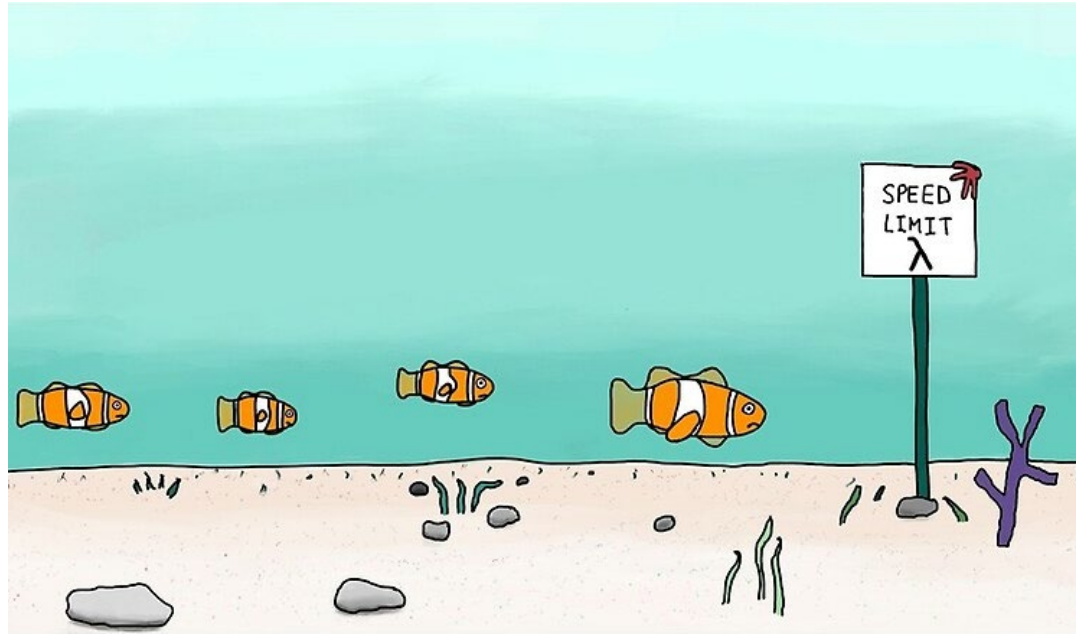


# Generalized Linear Models

## Regression methods for categorical response variables

ENTMLGY 6707 Entomological Techniques and Data Analysis





THE OHIO STATE UNIVERSITY

College of Food, Agricultural, and Environmental Science

Horticulture and Crop Science

# SYLLABUS

## HCS 7100

Data Visualization in R  
Autumn 2024 (full term)

2 credit hours

Tuesday 4:10-5:55pm and hybrid synchronous delivery

You can join class in person: Howlett 116 (Columbus) or Williams 123 (Wooster) or via Zoom (but I would recommend you come in person)

Zoom link for class: <https://go.osu.edu/dataviz-zoom> (Links to an external site.)

Full Zoom details: Meeting ID: 937 4016 0297 Password: dataviz

<https://osu.zoom.us/j/93740160297?pwd=2wPNkuaKbb2uaKFQ3jC8txf6tJIUHj.1>

Course website: <https://www.rdataviz.com/> used in conjunction with Carmen Canvas

## COURSE OVERVIEW

### Instructor

Instructor: Jessica Cooperstone

Email address: [cooperstone.1@osu.edu](mailto:cooperstone.1@osu.edu) (preferred contact method)

Phone number: 614-292-2843 (non-preferred contact method)

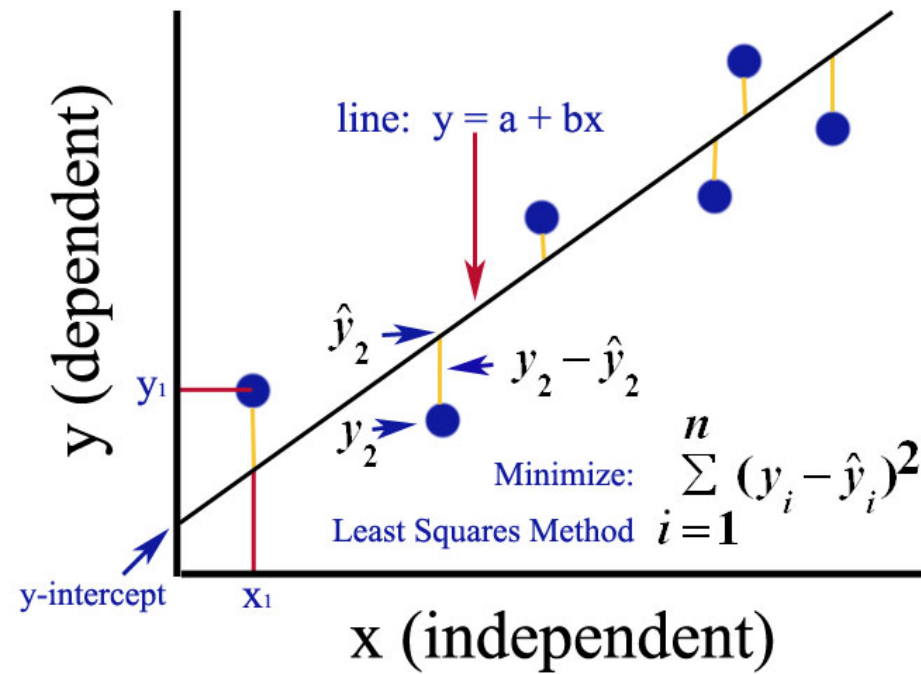
Office hours: I will hold 2, 1 hour-long virtual office hour sessions during the semester

# Learning objectives

- 1) Describe distributions used to model different types of categorical responses
- 2) State assumptions of Binomial and Poisson regressions
- 3) Understand how to fit Binomial and Poisson regressions using R
- 4) Interpret output from Binomial and Poisson regressions

# Regression analysis

Used to describe a relationship between a response variable and one or more predictor variables



# A quick review of simple linear regression

Assumptions of simple linear regression (a.k.a. general linear models)

- Model is correct
- Linearity of predictors  $E(Y|X) = \beta_0 + x_1\beta_1 + \dots + x_n\beta_n$
- Errors are independent and normally distributed
- Variance is constant (homoscedastic)

In simple linear regression, the response variable is continuous and may take on values from  $(-\infty, \infty)$ .

Interpretation of slope: a one unit change in  $X$  is associated with a  $\beta_1$  change in the mean of  $Y$ .

# Discrete responses

Historically, some types of discrete data have often been transformed to meet assumptions of normality.

With increases in computational power, we are now able to fit **generalized linear models** with statistical software.

We are no longer using ordinary least squares to estimate model coefficients. We are using maximum likelihood estimation.

# Activity

Here is a broad research question:

How do ladybeetle characteristics (e.g., size, species, age) affect their efficacy as predators?

Identify three **response variables** (one continuous, one binary, and one count) you might measure to begin answering this question.

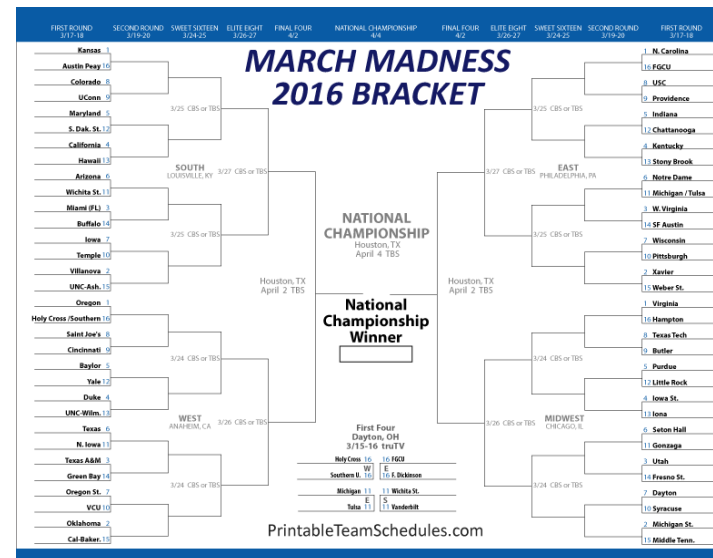
```
fit1 <- lm(Response_variable1 ~ beetle_character, data=df)
fit2 <- glm(Response_variable2 ~ beetle_character, data=df, family=binomial(link=logit))
fit3 <- glm(Response_variable3 ~ beetle_character, data=df, family=poisson(link=log))
```

# Binary data

A single binary variable can only take one of two, mutually exclusive forms (usually a 0 or 1)

- Presence of a disease
- Mortality (dead, alive)
- Success/failure
- Binning a variable into two groups (i.e. number of spores → presence/absence)

Win or lose?





# Distributions used to model binary data

- Bernoulli
- Binomial (a collection of Bernoulli trials)

Probability mass function:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

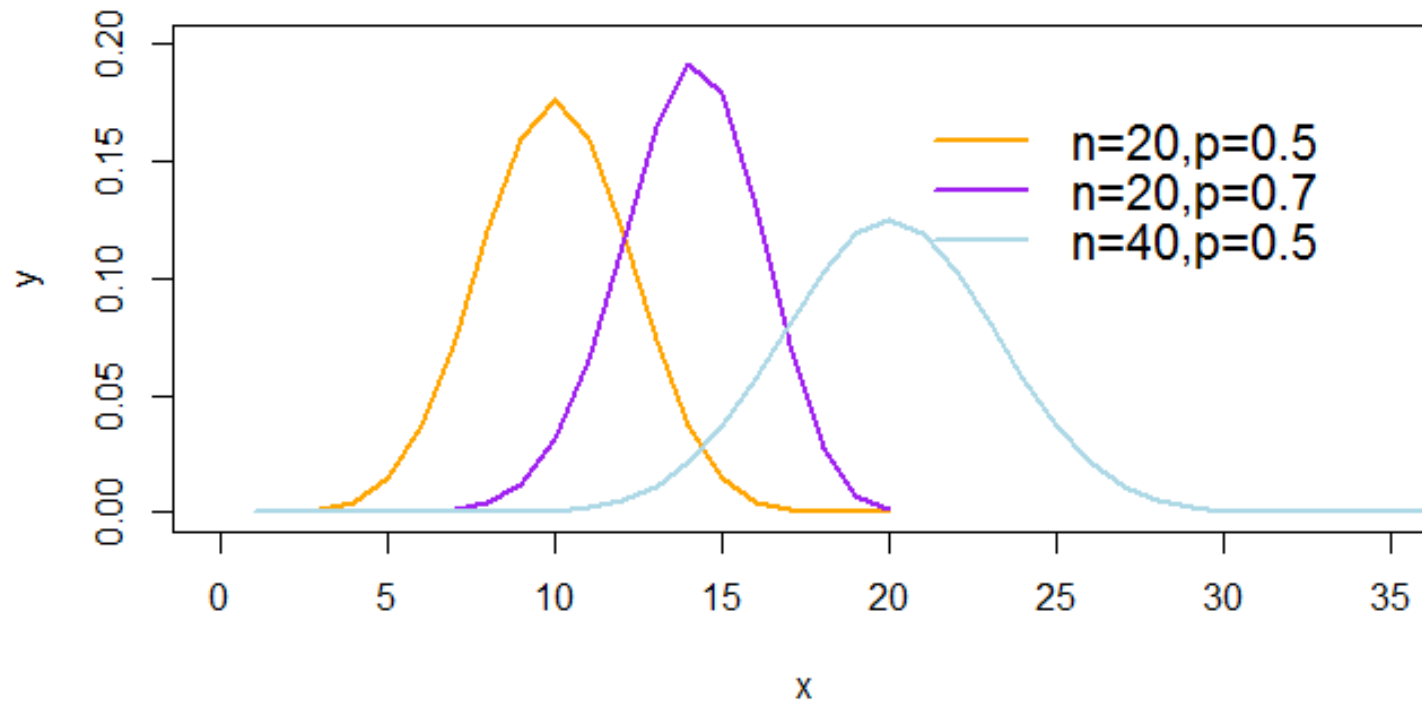
$$\text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

# Binomial distribution

- Mean:  $np$
- Variance:  $np(1 - p)$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$



# Count data

Variable of interest is measured/recorded as an integer (i.e. 1, 2, 3, 4...,n)

- Woodpeckers visiting a tree per day (rate)
- Mosquitoes caught in a trap



# Distributions used to model count data

Siméon Denis Poisson  
1781-1840



Probability mass function:

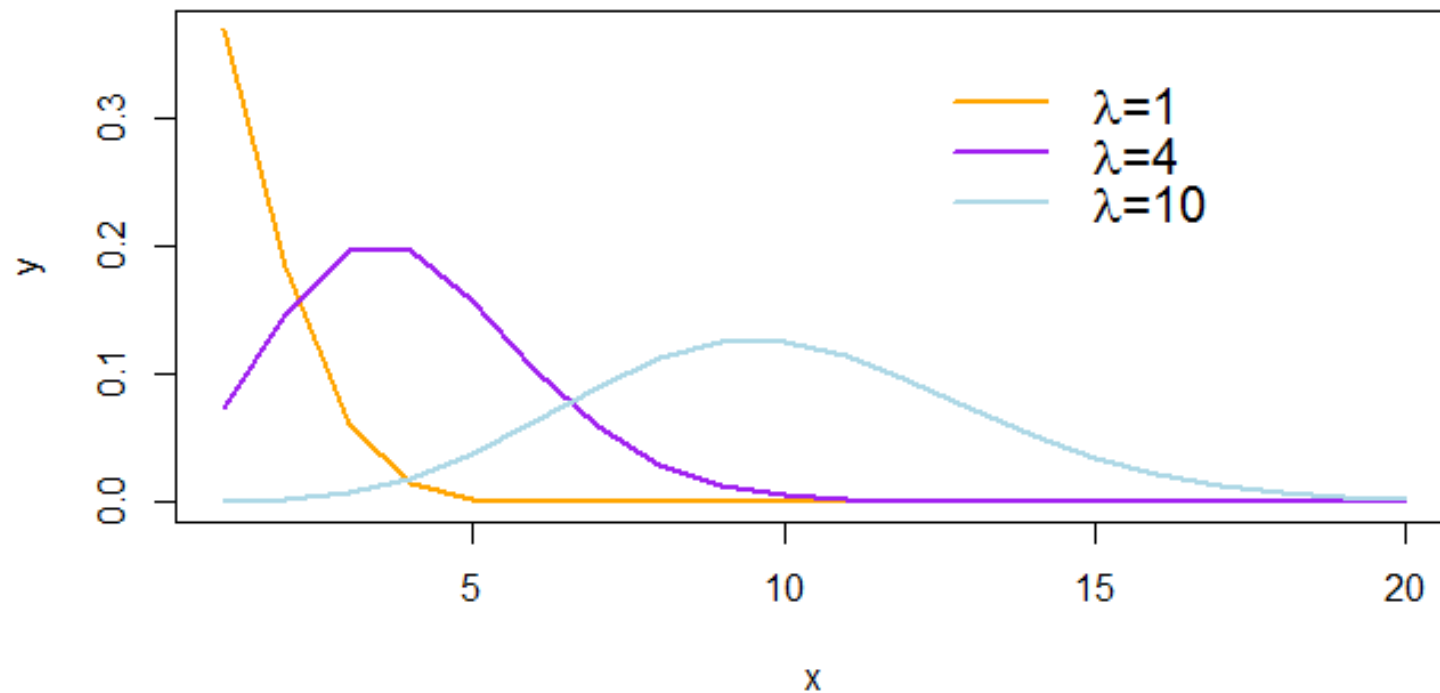
$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

# Poisson distribution

Mean:  $\lambda$

Variance:  $\lambda$

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$



# General vs. Generalized

**General linear models:** Refers to simple linear regressions, such as modeling a continuous response variable as a function of continuous and/or categorical predictors; includes ANOVA and ANCOVA.

```
> lm(Y~X, data=my_data)
```

**Generalized linear models:** Includes all general linear models, but also includes models in which we assume the residuals are non-normal (e.g., logistic regression, Poisson regression).

```
> glm(Y~X, data=my_data, family=binomial(link=logit))
```

```
> glm(Y~X, data=my_data, family=poisson(link=log))
```

# Flexibility



```
glmer(YesNo ~ X + (1|Group),  
data = df,  
family = binomial(logit))
```



```
glmer(Count ~ X + (1|Group),  
data = df,  
family = poisson(log))
```

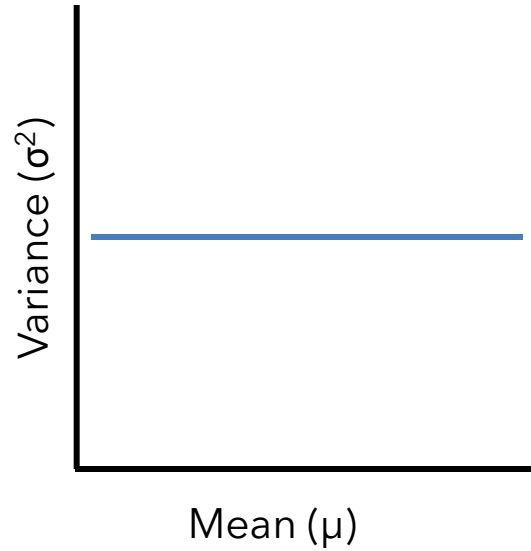
# Compare to Normal distribution

Distribution	Mean	Variance
Normal	$\mu$	$\sigma^2$
Binomial	$np$	$np(1-p)$
Poisson	$\lambda$	$\lambda$

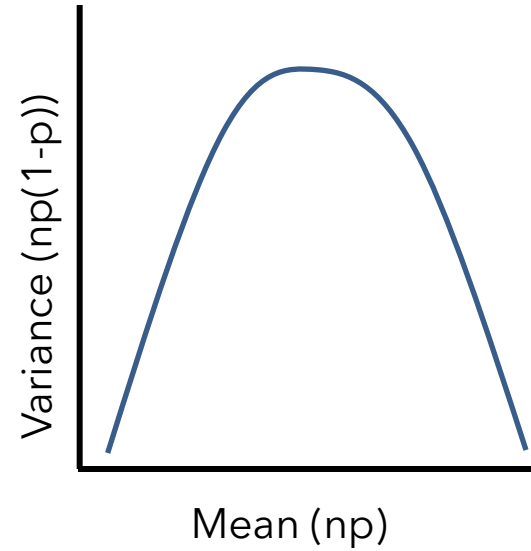
What is different about the last two distributions?



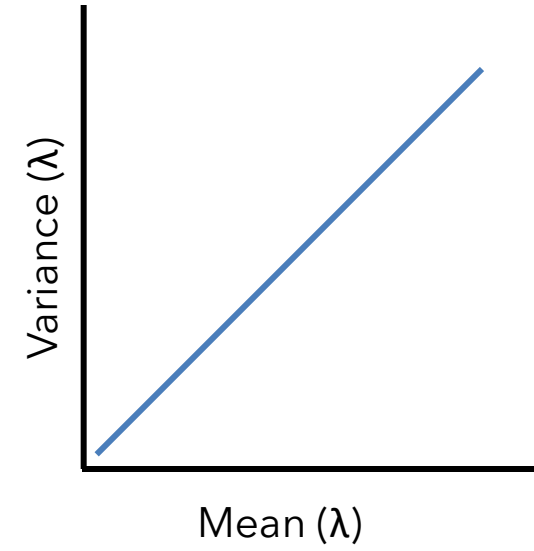
# Compare to Normal distribution



Normal



Binomial



Poisson

# Decisions to make when fitting a GLM

- 1) Determine the most likely distribution of your response variable (e.g., continuous (normal) vs. **binary (Binomial)** vs. **count (Poisson)**). If **Binomial** or **Poisson**, consider fitting a GLM.
- 2) Identify predictors ( $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ ).
- 3) Select a **link function**, often represented using  $\eta$  or  $g(\mu)$ . Link functions define how your *expected* response,  $E(X)$ , relates to your set of linear predictors,  $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ .

$$g(f(x)) = \beta_0 + x_1 \beta_1$$

## So...what are GLMs really doing?

- Allowing more appropriate specification of the structure of the errors
- Estimating the mean and variance separately

$$g(f(x)) = \beta_0 + x_1\beta_1$$

# Link functions

- A function that links our predictors to the mean of the response variable. A key ingredient for GLMs.
- Canonical ("most mathematically suitable") link functions:
  - Normal  $\rightarrow$  Identity
  - Binomial  $\rightarrow$  Logit (Logistic regression)
  - Poisson  $\rightarrow$  Log

Other links functions: probit, inverse, cloglog

# Link functions

Be aware of which link function you are using (this will come up later when we look at some examples), as the interpretation of model coefficients changes with the link function used.

# Link functions

- Normal  $\rightarrow$  identity

$$g(f(x)) = 1(f(x)) = \beta_0 + x_1\beta_1$$

- Binomial  $\rightarrow$  logit

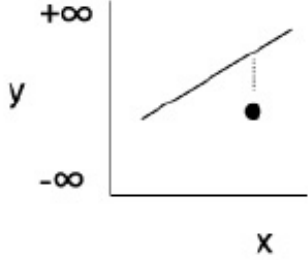
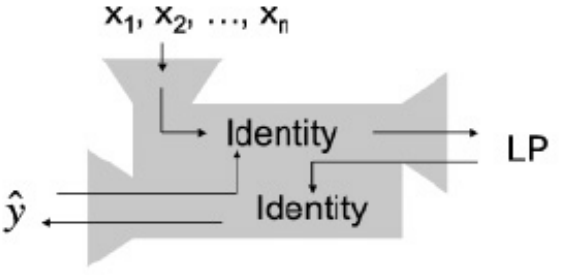
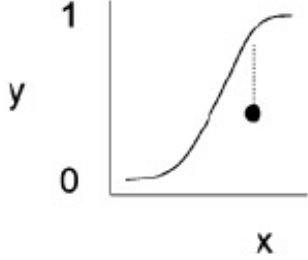
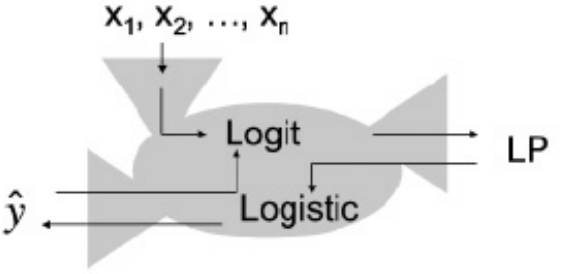
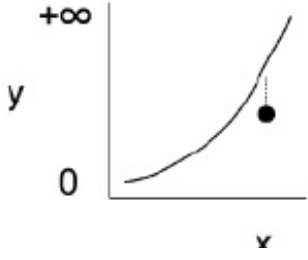
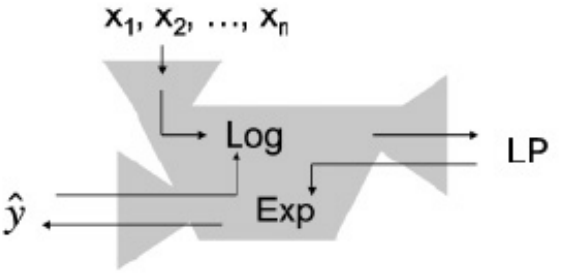
$$g(f(x)) = \log\left(\frac{p}{1-p}\right) = \beta_0 + x_1\beta_1$$

$$p = \frac{1}{1 + \exp(-(\beta_0 + x_1\beta_1))}$$

- Poisson  $\rightarrow$  log

$$g(f(x)) = \log(\mu) = \beta_0 + x_1\beta_1$$

$$\mu = \exp(\beta_0 + x_1\beta_1)$$

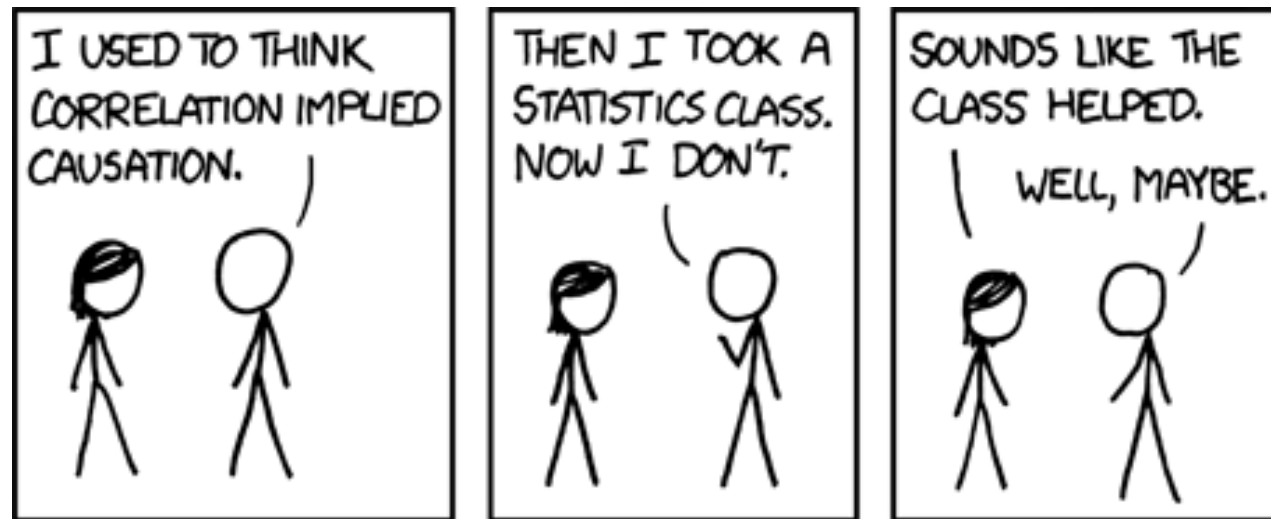
<b>Examples of Y</b>	<b>Input-output relationship</b>	<b>Error (residual) distribution</b>	<b>Link function and inverse</b>	<b>Meaning of the coefficients</b>
Left Ventricular Mass, LVM		Gaussian		Differences
Risk of a Binary Event		Binomial		Odds Ratios
Rates of a Count Event		Poisson		Rate Ratios

Ravani, Pietro & Parfrey, Patrick & Murphy, Sean & Gadag, Veeresh & Barrett, Brendan. (2008). Clinical research of kidney diseases IV: Standard regression models. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 23. 475-82. 10.1093/ndt/gfm880.

# Assumptions of *generalized* linear models (GLMs)

- Model is correct
- Errors are independent

The error structure specified is different





# Example: Binomial (logistic) regression



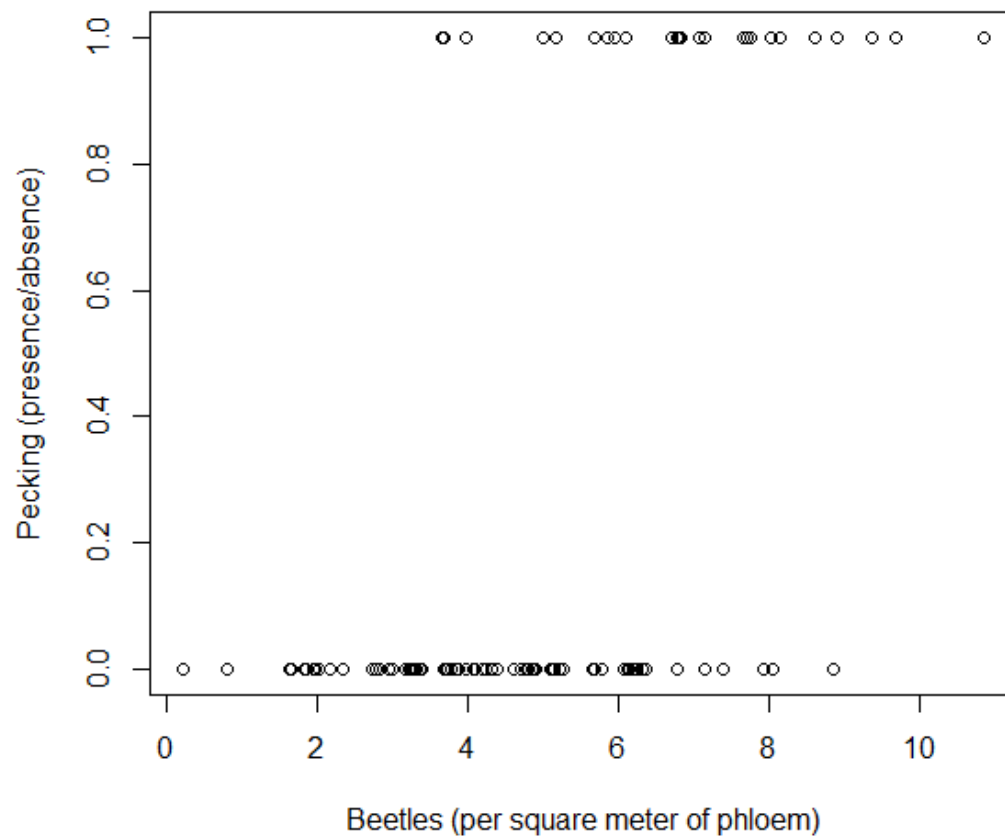
Emerald ash borer is a non-native phloem/woodboring beetle that infests ash trees.



Woodpeckers are typically the earliest responding natural enemies to new infestations of the beetle.

# The form of the data

```
> summary(tree_data)
  pecked    emerald_ash_borer
Min.   :0.00   Min.   : 0.00
1st Qu.:0.00   1st Qu.: 3.00
Median :0.00   Median : 5.00
Mean   :0.26   Mean   : 5.01
3rd Qu.:1.00   3rd Qu.: 6.00
Max.   :1.00   Max.   :11.00
```



# Model syntax in R

```
> fit1 <- glm(pecked ~ emerald_ash_borer, data=tree_data, family=binomial(link=logit))
> summary(fit1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9426	-0.6125	-0.2787	0.2918	2.2317

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.7047	1.1135	-5.123	3.0e-07 ***
emerald_ash_borer	0.8252	0.1788	4.617	3.9e-06 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 114.611 on 99 degrees of freedom  
Residual deviance: 78.544 on 98 degrees of freedom  
AIC: 82.544

Number of Fisher Scoring iterations: 5

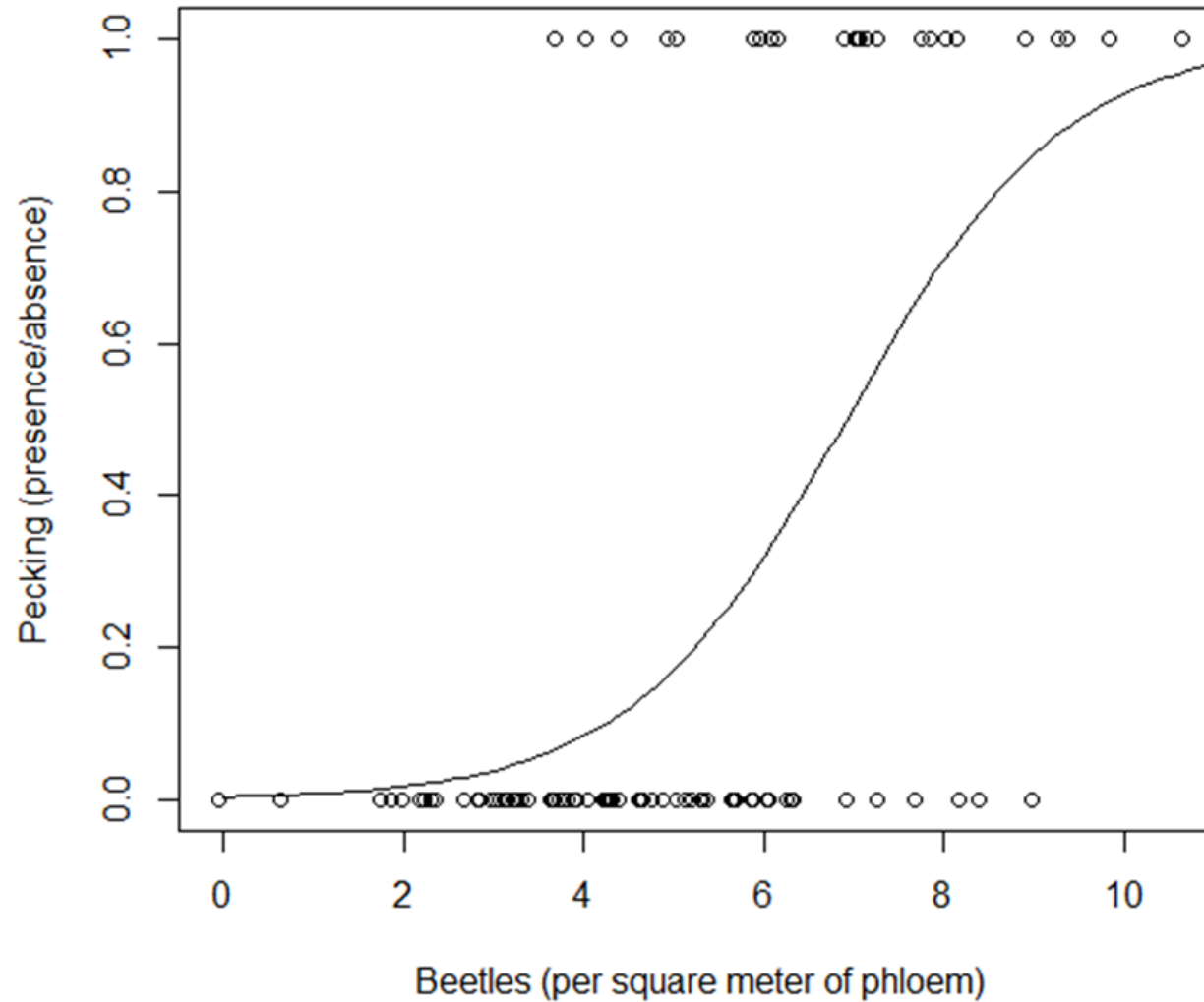
# How do we interpret the model?

When using the logit link: "The odds of finding woodpecking on a tree increase by 2.28 [=exp(0.8252)] with an increase in 1 insect per square meter of phloem."

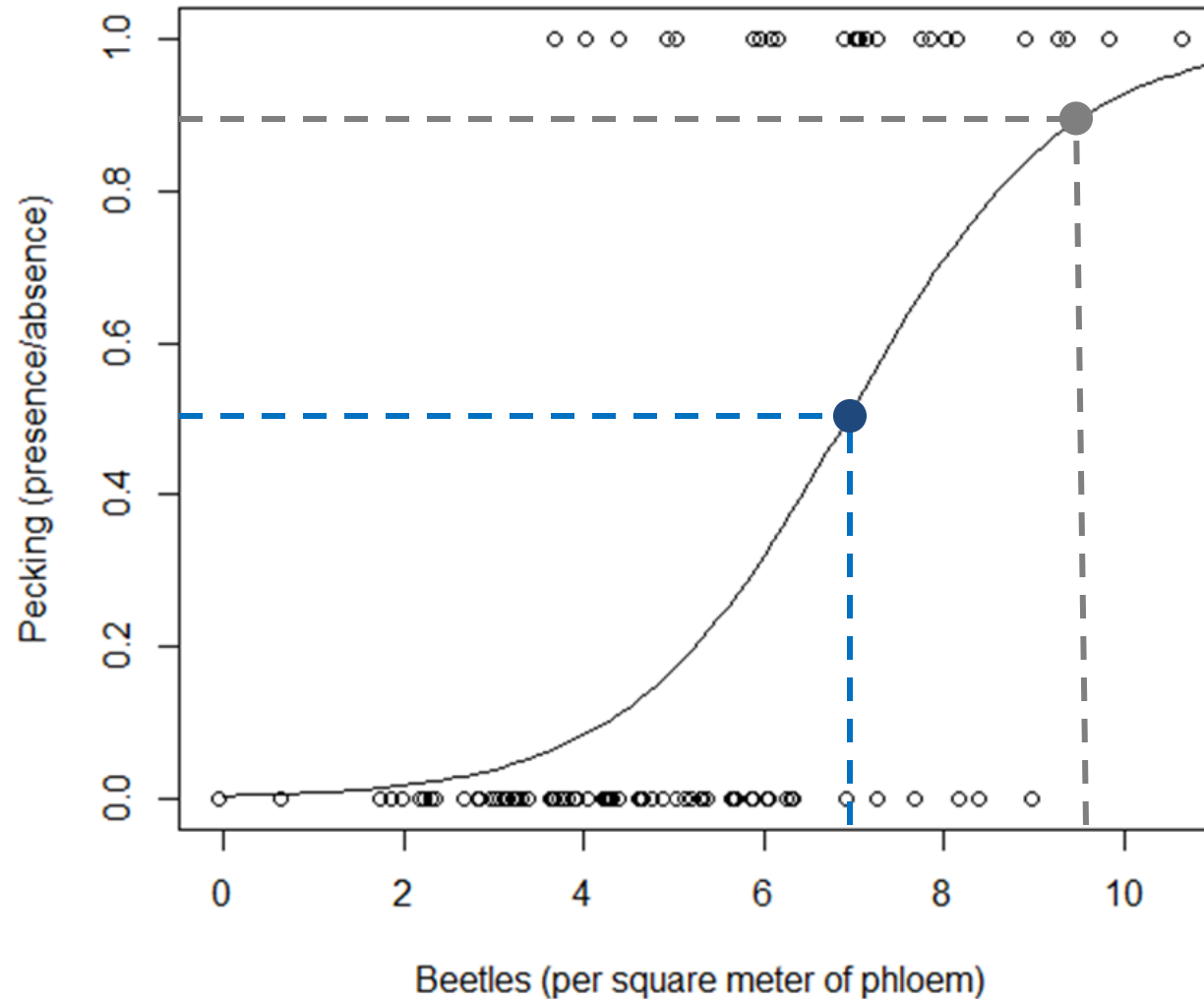
Odds ratio:

$$OR = \frac{P(Y=1|X=1)P(Y=0|X=1)}{P(Y=1|X=0)P(Y=0|X=0)}$$

P(Woodpecking) increases with densities of emerald ash borer



What density of beetles is associated with a 90% change of being pecked? What about a 50% chance?



# Model syntax in R

```
> fit1 <- glm(pecked ~ emerald_ash_borer, data=tree_data, family=binomial(link=logit))
> summary(fit1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9426	-0.6125	-0.2787	0.2918	2.2317

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.7047	1.1135	-5.123	3.0e-07	***
emerald_ash_borer	0.8252	0.1788	4.617	3.9e-06	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 114.611 on 99 degrees of freedom  
Residual deviance: 78.544 on 98 degrees of freedom  
AIC: 82.544

Number of Fisher Scoring iterations: 5

What density of beetles is associated with a 90% change of being pecked? What about a 50% chance?

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + x_1\beta_1$$

$$\log\left(\frac{0.9}{0.1}\right) = -5.7047 + x_1 * 0.8252$$

$$2.2 + 5.7047 = 0.8252 * x_1$$

$$x_1 = 9.6$$



What density of beetles is associated with a 90% change of being pecked? What about a 50% chance?

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + x_1\beta_1$$

$$\log\left(\frac{0.5}{0.5}\right) = -5.7047 + x_1 * 0.8252$$

$$5.7047/0.8252 = x_1$$

$$x_1 = 6.9$$

What density of beetles is associated with a 90% change of being pecked? What about a 50% chance?

Make R do the heavy lifting for you!

```
> library(MASS)
> dose.p(fit1, p=c(0.5,0.9))
```

	Dose	SE
p = 0.5:	6.912700	0.3937953
p = 0.9:	9.575202	0.8425177

# Activity

Interpret a logistic regression model predicting survival of individual trees as a function of defoliation intensity (unitless index).

Survival ~ defoliation

Note:  $\exp(0.096) \approx 1.10$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.860	0.185	10.05	<0.0001
defoliation	0.096	0.015	6.40	<0.0001

# Activity

Interpret a logistic regression model predicting survival of individual trees as a function of defoliation intensity (unitless index).

Survival ~ defoliation

Note:  $\exp(0.096) \approx 1.10$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.860	0.185	10.05	<0.0001
defoliation	0.096	0.015	6.40	<0.0001

**Answer:** "The odds of a tree dying increased by 1.10 with a 1 unit increase in the defoliation index."

# Example: Poisson regression

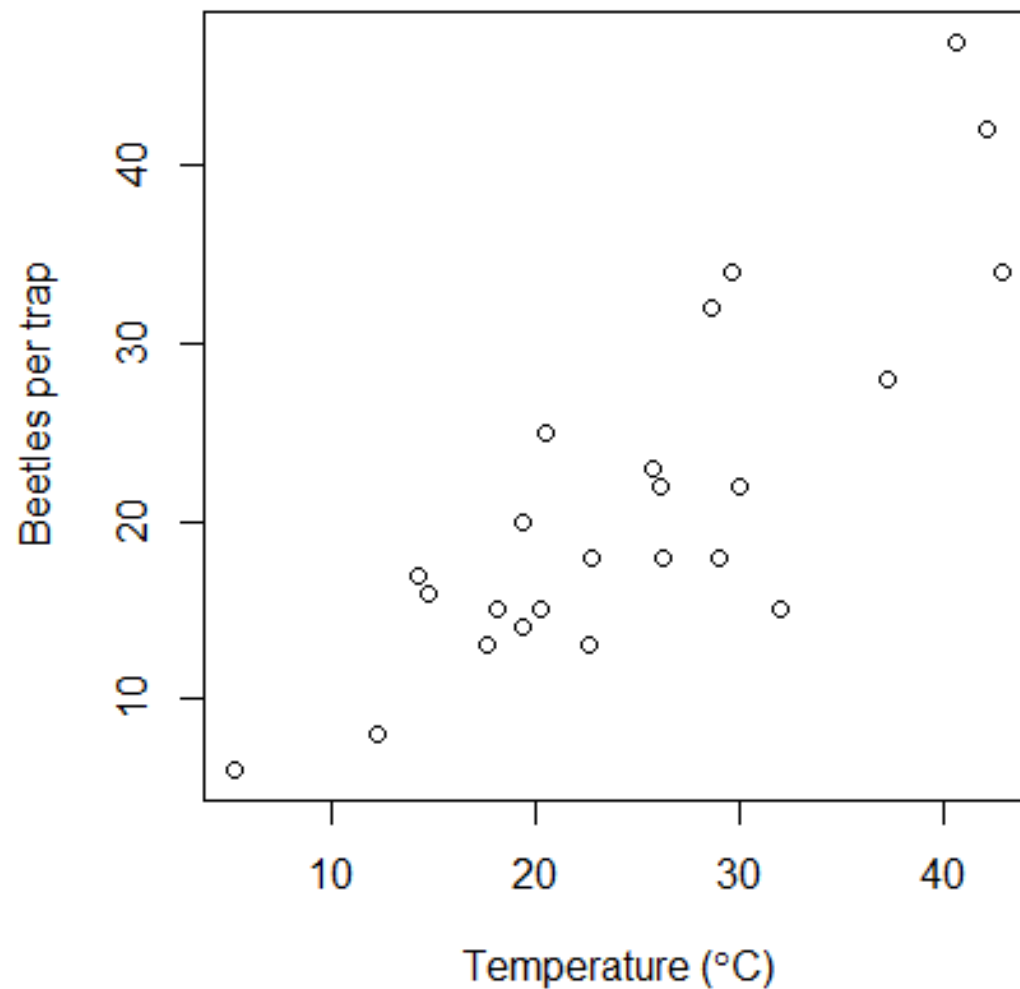


Number of eastern larch  
beetles caught in Lindgren  
funnel traps

# The form of the data

```
> summary(beetles)
```

ELB	temperature
Min. : 6.00	Min. : 5.334
1st Qu.:15.00	1st Qu.:19.079
Median :18.00	Median :24.263
Mean :21.46	Mean :24.913
3rd Qu.:25.75	3rd Qu.:29.701
Max. :47.00	Max. :42.869



# Model syntax and output in R

```
fit2 <- glm(ELB~temperature, data=beetles, family=poisson(link=log))
summary(fit2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6662	-0.8152	0.1128	0.8768	2.1676

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.869505	0.185303	10.089	< 2e-16 ***
temperature	0.045002	0.007716	5.833	5.46e-09 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 66.910 on 23 degrees of freedom  
Residual deviance: 32.277 on 22 degrees of freedom  
AIC: 147.64

Number of Fisher Scoring iterations: 4

# How do we interpret the model?

(note that  $\exp(0.0450)=1.046$ )

When using the log link...

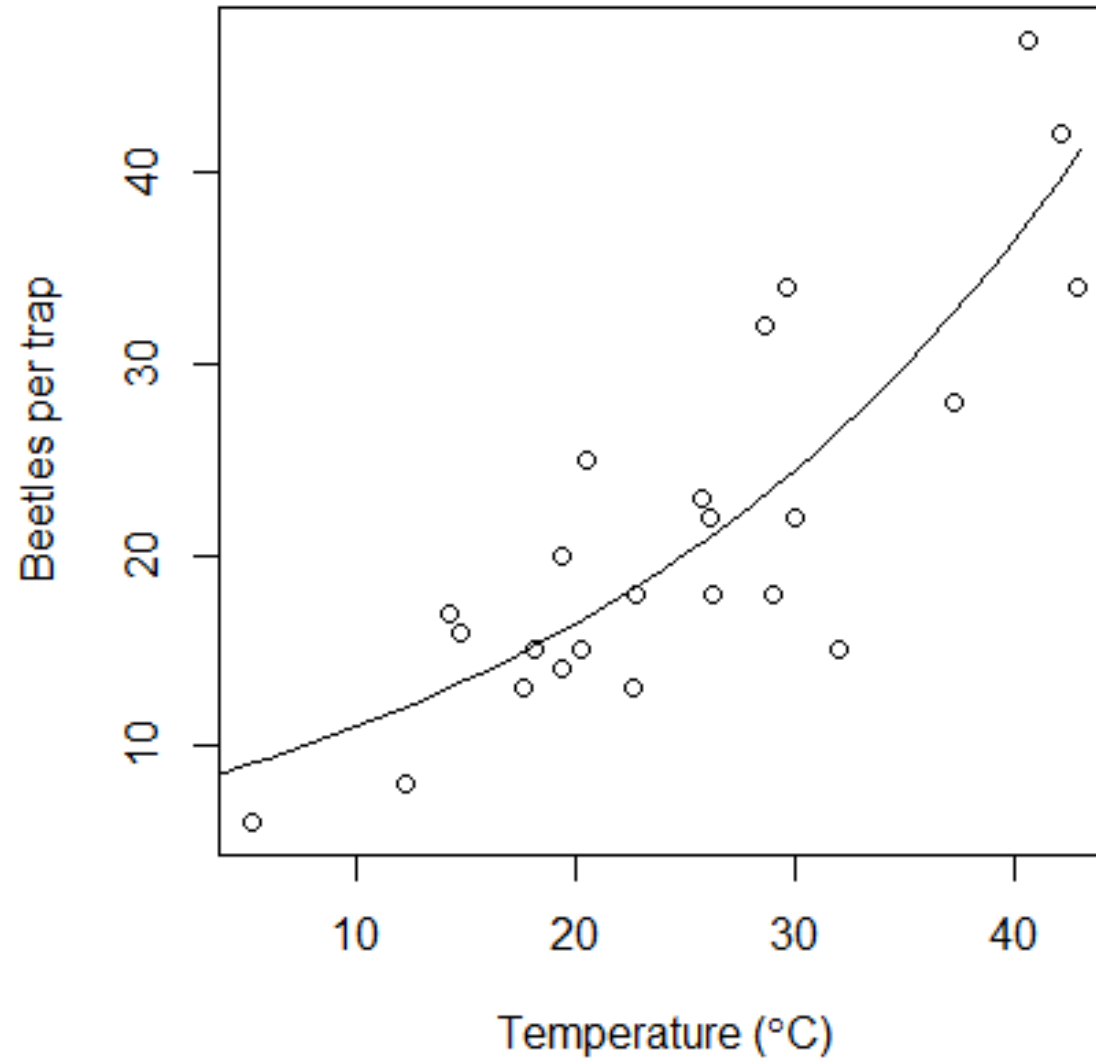
- "An increase in one degree Celsius is associated with a 4.6%  $[(1.046-1) * 100]$  increase in the number of beetles caught per trap."

OR

- "An increase in one degree Celsius is associated with a 1.046 times increase in the number of beetles caught per trap."



Trap catch increases with temperature



# Activity

Interpret a Poisson regression model predicting number of dead trees per site as a function of insect pressure (unitless index).

Number of dead trees ~ insect pressure

Note:  $\exp(0.096) \approx 1.10$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.860	0.185	10.05	<0.0001
insect_pressure	0.096	0.015	6.40	<0.0001

# Activity

Interpret a Poisson regression model predicting number of dead trees per site as a function of insect pressure (unitless index).

Number of dead trees ~ insect pressure

Note:  $\exp(0.096) \approx 1.10$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.860	0.185	10.05	<0.0001
insect_pressure	0.096	0.015	6.40	<0.0001

**Answer:** "An increase in one unit of the insect pressure index was associated with a 1.10 times (or 10%) increase in the number of beetles caught per trap."

# Exercise

Interpret this model (response: "prey capture", predictor: size of predator)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-17.5791	6.7584	-2.601	0.0123*
Pred.size	0.3932	0.0415	9.474	<0.0001***

As if it were a...

- 1) Simple linear regression
- 2) Binomial regression
- 3) Poisson regression

(Assume canonical links;  $\exp(0.93) \approx 1.50$ )

# Exercise

Interpret this model (response: "prey capture", predictor: size of predator)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-17.5791	6.7584	-2.601	0.0123*
Pred.size	0.3932	0.0415	9.474	<0.0001***

A one unit increase in predator size corresponds to...

- 1) 0.4 unit increase in prey capture
- 2) 1.5 times increase in the odds of prey capture
- 3) 50% (1.5 times) increase in prey capture

(Assume canonical links;  $\exp(0.93) \approx 1.50$ )