

Estendendo a regressão linear

Adicionando mais variáveis, interações e transformações.

Regressão linear é muito flexível!

Algumas transformações simples:

- O modelo linear que estamos usando consiste em fazer os parâmetros de uma distribuição mudar de acordo com alguma função.
- Às vezes, a relação entre parâmetros e variáveis não é linear.
- Nós podemos usar qualquer função, mas às vezes é útil usar uma transformação que lineariza a relação entre as variáveis.
- A função mais simples é uma função linear:

$$y_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

Regressão linear é muito flexível!

Algumas transformações simples:

- O modelo linear que estamos usando consiste em fazer os parâmetros de uma distribuição mudar de acordo com alguma função.
- Às vezes, a relação entre parâmetros e variáveis não é linear.
- Nós podemos usar qualquer função, mas às vezes é útil usar uma transformação que lineariza a relação entre as variáveis.
- A função mais simples é uma função linear:

$$y_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$



$$\log(y_i) \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

Regressão linear é muito flexível!

Incrementos multiplicativos.

- A transformação logarítmica da resposta é uma transformação comum quando o efeito do preditor sobre a resposta é multiplicativo.
 - mudança de uma unidade em x é associada com uma mudança de uma **porcentagem** constante de y .
- Muitos processos se beneficiam com uma transformação logarítmica:
 - Crescimento é proporcional ao tamanho anterior.
 - Qualquer processo multiplicativo.

$$y_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$



$$\Delta y \approx \beta \Delta x$$

$$\log(y_i) \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$



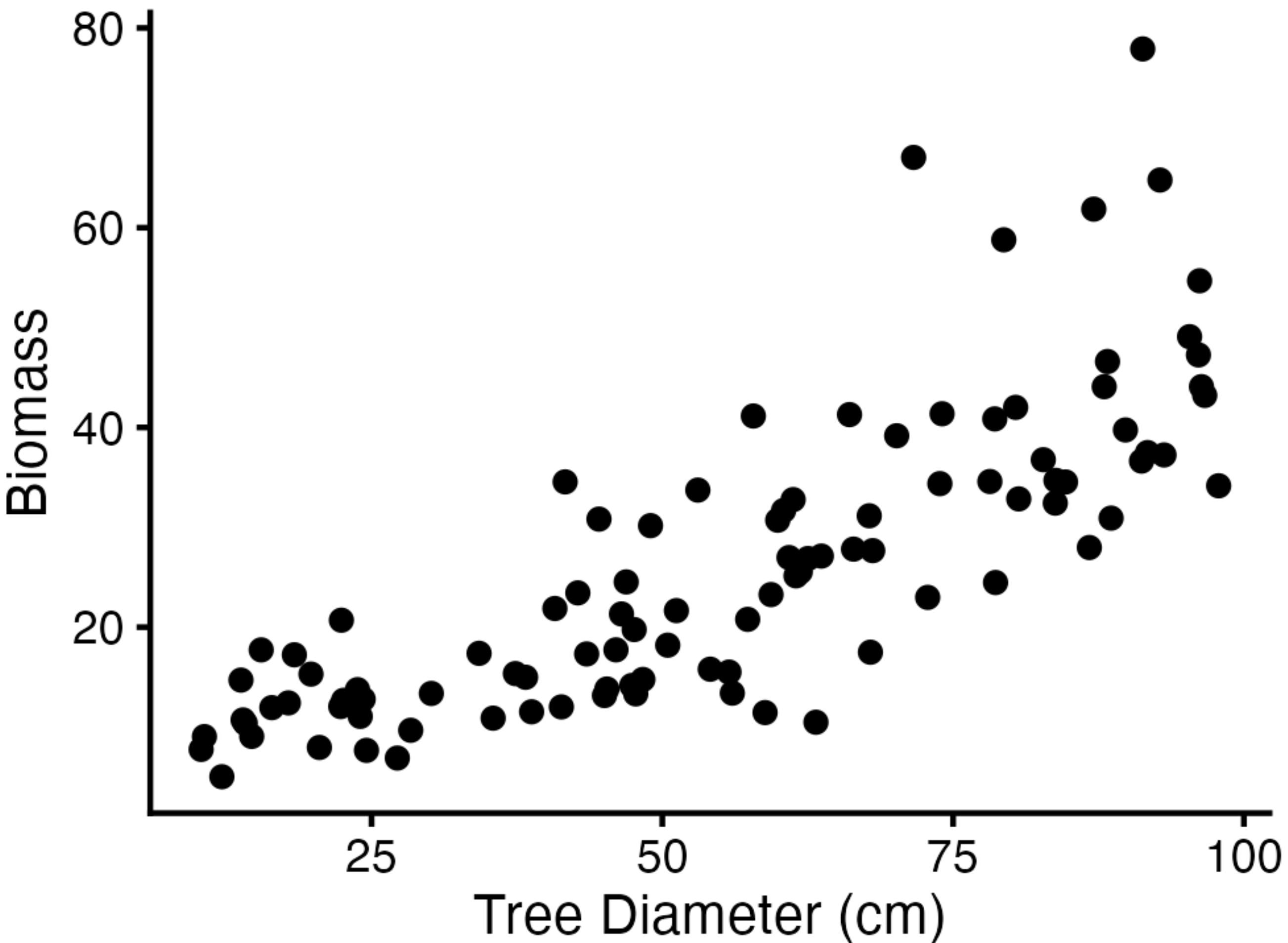
$$\% \Delta y \approx (100 \times \beta) \Delta x$$

Biomassa por diâmetro

Exemplo de relação não-linear.

- log transformar y antes de ajustar o modelo

```
df = data.frame(diameter, biomass)
stan_fit = stan_glm(log(biomass) ~ diameter,
                     data = df)
```

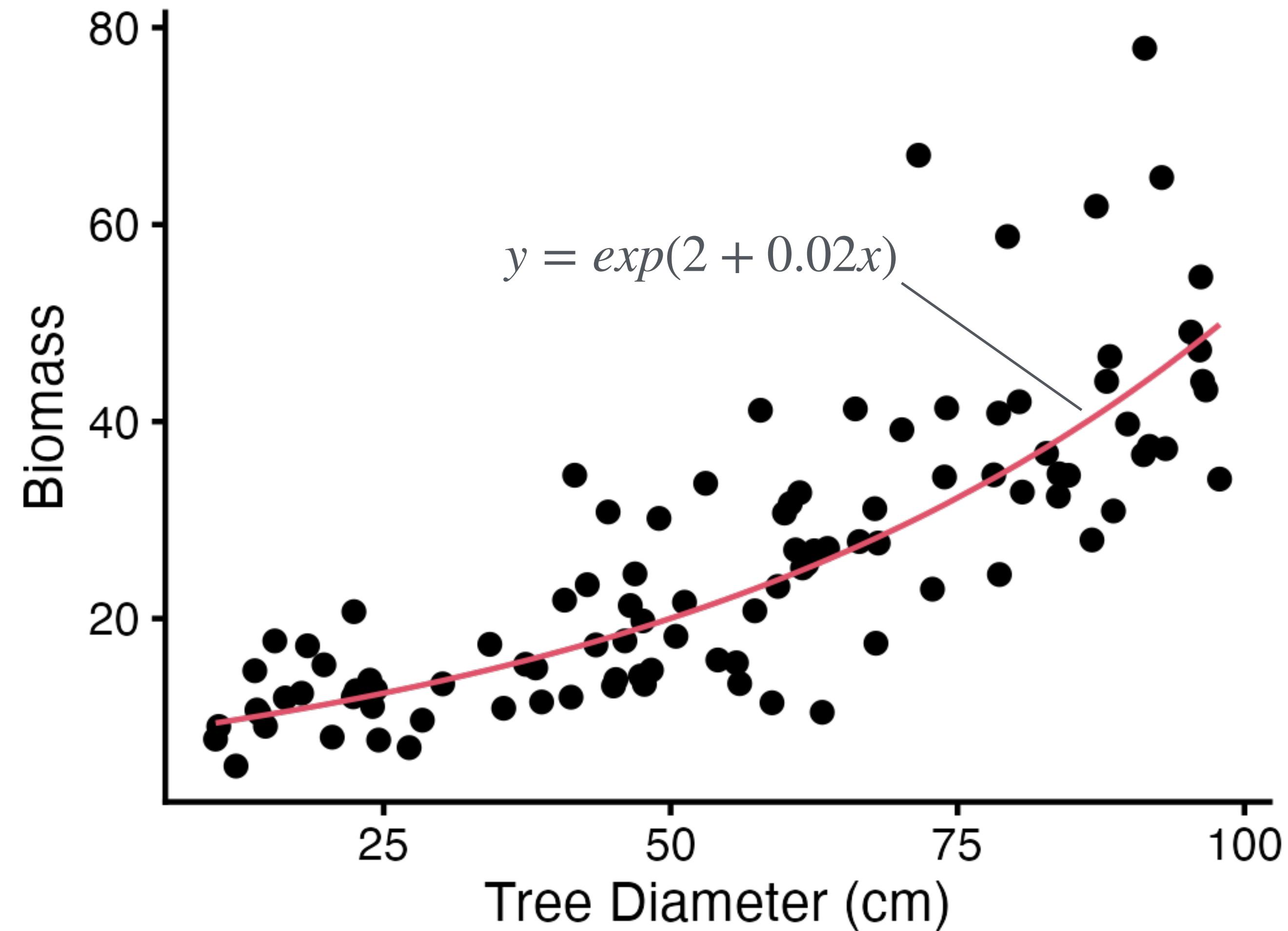


Biomassa por diâmetro

Exemplo de relação não-linear.

- log transformar y antes de ajustar o modelo

```
df = data.frame(diameter, biomass)
stan_fit = stan_glm(log(biomass) ~ diameter,
                     data = df)
```

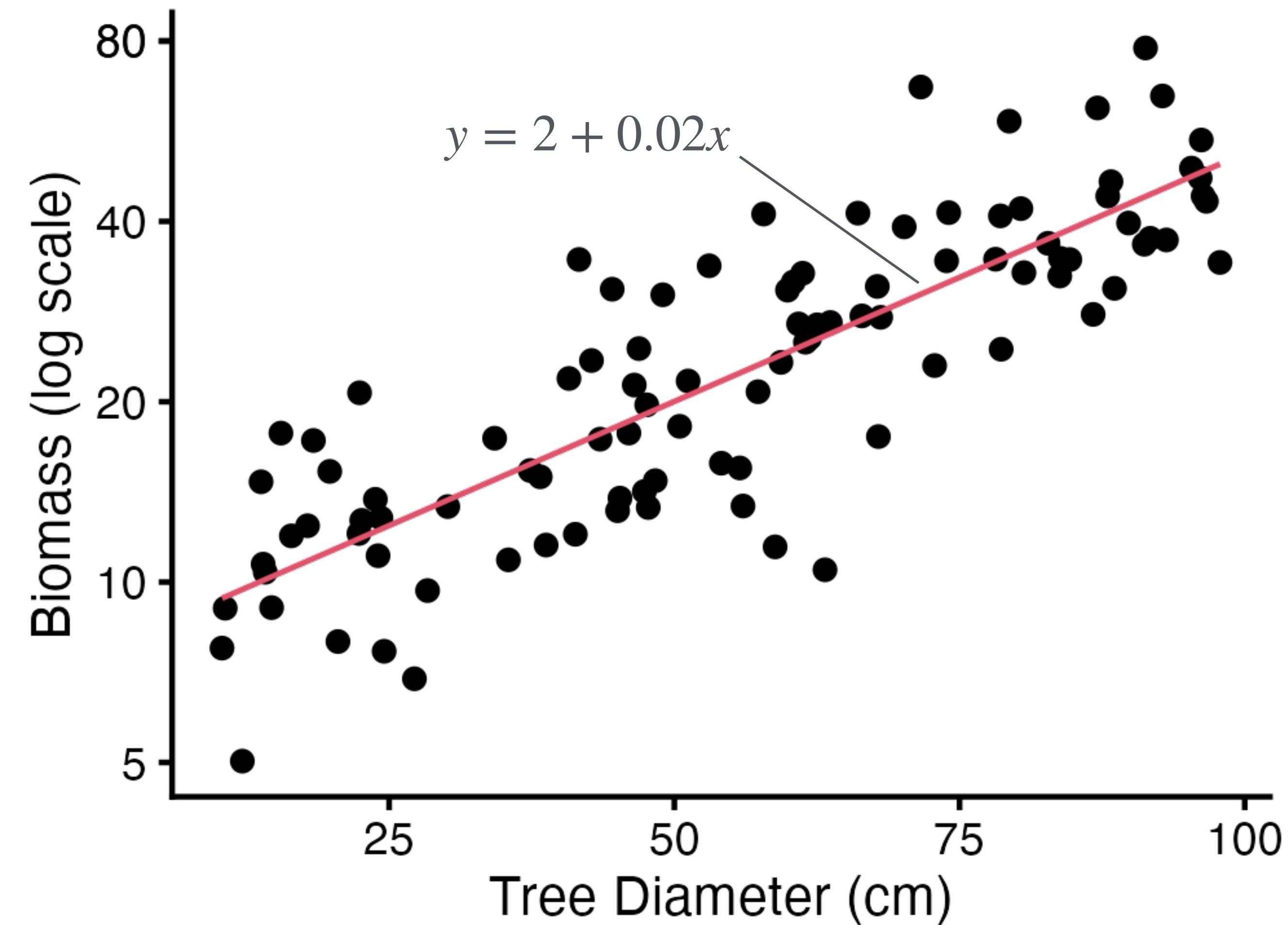


Biomassa por diâmetro

Exemplo de relação não-linear.

- log transformar y antes de ajustar o modelo

```
df = data.frame(diameter, biomass)
stan_fit = stan_glm(log(biomass) ~ diameter,
                     data = df)
```



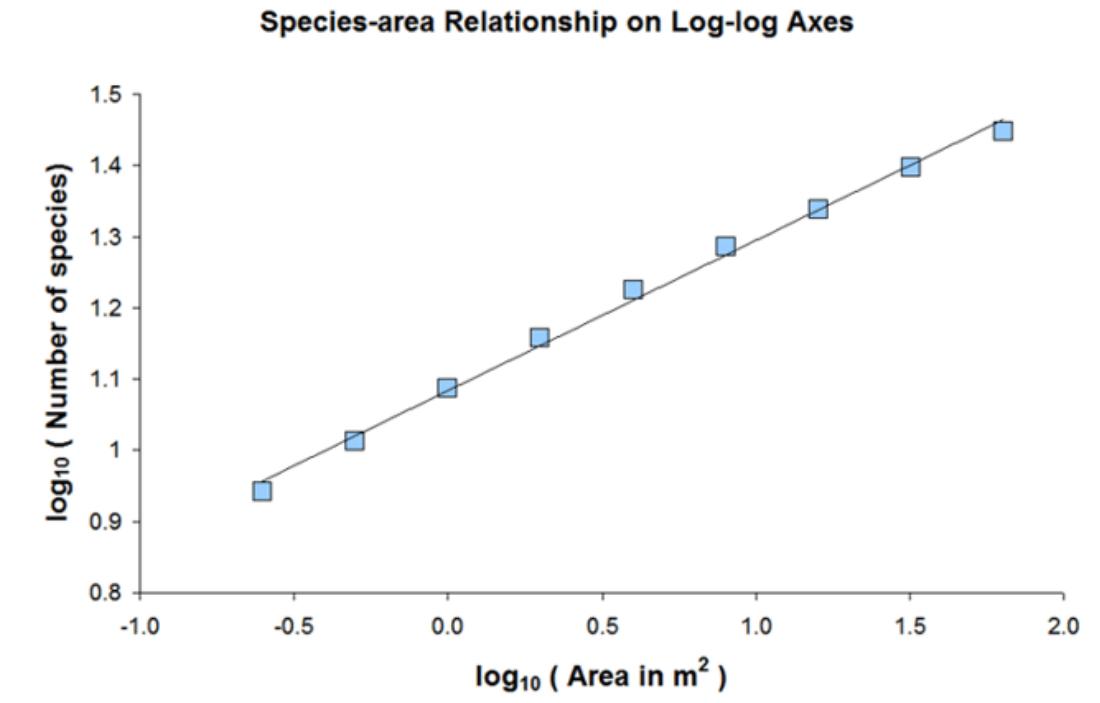
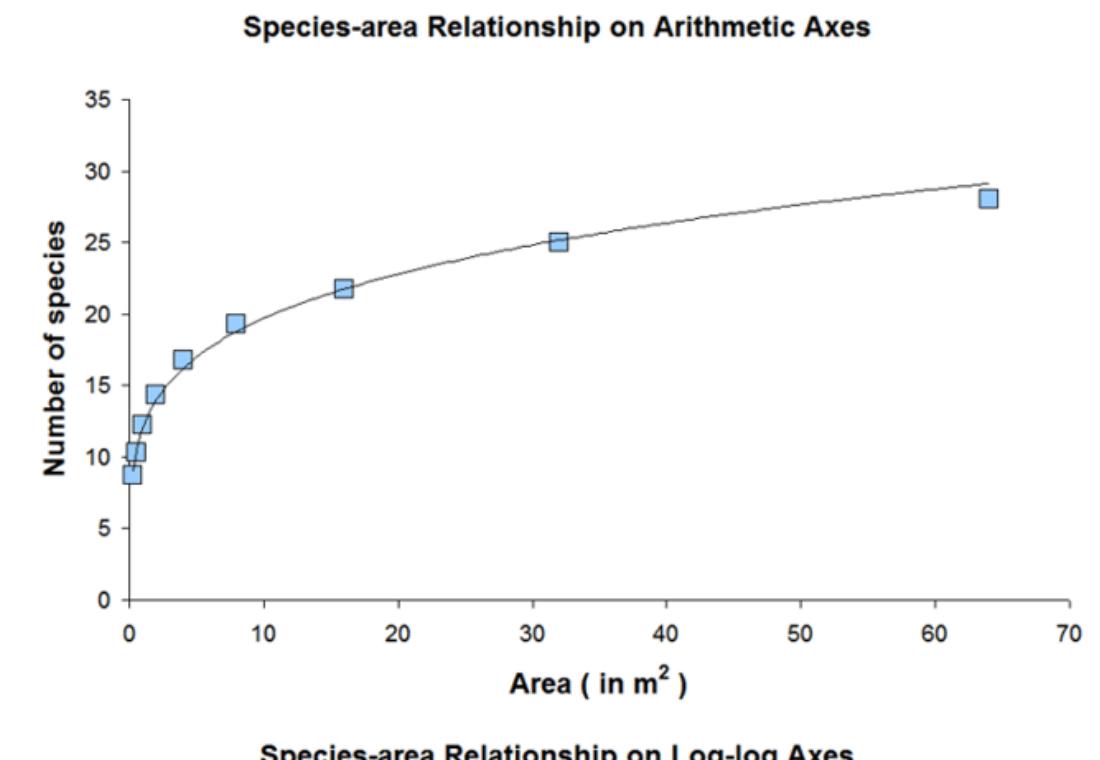
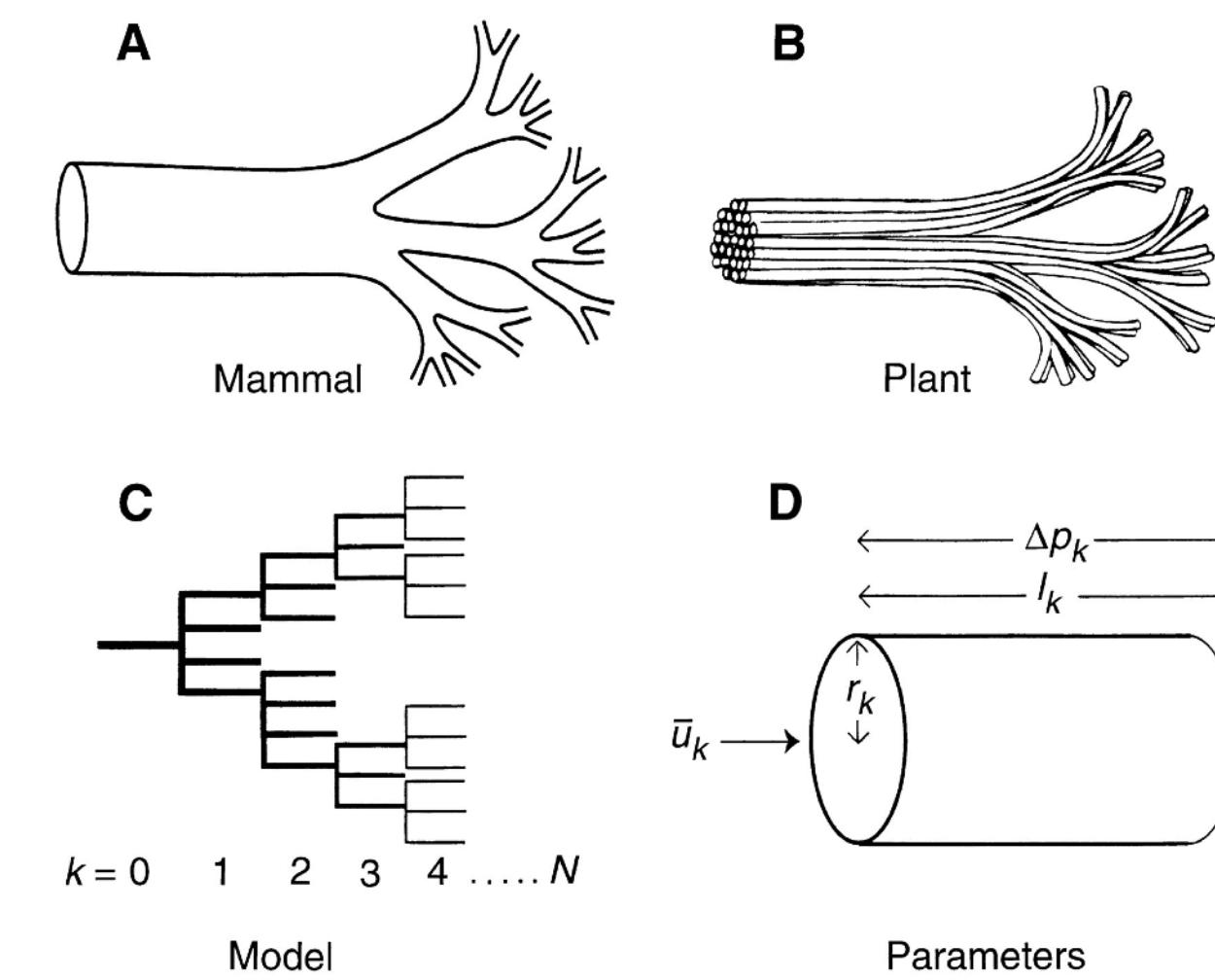
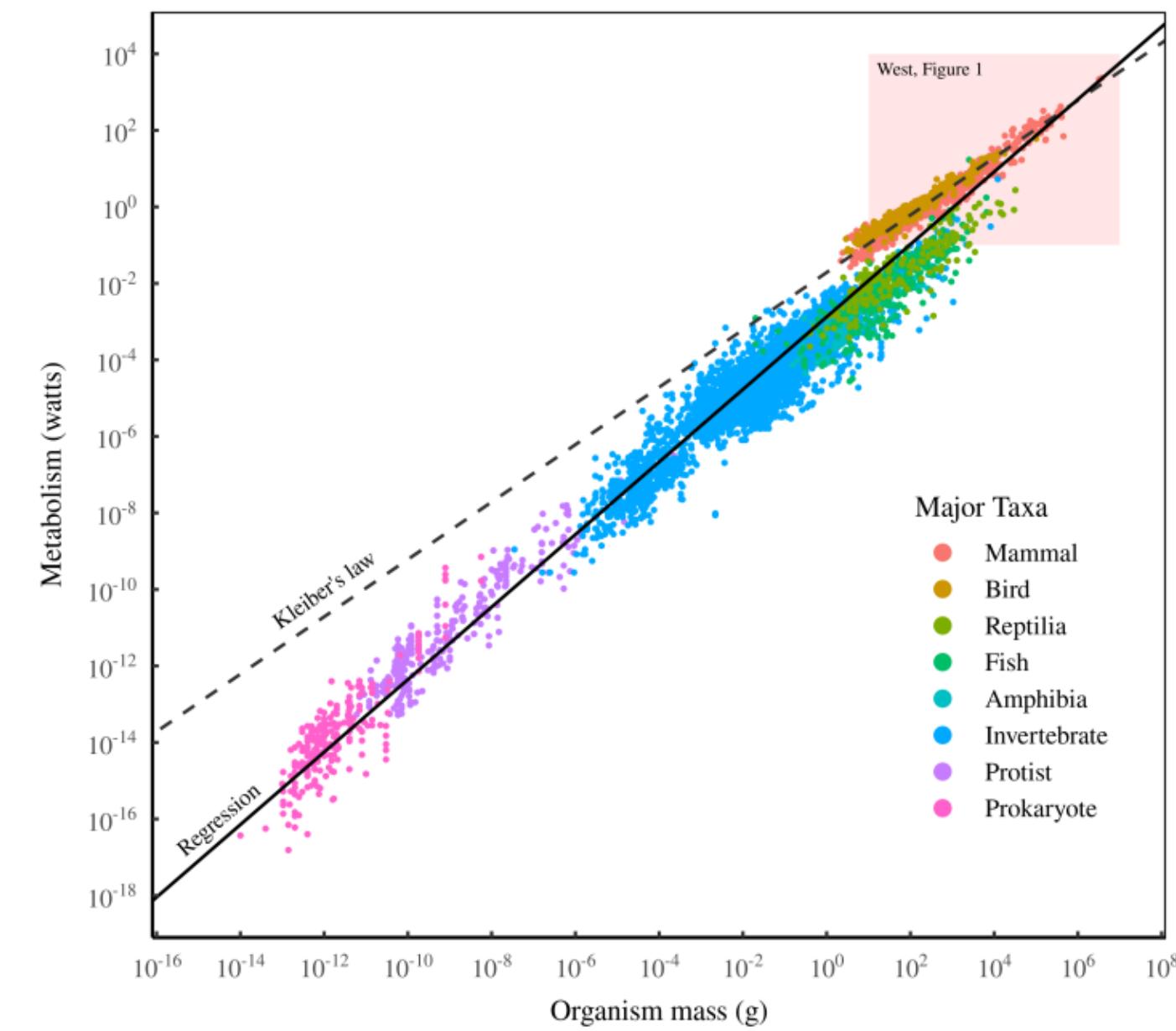
Relações de potência

Regressão Log-log

- Várias relações biológicas têm uma forma de relações de potência.

$$y \propto ax^b$$

- Essas relações aparecem por diferentes motivos:
 - West et al. (1997) atribuem a ubiquidade dessas relações ao padrão fractal de sistemas biológicos.
 - Sempre que um aumento proporcional leva a um aumento proporcional consistente, como em uma relação espécie-área, nós esperamos uma relação de potência.



Relações de potência

Regressão Log-Log

- Muitas relações biológicas têm a forma de uma relação de potência.
- Nós podemos linearizar essas relações usando uma transformação log-log.

$$y \propto ax^b$$



Tirando o log dos dois lados

$$\log(y) \propto \log(ax^b) = \log(a) + \log(x^b) =$$

$$\log(y) \propto \log(a) + b \log(x)$$

Relações de potência

Regressão Log-Log

- Muitas relações biológicas têm a forma de uma relação de potência.
- Nós podemos linearizar essas relações usando uma transformação log-log.
- Nesse modelo, a inclinação β é um estimativo do expoente da relação de potência.
- A interpretação da inclinação é que um aumento de 1% em x leva a um incremento de $\beta\%$ em y .

$$\log(y_i) \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \log(x_i)$$



$$\% \Delta y \approx \beta \% \Delta x$$

Tabela de referência rápida para transformação log

Modelo	Variável dependente	Variável preditora.	Interpretação de β
Level-level	y	x	$\Delta y \approx \beta \Delta x$
Level-log	y	$\log(x)$	$\Delta y \approx (\frac{\beta}{100}) \% \Delta x$
log-level	$\log(y)$	x	$\% \Delta y \approx (100 \beta) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y \approx \beta \% \Delta x$

Regressão linear é muito flexível!

Podemos modificar nossas funções como quisermos

- O modelo linear que estamos usando consiste em fazer os parâmetros de uma distribuição mudar de acordo com alguma função.
- A função mais simples é uma função linear com um único preditor:
- Se tivermos mais preditores, podemos simplesmente adicioná-los na equação linear, acrescentando um parâmetro para cada um.

$$y_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

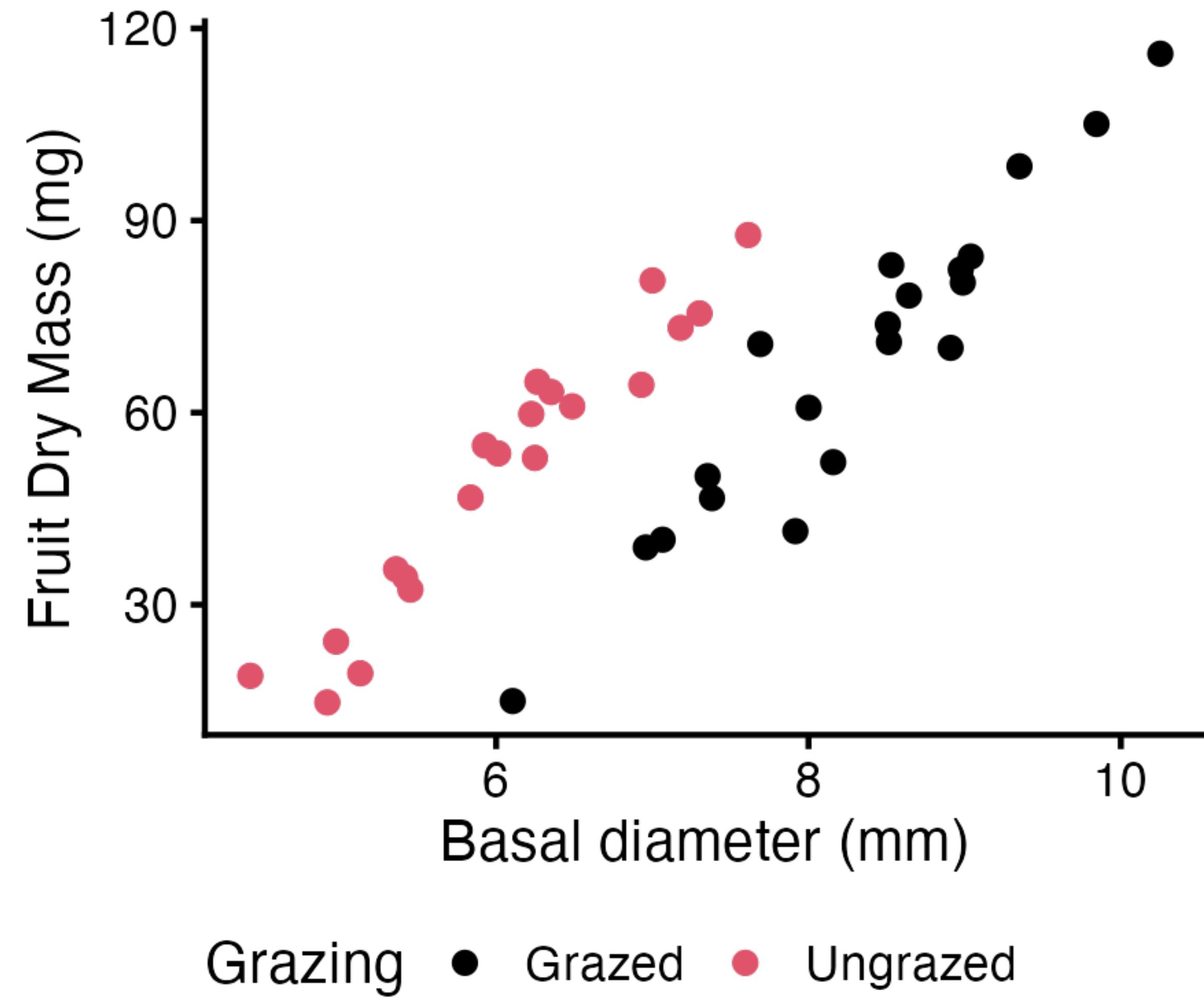


$$y_i \sim N(\mu_i, \sigma)$$

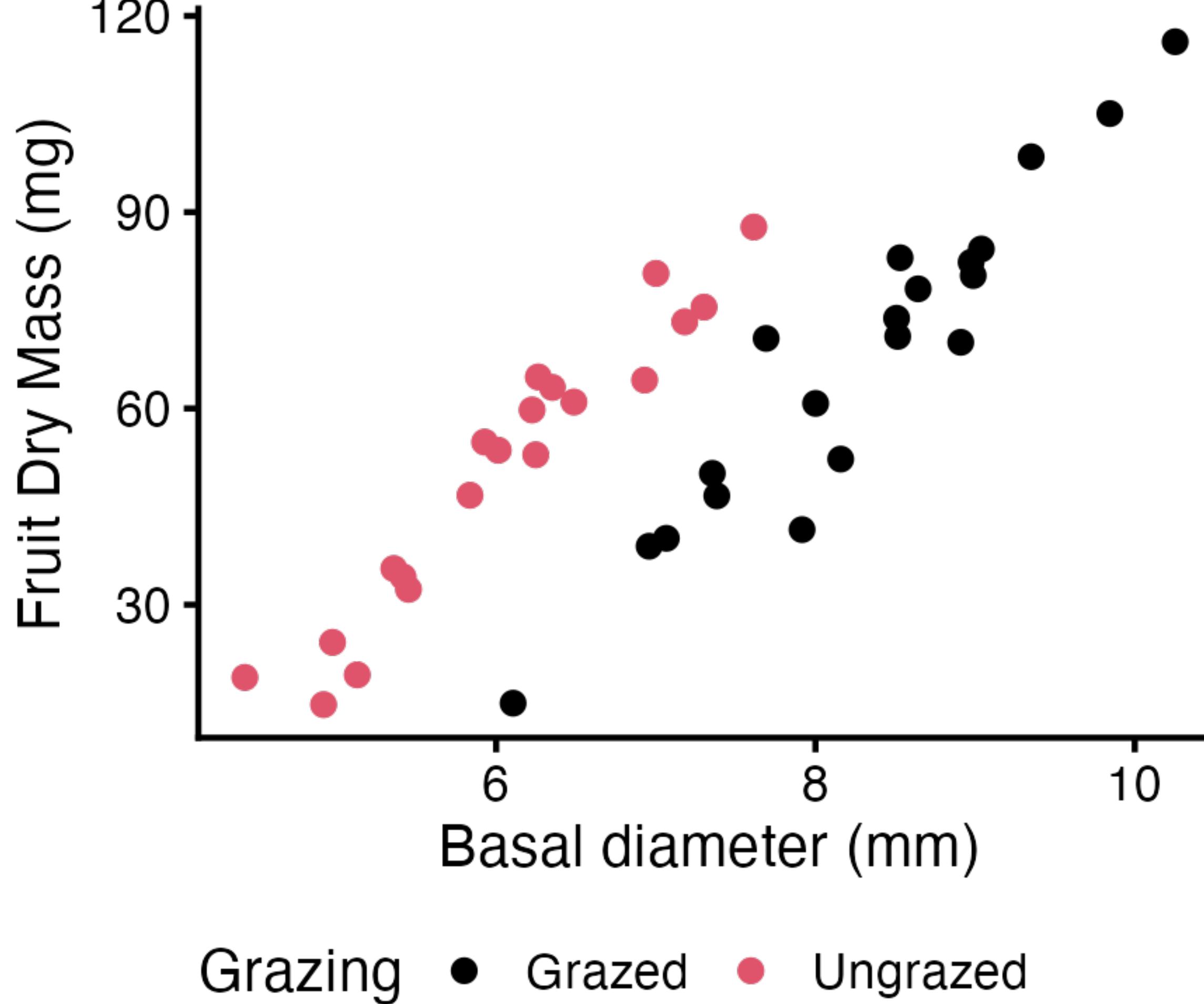
$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Exemplo com mais preditores.

- **Pergunta:** Qual o impacto da herbivoria na aptidão das plantas?
- **Experimento de campo:** 40 índividuos de *Ipomopsis aggregata* foram distribuídas ao acaso em dois tratamentos: **não protegidas da herbivoria** por coelhos e **protegidas da herbivoria** por uma gaiola.
- Variável resposta é o peso seco das frutas de cada planta.
- Variáveis preditoras:
 - Tratamento: fenced ou non-fenced
 - Diâmetro basal de cada planta em milímetros, medido antes do tratamento.



Z-scores



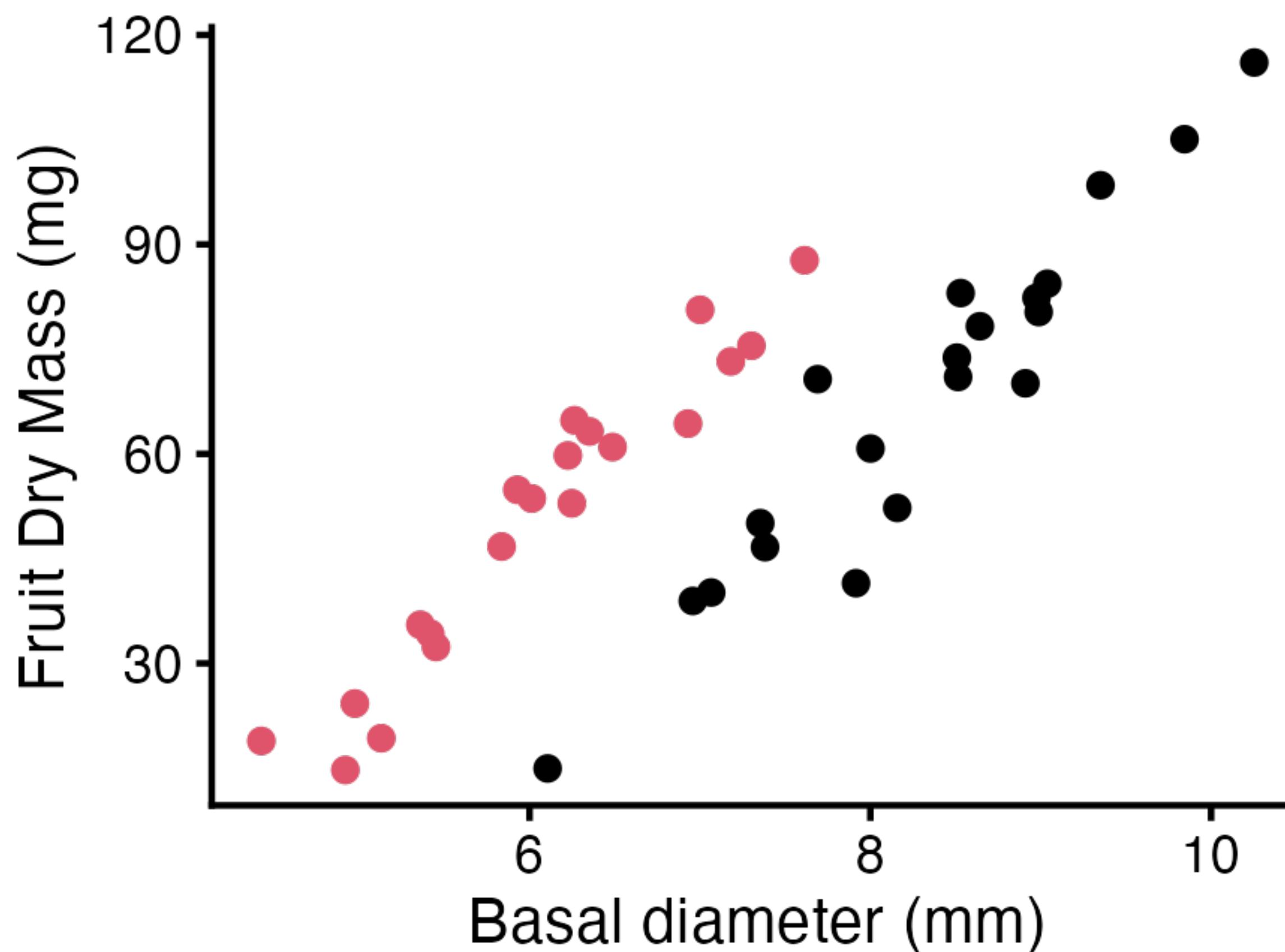
- É sempre bom escalar as variáveis, dividindo pelo desvio padrão e subtraindo a média:

$$\tilde{y}_i = \frac{y_i - \bar{y}}{sd(y)}$$

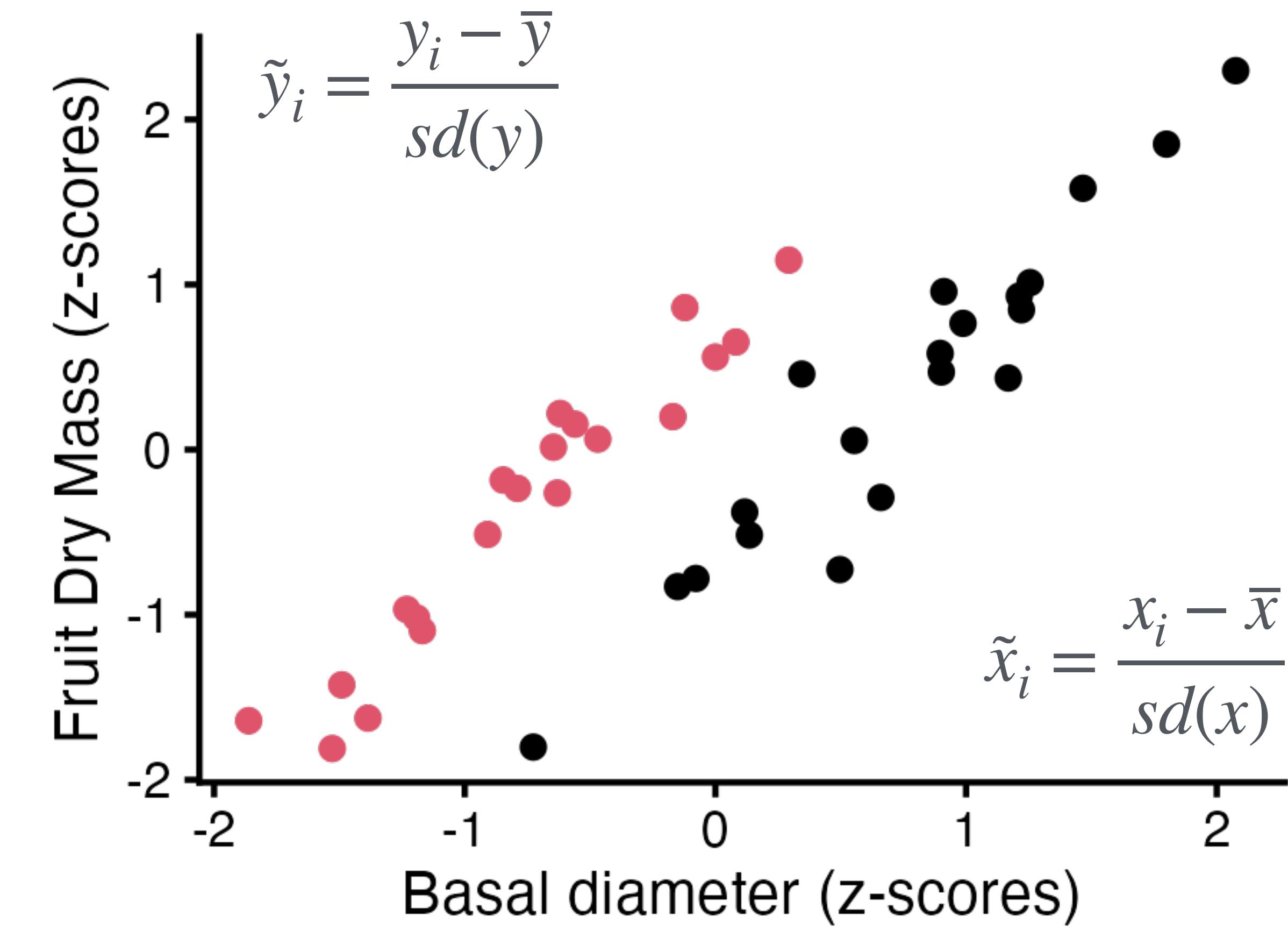
- A quantidade resultante, o **z-score**, uma variável contínua, é uma medida de quantos desvios padrões uma observação está da média.
- O uso de z-scores possibilita a comparação entre coeficientes de diferentes variáveis.
- Como a transformação é linear, podemos sempre calcular os parâmetros na escala original multiplicando pelo desvio padrão.

Variáveis escalonadas

Usar unidade de desvio padrão faz tudo ficar mais simples.



Grazing • Grazed • Ungrazed



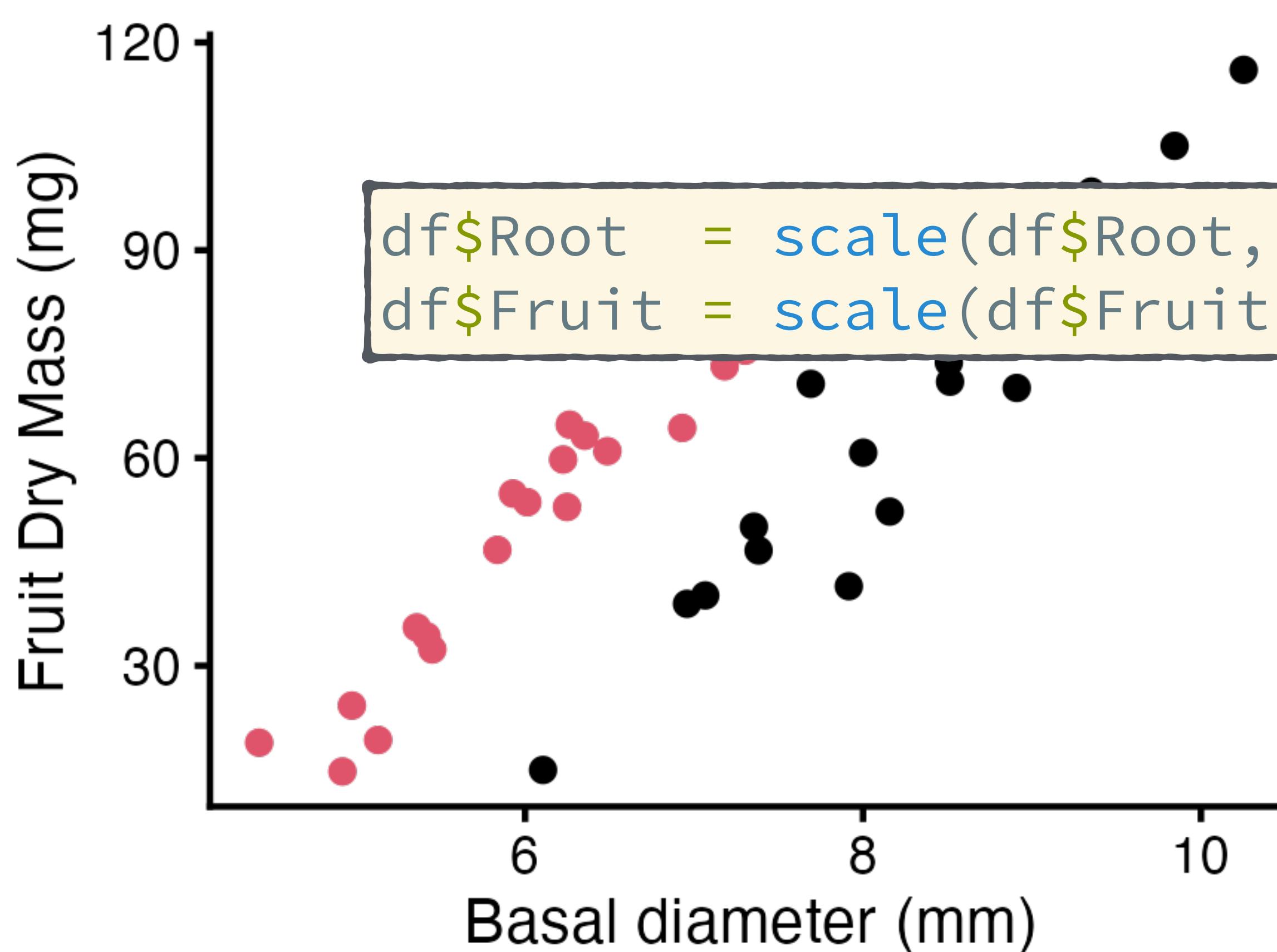
Grazing • Grazed • Ungrazed

$$\tilde{y}_i = \frac{y_i - \bar{y}}{sd(y)}$$

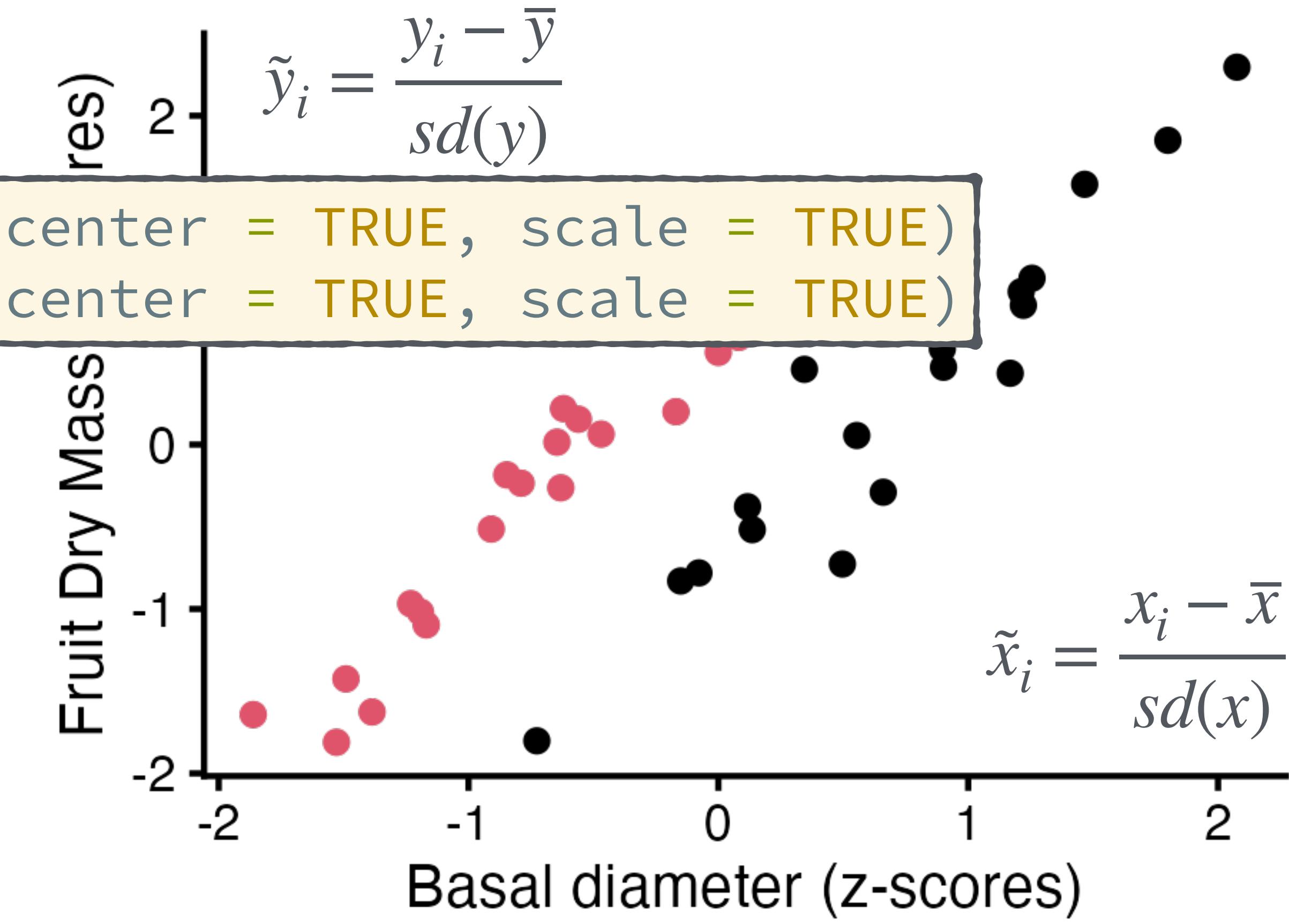
$$\tilde{x}_i = \frac{x_i - \bar{x}}{sd(x)}$$

Variáveis escalonadas

Usar unidade de desvio padrão faz tudo ficar mais simples.



Grazing • Grazed • Ungrazed



Grazing • Grazed • Ungrazed

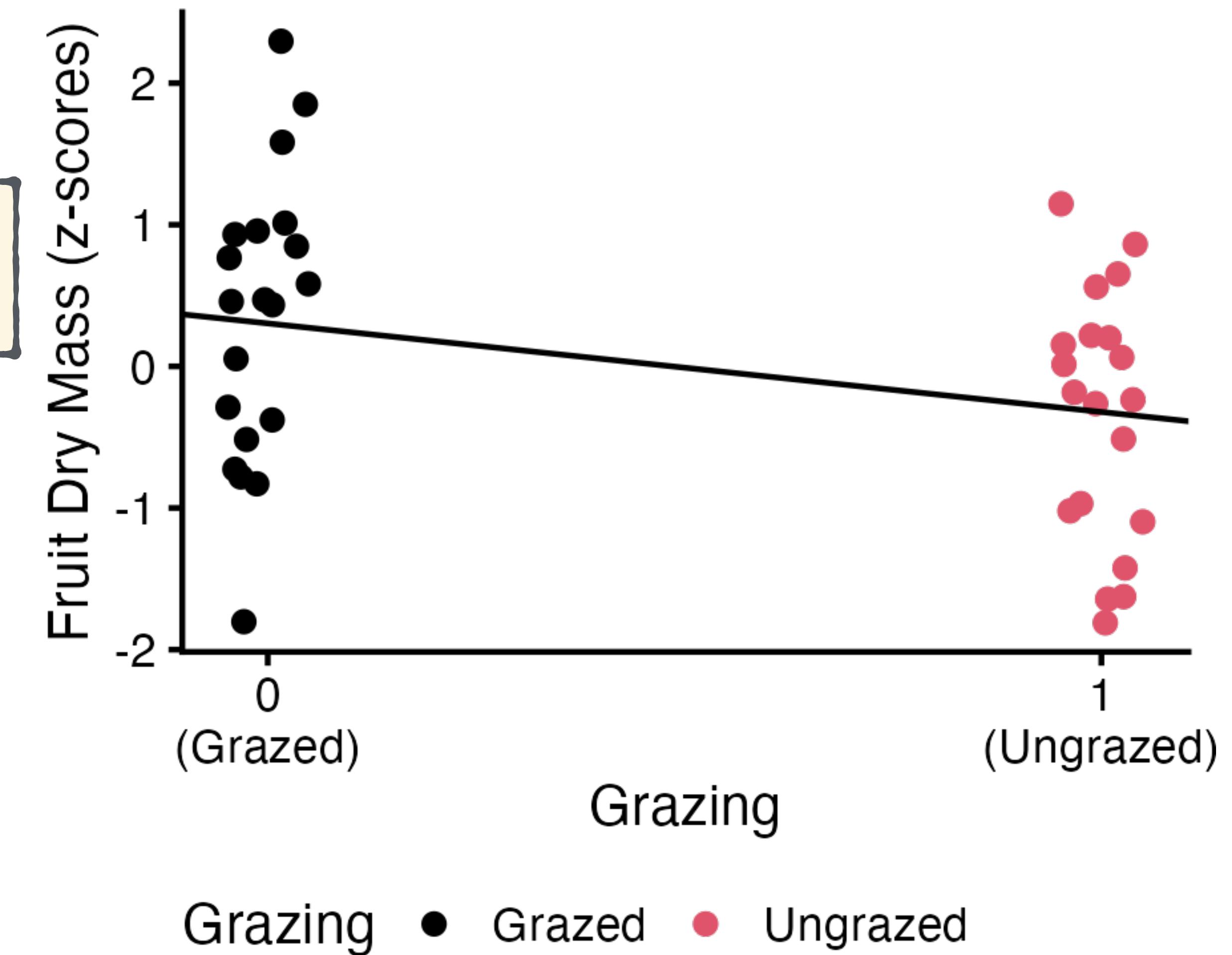
$$\tilde{y}_i = \frac{y_i - \bar{y}}{sd(y)}$$

$$\tilde{x}_i = \frac{x_i - \bar{x}}{sd(x)}$$

Modelo só com o tratamento

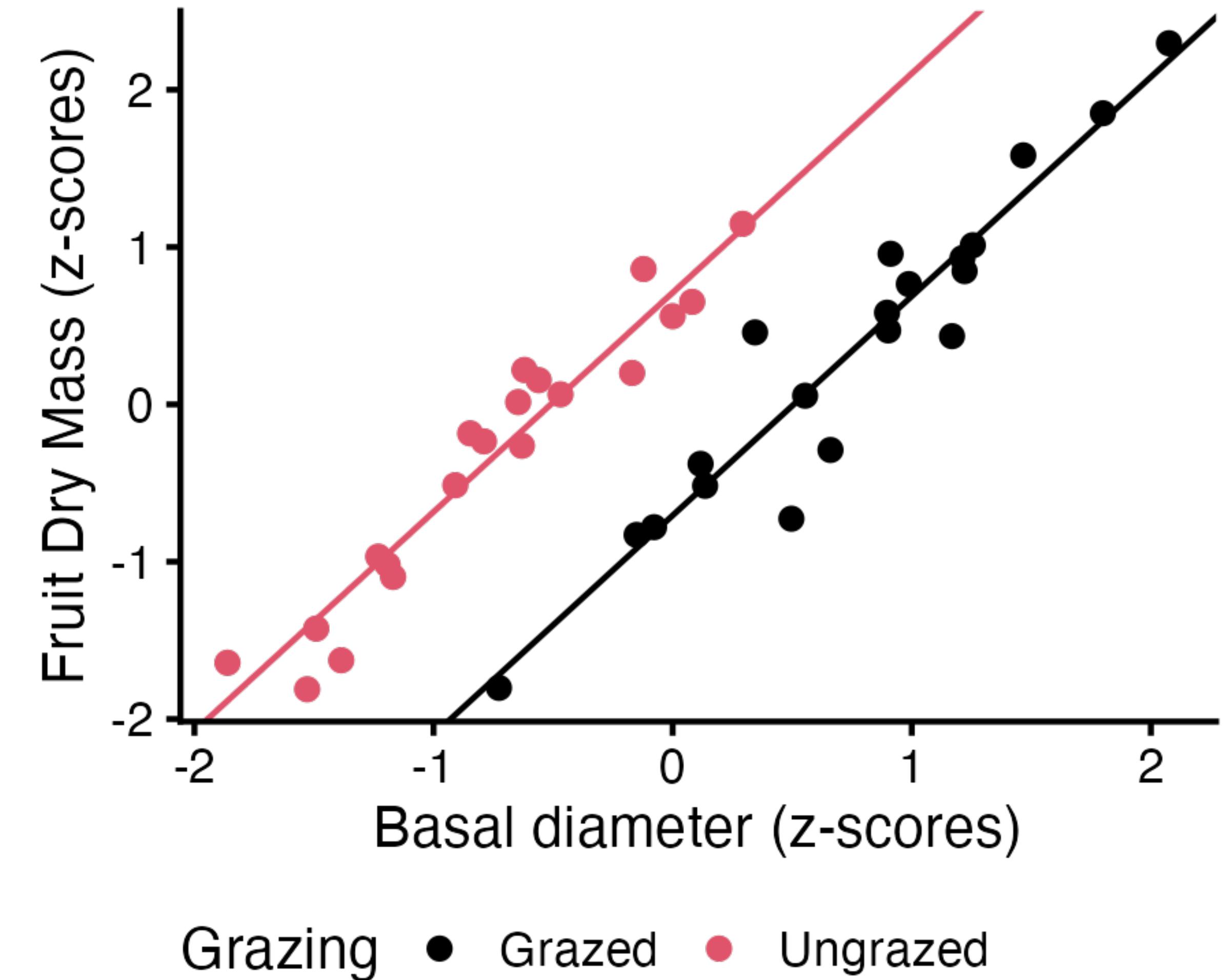
```
df$Grazing0 = ifelse(df$Grazing == "Ungrazed", 1, 0)  
m1 = lm(Fruit ~ Grazing0, data = df)
```

```
> precis(m1, prob = 0.97)  
      mean    sd  1.5% 98.5%  
(Intercept) 0.35  0.21 -0.11  0.81  
Grazing0     -0.69  0.30 -1.34 -0.04
```



O modelo de tratamento e a medida de tamanho

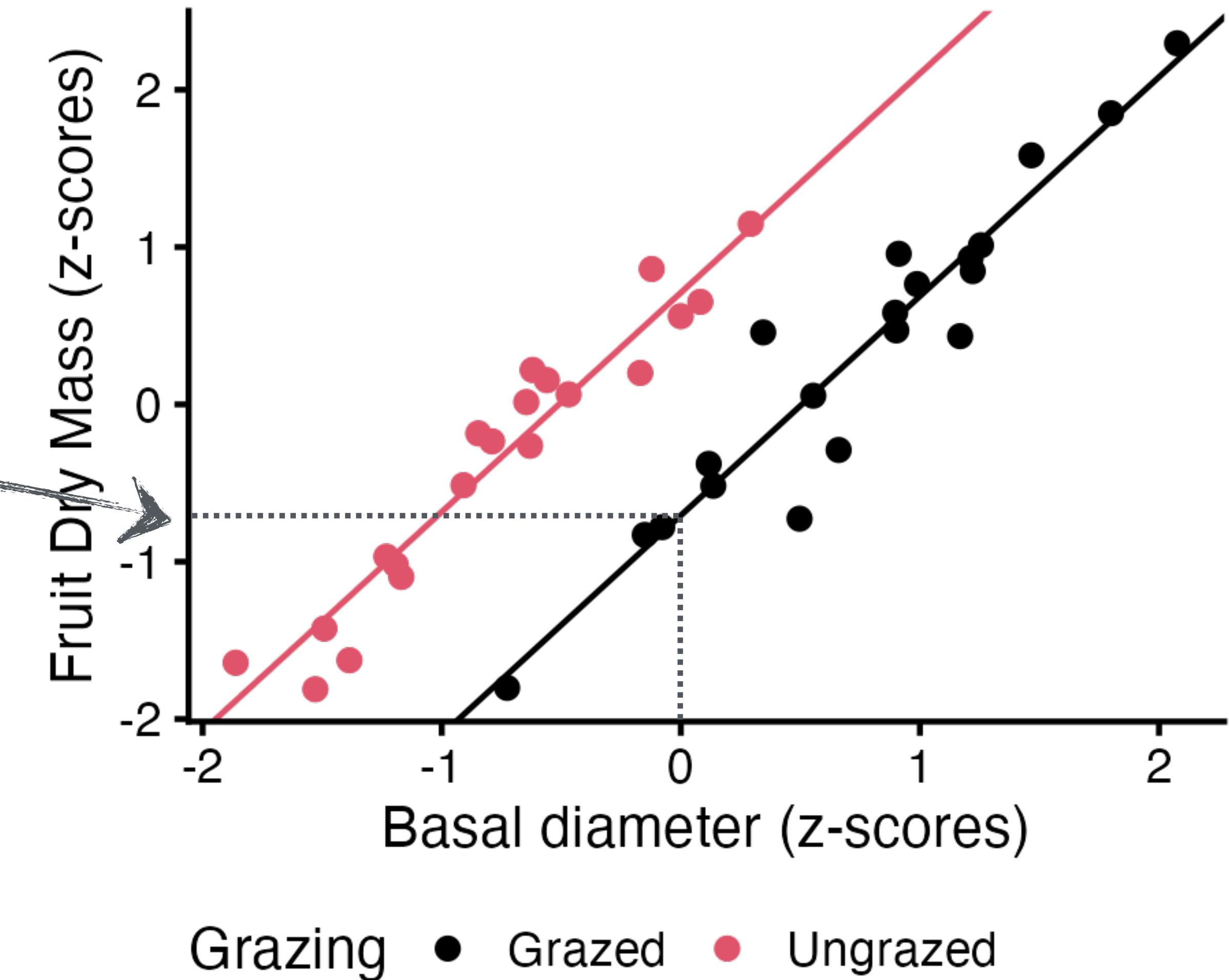
```
> m2 = lm(Fruit ~ Grazing0 + Root, data = df)
> precis(m2, prob = 0.97)
      mean   sd  1.5% 98.5%
(Intercept) -0.73 0.08 -0.91 0.56
Grazing0     1.46 0.14  1.17 1.76
Root         1.41 0.07  1.26 1.56
```



O modelo de tratamento e a medida de tamanho

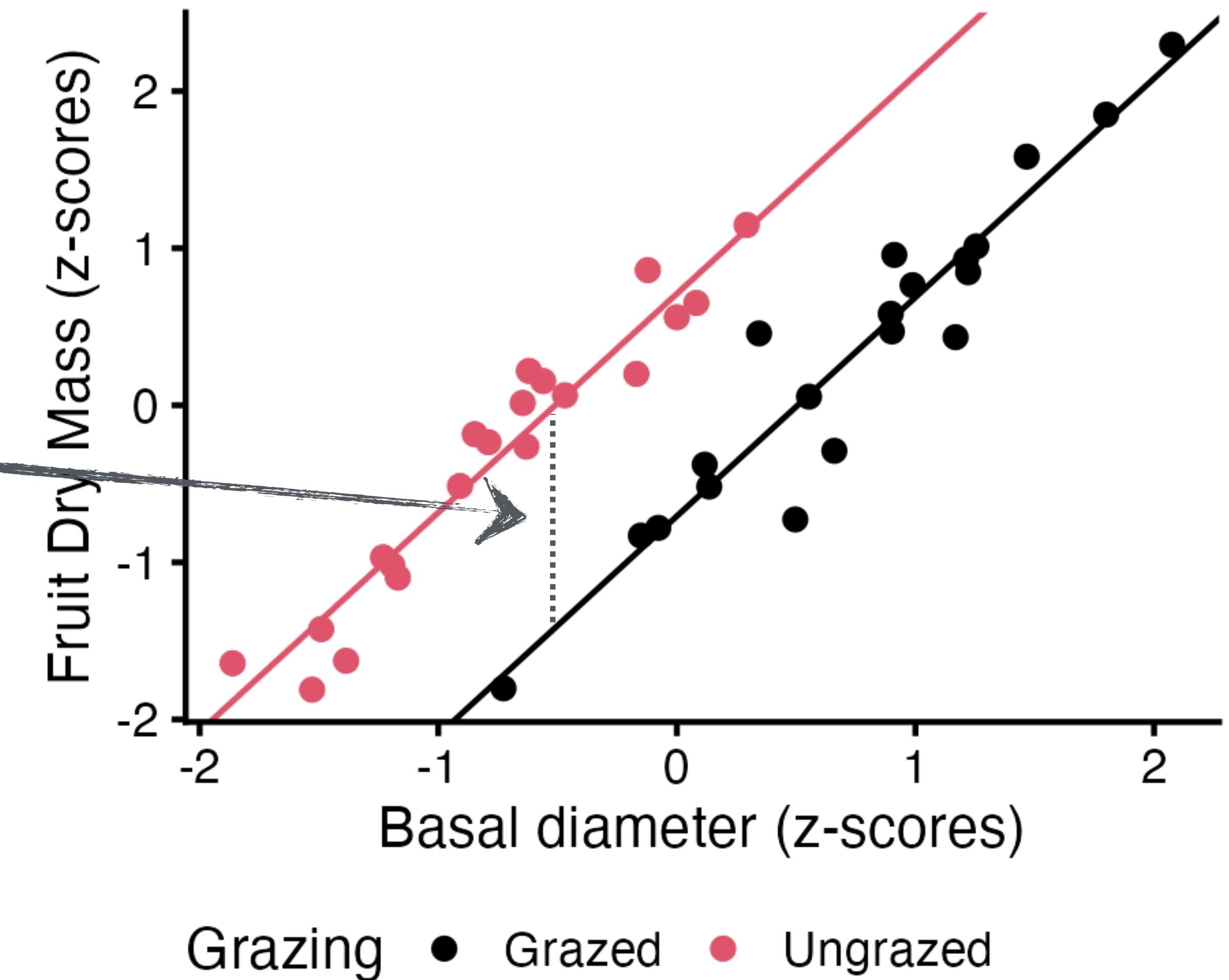
```
> m2 = lm(Fruit ~ Grazing0 + Root, data = df)
> precis(m2, prob = 0.97)

mean      sd   1.5% 98.5%
(Intercept) -0.73 0.08 -0.91 -0.56
Grazing0     1.46 0.14  1.17  1.76
Root         1.41 0.07  1.26  1.56
```



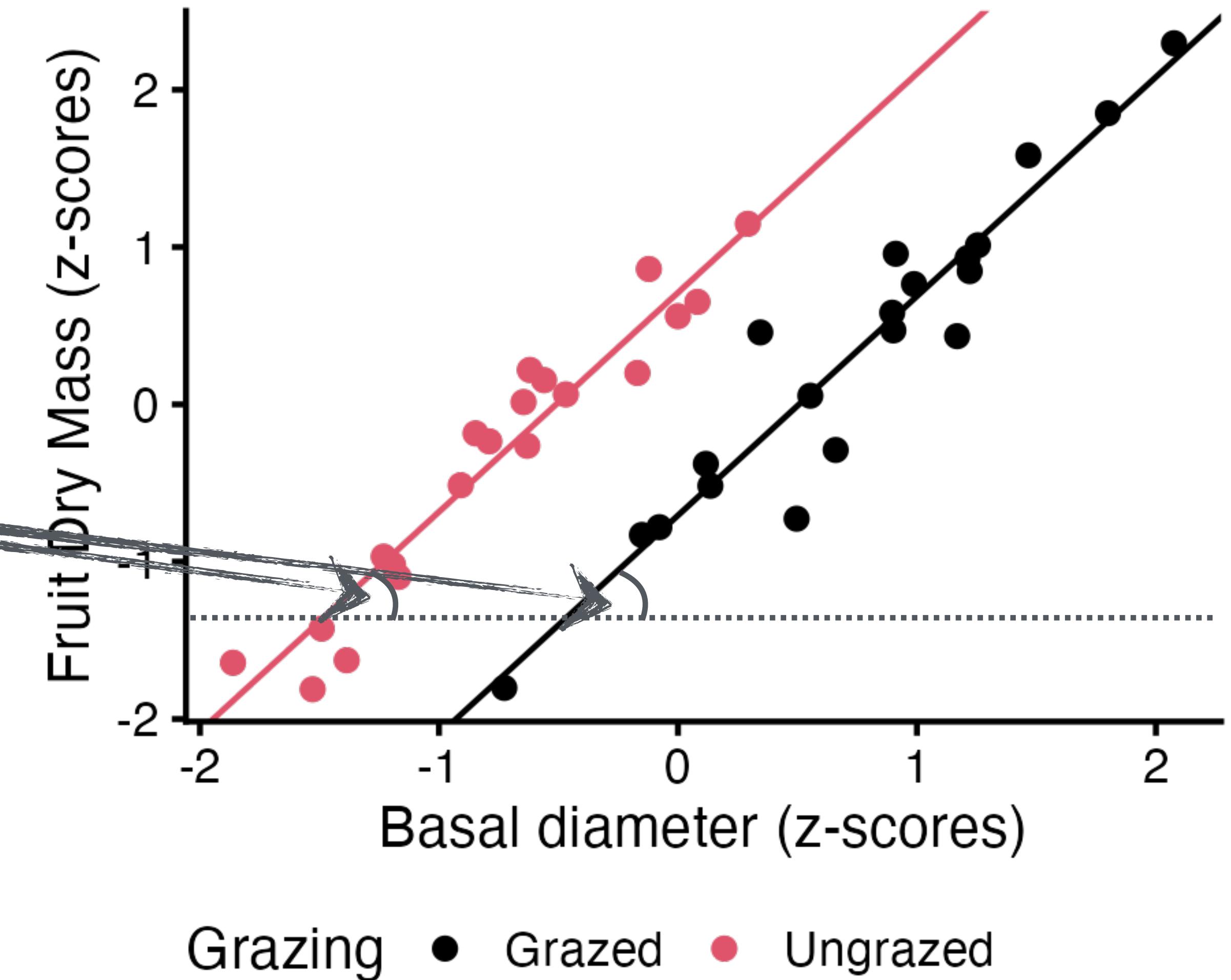
O modelo de tratamento e a medida de tamanho

```
> m2 = lm(Fruit ~ Grazing0 + Root, data = df)
> precis(m2, prob = 0.97)
      mean   sd  1.5% 98.5%
(Intercept) -0.73  0.08 -0.91  0.56
Grazing0     1.46  0.14  1.17  1.76
Root         1.41  0.07  1.26  1.56
```



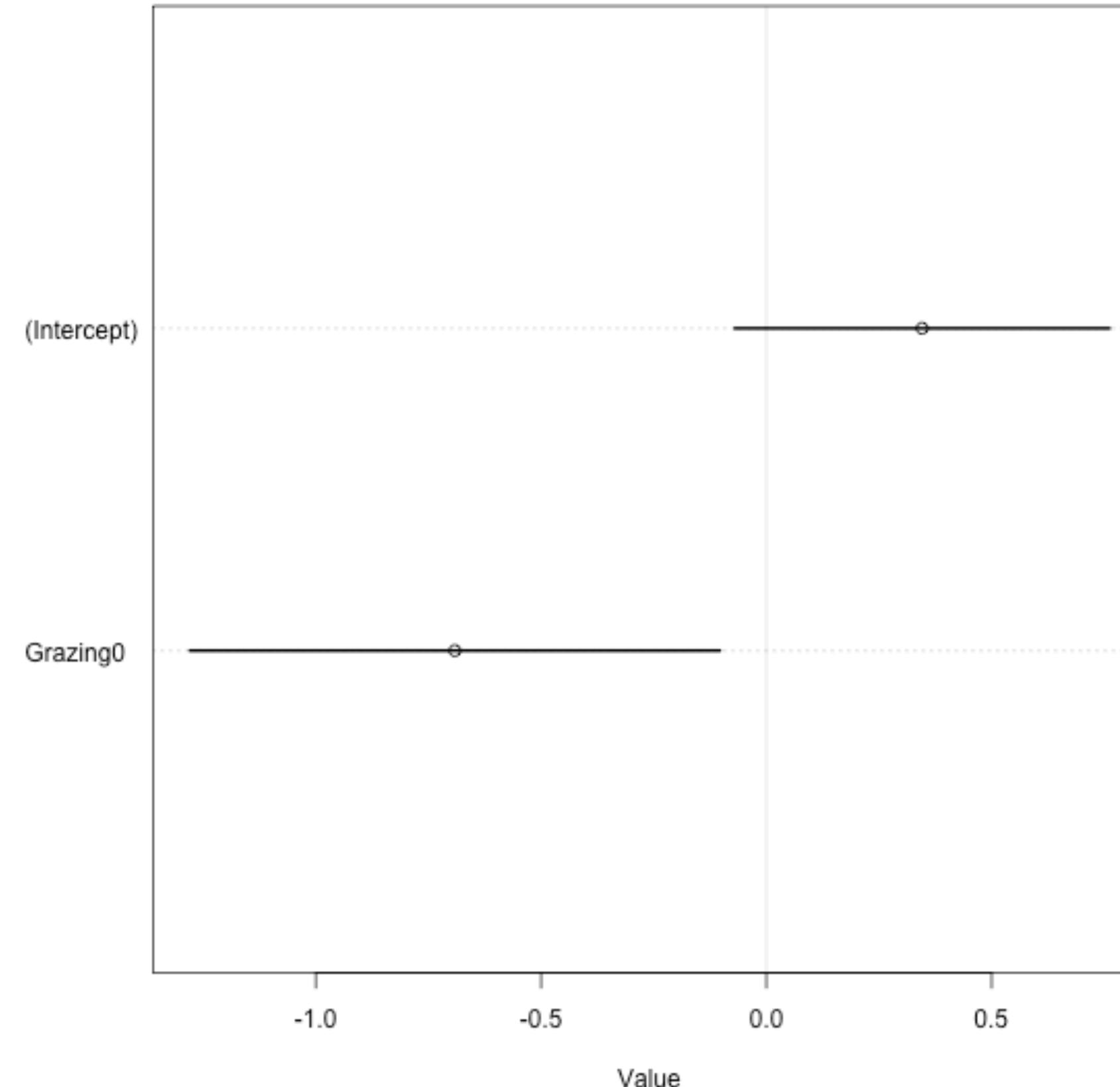
O modelo de tratamento e a medida de tamanho

```
> m2 = lm(Fruit ~ Grazing0 + Root, data = df)
> precis(m2, prob = 0.97)
      mean   sd  1.5% 98.5%
(Intercept) -0.73 0.08 -0.91  0.56
Grazing0     1.46 0.14  1.17  1.76
Root        1.41 0.07  1.26  1.56
```

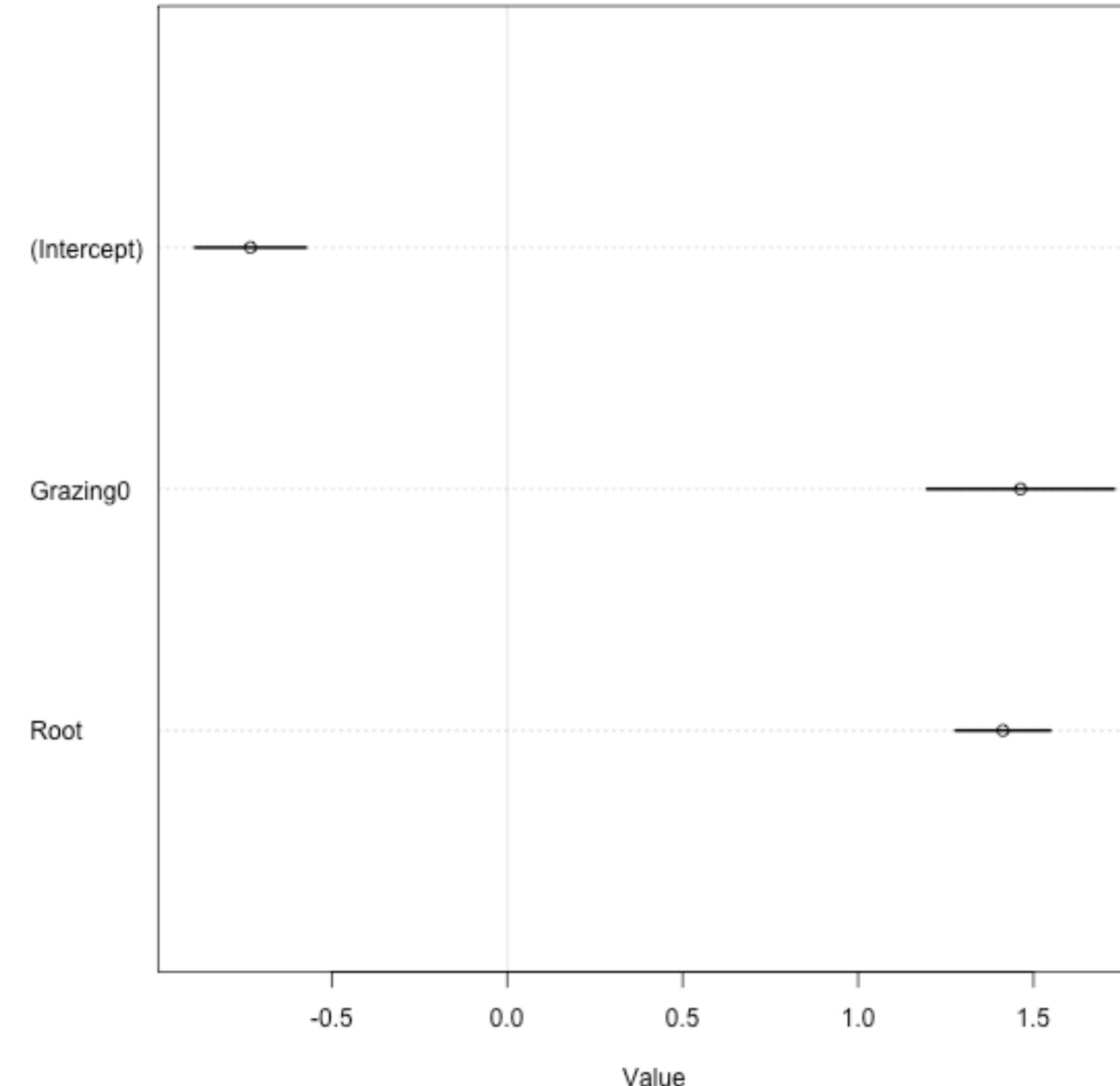


Comparando os dois modelos

$$y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta x_i$$



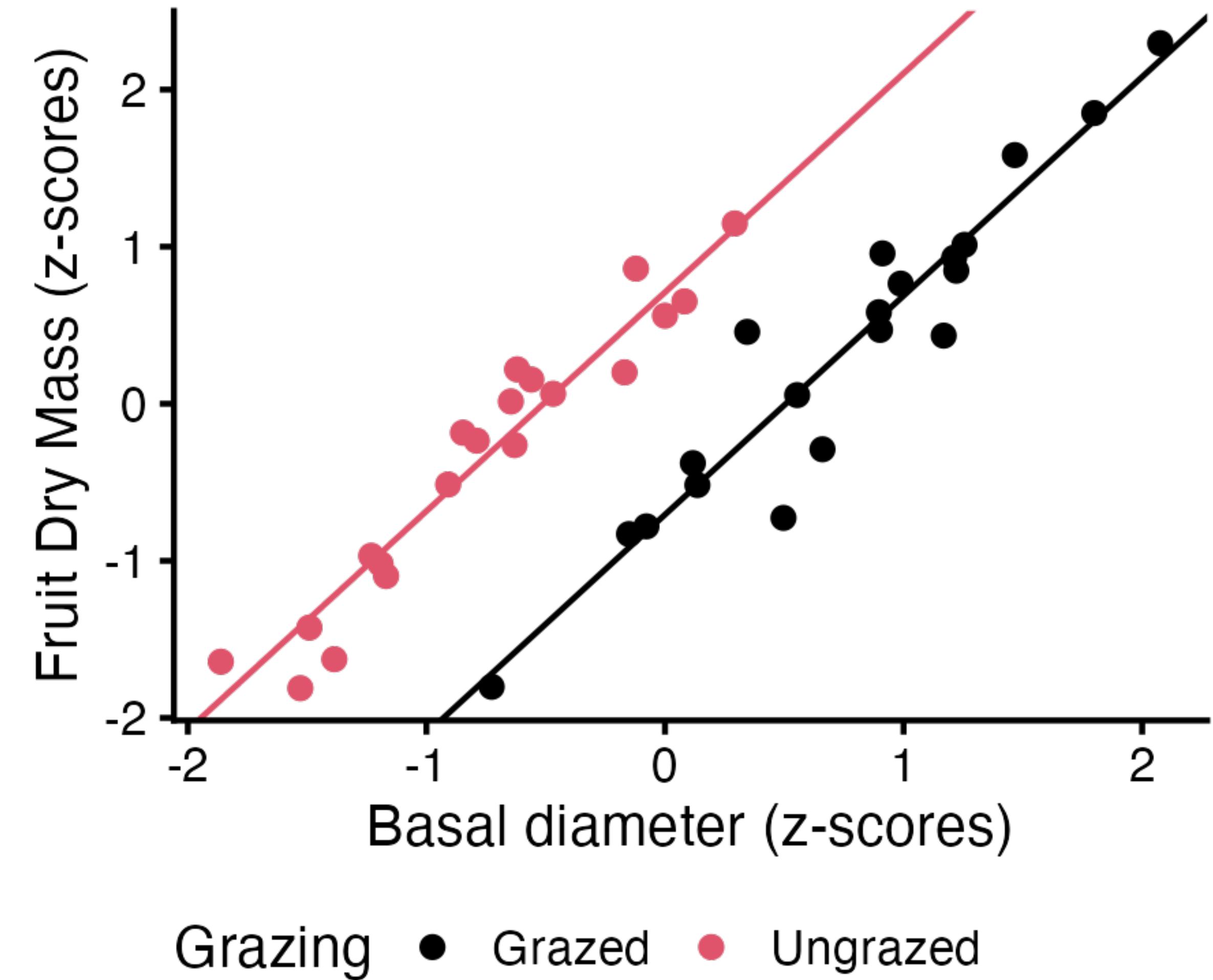
$$y_i \sim N(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2}$$



Por que os coeficientes mudam?

Régressão múltipla como uma estratégia de controle

- Coeficientes são comparações e adicionar preditores à regressão tem o mesmo efeito de **estratificar** os dados.
- O objetivo de adicionar mais preditores é: fazer comparações entre **observações equivalentes**:
 1. Efeito do tratamento para plantas com o mesmo tamanho?
 2. Qual o efeito do tamanho, dado um tratamento?

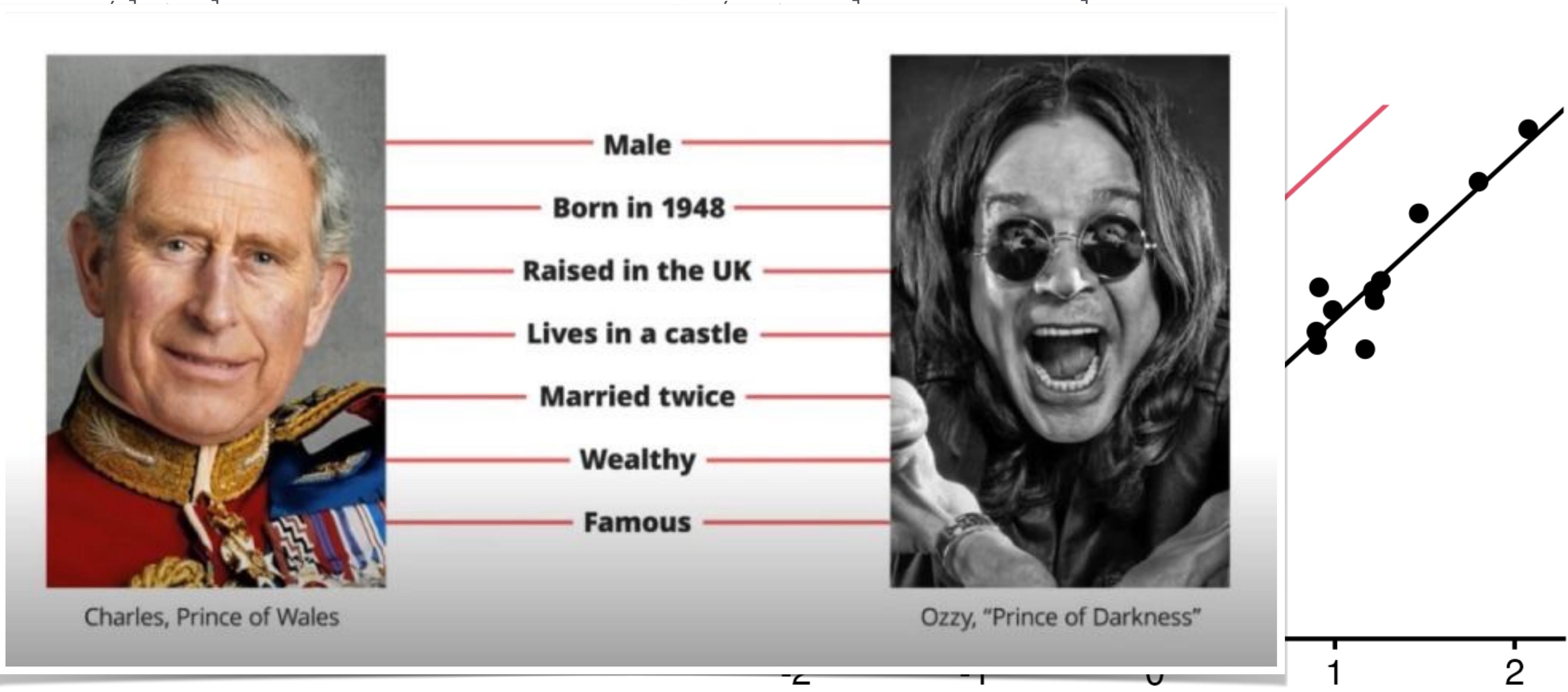


Por que os coeficientes mudam?

Regressão

- Coeficiente preditores da classe de **estratificação**
- O objetivo da regressão é fazer comparações entre classes **equivalentes**

1. Efeito do tratamento mesmo
2. Qual o efeito do tamanho, dado um tratamento?



Grazing • Grazed ● Ungrazed

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

$$x | z = \text{residuals} (\text{Im}(x \sim 1 + z))$$

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

$$x | z = \text{residuals} (\text{Im}(x \sim 1 + z))$$

$$z | x = \text{residuals} (\text{Im}(z \sim 1 + x))$$

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

$$x | z = \text{residuals} (\text{Im}(x \sim 1 + z))$$

$$z | x = \text{residuals} (\text{Im}(z \sim 1 + x))$$

$$y \sim \alpha + \beta_1 x | z$$

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

$$x | z = \text{residuals} (\text{Im}(x \sim 1 + z))$$

$$z | x = \text{residuals} (\text{Im}(z \sim 1 + x))$$

$$y \sim \alpha + \beta_1 x | z$$

$$y \sim \alpha + \beta_2 z | x$$

Relação entre regressão múltipla e regressão de resíduos

Os coeficientes de uma regressão múltipla podem ser obtidos usando regressões simples.

- Quando temos dois ou mais preditores, podemos interpretar o coeficiente associado a cada um como sendo o coeficiente de regressão da variável resposta no resíduo de cada preditor como função dos outros preditores.
- Que?
- Regressão múltipla só dá resultados diferentes quando os preditores são correlacionados!

$$y \sim \alpha + \beta_1 x + \beta_2 z$$

$$x | z = \text{residuals} (\text{Im}(x \sim 1 + z))$$

$$z | x = \text{residuals} (\text{Im}(z \sim 1 + x))$$

$$y \sim \alpha + \beta_1 x | z$$

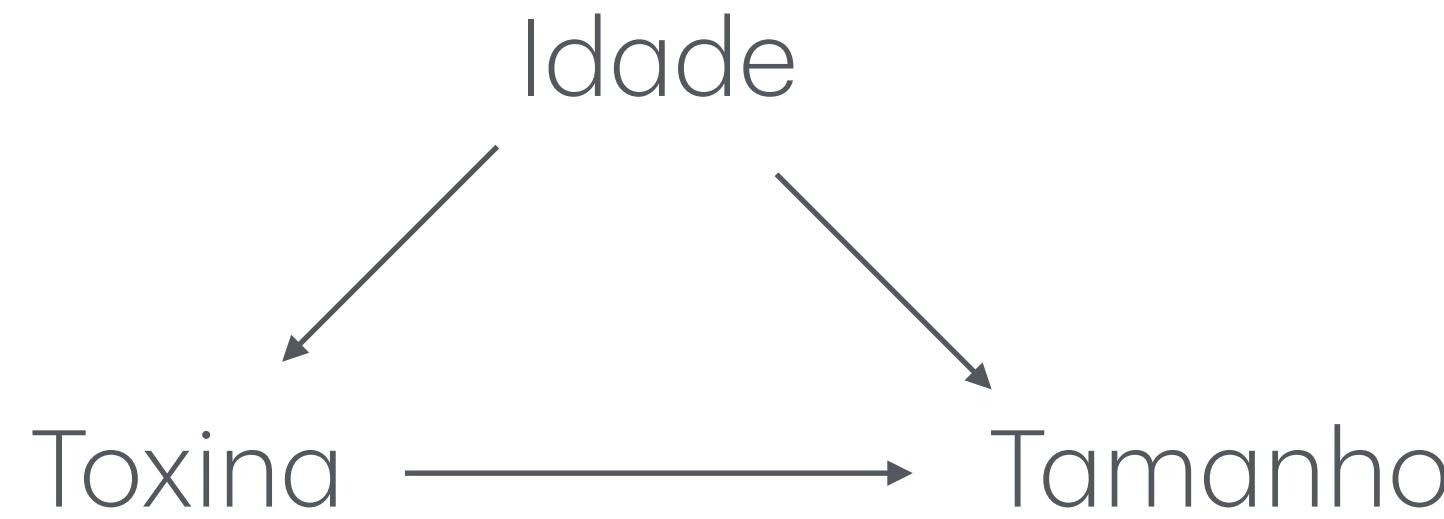
$$y \sim \alpha + \beta_2 z | x$$

Usamos simulações para entender a regressão múltipla

- Simulações são uma ferramenta poderosa para entender os nossos modelos.
- **Use simulações de forma liberal! Use simulações sempre!**
- Se pergunte:
 - Qual o processo gerador dos meus dados?
 - Qual modelo representa esse processo?
- Simule dados usando esse modelo, depois tente usar o modelo estatístico para ajustar esses dados simulados.

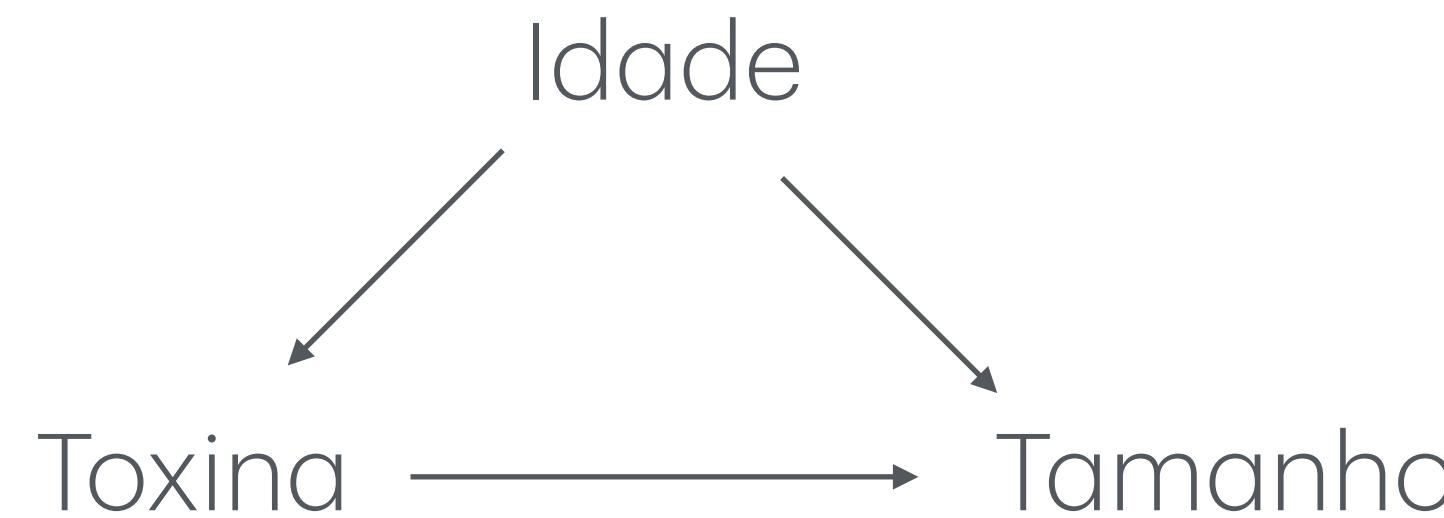
- Exemplo:

- **Pergunta:** Qual o efeito de exposição a uma toxina no tamanho dos indivíduos de uma população?
- Variável de confusão: A idade afeta tanto a exposição à toxina quanto o tamanho.



Usamos simulações para entender a regressão múltipla

- **Pergunta:** Qual o efeito de exposição a uma toxina no tamanho dos indivíduos de uma população?
- Variável de confusão: A idade afeta tanto a exposição à toxina quanto o tamanho.



Modelo possível:

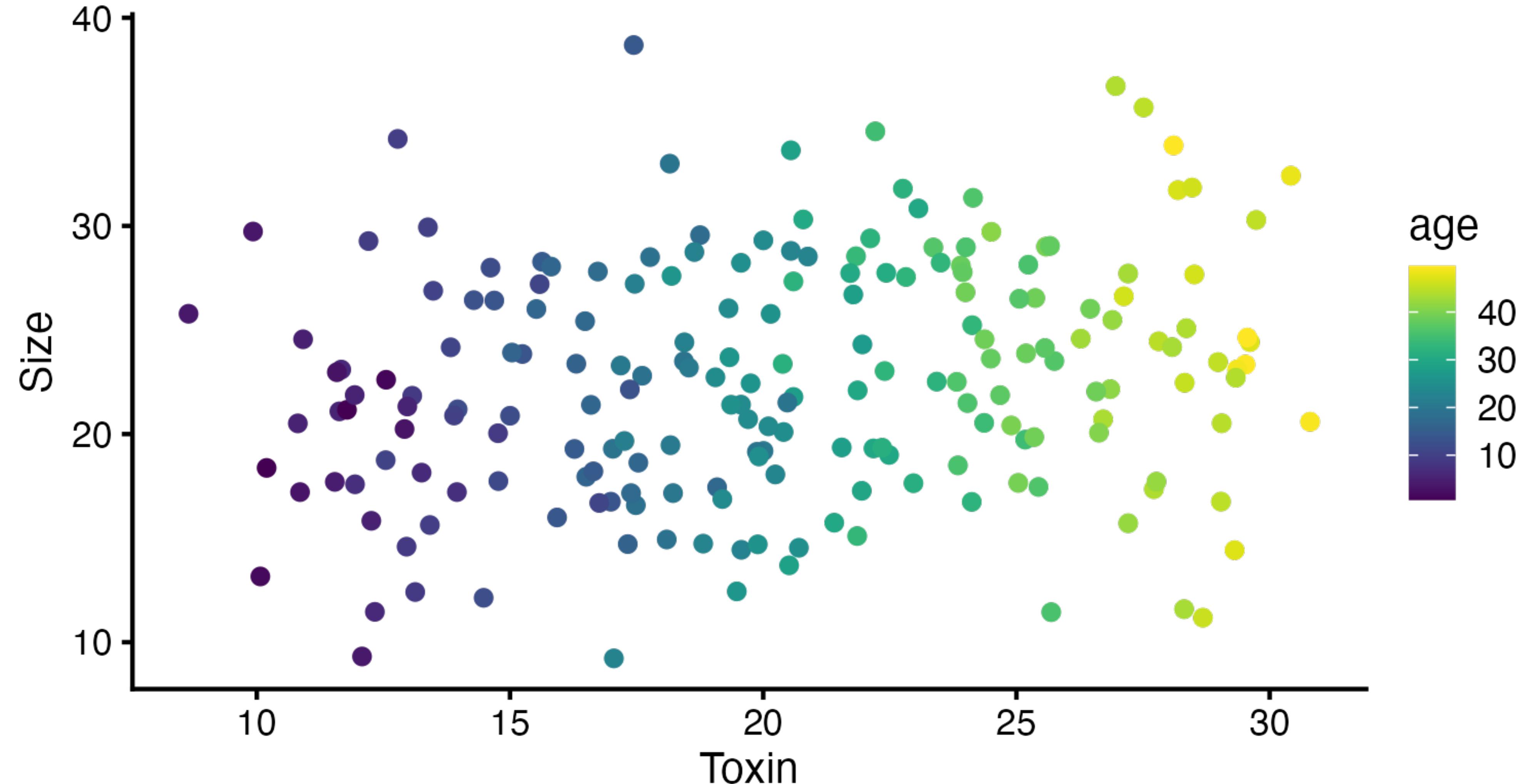
$$age_i = Uniform(0, 50)$$

$$toxin_i = Normal(10 + 0.4 \times age_i, 1)$$

$$size_i = Normal(30 - 1 \times toxin_i + 0.5 \times age_i, 5)$$

```
set.seed(1)
age <- runif(200, 0, 50)
toxin <- rnorm(200, 10 + 0.4*age, 1)
size = 30 - 1*toxin + 0.5*age + rnorm(200, 0, 5)
df = data.frame(age = age,
                 toxin = toxin,
                 size = size)
```

Dados simulados



Centralizar os dados e ajustar o modelo

```
df0 = df
df0$age = scale(df0$age, scale = FALSE)
df0$toxin = scale(df0$toxin, scale = FALSE)
df0$size = scale(df0$size, scale = FALSE)
```

$$size_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = a + b \times toxin_i + c \times age_i$$

```
rt_fit = ulam(alist(size ~ normal(mu, sigma),
                     mu <- a + b*toxin + c*age,
                     a ~ normal(0, 0.3),
                     b ~ normal(0, 1),
                     c ~ normal(0, 1),
                     sigma ~ exponential(1)),
               data = df0, chains = 4, cores = 4)
```

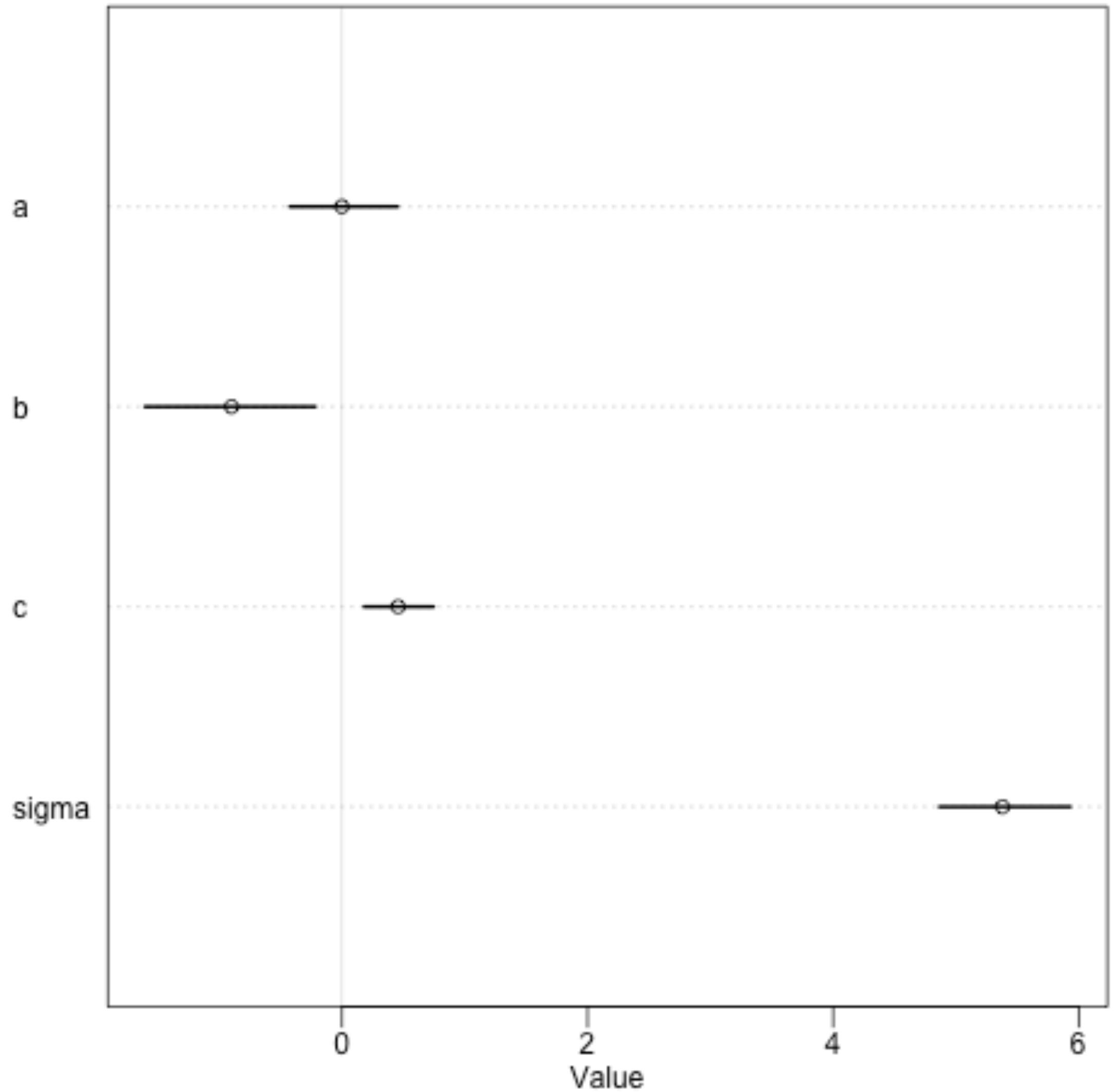
Ajuste do modelo

$$size_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = a + b \times toxin_i + c \times age_i$$

```
> precis(rt_fit, prob = 0.95)
    mean   sd 2.5% 97.5% rhat ess_bulk
a     0.00 0.22 -0.43  0.45     1 1679.38
b    -0.90 0.35 -1.60 -0.22     1  821.89
c     0.46 0.14  0.18  0.74     1  791.07
sigma 5.38 0.27  4.87  5.93     1 1225.56

> plot(precis(rt_fit, prob = 0.95))
```



Estimativas e valores reais.

Simulation:

$$age_i = Uniform(0, 50)$$

$$toxin_i = Normal(10 + 0.4 \times age_i, 1)$$

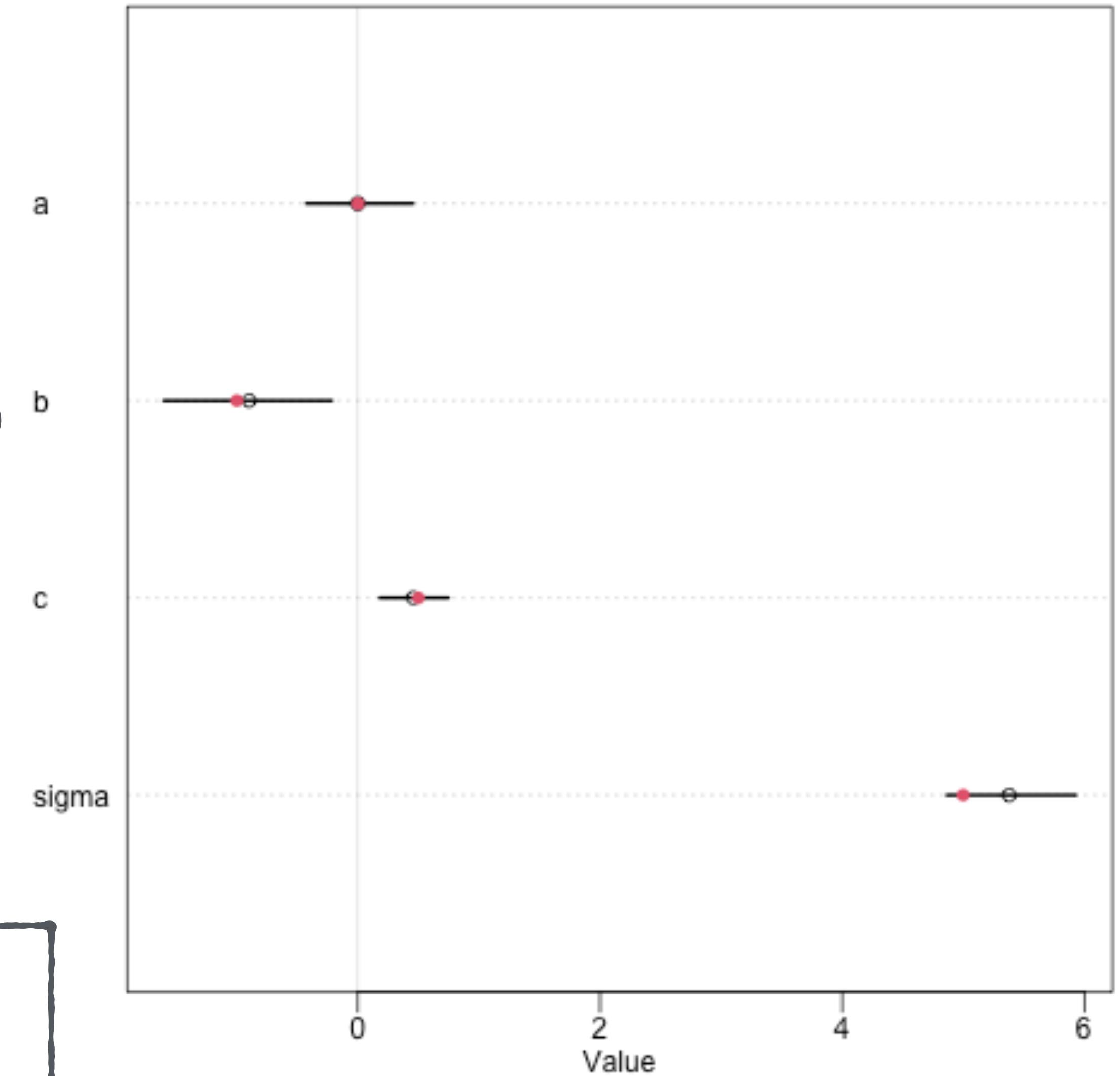
$$size_i = Normal(30 - 1 \times toxin_i + 0.5 \times age_i, 5)$$

Model:

$$size_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = a + b \times toxin_i + c \times age$$

Exercício: O que acontece se nós não incluirmos idade no modelo??



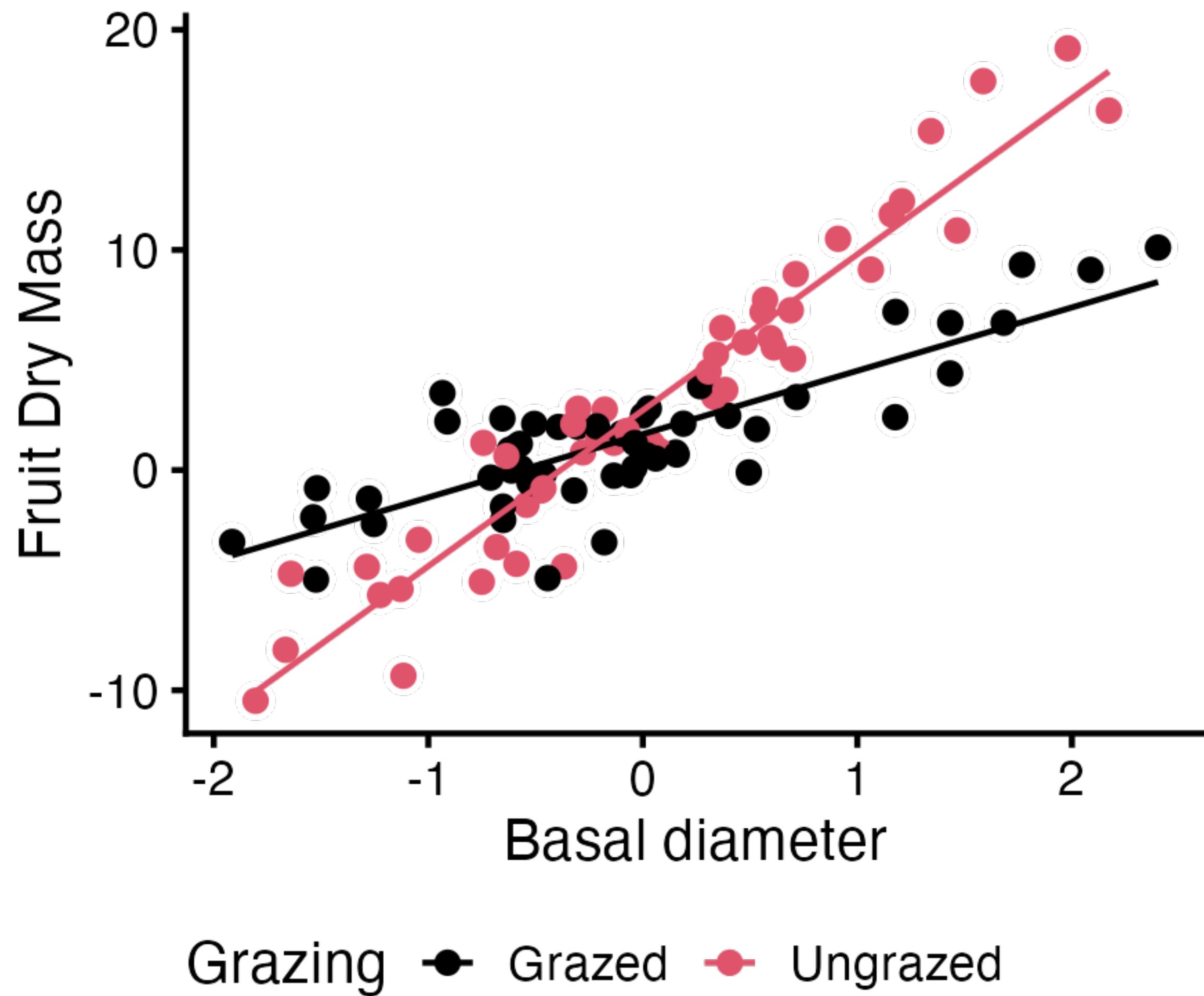
Se o seu modelo não funciona em
dados simulados, ele nunca vai
funcionar em dados reais!

Interações

Permitindo com que os coeficientes variem de acordo com outras variáveis

- Qual é a relação entre um preditor e uma resposta dependendo do valor de outro preditor?
- Nosso modelo pode incluir essa possibilidade com a inclusão do produto entre dois preditores como um terceiro preditor.
- Com dois preditores, x e z :

$$\mu = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$$



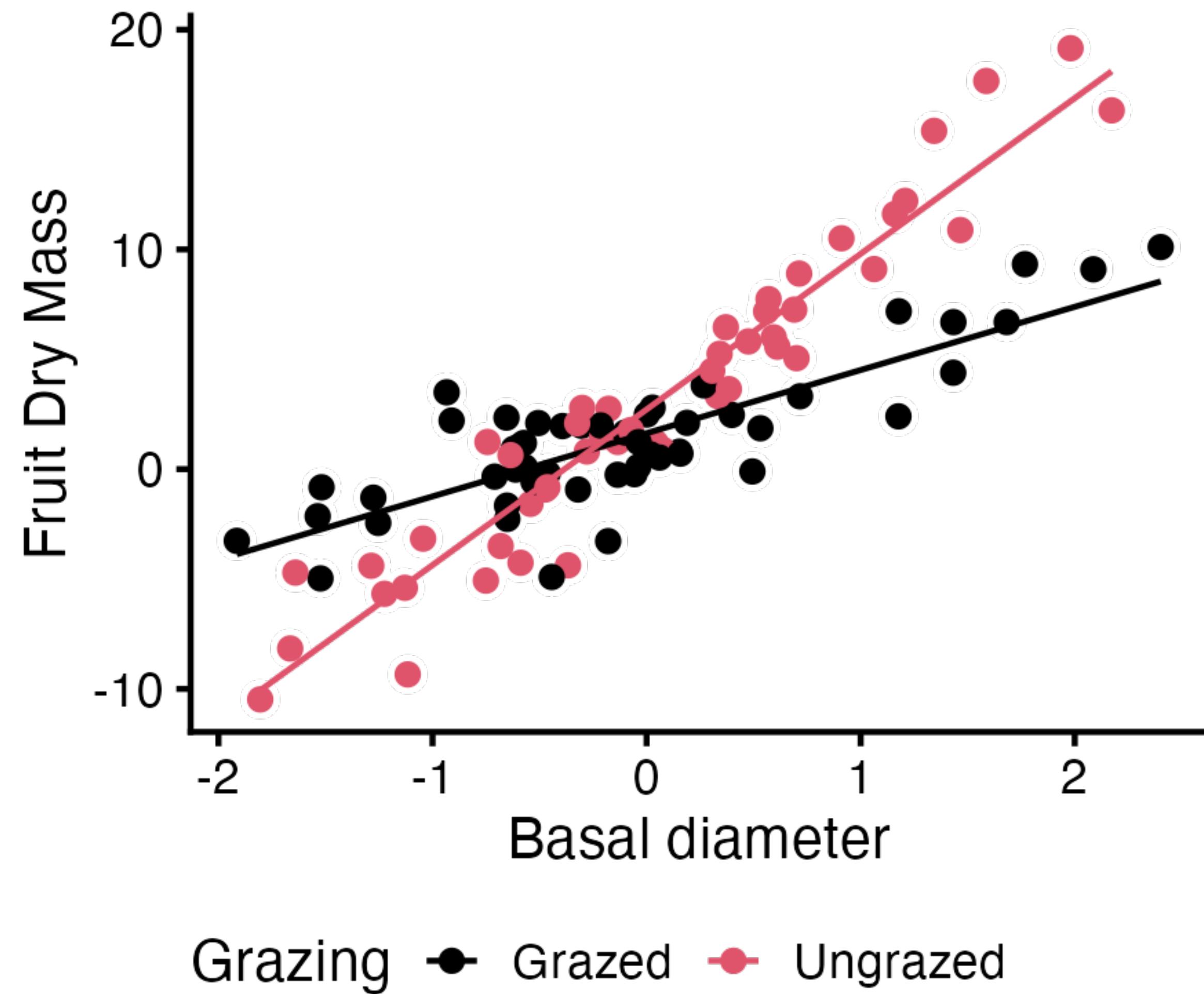
Interações

Permitindo com que os coeficientes variem de acordo com outras variáveis

- Qual é a relação entre um preditor e uma resposta dependendo do valor de outro preditor?
- Nosso modelo pode incluir essa possibilidade com a inclusão do produto entre dois preditores como um terceiro preditor.
- Com dois preditores, x e z :

$$\mu = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$$

Termo de interação!



Exemplo com interação

- **Pergunta:** qual é a melhor combinação de umidade e luminosidade para o crescimento de tulipas?
- **Experimento numa estufa:** tulipas mantidas em nove combinações de umidade do solo e sombreamento. Três repetições para cada combinação (total de 27 camas).
- **Variável resposta:** Altura média das plantas em cada pote.
- **Variáveis preditoras:**
 - Umidade: 3 níveis (low, med, high)
 - Sombreamento: 3 níveis (low, med, high)



Data from: Grafen & Hails (2002) Modern Statistics for the Life Sciences.

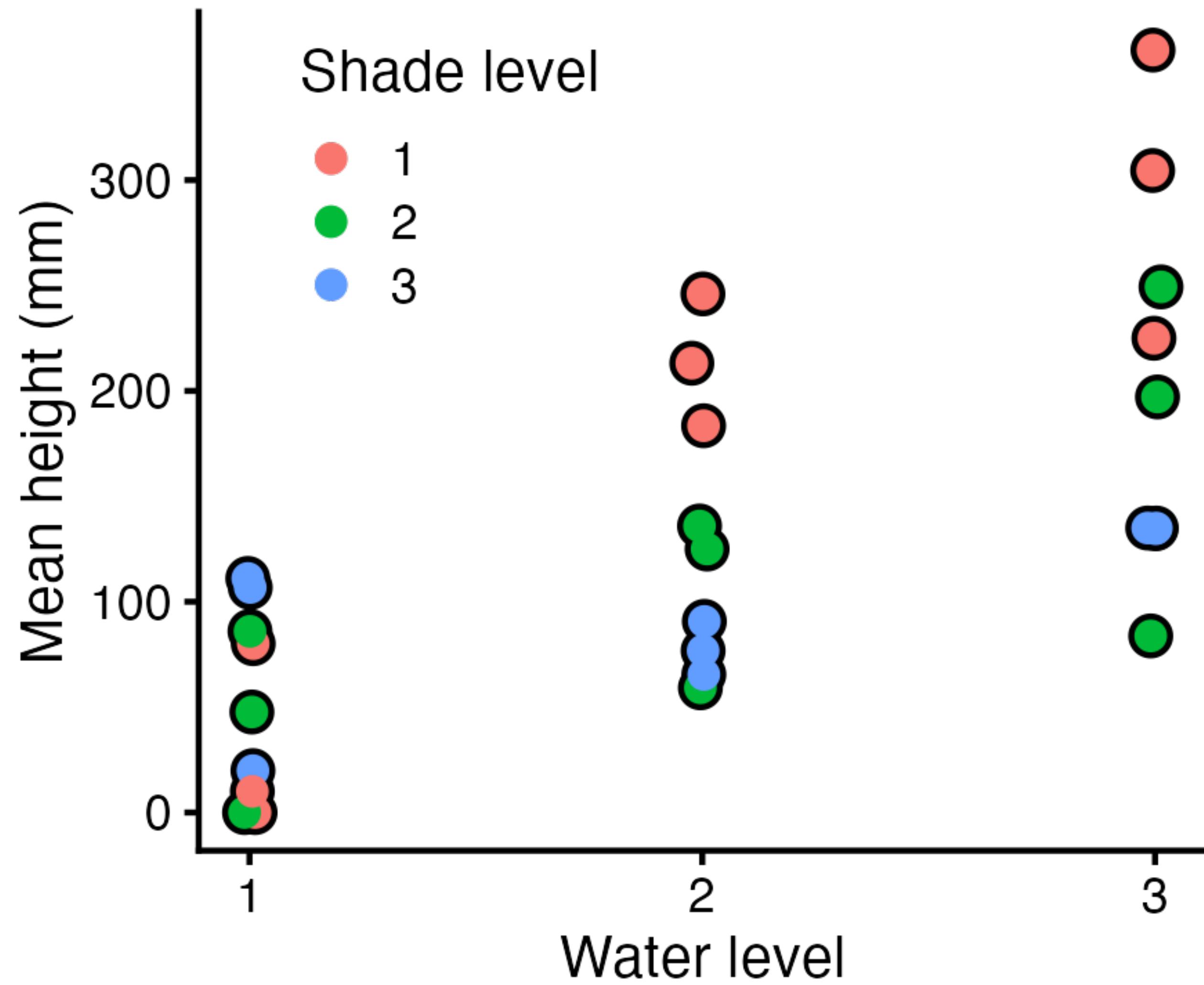
Tulips data

Shade level:

1. Low
2. Medium
3. High

Water level:

1. Low
2. Medium
3. High



Transformações simples

Sempre pense numa escala que torne os dados mais fáceis de interpretar!

Shade level:

-1 : Low

0: Medium

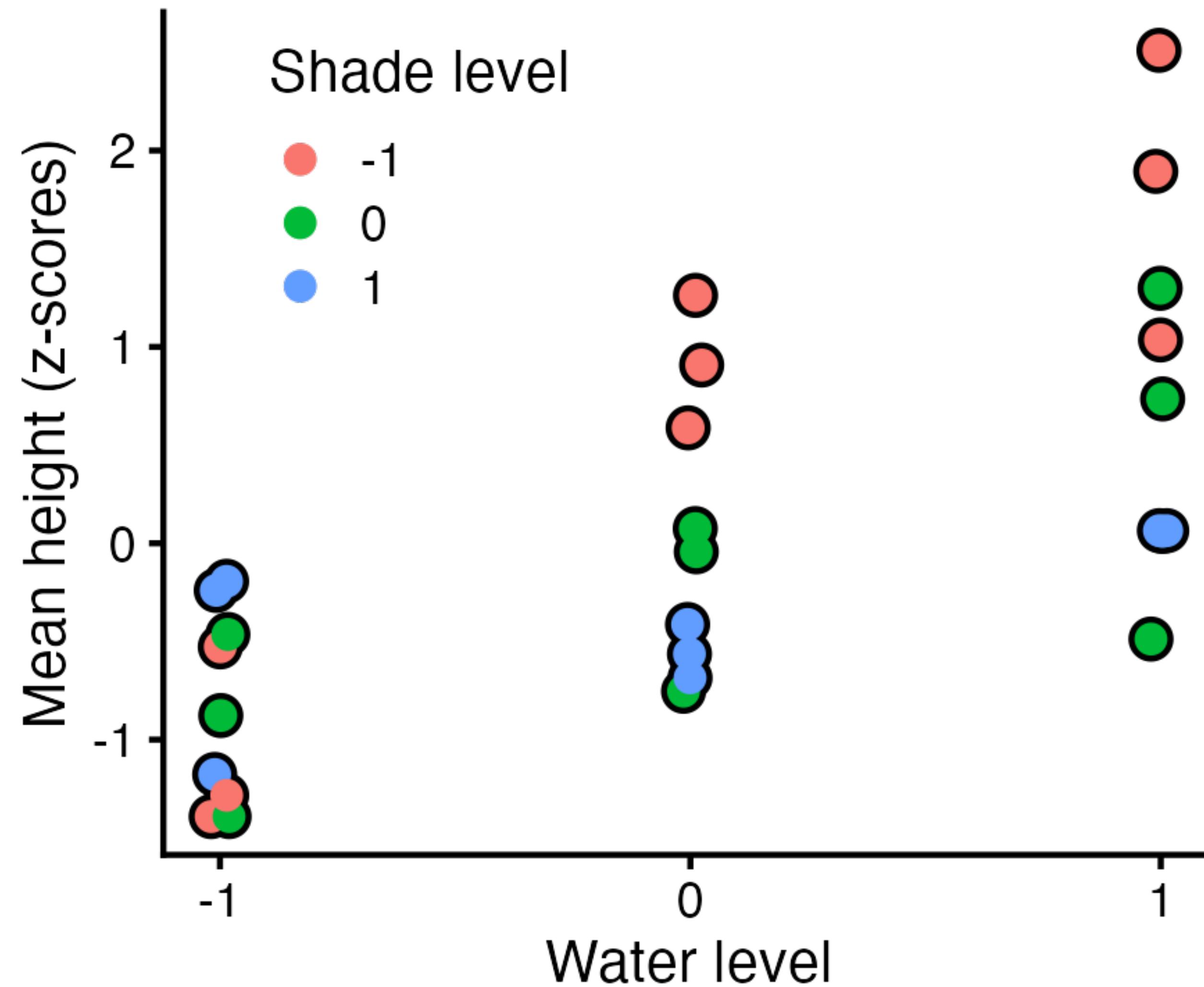
1: High

Water level:

-1 : Low

0: Medium

1: High

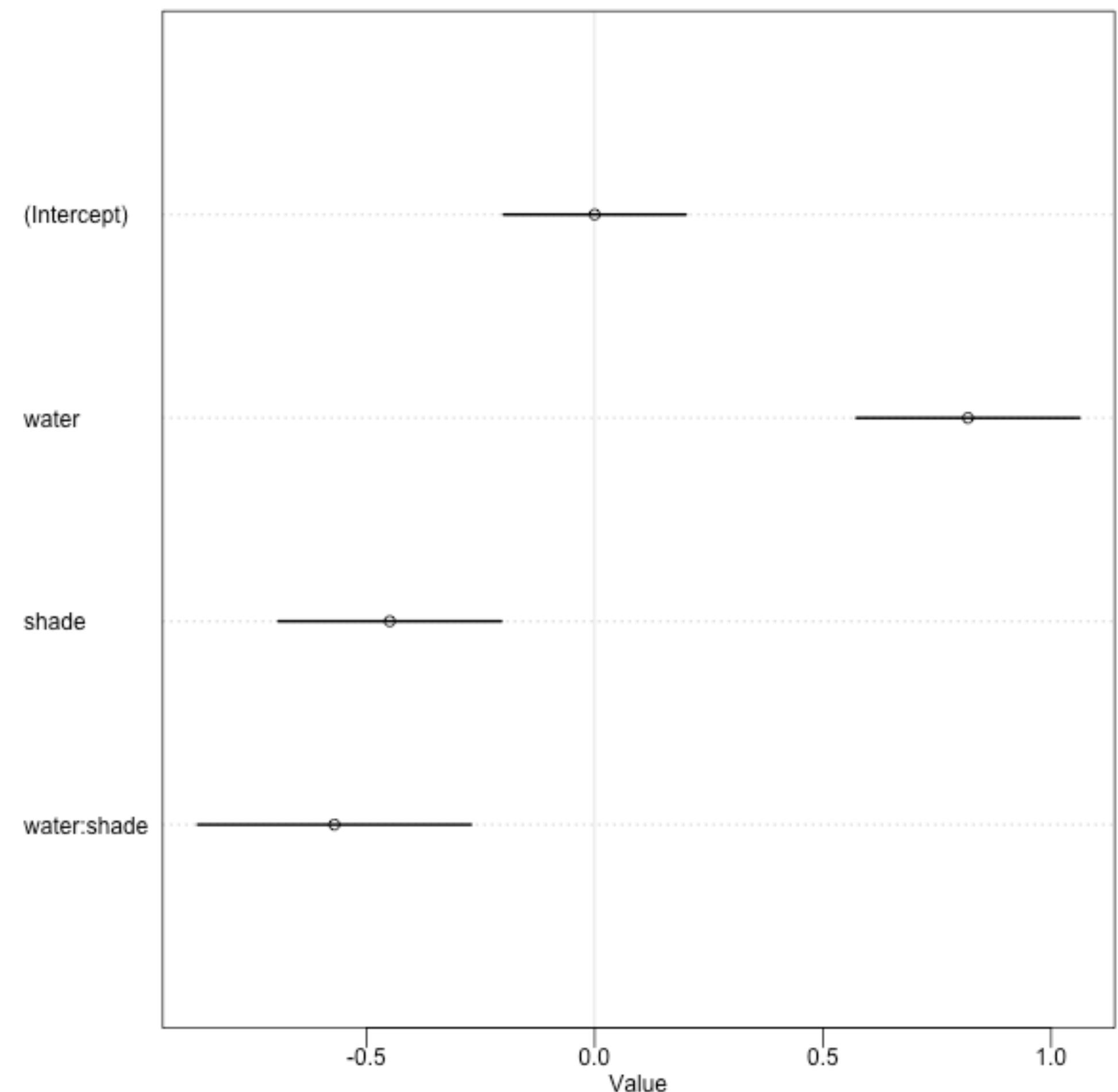


Modelo com interação

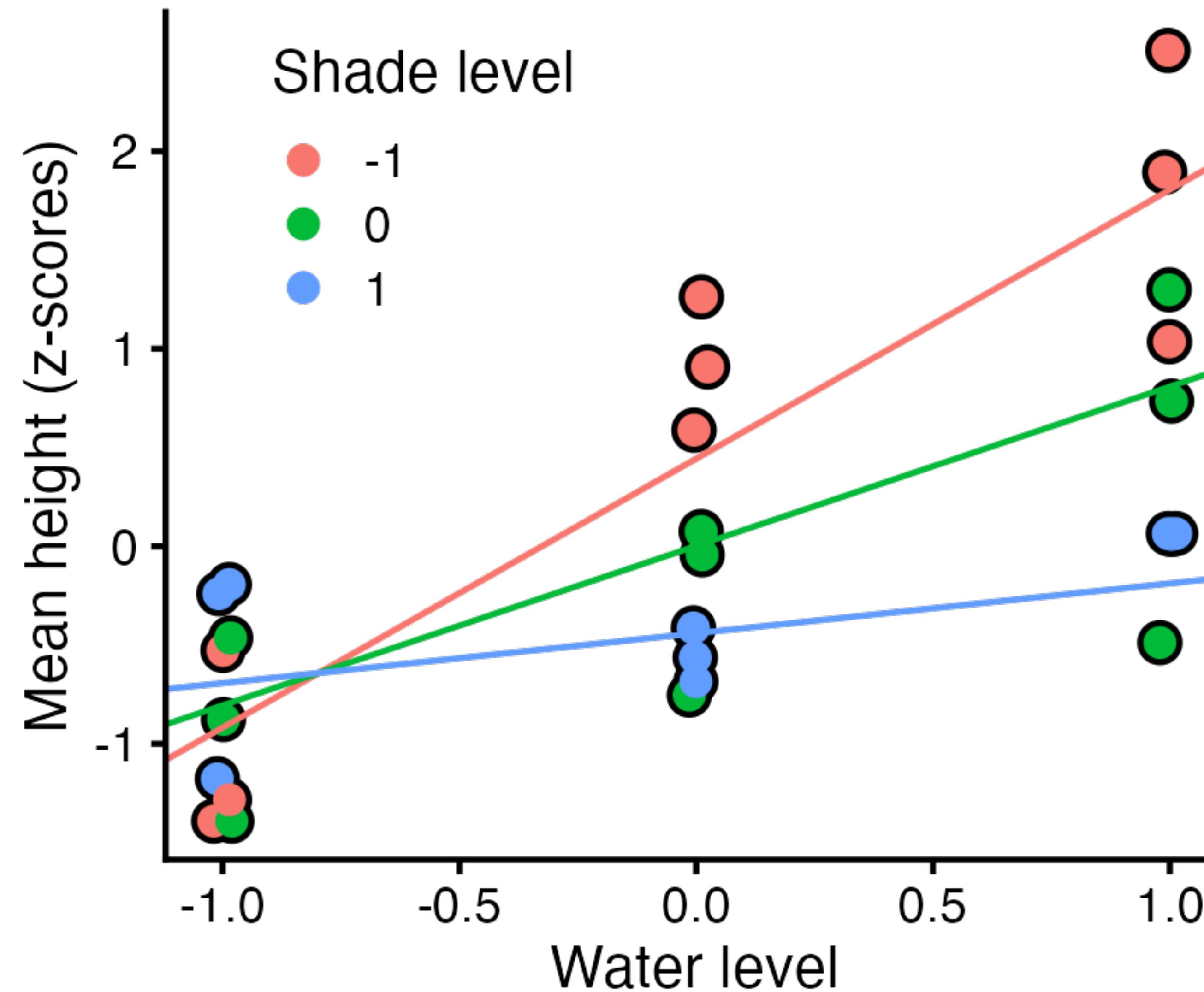
$$Height_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = a + b \text{ water}_i + c \text{ shade} + d \text{ water}_i \times \text{shade}_i$$

```
> data("tulips")
> df = tulips
> df$water = scale(df$water, scale = FALSE)
> df$shade = scale(df$shade, scale = FALSE)
> df$blooms = scale(df$blooms)
>
> rt_fit = lm(blooms ~ 1 + water + shade + water:shade,
              data = df)
> precis(rt_fit, prob = 0.95)
      mean   sd  2.5% 97.5%
(Intercept) 0.00 0.10 -0.20  0.20
water        0.82 0.12  0.57  1.06
shade       -0.45 0.12 -0.69 -0.20
water:shade -0.57 0.15 -0.87 -0.27
```



Linhas de previsão

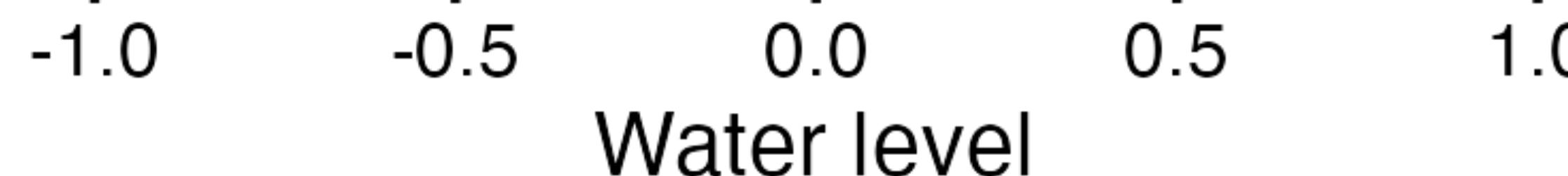


Linhas de previsão

```
cf = coef(rt_fit)
a = cf["(Intercept)"]; b = cf["water"]
c = cf["shade"]; d = cf["water:shade"]
col = scales::hue_pal()(3)

# Model: a + b*w + c*s + d*w*s

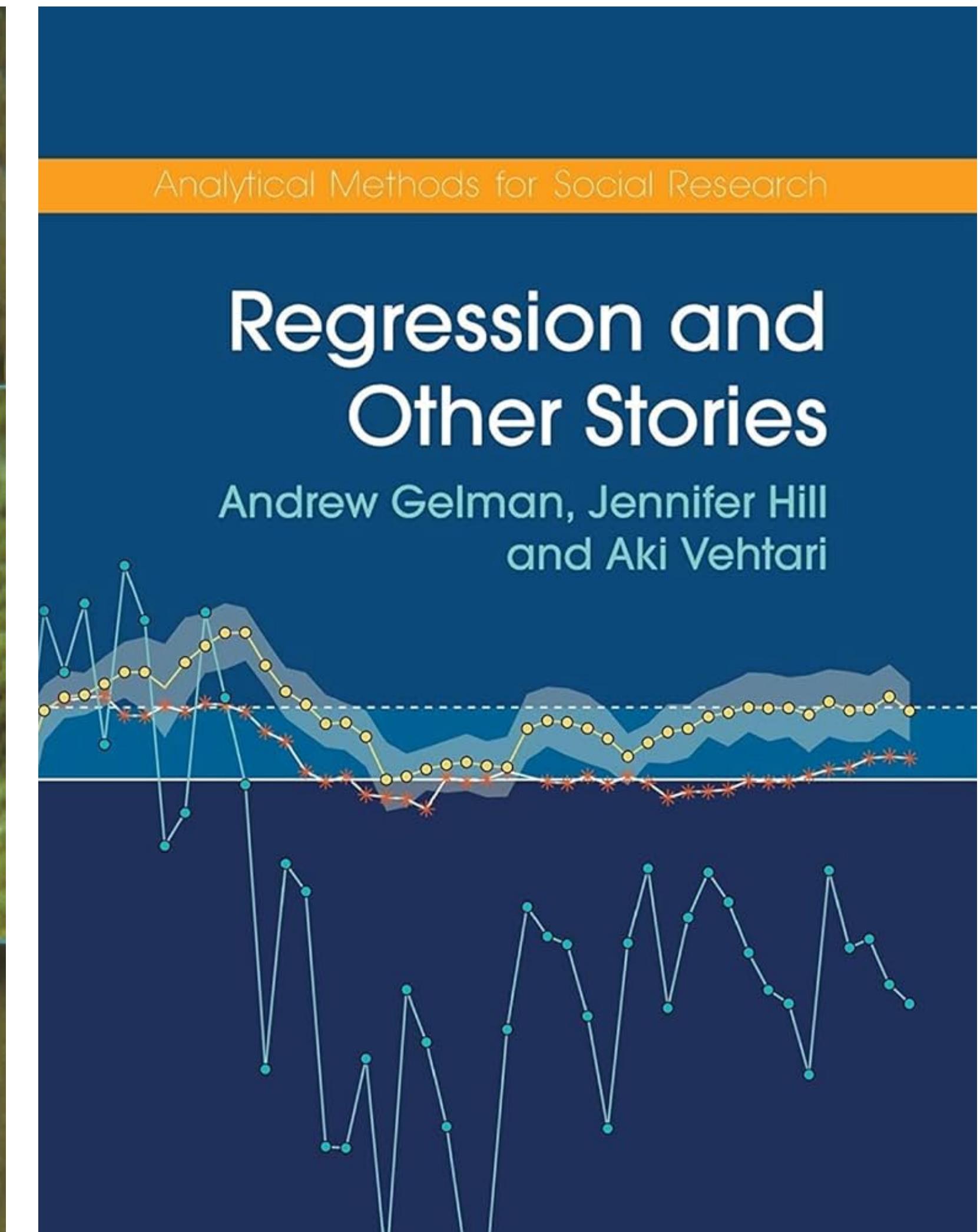
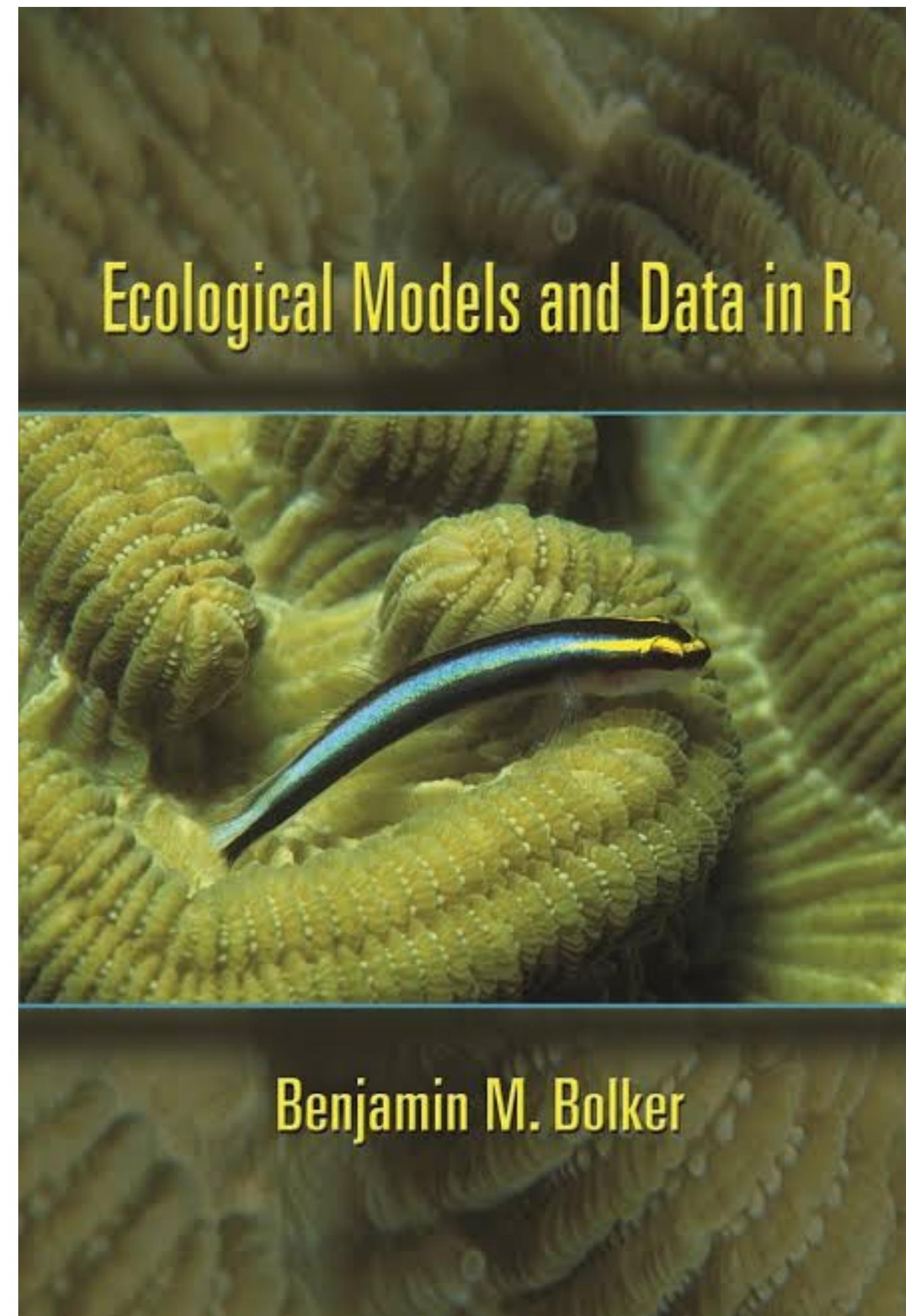
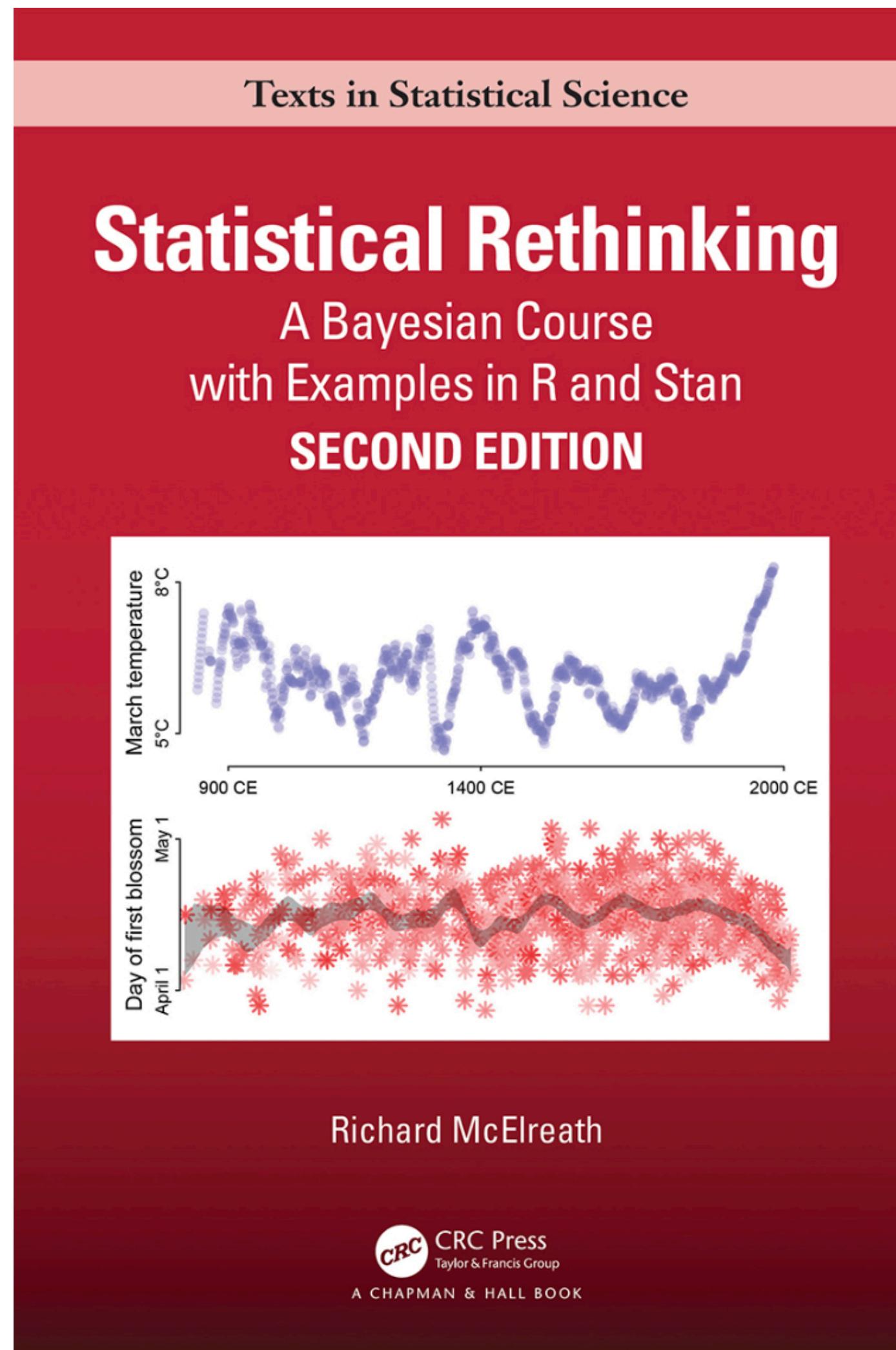
library("ggplot2")
p = ggplot(df, aes(x = water, y = blooms, color = factor(shade))) +
  geom_point(size = 2) +
  geom_abline(intercept = a - c, slope = b - d, col = col[1]) +
  geom_abline(intercept = a , slope = b , col = col[2]) +
  geom_abline(intercept = a + c, slope = b + d, col = col[3])
```



Summary

- Modelos lineares múltiplos nos permitem usar mais de um preditor no mesmo modelo.
- Eses modelos fazem uma forma automática de **estratificação**
 - **Ex:** Diferenças de tamanho para indivíduos da mesma idade, efeitos do tratamento para indivíduos do mesmo tamanho, e assim por diante.
- O objetivo é comparar observações equivalentes.
- Os coeficientes podem e devem mudar com a inclusão de preditores correlacionados entre si.
- Interpretar coeficientes é difícil. Use **gráficos, previsões e transformações** para facilitar a interpretação dos modelos.
- Amanhã, vamos falar sobre **como escolher as variáveis** que entram numa regressão.

Me apaixonei e quero saber mais



Me apaixonei e quero saber mais MESMO

