Regression and decomposition with ordinal health outcomes

Qian Wu*† David M. Kaplan‡

April 3, 2025

Abstract

Although ordinal health outcome values are categories like "poor" health or "moderate" depression, they are often assigned values $1, 2, 3, \ldots$ for convenience. We provide results on interpretation of subsequent analysis based on ordinary least squares (OLS) regression. For description, unlike for prediction, the OLS estimand's interpretation does not require that the $1, 2, 3, \ldots$ are cardinal values: it is always the "best linear approximation" of a summary of the conditional survival functions. Further, for Blinder–Oaxaca-type decomposition, the OLS-based estimator is numerically equivalent to a certain counterfactual-based decomposition of the survival function, again regardless of any cardinal values. Empirically, with 2022 U.S. data for working-age adults, we estimate a higher incidence of depression in the rural population, and we decompose the rural–urban difference. Including a nonparametric estimator that we describe, estimators agree that 33–39% of the rural–urban difference is statistically explained by income, education, age, sex, and geographic region. The OLS-based detailed decomposition shows this is mostly from income.

JEL classification: C25, I14

Keywords: Blinder–Oaxaca decomposition; Counterfactual distribution; Distribution regression; Survival function

*School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China; wuqj@swufe.edu.cn

†Big Data Laboratory on Financial Security and Behavior, SWUFE (Laboratory of Philosophy and Social Sciences, Ministry of Education), Chengdu 611130, China

‡Corresponding author; Department of Economics, University of Missouri, 615 Locust Street, Columbia, MO, USA; kaplandm@missouri.edu

1 Introduction

Ordinal variables are common in health economics. Such variables take values that are not numeric but rather categorical. For example, self-reported health status often takes values "poor," "fair," "good," "very good," and "excellent," which have an order from lowest to highest, but no numeric value. Mental health variables are also often ordinal. Our empirical analysis uses a measure of depression with values "none/minimal," "mild," "moderate," and "severe." Some variables are even coded with numeric values, but upon examination these values do not have a cardinal but merely ordinal meaning, like the Apgar score for newborns whose numbers are based on underlying categories like "no cyanosis" or "some flexion."

Even in raw data, ordinal outcome variables often come already coded with numeric values $1, 2, 3, \ldots$, making it easy to run ordinary least squares (OLS) regression and related analyses, but this raises questions about interpretation. For example, in the 2022 National Health Interview Survey (NHIS) data, we use the depression measure variable `PHQCAT\_A` with the categories listed above, but they are coded as numerical values $1, 2, 3, 4$. Ordinal variables do not even have a well-defined mean, because values like "mild" and "severe" cannot be summed or averaged. This seemingly suggests that in order for OLS results to be meaningful, we must interpret the $1, 2, 3, \ldots$ coding as cardinal values assigned to the respective categories. If those are indeed cardinal values, then we actually have a cardinal variable, so methods like OLS can be run and interpreted as usual. But what if those are not the true cardinal values?

Aspects of this general question have been addressed by several papers in health economics that take seriously the ordinal nature of such outcomes. These papers specifically consider measuring health inequality or polarization given an ordinal health outcome variable. For example, Allison and Foster (2004), Apouey (2007), Abul Naga and Yalcin (2008), and Kobus and Miłoś (2012) all agree that such measures should be "scale invariant" in the sense of not depending on whether we code the categories with cardinal values $1, 2, 3$ or $1, 2, 10$ or $1, 7, 8$, etc. The median-preserving spread of Allison and Foster (2004) and the

inequality indices proposed and studied by Apouey (2007), Abul Naga and Yalcin (2008), and Kobus and Miłoś (2012) can all be interpreted without any specific cardinal values, although it is assumed that each ordinal category corresponds to a single cardinal value. Building on the idea that first-order stochastic dominance provides a scale-invariant partial ordering of ordinal health distributions, Makdissi and Yazbeck (2017) incorporate a measure of socioeconomic status to determine when one population is "better" than another (in terms of a concentration or achievement index) robust to any possible scale (Theorem 1), and they refine their results in the case of shape restrictions on the scale (Section 4.2). Instead of the above papers' assumption that each category corresponds to a single cardinal value, Kaplan and Zhao (2023) model a continuous latent health distribution, so each ordinal category can correspond to a range of cardinal values. In that setting, they provide identification and inference results related to measures of both between-group and within-group health inequality. Similarly, our results are robust not only to any "scale" that assigns a single cardinal value to each ordinal category, but also to the lack of any such scale, as with a continuous latent health distribution. To emphasize this and the ordinal/cardinal distinction, we call such results *cardinalization-robust*.

In the same spirit, we make five contributions to the cardinalization-robust interpretation of OLS regression and OLS-based Blinder–Oaxaca decomposition with an ordinal outcome, looking toward our empirical decomposition of rural–urban mental health differences. First, for prediction, we have a negative result: the best predictor of Y is not cardinalization-robust, so there is no cardinalization-robust "best linear predictor" interpretation of OLS. That is, if we code the Y values as $1, 2, 3, \ldots$, then the interpretation of the OLS estimand as the best linear predictor crucially depends on $1, 2, 3, \ldots$ being true cardinal values (or an affine transformation thereof). Second, for description, the OLS estimand with Y coded $1, 2, 3, \ldots$ can be interpreted as the best linear approximation of the sum (across all Y categories) of conditional survival function values. This is explained in more detail in Section 3.3, but the main point is that unlike with prediction, the $1, 2, 3, \ldots$ coding here is innocuous: the OLS

interpretation does not require those values to have any cardinal meaning.

Third, as a practical contribution for ordinal decomposition, we describe how to apply the methodology of Chernozhukov et al. (2013) to get a cardinalization-robust decomposition of the survival function difference. This approach is based on a counterfactual distribution that combines the marginal $\boldsymbol{X}$ distribution of one group with the conditional distribution (of Y given $\boldsymbol{X}$) from the other group. This is essentially a generalization of the binary outcome decomposition of Fairlie (2005) that traces back at least to Even and Macpherson (1990) and Farber (1987); it is also more flexible than the ordered probit/logit decomposition of Bauer and Sinning (2008). We also describe how to implement a specific nonparametric version of this approach that we apply in our empirical analysis.

Fourth, again with Y coded as $1, 2, 3, \ldots$, we derive a numerical equivalence between two very different estimators: the OLS-based Blinder–Oaxaca "mean" decomposition, and the cardinalization-robust counterfactual survival function decomposition based on Chernozhukov et al. (2013) when using OLS to estimate linear probability models for the distribution regression step. Thus, despite seeming like it relies on $1, 2, 3, \ldots$ as cardinal values, the OLS-based Blinder–Oaxaca decomposition actually has a cardinalization-robust interpretation. Among other practical implications, this means we can reinterpret previously published Blinder–Oaxaca results in terms of a robust counterfactual survival function decomposition. It also means that going forward we can more confidently use OLS-based Blinder–Oaxaca, and more appropriately interpret its results, with only the caveat that other estimators may reduce functional form misspecification. The Blinder–Oaxaca approach also readily produces a "detailed decomposition" that shows how much of the overall difference is statistically explained by each variable individually.

Fifth, we empirically examine the mental health disparity between rural and urban groups in the U.S., decomposing the distributional difference in a measure of depression. Depression is important to study due to its large aggregate effects both personally and economically. For example, the annual total economic burden in the U.S. is estimated in the hundreds of billions

of dollars (Greenberg et al., 2021), and our Section 6.1 discusses several papers specifically on rural–urban mental health differences in the U.S., among others on closely related topics. In our NHIS 2022 sample, focusing on working-age adults, we find higher incidence of depression in rural residents, in the sense of first-order stochastic dominance, which itself contributes to a sparse literature with mixed results. For the decomposition, our explanatory variables are education, age, sex, income, and geographic region. Our various estimators attribute 33–39% of the depression difference to these variables. That is, these variables explain a substantial amount, but still leave a majority of the rural–urban difference unexplained. We include a nonparametric estimator that performs model selection among millions of candidate models, as we describe in detail in Section 6.3. Using the Blinder–Oaxaca approach, we report a detailed decomposition showing that income explains much more than any other covariate. Even though we do not believe that the depression categories correspond to cardinal values $1, 2, 3, 4$, our equivalence result implies that this detailed decomposition is still meaningful.

We include results for Blinder–Oaxaca-type decomposition because of its widespread use and importance. It is commonly used to decompose an overall mean difference in outcome between two groups into two components: one attributed to the group difference in explanatory variable means, and the other to differences in regression coefficients. Using the same idea published by Kitagawa (1955) and used earlier in the 1940s (see her footnote 3), the papers of Blinder (1973) and Oaxaca (1973) have over 20,000 citations in Google Scholar, with over 6000 of those coming since 2019, spanning the fields of economics, public health, sociology, medicine, demography, and others. Some examples in health include decomposing differences in various biomarkers by gender (Carrieri and Jones, 2017), self-reported health by age (Idler and Cartwright, 2018), various health outcomes by education or income (Kino and Kawachi, 2020), diabetes by Latinx identity (Cartwright, 2021), and obesity/BMI by race (Sen, 2014).

Despite the importance of both decomposition and ordinal variables, there is a limited literature on decomposition with ordinal outcomes. The extensive *Handbook of Labor Eco-*

nomics chapter on "Decomposition Methods in Economics" (Fortin et al., 2011) includes discussion of many population functionals and estimators and causal identification, but does not include the word "ordinal" anywhere in its 102 pages. (And "ordered" only appears in the context of parametric estimation of conditional distributions for a continuous outcome after "discretizing the outcome variable" (p. 70).) Some empirical work simply reduces the ordinal variable to a binary variable before doing a probit-based decomposition; for example, see Zhang et al. (2015, eqn. (7)) and Hauret and Williams (2017, p. 217). Although not "wrong," such simplification loses information and precision. Bauer and Sinning (2008) propose an ordered probit/logit decomposition, but it is used only to introduce nonlinearity while still treating the ordinal outcome as if had cardinal values $1, 2, 3, \ldots$, as seen in their equations on page 200. The same is true of Demoussis and Giannakopoulos (2007).[1] Similarly, empirical work often takes $1, 2, 3, \ldots$ as cardinal values and then runs the standard OLS-based Blinder–Oaxaca decomposition; for example, see Pan et al. (2015, §2.4), Awaworyi Churchill et al. (2020, §§2.1–2.2), Idler and Cartwright (2018), and Pilipiec et al. (2020, §2.2). Madden (2010, §2) acknowledges the cardinalization is not fully appropriate, yet his robustness check's ordered probit decomposition still uses the same cardinalization (p. 111). However, recall that our new results say that the $1, 2, 3, \ldots$ coding in all the above-cited work actually has a cardinalization-robust interpretation in terms of survival functions. That is, all the above methods and results are still valid, just with a somewhat different interpretation that we detail in our results.

One important limitation of using self-reported ordinal health outcomes is the potential for reporting heterogeneity, meaning that individuals with the same objective health condition may report different health statuses. This reporting difference could be due to differences in perception, cultural factors, or justification bias (Bound, 1991; King et al., 2004). For example, individuals from different socioeconomic backgrounds or geographic regions may interpret response categories differently, leading to biased estimates of health

[1] Their [7] and [8] have an important typo: the left-hand sides should have expectations of S rather than probabilities, as is clear from the text ("expected JS") and the right-hand sides, and equation [9] later.

inequalities (Kapteyn et al., 2007). Such discrepancies can undermine the validity of empirical results based on self-reported data, as they may reflect differences in reporting behavior rather than actual disparities in health outcomes. One well-established approach to addressing this issue is the use of anchoring vignettes, which provide respondents with hypothetical health scenarios to standardize response scales across individuals (King et al., 2004). This method has been widely applied in health economics to adjust for systematic differences in self-reported health measures (Bago d'Uva et al., 2008). Additionally, incorporating objective health indicators, such as biomarkers and clinical diagnoses, can help gauge the extent of reporting differences across the groups of interest (Jürges, 2007). In future research, it would be valuable to extend our framework by integrating such approaches when the ordinal health outcome is self-reported rather than clinically reported.

The remainder of this paper is organized as follows. Section 2 discusses the special case with a binary outcome. Section 3 characterizes the interpretation of OLS regression with an ordinal outcome, both in terms of prediction and description. Section 4 describes a framework for ordinal decomposition based on the counterfactual approach of Chernozhukov et al. (2013), as well as our new equivalence between "mean" and survival function decomposition. Section 5 describes estimation and inference, as well as our second equivalence result that provides a meaningful, robust interpretation for the naive OLS-based Blinder–Oaxaca decomposition. Section 6 contains our empirical contributions on rural–urban mental health disparities in the U.S. Appendix A contains an additional theoretical proof.

2 Special case: binary outcome

To build intuition for our approach to ordinal outcomes, we first consider the more familiar special case where the health outcome is binary. Specifically, outcome variable $Y \in \{0, 1\}$, with covariate vector $\boldsymbol{X}$. In practice, sometimes binary variables are generated from ordinal variables by grouping the lower categories as $Y = 0$ and the higher categories as $Y = 1$. For

example, $Y = 0$ could be low/no depression (none, minimal, or mild), and $Y = 1$ depression (moderate or severe), or $Y = 0$ could be "bad" general health (poor or fair) with $Y = 1$ "good" health (good, very good, or excellent).

We use the following notation. Uppercase denotes a random variable (scalar) like Y or a random vector like $\boldsymbol{X}$, whereas lowercase y and $\boldsymbol{x}$ are non-random scalar and vector values, respectively. The indicator function is $\mathbb{1}\{\cdot\}$: $\mathbb{1}\{A\} = 1$ if event A occurs, and $\mathbb{1}\{A\} = 0$ if not. For random variable Y, its cumulative distribution function (CDF) and survival function evaluated at value y are $F_Y(y) \equiv \mathrm{P}(Y \leq y)$ and $S_Y(y) \equiv \mathrm{P}(Y > y)$, respectively.

2.1 Identities, description, and normative properties

There are some convenient identities in this familiar special case. The mean is $\mathrm{E}(Y) = (0)\,\mathrm{P}(Y = 0) + (1)\,\mathrm{P}(Y = 1) = \mathrm{P}(Y = 1)$. This can further be written in terms of the CDF: $\mathrm{P}(Y = 1) = 1 - \mathrm{P}(Y \leq 0) = 1 - F_Y(0)$. More conveniently, we can use the survival function: $\mathrm{P}(Y = 1) = 1 - F_Y(0) = S_Y(0)$. This also holds conditional on any $\boldsymbol{X} = \boldsymbol{x}$. Altogether,

$$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = S_Y(0 \mid \boldsymbol{X} = \boldsymbol{x}). \tag{1}$$

That is, if we code our "low" and "high" binary variable with values $Y = 0$ and $Y = 1$, then the conditional "mean" function $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ equals the conditional survival function evaluated at zero, $S_Y(0 \mid \boldsymbol{X} = \boldsymbol{x})$. Although $\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ relies on Y to have cardinal values, the survival function does not: the notation $S_Y(0 \mid \boldsymbol{X} = \boldsymbol{x})$ is equivalent to $\mathrm{P}(\text{high} \mid \boldsymbol{X} = \boldsymbol{x})$, where "high" is the label of the higher category. That is, the conditional "mean" function has a cardinalization-robust interpretation. We extend these identities and interpretations from binary to general ordinal outcomes in Lemma 1 and Theorem 3.

Beyond description, the survival function value $S_Y(0)$ also has some normative properties relevant for policy decisions. Assume the higher category corresponds to a better health outcome; if not, simply replace "better" with "worse" below. Intuitively, with only two categories, one distribution is "better" if it has a higher probability of the high category

(higher $S_Y(0)$), which mechanically implies lower probability of the low category, too. More formally, higher $S_Y(0)$ corresponds to first-order stochastic dominance. Such a dominance relationship has traditionally been interpreted as an unambiguous indication that one distribution is better than another (e.g., Allison and Foster, 2004), although this is not necessarily true when interest is in an underlying latent distribution from which the ordinal categories are generated. However, even then, ordinal first-order stochastic dominance provides some evidence of the dominant latent distribution's superiority in terms of sets of quantiles being higher; see Theorem 2.2 of Kaplan and Zhao (2023). For example, consider a latent continuous measure of mental health, and the high category ($Y = 1$) corresponds to mental health above a given threshold. If one population has $S_Y(0) = 0.53 = 53\%$ in the high category, and a second population has $S_Y(0) = 0.49 = 49\%$ in the high category, then clearly we cannot conclude that the first latent distribution dominates the second, but we do know the first population's median is higher because it is higher than the high/low threshold, whereas the second population's median is below the threshold.

2.2 Prediction

Consider trying to predict if an individual has "low" health ($Y = 0$) given their covariate vector $\boldsymbol{X}$. To define optimal prediction, we first need to specify a loss function that quantifies how bad it is to guess value g when the true value is y. Conventionally (and without loss of generality), loss functions are normalized to set $L(y, g) = 0$ if $y = g$, and $L(y, g) > 0$ if $y \neq g$. If $Y = 0$ and $Y = 1$ are appropriate cardinal values for the low and high categories, then the mean $\mathrm{E}(Y) = S_Y(0)$ is the best prediction in the sense of minimizing the mean squared prediction error. That is, $\mathrm{E}(Y) = \arg\min_{g \in \mathbb{R}} \mathrm{E}[(Y - g)^2]$, which can also be interpreted as minimizing expected loss $\mathrm{E}[L(Y, g)]$ given the quadratic loss function $L(y, g) = (y - g)^2$. However, if "low" and "high" are categories that do not have any corresponding cardinal value, then it is nonsense to predict something like 0.62 when the options are "low" and "high." In that case, we must specify two non-zero values: $L_{01} \equiv$

$L(0,1)$ (short for $L(\text{low}, \text{high})$) and $L_{10} \equiv L(1,0)$. The expected loss if we predict $g = 0$ (low) is $\mathrm{E}[L(Y,0)] = L_{10}\,\mathrm{P}(Y = 1)$, and expected loss if we predict $g = 1$ (high) is $\mathrm{E}[L(Y,1)] = L_{01}\,\mathrm{P}(Y = 0) = L_{01}[1 - \mathrm{P}(Y = 1)]$, so the optimal prediction that minimizes expected loss is $g^* = \mathbb{1}\{\mathrm{P}(Y = 1) > L_{01}/(L_{01} + L_{10})\}$ (e.g., Kaplan, 2023, §14.3.2). Because $\mathrm{P}(Y = 1) = S_Y(0)$, this optimal prediction depends on the survival function, and similarly the best prediction of Y given $\boldsymbol{X}$ depends on the conditional survival function. However, the conditional survival function itself is not the best prediction, nor is there a scalar summary of the survival function that determines the best prediction in the general ordinal case.

2.3 Decomposition

Given the identities in Section 2.1, decomposition with a binary outcome coded $Y \in \{0,1\}$ can be interpreted meaningfully even if the 0 and 1 do not represent cardinal values. That is, decomposing the "mean" of Y is equivalent to decomposing the survival function evaluated at the lower category, $S_Y(0)$. Indeed, binary decomposition has been proposed by Fairlie (2005), building on even earlier work Even and Macpherson (1990); Farber (1987). For example, given two groups A and B, the "mean" difference in the binary outcome can be decomposed using the conventional Blinder–Oaxaca framework:

$$\mathrm{E}(Y^A) - \mathrm{E}(Y^B) = \overbrace{\mathrm{E}(\boldsymbol{X}^A)(\boldsymbol{\beta}^A - \boldsymbol{\beta}^B)}^{\text{unexplained}} + \overbrace{[\mathrm{E}(\boldsymbol{X}^A) - \mathrm{E}(\boldsymbol{X}^B)]\boldsymbol{\beta}^B}^{\text{explained}}.$$

Here, the first term is the "unexplained" component due to differences in coefficients $\boldsymbol{\beta}$, and the second term is the "explained" component due to differences in characteristics $\boldsymbol{X}$. Given (1), the left-hand side further equals $S_Y^A(0) - S_Y^B(0)$, so we can interpret the right-hand side as a cardinalization-robust decomposition of the difference in survival functions evaluated at the lower category.

3 Ordinal regression

Extending Section 2, we consider ordinary least squares (OLS) regression with a general ordinal outcome variable Y whose J categories have been assigned the numeric values $1, 2, \ldots, J$, respectively. In particular, we are interested in the OLS estimand's interpretation when these values do *not* represent cardinal values.

An equivalence for the "mean" is given first, followed by OLS results for both prediction and description. In later sections, we extend these results to Blinder–Oaxaca decomposition.

3.1 Mean

Although an ordinal random variable Y does not have a mean, if we assign the numeric values $1, \ldots, J$ to its J categories, then we can compute a "mean." Because we do not consider any other numeric assignments, throughout the paper we write this "mean" as $\mathrm{E}(Y)$. Lemma 1 shows that this "mean" has a meaningful, cardinalization-robust interpretation.

Lemma 1. *Given ordinal random variable Y, if we assign the numeric values $1, \ldots, J$ to its J categories, then its "mean" is $1 + \sum_{j=1}^{J-1} \mathrm{P}(Y > j)$.*

Proof. The "mean" is

$$
\begin{aligned}
\mathrm{E}(Y) &= \sum_{j=1}^{J} j\,\mathrm{P}(Y = j) \\
&= \mathrm{P}(Y = 1) + 2\,\mathrm{P}(Y = 2) + \cdots + J\,\mathrm{P}(Y = J) \\
&= \overbrace{[\mathrm{P}(Y = 1) + \mathrm{P}(Y = 2) + \cdots + \mathrm{P}(Y = J)]}^{=1} \\
&\quad + [\overbrace{\mathrm{P}(Y = 2) + \cdots + \mathrm{P}(Y = J)}^{=\mathrm{P}(Y>1)}] \\
&\quad + \cdots \\
&\quad + [\overbrace{\mathrm{P}(Y = J)}^{\mathrm{P}(Y>J-1)}]
\end{aligned}
$$

$$= 1 + \sum_{j=1}^{J-1} \mathrm{P}(Y > j). \qquad \square$$

Lemma 1 shows that the "mean" can be interpreted in terms of the survival function, which does not depend on any cardinal value assignment. The summand $\mathrm{P}(Y > j)$ is the survival function of Y evaluated at category j, $S_Y(j)$. Alternatively, the final expression could be rewritten in terms of the CDF as $J - \sum_{j=1}^{J-1} \mathrm{P}(Y \leq j)$. However, the survival function expression makes it more directly clear that higher values correspond to higher probabilities of higher-valued categories. For example, if one ordinal distribution first-order stochastically dominates another, then it has a higher survival function at all j, and thus has higher "mean" of $1 + \sum_{j=1}^{J-1} \mathrm{P}(Y > j)$.

The "1+" in Lemma 1 disappears if the scale is changed from $1, \ldots, J$ to $0, \ldots, J-1$. Letting $Z \equiv Y - 1$, then $\mathrm{E}(Z) = \mathrm{E}(Y-1) = \mathrm{E}(Y) - 1 = \sum_{j=1}^{J-1} S_Y(j) = \sum_{j=0}^{J-1} S_Z(j)$, where the second equality uses the linearity property of expectation, the third equality uses Lemma 1, and the fourth equality uses $\mathrm{P}(Y > j) = \mathrm{P}(Z + 1 > j) = \mathrm{P}(Z > j - 1)$. The scale $0, \ldots, J-1$ also more directly extends the binary scale $0, 1$ from Section 2. Thus, it is somewhat "better" to use $0, \ldots, J-1$ than $1, \ldots, J$, but we focus on $1, \ldots, J$ because that is the default coding in most datasets and the most commonly used in practice.

3.2 Prediction

With a cardinal-valued Y, it is well known that the mean provides the best predictor given a quadratic loss function:

$$\mathrm{E}(Y) = \arg\min_{g} \mathrm{E}[(Y - g)^2].$$

More generally, but still with quadratic loss, the conditional mean is the best predictor of Y given vector $\boldsymbol{X}$ (e.g., Wooldridge, 2010, Prop. CE.8, p. 32). It is also well known that the OLS estimand $\boldsymbol{\beta}$ is the "best" linear predictor (e.g., Wooldridge, 2010, Prop. LP.5, p. 35) in

the sense of

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathrm{E}[(Y - \boldsymbol{X}'\boldsymbol{b})^2].$$

Despite Lemma 1, we cannot derive any such best predictor results with an ordinal Y. This is true even for the simplest case of the unconditional mean. First, if each category does not have a corresponding cardinal value, then a prediction like 2.38 is nonsense, as it does not even correspond to one of the categories. Second, even if each category has a cardinal value and quadratic loss is used, the best predictor is sensitive to these values, as seen in the following simple example. Let $\mathrm{P}(Y = y) = 1/5$ given cardinal values $y = 1, 2, 3, 4, 5$. The true mean is thus $\mathrm{E}(Y) = 3$. If the cardinal values are really $1, \ldots, 5$, then this is indeed the best predictor of Y. However, imagine the cardinal values are instead $1, 2, 3, 4, 10$. In that case, the true mean is $(1 + 2 + 3 + 4 + 10)/5 = 4$, so the best predictor is 4. Thus, there is no cardinalization-robust interpretation of the "mean" in terms of prediction. We cannot generally interpret a nonparametric regression's conditional "mean" estimand as the best predictor of Y given $\boldsymbol{X}$, and we cannot interpret the OLS estimand as the best linear predictor.

Theorem 2. *If Y is an ordinal random variable, then generally there is no cardinalization-robust best predictor of Y, best predictor of Y given $\boldsymbol{X}$, or best linear predictor of Y given $\boldsymbol{X}$.*

Proof. The counterexample in the text preceding Theorem 2 shows the best predictor of Y is not invariant to the cardinal values of the categories even when such cardinal values are assumed to exist. Given scalar $X = 1$, this is also a special case of the best predictor of Y given $\boldsymbol{X}$ and the best linear predictor of Y given $\boldsymbol{X}$. □

3.3 Description

Unlike with prediction, the OLS estimand can be interpreted descriptively without any cardinalization. Lemma 1 immediately generalizes to the conditional "mean" function

$$m(\boldsymbol{x}) \equiv \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = 1 + \sum_{j=1}^{J-1} \mathrm{P}(Y > j \mid \boldsymbol{X} = \boldsymbol{x}). \tag{2}$$

That is, when coding Y categories as $1, \ldots, J$, we can interpret the conditional "mean" as a sum of conditional survival functions, analogous to the unconditional case. This conditional "mean" function is the estimand of a nonparametric regression of Y (coded $1, \ldots, J$) on $\boldsymbol{X}$.

When Y is cardinal-valued, it is well known that the OLS estimand $\boldsymbol{\beta}$ (the vector of linear projection coefficients) can be interpreted as the best linear approximation of the conditional mean function, with "best" again in terms of quadratic loss (e.g., Hansen, 2022, §2.25):

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathrm{E}\{[m(\boldsymbol{X}) - \boldsymbol{X}'\boldsymbol{b}]^2\}. \tag{3}$$

When Y is ordinal, the conditional "mean" $m(\boldsymbol{X})$ in (3) has the cardinalization-robust interpretation given in (2). Thus, the OLS estimand $\boldsymbol{X}'\boldsymbol{\beta}$ can be interpreted as the best linear approximation (in the mean squared error sense) of the sum of conditional survival functions in (2).

Theorem 3 shows that running OLS with an ordinal Y coded with values $1, \ldots, J$ yields a meaningful interpretation even if we do not believe the $1, \ldots, J$ represent cardinal values.

Theorem 3. *Let Y be an ordinal random variable whose J categories are assigned numeric values $1, \ldots, J$. Then, even if these do not represent cardinal values: a) the estimand of a nonparametric regression of Y on $\boldsymbol{X}$ can be written in terms of the conditional survival function as in* (2)*; b) for OLS regression of Y on $\boldsymbol{X}$, the population estimand $\boldsymbol{X}'\boldsymbol{\beta}$ is the best linear approximation in the sense of* (3)*.*

Proof. Combine (2) and (3). □

3.4 Normative properties

Extending Section 2.1, we discuss normative properties of the survival function sum $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$ that first appeared in Lemma 1 and continues to be a focus in later sections. Although our goal is not to advocate for its use in policy decisions, but rather to use it to properly interpret OLS-based analysis with the $1, \ldots, J$ coding common in practice, its normative properties are helpful to understand. Below, we assume higher Y is better; if not, simply replace "better" with "worse" below.

Like before, first-order stochastic dominance (FOSD) provides a helpful reference point. If we assume each ordinal category corresponds to a single cardinal value, then regardless of the values chosen, FOSD implies a higher mean (expected utility); for example, see Theorem 1 of Allison and Foster (2004). However, often this assumption does not seem realistic. For example, in our empirical study of depression, rather than assign a single cardinal value to everyone in the "moderate" depression category, it seems more realistic to assume that such individuals represent a continuum of latent depression severity. As noted in Section 2.1, if we imagine the ordinal outcome is reported based on a latent continuous health outcome, then ordinal FOSD does not imply latent FOSD, so it unfortunately does not provide an unambiguous ranking. However, ordinal FOSD does provide evidence that the dominant distribution is better with respect to a certain set of quantiles, without any such evidence in the opposite direction (Kaplan and Zhao, 2023, Thm. 2.2). Thus, although not totally definitive, FOSD still provides cardinalization-robust evidence of one distribution being better than another.

If FOSD holds, then $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$ represents some measure of the intensity of FOSD. The familiar CDF characterization of FOSD is that the dominant distribution has lower $F_Y(j)$ at each j, which is equivalent to higher $S_Y(j)$ at each j. In the context of comparing groups (populations) A and B, this means A dominates B if for each $j = 1, \ldots, J-1$, $S_Y^A(j) \geq S_Y^B(j)$ or equivalently $S_Y^A(j) - S_Y^B(j) \geq 0$. Summing such differences thus gives

some sense of the intensity of FOSD, and due to linearity of summation we can write

$$\sum_{j=1}^{J-1} S_Y^A(j) - S_Y^B(j) = \sum_{j=1}^{J-1} S_Y^A(j) - \sum_{j=1}^{J-1} S_Y^B(j),$$

the difference of the same survival function sum in Lemma 1. Even if there is not total FOSD and the term $S_Y^A(j) - S_Y^B(j)$ is negative for some j, but there is not too much deviation from FOSD, this may still be a reasonable summary. That said, we do not claim any decision-theoretic justification here, but merely show there is some interpretation regarding FOSD intensity of $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$, the cardinalization-robust object that implicitly arises when OLS-based methods are run with the $1, 2, 3, \ldots$ coded Y.

Interestingly, the expression $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$ is used by Jones et al. (2014, p. 527) for "normative evaluation" (p. 526) of ordinal self-assessed health distributions under different educational policy regimes, in the context of (in)equality of opportunity. However, they caution that "there is no unique way of doing this" (p. 526), and the motivation seems to be the integral of the survival function of a continuous (and implicitly cardinal) variable as in their (1), which relies on information about the horizontal axis scale that does not exist for ordinal variables. Nonetheless, they judged $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$ to be reasonable enough to be a policy-relevant metric.

If normative comparisons are the focus, then it may be more helpful to use methods to assess evidence regarding "stochastic monotonicity," meaning that as X increases, the ordinal Y gets "better" in the sense of FOSD. We provide such methods in other work (Wu and Kaplan, 2025). Meanwhile, we turn to decomposition in Section 4, where decomposing $\sum_{j=1}^{J-1} \mathrm{P}(Y > j)$ provides insights regardless of normative properties, and working toward the results most relevant for our empirical decomposition of rural–urban depression differences in the U.S.

4 Ordinal decomposition: framework and estimands

Turning attention to decomposition, this section introduces the counterfactual distribution framework used for both our practical and theoretical contributions. We use the framework of Chernozhukov et al. (2013), adapting their formulas to ordinal outcomes. Then, building on Lemma 1, we show how a naive "mean" decomposition is equivalent to a cardinalization-robust survival function decomposition.

4.1 Counterfactual distribution framework

First, we introduce notation for the main variables and functions. Ordinal outcome Y is a random variable with underlying categorical values like "low," "medium," and "high," labeled as $1, 2, \ldots, J$. This is partly for notational convenience, for example writing $\mathrm{P}(Y > 1)$ instead of $\mathrm{P}(Y > \text{low})$, and like before we will also consider methods that treat the $1, \ldots, J$ labels as cardinal values. Covariate vector $\boldsymbol{X}$ is a random vector including an intercept and other explanatory variables. Cumulative distribution functions (CDFs) have subscripts of the corresponding random variables: $F_Y(\cdot)$ for the CDF of Y, $F_{\boldsymbol{X}}(\cdot)$ for the CDF of $\boldsymbol{X}$, and $F_{Y|\boldsymbol{X}}(\cdot \mid \boldsymbol{x})$ for the conditional CDF of Y given $\boldsymbol{X} = \boldsymbol{x}$. The survival function is the complement of the CDF: $S_Y(y) \equiv \mathrm{P}(Y > y)$, or equivalently $S_Y(\cdot) = 1 - F_Y(\cdot)$. The two groups (populations) of interest are labeled A and B, generally used as superscripts. Thus, for group A: Y^A is the ordinal outcome with CDF $F_Y^A(\cdot)$ and survival function $S_Y^A(\cdot)$, $\boldsymbol{X}^A$ is the covariate vector with CDF $F_{\boldsymbol{X}}^A(\cdot)$ and support $\mathcal{X}^A$, and $F_{Y|\boldsymbol{X}}^A(\cdot \mid \cdot)$ is the conditional CDF. For group B, the A superscripts are all replaced with B superscripts. Similarly, a C superscript indicates the counterfactual distribution, introduced below.

Following Chernozhukov et al. (2013, §2.1), the population-level counterfactual distribution is defined as follows. The thought experiment is: starting from group B, what if we keep fixed the conditional distribution but change the covariate distribution to that of group A? Thus, we can see how much of a change in the outcome distribution is statistically explained

purely from the difference in covariate distributions. Because Y is ordinal with J categories, its distribution is fully characterized by the $J-1$ values of $F_Y(y)$ for $y \in \{1, \ldots, J-1\}$. Mathematically, as in (2.1) of Chernozhukov et al. (2013) or (27) of Fortin et al. (2011), the counterfactual CDF is

$$F_Y^C(y) \equiv \int_{\mathcal{X}^A} F_{Y|\boldsymbol{X}}^B(y \mid \boldsymbol{x}) \, \mathrm{d}F_{\boldsymbol{X}}^A(\boldsymbol{x}), \quad y \in \{1, \ldots, J-1\}. \tag{4}$$

A density version of this equation appears in (3) of DiNardo et al. (1996). As in (2.3) of Chernozhukov et al. (2013), the expression in (4) requires $\mathcal{X}^A \subseteq \mathcal{X}^B$; if instead $\mathcal{X}^B \subseteq \mathcal{X}^A$, then the A and B labels can be switched. For intuition about (4), consider the extreme cases: if $F_{\boldsymbol{X}}^A = F_{\boldsymbol{X}}^B$, then (4) yields $F_Y^C(y) = F_Y^B(y)$, and if $F_{Y|\boldsymbol{X}}^B = F_{Y|\boldsymbol{X}}^A$, then (4) yields $F_Y^C(y) = F_Y^A(y)$.

4.2 Summary statistic interpretations and equivalences

The full distributions F_Y^A, F_Y^B, and F_Y^C can and should be reported, but this requires reporting $3(J-1)$ values, so a summary can facilitate communication and understanding of results. We show an equivalence between a naive "mean" decomposition and a robust survival function decomposition. Everything in this section is still at the population level, to describe and understand the interpretation of different possible population objects of interest. Estimation and inference follow in Section 5.

Notationally, denote differences as Δ, with the total (subscript T), explained (E), and unexplained (U) differences respectively

$$\Delta_T, \ \Delta_E, \ \Delta_U. \tag{5}$$

For an ordinal outcome, a natural decomposition compares survival function differences summed (or averaged) across categories. This does not depend on the cardinal values of the categories. Although in a different context, this shares the spirit of Theorem 1 of Kobus and Miłoś (2012), who find that any health inequality index satisfying certain axioms can be

written as transformations of the category frequencies; our expression in (6) similarly depends only on category frequencies. Given survival functions $S_Y^A(\cdot)$ and $S_Y^B(\cdot)$, we summarize their difference as

$$\sum_{j=1}^{J}[S_Y^A(j) - S_Y^B(j)], \tag{6}$$

and similarly for other pairs of survival functions. Summing from $j = 1$ to $J-1$ is equivalent because $S_Y^A(J) = S_Y^B(J) = 0$. Taking the average (instead of sum) would multiply (6) by $1/J$, but ultimately the explained proportion would remain identical because the $1/J$ would cancel out in (8) below. Given (6), using the notation of (5) and adding superscript S for "survival," the corresponding differences are

$$\Delta_T^S = \sum_{j=1}^{J}[S_Y^A(j) - S_Y^B(j)], \quad \Delta_E^S = \sum_{j=1}^{J}[S_Y^C(j) - S_Y^B(j)], \quad \Delta_U^S = \sum_{j=1}^{J}[S_Y^A(j) - S_Y^C(j)], \tag{7}$$

and the explained proportion is

$$\frac{\Delta_E^S}{\Delta_T^S} = \frac{\sum_{j=1}^{J}[S_Y^C(j) - S_Y^B(j)]}{\sum_{j=1}^{J}[S_Y^A(j) - S_Y^B(j)]}. \tag{8}$$

This is equivalent to a CDF-based decomposition. The components in (7) equal the negative of their CDF-based analogs. For example,

$$\Delta_T^S = \sum_{j=1}^{J}[S_Y^A(j) - S_Y^B(j)] = \sum_{j=1}^{J}\{[1 - F_Y^A(j)] - [1 - F_Y^B(j)]\} = -\sum_{j=1}^{J}[F_Y^A(j) - F_Y^B(j)],$$

and similarly for the other differences in (7). Thus, the explained proportion remains the same because $(-\Delta_E^S)/(-\Delta_T^S) = \Delta_E^S/\Delta_T^S$.

Extending Lemma 1, the survival function decomposition is also equivalent to a naive "mean" decomposition after coding the Y categories as $1, \ldots, J$. We state this as a corollary to a more general result.

Theorem 4. *Let W and Z be discrete random variables with possible values $\{1, 2, \ldots, J\}$. Then, $\mathrm{E}(W) - \mathrm{E}(Z) = \sum_{j=1}^{J}[S_W(j) - S_Z(j)]$, where $S_W(j) \equiv \mathrm{P}(W > j)$ and $S_Z(j) \equiv \mathrm{P}(Z > j)$ are the survival functions.*

Proof. Plugging in for $\mathrm{E}(W)$ and $\mathrm{E}(Z)$ from Lemma 1,

$$\mathrm{E}(W) - \mathrm{E}(Z) = 1 + \sum_{j=1}^{J-1} S_W(j) - \left[1 + \sum_{j=1}^{J-1} S_Z(j)\right] = \sum_{j=1}^{J-1}[S_W(j) - S_Z(j)]$$
$$= \sum_{j=1}^{J}[S_W(j) - S_Z(j)]. \qquad \square$$

Corollary 5. *Given distributions of ordinal random variables Y^A, Y^B, and counterfactual Y^C, the survival function decomposition is equivalent to the "mean" decomposition after coding Y category values as $1, \ldots, J$, in the sense that the explained proportion in* (8) *equals the "mean"-based explained proportion.*

Proof. Theorem 4 implies the following for the "mean"-based decomposition components. The total difference is

$$\Delta_T^{\mu} \equiv \mathrm{E}(Y^A) - \mathrm{E}(Y^B) = \sum_{j=1}^{J}[S_Y^A(j) - S_Y^B(j)] = \Delta_T^S.$$

Similarly, for the explained difference,

$$\Delta_E^{\mu} \equiv \mathrm{E}(Y^C) - \mathrm{E}(Y^B) = \sum_{j=1}^{J}[S_Y^C(j) - S_Y^B(j)] = \Delta_T^S.$$

Thus, the explained proportions are also equal: $\Delta_E^{\mu}/\Delta_T^{\mu} = \Delta_E^S/\Delta_T^S$. $\square$

Corollary 5 serendipitously implies that we can interpret a "mean" decomposition as a survival function decomposition. Thus, if a paper reports results for an ordinal "mean" decomposition, then even if we disagree with a literal "mean" interpretation, we can still agree about the relative magnitude of explained and unexplained components.

4.3 Implications for regression-based decomposition

Corollary 5 applies to regression-based decomposition. Coding Y as $1, \ldots, J$, let $m^B(\boldsymbol{x}) \equiv \mathrm{E}(Y^B \mid \boldsymbol{X}^B = \boldsymbol{x})$, which by (2) can be interpreted more robustly as $1 + \sum_{j=1}^{J-1} S_Y^B(j \mid \boldsymbol{X}^B =$

$\boldsymbol{x}$). The counterfactual mean is $\mathrm{E}(Y^C) = \mathrm{E}[m^B(\boldsymbol{X}^A)]$. The decomposition is

$$\mathrm{E}(Y^A) - \mathrm{E}(Y^B) = \overbrace{\mathrm{E}(Y^A) - \underbrace{\mathrm{E}[m^B(\boldsymbol{X}^A)]}_{\mathrm{E}(Y^C)}}^{\text{unexplained}} + \overbrace{\underbrace{\mathrm{E}[m^B(\boldsymbol{X}^A)]}_{\mathrm{E}(Y^C)} - \mathrm{E}(Y^B)}^{\text{explained}}. \tag{9}$$

By Corollary 5, the decomposition in (9) can be interpreted in terms of survival functions. Thus, a nonparametric regression-based "mean" decomposition always has a survival function interpretation, without any assumption about cardinalization.

In Section 5.2, we provide an even more precise equivalence result for when OLS is used to estimate the decomposition.

5 Ordinal decomposition: computation and equivalence

Sections 5.1 and 5.3 closely follow the estimation and inference of Chernozhukov et al. (2013). We continue the notation introduced in Section 4.1. Theoretically, ordinal Y is simpler than continuous Y (as in Chernozhukov et al., 2013) because there are only $J-1$ values at which we need to estimate the counterfactual CDF, rather than a continuum of an infinite number of points. Thus, their asymptotic results all hold. Our first contribution in this section is to gather practical guidance, which we follow in our provided code.

Our second contribution is the new equivalence result in Section 5.2. This shows that the naive Blinder–Oaxaca "mean" decomposition estimator that seems to assume cardinal values $1, \ldots, J$ can be interpreted as a survival function decomposition estimator that is cardinalization-robust.

5.1 Estimation

The distribution regression model as in (3.1) of Chernozhukov et al. (2013) separately models the conditional CDF evaluated at each $y \in \{1, \ldots, J-1\}$ in turn. This is also the approach of Foresi and Peracchi (1995, §2). Generally, let $\Lambda(\cdot)$ be the link function, such as the

standard normal or logistic CDF, and let $\boldsymbol{P}(\boldsymbol{x})$ be a column vector of transformations of the original covariate vector $\boldsymbol{x}$. For example, $\boldsymbol{P}(\boldsymbol{x})$ can include squares, interactions, higher-degree polynomial terms, or other basis functions like B-splines. Let $\boldsymbol{\gamma}_y$ be the coefficient vector corresponding to category $y \in \{1, \ldots, J-1\}$. Then, the model is

$$F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \Lambda\big(\boldsymbol{P}(\boldsymbol{x})'\boldsymbol{\gamma}_y\big), \quad y \in \{1, \ldots, J-1\}. \tag{10}$$

Chernozhukov et al. (2013, p. 2217) note the link function $\Lambda(\cdot)$ is not as important as having a sufficiently flexible $\boldsymbol{P}(\boldsymbol{x})$. A popular choice of estimator is the series logit from Hirano et al. (2003, p. 1170), where $\Lambda(\cdot)$ is the logistic CDF and $\boldsymbol{P}(\boldsymbol{x})$ contains polynomials or other basis function transformations. Model selection techniques such as cross-validation can be used to select an appropriately flexible (but not too flexible) model in practice. More on basis expansions and model selection can be found in textbooks like that of Hastie et al. (2009, Chs. 5 and 7).

The model in (10) is estimated using (only) data from group B, yielding $\hat{\boldsymbol{\gamma}}_y^B$ for $y \in \{1, \ldots, J-1\}$. Weights can be used as appropriate. For given y and $\boldsymbol{x}$ values, similar to (10), the estimated conditional CDF is $\hat{F}^B_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \Lambda\big(\boldsymbol{P}(\boldsymbol{x})'\hat{\boldsymbol{\gamma}}_y^B\big)$.

The estimated conditional CDF for group B is then plugged into the counterfactual distribution formula from (4) along with the estimated marginal distribution of $\boldsymbol{X}^A$. Without sampling weights, integrating against $\hat{F}^A_{\boldsymbol{X}}$ is equivalent to averaging over the sample values of $\boldsymbol{X}^A$, so the estimated counterfactual CDF is as given at the end of Remark 3.1 of Chernozhukov et al. (2013):

$$\hat{F}^C_Y(y) = \int_{\mathcal{X}^A} \hat{F}^B_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) \,\mathrm{d}\hat{F}^A_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{n_A}\sum_{i=1}^{n_A} \Lambda\big(\boldsymbol{P}(\boldsymbol{X}_i^A)'\hat{\boldsymbol{\gamma}}_y^B\big), \quad y \in \{1, \ldots, J-1\}, \tag{11}$$

where $\boldsymbol{X}_i^A$ are the observations in the group A sample for $i = 1, \ldots, n_A$; see also page 71 of Fortin et al. (2011). If there are weights, then a weighted average can be taken:

$$\sum_{i=1}^{n_A} \tilde{w}_i^A \Lambda\big(\boldsymbol{P}(\boldsymbol{X}_i^A)'\hat{\boldsymbol{\gamma}}_y^B\big),$$

where $\tilde{w}_i^A \equiv w_i^A / \sum_{i=1}^{n_A} w_i^A$ normalizes the original weights w_i^A to sum to 1; the unweighted formula above is the special case with $\tilde{w}_i^A = 1/n_A$ for all i.

The actual group A and B outcome distributions can be estimated with the usual estimators. Without weights, for each $y \in \{1, \ldots, J-1\}$,

$$\hat{F}_Y^A(y) = \frac{1}{n_A} \sum_{i=1}^{n_A} \mathbb{1}\{Y_i^A \leq y\}, \quad \hat{F}_Y^B(y) = \frac{1}{n_B} \sum_{i=1}^{n_B} \mathbb{1}\{Y_i^B \leq y\},$$

where the Y_i^A are observations from the group A sample for $i = 1, \ldots, n_A$, and the Y_i^B are observations from the group B sample for $i = 1, \ldots, n_B$. With weights, similar to above,

$$\hat{F}_Y^A(y) = \sum_{i=1}^{n_A} \tilde{w}_i^A \, \mathbb{1}\{Y_i^A \leq y\}, \quad \hat{F}_Y^B(y) = \sum_{i=1}^{n_B} \tilde{w}_i^B \, \mathbb{1}\{Y_i^B \leq y\},$$

where again $\tilde{w}_i^A \equiv w_i^A / \sum_{i=1}^{n_A} w_i^A$ and similarly $\tilde{w}_i^B \equiv w_i^B / \sum_{i=1}^{n_B} w_i^B$ normalize the raw weights to sum to one in each sample.

Given the three estimated CDFs $\hat{F}_Y^A$, $\hat{F}_Y^B$, and $\hat{F}_Y^C$, the survival function decomposition and its explained proportion can be computed using (7) and (8), noting that estimated CDF $\hat{F}(\cdot)$ implies the corresponding estimated survival function $\hat{S}(y) = 1 - \hat{F}(y)$.

5.2 Equivalence with OLS-based Blinder–Oaxaca

Here, we establish a numerical equivalence between two seemingly very different estimators of the explained proportion of a decomposition. The first estimator uses the survival function decomposition in (8), where the counterfactual distribution is estimated as in Section 5.1 using the identity link function $\Lambda(a) = a$, i.e., by OLS with a linear probability model. The second estimator naively applies the conventional OLS-based Blinder–Oaxaca decomposition of the "mean," interpreting the coding $Y \in \{1, \ldots, J\}$ as cardinal values. For details about the conventional Blinder–Oaxaca decomposition that traces back to Kitagawa (1955), Blinder (1973), and Oaxaca (1973), see for example (15) and more generally Section 3.1 of Fortin et al. (2011).

Theorem 6. *Assuming both are well-defined given the data, the following two estimates of the explained proportion are numerically identical. First estimate: after coding Y with cardinal values $Y \in \{1, 2, \ldots, J\}$, estimate the conventional Blinder–Oaxaca mean decomposition, specifically the explained proportion*

$$\frac{(\bar{\boldsymbol{X}}^A - \bar{\boldsymbol{X}}^B)'\hat{\boldsymbol{\beta}}^B}{\bar{Y}^A - \bar{Y}^B},$$

where as usual $\hat{\boldsymbol{\beta}}^B$ is the OLS-estimated coefficient vector from regressing Y on $\boldsymbol{X}$ in sample B, and $\bar{\boldsymbol{X}}^A$ is the average of observed $\boldsymbol{X}$ values in the group A sample, with $\bar{Y}^A$ similarly the average of observed Y values in the group A sample, and with $\bar{\boldsymbol{X}}^B$ and $\bar{Y}^B$ defined similarly for group B. Second estimate: take the survival function decomposition's estimated explained proportion

$$\frac{\sum_{y=1}^{J}[\hat{S}^C(y) - \hat{S}^B(y)]}{\sum_{y=1}^{J}[\hat{S}^A(y) - \hat{S}^B(y)]}$$

as in (8)*, and compute $\hat{S}^C(\cdot)$ with the counterfactual distribution estimator in* (11) *with the special case $\Lambda(x) = x$ and $\boldsymbol{P}(\boldsymbol{x}) = \boldsymbol{x}$, with $\hat{\boldsymbol{\gamma}}_y^B$ estimated by OLS regression of $Z_y \equiv \mathbb{1}\{Y \leq y\}$ on $\boldsymbol{X}$ using data sample B.*

Proof. See Appendix A. □

Theorem 6 says that we can now more robustly and meaningfully reinterpret published results based on seemingly inappropriate application of Blinder–Oaxaca decomposition to ordinal outcomes. Specifically, even if a paper dubiously claims to decompose the "mean" of an ordinal outcome, we can interpret the estimated explained proportion in terms of survival functions and a counterfactual distribution that does not depend on any particular cardinalization. Although other estimators may help reduce functional form misspecification when estimating the counterfactual distribution, using the Blinder–Oaxaca estimate may still be useful for exploratory analysis. Additionally, if the functional form misspecification does not seem too large, Blinder–Oaxaca readily provides a "detailed decomposition" showing the separate contributions of each covariate; for example, see (17)–(18) and the surrounding

text of Fortin et al. (2011), which explains in more detail the idea first seen in Tables 3 and 4 of Oaxaca (1973) and Tables 1–4 and Section I of Blinder (1973).

5.3 Inference

Inference for the Δ components can use the bootstrap in Algorithm 2 of Chernozhukov et al. (2013). Their bootstrap (and the corresponding theory) is for $s(\hat{F}_Y^C)$, where $s(\cdot)$ summarizes a distribution like $\hat{F}_Y^C$. For $s(\cdot)$ like the "mean," analytic confidence intervals may be readily available, or as long as the bootstrap is being run anyway, they can be bootstrapped, too. Their Algorithm 2 bootstrap is a very general exchangeable weight bootstrap that includes the usual bootstrap as a special case. Per their Remark 5.1, the bootstrap weights (or resamples) should be done separately and independently for groups A and B. Given each bootstrap weight vector or sample, the full estimation procedure from Section 5.1 is run, and this is repeated many times. The many bootstrap-world estimates of the Δ components can then be used in any standard bootstrap confidence interval formula as desired.

6 Empirical results: mental health disparities

We apply our methodological results to an empirical analysis of rural–urban mental health disparity in the U.S.[2] Specifically, we decompose the overall rural–urban difference in depression using age, sex, education, income, and region.

6.1 Motivation and context

Depression is widely studied because of its prevalence and importance. For this reason, and inspired by the inclusion of mental health in the United Nation's Sustainable Development Goals, *The Lancet* formed a Commission on Global Mental Health, with Patel et al. (2018)

[2]Replication code is available online at `https://qianjoewu.github.io/`. Our analysis was performed in R (R Core Team, 2022) version 4.4.3, with help from packages `ggplot2` (Wickham, 2016), `ggmosaic` (Jeppson and Hofmann, 2023; Jeppson et al., 2023), and `fastglm` (Huling, 2022).

providing a detailed global overview. Economically, Greenberg et al. (2021) estimate that in the U.S. in 2018, the aggregate economic burden of adults with major depressive disorder exceeded $300 billion. This includes the costs of medical care, suicide, and decreased work hours as well as productivity. In the context of low-income and middle-income countries, Lund et al. (2011) find systematic evidence that mental health interventions are associated with improved economic outcomes, though weaker evidence that economic or financial interventions are associated with improved mental health.

Rural–urban differences in mental health are important, in terms of both causes and treatment. Because we analyze U.S. data, we focus here on other U.S. studies, while noting that studies from other countries have more often found better mental health outcomes in rural than urban residents (or no difference); for example, see the discussion and references in Nicholson (2008) and Breslau et al. (2014). In the U.S., the policy importance of considering rural–urban differences is reflected by the existence of the Federal Office of Rural Health Policy within the Department of Health and Human Services; with respect to mental health, the Office's role is discussed by Human and Wasem (1991), for example. Some challenges to delivering appropriate treatment are more common in rural areas. For example, compared to urban populations, rural populations in the U.S. have greater physical (distance) and cultural (stigma) barriers to mental health services (Gamm et al., 2010; Human and Wasem, 1991; Palomin et al., 2023), although Wang et al. (2004) find "no significant association" (p. 401) between rural/urban location and delay in seeking treatment for a mental disorder. Hastings and Cohn (2013) discuss multiple provider-side challenges in the context of rural central Appalachia.

Regarding rural–urban differences in mental health outcomes in the U.S., empirical evidence remains mixed and sparse. Breslau et al. (2014, p. 50) note that despite the received wisdom that rates of depression and other conditions are higher in urban areas, there is almost no high-quality empirical evidence of this, and some evidence of the opposite. Using data from the National Survey of Drug Use and Health, Breslau et al. (2014) find that

compared to large metropolitan populations, rural populations in the U.S. have higher rates of serious mental illness, but not statistically significantly so, and negligibly higher rates of major depression (odds ratio 1.01), although they also find statistically significantly higher rates of major depression in semi-rural areas; see Table 3 and related discussion. Summarizing some other papers in the literature, Gamm et al. (2010) note that rural men in the U.S. have significantly higher suicide rates, but rural–urban comparisons of other outcomes have shown less difference, albeit possibly due to underreporting in rural self-assessments. Using NHIS data from 1999, Probst et al. (2006) find higher overall rates of depression among rural residents, defining "rural" as living outside a metropolitan statistical area, and "urban" as living inside (p. 654). They find (Table 3) that this gap persists when controlling for demographic variables, but that it disappears when additionally controlling for other health measures and socioeconomic variables, although this includes a general self-reported health measure that may directly depend on depression. They use a binary measure of depression based on the now-obsolete Composite International Diagnostic Interview Short Form. We contribute empirical results to help fill this gap, using different data than Breslau et al. (2014) and much more recent NHIS data than Probst et al. (2006), as well as decomposing the overall rural–urban difference in the distribution of depression in terms of important demographic and economic factors.

Our choice of explanatory variables aligns with those studied in the literature. Income has been strongly linked to mental health outcomes like depression (e.g., Ettner, 1996; Lorant et al., 2003; Lund et al., 2011; Ridley et al., 2020). Education has also been found to have a strong link with mental health, specifically depression, by Cutler and Lleras-Muney (2006), for example. Besides income and education, we also use age, sex, and geographic region. Similar combinations are used in the literature. For example, Breslau et al. (2014) use age, sex, education, and location (rural/urban), plus race and marital status but without geographic region or income, which we find to be the most important explanatory variable. Probst et al. (2006) use the same variables as Breslau et al. (2014) plus additional measures

of health, English fluency, and (un)employment status, as well as a binary measure of income. As noted above, their inclusion of a general self-reported health measure seems potentially problematic if that measure depends directly on depression, making it unsurprising that it "explains" differences in depression. However, other variables above would be interesting to study in future work.

6.2 Data

We use the publicly available National Health Interview Survey (NHIS) 2022 data (National Center for Health Statistics, 2022), chosen for its inclusion of mental health assessment and recent availability. Our analysis targets individuals aged 24–64, focusing on those of working age and surpassing the average age at college graduation in the U.S.

The following variables are used. The outcome variable Y (`PHQCAT_A`) measures the severity of depressive symptoms, summarizing the eight-item Patient Health Questionnaire into four categories from low to high: "none/minimal," "mild," "moderate," and "severe." The rural and urban groups are defined using variable `URBRRL`: group A contains individuals who live in counties categorized as nonmetropolitan, while group B is large central metro counties. We use the provided variables for education (`EDUCP_A`), sex (`SEX_A`), age (`AGEP_A`), family income (`POVRATTC_A`), and geographic region (`REGION`) to construct our explanatory vector $\boldsymbol{X}$, as described in Section 6.3. The estimation uses the sampling weight variable (`WTFA_A`). We drop 282 observations (3.4%): those for which either age or urban group is missing, and those that fit our age and urban group restrictions but have another variable value missing. This leaves 7902 observations for our analysis.

Figure 1 shows a mosaic plot to visualize the distribution of depression levels for rural and urban groups. As a fraction of the total plot area, each cell's area represents the proportion of sampled individuals in that cell. For example, urban individuals with mild depression are 9.46% of the sample, so that rectangle's area is 9.46% of the total mosaic plot area, and labeled. Using the identity $\mathrm{P}(Y = y, X = x) = \mathrm{P}(Y = y \mid X = x)\,\mathrm{P}(X = x)$, the

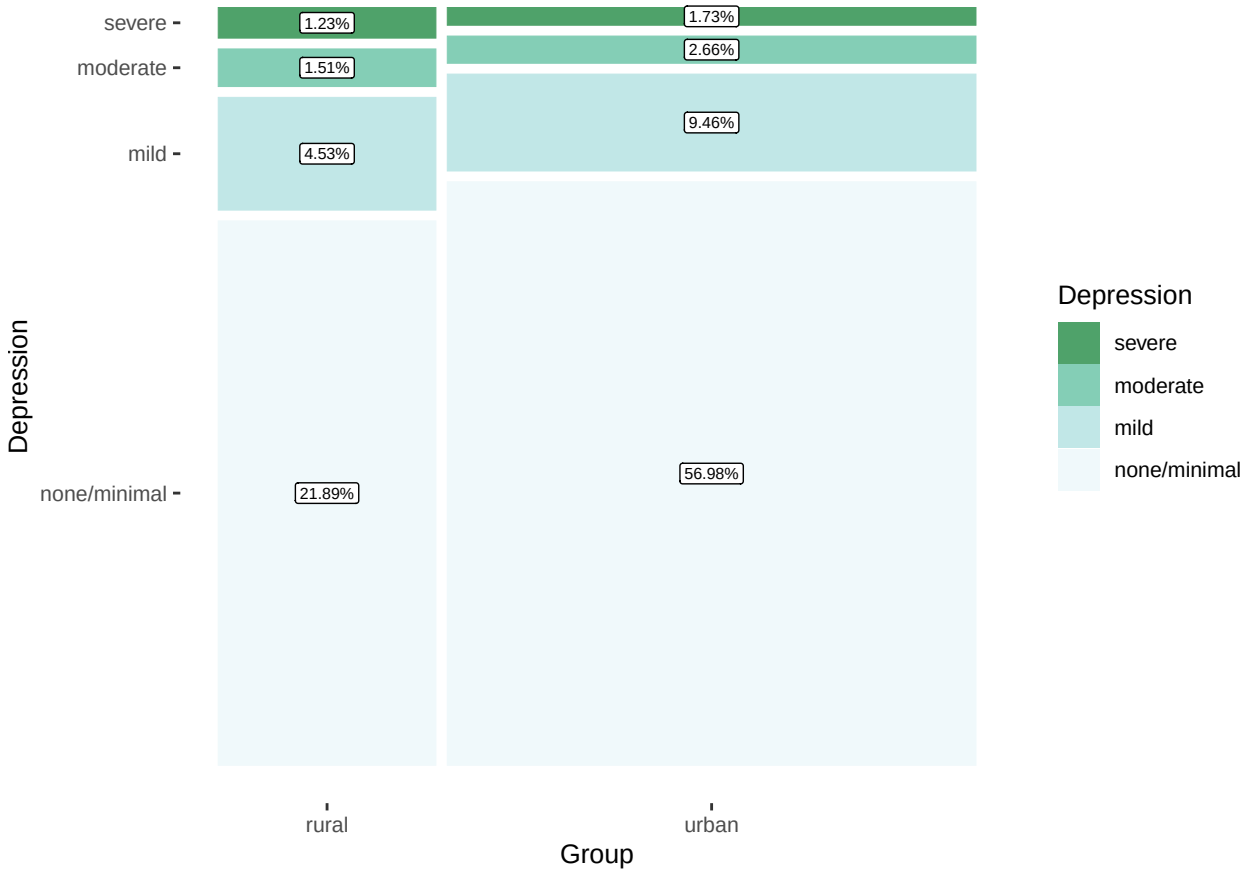

Figure 1: Urban and rural depression level distributions.

mosaic plot makes the width of the rectangle for cell (y, x) proportional to $\mathrm{P}(X = x)$, and its height proportional to $\mathrm{P}(Y = y \mid X = x)$. For the rectangle widths, there are $n_A = 2448$ individuals in the rural group A and $n_B = 5454$ in the urban group B. Because 5454 is a little over twice 2448, the urban column is a little over twice the width of the rural column. For the rectangle heights, each is proportional to the conditional (on group) probability of that depression category, which also helps us see the empirical conditional (on group) CDF values. For example, within the rural column, if we normalize the height of the column to 1, then the height of the bottom rectangle is $\hat{F}_Y^A(1)$, the empirical CDF for the rural group evaluated at category 1. Similarly, the height of the bottom two rectangles combined is $\hat{F}_Y^A(2)$, and the height of the bottom three rectangles combined is $\hat{F}_Y^A(3)$.

Figure 2 shows that the rural group first-order stochastically dominates the urban group

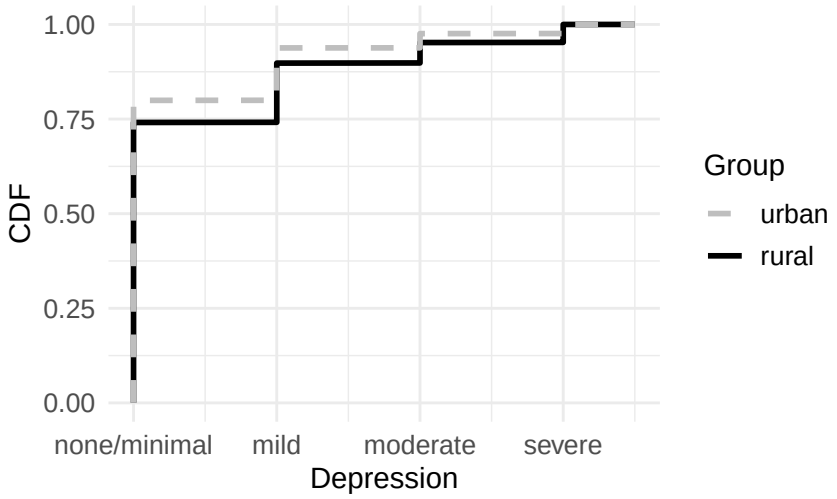

Figure 2: CDFs of Urban and rural depression level.

in depression level in the sample because the rural CDF is below the urban CDF at each category: $\hat{F}_Y^A(j) \leq \hat{F}_Y^B(j)$ at each $j \in \{1, 2, 3, 4\}$. This says that overall the rural group has higher levels of depression (worse mental health) than the urban group.

6.3 Estimation

To construct the counterfactual distribution for the rural group, we use distribution regression with three estimation methods: OLS with a linear probability model (LPM), logit, and nonparametric series logit.

For LPM/OLS and logit estimation, we include the following explanatory variables: a dummy variable for high education (equal to 1 if the individual has at least some college education), sex, age, squared age, income (family poverty ratio: family income divided by poverty threshold), a dummy variable for "high income" that equals 1 if the family poverty ratio has been top-coded (value 11), and region dummies for the Northeast, Midwest, and West (with South the base group).

For nonparametric estimation, we use the following model selection procedure. We run the procedure separately for each dependent variable $\mathbb{1}\{Y \leq y\}$, $y \in \{1, 2, 3\}$. The higher-order terms lack a natural ordering, so technically we do not use "series" logit because

we consider many non-nested subsets of higher-order terms as candidate models. That is, unlike with a scalar X for which a series estimator includes X^k for $k = 0, 1, \ldots, K$, here if we include two quadratic terms (for example), we try (X_1^2, X_2^2), (X_1^2, X_3^2), (X_2^2, X_3^2), (X_1^2, X_1X_2), (X_1^2, X_1X_3), etc. First, in every candidate model, we include a linear term for each explanatory variable described above. Second, we construct a candidate model for every possible combination of the quadratic terms, which include squared age, squared income, and interaction terms like high education*age. Third, we conduct logit estimation for each candidate model and compute the Akaike information criterion (AIC) of Akaike (1974). Fourth, we select the model with the lowest AIC as the optimal model. Fifth, we consider adding certain cubic terms if the corresponding quadratic terms are included in the selected model, and consider adding quartic terms if the cubic terms are selected, etc.

Instead of AIC, cross-validation could be used for model selection, but there are some disadvantages in our setting. First, if applied to the very large number of candidate models described above, the computation time for even 5-fold cross-validation would be prohibitive; it should take even longer than with AIC, which took nearly 20 hours to compute. Second, although computationally faster, there are drawbacks to applying 5-fold cross-validation to select the penalization hyperparameter for lasso (Tibshirani, 1996). One drawback is that the result is sensitive to the random number generator seed used (because the observations in each fold are chosen randomly). Additionally, the conventional error measures like those available through the `glmnet` package (Friedman et al., 2010; Tay et al., 2023) perform poorly for the more severe depression categories that comprise a relatively small fraction of the population. For example, with outcome $\mathbb{1}\{Y_i \leq 3\}$, the intercept-only model is selected as "best." In principle, this could be addressed by coding an alternative error measure based on weighted 0–1 loss (e.g., Kaplan, 2023, §14.3.1) that makes it relatively more important to correctly predict individuals who actually have severe depression. However, because the AIC model selection works well and such details are far from our main contributions, we do not pursue this alternative further.

In practice, due to computational constraints, in addition to linear terms we always include age*income, age*high income, squared age, and squared income in every candidate model. Each of the 22 additional quadratic terms may be included or excluded, yielding $2^{22} = 4{,}194{,}304$ candidate models. The selected model is different for each dependent variable $\mathbb{1}\{Y \leq y\}$.[3] For each $\mathbb{1}\{Y \leq y\}$, adding cubed income and/or cubed age to the selected quadratic model resulted in worse (higher) AIC, so we use the selected quadratic models for estimation.

In our decomposition result, we include bootstrapped standard errors, which were computed using the procedure outlined by Hlavac (2022, §2.4) and described here in Method 1.

Method 1. *[bootstrapped standard errors]*

1. *Take R random samples with replacement from the relevant set of observations, separately and independently for groups A and B (per Section 5.3).*
2. *In each approach, estimate and perform the decomposition for the sample from Step 1.*
3. *Calculate the bootstrapped standard error as the standard deviation of the R decomposition estimates from Step 2.*

We use $R = 1000$.

6.4 Results

We decompose the "mean" difference between rural group A and urban group B, equivalent to decomposing the survival function per Corollary 5. We use the three estimators in Section 6.3 as well as the naive conventional Blinder–Oaxaca decomposition estimator, to verify our Theorem 6.

[3]Beyond the baseline terms (high education, sex, age, income, high income, Northeast, Midwest, West, age*income, age*high income, squared age, and squared income), the selected model with dependent variable $\mathbb{1}\{Y \leq 1\}$ includes high education*age, high education*income, high education*high income, sex*income, sex*Northeast, age*West, income*Midwest; the selected model with $\mathbb{1}\{Y \leq 2\}$ includes high education*sex, high education*Northeast, sex*age, age*Midwest, and income*Northeast; and the selected model with $\mathbb{1}\{Y \leq 3\}$ includes sex*income, income*Midwest, high income*Northeast, and high income*West.

Table 1 displays the estimated rural, urban, and counterfactual CDFs. As before, the counterfactual starts from the group B urban distribution and substitutes in the group A rural distribution of $\boldsymbol{X}$, while keeping the group B urban conditional distribution of Y given $\boldsymbol{X}$. For the counterfactual, the estimated CDF values are similar across estimators, particularly OLS and logit. This suggests that OLS provides a reasonable approximation here, and with computation time in seconds instead of the many hours to compute our nonparametric estimator (almost 20 hours on a personal computer). At minimum, OLS seems very practical for exploratory analysis, although for the final analysis a nonparametric estimator may be preferred.

Table 1: Estimated actual and counterfactual CDFs.

Group	$\hat{F}(1)$	$\hat{F}(2)$	$\hat{F}(3)$
Rural	0.751	0.906	0.958
Urban	0.804	0.938	0.976
Counterfactual (OLS/LPM)	0.789	0.925	0.968
Counterfactual (logit)	0.790	0.926	0.968
Counterfactual (series logit)	0.786	0.923	0.968

Table 2: Decomposition results.

Model	Explained (%)	Unexplained (%)
Naive Blinder–Oaxaca	33.9	66.1
	(12.5)	(12.5)
OLS/LPM	33.9	66.1
	(12.5)	(12.5)
Logit	33.0	67.0
	(12.6)	(12.6)
Series logit	38.9	61.1
	(13.8)	(13.8)

Bootstrapped standard errors are in parentheses.

Table 2 displays the decomposition results. Given the similar counterfactual CDF estimates in Table 1, naturally the explained proportion estimates are also similar, all in the range of 33–39 percent. This suggests that education, sex, age, income, and region collec-

tively account for approximately 33–39 percent of the difference in the depression distribution between the rural and urban groups. This is a substantial amount, but still leaves over half unexplained.

To verify Theorem 6, we also compute the naive Blinder–Oaxaca decomposition using the $1, \ldots, J$ cardinalization. As expected, compared to the OLS counterfactual approach, the decomposition results are identical. Thus, even if researchers reported only the conventional Blinder–Oaxaca decomposition with this data, we could still interpret the results in terms of a survival function decomposition with a counterfactual CDF estimated by distribution regression, robust to any alternative cardinalization.

Table 3: Blinder–Oaxaca detailed decomposition results.

Variable	Rural mean	Urban mean	Explained (%)
Income (ratio)	3.274	4.715	41.6
	(0.055)	(0.050)	(12.1)
High income	0.020	0.102	-7.8
	(0.003)	(0.005)	(3.5)
Midwest	0.316	0.155	11.3
	(0.011)	(0.006)	(6.2)
West	0.152	0.342	-3.0
	(0.009)	(0.007)	(4.8)
Northeast	0.106	0.165	-2.1
	(0.009)	(0.006)	(2.1)
High education	0.528	0.685	-3.9
	(0.012)	(0.008)	(4.6)
Age	45.635	42.637	-58.0
	(0.283)	(0.184)	(28.0)
Age^2			56.3
			(28.1)
Female	0.493	0.499	-0.6
	(0.012)	(0.008)	(1.5)
Intercept	1.000	1.000	0.0
			(0.0)
Aggregate			33.9
			(12.5)

Bootstrapped standard errors are in parentheses.

Table 3 shows results of a "detailed decomposition" using the Blinder–Oaxaca estimates, which is another advantage of being able to use Blinder–Oaxaca, as our results justify. The detailed decomoposition shows how individuals variables contribute to the overall explained proportion. For each explanatory variable X_j in the vector $\boldsymbol{X}$, we show estimates of the rural mean $\mathrm{E}(X_j^A)$, urban mean $\mathrm{E}(X_j^B)$, and contribution

$$\frac{[\mathrm{E}(X_j^A) - \mathrm{E}(X_j^B)]\hat{\beta}_j^B}{\mathrm{E}(Y^A) - \mathrm{E}(Y^B)} \times 100\%$$

to the overall explained proportion. Note the sum of the estimated contributions equals the estimated explained proportion shown in Table 2. The rows are in decreasing order of absolute contributions, considering the combined contribution of the two income variables and the two age variables.

The main message is that income explains far more than any other variable. The combined contribution of income and the high-income dummy is $41.6 + (-7.8) = 33.8$, almost exactly the overall 33.9% explained. The other contributions are a mix of positive and negative values that nearly fully cancel out. As seen from the rural and urban means, urban incomes are higher, and higher incomes are associated with lower depression (negative $\hat{\beta}_j^B$ coefficient), so this partly explains the higher depression levels in rural areas. Although the contributions of Age and Age^2 initially seem even larger, their combined contribution is only -1.7. The midwest dummy has a contribution of 11.3 to the explained proportion. The means show that a much higher proportion of the rural individuals live in the midwest, and being in the midwest is associated with higher depression (positive $\hat{\beta}_j^B$ coefficient), so this also partly explains the higher depression levels in rural areas. The other contributions are all relatively small in magnitude, as well as all negative, so altogether they offset the midwest contribution. The small magnitudes are a combination of small regression coefficients and/or small differences in the rural and urban means of that X_j variable. Overall, income plays an important role in explaining the rural–urban difference in depression levels, but it still only accounts for about a third of the overall difference.

7 Conclusion

We have provided theoretical results about interpreting OLS-based analysis when an ordinal outcome Y is coded with numeric values $1, 2, 3, \ldots$. Although the "best linear predictor" interpretation of the OLS estimand requires such values to be the true cardinal values of the categories, a "best linear approximation" interpretation remains valid even when those are not cardinal values, where the approximation is of the sum of conditional survival function values. Further, the OLS-based Blinder–Oaxaca decomposition can be interpreted as a survival function decomposition the remains valid even if the $1, 2, 3, \ldots$ are not cardinal values. This suggests such "naive" OLS-based results can be interpreted robustly and can be practically useful when dealing with the commonly used ordinal variables in health economics, epidemiology, sociology, and related areas in health, medicine, and social science.

CRediT author contribution statement

Qian Wu: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing – original draft, Writing – review & editing, Visualization

David M. Kaplan: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing

Declaration of competing interest

We (the authors) declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

For helpful feedback, we thank Xing Ling, Zack Miller, Shawn Ni, Mike Pesko, Yuhao Yang, and especially Alyssa Carlson, as well as seminar participants from the Southwestern University of Finance and Economics and conference participants from the 2024 Midwest Econometrics Group (hosted by the University of Kentucky). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

A Additional proof

Proof of Theorem 6. Recall from Corollary 5 that decomposing the "mean" is equivalent to decomposing the average survival function difference; below, we use the "mean" for simplicity. For example, Corollary 5 allows us to write

$$\frac{\frac{1}{J}\sum_{y=1}^{J}[\hat{S}^C(y) - \hat{S}^B(y)]}{\frac{1}{J}\sum_{y=1}^{J}[\hat{S}^A(y) - \hat{S}^B(y)]} = \frac{\bar{Y}^C - \bar{Y}^B}{\bar{Y}^A - \bar{Y}^B},$$

where $\bar{Y}^A$ and $\bar{Y}^B$ are the sample means when coding Y with cardinal values $\{1, 2, \ldots, J\}$, and $\bar{Y}^C$ is similarly the mean of the estimated counterfactual distribution $\hat{F}^C$ with the same cardinal values.

Consider the standard Blinder–Oaxaca decomposition with the following notation. Let $\bar{Y}^A$ and $\bar{Y}^B$ denote the two sample means when coding $Y \in \{1, 2, \ldots, J\}$. Let $\bar{\boldsymbol{X}}^A$ and $\bar{\boldsymbol{X}}^B$ also denote sample means. These $\boldsymbol{X}$ vectors may include transformations of an original set of variables. Let $\hat{\boldsymbol{\beta}}^B$ be the OLS coefficient vector estimate. Let $\boldsymbol{Y}^B = (Y_1^B, Y_2^B, \ldots, Y_{n_B}^B)'$ be the column vector of observations Y_i^B for $i = 1, \ldots, n_B$. Let $\underline{\boldsymbol{X}}^B$ be the matrix with row i equal to the transpose of $\boldsymbol{X}_i^B$, so $\underline{\boldsymbol{X}}^B = (\boldsymbol{X}_1^B, \boldsymbol{X}_2^B, \ldots, \boldsymbol{X}_{n_B}^B)'$. The vector of residuals and its orthogonality property are

$$\hat{\boldsymbol{U}}^B \equiv \boldsymbol{Y}^B - \underline{\boldsymbol{X}}^B \hat{\boldsymbol{\beta}}^B \quad \text{with} \quad (\underline{\boldsymbol{X}}^B)' \hat{\boldsymbol{U}}^B = \boldsymbol{0}. \tag{A.1}$$

In the conventional Blinder–Oaxaca decomposition, the explained proportions in the

population and sample can respectively be written as

$$\rho_E \equiv \frac{\Delta_E}{\Delta_T} = \frac{[\mathrm{E}(\boldsymbol{X}^A) - \mathrm{E}(\boldsymbol{X}^B)]'\boldsymbol{\beta}^B}{\mathrm{E}(Y^A) - \mathrm{E}(Y^B)} = \frac{\mathrm{E}(\boldsymbol{X}^A)'\boldsymbol{\beta}^B - \mathrm{E}(Y^B)}{\mathrm{E}(Y^A) - \mathrm{E}(Y^B)},$$
$$\hat{\rho}_E = \frac{(\bar{\boldsymbol{X}}^A)'\hat{\boldsymbol{\beta}}^B - \bar{Y}^B}{\bar{Y}^A - \bar{Y}^B}. \tag{A.2}$$

It remains to show that the counterfactual distribution-based decomposition, when using OLS to estimate linear probability models for each category of Y, yields a term identical to $(\bar{\boldsymbol{X}}^A)'\hat{\boldsymbol{\beta}}^B$ in the numerator of (A.2), as seen below.

Now consider the OLS estimates of the counterfactual CDF. Define indicators $Z_j^B \equiv \mathbb{1}\{Y^B \leq j\}$ for $j \in \{1, \ldots, J-1\}$, so

$$Y^B = J - \sum_{j=1}^{J-1} Z_j^B. \tag{A.3}$$

Let $\hat{\boldsymbol{\gamma}}_j^B$ be the OLS coefficient vector estimate from regressing Z_j^B on $\boldsymbol{X}^B$. Analogous to $\boldsymbol{Y}^B$, let $\boldsymbol{Z}_j^B$ be the vector of n_B observations. For each $j \in \{1, \ldots, J-1\}$, the vector of residuals and its orthogonality property are

$$\hat{\boldsymbol{V}}_j^B \equiv \boldsymbol{Z}_j^B - \underline{\boldsymbol{X}}^B\hat{\boldsymbol{\gamma}}_j^B \quad \text{with} \quad (\underline{\boldsymbol{X}}^B)'\hat{\boldsymbol{V}}_j^B = \mathbf{0}. \tag{A.4}$$

Using the above and the orthogonality property of OLS residuals, we can derive the relationship between $\hat{\boldsymbol{\beta}}^B$ (from (A.1)) and the $\hat{\boldsymbol{\gamma}}_j^B$. Assume the constant term is the first element in vector $\boldsymbol{X}^B$; that is, $\boldsymbol{X}^B = (1, \ldots)'$, so the first column of matrix $\underline{\boldsymbol{X}}^B$ is all ones. Let $\boldsymbol{e}_1 \equiv (1, 0, \ldots, 0)'$ have the same length as the $\hat{\boldsymbol{\gamma}}_j^B$, and let $\mathbf{1} \equiv (1, \ldots, 1)'$ be a vector of n_B ones. Combining (A.3) and (A.4),

$$\boldsymbol{Y}^B = J\mathbf{1} - \sum_{j=1}^{J-1} \boldsymbol{Z}_j^B = J\mathbf{1} - \sum_{j=1}^{J-1}(\underline{\boldsymbol{X}}^B\hat{\boldsymbol{\gamma}}_j^B + \hat{\boldsymbol{V}}_j^B) = \underline{\boldsymbol{X}}^B \overbrace{\left(J\boldsymbol{e}_1 - \sum_{j=1}^{J-1}\hat{\boldsymbol{\gamma}}_j^B\right)}^{=\hat{\boldsymbol{\beta}}^B} + \sum_{j=1}^{J-1}(-\hat{\boldsymbol{V}}_j^B),$$

where the equality

$$J\boldsymbol{e}_1 - \sum_{j=1}^{J-1}\hat{\boldsymbol{\gamma}}_j^B = \hat{\boldsymbol{\beta}}^B \tag{A.5}$$

is implied by the orthogonality

$$(\underset{\sim}{\boldsymbol{X}}^B)' \sum_{j=1}^{J-1} (-\hat{\boldsymbol{V}}_j^B) = -\sum_{j=1}^{J-1} \overbrace{(\underset{\sim}{\boldsymbol{X}}^B)' \hat{\boldsymbol{V}}_j^B}^{=\mathbf{0}} = \mathbf{0}, \tag{A.6}$$

which follows from the orthogonality condition in (A.4).

Now consider the "mean" of the counterfactual distribution. Taking the expectation of (A.3) and using the linearity of the expectation operator,

$$\mathrm{E}(Y^B) = J - \sum_{j=1}^{J-1} \mathrm{E}(Z_j^B).$$

Thus, the estimated counterfactual "mean" is

$$\hat{\mathrm{E}}(Y^C) = J - \sum_{j=1}^{J-1} (\bar{\boldsymbol{X}}^A)' \hat{\boldsymbol{\gamma}}_j^B = (\bar{\boldsymbol{X}}^A)' \overbrace{\left(J\boldsymbol{e}_1 - \sum_{j=1}^{J-1} \hat{\boldsymbol{\gamma}}_j^B \right)}^{=\hat{\boldsymbol{\beta}}^B \text{ by (A.5)}} = (\bar{\boldsymbol{X}}^A)' \hat{\boldsymbol{\beta}}^B.$$

This final expression is identical to the term in the Blinder–Oaxaca decomposition numerator in (A.2), so the estimated explained proportion is identical. □

References

Abul Naga, R. H. and Yalcin, T. (2008). Inequality measurement for ordered response health data. *Journal of Health Economics*, 27(6):1614–1625.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Allison, R. A. and Foster, J. E. (2004). Measuring health inequality using qualitative data. *Journal of Health Economics*, 23(3):505–524.

Apouey, B. (2007). Measuring health polarization with self-assessed health data. *Health Economics*, 16(9):875–894.

Awaworyi Churchill, S., Munyanyi, M. E., Prakash, K., and Smyth, R. (2020). Locus of control and the gender gap in mental health. *Journal of Economic Behavior & Organization*, 178:740–758.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., and O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, 17(3):351–375.

Bauer, T. K. and Sinning, M. (2008). An extension of the blinder–oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis*, 92(2):197–206.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455.

Bound, J. (1991). Self-reported versus objective measures of health in retirement models. *Journal of Human Resources*, 26(1):106–138.

Breslau, J., Marshall, G. N., Pincus, H. A., and Brown, R. A. (2014). Are mental disorders more common in urban than rural areas of the United States? *Journal of Psychiatric Research*, 56:50–55.

Carrieri, V. and Jones, A. M. (2017). The income–health relationship 'beyond the mean': New evidence from biomarkers. *Health Economics*, 26(7):937–956.

Cartwright, K. (2021). Social determinants of the Latinx diabetes health disparity: a Oaxaca–Blinder decomposition analysis. *SSM - Population Health*, 15:100869.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Cutler, D. M. and Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence. NBER Working Paper 12352, National Bureau of Economic Research.

Demoussis, M. and Giannakopoulos, N. (2007). Exploring job satisfaction in private and public employment: Empirical evidence from Greece. *Labour*, 21(2):333–359.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Ettner, S. L. (1996). New evidence on the relationship between income and health. *Journal of Health Economics*, 15(1):67–85.

Even, W. E. and Macpherson, D. A. (1990). Plant size and the decline of unionism. *Economics Letters*, 32(4):393–398.

Fairlie, R. W. (2005). An extension of the blinder–oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement*, 30(4):305–316.

Farber, H. S. (1987). The recent decline of unionization in the United States. *Science*, 238(4829):915–920.

Foresi, S. and Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90(430):451–466.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4A, chapter 1, pages 1–102. Elsevier.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gamm, L., Stone, S., and Pittman, S. (2010). Mental health and mental disorders—a rural challenge: A literature review. *Rural Healthy People 2010*, 2:97–114.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Simes, M., Berman, R., Koenigsberg, S. H., and Kessler, R. C. (2021). The economic burden of adults with major depressive disorder in the United States (2010 and 2018). *PharmacoEconomics*, 39(6):653–665.

Hansen, B. E. (2022). Econometrics. Textbook draft.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition. Corrected 12th printing, January 13, 2017.

Hastings, S. L. and Cohn, T. J. (2013). Challenges and opportunities associated with rural

mental health practice. *Journal of Rural Mental Health*, 37(1):37–49.

Hauret, L. and Williams, D. R. (2017). Cross-national analysis of gender differences in job satisfaction. *Industrial Relations*, 56(2):203–235.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hlavac, M. (2022). *oaxaca: Blinder–Oaxaca Decomposition in R*. Social Policy Institute, Bratislava, Slovakia. R package version 0.1.5.

Huling, J. (2022). *fastglm: Fast and Stable Fitting of Generalized Linear Models using 'RcppEigen'*. R package version 0.0.3.

Human, J. and Wasem, C. (1991). Rural mental health in America. *American Psychologist*, 46(3):232–239.

Idler, E. and Cartwright, K. (2018). What do we rate when we rate our health? Decomposing age-related contributions to self-rated health. *Journal of Health and Social Behavior*, 59(1):74–93.

Jeppson, H. and Hofmann, H. (2023). Generalized mosaic plots in the ggplot2 framework. *The R Journal*, 14(4):50–78.

Jeppson, H., Hofmann, H., and Cook, D. H. (2023). *ggmosaic: Mosaic Plots in the 'ggplot2' Framework*. R package version 0.3.4.

Jones, A. M., Roemer, J. E., and Rosa Dias, P. (2014). Equalising opportunities in health through educational policy. *Social Choice and Welfare*, 43:521–545.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics*, 16(2):163–178.

Kaplan, D. M. (2023). PhD Core Econometrics II. Textbook draft.

Kaplan, D. M. and Zhao, W. (2023). Comparing latent inequality with ordinal data. *Econometrics Journal*, 26(2):189–214.

Kapteyn, A., Smith, J. P., and van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97(1):461–473.

King, G., Murray, C. J. L., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1):191–207.

Kino, S. and Kawachi, I. (2020). How much do preventive health behaviors explain education- and income-related inequalities in health? Results of Oaxaca–Blinder decomposition analysis. *Annals of Epidemiology*, 43:44–50.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272):1168–1194.

Kobus, M. and Miłoś, P. (2012). Inequality decomposition by population subgroups for ordinal data. *Journal of Health Economics*, 31(1):15–21.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., and Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American Journal of Epidemiology*, 157(2):98–112.

Lund, C., De Silva, M., Plagerson, S., Cooper, S., Chisholm, D., Das, J., Knapp, M., and Patel, V. (2011). Poverty and mental disorders: breaking the cycle in low-income and middle-income countries. *The Lancet*, 378(9801):1502–1514.

Madden, D. (2010). Gender differences in mental well-being: a decomposition analysis.

Social Indicators Research, 99(1):101–114.

Makdissi, P. and Yazbeck, M. (2017). Robust rankings of socioeconomic health inequality using a categorical variable. *Health Economics*, 26(9):1132–1145.

National Center for Health Statistics (2022). National Health Interview Survey. `https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm`. Public-use data file and documentation, accessed 2024.

Nicholson, L. A. (2008). Rural mental health. *Advances in Psychiatric Treatment*, 14(4):302–311.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709.

Palomin, A., Takishima-Lacasa, J., Selby-Nelson, E., and Mercado, A. (2023). Challenges and ethical implications in rural community mental health: The role of mental health providers. *Community Mental Health Journal*, 59(8):1442–1451.

Pan, J., Liu, D., and Ali, S. (2015). Patient dissatisfaction in China: What matters. *Social Science & Medicine*, 143:145–153.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., Prince, M., Rahman, A., Saraceno, B., Sarkar, B. K., De Silva, M., Singh, I., Stein, D. J., Sunkel, C., and Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157):1553–1598.

Pilipiec, P., Groot, W., and Pavlova, M. (2020). A longitudinal analysis of job satisfaction during a recession in the Netherlands. *Social Indicators Research*, 149(1):239–269.

Probst, J. C., Laditka, S. B., Moore, C. G., Harun, N., Powell, M. P., and Baxley, E. G. (2006). Rural-urban differences in depression prevalence: Implications for family medicine. *Family Medicine*, 38(9):653–660.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, 370(6522):eaay0214.

Sen, B. (2014). Using the Oaxaca–Blinder decomposition as an empirical tool to analyze racial disparities in obesity. *Obesity*, 22(7):1750–1755.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Wang, P. S., Berglund, P. A., Olfson, M., and Kessler, R. C. (2004). Delays in initial treatment contact after first onset of a mental disorder. *Health Services Research*, 39(2):393–416.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2nd edition.

Wu, Q. and Kaplan, D. M. (2025). Multiple testing of stochastic monotonicity. Working paper available at `https://kaplandm.github.io/`.

Zhang, H., Bago d'Uva, T., and van Doorslaer, E. (2015). The gender health gap in China:

A decomposition analysis. *Economics & Human Biology*, 18:13–26.