"99 percent of all statistics only tell 49 percent of the story"

Ron DeLegge II (Economist)

Now we can test for differences in adjusted means; but before that:

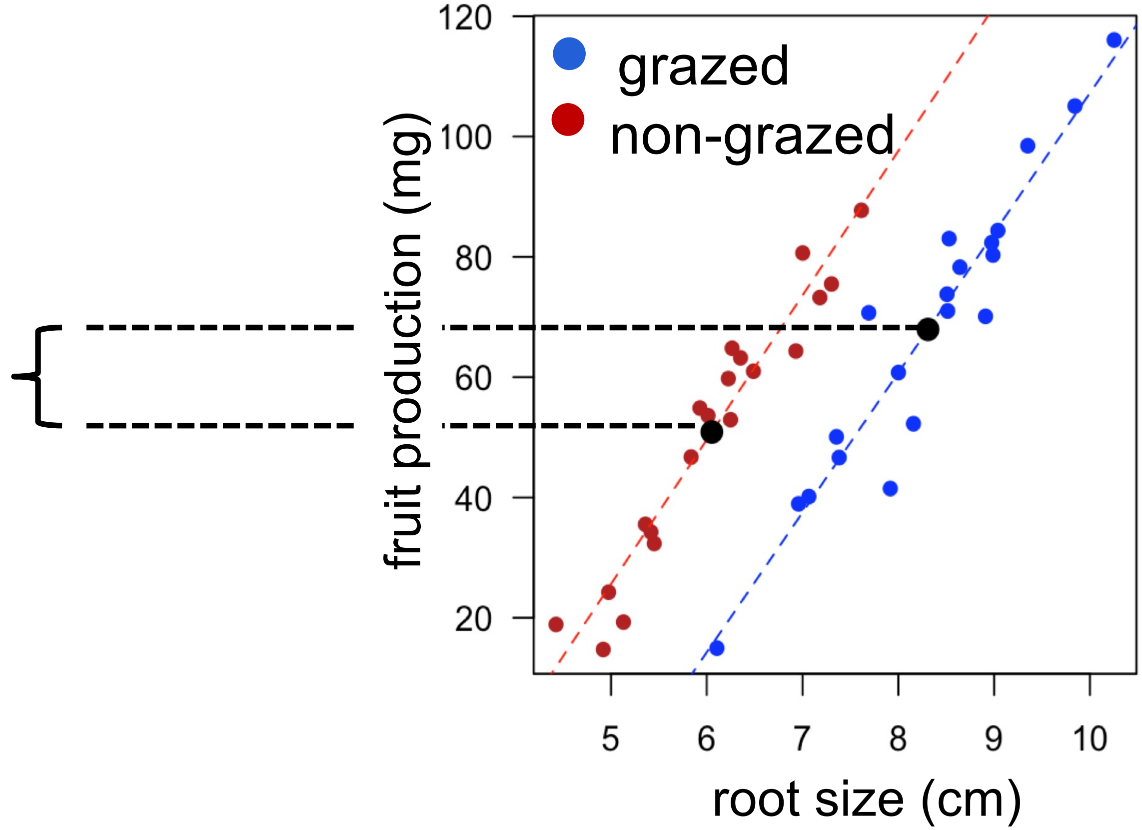
Critical statistical issues underlying General Linear Models (including ANCOVAs)

Lecture 10
(Type I and III sum-of-square)

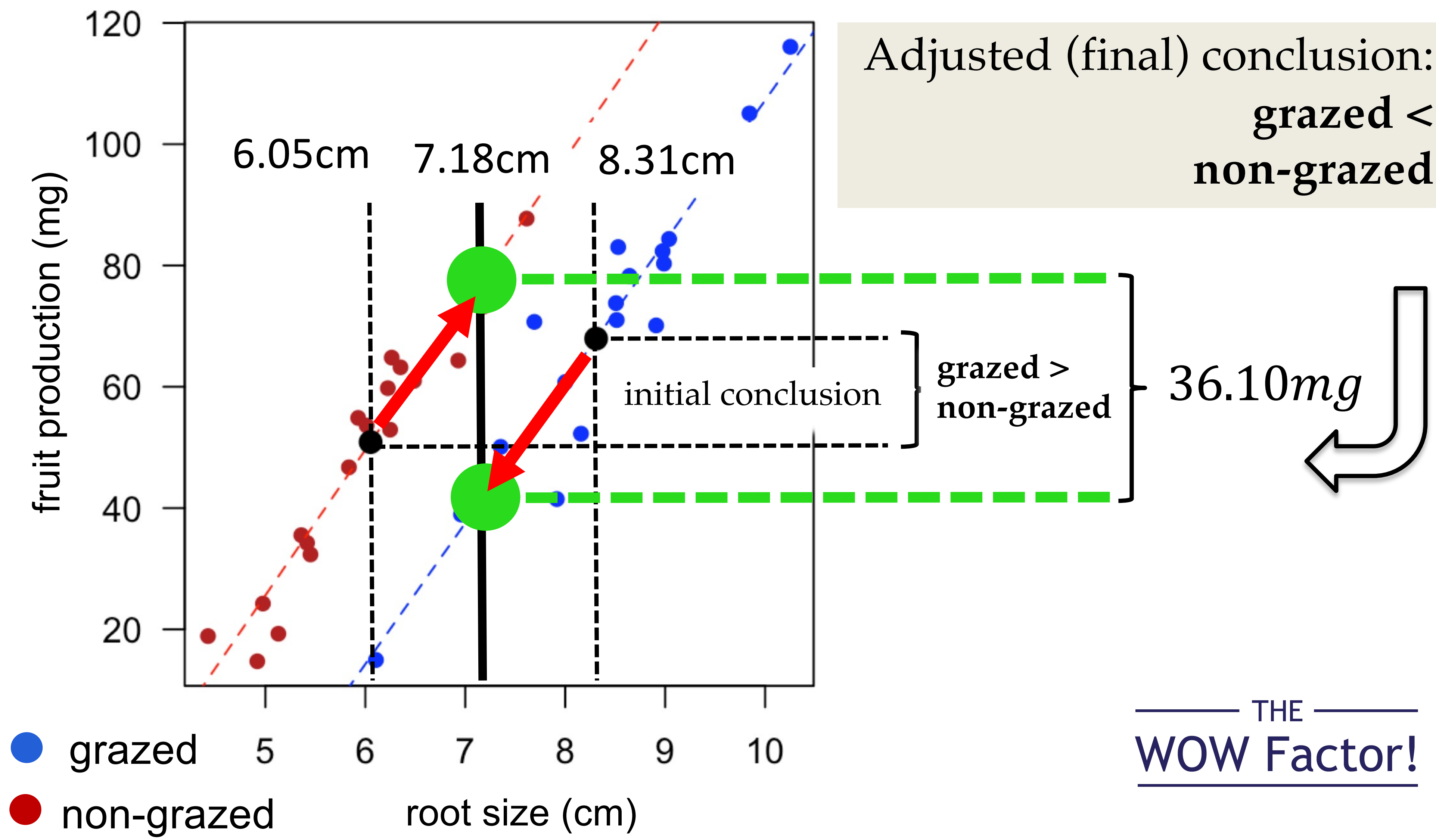
Initial conclusion: greater fruit production under grazing!

```
> anova(lm(Fruit~Grazing))
Analysis of Variance Table

Response: Fruit
          Df  Sum Sq Mean Sq F value  Pr(>F)
Grazing    1  2910.4 2910.44  5.3086 0.02678 *
Residuals 38 20833.4  548.25
```

What's the conclusion?



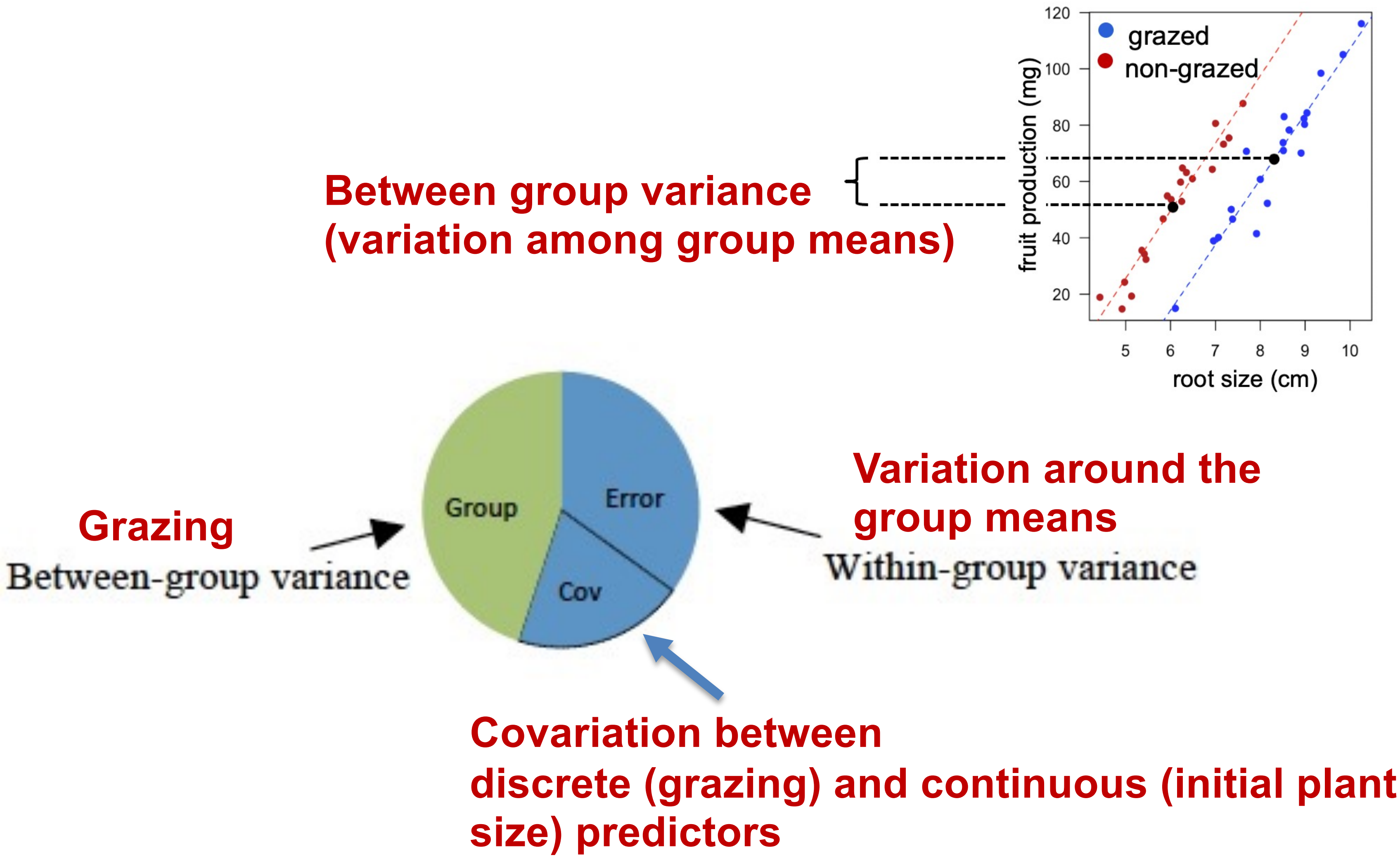
Grazing is significant - but in what direction?
Does grazing increase or reduce fruit production?



Adjusted (final) conclusion:
grazed < non-grazed

THE WOW Factor!

Analysis of covariance (ANCOVA) evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable (treatment), while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates or nuisance variables.



Adapted from https://en.wikipedia.org/wiki/Analysis_of_covariance

Before we proceed in testing for differences in adjusted means - a few really important issues

1) Since the two main conditions hold (1) Covariate can predict the response; and 2) Groups share a common slope, i.e., interaction is not significant), we can proceed to test the effect of grazing (categorical predictor) while controlling for initial plant size (root size). Also, given that the slopes are similar, we can drop the interaction in the final analysis (thought there are discussions about whether this is cautious or not).

$$Y = \mu + A_1 + X_1 \underline{(+A_1 \times X_1)}$$

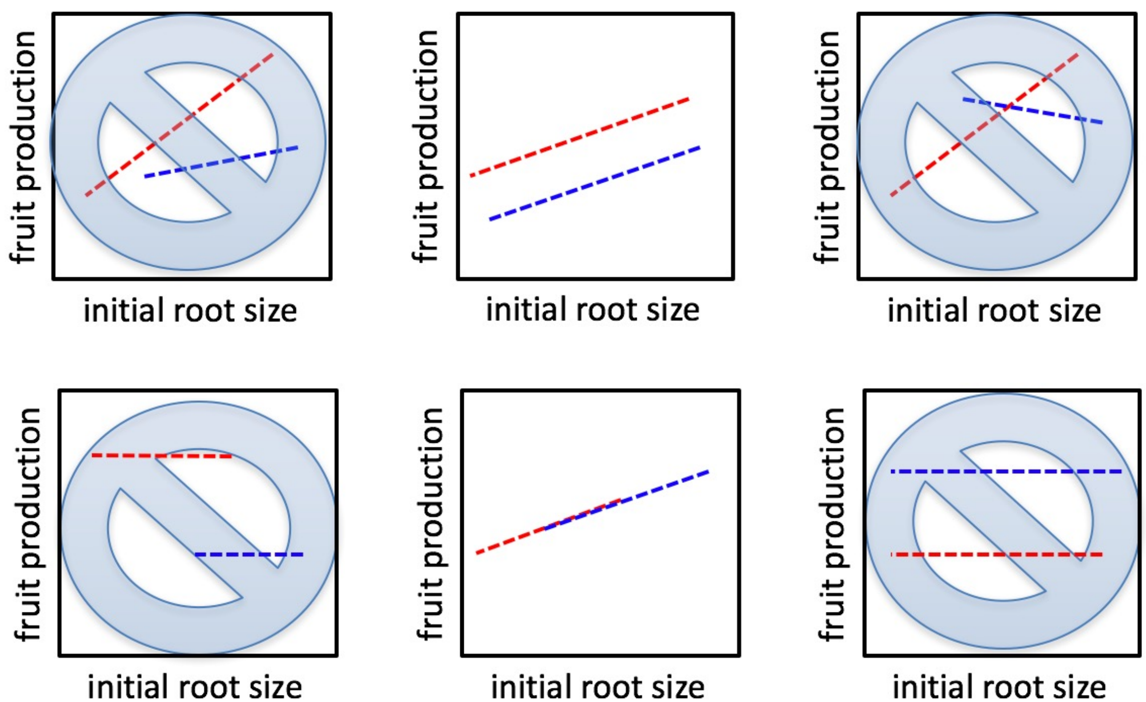
$$\text{Fruit production} = \mu + \mathit{Grazing} + \mathit{Root\ size}$$

$$\cancel{+ \mathit{Grazing} \times \mathit{Root\ size}}$$

Before we proceed in testing for differences in adjusted means - a few really important issues

$$Y = \mu + A_1 + X_1(+A_1 \times X_1)$$

$$\text{Fruit production} = \mu + \textit{Grazing} + \textit{Root size} + \textit{Grazing} \times \textit{Root size}$$



BIOL 422 & 680, Pedro Peres-Neto, Biology, Concordia University

ANOVA, Regression and types of sum-of-squares

Before we proceed in testing for differences in adjusted means - a few really important issues

1) Since the two main conditions hold (1) Covariate can predict the response; and 2) Groups share a common slope), we can proceed to test the effect of grazing (categorical predictor) while controlling for initial plant size (root size). Also, given that the slopes are similar, we can drop the interaction in the final analysis.

2) When multiple predictors are used, we estimate partial effects i.e., the total amount of variation explained by grazing once initial size (covariate) is controlled for (removed).

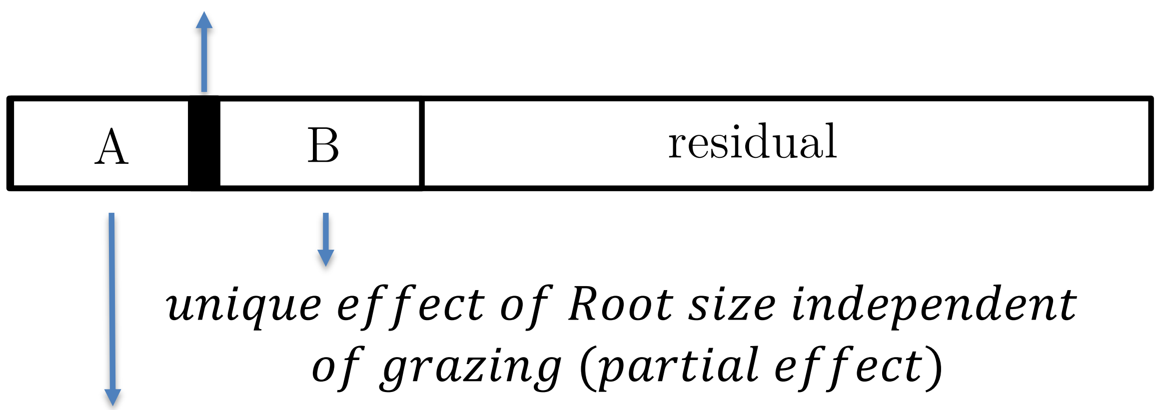
Before we proceed in testing for differences in adjusted means - a few really important issues

2) When multiple predictors are used, we estimate partial effects i.e., the total amount of variation explained by grazing once initial size (covariate) is controlled for (removed).

$$\text{Fruit production} = \mu + \mathit{Grazing} + \mathit{Root\ size}$$

$\boldsymbol{variation\ shared\ between\ Grazing\ and\ Root\ size\ (non-orthogonal)}$

Fruit production =



$unique\ effect\ of\ Root\ size\ independent\ of\ grazing\ (partial\ effect)$

$unique\ effect\ of\ Grazing$ independent of root size (partial effect)

Note that interaction can't be significant for adjustement; so assumed zero here

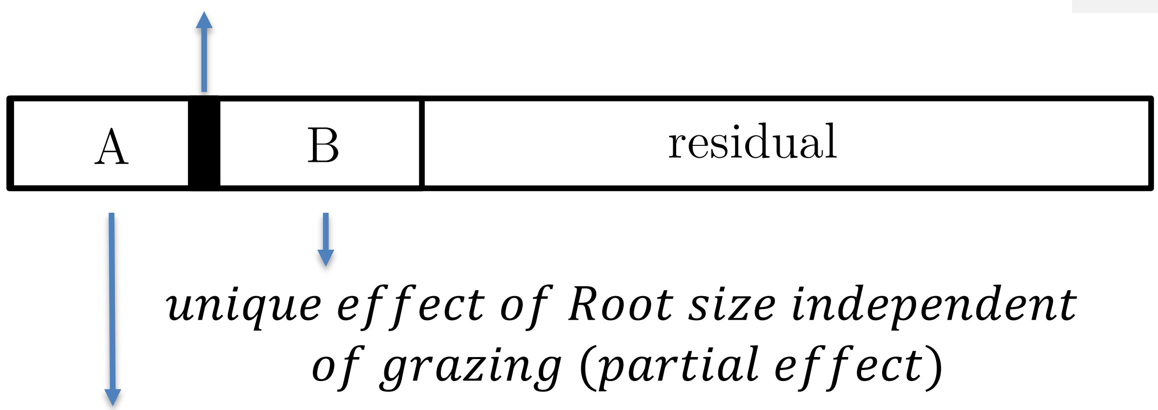
Before we proceed in testing for differences in adjusted means - a few really important issues

2) Remember that as in a regression model, partial effects are used, i.e., the total amount of variation explained by grazing once initial size (covariate) is controlled for (removed).

3) However, standard ANOVA assumes that categorical factors are orthogonal, and this is not possible when a categorical and a continuous variable are tested in the same model. After all, if grazing and Root would be orthogonal, there would be no correlation between them!

variation shared between Grazing and Root size (non − orthogona > 0

Fruit production =



unique effect of Root size independent of grazing (partial effect)

unique effect of Grazing independent of root size (partial effect)

Before we proceed in testing for differences in adjusted means - a few really important issues

2) Remember that as in a regression model, partial effects are used, i.e., the total amount of variation explained by grazing once initial size (covariate) is controlled for (removed).

3) However, standard ANOVA assumes that categorical factors are orthogonal, and this is not possible when a categorical and a continuous variable are tested in the same model. After all, if grazing and Root would be orthogonal, there would be no correlation between them!

grazing is a contrast (as seen in our last lecture)

```
> cor(as.numeric(Grazing),Root)
[1] -0.772087
```

Because of lack of orthogonality between categorical (grazed/non-grazed) and covariate (initial root size), the order of the categorical and covariate change the results when using a common ANOVA (which is based on type I Sum of squares).

```
> anova(lm(Fruit ~ Grazing+Root))
Analysis of Variance Table

Response: Fruit
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Grazing    1  2910.4  2910.4  63.929 1.397e-09 ***
Root       1 19148.9 19148.9 420.616 < 2.2e-16 ***
Residuals 37  1684.5    45.5
```

```
> anova(lm(Fruit ~ Root+Grazing))
Analysis of Variance Table

Response: Fruit
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Root       1 16795.0 16795.0  368.91 < 2.2e-16 ***
Grazing    1  5264.4  5264.4  115.63 6.107e-13 ***
Residuals 37  1684.5    45.5
```

Understanding the Type I sum-of-squares (sequential)

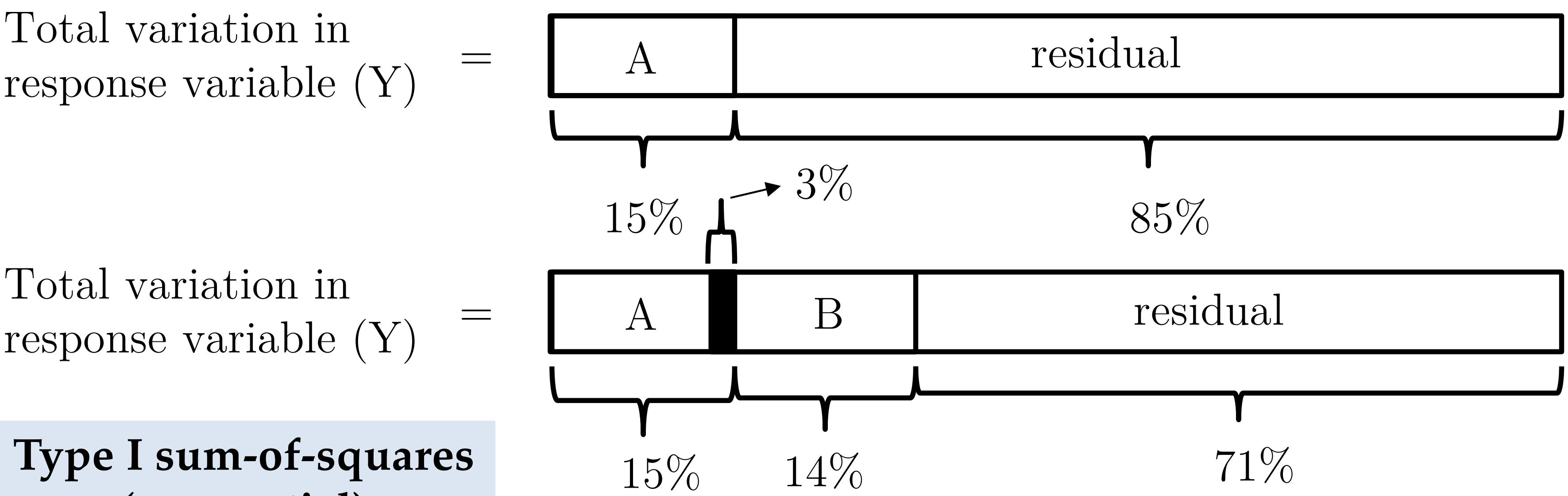
The Grazing treatment (A) is entered 1st into the model

Total variation in response variable (Y) =

Type I sum-of-squares (sequential)

$$\text{Fruit production} = \mu + Grazing$$

The Grazing treatment (A) is entered 1st into the model and then initial root size (B)

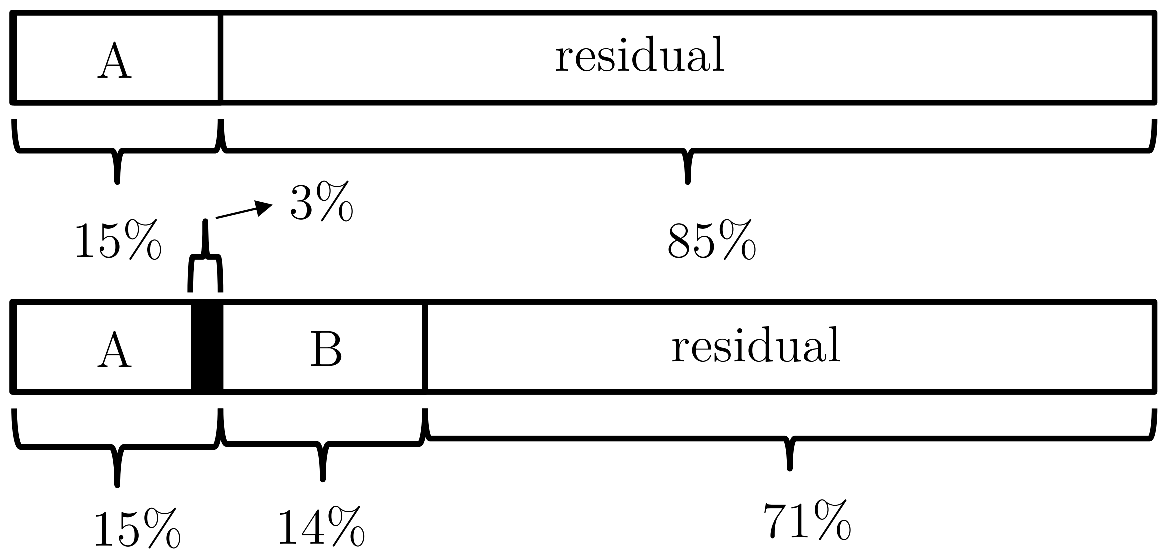


Type I sum-of-squares (sequential)

$$\text{Fruit production} = \mu + Grazing + Root\ size$$

The Grazing treatment (A) is entered 1st into the model and then initial root size (B)

Total variation in response variable (Y) =

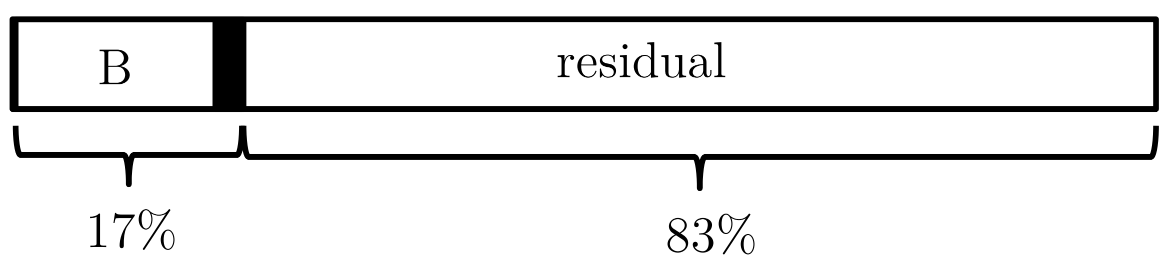


Total variation in response variable (Y) =

Type I sum-of-squares (sequential)

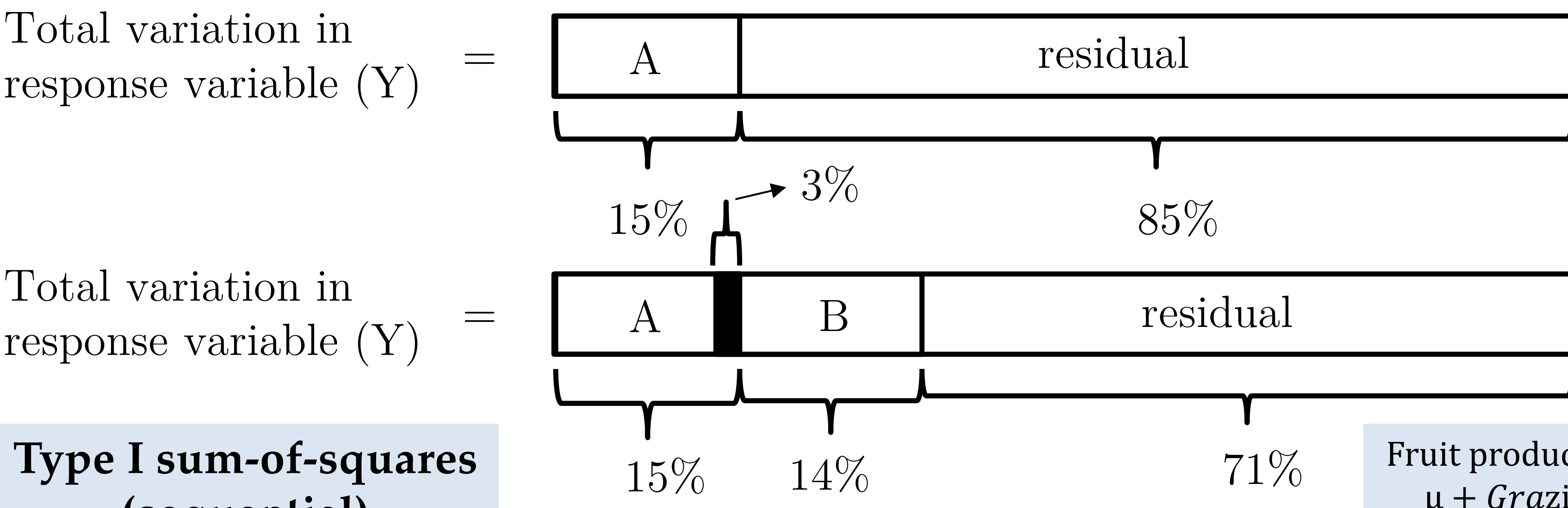
Initial root size (B) is entered 1st into the model and then the grazing treatment (A)

Total variation in response variable (Y) =



$$\text{Fruit production} = \mu + \textit{Root size}$$

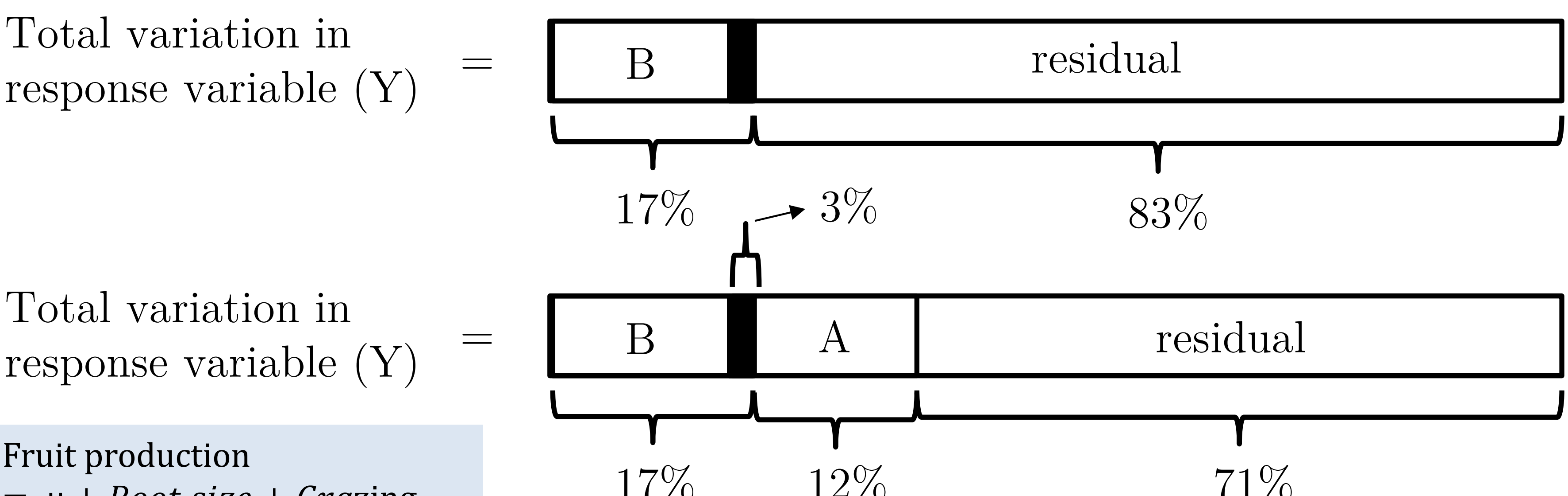
The Grazing treatment (A) is entered 1st into the model and then initial root size (B)



Type I sum-of-squares (sequential)

Fruit production = μ + *Grazing* + *Root size*

Initial root size (B) is entered 1st into the model and then the grazing treatment (A)

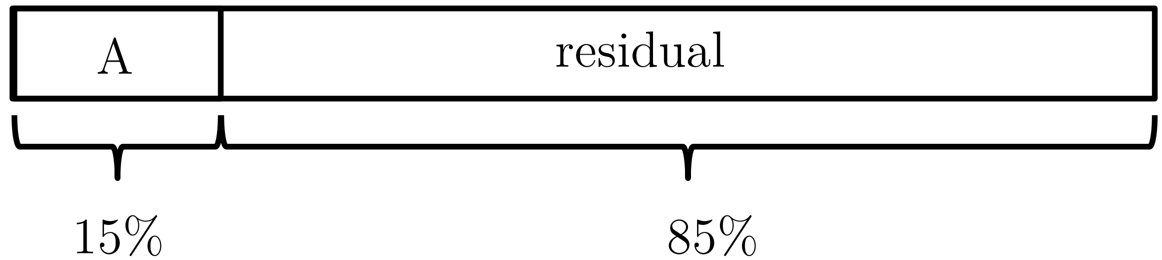


Fruit production = μ + *Root size* + *Grazing*

Understanding the Type III sum-of-squares (marginal or orthogonal)

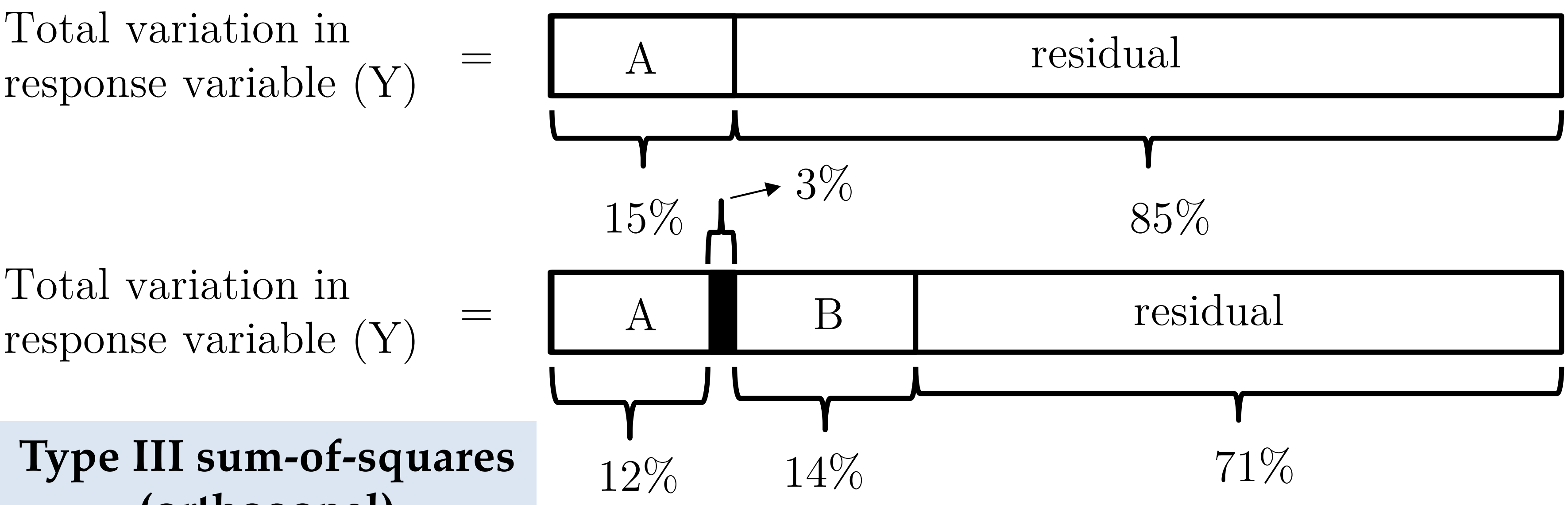
The Grazing treatment (A) is entered 1st into the model and then initial root size (B)

Total variation in response variable (Y) =



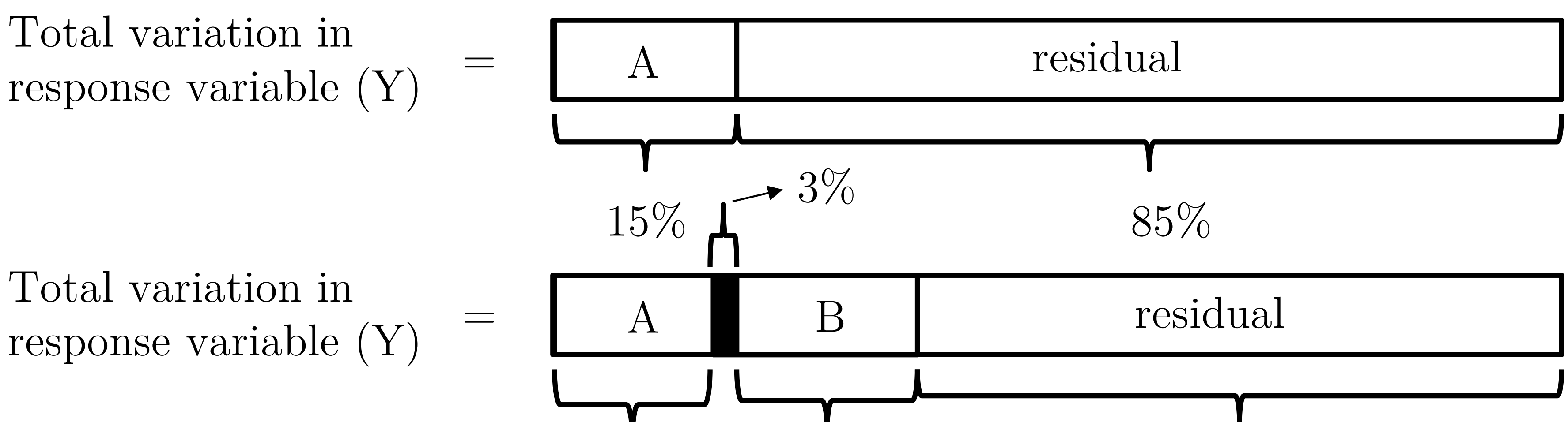
Type III sum-of-squares (orthogonal)

The Grazing treatment (A) is entered 1st into the model and then initial root size (B)



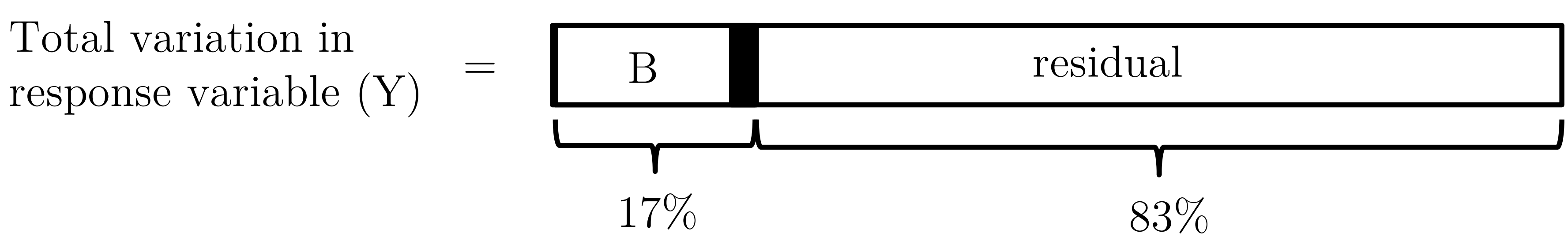
Type III sum-of-squares (orthogonal)

The Grazing treatment (A) is entered 1st into the model and then initial root size (B)

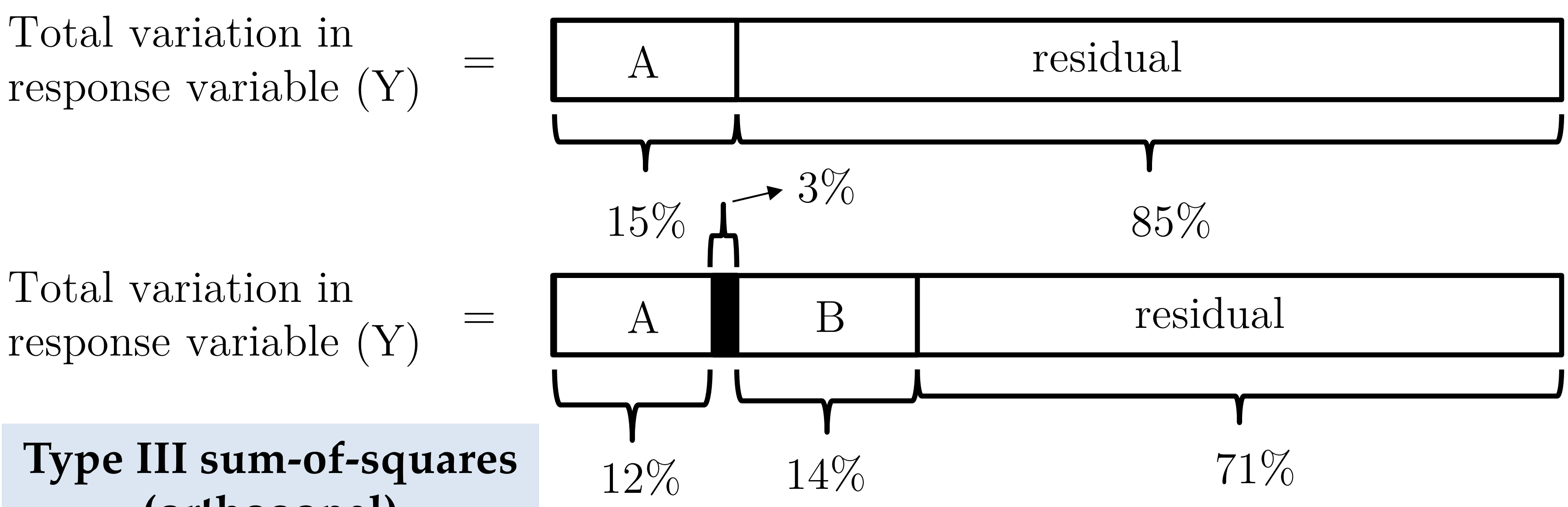


Type III sum-of-squares (orthogonal)

Initial root size (B) is entered 1st into the model and then the grazing treatment (A)

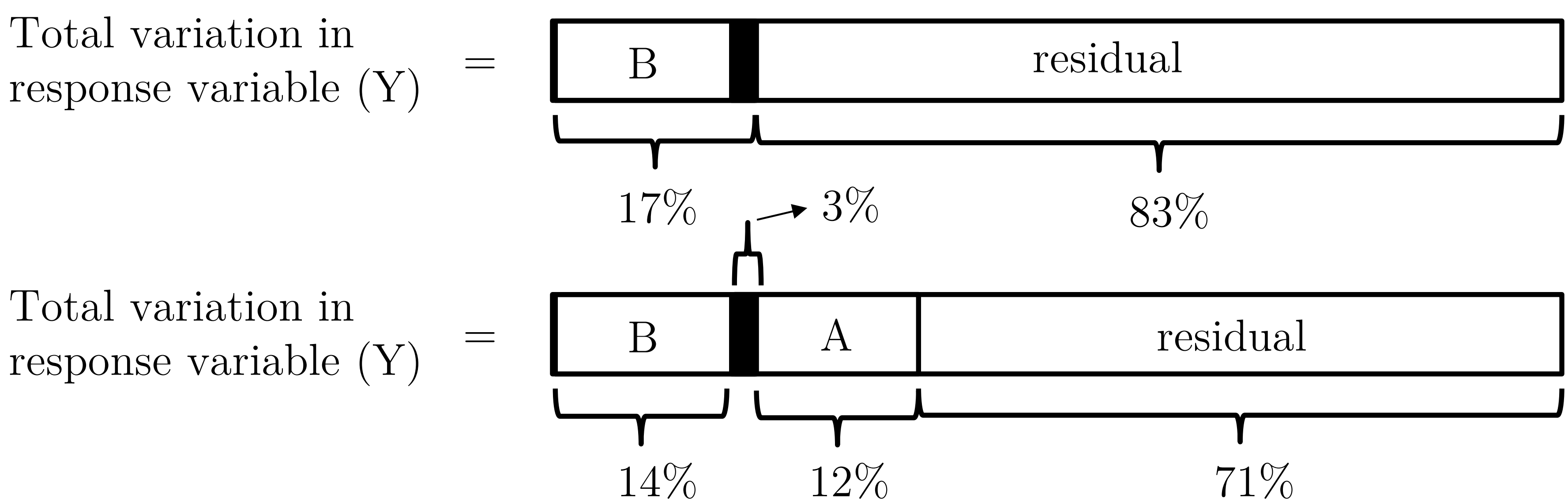


The Grazing treatment (A) is entered 1st into the model and then initial root size (B)



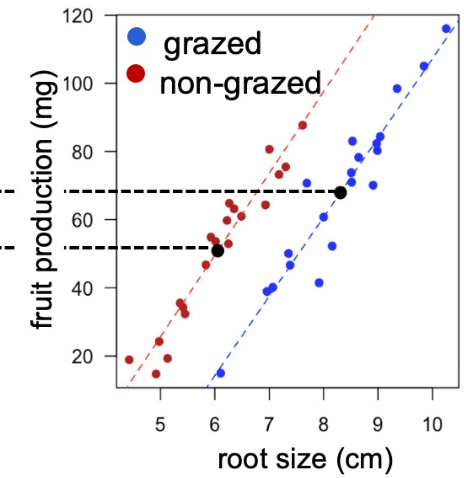
Type III sum-of-squares (orthogonal)

Initial root size (B) is entered 1st into the model and then the grazing treatment (A)



Analysis of covariance (ANCOVA) evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable (treatment), while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates or nuisance variables.

Between group variance
(variation among group means)



Grazing
Between-group variance

Variation around the group means
Within-group variance

Covariation between discrete (grazing) and continuous (initial plant size) predictors

Adapted from https://en.wikipedia.org/wiki/Analysis_of_covariance

Ecology, 87(10), 2006, pp. 2614–2625

VARIATION PARTITIONING OF SPECIES DATA MATRICES: ESTIMATION AND COMPARISON OF FRACTIONS

Pedro R. Peres-Neto,[1] Pierre Legendre, Stéphane Dray, and Daniel Borcard

Understanding semi-partial contributions via variation partitioning

Final test: Does grazing affect fruit production once controlled for initial root size?

Final test: Does grazing affect fruit production once controlled for initial root size?

H_0: Grazing treatments do not differ in fruit production.

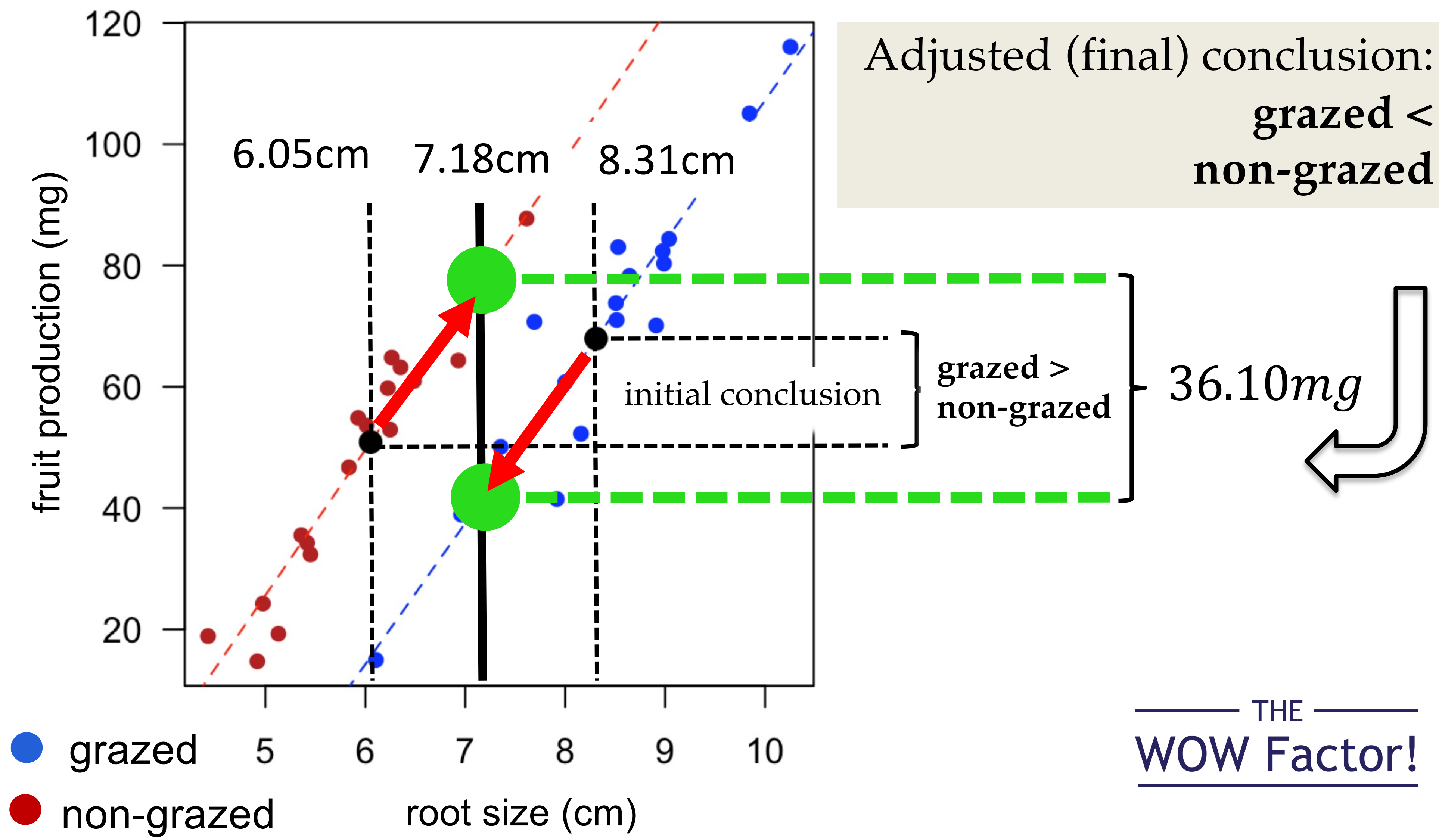
H_A: Grazing treatments differ in fruit production.

```
> Anova(lm.Fruit, type = "III") # note "A" in Anova is capitalized
Anova Table (Type III tests)

Response: Fruit
             Sum Sq Df F value    Pr(>F)
(Intercept)  7965.2  1  174.96 1.350e-15 ***
Grazing      5264.4  1  115.63 6.107e-13 ***
Root        19148.9  1  420.62 < 2.2e-16 ***
Residuals    1684.5 37
```

Type II and III Sum of squares so that order of entrance of categorical (grazing treatment) and continuous (covariate = initial root size).

Grazing is significant - but in which direction?

Does grazing increase or reduce fruit production?



THE WOW Factor!

Grazing is significant - but in which direction?

Does grazing increase or reduce fruit production?

$$\hat{Y}_{non-grazed (adjusted\ mean)} = -125.17mg + 36.1mg + 23.56mg/cm \times 7.18cm = 77.46212$$

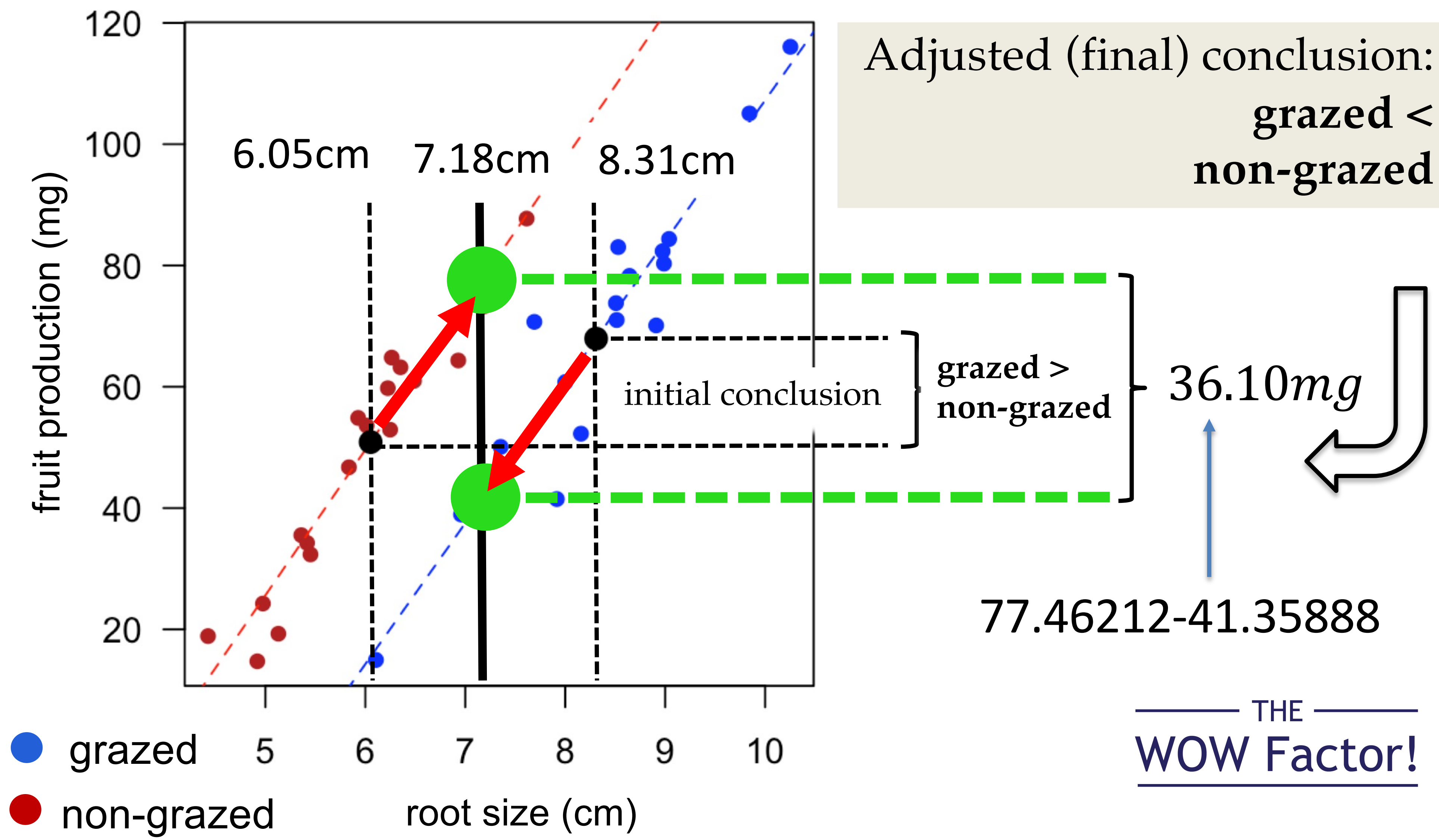
$$\hat{Y}_{grazed\ (adjusted\ mean)} = -125.17mg + 0.00mg + 23.56mg/cm \times 7.18cm = 41.35888$$

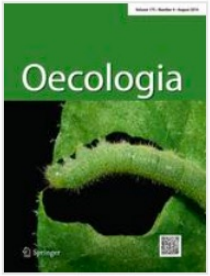
Adjusted (final) inference:

grazed fruit production < non-grazed fruit production

Grazing is significant - but in which direction?

Does grazing increase or reduce fruit production?





Oecologia
June 1992, Volume 90, Issue 3, pp 435–444 | Cite as

The effects of grazers on the performance of individuals and populations of scarlet gilia, *Ipomopsis aggregata*

Authors
Authors and affiliations

Joy Bergelson, Michael J. Crawley

I. aggregata exhibits considerable powers of regrowth following removal of its primary shoot by herbivores, but we found no evidence of overcompensation (i.e. of significantly higher plant performance where plants were exposed to ungulate herbivory) in a comparison between individuals on grazed and ungrazed sides of exclosure fences

Assessing if assumptions hold!

$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$	Analysis of Covariance (ANCOVA)

Testing for assumptions should be performed before reporting results – we did not do it here so that we paid attention to the problem first!

Assumptions

Assumption 1: linearity (more in the regression module)
The regression relationship between the dependent variable and concomitant variables must be linear.

Assumption 2: homogeneity of error variances (residual plot or the Breusch-Pagan test)

Equal variances for different treatment classes and observations.

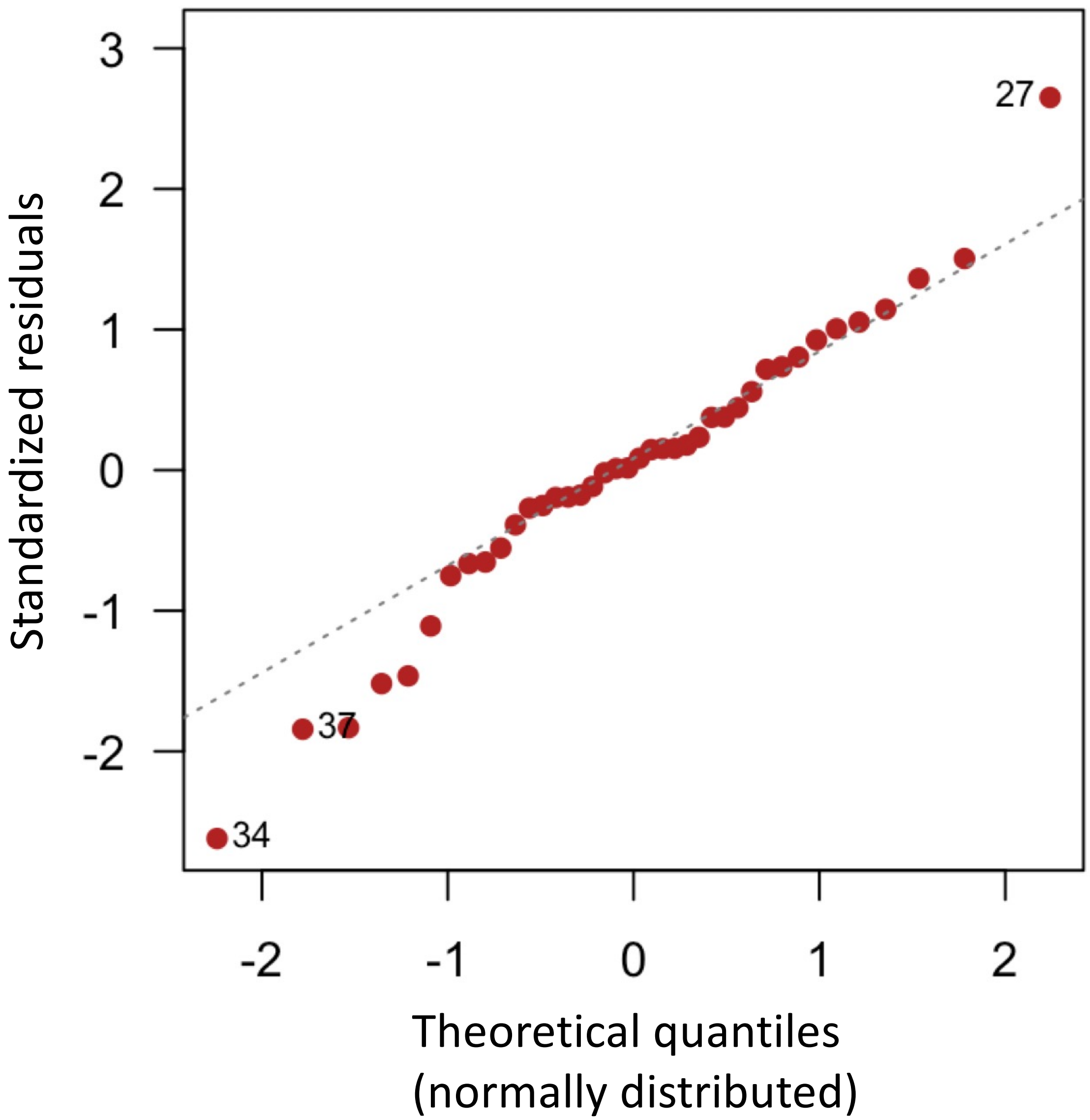
Assumption 3: independence of error terms (more in mixed models)
The errors are uncorrelated. That is, the error covariance matrix is diagonal.

Assumption 4: normality of error terms (Q-Q plot)
The residuals (errors) should be normally distributed.

Assumption 5: homogeneity of regression slopes (tested already).

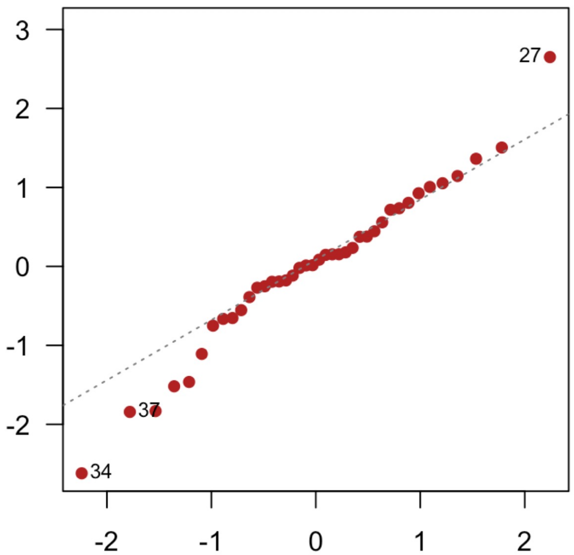
$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$ | Analysis of Covariance (ANCOVA)

Testing for normality assumptions (Q-Q normal residual plot)



$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$	Analysis of Covariance (ANCOVA)

Testing for normality assumptions (Q-Q normal residual plot)

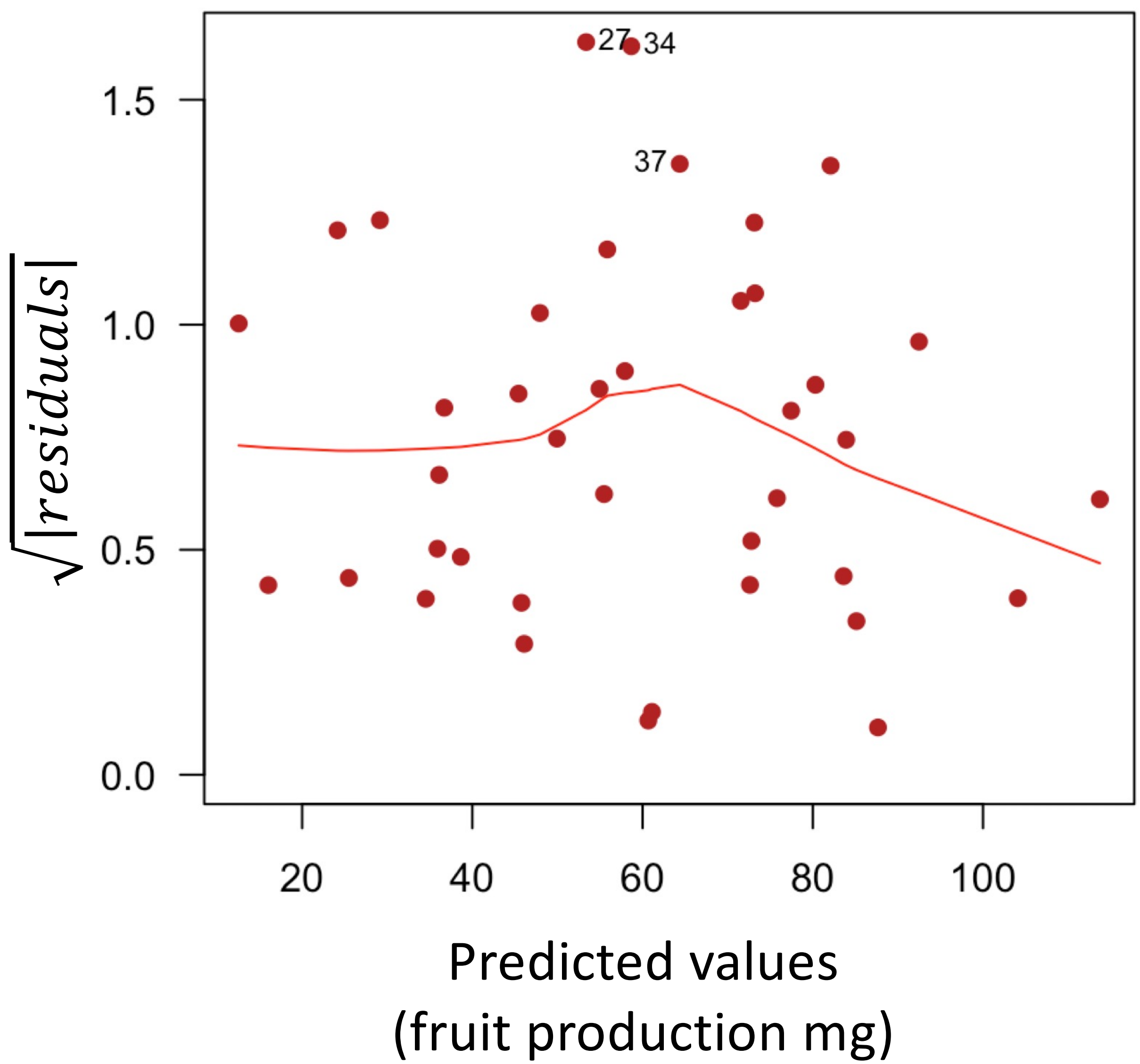



```
> shapiro.test(residuals(lm.result))

        Shapiro-Wilk normality test

data:  residuals(lm.result)
W = 0.97358, p-value = 0.4637
```

In doubt, resort to a formal test, though General Linear models (ANOVAs and regressions) are quite robust against non-normality.

$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$	Analysis of Covariance (ANCOVA)

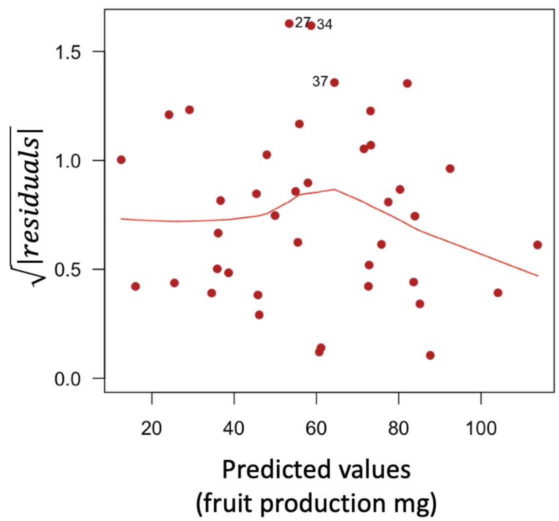
Testing for homoscedasticity



$$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$$

Analysis of Covariance (ANCOVA)

Testing for homoscedasticity

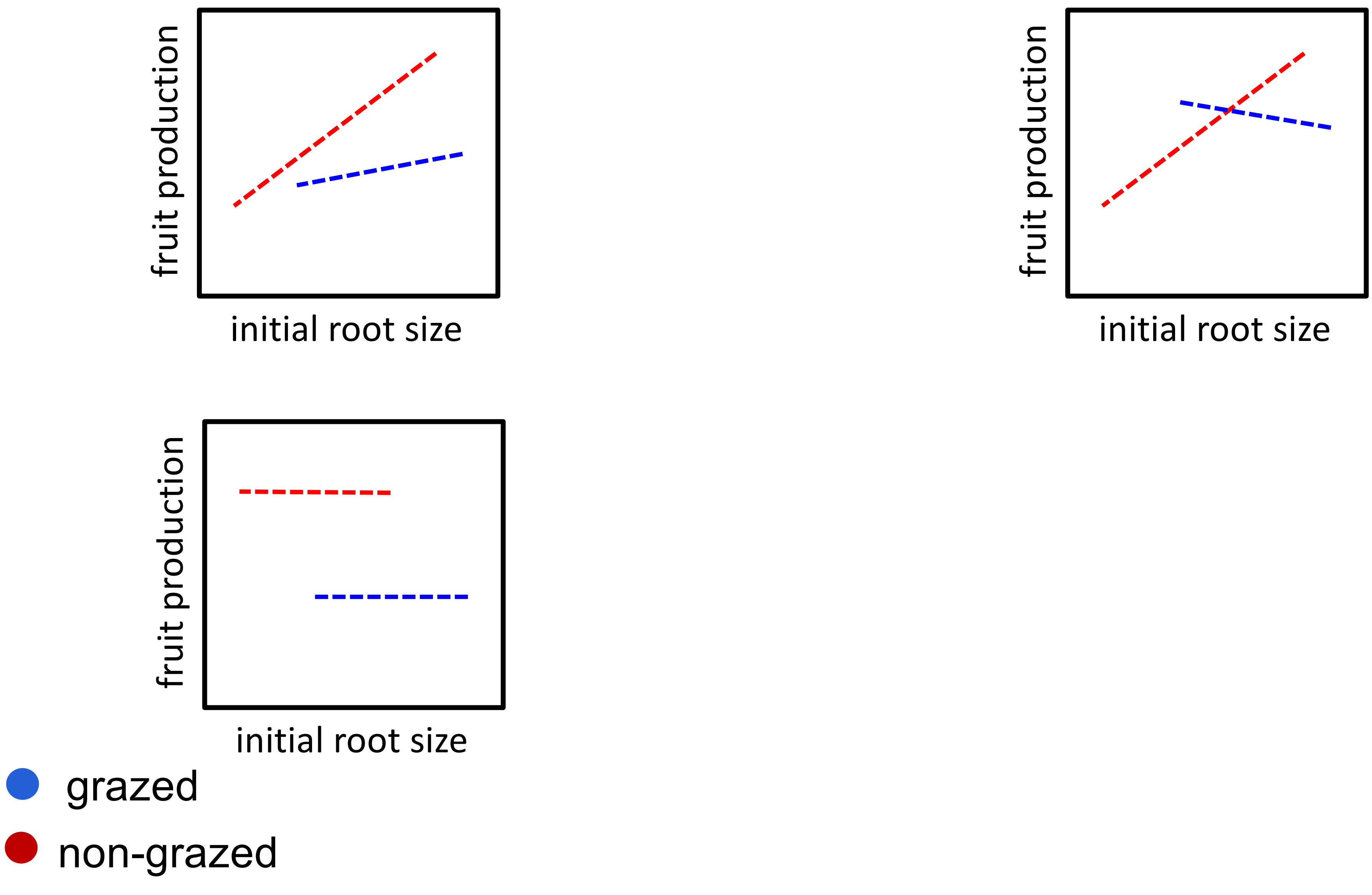



```
> bptest(lm.result)

        studentized Breusch-Pagan test

data:  lm.result
BP = 1.7063, df = 2, p-value = 0.4261
```

In doubt, resort to a formal test, General Linear models are sensitive to heteroscedasticity.

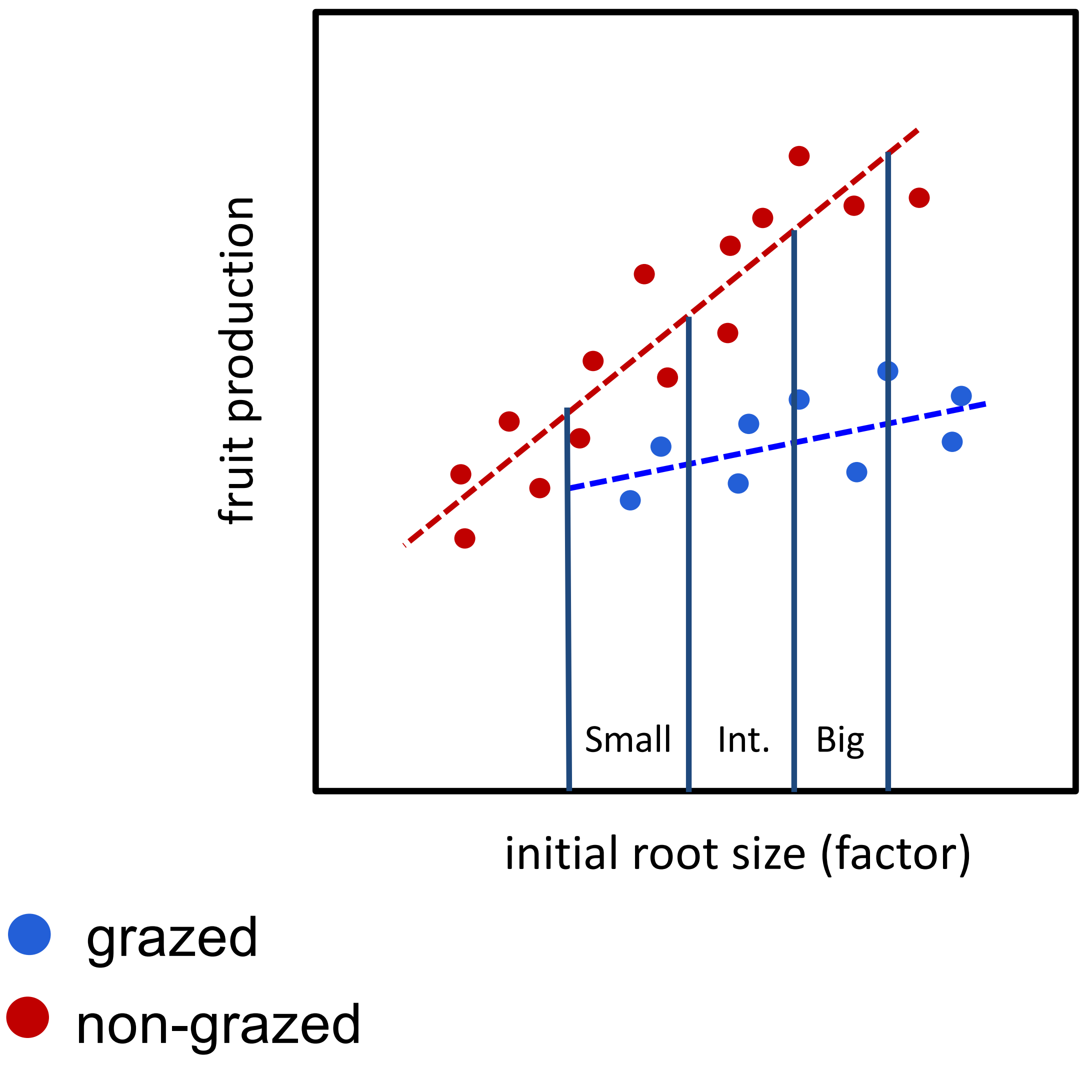
What to do in more complex cases?

There are approaches for the more complex cases when slopes differ between groups or when the response (Y) does not depend on the covariate (initial root size)



There are approaches for the more complex cases when slopes differ between groups or when the response (Y) does not depend on the covariate (initial root size) – later in the class.

When there is an interaction, then the differences in mean values between treatments vary as a function of the covariate, so we can't generalize to all initial root sizes.

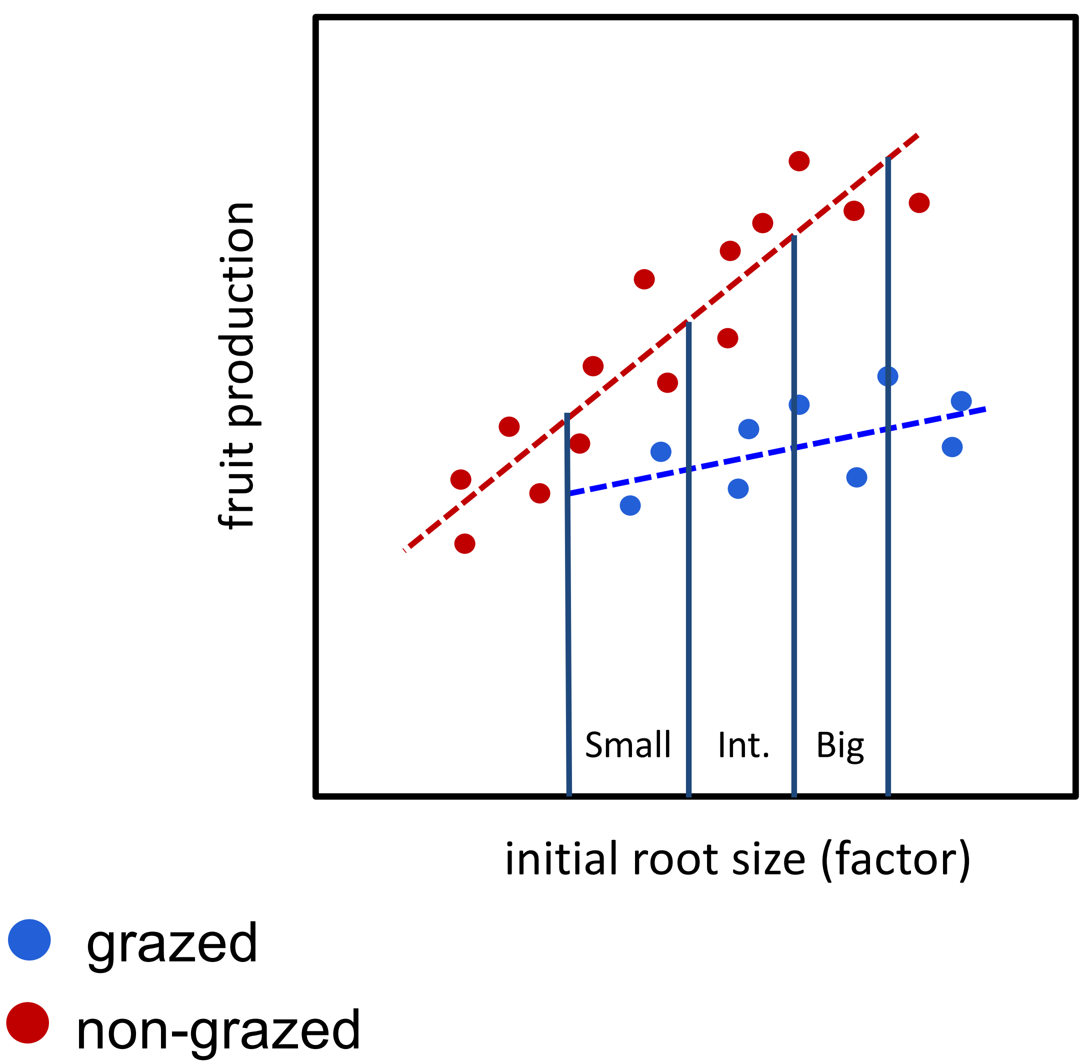


Solution: categorize the covariate (divide into few classes of initial root size) and use a two-factorial design:

1) the interpretation is more complex (i.e., there will be an interaction);

There are approaches for the more complex cases when slopes differ between groups or when the response (Y) does not depend on the covariate (initial root size) – later in the class.

When there is an interaction, then the differences in mean values between treatments vary as a function of the covariate, so we can't generalize to all initial root sizes.



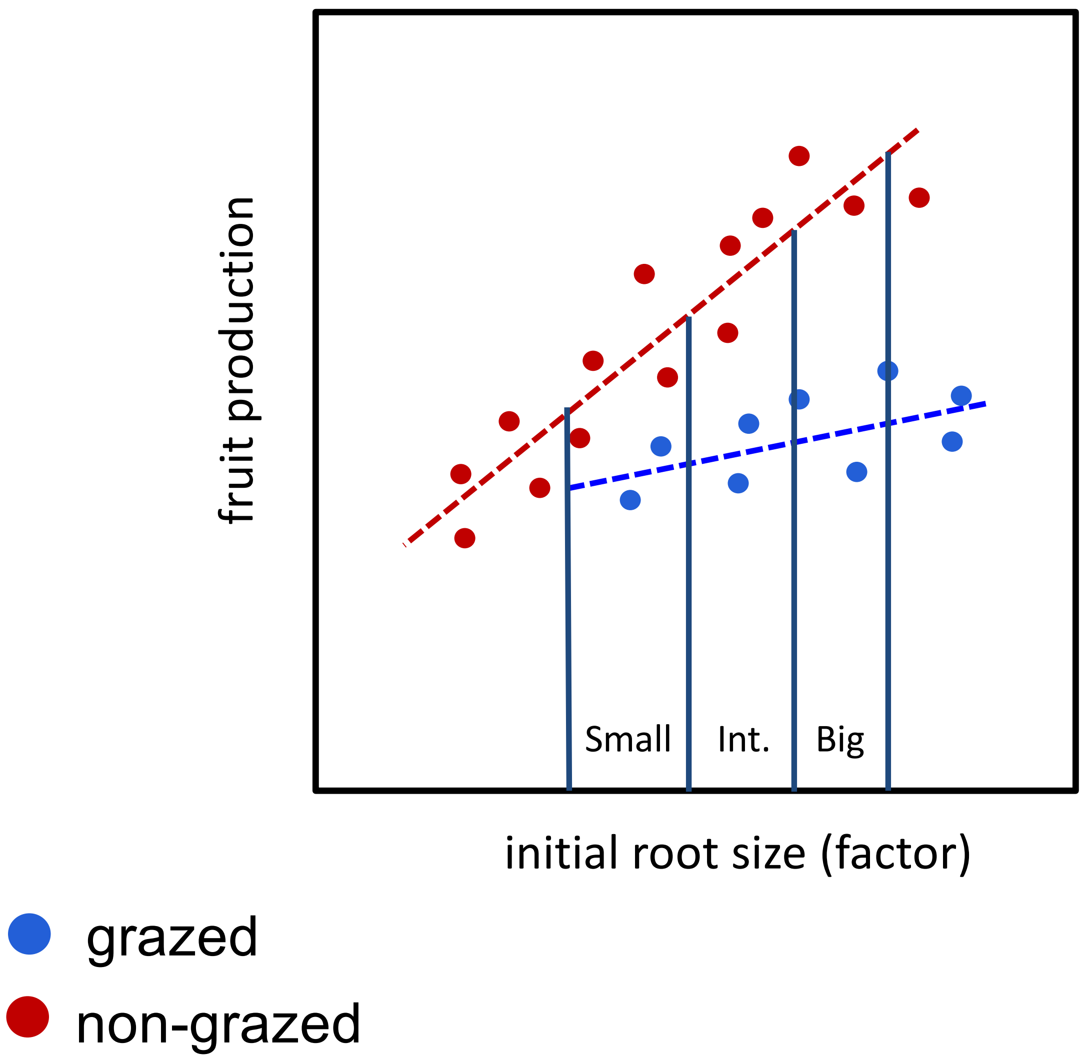
Solution: categorize the covariate (divide into few classes of initial root size) and use a two-factorial design:

1) the interpretation is more complex (i.e., there will be an interaction);

2) Loss of statistical power by decreasing the degrees of freedom via creating categories.

There are approaches for the more complex cases when slopes differ between groups or when the response (Y) does not depend on the covariate (initial root size) – later in the class.

When there is an interaction, then the differences in mean values between treatments vary as a function of the covariate, so we can't generalize to all initial root sizes.



Solution: categorize the covariate (divide into few classes of initial root size) and use a two-factorial design:

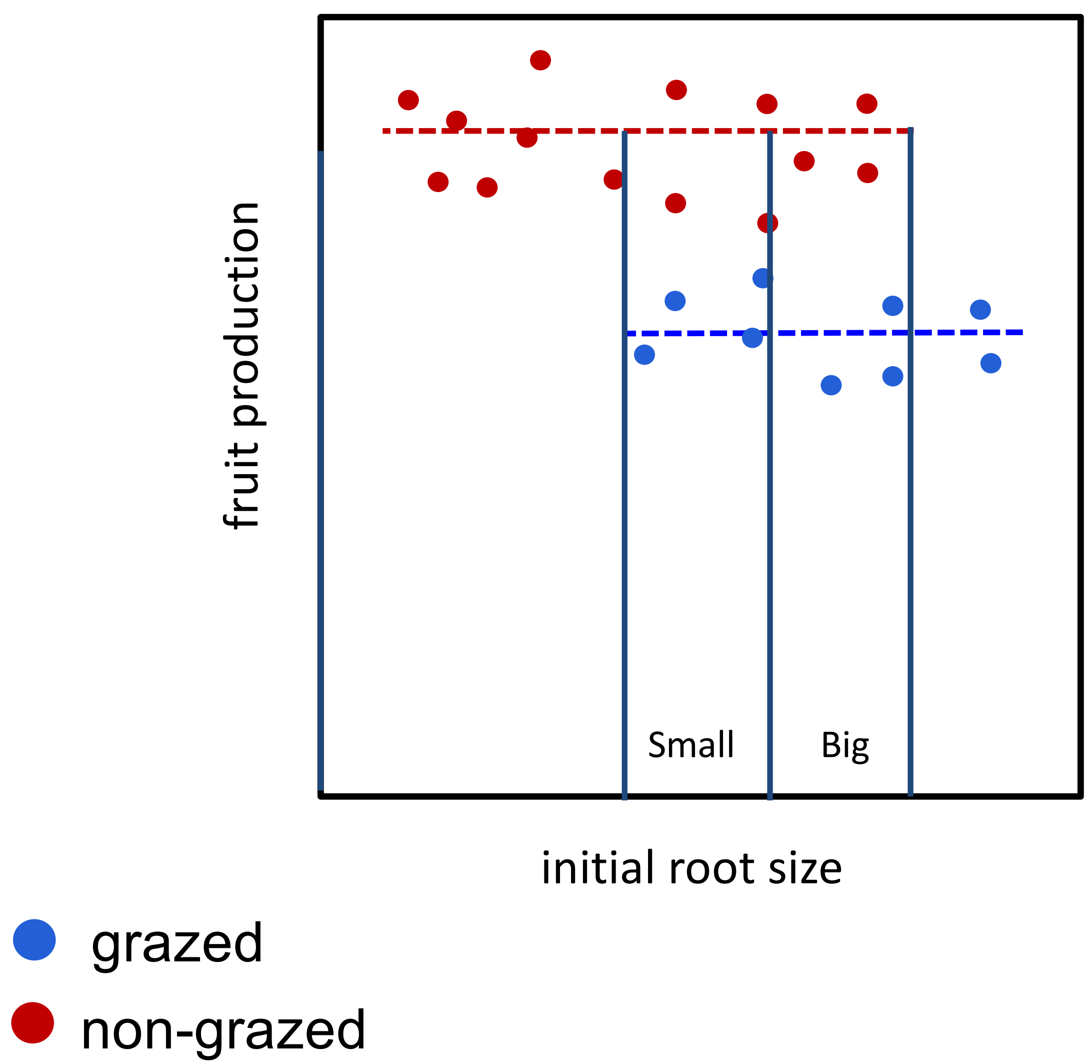
1) the interpretation is more complex (i.e., there will be an interaction);

2) Loss of statistical power by decreasing the degrees of freedom via creating categories.

3) The two series need to overlap substantially in their covariate values.

There are approaches for the more complex cases when slopes differ between groups or when the response (Y) does not depend on the covariate (initial root size) – later in the class.

When response variable (fruit production) is independent of continuous predictor (initial root size), but continuous differ in average between treatments.



Solution: categorize the covariate (divide into few classes of initial root size) and use a two-factorial design:

1) the interpretation is more complex (i.e., there will be an interaction);

2) Loss of statistical power by decreasing the degrees of freedom via creating categories.

3) The two series need to overlap substantially in their covariate values.

$Y = \mu + A_1 + X_1 + (A_1 \times X_1)$	Analysis of Covariance (ANCOVA)

- It is not always possible to randomize factors completely independent of each other. In the case of the fruit productivity, ideally the researchers should have made sure that the plants in grazing and no grazing plots should have had the same size.

- Confounding or nuisance (non-random) factors can often be the case, particularly in non-experimental studies.

- The terminology and some of the theory underlying "Type I, II & III" sum of squares seems to have been generated by SAS (Statistical Analysis System).

Doctor Tyrano, look for a covariate

Doctor Tyrano, stewed in the realization that he would win no accolades for finding the world's most medium-sized dinosaur!

General linear models (not Generalized linear model)

Linear Model	Common name
$Y = \mu + X$	Simple linear regression
$Y = \mu + A_1$	One-factorial (one-way) ANOVA
$Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
$Y = \mu + A_1 + X\ (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
$Y = \mu + X_1 + X_2 + X_3$	Multiple regression
$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
$Y_1 + Y_2 = \mu + A_1 + A_2 + A_1 \times A_2$	Multivariate ANOVA (MANOVA)

A represents categorical predictors (factors)

g represents groups of data (more on this later)

$(+A_1 \times X)$ - step 1 on an ANCOVA, but not in the final analysis

Multiple factors $A_1 + A_2 + \text{etc (and their interactions)}$