Review of Bayesian Regression

EES 5891-03

Bayesian Statistical Methods

Jonathan Gilligan

Class #22: Tuesday, November 15 2022

Linear Regression

Linear Regression

- Start with simple regression:
  - One predictor variable X
  - Predictor and outcome are both continuous (real numbers)

$$\begin{align} Y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \alpha + \beta X \end{align}$$

- Multiple linear regression
  - N predictor variables $X_1, \ldots, X_N$
  - Everything continuous

$$\begin{align} Y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \alpha + \sum_{i=1}^N \beta_i X_i \end{align}$$

- Polynomial regression
  - Just treat each power of X like a new predictor
  - You can have polynomials with more than one variable

$$\begin{align} Y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \alpha + \sum_{j=1}^N \beta_i X^j \end{align}$$

Scaling Variables

- Standardizing:
  - $X_{\text{std}} = (X = \bar X) / \sigma_X$
  - All variables on the same scale
  - Centered with 0 at the mean
  - Slopes (β) measure the effect of changing by 1 standard deviation.
- Log scaling
  - Good for outcome variables that **must** be positive
  - Good for predictor variables with a tail that covers a large range
    - Population is often log-scaled
  - Variable must be >0, so if some values are 0, add a small number to them (e.g., 0.01, 0.001).

Integer Models

- Boolean (Yes/No)
  - Coin toss: Heads or tails
  - Bernoulli distribution
  - Special case of binomial, with one trial: `dbinom(1, p)`
- Count data:
  - Binomial distribution
    - `rbinom(N, p)`
    - Maximum value = N
  - Poisson distribution
    - `dpois(lambda)`
    - No maximum value
    - Poisson(λ) is the limiting case of Binomial(N, p) when N is large and p is small, with $Np = \lambda$.

- Generalized Linear Models (GLMs):
  - Link functions:

    $$\begin{aligned} Y &\sim \text{Binomial}(N, p) \\ p &= \text{logit}^{-1}(\alpha + \beta X) \end{aligned}$$

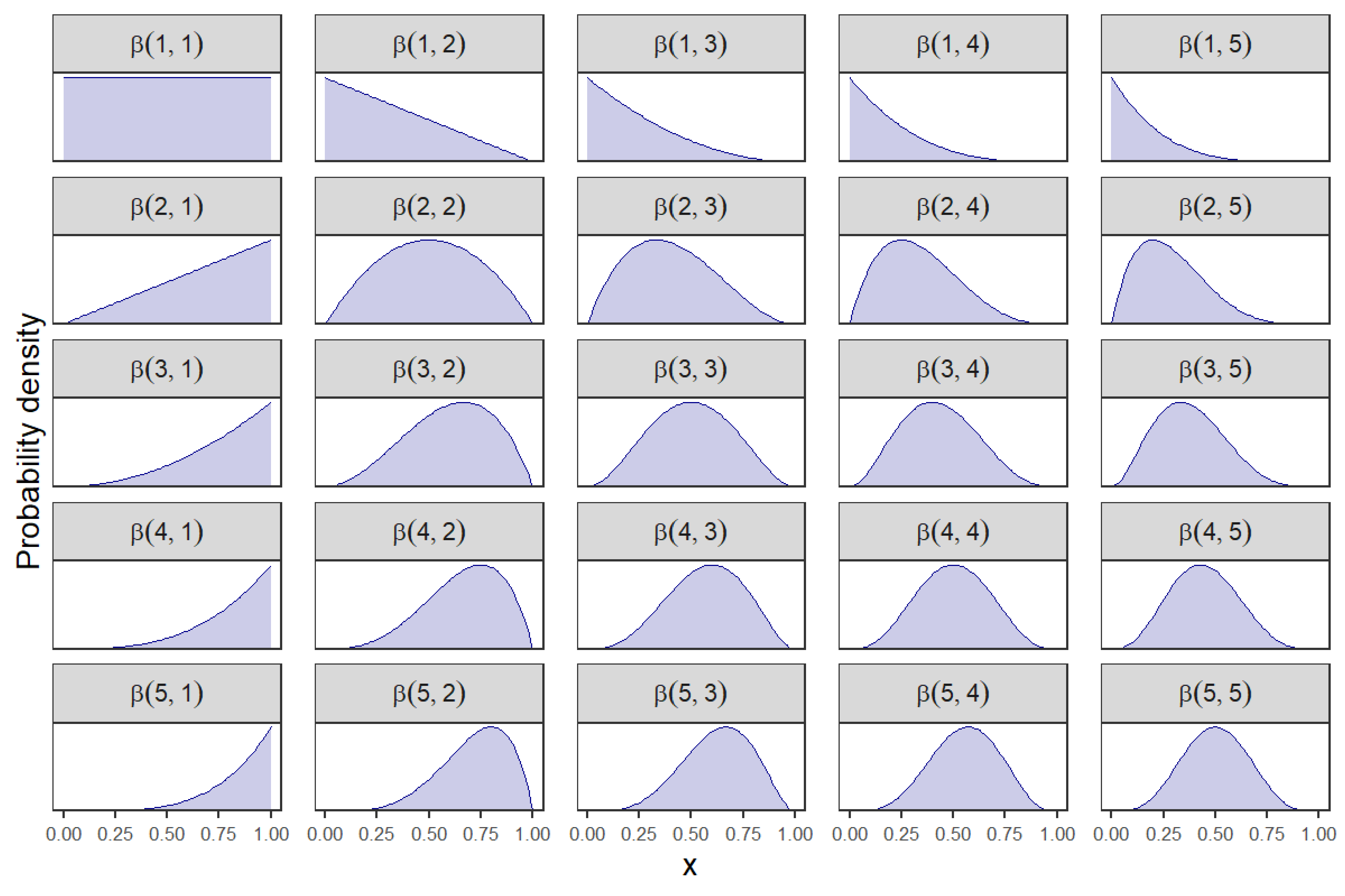
    ```
    y ~ dbinom(N, lambda),
    logit(p) = a + b * x
    ```

    $$\begin{aligned} Y &\sim \text{Poisson}(\lambda) \\ \lambda &= \exp(\alpha + \beta X) \end{aligned}$$

    ```
    y ~ dpois(lambda),
    log(lambda) = a + b * x
    ```

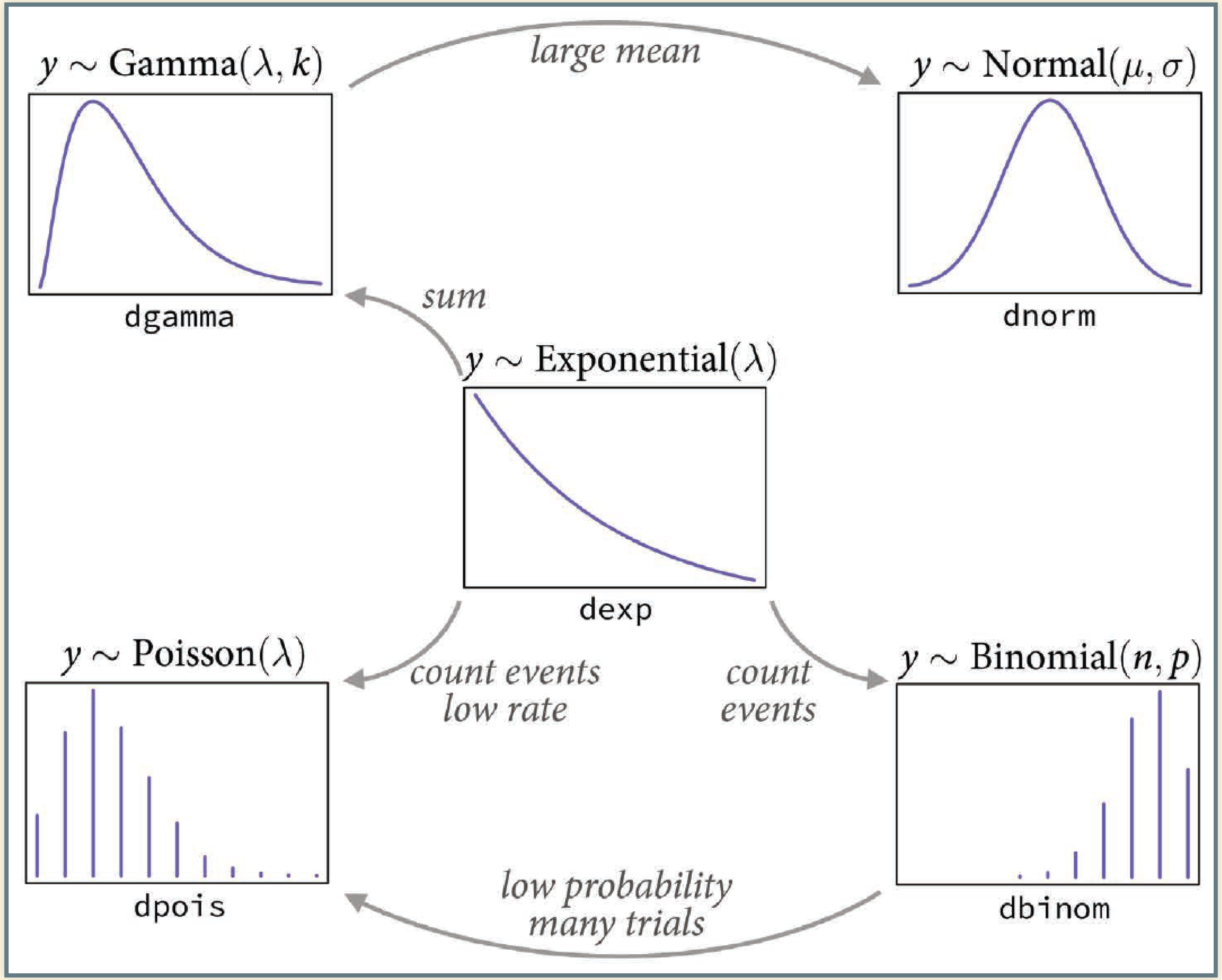
  - Priors for probabilities (range 0 to 1)
    - Beta functions: `dbeta(a, b)`
      - Mean = $a / (a + b)$

Beta Distributions



- Mean = \(a / (a + b)\)
- Variance = \(ab / ((a+b)^2 (a + b + 1))\)

Exponential Family of Distributions



Categorical Variables

- Categorical variables
  - Predictors (section 5.3):
    - Use index variables (0 or 1)
    - N levels: $N-1$ index-variables (default category)
      - Only one index is 1 (the one for the category)
      - If they're all zero, it's the default category
  - Outcomes (section 11.3):
    - Integer 0…*N*
    - Use multinomial (categorical) likelihood (`categorical()`)
    - Use Dirichlet priors (`dirichlet()`)
      - Dirichlet is like an N-dimensional generalization of the beta distribution
- Ordered categorical variables
  - Effects are cumulative
  - Predictors (section 12.4):
    - Use an integer variable (K) with values $0 \ldots N-1$
    - $N - 1$ variables δ_i ($i \in [1, \ldots, N-1]$)
      - $\mu = \alpha + \sum_{i=1}^{K} \delta_i$
    - Initially, no effect.
    - Every time you step to the next level, there's another effect, and they add up.
  - Outcomes (section 12.3):
    - Integer 0…*N*
    - Use multiple logistic regression
      - Initially at lowest value
      - As effect grows, step through values in sequence.

Multilevel Models

Multilevel Models

- Data is grouped into *clusters*
  - Geographical groupings (states, counties, etc.)
  - Temporal groupings (seasons)
  - Other categories:
    - Gender
    - Education
    - Profession
    - Species
    - Individual
    - ...
- Hyperpriors and hyperparameters:
  - Each group may have its own prior for slope, intercept, etc.
    - The parameters for that prior are drawn from a *hyperprior*

- Single-level model

  \[\small \begin{align} Y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \alpha + \beta X \\ \alpha &\sim \text{Normal}(0, 1) \\ \beta &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Exponential}(1) \\ \end{align} \]

- Two-level model (varying intercept)

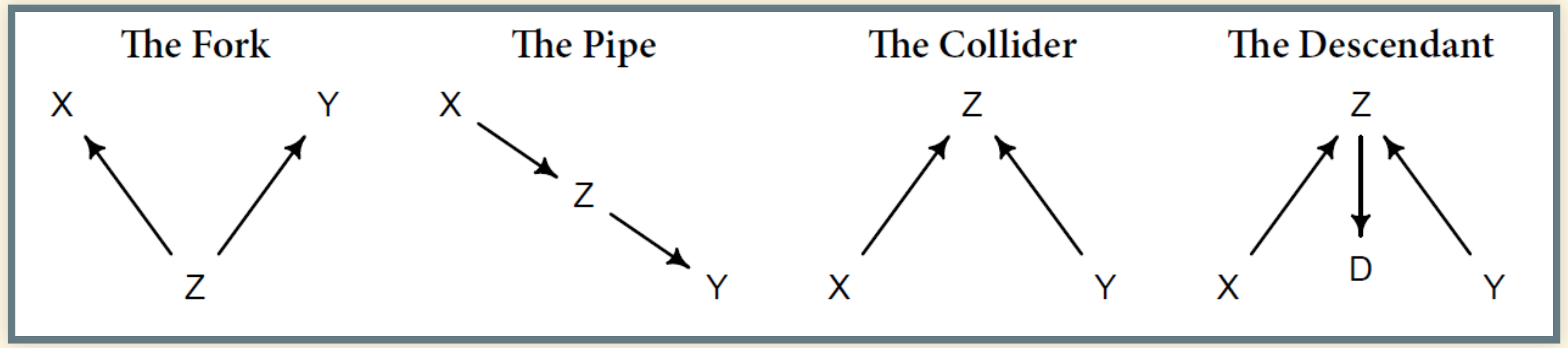
  \[\small \begin{align} Y &\sim \text{Normal}(\mu, \sigma) \\ \mu &= \alpha + \beta X \\ \alpha &\sim \text{Normal}(\bar \alpha, \sigma_\alpha) \\ \beta &\sim \text{Normal}(0, 1) \\ \bar \alpha &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Exponential}(1) \\ \sigma_\alpha &\sim \text{Exponential}(1) \\ \end{align} \]

Multilevel Models

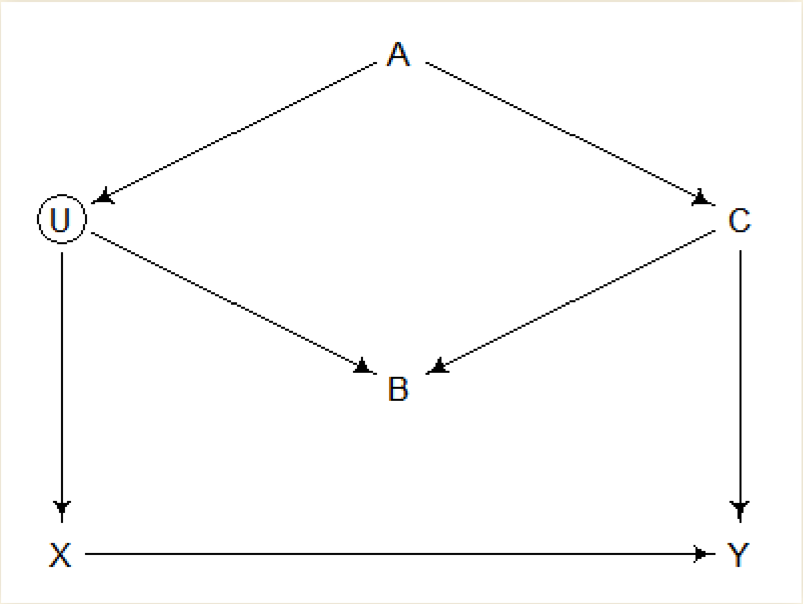
Designing Models

Designing Models

- Analyze relationships between variables
  - DAGs describe *causal relationships* among variables.
  - They can help us detect potential problems:
    - Spurious associations
    - Masked relationships
    - Multicollinearity
    - Post-treatment bias
    - Confounders
- Four fundamental types of confounding relationships:



- General rules:
  1. List all paths connecting *X* (potential cause) to *Y* (outcome)
  2. Classify each path as *open* or *closed*
     - A path is *open* unless it contains a *collider*
  3. Classify each path by whether it's a *backdoor path*.
     - A *backdoor path* has an arrow pointing at *X*
  4. If there are any *open backdoor paths*, try to close it by *conditioning* on a variable.
- Example (Section 6.4.2)



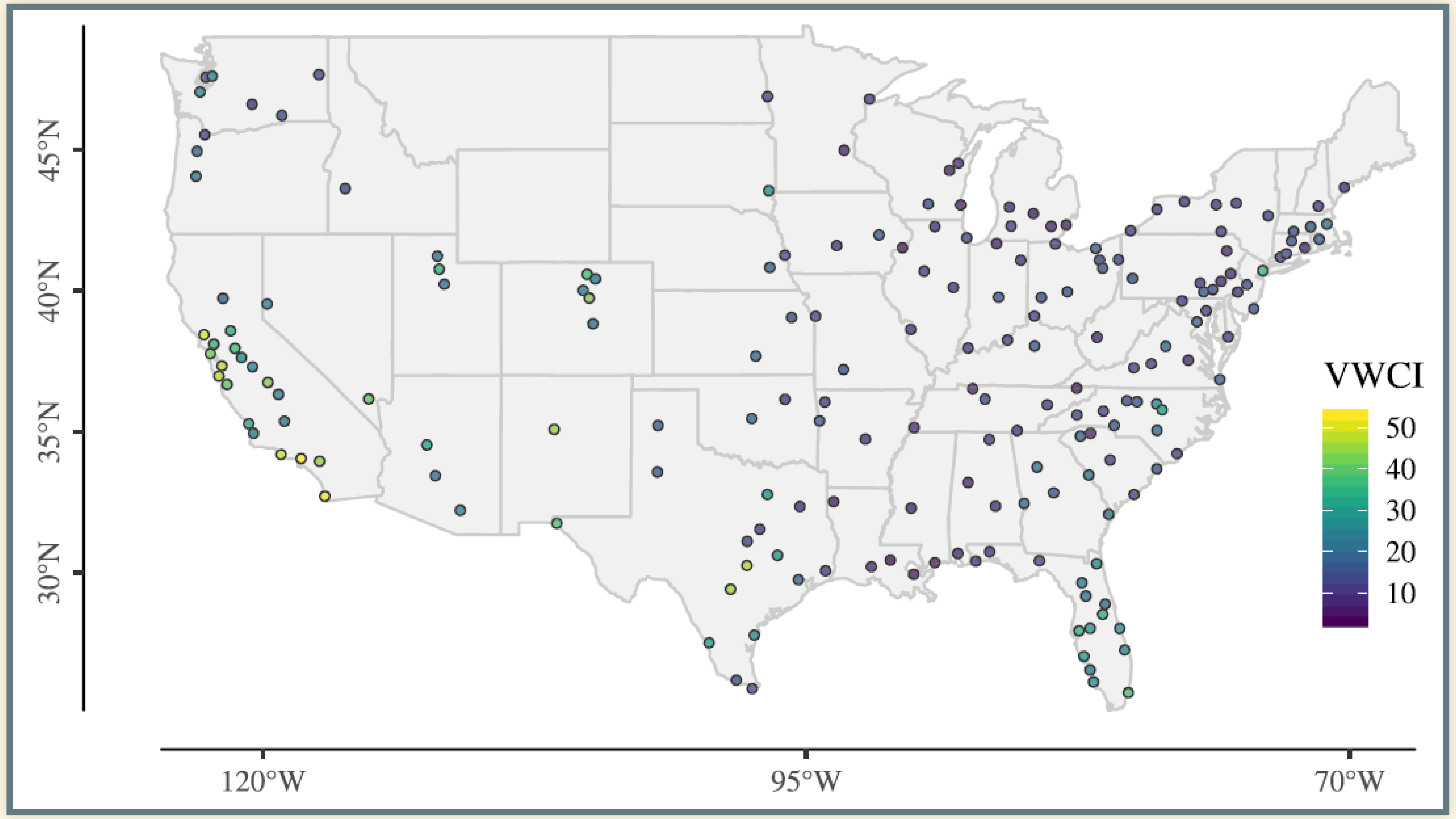
Testing DAGs

- How can you tell whether your DAG is correct?
  - You can't
  - There could always be important unobserved variables you don't know about
- Comparing DAGs with evidence
  - Analyze DAGs for *conditional independencies*
    - *Conditional independencies* are empirically testable statements
    - Use your model's posterior predictions to test predicted *conditional independencies*.
    - If DAG's *conditional independencies* are not observed, then it's probably not correct.
    - `dagitty`'s `impliedConditionalIndependencies()` function is your friend.
- DAG analysis is helpful, but it's not enough.
  - DAGs tell you about the *logical structure* of your model, but they don't tell you about the science.
  - What you know as a scientist is even more important:
    - *Why* do you think *X* influences *Y*?
    - *What* other variables might also play a role?

Example: Urban Water Conservation Policies

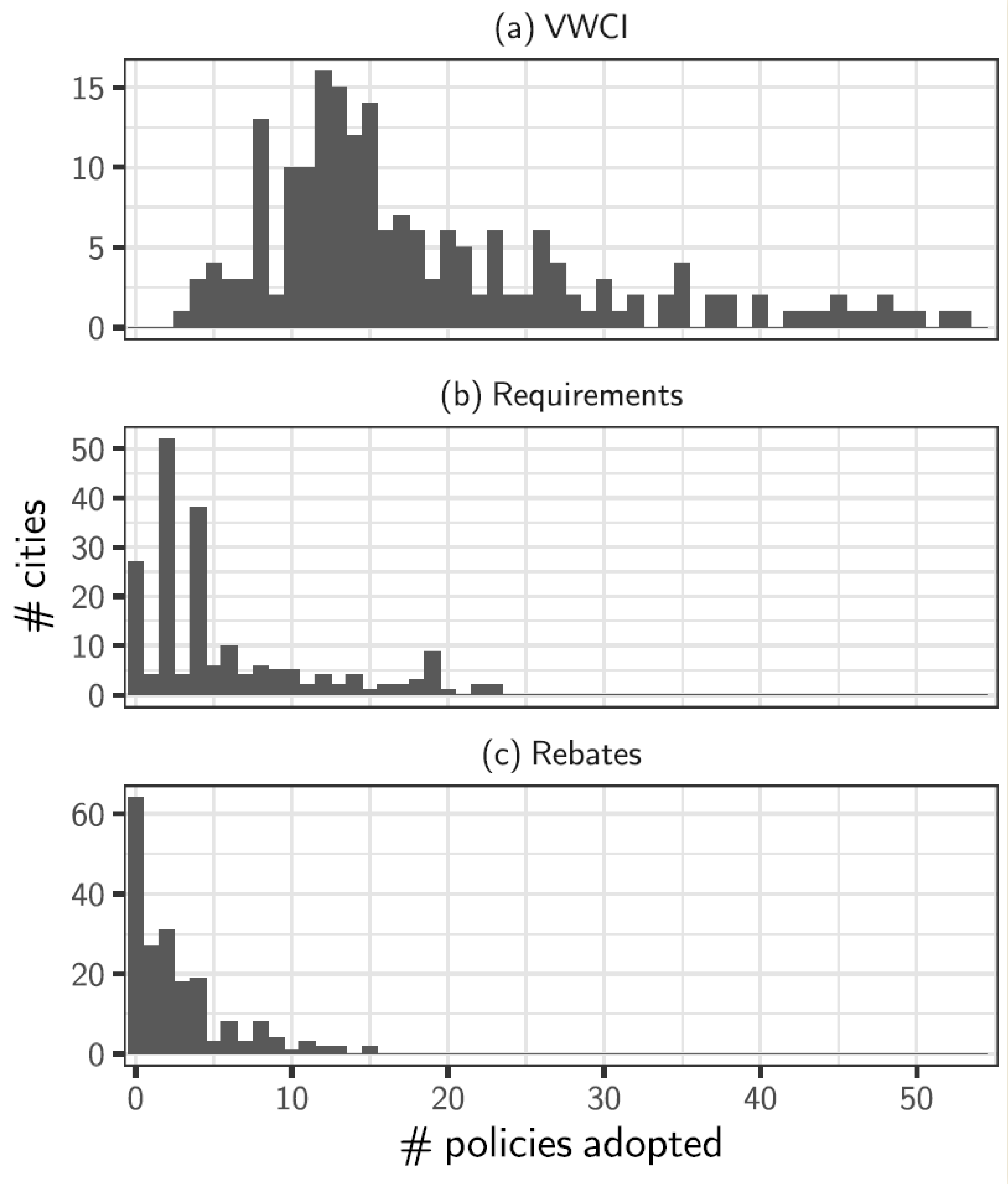
Urban Water Conservation Policies

- Build a database of water conservation policies for 197 largest cities in the US
  - Vanderbilt Water Conservation Index (VWCI)
    - List of 79 possible policies
      - 31 are *requirements*
      - 21 are *rebates* for voluntary actions
    - Each city gets a score based on how many policies it has adopted



Descxriptive Statistics

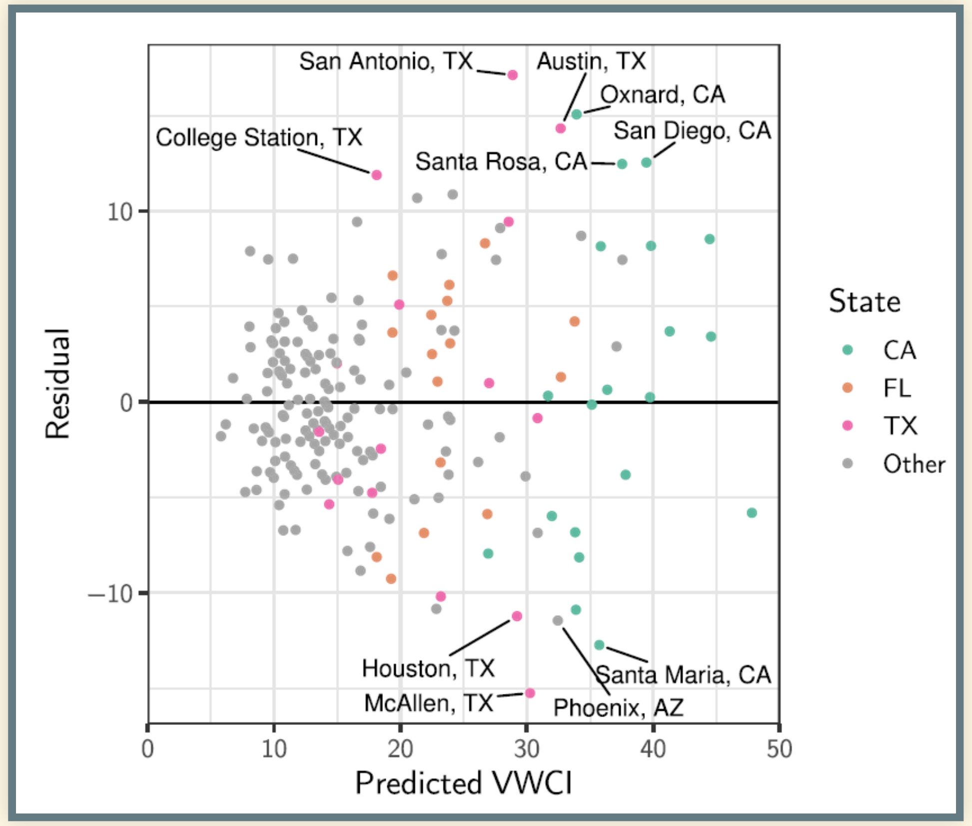
- Range of values: 3–53 (max. possible is 79)
- Mean: 18.7, median 15
- Model VWCI:
  - Predict score from
    - Temperature
    - Rainfall
    - Fraction of water supply from surface water
    - Population
    - Population growth
    - Personal income
    - Partisan voting index
  - Drop Honolulu HI and Anchorage AK and focus on 48 contiguous states
  - Temperature and Precipitation are collinear so use Köppen Aridity Index, which combines the two



Multilevel Model

- Multilevel variable-intercept model:
  - Predict each city's score from city-level data and state-level data
    - Priors for intercepts are based on state-level data

$$\begin{aligned} V_i &\sim \text{Binomial}(N_{\text{Actions}}, p_i) \\ \text{logit}(p_i) &= \alpha_j + \sum_{k \in \text{city variables}} \beta_k x_{ik} \\ \alpha_j &\sim \text{Normal}(\mu_j, \sigma_\alpha) \\ \mu_j &= \alpha_0 + \sum_{k \in \text{state variables}} \gamma_k w_{jk} \end{aligned}$$

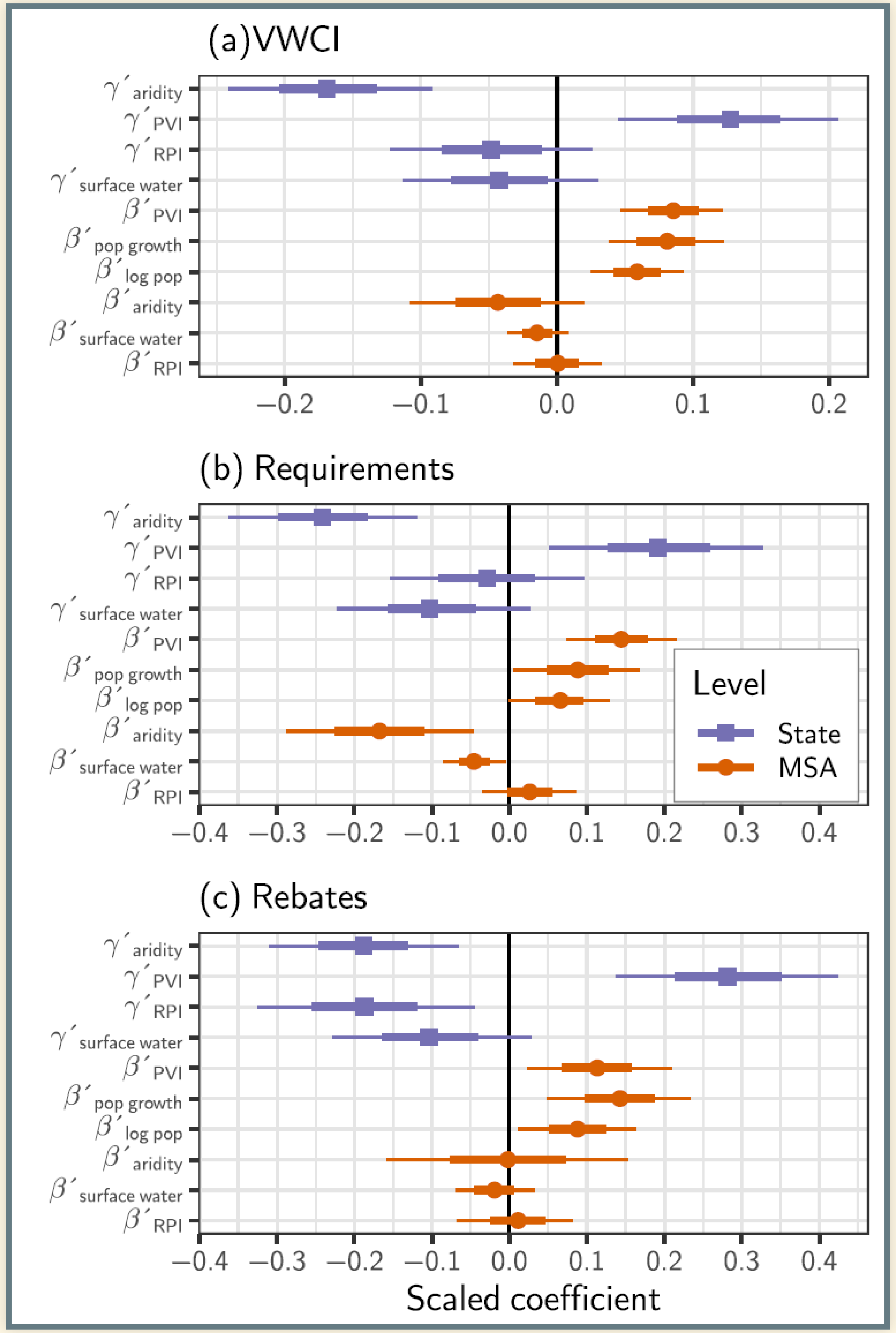


- Original residuals showed data were *overdispersed* (variance was too great for a Binomial), so we changed the model to use a *beta-binomial* distribution (see Chapter 12).

$$\begin{aligned} V_i &\sim \text{beta-Binomial}(N_{\text{Actions}}, \phi p_i, \phi (1 - p_i) \end{aligned}$$

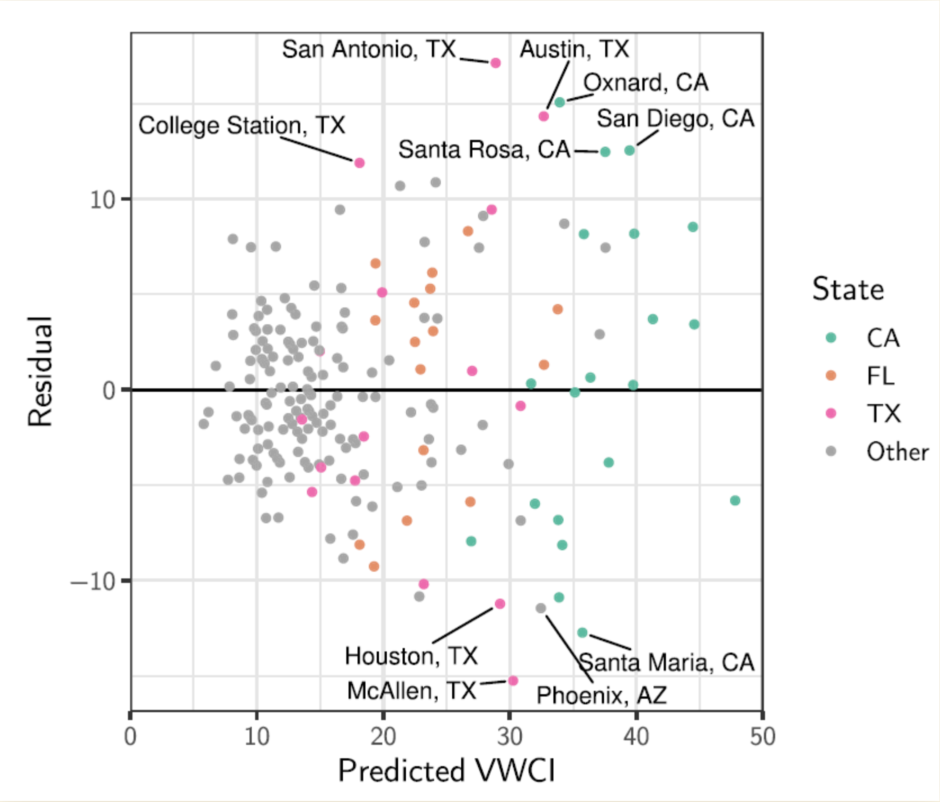
Results

- Most important state-level predictors:
  - Aridity
  - Partisan voting index (PVI)
- Most important city-level predictors:
  - PVI
  - Population growth
  - Log population
- City-level Aridity, surface water, and personal income don't matter after accounting for state-level effects



Checking model

- We did detailed interviews with water managers in San Antonio and Phoenix
  - San Antonio has a higher score than predictied.
    - Its Republican political leaning suggests low water conservation,
    - but the city doesn’t have so much choice:
      - A lawsuit over endangered species led to federal requirements to conserve water
  - Phoelix has a lower score than predicted.
    - Central Arizona Project brings water from Colorado River
    - Reduces water stress on Phoenix



Cities With the 10 Largest Residuals From VWCI Regression

Rank	City	VWCI	predicted VWCI	residual
1	San Antonio, TX	46	28.9	17.1
2	McAllen, TX	15	30.2	−15.2
3	Oxnard, CA	49	33.9	15.1
4	Austin, TX	47	32.7	14.3
5	Santa Maria, CA	23	35.7	−12.7
6	San Diego, CA	52	39.5	12.5
7	Santa Rosa, CA	50	37.5	12.5
8	College Station, TX	30	18.1	11.9
9	Phoenix, AZ	21	32.4	−11.4
10	Houston, TX	18	29.2	−11.2