Regressão e múltiplas variáveis preditoras

Métodos lineares

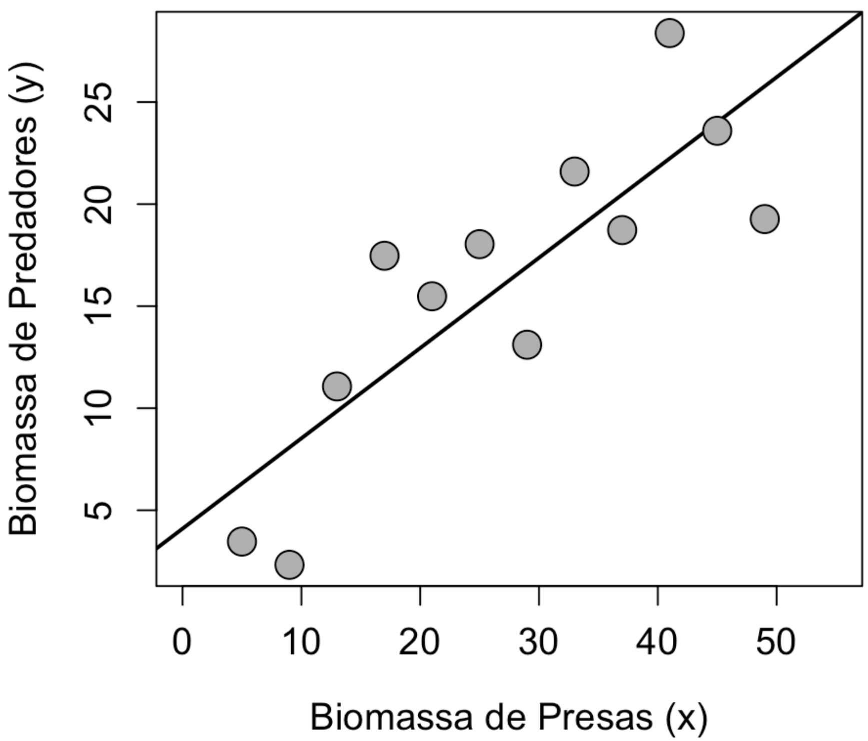
Nicholas A. C. Marino
github.com/nacmarino

Conteúdo da Aula

1. Regressão linear simples
2. Interações entre variáveis preditoras
    - Regressão Múltipla
    - ANOVA n-way
    - Análise de Covariância (ANCOVA) e similares

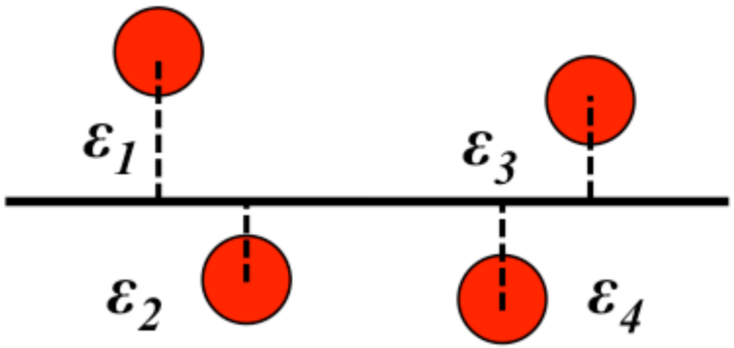
Regressão linear simples

- É uma análise na qual a variação na magnitude dos valores de uma variável **resposta y** é relacionada à variação na magnitude de uma outra **variável preditora X**.
  - Pode ser usada para determinar a forma de uma relação entre duas variáveis; ou,
  - Estimar os parâmetros (βs) de uma equação relacionando a magnitude de y àquela de x.



Pressupostos da regressão linear

1. Relação entre y e x é linear;
   - Relação linear: y ~ βx
   - Relação não linear: y ~ x^{β}
2. Independência espacial, temporal e individual de cada observação (valores de x e y);
3. Homogeneidade das variâncias nos resíduos e ao longo dos valores de x;
4. Normalidade dos resíduos associados à variabilidade em y.



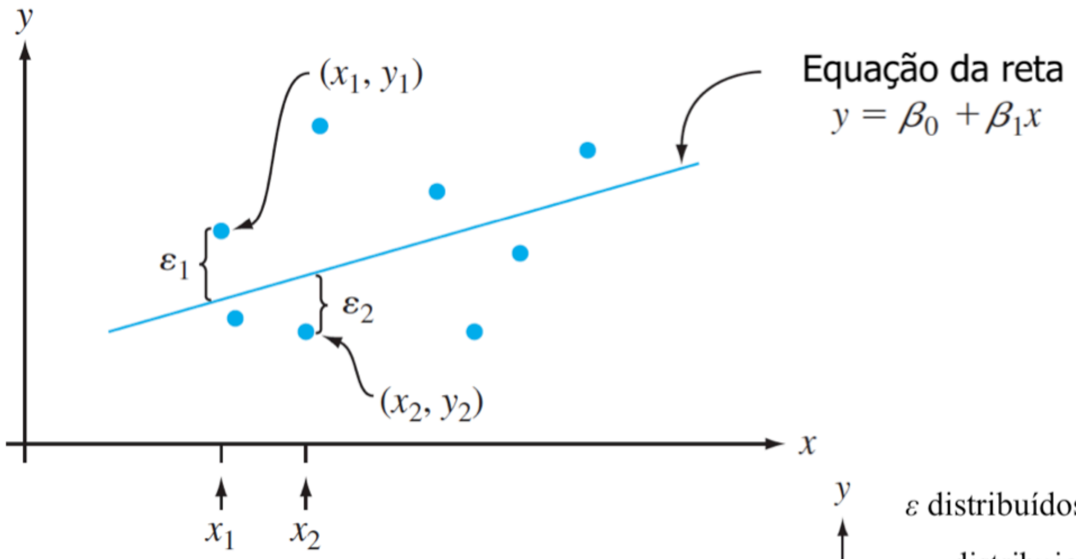
Representação de um modelo linear simples

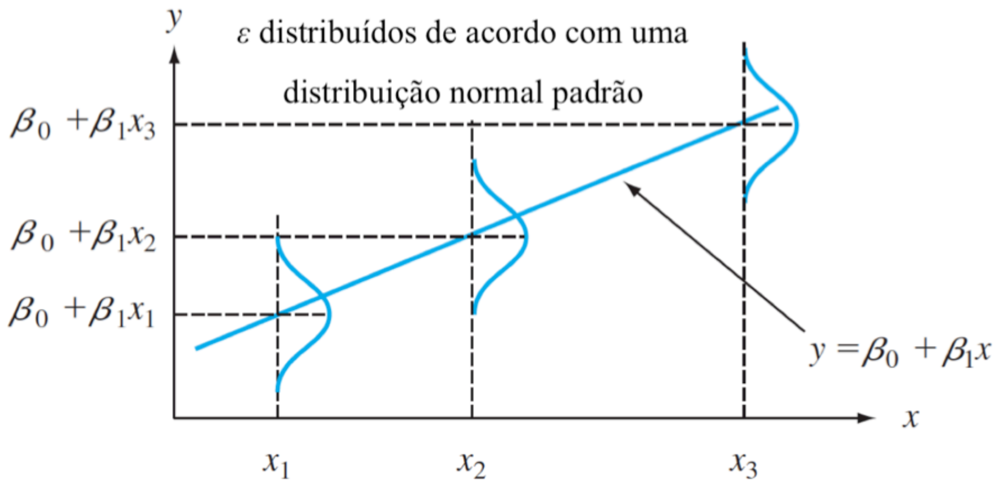
- Um modelo de regressão linear simples pode ser representado como:

$$y = \beta_0 + \beta_1 x + \epsilon$$

  - β_0 é o **intercepto** do modelo - o valor de y quando x é 0;
  - β_1 é a **inclinação** ou **slope** da regressão - a forma pela qual a magnitude de Y muda em função dos valores de x;
  - ϵ é o **erro** na estimativa do valor de Y que não pode ser explicados pela variação nos valores de x.

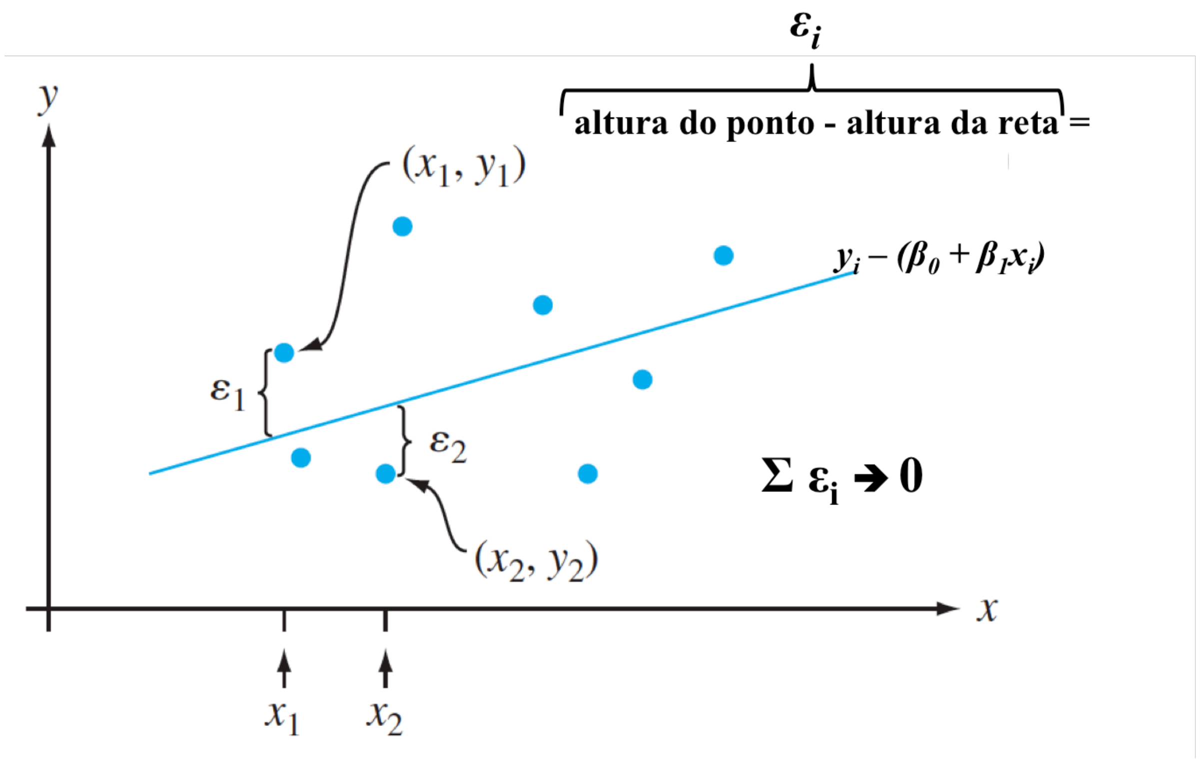
Representação de um modelo linear simples





A mecânica da regressão linear simples

- O princípio geral para determinar a equação da reta na regressão linear é minimizar a distância entre a localização da reta e cada combinação de valores de **x** e **y**.



A mecânica da regressão linear simples

- A inclinação da regressão, β_1, é calculada como a razão entre a **covariância x e y** pela **variância de x**.

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

- Já o intercepto da regressão, β_0, é calculada como a mudança no valor da média *y* a partir da magnitude do efeito de β_1 para o valor da média de *x*; em outras palavras, pela forma como o valor médio de x muda o valor médio de y.

$$\beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

A mecânica da regressão linear simples

- Uma vez que tenhamos os valores de β_0 e β_1, podemos calcular os **valores ajustados** de y para cada valor de x.
- Estes valores ajustados são os novos valores de y_i preditos pelos valores de x de acordo com o modelo estabelecido, sendo representados por $\hat{y}_i$.
- Com base nos desvios entre os valores observados, y_i e os valores preditos pelos modelo, $\hat{y}_i$, podemos calcular a **variação residual** para cada observação, isto é $\epsilon_i = y_i - \hat{y}_i$.
- Se elevarmos estas diferenças ao quadrado e as somarmos, teremos calculado a **Soma dos Quadrados Residual** da regressão:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

A mecânica da regressão linear simples

- De forma similar ao que aprendemos na ANOVA, também podemos calcular a variação total existente nos valores de *y* - a **Soma dos Quadrados Totais**.

$$SST = \sum(y_i - \bar{y})^2 = var(y)$$

- Note então que, com estas duas quantidade, já podemos realizar dos cálculos:
    - A variabilidade em y que pode ser explicada pelos valores de x, isto é a **Soma dos Quadrados da Regressão** (SSR):
        - SST = SSR + SSE (lembre-se da analogia à ANOVA);
    - O coeficiente de determinação da regressão, R^2; e,
    - Um valor de teste estatístico, baseado na distribuição de probabilidade F, que nos ajudará a determinar se a a variabilidade explicada pela regressão é maior do que aquela explicada pela variação residual nos dados.

A mecânica da regressão linear simples

- O coeficiente de determinação da regressão, R^2, descreve o quanto da variabilidade total dos dados é explicada pela regressão.
- Calculado como:

$$R^2 = 1 - \frac{SSE}{SST}$$

- Já o teste estatístico deve considerar a Soma dos Quadrados da Regressão (SSR) e a Soma dos Quadrados Residual (SSE) bem como seus graus de liberdade para a determinação de seus respectivos Quadrados Médios (F = MSR/MSE):
  - Graus de liberdade para SSR = 1, já que estamos estimando apenas 1 parâmetro, o β_1.
  - Graus de liberdade para SSE = n - 2, já que precisamos estimar dois parâmetros para a sua determinação, o β_0 e o β_1.

Exemplo

- Qual a relação entre a área das ilhas e a sua riqueza de espécies? A função `lm` no R pode ser usada para ajudar uma regressão aos dados e responder essa pergunta.

```
modelo <- lm(log(riqueza) ~ log(area), data = ilhas)
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: log(riqueza)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## log(area)  1 13.616 13.6156  30.791 3.804e-07 ***
## Residuals 78 34.491  0.4422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Teste de hipóteses sobre os βs

- Tão importante quanto estimar a significância estatística da regressão, é quantificar β_0 e β_1, bem como a sua variabilidade.
- Podemos calcular a estimativa do erro, S, para o intercepto (β_0) e para a inclinação da reta (β_1) como:

$$S_{\beta_0} = S_{\beta_1} \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$$

$$S_{\beta_1} = \sqrt{\frac{MSE}{var(x)}}$$

Teste de hipóteses sobre os βs

- Os βs seguem uma distribuição t de Student com *n - 2* graus de liberdade, portanto podemos usar o valor de t como o teste estatístico para testar hipóteses sobre os valores de β_0 e/ou β_1.

$$t = \frac{\beta_i - \beta_{H_0}}{S_{\beta_i}}$$

- O método para se obter as estimativas dos β na regressão linear é conhecido como **(Ordinary) Least Squares**, uma vez que ele encontra estas estimativas minimizando a Soma dos Quadrados dos Resíduos entre os valores preditos pelo modelo ($\hat{y}_i$) e os valores originais observados (y_i).

Exemplo

- No R, podemos ter acesso aos coeficientes (βs) e demais informações de interesse sobre o modelo através da função `summary`.

```
summary(modelo)
```

```
##
## Call:
## lm(formula = log(riqueza) ~ log(area), data = ilhas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5732 -0.5259  0.1650  0.5233  1.0441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.44720    0.10361  23.620  < 2e-16 ***
## log(area)    0.13100    0.02361   5.549  3.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.665 on 78 degrees of freedom
## Multiple R-squared:  0.283,  Adjusted R-squared:  0.2738
## F-statistic: 30.79 on 1 and 78 DF,  p-value: 3.804e-07
```

Exercício 1

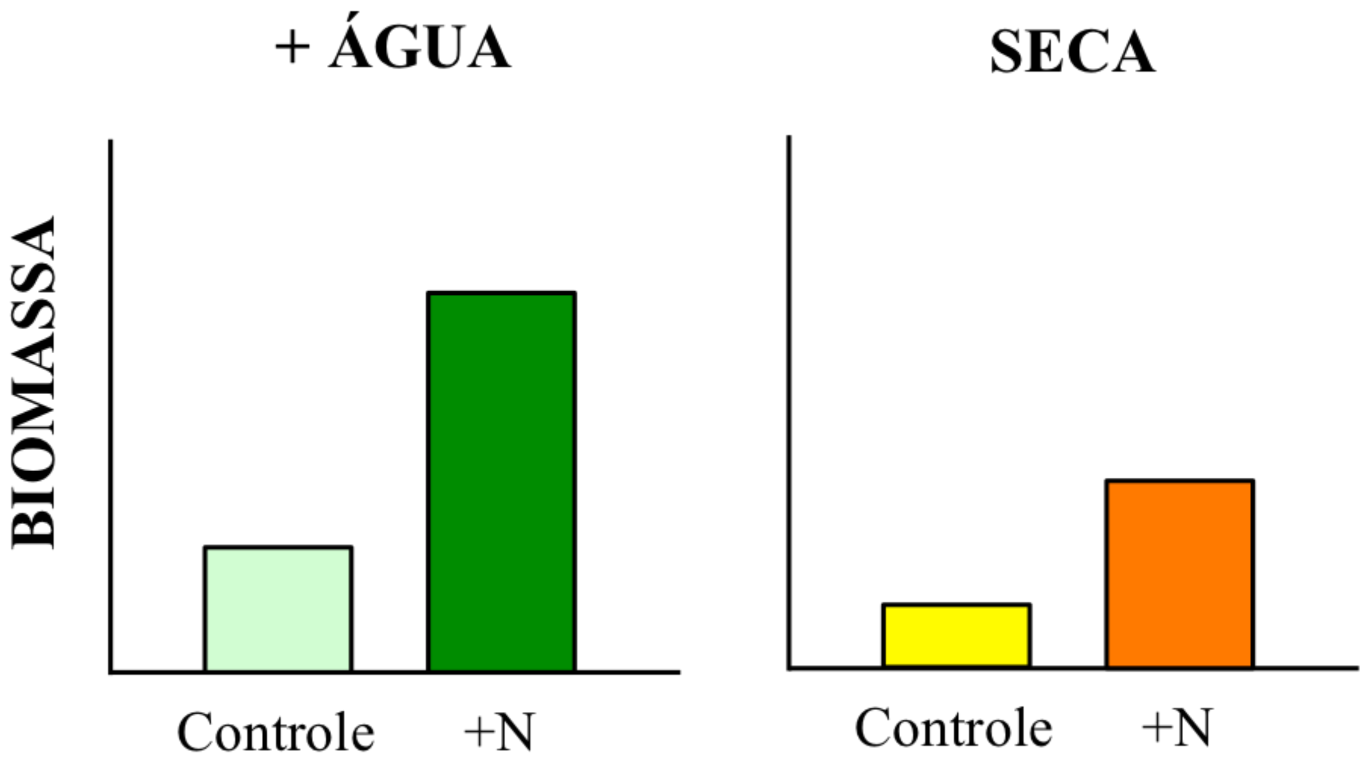
- Em duplas, discutam e interpretem os resultados obtidos através do `summary(modelo)`.
- O que seria um modelo que contenha apenas o intercepto? O que ele representaria?
- Qual a notação de um modelo que contenha apenas o intercepto?

Interações entre variáveis preditoras

- Na maioria das vezes, estamos interessados em determinar qual o efeito da magnitude de mais de uma variável preditora sobre a magnitude da variável resposta. Nesses casos, utilizar um modelo para cada variável resposta não faz sentido, hora porque não te permite testar diretamente a sua hipótese, hora porque pode aumentar a chance de erro do tipo I.
- Em tais casos, devemos criar modelos que contenham:
  - Os efeitos principais de todas as variáveis preditoras de interesse (**main effects**); e/ou,
  - Os efeitos principais de todas as variáveis preditoras e a **interação** entre elas (**main effects** e **interação**).

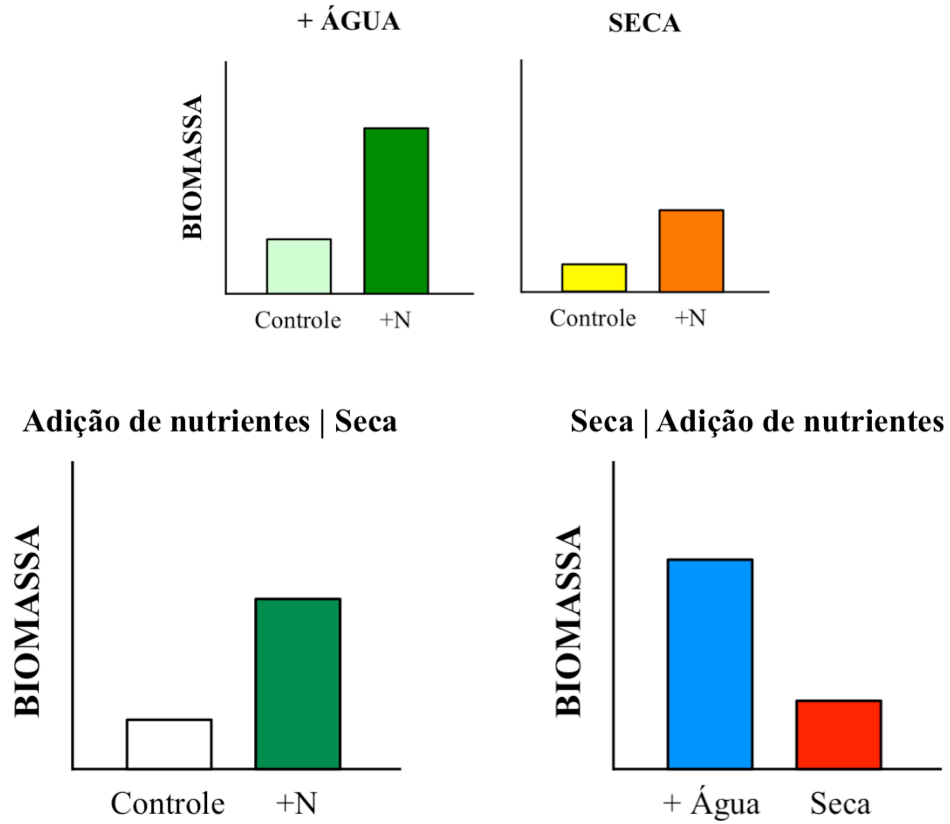
Efeitos principais *vs* interações

- O efeito principal de uma variável preditora (*main effect*) é aquele que descreve a forma pela qual esta variável modifica os valores da variável resposta, considerando o efeito de outras variáveis preditoras.



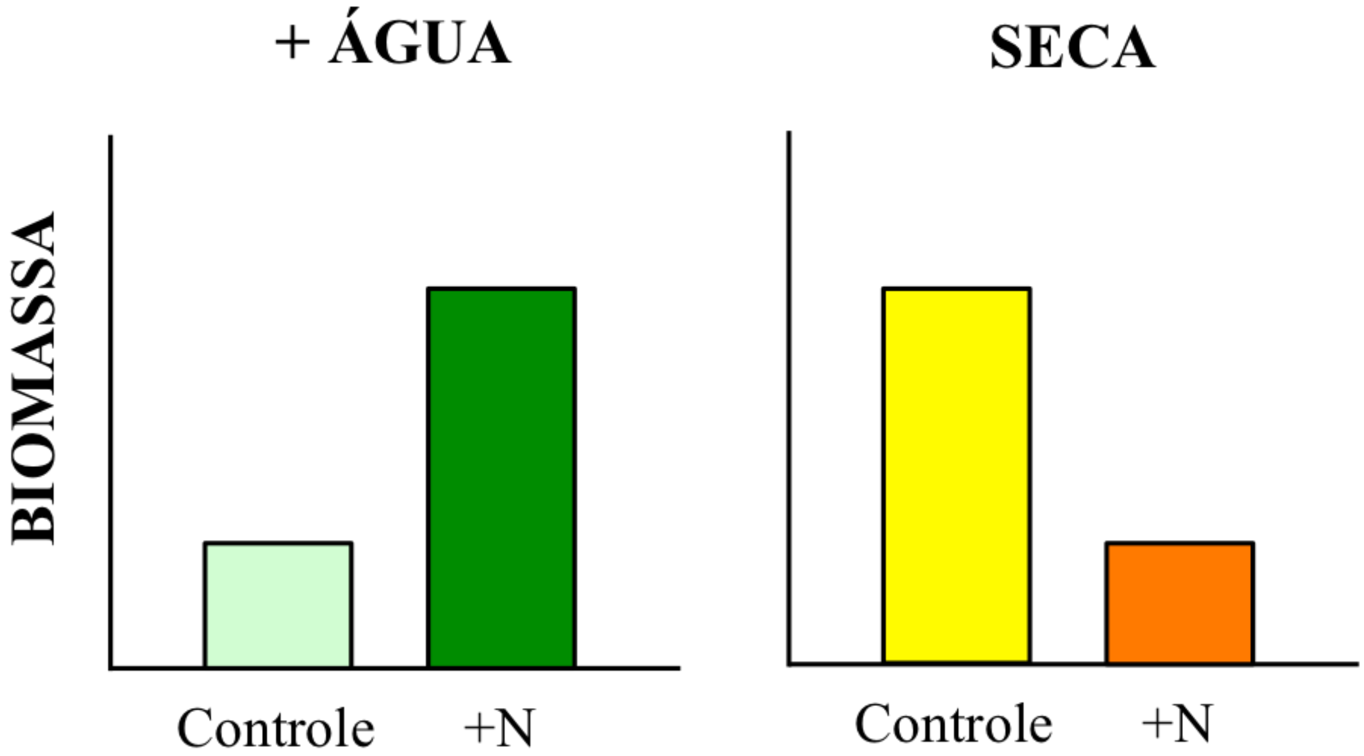
Efeitos principais *vs* interações

- Cada fator tem um efeito principal e individual na variável resposta: adição de nutrientes é aumentar a biomassa, enquanto que o efeito principal da seca é reduzir a biomassa.



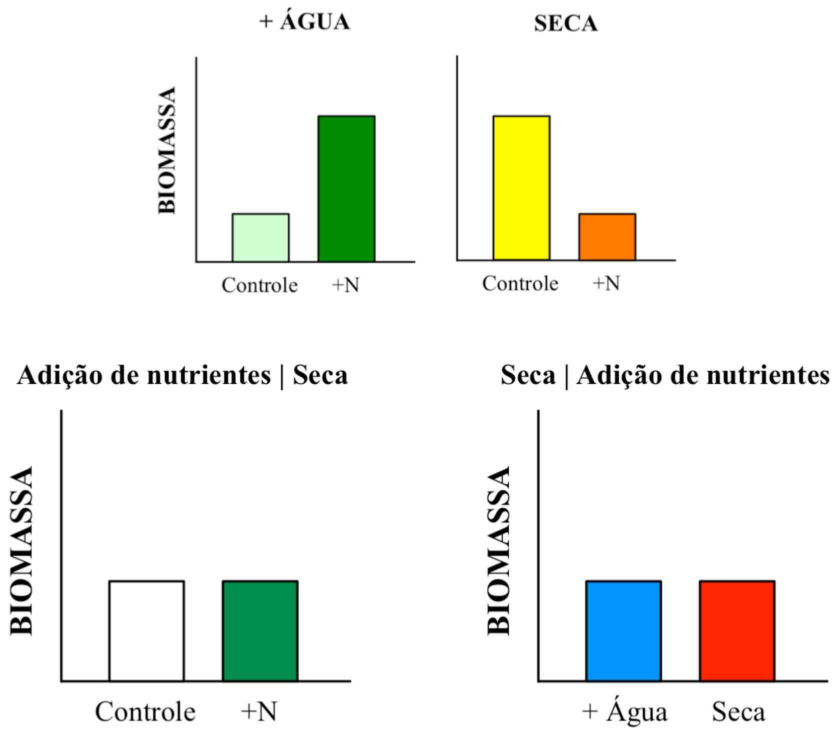
Efeitos principais *vs* interações

- Uma interação descreve a forma pela qual o efeito principal de uma variável preditora modifica o efeito principal de outra variável preditora sobre a variável resposta.

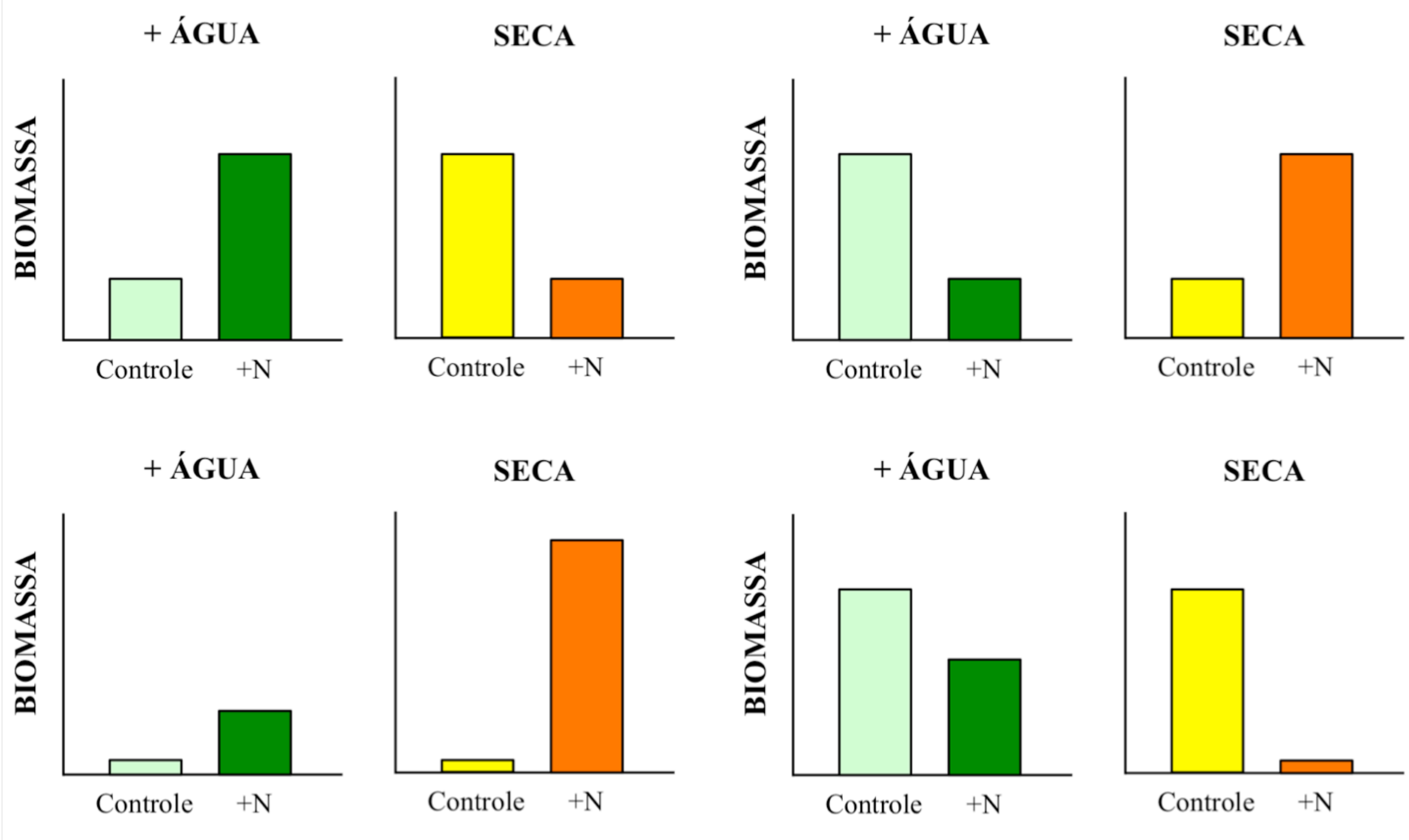


Efeitos principais *vs* interações

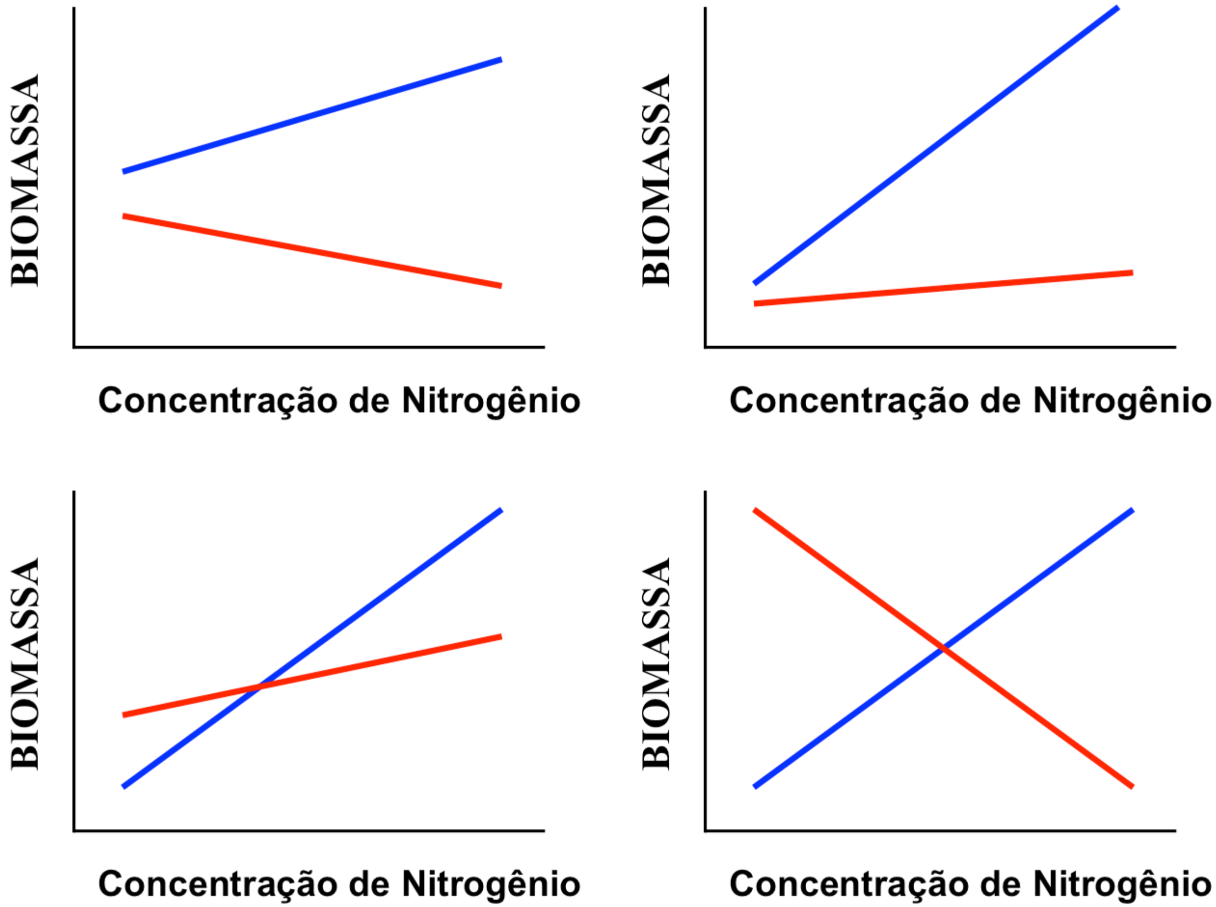
- Não podemos tirar conclusões sobre os efeitos principais de duas ou mais variáveis preditoras quando existe uma interação entre elas.
- Antagonismos e sinergismos nos efeitos variáveis preditoras são interações.



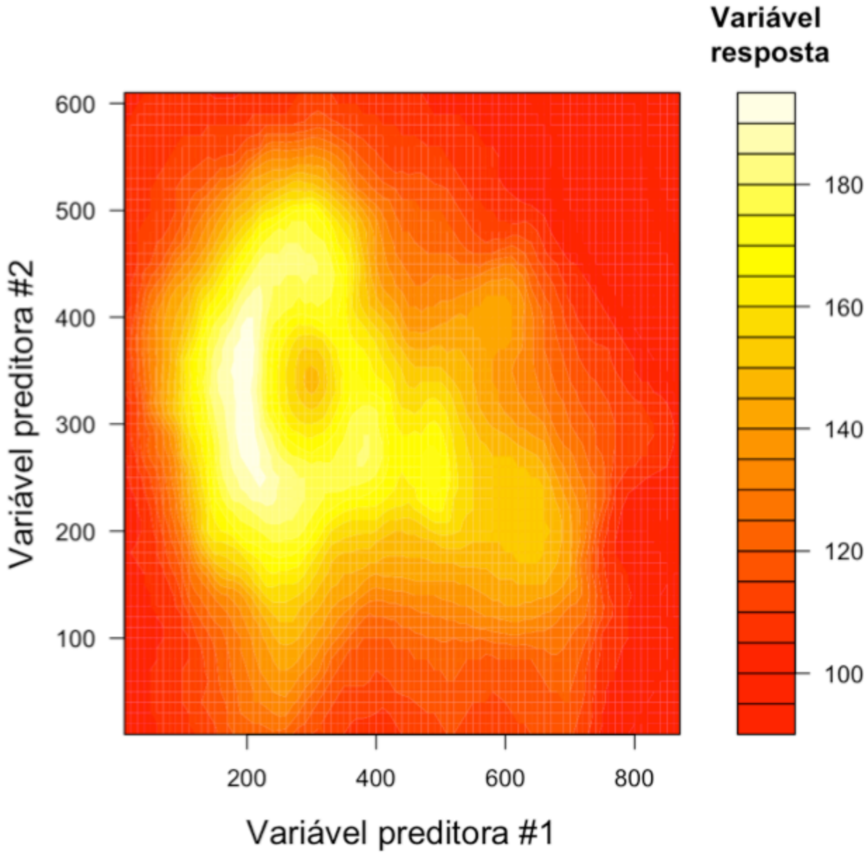
Interações entre variáveis categóricas

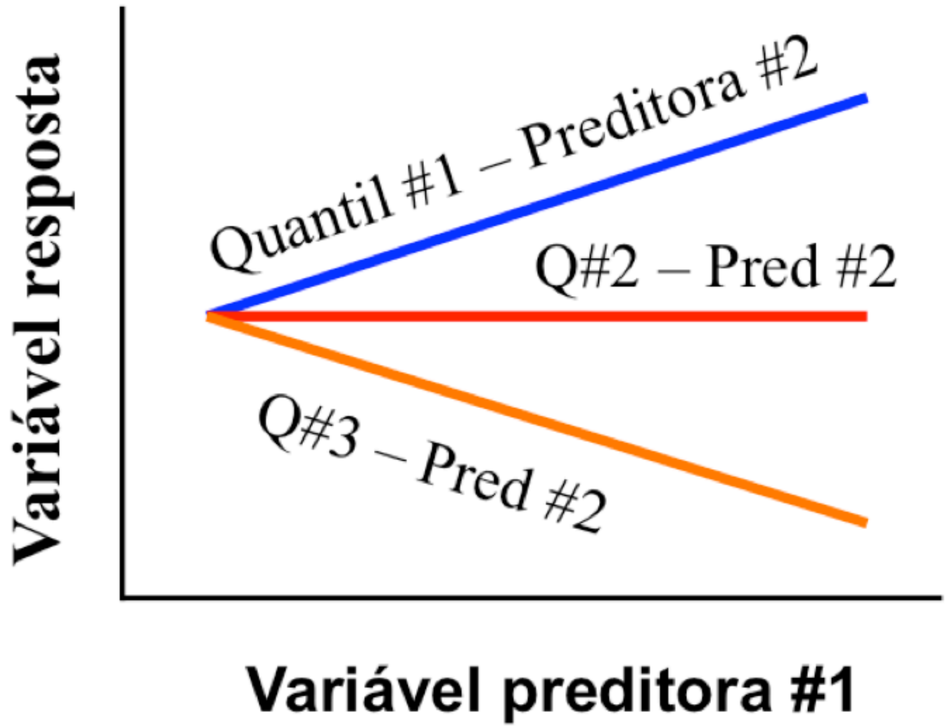


Interações entre uma variáveis categóricas e contínuas



Interações entre variáveis contínuas





Interações: como detectar e considerar?

- Podemos **detectar** a presença de uma interação:
    - Através da análise exploratória dos dados - figuras, figuras, figuras…
    - Adicionando uma interação ao modelo e testando:
        - Sua significância; e/ou,
        - Se sua inclusão melhora o ajuste do modelo aos dados.
- Devemos **considerar** uma interação no modelo quando:
    - Sua inclusão melhora o ajuste do modelo aos dados;
    - A sua pergunta/hipótese/predição incorpora envolve a existência de uma interação;
    - O seu experimento/delineamento contém uma interação.

Notação de modelos com múltiplas preditoras

- Vimos que um modelo de regressão linear simples pode ser escrito na forma:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Neste mesmo sentido, podemos descrever o modelo que contenha múltiplas variáveis preditoras x_i na forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_\mathrm{i} x_i + \epsilon$$

- Quando temos variáveis categóricas em um modelo, calculamos n - 1 valores de β_i (onde i $\neq$ 0).

- Isto ocorre pois o intercepto β_0 é tomado como um dos níveis da variável categórica e os outros β_i representaram o quanto os outros níveis alteram a estimativa de β_0.

Notação de modelos com múltiplas preditoras

- No exemplo abaixo, temos uma variável preditora categórica, com três níveis: a, b e c.

```
##   preditor resposta
## 1        a      4.5
## 2        b      8.4
## 3        c      7.2
## 4        a      9.0
## 5        b      2.0
## 6        c      8.0
## 7        a      6.6
## 8        b      3.5
## 9        c      3.7
```

Notação de modelos com múltiplas preditoras

- Quando criamos um modelo utilizando esta variável preditora, convertemos cada nível dela que não será o intercepto para valores de 0 ou 1 - isto é, ou a observação pertence aquele nível da variável preditora categórica ou não.

```
##   (Intercept) preditorb preditorc
## 1           1         0         0
## 2           1         1         0
## 3           1         0         1
## 4           1         0         0
## 5           1         1         0
## 6           1         0         1
## 7           1         0         0
## 8           1         1         0
## 9           1         0         1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$preditor
## [1] "contr.treatment"
```

- Nesse exemplo, nosso modelo seria, onde "ligaríamos"" ($x_i = 1$) ou "desligaríamos" ($x_i = 0$) os valores de x de acordo com o nível do fator:

$$y = \beta_0 + \beta_\mathrm{b} x_b + \beta_\mathrm{c} x_c + \epsilon$$

Notação de modelos com múltiplas preditoras

- Finalmente, quando temos uma interação em um modelo, calculamos um β_i para os efeitos principais de cada variável preditora, além de um outro β_i para a interação entre as variáveis preditoras:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

- Para termos uma interação em um modelo, precisamos considerar os seus efeitos principais também. Em outras palavras: **não pode existir um modelo somente com a interação sem haver também os efeitos principais**.

$$y = \beta_0 + \beta_3 x_1 x_2 + \epsilon$$

(não existe)

Regressão Múltipla

- É similar à regressão linear simples, mas incorpora os efeitos principais e, eventualmente, as interações entre múltiplas variáveis preditoras contínuas sobre a magnitude de uma variável resposta.
- Possui os mesmos pressupostos da regressão linear simples, com a adição de que as variáveis preditoras **não** podem ser **colineares**.
    - Colinearidade: implica em uma alta correlação entre duas variáveis preditoras que, normalmente, representam o mesmo fenômeno ou fenômenos muito similares. Por exemplo:
        - Concentração de O_2 na água *vs* saturação de O_2 na água;
        - Tamanho do corpo *vs* biomassa;
        - Temperatura máxima *vs* mínima *vs* amplitude; …
- Um modelo de regressão múltipla é:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \epsilon$$

Como calcular os βs na regressão múltipla

- A matemática pode ser complicada (método #1)…

$$\beta_1 = \frac{(\sum_{i=1}^{n} x_{2i}^2)(\sum_{i=1}^{n} x_{1i}y_i) - (\sum_{i=1}^{n} x_{1i}x_{2i})(\sum_{i=1}^{n} x_{2i}y_i)}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}$$

$$\beta_2 = \frac{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}y_i) - (\sum_{i=1}^{n} x_{1i}x_{2i})(\sum_{i=1}^{n} x_{1i}y_i)}{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2) - (\sum_{i=1}^{n} x_{1i}x_{2i})^2}$$

- Quanto mais _β_s, maior o número de termos no numerador e denominador para considerar o efeito de outras variáveis.

Como calcular os βs na regressão múltipla

- …ou mais simples (método #2).
- Calcular o coeficiente de correlação entre cada variável preditora x_n e a variável resposta y:

$$r_{ny} = \frac{\sum(x_{ni} - \bar{x_n})(y_i - \bar{y})}{\sqrt{\sum(x_{ni} - \bar{x_n})^2 - \sum(y_i - \bar{y})^2}}$$

- Calcular o coeficiente de correlação entre cada variável preditora no modelo:

$$r_{n_j\, n_k,\, k \neq j} = \frac{\sum(x_{nj,i} - \bar{x_{nj}})(x_{nk,i} - \bar{x_{nk}})}{\sqrt{\sum(x_{nj,i} - \bar{x_{nj}})^2 - \sum(x_{nk,i} - \bar{x_{nk}})^2}}$$

Como calcular os βs na regressão múltipla

- Através do método #2, o β do efeito da variável preditora 1 na variável resposta é, então:

$$\beta'_{1y,2} = \frac{r_{1y} - r_{2y}\ r_{12}}{(1 - r_{12}^2)}$$

- Enquanto que o β do efeito da variável preditora 2 na variável resposta é:

$$\beta'_{2y,1} = \frac{r_{2y} - r_{1y}\ r_{12}}{(1 - r_{12}^2)}$$

Diferença entre o método #1 e #2

- Através do método #1 obtemos uma estimativa do quanto a **mudança em uma unidade** na variável preditora #1 altera a magnitude da variável resposta, controlando o efeito da variável preditora 2 (β_1 e β_2).
- Já, através do método #2, obtemos uma estimativa do quanto a **mudança em uma unidade de desvio padrão** na variável preditora #1 **altera em um desvio padrão** a magnitude da variável resposta, controlando o efeito da variável preditora 2 (β'_1 e β'_2).
- Podemos converter entre as duas formas através da fórmula:

$$\beta_{jy,k} = \beta'_{jy,k} \frac{S_y}{S_{xj}}$$

Exemplo

- O R considera o método #1:

```
modelo1 <- lm(log(riqueza) ~ log(area) + precipitacao, data = ilhas)
summary(modelo1)
```

```
##
## Call:
## lm(formula = log(riqueza) ~ log(area) + precipitacao, data = ilhas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5244 -0.5187  0.1467  0.5404  0.9998
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.541e+00  2.501e-01  10.161 7.14e-16 ***
## log(area)     1.292e-01  2.415e-02   5.349 8.80e-07 ***
## precipitacao -5.367e-05  1.299e-04  -0.413    0.681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6685 on 77 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.266
## F-statistic: 15.32 on 2 and 77 DF,  p-value: 2.51e-06
```

Exemplo

- Mas podemos calcular o β' a partir do modelo:

```
## coeficientes do modelo
round(coef(modelo1), digits = 5)
```

```
##  (Intercept)    log(area) precipitacao
##      2.54117      0.12916     -0.00005
```

```
## mudando a igualdade
(coef(modelo1)[2] * sd(log(ilhas$area)))/sd(log(ilhas$riqueza))
```

```
## log(area)
## 0.5245438
```

```
(coef(modelo1)[3] * sd(ilhas$precipitacao))/sd(log(ilhas$riqueza))
```

```
## precipitacao
##  -0.04052776
```

Exemplo

- Ou padronizar e centralizar as variáveis preditoras e resposta **antes** de rodar o modelo (padronizar = transformação em valores de *Z*).
- Existem várias opções para automatizar esse processo, *e.g.* através da função `standardize` do pacote `arm`.

```
library(arm)
standardize(lm(log_riqueza ~ log_area + precipitacao, data = ilhas), standardize.y = TRUE)
```

```
##    (Intercept)     z.log_area z.precipitacao
##        0.00000        0.52454       -0.04053
```

Estimando a importância de uma variável na regressão

- β' é conhecido como o **coeficiente padronizado** e é uma estimativa não-enviesada da magnitude do efeito de uma variável preditora na variável resposta.
    - Coeficientes não são afetados pela sua magnitude: importante quando variáveis preditoras variam em ordens de grandeza.

```
##                     min        max
## area          -3.816713   11.84536
## precipitacao 299.024106 3244.36796
```

```
##  (Intercept)     log_area precipitacao
##      2.54117      0.12916     -0.00005
```

```
##    (Intercept)     z.log_area z.precipitacao
##        0.00000        0.52454       -0.04053
```

Estimando a importância de uma variável na regressão

- β' é conhecido como o **coeficiente padronizado** e é uma estimativa não-enviesada da magnitude do efeito de uma variável preditora na variável resposta.
    - Coeficientes das variáveis preditoras polinomiais e/ou que representem algum tipo de não-linearidade podem ser expressas de forma mais 'natural'.

```
##       (Intercept)          log_area      precipitacao I(precipitacao^2)
##         2.0636468         0.1296343         0.0006018        -0.0000002
```

```
##         (Intercept)          z.log_area      z.precipitacao
##             0.04340             0.52646            -0.03660
## I(z.precipitacao^2)
##            -0.17579
```

Estimando a importância de uma variável na regressão

- β' é conhecido como o **coeficiente padronizado** e é uma estimativa não-enviesada da magnitude do efeito de uma variável preditora na variável resposta.
    - A expressão do coeficiente da interação não afeta a expressão dos coeficientes dos efeitos principais.

```
## uma forma de representar a interacao
modelo2 <- lm(log_riqueza ~ log_area * precipitacao, data = ilhas)
## outra forma de representar a interacao
modelo2 <- lm(log_riqueza ~ log_area + precipitacao + log_area : precipitacao, data = ilhas)
```

```
##           (Intercept)              log_area          precipitacao
##             2.5934618             0.1137299            -0.0000819
## log_area:precipitacao
##             0.0000088
```

```
##               (Intercept)                z.log_area
##                   0.00192                   0.52079
##             z.precipitacao z.log_area:z.precipitacao
##                  -0.04151                   0.04218
```

Regressão: miscelânias e resumindo

1. Padronize as variáveis preditoras e resposta se você quiser ter uma noção da importância de cada variável em um modelo;
2. Padronização será especialmente importante quando houver interações, termos não-lineares e durante a seleção de modelos!;
3. Faça interpretações sobre a mudança na magnitude da variável y em função das variáveis preditoras usando os valores em escala natural;
4. Interações entre variáveis preditoras podem ser representadas como "variavel #1 * variavel #2" ou "variavel #1 + variavel #2 + variavel #1 : variavel #2";
5. Graus de liberdade de cada variável preditora contínua é sempre n - 1; para os resíduos será n - k - 1, onde k é o número de variáveis preditoras contínuas no modelo;
6. Existe uma diferença entre na forma como o teste estatístico é feito e testado no `summary` e na `anova`, vamos ver isso mais a frente;
7. Atente ao R^2 ajustado da regressão: leva em consideração o número de variáveis preditoras no modelo para estimar o coeficiente de determinação.

ANOVA n-way

- É a análise de variância utilizada quando estamos interessados em determinar o efeito de duas ou mais variáveis preditoras categóricas (x_1, x_2,…, x_n) sobre a magnitude de uma variável resposta contínua (_y) - portanto, possui todos os pressupostos da ANOVA one-way.
- Testa a hipótese nula de que os valores das médias entre os níveis de cada uma das variáveis categóricas não diferem entre si:

$$H_{0,A} = \mu_{A1} + \mu_{A2} + \ldots + \mu_{An} = 0 \quad OU \quad \mu_{A1} = \mu_{A2} = \ldots = \mu_{An}$$

$$H_{0,B} = \mu_{B1} + \mu_{B2} + \ldots + \mu_{Bn} = 0 \quad OU \quad \mu_{B1} = \mu_{B2} = \ldots = \mu_{Bn}$$

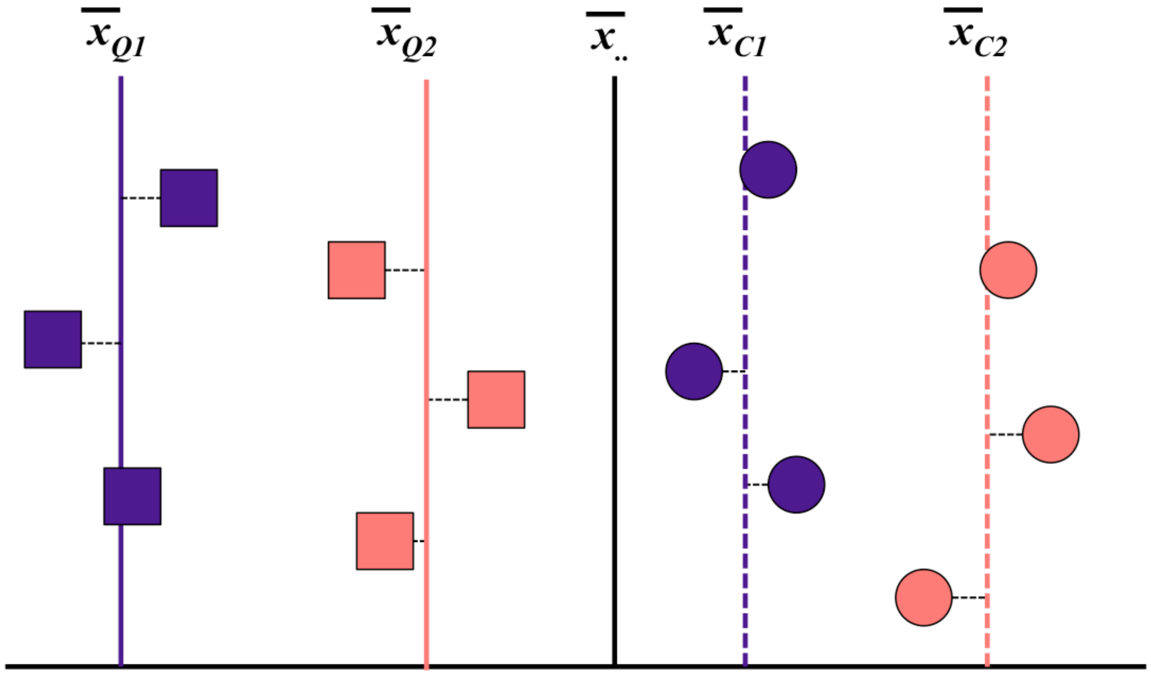
- No exemplo acima, temos dois fatores sendo testados simultaneamente na ANOVA; neste caso, chamamos este teste de **ANOVA two-way** ou **ANOVA de dois fatores**.
    - ANOVA three-way, ANOVA four-way, ANOVA five-way,…, ANOVA n-way.

ANOVA n-way

- Uma ANOVA n-way pode ser utilizada para testar apenas os efeitos principais de dois fatores *ou* os seus efeitos principais e interações.
  - ANOVA two-way: $y = \beta_0 + \beta_A x_A + \beta_B x_B + \epsilon$
  - ANOVA two-way fatorial: $y = \beta_0 + \beta_A x_A + \beta_B x_B + \beta_{AB} x_A x_B + \epsilon$
- Nem toda ANOVA n-way precisa ser uma ANOVA fatorial! As vezes, características da pergunta, hipóteses, predições, da amostragem e dos próprios dados previnem com que utilizemos uma combinação fatorial e tenhamos as interações no modelo.

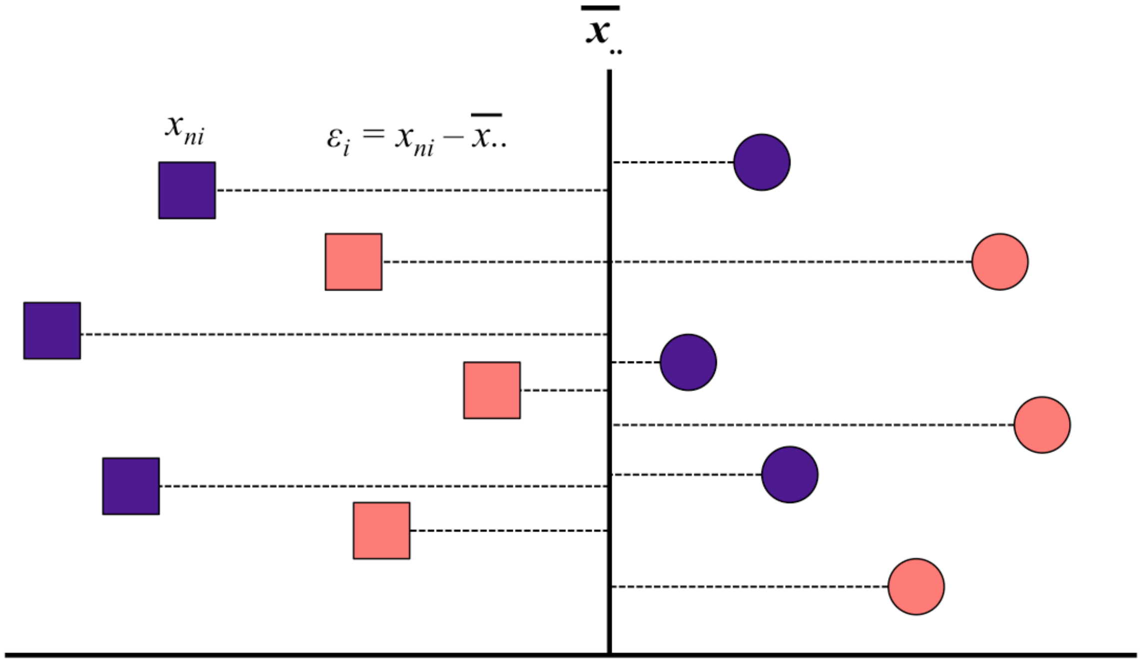
```
##            presenca_de_odor
## cor_da_flor nao sim
##     amarela   0   4
##     azul      2   2
##     verde     2   2
```

A mecânica de uma ANOVA n-way



SST = SSA + SSB + SSAB + SSE

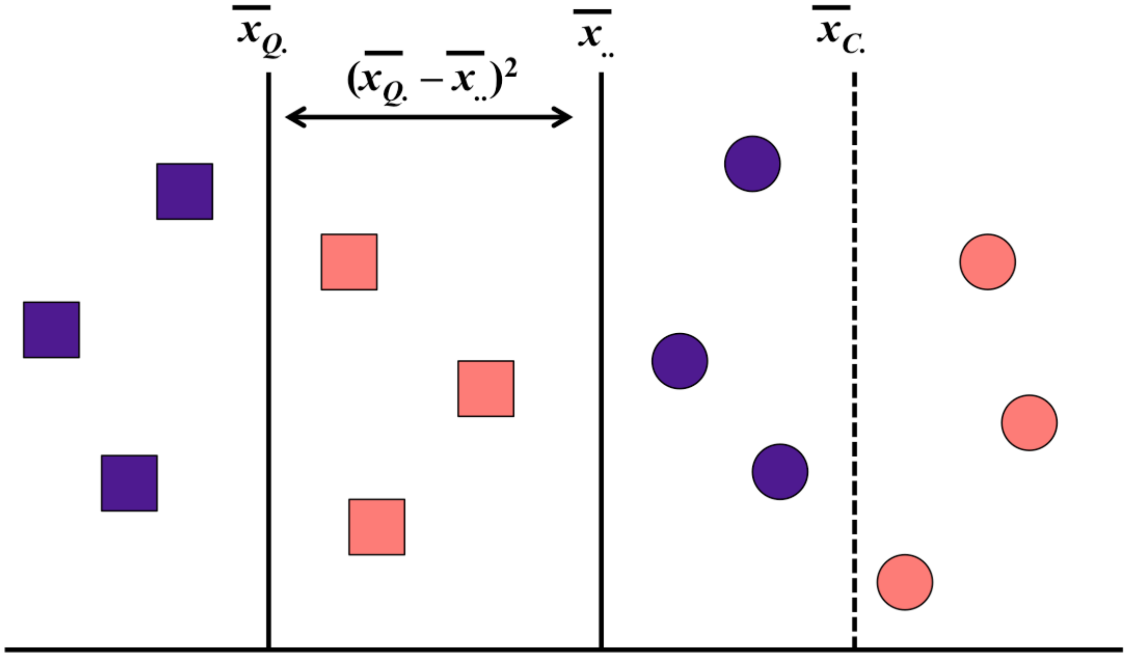
A mecânica de uma ANOVA n-way



Variância total nos dados = SST = $\Sigma\ \varepsilon_i^2$

A mecânica de uma ANOVA n-way

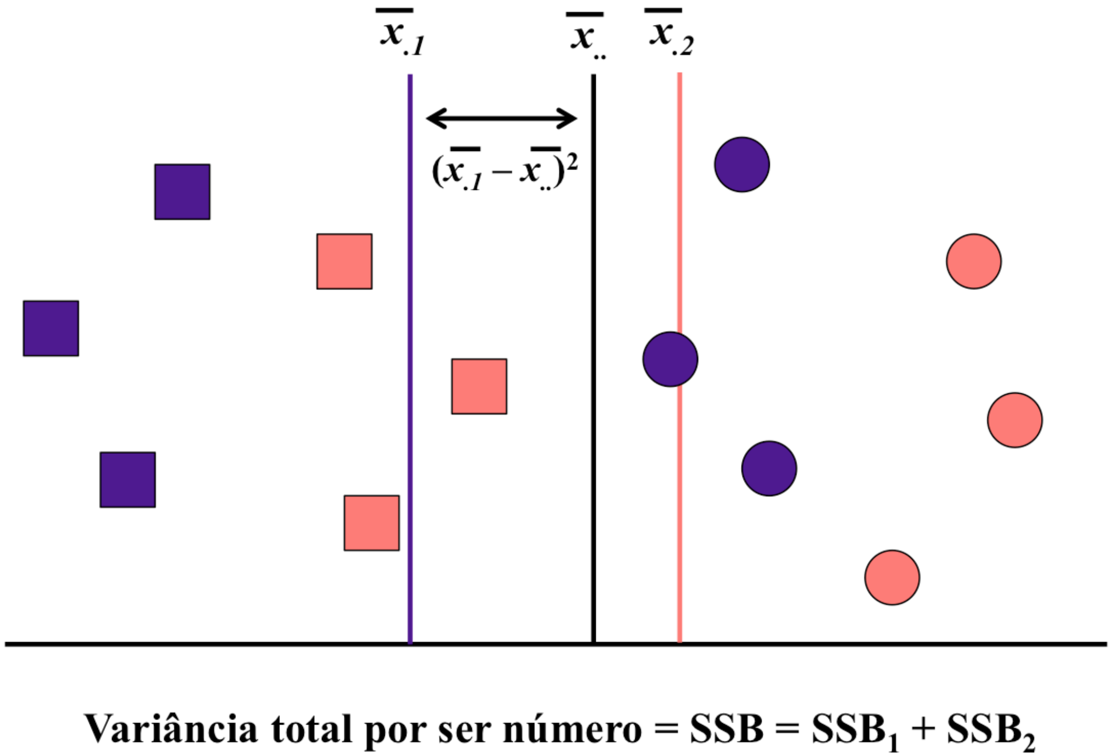
$$SSA_Q = (\bar{x_{Q.}} - \bar{x}_{..})^2 \qquad SSA_C = (\bar{x_{C.}} - \bar{x}_{..})^2$$



Variância total por ser letra = SSA = $\mathbf{SSA_Q + SSA_C}$

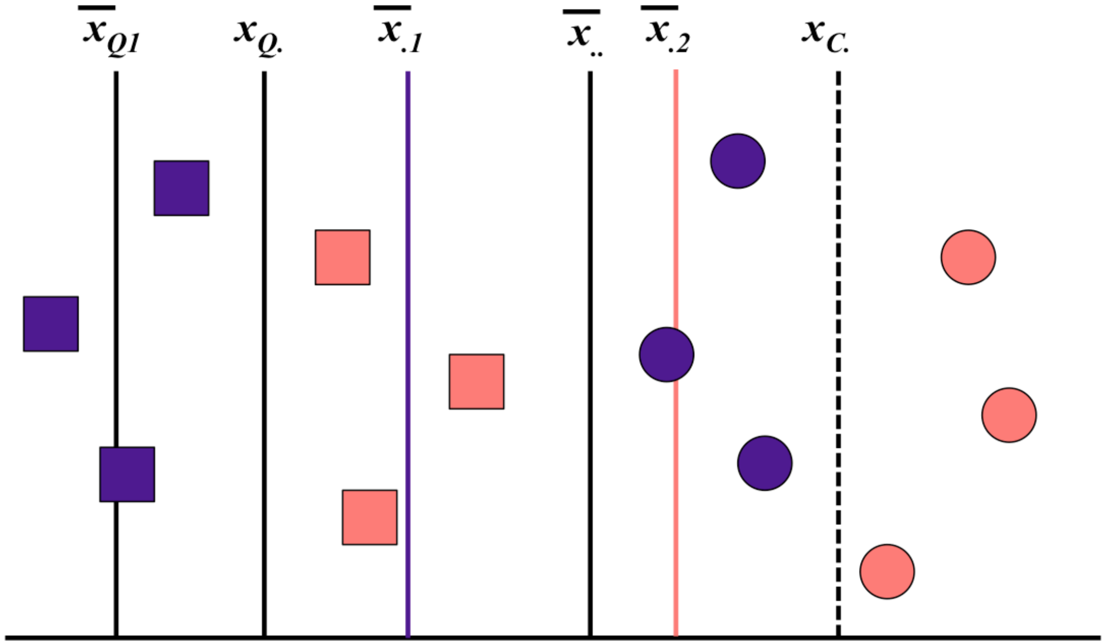
A mecânica de uma ANOVA n-way

$$SSB_1 = (\bar{x}_{.1} - \bar{x}_{..})^2 \qquad SSB_2 = (\bar{x}_{.2} - \bar{x}_{..})^2$$



A mecânica de uma ANOVA n-way

$$SSAB = \sum_{i} \sum_{j} (\bar{x_{ij}} - \bar{x_{i.}} - \bar{x_{.j}} + \bar{x_{..}})^2$$



Variância total por ser letra e número = SSAB

Teste estatístico para a ANOVA n-way

- Assim como fizemos para a ANOVA one-way, a ANOVA n-way utiliza como teste estatístico os valores de F para cada termo utilizado.
- Os valores de F são obtidos pela razão entre a Média dos Quadrados de cada termo e a Média dos Quadrados do Resíduo: F = $\frac{MSX}{MSE}$
- Os graus de liberdade para o cálculo de cada Média dos Quadrados é:
  - A - 1 ou B -1 para cada efeito principal, , onde A e B representam o número de níveis dentro de cada fator analisado;
  - (A - 1)(B - 1) para a interação;
  - AB(n - 1) para a variação residual, onde n é o número total de observações; e,
  - AB - 1 para a variação total.

Exemplo

- No R, podemos fazer uma ANOVA two-way associando a função `lm` à `anova`:

```
modelo3 <- lm(log(riqueza) ~ ilha * montanha, data = ilhas)
anova(modelo3)
```

```
## Analysis of Variance Table
##
## Response: log(riqueza)
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## ilha           1 18.0629 18.0629 60.5387 2.892e-11 ***
## montanha       1  6.1118  6.1118 20.4840 2.186e-05 ***
## ilha:montanha  1  1.2554  1.2554  4.2074   0.04369 *
## Residuals     76 22.6761  0.2984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercício 2

- Em duplas, comparem os resultados abaixo. O que você notou de diferente entre eles?
- Por que vocês acham que isso ocorreu?

```
anova(lm(log(riqueza) ~ ilha * montanha, data = ilhas))
anova(lm(log(riqueza) ~ montanha * ilha, data = ilhas))
```

Tipo de Soma dos Quadrados

- Apesar de intuitivo e fácil a obtenção dos valores de F para o teste estatístico, a maioria dos pacotes estatísticos oferece três formas principais para o cálculo das somas dos quadrados e valores de F.
    - Soma dos Quadrados do Tipo I ou **Soma dos Quadrados Sequencial**.
    - Soma dos Quadrados do Tipo II ou **Soma dos Quadrados Marginal**.
    - Soma dos Quadrados do Tipo III.
- A escolha do tipo de Soma dos Quadrados depende das suas perguntas, hipóteses e/ou predições.
- Mais sobre isso em:
    - https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/
    - https://www.uvm.edu/~dhowell/StatPages/More_Stuff/Type1-3.pdf

Soma dos Quadrados Sequencial (Tipo I)

- Útil quando sua hipótese preve que um efeito só aparece depois de controlarmos a influência de alguma outra variável.

$$y = \beta_0 + \beta_A \Rightarrow SS_A = SSA \Rightarrow SSA' = SS_A$$

$$y = \beta_0 + \beta_A + \beta_B \Rightarrow SS_{A+B} = (SSA + SSB) \Rightarrow SSB' = SS_{A+B} - SS_A$$

$$y = \beta_0 + \beta_A + \beta_B + \beta_{AB} \Rightarrow SS_{A+B+AB} = SSA + SSB + SSAB \Rightarrow SSAB' = SS_{A+B+AB} - SS_{A+B}$$

Soma dos Quadrados Marginal (Tipo II)

- Útil quando sua hipótese preve quer testar os efeitos de cada variável isolando o efeito das demais.

$$y = \beta_0 + \beta_A + \beta_B \Rightarrow SS_{A+B} = (SSA + SSB)$$

$$y = \beta_0 + \beta_A \Rightarrow SSA' = SS_{A+B} - SSA$$

$$y = \beta_0 + \beta_B \Rightarrow SSB' = SS_{A+B} - SSB$$

$$y = \beta_0 + \beta_A + \beta_B + \beta_{AB} \Rightarrow SS_{A+B+AB} = SSA + SSB + SSAB \Rightarrow SSAB' = SS_{A+B+AB} - SS_{A+B}$$

Soma dos Quadrados do Tipo III

- Útil quando sua hipótese preve quer testar os efeitos de uma interação.

$$y = \beta_0 + \beta_A + \beta_B + \beta_{AB} \Rightarrow SS_{A+B+AB} = SSA + SSB + SSAB$$

$$y = \beta_0 + \beta_A + \beta_B \Rightarrow SS_{A+B} = (SSA + SSB) \Rightarrow SSAB' = SS_{A+B+AB} - SS_{A+B}$$

$$y = \beta_0 + \beta_B + \beta_{AB} \Rightarrow SS_{B+AB} = SSB + SSAB \Rightarrow SSA' = SS_{A+B+AB} - SS_{B+AB}$$

$$y = \beta_0 + \beta_A + \beta_{AB} \Rightarrow SS_{A+AB} = SSA + SSAB \Rightarrow SSB' = SS_{A+B+AB} - SS_{A+AB}$$

Soma dos Quadrados no R

- No R, podemos utilizar a função `Anova` do pacote `car` para rodar ANOVAs com Soma dos Quadrados do Tipo II e III.

```
Anova(modelo3, type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: log(riqueza)
##                Sum Sq Df F value    Pr(>F)
## ilha          23.3085  1 78.1194 2.721e-13 ***
## montanha       6.1118  1 20.4840 2.186e-05 ***
## ilha:montanha  1.2554  1  4.2074   0.04369 *
## Residuals     22.6761 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Soma dos Quadrados no R

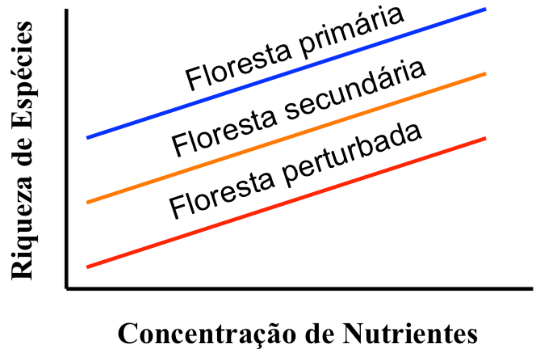
- No R, podemos utilizar a função `Anova` do pacote `car` para rodar ANOVAs com Soma dos Quadrados do Tipo II e III.

```
Anova(modelo3, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: log(riqueza)
##                Sum Sq Df  F value    Pr(>F)
## (Intercept)   231.177  1 774.7988 < 2.2e-16 ***
## ilha           15.160  1  50.8087 5.035e-10 ***
## montanha        1.258  1   4.2147   0.04352 *
## ilha:montanha   1.255  1   4.2074   0.04369 *
## Residuals      22.676 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

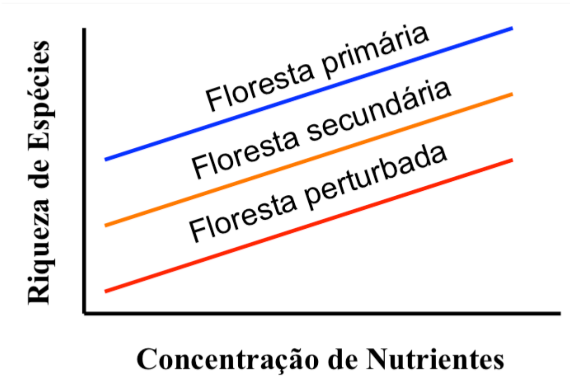
Análise de Covariância (ANCOVA) e similares

- Podemos combinar os efeitos principais e interações entre variáveis preditoras categóricas e contínuas em um mesmo modelo.
- Uma aplicação útil destes modelos é incorporar e controlar o efeito de uma covariável sobre a variável resposta, ao tentar estimar o efeito de um tratamento.
  - Variação natural no tamanho do habitat ou entre unidades amostrais, ao relacionar diversidade de habitats à riqueza de espécies;
  - Controlar a variação da biomassa dos indivíduos ao medir o efeito da temperatura no metabolismo;
  - Considerar o efeito da variação na concentração de nutrientes entre unidades amostrais ao examinar o efeito de uma outra variável no efeito das espécies;...



Análise de Covariância (ANCOVA) e similares

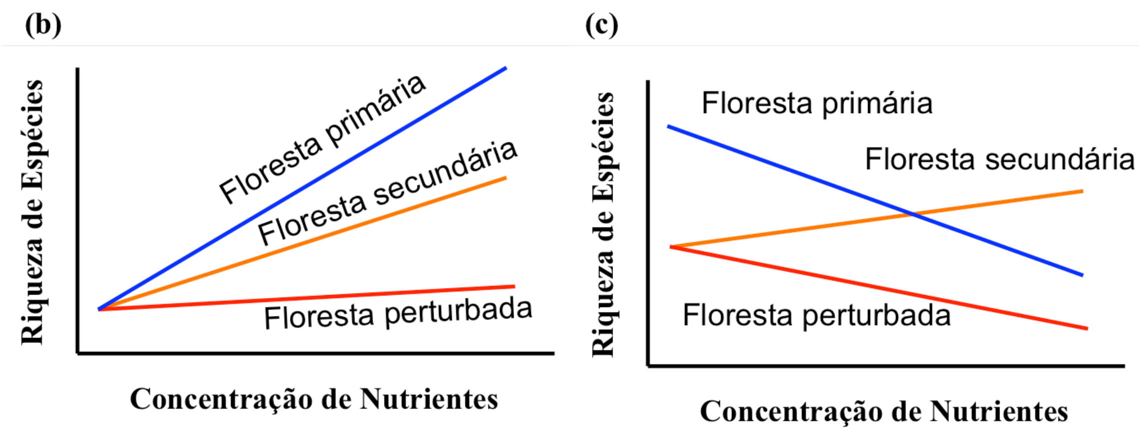
- Existem três hipóteses que normalmente testamos com desenhos do tipo "ANCOVA":



(a) Diferenças entre interceptos, mas inclinações similares;

(b) Similaridades no intercepto, mas inclinações diferentes;

(c) Diferenças no intercepto e nas inclinações.



Análise de Covariância (ANCOVA) e similares

- A ANCOVA é o caso especial no qual apenas os interceptos variam, mas as inclinações das restas entre os níveis da variável categórica preditora são similares;
- Na ANCOVA, a estimativa dos βs para cada nível da variável preditora são feitas para um valor fixo da variável preditora contínua x.
- No R:

```
## usado apenas para ANCOVA stricto sensu
modelo4 <- lm(log(riqueza) ~ log(area) + ilha, data = ilhas)
## usado para ANCOVA (quando a interação não é significativa) e similares
modelo4 <- lm(log(riqueza) ~ log(area) * ilha, data = ilhas)
```

Exemplo

```
modelo4 <- lm(log(riqueza) ~ log(area) * ilha, data = ilhas)
summary(modelo4)
```

```
##
## Call:
## lm(formula = log(riqueza) ~ log(area) * ilha, data = ilhas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08379 -0.23339  0.06018  0.23135  1.14284
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.07290    0.08189  37.523  < 2e-16 ***
## log(area)               0.09202    0.02072   4.441    3e-05 ***
## ilhaoceanica           -1.34773    0.11766 -11.454  < 2e-16 ***
## log(area):ilhaoceanica  0.09848    0.02722   3.618 0.000533 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3763 on 76 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7675
## F-statistic: 87.92 on 3 and 76 DF,  p-value: < 2.2e-16
```

Pós-testes

- Assim como para a ANCOVA, podemos realizar pós-testes para a ANOVA n-way e ANCOVA (e similares) utilizando o pacote `lsmeans`.

```
lsmeans(object = modelo3, specs = ~ ilha + montanha)
```

```
##  ilha     montanha   lsmean         SE df lower.CL upper.CL
##  costeira nao      3.170357 0.11389737 76 2.943511 3.397203
##  oceanica nao      1.695534 0.17273392 76 1.351505 2.039564
##  costeira sim      3.529035 0.13248087 76 3.265176 3.792893
##  oceanica sim      2.598095 0.09972797 76 2.399470 2.796721
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

Pós-testes

- Assim como para a ANCOVA, podemos realizar pós-testes para a ANOVA n-way e ANCOVA (e similares) utilizando o pacote `lsmeans`.

```
contrast(lsmeans(object = modelo3, specs = ~ ilha + montanha),
         by = "montanha", method = "tukey", adjust = "none")
```

```
## montanha = nao:
##  contrast             estimate        SE df t.ratio p.value
##  costeira - oceanica 1.4748227 0.2069049 76   7.128  <.0001
##
## montanha = sim:
##  contrast             estimate        SE df t.ratio p.value
##  costeira - oceanica 0.9309396 0.1658217 76   5.614  <.0001
##
## Results are given on the log (not the response) scale.
```

Pós-testes

- Assim como para a ANCOVA, podemos realizar pós-testes para a ANOVA n-way e ANCOVA (e similares) utilizando o pacote `lsmeans`.

```
lstrends(model = modelo4, specs = "ilha", var = "log(area)")
```

```
##  ilha     log(area).trend         SE df   lower.CL  upper.CL
##  costeira      0.09202304 0.02072268 76 0.05075024 0.1332958
##  oceanica      0.19050230 0.01765288 76 0.15534353 0.2256611
##
## Trends are based on the log (transformed) scale
## Confidence level used: 0.95
```

Pós-testes

- Assim como para a ANCOVA, podemos realizar pós-testes para a ANOVA n-way e ANCOVA (e similares) utilizando o pacote `lsmeans`.

```
contrast(lstrends(model = modelo4, specs = "ilha", var = "log(area)"), method = "tukey")
```

```
##  contrast              estimate        SE df t.ratio p.value
##  costeira - oceanica -0.09847926 0.0272223 76  -3.618  0.0005
```