Estatística Computacional - Lista 2

Rodrigo de Castro Ângelo

Exercício 1

Como as amostras são dependentes e estamos supondo normalidade dos dados, fazemos o teste t para duas amostras pareadas.

- H_0: $\mu_{atual} = \mu_{nova}$
- H_1: $\mu_{atual} > \mu_{nova}$

```r
atual <- c(24, 25, 27, 22, 23, 28, 26, 28, 29)
nova  <- c(21, 23, 28, 27, 24, 26, 25, 22, 23)
t.test(atual, nova, alternative = "greater", paired = T)
```

```
##
##  Paired t-test
##
## data:  atual and nova
## t = 1.2367, df = 8, p-value = 0.1256
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.7274866        Inf
## sample estimates:
## mean of the differences
##                1.444444
```

Como o valor-p $= 0.1256 > \alpha$, não rejeita-se a hipótese nula.

Conclusão:

Utilizando o teste t pareado, não rejeita-se a hipótese nula ao nível de 5% de significância, portanto devemos considerar que não houve diminuição no tempo médio para realização da tarefa.

Exercício 2

```r
tecnica1 <- c(1, 4, 4, 5, 6, 6)
tecnica2 <- c(2, 6, 6, 7, 7, 8)
```

Item a.

Como estamos supondo normalidade, usamos o teste F de Fisher.

- $H_0 : \frac{\sigma_{tecnica1}}{\sigma_{tecnica2}} = 1$
- $H_1 : \frac{\sigma_{tecnica1}}{\sigma_{tecnica2}} \neq 1$

```r
var.test(tecnica1, tecnica2)
```

```
##
##  F test to compare two variances
##
## data:  tecnica1 and tecnica2
## F = 0.78788, num df = 5, denom df = 5, p-value = 0.8
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1102486 5.6304827
## sample estimates:
## ratio of variances
##          0.7878788
```

Como o valor-p $> \alpha$, não rejeita-se a hipótese nula. Então devemos considerar que a razão entre as variâncias populacionais é 1, ou seja, as variâncias podem ser consideradas iguais.

Item b.

Como supomos normalidade e as variâncias populacionais são consideradas iguais, utilizamos o teste t para amostras independentes com mesma variância. Desejamos verificar se a Técnica 1 é mais eficiente, ou seja, se o tempo de recuperação é menor.

- H_0: $\mu_{tecnica1} = \mu_{tecnica2}$
- H_1: $\mu_{tecnica1} < \mu_{tecnica2}$

```r
t.test(tecnica1, tecnica2, var.equal = T, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  tecnica1 and tecnica2
## t = -1.4556, df = 10, p-value = 0.08809
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4086695
## sample estimates:
## mean of x mean of y
##  4.333333  6.000000
```

Como o valor-p $> \alpha$, não rejeita-se a hipótese nula. Portanto devemos considerar que a Técnica 1 não é mais eficiente que a Técnica 2.

Exercício 3

```
dinheiro <- c(56.00, 20.50, 37.37,  28.64)
cheque   <- c(80.90, 51.29, 40.95,  72.65, 132.47, 60.32, 60.00)
cartao   <- c(73.25, 56.65, 123.21, 56.50, 37.29,  44.65, 40.64)
```

Item a.

Análise descritiva das variáveis dinheiro, cheque e cartao:

```
mean(dinheiro)
```

```
## [1] 35.6275
```

```
sd(dinheiro)
```

```
## [1] 15.22872
```

```
summary(dinheiro)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.50   26.61   33.01   35.63   42.03   56.00
```

```
mean(cheque)
```

```
## [1] 71.22571
```

```
sd(cheque)
```

```
## [1] 30.01523
```

```
summary(cheque)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.95   55.65   60.32   71.23   76.78  132.47
```

```
mean(cartao)
```

```
## [1] 61.74143
```

```
sd(cartao)
```

```
## [1] 29.71949
```

```
summary(cartao)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37.29   42.65   56.50   61.74   64.95  123.21
```

Com base nos dados da análise descritiva, podemos observar que tanto a média quanto a mediana amostrais das compras pagas em dinheiro são bem menores que das outras formas de pagamento, sugerindo que na população isso também ocorra.

Item b.

Como temos 3 amostras , realiza-se a ANOVA. Tratando os dados:

```
forma.pagamento <- c(rep("Dinheiro", length(dinheiro)),
                     rep("Cheque", length(cheque)), rep("Cartão", length(cartao)))

compras <- data.frame(Valor = c(dinheiro, cheque, cartao), Forma_Pagamento = forma.pagamento)
```

```r
compras
```

```
##     Valor Forma_Pagamento
## 1   56.00        Dinheiro
## 2   20.50        Dinheiro
## 3   37.37        Dinheiro
## 4   28.64        Dinheiro
## 5   80.90          Cheque
## 6   51.29          Cheque
## 7   40.95          Cheque
## 8   72.65          Cheque
## 9  132.47          Cheque
## 10  60.32          Cheque
## 11  60.00          Cheque
## 12  73.25          Cartão
## 13  56.65          Cartão
## 14 123.21          Cartão
## 15  56.50          Cartão
## 16  37.29          Cartão
## 17  44.65          Cartão
## 18  40.64          Cartão
```

Teste de homocedasticidade dos dados. Será utilizado o teste de Bartlett.

- H_0: As variâncias são homogêneas
- H_1: As variâncias não são homogêneas

```r
bartlett.test(Valor ~ Forma_Pagamento, data = compras)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Valor by Forma_Pagamento
## Bartlett's K-squared = 1.4905, df = 2, p-value = 0.4746
```

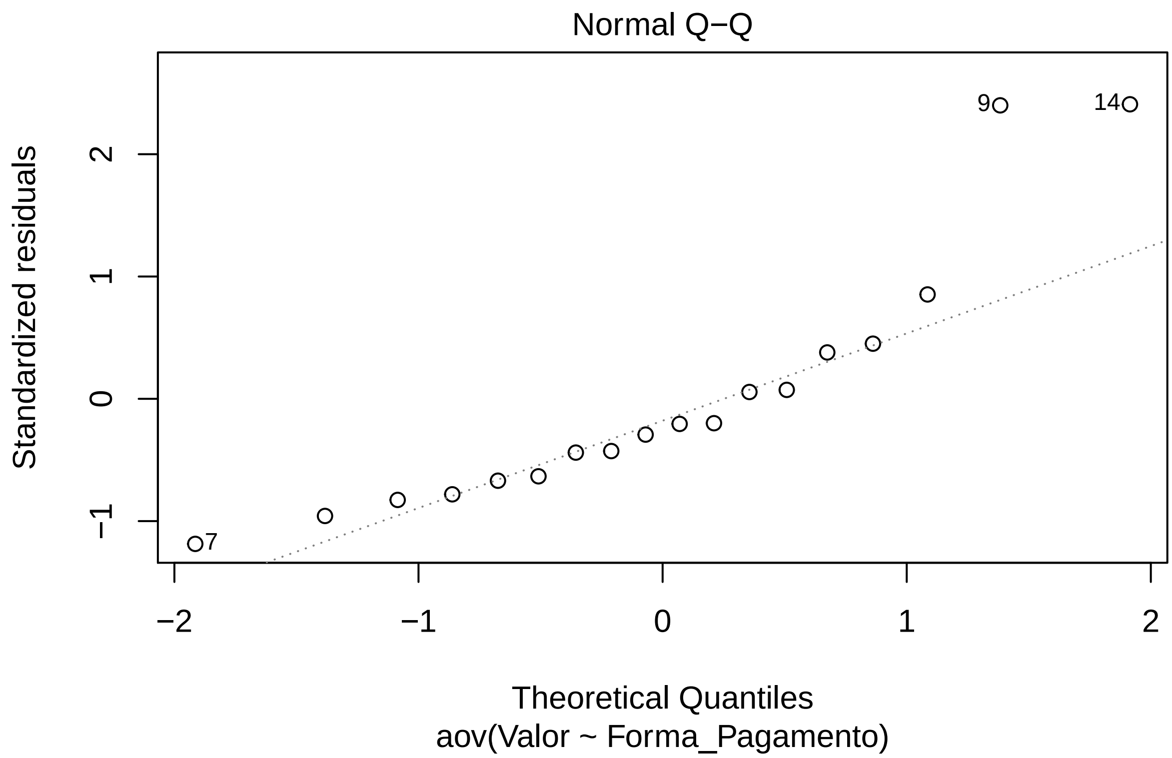
Como o valor-p $= 0.4746 > \alpha$, então não rejeita-se H_0 e pode-se prosseguir com a ANOVA.

- $H_0 : \mu_{dinheiro} = \mu_{cheque} = \mu_{cartao}$
- H_1 : Ao menos uma média é diferente

```r
compras.aov <- aov(Valor ~ Forma_Pagamento, data = compras)
summary(compras.aov)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Forma_Pagamento  2   3277    1638   2.156   0.15
## Residuals       15  11401     760
```

```r
plot(compras.aov, 2)
```




```r
shapiro.test(residuals(compras.aov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(compras.aov)
## W = 0.83074, p-value = 0.004272
```

Como os resíduos não apresentaram normalidade, como pode ser observado no gráfico, deve-se desconsiderar o resultado da ANOVA e procurar outra alternativa para comparar esses valores.

Exercício 4

Para este exercício, como não foi especificado, será fixado o nível de significância $\alpha = 0.05$

```r
aeusp <- read.csv("aeusp.txt", sep="", na.strings="")
aeusp$Renda=factor(aeusp$Renda)
aeusp$Sexo=factor(aeusp$Sexo, labels = c("Masculino", "Feminino"))
```

Item a.

```r
shapiro.test(aeusp[aeusp$Sexo == "Masculino",]$Itrab)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aeusp[aeusp$Sexo == "Masculino", ]$Itrab
## W = 0.94633, p-value = 5.11e-06
```

```r
shapiro.test(aeusp[aeusp$Sexo == "Feminino",]$Itrab)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aeusp[aeusp$Sexo == "Feminino", ]$Itrab
## W = 0.89362, p-value = 3.068e-11
```

```r
t.test(Itrab ~ Sexo, data = aeusp)
```

```
##
##  Welch Two Sample t-test
##
## data:  Itrab by Sexo
## t = 1.7247, df = 382.93, p-value = 0.08539
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1531239  2.3403363
## sample estimates:
## mean in group Masculino  mean in group Feminino
##                9.940828                8.847222
```

Fixando o nível de significância em 5%, temos que o valor-p $> \alpha$ e, portanto, não se rejeita H_0.

Conclusão

Não há diferenças significativas de Idade que começou a trabalhar entre homens e mulheres.

Item b.

Como temos 5 sub-populações neste caso, deve-se fazer o teste utilizando a ANOVA.

Primeiramente testa-se homocedasticidade dos dados. Será utilizado o teste de Bartlett.

- H_0: As variâncias são homogêneas
- H_1: As variâncias não são homogêneas

```r
bartlett.test(aeusp$Itrab ~ aeusp$Reproce)
```

```
##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  aeusp$Itrab by aeusp$Reproce
## Bartlett's K-squared = 1.7943, df = 4, p-value = 0.7735
```

Como o valor-p=0.7735 > α, então não rejeita-se H_0 e pode-se prosseguir com o teste.

- H_0: As médias da variável itrab nas sup-populações são iguais
- H_1: Ao menos uma sub-população tem média diferente de uma das demais para a variável itrab

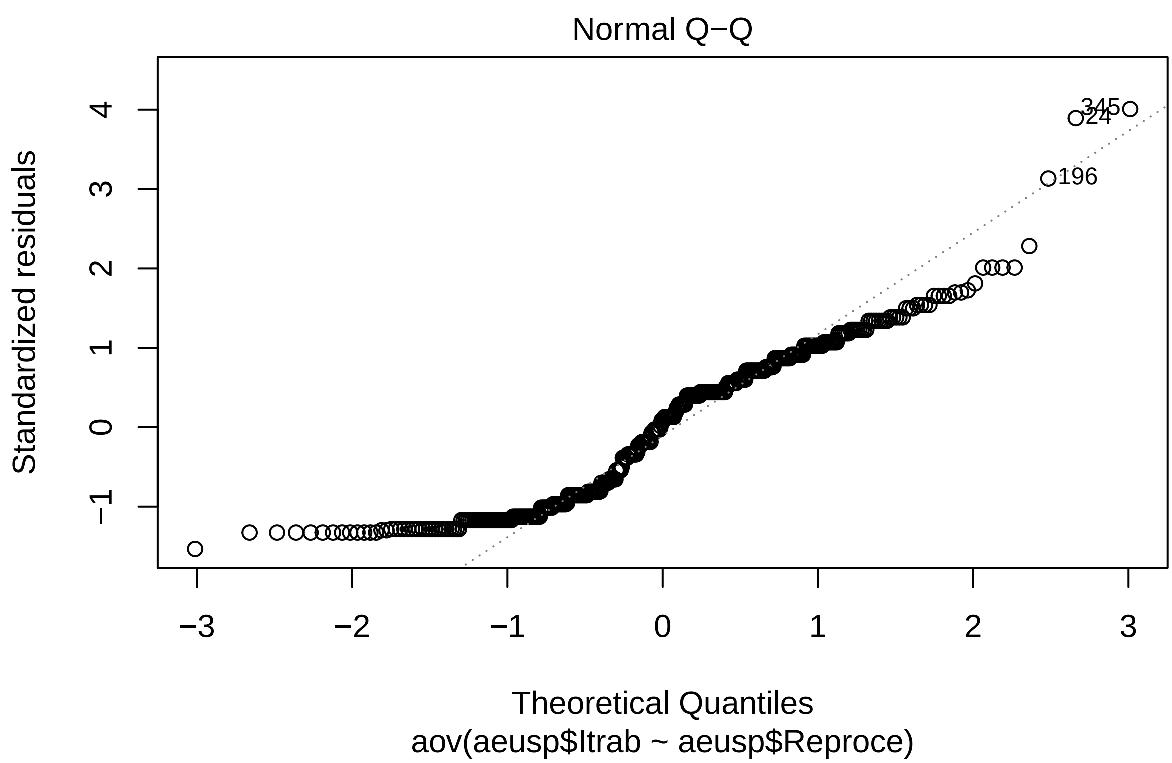
```
itrab_x_reproce.aov <- aov(aeusp$Itrab ~ aeusp$Reproce)
summary(itrab_x_reproce.aov)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## aeusp$Reproce   4     26    6.45   0.158  0.959
## Residuals     380  15527   40.86
```

```
shapiro.test(itrab_x_reproce.aov$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  itrab_x_reproce.aov$residuals
## W = 0.92825, p-value = 1.238e-12
```

```
plot(itrab_x_reproce.aov, 2)
```



Observando o teste de Shapiro-Wilk e o plot dos resíduos, pode-se verificar que os resíduos não seguem normalidade.

Neste caso, deve-se desconsiderar a ANOVA e procurar uma outra alternativa para comparar as médias das idades nas sub-populações

Conclusão:

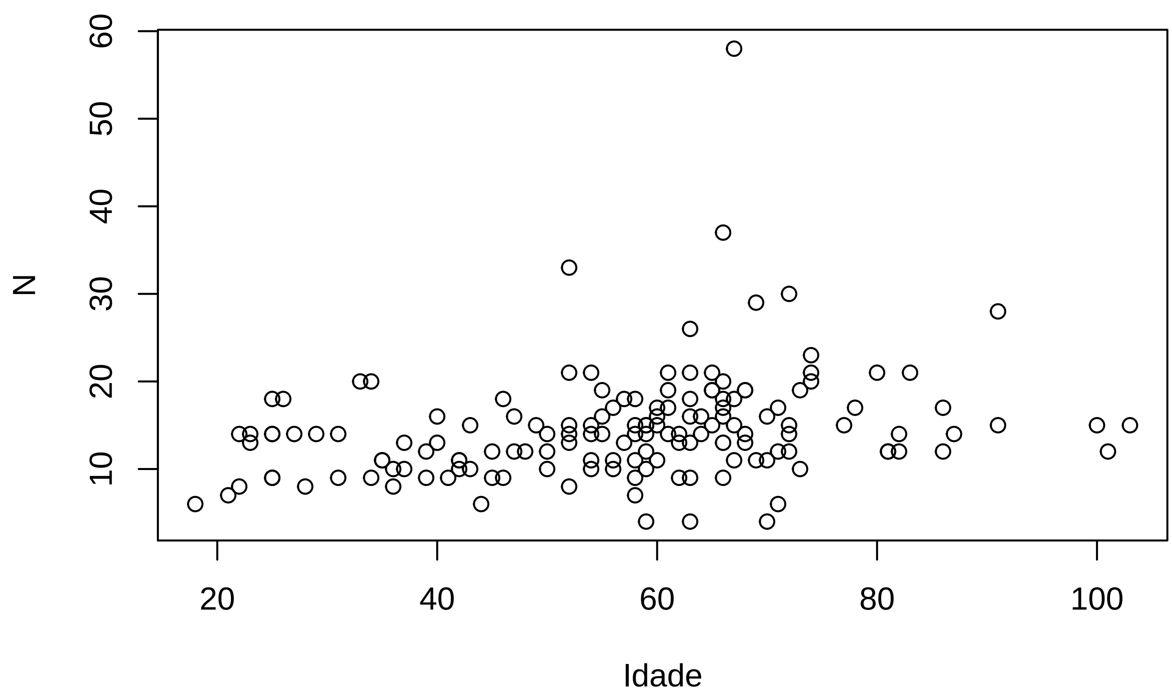
Não foi possível testar se as médias da variável itrab nas sub-populações definidas pela região de procedência são iguais.

Exercício 5

```
cancer <- read.table("cancer.txt", header = T)
cancer$Grupo <- factor(cancer$Grupo,
                       label = c("falso-negativo", "negativo", "positivo", "falso-positivo"),
                       levels = 1:4)
```

Item a.

```
casos_positivos <- cancer[cancer$Grupo == "falso-negativo" | cancer$Grupo == "positivo",]
plot(N ~ Idade, data = casos_positivos)
```



Com base no gráfico de dispersão para o N e Idade dos pacientes que têm a doença, pode-se dizer que a concentração de nitrogênio no sangue parece apresentar um leve aumento conforme aumenta a idade.

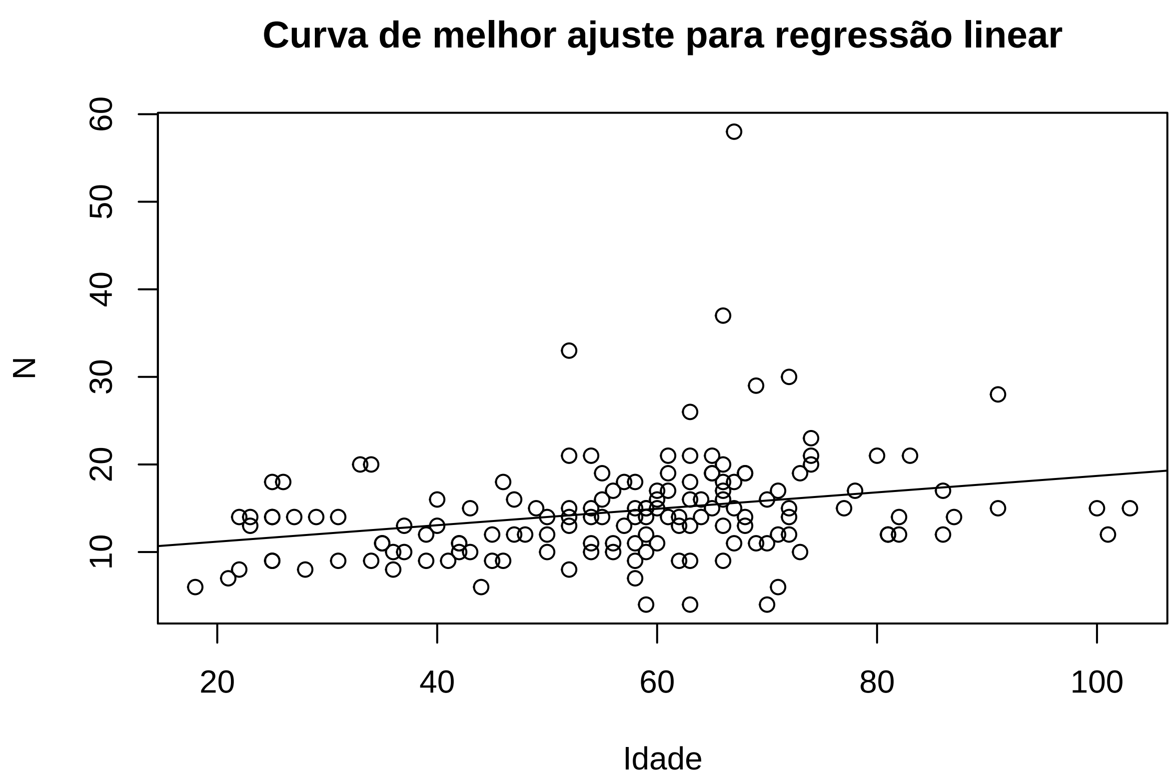
Item b.

```
n_x_idade.pos.modelo <- lm(N ~ Idade, data = casos_positivos)
summary(n_x_idade.pos.modelo)
```

```
##
## Call:
## lm(formula = N ~ Idade, data = casos_positivos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -11.879  -3.629  -0.847   2.419  42.403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31133    1.67129   5.571 1.15e-07 ***
## Idade        0.09382    0.02816   3.332  0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.099 on 149 degrees of freedom
## Multiple R-squared:  0.06935,	Adjusted R-squared:  0.06311
## F-statistic:  11.1 on 1 and 149 DF,  p-value: 0.001087
```

```r
plot(N ~ Idade, data = casos_positivos)
abline(n_x_idade.pos.modelo)
title("Curva de melhor ajuste para regressão linear")
```

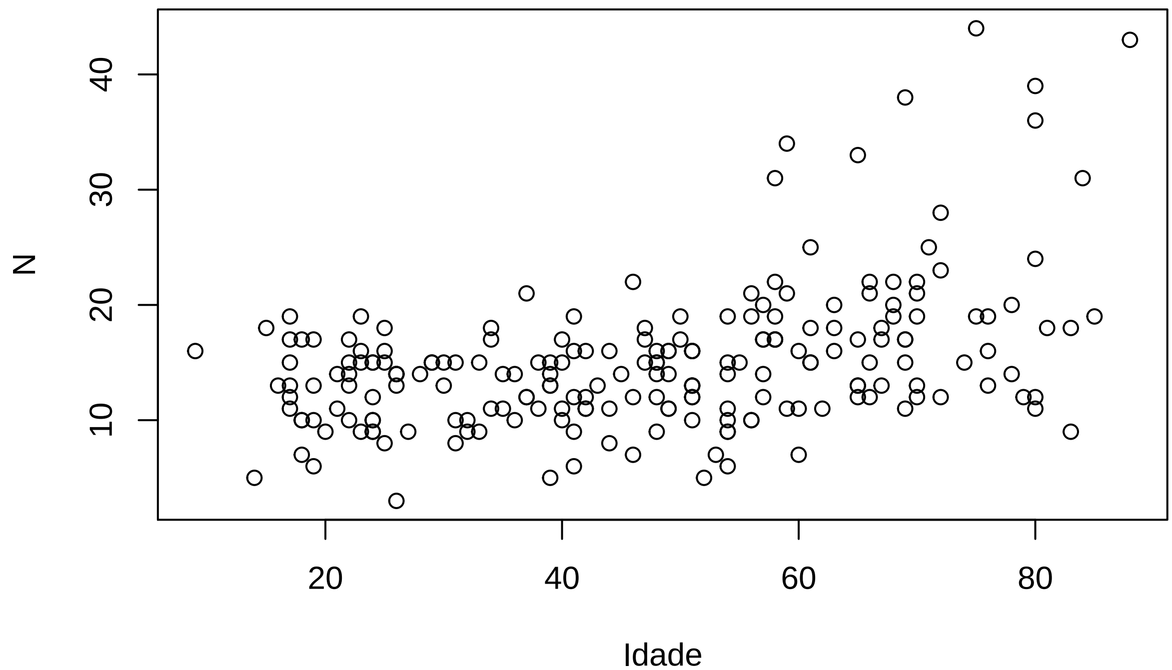


Pode-se observar que o resultado do modelo de regressão linear estima $\alpha = 9.31133$ e $\beta = 0.09382$.

Esse valor de 0.09382 para β indica que para cada unidade de aumento na variável Idade, a variável N aumenta em 0.09382. Ou seja, a concentração de nitrogênio no sangue de fato aumenta conforme a idade aumenta.

Item d.

```r
casos_negativos <- cancer[cancer$Grupo == "falso-positivo" | cancer$Grupo == "negativo",]
plot(N ~ Idade, data = casos_negativos)
```



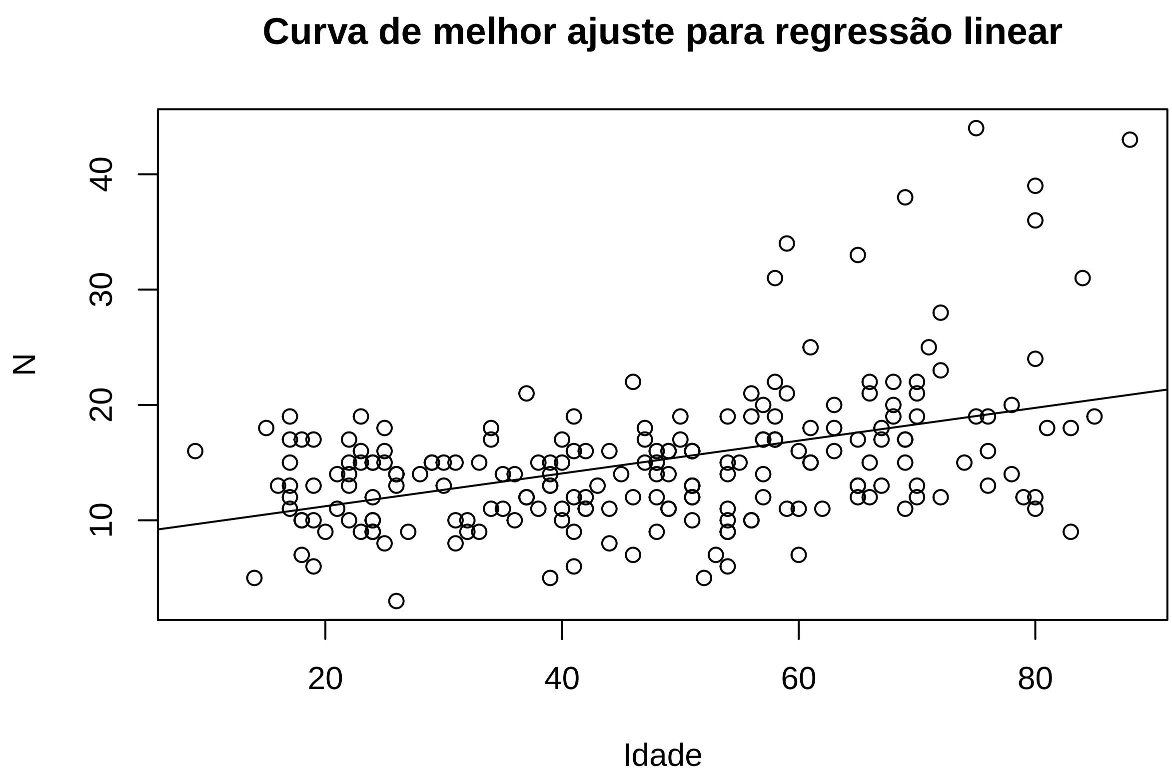
Pode-se observar neste gráfico que em pacientes que não têm a doença, à medida que a idade aumenta, a concentração de Nitrogênio teve um aumento mais expressivo do que nos pacientes que têm a doença.

Item e.

```r
n_x_idade.neg.modelo <- lm(N ~ Idade, data = casos_negativos)
summary(n_x_idade.neg.modelo)
```

```
##
## Call:
## lm(formula = N ~ Idade, data = casos_negativos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1729  -3.3466  -0.1992   2.5008  24.9639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.37812    1.03031   8.132 3.71e-14 ***
## Idade        0.14211    0.02019   7.037 2.76e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.612 on 209 degrees of freedom
## Multiple R-squared:  0.1916, Adjusted R-squared:  0.1877
## F-statistic: 49.52 on 1 and 209 DF,  p-value: 2.763e-11
```

```r
plot(N ~ Idade, data = casos_negativos)
abline(n_x_idade.neg.modelo)
title("Curva de melhor ajuste para regressão linear")
```



Pode-se observar que para este caso, o resultado do modelo de regressão linear estima $\alpha = 8.37812$ e $\beta = 0.14211$.

Esse valor de 0.14211 para β indica que para cada unidade de aumento na variável Idade, a variável N aumenta em 0.14211 Ou seja, a concentração de nitrogênio no sangue de fato aumenta conforme a idade aumenta.

Ao comparar este valor de β com o resultado do Item b., pode-se dizer que em pacientes que não têm a doença, a concentração de Nitorgênio aumenta mais com a idade do que em pacientes que têm a doença, pois $\beta_{negatvo} > \beta_{positivo}$.

Item f.

Teste χ^2 para independência

```r
shapiro.test(casos_negativos$Idade)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  casos_negativos$Idade
## W = 0.96953, p-value = 0.0001583
```

```r
shapiro.test(casos_negativos$N)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  casos_negativos$N
## W = 0.85372, p-value = 2.594e-13
```

```r
chisq.test(casos_negativos$Idade, casos_negativos$N)
```

```
## Warning in chisq.test(casos_negativos$Idade, casos_negativos$N): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  casos_negativos$Idade and casos_negativos$N
## X-squared = 2273.7, df = 2070, p-value = 0.001042
```

Como o valor-p $= 0.001042 < \alpha$, rejeita-se H_0. Então deve-se considerar que a idade influencia a concentração de nitrogênio para os pacientes sem a doença.

Item g.

Os itens anteriores sugerem que pacientes com a doença apresentam um menor crescimento da concentração de nitrogênio com a idade do que pacientes sem a doença. Então sim, pode-se dizer que esse efeito é um dado importante para discriminar entre pacientes com e sem a doença.

Exercício 6

Item a.

```r
# P(X = x), em que X ~ Zeta(a)
dzeta <- function(x, a) {
  c = 0
  for(i in 1:10^6) {
    c = c + i^a
  }
  return(1/(c*x^(a)))
}
```

Item b.

```r
# P(X <= x), em que X ~ Zeta(a)
pzeta <- function(x, a) {
  soma = 0
  for(i in 1:x) {
    soma = soma + dzeta(i, a)
  }
  return(soma)
}
```