Latent Variable Methods to Address Measurement Error Bias from Ordered Categorical Predictors

Ramses Llobet*

Working paper - Draft

Abstract

In many research settings, estimating the effect of a continuous predictor Z on an outcome Y is complicated by the fact that Z is often not directly observed but instead measured through an ordered categorical variable R. A common strategy is to treat R as a continuous variable or, when Z has an interpretable metric (such as income or time), to impute Z using interval midpoint imputation. However, both approaches can introduce or exacerbate measurement error, resulting in biased and inefficient estimates. This study evaluates three methods for recovering Z from R, using binned continuous variables as a motivating example: (1) midpoint imputation, (2) interval regression, and (3) Bayesian rank likelihood. Monte Carlo experiments compare these methods in terms of prediction error for Z and the resulting bias and efficiency when Y is regressed on $\hat{Z}$. Preliminary results indicate that parametric estimation and prediction of Z using model-based approaches outperform midpoint imputation in both bias and efficiency. Furthermore, these results show that researchers should avoid treating categorical variables as continuous predictors, as this strategy produces highly biased and inconsistent estimates, regardless of the number of k categorical levels.

*Ph.D. candidate, Department of Political Science, University of Washington. contact: rllobet@uw.edu

Introduction

Ordered categorical indicators are ubiquitous in survey research, offering a pragmatic solution when precise measurement is impractical or intrusive. Whether assessing educational attainment, attitudes, or economic status, researchers frequently rely on respondents' self-placement into a set of predefined, rank-ordered categories. Among these, income stands out as both one of the most theoretically important and the most frequently coarsened variables. Instead of recording exact amounts, surveys routinely ask individuals to select the interval that best represents their earnings, resulting in data that is ordinal rather than continuous.

This widespread measurement practice introduces subtle but important statistical challenges. Treating these ordered categories as continuous predictors—a common shortcut—can lead to inefficient or biased inference. The issue is further compounded by censoring mechanisms inherent in survey design. Top-coding, for example, places all high-income respondents in a single upper bracket, obscuring genuine variation at the top of the distribution. Likewise, discretizing income into intervals causes information loss, as the true value is only known to fall within a certain range.

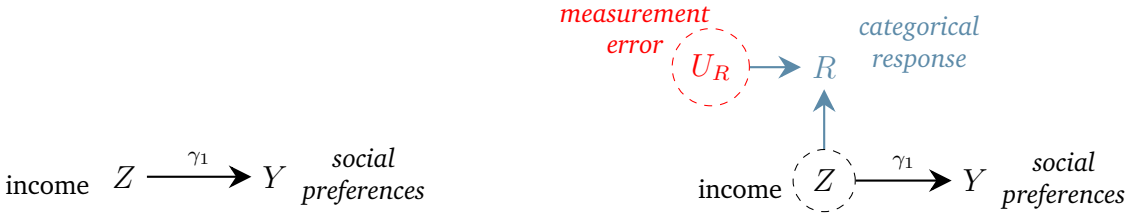
Despite the ubiquity of these problems, measurement error induced by using grouped income variables is rarely addressed explicitly in applied research. Most practitioners resort to one of two ad hoc solutions: treating the ordinal indicator as a continuous variable, or imputing the latent value of income by substituting the interval midpoint—sometimes adjusting the upper category via a parametric formula such as Pareto extrapolation (Hout, 2004). Both approaches are widely used, but their consequences for bias and statistical efficiency are not well understood.

This paper tackles these pervasive but underappreciated methodological issues. I systematically assess the measurement error bias that arises from these two common strategies and evaluate their performance alongside two formal latent variable modeling ap-

proaches: interval regression and Bayesian rank likelihood. Using Monte Carlo experiments designed to mimic typical survey data, I compare the bias, efficiency, and prediction error of each approach. The findings provide practical guidance for applied researchers on how to obtain more reliable inferences when working with ordered categorical predictors that represent coarsened continuous variables such as income.

Measurement Error in Categorical Predictors

Figure 1 illustrates the core measurement problem motivating this study. In much of social science, individual income is theorized as a key determinant of attitudes and behaviors, serving as a proxy for social class in models of gender role attitudes (Desai, Chugh and Brief, 2014), emotional well-being (Yang and Ma, 2020), racial-ethnic differences (Riegle-Crumb and Grodsky, 2010), environmental policy support (Pampel and Hunter, 2012), and preferences for inequality and redistribution (Rueda and Stegmueller, 2019). Causal frameworks, like that depicted in panel (a), assume that true income Z—a continuous, meaningful metric—directly shapes individual preferences.



(a) No measurement error.

(b) Measurement error.

Figure 1: **Measurement Error.** The left panel (a) depicts a causal model without measurement error, while the right panel (b) illustrates measurement error arising from item non response, classification error, underreporting, and censoring. Nodes represent variables, and dashed-line circles indicate unobservable variables.

Yet, in practice, survey income measurements almost never provide direct, continuous values. Instead, as depicted in the right-hand panel of Figure 1, researchers must rely on respondents' self-assignment into ordered income brackets R, which are simply coarsened representations of the underlying metric variable Z. While Z (true income) is continuous

and meaningful in real-world units, what is observed is a censored or grouped outcome: respondents indicate only the interval their income falls into, or in the case of top-coding, whether it exceeds an upper threshold.

This coarsening of continuous variables is ubiquitous in the social sciences and beyond. Grouped or interval-censored variables are found not only in survey research but also in experimental, clinical, and administrative data, especially when practical or ethical reasons make precise measurement infeasible. Each instance of converting a continuous latent trait to categories results in loss of information.

Measurement error in such settings arises mainly from two sources: misclassification—including forms of nonresponse and underreporting—and censoring due to incomplete observation of the metric. Both forms of error tend to be mean-reverting, inducing bias toward zero in regression coefficients when such variables are used as predictors (Bound, Brown and Mathiowetz, 2001; Lustig, 2020). This paper focuses specifically on the bias arising from censoring.

While outcome-censoring and its inferential challenges have been widely analyzed, the effect of censoring in predictors—such as income—is far less appreciated. The problem is especially acute for self-reported income, where top-coding and underreporting are common and often unaddressed (Lustig, 2020).

Despite these issues, the prevailing analytic practice is to treat ordinal variables—such as grouped income—as if they were continuous. Two strategies are especially common: treating R as a continuous variable when it has many categories, or imputing the latent metric Z by using the midpoints of reported intervals.

One might ask: why not always fit these variables using dummy variables, as is standard for other categorical predictors? In many applications, researchers need to recover the latent continuous scale, for example, to harmonize income data across surveys and years (e.g., inflation adjustments, Hout 2004), or to aggregate across measures that share a common metric (Nakagawa and Sozu, 2024).

However, both strategies—fitting R as continuous or applying midpoint imputation—are likely to introduce bias and inconsistency in regression estimates for the effect of income (γ), as shown in Figure 1. These biases are expected to be most severe when the latent distribution of Z departs from normality, as is frequently the case for income and other economic variables.

I therefore hypothesize that both the practice of treating ordered categorical variables as continuous and the use of midpoint imputation yield biased and inconsistent estimates of γ. This bias is expected to increase as the distribution of Z becomes more skewed or heavy-tailed—precisely the scenario with income data.

To systematically investigate these issues, the remainder of this paper proceeds as follows: First, I present Monte Carlo simulation evidence on the magnitude of measurement error bias arising from fitting categorical predictors as continuous or applying midpoint imputation. Next, I introduce and evaluate latent variable models for R that make inferences to Z and allow to predict the latent variable with reduced measurement error—specifically, interval regression and Bayesian rank likelihood methods. Subsequent simulation analyses compare these approaches in terms of bias and efficiency of γ, and prediction error of Z. I provide preliminary conclusions and a discussion of these results future research steps of this project.

First Monte Carlo Experiment: Bias in Categorical Predictors

This section presents a series of Monte Carlo experiments designed to examine how the estimation of γ_1, the marginal effect of Z on Y, is affected when Z is unobserved and only an ordered categorical proxy R is available. Building on the data-generating process illustrated in Figure 1(b), the simulations focus on two widely used approaches: (1) treating R as a continuous predictor, and (2) constructing a midpoint-imputed variable $\hat{Z}_{\text{mp}}$ based on R. The experiments systematically explore how the performance of these methods varies with (a) the number of categories k used to discretize Z, and (b) the

distributional shape of the underlying latent variable Z.

In each scenario, the outcome Y is generated as a linear function of Z with a true marginal effect $\gamma_1 = 1$ and standard normal errors ($\epsilon \sim \mathcal{N}(0, 1)$). To assess the impact of distributional assumptions, three distinct forms for Z are considered: normal ($\mathcal{N}(0, 1)$), lognormal (Lognormal$(0, 0.2)$), and Pareto (Pareto$(0, 1.5)$). Additional sensitivity analyses explore variations in the residual variance for the lognormal case and in the tail parameter α for the Pareto case, allowing a broader assessment of the conditions under which bias may arise.

After sampling Z, it is discretized into k ordered categories R, defined by threshold values (τ) and their associated interval bounds $[L, U)$. This process is visualized in Figure 3. For midpoint imputation, each observation in category k receives the midpoint of its interval $[L_i, U_i)$. For the top-coded category, right-censoring is addressed using either a Pareto-based extrapolation (Hout, 2004) or, for the normal and lognormal cases, a constant spacing approach. The simulations evaluate estimation performance as the number of categories k increases from 3 to 30, encompassing the common empirical practice of treating ordinal variables with five or more categories as continuous (Rhemtulla, Brosseau-Liard and Savalei, 2012; Robitzsch, 2020).

Figure 2 summarizes the bias in estimating γ_1 across all simulation conditions, with the top panel showing results for R as a continuous predictor and the bottom panel for midpoint imputation. All point estimates are mean-centered on the true γ_1, so departures from zero indicate bias.

The findings are striking: when R is modeled as continuous, the resulting estimates of γ_1 are systematically biased and inconsistent, regardless of the number of categories. Increasing k does not alleviate the bias; in fact, estimates converge further from the true value, with mean-centered bias approaching -1 (attrition towards 0). This result clearly demonstrates that treating ordinal variables as continuous is inappropriate as far as we assume that the categorical predictor is a discretized realization of an underlying continuous

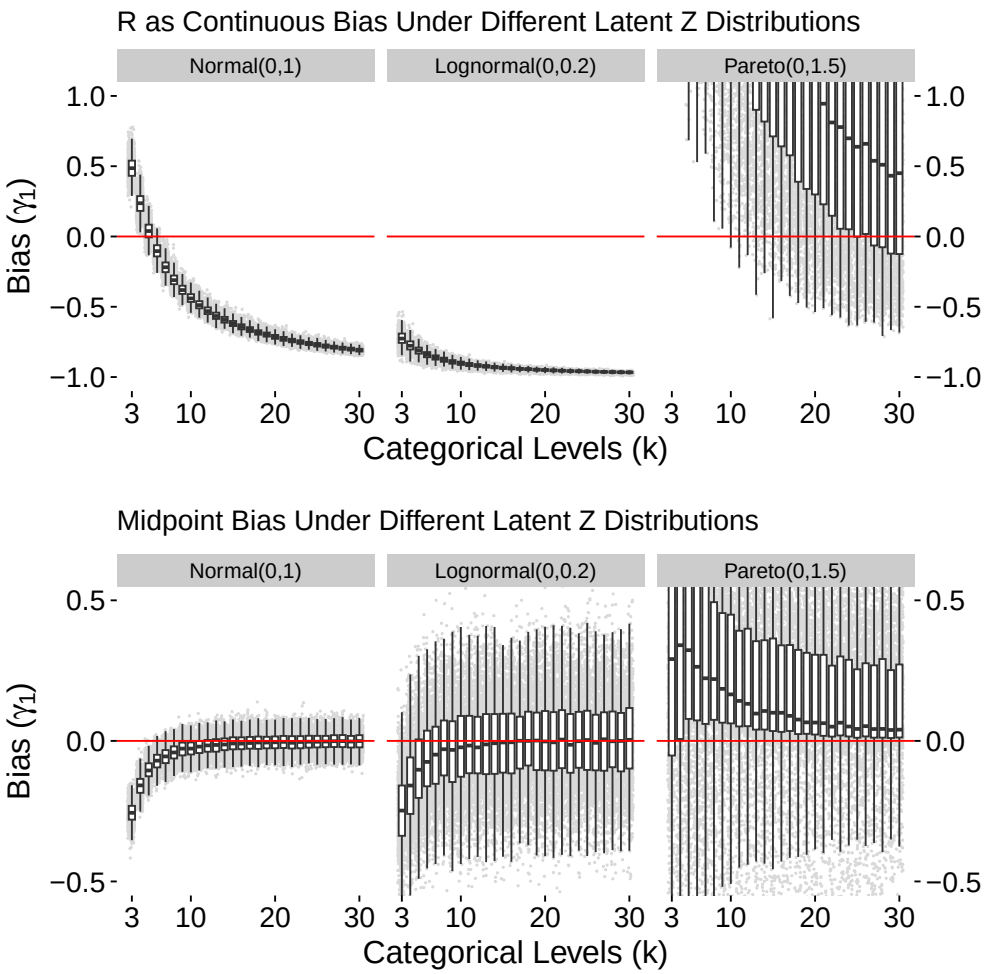


Figure 2: **Bias in γ_1 for continuous and midpoint imputation approaches across categorical levels.** Each panel compares the distribution of bias in estimating γ_1 as a function of the number of categories k, under different latent Z distributions (Normal, Lognormal, and Pareto). Boxplots show the distribution of estimator bias at each k, with added jittered points illustrating the variability and coverage of individual simulation runs. All bias values are mean-centered relative to the true value ($\gamma_1 = 1$), so deviations from zero represent bias. The y-axis is truncated uniformly across panels to facilitate direct comparison between the continuous (top) and midpoint (bottom) estimators under each condition.

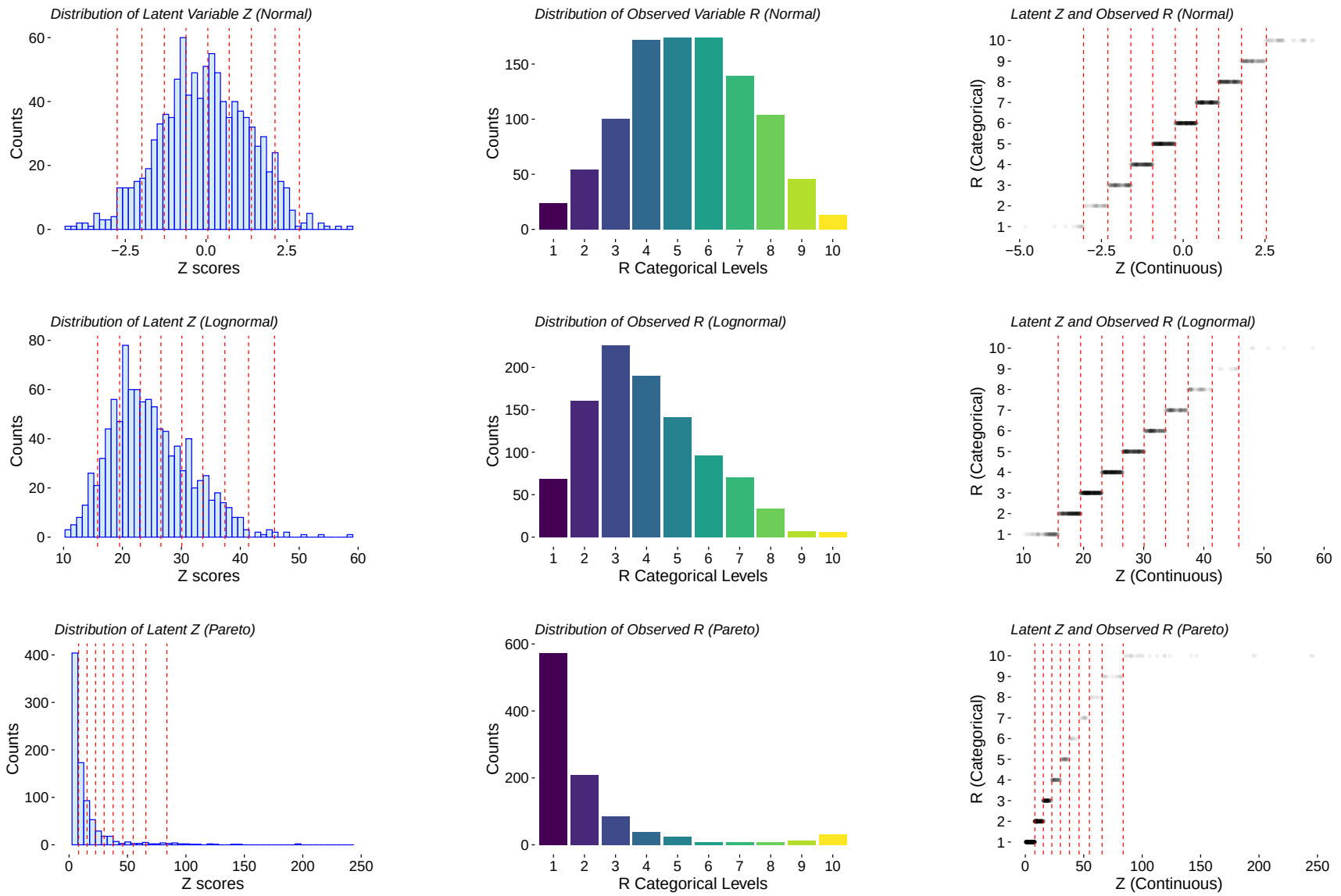


Figure 3: Simulated latent variable Z and response R under three distributions: (top) Normal, (middle) Lognormal, (bottom) Pareto. Columns show the marginal Z distribution, the marginal R, and the joint (Z, R).

distribution.

Midpoint imputation, by contrast, yields estimators whose bias generally declines as k increases, but the rate and direction of this convergence depend heavily on the distribution of Z. For normal and lognormal Z, bias becomes negligible for $k > 14$, though the lognormal case shows higher estimator variance and less efficiency. For intermediate values of k (between 7 and 14), there is still some minor downward bias. In the Pareto case, however, midpoint imputation produces persistent upward bias and substantial estimator variance, even as k approaches 20, reflecting the influence of heavy tails and skewness in the latent variable.

Figure 4 provides further detail. For lognormal Z, bias is minimal for low residual variances ($\text{Var}(\epsilon) < 0.6$) and when $k > 6$, but increases and turns positive as variance and skewness grow. For Pareto Z, heavier tails (lower α) are associated with even higher, persistent upward bias, regardless of the number of categories.

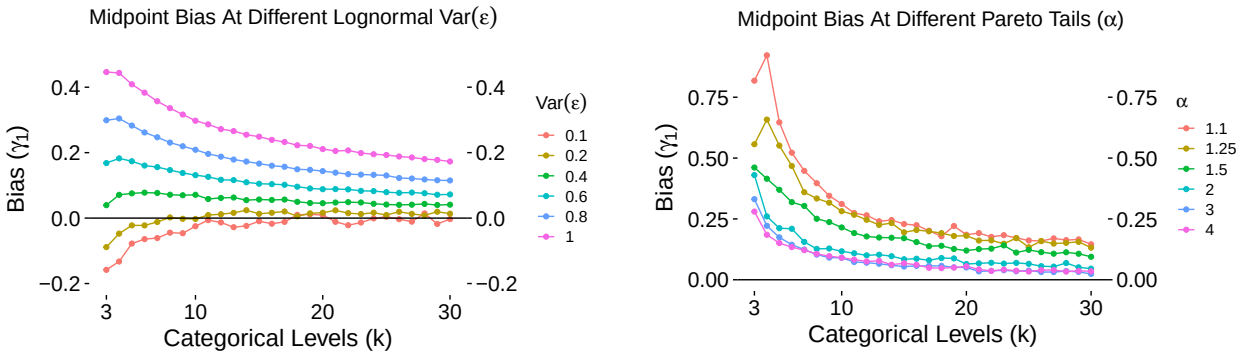


Figure 4: Bias under different latent-Z distributions. The figures show mean

In summary, the simulations yield two key conclusions. First, ordinal categorical variables should never be modeled as continuous predictors, since this approach is always biased and inconsistent—with attenuation toward zero becoming more severe as k increases. Second, midpoint imputation can produce unbiased and consistent estimates, but only when the number of categories is sufficiently large ($k \geq 10$) and the latent Z is approximately normal. When Z is skewed or heavy-tailed, or when k is small, both the direction

and magnitude of bias from midpoint imputation become unpredictable—highlighting the importance of careful model specification and sensitivity analysis in applied research.

Latent Variable Methods for Ordered Categories

The preceding analysis showed that midpoint imputation, though attractive for its simplicity and minimal assumptions, yields biased estimates of γ_1 when the number of ordered categories k is less than 10—even when the latent variable Z is close to normal. As Z becomes more variable or skewed, this bias remains substantial even with a larger number of categories. While midpoint imputation requires little more than knowledge of interval thresholds and a top-category extrapolation rule, its performance is highly sensitive to the underlying distribution of Z.

Given these limitations, a natural question emerges: can we achieve more accurate and efficient inference by explicitly modeling the relationship between the observed categories and their latent origins? What do we gain or risk by introducing parametric or semiparametric assumptions?

This section introduces two modeling frameworks that treat observed ordered categories as manifestations of a latent continuous variable: interval regression and the Bayesian rank likelihood. Both methods aim to leverage additional information—whether about the interval structure or the rank ordering of categories—to recover Z and, in turn, improve estimation of γ_1. I detail each approach below and then assess, through further simulation, whether these methods can reduce measurement error bias relative to midpoint imputation.

Ordinal and Interval Regression

The relationship between the observed ordered categorical variable R and the latent continuous variable Z can be formalized via a monotonic function $g(\cdot)$, where $R = g(Z)$. This transformation partitions partitions the range of Z into discrete, ordered categories

labeled $k = 1, 2, \ldots, K$, , determined by a set of thresholds (cutpoints) $\tau = \tau_1, \tau_2, \ldots, \tau_{K-1}$, satisfying the ordering:

$$-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K-1} < \tau_K = +\infty.$$

An individual i is observed in category k if their latent value Z_i falls between the corresponding thresholds:

$$R_i = k \quad \text{if and only if} \quad \tau_{k-1} \leq Z_i < \tau_k.$$

Accordingly, the probability that respondent i selects category k is

$$\Pr(R_i = k \mid X_i) = \Pr(\tau_{k-1} \leq Z_i < \tau_k \mid X_i)$$

When Z represents an abstract construct lacking a natural metric (as is common with subjective constructs such as attitudes or opinions), model identification requires normalizing its scale. The standard approach assumes $Z_i \sim F(\mu, 1)$, where F is a cumulative distribution function (CDF), and $\mu = X_i'\beta$ is the systematic component:

$$Z_i \sim F(X_i'\beta, 1)$$

The standarized latent variable is

$$W_i = \frac{Z_i - \mu}{1} \sim F(0, 1)$$

so the probability of observing category k is

$$\Pr(R_i = k) = F(\tau_k - X_i'\beta) - F(\tau_{k-1} - X_i'\beta)$$

If F is chosen as the standard normal (Φ), this is the ordered probit model; with the

logistic CDF (Λ), it is the ordered logit model. Here, the coefficients β and thresholds τ are estimated via maximum likelihood, with the variance fixed (typically $\sigma^2 = 1$), since the latent scale is arbitrary.

In contrast, if the observed categories correspond to explicit intervals of a latent variable with a known metric (e.g., income brackets, age groups, etc), each observed response maps directly to a known interval $[L_i, U_i)$ for Z_i. Here, the latent variable is modeled with estimable conditional variance:

$$Z_i \sim F(X_i'\beta,\, \sigma^2).$$

The probability of Z_i falling within its interval is:

$$\Pr(L_i \leq Z_i < U_i) = F\left(\frac{U_i - X_i'\beta}{\sigma}\right) - F\left(\frac{L_i - X_i'\beta}{\sigma}\right).$$

With fixed interval bounds $[L_i, U_i)$, both β and σ^2 can be estimated jointly via maximum likelihood, since the intervals provide information on the latent scale. This additional scale information makes interval regression more efficient than ordinal regression, resulting in smaller standard errors.

A further benefit is that, after estimating $\hat{\beta}$ and $\hat{\sigma}^2$, one can predict the latent variable Z using any desired summary (mean, median, quantiles), conditional on the assumed parametric choice of $F(.)$. Unlike ordinal regression, which relies on scale normalization, interval regression exploits the observed metric, providing both more flexible inference and a more interpretable scale for the latent variable.

Bayesian Rank Likelihood

The Bayesian rank likelihood approach provides a semiparametric alternative for modeling ordered categorical data, enabling inference based exclusively on the ordinal information (ranks) of the observed responses without imposing explicit assumptions about

threshold locations or spacing between categories. This method is particularly useful when the precise mapping from the latent variable to observed categories is unknown or difficult to specify (Pettitt, 1982; Hoff, 2008, 2009).

As before, let $R = (R_1, \ldots, R_n)$ denote the observed ordered categorical outcome derived from an underlying continuous latent variable $Z = (Z_1, \ldots, Z_n)$, which follows the latent regression model

$$Z_i = X_i'\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where X_i is a vector of covariates, β is a vector of regression coefficients, and σ^2 is the residual variance. The observed ordinal response R_i is linked to Z_i via an unknown monotonic transformation $g(\cdot)$, such that $R_i = g(Z_i)$. This means that the exact numeric thresholds or spacing between categories are unknown, but the rank ordering is preserved. Although both the transformation $g(\cdot)$ and the scale of Z remain unspecified, the observed ranks impose order constraints on the latent variable. Specifically, if $R_{i_1} < R_{i_2}$, then necessarily $Z_{i_1} < Z_{i_2}$.

The vector of latent variables $\mathbf{Z} \in \mathbb{R}^n$ is therefore constrained to lie within the set

$$\mathcal{R}(R) = \{\mathbf{z} \in \mathbb{R}^n : z_{i_1} < z_{i_2} \quad \text{whenever} \quad R_{i_1} < R_{i_2}\},$$

which includes all possible latent configurations consistent with the observed ordering of R. Consequently, the rank likelihood for (β, σ^2) is defined as the probability that a realization of $\mathbf{Z}$ from the regression model falls within this permissible set:

$$L(\beta, \sigma^2) = P\left(\mathbf{Z} \in \mathcal{R}(R) \mid X, \beta, \sigma^2\right).$$

Direct maximum likelihood estimation of this rank likelihood is computationally challenging. Instead, Bayesian estimation using Markov Chain Monte Carlo (MCMC) provides a robust and efficient solution (Hoff, 2008). We assign independent priors $\beta \sim \mathcal{N}(0, I)$

and $\sigma^2 \sim$ Inv-Gamma(a_0, b_0), yielding the joint posterior distribution

$$p(\beta, \sigma^2, \mathbf{Z} \mid R, X) \propto p(\beta)\, p(\sigma^2)\, p(\mathbf{Z} \mid X, \beta, \sigma^2)\, I(\mathbf{Z} \in \mathcal{R}(R)),$$

where $I(\cdot)$ is the indicator function enforcing the ordering constraint. Posterior inference is performed via Gibbs sampling. At each iteration, the latent variable Z_i is updated by sampling from a truncated normal distribution

$$Z_i \mid \beta, \sigma^2, \mathbf{Z}_{-i}, R, X \sim \mathcal{N}_{[a_i,\, b_i]}(X_i'\beta, \sigma^2),$$

where $\mathcal{N}_{[a_i,b_i]}(\cdot)$ denotes a normal distribution truncated to the interval (a_i, b_i). The truncation bounds are determined by the observed ranks: the lower bound a_i is the maximum of Z_j over all j such that $R_j < R_i$, or $-\infty$ if no such j exists; similarly, the upper bound b_i is the minimum of Z_j over all j such that $R_j > R_i$, or $+\infty$ if none exists.

When additional information about category-specific interval boundaries $[L_k, U_k)$ for category k is available, these known metric bounds further refine the truncation limits for each observation i belonging to category k:

$$a_i = \max\left(\max_{j:R_j<R_i} Z_j,\, L_k\right), \quad b_i = \min\left(\min_{j:R_j>R_i} Z_j,\, U_k\right).$$

This allows the sampler to leverage both the ordinal ranking and available interval information, improving estimation efficiency and helping to identify the scale of Z when metric data are accessible.

After updating all latent variables $\{Z_i\}$, the regression coefficients β and residual variance σ^2 are sampled from their respective full conditional distributions. Given $\mathbf{Z}$, β has a multivariate normal posterior, and σ^2 has an inverse-gamma posterior, reflecting the Gaussian linear regression structure. Estimation of σ^2 is critical, as it governs the spread and uncertainty of the latent variable and influences inference on β.

Sampling from the truncated normal distribution is central to the rank likelihood ap-

proach: it restricts latent draws to the regions consistent with observed ordinal data and interval bounds. This enforces the ordinal structure implied by the data and ensures the posterior reflects all known constraints.

Although many applications assume $Z \sim \mathcal{N}(\mu, \sigma^2)$, the approach generalizes to other continuous latent distributions $F(\cdot)$. For example, if Z is lognormally distributed, the latent variable is reparameterized as $w_i = \log(Z_i)$ and sampled analogously on the log scale.

Unlike traditional interval regression, which relies on explicit knowledge of category boundaries and models β and σ^2 on a known metric scale, the Bayesian rank likelihood is robust to unknown or misspecified threshold locations and exploits only the ordinal information. When interval boundaries are known and incorporated, the rank likelihood method converges toward the interval regression approach, but it remains valid even when only ordinal data are observed.

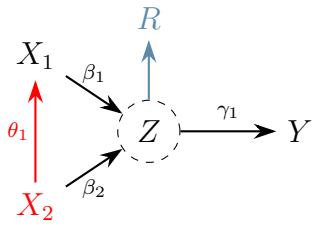
In summary, the Bayesian rank likelihood method is a flexible and semiparametric framework that accommodates a wide variety of continuous and ordinal data scenarios, as it only requires the outcome's rank-ordering to perform inference. It is particularly advantageous when threshold locations are unknown or modeling robustness is desired. Moreover, it often exhibits favorable finite-sample properties in terms of efficiency and bias compared to fully parametric alternatives (Adolph, 2011), making it a valuable tool for applied researchers working with coarsened, grouped, or ordered categorical data.

Second Monte Carlo Experiment: Assessing Latent Variable Methods

Building on the first set of experiments, where treating R as continuous or using midpoint imputation (Z_{mp}) was shown to yield bias—sometimes substantial—in estimating regression effects, I now turn to a direct comparison of midpoint imputation with two latent variable modeling strategies: interval regression ($\hat{Z}_{\text{inter}}$) and Bayesian rank likelihood ($\hat{Z}_{\text{rank}}$). The aim is to evaluate whether and under what conditions these (semi)parametric

approaches deliver less biased and more reliable predictions of Z.

Figure 5 illustrates the more complex data-generating process employed in this experiment. Here, the latent variable Z is modeled as a function of two predictors, X_1 and X_2, each with marginal effects (β_1, β_2). Importantly, X_1 itself is endogenous, generated as a function of X_2 with effect θ_1. The explicit causal structure and corresponding equations are shown in panels (a) and (b) of the figure.



(a) Causal Model of the Data-Generating Process.

$$
\begin{aligned}
X_2 &= \varepsilon_{X_2} \\
X_1 &= \theta_1 X_2 + \varepsilon_{X_1} \\
Z &= \beta_1 X_1 + \beta_2 X_2 + \varepsilon_Z \\
R &= g(Z) \\
Y &= \gamma_1 Z + \varepsilon_Y
\end{aligned}
$$

(b) Equations implied by the DAG.

Figure 5: **Data-Generating Process of Monte Carlo Experiments.**

As in the previous experiments, I keep the setup straightforward: error terms for X_1 and X_2 are drawn from standard normals, and all coefficients are set to unity except for $\theta_1 = -0.5$. The main distinction is the inclusion of multiple predictors, providing a more realistic scenario where Z is shaped by several observed covariates. Z itself is generated under three distributional regimes: normal, lognormal (with residual variance 0.2), and Pareto (with tail parameter $\alpha = 1.5$). The observed categorical variable R results from discretizing Z into k categories, where k ranges from 3 to 20. The experimental procedure unfolds in two stages:

1. **Stage One**: Predict the latent variable Z from R and covariates X_1 and X_2 using both interval regression and Bayesian rank likelihood. The specification of Z's distribution (normal, lognormal, or Pareto) is crucial at this step.
2. **Stage Two**: Stage Two: Quantify measurement error bias by regressing Y on each estimate of the latent variable: Z_{mp}, $\hat{Z}_{\text{inter}}$, and $\hat{Z}_{\text{rank}}$.

The key goal is to evaluate both the prediction error of $\hat{Z}$ and the bias in parameter

estimates $\boldsymbol{\pi}$ for both R and Y, where $\boldsymbol{\pi}$ refers to the vectors of coefficients ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$). The comparison relies on two performance metrics:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi} - \pi), \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{z}_i - z_i|$$

Here, bias measures the average deviation of estimated coefficients from their true values, while mean absolute error (MAE) captures the average prediction error for the latent variable. Estimator variability and coverage are illustrated using boxplots and jittered points across all k and distributional conditions.

To deepen the assessment, I further compare parametric methods under two scenarios[1]: (1) when the true covariate model for Z is known and used, and (2) when only an intercept is included (i.e., no covariates).

I hypothesize that both interval regression and Bayesian rank likelihood will outperform midpoint imputation in terms of both prediction error for $\hat{Z}$ and measurement error bias in γ_1, since they leverage information from interval boundaries and can flexibly incorporate model structure. However, including relevant covariates should further improve the performance of these methods—particularly when the predictors capture meaningful variation in Z—highlighting the value of thoughtful model specification in recovering latent continuous variables from categorical indicators.

Simulation results

The Monte Carlo simulations provide a systematic comparison of three approaches for recovering a latent continuous variable Z from ordered categorical data R: midpoint imputation (Z_{mp}), interval regression (Z_{inter}), and Bayesian rank likelihood (Z_{rank}). Results are

[1]Originally, I conducted several experiments to examine the impact of omitted variable bias and assess its sensitivity using partial R^2 analysis. For simplicity, however, I report only the intercept-only model here, as I recently discovered that its results are substantial and interesting enough to discuss preliminary conclusions without introducing additional complexity related to omitted variable bias. In future iterations of this project, I plan to include a formal sensitivity analysis for omitted variable bias. See the section on preliminary conclusions and future steps for more details.

presented for three types of latent distributions—normal, lognormal, and Pareto—while varying the number of categories k from 3 to 20. The performance of each method is evaluated by the measurement error bias in the regression coefficient γ_1 (visualized in Figure 6), as well as by bias in $\boldsymbol{\beta}$ and the mean absolute error (MAE) for $\hat{Z}$ (summarized in Table 1).

Normal Distribution:

When the latent variable Z is normally distributed, both interval regression and Bayesian rank likelihood demonstrate clear and consistent advantages over midpoint imputation. Across all k, these parametric methods yield essentially unbiased estimates of γ_1 and low prediction error for $\hat{Z}$—regardless of whether the model includes covariates or is intercept-only. In contrast, midpoint imputation displays notable downward bias at lower k (from -0.32 to -0.12 for $k = 3$ to $k = 6$), with bias shrinking and becoming negligible for $k \geq 13$. Moreover, interval regression and rank likelihood maintain superior coverage properties across the entire range of k.

Lognormal Distribution:

For a lognormally distributed Z, the two parametric methods again outperform midpoint imputation. Both interval regression and Bayesian rank likelihood produce only minor biases at small k—mainly in intercept-only models—but quickly become unbiased and efficient for $k \geq 7$, once the interval boundaries provide enough information to offset skewness in Z. In contrast, midpoint imputation remains biased across almost all k and exhibits poor probability coverage even at higher category counts, further emphasizing the robustness of the parametric approaches.

Pareto Distribution:

The Pareto case, with its heavy tails, poses a substantial challenge for all three methods. Midpoint imputation is the least reliable, with bias ranging from -1 to over 2 for $k = 10$, reflecting its sensitivity to sparsely populated upper categories and distributional extremes. Interval regression and Bayesian rank likelihood are also affected, showing persistent (typ-

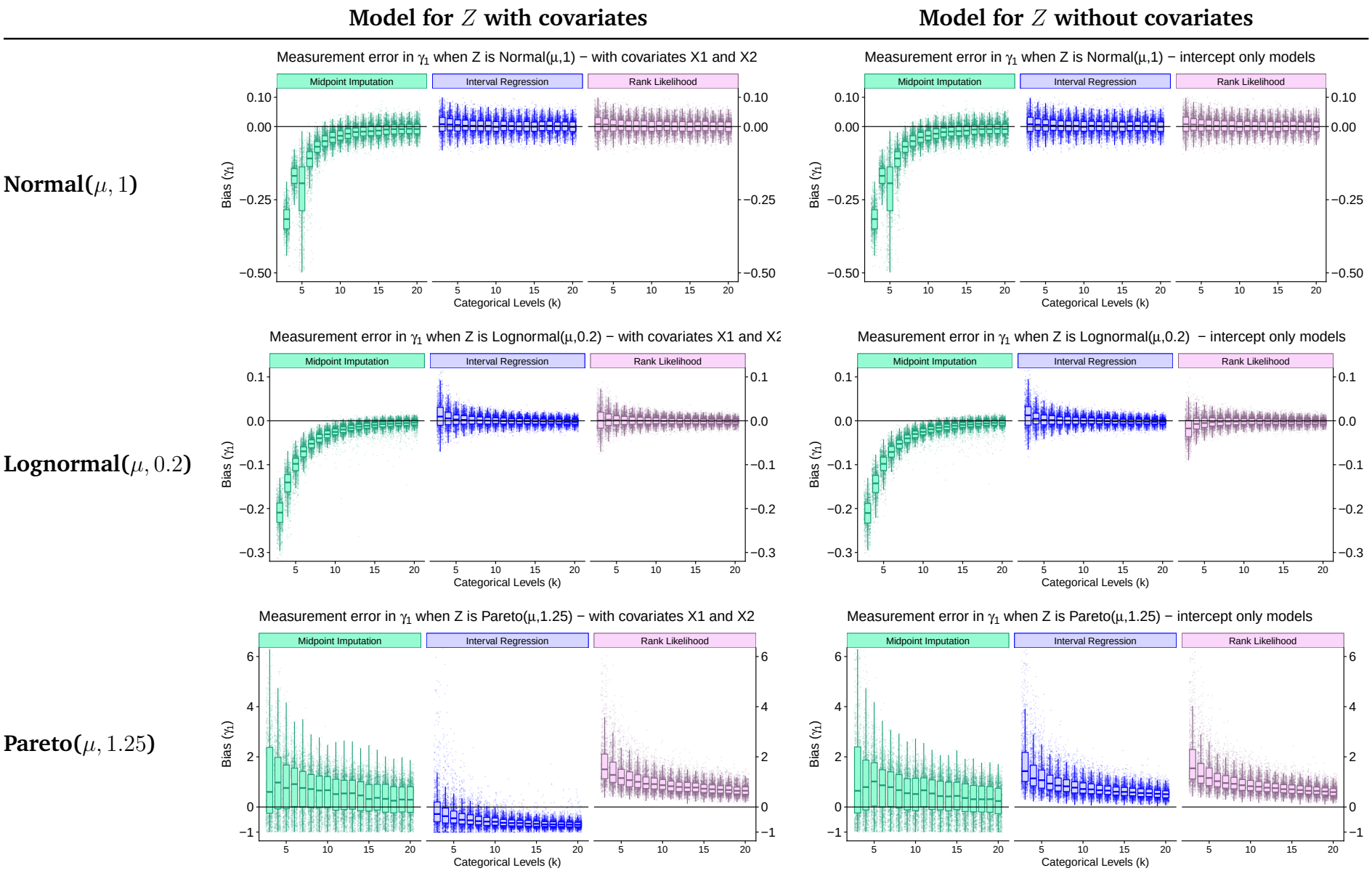


Figure 6: Bias of γ_1 under three latent-Z distributions, comparing models *with* vs. *without* covariates.

ically upward) bias in intercept-only models. Notably, including covariates in the interval regression does not eliminate this bias; the approach suffers from instability in prediction error, as shown by large and erratic MAE values in Table 1. This instability suggests that the interval regression function requires further refinement for extreme heavy-tailed data.

It is important to add that, for the Pareto scenario, the current implementation of Bayesian rank likelihood uses a lognormal target distribution for Z, since a stable routine for Pareto updating in the Gibbs sampler has not yet been developed. Consequently, the results for Bayesian rank likelihood in this case should be interpreted with caution, as they do not reflect a true Pareto-based update.

Overall, these simulation results indicate that interval regression and Bayesian rank likelihood provide accurate and efficient recovery of Z and γ_1 in both normal and lognormal settings, consistently outperforming midpoint imputation—especially as the number of categories increases. However, for extremely heavy-tailed distributions like the Pareto, all methods exhibit limitations: midpoint imputation becomes highly variable, the Bayesian rank likelihood is limited by its current implementation, and interval regression can yield unstable estimates. Thus, researchers should be particularly cautious when applying these methods to data with pronounced skewness or heavy tails, and should avoid treating ordered categorical variables as continuous predictors without appropriate modeling and diagnostic checks.

Preliminary Conclusions and Future Steps

Although this working paper is still under development, the evidence presented so far allows several key preliminary conclusions.

1. As demonstrated by the Monte Carlo simulations in Figure 2, ordered categorical variables should not be modeled as continuous predictors if we view them as discrete realizations of an underlying latent variable, regardless of whether or not they possess an intrinsic metric. Instead, these variables should be included as sets of

indicator (dummy) variables. Treating them as continuous results in biased estimates, with coefficients exhibiting increasing attenuation towards zero as the number of categories k grows. The findings here provide a clear warning against the widespread—but misguided—practice, often cited in the psychometrics literature, of treating ordinal variables as continuous when $k > 5$.

2. When the ordered categorical variable possesses a known metric and it is important to recover its underlying values—such as when harmonizing or adjusting income data for inflation across survey years (Hout, 2004)—the interval boundaries can be used to approximate the latent variable. The simulations here show that, at least for large samples, both interval regression and Bayesian rank likelihood allow Z to be recovered accurately and efficiently from these bounds. Even under normality, midpoint imputation remains relatively biased and inefficient, and is therefore not recommended.
3. Notably, both parametric models can predict the latent variable Z with minimal prediction error and measurement error bias—even without covariates—when compared to midpoint imputation, at least in the normal and lognormal scenarios. This result is encouraging, as it reduces the need for complex model specification; researchers need only make distributional assumptions about Z. It is also worth noting that even midpoint imputation is not free from distributional assumptions, since it requires a choice of extrapolation formula for top-coded categories.

There are, however, several important directions for future research. First, the current routines for fitting the parametric models are not stable when estimating and predicting a latent variable with a Pareto distribution. For Bayesian rank likelihood, a promising improvement may be to replace the Gibbs sampler with Hamiltonian Monte Carlo, which could enhance both efficiency and stability.

It is also somewhat unexpected that including covariates in the parametric models does not substantially reduce prediction error or increase the efficiency of latent variable esti-

		Model with covariates							Model without covariates		
Dist.	**cat_level**	**b1_inter**	**b2_inter**	**b1_rank**	**b2_rank**	**mae_inter**	**mae_rank**	**mae_mp**	**mae_inter**	**mae_rank**	**mae_mp**
Normal	4	0.992	0.992	0.990	0.990	0.395	0.396	0.577	0.427	0.427	0.577
	5	0.996	0.996	0.994	0.994	0.333	0.333	1.310	0.352	0.352	1.310
	6	0.998	0.996	0.996	0.994	0.286	0.286	0.344	0.297	0.297	0.344
	7	0.998	0.998	0.996	0.996	0.250	0.250	0.277	0.258	0.258	0.277
	8	0.996	0.997	0.994	0.995	0.222	0.222	0.238	0.227	0.227	0.238
	9	0.997	0.997	0.995	0.995	0.199	0.199	0.210	0.203	0.203	0.210
	10	0.999	0.998	0.997	0.996	0.181	0.181	0.189	0.184	0.184	0.189
	11	1.000	0.998	0.998	0.996	0.165	0.165	0.171	0.167	0.167	0.171
	12	0.999	0.998	0.997	0.996	0.151	0.151	0.156	0.153	0.153	0.156
	13	0.999	0.999	0.997	0.997	0.140	0.140	0.144	0.141	0.141	0.144
	14	0.997	0.998	0.995	0.996	0.130	0.130	0.133	0.131	0.131	0.133
	15	0.998	0.998	0.996	0.996	0.122	0.122	0.124	0.123	0.123	0.124
	16	0.998	0.998	0.996	0.996	0.114	0.114	0.116	0.115	0.115	0.116
	17	0.998	0.998	0.996	0.996	0.108	0.108	0.110	0.108	0.108	0.110
	18	1.000	0.999	0.998	0.997	0.101	0.101	0.103	0.102	0.102	0.103
	19	1.000	0.999	0.998	0.997	0.0962	0.0962	0.0977	0.0967	0.0968	0.0977
	20	0.999	0.999	0.997	0.997	0.0913	0.0913	0.0926	0.0918	0.0918	0.0926
Lognormal	3	0.983	0.983	0.922	0.922	2.420	2.440	3.390	2.780	2.800	3.400
	4	0.990	0.991	0.931	0.932	2.010	2.020	2.590	2.210	2.220	2.600
	5	0.994	0.994	0.938	0.938	1.710	1.720	2.070	1.820	1.830	2.060
	6	0.996	0.997	0.942	0.943	1.480	1.480	1.700	1.560	1.560	1.710
	7	0.996	0.995	0.944	0.943	1.300	1.300	1.450	1.350	1.350	1.460
	8	0.995	0.996	0.944	0.945	1.150	1.160	1.270	1.200	1.200	1.270
	9	0.997	0.996	0.946	0.946	1.040	1.040	1.120	1.060	1.060	1.110
	10	0.997	0.998	0.947	0.948	0.942	0.943	1.000	0.964	0.965	1.000
	11	0.996	0.999	0.947	0.949	0.859	0.860	0.905	0.878	0.879	0.908
	12	0.998	0.997	0.949	0.948	0.788	0.788	0.825	0.803	0.804	0.827
	13	0.995	0.995	0.947	0.946	0.729	0.729	0.758	0.743	0.743	0.762
	14	0.998	0.998	0.949	0.949	0.681	0.681	0.705	0.687	0.687	0.702
	15	0.999	0.999	0.950	0.950	0.635	0.636	0.655	0.639	0.640	0.652
	16	0.999	0.999	0.951	0.950	0.596	0.596	0.612	0.599	0.599	0.609
	17	0.998	0.998	0.949	0.949	0.560	0.560	0.574	0.565	0.565	0.574
	18	0.998	0.998	0.949	0.950	0.528	0.528	0.540	0.532	0.532	0.539
	19	1.000	1.000	0.953	0.951	0.500	0.500	0.510	0.504	0.504	0.509
	20	0.998	0.998	0.950	0.950	0.473	0.473	0.482	0.478	0.478	0.483
Pareto	3	0.891	0.893	0.574	0.575	894104	9.26	30.9	12.20	10.40	31.60
	4	0.906	0.909	0.618	0.621	157551	7.80	21.1	8.78	8.00	20.10
	5	0.927	0.919	0.682	0.676	39730	6.56	15.7	7.50	7.07	21.90
	6	0.946	0.941	0.731	0.726	19486	5.95	14.2	6.32	6.14	17.80
	7	0.953	0.949	0.777	0.774	4221	5.43	12.9	5.78	5.71	10.20
	8	0.957	0.957	0.812	0.813	132	4.96	13.5	5.25	5.25	10.10
	9	0.967	0.967	0.846	0.847	1067	4.72	10.5	4.79	4.83	13.30
	10	0.971	0.972	0.875	0.879	1758	4.40	10.9	4.49	4.55	11.90
	11	0.978	0.979	0.903	0.904	493	4.14	22.0	4.26	4.32	8.24
	12	0.983	0.983	0.930	0.933	2413	3.92	11.1	3.93	4.00	7.70
	13	0.983	0.984	0.943	0.946	14838985	3.74	10.6	3.75	3.82	9.70
	14	0.986	0.985	0.956	0.953	9612044	3.62	7.41	3.54	3.61	9.79
	15	0.989	0.990	0.970	0.970	90877181	3.42	14.9	3.43	3.50	14.90
	16	0.992	0.993	0.980	0.978	5021661254	3.28	18.6	3.28	3.34	24.40
	17	0.991	0.992	0.987	0.989	91.0	3.17	7.54	3.09	3.15	8.72
	18	0.991	0.994	0.988	0.992	945457778	3.02	5.69	2.98	3.04	14.90
	19	0.996	0.996	0.995	0.996	458738569	2.89	10.9	2.84	2.89	10.60
	20	0.996	0.996	1.000	1.000	3413998703	2.84	21.5	2.70	2.75	27.00

Table 1: Summary of β_1 and β_2 bias and mean absolute error (MAE) across three conditions latent Z distributions.

mation. This may be due to the simplicity of the current data-generating processes, which feature only weak effects. In more realistic settings—with richer sets of predictors, particularly those that can inform the distribution's tails, as with the Pareto case—the benefits of covariate modeling could become more pronounced.

Another important motivation for incorporating covariates is to extend the Bayesian rank likelihood model for simultaneous treatment of measurement error and missing data—a common problem in income questions on surveys. Such extensions could help address nonresponse bias in addition to measurement error.

Additionally, I am considering adopting a more flexible parameterization of the Pareto distribution. While the current routines are based on the Pareto Type I distribution, previous research (Charpentier and Flachaire, 2022) suggests that a Generalized Pareto distribution may be less sensitive to threshold choice and better suited for reliably estimating heavy tails.

Moreover, both the rank likelihood and interval regression models require the analyst to specify a distributional form for the latent variable—typically normal, lognormal, or Pareto. For the Bayesian rank likelihood, an appealing extension would be to use nonparametric methods, such as empirical cumulative distribution functions or Dirichlet processes, to estimate the latent distribution directly from the data.

It is also worth noting that in these Monte Carlo experiments, some design choices have been held constant—such as equal-interval (rather than quantile-based) discretization and a fixed sample size of 1,000. Future work should explore how these results change with alternative discretization schemes and smaller samples.

In the next stages of this project, I plan to address these remaining challenges. If the findings remain robust, I will also replicate published studies that rely on midpoint imputation, to assess whether the inefficiencies and biases uncovered here are large enough to alter substantive conclusions in applied research.

References

Adolph, Christopher. 2011. "Small Sample Properties of Partially-Observed Rank Data Estimators.".

Bound, John, Charles Brown and Nancy Mathiowetz. 2001. *Measurement Error in Survey Data*. pp. 3705–3843.

Charpentier, Arthur and Emmanuel Flachaire. 2022. "Pareto models for top incomes and wealth." *The Journal of Economic Inequality* 20:1–25.

Desai, Sreedhari D., Dolly Chugh and Arthur P. Brief. 2014. "The Implications of Marriage Structure for Men's Workplace Attitudes, Beliefs, and Behaviors toward Women." *Administrative Science Quarterly* 59:330–365.

Hoff, Peter. 2008. "Rank Likelihood Estimation for Continuous and Discrete Data." *ISBA Bulletin* 14.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods*. Springer New York.

Hout, Michael. 2004. "Getting the Most Out of the GSS Income Measures.".

Lustig, Nora. 2020. "The "Missing Rich" In Household Surveys: Causes and Correction Approaches.".

Nakagawa, Yuki and Takashi Sozu. 2024. "Improvement of Midpoint Imputation for Estimation of Median Survival Time for Interval-Censored Time-to-Event Data." *Therapeutic Innovation & Regulatory Science* 58:721–729.

Pampel, Fred C. and Lori M. Hunter. 2012. "Cohort Change, Diffusion, and Support for Environmental Spending in the United States." *American Journal of Sociology* 118:420–448.

Pettitt, A. N. 1982. "Inference for the Linear Model Using a Likelihood Based on Ranks." *Journal of the Royal Statistical Society: Series B (Methodological)* 44:234–243.

Rhemtulla, Mijke, Patricia É. Brosseau-Liard and Victoria Savalei. 2012. "When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions." *Psychological Methods* 17:354–373.

Riegle-Crumb, Catherine and Eric Grodsky. 2010. "Racial-Ethnic Differences at the Intersection of Math Course-taking and Achievement." *Sociology of Education* 83:248–270.

Robitzsch, Alexander. 2020. "Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods." *Frontiers in Education* 5.

Rueda, David and Daniel Stegmueller. 2019. *Who Wants What?* Cambridge University Press.

Yang, Haiyang and Jingjing Ma. 2020. "How an Epidemic Outbreak Impacts Happiness: Factors that Worsen (vs. Protect) Emotional Well-being during the Coronavirus Pandemic." *Psychiatry Research* 289:113045.