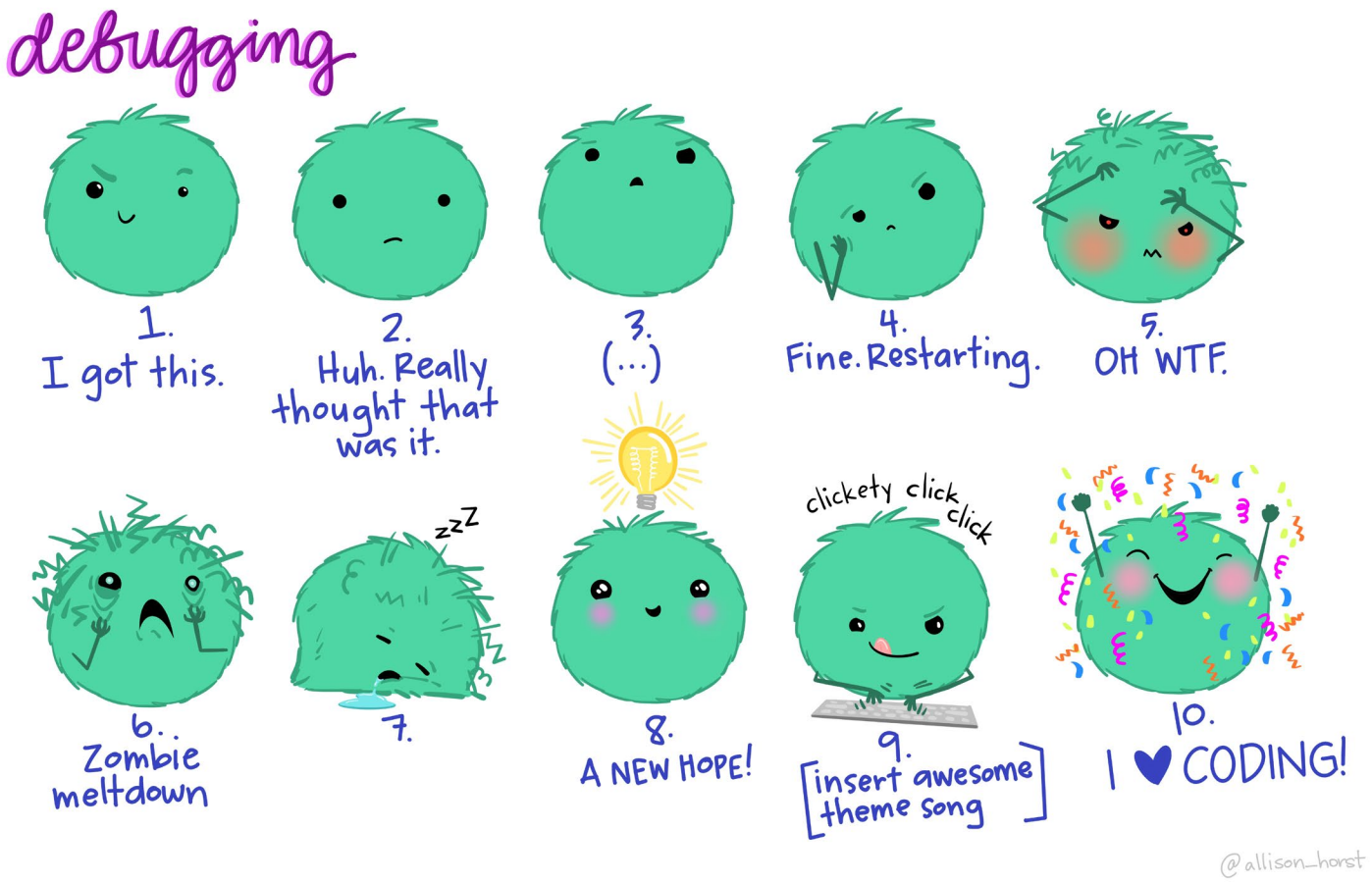
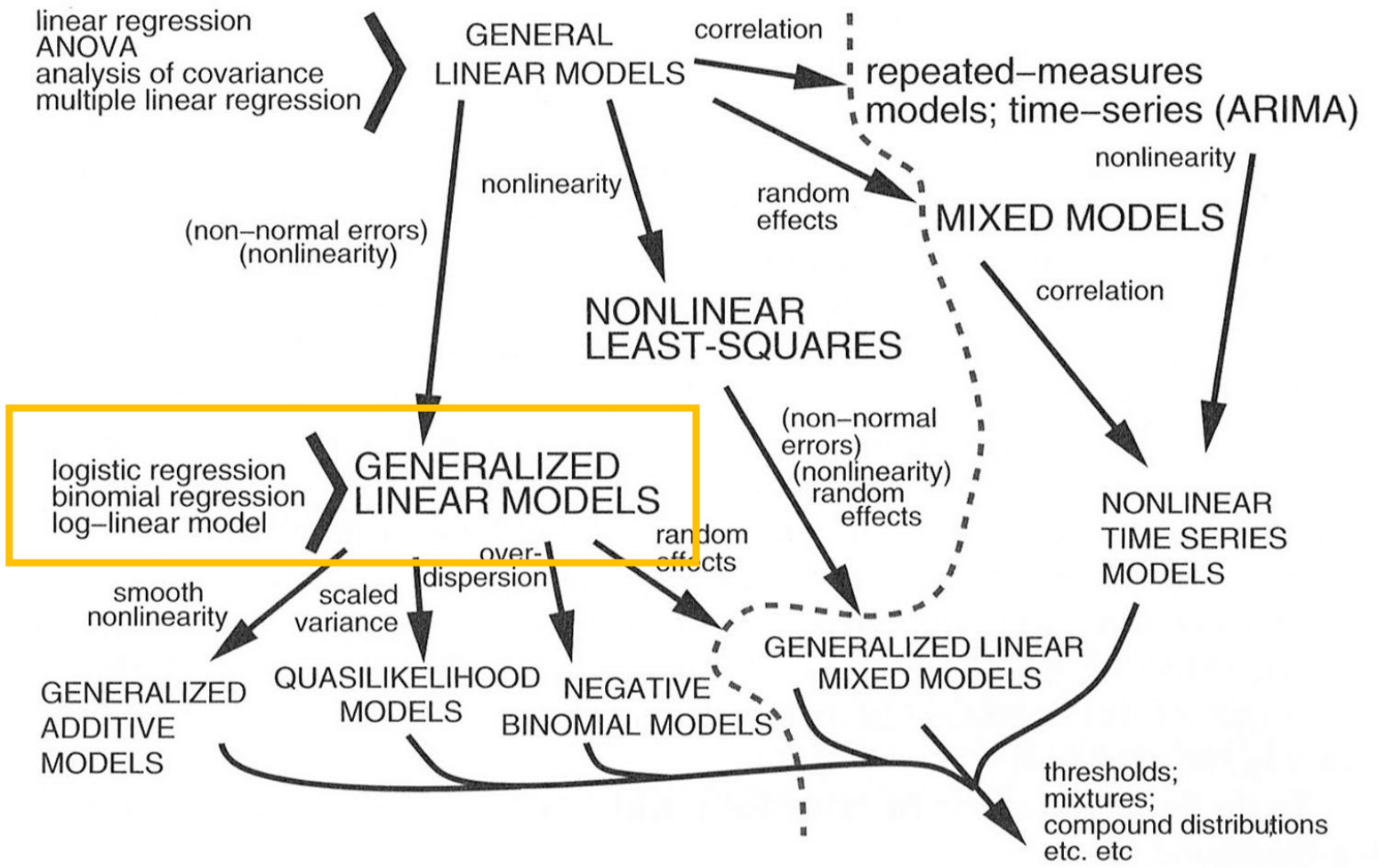
Questions from last week?

- What are generalized linear models (GLMs)?
- Practice using common ones: binomial and poisson
- Learn how to interpret regression coefficients within the context of generalized models
- Learn how to effectively visualize GLMs



A road map for the next ~10 weeks



This week

- More glm practice:
  - Interpreting coefficients
  - hypothesis testing
  - interactions
- Hurdle models to deal with zero-inflated data

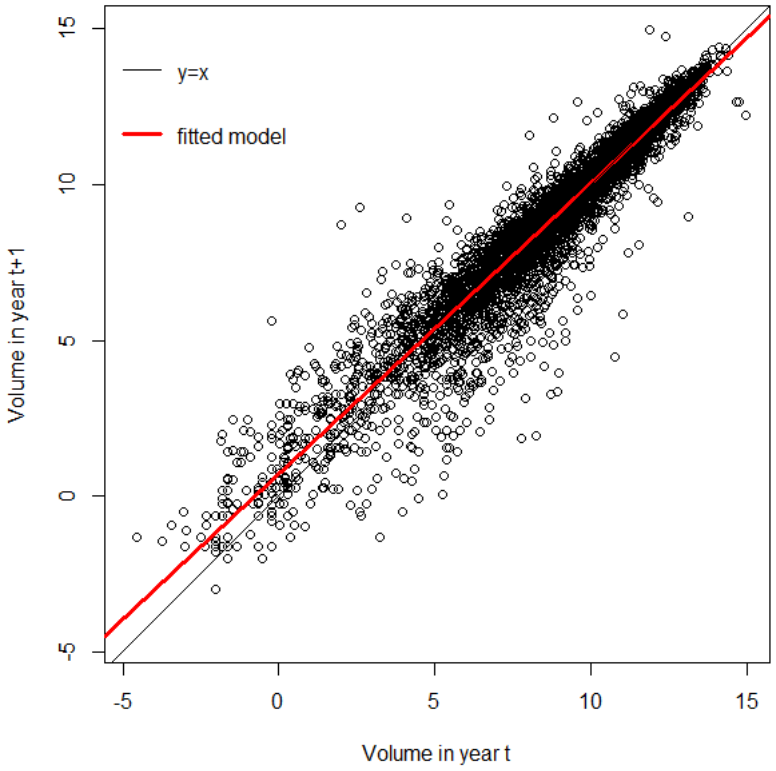
Deeper dive into interpreting GLM coefficients

- What do the intercepts and slopes actually mean?
- More practice going between the original and transformed scales
- Return to cholla demography models

Growth

$$Size_{t+1,i} = \beta_0 + \beta_1 Size_{t,i} + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

```
## Call:
## lm(formula = log(vol_t1) ~ log(vol_t), data = cholla)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4402 -0.2518  0.0863  0.3789  6.1755
##
## Coefficients:
##                  Estimate  Std. Error  t value   Pr(>|t|)
## (Intercept)  0.706465  0.035562  19.87    <2e-16 ***
## log(vol_t)   0.932831  0.003766  247.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8641 on 5661 degrees of freedom
##   (1127 observations deleted due to missingness)
## Multiple R-squared:  0.9155, Adjusted R-squared:  0.9155
## F-statistic: 6.135e+04 on 1 and 5661 DF,  p-value: < 2.2e-16
```



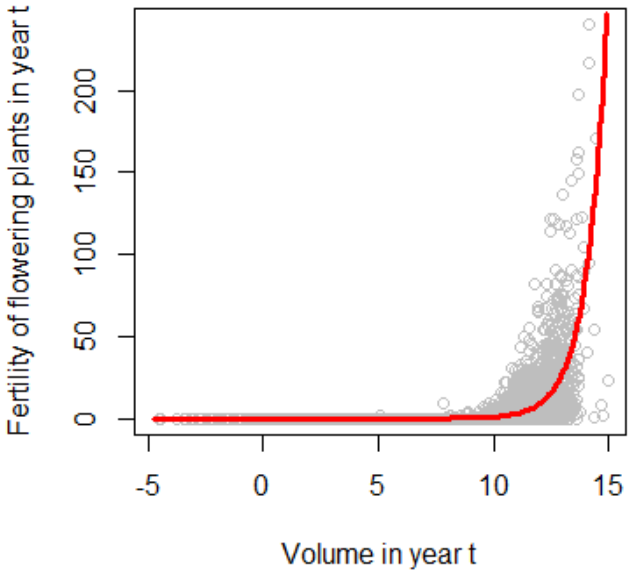
Fertility

$$Fert_{t,i} \sim Poisson(ln(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$ln(fertility) = -11.55 + 1.14 * Size_t$$

```
Call:
glm(formula = Goodbuds_t ~ log(vol_t), family = "poisson", data = cholla)
Deviance Residuals:
Min 1Q Median 3Q Max -19.4929 -1.3081 -0.4177 -0.0513 17.2248

Coefficients:
            Estimate   Std. Error z value Pr(>|z|)
(Intercept) -11.550855 0.075445   -153.1  <2e-16 ***
log(vol_t)  1.143578   0.006056   188.8   <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: 89565 on 6089 degrees of freedom
Residual deviance: 30476 on 6088 degrees of freedom
(700 observations deleted due to missingness)
AIC: 36147
Number of Fisher Scoring iterations: 6
```

We used the exponential to back-transform the predicted values to the observed scale



Some math for thinking about poisson intercept

$$ln(p_i) = \beta_0 + \beta_1 Size_{t,i}$$

$$ln(p_i) = \beta_0 + \beta_1 \underbrace{Size_{t,i}}_{0}$$

When the size measure is zero, the expected value of fertility is

$$p_i = e^{\beta_0}$$

Fertility

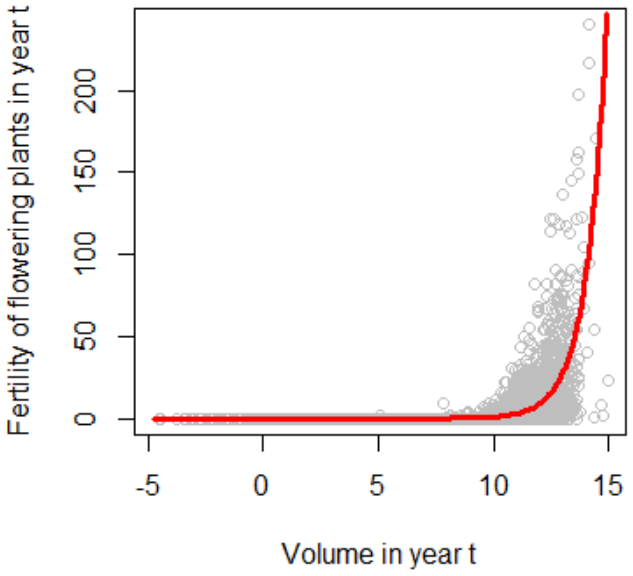
$$Fert_{t,i} \sim Poisson(ln(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$ln(fertility) = -11.55 + 1.14 * Size_t$$

```
Call:
glm(formula = Goodbuds_t ~ log(vol_t), family = "poisson", data = cholla)
Deviance Residuals:
Min 1Q Median 3Q Max -19.4929 -1.3081 -0.4177 -0.0513 17.2248

Coefficients:
            Estimate   Std. Error z value Pr(>|z|)
(Intercept) -11.550855 0.075445   -153.1  <2e-16 ***
log(vol_t)  1.143578   0.006056   188.8   <2e-16 ***
```

Intercept coefficient: at log(vol_t)=0, the expected fertility is exp(-11.55)=0.0000096 offspring

We used the exponential to back-transform the predicted values to the observed scale



Some math for thinking about poisson slope

$$ln(p_i) = \beta_0 + \beta_1 Size_{t,i}$$

Compare equations for when we increase the predictor by 1 unit

$$ln(p_0) = \beta_0 + \beta_1 Size_{t,0}$$

$$\begin{aligned} ln(p_1) &= \beta_0 + \beta_1 (Size_{t,0} + 1) \\ &= \beta_0 + \beta_1 Size_{t,0} + \beta_1 \\ &= ln(p_0) + \beta_1 \end{aligned}$$

So when we increase the predictor by 1 unit, the natural log of the response increases by the slope
Which is the same as

$$p_1 = p_0 e^{\beta_1}$$

When we increase the predictor by 1 unit, the response is *multiplied* by the exponential of the slope

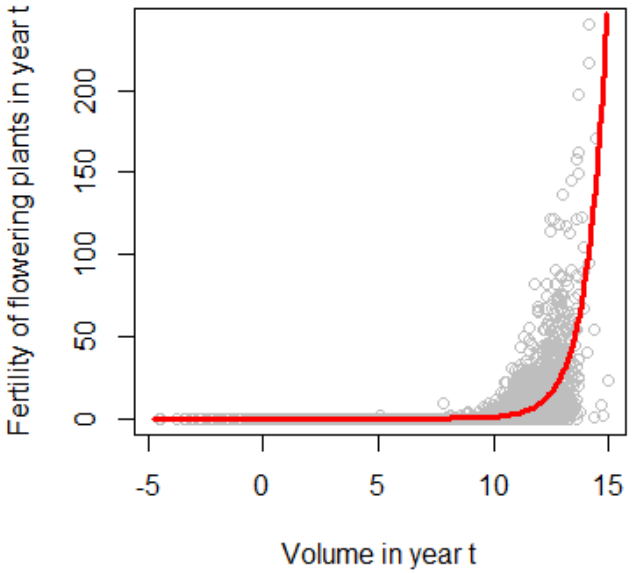
Fertility

$$Fert_{t,i} \sim Poisson(ln(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$ln(fertility) = -11.55 + 1.14 * Size_t$$

```
Call:
glm(formula = Goodbuds_t ~ log(vol_t), family = "poisson", data = cholla)
Deviance Residuals:
Min 1Q Median 3Q Max -19.4929 -1.3081 -0.4177 -0.0513 17.2248

Coefficients:
            Estimate   Std. Error z value Pr(>|z|)
(Intercept) -11.550855 0.075445   -153.1  <2e-16 ***
log(vol_t)  1.143578   0.006056   188.8   <2e-16 ***
```

We used the exponential to back-transform the predicted values to the observed scale



Intercept coefficient: at log(vol_t)=0, the expected fertility is exp(-11.55)=0.0000096 offspring

Slope coefficient: for each unit increase in log(vol_t), the expected fertility is *multiplied by* exp(1.14)=3.13 offspring

Survival

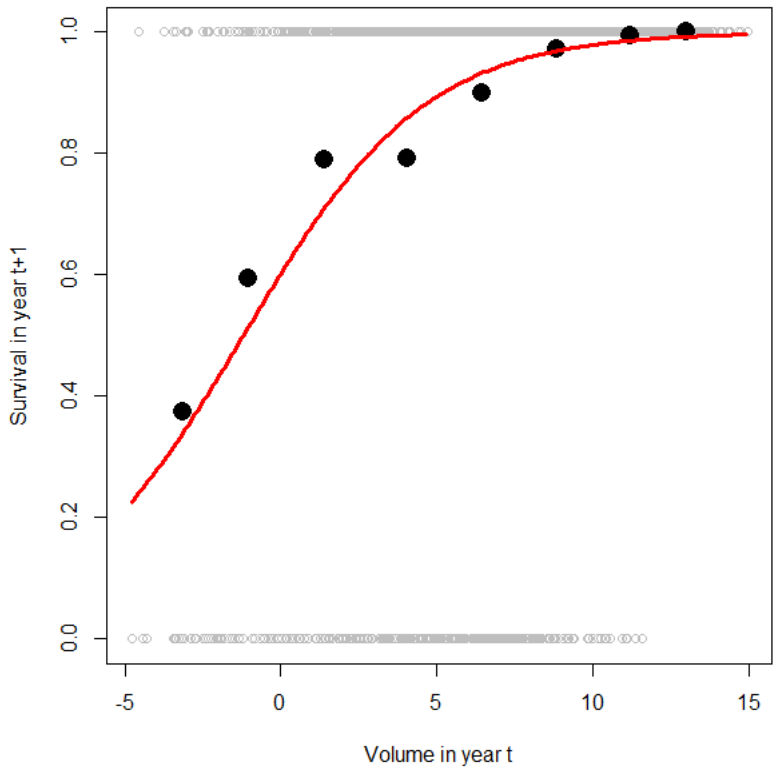
$$Surv_{t+1,i} \sim Bernoulli(logit(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$logit(surv) = 0.41 + 0.34 * Size_t$$

```
Call: glm(formula = Survival_t1 ~ log(vol_t), family = "binomial", data = cholla)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.9670   0.1611   0.2103   0.3115   1.6906

Coefficients:
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  0.41045   0.09214    4.455    8.4e-06 ***
log(vol_t)   0.34344   0.01451    23.672   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2833.2  on 6046  degrees of freedom
Residual deviance: 2194.1  on 6045  degrees of freedom
  (743 observations deleted due to missingness)
AIC: 2198.1

Number of Fisher Scoring iterations: 6
```

We used inverse-logit to back-transform the predicted values to the observed scale

$$invlogit(x) = \frac{e^x}{1 + e^x}$$



Survival

$$Surv_{t+1,i} \sim Bernoulli(logit(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$logit(surv) = 0.41 + 0.34 * Size_t$$

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

Call: glm(formula = Survival_t1 ~ log(vol_t), family = "binomial", data = cholla)

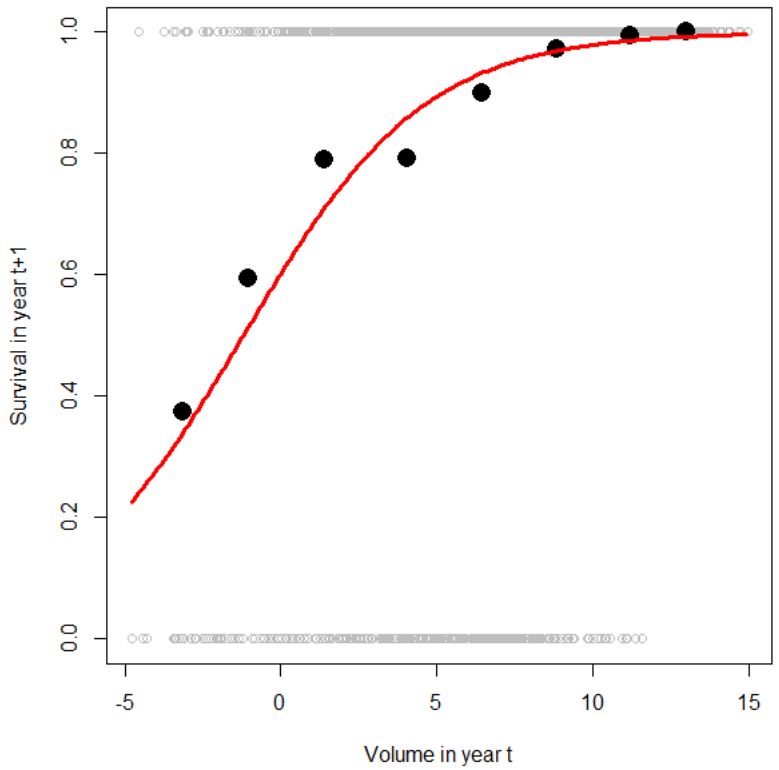
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.9670	0.1611	0.2103	0.3115	1.6906

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.41045	0.09214	4.455	8.4e-06 ***
log(vol_t)	0.34344	0.01451	23.672	< 2e-16 ***

A mechanistic way of thinking about the logit link function is that it calculates the “logged-odds ratio”
-> exponentiate coefficients to get the “odds ratio”



Some math for thinking about binomial intercept

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Size_{t,i}$$

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cancel{Size_{t,i}} \rightarrow 0$$

When the size measure is zero, the expected odds ratio of survival is

$$\frac{p_i}{1-p_i} = e^{\beta_0}$$

Or, the odds of survival p_i are e^{β_0} times the odds of mortality $1 - p_i$

$$p_i = e^{\beta_0}(1-p_i)$$

Survival

$$Surv_{t+1,i} \sim Bernoulli(logit(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$logit(surv) = 0.41 + 0.34 * Size_t$$

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

Call: glm(formula = Survival_t1 ~ log(vol_t), family = "binomial", data = cholla)

Deviance Residuals:

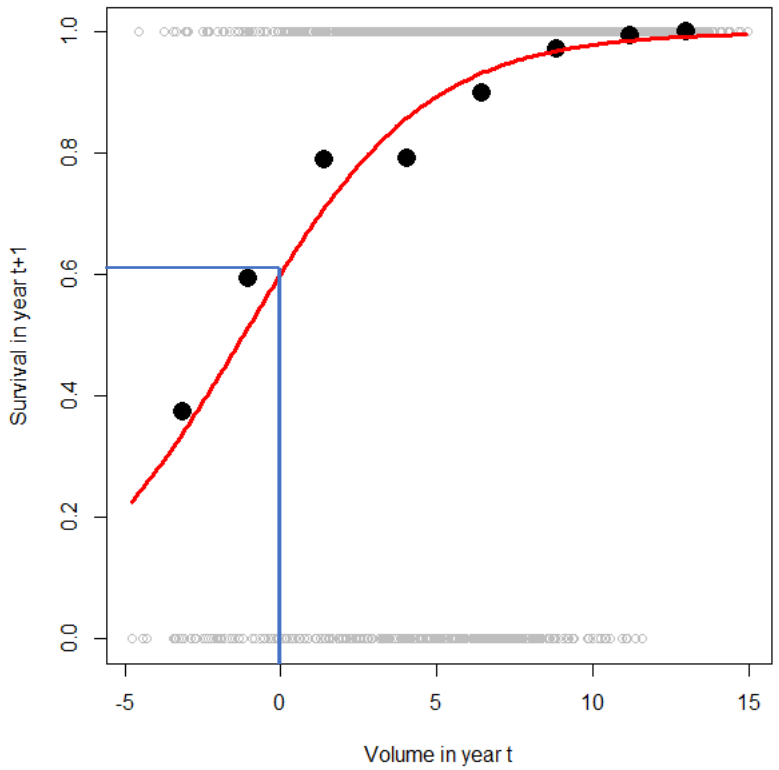
Min	1Q	Median	3Q	Max
-2.9670	0.1611	0.2103	0.3115	1.6906

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.41045	0.09214	4.455	8.4e-06 ***
log(vol_t)	0.34344	0.01451	23.672	< 2e-16 ***

A mechanistic way of thinking about the logit link function is that it calculates the "logged-odds ratio"
-> exponentiate coefficients to get the "odds ratio"

Intercept coefficient: at log(vol_t)=0, the odds of survival are exp(0.41)=1.5 times higher than the odds of mortality



Some math for thinking about binomial slope

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Size_{t,i}$$

Compare equations for when we increase the predictor by 1 unit

$$ln\left(\frac{p_0}{1-p_0}\right) = \beta_0 + \beta_1 Size_{t,0}$$

$$\begin{aligned} ln\left(\frac{p_1}{1-p_1}\right) &= \beta_0 + \beta_1(Size_{t,0} + 1) \\ &= \beta_0 + \beta_1 Size_{t,0} + \beta_1 \\ &= ln\left(\frac{p_0}{1-p_0}\right) + \beta_1 \end{aligned}$$

So when we increase the predictor by 1 unit, the log-odds ratio increases by the slope
Which is the same as

$$\frac{p_1}{1-p_1} = (\frac{p_0}{1-p_0})e^{\beta_1}$$

When we increase the predictor by 1 unit, the odds ratio is *multiplied* by the exponential of the slope

Survival

$$Surv_{t+1,i} \sim Bernoulli(logit(p_i) = \beta_0 + \beta_1 Size_{t,i})$$

$$logit(surv) = 0.41 + 0.34 * Size_t$$

Call: glm(formula = Survival_t1 ~ log(vol_t), family = "binomial", data = cholla)

Deviance Residuals:

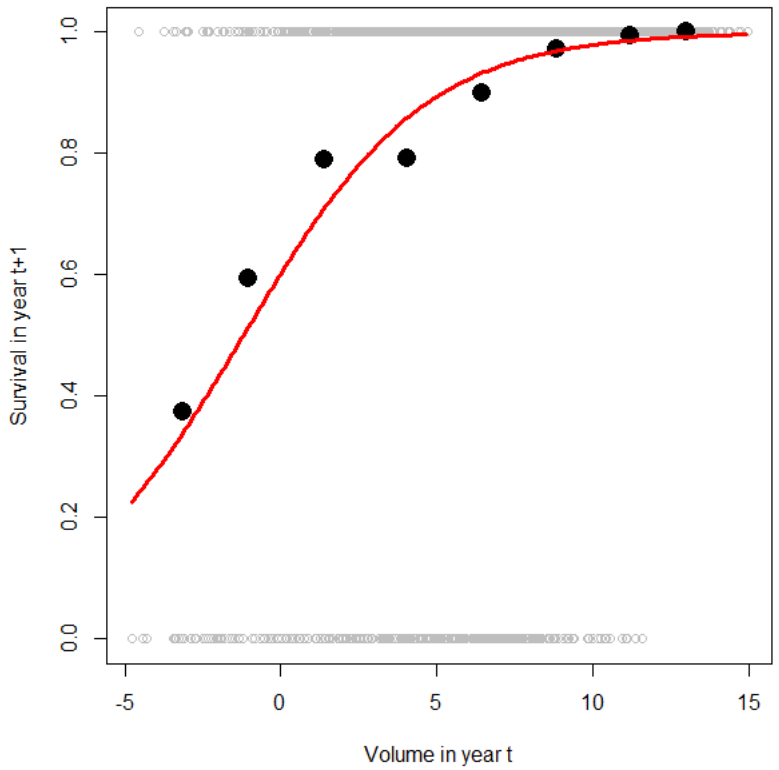
Min	1Q	Median	3Q	Max
-2.9670	0.1611	0.2103	0.3115	1.6906

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.41045	0.09214	4.455	8.4e-06 ***
log(vol_t)	0.34344	0.01451	23.672	< 2e-16 ***

A mechanistic way of thinking about the logit link function is that it calculates the "logged-odds ratio"
-> exponentiate coefficients to get the "odds ratio"

Slope coefficient: for each unit increase in log(vol_t), the odds ratio of survival changes by a multiplicative factor of exp(0.34)=1.4

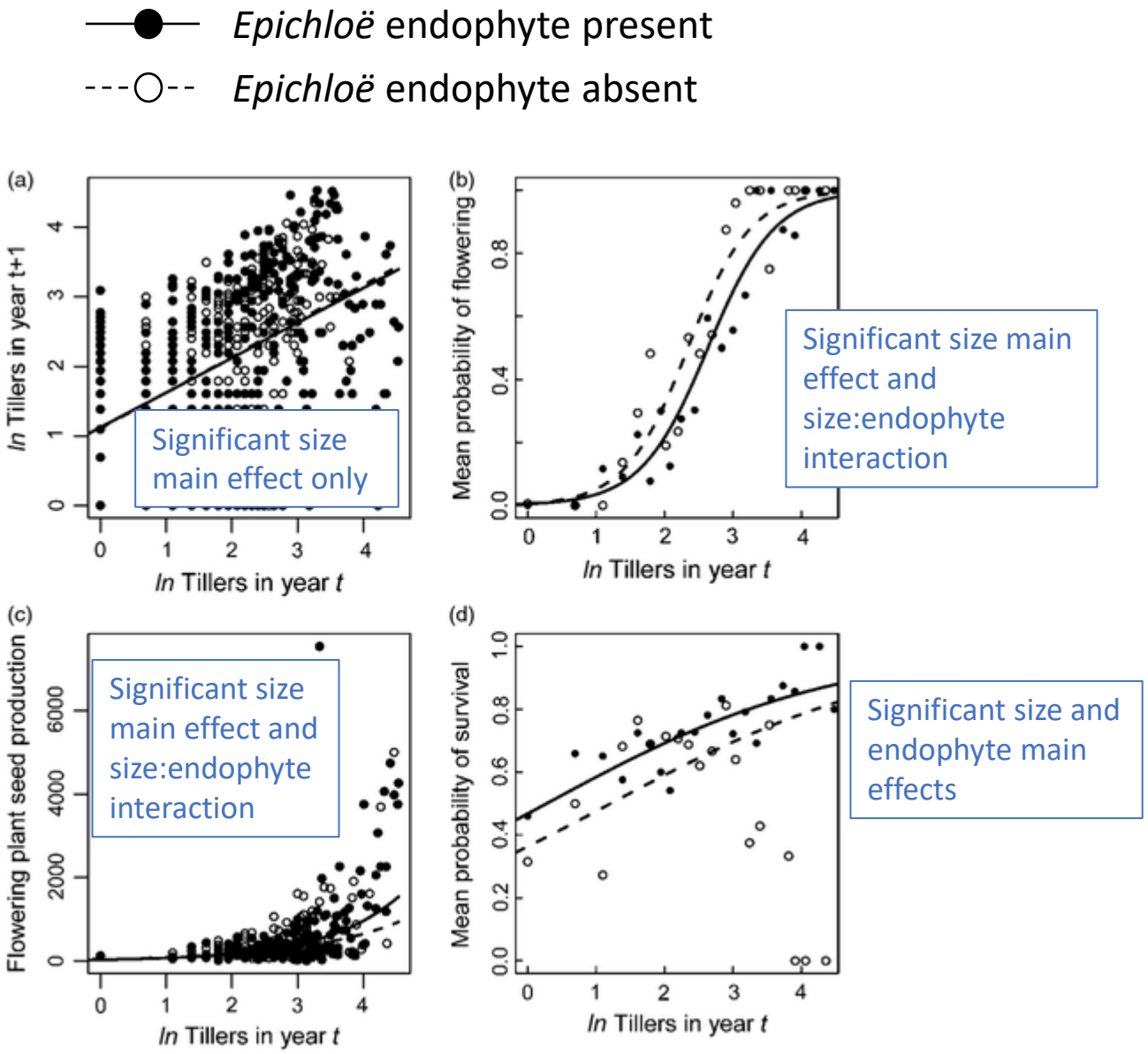


Adding complexity

Just like LMs, GLMs can be extended to include multiple combinations of categorical and continuous predictors, as well as interactions

Example:

Vital rate ~ size*endophyte



Model: Vital rate ~ size*endophyte

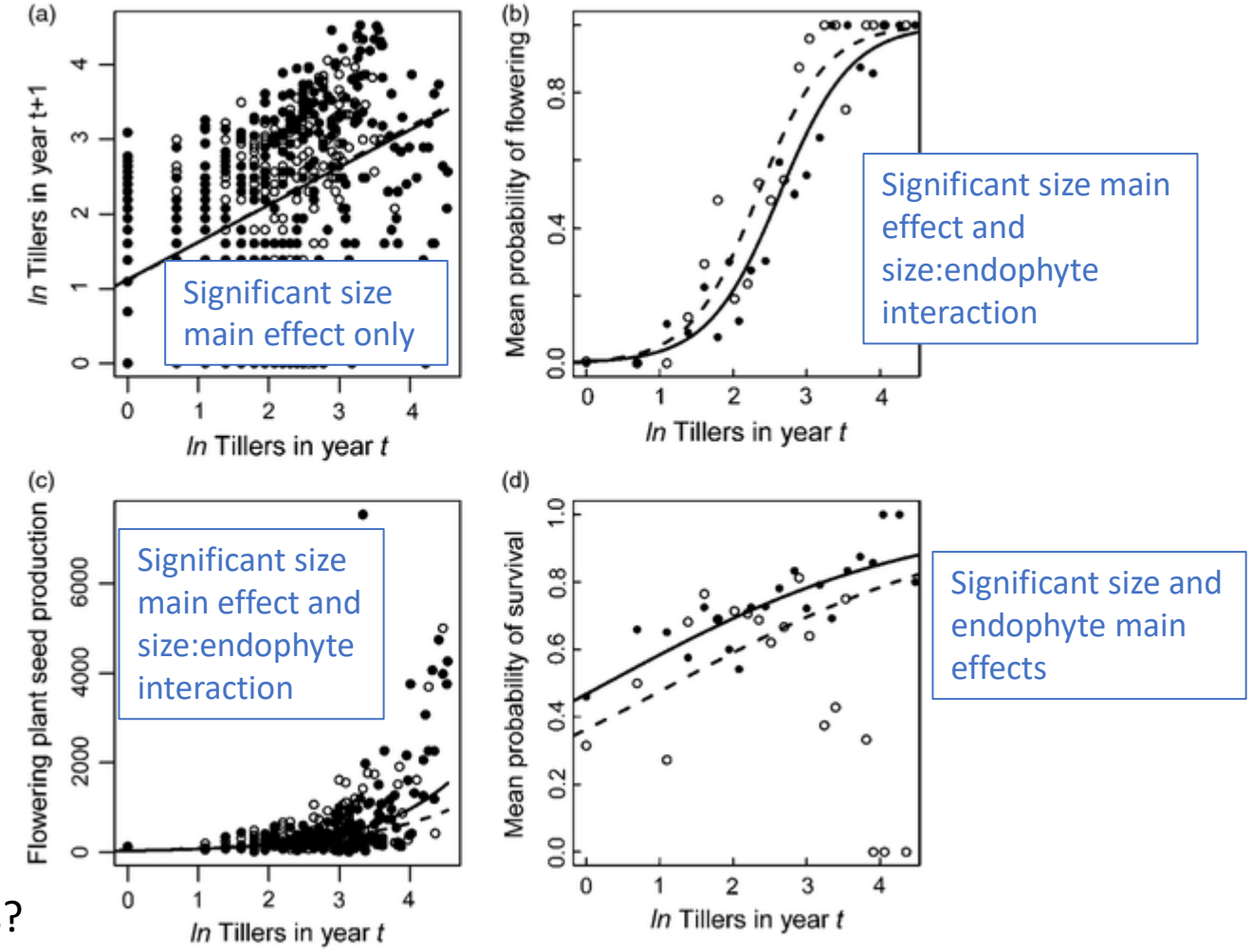
P. alsodes		
E^-	E^+	Description
−0.56	−0.13	Survival intercept
0.46	0.47	Survival slope
1.11	1.12	Growth intercept
0.51	0.50	Growth slope
−5.09	−5.37	Flowering intercept
2.17	2.04	Flowering slope
3.76	3.36	Seeds intercept
0.68	0.88	Seeds slope

How do we interpret these coefficients?

What are the vital rate functions for E+ vs. E- populations?

Epichloë endophyte present

Epichloë endophyte absent

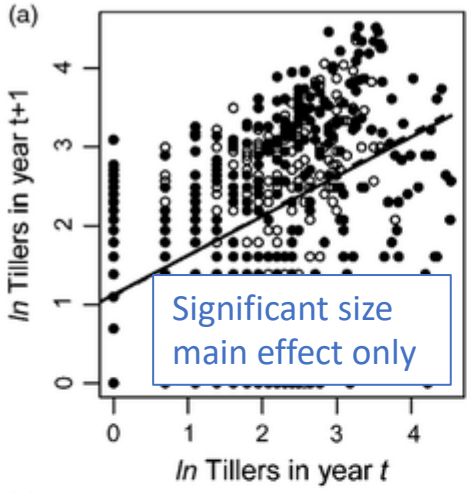


Model: Vital rate ~ size*endophyte

P. alsodes		
E^-	E^+	Description
−0.56	−0.13	Survival intercept
0.46	0.47	Survival slope
1.11	1.12	Growth intercept
0.51	0.50	Growth slope
−5.09	−5.37	Flowering intercept
2.17	2.04	Flowering slope
3.76	3.36	Seeds intercept
0.68	0.88	Seeds slope

Epichloë endophyte present

Epichloë endophyte absent



How do we interpret these coefficients?

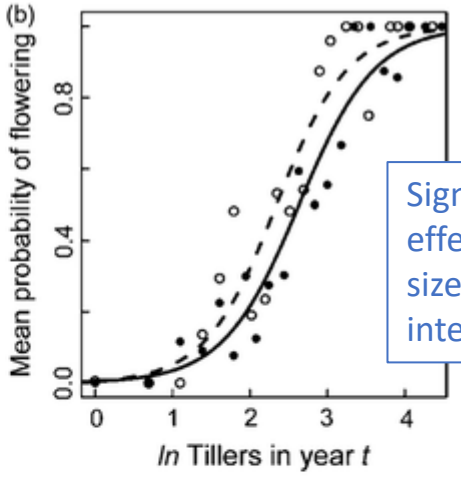
What are the vital rate functions for E+ vs. E- populations?

Model: Vital rate ~ size*endophyte

P. alsodes		
E^-	E^+	Description
−0.56	−0.13	Survival intercept
0.46	0.47	Survival slope
1.11	1.12	Growth intercept
0.51	0.50	Growth slope
−5.09	−5.37	Flowering intercept
2.17	2.04	Flowering slope
3.76	3.36	Seeds intercept
0.68	0.88	Seeds slope

Epichloë endophyte present

Epichloë endophyte absent



Significant size main effect and size:endophyte interaction

How do we interpret these coefficients?

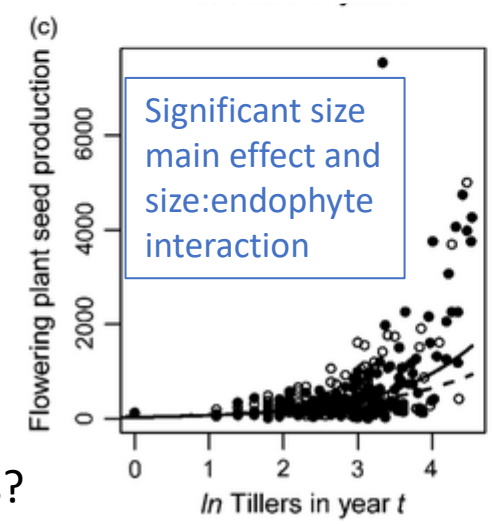
What are the vital rate functions for E+ vs. E- populations?

Model: Vital rate ~ size*endophyte

—●— *Epichloë* endophyte present

---○-- *Epichloë* endophyte absent

P. alsodes		
E^-	E^+	Description
−0.56	−0.13	Survival intercept
0.46	0.47	Survival slope
1.11	1.12	Growth intercept
0.51	0.50	Growth slope
−5.09	−5.37	Flowering intercept
2.17	2.04	Flowering slope
3.76	3.36	Seeds intercept
0.68	0.88	Seeds slope



How do we interpret these coefficients?

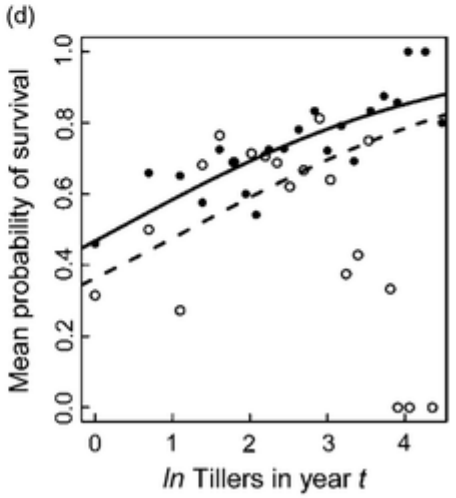
What are the vital rate functions for E+ vs. E- populations?

Model: Vital rate ~ size*endophyte

Epichloë endophyte present

Epichloë endophyte absent

P. alsodes		
E^-	E^+	Description
−0.56	−0.13	Survival intercept
0.46	0.47	Survival slope
1.11	1.12	Growth intercept
0.51	0.50	Growth slope
−5.09	−5.37	Flowering intercept
2.17	2.04	Flowering slope
3.76	3.36	Seeds intercept
0.68	0.88	Seeds slope



Significant size and endophyte main effects

How do we interpret these coefficients?

What are the vital rate functions for E+ vs. E- populations?

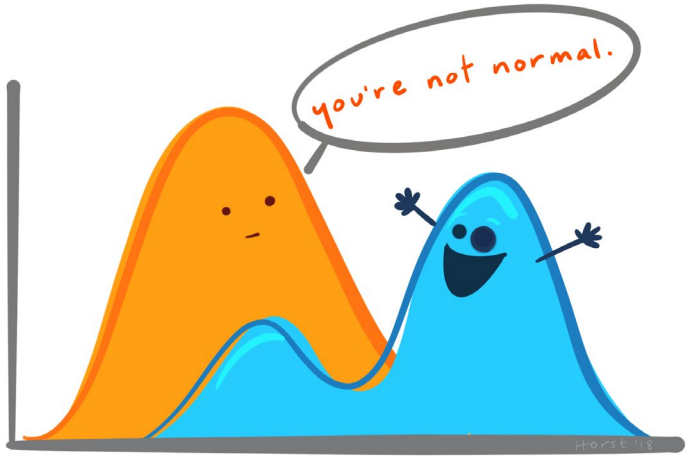
Break

glm assumptions

- Appropriate model and lack of outliers
- You are using the correct distribution and link function
- Explanatory variables included in linear predictor on correct scale
- Correct variance function
- No overdispersion beyond mean-variance relationship expected from the specified distribution
  - E.g. Poisson distribution: variance equal to mean
- Independent observations

What is overdispersion?

Recall, in linear models, we were restrained to normally-distributed residuals



To relax this assumption, especially for data whose process of generation are expected to not be normal, we use glms

In a glm, what we do instead is specify the expected distribution, and use a linearizing link function to transform the response variable to a linear/normal scale

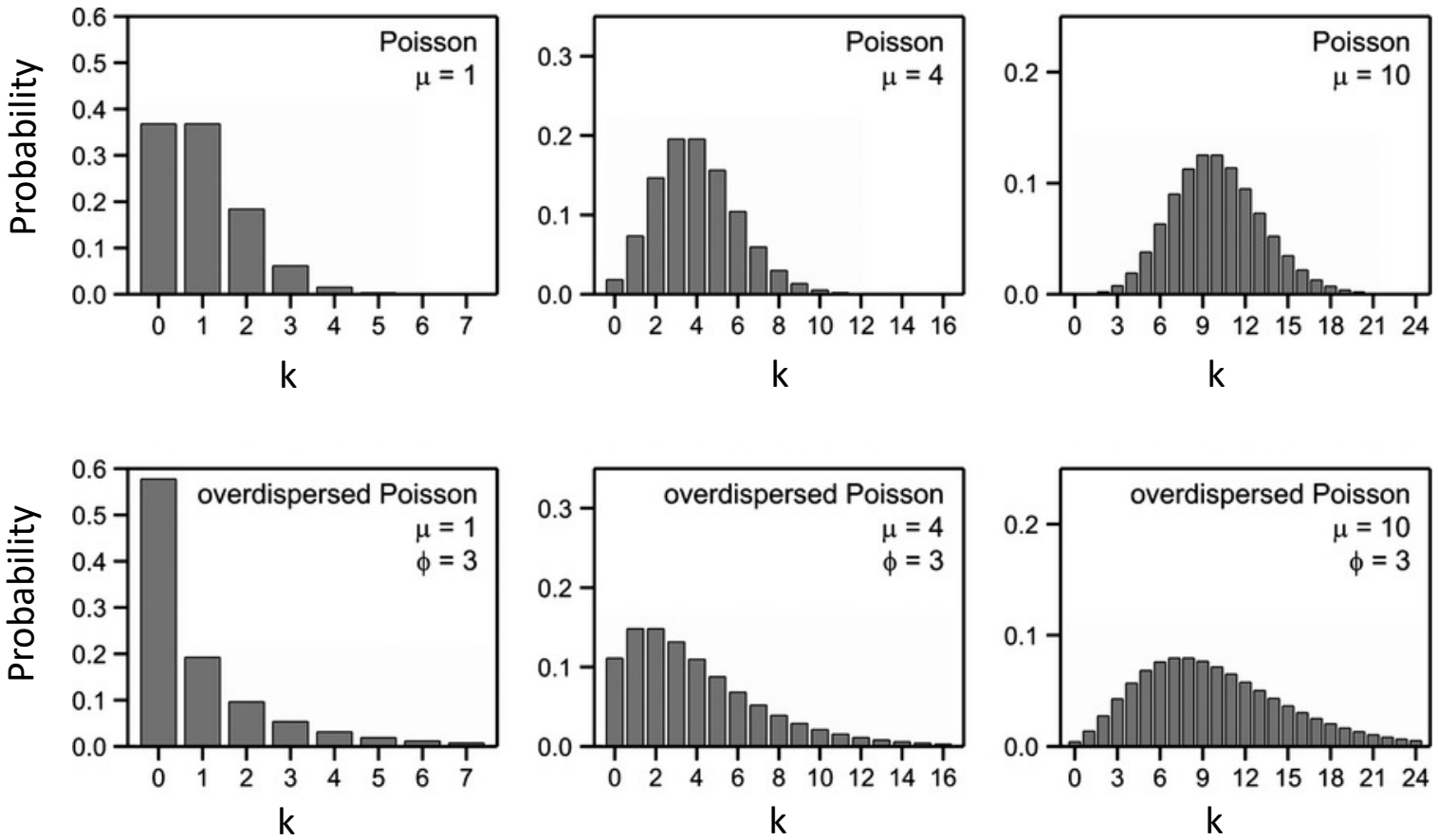
But when we specify a distribution, those distributions come with specific mean-variance relationships

What is overdispersion?

For example, the probability mass function of a poisson distribution is given by:

$$\frac{\lambda^k e^{-\lambda}}{k!}$$

Where *k* is the number of occurrences and λ is the expected value (mean)



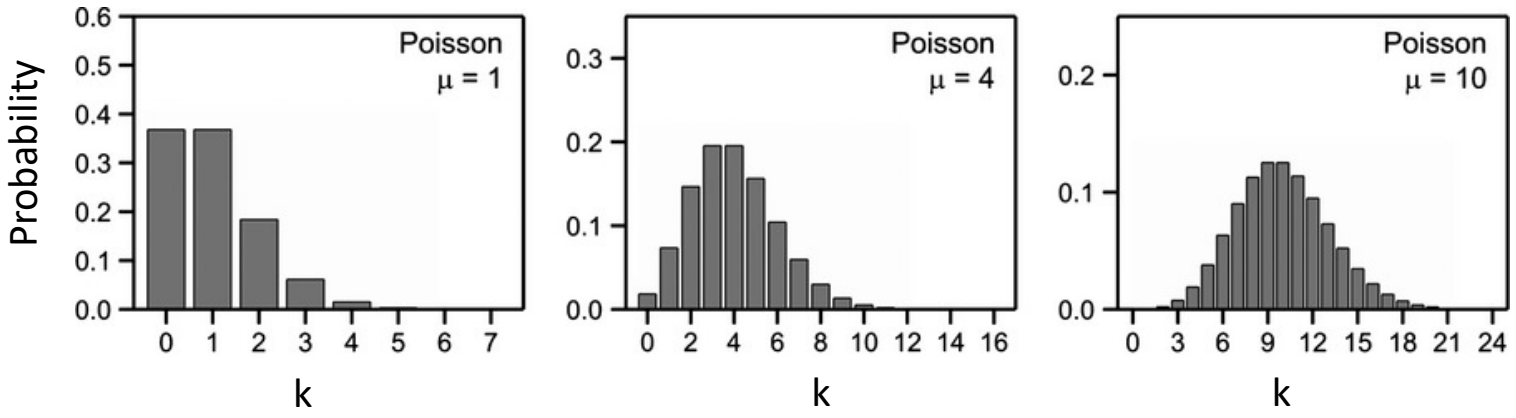
Importantly, for the Poisson distribution, mean=variance

Examples of what overdispersed Poisson distributions may look like

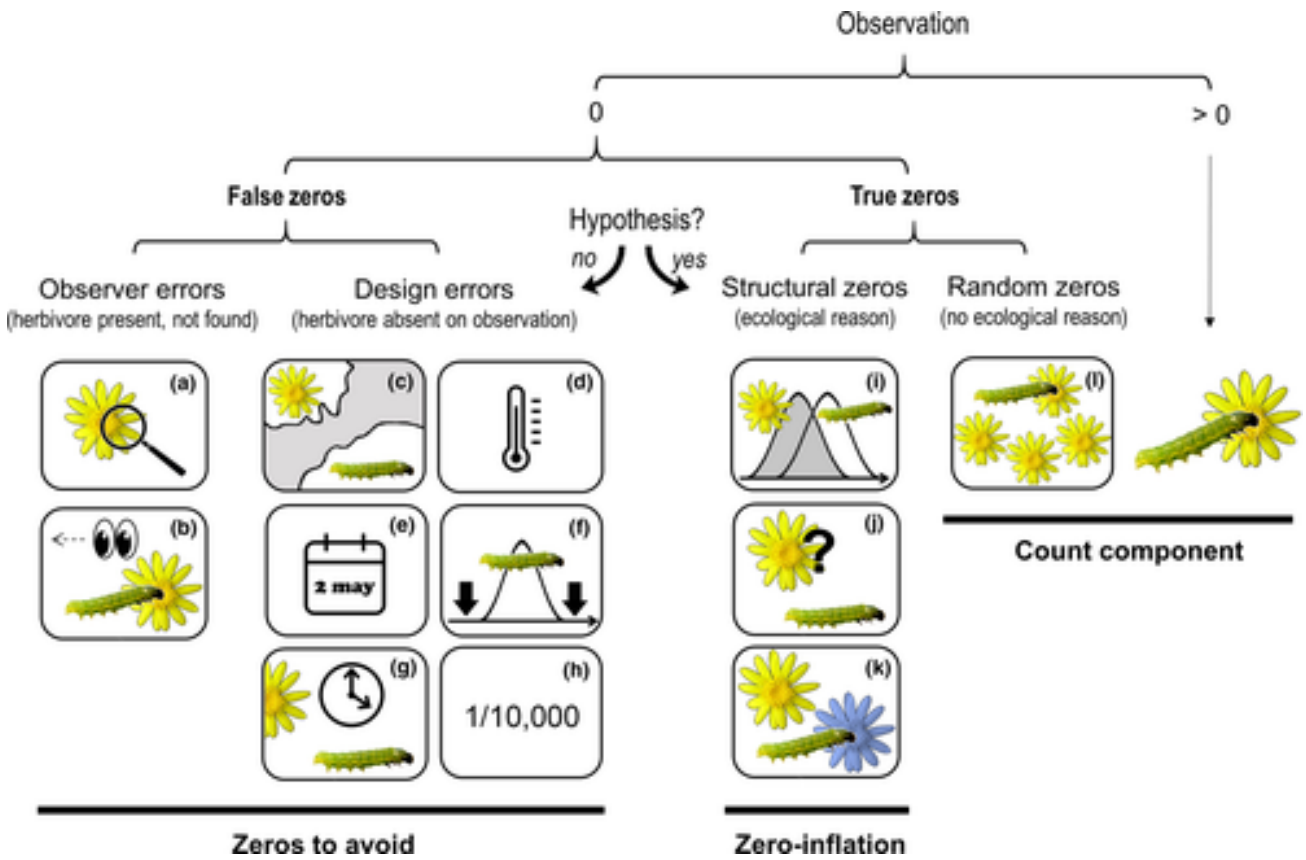
Figure from Goedhart et al. 2014

What causes overdispersion?

- So many things…data are often not well behaved
- One specific reason that comes up often: too many zeros (zero inflation)

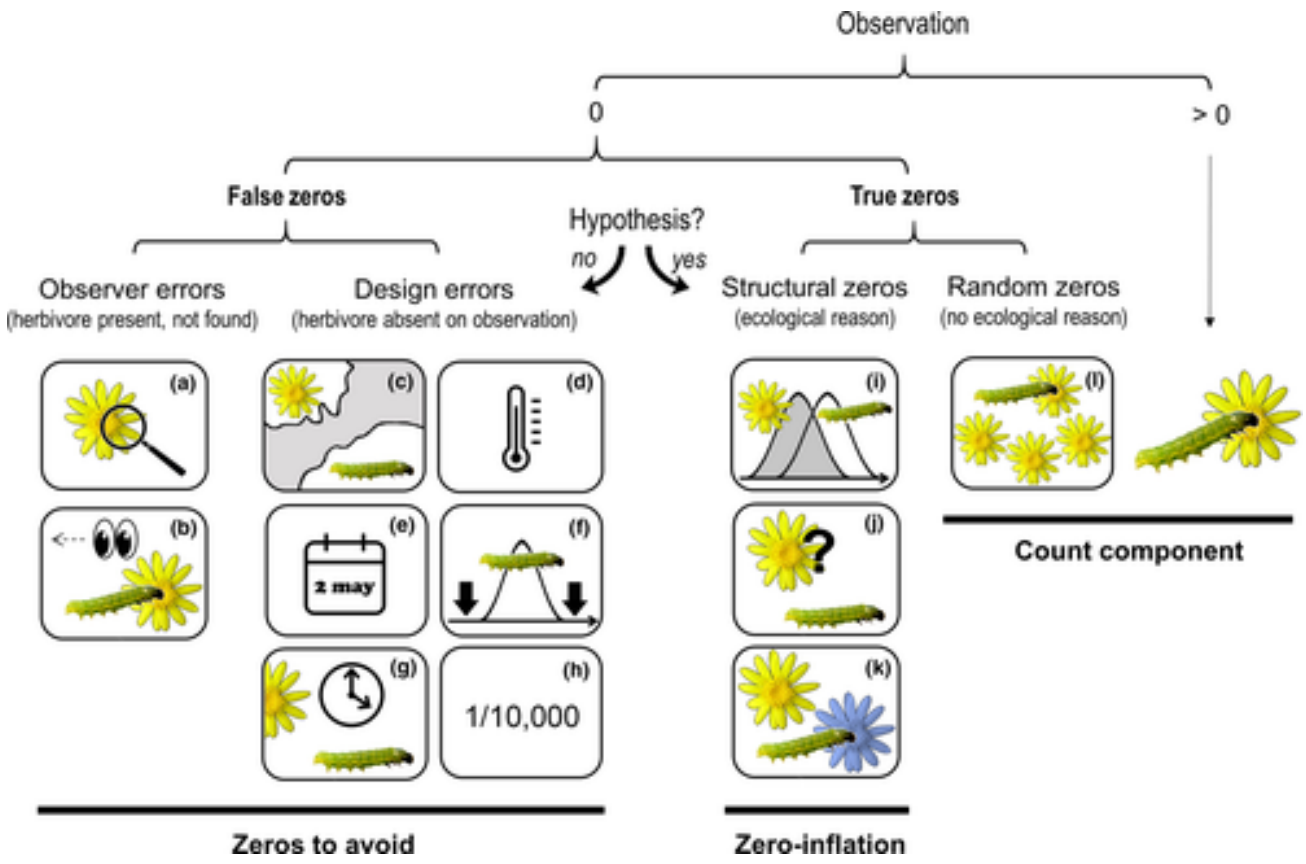


Why are there so many zeros in your data?



(Blasco-Moreno et al. 2019)

How many zeros is too many zeros?



Depends on the distribution you are using to model your residual variation

Can check dispersion parameter for your model against the assumed value based on the distribution

e.g. Dispersion for poisson family is taken to be 1

Where $\varphi = \frac{deviance}{df}$

Can also more specifically check for zero inflation using zero inflation index

(Blasco-Moreno et al. 2019)

What to do with overdispersed and/or zero-inflated data?

Type of zeros	Source	Generator process	Over-dispersion	Zero inflation	Modelling approach
False zeros	Design errors	Poor experimental design	—	—	Remove before analysis
	Observer errors	Lack of experience	—	—	Remove before analysis
True zeros	Random	Sampling variability	No	No	Poisson
			Yes	No	NB
	Structural	Outside the count process	No	Yes	ZIP or ZAP
			Yes	Yes	ZINB or ZANB

NB, negative binomial (McCullagh & Nelder, 1989).
ZIP, zero-inflated poisson (Lambert, 1992) and ZINB, zero-inflated negative binomial (Greene, 1994).
ZAP, zero-altered poisson and ZANB, zero-altered negative binomial (Mullahy, 1986).

(Blasco-Moreno et al. 2019)

Zero-altered (ZA) models, aka hurdle models

- Separate count data into zero and non-zero counts, analyze separately
- First do a ~binomial model for zeros vs. non-zeros
- Then for the subset of data with all non-zero counts, analyze magnitude

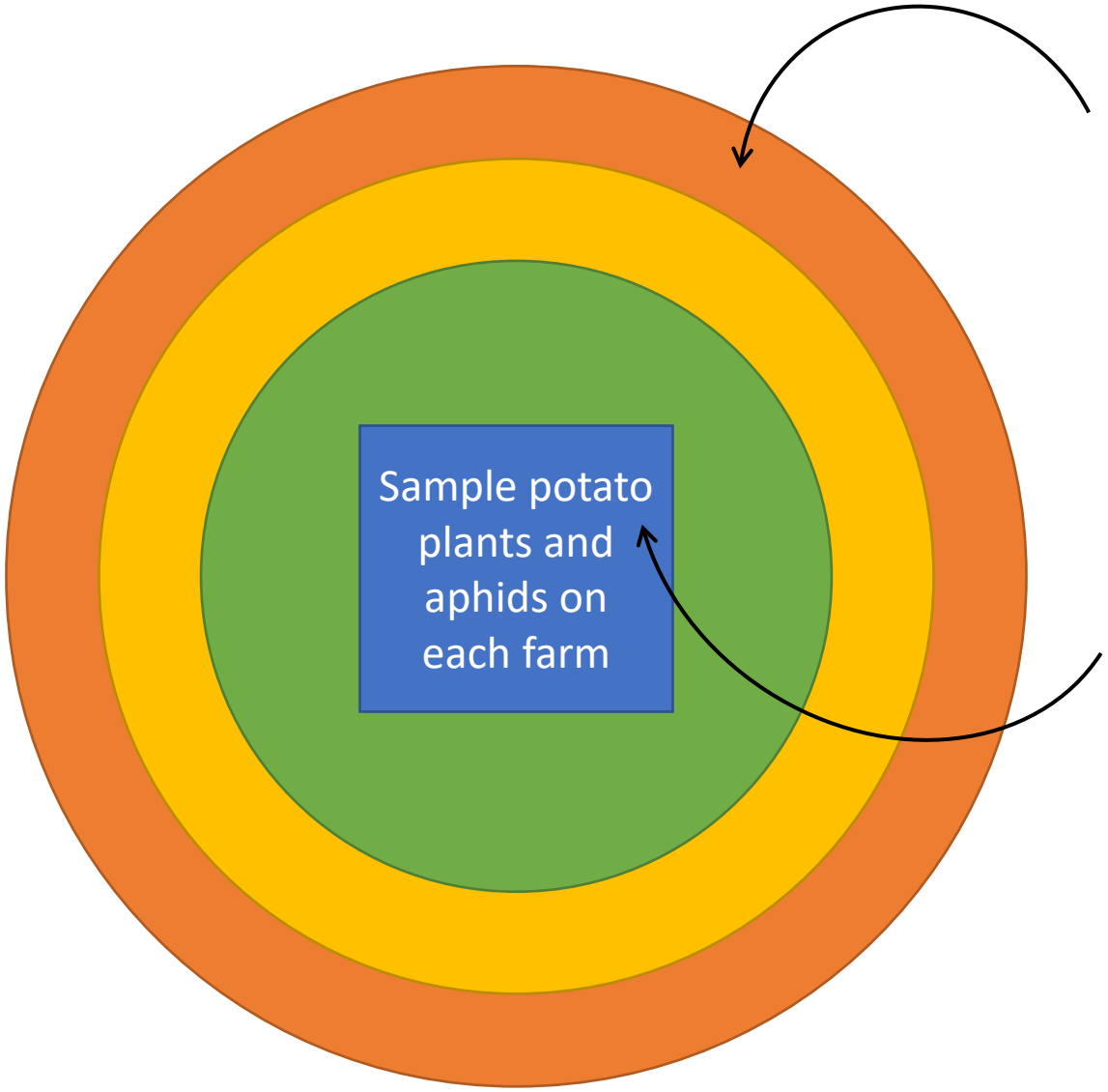
- Assumes no random/error-generated zeros

Example: potato virus Y prevalence

- Causes potato tuber necrotic ringspot disease
- Vectored by aphids
- Claflin et al. 2016: What is the role of landscape composition and aphid abundance on PVY prevalence?

(Cornell Extension photos)

Experimental setup



Use % agricultural land at various radii from farm to quantify landscape composition at different scales as predictor

Prevalence: proportion of plants sampled that tested positive for PVY

To R

Example stats methods section

Sequential models for over-dispersed, zero-inflated data

Because the data from our study are zero-inflated (characterized by excess zeros), *we employ a two-step modelling process for our statistical analysis*. Excess zeros are assumed to be generated by a process separate from the process responsible for non-zero count values, and can be modelled independently (Zuur et al. 2009).

We assume that Bernoulli probability governs the binary outcome of whether a count variate has a zero (no observed infection = 0) or a positive (observed infection = 1) outcome. In two-step models, it is typically then assumed that the positive count data are governed either by a Poisson process or by a Binomial (success/failure) process. As our data are both zero-inflated and over-dispersed, we cannot assume a Poisson process for the count data, and we model it as binomial success/failure.

The two-step modelling therefore *employs an initial model for the binary 'presence' vs. 'absence' of disease*, and includes information on presence and absence of disease from both infected and non-infected farms, thus the entire data set. This step shows that, as expected, there was no relationship between landscape parameters and the presence or absence of infection (Table S2). *The second step utilizes a reduced data set including only positive count data from infected farms in a binomial model with disease 'successes' (number of infected samples) vs. 'failures' (number of uninfected samples) regressed against landscape parameters*. In both 2012 and 2013, eight farms with no observed PVY were excluded from the reduced data set. We also performed success/failure analysis on the full data set and get consistent, though less significant results (see Table S4).

Claflin et al. 2016

Results

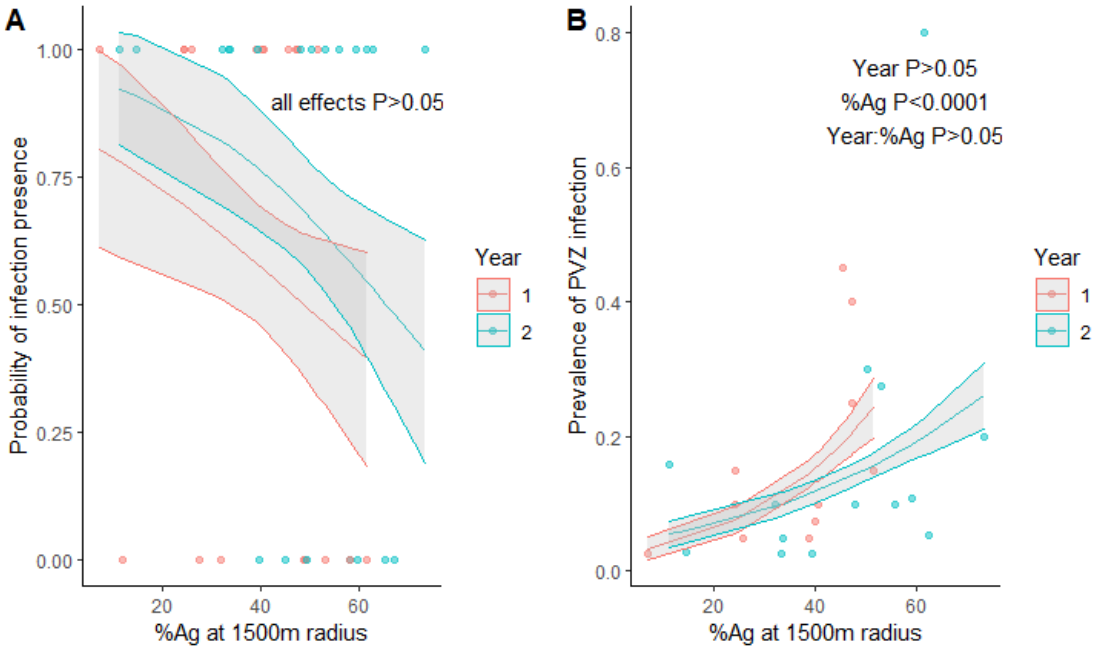


Fig. 1 Data and fitted hurdle models for (A) probability of infection presence and (B) prevalence of infection given its presence depending on year and % agriculture on landscape in 1500m radius. Model statistics are given in panel. Points show raw data, and line and ribbons show fitted model with SE.

Results

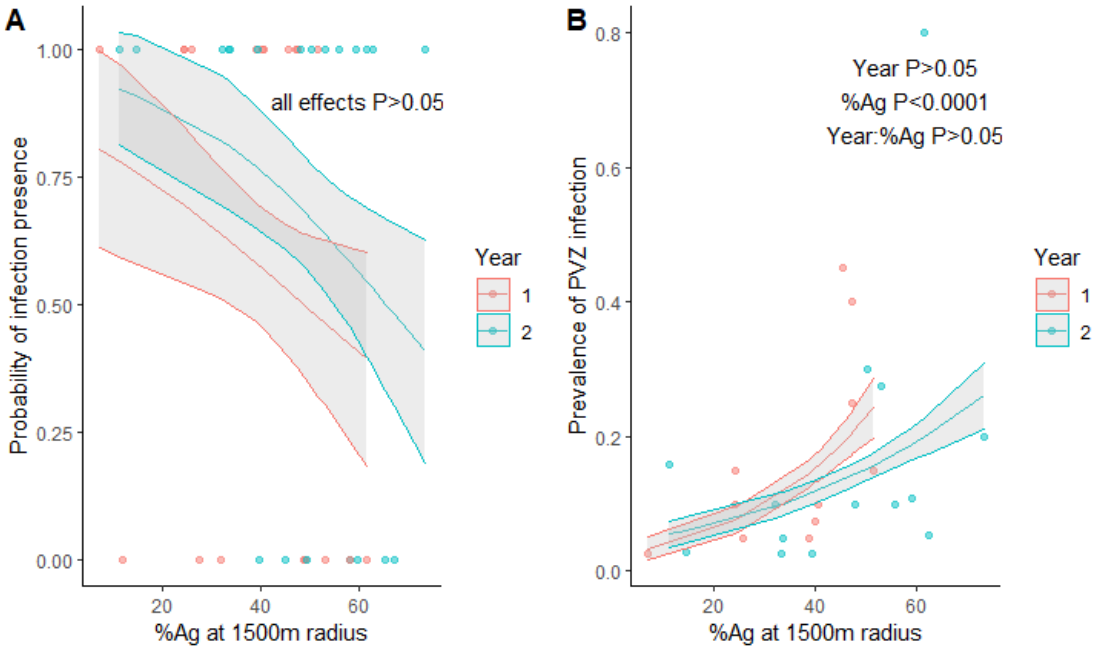


Fig. 1 Data and fitted hurdle models for (A) probability of infection presence and (B) prevalence of infection given its presence depending on year and % agriculture on landscape in 1500m radius. Model statistics are given in panel. Points show raw data, and line and ribbons show fitted model with SE.

Using a two-step hurdle model, we found that the presence of PVY infection did not depend on year, the percentage of agricultural land use at 1500m scale, or their interaction (all P>0.05, model pseudo R^2=0.06). However, we found that PVY prevalence in farms where disease was present significantly increased with the percentage of agricultural land use at 1500m scale (X^2 test on 1 and 22 *df*, P=1.27e-06), but did not differ between years (year main effect and interaction P>0.05; model pseudo R^2=0.22; see Fig. 1).