Estatistica Descritiva - lista 3

Bruna Umino, Beatriz Vianna *IME - USP*

professora Marcia D'Elia Branco

Questão 1

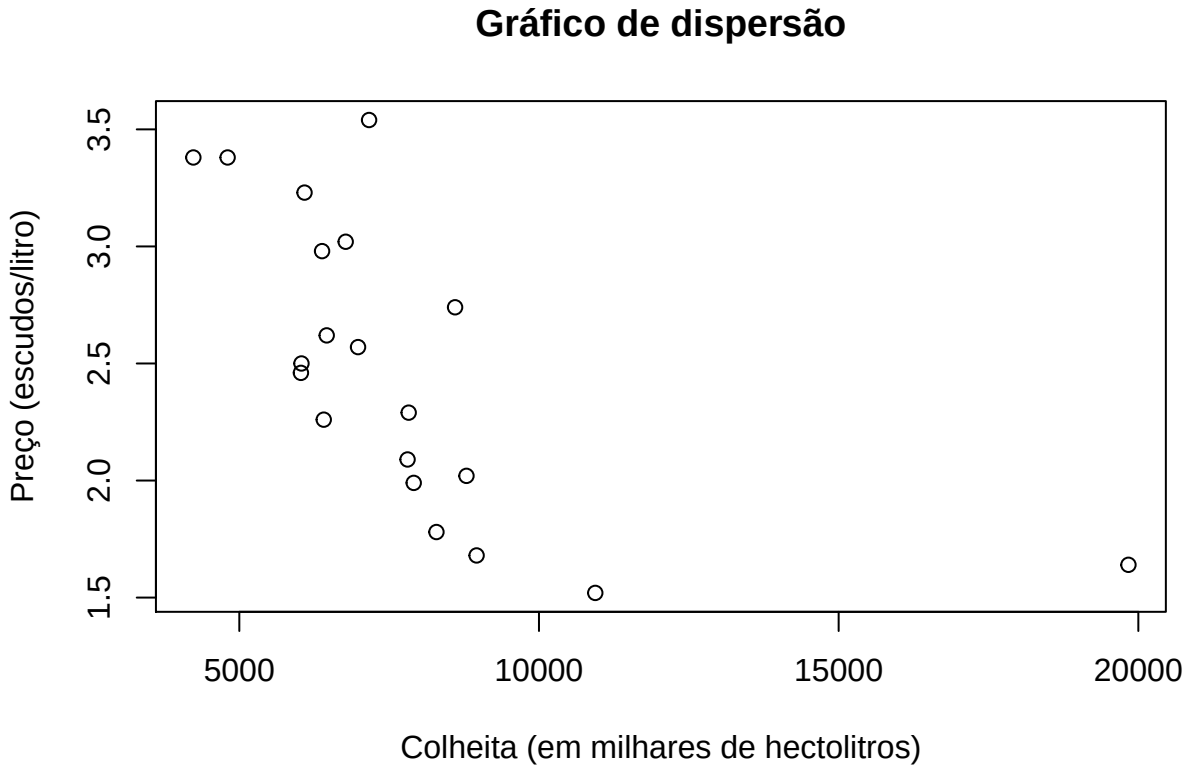
```
ano<-c(1942,1943,1944,1945,1946,1947,1948,1949,1950,1951,
       1952,1953,1954,1955,1956,1957,1958,1959,1960,1961)

colheita<-c(6409, 19835, 10939, 7826, 7165, 7807, 6028, 6037, 6458, 6981,
            4233, 8790, 8959, 8289, 7910, 6775, 6088, 6381, 8600, 4805)

preco<-c(2.26, 1.64, 1.52, 2.29, 3.54, 2.09, 2.46, 2.50, 2.62, 2.57,
         3.38, 2.02, 1.68, 1.78, 1.99, 3.02, 3.23, 2.98, 2.74, 3.38)
tabela <- data.frame(ano, colheita, preco)
```

1a)

```
#Grafico de dispersao
plot( preco ~ colheita,
      xlab = "Colheita (em milhares de hectolitros)",
      ylab = "Preço (escudos/litro)",
      main = "Gráfico de dispersão")
```



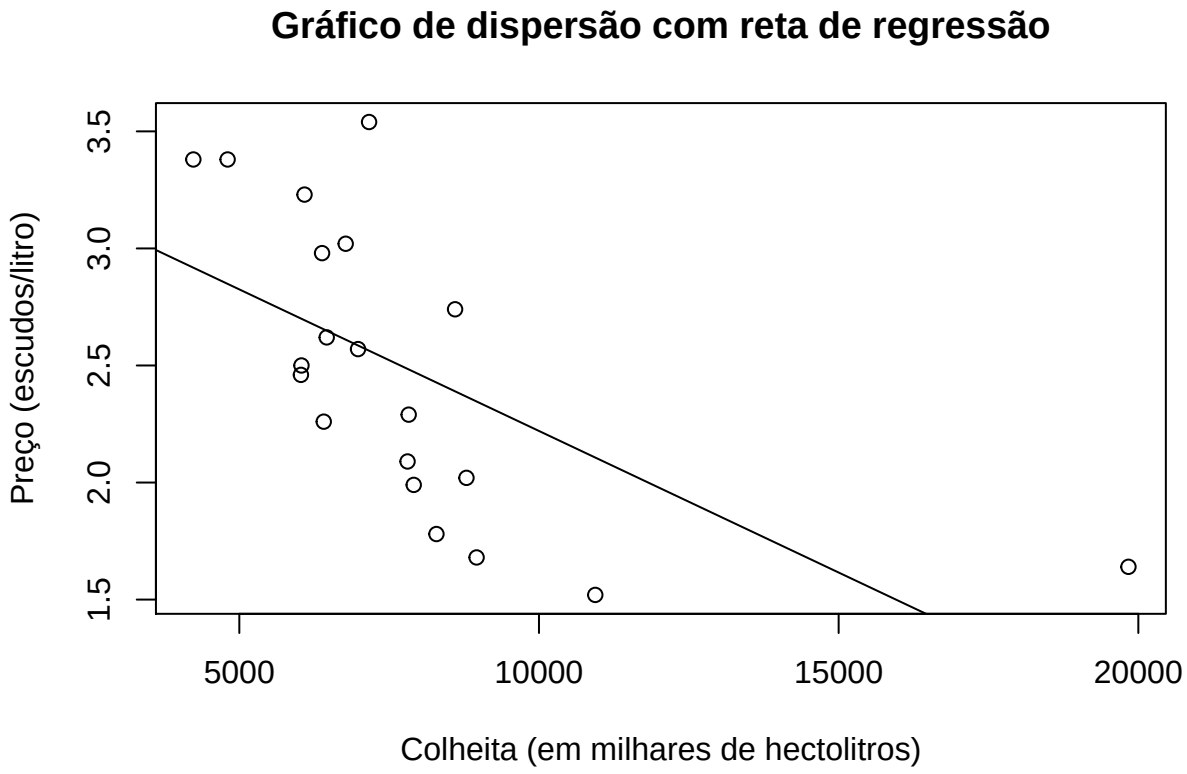

```
#Correlacao linear
cor(preco, colheita)
```

```
## [1] -0.6239457
```

Como podemos observar, o gráfico apresenta uma correlação linear significativa e negativa (ou seja, quanto menor a colheita, mais alto fica o preço), que está sendo influenciada pelo dado do ano de 1943. Devido a este valor, a correlação está mais elevada.

1b)

```
plot( preco ~ colheita,
      xlab = "Colheita (em milhares de hectolitros)",
      ylab = "Preço (escudos/litro)",
      main = "Gráfico de dispersão com reta de regressão")
abline(lm (preco ~ colheita))
```




```
lm (preco ~ colheita)
```

```
##
## Call:
## lm(formula = preco ~ colheita)
##
## Coefficients:
## (Intercept)     colheita
##   3.4297758   -0.0001209
```

Dado que o resultado do coeficiente angular foi igual a -0.0001209, podemos observar que consiste em um valor negativo, ou seja, o preço e a colheita são inversamente proporcionais. O valor deste coeficiente angular parece baixo (praticamente uma reta horizontal) devido à diferença de escala. Mas ainda assim a correlação é forte como observado na questão anterior.

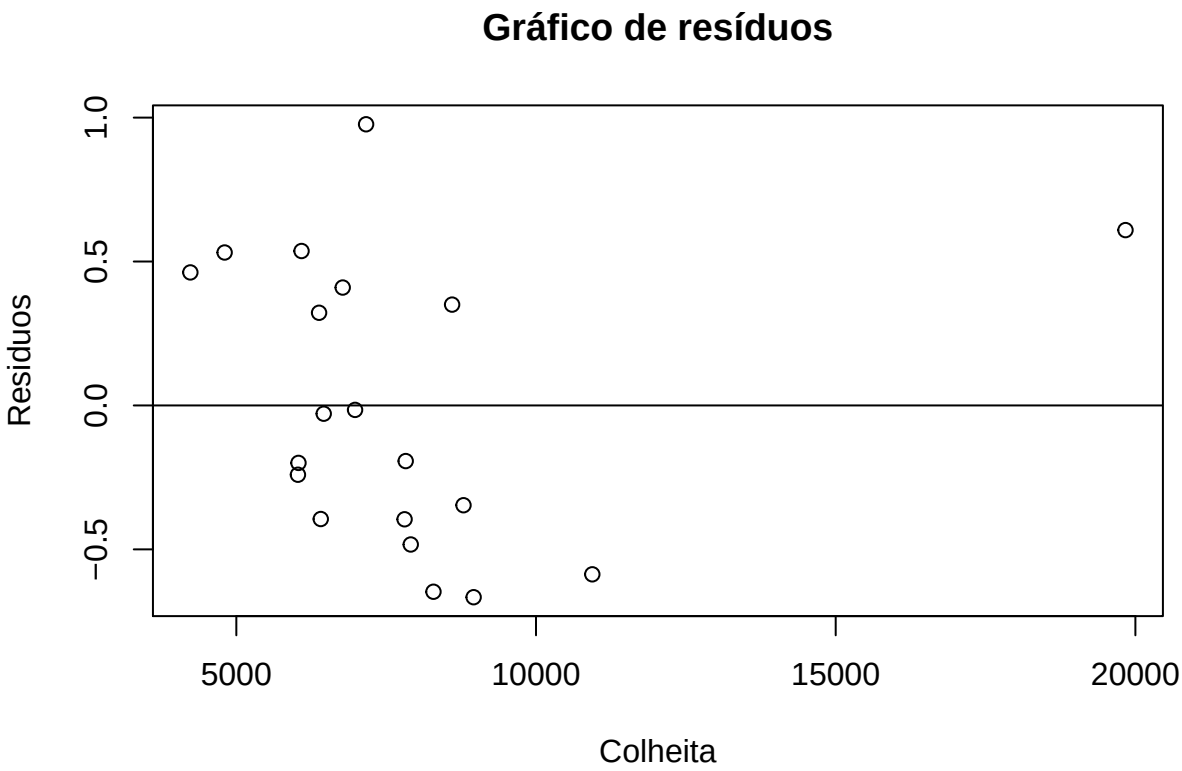
1c)

```
lm(preco~colheita)
```

```
##
## Call:
## lm(formula = preco ~ colheita)
```

```
##
## Coefficients:
## (Intercept)     colheita
##   3.4297758   -0.0001209
```

```
residuos <- resid(lm(preco~colheita))
plot(colheita, residuos,
     ylab="Residuos",
     xlab="Colheita",
     main="Gráfico de resíduos")
abline(0,0)
```

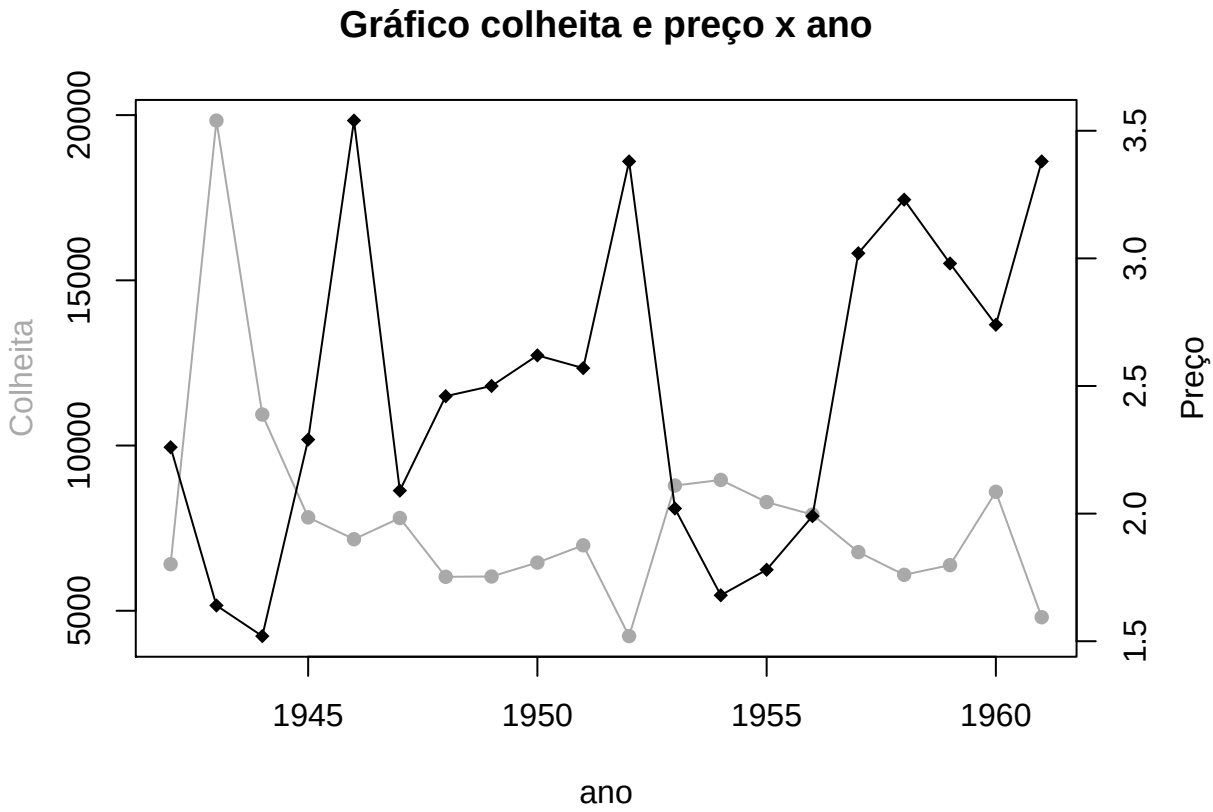


Observando o gráfico, podemos ver que aumentando o valor da colheita, não aumenta a variabilidade dos dados, então o gráfico é homocedástico.

1d)

```
par(mar=c(4,4,4,4))
#grafico de dispersão ano x colheita
plot (colheita~ano, type="o", col="darkgray", ylab = "", pch=16,
      main = "Gráfico colheita e preço x ano")
mtext("Colheita", side = 2, line = 2.5, col="darkgray")
par(new=TRUE)
#grafico de dispersão ano x preço
```

```
plot(preco~ano, axes=FALSE, type='o', col='black', ann=FALSE, pch=18)
mtext("Preço", side = 4, line = 2.5, col="black")
axis(4)
```



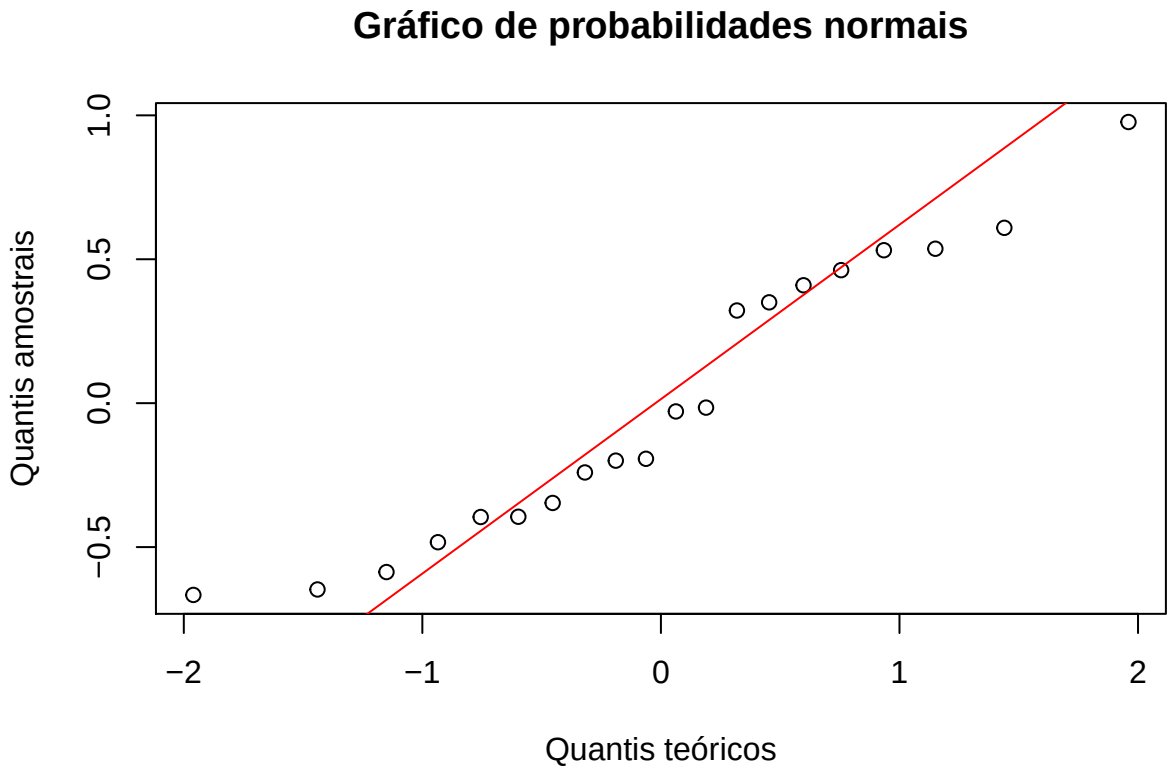
Através dos gráficos podemos observar que ocoreu uma considerável colheita no ano de 1943, que resultou no pior preço e no ano de 1952 ocorreu a menor colheita deste intervalo de tempo. O resto dos valores coletados está na faixa de 5000 a 10900 milhares de hectolitros.

Em relação ao preço, há picos nos anos de 1946, 1961, 1952 e 1958, que são os anos nos quais ocorreram as piores colheitas, e nos anos de 1944, 1943 e 1954, quando foram relatados os menores preços.

É fácil observar neste gráfico a correlação negativa entre preço e colheita, quando a colheita apresenta um pico alto ou baixo em um ano, o preço apresenta pico inverso no mesmo ano.

1e)

```
qqnorm(residuos,
       main = "Gráfico de probabilidades normais",
       xlab = "Quantis teóricos",
       ylab = "Quantis amostrais")
qqline(residuos, col="red")
```



Os quantis dos resíduos não se aproximam muito para uma normal, mas podemos observar que os valores do meio estão mais próximos da linha $Y = X$ e as caudas se afastam consideravelmente.

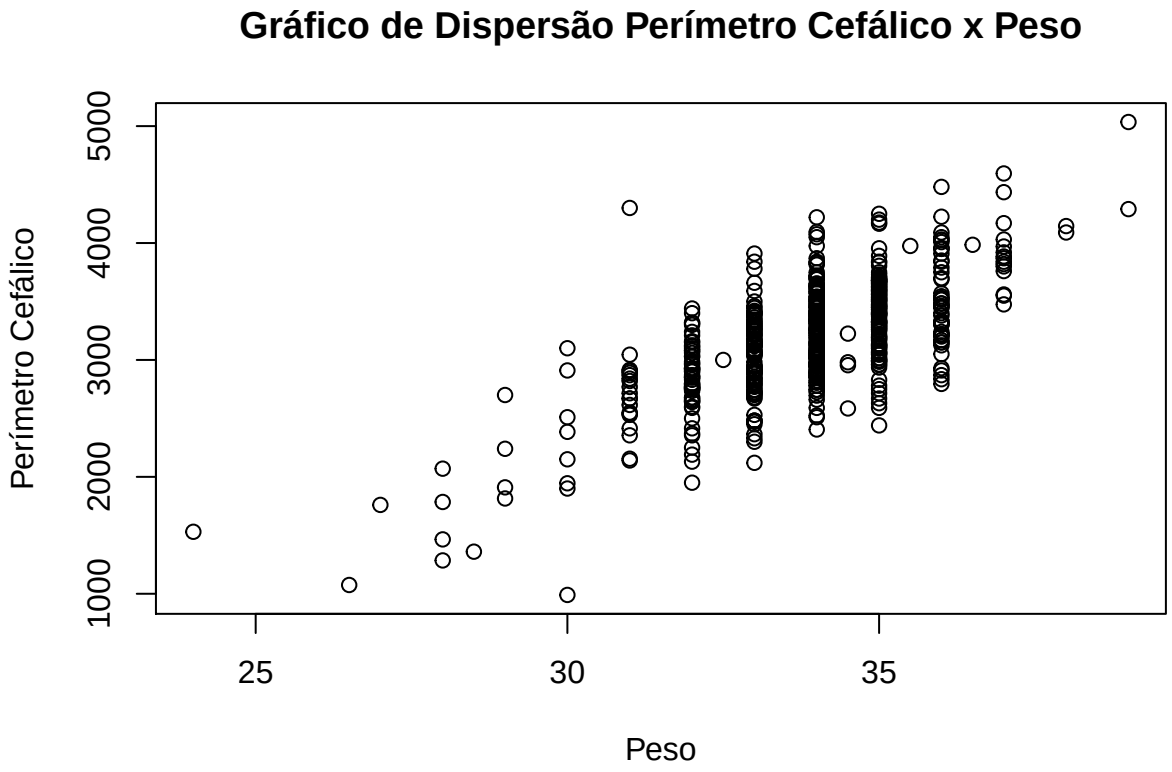
Questao 2

2a)

```
library (magrittr)
dados <- read.csv2("/home/be/viabianna/Downloads/dadosmalariaCEA15P14.csv")
#Retirar os dados que contém N/A
dados <- dados %>% subset(!is.na(pc)) %>% subset(!is.na(peso)) %>% subset(!is.na(est))
```

2a)

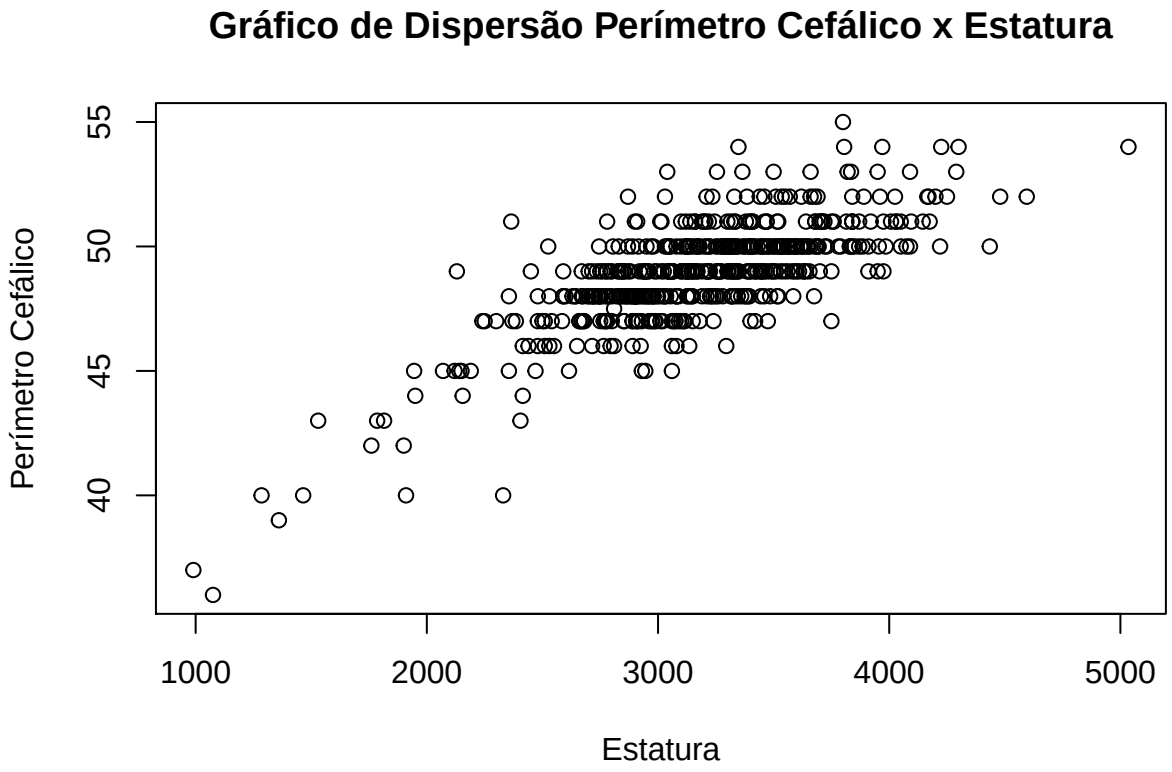
```
#Gráfico de Dispersão Perímetro Cefálico x Peso
plot(dados$peso~dados$pc, xlab="Peso", ylab="Perímetro Cefálico",
     main= "Gráfico de Dispersão Perímetro Cefálico x Peso")
```




```
#Correlação Perímetro Cefálico x Peso
cor(dados$peso,dados$pc)
```

```
## [1] 0.7036036
```

```
#Gráfico de Dispersão Perímetro Cefálico x Estatura
plot(dados$est~dados$peso, xlab="Estatura", ylab="Perímetro Cefálico",
     main= "Gráfico de Dispersão Perímetro Cefálico x Estatura")
```




```
#Correlação Perímetro Cefálico x Estatura
cor(dados$pc,dados$est)
```

```
## [1] 0.6101827
```

2b)

```
equação <- (lm(dados$pc~dados$peso))
equação
```

```
##
## Call:
## lm(formula = dados$pc ~ dados$peso)
##
## Coefficients:
## (Intercept)   dados$peso
##   26.061047     0.002441
```

A partir destes dados, sabemos que a reta de regressão para Perímetro Cefálico x Peso (que é a variável que apresenta maior correlação) será

$y = 0,0024x + 26,0610$

```
equação <- (lm(dados$pc~dados$est+dados$peso))
equação
```

```
##
## Call:
## lm(formula = dados$pc ~ dados$est + dados$peso)
##
## Coefficients:
## (Intercept)    dados$est   dados$peso
##   20.685784     0.140293     0.001972
```

Já a reta Perímetro Cefálico x Estatura e Peso será

$y = 0,1402x_e + 0,0019x_p + 20,6857$

Assim sendo, o perímetro cefálico esperado, em centímetros, para um recém nascido de 50cm e 3kg será:

$y = 0,1402 * 50 + 0,0019 * 3000 + 20,6857$

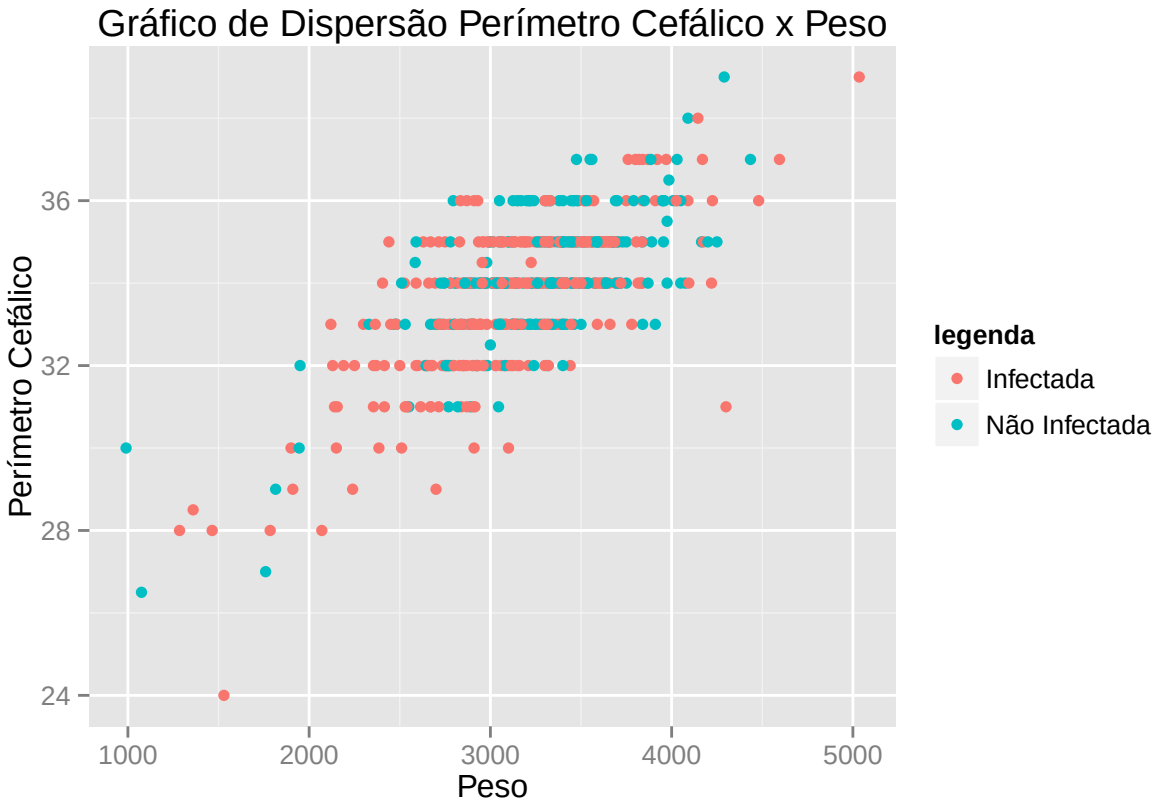
$y = 7,0100 + 5,7000 + 20,6857$

$y = 33,3957$

2c)

```
#organização dos dados a serem usados,
#transformar grupo em variável binária
dados2 <- data.frame(dados$peso, dados$grupo, dados$pc)
dadosgrupo <- vector(length=length(dados2$dados.grupo))
dadosgrupo[which(dados2$dados.grupo!=0)] <- 'Infectada'
dadosgrupo[which(dados2$dados.grupo==0)] <- "Não Infectada"
dados2$dados.grupo <- dadosgrupo
```

```
library(ggplot2)
legenda <- as.factor(dados2$dados.grupo)
ggplot(data = dados2,
      aes(x = dados2$dados.peso, y =dados2$dados.pc, colour = legenda)) +
      geom_point()+
      xlab("Peso")+
      ylab("Perímetro Cefálico")+
      labs(title="Gráfico de Dispersão Perímetro Cefálico x Peso")
```



2d)

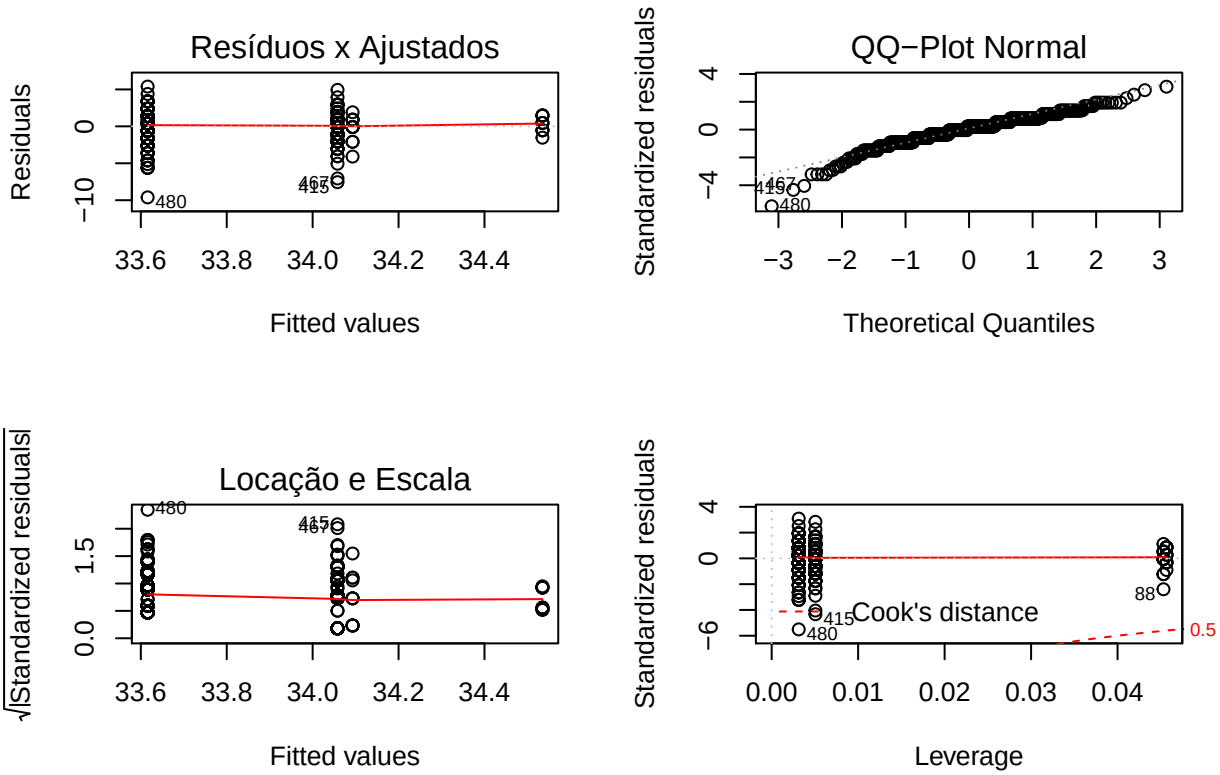
```
#organização dos dados a serem utilizados, transformar grupo em binário
#transformar idade em binária (maior que 35/ menor ou igual a 35)
dados3 <- data.frame( dados$grupo, dados$pc, dados$idade)
dadosgrupo <- vector(length=length(dados3$dados.grupo))
dadosgrupo[which(dados3$dados.grupo!=0)] <- 1
dados3$dados.grupo <- dadosgrupo
dadosgrupo2 <- vector(length=length(dados3$dados.idade))
dadosgrupo2[which(dados3$dados.idade<=35)] <- 0
dadosgrupo2[which(dados3$dados.idade>35)] <- 1
dados3$idadecat<- dadosgrupo2
```

```
fit1 <- lm(dados3$dados.pc~dados3$dados.grupo+dados3$idadecat)
summary(fit1)
```

```
##
## Call:
## lm(formula = dados3$dados.pc ~ dados3$dados.grupo + dados3$idadecat)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.6159 -1.0579  0.3841  1.3841  5.3841
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          34.0579     0.1244 273.840  < 2e-16 ***
## dados3$dados.grupo   -0.4420     0.1565  -2.825  0.00491 **
## dados3$idadecat       0.4771     0.3731   1.279  0.20157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.748 on 526 degrees of freedom
## Multiple R-squared:  0.01853,    Adjusted R-squared:  0.0148
## F-statistic: 4.967 on 2 and 526 DF,  p-value: 0.007298
```

```
par(mfrow=c(2,2))
plot(fit1, caption = c("Resíduos x Ajustados", "QQ-Plot Normal",
                       "Locação e Escala", "resìduos e alavancagem"))
```



O modelo está ajustando o perímetro cefálico em relação ao grupo da gestante (0=não infectada e 1=infectada) e a idade da gestante (0= até 35 anos e 1=mais de 35 anos)

O desvio padrão dos resíduos é 1.748

Devido aos dados estarem agrupados de forma binária, os gráficos com exceção do qqnorm saem alinhados verticalmente. Em relação ao gráfico de residuos podemos observar que a medida que aumenta o eixo x (a idade e o grupo), diminui a variabilidade dos dados, mostrando uma não homocedastidade. No entanto, analisando o gráfico qqnorm, o modelo se aproxima de uma normal, exceto nas caudas.

Questão 3

Ao se fazer um diagnóstico binário, no qual Y assume apenas dois valores (positivo e negativo), queremos uma regra de predição que minimize os erros cometidos. Se tomarmos $\pi = 1$ por exemplo, nosso Y_i sempre será positivo. Isso irá maximizar o diagnóstico de verdadeiros positivos, mas também irá minimizar o diagnóstico dos quadros negativos, ou seja, também teremos muitos falsos positivos (valores negativos que foram diagnosticados erroneamente como positivos).

Para a escolha do valor de π é utilizada a curva ROC (do inglês Receiver Operating Characteristic - Característica de operação do receptor). Este gráfico apresenta em seu eixo vertical $P(Y_i = 1|Y = 1)$ - chamado sensibilidade - e em seu eixo horizontal $1 - P(Y_i = 0|Y = 0)$ - chamado especificidade. A curva apresenta a associação entre as duas variáveis (sensibilidade e especificidade) para cada valor de π entre 0 e 1. O que procuramos é o ponto da curva que apresenta um valor muito alto para a variável do eixo y e um muito baixo para a variável do eixo x.

A tabela abaixo mostra os possíveis resultados de um teste:

Resultado do teste	positivos	negativos
positivo	verdadeiros positivos (VP)	falsos positivos(FP)
negativo	falsos negativos (FN)	verdadeiros negativos (VN)
total	total positivos (VP+FN)	total negativos (FP+VN)
desempenho	sensibilidade $S = \frac{VP}{(VP+FN)}$	especificidade $E = \frac{VN}{(FP+VN)}$

Não existe uma fórmula pré-definida para a escolha do ponto ótimo, com melhor desempenho (sensibilidade e especificidade altas), pois essa escolha também depende do teste que está sendo feito e da diferença entre a gravidade da consequência de um FP ou um FN. Há testes por exemplo nos quais é melhor manter o valor de π alto, e diminuir a especificidade para ter maior sensibilidade, pois as consequências de um falso negativo (diagnóstico errado de um exame positivo) seriam piores que as de um falso positivo.

A curva ROC também é útil na hora de analisar a acurácia de um teste. Um teste de resultados aleatórios com $P(Y_i = 1) = P(Y_i = 0) = \frac{1}{2}$ teria como curva ROC esperada a reta $x = y$. Assim sendo, quanto mais a curva ROC se afasta da reta $x = y$, aproximando-se dos cantos esquerdo e superior do gráfico, mais acurado é o teste (com altas sensibilidade e especificidade).

Questão 4

MQ $\rightarrow$ mínimos quadrados

$y_i = \alpha + \beta x_i + \varepsilon_i \rightarrow \varepsilon_i = y_i - \alpha - \beta x_i$

Como x tem valores 0 ou 1:

$x = 0 \rightarrow \varepsilon_i = y_i - \alpha$

$x = 1 \rightarrow \varepsilon_i = y_i - \alpha - \beta$

$\varepsilon_i = y_i - \alpha - \beta x_i \rightarrow \sum_{i=i}^{n} (\varepsilon_i)^2$

$\varepsilon_i = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$
$\varepsilon_i = \sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i)$
Fazendo os mínimos quadrados:
$\frac{d}{d\alpha}\sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i)$
$= \sum_{i=1}^{n} 2\alpha - 2y_i + 2\beta x_i = 2[n\alpha - \sum_{i=1}^{n} y_i + \beta\sum_{i=1}^{n} x_i]$
Para provar que é o mínimo:
$\frac{d}{d\alpha}\sum_{i=1}^{n} 2\alpha - 2y_i + 2\beta x_i = 2 > 0$

$\frac{d}{d\beta}\sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i) = \sum_{i=1}^{n} 2\beta x_i^2 - 2x_i y_i + 2\alpha x_i$
$= 2[\beta\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \alpha\sum_{i=1}^{n} x_i]$
Para provar que é o mínimo:
$\frac{d}{d\beta}\sum_{i=1}^{n} 2\beta x_i^2 - 2x_i y_i + 2\alpha = 2\sum_{i=1}^{n} x_i^2 > 0$ (é sempre positiva dado que x_i está ao quadrado)

Se:
$n\alpha - \sum_{i=1}^{n} y_i + \beta\sum_{i=1}^{n} x_i = 0$
$\beta\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \alpha\sum_{i=1}^{n} = 0$
então
$\alpha = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\beta}{n}\sum_{i=1}^{n} x_i$
$\beta = \frac{\sum_{i=1}^{n} x_i y_i - \alpha\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$
Substituindo α em β temos:
$\beta = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \frac{\beta}{n}\sum_{i=1} n x_i \sum_{i=1}^{n}}{\sum_{i=1}^{n} x_i^2}$
$\rightarrow \beta = \frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}$

$\sum_{i=1}^{n} y_i = \sum_{j=1}^{n} y_{1j} + \sum_{j=1}^{n} y_{2j}$
sendo y_1 os valores de y quando $x = 0$ e y_2 os valores de y correspondentes a $x = 1$.
Sabemos que $\sum_{i=1}^{n} x_i y_i = \sum_{j=1}^{n} y_{2j}$ e que $\sum_{i=1}^{n} x_i = n_2 = \sum_{i=1}^{n} x_i^2$
$\beta = \frac{\sum_{i=1}^{n} y_{2j} - \frac{1}{n}\sum_{i=1}^{n} y_i(n_2)}{n_2 - \frac{(n_2)^2}{n}}$
$\beta = \frac{n\sum_{i=1} n y_{2j} - \sum_{i=1}^{n} y_i(n_2)(\frac{1}{n_2})}{n(n_2) - (n_2)^2}$
$\beta = \frac{n\bar{y}_2 - \sum_{i=1}^{n} y_i}{n - n_2}$
$\beta = \frac{n\bar{y}_2 - \sum_{i=1}^{n} y_i}{n_1}$
$\beta = \frac{n\bar{y}_2}{n_1} - \frac{\sum_{i=1}^{n} y_{1j}}{n_1} - \frac{\sum_{i=1}^{n} y_{2j}}{n_1}$
$\beta = \frac{n}{n_1}\bar{y}_2 - \bar{y}_1 - \frac{n_2\bar{y}_2}{n_1}$
$\beta = \frac{(n_1+n_2)\bar{y}_2 - n_1\bar{y}_1 - n_2\bar{y}_2}{n_1}$
$\rightarrow \beta = \bar{y}_2 - \bar{y}_1$

Substituindo $\beta = \bar{y}_2 - \bar{y}_1$ em α
$\alpha = \frac{1}{n}\sum_{i=1}^{n} y_1 - \frac{\bar{y}_2 - \bar{y}_1}{n}\sum_{i=1}^{n} x_i$
$\alpha = \frac{\sum_{i=1}^{n} y_1 - n_2\bar{y}_2 + n_2\bar{y}_1}{n}$
$\alpha = \frac{\sum_{j=1}^{n} y_{1j} + \sum_{j=1}^{n} y_{2j} - n_2\bar{y}_2 + n_2\bar{y}_1}{n} = \frac{n_1\frac{\sum_{j=1}^{n} y_{1j}}{n_1} + n_2\frac{\sum_{j=1}^{n} y_{2j}}{n_2} - n_2\bar{y}_2 + n_2\bar{y}_1}{n}$
$\alpha = \frac{n_1\bar{y}_1 + (n_2\bar{y}_2 - n_2\bar{y}_2) + n_2\bar{y}_1}{n} = \frac{(n_1+n_2)\bar{y}_1}{n}$
$\rightarrow \alpha = \bar{y}_1$