Assumptions of Linear Model: Part II

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

Latent Variables

- Variables that are not directly observed
- Values are inferred from model
  - Parameter model: prior on value
  - Data and Process models provide constraint

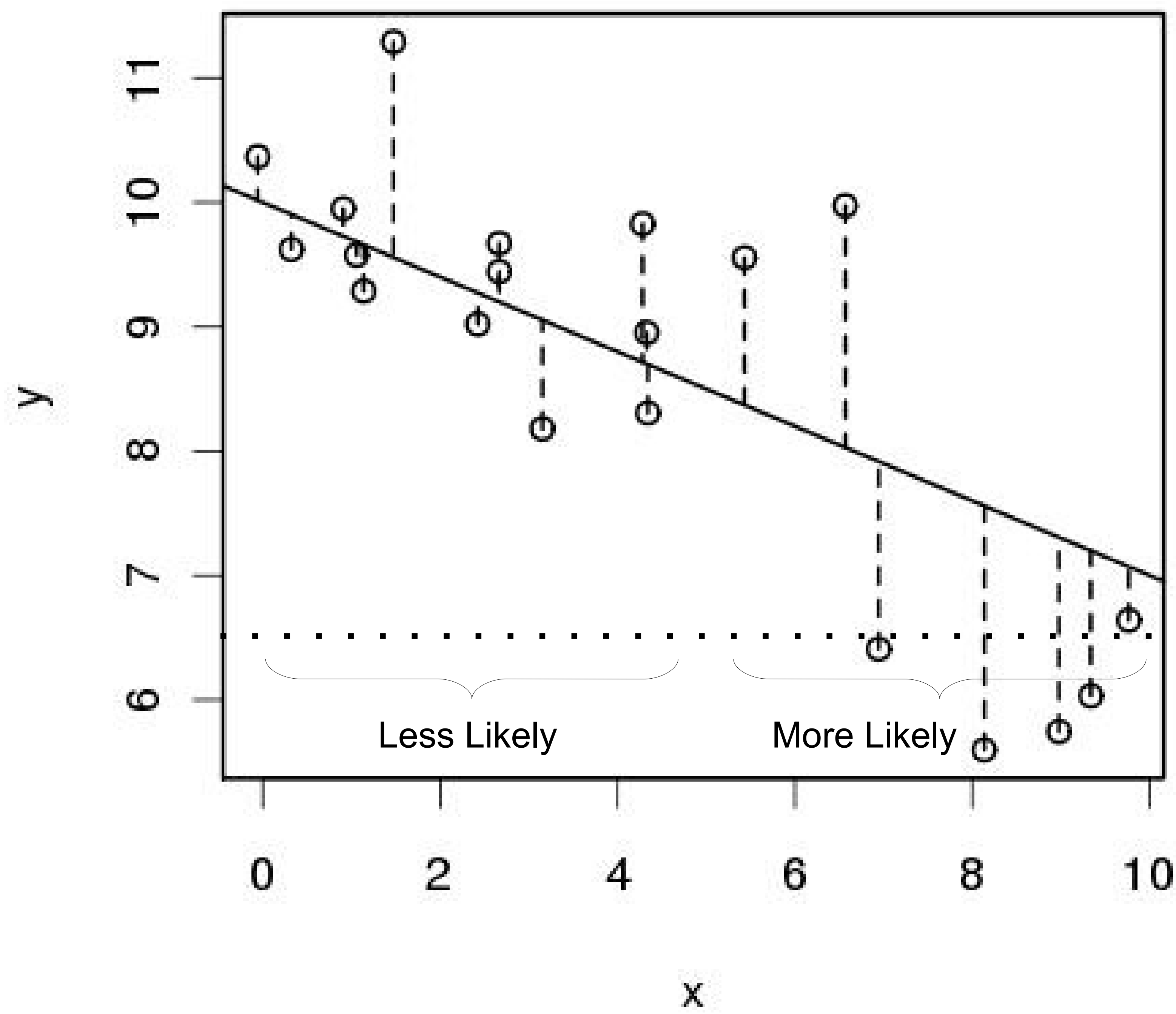
$$p(\boldsymbol{X}|\ldots) \propto N(y|\beta_0+\beta_1 x, \sigma^2) N(x^{(o)}|x, \tau^2) N(x|X_{0,} V_X)$$

- MCMC integrates over (by sampling) the values the unobserved variable could take on
- Contribute to uncertainty in parameters, model
- Ignoring this variability can lead to falsely <u>overconfident</u> conclusions

Missing data models

$$\vec{y} \sim N(\boldsymbol{X}\vec{\beta}, \sigma^2)$$

- Let's assume a standard multiple regression model (homoskedastic, no error in X)
- If some of the y's are missing
  - Can just predict the distribution of those values using the model PI
- What if some of the X's are missing
  - The observed y is more likely to have come from some values of X than others



y
x
11
10
9
8
7
6
0
2
4
6
8
10
Less Likely
More Likely

Missing Data

$$\mu = X\beta$$ **Process model**

$$y \sim N(\mu, \sigma^2)$$ **Data model for y**

$$\vec{\beta} \sim N(B_{0,} V_B)$$ **Prior for beta**

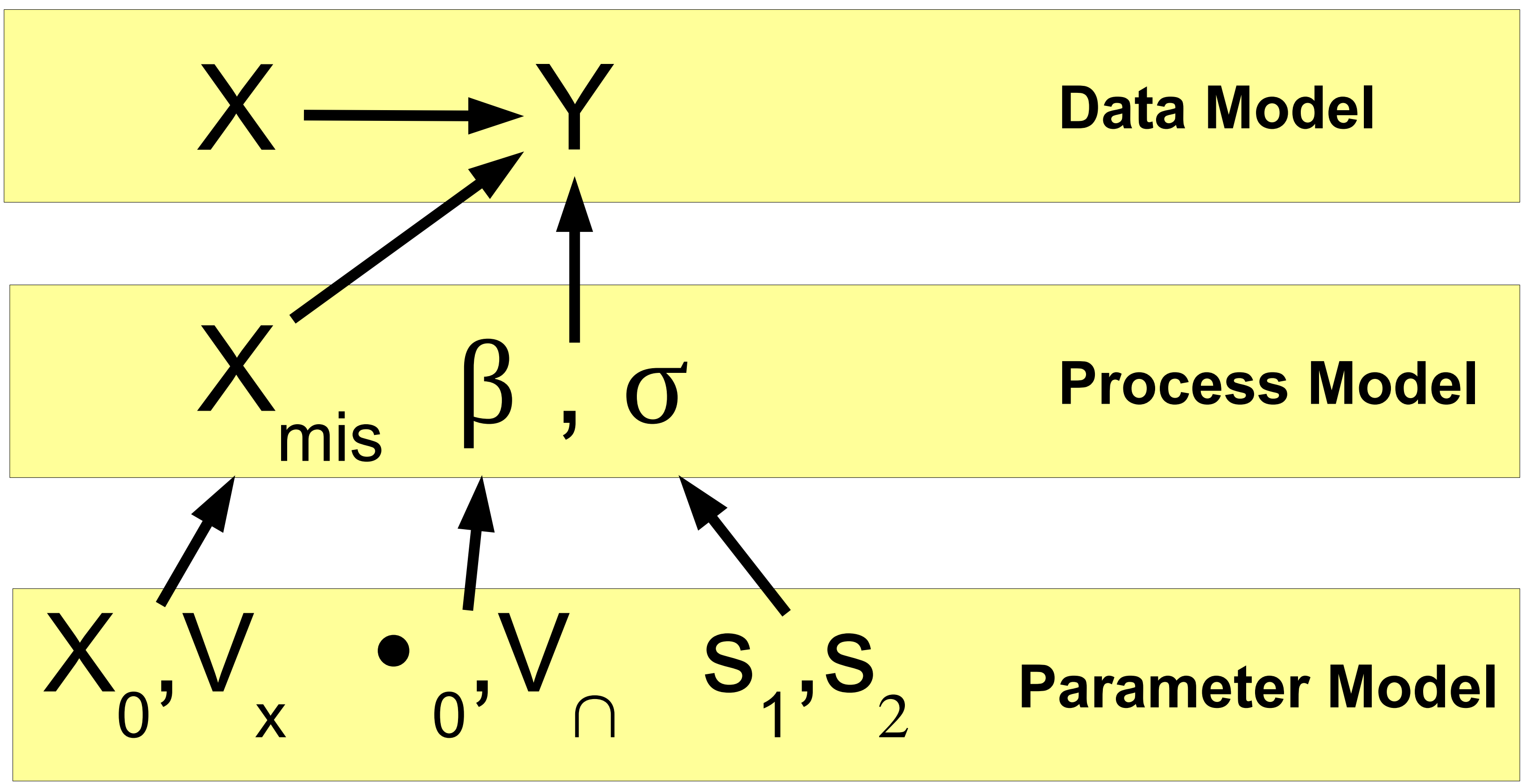
$$\sigma^2 \sim IG(s_{1,} s_2)$$ **Prior for sigma**

$$x_{mis} \sim N(X_{0,} V_X)$$ **Prior for missing X**

$$p(x_{mis}|\ldots) \propto N(Y|X\beta, \sigma^2) N(x|X_{0,} V_X)$$

Missing Data Model

$$\vec{y} \sim N(\boldsymbol{X}\vec{\beta}, \sigma^2)$$



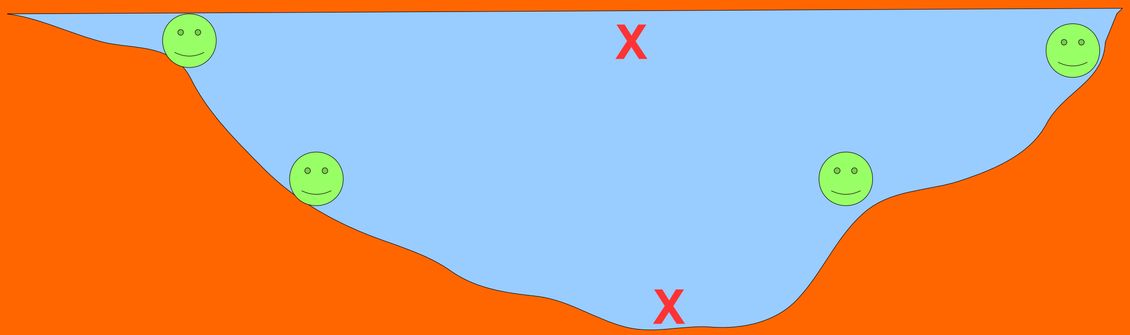
Conceptually within the MCMC

- Update the regression model based on ALL the rows of data conditioned on the current values of the missing data
- Update the missing data based on the current regression model and the values that all other covariates take on

- Overall, integrate over the uncertainty in missing X's
- Model uncertainty increases, but less so than if whole rows of data was dropped (partial info.)

ASSUMPTION!!

- Missing data models assume that the data is *missing at random*
- If data is missing SYSTEMATICALLY it can not be estimated



JAGS example: Simple Regression

```
model{
  ## priors
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
  sigma ~ dgamma(0.1,0.1)
  for(i in mis) { x[i] ~ dunif(0,10)}

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

Vector giving indices of missing values

X	Y
4.68	8.46
2.95	8.55
9.09	7.01
8.15	9.06
1.76	11.38
4.23	9.12
7.73	7.3
2.43	8.02
6.46	8.45
4.06	8.95
2.42	9.62
0.6	9.15
8.17	7.51
0.22	10.8
4.93	9.78
2.99	10.71
8.36	8.89
6.4	8.21
8.17	6.22
6.46	5.4
1.82	10.05
9.52	7.96
2.44	9.63
6.84	7.05
7.42	8.73
NA	7.5

JAGS example: Simple Regression

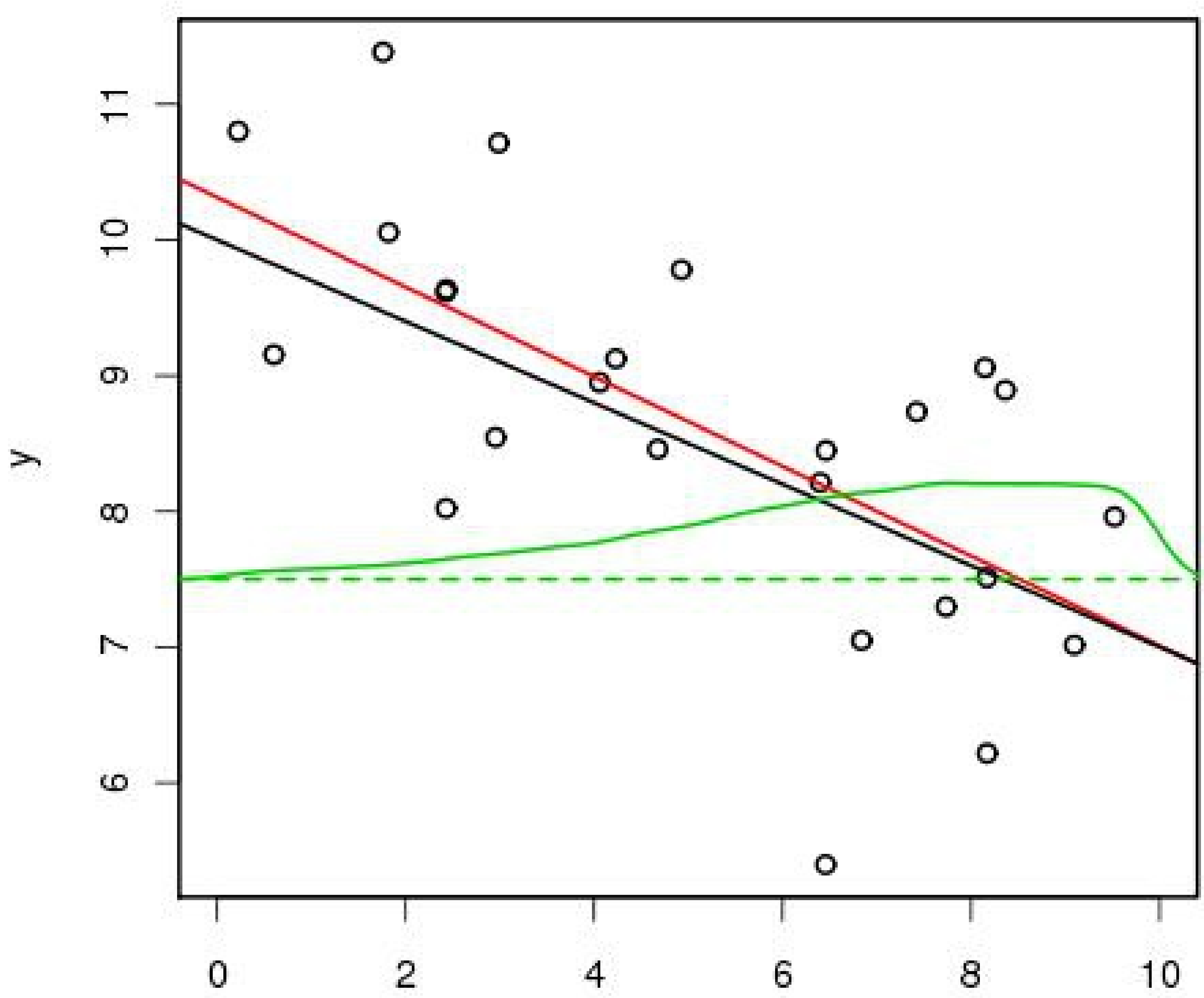
```
model{
  ## priors
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
  sigma ~ dgamma(0.1,0.1)
  for(i in mis) { x[i] ~ dunif(0,10)}

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

Vector giving indices of missing values

X	Y
4.68	8.46
2.95	8.55
9.09	7.01
8.15	9.06
1.76	11.38
4.23	9.12
7.73	7.3
2.43	8.02
6.46	8.45
4.06	8.95
2.42	9.62
0.6	9.15
8.17	7.51
0.22	10.8
4.93	9.78
2.99	10.71
8.36	8.89
6.4	8.21
8.17	6.22
6.46	5.4
1.82	10.05
9.52	7.96
2.44	9.63
6.84	7.05
7.42	8.73
NA	7.5

Example



Assumptions of Linear Model

- Homoskedasticity **Model variance**
- No error in X variables **Errors in variables**
- No missing data **Missing data model**
- **Normally distributed error**
- 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

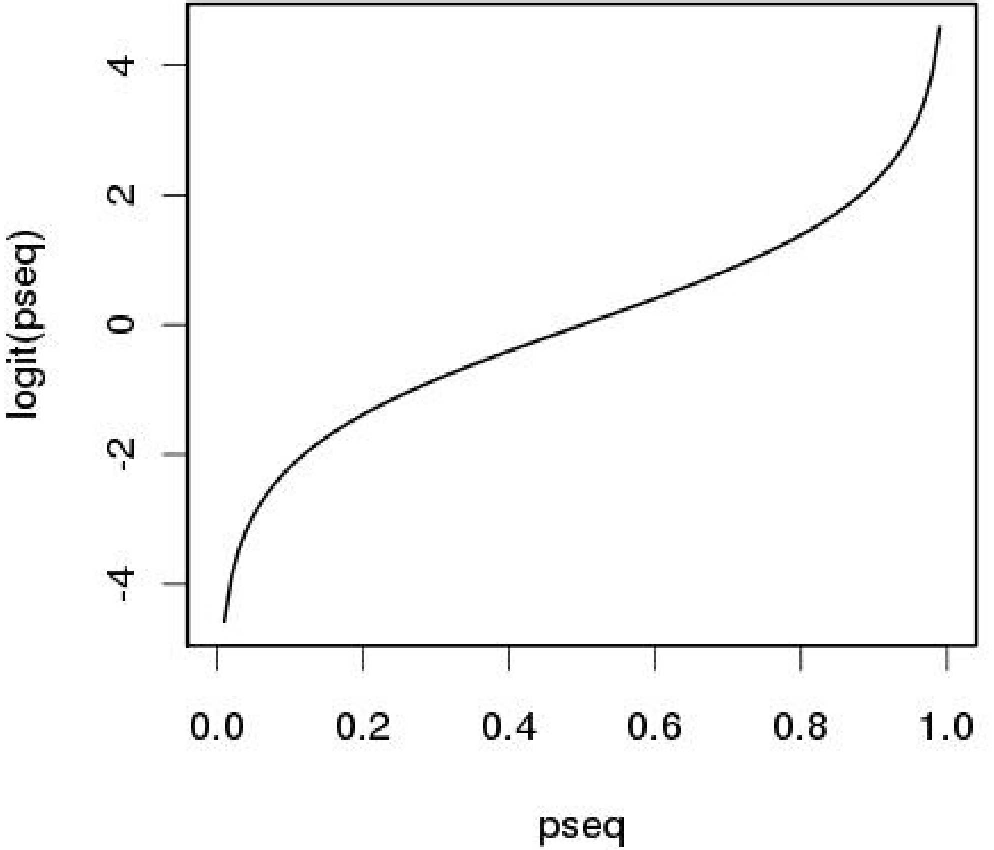
- "Canonical" Link Functions

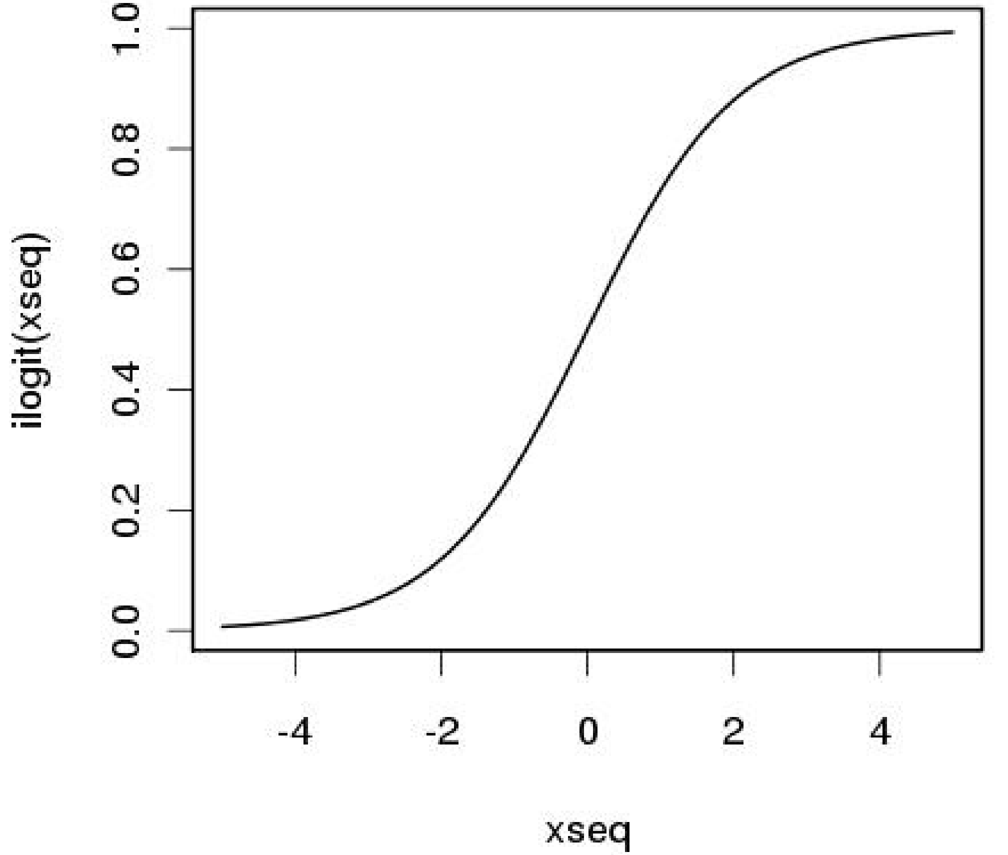
Distribution	Link Name	Link Function	Mean Function
Normal	Identity	$Xb = \mu$	$\mu = Xb$
Exponential	Inverse	$Xb = \mu^{-1}$	$\mu = (Xb)^{-1}$
Gamma			
Poisson	Log	$Xb = \ln(\mu)$	$\mu = \exp(Xb)$
Binomial	Logit	$Xb = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(Xb)}{1+\exp(Xb)}$
Multinomial			

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

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

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





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

$$y \sim Bern(\theta)$$ **Data Model**

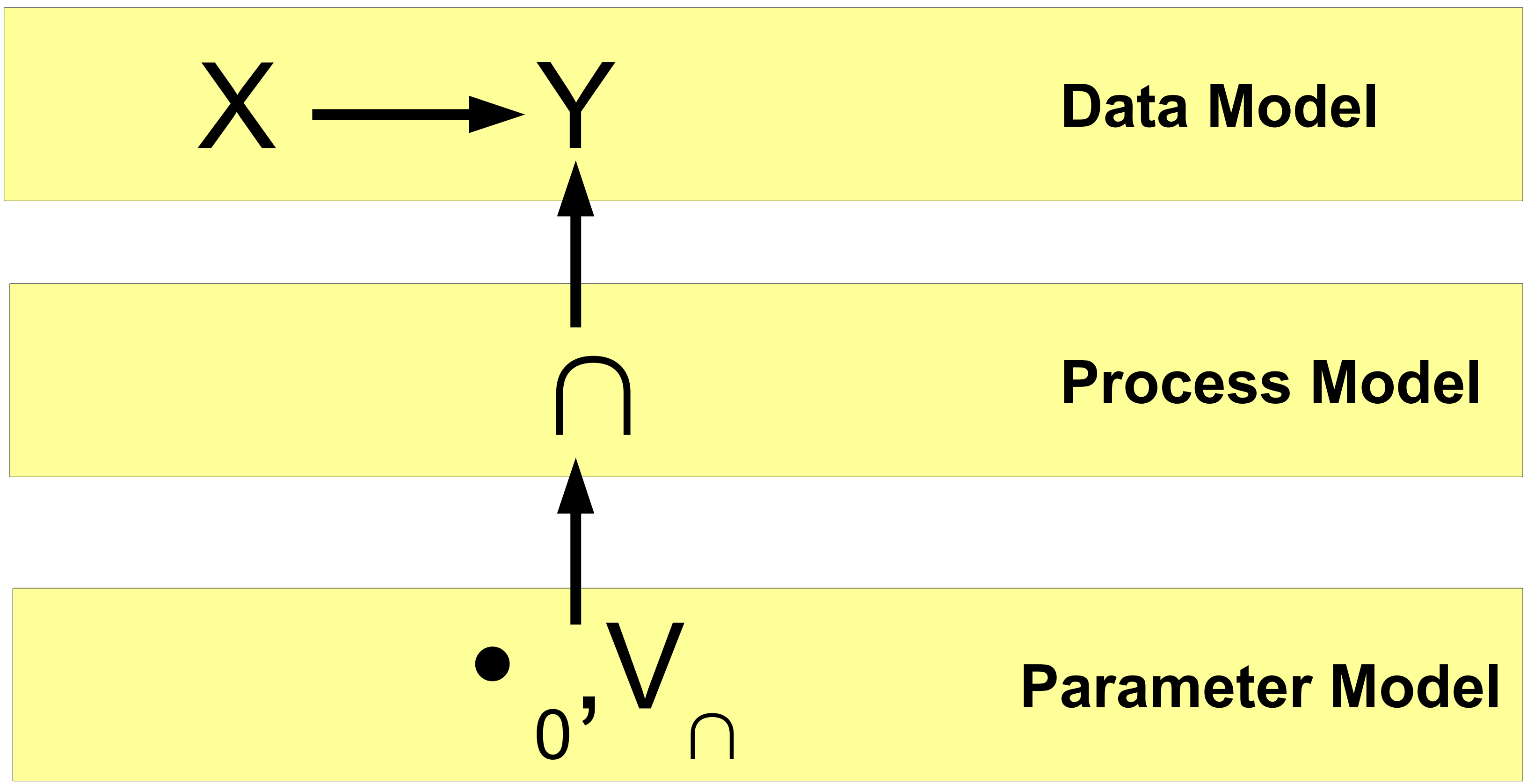
$$logit(\theta) = X\beta$$ **Process Model**

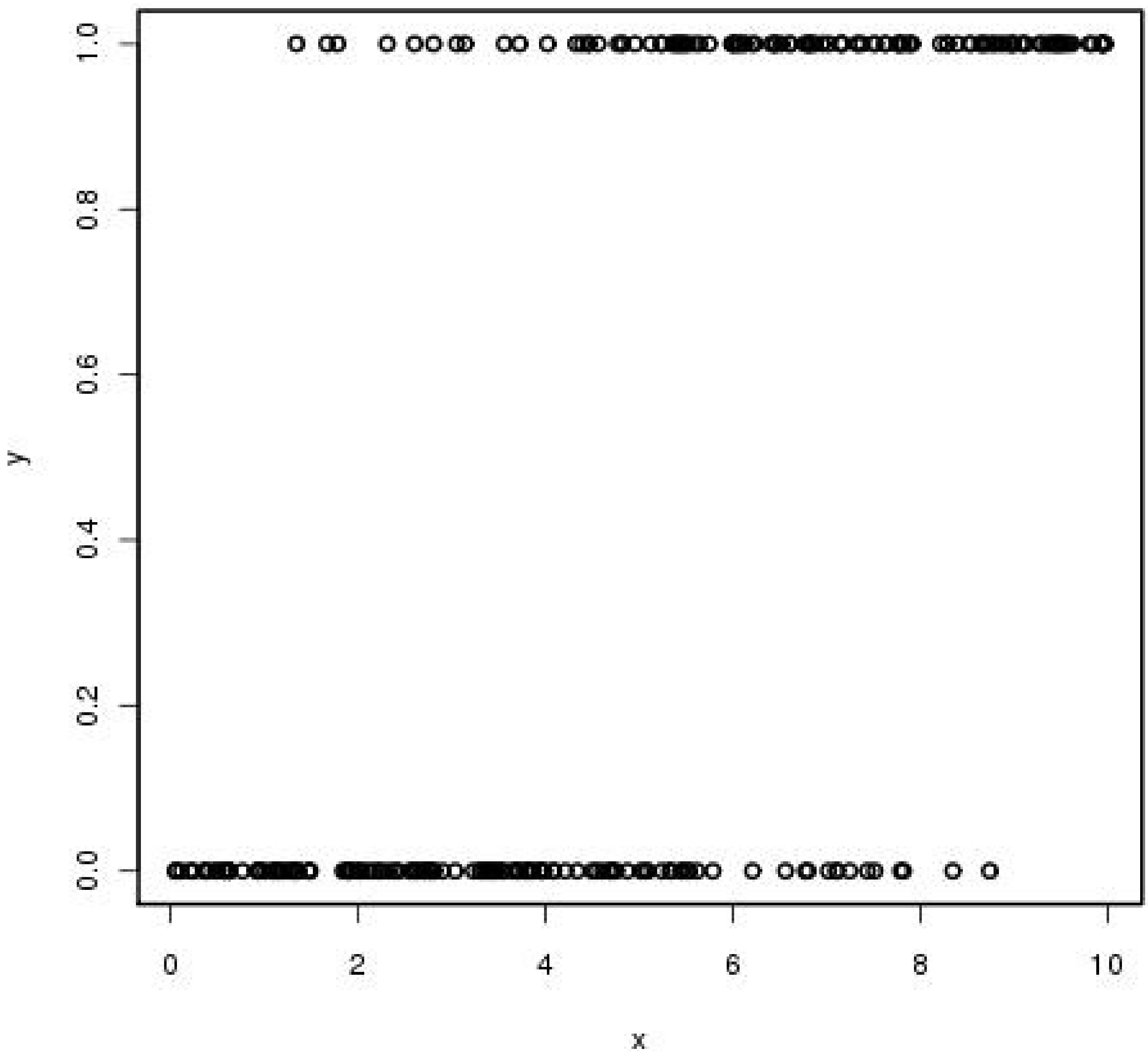
- Bayesian

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

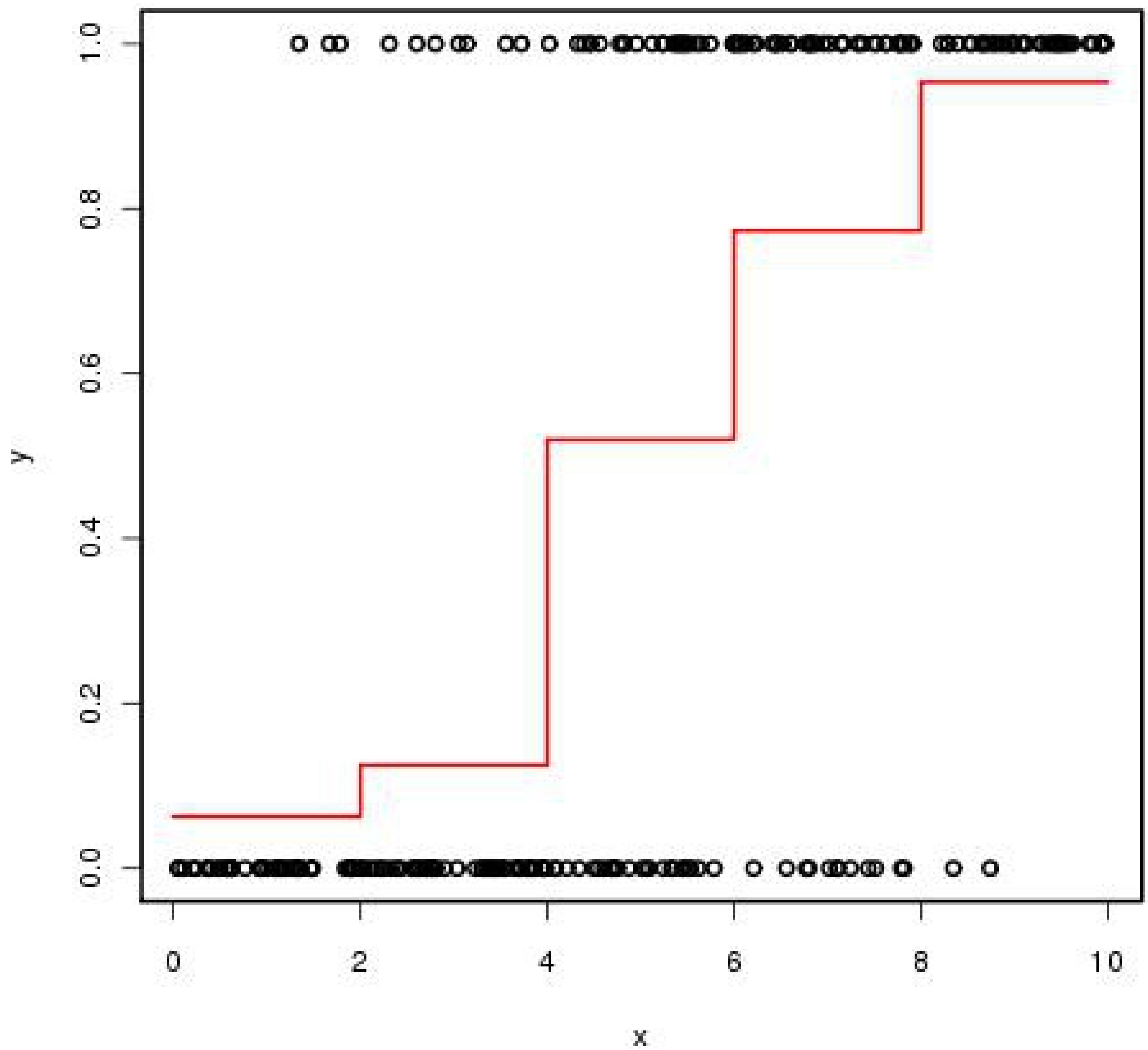
Logistic Regression

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





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



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

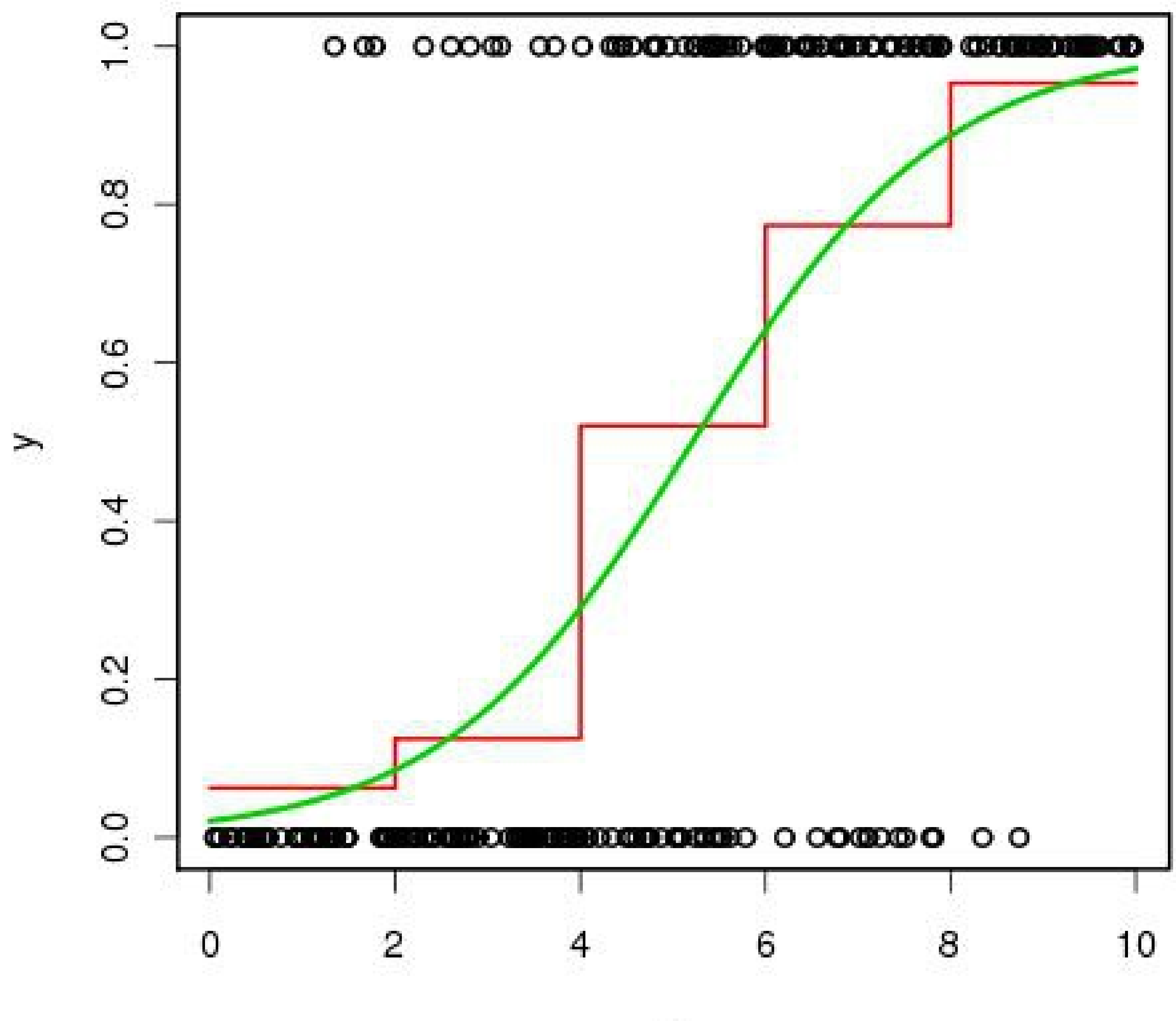
Logistic Regression in R

- 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)
}
```



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

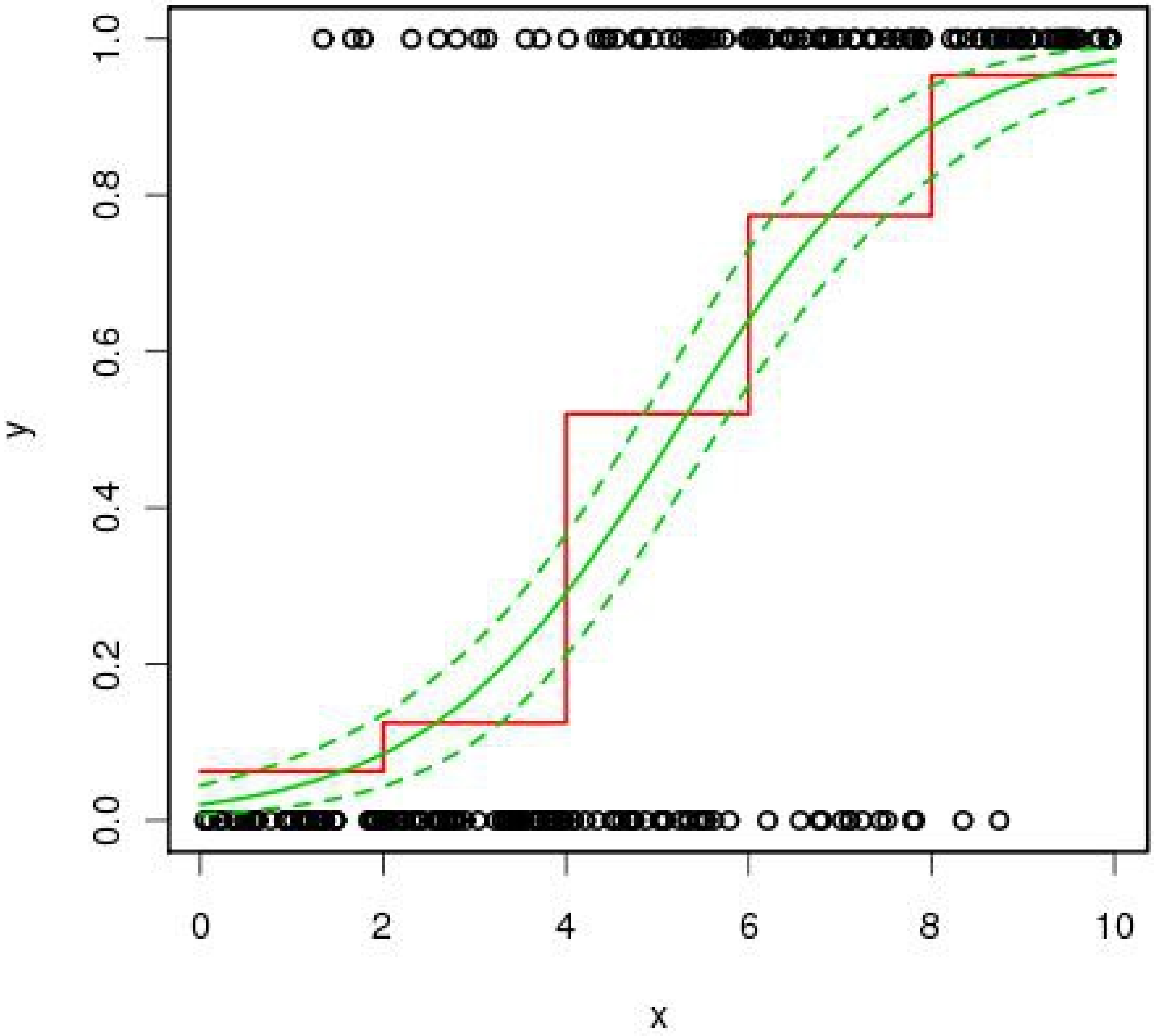
```
Call:
glm(formula = y ~ x, family = binomial())

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
```

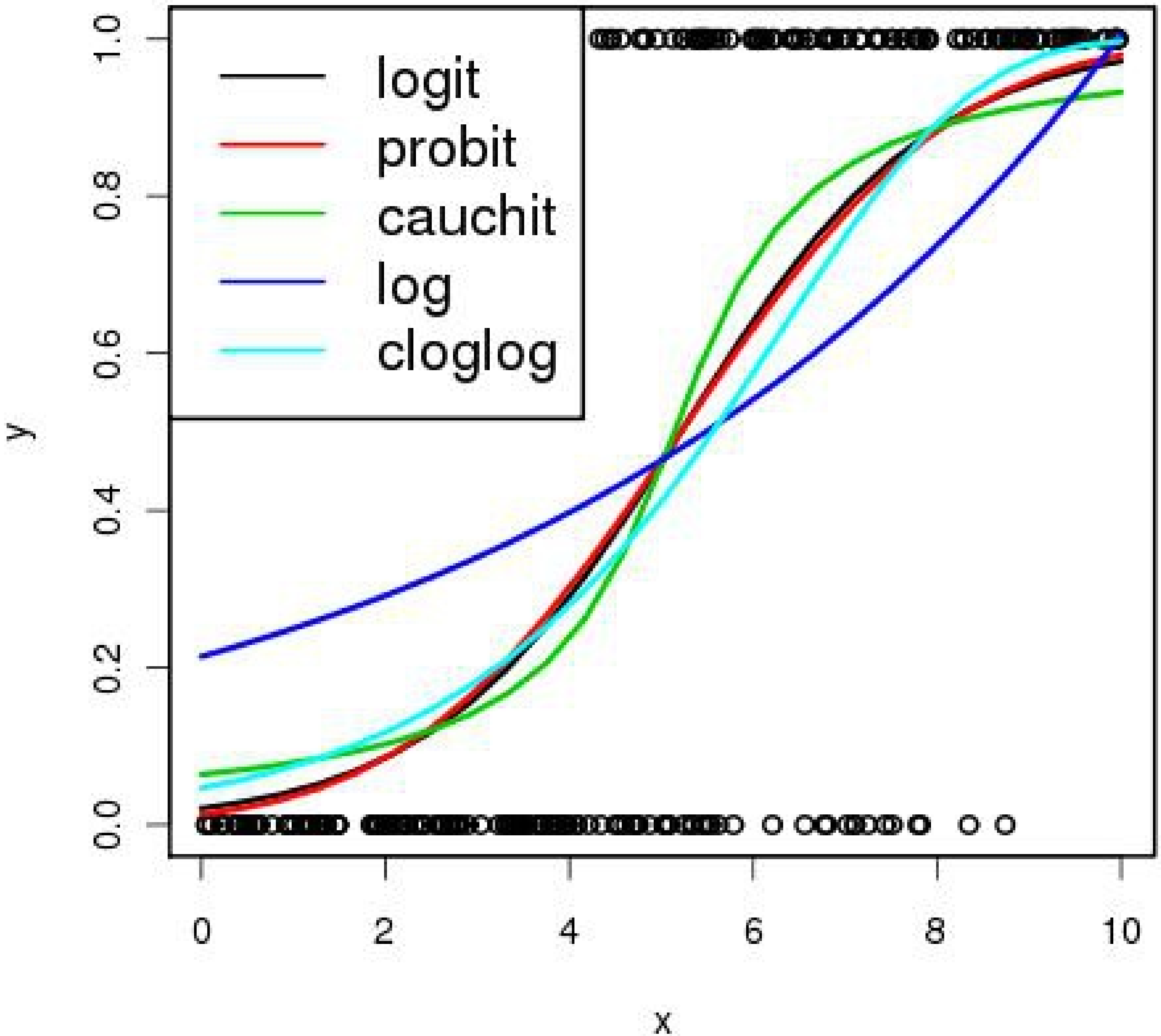


1.0
0.8
0.6
0.4
0.2
0.0
y
0
2
4
6
8
10
x

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

Coming next...

- GLM
  - Bayesian Logistic
  - Poisson Regression
  - Multinomial
- Continuing our exploration of relaxing the assumptions of linear models