A look into the Age and Expenditures during the Holidays

Sanjaya J Shetty

28/05/2021

Answers

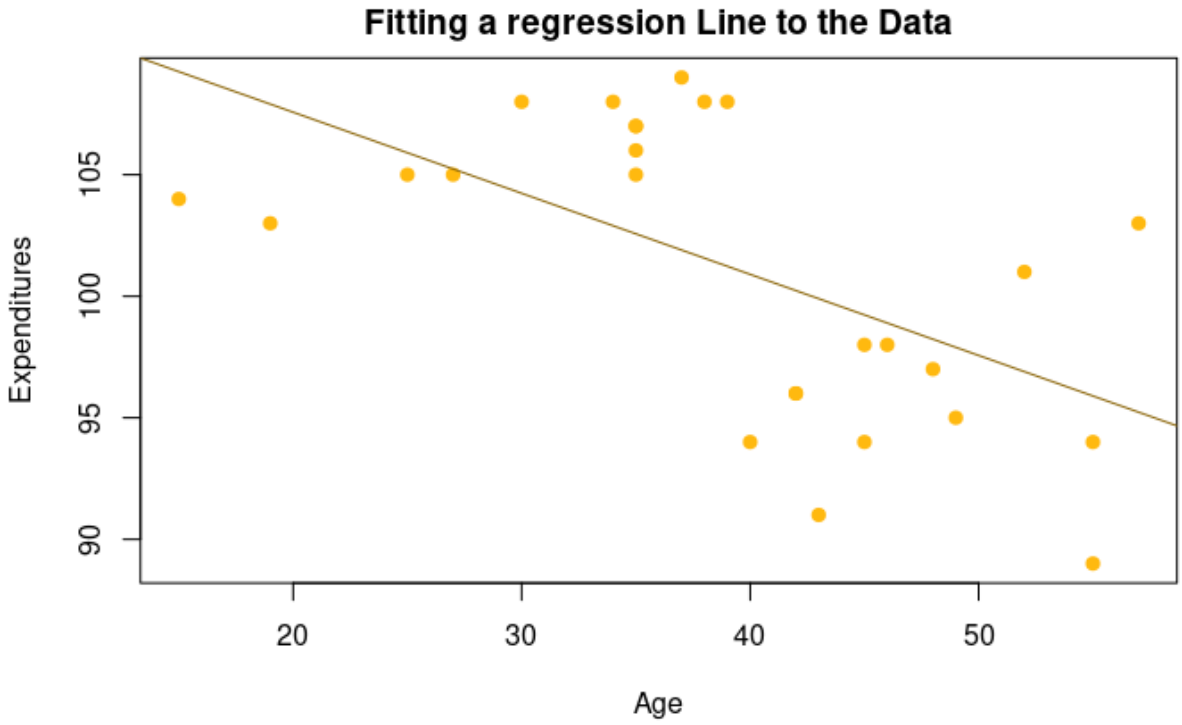
a. Use all data to estimate the coefficients a and b in a simple regression model, where expenditures is the dependent variable and age is the explanatory factor. Also compute the standard error and the t-value of b.

Answer:

Table 1	Estimate Std.	Error	t-value
(Intercept)	114.24111	3.88208	29.428
Age	-0.33360	0.09537	-3.498

b. Make the scatter diagram of expenditures against age and add the regression line y = a + bx of part (a) in this diagram. What conclusion do you draw from this diagram?

Answer:



c. It seems there are two sets of observations in the scatter diagram, one for clients aged 40 or higher and another for clients aged below 40. Divide the sample into these two clusters, and for each cluster estimate the coefficients a and b and determine the standard error and t-value of b.

Answer:

For People Below 40:

Table 2	Estimate Std.	Error	t-value
(Intercept)	100.23228	1.41590	70.79
Age	0.19797	0.04438	4.46

For People Above 40:

Table 2	Estimate Std.	Error	t-value
(Intercept)	88.8719	9.4585	9.396
Age	0.1465	0.1974	0.742

d. Discuss and explain the main differences between the outcomes in parts (a) and (c). Describe in words what you have learned from these results.

Answer: The main difference between the outcome in part (a) and (c) is that the slope of the diagram. The plot seen in the part(a) (*i.e plot 4*) would show an inaccurate estimation of expenditure of people near the age 40. Whereas the plot seen the part (c) (*i.e plot 5 and plot 6*) would be an more accurate fit.

Method:

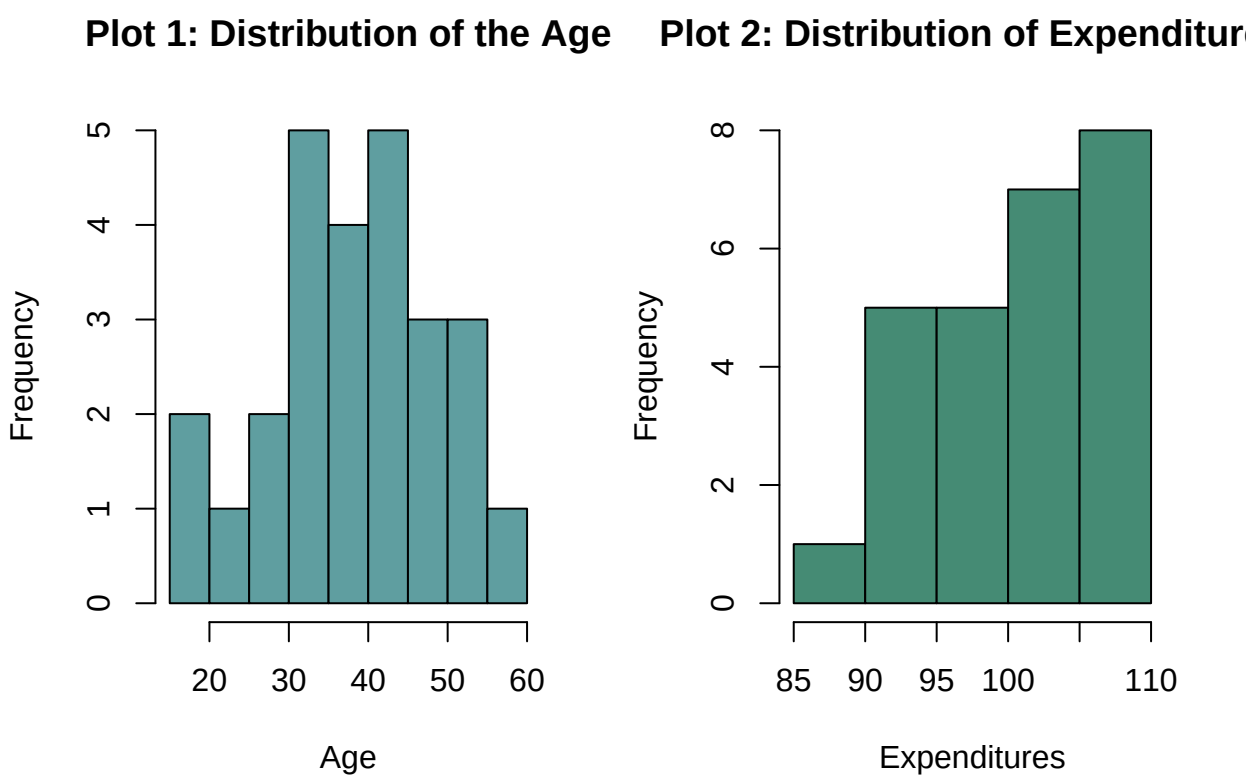
Load the library

```
library(plyr)
library(readxl)
```

Load the Dataset

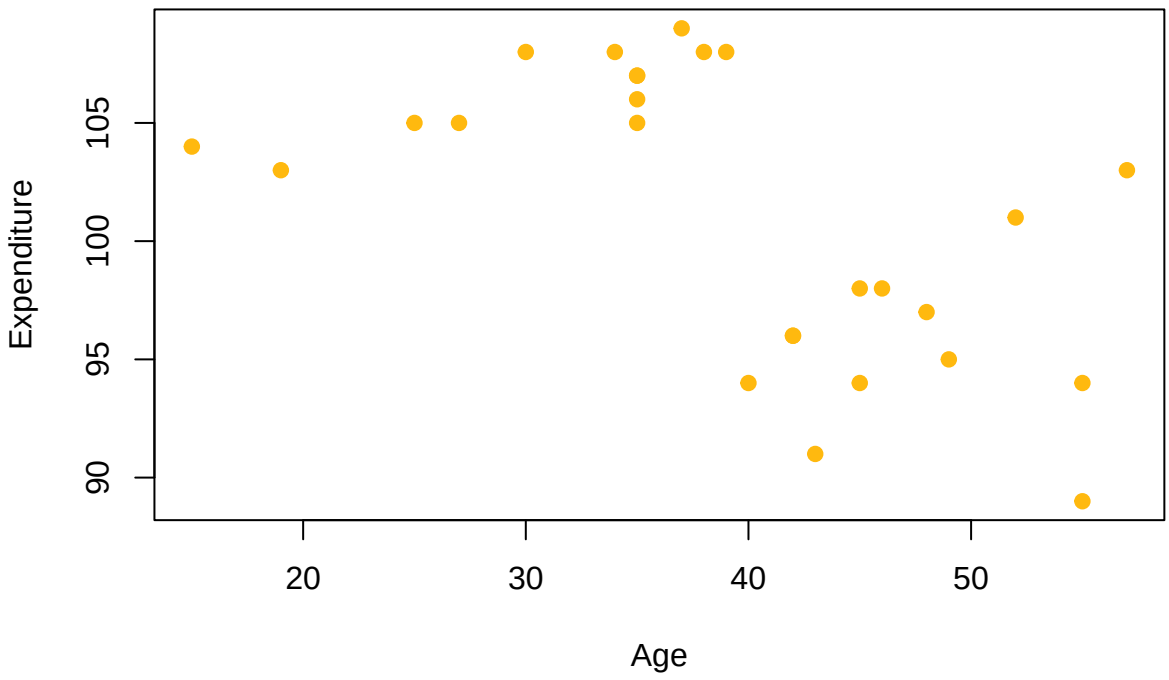
```
data <- read_xls(paste0(getwd(),"/data/TestExer1-holiday expenditures-round2.xls"))
```

```
par(mfrow = c(1,2))
hist(data$Age, main = 'Plot 1: Distribution of the Age', col = 'cadetblue',
     xlab = 'Age')
hist(data$Expenditures, main = 'Plot 2: Distribution of Expenditures', col = 'aquamarine4',
     xlab = 'Expenditures')
```




```
plot(x = data$Age, y = data$Expenditures, main = 'Plot 3: Expenditure of people across different age gr
     xlab = 'Age', ylab = 'Expenditure', col = 'darkgoldenrod1', pch = 19)
```

Plot 3: Expenditure of people across different age group



Since the above plot shows the clear clustering of data around two corner, hence drawing a single regression line for prediction would be unwise(see plot 4 below). So, In order to produce the meaningful regression line, we could cut off the dataset at Age 40 and would fit a regression model separately.

```
model <- lm(Expenditures~Age, data)

summary(model)
```

```
##
## Call:
## lm(formula = Expenditures ~ Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8965 -4.2301 -0.8984  4.3525  7.7739
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.24111    3.88208  29.428  < 2e-16 ***
## Age          -0.33360    0.09537  -3.498  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.073 on 24 degrees of freedom
## Multiple R-squared:  0.3377, Adjusted R-squared:  0.3101
## F-statistic: 12.24 on 1 and 24 DF,  p-value: 0.001852
```

Creating two separate the dataset.

```
dataA40 <- data[data$Age >= 40,] # Above 40

dataB40 <- data[data$Age < 40,] # Below 40
```

Fitting linear regression to the dataset containging people above 40 years old.

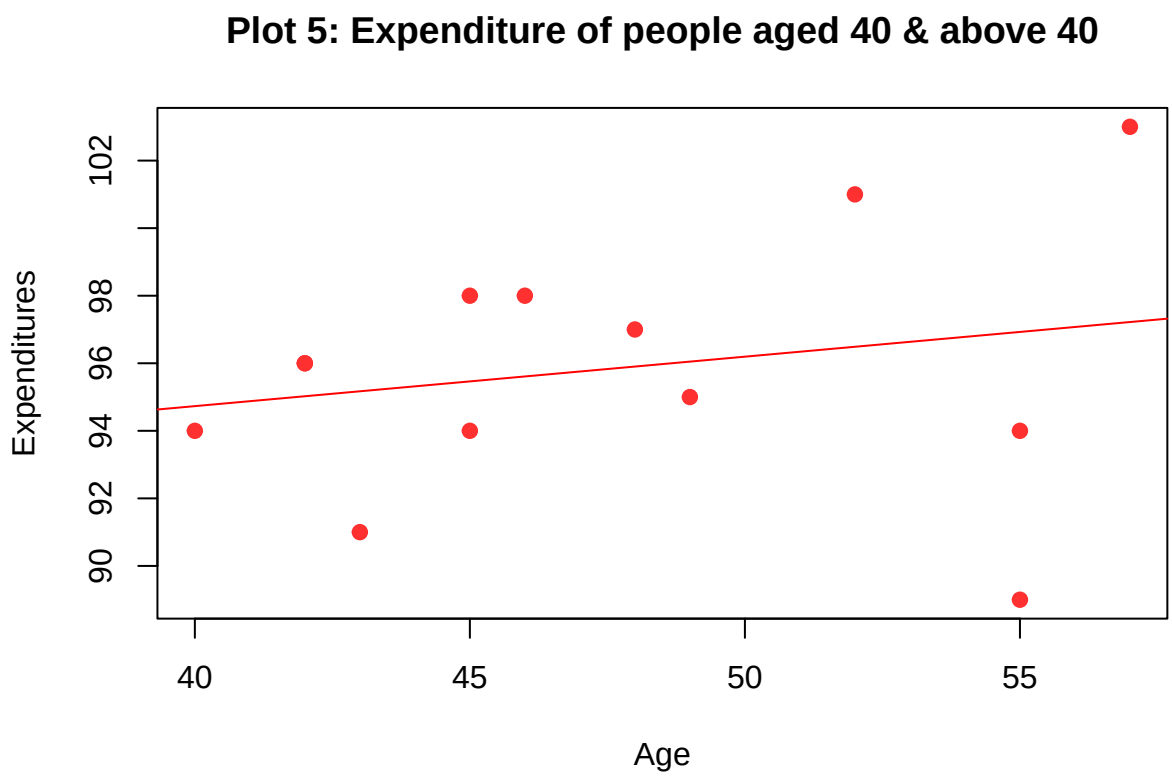
```
modelA40 <- lm(Expenditures~Age, dataA40)

summary(modelA40)
```

```
##
## Call:
## lm(formula = Expenditures ~ Age, data = dataA40)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9278 -1.4631  0.9763  2.3905  5.7793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  88.8719     9.4585   9.396 1.37e-06 ***
## Age           0.1465     0.1974   0.742    0.474
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 11 degrees of freedom
## Multiple R-squared:  0.04767,    Adjusted R-squared:  -0.0389
## F-statistic: 0.5507 on 1 and 11 DF,  p-value: 0.4736
```

```
plot(dataA40$Age, dataA40$Expenditures, xlab = 'Age', ylab = 'Expenditures',
     main = 'Plot 5: Expenditure of people aged 40 & above 40', pch = 19, col = 'firebrick1')
abline(modelA40, col = 'red')
```



Fitting linear regression to dataset containing people below 40.

```
modelB40 <- lm(Expenditures~Age, dataB40)

summary(modelB40)
```

```
##
## Call:
## lm(formula = Expenditures ~ Age, data = dataB40)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1613 -0.5775 -0.1613  0.7982  1.8286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.23228    1.41590   70.79 5.55e-16 ***
```

```
## Age           0.19797    0.04438    4.46 0.000962 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.153 on 11 degrees of freedom
## Multiple R-squared:  0.644,  Adjusted R-squared:  0.6116
## F-statistic:  19.9 on 1 and 11 DF,  p-value: 0.0009619
```

```
plot(dataB40$Age, dataB40$Expenditures, xlab = 'Age', ylab = 'Expenditure',
     main = 'Plot 6: Expenditure of people aged below 40', pch = 19, col = 'darkorange')
abline(modelB40, col = 'orange')
```