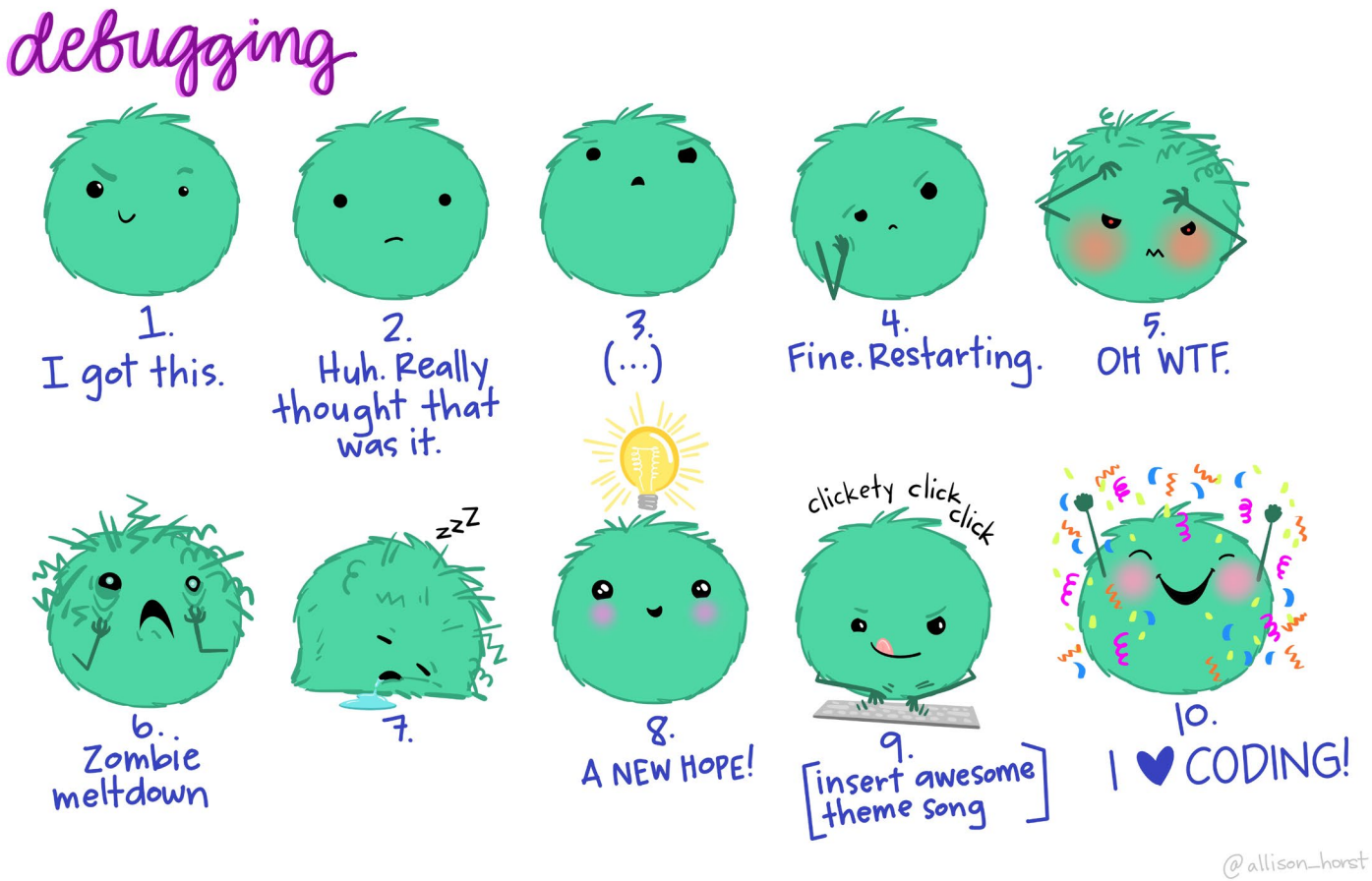
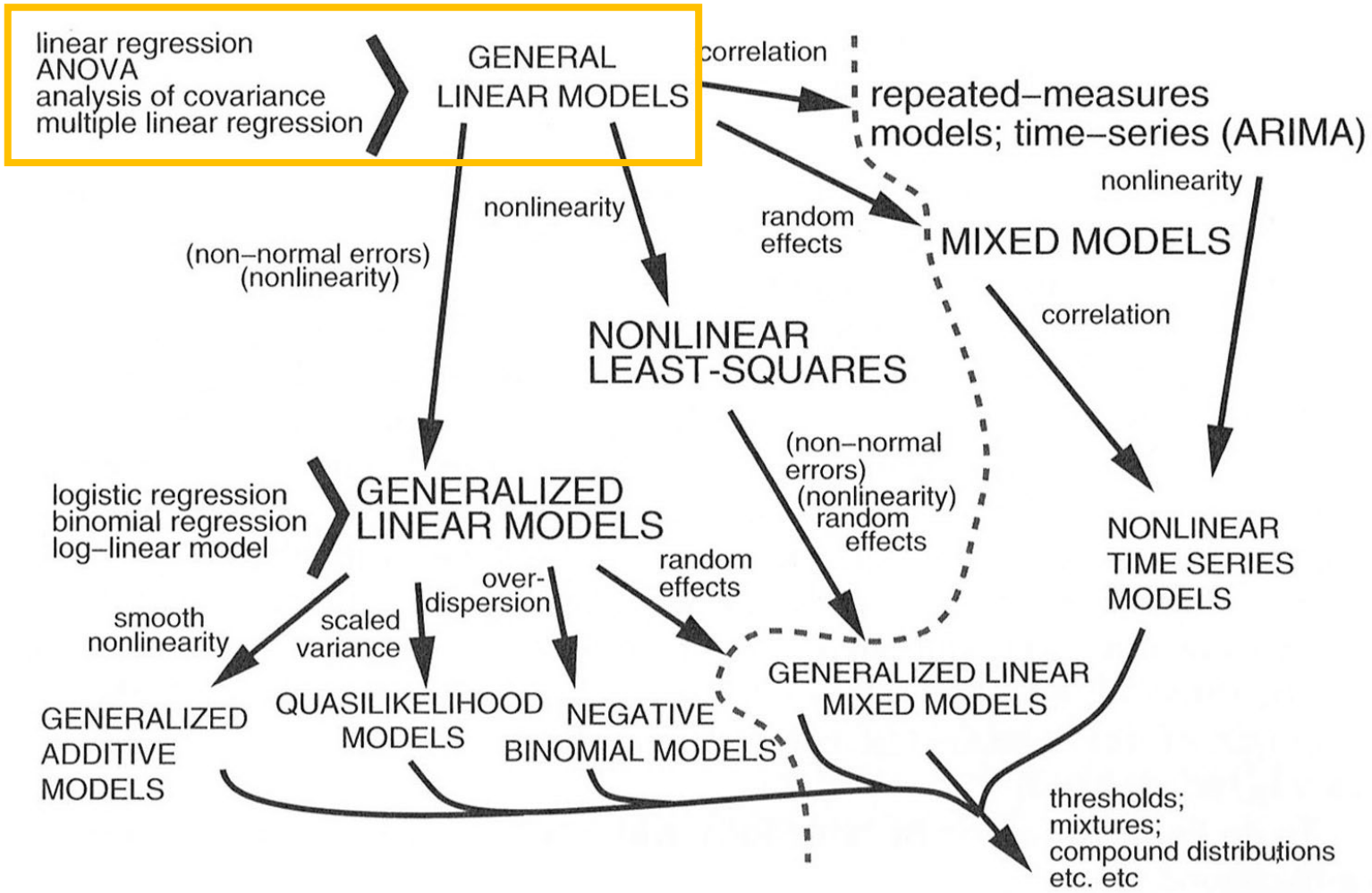
Questions from last week?

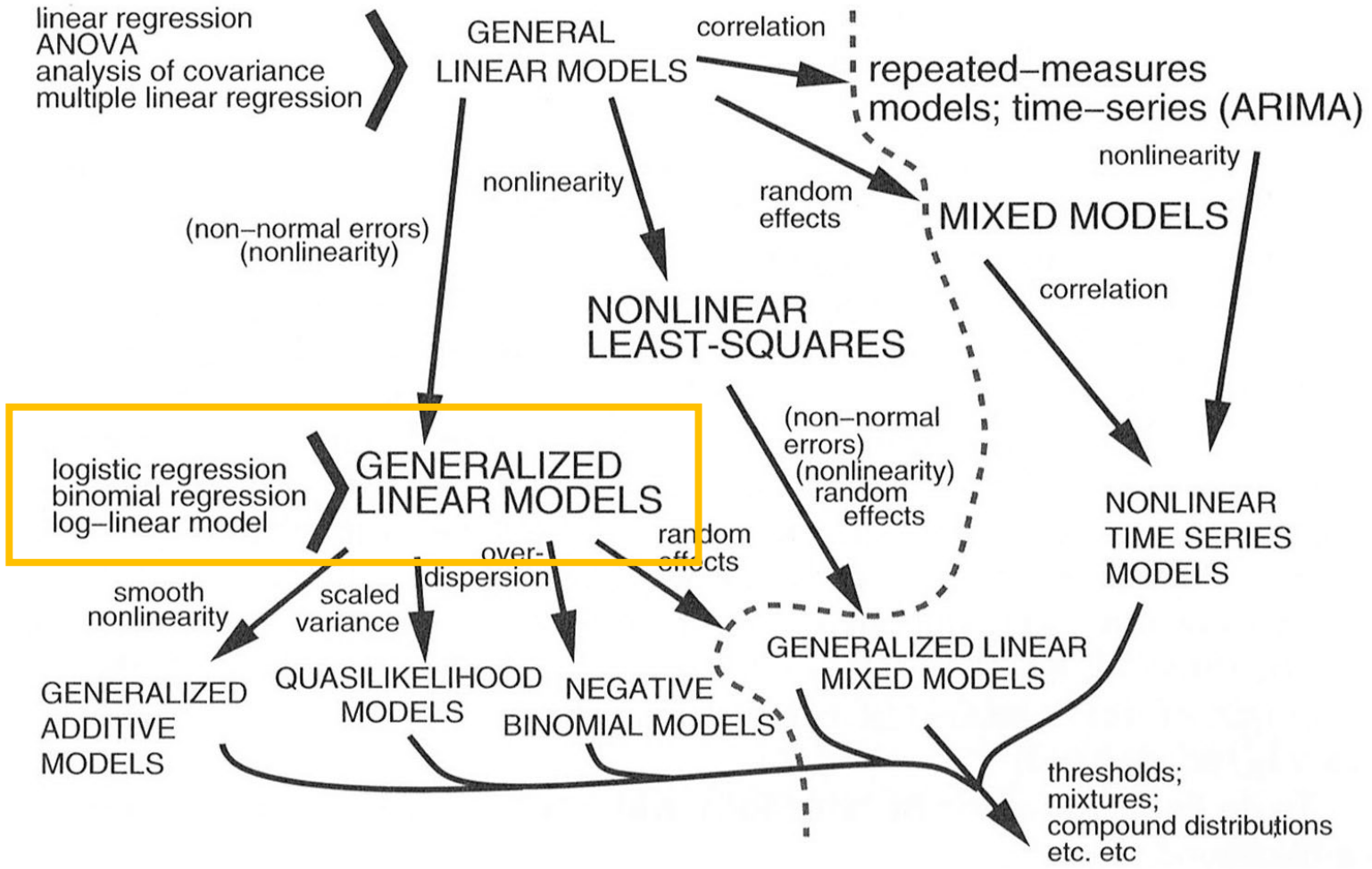
- Statistical interactions
- Introducing alternative methods to do multiple comparisons and control for multiple testing.
- Messier data, more complex models, how to interpret results and visualize them?
  - Violations of normality → data transformation
  - Unbalanced designs



A road map for the next ~10 weeks



A road map for the next ~10 weeks



This 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

Recall: (general) linear model

$$One\ response\ variable = one\ or\ more\ linear\ combinations\ of\ predictor\ variables + error$$

For any given value of the predictor, the sampled response are independent with normally distributed errors that have equal variance (homoscedasticity)

The “error” is the unmeasurable amount of deviation of the observed value from the true value

The “residual” is an observable estimate of the unobservable statistical error, based on the sample

Generalized linear models

- GLMs “generalize” the (general) linear models
- Fit *nonlinear relationships* that have a *linearizing transformation*
- That transformation is called a *link function*
- For example: log is the link function for a poisson relationship

What types of data can you think of that may not have a normal (Gaussian) distribution?

Generalized linear models architecture

1. A statistical distribution used to describe random variation in the response. This is the stochastic part of the model.
2. A linear predictor for the expected value of the response. This is the deterministic part of the model.
3. A “link function” that converts the expected value of the response to the appropriate scale.

$$link\ fn(One\ response\ variable) = one\ or\ more\ linear\ combinations\ of\ predictor\ variables + error$$

Exponential dispersion model family of distributions

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right).$$

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Exponential dispersion model family of distributions

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right).$$

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$
Gamma				
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$
Categorical	integer: $[0, K)$	outcome of single K-way occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1			
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences		

Examples

Mass, height, length

Population growth

Number of individuals in a plot

Germination, survival, infection status of a single individual/trial
As above, for multiple individuals/trials i.e. proportions

Susceptible/infected/recovered

Case study: fitting vital rate functions in a population demography study

- A common technique in population modeling is to use size or age-based vital rate probabilities
- From year *t* to year *t+1* what is the probability that an individual…
  - Reproduces (bernoulli/binomial)
    - How many offspring (fertility)? (poisson)
  - Survives (bernoulli/binomial)
  - Grows (normal)
- That probability also differs by individual size/age

Cholla cactus demography by Tom Miller and lab members

Let’s start somewhere familiar: growth

Growth model:

$$Size_{t+1,i} = \beta_0 + \beta_1 Size_{t,i} + \varepsilon_i$$

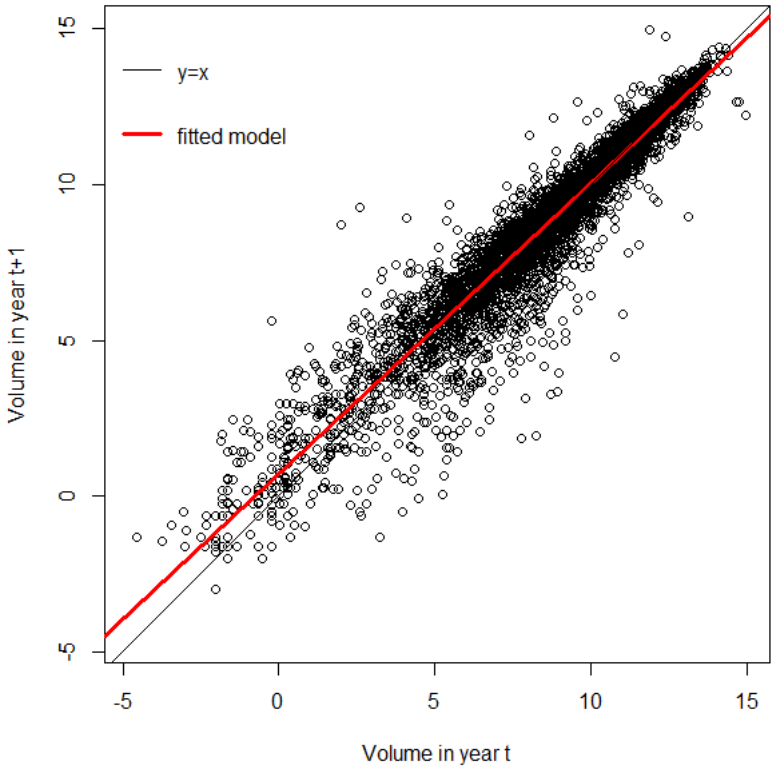
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

To R!

Growth

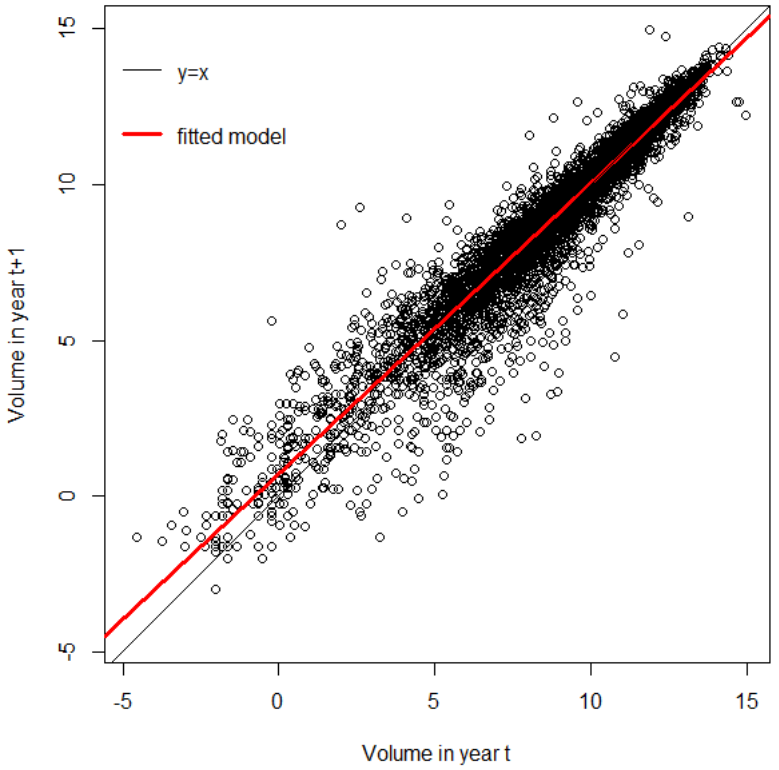
$$Size_{t+1,i} = \beta_0 + \beta_1 Size_{t,i} + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

```
## 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
```



How to report results?

- Cholla size in year t+1 was well-predicted by size in year t ($F_{1,5661}$ = 6.1e+04, P < 2.2e-16, adj. R^2 = 0.92).
- Fitted intercept was 0.71 (t = 19.87, df = 5661, P < 2.2e-16) slope was 0.93 (t = 247.69, df = 5661, P < 2.2e-16), indicating that smaller plants were more likely to increase in size compared to larger plants.



What about survival?

Survival function:

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

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

p = probability of success

(in our case, is what we are trying to fit, which is $Surv_{t+1}$)

What about survival?

Survival model:

Statistical distribution

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

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

p = probability of success

(in our case, is what we are trying to fit, which is $Surv_{t+1}$)

What about survival?

Linear predictor
(deterministic/systematic portion)

Survival model:

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

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

p = probability of success

(in our case, is what we are trying to fit, which is $Surv_{t+1}$)

What about survival?

Survival model:

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

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

Link function

p = probability of success

(in our case, is what we are trying to fit, which is $Surv_{t+1}$)

What about survival?

Survival model:

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

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

p = probability of success

(in our case, is what we are trying to fit, which is $Surv_{t+1}$)

To R!

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
- Constant dispersion (no overdispersion)
- Independent observations

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

Survival

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

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:

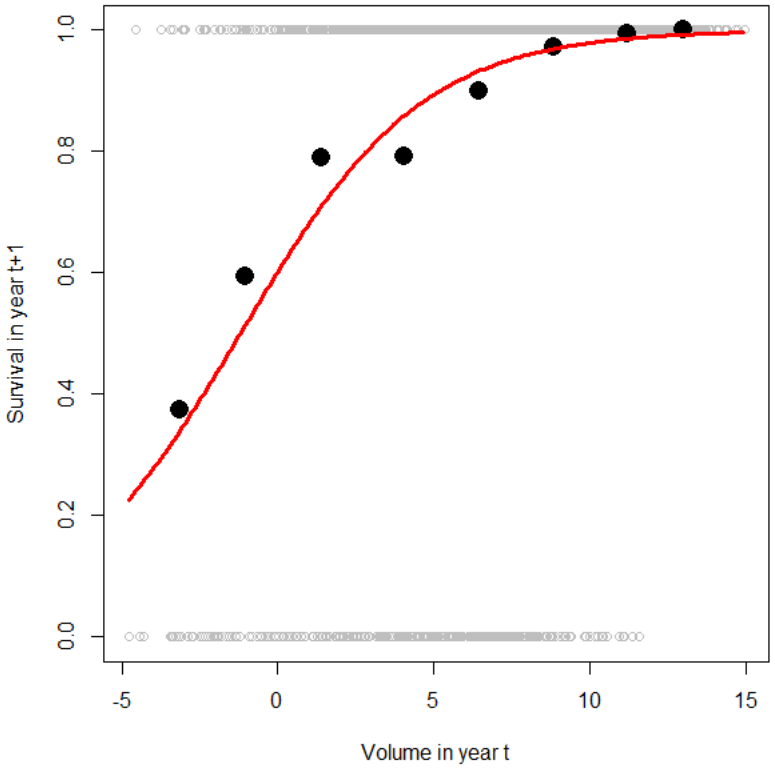
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.41045	0.09214	4.455	8.4e-06 ***
log(vol_t)	0.34344	0.01451	23.672	< 2e-16 ***

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



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

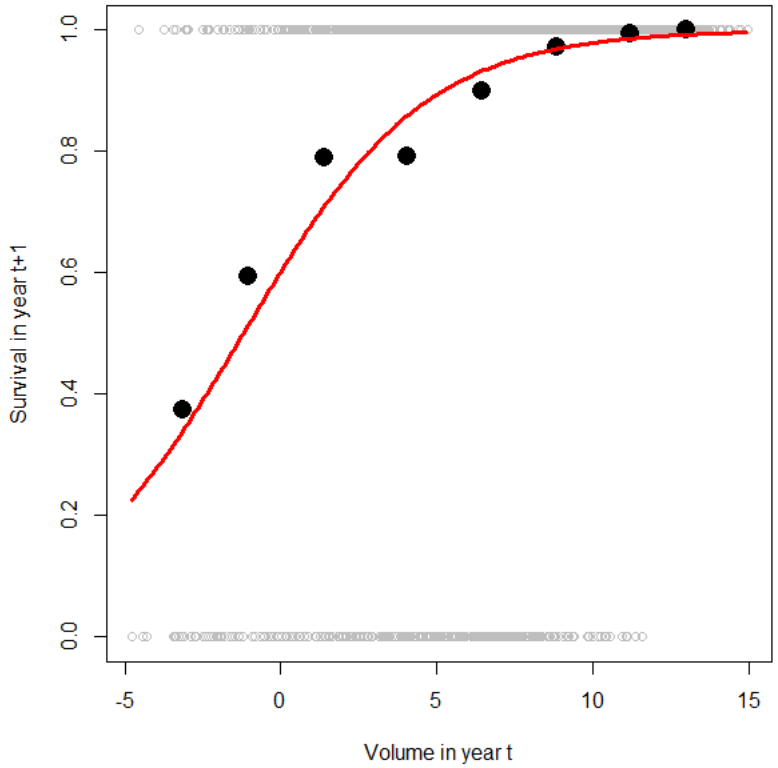
(Dispersion parameter for binomial family taken to be 1)

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

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

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

Deviance: kind of like maximum likelihood versions of "errors" in lm

- Null deviance is similar to total sums of squares (SS_Y)
- Residual deviance is similar to residual sums of squares (RSS)

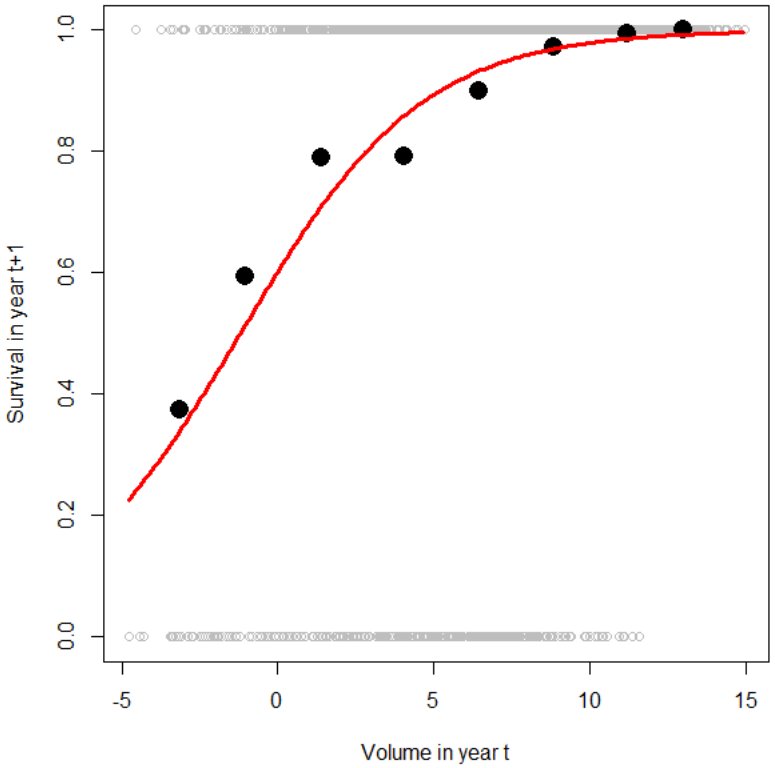
There is no R^2 for GLM, but the closest analog is "explained deviance" or pseudo R^2 =

$$1 - \frac{residual\ deviance}{null\ deviance}$$

In this case = 1- (2194.1/2833.2)=0.2256

How to report results?

- Cholla survival in year t+1 depended on size in year t (pseudo-R^2 = 0.22).
- Fitted intercept was 0.41 (z = 4.46, df = 6045, P = 8.4e-06) slope was 0.93 (z = 23.67, df = 6045, P < 2e-16), indicating that probability of survival increased with size.



break

Try fitting reproduction (whether it produced flower buds) based on size for lab this week

Reproduction function:

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

p = probability of success

(in our case, is what we are trying to fit, which is $Repro_{t+1}$)

Fertility

Fertility model:

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

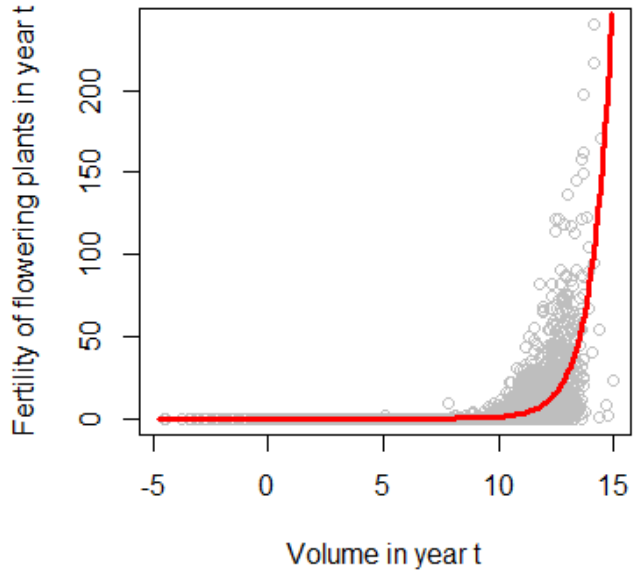
p = counts
(we are using number of buds to represent fertility)

To R!

Fertility

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

```
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
```



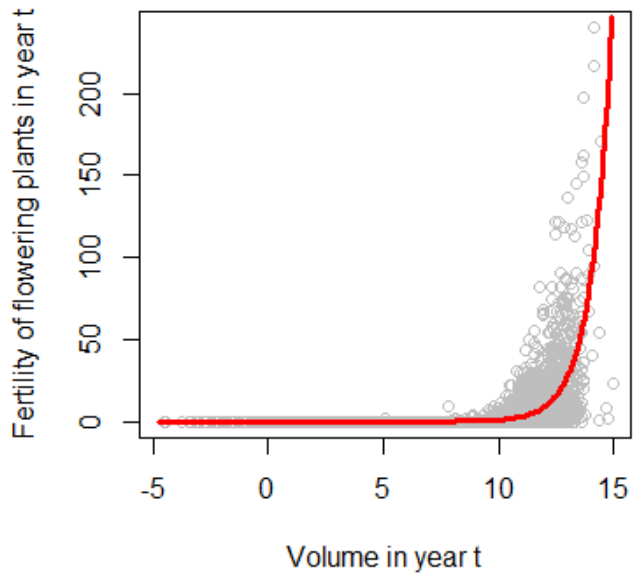
Fertility

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

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

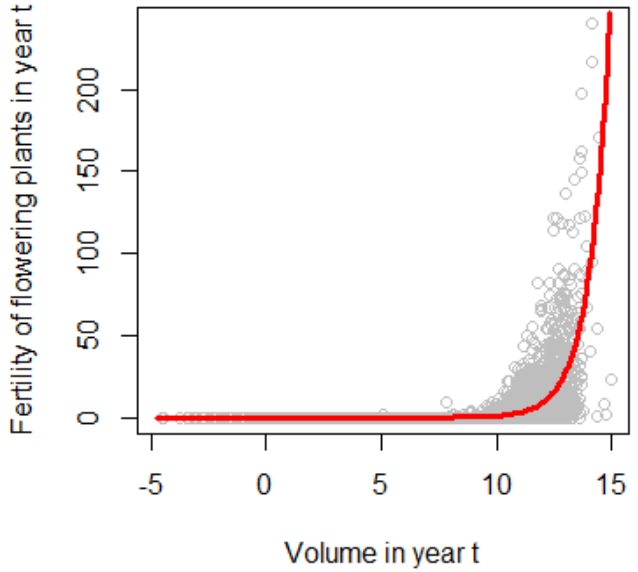
```
Call:
glm(formula = Goodbuds_t ~ log(vol_t), family = "poisson", data = cholla)
Deviance Residuals:
Min 1Q Median 3Q Max -19.4929 -1.3081 -0.4177 -0.0513 17.2248

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



How to report results?

- Cholla fertility in year t depended on size (pseudo-R^2 = 0.66).
- Fitted intercept was -11.55 (z = -153.1, df = 6088, P < 2e-16) slope was 1.14 (z = 188.8, df = 6088, P < 2e-16), indicating that number of potential offspring as measured by flower bud production increased with size.



Hold up, why don't we just log-transform the data and fit lm? Didn't we do that last week?

- Is there a difference?

$$lm(\log(response) \sim predictors)\ vs.\ glm(response \sim predictors, family = poisson)$$

- Is one better than the other? How do you decide?

Other options for overdispersed residuals

- Negative binomial
- Quasipoisson

- Next week we will talk more about zero-inflation
  - Hurdle models
  - Reproduction example in our case study: fit one model for reproduction (yes/no), and if reproduces, fit another model for how many offspring (zero-truncated version of our fertility model).