Law Firm Absenteeism - Multiple Linear Regression

Faith Kabanda

Table of Contents

[Project Summary 1](#_Toc116474650)

[Import Dataset Using Readxl 1](#_Toc116474651)

[Understanding The Data Using Summary Statistics 4](#_Toc116474652)

[Understanding The Relationship Between Absence And Commitment 5](#_Toc116474653)

[Multiple Linear Regression 5](#_Toc116474654)

[Statistical Significance Of The Independent Variables 6](#_Toc116474655)

[Significance of the Regression Model at 0.05 Level of Significance 7](#_Toc116474656)

[Prediction And Forecasting 7](#_Toc116474657)

## Project Summary

A Director at a hypothetical small Zimbabwean law firm is concerned about employee absenteeism among staff. She believes that organisation commitment is probably the most significant contributing factor. She selected a random sample of 31 employee files and noted their level of absenteeism (in days p.a.). She sent a confidential questionnaire to each of the selected employees from which an organisation commitment index was derived. The Director is interested in the possible effect that other factors such as job tenure (time in months at the organisation), and grade (1 = clerk, 2 = consultant, 3 = lawyer) have on the level of employee absenteeism. Multiple linear regression of the data with commitment, job tenure and grade as independent variables is used to help the Director understand factors, if any, that are contributing to employees’ absenteeism. NB: For the variable grade, two dummy variables were used.

### Import Dataset Using Readxl

library(readxl)  
law\_firm\_data <- read\_excel("C:/Users/Faith Kabanda/OneDrive/Documents/Data Analysis Portfolio By Faith Kabanda/Data-Analyst-Portfolio-By-Faith-Kabanda/Project 4 - Multiple Regression Analysis in R/Law Firm Absenteeism Data - Multiple Linear Regression.xlsx",   
 sheet = "Data")  
head(law\_firm\_data) #to get the first few rows of the data

## # A tibble: 6 × 7  
## Absence Commitment Tenure `Employee Grade` Clerk Consultant Lawyer  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 23 49 143 2 0 1 0  
## 2 26 40 125 2 0 1 0  
## 3 18 69 95 2 0 1 0  
## 4 12 72 84 1 1 0 0  
## 5 30 26 164 3 0 0 1  
## 6 9 78 67 1 1 0 0

#Create a table  
  
library(kableExtra)

## Warning: package 'kableExtra' was built under R version 4.2.1

## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output  
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'

library(knitr)

## Warning: package 'knitr' was built under R version 4.2.1

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:kableExtra':  
##   
## group\_rows

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.2.1

law\_firm\_data %>% head() %>% kable() %>% column\_spec(1:6,border\_left = T, border\_right = T) %>% kable\_styling("striped")

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Absence | Commitment | Tenure | Employee Grade | Clerk | Consultant | Lawyer |
| 23 | 49 | 143 | 2 | 0 | 1 | 0 |
| 26 | 40 | 125 | 2 | 0 | 1 | 0 |
| 18 | 69 | 95 | 2 | 0 | 1 | 0 |
| 12 | 72 | 84 | 1 | 1 | 0 | 0 |
| 30 | 26 | 164 | 3 | 0 | 0 | 1 |
| 9 | 78 | 67 | 1 | 1 | 0 | 0 |

#This code gives 6 columns in total and kable\_styling gives proportional rows and columns giving rise equally size cells/small boxes. Lastly, removing %>% head() will give the full tabl  
# **Generate table [from kable() function] in MS Word: Highlight raw data > insert > insert table > convert text to table > select the right # of columns > enter**.

### Understanding The Data Using Summary Statistics

summary(law\_firm\_data$Absence)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 13.50 20.00 19.65 25.50 36.00

summary(law\_firm\_data$Commitment)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 40.00 49.00 51.74 65.00 82.00

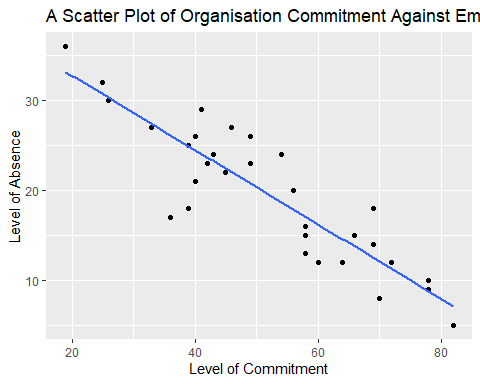
summary(law\_firm\_data$Tenure)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 24.0 65.5 120.0 107.6 144.5 189.0

### Understanding The Relationship Between Absence And Commitment

library(ggplot2)  
scatterplot<- ggplot(law\_firm\_data, aes(x = Commitment, y = Absence)) + geom\_point()+ geom\_smooth(method = lm, se = FALSE) + xlab("Level of Commitment") + ylab("Level of Absence") + ggtitle("A Scatter Plot of Organisation Commitment Against Employee Absenteeism") #geom\_smooth(method = lm) adds the regression line. Note that "lm" gives the straight line and "se" gives a shaded region around the line.  
scatterplot

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



According to the scatter plot above, as the level of absenteeism increases, the level of commitment decreases. This shows that absenteeism and commitment have a strong negative linear relationship with one another, as commitment increases, the absenteeism reduces. In addition, the data points come to forming somewhat of a downwards diagonal line along the negative gradient line when plotted, proving that there is a correlation between the two variables. There are no visible outliers.

### Multiple Linear Regression

#generate the model  
model<- lm(law\_firm\_data$Absence ~ law\_firm\_data$Commitment + law\_firm\_data$Tenure + law\_firm\_data$Clerk + law\_firm\_data$Consultant, data = law\_firm\_data)  
summary(model)

##   
## Call:  
## lm(formula = law\_firm\_data$Absence ~ law\_firm\_data$Commitment +   
## law\_firm\_data$Tenure + law\_firm\_data$Clerk + law\_firm\_data$Consultant,   
## data = law\_firm\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4191 -1.9394 0.0302 1.2563 6.0740   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.09777 4.41690 7.041 1.78e-07 \*\*\*  
## law\_firm\_data$Commitment -0.33890 0.07005 -4.838 5.15e-05 \*\*\*  
## law\_firm\_data$Tenure 0.04615 0.01818 2.538 0.0175 \*   
## law\_firm\_data$Clerk 1.21395 2.42045 0.502 0.6202   
## law\_firm\_data$Consultant 1.86753 1.67983 1.112 0.2764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.207 on 26 degrees of freedom  
## Multiple R-squared: 0.8496, Adjusted R-squared: 0.8264   
## F-statistic: 36.71 on 4 and 26 DF, p-value: 2.434e-10

### Statistical Significance Of The Independent Variables

H0: βi = 0 & H1: βi ≠ 0  
Rejection criteria is as follows: Reject H0 if |Tcal| > Tα/2(n-k)

Commitment

* Let i= 1 = commitment
* Tα/2(31– 4 =27, 0.025) = 2.05
* Tcal = -4.838

In this case, with commitment, the |Tcal| = -4.838 > Tα/2(27, 0.025) = 2.05, therefore we reject H0 and conclude that the coefficient β1 is not significantly different from zero at 5%. The implication is that commitment has no influence on absenteeism.

Tenure

* Let i= 2 = tenure
* Tα/2(27) = 2.05
* Tcal = 2.538

In this case, with job tenure, the |Tcal| = 2.538 > Tα/2(27, 0.025) = 2.05, therefore we reject H0 and conclude that the coefficient β2, Tenure is not significantly different from zero at 5%. The implication is that job tenure has no influence on absenteeism.

Clerk Level

* Let i= 3 = Clerk
* Tα/2(27) = 2.05
* Tcal = 0.502

In this case, with lecturer level, the |Tcal| = 0.501537884 < Tα/2(27, 0.025) = 2.05, therefore we fail to reject H0 and conclude that the coefficient β3 is significantly different from zero at 5%. The implication is that clerk level of employment has significant influence on absenteeism.

Consultant Level

* Let i= 4 = Consulatant
* Tα/2(27) = 2.05
* Tcal = 1.112

In this case, with senior lecturer level, the |Tcal| = 0. 1.112 < Tα/2(27, 0.025) = 2.05, therefore we fail to reject H0 and conclude that the coefficient β4 is significantly different from zero at 5%. The implication is that the consultant level of employment has significant influence on absenteeism.

### Significance of the Regression Model at 0.05 Level of Significance

* Y = Β0 + Β1X1 B2X2 + B3X3
* H0: β1 = β2 = β3 = … βk= 0
* H1: Not all slope coefficients are simultaneously zero.

For a given level of significance α = 0.05, numerator degrees of freedom k −1 and denominator degrees of freedom n − k, the critical value is given by F (k −1, n − k)α. Therefore, we will use F (4,26) α = 0.05 = 2.7426. Fcal from the ANOVA table is 36.70775776

Since Fcal = 36.70775776 > F (4,26)0.05 = 2.7426, we reject H0 and conclude that not all slope coefficients are simultaneously zero. Hence there is a significant relationship between the variables in the linear regression model of the data set.

### Prediction And Forecasting

Predicting the level of absence of an employee who has a tenure of 10 years and an organisational commitment index of 50 for all 3 levels of commitment are as follows:

*Absence = 1.214Clerk + 1.868Consulatant + 0.04615Tenure -0.3389 Commitment + 31.098*

1. Absence for Clerk Absence = 1.214(1) + 1.868(0) + 0.04615(10 x12) + -0.3389 (50) + 31.098 Absence = 20.90517271
2. Absence for Consultant Absence = 1.214(0) + 1.868(1) + 0.04615(10 x12) + -0.3389 (50) + 31.098 Absence = 21.55875238
3. Absence for Lawyer Absence = 1.214(0) + 1.868(0) + 0.04615(10 x12) + -0.3389 (50) + 31.098 Absence = 19.69122602