Diagnostics for Regression Models with Discrete Outcomes Using Surrogate Empirical Residual Distribution Functions

Lu Yang
Amsterdam School of Economics
University of Amsterdam
Email: l.yang@uva.nl
February 21, 2019

Abstract

Making informed decisions about model adequacy has been an outstanding issue for regression models with discrete outcomes. Standard residuals for such outcomes show a large discrepancy from the hypothesized pattern even under the true model and are often not informative especially when data are highly discrete. To fill this gap, we propose a surrogate empirical residual distribution function for general discrete (e.g. ordinal and count) outcomes that serves as an alternative to the empirical Cox-Snell residual distribution function. The diagnostic tool we propose is a principled approach and does not require injecting noise to the data. When at least one continuous covariate is available, we show asymptotically that the proposed function converges uniformly to the identity function under the correctly specified model, even with highly discrete (e.g. binary) outcomes. Through simulation studies, we demonstrate empirically that the proposed surrogate empirical residual distribution function outperforms other commonly used residuals for various diagnostic tasks, since it is close to the hypothesized pattern under the true model and significantly departs from this pattern with model misspecification, and is thus an effective diagnostic tool.

Keywords: Cox-Snell residuals; Goodness-of-fit; Generalized linear models; Insurance claim frequency; $m$-asymptotics; Empirical process.
1. INTRODUCTION

Regression models summarize researchers’ and analysts’ knowledge about relationships between covariates and an outcome of interest. For example, in auto insurance applications, when modeling the number of claims from each policyholder, Poisson distributions are commonly adopted by actuaries, and a set of typical covariates including drivers’ age and car model is usually used to predict the number of claims. However, researchers’ prior knowledge may not sufficiently capture the patterns in the data. Given the potential pernicious consequences of model misspecification, in many fields, analysts are routinely tasked with demonstrating that their models have sufficiently characterized all pertinent features of the data. Hence, it is of prime importance to check the adequacy of the model fit and if necessary, further refine the model. To illustrate, a quote from George E. P. Box (DeGroot 1987) relates to this point:

When you look at residuals or you look at discrepancies between what you thought and what happened, that’s the part where you say, “Aha, it’s clear that what I was thinking was wrong, and it’s wrong in the following kinds of ways; which suggest to me (or suggest to this guy I’m working with, who’s an engineer and understands these things) that perhaps this is what’s happening.”

And so we have to go and do something else, and perhaps run experiments on variables that we have never even considered before. So in a sense this is the only part where something really new is created.

Towards the aim of model diagnostics and refinement, there are two main streams of literature on evaluating goodness-of-fit for regression models. The first one is on model comparison by conducting formal tests among nested (e.g. likelihood ratio tests) or non-nested (e.g. Vuong’s tests as in Vuong (1989)) models. The second type of approach focuses on diagnostics for a specific model at hand and does not require alternative models, which is the focus of this paper. Residuals are a central component of this second class of diagnos-
tic methods. One can use graphical techniques (Ben and Yohai 2004) or construct overall goodness-of-fit tests using residuals (e.g. Randles 1984) to assess the adequacy of a model.

Residuals, originally rooted in linear models, have been used in regression model diagnostics extensively, e.g., identifying outliers, detecting further factors, etc. Cox and Snell (1968) generalized the idea of residuals beyond normality by creating independent and identically distributed (i.i.d.) variables homogeneous across covariates. Their framework is effective for continuous outcomes. For example, for continuous observations $Y_i, i = 1, \ldots, n$, the uniform Cox-Snell residuals are defined as $\hat{F}_i(Y_i)$ with $\hat{F}_i$ as the fitted distribution function of $Y_i$. Given a well-fitting model, the Cox-Snell residuals should present a uniform trend, and otherwise lack of fit is implied.

However, the effectiveness of Cox-Snell residuals does not carry over to discrete outcomes, which in general cannot be expressed as transformations of i.i.d. variables. Yet regression models for discrete outcomes have been long and widely applied in many areas of research, including insurance, biology, and education, among many others. The focus of diagnostics for models with discrete outcomes has been to create approximately identically distributed variables. Following this line of thought, there is significant literature generalizing the idea of residuals to generalized linear models (GLMs) for discrete outcomes searching for the optimal transformations (e.g., Pierce and Schafer 1986). In practice, there are two types of established residuals commonly adopted for discrete observations (McCullagh and Nelder 1989). The first type is Pearson residuals defined as the signed square root of the contribution of each point to the Pearson goodness-of-fit statistic

$$r_P(Y_i; \theta) = \frac{Y_i - \mu_i}{V(\mu_i)^{1/2}}$$

where $\mu_i$ is the mean of the outcome $Y_i$ calculated using parameter $\theta$, and $V(\cdot)$ is the function relating the variance to the mean. The second type is deviance residuals which are based on the contribution of each observation to the likelihood and are the signed difference
between the likelihood of the fitted model and the saturated model for which the number of parameters is same as the number of observations, i.e.,

\[ r_D(Y_i; \theta) = \text{sgn} (Y_i - \mu_i) \left\{ 2 \left[ l(Y_i; \hat{\theta}) - l(Y_i; \theta) \right] \right\}^{1/2}, \]

where \( l(Y_i; \theta) \) denotes the log-likelihood function taking value at \( Y_i \) when the parameter is set to be \( \theta \), and \( \hat{\theta} \) is the parameter of the saturated model. There exist other well-known residuals, for instance Anscombe residuals (Anscombe 1961), though it has been shown that Anscombe residuals and deviance residuals behave similarly (McCullagh and Nelder 1989).

Dunn and Smyth (1996) proposed a randomized quantile residuals based on the idea of continuization, or jittering. Let \( a_i = \lim_{y \downarrow Y_i} \hat{F}_i(y) \) and \( b_i = \hat{F}_i(Y_i) \), then the randomized quantile residual for \( Y_i \) is defined as

\[ r_R(Y_i) = \Phi^{-1}(U_i), \]

where \( U_i \) is a uniform random variable on the interval \((a_i, b_i]\) independent of \( Y_i \). There are also residuals available for specific types of discrete data, e.g., binary (Landwehr et al. 1984) and ordinal data (Li and Shepherd 2012; Liu and Zhang 2018).

As an informative diagnostic tool, residuals should have the following two desirable properties. First, it is crucial that the residuals follow a known shape or pattern under the true model and deviate from this shape with model misspecification. This pattern of behavior is the foundation of diagnostics. Percentile-percentile (P-P) plots are commonly employed to visually assess discrepancies of the empirical distribution of residuals with the hypothesized distribution. However, residuals of discrete outcomes often deviate dramatically from the null shape even under the true model. It has been noted that the level of discreteness plays a key role in the behavior of residuals, so called \( m \)-asymptotics, in addition to the typical \( n \)-asymptotics (Pierce and Schafer 1986). Here \( m \) could be the size of binomial distributions, or the Poisson means, which controls the discreteness level. When \( m \) is small, deviance resid-
uals and Pearson residuals could have a large discrepancy with the null pattern (a normal distribution) even under the true model, and a large sample size \( n \) does not relieve this concern. As a result, diagnostics are not informative in this case. Second, diagnostic tools should be sensitive to model misspecification so as to provide effective detection of inadequacy of fitting. This may not be the case for randomized quantile residuals with another layer of randomness introduced (see, e.g. Figure 9).

In this paper, we revisit Cox-Snell residuals under discreteness. We construct a surrogate empirical residual distribution function which serves as a substitute for the empirical distribution of Cox-Snell residuals. Instead of attempting to construct residuals themselves as in most existing literature, we build a function for discrete outcomes based on the idea of local averaging. We show asymptotically that the proposed surrogate empirical residual distribution function converges to the identity function uniformly under the true model. The proof of such is nontrivial due to the unique form of the estimator. Under many types of misspecification including missing covariates, overdispersion, and incorrect link function, we show empirically the proposed surrogate empirical residual distribution function deviates significantly from the identity function. Hence, it can be used as an alternative to the empirical residual distribution function in P-P plots.

We highlight the contributions of the proposed diagnostic tool as follows. First, theoretically, the surrogate empirical residual distribution function converges to the identity function uniformly under a correctly specified model, even with small \( m \), which guarantees the effectiveness of diagnostics. In contrast, other standard residuals including deviance residuals are normally distributed with an error term of order at least \( O_p(m^{-1/2}) \) which cannot be fixed by large sample sizes. Second, the diagnostic tool we propose is a principled approach and does not require injecting noise to the data, such as is done for randomized quantile residuals. To the best of our knowledge, this is the only diagnostic approach which is not simulation-based and still guarantees the asymptotic convergence to the null shape for discrete outcomes. Third, we demonstrate empirically that the proposed tool outperforms
other diagnostic approaches in terms of the two key aspects of diagnostic tools as described above in various settings when at least one continuous covariate is available. Lastly, our tool works for general discrete outcomes including ordinary (binary) and count data, and it provides an overall check on goodness-of-fit. If insufficiency of fit is detected and one is interested in exploring the potential causes, more specific diagnostic tools in the literature can be further applied, e.g., detecting outliers (Pregibon 1981).

The rest of the paper is organized as follows. In Section 2, we present the surrogate empirical residual distribution function and its asymptotic properties. In Section 3, we demonstrate the usage and properties of the proposed tool in simulated examples, and Section 4 contains an application of the proposed tool on an insurance dataset. Discussion and conclusions are presented in Section 5. The appendix includes additional theoretical, simulation, and data analysis results. The proofs for theoretical results are included in the supplementary material.

2. METHODOLOGY

2.1 Why Not Directly Apply Cox-Snell Residuals?

Let $Y$ be the outcome of interest. Denote the distribution function of $Y$ conditional on covariates $X = x$ as $F(y|x) = P(Y \leq y|X = x)$, where $F$ belongs to a parametric family indexed by parameters $\beta$. The parameters $\beta$ can potentially relate to location, scale, and shape parameters.

If $Y$ is continuous, plugging $(X, Y)$ in $F$, the variable $F(Y|X)$ is known as the probability integral transform. For any fixed value $s \in (0, 1)$,

\[ P(F(Y|X = x) \leq s) = s, \tag{1} \]
and taking expectation with respect to $X$ yields

$$P(F(Y|X) \leq s) = s,$$  \hfill (2)

i.e. $F(Y|X)$ is uniformly distributed. Given an i.i.d sample $(X_i, Y_i), i = 1, \ldots, n$, with a fitted model $\hat{F}$ depending on fitted parameters $\hat{\beta}$, one can obtain a sequence of Cox-Snell residuals $\hat{F}(Y_i|X_i)$ and their empirical residual distribution function

$$\hat{U}_C(s; \hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} 1(\hat{F}(Y_i|X_i) \leq s).$$  \hfill (3)

Under a correctly specified model, $\hat{U}_C(\cdot; \hat{\beta})$ should be approximately an identity function, and thus a large discrepancy indicates misspecification.

Due to this property, it is common practice to use P-P plots to visualize the comparison between the empirical distribution of the Cox-Snell residuals and the null function under the hypothesized model, i.e., the identity function. Figure 1 portrays the P-P plots of the Cox-Snell residuals in simulated examples. On the left panel, the data are generated with a gamma regression model and the Cox-Snell residuals are calculated using the underlying model. As anticipated, the Cox-Snell residuals appear to be uniform.

Figure 1: P-P plots of Cox-Snell residuals. Left panel: gamma regression. Right panel: Poisson regression.
However, when $Y$ is a discrete variable taking integer values, without loss of generality, the Cox-Snell residuals are not uniformly distributed even under the true model. As in the right panel of Figure 1, when the data are generated from a Poisson GLM, the Cox-Snell residuals are far apart from uniformity even with the knowledge of the underlying model.

The uniformity of probability integral transforms invalidates under discreteness due to the fact that (1) is not true for some values of $s \in (0, 1)$, in contrast to the continuous cases. The lemma below gives the condition under which (1) holds for discrete outcomes.

**Lemma 2.1.** Conditioning on $X = x$, (1) holds for discrete $Y$ if and only if $s = F(k|x)$ for some integer $k$.

**Proof.** "If." Assume $s = F(k|x)$, then

$$P(F(Y|x) \leq s) = P(F(Y|x) \leq F(k|x)) = P(Y < k + 1|x) = P(Y \leq k|x) = s.$$ 

“Only if.” Now suppose $P(F(Y|x) \leq s) = s$. If $s \neq F(k|x)$ for any integer $k$, without loss of generality, assume there exists a $k_0$ such that $F(k_0|x) < s < F(k_0 + 1|x)$, as demonstrated in Figure 2. Then

$$P(F(Y|x) \leq s) = P(Y < k_0 + 1|x) = P(Y \leq k_0|x) = F(k_0|x) < s$$

which is contradictory. Therefore, it holds that $s = F(k|x)$ for some integer $k$. 

\[\square\]

### 2.2 Construction of Surrogate Empirical Residual Distribution Functions

To construct an alternative to $\hat{U}_C(\cdot; \beta)$ under discreteness, intuitively, for each point $s$, if one could find a subset of observations for which (1) holds, this subset can be plugged in (3) to obtain the identity function. Without loss of generality, assume the support of $Y$
Figure 2: Demonstrative plot of the distribution function of a discrete random variable.

is nonnegative integers. Motivated by Lemma 2.1, we then define the conditional range of the distribution function given \( X = x \) as a grid \( \Lambda(x) = \{ F(k|x) : k = 0, 1, \ldots \} \). Note that the range of \( Y \) can be finite, e.g. for binary variables, \( \Lambda(x) = \{ F(0|x), 1 \} \). From Lemma 2.1, (1) is true if and only if \( s \in \Lambda(x) \). If \( X \) contains continuous components, a common situation in practice, when \( X \) varies in regression, there might be a subset of observations for which \( s \in \Lambda(X) \) holds approximately, or equivalently, the distance between \( s \) and \( \Lambda(X) \), \( d(s, \Lambda(X)) = \min\{ |s - \eta|, \eta \in \Lambda(X) \} \), is small. Hence, we need to carefully characterize \( d(s, \Lambda(X)) \).

Conditioning on \( X = x \), denote \( F^{(-1)}(\cdot|x) \) as the general inverse function of \( F(\cdot|x) \) such that \( F^{(-1)}(s|x) = \inf\{ y : s \leq F(y|x) < 1 \} \) for \( s \in (0, 1) \). Here we exclude \( \{1\} \) to avoid boundary effects. Removing this point is not a concern because (1) always holds on the boundary. Denote

\[
H^+(s; x) = F(F^{(-1)}(s|x)|x).
\]
It can be seen that $H^+(s; x) = \min\{\eta \in \Lambda(x) \setminus \{1\} : \eta \geq s\}$, i.e., $H^+(s; x)$ is the smallest interior point on the grid $\Lambda(x)$ that is larger than or equal to $s$. In the same way, one can define the largest interior point on the grid $\Lambda(x)$ that is smaller than or equal to $s$ as

$$H^-(s; x) = \max\{\eta \in \Lambda(x) \setminus \{1\} : \eta \leq s\}.$$ 

To combine these two cases, define the interior grid point closest to $s$

$$H(s; x) = \begin{cases} 
H^+(s; x) & H^+(s; x) + H^-(s; x) \leq 2s \text{ or } s < F(0|x), \\
H^-(s; x) & H^+(s; x) + H^-(s; x) > 2s \text{ or } s > \max(\Lambda(x) \setminus \{1\}). 
\end{cases} \tag{4}$$

One can view $H(s; x)$ as the proximal interpolator which maps $s$ to its nearest neighbor on $\Lambda(x)$. It follows that $d(s, \Lambda(x) \setminus \{1\}) = |H(s; x) - s|$.

When $s$ is “close to” being on the grid given $x$ in the sense that $H(s; x) \approx s$, we have an approximation to (1)

$$P \left( F(Y|x) \leq H(s; x) \right) = H(s; x) \approx s,$$

where the first equation holds due to the fact that $H(s; x) \in \Lambda(x)$ by its definition and Lemma 2.1. Therefore, we can focus on the empirical residual distribution function among the subset of observations for which $H(s; X) \approx s$.

Now consider a sample $(X_i, Y_i), i = 1, \ldots, n$ and a fixed value of $s$. To realize the above idea, we use a kernel function $K(\cdot)$ to select the subset of observations whose grid is close to $s$ by assigning weights depending on the normalized distance between $s$ and $H(s; X_i)$, i.e., $K \left( \frac{(H(s; X_i) - s)}{\epsilon_n} \right)$, where $\epsilon_n$ is a small bandwidth. Then, we focus on the empirical distribution function of the residuals using the selected subset of data and define the surrogate empirical residual distribution function

$$\hat{U}(s; \beta) = \sum_{i=1}^{n} W_{ni}(s; X_i; \beta) 1 \left[ F(Y_i|X_i) \leq H(s; X_i) \right], \tag{5}$$
where

\[ W_{ni}(s; X_i, \beta) = \frac{K \left[ (H(s; X_i) - s)/\epsilon_n \right]}{\sum_{i=1}^{n} K \left[ (H(s; X_i) - s)/\epsilon_n \right]}, \]

and \( K \) is a bounded, symmetric, and Lipschitz continuous kernel.

The function \( \hat{U}(\cdot; \beta) \) should be close to the identity function under the true model and as a result, discrepancies from identity indicate lack of fit. Hence, \( \hat{U}(\cdot; \beta) \) can be used as a diagnostic tool in place of the empirical Cox-Snell residual distribution function for discrete outcomes. To demonstrate the effectiveness of the proposed tool immediately, Figure 3 (right panel) shows the curve of \( \hat{U}(\cdot; \beta) \) for the Poisson example; a more thorough simulation study is included in Section 3. In practice, \( \beta \) is unknown; let \( \hat{\beta} \) be the corresponding estimator. By plugging \( \hat{\beta} \) in (5), we may obtain \( \hat{U}(\cdot; \hat{\beta}) \). We will show in Section 2.4 that the uncertainty in the coefficients is negligible asymptotically.

![Figure 3](image1.png)

**Figure 3:** Left panel: P-P plot of Cox-Snell residuals of Poisson regression. Right panel: curve of the proposed surrogate empirical residual distribution function.

An important choice left to be made is the bandwidth \( \epsilon_n \). Here we propose to select a data-driven bandwidth by minimizing the \( L_2 \)-norm distance of \( \hat{U}(\cdot; \hat{\beta}) \) and the null pattern under true model, i.e., the identity function

\[ \min_{\epsilon_n} \int_{s} \left( \hat{U}(s, \hat{\beta}) - s \right)^2. \]

In practice, the minimizer can be searched numerically.
2.3 Examples

As an example, when $Y$ is a binary outcome in the most discrete case, its distribution grid only contains two points, i.e., $\Lambda(x) = \{F(0|x), 1\}$ and thus $H(s; x) = F(0|x)$. Then, (5) becomes

$$\sum_{i=1}^{n} K \left[ \frac{(F(0|X_i) - s)}{\epsilon_n} \right] 1(Y_i = 0) \sum_{i=1}^{n} K \left[ \frac{(F(0|X_i) - s)}{\epsilon_n} \right].$$

In the binary case, the proposed function takes the form of the Nadaraya-Watson estimator. In contrast, when $Y$ is continuous, which can be viewed as the limiting case when $Y$ gets less discrete, it is always true that $H(s; X_i) = s$. It follows that $W_{ni}(s; X_i, \beta) = 1/n$, and (5) degenerates to (3). In both extreme cases, the resulting (surrogate) empirical residual distribution function has been extensively studied in the literature.

Technical difficulties and major departures from existing methods are pronounced when $Y$ is discrete with an infinite range, for instance a Poisson variable, under which $\hat{U}(\cdot; \beta)$ is nonstandard in the following aspects. First, for a fixed point $s$, $H(s; X_i)$ is a noncontinuous variable. We include an example for illustration assuming that $Y$ follows the commonly used Poisson GLM with the log link, i.e., $Y|X \sim \text{Poisson}(\exp(X'\beta))$. Figure 4 shows $H(s; X)$ (red solid curves) for fixed $s$ as a function of the random location parameter $\mu = X'\beta$. In this example, for a fixed $k$, $F(k|X)$ (dashed lines) is a monotone decreasing function of $\mu$. The curve of $H(s; X)$ as a function of $\mu$ is comprised of continuous pieces from the curves of $F(k|X), k = 0, 1, \ldots$, and the transitions occur when the nearest neighbor of $s$ changes from $F(k)$ to $F(k + 1)$ for integer $k$ as $\mu$ increases. The random variable $H(s; X)$ is a continuous function of $\mu$ almost everywhere except at a countable number of points under which there are two nearest neighbors of $s$ on $\Lambda(X)$. We will further address the issue of discontinuity in Section 2.4, which complicates the proof for asymptotic properties.

The second complicating factor is that $H(s; X)$, the proximal interpolator of $s$, certainly changes with $s$. It implies that when focusing on different points, we plug different variables into the kernel function, which distinguishes the proposed function from traditional
Figure 4: $H(s; X)$ (solid red curve) for fixed $s$ as a function of $\mu = X'\beta$ for Poisson GLM with the log link. Dashed black curves: $F(k|X)$, from left to right $k = 0, 1, 2, 3, 4, 5, 15, 16$. The curve of $H(s; X)$ is comprised of pieces from the curves of $F(k|X), k = 0, \ldots$. Blue horizontal lines: $s + \epsilon, s - \epsilon$.

nonparametric regression models. Assuming continuity of $\mu$, $F(k|X)$ is random with a density denoted as $f_{F(k|X)}$, and $f_{H(s;X)}$ is the density of $H(s; X)$. From the form of (5), the weights $W_{ni}(s; X_i; \beta)$ relate to the density of $f_{H(s;X)}$ at $s$, i.e., $f_{H(s;X)}(s)$. By transformation of random variables

$$f_{H(s;X)}(s) = \sum_{k=0}^{\infty} f_{F(k|X)}(s) = g(s).$$

Note that $f_{H(s;X)}(t)$ is a density function with respect to $t$, while $g(s)$ is not a density function with respect to $s$. According to (6), unlike typical nonparametric regression methods for which one assigns weights to observations depending on one variable, here all the $F(k|X), k = 0, \ldots$ contribute to the weights, and this dynamic scheme increases efficiency. Meanwhile, in Section 2.4, by making realistic assumptions, we make sure $g(s)$ is bounded. For ordinal (binary) outcomes with finite support, the summation is up to the second largest possible value. For example, in Figure 5, $f_{H(s;X)}(s) = f_{F(0|X)}(s)$ for the left panel and $f_{H(s;X)}(s) = \sum_{k=0}^{2} f_{F(k|X)}(s)$ for the right panel.
Figure 5: $H(s; X)$ (solid curve) as a function of $\mu = X'\beta$ for logistic regression (left panel) and ordinal regression with 4 levels (right panel). Dashed curves for right panel: $F(k|X)$, from left to right $k = 0, 1, 2$. Horizontal lines: $s + \epsilon, s, s - \epsilon$.

2.4 Asymptotics

In this section, we study the asymptotic properties of the surrogate empirical residual distribution function $\hat{U}(\cdot; \beta)$ defined in (5). We first analyze $\hat{U}(\cdot; \beta)$ with known $\beta$ and then analyze it when a $\sqrt{n}$ consistent estimator is plugged in. For the sake of simplicity, here we only show the asymptotic properties when regression is conducted on a single location parameter, though our methodology is applicable to general settings, e.g., double GLMs with the dispersion parameter as a function of covariates as well, or zero-inflated Poisson models with more than one location parameter. We will demonstrate the usage of our methodology in more general situations through numerical examples in Section 3.

Let $\mu = X'\beta$ be the random location parameter and a monotone increasing function of the mean. For example, in Poisson regression, the mean $\lambda = \exp(\mu)$, and in logistic regression the mean is $1/(1 + \exp(-\mu))$. For a fixed $k$, we assume $F(k|X)$ is a monotone decreasing function of $\mu$, which is satisfied for many commonly used models including logistic, Poisson, and negative binomial models.

As pointed out in the previous section, $H(s; X)$ is not a continuous function of $\mu$. Consequently, the density of $H(s; X)$ (i.e., $f_{H(s; X)}$) is not smooth. The density of $H(s; X)$ at a point other than $s$, i.e., $f_{H(s; X)}(s + \epsilon)$ has a different form from $f_{H(s; X)}(s)$ for $\epsilon \neq 0$. For
a small $k$ such as $k \leq 5$ in Figure 4, $f_{F(k|X)}$ contributes to $f_{H(s,X)}$ at $s + \epsilon$ when applying transformation of random variables. While for a large $k$ such as $k = 15$, the density of $F(15|X)$ does not contribute to the density of $f_{H(s,X)}$ at $s + \epsilon$. Therefore, in this example,

$$\sum_{k=0}^{5} f_{F(k|X)}(s + \epsilon) \leq f_{H(s,X)}(s + \epsilon) < \sum_{k=0}^{\infty} f_{F(k|X)}(s + \epsilon).$$

(7)

That is, compared with (6), $f_{H(s,X)}$ is not smooth due to loss of $f_{F(k|X)}$ curves contributing to $f_{H(s,X)}$ at $s + \epsilon$. The non-smoothness issue is less of a concern for variables with a finite range. When $\epsilon$ takes a small value $\epsilon_n$ which goes to 0, a finite number of jump points would be excluded from the small neighborhood of $s$, as in Figure 5.

To exclude the boundary effect and achieve uniform convergence, we focus on a closed subset of $(0,1)$ denoted as $[a_L, a_U]$. Let $V$ be a subset of $[a_L, a_U]$ such that for $s \in V$, $g(s) > 0$. Let the bandwidth $\epsilon_n$ satisfy that $\epsilon_n \to 0$ and $n\epsilon_n \to \infty$ as $n \to \infty$, and $n\epsilon_n^5 = O(1)$. Assume we can interchange the derivatives and the limits, then the derivatives $g'(s) = \sum_{k=0}^{\infty} f'_{F(k|X)}(s)$, $g''(s) = \sum_{k=0}^{\infty} f''_{F(k|X)}(s)$. Let $K$ be a symmetric kernel function with compact support, and denote $R_2(K) = \int K(u)^2 du$, $\kappa_2 = \int u^2 K(u) du$, we have the following property with regularity conditions described in Appendix A.

**Theorem 2.1** (Weak Convergence). Define $Z_n(s) = \sqrt{n\epsilon_n} \left( \hat{U}_n(s; \beta) - s \right)$. Under Assumptions A.1 and A.2, and if we further assume that $n\epsilon_n^5 \to R_\epsilon$, uniformly for $V$,

$$Z_n \Rightarrow Z$$

where $Z$ is a Gaussian process with pointwise mean $\kappa_2 g'(s) \sqrt{R_\epsilon}/g(s)$ and variance $R_2(K)s(1-s)/g(s)$. In addition, the covariance of $Z$ is zero for any two distinct points.

With additional assumptions described in the appendix, we guarantee the weak convergence of the surrogate empirical residual distribution function when the estimated coefficients are plugged in. In addition, we adopt kernel functions satisfying Hölder conditions with ex-
ponents 2, i.e., there is a constant $\alpha_1$ such that uniformly for $u, v$,

$$|K(u) - K(v)| \leq \alpha_1 |u - v|^2.$$  

Kernels with high order of smoothness, e.g., Epanechnikov and quartic kernels, satisfy this condition.

**Theorem 2.2** (Weak convergence with estimated coefficients). With the estimator of the model plugged in, $\hat{Z}_n(s) = \sqrt{n\epsilon_n} \left( \hat{U}(s; \hat{\beta}) - s \right)$. Under Assumptions A.1, A.2, A.3, and A.4, and further, if we adopt kernel functions satisfying Hölder conditions with exponents 2, then uniformly for $s \in V$, $\hat{Z}_n$ weakly converges to the same Gaussian process as in Theorem 2.1.

To analyze the asymptotic results, first, the discrepancy of our proposed empirical process with the null pattern depends on both $m$ and $n$, i.e., pointwise bias $\kappa_2 g'(s) \epsilon_n^2 / g(s)$ and variance $s(1 - s) R_2(K) / [n \epsilon_n g(s)]$. Even if the data are highly discrete, i.e. $m$ is small, the bias and variance go to 0 when the sample size is large. In contrast, the deviance residuals are normally distributed with an error term of order at least $O_p(m^{-1/2})$ which cannot be improved by large sample sizes. Meanwhile, from the form of the variance, a large $m$ leads to large values of $g$ and thus a small variance. Second, the proposed tool has a slower convergence rate than $n^{-1/2}$ as for deviance residuals. Therefore, the proposed tool requires a larger sample size for satisfactory performance.

## 3. SIMULATION

In this section, we use a variety of numerical examples to demonstrate model diagnostics using our surrogate empirical residual distribution functions. We examine two important aspects: the proximity of $\hat{U}(\cdot; \hat{\beta})$ to null patterns under true models and its discrepancy with null patterns under misspecified models. Throughout this section, the bandwidth is selected
using the approach proposed in Section 2, and the Epanechnikov kernel is used.

3.1 Closeness to Null Patterns under Correct Model Specification

Poisson examples. As a valid diagnostic tool, it is essential to guarantee the closeness to the null pattern if the model is correctly specified. This has been an issue for commonly used residuals including deviance and Pearson residuals when the data are highly discrete. We first explore the effect of the discreteness level on different residuals, i.e., \( m \)-asymptotics, through Poisson examples under GLMs with log link. The location parameter \( \mu = X' \beta = \beta_0 + X_1 \beta_1 + X_2 \beta_2 \), where \( X_1 \sim N(0, 1) \), and \( X_2 \) is a dummy variable with probability of 1 as 0.7. The covariates \( X_1 \) and \( X_2 \) are independent. We conduct simulations with three levels of discreteness:

- Small mean: \( \beta_0 = -2, \beta_1 = 2, \beta_2 = 1 \).
- Medium mean: \( \beta_0 = 0, \beta_1 = 2, \beta_2 = 1 \).
- Large mean: \( \beta_0 = 5, \beta_1 = 2, \beta_2 = 1 \).

For each of the experiments, we generate data, fit the correct regression model, and compute the residuals or surrogate empirical residual distribution function. We then summarize the results graphically by providing the curve of \( \hat{U}(\cdot; \hat{\beta}) \) and the PP-plots of other types of residuals. Given a correct model specification, the null pattern should be along the diagonal.

Figure 6 presents the results with sample size 500. The upper row corresponds to the small mean scenario. As anticipated, the deviance and Pearson residuals are far apart from normality due to small \( m \), while the proposed surrogate empirical residual distribution function is much closer to the identity function. That is, our method provides more reliable conclusions for cases with high level of discreteness. When we move to the middle row corresponding to the medium mean level, our method keeps the pattern along the diagonal, and the other two residuals are getting closer to being normally distributed. As the mean increases to large case (lower row of Figure 6), all three methods appear close to the null
pattern. This also draws our attention that under the large mean, all the residuals behave similarly and equally well.

The demonstrative results in Figure 6 are based on one sample. We further include the results with replications in the appendix as Figure 13 to visualize the standard error. As one can see, the diagonal is within the confidence band of the proposed method even under high level of discreteness, though the proposed method has a large variance under the small mean scenario compared with the other two types residuals. When the mean increases to medium level, the variance gets smaller, which is consistent with the theoretical result of Theorem 2.2.

**Binary examples.** We also include the results for binary outcomes with logistic regression model, which is the most discrete possible case. We set \( \text{logit}(p_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \) where \( p_1 \) is the probability of 1, \( \beta_0 = -2, \beta_1 = 2, \beta_2 = 1 \), \( X_1 \sim N(0, 1) \), and \( X_2 \) is a dummy variable with probability of one as 0.7. Figure 7 summarizes the results. From the first row, when the sample size is 100, the difficulty resulting from high level of discreteness is clear since the deviance and Pearson residuals are far apart from normality. However, the proposed method is reasonably close to the null pattern. When we increase the sample size to 2000, the deviance and Pearson residuals do not improve, whereas the proposed method gets closer to the null pattern. This is consistent with our theoretical results in Section 2.4. For the proposed method, a large sample size is a remedy for the poor performance induced by high level of discreteness, whereas the discrepancy of the deviance and Pearson residuals with normality cannot be fixed by increasing sample size. This property of the proposed method is especially useful under the current trend of utilizing large datasets. We can draw similar conclusions from other examples including negative binomial distributions, which we will include in Section 3.2 for a comparison with misspecified models.
Figure 6: Diagnostic plots for Poisson outcomes under correct models. The three rows correspond to small, medium, and large mean levels. Sample size 500.

### 3.2 Diagnostics for Misspecification

As concluded in the previous section, the proposed method is closer to null pattern under the true model compared with other residuals in various settings. In this section we check the ability of the proposed method to detect important causes of model misspecification including omission of covariates, overdispersion, and incorrect link functions.

**Missing covariates.** We first demonstrate that our surrogate empirical residual distribution function is a useful diagnostic tool for detecting missing covariates through a Poisson
Figure 7: Diagnostic plots for binary outcomes under correct models. The two rows correspond to sample sizes 100 and 2000.

example. The underlying model is \( \mu = X' \beta = \beta_0 + X_1 \beta_1 + X_2 \beta_2 \), where \( X_1, X_2 \sim N(0,1) \) independently, and \( \beta_0 = -2, \beta_1 = 2, \beta_2 = 1.5 \). Under the misspecified model, the covariate \( X_2 \) is missing. Figure 8 includes the results in which we compare the proposed method with deviance residuals and randomized quantile residuals. Given the comparable performances of deviance and Pearson residuals, we omit Pearson residuals. By comparing the top row which are for the true model with the bottom row corresponding to the misspecified model, we can see that our tool is illuminating in the sense that the curve is close to the null pattern under the true model while it shows a large discrepancy when covariates are missing even with high level of discreteness. For deviance residuals and randomized quantile residuals, there is no significant improvement from misspecified model to the true model, and thus it is hard to draw a conclusion about whether the model is sufficient for the data. The results for medium mean level with \( \beta_0 = 0 \) is included in the appendix as Figure 15, in which all the methods become more informative, and the proposed methods still outperforms.
Overdispersion. For discrete outcomes, especially count data, one of the most common issues is overdispersion. We next use numerical examples to examine the ability of the proposed method as well as other residuals to detect overdispersion under different scenarios. We first generate data using negative binomial distributions with the same mean structure as for the Poisson outcomes in Section 3.1, and the size parameter is set to be 2. For the misspecified model, we fit the data with a Poisson GLM and thus overdispersion is present. Figure 9 shows the results for the small mean scenario. The top row includes the results for the true model, while the bottom row shows the results under the misspecified model. By comparing the first column, which corresponds to the proposed method, we can see that the proposed tool is again informative. In contrast, the deviance residuals show a large discrepancy under both models, while randomized residuals are not sensitive to misspecification in this case.
Figure 9: Diagnostic plots for overdispersion with negative binomial outcomes under small mean. The top row corresponds to the true model, and the bottom row shows the results when Poisson GLM is mistakenly used. Sample size: 500.

Zero-inflated Poisson models are also commonly adopted to tackle the overdispersion issue. We now include an example of the zero-inflated Poisson model to demonstrate the usage of the proposed method for non-exponential distributions. The probability of excess zero is modeled with $\logit(p_0) = \beta_{00} + \beta_{10} X_1$, and the Poisson component has a mean $\lambda = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ where $X_1 \sim N(0, 1)$ and $X_2$ is a dummy variable with probability of 1 as 0.7, and $(\beta_{00}, \beta_{10}, \beta_0, \beta_1, \beta_2) = (-2, 2, 0, 2, 1)$. We compare the results of the true model with the results from a Poisson model through the proposed approach and Pearson residuals, since deviance residuals are not as well-defined for models with more than one location parameter. Figure 10 shows the results from which we can see the proposed method can help detect the insufficiency of fitting. It is noticeable that the curve of $\hat{U}(\cdot; \hat{\beta})$ seems unstable at lower left corner. This is due to the fact that this part is out of the $V$ area as in Theorem 2.2. When constructing formal goodness-of-fit tests, one can diminish the influence.
of this part by downweighting the corresponding area. In contrast, Pearson residuals show a large discrepancy over the whole range under the true model and thus is not revealing.

![Diagnostic plots for zero-inflated Poisson model. Sample size 500.](image)

**Incorrect link function.** To provide diagnostics for link functions, we include a Poisson example. The true link function is the square root function, i.e., the mean \( \lambda = (\beta_0 + \beta_1 X_1 + \beta_2 X_2)^2 \) where \( X_1 \sim N(0, 1) \) and \( X_2 \) is a binary variable with probability of 1 as 0.7, and \( (\beta_0, \beta_1, \beta_2) = (0, 1, 1) \). In the misspecified model, the log link is used. The comparative results are included in Figure 11. By comparing across the two rows, we can see the proposed method and the randomized quantile residuals show the transition from being close to the 45 degree line to a disagreement. The deviance residuals also show a larger discrepancy with incorrect link function, though there is a noticeable difference with the null pattern under the true model. Nonetheless we also notice that under many scenarios, the log link can provide reasonable fitting even if the true link function is a square root or identity.
function.

Figure 11: Diagnostic plots for link function for Poisson models. Sample size 500. The two rows correspond to right link (square root) and wrong link (log).

4. ANALYSIS OF INSURANCE CLAIM FREQUENCY DATA

In this section, we present an application of the proposed surrogate empirical residual distribution function to insurance claim frequency data. Frequency, the number of reported claims for each policyholder, is an important component of insurance claim data and largely reveals the riskiness of policyholders. Here we use a dataset from the Local Government Property Insurance Fund (LGPIF) in the state of Wisconsin. The LGPIF was established to provide property insurance for local government entities and provides different types of coverage including government buildings, vehicles, and equipments. In this paper we focus on building and contents (BC) insurance, which is the major coverage offered by the LGPIF. The
dataset contains 5660 observations from year 2006 to 2010. Table 1 provides the empirical numbers of observations. Covariates together with their summary statistics are displayed in the appendix Table 3. Among them, coverage and deductible are continuous covariates.

Here we fit several commonly employed count regression models to the claim frequency data: Poisson, negative binomial (NB), and zero-inflated Poisson. In addition, it can be seen from Table 1 that the data contain a large number of zeros and a significant amount of ones. This motivates the usage of zero-one-inflated Poisson models in Frees et al. (2016) whose distribution function can be expressed as

\[
F(k) = \begin{cases} 
\pi_0 + (1 - \pi_0 - \pi_1) \exp(-\lambda) & k = 0, \\
\pi_0 + \pi_1 + (1 - \pi_0 - \pi_1) \sum_{i=0}^{k} \frac{(\exp(-\lambda))^{i}}{i!} & k > 0,
\end{cases}
\]

where \(\pi_0\) and \(\pi_1\) are the probabilities of extra zeros and ones, respectively, and \(\lambda\) is the expected Poisson count.

We then apply the proposed diagnostic tool on the models. Figure 12 summarizes the curves of \(\hat{U}(\cdot; \hat{\beta})\) defined in (5). As we can see, the zero-one-inflated Poisson model and the negative binomial model provide satisfactory results, while the Poisson and zero-inflated Poisson models fit the data poorly. This is further confirmed by numerical summary in Table 2, which displays the \(L_2\)-norm distances between \(\hat{U}(\cdot; \hat{\beta})\) and the 45 degree line

\[
\int_{[0.3,0.9]} \left( \hat{U}(s, \hat{\beta}) - s \right)^2,
\]

where the integration range is selected as a subset of \(V\) in Theorem 2.2. We can see the zero-one-inflated Poisson model outperforms other models with smallest distance, which is
Table 2: $L_2$ norm distances of $\hat{U}(\cdot; \hat{\beta})$ from different models with the diagonal (multiplied by 1000)

|           | Poisson | NB     | Zero-Inflated Poisson | Zero-One-Inflated Poisson |
|-----------|---------|--------|-----------------------|---------------------------|
|           | 13.300  | 0.455  | 3.988                 | 0.445                     |

consistent with the results depending on goodness-of-fit test statistics in Frees et al. (2016), though the negative binomial model is also reasonably good. The coefficients of the selected model (zero-one-inflated Poisson) are provided in the appendix in Table 4.

![Diagnostic plots for Poisson, negative binomial, zero-inflated Poisson, and zero-one-inflated Poisson models on the LGPIF data.](image)

Figure 12: Diagnostic plots for Poisson, negative binomial, zero-inflated Poisson, and zero-one-inflated Poisson models on the LGPIF data.

5. CONCLUSIONS

In this paper we proposed a surrogate empirical residual distribution function for diag-
nosing regression models with discrete outcomes when at least one continuous covariate is available. We showed the weak convergence of the proposed surrogate empirical residual distribution function, and through simulation studies and empirical analysis, we demonstrated that the proposed method possesses the appealing properties for diagnostic tools that it is close to the hypothesized pattern under the true model, and under the misspecified model, it shows a significant discrepancy. It was also highlighted that under a high level of discreteness (e.g., binary outcomes and Poisson outcomes with small means), the proposed method gives reasonable results, and as the sample size increases, the performance improves, whereas for other commonly used residuals such as deviance residuals, there is a significant error term which cannot be fixed by large sample sizes.

Besides the graphical diagnostics of regression models, the proposed tool may be a reasonable starting point for the construction of goodness-of-fit tests to obtain conclusions with statements of statistical confidence. However, this is beyond the scope of this paper. In this paper, we have shown the weak convergence of the proposed function, which builds essential foundations for goodness-of-fit tests; we leave it for future research.

A. ADDITIONAL THEORETICAL RESULTS

To formalize the discontinuity patterns in $H(s; X)$, denote $M^k_s$ as the jump point of $H(s; X)$ transiting from $F(k - 1|X)$ to $F(k|X)$, then

$$H(s; X) = F(k|X) \text{ when } M^k_s \leq \mu < M^{k+1}_s.$$  (8)

When $M^0_s = -\infty < \mu < M^1_s$, for example, $F(0|X)$ is closest to $s$, and $H(s; X) = F(0|X)$ from the definition of $H(s; X)$. When $\mu = M^1_s$, $F(0|X)$ and $F(1|X)$ are equidistant from $s$. While $M^1_s < \mu < M^2_s$, $F(1|X)$ is closest to $s$ and thus $H(s; X) = F(1|X)$.

The following Lemma A.1 and Assumption A.1 are made to handle the non-smoothness issue for variables with an infinite range. Lemma A.1 guarantees the summation on the left
of (7) can be up to a large number \( a_n \) going to \( \infty \), and \( f_{H(s;X)}(\cdot) \) can be approximated by \( \sum_{k=0}^{a_n} f_{F(k|X)}(\cdot) \), which is smooth in the \( \epsilon_n \) neighborhood of \( s \).

**Lemma A.1.** There exists a sequence \( a_n \) going to infinity such that for all \( k \leq a_n \), for any \( s \in [a_L, a_U] \), \( f_{H(s;X)}(s + \epsilon_n) \geq \sum_{k=0}^{a_n} f_{F(k|X)}(s + \epsilon_n) \).

Lemma A.1 can be satisfied by choosing right order of \( a_n \) depending on \( \epsilon_n \). The proofs for all the theoretical results can be found in the supplementary material. The following assumption constrains the tail probability of \( \mu \) to ensure \( \sum_{k=0}^{a_n} f_{F(k|X)}(\cdot) \) is a good approximation to \( f_{H(s;X)}(\cdot) \).

**Assumption A.1.** Let \( a_n \) be the sequence in Lemma A.1 for \( H(\cdot;X) \), then \( \epsilon_n^{-2} P(\mu > M_{s_n}^a) \rightarrow 0 \), for any \( s \in [a_L, a_U] \).

We make the following regularity assumption to ensure \( f_{F(k|X)} \) is sufficiently smooth.

**Assumption A.2.** For fixed \( k \), \( f_{F(k|X)} \) is twice continuously differentiable, and \( g \) and its derivatives \( g' \) and \( g'' \) are bounded uniformly. In addition, for any \( k \) and \( k' \), the joint density of \( (F(k|X), F(k'|X)) \) are bounded.

A necessary condition for Assumption A.2 is that there exists at least one continuous regressor whose coefficient is not 0. When \( Y \) follows a Poisson distribution with mean \( \lambda = \exp(\mu) \), Assumptions A.1 and A.2 are satisfied if \( E_X \lambda \) is finite, and they hold for negative binomial distributions if \( E_X \lambda^2 \) is finite; see the supplementary material for verification. Therefore, if there are highly right-skewed covariates, log transformation is suggested. For binary and ordinal variables, Assumption A.2 is satisfied given that the density of \( \mu \) is twice continuously differentiable.

Now, we make the following assumption in order to guarantee the convergence of the surrogate empirical residual distribution function when the estimated coefficients are plugged in. Denote \( H(s;X, \theta) \) as closet interior grid point to \( s \) when the coefficients are set to be \( \theta \).
**Assumption A.3** (Lipschitz condition). There exists a constant $\alpha_2$ uniformly such that for all for bounded $\theta$ and $\theta'$, when $|\theta - \theta'|$ is small enough, for any $s \in V$

$$||H(s; X, \theta) - s| - |H(s; X, \theta') - s|| \leq \alpha_2|\theta - \theta'|$$

almost surely.

This assumption is satisfied when $Y$ follows Poisson GLMs with the log link and bounded covariates. The following assumption guarantees the model estimation is well taken care of.

**Assumption A.4.** $\sqrt{n}(\hat{\beta} - \beta) = O_p(1)$.

### B. ADDITIONAL SIMULATION AND DATA ANALYSIS RESULTS

The randomized quantile residuals (Dunn and Smyth (1996)) are based on the idea of continuization and are normally distributed under the true model theoretically. However, in order to produce continuous variables, randomness is introduced by adding an external uniform variable, as mentioned in Section 1. Hence, it is not robust if the sample size is small as a result of an extra layer of randomness. As indicated in Figure 14 with small mean Poisson outcomes, for a given dataset, the randomized quantile residuals give quite different conclusions with different random numbers (middle and right panels), though this randomness would vanish as the sample size increases or the discreteness level reduces.

Figure 15 includes the results for detecting missing covariates for medium discreteness level. We can see that all the methods become more informative compared with the small mean scenario (Figure 8). Though deviance residuals show a slight discrepancy under the true model, and the randomized quantile residuals appear to be reliable in this case.

For additional data analysis results, Table 3 summarizes the rating variables, and Table
Figure 13: Diagnostic plots for Poisson outcomes under correct models with replications. Sample size 500. The black curves correspond to the samples shown in Section 3, and the grey curves are for replications.
Figure 14: Diagnostic plots for Poisson outcomes with high discreteness level using the proposed method and randomized residual distribution function with two different random numbers. Sample size 100.

Figure 15: Diagnostic plots for Poisson outcomes when covariates are missing under medium mean scenario. Top row: correct model. Bottom row: $X_2$ is omitted. The sample size is 500.
Table 3: Description and summary statistics of covariates.

| Variable   | Description                           | Mean (s.d.) |
|------------|---------------------------------------|-------------|
| TypeCity   | =1 if entity type is city             | 0.140       |
| TypeCounty | =1 if entity type is county           | 0.058       |
| TypeSchool | =1 if entity type is school           | 0.282       |
| TypeTown   | =1 if entity type is town             | 0.173       |
| TypeVillage| =1 if entity type is village          | 0.237       |
| TypeMisc   | =1 if entity type is other            | 0.110       |
| NoClaimCredit | =1 if no building and content claims in prior year | 0.328 |
| lnCoverage | Coverage of BC line in logarithmic millions of dollars | 2.119 (2.000) |
| lnDeduct   | BC deductible level in logarithmic millions of dollars | 7.155 (1.174) |

Table 4: Coefficient estimates.

| Variable Name | Coef. | s.e. |
|---------------|-------|------|
| Count (Intercept) | -1.540 | 0.125 |
| lnCoverage    | 0.751 | 0.023 |
| lnDeduct      | -0.020 | 0.017 |
| NoClaimCredit | -0.395 | 0.131 |
| TypeCity      | -0.143 | 0.079 |
| TypeCounty    | -0.250 | 0.087 |
| TypeMisc      | -0.195 | 0.179 |
| TypeSchool    | -1.157 | 0.085 |
| TypeTown      | 0.186  | 0.175 |
| Zero (Intercept)| -4.755 | 0.448 |
| lnCoverage    | -0.580 | 0.078 |
| lnDeduct      | 0.879  | 0.062 |
| NoClaimCredit | 0.536  | 0.280 |
| One (Intercept)| -5.533 | 0.639 |
| lnCoverage    | -0.047 | 0.094 |
| lnDeduct      | 0.577  | 0.084 |
| NoClaimCredit | 0.300  | 0.353 |
4 includes the fitted coefficients of the selected model.

SUPPLEMENTARY MATERIALS

The supplementary materials include proofs for theoretical results in Sections 2.4 and Appendix A.

References

Anscombe, F. J. (1961). Examination of residuals. Technical report, PRINCETON UNIVERSITY PRINCETON United States.

Ben, M. G. and V. J. Yohai (2004). Quantile–quantile plot for deviance residuals in the generalized linear model. *Journal of Computational and Graphical Statistics* 13(1), 36–47.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275.

DeGroot, M. H. (1987). A conversation with george box. *Statistical Science* 2(3), 239–258.

Dunn, P. K. and G. K. Smyth (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5(3), 236–244.

Frees, E. W., G. Lee, and L. Yang (2016). Multivariate frequency-severity regression models in insurance. *Risks* 4(1), 4.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* 79(385), 61–71.

Li, C. and B. E. Shepherd (2012). A new residual for ordinal outcomes. *Biometrika* 99(2), 473–480.

Liu, D. and H. Zhang (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 1–10.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*, Volume 37. CRC press.

Pierce, D. A. and D. W. Schafer (1986). Residuals in generalized linear models. *Journal of the American Statistical Association* 81(396), 977–986.

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

Randles, R. H. (1984). On tests applied to residuals. *Journal of the American Statistical Association* 79(386), 349–354.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.
