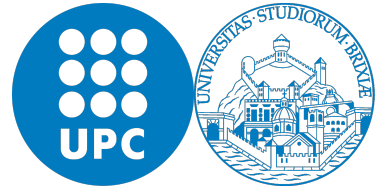
Module 3

Beyond Standard Approaches: Modeling Ordinal Data with CUB Models



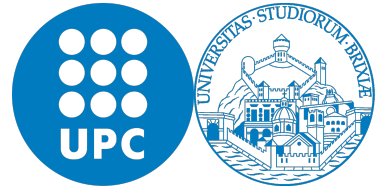
MESIO Summer School

Introduction to CUB Models

CUB Model with covariates

CUB Model with Shelter Option

Treatment of ”Don’t Know” (DK) Options



Limitations of Classical Models

- While powerful and widely used, ordered logit/probit models operate under specific assumptions:
  - **Proportional Odds Assumption:** This assumption can be restrictive and, if violated, can lead to misleading conclusions.
  - **Latent Variable Interpretation:** While mathematically convenient, the interpretation of the latent variable might not always align perfectly with the psychological process of how an individual actually chooses an ordered category. It doesn't explicitly account for aspects like uncertainty or indecision.

The Need

These limitations highlight a clear need for alternative statistical models designed to capture the unique characteristics of ordinal data and reflect underlying cognitive processes. This is where **CUB models** come into play.

MESIO SUMMER SCHOOL

CUB Models: An Innovative Approach

- CUB models, developed starting in the early 2000s, offer a distinct and innovative approach.
- They stand apart from traditional methods by explicitly incorporating **psychological interpretations** into their statistical structure.
- At their core, CUB models are **mixture models** specifically tailored for rating data.
- The fundamental idea: a respondent's observed rating is not solely a direct and deterministic mapping of their true "feeling" or "utility."
- Instead, it is also significantly affected by an element of **"uncertainty" or "indecision"** that can influence the final choice.

Key Insight

This acknowledgment of psychological complexity in the response process is what makes CUB models particularly insightful for rating data.

MESIO SUMMER SCHOOL

The Psychological Reasons Behind the CUB Model

- The innovation of the CUB model lies in its attempt to statistically represent **two fundamental psychological components** that influence a respondent's choice on an ordinal scale.
- The model posits that an observed rating R is a probabilistic outcome of a decision process that weighs these two components:
  1. **Feeling**
  2. **Uncertainty**

MESIO SUMMER SCHOOL

1. Feeling

- This component represents the **conscious, rational, and deliberative** aspect of the decision-making process.
- It reflects the respondent's **genuine evaluation, opinion, or perception** regarding the item being rated.
- The "feeling" directs the respondent towards a specific category on the scale that best aligns with their internal assessment.
- This is the component that captures the respondent's **true position** on the matter at hand.

1. Feeling: Interpretations

- Depending on the specific context of the rating, ”feeling” can be interpreted as:
  - **Agreement/Disagreement**: How much a person agrees or disagrees with a statement.
  - **Satisfaction/Dissatisfaction**: The level of contentment or discontent with a product, service, or experience.
  - **Liking/Disliking**: The degree of preference or aversion towards an item.
  - **Perceived quality, importance, risk, etc.**: The subjective assessment of various attributes.

The feeling component drives the respondent towards a **particular region** of the ordinal scale, reflecting their underlying preference or ”attraction” to certain categories.

2. Uncertainty (or Indecision/Fuzziness)

- This component captures the **hesitation, randomness, or lack of decisiveness** that can accompany the choice process.
- It acknowledges a crucial psychological reality: respondents may not always have a perfectly clear and precise mapping of their internal feeling onto the provided scale categories.
- This can introduce a degree of **randomness or "noise"** into the selection process.

2. Uncertainty: Sources

- Sources of this uncertainty can be varied and include:
  - **Lack of Information or Knowledge**: Insufficient information to form a strong opinion.
  - **Ambiguity**: Unclear question wording or scale category definitions.
  - **Personal Tendencies**: Some individuals are inherently more indecisive.
  - **Cognitive Effort/Satisficing**: Choosing a plausible but not necessarily precise option to save mental effort.
  - **Time Pressure or Fatigue**: Being rushed or tired can reduce decision precision.
  - **Emotional State or Mood**: A respondent's transient emotional state can introduce variability.

The uncertainty component effectively describes the probability that the respondent's choice is influenced by **random factors** rather than a specific preference.

The CUB Model: Statistical Formulation

- The basic CUB model, designed for m ordered categories (typically $r = 1, 2, \ldots, m$), is a finite mixture of two discrete probability distributions:
  - A **shifted Binomial distribution** to model the **feeling** component.
  - A **discrete Uniform distribution** to model the **uncertainty** component.
- The **Probability Mass Function (PMF)** for an observed rating $R = r$ is given by:

$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r) +$$

where

- m: The number of ordered categories on the scale (e.g., $m = 5$ for a 5-point Likert scale).
- r: The selected category by a respondent, where $r \in \{1, 2, \ldots, m\}$.

$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

$\pi \in (0, 1]$ acts as a mixture weight.

- π represents the probability that the observed choice is driven by the **feeling component** (shifted Binomial).
- Consequently, $(1 - \pi)$ is called the **Uncertainty parameter** and represents the probability that the observed choice is driven by the **uncertainty component** (discrete Uniform). A higher $(1 - \pi)$ means greater indecision or randomness.

$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

$\xi \in [0, 1]$ is directly related to the shifted Binomial distribution while $(1 - \xi)$ is the success parameter of the distribution and it is called **"Feeling parameter"**.

- If $(1 - \xi)$ is **high** (e.g., close to 1, meaning ξ is close to 0), there's a strong underlying feeling towards the **higher end of the scale**.
- If $(1 - \xi)$ is **low** (e.g., close to 0, meaning ξ is close to 1), there's a strong underlying feeling towards the **lower end of the scale**.
- If $(1 - \xi) = 0.5$ (meaning $\xi = 0.5$), the feeling component is **neutral or centered**, implying a symmetric preference.

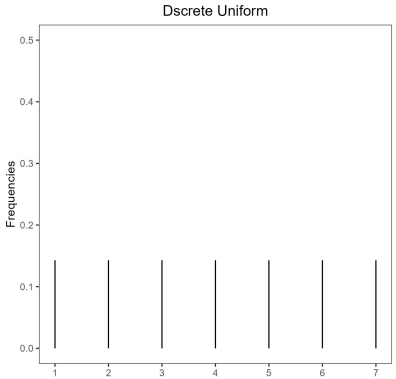
$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

$U(r \mid m)$ is the **Uniform Component**

- Probability of choosing category r from a discrete Uniform distribution:

$$U(r \mid m) = \frac{1}{m}$$

- Represents complete randomness or lack of specific preference.



$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

$B(r \mid \xi)$ is the probability of choosing category r according to a **shifted Binomial distribution**.

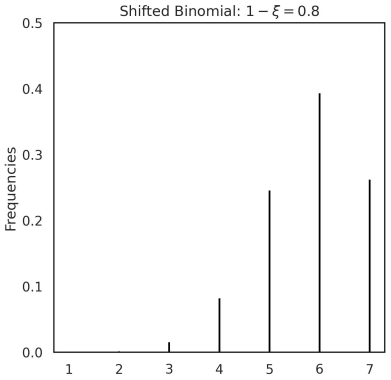
- Specifically, this refers to $P(X = r - 1)$ where $X \sim \text{Binomial}(m - 1, 1 - \xi)$.
- The "shifted" aspect arises because the rating scale typically starts from 1, while a standard Binomial distribution's trials start from 0. To model a choice of r on a scale $1, \ldots, m$, we consider $r - 1$ "successes" out of $m - 1$ "trials".
- So, the PMF is:

$$B(r \mid m - 1, 1 - \xi) = \binom{m-1}{r-1} (1 - \xi)^{r-1} \xi^{m-r}$$

- This component models the **feeling** towards a particular category, allowing for various shapes (unimodal, skewed left/right).

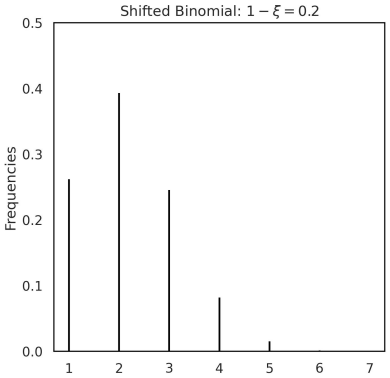
$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

If $(1 - \xi)$ is **high** (e.g., close to 1, meaning ξ is close to 0), there's a strong underlying feeling towards the **higher end of the scale**.



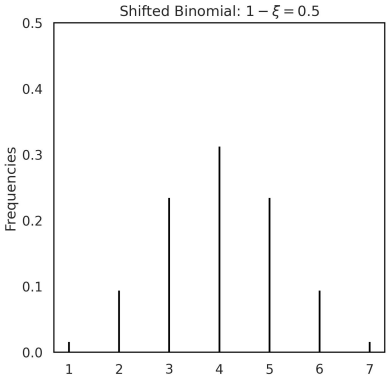
$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

If $(1 - \xi)$ is **low** (e.g., close to 0, meaning ξ is close to 1), there's a strong underlying feeling towards the **lower end of the scale**.



$$P(R = r \mid \pi, \xi) = \pi B(r \mid \xi) + (1 - \pi)U(r)$$

If $(1-\xi) = 0.5$ (meaning $\xi = 0.5$), the feeling component is **neutral or centered**, implying a symmetric preference.



The CUB Model: Final PMF

Final Probability Mass Function

The CUB model's PMF for an observed rating $R = r$ is:

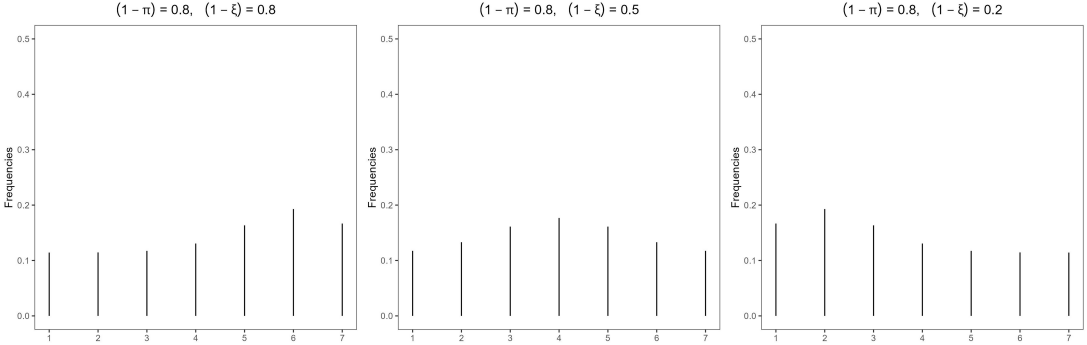
$$P(R = r \mid \pi, \xi) = (1 - \pi)\frac{1}{m} + \pi \binom{m-1}{r-1} (1 - \xi)^{r-1} \xi^{m-r}$$

The role of the Uncertainty Parameter

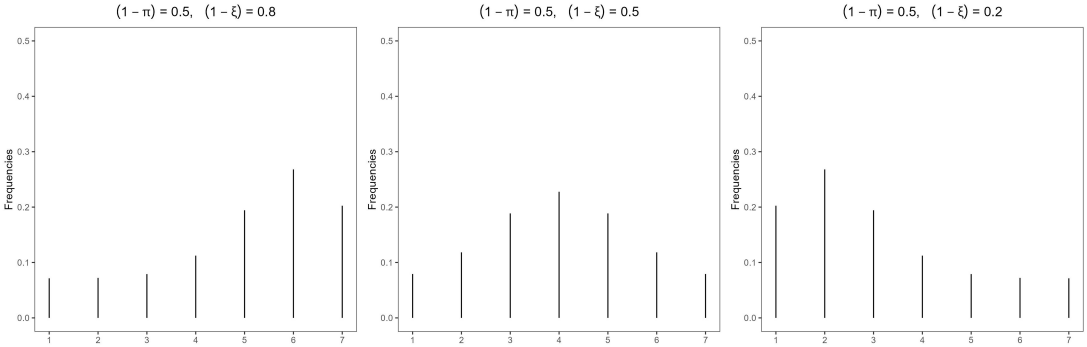
The Uncertainty Parameter $(1-\pi)$ directly quantifies the **level of uncertainty or indecision** in the respondent's choice.

- **If** $(1-\pi)=0$ **(i.e.,** $\pi=1$**):** The choice is entirely determined by the feeling (Binomial component). **No uncertainty**. The response distribution reflects the Binomial shape.
- **If** $(1-\pi)=1$ **(i.e.,** $\pi=0$**):** The choice is entirely determined by "uncertainty" (Uniform component). Feeling plays no role. The response distribution will be perfectly flat.
- **Values between 0 and 1:** Indicate a mix. A higher $(1-\pi)$ "flattens" the observed distribution towards a uniform shape.

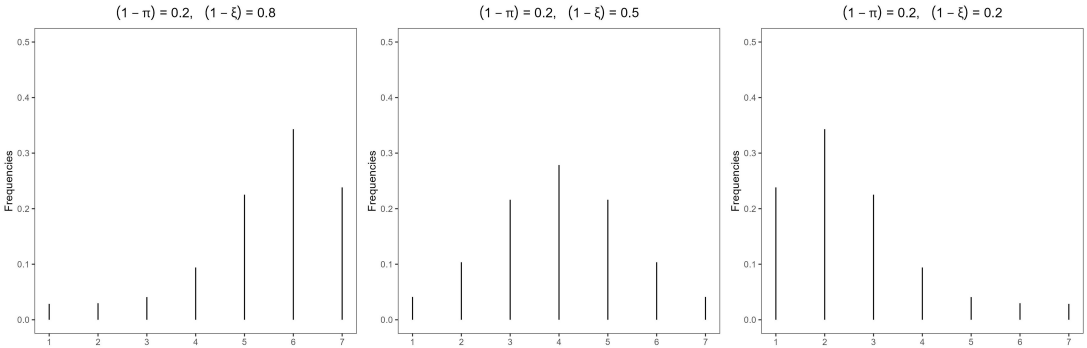
CUB Model with high uncertainty



CUB Model with medium uncertainty



CUB Model with low uncertainty



Model Identifiability and Estimation

- **Identifiability:**
  - A model is identifiable if different sets of parameter values lead to different probability distributions for the observed data.
  - The CUB model is identifiable if the number of categories $m > 3$.
- **Maximum Likelihood Estimation (MLE):**
  - The parameters (π, ξ) are typically estimated using MLE.
  - Since the CUB model is a mixture model, direct maximization of the log-likelihood is complex.
  - Therefore, the **Expectation-Maximization (EM) algorithm** is a common and robust iterative method for finding the MLEs.

Assessing the Goodness of Fit

- A particularly common and intuitive measure for assessing how well a CUB model fits the observed data is the **Dissimilarity ($Diss$) Index**.

$$Diss = \frac{1}{2} \sum_{r=1}^{m} | f_r - p_r(\hat{\theta}) |$$

where f_r are the observed relative frequencies and $p_r(\hat{\theta})$ are the estimated probabilities for the response categories.

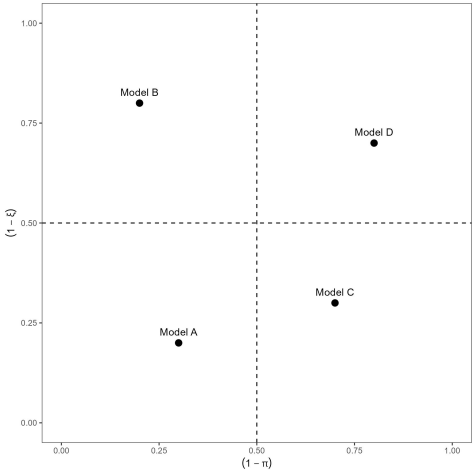
- **Interpretation:**
  - Values of $Diss$ closer to 0 indicate a **better fit**, with a perfect fit yielding $Diss = 0$.
  - The maximum value of $Diss$ is 1 (no overlap).
  - It measures the proportion of responses to be changed to achieve a perfect fit.

Advantage

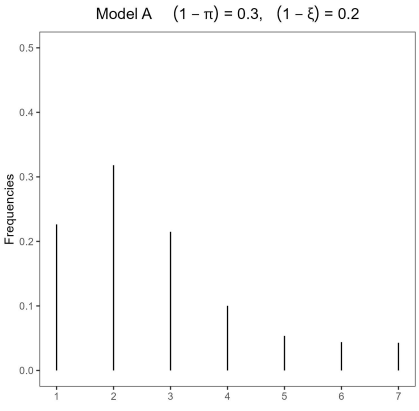
The Dissimilarity Index is less sensitive to low expected frequencies than Pearson's Chi-squared test and gives a clear indication of prediction accuracy.

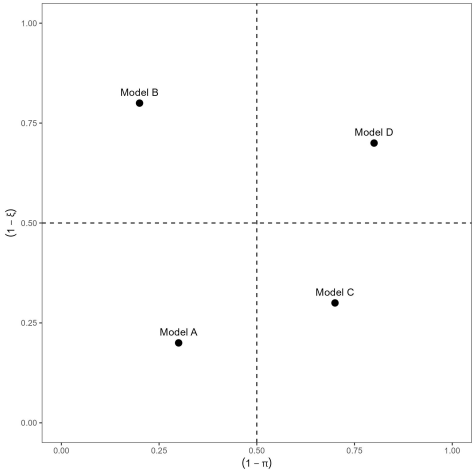
Parameter Space Visualization

- It is typically represented on a unit square where:
  - The **x-axis** represents $(1 - \pi)$, the **uncertainty level**.
  - The **y-axis** represents $(1 - \xi)$, the **feeling or attraction level**.
- This plot is helpful for comparing different datasets, subgroups, or even changes within a single group over time.

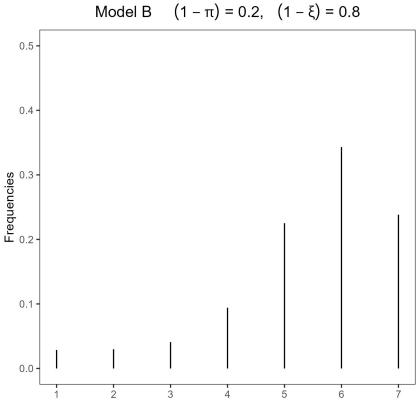


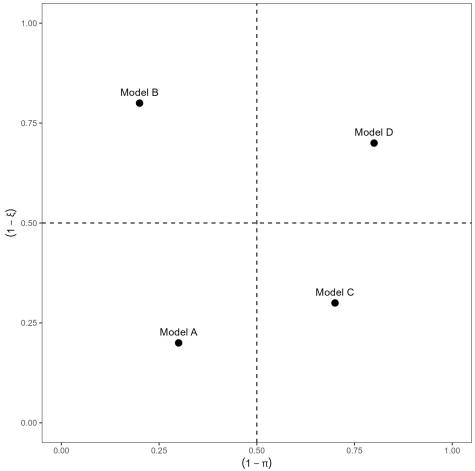
Parameter Space Visualization



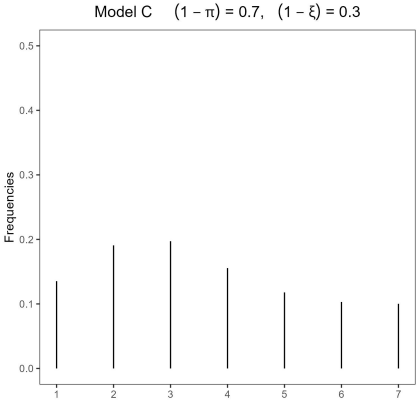


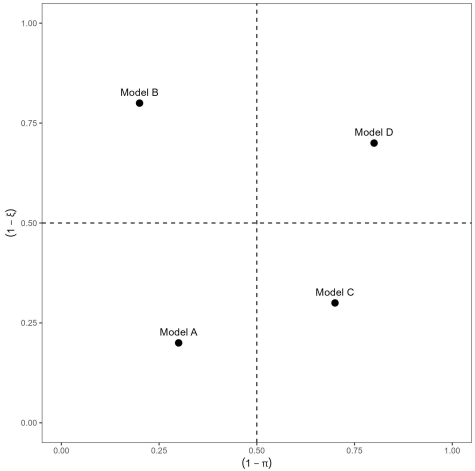
Parameter Space Visualization



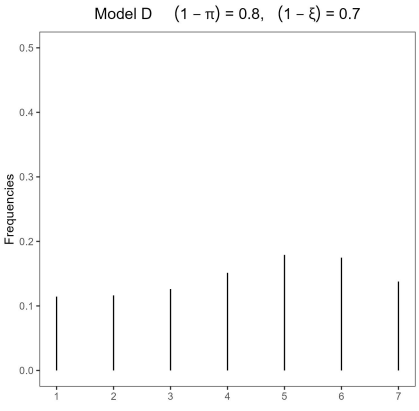


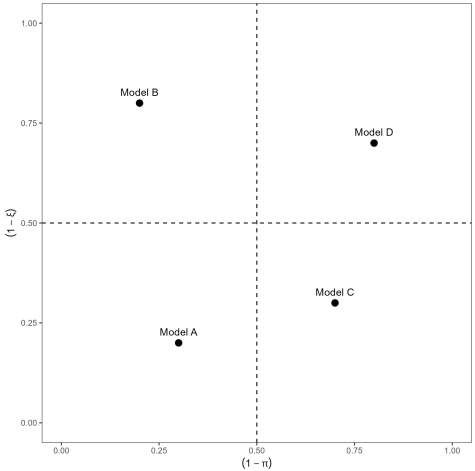
Parameter Space Visualization





Parameter Space Visualization



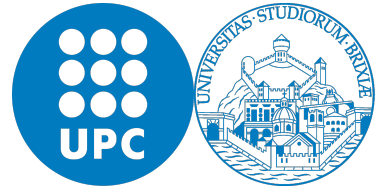


Introduction to CUB Models

CUB Model with covariates

CUB Model with Shelter Option

Treatment of ”Don’t Know” (DK) Options



Extensions of the CUB Model: Incorporating Covariates

- The basic CUB model assumes population **homogeneity** (parameters are constant for all individuals).
- This oversimplifies the reality of human responses.
- In real-world applications, psychological components (feeling and uncertainty) often **vary across individuals or groups**.
- These variations are systematically related to **observable characteristics** (covariates) of respondents or context.

Why Incorporate Covariates?

Moving beyond simple description to nuanced and insightful analysis.

Why Incorporate Covariates?

Incorporating covariates allows us to:

- **Explain heterogeneity in response patterns:**
  - Identify systematic factors explaining why different individuals/groups exhibit varying levels of feeling and uncertainty, instead of just seeing random noise.
- **Understand how specific characteristics influence feeling and uncertainty:**
  - Quantify the impact of variables like age on certainty (π) or education level on the tendency to rate highly ($1 - \xi$).
  - This shifts the analysis from mere description to **explanation and inference**.

Statistical Formulation: CUB Model with Covariates

- Core idea: Parameters π and ξ are **no longer fixed constants**.
- They are modeled as **functions of explanatory variables**.
  - Let γ_i be covariates for subject i influencing **uncertainty** (π_i).
  - Let w_i be covariates for subject i influencing **feeling** (ξ_i).
  - These covariate vectors (γ_i, w_i) can be the same, overlap, or be different.
- A **logistic (logit) link function** is used to ensure $\pi_i, \xi_i \in [0, 1]$.

Modeling the Uncertainty Parameter

- The log-odds of uncertainty are linked to covariates $\boldsymbol{\gamma}_i$:

$$\text{logit}(1-\pi_i) = \log\left(\frac{1-\pi_i}{\pi_i}\right) = \boldsymbol{\gamma}_i^T\boldsymbol{\beta}$$

  where $\boldsymbol{\beta}$ is a vector of coefficients for covariates $\boldsymbol{\gamma}_i$.

- From this, we derive the direct relationship for $(1-\pi_i)$:

$$(1-\pi_i) = \frac{\exp(\boldsymbol{\gamma}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{\gamma}_i^T\boldsymbol{\beta})}$$

- And consequently for π_i:

$$\pi_i = 1-(1-\pi_i) = \frac{1}{1+\exp(\boldsymbol{\gamma}_i^T\boldsymbol{\beta})}$$

Modeling the Feeling Parameter

- Similarly the log-odds of feeling for higher scores are linked to covariates $\boldsymbol{w}_i$:

$$\text{logit}(1-\xi_i) = \log\left(\frac{1-\xi_i}{\xi_i}\right) = \boldsymbol{w}_i^T\boldsymbol{\gamma}$$

  where γ is a vector of coefficients for covariates $\boldsymbol{w}_i$.

- This implies the following for $(1-\xi_i)$:

$$(1-\xi_i) = \frac{\exp(\boldsymbol{w}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{w}_i^T\boldsymbol{\gamma})}$$

- And for ξ_i:

$$\xi_i = 1-(1-\xi_i) = \frac{1}{1+\exp(\boldsymbol{w}_i^T\boldsymbol{\gamma})}$$

CUB Model with Covariates

The probability mass function for an observed rating $R_i = r$ for subject i becomes:

$$P(R_i = r \mid \pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} (1-\xi_i)^{r-1} \xi_i^{m-r} + (1-\pi_i)\frac{1}{m}$$

where π_i and ξ_i are now dynamically determined by the covariates $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ defined as follows:

$$\begin{cases} \text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{y}_i\boldsymbol{\beta}; \\ \text{logit}(\xi_i) = \ln\left(\frac{\xi_i}{1-\xi_i}\right) = \mathbf{w}_i\boldsymbol{\gamma}; \end{cases} \iff \begin{cases} \pi_i = \frac{1}{1+e^{-\mathbf{y}_i\boldsymbol{\beta}}}; \\ \xi_i = \frac{1}{1+e^{-\mathbf{w}_i\boldsymbol{\gamma}}}; \end{cases}$$

Interpretation of Covariate Effects

- Remember: coefficients operate on the **log-odds scale** due to the logistic link.

Interpretation of Covariate Effects

- Remember: coefficients operate on the **log-odds scale** due to the logistic link.
- **Coefficients β (for Uncertainty):**
  - Describe the impact of covariates on the log-odds of uncertainty.
  - A **positive** β_k: Increase in covariate y_{ik} (holding others constant) **increases log-odds of uncertainty**. This means $(1 - \pi_i)$ increases, making the choice more uniform/random.
  - A **negative** β_k: Increase in y_{ik} **decreases log-odds of uncertainty**. This means $(1 - \pi_i)$ decreases, making the choice more driven by feeling/less uncertain.

Interpretation of Covariate Effects

- **Coefficients γ (for Feeling):**
  - Describe the impact of covariates w_i on the log-odds of feeling towards higher scores.
  - A **positive** γ_k: Increase in covariate w_{ik} (holding others constant) **increases log-odds of feeling for higher scores**. This means $(1 - \xi_i)$ increases, shifting preference towards the higher end of the scale.
  - A **negative** γ_k: Increase in w_{ik} **decreases log-odds of feeling for higher scores**. This means $(1 - \xi_i)$ decreases, shifting preference towards the lower end of the scale.

Interpretation of Covariate Effects

- **Coefficients γ (for Feeling):**
  - Describe the impact of covariates w_i on the log-odds of feeling towards higher scores.
  - A **positive** γ_k: Increase in covariate w_{ik} (holding others constant) **increases log-odds of feeling for higher scores**. This means $(1 - \xi_i)$ increases, shifting preference towards the higher end of the scale.
  - A **negative** γ_k: Increase in w_{ik} **decreases log-odds of feeling for higher scores**. This means $(1 - \xi_i)$ decreases, shifting preference towards the lower end of the scale.

Important Note

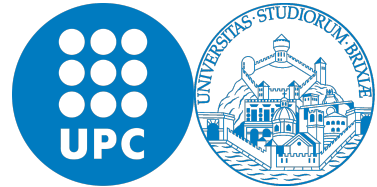
CUB models with covariates are **not Generalized Linear Models (GLMs)** in the strict sense. While π_i and ξ_i use GLM-like structures (logistic regressions), the response variable itself (the mixture) doesn't belong to the exponential family.

Introduction to CUB Models

CUB Model with covariates

CUB Model with Shelter Option

Treatment of ”Don’t Know” (DK) Options



CUB Model with Shelter Option

- While the basic CUB model captures feeling and uncertainty, real-world data often show additional complexities.
- One common phenomenon: respondents disproportionately select a specific category.
- It suggests this category acts as a **”shelter”** or **”refuge”** for some respondents
- This is called **shelter effect**, which describes a specific category receiving an ”extra” probability mass.
  - This goes beyond what’s predicted by feeling or pure uncertainty.
  - Occurs because some respondents gravitate towards this category for reasons **unrelated to precise preference or indecision**.

MESIO SUMMER SCHOOL

Reasons for ”Shelter-Seeking” Behavior

- **Cognitive Simplification:** Choosing an easy, less demanding option to reduce mental effort (e.g., the middle category, a neutral option, or even first/last if easy default).
- **Fatigue or Boredom:** In long questionnaires, respondents may disengage and pick a convenient category instead of carefully considering.
- **Social Desirability or Privacy Concerns:** Selecting a ”safe”, non-committal, or socially acceptable answer to avoid strong opinions or protect privacy. The neutral option often serves this.
- **Questionnaire Design:** Poorly worded questions, ambiguous scale anchors, or overwhelming options might inadvertently guide respondents to a default or ambiguous middle ground.
- **Satisficing:** Tendency to select a minimally acceptable response rather than the optimal one, often to save cognitive resources. The shelter option becomes the ”good enough” answer.

Implication of the Shelter Effect

- The critical implication: the chosen category c has an **"extra" probability mass**.
- The observed frequency for category c is higher than what a standard two-component CUB model would predict.
- Effectively, a subset of respondents might be selecting c **directly**, bypassing the usual feeling-uncertainty decision process.

MESIO SUMMER SCHOOL

Defining the Shelter Category

- The shelter category, denoted by c ($c \in \{1, 2, \ldots, m\}$), is a specific category with this inflated probability.
- Identifying this category is a crucial preliminary step.
- It can be:
  - **Hypothesized *a priori*:** Based on scale design.
    - E.g., on a 5-point Likert scale, $c = 3$ ("Neutral") is a common candidate.
    - Min ($c = 1$) or max ($c = m$) could also function as defaults.
  - **Identified Empirically:** If no *a priori* hypothesis exists.
    - Observe unusually high frequency for a category not well explained by a simple CUB model.
    - A large positive residual for a specific category from a basic CUB fit suggests a potential shelter effect.
- Once identified, the shelter category c is treated as fixed in model estimation.

Statistical Formulation: CUB Model with Shelter Option

- To account for the shelter effect, the basic CUB model is extended into a **three-component mixture model**.
- A common approach is the **CUB with Shelter** model.
- This model introduces a third component: a **degenerate distribution** that assigns all probability mass exclusively to the shelter category c.
- The **Probability Mass Function (PMF)** for the CUSH model for a rating $R = r$ is:

$$P(R = r \mid \pi, \xi, \delta) = \delta \Big[D_c(r) \Big] + (1 - \delta) \Big[\pi B(r \mid m - 1, 1 - \xi) + (1 - \pi) U(r \mid m) \Big]$$

Degenerate Distribution for Shelter Category ($D_c(r)$)

- This is the additional component compared to the basic CUB model.
- It's a degenerate distribution that places all probability on the shelter category c:

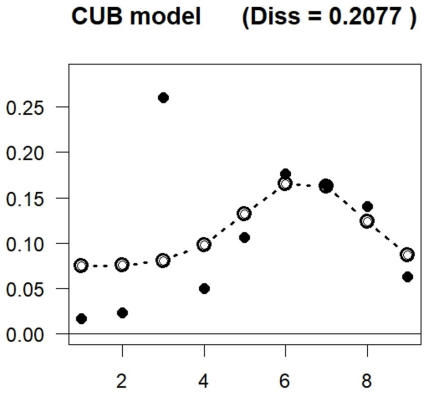
$$D_c(r) = \begin{cases} 1, & \text{if } r = c; \\ 0, & \text{otherwise;} \end{cases} \qquad r = 1, \ldots, m.$$

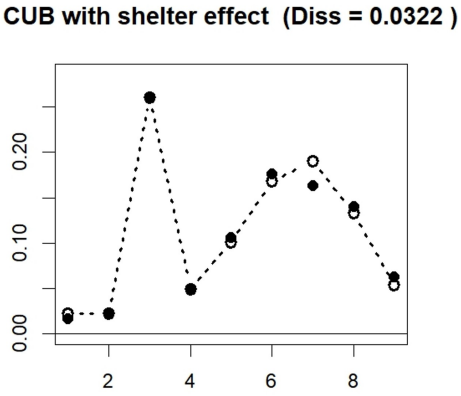
- The parameter $\delta \in [0, 1]$ represents the **probability of choosing the shelter category c directly**.
  - This is the weight assigned to the degenerate distribution.
  - A higher δ indicates a stronger tendency for respondents to opt for the designated shelter category, irrespective of their true feeling or general uncertainty.

Model Selection: CUB vs CUB with shelter

- **Information Criteria:**
  - **Akaike Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)**.
  - **BIC is often preferred for CUB models** as it penalizes complexity more heavily, aiding in choosing more parsimonious models.
  - Lower values indicate a better-fitting model.
- **Residual Analysis:**
  - After fitting a basic CUB, examine residuals (differences between observed and fitted probabilities).
  - A **large positive residual** for a specific category strongly suggests a shelter effect for that category.
  - The **Dissimilarity Index** ($Diss$) can be used to compare the overall fit improvement when a shelter component is added.

Model Selection: CUB vs CUB with shelter



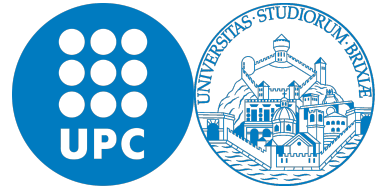


Introduction to CUB Models

CUB Model with covariates

CUB Model with Shelter Option

Treatment of ”Don’t Know” (DK) Options



Treatment of ”Don’t Know” (DK) Options within CUB Framework

- Surveys often include ”Don’t Know” (DK), ”No Opinion,” or ”Not Applicable” options.
- These non-substantive responses pose a significant challenge for traditional statistical modeling.
- They don’t fit neatly into standard analytical frameworks.
- The **CUB framework** offers a theoretically robust and psychologically insightful approach to incorporating information from DK responses.
- DK responses are not just missing data; they represent an **active choice** reflecting a specific cognitive or attitudinal state.

Why DK Responses Are Problematic

- **Not Simply Missing Data:**
  - Discarding DKs (listwise deletion) can lead to **biased samples and results**.
  - If DK responders differ systematically from substantive responders, removing them distorts sample representativeness and affects generalizability.
- **Imputation Issues:**
  - Imputing a value for DK is inherently difficult and relies on **strong, untestable assumptions**.
  - Assigning a central value (mean/median) might mask genuine uncertainty.
  - Complex imputation methods (e.g., multiple imputation) can be challenging for ordinal data.
- **Adding DK as a Category:**
  - Including DK as another ordinal category **breaks the ordinal nature** of the scale.
  - A sequence like ”Strongly Disagree, Disagree, DK, Agree, Strongly Agree” is conceptually problematic.
  - DK is not naturally ordered between ”Disagree” and ”Agree”; it’s a different type of response altogether.

Understanding Underlying Meanings of DK

The varied meanings of DK emphasize that it's a **rich source of information**, not just data to be discarded or arbitrarily filled in.

- **True Lack of Knowledge/Opinion:** The respondent genuinely has no information, experience, or hasn't formed an opinion.
- **Unwillingness to Answer:** The respondent might have an opinion but chooses not to express it (e.g., sensitive topic, privacy, social desirability).
- **Inability to Map Opinion to Scale:** The respondent has an opinion but feels the provided categories are inadequate, too vague, or don't fit their nuanced view (e.g., feeling falls "between" categories).
- **Question Ambiguity:** The respondent doesn't understand the question, its assumptions, or terms, leading them to pick DK as an escape.

MESIO SUMMER SCHOOL

CUB Framework Approach to ”Don’t Know”

- The CUB framework approaches DK responses by thinking of the total population as having two unobserved (**latent**) groups:
  1. **Group A=0: Those who *can* give a substantive rating** on the m-point scale.
     - These people have a genuine underlying feeling or are able to make a choice.
  2. **Group A=1: Those who *cannot* (or would not) and would genuinely choose DK** if it were an option.
     - This group essentially represents the ”true” non-responders when it comes to having a substantive opinion.
- We use p_{DK} to represent the proportion of individuals in the population who would choose DK.

Modeling Assumptions for Latent Groups

This approach makes specific assumptions about how each latent group generates responses:

- **For those who can answer (Group A=0, proportion $(1 - p_{DK})$):**
  - Their responses ($R = r$) are assumed to follow a **standard CUB model**.
  - This CUB model has its own parameters: π_0 (uncertainty within this group) and ξ_0 (feeling within this group).

$$P(R = r \mid A = 0, \pi_0, \xi_0) = \pi_0 B(r \mid m - 1, 1 - \xi_0) + (1 - \pi_0) U(r \mid m)$$

- **For those who would choose DK (Group A=1, proportion p_{DK}):**
  - If forced to pick from the m-point scale (e.g., DK option unavailable), their choice is not based on genuine "feeling."
  - Responses are driven purely by randomness or uncertainty across available options.
  - Modeled by a **discrete Uniform distribution**:

$$P(R = r \mid A = 1) = U(r \mid m) = \frac{1}{m}$$

Overall Observed Distribution

- The overall observed distribution of ratings, if all respondents are forced to choose from the m-point scale (i.e., no explicit DK option, or considering hypothetical responses of those who would pick DK):
- It's a **mix of the CUB model for "knowers" and the Uniform distribution for those "forced" to choose**.

Adjusting CUB Parameters using Observed DKs

When DK responses are explicitly allowed in a survey and present in the data, the approach proceeds as follows:

1. **Estimate $\hat{p}_{DK}$:** The observed proportion of DK responses in the sample.

$$\hat{p}_{DK} = \frac{\text{Total number of DK responses}}{\text{Total number of responses}}$$

2. **Focus on Substantive Responders:** The remaining $(1 - \hat{p}_{DK})$ proportion of the sample consists of individuals who gave a rating on the m-point scale $(1, \ldots, m)$. Let N_{sub} be this number.
3. **Model for Substantive Responses:** A standard CUB model is fitted **only to these N_{sub} substantive responses**.
   - This gives estimates for their underlying parameters: π_S (uncertainty among substantive responders) and ξ_S (feeling among substantive responders).

$$P(R = r \mid \text{Substantive}) = \pi_S B(r \mid m - 1, 1 - \xi_S) + (1 - \pi_S) U(r \mid m)$$

Adjusting CUB Parameters using Observed DKs (Cont.)

4. **Relating to Overall Population Parameters:** The crucial step is linking (π_S, ξ_S) back to the overall population's true feeling and uncertainty, accounting for $\hat{p}_{DK}$.
   - The **feeling parameter for the overall population**, $(1 - \xi)$, is best represented by the feeling of those who gave a substantive rating:

$$(1 - \xi) = (1 - \xi_S)$$

     (People choosing DK don't contribute to the "feeling" aspect.)
   - **The overall uncertainty parameter for the population**, $(1 - \pi)$, comes from two sources:
     - Inherent uncertainty among those who could answer (captured by $1 - \pi_S$).
     - Complete uncertainty of those who chose DK (considered 100% uncertain regarding the m-point scale).

Adjusting the Uncertainty using Observed DKs

- The overall π for the population (probability a response is driven by feeling) is:

$$\pi = (1 - \hat{p}_{DK}) \cdot \pi_S$$

This means: (probability not a DK type) $\times$ (probability, given not DK, of responding based on feeling).

- Consequently, the overall uncertainty for the population is:

$$(1 - \pi) = 1 - (1 - \hat{p}_{DK}) \cdot \pi_S$$

This can be rewritten as:

$$(1 - \pi) = \hat{p}_{DK} + (1 - \hat{p}_{DK}) \cdot (1 - \pi_S)$$

This means that the overall uncertainty is the sum of the DK proportion and the weighted uncertainty among substantive responders.