SAS Tutorial for STAT 350 Lab 9

Author: Leonore Findsen, Cheng Li

Example: (Data Set: loc.txt)

Job Stress and Locus of Control Many factors, such as the type of job, education level, and job experience, can affect the stress felt by workers on the job. Locus of control (LOC) is a term in psychology that describes the extent to which a person believes he or she is in control of the events that influence his or her life. Is feeling "more in control" associated with less job stress? A recent study examined the relationship between LOC and several work-related behavioral measures among certified public accountants in Taiwan. LOC was assessed using a questionnaire that asked respondents to select one of two options for each of 23 items. Scores ranged from 0 to 23. Individuals with low LOC believe that their own behavior and attributes determine their rewards in life. Those with high LOC believe that these rewards are beyond their control. Each accountant's job stress was assessed using the averaged score on 22 items, each scored on a five-point scale. The higher the score, the higher the perceived job stress. We will consider a random sample of 100 accountants.

a) Make a scatterplot of the data (including the least squares regression line) with LOC on the x axis and Stress on the Y axis. Briefly describe the relationship between the job stress and LOC.
b) Compute the correlation coefficient between Stress vs. LOC.
c) Find the equation of the least-squares regression line for predicting Stress from LOC.
d) What is r^2 for these data?
e) Obtain the residuals and plot them versus LOC. Is there anything unusual to report? Please explain.
f) Do the residuals appear to be approximately Normal? Explain your answer.
g) Based on your answers for parts (a), (e) and (f), do the assumptions for the linear regression analysis appear reasonable? Explain your answer.
h) Construct and interpret the 95% confidence interval for the slope and y-intercept.
i) Does Job Stress increase with LOC? Carry out a test of significance on the slope. State hypotheses, give a test statistic and *P*-value, and state your conclusion.
j) Briefly summarize what your data analysis shows.

Solution:

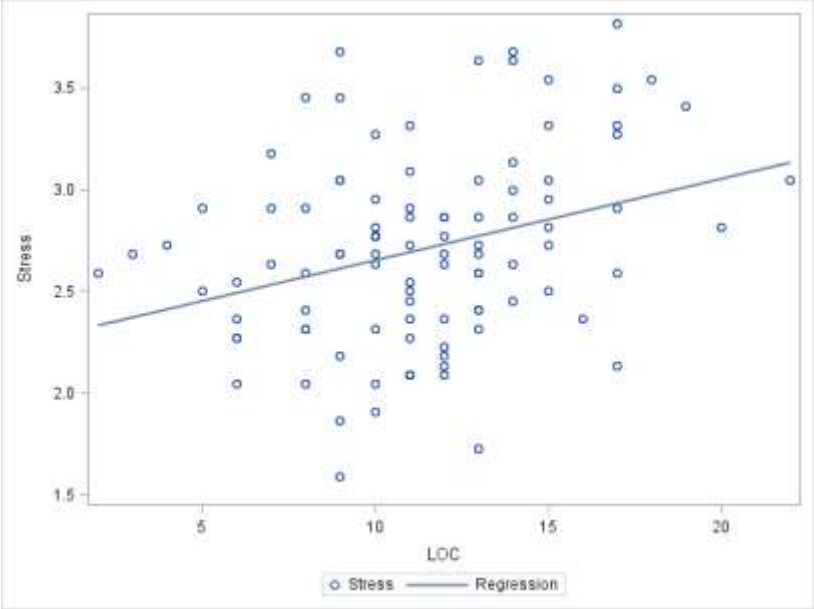
```
data job;
  infile 'W:\PC-Text\loc.txt' firstobs = 2 delimiter = '09'x;
  input Subject LOC Stress;
run;

*Scatter plot;
proc sgplot data=job;
   scatter y=Stress x = LOC;
   reg y=Stress x = LOC;
run;

proc corr data=job;
  var LOC Stress; *Only these variables will be printed out;
  run;
```

SAS Tutorial for STAT 350 Lab 9

Author: Leonore Findsen, Cheng Li

```
*linear regression;
*   The plot command allows more data points to be printed in
      the diagnostic plots;
proc reg data=job plot(maxpoints = 10000);
  model Stress = LOC/clb;
  *clb performs the confidence interval of the 'b's
    that is, the parameters;
run;
```

a) Make a scatterplot of the data (including the least squares regression line) with LOC on the x axis and Stress on the Y axis. Briefly describe the relationship between the job stress and LOC.



The plot looks linear with a positive correlation. However, there may be a problem with constant standard deviation at the low and high values of LOC. I am not sure about the strength because the scale on the y-axis is so small. I do not see any outliers.

b) Find the correlation between Stress vs. LOC.

The SAS System

The CORR Procedure

2 Variables:	LOC Stress

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
LOC	100	11.40000	3.69821	1140	2.00000	22.00000
Stress	100	2.71045	0.47263	271.04544	1.59091	3.81818

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	LOC	Stress
LOC	1.00000	0.31228
		0.0016
Stress	0.31228	1.00000
	0.0016	

The correlation coefficient between Stress vs. LOC is 0.31228.
This looks like there is a weak association between Stress and LOC. Therefore, the strength is low.

Note: only include the last table circled in red in your report.

c) Find the equation of the least-squares regression line for predicting Stress from LOC.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2.25550	0.14691	15.35	<.0001	1.96395	2.54704
LOC	1	0.03991	0.01226	3.25	0.0016	0.01557	0.06425

Stress = 2.2550 + 0.03991 LOC

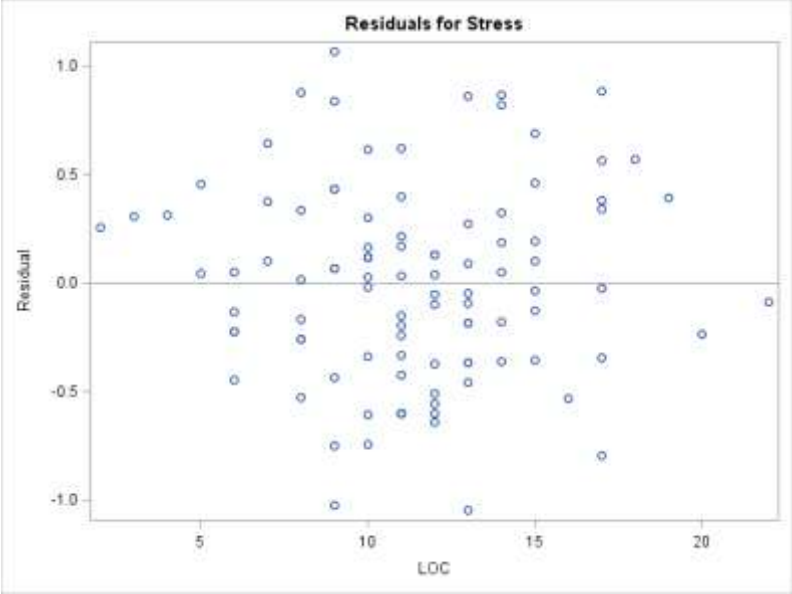
d) What is r^2 for these data?

Root MSE	0.45128	R-Square	0.0975
Dependent Mean	2.71045	Adj R-Sq	0.0883
Coeff Var	16.64948		

$r^2 = 0.0975$

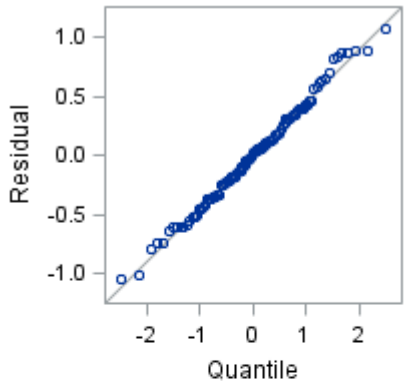
This does not look very good.

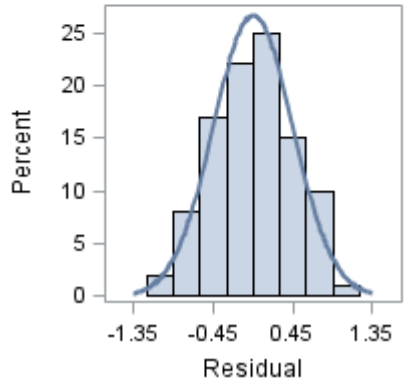
e) Obtain the residuals and plot them versus LOC. Is there anything unusual to report? Please explain.



I see no pattern here so the association seems to be linear. There might be a problem with constant standard deviation at the high and low values. Since the scale of the residuals is so small, I would say that constant standard deviation is valid. I do not see any outliers.

f) Do the residuals appear to be approximately Normal? Explain your answer.





It looks like the residuals are normal because on the QQ plot the points are close to the line and the line on the histogram seems to match the histogram.

g) Based on your answers for parts (a), (e) and (f), do the assumptions for the linear regression analysis appear reasonable? Explain your answer.

Assuming that we have an SRS, the three other assumptions are met; linear, constant standard deviation of the residuals and normality of the residuals, therefore linear regression analysis appears to be reasonable.

h) Construct and interpret the 95% confidence interval for the slope and y-intercept.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2.25550	0.14691	15.35	<.0001	1.96395	2.54704
LOC	1	0.03991	0.01226	3.25	0.0016	0.01557	0.06425

SAS Tutorial for STAT 350 Lab 9

Author: Leonore Findsen, Cheng Li

Slope:
95% CI (0.01557, 0.06425)
We are 95% confident that the population slope of Stress vs. LOC is between 0.01557 and 0.06425,

Intercept:
95% CI (1.96395, 2.54704)
We are 95% confident that the population y-intercept is between of Stress vs. LOC is between 1.96395 and 2.54704.

i) Does Job Stress increase with LOC? Carry out a test of significance on the slope. State hypotheses, give a test statistic and *P*-value, and state your conclusion.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2.25550	0.14691	15.35	<.0001	1.96395	2.54704
LOC	1	0.03991	0.01226	3.25	0.0016	0.01557	0.06425

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.15651	2.15651	10.59	0.0016
Error	98	19.95776	0.20365		
Corrected Total	99	22.11426			

Step 1: Definition of the terms
β_1 is the population slope

Step 2: State the hypotheses
H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

Step 3: Find the *Test Statistic, p-value, report DF*
$t_{ts} = 3.25$
DF = 98
P-value = 0.0016
(Note that the F test statistic = 10.59 = 3.25^2 and the P-values are identical)

Step 4: Conclusion:
$\alpha = 0.05$
Since 0.0016 ≤ 0.05, we should reject H_0
The data provides strong evidence (P-value = 0.0016) that there is an association between job stress and LOC.

j) Briefly summarize what your data analysis shows.

Assuming that the standard deviation is constant, the assumptions are met. The data shows that there is a slight association between Stress and LOC. The weak association is also seen by the small values of r and r^2. Therefore, there is a possibility that there is a slight association, but prediction is not recommended from this study because of the small value of r^2.