Web-based Supplementary Materials for "On Assessing Binary Regression Models Based on Ungrouped Data" by Lu and Yang

Chunling Lu*[1] and Yuhong Yang†[2]

[1]Division of Global Health, Brigham and Women's Hospital, & Department of Global Health and Social Medicine, Harvard University
[2]School of Statistics, University of Minnesota

1 Web Appendix A: Assumptions for Theorem 1

We state below the assumptions for Theorem 1 in Section 2 of the main paper.

Define the L_q norm $\|f\|_q = \left(\int |f(x)|^q P_X(dx)\right)^{1/q}$ for $1 \le q < \infty$.

Assumption 1.1 *The parameter space Θ and sample space $\mathcal{X}$ of the covariates are such that there exists a small positive constant $c < 1/2$ such that $c \le f(x;\theta) \le 1-c$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$. The nonparametric estimators of f are also assumed to be between c and $1-c$ with probability 1.*

This assumption is often made to analyze parametric and nonparametric estimators, e.g., in van der Lann, Dudoit and Keles (2004).

Recall that for a positive sequence a_n, an estimator $\{\widehat{f}_n\}_{n=1}^{\infty}$ is said to converge exactly at rate $\{a_n\}$ in probability under the L_2 loss if (i) $\left\|f - \widehat{f}_n\right\|_2 = O_p(a_n)$, and (ii) for every $0 < \epsilon < 1$, there exists $c_\epsilon > 0$ such that when n is large enough, $P\left(\left\|f - \widehat{f}_n\right\|_2 \ge c_\epsilon a_n\right) \ge 1 - \epsilon$.

Assumption 1.2 *Under H_0 and H_1, the nonparametric estimator $\widehat{f}_{n,2}$ converges exactly at rate p_n and q_n respectively in probability under the L_2 loss, where $p_n \to 0$, $q_n \to 0$ and $\sqrt{n}\min(p_n, q_n) \to \infty$. Under H_0, the parametric estimator converges exactly at rate $1/\sqrt{n}$ under the L_2 loss, and under H_1, it does not converge in that the L_2 loss is bounded away from zero in probability.*

The assumption is natural because nonparametric estimators converge to the true function under very mild conditions, but at a rate slower than the parametric rate $1/\sqrt{n}$, under a quadratic-type of loss. For

*E-mail: chunling_lu@hms.harvard.edu
†E-mail: yyang@stat.edu.au

the parametric estimator, if the model is correct, then typically the estimator is at the parametric rate, but if the model is wrong, then it does not converge at all. See e.g., Devroye, Györfi and Lugosi (1996); Yang (1999); Biau (2012) and references given there for general results on rates of convergence for estimating f for parametric and nonparametric situations. In particular, convergences of histogram, tree-based methods (such as random forest) as well as those based on series expansion are presented.

Assumption 1.3 *For both the parametric and the nonparametric estimators, we have* $\left\|f-\widehat{f}_{n,j}\right\|_4 / \left\|f-\widehat{f}_{n,j}\right\|_2 = O_p(1)$ *for* $j=1,2$.

This regularity assumption is used to avoid unruly tail behaviors of the estimators.

2 Web Appendix B: Proof of Theorem 1

It suffices to prove the result with a single data splitting, as shown below. Suppose that for a single splitting, we have, for instance, $P(CV(\widehat{f}_{n_1,1}) \geq CV(\widehat{f}_{n_1,2})) \to 1$ (the other case follows from the same argument). Then due to the iid nature of the observations, we must have $E\left(\sum_{\pi\in\Pi}\tau_\pi/|\Pi|\right) \to 1$. With $\sum_{\pi\in\Pi}\tau_\pi/|\Pi|$ being between 0 and 1, for its expectation to converge to 1, it is necessary to have $\sum_{\pi\in\Pi}\tau_\pi/|\Pi| \to 1$ in probability. As a result, $P\left(\sum_{\pi\in\Pi}\tau_\pi \geq |\Pi|/2\right) \to 1$. Below we prove the conclusion of Theorem 1 with a single data splitting.

We first show that the probability of type I error goes to zero as $n\to\infty$.

Let $H(f,g)(x) = f(x)\log\frac{f(x)}{g(x} + (1-f(x))\log\frac{1-f(x)}{1-g(x)}$ and

$$\begin{aligned} L_n(f,g) &= \sum_{i=n_1+1}^{n}\left(f(X_i)\log\frac{f(X_i)}{g(X_i)} + (1-f(X_i))\log\frac{1-f(X_i)}{1-g(X_i)}\right) \\ &= \sum_{i=n_1+1}^{n} H(f,g)(X_i). \end{aligned}$$

Then conditional on Z^1 and $X^2 = (X_{n_1+1},...,X_n)$, assuming, for a moment, $L_n(f,\widehat{f}_{n_1,2}) > L_n(f,\widehat{f}_{n_1,1})$, by

Chebyshev's inequality, we have

$$
\begin{aligned}
& P\left(CV(\widehat{f}_{n_1,1}) < CV(\widehat{f}_{n_1,2}) \,|Z^1, X^2\right) \\
= \;& P\left(\sum_{i=n_1+1}^{n}\left(Y_i \log\frac{\widehat{f}_{n_1,2}(X_i)}{\widehat{f}_{n_1,1}(X_i)} + (1-Y_i)\log\frac{(1-\widehat{f}_{n_1,2}(X_i))}{(1-\widehat{f}_{n_1,1}(X_i))}\right) + (L_n(f,\widehat{f}_{n_1,2}) - L_n(f,\widehat{f}_{n_1,1}))\right. \\
& \left. > (L_n(f,\widehat{f}_{n_1,2}) - L_n(f,\widehat{f}_{n_1,1}))\,|Z^1, X^2\right) \\
\leq \;& \min\left(1, \frac{\sum_{i=n_1+1}^{n} f(X_i)(1-f(X_i))\left(\log\frac{\widehat{f}_{n_1,2}(X_i)(1-\widehat{f}_{n_1,1}(X_i))}{\widehat{f}_{n_1,1}(X_i)(1-\widehat{f}_{n_1,2}(X_i))}\right)^2}{\left(L_n(f,\widehat{f}_{n_1,2}) - L_n(f,\widehat{f}_{n_1,1})\right)^2}\right). \qquad (1)
\end{aligned}
$$

Thus to show the probability of type I error goes to zero, we only need to prove i) $L_n(f,\widehat{f}_{n_1,2}) > L_n(f,\widehat{f}_{n_1,1})$ occurs with probability going to 1; ii) the last expression above converges to zero in probability. Note that

$$
\log\frac{\widehat{f}_{n_1,2}(X_i)(1-\widehat{f}_{n_1,1}(X_i))}{\widehat{f}_{n_1,1}(X_i)(1-\widehat{f}_{n_1,2}(X_i))} = \log\frac{\widehat{f}_{n_1,2}(X_i)}{f(X_i)} - \log\frac{\widehat{f}_{n_1,1}(X_i)}{f(X_i)} + \log\frac{1-\widehat{f}_{n_1,1}(X_i)}{1-f(X_i)} - \log\frac{1-\widehat{f}_{n_1,2}(X_i)}{1-f(X_i)}.
$$

Applying the simple inequality $\log x \leq x-1$ for $x>0$, together with the assumption that f and its estimates are all bounded between c and $1-c$, it can be easily verified that

$$
\left|\log\frac{\widehat{f}_{n_1,2}(X_i)(1-\widehat{f}_{n_1,1}(X_i))}{\widehat{f}_{n_1,1}(X_i)(1-\widehat{f}_{n_1,2}(X_i))}\right| \leq \frac{2\left|f(X_i)-\widehat{f}_{n_1,2}(X_i)\right| + 2\left|f(X_i)-\widehat{f}_{n_1,1}(X_i)\right|}{c}.
$$

Together with $x(1-x)\leq 1/4$, it follows that the numerator in the last expression in Equation (1) is upper bounded by $2/c^2$ times

$$
\sum_{i=n_1+1}^{n}\left((f(X_i)-\widehat{f}_{n_1,1}(X_i))^2 + (f(X_i)-\widehat{f}_{n_1,2}(X_i))^2\right).
$$

Now $L_n(f,\widehat{f}_{n_1,j}) = \sum_{i=n_1+1}^{n} H(f,\widehat{f}_{n_1,j})(X_i)$, where $H(f,\widehat{f}_{n_1,j})(X_i)$ has mean $L(f,\widehat{f}_{n_1,j})$ conditional on Z^1. Consider $W = H(f,\widehat{f}_{n_1,j})(X) - L(f,\widehat{f}_{n_1,j})$. Let E_{Z^1} (and P_{Z^1}) denote the conditional expectation given Z^1. Then $E_{Z^1}W = 0$, $Var_{Z^1}W \leq E_{Z^1}H^2(f,\widehat{f}_{n_1,j})(X)$ and it is readily seen that $|W| \leq 2\log\frac{1-c}{c}$ under Assumption 1. For $c\leq p,q\leq 1-c$, it can be verified that

$$
c_2(p-q)^2 \leq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \leq c_1(p-q)^2,
$$

where c_1 and c_2 are positive constants that depend only on c. Thus

$$Var_{Z^1} W \leq c_1^2 E_{Z^1} \left(f(X) - \widehat{f}_{n_1,j}(X) \right)^4 = c_1^2 \left\| f - \widehat{f}_{n_1,j} \right\|_4^4.$$

Then by Bernstein's inequality (see, e.g., Pollard (1984, page 193)), for each $x > 0$, we have

$$\begin{aligned} & P_{Z^1} \left(L_n(f, \widehat{f}_{n_1,1}) - n_2 L(f, \widehat{f}_{n_1,1}) \geq x \right) \\ \leq & \exp\left(-\frac{1}{2} \frac{x^2}{n_2 c_1^2 \left\| f - \widehat{f}_{n_1,1} \right\|_4^4 + \frac{2x \log \frac{1-c}{c}}{3}} \right) \\ \leq & \exp\left(-\frac{1}{2} \frac{Ax^2}{n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_4^4 + \frac{2x}{3}} \right), \end{aligned}$$

for some constant $A > 0$ that depends only on c. Note that

$$L(f, \widehat{f}_{n_1,1}) = \int \left(f(x) \log \frac{f(x)}{\widehat{f}_{n_1,1}(x)} + (1 - f(x)) \log \frac{1 - f(x)}{1 - \widehat{f}_{n_1,1}(x)} \right) P_X dx \leq c_1 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2.$$

Then, with $x = \beta n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2$ for some $\beta > 0$, we have

$$\begin{aligned} & P_{Z^1} \left(L_n(f, \widehat{f}_{n_1,1}) \geq (c_1 + \beta) n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2 \right) \\ \leq & P_{Z^1} \left(L_n(f, \widehat{f}_{n_1,1}) - n_2 L(f, \widehat{f}_{n_1,1}) \geq \beta n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2 \right) \\ \leq & \exp\left(-\frac{1}{2} \frac{A\beta^2 n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^4}{\left\| f - \widehat{f}_{n_1,1} \right\|_4^4 + \frac{2\beta}{3} \left\| f - \widehat{f}_{n_1,1} \right\|_2^2} \right). \end{aligned}$$

For the other estimator, for $x > 0$, similarly, we have

$$\begin{aligned} & P_{Z^1} \left(L_n(f, \widehat{f}_{n_1,2}) - n_2 L(f, \widehat{f}_{n_1,2}) \leq -x \right) \\ \leq & \exp\left(-\frac{1}{2} \frac{x^2}{n_2 c_1^2 \left\| f - \widehat{f}_{n_1,2} \right\|_4^4 + \frac{2x \log \frac{1-c}{c}}{3}} \right), \end{aligned}$$

and since $L(f, \widehat{f}_{n_1,2}) \geq c_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2$,

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,2}) - n_2 c_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2 \leq -x \right) \\ \leq \quad & \exp\left(-\frac{1}{2} \frac{Ax^2}{n_2 \left\| f - \widehat{f}_{n_1,2} \right\|_4^4 + \frac{2x}{3}} \right). \end{aligned}$$

Then, with $x = \widetilde{\beta} c_2 n_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2$ for some $0 < \widetilde{\beta} < 1$, we have

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,2}) \leq (1-\widetilde{\beta}) n_2 c_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2 \right) \\ \leq \quad & \exp\left(-\frac{1}{2} \frac{A\widetilde{\beta}^2 c_2^2 n_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^4}{\left\| f - \widehat{f}_{n_1,2} \right\|_4^4 + \frac{2c_2\widetilde{\beta}}{3} \left\| f - \widehat{f}_{n_1,2} \right\|_2^2} \right). \end{aligned}$$

From above, we know $L_n(f, \widehat{f}_{n_1,1}) = O_p(n_2/n_1) = O_p(1)$ and $L_n(f, \widehat{f}_{n_1,2})$ is lower bounded in probability by $n_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2$, which is of a higher order than n_2/n_1 and thus approaches ∞ in probability. Furthermore, $\sum_{i=n_1+1}^{n} \left((f(X_i) - \widehat{f}_{n_1,1}(X_i))^2 + (f(X_i) - \widehat{f}_{n_1,2}(X_i))^2 \right)$ is of order $n_2 \left\| f - \widehat{f}_{n_1,2} \right\|_2^2$ in probability. Putting things together, under the assumptions of the theorem, the two conditions after Equation (1) are satisfied. We conclude that under the null hypothesis, the probability of type I error indeed converges to zero.

Now we handle the case under H_1, i.e., we show that the power of the DRYV method is asymptotically 1. As before, for $x > 0$, we have

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,1}) - n_2 L(f, \widehat{f}_{n_1,1}) \leq -x \right) \\ \leq \quad & \exp\left(-\frac{1}{2} \frac{x^2}{n_2 c_1^2 \left\| f - \widehat{f}_{n_1,1} \right\|_4^4 + \frac{2x \log \frac{1-c}{c}}{3}} \right). \end{aligned}$$

Because $L(f, \widehat{f}_{n_1,1}) \geq c_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2$, with $x = \widetilde{\beta}_1 c_2 n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2$ for some $0 < \widetilde{\beta}_1 < 1$, we know

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,1}) \leq (1-\widetilde{\beta}_1) n_2 c_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^2 \right) \\ \leq \quad & \exp\left(-\frac{1}{2} \frac{A\widetilde{\beta}_1^2 c_2^2 n_2 \left\| f - \widehat{f}_{n_1,1} \right\|_2^4}{\left\| f - \widehat{f}_{n_1,1} \right\|_4^4 + \frac{2c_2\widetilde{\beta}_1}{3} \left\| f - \widehat{f}_{n_1,1} \right\|_2^2} \right). \end{aligned}$$

For the nonparametric estimator, similarly as before, we have for $x > 0$,

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,2}) - n_2 L(f, \widehat{f}_{n_1,2}) \geq x\right) \\ \leq & \exp\left(-\frac{1}{2}\frac{Ax^2}{n_2\left\|f - \widehat{f}_{n_1,2}\right\|_4^4 + \frac{2x}{3}}\right). \end{aligned}$$

Again, since $L(f, \widehat{f}_{n_1,2}) \leq c_1\left\|f - \widehat{f}_{n_1,2}\right\|_2^2$, with $x = \beta_1 n_2\left\|f - \widehat{f}_{n_1,2}\right\|_2^2$ for some $\beta_1 > 0$, we have

$$\begin{aligned} & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,2}) \geq (c_1 + \beta_1) n_2\left\|f - \widehat{f}_{n_1,2}\right\|_2^2\right) \\ \leq & P_{Z^1}\left(L_n(f, \widehat{f}_{n_1,2}) - n_2 L(f, \widehat{f}_{n_1,2}) \geq \beta_1 n_2\left\|f - \widehat{f}_{n_1,2}\right\|_2^2\right) \\ \leq & \exp\left(-\frac{1}{2}\frac{A{\beta_1}^2 n_2\left\|f - \widehat{f}_{n_1,2}\right\|_2^4}{\left\|f - \widehat{f}_{n_1,2}\right\|_4^4 + \frac{2\beta_1}{3}\left\|f - \widehat{f}_{n_1,2}\right\|_2^2}\right). \end{aligned}$$

Thus, under H_1, $L_n(f, \widehat{f}_{n_1,1})$ is lower bounded in probability by order n_2 while $L_n(f, \widehat{f}_{n_1,2})$ is upper bounded in probability by $n_2\left\|f - \widehat{f}_{n_1,2}\right\|_2^2 = o_p(n_2)$. Thus the relative performance between the two estimators is reversed from the earlier H_0 case. This completes the proof of Theorem 1.

Remark 2.1 *The theoretical result holds regardless of how the multiple data splittings are done. Based on our simulations, 100 random data splittings typically work very well.*

Remark 2.2 *Consistency of CV is studied in Yang (2006) also in the context of binary regression. However, it focuses only on classification error, which is related to but improper for assessing the goodness of fit of a model.*

Remark 2.3 *From the proof above, since the parametric estimator of the conditional probability function f based on the model to be assessed and the nonparametric estimator have different rates of convergence under H_0 and H_1 respectively, we know that for a single data splitting based CV comparison, $CV(\widehat{f}_{n_1,1})$ and $CV(\widehat{f}_{n_1,2})$ are in the right order with probability going to one. Then for any cutoff $c \in (0,1)$, if we accept H_0 when $\sum_{\pi\in\Pi}\tau_\pi/|\Pi| \geq c$ (and reject H_0 otherwise), then the same conclusion of Theorem 1 holds, i.e., under H_0, the probability of type I error goes to 0 and under H_1, the power of the test goes to 1.*

Remark 2.4 *In the theoretical examination of the performances of the MTA and the chosen nonparametric method through CV, in the asymptotic sense, their relative performances at the full sample size n and the*

reduced sample size n_1 (due to data splitting) stay the same in ordering (i.e., under no lack of fit of the MTA, the MTA beats the nonparametric method with high probability at both sample sizes; under a lack of fit, the MTA is beaten by the nonparametric method with high probability at both sample sizes. Hence, in the limit sense, the DRYV approach works. In a finite sample situation, however, the sample size reduction from n to n_1 can lead to the opposite relative performances of the two competitors. See e.g., Burman (1990) and Arlot and Celisse (2010) for work and more references on correcting this bias of CV for the purpose of **risk estimation**. *Note, however, it is unclear how better risk estimation there is related to the capability of assessing the parametric model in our context.*

3 Web Appendix C: Additional analysis on the low birth weight data

In this section, we provide some additional results on the models/methods considered for the low birth weight data in Section 4 of the main paper. Recall that both AIC and BIC models are evaluated. As is well-known, BIC is consistent in selecting the true model if it is among the candidates. On this ground, many researchers prefer the use of BIC when the goal is to identify the best parametric model (instead of seeking the best predictive performance). However, as shown in Yang (2005) and Liu and Yang (2011), the choice between BIC and AIC should not be dictated by the asymptotic properties they may or may not have. What matters for the application is whether the data generating process is better described as practically parametric or practically nonparametric at the actual sample size.

In the main paper, DRYV pairwise comparison tables at 50% and 90% training fractions are given to provide very insightful information. To see the trend there more clearly, the table for training fraction of 75% (141:48) is given here, which is indeed consistent with and supportive of the observations there.

Table 1: DRYV Pairwise Comparison with 75% Training. The (i, j)-th entry records the percentages of times among the random data splittings the i-th model/method has a strictly larger predictive likelihood than the j-th model/method.

	Age-only	BIC	AIC	RF	BAG
Age-only	0	19.7	25.0	50.0	73.7
BIC	80.3	0	43.3	82.0	93.3
AIC	75.0	56.7	0	84.3	94.7
RF	50.0	18.0	15.7	0	90.0
BAG	26.3	6.7	5.3	10.0	0

Besides pairwise comparisons, a multi-horse racing done in the same spirit of DRYV can add further value for the understanding of the relative performances of the models/methods. Applying the DRYV method on

the three models together with random forest and bagging provide valuable information, with 300 data splittings at 94:95, 141:48 and 170:19 ratios of training versus evaluation, the number of times a method gives the highest predictive likelihood in each case is recorded in Table 2 for the five competitors:

Table 2: Percentages of Having the Highest Predictive Likelihood

Training Fraction	Age-only	BIC	AIC	RF	BAG
50%	12.7	38.3	40.0	8.3	0.7
75%	12.7	33.3	47.7	6.3	0.0
90%	16.0	25.3	48.7	8.0	2.0

The simple table from the DRYV approach immediately tells that the nonparametric methods are not competitive at all, strongly supporting that a parametric approach based on logistic regression is suitable for the data. The top competitors are the BIC and AIC models. With the AIC model consistently performing better than the BIC model in Table 2, especially with the observation that as the training size increases, the advantage of the AIC model is getting stronger, the AIC model may be somewhat more favorable unless a more parsimonious models is sought from practical perspectives.

It is worth pointing out that when choosing between the BIC and AIC models, which luckily are nested in this example, one may use the deviance-difference based chi-square test, and the result also strongly supports the AIC model (p-value 0.015). However, the age-only model is not nested within the BIC or AIC model and it is unclear how to do an asymptotically valid test between them.

In summary, for this example, we see that the HL test is not able to reveal the poor performance of the age-only model. In contrast, by our DRYV approach, the model is seen to be unfit. Furthermore, the DRYV pairwise comparison tables may offer helpful insight on comparison of relevant models or methods.

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40-79.

Biau, G. (2012). Analysis of a random forests model, *Journal of Machine Learning Research*, 13, 1063-1095.

Burman, P. (1990) Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods, *Sankhya* Series A, Indian Journal of Statistics, 52, 314-345.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Liu, W. and Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Annals of Statistics*, 39, 2074-2102.

Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer, New York.

van der Lann, M. J., Dudoit, S. and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Stat Appl Genet Mol Biol*, 3, Article 4.

Yang, Y. (1999). Minimax nonparametric classification—Part I: rates of convergence; Part II: model selection for adaptation. *IEEE Transaction on Information Theory*, 45, 2271-2292.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, 92, 937-950.

Yang, Y. (2006). Comparing learning methods for classification. *Statistica Sinica*, 16, 635-657.