# Regressão Linear

Prática

# Agenda

- Lembretes
- Laboratório

# Lembrete - Regressão Linear

# Lembrete - Teoria

---

Prever Y utilizando **variáveis (uma ou mais)**

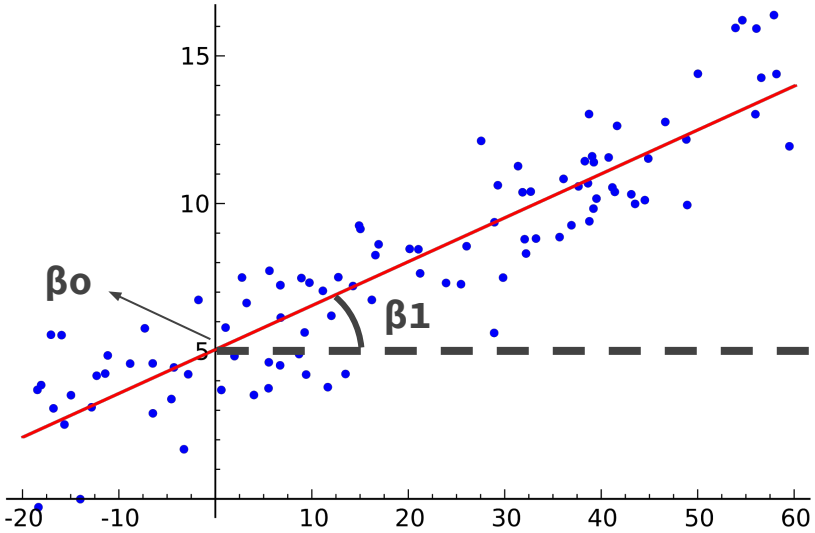
variáveis

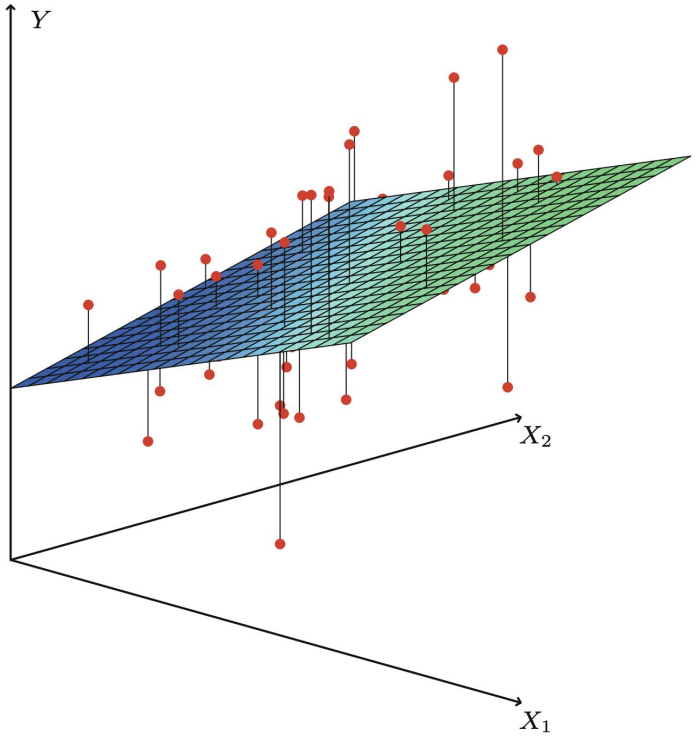
termo de erro

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Preço da casa $= \beta_0 + \beta_1 \times$ Área $+ \beta_2 \times$ Quartos $+ \beta_3 \times$ Vagas de garagem $+ \epsilon$

# Lembrete - Resultado



# Lembrete - Premissas

---

***Premissa aditiva***

O efeito da variável Xj em Y é independente dos valores das outras variáveis.

***Premissa linear***

A mudança no valor de Y em relação ao aumento em uma unidade em Xj é constante

# Significância dos coeficientes

---

Média de y

Predição

$$F = \frac{\left(\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2\right)/p}{\sum(y_i - \hat{y}_i)^2/(n - p - 1)}$$

Real

Variáveis

O que isso significa?
**Se não há relação entre Y e as variáveis, F≈1**

Se **F>1**, pode-se esperar que **há relação** entre as variáveis e Y!

Com o valor de F, estimamos o p-valor.

Quanto **menor** o p-valor, **maior** a chance de se rejeitar H0!

# Qualidade do modelo

$$\text{RSE} = \sqrt{\frac{1}{n - p - 1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Predição

Variáveis

Bom modelo: RSE **pequeno!**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Valor real

Valor estimado

Valor real

Média de y

O que isso significa?
**O quanto os valores de Y podem ser explicados por X**

Quanto maior o R2, mais *explicabilidade* tem o seu modelo

# Laboratório: Regressão Linear Simples

# Bibliotecas utilizadas

```
> library(MASS) #datasets + funções

> library(ISLR) #datasets do livro


> install.packages("ISLR") #para instalar
```

# Ajuste

```
> attach(Boston)

> lm.fit = lm(medv~lstat)
                Y     X

> lm.fit

Call:
lm(formula = medv ~ lstat)

Coefficients:
(Intercept)        lstat
      34.55        -0.95
```

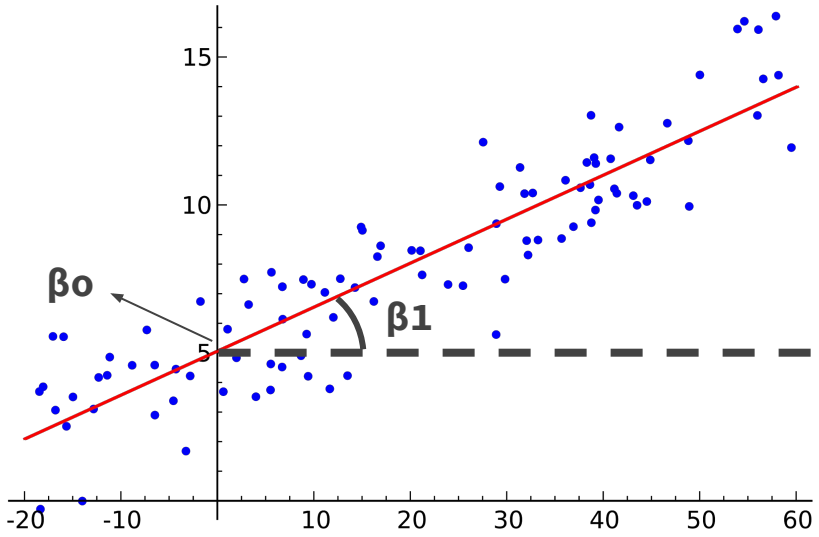
$\beta_0$ $\beta_1$

Utilizaremos o dataset Boston, com registro de medianas de valor **(medv)** de preços de casas em bairros de Boston.

rm = número médio de quartos por casa

age = idade das casas

lstat = % de moradias populares



# Sumário

```
> summary (lm.fit)

Call:
lm(formula = medv ~ lstat)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

RSE: 6.216 (quanto menor, melhor)

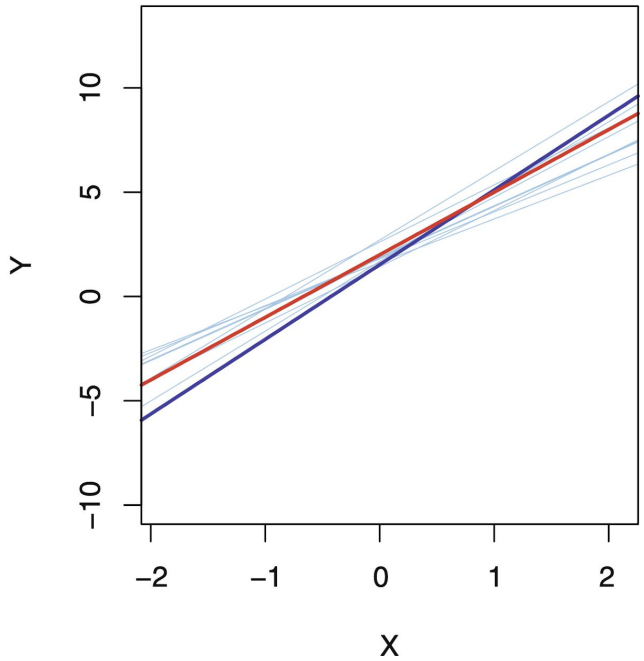
R2: 0.5432 (quanto maior, melhor)

F-estatística: 601.6 (quanto maior, melhor)

**Diferença entre os R2: o adjusted adiciona penalidades para o aumento no número de variáveis**

# Intervalos de confiança

```
> confint(lm.fit)

                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```



# Predições

```
>#predizendo medv para os valores de 5%, 10% e 15% de moradias populares

> predict (lm.fit ,data.frame(lstat=c(5,10 ,15)), interval="confidence")
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461


> predict (lm.fit ,data.frame(lstat=c(5,10 ,15)), interval="prediction")
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

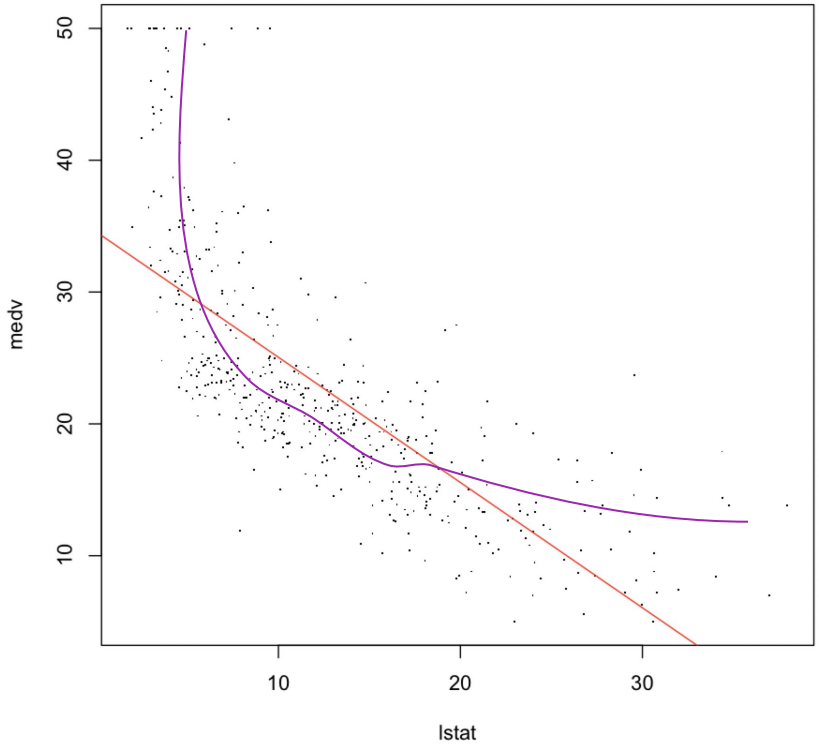
Intervalo de confiança: variação do medv em relação a amostras da população analisada

Intervalo de predição: variação do medv em relação a um ponto específico **(não visto antes)**

Intervalo de predição é sempre **maior** que o intervalo de confiança pelo erro irredutível do **novo ponto** diferir da amostra

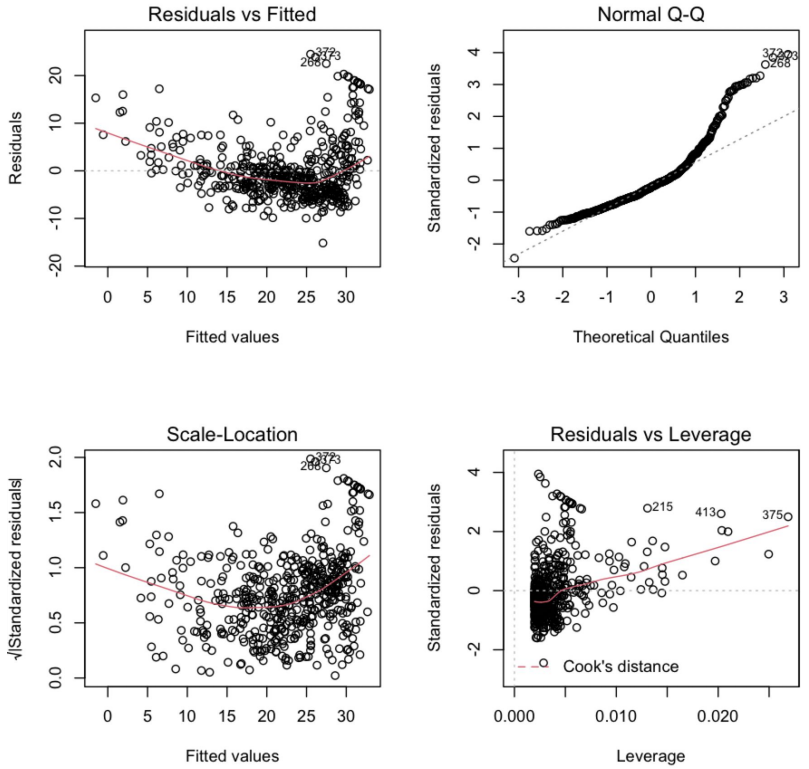
# Premissas atendidas?

```
> plot(lstat, medv, pch=".")

> abline(lm.fit, col="red")

> #sinais de não linearidade entre os dados!!

> #precisamos fazer transformações nos nossos dados!!
```

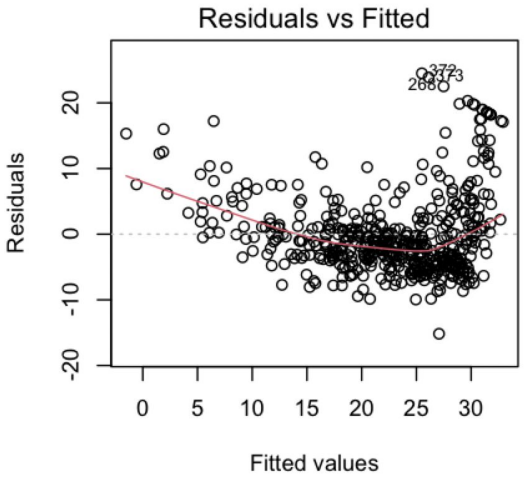


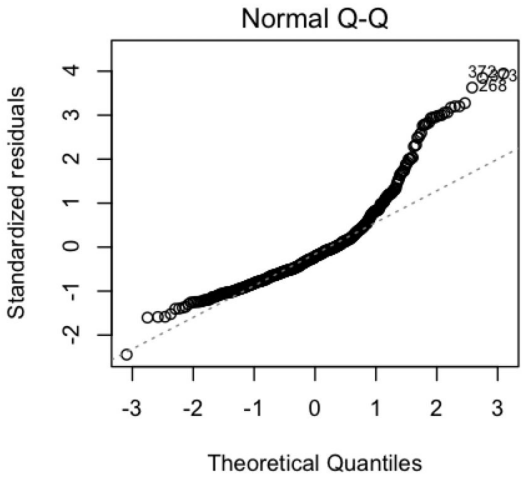
# Gráficos de diagnóstico

```
> #cria um grid 2x2 de imagens

> par(mfrow=c(2,2))

> #retorna gráficos de diagnóstico

> plot(lm.fit)
```



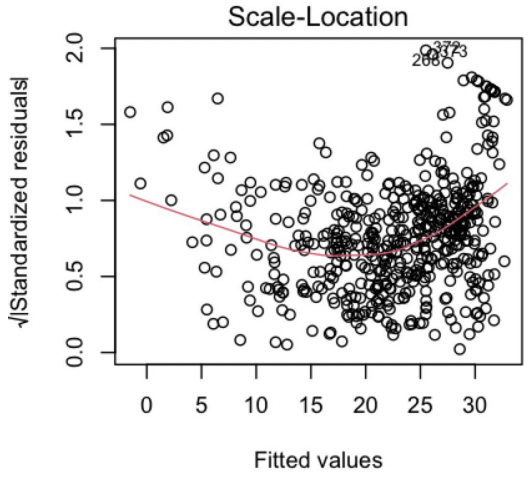
Padrão de resíduos
**Ideal: linha**

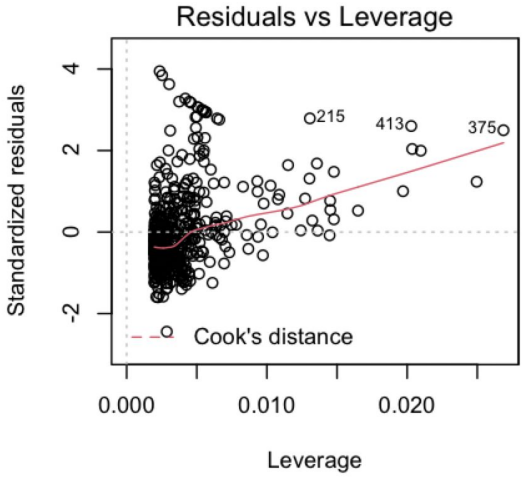



Verificar se os resíduos são normalmente distribuídos
**Ideal: linha**

Distribuição de resíduos pela variedade de preditores
**Ideal: sem padrão**




Identificar pontos de grande influência
**Ideal: Remover pontos**

# Laboratório: Regressão Linear Múltipla

# Ajuste - 2 variáveis

```
> lm.fit = lm(medv~lstat+age,data=Boston)

> summary(lm.fit)

Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.981  -3.978  -1.283   1.968  23.158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,    Adjusted R-squared:  0.5495
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

| | 1 variável | 2 variáveis |
|---|---|---|
| RSE | 6.216 | 6.173 |
| R2-adj | 0.5432 | 0.5495 |
| F-stat | 601.6 | 309 |

# Ajuste - Todas as variáveis

```
> lm.fit = lm(medv~.,data=Boston)

> summary(lm.fit)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

| | 1 variável | 13 variáveis |
|---|---|---|
| RSE | 6.216 | 4.745 |
| R2-adj | 0.5432 | 0.7338 |
| F-stat | 601.6 | 108.1 |

# Ajuste - Retirando variáveis

```
> lm.fit = lm(medv~.-age,data=Boston)

> summary(lm.fit)

Call:
lm(formula = medv ~ . - age, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6054  -2.7313  -0.5188   1.7601  26.2243

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.436927   5.080119   7.172 2.72e-12 ***
crim         -0.108006   0.032832  -3.290 0.001075 **
zn            0.046334   0.013613   3.404 0.000719 ***
indus         0.020562   0.061433   0.335 0.737989
chas          2.689026   0.859598   3.128 0.001863 **
nox         -17.713540   3.679308  -4.814 1.97e-06 ***
rm            3.814394   0.408480   9.338  < 2e-16 ***
dis          -1.478612   0.190611  -7.757 5.03e-14 ***
rad           0.305786   0.066089   4.627 4.75e-06 ***
tax          -0.012329   0.003755  -3.283 0.001099 **
ptratio      -0.952211   0.130294  -7.308 1.10e-12 ***
black         0.009321   0.002678   3.481 0.000544 ***
lstat        -0.523852   0.047625 -10.999  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.74 on 493 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7343
F-statistic: 117.3 on 12 and 493 DF,  p-value: < 2.2e-16
```

| | 13 variáveis (todas) | 12 variáveis (retirando age) |
|---|---|---|
| RSE | 4.745 | 4.74 |
| R2-adj | 0.7338 | 0.7343 |
| F-stat | 108.1 | 117.3 |

# Inspeção de colinearidade

```
> lm.fit = lm(medv~.,data=Boston)

> library(car)

> vif(lm.fit)
```

```
    crim       zn    indus     chas      nox       rm      age      dis      rad      tax  ptratio    black
1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945 7.484496 9.008554 1.799084 1.348521
   lstat
2.941491
```

Quanto maior o VIF, maior a colinearidade

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

R2 da regressão de Xj pelas outras variáveis

# Termos de interação

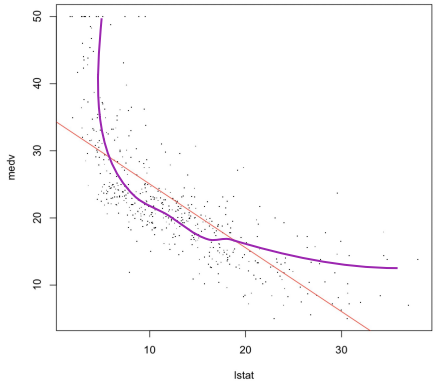
```
> summary(lm(medv~lstat*age,data=Boston))

Call:
lm(formula = medv ~ lstat * age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.806  -4.045  -1.333   2.085  27.552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
age         -0.0007209  0.0198792  -0.036   0.9711
lstat:age    0.0041560  0.0018518   2.244   0.0252 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.149 on 502 degrees of freedom
Multiple R-squared:  0.5557,    Adjusted R-squared:  0.5531
F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

A variável X1*X2 adiciona X1, X2 e X1*X2 na regressão

| | Sem termo de interação | Com termo de interação |
|---|---|---|
| **RSE** | 6.216 | 6.149 |
| **R2-adj** | 0.5432 | 0.5531 |
| **F-stat** | 601.6 | 209.3 |

# Transformações não lineares

```
> lm.fit2=lm(medv~lstat+I(lstat^2))

> summary(lm.fit2)

Call:
lm(formula = medv ~ lstat + I(lstat^2))

Residuals:
     Min       1Q   Median       3Q      Max
-15.2834  -3.8313  -0.5295   2.3095  25.4148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.524 on 503 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6393
F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```




| | Sem transformação não linear | Com transformação não linear |
|---|---|---|
| **RSE** | 6.216 | 5.524 |
| **R2-adj** | 0.5432 | 0.6407 |
| **F-stat** | 601.6 | 448.5 |

# Comparando modelos

```
> lm.fit=lm(medv~lstat)

> lm.fit2=lm(medv~lstat+I(lstat^2))

> anova(lm.fit ,lm.fit2)
```

```
Analysis of Variance Table

Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A função anova() realiza um teste de hipótese comparando 2 ou mais modelos.

Hipótese nula: modelos equivalentes

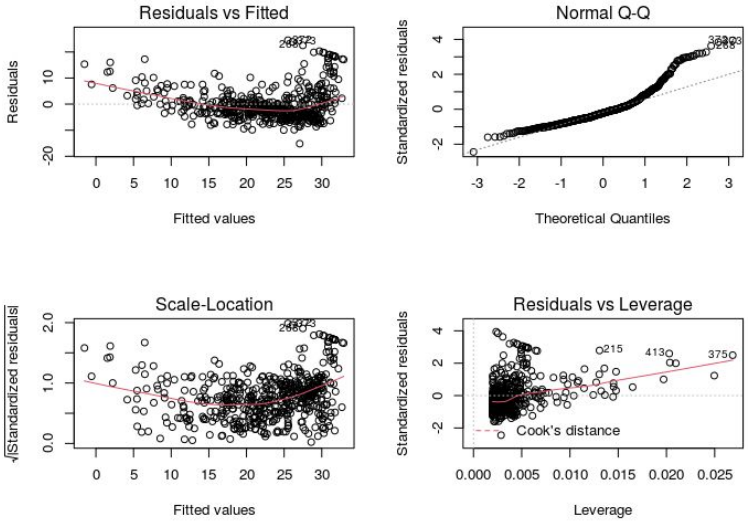
Hipótese alternativa: um modelo é melhor

plot(lm.fit2)

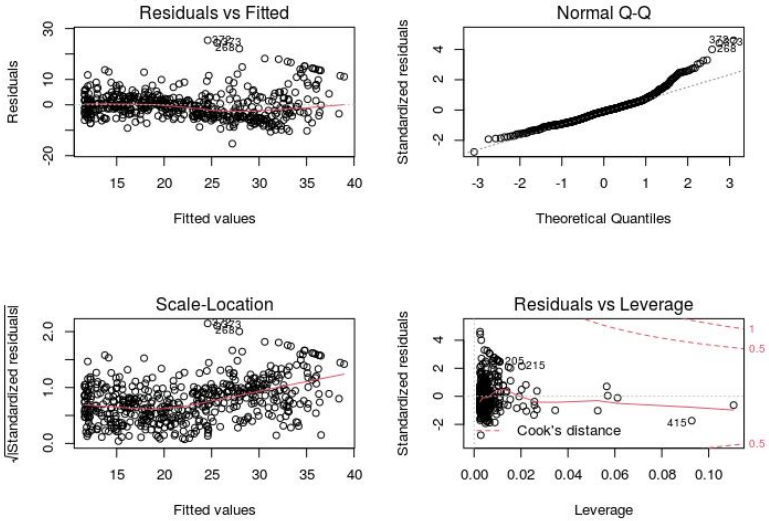
Quanto maior a F-stat, menor o p-valor, maior a certeza em rejeitar a hipótese nula!

# Comparando modelos

Sem tranformação




Com tranformação



# (mais) Transformações não lineares

```
> lm.fit5=lm(medv~poly(lstat ,5))

> summary(lm.fit5)

Call:
lm(formula = medv ~ poly(lstat, 5))

Residuals:
     Min       1Q   Median       3Q      Max
-13.5433  -3.1039  -0.7052   2.0844  27.1153

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.215 on 500 degrees of freedom
Multiple R-squared:  0.6817,    Adjusted R-squared:  0.6785
F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

| | Com 1 transformação não linear | Com 5 transformações não lineares |
|---|---|---|
| **RSE** | 5.524 | 5.215 |
| **R2-adj** | 0.6407 | 0.6785 |
| **F-stat** | 448.5 | 214.2 |