Latent Variable Methods to Address Measurement Error from Ordered Categorical Data

Ramses Llobet, Ph.D. Candidate

Political Science, University of Washington

Survey data – Germany (ESS 5)

2. Income deciles

	Source data	Adjusted data		Income deciles
01	less than 945 € per month		J	0-220 (week) 0-945 (month) 0-11340 (year)
02	946 €-1285 €		R	221-300 (week) 946-1290 (month) 11341-15420 (year)
03	1286 € - 1579 €		C	301-360 (week) 1291-1580 (month) 15421-18950 (year)
04	1580 € - 1886 €		M	361-430 (week) 1581-1890 (month) 18951-22630 (year)
05	1887 €- 2208 €		F	431-510 (week) 1891-2210 (month) 22631-26500 (year)
06	2209 € - 2558 €		S	511-590 (week) 2211-2560 (month) 26501-30700 (year)
07	2559 € - 2976 €		K	591-690 (week) 2561-2980 (month) 30701-35710 (year)
08	2977 € - 3532 €		P	691-820 (week) 2981-3530 (month) 35711-42380 (year)
09	3533 €- 4481 €		D	821-1030 (week) 3531-4480 (month) 42381-53770 (year)
10	4482 € or more		H	1031 or more (week) 4481 or more (month) 53771 or more (year)

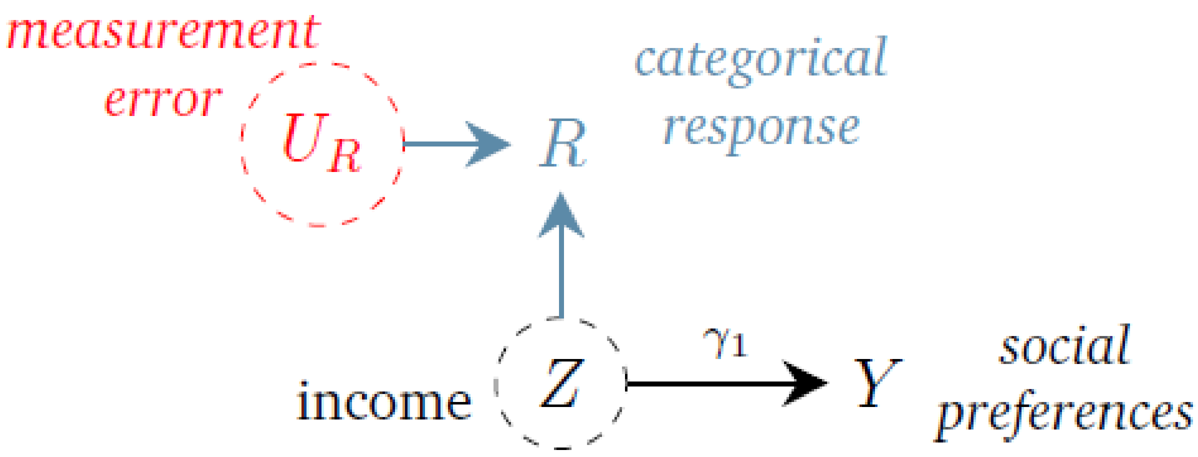
Income decile table refers to:

Ordered Categories and Measurement Error

income $Z \xrightarrow{\gamma_1} Y$ *social preferences*

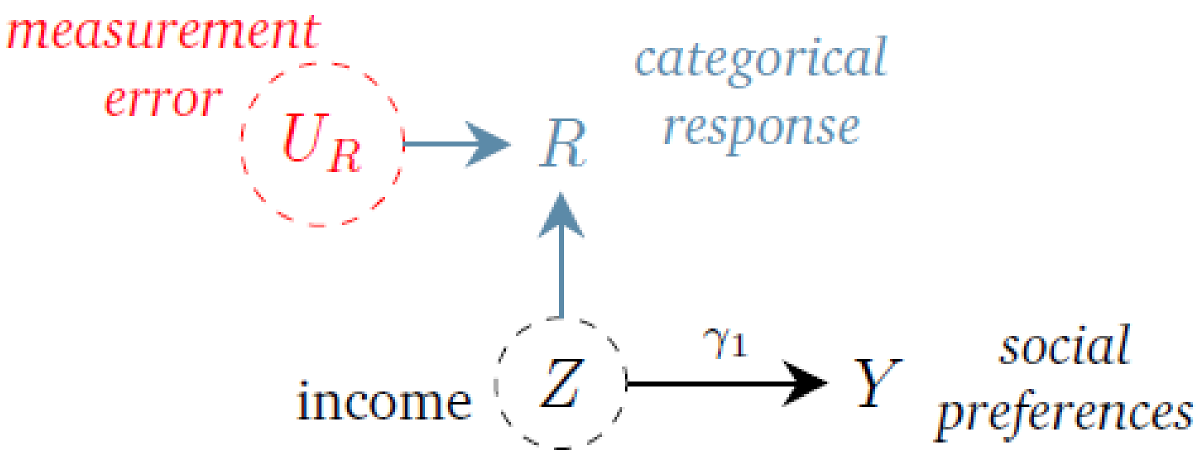
(a) No measurement error.

Ordered Categories and Measurement Error



(b) Measurement error.

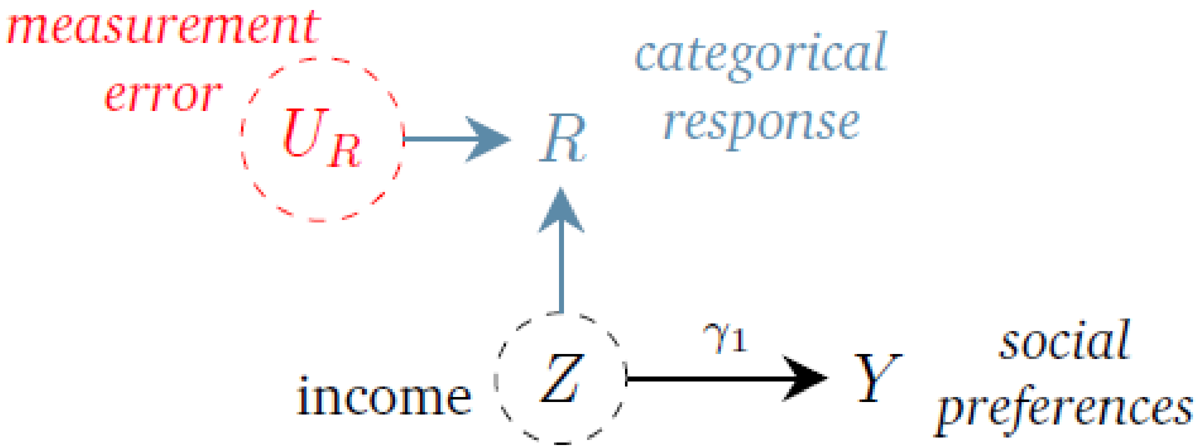
Ordered Categories and Measurement Error



(b) Measurement error.

1. **Misclassification**
   i. Item non-response
   ii. Underreporting

Ordered Categories and Measurement Error



(b) Measurement error.

1. **Misclassification**
   i. Item non-response
   ii. Underreporting
2. **Censoring**
   i. Interval-censoring
   ii. Top-coding problem

Measurement Error: Midpoint Imputation

ID	True income Z	Category R	Lower L	Upper U
57	55,300	5	40,000	60,000
12	3,900	1	0	5,000

Measurement Error: Midpoint Imputation

ID	True income Z	Category R	Lower L	Upper U
57	55,300	5	40,000	60,000
12	3,900	1	0	5,000
83	72,800	6	60,000	80,000
145	9,600	2	5,000	10,000
230	277,000	10	200,000	$+\infty$
34	16,200	3	10,000	25,000

Measurement Error: Midpoint Imputation

ID	True income Z	Category R	Lower L	Upper U	Midpoint
57	55,300	5	40,000	60,000	50,000
12	3,900	1	0	5,000	2,500
83	72,800	6	60,000	80,000	70,000
145	9,600	2	5,000	10,000	7,500
230	277,000	10	200,000	$+\infty$	–
34	16,200	3	10,000	25,000	17,500

First Monte Carlo Experiment

1. Fit R as a numeric continuous. $Y = \gamma_0 + \gamma_1 R + \epsilon$

2. Fit Midpoint regressor. $Y = \gamma_0 + \gamma_1 Z_{MP} + \epsilon$

ID	True income Z	Category R	Lower L	Upper U	Midpoint
57	55,300	5	40,000	60,000	50,000
12	3,900	1	0	5,000	2,500
83	72,800	6	60,000	80,000	70,000
145	9,600	2	5,000	10,000	7,500
230	277,000	10	200,000	$+\infty$	–
34	16,200	3	10,000	25,000	17,500

First Monte Carlo Experiment

First Monte Carlo Experiment

- We set: $\gamma_1 = 1$

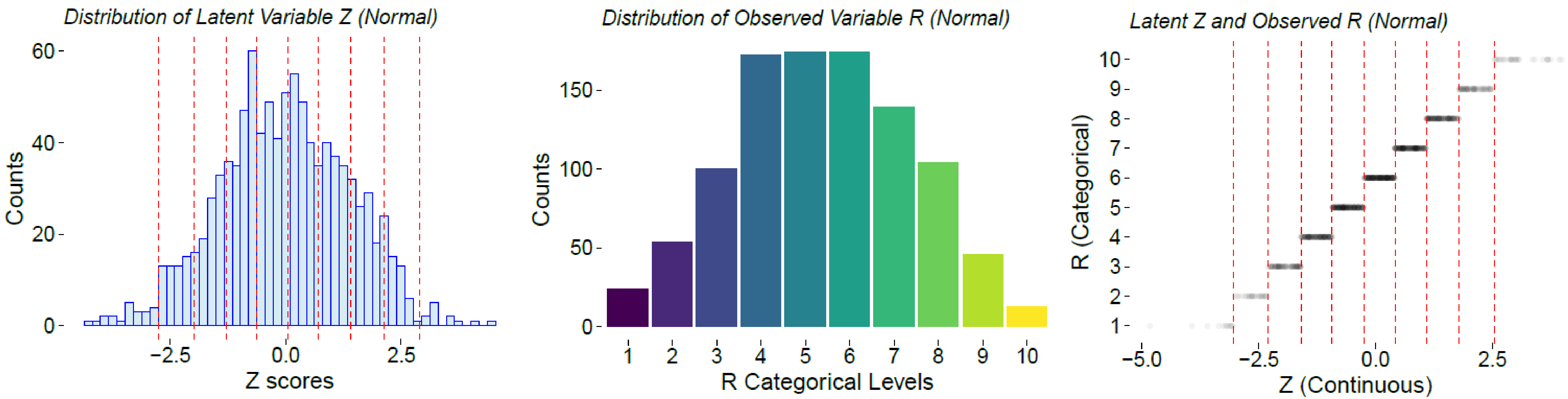
First Monte Carlo Experiment

- We set: $\gamma_1 = 1$
- Three distributional assumptions on **Z**:
  - *Normal*(0,1)
  - *Lognormal* (mu, 0.2)
  - *Pareto*(mu,1.5)

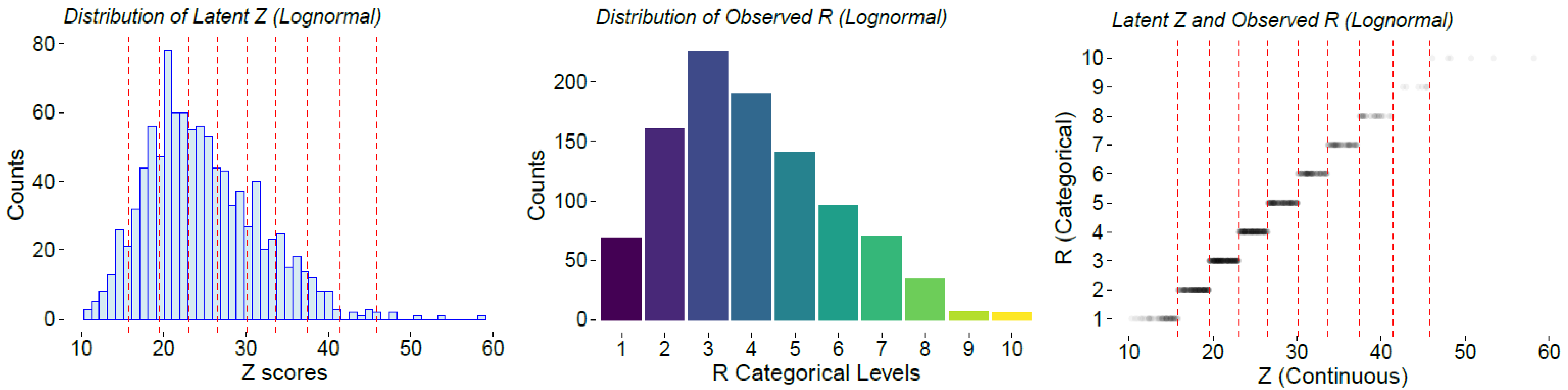
First Monte Carlo Experiment

- We set: $\gamma_1 = 1$
- Three distributional assumptions on **Z**:
  - *Normal*(0,1)
  - *Lognormal* (mu, 0.2)
  - *Pareto*(mu,1.5)
- From 3 to 30 categorical levels (*k*)
  - 1,000 simulations (*sims*)
  - In every sims, a sample of 1,000 observations (n)

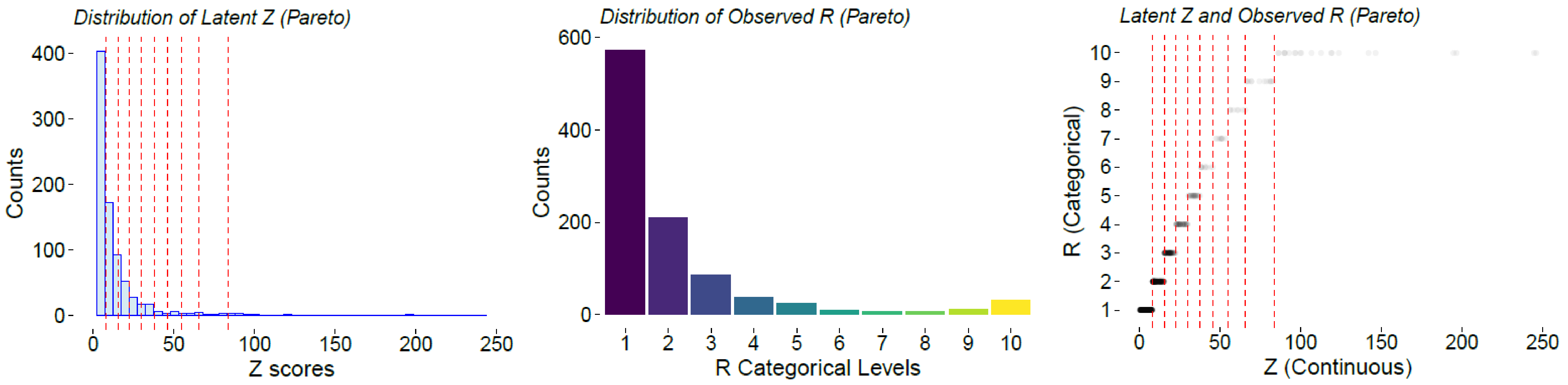
Sampling from Normal(0,1), *k*=10



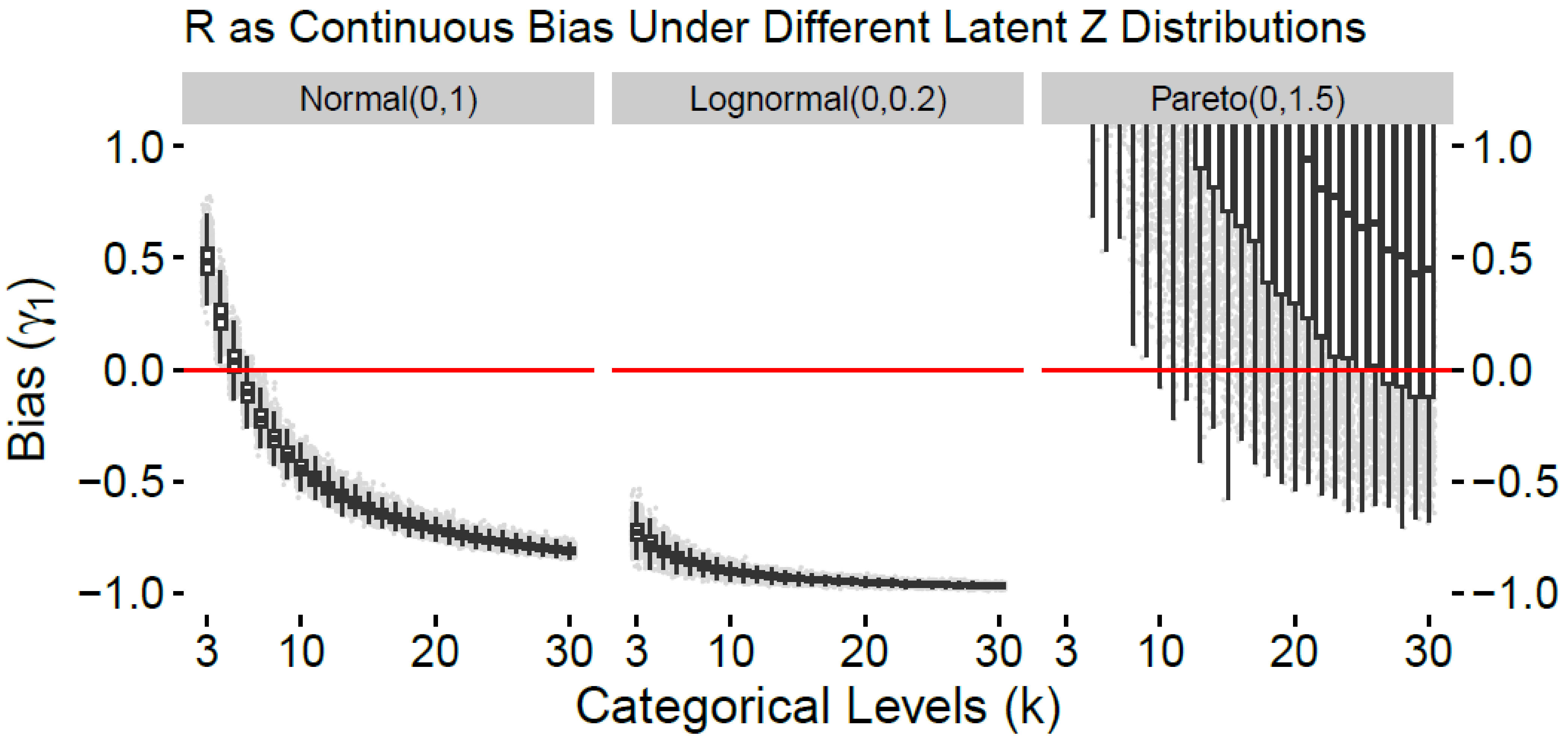
Sampling from Lognormal(24,0.2), *k*=10



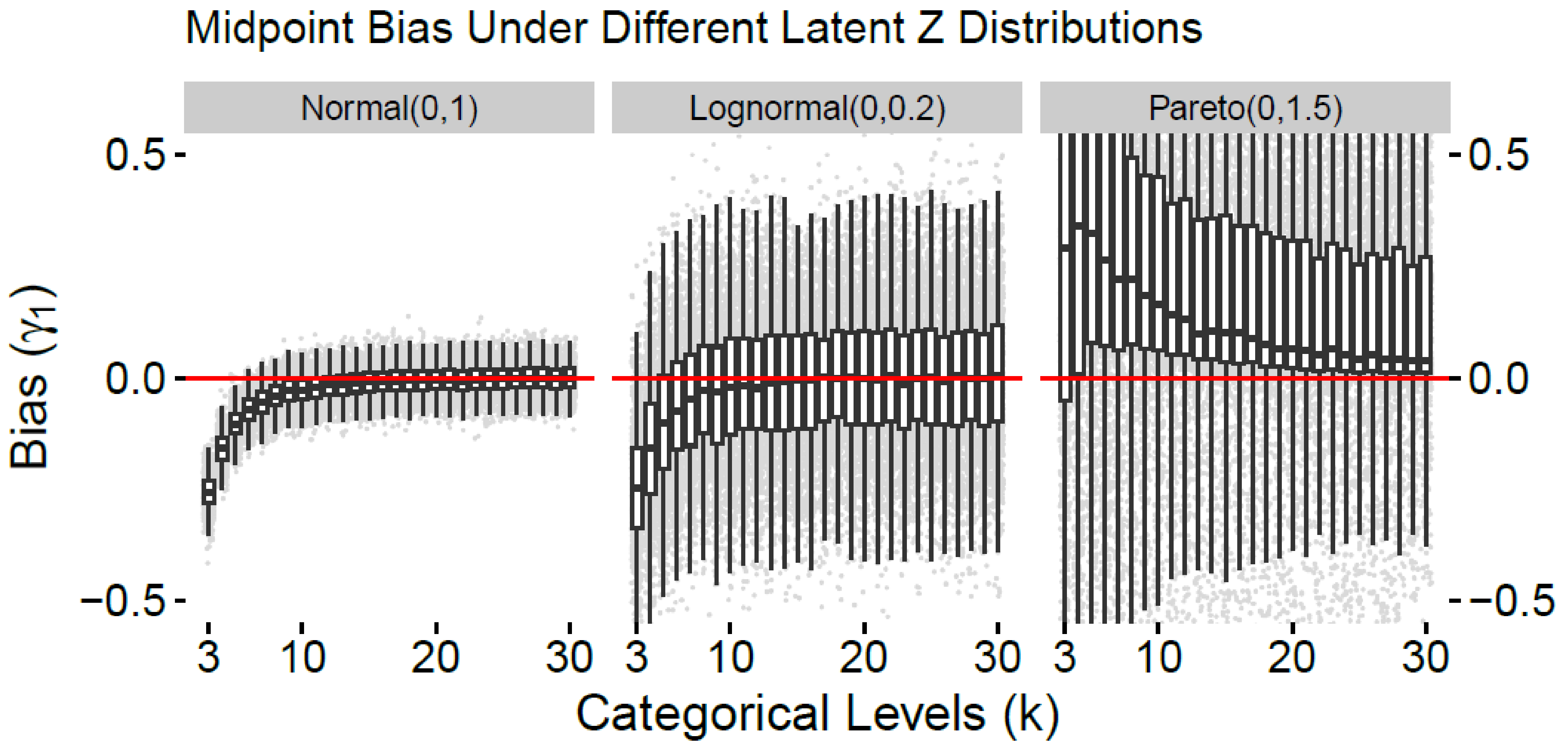
Sampling from Pareto(24,1.5), k=10



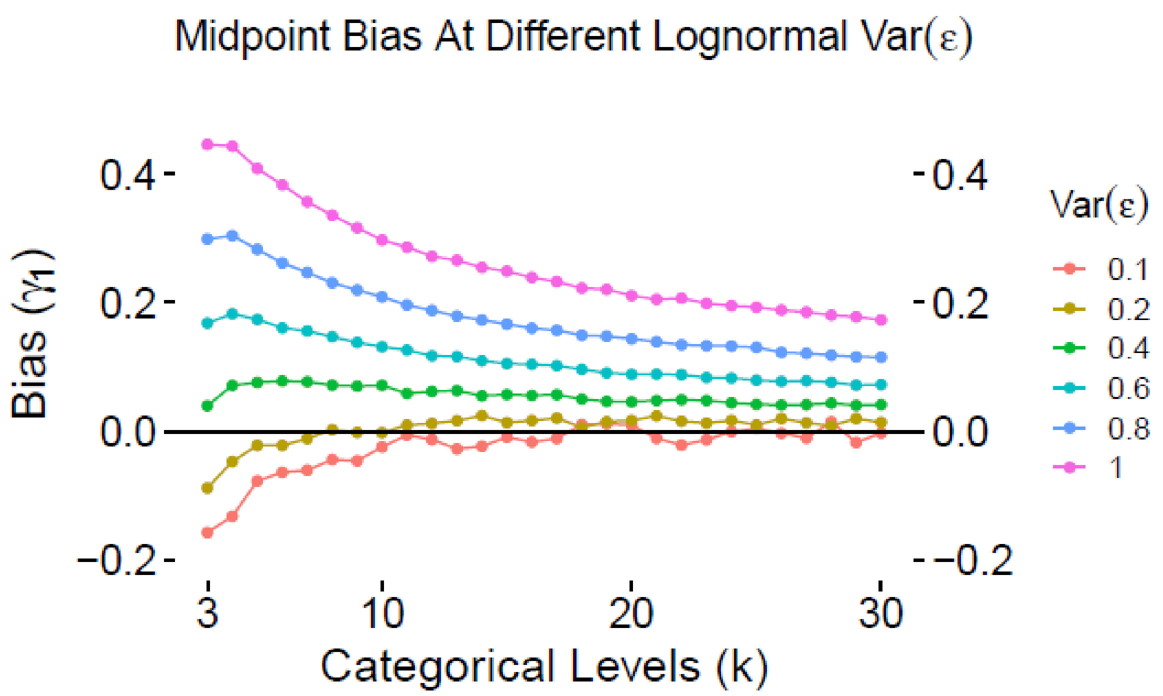
Results: R as *Continuous*

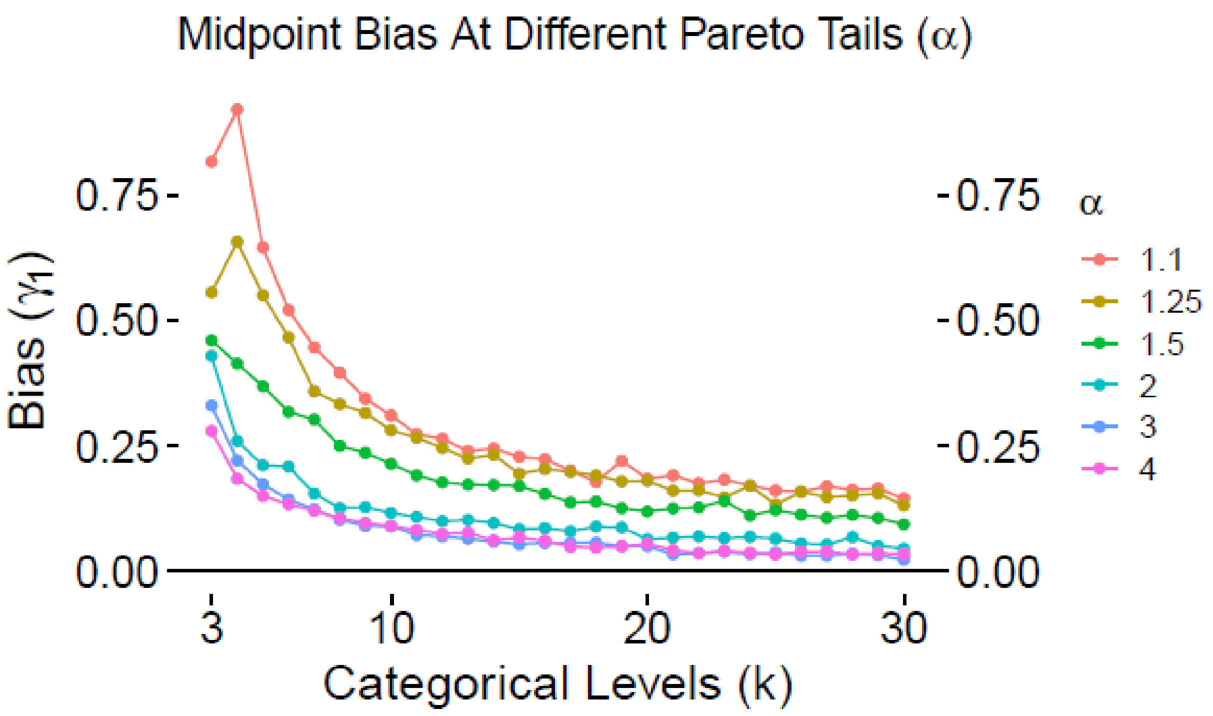


Results: Midpoint Imputation



Results Sensitivity at Different Scenarios





Results: Conclusions

Results: Conclusions

- When R is fitted as continuous:
  - **Bias** and **Inconsistent** Estimates.
  - As we increase *k*, higher attrition towards 0.

Results: Conclusions

- When R is fitted as continuous:
  - **Bias** and **Inconsistent** Estimates.
  - As we increase *k*, higher attrition towards 0.
- When Z is midpoint imputed:
  - **Biased** estimates, but it **decreases** as *k* **increases**.
  - Skewness and variance of Z makes bias unpredictable.

Results: Conclusions

- When R is fitted as continuous:
  - **Bias** and **Inconsistent** Estimates.
  - As we increase *k*, higher attrition towards 0.
- When Z is midpoint imputed:
  - **Biased** estimates, but it **decreases** as *k* **increases**.
  - Skewness and variance of Z makes bias unpredictable.

Can we do better? Maybe model Z? **Two candidates**

Interval Regression

- **Interval regression** is a special case of ordinal regression.
  - The **metric** and **interval bounds** of the cutpoints are known.

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

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

Bayesian Rank Likelihood

- A semiparametric approach to inference that uses the ordering of the outcome to make inference

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

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

- Inference is via **Gibbs sampling**:
  - a prior distribution must be provided to Z
  - truncated continuous distribution is used to update intervals

Second Monte Carlo Experiment

Second Monte Carlo Experiment

- First experiment, but now we make Z endogenous to a model.

Second Monte Carlo Experiment

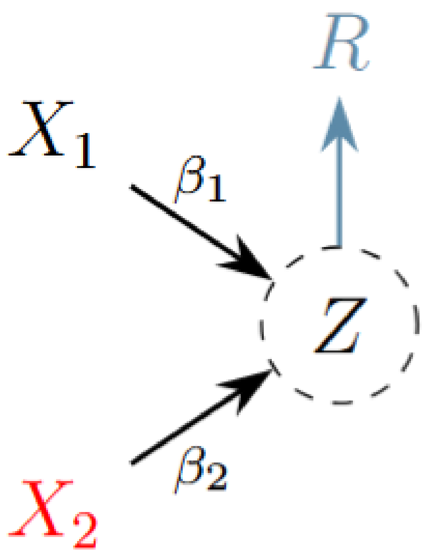
- First experiment, but now we make Z endogenous to a model.
  - Z is a function of **two variables**: X1 and X2
  - where beta_1 = beta_2 = 1

Second Monte Carlo Experiment

- First experiment, but now we make Z endogenous to a model.
  - Z is a function of **two variables**: X1 and X2
  - where beta_1 = beta_2 = 1

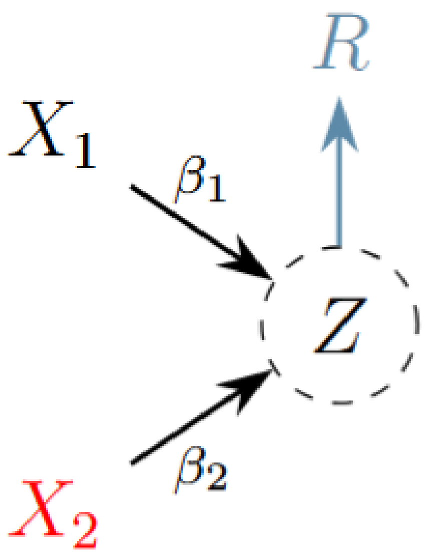
- From 3 to 20 categorical levels (*k*)

First Stage Regression



$$
\begin{aligned}
Z &= \beta_1 X_1 + \beta_2 X_2 + \varepsilon_Z,\\
R &= g(Z;\, \tau_k),
\end{aligned}
$$

First Stage Regression

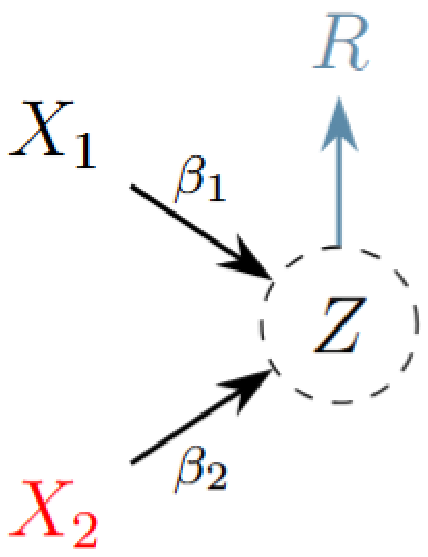


$$Z = \beta_1 X_1 + \beta_2 X_2 + \varepsilon_Z,$$
$$R = g(Z;\, \tau_k),$$

This model is estimated with:

1. **Interval regression** or
2. **Bayesian Rank Likelihood**

First Stage Regression



$$Z = \beta_1 X_1 + \beta_2 X_2 + \varepsilon_Z,$$
$$R = g(Z;\, \tau_k),$$

This model is estimated with:

1. **Interval regression** or
2. **Bayesian Rank Likelihood**

Two scenarios:

1. With **full model** (Xs)
2. With no variables, **intercept-only**

Second Stage Regression

$$Z^{\star} \xrightarrow{\gamma_1} Y \qquad\qquad Y = \gamma_1 Z^{\star} + \varepsilon_Y,$$

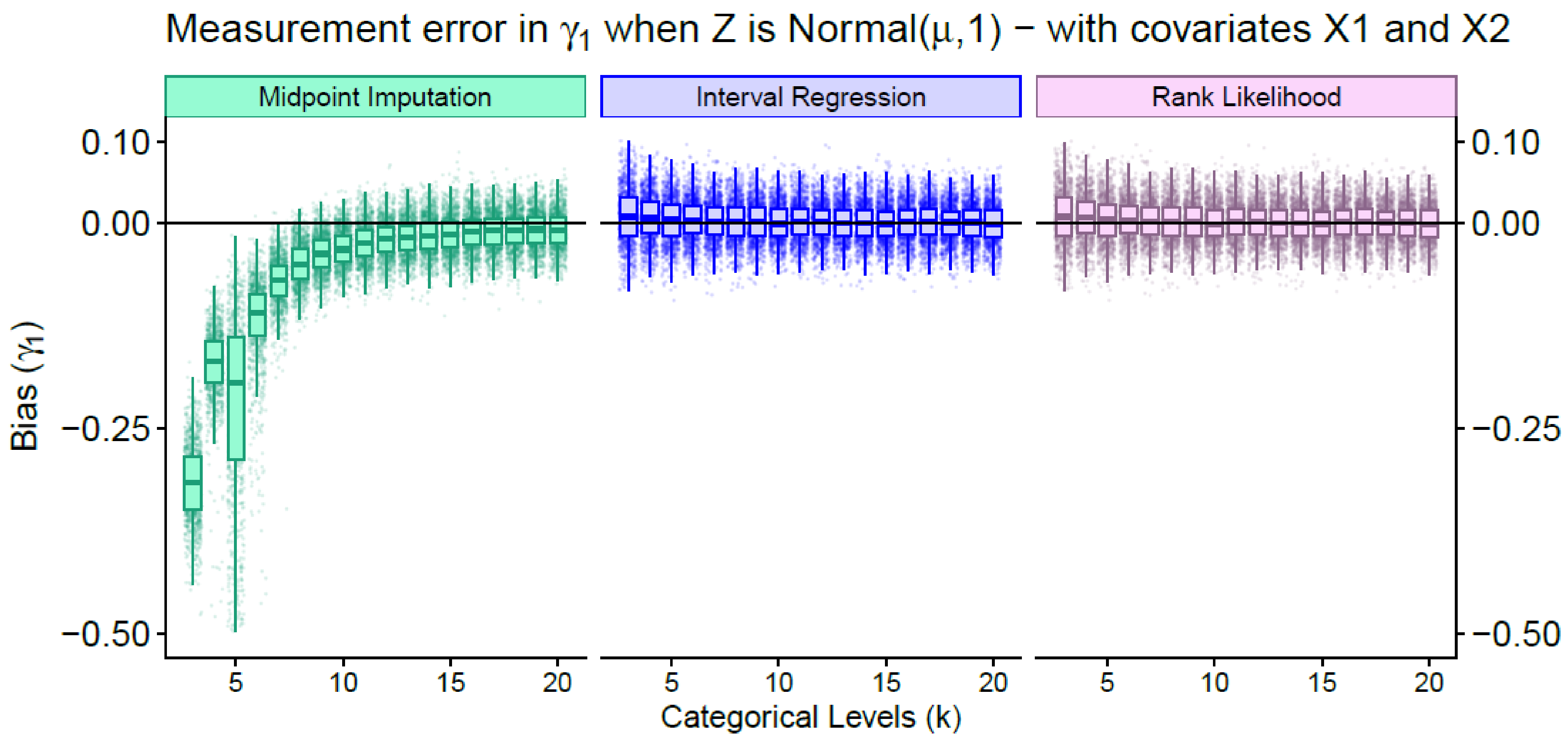
- We predict Z from the First Stage.

Second Stage Regression

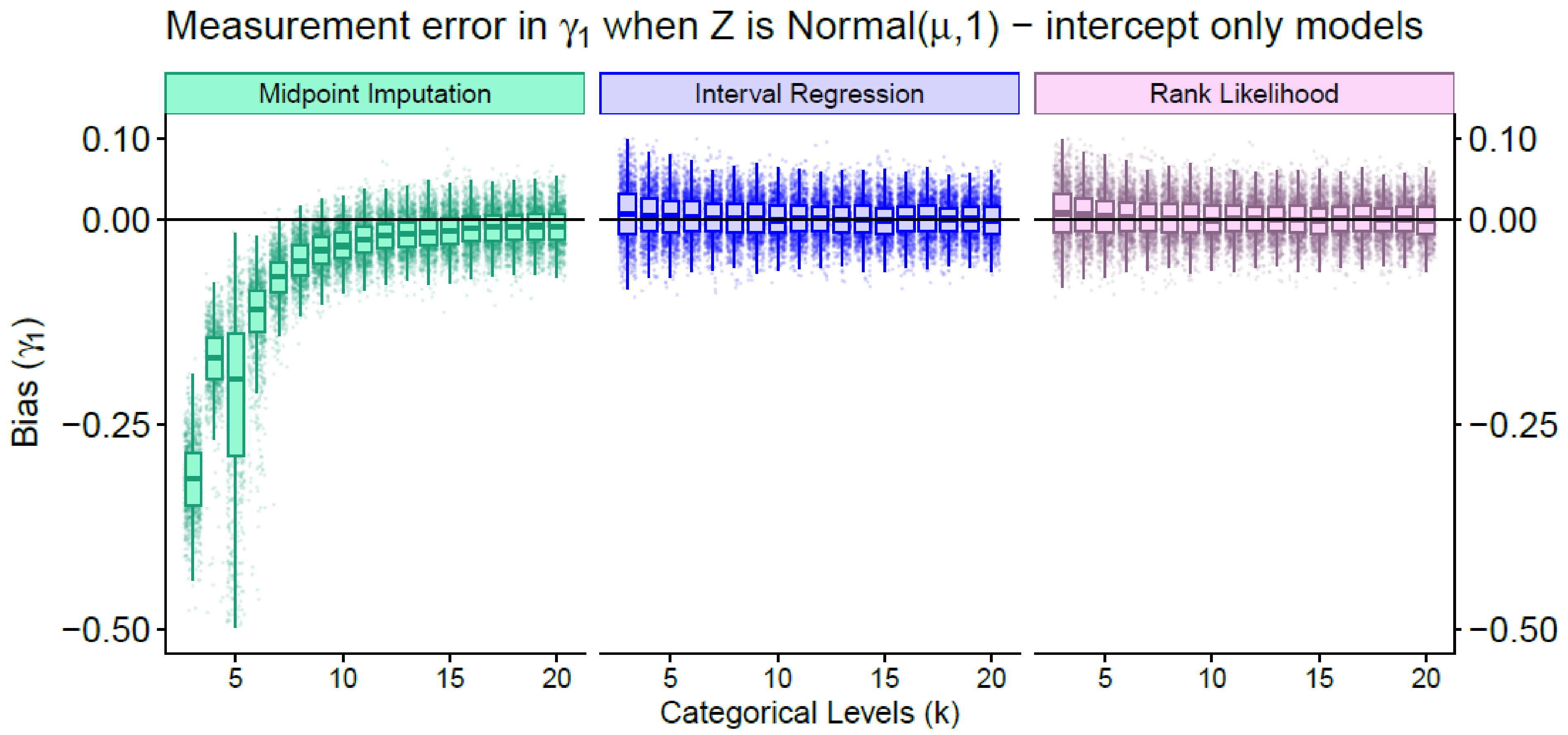
$$Z^\star \xrightarrow{\gamma_1} Y \qquad\qquad Y = \gamma_1 Z^\star + \varepsilon_Y,$$

- We predict Z from the First Stage.
- Then we evaluate gamma_1 under **three predictions**:
    1. Z from **Interval Regression**.
    2. Z from **Bayesian Rank Likelihood**.
    3. Z from **Midpoint**.

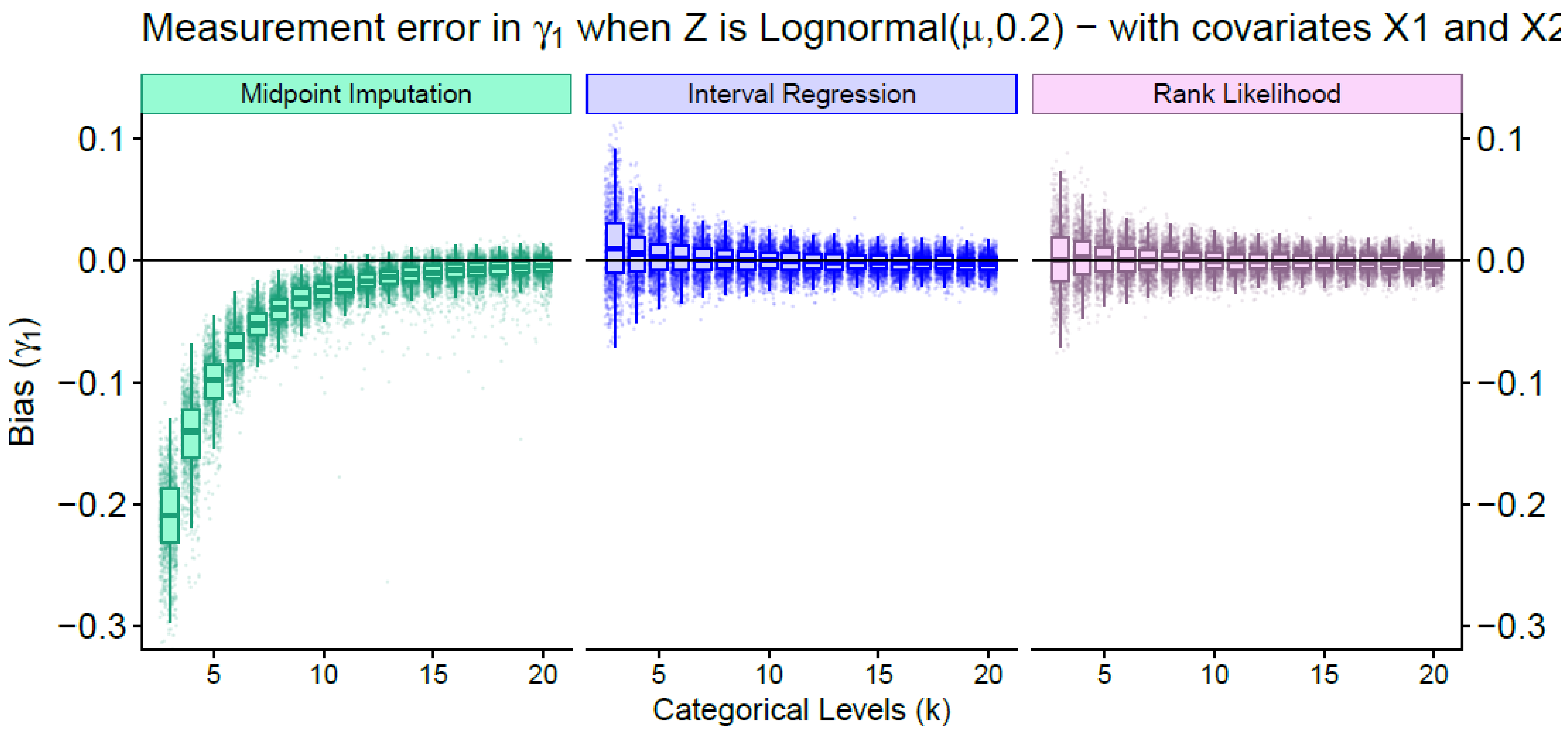
MC Results – Normality (with covariates)



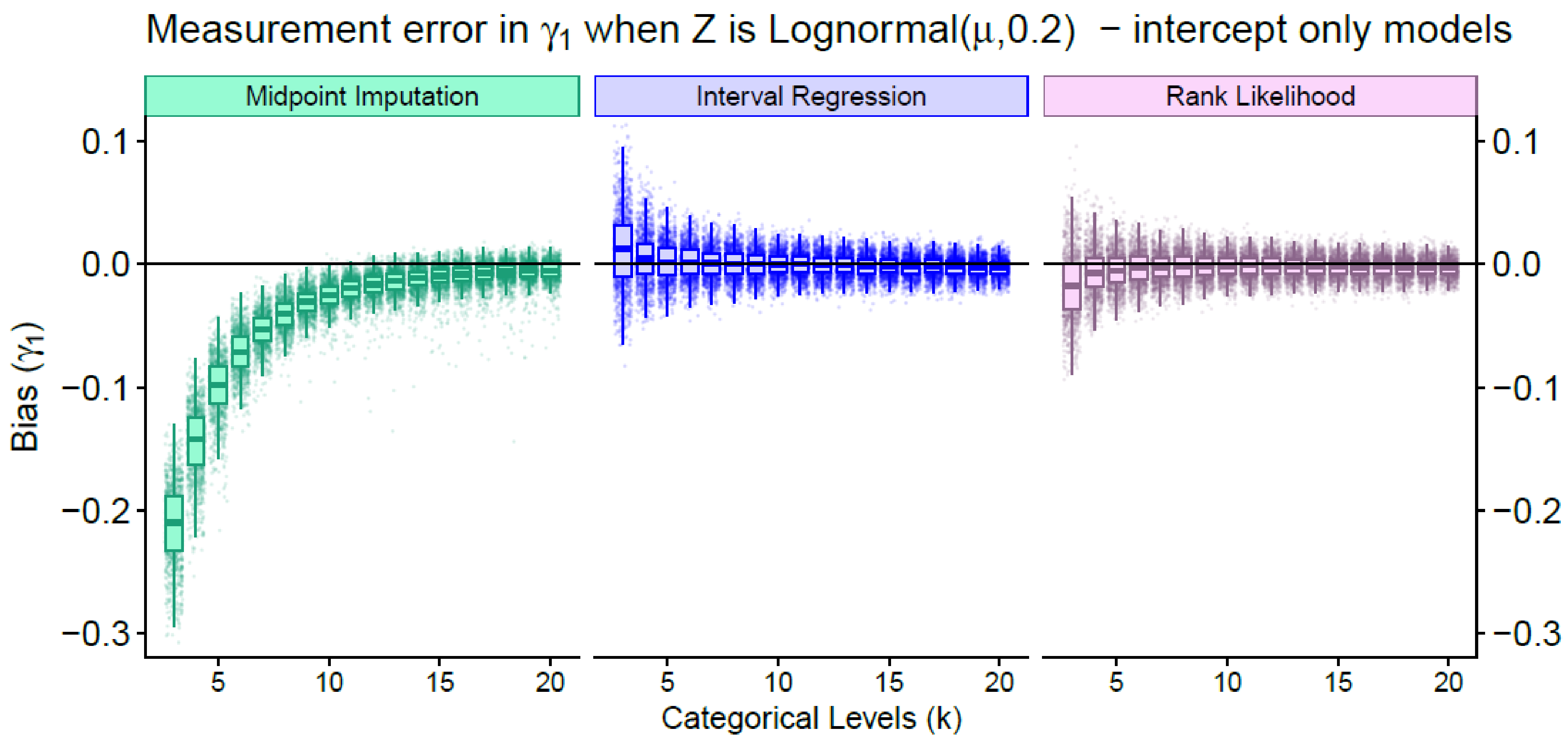
MC Results – Normality (intercept only)



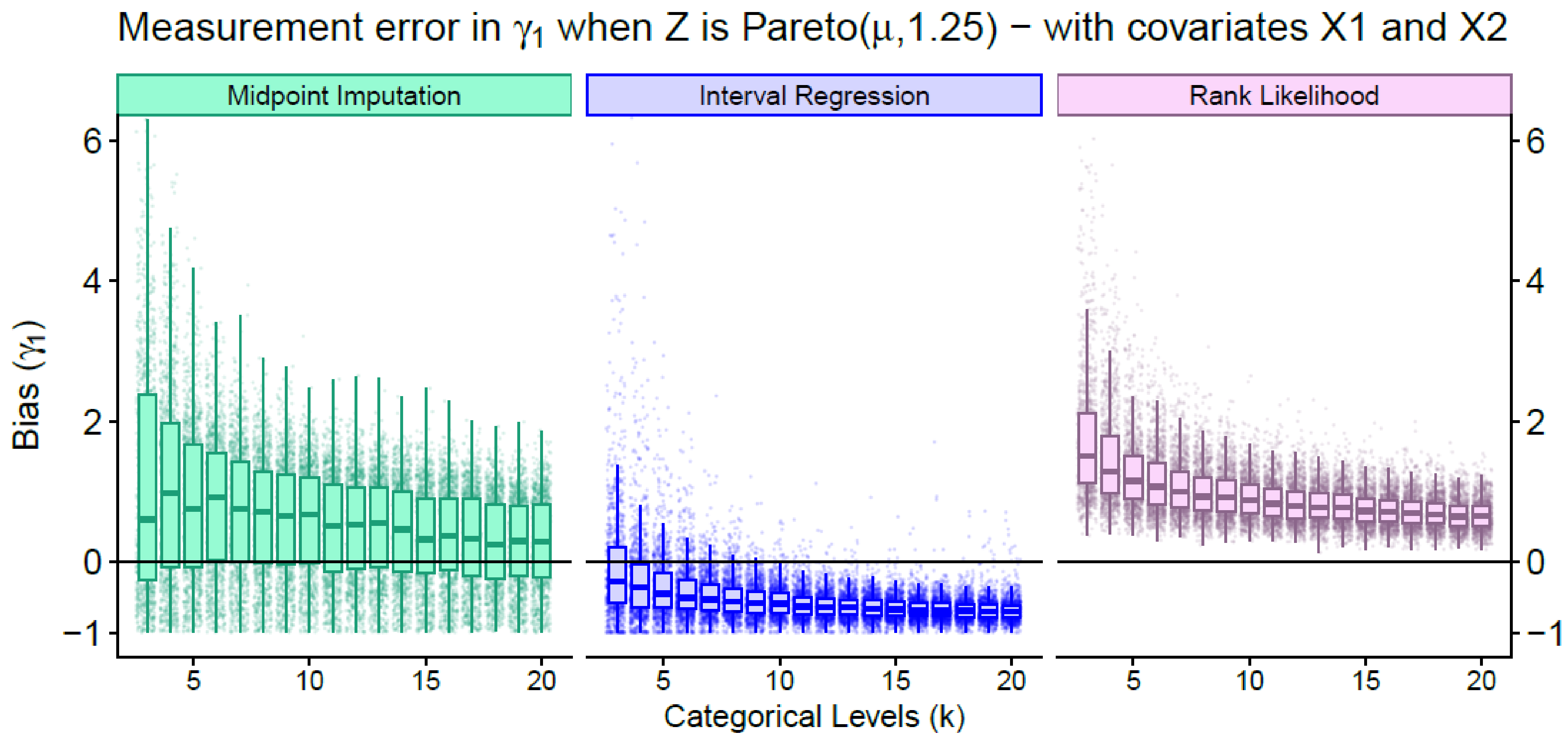
MC Results – Lognormal (with covariates)



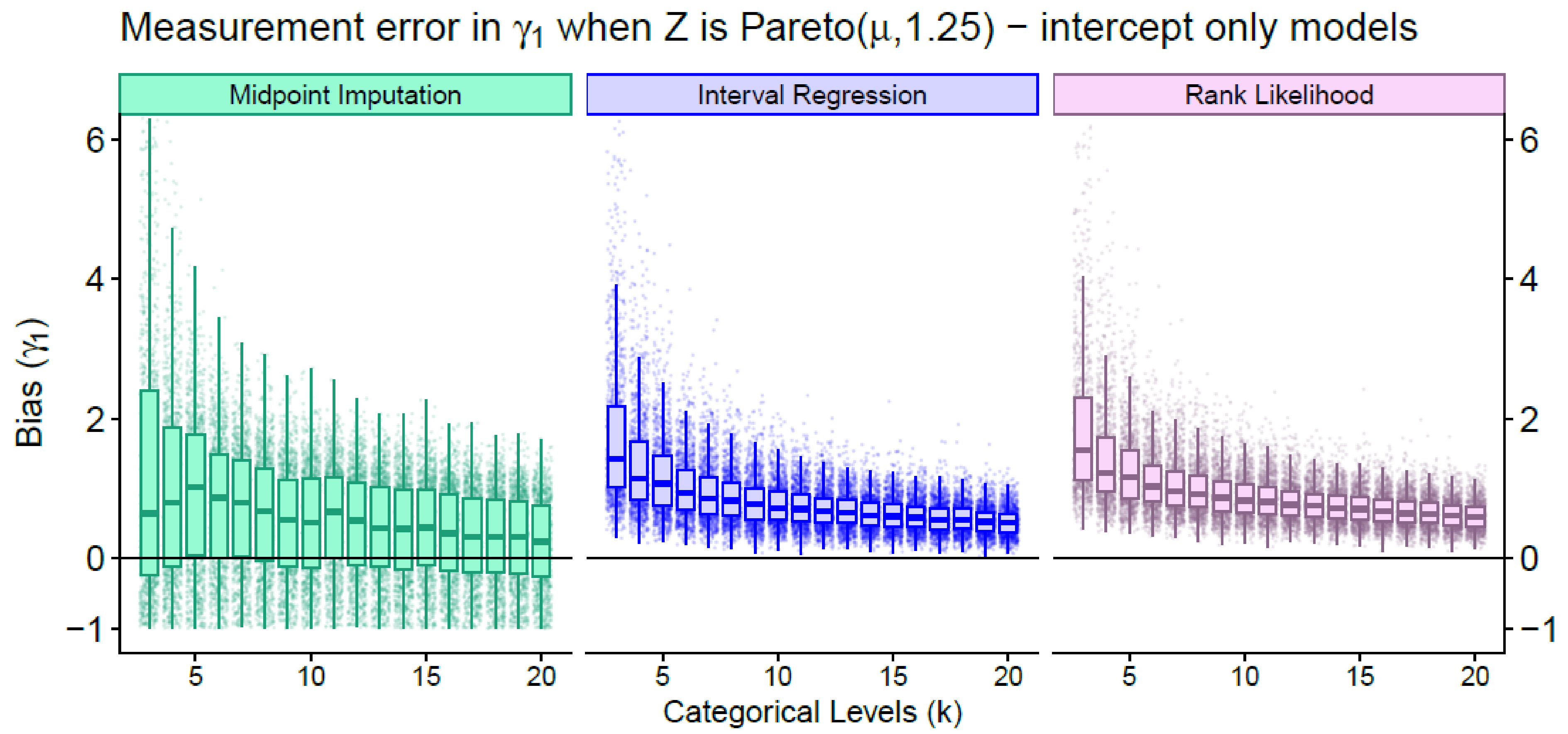
MC Results – Lognormal (intercept only)



MC Results – Pareto (with covariates)



MC Results – Pareto (intercept only)



Conclusions

Conclusions

- Ordinal variables should **not** be fitted as continuous variables.
  - Increasing the **number of categories** (*k*) does not help.

Conclusions

- Ordinal variables should **not** be fitted as continuous variables.
  - Increasing the **number of categories** (*k*) does not help.
- If metric is of interest, should you use midpoint imputation?
  - If you assume **Z is normal**, and you have **more than 12 categories**, OK!
  - Otherwise, use **interval regression** or **Bayesian Rank Likelihood** to **predict Z**!

Conclusions

- Ordinal variables should **not** be fitted as continuous variables.
  - Increasing the **number of categories** (*k*) does not help.
- If metric is of interest, should you use midpoint imputation?
  - If you assume **Z is normal**, and you have **more than 12 categories**, OK!
  - Otherwise, use **interval regression** or **Bayesian Rank Likelihood** to **predict Z**!
- The results show that an intercept-only model, is good enough,
- But perhaps a **model** can be more relevant if the DGP is more complex (like in a Pareto).

Thanks for your attention!

FIN