Other useful models

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Models so far
- Other GLMs
- Other useful models
- Some other approaches (**tomorrow**)

Covered GLMs

- Linear regression
- Binomial regression
- Poisson regression
- Negative-binomial regression (for fixed dispersion)

Other models

- Log-normal regression
- Inverse Gaussian regression
- Gamma regression
- Tweedie regression
- beta-binomial regression
- Multinomial regression
- Ordinal regression
- Zero-inflated regression
- Zero-truncated regression
- Hurdle models
- Conway-Maxwell-Poisson regression
- Beta regression

is EF and GLM, is EF but not GLM, is not EF or GLM

Positive continuous data

Data bounded at zero $y_i > 0$, and might or might not include zeros.

- Usually (right) skewed
- Variance increases with mean

E.g.,:

- Biomass
- Size measurements
- Nitrogen in soil
- Amount of rainfall
- Insurance claims

Options:

- Log-normal regression
- Inverse Gaussian GLM
- Gamma GLM
- **Tweedie GLMs**

Log-normal regression

Connected to the normal distribution, in the sense:

$$\begin{aligned} \log(y_i) &\sim \mathcal{N}(\mu_i, \sigma^2) \\ y_i &\sim \mathcal{L}ognormal(\mu_i, \sigma^2) \end{aligned} \tag{1}$$

- So, easy to fit!
- Two parameters: μ_i and σ^2
- $\mathbb{E}(y_i|x_i) = \exp(\mu_i + \frac{\sigma^2}{2})$
- $\text{var}(y_i|x_i) = \{\exp(\sigma^2) - 1\} \exp(2\mu_i + \sigma^2)$

We are not modeling the mean but the location parameter μ_i

Inverse Gaussian GLM

- Sharper peak and heavier tails than log-normal
- Often used in presence of extreme events

$$\frac{1}{\mathbb{E}(y_i|\mathbf{x}_i)^2} = \alpha + \mathbf{x}_i\boldsymbol{\beta} \qquad \mathbb{E}(y_i|\mathbf{x}_i) = \frac{1}{(\alpha + \mathbf{x}_i\boldsymbol{\beta})^2} \tag{2}$$

- parameters: location μ_i and dispersion ϕ
- canonical link: squared inverse
- alternatively inverse, log-link

$$\text{var}(y_i|\mathbf{x}_i) = \frac{\mu_i^3}{\phi}$$

Gamma GLM

- (negative) Inverse link is canonical
  - provide starting positive values `mustart`
- Log-link is more stable

Tweedie distributions

Many of the distributions so far are special forms of the *Tweedie family* of distributions:

$$\text{var}(y_i|\mathbf{x}_i) = \mu_i^{\xi}\phi$$

- Normal distribution $\xi = 0$
- Poisson distribution $\xi = 1$
- Gamma distribution $\xi = 2$
- Inverse Gaussian distribution $\xi = 3$

With Tweedie we can:

- Analyse positive continuous data (with zeros!)
- Counts

Our example data: soil nitrogen

Originally by Lane et al. (2002)

- Soil nitrogen in kilograms per hectare
- Nitrogen fertilizer dose in kilograms per hectare

Fert	Source	SoilN
0	0	4.53
0	0	5.46
0	0	4.77
48	0	6.17
48	0	9.30
48	0	8.29
96	0	11.30
96	0	16.58

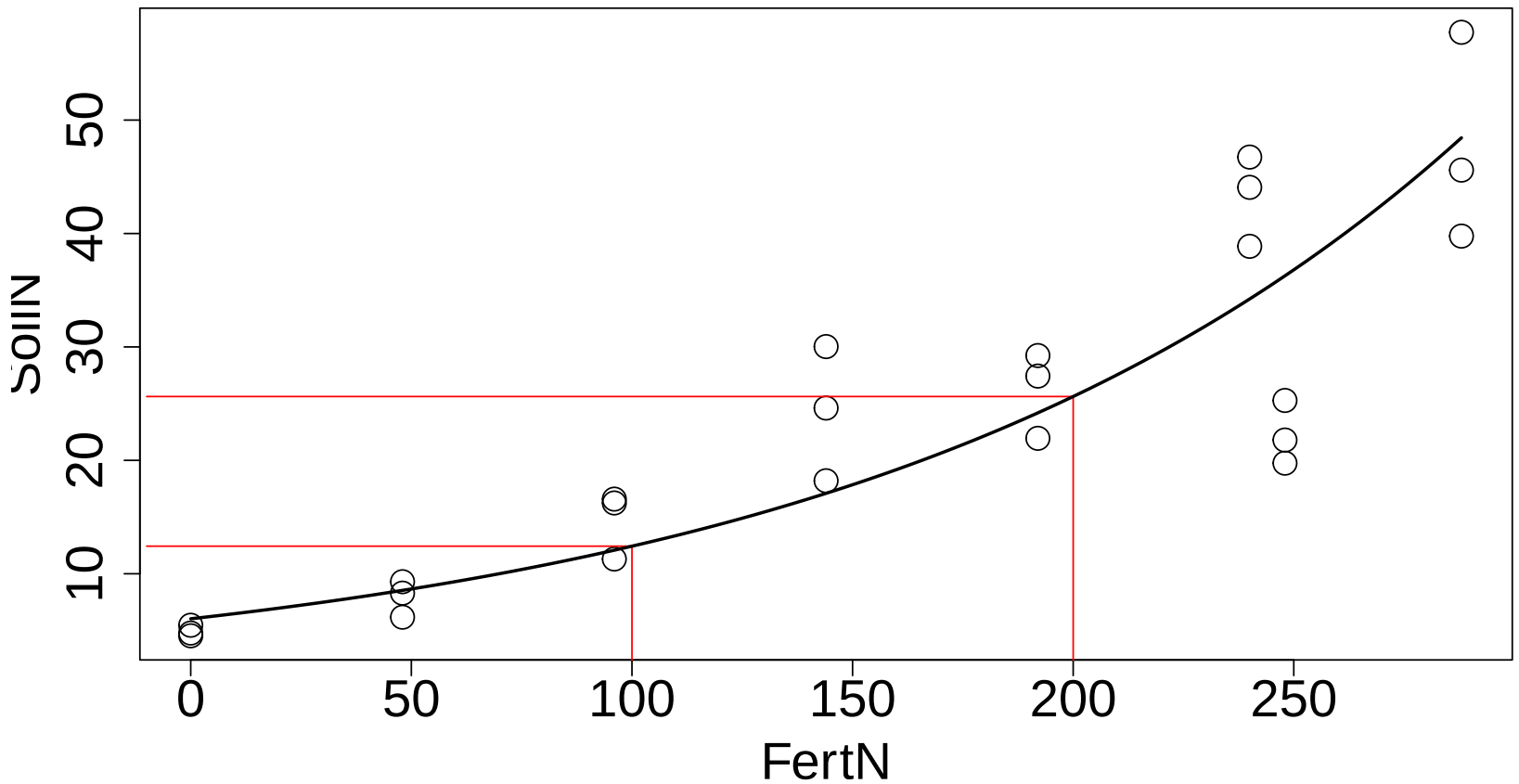
Log-normal regression in R

```
model1 <- lm(log(SoilN)~Fert, data = nitrogen)
fitted <- exp(predict(model1)) #We need to backtransform
exp(cbind(est=coef(model1), confint(model1)))
```

```
##                    est    2.5 %   97.5 %
## (Intercept) 6.030274 4.729154 7.689367
## Fert        1.007261 1.005932 1.008591
```

- Soil nitrogen increases with supplied nitrogen
- $\log(2)/0.007234334 \approx 96$
- Soil nitrogen doubles for almost every 100 kg/hectare fertilizer

Log-normal regression: results



Data-transformed linear regression

- I don't generally recommend response transformations
- Works poorly for a range of data-types
- GLMs are there to avoid this sort of thing
- But it usually works well for data with a large mean

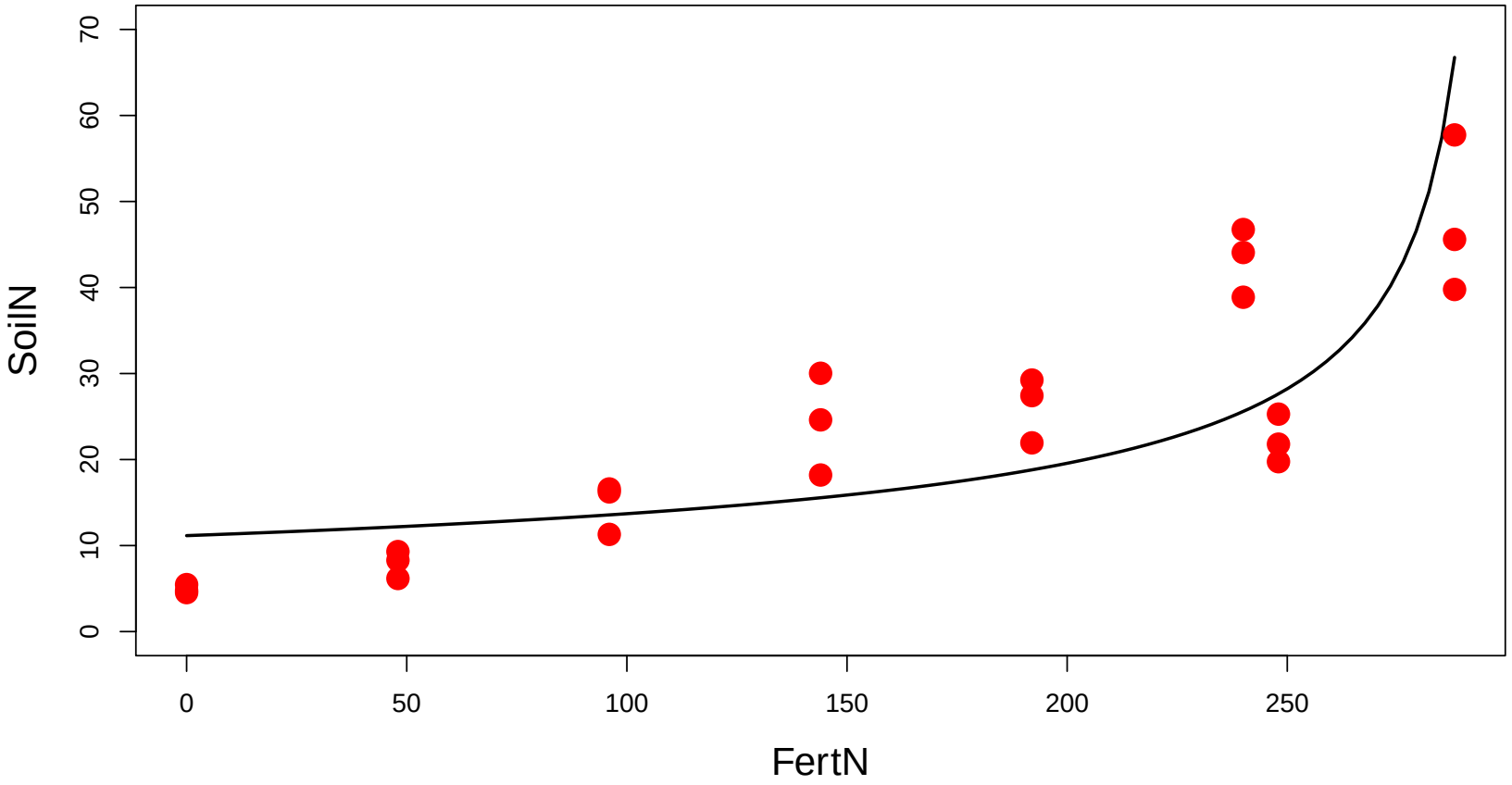
Inverse Gaussian GLM in R

```
model2 <- glm(SoilN~Fert, data = nitrogen, family = "inverse.gaussian")
summary(model2)
```

```
##
## Call:
## glm(formula = SoilN ~ Fert, family = "inverse.gaussian", data = nitrogen)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.039e-03  1.728e-03   4.652 0.000123 ***
## Fert        -2.714e-05  6.202e-06  -4.375 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.01367622)
##
##     Null deviance: 0.72554  on 23  degrees of freedom
## Residual deviance: 0.37002  on 22  degrees of freedom
## AIC: 185.12
##
## Number of Fisher Scoring iterations: 6
```

Interpretation is -very- difficult with inverse-type link functions

Inverse Gaussian GLM: results



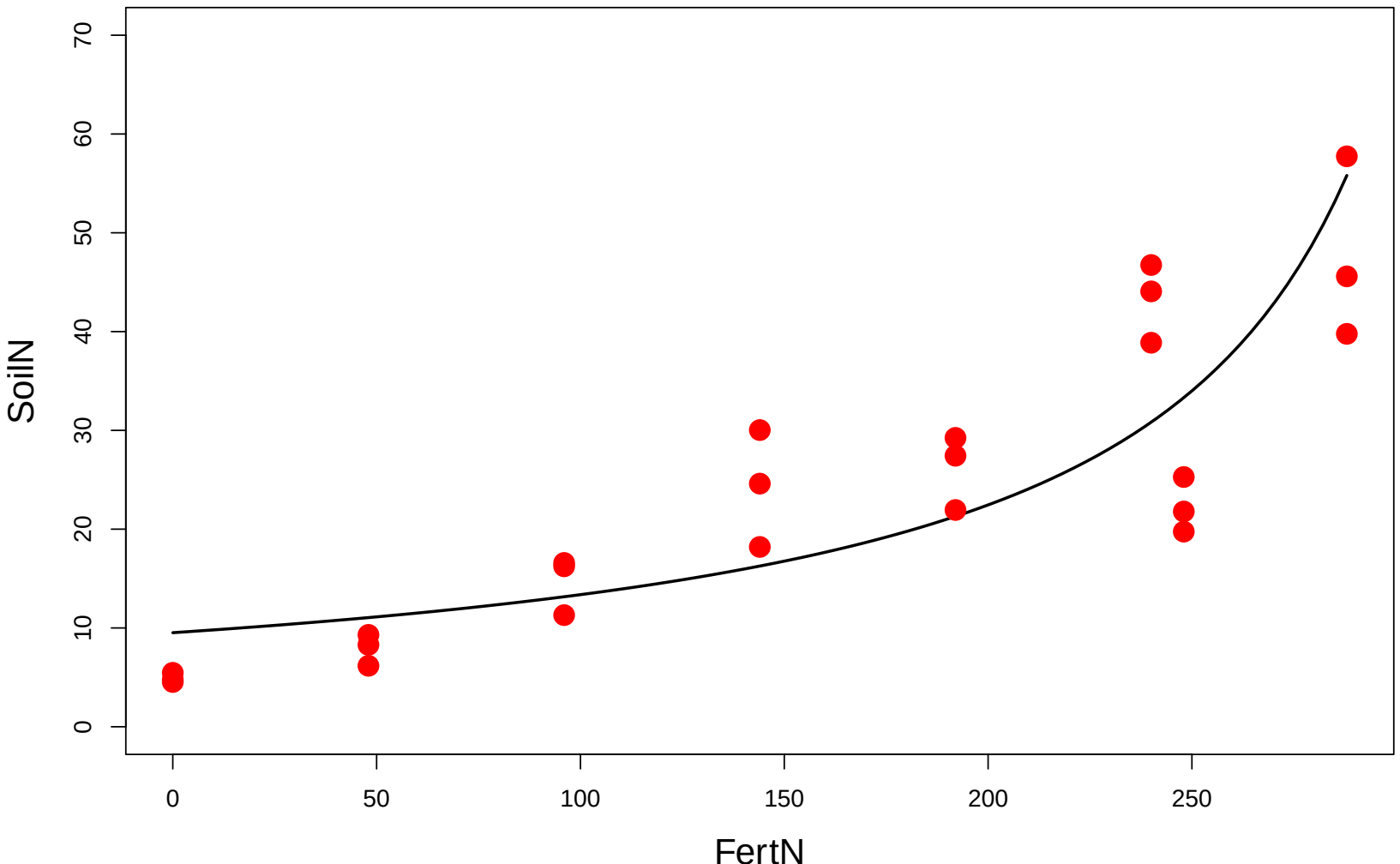
Gamma GLM

(note, uppercase Gamma)

```
model3 <- glm(SoilN~Fert, data = nitrogen, family = "Gamma")
summary(model3)
```

```
##
## Call:
## glm(formula = SoilN ~ Fert, family = "Gamma", data = nitrogen)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.050e-01  1.170e-02   8.977 8.28e-09 ***
## Fert        -3.025e-04  4.605e-05  -6.569 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1512009)
##
##     Null deviance: 11.5897  on 23  degrees of freedom
## Residual deviance:  3.4555  on 22  degrees of freedom
## AIC: 168.93
##
## Number of Fisher Scoring iterations: 5
```

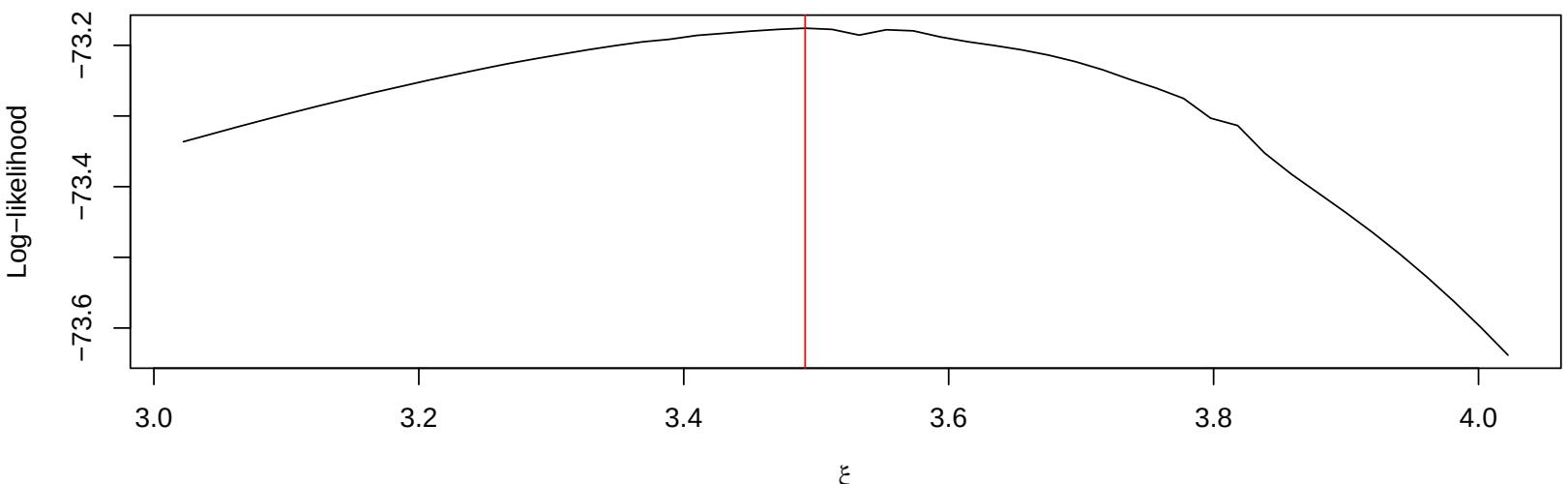
Gamma GLM: results



Tweedie GLM in R: finding ξ

```
## 2.01 2.06 2.11 2.16 2.21 2.26 2.31 2.36 2.41 2.46 2.51 2.5
## .....................................Done.
```

```
## 3.022245 3.032245 3.042245 3.052245 3.062245 3.072245 3.08
## ..................................................
```

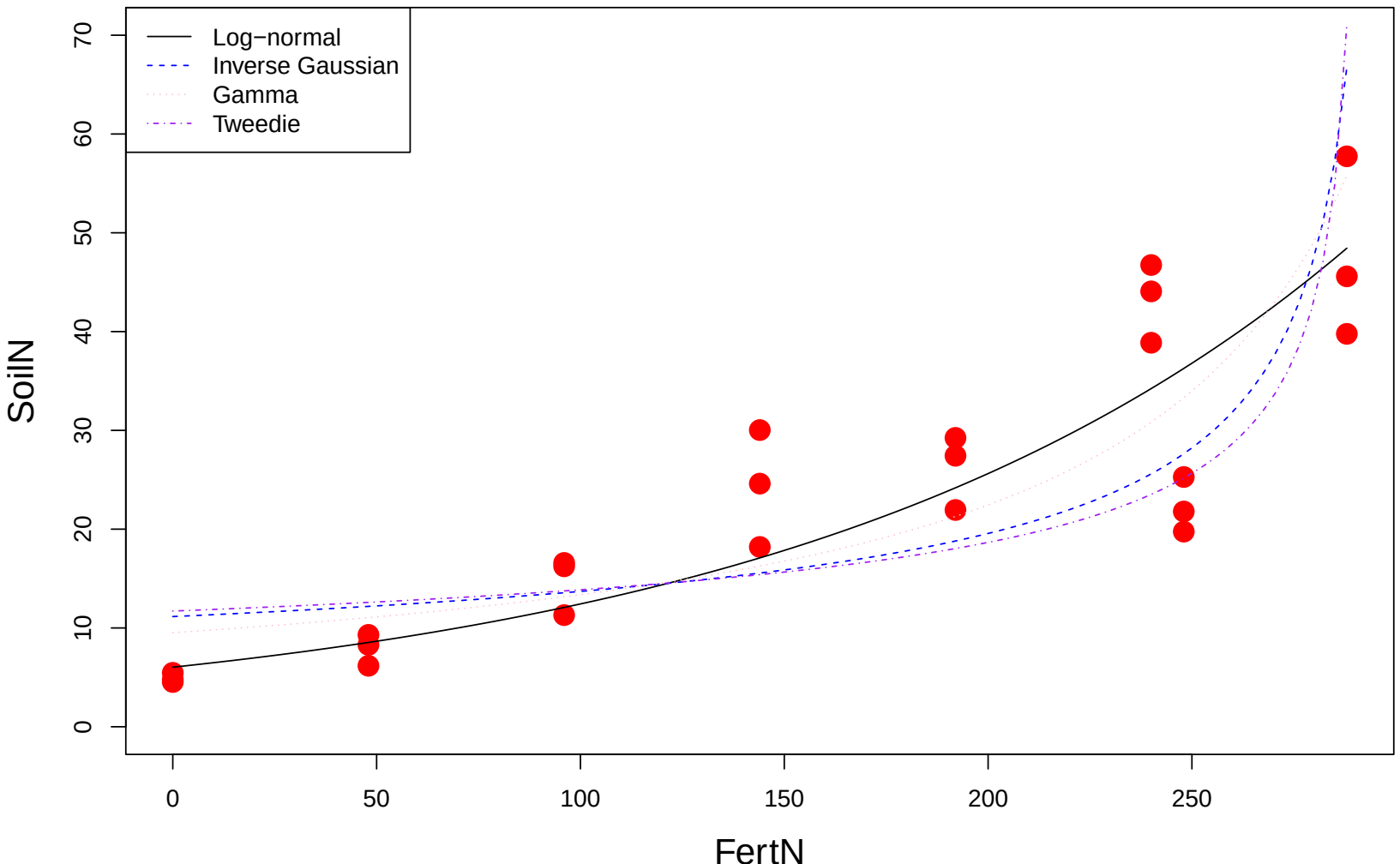


Tweedie GLM in R: fitting

```
model4 <- glm(SoilN~Fert, data = nitrogen, family = statmod::tweedie(var.power = out.est2$xi.max))
summary(model4)
```

```
##
## Call:
## glm(formula = SoilN ~ Fert, family = statmod::tweedie(var.power = out.est2$xi.max),
##     data = nitrogen)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.175e-03  6.002e-04   3.624  0.00150 **
## Fert        -7.467e-06  2.113e-06  -3.535  0.00186 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Tweedie family taken to be 0.004092795)
##
##     Null deviance: 0.20212  on 23  degrees of freedom
## Residual deviance: 0.12582  on 22  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

All together



Conclusion

- They can be relatively similar, but link function matters.
- **Mind the tail behavior**

Tweedie for flexible modeling of positive data (potentially with zeros)

Categories

Binomial models assume **dichotomy**
Alternative: **polychotomy**, multiple groups

Multinomial

Gives the number of successes for a set of outcomes (instead of either success or fail)

Figure 1: We roll a dice many times, and count the number of times it hits each side. If we only throw the dice once, the throw follows a **categorical distribution**.

Multinomial logistic regression

The model we assume is:

$$\begin{aligned} \text{pr}(y_{ij}) = \pi_{ij} &= \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} \\ \log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \eta_{ij} &= \mathbf{x}_{ij}\boldsymbol{\beta}_j \end{aligned} \tag{3}$$

- E.g., probability of observing species at a site.
- We could fit multiple binomial models
  - Gets a bit messy: $y_{i1}|N_i \sim \mathcal{B}in(\pi_{i1}; N_i)$
  - $y_{i2}|N_i - y_{i1} \sim \mathcal{B}in(\pi_{ij}; N_i - y_{i1})$
- Equivalent to using a Poisson regression with fixed total

Cumulative link models, such as:

For ordered data

-
-
-

Ordinal: proportional odds

More commonly, the categories are ordered.

$$\text{pr}(y_i \leq j) = \sum_{k=1}^{j} \text{pr}(y_k) = \frac{\exp(\theta_j + \boldsymbol{\beta}^\top \mathbf{x}_i)}{1 + \exp(\theta_j + \boldsymbol{\beta}^\top \mathbf{x}_i)} \tag{4}$$

- E.g., cover classes

Ordinal: proportional odds

Referred to as "proportional odds" because:

$$\frac{\pi_{ij}/(1-\pi_{ij})}{\pi_{i+1j}/(1-\pi_{i+1j})} = \exp\{\boldsymbol{\beta}(\mathbf{x}_{ij} - \mathbf{x}_{i+1j})\} \tag{5}$$

the log-odds is proportional to the difference in the covariates.

Ordinal example: vegetation data

```
data(dune,dune.env,package="vegan")
data <- cbind(y=as.factor(dune$Bracruta), dune.env)
(model5 <- MASS::polr(y~A1, data = data))
```

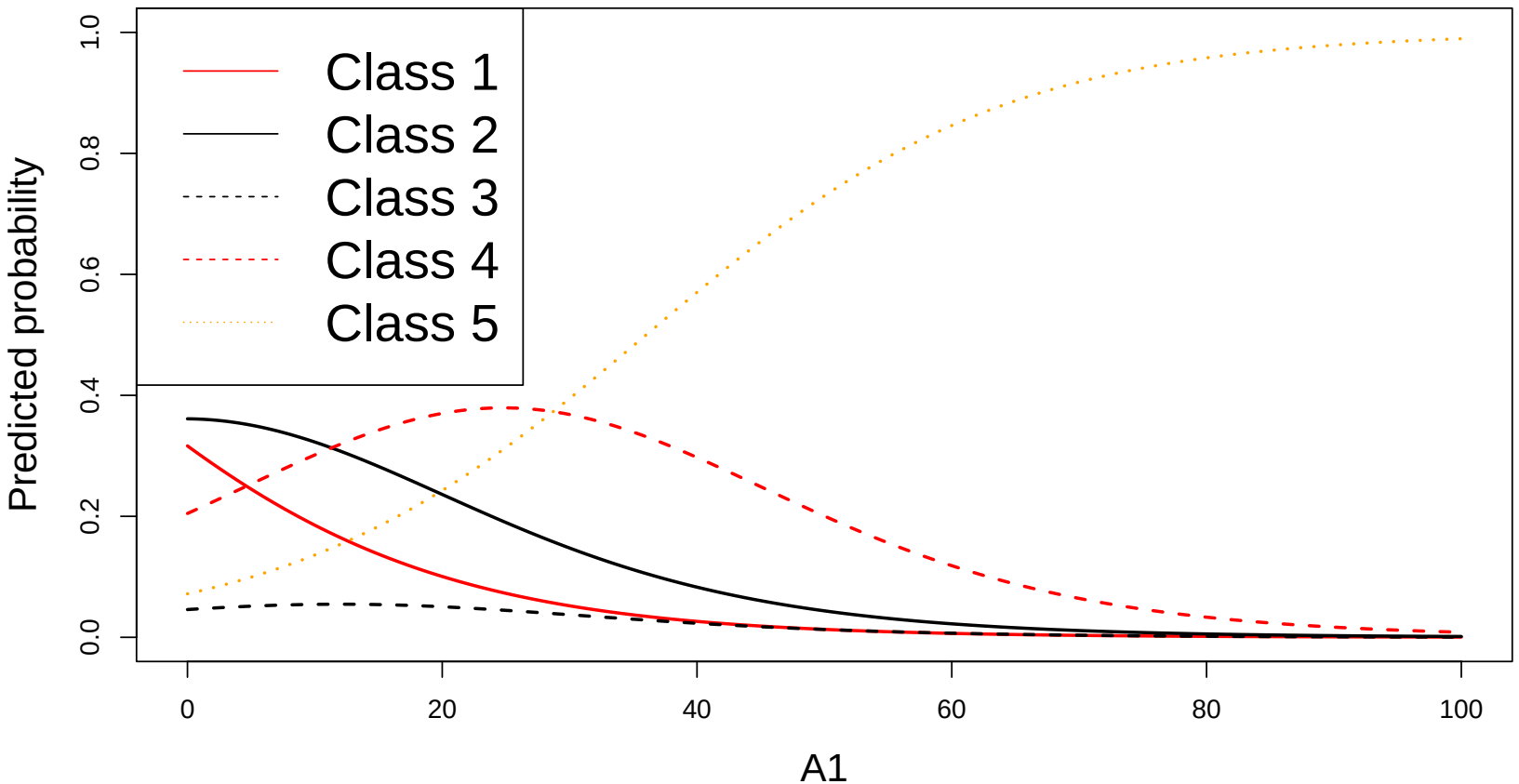
```
## Call:
## MASS::polr(formula = y ~ A1, data = data)
##
## Coefficients:
##         A1
## 0.07104338
##
## Intercepts:
##       0|2       2|3       3|4       4|6
## -0.7704179 0.7426109 0.9615655 2.5585698
##
## Residual Deviance: 57.48771
## AIC: 67.48771
```

Example: vegetation data

```
predict(model5,type="probs")
```

```
##             0         2          3         4          6
## 1   0.2750122 0.3576581 0.04925726 0.2317706 0.08630192
## 2   0.2652086 0.3558300 0.05000602 0.2386505 0.09030488
## 3   0.2542818 0.3532918 0.05080543 0.2465371 0.09508384
## 4   0.2556313 0.3536348 0.05070893 0.2455506 0.09447432
## 5   0.2282902 0.3449388 0.05251291 0.2662255 0.10803261
## 6   0.2542818 0.3532918 0.05080543 0.2465371 0.09508384
## 7   0.2750122 0.3576581 0.04925726 0.2317706 0.08630192
## 8   0.2556313 0.3536348 0.05070893 0.2455506 0.09447432
## 9   0.2624490 0.3552399 0.05021166 0.2406205 0.09147894
## 10  0.2679868 0.3563901 0.04979663 0.2366821 0.08914442
## 11  0.2652086 0.3558300 0.05000602 0.2386505 0.09030488
## 12  0.2346085 0.3472870 0.05212735 0.2613202 0.10465704
```

Ordinal example: vegetation data



Note that ordering has introduced some shapes.

Proportions

We cannot analyse proportion data with linear regression; might get something below zero or above one

- Proportions usually include 0 and/or 1, but also everything between
- Binomial responses are dichotomous (either 0 or 1), not proportions
- If we know the total, we can specify the Binomial directly in terms of proportions

Recall: binomial GLM

$$\mathcal{L}(y_i;\Theta) = \exp\left\{ \log \binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N} \right\} \quad (6)$$

Recall: binomial GLM

$$\mathcal{L}(y_i;\Theta) = \exp\left\{ \log \binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N} \right\} \quad (6)$$

in terms of proportion of successes:

$$\mathcal{L}(Ny_i;\Theta) = \exp\left\{ \log \binom{N}{Ny_i} + \frac{y_i\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N} \right\} \quad (7)$$

Recall: binomial GLM

$$\mathcal{L}(y_i;\Theta) = \exp\left\{\log\binom{N}{r} + \frac{\frac{r}{N}\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N}\right\} \quad (6)$$

in terms of proportion of successes:

$$\mathcal{L}(Ny_i;\Theta) = \exp\left\{\log\binom{N}{Ny_i} + \frac{y_i\log(\frac{p_i}{1-p_i}) + \log(1-p_i)}{1/N}\right\} \quad (7)$$

So, we still model the successes and failures

Unknown total

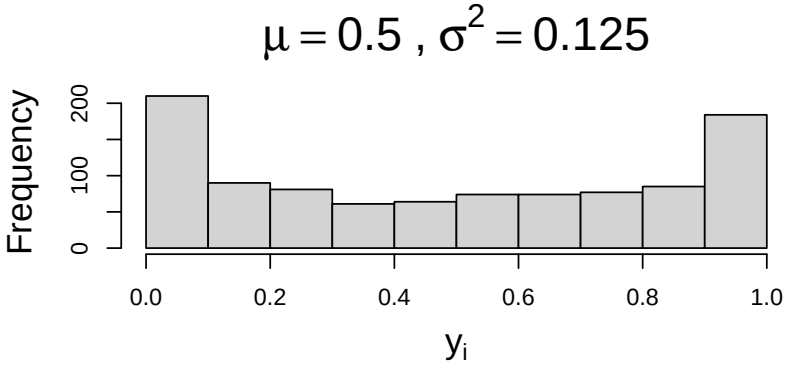
If the number of trials is unknown, we cannot specify a binomial glm

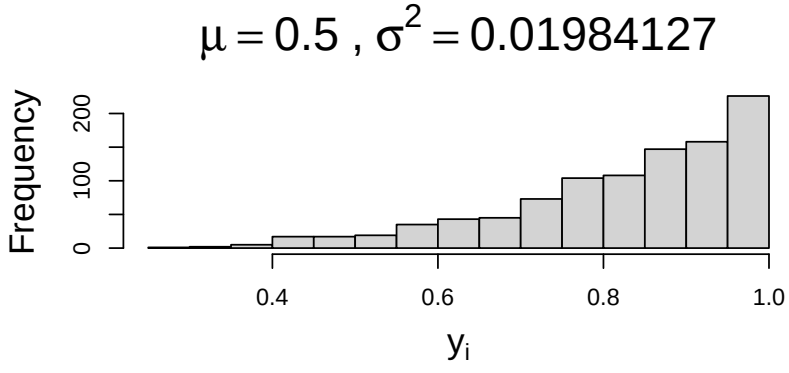
- We can model the proportions directly
- "beta regression", which is not a GLM
- Various packages implement it, with logit or probit link-functions
- Requires data between 0 and 1 (i.e., without 0 or 1)

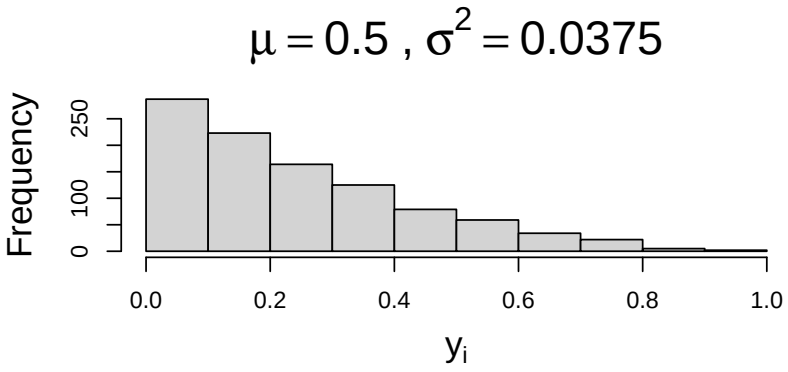
When there are zeros and/or ones..zero-inflated models or hurdle models (ZIB or ordered beta regression).

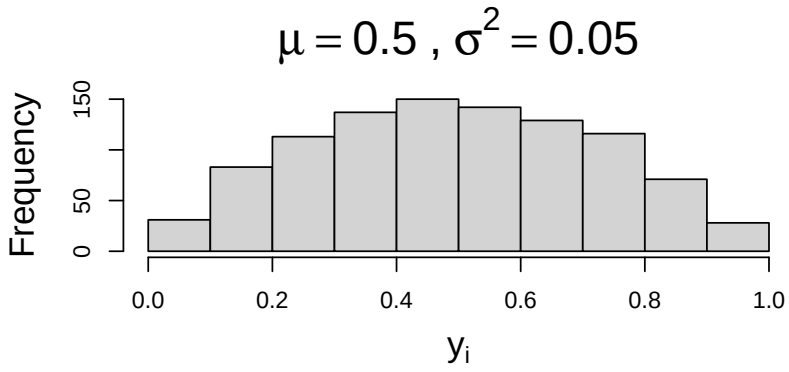
Beta distribution

Two parameter distribution with mean and variance









Example: Baseball game wins

Wins of baseball games

Baseball wins: the data

Data from Cochran (2002)

- 838 games
- Many variables: team, league, division, year, runs scores, wins
- Response variable: attendance

franchise	league	division	year	attendance	runs.scored	runs.allowed	wins	losses	games.behind
BAL	AL	EAST	69	1062069	779	517	109	53	0.0
BOS	AL	EAST	69	1833246	743	736	87	75	22.0
CLE	AL	EAST	69	619970	573	717	62	99	46.5
DET	AL	EAST	69	1577481	701	601	90	72	19.0
NYA	AL	EAST	69	1067996	562	587	80	81	28.5
WAS	AL	EAST	69	918106	694	644	86	76	23.0
CAL	AL	WEST	69	758388	528	652	71	91	26.0

Note the new column $prop_wins = \frac{wins}{wins+losses}$

Baseball wins: binomial GLM

Without weights

```
model6 <- glm(prop_wins ~ attendance, family="binomial", data = MLBattend)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

With weights

```
model6 <- glm(prop_wins ~ scale(attendance), family="binomial", weights = wins + losses, data = MLBattend)
```

Baseball wins: beta regression

```
library(glmmTMB)
model7 <- glmmTMB(prop_wins ~ scale(attendance), family=beta_family(), data = MLBattend)
```

```
##                      binomial          beta
## (Intercept)       -0.001815271 -0.0002906077
## scale(attendance)  0.121796467  0.1209277075
```

Excess zeros

When there are more zeros than the selected distribution can generate.

- Can be accounted for with additional model terms (covariate)
- Alternatively:
  - Zero-inflated models
  - Hurdle models
  - Negative-binomial models: can handle excess zeros very well

1) Zero-inflated models: another process generates additional zeros (e.g., colonisation vs. abundance)

- e.g., Species abundance

2) Hurdle models: only one of the two processes generates values below the hurdle

Zero-inflation (Poisson)

$$y_i \sim ZIP(\lambda_{ij}, \pi_i) \tag{8}$$

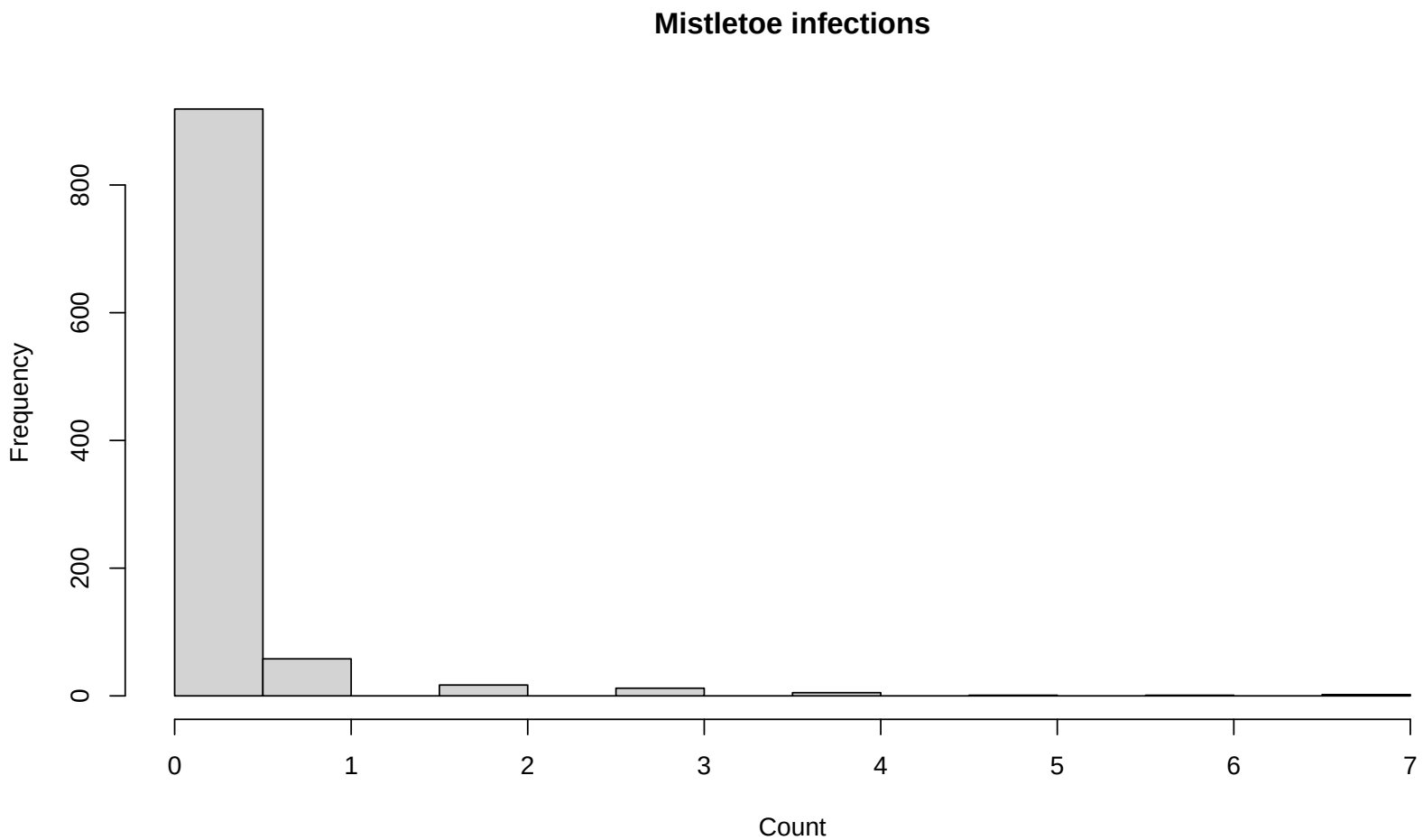
$$y_i|\nu_j \sim \begin{cases} pPois(\lambda_i) & \text{if } y_i > 0 \\ \begin{cases} p\exp(-\lambda_i) & \text{if } \nu_i = 1 \\ 1-p & \text{if } \nu_i = 0 \end{cases} & \text{if } y_i = 0 \end{cases} \tag{9}$$

- can choose to separately model p as a logistic regression
- quadratic mean-variance as in NB

Example: Mistletoe infections

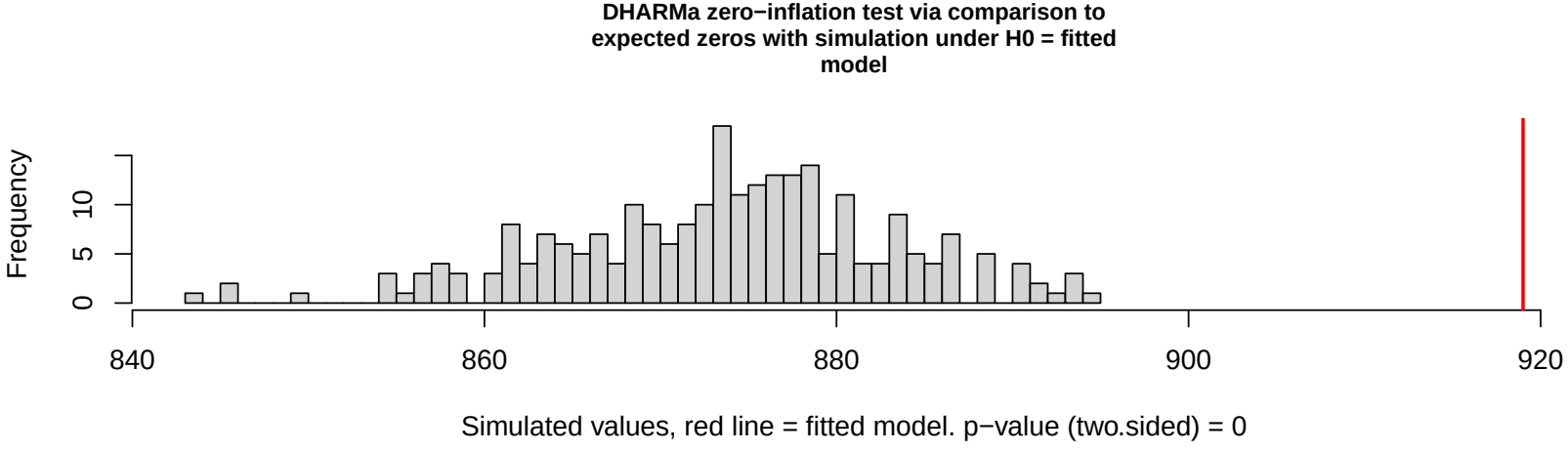
Counts of mistletoe infections

Mistletoe: visually



Mistletoe: poisson regression

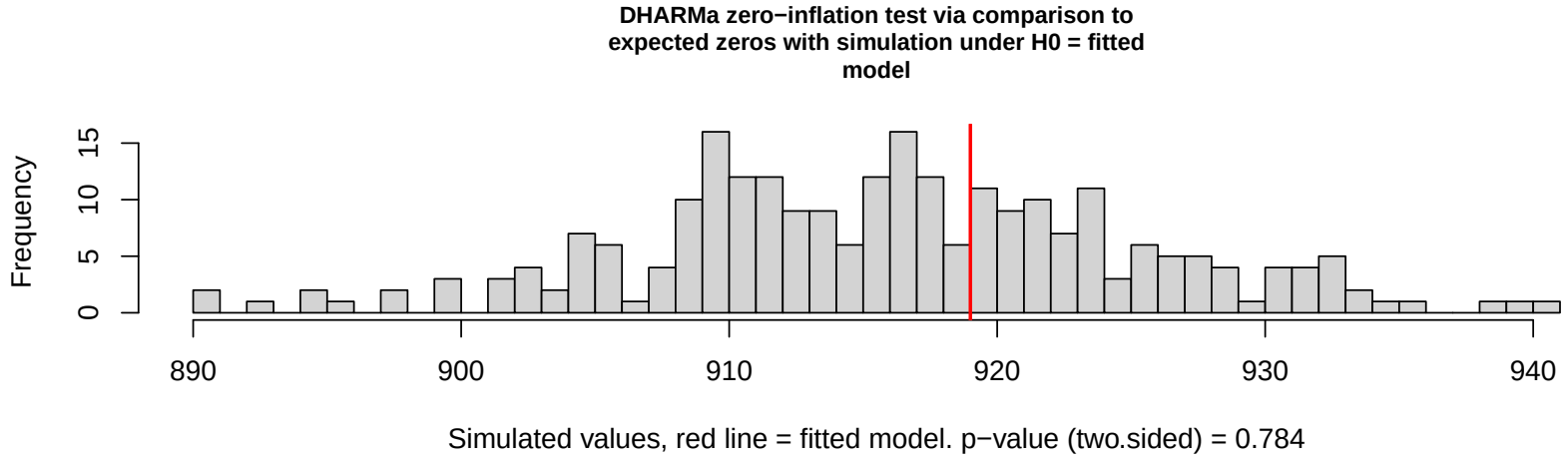
```
model8 <- glm(No.of.mistletoes~DBH, family="poisson", data = data)
```




```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 1.0513, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Mistletoe: zero-inflated poisson regression

```
model9 <- glmmTMB::glmmTMB(No.of.mistletoes~DBH,ziformula=~1, family="poisson", data = data)
```




```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 1.0027, p-value = 0.784
## alternative hypothesis: two.sided
```

Summary

- At least not mentioned: Generalized and Conway-Maxwell Poisson
- Ordered beta
- Truncation
- Many, many others