2 exercices Rstudio

Lynda.M

2022-05-29

library(tidyverse)  
library(openintro)  
library(car)

## import: the .csv file to the global environment (it’s relative to where you put the employee.csv, for example I put it in desktop (in the office)

employee <- read.csv("employee.csv", TRUE, ",")  
class(employee)

## [1] "data.frame"

View(employee)

## Exercise 1

### a

1. Check that the sample size is n = 71 observations.

nrow(employee)

## [1] 71

The number of rows above shows that sample size is 71.

### b

1. Noting Td.d.l. a random variable following a Student’s t law at d.d.l. degrees of freedom, calculate the probability P (T69 > 0) . Comment.

Answer: The formula for calculating probability is pt(q=T,df). Here we have df=69 as given above so we calculate the probability as

pt(0,df=69,lower.tail=F)

## [1] 0.5

The result shows that one sided p value is 0 which shows that there is 50% chance of getting distribution on positive side of the t-curve.

### c

1. Find the point q0.8 such that P (T69 < q0.8) = 0.8 (80% of the observations are below).

Answer:

pt(0.8,df=69,lower.tail = T)

## [1] 0.7867717

### d

1. Find the point t⋆69 such that P (|T69| ≥ t⋆69 ) = α where α = 5%. Answer:

1-pt(0.69,69,lower.tail = T)

## [1] 0.2462542

## Exercice 2

Consider experience as the explanatory variable and salary as the explained variable.

donc x= experience (explanatory variable), y= salary (explained variable).

### a

1. Calculate the mean of each of these variables as well as their respective standard deviation.

#### a.1

Answer: La moyenne de la variable expérience est : 5.746479 l’écart type de la variable expérience est : 3.241333

mean(employee$experience)

## [1] 5.746479

sd(employee$experience)

## [1] 3.241333

Mean value for experience is 5.74 Standard deviation for experience is 3.24.

#### a.2

mean(employee$salary)

## [1] 45141.51

sd(employee$salary)

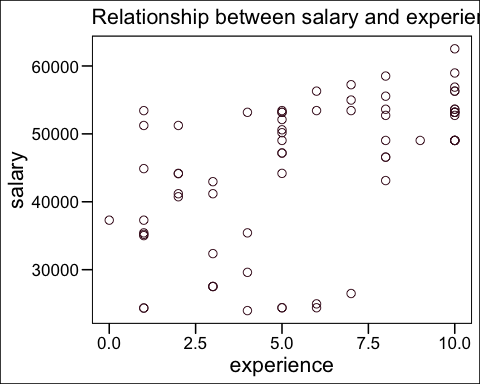
## [1] 10805.85

La moyenne de la variable salary est : 45141.51 l’écart type de la variable expérience est : 10805.85

### b

1. Use a scatter plot to represent the relationship between these two variables and describe the relationship between them.

library(ggplot2)  
  
ggplot(employee) +  
 aes(x = experience, y = salary) +  
 geom\_point(shape = "circle open", size = 2.6, colour = "#461124") +  
 labs(subtitle = "Relationship between salary and experience") +  
 ggthemes::theme\_base()



Generally with increase in experience salary also increases according to the plot. The relationship between two variable is not strictly linear. In some cases the salary is high even with less experience. The employee with 10 years of experience have highest salaries in the dataset.

### c

1. Calculate and interpret their correlation coefficient. Comment against the scatter plot.

cor(employee$experience,employee$salary,method=c("pearson", "kendall", "spearman"))

## [1] 0.5515126

Corelation coefficient value is 0.55 between the two variables. Generally the coefficient value > 0.5 means that two variables are weakly positive correlated with each other. It can be observed from the scatter plot as well where with increase of experience salary increases as well. The scatter plots shows that there are some cases where salary is not increasing with experience for employee and it can be an underlying reason for not getting a much value of correlation coefficient.

### d

1. Give the equation of the regression line that connects the two variables and plot it on the scatter plot.

a <- employee$experience  
b <- employee$salary  
model <- lm(a~b,data=employee)  
model

##   
## Call:  
## lm(formula = a ~ b, data = employee)  
##   
## Coefficients:  
## (Intercept) b   
## -1.7213787 0.0001654

Answer:

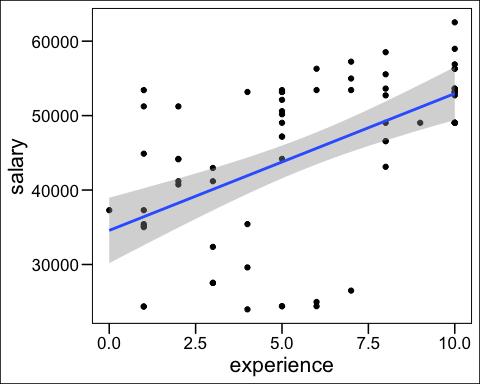
* the estimated regression line equation can be written as follows:

salary = -1.7213787 + 0.0001654\*experience

The intercept value is negative which means that the

ggplot(data = employee, aes(x = experience, y = salary)) + geom\_point() + stat\_smooth(method = "lm", se = TRUE) +  
 ggthemes::theme\_base()

## `geom\_smooth()` using formula 'y ~ x'



### e

1. Calculate the standard error Sb1 of the slope b1 of the regression line.

sqrt(deviance(model)/df.residual(model))

## [1] 2.723334

standard error of slope : 2.72 on 69 degrees of freedom

### f

1. Deduce the 95% confidence interval of the slope b1.

confint(model,'employee$salary',level=0.95)

## 2.5 % 97.5 %  
## employee$salary NA NA

### g

1. Test at the 5% threshold if the slope is significantly different from 0. Interpret the result.

For this purpose we need to define 2 hypothesis at 5% sigi=nficance level which can tested afterwards; **Null hypothesis H0**: Slope is significantly different than 0 **Alternate Hypothesis HA**: Slope is not significantly different than 0

The result of t-test below shows that p-value is less than 0.05 (significance level) so we reject our null hypothesis.

Since we rejected the null hypothesis, we have sufficient evidence to say that the true average increase in salary for experience is not zero.

t.test(employee$experience,employee$salary,data=employee)

##   
## Welch Two Sample t-test  
##   
## data: employee$experience and employee$salary  
## t = -35.196, df = 70, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -47693.46 -42578.06  
## sample estimates:  
## mean of x mean of y   
## 5.746479 45141.507042

### h

1. How much of the variability in wages can be explained by the fact that some employees have more experience than others?

summary(model)

##   
## Call:  
## lm(formula = a ~ b, data = employee)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.1165 -2.0753 0.0411 2.5985 4.3389   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.721e+00 1.398e+00 -1.232 0.222   
## b 1.654e-04 3.012e-05 5.492 6.21e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.723 on 69 degrees of freedom  
## Multiple R-squared: 0.3042, Adjusted R-squared: 0.2941   
## F-statistic: 30.16 on 1 and 69 DF, p-value: 6.206e-07

The summary of our regression model shows that R2 is 0.29 which means for 29% percent of time the salary can be predicted based on the experience value. Hence 29% of variability in one can be explained by the differences in the other variable.

### i

1. What annual salary could be expected from an employee with 8 years of experience? And an employee with 3 years of experience?

We can deduce that from our linear equation of the model.

#### For 8 years of experience

predict(model,data.frame(b=8),interval = 'confidence',level=0.95)\*10000

## fit lwr upr  
## 1 -17200.55 -45078.49 10677.38

so for an employee with 8 years of experience, salary of 17201 is expected.

#### for 3 years of experience

predict(model,data.frame(b=3),interval = 'confidence',level=0.95)\*10000

## fit lwr upr  
## 1 -17208.82 -45089.68 10672.03

so for an employee with 3 years of experience, salary of 17209 is expected.

### j

1. Examine the regression conditions based on the residuals.

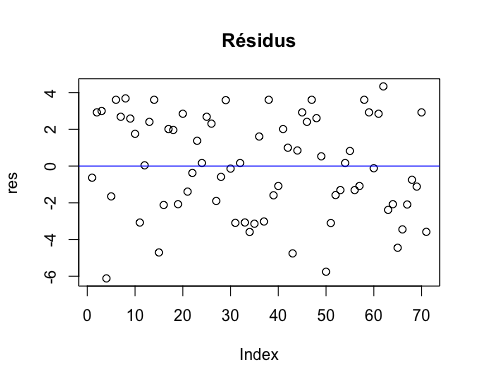
Extraction des résidus

head(resid(model))

## 1 2 3 4 5 6   
## -0.6333298 2.9246884 2.9994637 -6.1165043 -1.6498201 3.6097430

It is observed that the residuals seem to follow a trend. On time series data (traditionally ordered chronologically), this could indicate an auto-correlation of errors (contrary to the independence hypothesis), and therefore of condition not taken into account (ex: age, training, etc.) .

res<-resid(model)  
plot(res,main="Résidus") + abline(h=0,col="blue")



## integer(0)

#### k

1. Create employee\_f a base that includes all employees and employee\_m the rest of the sample. Calculate the slope of the regression in each subgroup and comment.

#subtracting 1st id column  
p <- employee[,-1]  
  
# creating a dataframe for female employee  
employee\_f <- p %>% filter(gender==0)  
  
#creating a dataframe for male employee  
employee\_m <- p %>% filter(gender==1)  
  
#fitting regression model on female employee  
model\_f <- lm(employee\_f$experience~employee\_f$salary,data=employee\_f)  
  
#fitting regression model on male employee  
model\_m <- lm(employee\_m$experience~employee\_m$salary,data=employee\_m)  
  
# slope of regression for female subgroup  
  
model\_f$coef[2]

## employee\_f$salary   
## 0.0001175642

# slope of regression for male subgroup  
model\_m$coef[2]

## employee\_m$salary   
## 0.0002276517

The slope of regression for female subgroup is 1.175642^{-4} and the slope of regression for male subgroup is 2.276517^{-4}.