Rare events data and zero reduction sampling [1] [2]

HaiYing Wang

University of Connecticut

2024 Joint Research Conference

[1] Suppoted by NSF CCF 2105571

[2] Panyi Dong, Jiaqi Liu, Zhiyu Quan

Outline

1. Introduction
2. Model setup and full data results
3. Zero Reducing Sampling
   - Inverse probability weighting (IPW)
   - Conditional Likelihood Estimator
   - Test for goodness of fit
4. Numerical results
   - Simulated data
   - PANDOR data

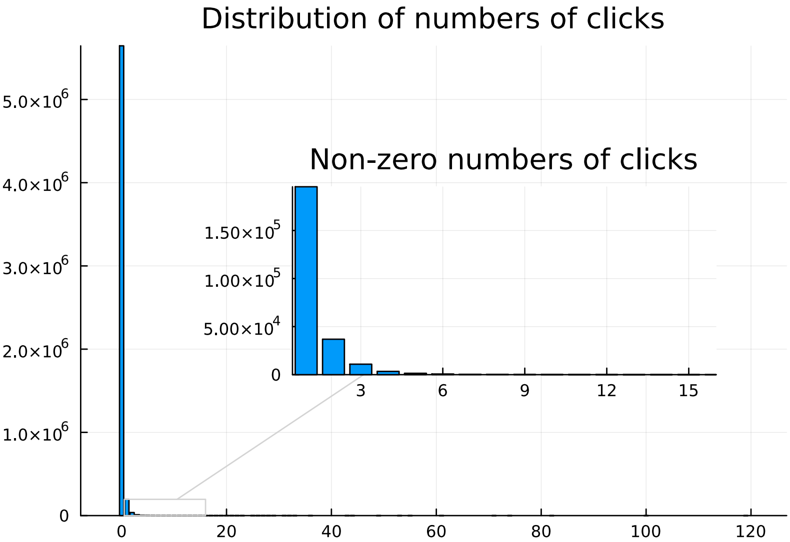
Outline

Introduction

- Rare events data occurs when the outcome of interest occurs infrequently, and the majority of observed responses are zeros.
- A typical example is very imbalanced binary data where the number of cases is much smaller than the number of controls.
- Rare events data beyond binary responses are also common.
  - About 5% of the 678,013 insurance policies in the French Motor Third-Party Liability (MTPL) dataset incurred at least one claim.
  - From 2006-2015, less than 1% of the insured homes in Connecticut had one or more claims on weather-related damages. [3]
  - In large online recommendation systems, most users do not click on any offers. For example, in the PANDOR data (Sidana *et al.*, 2018), less than 4% of the 5,894,430 users made one or more clicks on the offers shown.

[3] `https://www.iii.org/fact-statistic/facts-statistics-homeowners-and-renters-insurance`

Numbers of clicks in the PANDOR data



Some questions

- For non-binary responses, can we treat all non-zeros as ones to convert the data into binary rare events data?
- Whether the available information in the data is limited by the number of non-zeros?
- Weather all zeros are the same? Are there rare zeros?
- Will optimal subsampling designs prefer rare zeros?

Outline

Zero inflated regression

Let $\mathcal{D}_N = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ be observed data from the distribution of (X, Y). Let the conditional density of Y given $X = \boldsymbol{x}$ be

$$d(y \mid \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = p_0(\boldsymbol{z}, \boldsymbol{\theta}) I(y = 0) + \{1 - p_0(\boldsymbol{z}, \boldsymbol{\theta})\} h(y \mid \boldsymbol{v}; \boldsymbol{\gamma}), \quad (1)$$

where $\boldsymbol{z}$ and $\boldsymbol{v}$ are components of $\boldsymbol{x}$ and are allowed to overlap;

- $p_0(\boldsymbol{z}, \boldsymbol{\theta})$ generates dominating zeros;

$$p_0(\boldsymbol{z}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-g(\boldsymbol{z}, \boldsymbol{\theta})}} = \frac{1}{1 + e^{-\alpha - f(\boldsymbol{z}, \boldsymbol{\beta})}}; \quad (2)$$

- $h(y \mid \boldsymbol{v}; \boldsymbol{\gamma})$ is a density that generates rare observations;
- If $\mathbb{P}_h(Y = 0 \mid \boldsymbol{v}; \boldsymbol{\gamma}) > 0$, then model (1) has two types of zeros:
  - the dominating zeros from $p_0(\boldsymbol{z}, \boldsymbol{\theta})$
  - the rare zeros from $h(y \mid \boldsymbol{v}; \boldsymbol{\gamma})$;
- $\boldsymbol{\eta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ are unknown parameter vector.

Assumption on rareness

- Assume the true $\alpha \to \infty$ [4] so that

$$\frac{N_{nz}}{N} \xrightarrow{P} 0; \tag{3}$$

$$N_{nz} \xrightarrow{P} \infty \tag{4}$$

as $N \to \infty$, where N_{nz} is the number of nonzeros in the data.

[4] Wang (2020); Wang *et al.* (2021)

Full data estimator

Theorem 2.1

The full data MLE $\hat{\boldsymbol{\eta}}$ *satisfies that*

$$\sqrt{N_{nz}}(\hat{\boldsymbol{\eta}}_{\text{full}} - \boldsymbol{\eta}) \to \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\text{full}}), \tag{5}$$

where $\mathbf{V}_{\text{full}} = \mathbb{E}[e^{-f}(1 - h_0)]\boldsymbol{\Sigma}_{\text{full}}^{-1}$,

$$\boldsymbol{\Sigma}_{\text{full}} = \mathbb{E}\left(e^{-f}\begin{bmatrix}(1-h_0)\dot{g}^{\otimes 2} & \dot{h}_0\dot{g}^{\mathrm{T}} \\ \dot{h}_0\dot{g}^{\mathrm{T}} & \mathbf{M}_\gamma(V) - h_0\dot{l}^{\otimes 2}\end{bmatrix}\right), \tag{6}$$

and $h_0 = h(0 \mid V, \boldsymbol{\gamma})\mathrm{d}0$.

- The consistent rate is $\sqrt{N_{nz}}$ instead of $\sqrt{N}$.
- Treating non-zeros as ones forces $h_0 = 0$, making modeling non-zeros unrelated to zeros;
- it model zeros and non-zeros separately.

Outline

Subsampling zeros

Algorithm 1 zero reducing sampling

For $i = 1, ..., N$:

1. if $y_i = 0$,
   1. calculate $\pi(\boldsymbol{x}_i)$ and generate $u_i \sim \mathbb{U}(0,1)$;
   2. if $u_i \leq \pi(\boldsymbol{x}_i)$, include $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i) = \pi(\boldsymbol{x}_i)\}$ in the sample.
2. if $y_i \neq 0$, include $\{\boldsymbol{x}_i, y_i, \pi(\boldsymbol{x}_i, y_i) = 1\}$ in the sample;

- $\pi(\boldsymbol{x})$: sampling probability for the non-zeros.
- $\pi(\boldsymbol{x}_i, y_i) = y_i + (1 - y_i)\pi(\boldsymbol{x}_i)$: inclusion probability of $(\boldsymbol{x}_i, y_i)$.
- $\delta_i = 1$ if the i-th data point is selected and $\delta_i = 0$ otherwise.

Inverse probability weighting (IPW)

The selected subsample is biased. Consider the IPW estimator

$$\hat{\boldsymbol{\theta}}_{\text{ipw}}, \hat{\boldsymbol{\gamma}}_{\text{ipw}} = \arg\max_{\boldsymbol{\theta},\boldsymbol{\gamma}} \sum_{i=1}^{N} \delta_i \frac{\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{v}_i, y_i)}{\pi(\boldsymbol{x}_i, y_i)}. \tag{7}$$

Theorem 3.1

Let $r = \lim_{N\to\infty} N_{nz}/N_0^*$ *and* $\pi(\boldsymbol{x}) = \rho\varphi(\boldsymbol{x})$ *with* $\mathbb{E}\{\varphi(\boldsymbol{x})\} = 1$. *Under some moment assumptions,*

$$\sqrt{N_{nz}}(\hat{\boldsymbol{\eta}}_{\text{ipw}} - \boldsymbol{\eta}) \to \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\text{ipw}}). \tag{8}$$

where $\mathbf{V}_{\text{ipw}} = \mathbf{V}_{\text{full}} + r\boldsymbol{\Sigma}_{\text{full}}^{-1}\mathbf{V}_{\pi}\boldsymbol{\Sigma}_{\text{full}}^{-1}$, *and*

$$\mathbf{V}_{\pi} = \mathbb{E}\left\{\frac{e^{-2f}}{\varphi(\boldsymbol{x})}\begin{bmatrix}(1-h_0)\dot{g}\\ \dot{h}_0\end{bmatrix}^{\otimes 2}\right\}. \tag{9}$$

Optimal sampling probabilities

The $\varphi(\boldsymbol{x})$ that minimizes the variance inflation is

$$\varphi_{\mathrm{os}}(\boldsymbol{x}) = \frac{\|\text{Ł}\boldsymbol{\Sigma}_{\mathrm{full}}^{-1}\dot{\ell}(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{x}, 0)\|}{\mathbb{E}\{\|\text{Ł}\boldsymbol{\Sigma}_{\mathrm{full}}^{-1}\dot{\ell}(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{x}, 0)\|\}}. \tag{10}$$

- If $\text{Ł} = \mathbf{I}$, then $\varphi_{\mathrm{os}}(\boldsymbol{x})$ is A-optimal.
- If $\text{Ł} = \boldsymbol{\Sigma}_{\mathrm{full}}$, then $\varphi_{\mathrm{os}}(\boldsymbol{x})$ requires the least computational cost.
- $\varphi_{\mathrm{os}}(\boldsymbol{x})$ depends on unknown parameters, so a pilot estimate $\tilde{\boldsymbol{\eta}}$ is required.

The IPW is not efficient

$$\hat{\boldsymbol{\theta}}_{\text{ipw}}, \hat{\boldsymbol{\gamma}}_{\text{ipw}} = \arg\max_{\boldsymbol{\theta},\boldsymbol{\gamma}} \sum_{i=1}^{N} \delta_i \frac{\ell(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{v}_i, y_i)}{\pi(\boldsymbol{x}_i, y_i)}. \tag{11}$$

1. The IPW down-weights more informative data points.
2. A naive unweighted estimator is biased and inconsistent.

Likelihood based estimator

The conditional log-likelihood of $Y \mid \boldsymbol{x}, \delta = 1$ for the subsample is

$$\ell_{\text{cle}}(\boldsymbol{\eta}) = \sum_{i=1}^{N} \delta_i \Bigg[\log(1 + e^{-g_i} h_{0i}) I(y_i = 0) - (g_i - \log h_i) I(y_i \neq 0) \\ - \log\left\{(1 - h_{0i})e^{-g_i} + (1 + h_{0i}e^{-g_i})\pi(\boldsymbol{x}_i)\right\}\Bigg],$$

where $h_{0i} = h(0 \mid \boldsymbol{v}_i; \boldsymbol{\gamma})\mathrm{d}0$, $g_i = g(\boldsymbol{z}_i, \boldsymbol{\theta})$, and $h_i = h(y_i \mid \boldsymbol{v}_i; \boldsymbol{\gamma})$.

- Here, $\ell_{\text{cle}}(\boldsymbol{\eta})$ has an explicit expression.
- The conditional likelihood estimator is

$$\hat{\boldsymbol{\eta}}_{\text{cle}} = \arg\max_{\boldsymbol{\eta}} \ell_{\text{cle}}(\boldsymbol{\eta}). \tag{12}$$

Theoretical analysis of $\hat{\boldsymbol{\eta}}_{\text{cle}}$

Theorem 3.2

Under some moment assumptions,

$$\sqrt{N_{nz}}(\hat{\boldsymbol{\eta}}_{\text{cle}} - \boldsymbol{\eta}) \rightarrow \mathbb{N}(\mathbf{0},\ \mathbf{V}_{\text{cle}}), \tag{13}$$

where $\mathbf{V}_{\text{cle}} = \mathbb{E}[e^{-f}(1-h_0)]\boldsymbol{\Sigma}_{\text{cle}}^{-1}$, $\boldsymbol{\Sigma}_{\text{cle}} = \boldsymbol{\Sigma}_{\text{full}} - \boldsymbol{\Sigma}_I$, *and*

$$\boldsymbol{\Sigma}_I = r\mathbb{E}\left(\frac{e^{-2f}}{\frac{r\{1-h_0\}e^{-f}}{\mathbb{E}[\{1-h_0\}e^{-f}]} + \varphi(X)}\begin{bmatrix}(1-h_0)\dot{g}\\ \dot{h}_0\end{bmatrix}^{\otimes 2}\right). \tag{14}$$

Furthermore,

$$\mathbf{V}_{\text{cle}} \leq \mathbf{V}_{\text{ipw}} \tag{15}$$

The equality holds when $r = 0$ *and in this case* $\mathbf{V}_{\text{cle}} = \mathbf{V}_{\text{ipw}} = \mathbf{V}_{\text{full}}$.

CLE vs IPW

- The CLE has a higher estimation efficiency than the IPW.
- The CLE is less sensitive to the choice of $\varphi(\boldsymbol{x})$ and the pilot estimates.
- $\varphi_{\rm os}(\boldsymbol{x})$ is optimal for the IPW estimator, not for the CLE.
- An optimal $\varphi(\boldsymbol{x})$ for the CLE should be nonrandom binary and based on an optimal design.
- The CLE rely on the correct model assumption; it may not be consistent to $\hat{\boldsymbol{\eta}}_{\rm full}$ when the model is mis-specified.
- The IPW estimator is always consistent to $\hat{\boldsymbol{\eta}}_{\rm full}$, and thus may be preferred under model mis-specification.
- The full data MLE $\hat{\boldsymbol{\eta}}_{\rm full}$ minimizes the Kullback-Leibler distance between the mis-specified model class and the true model [5].

[5] White (1982)

Test for model correctness

- With a correct model, both $\hat{\boldsymbol{\eta}}_{\mathrm{ipw}}$, and $\hat{\boldsymbol{\eta}}_{\mathrm{cle}}$ estimate the true parameter.
- With model mis-specification, $\hat{\boldsymbol{\eta}}_{\mathrm{ipw}}$ is consistent to $\hat{\boldsymbol{\eta}}_{\mathrm{full}}$, and $\hat{\boldsymbol{\eta}}_{\mathrm{cle}}$ estimates something else.
- With a correct model,

$$\hat{\boldsymbol{\eta}}_{\mathrm{ipw}} - \hat{\boldsymbol{\eta}}_{\mathrm{cle}} \dot{\sim} \mathbb{N}(\mathbf{0},\ \mathbf{V}_T) \tag{16}$$

- Define the test statistics as

$$H = (\hat{\boldsymbol{\eta}}_{\mathrm{ipw}} - \hat{\boldsymbol{\eta}}_{\mathrm{cle}})^{\mathrm{T}} \hat{\mathbf{V}}_T^{-1} (\hat{\boldsymbol{\eta}}_{\mathrm{ipw}} - \hat{\boldsymbol{\eta}}_{\mathrm{cle}}) \dot{\sim} \chi^2_d, \tag{17}$$

- Use $\hat{\boldsymbol{\eta}}_{\mathrm{cle}}$ if H fails to reject and use $\hat{\boldsymbol{\eta}}_{\mathrm{ipw}}$ otherwise, i.e.,

$$\hat{\boldsymbol{\eta}}_{\mathrm{test}} = I(H \le \chi^2_{d,c}) \hat{\boldsymbol{\eta}}_{\mathrm{cle}} + I(H > \chi^2_{d,c}) \hat{\boldsymbol{\eta}}_{\mathrm{ipw}}. \tag{18}$$

Outline

Simulation setup

Working model: zero-inflated Poisson regression.

$$g(\boldsymbol{z};\boldsymbol{\theta}) = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\theta}. \tag{19}$$

$$h(y \mid \boldsymbol{v};\boldsymbol{\gamma}) = \frac{e^{-\mu}\mu^{y}}{y!} \text{ with } \mu = e^{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\gamma}}. \tag{20}$$

- Covariate $X = (Z^{\mathrm{T}}, V^{\mathrm{T}})^{\mathrm{T}} = \Sigma^{1/2}U$, where elements of U are i.i.d. with the following distributions.
  - a. standard normal: symmetric with light tails;
  - b. standard exponential: positively skewed;
  - c. t_5: symmetric with heavier tails;
- Full data sample size: $N = 5 \times 10^5$.
- Percentage of non-zeros is around 0.6%.
- Sampling rate as $\rho = 0.006, 0.01$, 0.02, and 0.04.
- A pilot sample of size 200 is used in each repetition of the simulation.

Methods considered

- Sampling probabilities:
  1. uni: uniform sampling
  2. opt: optimal sampling under L-optimality
- Estimation methods:
  - ipw: inverse probability weighting
  - lik: conditional likelihood
  - pre: the estimator base on a pre-test defined in (18).

Probabilities ($\times 10^2$) of a zero being selected

	uni	opt					
		Case 1		Case 2		Case 3	
n	0	**0**	**0**	**0**	**0**	**0**	**0**
600	0.60	0.67	0.60	0.81	0.60	0.81	0.60
1000	1.01	1.11	1.00	1.27	0.97	1.30	0.99
2000	2.01	2.22	2.01	2.34	1.85	2.43	1.94
4000	4.03	4.41	4.00	4.15	3.42	4.50	3.77

- **0**: rare zeros
- **0**: dominating zeros

Note

- The probability that maximizes the selection rare zeros is:

$$\pi(\boldsymbol{x}) \propto h(0 \mid \boldsymbol{v}; \boldsymbol{\gamma})\{1 - p_0(\boldsymbol{z}; \boldsymbol{\theta})\} \tag{21}$$

- It does not work well on parameter estimation.

MSE ($\times 10^2$) for estimating $\boldsymbol{\eta}$ with the correct model

Case 1: full data estimator MSE is 1.288

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	3.050	2.154	2.300	2.333	2.170	2.190
1000	2.362	1.901	1.948	1.963	1.893	1.897
2000	1.855	1.649	1.665	1.625	1.600	1.601
4000	1.538	1.456	1.463	1.422	1.416	1.418

Case 2: full data estimator MSE is 1.131

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	9.620	2.706	8.035	2.981	2.413	2.463
1000	6.380	2.283	5.079	2.541	2.113	2.155
2000	3.931	1.921	2.941	1.782	1.642	1.648
4000	2.380	1.566	1.919	1.436	1.412	1.413

Case 3: full data estimator MSE is 0.953

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	4.378	1.978	3.342	1.636	1.528	1.539
1000	3.107	1.712	2.494	1.365	1.297	1.297
2000	2.030	1.408	1.717	1.168	1.130	1.144
4000	1.552	1.236	1.352	1.065	1.057	1.057

Wrong model

The link function for generating the dominating zeros is the probit link instead of the logit link, i.e.,

$$g(\boldsymbol{z};\boldsymbol{\theta}) = \log\Phi(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\theta}) - \log\{1-\Phi(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\theta})\}, \tag{22}$$

where Φ is the standard normal distribution function.
And the non-zero generating distribution has an additional quadratic term. Specifically,

$$h(y \mid \boldsymbol{v};\boldsymbol{\gamma}) = \frac{e^{-\mu}\mu^{y}}{y!} \text{ with } \mu = e^{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\gamma}+\gamma_q v_4^2}. \tag{23}$$

- When the working model is wrong, the parameter $\boldsymbol{\eta}$ lose its meaning in this model class.
- The full data estimator $\hat{\boldsymbol{\eta}}_{\mathrm{full}}$ minimize the Kullback-Leibler distance between the working model and the true data generating model [6].

[6] White (1982)

MSE ($\times 10^2$) for approximating $\hat{\boldsymbol{\eta}}_{\text{full}}$ with a wrong model

Case 5: full data estimator MSE is 0.0

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	19.577	24.992	24.092	3.966	8.391	6.186
1000	11.361	19.116	15.583	2.581	5.931	4.190
2000	5.504	12.620	8.900	1.203	3.151	1.866
4000	2.535	7.288	4.394	0.530	1.537	0.739

Case 6: full data estimator MSE is 0.0

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	210.753	342.757	225.248	36.608	99.894	64.126
1000	107.914	265.119	124.876	26.827	76.177	46.264
2000	46.076	177.040	62.609	16.722	52.910	28.790
4000	19.286	110.633	29.460	8.969	33.945	15.823

Case 7: full data estimator MSE is 0.0

n	uni-ipw	uni-lik	uni-pre	opt-ipw	opt-lik	opt-pre
600	30.408	36.657	34.024	3.963	12.592	7.998
1000	17.542	28.729	21.718	2.440	8.804	4.409
2000	8.317	19.408	12.623	1.073	4.429	1.782
4000	4.054	12.362	7.683	0.572	2.336	0.850

The PANDOR dataset

- It contains information and clicks of users on Purch's high-tech websites over the ads showed to them for one month.
- The data available here [7] contains 48,602,664 events for 5,894,431 users, and the raw data file is over 160GB.
- Among the 5,894,431 users, about 4% of them clicked on the ads one or more times.
- We model the number of clicks using the working model in 19 and 20 with
  - Z_0 the intercept
  - Z_1 the number of pages viewed by the user
  - Z_2 the average number of keywords in the offers to the user
  - V_0 the intercept
  - V_1 the number of offers to the user

[7] https://archive.ics.uci.edu/dataset/460/pandor

MSE ($\times 10^4$) for approximate $\hat{\boldsymbol{\eta}}_{\text{full}}$

uni				**opt**			
ρ	**ipw**	**lik**	**pre**	**ipw**	**lik**	**pre**	ρ
0.044	1.333	91.461	1.333	0.205	84.397	0.205	0.040
0.089	0.722	45.194	0.722	0.089	23.802	0.089	0.076
0.177	0.267	17.129	0.267	0.042	5.059	0.042	0.142

Prediction on testing data

Prediction mean squared error (PMSE)

full data PMSE: 0.1083

uni				opt			
ρ	**ipw**	**lik**	**pre**	**ipw**	**lik**	**pre**	ρ
0.0443	0.1087	0.1397	0.1087	0.1083	0.1294	0.1083	0.0397
0.0886	0.1084	0.1316	0.1084	0.1083	0.1140	0.1083	0.0758
0.1771	0.1083	0.1236	0.1083	0.1083	0.1098	0.1083	0.1419

Prediction AUC

full data AUC: 0.6703

uni				opt			
ρ	**ipw**	**lik**	**pre**	**ipw**	**lik**	**pre**	ρ
0.0443	0.6694	0.6694	0.6694	0.6699	0.6576	0.6699	0.0397
0.0886	0.6699	0.6694	0.6699	0.6700	0.6585	0.6700	0.0758
0.1771	0.6699	0.6591	0.6699	0.6701	0.6585	0.6701	0.1419

Thank you!

Sidana, S., Laclau, C., and Amini, M.-R. (2018). Learning to recommend diverse items over implicit feedback on pandor. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 427–431.

Wang, H. (2020). Logistic regression for massive data with rare events. In H. D. III and A. Singh, eds., *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 9829–9836. PMLR.

Wang, H., Zhang, A., and Wang, C. (2021). Nonuniform negative sampling and log odds correction with rare events data. In *Proceedings of The 35 Conference on Neural Information Processing Systems (NeurIPS 2021).*, Proceedings of Machine Learning Research. PMLR.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.