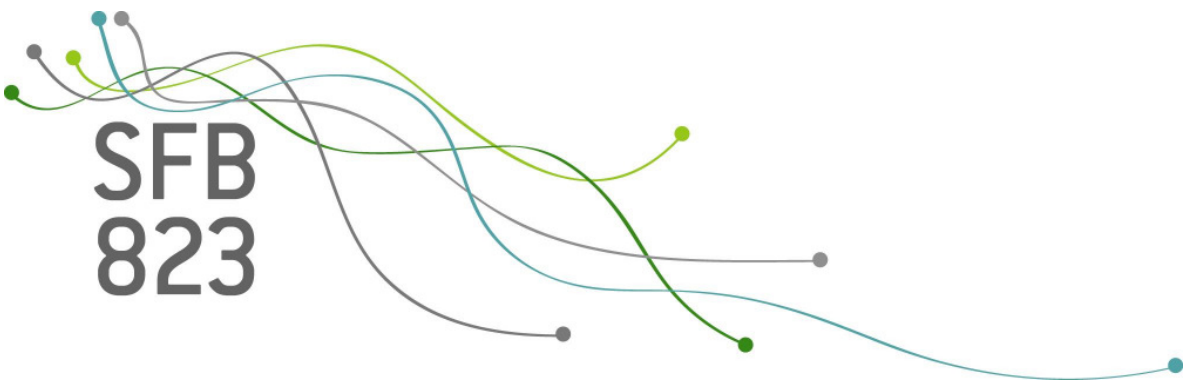
SFB 823

Reject inference in consumer credit scoring with nonignorable missing data

Michael Bücker, Walter Krämer

Nr. 1/2011

Discussion Paper



Reject inference in consumer credit scoring with nonignorable missing data ⋆

Michael Bücker [a], Walter Krämer [a]

[a] *Fakultät Statistik, TU Dortmund, D-44221 Dortmund, Germany*

Abstract

We generalize an empirical likelihood approach to missing data to the case of consumer credit scoring and provide a Hausman test for nonignorability of the missings. An application to recent consumer credit data shows that our model yields parameter estimates which are significantly different (both statistically and economically) from the case where customers who were refused credit are ignored.

Key words: Missing data; reject inference; credit scoring; logistic regression.

1 Introduction and summary

Statistical models for predicting defaults in the consumer credit industry and elsewhere suffer from the non-availability of default information for customers who were denied credit in the first place (Crook and Banasik 2004; Hand and Henley 1993). This does not matter as far as estimation goes if such data are missing at random (MAR) in the sense of Rubin (1976), i.e. if the probability of default, given all the relevant exogenous variables of the model, is the same whether an applicant is granted a credit or not. In applications, this can reasonably be assumed if creditors base their decision on the same statistical model (or a preliminary version thereof) which is to be estimated.

However, such procedures are illegal in many countries. In Germany, for instance, the federal data privacy act ("Bundesdatenschutzgesetz") explicitly forbids banks to grant consumer credits solely on the basis of a statistical

⋆ Financial support from Deutsche Forschungsgemeinschaft (SFB 823) is gratefully acknowledged. We thank Jing Qin for helpful discussions and comments.

Email addresses: buecker@statistik.tu-dortmund.de (Michael Bücker), walterk@statistik.tu-dortmund.de (Walter Krämer).

model – there must be some human judgement involved as well. This means that in practice, among applicants with otherwise identical sets of "official" explanatory variables, some may be granted a credit and some may not. Or technically speaking, the probability of not being granted a credit, given the observed regressors, is not the same as the probability given the observed regressors *and* future default information whenever human judgement adds any additional information on the latter. Therefore, data are missing not at random (MNAR) in the Rubin (1976) sense.

The present paper proposes a new approach to cope with this. It is based on Qin *et al.* (2002), who show how to reweight observations in the light of missing data, given a parametric model for the missings, using empirical likelihood (Owen 2001). In the context of a logistic regression model for defaults, we show that this reweighting delivers consistent and asymptotically normal parameter estimates even when credit decisions and defaults are still correlated, given all regressors. We also propose a Hausman test to check whether this dependency prevails. Finally we apply our model to a recent data set of 9,000 individuals applying for credit with a major German bank and show that it yields parameter estimates which are significantly different both in a statistical and in an economic sense.

2 Reweighting observations in the context of missing data

We consider iid data sets of the type $(Y_i, \boldsymbol{X}_i, R_i)$ $(i = 1, \ldots, N)$, where $Y_i = 1$ in case of default and $Y_i = 0$ in case of no default, $\boldsymbol{X}_i$ $(k \times 1)$ is a vector of regressors, and $R_i = 1$ if credit is granted and $R_i = 0$ if credit is denied. Without loss of generality, we assume that $R_i = 1$ for the first $n < N$ data sets, i.e. Y_i is missing for $i = n+1, \ldots, N$. We also assume that the dependence of Y on $\boldsymbol{X}$ can be described by a logistic regression model

$$P(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta}) := \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}. \quad (1)$$

Ignoring all data beyond n produces inconsistent ML-estimates for the model (1) whenever data are missing not at random in the sense that

$$P(R = 1|\boldsymbol{X}, Y) \neq P(R = 1|\boldsymbol{X}). \quad (2)$$

We now show, following Qin *et al.* (2002), how this inconsistency can be removed. To that purpose, let $F(y, \boldsymbol{x})$ be the joint distribution function of $(Y, \boldsymbol{X})$ (no parametric model is needed for this), let

$$w(y, \boldsymbol{x}, \boldsymbol{\theta}) := P(R = 1|Y, \boldsymbol{X}, \boldsymbol{\theta})$$

be some parametric model for observability, let $W := P(R = 1)$, and consider the following semiparametric likelihood for $\boldsymbol{\theta}$, W, and F:

$$L_n(\boldsymbol{\theta}, W, F) = \left[\prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)\right] \cdot (1 - W)^{N-n}. \tag{3}$$

This function is maximized under the constraints

$$\begin{aligned} & p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 0, \quad \sum_{i=1}^{n} p_i \left[\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}\right] = 0, \\ \text{and} \quad & \sum_{i=1}^{n} p_i \left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] = 0, \end{aligned} \tag{4}$$

where $p_i = dF(y_i, \boldsymbol{x}_i) = F(y_i, \boldsymbol{x}_i) - F_-(y_i, \boldsymbol{x}_i)$, i.e. p_i is the increase in the joint distribution function at $(y_i, \boldsymbol{x}_i)$ and $\boldsymbol{\mu}_{\boldsymbol{X}}$ is either the known expectation or the empirical mean of $\boldsymbol{X}$. By introducing Lagrange multipliers and profiling for all values of p_i, it is seen that

$$p_i = \frac{1}{n\left[1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)\right]},$$

where $\boldsymbol{\lambda}_1$ and λ_2 are Lagrange multipliers. Substituting p_i into (3) results in a profile likelihood that can be maximized numerically. Qin *et al.* (2002, Theorem 1) show that under mild regularity conditions, the resulting empirical likelihood estimates for $\boldsymbol{\theta}$ and W are consistent and asymptotically normal.

This however shall not concern us here. We are interested in the plug-in estimate $\hat{p}_i$ of p_i in order to reweight the likelihood derived from (1) to obtain

$$L_n^\star(\boldsymbol{\beta}) = \prod_{i=1}^{n} \hat{p}_i f(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}), \tag{5}$$

where $f(y_i | \boldsymbol{x}_i, \boldsymbol{\beta}) = \left[P(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta})\right]^{y_i} \cdot \left[1 - P(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{\beta})\right]^{1-y_i}$.

The conventional ML-estimator $\hat{\boldsymbol{\beta}}$, which ignores all missings, is the solution to (5) without the weights $\hat{p}_i$. Our main theoretical result is that maximizing (5) yields a consistent and asymptotically normal estimator $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ even in the case of (2), i.e. when missingness cannot be ignored.

Theorem 1 *Let $\boldsymbol{\beta}$ be from some compact subset of $\mathbb{R}^{k+1}$. Also, the marginal distribution of $\boldsymbol{X}$ must not depend on $\boldsymbol{\beta}$. Then, under mild additional regularity conditions to be specified in the appendix, the modified ML-estimator $\tilde{\boldsymbol{\beta}}$ is weekly consistent and*

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbf{0}}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{V}\right),$$

where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$.

The proof of this theorem and the description of the limiting covariance matrix $\boldsymbol{V}$ are in the appendix.

Table 1 provides some finite sample Monte Carlo evidence for $N = 10,000$, a common sample size in consumer credit scoring applications, $k = 1$, $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 4)$ and

$$P(Y_i = 1|X_i = x_i, \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (i = 1 \ldots, N).$$

The observability of Y is governed by

$$w(y_i, x_i, \boldsymbol{\theta}) = \frac{\exp(\theta_0 + \theta_1 y_i)}{1 + \exp(\theta_0 + \theta_1 y_i)} \quad (i = 1 \ldots, N). \tag{6}$$

In the table, we keep β_1, β_2 and θ_1 fixed at 2, -1, and 1, respectively, and report the empirical bias and the empirical mean square error for various values of the crucial parameter θ_0 which determines the proportion of missing data (the larger θ_0, the smaller the proportion of missing y's). 1,000 runs are performed for each parameter combination.

The table documents a considerable gain in efficiency for our new estimator when the percentage of missings is large, both in terms of bias and mean

Table 1. Bias and Mean Square Error of new and conventional parameter estimates (each multiplied by 1,000)

	θ_0 (resulting percentage of missings in parentheses)				
	-2 (76.5)	-1 (55.2)	0 (32.1)	1 (15.3)	2 (6.4)
bias($\hat{\beta}_1$)	818.276	618.419	380.993	187.864	77.567
bias($\tilde{\beta}_1$)	3.200	-2.432	2.108	1.275	-0.202
bias($\hat{\beta}_2$)	-3.757	-0.568	-0.356	-0.127	0.543
bias($\tilde{\beta}_2$)	-4.291	-0.786	-0.322	-0.151	0.538
var($\hat{\beta}_1$)	11.883	5.406	2.778	2.149	1.834
var($\tilde{\beta}_1$)	21.521	8.537	4.089	2.475	1.988
var($\hat{\beta}_2$)	3.476	1.459	0.906	0.664	0.599
var($\tilde{\beta}_2$)	3.645	1.492	0.919	0.666	0.600
mse($\hat{\boldsymbol{\beta}}$)	684.933	389.301	148.836	38.102	8.448
mse($\tilde{\boldsymbol{\beta}}$)	25.169	10.026	5.007	3.140	2.586

square error. This advantage becomes smaller as the percentage of nonignorable missings decreases, but the MSE of the standard ML-estimator is still more than three times the MSE of our modified estimator for a percentage of missings as small as 6.4%.

Similar results were obtained for other parameter combinations and can be obtained from the authors upon request.

A major drawback of the proposed estimation method is the non-identifiability in the case of too many covariates in the missing data process. More precisely, the parameter $\boldsymbol{\theta}$ of the missing data process must not have length larger than $k+1$ since the number of free parameters must not exceed the number of estimation equations in (4). Therefore, if $w(y, \boldsymbol{x}, \boldsymbol{\theta})$ is a logistic regression model and the missingness depends on Y we can only identify the parameters of $k-1$ covariates in addition to the intercept and the parameter of Y.

3 A Hausman test for nonignorability

If the y's are missing at random, the conventional ML-estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal and efficient with covariance matrix $\boldsymbol{V}^\star$, say, so

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V} - \boldsymbol{V}^\star)$$

(Hausman 1978). This follows from the fact that the difference $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ must be asymptotically uncorrelated with the modified ML-estimator $\tilde{\boldsymbol{\beta}}$ due to the efficiency of the conventional ML-estimator $\hat{\boldsymbol{\beta}}$. In the MNAR case, however, $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ is inconsistent, so the statistic

$$h := N\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)^\top (\boldsymbol{V} - \boldsymbol{V}^\star)^- \left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right), \tag{7}$$

where $(\quad)^-$ denotes some generalized inverse, provides a consistent test of the MAR null hypothesis.

Under the null and some regularity conditions, h is asymptotically χ^2, with degrees of freedom equal to the rank of $\boldsymbol{V} - \boldsymbol{V}^\star$. This leads to all sorts of complications in applications where $\boldsymbol{V} - \boldsymbol{V}^\star$ is singular, but the estimate for $\boldsymbol{V} - \boldsymbol{V}^\star$ that is used in finite samples for the statistic (7) has full rank nevertheless (Krämer and Sonnberger 1986). Also, the estimate of the differences of the covariance matrices can fail to be positive-definite (Schreiber 2008). We therefore propose to estimate the finite sample covariance matrix of $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$ via the bootstrap. Table 2 provides some Monte Carlo results. The model is the same as in section 2, with an additional regressor $X_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(0, 4)$ independent of the first and also of the missingness. Thus, as in section 2, we now allow

for random missings by generalizing (6) to

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp(\theta_0 + \theta_1 y_i + \theta_2 x_{1,i})}{1 + \exp(\theta_0 + \theta_1 y_i + \theta_2 x_{1,i})} \quad (i = 1 \ldots, N).$$

We let $\theta_2 = -1$ in all experiments. The MAR-case corresponds to $\theta_1 = 0$, the MNAR case corresponds to $\theta_1 = -1$. As in section 2, we let θ_0 vary across some range of values, and perform 1,000 runs for each parameter combination.

Table 2. Empirical rejection frequencies of the Hausman test for various percentages of missing data

a) MAR ($\theta_1 = 0$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	-2 (77.5)	-0.5 (57.5)	0.5 (42.5)	2 (22.5)
0.01	0.010	0.017	0.018	0.028
0.05	0.067	0.059	0.063	0.076
0.10	0.116	0.100	0.110	0.125

b) MNAR ($\theta_1 = -1$)

	θ_0 (resulting percentage of missings in parentheses)			
nominal significance level α	-2 (85.8)	0 (61.5)	1.5 (38.1)	3 (18.4)
0.01	0.265	0.956	0.996	1.000
0.05	0.481	0.989	1.000	1.000
0.10	0.596	0.995	1.000	1.000

4 An Application to consumer credit

Next we analyze 9,780 credit histories provided to us by a major German bank. For 4,000 clients the repayment status is known, all other clients have been denied credit. The lending institution holds information about the following covariates: age, marital status, number of children, job, working time, household income, potential bail, purchasing power of the area where the applicant lives, number of credits raised, a rating by the German General Credit Protection Agency "Schufa", new customer or not, housing type, and habitation time at the current address. The variables age (in years), working time (in

months), income (in Euro), and habitation time (in months) are metric, the other covariates are measured on a categorical scale.

The variable habitation time turns out to be independent of the missingness and can thus be disregarded as a covariate for the model w. Hence we are able to estimate the weights $\hat{p}_i$. By means of these weights we compute the estimators for $\boldsymbol{\beta}$. Table 3 reports the parameter estimates of a conventional logistic regression model ignoring all missings as well as the estimates resulting from our new approach. The results indicate that the new approach leads to significantly different estimates, for some variables even the sign of the estimate is reversed. For instance, working time has a negative effect on repayment in the conventional model whereas the effect is positive in our new model. Similarly, potential bail shows a positive effect in the standard model, which turns negative with our new estimator. A possible explanation for this reversal is the requirement of a co-signer only for clients with a high default risk.

To perform the Hausman test for nonignorability of the rejects, we need to estimate the covariance matrix of $\boldsymbol{q} := \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$. As the bootstrap is known to lead to poor results in models with categorical covariates we use a jackknife-estimator proposed by Shao (1992) and obtain

$$\widehat{\mathrm{Var}}_{\mathrm{J}}(\boldsymbol{q}) = \frac{n-(k+1)}{n} \sum_{i=1}^{n} (\boldsymbol{q}_{(i)} - \boldsymbol{q})(\boldsymbol{q}_{(i)} - \boldsymbol{q})^{\top},$$

where $\boldsymbol{q}_{(i)}$ comprises the estimators obtained from all but the ith observation. The resulting test statistic is 478.367 ($p = 0.000$), so the null hypothesis of nonignorability of the rejects is indeed rejected

The goodness of fit of both models can be compared by McFadden's R^2. The conventional model yields $R^2_{\mathrm{McF}} = 0.103$ and for the new model we have $R^2_{\mathrm{McF}} = 0.341$.

5 Discussion

The technique developed here can be applied to arbitrary parametric models of the dependency of Y on $\boldsymbol{X}$. Future research could include an enhancement of the method so that parameters for all covariates are identifiable in the missingness model. Additional applications for the new estimator can be imagined like clinical studies with nonrespondents or further inquiries where missing response occur.

Table 3. Conventional and new parameter estimates of the logistic regression model for creditworthiness. For categorical variables the reference class is given in parentheses.

		$\hat{\boldsymbol{\beta}}$	sd	$\tilde{\boldsymbol{\beta}}$	sd
Intercept		5.726	0.930	4.638	0.983
Age		-0.032	0.013	-0.053	0.012
Marital status	Widowed	0.264	0.873	0.997	0.806
(Married)	Single	-0.299	0.269	-0.585	0.287
	Cohabitee	1.632	1.030	3.134	1.178
	Divorced	-0.121	0.336	-0.053	0.370
	Separated	-0.663	0.424	-1.028	0.505
Children	1 child	0.007	0.271	0.240	0.333
(No child)	2 children	-0.450	0.293	-1.033	0.450
	3-6 children	-0.124	0.517	-0.754	0.325
Job	Employee	-0.826	0.447	-0.997	0.431
(Civil servant)	Crafter	-1.336	0.468	-1.788	0.301
	Self-Employed	-1.958	0.578	-3.296	0.618
	Retiree	0.635	1.321	3.459	1.229
	Other	-0.621	0.876	-1.433	1.240
Working time		0.003	0.001	-0.002	0.001
Household income		0.000	0.000	0.001	0.000
Bail (not available	available	0.102	0.284	-0.240	0.267
Purchasing power	high	-0.188	0.327	-0.178	0.391
(Very high)	medium	-0.219	0.311	-0.941	0.348
	low	-0.129	0.353	-0.067	0.414
	Other	-0.669	0.401	-2.248	0.385
Credits	2 credits	-0.119	0.259	-0.058	0.340
(1 credit)	3 credits	-0.113	0.312	-0.222	0.450
	4 or more	-0.365	0.290	-1.165	0.306
	no credit	-0.294	0.239	-1.360	0.287
	Other	-0.800	0.353	-1.574	0.403
Schufa score	B-E	-0.830	0.205	-0.922	0.229
(A-D)	F-J	-1.548	0.271	-2.741	0.295
	K-M	-4.202	0.615	-7.086	1.541
	P	-0.804	0.477	-2.143	0.393
New customer (no)	yes	0.068	0.329	-0.091	0.500
House type	1 family	0.340	0.496	0.959	0.741
(No family)	2 families	-0.165	0.506	0.394	0.691
	3-5 families	0.331	0.503	0.572	0.701
	6-10 families	-0.069	0.481	0.682	0.772
	11-14 families	0.124	0.543	0.875	0.770
	15-20 families	-0.165	0.633	-0.229	0.719
	>20 families	0.334	0.604	1.551	0.953
	Other	0.256	0.525	2.686	0.612
Habitation time		-0.001	0.001	0.001	0.001

References

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, **28**, 857–874.

Hand, D. J. and Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business & Industry*, **5**, 45–55.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**(6), 1251–1271.

Krämer, W. and Sonnberger, H. (1986). Computational pitfalls of the Hausman test. *Journal of Economic Dynamics and Control*, **10**(1-2), 163–165.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

Qin, J., Leung, D., and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, **97**(457), 193–200.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **36**(3), 581–592.

Schreiber, S. (2008). The Hausman test statistic can be negative even asymptotically. *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)*, **228**(4), 394–405.

Shao, J. (1992). Jackknifing in generalized linear models. *Annals of the Institute of Statistical Mathematics*, **44**(4), 673–686.

A Proof of theorem

Let

$$\boldsymbol{\Psi}^n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} n\widehat{p}_i \frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

be the derivative of the log of the likelihood (5). In addition, let

$$\boldsymbol{\psi}_i^n(\boldsymbol{\beta}) := n\widehat{p}_i \frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

and

$$\boldsymbol{s}_i^n(\boldsymbol{\beta}) := \frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

be functions in $\boldsymbol{\beta}$. From Qin *et al.* (2002) it can be seen that

$$n\widehat{p}_i \xrightarrow{p} \frac{W_0}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}.$$

Let E be the expectation with respect to $F(y,\boldsymbol{x})$ and E_C as the expectation with respect to the conditional distribution $w(y,\boldsymbol{x},\boldsymbol{\theta}_0)dF(y,\boldsymbol{x})/W_0$, where $\boldsymbol{\theta}_0$

and W_0 represent the true values of $\boldsymbol{\theta}$ and W respectively. Then it is easily verified that

$$\mathrm{E}_C\left(\frac{W_0}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}\frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)=\mathrm{E}\left(\frac{\partial \ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right).$$

Similar to Qin *et al.* (2002), let

$$\begin{aligned}\boldsymbol{\gamma}&=\boldsymbol{\lambda}_1(1-W), &\qquad \boldsymbol{\eta}&=(\boldsymbol{\theta}^\top,W,\boldsymbol{\gamma}^\top)^\top,\\ \boldsymbol{\eta}_0&=(\boldsymbol{\theta}_0^\top,W_0,\mathbf{0}^\top)^\top, &\qquad a_N&=\frac{N}{n}-\frac{1}{W_0}.\end{aligned}$$

Then

$$np_i=\frac{1-W}{1-\frac{W}{W_0}+\frac{1-W_0}{W_0}w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})+\boldsymbol{\gamma}^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_{\boldsymbol{X}})+a_N(w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W)}.$$

Defining

$$\boldsymbol{\Xi}^n(\boldsymbol{\beta},\boldsymbol{\eta},a_N):=\frac{1}{n}\sum_{i=1}^n\boldsymbol{\xi}_i(\boldsymbol{\beta},\boldsymbol{\eta},a_N) \tag{A.1}$$

where

$$\begin{aligned}&\boldsymbol{\xi}_i(\boldsymbol{\beta},\boldsymbol{\eta},a_N)\\ &:=\frac{1}{n}\sum_{i=1}^n\frac{1-W}{1-\frac{W}{W_0}+\frac{1-W_0}{W_0}w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})+\boldsymbol{\gamma}^\top(\boldsymbol{x}_i-\boldsymbol{\mu}_{\boldsymbol{X}})+a_N(w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W)}\boldsymbol{s}_i(\boldsymbol{\beta})\end{aligned}$$

we have

$$\boldsymbol{\Xi}^n(\boldsymbol{\beta},\widehat{\boldsymbol{\eta}},a_N)=\boldsymbol{\Psi}^n(\boldsymbol{\beta}).$$

Also, let $\Xi_j^n(\boldsymbol{\beta},\boldsymbol{\eta},a_N)$ denote the jth component of $\boldsymbol{\Xi}^n(\boldsymbol{\beta},\boldsymbol{\eta},a_N)$.

For the proof of theorem 1, we impose the following assumptions:

(A1) $\boldsymbol{\psi}^n(\boldsymbol{\beta})$ is asymptotically uniformly integrable,
(A2) $\mathrm{E}_C\left(\boldsymbol{\psi}_i^n(\boldsymbol{\beta})\right)$ is equicontinuous,
(A3) $\boldsymbol{\psi}^n(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ for almost all $y,\boldsymbol{x}$,
(A4) $\exists d(\boldsymbol{x},y)$ with $\mathrm{E}_C(d(\boldsymbol{X},Y))<\infty$ and $\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\|\le d(\boldsymbol{x},y)\ \forall\boldsymbol{\beta}$,
(A5) the operations of integration with respect to $y,\boldsymbol{x}$ and differentiation with respect to $\boldsymbol{\beta}$ can be interchanged,
(A6) $\mathrm{E}\left(\partial\ln f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}\right)$ has a unique root,
(A7) $\boldsymbol{\Psi}^n(\hat{\boldsymbol{\beta}})=o_p(1)$,
(A8) $\frac{\partial^2\Xi_j^n(\boldsymbol{\beta}^j,\boldsymbol{\eta}^j,a_n^j)}{\partial\boldsymbol{\beta}^\top\partial\boldsymbol{\beta}}$, $\frac{\partial^2\Xi_j^n(\boldsymbol{\beta}^j,\boldsymbol{\eta}^j,a_n^j)}{\partial\boldsymbol{\eta}^\top\partial\boldsymbol{\eta}}$ und $\frac{\partial^2\Xi_j^n(\boldsymbol{\beta}^j,\boldsymbol{\eta}^j,a_n^j)}{(\partial a_N)^2}$ exist and are bounded by an integrable function $\forall\ j$,
(A9) $\mathrm{E}\left(\frac{\partial s(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^\top}\right)$ exists and is invertible.

Proof of Theorem 1 We prove the consistency of the M-estimator $\hat{\boldsymbol{\beta}}$ by showing that under the additional conditions (A1)-(A7), $\boldsymbol{\Psi}^n(\boldsymbol{\beta})$ converges uniformly in probability to $\boldsymbol{\Psi}(\boldsymbol{\beta}) := \mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right)$ with unique root $\boldsymbol{\beta}_0$.

From the uniform integrability of $\boldsymbol{\psi}^n(\boldsymbol{\beta})$ and the equicontinuity of $\mathrm{E}_C\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ it follows that

$$\sup_{\boldsymbol{\beta}} \left\| \mathrm{E}_C\left(\boldsymbol{\psi}_i^n(\boldsymbol{\beta})\right) - \mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right) \right\| \xrightarrow{p} 0.$$

Also, by the uniform law of large numbers,

$$\sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta}) - \mathrm{E}_C\left(\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta})\right) \right] \right\| \xrightarrow{p} 0.$$

Finally, the consistence of $\hat{\boldsymbol{\beta}}$ follows from the fact that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}} \left\| \boldsymbol{\Psi}^n(\boldsymbol{\beta}) - \boldsymbol{\Psi}(\boldsymbol{\beta}) \right\| \\
=&\sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta}) - \mathrm{E}_C\left(\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta})\right) \right] + \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_C\left(\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta})\right) - \mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right) \right\| \\
\leq&\sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta}) - \mathrm{E}_C\left(\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta})\right) \right] \right\| + \sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_C\left(\boldsymbol{\psi}_{\boldsymbol{i}}^n(\boldsymbol{\beta})\right) - \mathrm{E}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right) \right\| \xrightarrow{p} 0,
\end{aligned}$$

where the second term converges as it contains a Cesàro mean.

The normality of the estimator can be derived by a componentwise Taylor expansion of (A.1) similar to the proof of Theorem 2 in Qin *et al.* (2002). We have

$$\begin{aligned}
0 &= \Xi_j^n(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}, a_N) \\
&= \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} \left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} \left(a_N - 0 \right) + o_p(1) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) + o_p(N^{-\frac{1}{2}}) \\
&= \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) + \left(\frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} + o_p(1) \right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} \left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \right) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} \left(a_N - 0 \right) + o_p\left(N^{-\frac{1}{2}} \right),
\end{aligned}$$

and so from the central limit theorem

$$\sqrt{N} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) = \sqrt{N} \boldsymbol{C}_n^{-1} \boldsymbol{\zeta}_n + o_p(1),$$

where

$$\boldsymbol{\zeta}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial\boldsymbol{\eta}^\top}\left(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial a_N}\left(a_N-0\right)\right]$$

and

$$\boldsymbol{C}_n = -\left(\frac{\partial\boldsymbol{\Xi}^n(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial\boldsymbol{\beta}^\top} + o_p(1)\right) \xrightarrow{p} \mathrm{E}\left(-\frac{\partial^2 \ln f(Y|\boldsymbol{X},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right) =: \boldsymbol{F}.$$

In addition,

$$\begin{aligned}
\boldsymbol{\zeta}_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial\boldsymbol{\eta}^\top}\left(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial a_N}\left(a_N-0\right)\right] \\
&= \frac{1}{W_0}\frac{1}{N}\sum_{i=1}^{N} r_i\left(\frac{W_0}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right) + \boldsymbol{b}_1\boldsymbol{U}_N^{-1}\boldsymbol{\phi}_N + \boldsymbol{b}_2 a_N + o_p\left(N^{-\frac{1}{2}}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} r_i\left(\frac{\boldsymbol{s}_i(\boldsymbol{\beta}_0)}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}\right) + \frac{1}{N}\sum_{i=1}^{N}\left[r_i\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{g}_i + \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h}+\boldsymbol{b}_2\right)a_N\right] + o_p\left(N^{-\frac{1}{2}}\right),
\end{aligned}$$

where

$$\boldsymbol{b}_1 = \mathrm{E}_C\left(\frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial\boldsymbol{\eta}^\top}\right) \quad \text{und} \quad \boldsymbol{b}_2 = \mathrm{E}_C\left(\frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial a_N}\right)$$

and $\boldsymbol{g}_i$, $\boldsymbol{U}$ and $\boldsymbol{h}$ given as in Qin *et al.* (2002). That means the components of $\boldsymbol{g}_i$ are given by

$$\begin{aligned}
\boldsymbol{g}_i^1(\boldsymbol{\eta},a_N) &= \frac{\partial \ln w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\
&- \frac{\left[a_N + \frac{1-W_0}{W_0}\right]\frac{\partial w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}}{1-\frac{W}{W_0}+\left(\frac{1}{W_0}-1\right)w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})+\boldsymbol{\gamma}^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})+a_N\left(w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W\right)},
\end{aligned}$$

$$\begin{aligned}
&g_i^2(\boldsymbol{\eta},a_N) \\
=&\frac{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W}{1-\frac{W}{W_0}+\left(\frac{1}{W_0}-1\right)w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})+\boldsymbol{\gamma}^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})+a_N\left(w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W\right)},
\end{aligned}$$

and

$$\begin{aligned}
&\boldsymbol{g}_i^3(\boldsymbol{\eta},a_N) \\
=&\frac{\boldsymbol{x}_i-\boldsymbol{\mu}_{\boldsymbol{X}}}{1-\frac{W}{W_0}+\left(\frac{1}{W_0}-1\right)w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})+\boldsymbol{\gamma}^\top(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{X}})+a_N\left(w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta})-W\right)},
\end{aligned}$$

the matrix $\boldsymbol{U}$ is defined by

$$\boldsymbol{U} = W_0\cdot\mathrm{E}_C(\boldsymbol{T})$$

with

$$\boldsymbol{T} = \begin{pmatrix} \mathbf{0} & \frac{\partial w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}}{(1-W_0)w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} & -\frac{W_0(\boldsymbol{X}-\boldsymbol{\mu_X})\partial w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}}{(1-W_0)w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} \\ -\frac{W_0^2\partial w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}}{(1-W_0)w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} & -\frac{W_0^2(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)-1)}{(1-W_0)^2w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} & \frac{W_0^2(\boldsymbol{X}-\boldsymbol{\mu_X})(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)-W_0)}{(1-W_0)^2w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} \\ \frac{W_0(\boldsymbol{X}-\boldsymbol{\mu_X})\partial w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}}{(1-W_0)w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} & -\frac{W_0(\boldsymbol{X}-\boldsymbol{\mu_X})}{(1-W_0)^2w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} & \frac{W_0^2(\boldsymbol{X}-\boldsymbol{\mu_X})^2}{(1-W_0)^2w^2(Y,\boldsymbol{X},\boldsymbol{\theta}_0)} \end{pmatrix}$$

and finally

$$\boldsymbol{h} = -\frac{W_0^3}{(1-W_0)^2}\mathrm{E}_C\begin{pmatrix} \frac{1-W_0}{(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0))^2}\frac{\partial w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} \\ \frac{(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)-W_0)^2}{(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0))^2} \\ \frac{(\boldsymbol{X}-\boldsymbol{\mu_X})(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)-W_0)}{(w(Y,\boldsymbol{X},\boldsymbol{\theta}_0))^2} \end{pmatrix}.$$

Therefore, we have

$$\sqrt{N}\boldsymbol{\zeta}_n \xrightarrow{P} \mathcal{N}(\mathbf{0}, \boldsymbol{H}),$$

where

$$\boldsymbol{H} = \mathrm{Var}\left(R\left(\frac{\boldsymbol{s}(\boldsymbol{\beta}_0)}{w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}\right) + R\boldsymbol{d}_1\boldsymbol{g}_i + \boldsymbol{d}_2\frac{1}{W_0}\left(1-\frac{R}{W_0}\right)\right)$$

and

$$\boldsymbol{d}_1 = \boldsymbol{b}_1\boldsymbol{U}^{-1} \quad \text{und} \quad \boldsymbol{d}_2 = \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h} + \boldsymbol{b}_2\right).$$

Finally, we get

$$\sqrt{N}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{p} \mathcal{N}(\mathbf{0}, \boldsymbol{V})$$

with $\boldsymbol{V} = \boldsymbol{F}^{-1}\boldsymbol{H}\boldsymbol{F}^{-1}$. □

In the proof of Theorem 1 the exact form of $f(y|\boldsymbol{x},\boldsymbol{\beta})$ need not to be a logistic regression model but can be rather arbitrary as long as it meets the assumptions. For a logistic regression, however, (A1), (A2), and (A4) can be simplified. In this case we have

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right).$$

so that the uniform integrability follows from $\mathrm{E}_C\left((n\widehat{p})^2\left\|\boldsymbol{X}\right\|^2\right) < \infty$ because

$$\mathrm{E}_C\left(\left\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\right\|^2\right) = \mathrm{E}_C\left(\left\|n\widehat{p}\boldsymbol{X}\underbrace{\left(Y - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}\right)}_{\in(-1,1)}\right\|^2\right) \leq \mathrm{E}_C\left((n\widehat{p})^2\left\|\boldsymbol{X}\right\|^2\right).$$

In a similar manner one can show that (A4) holds if

$$\mathrm{E}_C\left((n\widehat{p})\,\|\boldsymbol{X}\|\right) < \infty$$

and (A2) is valid if there exists an $M < \infty$, so that for all $n \in \mathbb{N}$

$$\mathrm{E}_C\left(n\widehat{p}\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)\right) \leq M,$$

where $\lambda_{\max}(\boldsymbol{A})$ denotes the greatest eigenvalue of $\boldsymbol{A}^\top\boldsymbol{A}$.