lista 4 - descritiva

Lista 4 - Estatística Descritiva

Professora: Márcia D'Elia Branco

Nomes:

`Bruna Umino`

`Beatriz Vianna`

Questão 1

```
UPM <- c(100, 100, 125, 125, 150, 150, 175, 175, 200, 200, 225, 225)
clientes <- c(30, 44, 114, 138, 155, 163, 145, 163, 158, 126, 126, 106)
```

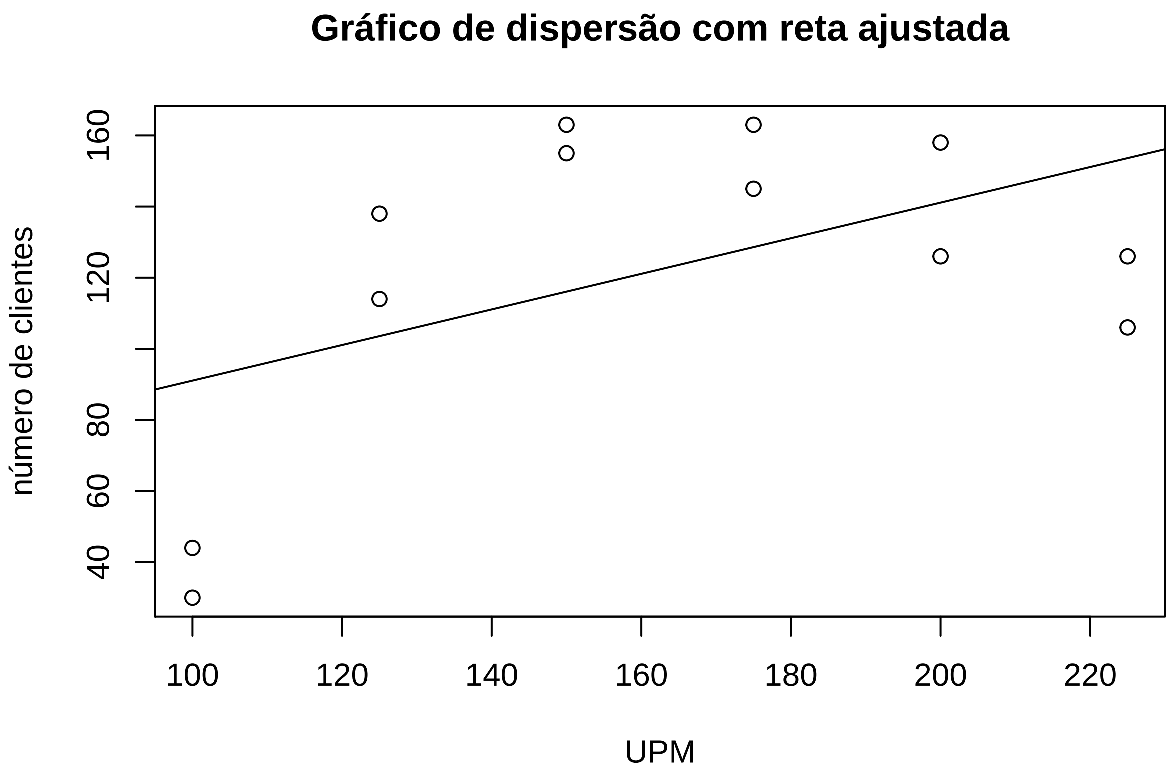
1a

```
lm(clientes~UPM)
```

```
##
## Call:
## lm(formula = clientes ~ UPM)
##
## Coefficients:
## (Intercept)          UPM
##     40.9905       0.5006
```

1b

```
modelo1 <- lm(clientes~UPM)
plot(UPM, clientes, xlab = "UPM", ylab = "número de clientes", main="Gráfico de dispersão com reta ajus
abline(modelo1)
```



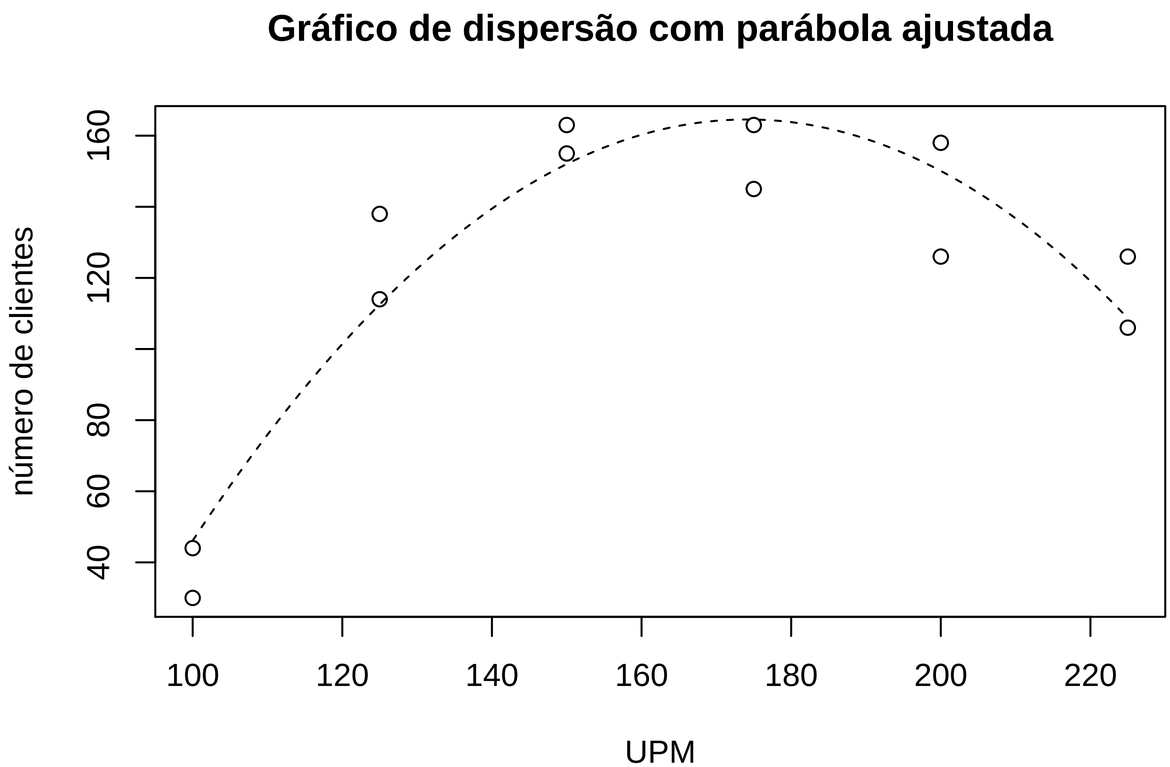
Como podemos observar, a reta ajustada não é o melhor modelo para ajustar os dados, pois não se comportam linearmente

1c

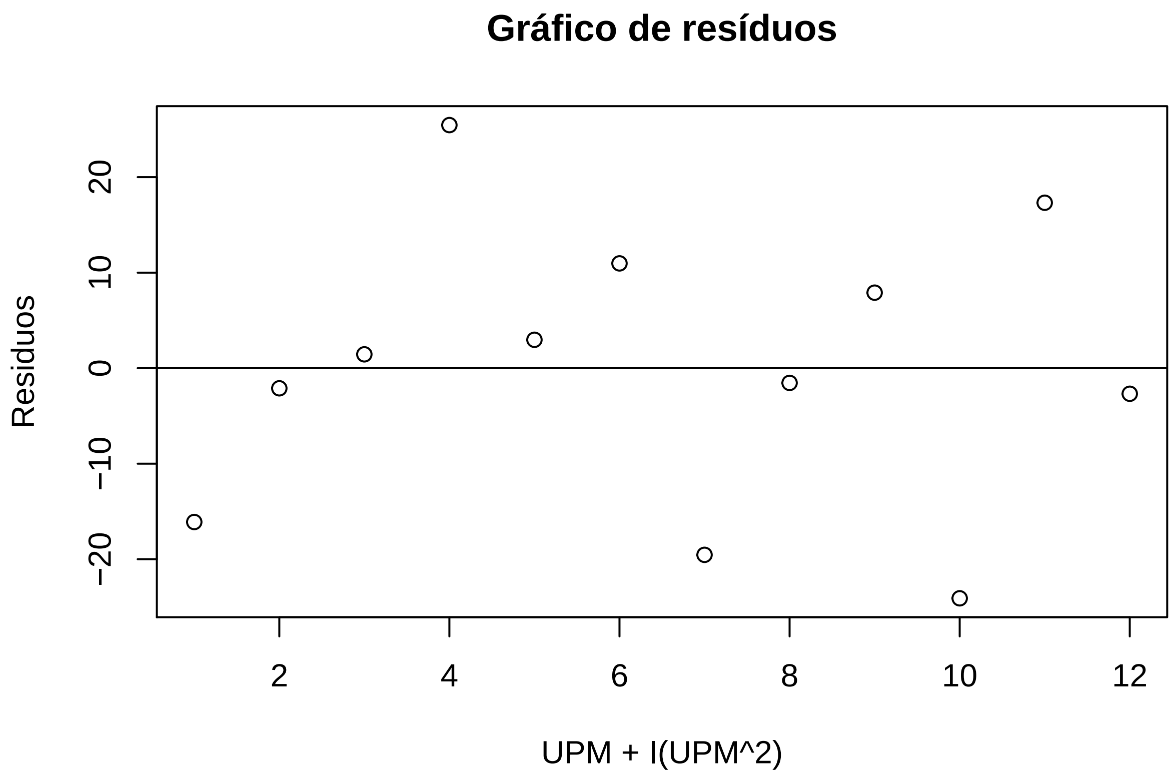
```
plot(clientes~UPM, xlab = "UPM", ylab = "número de clientes", main="Gráfico de dispersão com parábola a
modelo1 <- lm(clientes~UPM)
modelo2 <- update(modelo1,.~. +I(UPM^2))
anova(modelo1,modelo2)
```

```
## Analysis of Variance Table
##
## Model 1: clientes ~ UPM
## Model 2: clientes ~ UPM + I(UPM^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     10 15949.4
## 2      9  2377.4  1     13572 51.379 5.263e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#abline(modelo1)
cf.m2 <- coef(modelo2)
curve(cf.m2[1]+cf.m2[2]*x+cf.m2[3]*x^2, add=T, lty=2)
```




```
residuos <- residuals(modelo2)
plot(residuos,
     ylab="Residuos",
     xlab="UPM + I(UPM^2)",
     main="Gráfico de resíduos")
abline(0,0)
```



Observando o gráfico, podemos ver que mesmo com o aumento o valor do eixo x, não aumenta a variabilidade dos dados, então o gráfico é homocedástico.

Questão 2

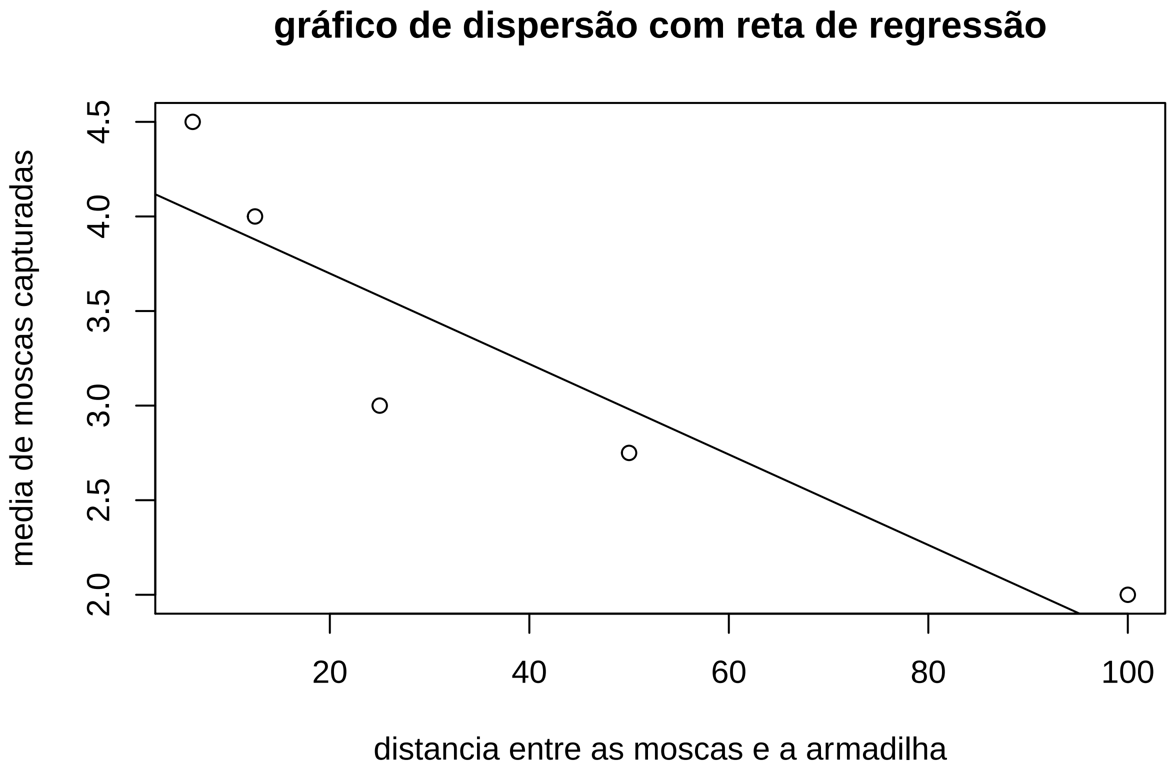
```
distancia <- c(6.25, 12.5, 25.0, 50.0, 100.0)
primeiro <- c(5, 5, 4, 3, 1)
segundo <- c(3,2,5,4,2)
terceiro <- c (4,5,3,2,2)
quarto <-  c(6,4,0,2,3)
medias <- c (0,0,0,0,0)

tabela <- data.frame (distancia, primeiro, segundo, terceiro, quarto)
medias <- c(rowMeans (tabela[,2:5]))
tabela$medias <- medias
tabela
```

```
##   distancia primeiro segundo terceiro quarto medias
## 1      6.25        5       3        4      6   4.50
## 2     12.50        5       2        5      4   4.00
## 3     25.00        4       5        3      0   3.00
## 4     50.00        3       4        2      2   2.75
## 5    100.00        1       2        2      3   2.00
```

Pela tabela, é possível notar que existe uma relação entre a distância das moscas e a média de moscas capturadas (a medida que uma aumenta, a outra diminui). Iremos ajustar uma reta para verificar se essa relação é linear:

```r
plot (tabela$medias~tabela$distancia,
  xlab="distancia entre as moscas e a armadilha",
  ylab="media de moscas capturadas",
  main="gráfico de dispersão com reta de regressão")
abline (lm(tabela$medias~tabela$distancia))
```

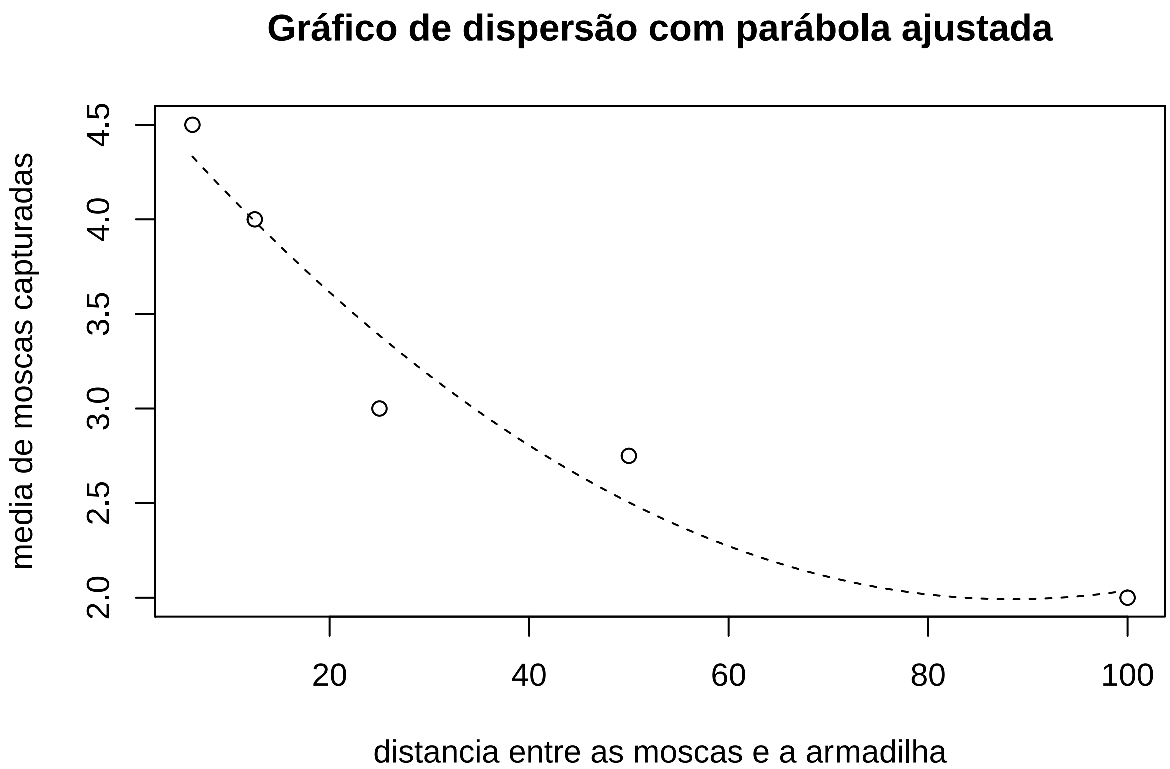


Ao observar o gráfico obtido e a relação entre os pontos e a reta de regressão, é fácil ver sem mais cálculos que a dispersão dos pontos não é linear, e parece aproximar-se de uma parábola. Façamos o gráfico a seguir:

```r
plot(tabela$medias~tabela$distancia, xlab = "distancia entre as moscas e a armadilha", ylab = "media de
modelo1 <- lm(tabela$medias~tabela$distancia)
modelo2 <- update(modelo1,.~. +I(tabela$distancia^2))
anova(modelo1,modelo2)
```

```
## Analysis of Variance Table
##
## Model 1: tabela$medias ~ tabela$distancia
## Model 2: tabela$medias ~ tabela$distancia + I(tabela$distancia^2)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      3 0.67297
## 2      2 0.23948  1   0.43348 3.6201 0.1974
```

```r
cf.m2 <- coef(modelo2)
curve(cf.m2[1]+cf.m2[2]*x+cf.m2[3]*x^2, add=T, lty=2)
```



Esta distribuição parece mais próxima do gráfico de dispersão com parábola ajustada. Temos que estar atentos ao fato de que há apenas cinco observações, portanto é difícil chegar a dados conclusivos sobre o modelo. O que pode-se dizer é que existe uma relação entre a distância da armadilha e a quantidade de moscas capituradas, e que esta relação parece ser quadrática.

Questão 3

10	11	12	13	14
-2	26	-2	0	-4
0	-4	-6	4	0
-4	-2	2	-2	-4
12	-6	8	0	4
-2	2	-2	-4	-6
		2		-2

A partir da análise dos valores dos resíduos, é possível notar muitos valores negativos, mais próximos de zero, alguns positivos próximos de zero também e uns poucos positivos mais distantes de zero, como 12 e principalmente 26. Este valor, 26, está muito distante da reta de regressão linear em comparação a todos os outros, o que é um forte indicativo de que pode ser um vilão. Se esta observação fosse excluída ao se ajustar a reta, esta reta ficaria um pouco mais para baixo, e os valores de seus resíduos seriam mais próximos de zero.

Questão 4

```
escore <- c(9,13,6,8,10,4,14,8,11,7,9,7,5,14,13,16,10,12,11,14,15,18,7,16,9,9,11,13,15,13,10,11,6,17,14
resposta <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

```
ajuste_glm <- glm(resposta ~ escore, family = binomial())
summary(ajuste_glm)
```

```
##
## Call:
## glm(formula = resposta ~ escore, family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6702  -0.7402  -0.4749   0.5200   2.1157
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.4040     1.1918   2.017  0.04369 *
## escore       -0.3235     0.1140  -2.838  0.00453 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.806  on 53  degrees of freedom
## Residual deviance: 51.017  on 52  degrees of freedom
## AIC: 55.017
##
## Number of Fisher Scoring iterations: 5
```

A partir dos dados podemos observar que a estimativa do intercepto é 2.4040 e do coeficiente angular é -0.3235, ou seja, quanto maior o seu escore no exame psicológico, maior será sua chance de não ter ocorrência de sintomas de demência senil. Além disso, o Pr(>|z|) mostra os p valores correspondentes aos z values (quociente da estimativa pelo erro padrão) em uma distrubuição normal, observando os p valores e pela quantidade de * sabemos que o valor 0.00453 se aproxima mais do centro da normal que o valor 0.04369, ou seja, é mais importatnte para a análise que o valor dado pelo intercepto. Por fim, o residual deviance apresenta a falta de ajuste do modelo como um todo e o null deviance é a mesma medida reduzida à apenas o intercepto.

Questão 5

```
library(car)
dados <- Duncan
```