Matéria de Análise de Dados – Ciências Biomédicas

Aula 11 – Regressão 2

James Hunter

24 de março de 2017

Uma revisão rápida da última aula para começar. Regressão linear é a técnica básica entre os modelos estatísticos. Esta vez, usaremos um outro exemplo. Mostrarei todos os passos de preparação e analise deste dataset simples.

R tem uma gama variada de datasets para uso de teste e ensino. Um é chamado `women` e contem as alturas e pesos de 15 mulheres. Nós vamos ver se podemos construir um modelo linear que mostrará uma relação entre altura e peso e se altura pode prever o peso das mulheres. Lógico, este modelo é 1) trivial e 2) super-simplificado, mas oferece uma visão clara de como construir um modelo linear.

Passo 1: Preparar os Dados

Neste caso, traduzirei os dados em português e medidas métricas usando o pacote `dplyr`.

```
# entrada de dados acontece diretamente porque `women` fica dentro de R
data(women)
kable(women)
```

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

```
str(women) #mostrar a estrutura de data.frame
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

A comanda `transmute` cria novas variáveis e apaga as originais (diferente do `mutate` que não apaga as originais). Colocarei o resultado num dataset novo `mulheres`.

```
mulheres <- transmute(women, alturacm = round(height*2.54, 2),
                      pesokg = round(weight/2.2, 2))
kable(mulheres)
```

alturacm	pesokg
147.32	52.27
149.86	53.18
152.40	54.55
154.94	55.91
157.48	57.27
160.02	58.64
162.56	60.00
165.10	61.36
167.64	63.18
170.18	64.55
172.72	66.36
175.26	68.18
177.80	70.00
180.34	72.27
182.88	74.55

```
str(mulheres)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ alturacm: num  147 150 152 155 157 ...
##  $ pesokg  : num  52.3 53.2 54.5 55.9 57.3 ...
```

Agora, estamos prontos para segunda etapa, a analise exploratória dos dados.

Passo 2: Explorar dos Dados

Daremos uma olhada nos dados especialmente para determinar se podemos usar regressão linear para construir um modelo explicativo e preditivo.

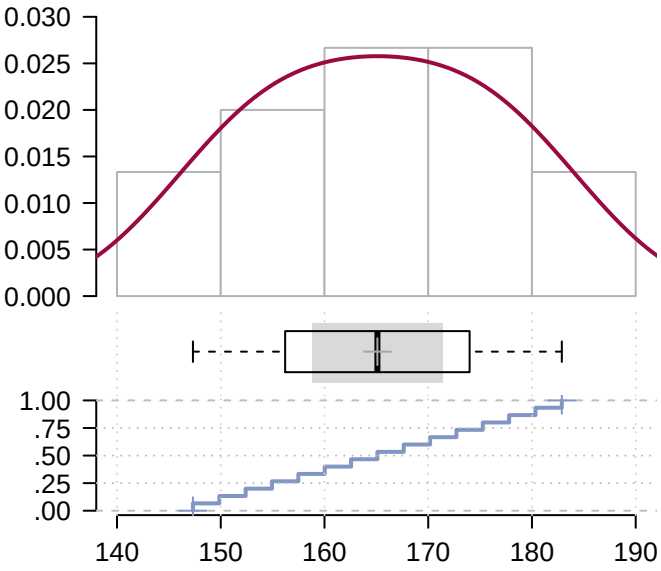
Usarei o pacote `DescTools` cujo ferramenta principal `Desc` é muito mais completa que a função `summary` no sistema base de R. `Desc` também prepara gráficos úteis que mostram a distribuição das variáveis. Além de `Desc`, vou apresentar uma plotagem das 2 variáveis, um scatterplot ('gráfico de dispersão') de um terceiro pacote, `car`. Este scatterplot também mostra uma linha de regressão simples e uma linha suavizada para podemos ver qual tipo de regressão serve melhor para analise desta relação.

```
Desc(mulheres, plotit = TRUE)
```

```
## ------------------------------------------------------------------------
## Describe mulheres (data.frame):
##
## data.frame:  15 obs. of  2 variables
##
##   Nr  ColName   Class    NAs  Levels
##   1   alturacm  numeric  .
##   2   pesokg    numeric  .
##
##
```

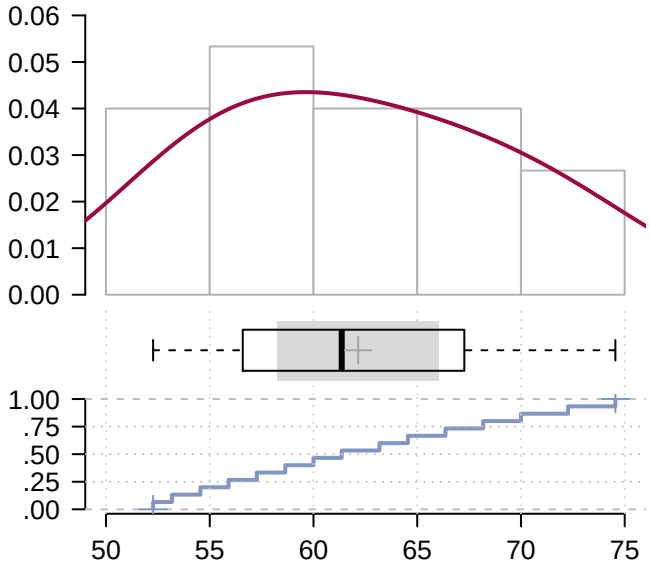
```
## ------------------------------------------------------------------------
## 1 - alturacm (numeric)
##
##    length        n      NAs   unique       0s       mean   meanCI
##        15       15        0      = n        0    165.100  158.809
##             100.0%     0.0%             0.0%            171.391
##
##       .05      .10      .25   median      .75       .90      .95
##   149.098  150.876  156.210  165.100  173.990   179.324  181.102
##
##     range       sd    vcoef      mad      IQR      skew     kurt
##    35.560   11.359    0.069   15.063   17.780  1.696e-15   -1.441
##
## lowest : 147.32, 149.86, 152.4, 154.94, 157.48
## highest: 172.72, 175.26, 177.8, 180.34, 182.88
```

1 – alturacm (numeric)




```
## ------------------------------------------------------------------------
## 2 - pesokg (numeric)
##
##   length       n     NAs  unique      0s    mean  meanCI
##       15      15       0     = n       0  62.151  58.250
##           100.0%    0.0%            0.0%          66.053
##
##      .05     .10     .25  median     .75     .90     .95
##   52.907  53.728  56.590  61.360  67.270  71.362  72.954
##
##    range      sd   vcoef     mad     IQR    skew    kurt
##   22.280   7.045   0.113   8.080  10.680   0.228  -1.344
##
## lowest : 52.27, 53.18, 54.55, 55.91, 57.27
## highest: 66.36, 68.18, 70.0, 72.27, 74.55
```

2 − pesokg (numeric)



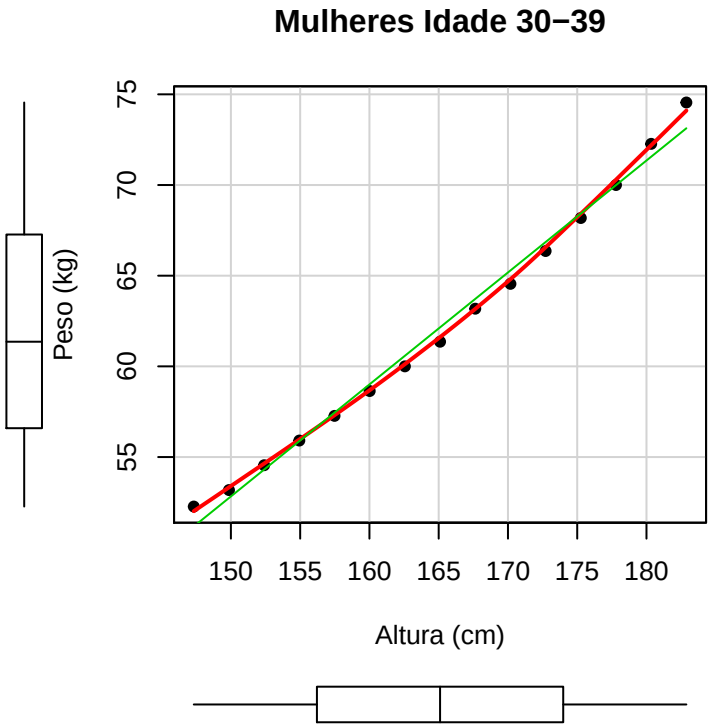

```
altnorm <- unlist(shapiro.test(mulheres$alturacm)[2])
pesonorm <- unlist(shapiro.test(mulheres$pesokg)[2])
paste("Normalidade de altura per Shapiro-Wilks (valor-p):",
      round(altnorm,3))
```

```
## [1] "Normalidade de altura per Shapiro-Wilks (valor-p): 0.755"
```

```
paste("Normalidade de peso per Shapiro-Wilks (valor-p):",
      round(pesonorm,3))
```

```
## [1] "Normalidade de peso per Shapiro-Wilks (valor-p): 0.7"
```

```
scatterplot(pesokg ~ alturacm,
            data=mulheres,
            spread=FALSE, smooth = TRUE,
            pch=19,
            main="Mulheres Idade 30-39",
            xlab="Altura (cm)",
            ylab="Peso (kg)")
```



Os dados parecem de ser normalmente distribuídos. Também, tem um formato quase linear apesar a curva suavizada faz um fit um pouco melhor. Esta analise prova que vale a pena de construir um regressão linear. Mais tarde, podemos testar outras regressões que ficam mais perto a linha vermelha no scatterplot.

Passo 3 - Regressão Simples Linear

Usamos a função `lm` para construir uma regressão linear. Esta função tem muitas opções, mas vamos usar a forma mais simples. (Muitas das opções tem a ver com o tratamento das matrizes usadas no cálculo do modelo.) Especificaremos uma formula que fica no formato de `<resposta> ~ <termos>`, ou seja, com a variável dependente antes do til e a variável (ou variáveis) independente depois. O segundo parâmetro seria o nome de dataframe ou tibble que contem os dados. Nós salvamos o resultado do modelo num objeto de R para que possamos estudar todos os detalhes do modelo. A função `summary` resuma os resultados principais do modelo e a função `anova` lista a tabela ANOVA para o modelo. Ao final, usamos `plot` para iniciar nossa verificação do fit do modelo. Este plotagem consiste de 4 gráficos. Vamos conversar sobre a informação nesses gráficos na próxima aula.

```
fit <- lm(pesokg ~ alturacm, data = mulheres)
summary(fit)
```

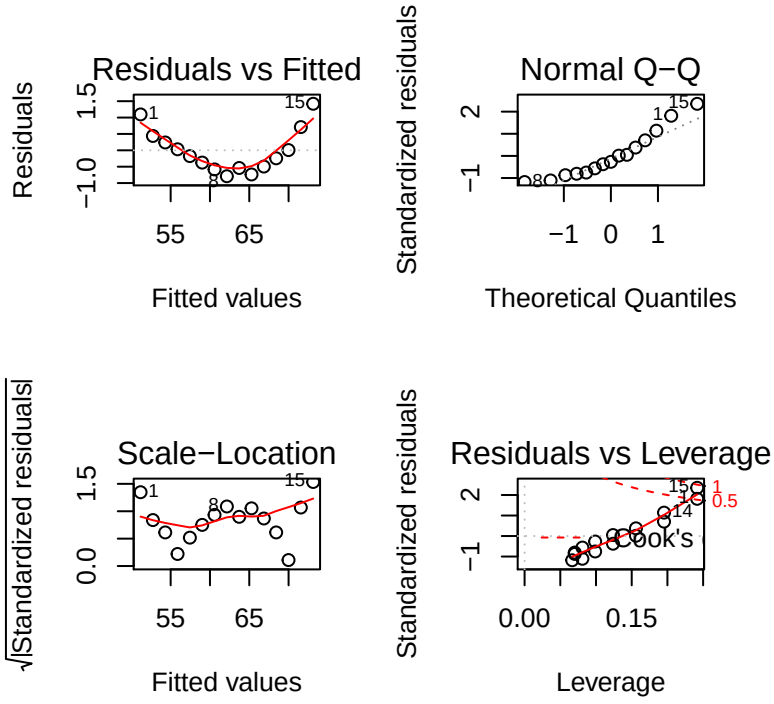
```
##
## Call:
## lm(formula = pesokg ~ alturacm, data = mulheres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7913 -0.5178 -0.1767  0.3388  1.4212
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.78260    2.70019  -14.73 1.72e-09 ***
```

```
## alturacm      0.61741    0.01632   37.83 1.10e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6936 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1431 on 1 and 13 DF,  p-value: 1.099e-14
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: pesokg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## alturacm   1 688.60  688.60  1431.4 1.099e-14 ***
## Residuals 13   6.25    0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
plot(fit)
```




```
par(mfrow=c(1,1))
```

Passo 4 - O Que Significa os Resultados

O que é significativo deste modelo? Qual é a relação entre altura e peso? O `summary` do modelo sugere que altura explica quase toda a variância na variável peso. Podemos ver isso em dois elementos do `summary`.

- A estatística F mostra que o modelo como um todo é significativo (F = 1431 com 1 e 13 graus de liberdade, p-value = $<1.0 \times 10^{-}6$). [Qualquer valor p menor que isso não pode ser diferenciado de 0. Então, a prática normal é usar este valor como limite.]

- O coeficiente de determinação ("R-squared") mostra que a variável independente expressa 99.1% da variância na variável peso.

O modelo diz que para cada centímetro de altura adicional que uma mulher cresce, pode esperar um aumento do peso de 0.61741 quilos. O intercepto diz que se uma mulher tem 0 cm de altura, ele vai pesar -39.8 kg. Neste caso, este tem significado nenhum, mas na maioria dos modelos, tem utilidade.

A tabela ANOVA também mostra que a soma de quadrados da regressão (SSR) é 688.60 e a SSE (soma relacionada a erros) é 6.25. Este valor será divido por o número de graus de liberdade dos residuais (13), que cria uma média (MSE) de só 0.48.

Mas, você acredita que altura explica 99% ou mais da variância no peso das mulheres? Talvez, temos um caso em que existe vários fatores de "confounding", onde temos um modelo que parece válido, mas existem outras variáveis escondidas que influenciam os resultados.

Passo 5 - Apuração das Premissas do Modelo

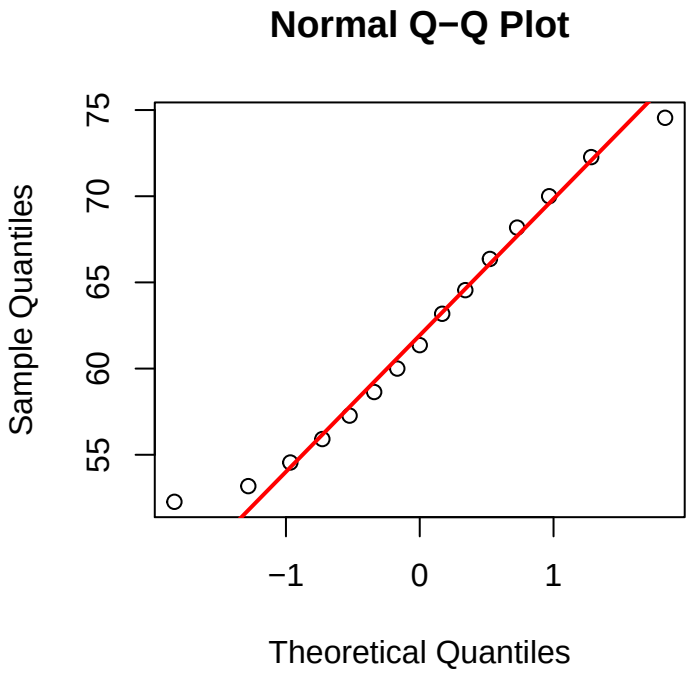
Vamos visar as premissas que o modelo deve satisfazer para ser válido. Nesta apuração, os quatro gráficos oferece bastante ajuda.

Normalidade

Como sabemos o modelo deve ser construído das variáveis que têm distribuições normais. O gráfico Q-Q (a direto em cima dos quatro) mostra isso. O gráfico Q-Q mede os residuais padronizados contra quantis teóricos de uma distribuição normal padronizada (média = 0, dp = 1). Um modelo perfeito terá os pontos diretamente em cima da linha. Neste caso, no meio da distribuição, os valores ficam bem perto a linha de igualdade, mas na cauda direto (especialmente casos 1 e 15), eles ficam acima da linha. Este replica o que aprendemos na apuração das variáveis individuais em Passo 2 e não apresentam grandes problemas.

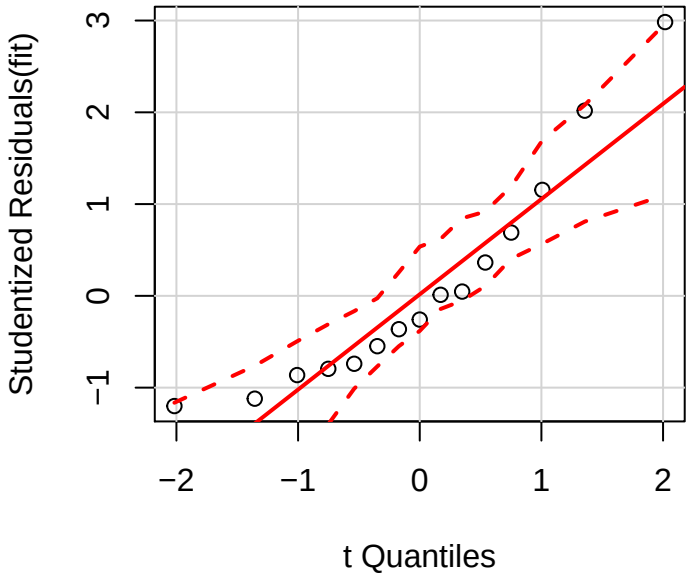
Podemos também construir o gráfico Q-Q diretamente com as comandas `qqnorm` e `qqline`. Neste caso, não usamos o resultado `fit`, mas a variável dependente original. O sistema fará os cálculos necessários para determinar os pontos.

```
qqnorm(mulheres$pesokg)
qqline(mulheres$pesokg, col = 2, lwd = 2)
```



Uma terceira maneira de construir um gráfico Q-Q é de usar a função `qqPlot` no pacote `car`. Esta versão da função inclui um intervalo de confiança que mostra só os pontos extremos ficam fora da linha e do IC. Pode anotar que esta versão do gráfico coloca todos os valores em termos da distribuição `t` invés de estritamente normal, lembrando que esta família de distribuições tem as mesma características básicas de uma distribuição normal, mas com caudas mais grossas. A palavra "studentized" é uma referência a William Gosset, que assinou o artigo que primeiro descreveu esta distribuição com o pseudônimo de "Student".

```
car::qqPlot(fit)
```



Linearidade

A variável dependente e a variável independente devem ter uma relação linear. Esta é a premissa básica da regressão linear. O gráfico `Residuals vs Fitted` ajuda na determinação disso. Se for linear, os pontos

de dados seriam numa banda horizontal sem outra forma clara. O gráfico só mostraria o barulho aleatório sobrando depois da variância do modelo de regressão foi acomodado. No caso de `mulheres`, o gráfico indica que eles assumam uma forma quadrática.

Nós podemos lidar com esse problema com a adição de um termo independente levantando `altura` para quadrado. Na próxima versão do modelo, usaremos este termo.

Homoscedasticidade

Além de ter uma forma normal, os dados também tem que ter variâncias iguais aos todos os níveis da variável independente. Nós usamos o gráfico `Scale - Location` para determinar isso. Aqui estamos procurando também uma banda horizontal dos pontos sem uma direção clara. A linha aqui mostra que apesar de ser perto de uma tendência clara, a linha cabe dentro dos parâmetros desta premissa.

O quarto gráfico, `Residuals vs Leverage` identifica **outliers** e pontos que exercem uma influência muito grande no resultado. Aprenderemos a interpretação dessas fatores na próxima aula.

Opção de Regressão Polinomial

Para tomar em conta a presença de uma distribuição claramente não-linear indicada no gráficos `Residuals vs Fitted` e scatterplot dos dados, podemos utilizar um outro tipo de regressão linear, chamado regressão *polynomial*. Neste caso, nós aumentamos ao modelo um termo que expressa a variável independente levantada ao quadrado.

Para fazer isso, precisamos usar a função `I()`. O operador `^` tem significado na elaboração das formulas. Portanto, precisamos uma função que inibe este comportamento e força R para reconhecer `^` no significado tradicional como simbolo para levantar um número para um poder. Este é o papel da função `I()`.

Nossa formula assim torna: `peso ~ altura + I(altura^2)`. A regressão é polinomial porque é equivalente a formula $y = x + x^2$, que lembramos da álgebra como uma equação polinomial. Nós veremos que este modelo oferece um fit melhor para os dados e cumpre mais apropriadamente as premissas da regressão.

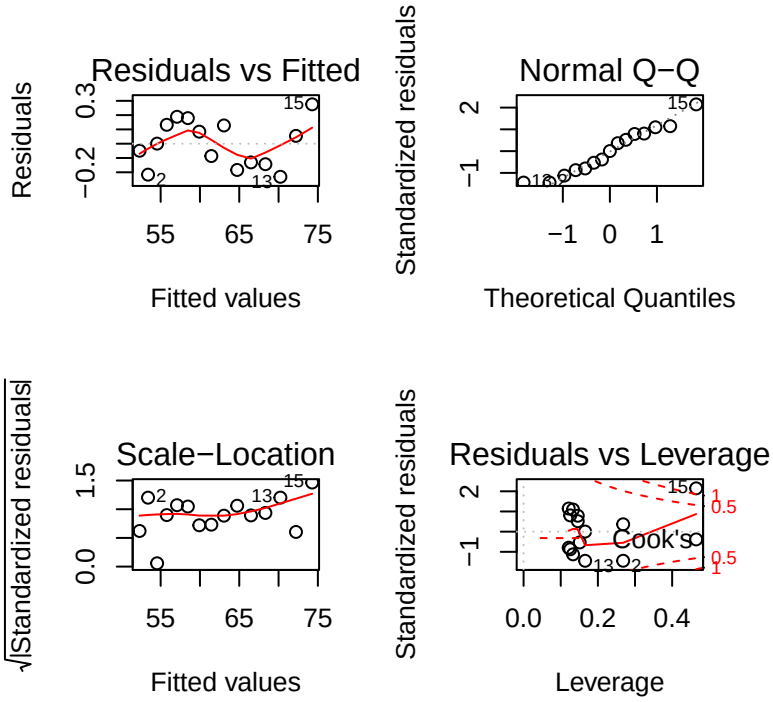
```
fitpoly <- lm(pesokg ~ alturacm + I(alturacm^2), data = mulheres)
summary(fitpoly)
```

```
##
## Call:
## lm(formula = pesokg ~ alturacm + I(alturacm^2), data = mulheres)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.231604 -0.137179  0.000539  0.129965  0.275529
##
## Coefficients:
##                  Estimate  Std. Error t value      Pr(>|t|)
## (Intercept)   119.0833006  11.5095658  10.346 0.00000024757 ***
## alturacm       -1.3156140   0.1398583  -9.407 0.00000069060 ***
## I(alturacm^2)   0.0058541   0.0004234  13.827 0.00000000982 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1754 on 12 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 1.128e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

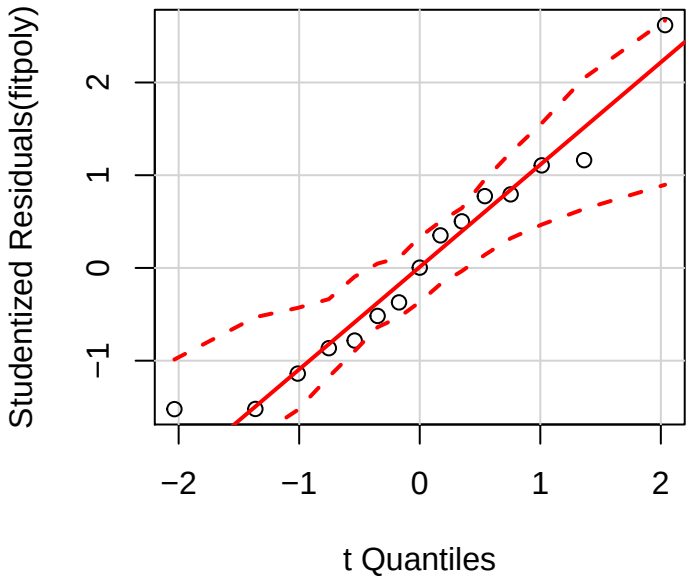
```r
anova(fitpoly)
```

```
## Analysis of Variance Table
##
## Response: pesokg
##                Df Sum Sq Mean Sq  F value         Pr(>F)
## alturacm        1 688.60  688.60 22373.36      < 2.2e-16 ***
## I(alturacm^2)   1   5.88    5.88   191.19 0.000000009822 ***
## Residuals      12   0.37    0.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(2,2))
plot(fitpoly)
```




```r
par(mfrow=c(1,1))
qqPlot(fitpoly)
```



Agora, podemos ver que explicamos mais um pouco da variância dos dados e os dois termos independentes ficam significativos. Também, o gráfico problemático, de `Residuals vs Fitted` não mostra nenhuma tendência quadrática e podemos concluir que este modelo descreve bem a relação entre as variáveis.

Próxima Aula - Regressão Múltipla e Apuração dos Modelos Mais Detalhadas

Na próxima aula, vamos estender o conceito da regressão para incluir múltiplas variáveis independentes e aprofundar nosso entendimento da interpretação e apuração dos modelos de regressão.