Supplementary Materials

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

A Proofs of the theoretical results

A.1 Proofs for Section 3

Proof of Theorem 1. Let y_{max} be the maximum of all the possible values of Y ($y_{max} = \infty$ for count data), and y_t be such that $\pi(y_t - 1; \boldsymbol{x}) \leq t \leq \pi(y_t; \boldsymbol{x})$. Then, we can write

$$\begin{aligned}
E_{(Y|\boldsymbol{X}=\boldsymbol{x})} Res(t; Y, \boldsymbol{x}) &= \sum_{y=0}^{y_{max}} Res(t; y, \boldsymbol{x}) \Pr\{Y = y\} \\
&= \sum_{y=0}^{y_{max}} \Pr\{U(\pi(y-1; \boldsymbol{x}), \pi(y; \boldsymbol{x})) \leq t\} \Pr\{Y = y\} \\
&= \sum_{y=0}^{y_t - 1} \Pr\{Y = y\} + \Pr\{U(\pi(y_t - 1; \boldsymbol{x}), \pi(y_t; \boldsymbol{x})) \leq t\} \Pr\{Y = y_t\} \\
&= \pi_0(y_t - 1; \boldsymbol{x}) + \frac{t - \pi(y_t - 1; \boldsymbol{x})}{\pi(y_t; \boldsymbol{x}) - \pi(y_t - 1; \boldsymbol{x})} \big\{\pi_0(y_t; \boldsymbol{x}) - \pi_0(y_t - 1; \boldsymbol{x})\big\} \\
&= t.
\end{aligned}$$

The last equation is a result of the assumption $\pi(\cdot; \boldsymbol{x}) \equiv \pi_0(\cdot; \boldsymbol{x})$. This completes the proof. $\square$

Proof of Theorem 2. By the law of iterated expectations, we have

$$\begin{aligned} E_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{x}) &= E_{\boldsymbol{X}}\{E_{(Y|\boldsymbol{X}=\boldsymbol{x})}Res(t;Y,\boldsymbol{x})\} \\ &= \int E_{(Y|\boldsymbol{X}=\boldsymbol{x})}Res(t;Y,\boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}) = \int t\, dF_{\boldsymbol{X}}(\boldsymbol{x}) = t \end{aligned}$$

The establishment above holds for every $t \in [0,1]$. This completes the proof. □

Proof of Corollary 1. To prove Corollary 1, we only need to verify the finite variance assumption of the strong law of large numbers. The expression for the variance can be written as follows:

$$Var_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{X}) = E_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{X})^2 - \{E_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{X})\}^2.$$

Since $Res(t;Y,\boldsymbol{X}) \in [0,1]$, we immediately have $E_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{X})^2 \leq 1$. Therefore,

$$Var_{(Y,\boldsymbol{X})}Res(t;Y,\boldsymbol{X}) \leq 1 - t^2 \leq 1,$$

which proves the variance of $Res(t;Y,\boldsymbol{X})$ is finite. □

Proof of Theorem 3. Since $(Y_1,\boldsymbol{X}_1),(Y_2,\boldsymbol{X}_2),\ldots$ is an infinite sequence of i.i.d. random variables, and we have that

$$\begin{aligned} Var(\frac{1}{n}\sum_{i=1}^{n} Res(t;Y_i,\boldsymbol{X}_i)) &= \frac{1}{n^2}Var(Res(t;Y_1,\boldsymbol{X}_1)+\cdots+Res(t;Y_n,\boldsymbol{X}_n)) \\ &= \frac{nVar(Res(t;Y,\boldsymbol{X}))}{n^2} \leq \frac{1}{n}. \end{aligned}$$

The last inequality is due to the fact that $Var(Res(t;Y_n,\boldsymbol{X}_n)) \leq 1$ as proved in Theorem 2. Therefore, we can use Chebyshev's inequality on $\frac{1}{n}\sum_{i=1}^{n} Res(t;Y_i,\boldsymbol{X}_i)$, which results in

$$\Pr\left\{|\frac{1}{n}\sum_{i=1}^{n} Res(t;Y_i,\boldsymbol{X}_i) - t| \geq \epsilon\right\} \leq \frac{Var(\frac{1}{n}\sum_{i=1}^{n} Res(t;Y_i,\boldsymbol{X}_i))}{\epsilon^2} \leq \frac{1}{n\epsilon^2}.$$

The inequality above is equivalent to

$$\Pr\left\{|\overline{Res}(t; Y_n, \boldsymbol{X}_n) - t| < \epsilon\right\} \geq 1 - \frac{1}{n\epsilon^2}.$$

As the right side of the inequality does not depend on t, we have

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

This completes the proof. □

Proof of Theorem 4. For the null model $\mathcal{M}_0$, we have proved in Theorem 3 that

$$\Pr\left\{|\overline{Res}_n^{(\mathcal{M}_0)}(t; Y_n, \boldsymbol{X}_n) - t| < \epsilon\right\} \geq 1 - \frac{1}{n\epsilon^2}$$

for any $t \in (0, 1)$ and $\epsilon > 0$. We let $\epsilon = n^{-1/2+c/2}$, then,

$$\inf_{t\in(0,1)} \Pr\left\{D_n^{(\mathcal{M}_0)}(t) < n^{-1/2+c/2}\right\} \geq 1 - n^{-c}.$$

As a result,

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

On the other hand, for the alternative model,

$$\Pr\left\{(n^{1/2-c/2})D_{n,max}^{(\mathcal{M}_a)} > 1\right\} = \Pr\left\{(n^{1/2-c})D_{n,max}^{(\mathcal{M}_a)} > n^{-a/2}\right\} \to \Pr\{D > n^{-c/2}\} \to 1.$$

Therefore,

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

This completes the proof.

□

A.2 Consistency of the functional residual and parallel results with $\hat{\beta}$

For the ease of presentation, the main paper uses 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 $\hat{Res}(t; y, \boldsymbol{x}) \equiv Res_{\hat{\boldsymbol{\beta}}}(t; y, \boldsymbol{x})$. In this section, we establish a consistency result of $\hat{Res}(t; y, \boldsymbol{x})$. The only assumption needed is the consistency of the probability estimate $\hat{\pi}(y; \boldsymbol{x}, \hat{\boldsymbol{\beta}})$.

Assumption A.1. *The estimate $\hat{\boldsymbol{\beta}}$ is consistent such that $\hat{\pi}(y; \boldsymbol{x}, \hat{\boldsymbol{\beta}}) = \pi(y; \boldsymbol{x}, \boldsymbol{\beta}) + o_p(1)$.*

Theorem A.1 (Consistency of the Functional Residual)**.** *Under Assumption A.1, $\hat{Res}(t; y, \boldsymbol{x}) = Res(t; y, \boldsymbol{x}) + o_p(1)$.*

Proof. Based on Assumption A.1, we let $\hat{\pi}(y-1; \boldsymbol{x}, \hat{\boldsymbol{\beta}}) = \pi(y-1; \boldsymbol{x}, \boldsymbol{\beta}) + e_1$ and $\hat{\pi}(y; \boldsymbol{x}, \hat{\boldsymbol{\beta}}) = \pi(y; \boldsymbol{x}, \boldsymbol{\beta}) + e_2$, where $e_1 = o_p(1), e_2 = o_p(1)$. We write

$$\begin{aligned}\hat{Res}(t; y, \boldsymbol{x}) = F_{U(\hat{\pi}(y-1;\boldsymbol{x}), \hat{\pi}(y;\boldsymbol{x}))}(t) &= \Pr\{U(\hat{\pi}(y-1; \boldsymbol{x}), \hat{\pi}(y; \boldsymbol{x})) \leq t\} \\ &= \Pr\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\}.\end{aligned}$$

In the following, we consider three scenarios depending on the value of t.

Scenario 1. If $t < \pi(y-1; \boldsymbol{x})$, let $d = \pi(y-1; \boldsymbol{x}) - t$. We have

$$\begin{aligned}\hat{Res}(t; y, \boldsymbol{x}) - Res(t; y, \boldsymbol{x}) &= \Pr\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\} - 0 \\ &= \Pr\Big\{\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\} \cap \{\{|e_1| \geq d\} \cup \{|e_1| < d\}\}\Big\} \\ &= \Pr\Big\{\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\} \cap \{|e_1| \geq d\}\Big\} + \\ &\quad \Pr\Big\{\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\} \cap \{|e_1| < d\}\Big\} \\ &= o_p(1) + 0 = o_p(1).\end{aligned}$$

Similarly, if $t > \pi(y; x)$, we have

$$\begin{aligned}\hat{Res}(t; y, \boldsymbol{x}) - Res(t; y, \boldsymbol{x}) &= \Pr\{U(\pi(y-1; \boldsymbol{x}) + e_1, \pi(y; \boldsymbol{x}) + e_2) \leq t\} - 1 \\ &= 1 + o_p(1) - 1 = o_p(1).\end{aligned}$$

Scenario 2. If $\pi(y-1;\boldsymbol{x}) < t < \pi(y;\boldsymbol{x})$, let $d = \min\{t-\pi(y-1;\boldsymbol{x}), \pi(y;\boldsymbol{x})-t\}$. We have

$$\hat{Res}(t;y,\boldsymbol{x}) = \Pr\Big\{\{U(\pi(y-1;\boldsymbol{x})+e_1, \pi(y;\boldsymbol{x})+e_2) \le t\} \cap \\ \big\{\{|e_1| < d \wedge |e_2| < d\} \cup \{|e_1| \ge d \vee |e_2| \ge d\}\big\}\Big\} = \bigcirc_1 + \bigcirc_2,$$

where $\bigcirc_1 = \Pr\big\{\{U(\pi(y-1;\boldsymbol{x})+e_1, \pi(y;\boldsymbol{x})+e_2) \le t\} \cap \{|e_1| < d \wedge |e_2| < d\}\big\}$, and $\bigcirc_2 = \Pr\big\{\{U(\pi(y-1;\boldsymbol{x})+e_1, \pi(y;\boldsymbol{x})+e_2) \le t\} \cap \{|e_1| \ge d \vee |e_2| \ge d\}\big\}$. We have

$$\bigcirc_1 = \frac{t-\pi(y-1;\boldsymbol{x})+o_p(1)}{\pi(y;\boldsymbol{x})-\pi(y-1;\boldsymbol{x})+o_p(1)} = \frac{t-\pi(y-1;\boldsymbol{x})}{\pi(y;\boldsymbol{x})-\pi(y-1;\boldsymbol{x})} + o_p(1),$$

and

$$\bigcirc_2 \le \Pr\{|e_1| \ge d\} + \Pr\{|e_2| \ge d\}\big\} = o(1).$$

Therefore, $\hat{Res}(t;y,\boldsymbol{x}) - Res(t;y,\boldsymbol{x}) = \bigcirc_1 + \bigcirc_2 - \frac{t-\pi(y-1;\boldsymbol{x})}{\pi(y;\boldsymbol{x})-\pi(y-1;\boldsymbol{x})} = o_p(1)$.

Scenario 3. If $t = \pi(y-1;\boldsymbol{x})$,

$$\begin{aligned}\hat{Res}(t;y,\boldsymbol{x}) - Res(t;y,\boldsymbol{x}) &= \Pr\{U(\pi(y-1;\boldsymbol{x})+e_1, \pi(y;\boldsymbol{x})+e_2) \le t \wedge e_1 < 0\} + \\ &\quad \Pr\{U(\pi(y-1;\boldsymbol{x})+e_1, \pi(y;\boldsymbol{x})+e_2) \le t \vee e_1 \ge 0\} - 0 \\ &= \frac{o_p(1)}{\pi(y;\boldsymbol{x})-\pi(y-1;\boldsymbol{x})+o_p(1)} = o_p(1).\end{aligned}$$

Similarly, if $t = \pi(y;x)$, we have $\hat{Res}(t;y,\boldsymbol{x}) - Res(t;y,\boldsymbol{x}) = 1 + o_p(1) - 1 = o_p(1)$. This completes the proof. □

As a result of Theorem A.1, we can establish theoretical result for $\hat{Res}(t;y,\boldsymbol{x})$ in a parallel way to those in Theorems 1-3 as presented in Section 3 of the main paper. The distributional conclusions remain the same except a vanishing term.

Theorem 1′ (Conditional Expectation under the Null)**.** If Assumption A.1 holds, given $\boldsymbol{X} = \boldsymbol{x}$, the conditional expectation of the functional residual $\hat{Res}(\cdot;Y,\boldsymbol{x})$ is the CDF of a $U(0,1)$ distribution plus a vanishing term $o(1)$, i.e., $E_Y\hat{Res}(t;Y,\boldsymbol{x}) = F_{U(0,1)}(t) + o(1) = t + o(1)$ for any $t \in [0,1]$.

Theorem 2′ (Unconditional Expectation under the Null)**.** If Assumption A.1 holds, the unconditional expectation of the functional residual $\hat{Res}(\cdot; Y, \boldsymbol{X})$ is the CDF of a $U(0,1)$ distribution plus a vanishing term $o(1)$, i.e., $E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X}) = F_{U(0,1)}(t) + o(1) = t + o(1)$ for any $t \in [0,1]$.

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. If Assumption A.1 holds, the average of the functional residuals $\hat{Res}(t; Y_i, \boldsymbol{X}_i)$ converges to the function $F_{U(0,1)}(t) = t$ uniformly; that is, for any $\epsilon > 0$,

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

We prove Theorems 1′-3′ in the following.

Proof of Theorem 1′. Based on Theorem A.1, we have $E_Y\hat{Res}(t; Y, \boldsymbol{x}) = E_Y(Res(t; Y, \boldsymbol{x}) + o_p(1)) = t + E(o_p(1))$. So we only need to prove $E(o_p(1)) = o(1)$. We let $Z_n = o_p(1)$. By definition, $\forall\ \epsilon, \delta > 0$, $\exists\ N_{\epsilon,\delta}$ such that $\forall\ n \geq N_{\epsilon,\delta}$, we have $\Pr(|Z_n| \geq \delta) < \epsilon$. Therefore, $E(|Z_n|) = E(|Z_n|\mathbf{1}_{|Z_n|<\delta}) + E(|Z_n|\mathbf{1}_{|Z_n|\geq\delta}) \leq \delta + \epsilon$. Thus, $\lim_{n\to\infty} E(|Z_n|) = 0$. This completes the proof. □

Proof of Theorem 2′. We write

$$E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{x}) \;=\; E_{\boldsymbol{X}}\{E_{(Y|\boldsymbol{X}=\boldsymbol{x})}\hat{Res}(t; Y, \boldsymbol{x})\} = \int E_{(Y|\boldsymbol{X}=\boldsymbol{x})}\hat{Res}(t; Y, \boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}).$$

Based on Lebesgue dominated convergence theorem, we have $\int E_{(Y|\boldsymbol{X}=\boldsymbol{x})}\hat{Res}(t; Y, \boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}) \to \int E_{(Y|\boldsymbol{X}=\boldsymbol{x})}Res(t; Y, \boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}) = t$. Thus, $E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X}) = t + o(1)$. This completes the proof. □

Proof of Theorem 3′. Since $\hat{Res}(t; Y, \boldsymbol{X})$ is a measurement of probability, $E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X})^2 \leq 1$. According to Theorem 2′,

$$Var_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X}) = E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X})^2 - \{E_{(Y,\boldsymbol{X})}\hat{Res}(t; Y, \boldsymbol{X})\}^2 \leq 1 - (t + o(1))^2 \leq 1 + o(1).$$

Then, we have

$$\begin{aligned} Var(\frac{1}{n}\sum_{i=1}^{n}\hat{Res}(t;Y_i,\boldsymbol{X}_i)) &= \frac{1}{n^2}Var(\hat{Res}(t;Y_1,\boldsymbol{X}_1)+\cdots+\hat{Res}(t;Y_n,\boldsymbol{X}_n)) \\ &= \frac{nVar(\hat{Res}(t;Y,\boldsymbol{X}))}{n^2} \leq \frac{1+o(1)}{n}. \end{aligned}$$

Therefore, we can use Chebyshev's inequality which results in

$$\Pr\left\{|\frac{1}{n}\sum_{i=1}^{n}\hat{Res}(t;Y_i,\boldsymbol{X}_i)-t|\geq\epsilon\right\} \leq \frac{Var(\frac{1}{n}\sum_{i=1}^{n}\hat{Res}(t;Y_i,\boldsymbol{X}_i))}{\epsilon^2} \leq \frac{1+o(1)}{n\epsilon^2}.$$

The inequality above is equivalent to $\Pr\left\{|\overline{\hat{Res}}(t;Y_n,\boldsymbol{X}_n)-t|<\epsilon\right\} \geq 1-\frac{1}{n\epsilon^2}+o(1)$. As the right side of the inequality does not depend on t, we have

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

This completes the proof. □

A.3 Proofs for Section 6

Proof of Theorem 5. First, by the definitions of latent variable and probability density function of truncated distribution, we can get the probability density function of the surrogate variable:

$$f_{R_s}(r_s|Y=y) = \frac{g(r_s)}{G(\alpha_y-\boldsymbol{X}\beta)-G(\alpha_{y-1}-\boldsymbol{X}\beta)}, \tag{S1}$$

where $g(\cdot)=G'(\cdot)$ is the probability density function(PDF).

Secondly, the functional residual defined in (5) follows $F_{U(\pi(y-1;\boldsymbol{x}),\pi(y;\boldsymbol{x}))}(t)$. Therefore, the PDF of R_f can be written as:

$$f_{R_f}(r_f) = \frac{1}{\pi(y;\boldsymbol{x})-\pi(y-1;\boldsymbol{x})} = \frac{1}{G(\alpha_y-\boldsymbol{X}\beta)-G(\alpha_{y-1}-\boldsymbol{X}\beta)}$$

Now we define a random variable $R_H = G^{-1}(R_f)$, the PDF of R_H is:

$$\begin{aligned} f_{R_H}(r_H) &= f_{R_f}(G(r_H)) \\ &= \frac{|G(r_H)|'}{G(\alpha_y - \boldsymbol{X}\beta) - G(\alpha_{y-1} - \boldsymbol{X}\beta)} \\ &= \frac{g(r_H)}{G(\alpha_y - \boldsymbol{X}\beta) - G(\alpha_{y-1} - \boldsymbol{X}\beta)} \end{aligned} \tag{S2}$$

The comparison between (S1) and (S2) suggests that $G^{-1}(R_f)$ and R_s follow the same distribution which completes the proof. □

Proof of Theorem 6. First, we write the equation of R_{SBS} based on $F_a(\cdot)$:

$$\begin{aligned} R_{SBS} &= \Pr\{y > Y\} - \Pr\{y < Y\} \\ &= \Pr\{Y < y\} - \Pr\{Y > y\} \\ &= \Pr\{Y \leq y - 1\} - (1 - \Pr\{Y \leq y\}) \\ &= (F_a(y) + F_a(y-1)) - 1 \end{aligned}$$

Secondly, since $R_{func} \sim Res(t; y, \boldsymbol{x}) = F_{U(\pi(y-1;\boldsymbol{x}),\pi(y;\boldsymbol{x}))}(t)$, the conditional expectation of $E\{R_{func}|(y, \boldsymbol{x})\}$ is $\frac{1}{2}(F_a(y) + F_a(y-1))$. Therefore, $2E\{R_{func}|(y, \boldsymbol{x})\} - 1 = (F_a(y) + F_a(y-1)) - 1 = R_{SBS}$. This completes the proof. □

B Figures

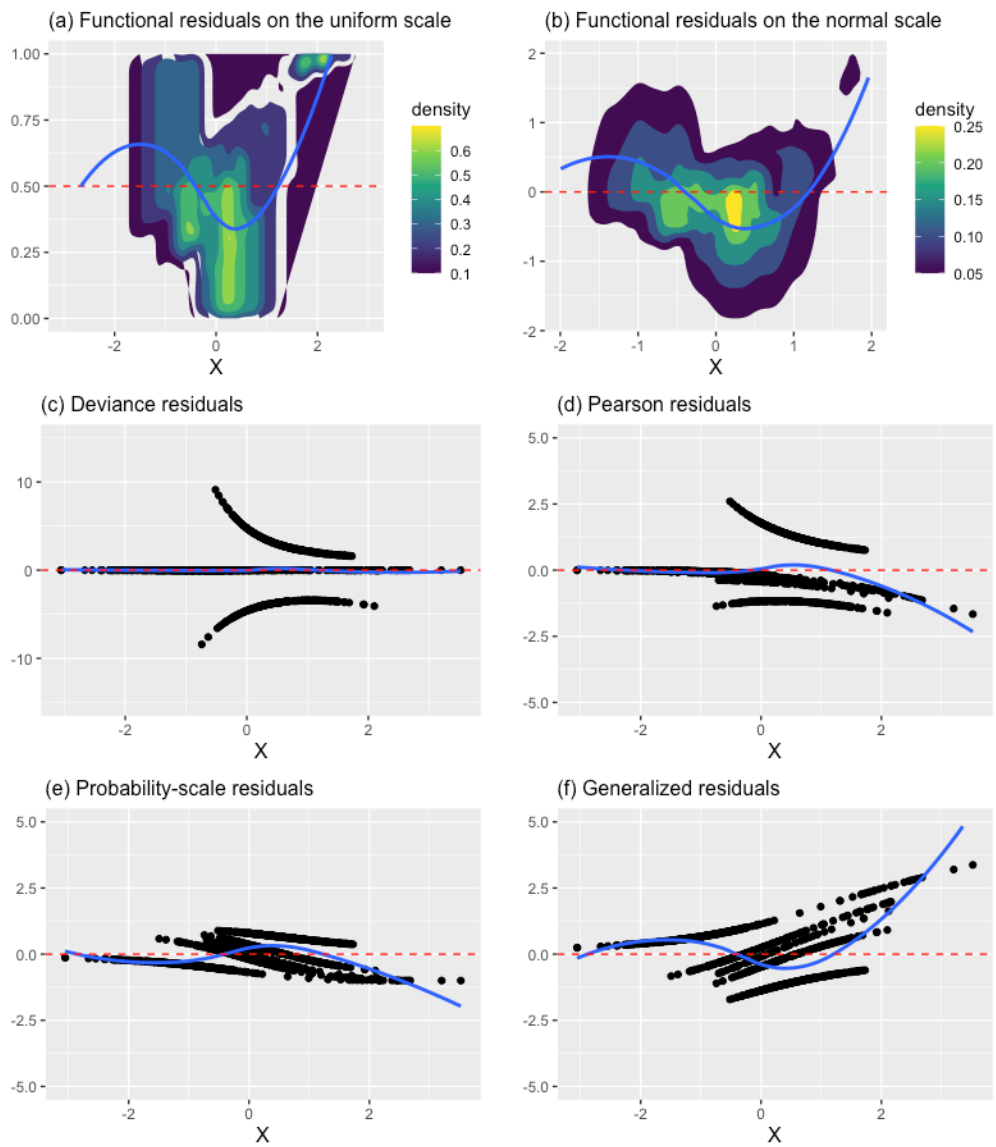

Figure S1: Proposed functional-residual-*vs*-covariate plots (top row) and other residual-*vs*-covariate plots (middle and bottom rows) when the quadratic term X^2 is missing in the working adjacent-category logit model (Example 2).

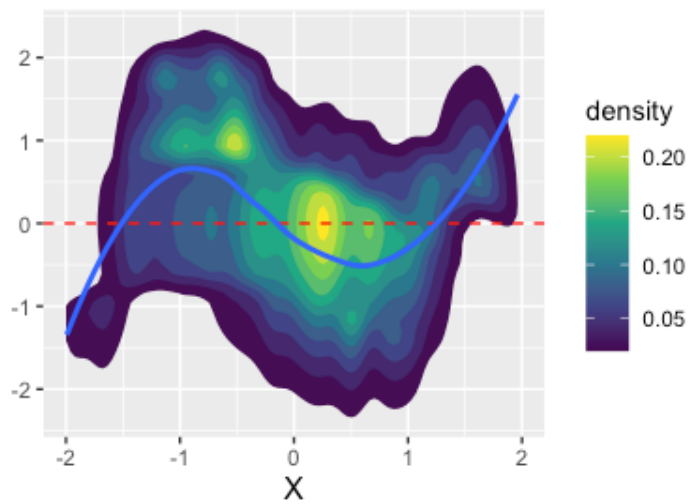

Figure S2: Functional-residual-*vs*-covariate plot when the cubic term X^3 is missing in the working adjacent-category logit model (Example 2).

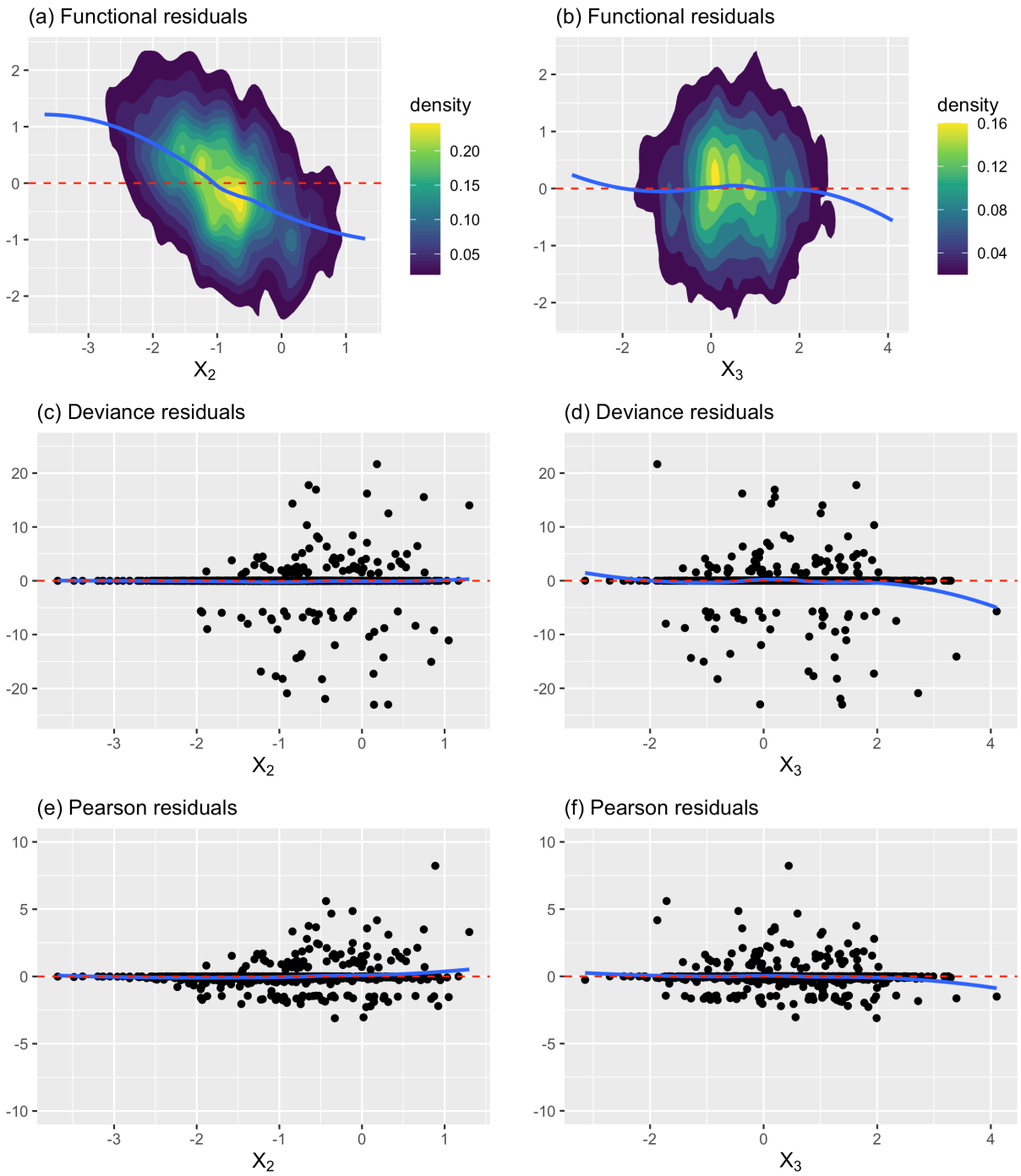

Figure S3: Residual-*vs*-covariate plots when X_2 is correlated with ordinal data Y (the left panel) whereas X_3 is not (the right panel) in the setting of Example 3.

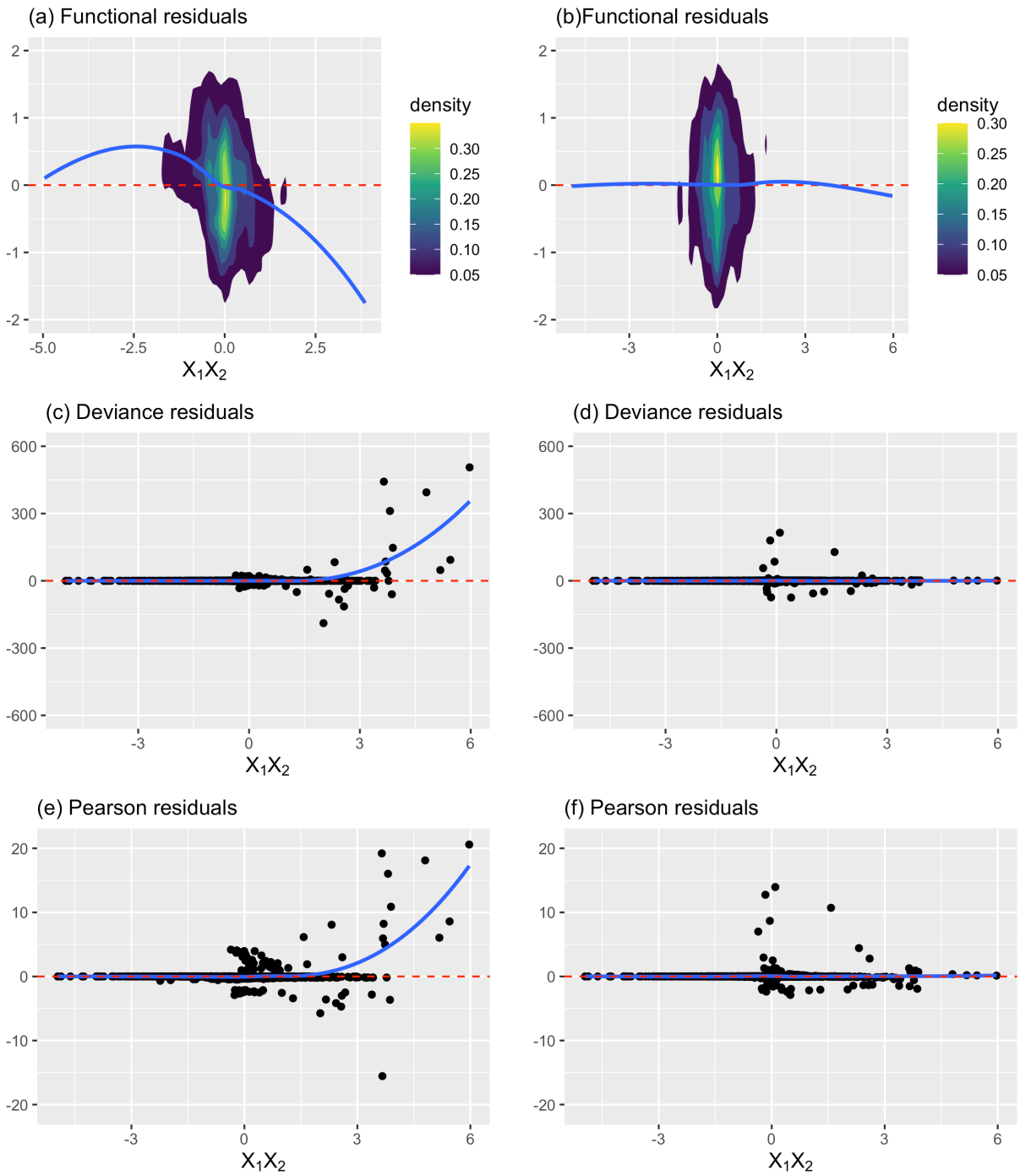

Figure S4: Residual-*vs*-covariate plots before (left) and after (right) the interaction term X_1X_2 is included in the working adjacent-category logit model (Example 4).

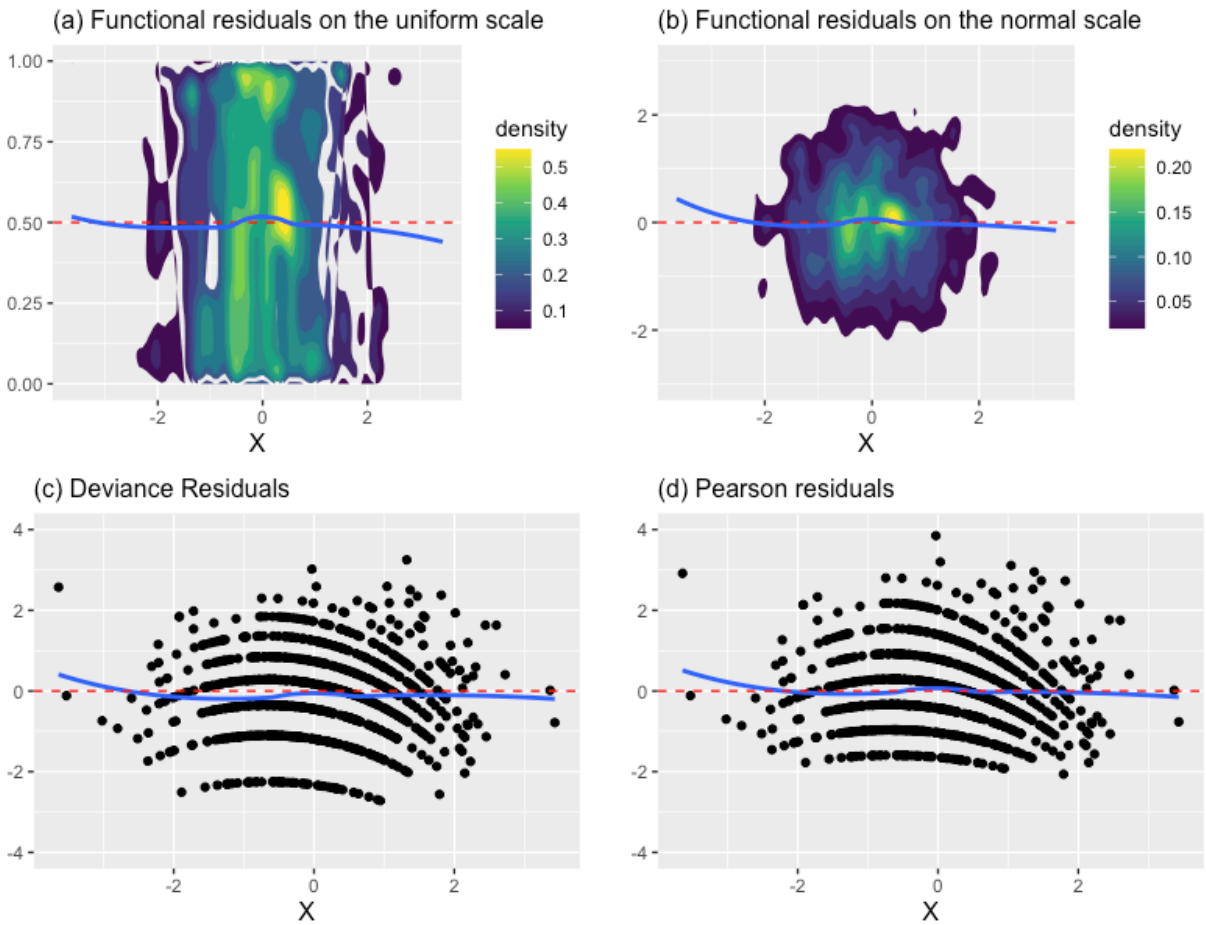

Figure S5: Plots of the proposed functional residual (upper row) and traditional deviance and Pearson residuals (lower row) against the covariate X when the working Poisson model is specified correctly for count data in the setting of Example 5.

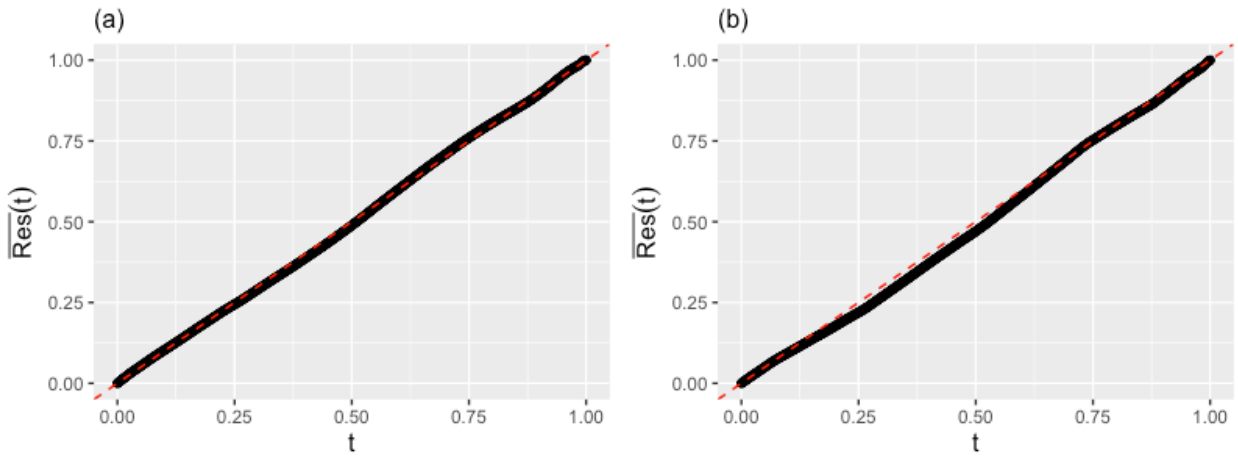

Figure S6: Proposed *Fn-Fn* plots when the working model is specified correctly for count data in the setting of Example 5. The left and right panels are obtained using the full sample (1000 observations) and the subsample $\{(y_i, x_i) \mid x_i < 0\}$, respectively.

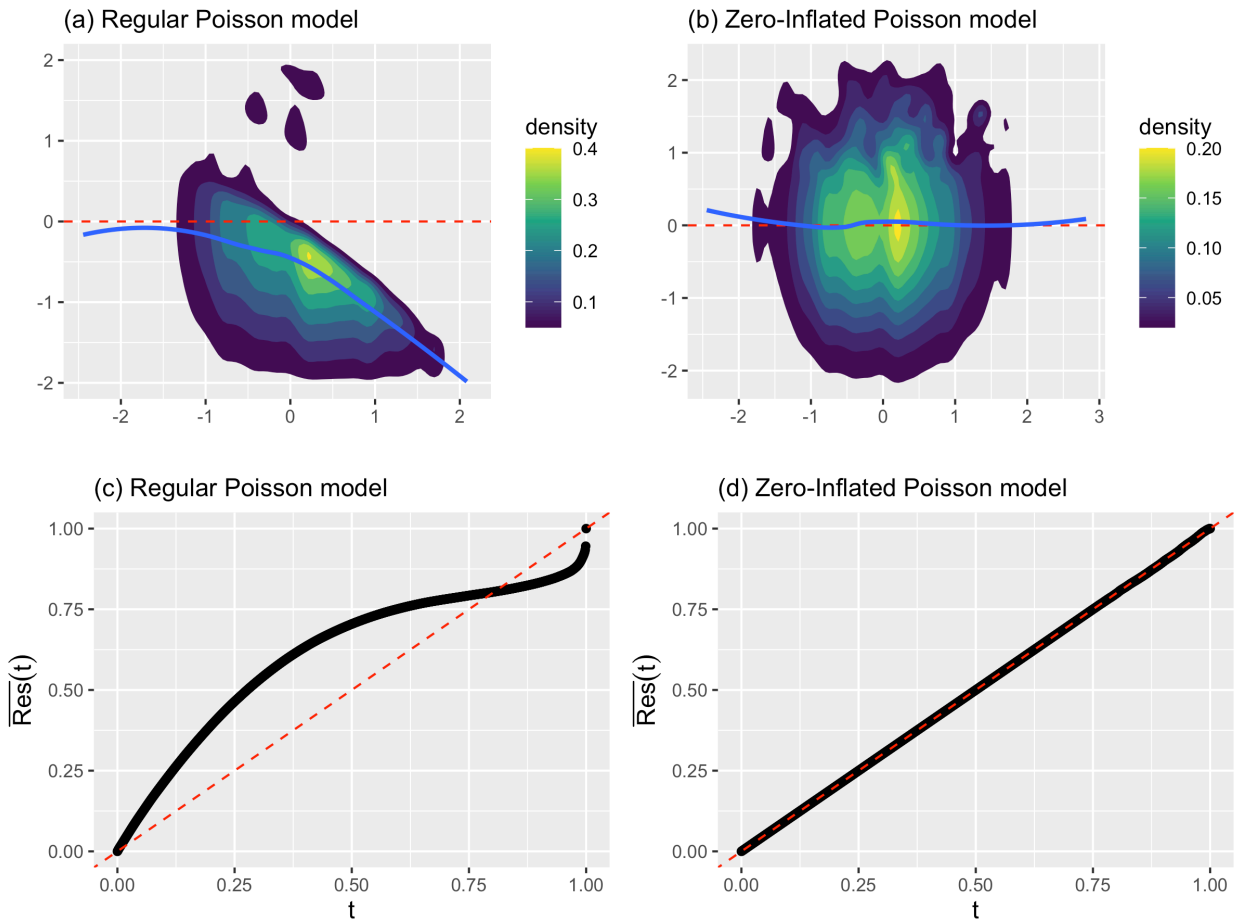

Figure S7: Functional-residual-*vs*-covariate plots (upper row) and *Fn*-*Fn* plots (lower row) in the presence of excessive zeros. Shown are the results for the regular Poisson model (left) and zero-inflated Poisson model (right) used in Example 6.

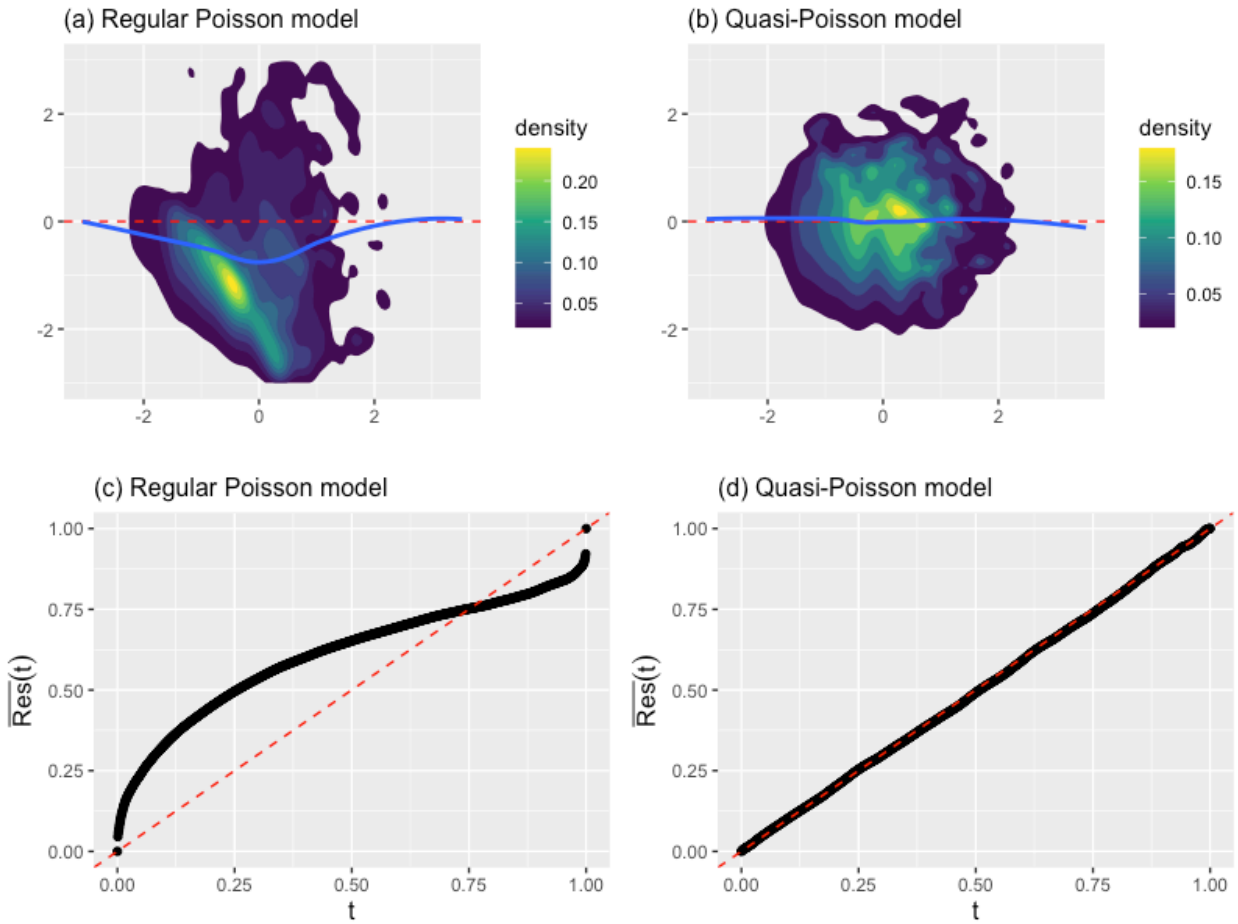

Figure S8: Functional-residual-*vs*-covariate plots (upper row) and *Fn*-*Fn* plots (lower row) in the presence of over-dispersion. Shown are the results for regular Poisson model (left) and Quasi-Poisson model (right) used in Example 7.

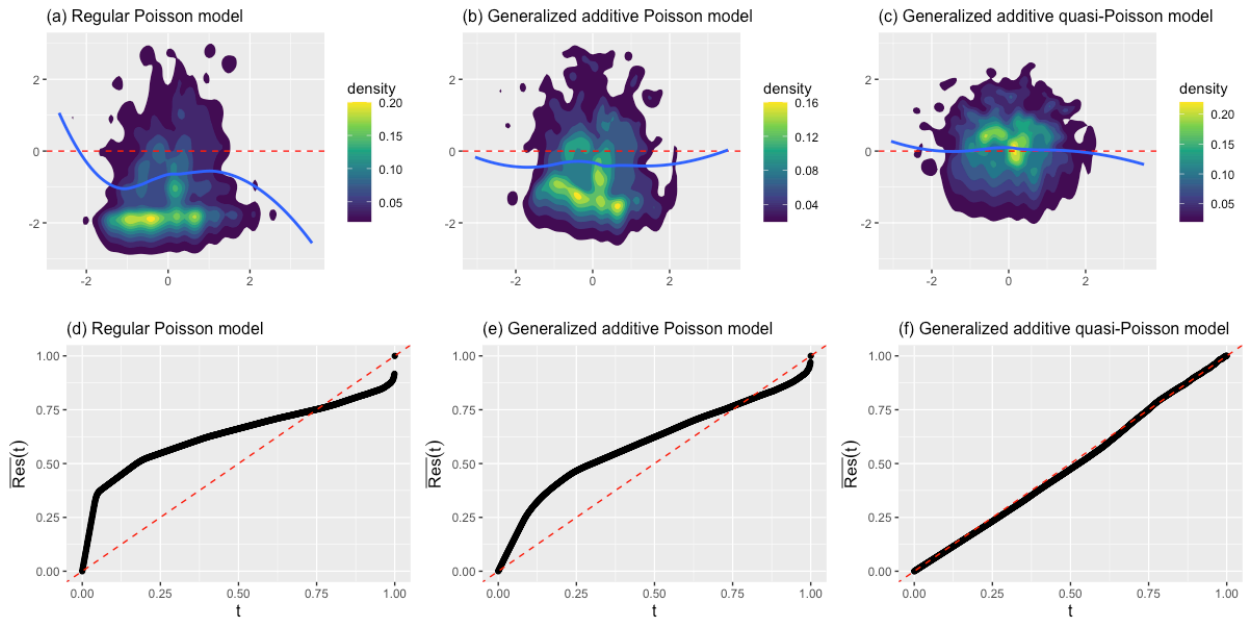

Figure S9: Functional-residual-vs-covariate plots ((a)-(c)) and *Fn-Fn* plots ((d)-(f)) for building a semiparametric model for count data with a sine effect of the predictor (Example 8).

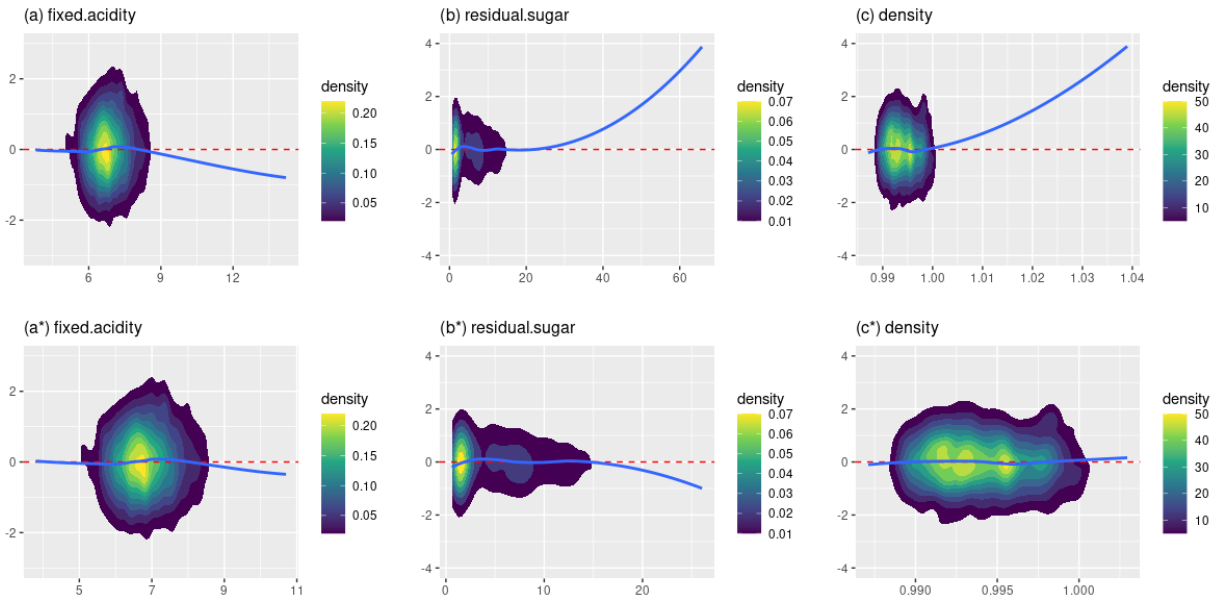

Figure S10: Functional-residual-*vs*-covariate plots before (upper row) and after (lower row) deleting the outliers from the wine quality dataset.

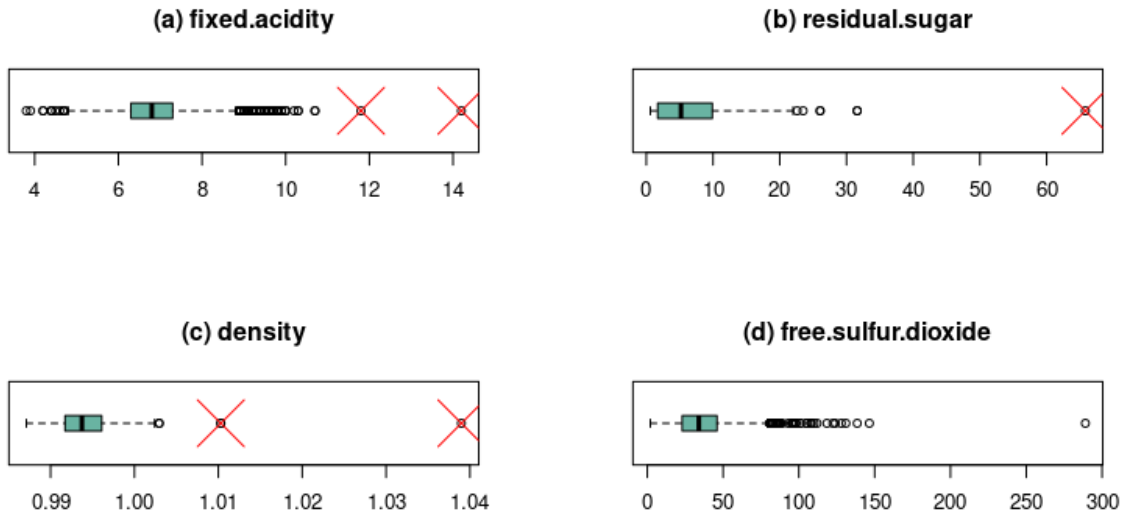

Figure S11: Boxplots of the variables that may contain outliers in the wine quality dataset.

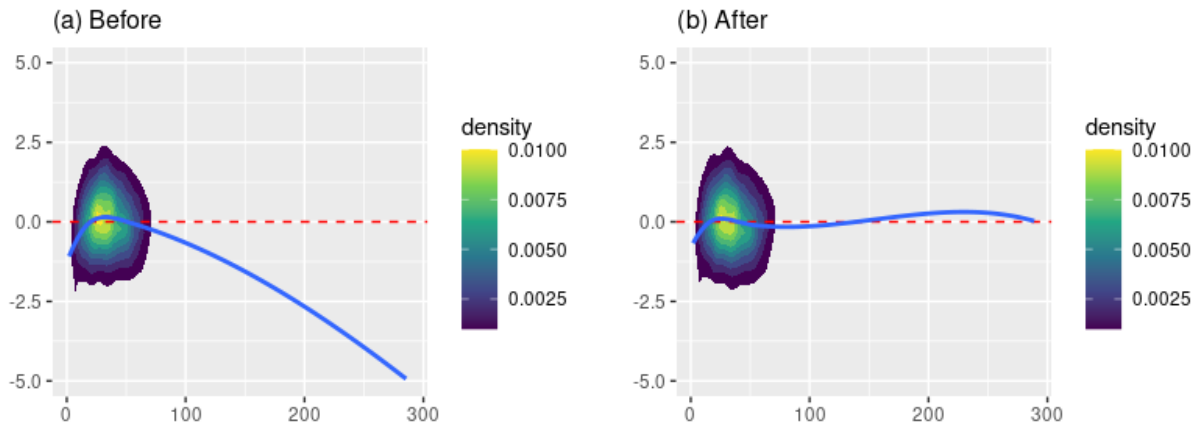

Figure S12: Plots of functional residuals versus *free.sulfur.dioxide* before and after adding its quadratic term to the model.

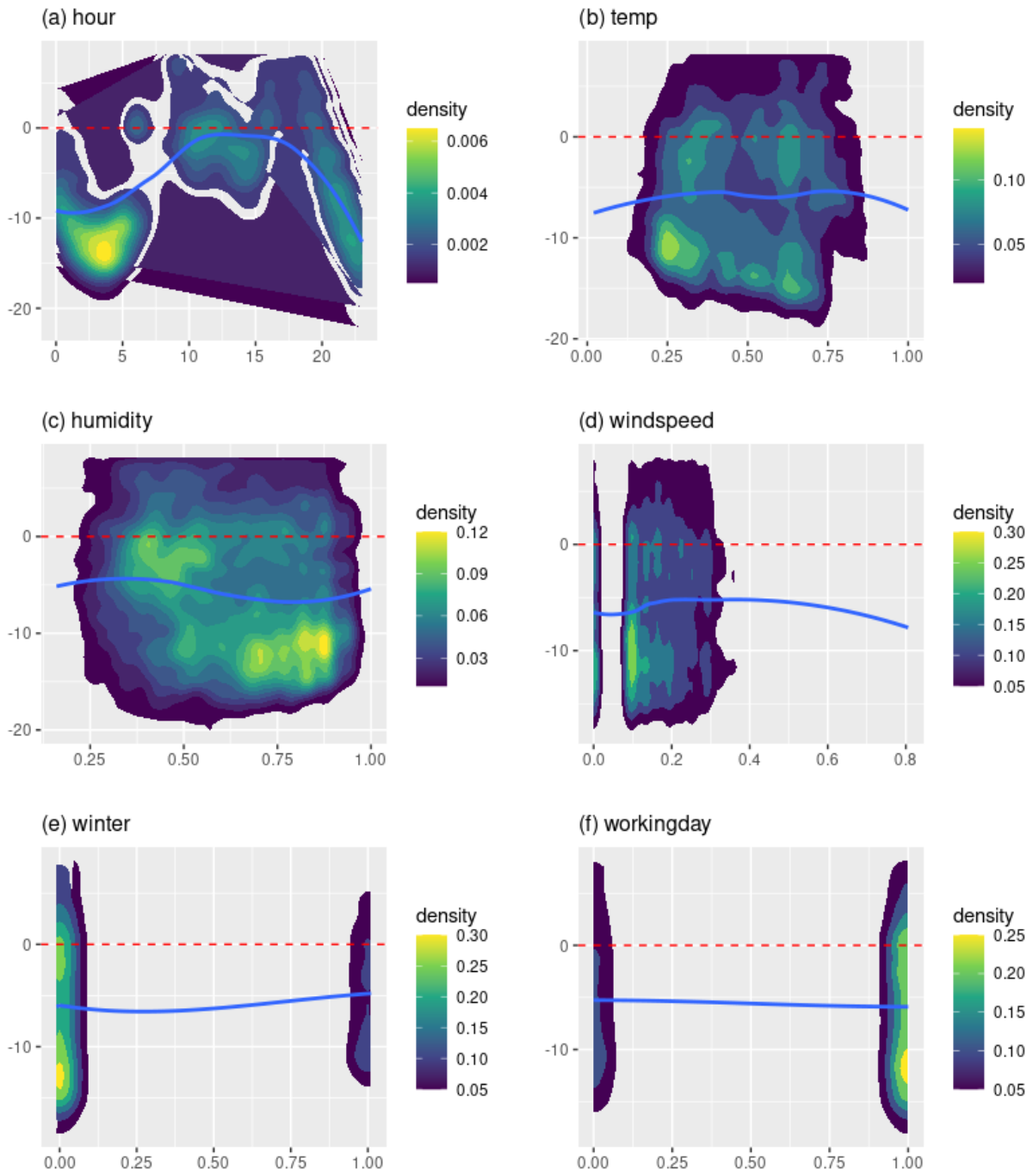

Figure S13: Functional-residual-*vs*-covariate plots for the initial Poisson model fitted to the Captial Bikeshare dataset.

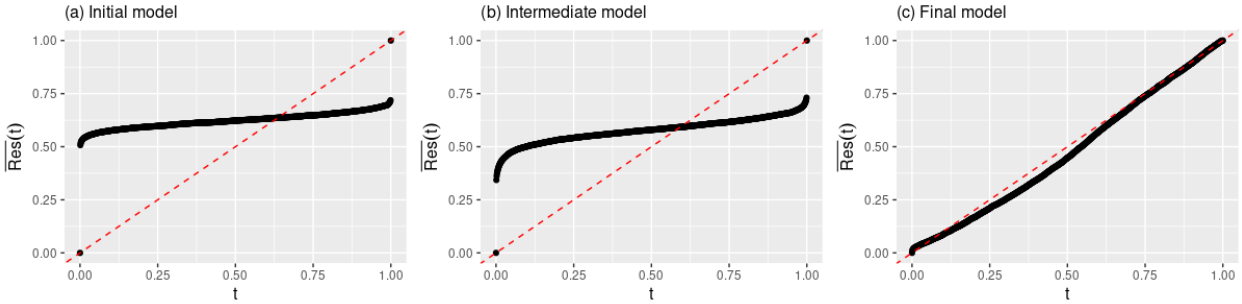

Figure S14: The *Fn-Fn* plots for the initial, intermediate, and final models developed in the model building process.

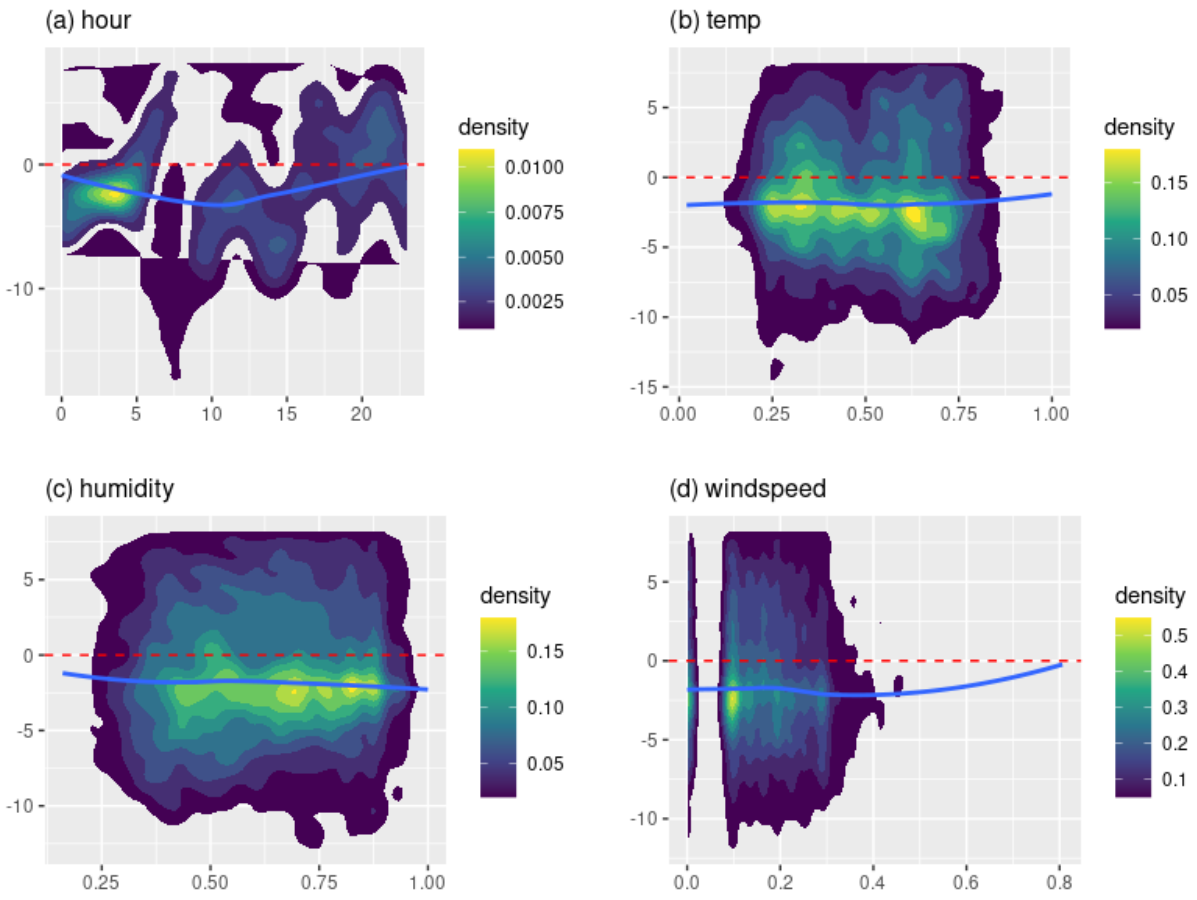

Figure S15: Functional-residual-*vs*-covariate plots after adding the smoothing functions of the variables *hour, temp, humidity,* and *windspeed* to the Poisson model.

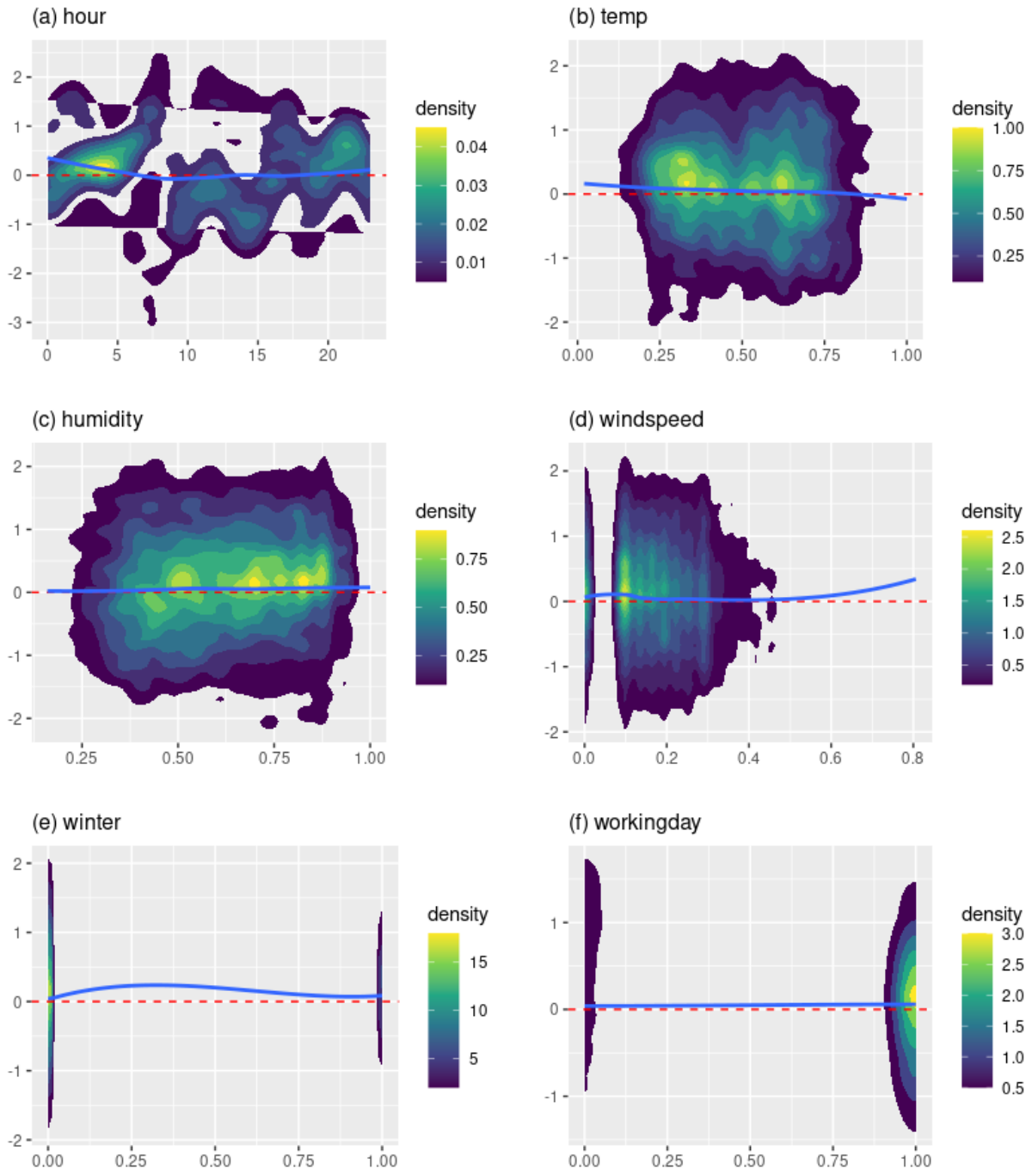

Figure S16: Functional-residual-*vs*-covariate plots for the final generalized additive quasi-Poisson model.

C Tables

Table S1: Summary of the initial and final models for the wine quality data.

Variable Name	Initial model			Final model		
	Estimate	Std.error	P-value	Estimate	Std.error	P-value
volatile.acidity	3.458	0.210	<0.001	3.440	0.216	<0.001
alcohol	-0.328	0.049	<0.001	-0.184	0.054	<0.001
sulphates	-1.121	0.178	<0.001	-1.362	0.184	<0.001
fixed.acidity	-0.136	0.039	<0.001	-0.246	0.043	<0.001
residual.sugar	-0.152	0.014	<0.001	-0.188	0.016	<0.001
free.sulfur.dioxide	-0.006	0.001	<0.001	-0.052	0.004	<0.001
pH	-1.295	0.193	<0.001	-1.647	0.206	<0.001
density	284.5	37.38	<0.001	417.4	42.13	<0.001
free.sulfur.dioxideˆ2	-	-	-	0.0005	<0.001	<0.001

Table S2: The explanatory variables in the bike sharing data.

Variable Name	Possible Values	Description
hour	{0, 1, ... , 23}	Time in 24-hour format.
temp	[0, 1]	The normalized temperature in Celsius, i.e., (t-tmin)/(tmax-tmin), tmin=-8, tmax=39.
humidity	[0, 1]	Normalized humidity. The values are derived via h/hmax , hmax =100.
windspeed	[0, 1]	Normalized wind speed. The values are derived via w/wmax, wmax=67.
winter	{0, 1}	1: winter; 0: otherwise.
workingday	{0, 1}	1: neither weekend nor holiday; 0: otherwise.
weather	{1, 2, 3, 4}	Weather situation: 1: clear, few clouds, partly cloudy. 2: mist, mist + few clouds, mist + broken clouds, mist + cloudy. 3: light snow, light Rain + scattered clouds, light Rain + thunderstorm + scattered clouds. 4: snow + fog, heavy rain + ice pallets + thunderstorm + mist.

Table S3: Summary of the initial and final models for the bike sharing data.

	Initial model			Final model		
Variable Name	Estimate	Std.error	P-value	Estimate	Std.error	P-value
Intercept	4.848	0.005	<0.001	5.202	0.017	<0.001
winter	-0.323	0.002	<0.001	-0.284	0.016	<0.001
workingday	0.061	0.002	<0.001	0.056	0.010	<0.001
weather	-0.025	0.001	<0.001	-0.139	0.010	<0.001
				Edf	Ref.df	P-value
hour	0.046	<0.001	<0.001	8.986	9.000	<0.001
temp	1.157	0.005	<0.001	4.613	5.632	<0.001
humidity	-1.005	0.005	<0.001	8.240	8.835	<0.001
windspeed	0.223	0.006	<0.001	3.582	4.409	<0.001
Dispersion Parameter		1.000			42.998	

D Additional examples for Section 4.2

Example S1 (Missing of a higher order term)**.** Suppose that the data are generated from the same Poisson model used in Example 5, whereas the quadratic term X^2 is not included in the working model. Our functional-residual-*vs*-covariate relationship is displayed in Figure S17(a). The U-shape of Lowess curve unequivocally indicates the missing of a quadratic term in the working model. It is worth noting that Figure S17(a) here is similar to Figure S1(b) where a quadratic form is missing in modeling ordinal data. We stress that this similarity across different types of data is not a coincidence but an appealing feature consistently seen throughout the following examples.

As a further demonstration, we simulate 1000 data points (x_i, y_i) from the model with a cubic term $log(E(Y|X)) = \alpha+\beta_1 X+\beta_2 X^2+\beta_3 X^3$, where $X \sim \mathcal{N}(0, 0.5)$ and $(\alpha, \beta_1, \beta_2, \beta_3) = (0.8, -0.2, 0.5, -0.5)$. When the X^3 term is not included in the working model, the LOWESS curve in Figure S17(b) shows a cubic pattern, which indicates the missing of an X^3 term in the model.

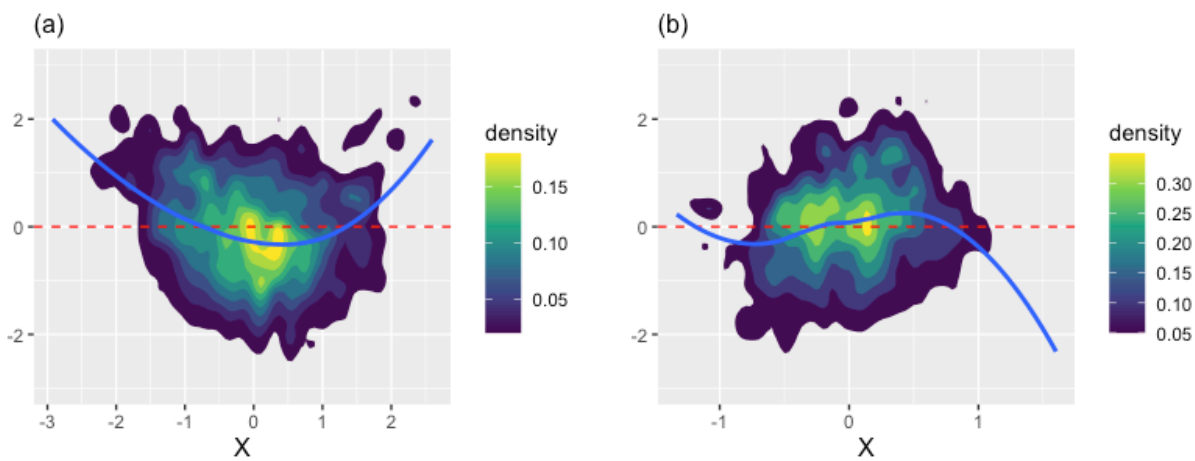

Figure S17: Functional-residual-*vs*-covariate plots when the quadratic term X^2 is missing in the working model (left) and cubic term X^3 is missing in the working model (right) for the count data (Example S1).

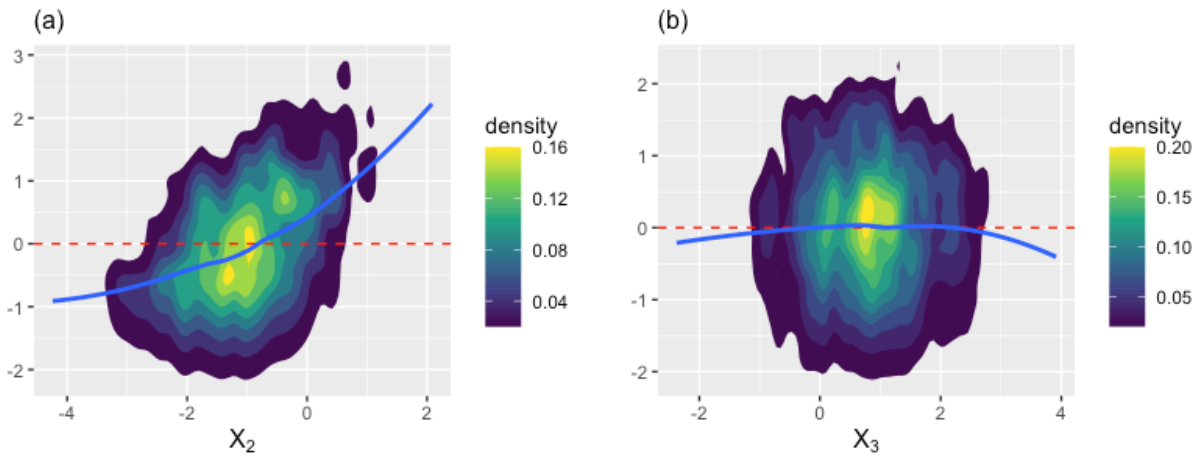

Figure S18: Functional-residual-*vs*-covariate plots when a covariate X_2 (left) is correlated with the count response Y whereas X_3 (right) is not in the setting of Example S2.

Example S2 (Missing of relevant covariates)**.** We simulate 1000 samples from a Poisson model with multiple covariates $log(E(Y|X)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, where $X_1 \sim \mathcal{N}(0, 0.8)$, $X_2 \sim \mathcal{N}(-1, 1)$, and $X_3 \sim \mathcal{N}(0.8, 0.9)$. We set the coefficients $(\alpha, \beta_1, \beta_2, \beta_3) = (0.5, 0.25, 0.5, 0)$ such that both X_1 and X_2 are correlated with the count outcome Y while X_3 is not related as $\beta_3 = 0$. We fit a Poisson model to the simulated data using X_1 only. We investigate the roles of X_2 and X_3, which are not modeled, through the lens of the functional-residual-*vs*-covariate plots presented in Figure S18. We observe in Figure S18(a) a positive relationship between our functional residuals and X_2, which indicates that X_2 may help explain a significant proportion of the residual variability. To the contrary, the symmetrically distributed "heat" in Figure S18(b) does not suggest any correlation between our functional residuals and X_3. This observation is in line with the way we set X_3 as independent of Y.

Example S3 (Missing of an interaction)**.** We simulate 1000 samples from a Poisson model with an interaction term $log(E(Y|X)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, where $X_1 \sim \mathcal{N}(0.5, 1)$, $X_2 \sim \mathcal{N}(-1, 0.7)$, and $(\alpha, \beta_1, \beta_2, \beta_3) = (-0.1, 0.8, -0.5, 0.6)$. We fit a Poisson model without the interaction term $X_1 X_2$. To see if our approach can detect the missing of interactions, we plot our functional residuals versus $X_1 X_2$ in Figure S19(a). The LOWESS curve exhibits a

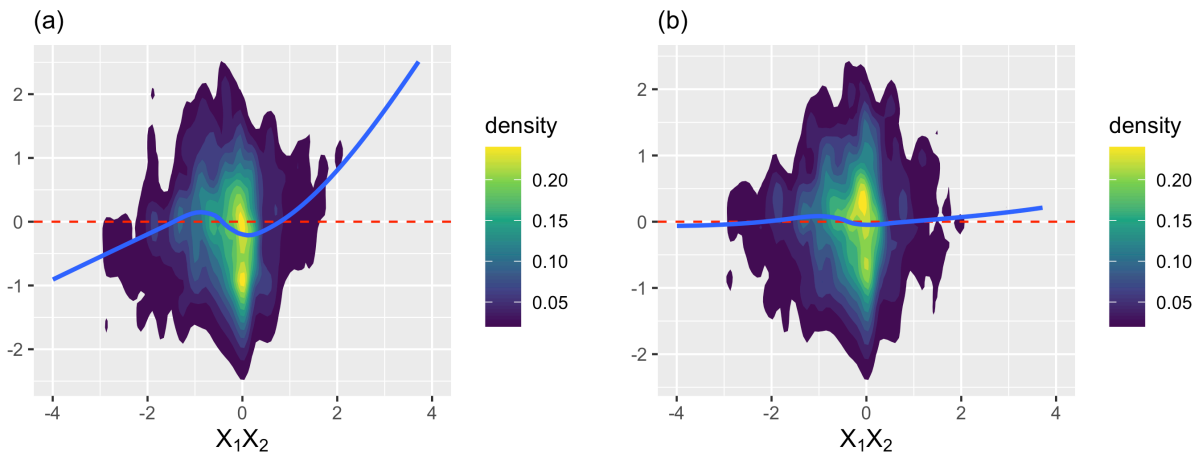

Figure S19: Functional-residual-*vs*-covariate plots before (left) and after (right) the interaction term X_1X_2 is included in the Poisson model for the count data (Example S3).

clear increasing pattern compared to the plot in Figure S19(b) where the model is specified correctly including the interaction term.

E Residual temporal effects in time series models

Our functional residuals apply to time series models. To see this, recall the general model form we have considered in the paper is

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

This general form subsumes time series models by allowing $\boldsymbol{X} = (\boldsymbol{X}_{\tilde{t}}, Y_{\tilde{t}-1}, \cdots, Y_{\tilde{t}-\tilde{p}})$ to include observations $Y_{\tilde{t}-1}, \cdots, Y_{\tilde{t}-\tilde{p}}$ at previous time points. Specifically, a time series model for discrete outcomes can be written as general as

$$Y_{\tilde{t}} \mid (\boldsymbol{X}_{\tilde{t}}, Y_{\tilde{t}-1}, \cdots, Y_{\tilde{t}-\tilde{p}}) \sim \pi(y_{\tilde{t}}; \boldsymbol{x}_{\tilde{t}}, y_{\tilde{t}-1}, \cdots, y_{\tilde{t}-\tilde{p}}, \boldsymbol{\beta}).$$

With a minimal change, the functional residual becomes

$$Res(h; y_{\tilde{t}}, \cdots, y_{\tilde{t}-\tilde{p}}, \boldsymbol{x}_{\tilde{t}}) = Pr\{U(\pi(y_{\tilde{t}} - 1; y_{\tilde{t}-1}, \cdots, y_{\tilde{t}-\tilde{p}}, \boldsymbol{x}_{\tilde{t}}), \pi(y_{\tilde{t}}; y_{\tilde{t}-1}, \cdots, y_{\tilde{t}-\tilde{p}}, \boldsymbol{x}_{\tilde{t}})) \leq h\}.$$

As the change is only symbolic (i.e., replacing $\boldsymbol{X}$ with $(\boldsymbol{X}_{\tilde{t}}, Y_{\tilde{t}-1}, \cdots, Y_{\tilde{t}-\tilde{p}})$), all the properties established in the paper remain valid.

Through the lens of our diagnostic tools, we are able to show that a time series model may not be sufficient to capture all temporal dependencies in the data. We use the same bike-sharing dataset analyzed in Section 5.3 to demonstrate this point. Since the outcome data are integer-valued hourly bike rentals, we use the integer-valued GARCH model, abbreviated as INGARCH($\tilde{p}, \tilde{q}$), where $\tilde{p}$ and $\tilde{q}$ are the autoregressive and the moving average orders (Ferland et al., 2006; Shumway and Stoffer, 2017). Based on the knowledge acquired in our analysis presented in Section 5.3, we use the negative binomial distribution, instead of Poisson distribution, to model the response. Specifically, $Y_{\tilde{t}} \sim NB(\lambda_{\tilde{t}}, \phi)$ where $\lambda_{\tilde{t}} = E(Y_{\tilde{t}} \mid (\boldsymbol{X}_{\tilde{t}}, Y_{\tilde{t}-1}, \cdots, Y_{\tilde{t}-\tilde{p}}))$ is the conditional mean and ϕ the dispersion parameter.

We start with the simplest INGARCH(1,0) model

$$log(\lambda_{\tilde{t}}) = log(E(Y_{\tilde{t}} | \boldsymbol{X}_{\tilde{t}}, Y_{\tilde{t}-1})) = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{X}_{\tilde{t}} + \beta_2 \, log(Y_{\tilde{t}-1} + 1),$$

which uses one past value $Y_{\tilde{t}-1}$ as a predictor. To see if this time series model has adequately captured all temporal dependencies, we plot the functional residuals versus the variable *hour* in Figure S20 (a). We observe that the functional residuals all fall blow the center horizontal line in the time interval 0-5 AM, while the LOWESS curve is well above the center horizontal line during 8AM-5PM. After 8PM, the LOWESS curve decreases rapidly until midnight. These patterns in the functional residuals provide compelling evidence that the INGARCH(1,0) model is not able to capture all temporal trends in the data. This model inadequacy issue cannot be solved by increasing the order of autoregressive and moving average terms. For example, we refit the data using INGARCH(2,0), INGARCH(3,0), and INGARCH(3,3) models. Similar residual patterns remain in the functional-residual-*vs*-hour plots displayed in Figures S20(b)-(d). The implication is that hourly bike rentals cannot be simply explained or predicted by data in the past hours in a day. A closer look at the plots in Figure S20 reveals that the distributions of hourly rentals at different times of a day (e.g., 0-5AM versus 8AM-5PM) are intrinsically different. This heterogeneity in distributions (conditional on various times) has been discussed in Section 5.2; see the paragraph entitled "Additional insights into the final semiparametric model".

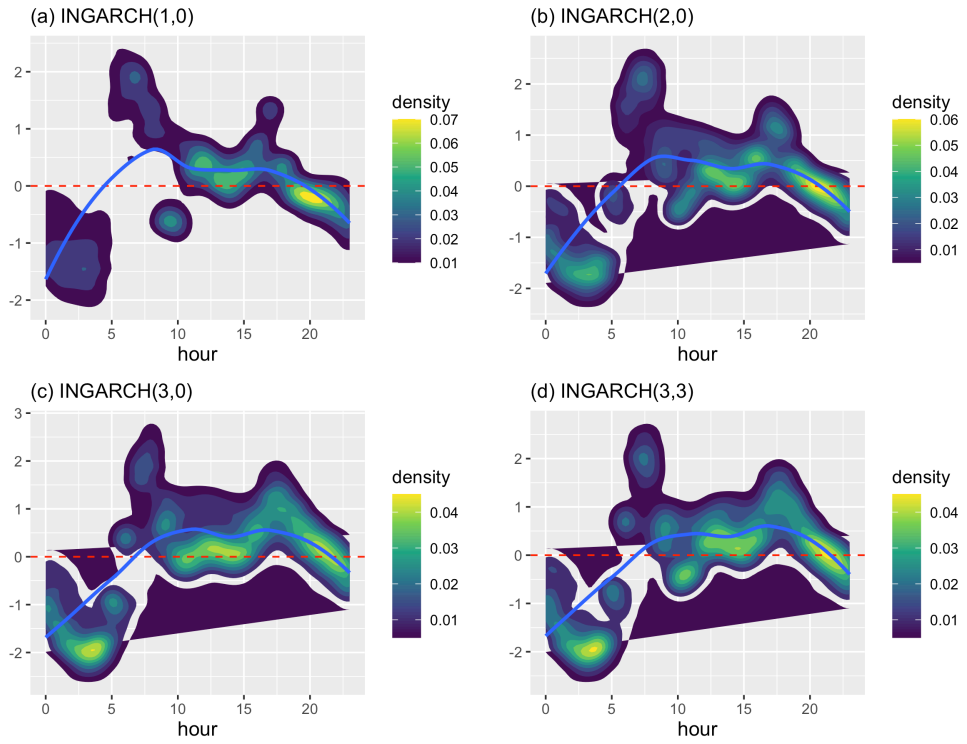

Figure S20: Functional-residual-*vs*-hour plots using different INGARCH models in the analysis of the bike sharing dataset in Section 5.3.

F Comparison with other diagnostic methods

We provide a more comparison with Liu and Zhang (2018)'s surrogate residual, Li and Shepherd (2012)'s probability-scale residual, and Yang (2021)'s quasi-empirical residual distribution function. Theoretical merits and limitations of each method are summarized below with numerical support using Examples 3-4.

- The surrogate residual proposed in Liu and Zhang (2018) focused on parametric models with a latent continues response. Used in Example 1-4 are, however, adjacent-category logit models, which are derived from the odds relating to response categories instead of using cumulative probabilities of a latent variable (Liu and Agresti, 2005; Agresti, 2010). Figures S21(c)-(d) and Figures S22(c)-(d) cast surrogate residuals versus covariates by fitting an ordinal logistic model (i.e., proportional odds model) while the true model is an adjacent-category logit model. While the plots suggest trends similar to those observed with functional residuals, caution is advised because surrogate residuals depend on simulation. As a result, a residual plot may vary from one simulation to another. When the sample size is not large, the pattern robustness needs to be checked.

- As elaborated in Section 6, the value of the probability-scale residual lies in its first-moment property, i.e., $E(R_{PS} \mid X) = 0$ when the model is specified correctly. Our numerical experience suggests that it is effective in capturing misspecifications in the mean structure of the fitted model; see Figure S21(e) and Figure S22(e) as examples where the omissions of X_2 and X_1X_2 are captured respectively. However, the visual examination of the probability-scale residual should be limited to its conditional mean only. So far, there is no theoretical ground to interpret any trends or patterns unrelated to its mean. For example, the patterns observed in Figure S21(e) (e.g., varying variances and distributions across different values of X) should be ignored. Since we

cannot interpret its distributional information beyond the mean, QQ plots are not feasible (see examples in Liu and Zhang, 2018).

- Yang (2021)'s quasi-empirical residual distribution function converges to the identity function if the model is specified correctly. This result leads to a QQ-like plot, which provides an overall assessment of the fitted model (see Figure S21(g) and Figures S22(g)-(h) for example). The plot is a useful tool to gauge practical significance of model misspecification, if it exists (see (A1) in the introduction). On the other hand, as discussed in Yang (2021), there is no justification yet to use it to inspect the relationship between residuals and covariates, which may limit its utility in (i) detecting model misspecifications – Figure S21(g) shows an adequate fit, whereas other residual-versus-covariate plots indicate the opposite; or (ii) revealing the cause for model misspecifications – it is not obvious to tell that the cause for the upper-tail pattern in Figures S22(g) is the omission of an interaction term.

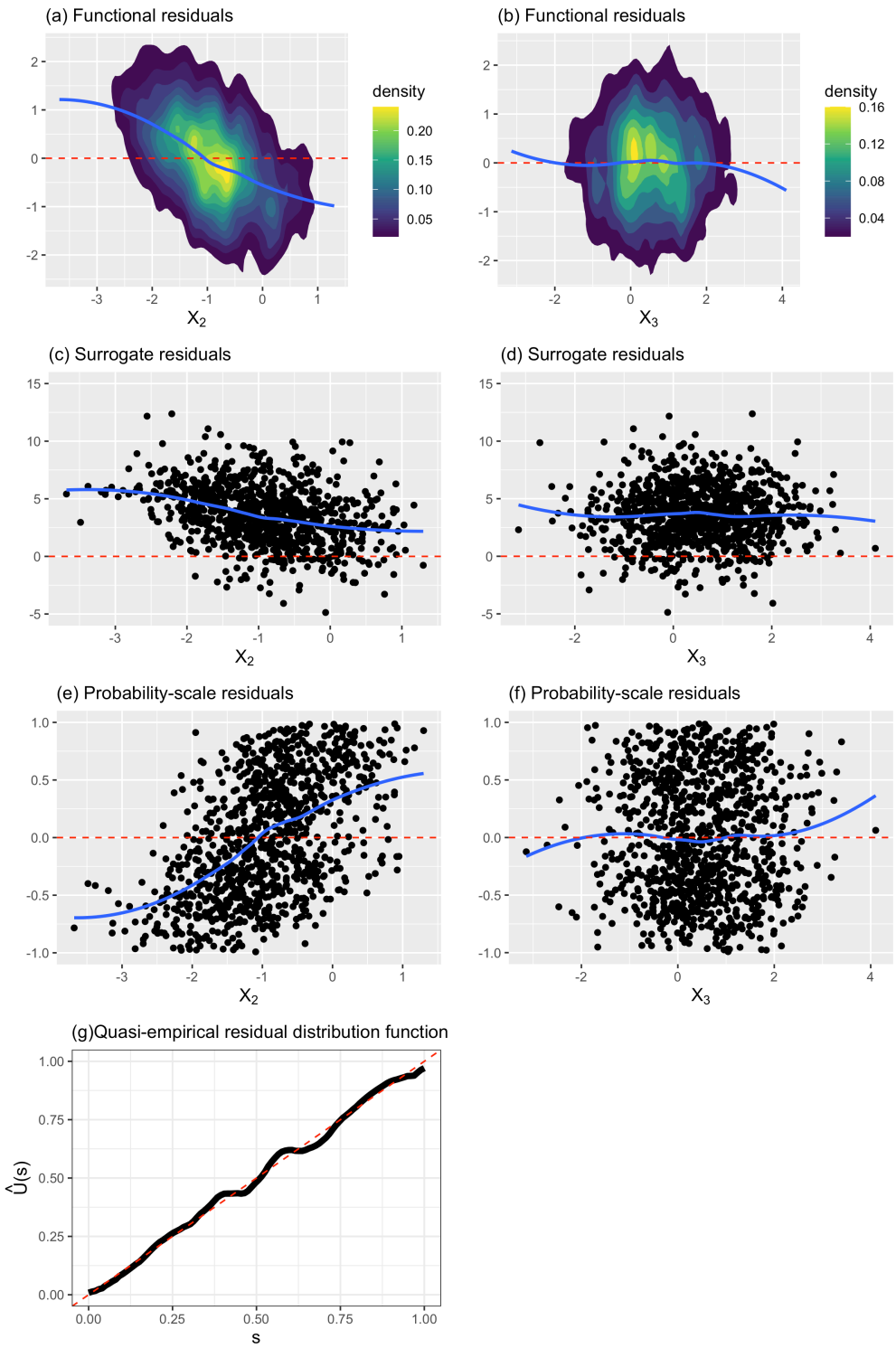

Figure S21: Residual-*vs*-covariate plots (a)-(f) when X_2 is correlated with ordinal data Y (the left panel) whereas X_3 is not (the right panel) in the setting of Example 3. Plotted in (g) is an overall assessment of the fitted model that does not include X_2 or X_3.

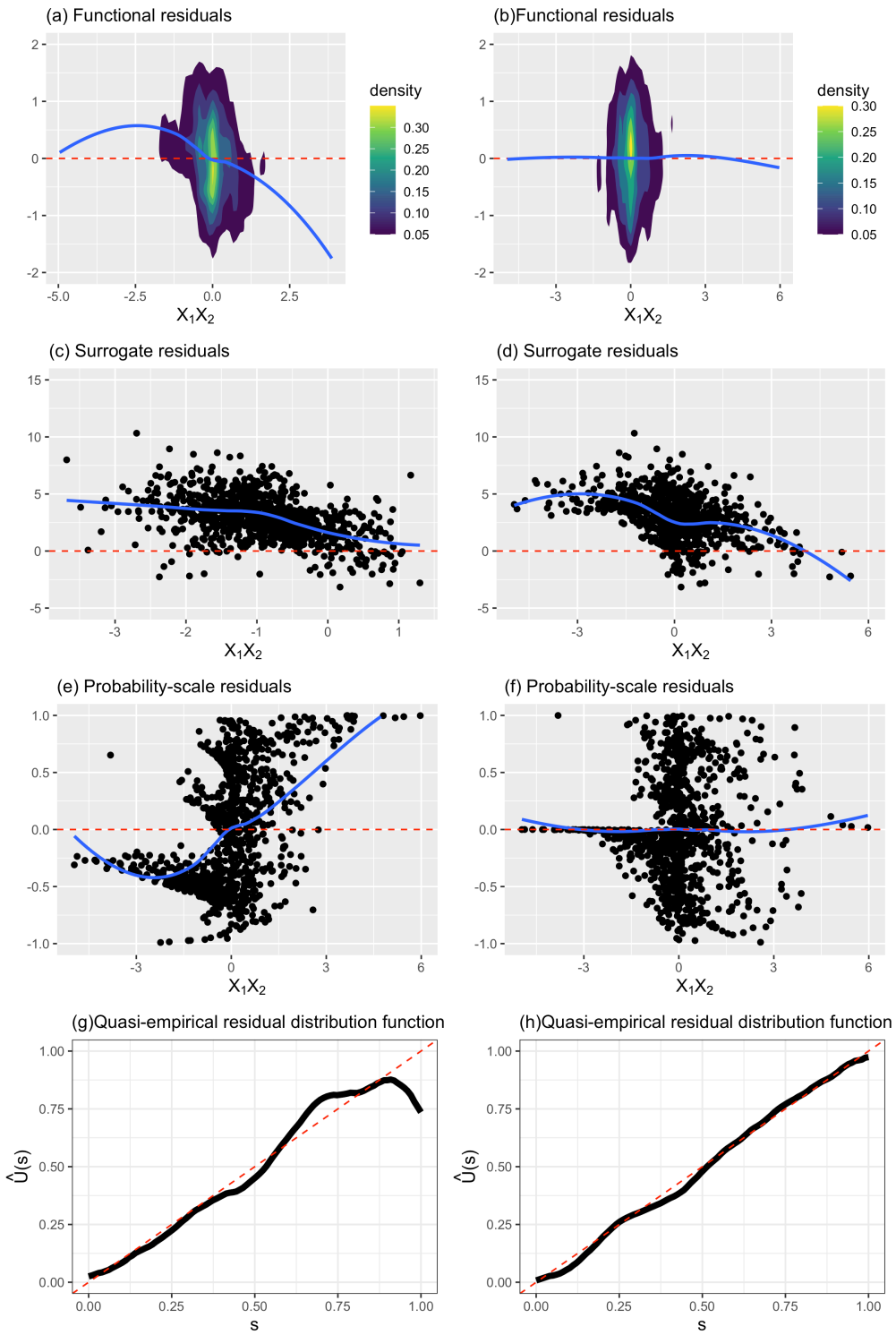

Figure S22: Functional-residual-*vs*-covariate plots and existing diagnostic plots before (left) and after (right) the interaction term X_1X_2 is included in the working adjacent-category logit model (Example 4).

G Graphical test

The functional residual can be used to develop goodness-of-fit tests. As our focus is not on a dichotomous decision but insights into practical significance as well as model diagnosis, we adopt the graphical test approach (Davison and Hinkley, 1997, pp.150-154). This approach is well suited for our purposes because it embeds hypothesis testing in graphical inspections.

The idea of the graphical test is to generate a "probable envelop" for comparison against the original plot. In our specific context, the original plot is an F_n-F_n plot in which the curve $(t, \overline{Res}(t))$ is displayed alongside the reference 45-degree line. For any given t in the interval $(0,1)$, $\overline{Res}(t) = \sum_{i=1}^{n} Res(t; y_i, \boldsymbol{x}_i)/n$ can be viewed as a test statistic. Under the null hypothesis, it converges to t (see Corollary 1) and its distribution can be approximated by resampling the data using the bootstrap method. Given an error rate α, a two-sided bootstrap test yields the lower and upper critical values $L^*_{\alpha/2}(t)$ and $U^*_{1-\alpha/2}(t)$. Varying the value of t, the set $\mathcal{E}_{1-\alpha} = \{(L^*_{\alpha/2}(t), U^*_{1-\alpha/2}(t)), t \in (0,1)\}$ forms a test envelop. Excursion of the 45-degree line outside $\mathcal{E}_{1-\alpha}$ may be regarded as evidence against the hypothesis that the model is correct. This simultaneous comparison across t is called a graphical test.

Below we describe how to compute the test envelop $\mathcal{E}_{1-\alpha}$ using the bootstrap method.

Algorithm 1 Computing the lower and upper critical values $L^*_{\alpha/2}(t)$ and $U^*_{1-\alpha/2}(t)$ of $\mathcal{E}_{1-\alpha}$, given a size-n sample $(\boldsymbol{y}, \boldsymbol{x})$ and a confidence level $100(1-\alpha)\%$.

Step 1: To obtain the b-th bootstrap copy $\overline{Res}^*_{(b)}(t)$, we

(1) resample a size-n sample $(\boldsymbol{y}^*_{(b)}, \boldsymbol{x}^*_{(b)})$ from the given sample $(\boldsymbol{y}, \boldsymbol{x})$;

(2) refit the same model to the bootstrap sample $(\boldsymbol{y}^*_{(b)}, \boldsymbol{x}^*_{(b)})$ and obtain a bootstrap estimate $\hat{\boldsymbol{\beta}}^*_{(b)}$ of the regression coefficient $\boldsymbol{\beta}$;

(3) obtain the bootstrap functional residual $\overline{Res}^*_{(b)}(t) \equiv Res^*_{(b)}(t; \boldsymbol{y}^*_{(b)}, \boldsymbol{x}^*_{(b)})$;

End the For-loop.

Step 2: Set $L^*_{\alpha/2}(t) = q_{\alpha/2}(t)$ and $U^*_{1-\alpha/2}(t) = q_{1-\alpha/2}(t)$, where $q_\nu(t)$ is the (ν)-th quantile of the bootstrap copies $\{\overline{Res}^*_{(1)}(t), \ldots, \overline{Res}^*_{(B)}(t)\}$.

In Figure S23, we illustrate the test envelop $\mathcal{E}_{1-\alpha}$ in the presence of overdispersion and zero-inflation in the analysis of count data. We use the same settings as those in Examples 6-7 presented in the main paper. When the count data are overdispersed but fitted with a regular Poisson model where the overdispersion parameter is set to 1, Figure S23(a) shows that the F_n-F_n curve deviates remarkably from the 45-degree line. The surrounding dotted curves form a test envelop with the type I error rate α being 0.05. The envelop is thin, and it does not cover the 45-degree line except for a very small section around $t = 0.75$. Moreover, there is a notable gap between the envelop and the 45-degree line when $0 < t < 0.75$. These observations can be taken as strong evidence indicating significance of model misspecification in *both* statistical and practical terms. When the dispersion parameter is allowed be estimated from the data and a quasi-Poisson model is fitted instead, Figure S23(b) shows that the 45-degree line is well contained in the test envelop, indicating a nice fit of the model to the data. Similar observations can be said to the lower row of Figure S23, where the number of zeros is much inflated. In the presence of zero-inflation, the thin and skewed test envelop in Figure S23(c) is a clear indication of the lack of fit of the regular Poisson model. When the zero inflation is taken into account, the test envelop in Figure S23(d) shrinks to the 45-degree line, which indicates an almost perfect alignment with the benchmark.

The test envelop $\mathcal{E}_{1-\alpha}$ illustrated in Figure S23 is a result of multiple comparisons. Specifically, for each $t = 0.02, 0.04, \ldots, 0.98$, we conduct a test and obtain a confidence interval. The 49 intervals altogether form an envelop. If a single binary decision about "rejecting or not" is really wanted, the significance threshold needs to be adjusted to control the type I error rate. Instead of using Bonferroni correction which is known to be too conservative, we use the Benjamini-Hochberg procedure to control the false discovery rate. In our context, a "false discovery" means rejecting the model while it is correct. Note that the multiple comparisons entailed in the test envelop examine the same hypothesis and aims for the same "discovery" which is the model being misspecified. Our numerical studies show that the Benjamini-Hochberg procedure reasonably controls the overall false rejection rate.

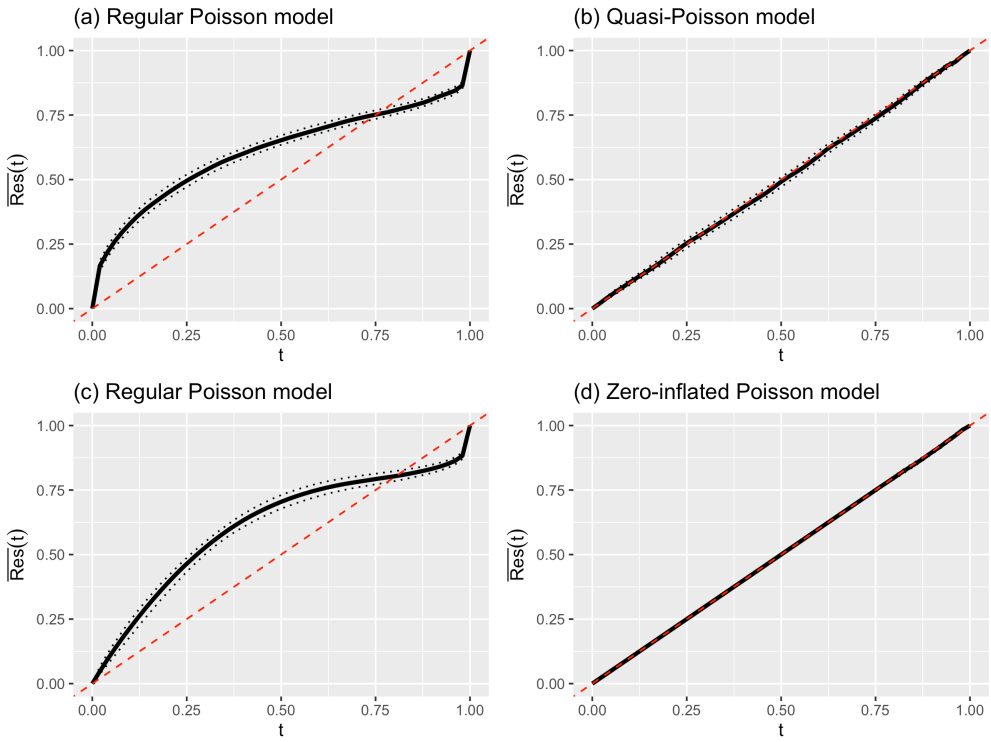

Figure S23: F_n-F_n plots with test envelops, depicted by dotted curves surrounding the solid curve, in the presence of overdispersion (the upper row) and zero-inflation (the lower row).

In Table S4, we examine the type I error and the power of our test procedure in comparison with the classical Pearson's χ^2 test. Note that Pearson's χ^2 test directly measures the discrepancy between the observed and fitted count values, and therefore it may be viewed as a "gold standard" for testing the goodness of fit for count data. Table S4 contains the simulation result under the settings of Examples 6-7, the same settings used to illustrate our graphical test as in Figure S23. In the presence of overdispersion, Panel A shows that Pearson's χ^2 test completely fails since its type I error rate is 100%. In other words, Pearson's χ^2 test is not valid without the knowledge of the dispersion parameter (Faraway, 2016, pp.158). In contrast, our method effectively controls the type I error rate while being able to detect misspecification 100% of the time, even with a sample size as small as 200. In the presence of zero inflation, Panel B shows that our method and Pearson's χ^2 test perform similarly under the setting of Example 6. Both well control the type I error rate for $n = 200, \ldots, 500$, and neither misses the detection of misspecification even with a sample size as small as 200. To better uncover any difference, we reduce the severity of zero inflation. When zeros take a proportion of 16% $\sim$ 34.5%, Panel C shows that Pearson's χ^2 test has notably higher power when $n = 200$. The power of our method catches up quickly when the sample size increases.

To summarize, our method offers distinct advantages over the classical Pearson's χ^2 test: (i) it reveals the practical significance of model deviations through the lens of a 'probable envelope' generated by the graphical test, and (ii) it provides a goodness-of-fit test applicable to general situations, including those where the dispersion parameter is unknown. On the other hand, if the inference is solely on statistical significance, our method may not be as powerful as Pearson's χ^2 test where applicable. This is not surprising given the trade-off between general applicability and higher power in specific situations.

Table S4: Type I error and power: a comparison of our test versus Pearson's χ^2 test.

		n=200		n=300		n=400		n=500	
	α	The null overdispersed model							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
Panel A Overdispersion: Example 7 ($\theta = 7$)	0.05	1	0.013	NV*	0.015	NV	0.033	NV	0.022
	0.10	1	0.026	NV	0.030	NV	0.048	NV	0.044
		Missing of modeling dispersion							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
	0.05	NV	0.970	NV	0.997	NV	1	NV	1
	0.10	NV	0.989	NV	0.998	NV	1	NV	1
		The null zero-inflated model							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
Panel B Zero-inflation: Example 6	0.05	0.021	0.038	0.020	0.037	0.019	0.033	0.026	0.039
	0.10	0.051	0.064	0.042	0.077	0.048	0.066	0.047	0.066
		Missing of a zero-inflated component							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
	0.05	1	1	1	1	1	1	1	1
	0.10	1	1	1	1	1	1	1	1
		The null zero-inflated model							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
Panel C Zero-inflation: Example 6 with Reduced Zeros	0.05	0.031	0.034	0.034	0.049	0.034	0.036	0.031	0.044
	0.10	0.095	0.057	0.065	0.072	0.076	0.058	0.088	0.084
		Missing of a zero-inflated component							
		χ^2	Ours	χ^2	Ours	χ^2	Ours	χ^2	Ours
	0.05	0.823	0.576	0.904	0.843	0.976	0.958	0.992	0.983
	0.10	0.890	0.662	0.948	0.887	0.985	0.972	0.996	0.987

H Robustness of residual patterns with reduced sample sizes

In our two case studies, the sample sizes are not small (4898 and 8734). To examine the effectiveness of our approach with a small sample size, we randomly select a subset of 250 observations out of each of the two data sets. We repeat the same analysis and find that our approach is still able to identify key misspecifications. For example, in the initial adjacent-categorical logit model for the wine quality analysis, Figure S24(a) captures the missing of the quadratic term of *free.sulfur.dioxide*. In the initial Poisson model for bike sharing analysis, Figure S24(b) exhibits a residual fluctuation across the 24 hours, which indicates a cyclical effect of the factor *hour*. In the updated semiparametric additive Poisson model, Figure S24(c) shows a considerably stretched *Fn-Fn* curve, an indication of extra large variability. In the final semiparametric additive quasi-Poisson model, Figure S24(d) reflects the heterogeneity in residual distributions across the 24 hours, which offers deeper insights into the an already-complex count regression model. All of these issues have been discussed in the original analysis, and they can still be uncovered in our analysis of the subsets that are only a small fraction (5.1% and 2.9%) of the original data sets.

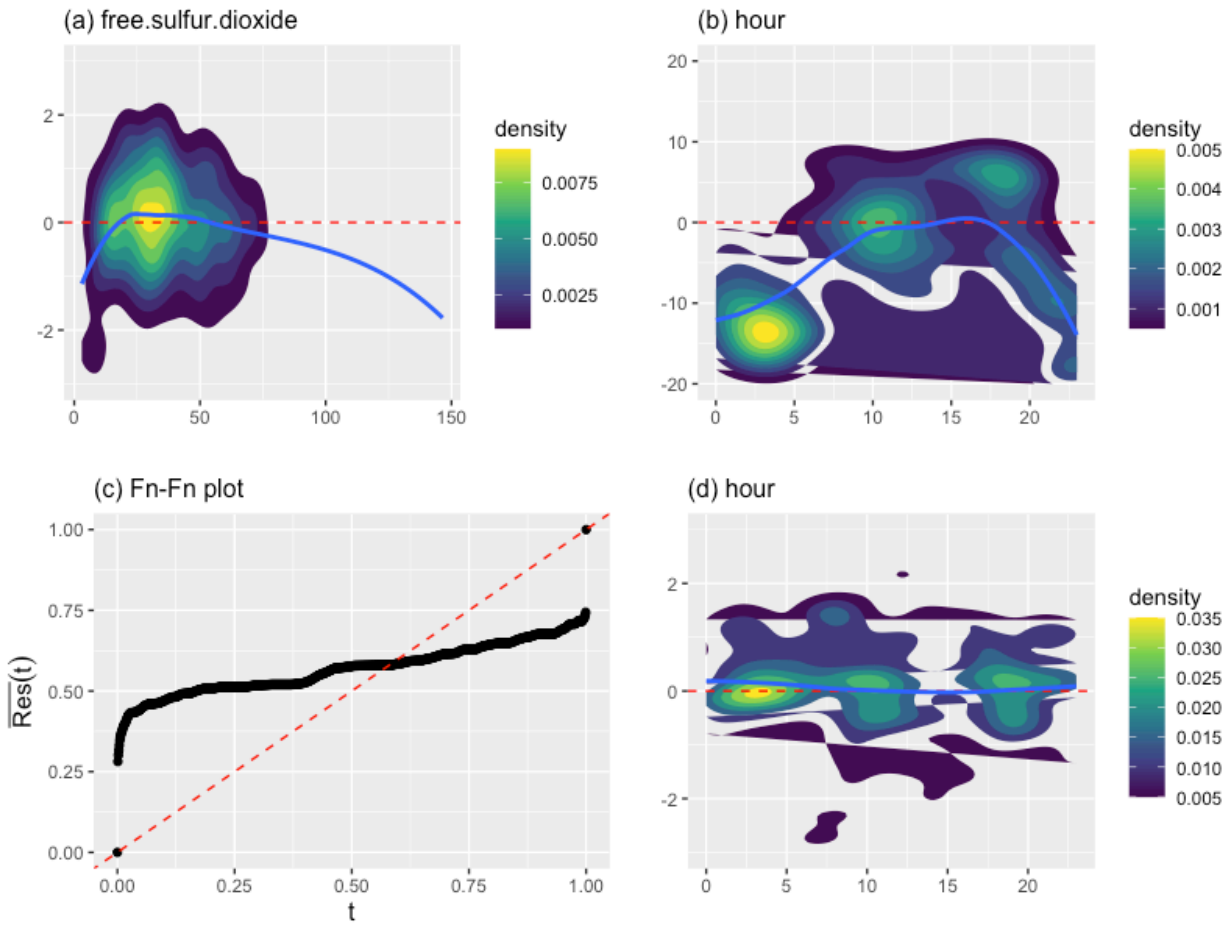

Figure S24: Model diagnostics for the two case studies when a much smaller subset ($n = 250$) is randomly selected from each of the original data sets. The same analysis is repeated following the procedures presented in Section 6.

References

Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, John Wiley & Sons: Hoboken, New Jersey, 2nd ed.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press: NY, USA.

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

Ferland, R., Latour, A., and Oraichi, D. (2006), "Integer-valued GARCH process," *Journal of Time Series Analysis*, 27, 923–942.

Li, C. and Shepherd, B. (2012), "A new residual for ordinal outcomes," *Biometrika*, 99, 473–480.

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, I. and Agresti, A. (2005), "The analysis of ordered categorical data: An overview and a survey of recent developments (with discussion)," *Test*, 14, 1–73.

Shumway, R. H. and Stoffer, D. S. (2017), *Time Series Analysis and Its Applications : with R Examples*, Cham, Switzerland: Springer.

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.