# Model Questions

STA 506 2.0 Linear Regression Analysis

Thiyanga Talagala

05/12/2020

Answers: in class discussion on 12 Dec 2020.

**Use 5% significance level for all tests.**

## Question 1

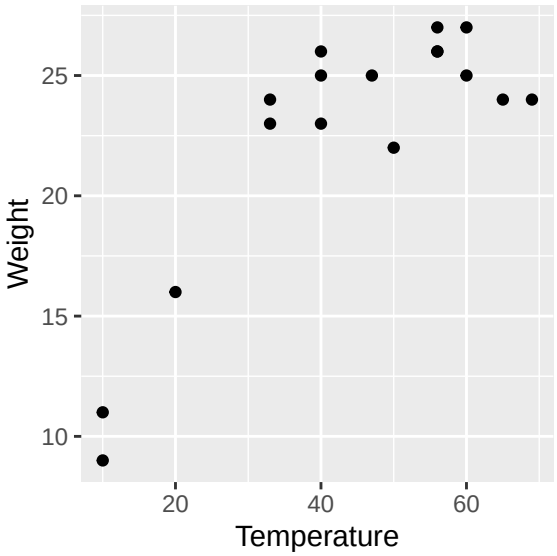
A chemical reaction is performed at different levels of temperature (Celsius) and the end product is weighed (g). The following results were obtained for the purpose of finding a regression model to represent the relationship of the two variables.

| | Temperature | Weight |
|---|---|---|
| 1 | 10 | 11 |
| 2 | 10 | 9 |
| 3 | 20 | 16 |
| 4 | 33 | 23 |
| 5 | 33 | 24 |
| 6 | 40 | 25 |
| 7 | 40 | 26 |
| 8 | 40 | 23 |
| 9 | 47 | 25 |
| 10 | 50 | 22 |
| 11 | 56 | 26 |
| 12 | 56 | 27 |
| 13 | 56 | 26 |
| 14 | 60 | 25 |
| 15 | 60 | 27 |
| 16 | 65 | 24 |
| 17 | 69 | 24 |

i) The two variables are supposed to have a linear relationship. Write the model you would fit to these data.

A regression analysis was performed with these data and the following outputs were obtained using R.

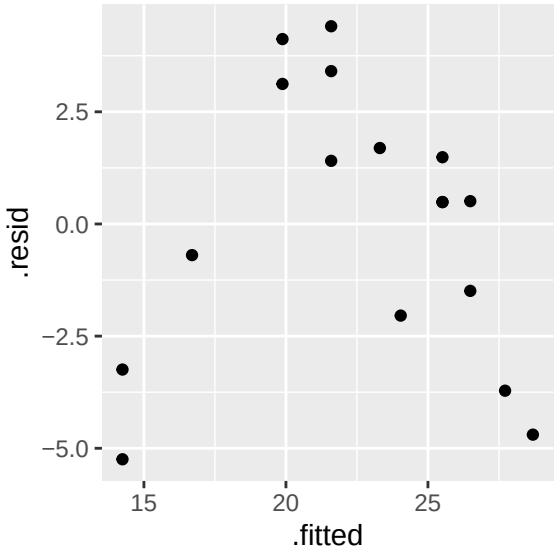
**Output a**



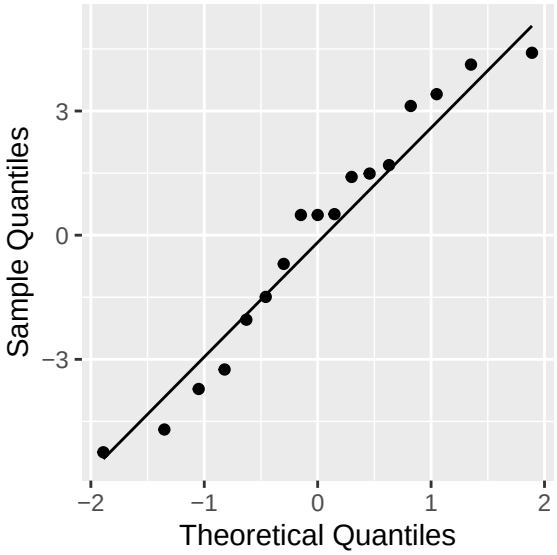

**Output b**

```
Call:
lm(formula = Weight ~ Temperature, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2450 -2.0422  0.4882  1.6926  4.4071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79572    2.03828   5.787 3.58e-05 ***
Temperature  0.24493    0.04318   5.672 4.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.123 on 15 degrees of freedom
Multiple R-squared:  0.682, Adjusted R-squared:  0.6608
F-statistic: 32.18 on 1 and 15 DF,  p-value: 4.429e-05
```

**Output c**




**Output d**




**Output e**

```
    Shapiro-Wilk normality test

data:  fitmodel$.resid
W = 0.95278, p-value = 0.502
```

ii) Two undergraduates studying statistics were looking at this analysis.

(A) One said that the results strongly suggest that this model is highly significant and can be used for prediction purposes.

(B) The other said that the results show the fitted model is not appropriate for this case and this model cannot be used for prediction.

With whom would you agree? Justify your argument using each part ((a) to (e)) of the results given.

## Question 2

In a soap production factory, there are two machines used for the production. Using 27 production runs; 15 of line 1 and 12 of line 2, the management wanted to find the relationship between the machine speed and the amount of scrap produced during the production process. To allow the two machines have different regression lines with different intercepts and slopes the following model was fitted with all 27 observations.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

where,

$X_1$ is line speed and

$$X_2 = \begin{cases} 1 & \text{if line 1} \\ 0 & \text{if line 2} \end{cases} \tag{1}$$

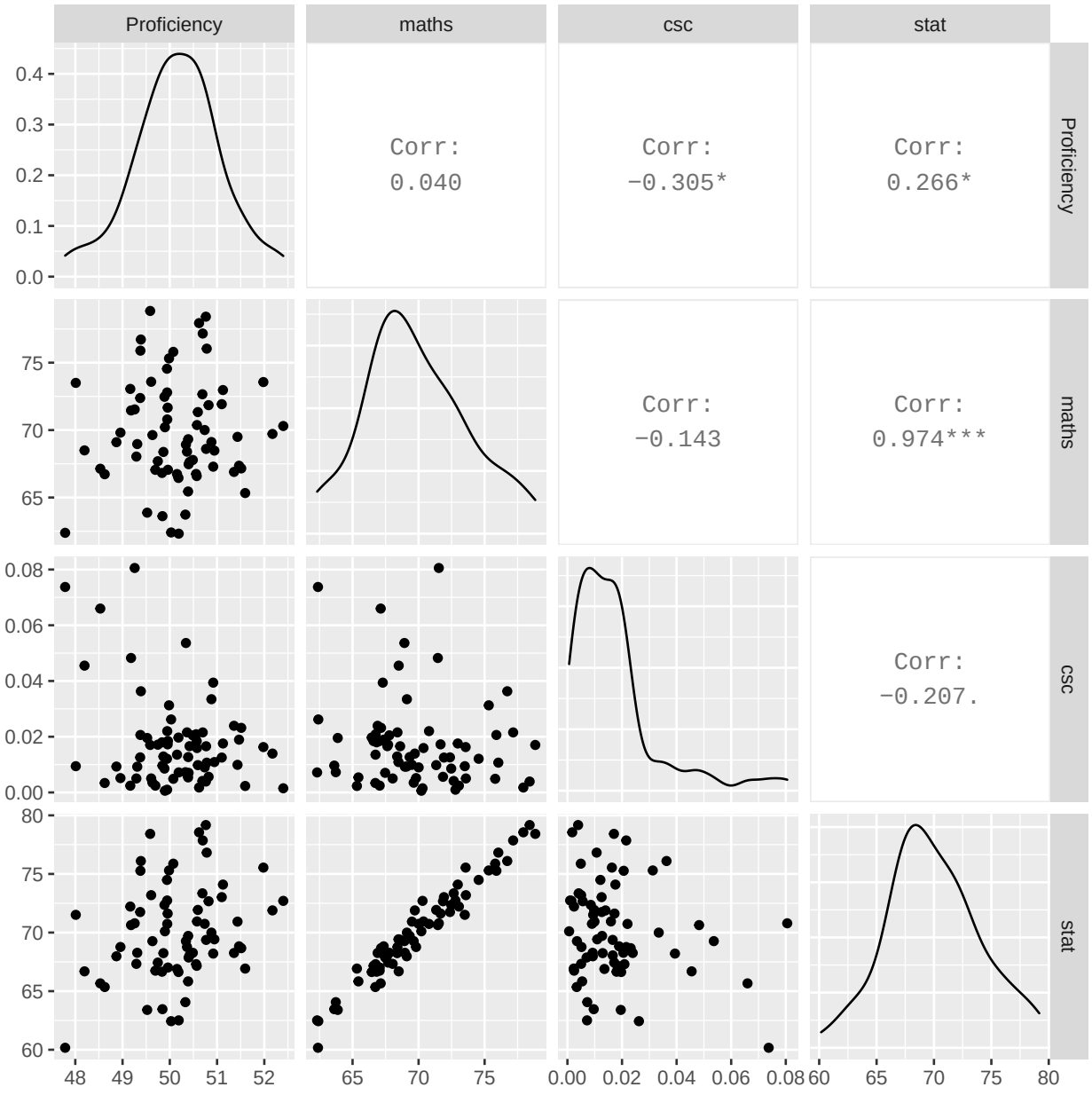
i) Draw a sketch of the scatter plot which is expected with the above model.

ii) Write the model for each machine.

iii) Write the hypotheses that should be tested to find whether the two machines have the same regression model or not, i.e. whether both the intercept and the slope are the same of the two models you wrote in ii) in the above.

## Question 3

A group of new graduates who have studied Statistics, Mathematics and Computer Science at the Faculty of Applied Sciences of University of Jayewardenepura joined a company. They were given three tests in the three subjects they have studied for the degree at the final interview at which they were selected for the job. After three months of a probationary period, their proficiency for the job was measured. The tests scores and the measure of proficiency were analyzed to find a model to predict proficiency by the test scores. Some results are shown below.




```
model.sjp <- lm(Proficiency ~ maths + csc + stat, data=df)
summary(model.sjp)
```

```
Call:
lm(formula = Proficiency ~ maths + csc + stat, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-1.136e-13  5.390e-16  2.112e-15  2.632e-15  9.808e-15

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
```

```
(Intercept)  5.000e+01  3.311e-14  1.510e+15   <2e-16 ***
maths       -1.000e+00  2.113e-15 -4.732e+14   <2e-16 ***
csc          1.647e-14  1.175e-13  1.400e-01    0.889
stat         1.000e+00  2.062e-15  4.849e+14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.51e-14 on 66 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 8.644e+28 on 3 and 66 DF,  p-value: < 2.2e-16
```

```r
car::vif(model.sjp)
```

```
    maths       csc      stat
20.786453  1.123955 21.276288
```

A statistician examined these results and claimed that "multicollinearity" has affected this model.

i) What is meant by multicollinearity?

ii) Do you agree with statistician claim. Justify your answer.

iii)

## Question 4

it is required to study the relationship between age ($X$) and girth ($Y$) of teak trees growing in a plantation. Note that girth is the diameter of the tree (in inches) measured at 5 inches above the ground. The girth of the trees and the ages (in years) have been recorded from a sample of 25 trees. Assume that the scatterplot of the data clearly shows a linear relationship between the two variables with an intercept.

i) Write the simple linear regression model that you would be fitted for the above variables. Define all terms in it and state any assumptions regarding the model.

ii) Later it was suggested that a linear model goes through the origin is suitable for this situation. Write the new model using the usual notation.

iii) The estimated regression model in part (ii) satisfied all of the assumptions regarding the error term. Sketch the residual plot vs fitted values and Q-Q normality plot of residuals.

## Question 5

An experiment was conducted to determine the influence of sulfide concentration ($X_1$) on the whiteness of rayon ($Y$). The results obtained through R are given below.

```r
x1 <- rnorm(15, mean=40)
y <- 13 + (2*x1) + rnorm(15)
df5 <- data.frame(x1=x1, Y=y)
mod5 <- lm(Y ~ x1, data=df5)
summary(mod5)
```

```
Call:
lm(formula = Y ~ x1, data = df5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69929 -0.48179  0.02163  0.66530  1.31226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.718     12.780   1.699    0.113
x1             1.786      0.321   5.563 9.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9087 on 13 degrees of freedom
Multiple R-squared:  0.7042,    Adjusted R-squared:  0.6814
F-statistic: 30.94 on 1 and 13 DF,  p-value: 9.185e-05
```

```
anova(mod5)
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 25.550 25.5497  30.944 9.185e-05 ***
Residuals 13 10.734  0.8257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i) Construct the ANOVA table using the above results.

ii) Wtite the hypothesis to be tested in the ANOVA in part i.

iii) What is your decision about the fitted model.