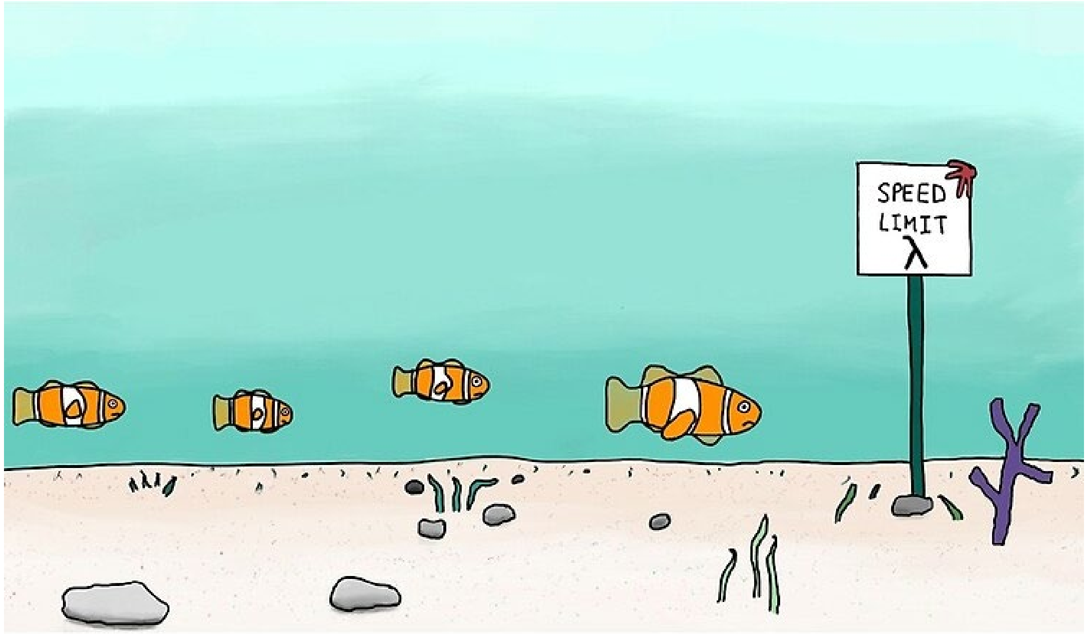
Generalized Linear Models

Regression methods for categorical response variables

ENTMLGY 6702 Entomological Techniques and Data Analysis

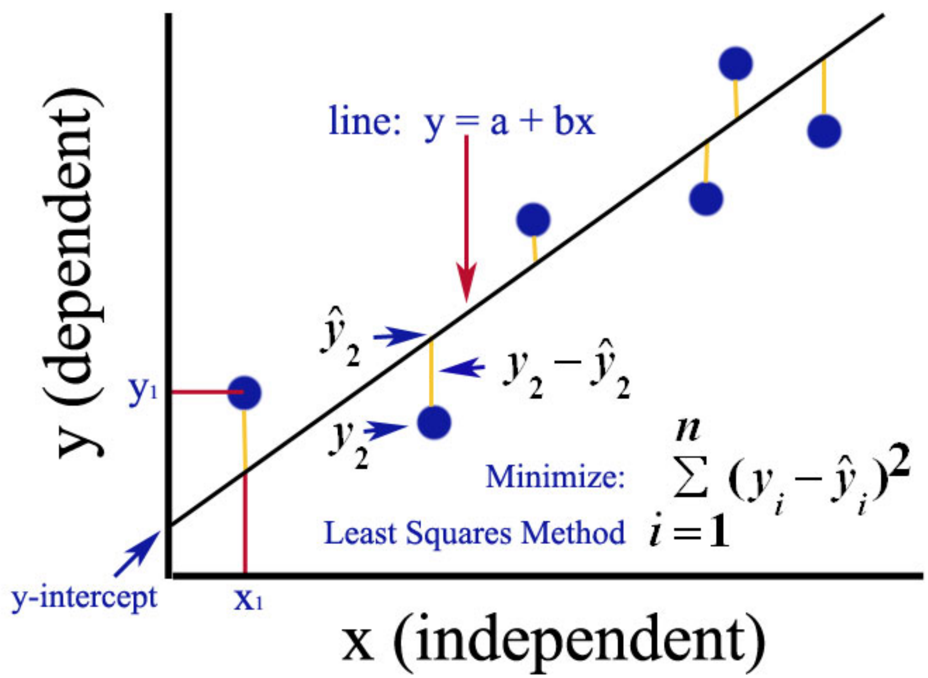


Learning objectives

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(s) between a response variable and one or more predictor variables



http://www2.cedarcrest.edu/

A quick review of simple linear regression

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

- Model is correct
- Linearity of predictors $E(Y|X) = \beta_0 + x_1\beta_1 + \ldots + x_n\beta_n$
- 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: A one unit change in X is associated with a β_1 change in the mean of Y.

Discrete responses

Historically, discrete data has 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 does ladybeetle size affect their efficacy as predators?

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

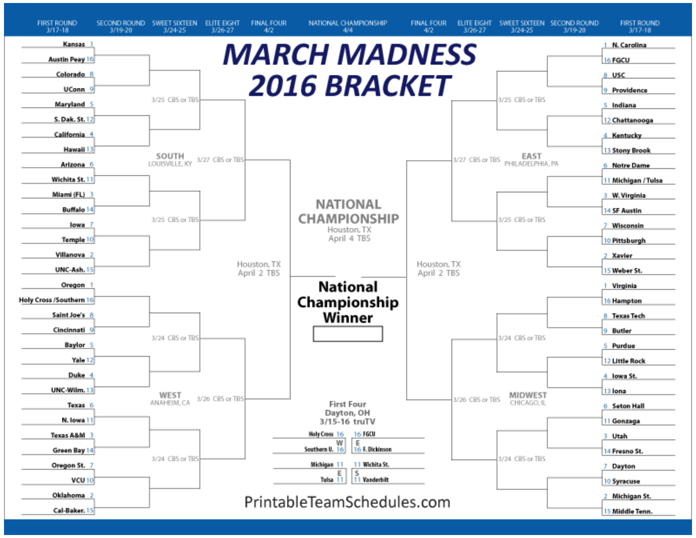
Response variable ~ Ladybeetle size

Binary data

A single binary variable may 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:

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

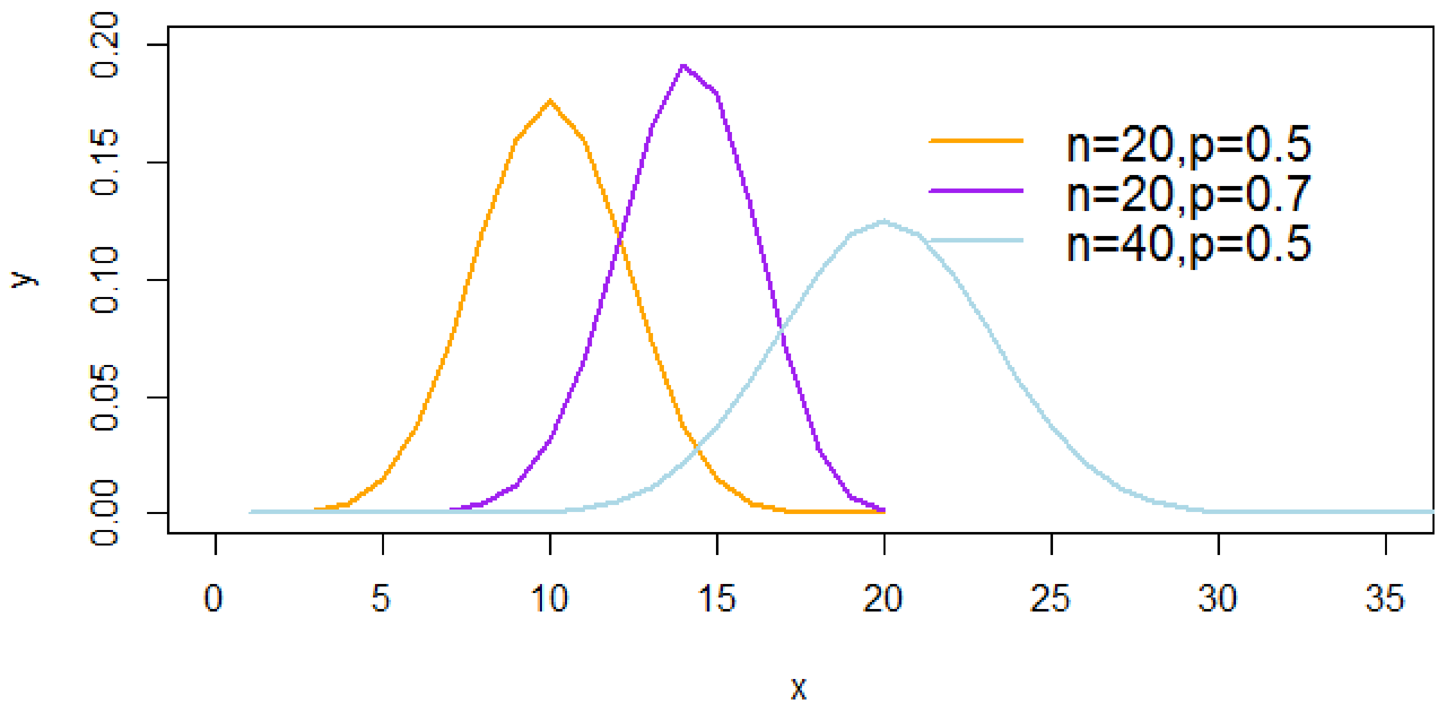
$$\text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

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:

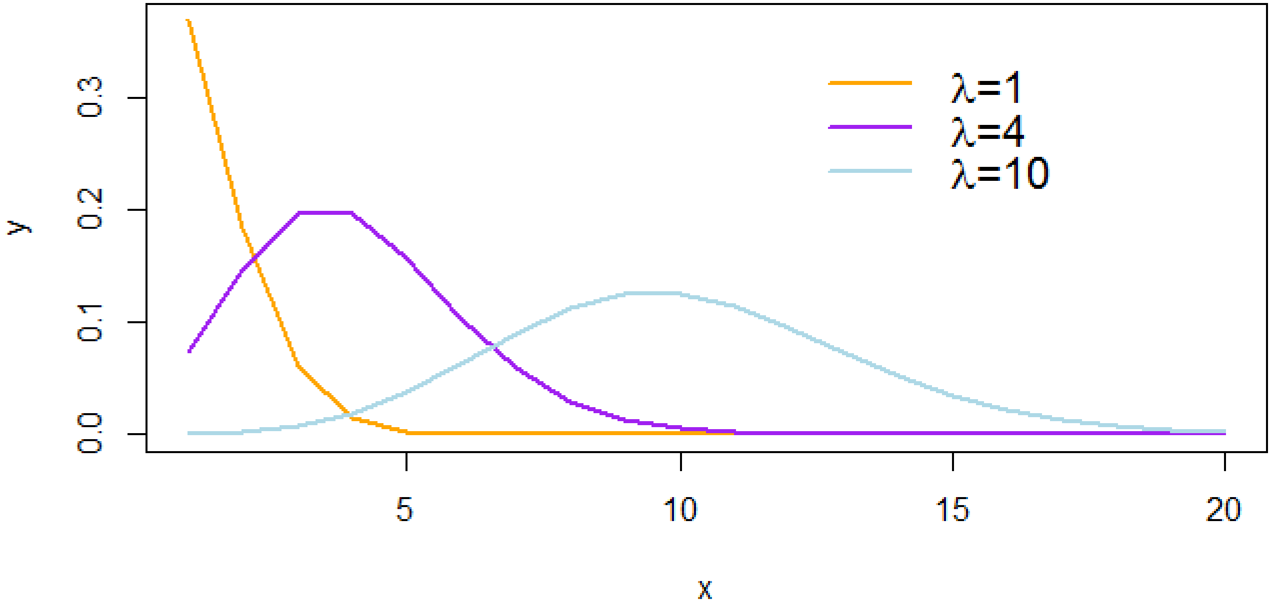
$$P(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}$$

Poisson distribution

Mean: λ

Variance: λ

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



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))
```

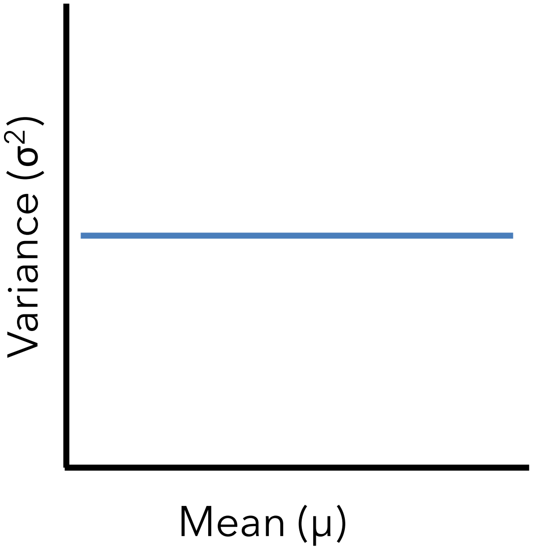
```
glmer(Count ~ X + (1|Group),
data = df,
family = poisson(log))
```

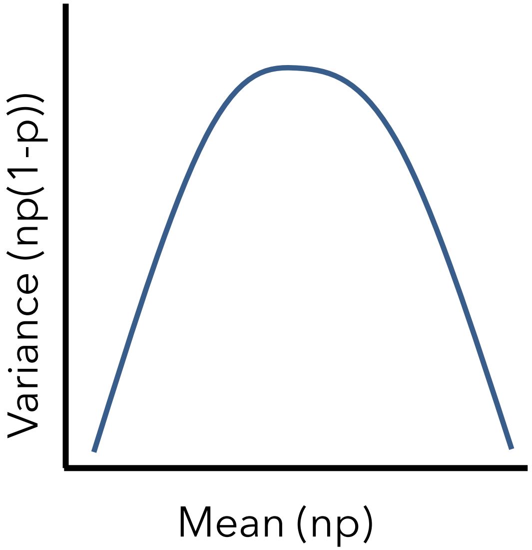
Compare to Normal distribution

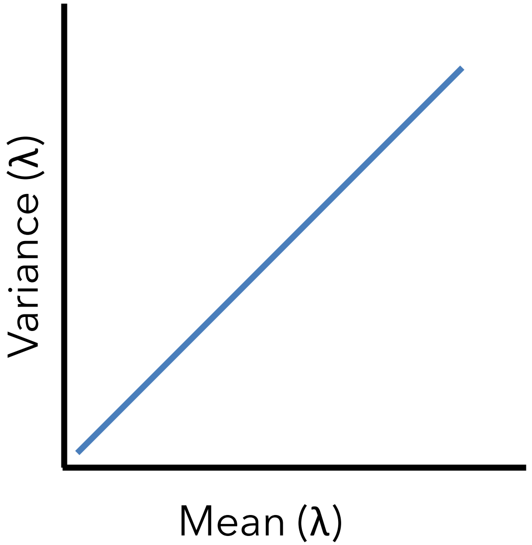
Distribution	Mean	Variance
Normal	μ	σ^2
Binomial	np	np(1-p)
Poisson	λ	λ

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 ($\beta_0+\beta_1X_1+\beta_2X_2$).

3) Select a link function, often represented using η or $g(\mu)$. Link functions define how your expected response, $E(X)$, relates to your set of linear predictors, $\beta_0+\beta_1X_1+\beta_2X_2$.

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

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 → Identity
  - Binomial → Logit (Logistic regression)
  - Poisson → 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), but don't get too hung up on the underlying math.

- Interpretations of model coefficients change with the link function used

Link functions

- Normal → identity

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

- Binomial → logit

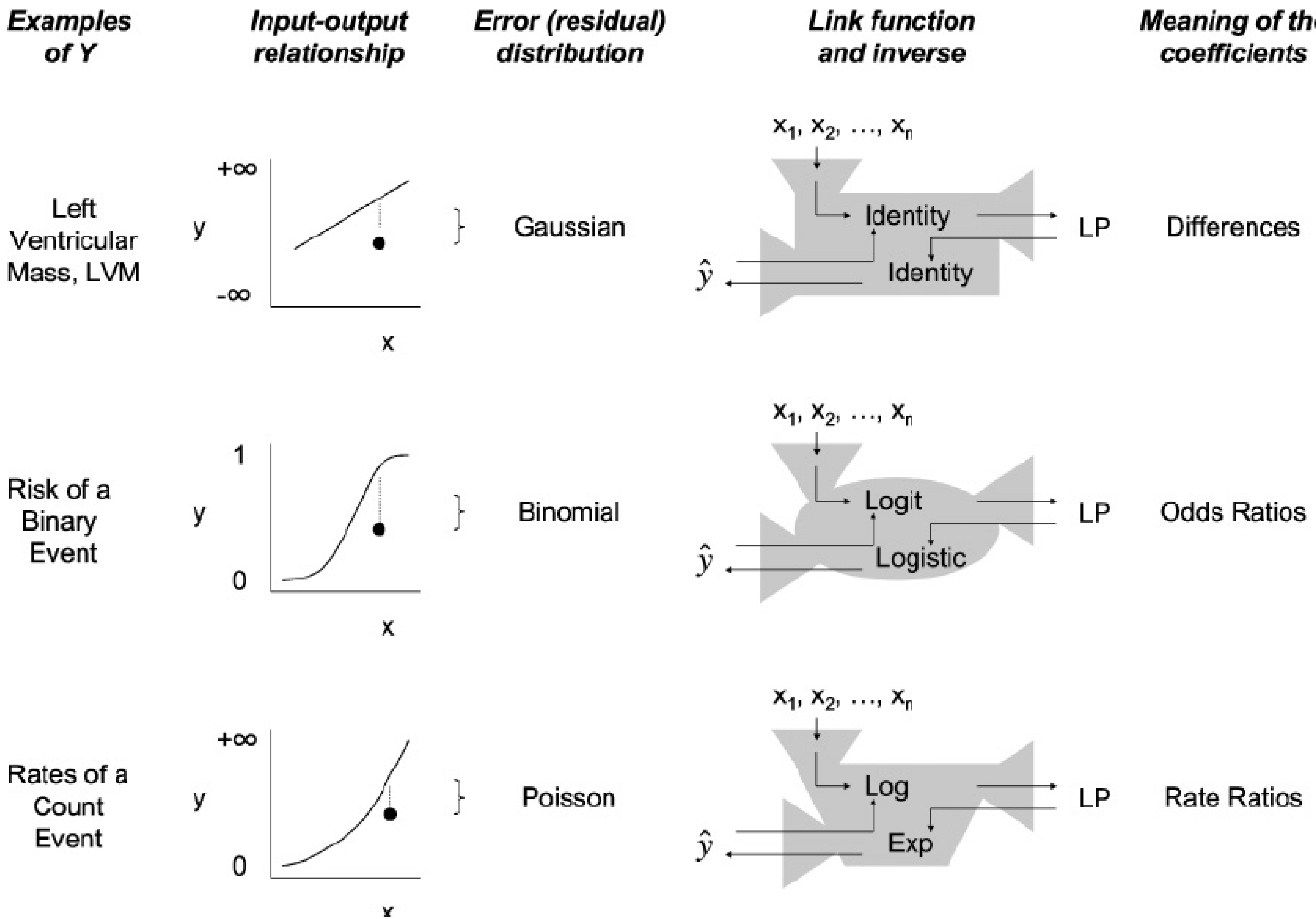
$$g(f(x)) = log(\frac{p}{1-p}) = \beta_0 + x_1\beta_1$$

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

- Poisson → log

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

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

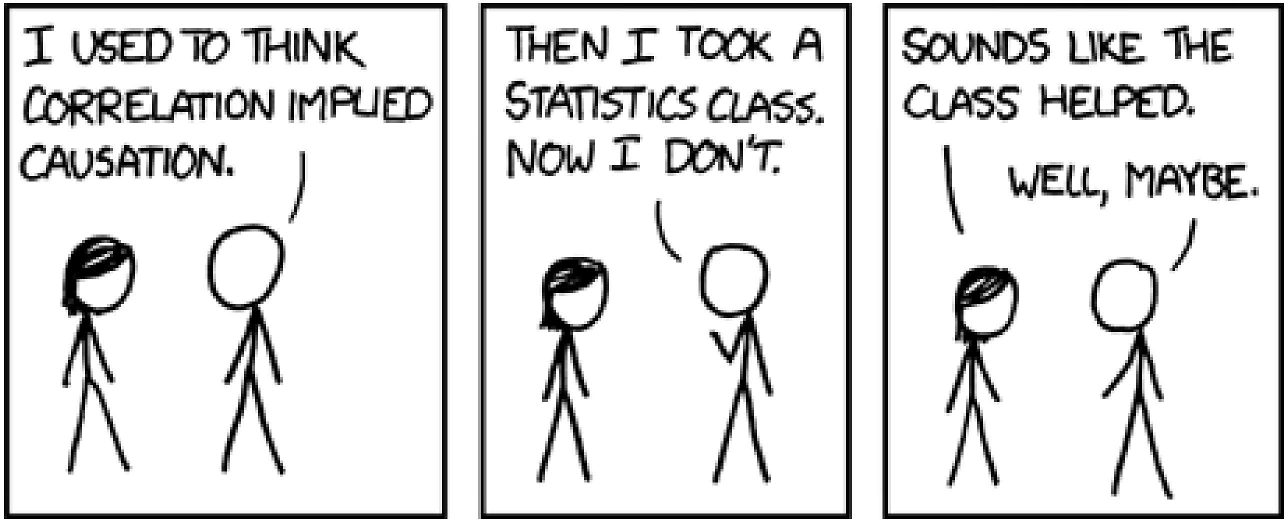


Ravani, Pietro & Parfrey, Patrick & Murphy, Sean & Gadag, Veeresh & Barrett, Brendan. (2008). Clinical research of kidney diseases IV: Standard regression models. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association.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



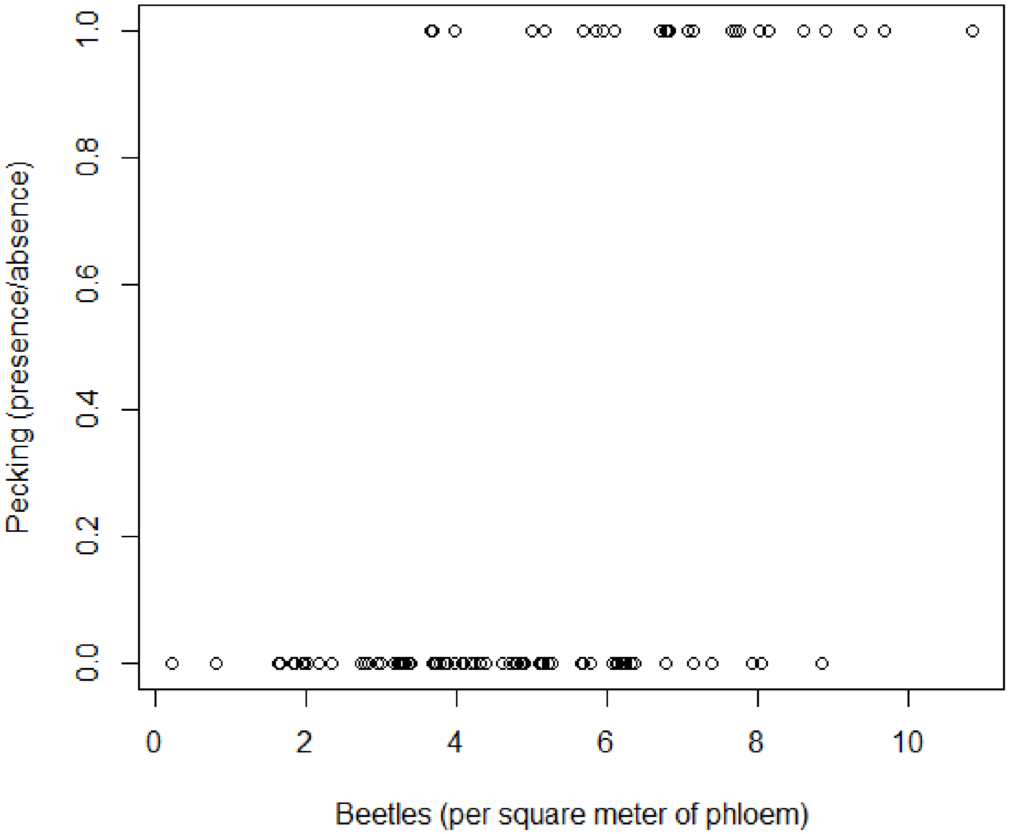
Example: Binomial (logistic) regression

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
```

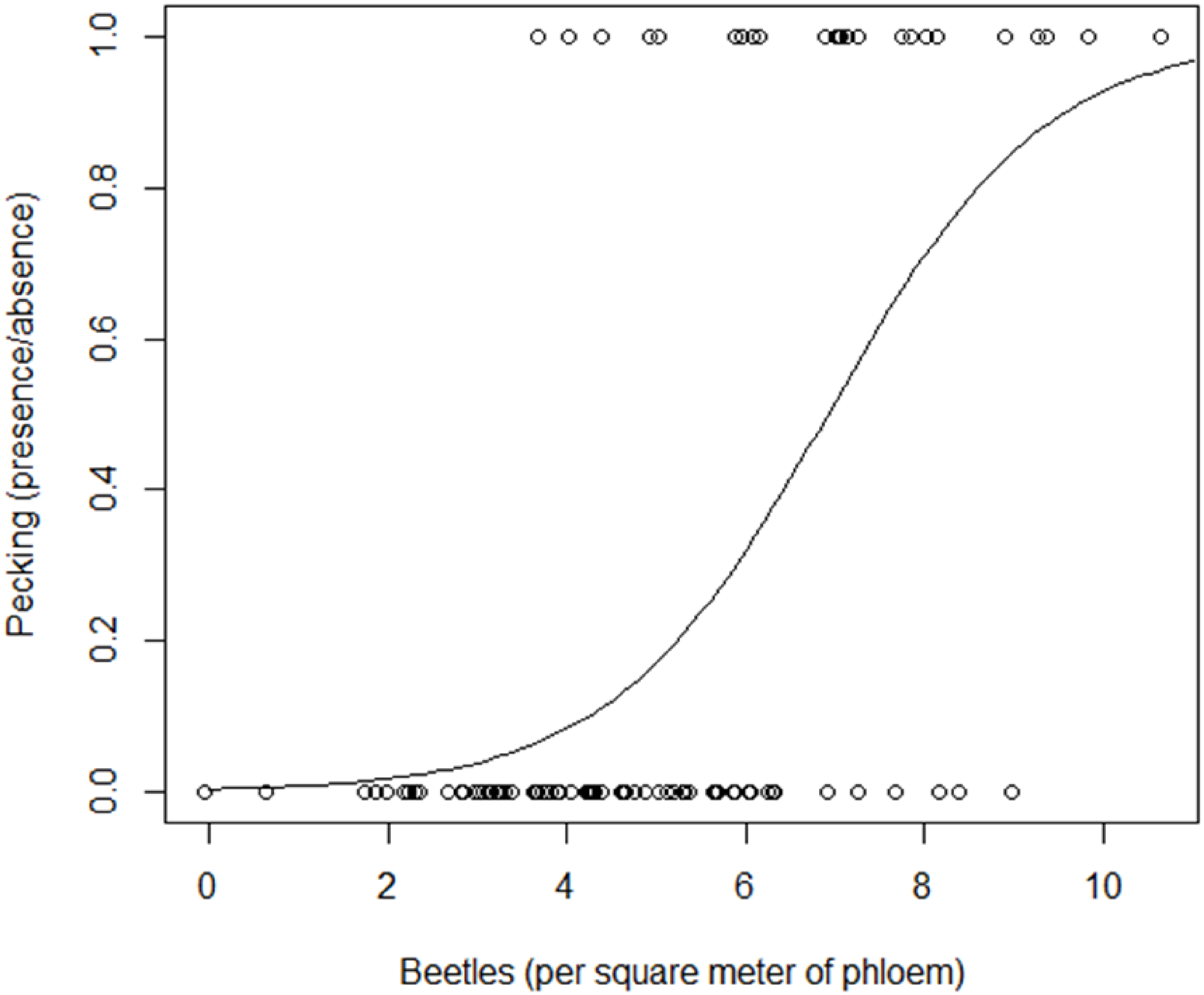
How do we interpret the model?

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."

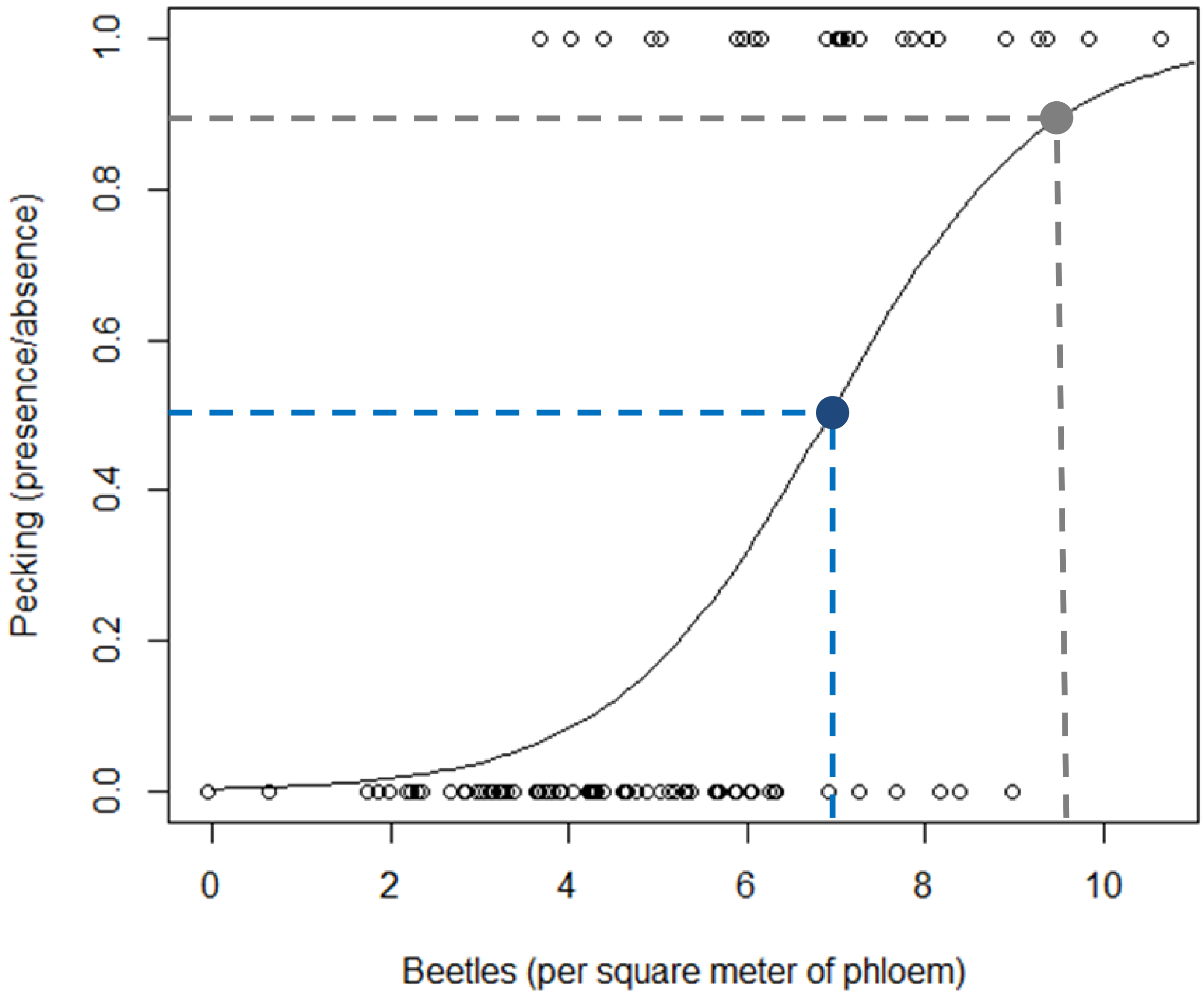
Odds ratio:

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

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?

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

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

$$2.2 + 5.7047 = 0.8252 * x_1$$

$$x_1 = 9.6$$

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

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

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

$$5.7047/0.8252 = x_1$$

$$x_1 = 6.9$$

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) ≈ 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
```

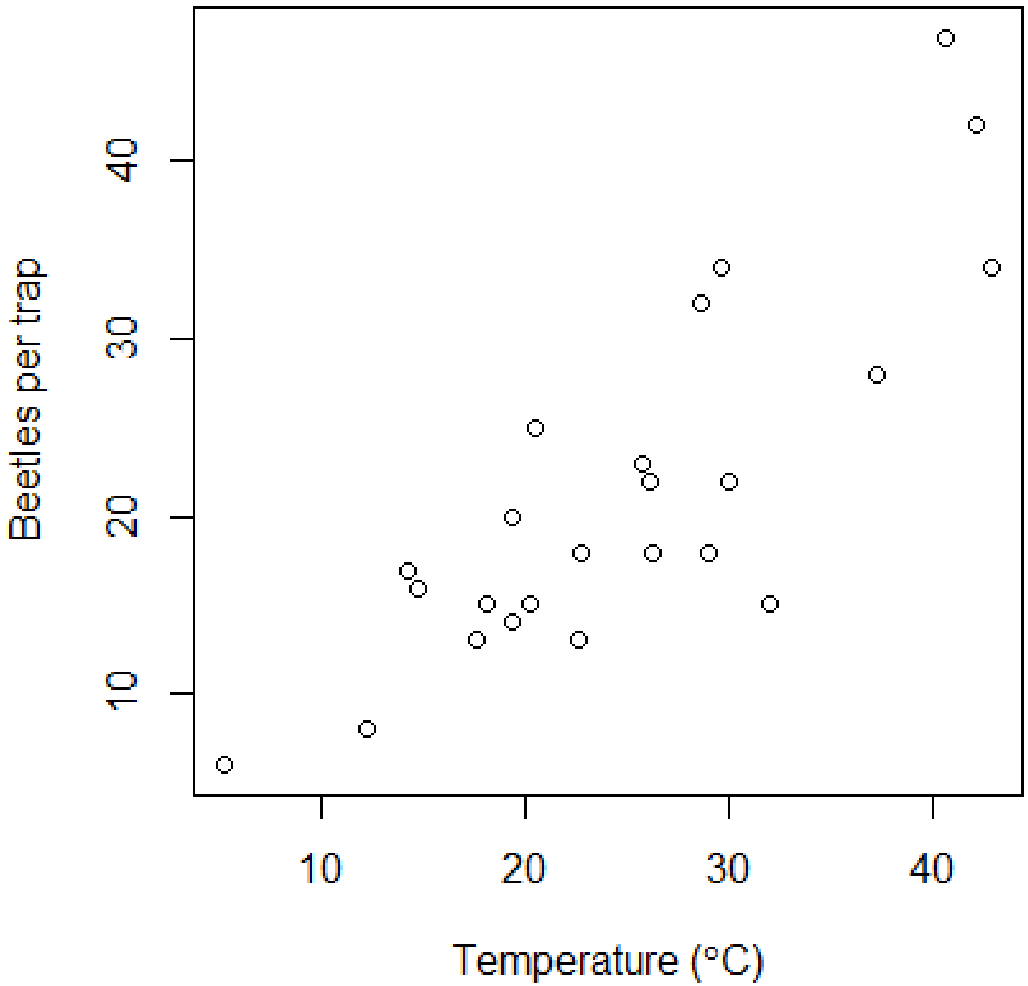
Answer: "The odds of a tree dying increased by 1.10 with a 1 unit increase in the defoliation index."

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)

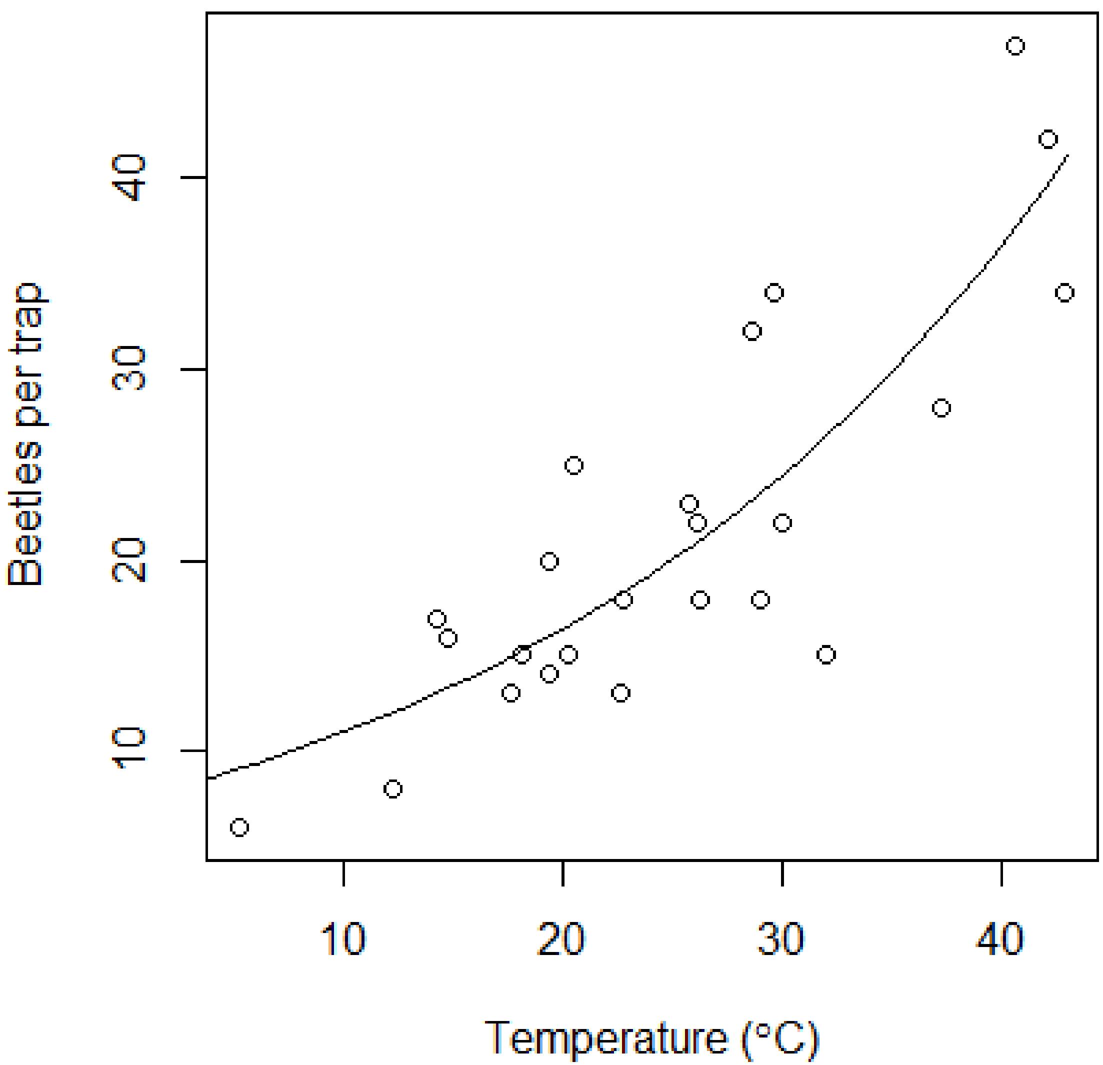
When using the log link…

- "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) ≈ 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) ≈ 1.50)

Exercise

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

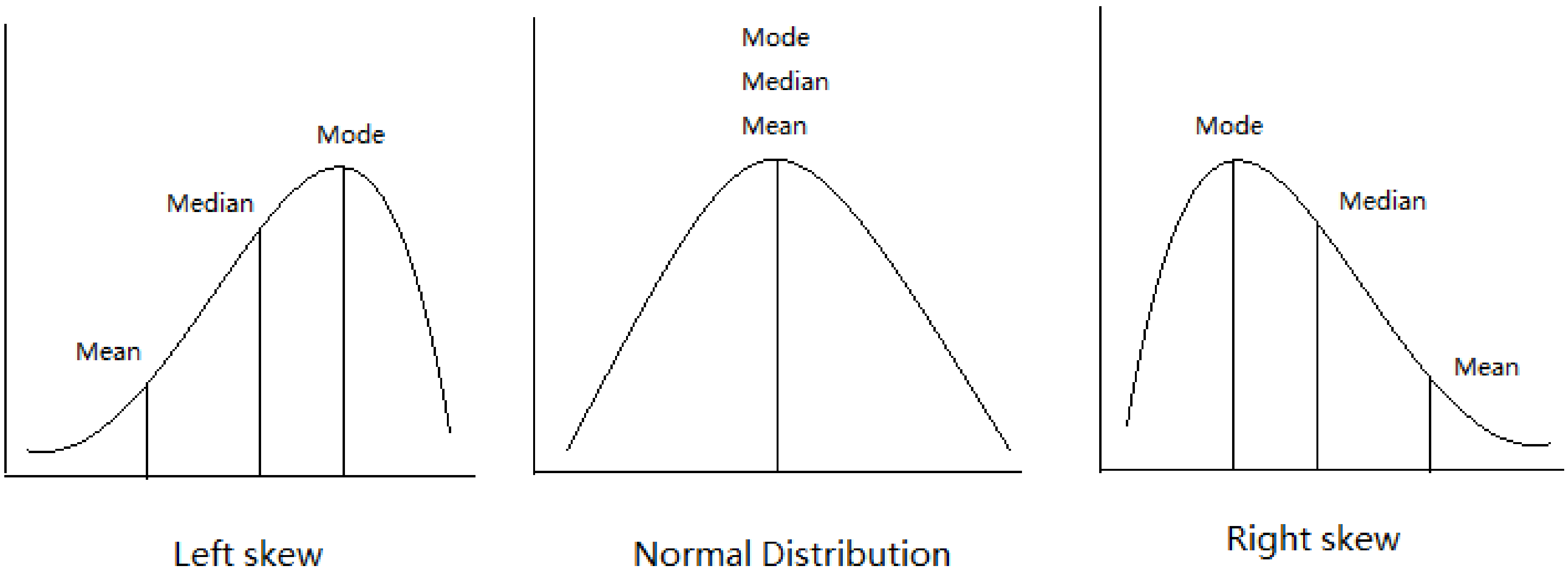
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123*
Pred.size     0.3932     0.0415   9.474  <0.0001***
```

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

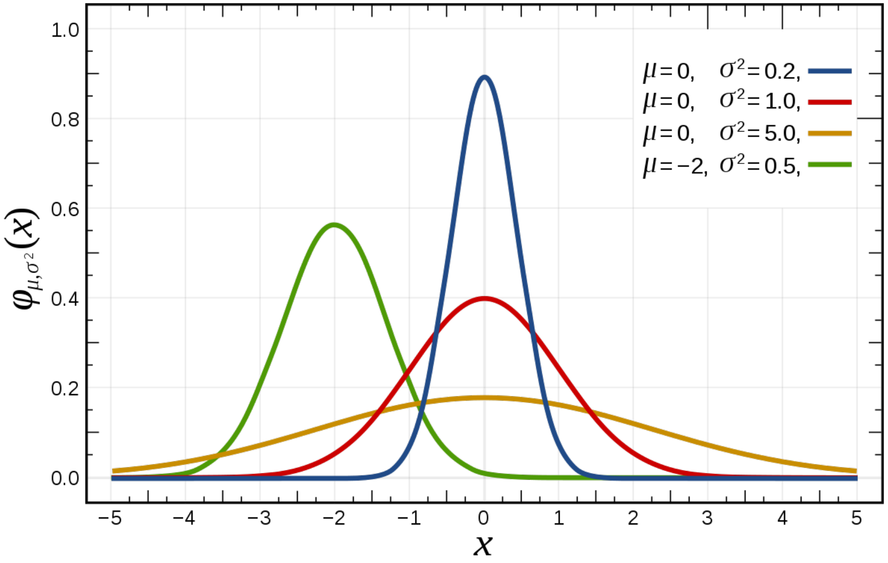
(Assume canonical links; $\exp(0.93) \approx 1.50$)

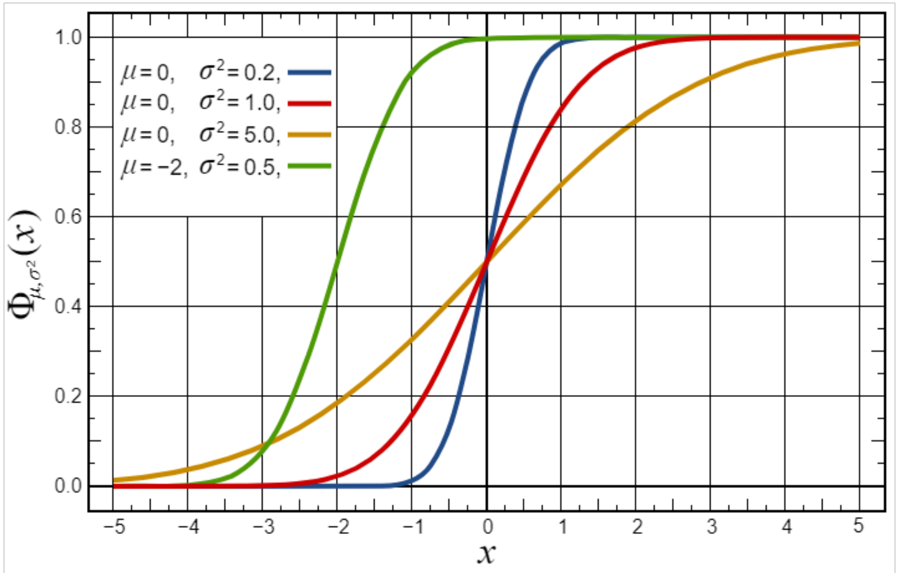
“Bonus” model: Quantile regression



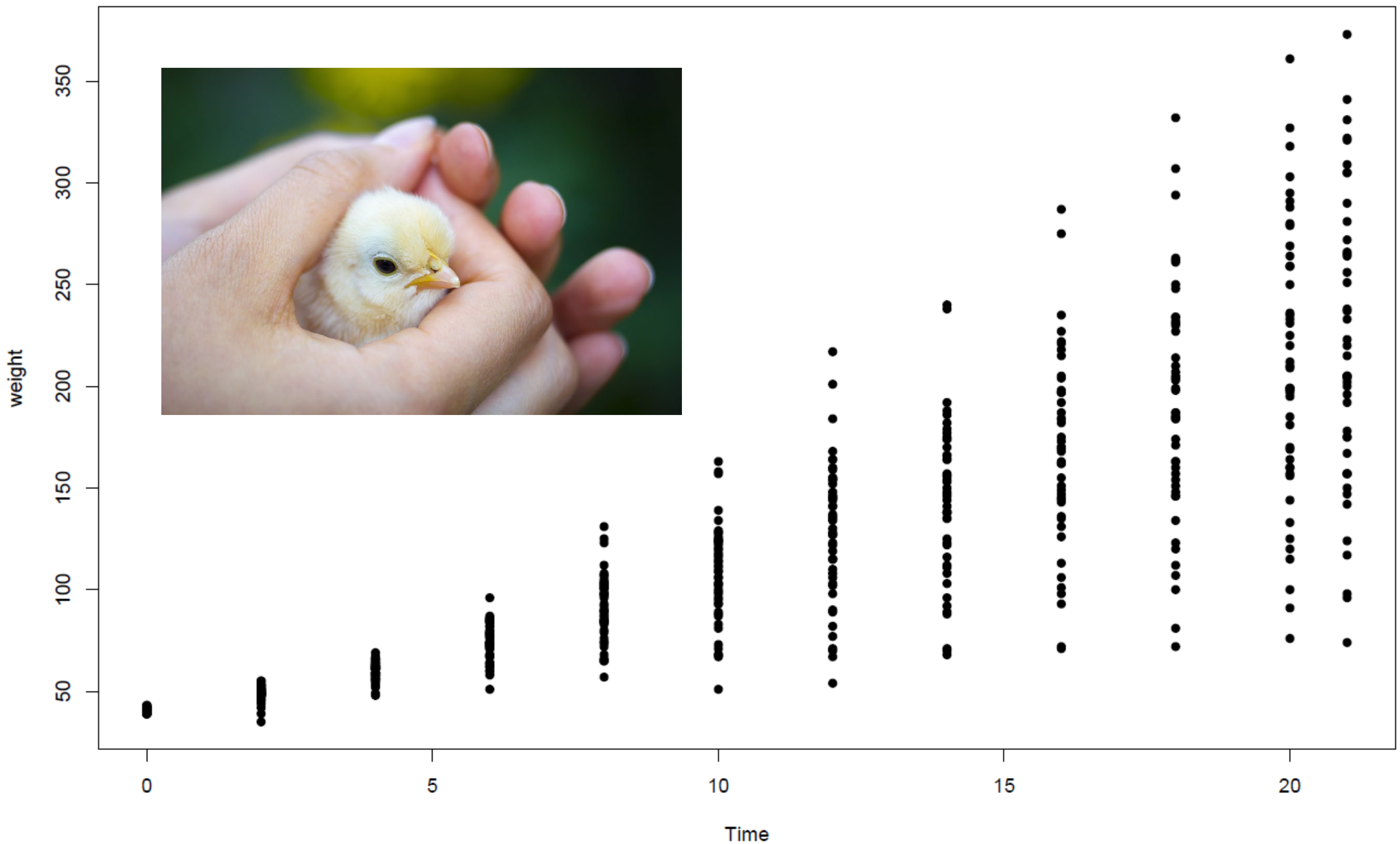
Nhan Tran (medium.com)

"Bonus" model: Quantile regression





"Bonus" model: Quantile regression



“Bonus” model: Quantile regression

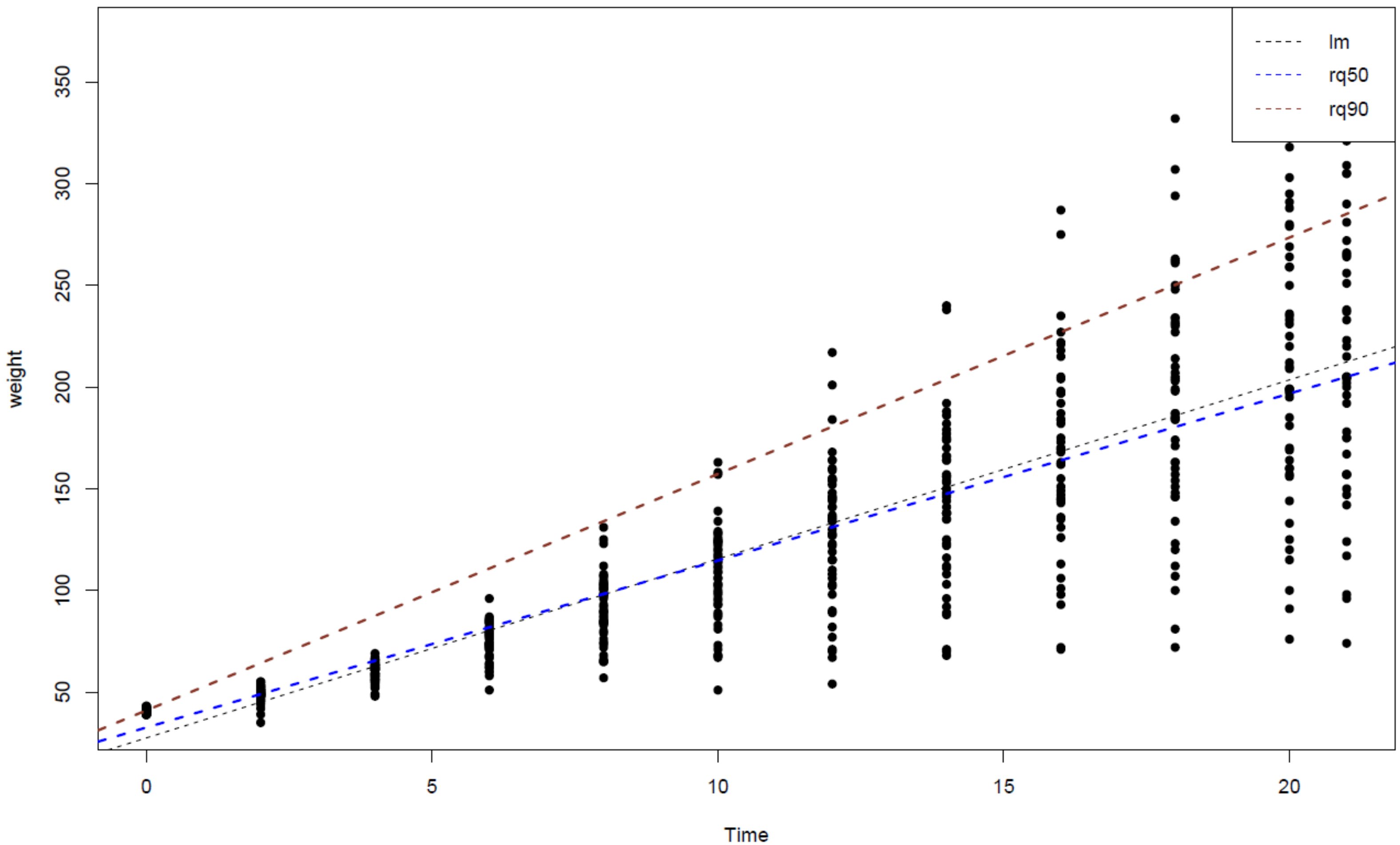
```
> library(quantreg)
> data(ChickWeight)
> rqfit_50 <- rq(weight ~ Time, data = ChickWeight)
> summary(rqfit_50)

Call: rq(formula = weight ~ Time, data = ChickWeight)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 32.57895     30.35894 33.59759
Time         8.21053      7.83257  8.56265
Warning message:
In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) : Solution may be nonunique
```

“Bonus” model: Quantile regression

```
> rqfit_90 <- rq(weight ~ Time, data = ChickWeight, tau=0.9)
> rqfit_90
Call:
rq(formula = weight ~ Time, tau = 0.9, data = ChickWeight)

Coefficients:
(Intercept)        Time
     41.000      11.625

Degrees of freedom: 578 total; 576 residual
> summary(rqfit_90)

Call: rq(formula = weight ~ Time, tau = 0.9, data = ChickWeight)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 41.00000     41.00000 42.19027
Time        11.62500     11.41475 11.92790
```

Interpretation: A one unit change in X is associated with a β_1 change in the median of Y.

“Bonus” model: Quantile regression



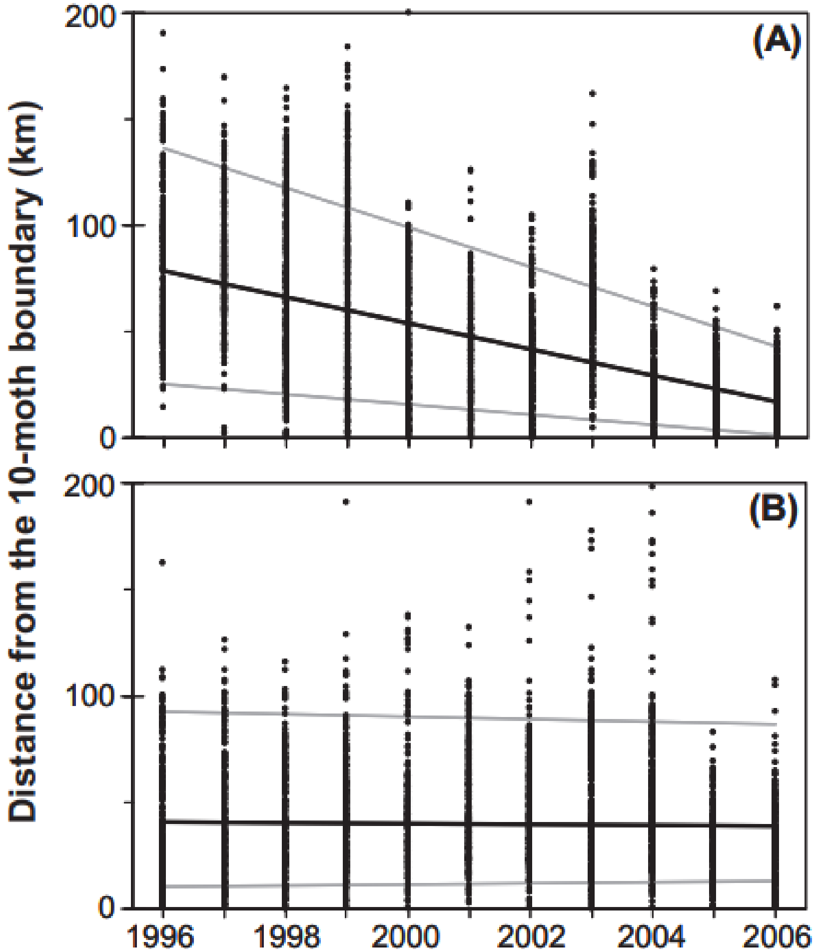


Fig. 3. Distance between cells recording high-low association (identified by the Local Moran, equation 1, and the dynamic 10-moth population boundary in Wisconsin [A], and West Virginia and Virginia [B], 1996–2006). Dots are observed distances, and the three lines are the predictions from the quantile regression fit to the 90th percentile (top gray line), the 50th percentile (middle black line), and the 10th percentile (bottom gray line). Note how the distance in A sharply declines through time although not in B.

Long-Distance Dispersal of the Gypsy Moth (Lepidoptera: Lymantriidae) Facilitated Its Initial Invasion of Wisconsin

PATRICK C. TOBIN[1] AND LAURA M. BLACKBURN

Forest Service, U.S. Department of Agriculture, Northern Research Station, 180 Canfield St., Morgantown, WV 26505-3101

Environ. Entomol. 37(1): 87–93 (2008)