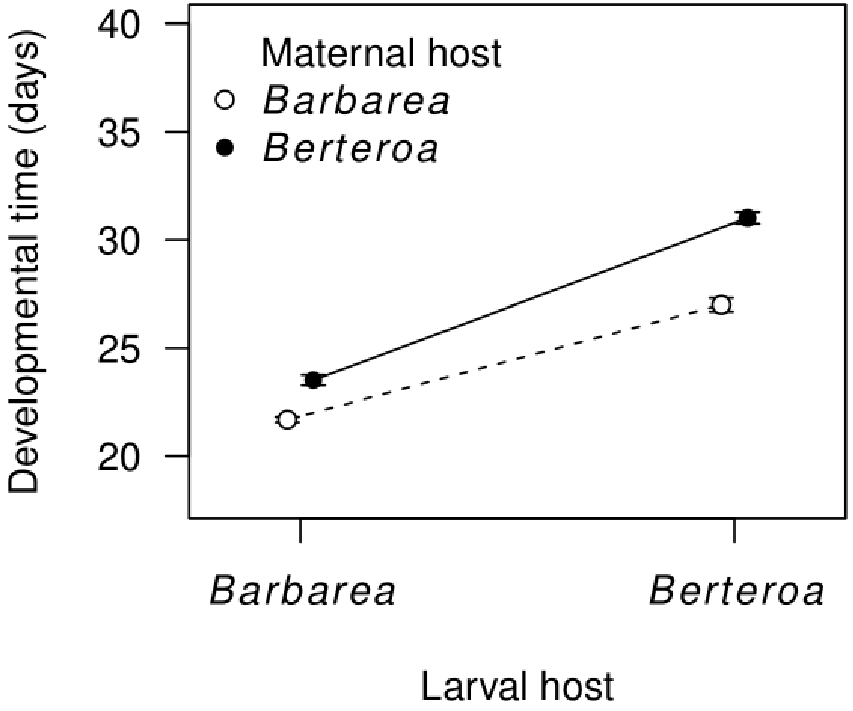
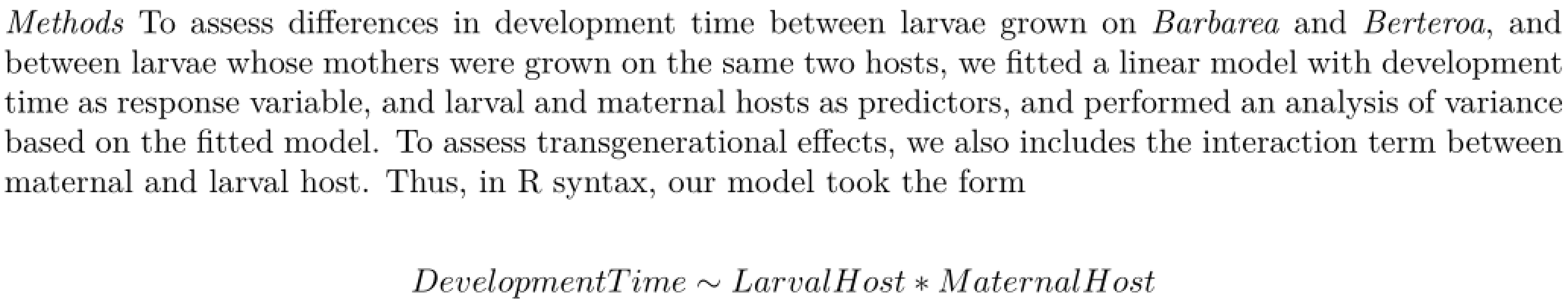
Discussion of exercise 3

- ANOVA analysis of factorial experiment

- “Interaction plots” illustrate the interactive effect of two factors
- Here, the effect of the maternal host is slightly stronger when the larval host is *Berteroa*, than when the larval host is *Barbarea*



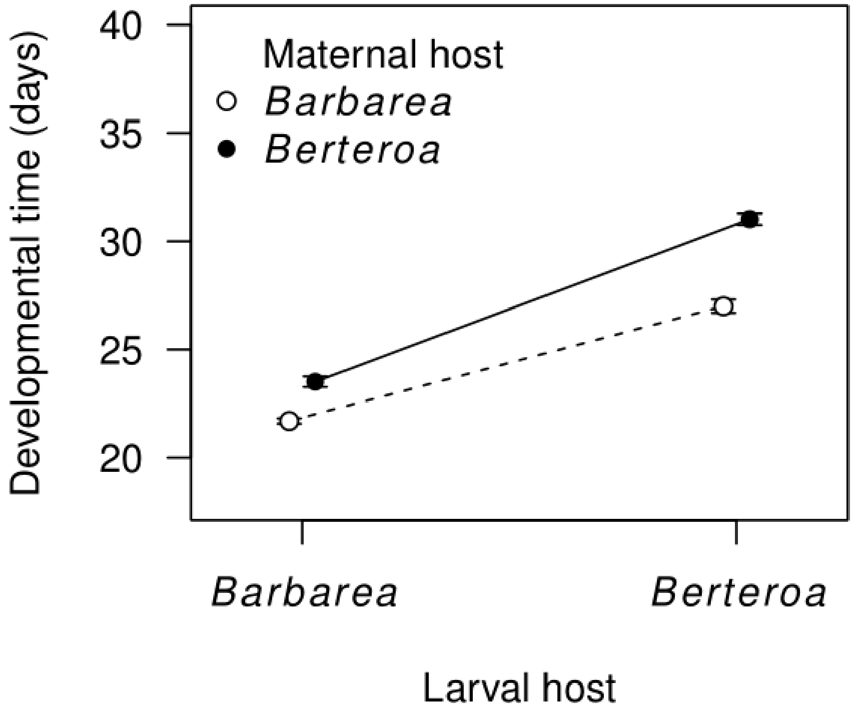
Suggested analysis methods



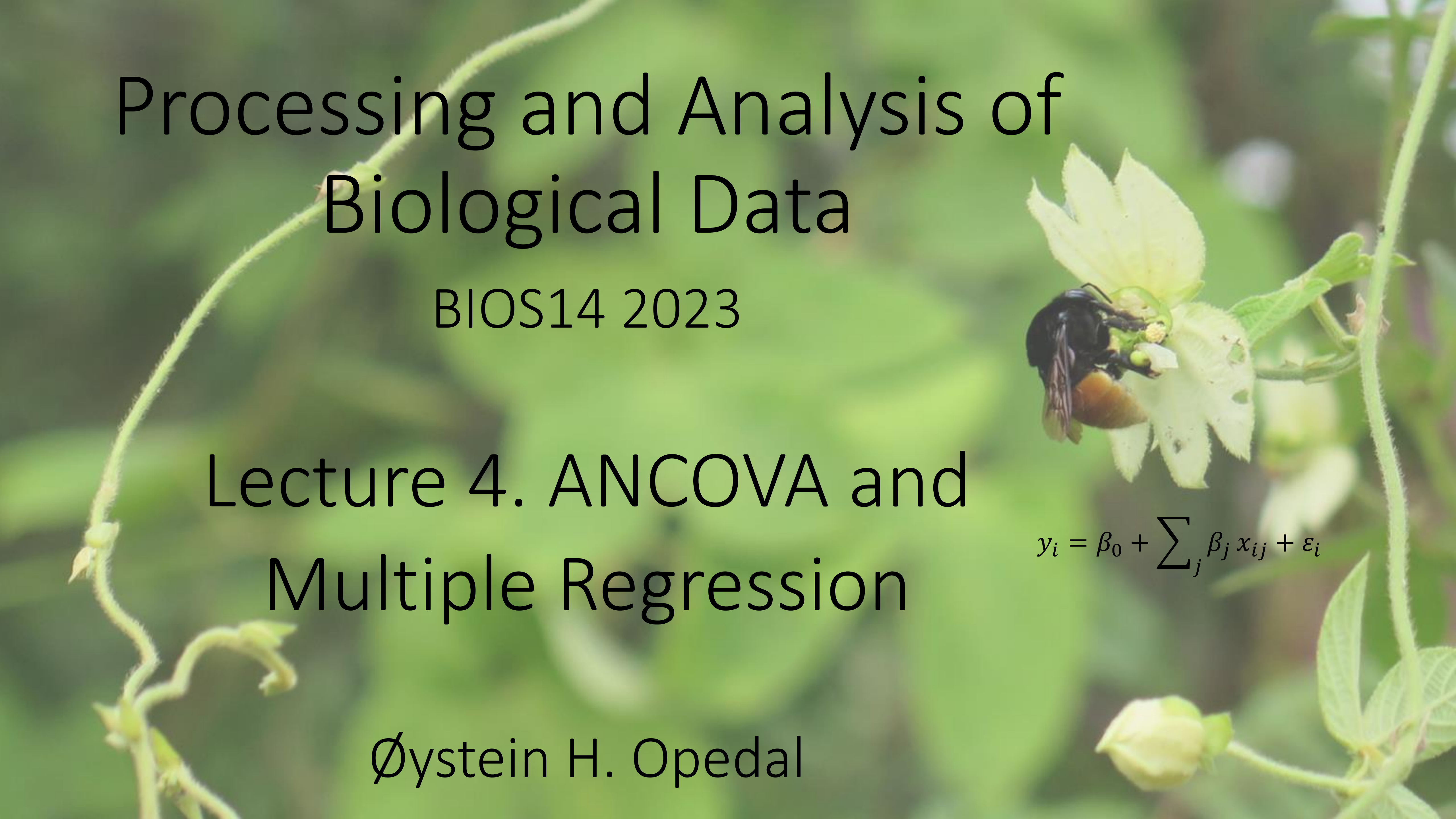
Methods To assess differences in development time between larvae grown on *Barbarea* and *Berteroa*, and between larvae whose mothers were grown on the same two hosts, we fitted a linear model with development time as response variable, and larval and maternal hosts as predictors, and performed an analysis of variance based on the fitted model. To assess transgenerational effects, we also includes the interaction term between maternal and larval host. Thus, in R syntax, our model took the form

$$DevelopmentTime \sim LarvalHost * MaternalHost$$

Suggested results



Results The larvae developed 22.1% faster when grown on Barbarea than when grown on Berteroa (mean development time = 22.6 and 29.0 days, respectively, $F_{1,283} = 765.21$, Fig. 1). Larvae whose mothers were grown on Barbarea developed 10.7% faster (mean development time = 24.3 and 27.3 days, respectively, $F_{1,283} = 177.90$). The difference in development time between larval host plants was slightly larger when the mother was grown on Berteroa than when the mother was grown on Barbarea (24.2% vs. 19.6% reduction in developmental time on Barbarea, respectively).



Processing and Analysis of Biological Data

BIOS14 2023

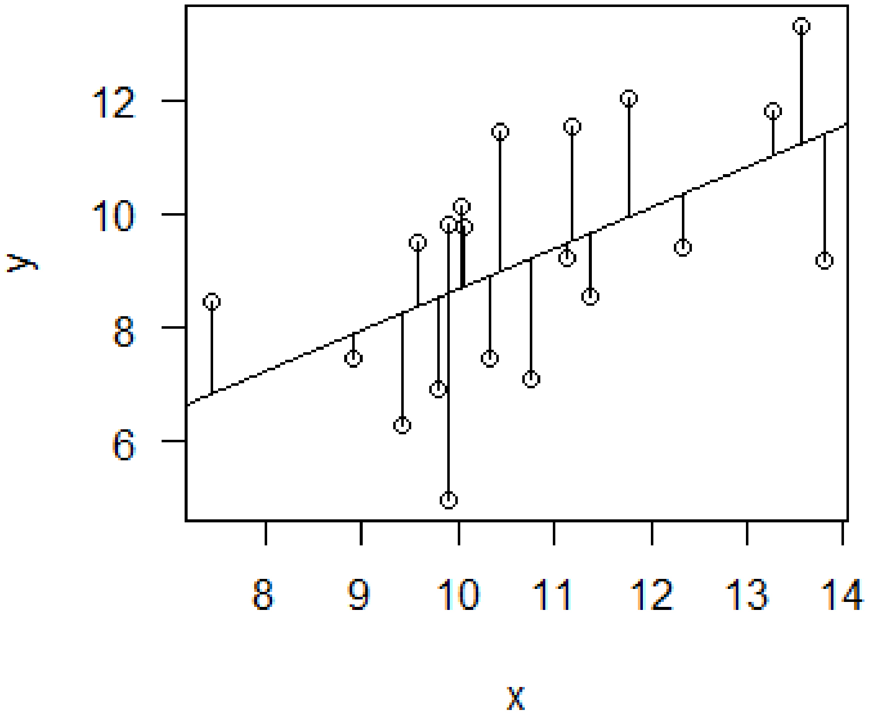
Lecture 4. ANCOVA and Multiple Regression

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i$$

Øystein H. Opedal

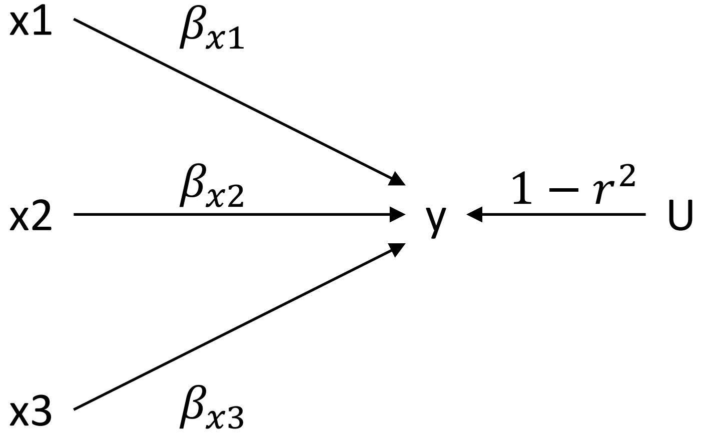
The linear model

- Most of the models we will work with in this course are linear models, that describe how a linear set of predictors relate to a response variable
- A key element of the model is the so-called linear predictor:
- $y_i = \beta_0 + \sum_j \beta_j \, x_{ij} + \varepsilon_i, \, \varepsilon \sim N(0, \sigma^2)$
- The term $\varepsilon \sim N(0, \sigma^2)$ means that the residuals (epsilon) are assumed to follow a normal distribution



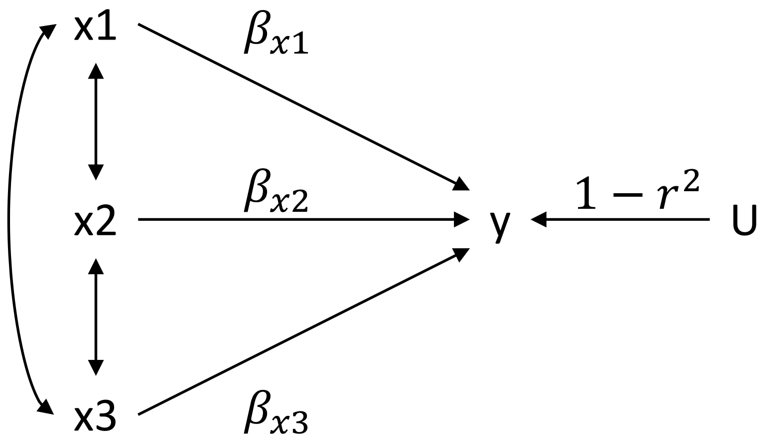
Multiple regression

- A linear model with multiple continuous predictors is called a multiple regression.
- Each slope is estimated while holding the other predictors constant, and are thus **marginal effects**.



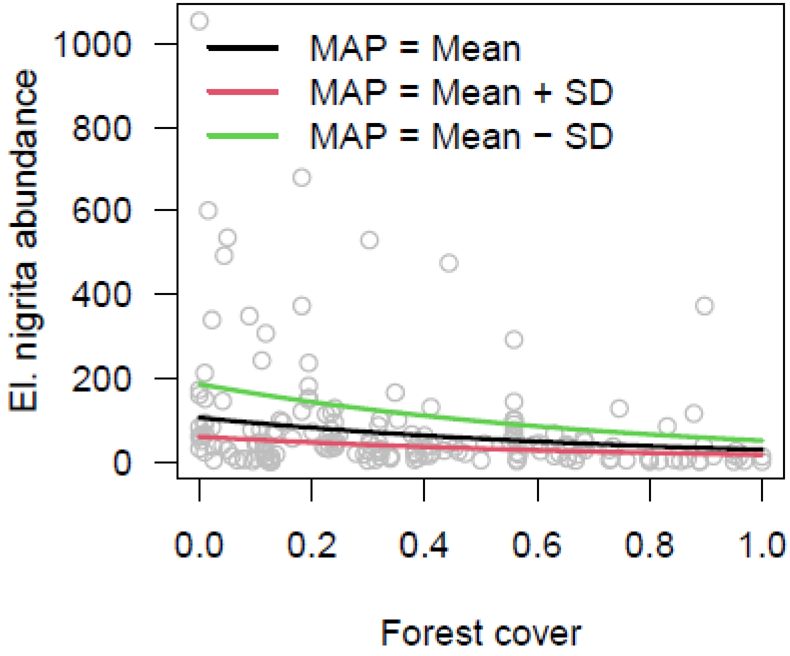
Multiple regression

- A linear model with multiple continuous predictors is called a multiple regression.
- Each slope is estimated while holding the other predictors constant, and are thus **marginal effects**.
- The **net effect** of a predictor (as detected in a univariate model) will also include indirect effects of other variables

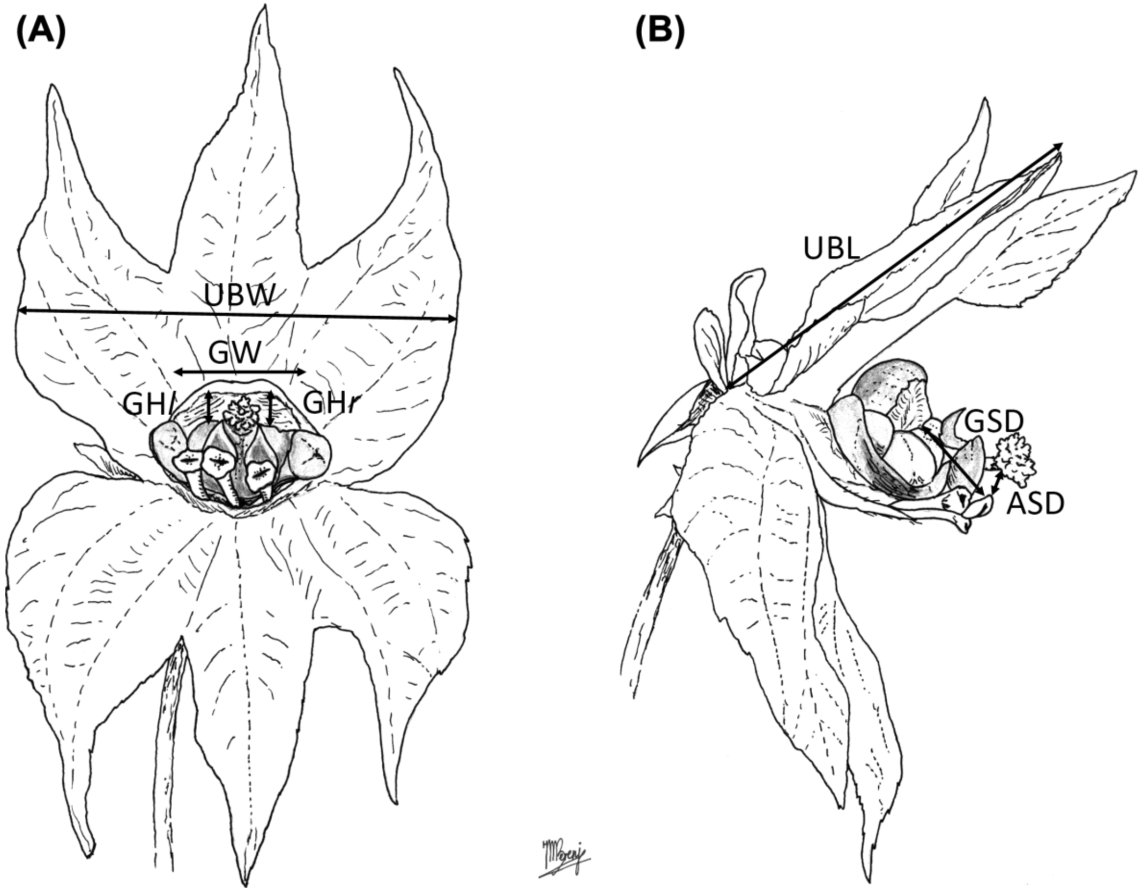


Multiple regression

- A linear model with multiple continuous predictors is called a multiple regression.
- Each slope is estimated while holding the other predictors constant, and are thus **marginal effects**.
- The **net effect** of a predictor (as detected in a univariate model) will also include indirect effects of other variables



Example: multivariate selection gradients



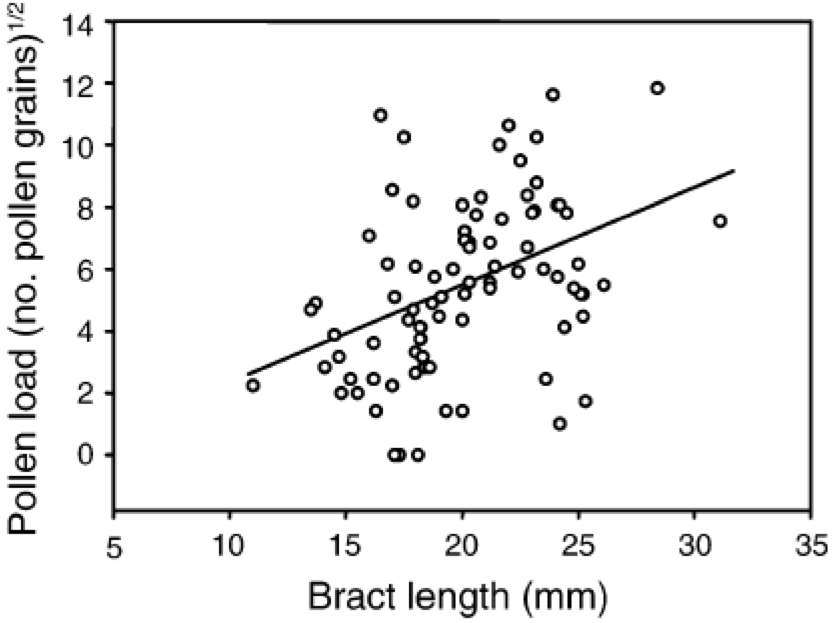
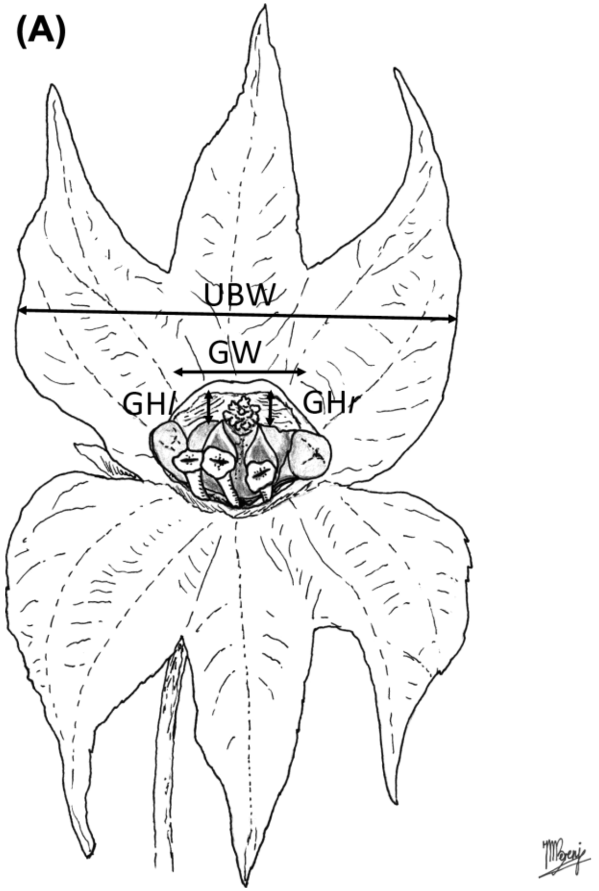
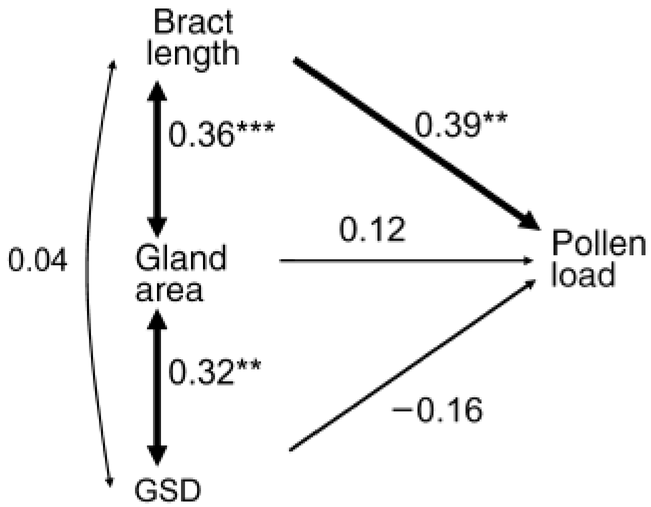


FIG. 4. Effect of natural variation in bract length on the amount of pollen found on the stigmas at the end of the female phase (i.e., pollen transported to stigmas by pollinators). The coefficients (±SE) for the regression line are: intercept $=-0.88 \pm 1.55$, slope $= 0.31 \pm 0.07$; $R^2 = 0.18$. See Fig. 5 for significance testing.

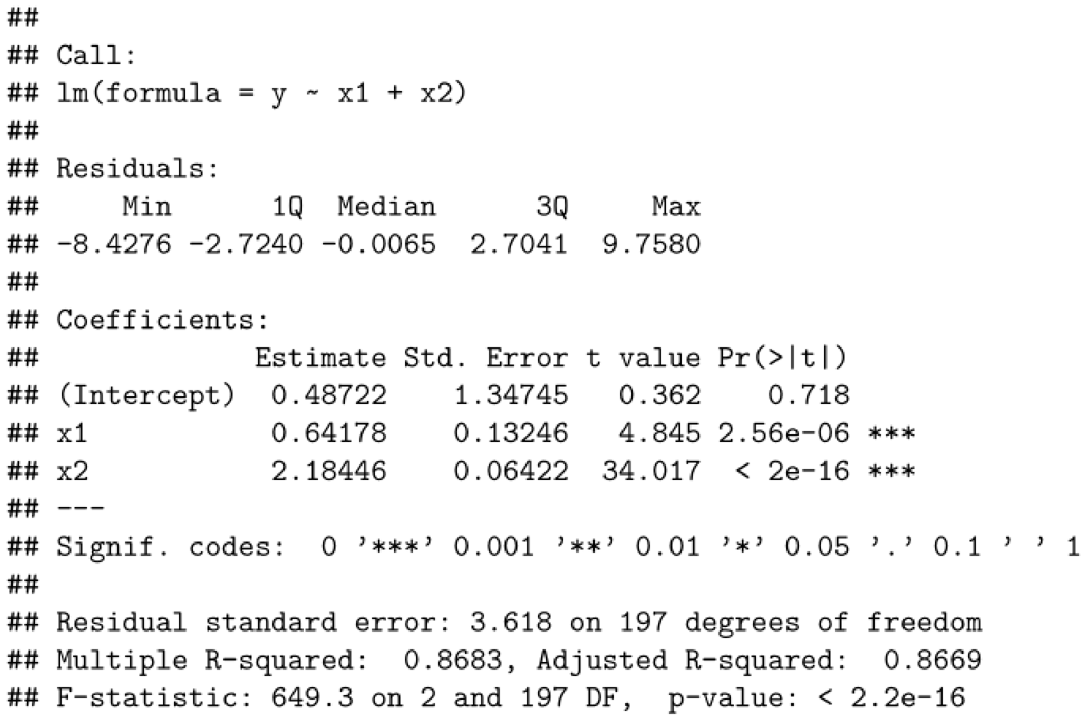
Example: multivariate selection gradients





Multiple-regression model in R

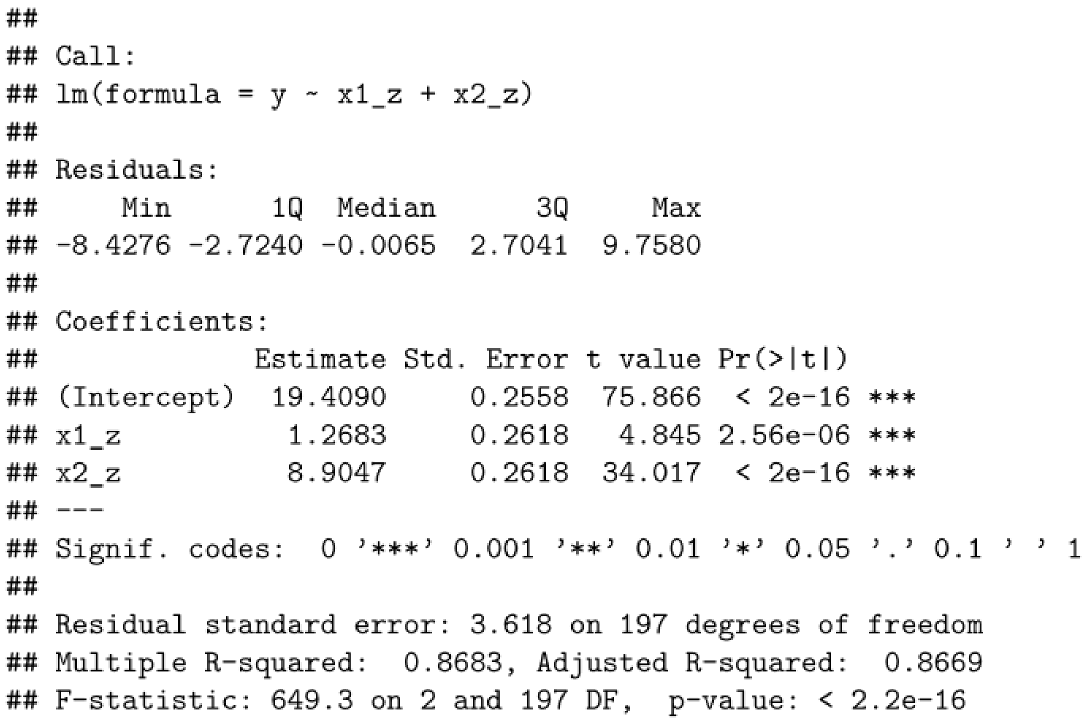
- The parameter estimates from a multiple-regression model are marginal effects, i.e. the effect with all other variables held constant



```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4276 -2.7240 -0.0065  2.7041  9.7580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48722    1.34745   0.362    0.718
## x1           0.64178    0.13246   4.845 2.56e-06 ***
## x2           2.18446    0.06422  34.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 197 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8669
## F-statistic: 649.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Multiple-regression model in R

- The parameter estimates from a multiple-regression model are marginal effects, i.e. the effect with all other variables held constant
- If we standardize the predictors, we can compare the strength of effects across variables (for example in units of standard deviations)



```
##
## Call:
## lm(formula = y ~ x1_z + x2_z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4276 -2.7240 -0.0065  2.7041  9.7580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4090     0.2558  75.866  < 2e-16 ***
## x1_z          1.2683     0.2618   4.845 2.56e-06 ***
## x2_z          8.9047     0.2618  34.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.618 on 197 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8669
## F-statistic: 649.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

$$z = \frac{x - \bar{x}}{\sigma(x)}$$

Overfitting and multicollinearity

- When we increase the number of variables in the model, we risk problems with “overfitting”, that is fitting a model that explains much variance, but makes poor predictions for independent data
- If the independent (predictor) variables are strongly correlated, this can lead to imprecise estimates (multicollinearity).
- Thus, we often want to select the simplest, most parsimonous model we can (cf. “Occam´s Razor”).
- We can quantify these effects through variance inflation factors, or through cross-validation.

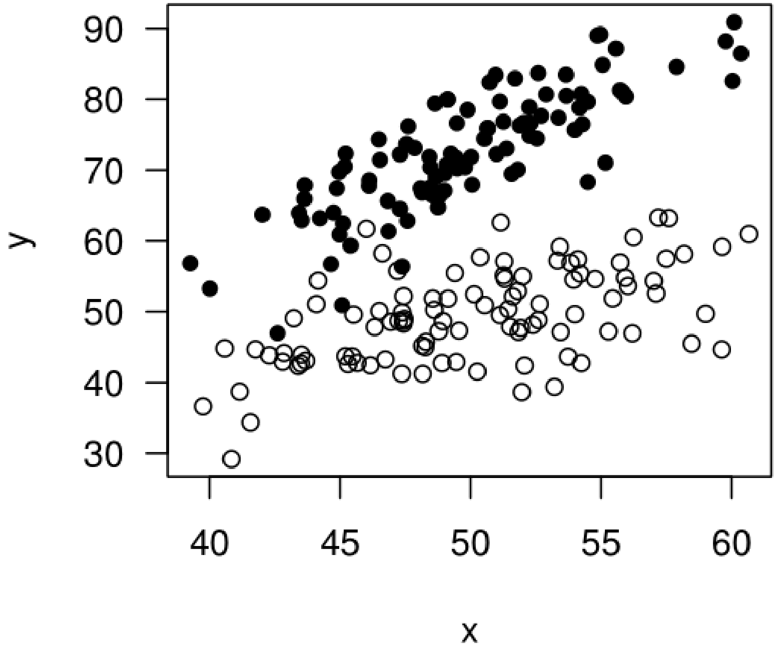
Variance inflation factors

- Variance inflation factors quantify potentialy problematic multicollinarity with values greater than 3(5) considered problematic
- Can be quantified by the ratio 1/(1- r^2), where the r^2 is for a model predicting the focal variable, with all other predictors as explanatory variables

$$VIF_i = \frac{1}{1-r_i^2}$$

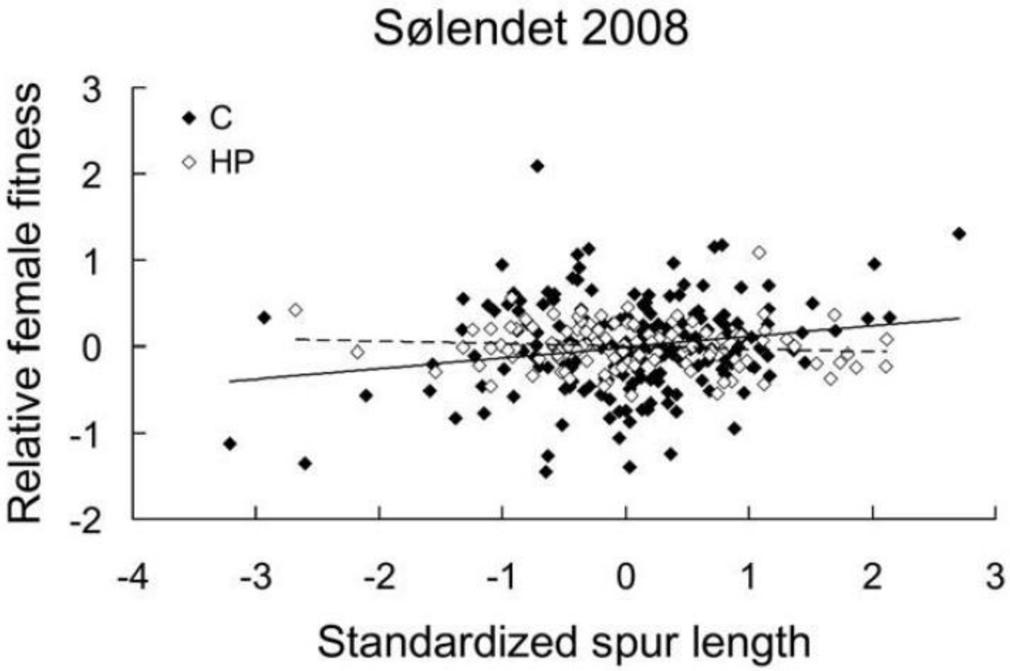
Analysis of Covariance (ANCOVA)

- Analyses of covariance are linear models with both continuous and categorical predictors
- We can use these models to assess whether the slope of a regression differs between groups (e.g. treatments)
- Residuals are assumed to be normally distributed within each group, and variances are assumed to be equal



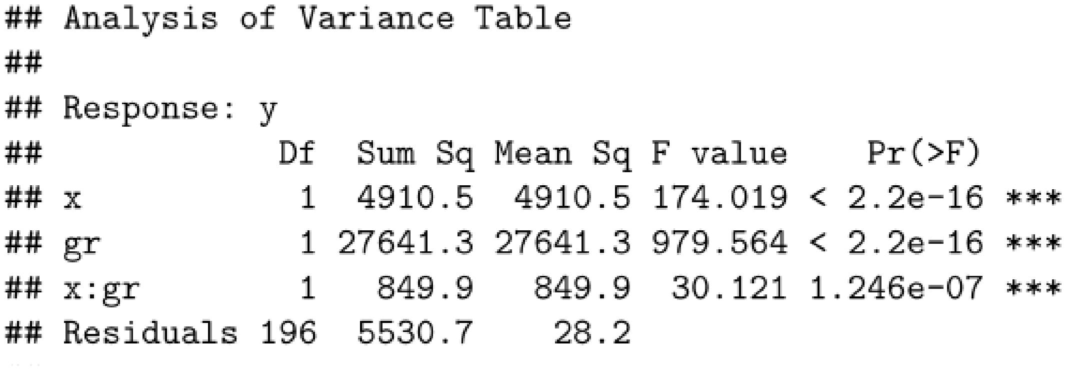
Example: pollinator-mediated selection

- Analysis of covariance can be used to assess differences in slopes between experimental treatments
- In this case, to show that the relationship between floral spur length and fitness is less steep, as expected, when plants are hand-pollinated



ANCOVA model in R

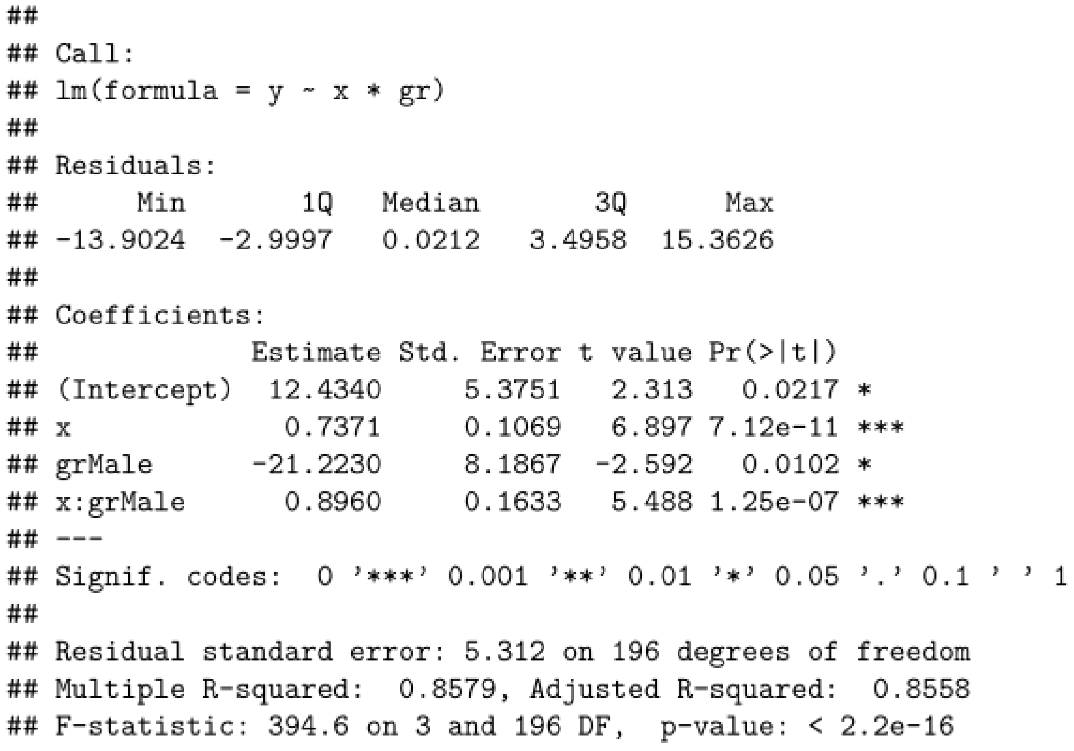
- The ANOVA table gives the **sums of squares** associated with each predictor, i.e. the sum of square deviations from the predicted value (mean).
- The interaction term tests for heterogeneity of slopes, and the main effect of the grouping variable tests for different intercepts



```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## x           1  4910.5  4910.5 174.019 < 2.2e-16 ***
## gr          1 27641.3 27641.3 979.564 < 2.2e-16 ***
## x:gr        1   849.9   849.9  30.121 1.246e-07 ***
## Residuals 196  5530.7    28.2
```

ANCOVA model in R

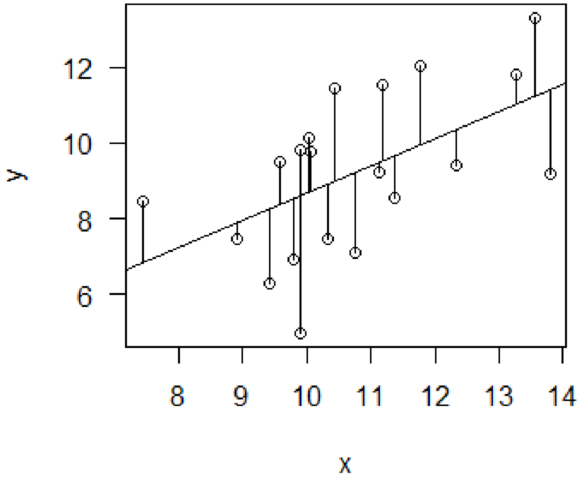
- As for all linear models, the summary table gives the parameter estimates, their standard errors, and other model statistics
- In an ANCOVA model, the intercept gives the intercept for the categorical reference level (first level of the factor “gr”, here “Female”).
- The parameter “grMale” gives the contrast between the male and female intercepts.

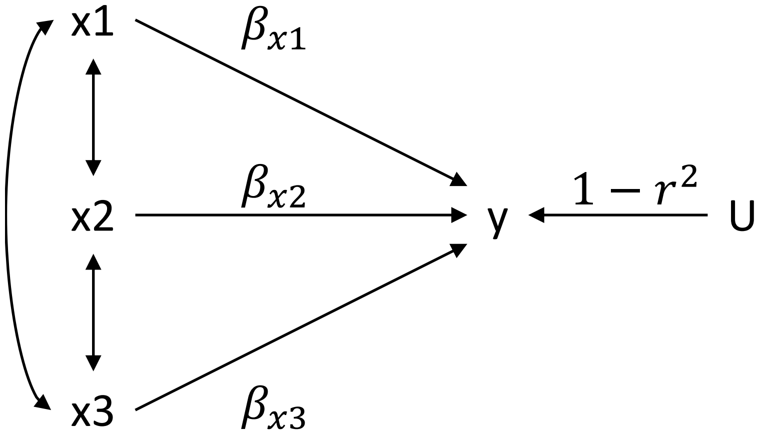


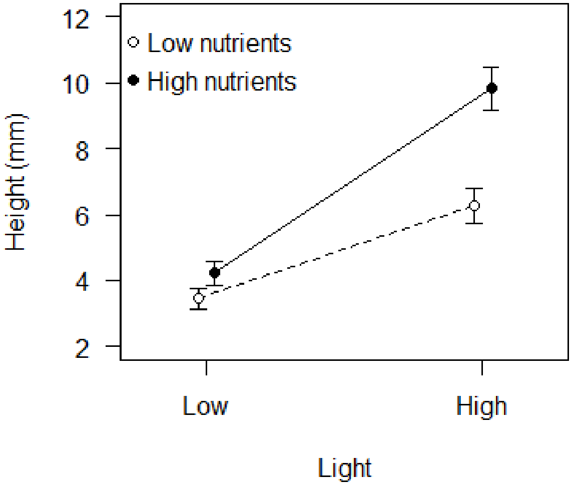
```
##
## Call:
## lm(formula = y ~ x * gr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9024  -2.9997   0.0212   3.4958  15.3626
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4340     5.3751   2.313   0.0217 *
## x             0.7371     0.1069   6.897 7.12e-11 ***
## grMale      -21.2230     8.1867  -2.592   0.0102 *
## x:grMale      0.8960     0.1633   5.488 1.25e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.312 on 196 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8558
## F-statistic: 394.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

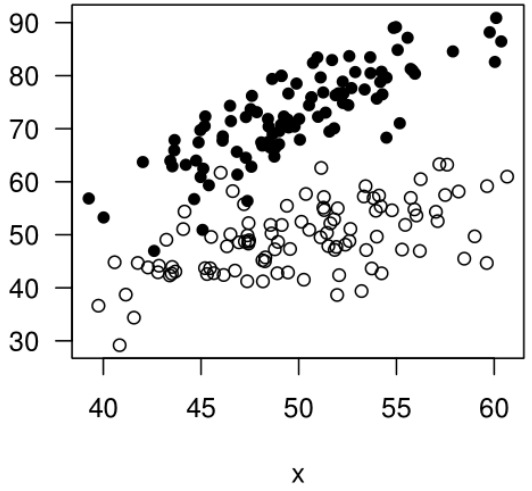
Overview of linear models

- Continuous covariates: (multiple) regression
- Categorical covariates: N-way ANOVA
- Continuous and categorical covariates: ANCOVA









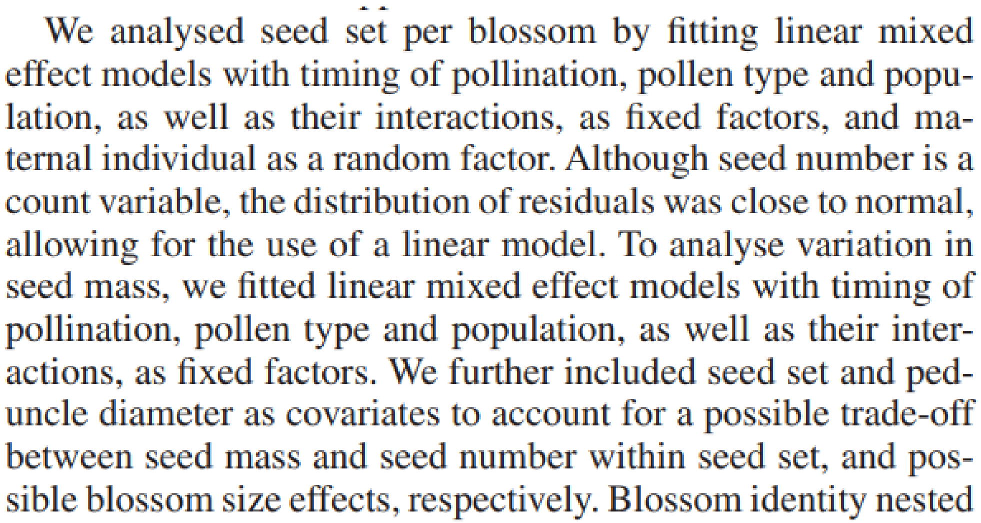
General intro to exercises

- You can work with any dataset you may have!
- We will provide one or more potential datasets
- All exercises will involve formulating a research question, choosing analysis methods, performing the analysis, interpreting the results, and writing relevant Methods and Results
- The mid-term exercise and the exam will be similar
- Towards the end of each exercise session, we will show some examples of how the data could be analysed and the results presented
- No written feedback on the exercises (except mid-term), but we strongly encourage you to give feedback on each other’s work.

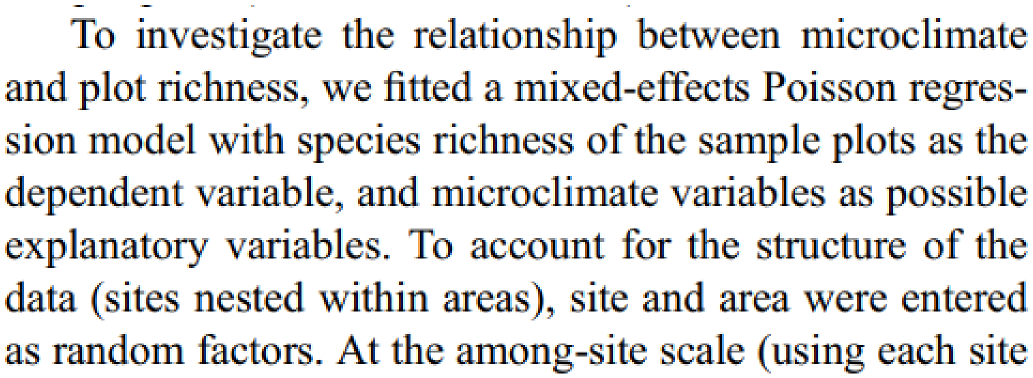
Writing (analysis) methods

- Normally a specific section at the end of the Methods (but sometimes integrated throughout)
- Focus on the aim of the analysis before technical details
- For models, list terms in words, sometimes include model equation
- R syntax is increasingly acceptable

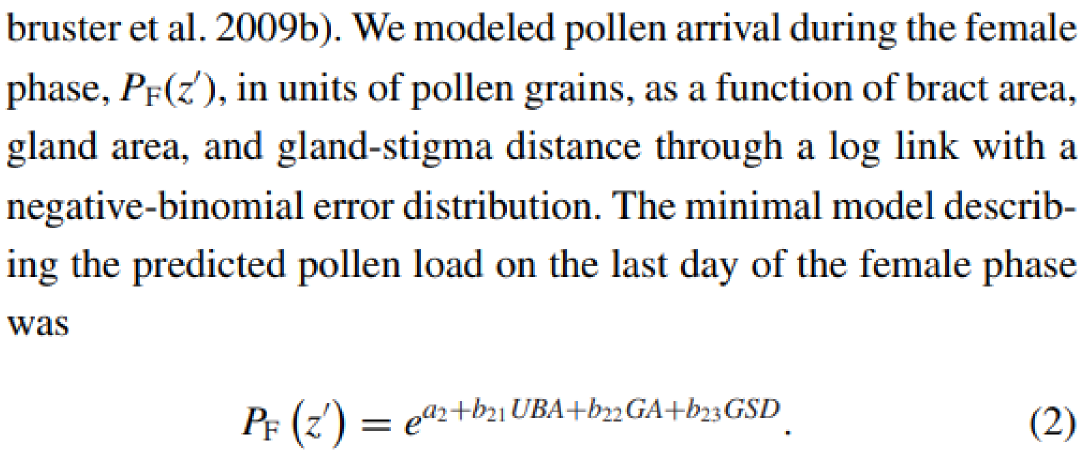
Examples of model description



We analysed seed set per blossom by fitting linear mixed effect models with timing of pollination, pollen type and population, as well as their interactions, as fixed factors, and maternal individual as a random factor. Although seed number is a count variable, the distribution of residuals was close to normal, allowing for the use of a linear model. To analyse variation in seed mass, we fitted linear mixed effect models with timing of pollination, pollen type and population, as well as their interactions, as fixed factors. We further included seed set and peduncle diameter as covariates to account for a possible trade-off between seed mass and seed number within seed set, and possible blossom size effects, respectively. Blossom identity nested



To investigate the relationship between microclimate and plot richness, we fitted a mixed-effects Poisson regression model with species richness of the sample plots as the dependent variable, and microclimate variables as possible explanatory variables. To account for the structure of the data (sites nested within areas), site and area were entered as random factors. At the among-site scale (using each site

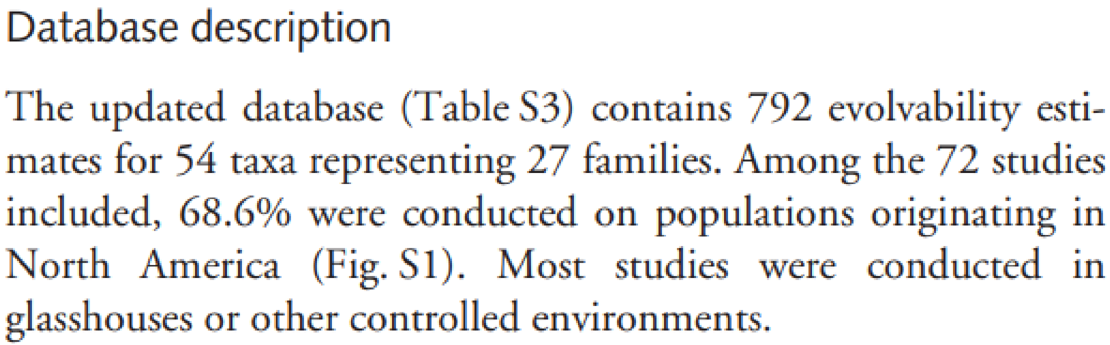


bruster et al. 2009b). We modeled pollen arrival during the female phase, $P_F(z')$, in units of pollen grains, as a function of bract area, gland area, and gland-stigma distance through a log link with a negative-binomial error distribution. The minimal model describing the predicted pollen load on the last day of the female phase was

$$P_F\left(z'\right) = e^{a_2+b_{21}UBA+b_{22}GA+b_{23}GSD}. \qquad (2)$$

Writing results

- Start with summary statistics/patterns of variation

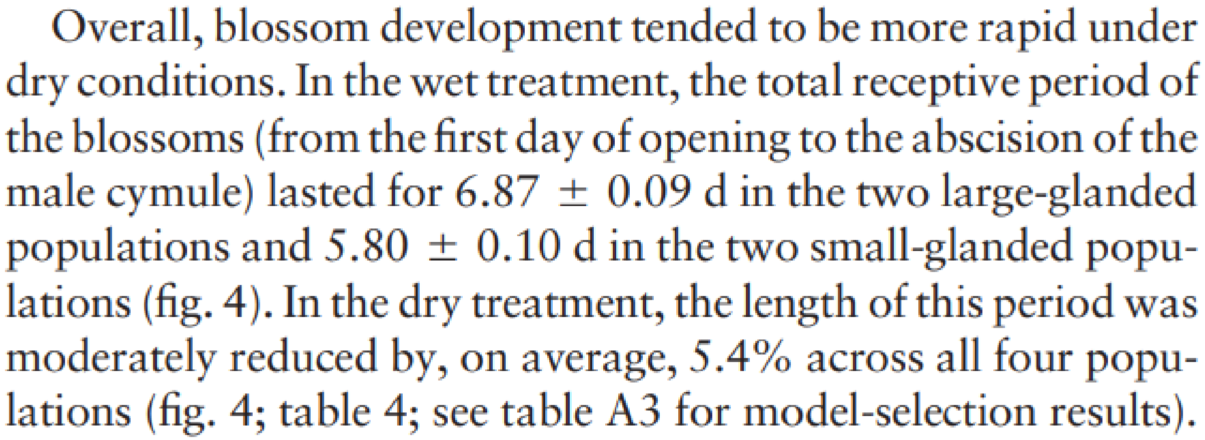


Database description

The updated database (Table S3) contains 792 evolvability estimates for 54 taxa representing 27 families. Among the 72 studies included, 68.6% were conducted on populations originating in North America (Fig. S1). Most studies were conducted in glasshouses or other controlled environments.

Writing results

- Start with summary statistics/patterns of variation
- Focus on biology over statistics (explain the result in biological terms)
- Quantify and exemplify (*y* increased by 0.5 mm per mm increase in *x*)



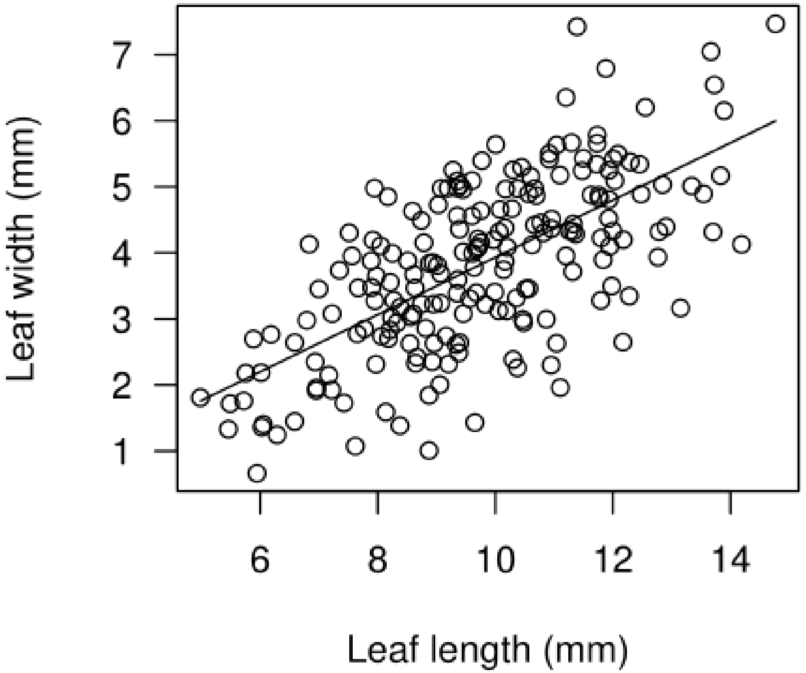
Overall, blossom development tended to be more rapid under dry conditions. In the wet treatment, the total receptive period of the blossoms (from the first day of opening to the abscision of the male cymule) lasted for 6.87 ± 0.09 d in the two large-glanded populations and 5.80 ± 0.10 d in the two small-glanded populations (fig. 4). In the dry treatment, the length of this period was moderately reduced by, on average, 5.4% across all four populations (fig. 4; table 4; see table A3 for model-selection results).

Writing results

- Start with summary statistics/patterns of variation
- Focus on biology over statistics (explain the result in biological terms)
- Quantify and exemplify (*y* increased by 0.5 mm per mm increase in *x*)
- No discussion (normally)
- Avoid unnecessary introductory sentences like “Population means are reported in Table 1”.
- Refer to tables and figures after a statement (“Body size varied substantially among populations (range = X.X to Y.Y, Table 1)”.

Making nice figures

- Units!
- Keep regression lines within the data range
- If several symbols/colors: legend inside the graph



Making nice tables

- No vertical lines
- All parameters explained in legend

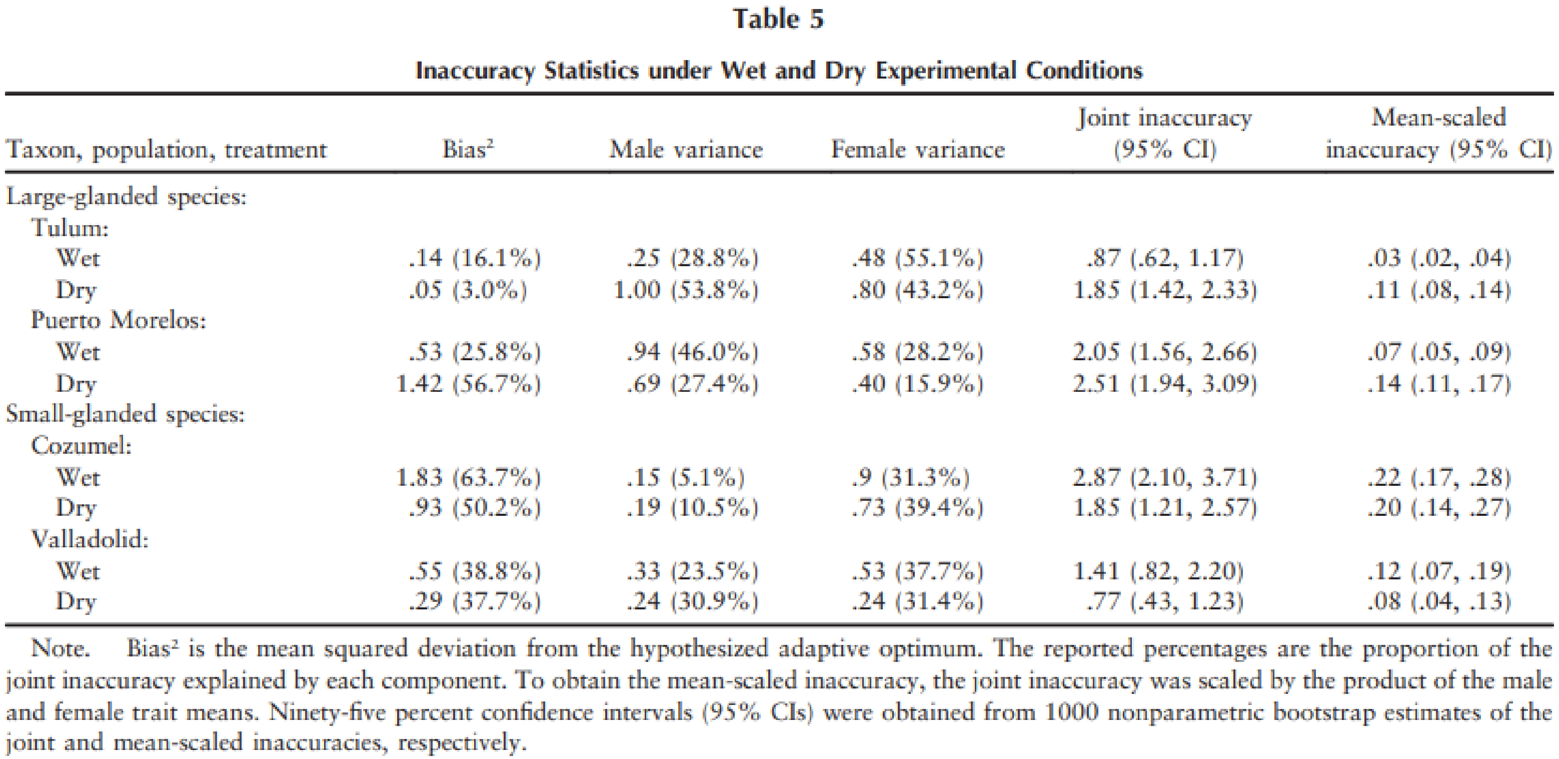


Table 5

Inaccuracy Statistics under Wet and Dry Experimental Conditions

Taxon, population, treatment	Bias^2	Male variance	Female variance	Joint inaccuracy (95% CI)	Mean-scaled inaccuracy (95% CI)
Large-glanded species:					
Tulum:					
Wet	.14 (16.1%)	.25 (28.8%)	.48 (55.1%)	.87 (.62, 1.17)	.03 (.02, .04)
Dry	.05 (3.0%)	1.00 (53.8%)	.80 (43.2%)	1.85 (1.42, 2.33)	.11 (.08, .14)
Puerto Morelos:					
Wet	.53 (25.8%)	.94 (46.0%)	.58 (28.2%)	2.05 (1.56, 2.66)	.07 (.05, .09)
Dry	1.42 (56.7%)	.69 (27.4%)	.40 (15.9%)	2.51 (1.94, 3.09)	.14 (.11, .17)
Small-glanded species:					
Cozumel:					
Wet	1.83 (63.7%)	.15 (5.1%)	.9 (31.3%)	2.87 (2.10, 3.71)	.22 (.17, .28)
Dry	.93 (50.2%)	.19 (10.5%)	.73 (39.4%)	1.85 (1.21, 2.57)	.20 (.14, .27)
Valladolid:					
Wet	.55 (38.8%)	.33 (23.5%)	.53 (37.7%)	1.41 (.82, 2.20)	.12 (.07, .19)
Dry	.29 (37.7%)	.24 (30.9%)	.24 (31.4%)	.77 (.43, 1.23)	.08 (.04, .13)

Note. Bias^2 is the mean squared deviation from the hypothesized adaptive optimum. The reported percentages are the proportion of the joint inaccuracy explained by each component. To obtain the mean-scaled inaccuracy, the joint inaccuracy was scaled by the product of the male and female trait means. Ninety-five percent confidence intervals (95% CIs) were obtained from 1000 nonparametric bootstrap estimates of the joint and mean-scaled inaccuracies, respectively.