Residuals and Diagnostics for Ordinal Regression Models" An Introduction to the sure package

by Brandon Greenwell, Andrew McCarthy, Bradley Boehmke and Dungang Liu

Abstract Residual diagnostics is an important topic in the classroom, but it is less often used in practice. Part of the reason for this is that more complex models, like cumulative link models and logistic regression, do not produce standard residuals that are easily interpreted as those in ordinary linear regression. In this paper, we introduce the concept of surrogate residuals and demonstrate there use through the R package **sure**.

Introduction

Categorical outcomes are encountered frequently in practice across different fields. For example, in medical studies, the outcome of interest is often binary (e.g., presence or absence of a particular disease after applying a treatment). It is also not uncommon for a categorical outcome to have a natural ordering. For instance, in an opinion poll, the response may be satisfaction such as low, medium, and high. In this case, the response is ordered: low < medium < high.

The *cumulative link* model is a natural choice for modelling an ordinal outcome. Consider an ordinal categorical outcome $\mathcal{Y}$ with ordered categories $1 < 2 < \cdots < J$. In a cumulative link model, the cumulative probabilities are linked to the linear predictor according to

$$G^{-1}\left(\Pr\left\{\mathcal{Y} \le j\right\}\right) = \alpha_j + f\left(\boldsymbol{X}, \boldsymbol{\beta}\right), \tag{1}$$

where G is a continuous cumulative distribution function, α_j are the category-specific intercepts, $\boldsymbol{X}$ is a matrix of covariates, and $\boldsymbol{\beta}$ is a vector of fixed regression coefficients. The intercept parameters satisfy $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J = \infty$. We should point out that some authors (and software) use the alternate formulation

$$G^{-1}\left(\Pr\left\{\mathcal{Y} \ge j\right\}\right) = \alpha_j^\star + f\left(\boldsymbol{X}, \boldsymbol{\beta}^\star\right), \tag{2}$$

This formulation provides coefficients that are consitent with the ordinary logistic regression model. The estimated coefficients from model (2) will have the opposite sign as those in model (1).

Another way to interpret the cumulative link model is through a *latent* continuous random variable $\mathcal{Z} = -f\left(\boldsymbol{X}, \boldsymbol{\beta}\right) + \epsilon$, where ϵ is a continuous random variable with location parameter 0, scale parameter 1, and cumulative distribution function $G\left(\cdot\right)$. We then construct an ordered factor according to the rule

$$y = j \quad if \quad \alpha_{j-1} < z \le \alpha_j.$$

For $\epsilon \sim N\left(0, 1\right)$, this leads to the usual probit model for ordinal responses

$$\Pr\left\{\mathcal{Y} \le j\right\} = \Pr\left\{\mathcal{Z} \le \alpha_j\right\} = \Pr\left\{-f\left(\boldsymbol{X}, \boldsymbol{\beta}\right) + \epsilon \le \alpha_j\right\} = \Phi\left(\alpha_j + f\left(\boldsymbol{X}, \boldsymbol{\beta}\right)\right).$$

Common choices for the link function and the implied (standard) distribution for ϵ are described in Table 1.

Link	Distribution of ϵ	$G\left(y\right)$	$G^{-1}\left(p\right)$
logit[1]	logistic	$\exp\left(y\right) / \left[1 + \exp\left(y\right)\right]$	$\log\left[p / \left(1 - p\right)\right]$
probit	standard normal	$\Phi\left(y\right)$	$\Phi^{-1}\left(p\right)$
log-log	Gumbel (max)	$\exp\left[-\exp\left(-y\right)\right]$	$-\log\left[-\log\left(p\right)\right]$
complimentary log-log	Gumbel (min)	$1 - \exp\left[-\exp\left(y\right)\right]$	$\log\left[-\log\left(1 - p\right)\right]$
cauchit	Cauchy	$\pi^{-1}\arctan\left(y\right) + 1/2$	$\tan\left(\pi p - \pi/2\right)$

Table 1: Common link functions.

There a number of R packages that can be used to fit cumulative link models (1). The recommended package **MASS** (Venables and Ripley, 2002) has the function `polr` (proportional odds logistic regression) which, despite the name, can be used with all of the above link functions. The **VGAM** package

(Yee, 2017) has the `vglm` function for fitting vector generalized linear models, which includes the broad class of cumulative link models. By default, `vglm` uses the same parameterization as in Equation (1), but provides the option for fitting (2) instead; this will result in the estimated coefficients having the opposite sign. Package **ordinal** (Christensen, 2015) has the `clm` function for fitting cumulative link models. The popular **rms** package (Harrell Jr, 2017) has two functions: `lrm` for fitting logistic regression models and cumulative link models of the form (2) using the logit link, and `orm` for fitting ordinal regression models of the form (2).

For a continuous outcome $\mathcal{Y}$, the residual is traditionally defined as the difference between the observed and fitted values. For categorical outcomes, the residuals are more difficult to define, and few solutions have been proposed in the literature. Liu et al. (2009) proposed using the cumulative sums of residuals derived from collapsing the ordered categories into multiple binary outcomes. Unfortunately, this method leads to multiple residuals for the ordinal outcome and therefore difficult to interpret. Li and Shepherd (2012) showed that the sign-based statistic (SBS)

$$R_{SBS} = E\left\{sign\left(y - \mathcal{Y}\right)\right\} = Pr\left\{y > \mathcal{Y}\right\} - Pr\left\{y < \mathcal{Y}\right\}, \tag{3}$$

can be used as a residual for proportional odds regression models; these are referred to later by Li and Shepherd as *probability-based residuals*, but we will follow Liu and Zhang (2017) and refer to them as SBS residuals. For an overview of the theoretical and graphical properties of the SBS residual (3), see Liu and Zhang (2017). These are available in the **PResiduals** package (Dupont et al., 2016). A limitation with the SBS residuals is that they are based on a discrete outcome and hence, discrete themselves. This makes using them in various diagnostic plots far less useful.

Surrogate-based residuals

Liu and Zhang (2017) propose a new type of residual that is based on a continuous variable $\mathcal{S}$ that acts as a surrogate for the ordinaly outcome $\mathcal{Y}$.

$$R_{\mathcal{S}} = \mathcal{S} - E\left(\mathcal{S}|\boldsymbol{X}\right). \tag{4}$$

The benefit of the surrogate-based residual (4) is that is based on a continuous variable $\mathcal{S}$. As a consequence R_S will also be continuous. The continuous variable $\mathcal{S}$ is based on the conditional distribution of the latent variable $\mathcal{Z}$ given $\mathcal{Y}$. In particular, given $\mathcal{Y} = y$, Liu and Zhang (2017) show that $\mathcal{S}$ follows a trunacted distribution obtained by truncating the distribution of $\mathcal{Z} = -f\left(\boldsymbol{X}, \boldsymbol{\beta}\right) + \epsilon$ using the interval $\left(\alpha_{y-1}, \alpha_y\right)$.

If the assumed model agrees with the true model, then the following hold:

symmetry around zero $E\left(R_{\mathcal{S}}|\boldsymbol{X}\right) = 0$;

homogeneity $Var\left(R_{\mathcal{S}}|\boldsymbol{X}\right)$ is constant and independent of $\boldsymbol{X}$;

reference distribution the empirical distribution of $R_{\mathcal{S}}$ approximates an explicit distribution that is related to the link function.

These properties allow for a thorough examination of the residuals to check model adequacy and misspecification of the mean structure and link function.

Jittering for general models

The latent method discussed in Section 2.2 applies to cumulative link models for ordinal outcomes. For more general models, we can define a surrogate using a technique called *jittering*. Suppose the true model for a categorical outcome $\mathcal{Y}$

$$\mathcal{Y} \sim F_a\left(y; \boldsymbol{X}, \boldsymbol{\beta}\right), \tag{5}$$

where $F\left(\cdot\right)$ is a discrete cumulative distribution function. This model is general enough to cover the cumulative link model (1), and nearly any pararmetric or nonparametric model for categorical outcomes (e.g., logistic regression).

Liu and Zhang (2017) suggest defining the surrogate $\mathcal{S}$ using either of the following two approaches:

1. jittering on the outcome scale: $\mathcal{S}|\mathcal{Y} = y \sim \mathcal{U}\left[y, y+1\right]$;
2. jittering on the probability scale: $\mathcal{S}|\mathcal{Y} = y \sim \mathcal{U}\left[F_a\left(y-1\right), F_a\left(y\right)\right]$.

Once a surrogate is obtained, we define the surrogate residuals in the same way as Equation 4. In either case, if the hypothesized model is correct, then symmetry around zero still holds; that is $E\left(R_{\mathcal{S}}|\boldsymbol{X}\right) = 0$.

For the later case, if the hypothesized model is correct then $R_{\mathcal{S}}|\boldsymbol{X} \sim \mathcal{U}\left(-1/2, 1/2\right)$. In other words, jittering on the probability scale has the additional property that the conditional ditribution of $R_{\mathcal{S}}$ given $\boldsymbol{X}$ has an explicit form. This allows for a full examination of the distributional information of the residual.

Bootstrapping

Since surrogate residuals are based on sampling, additional error is introduced. One way to minimize this sampling error and help stabilize any patterns in diagnostic plots is to use the bootstrap (Efron, 1979).

The procedure for bootstrapping surrogate residuals is similar to the model-based bootstrap algorithm used in linear regression. To obtain the b-th boostrap replicate of the residuals, Liu and Zhang (2017) suggest the following algorithm:

Step 1 Perform a standard case-wise bootstrap of the original data to obtain the bootstrap sample $\left\{\left(\boldsymbol{X}_{1b}^{\star}, \mathcal{Y}_{1b}^{\star}\right), \dots, \left(\boldsymbol{X}_{nk}^{\star}, \mathcal{Y}_{nk}^{\star}\right)\right\}$.

Step 2 Using the procedure outlined in the previous section, obtain a sample of surrogate residuals $R_{\mathcal{S}_{1b}}^{\star}, \dots, R_{\mathcal{S}_{nb}}^{\star}$ using the bootstrapped data obtained in **Step 2**.

In diagnostic plots, ... For Q-Q plots, Liu and Zhang (2017) suggest using the median of the B empirical distributions.

Surrogate-based residuals in R

The **sure** package supports a variety of R packages for fitting cumulative link and other types of models. The supported packages and their corresponding functions are described in Table 2.

Package	Function(s)	Model	Parameterization
stats	`glm`	binary regression	NA
MASS	`polr`	cumulative link	$Pr\left\{\mathcal{Y} \le j\right\}$
rms	`lrm`	cumulative link	$Pr\left\{\mathcal{Y} \ge j\right\}$
	`lrm`	logistic regression	NA
	`orm`	cumulative link	$Pr\left\{\mathcal{Y} \ge j\right\}$
ordinal	`clm`	cumulative link	$Pr\left\{\mathcal{Y} \le j\right\}$
VGAM	`vglm`	cumulative link	$Pr\left\{\mathcal{Y} \le j\right\}$[2]

Table 2: Supported packages.

The **sure** package currently only exports three functions:

- `resids`—construct (surrogate-based) residuals for fitted model objects of class `"clm"`, `"polr"`, and `"vglm"`;
- `autoplot`—produce various diagnostic plots using **ggplot2** graphics (Wickham, 2009);
- `gof`—simulate p-values from a goodness-of-fit test.

In addition, the package also includes three simulated data sets: `df1`, `df2`, and `df3`. These data sets are used to demonstrate various uses of the surrogate residual approach throughout this paper.

Detecting a misspecified mean structure

For illustration, the data frame `df1` contains $n = 2000$ observations from the following cumulative link model:

$$Pr\left\{\mathcal{Y} \le j\right\} = \Phi\left(\alpha_j + \beta_1 X + \beta_2 X^2\right), \quad j = 1, 2, 3, 4, \tag{6}$$

where $\alpha_1 = -16$, $\alpha_2 = -12$, $\alpha_3 = -8$, $\beta_1 = 8$, $\beta_2 = -1$, and $X \sim \mathcal{U}\left(1, 7\right)$. These simulated data are available in the `df1` data frame from the **sure** package and are loaded automatically with the package; see `?df1` for details.. Below, we fit a (correctly specified) probit model using the `polr` function from the **MASS** package.

```
# Load required package(s)
library(MASS)
```

```
# Fit a cumulative link model with probit link
fit.polr <- polr(y ~ x + I(x ^ 2), data = df1, method = "probit")
```

The code chunk below obtains the SBS residuals (3) from the previously fitted probit model `fit.polr` using the **PResiduals** package and constructs a couple of diagnostic plots. The results are displayed in Figure 1.

```
# Load required package(s)
library(PResiduals)

# Obtain the SBS/probability-scale residusls
pres <- presid(fit.polr)

# Residual vs. covariate plot
p1 <- ggplot(data.frame(x = df1$x, y = pres), aes(x, y)) +
  geom_point(alpha = 0.5) +
  geom_smooth(color = "red", se = FALSE) +
  ylab("Probability-scale residual")

# Q-Q plot of the residuals
p2 <- ggplot(data.frame(y = pres), aes(sample = y)) +
  stat_qq(distribution = qunif, dparams = list(min = -1, max = 1), alpha = 0.5) +
  xlab("Sample quantile") +
  ylab("Theoretical quantile")

# Figure 1
grid.arrange(p1, p2, ncol = 2)
```

(**Note:** the reference distribution for the SBS residual is the $\mathcal{U}(-1, 1)$ distribution.) As can be seen in Figure 1, the SBS residuals, which are inherently discrete, often display unusual patterns in diagnostic plots, making them less useful as a diagnostic tool. There is a pattern for each of the $J = 4$ classes!

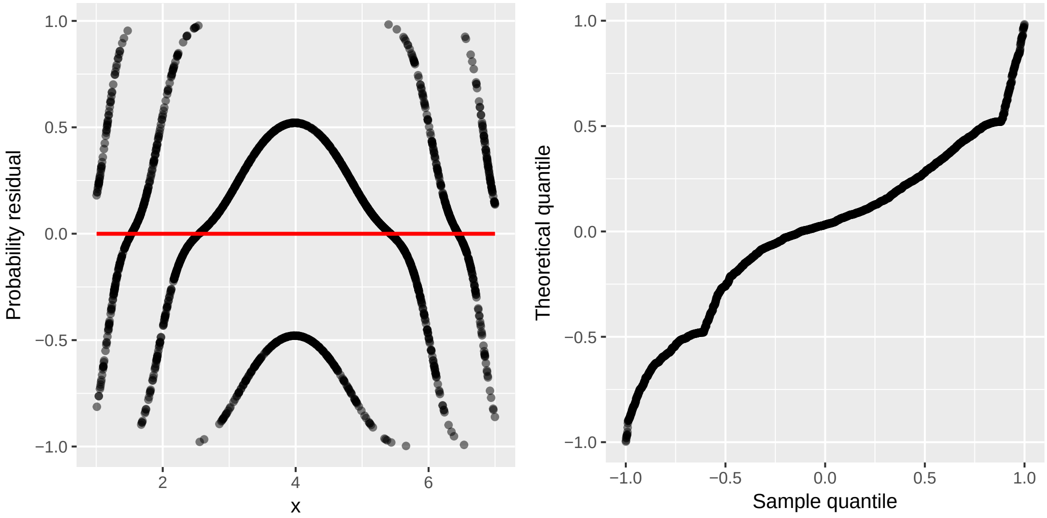

Figure 1: SBS residual plots for the (correctly specified) probit model fit to the `df1` data set. *Left*: Residual vs. covariate plot. *Right*: Q-Q plot of the residuals. Nonparametric smooths are indicated by red curves.

Similarly, we can use the `resids` function in package **sure** to obtain the surrogate-based residuals discussed in Section 2.2. This is illusratted in the following code chunk. the results are displayed in Figure 2.

```
# Load required package(s)
library(ggplot2)
library(sure)

# Obtain surrogate-based residuals
```

```
set.seed(101)  # for reproducibility
sres <- resids(fit.polr)

# Residual vs. covariate plot
p1 <- autoplot(sres, what = "covariate", x = df1$x, xlab = "x")

# Q-Q plot of the residuals
p2 <- autoplot(sres, what = "qq", distirbution = pnorm)

# Figure ?
grid.arrange(p1, p2, ncol = 2)
```

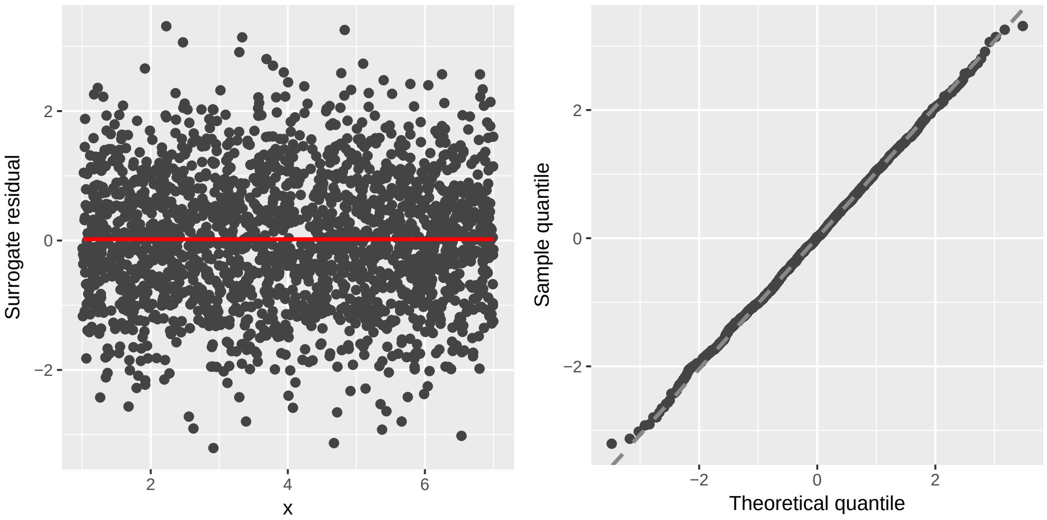

Figure 2: Surrogate-based residual plots for the (correctly specified) probit model fit to the `df1` data set. *Left*: Residual vs. covariate plot. *Right*: Q-Q plot of the residuals. Nonparametric smooths are indicated by red curves.

We also wrote `autoplot` methods for various classes of models listed in Table 2, so you can just give `autoplot` the fitted model directly. The benefit of this approach is that the fitted values and reference distirbution (used in quantile-quantile plots) are automatically extracted. For example, to reproduce the Q-Q plot in Figure 2, we could have just used

```
set.seed(101)  # for reproducibility
autoplot(fit.polr, what = "qq")  # same as top right of Figure 1
```

Suppose that we did not include the quadratic term in our fitted model. We could expect a residual-vs-x plot to clearly indicate that such a (correct) quadratic term is missing. Below we update the previously fitted model by removing the quadratic term, then update the residual-vs-covariate plots (code not shown). The updated residual plots are displayed in Figure 3.

```
fit.polr <- update(fit.polr, y ~ x)  # remove quadratic term
```

The SBS residuals gives some indication of a misspecified mean structure, but this only becomes more clear with increasing J, and the plot is still discrete. This is overcome by the surrogate residuals which produces a residual plot not unlike those seen in ordinary linear regresion models.

Detecting heteroscedasticty

One issue that oftens raises concerns in stataistical inference is that of heteroscedasticity; that is, when the error term has non constant variance. Heteroscedasticity can bias the statistical infeence and lead to improper standard errors, confidence intervals, and p-values. Therefore, it is imperative to identify heteroscedacticity whenever present and take appropriate action (e.g., transformations, etc.). In ordinary linear regression, this topic has been covered extensively. For categorical models, on the other hand, not much has been proposed in the literature.

As discussed in Section 2.2, one of the properties of the surrogate-based residual $R_{\mathcal{S}}$ is that, if the model is specified correctly, then $Var\left(R_{\mathcal{S}}|X\right) = c$, where c is a contant.

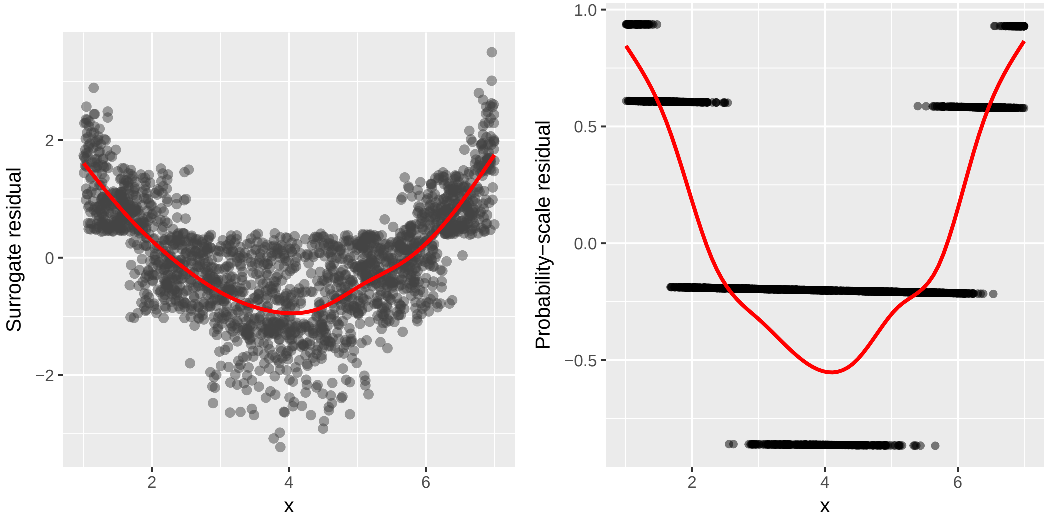

Figure 3: Residual-vs-covariate plots for a probit model with a misspecified mean structure fit to the simulated data from model (6). *Left*: Surrogate residuals. *Right*: SBS residuals. Nonparametric smooths are indicated by red curves.

For this example, we generated $n = 2000$ observations from the following ordered probit model:

$$Pr\left\{\mathcal{Y} \leq j\right\} = \Phi\left\{\left(\alpha_j + \beta X\right)/\sigma_X\right\}, \quad j = 1, 2, 3, 4, 5,$$

where $\alpha_1 = -36$, $\alpha_2 = -6$, $\alpha_3 = 34$, $\alpha_4 = 64$, $\beta = -4$, $X \sim \mathcal{U}(2, 7)$, and $\sigma_X = X^2$. Notice how the variability is an increasing function of X. These data are available in the df2 data frame that is automatically loaded with the **sure** package; see ?df2 for details.

The following block of code uses the orm function from the popular **rms** package to fit a probit model to the simulated data. **Note** that we had to set x = TRUE in the call to orm in order to use the presid function later.

```
# Load required package(s)
library(rms)

# Fit a cumulative link model with probit link
fit.orm <- orm(y ~ x, data = df2, family = "probit", x = TRUE)
```

If heteroscedasticity is present, we would expect this to show up in various diagnostic plots, such as a residual vs. covariate plot in this case. Below we obtain the SBS and surrogate residuals as before and plot them against X. The results are displayed in Figure 4.

```
pres <- presid(fit.orm)  # SBS residuals
set.seed(102)  # for reproducibility
sres <- resids(fit.orm)  # surrogate residuals

# Residual vs. covariate plots
p1 <- autoplot(sres, what = "covariate", x = df2$x, xlab = "x")
p2 <- ggplot(data.frame(x = df2$x, y = presid(fit.orm)), aes(x, y)) +
  geom_point(size = 2, alpha = 0.25) +
  geom_smooth(col = "red", se = FALSE) +
  ylab("Probability scale residual")

# Figure ?
grid.arrange(p1, p2, ncol = 2)
```

In this case, it is less clear that there is an issue with constant variance from the probability-scale residual plot...

Summary

TBD.

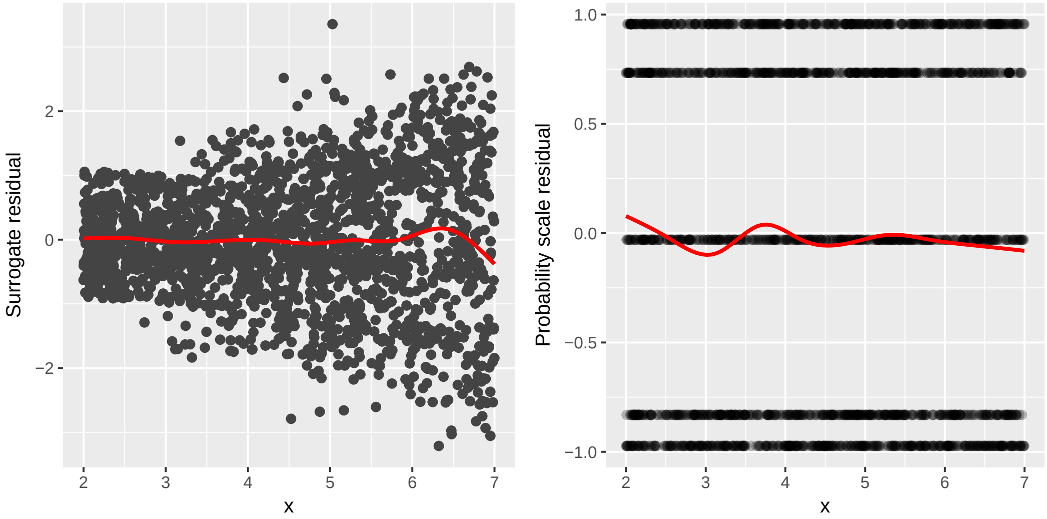

Figure 4: Residual vs. covariate plots for the simulated data. *Left*: Surrogate residuals. *Right*: SBS residuals.

Acknowledgments

TBD.

Bibliography

R. H. B. Christensen. ordinal—regression models for ordinal data, 2015. URL http://www.cran.r-project.org/package=ordinal. R package version 2015.6-28. [p2]

C. Dupont, J. Horner, C. Li, Q. Liu, and B. Shepherd. *PResiduals: Probability-Scale Residuals and Residual Correlations*, 2016. URL https://CRAN.R-project.org/package=PResiduals. R package version 0.2-4. [p2]

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals fo Statistics*, 7(1):1–26, 1979. URL http://dx.doi.org/10.1214/aos/1176344552. [p3]

F. E. Harrell Jr. *rms: Regression Modeling Strategies*, 2017. URL https://CRAN.R-project.org/package=rms. R package version 5.1-1. [p2]

C. Li and B. E. Shepherd. A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480, 2012. URL http://dx.doi.org/10.1093/biomet/asr073. [p2]

D. Liu and H. Zhang. Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, X(Y):XX–YY, 2017. URL http://dx.doi.org/10.1080/01621459.2017.1292915. [p2, 3]

I. Liu, B. Mukherjee, T. Suesse, D. Sparrow, and S. K. Park. Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in Medicine*, 28(3):412–429, 2009. URL http://dx.doi.org/10.1080/01621459.2017.1292915. [p2]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0. [p1]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org. [p3]

T. W. Yee. *VGAM: Vector Generalized Linear and Additive Models*, 2017. URL https://CRAN.R-project.org/package=VGAM. R package version 1.0-3. [p2]

Author One
Affiliation
Address

Country
(ORCiD if desired)
author1@work

Author Two
Affiliation
Address
Country
(ORCiD if desired)
author2@work

Author Three
Affiliation
Address
Country
(ORCiD if desired)
author3@work