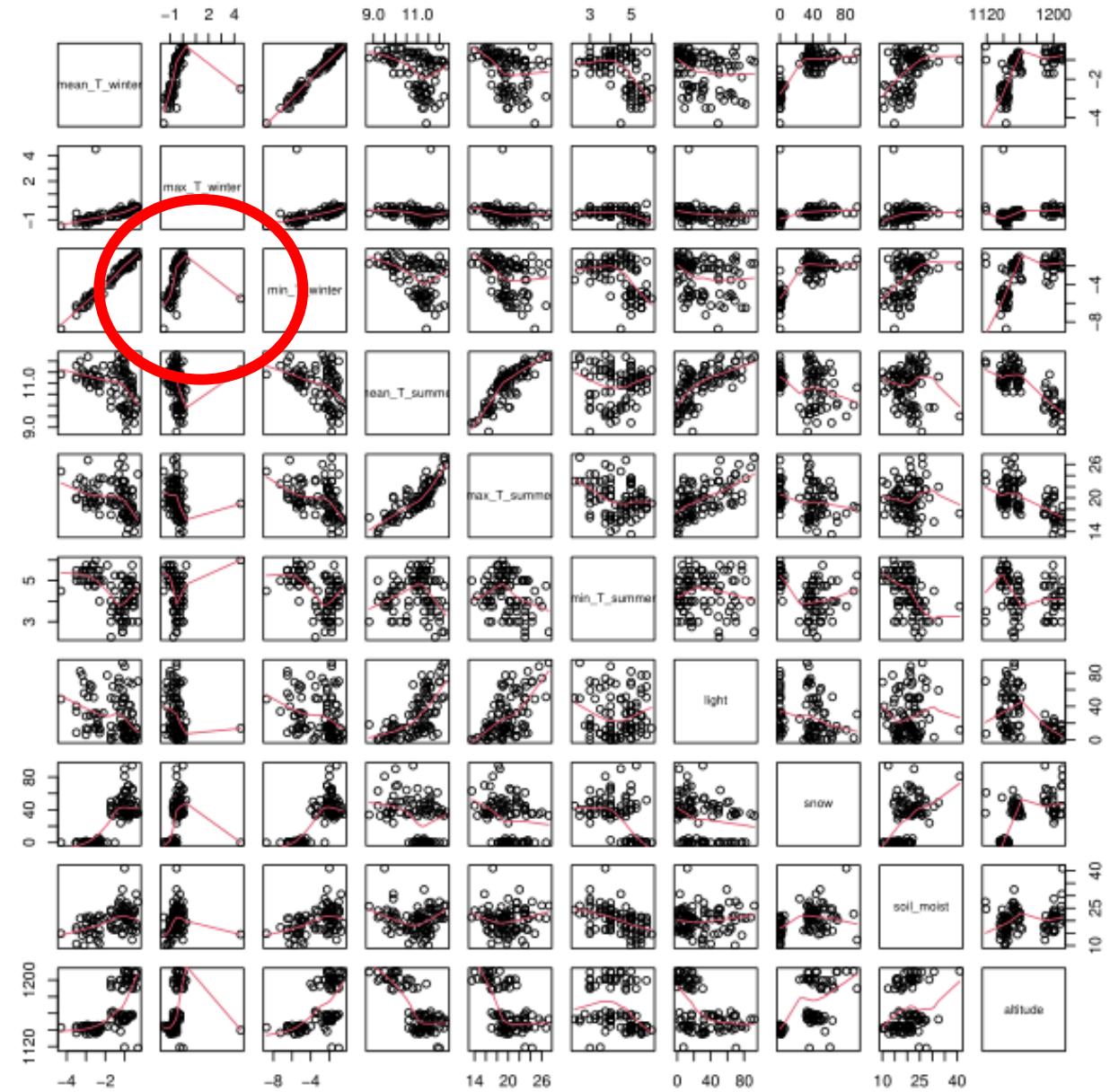
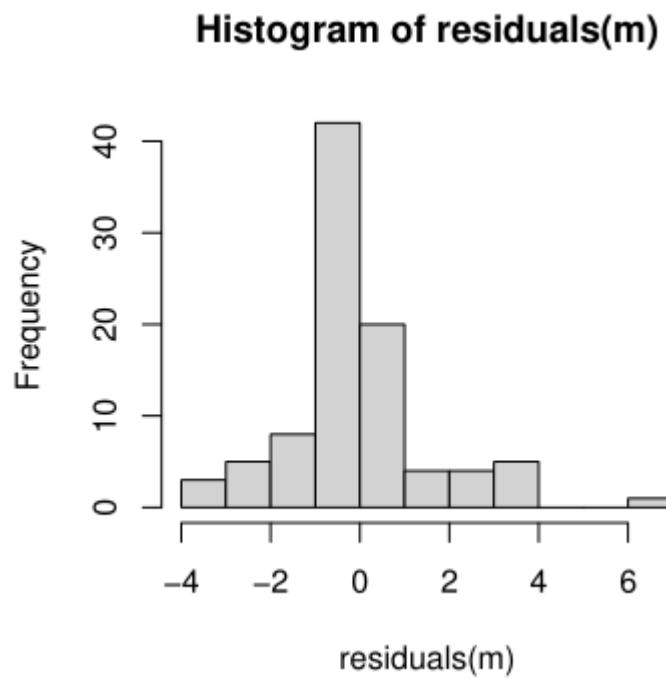


Discussion of exercise 4

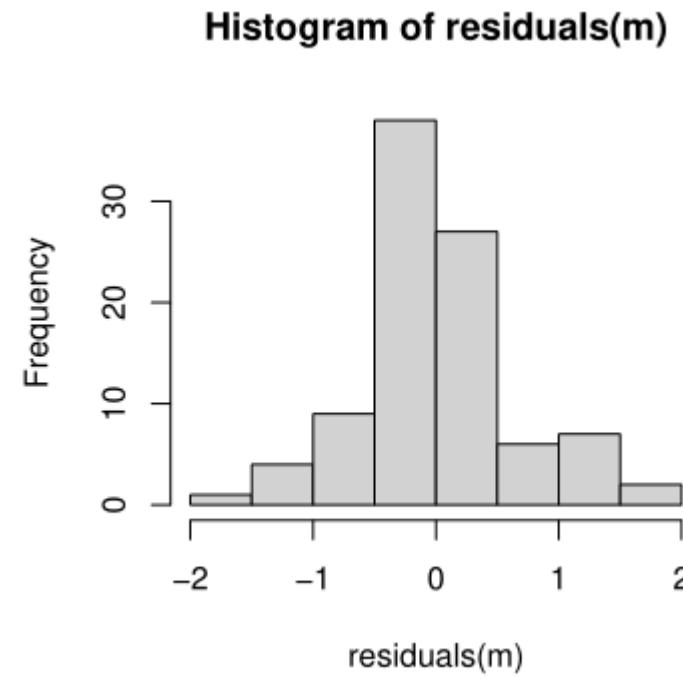
- The importance of data exploration
- Backward selection vs. biologically justified models



Untransformed



Sqrt-transformed



Backward selection

```
##  
## Call:  
## lm(formula = sqrt(Thalictrum.alpinum) ~ mean_T_winter + max_T_winter +  
##       min_T_winter + mean_T_summer + max_T_summer + min_T_summer +  
##       light + snow + altitude, data = plants, na.action = na.exclude)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.64774 -0.32645 -0.09373  0.32101  1.60726  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  2.262390  8.600678  0.263  0.79316  
## mean_T_winter 0.534162  0.561831  0.951  0.34446  
## max_T_winter -0.031941  0.424243 -0.075  0.94016  
## min_T_winter -0.111548  0.238945 -0.467  0.64182  
## mean_T_summer 0.766547  0.275480  2.783  0.00666 **  
## max_T_summer -0.129458  0.065572 -1.974  0.05163 .  
## min_T_summer -0.179251  0.126460 -1.417  0.16005  
## light         0.007020  0.003649  1.924  0.05779 .  
## snow          0.012631  0.005140  2.457  0.01605 *  
## altitude     -0.005797  0.005905 -0.982  0.32903  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.6468 on 84 degrees of freedom  
##   (1 observation deleted due to missingness)  
## Multiple R-squared:  0.546, Adjusted R-squared:  0.4973  
## F-statistic: 11.22 on 9 and 84 DF,  p-value: 2.528e-11
```

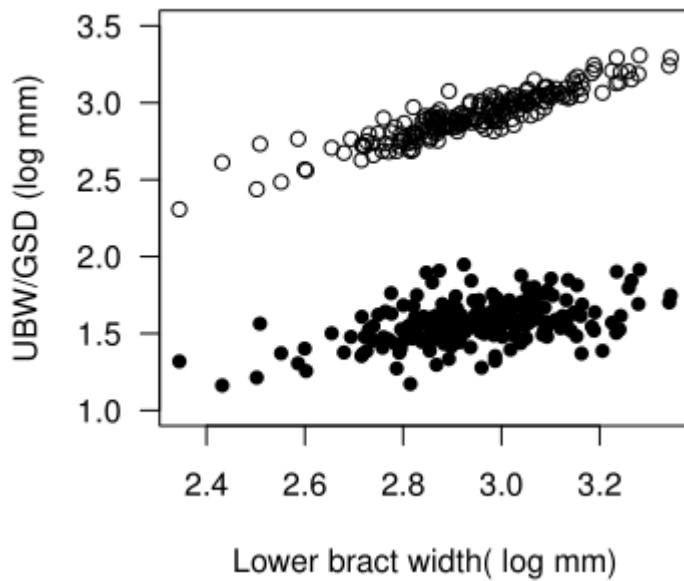
```
##  
## Call:  
## lm(formula = sqrt(Thalictrum.alpinum) ~ mean_T_winter + min_T_winter +  
##      mean_T_summer + max_T_summer + min_T_summer + light + snow +  
##      soil_moist + altitude, data = plants, na.action = na.exclude)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.62694 -0.32380 -0.07332  0.29264  1.62183  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.411671  8.576529  0.281  0.77927  
## mean_T_winter 0.569210  0.446838  1.274  0.20631  
## min_T_winter -0.143767  0.218489 -0.658  0.51238  
## mean_T_summer 0.768061  0.275298  2.790  0.00655 **  
## max_T_summer -0.132171  0.066057 -2.001  0.04871 *  
## min_T_summer -0.205011  0.126349 -1.623  0.10852  
## light         0.007255  0.003712  1.954  0.05408 .  
## snow          0.013162  0.005162  2.550  0.01263 *  
## soil_moist    -0.005695  0.016416 -0.347  0.72953  
## altitude      -0.005748  0.005857 -0.981  0.32929  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.6415 on 82 degrees of freedom  
##   (3 observations deleted due to missingness)  
## Multiple R-squared:  0.5455, Adjusted R-squared:  0.4956  
## F-statistic: 10.94 on 9 and 82 DF,  p-value: 5.347e-11
```

Minimal adequate model

```
##  
## Call:  
## lm(formula = sqrt(Thalictrum.alpinum) ~ mean_T_winter + mean_T_summer +  
##       max_T_summer + light + snow, data = plants, na.action = na.exclude)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -1.50452 -0.40421 -0.04351  0.27470  1.60612  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -6.437373  1.316222 -4.891 4.49e-06 ***  
## mean_T_winter  0.360464  0.121818  2.959  0.00396 **  
## mean_T_summer  0.794341  0.174225  4.559 1.65e-05 ***  
## max_T_summer -0.080261  0.045387 -1.768  0.08046 .  
## light         0.006737  0.003594  1.874  0.06419 .  
## snow          0.012148  0.004627  2.625  0.01021 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.6445 on 88 degrees of freedom  
##   (1 observation deleted due to missingness)  
## Multiple R-squared:  0.5277, Adjusted R-squared:  0.5008  
## F-statistic: 19.66 on 5 and 88 DF,  p-value: 4.199e-13
```

- Note that r^2 is very high (for ecological analyses), and dropped just a tiny bit from the full (saturated) model.

ANCOVA/Floral integration/allometry



```
mUBW = lm(log(UBW)-log(LBW), data=blossoms)
mGSD = lm(log(GSD)-log(LBW), data=blossoms)
summary(mUBW)$coef
```

```
##             Estimate Std. Error   t value   Pr(>|t|) 
## (Intercept) 0.3193964 0.07851616 4.067907 6.822632e-05
## log(LBW)    0.8819832 0.02662027 33.132018 3.727171e-83
```

```
summary(mGSD)$coef
```

```
##             Estimate Std. Error   t value   Pr(>|t|) 
## (Intercept) 0.3782970 0.15788784 2.395986 1.749721e-02
## log(LBW)    0.4047488 0.05353059 7.561073 1.416389e-12
```

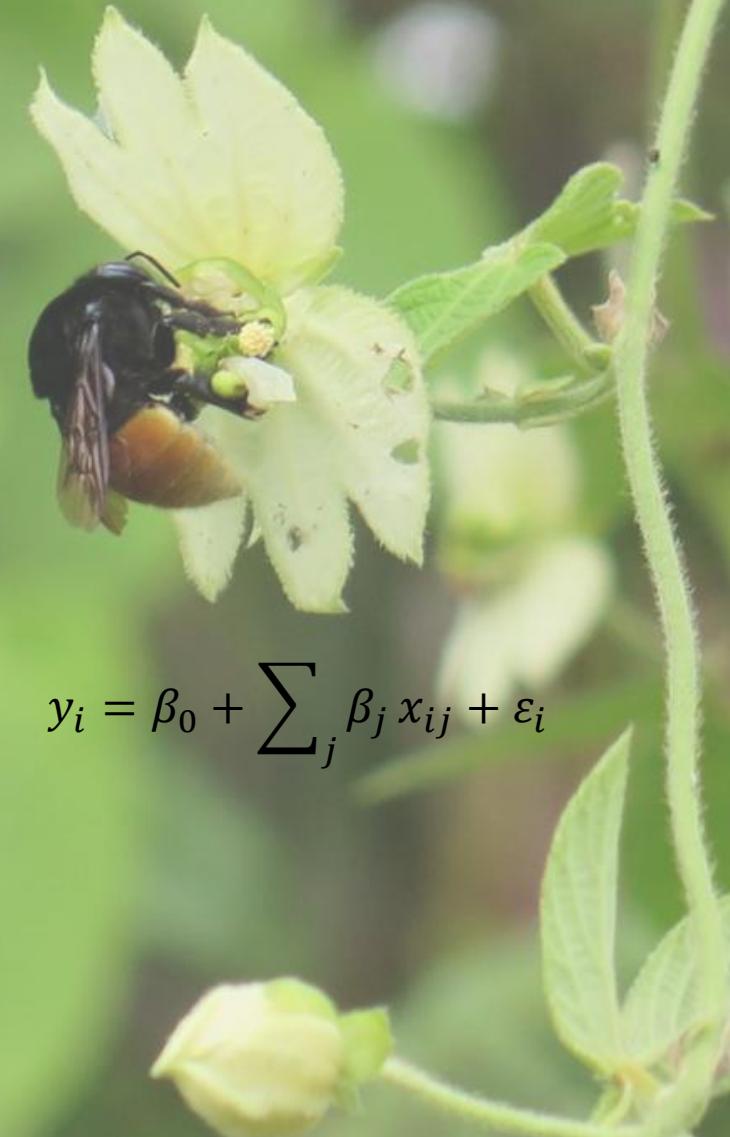
Processing and Analysis of Biological Data

BIOS15 2025

Lecture 5. GLM I: Logistic regression

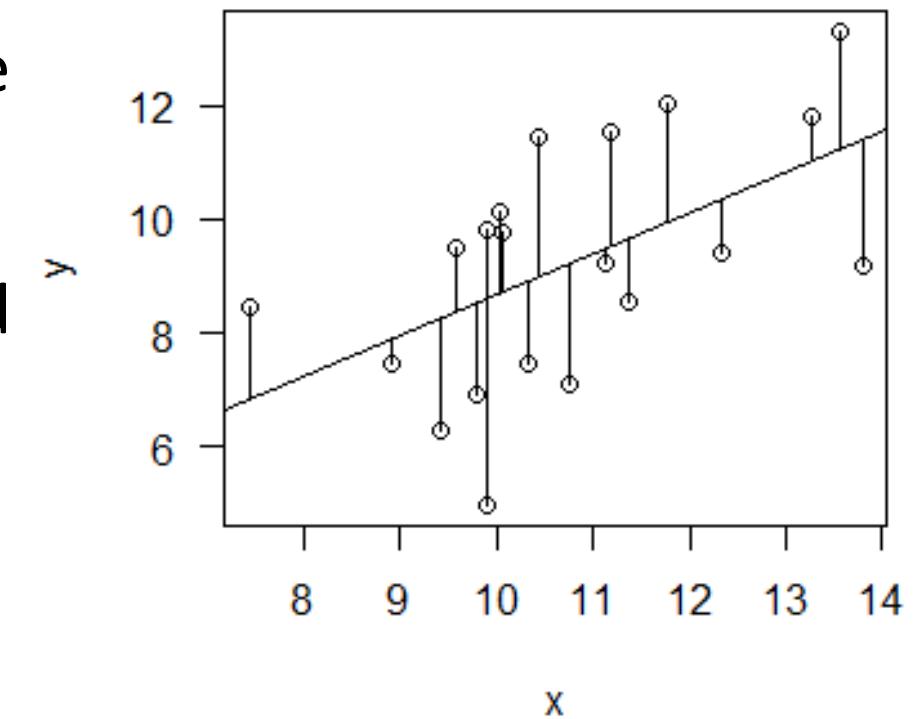
Øystein H. Opedal

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



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



Generalized linear models

- Generalized linear models extend the linear model by relaxing the assumption of normally distributed residuals
- The model connects a response variable to the familiar linear predictor (η) through a **link function** (g)
- The link functions are specific to different **error distributions**, the most common are **Binomial** and **Poisson** errors

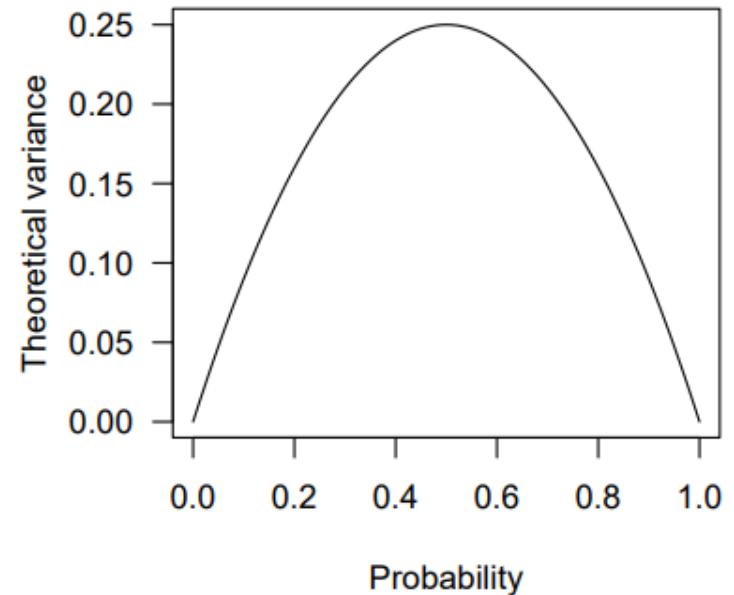
$$\eta = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i, \varepsilon \sim N(0, \sigma^2)$$

$$y = g^{-1}(\eta)$$

Analysis of binary (0/1) data

- Binary data can be analysed with a binomial error distribution
- The binomial distribution summarizes a set of Bernouli trials yielding 0/1 responses, where 1 is “success” and 0 is “failure”.
- The distribution is defined by two parameters, the number of trials n and the probability p .

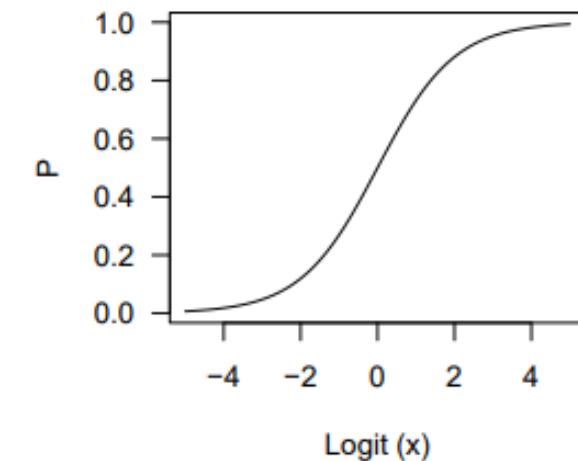
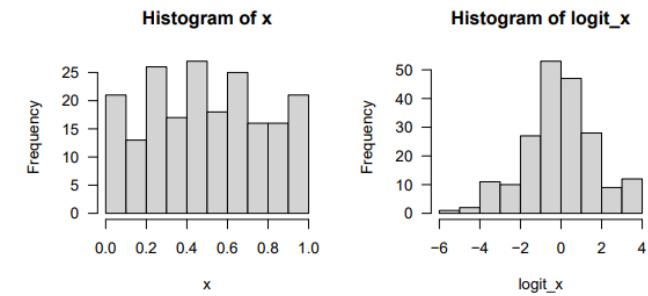
$$\sigma^2 = np(1 - p)$$



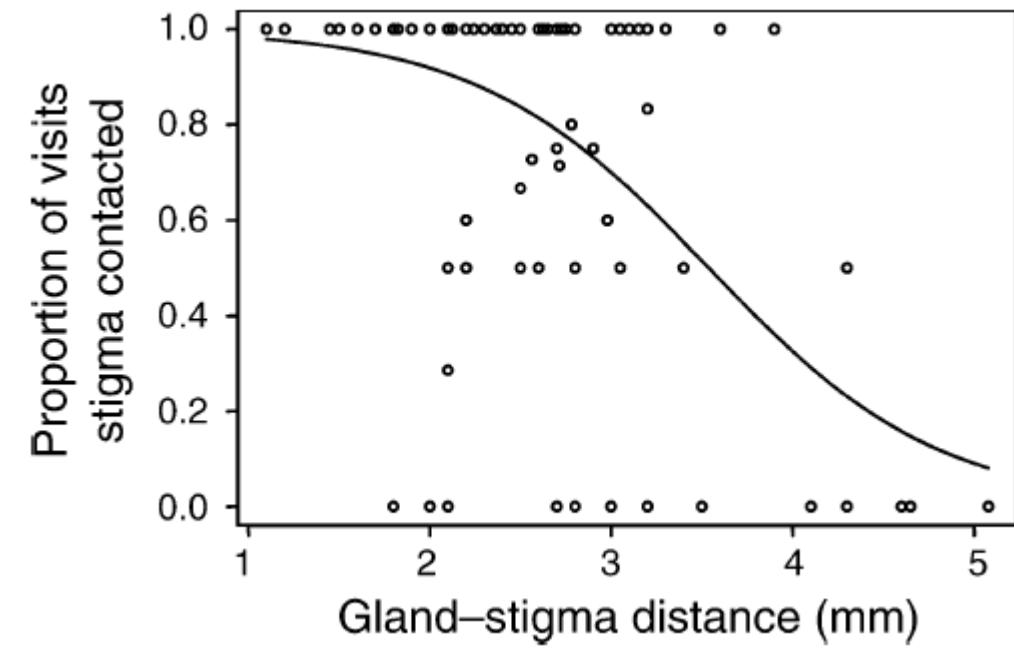
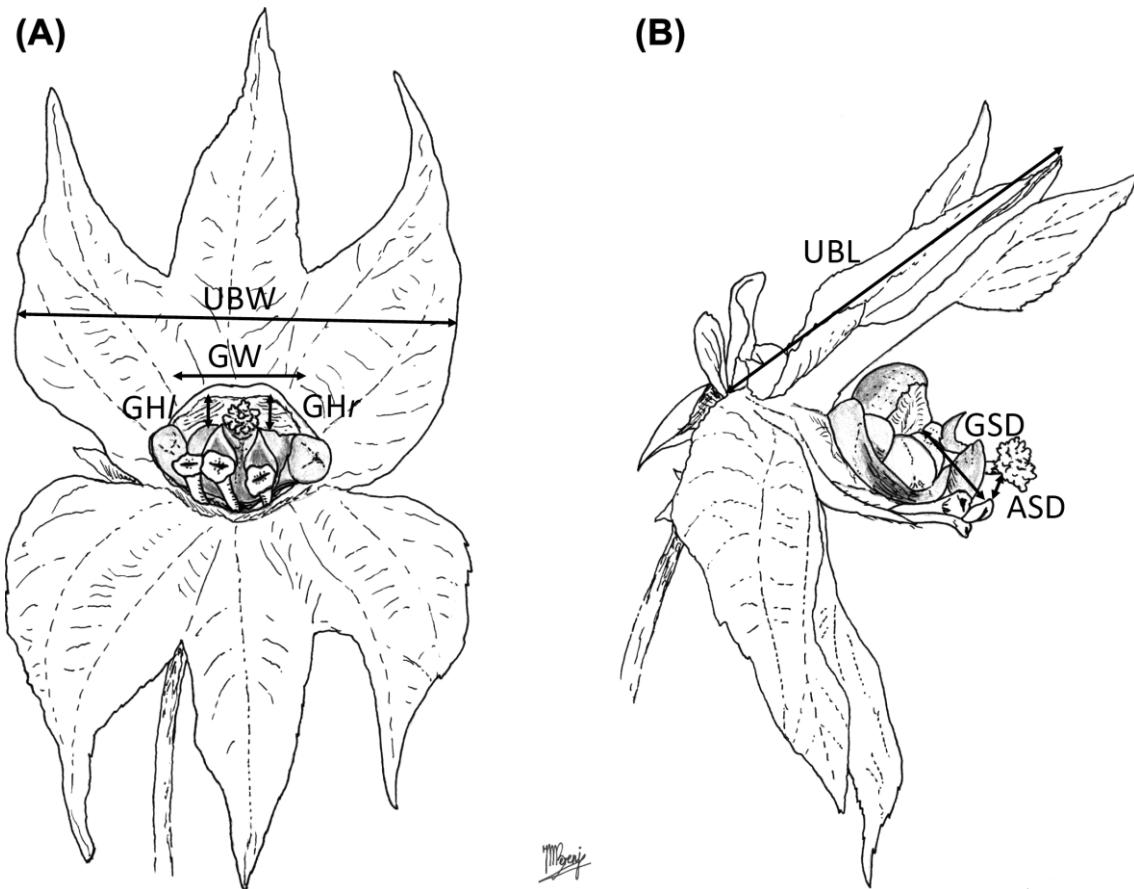
Logistic regression

- A GLM with binomial errors is called a logistic regression
- The most common link function is the logit (log odds) link
- The data can be binary (0/1), sets of binary variables (number of successes and failures), or proportions

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

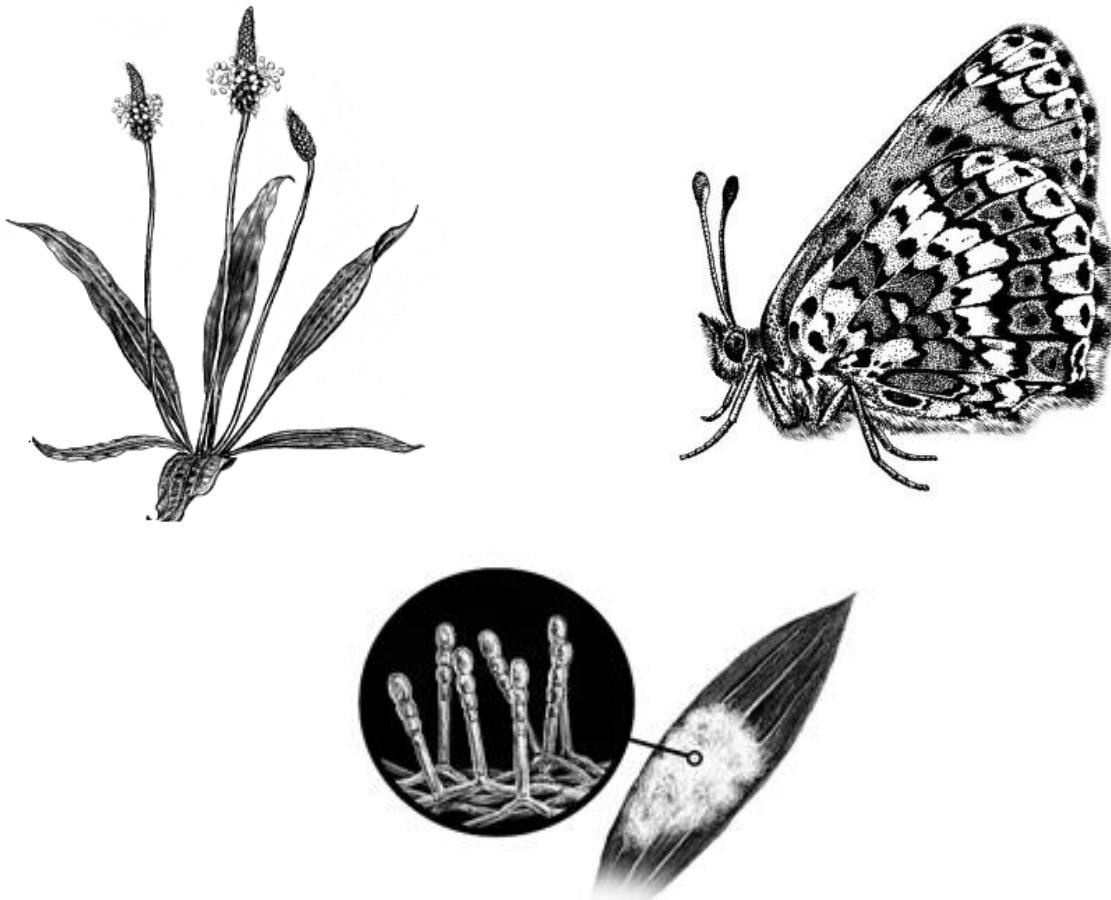


Example: functional pollination studies



Example: plant-animal interactions

- The Glanville Fritillary butterfly (*Melitaea cinxia*) and the powdery mildew *Podosphaera plantaginis* share *Plantago lanceolata* as a host plant
- A tripartite interaction between a plant, a herbivore, and a plant pathogen



A tripartite interaction between a plant, a herbivore, and a plant pathogen

- The Glanville Fritillary butterfly (*Melitaea cinxia*) and the powdery mildew *Podosphaera plantaginis* share *Plantago lanceolata* as a host plant
- Butterfly survival tends to be lower on mildew-infected host plants

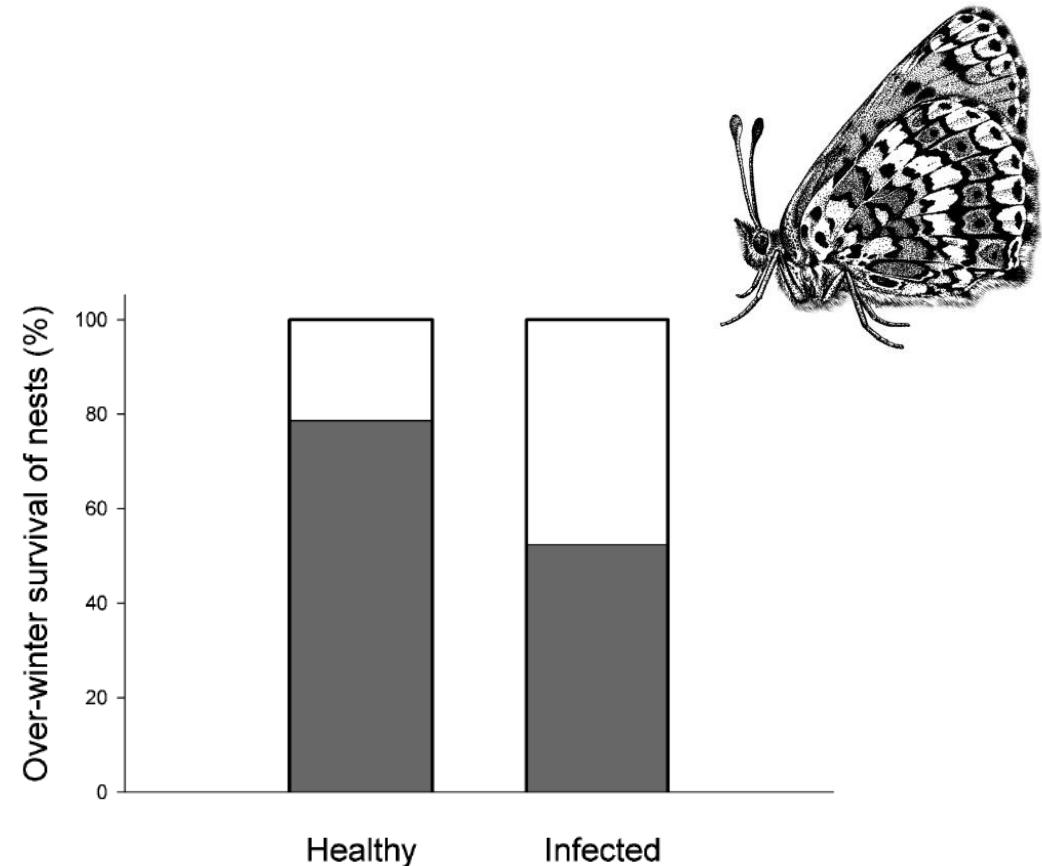
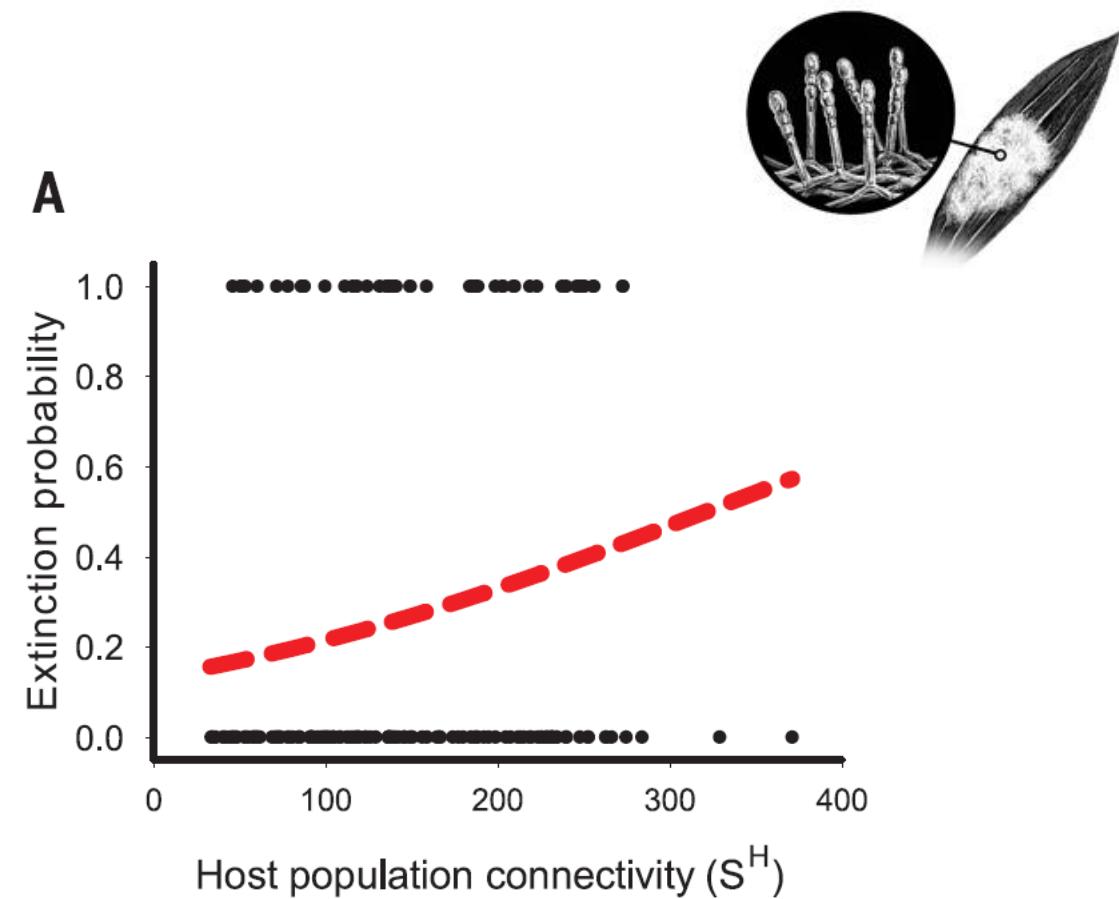


Fig. 4. Over-winter survival of *M. cinxia* larval groups was 26% higher in non-infected than in infected host populations.

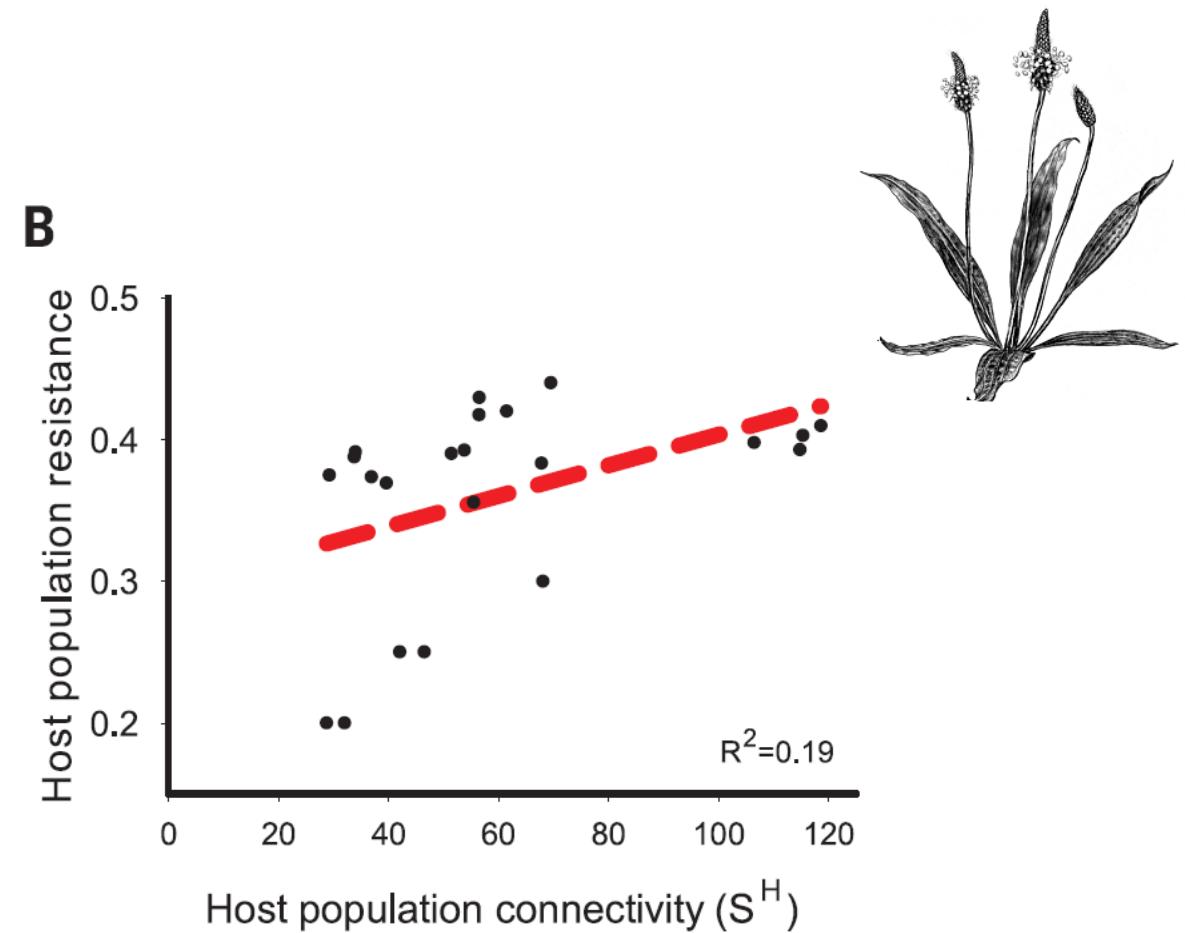
Host plants can evolve pathogen resistance

- The mildew is more likely to go locally extinct when infecting well-connected *Plantago* populations



Host plants can evolve pathogen resistance

- The mildew is more likely to go locally extinct when infecting well-connected *Plantago* populations
- These host populations have evolved greater resistance towards the pathogen



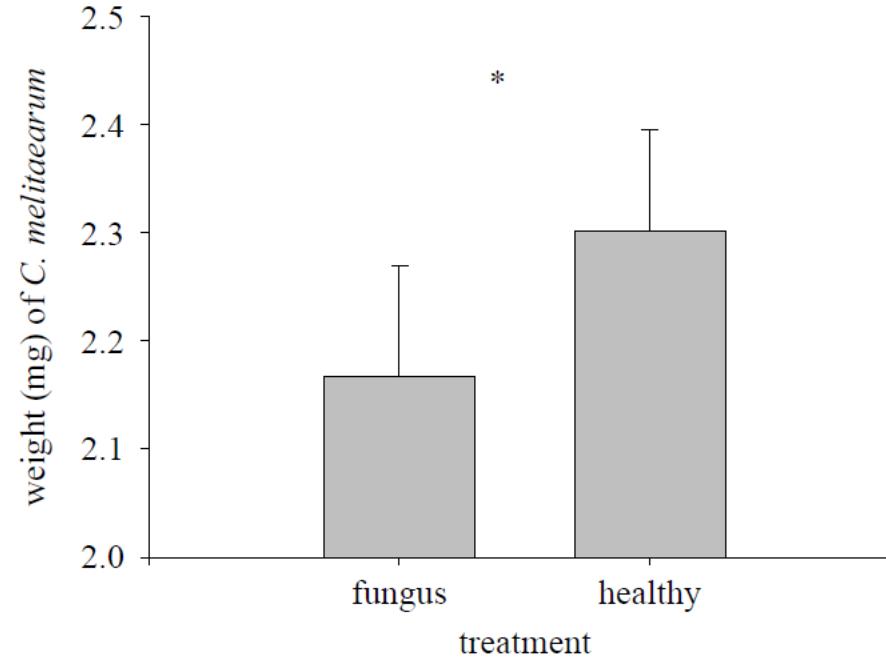
A tritrophic interaction between a plant, a herbivore, and a parasitoid

- The wasp *Cotesia melitaearum* is a parasitoid of the Glanville Fritillary



A tritrophic interaction between a plant, a herbivore, and a parasitoid – mediated by a pathogen

- The wasp *Cotesia melitaearum* is a parasitoid of the Glanville Fritillary
- The wasp grows better when host larvae feed on healthy host plants

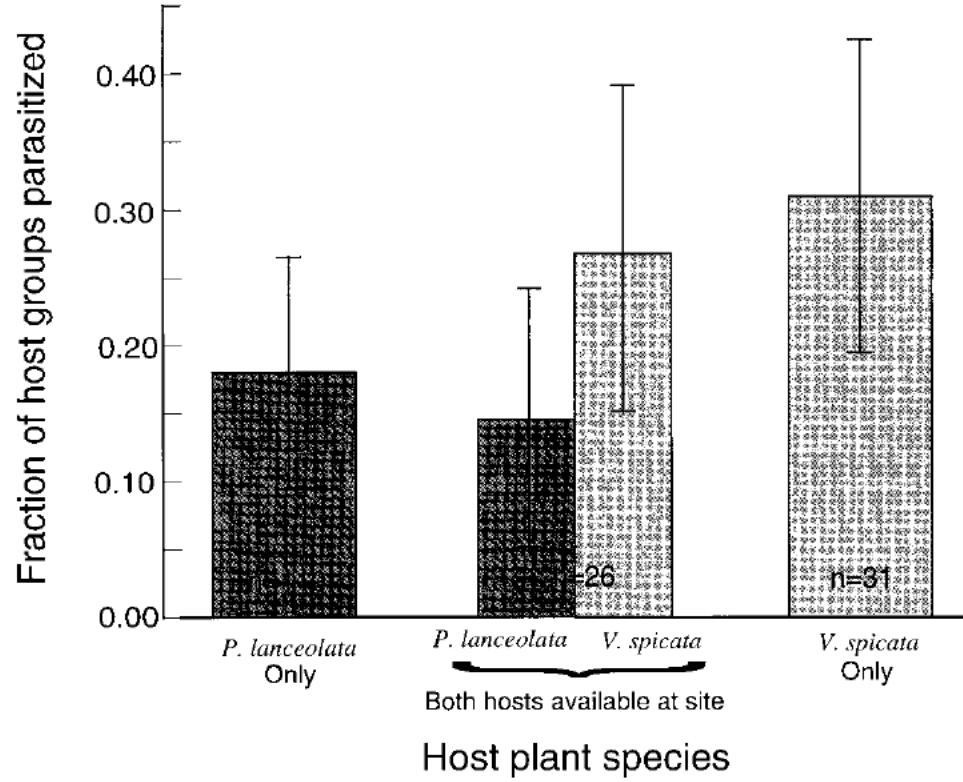


Butterfly host plant choices

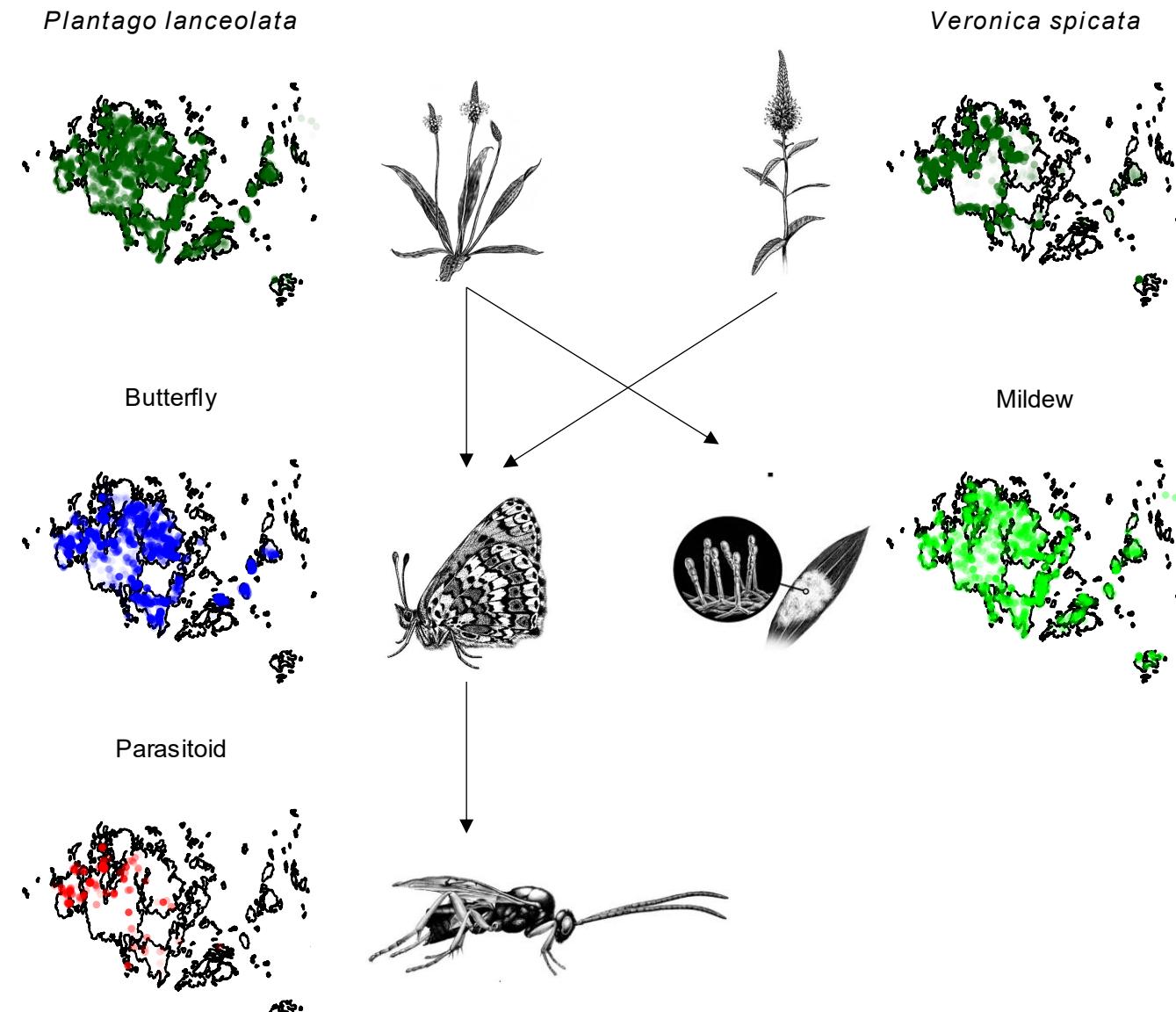
- The Glanville Fritillary uses two different host plants in the Åland Islands: *Plantago lanceolata* and *Veronica spicata*



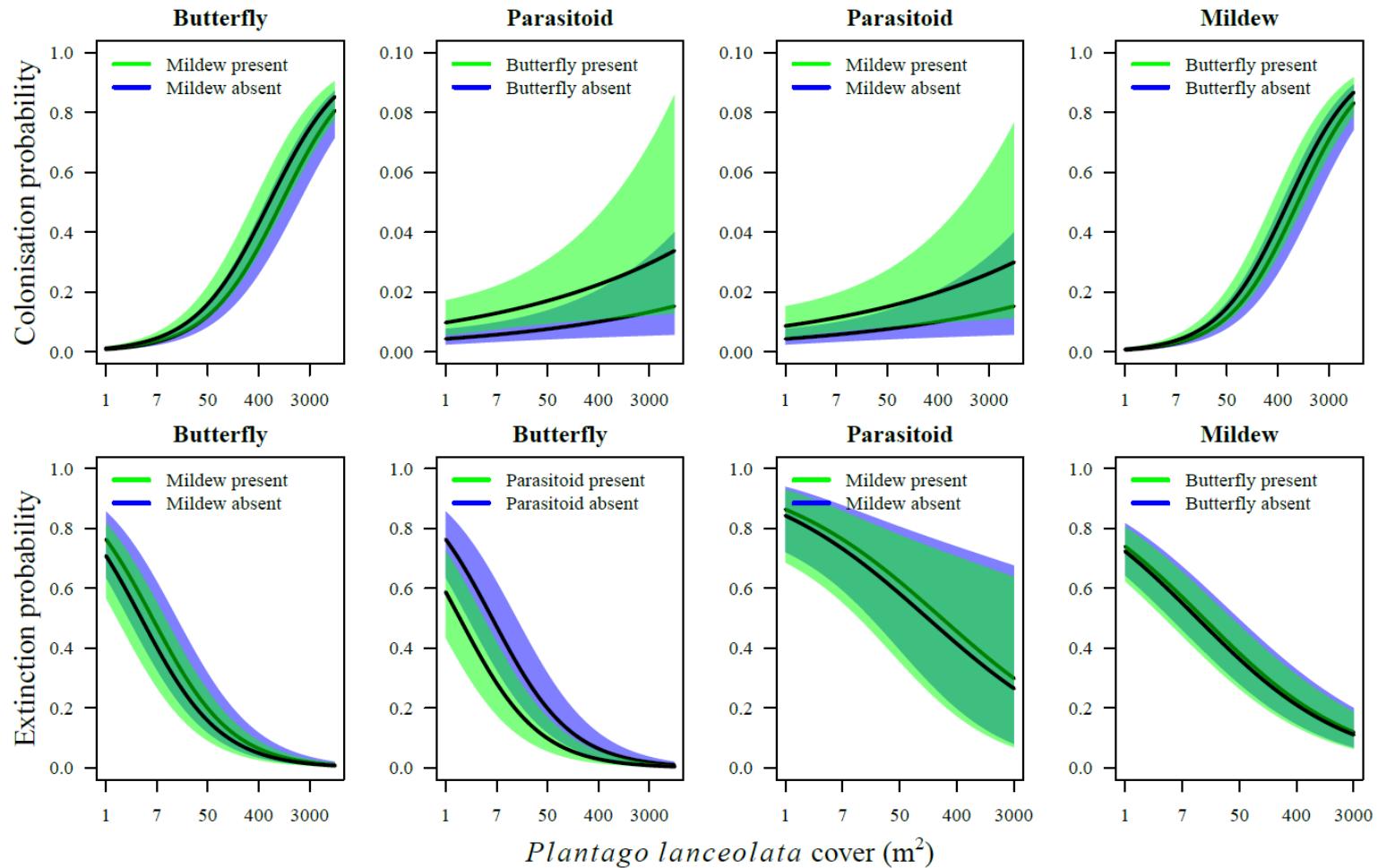
Butterfly host plant choices have consequences



Does this affect metacommunity dynamics?



Weak and unexpected effects of the presence of interacting species within patches



Logistic regression model in R

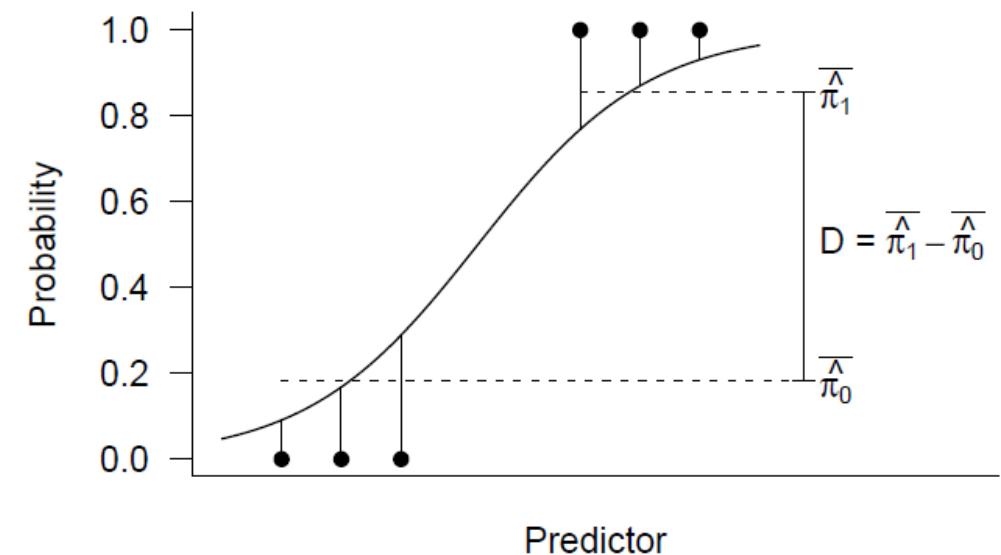
- The parameter estimates from a GLM are on the link scale, i.e. they describe in this case the change in the log odds of y per unit change in x
- The deviance measures the deviation of the model from a “perfect” model
- The normal r^2 is not valid, though there are options

```
##  
## Call:  
## glm(formula = y ~ x, family = binomial(link = "logit"))  
##  
## Deviance Residuals:  
##      Min       1Q   Median       3Q      Max  
## -2.0156  -1.1761   0.6573   0.7813   1.2629  
##  
## Coefficients:  
##                 Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.60084   0.53714  -1.119   0.2633  
## x           0.17488   0.05543   3.155   0.0016 **  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 227.10 on 199 degrees of freedom  
## Residual deviance: 216.37 on 198 degrees of freedom  
## AIC: 220.37  
##  
## Number of Fisher Scoring iterations: 4
```

The r^2 in logistic regression

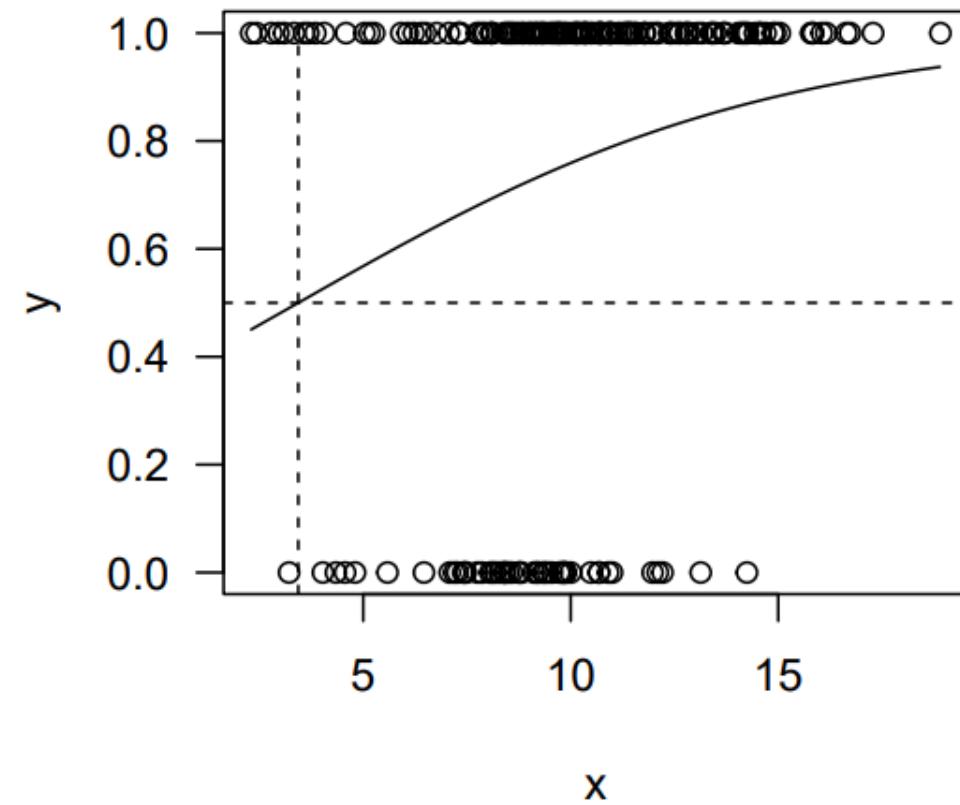
- We can quantify model fit through various “Pseudo r^2 ” metrics (see lecture notes)
- Another important measure is the coefficient of discrimination D, or Tjur’s r^2 . This measures the difference in the model-predicted probabilities between observed 1’s and observed 0’s.

$$D = \hat{\pi}_1 - \hat{\pi}_0,$$



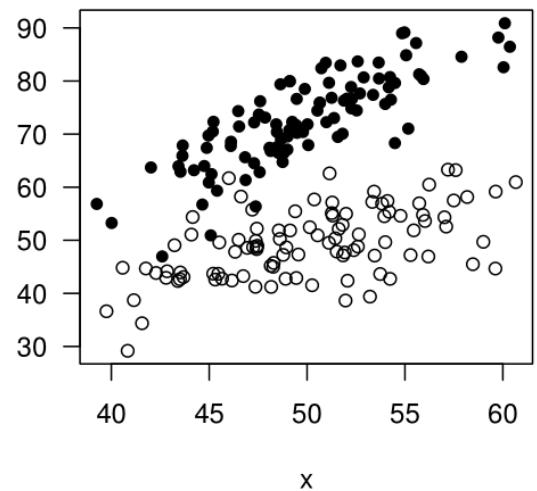
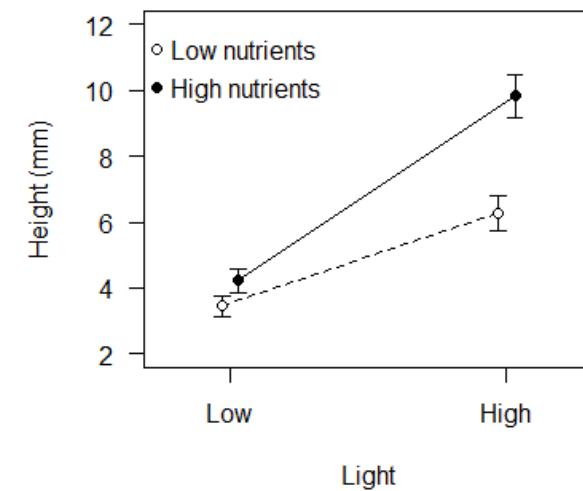
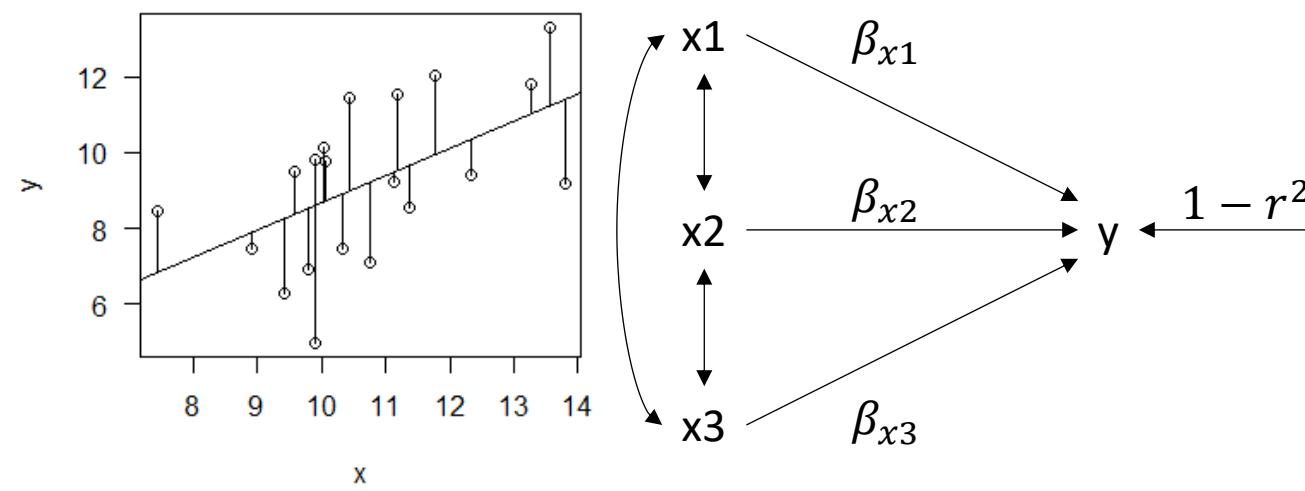
Logistic regression

- To quantify effects, we can backtransform the parameter estimates to the probability scale to illustrate effects in e.g. a graph



Overview of (generalized) linear models

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



Overview of (generalized) linear models

- Binary/proportional data: Logistic regression

