Surrogate-based Residuals and Diagnostics in R: An Introduction to the sure package

by Brandon M. Greenwell, Author Two and Author Three

Abstract An abstract of less than 150 words.

Introduction

Categorical outcomes are encountered frequently in practice accross different fields. For example, in medical studies, the outcome of interest is often binary (e.g., presence or absense of a particular disease after applying a treatment). It is also not uncommon for a categorical outcome $\mathcal{Y}$ to have a natural ordering. For instance, in an opinon poll, the response may be satisfaction (e.g., $\mathcal{Y} \in \{Low, Medium, High\}$).

Logistic and probit regression are popular choices for modelling a binary outcome. The surrogate approach to constructing residuals actually applies to a wide class of general models of the form

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

where $F_a\left(\cdot\right)$ is a discrete cumulative distribution function. This includes binary regression as a special case. For example, the probit model has

$$\mathcal{Y} \sim bernoulli\left[\Phi\left(\boldsymbol{x}^\top \boldsymbol{\beta}\right)\right],$$

where $\Phi\left(\cdot\right)$ is the cumulative distribution function for the standard normal distribution.

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

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

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 $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J = \infty$. Common choices for the link function G^{-1} include:

logit: $G^{-1}\left(p\right) = \log\left[p/\left(1-p\right)\right]$;

probit: $G^{-1}\left(p\right) = \Phi^{-1}\left(p\right)$ (i.e., the quantile function for the standard normal distribution);

log-log: $G^{-1}\left(p\right) = \log\left[-\log\left(p\right)\right]$;

complimentary log-log: $G^{-1}\left(p\right) = \log\left[-\log\left(1-p\right)\right]$;

cauchit: $G^{-1}\left(p\right) = \tan\left(\pi p - \pi/2\right)$.

Another way to interpret the cumulative link model is through a *latent* continuous random variable $\mathcal{Z} = -\boldsymbol{X}\boldsymbol{\beta} + \epsilon$, 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

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

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

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

There a number of R packages that can be used to fit models of the form (1). The recommended package **MASS** (Venables and Ripley, 2002) has the function `polr` (proportional odds logistic regression) which can be used with all of the above link functions. The **VGAM** (Yee, 2017) package has the `vglm` function for fitting *vector generalized linear models*, which includes the broad class of cumulative link models. 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 which allows the response to be an ordinal factor, and `orm` for fitting ordinal regression models of the form (1).

For a continuous outcome, the residual is traditionally defined as the observed and fitted values. For categorical outcomes, the residuals are more difficult to define.

Very few residuals for ordinal regression models 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

$$E\left\{sign\left(y-\mathcal{Y}\right)\right\}=Pr\left\{y>\mathcal{Y}\right\}-Pr\left\{y<\mathcal{Y}\right\}, \tag{2}$$

can be used as a residual for proportional odds regression models. For an overview of the theoretical and graphical properties of (2), see Liu and Zhang (2017). These are available in the **PResiduals** package (Dupont et al., 2016). A limitation of the probability-scale residuals is that they are discrete...

Surrogate-based residuals

The problem with the LS 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. The surrogate based residual, on the other hand, is based on a continuous (unobserved) latent variable S and is far better suited for use in visual diagnostics.

Proposed in Liu and Zhang (2017).

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

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

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

reference distribution the emprical distribution of R 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 the more general model, we can define a surrogate through a technique called *jittering*. We offer two approaches for defining a surrogate S:

outcome scale $\mathcal{S}|\mathcal{Y}=y\sim\mathcal{U}\left[y,y+1\right]$;

probability scale $\mathcal{S}|\mathcal{Y}=y\sim\mathcal{U}\left[F_a\left(y-1\right),F_a\left(y\right)\right]$.

Then, we can define a residual by

$$R=S-E\left(S|\boldsymbol{X}\right).$$

For the case of binary regression, all we need are the fitted probabilities $G^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}\right)$ and the ability to simiulate from the uniform distribution. For example, if `fit.glm` is a `"glm"` object with the `binomial` family, and y is an integer representing the binary outcome in $\{0,1\}$, then the residuals from method (2) can be constructed as follows:

```
p1 <- pbinom(y - 1, size = 1, prob = fit.glm$fitted)  # F(y-1)
p2 <- pbinom(y, size = 1, prob = object$fitted)       # F(y)
runif(length(y), min = p1, max = p2) - 0.5            # S - E(S|X)
```

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

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

reference distribution for method (2) the $R|\boldsymbol{X}\sim\mathcal{U}\left[-1/2,1/2\right]$.

Both methods have the zero mean property, but only method (2) allows for examinination of the full distributional information in the residual.

Residual-based OR diagnostics in R

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

- `resids`—construct (surrogate-based) residuals for fitted model objects of class `"clm"`, `"polr"`, and `"vglm"`;

Package	Function(s)	Model	Parameterization
stats	`glm`	binary regression	NA
MASS	`polr`	cumulative link	$Pr\{\mathcal{Y} \leq j\}$
rms	`lrm`	cumulative link	$Pr\{\mathcal{Y} \geq j\}$
	`lrm`	logistic regression	NA
	`orm`	cumulative link	$Pr\{\mathcal{Y} \geq j\}$
ordinal	`clm`	cumulative link	$Pr\{\mathcal{Y} \leq j\}$
VGAM	`vglm`	cumulative link	$Pr\{\mathcal{Y} \leq j\}$[1]

Table 1: Supported packages.

- `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 demonstarte various uses of the surrogate residual appproach throughout this paper.

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

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

where $\alpha_1 = -16$, $\alpha_2 = -12$, $\alpha_3 = -8$, $\beta_1 = 8$, $\beta_2 = -1$, and $X \sim \mathcal{U}(1, 7)$. These simulated data are available in the `df.quadratic` data frame from the **sure** package. Below, we fit a (correctly specified) probit model using the `polr` function from the **MASS** package.

```
library(MASS)
data(df1, package = "sure")  # load the data
head(df1)  # inspect the first 6 rows
fit.polr <- polr(y ~ x + I(x ^ 2), data = df1, method = "probit")
```

The code chunk below obtains the probability-scale residuals (2) from the previously fitted probit model `fit.polr` using the **PResiduals** package.

```
library(PResiduals)
pres <- presid(fit.polr)  # probability-scale residuals
```

A couple diagnostic plots based on these residuals are shown in the bottom row of 1. (**Note:** the reference distribution for the probability-based residual is the $\mathcal{U}(-1, 1)$ distribution.) As can be seen, these residuals, which are inherently discrete, often display unusual patterns in diagnostic plots, making them less useful as a diagnostic tool for ordinal regression models.

Similarly, we can use the `resids` function in package **sure** to obtain the surrogate-based residuals. This is illustarted in the following code chunk. the results are displayed in the top row of Figure 1.

```
library(ggplot2)  # for autoplot function
library(sure)
sres <- resids(fit.polr)

# Residual vs covariate and Q-Q plots using the surrogate residuals
p1 <- autoplot(sres, what = "covariate", x = df1$x, xlab = "x")
p2 <- autoplot(sres, what = "qq", distirbution = pnorm)

# Residual vs covariate and Q-Q plots using the probability-scale residuals
p3 <- 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")
p4 <- 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, p3, p4, ncol = 2)
```

Alternatively, we wrote `autoplot` methods for various OR model classes, 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.

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

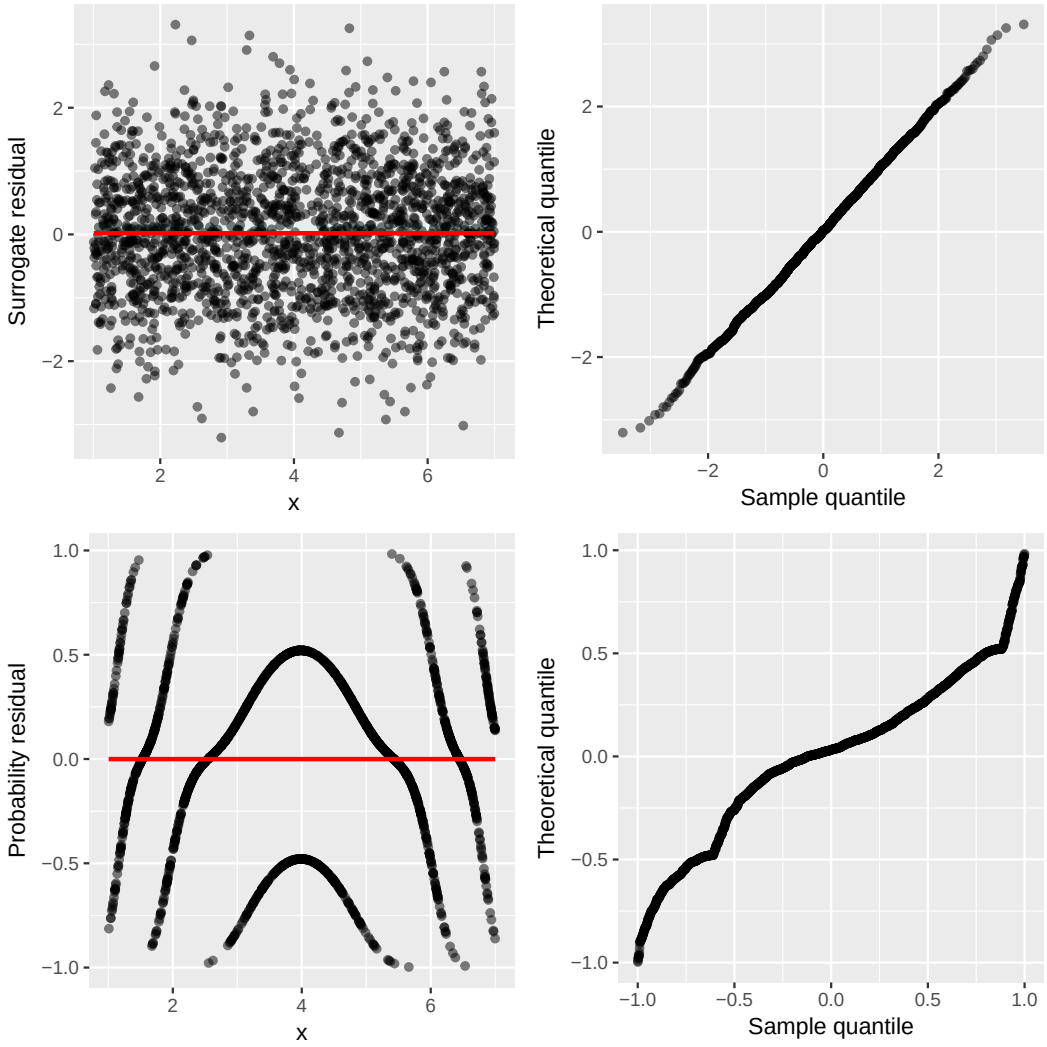

Figure 1: Various diagnostic plots for a (correctly specified) probit model fit to the simulated data from model (3). *Top left*: Surrogate residual vs. covariate plot. *Top right*: Quantile-quantile plot of the surrogate residuals. *Bottom left*: Probability-scale residual vs. covariate plot. *Bottom right*: Quantile-quantile plot of the probability-scale residuals.

Detecting a misspecified mean structure

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... The probability-scale residual 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 oversome by the surrogate residuals which produces a residual plot not unlike those seen in ordinary linear regresion models...

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

Detecting heteroscedasticty

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

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

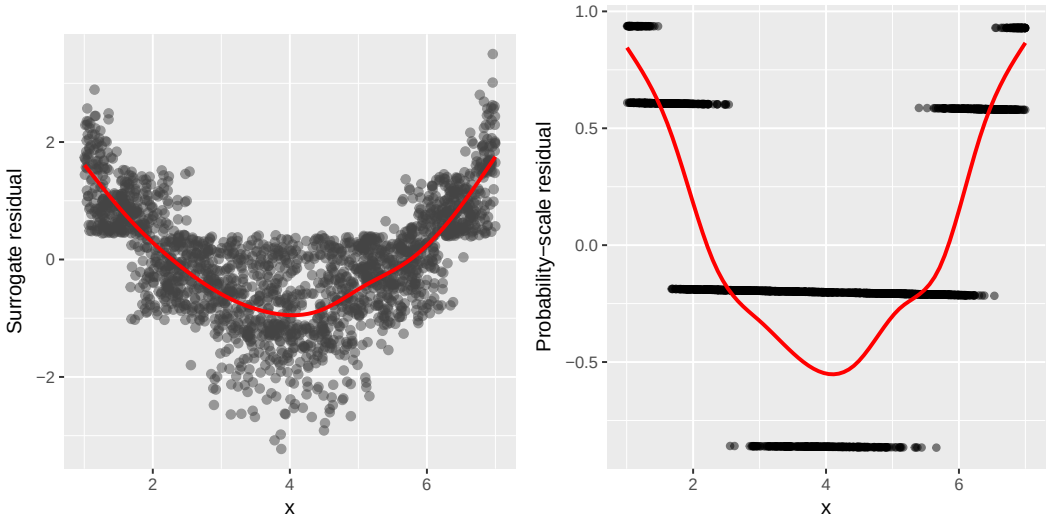

Figure 2: Various diagnostic plots for a probit model with a misspecified mean structure fit to the simulated data from model (3). *Left*: Surrogate residual vs. covariate plot. *Right*: Probability-scale residual vs. covariate plot.

where $\alpha_1 = -36$, $\alpha_2 = -6$, $\alpha_3 = 34$, $\alpha_4 = 64$, $\beta = -4$, $X \sim \mathcal{U}(2,7)$, and $\sigma_X = X^2$.

The following block of code uses the **MASS** package function `polr` to fit a probit model to the simulated `df2` data.

```
library(ggplot2)
library(MASS)
library(sure)
fit.polr <- polr(y ~ x, data = df2, method = "probit")
set.seed(101)  # for reproducibility
sres <- resids(fit.polr)  # surrogate-based residuals

# Figure 1 (left)
ggplot(data.frame(x = df2$x, y = sres), aes(x, y)) +
  geom_point(size = 2, alpha = 0.25) +
  geom_smooth(color = "red", se = FALSE) +
  ylab("Surrogate residual")
```

Alternatively, we can plot the residuals directly from the fitted model using the `autoplot` function:

```
  autoplot(fit.polr, what = "covariate", x = df2$x)  # plot not shown
```

We can also easily obtain and plot the standard Li-Shepherd residuals against x using the **PResiduals** package function `presid`:

```
library(PResiduals)
pres <- presid(fit.polr)  # probability-scale residuals

# Figure 1 (right)
ggplot(data.frame(x = df2$x, y = pres), aes(x, y)) +
  geom_point(size = 2, alpha = 0.25) +
  geom_smooth(color = "red", se = FALSE) +
  ylab("probability-scale residual")
```

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

Detecting a misspecified link function

For this example, we simulated $n = 2000$ observations from the quadratic model, but using a log-log link.

```
data(df3, package = "sure")
fit.probit <- polr(y ~ x + I(x ^ 2), data = df3, method = "probit")
```

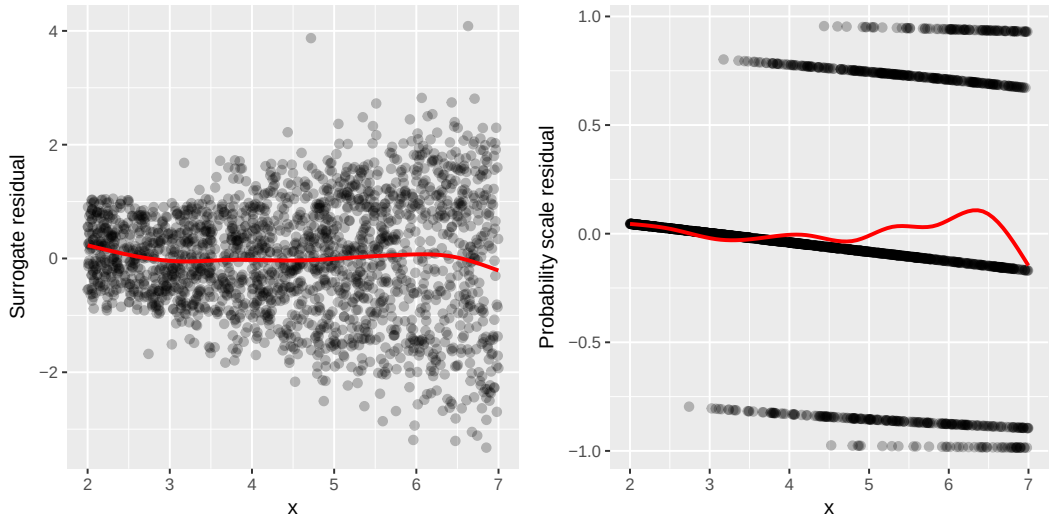

Figure 3: Residual vs. covariate plots for the simulated heteroscedastid data. *Left*: Surrogate residuals. *Right*: probability-scale residuals.

```
fit.logistic <- polr(y ~ x + I(x ^ 2), data = df3, method = "logistic")
fit.loglog <- polr(y ~ x + I(x ^ 2), data = df3, method = "loglog")  # correct link
fit.cloglog <- polr(y ~ x + I(x ^ 2), data = df3, method = "cloglog")

# Figure ?
p1 <- autoplot(fit.probit, nsim = 100, what = "qq")
p2 <- autoplot(fit.logistic, nsim = 100, what = "qq")
p3 <- autoplot(fit.loglog, nsim = 100, what = "qq")
p4 <-  autoplot(fit.cloglog, nsim = 100, what = "qq")
grid.arrange(p1, p2, p3, p4, ncol = 2)  # bottom left plot is correct model
```

Checking the proportionality assumption

Coming soon!

Assessing goodness-of-fit

Coming soon!

```
plot(gof(houses.polr, nsim = 1000))
```

Bitterness of wine

```
library(ordinal)
data(wine, package = "ordinal")
wine.clm <- clm(rating ~ temp * contact, data = wine)  # default logit link

set.seed(101)  # for reproducibility
grid.arrange(
  autoplot(wine.clm, nsim = 10, what = "qq"),
  autoplot(wine.clm, nsim = 10, what = "fitted"),
  autoplot(wine.clm, nsim = 10, what = "cov", x = wine$temp),
  autoplot(wine.clm, nsim = 10, what = "cov", x = wine$contact),
  ncol = 2
)
```

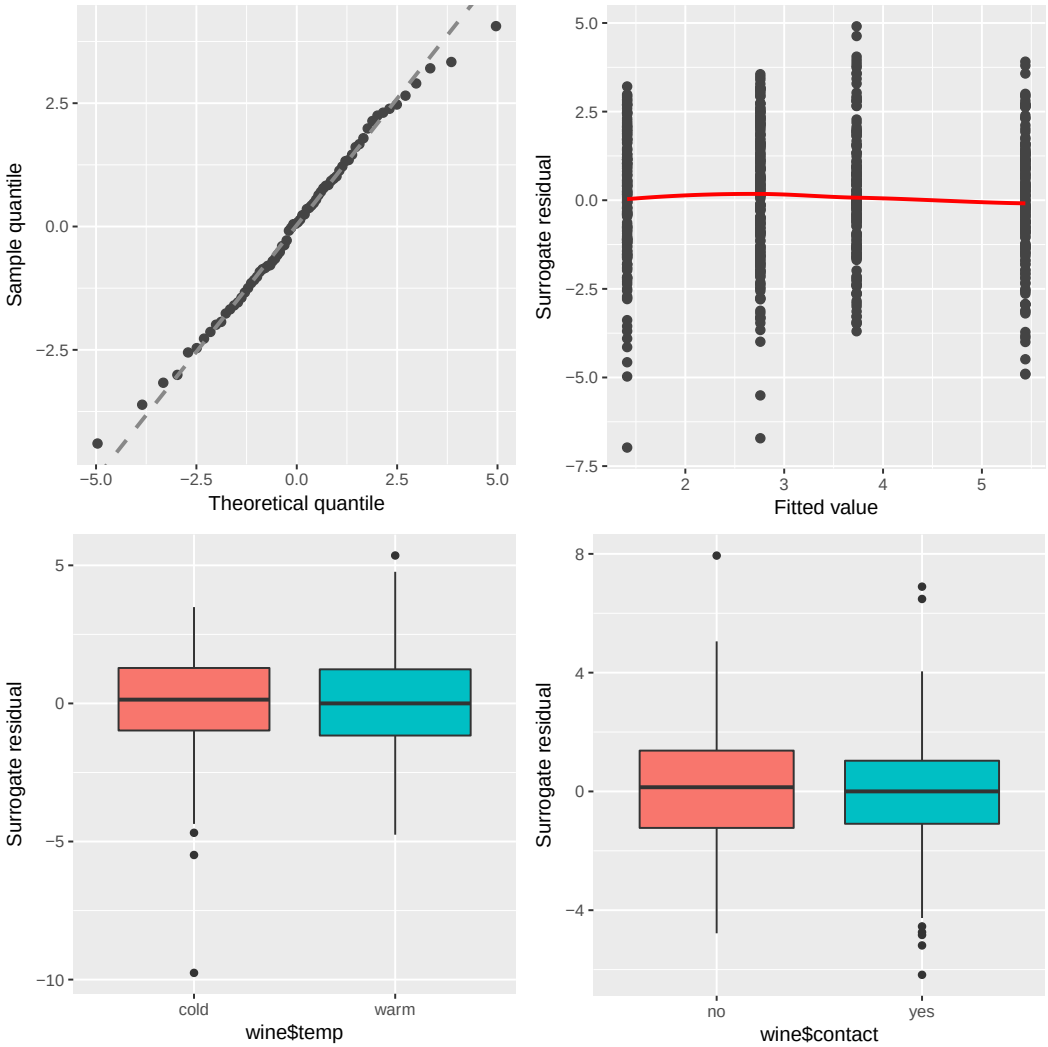

Figure 4: Residual diagnostic plots for the quality of wine example.

Summary

This file is only a basic article template. For full details of *The R Journal* style and information on how to prepare your article for submission, see the Instructions for Authors.

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. [p1]

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]

F. E. Harrell Jr. *rms: Regression Modeling Strategies*, 2017. URL https://CRAN.R-project.org/package=rms. R package version 5.1-1. [p1]

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*, ?(?):?–?, 2017. URL http://dx.doi.org/10.1080/01621459.2017.1292915. [p2]

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. [p1]

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