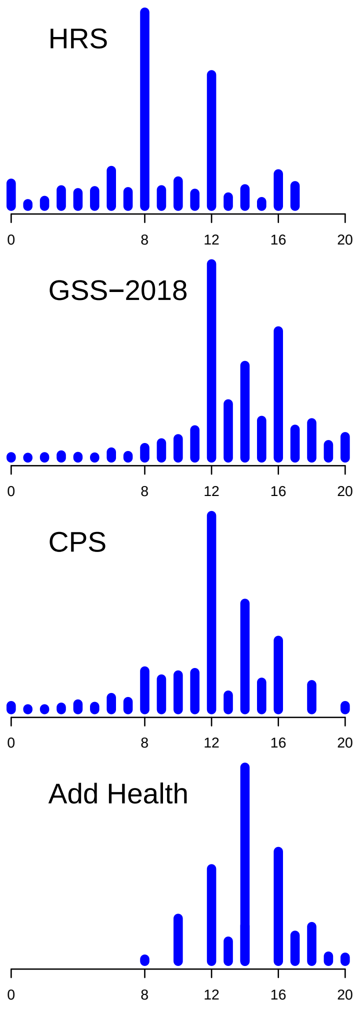
An ordinal model for analysis of years of education

Benjamin W. Domingue[1,†], Klint Kanopka[1], Sam Trejo[2], and Jeremy Freese[3]

[1]Stanford Graduate School of Education
[2]La Follette School of Public Affairs, University of Wisconsin Madison
[3]Department of Sociology, Stanford University
[†]ben.domingue@gmail.com

Years of education is a commonly used outcome variable in many lifecourse studies. We argue that such studies may derive additional insights from a treatment of years of education as an ordinal outcome rather than the standard treatment using the linear model. Via simulation, we show that the ordinal approach performs well if the linear model is actually the true model while, in the reverse scenario, estimates from the linear model may be somewhat suboptimal when the ordinal model is the true model. We use data from the Health and Retirement Study to illustrate additional insights that are readily derived from application of the ordinal model and offer a suggested workflow for future analysis.

Figure 1: Distribution of educational attainment in various data.



Introduction

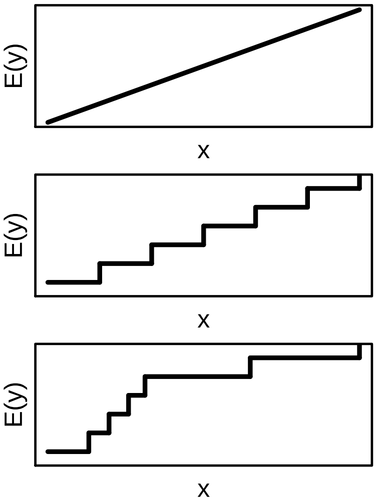
Numerous studies focus on years of education (i.e., educational attainment) as a key outcome. Educational attainment is an important variable given that one's level of schooling is the a frequent proxy for the culmination of development (and the inputs that affect it) early in the life and a key predictor of job market success, socioeconomic status, and health later in the life course. Given its centrality as an endpoint in understanding child development and wellbeing and its role as a key point of linkage between early-life and later-life, it is a key variable in many studies of education and health in fields such as psychology, sociology, and economics.

However, we suspect that the standard approach for modeling years of education as an outcome is flawed. Consider Figure 1 which shows histograms of the years of education in several datasets (CPS [1], GSS [2], Add Health [3], & HRS [4]). These distributions show a "clumping" of values that reflects the fact that people tend to leave school at specified end-points; in particular, note the cluster at 12 and 16 years reflecting completion of high school and 4-year college respectively. These distributions are clearly not normal. The standard approach used when dealing with this outcome—the linear model—does not assume that they are, only that they be conditionally normal. However, we argue that it is probably inappropriate to suppose that they are conditionally normal for most reasonable conditioning variables.

Further, there are reasons to suspect that treating the scale as "interval" may distort subsequent inference. That is, differences in the scale may not be consistently meaningful. Although this scale is an interval scale of time, it may not be "interval" in its association with other variables.[1] For example, an extra year of high school is quite inexpensive both in terms of out-of-pocket (i.e., there is no tuition at public schools) and opportunity costs while an extra year of college may be quite expensive in terms of out-of-pocket costs. Or, going from 11 to 12 years of schooling may be associated with a decrease in some disease later in life while there may be no predicted change associated with going from 12 to 13 years. Application of the linear model does not readily allow for us to identify such variation.

Given these facts, we consider an alternative to the standard approach for modeling educational attainment. Below we contrast the standard approach for modeling years of education with an alternative model based on treating educational attainment as an ordinal outcome. We then show (i) that the ordinal alternative behaves appropriately when the linear model is in fact true, (ii) that we can adjudicate between whether the linear or ordinal model in fact generated the data, and (iii) show that the linear model may produce suboptimal estimates when the ordinal model is in fact true. We then illustrate the potential benefits deriving from fitting the ordinal model to empirical data using examples from the HRS.

Figure 2: Conceptual models



Contrasting Conceptual Models

We begin with a discussion of the key conceptual components that we will consider below. These components—all based on differences in how the expectation of some outcome y varies as a function of a predictor x—are shown in Figure 2. Here, we are discussing the true association between x and y; following this, we turn to a discussion of how to model these differences. The top panel begins with a conventional scenario wherein $\mathbb{E}(y)$ is a linear function of x. In this case, the linear model is clearly the appropriate choice for understanding variation in y as a function of x.

Suppose that we do not observe such continuous variation in $\mathbb{E}(y)$. Instead, as in the second panel, consider a scenario wherein y is only observed at discrete (rather than continuously-varying) values. In this case, a linear approximation will be appropriate especially if the discrete elements of y are relatively close (so that the individual "steps" in $\mathbb{E}(y)$ are not "too big"). If there is some latent variable y'—

[1] In a technical sense, the issue here is not that the scale is non-interval but that associations are nonlinear (for example, see [5].

wherein y is merely a coarsened (so that it only takes integral values) version of y' and y' is a linear function of x—then this second panel would be the result. Moving to the third panel, we now observe variation in the size of the horizontal steps (the vertical steps are still homogeneous). In the terminology we introduce below, the underlying steps in x resulting in changes in $\mathbb{E}(y)$ are not "equidistant". Our argument is that a treatment of the differences in $\mathbb{E}(y)$ as ordinal in the last two panels may allow for more accurate understanding of the relationship between x and y.

We now turn to the question of what model to use in each case. As noted above, a linear model is clearly appropriate for the top panel. In subsequent panels, a linear approximation may be acceptable. This will be true especially if the steps in x and $\mathbb{E}(y)$ are relatively small. As we show below, our preferred ordinal approach will allow us to determine which of the panels is the most appropriate model at relatively little cost. Attempting to evaluate which of the panels in Figure 2 is relevant in a given scenario yields estimates with potentially improved statistical properties as well as important additional insight into the substantive issues in a given context as compared to just fitting the linear model suggested in the first panel.

Formal Models

We now introduce the relevant formal models underlying the issues shown in Figure 2. Throughout this discussion and the paper, we distinguish between the data-generating model (DGM) and the data-analytic model (DAM). The models we now introduce will act as both DGM and DAM. We focus on the simple case of a single predictor x.

Cumulative link model (CLM). We begin with a model for ordinal outcomes. The cumulative link model (CLM) [6], treats y_i as an ordinal variable rather than an intervally-scaled continuous variable.[2] This model asserts that

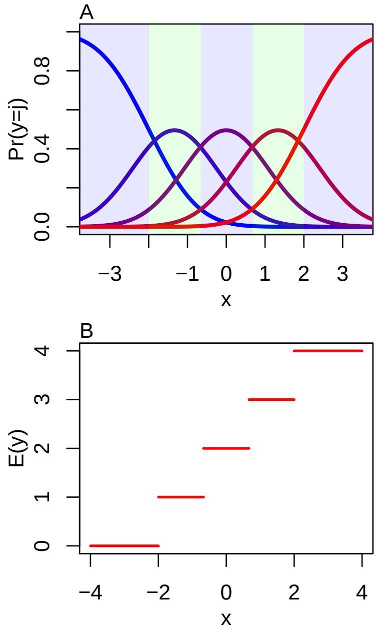
$$\Pr(y_i \leq j | x_i) = F(\theta_j - \beta x_i) \tag{1}$$

for each category j into which y might fall; for years of education, perhaps $j \in \{8, ..., 20\}$; note that the actual values of j are irrelevant and we will arbitrarily write $j \in \{1, ..., J\}$. The F function is a link function; here, we assume that the link is the probit, i.e., $F(\theta_j - \beta x_i) = \Phi(\theta_j - \beta x_i)$ where Φ is the CDF for the standard normal. Other choices are possible; we focus on the probit given that it offers a straightforward method for comparing estimates from the CLM to linear alternatives. As one key point of intuition, if y and x are positively correlated, then we will observe $\beta > 0$. That is, in-

[2] The CLM is also known as the ordered probit model when the normal CDF is used as the link function [7].

creased values of x will effectively translate into downward shifts in the thresholds.

Figure 3: Illustration of CLM. Curves represent $\Pr(y = j|x)$ where $J = 5, \beta = 1, x \sim \mathrm{N}(0,1)$, and $\theta = \{-2, -0.67, 0.67, 2\}$. Shaded areas represent regions between different values of θ. Color of curve (from blue to red) corresponds to increasing value of j.



It is perhaps more straightforward to understand Eqn 1 in terms of the probabilities of responses in each category. Using the notion of a "category response function" from item response theory [8], if we define the quantity in Eqn 1 as $\Pr(y_i \leq j|x_i) = \psi_j$ then we can write

$$\Pr(y_i = j|x_i) = \psi_j - \psi_{j-1}. \tag{2}$$

Figure 3 Panel A considers such curves for a specific set of parameters; note that for a given value of x, we have $\sum_j \Pr(y_i = j|x_i) = 1$ given that y_i must be in one of the j categories. For $j = 1$ and $j = J$ (the boundary categories), the curves are monotonically decreasing and increasing respectively thus reflecting the fact that very small and very large values of x are nearly certain to be in the first or last category respectively. For the middle categories, the curves represent small probabilities at the extremes with some local maxima. Figure 3 Panel B considers $\mathbb{E}(y)$ as a function of x based on the category response probabilities[3]; note the similarity to the second panel of Figure 2.

Linear model (LM). So as to contrast some of its features with the CLM, we now consider the linear model (LM) for years of education, $y = \beta_0 + \beta_1 x_i$. For comparison to the CLM, we derive an expression similar to Eqn 2 for the LM case. The standard version of this model asserts that the probability density associated with an observation of y_i is

$$f(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right). \tag{3}$$

This is the standard linear model, $\mathbb{E}(y_i|x_i) = \beta_0 + \beta x_i$, supplemented with the assumptions that errors are homoscedastic. We can then compute

$$\Pr(y \leq j|x_i) = \int_{-\infty}^{j} f(t|x_i)dt. \tag{4}$$

Using that, we derive

$$\Pr(y_i \leq j|x_i) - \Pr(y_i \leq j-1|x_i) = \int_{j-1}^{j} f(t|x_i)dt. \tag{5}$$

In a manner similar to that of Figure 3, we can compute these values; see Figure 4. These pseudo-category response functions[4] are similar to those in Figure 3, a function of the fact that both models rely on the CDF of the normal distribution and the fact that the θ values are evenly spaced in Figure 3. The key difference is that here there are not category response functions for the top and bottom values as the DGM here leads to unbounded outcome values; this is an important difference but we would also note that outcomes are typically

[3] What is the smallest j such that $\sum_{i \leq j} \psi_i > 0.5$?

[4] Because y_i here is not integral-valued.

bounded in many settings thus somewhat muting the differences between the approaches in real-world settings.

Extensions of the CLM. We now consider two extensions of the CLM. The first pertains to a special property of CLM shown in Figure 3. The θ values are evenly spaced (i.e., differences between consecutive values are equivalent); we now further discuss the role of these parameters. The θ values are frequently referred to as "thresholds". We emphasize their role in Figure 3 by changing the background shading at these values. If $F(0) = 0.5$ (which is true for the probit link), then θ_j reflects the value at which a value of $x\beta$ has even odds of being in category j or below versus category $j+1$ or above. Note that the differences in color do not perfectly correspond to points where the category response curves cross. This is due to the fact that the thresholds are relative to statements about where there are even odds for the cumulative probabilities not the category probabilities.

The thresholds need not be evenly distributed as they are in Figure 3. We illustrate the case of non-uniform thresholds in Figure 5. Here, values of $x \in [-2, 0.28]$ correspond to an expected value of 1; in contrast, there is a much smaller interval corresponding to $\mathbb{E}(y) = 2$ and an even smaller interval corresponding to $\mathbb{E}(y) = 3$. Thus, we can resolve the differences between panels 2 and 3 in Figure 2 into a question about the values of θ. In particular, we can test if they are equidistant; we refer to p-values from such a test as p_θ.

The second extension of the CLM is a way of allowing for flexibility into the CLM. In particular, note that Eqn 1 posits a uniform shift as a function of the covariate.[5] This need not be the case. We can relax this assumptions by allowing for "nominal effects" wherein the effect of a covariate varies across the categories. This model—the LCM with nominal effects (CLMn)—asserts that

$$\Pr(y_i \leq j | x_i) = F(\theta_j - \beta_j x_i). \tag{6}$$

In this model, the distances between thresholds effectively vary as a function of x (see Figure 2 in [6]). As we discuss below in the context of an empirical example, this extension allows for a variety of empirical features of our data to become apparent.

Simulation Studies

In the below simulation studies, we consider estimates generated via LM and CLM as DAM for different choices of the DGM. In particular, we consider the sequence of studies suggested in Table 1. We first use the LM as the DGM and show that the CLM acts functionally equivalently to the LM for a DAM. We then turn to the CLM as DGM

Figure 4: Illustration of pseudo-category response functions for LM (i.e., curves represent: $\Pr(y \leq j|x) - \Pr(y \leq j-1|x)$) where $\beta = 1$, $x \sim \mathrm{N}(0,1)$ and Eqn 3 is the DGM.

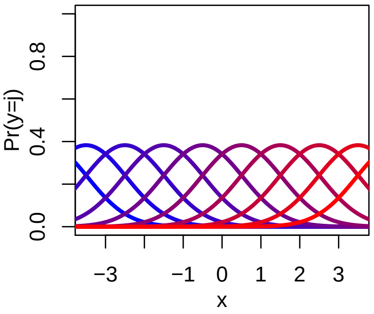
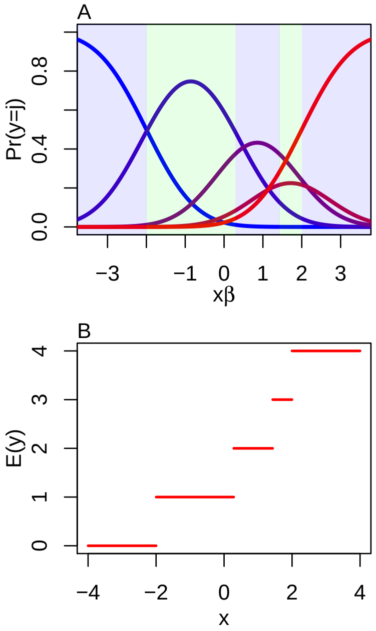


Figure 5: Illustration of CLM. Curves represent $\Pr(y = j|x)$ where $J = 5, \beta = 1$, $x \sim \mathrm{N}(0,1)$, and $\theta = \{-2, 0, 1, 1.5\}$. Shaded areas represent regions between different values of θ.



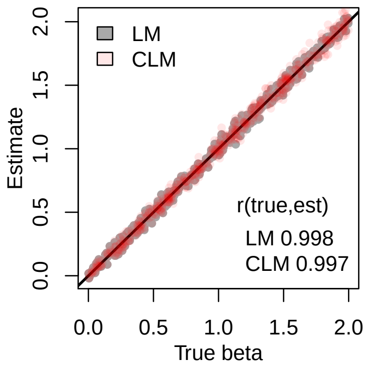
[5] This phenomenon pertains to the "proportional odds" assumption that the effect of x is constant across levels of y in the case of the logit link. Given our focus on the probit link, we do not use this language.

Table 1: Design of simulation studies

	DAM	
DGM	LM	CLM
LM		1
CLM	2	

and illustrate shortcomings of the LM when it is used as the DAM. Estimation of CLM is performed via [9].

Figure 6: Scatterplot of true β parameters versus LM and CLM estimates for $N = 1000$ and $\sigma^2 = 1$.



1. *When LM is the DGM, the CLM is an efficient alternative.*

We simulate data as $y_i^\star \sim \mathrm{N}(\beta x_i, \sigma^2)$ where $\beta \sim \mathrm{Unif}(0,2)$, $x \sim \mathrm{N}(0,1)$, and $\sigma^2 \in \{1,4,25\}$ and $1 \leq i \leq N$. We then round $y_i^\star$ to the nearest integer, this is y_i. Using y_i, we then estimate the LM and CLM; a comparison of point estimates relative to the true β values is in Figure 6 for a subset of the generating conditions. Both sets of estimates are relatively accurate (i.e., they track the 45 degree line). The CLM estimates have a slightly smaller correlation with the true values as compared to the LM estimates, but the difference is indeed quite small.

This evidence suggests that the CLM can be used in most situations without a severe reduction in estimation quality (this is even true for smaller sample sizes, see Table 2). To account for uncertainty in estimates, we also compare t statistics from LM and CLM estimates. When σ^2 is small, CLM test statistics are roughly 80% the size of their LM peers; differences decrease as σ^2 increases. Thus, estimates are nearly as precise when using CLM as LM and there is relatively little increase in uncertainty around these estimates.

Table 2: Correlations between true parameters and LM/CLM estimates (when LM is the DGM) for different values of N.

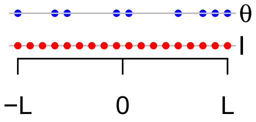
		$r(\text{true}, \text{est})$		Mean	
N	$\bar{J}$	LM	CLM	$\frac{t_{\text{CLM}}}{t_{\text{LM}}}$	$p_\theta < .05$
$\sigma^2 = 1$					
100	8	0.981	0.969	0.798	0.040
1000	10	0.998	0.997	0.800	0.056
10000	12	1.000	1.000	0.800	0.038
$\sigma^2 = 4$					
100	11	0.944	0.931	0.878	0.028
1000	15	0.994	0.993	0.875	0.036
10000	18	0.999	0.999	0.929	0.044
$\sigma^2 = 25$					
100	22	0.747	0.735	0.992	0.002
1000	32	0.967	0.966	0.988	0.032
10000	38	0.996	0.996	0.987	0.034

Further, we argue that any inadequacies of using the CLM as DAM when the LM is in fact the DGM can be mitigated by being able to detect the appropriate DGM (i.e., the LM). To do this, we can examine the θ values. In particular, the CLM can be used to adjudicate whether the LM is an appropriate model in a given scenario a test of whether such thresholds are equidistant. Under the LM, they should be given that

$$\Pr(y_i \leq j) - \Pr(y_i \leq j-1) = \Pr(y_i \leq j-1) - \Pr(y_i \leq j-2) \qquad (7)$$

for all values of j. In the final column of Table 2, we show the proportion of tests of equidistant thresholds that would be rejected when the LM is the DGM. Values are very near the alpha level (0.05) for modestly sized samples suggesting that this test can be used to reliably identify cases wherein thresholds are equidistant and, consequently, the LM can be used as DAM.

Figure 7: A schematic comparing l and θ for $J = 10$ and $s = 2$.



2. *When the CLM is the DGM, LM-based estimates are less accurate.*

In contrast, when the LM is not the DGM, estimates derived from the LM can be suboptimal. We illustrate this by simulating data from the CLM where $\beta \sim \mathrm{Unif}[0,1]$, $J \in \{5, 10, 20\}$, and values of θ_j are manipulated via two parameters, L and s. The first parameter controls the range over which the non-boundary categories predominate. The second parameter controls the level of regularity in the sequence of thresholds (for $s = 1$, the thresholds are evenly spaced). Details are as follows (a schematic comparing a hypothetical l and θ for $J = 10$ and $s = 2$ is shown in Figure 7):

- For a given J, we first create a lattice of $s(J-1)$ evenly spaced θ_j values between $-L$ and L. Denote these points as $l = \{l_1, ..., l_{s(J-1)}\}$ (and note that $l_1 = -L$ and $l_{s(J-1)} = L$ by design).
- Create a subset of lattice points, l' by removing l_1 and $l_{s(J-1)}$.
- From the remaining lattice points in $l' = \{l_2, ..., l_{s(J-1)-1}\}$, we sample $J - 3$ values. Define this as l_{sample}. Note that when $s = 1$, $l' = l_{\text{sample}}$.
- Define the set of θ values as the set $\{l_1, l_{\text{sample}}, l_{s(J-1)}\}$.

Note that higher values of s lead to sets of θ values that are increasingly sparse subsets of l; that is, Figure 3 would correspond to $s = 1$ whereas Figure 5 is consistent with $s > 1$.

Results for simulations based on 1000 iterations for each configuration of parameters are in Table 3. Estimates from the CLM are generally more highly correlated with the true parameters than are estimates from LM although differences are relatively small for larger N. However, even for larger N, estimates from LM can be substantially noisier if either L or s is relatively large especially for small J. When $N = 1000$ and $J = 5$, LM-based estimates are relatively weakly correlated, $r = 0.66$, with true parameters when $L = 4$ and $s = 3$ whereas the CLM-based estimates are highly correlated with true parameters, $r = 0.97$. Even when $J = 10$, estimates from LM have an appreciably larger amount of noise when $L = 4$ and $s = 3$.

Table 3: Correlations (r) between LM and CLM estimates when the CLM is the DGM.

			$r(\text{true}, \text{est})$	
J	L	s	LM	CLM
$N = 250$				
5	2	1	0.97	0.97
5	2	3	0.93	0.96
5	4	1	0.96	0.96
5	4	3	0.66	0.90
10	2	1	0.98	0.97
10	2	3	0.96	0.97
10	4	1	0.97	0.97
10	4	3	0.84	0.96
20	2	1	0.97	0.97
20	2	3	0.97	0.97
20	4	1	0.98	0.97
20	4	3	0.92	0.97
$N = 1000$				
5	2	1	0.99	0.99
5	2	3	0.96	0.99
5	4	1	0.99	0.99
5	4	3	0.66	0.97
10	2	1	0.99	0.99
10	2	3	0.97	0.99
10	4	1	0.99	0.99
10	4	3	0.88	0.99
20	2	1	0.99	0.99
20	2	3	0.98	0.99
20	4	1	0.99	0.99
20	4	3	0.94	0.99
$N = 10000$				
5	2	1	1.00	1.00
5	2	3	0.96	1.00
5	4	1	1.00	1.00
5	4	3	0.70	1.00
10	2	1	1.00	1.00
10	2	3	0.98	1.00
10	4	1	1.00	1.00
10	4	3	0.88	1.00
20	2	1	1.00	1.00
20	2	3	0.99	1.00
20	4	1	1.00	1.00
20	4	3	0.94	1.00

Summary of Simulation Studies

Results from simulation suggest that if the LM is the DGM, then there is little to lose if we use the CLM as the DAM. Estimates from CLM are nearly as efficient as those from the LM. Further, we can determine whether the LM is appropriate by examining the CLM thresholds. In contrast, if the CLM is the DGM, estimates from the LM are less precise. The differences in many cases are modest but can be more pronounced in cases with either relatively small J or highly varying θ. Collectively, this evidence suggests that ordinal approaches can be considered in at least complementary roles (if not outright substitutes) alongside traditional approaches for analysis of potentially ordinal variables such as years of education.

We further illustrate the benefits of the ordinal approach below in empirical analysis. We demonstrate two key facts. We first show that thresholds are not equidistant which thus suggests, based on simulation evidence, that the CLM may produce more accurate estimates. We then use the CLMn to emphasize that the homogeneity assumption of the linear model is often suboptimal and that more precise inferences may be obtained from the CLMn approach.

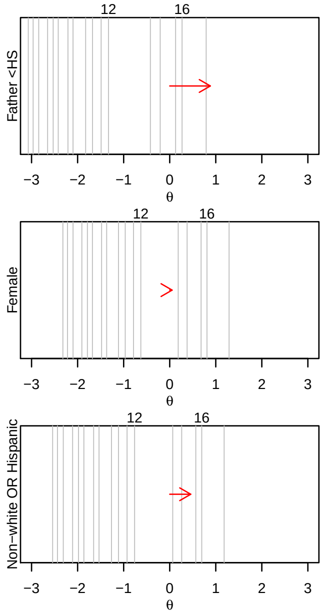
Empirical Illustration

We now illustrate the potential benefits derived from the CLM approach using data from the HRS [4]. The HRS is a biannual survey of Americans over 50 and their spouses; it is widely used to study the health and economic wellbeing of older Americans. This study asks about the number of years the respondent spent in school.[6]

We consider three predictors of years of education: father's not having a HS diploma, the respondent being female, and an indicator of the respondent's race and ethnicity (i.e., the respondent not being white or Hispanic). In the sample, 64% of respondents' fathers did not have a HS diploma, 56% were female, and 35% were non-white. Correlations between years of education and these variables are -0.37, -0.01, and -0.23 respectively. These correlations would be the effective outcomes of an approach based on LM; we now illustrate some additional insights that are easily derived using the ordinal approach.

We first estimate the CLM model in Figure 8. Consider the threshold estimates (i.e,. θ_j), shown as vertical gray lines. Note first the fact that, in all cases, the thresholds are not equidistant (this is visually apparent; all $p_\theta < 1^{-10}$). In all cases, the largest gap in the θ values occurs for the threshold between 12 and 13 years of education (i.e., no postsecondary schooling versus any attendance of postsecondary schooling). The non-equidistant thresholds suggest that the LM is not the DGM and thus the LM will not perform as well in terms of precise parameter recovery as will the CLM. The estimate of β from Eqn 1 is emphasized as a red arrow (note that the arrow points from 0 to $-\beta$ so as to emphasize the shifts in intercepts in the appropriate direction). All estimates are significant and suggest rightward shifts in the intercepts; going from a 0 to 1 on any predictor would effectively move the thresholds up by a uniform amount with the effect of increasing the proportion of the distribution at lower levels of education.

Figure 8: Estimates of θ and nominal offsets based on three models of educational attainment.



We now consider CLMn estimates, see Figure 9. The estimate from the CLM model—the red arrow in Figure 8—is shown as a vertical red line. Heterogeneity in the effect at the different steps—i.e.,

[6] See the raedyrs variable, https://www.rand.org/content/dam/rand/www/external/labor/aging/dataprod/randhrs1992_2016v2.pdf.

the β_j estimates in Eqn 6—are shown by the blue segments (which represent confidence intervals for the blue points). We focus on the patterns for being female or nonwhite. For being female, transitions at lower grades are below the mean effect and, in fact, below zero. This suggests a decrease in the thresholds for females. In contrast, at the higher levels, thresholds for females are relatively increased. Consequently, the distribution of the females' years of education is not well-characterized as a uniform shift of the males'. Rather, the female distribution has less variation than the male distribution; respondents with relatively little education are more likely to have a year or so more if they are female whereas relatively well-educated respondents are more likely to have a year or two less if female.

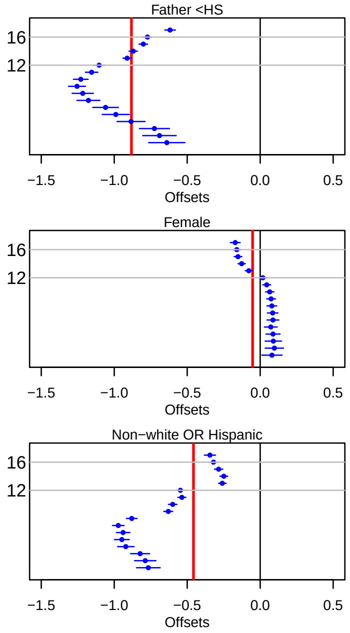
For nonwhite respondents, the effect of being nonwhite is a non-uniform but consistently negative shift in the offsets. The specific patterning of these results suggests that nonwhite respondents are much more likely to have fewer years of education but that being nonwhite being an increasingly muted predictor of educational differences amongst the well-educated.[7]

Discussion

Educational attainment is both a conceptually important variable and one that is easily and consistently measured in many kinds of surveys. Thus, it is widely used as an outcome. We present evidence to suggest that modeling it as an ordinal outcome may yield benefits. If a linear model is the DGM, the ordinal approach will be essentially equivalent in terms of resulting inferences and can also be used to test whether the interval-based LM approach is appropriate by testing whether thresholds are equidistant. If not, we show that there are benefits—both statistical and conceptual—to be had from application of the cumulative link model. The approach that we suggest could also be used to study alternative approaches to educational attainment (e.g., the same approach could be used if a person's highest educational credential is the outcome of interest).

We demonstrate the efficacy of our CLM approach using data from the HRS [4]. Although we considered relatively straightforward cases, the evidence presented suggests that CLM estimates would be more precise than LM estimates. Furthermore, the CLMn approach is able to identify nuanced higher-order effects that offer us increased insight about the associations in question. In particular, we are able to readily detect that the main difference between the male and female distributions of years of education is in the second moment and that the difference between years of education amongst white and nonwhite respondents is most pronounced at lower-levels of educa-

Figure 9: Comparison of CLM and CLMn estimates from HRS data. Main effect estimates from CLM are shown as vertical red lines. Category-specific offsets are shown as blue points (plus 95% confidence intervals).



[7] An alternative representation of these stylized facts could be framed via quantile regression [10]. If $Q_\tau(Y|X) = \beta_\tau$ indicates an estimated association of β_τ between an outcome Y and a predictor X at the τ-th quantile of Y, then: for female, β_τ would be positive for small τ and negative for large τ; for nonwhite respondents, β_τ would decline as a function of τ.

tion. There are, of course, other ways of noting these facts. However, we emphasize that these sets of findings both follow directly from a single analytic approach.

In the future, we would advocate for the following workflow when researchers consider educational attainment as an outcome:

- Fit CLMn. Examine the nominal effects.
  - If the nominal effects show substantial variation, use this saturated model
  - If not, go to:
- Fit CLM. Examine the thresholds.
  - If thresholds vary, use this model.
  - If thresholds are equidistant, go to:
- Utilize LM.

We think that this strategy for modeling educational attainment is fairly straightforward and easy to implement.

Future work may consider other issues; we emphasize two possibilities here. First, other variables should perhaps be treated similarly. One potential class of outcomes that may merit such treatment is sum scores. Such scores have even less claim to being interval [11] and are thus even more problematic from the perspective of handling them as continuous outcomes. The ordinal approach we advocate for here may yield benefits. One outcome in particular that may merit such treatment is depression which is frequently modeled as such a sum score. Previous work has suggested that such modeling may be ineffective due to indicator-specific effects [12] but we also note that such an outcome, if modeled using the LM, may also be suboptimal from a statistical perspective. Second, the cumulative link model offers additional flexibility that we do not discuss here. In particular, incorporation of scale effects [6] may allow for identification of additional higher-order effects that would further illuminate mechanisms of relevance for the study of educational attainment.

Acknowledgements

The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

References

[1] Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. Integrated public use microdata series, current population survey: Version 8.0 [dataset]., 2020.

[2] James A Davis and Tom W Smith. *The NORC general social survey: A user's guide*, volume 1. SAGE publications, 1991.

[3] Kathleen Mullan Harris, Carolyn Tucker Halpern, Eric A Whitsel, Jon M Hussey, Ley A Killeya-Jones, Joyce Tabor, and Sarah C Dean. Cohort profile: The national longitudinal study of adolescent to adult health (add health). *International Journal of Epidemiology*, 48(5):1415–1415k, 2019.

[4] Amanda Sonnega, Jessica D Faul, Mary Beth Ofstedal, Kenneth M Langa, John WR Phillips, and David R Weir. Cohort profile: the health and retirement study (hrs). *International journal of epidemiology*, 43(2):576–585, 2014.

[5] Jennifer Karas Montez, Robert A Hummer, and Mark D Hayward. Educational attainment and adult mortality in the united states: A systematic analysis of functional form. *Demography*, 49(1):315–336, 2012.

[6] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Journal of Statistical Software*, 2018.

[7] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*, volume 7. Stata press, 2006.

[8] Fumiko Samejima. Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer, 1997.

[9] R. H. B. Christensen. ordinal—regression models for ordinal data, 2019. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

[10] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[11] Ben Domingue. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1):1–19, 2014.

[12] Eiko I Fried and Randolph M Nesse. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC medicine*, 13(1):1–11, 2015.