# Statistics with SpaRRows II

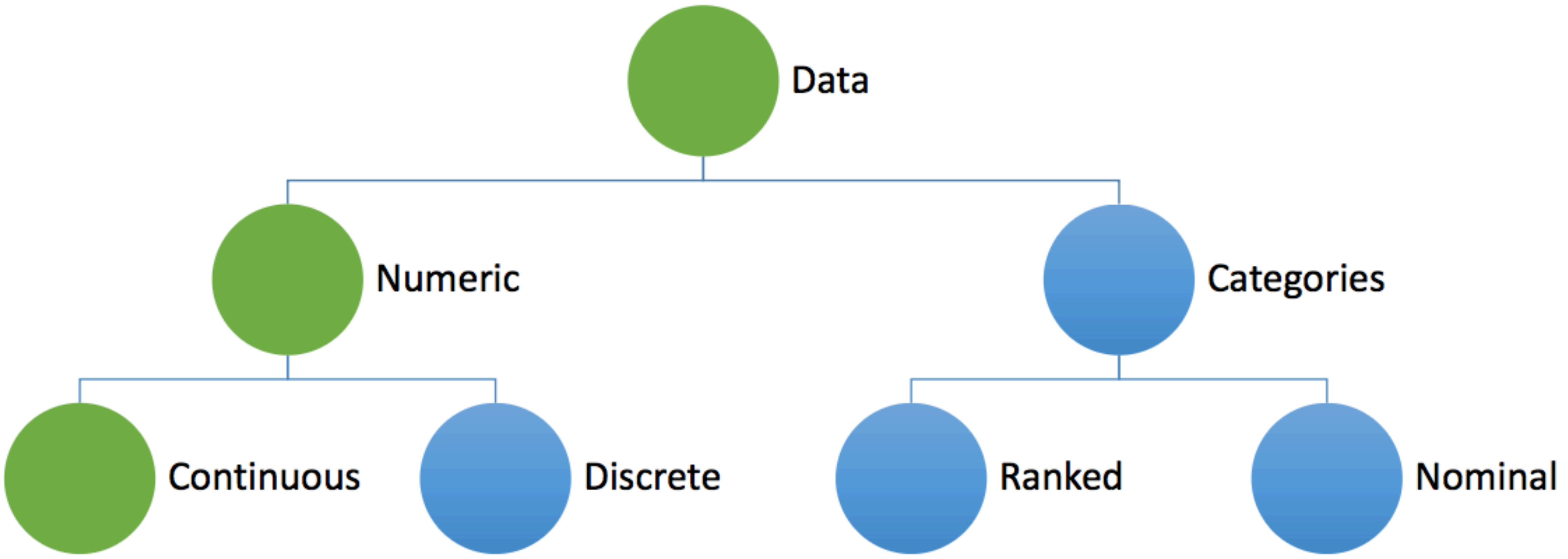
## Many models, matrices, and magic

Julia Schroeder

Julia.schroeder@imperial.ac.uk

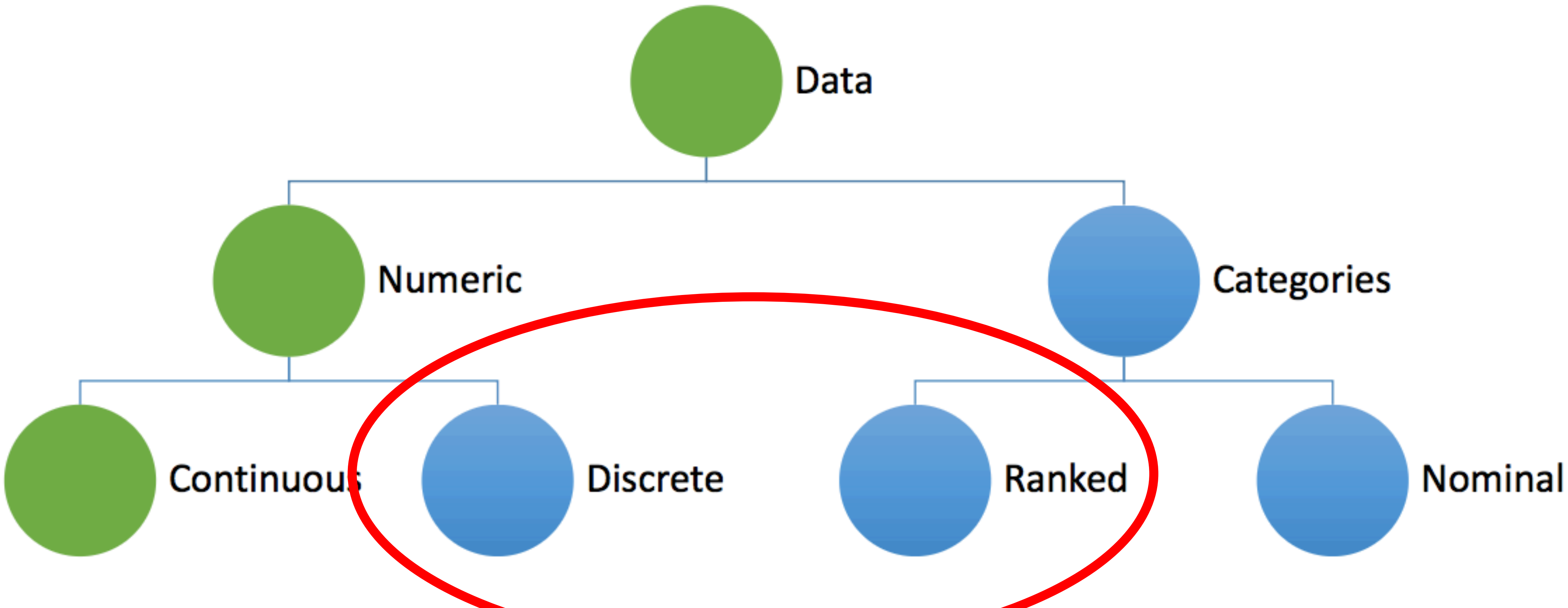
# Remember this?

## Data types



# Remember this?

## Data types



# Generalized Linear Models

# Generalized Linear Models

- Don’t panic

# Generalized Linear Models

- Don’t panic
- Extension of Linear Models

# Generalized Linear Models

- Don't panic
- Extension of Linear Models
- General philosophy is the same

# Generalized Linear Models

- Don’t panic
- Extension of Linear Models
- General philosophy is the same

# Problematic response variables

- Count data

# Count data

- Number of trees in a plot

# Count data

- Number of trees in a plot
- Number of offspring

# Count data

- Number of trees in a plot
- Number of offspring
- Number of species observed

# Count data

- Number of trees in a plot
- Number of offspring
- Number of species observed
- …

# Count data

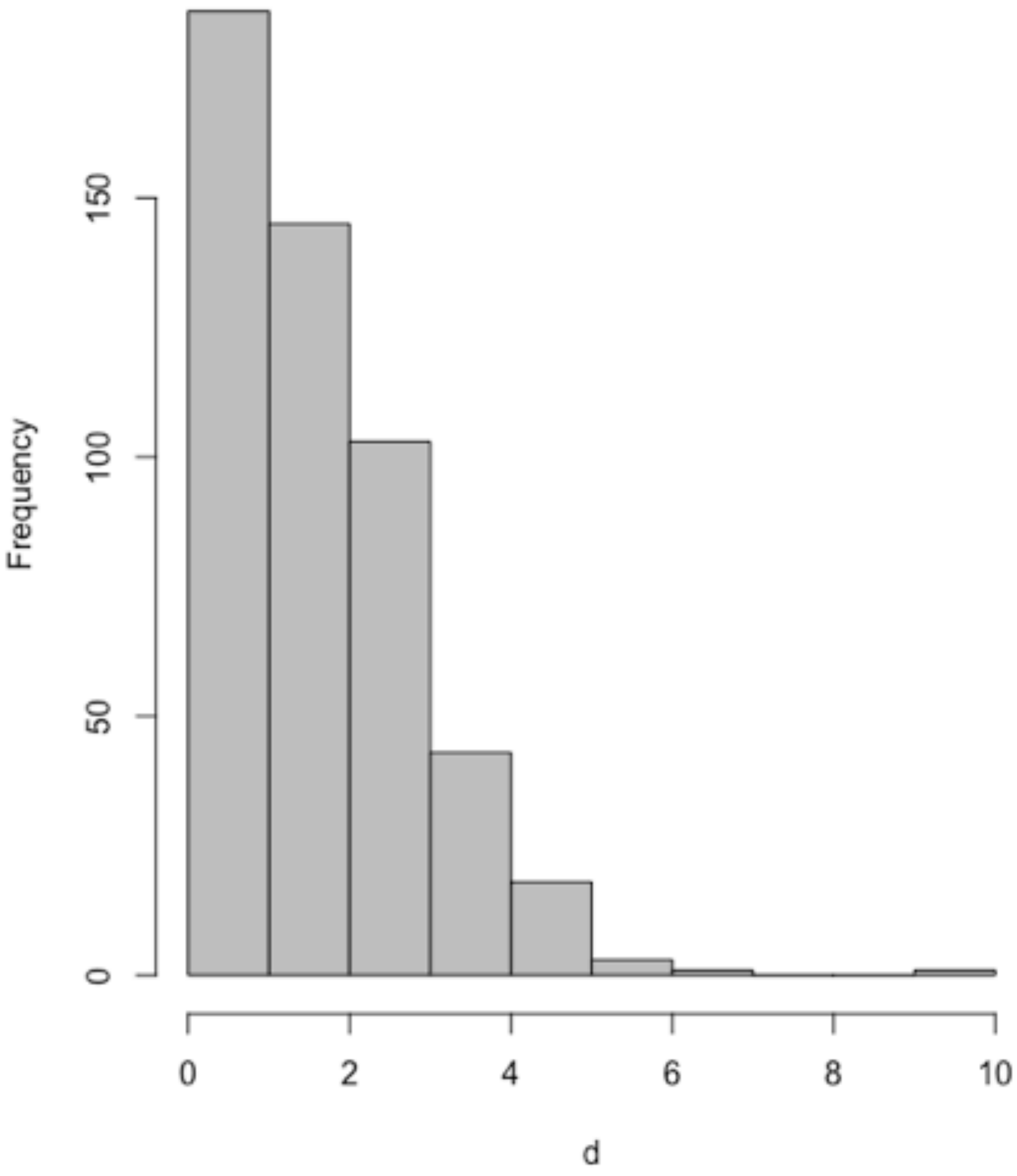
- Number of trees in a plot
- Number of offspring
- Number of species observed
- …

- Cannot be less than zero
- $\geq 0$

# Count data

- Number of trees in a plot
- Number of offspring
- Number of species observed
- …

- Cannot be less than zero
- $\geq 0$
- Right-tailed




- Cannot be less than zero
- $\geq 0$
- Right-tailed

# Problematic response variables

- Count data
- Binary data

# Problematic response variables

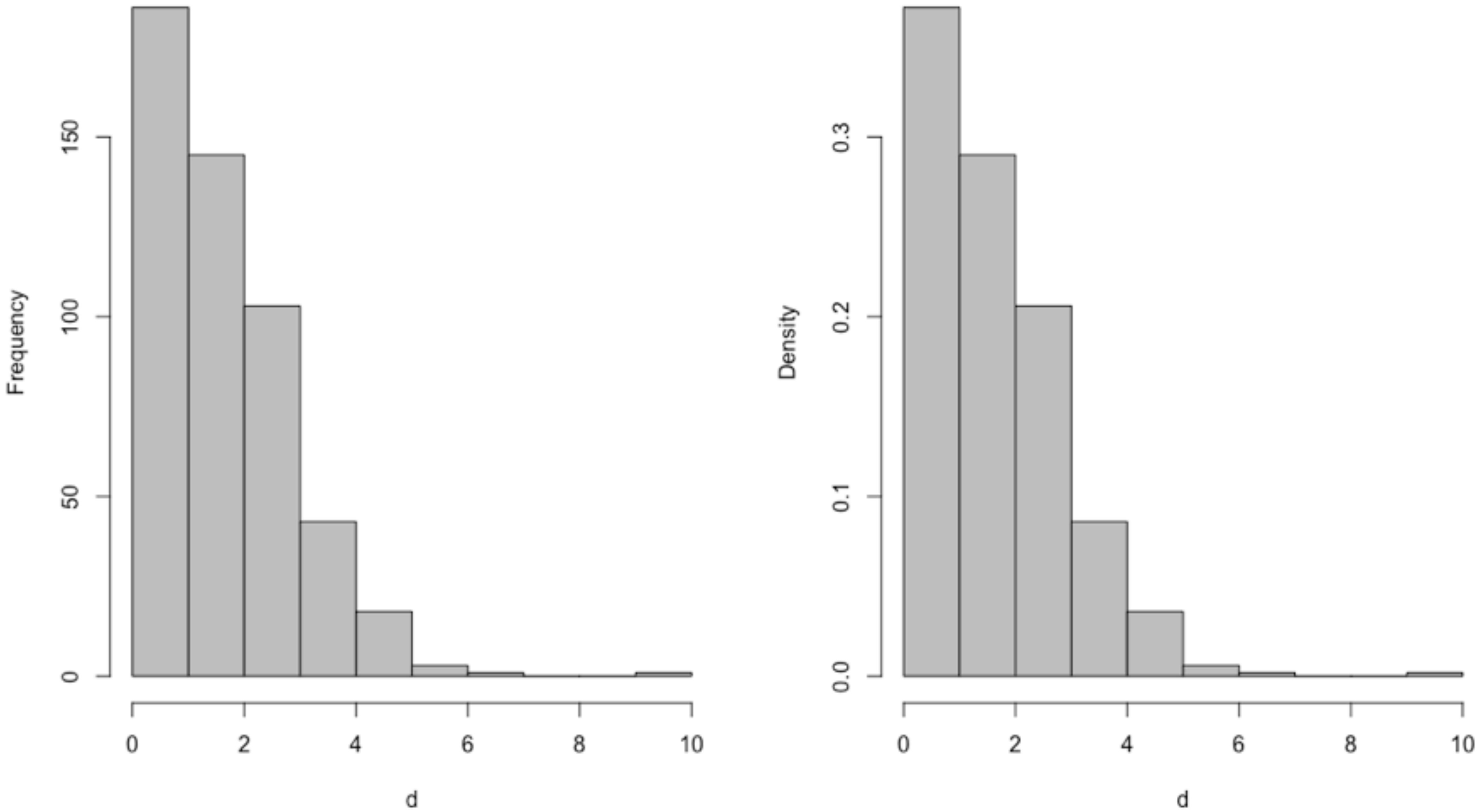
- Count data
- Binary data
- Percentage data

# Problematic response variables

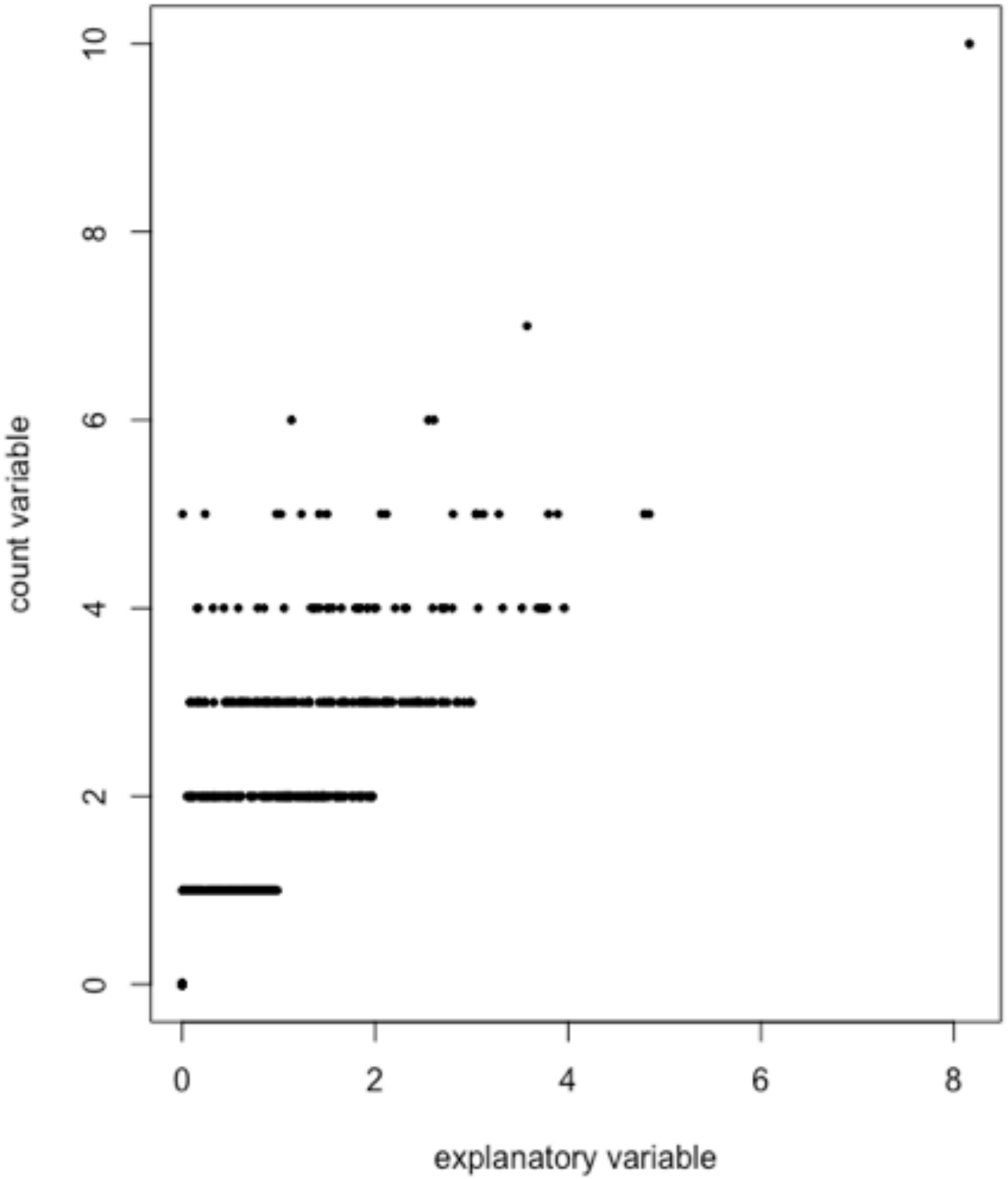
- Count data
- Binary data
- Percentage data
- Ratio data

# Count data

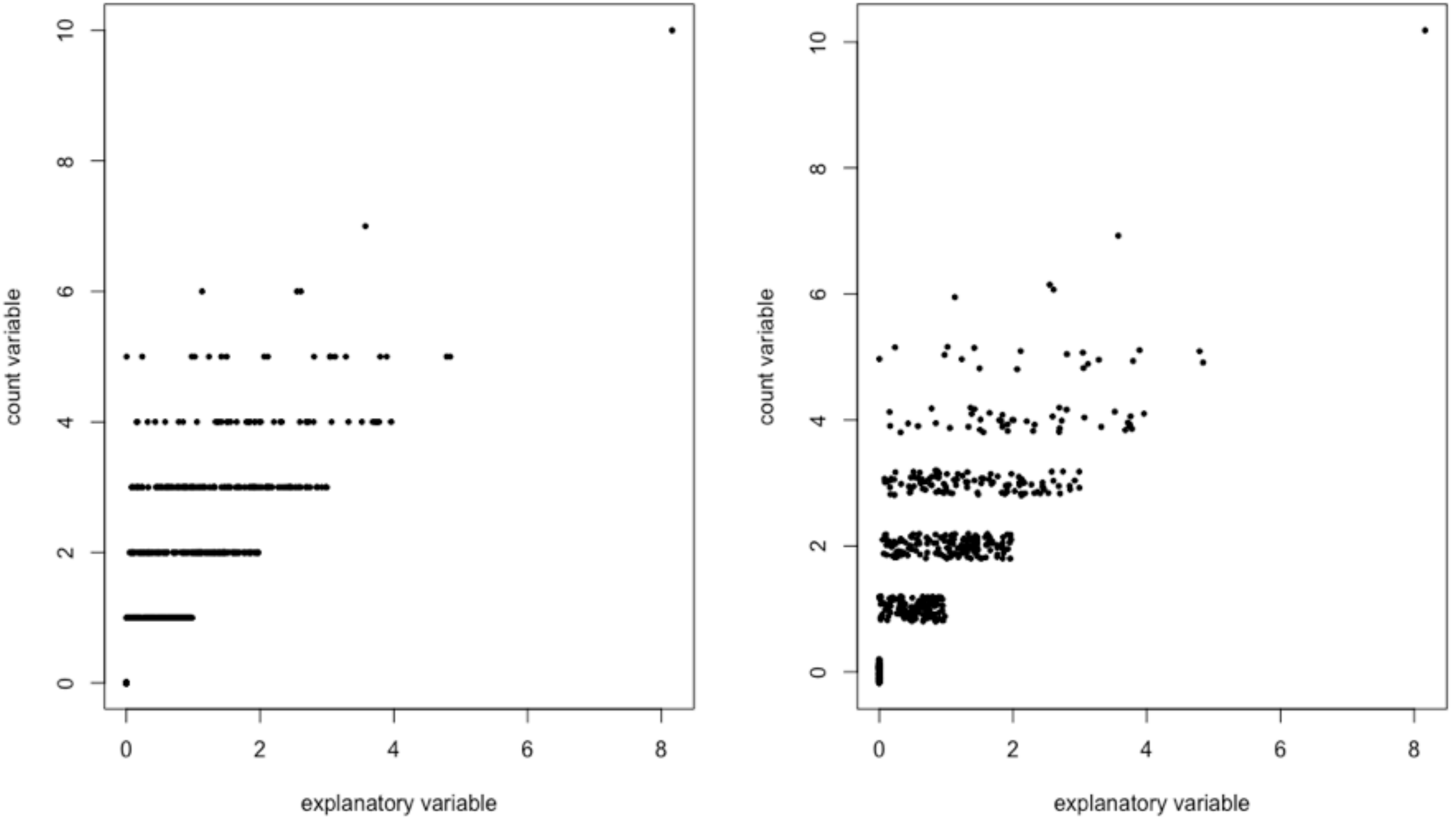
# Count data



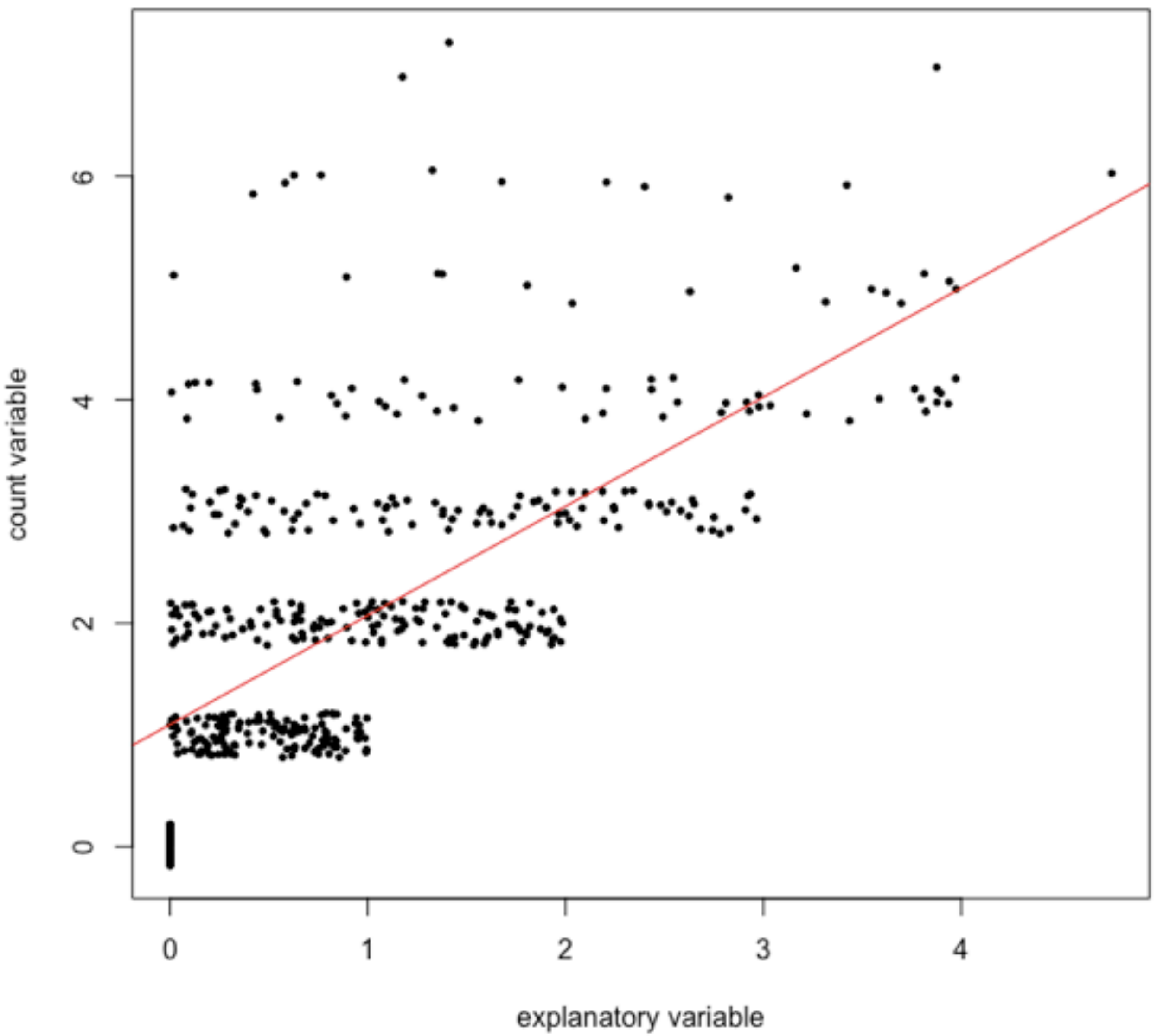
# Count data as response variable



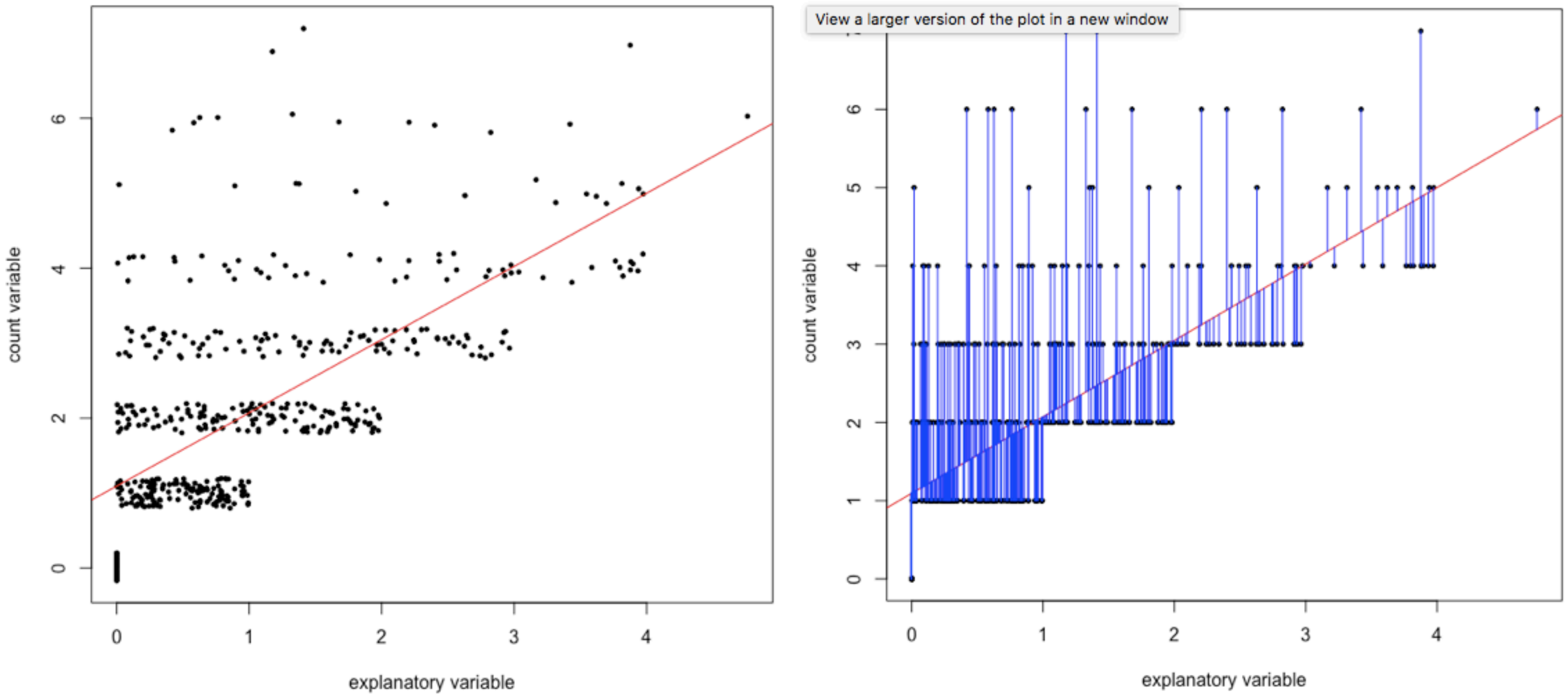
# Count data as response variable



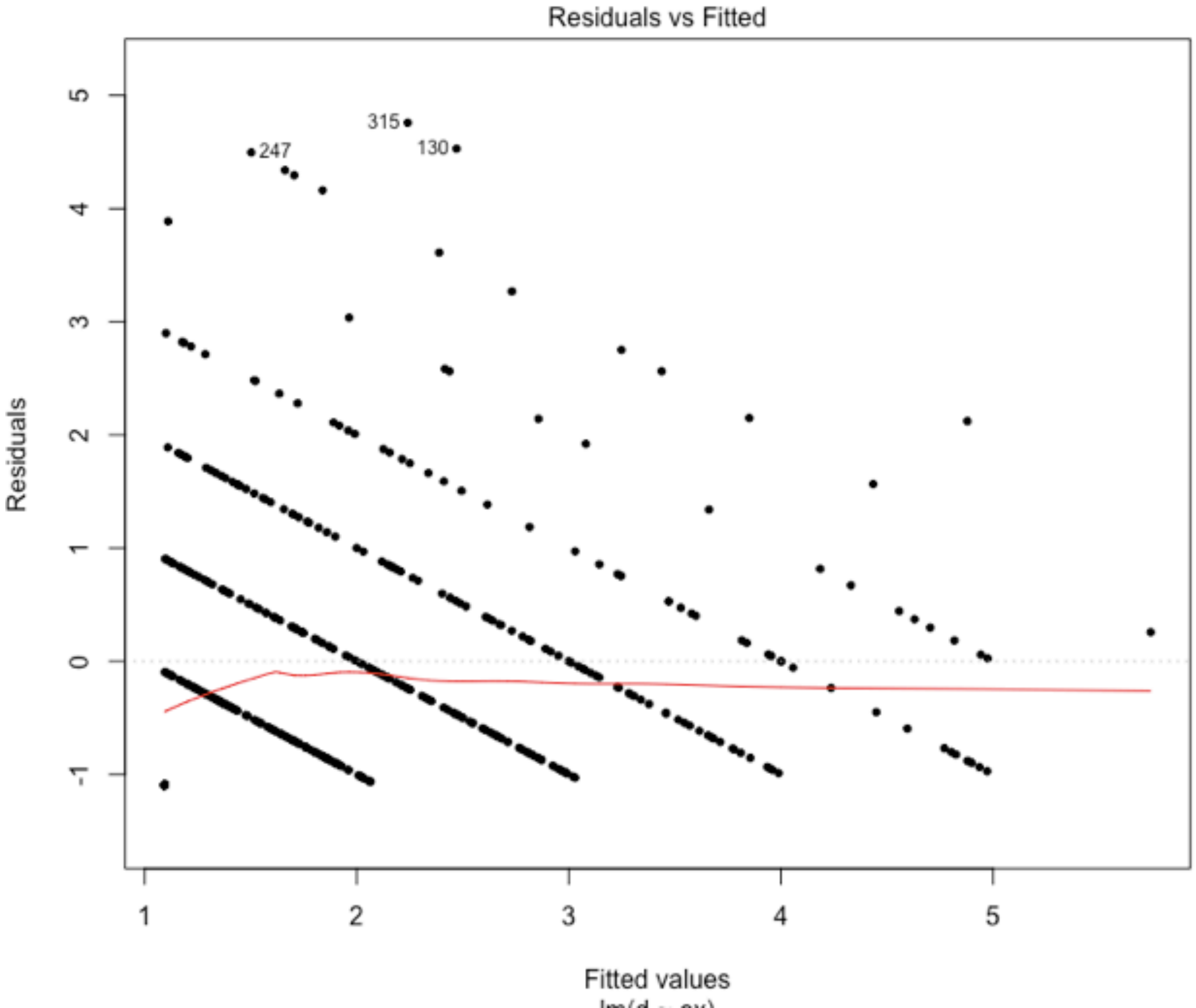
# Count data as response variable



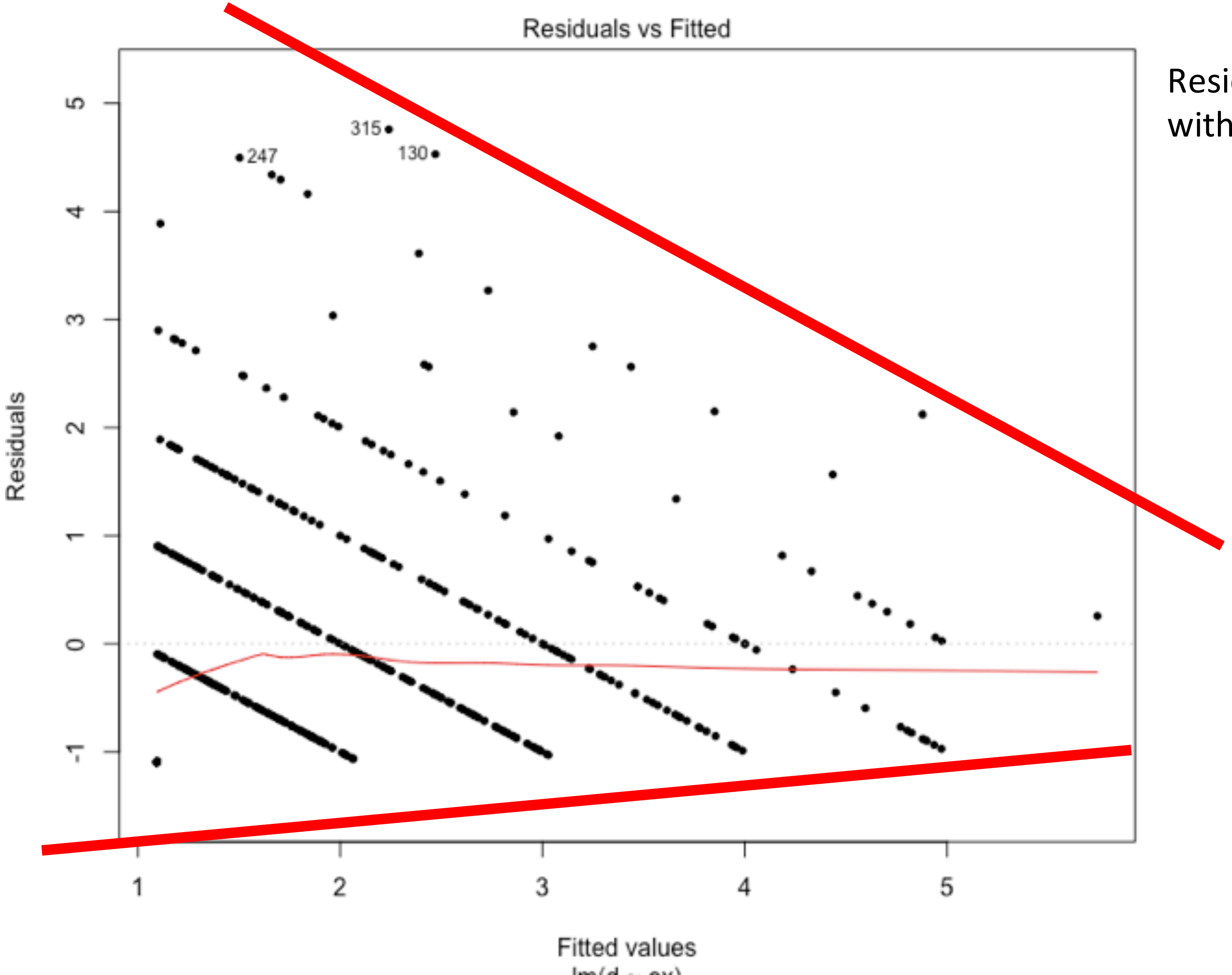
# Count data as response variable



# Count data as response variable - validation

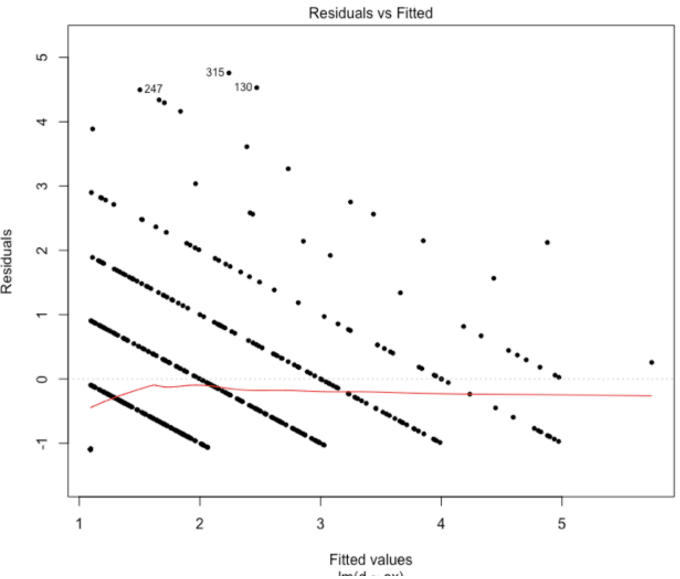


# Count data as response variable - validation

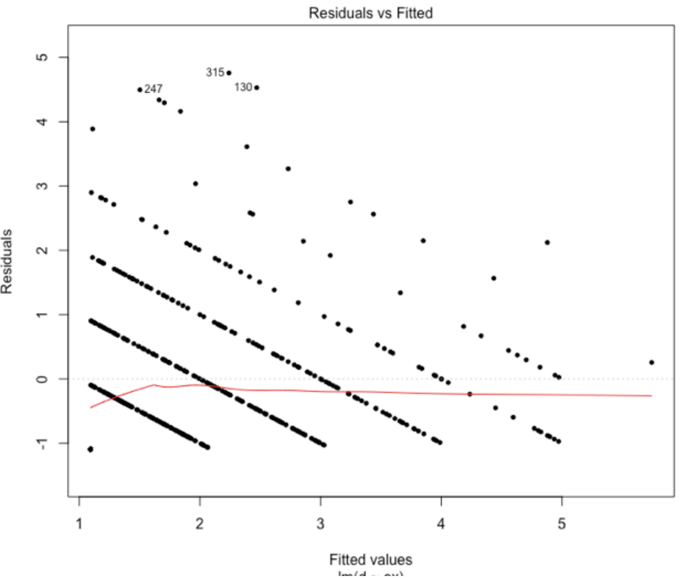


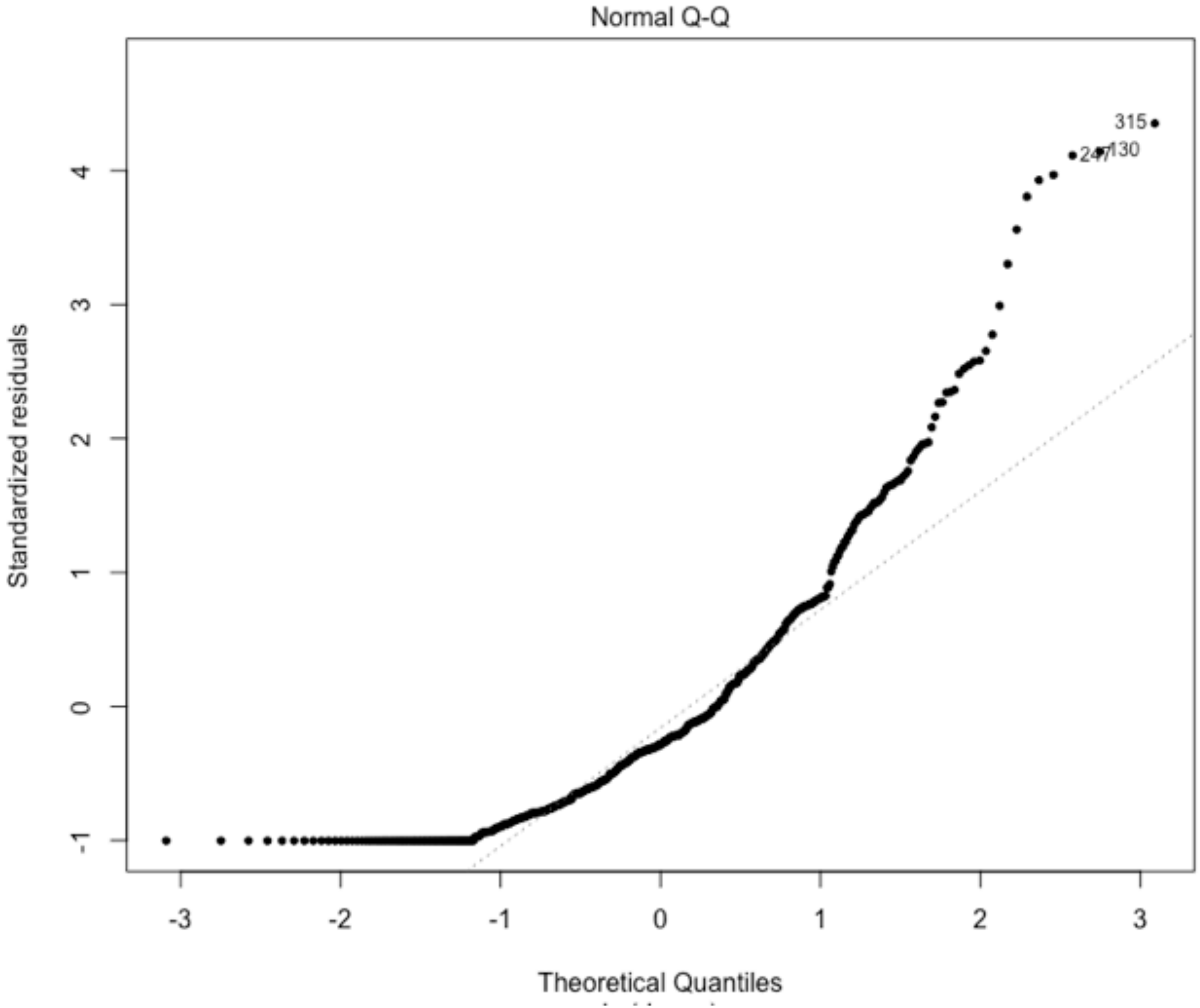

Residual variance decreases with Y

# Count data as response variable - validation

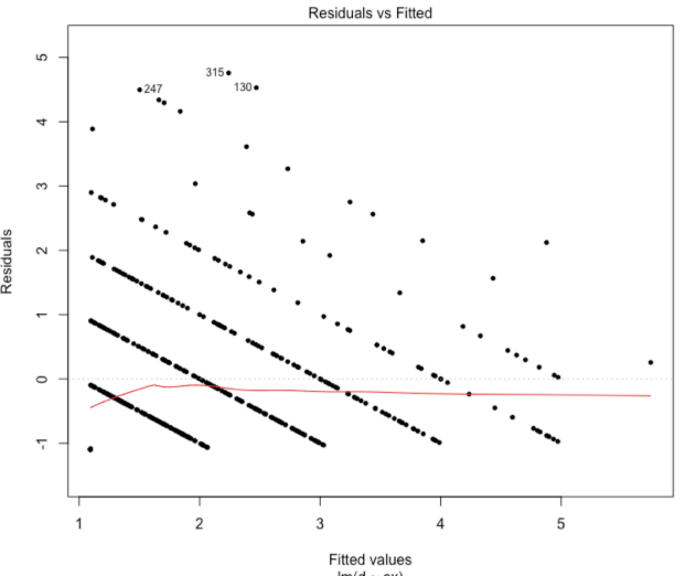


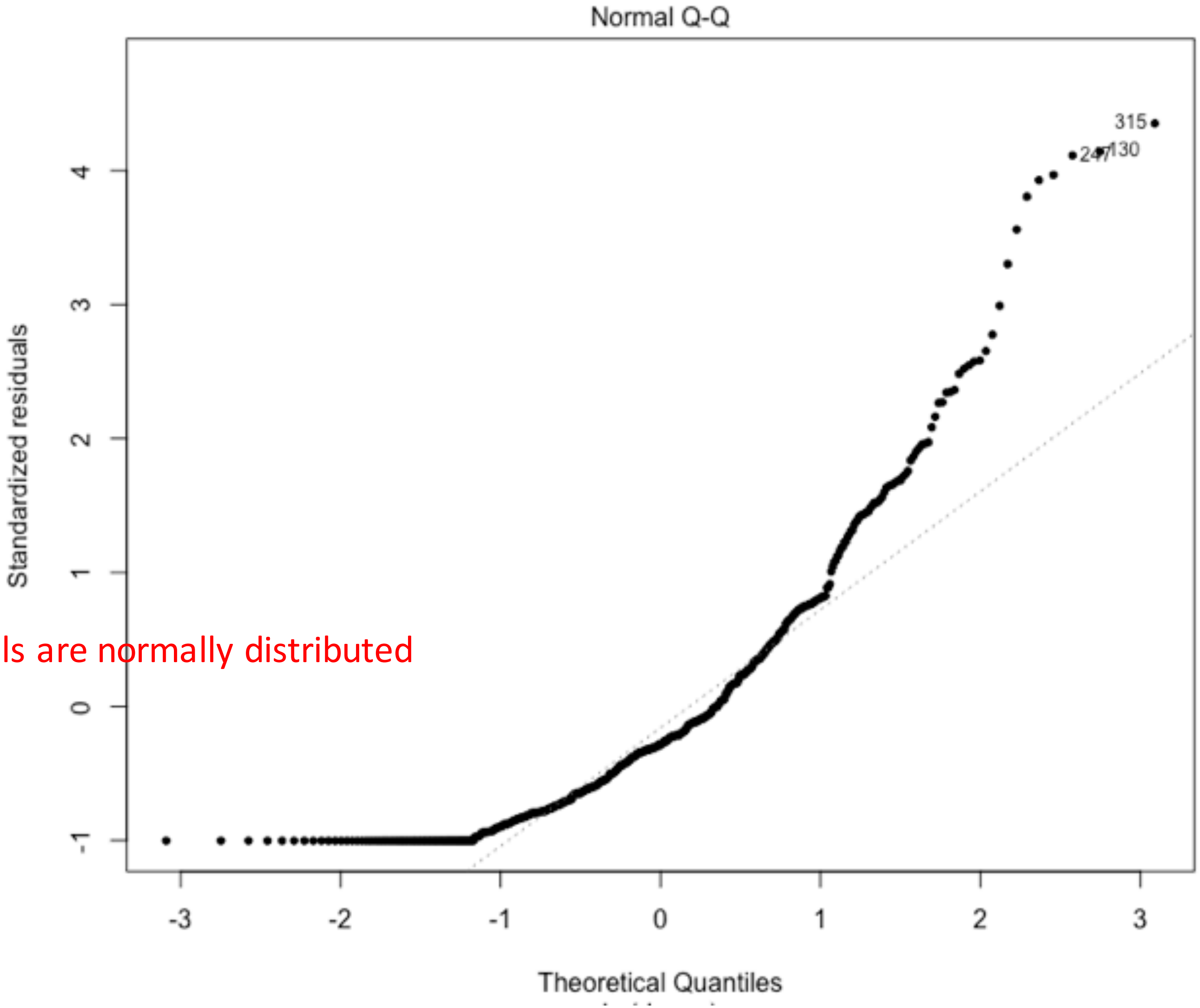
# Count data as response variable - validation



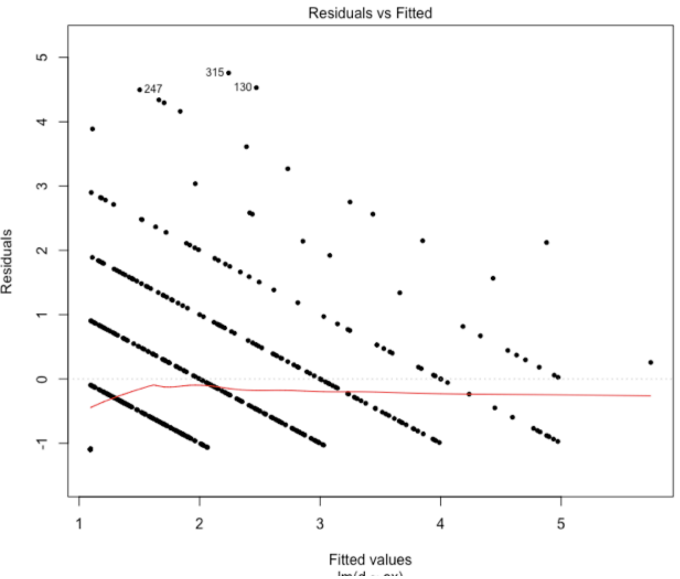
# Count data as response variable - validation

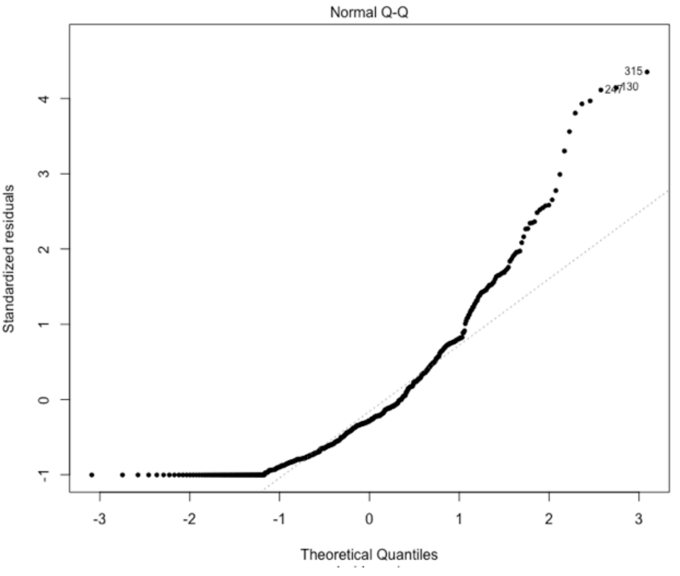


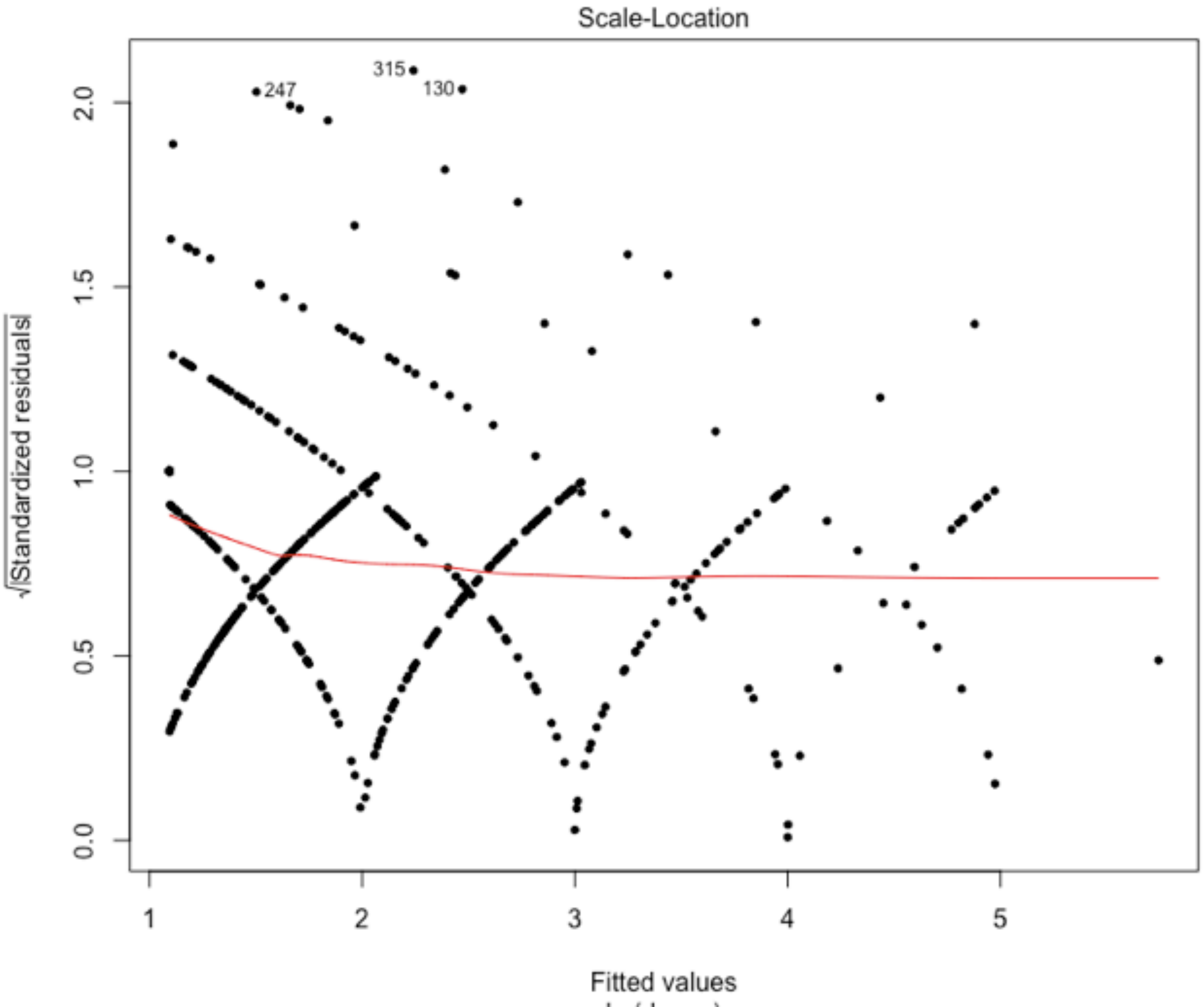




Violation of the assumption that residuals are normally distributed

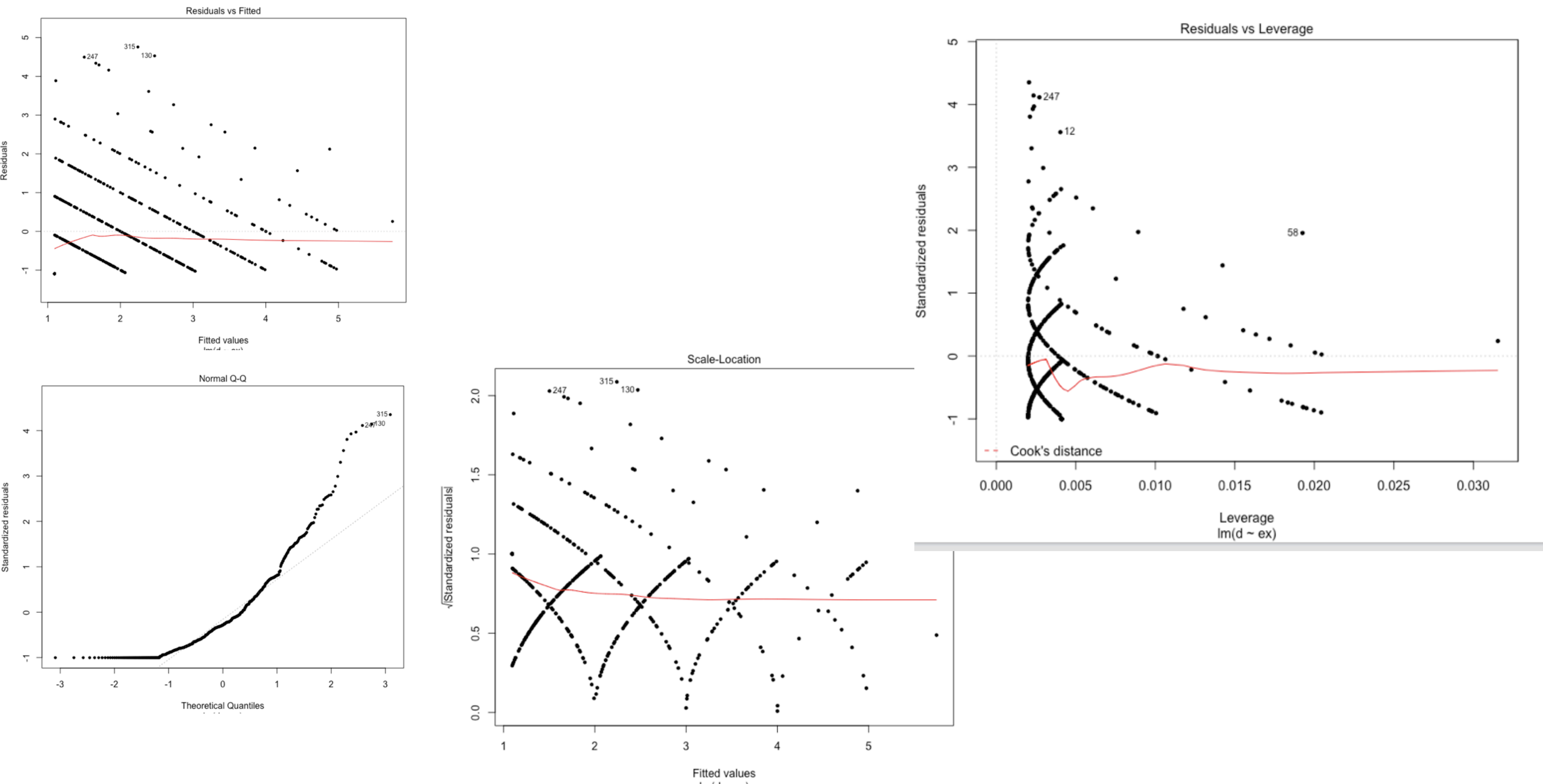
# Count data as response variable - validation



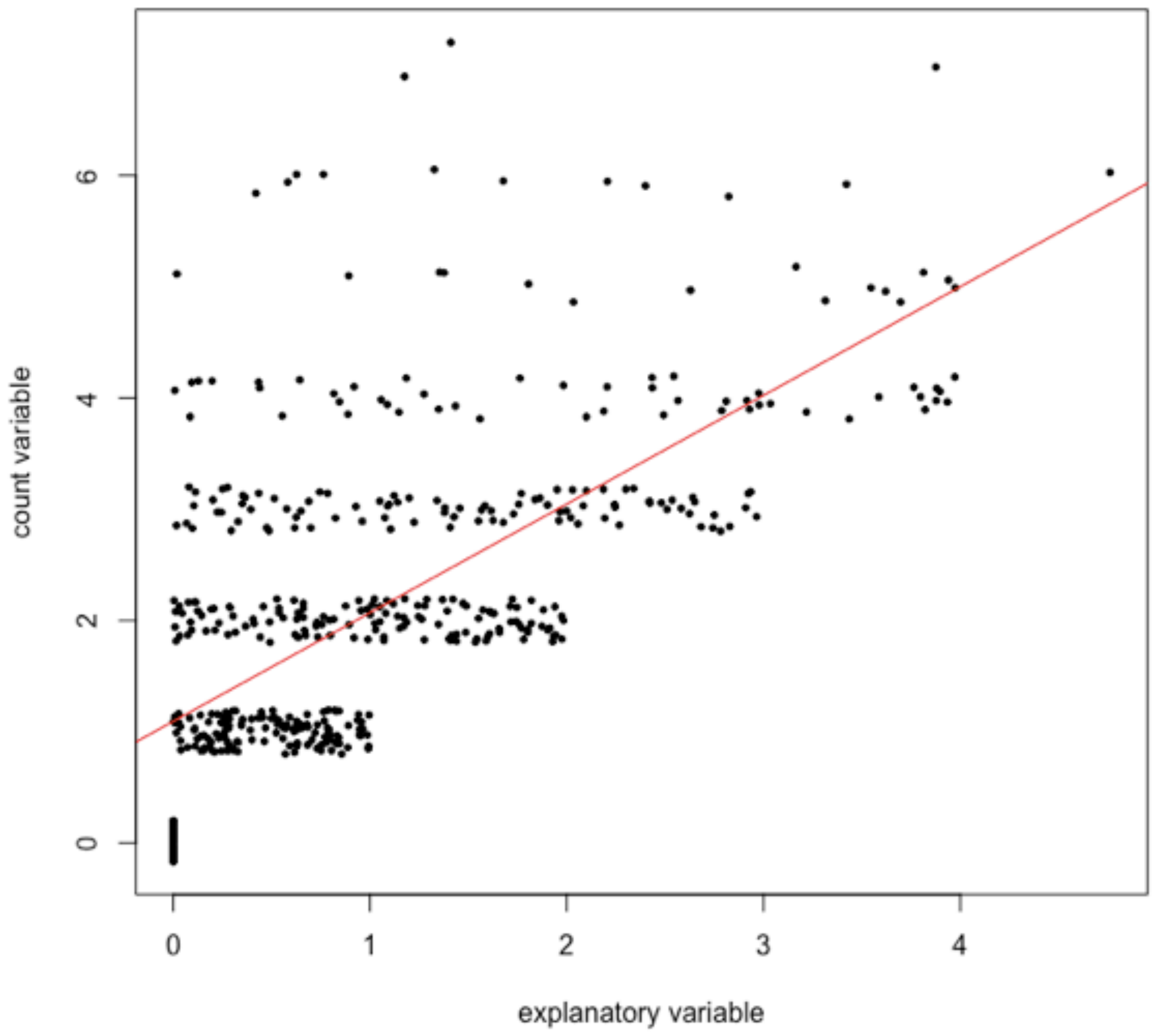



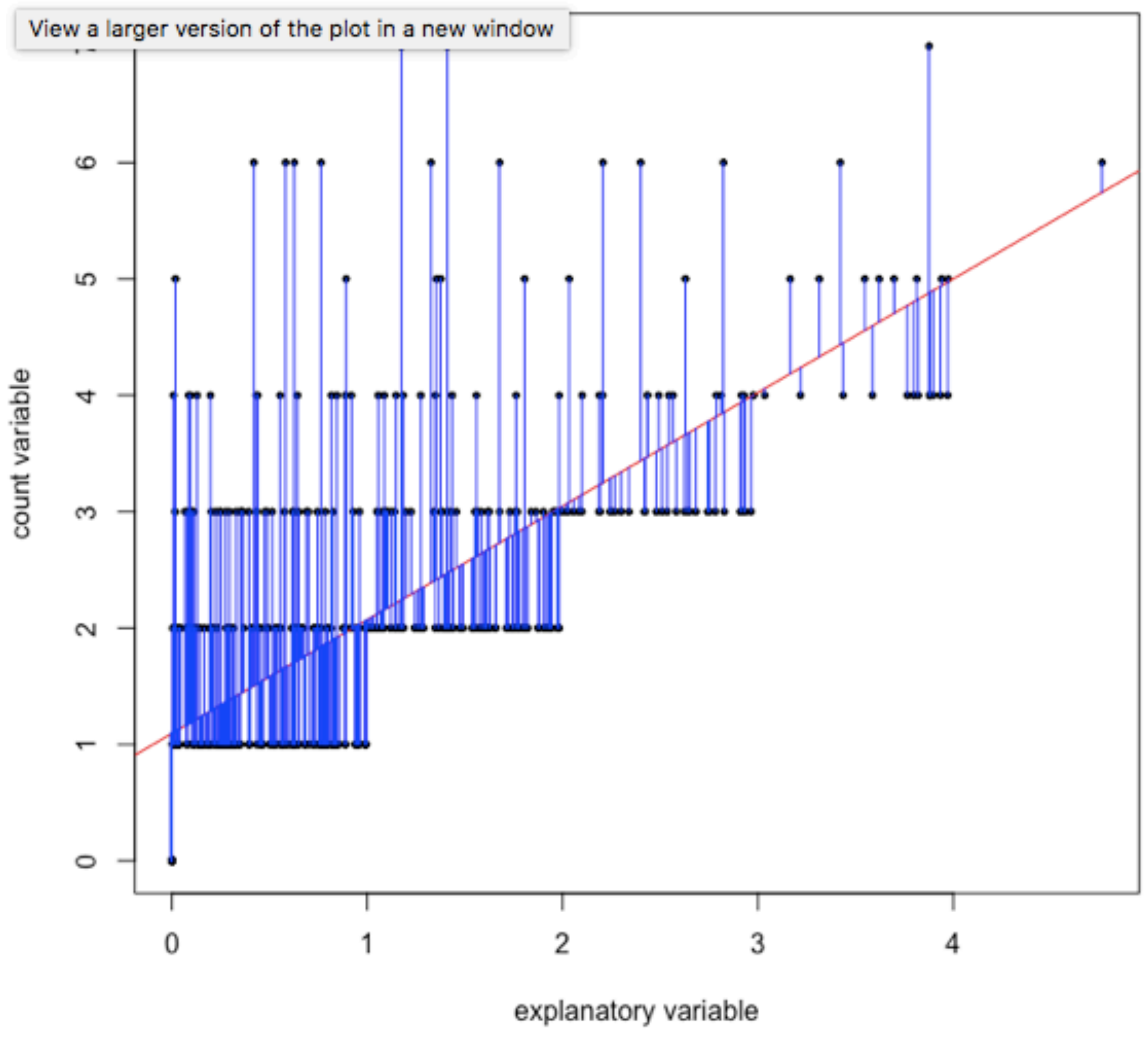



# Count data as response variable – validation

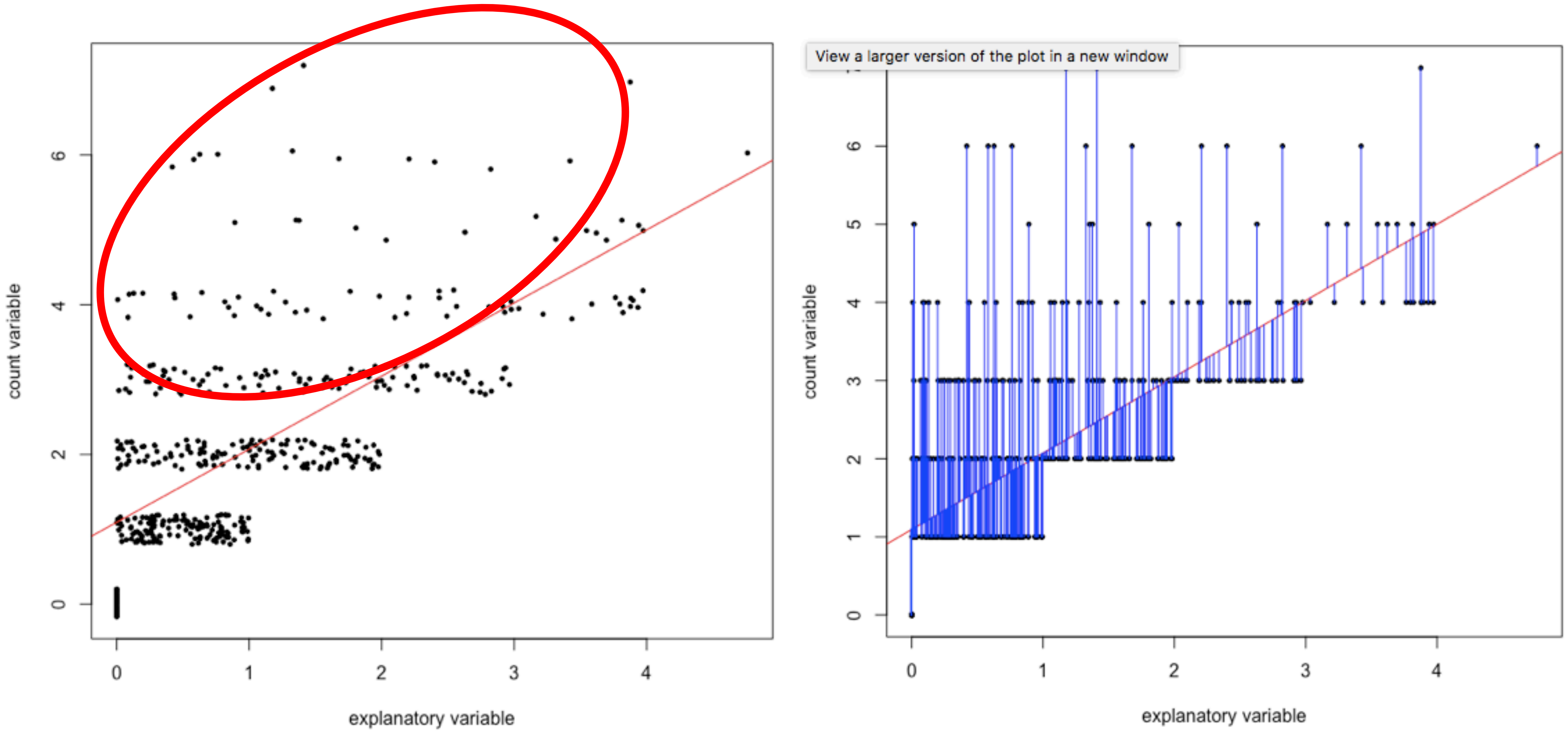


# Count data as response variable

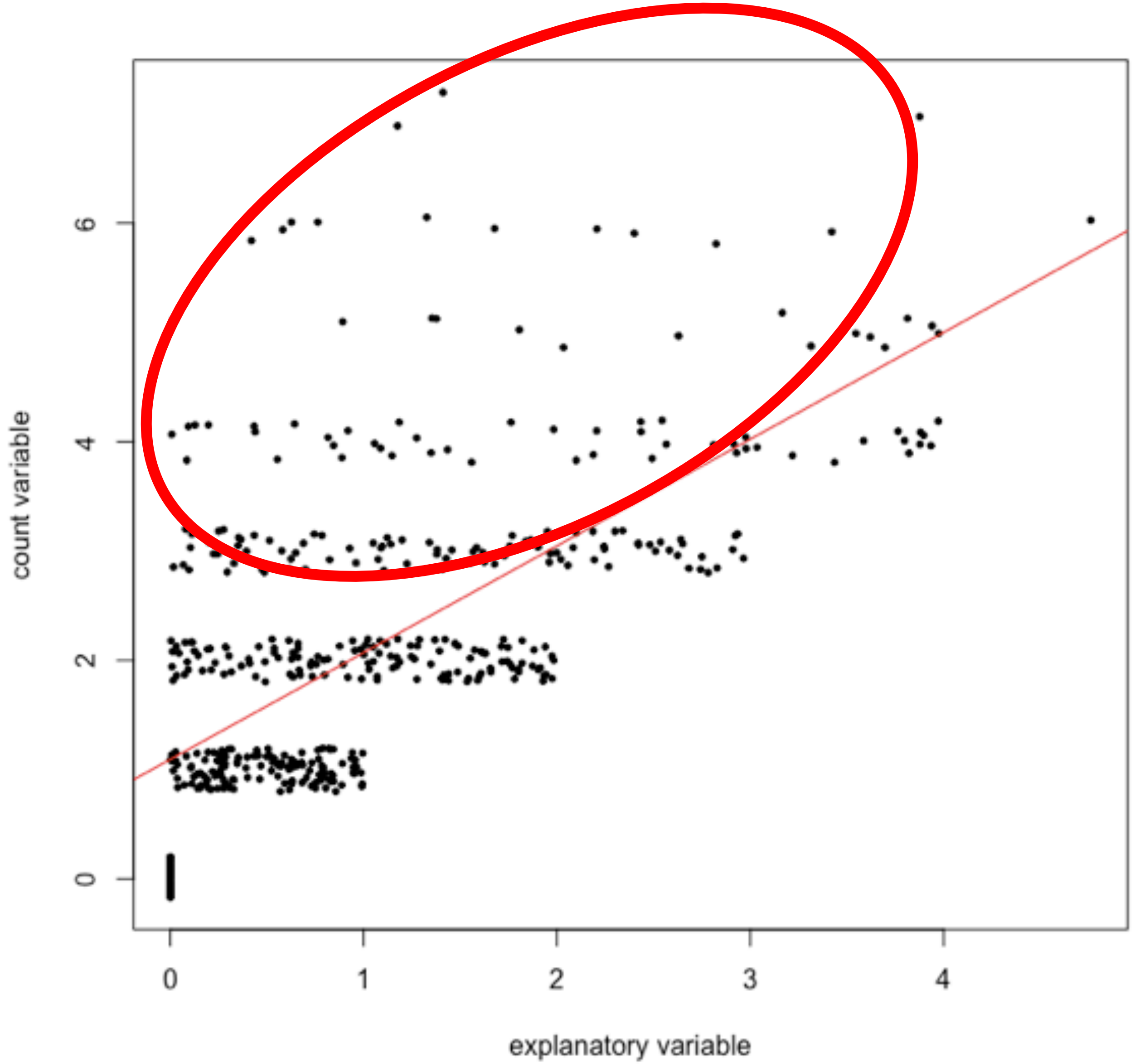


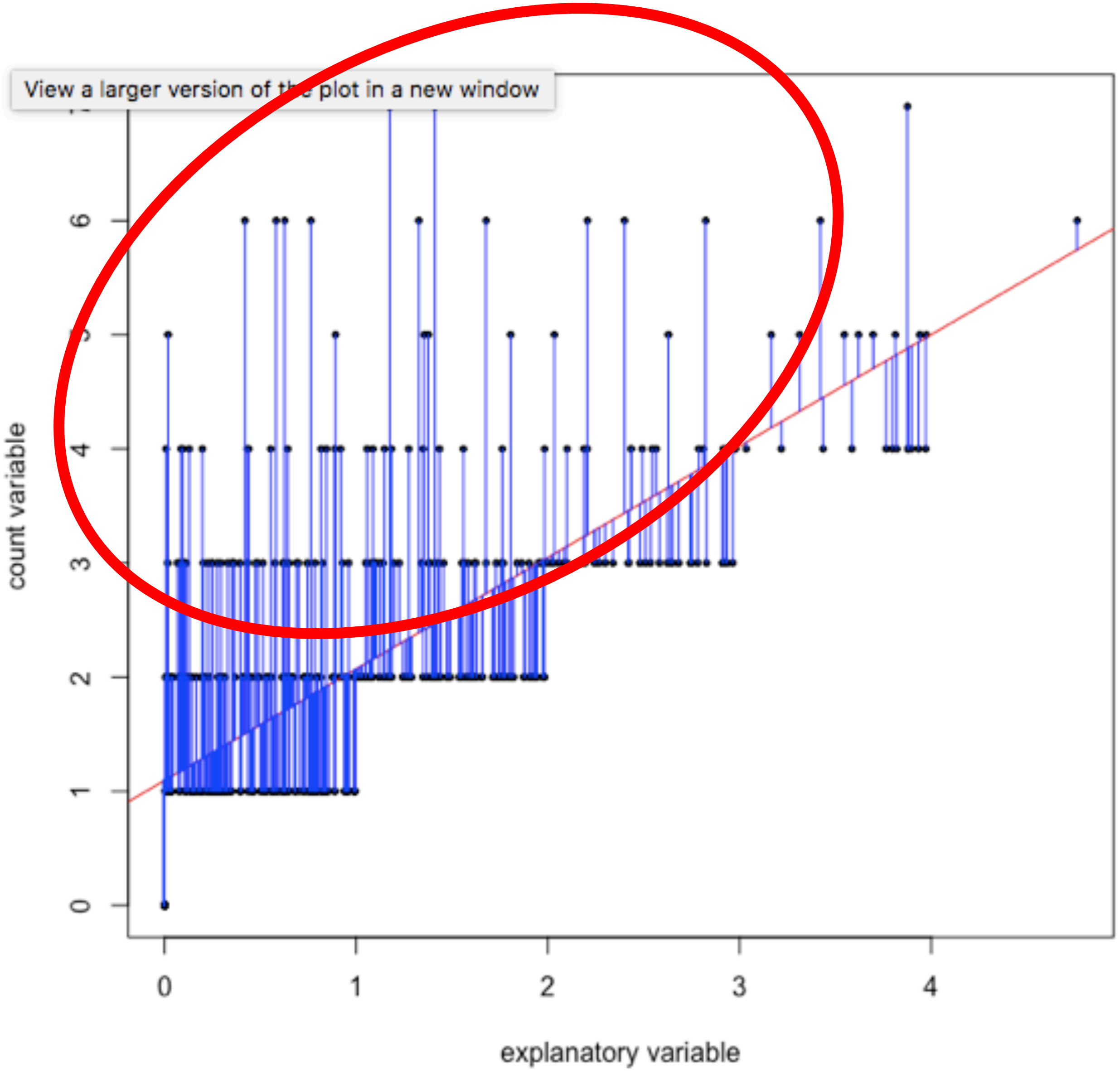



# Count data as response variable

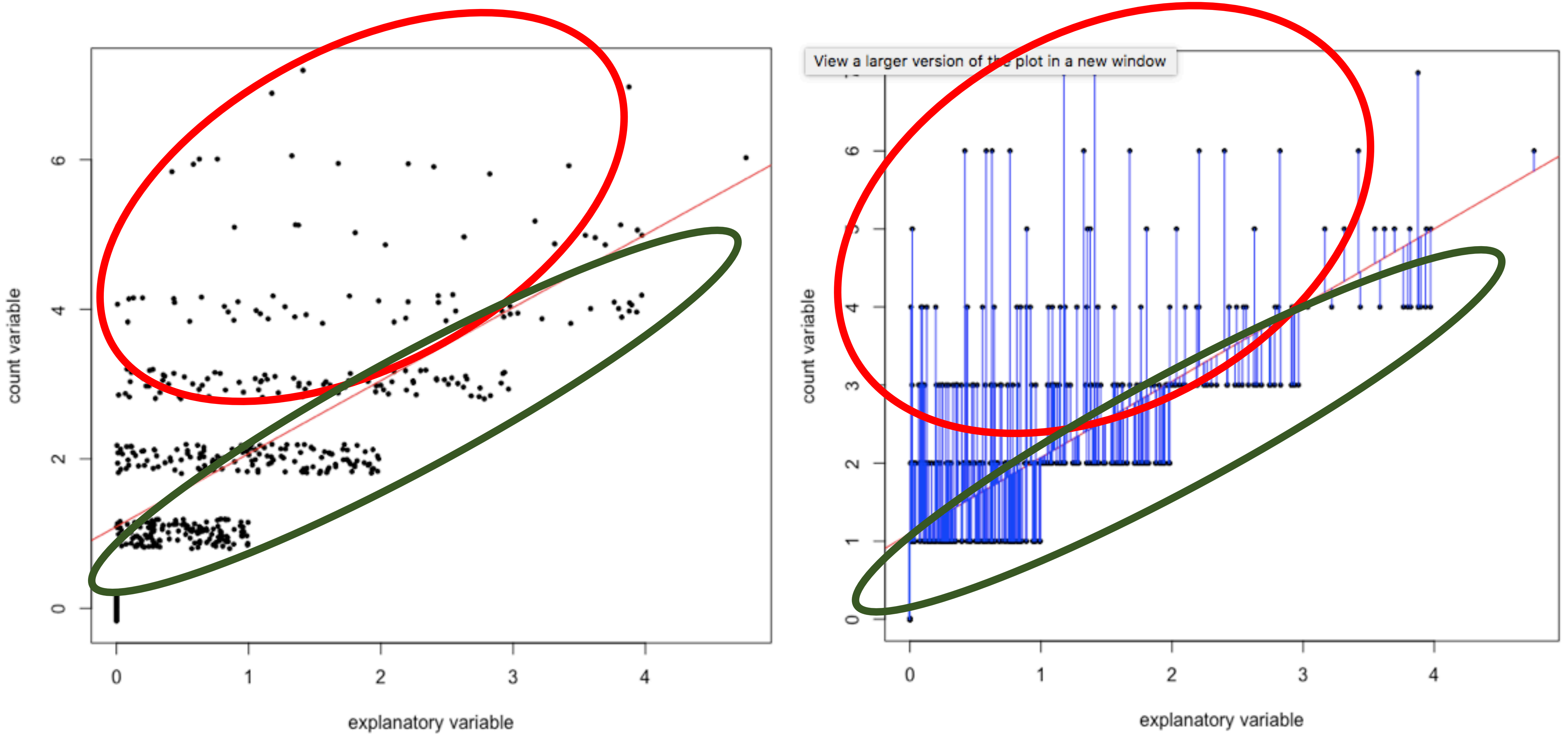


# Count data as response variable



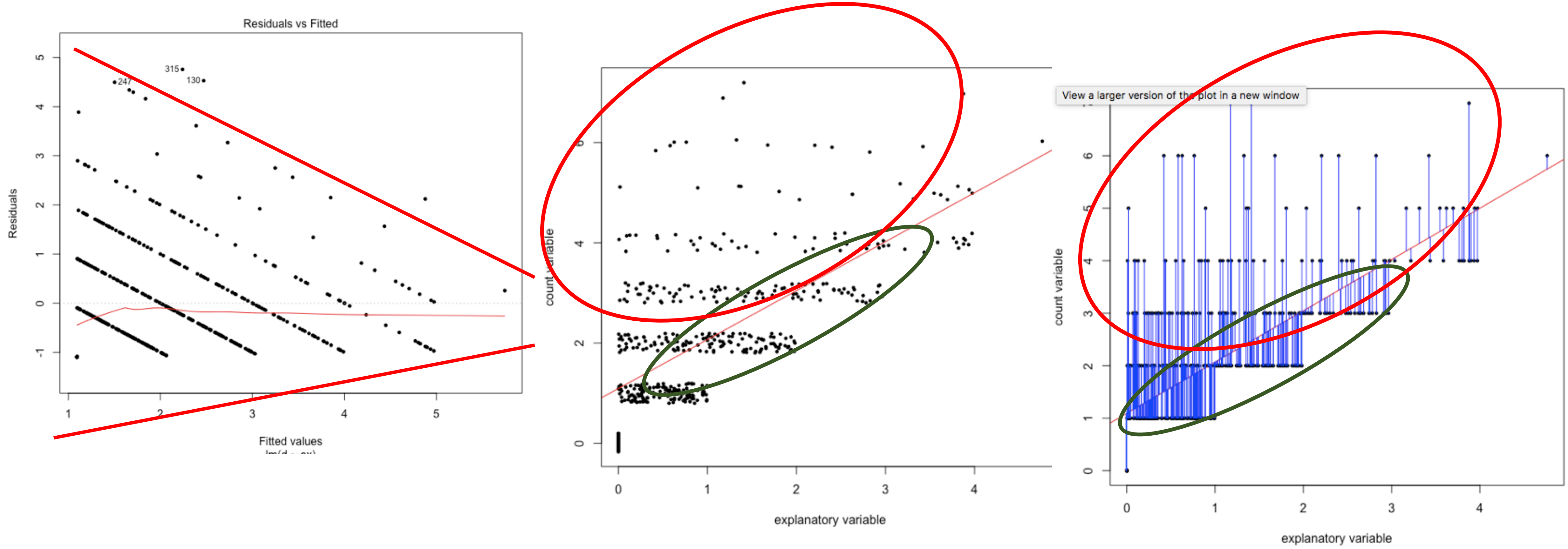
# Count data as response variable



# What is the problem here?

# What is the problem here?

- Residual variance decreases with increasing Y



# What is the problem here?

- Residual variance decreases with increasing Y
- Still a problem when transformed
  - And then, zeroes are hard to account for

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

# Assumptions of a linear model:

$$\begin{pmatrix} \widehat{Y_1} \\ \widehat{Y_2} \\ \\ .. \\ \\ \widehat{Y_i} \end{pmatrix}$$

$$\hat{Y} = \text{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix}$$

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

# Assumptions of a linear model:

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \\ .. \\ \beta_i \end{bmatrix}$$

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \quad \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ .. \\ \varepsilon_i \end{pmatrix} \quad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

$$\begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_i \end{pmatrix}$$

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

follows approximately

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

follows approximately

a normal distribution

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Parameter estimates of linear model

Predicted values

Explanatory variable values

follows approximately

With a mean of the mean of the response

a normal distribution

Observed data

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2) \quad \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \\ .. \\ \overline{y_i} \end{bmatrix}$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

follows approximately

a normal distribution

With a mean of the mean of the response

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

follows approximately

a normal distribution

With a mean of the mean of the response

And a constant variance

# Assumptions of a linear model:

Error

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

Predicted values

Explanatory variable values

Parameter estimates of linear model

Observed data

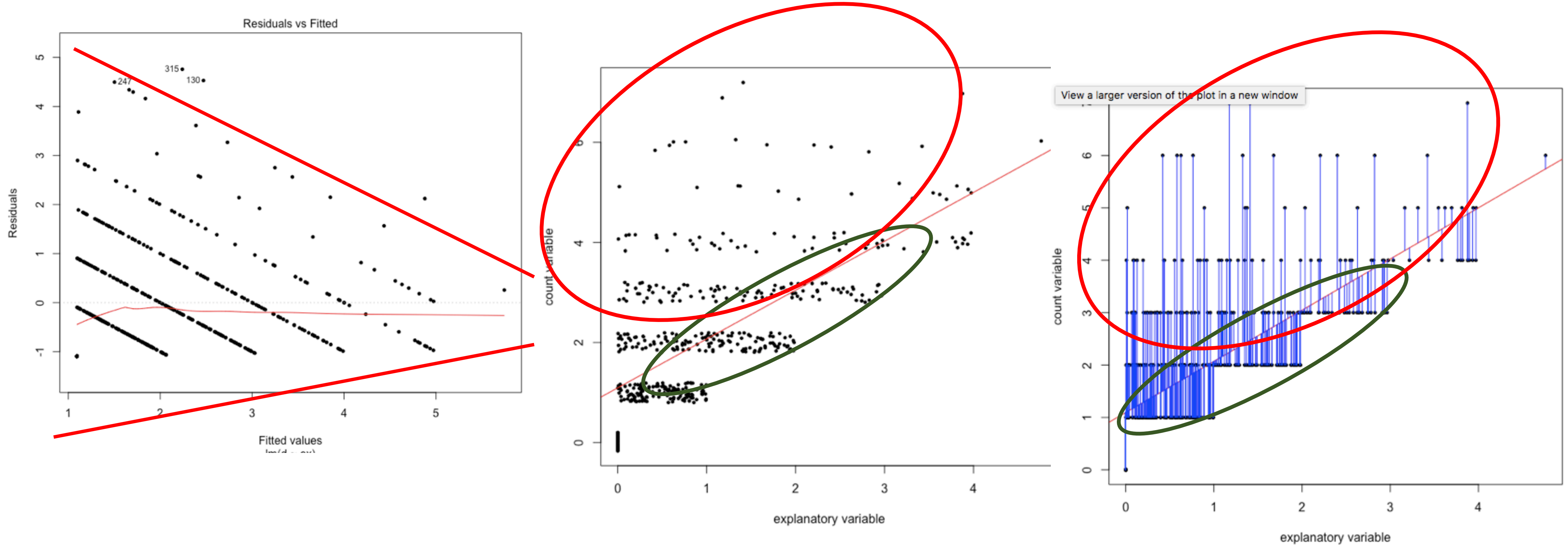
follows approximately

a normal distribution

With a mean of the mean of the response

And a **constant** variance

# Variance CHANGES with y



# This clearly does not work for count data

- Variance increases with predicted mean

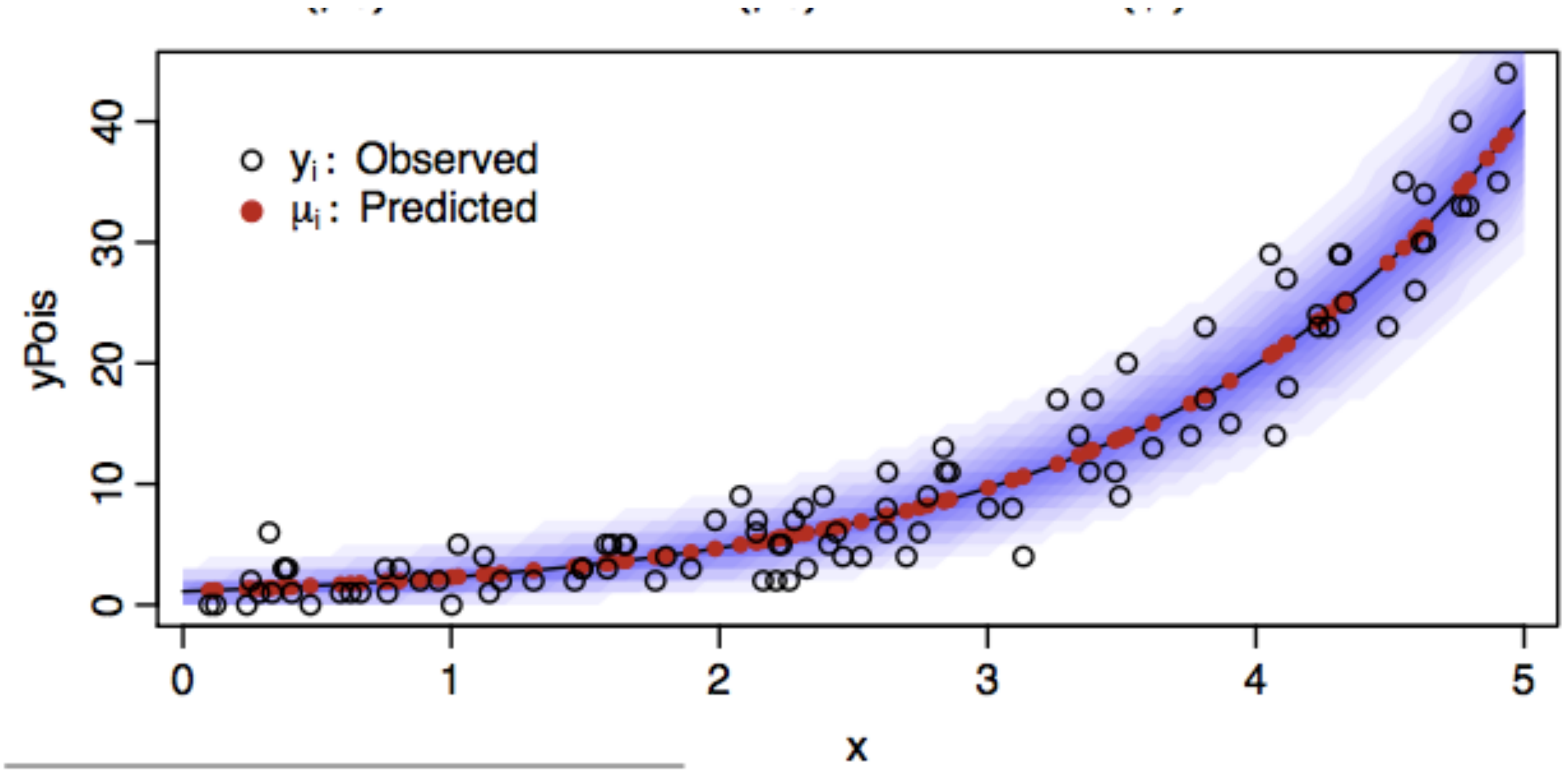
# This clearly does not work for count data

- Variance increases with predicted mean
- Errors follow *Poisson* distribution

# What is a *Poisson* distribution?

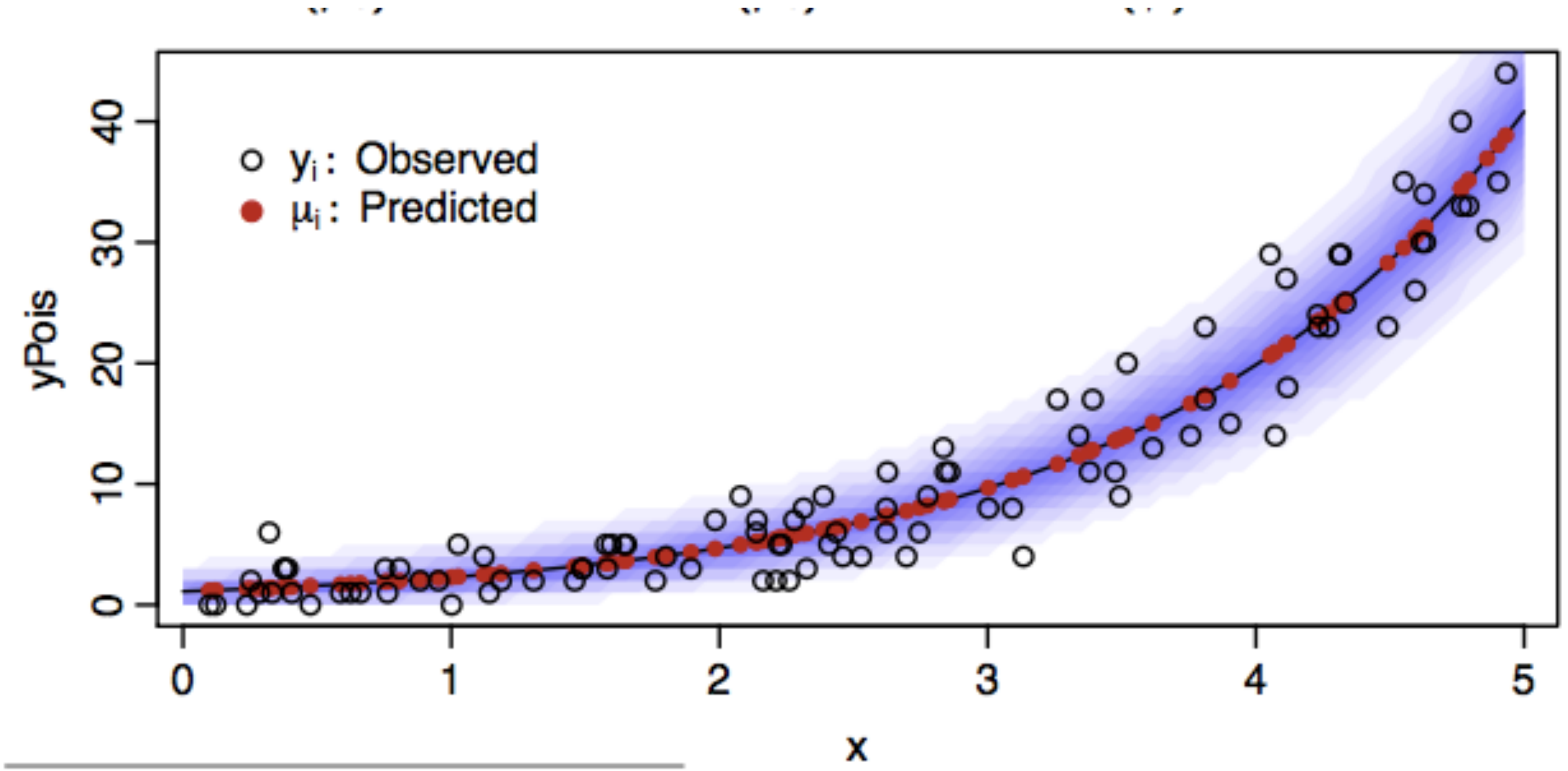
# What is a *Poisson* distribution?

- Variance increases with predicted mean



# What is a *Poisson* distribution?

- Variance increases with predicted mean




Variance = mean

# GLMs to the rescue!

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models
- Allow response to have arbitrary distributions

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models
- Allow response to have ”arbitrary” distributions
- And an “arbitrary” link function of the response that then varies linearly with the predicted values

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models
- Allow response to have ”arbitrary” distributions
- And an “arbitrary” link function of the response that then varies linearly with the predicted values

- “arbitrary” because there are some commonly used ones that we’ll cover:

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models
- Allow response to have ”arbitrary” distributions
- And an “arbitrary” link function of the response that then varies linearly with the predicted values

- “arbitrary” because there are some commonly used ones that we’ll cover:
- *Log link* and *Logit link*

# GLMs to the rescue!

- GLM – Genera**lized** Linear Models
- Allow response to have ”arbitrary” distributions
- And an “arbitrary” link function of the response that then varies linearly with the predicted values

- “arbitrary” because there are some commonly used ones that we’ll cover:
- *Log link* and *Logit link* for
- COUNT and BINARY data

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB} + E, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \mathrm{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \mathrm{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

A function of the response

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \mathrm{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

A function of the response

Poisson distribution

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \mathrm{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

A function of the response

Poisson distribution

With mean = variance = expected

# Count data – log link– *Poisson* model

$$\hat{Y} = \text{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \text{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

$$\begin{pmatrix} \overline{y_1} \\ \overline{y_2} \\ .. \\ \overline{y_i} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ .. \\ \sigma_i^2 \end{pmatrix}$$

A function of the response

Poisson distribution

With mean = variance = expected

# Count data – log link– *Poisson* model

$$\hat{Y} = \mathrm{XB}, \qquad y_i \sim N(\bar{Y}_i, \sigma^2)$$

$$g(\hat{Y}) = \mathrm{XB}, \qquad y_i \sim Poi(\bar{Y}_i, \phi)$$

A function of the response

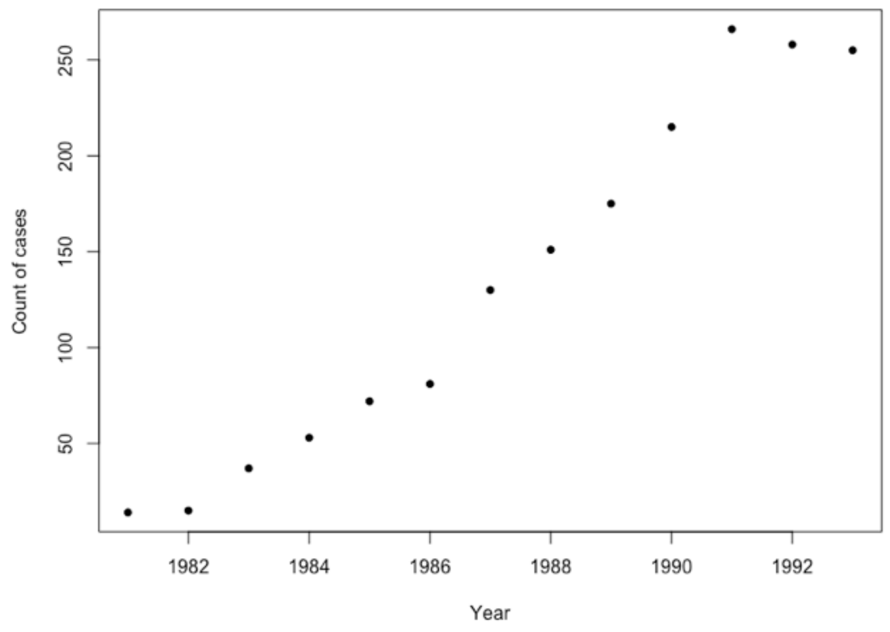
Poisson distribution

With mean = variance = expected

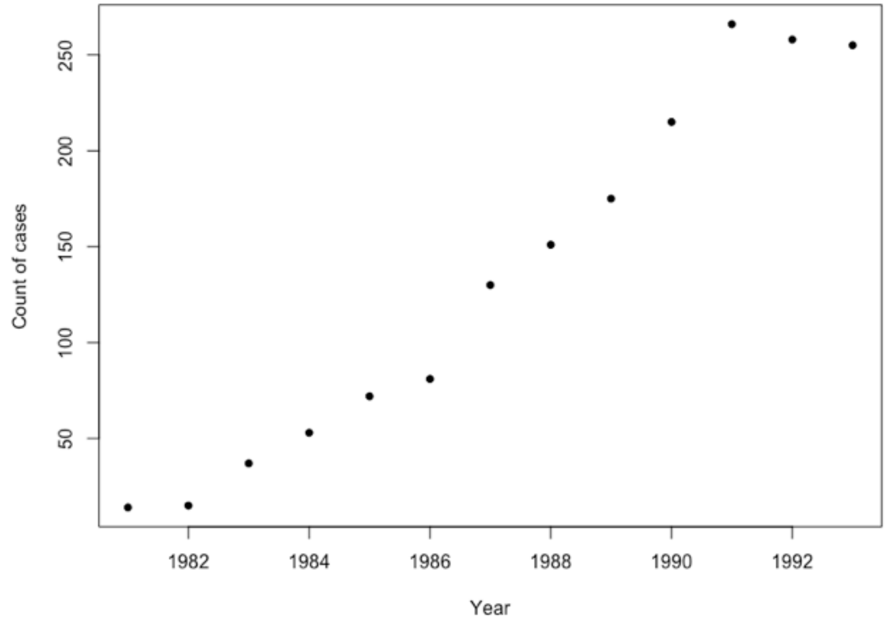
Scale parameter

# How to interpret?

# How to interpret?



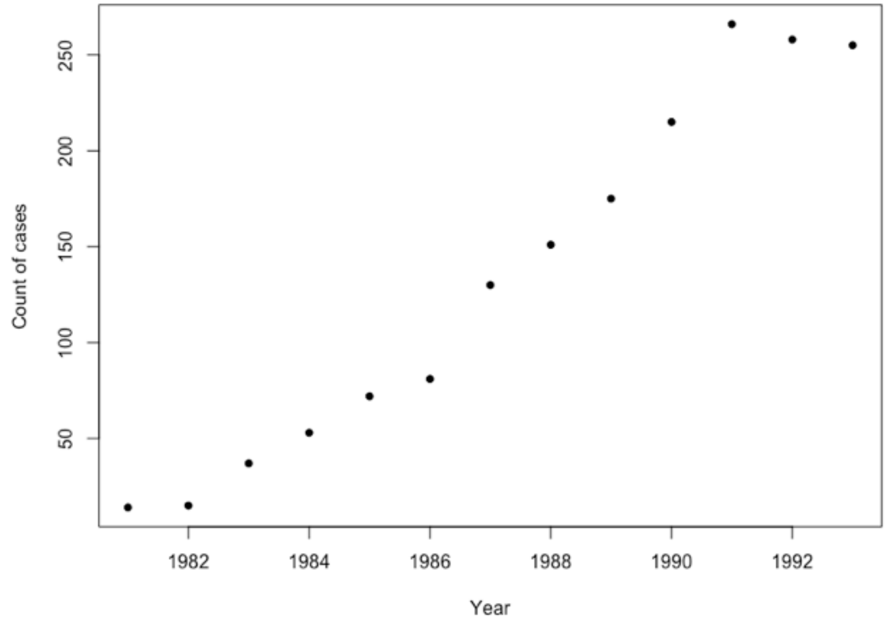
# How to interpret?




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

# How to interpret?

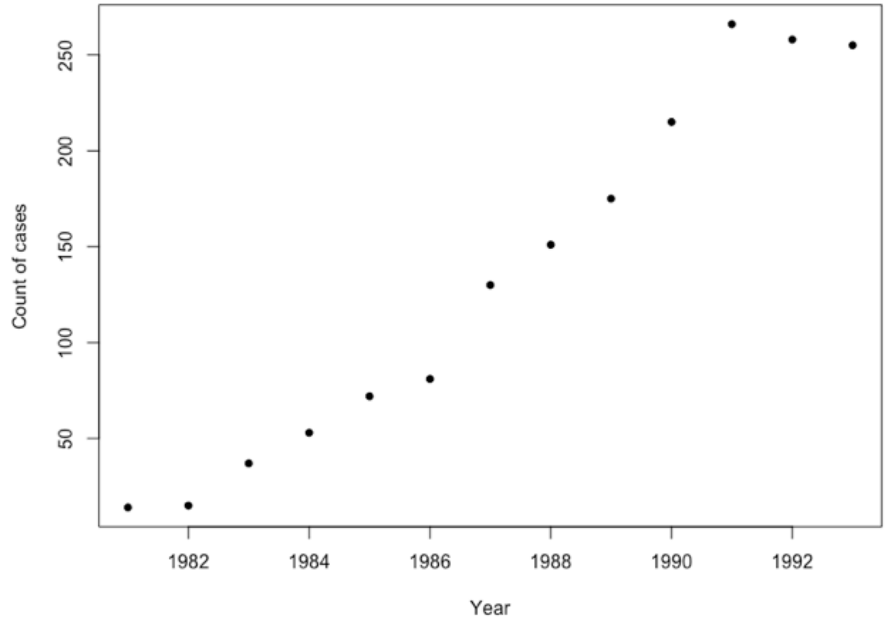
$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

# How to interpret?

$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$

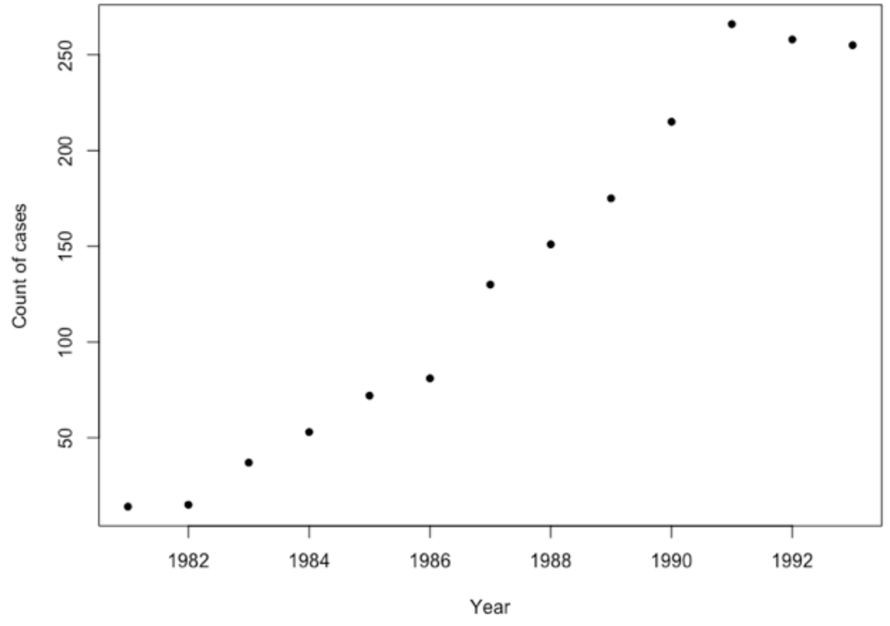



```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

Intercept. Cases when year = 0. Irrelevant

# How to interpret?

$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

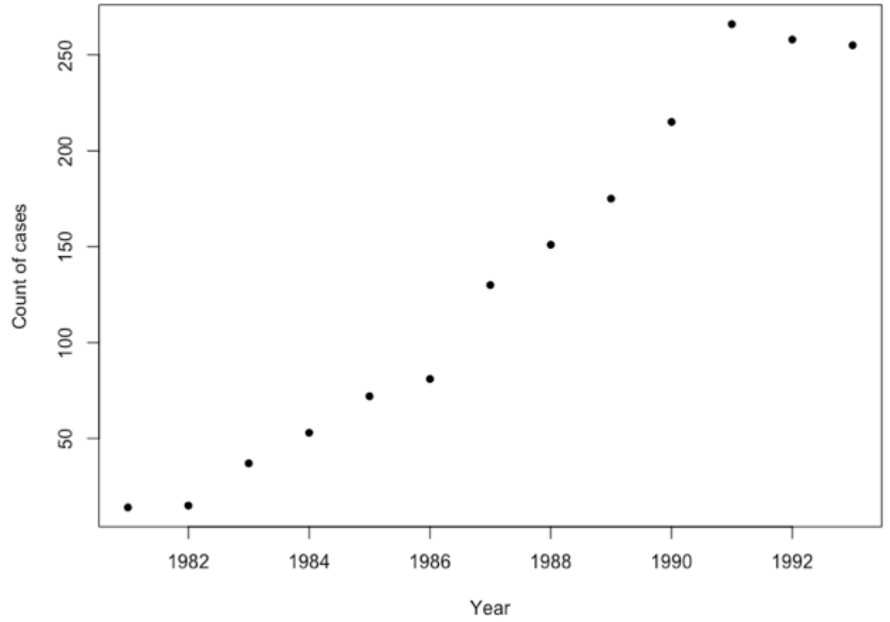
Intercept. Cases when year = 0. Irrelevant

Slope. Increase in cases over time. On LINK scale.

# How to interpret?

$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$

$$e^{b_1} = e^{0.2}$$




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

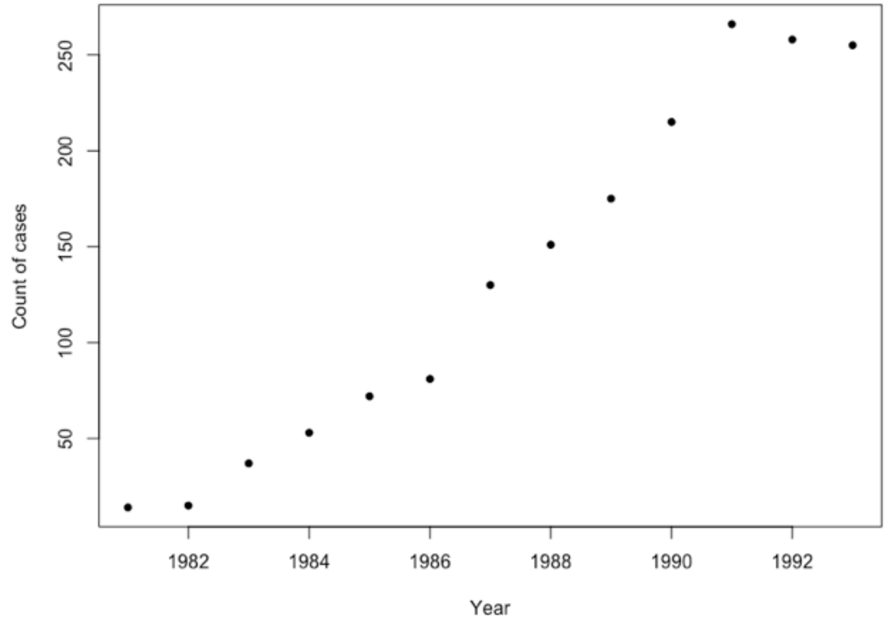
Intercept. Cases when year = 0. Irrelevant

Slope. Increase in cases over time. On LINK scale.

# How to interpret?

$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$

$$e^{b_1} = e^{0.2} = 1.22$$




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

Intercept. Cases when year = 0. Irrelevant

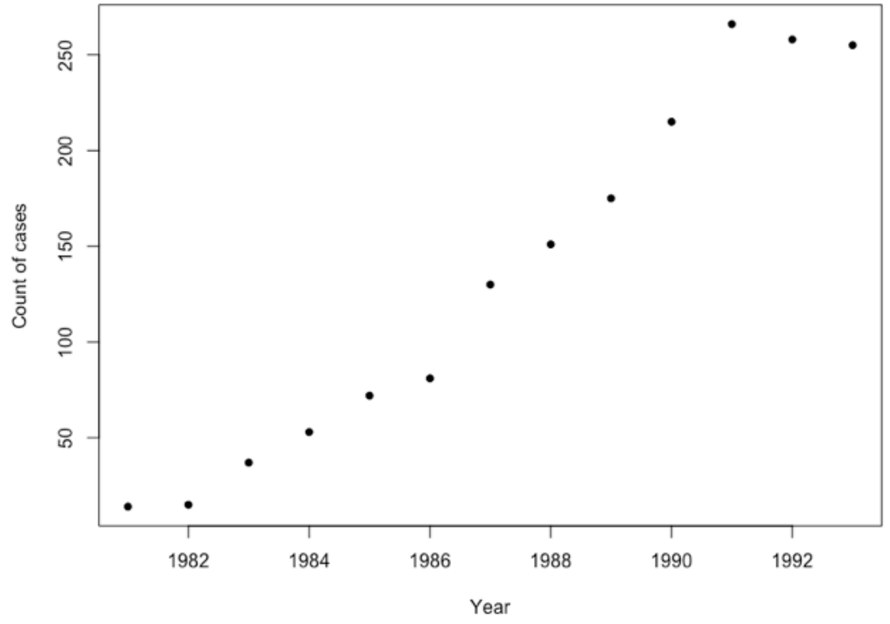
Slope. Increase in cases over time. On LINK scale.

# How to interpret?

$$y_i = Poisson(\exp(-39.15 + 0.20x_{i1}))$$

$$e^{b_1} = e^{0.2} = 1.22$$

Every consecutive year, there are 1.22% more cases than the year before




```
> summary(m1)

Call:
glm(formula = cases ~ yr, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6398  -1.3494  -0.2574   2.1500   2.6866

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.915e+02  1.495e+01  -26.19   <2e-16 ***
yr           1.994e-01  7.513e-03   26.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

Intercept. Cases when year = 0. Irrelevant

Slope. Increase in cases over time. On LINK scale.

# GLMs model selection

- We use DEVIANCE to select models

# GLMs and residuals

- We use DEVIANCE to select models
- Residuals are deviance
- But not all deviance are residuals

# GLMs and residuals

- We use DEVIANCE to select models
- Residuals are deviance

# GLMs and residuals

- We use DEVIANCE to select models
- Residuals are deviance
- But not all deviance are residuals

# GLMs and residuals

- We use DEVIANCE to select models
- Residuals are deviance
- But not all deviance are residuals

- Deviance is a similar idea, but slightly different estimation
- Not based on sums of squares

# GLMs and residuals

- We use DEVIANCE to select models
- Residuals are deviance
- But not all deviance are residuals

- Deviance is a similar idea, but slightly different estimation
- Not based on sums of squares
- You used deviance in likelihood ratio test, or AIC to select best models

# Detour - AIC

# Akaike information criterion

- Better than logL test as it includes complexity

# Akaike information criterion

- Better than logL test as it includes complexity
- Penalizes for low df

# Akaike information criterion

- Better than logL test as it includes complexity
- Penalizes for low df

$$AIC = 2\,(n - df) - 2lnL$$

# Akaike information criterion

- Better than logL test as it includes complexity
- Penalizes for low df
- Model with lower AIC = better

$$AIC = 2\,(n - df) - 2lnL$$

# Akaike information criterion

- Better than logL test as it includes complexity
- Penalizes for low df
- Model with lower AIC = better

$$AIC = 2\,(n - df) - 2lnL$$

- Rule of thumb: $\Delta AIC < 2$ not statistically significant

# AIC in R

```
> m0<-lm(y~1)
> m1<-lm(y~x)
```

# AIC in R

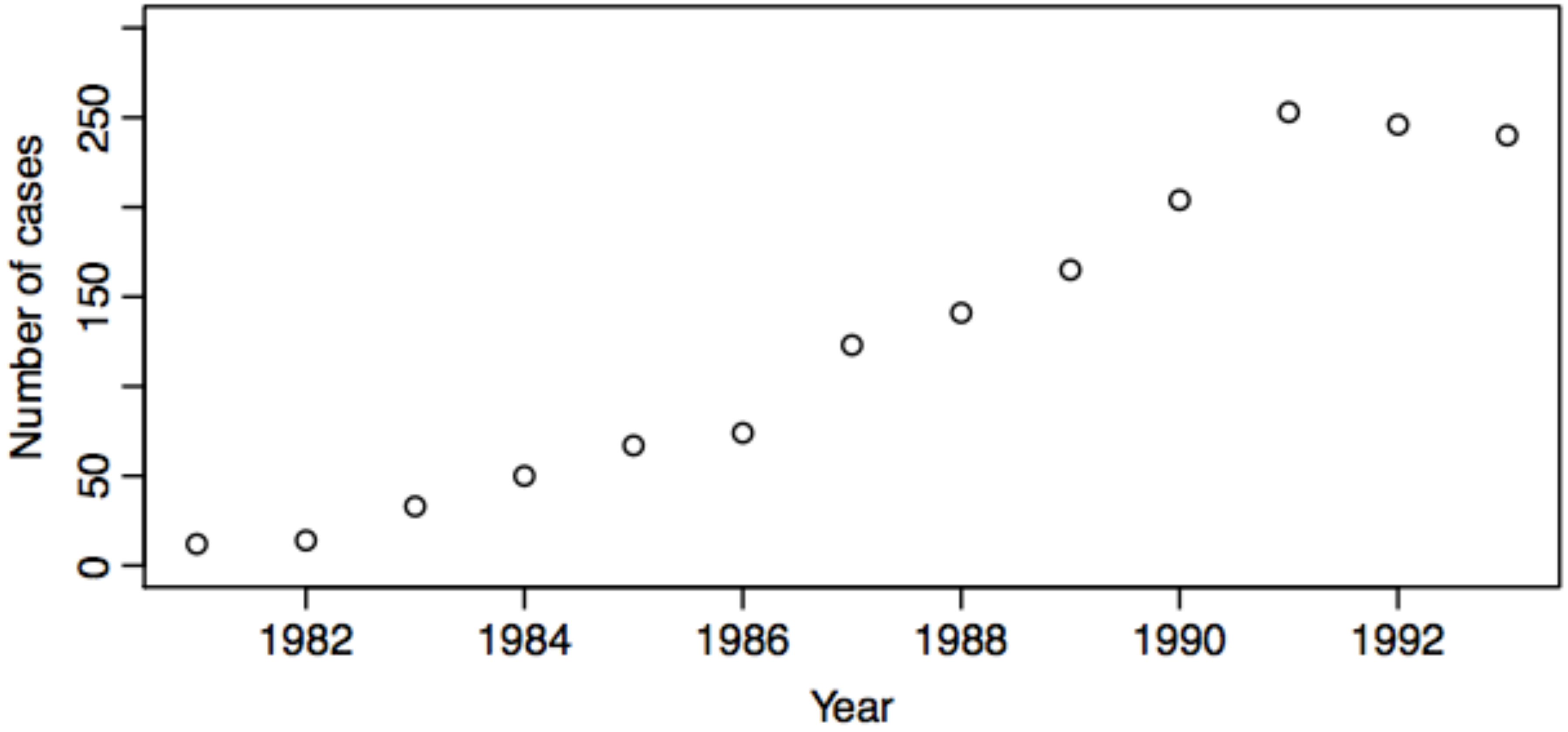
```
> m0<-lm(y~1)
> m1<-lm(y~x)
> AIC(m0)
[1] 38.97657
> AIC(m1)
[1] 17.36793
```

# AIC in R

```
> m0<-lm(y~1)
> m1<-lm(y~x)
> AIC(m0)
[1] 38.97657
> AIC(m1)
[1] 17.36793
> AIC(m0)-AIC(m1)
[1] 21.60864
> |
```

# Example

- Number of AIDS cases in Belgium from 1981 to 2013.
- Models with linear, quadratic and cubic linear predictors.



# Example

- The code to fit a GLM in R is very similar to a linear model.
- Specify an error structure and a link function.

# Example

- The code to fit a GLM in R is very similar to a linear model.
- Specify an error structure and a link function.

```
null <- glm(cases ~ 1, data = belg.aids,
    family = poisson(link = log))
```

# Example

- The code to fit a GLM in R is very similar to a linear model.
- Specify an error structure and a link function.

```
null <- glm(cases ~ 1, data = belg.aids,
    family = poisson(link = log))
am1 <- glm(cases ~ year, data = belg.aids,
    family = poisson(link = log))
```

# Example

- The code to fit a GLM in R is very similar to a linear model.
- Specify an error structure and a link function.

```
null <- glm(cases ~ 1, data = belg.aids,
    family = poisson(link = log))
am1 <- glm(cases ~ year, data = belg.aids,
    family = poisson(link = log))
am2 <- glm(cases ~ year + I(year^2), data = belg.aids,
    family = poisson(link = log))
```

# Example

- The code to fit a GLM in R is very similar to a linear model.
- Specify an error structure and a link function.

```
null <- glm(cases ~ 1, data = belg.aids,
    family = poisson(link = log))
am1 <- glm(cases ~ year, data = belg.aids,
    family = poisson(link = log))
am2 <- glm(cases ~ year + I(year^2), data = belg.aids,
    family = poisson(link = log))
am3 <- glm(cases ~ year + I(year^2) + I(year^3),
    data = belg.aids, family = poisson(link = log))
```

# Example

```
AIC(null, am1, am2, am3)
##       df   AIC
## null   1 955.9
## am1    2 166.4
## am2    3  96.9
## am3    4  98.7
```

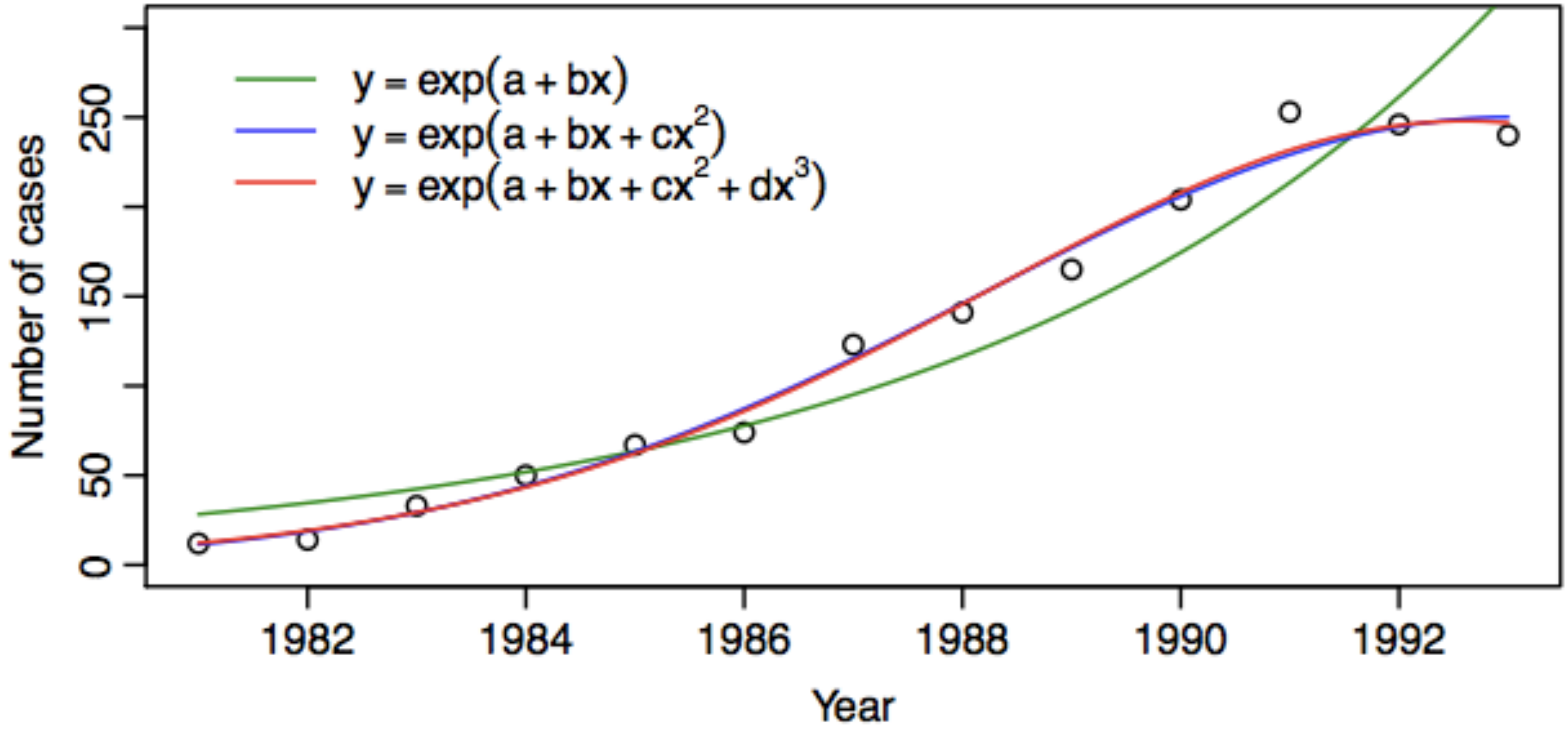
# Example

- And we can use linear models summaries to look at the significance of coefficients
- Importantly, model coefficients are estimated and reported *on the scale of the linear predictor*

```
summary(am2)

...
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.48e+04   1.05e+04   -8.07  7.3e-16
## year         8.51e+01   1.06e+01    8.05  8.4e-16
## I(year^2)   -2.14e-02   2.66e-03   -8.03  9.8e-16
##
...
```

# Example

- The model is described on the scale of the linear predictor
- Showing those models on the original data:



# HO – Poisson models