lista 4 - Estatística descritiva

Professora: Márcia D'Elia Branco

Bruna Umino

Beatriz Vianna

Questão 1

```r
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)
```

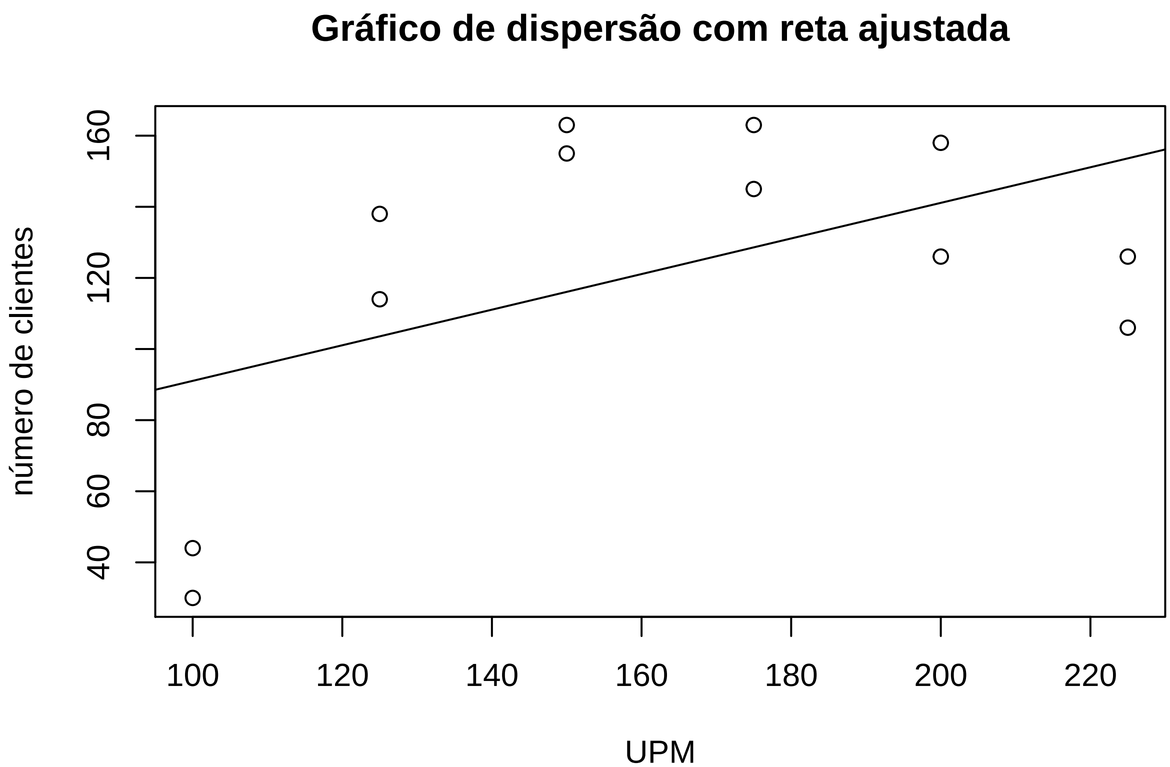
1a

```r
summary(lm(clientes~UPM))
```

```
##
## Call:
## lm(formula = clientes ~ UPM)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -61.05 -32.48  13.42  34.42  46.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.9905    45.3678   0.904   0.3875
## UPM           0.5006     0.2700   1.854   0.0935 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.94 on 10 degrees of freedom
## Multiple R-squared:  0.2558, Adjusted R-squared:  0.1813
## F-statistic: 3.437 on 1 and 10 DF,  p-value: 0.09346
```

1b

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



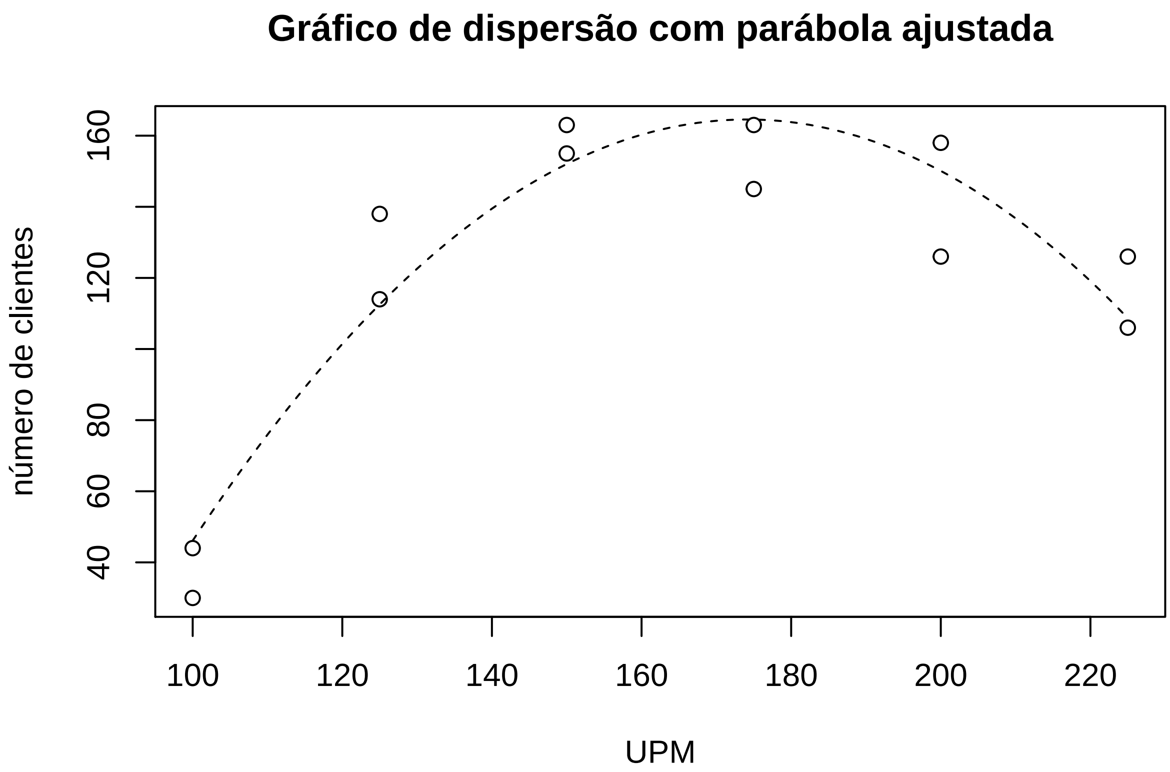
Como podemos observar, a reta ajustada não é o melhor modelo para ajustar os dados, pois não se comportam linearmente

1c

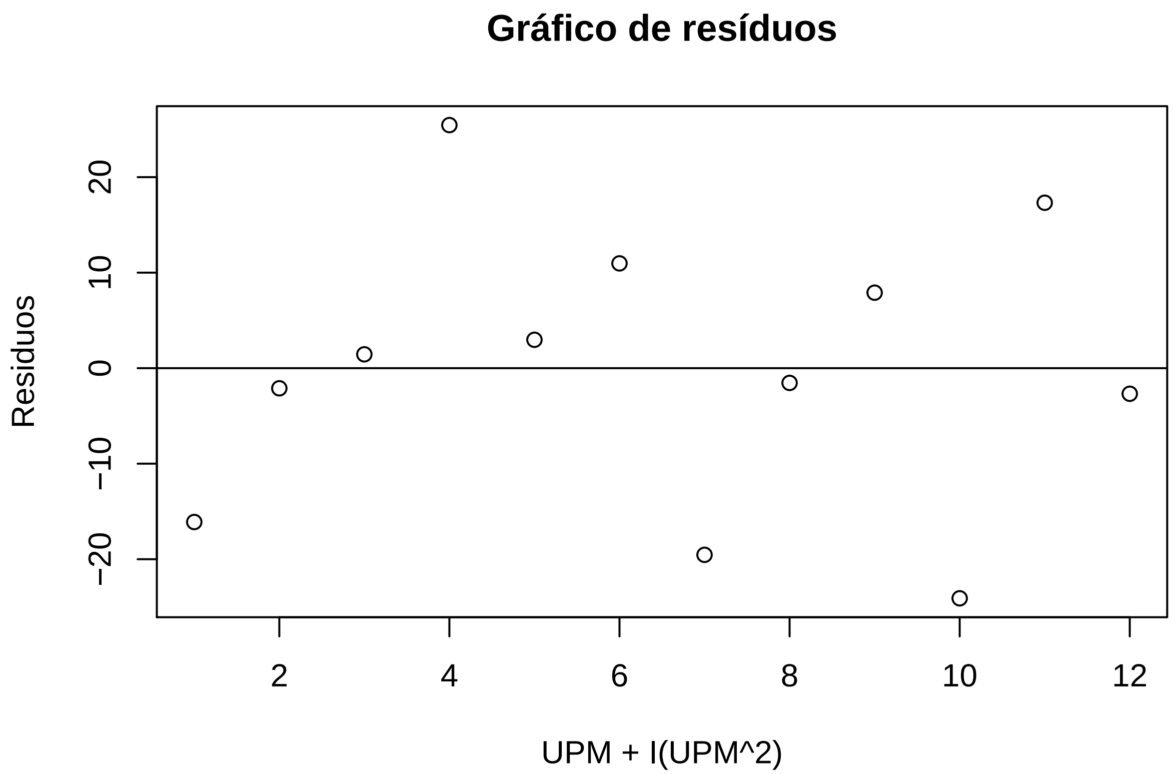
```
plot(clientes~UPM, xlab = "UPM", ylab = "número de clientes",
     main="Gráfico de dispersão com parábola ajustada")
modelo1 <- lm(clientes~UPM)
modelo2 <- update(modelo1,.~. +I(UPM^2))
anova(modelo1,modelo2)
```

```
## 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
```

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




```r
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

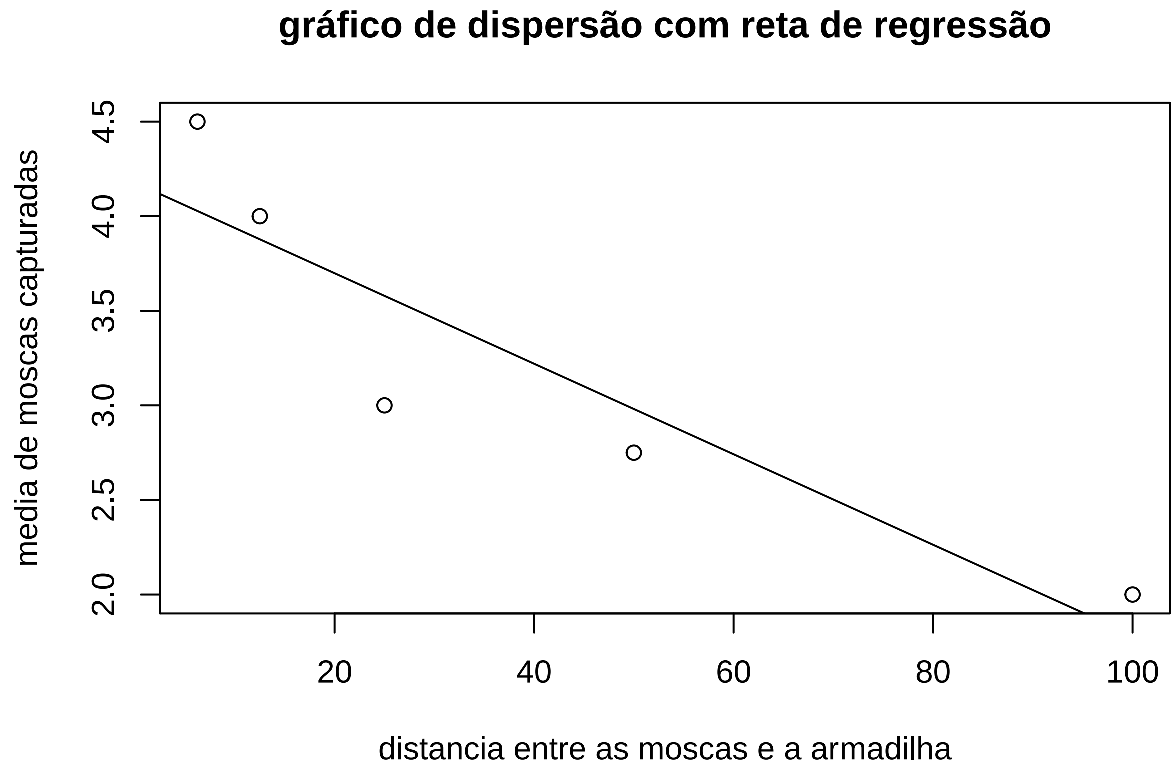
```
distancia <- c(6.25, 12.5, 25.0, 50.0, 100.0)
primeiro <- c(5, 5, 4, 3, 1)
segundo <- c(3,2,5,4,2)
terceiro <- c (4,5,3,2,2)
quarto <-  c(6,4,0,2,3)
medias <- c (0,0,0,0,0)

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))
```

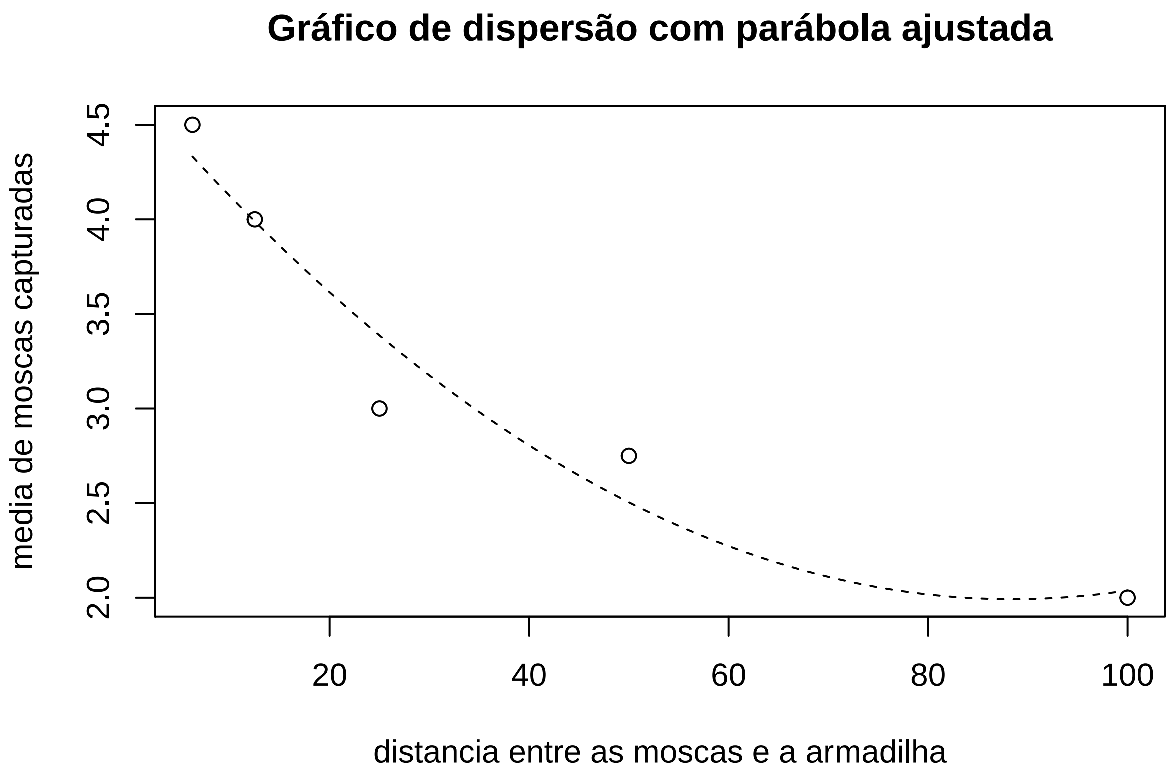


Ao observar o gráfico obtido e a relação entre os pontos e a reta de regressão, é fácil ver sem mais cálculos que a dispersão dos pontos não é linear, e parece aproximar-se de uma parábola. Façamos o gráfico a seguir:

```r
plot(tabela$medias~tabela$distancia, xlab = "distancia entre as moscas e a armadilha",
     ylab = "media de moscas capturadas", main="Gráfico de dispersão com parábola ajustada")
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,
            16,9,9,11,13,15,13,10,11,6,17,14,19,9,11,14,10,16,10,16,14,
            13,13,9,15,10,11,12,4,14,20)
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,
              0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
```

```
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 asteriscos sabemos a importância do valor estimado pois quanto menor esse valor (e maior a quantidade de asteriscos indicativos), maior a chance de rejeitar a hipótese nula sendo assim o valor estimado correspondente (ao intercepto ou ao escore) é mais importante. 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
```

5a

O conjunto de dados Duncan é uma tabela com 45 observações de 45 profissões diferentes relacionadas a quatro variáveis:

- *type (tipo)* - tipo do trabalho, dividido em três grupos: profissional e gerente; blue collar (trabalhos manuais); white collar (trabalhos de escritórios)
  - *income (renda)* - varia de 0 a 100 e corresponde à porcentagem de homens nessa profissão que recebiam mais de $3500 em 1950;

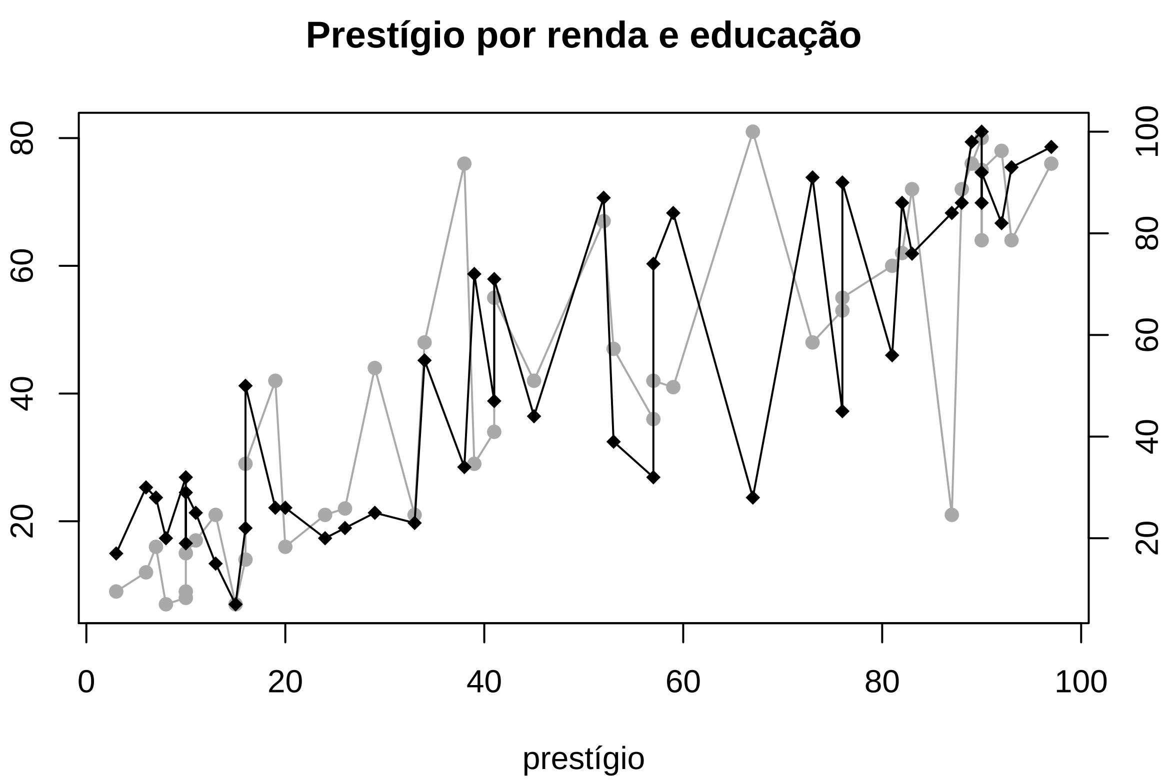
    - *education (educação)* - varia de 0 a 100 e corresponde à porcentagem de homens nessa profissão que tinham ensino médio completo;

      - *prestige (prestígio)* - também varia de 0 a 100 e apresenta a porcentagem de "bom" e "excelente" que esta profissão recebeu no estudo NORC (um estudo no qual voluntários ranqueam as profissões com base em características como renda estimada, liberdade de escolha e o quão interessante é o trabalho).

5b

```
dados <- dados[order(dados$prestige, decreasing=TRUE),]

plot(dados$income~dados$prestige, type="o", col="darkgray", ylab=" ",xlab="prestígio",
     pch=16, main = "Prestígio por renda e educação")
par(new=TRUE)
plot(dados$education~dados$prestige, axes=FALSE, type="o", col="black", ann=FALSE, pch=18)
axis(4)
```



No gráfico acima, procuramos a *renda* (em cinza escuro, com escala à esquerda) e a *educação* (preto, com escala à direita) como indicadores de *prestígio*. Como esperado, notamos uma forte associação (profissões que exigem maior educação e que pagam melhor são consideradas de maior prestígio), mas podemos observar algumas profissões que têm alto prestígio mesmo com baixa renda ou baixo nível de educação.

Para isso, vamos observar as 15 profissões com maior prestígio:

```
dados[1:16,]
```

```
##               type income education prestige
## physician     prof     76        97       97
## professor     prof     64        93       93
## banker        prof     78        82       92
## architect     prof     75        92       90
## chemist       prof     64        86       90
## dentist       prof     80       100       90
## lawyer        prof     76        98       89
## engineer      prof     72        86       88
## minister      prof     21        84       87
## pilot         prof     72        76       83
## accountant    prof     62        86       82
## factory.owner prof     60        56       81
## author        prof     55        90       76
## contractor    prof     53        45       76
## teacher       prof     48        91       73
## RR.engineer     bc     81        28       67
```

A partir desta tabela, podemos notar que a profissão "ministro" tem renda baixa (apenas 21% dos profissionais de sexo masculino recebiam mais de $3500), comparado a outras profissões de mesmo prestígio, e que donos de fábrica e empreiteiro mesmo com menor educação (56% e 45% respectivamente, com ensino médio concluído) têm renda e prestígio similares às de profissões que contam com 90% dos profissionais tendo ensino superior completo.

Vamos analisar agora as o prestígio das profissões segundo o tipo:

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit
```

```
describeBy (dados$prestige, group= dados$type, na.rm=TRUE, skew=FALSE)
```

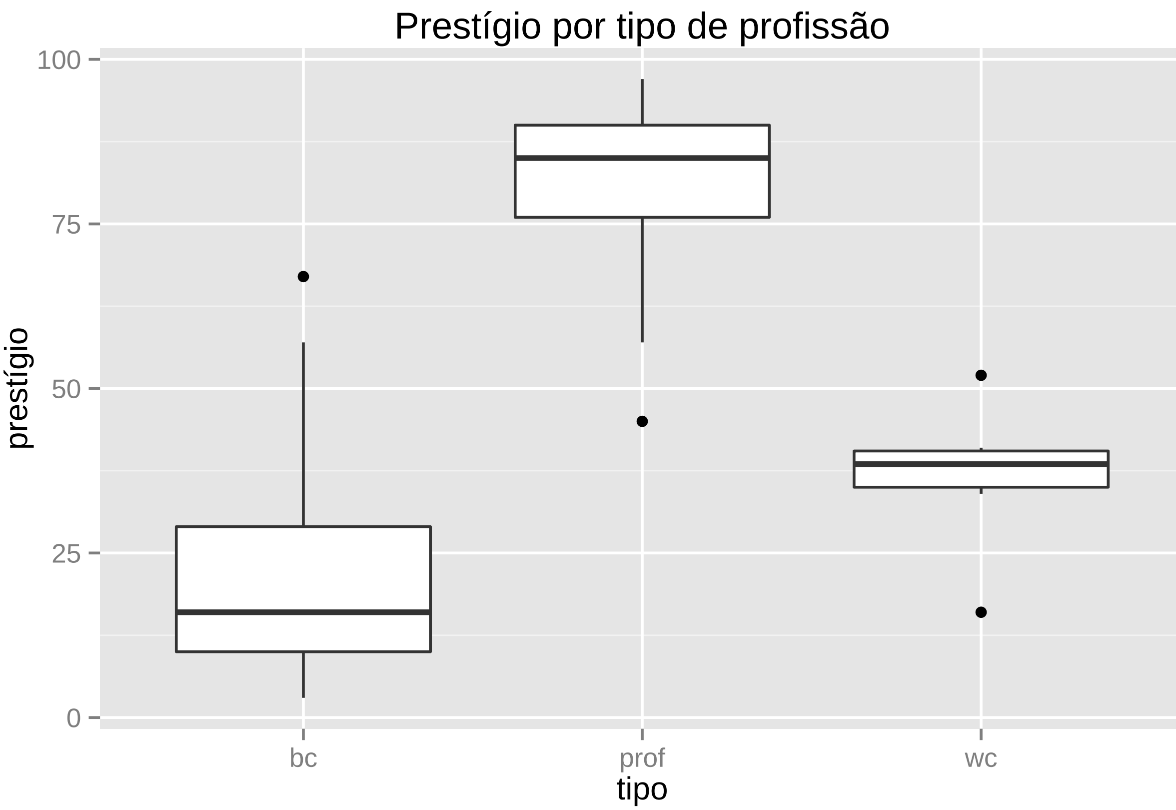
```
##
##  Descriptive statistics by group
## group: bc
##    vars  n  mean    sd min max range   se
## X1    1 21 22.76 18.06   3  67    64 3.94
## ---------------------------------------------------------
## group: prof
##    vars  n  mean    sd min max range   se
## X1    1 18 80.44 14.11  45  97    52 3.32
## ---------------------------------------------------------
## group: wc
##    vars n  mean    sd min max range   se
## X1    1 6 36.67 11.79  16  52    36 4.81
```

```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:psych':
##
##     %+%
```

```r
ggplot(dados, aes(x=as.factor(type),
                  y=prestige))+
  geom_boxplot()+
  labs(x="tipo", y="prestígio")+
  ggtitle("Prestígio por tipo de profissão")
```



Como esperado, os trabalhos manuais têm menos prestígio, mas também são os que apresentam maior variação. Os trabalhos com maior prestígios são os mais profissionalizados. Os trabalhos de escritório apresentam menor variação, estando quase todos em torno de 32% de prestígio, exceto por dois outliers: repórter, com 52% de prestígio, e balconista, com 16%. Dentre os trabalhos profissionalizados, gerente de loja é oúnico outlier, apresentando 52% de prestígio, e dentre os trabalhadores manuais, o outlier é engenheiro de trem (maquinista) com 67% de prestígio.

5c

```r
library(pander)
fit1 <- lm(dados$prestige~dados$education+dados$income+dados$type)
modelo3 <- summary(fit1)
mod3_coef <- modelo3$coefficients
```

```r
colnames(mod3_coef) <- c("Estimativa", "erro padrão", "valor t", "p-valor")
rownames(mod3_coef) <- c("Intercepto", "education", "income", "typeprof", "typewc")
pander(mod3_coef)
```

	Estimativa	erro padrão	valor t	p-valor
Intercepto	-0.185	3.714	-0.04982	0.9605
education	0.3453	0.1136	3.04	0.004164
income	0.5975	0.08936	6.687	5.124e-08
typeprof	16.66	6.993	2.382	0.02206
typewc	-14.66	6.109	-2.4	0.02114

Tabela de medidas resumo

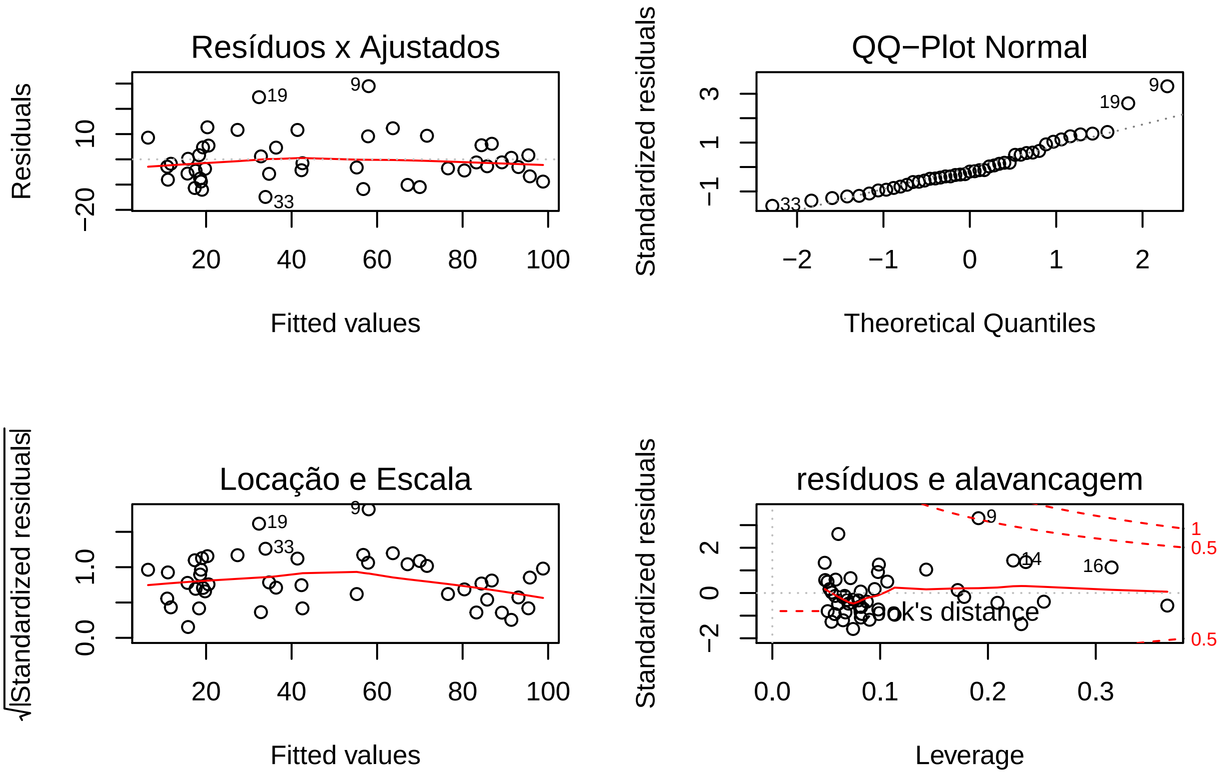
Minimo	1Q	Mediana	3Q	Máximo
-14.890	-5.740	-1.754	5.442	28.972

erro padrão residual: 9.744 R quadrado multiplo: 0.9131 R quadrado ajustado: 0.9044 Estatística F: 105 em 4 com 40 graus de liberdade, p-valor: < 2.2e-16

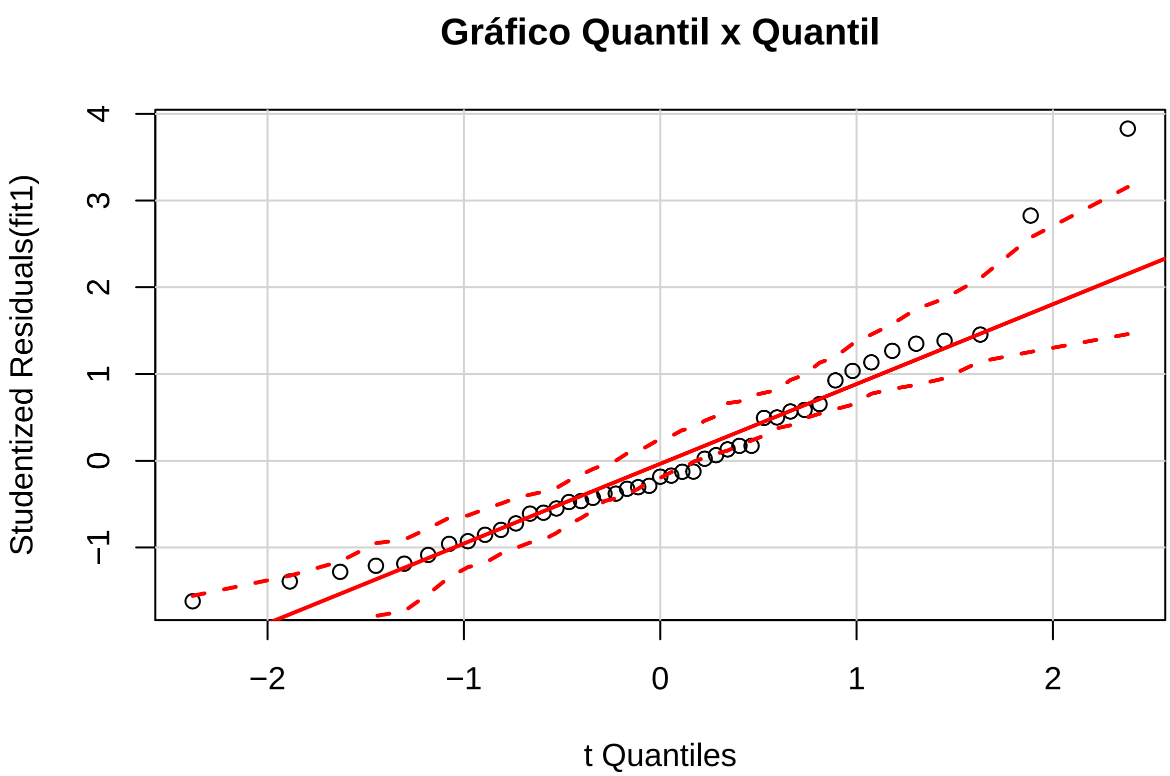
Observando os dados obtidos na tabela de coeficientes, com o enfoque no p-valor, podemos dizer que todos os valores estimados, com exceção do intercepto, são significativo, pois os p-valores são baixos, ao contrario do valor obtido no intercepto que com 0,96051 temos uma grande chance de aceitar a hipótese, ou seja, o intercepto ser igual a zero. Além disso, o R quadrado múltiplo foi igual à 0.9131, bem próximo de 1, que significa que o modelo foi bem ajustado, sendo que 90% dos dados podem ser explicada pelo modelo.

5d

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




```r
qqPlot(fit1 ,main = "Gráfico Quantil x Quantil")
```



Pela analise de resíduos não existe nenhuma tendência aparente, então a variabilidade parece constante, mostrando uma homocedasticidade e pelos gráficos qqplots, ambos possuem a maior parte dos dados proximos da reta $Y = X$ com apenas alguns outliares, com isso concluimos que o modelo se aproxima de uma normal.

5e

O modelo pode sim ser utilizado, pois analisando os resíduos, encontramos p-valores baixos, além de um elevado valor do R quadrado que nos mostra através dos dados que o modelo é bem ajustado, o que se observa também pelos gráficos plotados, os quais pelo gráfico de resíduos concluimos uma homocedastidade dos dados e pelos qqplots uma aproximação do modelo para uma normal. Assim, podemos dizer que em 1950, o prestígio pode ser explicado pela profissão, pela renda e pela escolaridade. Porém, devemos notar que os dados são antigos, então não é aconselhável utilizar esses resultados para tentar explicar um préstigo nos dias atuais, mas é possível utilizá-lo como base para uma nova modelagem.