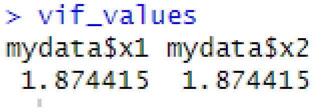
JACOB S. SANTOS February 6, 2024

BSCOE 3-1 Dr. Ariel Sison

Assignment

1. Determine the multicollinearity of the independent variables.



```
> vif_values
mydata$x1 mydata$x2
 1.874415  1.874415
```

$$VIF = \frac{1}{1 - R_i^2}$$

VIF Values

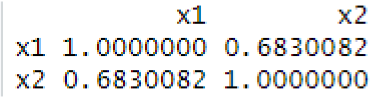
As the VIF values of the independent variables are both below 5, this indicates that there is no to little multicollinearity.

2. Determine the accepted threshold for multicollinearity problem and p-value. Cite references.

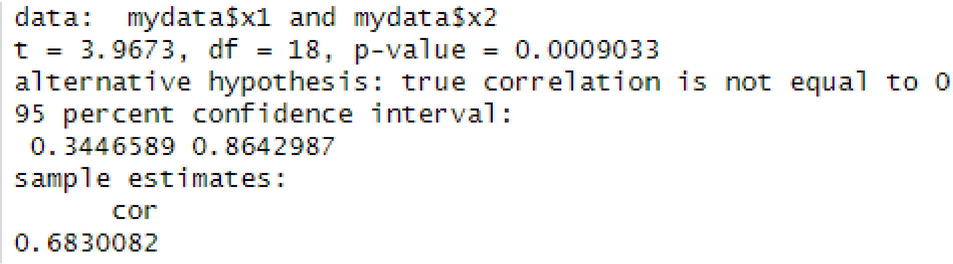
   Multicollinearity:
   The variance inflation factor is a measure to analyze the magnitude of multicollinearity of model terms. A VIF **less than 5** indicates a low correlation of that predictor with other predictors. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity.

   P-value:
   p-values help assess the statistical significance of independent variables in a regression model, while multicollinearity examines the correlation between independent variables. If there is low multicollinearity in the regression model, it generally indicates that the independent variables are not highly correlated with each other and has low p-value. In our case the common accepted threshold p-value is **0.05**

3. Determine the correlation of the independent variables.



```
          x1        x2
x1 1.0000000 0.6830082
x2 0.6830082 1.0000000
```



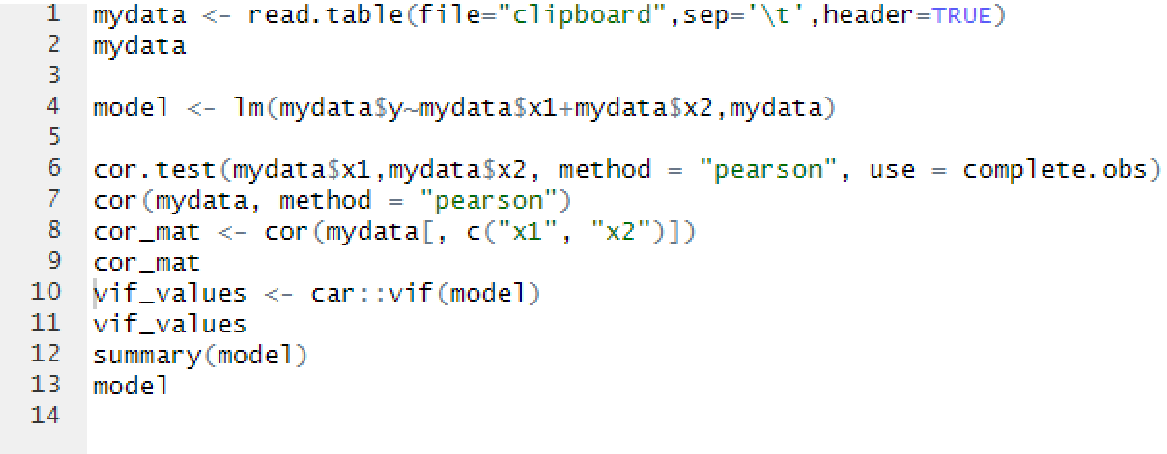
```
data:  mydata$x1 and mydata$x2
t = 3.9673, df = 18, p-value = 0.0009033
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3446589 0.8642987
sample estimates:
      cor
0.6830082
```

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Correlation Matrix (Pearson)

This is a moderate positive correlation, which means there is a tendency for high x1 variable scores go with high x2 variable scores (and vice versa). A correlation coefficient of 1 indicates a perfect positive linear relationship, while a coefficient of -1 indicates a perfect negative linear relationship. In this case, the coefficient is positive, but it's less than 1 **(0.6830082)** indicating a strong positive correlation, but not a perfect one.

Code:



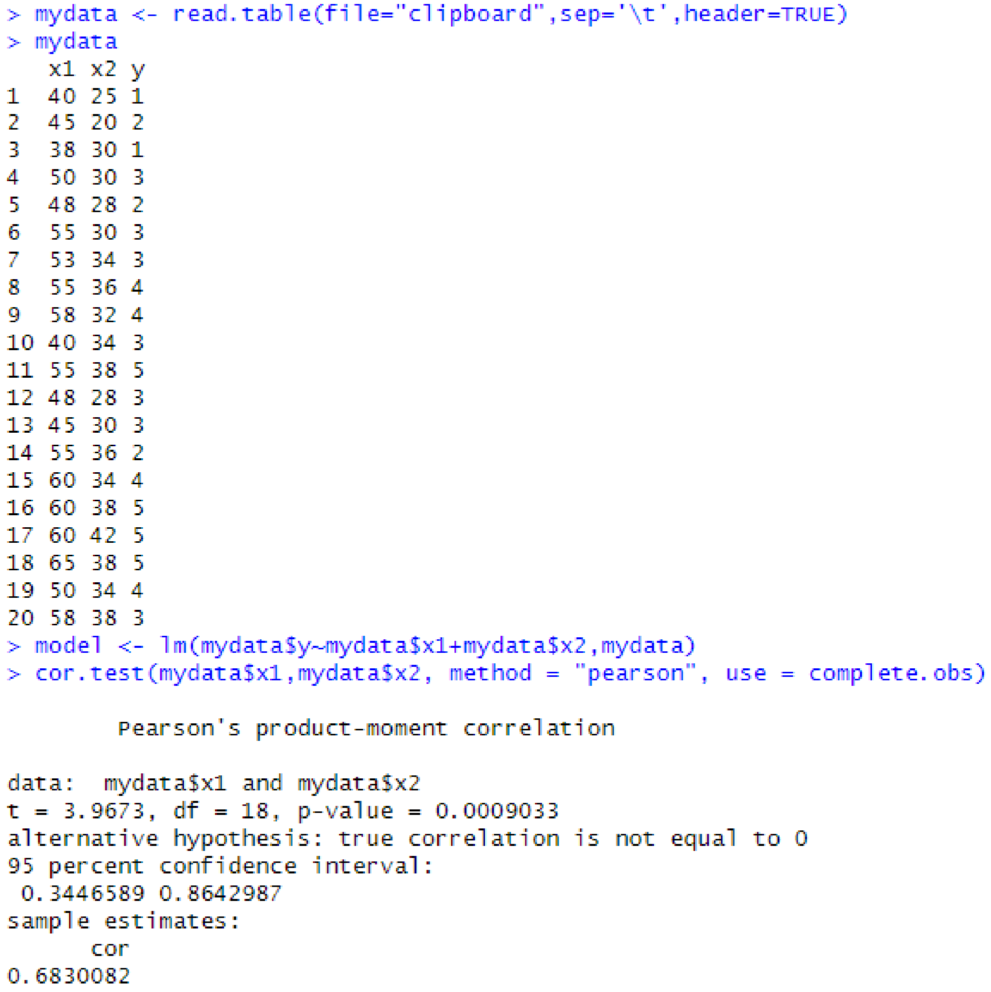
```
mydata <- read.table(file="clipboard",sep='\t',header=TRUE)
mydata

model <- lm(mydata$y~mydata$x1+mydata$x2,mydata)

cor.test(mydata$x1,mydata$x2, method = "pearson", use = complete.obs)
cor(mydata, method = "pearson")
cor_mat <- cor(mydata[, c("x1", "x2")])
cor_mat
vif_values <- car::vif(model)
vif_values
summary(model)
model

```

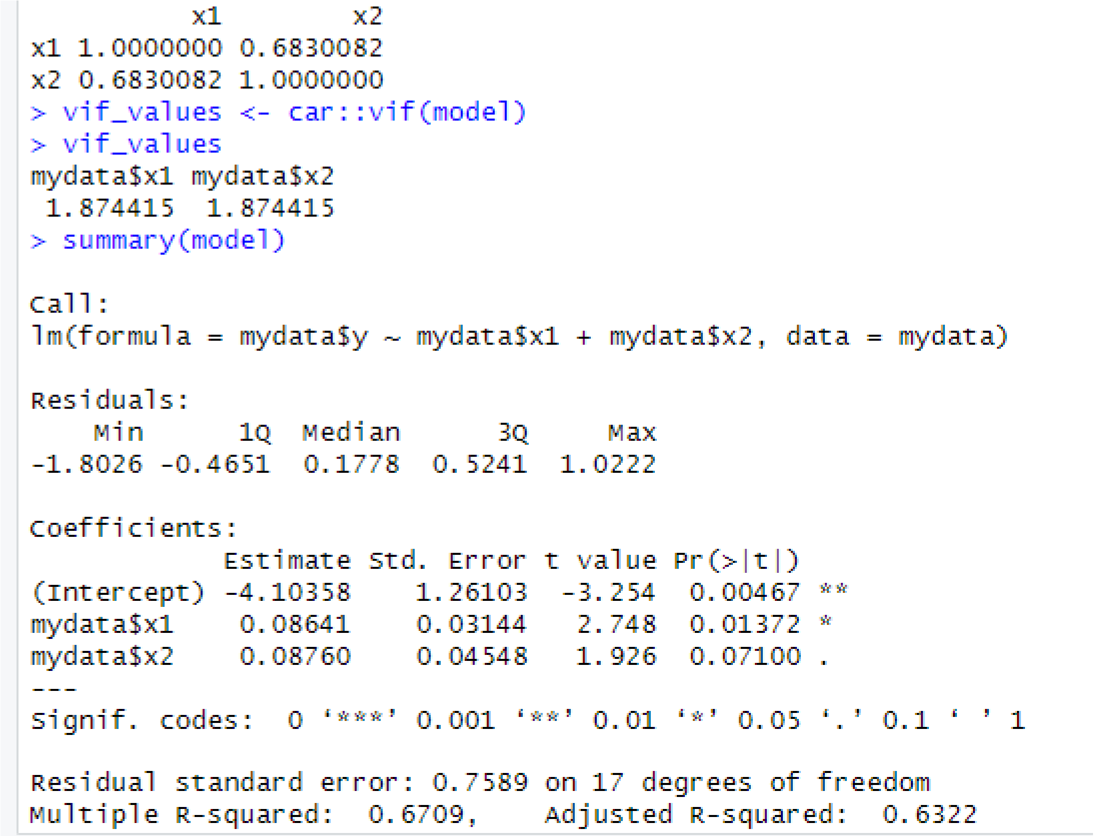
Output:



```
> mydata <- read.table(file="clipboard",sep='\t',header=TRUE)
> mydata
   x1 x2 y
1  40 25 1
2  45 20 2
3  38 30 1
4  50 30 3
5  48 28 2
6  55 30 3
7  53 34 3
8  55 36 4
9  58 32 4
10 40 34 3
11 55 38 5
12 48 28 3
13 45 30 3
14 55 36 2
15 60 34 4
16 60 38 5
17 60 42 5
18 65 38 5
19 50 34 4
20 58 38 3
> model <- lm(mydata$y~mydata$x1+mydata$x2,mydata)
> cor.test(mydata$x1,mydata$x2, method = "pearson", use = complete.obs)

        Pearson's product-moment correlation

data:  mydata$x1 and mydata$x2
t = 3.9673, df = 18, p-value = 0.0009033
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3446589 0.8642987
sample estimates:
      cor
0.6830082
```



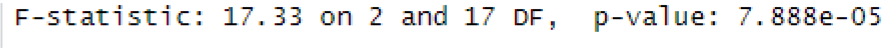
```
           x1        x2
x1 1.0000000 0.6830082
x2 0.6830082 1.0000000
> vif_values <- car::vif(model)
> vif_values
mydata$x1 mydata$x2 
 1.874415  1.874415 
> summary(model)

Call:
lm(formula = mydata$y ~ mydata$x1 + mydata$x2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8026 -0.4651  0.1778  0.5241  1.0222 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -4.10358    1.26103  -3.254  0.00467 **
mydata$x1    0.08641    0.03144   2.748  0.01372 * 
mydata$x2    0.08760    0.04548   1.926  0.07100 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7589 on 17 degrees of freedom
Multiple R-squared:  0.6709,	Adjusted R-squared:  0.6322 
```



```
F-statistic: 17.33 on 2 and 17 DF,  p-value: 7.888e-05
```

References:

Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Qual Quant. 2018;52(4):1957-1976. doi:10.1007/s11135-017-0584-6 https://www.researchgate.net/publication/321039845_Confounding_and_collinearity_in_regression_analysis_a_cautionary_tale_and_an_alternative_procedure_illustrated_by_studies_of_British_voting_behaviour

https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/#:~:text=The%20Variance%20Inflation%20Factor%20(VIF)%20measures%20the%20severity%20of%20multicollinearity,as%20a%20result%20of%20collinearity.

https://easystats.github.io/performance/reference/check_collinearity.html#:~:text=A%20VIF%20less%20than%205,2013).

https://medium.com/@shouryareddy306/variance-inflation-factor-vif-c0a392efbd4f

https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

https://www.scribbr.com/statistics/p-value/#:~:text=The%20most%20common%20threshold%20is,of%200.01%2C%20or%20even%200.001.