# R Tutorial for STAT 350 for Lab 9

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

**Example: (Data Set: loc.txt – website)**

**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). Briefly describe the relationship between the job stress and LOC.
b) Obtain the residuals and plot them versus LOC. Is there anything unusual to report? Please explain.
c) Do the residuals appear to be approximately Normal? Explain your answer.
d) Based on your answers for parts (a), (b) and (c), do the assumptions for the linear regression analysis appear reasonable? Explain your answer.
e) Find the equation of the least-squares regression line for predicting Stress from LOC.
f) What is $r^2$ for these data?
g) Briefly summarize what your data analysis shows.
h) Construct and interpret the 95% confidence interval for the slope and y-intercept.
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.
j) Seung Jin Lim has worked as an accountant for several years and has LOC = 16. Predict his level of job stress.
k) What is the 95% confidence interval for the mean stress level of all accountants with LOC = 16?
l) What is the 95% prediction interval for the next accountant with LOC = 16?

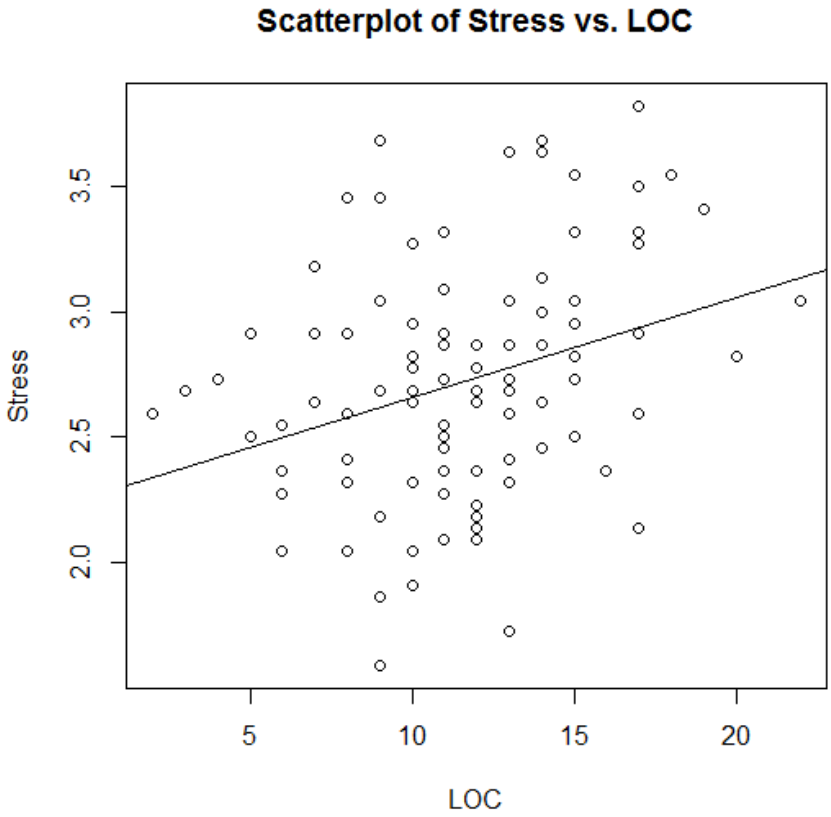
**Solution:**

First you need to read in the data as before; I put it in a variable called 'job'.

```
> job <- read.table(file = "loc.txt", header = TRUE)
> attach(job)
> #a)
> library(lattice)
> xyplot(STRESS ~ LOC,
      data = job,
```

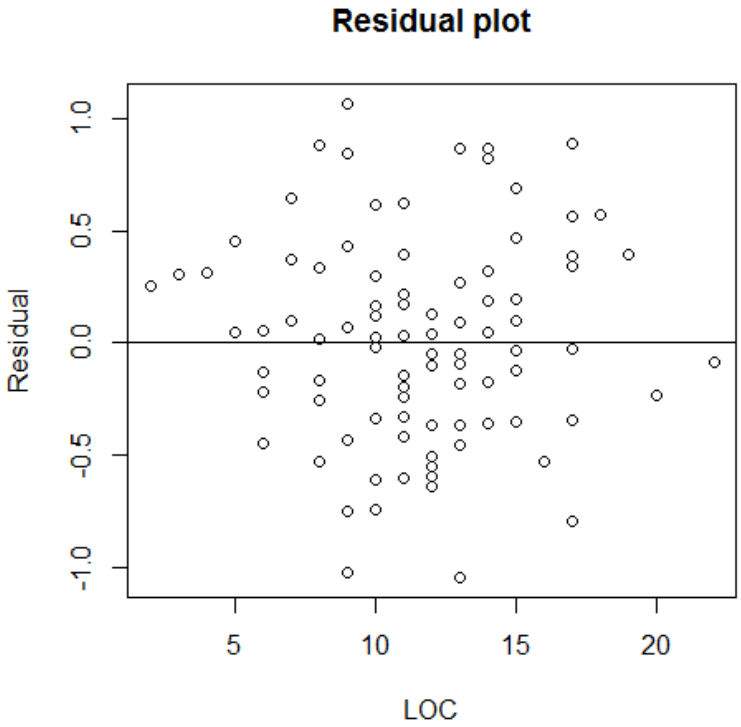
```
      panel = function(x, y){
            panel.xyplot(x, y)
            panel.lmline(x, y)
            })
> #b) to generate the linear regression and calculate the residuals
> job.lm = lm(STRESS ~ LOC)
> job.resid = job.lm$res #Extract residuals obtained in job.lm operation
> xyplot(job.resid ~ LOC,
      data = job,
      main="Residual plot",
      ylab = "Residual",
      panel = function(x, y){
            panel.xyplot(x, y)
            panel.abline(h = 0)
      })
> #d) Calculate the histogram and qqplot on the residuals
> #e)-j) Results from linear regression
> summary(job.lm)
> #h)-i) 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)
> #j)-k) 95% CI of the true population mean
> newdata <- data.frame(LOC=16)
> predict(job.lm, newdata, interval = "confidence")
> # l) 95% Prediction Interval (PI)...more variable than CI
> # l) Why? More unknown about a prediction than confidence/CLT
> predict(job.lm, newdata, interval = "predict")
```

**a) Make a scatterplot of the data (including the least squares regression line). Briefly describe the relationship between the job stress and LOC.**



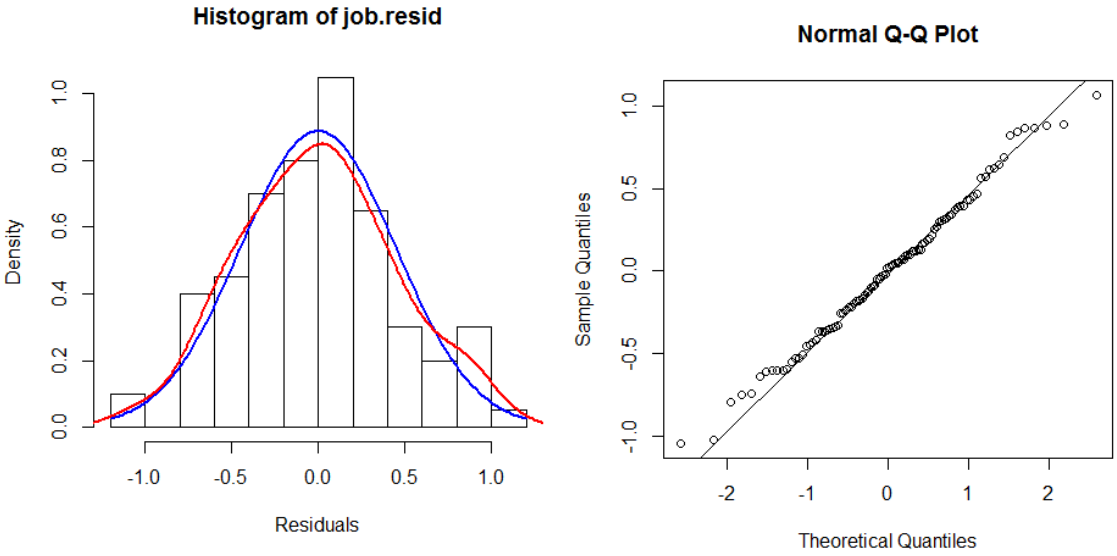

This looks linear with a positive correlation. However, there may be a problem with constant standard deviation at the low and high values of LOC.

**b) 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. Since the scale of the residuals is so small, I would say that constant standard deviation is valid.

**c) 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.

**d) Based on your answers for parts (a), (b) and (c), 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.

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

```
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

**f) What is $r^2$ for these data?**

$R^2$ = 0.09752

**g) Briefly summarize what your data analysis shows.**

Since the assumptions are met, we can use inference in this situation. However, because the $R^2$ value is so low, I would say that the scatter would prevent valid prediction.

**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 is between 0.0155099 and 0.06424615

Intercept:
95% CI (1.96395317, 2.54704023)
We are 95% confident that the population y-intercept 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.**

**Step 0: Definition of the terms**
$\beta_1$ is the population slope

**Step 1: State the hypotheses**
$H_0$: $\beta_1 = 0$
$H_a$: $\beta_1 \neq 0$

**Step 2: Find the *Test Statistic, report DF*.**
$t_t = 3.254$
DF = 98

**Step 3: Find the *p-value:***
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) that there is an association between job stress and LOC.

**j) Seung Jin Lim has worked as an accountant for several years and has LOC = 16. Predict his level of job stress.**

```
       fit      lwr      upr
1 2.894034 2.750669 3.037399
```

The predicted value is 2.894037

**k) What is the 95% confidence interval for the mean stress level of all accountants with LOC = 16? Please interpret the interval.**

95% CI for mean: (2.750669, 3.037399)
We are 95% confidence that the population mean stress is between 2.750669 and 3.037399 when LOC = 16.

**j) What is the 95% prediction interval for the next accountant with LOC = 16?**

```
       fit      lwr     upr
1 2.894034 1.987087 3.80098
```

95% prediction interval: (1.987087, 3.80098)
We are 95% confidence that the next stress value is between 1.987087 and 3.80098 when LOC = 16.