Small Sample Properties of Partially-Observed Rank Data Estimators

Christopher Adolph, University of Washington, cadolph@uw.edu

October 6, 2011

Abstract. Several estimators from the social science toolkit might be used to model the relationship between imprecisely-observed ranks in a hierarchy, and covariates explaining those ranks. But application of standard methods—such as linear regression, ordered probit, or censored regression—is complicated by the interdependence of rank observations. Monte Carlo evidence shows that estimators which either delete partially observed ranks and/or inappropriately assume ranks are iid perform poorly, yielding inefficient and sometimes biased estimates, and wildly inaccurate confidence intervals. In contrast, a Bayesian partial rank model—designed to impute missing ranks within known bounds, and account for interdependence across ranks—performs well, even when most or all ranks are observed imprecisely.

This note explores the properties of several estimators an analyst might use to model a dependent variable which takes the form of a ranking of the observed cases. Variables of this type are ubiquitous in the study of power and position within hierarchical organizations, but are under-exploited in political science and related fields, perhaps due to a lack of analytic methods, and a related (and unwarranted) assumption that precise observation of ranks is required for modeling their relationship to explanatory covariates.

Partially-observed rank data refers to ranks which are, at least in some cases, known only up to a logically consistent set of bounds. For example, expert observers of an organization may consider the exact ranks of top officials to be common knowledge, but may be willing to state the ranks of mid-level officials only up to bounds implied by their general level within an organizational chart. Here we consider the application of seven different statistical models to partially-observed ranks of this type. Some of these models are based on linear regression, and simply delete partially observed cases. Other models are based on ordered probit, and instead delete the precise ranks, and analyze tiers, instead. Some models use maximum likelihood estimators for censored data to simultaneous incorporate precise and bounded ranks.

But all the foregoing models have a problem: they assume ranks are identically and independently distributed (iid), when ranks are, of course, logically interdependent. If one observed individual secures a specific rank, no other person can have that rank, and, more insidiously, the rank of each individual depends on their strength relative to all other individuals. We consider a seventh approach which recognizes and incoporates this interdependence, a Bayesian model of partial rank data, discussed in further detail in Author (2011), and show that this model has superior properties in analyzing data of this kind.

Before proceding to the evidence for this claim, we note that for estimators applied to ranked individuals in a hierarchy, asymptoptic properties are irrelevant, if not entirely fictive. This is true for two reasons. First, analysts often have complete universes of ranked individuals, at least at the top, so appeals to repeated sampling from a larger universe are empty. Second, and arguably more important, it is difficult to sustain the claim that relationships inside a ranked hierarchy will stay the same if the size or scope of that hierarchy is expanded—whether in fact, or in an inferential thought experiment—and any effort to reach desirable asymptoptic properties by "adding more data" may do no more than alter the relationships the model sought to uncover. In a simple example, if we want to improve the precision of our estimates of the determinants of elite civil servant ranks, past a certain point, we cannot appeal to adding more individuals from the bottom of the organization: the variables which determine who gets to be minister, vice minister, or assistant vice minister are likely very different from those which determine who gets to be a janitor or unpaid intern. For these reason, we focus from the start on the small sample properties of our estimators. To do so, we use an exploratory, but nonethless illustrative, set of Monte Carlo experiments.

1 Design of the Monte Carlo experiment

We consider a model of rank data where y_i, $i \in 1, \ldots, N$, indicates the rank data actually observed. The true ranks, which may sometimes be unobserved, are denoted $\breve{y}_i$ and are a function of a continuous latent strength, y_i^*. Using $\text{rank}(a; \mathbf{b})$ to indicate the rank of element a in the set of elements $\mathbf{b}$, we define the true rank of individual i in terms of latent strength, such that $\breve{y}_i = \text{rank}(y_i^*; \mathbf{y}^*)$. The latent strength itself is a function of covariates, and is given by the true model:

$$y_i^* = 0.25x_{1i} + 0.5x_{2i} + 0.25x_{3i} + \varepsilon_i, \qquad \varepsilon \sim \mathcal{N}(0, 1) \tag{1}$$

The covariates are drawn from a multivariate normal distribution, and each has zero mean and unit variance. The first two covariates are correlated, $\text{cor}(x_1, x_2) = 0.5$, while the others are not, $\text{cor}(x_1, x_3) = \text{cor}(x_2, x_3) = 0$.

In situations where we completely observe the ranks, $y_i = \breve{y}_i$ for all i. But recall that rank data are often only partially observed, such that there are "tiers" of observations whose ranks are known only up to a common set of upper and lower bounds. In these cases, y_i will be missing, but we will observe the tier bounds, y_i^{lower} and y_i^{upper}, which contain the true $\breve{y}_i$.

In our Monte Carlo experiments, we draw random datasets from the true model, and attempt to recover the conditional relationship between $\breve{y}_i$ and $x_{\cdot i}$ even when y_i is only partially observed. We consider datasets across a range of sizes likely to be encountered in applications to political hierarchies. We also partition the ranks into three equally sized tiers, and allow the ranks within each tier to be observed or unobserved. Our estimators always know the tier of each observation, even when they do not know the rank.

Because we partition our data into three parts, it is convenient to consider datasets where N is a power of 3. Specifically, we create datasets with $N = 9$, 27, 81, 243, or 729 observations, and for each size consider cases where no tiers are observed, only the top tier is observed, the top two tiers are observed, or the full dataset is observed. Testing all combinations of N and these tier observation patterns results in 20 different scenarios. For each, we generate 100 datasets from the true model above, for a total of 2000 Monte Carlo runs.

2 The quantity of interest

In many Monte Carlo experiments, interest centers on the ability of estimators to recover the model parameters contained in equation 1 above. However, these parameters are not the quantities of interest in rank data applications. First, the scale of y_i^* cannot be identified, so the scale of these parameters is similarly unknown. Second, and more important, the effect of $\mathbf{x}$ on y_i depends on the full vector of latent strengths, y_i^*. That is, the number of ranks gained by an increase in x_1 depends not only on the parameter of x_1, but also on how close the individual's latent strength is to the strength of those individuals immediately above.

Instead of the raw parameters, the quantity of interest in rank data models is the expected (percentile) change in rank given an increase in the covariate x_1 for an individual who initially has average values on all observed covariates. This quantity of interest exactly matches the quantities of interest considered in applications like Author (2011), and, because we measure it in percentiles rather than raw ranks, this quantity is also comparable across datasets with different N.

3 Measuring estimator performance

We judge the performance of each estimator's conditional expectations of the effect of x_1 on rank $\breve{y}_i$ using three criteria:

1. **Bias**, or the average difference between the true and estimated percentile change in rank, given a one-unit increase in the covariate x_1 across our Monte Carlo runs. For ease of comparison across models, we show only the magnitude of bias here. Smaller values are better, and indicate less bias.

2. **Root mean squared error (RMSE)**, or the expected error between the true and estimated percentile change in rank, given an increase in the covariate x_1 across our Monte Carlo runs. Smaller values are better, and indicate greater efficiency. Because RMSE incorporates bias, it is our preferred indicator of which model produces the best estimates.

3. **Coverage of confidence intervals**, which shows, for each interval from 1% to 99%, how often the true lies inside the reported confidence interval. This fraction should match the stated interval for us to fully trust reported confidence intervals.

4 Estimators Considered

We consider seven competing estimators of the relationship between partially observed ranks and covariates:

1. **Bayesian Partial Rank Model.** This is the model described in Author (2011) and associated appendices. Unlike all other models tested, the Bayesian Partial Rank Model makes no assumption of independence across rank observations, and unlike most of the other models, it makes full use of available information about partially observed ranks. We expect the Bayesian partial rank model to have at least as low bias and efficiency as other models, and expect it to outperform other models when many ranks are known only up to a bound. Finally, we expect only this model to produce unbiased confidence intervals.

2. **Linear Regression.** Applying linear regression to ranks allows us to produce an estimate of the conditional expectation of rank given x_1:

$$\mathrm{E}(y_i|x_{1i} = 1) - \mathrm{E}(y_i|x_{1i} = 0) \tag{2}$$

   To the extent that ranks are fully observed, this formula should provide an unbiased and efficient estimate of conditional relationships, by the usual best linear unbiased estimator (BLUE) properties of least squares. However, because the ranks are not iid, the standard errors and associated confidence intervals reported by linear regression will be incorrect, even when ranks are fully observed. Moreover, when many ranks are known only up to a set of bounds—as is often the case in practice—linear regression on the listwise deleted data may be quite inefficient, or even biased if the missingness is correlated with the data (King et al., 2001).

3. **Ranked Linear Regression.** Ranks themselves cannot be Normally distributed or even iid. Moreover, the conditional expectations produced by linear regression are obviously not true ranks, as these conditional expectations are not unique for each individual and do not condition on the expected ranks of other individuals. We might improve (but not completely salvage) linear regression standard errors and confidence intervals if we instead

compute the expected change in the *rank* of expected rank. That is, rather than reporting $\mathrm{E}(y_i|x_{1i} = 1) - \mathrm{E}(y_i|x_{1i} = 0)$ under linear regression, we compute from the linear regression estimates:

$$\mathrm{rank}\left(\mathrm{E}(y_i|x_{1i} = 1); \mathrm{E}(\mathbf{y}|\mathbf{x})\right) - \mathrm{rank}\left(\mathrm{E}(y_i|x_{1i} = 0); \mathrm{E}(\mathbf{y}|\mathbf{x})\right), \tag{3}$$

a quantity which is a change in ranks. We conjecture that standard errors and confidence intervals calculated (using standard simulation techniques) for this *ad hoc* rank adjustment of the linear regression conditional expectations may better match the true rank distribution, and with minimal effort might mitigate the iid issues plaguing linear regression. However, because the estimator itself still assumes y_i is iid, we do not expect to full correction of bias in confidence intervals.

4. **Ordered Probit.** An alternative approach is to abandon direct modeling of the ranks, and instead use an ordered choice model to predict the tier to which each individual belongs. This might mitigates, but cannot eliminate, the iid problem: tiers are still interdependent, as no more than one-third of our observations can lie in each tier. This violation of iid means the usual nice properties of maximum likelihood estimators need not apply. Compounding the problem, ordered probit discards more and more information the better observed are ranks. Yet while ordered probit discards rank information in estimation, it still *produces* estimates of each individuals rank: the expected latent strength of each individual. Ordered probit essentially trades some precision in hope of potentially more reliable standard errors.

5. **Ranked Ordered Probit.** Just as with linear regression, we conjecture that we can mitigate the problematic iid assumption of ordered probit by converting the conditional expectations of latent strength produced by the model to ranks, using formula 3 above. We suspect that standard errors and confidence intervals computed from the ranked conditional expectation will have better properties, while still falling short of the ideal.

6. **Censored Regression.** Censored regression is a generalization of Tobit which allows for an arbitrary number of alternating ranges of fully observed and partially observed cases, analogous to observed and partially observed tiers in rank data (Schnedler, 2005). Thus censored regression may combine the best of linear regression and ordered probit when at

least some ranks are known exactly. However, censored regression assumes our data are iid, and so the usual good properties of maximum likelihood fail to hold. Confidence intervals in particular should be biased.

7. **Ranked Censored Regression.** As with linear regression and ordered probit, we conjecture that using formula 3 to rank the conditional expectations of rank produced by censored regression may improve, but not completely repair, the standard errors and confidence intervals yielded by this technique at minimal cost.

Although the six alternatives to the Bayesian partial rank model are easier to estimate in terms of computer time, they are less widely applicable. Because ordered probit requires estimation of cutpoints dividing the tiers, it involves estimating a total of five parameters, and is not estimable for the $N = 9$ datasets. And because linear regression and censored regression both require at least some fully observed cases to estimate, these methods are unavailable when all tiers are missing. Only the Bayesian partial rank model is identified and estimatable in all scenarios, and only the Bayesian partial rank model is expected in theory to have efficient results with meaningful confidence intervals.

5 Results

The top panel of Figure 1 shows patterns of bias in estimation of conditional expectations of rank, given the number of ranked individuals and the pattern of partial observation of rank. Moving from left to right within each plot, we see how bias changes as N increases. As we move to the right across the figure to different plots, we see how bias changes as more of these cases are fully observed, rather than simply bounded. Lower values of bias, and thus lower lines, are better.

In general, the Bayesian partial rank model (solid black lines) is consistently at or near the lowest levels of bias; over the broad sweep of scenarios, no other model comes close. The only notable competitor to the partial rank model is unranked linear regression (dashed green line), but only when ranks are fully observed, and N is not too small. In this case, linear regression (and the trivially equivalent censored regression) are slightly less biased than the Bayesian partial rank model, but the advantage is small.

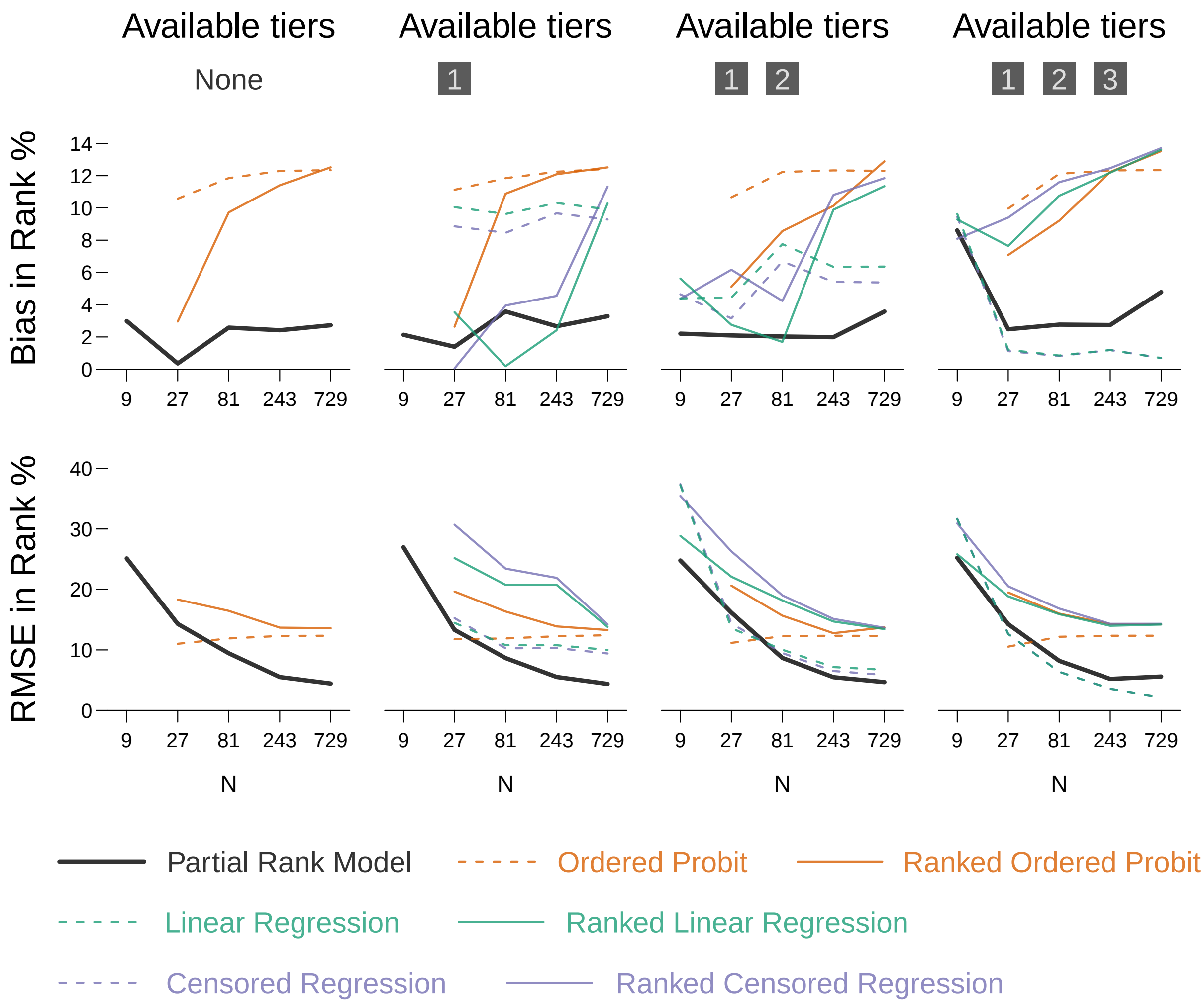

Figure 1: **Bias and efficiency of rank data estimators.** The plots show bias and efficiency in the estimation of conditional relationships between ranks and covariates, and are averaged over 100 simulated datasets. "Available tiers" indicates which of three equally-sized tiers of the ranks are fully observed. The bounds of tiers are always available.

Ordered probit (dashed orange) is always substantially biased. This bias does not shrink, even in relatively large datasets. The ranked method for linear regression (solid green) and censored regression (solid purple) substantially reduce bias in linear regression and censored regression in small, partially observed samples, but ranking does not help if the number of fully observed cases is large. Unranked censored regression outperforms unranked linear regression when some ranks are known only up to a bound, but the reduction in bias is small relative to that offered by the partial rank model.

The efficiency of these models follows a similar pattern: the Bayesian partial rank model is consistently the most efficient estimator, or at least essentially tied for most efficient. Ordered probit's performance is mediocre, and does not improve as N increases. Linear regression and censored regression are distinctly less efficient than the partial rank model when ranks are partially observed. With partial observation, censored regression once again marginally outperforms linear regression. In small, partially observed samples, applying ranks to the conditional expectations of linear regression and censored regression exacts a large efficiency penalty, exposing a trade-off between bias and efficiency in the use of this technique. However, because the Bayesian partial rank model is clearly superior in bias and efficiency given partial observation, and a strong competitor when ranks are complete, this trade-off can be avoided altogether.

When we turn to the properties of confidence intervals for the estimators, it becomes clear that the Bayesian partial rank model is the only viable approach. It is not enough to have unbiased or efficient estimates of the relationship between covariates and rank: we also need to know whether we can trust these estimates, and for that we need unbiased estimates of confidence intervals.

Figure 2 shows the coverage of each confidence interval from 1% to 99% for each model. For example, if the 95% confidence intervals given by a model is unbiased, it should include the true effect of x_1 on $\breve{y}$ in approximately 95 out of 100 simulations. If all the confidence intervals given by a model are unbiased, then its line in Figure 2 will lie on the 45° line. If the model's line falls into the upper triangle, its confidence intervals are too conservative. But if the model falls into the lower triangle, its confidence intervals encourange too much confidence. Because we have only 100 simulated datasets, we do not expect great precision here. Instead, we look for general patterns that may be confirmed in further Monte Carlo work.

Three observations emerge from Figure 2:

First, in almost every case, the Bayesian partial rank model has appropriate coverage of confidence intervals, while in most cases, other models have poor coverage. The invalid iid assumption behind other models' confidence intervals imposes a large cost: we typically cannot trust inference about ranks from linear regression, censored regression, or ordered probit. By avoiding this assumption, the partial rank model provides reliable measures of the uncertainty of its estimates, and thus makes scientific analysis of partially-observed rank data possible.

Second, unranked models in most cases have awful coverage, and encourage potentially dramatic over-confidence in results. Unranked ordered probit performs terribly across the board, contrary to hopes that it might mitigate interdependence by omitting explicit rank data. Because tiers themselves are interdependent, ordered probit's iid assumption is still violated, yet there is also less information available to the model to produce a good estimate. Unranked linear regression and censored regression give excessive confidence in comparison to their ranked counterparts whenever some ranks are missing; as N increases, this overconfidence becomes quite severe.

Third, ranking the conditional expectations of linear regression, censored regression, and to a lesser extent, ordered probit, does improve the coverage of their confidence intervals, as conjectured above. However, as expected, this repair is not complete, and the Bayesian partial rank model has better coverage than these *ad hoc* alternatives in all scenarios.

Finally, we note an anomaly in large ranked datasets, in particular those with *only* fully observed cases: in these situations, the confidence intervals offered by the Bayesian partial rank model appear somewhat overconfident, while with full observation, linear regression confidence intervals appear much less biased. We leave the source of this puzzle to future work, but note that for partially observed data, the properties of the Bayesian partial rank model remain good.

6 Conclusion

Monte Carlo evidence suggests analysts of rank data should use the Bayesian partial rank model, rather than linear regression, censored regression, or ordered probit, whenever some cases are missing or observed up to a bound. Ordered probit in particular should never be used to analyze

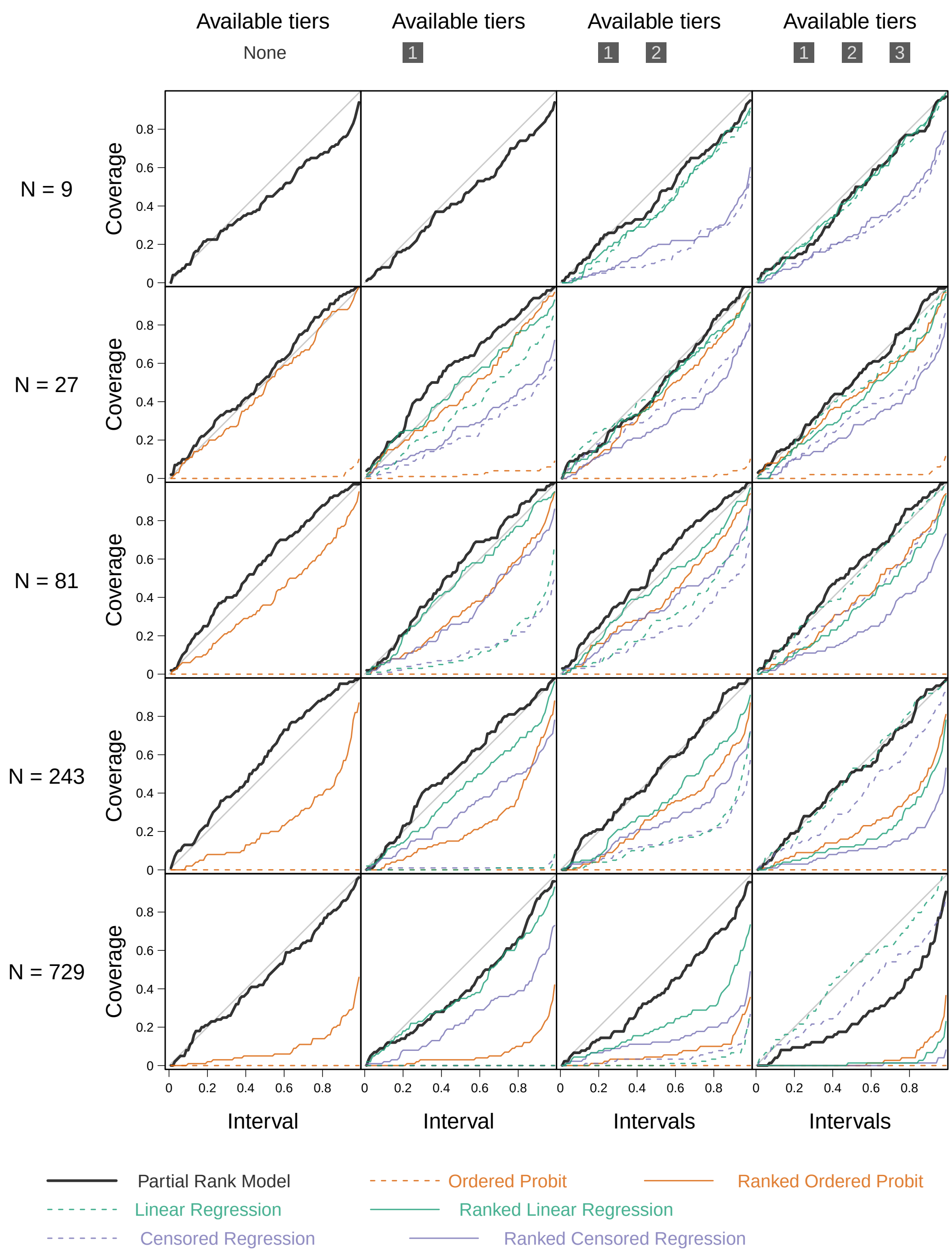

Figure 2: **Coverage of confidence intervals in rank data estimators.** The plots show how often, across 100 simulated datasets, the reported confidence intervals around conditional expectations contain the true effect of covariates on rank. "Available tiers" indicates which of three equally-sized tiers of the ranks are fully observed. The bounds of tiers are always available.

rank data, even if ranks are aggregated up to tiers. *Ad hoc* ranking of conditional expectations from these alternative models can partly repair some of the bias in their confidence intervals, but at a significant efficiency cost; in any case, the Bayesian partial rank model generally outperforms these *ad hoc* models on all measures in small samples. Our recommendation to use the partial rank model holds even if all ranks are observed only partially: even in this case, the ordered probit alternative gives more biased estimates and dramatically incorrect confidence intervals compared to the Bayesian partial rank model.

When rank datasets are large, and all ranks are fully observed, linear regression may be a practical alternative to the partial rank model, despite the technically incorrect iid assumption. However, large, fully observed rank datasets are the exception in studies of political rank, and for most applications, the Bayesian partial rank model remains the safe choice in terms of bias, efficiency, and coverage.

References

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American Political Science Review* 95(1):49–69.

Schnedler, Wendelin. 2005. "Likelihood estimation for censored random vectors." *Econometric Reviews* 24(2):195–217.