R Tutorial for STAT 350 for Lab 9

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke

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.

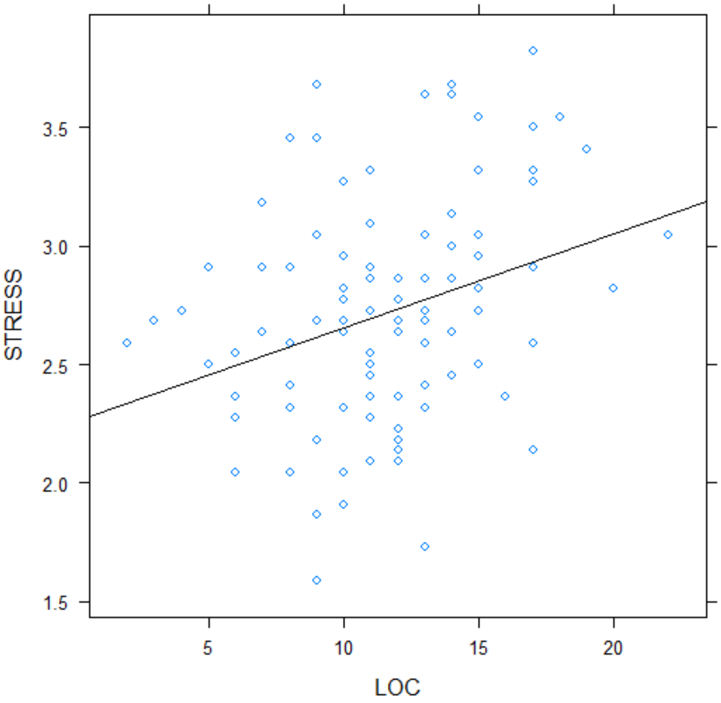
a) Make a scatterplot of the data (including the least squares regression line) with LOC on the x axis and Stress on the Y axis. Briefly describe the relationship between the job stress and LOC.
b) Compute the correlation coefficient between Stress vs. LOC.
c) Find the equation of the least-squares regression 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:

code:

```
> job <- read.table(file = "loc.txt", header = TRUE)
> attach(job)
> #a)
> library(lattice)
> xyplot(STRESS ~ LOC,
      data = job,
      panel = function(x, y){
            panel.xyplot(x, y)
            panel.lmline(x, y)
            })
> #b) correlation
> cor(LOC, STRESS)
> #c), d), i) calculate linear regression and get results
> job.lm = lm(STRESS ~ LOC)
> summary(job.lm)
> #e) calculate the residuals
> job.resid = job.lm$res #Extract residuals obtained in
     job.lm operation
> xyplot(job.resid ~ LOC,
      data = job,
      main=”Residual plot”,
      ylab = "Residual",
      panel = function(x, y){
            panel.xyplot(x, y)
            panel.abline(h = 0)
      })
> #f) Calculate the histogram and qqplot on the residuals
> #     please see previous labs for this
> #     Note: this is a single sample
> #h) Generate the 2-sided Confidence Interval (CI) for the parameters
> confint(job.lm, level = 0.95)
> #NOTE: This can also be done by hand from output of summary(job.lm)
> #   However, in this lab, you must use the code above.
```

a) Make a scatterplot of the data (including the least squares regression line). 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) Compute the correlation coefficient between Stress vs. LOC.

```
> cor(LOC, STRESS)
[1] 0.3122765
```

The correlation coefficient between Stress vs. LOC is 0.3122765.
This looks like there is a weak association between Stress and LOC. Therefore, the strength is low.

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

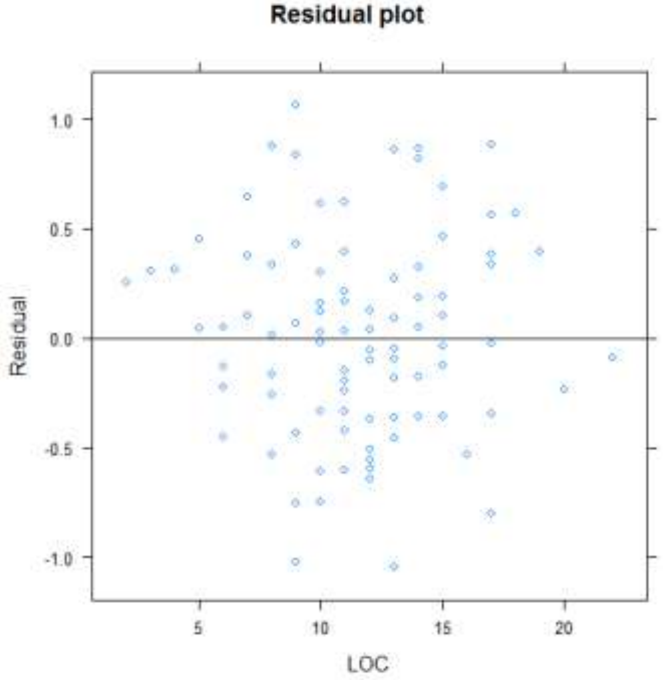
```
> summary(job.lm)

Call:
lm(formula = STRESS ~ LOC)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.04704 -0.33806  0.02169  0.30798  1.06715 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.25550    0.14691  15.353  < 2e-16 ***
LOC          0.03991    0.01226   3.254  0.00156 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4513 on 98 degrees of freedom
Multiple R-squared:  0.09752,	Adjusted R-squared:  0.08831 
F-statistic: 10.59 on 1 and 98 DF,  p-value: 0.001562
```

Stress = 2.25550 + 0.03991 LOC

d) What is r^2 for these data?

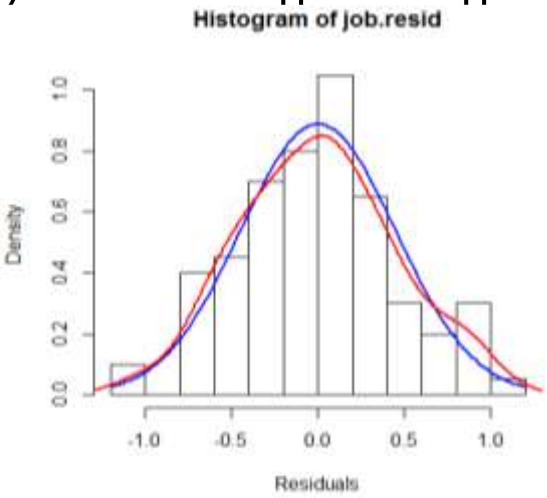
R^2 = 0.09752
This does not look very good.

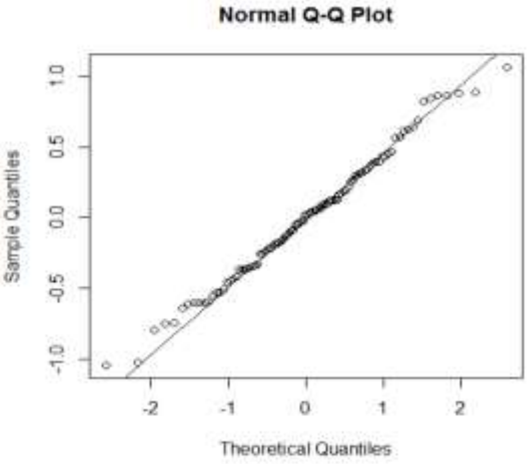
e) Obtain the residuals and plot them versus LOC. Is there anything unusual to report? Please explain.



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 blue/red lines on the histogram seems to be close.

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.

```
                    2.5 %     97.5 %
(Intercept) 1.96395317 2.54704023
LOC         0.01557099 0.06424615
```

Slope:
95% CI (0.01557099, 0.06424615)
We are 95% confident that the population slope of Stress vs. LOC is between 0.0155099 and 0.06424615.

Intercept:
95% CI (1.96395317, 2.54704023)
We are 95% confident that the population y-intercept of Stress vs. LOC is between 1.96395317 and 2.54704023.

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

You do not need to repeat the output. I am doing it so that I can indicate what values that I am using.

```
Call:
lm(formula = STRESS ~ LOC)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04704 -0.33806  0.02169  0.30798  1.06715

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.25550    0.14691  15.353  < 2e-16 ***
LOC          0.03991    0.01226   3.254  0.00156 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4513 on 98 degrees of freedom
Multiple R-squared:  0.09752,   Adjusted R-squared:  0.08831
F-statistic: 10.59 on 1 and 98 DF,  p-value: 0.001562
```

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*

$t_{ts} = 3.254$
DF = 98
P-value = 0.00156
(Note that the F test statistic = 10.59 = 3.254^2 and the P-values are identical)

Step 4: Conclusion:

$\alpha = 0.05$
Since $0.00156 \leq 0.05$, we should reject H_0
The data provides sufficiently strong evidence (P-value = 0.00156) to the claim 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.