Model diagnostics of discrete data regression: a unifying framework using functional residuals

Zewei Lin

Ph.D. Candidate in Business Analytics,
Department of Operations, Business Analytics, & Information Systems,
University of Cincinnati, Cincinnati, Ohio, U.S.A

SDSS 2022

*The project is coauthored with my advisor Dr. Dungang Liu.

Background

- Binary data: bankruptcy/default status or disease prevention/treatment outcome
- Ordinal data: ratings of bonds, school districts, and pain severity
- Integer-valued count data: records of insurance claims, emergency room visits, and frequency of product/device usage

Motivation

- There is no well-established model diagnostic tools for discrete data regression models.
  - Pearson/deviance residual analysis and goodness-of-fit tests have limited utility in model diagnostics and treatment.
- The general interest has shifted from making a “yes/no” decision to knowing *why, how, and what to do*.
  - The fact that a *p*-value dose not measure the size of an effect undermines the usefulness of goodness-of-t tests in general.

Our Contribution

Unlike the literature defining a single-valued quantity as the residual, we propose to use a **function** as a vehicle to retain the residual information. In the presence of data discreteness, we show that

- a functional residual is appropriate for summarizing the residual randomness that cannot be captured by the structural part of the model
- its theoretical properties lead to the **innovation of new diagnostic tools** including the functional-residual-vs-covariate plot and Function-to-Function (Fn-Fn) plot
- it broadens the diagnostic scope as it applies to virtually **all parametric models for binary, ordinal and count data, all in a unified diagnostic scheme**
- 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.

Definition

The most general form of *working model* for discrete data can be written as

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

where $\pi(y; \boldsymbol{X}, \boldsymbol{\beta}) = \Pr\{Y \le y \mid \boldsymbol{X}, \boldsymbol{\beta}\}$ is a discrete distribution function.

Definition: 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 $\Pi = \{F(t) : 0 \le t \le 1; 0 \le F(t) \le 1;$ and $F(t_1) \le F(t_2)$ for any $t_1 < t_2\}$. Specifically,

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

Illustration

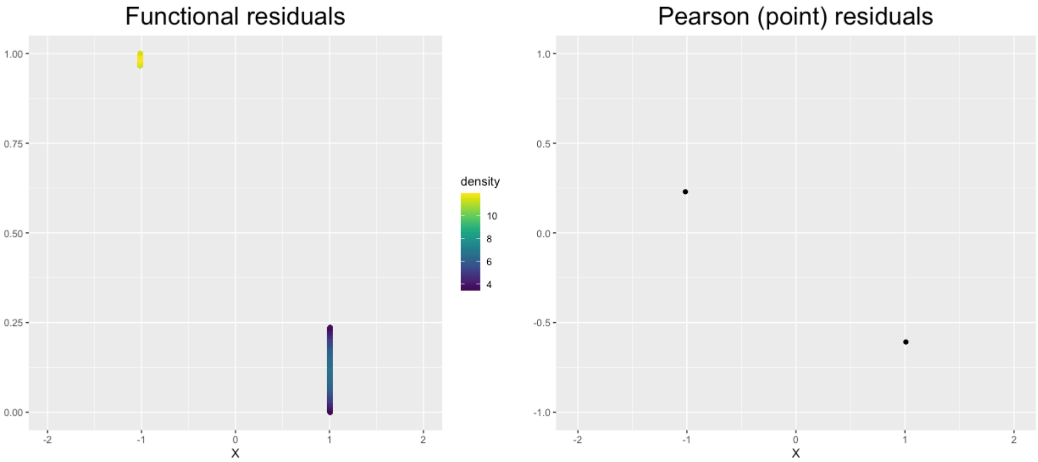

Figure: A comparison of functional residuals and Pearson (point) residuals when the observation is $\left(y = 0, x = 1\right)$ or $\left(y = 1, x = -1\right)$.

Theoretical properties

We propose to use a **functional-residual-*vs*-covariate plot** to examine the working model.

Theorem 1 (Conditional Expectation under the Null). Given $\boldsymbol{X} = \boldsymbol{x}$, the conditional expectation of the functional residual $Res(\cdot; Y, \boldsymbol{x})$ 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 $\pi(\cdot; \boldsymbol{x}) \equiv \pi_0(\cdot; \boldsymbol{x})$.

Simulated example

We simulate 1000 ordinal data points y_i $\left(= 0, 1, 2, 3, 4\right)$ 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 $\left(\alpha_0, \alpha_1, \alpha_2, \alpha_3\right) = \left(1.5, 1.5, -1, 1\right)$, $\left(\beta_1, \beta_2\right) = \left(1.5, -1\right)$ and the covariate $X \sim \mathcal{N}\left(0, 1\right)$.

New diagnostic tool

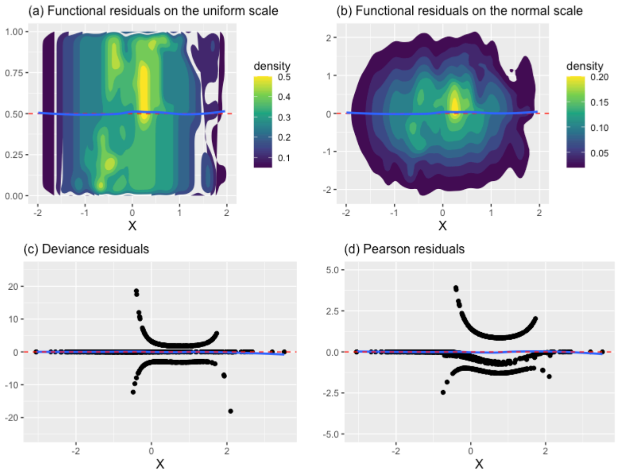

Figure: Proposed functional-residual-*vs*-covariate plots (upper row) and traditional residual-*vs*-covariate plots (lower row) when the working model is specified correctly for ordinal data.

Theoretical properties

We propose to draw the *Function-Function (Fn-Fn) plot* which is an **analogy** to Q-Q plot.

Theorem 2. 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 $t \in (0, 1)$,

$$\overline{Res}(t) = \frac{1}{n} \sum_{i=1}^{n} Res(t; Y_i, \boldsymbol{X}_i) \to F_{U(0,1)}(t) = t \quad \text{almost surely}, \tag{3}$$

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

Function-Function (Fn-Fn) plot

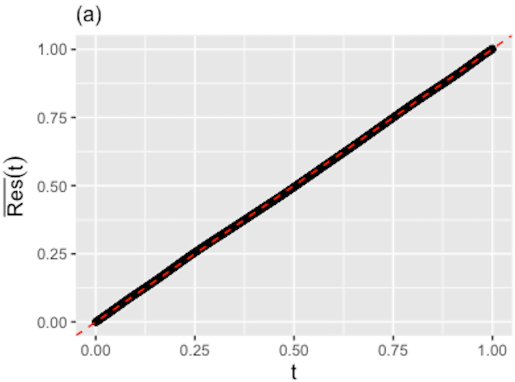

Figure: Proposed *Fn-Fn* plots when the working model is specified correctly for ordinal data.

Model misspecification detection

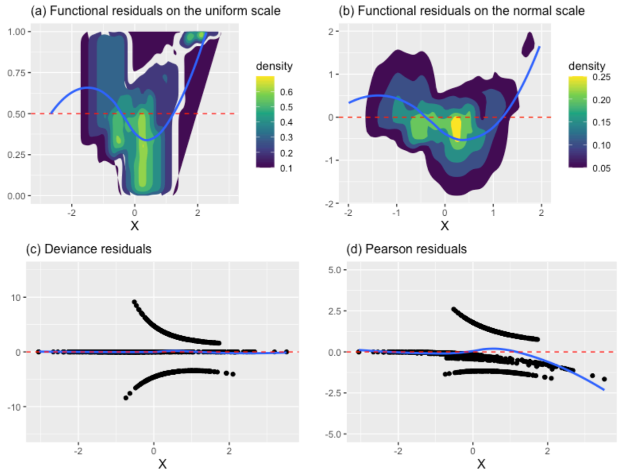

Figure: Proposed functional-residual-*vs*-covariate plots (upper row) and traditional residual-*vs*-covariate plots (lower row) when the quadratic term X^2 is missing in the working adjacent-category logit model.

Missing of a higher order

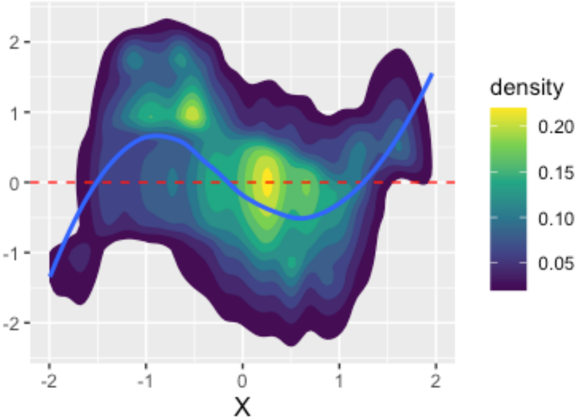

Figure: Functional-residual-*vs*-covariate plot when the cubic term X^3 is missing in the working adjacent-category logit model.

Missing of relevant covariates

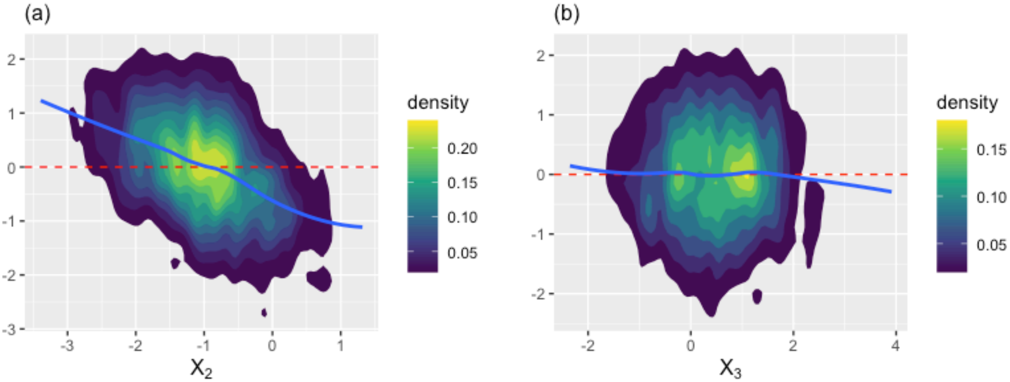

Figure: Functional-residual-*vs*-covariate plots when X_2 is correlated with ordinal data Y (the left panel) whereas X_3 is not (the right panel).

Missing of interaction

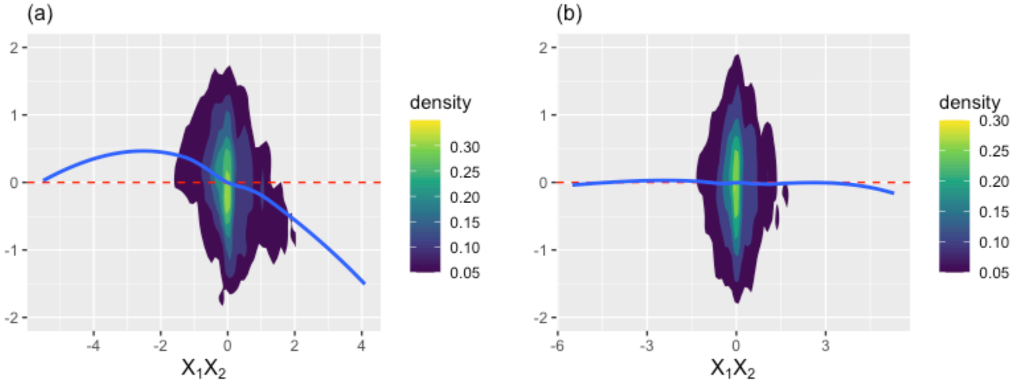

Figure: Functional-residual-*vs*-covariate plots before (left) and after (right) the interaction term X_1X_2 is included in the working adjacent-category logit model.

University of CINCINNATI

Bike sharing modeling

- We perform statistical modeling of the data from Capital Bike Sharing System at Washington D.C. The data set contains 8734 observations of the hourly bike rentals in 2012.
- To improve the efficiency of the rental system, it is crucial to examine how weather conditions and time/day influence consumer behavior.
- As the outcome is the number of hourly rentals during a day, Poisson regression models can generate interpretable insights for system managers. But again, a general diagnostic procedure is needed to guide, assess, and refine the model building process.

Initial Poisson regression model

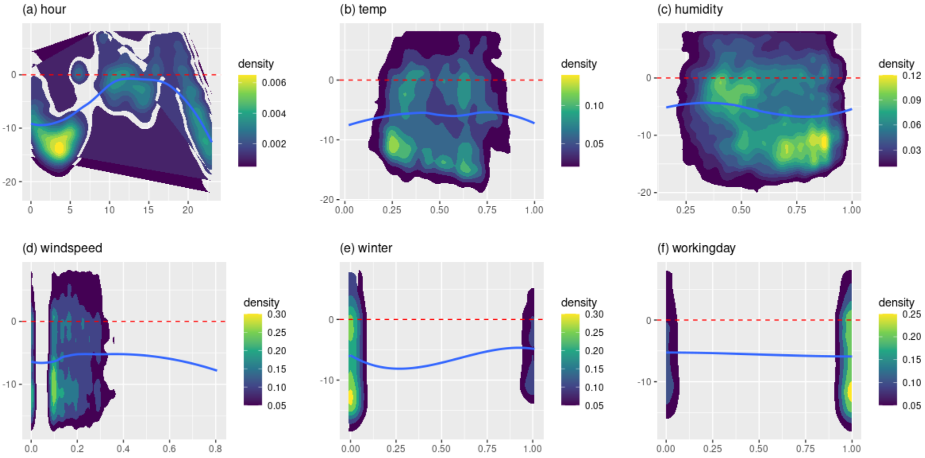

Figure: Functional-residual-*vs*-covariate plots for the initial Poisson model fitted to the Captial Bikeshare dataset.

After adding smoothing functions

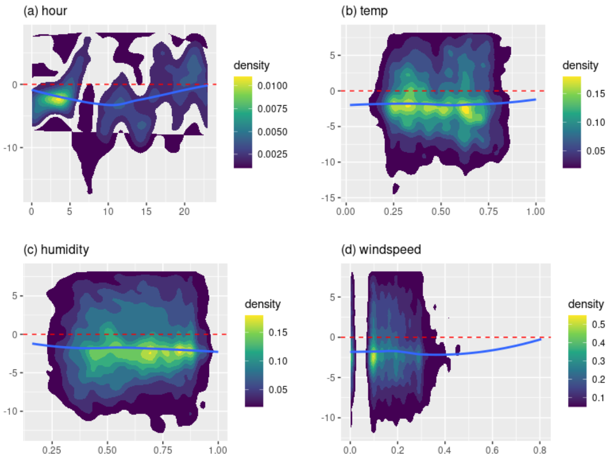

Figure: Functional-residual-*vs*-covariate plots after adding the smoothing functions of the variables *hour, temp, humidity,* and *windspeed* to the Poisson model.

Modeling dispersion parameter

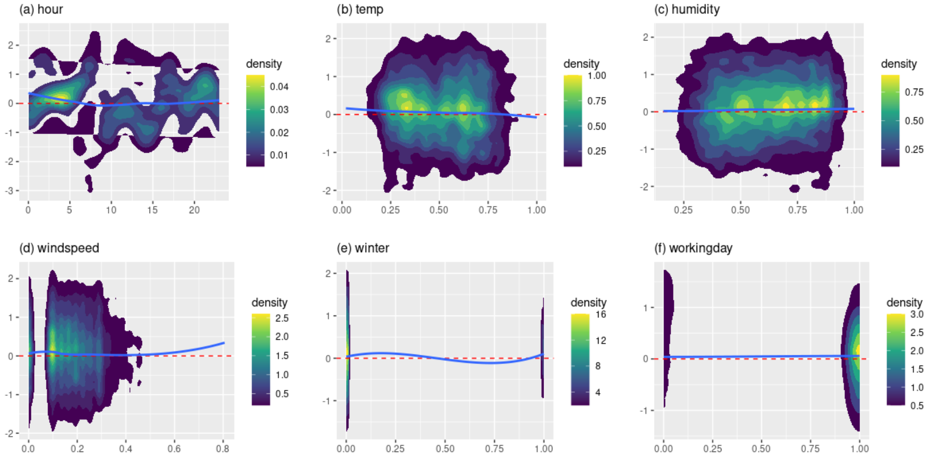

Figure: Functional-residual-*vs*-covariate plots for the final generalized additive quasi-Poisson model.

Fn-Fn plots

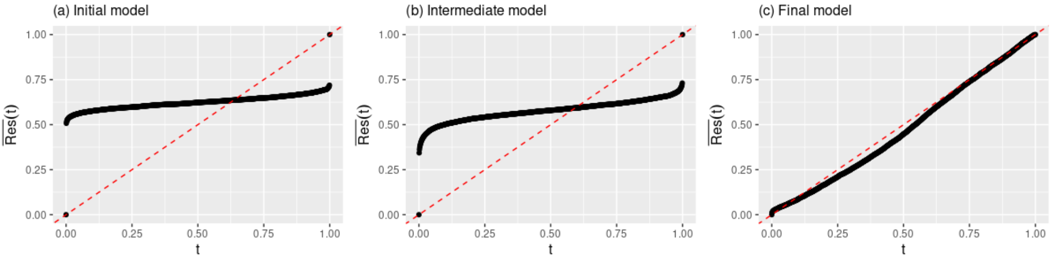

Figure: The *Fn-Fn* plots for the initial, intermediate, and final models developed in the model building process.

Reference

- Zewei Lin and Dungang Liu. Model diagnostics of discrete data regression: a unifying framework using functional residuals. *Submitted to Journal of the American Statistical Association.*
- Liu, Dungang, Shaobo Li, Yan Yu, and Irini Moustaki. Assessing partial association between ordinal variables: quantification, visualization, and hypothesis testing.*Journal of the American Statistical Association* 116(534):955-968, 2021.
- Dungang Liu and Heping Zhang. Residuals and diagnostics for ordinal regression models: a surrogate approach. *Journal of the American Statistical Association*, 113(522):845854, 2018.
- C.Li and B.E.Shepherd. A new residual for ordinal outcomes. *Biometrika*, 99(2):473480, 2012.