A unified framework for residual diagnostics in generalized linear models and beyond

Dungang Liu, Zewei Lin and Heping Zhang*

September 26, 2024

*Dungang Liu is Professor of Business Analytics and Academic Director for the Center for Business Analytics, Department of Operations, Business Analytics and Information Systems, Lindner College of Business, University of Cincinnati, Cincinnati, Ohio, USA (email: dungang.liu@uc.edu); Zewei Lin is Assistant Professor of Analytics, Department of Information Systems and Analytics, McCoy College of Business, Texas State University, San Marcos, Texas, USA (email: pqt19@txstate.edu); Heping Zhang is Susan Dwight Bliss Professor of Biostatistics, Professor in the Child Study Center and Professor of Statistics and Data Science, Professor of Obstetrics, Gynecology, and Reproductive Sciences, Yale University, New Haven, Connecticut, USA (email: heping.zhang@yale.edu).

A unified framework for residual diagnostics in generalized linear models and beyond

Abstract

Model diagnostics is an indispensable component in regression analysis, yet it has not been well addressed in generalized linear models (GLMs). When outcome data are discrete, classical Pearson and deviance residuals have limited utility in generating diagnostic insights. This paper establishes a novel diagnostic framework for GLMs and their extensions. Unlike the convention of using a point statistic as a residual, we propose to use a function as a vehicle to retain residual information. In the presence of data discreteness, we show that such a functional residual is appropriate for summarizing the residual randomness that cannot be captured by the structural part of the model. We establish its theoretical properties, which lead to the innovation of new diagnostic tools including the functional-residual-*vs*-covariate plot and Function-to-Function plot (similar to a Quantile-Quantile plot). Our numerical studies demonstrate that the use of these tools can reveal a variety of model misspecifications, such as not properly including a higher-order term, an explanatory variable, an interaction effect, a dispersion parameter, or a zero-inflation component. As a general notion, the functional residual considerably broadens the diagnostic scope as it applies to GLMs for binary, ordinal and count data as well as semiparametric models (e.g., generalized additive models), all in a unified framework. Its functional form provides a way to unify point residuals such as Liu-Zhang's surrogate residual and Li-Shepherd's probability-scale residual. As its graphical outputs can be interpreted in a similar way to those for linear models, our framework also unifies diagnostic interpretation for discrete data and continuous data.

Keywords: categorical data, count data, goodness of fit, Poisson regression, ordinal regression, QQ plot, residual plot, semiparametric model, surrogate residual.

1 Introduction

Model diagnostics is indispensable in statistical modeling. It helps domain practitioners assess their working models so as to better inform scientific conclusions and business decisions. However, except for the limited class of linear models, model diagnostics have not been well addressed when outcome data are discrete, and yet such outcome data are prevalent in scientific research, business practice, and policy making. For example, binary data are used to indicate disease prevention/treatment outcome or bankruptcy/default status; ordinal data are commonly seen in ratings of bonds, school districts, and pain severity; and integer-valued count data are prevalent in (electronic) records of insurance claims, emergency room visits, and frequency of product/device usage (Franses and Paap, 2001; De Jong and Heller, 2008; Frees, 2009; Faraway, 2016). When outcome data are discrete, generalized linear models (GLMs) are the common framework for statistical inference. But for these models, classical assessment methods, such as Pearson/deviance residual analysis and goodness-of-fit tests, have limited utility in generating insights and advising modeling strategies. The lack of an effective diagnostic tool results in an elevated risk of model misspecification and unmeasurable bias in inference. We establish a novel framework for diagnostics of GLMs with the goal of generating actionable insights in statistical modeling.

Traditional goodness-of-fit assessments of a regression model heavily rely on hypothesis testing (see, e.g., Pyne, 1979; Lipsitz et al., 1996; Hosmer et al., 1997; Archer et al., 2007; Blöchlinger and Leippold, 2011; Fagerland and Hosmer, 2016; Fernández and Liu, 2016). In recent years, there has been a line of research specifically devoted to addressing the following model assessment issues in GLMs for discrete data:

(A1) practical significance of model misspecification, if it exists;
(A2) what components of the working model are not specified correctly; and
(A3) what treatments may be needed to reduce the extent of model misspecification.

The methodological developments are partially driven by the increasingly easier access to

larger samples of data, in which situation traditional goodness-of-fit tests are likely to lead to rejection of the hypothesis of a perfect fit (Wasserstein and Lazar, 2016). To assess whether a logistic regression model is acceptable for practical purposes (A1), for example, Nattino et al. (2020) proposed a modification of the Hosmer-Lemeshow test that is independent of sample size. For the general class of discrete data, Yang (2021) proposed an alternative to the empirical distribution of Cox-Snell residuals, which allows the use of QQ plots to visualize the practical significance of model misspecification (A1). To better understand (A2) and (A3), diagnostic tools with comprehensive utility are much needed to reveal causes and inform actions with interpretable evidence. Li and Shepherd (2012) proposed the probability-scale residual with the zero-mean property under a correct model, which allows inspection of the mean structure using residual-versus-covariate plots (see also Li and Shepherd, 2010; Shepherd et al., 2016; Liu et al., 2018). To fully examine the distributional assumptions, Liu and Zhang (2018) proposed the surrogate residual that can generate residual-versus-covariate plots as well as Quantile-Quantile plots (QQ-plots) for categorical data (see also Liu et al., 2021a, 2023; Cheng, Wang, and Zhang, 2021; Li et al., 2024).

To achieve (A1)-(A3) altogether, we develop a new notion of residual that applies to GLMs and beyond. Unlike the literature that always defines a single-valued quantity as the residual, we propose to *use a function as a vehicle to retain the residual information*. This idea is similar to that used in confidence distribution inference where a function-based inference is shown to be more effective than a single-statistic-based inference (see the review by Xie and Singh, 2013 and various methods in Liu, Liu, and Xie, 2014, 2015, 2021b; Shen, Liu, and Xie, 2020; Luo et al., 2021). In the presence of data discreteness, we show that a functional residual is an appropriate tool for summarizing the residual randomness that cannot be captured by the structural part of the model. We establish theoretical properties of the functional residual for general discrete data regression models. These properties lead to the innovation of new diagnostic tools, such as the functional-residual-*vs*-covariate plot and Function-to-Function (*Fn-Fn*) plot. The graphical results can be interpreted in a

similar way to those from traditional residual-*vs*-covariate plots and QQ plots for assessing linear regression models. Therefore, our visualization tools enable us to evaluate practical significance of model misspecification, in a similar way to what we do for linear models.

Furthermore, our new diagnostic tools can reveal causes of model misspecification and suggest remedies for a better modeling fit. In our extensive simulation studies, we identify a variety of model misspecifications, such as not properly including a higher-order term, an explanatory variable, an interaction effect, a dispersion parameter, or a zero-inflation component. The effectiveness of our methodology is also demonstrated in the two case studies of wine marketing ($n = 4898$) and bike sharing ($n = 8734$). In each case study, we show that our tools generate not only statistical insights that inform model refinement but also business insights that guide the entire model building and delivering process.

Our method is built upon the surrogate analysis framework for discrete data (Liu and Zhang, 2018; Liu et al., 2021a, 2023; Cheng et al., 2021; Li et al., 2024). The surrogate idea is to simulate a continuous variable and use it as a surrogate of the original discrete variable to carry out statistical inference. The reliance on simulation adds a layer of randomness on top of the randomness in the observed data. The additional randomness is passed to graphical outputs, which may hinder robust inference when the sample size is not large. Our introduction of the functional residual eliminates the need of simulation and produce deterministic graphical outputs given the observed data. This is accomplished by expanding Liu-Zhang's surrogate residuals from the space of point statistics to the space of cumulative distribution functions (CDFs), aligning with the idea adopted by confidence distribution methodologies (Xie and Singh, 2013). Second, compared to Liu and Zhang (2018) that relies on the latent variable structure of ordinal regression models, our diagnostic framework is broad enough to cover all GLMs and even beyond. The setup and development in this paper lend themselves to semiparametric models. In one of our case studies, we show how a parametric model evolves to a semiparametric one under the guidance of our diagnostic

procedure. Last, our theoretical and graphical results can be interpreted in the same way irrespective to what type of data or model are dealt with. This coherence of interpretability is a key strength that distinguishes our method from the vast majority of model assessment methods that are either data-specific or model-specific.

2 Methodology

2.1 A general model setting for discrete data

Let Y be a discrete outcome and Ξ the set of possible values. Examples include

- a binary variable Y with $\Xi = \{0, 1\}$;
- an ordinal variable Y with $\Xi = \{0, 1, \ldots, J\}$; and
- a count variable Y with $\Xi = \{0, 1, 2, \ldots\}$.

Given a set of explanatory or predictive variables $\boldsymbol{X} = \{X_1, \ldots, X_p\}$, researchers and practitioners may use their domain knowledge and conventions to initiate a statistical model to quantitatively describe the influence of $\boldsymbol{X}$ on the probabilistic distribution of Y. In its most general form, such a *working model* can be written as

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \pi(y; \boldsymbol{x}, \boldsymbol{\beta}). \tag{1}$$

Here, $\pi(y; \boldsymbol{x}, \boldsymbol{\beta}) = \Pr\{Y \leq y \mid \boldsymbol{x}, \boldsymbol{\beta}\}$ is a discrete distribution function defined on the sample space Ξ, and $\boldsymbol{\beta}$ is a vector that includes all the unknown parameters (e.g., regression slope parameters). Model (1) encompasses all GLMs for discrete variables, such as commonly used logistic/probit models for binary data, proportional odds or ordered probit models for ordinal data, and Poisson or negative binomial models for count data. Less frequently used adjacent-category logit or stereotype logit models are also examples of Model (1). It is worth stressing that Model (1) is broader than the class of typical GLMs. Its general form lends itself to examining semiparametric models, such as generalized additive models.

Our goal is to assess whether or not the working Model (1) is consistent with the observed data set $\{(y_i, \boldsymbol{x}_i) : i = 1, 2, \ldots, n\}$. Assuming that data are generated from the true distribution $\pi_0(y; \boldsymbol{X})$, we are seeking statistical evidence against the assumption $\pi(\cdot; \boldsymbol{X}) \equiv \pi_0(\cdot; \boldsymbol{X})$.

2.2 Notion of functional residuals

We introduce a new notion of residuals for Model (1) and apply it to GLMs and other discrete data regression. Unlike the classical methods (Pearson or deviance residuals included) that use a *single-point* statistic to define a residual, we propose to use a *functional form* to represent residual information, which we term as a *functional residual*.

Definition 1 (Functional Residual)**.** *For a discrete outcome Y believed to follow Model* (1) *with a set of explanatory variables $\boldsymbol{X}$, a functional residual for an observation $(y, \boldsymbol{x})$ is a mapping from the sample space Ω to the function space $\mathcal{F} = \{F(t) : 0 \leq t \leq 1; 0 \leq F(t) \leq 1;$ and $F(t_1) \leq F(t_2)$ for any $t_1 < t_2\}$. Specifically,*

$$(y, \boldsymbol{x}) \to Res(t; y, \boldsymbol{x}) = F_{U(\pi(y-1;\boldsymbol{x}),\pi(y;\boldsymbol{x}))}(t) = \Pr\{U(\pi(y-1;\boldsymbol{x}), \pi(y;\boldsymbol{x})) \leq t\}, \qquad (2)$$

where $U(a, b)$ denotes a random variable uniformly distributed on the interval (a, b) .

Basically, given an observation $(y, \boldsymbol{x})$, we propose to use (2) as a residual, which is the CDF of a random variable U on the interval (0,1) with support on the section $(\pi(y-1;\boldsymbol{x}), \pi(y;\boldsymbol{x}))$. Heuristically, a GLM or its semiparametric extension, regardless of its specific form, yields probabilities $\{\pi(0;\boldsymbol{x}), \pi(1;\boldsymbol{x}), \ldots\}$, which forms a partition of the probability interval $(0, 1)$. The partition consists of sections $\{(0, \pi(0;\boldsymbol{x})), (\pi(0;\boldsymbol{x}), \pi(1;\boldsymbol{x})), (\pi(1;\boldsymbol{x}), \pi(2;\boldsymbol{x})), \ldots\}$. Given an observed outcome y, we will show in the next subsection that the residual uncertainty can be expressed by the CDF of a random variable U with support on the section $(\pi(y-1;\boldsymbol{x}), \pi(y;\boldsymbol{x}))$, which is the function defined in (2). In Section 6, we show the functional residual in (2) is in fact the CDF of the general surrogate residual (see Liu and Zhang, 2018; Liu et al., 2021a; Li et al., 2024). Therefore, it is an extension of

the surrogate residual from the space of point statistics to the space of CDFs. We defer a detailed explanation of the underlying intuition for this notion to the next subsection. In what follows, we use a simple example to illustrate this notion.

Consider a binary outcome which follows a logistic regression model

$$\text{logit}(\Pr\{Y = 1\}) = -1 + 2X. \tag{3}$$

Given an observation (y, x) from this model, its functional residual is

$$Res(t; y, x) = \begin{cases} F_{U(0,\pi(0;x))}(t) = \Pr\{U(0, \pi(0; x)) \leq t\} & \text{if } y = 0, \\ F_{U(\pi(0;x),1)}(t) = \Pr\{U(\pi(0; x), 1) \leq t\} & \text{if } y = 1. \end{cases}$$

Apparently, a functional residual is a mapping from the sample space Ω to a function space consisting of all CDFs. More specifically, the functional residual for $(y = 1, x = -1)$ is

$$Res(t; y = 1, x = -1) = \Pr\{U(0.95, 1) \leq t\} = \begin{cases} 0 & \text{if } 0 < t \leq 0.95, \\ \frac{t-0.95}{1-0.95} & \text{if } 0.95 < t \leq 1. \end{cases}$$

The functional residual for $(y = 0, x = 1)$ is

$$Res(t; y = 0, x = 1) = \Pr\{U(0, 0.27) \leq t\} = \begin{cases} \frac{t}{0.27} & \text{if } 0 < t \leq 0.27, \\ 1 & \text{if } 0.27 < t \leq 1. \end{cases}$$

The two functional residuals are CDFs with support on $(\pi(0; \boldsymbol{x}), 1)$ and $(0, \pi(0; \boldsymbol{x}))$, respectively, as illustrated in Figure 1(a)-(b). Their density forms can be visualized via a heat map in Figure 2(a), where the brighter colors indicate higher densities (with shorter support). For comparison purposes, we plot Pearson residuals in Figure 2(b). It reveals that the use of the whole function is distinct from the traditional use of a single point as the residual statistic. Intuitively, a function is loaded with more information than a single-valued statistic. In the next subsection, we will explain the intuition behind the new residual notion.

2.3 Intuition behind the notion

Our intuition originates from two fundamental observations.

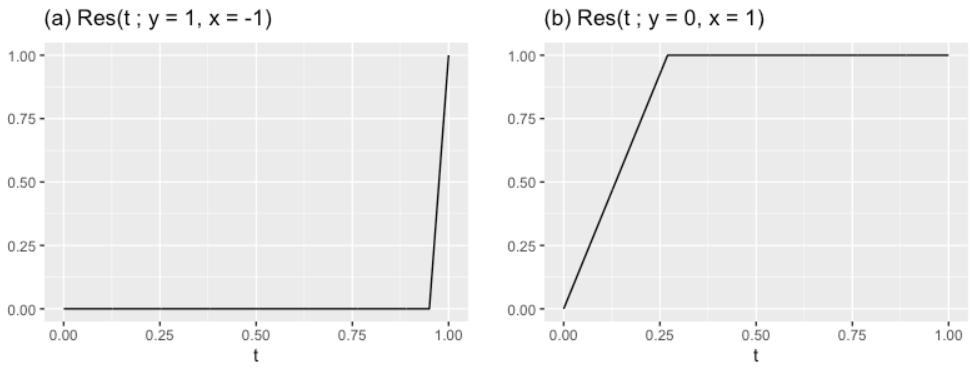

Figure 1: An illustration of the functional residuals for the logistic regression model (3).

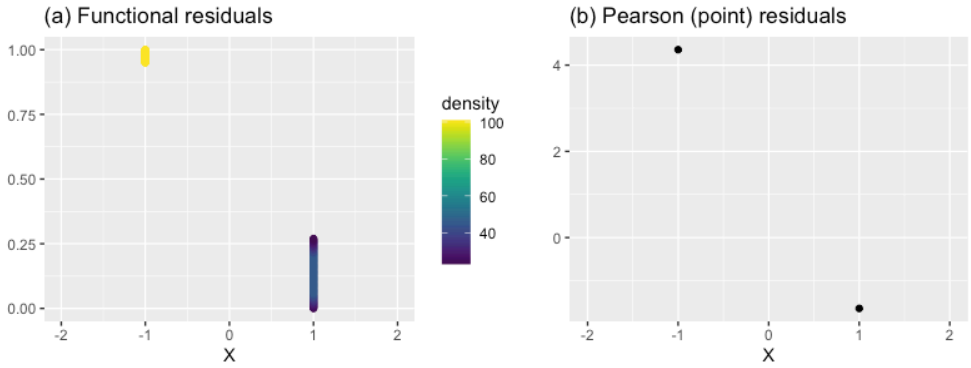

Figure 2: A comparison of functional residuals and Pearson (point) residuals when the observation is $(y = 0, x = 1)$ or $(y = 1, x = -1)$.

(Obs-1) *A residual is a quantity describing the residual randomness that cannot be captured by the structural part of the model.*

This principle is followed in defining a residual for the linear regression model $Y = \alpha + \beta X + \epsilon$. The structural part of the model is $(\alpha + \beta x)$, meaning that other than this conditional mean $(\alpha + \beta x)$ (and conditional variance σ^2) we know nothing about where Y would be. The statistic $\{y - (\alpha + \beta x)\}$, the difference between the observed y and $(\alpha + \beta x)$, reflects the residual randomness that cannot be captured by the structural part $(\alpha + \beta x)$. It is thus defined as a residual. This rationale is illustrated in the left panel of Figure 3.

For discrete data, (Obs-1) serves as the basis for our notion of functional residuals as well. To see this, we need to recognize the structural part of the general Model (1).

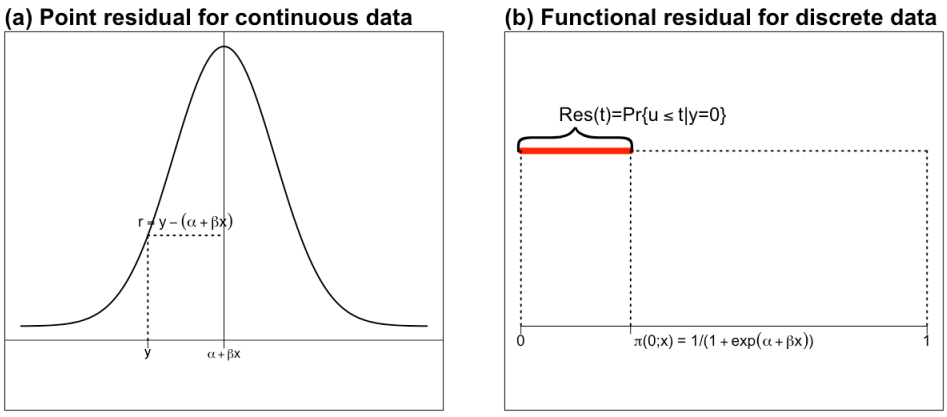

Figure 3: An illustration of our idea of constructing a functional residual for a logistic regression model (right panel), in comparison to the traditional method of deriving a point residual for a linear regression model (left panel). The entire thickened red line segment represents the functional residual for the observation $(y = 0, x = 1)$.

(Obs-2) *Given* $\boldsymbol{X} = \boldsymbol{x}$*, the distribution* $\pi(\cdot; \boldsymbol{x})$ *specified by Model* (1) *gives nothing but a partition* Π *of the probability interval [0,1]. The structural part of this model is no more than the totality of the cutpoints* $\{\pi(0; \boldsymbol{x}), \pi(1; \boldsymbol{x}), \ldots\}$*, which determines a partition* $\Pi = \{(0, \pi(0; \boldsymbol{x})), (\pi(0; \boldsymbol{x}), \pi(1; \boldsymbol{x})), (\pi(1; \boldsymbol{x}), \pi(2; \boldsymbol{x})), \ldots\}$.

To elaborate on (Obs-2), we use the logistic regression model (3). Given $X = x = 1$, for instance, this model essentially gives a probability $\pi(0; x) = \Pr\{Y = 0 \mid x\} = 0.27$. As illustrated in Figure 3(b), this probability serves as a cutpoint that yields a partition of the probability interval [0,1], i.e., $\Pi = \{(0, \pi(0; \boldsymbol{x})), (\pi(0; \boldsymbol{x}), 1)\}$. Figure 3 suggests that the partition Π plays a role similar to $(\alpha + \beta x)$ in the linear model in the sense that we know nothing but Π from the model. Therefore, we view the partition Π as the structural part of the logistic model. The next question is: given an observed outcome, say, $y = 0$, how to describe the residual randomness that *cannot* be captured by Π?

To answer this question, we image a needle is randomly dropped onto the probability interval $(0, 1)$. Let U be a random variable whose value records where the needle lands. Given an outcome $y = 0$, what we know is that a realization of U landed in the section $(0, \pi(0; x))$.

Yet what we do not know is where it is actually located. Following this line of thinking, the residual randomness can be summarized by a truncated distribution $U \mid 0 < U \leq \pi(0; x)$, which is a uniform distribution with support on $(0, \pi(0; x))$ (see Figure 3(b)). Therefore, we propose to use the entire CDF of this uniform distribution to retain the residual information, which leads to the notion of functional residual in (3). Similarly, if $y = 1$, we use the CDF of the uniform distribution with support on $(\pi(0; x), 1)$ to retain residual information.

The idea explained by Figure 3 extends to GLMs for ordinal and count outcomes. For an ordinal outcome, a GLM yields a partition with J sections, i.e.,

$$\Pi = \{(0, \pi(0; \boldsymbol{x})), (\pi(0; \boldsymbol{x}), \pi(1; \boldsymbol{x})), \ldots, (\pi(J-1; \boldsymbol{x}), 1)\}.$$

For a count outcome, a GLM yields a partition with infinitely many sections, i.e.,

$$\Pi = \{(0, \pi(0; \boldsymbol{x})), (\pi(0; \boldsymbol{x}), \pi(1; \boldsymbol{x})), (\pi(1; \boldsymbol{x}), \pi(2; \boldsymbol{x})), \ldots, \}.$$

When $Y = y$ is observed, the CDF of a uniform distribution with support on the section $(\pi(y-1; \boldsymbol{x}), \pi(y; \boldsymbol{x}))$ is used as the residual function. As the arguments do not require a specific form of the model, the same idea extends further to semiparametric models.

Remark 1. *The idea of using a function, rather than a point, to make inference is employed in the confidence distribution methodologies (Xie and Singh, 2013; Schweder and Hjort, 2016) but in a different context. A confidence distribution uses a CDF to describe the uncertainty in our knowledge about the parameter. Here, we use a CDF to describe the uncertainty in our knowledge about the discrete outcome that cannot be captured by a GLM.*

Remark 2. *For the ease of understanding, our discussion so far has used the residual random variable $R \equiv R_{\boldsymbol{\beta}}$ for linear models by assuming $\boldsymbol{\beta}$ is known. Similarly, we have used the true $\boldsymbol{\beta}$ in the functional residual $Res(t; y, \boldsymbol{x}) \equiv Res_{\boldsymbol{\beta}}(t; y, \boldsymbol{x})$. In practice, we estimate $\boldsymbol{\beta}$ and use the point residual $\hat{R} \equiv R_{\hat{\boldsymbol{\beta}}}$ or functional residual $\hat{Res}(t; y, \boldsymbol{x}) \equiv Res_{\hat{\boldsymbol{\beta}}}(t; y, \boldsymbol{x})$. If $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$, we typically have $\hat{R} \to R$ and $\hat{Res}(t; y, \boldsymbol{x}) \to Res(t; y, \boldsymbol{x})$. Analogous presentation logic is used in Li and Shepherd (2012) and Liu and Zhang (2018).*

The notion of functional residual is rooted in the surrogate idea proposed for discrete data analysis. The surrogate idea is to simulate a continuous variable and use it as a surrogate of the original discrete variable to make inference (Liu and Zhang, 2018; Liu et al., 2021a, 2023; Cheng et al., 2021). The simulation of surrogate variables induces another layer of randomness, which may make graphical results (e.g., residual-versus-covariate plots or QQ-plots) vary from one simulation to another. It remains unresolved how to quantify this variability and how to make robust graphical inference. The notion of functional residual provides a solution as it eliminates the simulation process. As a result, our graphical presentation of such residuals is deterministic given the observed data. This is achieved by using the entire CDF function to retain the residual information. Furthermore, the use of a function, instead of a point statistic, can preserve more information for inference. This point will be demonstrated in the establishment of theoretical results (e.g., Theorem 3: uniform convergence in QQ-like plots), which may be otherwise unavailable using existing point residual statistics.

3 Theoretical results

3.1 Properties under the null

We examine the conditional and unconditional expectation of the functional residual defined in (2) under the null. The theoretical results lead to the development of function-based diagnostic plots, which will be examined in simulated examples and real case studies.

Theorem 1 (Conditional Expectation under the Null)**.** *Given* $\boldsymbol{X} = \boldsymbol{x}$*, the conditional expectation of the functional residual* $Res(\cdot; Y, \boldsymbol{x})$ *in* (2) *is the CDF of a U(0,1) distribution, i.e.,*

$$E_Y Res(t; Y, \boldsymbol{x}) = F_{U(0,1)}(t) = t \text{ for any } t \in [0, 1],$$

provided that $\pi(\cdot; \boldsymbol{x}) \equiv \pi_0(\cdot; \boldsymbol{x})$.

On the basis of Theorem 1, we propose to use a functional-residual-*vs*-covariate plot to

examine the working model. Unlike the traditional residual-*vs*-covariate plot that displays point residual statistics, a functional-residual-*vs*-covariate plot exhibits the density function $\partial Res(t; y, x)/\partial t$ versus x. Such a new tool can be better visualized through the lens of a heat map, which is demonstrated in the following example.

Example 1 (Working model specified correctly)**.** We simulate 1000 ordinal data points y_i $(= 0, 1, 2, 3, 4)$ from an adjacent-category logit model

$$log\frac{\Pr\{Y = j\}}{\Pr\{Y = j + 1\}} = \alpha_j + \beta_1 X + \beta_2 X^2, \qquad j = 0, 1, 2, 3$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (1.5, 1.5, -1, 1)$, $(\beta_1, \beta_2) = (1.5, -1)$ and the covariate $X \sim \mathcal{N}(0, 1)$. Suppose the working model is specified correctly. We obtain functional residuals as defined in (2), and cast its density form $\partial Res(t; y_i, x_i)/\partial t$ against x_i. This functional-residual-*vs*-covariate plot is rendered in a heat map as shown in Figure 4(a). It indicates that the "heat" is evenly distributed on the two sides of the center horizontal line. This observation confirms the result in Theorem 1; that is, the null distribution of the "heat" is uniformly distributed on the (0,1) interval. The functional-residual-*vs*-covariate plot can also be rendered on the scale of a normal distribution as seen in the Figure 4(b), which exhibits a similar pattern of a symmetric distribution of the "heat". In contrast, when the classical deviance and Pearson residuals are used, the traditional residual-*vs*-covariate plots exhibit *asymmetric* distributions around the center horizontal line (see Figure 4(c) and (d)). This observation once again confirms a phenomenon discussed in Liu and Zhang (2018) and Liu et al. (2021a); that is, *classical residuals may exhibit unusual patterns even when the model is specified correctly.* The same problem is seen in other point residuals, such as the probability-scale residual (Li and Shepherd, 2012; Shepherd et al., 2016) in Figure 4(e) and the generalized residual (Franses and Paap, 2001) in Figure 4(f). The root cause is that the conditional distribution (on X) of these residuals may be heterogeneous, varying from x_i to x_j.

In what follows, we examine the unconditional expectation of the functional residual $Res(\cdot; Y, \boldsymbol{X})$. The result can be used to develop a new diagnostic tool that plays a similar

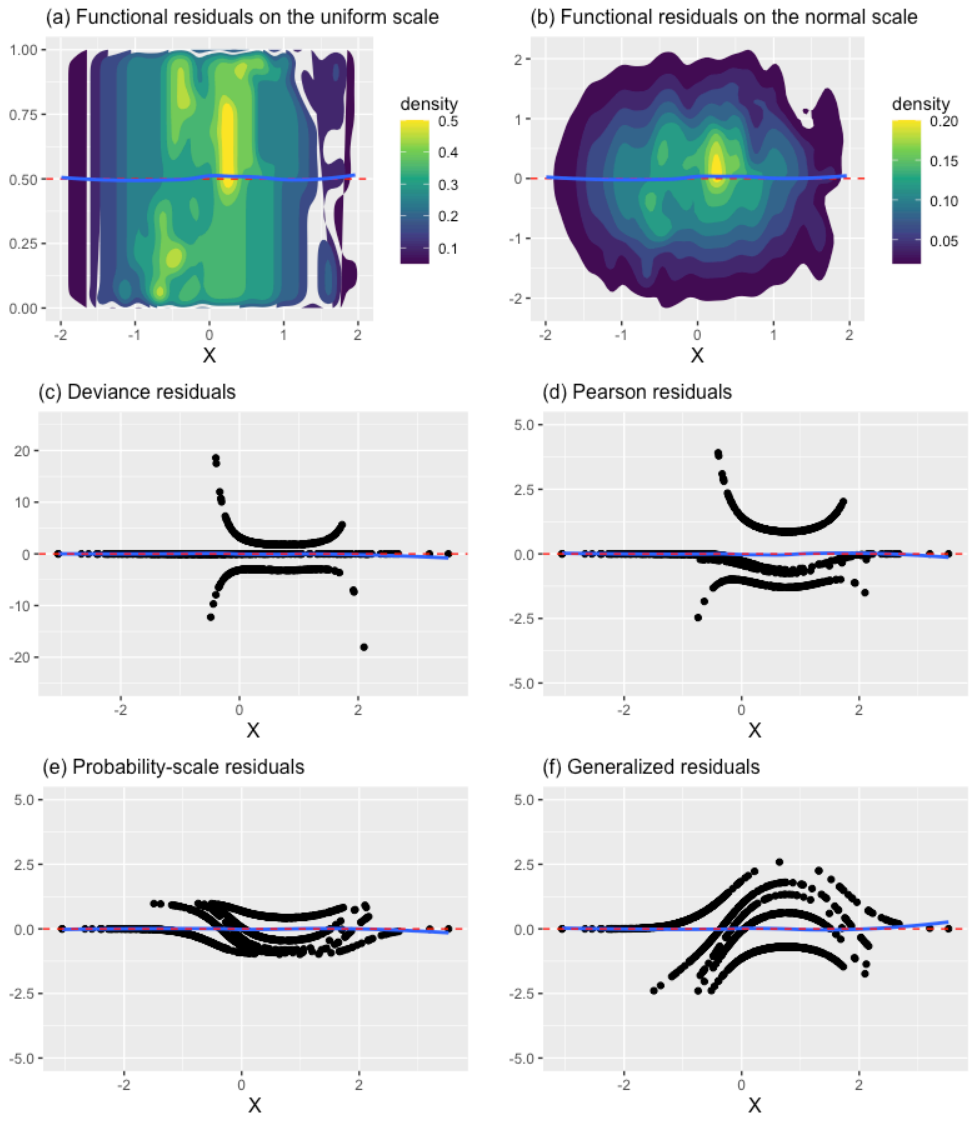

Figure 4: Proposed functional-residual-*vs*-covariate plots ((a)-(b)) and existing residual-*vs*-covariate plots ((c)-(f)) when the working model is specified correctly for ordinal data in the setting of Example 1.

role of QQ plots in traditional regression analysis.

Theorem 2 (Unconditional Expectation under the Null)**.** *The unconditional expectation of the functional residual* $Res(\cdot; Y, \boldsymbol{X})$ *in* (2) *is the CDF of a U(0,1) distribution, i.e.,*

$$E_{(Y,\boldsymbol{X})} Res(t; Y, \boldsymbol{X}) = F_{U(0,1)}(t) = t \text{ for any } t \in [0, 1],$$

provided that $\pi(\cdot; \boldsymbol{X} = \boldsymbol{x}) \equiv \pi_0(\cdot; \boldsymbol{X} = \boldsymbol{x})$ *for any* $\boldsymbol{x}$*.*

Following the law of large numbers, we have the following corollary immediately.

Corollary 1. *Suppose* $(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots$ *is an infinite sequence of i.i.d. random variables. Then, for any* $t \in (0, 1)$*,*

$$\overline{Res}(t) = \frac{1}{n} \sum_{i=1}^{n} Res(t; Y_i, \boldsymbol{X}_i) \to F_{U(0,1)}(t) = t \quad \textit{almost surely,} \tag{4}$$

provided that $\pi(\cdot; \boldsymbol{X} = \boldsymbol{x}) \equiv \pi_0(\cdot; \boldsymbol{X} = \boldsymbol{x})$ *for any* $\boldsymbol{x}$*.*

On the basis of Corollary 1, we propose to draw the function $\overline{Res}(t)$ against its null function $F_{U(0,1)}(t) = t$, which results in a *Function-Function (Fn-Fn) plot*. If the working model is specified (approximately) correctly, the *Fn-Fn* plot should yield a curve aligning (approximately) with the 45-degree straight line. With this regard, the *Fn-Fn* plot can be used in analogy to the classical QQ plot for linear regression models.

Example 1 (Continued)**.** We draw *Fn-Fn* plots for the same adjacent-category logit model considered previously. When the model is specified correctly, Figure 5(a) shows the *Fn-Fn* curve aligns almost perfectly with the 45-degree line. This observation confirms the result in Corollary 1. As a futher investigation, we restrict the sample to a subset $\{(y_i, x_i) \mid x_i < 0\}$ and reproduce the *Fn-Fn* plot in Figure 5(b). The *Fn-Fn* curve still aligns very well with the 45-degree line, which suggests that our method allows the probe of specific subgroups.

Theorem 3 (Uniform convergence)**.** *Suppose* $(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots$ *is an infinite sequence of i.i.d. random variables. The average of the functional residuals* $Res(t; Y_i, \boldsymbol{X}_i)$ *in* (2)

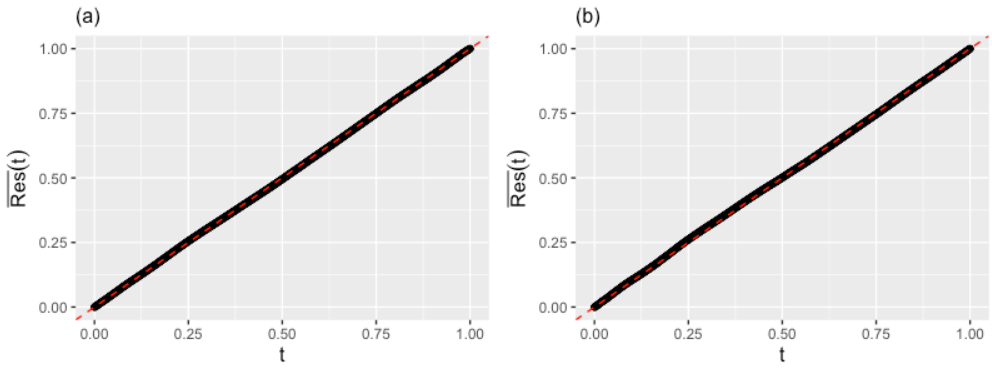

Figure 5: Proposed *Fn-Fn* plots when the working model is specified correctly for ordinal data in the setting of Example 1. The left and right panels are obtained using the full sample (1000 observations) and the subsample $\{(y_i, x_i) \mid x_i < 0\}$, respectively.

converges to the function $F_{U(0,1)}(t) = t$ *uniformly; that is, for any* $\epsilon > 0$,

$$\inf_{t\in[0,1]} \Pr\left\{|\overline{Res}(t; Y_n, \boldsymbol{X}_n) - t| < \epsilon\right\} \to 1, \quad \text{as } n \to \infty.$$

As compared to Theorem 2 and Corollary 1, which only justify the point-to-point convergence of $\overline{Res}(t; Y_n, \boldsymbol{X}_n)$ to t, Theorem 3 further shows the convergence is uniform for the entire set $t \in [0, 1]$. The implication is when the model is right, the entire curve $(t, \overline{Res}(t))$ will move toward the 45-degree line at the same rate as the sample size increases. In other words, if the curve $(t, \overline{Res}(t))$ deviates partly from the 45-degree line, it is not a result of different convergence rates over $t \in [0, 1]$ but an indication of model misspecification.

Remark 3. *Due to the heterogeneous conditional distributions given* x *as observed in Figure 4(c)-(f), Pearson, deviance, probability-scale, and generalized residuals do not have an explicit unconditional distribution. The lack of a clear reference distribution hinders the development of QQ-plots based on these residuals. On the other hand, the surrogate residual in Liu and Zhang (2018) relies on simulation given the observed data. To control the simulation (or sampling) error and graphical variability from one QQ-plot to another, Liu and Zhang (2018) proposed to use multiple sampling, which however increases technical sophistication and convolute practical interpretation. The use of functional residuals eliminate the need of simulation and produce non-random (give the observed data) QQ-like plots. In Section 6, we reveal the connection between the surrogate residual and functional residual.*

Remark 4. *In Supplementary Materials A.2, we show that* $\hat{Res}_{\hat{\boldsymbol{\beta}}}(t; y, \boldsymbol{x}) = Res_{\boldsymbol{\beta}}(t; y, \boldsymbol{x}) + o_p(1)$ *as long as* $\hat{\boldsymbol{\beta}}$ *is a consistent estimate such that* $\hat{\pi}(y; \boldsymbol{x}, \hat{\boldsymbol{\beta}}) = \pi(y; \boldsymbol{x}, \boldsymbol{\beta}) + o_p(1)$. *We also establish Theorems* 1′-3′ *that are similar to Theorems 1-3 except a vanishing term.*

3.2 Distinguishing power against alternative models

A theoretical "power study" is a challenging task in the context of model diagnostics. It is absent in the traditional GLM diagnostics studies (Pregibon, 1981; Landwehr et al., 1984; Pierce and Schafer, 1986; Davison et al., 1989; Ben and Yohai, 2004) as well as in the recent major contributions to the field (Li and Shepherd, 2012; Liu and Zhang, 2018; Yang, 2021). The challenges stem from the following: (i) diagnostic procedures often do not pre-specify alternative models; (ii) the model space is vast, and both the null and alternative models may not be characterized by a common parameter θ within Euclidean space; and (iii) there is no well-accepted method for defining the "distance" between different models (e.g., between the null model $\mathcal{M}_0$ and an alternative model $\mathcal{M}_a$).

Based on the result that the F_n-F_n curve of the null model $\mathcal{M}_0$ converges to the 45-degree line (Theorem 3), we propose to use the distance between the F_n-F_n curve for $\mathcal{M}_a$ and the 45-degree line as a measure of the distance between $\mathcal{M}_a$ and $\mathcal{M}_0$. Under this consideration, the 45-degree line plays a role similar to the null $H_0 : \theta = \theta_0$ in hypothesis testing. A power study for hypothesis testing often entails a power examination when the alternative $\theta_a \to \theta_0$ at a certain rate as the sample size n increases (DasGupta, 2008). When the local alternative θ_a is close to the null, the power shows the test's ability to detect a small departure. It is of particular interest in situations where the effect size is small and only becomes noticeable with a large sample size. We adapt the idea and modify it to fit our context of model diagnostics. For any model $\mathcal{M}$, we define

$$D_n^{(\mathcal{M})}(t) = \left| \overline{Res}_n^{(\mathcal{M})}(t) - t \right|$$

as the distance function between the F_n-F_n curve for $\mathcal{M}$ and the 45-degree line. Let

$$D_{n,max}^{(\mathcal{M})} = \max_{t\in[0,1]} D_n^{(\mathcal{M})}(t)$$

denote the maximum deviation of the F_n-F_n curve. Then, we have the result as below.

Theorem 4. *Assume that for the alternative model $\mathcal{M}_a$, its maximum deviation from the 45-degree line diminishes at the rate $O_p(n^{-(1/2-c)})$, i.e., $n^{1/2-c}D_{n,max}^{(\mathcal{M})} \to D$ in distribution, where D is a non-degenerate random variable and c is an arbitrary positive constant. Then,*

$$\Pr\left\{D_{n,max}^{(\mathcal{M}_a)} > n^{-1/2+c/2} \ \&\ D_{n,max}^{(\mathcal{M}_0)} < n^{-1/2+c/2}\right\} \to 1 \quad \text{as } n \to \infty.$$

Basically, the result says that our approach can distinguish the alternative model $\mathcal{M}_a$ from the null model $\mathcal{M}_0$ in the following sense:

> *Under the assumption in Theorem 4, with probability one, there exists a "graphical envelop" (of size $n^{-1/2+c/2}$) centered on the 45-degree line, such that the F_n-F_n curve for the null model $\mathcal{M}_0$ lies entirely within the envelop, while the F_n-F_n curve for the alternative model $\mathcal{M}_a$ makes excursions outside the envelop.*

As the constant c can be arbitrarily small, the result here has a practical implication; that is, if the model deviation, measured by its the distance between the F_n-F_n curve for $\mathcal{M}_a$ and the 45-degree line, is greater than the order of $n^{-1/2}$, our diagnostic tool can capture its excursion and distinguish it from the plot for the null model. The result describes the sensitivity of our tool, as a general diagnostic approach for any model in the form of (1), to the alternative models. In Supplementary Materials G, we show how to produce a graphical envelop and study its distinguishing power numerically.

We stress that the applicability of Theorems 1-4 is not dependent on the type of outcome data or the form of working models. They apply to general Model (1) for any discrete outcome data, as opposed to Liu and Zhang (2018) that relies on the latent variable structure of parametric models. In our numerical investigation, we examine cases where latent variables may not exist, such as the adjacent-category logit model for ordinal data, Poisson/negative binomial model for count data, and even semiparametric models.

4 Simulation studies of diagnostic utility

We use simulated examples to demonstrate the utility of our diagnostic tools in identifying a variety of model misspecifications. The examples will manifest two features of our framework:

- Unlike hypothesis testing that yields a "yes/no" answer, our approach reveals the *extent* to which the working model deviates from the true model, and furthermore, it gives *actionable clues* as to what model component is misspecified as well as what treatment should be applied to improve the model.
- Unlike other model assessment procedures that are data- or model-specific, our approach provides a unifying treatment in the sense that diagnostic results can be interpreted in a similar fashion *regardless of the type of response data (binary, ordinal, or count) and the choice of regression models*.

4.1 Parametric models for ordinal data

Different from Liu and Zhang (2018) which focused on the cumulative link model, our approach is not constrained to latent-variable-induced models. Instead, it applies to any model for ordinal data. In this subsection, we use the adjacent-category logit model to study an ordinal outcome ($Y = 0, 1, 2, 3, 4$) consistently throughout the examples.

Example 2 (Missing of a higher order term)**.** The ordinal data are generated from the same adjacent-category logit model as in Example 1, where the X^2 term is a crucial component in the underlying model. Initial analysis, however, often begins with a simple adjacent-category logit model that only contains a linear term of X, such as $log(p_j/p_{j+1}) = \alpha_j + \beta X$, where $p_j = \Pr\{Y = j\}$. For this working model, we obtain its functional-residual-*vs*-covariate plots, which are displayed in Figure S1(a) in Supplementary Materials B. The "heat" in the upper left panel is unevenly distributed around the center line, being quite different from that in Figure 4(a) where the model is specified correctly. Moreover, the LOWESS curve clearly

indicates the missing of the quadratic term in the working model. Visualized in Figure S1(b) are functional residuals *in the standard normal scale* (see Remark 5), which exhibits a similar pattern. For comparison purposes, classical residual-*vs*-covariate plots are included in the lower row of Figure S1. The deviance residuals in Figure S1(c) barely yield any evidence against the use of the working model without the quadratic term. In fact, the comparison of the plot here with that in Figure 4(c) manifests a "fatal defect" of deviance residuals, as we can hardly tell which residual-*vs*-covariate plot is normal and which is abnormal. In Figure S1(d), Pearson residuals exhibit a declining trend in the region where $X > 0$. But it may be confusing if we compare this plot with Figure 4(d) as "U"-shaped curves are also seen in the latter.

As a further study, we simulate 1000 samples from a adjacent-category logit model with a cubic term $log(p_j/p_{j+1}) = \alpha_j + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$, where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-1, 1.5, 2, 3)$, $(\beta_1, \beta_2, \beta_3) = (2, -1, -1.5)$, and $X \sim \mathcal{N}(0, 1)$. Assuming that the quadratic term is included in the working model while the cubic term is not, we cast its functional residuals to the normal scale in Figure S2. The cubic LOWESS curve there informs that a quadratic polynomial is not adequate for explaining $log(p_j/p_{j+1})$ and a high-order term should be added.

Remark 5. *Although the functional-residual-vs-covariate plots on the uniform and normal scales are loaded with the same amount of information, we recommend to use the normal scale for better visualization purposes. We may hardly distinguish points, such as 0.99 and 0.999, on the edges of the (0,1) interval, but the difference is apparent on the normal scale as* $\Phi^{-1}(0.99) = 2.33$ *and* $\Phi^{-1}(0.999) = 3.09$. *We will use the normal scale to present functional residuals in the rest of our paper.*

Example 3 (Missing of relevant covariates)**.** We simulate 1000 samples from the adjacent-category logit model $log(p_j/p_{j+1}) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. Here, $(\beta_1, \beta_2, \beta_3) = (1.5, 1, 0)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-1, -2, 0.5, 2)$, $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(-1, 0.8)$, and $X_3 \sim \mathcal{N}(0.5, 1)$. In this setting, both X_1 and X_2 are correlated with the ordinal response Y, whereas X_3 is irrelevant as $\beta_3 = 0$. We fit an adjacent-category logit model with only X_1. Figure S3 casts

functional residuals against X_2 and X_3 that are not included in the working model. The strong correlation seen in Figure S3(a) indicates that the missing covariate X_2 may help explain a significant proportion of the residual variability, and thus it should be added to the model. To the contrary, the symmetrically distributed "heat" in Figure S3(b) does not exhibit any correlation between our functional residual and X_3, which is consistent with the fact that X_3 is irrelevant ($\beta_3 = 0$). Examining plots in Figure S3(c)-(f) in the same way, we find out that neither deviance nor Pearson residuals reveal the existence or lack of correlation as the LOWESS curves in both plots, for X_2 and X_3, are flat.

Example 4 (Missing of interaction)**.** We simulate 1000 samples from the adjacent-category logit model $log(p_j/p_{j+1}) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, which contains an interaction term between the two covariates. Here, $(\beta_1, \beta_2, \beta_3) = (1, 2, 2)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-1, -2, 0.5, 2)$, $X_1 \sim \mathcal{N}(0, 1)$, and $X_2 \sim \mathcal{N}(-1, 0.8)$. We fit an adjacent-category logit model without the interaction. The functional residuals are plotted against $X_1 X_2$ in Figure S4(a). The decreasing trend is clear evidence that the residual variability can be partially explained by the missing interaction term. After adding $X_1 X_2$ to the working model, the decreasing trend disappears in the updated functional-residual-*vs*-covariate plot (Figure S4(b)). On the other hand, the LOWESS curves in both the deviance residual plot (c) and the Pearson residual plot (e) indicate the interaction effect exists only when $X_1 X_2 > 3$, which nevertheless is not true. According to the simulation setting, the interaction effect exists regardless of the value of $X_1 X_2$. Our functional residual plot in (a) reveals this interaction effect in the region where the majority of data points reside (i.e., $-2 < X_1 X_2 < 2$).

4.2 Parametric models for count data

We continue to study the diagnostic utility of our functional residuals for integer-valued count data. Considered in the numerical examples are a variety of Poisson regression models.

Example 5 (Correct specification of the model)**.** We generate 1000 samples from the Poisson

regression model with a quadratic term $log(E(Y|X)) = \alpha + \beta_1 X + \beta_2 X^2$, where $(\alpha, \beta_1, \beta_2) = (1, 0.2, 0.15)$ and the covariate $X \sim \mathcal{N}(0, 1)$. When the model is specified correctly, we obtain our functional residuals as well as traditional deviance and Pearson residuals. Throwing these residuals against the covariate X yields diagnostic plots in Figure S5. The upper row of Figure S5 shows that our functional residuals are evenly distributed around the dashed horizontal line, regardless of the scales (being uniform or normal) used for visualization. Our F_n-F_n plot in Figure S6 reveals nothing worth attention either. This no-show of unusual patters is what we anticipate as the model is specified correctly. In contrast, the deviance residuals in Figure S5(c) (and Pearson residuals in Figure S5(d)) cluster together and form multiple parabolas opening downwards. This pattern may mislead analysts to believe that certain high-order terms of X need to be added to the model. This observation once again reinforces our finding in Example 1; that is, the discreteness of data can cause deceptive clusters and shapes if traditional point residuals are used to produce diagnostic plots.

In Supplementary Materials D, we present results when a higher order term, a relevant covariate, or an interaction is missing in the working Poisson model. These are the types of misspecification studied previously for ordinal data. Our finding is that our functional residual yields diagnostic plots that can be interpreted in the same fashion, although the data type and parametric model are different. In what follows, we focus on two misspecification cases that are common specifically in modeling count data.

Example 6 (Missing of a zero-inflated component)**.** In the case of an excess of zeros in count data, the regular Poisson model is inadequate and a new component for modeling zeros needs to be introduced. For example, the count of zeros follows a logistic model $logit\{P(Y = 0)\} = 1 + 0.2X$, while the non-zero counts follow a truncated Poisson model $log(E(Y|X)) = 1 + X$ (for $Y \geq 1$). We simulate 1000 samples from this so-called zero-inflated Poisson model with $X \sim \mathcal{N}(0, 0.8)$. We fit, however, a regular Poisson model to the simulated data. The functional-residual-*vs*-covariate plot is displayed in Figure S7(a). It is observed that the vast majority of the functional residuals falls below the horizontal

line at zero, and a long slice of the residual cloud approaches very close to -2. This is an indication that the proportions of smaller count values, such as zeros, are underestimated by the regular Poisson model. This observation is confirmed in Figure S7(c), where the lower tail of the *Fn-Fn* curve goes far above the 45-degree line. We therefore update the model with an addition of a zero-inflated component and reproduce the functional-residual-*vs*-covariate plot in Figure S7(b). The residual cloud becomes symmetrically distributed around zero and the LOWESS curve is almost flat. Figure S7(d) confirms the appropriateness of the model by showing that the *Fn-Fn* curve aligns well with the 45-degree line.

Example 7 (Missing of modeling dispersion)**.** We simulate 1000 count data samples from a Poisson distribution with the dispersion parameter being $\frac{1+\phi}{\phi} = 7$ where $log(E(Y|X)) = 1.2 + 1.3X$ and $X \sim \mathcal{N}(0, 1)$. The simulated data are fitted using a regular Poisson model with the default setting that the dispersion parameter is set as 1. In this case, the functional-residual-*vs*-covariate plot in Figure S8(a) shows that the residual cloud is stretched toward both the lower and higher ends. This is in contrast with the well-rounded residual cloud as seen in Figure S8(b), where the over-dispersion is correctly modeled using a quasi-Poisson model. We also note that the over-dispersion in count data may result in inflated proportions of smaller count values (e.g., 0, 1, 2). This explains why in Figure S8(c) the *Fn-Fn* curve is highly skewed to the left. This skewness disappears in Figure S8(d) when the dispersion parameter is introduced to capture the additional variability.

4.3 Semiparametric models

A generalized linear model assumes $E(Y|\boldsymbol{X}) = g^{-1}(\eta) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$, where $\eta = \boldsymbol{X}\boldsymbol{\beta}$ is called the linear predictor. Basically, it assumes that the transformed conditional expectation $g(E(Y|\boldsymbol{X}))$ can be modeled by a linear combination of $\boldsymbol{X}$. However, nonlinear effects of covariates are not uncommon in real world applications, where it may be more appropriate to model $\eta = f(\boldsymbol{X})$ nonparametrically (Faraway, 2016). Here, $f(\cdot)$ is an unknown smooth

function that needs to be estimated. Diagnostics of such semiparametric models have been rarely discussed in the literature. As our model setup in (1) lends itself to semiparametric models as such, our development so far applies to diagnostics of semiparametric models. Below we use a simulated example to illustrate how to use functional residuals to diagnose semiparametric models. A real case study can be found in Section 5.2.

Example 8. The negative binomial distribution is commonly used to simulate count data with overdispersed Poisson distribution. We simulate 1000 count data samples of Y from a negative binomial distribution NB($\frac{\lambda}{\phi-1}, \frac{1}{\phi}$). The logarithm of Y's expectation is set as $log(\lambda) = log(E(Y|X)) = 1.2 + 1.3sin(X) - 0.8X$. The sine function in the mean structure is used to simulate the cyclical effects of the covariate $X \sim \mathcal{N}(0, 1)$. The variance of Y is $\lambda\phi$, where ϕ is the dispersion parameter set as $\phi = 7$. We fit a regular Poisson model $log(E(Y|X)) = \alpha + X\beta$ to the simulated data. The functional-residual-*vs*-covariate plot is presented in Figure S9(a), where the LOWESS curve displays a polynomial pattern. To incorporate this effect, we add a cubic smoothing spline term to the model, which results in a generalized additive model, a semiparametric model. Figure S9(b) shows that the nonparametric term helps pull the LOWESS curve toward the center horizontal line and reduce its fluctuation. However, the *Fn-Fn* plot in Figure S9(e) only shows a slight improvement over Figure S9(d). The gap between the *Fn-Fn* curve and the 45 degree line may indicate extra variability that has not been modeled. Note that the nonparametric term adjusts the mean structure of the model, and it cannot address the extra variability on top of the mean structure. Our next step is to add the dispersion parameter to the model and let data decide its value. Our *Fn-Fn* plot of this quasi-Poisson model in Figure S9(f) shows the gap is closed as a result of adding the dispersion parameter. At the same time, the functional-residual-*vs*-covariate plot in Figure S9(c) shows that the LOWESS curve is close to the center horizontal line except at the two ends. Therefore we may conclude that this generalized additive quasi-Poission model provide an adequate fit to the data.

5 Case studies

We use our methodology in parametric and semiparametric modeling for business studies. Presented in our first case study is modeling and diagnostics for wine quality to facilitate wine marketing and manufacturing, where the outcome variable expert rating is ordinal. The second case study also addresses a marketing problem, but the focus is on bike sharing and the outcome is the number of hourly rentals. Our analyses show that our diagnostic tools can (i) detect a variety of model misspecifications, which are otherwise concealed due to the discreteness of data; (ii) provide actionable clues that can guide model treatment; and (iii) be interpreted in the same way regardless of the type of outcomes or models.

5.1 Diagnostics for wine quality modeling

In order to better serve customers around the globe, the wine industry is investing and developing new technologies used in its manufacturing process. Among many key elements that can be analytically studied in this process are wine certification and quality assessment (Cortez et al., 2009). The wine certification is performed in labs by using physicochemical tests which measure *pH*, *alcohol*, *density*, and other characteristics of wine. However, tests of wine quality heavily rely on human experts, which is time-consuming and expensive. It is therefore desirable to establish the correlation between human preference and physicochemical properties. Understanding of this relationship can automate the assessment of wine quality and inform the entire wine manufacturing process. To this end, Cortez et al. (2009) collected data of white vinho verde wine from the north of Portugal (`http://archive.ics.uci.edu/ml/datasets/Wine+Quality`). This data set consists of 4898 observations and 12 variables. The outcome *quality* is measured on a rating scale from 0 to 10, with 10 indicating the best quality. The observed ratings, ranging from 3 to 9, are derived from the judgment of wine experts. To understand how human ratings are influenced by the physicochemical characteristics of wine, we fit an adjacent-categorical model using 8

predictors listed in Table S1 in Supplementary Materials C.

(*Detecting outliers*) As the existence of outliers may be a cause of graphic patterns that may be misinterpreted as model misspecification, we identify and remove them before proposing any model adjustment. From the diagnostic plots in the upper row of Figure S10, we observe the presence of outliers in variables (a) *fixed.acidity*, (b) *residual.sugar*, and (c) *density*. These outliers drag the LOWESS curves farther away from the center horizontal line. The detected outliers are confirmed in Figure S11(a)-(c). For example, in Figure S11(b), the highest value of *residual.sugar* is 65.8, which is quite distant from the second highest value 31.6. We emphasize, however, that being an outlier in its marginal distribution as seen in Figure S11 does not necessarily suggest dropping it from the model. What really matters is its impact on the model fitting, which can be assessed through the lens of our proposed functional-residual-*vs*-corvariate plots. We follow this argument and decide to remove 5 outliers labeled with the cross ($\times$) symbol. After the removal, the updated LOWESS curves in the lower row of Figure S10(a*)-(c*) are moving much closer to the center horizontal line. Nevertheless, we keep the outlier sitting on the far right of Figure S11(d). The reason is that it does not show to have a negative impact on the model fitting (see Figure S12(b)) as long as we make a further adjustment of the model, which will be discussed below.

(*Adding a quadratic term of free.sulfur.dioxide*) The parabola-shaped LOWESS curve in Figure S12(a) indicates the missing of a higher order term of *free.sulfur.dioxide* in the model. Figure S12(b) shows that adding a quadratic term of this variable results in the LOWESS curve being flattened. The *t*-test also confirms the added quadratic term is significant at $\alpha = 0.001$ level (see Table S1). Using the AIC criterion, we also see a notable improvement in this final model. With the added term, the AIC value decreases from 10998 to 10813, which may imply an elevated predictive power.

5.2 Diagnostics for bike sharing modeling

In recent years, bike-sharing business has become increasingly popular in populous cities as a convenient transportation alternative. This relatively new business has potential to decrease the dependence on automobile and reduce greenhouse gas emissions (Contardo et al., 2012). Around the globe, there are more than 500 bike-sharing programs which deploy over 500 thousand bicycles per year. Residents and travelers can easily rent a bike from and return it to designated stations using self service (Fanaee-T and Gama, 2014). However, existing data show that the hourly rental of bikes is highly variable, and it heavily depends on the time of a day and weather conditions. In order to reduce operational cost, it is imperative to understand what factors and how they drive the demand of bike rentals. To achieve this, we perform statistical modeling of the data from Capital Bike Sharing System at Washington D.C. (Fanaee-T and Gama, 2014). The data set contains 8734 observations of the hourly bike rentals in 2012. The explanatory variables are listed in Table S2. We start the modeling process by using a regular Poisson regression model with all of the explanatory variables. The model fitting result can be found in Table S3 under the title "Initial model".

(*Adding nonparametric terms*) To check the model fitting, we plot the functional residuals versus each of the covariates in Figure S13. The LOWESS curves in all the plots are falling far below the center horizontal line. This indicates that the group of small functional residuals is disproportionately large, which is also captured by heavily stretched tails of the *Fn-Fn* plot in Figure S14(a). In particular, the LOWESS curve in Figure S13(a) is neither linear nor monotonic, but instead it exhibits a pattern of strong fluctuations throughout the 24 hours in a day. This observation reflects that the factor *hour* may be a major determinant that drives the number of rentals in a cyclical way. This cyclical influence is not captured by the linear mean structure in the regular Poisson model. We therefore decide to model the factor *hour* nonparametrically using cubic smoothing splines. We apply the same change to the variables *temp, humidity,* and *windspeed*, considering that they all show some degrees

of fluctuations in Figure S13(b)-(d). The use of nonparametric terms leads to a generalized additive model. Figure S15 displays the updated diagnostic plots for *hour, temp, humidity,* and *windspeed*. With respect to the shape of LOWESS curves, we observe that the degrees of fluctuations have all been reduced. At the same time, the use of smoothing functions has also helped pull all the LOWESS curves toward the center horizontal line. In what follows, we continue to show how our diagnostic tools can lead us to eventually close the remaining gap between the LOWESS curves and the center horizontal line.

(*Adding a dispersion parameter*) Although improved, the *Fn-Fn* curve in Figure S14(b) still shows a large deviation from the 45-degree line. This is probably due to the presence of a large amount of small value data, which is also reflected in Figure S15 where all of the LOWESS curves are still below the center horizontal line. Note that a Poisson distribution forces its variance to be the same as the mean. However, the disproportionate amount of low bike demand (e.g., 1-5 AM) may have inflated the variance and made it much larger than the mean. Therefore, the variability of bike rentals may be underestimated in the generalized additive Poisson model. To this end, we add a dispersion parameter to the model and allow data to give an estimate instead of using the default value which is 1. We call this a generalized additive quasi-Poisson model. The model fitting result can be found in Table S3 under the title "Final model". The estimate of the dispersion parameter is 42.998. Its considerable size confirms that the variance adjustment, suggested by our diagnostic tools, is necessary. After this update, the *Fn-Fn* curve in Figure S14(c) turns out to be very close to the 45-degree line. This is an indication that the new model captures the inflated variability caused by very low bike demand in early AM time period. The improvement of the model is also reflected in Figure S16, where the LOWESS curves are now very close to the center horizontal line for the entire set of variables.

(*Additional insights into the final semiparametric model*) By adding nonparametric terms and a dispersion parameter, the final generalized additive quasi-Poisson model has shown

a remarkable improvement over the initial Poisson model. We conclude the discussion by offering an additional insight drawn from our diagnostic tools. We observe in Figure S16(a) that the conditional distributions of the functional residual are not homogeneous across the 24 hours. For example, the distributions (conditional on *hour*) in early AM hours and peak hours are very different. This heterogeneity in distribution leads to the "residual patterns" in Figure S16(a), and it is clearly not captured by the final semiparametric model. The reason is that all the adjustments we have made so far (i.e., adding smoothing functions and a dispersion parameter) only addressed the issues in the first and second moments (i.e., mean and variance). To completely eliminate the remaining patterns in Figure S16(a), we may consider options such as introducing new parameters to model the skewness and other higher-order moments or using a time series model to capture the remaining temporal patterns. In Supplementary Materials E, we justify the applicability of our framework to time series models and carry out analysis of the same bike rental dataset. We observe remaining temporal patters in functional residuals, which indicates that even time series models are not able to capture all temporal dependencies in the data. This result is not surprising, however. As Li et al. (2015) pointed out, a challenge of bike sharing business is the highly varying and skewed bike usage during 24 hours and across four seasons. This challenge becomes evident in our diagnostic analysis, indicating the need for increased focus from domain experts.

6 Unifying point residuals in the same framework

We examine the connection of our functional residual to point residuals developed in recent years. We show that both the surrogate residual (Liu and Zhang, 2018; Liu et al., 2021a) and the probability-scale residual (Li and Shepherd, 2010, 2012) are point statistics that can be drawn or derived from our functional residual. To see this, we define a *functional residual variable* as

$$R_{func} \mid (y, \boldsymbol{x}) \sim Res(t; y, \boldsymbol{x}). \tag{5}$$

Basically, given the data $(y, \boldsymbol{x})$, as the functional residual $Res(t; y, \boldsymbol{x})$ in (2) is a distribution function, we let R_{func} follow this distribution.

Liu and Zhang (2018) considered the class of cumulative link models

$$G^{-1}\left\{\Pr(Y \leq j)\right\} = \alpha_j - \boldsymbol{X}\boldsymbol{\beta}, \tag{6}$$

where $G^{-1}(\cdot)$ is the link function and the intercept parameters $-\infty = \alpha_0 < \alpha_1 \cdots < \alpha_J = \infty$. They use the latent structure of Model (6) to generate a surrogate outcome $S \sim Z \mid \alpha_{j-1} < Z \leq \alpha_j$ if $Y = j$, where $Z = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ ($\epsilon \sim G$) is a latent outcome. The residual is defined as the difference between the surrogate outcome S and its expectation, i.e., $R_{sur} = S - E(S \mid \boldsymbol{X})$. The following result shows that R_{sur} follows the same distribution of $G^{-1}(R_{func})$.

Theorem 5. *For the cumulative link model in* (6)*, our transformed functional residual variable* $G^{-1}(R_{func})$ *and the surrogate residual variable* R_{sur} *follow the same conditional distribution, i.e.,* $G^{-1}(R_{func}) \stackrel{d}{=} R_{sur}$*, given the data* $(y, \boldsymbol{x})$*.*

Theorem 5 states that given the data, $G^{-1}(R_{func})$ and R_{sur} have the same conditional distribution, which implies that they have the same utility in model diagnostics. As R_{func} is a variable drawn from $Res(t; y, \boldsymbol{x})$, our functional residual can be viewed as an extended notion of the surrogate residual. Basically, a functional residual in (2) extends a surrogate residual from the point space to the functional space. Such an extension has several merits.

First, the functional residual applies to cumulative link models as well as non-cumulative-link models, such as the adjacent-category logit model (Section 4.1 and 5.1) and Poisson model (Section 4.2 and 5.2). Furthermore, this notion applies to semiparametric models when nonparametric terms are included (Section 4.3 and 5.2). It therefore broadens the scope of applicable models and provides a unified treatment for binary, ordinal, and count data. Third, the functional residual does not rely on simulation, and thus it eliminates the randomness (given the data) as seen in generation of surrogate residuals. As the entire CDF

function is used to preserve the residual information, a functional residual may be viewed as the result of performing multiple sampling of surrogate residuals for infinitely many times.

Remark 6. *Theorem 5 also implies that the surrogate residual and functional residual should generate similar patterns in residual-vs-covariate plots (other than a $G(\cdot)$ transformation) when both residuals apply (e.g., in cumulative link models) and when the sample size is large.*

The result below shows that the probability-scale residual (Li and Shepherd, 2012)

$$R_{PS} = E\{sign(y - Y)\} = Pr\{y < Y\} - Pr\{y > Y\}$$

can also be derived from our functional residual. This residual is the difference between the probabilities of the variable greater than or less than the observed value.

Theorem 6. *Given the data $(y, \boldsymbol{x})$, the probability-scale residual R_{SBS} can be expressed using the expectation of the functional residual variable, i.e., $R_{PS} = 2E\{R_{func}|(y, \boldsymbol{x})\} - 1$.*

The usefulness of the probability-scale residual relies on its first-moment property, i.e., $E(R_{PS}) = 0$ when the model is correct. The residual examination should be limited to the conditional mean. Other unusual patterns, as observed in Figure 4(e), may not be indications of model misspecification (see the discussion in Liu and Zhang, 2018 and Liu et al., 2021a). Theorem 6 explains where this limitation stems from. Basically, it says the probability-scale residual extracts the mean information of the functional residual. As a result, it preserves the first-moment information of the CDF $Res(\cdot; y, \boldsymbol{x})$ but may lose other distributional information, such as variance, skewness, and other higher-order moments.

In Supplementary Materials F, we provide a more comprehensive comparison between our approach and other recently developed diagnostic tools for discrete data. The discussion focuses on theoretical merits and limitations of each method with numerical evidence.

7 Summary

In this article, we have established a new framework of model diagnostics for discrete data regression. The framework is based on a new notion of functional residual, which is fundamentally different from the traditional residual concept. We have developed theoretical properties of functional residuals to support the use of the residual-*vs*-covariate plot and *Fn*-*Fn* plot as diagnostic tools. The simulated examples in a variety of settings and case studies have shown that our framework has the power to detect misspecification of many important components of GLMs for ordinal/count data and their semiparametric extensions.

The concept of functional residual offers a unifying framework in the following respects: (i) it unifies GLM diagnostics for binary, ordinal, and count data, and it lends itself to diagnostics for semiparametric models; (ii) it unifies two useful point residuals, the surrogate residual and probability-scale residual, in the same framework; (iii) it unifies the interpretation of diagnostic plots for discrete outcomes and continuous outcomes. We close the paper by elaborating on (iii). It has been demonstrated throughout the paper that the interpretation of our diagnostic plots for discrete data is similar to that for continuous data when linear models are used. An immediate and important benefit is that practitioners can interpret and draw insights from our diagnostic plots as if the outcome data were continuous and a linear model were used. The similarity between ours and conventional diagnostics for linear models can certainly lower the learning/comprehension cost and facilitate easier communications with decision/policy makers in industries, governments and nonprofit organizations.

Our focus is not on hypothesis testing but residual diagnostics. The notion of functional residuals, however, can be used to develop goodness-of-fit tests. In Supplementary Materials G, we perform graphical tests based on the F_n-F_n plot and the result in Theorem 3. As discussed in the introduction, goodness-of-fit tests concern *statistical significance* of model deviations, which often depends on the sample size n. When the working model is not a

perfect fit, a "non-concerning" deviation may lead to a p-value being very small if n is large. Residual diagnostics aims to visualize *practical significance* of model deviations. A notable deviation, if discovered, might not become more prominent if more data were available; see Supplementary Materials H for examples where diagnostic plots are generated using different sample sizes. Furthermore, goodness-of-fit tests in general do not generate insights into what components of the model are specified incorrectly and what treatment may be needed. In this regard, our residual diagnostics serve as a tool to *generate hypotheses* and *recommend actions*. This utility is poised to play a crucial role in the broader analytics domain (e.g., business analytics) where actionable insights are favored over simple binary decisions.

References

Archer, K. J., Lemeshow, S., and Hosmer, D. W. (2007), "Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design," *Computational Statistics and Data Analysis*, 51, 4450–4464.

Ben, M. G. and Yohai, V. J. (2004), "Quantile–quantile plot for deviance residuals in the generalized linear model," *Journal of Computational and Graphical Statistics*, 13, 36–47.

Blöchlinger, A. and Leippold, M. (2011), "A new goodness-of-fit test for event forecasting and its application to credit defaults," *Management Science*, 57, 487–505.

Cheng, C., Wang, R., and Zhang, H. (2021), "Surrogate residuals for discrete choice models," *Journal of Computational and Graphical Statistics*, 30, 67–77.

Contardo, C., Morency, C., and Rousseau, L.-M. (2012), *Balancing a Dynamic Public Bike-sharing System*, vol. 4, Cirrelt: Montreal.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, 47, 547–553.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, vol. 180, Springer:

New York.

Davison, A., Gigli, and A (1989), "Deviance residuals and normal scores plots," *Biometrika*, 76, 211–221.

De Jong, P. and Heller, G. Z. (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press: NY, USA.

Fagerland, M. W. and Hosmer, D. W. (2016), "Tests for goodness of fit in ordinal logistic regression models," *Journal of Statistical Computation and Simulation*, 86, 3398–3418.

Fanaee-T, H. and Gama, J. (2014), "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, 2, 113–127.

Faraway, J. J. (2016), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman and Hall/CRC.

Fernández, D. and Liu, I. (2016), "A goodness-of-fit test for the ordered stereotype model," *Statistics in Medicine*, 35, 4660–4696.

Franses, P. H. and Paap, R. (2001), *Quantitative Models in Marketing Research*, Cambridge University Press: Cambridge, UK.

Frees, E. W. (2009), *Regression Modeling with Actuarial and Financial Applications*, Cambridge University Press: NY, USA.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997), "A comparison of goodness-of-fit tests for the logistic regression model," *Statistics in Medicine*, 16, 965–980.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), "Graphical methods for assessing logistic regression models," *Journal of the American Statistical Association*, 79, 61–71.

Li, C. and Shepherd, B. (2010), "Test of association between two ordinal variables while adjusting for covariates," *Journal of the American Statistical Association*, 105, 612–620.

— (2012), "A new residual for ordinal outcomes," *Biometrika*, 99, 473–480.

Li, S., Fan, Z., Liu, I., Morrison, P. S., and Liu, D. (2024), "Surrogate method for partial as-

sociation between mixed data with application to well-being survey analysis," *The Annals of Applied Statistics*, 18, 2254–2276.

Li, Y., Zheng, Y., Zhang, H., and Chen, L. (2015), "Traffic Prediction in a Bike-Sharing System," in *Proceedings of the 23rd ACM International Conference on Advances in Geographical Information Systems*, ACM SIGSPATIAL 2015.

Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (1996), "Goodness-of-fit tests for ordinal response regression models," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45, 175–190.

Liu, D., Li, S., Yu, Y., and Moustaki, I. (2021a), "Assessing partial association between ordinal variables: quantification, visualization, and hypothesis testing," *Journal of the American Statistical Association*, 116, 955–968.

Liu, D., Liu, R., and Xie, M. (2014), "Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events," *Journal of the American Statistical Association*, 109, 1450–1465.

— (2015), "Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness," *Journal of the American Statistical Association*, 110, 326–340.

— (2021b), "Nonparametric fusion learning for multiparameters: Synthesize inferences from diverse sources using data depth and confidence distribution," *Journal of the American Statistical Association*, 117, 20862104.

Liu, D. and Zhang, H. (2018), "Residuals and diagnostics for ordinal regression models: a surrogate approach," *Journal of the American Statistical Association*, 113, 845–854.

Liu, D., Zhu, X., Greenwell, B., and Lin, Z. (2023), "A new goodness-of-fit measure for probit models: Surrogate R^2," *British Journal of Mathematical and Statistical Psychology*, 76, 192–210.

Liu, Q., Li, C., Wanga, V., and Shepherd, B. E. (2018), "Covariate-adjusted Spearman's

rank correlation with probability-scale residuals," *Biometrics*, 74, 595–605.

Luo, X., Dasgupta, T., Xie, M., and Liu, R. Y. (2021), "Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83, 777–797.

Nattino, G., Pennell, M. L., and Lemeshow, S. (2020), "Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test," *Biometrics*, 76, 549–560.

Pierce, D. A. and Schafer, D. W. (1986), "Residuals in generalized linear models," *Journal of the American Statistical Association*, 81, 977–986.

Pregibon, D. (1981), "Logistic regression diagnostics," *The Annals of Statistics*, 9, 705–724.

Pyne, D. A. (1979), "Single-variable Poisson regression: A goodness-of-fit test and the comparison of regression coefficients," *Journal of the American Statistical Association*, 74, 489–493.

Schweder, T. and Hjort, N. L. (2016), *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*, United Kingdom: Cambridge University Press.

Shen, J., Liu, R. Y., and Xie, M.-g. (2020), "i fusion: individualized fusion learning," *Journal of the American Statistical Association*, 115, 1251–1267.

Shepherd, B. E., Li, C., and Liu, Q. (2016), "Probability-scale residuals for continuous, discrete, and censored data," *Canadian Journal of Statistics*, 44, 463–479.

Wasserstein, R. L. and Lazar, N. A. (2016), "The ASA statement on p-values: context, process, and purpose," *The American Statistician*, 70, 129–133.

Xie, M. and Singh, K. (2013), "Confidence distribution, the frequentist distribution estimator of a parameter: a review (with discussion)," *International Statistical Review*, 81, 2–39.

Yang, L. (2021), "Assessment of regression models with discrete outcomes using quasi-empirical residual distribution functions," *Journal of Computational and Graphical Statistics*, 30, 1019–1035.