Technical notes: Robust regression for at least ordinal outcomes

Förster, Martin

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:
Förster, M. (2025). *Technical notes: Robust regression for at least ordinal outcomes.* (Schriftenreihe für erweiterte Replikationen, Crowdsourcing und empirische Theorieüberprüfung, 4). https://doi.org/10.17605/OSF.IO/4HWXU

EReCT
No. 4, 2025
ISSN 2366-5041
Current version and supplementary material: DOI 10.17605/OSF.IO/4HWXU

Technical Notes: Robust regression for at least ordinal outcomes

Martin Förster*

February 28, 2025

If the task is to conduct a regression analysis in order to examine the statistical associations between variables, many scientists think, of course, of OLS first. And it makes perfect sense: it's probably the first regression technique you'll encounter - it's part of the basic repertoire of statistical analysis. Furthermore, OLS results are very informative. However, OLS needs a bunch of assumptions regarding the data at hand to be satisfied (Wooldridge 2010). If these assumptions are not met, results from OLS can be unstable, biased, or misleading. It is important to note that the assumptions of OLS are seldom fully met. Hence, to get stable results, we must apply robust regression techniques - that is, techniques which do not need some of the assumptions to be satisfied.

I will discuss two scenarios where alternative regression techniques provide more robust results compared to OLS. Both are about handling certain characteristics of the dependent variable. First, we consider a scenario where the measurement level of the dependent variable is ordinal. In the second scenario, the outcome is metric, but its distribution is strongly skewed. Finally, it is outlined how robust inference statistics can be achieved for both scenarios.

The techniques discussed are not only relevant but particularly so in the context of extended replications. Since replications aim to assess the stability of results, the application of alternative techniques aids in identifying methodological artifacts.

Modeling an ordinal outcome

Starting with a scenario where the outcome variable is ordinal, it is evident that OLS is unsuitable, as OLS requires the outcome variable to be at least interval-scaled. To apply a robust regression technique here means, in a very basic sense, just not to use OLS but an alternative.

As an ordinal variable, the values of the dependent variable represent ordered categories. Only the order of the categories has any meaning, the distances

*Europa-Universität Flensburg, ✉ martin.foerster@uni-flensburg.de

between the values are meaningless. Suppose an ordinal dependent variable Y with m categories. For each of the m categories we can assign a rank. Saying that category 2 (Y_2) is higher than category 1 (Y_1), and lower than category 3 (Y_3).

$$Y_1 < Y_2 < \ldots < Y_m$$

And that's it. We cannot say how much more or less is a category than another. But the OLS regression coefficient would try to tell us this: according to the model, by how many units is Y predicted to be larger or smaller if an independent variable X is one unit larger. By regression of an ordinal outcome, instead of modeling Y on a continuous value range, the leap from one category of Y to another is modeled. A positive or negative effect of X then means, that the higher the value of X, the higher or lower the category of Y predicted by the model respectively. In detail, ordinal regression models the probability of Y takes the higher category across all adjacent categories. From $k = 2$ to $k = m$, the probability

$$P\left(Y = k \mid X\right) = 1 - P\left(Y = k - 1 \mid X\right)$$

is determined by the model

$$P\left(Y = k \mid X\right) = g\left(\tau_k + BX\right)$$

where X is a vector of independent predictor variables and B the vector of the respective regression coefficients (each variable in X is assigned a corresponding coefficient in B). The model wasn't a model unless it was able to predict the specific category Y takes conditional on X. Therefore, the cut points $-\tau_k$ must be estimated. The cut points could be read like how large the scalar BX must be to predict Y is taking the categorical value k or a higher. The cut point of the first category is defined as $-\tau_1 = -\infty$. The scalar BX is always larger than minus infinity in the same way that Y is taking the smallest value or a higher, which is Y is taking a category at all. The link function is denoted by $g(\cdot)$.

As the modeling of Y is appropriate as an ordinal outcome, a model of the *cumulative* probabilities is estimated. Specifically, if $-\tau_{k-1} < -\tau_k$ for all categories, the model can also be formulated as follows:

$$P\left(Y \geq k \mid X\right) = g\left(\tau_k + BX\right)$$

Example I: I have selected the PISA 2022 data provided by the OECD (2022). Not the whole data from all participating countries is used but the German subset only. As dependent ordinal variable I used the proficiency level in the math domain (PLM). Six levels or categories of PLM were defined. For the ease of demonstration just one independent variable is used, specifically the interval scaled mathematics self-efficacy of the students ($MATHEFF$).

By having a look on the distribution of the dependent variable PLM (figure 1), the probit link function seems appropriate to model PLM: $g\left(\cdot\right) = \phi\left(\cdot\right)$.

Figure 1: PLM distribution

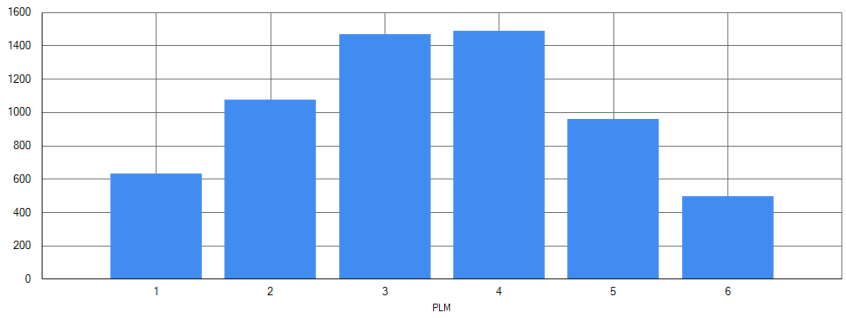


Table 1: Ordinal regression estimates (PLM model)

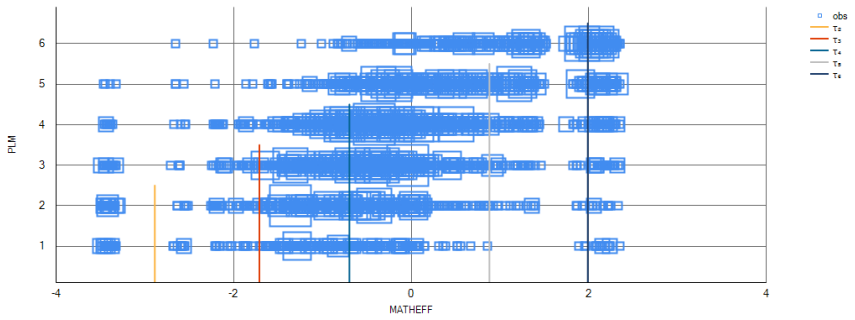
Parameter	**Estimate**
b_1	$0,257$
$-\tau_2$	$-0,743$
$-\tau_3$	$-0,4379$
$-\tau_4$	$-0,179$
$-\tau_5$	$0,226$
$-\tau_6$	$0,512$

So, the probability of PLM taking category k or a higher one, conditional on the independent variable $MATHEFF$, is modeled.

$$P\left(PLM \geq k \mid MATHEFF\right) = \phi\left(\tau_k + b_1 MATHEFF\right)$$

The estimates for the model are shown in table 1. At first, b_1, the regression coefficient of $MATHEFF$ shows a positive effect on PLM. The higher the mathematics self-efficacy of the students, the higher their proficiency level in the math domain. The first cut point $(-\tau_1)$, of course by definition, is minus infinity, since the probability of PLM taking level 1 or higher must be 1, regardless of $MATHEFF$. The second cut point is an estimate and it is $-0,743$. That is, to predict PLM taking category 2 or a higher, BX must exceed $-0,743$. Because in this example there is only one independent variable in the model, we can calculate the $MATHEFF$ score which equals this cut point. We doing so by dividing the respective cut point by b_1. For cut point τ_2 this threshold is a $MATHEFF$ score of $-2,89$. Each of the five estimated cut points was translated from the probit to the $MATHEFF$ scale the same way $(\frac{\tau_k}{b_1})$ (figure 2).

Figure 2: Plot PLM by $MATHEFF$ and cut points



Modeling a metric outcome

OLS assumptions also can be violated with a metric dependent variable, specifically by a heavily skewed or in some way non-symmetric distribution of Y. In such a case ordinal regression as just discussed is not an option, because the values of Y do not represent categories and the range of possible values can be very large, maybe infinite. But in an ordinal regression model cut points must be estimated for each value of Y. Doing this for a metric outcome variable would rise the number of parameters to be estimated and hence decreases the degrees of freedom dramatically. Robustness must therefore be introduced in another way, namely by transforming Y and by adjusting the regression technique.

First, consider the transformation of the dependent variable. In fact, Y is no longer supposed to used as dependent variable, but instead its *cumulative density*, represented by the mean relative ranks of the values of Y.

$$y_i^* = \frac{R(y_i)}{n}$$

$R(y_i)$ denotes the mean absolute rank of the value of the corresponding variable ($1 \leq R(y_i) \leq n$).

This transformation or alternative to the original dependent variable reduces skewness by flattening out the distribution. Additionally, the distribution of Y^* has no outliers.

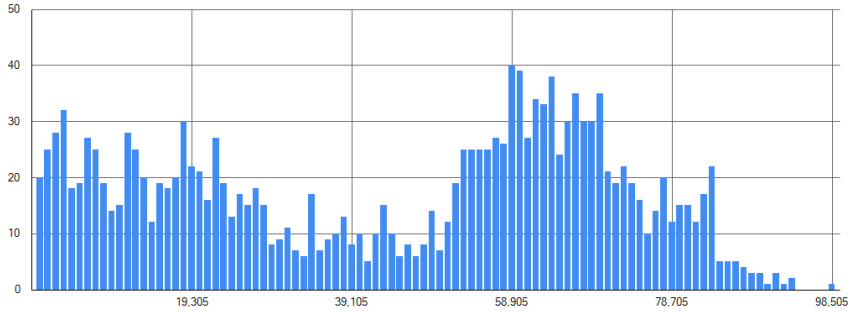
The model formulation is the conventional linear form: a constant value (b_0), a vector of independent variables (X), and a vector of regression coefficients (B).

$$\hat{Y}^* = b_0 + BX$$

As it is seen easily, this model is very parsimonious compared to an ordinal regression model, because it may not require estimating many cut points.

Next, the limited range of values of the (transformed) dependent variable ($0 < y_i^* \leq 1$) is considered by correcting the residuals ($e_i = y_i^* - \hat{y}_i^*$) while estimating the parameters of the model iteratively.

Figure 3: $dh17$ distribution



$$\hat{e}_i = \begin{cases} e_i + (\hat{y}_i^* - \max(y_i^*)) & \text{if } \hat{y}_i^* > \max(y_i^*) \\ e_i - (\min(y_i^*) - \hat{y}_i^*) & \text{if } \hat{y}_i^* < \min(y_i^*) \end{cases}$$

Finally, to stabilize the estimation, the importance of outliers in terms of residuals can be reduced. There are multiple methods to accomplish this. Among other things, the construction of weights and median regression. Concerning the application of weights, these can be derived from the corrected residuals in order to perform a reweighted least squares (RLS) regression.

$$w_i = \left((1 + \hat{e}_i) \sum_{i=1}^{n} \left(\frac{1}{(1 + \hat{e}_i)\, n} \right) \right)^{-1}$$

The weights (w_i) directly counteract the leverage effect of the residuals.

Another way is to perform median regression, that is to calculate the residuals relating to the conditional median (instead of residuals with respect to the conditional arithmetic mean). This regression technique also increases robustness, as the median is unreceptive to outliers.

Example II: To demonstrate this with an example, the ALLBUS 2023 data (GESIS Leibniz-Institut für Sozialwissenschaften 2025) were used. Once again, not the whole data set is used, but the subset of the eastern states of Germany. The dependent variable Y is the age of the youngest member of the household, $dh17$. As independent predictor variable X the *age* of the respondents is used. The level of measurement for both variables, $dh17$ and *age*, is the ratio scale.

The distribution plot of the dependent variable $dh17$ (figure 3) shows that it is strongly skewed. The first step was to transform $dh17$ into its mean ranks. While the distribution of $dh17^*$ does not conform to a normal distribution, it is less peculiar than before (figure 4). In a certain sense, the distribution of $dh17^*$ is symmetrical and no longer has any outliers. Figure 5 presents a direct comparison of the densities of the variables $dh17$ and $dh17^*$.

The model to be estimate is stated as follows:

Figure 4: $dh17$-$dh17^*$ transformation plot and $dh17^*$ distribution

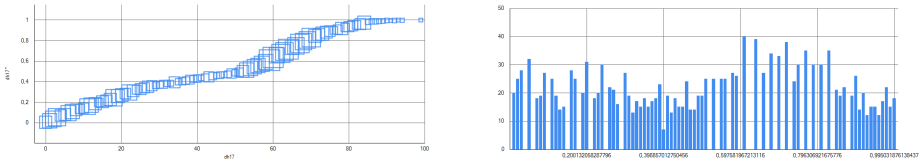
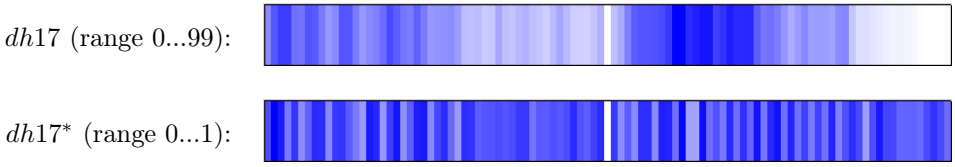


Figure 5: Density plots $dh17$ and $dh17^*$



$$\widehat{dh17}^* = b_0 + b_1 age$$

The estimates according to RLS and median regression are shown in table 2. The RLS estimate of the effect of *age* on $dh17$ is positive. The older the respondent, the older the youngest member of the respective household. With respect to the regression coefficient b_1, the median regression comes to the same result.

Inferential robustness: bootstrapping confidence intervals

Finally, regarding inferential statistics, bootstrapping the confidence intervals (CI) is a method for achieving robustness that works for both ordinal regression and cumulative density regression. To avoid making assumptions in terms of

Table 2: Cumulative density regression ($dh17$)

Parameter	**Estimate**
RLS	
b_0	-0.241
b_1	$0,014$
Median regression	
b_0	$-0,222$
b_1	$0,014$

Table 3: Estimate and bootstrapped quantiles (median, CI-quantiles)

		Bootstrap quantiles	
Parameter	**Estimate**	**Median**	**95% CI**
	Example I (PISA)		
b_1	$0,257$	$0,258$	$[0,223; 0,282]$
	Example II (ALLBUS)		
RLS			
b_1	$0,0138$	$0,0138$	$[0,0135; 0,0139]$
Median regression			
b_1	$0,014$	$0,014$	$[0,013; 0,015]$

distribution, I suggest the corrected percentile method (Efron 1981). According to this, a 95% CI for a regression coefficient b can determined as per

$$CI = \left[\widehat{CDF}^{-1}\left(\phi\left(2z - 1,96\right)\right),\ \widehat{CDF}^{-1}\left(\phi\left(2z + 1,96\right)\right)\right],$$

where the correction term $z = \phi^{-1}\left(P\left(\hat{b} < b\right)\right)$. The bootstrapped estimate is denoted by $\hat{b}$, the estimate from the original sample is denoted by b. $\widehat{CDF}^{-1}$ is the quantile of the bootstrap distribution of the regression coefficient.

Examples I and II *(continued):* For each regression analysis, the obtained confidence intervals (table 3) indicating statistically significant effects: the proficiency level in the math domain is positively affected by the mathematics self-efficacy (b_1 in example I), and the effect of the respondents age on the age of the youngest member of the household (b_1 in example II) is also positive.

References

Efron, Bradley (Jan. 1981). "Nonparametric standard errors and confidence intervals". In: *Canadian Journal of Statistics* 9.2, pp. 139–158. ISSN: 1708-945X. DOI: 10.2307/3314608.

GESIS Leibniz-Institut für Sozialwissenschaften (2025). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUScompact 2023*. de. DOI: 10.4232/1.14481.

OECD (2022). *Programme for International Student Assessment (PISA)*.

Wooldridge, Jeffrey M. (2010). *Econometric analysis of cross section and panel data*. Second edition. Cambridge, Massachusetts: MIT Press. 1974 pp. ISBN: 9780262296793.