GENERAL ASSEMBLY

GENERALIZED LINEAR MODELS

Matt Brems

Data Science Immersive

LINEAR AND LOGISTIC REGRESSION

- So far, we've learned how to fit linear and logistic regression models.

- Linear and logistic regressions are both types of **linear models**.
  - What do we mean when we say **linear model**?
    - We want to model some outcome by a series of predictors $X_1, \dots, X_p$, where a one-unit change in an independent variable has a fixed effect on the outcome.
    - You might also hear the phrase **linear in the coefficient**.

LINEAR AND LOGISTIC REGRESSION

- If I want to predict some continuous Y value between $(-\infty, \infty)$, I'll probably use a linear regression model.
- If I want to predict some continuous Y value between $(0,1)$, I'll probably use a logistic regression model.
  - Remember that logistic regression models predict probabilities between 0 and 1 and that we can use these probabilities to classify observations into one of two classes.

- What are examples of things we might like to predict that don't fall in either a $(-\infty, \infty)$ or $(0,1)$ range?

GOING BEYOND LINEAR AND LOGISTIC REGRESSION

- What are examples of things we might like to predict that don't fall in either a $(-\infty, \infty)$ or $(0,1)$ range?

GENERALIZED LINEAR MODELS

- This is where **generalized linear models** comes into play.
    - We can create models that take the benefits of being linear in the coefficient and "bend" it to match the range of the Y we want to model.
    - In order to make this happen, we need three things:
        - A linear piece.
        - A "bending" piece.
        - A random piece.

GENERALIZED LINEAR MODELS

- Generalized linear models have the following form:

$$Y = g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) + \varepsilon$$

GENERALIZED LINEAR MODELS

- Generalized linear models have the following form:

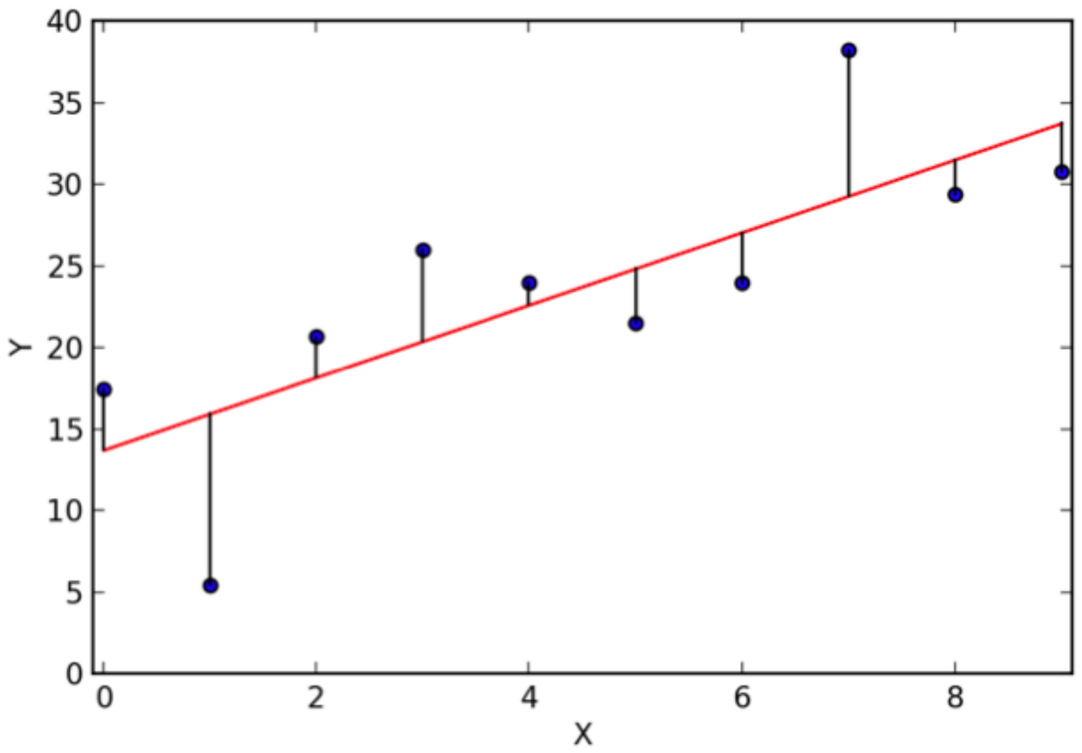
$$Y = g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) + \varepsilon$$

- $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ is our **linear component**.
  - We want a one-unit change in X to change Y in a certain way.
- $g(\cdot)$ is our **link component**.
  - We need to "bend" our linear component to match the range of Y.
- ε is our **random component**.
  - We need our errors to also match our Y.

GENERALIZED LINEAR MODELS

$$Y = \ g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) \ + \varepsilon$$

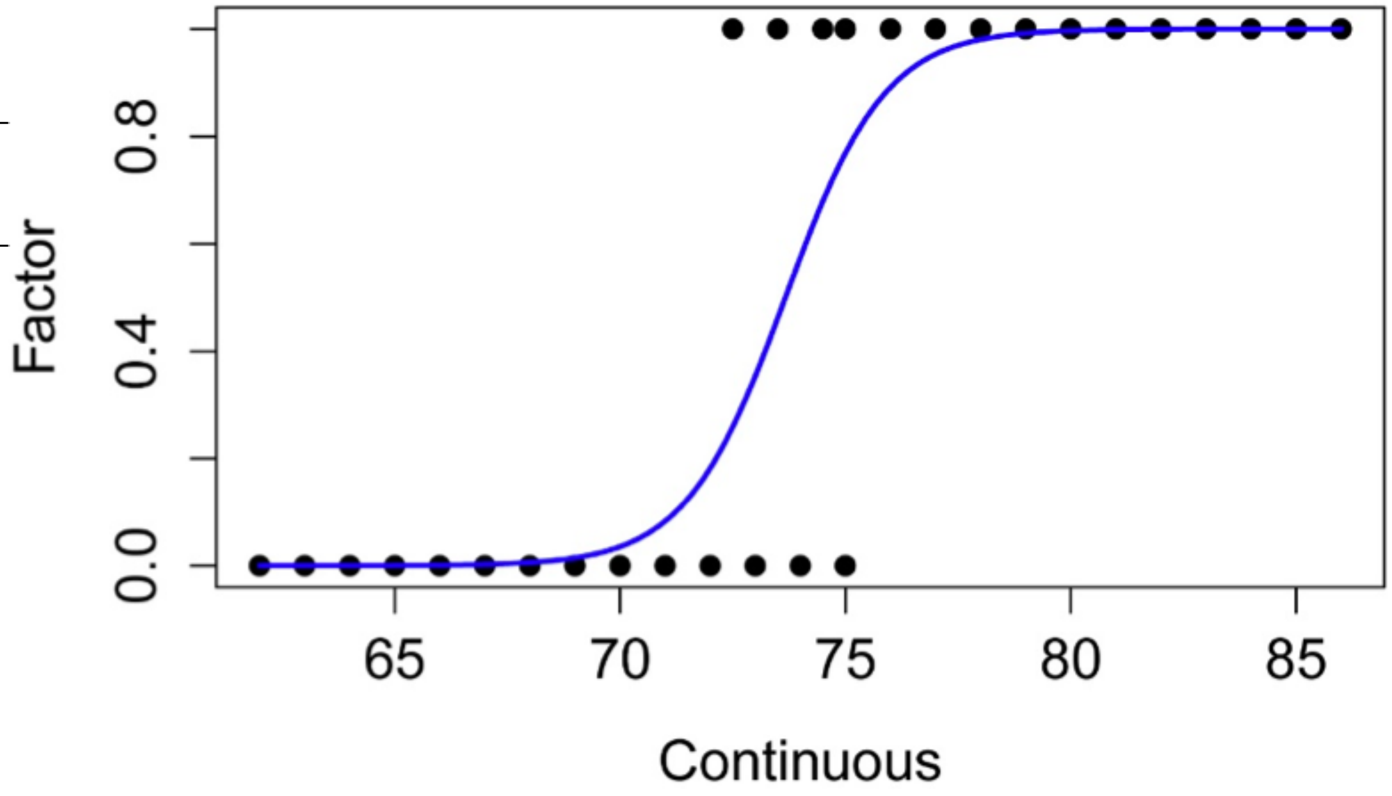
- **Linear regression:**
  - Linear Component:
  - Link Component:
  - Random Component:



GENERALIZED LINEAR MODELS

$$Y = \ g(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \ + \varepsilon$$

- **Logistic regression:**
  - Linear Component:
  - Link Component:
  - Random Component:



GENERALIZED LINEAR MODELS

- Generalized linear models have the following form:

$$Y = g\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) + \varepsilon$$

- What does the link function do?
  - It transforms the linear component to match the range of Y.
  - It also affects how we interpret our coefficients.

- What does the random component do?
  - It relates our predicted values $\hat{y}$ to our observed values y.

GENERALIZED LINEAR MODELS

- If your Y values are $\{0,1,2,\ldots\}$:
  - Poisson Regression (if $Var[X] \approx E[X]$)
  - Negative Binomial Regression (if $Var[X] \gg E[X]$)

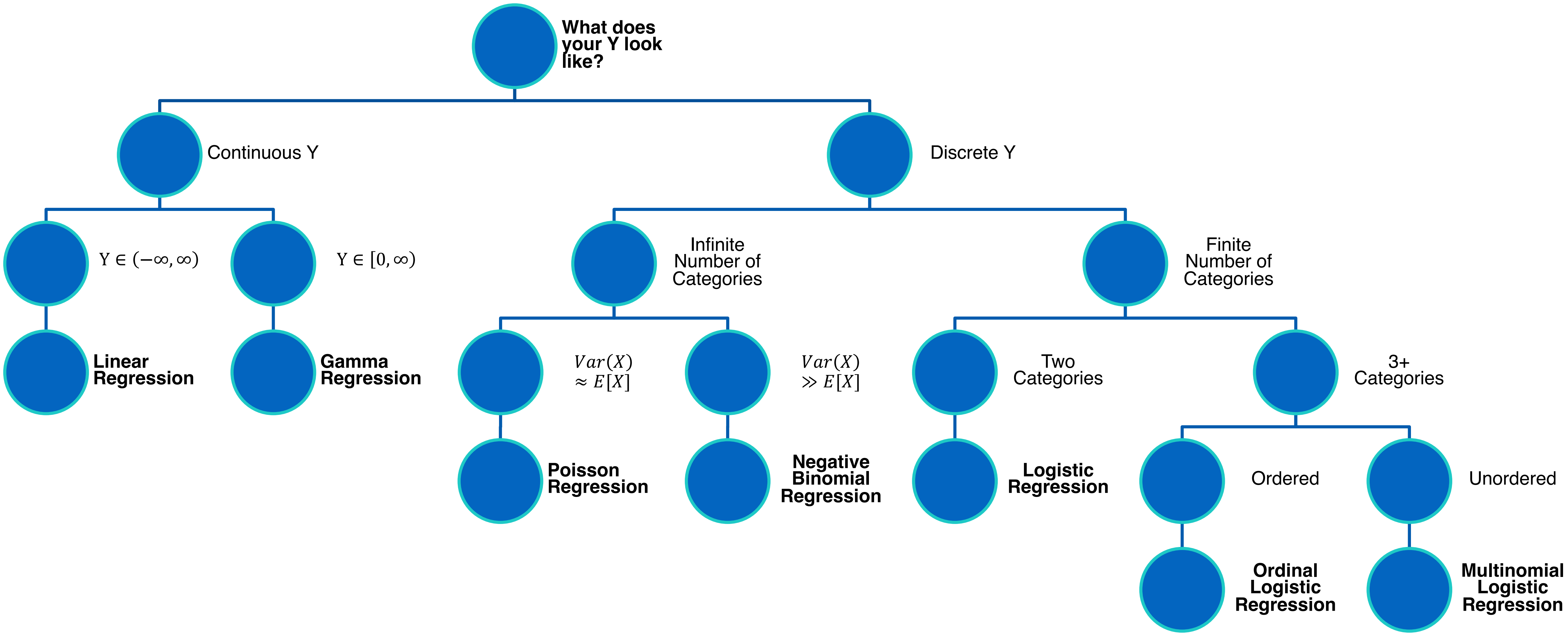
GENERALIZED LINEAR MODELS

- If your Y values are $\{0,1,2 \ldots, k\}$:
  - Multinomial Logistic Regression (if categories are unordered)
  - Ordinal Logistic Regression (if categories are ordered)

GENERALIZED LINEAR MODELS

- If your Y values are $[0, \infty)$:
    - Gamma Regression

GENERALIZED LINEAR MODELS CHEAT SHEET



ITERATIVELY REWEIGHTED LEAST SQUARES

- When we fit a linear regression model using ordinary least-squares, the values of $\widehat{\boldsymbol{\beta}}$ are directly calculated.

- However, there exists no closed-form solution for $\widehat{\boldsymbol{\beta}}$ with a generalized linear model.
    - Instead, we must calculate a first guess $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$, update our weights, calculate $\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$, and iterate until $|\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{i+1}}|$ is sufficiently small.

WHY USE GLMS?

- Generalized linear models are a series of models that:
  - Allow us to easily infer relationships between X and Y.
  - Have convenient parametric distributions.
  - Are flexible enough to handle a variety of scenarios.
    - Hierarchical models.
    - "Zero-inflated" or "zero-truncated" models.