Generalized Linear Models

Assumptions of Linear Model

- Homoskedasticity **Model variance**
- No error in X variables **Errors in variables**
- No missing data **Missing data model**
- **Normally distributed error** **GLM**
- Error in Y variables is measurement error
- Observations are independent

Generalized Linear Models

- Retains linear function
- Allows for alternate PDFs to be used in likelihood
- However, with many non-Normal PDFs the range of the model parameters does not allow a linear function to be used safely
  - Pois(λ): $\lambda > 0$
  - Binom(n,θ) $0 < \theta < 1$
- Typically a *link* function is used to relate linear model to PDF

Link Functions

- "Canonical" Link Functions

Distribution	Link Name	Link Function	Mean Function
Normal	Identity	$Xb = \mu$	$\mu = Xb$
Exponential	Inverse	$Xb = \mu^{-1}$	$\mu = (Xb)^{-1}$
Gamma			
Poisson	Log	$Xb = \ln(\mu)$	$\mu = \exp(Xb)$
Binomial	Logit	$Xb = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(Xb)}{1+\exp(Xb)}$
Multinomial			

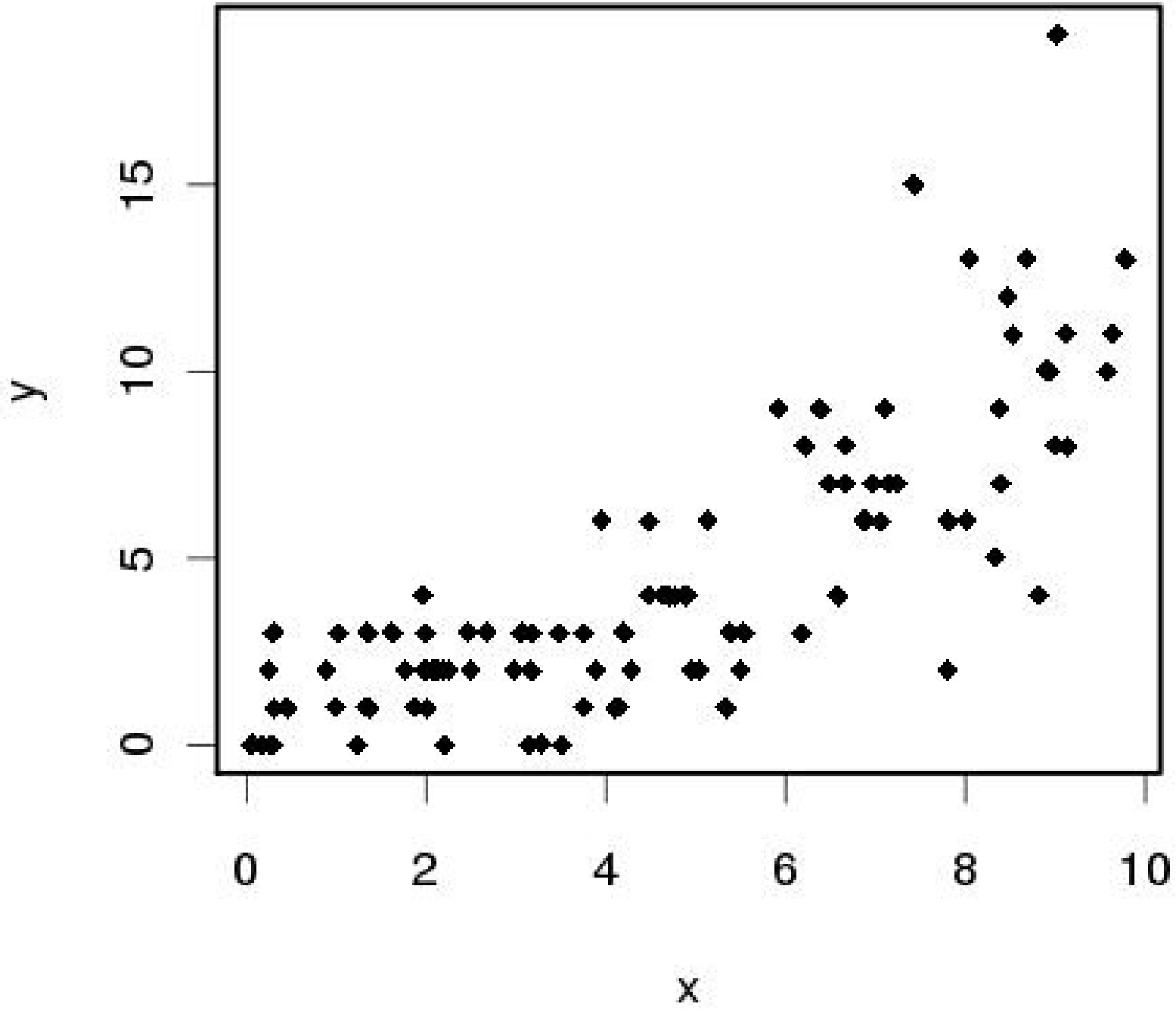
- Can use most any function as a link function but may only be valid over a restricted range
- Most are technically nonlinear functions

Poisson Regression

- Most common link is log(λ)

$$y \sim Poisson(\lambda)$$
$$\log(\lambda) = X\beta$$

- Commonly used to model count data
  - Especially for low counts were poor normal approx
- Easily generalized to Negative Binomial regression (not canned)
- Flexible in alternative link functions ($\lambda > 0$)
  - Lab 4: cone counts


y
0
5
10
15
0
2
4
6
8
10
x

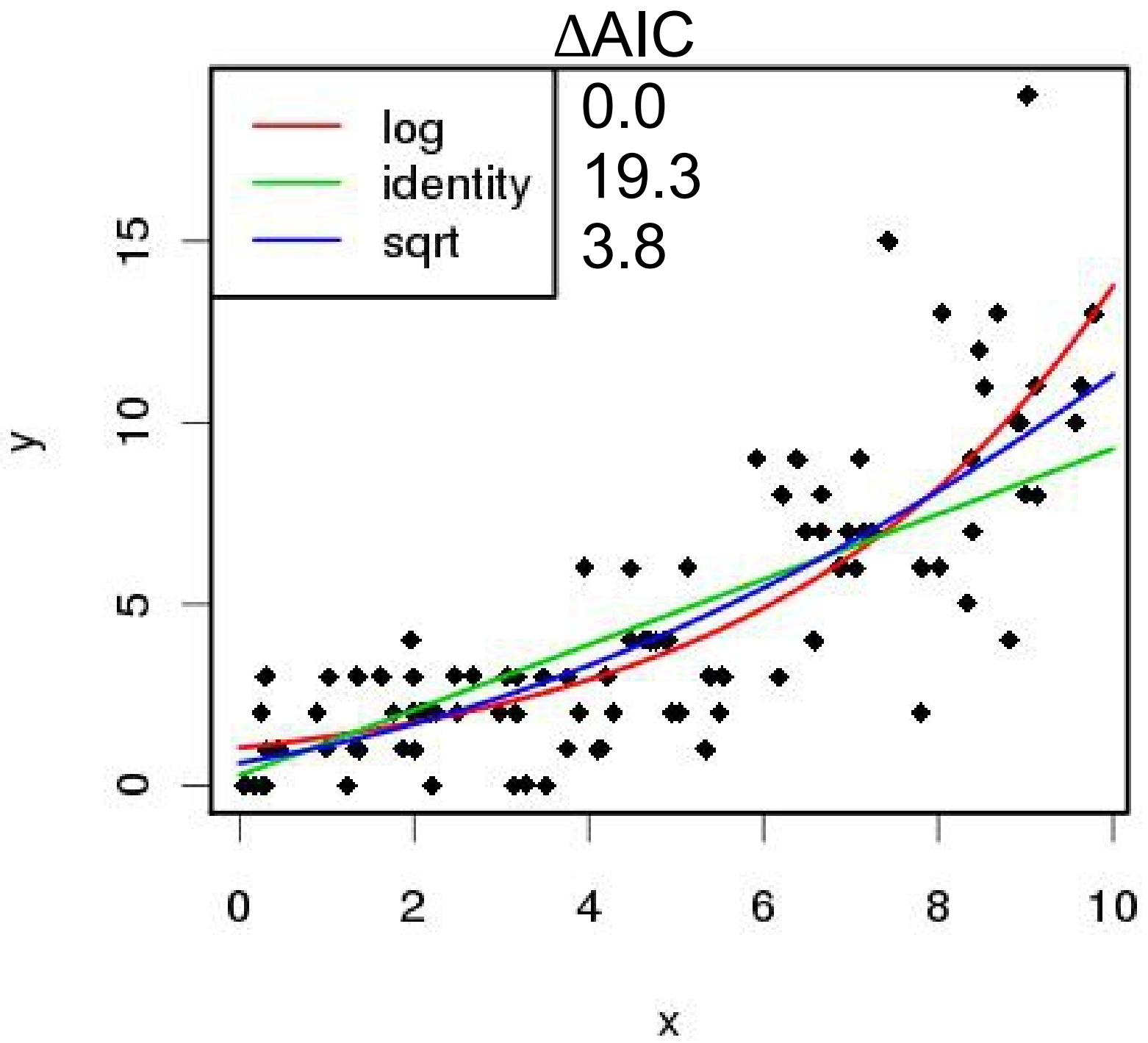
Likelihood in R

- Option 1

```
glm(y ~ x,family=poisson(link="log"))
```

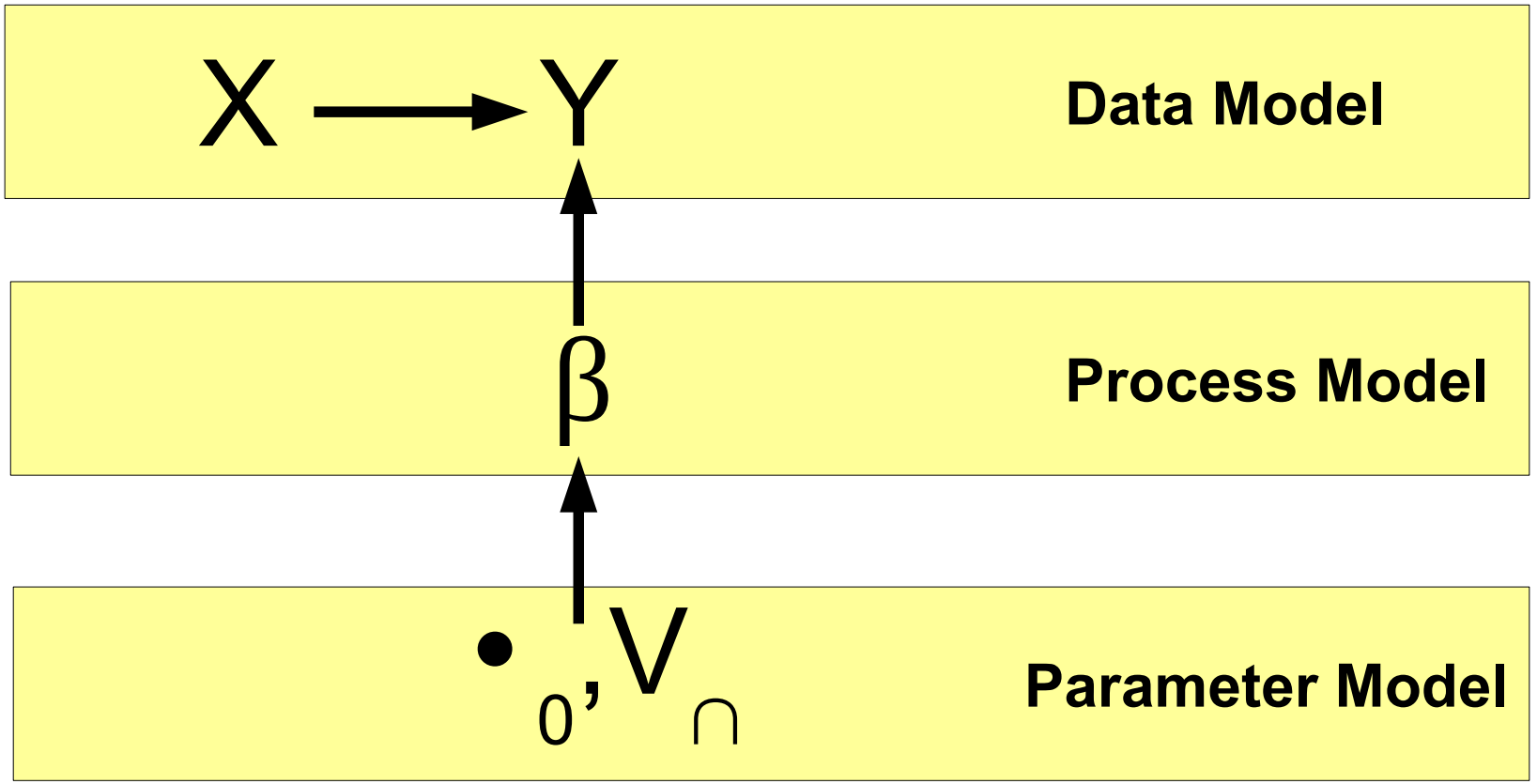
- Option 2

```
lnL = function(beta){
    -sum(dpois(y,exp(beta[0] + beta[1]*x),log=T))
}
```


ΔAIC
log
identity
sqrt
0.0
19.3
3.8
y
x
0
5
10
15
0
2
4
6
8
10

Poisson Regression

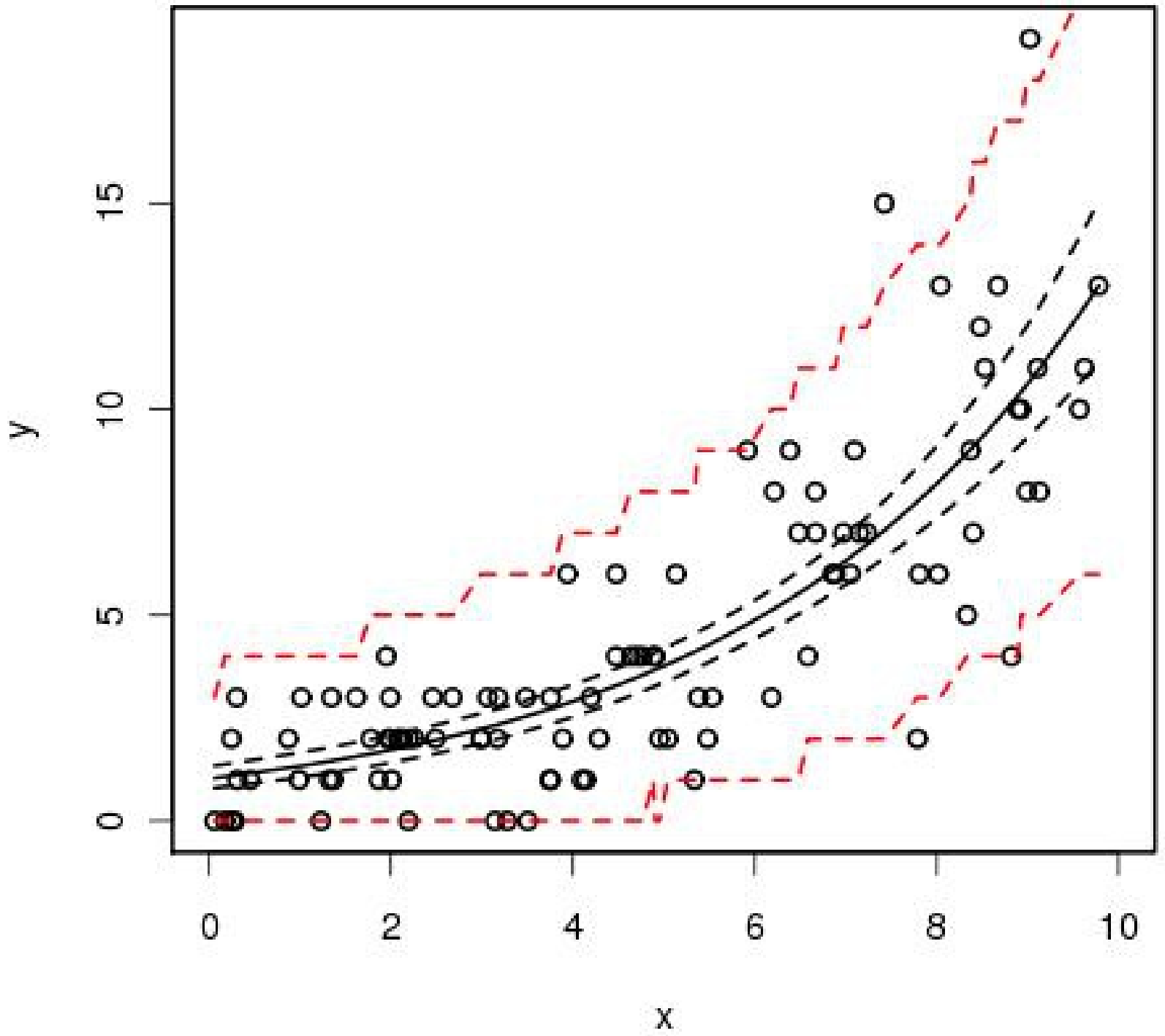
$$\vec{y} \sim \text{Pois}\left(\exp\left(\boldsymbol{X}\vec{\beta}\right)\right)$$



Bayesian Poisson Regression

$$y \sim Poisson(\lambda)$$
$$\log(\lambda) = X\beta$$
$$\beta \sim N(B_{0,}, V_B)$$

```
model{
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
  ## no sigma
  for(i in 1:n){
    log(mu[i]) <- beta[1]+beta[2]*x[i]
    y[i] ~ dpois(mu[i])
  }
}
```


y
0
5
10
15
0
2
4
6
8
10
x

Logistic Regression

- Common model for the analysis of boolean data (0/1, True/False, Present/Absent)
- Assumes a Bernoulli likelihood
  - Bern(θ) = Binom(1,θ)
- Likelihood specification

$$y \sim Bern(\theta) \qquad \textbf{Data Model}$$

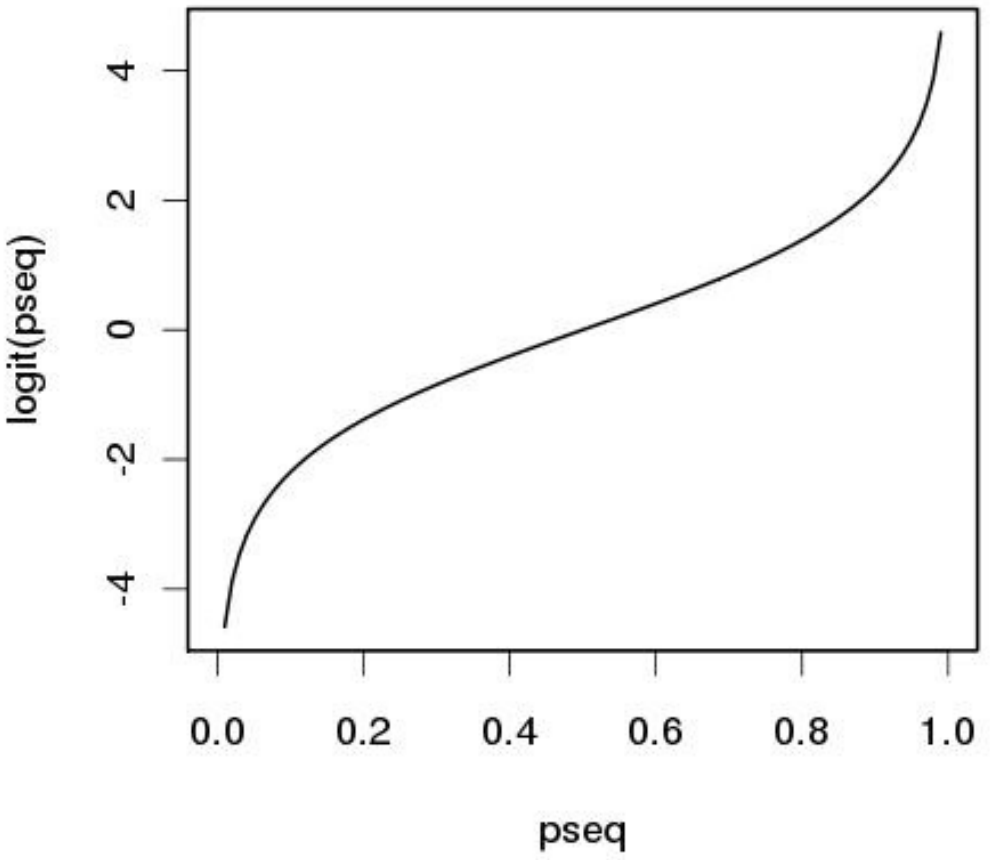
$$logit(\theta) = X\beta \qquad \textbf{Process Model}$$

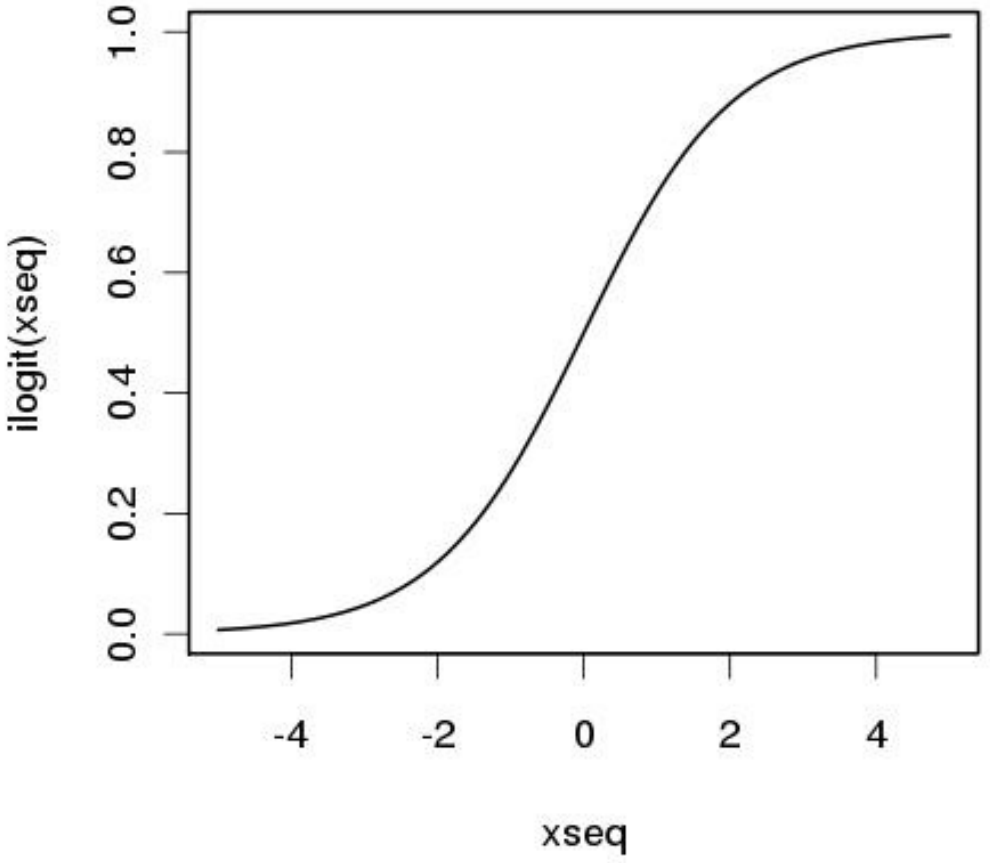
- Bayesian

$$\beta \sim N(B_{0,}, V_B) \qquad \textbf{Parameter Model}$$

Logit $Xb = \ln\left(\frac{\mu}{1-\mu}\right)$

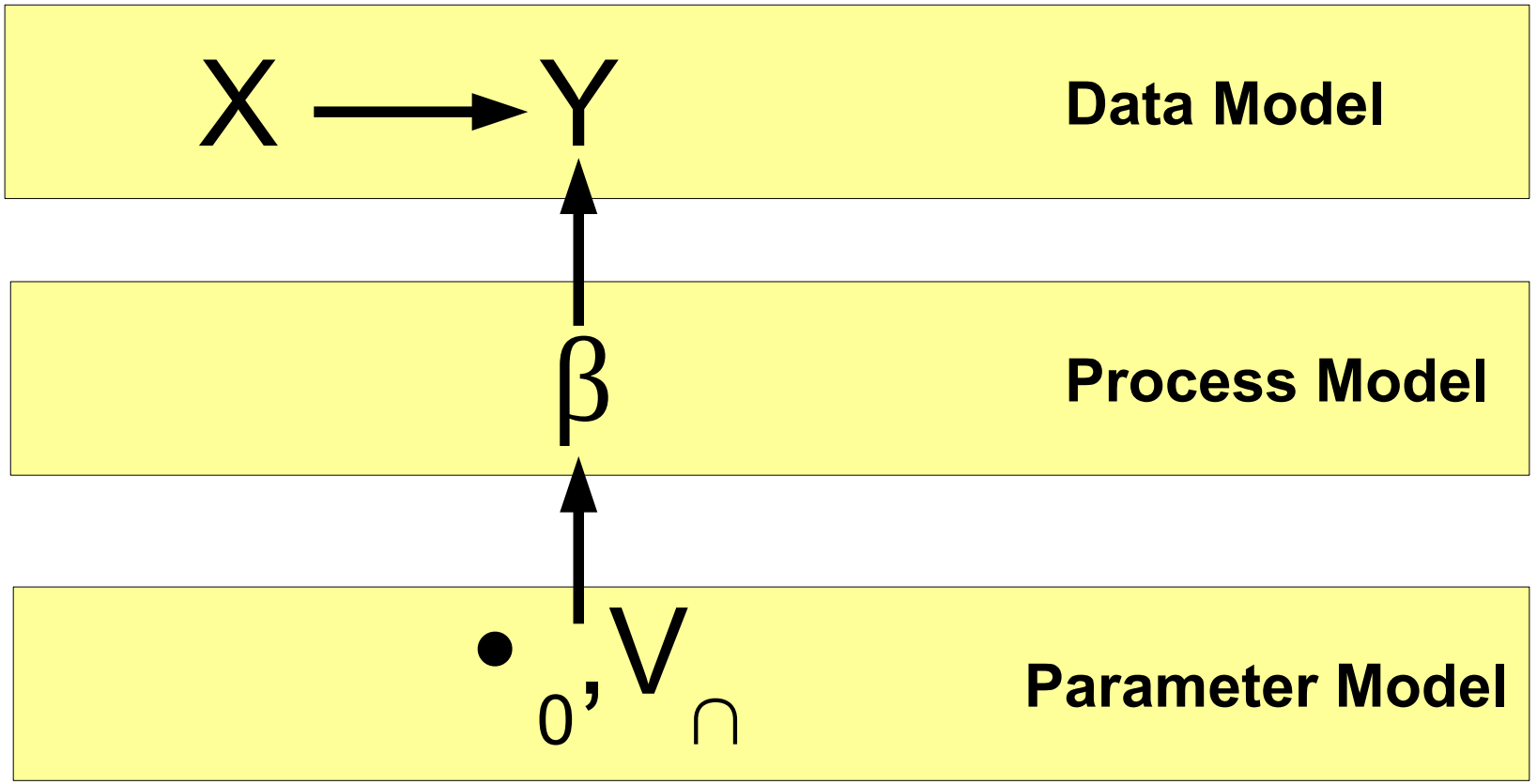
- Interpretation: Log of the ODDS RATIO
- logit(0.5) = 0.0

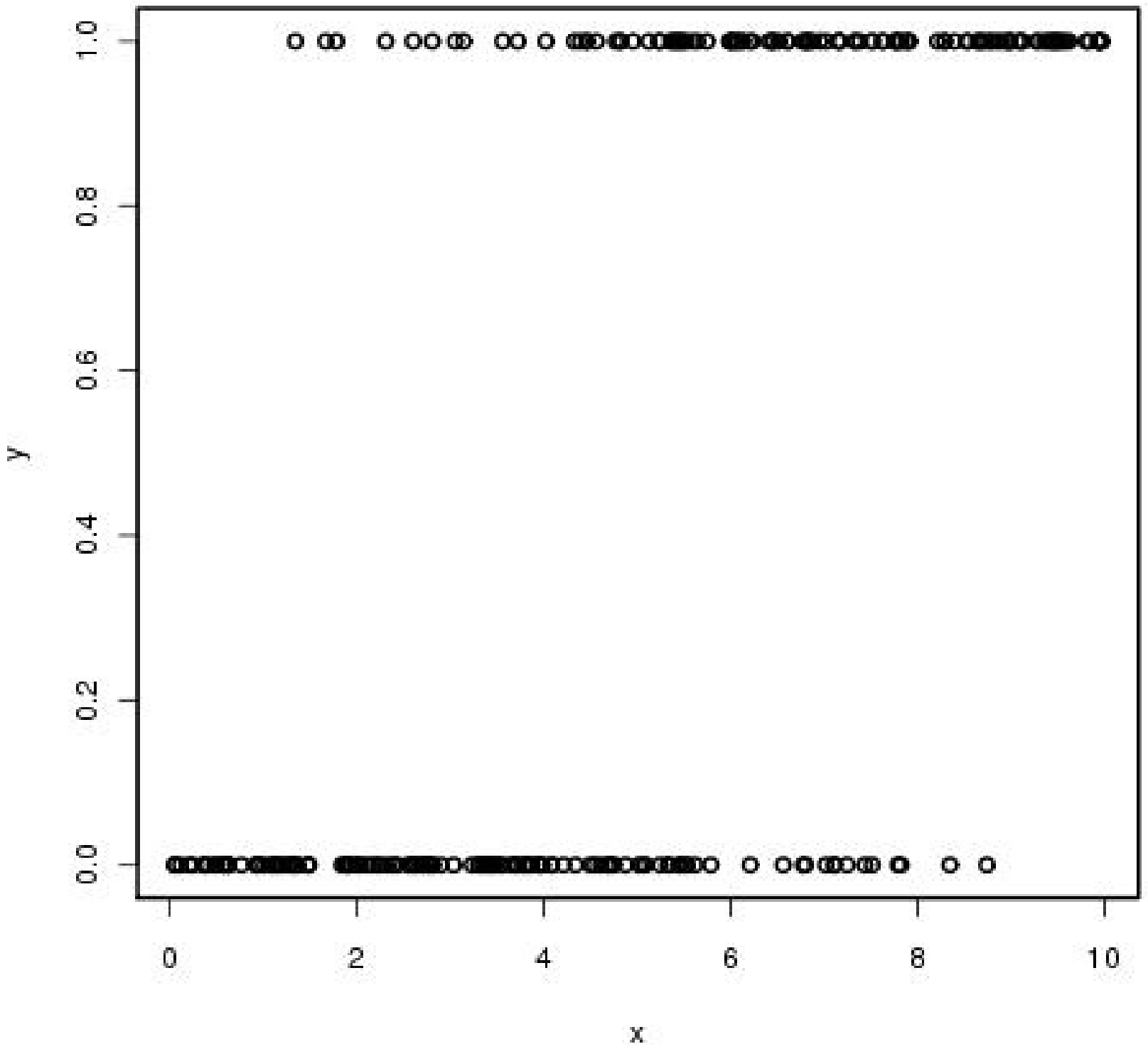


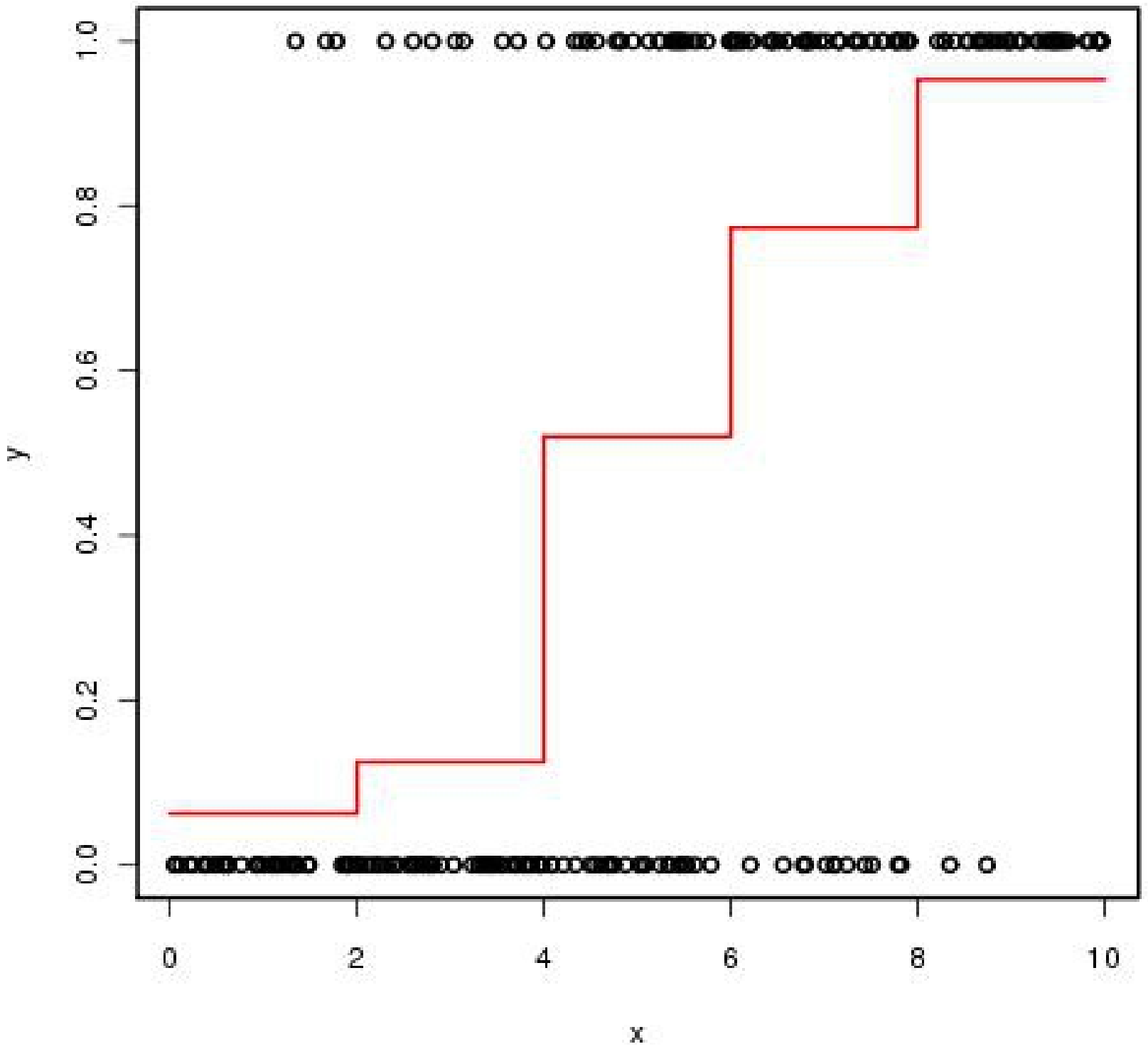


Logistic Regression

$$\vec{y} \sim \text{Binom}\left(1, logit^{-1}\left(\boldsymbol{X}\vec{\beta}\right)\right)$$




y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x


y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

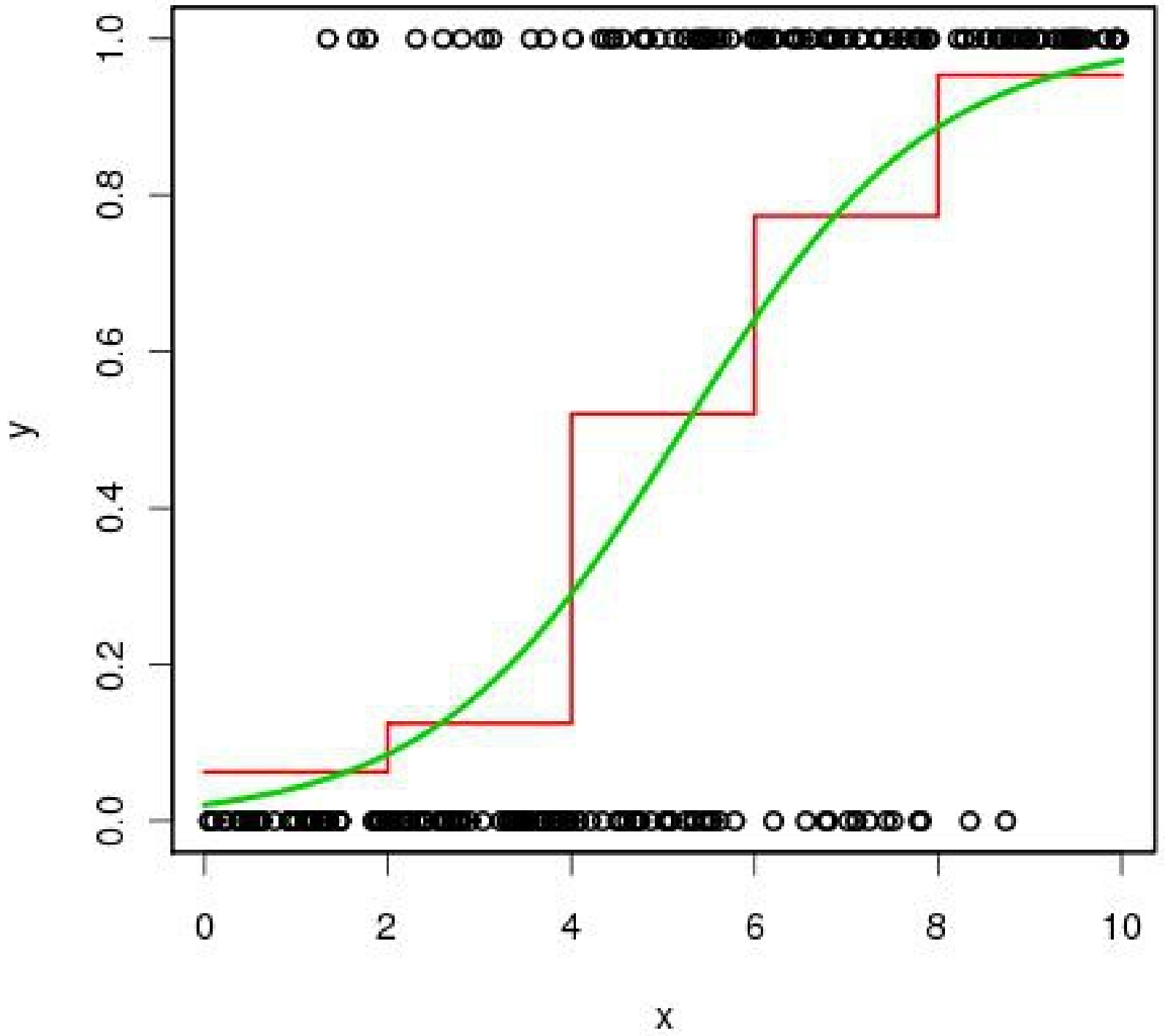
Logistic Regression in R

- Option 1 – built in function

```
glm(y ~ x, family = binomial(link="logit"))
```

- Option 2 – homebrew

```
lnL = function(beta){
    -dbinom(y,1,ilogit(beta[0] + beta[1]*x),log=T)
}
```


y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

Call:
glm(formula = y ~ x, family = binomial())

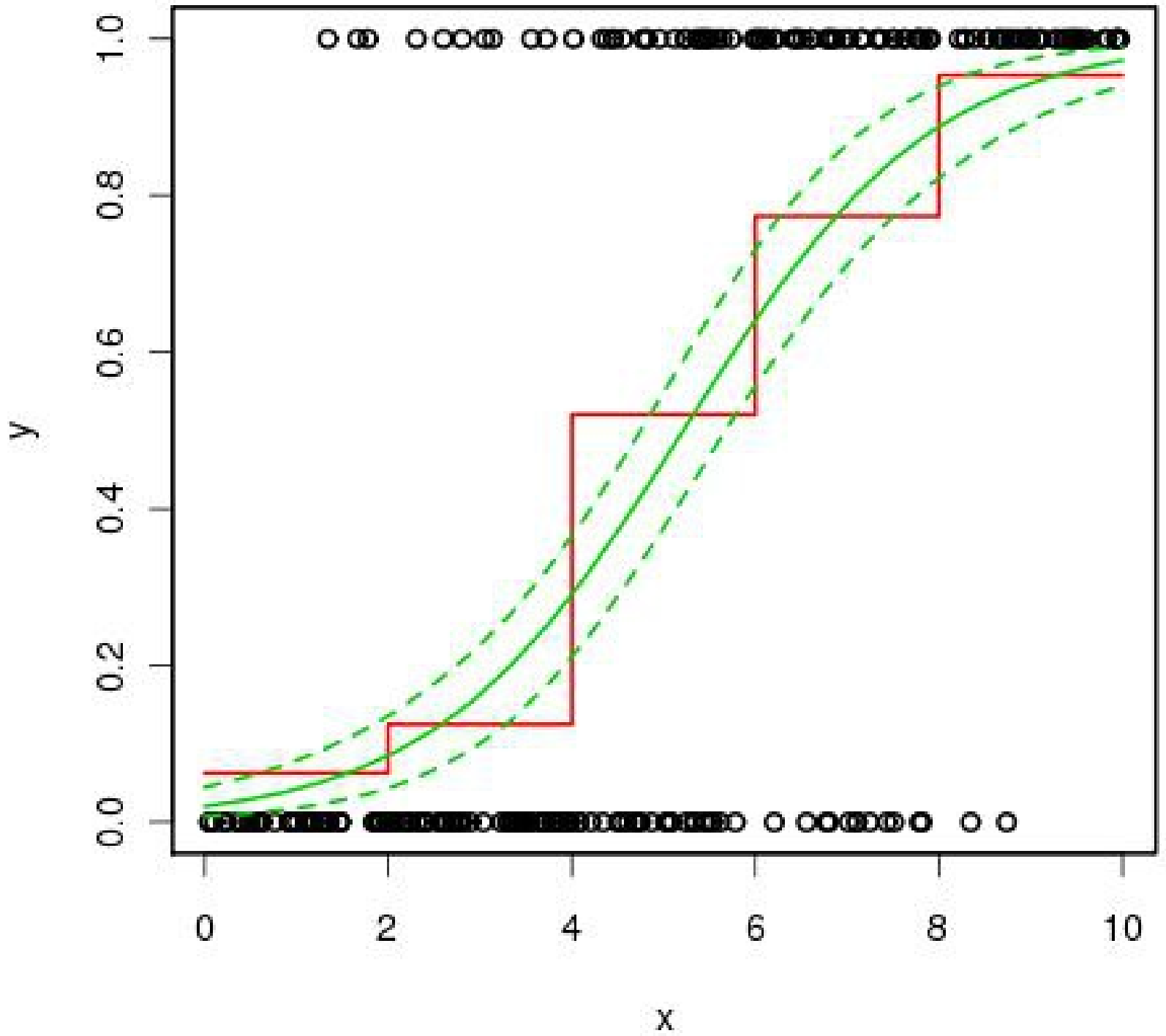
Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.3138 -0.6560 -0.2362   0.6169   2.4143

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) **-3.85078** **0.48091** -8.007 1.17e-15 ***
x   **0.73874** **0.08779**   8.415   < 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: 345.79 on 249 degrees of freedom
Residual deviance: 209.40 on 248 degrees of freedom
AIC: 213.40

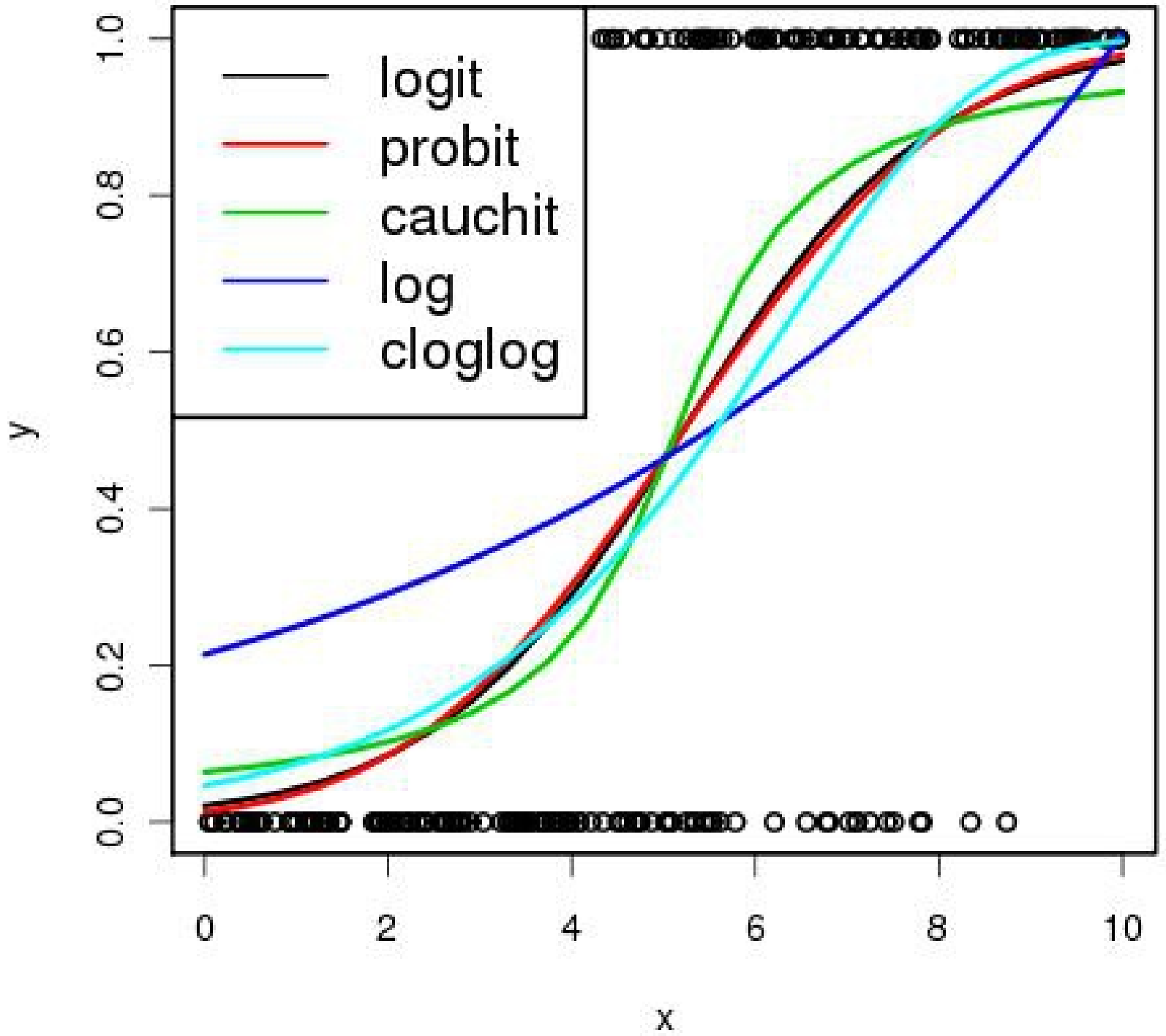

y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

Alternative link functions

- “probit” – Normal CDF
- “cauchit” - Cauchy CDF
- “log” -- $\mu = \exp(X\beta)$
- “cloglog” - Complimentary log-log
  - Asymmetric, often used for high or low probabilities

$$\mu = 1 - \exp(-\exp(X\beta))$$

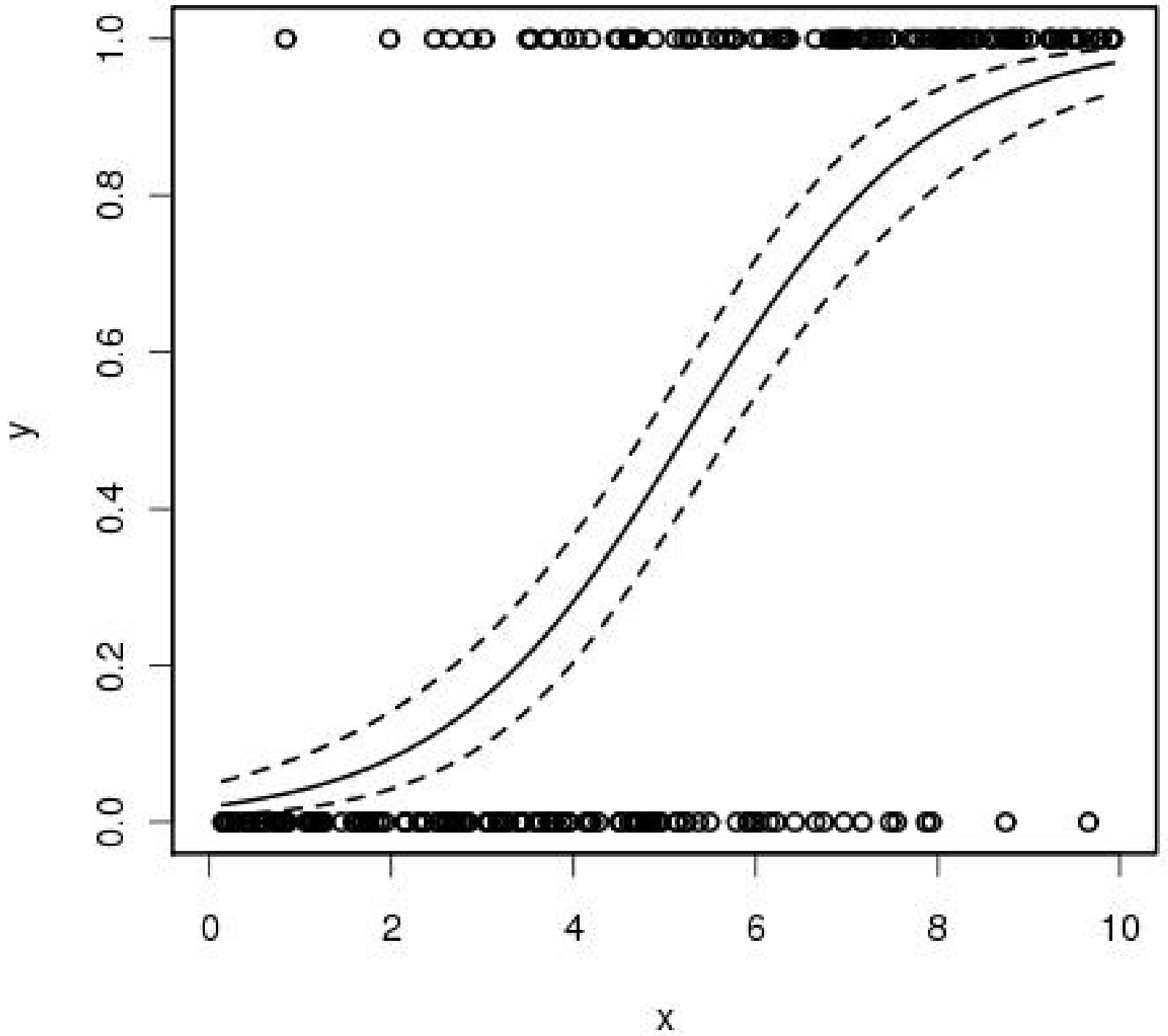
- If you code yourself, any function that projects from Real to (0,1)


logit
probit
cauchit
log
cloglog
y
x
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

Bayesian Logistic Regression

```
model {
  ## priors
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}

  for(i in 1:n){
    logit(mu[i]) <- beta[1]+beta[2]*x[i]
    y[i] ~ dbern(mu[i])
  }
}
```


y
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
x

Logistic + Exponential: Tree Mortality

- Logistic model of annual survival probability

$$logit(\rho) = X\beta$$

- Bernoulli likelihood

$$Y \sim Bern(\rho^t)$$

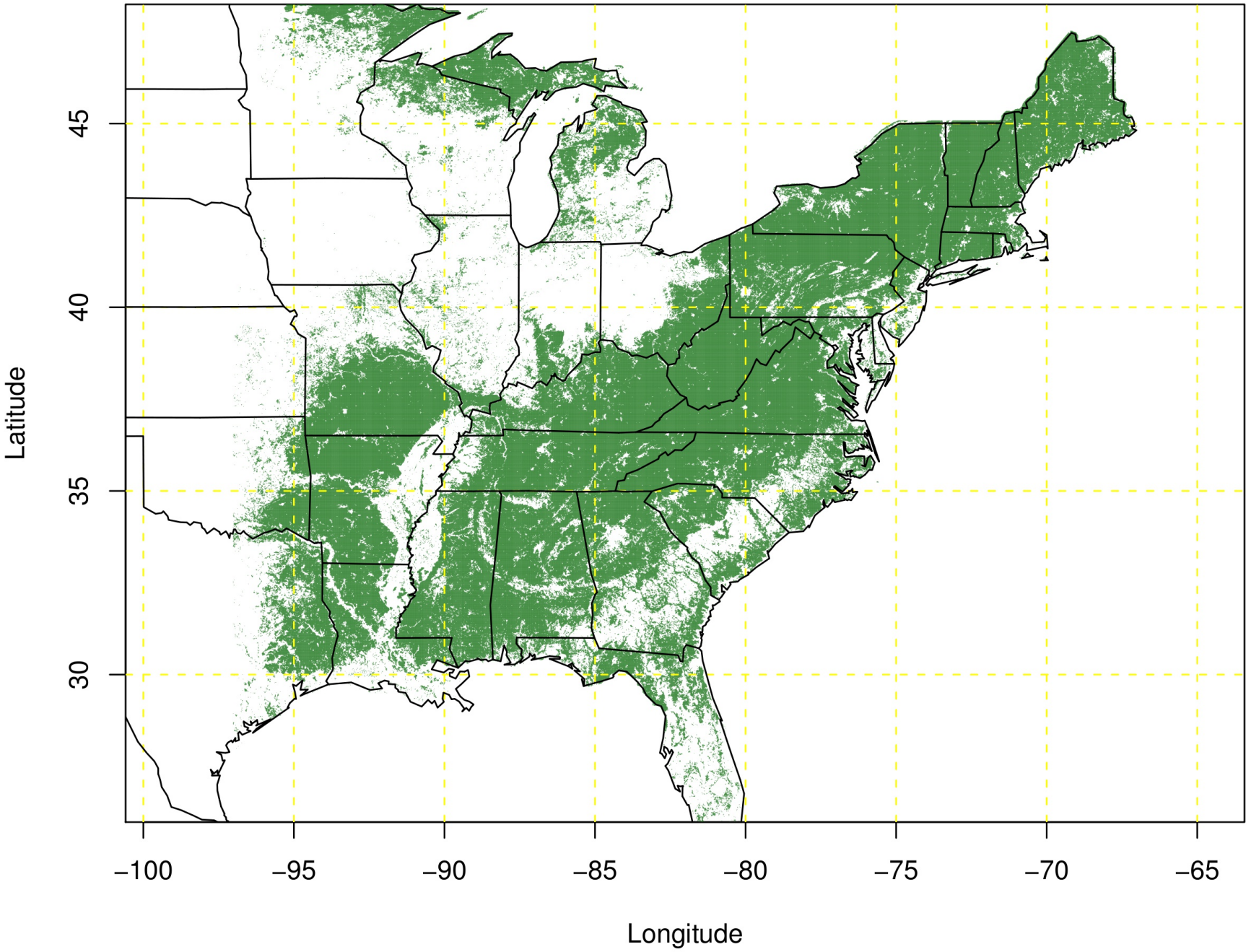
- Normal prior

$$\beta \sim N(B_{0,} V_0)$$

- Metropolis MCMC
  - 50K steps, 6-20K burn-in, thin 1/3

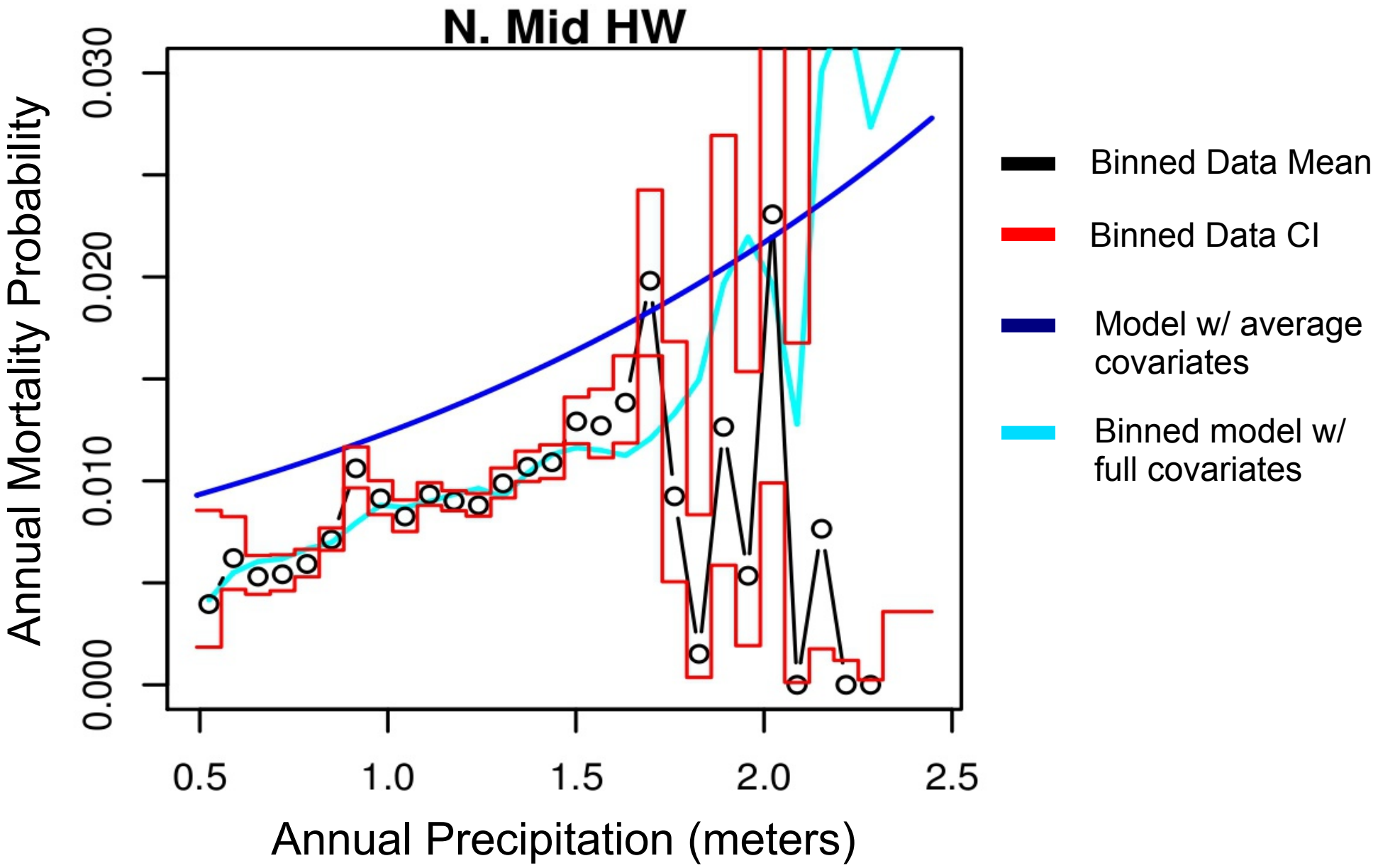
Dietze and Moorcroft 2011 Global Change Biology

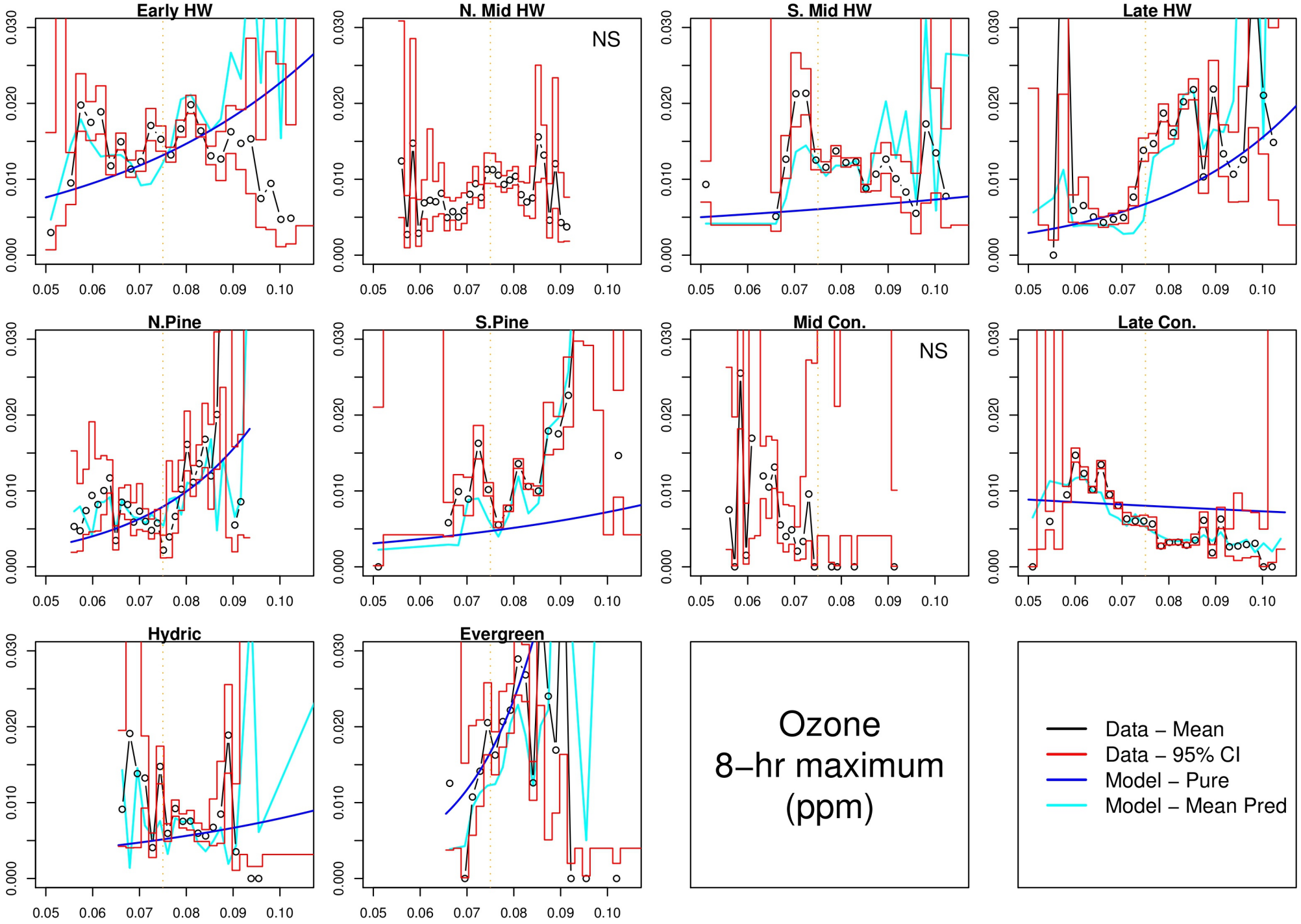
Distribution of Eastern Forest

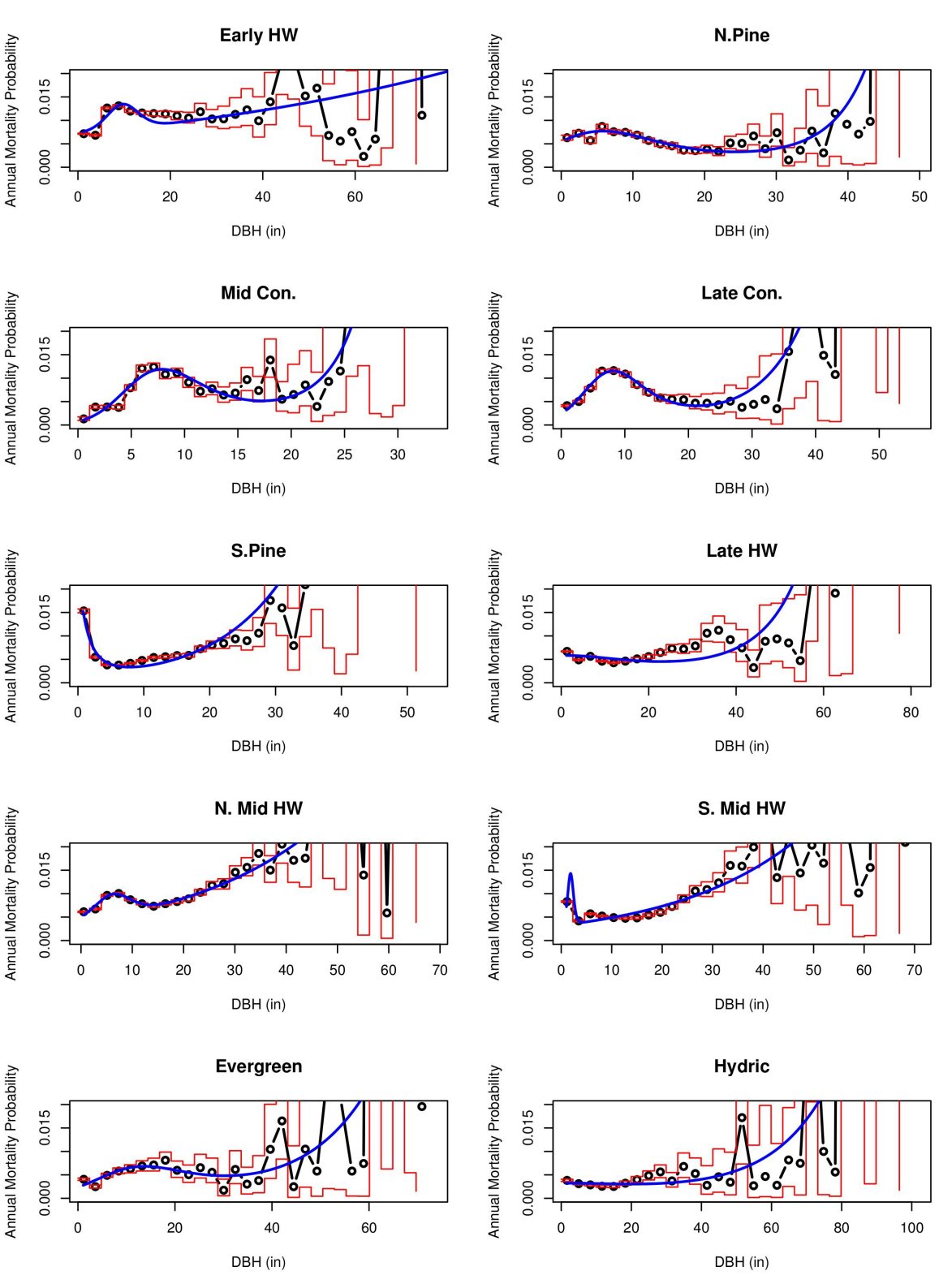


Mortality Covariates

- Climate
  - Precipitation
  - Summer Max Temp
  - Winter Min Temp
- Air Pollutants
  - NO3 deposition
  - SO4 deposition
  - Ozone
- Landscape/abiotic
  - Elevation
  - Slope
  - Radiation index
  - Topographic Moisture
- Stand scale biotic
  - DBH
  - Stand Basal Area
  - Stand Age


N. Mid HW
Annual Mortality Probability
0.000
0.010
0.020
0.030
0.5
1.0
1.5
2.0
2.5
Annual Precipitation (meters)
Binned Data Mean
Binned Data CI
Model w/ average covariates
Binned model w/ full covariates


Early HW
N. Mid HW
NS
S. Mid HW
Late HW
N.Pine
S.Pine
Mid Con.
NS
Late Con.
Hydric
Evergreen
0.000
0.010
0.020
0.030
0.05
0.06
0.07
0.08
0.09
0.10
Ozone
8–hr maximum
(ppm)
Data – Mean
Data – 95% CI
Model – Pure
Model – Mean Pred


Early HW
N.Pine
Mid Con.
Late Con.
S.Pine
Late HW
N. Mid HW
S. Mid HW
Evergreen
Hydric
Annual Mortality Probability
DBH (in)
0.000
0.015

Survival Covariate Effects

		Climate			Pollution			Stand		Topography				
PFT	int	Precip	Tmin	Tmax	no3	so4	O3	age	BA	slope	rad	elev	TCI	n
Early HW	-	+	-	+	+	-	-	+	-	-	ns	+	-	78,578
N Mid HW	+	-	+	-	-	ns	ns	+	-	-	ns	-	ns	144,168
S Mid HW	-	+	-	+	+	-	-	+	-	-	+	+	-	117,437
Late HW	-	+	-	+	+	-	-	+	-	-	ns	+	-	78,635
N.Pine	+	-	+	-	+	-	-	ns	+	-	ns	-	ns	18,973
S. Pine	-	+	-	+	+	-	-	-	-	-	ns	-	-	159,476
Mid Con	+	+	ns	ns	+	ns	ns	-	ns	ns	ns	-	ns	10,620
Late Con	+	ns	+	+	+	+	+	+	ns	+	-	ns	ns	149,397
Evergreen	-	+	-	+	+	-	-	+	-	ns	ns	+	-	7,891
Hydric	-	+	-	ns	+	-	-	+	ns	ns	ns	ns	-	22,352

Parameter Sensitivity

	Climate			Pollution				Stand			Topography				
PFT	Precip	Tmin	Tmax	no3	so4	NxS	O3	age	BA	dia	slope	rad	elev	TCI	MEAN
Early HW	7.70	3.97	13.34	12.99	21.46	7.48	2.37	1.52	2.50	9.40	1.37	NS	1.56	1.97	6.17
N Mid HW	1.82	1.56	1.63	2.02	NS	NS	NS	NS	1.97	6.00	0.71	NS	1.07	NS	1.35
S Mid HW	1.08	0.58	4.28	8.44	20.52	4.76	0.23	1.25	1.21	13.55	0.46	0.15	0.87	0.77	4.09
Late HW	1.50	0.83	3.71	18.18	38.02	9.19	1.78	2.08	1.50	14.42	0.93	NS	0.65	0.49	6.47
N.Pine	0.79	6.31	3.18	2.78	8.25	5.74	2.71	NS	0.36	1.87	0.74	NS	2.79	NS	2.29
S. Pine	0.91	0.77	3.52	5.54	42.48	4.06	0.37	1.59	1.87	29.56	0.36	NS	0.48	0.27	6.75
Mid Con	0.92	NS	NS	1.17	NS	NS	NS	1.38	NS	5.98	NS	NS	8.55	NS	1.38
Late Con	NS	0.91	1.59	1.20	1.00	1.99	0.26	1.64	NS	3.78	0.22	0.35	NS	NS	0.82
Evergreen	2.48	3.37	19.19	31.81	52.42	9.82	7.22	10.70	3.17	28.66	NS	NS	3.52	2.19	12.67
Hydric	1.05	0.99	NS	4.74	1.08	2.79	0.38	3.40	NS	24.92	NS	NS	NS	0.54	2.85
MEAN	1.82	1.93	5.04	8.89	18.52	4.58	1.53	2.43	1.26	13.81	0.48	0.05	1.95	0.62	

$$S_i = 1000 \cdot SD\left(Y_{pred} \middle| \bar{X}_{-i}\right)$$

Logit-Normal Model

- Bernoulli likelihood only accounts for *sampling* error
- Can add additional "extrabinomial" variation to account for the fact that the covariates do not full account for the mean risk

$$y \sim Bern(\theta)$$

$$logit(\theta) = X\beta + \epsilon$$

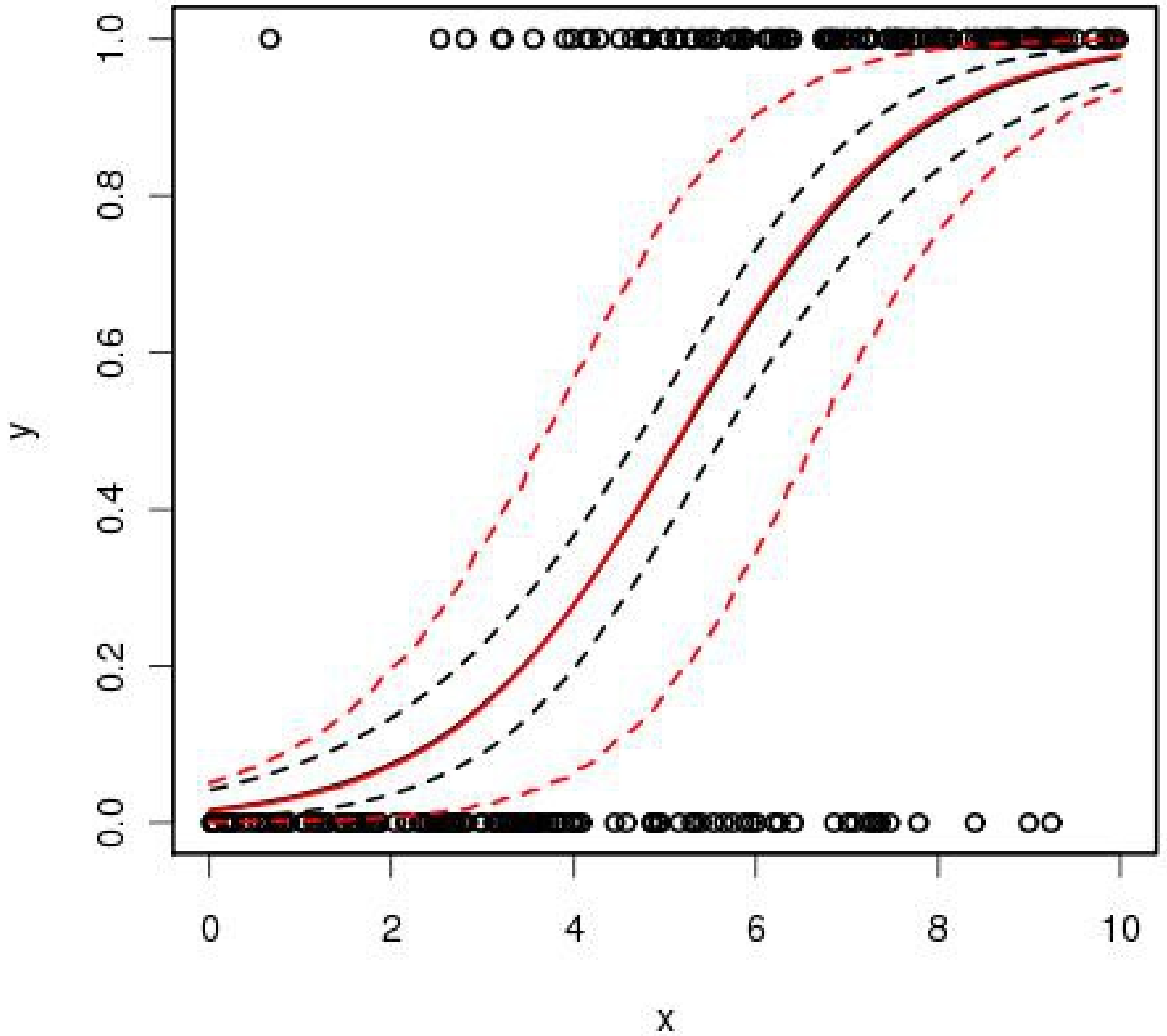
$$\epsilon \sim N(0, \sigma^2)$$

- Fairly sensitive to prior on sigma

Logit-Normal Regression

```
model <- function(){
  for(i in 1:2) { beta[i] ~ dnorm(0,0.01)}
  sigma ~ dgamma(1,0.1)

  for(i in 1:n){
    Ey[i] <- beta[1]+beta[2]*x[i]
    mu[i] ~ dnorm(Ey[i],sigma)
    logit(theta[i]) <- mu[i]
    y[i] ~ dbern(theta[i])
  }
}
```


y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

Multinomial Regression

- Categorical variable as a function of covariates in a linear model
  - Probability of falling within each group
  - Categories can be ordered (ordinal) or unordered (nominal)
- Multivariate extension of the logistic regression to >2 categories
  - “cumulative logit model”
- End up with K-1 regression models for K classes

Logistic Regression

$$y \sim \text{Binom}(1, \theta) = \binom{n}{y} \prod \theta^{y} (1-\theta)^{n-y}$$

$$logit(\theta) = \log\left(\frac{\theta}{1-\theta}\right) = X\beta$$

Multinomial Regression

$$y \sim Multinom(1, \theta_{1,}, \ldots, \theta_{K}) = \binom{n}{y_1 \cdots y_K} \prod_{k=1}^{K} \theta_k^{y_k}$$

For the k^{th} class:

$$\log\left(\frac{\theta_1 + \cdots + \theta_k}{1-(\theta_1 + \cdots + \theta_k)}\right) = X\beta_k$$

Inverting the link function for the FIRST class

$$\theta_1 = \frac{\exp(X\beta_1)}{1+\exp(X\beta_1)}$$

For the middle classes

$$\theta_k = \frac{\exp(X\beta_k)}{1+\exp(X\beta_k)} - \sum_{j=1}^{k-1} \theta_j$$

For the LAST class

$$\theta_K = 1 - \sum_{j=1}^{K-1} \theta_j$$

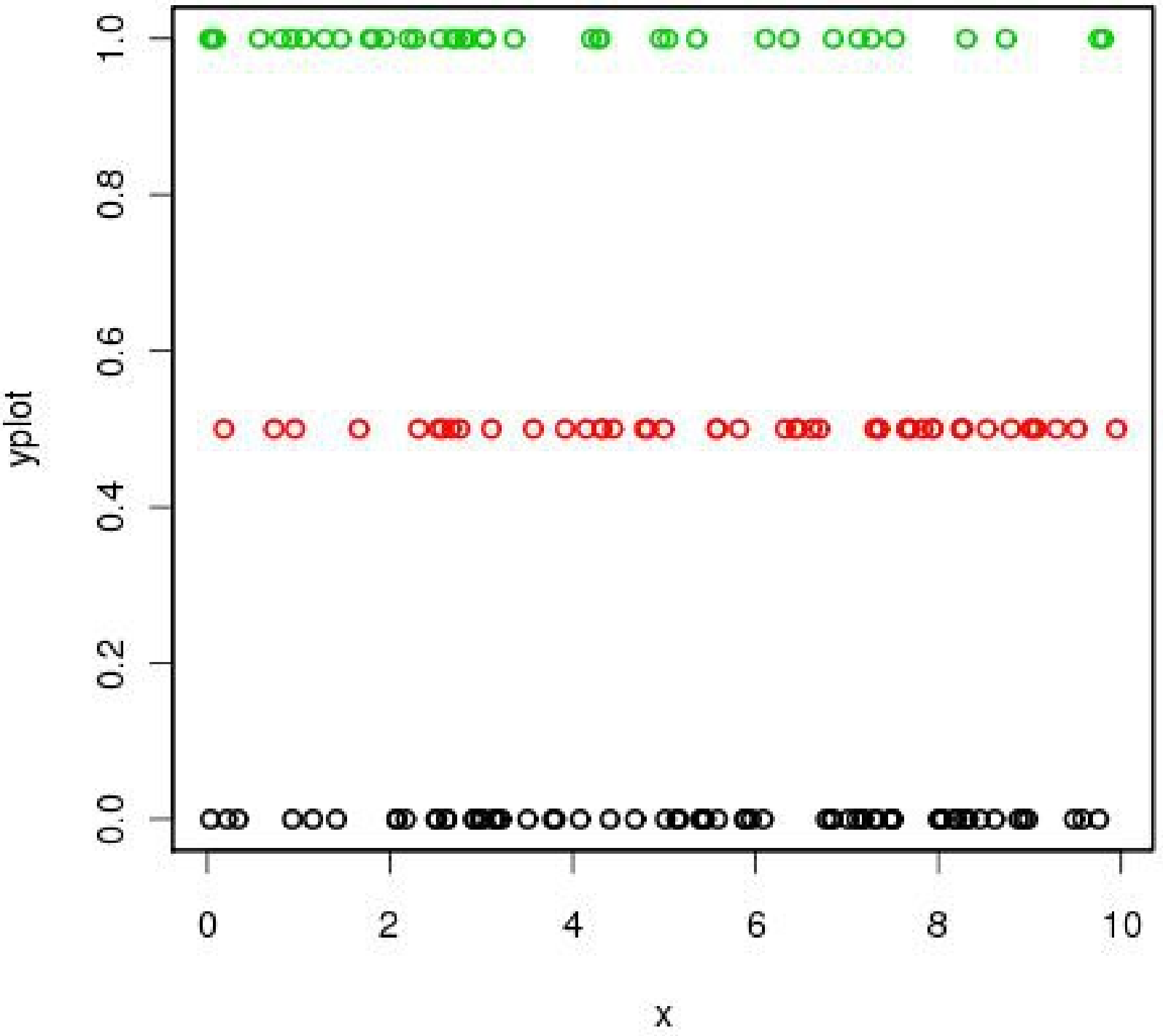
Additional Restrictions

- Because the regression model for each of the K-1 categories is expressed in terms of the cumulative probability of all categories up to that one, these curves cannot cross each other
  - Slopes and intercepts must be ordered
  - $\cap_{0,k} < \cap_{0,k+1}$ and $\cap_{1,k} < \cap_{1,k+1}$ for all k
- Range restrictions can be accommodated in the prior using an indicator function $\mathbf{I}(A)$
- $\mathbf{I}(A) = 1$ if A is true and $\mathbf{I}(A) = 0$ if A is false

Multinomial Regression Priors

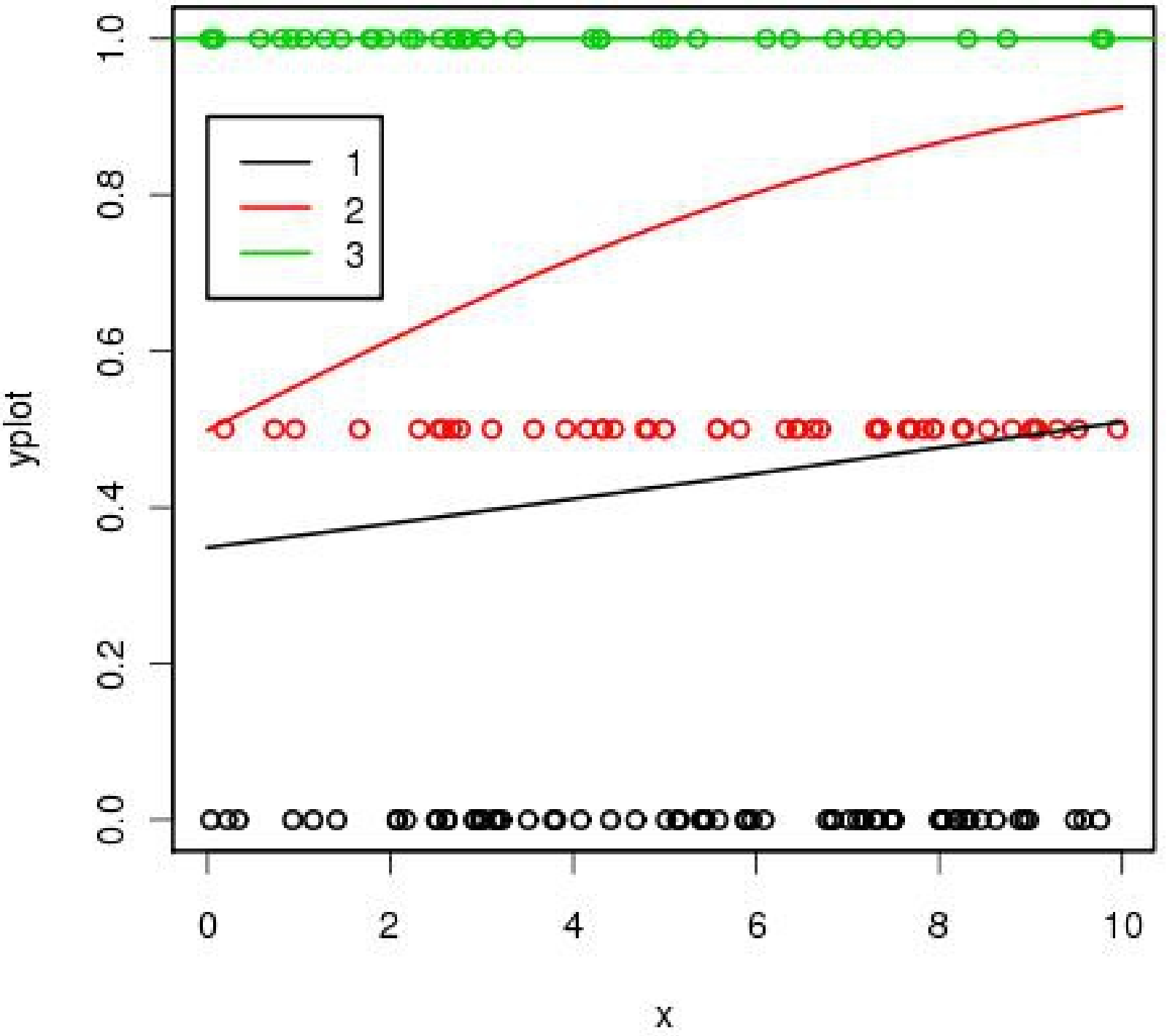
$$\beta_{0,k} \sim \begin{cases} N(b_{0,1}, V_b) I(\beta_{0,1} < \beta_{0,2}) & k = 1 \\ N(b_{0,k}, V_b) I(\beta_{0,k-1} < \beta_{0,k} < \beta_{0,k+1}) & 1 < k < K-1 \\ N(b_{0,K-1}, V_b) I(\beta_{0,K-2} < \beta_{0,K-1}) & k = K-1 \end{cases}$$

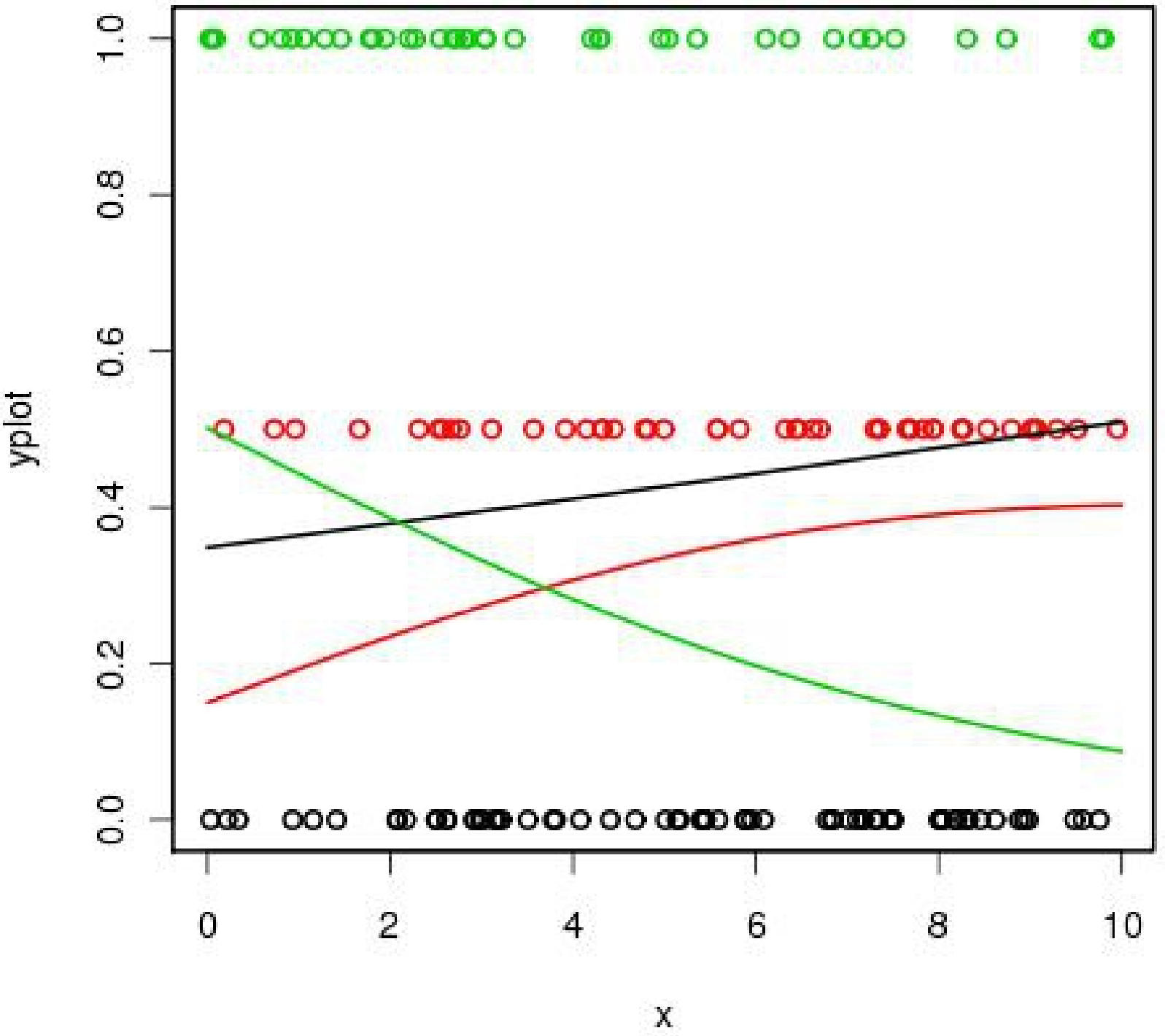
$$\beta_{1,k} \sim \begin{cases} N(b_{1,1}, V_b) I(\beta_{1,1} < \beta_{1,2}) & k = 1 \\ N(b_{1,k}, V_b) I(\beta_{1,k-1} < \beta_{1,k} < \beta_{1,k+1}) & 1 < k < K-1 \\ N(b_{1,K-1}, V_b) I(\beta_{1,K-2} < \beta_{1,K-1}) & k = K-1 \end{cases}$$


yplot
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
x

Multinomial JAGS code

```
model{
    beta[1, 1] ~ dnorm(0.0, 0.001) T(,beta[2, 1])
    beta[2, 1] ~ dnorm(0.0, 0.001)
    beta[1, 2] ~ dnorm(0.0, 0.001) T(,beta[2, 2])
    beta[2, 2] ~ dnorm(0.0, 0.001)
    for (i in 1:n) {
        logit(mu[i, 1]) <- beta[1, 1] + beta[1, 2] * x[i]
        logit(cmu2[i]) <- beta[2, 1] + beta[2, 2] * x[i]
        mu[i, 2] <- cmu2[i] - mu[i, 1]
        mu[i, 3] <- 1 - cmu2[i]
        y[i, ] ~ dmulti(mu[i, ], 1)
    }
}
```


yplot
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
0
2
4
6
8
10
x


yplot
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
x

Assumptions of Linear Model

- Homoskedasticity **Model variance**
- No error in X variables **Errors in variables**
- No missing data **Missing data model**
- Normally distributed error **GLM**
- Error in Y variables is measurement error
- Observations are independent

Example

- Are banana slugs susceptible to soil mercury contamination?
- Data:
  - Y = slug counts, n= 35
  - X^{O} = soil mercury concentration (0-5000 μg/kg)
    - Lost samples 9, 13, 27
    - Manufacture reports a 10% accuracy
    - Further digging reveals this refers to a 95% CI which was calibrated on n=150
- Graph? JAGS Code?