Limitations of linear models

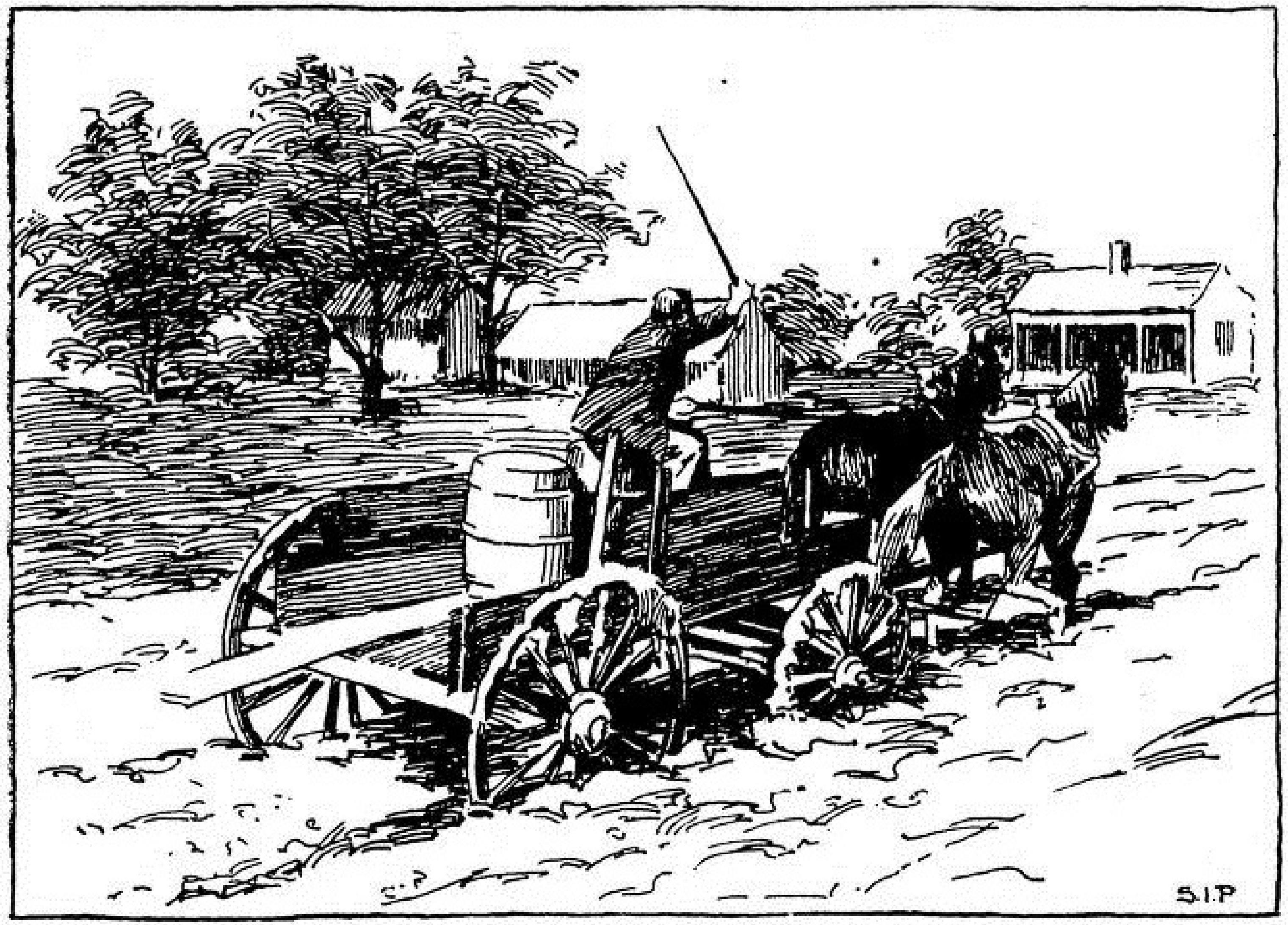
GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

Course overview

- Chapter 1: Review and limits of linear model and Poisson regressions
- Chapter 2: Logistic (Binomial) regression
- Chapter 3: Interpreting and plotting GLMs
- Chapter 4: Multiple regression with GLMs

Workhorse of data science



[1] US Department of Agriculture https://www.nal.usda.gov/exhibits/ipd/localfoods/exhibits/show/farmto-table/the-roads-of-rural-america

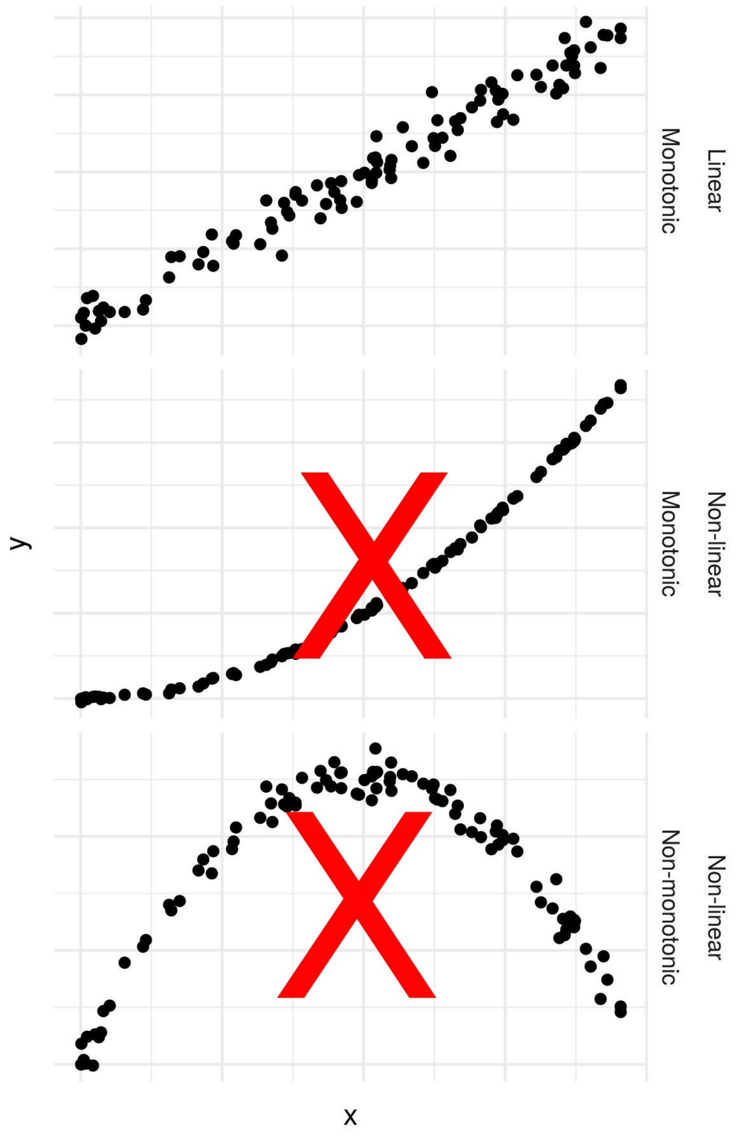
Linear models

- How can linear coefficients explain the data?
- Intercept for baseline effect
- Slope for linear predictor
- $y = \beta_0 + \beta_1 x + \epsilon$

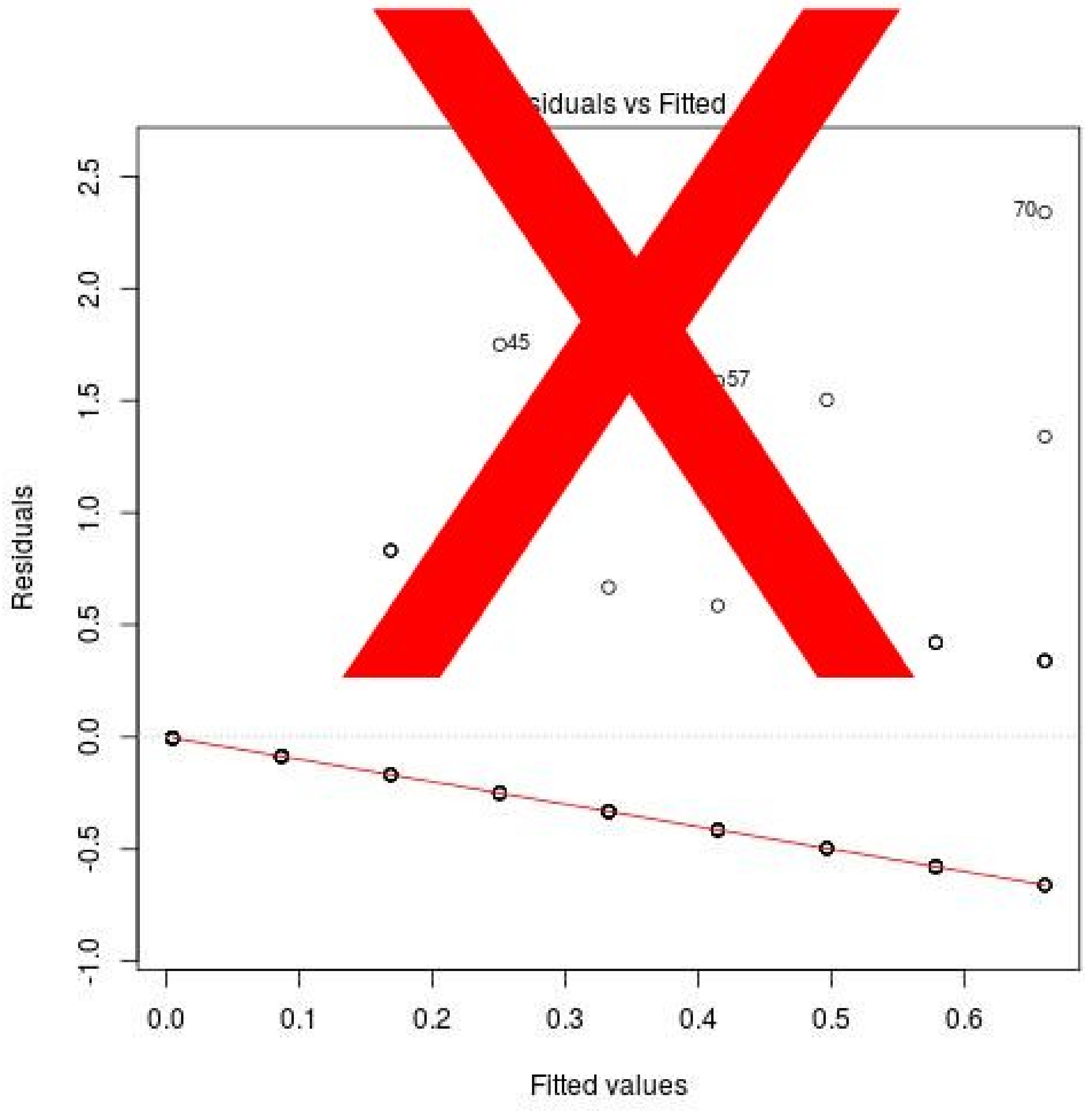
Linear models in R

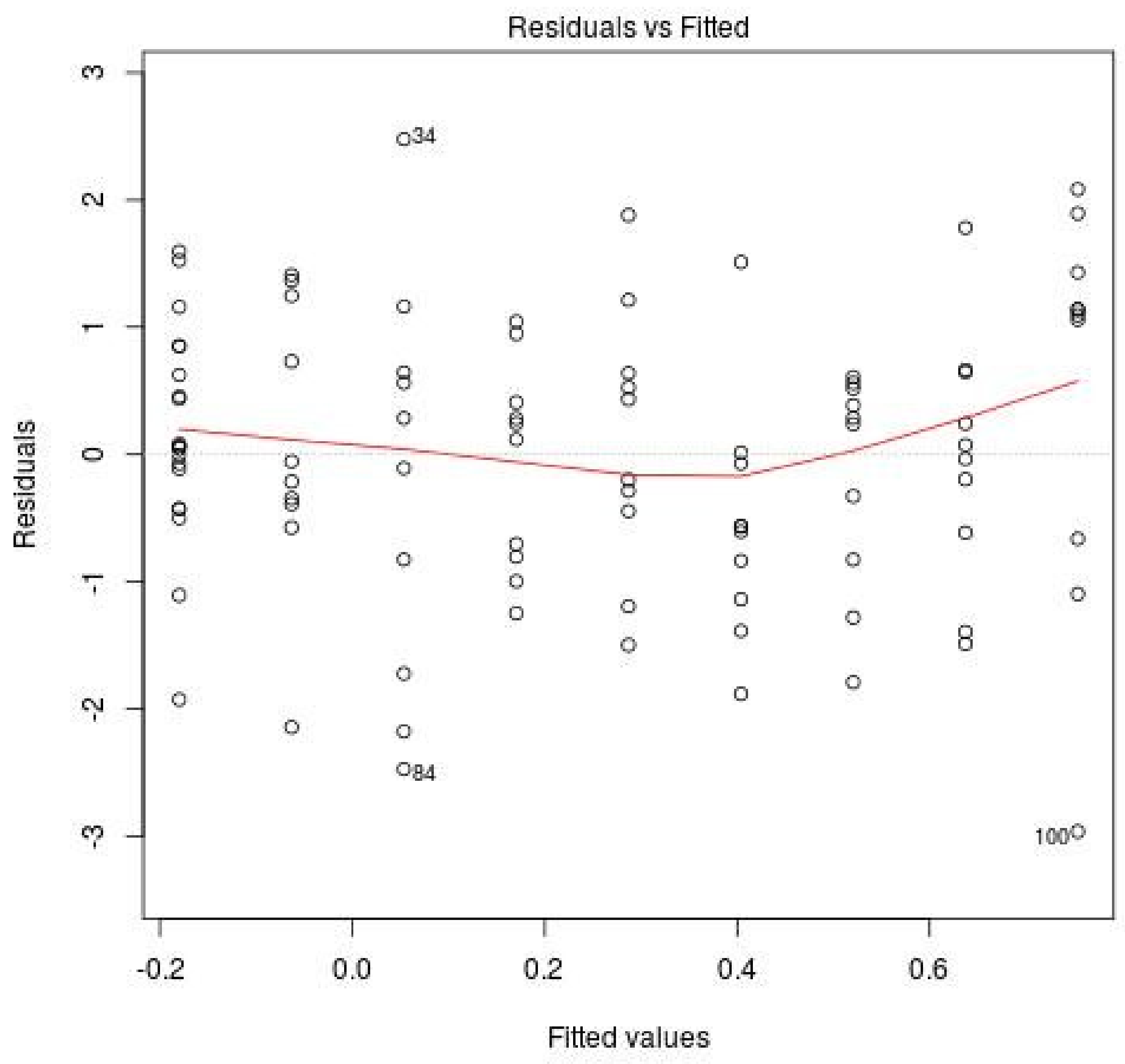
```
lm(y ~ x, data = dat)
```

Assumption of linearity

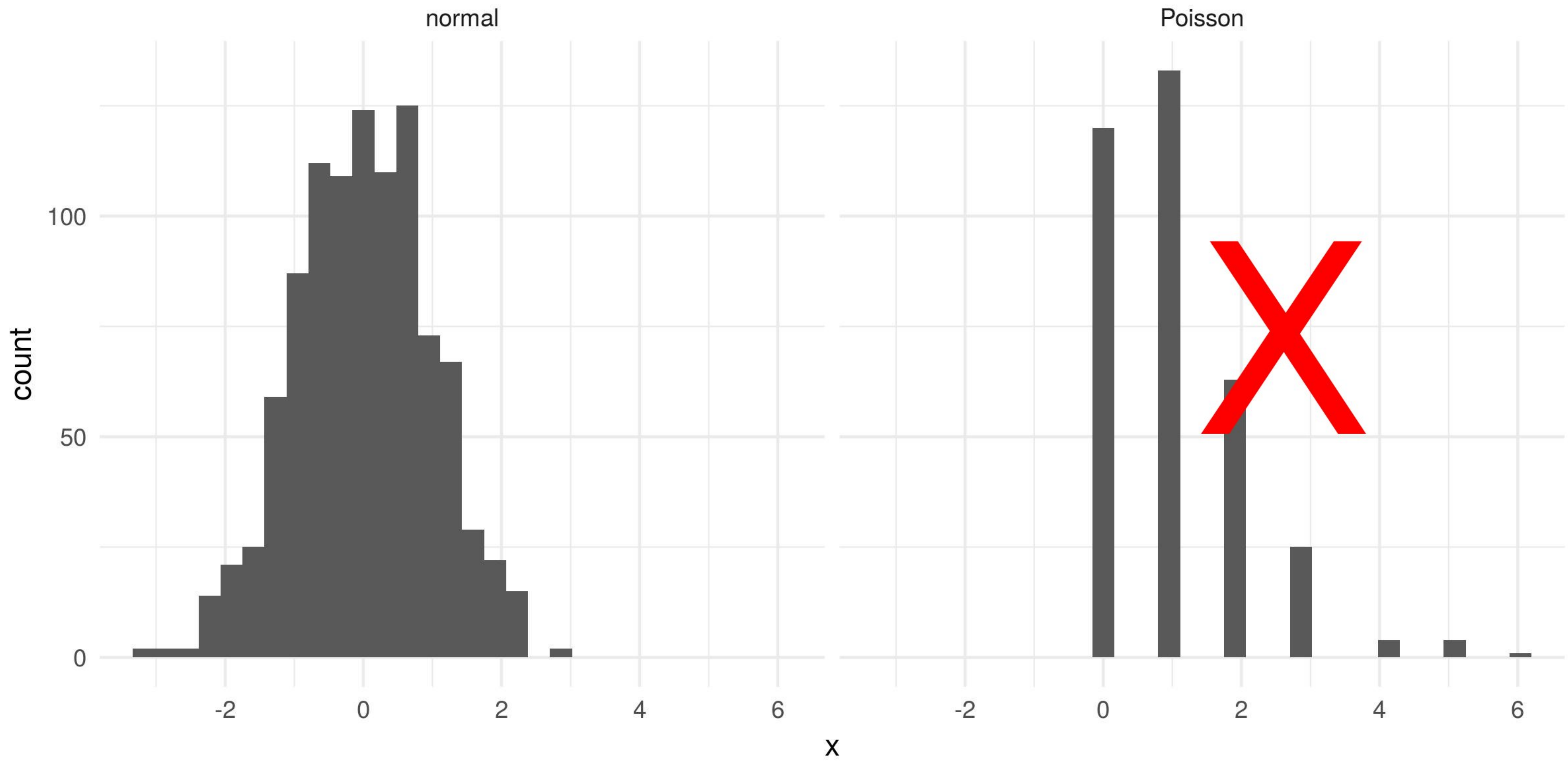


Assumption of normality

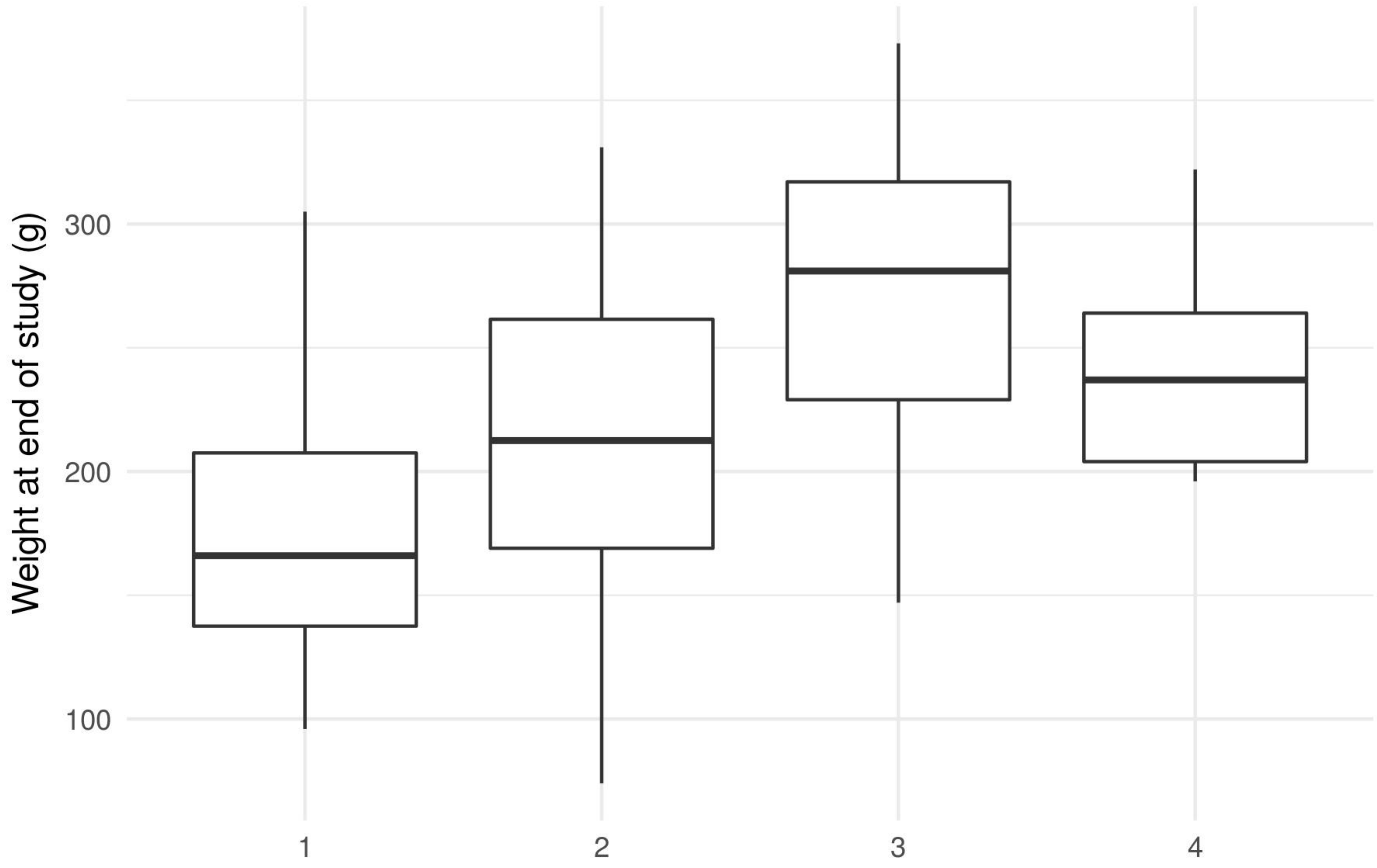




Assumption of continuous variables



-


Weight at end of study (g)
300
200
100
1
2
3
4

Chick diets impact on weight

- `ChickWeight` data from `datasets` package
- `ChickWeightsEnd` last observation from study
- How do diets 2, 3, and 4 compare to diet 1?

```
lm(formula = weight ~ Diet, data = ChickWeightEnd)
```

```
Call:
lm(formula = weight ~ Diet, data = ChickWeightEnd)
Coefficients:
(Intercept)        Diet2        Diet3        Diet4
     177.75        36.95        92.55        60.81
```

What about survivorship or counts?

- What about chick survivorship or chick counts?
- Neither are continuous!
- We need a new tool
- The generalized linear model

Generalized linear model

- Similar to linear models
- **Non-normal error distribution**
- **Link functions:** $y = \psi(b_0 + b_1 x + \epsilon)$

GLMs in R

```
glm( y ~ x, data = data, family = "gaussian")
```

- `lm()` same as `glm( ..., family = "gaussian")`

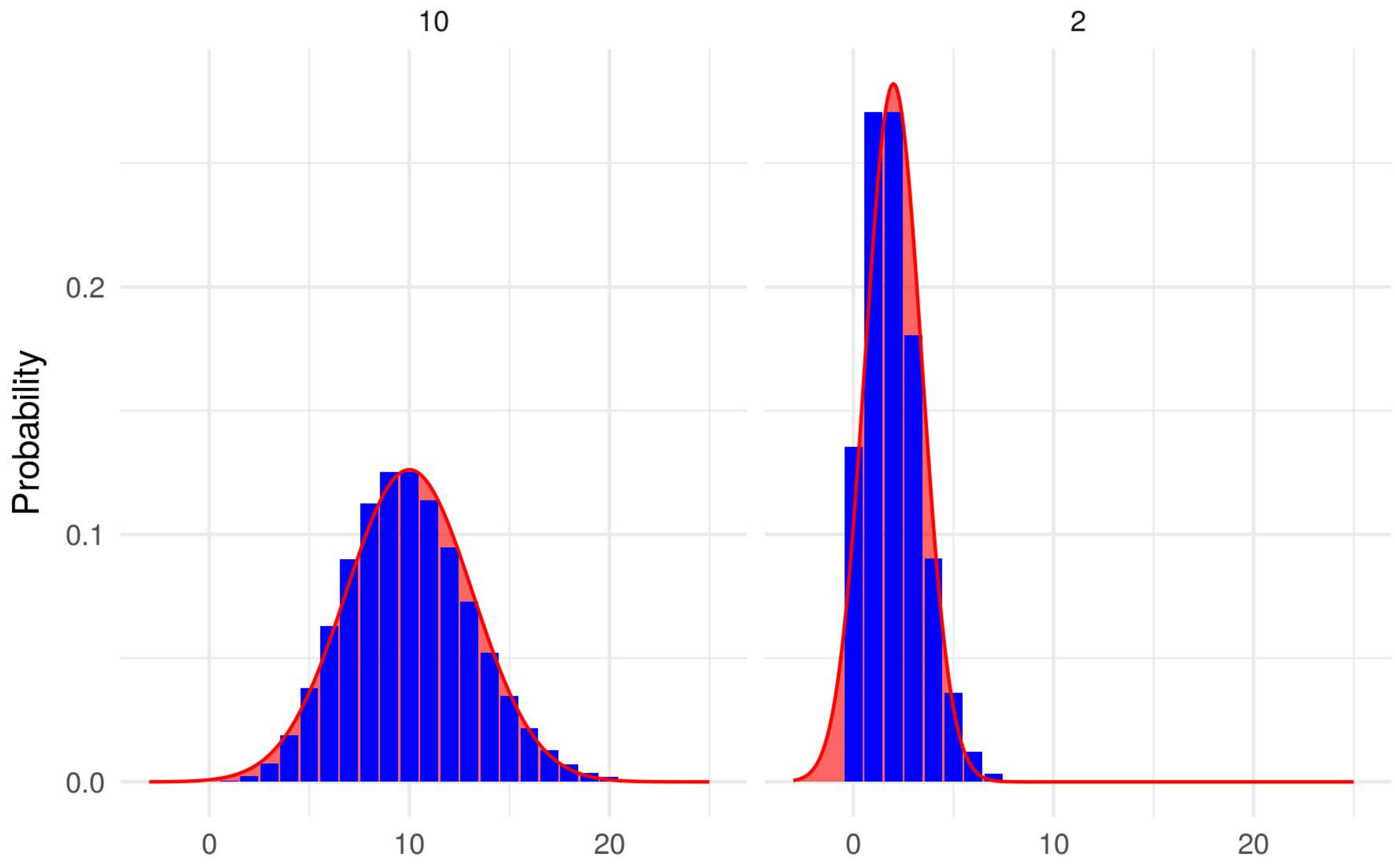
Let's practice!!

GENERALIZED LINEAR MODELS IN R

Poisson regression

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor


10
2
Probability
0.2
0.1
0.0
0
10
20
0
10
20

Poisson distribution

- Discrete integers: x = 0, 1, 2, 3, ...
- Mean and variance parameter λ
- $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- Fixed area/time (e.g., goal per one game)

Poisson distribution in R

```
dpois(x = ..., lambda = ...)
```

GLM with R requirements

- Discrete counts: 0, 1, 2, 3...
- Defined area and time
- Log-scale coefficients

GLM with Poisson in R

```
glm(y ~ x, data = dat, family = 'poisson')
```

When not to use Poisson distribution

- Non-count or non-positive data (e.g., 1.4 or -2)
- Non-constant sample area or time (e.g., trees km^{-1} vs. trees m^{-1})
- Mean $\gtrsim$30
- Over-dispersed data
- Zero-inflated data

Formula intercepts

- Comparison or intercept
- Comparison `formula = y ~ x`
- Intercept `formula = y ~ x - 1`

Goals per game

- Two players, which approach do we use?
- If we want to know difference between players, use comparison:

```
glm(goal ~ player, data = scores, family = "poisson")
```

- If we want to know average per player, use intercepts:

```
glm(goal ~ player - 1, data = scores, family = "poisson")
```

Let's practice!

GENERALIZED LINEAR MODELS IN R

Basic lm() functions with glm()

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

Interacting with model objects

- Allow interaction with outputs
- Base R functions apply to `glm()`
- Useful shortcuts

Model print

- `print()` usually default

```
print(poisson_out)
```

```
Call:  glm(formula = y ~ x, family = "poisson", data = dat)

Coefficients:
(Intercept)            x
   -1.43036      0.05815

Degrees of Freedom: 29 Total (i.e. Null);  28 Residual
Null Deviance:        35.63
Residual Deviance: 30.92     AIC: 66.02
```

Model summary

- `summary()` provides more details

```
summary(poisson_out)
```

```
#...
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6547  -0.9666  -0.7226   0.3830   2.3022
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.43036    0.59004  -2.424   0.0153 *
x            0.05815    0.02779   2.093   0.0364 *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 35.627  on 29  degrees of freedom
Residual deviance: 30.918  on 28  degrees of freedom
AIC: 66.024
Number of Fisher Scoring iterations: 5
```

Tidy output

- **Tidyverse** provides standardized model outputs
- `tidy()` from **Broom package**

```
library(broom)
tidy(poisson_out)
```

```
          term    estimate  std.error statistic    p.value
1 (Intercept) -1.43035579 0.59003923 -2.424171 0.01534339
2           x  0.05814858 0.02778801  2.092578 0.03638686
```

Regression coefficients

- `coef()` prints regression coefficients

```
coef(poisson_out)
```

```
(Intercept)           x
-1.43035579  0.05814858
```

Confidence intervals

- `confint()` estimates the confidence intervals

```
confint(poisson_out)
```

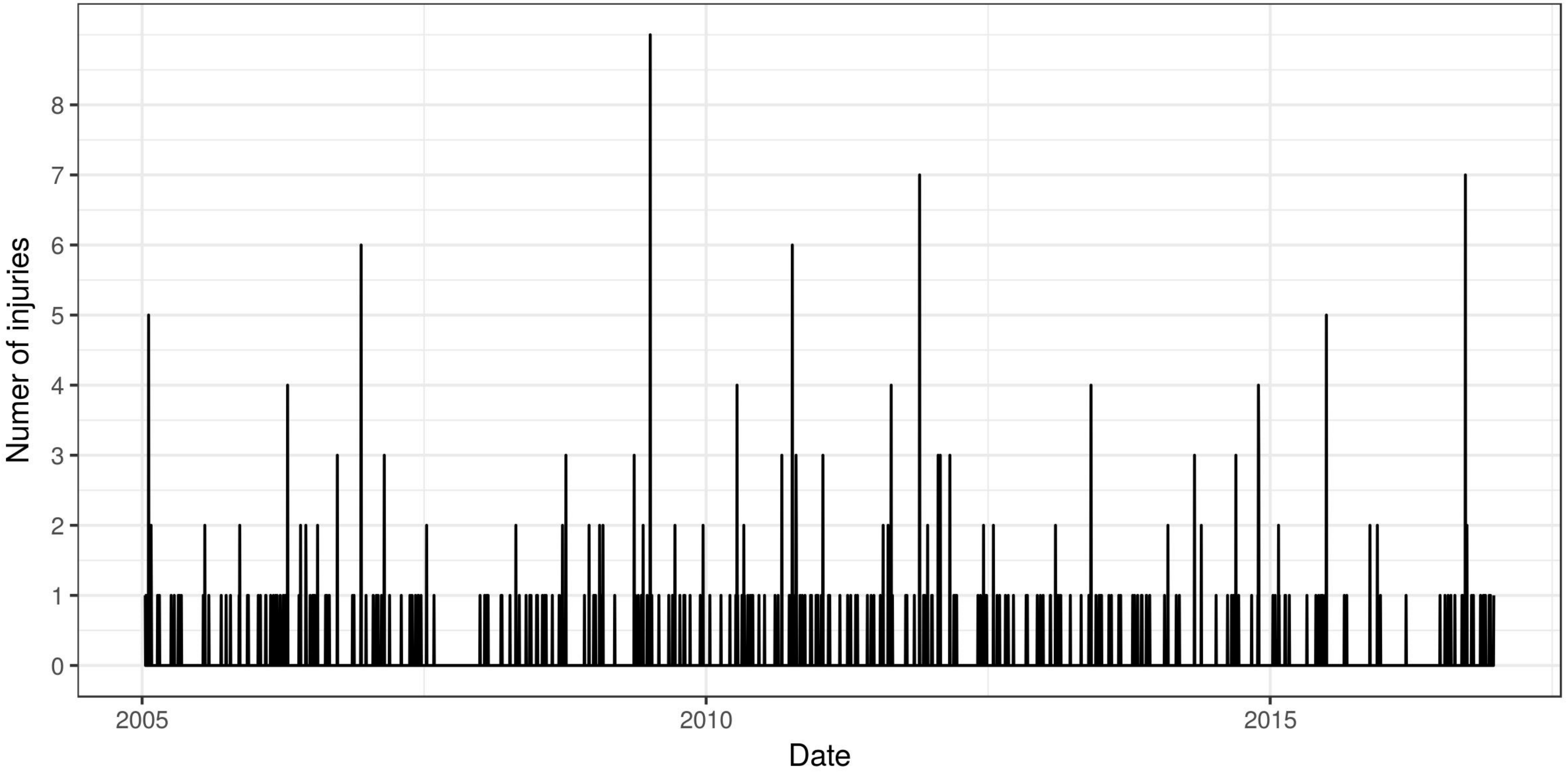
```
Waiting for profiling to be done...
                   2.5 %     97.5 %
(Intercept) -2.725545344 -0.3897748
x            0.005500767  0.1155564
```

Predictions

- `predict(model, new_data)`
- `new_data` argument:
  - Unspecified: `predict()` returns predictions based on original data used to fit the model.
  - Specified: `predict()` returns predictions for `new_data` .

Fire injury dataset

- Daily civilian injuries
- Louisville, KY
- Count data, many zeros



Let's practice!

GENERALIZED LINEAR MODELS IN R