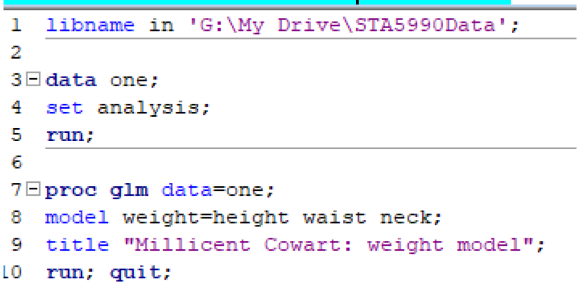
Pearson Correlation Statistics (Fisher's z Transformation)									
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
rainfall	burned	10	-0.91727	-1.57157	-0.05096	-0.90880	-0.978516	-0.652598	<.0001

The 95% confidence interval for the correlation coefficient is (-0.9785, -0.6526).

2. 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.

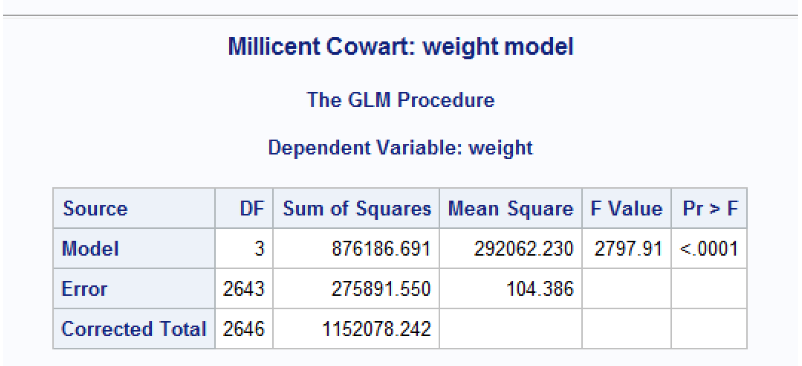
- Determine whether the regression is significant. Include your SAS output, hypothesis, and conclusion.
  H0: $\beta_1 = \beta_2 = \beta_3 = 0$
  H1: At least one is not equal to zero



```
libname in 'G:\My Drive\STA5990Data';

data one;
set analysis;
run;

proc glm data=one;
model weight=height waist neck;
title "Millicent Cowart: weight model";
run; quit;
```



Millicent Cowart: weight model

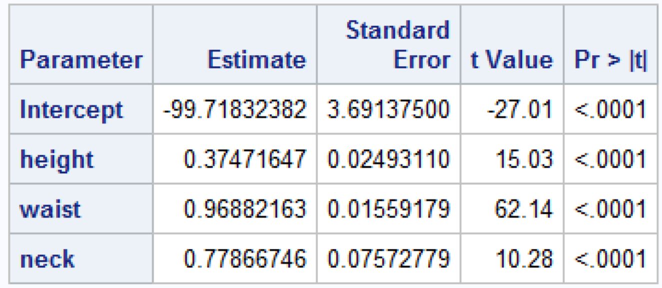
The GLM Procedure

Dependent Variable: weight

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			

P-value: <0.0001
Conclusion: Reject the null hypothesis. The slope is not equal to zero, so height, waist, and neck are significant predictors of weight.

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

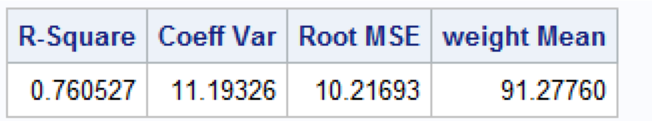


Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-99.71832382	3.69137500	-27.01	<.0001
height	0.37471647	0.02493110	15.03	<.0001
waist	0.96882163	0.01559179	62.14	<.0001
neck	0.77866746	0.07572779	10.28	<.0001

$$\hat{y} = -99.72 + 0.37x_h + 0.97x_\omega + 0.78x_n$$

For each increase in height, there is an expected 0.37 increase in weight.
For each increase in waist, there is an expected 0.97 increase in weight.
For each increase in neck, there is an expected 0.78 increase in weight.

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



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

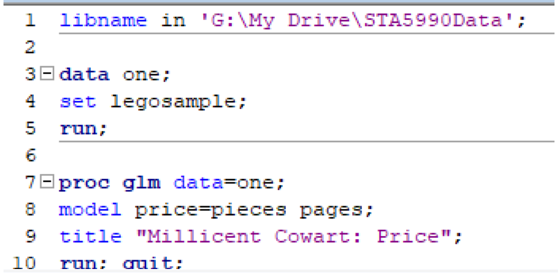
76% of the variability in weight is explained by the regression.

---

3. 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.

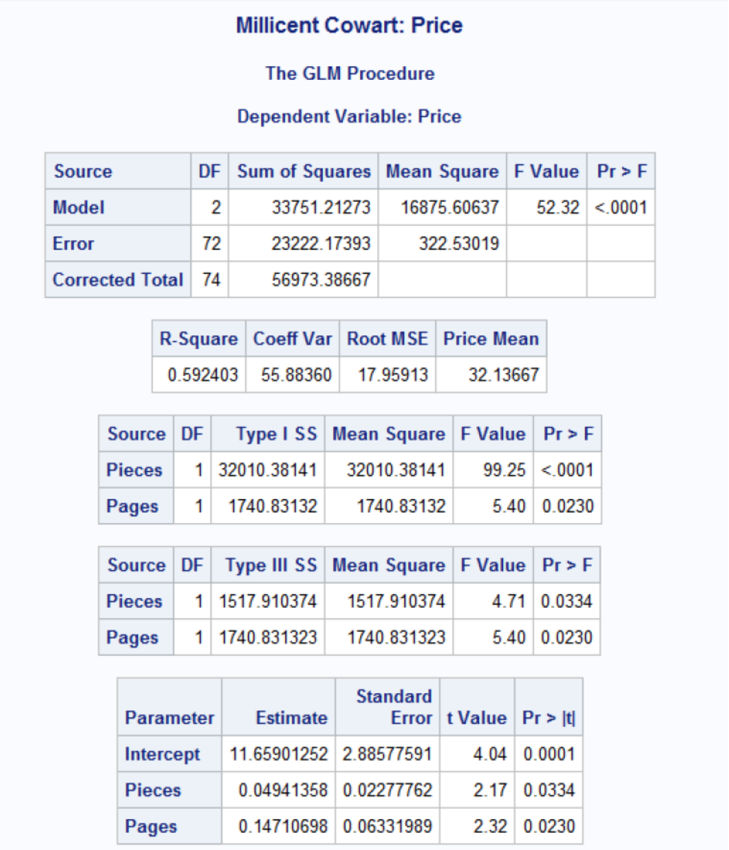
- Determine whether the regression is significant. Include your SAS output, hypothesis, and conclusion.
  H0:$\beta_1 = \beta_2 = 0$
  H1: At least one is not equal to zero



```
libname in 'G:\My Drive\STA5990Data';

data one;
set legosample;
run;

proc glm data=one;
model price=pieces pages;
title "Millicent Cowart: Price";
run; quit;
```



Millicent Cowart: Price

The GLM Procedure

Dependent Variable: Price

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
Pieces	1	32010.38141	32010.38141	99.25	<.0001
Pages	1	1740.83132	1740.83132	5.40	0.0230

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

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

P-value: <0.001
Conclusion: Reject the null hypothesis. Pieces and pages are significant predictors of price.

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

  $$\hat{y} = 11.66 + 0.049x_{pi} + 0.147x_{pa}$$

  For each increase in piece, there is an expected 0.049 *increase in price*.

  For each increase in pages, there is an expected 0.147 *increase in price*.

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

  59% of the variability in price is explained by the regression.