Exercise 1.5

In [1]:
```
data <- read.csv("./data/TestExer1.csv")
names(data) <- c('Observation', 'Age', 'Expenditure')
```

In [2]:
```
summary(data)
```

```
 Observation         Age         Expenditure
Min.   : 1.00   Min.   :15.00   Min.   : 89.0
1st Qu.: 7.25   1st Qu.:35.00   1st Qu.: 96.0
Median :13.50   Median :39.50   Median :103.0
Mean   :13.50   Mean   :39.35   Mean   :101.1
3rd Qu.:19.75   3rd Qu.:45.75   3rd Qu.:106.8
Max.   :26.00   Max.   :57.00   Max.   :109.0
```

(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

In [3]:
```
fit <- lm(Expenditure ~ Age,data=data)

sprintf("Answer:")
sprintf("a = %.2f", coef(fit)[1])
sprintf("b = %.2f", coef(fit)[2])
sprintf("standard-error = %.2f", summary(fit)$sigma)
sprintf("t-value of b = %.5f", coef(summary(fit))[2, "t value"])
```

'Answer:'

'a = 114.24'

'b = -0.33'

'standard-error = 5.07'

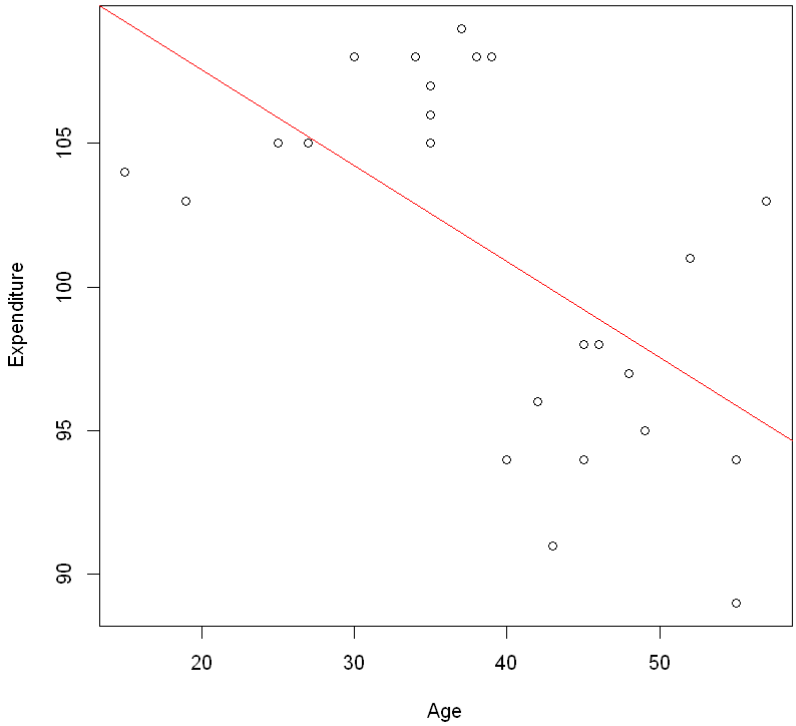
't-value of b = -3.49794'

(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:

The regression line does not fit the data well. The deviation of sample to the regression line is relatively large. As we have observed in Exercise 1.1, the data possesses different trends for younger and older age group, which the linear regression line unable to capture.

In [5]:
```
plot(data$Age, data$Expenditure, xlab="Age", ylab="Expenditure")
abline)fit, col="red)"
```



(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

In [6]:
```
below40 <- data[data$Age < 40, ]
fit1 <- lm(Expenditure ~ Age,data=below40)

sprintf("Regression values (for clients below age 40):")
sprintf("a = %.2f", coef(fit1)[1])
sprintf("b = %.2f", coef(fit1)[2])
sprintf("standard-error = %.2f", summary(fit1)$sigma)
sprintf("t-value of b = %.5f", coef(summary(fit1))[2, "t value"])
```

'Regression values (for clients below age 40):'

'a = 100.23'

'b = 0.20'

'standard-error = 1.15'

't-value of b = 4.46045'

In [7]:
```
above40 <- data[data$Age >= 40, ]
fit2 <- lm(Expenditure ~ Age,data=above40)

sprintf("Regression values (for clients above age 40):")
sprintf("a = %.2f", coef(fit2)[1])
sprintf("b = %.2f", coef(fit2)[2])
sprintf("standard-error = %.2f", summary(fit2)$sigma)
sprintf("t-value of b = %.5f", coef(summary(fit2))[2, "t value"])
```

'Regression values (for clients above age 40):'

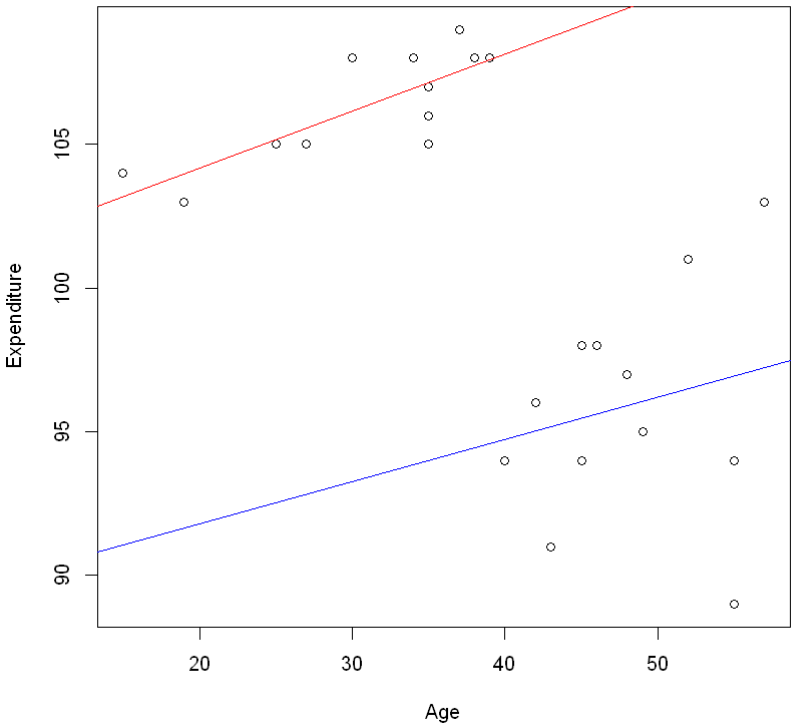
'a = 88.87'

'b = 0.15'

'standard-error = 3.83'

't-value of b = 0.74206'

In [8]:
```
plot(data$Age, data$Expenditure, xlab="Age", ylab="Expenditure")
abline(fit1, col="red")
abline(fit2, col="blue")
```



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.

In part (a) we fit the all the observations into a single linear regression line. In part (c), we split the observations into two groups and fit a linear regression to each group separately. The result in (c) provides a much better fit for each group.

Both the groups show that their willingness to spend increases with age. However, in the younger age group, their expenditure has significantly higher baseline than in older age group. This characteristic might be borne out of generational differences. For example, millenials values travel experiences highly and therefore, are willing to spend a majority of their disposable income to travel. On the other hand, baby boomers, who put more attention to financial stability, are usually more conservative in spending their money to travel. As such, the trends in the data seem to be reasonable and correspond to a possible phenomemon in reality.