Causal Inference with Ordinal Outcome: Semiparametric Approach

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

2.1 Identification Problem

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, assume following partially linear true data-generating process (DGP):

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

D is a binary treatment indicator, where 1 denotes a treatment group and 0 denotes a control group, X is a vector of regressors, and ϵ is a latent error term and its density is defined with f, and its distribution is defined by F.

Now, 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. We put following standard causal inference assumptions regarding the latent Y^*.

Assumption 1. *SUTVA A unit's potential outcomes are not affected by the treatment given to other units.* $Y_i^* = Y^*(D_i)$

Assumption 2. *Ignorability The treatment assignment is conditionally independent to the potential outcomes.* $Y_i^* \perp\!\!\!\perp D_i \mid X_i$

Assumption 3. *Positivity There is non-zero probability of being treated.* $\mathbb{P}(D_i = 1) > 0$

The target causal estimand considered in this paper is the average treatment effect (ATE) in terms of the latent outcome Y^*.

If all above assumptions are met, ATE on Y^* can be written as

$$\mathbb{E}\left[Y_i^*(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i^*(0) \mid D_i = 0\right] = \tau$$

.

The problem arises from the fact that continuous Y^* can not be observed directly, but we only can observe some (monotonic) transformation of it. Denote the observed ordinal outcome as Y and the transformation (often called reporting function) as g.

$$Y = g(Y^*)$$

One may concern the inter- and intra-personal differences in g. In other words, each respondent may use different internal process in transforming Y^* into Y. This hampers the identification, but the concern may be attenuated by careful survey design utilizing anchored vignettes suggested in (King et al., 2004; King and Wand, 2007). Throughout the paper, I assume everyone shares same g, ignoring the additional problems arising from inter- and intra-personal differences.

Assumption 4. *Uniform Monotonic Reporting function The monotonic reporting function g is invariant across units.*

g can also be thought of as a function that cuts Y^* into Y with some thresholds,

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

, where α_k denotes the threshold points for each ordinal category.

Under this settings, we only can observe Y directly, and have no good knowledge on either Y^* nor g. Thus, the naive estimator of ATE, $\mathbb{E}\left[Y^*(1)\right] - \mathbb{E}\left[Y^*(0)\right]$ cannot be obtained. Also, since

g is unknown, and scaling by positive constant would not change the ordering of Y, we only can identify the target causal parameter up to scale. In other words, we only can identify the size of the treatment effect relative to a coefficient of another covariate. Let us denote β_a as a coefficient of the covariate that is going to work as an anchor.

Definition 1. *Anchored Latent Treatment Effect* $\tau_{ALTE} = \tau_0/\beta_a$

2.2 Estimation

Regarding the setting, 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.

The first approach assigns numerical values to ordinal outcome Y. The assumption here is that the numerical labels would be a proper representation of the inverse of the unobserved reporting function, g^{-1}. Let us denote this cardinalization scheme as l

Assumption 5. *Proper Cardinalization Researchers can properly assign numerical values to the ordinal outcome Y which can then approximate the inverse of the reporting function. $l \sim g^{-1}$*

This enables researcher to use usual causal inference tools such as means-difference, least squares, and difference-in-differences, because now the numerical values are assumed to capture the latent variable Y^*. One critical downside for this approach comes from the fact that the numerical labels are arbitrary, and theoretically any labels should work if they preserve the order. In some problems, researchers may find some numerical labels that can is believed to be the true values on the latent space, but in many cases it is hard to justify one labels to the other. 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 $\{-1, 3, 15, 50, 100\}$. Thus, 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. Even worse, 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 estimated causal effect. Bloem (2022) partly deals with this problem by providing a sensitivity test and a partial identification method based on the test.

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.

Assumption 6. *Distributional Assumption on Error Term The error term, ϵ follows certain distribution. For example, ordered logit model assumes standard logistic distribution and ordered probit model assumes standard normal.*

While these assumptions facilitate fast and efficient estimation through 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 KDE-based Semiparametric Ordinal 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. This allow us to identify the ALTE only with Assumption 1 to 4 and following assumption on the distribution of covariates.

Assumption 7. *Differentiable Conditional Density Conditional density of X given (β, τ, D, Y) is strictly positive and differentiable.*

Let us denote the true log-likelihood as Q. If we can estimate Q consistently, then maximum likelihood would guarantee us consistent estimation of τ. Define the quasi-log-likelihood function $\hat{Q}$ as following:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}(X_i)\sum_{j=0}^{J}1_{\{Y=j\}}\log\left[\hat{\mathbb{P}}\left(Y\leq j \mid X_i,\beta,D_i,\tau\right)-\hat{\mathbb{P}}\left(Y\leq j-1 \mid X_i,\beta,D_i,\tau\right)\right] \tag{2}$$

where $\hat{\mathbb{P}}_{-1}=0$, $\hat{\mathbb{P}}_{J+1}=1$, $\hat{\mathbb{P}}$ is a kernel regression estimator of the CDF of the error term, and the trimming function $\hat{m}$ is to rule out the extreme data points where kernel regressor may not work poorly. The only part that needs estimation in $\hat{Q}$ is $\hat{\mathbb{P}}$, so we now focus on estimation of it.

For notational ease, let's denote $f(X,\beta)+D^T\tau$ as V. By the Bayes' rule, $\mathbb{P}\left(Y\leq j \mid V\right))$ can be

expressed as:

$$\mathbb{P}\left(Y \leq j \mid V)\right) = \frac{\mathbb{P}\left(Y \leq j\right) \times g_1(V \mid Y \leq j)}{\mathbb{P}\left(Y \leq j\right) \times g_1(V \mid Y \leq j) + \mathbb{P}\left(Y > j\right) \times g_0(V \mid Y > j)}$$

, where $g_0(\cdot)$ and $g_0(\cdot)$ denote conditional density of V given $(Y > j)$ or $(Y \leq j)$ respectively.

While both $\mathbb{P}\left(Y \leq j\right)$ and $\mathbb{P}\left(Y > j\right)$ can be approximated as a sample probability, we suggest to utilize kernel density estimation (KDE) in estimating $g_1(\cdot)$ and $g_0(\cdot)$. KDE is a non-parametric method to estimate the probability density function of a random variable based on selected kernel and bandwidth. Bandwidth selection is an important step in KDE, because both estimation quality and convergence rate are depend on bandwidth. In this paper, slightly modified version of the local smoothing bandwidth developed by Abramson (1982) and Silverman (1986) is used. As an overview, this bandwidth selection method calculates weights for each data points based on pilot KDE and the final KDE bandwidth are derived based on the weights. This process helps reducing bias in estimation.

All suggestions on the pilot bandwidth, weights function parameters, and final bandwidth are to guarantee the consistency of the $\hat{Q}$. To be more specific, $\hat{\mathbb{P}}$ should converge to the true $\mathbb{P}$ fast enough to ensure the performance in small sample, but also must not go too fast that the derivatives of it vanishes. If $\hat{P}$ converges too fast, then the derivatives may not exist which make it harder to establish the consistency of the $\hat{Q}$ based on maximum likelihood.

The estimation of $g_1(\cdot)$ starts with pilot density estimation. Let $\hat{\sigma}_1$ be the sample standard deviation of the $V \cdot 1_{\{Y_k \leq j\}}$ and $n_1 = \sum_k 1_{\{Y_k \leq j\}}$. 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 $g_1(\cdot)$ to be:

$$\hat{\pi}_{1i} = \frac{1}{n_1} \sum_{k \neq i} \left[1_{\{Y_k \leq j\}} \cdot K\left(\frac{V_i - V_k}{\hat{\sigma}_1 h_p} \right) \cdot \frac{1}{\hat{\sigma}_1 h_p} \right]$$

This is essentially same as estimating the conditional density of $\mathbb{P}\left(V \mid Y \leq j\right)$ based on the sub-sample of V_k, where $(Y_k \leq j)$.

Next, based on the pilot density weights for the final KDE bandwidth are calculated for each data points. These weights control the rate at which final bandwidth converges to 0 as $n \to \infty$.

In other words, the weight functions can stretch the final bandwidth for some data points if data points around them are relatively scare area and thus need more smoothing (i.e. tail area), but at the same time prevent the final bandwidth from being too wide.

First, estimated 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 $\hat{\pi}_{1i}$. This widens the final bandwidth in regions where the estimated pilot density is relatively low. This can be especially helpful in capturing tail regions.

To control the bandwidth to be too wide and thus converges too fast, define the estimated damping function, $\hat{d}_1$ for $l > 0$ which approximates the indicator function $1_{\{l > a_{n_1}\}}$. Set $a_{n_1} \propto [\ln n_1]^{-1}$ and choose $\epsilon \in (0, \frac{1}{40} - \frac{\delta}{20})$ gives:

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

Lastly, calculate the weights for the final KDE, w_{1i} with parameters above:

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

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

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

$\hat{g}(\cdot)$ is defined analogously, replacing $Y_k \le j$ with $Y_k > j$. Note again that the final kernel estimation of $g_1(\cdot)$ is depend on the subsample of V where $Y \le j$.

Even if we used local smoothing technique in estimation of the $\hat{\mathbb{P}}$, there is still a concern that the some edge cases of V may drag the estimation. In this situation trimming out some extreme data points can be a solution. Trimming function $\hat{m}$ does this by trimming out the data points

whose absolute value are not within the range of qth quantile of $|X_i|$.

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

The final quasi-likelihood function in Equation (2) can now be constructed with above estimated values and the trimming function.

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 $\mathbb{P}\left(\hat{Y} \leq j\right)$. 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).

4 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:

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

, 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 $\mathcal{N}(0, 5)$ 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 standard normal distribution $\mathbb{N}(0, 1)$ and the t-distribution with degrees of freedom of 1. The t-distribution has heavier tail regions than both the standard logistic and the standard normal distribution. Four sample sizes of 250, 500, 750 and 1,000 are considered. I first generate a population of size $1e+5$ and draw random samples for Monte Carlo replication from this population. The number of Monte Carlo replication is 1,000, and all estimations are the means of these 1,000 replication results. Regarding the Klein and Sherman estimator, I set $\delta = \frac{1}{8}$ and fix other parameters to the middle points of the suggested intervals. For instance, the final bandwidth, h_f for $n = 250$ is selected as $n^{\frac{\frac{(3+\delta)}{20} + \frac{1}{6}}{2}}$

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.

5 Applications

5.1 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

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

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

panels on the second column further confirm that the violation of the distributional assumption is 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

6 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.