Discussion of Exercise 1

- General strategy for programming exercises
- The standard normal distribution

Strategy for programming

- Start by performing the relevant operation once, then add indices etc. (i.e. start writing functions from the inside)
- Stepwise strategy:
  - 1. Verbal code
  - 2. Pseudocode
  - 3. Code

Strategy for programming

Verbal code:

- Generate random normal variable with a range of CV
- Compute CV
- Compute SD of logs
- Save results
- Plot results

Pseudocode:

- rnorm()
- function(x, mean, sd)
- sd()
- vals[i,] =
- plot(cvvals, sdvals)

Code:

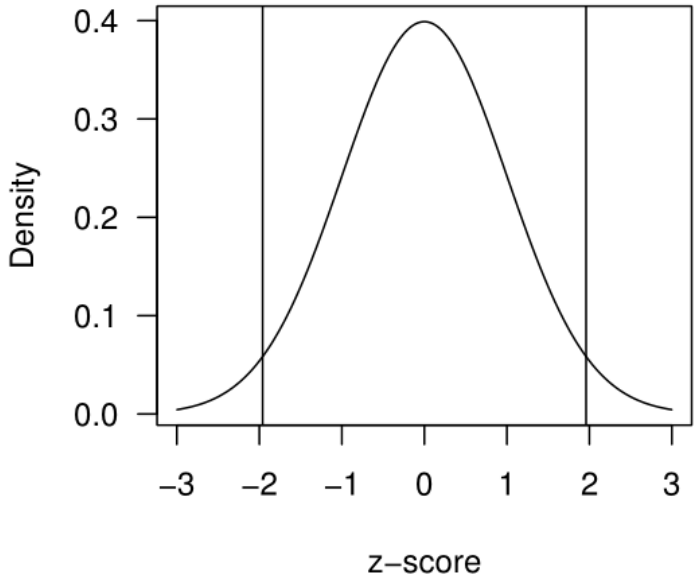
```
set.seed(1)
out = matrix(NA, nrow=200, ncol=2)
sdvals = runif(200, 2, 5)

for(i in 1:200){
x = rnorm(200, 20, sdvals[i])
cv = sd(x)/mean(x)
sd_log = sd(log(x))
out[i,1] = cv
out[i,2] = sd_log
}

plot(out[,1], out[,2],
     xlab="CV(x)",
     ylab="SD(log[x])", las=1)
lines(0:1, 0:1)
```

The standard normal distribution

- Normal distribution with mean of zero and standard deviation of one
- The 2.5 and 97.5 quantiles (percentiles) fall at ~1.96 SDs from the mean
- 95% CI = ±1.96*SE*



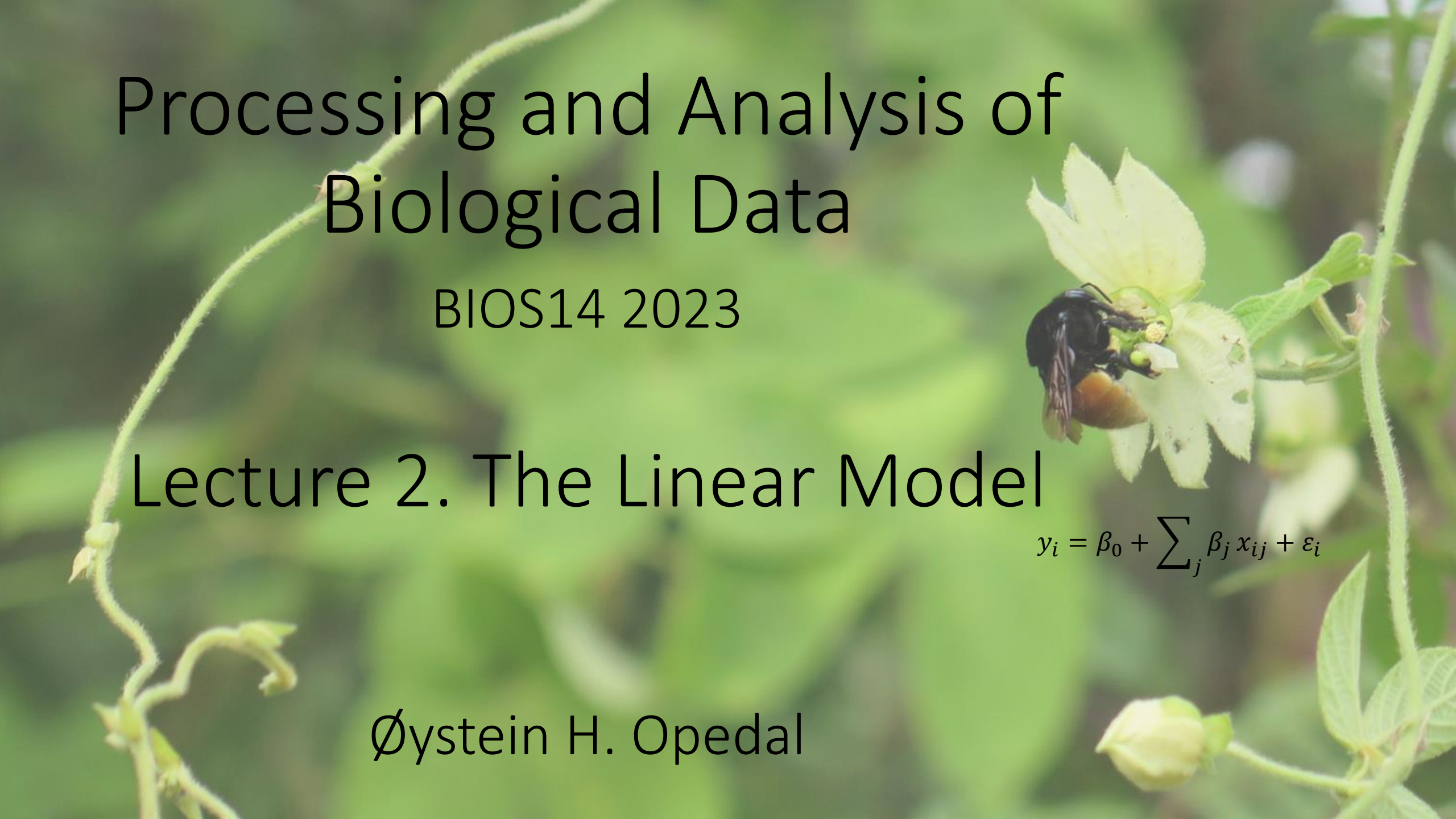
The standard normal distribution

- Normal distribution with mean of zero and standard deviation of one
- The 2.5 and 97.5 quantiles (percentiles) fall at ~1.96 SDs from the mean
- 95% CI = ±1.96*SE*
- Quantiles in table, or `qnorm` in `R`

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Processing and Analysis of Biological Data

BIOS14 2023



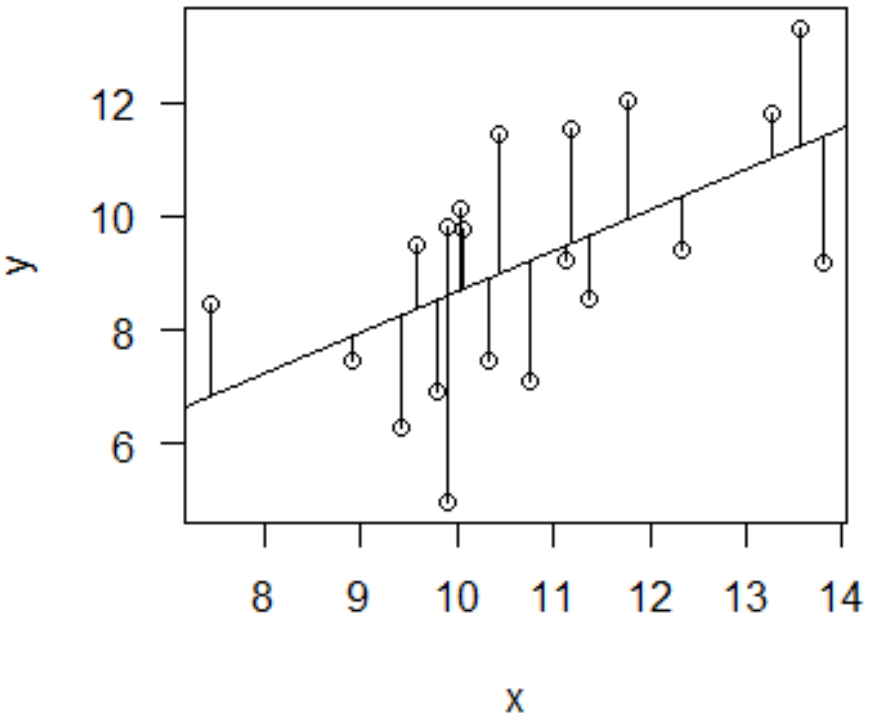
Lecture 2. The Linear Model

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

Øystein H. Opedal

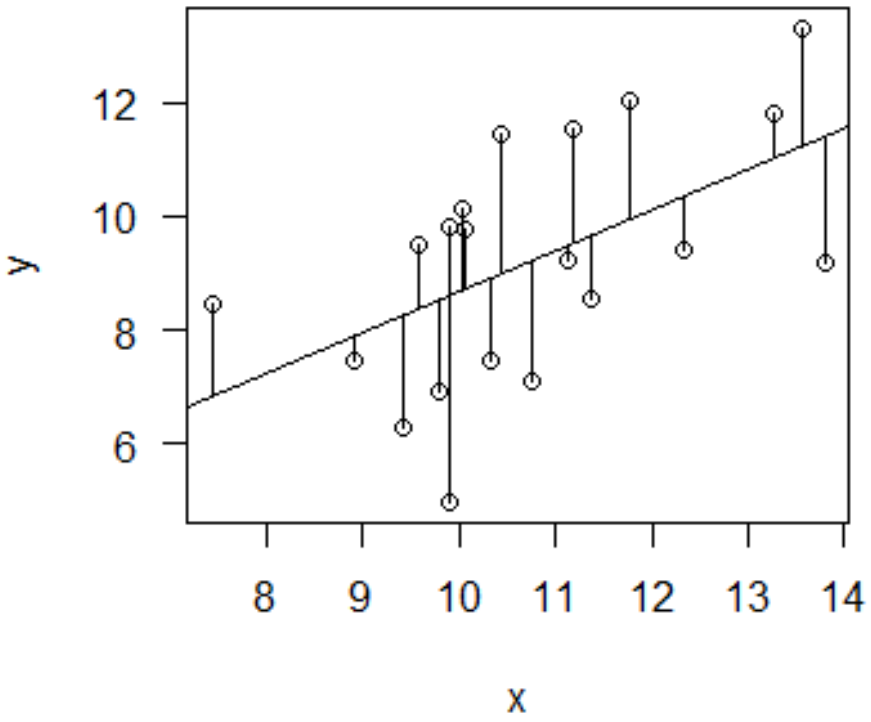
The linear model

- 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



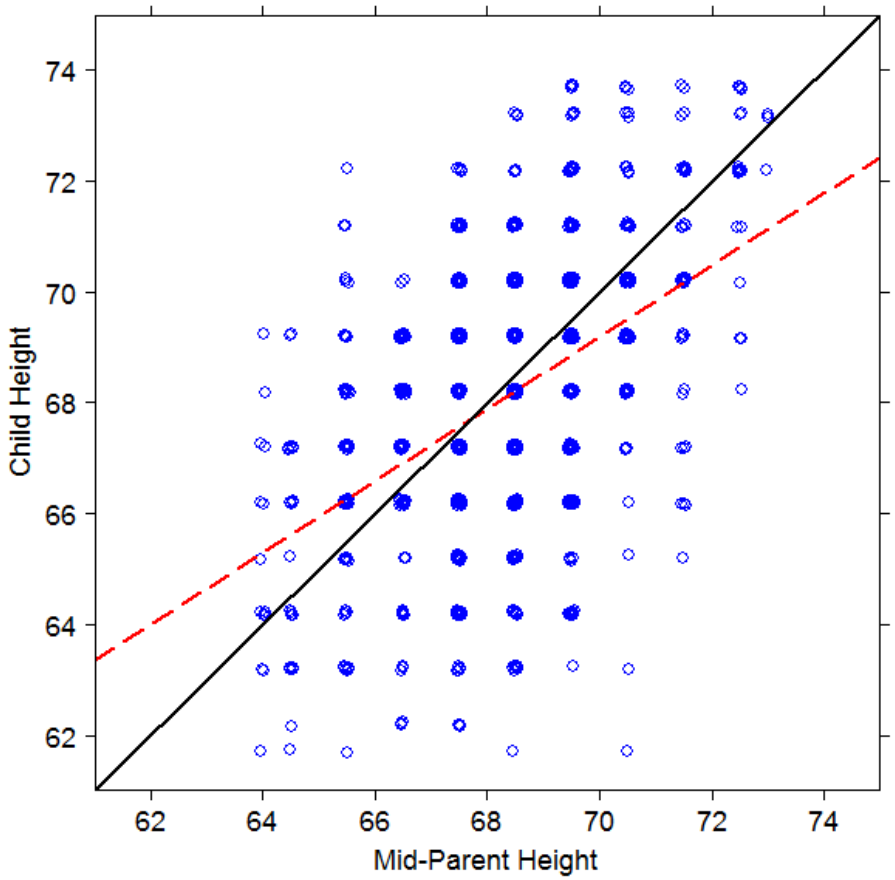
Linear regression

- When the predictor variable is continuous, a linear model represents the regression of *y* on *x*.
- Linear regression is a common tool in biology, used to analyse the (potentially causal) effect of *x* on *y*.
- Also used in an “applied” way, to predict unknown values of y (for example estimating bird body size from tarsus length, which is easy to measure in the field).
- The basis of more flexible models such as GLMs and mixed models.



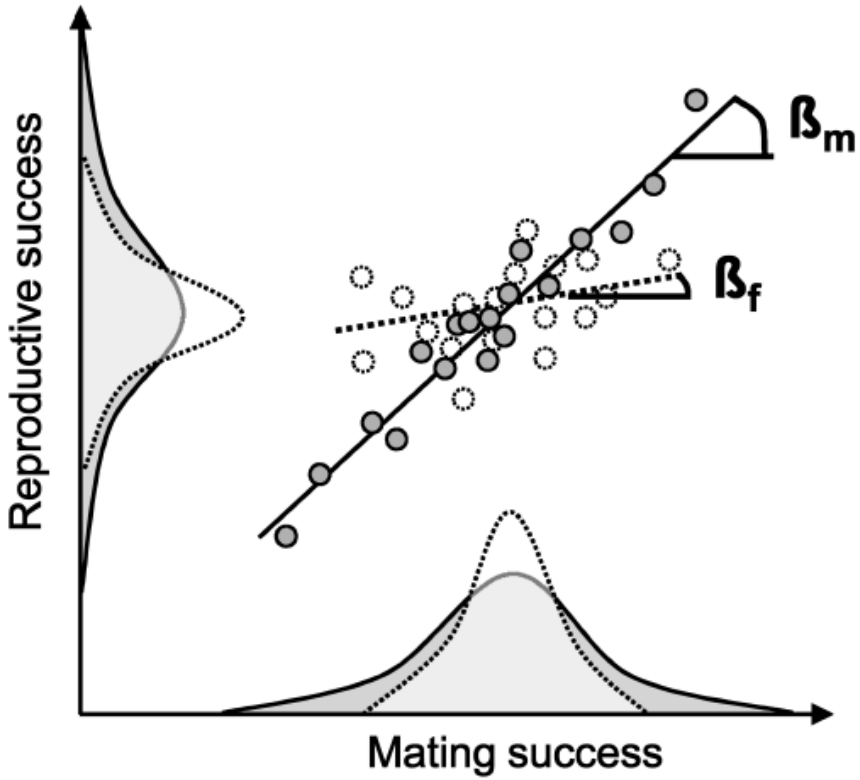
The first regression

- Francis Galton developed least-square regression as part of his work on the heredity of human height
- In parent-offspring regression, the slope gives an estimate of the heritability of a phenotypic trait



Measuring sexual selection

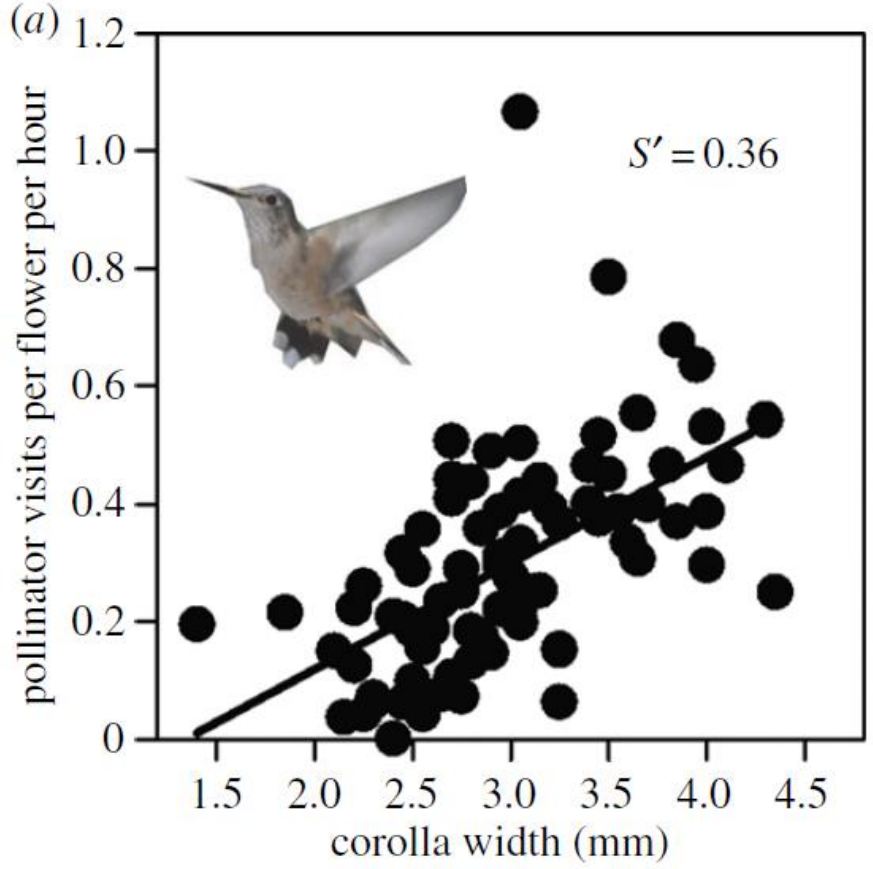
- Linear regression can be used to estimate Bateman gradients, describing how reproductive success scale with mating success
- The sex with the steeper Bateman gradient will usually have greater variance in fitness and thus greater opportunity for natural selection



Anthes et al. 2010 Am Nat

Measuring natural selection

- Linear regression can be used to estimate the strength and shape of natural selection on phenotypic traits
- If the response variable is relative fitness, and the predictor is scaled to unit variance (SD=1), the slope gives the *selection intensity*



Campbell & Powers 2015 PRS

Misuse of regression

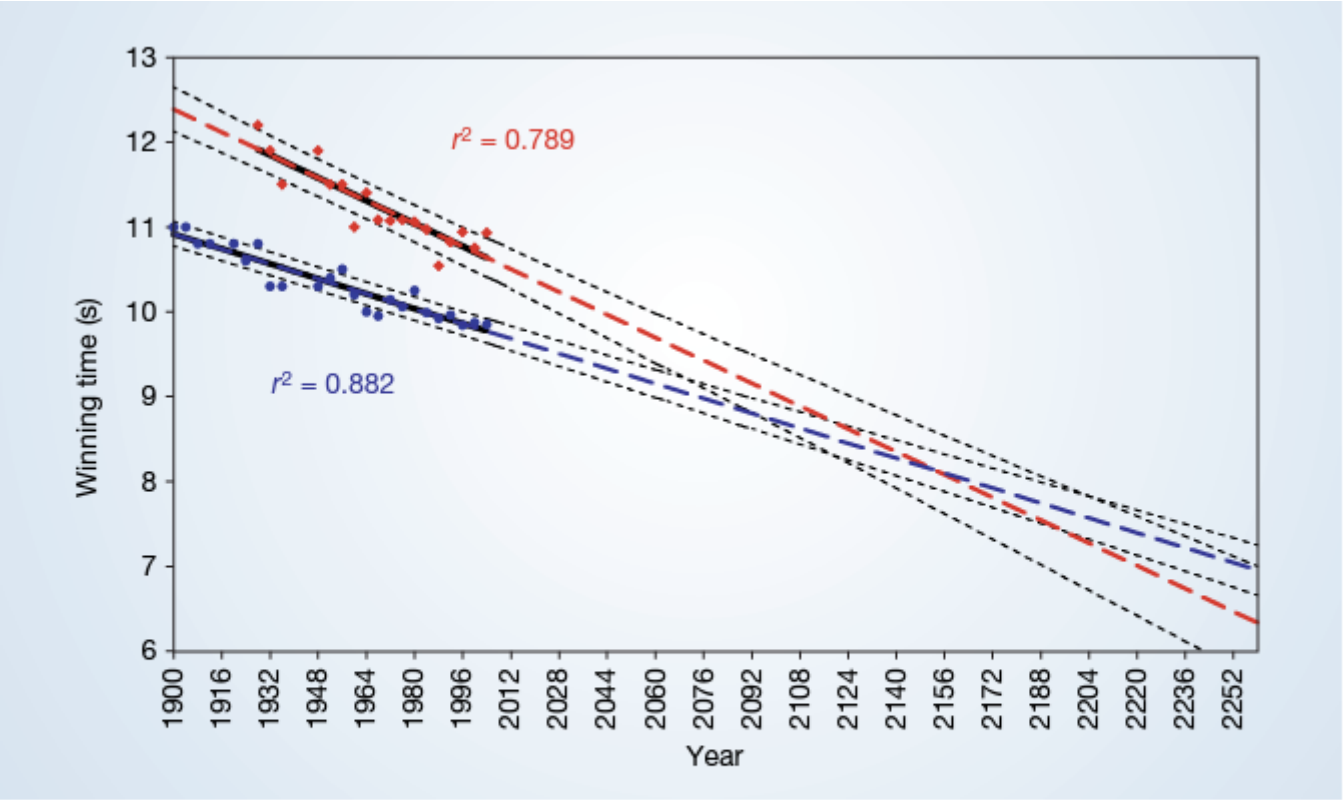


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Linear model in R

- The model summary gives some quantiles of the residual distribution, the parameter estimates, and the results of some hypothesis tests

```
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.45122 -0.68319  0.02913  0.69861  2.88937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.40114    0.35186
## x            0.43330    0.03538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9912 on 198 degrees of freedom
## Multiple R-squared:  0.4311, Adjusted R-squared:  0.4282
```

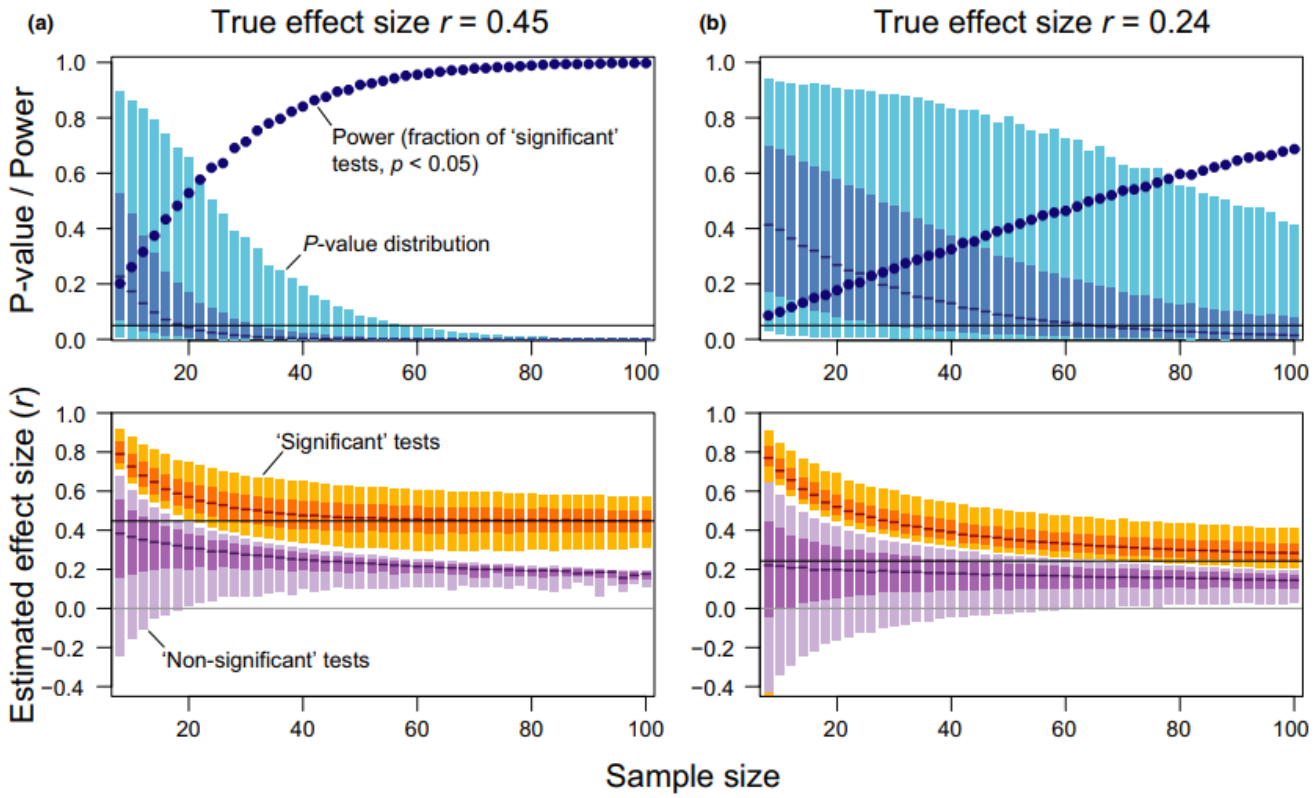
Hypothesis testing in linear regression

- The p-value depends on the strength of the effect, the variation in the data, and the sample size
- All else being equal, the p-value will decrease with increasing sample size

$$t = \frac{\hat{\beta}}{\frac{\sigma(\hat{\beta})}{\sqrt{n}}}$$

The problem with p-values

- For moderate sample sizes, even important effects can go statistically undetected
- Detectable effects at small sample sizes will often be upwardly biased



The problem with p-values

- For moderate sample sizes, even important effects can go statistically undetected
- Detectable effects at small sample sizes will often be upwardly biased

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

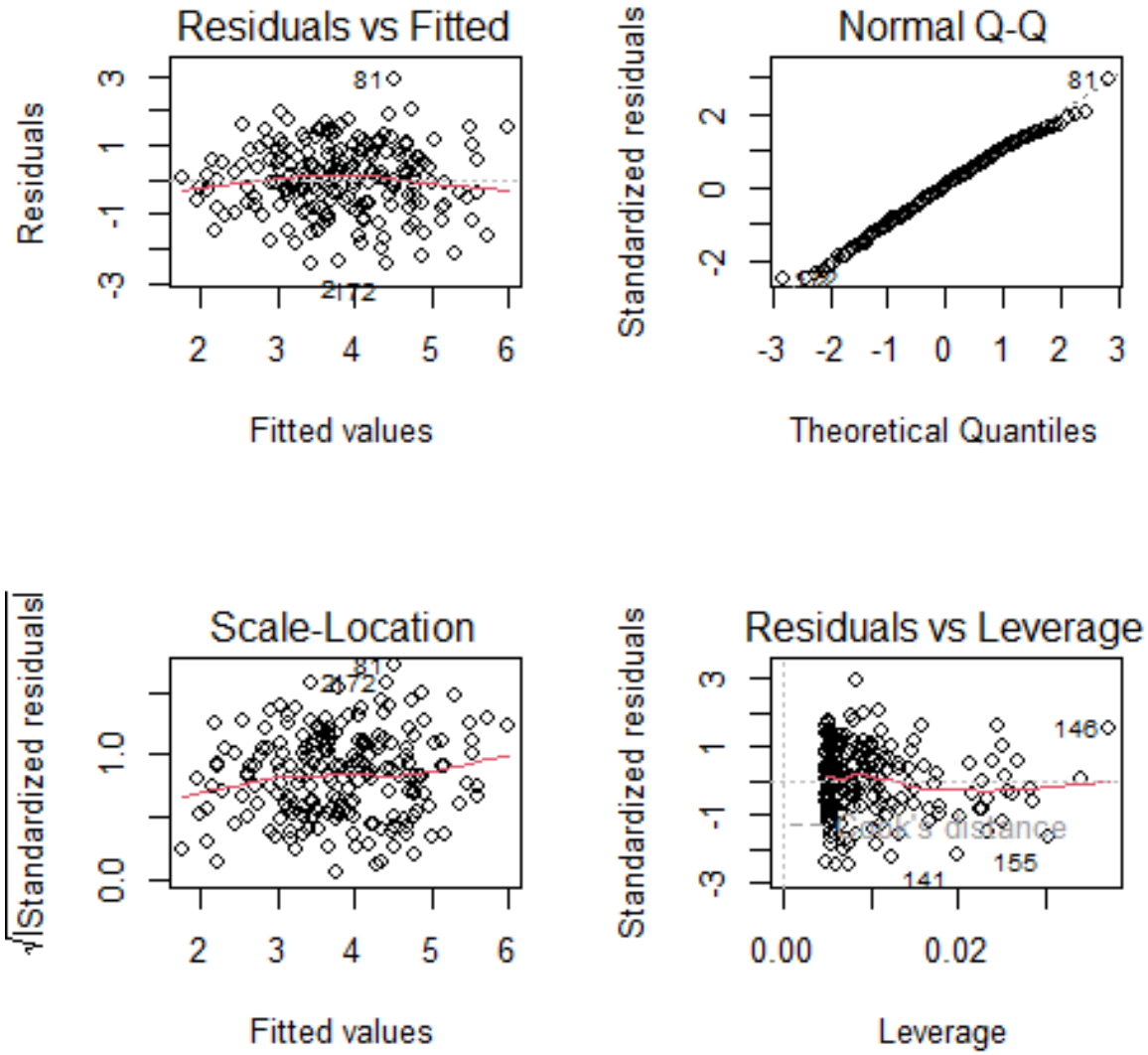
The problem with p-values

- Beyond these statistical issues, the biological issue is that focusing on p-values stops people from thinking.
- Estimation and quantification are the key aspects of modelling in biology.

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Model diagnostics

- It´s important to know the assumptions of our models, like normally distributed residuals for a linear regression
- However, small deviations from normality is generally not a problem, and testing statistically for such deviations have all the usual problems of hypothesis testing



Variation explained

- The model r^2 (coefficient of determination) gives the proportion of variance explained by the model
- Can be computed as the variance in the model-predicted values over the total variance in the response variable

$$V(\hat{y}) = V(X\beta)$$

$$r^2 = V(\hat{y})/V(y)$$

General intro to exercises

- You can work with any dataset you may have!
- We will provide one or more example 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 reports, mid-term exercise and the exam will be similar
- During the morning after each exercise session, we will show some examples of how the data could be analysed and the results presented
- Written feedback on the exercise reports + mid-term, and we 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

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

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*)

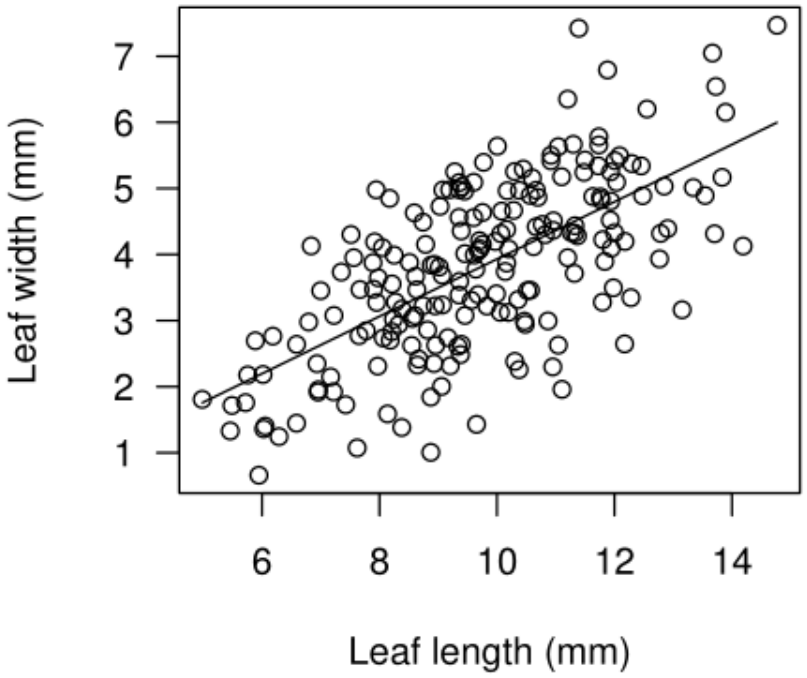
Overall, blossom development tended to be more rapid under dry conditions. In the wet treatment, the total receptive period of the blossoms (from the first day of opening to the abscision of the male cymule) lasted for 6.87 ± 0.09 d in the two large-glanded populations and 5.80 ± 0.10 d in the two small-glanded 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

- Labels and units!
- Keep regression lines within the data range
- If several symbols/colors: legend inside the graph
- Usually quite square panels look the best (not very rectangular)



There are also other packages for graphics

ggPlot

base R

Making nice tables

- No vertical lines
- All parameters explained in legend

Table 5

Inaccuracy Statistics under Wet and Dry Experimental Conditions

Taxon, population, treatment	Bias²	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. Bias² 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.