Multiple logistic regression

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

Chapter overview

- Multiple logistic regression
- Formulas in R
- Model assumptions

Why multiple regression?

Problem: Multiple predictor variables. Which one should I include?

Solution: Include all of them using multiple regression.

Multiple predictor variables

- Simple linear models or simple GLM:
  - Limited to 1 Slope and 1 intercept
  - $y \sim \beta_0 + \beta_1 x + \epsilon$
- Multiple regression
  - Multiple slopes and intercepts:
  - $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x + \beta_3 x_3 \ldots + \epsilon$

Too much of a good thing

Theoretical maximum number of coefficients:

Maximum number of βs = Number of observations

Over-fitting:

Using too many predictors compared to number of samples

Practical maximum number of coefficients:

Number of $\beta \times 10 \approx$ Number of observations

Bus data: Two possible predictors

- With bus commuter data, 2 possible predictors
  - Number of days one commutes: `CommuteDay`
  - Distance of commute: `MilesOneWay`
- Possible to build a model with both

```
glm(Bus ~ CommuteDay + MilesOneWay, data = bus,  family = 'binomial')
```

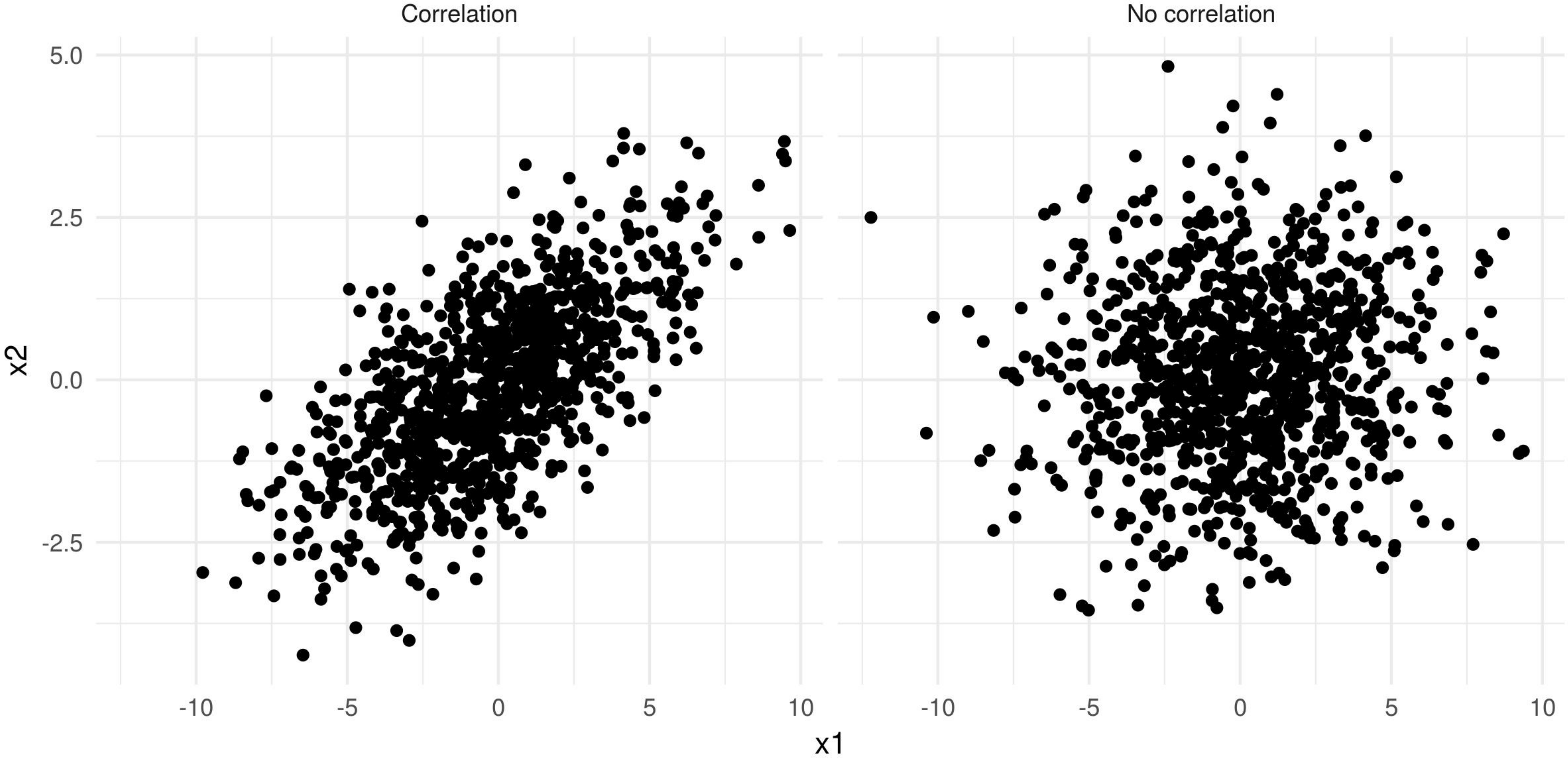
Summary of GLM with multiple predictors

```
Call:
glm(formula = Bus ~ CommuteDays + MilesOneWay, family = "binomial",
    data = bus)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0732  -0.9035  -0.7816   1.3968   2.5066

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.707515   0.119719  -5.910 3.42e-09 ***
CommuteDays  0.066084   0.023181   2.851  0.00436 **
MilesOneWay -0.059571   0.003218 -18.512  < 2e-16 ***
#...
```

Correlation between predictors



Order of coefficients

No correlation between predictors

- Order not important
- $y \sim x_1 + x_2 + \epsilon \approx y \sim x_2 + x_1 + \epsilon$

Correlation between predictors

- Order may changes estimates
- $y \sim x_1 + x_2 + \epsilon \neq y \sim x_2 + x_1 + \epsilon$

Let's practice!

GENERALIZED LINEAR MODELS IN R

Formulas in R

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

Why care about formulas for multiple logistic regression?

- Formulas backbone of regression
- Tricky to figure out
- Understanding `model.matrix()` key

Slopes

- Estimates coefficient for continuous variable
  - e.g., `height = c(72.3, 21.1, 3.7, 1.0)`
- Formula also requires a global intercept
- Multiple slopes: Slope for each predictor

Intercepts

- Discrete groups used to predict
- factor or character in R: `fish = c("red", "blue")`
- Single intercept has two options:
  - Reference intercept + contrast: `y ~ x`
  - Intercept for each group: `y ~ x -1`

Multiple intercepts

- Estimates effect of each group compared to reference group
- The first group, alphabetically, in the factor
- Default has one reference group per variable
  - `y ~ x1 + x2`
- Can specify one group to estimate an intercept for all groups
  - `y ~ x1+ x2 - 1`
- First variable has intercept estimated for each group

Dummy variables

- Codes group membership
- Used under the hood (i.e., `model.matrix()`)
- 0s and 1s for each group
- Example input: `color = c("red", "blue")`
- Dummy variables for `y ~ colors` :
  - `intercept = c(1, 1)`
  - `blue = c(0, 1)`
- Dummy variables for `y ~ colors - 1` :
  - `red = c(1, 0)`
  - `blue = c(0, 1)`

model.matrix()

- `model.matrix()` does legwork for us
- Foundation for formulas in R

```
model.matrix( ~ colors)
```

```
  (Intercept) colorsred
1           1         1
2           1         0
```

```
attr(,"assign")
```

```
[1] 0 1
```

```
attr(,"contrasts")
attr(,"contrasts")$colors
```

```
"contr.treatment"
```

- Order determined by factor order
- Change order change with Tidyverse or `factor()`

Factor vs numeric caveat

- R thinks variable is numeric
  - e.g., `month = c(1, 2, 3)`

```
month <- c( 1, 2, 3)
model.matrix( ~ month)
```

```
  (Intercept) month
1           1     1
2           1     2
3           1     3
```

```
attr(,"assign")
```

```
0 1
```

- Need to specify factor or character
  - e.g., `month = factor(c( 1, 2, 3))`

```
model.matrix( ~ month)
```

```
  (Intercept) month2 month3
1           1      0      0
2           1      1      0
3           1      0      1
```

```
attr(,"assign")
```

```
0 1 1
```

```
attr(,"contrasts")$month
```

```
"contr.treatment"
```

Let's practice!

GENERALIZED LINEAR MODELS IN R

Assumptions of multiple logistic regression

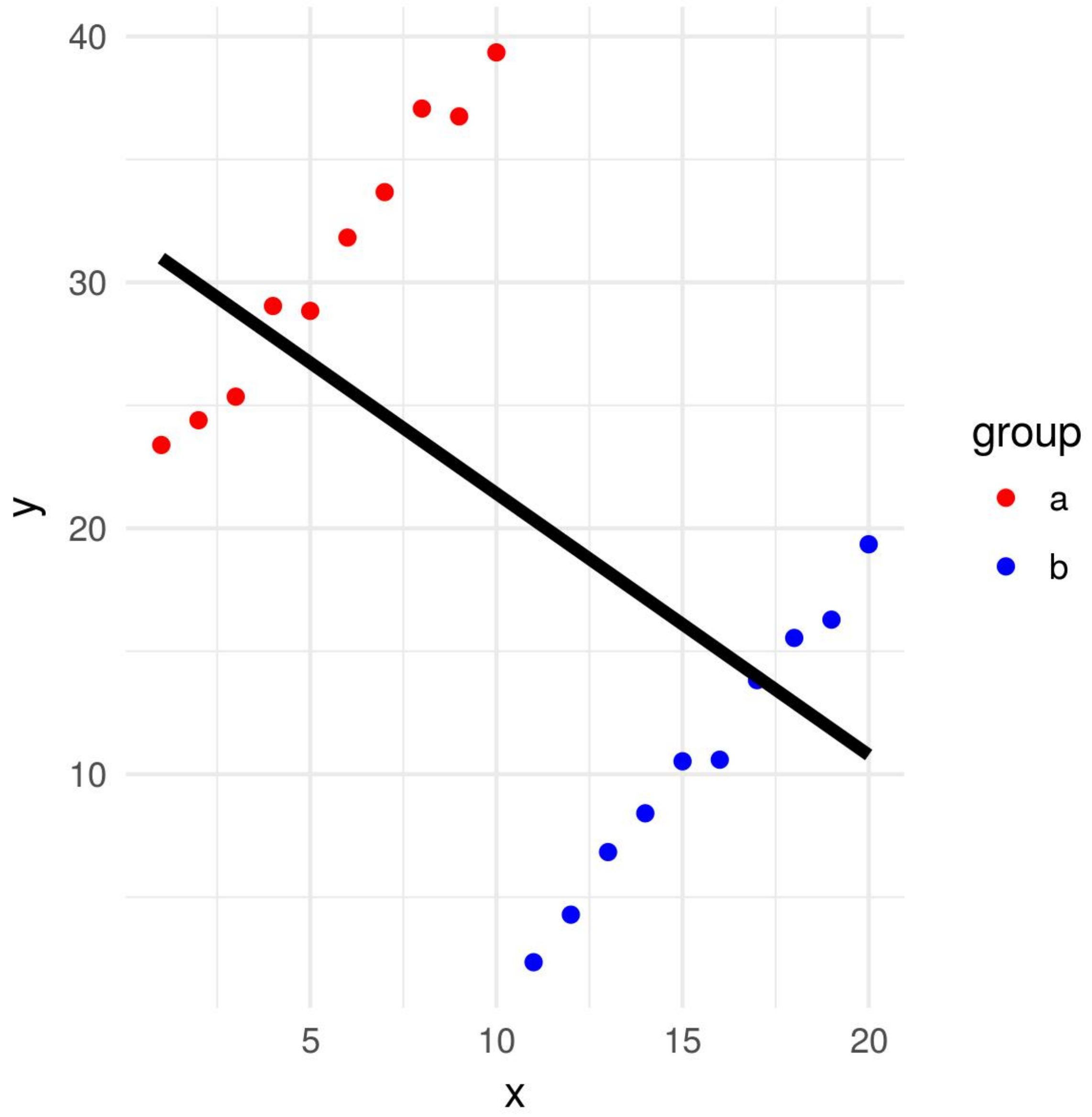
GENERALIZED LINEAR MODELS IN R

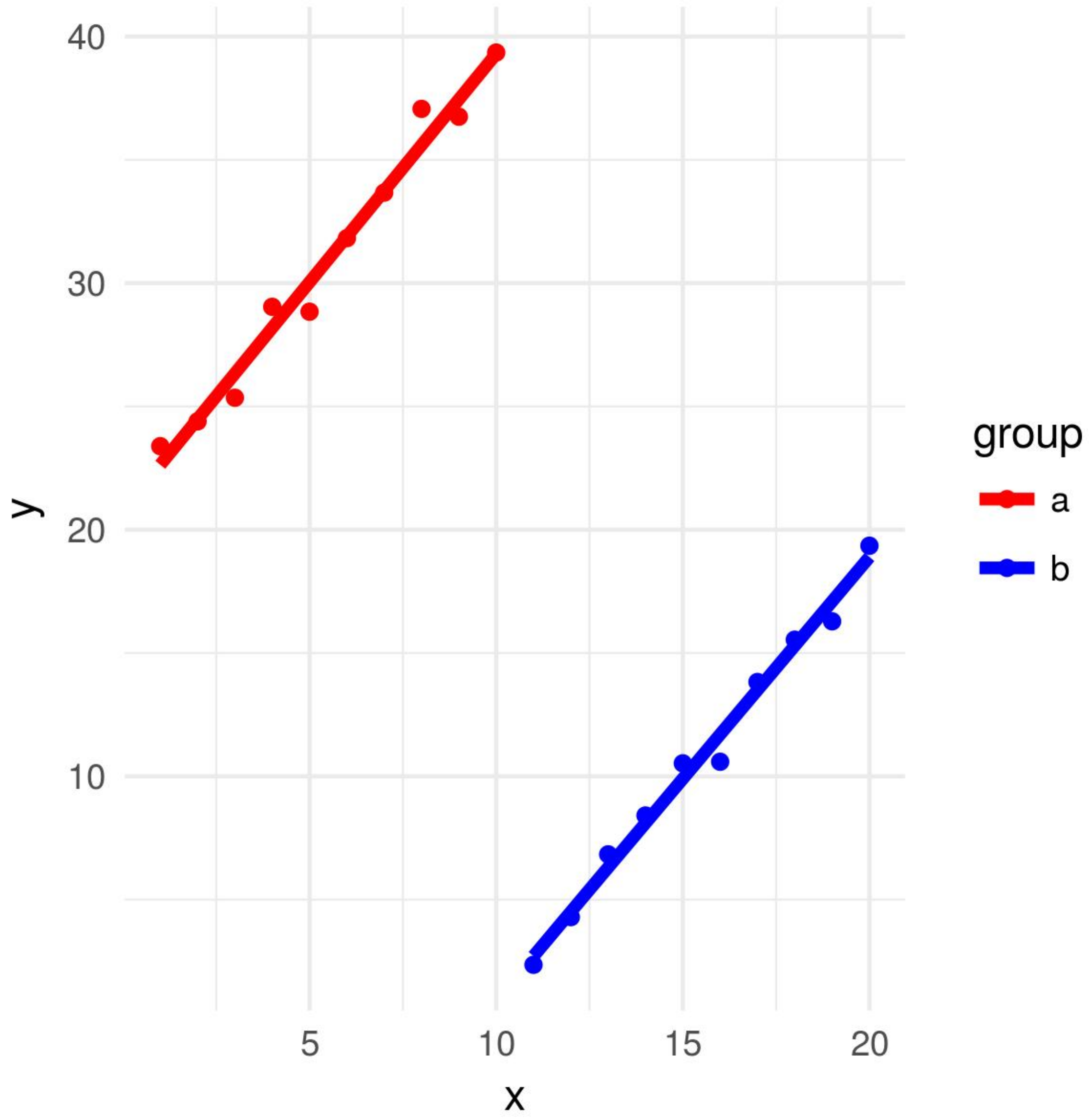
Richard Erickson
Instructor

Assumptions

- Limitations also apply to Poisson and other GLMs
- Important assumptions:
  - Simpson's paradox
  - Linear, monotonic
  - Independence
  - Overdispersion

Example Simpson's paradox





Simpson's paradox

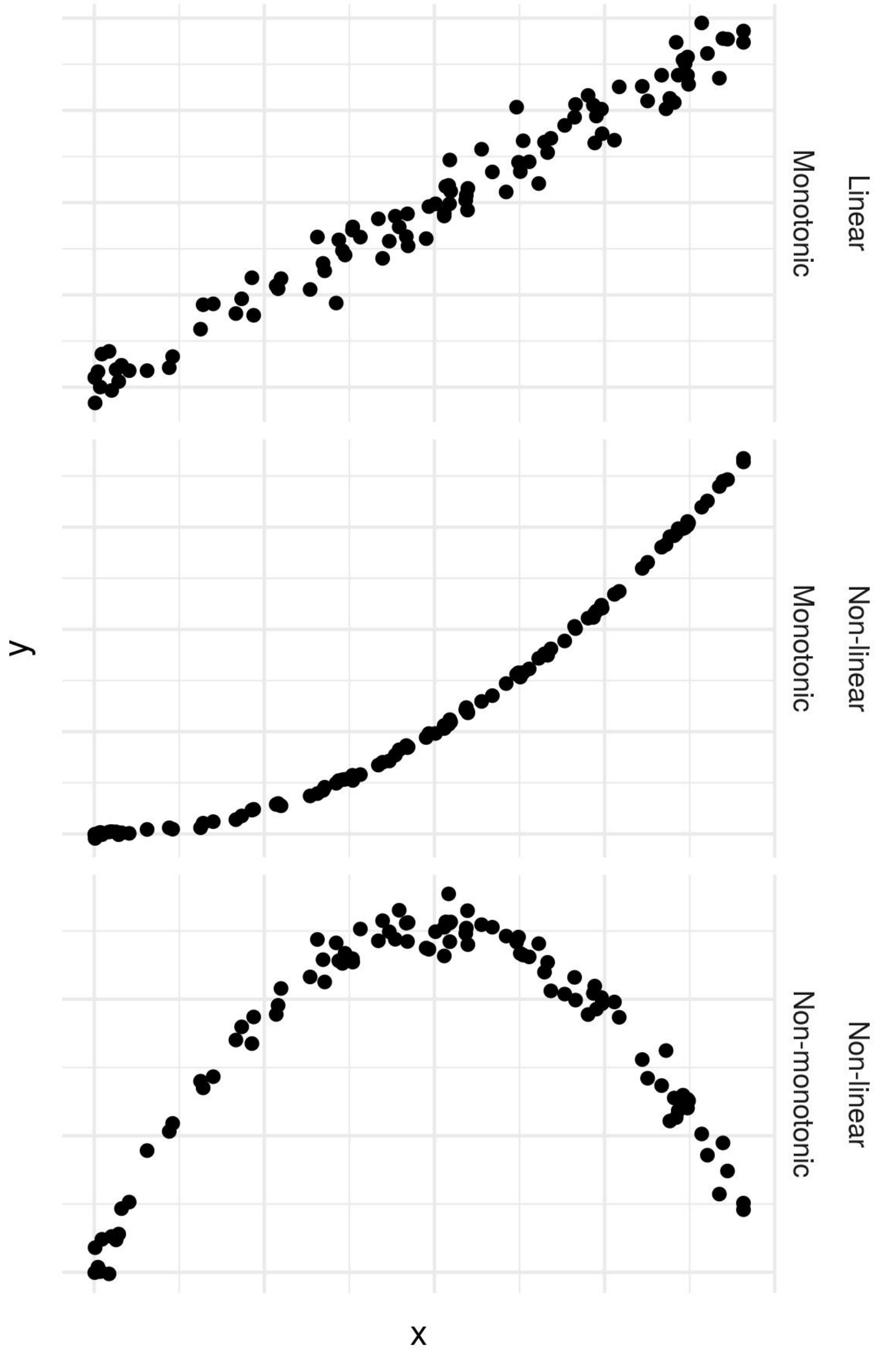
Key points

- Missing important predictor
- Inclusion changes outcome
- Easy to visualize with `lm()`

Simpson's paradox and admission data

Admissions data

- University of California Berkeley
- Graduate admission
- Rate of admission by department and gender
- Does bias exist?


Linear
Monotonic
Non-linear
Monotonic
Non-linear
Non-monotonic
y
x

Independence

Predictors

- If all independent, order has no effect on estimates
- If non-independent, order can change estimates

Response

- What is unit of focus?
- Individual, groups, group of groups?
- Test scores
  - Individual student?
  - Teacher? School? District?

Overdispersion

- Too many zeros or one (Binomial)
- Too many zeros, too large variance (Poisson)
- Variance changes
- Beyond scope of this course

Let's practice!

GENERALIZED LINEAR MODELS IN R

Conclusion

GENERALIZED LINEAR MODELS IN R

Richard Erickson
Instructor

What you've learned

- How GLM extends LM:
  - Poisson Error term
  - Binomial Error term
- Understanding and plotting results
- GLM with multiple regression

Where to from here?

- DataCamp **Multiple (linear) regression course in R** (if you missed it)
- Extending to include random effects with **Hierarchical and mixed-effect models in R**
- Fit **generalized additive models in R** (GAMs) to non-linear models
- Decide what coefficients to use with model selection such as AIC
- Many other types of regression
- Searching and R packages documentation to learn more

Happy coding!

GENERALIZED LINEAR MODELS IN R