Regressão linear múltipla

Atsler Luana Lehun

Lidiany Doreto Cavalcanti

- É uma extensão da regressão linear simples ➔ usada para prever uma variável resposta (y) com base em várias variáveis preditoras distintas (x)

$$Y = b0 + b1x1 + b2x2 + b3x3 + e$$

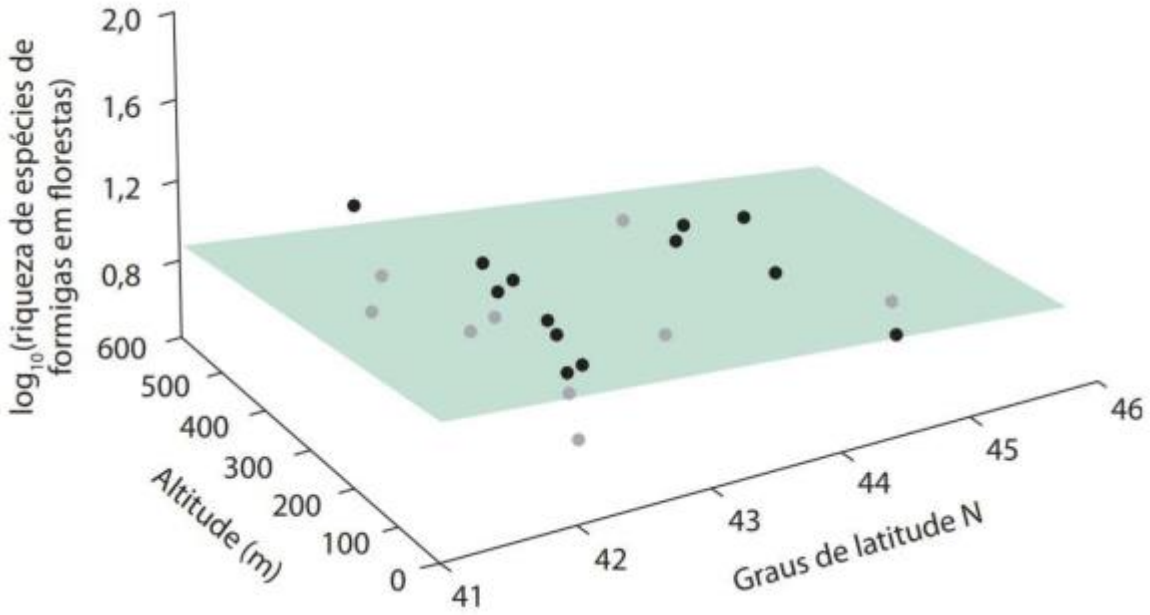
Variável resposta — Preditor 1 — Preditor 2 — Preditor 3

- Os valores de b ➔ são coeficientes da regressão que mostram a relação entre o Y e o X correspondente
- e ➔ erro

Calculados a partir dos resíduos

Reta de ajuste x plano de ajuste

- Assim como a regressão linear simples, o ajuste do modelo da regressão linear múltipla também é calculado a partir do método dos mínimos quadrados
  - Ao invés de uma reta que se ajusta à duas variáveis teremos um plano



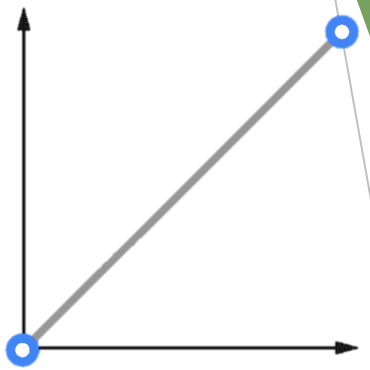
O plano também é calculado para que minimize a proporção de resíduos não explicada pela relação entre as variáveis

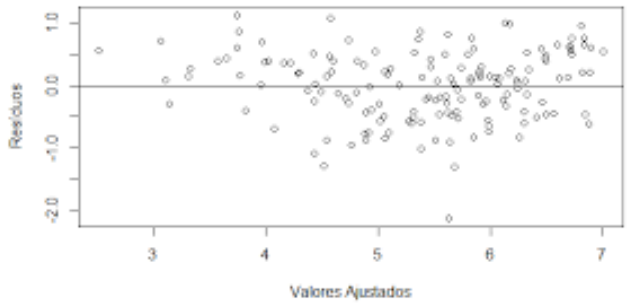
- Os resíduos são calculados a distancia vertical de cada dado até o plano predito

- Assim como a regressão linear simples teremos:
  - Estimativa dos parâmetros do modelo pelo **método dos mínimos quadrados**
  - R^2
  - Estatística F para testar a significância do modelo geral
  - As variâncias do erro
  - Intervalo de confiança
  - Teste de hipóteses para cada um dos coeficientes (b1,b2...)

Pressupostos

- **Linearidade**
- **Normalidade (Shapiro-Wilk do modelo)**
- **Homocedasticidade (inspeção visual dos resíduos)**
- **Multicolinearidade**





Multicolinearidade

- Quando duas ou mais variáveis preditoras são correlacionadas entre si

- Qual o problema?
  - Quando acontece correlação entre as variáveis dificulta a separação das contribuições de cada variável para a variável resposta
  - Matematicamente ➔ as estimativas dos mínimos quadrados ficam instáveis e difíceis de calcular

O que fazer então?

- Testes para verificar as variáveis correlacionadas e retirá-las
- Escolher de acordo com critérios ecológicos
- Combinar matematicamente um conjunto de variáveis preditoras correlacionadas em um número menor de variáveis usando métodos multivariados (eixos da PCA)

Testes no R para multicolinearidade

- VIF (fator de inflação de variação) ➔ mede quanto a variação de um coeficiente de regressão é inflada devido à multicolinearidade no modelo
  - O menor valor é 1 (ausência de multicolinearidade)
  - Regra geral: um VIF maior que 5 ou 10, tirar a variável

- Teste de correlação
- Visualização gráfica

Métodos de seleção de modelos para regressão múltipla

O que fazer então?

- O critério de informação de Akaike (AIC) pode ser uma saída para seleção de modelos!

- AIC ➔ Uma medida do poder explicativo de um modelo estatístico que considera o número de parâmetros do modelo

- Critério de parcimônia ➔ menor número de variáveis que tem o mesmo poder se explicação

O que fazer então?

- AIC ➔ quanto maior o AIC, a inclusão/exclusão da co-variável não melhora a qualidade do modelo

- Vamos sempre considerar os modelos que são significativos ($p<0,05$) e com menor valor de AIC (delta AIC <2)

Exemplo de regressão linear múltipla

- Instalar o pacote "*datarium*" do R
- Vamos usar a planilha Marketing ➔ contém os gastos com mídias para cada rede social

	youtube	facebook	newspaper	sales
1	276.12	45.36	83.04	26.52
2	53.40	47.16	54.12	12.48
3	20.64	55.08	83.16	11.16
4	181.80	49.56	70.20	22.20
5	216.96	12.96	70.08	15.48
6	10.44	58.68	90.00	8.64
7	69.00	39.36	28.20	14.16
8	144.24	23.52	13.92	15.84
9	10.32	2.52	1.20	5.76
10	239.76	3.12	25.44	12.72

- Nós precisamos construir um modelo assim:

$$sales = b0 + b1*youtube + b2*facebook + b3*newspaper$$

	youtube	facebook	newspaper	sales
1	276.12	45.36	83.04	26.52
2	53.40	47.16	54.12	12.48
3	20.64	55.08	83.16	11.16
4	181.80	49.56	70.20	22.20
5	216.96	12.96	70.08	15.48
6	10.44	58.68	90.00	8.64
7	69.00	39.36	28.20	14.16
8	144.24	23.52	13.92	15.84
9	10.32	2.52	1.20	5.76
10	239.76	3.12	25.44	12.72

Construindo o modelo...

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

1º Observar o valor de F e p

O p significativo indica que pelo menos uma das variáveis preditoras tem influência sobre o Y

Qual variável preditora tem relação?

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

O gasto com publicidade no jornal, não faz aumentar o número de vendas do produto

Exemplo: para cada aumento de 1 unidade (1000 dólares) no orçamento de publicidade do youtube, mantendo todos os outros preditores constantes, podemos esperar um aumento de 0,045 * 1000 = 45 unidades de vendas, em média.

Como jornal não tem efeito, podemos tirá-lo!

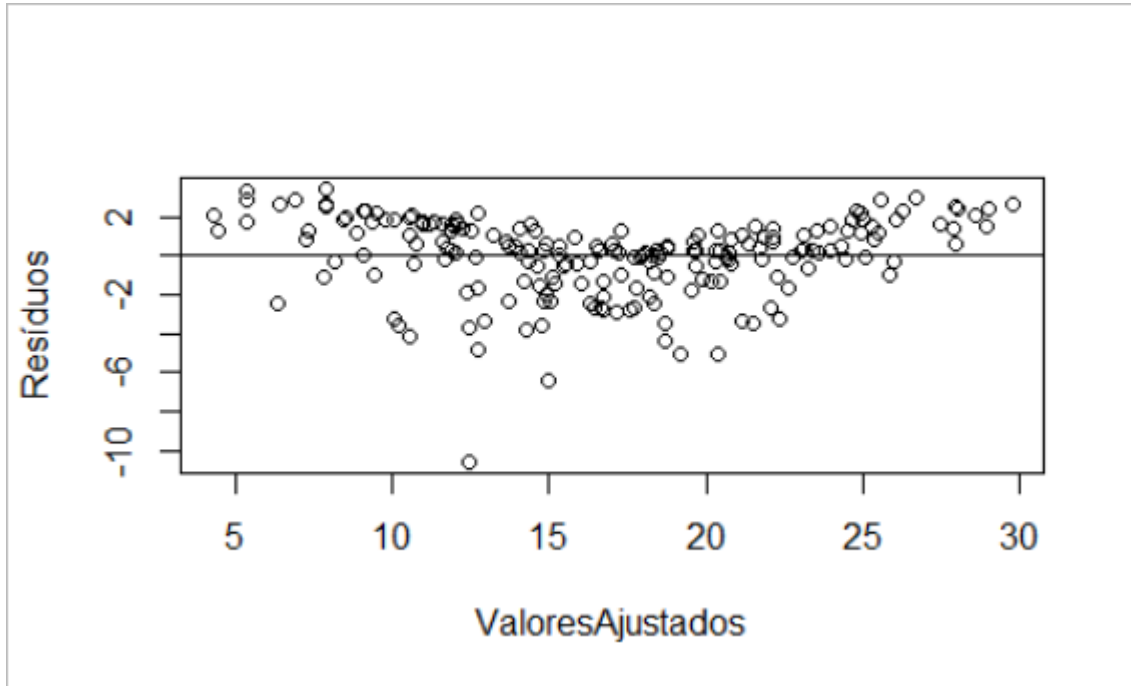
```
Call:
lm(formula = sales ~ youtube + facebook, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5572  -1.0502   0.2906   1.4049   3.3994

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50532    0.35339   9.919   <2e-16 ***
youtube      0.04575    0.00139  32.909   <2e-16 ***
facebook     0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Testar os pressupostos

- Normalidade: Shapiro-Wilks (p>0,05)

```
> shapiro.test(modelo1$residuals) #normalidade

        Shapiro-Wilk normality test

data:  modelo1$residuals
W = 0.91804, p-value = 4.19e-09
```

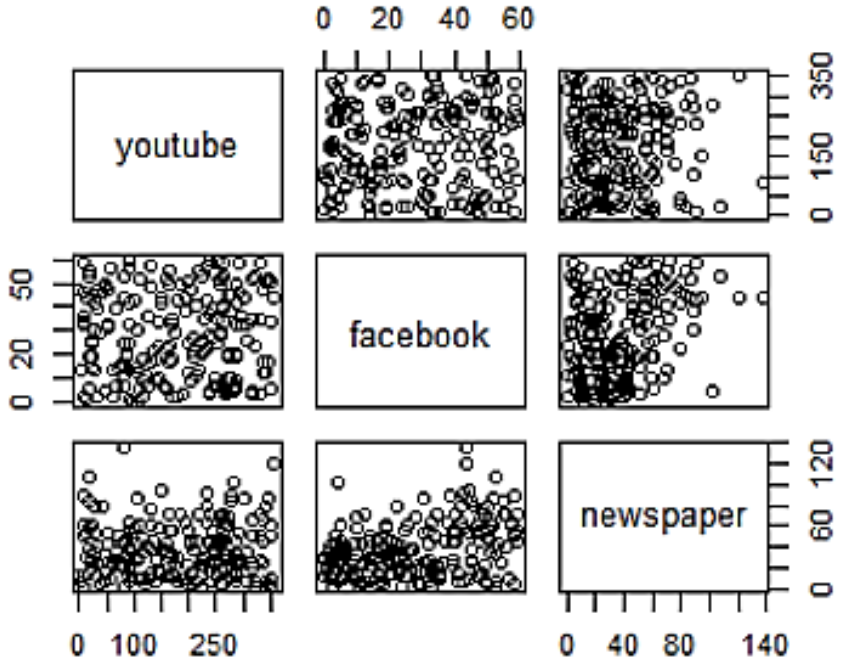
- Homocedasticidade



- Multicolinearidade
  - VIF
  - Correlação
  - Plotar um gráfico

```
> vif(marketing[,-4])
  youtube  facebook newspaper
 1.004611  1.144952  1.145187
```

```
             youtube   facebook  newspaper
youtube   1.00000000 0.05480866 0.05664787
facebook  0.05480866 1.00000000 0.35410375
newspaper 0.05664787 0.35410375 1.00000000
```



Precisão do modelo (R^2)

```
Residual standard error: 2.018 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Significa que 89% da variação na medida de vendas pode ser prevista pelos orçamentos de publicidade do youtube e do facebook

E agora?

Transformamos em LOG

A transformação logarítmica é uma técnica amplamente utilizada para manipular dados e torná-los mais adequados para análise:

Normalizar distribuições,

Reduzir efeitos de outliers,

Facilitar interpretações.

O que é a transformação logarítmica?

consiste em aplicar a função logaritmo aos valores de uma variável.

```
R 4.2.2 · C:/Users/User/Desktop/Disciplina_PEA/
log(facebook + 1)    3.5139     0.1905   18.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.487 on 197 degrees of freedom
Multiple R-squared:  0.8438,    Adjusted R-squared:  0.8422
F-statistic: 532.2 on 2 and 197 DF,  p-value: < 2.2e-16

> shapiro.test(modelo2$residuals) #normalidade

        Shapiro-Wilk normality test

data:  modelo2$residuals
W = 0.84984, p-value = 4.224e-13

>
```

USAR MODELOS COM OUTRA FAMÍLIA DE DISTRIBUIÇÃO GLM

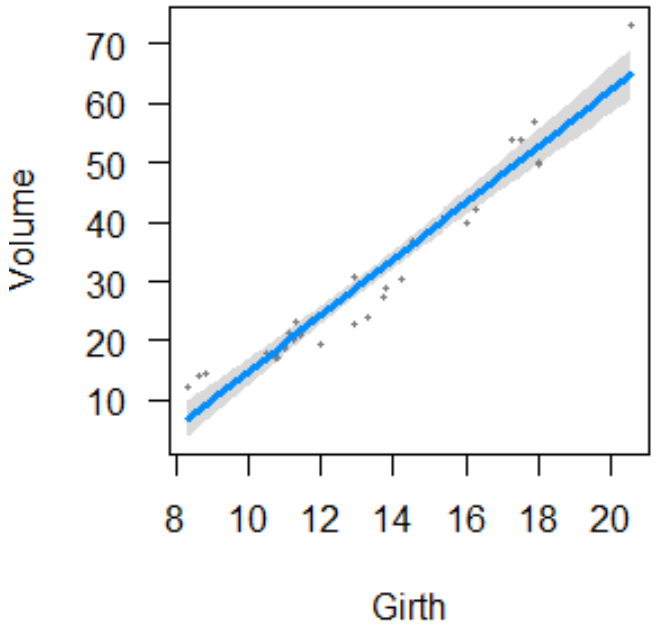
Exemplo 2

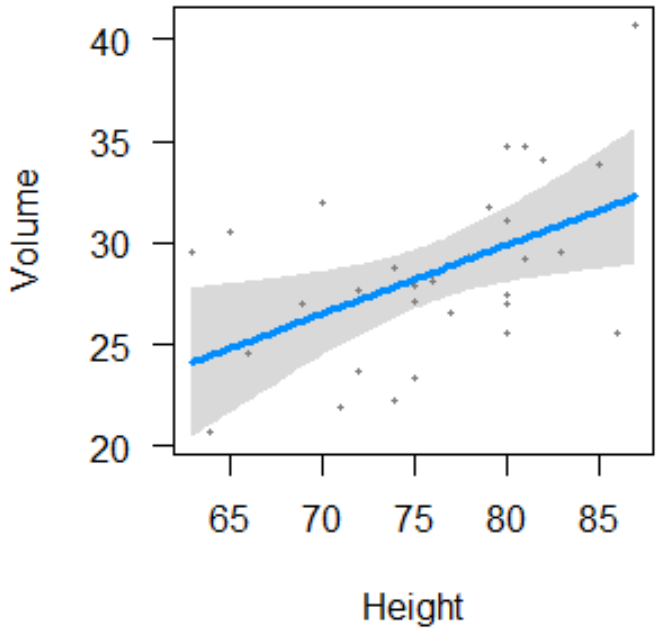
```
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

>
```

Girth (circunferência em polegadas)
Height (altura em pés)
Volume (volume da árvore em pés cúbicos)

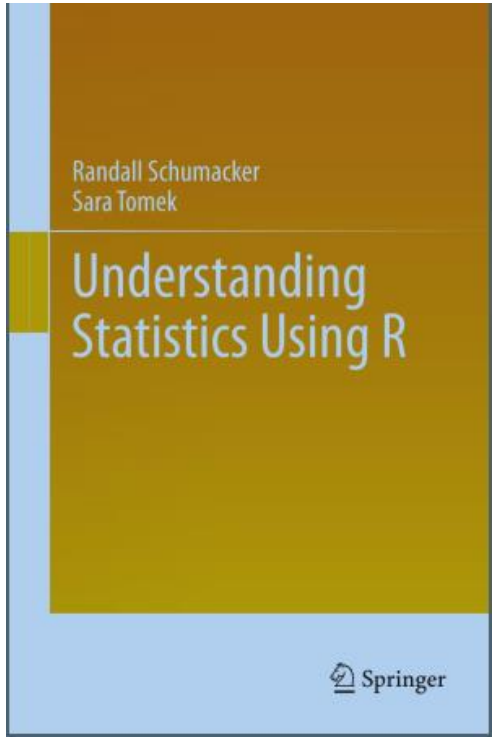
```
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

>
```

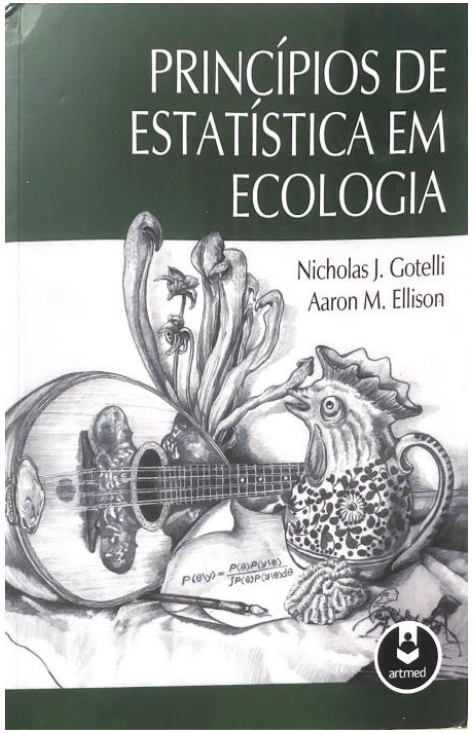
```
Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

> ####pressuposto de normalidade
> #se > 0.05 ? dist. normal
> shapiro.test(modelo3$residuals) #normalidade

        Shapiro-Wilk normality test

data:  modelo3$residuals
W = 0.97431, p-value = 0.644

> |
```


Volume
70
60
50
40
30
20
10
8
10
12
14
16
18
20
Girth


Volume
40
35
30
25
20
65
70
75
80
85
Height

Referências





http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r