Plan for this week

- Today
  - General comments on regression/ANOVA reports
  - Discussion of exercise 5 (logistic regression)
  - Lecture on Poisson and Negative Binomial regression for count data
  - Seminar, presentations by Emile and Allyson from Melbourne
  - Exercise (Poisson/negative binomial regression) (Hand-in, deadline Sunday)
- Wednesday
  - Discussion session Poisson exercise
  - Lecture on mixed models (Masahito)
  - Seminar: How can we improve this course?
  - Exercise on mixed models
- Sunday (Friday)
  - Deadline for Poisson exercise report

ANOVA reports, general comments

- Biological questions/predictions (e.g. why do we expect an effect of maternal host on larval performance?)
- “To assess differences in larval performance between the two host plants, we fitted…”
- Models are **fitted** to data (not “made”, “created” etc.)
- R code in Appendix

ANOVA reports, general comments

- Units! (Also in tables, text)
- Proper tables (no copying from R)
- Table headers: **Parameter estimates** (not “coefficients”, “model summary” etc)
- Explain all parameters/properties (e.g. for boxplots, explain median, quartiles, whiskers, individual data points)

Discussion of exercise 5

- Seed germination data

Seed germination analysis

- Probability of seed germination increases with time to sowing (after-ripening time): seeds mature gradually after dispersal
- Negative effect of seed size: larger seeds require longer periods of after-ripening

```
##
## Call:
## glm(formula = germ2 ~ timetosowing + MCseed, family = "binomial",
##     data = subdat, weights = nseed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7999  -0.7108  -0.4028   0.8715   3.3335
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.175210   0.369903 -11.287  < 2e-16 ***
## timetosowing  0.039120   0.003308  11.825  < 2e-16 ***
## MCseed       -0.217828   0.035230  -6.183 6.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 536.70  on 230  degrees of freedom
## Residual deviance: 293.46  on 228  degrees of freedom
## AIC: 345.2
##
## Number of Fisher Scoring iterations: 5
```

To calculate the duration of after-ripening needed for a 50% germination rate, we use the equation above to find that this would be 106.7 days in this population.

```
-coefs[1,1]/coefs[2,1]
```

```
## [1] 106.7274
```

To quantify the seed size effect, we can ask how this changes for a seed that is one standard deviation larger or smaller than the mean.

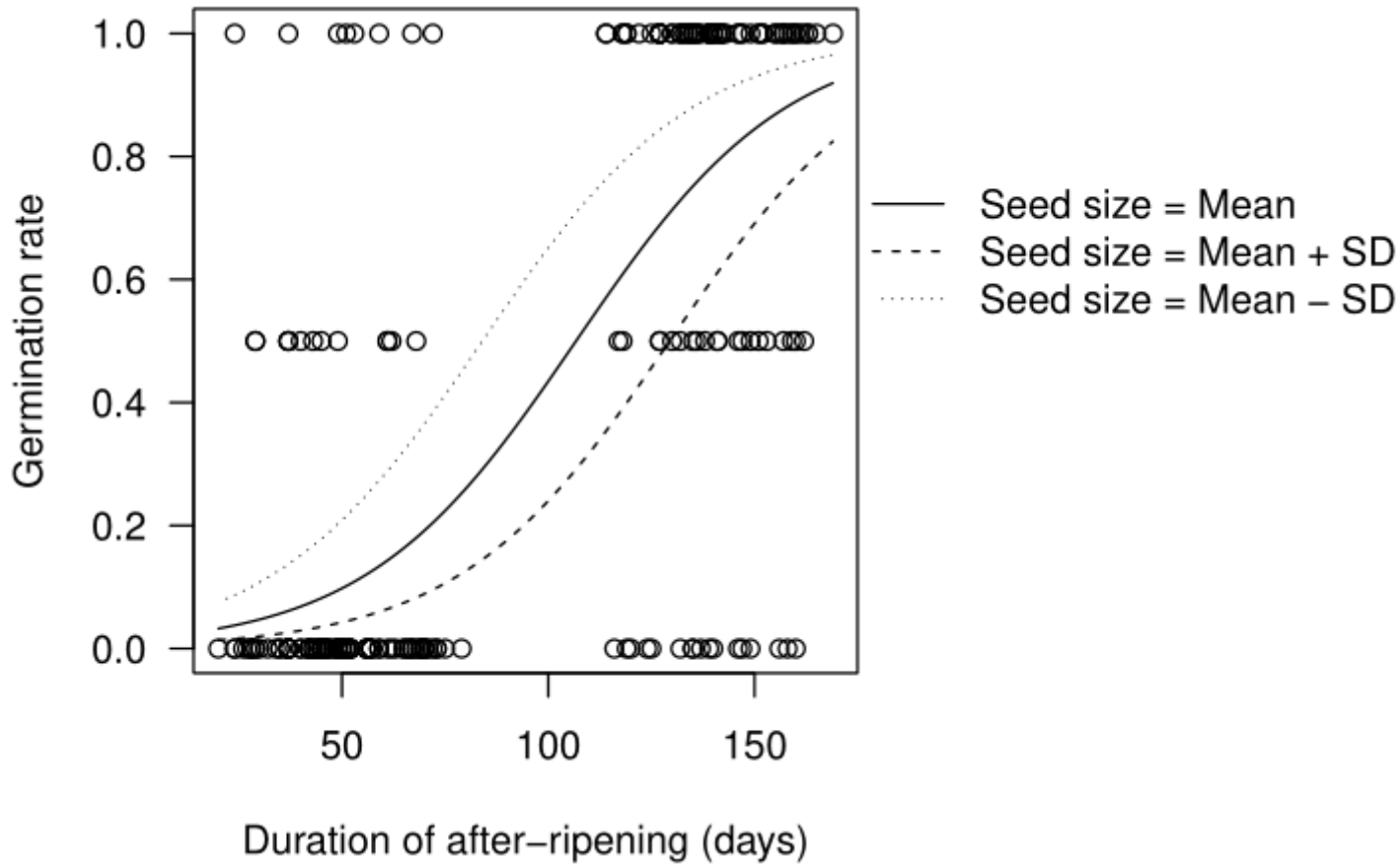
```
-(coefs[1,1] + coefs[3,1]*sd(subdat$MCseed))/coefs[2,1]
```

```
## [1] 129.424
```

```
-(coefs[1,1] - coefs[3,1]*sd(subdat$MCseed))/coefs[2,1]
```

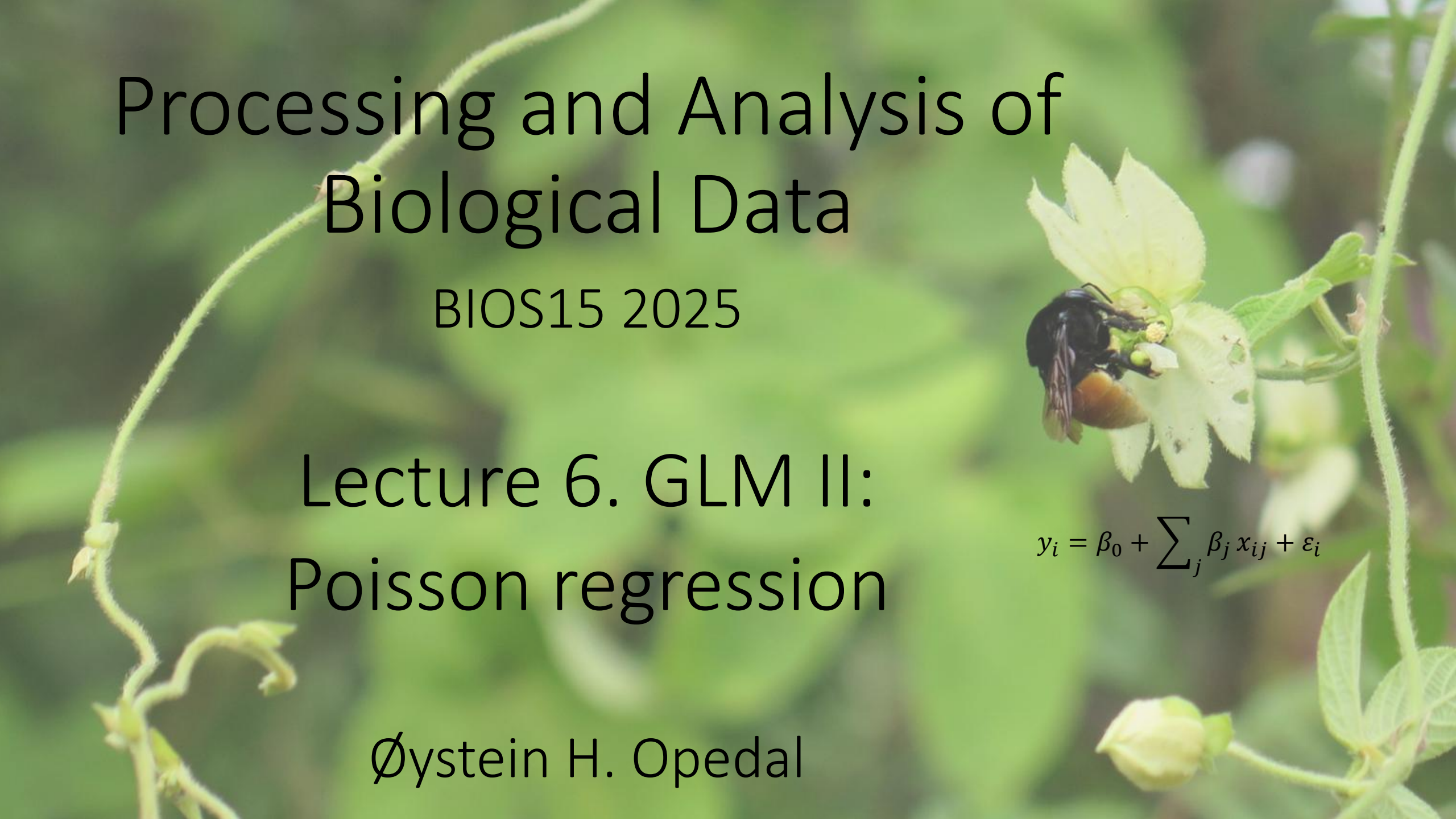
```
## [1] 84.03079
```

We could write the results like this: The probability of germination increased with longer duration of after-ripening (Fig. 1, Table 1). A seed of average size would have 50% probability of germinating when sown after 106.7 days of after-ripening. For a seed one standard deviation larger or smaller than the mean, this period would change to 129.4 days and 84.0 days, respectively.


1.0
0.8
0.6
0.4
0.2
0.0
Germination rate
50
100
150
Duration of after−ripening (days)
Seed size = Mean
Seed size = Mean + SD
Seed size = Mean − SD

Processing and Analysis of Biological Data

BIOS15 2025



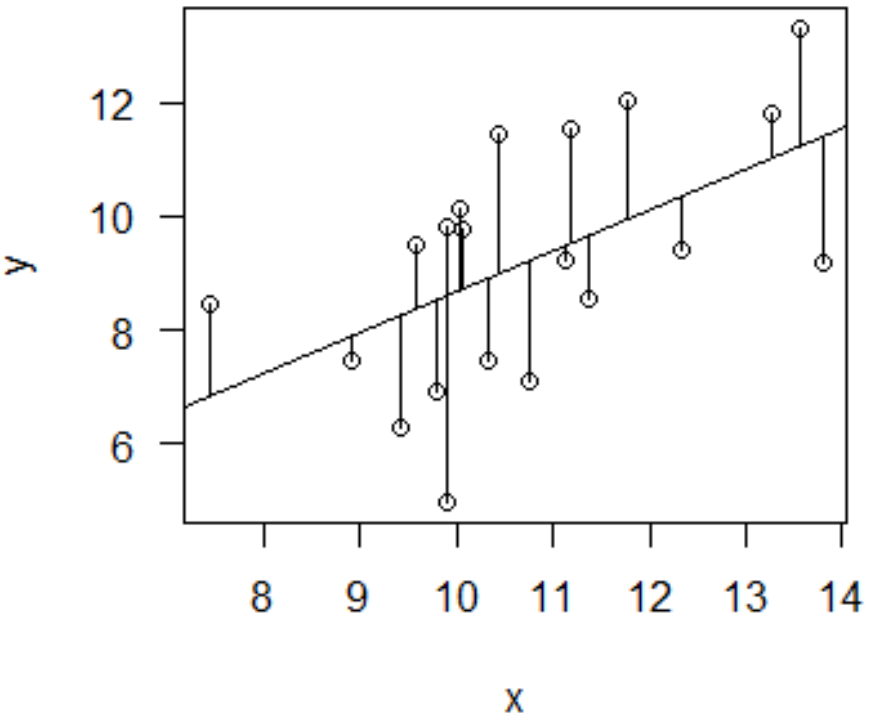
Lecture 6. GLM II: Poisson regression

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

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



Generalized linear models

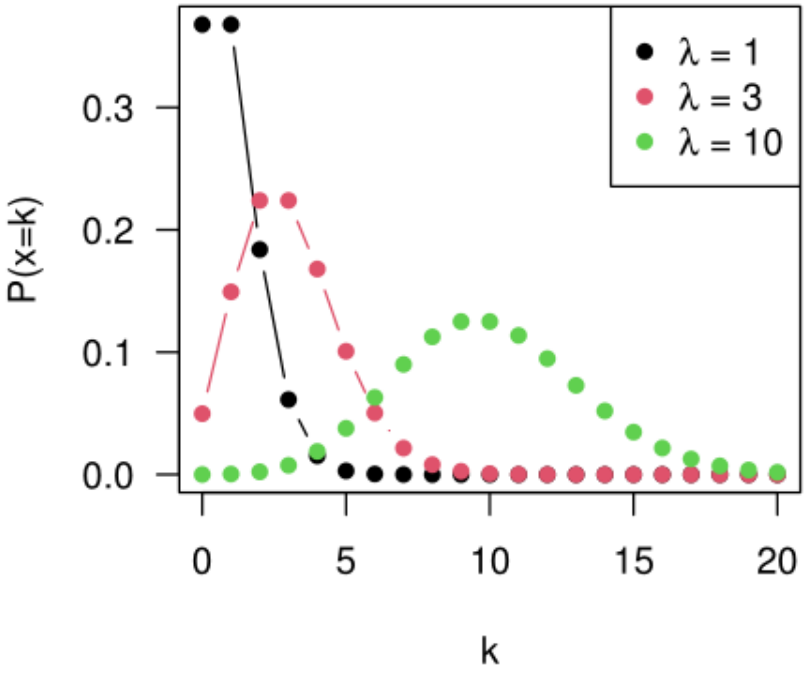
- Generalized linear models extend the linear model by relaxing the assumption of normally distributed residuals
- The model connects a response variable to the familiar linear predictor (η) through a **link function** (g^{-1})
- The link functions are specific to different **error distributions**, the 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 count data

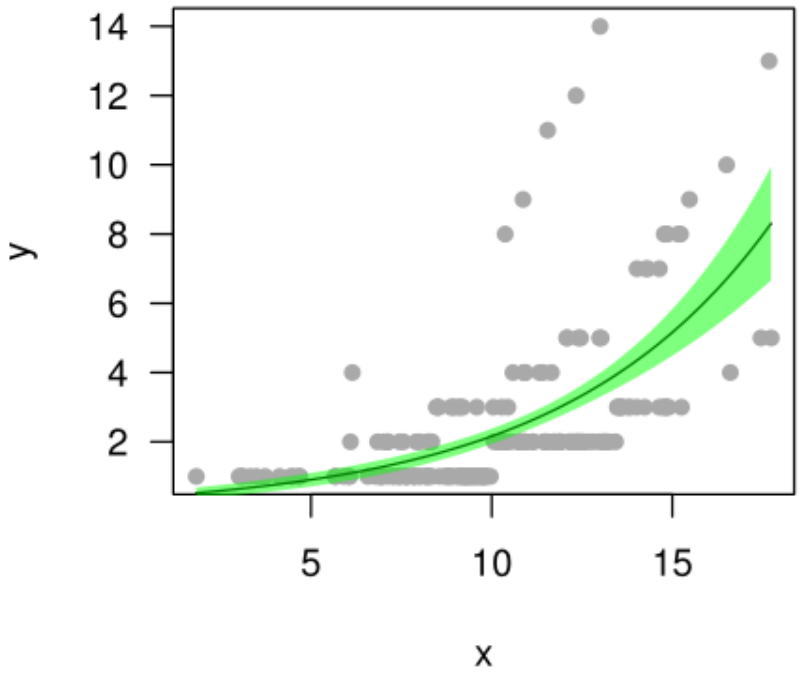
- Count data (not only for fish) can be analysed with a Poisson error distribution
- The Poisson distribution has a single parameter λ giving both the mean and the variance.
- When λ is small, the distribution is skewed.
- When λ is large, the Poisson approaches the normal distribution



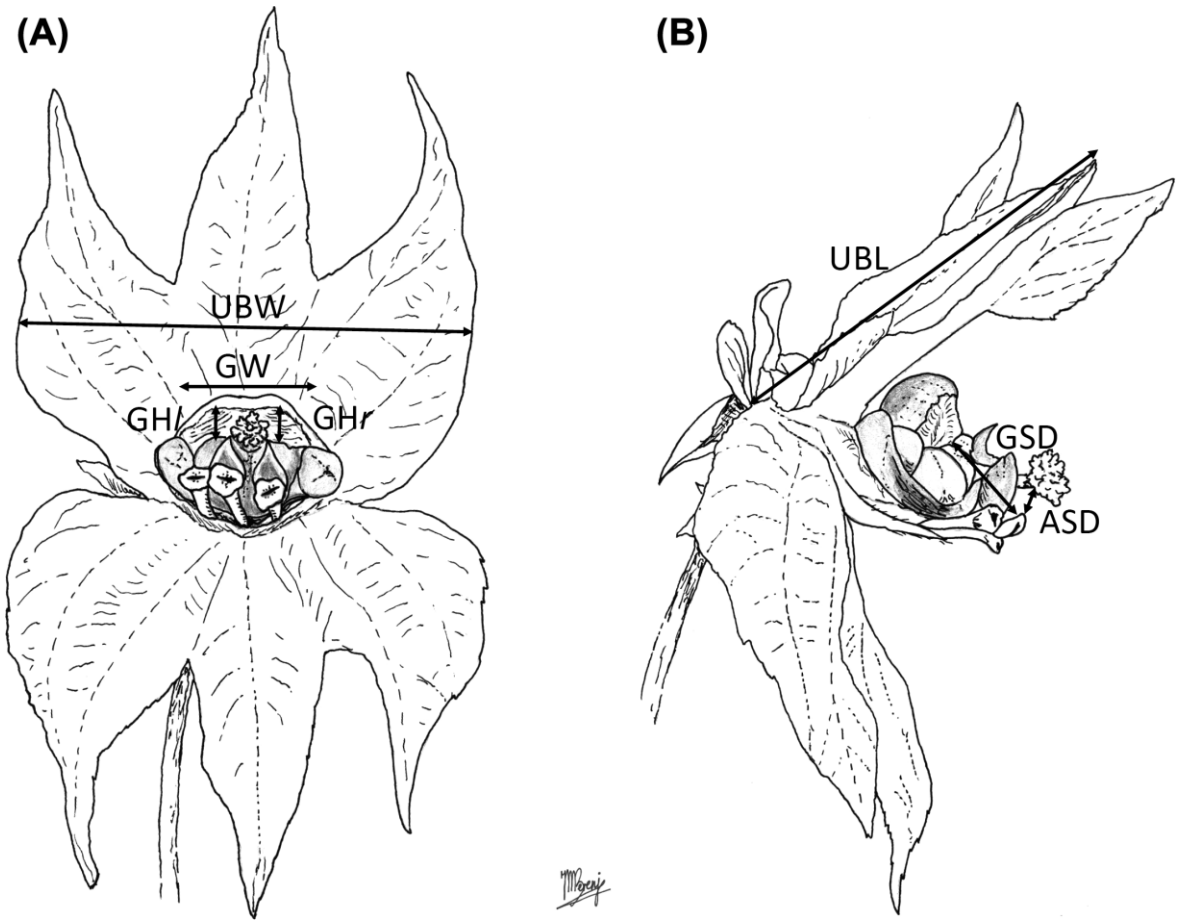
k = number of occurrences

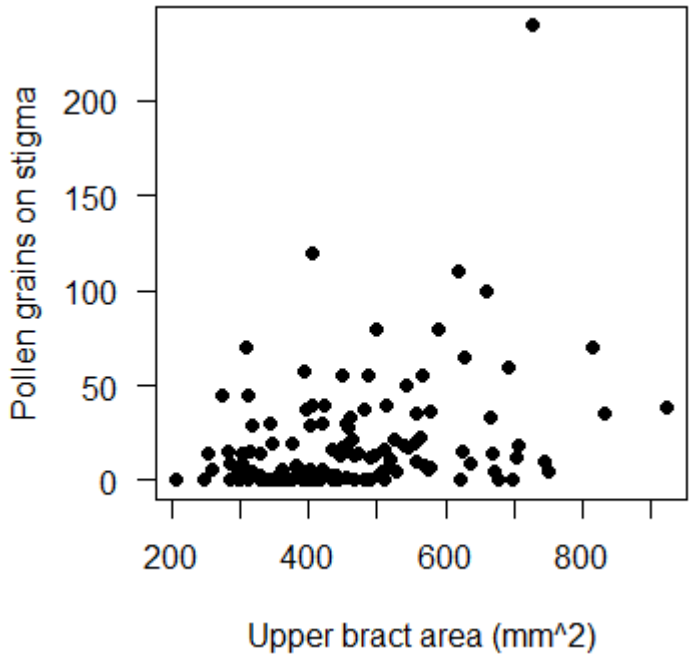
Poisson regression

- A GLM with Poisson errors is called a Poisson regression
- The link function is the natural log, and the inverse link is the exponential
- The data must be integers, i.e. whole numbers



Example: number of pollen grains on stigma





Poisson 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 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, but there are options

```
##
## Call:
## glm(formula = y ~ x, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0059  -0.6632  -0.3636   0.4005   5.5808
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.89507    0.18198  -4.918 8.72e-07 ***
## x            0.16570    0.01491  11.116  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 286.20  on 199  degrees of freedom
## Residual deviance: 157.49  on 198  degrees of freedom
## AIC: 689.1
##
```

Poisson regression model in R

- Because of the log link function, we can interpret the slope as the proportional change in y per unit change in x
- (Recall that log-transformation and mean-scaling have very similar effects)
- Here, a unit change in x increases log(y) by 0.17, and y thus increases by exp(0.17) = 1.185 = 18.5%

```
##
## Call:
## glm(formula = y ~ x, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0059  -0.6632  -0.3636   0.4005   5.5808
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.89507    0.18198  -4.918 8.72e-07 ***
## x            0.16570    0.01491  11.116  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 286.20  on 199  degrees of freedom
## Residual deviance: 157.49  on 198  degrees of freedom
## AIC: 689.1
##
```

The r^2 in Poisson regression

- We can quantify model fit through various "Pseudo r^2" metrics (see lecture notes)

$$1 - \frac{Residual\ deviance}{Null\ deviance}$$

Poisson regression model in R

- If the variance increase disproportionally with the mean, there is **overdispersion** in the data
- Overdispersion is a problem if the residual deviance is much higher than the residual degrees of freedom

```
##
## Call:
## glm(formula = y ~ x, family = "poisson")
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -7.339  -3.851  -3.015  -2.147  59.211
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.054938   0.094237  -0.583     0.56
## x            0.217419   0.007729  28.129   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8793.6  on 199  degrees of freedom
## Residual deviance: 8005.2  on 198  degrees of freedom
## AIC: 8452.8
##
```

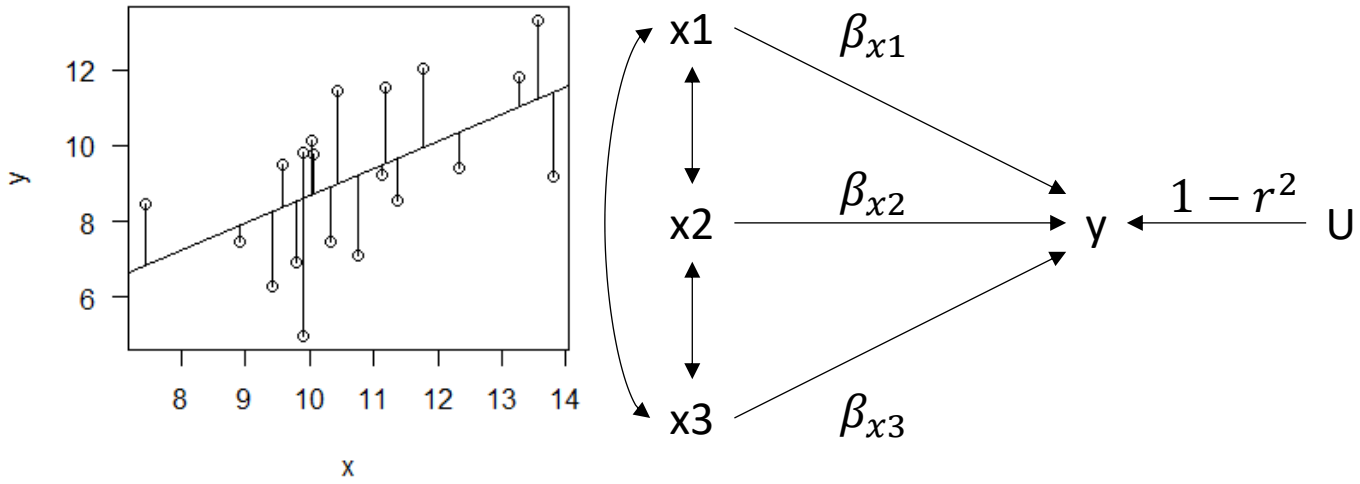
Poisson regression model in R

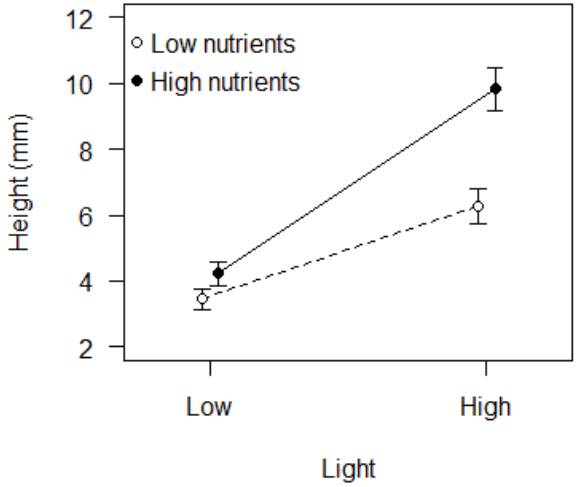
- If the variance increase disproportionally with the mean, there is **overdispersion** in the data
- Overdispersion is a problem if the residual deviance is much higher than the residual degrees of freedom
- In this case, we fit the model with **negative binomial** errors instead, which allows the variance to increase disproportionally

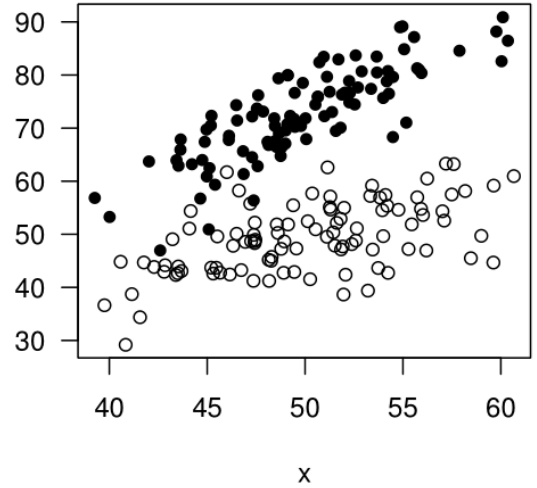
```
##
## Call:
## glm.nb(formula = y ~ x, init.theta = 0.2993347963, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3934  -1.0868  -0.8307  -0.5080   4.6882
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17822    0.50977  -2.311   0.0208 *
## x            0.31932    0.04817   6.628 3.39e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2993) family taken to be 1)
##
##     Null deviance: 248.15  on 199  degrees of freedom
## Residual deviance: 215.64  on 198  degrees of freedom
## AIC: 1083.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2993
##           Std. Err.:  0.0304
##
##  2 x log-likelihood:  -1077.2360
```

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
- Count data: Poisson GLM
- Overdispersed count data: Negative binomial GLM