When Ordinal Models Mislead: Diagnosing Distributional Assumptions and Improving Inference

Chanhyuk Park

Abstract

Many of the political science research aiming the identification of causal parameters also rely heavily on ordinal scales in measuring outcomes such as attitudes, preferences, or perceptions. While emphasis on causal identification increases, there has been less discussion on how we can do causal inference with these ordinal outcomes. Simple cardinalization only identifies the direction of the causal effect at best and generally cannot identify the causal effect directly. IRT based methods require multiple items of questions which is usually not the case in practice. Ordered logit and probit models are widely used but, they rely on strong and often untested assumptions about the distribution of unobserved error terms, and when these assumptions are violated, the estimates are inconsistent and biased, leading to incorrect inferences. This paper introduces a diagnosis for the distributional assumptions by utilizing surrogate-based residuals, and suggests a semiparametric identification strategy as an alternative model that is more robust than standard Ordered Probit or logit models. Through simulation studies and an empirical application to political attitude data, this paper shows how departures from assumed error distributions can lead to substantively misleading conclusions.

1 Introduction

With growing emphasis on the causal inference framework, political science has sought on identifying the causal effect on policies and other political factors. Survey experiment, observational studies with designs such as difference-in-difference and regression discontinuity become some of the prevalent empirical strategy in political science. For example, during the period from 2021 to 2024, about 20% of articles published in top 3 general political science journals (*American Journal of Political Science, American Political Science Journal, and Journal of Politics*) included survey experiment as one of their empirical strategy.

A lost of research with causal inference framework deal with attitudes, opinions and preferences as outcomes, and they often measured with ordinal scales. Researchers are interested in identifying causal effect on people's preferences such as politicians' approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009), support for redistribution (Alt and Iversen, 2017; Magni, 2021), and foreign policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005). Researchers frequently rely on ordered responses, measured with Likert-type choices (e.g., *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree*). Ordinal scales, unlike interval or cardinal scales, carry only information on orders. The common operation with labels of $\{1, 2, 3, 4, 5\}$ is arbitrary and carries no numerical meaning, and the distances between each values are not necessarily equal.

The problem is that commonly used causal inference tools mostly assumes cardinal or at least interval outcomes. Thus, it is not guarantees that tools and designs such as means-difference, difference-in-differences, and regression discontinuity to work on ordinal outcomes. One commonly used approach is ignoring the ordinal nature and treat ordinal outcomes as cardinal variables by assigning numeric labels to each responses. This enables us to use all causal inference tools as usual. However, since the numeric labels are arbitrary, this approach, at best, only can uncover the direction of the causal effect, and the size of the estimates are hard to interpret (Schroder and Yitzhaki, 2017; Bond and Lang, 2018; Bloem, 2022).

Another approach is based on IRT. IRT based methods utilize multiple items of questions on one latent outcome variable to recover the latent outcome variable itself in the first step and then use causal inference tools with the recovered latent outcome variable. Recently developed hierarchical IRT method (Zhou, 2019; Stoetzer, Zhou and Steenbergen, 2025) effectively merges the two steps with EM algorithm. Although this methods can identify the treatment effect up to scale, but in practice, it is hard to find multiple items on political concepts.

Lastly, common ordered logit model and ordered probit model are built on a latent variable framework, and thus useful tool in identifying the causal parameter in latent space. They effectively incorporate the ordinal character of the outcome variable and also guarantees fast estimation through the maximum likelihood estimation (MLE). However, both models rely on strict distributional assumptions about the error term in the latent variable model – logistic for ordered logit, and standard normal for ordered probit. When these assumptions are violated, estimates can be inconsistent and biased, even in large samples. Critically, this bias is not just a matter of inefficiency; in many cases, it can alter the sign of estimated treatment effects or covariate associations, leading to incorrect substantive conclusions (Manski, 1988; Greene and Hensher, 2010).

This paper makes two contributions regarding these problems. First, it introduces a practical and accessible framework for diagnosing distributional assumptions in ordinal regression models using surrogate residuals. These diagnostics allow researchers to detect skewness and other departures from assumed error distributions, implemented via the sure package in R. Second, this paper also proposes a semiparametric alternative to conventional ordinal models when distributional assumptions appear to be violated. Specifically, I focus on the Klein–Spady estimator, which does not assume a specific error distribution and remains consistent under a broader set of conditions. Utilizing Kernel Density Estimation (KDE) this estimator offers a compelling balance between robustness and interpretability, making it a promising option for applied survey researchers.

To evaluate the performance of these methods, the paper presents Monte Carlo simulations that vary the shape of the latent error distribution and the precision of covariate measurement. The simulations show that ordered logit and probit estimators become biased under skewed distributions, while the semiparametric estimator maintains unbiasedness. Finally, this paper apply these tools to a real-world political survey, illustrating how distributional misspecification can meaningfully affect substantive conclusions and showing how semiparametric methods provide a more robust alternative.

By offering both a diagnostic strategy and an estimation solution, this paper provides researchers with tools to improve inference in a wide class of models using ordinal outcomes—a common yet under-scrutinized challenge in political science.

2 Causal Inference with Ordinal Outcomes

With growing emphasis on the causal inference framework, political science has sought on identifying the causal effect on policies and other political factors. Survey experiment, observational studies with designs such as difference-in-difference and regression discontinuity become some of the prevalent empirical strategy in political science. For example, during the period from 2021 to 2024, about 20% of articles published in top 3 general political science journals (*American Journal of Political Science, American Political Science Journal, and Journal of Politics*) included survey experiment as one of their empirical strategy.

Many of political science research with causal inference framework target the causal effect on abstract and subjective concepts and preferences, and they often operationalized as ordinal variables.Researchers are interested in identifying causal effect of policies and other factors on people's preferences such as politicians' approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009), support for redistribution (Alt and Iversen, 2017; Magni, 2021), and foreign policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005). We cannot observe these outcomes directly, but we believe many of them can be expressed on a uni-dimensional space; for example, we can imagine to line up people based on their opinion on pension reform. Researchers frequently rely on ordered responses, measured with Likert-type choices (e.g., *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree*).

Although this combination of ordinal outcome and causal inference becomes more common, there has been less discussion on exactly what it means to do causal inference with ordinal outcome. Ordinal scales, unlike interval or cardinal scales, carry only information on orders. The common operational labels of $\{1, 2, 3, 4, 5\}$ are arbitrary and carry no numerical meaning, and the distances between each values are not necessarily equal. This unique characteristic of ordinal outcomes poses question on what should be the estimand, and which causal inference tools to be used to estimate, and what the estimates mean, because many of the causal inference tools assume that the outcomes are cardinal or at least interval.

To be more specific, let us suppose a following simple model. There is a latent continuous outcome variable of Y^* in a latent space. This can be thought as lining up people based on their preferences or opinion on a certain political issue. Then, the true data-generating process (DGP) is captured as:

$$Y^* = X^T\beta + D^T\tau + \epsilon \tag{1}$$

D is a binary treatment indicator, where 1 denotes a treatment group and 0 denotes a control group. τ is the coefficient for the treatment status, X is a vector of regressors and β is a vector of coefficients and ϵ is the latent error term.

The continuous Y^* is not observed directly, but only the ordinal labels, $Y \in \{1, 2, \ldots, J\}$ is observed.

$$Y = \begin{cases} 0 & \text{if } \alpha_{-1} < Y^* \leq \alpha_0 \\ 1 & \text{if } \alpha_0 < Y^* \leq \alpha_1 \\ \vdots & \vdots \\ J & \text{if } \alpha_{J-1} < Y^* \leq \alpha_J \end{cases}$$

, where α_k denotes the threshold points for each ordinal category. Conventionally, $-\infty = \alpha_{-1} \leq \alpha_0 \leq \ldots \leq \alpha_k = \infty$. Here I assume that everyone has same thresholds, ignoring the additional problems arising from inter- and intra-personal differences in scale use (King et al., 2004; King and Wand, 2007).

Borrowing from the potential outcome framework (Rubin, 1974), let us denote the potential outcome in the latent space as $Y^*(d)$ and the potential outcome in observed data as $Y(d)$, where $d \in \{0, 1\}$ denotes the treatment status. Standard causal inference assumptions such as stable unit treatment value assumption and ignorability are met by construction regarding the latent Y^*. Since

D is binary, the target treatment effect here is τ.

Regarding the causal estimand τ, there have been three common approaches in political science: 1) ignore the ordinal nature and regard it as cardinal value 2) IRT based approaches, and 3) ordered regression. Assigning numerical values to ordinal outcomes enables using usual causal inference tools such as means-difference, least squares, and difference-in-differences. One downside for this approach comes from the fact that the numerical labels are arbitrary, and theoretically any labels should work if the preserve the order. For example, labeling *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree* as $\{1, 2, 3, 4, 5\}$ is no more reasonable than labeling them as $\{1, 3, 15, 50, 100\}$. As discussed in Bond and Lang (2018) and Schroder and Yitzhaki (2017), in most cases, there exists at least one labeling scheme that can flip the sign of the causal effect. Bloem (2022) partly deals with this problem by providing a sensitivity test and a partial identification method based on the test.

However, the more serious problem of cardinalization approach is the interpretation of the estimates. For instance, means-difference and difference-in-differences are comparing numerical expectations of outcomes from different groups. However, if the observed outcomes are *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree*, even though we have transformed them into numbers, what does the expectations of these responses, mean in terms of the latent variable Y^*? In a potential outcome framework, the relationship between $\mathbb{E}\left[Y(d) \mid X\right]$ and $\mathbb{E}\left[Y^*(d) \mid X\right]$ is not clear. Furthermore, even if we safely put some meanings to the expectations, the size of the difference between expectations are hardly interpretable, since they depend on the labeling scheme. For example, in case of 5 point Likert scale, if we assign $\{2, 4, 6, 8, 10\}$, instead of $\{1, 2, 3, 4, 5\}$, the size of the difference will be doubled. This may significantly misleading, since the estimates we have only have very loose link with the treatment effects, and may over- or under-estimate them.

Another approach in estimating causal effect with ordinal variable is through IRT. IRT based Two-step approach first estimate the latent variable using the serious of items that were designed to measure the same concept in different angles. In the context of the model setting above, using multiple Y, IRT estimates the Y^*. After we get the estimate for the latent variable Y^*, then we can employ the usual causal inference tools to identify τ up to scale since it is now have numerical meanings. The second variant of IRT based approach, the hierarchical IRT is recent discussed in Stoetzer, Zhou and Steenbergen (2025). Instead of estimating the latent variable in the first step, they effectively merge the two steps in to EM algorithm, and produces more consistent estimates of the causal effects. Since IRT based methods do not put any numerical meaning to the ordinal outcomes themselves, the estimates from the methods can be interpreted as the target causal effect up to scale. One downside of the approach is that it is advised to have at least 3 different items for IRT to work properly, but most of the political science research rely on 1 or 2 items in measuring their outcomes. Therefore, when researcher is focusing on the treatment effect on specific dimension such as opinion on gender inequality or immigrant issue, where outcomes are not usually measured with multiple items, IRT might not be a good choice.

The third approach is using ordinal regression, and by far the most common methods are ordered logit and probit regressions. Similar to the IRT based methods, both ordered logit and probit utilize the ordered nature of the outcomes measured in ordinal scales, and designed to identify the actual target causal effect in unobserved, latent space up to scale. However, these models require strong distributional assumptions on the error term in the latent space, ϵ; ordered probit assumes a normal distribution and ordered logit assumes a logistic distribution. While these assumptions facilitate fast and efficient estimation using maximum likelihood, when the distributional assumptions fail, the estimates are statistically inconsistent and biased.

Let i be the index for ith data. To construct likelihood function, the probability of observing $Y = j$ given X_i is:

$$\begin{aligned}
\mathbb{P}\left(Y = j \mid X_i\right) &= \mathbb{P}\left(Y^* \leq \alpha_j\right) - \mathbb{P}\left(Y^* > \alpha_{j-1}\right) \\
&= \mathbb{P}\left(\epsilon \leq \alpha_j - (X_i^T\beta + D_i^T\tau)\right) - \mathbb{P}\left(\epsilon > \alpha_{j-1} - (X_i^T\beta + D_i^T\tau)\right) \\
&= F(\alpha_j - (X_i^T\beta + D_i^T\tau)) - F(\alpha_{j-1} - (X_i^T\beta + D_i^T\tau))
\end{aligned}$$

, where $F(\cdot)$ is a unknown but assumed distributional function (CDF).

Based on this, the parameters can be easily estimated with maximum likelihood.

$$(\beta,\tau)=\arg\max_{\beta,\tau}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(F(\alpha_j-(X_i^T\beta+D_i^T\tau))-F(\alpha_{j-1}-(X_i^T\beta+D_i^T\tau))\right)\right]$$

To solve this MLE, standard ordered logit and probit models put assumption on $F(\cdot)$. For example, Ordered Probit model assume that the error term (ϵ) follows the standard normal distribution. Let $\Phi(\cdot)$ denote the standard normal distribution function. Then the MLE becomes:

$$(\beta,\tau)=\arg\max_{\beta,\tau}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\Phi(\alpha_j-(X_i^T\beta+D_i^T\tau))-\Phi(\alpha_{j-1}-(X_i^T\beta+D_i^T\tau))\right)\right]$$

However, it is not guaranteed that the distributional assumption that $F(\cdot)=\Phi(\cdot)$, and if they are different, the estimators from ordered probit model are to be biased and inconsistent, because it will converge to some value $(\tilde{\beta},\tilde{\alpha})\neq(\beta,\alpha)$ and increase in sample size does not make the two distributions closer, and therefore the estimators from two MLE will not converge to the true values (Manski, 1988; Greene and Hensher, 2010; Bond and Lang, 2018). Considering ordered logit model, logistic distribution is assumed and exactly same logic will lead to the biased and inconsistent estimation of parameters. The size and direction of the bias depend on the difference between $F(\cdot)$ and the assumed distribution, but in recent empirical works, right (positively) skewed error distributions would generally attenuate the size of the β estimations (Johnston, McDonald and Quist, 2020; Smits et al., 2020). Theoretically, the estimates from a misspecified model may have opposite sign to the true treatment effect, leading researchers toward substantially different inferences (Manski, 1988; Greene and Hensher, 2010). This issue even harder to detect because ordered logit and ordered probit models often produce similar results.

This paper extends the ordinal regression approach, first by introducing a diagnostic tool for the validity of the distributional assumptions for the ordered logit and probit models, and by introducing a semiparametric method that can estimate the true treatment effect without strong distributional assumptions.

3 Test for the Distributional Assumptions: Surrogate-based Residuals Approach

The researchers generally overlook the inconsistent bias in ordered logit and probit regression models partly because the absence of appropriate diagnostic tools that can test the validity of the distributional assumptions.

Although statistical tests for distributional assumptions has been existed (Bera and Jarque, 1982; Glewwe, 1997; Weiss, 1997), most of them are limited to probit models because most of the tests based on the moment conditions and moment conditions for standard normal is much clearer. More recently, Li and Shepherd (2010) suggested a general residual analysis approach using the sign-based statistics (SBS), by collapsing ordered choices into multiple binary choices, but the result from the analysis is hard to interpret and fails to provide consistent diagnosis because the calculates residuals are dependent on the values of the covariates.

Liu and Zhang (2018) extend the residual approach suggested by Li and Shepherd (2010), but instead of using the sign-based statistic, it generates a surrogate variable for the error term in latent variable space and use that to test for the distributional assumptions. Since the surrogate residuals are constructed to be continuous, it shares the similar properties to that of the common residuals for continuous outcomes. Thus, one can use this surrogate residuals to either graphically diagnosis the validity of the distributional assumption or to construct numerical test including the Kolmogorov-Smirnov test or the Anderson-Darling test.

To illustrate the construction of the surrogate residuals, let's suppose a standard problem setting with latent variable space mentioned in Section 1 and further suppose that we consider a ordered regression model with an assumption that the error term in latent space follows $F(\cdot)$. Then, define Z as $-(X^T\beta + D^T\tau) + \epsilon$, where ϵ follows the error distribution assumed by the model, $F(\cdot)$. If the specified ordered regression model is correct, the marginal distribution of Z should closely follow that of the true latent outcome variable Y^*. Based on Z and the observed outcome Y, then the surrogate variable S is sampled from a conditional distribution of hypothetical variable Z given the observed ordinal outcomes Y. This results in truncating Z at each of the estimated threshold points by the model. To be more specific, S follows below distribution:

$$S \sim \begin{cases} Z \mid \alpha_0 < Z \leq \alpha_1 & \text{if } Y = 1 \\ Z \mid \alpha_2 < Z \leq \alpha_2 & \text{if } Y = 2 \\ \vdots & \vdots \\ Z \mid \alpha_{J-1} < Z \leq \alpha_J & \text{if } Y = J \end{cases}$$

The constructed surrogate S is to be a continuous variable, thus the surrogate-based residuals can be calculated just like the other continuous variable cases:

$$R_S = S - \mathbb{E}\left[S|X\right] = S - \mathbb{E}\left[Z|X\right]$$

Since the surrogate S is continuous, graphical analysis such as Q-Q plot against the assumed distribution can provide first step test for the distributional assumption and bootstrapped goodness-of-fit tests such as Kolmogorov-Smirnov test can be done [1]. Kolmogorov-Smirnov test used to check the equality between the CDF of the assumed distribution and the empirical CDF of the surrogate-based residuals by summing up the distance between two distributions. If the surrogate-based residuals from the considered ordered regression model do not seem to agree with the assumed distribution, one should consider using alternative models [2]. One may want to try alternative parametric regression models such as skewed logit distribution (Nagler, 1994) or a generalized version of ordered logit or probit models (Johnston, McDonald and Quist, 2020) and pick a model with the best goodness-of-fit score using the surrogate-based residuals.

Alternative to the parametric approach rely on semiparametric approach with minimum distributional assumptions. Semiparametric approaches have been developed to mitigate the potential distributional misspecification. The literature can be roughly divided in two categories. The rank-

[1] R package sure provides useful tools such as functions for surrogate residual calculation and plotting.

[2] However, it should be noted that since this diagnosis based on (surrogate) residuals, distributional misspecification is not the only reason that can cause a bad fit. As Glewwe (1997) and Greene and Hensher (2010) noted earlier, there can be multiple reasons including omitted variable with wildly different distributions or misspecification of a functional form of the latent equation. Both misspecified functional form and omitted variable are important issues require attention but for the purpose of this paper, I focus on the issues arise from the misspecification regarding distribution of the error terms.

based maximum score approach put minimum assumptions on distributions that can enable quantile regressions and uses that property to estimate the parameters. One downside of this approach is that it is usually the case that the convergence is very slow (Manski, 1988; Lee, 1992). On the other hand, Kernel-based approaches estimate the error distribution nonparametrically using kernel estimation strategy. Lewbel (2000) was one of the first to attempt relaxing both assumptions in this approach. Klein and Sherman (2002) introduced a shift-restriction-based approach that uses kernel density estimation (KDE) for both cut points and regression coefficients, providing greater flexibility. In the next section, I would like to introduce a Kernel-based semiparametric approach suggested by Klein and Sherman (2002).

4 One Alternative Model: KDE-based Semiparametric Regression

Instead of assuming the distribution of the error term, Klein and Sherman (2002) suggested a method that can identify the treatment effect based on the estimated error distribution. The method first estimate the error distribution based on kernel density estimation (KDE) and then construct a quasi-likelihood function based on it. Beginning with the standard latent variable space setting discussed in Section 2, recall that

$$\begin{aligned}\mathbb{P}\left(Y = j \mid X_i\right) &= \mathbb{P}\left(Y^* \leq \alpha_j\right) - \mathbb{P}\left(Y^* > \alpha_j - 1\right) \\ &= \mathbb{P}\left(Y \leq j \mid (X_i^T\beta + D_i^T\tau)\right) - \mathbb{P}\left(Y \leq j-1 \mid (X_i^T\beta + D_i^T\tau)\right)\end{aligned}$$

Using the Bayes' rule, $\mathbb{P}\left(Y \leq j \mid (X_i^T\beta + D_i^T\tau)\right)$ can be expressed as:

$$\mathbb{P}\left(Y \leq j \mid (X_i^T\beta + D_i^T\tau)\right) = \frac{\mathbb{P}\left(Y \leq j\right) \times g_1((X^T\beta + D^T\tau) \mid Y \leq j)}{\mathbb{P}\left(Y \leq j\right) \times g_1((X^T\beta + D^T\tau) \mid Y \leq j) + \mathbb{P}\left(Y > j\right) \times g_0((X^T\beta + D^T\tau) \mid Y > j)}$$

, where $g_0(\cdot)$ and $g_0(\cdot)$ denote conditional density of $(X^T\beta+D^T\tau)$ given $Y > j$ or $Y \leq j$ respectively.

Both $\mathbb{P}\left(Y \leq j\right)$ and $\mathbb{P}\left(Y > j\right)$ are approximated as a sample probability, and based on local smoothing technique developed by Abramson (1982) and Silverman (1986), they estimate the $g_1(\cdot)$

and $g_0(\cdot)$ density using KDE, and use them to approximate $\mathbb{P}\left(Y \le j \mid (X_i^T\beta + D_i^T\tau)\right)$.

The estimation of $g_1(\cdot)$ starts with pilot density estimation. Let $\hat{\sigma}_1$ be the sample standard deviation of the $(X_k^T\beta + D_k^T\tau 1_{\{Y_k \le j\}}$ and $n_1 = \sum_k 1_{\{Y_k \le j\}}$. Fix $\delta \in (0, \frac{1}{3)}$, and define a pilot bandwidth $h_p = n^\gamma$, where $\frac{1}{10} < \gamma < \frac{1}{3(3+\delta)}$. Then the pilot KDE of $g_1(\cdot)$ can be done in a leave-one-out estimation:

$$\hat{\pi}_{1i} = \frac{1}{n_1}\sum_{k \ne i}\left[1_{\{Y_k \le j\}} \cdot K\left(\frac{(X_i^T\beta + D_i^T\tau) - (X_k^T\beta + D_k^T\tau)}{\hat{\sigma}_1 h_p}\right) \cdot \frac{1}{\hat{\sigma}_1 h_p}\right]$$

Next, to make final KDE to be a smooth density function, local smoothing parameter $\hat{l}_{1i}$ is defined as:

$$\hat{l}_{1i} = \frac{\hat{\pi}_{1i}}{\hat{m}_1}$$

, where $\hat{m}_1$ is a geometric mean of the $\hat{\pi}_{1i}$.

Then define $\hat{d}_1$ for $l > 0$ which approximates the indicator function $1_{\{l > a_{n_1}\}}$, where $a_{n_1} \propto [\ln n_1]^{-1}$.

$$\hat{d}_1 = \frac{1}{1 + \exp\left(-n_1^\epsilon[l - a_{n_1}]\right)}$$

, where $\epsilon \in (0, \frac{1}{40} - \frac{\delta}{20})$.

Lastly, calculate the bandwidth for the final KDE, h_f with parameters above:

$$h_f = \hat{\sigma_1}\left[\hat{l}_{1i}\hat{d}_{1i} + a_{n_1}[1 - \hat{d}_{1i}]\right]^{-\frac{1}{2}}$$

Choose $\alpha \in (\frac{3+\delta}{20}, \frac{1}{6})$. Let $h = n^{-\alpha}$. Define

$$\hat{g}_1(\cdot) = \frac{1}{n_1}\sum_k 1_{\{Y_k \le j\}} \cdot K\left(\frac{(X_i^T\beta + D_i^T\tau) - f(X_k - \beta)}{\hat{h}_f \cdot h}\right) \cdot \frac{1}{\hat{h}_f \cdot h}$$

$\hat{g}(\cdot)$ is defined analogously, replacing $Y_k \le j$ with $Y_k > j$.

Since the KDE may not perform well on the edge cases, trimming out the data points that the absolute value is not within the range of qth quantile of $|X_i|$. Let $\hat{m}$ be the indicator of this trimming:

$$\hat{m}(X_i) = 1_{\{|X_i| < q\text{th quantile of } |X|\}}$$

The final quasi-likelihood function is:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i)\sum_{j=0}^{J}1_{\{Y=j\}}\log\left[\hat{\mathbb{P}}\left(Y \leq j \mid f(X_{i,\beta})\right) - \hat{\mathbb{P}}\left(Y \leq j-1 \mid (X_i^T\beta + D_i^T\tau)\right)\right]$$

The maximum likelihood estimation using this quasi-likelihood function gives estimates for β and τ up to scale.

Theorem 1 (Consistency of the estimator)**.** *For an consistent KDE, as* $n \to \infty$*,* $\hat{\tau} - \tau = o_p(1)$

The consistency of the estimates can be guaranteed by the fact that the KDE method is consistent. As the n goes to infinity, the KDE estimates of $\hat{g}_1(\cdot)$ and $\hat{g}_0(\cdot)$ approach the true density and as the $\mathbb{P}\,(Y\,\hat{}\leq j)$. Then the quasi-likelihood function approach the true likelihood function and the maximum likelihood then guarantees the consistency. The KDE method used in this estimator is locally smoothed KDE, and for the consistency of the locally smoothed KDE, see Klein and Spady (1993), Abramson (1982), and Silverman (1986).

5 Monte Carlo Simulation

I tried several Monte Carlo experiments to test the surrogate-based residuals and compare the performance of the Klein and Sherman semiparametric estimator compared with standard ordered probit and logit models. Here I consider the case where the outcome is a ordinal variable with 3 levels.

The latent variable Y^* follows the relationship:

$$Y^* = D^T\tau + X^T\beta + \epsilon$$

, and we observe Y that is constructed as:

$$Y = \begin{cases} 1 & \text{if} \quad \alpha_{-1} \le Y^* \le \alpha_0 \\ 2 & \text{if} \quad \alpha_0 \le Y^* \le \alpha_1 \\ 3 & \text{if} \quad \alpha_1 \le Y^* \le \alpha_2 \end{cases}$$

, where $\alpha_{-1} = -\infty$, and $\alpha_2 = \infty$.

D simulates treatment status, so each data point is randomly assigned 0(control) or 1(treatment), X is drawn from $\mathcal{N}(0,1)$ and τ set to be 0.5. Since the estimators only identify τ up to scale, without lose of generality, I set β to be 1 for easier interpretation.

I tested for two different distribution for the error term, ϵ, the lognormal distribution with $\log(sd) = 10$ (skewed right or positively skewed) and the negative lognormal distribution with the same parameter setting(skewed left or negatively skewed). Both distributions have relatively fat tailed and skewed compared to the standard normal distribution.Two sample sizes of 1,000 and 2,000 are considered. I first generate a population of size $1e + 7$ and draw random samples for Monte Carlo replication from this population. The number of Monte Carlo replication is 100, and all estimations are the means of the 100 replication results.

By construction, the distributional assumptions for both ordered logit and ordered probit models are violated. I would like to compare the estimation results from both ordered logit and probit model and OLS with Klein and Sherman estimator proposed in Section 4 to show the performance of the semiparametric estimator. For the OLS, I employed two labeling scheme:

Table 1 summarizes the Monte Carlo simulation results. Since all models identify β coefficients up to the scale, I set $\beta_2 = 1$. Recall that the true values for β_1/β_2 is -1 and β_D/β_2 is 0.01. The coefficient estimates from OLS is largely overestimated in size but the signs are correct. The two ordered regression models generally accurate regarding the coefficient estimates for β_1. However, the sign of the coefficient estimates for the treatment variable sometimes go reverse to the true value. These are somewhat expected from the previous diagnosis using surrogate-based residuals. Since the models cannot aptly capture the true error distribution, the coefficient estimates become biased and inconsistent. Considering the standard errors, these coefficients would not carry statistically significant meanings. Klein and Sherman estimator generally does well, although it overestimates the coefficients of the treatment variable and fails to achieve statistically significant results.

6 Applications

6.1 Application 1: Ballard-Rosa, Goldstein and Rudra (2024)

The first example replicates the analysis by Ballard-Rosa, Goldstein and Rudra (2024). The main argument of the study was that the exposure to the framing linking globalization and reduced American Dream would lower the support for trade expansion, and the effect would be larger among the citizens who believe in meritocratic values. The treatment group was shown a statement that some of the US's foreign policy reduces chances of achieving American Dreams, while control group was given that some of the US's foreign policy increases the domestic economic inequality. The outcome of interest is the view on the expansion of trade relationship, and this was measured with 5 point Likert-type choices. The authors utilized OLS and the results generally confirmed their claim; that the American Dream framing greatly reduce individual's support for trade expansion and the effect was larger if a respondent believes more in meritocratic values.

Even though the outcome variable is measured in ordinal scales, the authors analyze the data using OLS, which assumes that the outcome to be at least interval. This may lead to potential bias due to model misspecification, so I retest their main models in Table 2 (column 1 and 2) with ordered regression models. The results reported in Table 2. The first two columns show the replicated result of their original analysis and columns 3 to 6 report the ordered logit and probit model results.

To further confirm the results from the ordered logit and probit model, diagnostic test on distributional assumptions using surrogate-based residuals are employed. Figure 3 graphically depicts the tests for the distributional assumptions. The panels on the first column shows the Q-Q plot of the surrogate-based residuals and the assumed distributions and two seem to be aligned well. The panels on the second column further confirm that the violation of the distributional assumption is

Table 1: Replication of Ballard-Rosa, Goldstein and Rudra (2024) Table 2

	Original – OLS		Ordered Logit		Ordered Probit	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
American Dream Treatment	-0.05	-0.08	-0.13	-0.18	-0.06	-0.09
	(0.05)	(0.05)	(0.09)	(0.10)	(0.06)	(0.06)
meritocrat	0.04	0.06	-0.04	-0.00	-0.00	0.02
	(0.08)	(0.09)	(0.14)	(0.15)	(0.08)	(0.08)
AD Treatment x meritocrat	-0.29^{*}	-0.33^{**}	-0.44^{*}	-0.52^{*}	-0.30^{**}	-0.35^{**}
	(0.12)	(0.12)	(0.20)	(0.21)	(0.11)	(0.12)
Deviance	1929.01	1927.39	5322.79	5321.19	5327.36	5325.32
Dispersion	0.99	0.99				
Num. obs.	1954	1954	1954	1954	1954	1954
AIC			5366.79	5367.19	5371.36	5371.32
BIC			5489.49	5495.47	5494.07	5499.61
Log Likelihood			-2661.39	-2660.59	-2663.68	-2662.66

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

not an issue in these models. The p-values from 50 Kolmogorov-Smirnov tests are almost uniformly distributed in both models. However, the ordered logit model aligns more tightly with the 45° line, thus it might be the best fitted model in this case.

Need to include Klein and Sherman Results

7 Conclusion

This paper underscores the importance of scrutinizing distributional assumptions in ordinal regression models, particularly in the context of survey-based causal inference. While ordered logit and probit models have become standard tools in political science, their reliance on specific error distributions poses significant risks when those assumptions do not hold. The violation of distributional assumptions leads to the biased and inconsistent estimation which may lead to inappropriate inference. This point has been demonstrated through simulation studies and empirical analysis. Although it is not statistically meaningful, the ordered regression models show misleading signs of treatment effects in simulation settings.

To address this challenge, I introduced a diagnostic approach using surrogate residuals, offering researchers a practical means to assess the validity of distributional assumptions in ordinal models. The suggested method approximates the residuals in latent variable space in continuous manner, which enables further diagnosis using graphical tools such as Q-Q plot and standard tests for distributional similarity such as Kolmogorov-Smirnov test and Anderson-Darling test.

When diagnostics suggest substantial departures, this paper recommends a semiparametric alternative – the Klein and Sherman estimator – which avoids reliance on a fully specified error distribution and performs robustly under a wide range of conditions. Monte Carlo simulation reaffirms that Klein and Sherman estimator can outperform standard ordered regression models when distributional assumptions are violated.

These tools empower researchers to move beyond mechanical application of ordered logit and probit models and to make informed modeling choices that better reflect the structure of their data. By integrating diagnostics and semiparametric methods into the applied researcher's toolkit, this work contributes to more reliable causal inference and better empirical practice in the analysis of ordinal survey outcomes.

References

Abramson, Ian S. 1982. "On Bandwidth Variation in Kernel Estimates-A Square Root Law." *The Annals of Statistics* 10(4).

Alt, James and Torben Iversen. 2017. "Inequality, Labor Market Segmentation, and Preferences for Redistribution." *American Journal of Political Science* 61(1):21–36.

Ballard-Rosa, Cameron, Judith L. Goldstein and Nita Rudra. 2024. "Trade as Villain: Belief in the American Dream and Declining Support for Globalization." *The Journal of Politics* 86(1):274–290.

Bera, Anil K and Carlos M Jarque. 1982. "Model specification tests: A simultaneous approach." *Journal of Econometrics* 20(1):59–82.

Bloem, Jeffrey R. 2022. "How Much Does the Cardinal Treatment of Ordinal Variables Matter? An Empirical Investigation." *Political Analysis* 30(2):197–213.

Bond, Timothy N. and Kevin Lang. 2018. "The Sad Truth about Happiness Scales." *Journal of Political Economy* 127(4):1629–1640.

Canes-Wrone, Brandice and Scott De Marchi. 2002. "Presidential Approval and Legislative Success." *Journal of Politics* 64(2):491–509.

Glewwe, P. 1997. "A test of the normality assumption in ordered probit model." *Econometric Reviews* 16(1):1–19.

Greene, William H. and David A. Hensher. 2010. *Modeling Ordered Choices : A Primer*. Cambridge: Cambridge University Press.

Johnston, Carla, James McDonald and Kramer Quist. 2020. "A generalized ordered Probit model." *Communications in Statistics - Theory and Methods* 49(7):1712–1729.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon and Ajay Tandon. 2004. “Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research.” *American Political Science Review* 98:191–207.

King, Gary and Jonathan Wand. 2007. “Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes.” *Political Analysis* 15(1):46–66.

Klein, Roger W. and Richard H. Spady. 1993. “An Efficient Semiparametric Estimator for Binary Response Models.” *Econometrica* 61(2):387–421.

Klein, Roger W. and Robert P. Sherman. 2002. “Shift Restrictions and Semiparametric Estimation in Ordered Response Models.” *Econometrica* 70(2):663–691.

Kriner, Douglas and Liam Schwartz. 2009. “Partisan Dynamics and the Volatility of Presidential Approval.” *British Journal of Political Science* 39(3):609–631.

Lee, Myoung-jae. 1992. “Median regression for ordered discrete response.” *Journal of Econometrics* 51(1):59–77.

Lewbel, Arthur. 2000. “Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables.” *Journal of Econometrics* 97(1):145–177.

Li, Chun and Bryan E. Shepherd. 2010. “Test of Association Between Two Ordinal Variables While Adjusting for Covariates.” *Journal of the American Statistical Association* 105(490):612–620. PMID: 20882122.

Liu, Dungang and Heping Zhang. 2018. “Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach.” *Journal of the American Statistical Association* 113(522):845–854. PMID: 30220754.

Magni, Gabriele. 2021. “Economic Inequality, Immigrants and Selective Solidarity: From Perceived Lack of Opportunity to In-group Favoritism.” *British Journal of Political Science* 51(4):1357–1380.

Manski, Charles F. 1988. “Identification of Binary Response Models.” *Journal of the American Statistical Association* 83(403):729–738.

Mayda, Anna Maria and Dani Rodrik. 2005. “Why are some people (and countries) more protectionist than others?” *European Economic Review* 49(6):1393–1430.

Nagler, Jonathan. 1994. “Scobit: An Alternative Estimator to Logit and Probit.” *American Journal of Political Science* 38(1):230–255.

Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of educational Psychology* 66(5):688.

Scheve, Kenneth F and Matthew J Slaughter. 2001. “What determines individual trade-policy preferences?” *Journal of International Economics* 54(2):267–292.

Schroder, Carsten and Shlomo Yitzhaki. 2017. “Revisiting the evidence for cardinal treatment of ordinal variables.” *European Economic Review* 92:337–358.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. Routledge.

Smits, Niels, Oguzhan Ogreden, Mauricio Garnier-Villarreal, Caroline B Terwee and R Philip Chalmers. 2020. “A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement.” *Statistical Methods in Medical Research* 29(4):1030–1048.

Stoetzer, Lukas F., Xiang Zhou and Marco Steenbergen. 2025. “Causal inference with latent outcomes.” *American Journal of Political Science* 69(2):624–640.

Weiss, Andrew A. 1997. “Specification tests in ordered logit and probit models.” *Econometric Reviews* 16(4):361–391.

Zhou, Xiang. 2019. “Hierarchical Item Response Models for Analyzing Public Opinion.” *Political Analysis* 27(4):481–502.