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.

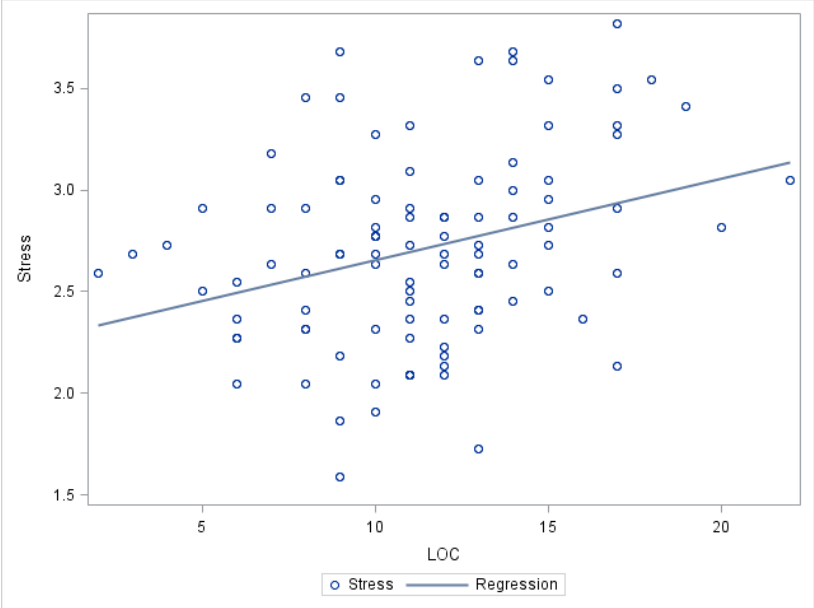
a) Make a scatterplot of the data (including the least-squares regression line) with LOC on the x-axis and Stress on the y-axis. Briefly describe the relationship between Stress and LOC.
b) Compute the correlation coefficient between Stress and LOC.
c) Find the equation of the least-squares regression line for predicting Stress from LOC.
d) What is R^2 for these data?
e) Plot the residuals 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 intervals for the slope and y-intercept.
i) Is Stress associated 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:\loc.txt' firstobs = 2 delimiter = '09'x;
  input Subject LOC Stress;
run;

/* This data does not need to be subsetted in any way, but we will show
you how to subset such that only a specified range of values remains in
the data. */
/* Suppose we want to select the data with Stress level being greater
than 0 and less than 5 below. Note that common logical operators in SAS
were introduced in Lab 7. */

data job_subset;
  set job;
  if 0 < Stress AND Stress < 5;
run;

*Scatter plot;
proc sgplot data = job_subset;
```

```
  scatter y = Stress x = LOC;
  reg y = Stress x = LOC;
run;

*Correlation
 The 'noprob' option prevents output of p-values for the correlations;
proc corr data = job_subset noprob;
  var LOC Stress;
run;
*Linear regression and the rest of the diagnostic plots;
proc reg data = job_subset;
  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 Stress and LOC.



The plot looks linear with a positive direction. I am not sure about the strength because the scale on the y-axis is so small. I do not see any outliers.

b) Compute the correlation coefficient between Stress and LOC.

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

	LOC	Stress
LOC	1.00000	0.31228
Stress	0.31228	1.00000

The correlation coefficient between Stress and LOC is 0.31228.
This looks like there is a weak but nonnegligible association between Stress and LOC.

Note: only include the last table 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

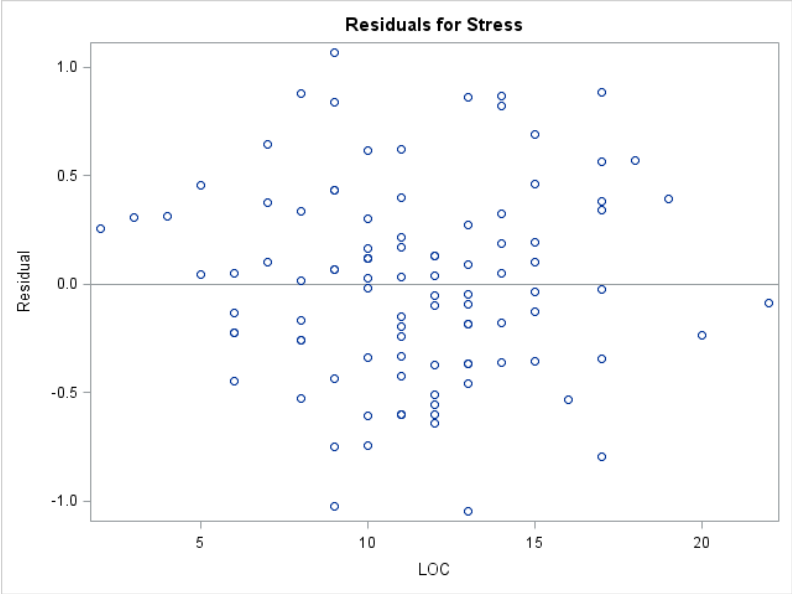
Stress = 2.25550 + 0.03991 LOC

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		

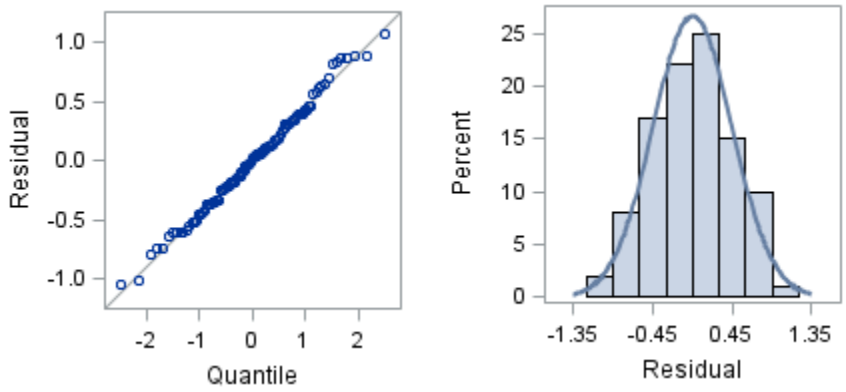
R^2 = 0.0975
This does not look very good.

e) Plot the residuals versus LOC. Is there anything unusual to report? Please explain.



I see no pattern here so the association between Stress and LOC seems to be linear. There is a possibility that the standard deviation is not constant, but that could be due to the fact that there are only a few points at higher and lower ranges, making it harder to assess the true variability. I do not see any outliers.

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



I know that we stated that all histograms should have the 'two' lines. However, this is not required if the histogram is automatically generated as a diagnostic from a procedure.

It looks like the residuals are normal because on the QQ plot the points are close to the line without systematic deviation. The histogram reveals a symmetric, unimodal pattern, and the shape closely matches that of the overlaid estimated normal density curve.

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 relationship, constant standard deviation of the residuals and normality of the residuals, the linear regression analysis appears to be reasonable.

h) Construct and interpret the 95% confidence intervals 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

Slope:
95% CI (0.01557, 0.06425)
We are 95% confident that the population slope of Stress vs. LOC is covered by the interval from 0.01557 to 0.06425.

Note that the slope is indicated by the x variable in the output.

y-intercept:
95% CI (1.96395, 2.54704)
We are 95% confident that the population y-intercept of Stress vs. LOC is covered by the interval from 1.96395 to 2.54704.

i) Is Stress associated 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			

Be sure to include output that shows what the degrees of freedom are.

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 = $10.59 \approx 3.25^2 = t_{ts}^2$, and the *p*-values are identical.)

Step 4: Conclusion:
$\alpha = 0.05$
Since $0.0016 \leq 0.05$, we should reject H_0
The data provide evidence (*p*-value = 0.0016) to the claim that there is an association between Stress and LOC.

j) Briefly summarize what your data analysis shows.

Assuming that the standard deviation is close to being constant, the assumptions are met. The data show that there is an association between Stress and LOC. However, the small values of correlation *r* and R^2 indicate that the association is weak. Therefore, the study shows that there is a slight association, but prediction is not recommended because of the small value of R^2.