Salary Prediction

Introduction to data Science

CCDS-211(19774)

Linear Regression

Joud Abdullah Aljehani

2111644

Predict the salary based on Years of Experience

Problem:

Company want to predict the Salary of the employee based on their years of experience ,I will use the xlsx files that have data of the salary and age , years of experience we'll use the linear regression to predict the salary.

Question:

What will happen if the years of experience increased?

how much does the salary depend in years of experience ?

Answer:

we can develop a linear regression model based on salary and years experience by using RStudio.

the salary is the dependent variable and the years experience will be the independent variable.

This is my dataset I will work on

```
# A tibble: 30 × 3
   YearsExperience   Age Salary
             <dbl> <dbl>  <dbl>
 1             1.1  21    39343
 2             1.3  21.5  46205
 3             1.5  21.7  37731
 4             2    22    43525
 5             2.2  22.2  39891
 6             2.9  23    56642
 7             3    23    60150
 8             3.2  23.3  54445
 9             3.2  23.3  64445
10             3.7  23.6  57189
# … with 20 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

this picture represent the summary of the linear regression relation

```
> summary(salarylm)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7958.0 -4088.5  -459.9  3372.6 11448.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25792.2     2273.1   11.35 5.51e-12 ***
x             9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared:  0.957,     Adjusted R-squared:  0.9554
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16

>
```

Now we know the value of average error in prediction (residuals) and the median ,maximum,minimum for our linear regression.

this picture represent the summary of the dataset

```
> summary(setdf)
 YearsExperience       Age             Salary
 Min.   : 1.100   Min.   :21.00   Min.   : 37731
 1st Qu.: 3.200   1st Qu.:23.30   1st Qu.: 56721
 Median : 4.700   Median :25.00   Median : 65237
 Mean   : 5.313   Mean   :27.22   Mean   : 76003
 3rd Qu.: 7.700   3rd Qu.:30.75   3rd Qu.:100545
 Max.   :10.500   Max.   :38.00   Max.   :122391
>
```

As we see here this summary show as the median ,maximum, minimum for each column in our dataset

The dataset contains the following columns:

years Experience : the time spend on work and gain experience (minimum=1 year, maximum=10 years, median=4 years and 7 month, mean=5)

Age :age of each employee (minimum=21, maximum=38, median=25 , mean=27)

Salary :amount of money does he get(minimum=37k , maximum=112k, median=65k , mean=76k)

We will focus on years Experience and Salary

Find the correlation between variables

```
> cor(y,x)
[1] 0.9782416
```

As we can see here that there is a high positive correlation between salary and years experience, then we can use the linear regression

Checking the null values

```
> sum(is.na(setdf))
[1] 0
> colSums(is.na(setdf))
YearsExperience             Age          Salary
              0               0               0
>
```

after we check that we don’t have any missing values to deal with it we can continue

with this interception we can predict the salary

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      25792         9450
```

Linear Regression Equation for prediction is: salary = α + β years experience

Salary = 25792 + 9450 $*$ years experience

we can see here when they don't have any experience how much the salary will become

```
> sp = data.frame(x = 0);
> salary = predict(salarylm,sp);
> print(salary);
       1
25792.2
>
```

we can see here when they have some experience how much the salary will become

```
       1
25792.2
> sp = data.frame(x = 5);
>  salary = predict(salarylm,sp);
>  print(salary);
       1
73042.01
>
```

these two predictions prove that it have a positive correlation

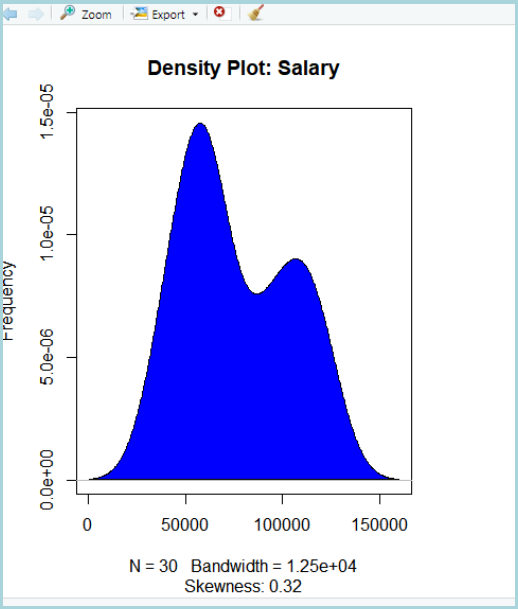
scatter plot

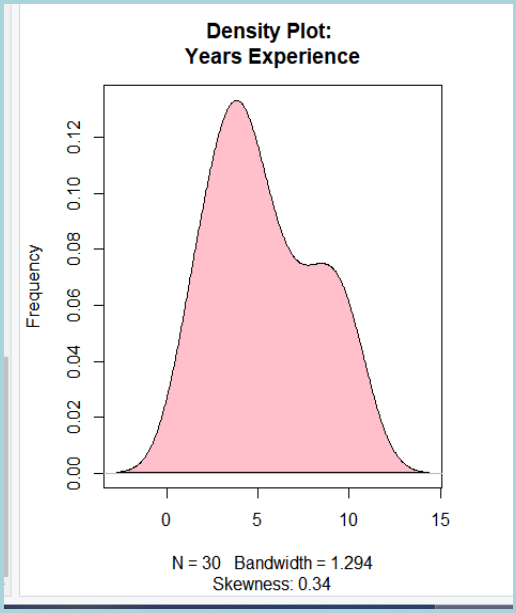
This is Visualize the linear relationship between the predictor (salary) and response (years experience),the values close to the line and the rest around the line.

Boxplot

As we see that the graph does not represent any outlier
The outlier can considerably affect the predictions as they can easily affect the direction

Density plot





Because the skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

LR model

As We can see the plot shows the strong positive correlation between the Salary and Years experience. whenever the Years experience is high the Salary will be also high .

The Cod

```
R 4.2.1 · C:/Users/jjood/3D Objects/MyR_Dir/
> setdf=read_excel("last.xlsx",sheet=1)
> setdf
# A tibble: 30 × 3
   YearsExperience   Age Salary
             <dbl> <dbl>  <dbl>
 1             1.1  21    39343
 2             1.3  21.5  46205
 3             1.5  21.7  37731
 4             2    22    43525
 5             2.2  22.2  39891
 6             2.9  23    56642
 7             3    23    60150
 8             3.2  23.3  54445
 9             3.2  23.3  64445
10             3.7  23.6  57189
# … with 20 more rows
# ℹ Use `print(n = ...)` to see more rows
> sum(is.na(setdf))
[1] 0
> colSums(is.na(setdf))
YearsExperience             Age          Salary
              0               0               0
> x=setdf$YearsExperience
> y=setdf$Salary
> salarylm = lm(y~x)
> print(salarylm)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      25792         9450

> summary(salarylm)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7958.0 -4088.5  -459.9  3372.6 11448.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25792.2     2273.1   11.35 5.51e-12 ***
x             9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
R 4.2.1 · C:/Users/jjood/3D Objects/MyR_Dir/
x            9450.0      378.8   24.95  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5788 on 28 degrees of freedom
Multiple R-squared:  0.957,     Adjusted R-squared:  0.9554
F-statistic: 622.5 on 1 and 28 DF,  p-value: < 2.2e-16

> cor(x,y)
[1] 0.9782416
> scatter.smooth(x=setdf$YearsExperience, setdf$Salary, main="salary ~ Years Experience")
> sp = data.frame(x = 0);
>  salary = predict(salarylm,sp);
>  print(salary);
       1
25792.2
> sp = data.frame(x = 5);
>  salary = predict(salarylm,sp);
>  print(salary);
       1
73042.01
> par(mfrow=c(1, 2))
>  boxplot(setdf$YearsExperience, main="Years Experience", sub=paste("Outlier rows: ", boxplot.stats(setdf$YearsExperience)
$out)) ; # box plot for 'speed'
>  boxplot(setdf$Salary, main="Salary", sub=paste("Outlier rows: ", boxplot.stats(setdf$Salary)$out))
> plot(setdf$Salary,setdf$YearsExperience,
+      +      col = "green",
Error: unexpected '=' in:
"plot(setdf$Salary,setdf$YearsExperience,
     +      col ="
> plot(setdf$YearsExperience, setdf$Salary,
+      col = "green",
+      main = "Speed & Distance Regression",
+      abline(lm(setdf$Salary ~ setdf$YearsExperience)),
+      cex = 1.0,
+      pch = 16 ,
+      xlab = "Salary",
+      ylab = "Year sExperience");
>
> install.packages("e1071",dep=TRUE)
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate
```

```
> library(e1071);
Warning message:
package 'e1071' was built under R version 4.2.2
> plot(density(setdf$YearsExperience), main="Density Plot:
+  Years Experience", ylab="Frequency", sub=paste("Skewness:",
+ round(e1071::skewness(setdf$YearsExperience), 2)));  polygon(density(setdf$YearsExperience), col="pink");
>
> plot(density(setdf$Salary), main="Density Plot: Salary",
+  ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(setdf$Salary), 2)));
>      polygon(density(setdf$Salary), col="blue");
>
```