Some technical details on confidence intervals for LIFT measures in data mining *

Wenxin Jiang and Yu Zhao †

June 2, 2014

Abstract

A LIFT measure, such as the response rate, lift, or the percentage of captured response, is a fundamental measure of effectiveness for a scoring rule obtained from data mining, which is estimated from a set of validation data. The LIFT measures are related to the ROC (Receiver Operator Characteristic), but there exist some important differences. In this paper, we study how to construct confidence intervals of the LIFT measures. We point out the difficulty of this task and explain how simple binomial confidence intervals can have incorrect coverage probabilities, due to omitting variation from the sample percentile of the scoring rule. We derive the asymptotic distribution using some advanced empirical process theory and the functional delta method in

*Technical Report 14-02, Department of Statistics, Northwestern University.

†Wenxin Jiang is Professor of Department of Statistics, Northwestern University, Evanston, IL 60208 (email: wjiang@northwestern.edu); and Yu Zhao is Statistician at Amazon (email: yuzhaonwu@gmail.com).

Appendix B. The additional variation is shown to be related to a conditional mean response, which can be estimated by a local averaging of the responses over the scores from the validation data. Alternatively, a subsampling method is shown to provide a valid confidence interval, without needing to estimate the conditional mean response. A simple nonparamatric boostrap confidence interval can also be used. Numerical experiments are conducted to compare these different methods regarding the coverage probabilities and the lengths of the resulting confidence intervals.

Keywords: bootstrap, confidence interval, empirical process, functional delta method, LIFT, local average, %response, ROC (Receiver Operator Characteristic), subsampling, validation data

1 Introduction

In data mining, predictive models can be used to detect the likely responders to marketing campaigns. The capability of a predictive model to capture the responders can be evaluated by a number of measures on the validation data. For example, SAS Enterprise Miner, a popular data mining software, is capable of presenting various kind of "LIFT charts", which is actually a general name that includes $lift$, $\%response$, and $\%captured.response$, among others, for describing the effectiveness of a predictive model in identifying responders, and also for comparing different predictive models (SAS Institute, 2003). For example, a predictive model (say, logistic regression, or neural network) is used to rank the subjects in the validation data according to a score, which can be related to an estimated probability of responding to a marketing campaign. A measure such as the $\%response$ is then computed in SAS from a fixed percentage, say, $100r\%$ of validation data ($r \in (0, 1)$), where the

predictive model scores above a corresponding cut-off point $\tilde{c}_r$ estimated from the validation data. Figure 1 presents such a performance measure $lift$ (for a logistic regression model) plotted against the 10 *decile* values of $r \in \{0.1, 0.2, 0.3, ..., 1.0\}$, on a validation data set explained in more detail in Section 6.

Despite their decade-long popularity in data mining practices, the LIFT measures have not been studied very thoroughly in a statistical perspective. For example, a recent literature review on "Confidence interval of lift" provided little relevant work, and SAS Enterprise Miner still does not include the option of confidence intervals for the LIFT measures in a typical output such as Figure 1. The current paper attempts to fill this void and enable the users to add valid confidence intervals to Figure 1. We will illustrate this with a real data application in Section 6.

The problem of finding a confidence interval, for example, for $\%response$, may be deceptively simple, since the parameter can be estimated by a sample proportion, which would suggest a binomial confidence interval. However, we will show that the proper solution will turn out to be much more complicated. Binomial confidence intervals would not be appropriate to account for all the inherent variations, since the sample proportion turns out to be *not* computed over statistically independent subjects - these subjects all have model scores above a common cut-off point estimated from the entire validation sample. We need to deal with the extra variation of the cut-off point $\tilde{c}_r$, of the $100r\%$ top model scores, as estimated from the validation sample. This turns out to be a mathematically challenging task, since a parameter such as $\%response$ will be shown to be a discontinuous function of the estimated cutoff point. We will use some advanced empirical process theory and the functional delta method in Appendix B to derive the asymptotic distribution properly.

The phenomenon of improper coverage of the binomial confidence intervals was first reported in Rosset et al. (2001), which is the only reference directly related

to our paper. Their paper studied the behavior of binomial confidence intervals for a LIFT measure (%*captured.response*), and found empirically that the they are overly conservative and too loose compared to the bootstrap confidence intervals. Since the emphasis of their paper is different (on the effect of sampling balance on the measures), there empirical finding was not very noticeable and was placed in a very short subsubsection (Rosset et al. 2001, Section 3.2.3 and Table 2).

Our paper will provide a theoretical explanation to the empirical phenomenon reported in Rosset et al. (2001). We focus on how to derive a correct asymptotic theory to account for the additional sources of variations in the confidence intervals. Although little previous work was done on the asymptotic distributions and the confidence intervals of the LIFT measures, there have been extensive studies on a related performance measure ROC (Receiver Operator Characteristic), which is commonly used in medical diagnosis (see, e.g., Ma and Hall 1993; Hsieh and Turnbull 1996; Hall et al. 2005ab; Horváth et al. 2008; Su et al. 2009). Our theoretical work is most related to the theoretical works of Hsieh and Turnbull (1996) and Hall et al. (2004) on ROC. The relations and differences, as well as some other related works, are described in more details in a separate subsection (Section 2.3) later.

The following is an outline of this paper. In Section 2, we first define various LIFT measures used in SAS and point out how these LIFT measures are related to each other, and how they are estimated in data mining practices using a validation data set. We then discuss the relation and difference between the LIFT measures and ROC measures and discuss some related works in Section 2.3. Then in Section 3, we describe the asymptotic distributions of the validation sample estimates of the LIFT measures. The asymptotic variances are shown to be related to a mean response conditional on the score at the cutoff point. At this point, two different approaches are considered to derive confidence intervals (in Section 4): One uses a local averaging

method to estimate the conditional mean parameter, the other uses a subsampling method to bypass the step of local averaging. Simulations are conducted in Section 5 to compare the coverage probabilities and the lengths of the confidence intervals obtained from the binomial method, the local averaging method, and the subsampling method. A real data application is described in Section 6. Appendix A describes how to extend our theory to compute confidence intervals that are simultaneously valid for all the deciles r. Technical proofs and more general distributional results based on empirical processes are included in Appendix B.

The current paper focuses on the large sample asymptotics, which is certainly very appropirate in the current era of big data. For smaller sample sizes, some adjustment on the variance estimation can lead to further improvements. These details are included in Section 8, which also has included a nonparametric bootstrap method in the numerical comparisons.

2 Lift measures

2.1 Notation and definitions

Before proceeding, we introduce some mathematical notation. Let $Y \in \{0, 1\}$ be the random response variable which will be 1 if a subject responds to the marketing campaign and 0 if otherwise. Let X be a random input vector, which can include demographic variables or household status variables that can be used to predict the response. Let $S = S(X)$ be a scalar function of X used to score the subjects. This score S in the ideal case would be the same as the response probability $P(Y = 1|X)$, but in practice an estimated version (from a training data set based on logistic regression, neural networks, or a decision tree, for example) of the response probability (or its monotone transformation) is often used instead. In this paper we will not

investigate how S is obtained, but only consider how it performs once the scoring rule $S(X)$ is given. Therefore, sometimes it may be simpler to directly treat S as a random variable of interest and consider the joint distribution for (Y, S), instead of for (Y, X).

Consider a marketing campaign that aims to contact the top $100r\%$ ($r \in (0,1)$) of the subjects according to score S. Correspondingly, we define an "action indicator" $A = I(S > c)$ whose expectation is $EA = r$. Here an individual would be contacted by the marketing campaign if and only if its action value $A = 1$. The parameter c is a cutoff for the score S, which is the $100(1-r)\%$ percentile of S. The performance of the scoring method S can be described by the following LIFT measures:

(i) $\pi \equiv \%response = P(Y = 1|A = 1) = E(YA)/EA = E(YA)/r$ which measures the percentage of responders among all the contacted people;

(ii) $\kappa \equiv \%captured.response = P(A = 1|Y = 1) = E(YA)/EY$ which measures the percentage of contacted people among all people who would respond;

(iii) $lift = P(Y = 1|A = 1)/P(Y = 1) = E(YA)/(EYEA)$ which is the ratio of the $\%response$ achieved by action A and the baseline $\%response$ $P(Y = 1)$;

(iv) $expected.profit = Eg(Y, A)$, according to some pay-off function $g(.,.)$.

All these measures can be shown to be a monotone function of $\%response$ and can be regarded as equivalent. So, for example, for any two actions A_1 and A_2, $lift(A_1) > lift(A_2)$ if and only if $\%response(A_1) > \%response(A_2)$, since $lift = \%response/EY$. Likewise, $\%captured.response = (r/EY)\%response$, and $expected.profit = r[(g(1,1) + g(0,0) - g(1,0) - g(0,1)) \cdot \%response + g(0,1) - g(0,0)] + [(g(1,0) - g(0,0))EY + g(0,0)]$ are also monotone in $\%response$. These relations immediately imply the following result:

Proposition 1 *For two actions A_1 and A_2 with the same percentage of contacted population $EA_1 = EA_2 = r > 0$, we have*

$$\frac{\%response(A_1)}{\%response(A_2)} = \frac{\%captured.response(A_1)}{\%captured.response(A_2)} = \frac{lift(A_1)}{lift(A_2)},$$

whenever these ratios are finite.

2.2 Estimated LIFT measures from the validation sample

The quantities defined above are population quantities related to an unknown target population. In practice, all these quantities will need to be estimated on a sample (validation data). We will replace all P by $\tilde{P}$, E by $\tilde{E}$, where *tilde* denotes the empirical version of the corresponding measure for the validation sample. Therefore, for example, $\tilde{E}g(Y, A) = m^{-1}\sum_{i=1}^{m} g(Y_i, A_i)$ for a validation sample $(Y_i, S_i)_{i=1}^{m}$ with size m, which are assume to be iid (independent and identically distributed) with (Y, S). In particular, the cut-off parameter c will be estimated by $\tilde{c}$, which is the $100(1-r)$ sample percentile. The corresponding estimated action rule is $\tilde{A} = I[S > \tilde{c}]$ with sample probability $\tilde{E}(\tilde{A}) = r$[1]. Then $\pi = \%response$ is estimated by

$$\tilde{\pi} = \tilde{E}(Y\tilde{A})/\tilde{E}\tilde{A}, \tag{1}$$

where $\tilde{E}(\tilde{A}) = r$ and $\tilde{E}(Y\tilde{A}) = m^{-1}\sum_{i=1}^{m} Y_i I(S_i > \tilde{c})$.

2.3 Relation to the literature on ROC

The LIFT measures are intimately related to the ROC, but there exist some important differences. Using a notation similar to Hsieh and Turnbull (1996) and Hall et

[1]The two sides of this equation can actually be slightly different if r is not divisible by m. However, we will ignore this difference in the notation for simplicity, since the size of the difference is at most $1/m$ and does not change asymptotics in the leading order $O_p(1/\sqrt{m})$.

al. (2004), the ROC measures the True Positive rate $1 - F_1(s) = P(S > s|Y = 1)$ for a given False Positive rate $p = 1 - F_0(s) = P(S > s|Y = 0)$. Therefore the ROC function can be expressed as $ROC = 1 - F_1 \circ F_0^{-1}(1 - p)$. This is similar to a LIFT measure $\%captured.response = P(S > c|Y = 1)$, which can be expressed as $\%captured.response = 1 - F_1 \circ F^{-1}(1 - r)$ where $r = 1 - F(c) = P(S > c)$ is the percentage of contacted population. In data mining, the later approach is more natural, since it is more natural to examine the performance against r (the percentage of contacted population due to a limited budget), say, in a marketing campaign, than to control the false positive rate p.

Our estimates of the LIFT measures in Section 2.2 are based on the empirical distribution from the validation sampl, similar to Hsieh and Turnbull (1996) in the ROC literature. Although the kernel-smoothed estimate by Hall et al. (2004) may also be used in principle, we use the simplest empirical distribution estimate here, since this is currently the prevalent practice in data mining softwares.

There are many methods to construct confidence intervals and simultaneous confidence bands for ROC measures, see a good literature review and an extensive numerical comparison of the performances by Macskassy et al. (2005a,b). For nonparametric estimates of the ROC measures, the asymptotic distribution theory is given by, e.g., Hsieh and Turnbull (1996), and pointwise confidence intervals are provided by, e.g., Hall et al.(2004).

Our situation has some important differences from that of Hsieh and Turnbull (1996) and Hall et al.(2004). Their nonparametric estimates of the two distributions F_0 and F_1 are assumed to be independent, since they correspond to mutually nonoverlapping populations with $Y = 0$ and $Y = 1$, respectively. This is used in their derivation of a first order approximation in a sum of two independent components. On the other hand, our estimates of F and F_1 are *not* independent, since F

is the distribution of the combined population. In addition, Y is generally regarded as random in data mining, and some other LIFT measures, such as the $\%response$, are not only related to the two distributions F_1 and F as is the $\%captured.response$, but also related to the mixing proportion $P(Y = 1)$.

In this paper, we focus on asymptotic distributions for estimated LIFT measures with a given contacted proportion r, and provide methods for constructing pointwise confidence intervals (similar to those of Hall et al. 2004). However, we also outline in Appendix A how to derive confidence intervals that are simultaneously valid for all decile values of r. In addition, a joint asymptotic distribution for the entire LIFT curve estimate (similar to Hsieh and Turnbull 1996 in the context of ROC) is also provided in Appendix B.

3 Asymptotic distribution for the estimated LIFT measures

In general, below we will denote θ for any of the four LIFT measures and $\tilde{\theta}$ as its validation sample estimate.

We will consider how to construct a pointwise confidence interval for θ based on the validation sample estimate $\tilde{\theta}$. We show that for large m, $\tilde{\theta}$ converges in distribution to a normal distribution $N(\theta, var(\tilde{\theta}))$. The computation of asymptotic variance $var(\tilde{\theta})$, however, is not as straightforward as it looks. For example, although the estimated $\%response$ $\tilde{\pi} = \tilde{E}(Y\tilde{A})/\tilde{E}\tilde{A}$ can be regarded as a sample proportion of individuals with $Y = 1$ out of all individuals being contacted (with $\tilde{A} = 1$), it is not proper to use a binomial distribution to compute its variance, since $\tilde{A}$ depends on an estimated cutoff $\tilde{c}$, which is the $(1 - r)$th empirical quantile of the scoring rule S, dependent on all m random individuals. The terms being averaged in $\tilde{E}(\cdot)$ are

therefore no longer independent. Moreover, the estimator

$$\tilde{\pi} = m^{-1} \sum_{i=1}^{m} Y_i I(S_i > \tilde{c})/r \equiv \tilde{G}(\tilde{c})/r$$

depends on $\tilde{c}$ in a discontinuous way, so $\tilde{G}'(c)$ does not exist, and we cannot use the usual δ technique to derive the asymptotic distribution based on a first order Taylor expansion in $\tilde{c}$. We will use the functional delta method (as described in, e.g., Section 3.9, van der Vaart and Wellner 1996) to solve this problem. The key observation is that although $\tilde{\pi}$ is not continuous in $\tilde{c}$, it is differentiable (in Hadarmard's sense) as a *functional* of the empirical process $\tilde{G}$ and the empirical quantile process $\tilde{c}$, at the limiting points of these empirical quantities, and we can essentially do a functional delta method to derive the asymptotic distribution of $\tilde{\pi}$ from the joint distribution of $\tilde{G}$ and $\tilde{c}$. We will show the derivation of the following results in Appendix B.

Proposition 2 *Assume that $P(Y = 1) > 0$, and that the conditional probability densities $p(S|Y)$ exist and are positive and continuous differentiable in a neighborhood of $S = c$, for $Y = 1$ and $Y = 0$. Then we have the following results: for any of the four LIFT measures θ, as $m \to \infty$, the sample LIFT measure $\tilde{\theta}$ converges in distribution to a normal distribution $N(\theta, var(\tilde{\theta}))$, where*

$var(\tilde{\theta}) = m^{-1} var(H) = m^{-1} E(H - EH)^2$,

where $H = (Y - \Lambda)(aA + b)$,

$A = I(S > c)$, *S is a score, c is the cutoff such that $r = EI(S > c)$, $r \in (0, 1)$,*

and the symbol Λ denotes the conditional mean $\Lambda = E(Y|S = c)$,.

The parameters (a, b) take different values for the four LIFT measures:

(i) For $\theta = \%response$, use $(a, b) = (r^{-1}, 0)$;

(ii) For $\theta = \%captured.response$, use $(a, b) = ((EY)^{-1}, -(EY)^{-1}\%captured.response)$;

(iii) For $\theta = lift$, use $(a, b) = ((rEY)^{-1}, -(rEY)^{-1}\%captured.response)$;

(iv) For $\theta = expected.profit = Eg(Y, A)$, *use* $(a, b) = (g(1,1) + g(0,0) - g(1,0) - g(0,1), g(1,0) - g(0,0))$.

Remark 1 In practice, the deciles used are $r \in \{0.1, 0.2, 0.3, ..., 0.9, 1.0\}$, and the last point $r = 1.0$ is not included in the domain $(0, 1)$ of the Proposition. However, the derived asymptotic variance formula is still correct for $r = 1.0$. This is because when $r = 1$, all people are contacted, and the LIFT parameters have very simple forms. e.g., the $\%response = P(Y = 1)$, $lift = \%captured.response = 1$ always. The variance of their sample estimates are trivial and can be checked individually to satisfy the formulas of Proposition 2.

Corollary 1 *Assume the conditions of Proposition 2. For the estimated* $\%response$ $\tilde{\pi} = \tilde{E}(Y\tilde{A})/\tilde{E}\tilde{A}$, *the asymptotic variance is*

$$var(\tilde{\pi}) = (mr)^{-1}\pi(1-\pi) * [1 + (1-r)(\pi - \Lambda)^2/(\pi(1-\pi))]. \tag{2}$$

For the estimated $\%captured.response$ $\tilde{\kappa} = \tilde{E}(Y\tilde{A})/\tilde{E}Y$, *the asymptotic variance is*

$$var(\tilde{\kappa}) = (m\pi_0)^{-1}\kappa(1-\kappa) * [1 - 2\Lambda + \Lambda^2(1-r)/(\pi(1-\kappa))] \tag{3}$$

For the esimated $lift$ $\tilde{\kappa}/r$, *the asymptotic variance is* $var(\tilde{\kappa})/r^2$. *Here* $\pi_0 = EY$, $\pi = E(YI(S > c))/EI(S > c)$ *is the* $\%response$, $\tilde{\pi}$ *is its estimator based on (1),* $\kappa = E(YI(S > c))/EY$ *is the* $\%captured.response$, $\tilde{\kappa}$ *is its estimator,* S *is a score,* c *is the cutoff such that* $r = EI(S > c)$, $r \in (0, 1)$, *and* $\Lambda = E(Y|S = c)$.

Remark 2 *(Theoretical behaviors of the binomial variances.)* In the formula for the asymptotic variance of the estimated $\%response$, the term $(mr)^{-1}\pi(1-\pi) \equiv var_B(\tilde{\pi})$ is the variance of a binomial proportion based on $Bin(mr, \pi)$, where mr is the total number of contacted people. In the formula for the asymptotic variance of the estimated $\%captured.response$, the term $(m\pi_0)^{-1}\kappa(1-\kappa) \equiv var_B(\tilde{\kappa})$ is the variance

of a binomial proportion $Bin(\sum_{i=1}^{m} Y_i, \kappa)$, where $m\pi_0$ can be estimated by the total number of responders. (See Rosset et al. 2001, equations (8) and (10)).

Now consider the ratio $var(\tilde{\kappa})/var_B(\tilde{\kappa}) = [1-2\Lambda+\Lambda^2(1-r)/(\pi(1-\kappa))]$ (which is also the variance ratio for the $lift$ estimate). There is a negative sign, which suggests that the correct variance may be smaller than the binomial variance, leading to more accurate (shorter) confidence intervals. Note that the ratio $var(\tilde{\kappa})/var_B(\tilde{\kappa}) = 1 - \pi + O(r)$.[2] So the ratio may become nearly 0, and the binomial variance may become much too large, for small r corresponding to a high $\%response$ π (i.e., when a small percentage of highly likely responders are contacted). In this situation, our theoretical result can provide an expanation about the empirical findings reported in Rosset et al (2001, Table 2), that the binomial confidence intervals are loose at small r values (such as 10%, 5%, 3% and 1%).

Regarding the ratio $var(\tilde{\pi})/var_B(\tilde{\pi}) = [1 + (1 - r)(\pi - \Lambda)^2/(\pi(1 - \pi))]$, note that our result suggests that $var(\tilde{\pi}) \geq var_B(\tilde{\pi})$ and the difference depends on the difference $\pi - \Lambda = E(Y|S > c) - E(Y|S = c)$, which would be small if $E(Y|S = s)$ varies slowly for $s \geq c$. However, when the mean function $E(Y|S = s)$ changes steeply, the binomial variance can be too small, leading to a lower-than-nominal coverage probability for the resulting confidence interval.

We will provide two examples later (in Section 5.1), where the variance ratios can be computed analytically, to show these two types of biases on the coverage probabilities of the binomial confidence intervals. They will be named the Case I example (for the too-long intervals for $lift$ or $\%captured.response$) and the Case II example (for the too-short intervals for $\%response$), respectively.

[2]This is established by expanding the ratio for small r and noticing that for smooth $\Lambda(r)$, we have $\Lambda - \pi = O(r)$ for small r, and that $\kappa = r\pi/\pi_0$.

4 Confidence intervals

4.1 Local estimation method

We now consider how to apply Proposition 2 to construct valid asymptotic confidence intervals. Suppose we have a consistent variance estimate $\tilde{var}(\tilde{\theta})$ based on the validation sample, such that $\tilde{var}(\tilde{\theta})/var(\tilde{\theta}) \to 1$ in probability as $m \to \infty$. Then due to the Slutsky's theorem, as $m \to \infty$, Proposition 2 implies that $(\tilde{\theta}-\theta)/\sqrt{\tilde{var}(\tilde{\theta})} \to N(0,1)$ in distribution, which implies that $P(\theta \in \tilde{\theta} \pm z_\alpha \sqrt{\tilde{var}(\tilde{\theta})}) \to 1-\alpha$ for any $\alpha \in (0,1)$, if z_α has a standard normal cumulative distribution function value $\Phi(z_\alpha) = 1-\alpha/2$. Therefore an asymptotic $100(1-\alpha)\%$ confidence interval for θ is $\tilde{\theta} \pm z_\alpha \sqrt{\tilde{var}(\tilde{\theta})}$.

We now consider the problem of consistently estimating $var(\tilde{\theta})$ by $\tilde{var}(\tilde{\theta})$ based on the validation sample. In the variance formula, $var(H)$ can be estimated by the validation sample variance $\tilde{var}(H) = \tilde{E}(H - \tilde{E}H)^2$, where the unknown parameters $(EY, \%captured.response, c, \Lambda)$ are replaced by their consistent estimates $(\widetilde{EY}, \widetilde{\%captured.response}, \tilde{c}, \tilde{\Lambda})$ based on the validation sample. The parameter r is known (such as 20% to be contacted by a marketing campaign). Now we consider estimation of $\Lambda = E(Y|S=c)$. In general, Λ needs to be estimated from a local regression of Y on S based on the validation data around $S = c$, where c is estimated by $\tilde{c}$. For example, a local average estimate $\tilde{\Lambda} = \tilde{E}YI(S \in \tilde{c} \pm h)/\tilde{E}I(S \in \tilde{c} \pm h)$ is well known to be consistent when h decreases with m such that $h \to 0$ and $mh \to \infty$.

4.2 Subsampling method

In the previous subsection, we described a method to construct a confidence interval for a LIFT parameter θ. The asymptotic variance involves a conditional mean response parameter Λ, which can be estimated by a local averaging of the responses against the scores. An alternative method we consider here is to bypass the problem

of estimating Λ by using a subsample estimate of the asymptotic variance, similar to Ibragimov and Müller (2010).

Let $\tilde{\theta}_j, j = 1, ..., q$ be q estimated LIFT measures based on $q(> 1)$ *independent* sub-samples of size $m_j, j = 1, ..., q$. The total sample size is $m = \sum_{j=1}^{q} m_j$ and we consider the asymptotics when $\lim_{m\to\infty} m_j/m = 1/q$ (which will be valid almost surely, for example, when observations are randomly assigned to the q groups with equal probability).

Let $\bar{\theta}$ and s be, respectively, the sample mean and sample variance of $\{\tilde{\theta}_j, j = 1, ..., q\}$. Then we have the following proposition.

Proposition 3 *Assume the regularity condition of Proposition 2, and assume that* $\lim_{m\to\infty} m_j/m = 1/q$ *for all* $j = 1, ..., q$. *We have the following results as* $m \to \infty$*:* $\sqrt{q}(\bar{\theta} - \theta)/s$ *converges in distribution to* t_{q-1} *(the t-distribution with* $q - 1$ *degrees of freedom), and for any* $\alpha \in (0, 1)$, $P(\theta \in \bar{\theta} \pm t_{q-1,\alpha} s/\sqrt{q}) \to 1 - \alpha$, *where* $t_{q-1,\alpha}$ *is the* $(1 - \alpha/2)$*th quantile of the* t_{q-1} *distribution.* [3]

Proof:

Applying the asymptotic normality result of Proposition 2, noting that the sub-samples are independent whereas under this method, we have (*) $\sqrt{m/q}(\tilde{\theta}_1 - \theta, ..., \tilde{\theta}_q - \theta)^T$ converges in distribution to $N(\mathbf{0}, diag(\sigma^2, ..., \sigma^2))$ where $\sigma^2 = var(H)$. (Note that here the individual sample sizes m_j have all been replaced by m/q due to the Slutsky's Theorem.) Note that the result (*) satisfies the basic assumption (4) made in Ibragimov and Müller (2010).

Due to the continuous mapping theorem, we derive from (*) that $\sqrt{q}(\bar{\theta} - \theta)/s$ converges in distribution to t_{q-1} (the t-distribution with $q - 1$ degrees of freedom).

[3] This suggests that we can construct a $100(1 - \alpha)\%$ asymptotic confidence interval for θ by $\bar{\theta} \pm t_{q-1,\alpha} s/\sqrt{q}$. Alternatively, one can center the interval at the original whole sample estimater $\tilde{\theta}$ to obtain $\tilde{\theta} \pm t_{q-1,\alpha} s/\sqrt{q}$, as described in Section 8.

Then the coverage probability result holds.

Q.E.D.

5 A simulation study

5.1 Two analytic examples

In the remarks after Corollary 1, we have described the behavior of the binomial confidence intervals, based on a theoretical study on the ratios of the asymptotic variances. In this subsection, we will use the notation of Corollary 1 and present two examples where the asymptotic variances can be analytically computed, in order to have a sense of the size of the numerical differences that can be observed. The simulation study later will be based on a smoothed version of the two models presented here.

Case I: (gradual case)

Let $S = X$ and $X \sim Unif(0,1)$, $E(Y|X) = X$. we consider contacting top $r = 10\%$ of the X-scores. Then the threshold valus is $c = 0.9$, since $P(X > c) = 10\%$. The other useful parameters include $\pi_0 = EY = 0.5$, $\Lambda = E(Y|X = c) = 0.9$, $\pi = P(Y = 1|X > c) = 0.95$, $\kappa = P(X > c|Y = 1) = r\pi/\pi_0 = 0.19$. This is a case with a small $\pi - \Lambda = 0.05$. The ration $var(\tilde{\pi})/var_B(\tilde{\pi}) = [1 + (1-r)(\pi - \Lambda)^2/(\pi(1-\pi))] = 1.0473684$, which is very close to 1. So there should be not much difference in the coverage probability and the length when one uses the binomial confidence interval for $\%response$. However, the ratio $var(\tilde{\kappa})/var_B(\tilde{\kappa}) = [1-2\Lambda+\Lambda^2(1-r)/(\pi(1-\kappa))] = 0.1473684$ is very small. The length of the binomial confidence interval for the $lift$ or for the $\%captured.response$ will much be longer than needed, at a ratio

$1/\sqrt{\{var(\tilde{\kappa})/var_B(\tilde{\kappa})\}} = 2.60494$. The coverage probability will be overly conservative.

Case II: (steep case)

The structure is similar to Case I except that we $E(Y|X) = \max\{\min\{3(X - 1/3), 1\}, 0\}$, which is a three-piece linear model that is 0 below $X < 1/3$, and 1 above $X > 2/3$. We conider contacting half of the people, $r = 0.5$. Then $c = 0.5$ since $P(X > 0.5) = 0.5$. The other useful parameters include $\pi_0 = EY = 0.5$, $\Lambda = E(Y|X = c) = 0.5$, $\pi = P(Y = 1|X > c) = 11/12 = 0.9166667 = P(X > c|Y = 1) = \kappa$. This is a case with a large $\pi - \Lambda = 0.4166667$, caused by a steep change in the middle part of the X domain. The variance ratio $var(\tilde{\pi})/var_B(\tilde{\pi}) = [1 + (1 - r)(\pi - \Lambda)^2/(\pi(1 - \pi))] = 2.136364$, which is much larger than 1. So the binomial confidence interval for the $\%response$ is too short and will under-cover in probability.

(In this case the binomial confidence interval for the $lift$ or $\%captured.response$ will also be too short since $var(\tilde{\kappa})/var_B(\tilde{\kappa}) = [1 - 2\Lambda + \Lambda^2(1 - r)/(\pi(1 - \kappa))] = 1.636364 > 1.$),

When the nominal confidence level =0.95, the asymptotic coverage probability for the binomial confidence interval (of the form $\tilde{\pi} \pm 1.96\sqrt{\tilde{var}_B(\tilde{\pi})}$) is only $\Phi(+1.96\sqrt{1/2.136364}) - \Phi(-1.96\sqrt{1/2.136364}) \approx 0.820$.

Although these two cases are only theoretical predictions based on piecewise constant response curves, we will verify in the later simulations that these predictions are qualitatively correct even with more realistic smooth response curves following the logistic regression models.

5.2 Simulations

We let the score $S = X \sim Unif[0, 1]$ in the simulations. We will use two different logistic regression models to generate simulated data. These two cases will be called the 'gradual' case and the 'steep' case, respectively.

In the first (gradual) case, the true response probability (logit model) is $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{2.75-5.4x}}$, population with the top 10% of S is contacted, and μ^* varies little (from 0.892 to 0.934) in the contacted region $S \in (0.9, 1.0]$.

In the second (steep) case, the true response probability (logit model) is $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{9-18.5x}}$, population with the top 50% of S is contacted, and μ^* varies a lot (from 0.562 to 1.000) in the contacted region $S \in (0.5, 1.0]$.

These are similar to the two cases discussed in Section 5.1 but are now smoothed.

We will compare the performance of three confidence intervals (CIs): the binomial CI, the local estimation CI (as explained in Section 4.1), and the subsampling CI (as explained in Section 4.2). For the subsampling method, we use $q = 10$ subsamples in the simulation. For the local estimation method, we need to estimate the conditional mean function $\Lambda = E(Y|S = c)$. We will use a local average estimator $\tilde{\Lambda}$, which is the sample average of Y for $S \in [c - h, c + h]$. We will use $h = m^{-1/3}$ in the simulations (which will lead to the optimal convergence rate for estimating Λ). So, for example, when $m = 1000$, we use $h_{optimal} = 0.1$.

In the Tables (1 to 6), we compare the empirical coverage probabilities with the nominal ones for a thousand CIs obtained from the three methods (binomial, local estimation, and subsampling), and also the average widths of the CIs, as well as the algorithm complexities in terms of the total execution times.[4] Each of these thousand

[4]The hardware information of the computer used to derive the results is as follows for reference of timing. CPU: Intel® Core™ i5-3210M CPU 2.50GHz; RAM: 8.00GB; OS: Windows® 7 64 bits;

CIs is derived from a common sample size m. We report the results for two choices of the sample size: $m = 1000$ and $m = 10000$, which are quite typical sample sizes for the real data sets used in data mining. (Behaviors at smaller sample sizes, and with varying tuning parameters h and q, are summarized in Section 8.)

We first look at the performance of the Binomial CI. For $\%response$, our theoretical predictions in Section 5.1 turn out to be very close to the actual simulation results. In Case 1 (the gradual case), we predicted that the coverage probabilities would be close to the nominal ones. The simulation results (based on 1000 CIs) are 93.3% (nominal: 95%) and 91.9%(nominal: 90%). In Case 2 (the steep case), we predicted that the CI would severely undercover. The simulation results are 81.2%(nominal: 95%) and 74.0%(nominal: 90%).

For $lift$ and $\%captured.response$, since they only differ by a factor r, the CI coverage probabilities must be exactly the same in simulations, while the widths of their CIs must differ exactly by a factor of r. In the gradual case (where the binomial CI performs well for the $\%response$), we notice that the binomial CIs for both $lift$ and $\%captured.response$ tend to be overly conservative and too loose (similar to the findings reported in Rosset et al. 2001, Section 3.2.3). This can be seen in the simulation results for Case 1, where for $lift$ and $\%captured.response$ all the coverage probabilities (based on 1000 CIs) are 100%. On the other hand, for Case 2, where the binomial CI undercovers for $\%response$, it also undercovers for $lift$ and $\%captured.response$. The simulation coverage probabilities are 85.5%(nominal: 95%) and 77.1%(nominal: 90%).

In summary, the results above verify our theoretical predictions. Regarding the sample sizes (m) needed for the asymptotic CIs to have satisfactory coverage prob-

Pseudo-random number generator: R®; Programming language and major software component: R®.

abilities (being close to the nominal ones), both the local estimation method and the subsampling method perform quite well when $m = 1000$ already, and they do even better when $m = 10000$. (In comparison, increasing m does not improve the performance of the binomial method. In other words, the binomial CIs can either overcover or undercover *even* with $m = 10000$.) The local estimation method tends to produce slightly shorter CIs than the subsampling method, but they both work well in terms of the coverage probabilities.

Table 1: CIs for %response, case 1, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{2.75-5.4x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	0.933	0.933	0.941	0.952	0.952	0.947
width	0.108	0.109	0.124	0.0346	0.0348	0.0389
coverage(90%CI)	0.919	0.919	0.890	0.888	0.904	0.898
width	0.0908	0.0916	0.100	0.0290	0.0292	0.0315
time(sec)	12.66	18.17	13.43	124.91	184.1	136.37

Table 2: CIs for %response, case 2, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{9-18.5x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	0.812	0.936	0.942	0.821	0.947	0.947
width	0.0424	0.0612	0.0731	0.0134	0.0199	0.0226
coverage(90%CI)	0.740	0.892	0.881	0.737	0.899	0.905
width	0.0355	0.0514	0.0592	0.0113	0.0167	0.0183
time(sec)	14.53	22.23	20.89	142.63	224.25	182.43

Table 3: CIs for Lift, case 1, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{2.75-5.4x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	1.00	0.946	0.939	1.00	0.939	0.942
width	0.686	0.290	0.332	0.217	0.0915	0.103
coverage(90%CI)	1.00	0.897	0.888	1.00	0.887	0.890
width	0.576	0.243	0.270	0.182	0.0768	0.0832
time(sec)	12.66	18.17	13.43	124.91	184.1	136.37

Table 4: CIs for Lift, case 2, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{9-18.5x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	0.855	0.962	0.946	0.851	0.949	0.943
width	0.0972	0.135	0.142	0.0308	0.0415	0.0460
coverage(90%CI)	0.771	0.910	0.894	0.769	0.897	0.893
width	0.0816	0.113	0.115	0.0259	0.0348	0.0373
time(sec)	14.53	22.23	20.89	142.63	224.25	182.43

Table 5: CIs for %captured response, case 1, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{2.75-5.4x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	1.00	0.946	0.939	1.00	0.939	0.942
width	0.0686	0.0290	0.0332	0.0217	0.00915	0.0103
coverage(90%CI)	1.00	0.897	0.888	1.00	0.887	0.890
width	0.0576	0.0243	0.0270	0.0182	0.00768	0.00832
time(sec)	12.66	18.17	13.43	124.91	184.1	136.37

Table 6: CIs for %captured response, case 2, $\mu^*(x) = E(Y|X = x) = \frac{1}{1+e^{9-18.5x}}$

sample size	n=1000	n=1000	n=1000	n=10000	n=10000	n=10000
method	Binomial	Local	Subsample	Binomial	Local	Subsample
coverage(95%CI)	0.855	0.962	0.946	0.851	0.949	0.943
width	0.0486	0.0674	0.0711	0.0154	0.0207	0.0230
coverage(90%CI)	0.771	0.910	0.894	0.769	0.897	0.893
width	0.0408	0.0565	0.0576	0.0129	0.0174	0.0186
time(sec)	14.53	22.23	20.89	142.63	224.25	182.43

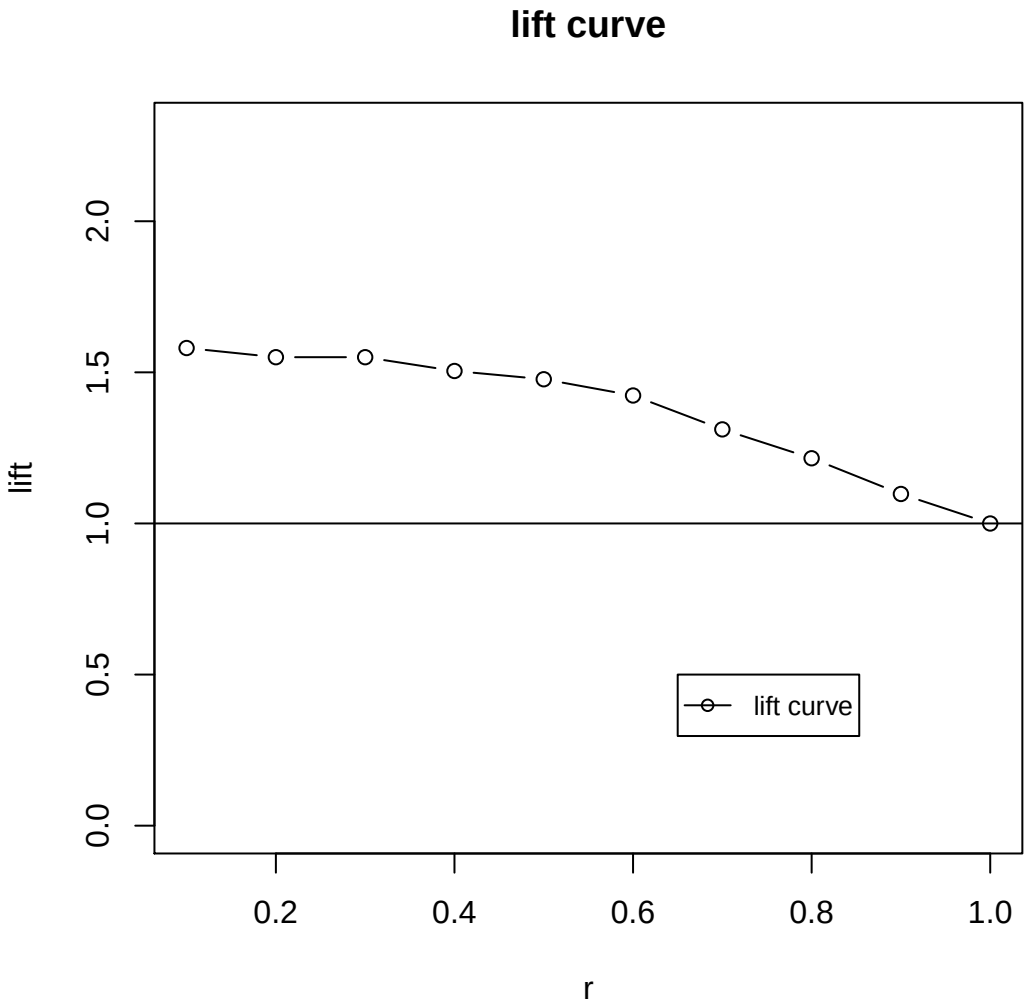


Figure 1: $lift$ estimates without confidence intervals. (The solid horizontal line represents the "baseline $lift$", which is equal to the $lift$ estimate at $r = 1.0$.)

6 Real data application

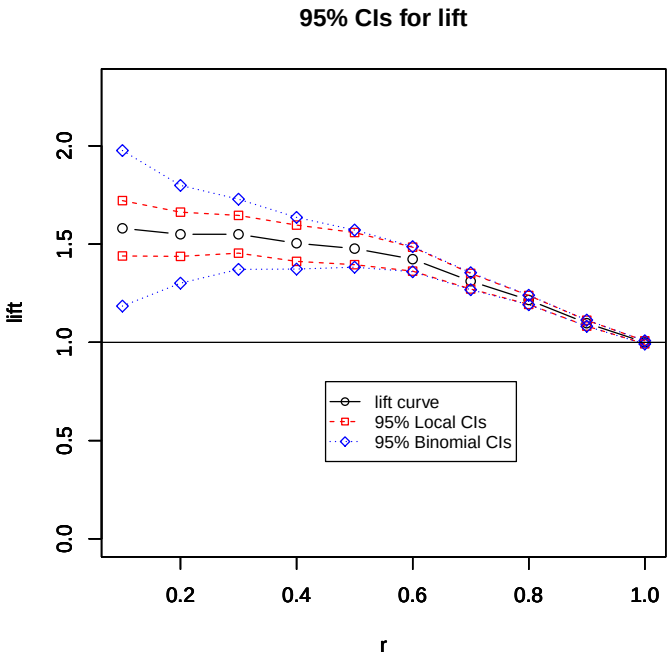


Figure 2: $lift$ estimates and their CIs (The solid horizontal line with height 1 represents the "$lift$ with campaign action completely random".)

To illustrate the application of the methodology, in this section we consider the Orange Juice Data (Stine, Foster and Waterman 1998). The data contains purchases where the customer either purchased Citrus Hill ($Y = 1$) or Minute Maid ($Y = 0$) Orange Juice. A number of characteristics (X) of the customer and product are recorded, such as "price difference" and "customer brand loyalty". The total number of observations is 1070, from which we save half randomly for the training data. We use logistic regression with the default option in SPSS to obtain an estimate of the probability $P(Y = 1|X)$, which will be used as the scoring rule $S(X)$. The rule $S(X)$ is applied to the validation sample with $m = 535$, and the

sample $lift$ values are computed and plotted in Figure 2, according to the deciles $r \in \{0.1, 0.2, 0.3, ..., 0.9, 1.0\}$.

In this figure, the pointwise asymptotic 95% confidence intervals are presented as little squares (connected by dashed lines). They are computed according to the local estimation method with bandwidth $h = m^{-1/3}$. For comparison, we have also presented the binomial confidence intervals as little diamonds (connected in solid lines). We notice that the local estimation method can lead to much more accurate confidence intervals, especially at small r. At $r = 0.1$, the width of the confidence interval is reduced to 36% of the width from the binomial method.

Therefore, we have demonstrated that our method can be used to provide valid confidence intervals to the LIFT chart function of the standard data mining software, which will be a very useful addition for statisticians.

7 Discussions

Despite the popularity of the data mining practices, we find very little previous work in statistical inference for the LIFT measures, which are commonly used in data mining. In this paper, we have provided an asymptotic distribution theory for the validation sample estimates of the LIFT measures. We have also discussed several methods for constructing valid asymptotic confidence intervals for some common LIFT measures, including $\%response$, $lift$, and $\%captured.response$.

The current work focuses on single confidence interval with one given scoring rule applied to a fixed percentage of the targeted population. However, after adjusting for multiplicity as outlined in Appendix A, our asymptotic distribution results may also be applied to study simultaneous confidence intervals, with several different percentiles of the contacted population, as is commonly plotted in the LIFT charts

in data mining practices. This is illustrated in a real data example in Zhao (2014, Ph.D. thesis). It will also be of interest to study the statistical comparison of several scoring rules, say, one obtained from logistic regression, one from neural networks, and one from a decision tree. We will leave this as future work.

8 Further technical details

More extensive simulation studies and additional technical details are contained in this section, the contents are of which are summized below. We have studied the sensitiy of the choice of the tuning parameters (h for local average and q for subsampling), and we have studied the performance of the confidence intervals for smaller sample sizes such as $m = 200, 600, 1000$. We have also studied the coverage property for all r values from 0.1 to 1.0 and considered small sample corrections to improve the performance for very small and very large r. In addition, we have also studied the performance of a simple nonparametric bootstrap confidence interval. We found that the bootrap method is not significantly better than the proposed methods and can be slower by orders of magnitude. We found that the proposed methods are robust under different choices of the tuning parameters, and have good coverage performance for smaller sample sizes, at all decile values of r, after using the plus four-type corrections which are similar to the ones used in common freshman textbooks (e.g., Moore 2010, p.508).

We have studied the nominal 95% confidence intervals in all these results. Also, we have only reported results on $lift$ and on $\%response$. (We omitted $\%captured.response$ since it is proportional to the $lift$ by a nonrandom constant.) Here is a detailed summary.

1. (Bootstrap). As a referee suggested, we study the percentile bootstrap confidence

interval, which is a simple nonparametric bootstrap method based on the 2.5% and 97.5% bootstrap quantiles. We use 1000 bootstrap repetitions, which is a typical number of bootstrap repetitions, as is used in Section 2.2 of Macskassy et al. 2005.

A theoretical note: We believe that the bootstrap method is asymptotically valid, since our results in Appendix B can be used to establish the Hadamard differentiablity of the LIFT estimators and we can use the idea of Van der Vaart (2000, Example 23.11) to prove the validity of the boostrap method. However, we will describe later that the bootstrap method is much slower and does not perform better than the proposed methods. Below, we will first point out that the bootstrap method (as well as all other methods) can work very poorly in some situations, and a plus four correction (explained below) can significantly improve its performance.

2. (Plus four correction).

   We notice that the asymptotic confidence intervals can work very poorly for some finite sample situations (especially in Case II simulations). Plus four corrections can significantly improve the finite sample performance.

   For example, for the bootstrap method, initially, percentiles obtained were directly based on resampling the original estimates, without the plus four correction. Then we notice that they perform poorly (in having very low coverage probabilities) for small or moderate m, e.g., $m = 500$, for some r values in Case II. After some experiments, we found the reason. In Case II, $P(A = 1|Y = 1)$ is nearly 1 for high r (e.g., for $r = 0.9$), and $P(Y = 1|A = 1)$ is nearly 1 for small r (e.g., 0.1). Over resampling, the sample proportions are almost always one, which will often lead to zero-width confidence intervals, missing the true

values by a little amount.

Similarly, in Case II for these r values, confidence intervals without plus four do not work well for all other three methods. E.g., for the binomial method, a binomial variance estimate such as $\sqrt{\tilde{\pi}(1-\tilde{\pi})/50}$ is zero when the sample proportion $\tilde{\pi}$ is, say, 50/50, without plus four correction. This often happens for large success probability very close to one (such as $\pi = 0.9997$). Which will be missed by the zero-width confidence interval $[1, 1]$. Even though it misses only by a small amount, it will almost always happen and will lead to a very low coverage probability. For remedy, one can use a plus four-type corrections that are commonly used in freshman textbooks (e.g., Moore 2010, p.508), by adding 4 observations with 2 Yeses and 2 Nos, then the $\tilde{\pi}$ in the sample variance estimate will be replaced by 52/54, and the resulting confidence interval, even though changes very little, will have a nonzero width enough to cover the true value. The plus four approach can sometimes be overly conservative, but it is better to overcover than to undercover, and the increase in the width of the confidence interval is often very little anyway. The plus four correction does not affect the asympotics and any differences will go away in the large sample limit.

For the local estimation method, in the variance formulas we can use the plus four estimates for all proportion parameters, including Λ (when estimated from the observations falling within the bandwith).

Now we consider the plus four correction for bootstrap, i.e., bootstrapping the plus four estimates. A random plus four method is needed here, since the fixed plus four estimates (say, (50+2) out of (50+4),) would remain the same over bootstrap resampling, if the resampled original estimates were all the same (say, 50 out of 50) to begin with. The random method adds 4 observations with

$Bin(4, 0.5)$ Yeses and $4 - Bin(4, 2)$ No's, to the sample proportion estimates, independently over all the bootstrap repetitions.

Similarly, a fixed plus four method would not work for subsampling. We considered a random plus four method that adds a total number of 2 Yeses and 2 Nos, randomly assigned to the q subsample proportion estimates, when estimating the $\%response$ and the $\%captured.response$. Alternatively, we considered a method of an approximate plus four type correction to the"variance estimate". We found that both methods work very similarly, with the aproximate plus four method working slightly better. All the results reported in this section for subsampling are based on this approximate plus four method. This method involves increasing the original "variance estimate" (the sample variance of q subsample estimates, without plus four correction, divided by q) by a small correction $2/n^2$, where n is the sample size used in the original (whole sample) proportion estimate (which is the total number of responses for $\%captured.response$, or the total number of contacted people for $\%response$). This correction does not affect the asymptotics since the asymptotic variance is of order $1/n$. It is based on an analogy related to the following approximation to the plus four correction for the binomial method:

A theoretical note: Notice that for a binomial data with y Yeses and $n - y$ Nos, the effect of plus four correction on the binomial variance estimate is an increase upto $2/n^2$. Let $v4 = (y+2)/(n+4) * (1-(y+2)/(n+4))/(n+4)$ and $v = (y/n)(1 - y/n)/n$. Then $v4 = v[n/(n+4)^3] + (2/n^2)[(n+2)n^2/(n+4)^3]$, where the factors in the square brackets are less than 1 and are very close to 1 for large n.

The plus four corrections lead to signficant improvement for all four methods in Case II. See Figure 3 (for Case I) and Figure 4 (for Case II) for a

comparison of the four methods: binomial, local estimation, subsampling and bootstrap, with and without plus four corrections. We tried sample sizes $m = 100, 300, 500, 1000$ and reported the results at $m = 500$. The results at other sample sizes are all similar.

3. (Performance Comparisons).

   As seen in Figure 3 (for Case I) and Figure 4 (for Case II), all the four methods (binomial, local estimation, subsampling and bootstrap) work well in Case I, with or without plus four correction. In Case ll, all four methods work very poorly without plus 4 correction, for $\%response$ at small r, and for $\%captured.response$ at large r. All four methods work much better with plus four corrections.

   The proposed methods work very well in both Case I and Case II, compared to the binomial method and the bootstrap method. The subsampling confidence intervals are wider and more conservative than the local estimation confidence intervals, probably due to the use of the t distribution instead of the z distribution. In terms of the coverage probabilities, there are no serious undercoverages in either method, and any overcoverages do not lead to significant widening of the confidence intervals. These hold for all decile values of r. This is in contrast to the binomial method, which can also overcover for $lift$ at small r values, but the widening of the confidence intervals is much more severe.

4. (Time Comparison). In terms of the time needed to compute the confidence intervals, the binomial method is the fasted, next the subsampling method, next the local estimation method, and the bootstrap method is the slowest. Over various runs, we observe that typical ratios of the computational time needed are about 1 : (1.2 to 1.6) : (1.5 to 1.7) : (30 to 100).

5. (Center of the confidence intervals). Whenever possible, we use the unmodified original sample proportion estimates as the center of the confidence intervals, since sample proportions are most commonly used to estimate the LIFT measures (e.g., in standard data mining softwares)

   This means that we do not use the plus four correction on the center of the (binomial, local, or subsampling) confidence intervals, but only use the correction for improving the variance estimates.

   This also means that for the subsampling method, we do not use the average subsample estimates as the center of the confidence interval, but rather the original whole sample estimates which are used in standard data ming softwares, which are easier to compute, and can perform better (see below). [The subsample estimates are only used to "estimate the variance" (by the sample variance of q subsample esimates, divided by q), with a plus four correction.]

   A theoretical note: We believe that the whole sample estimate $\tilde{\theta}$ and the average of the subsample estimates $\bar{\theta}$ are asymptotically equivalent. A heuristic argument is that they have the same infuence function and the same Hadamard differential with respect to the underlying subsample empircal distribution functions. However, in results not presented here, we found that the t confidence intervals centered at the whole sample estimates $\tilde{\theta}$ perform better than the intervals centered at $\bar{\theta}$. This may be because the latter has a larger asymptotic bias (or order $O(1/(m/q))$) than that of the former (of order $O(1/m)$). (See, e.g., Haas 2006.) However we expect that the difference will disappear asymptotically. For large sample sizes such as those used in the earlier simulations in Section 5, the using confidence intervals centered at the average subsample estimates also perform very well.

6. (Effect of tuning parameters under various sample sizes).

   In Figure 5 (for Case I) and Figure 6 (for Case II), we consider the effect of using different bandwith parameters h for the local estimation method. We consider $h = 0.5m^{1/3}$, $m^{1/3}$, $1.5m^{1/3}$, for $m = 200, 600, 1000$. In Figure 7 (for Case I) and Figure 8 (for Case II), we consider the effect of using different number of subsample q for the subsampling method. We consider $q = 5, 10, 20$, for $m = 200, 600, 1000$. In all these simulations, we found that the results change very little. The coverage probabilities are not significantly affected by different choices of h or q. In simulation results not presented here, we also found that the widths of confidence intervals are not significantly affected by different choices of h or of q. However, the widths of the subsampling confidence intervals vary somewhat more than the widths of the local estimation confidence intervals, and they tend to be wider and more conservative. In terms of the coverage probabilities, there are no serious undercoverages for either method, for all these choices of tuning parameters and sample sizes.

Appendix A: Simultaneous confidence intervals

Consider any LIFT measure θ, such as $\theta = \%response = E(Y|S > c)$. It is dependent on the scoring rule S, as well as the cutoff c which is determined by the percentage contacted $r = P(S > c)$. SAS Enterprise Miner plots the LIFT charts to display multiple θ's simultaneously. These θ's can include, for example, $\%response_k$ (or $lift_k$) at all the deciles $r_k \in \{0.1, 0.2, 0.3, ..., 1.0\}$. Let us label the corresponding performance measures as θ_k, $k = 1, ..., p$. The previous confidence intervals are of the form $\tilde{\theta}_k \pm z_\alpha s_k$ for the performance measure θ_k, $k = 1, ..., p$, where $s_k = \sqrt{\tilde{var}(\tilde{\theta}_k)}$. These confidence intervals only have *individually* correct asymptotic coverage probabilities, i.e., $P(\theta_k \in \tilde{\theta}_k \pm z_\alpha s_k) \approx 1 - \alpha$, for each $k = 1, ..., p$. However,

Figure 3: Comparison of four CI methods, without and with plus 4 correction, for Case I. ($m = 500$)

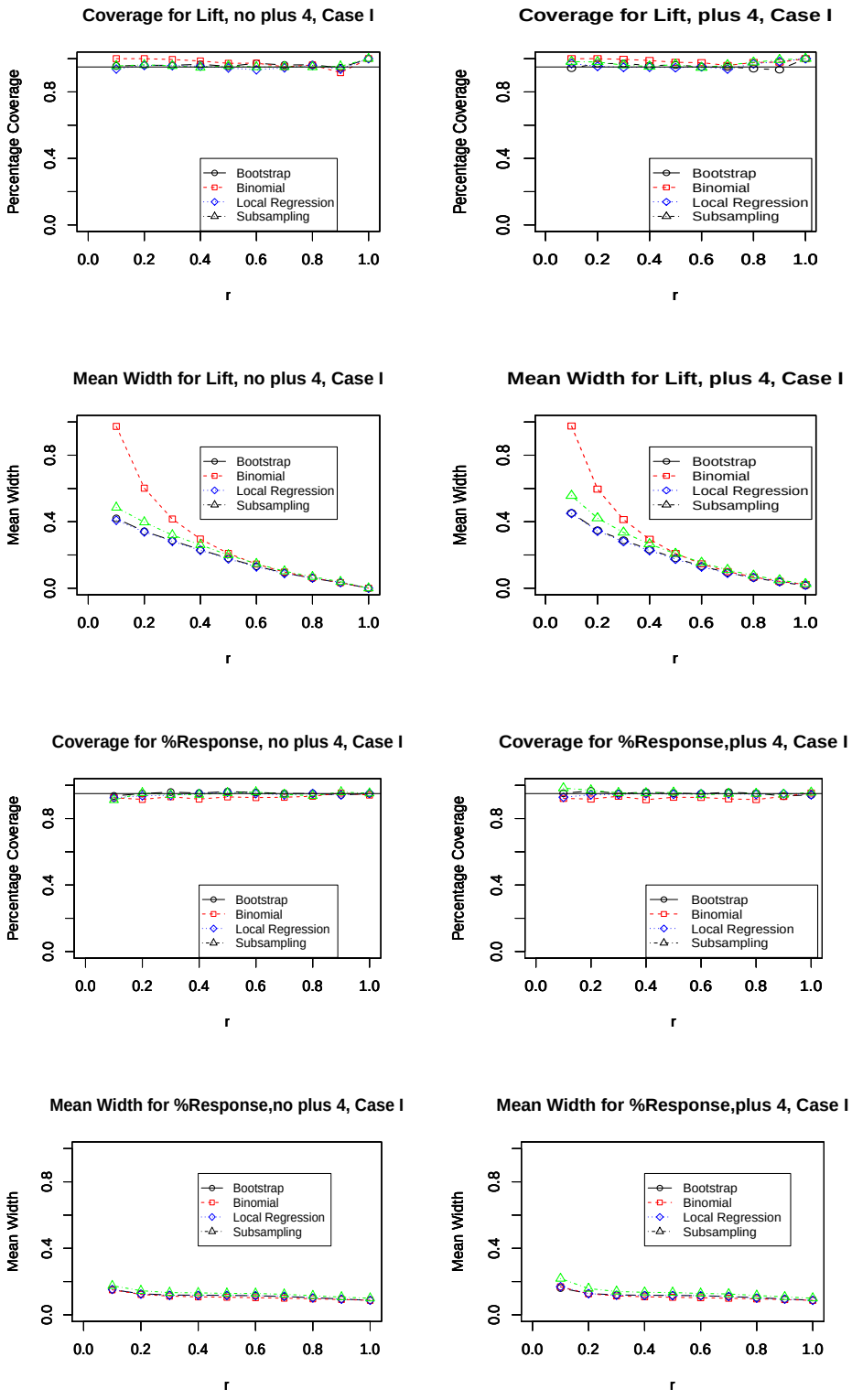


Figure 4: Comparison of four CI methods, without and with plus 4 correction, for Case II. ($m = 500$)

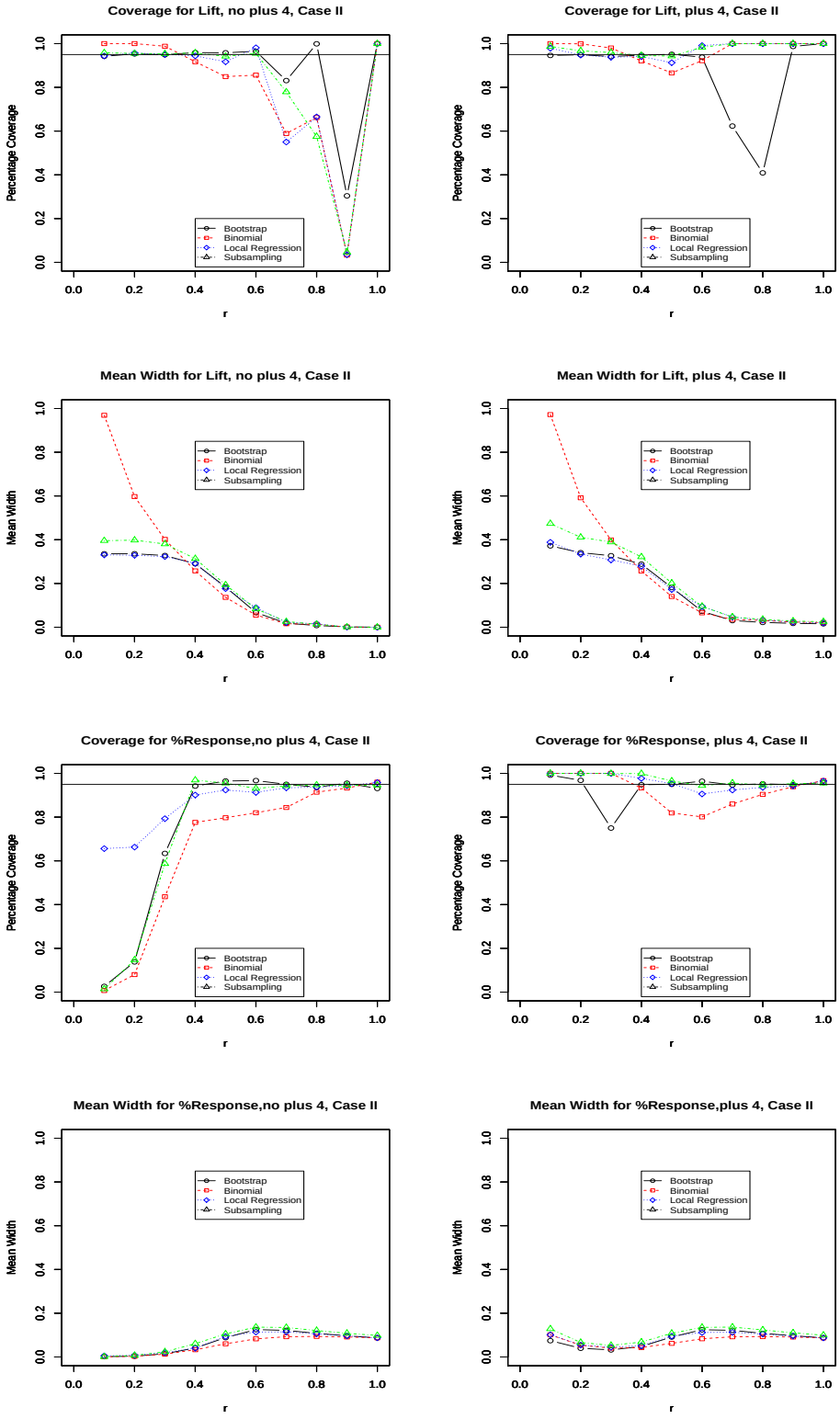


Figure 5: Different h for local method, with plus 4 correction, for Case I.

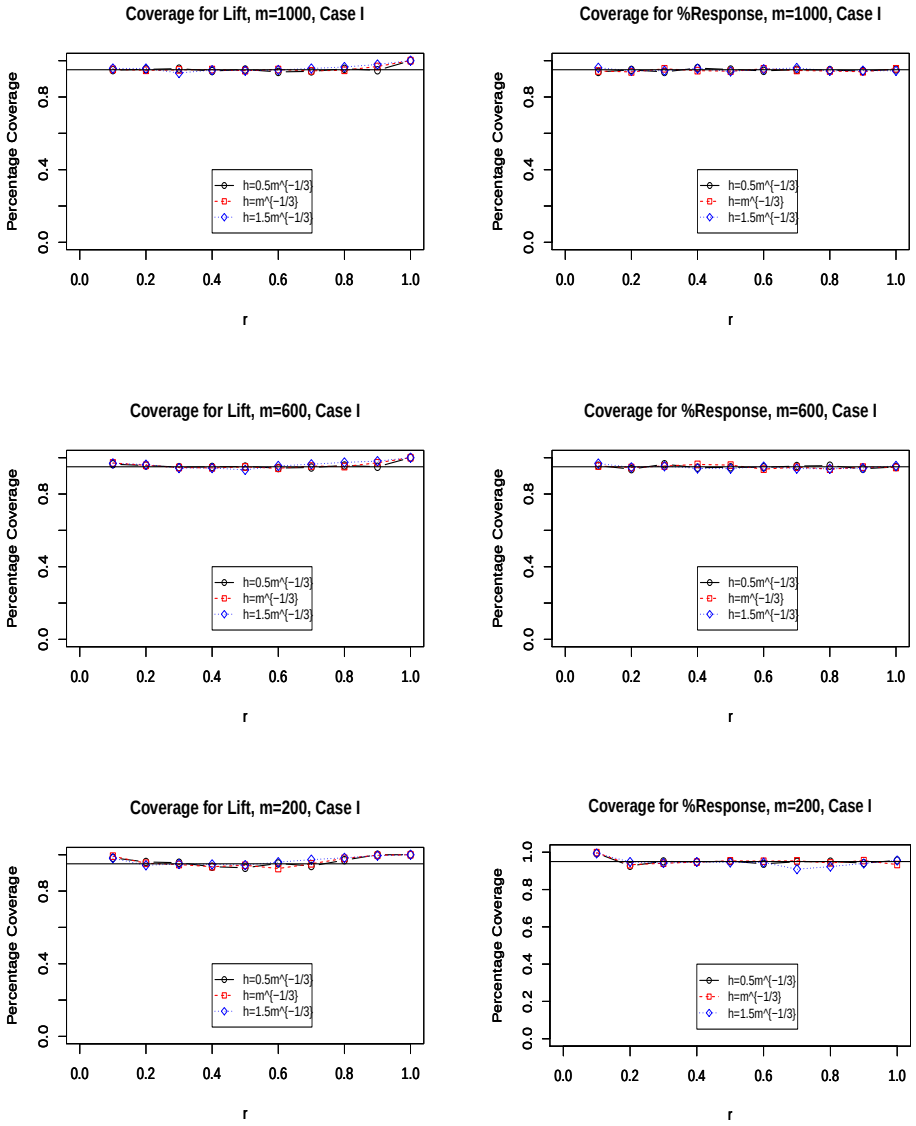


Figure 6: Different h for local method, with plus 4 correction, for Case II.

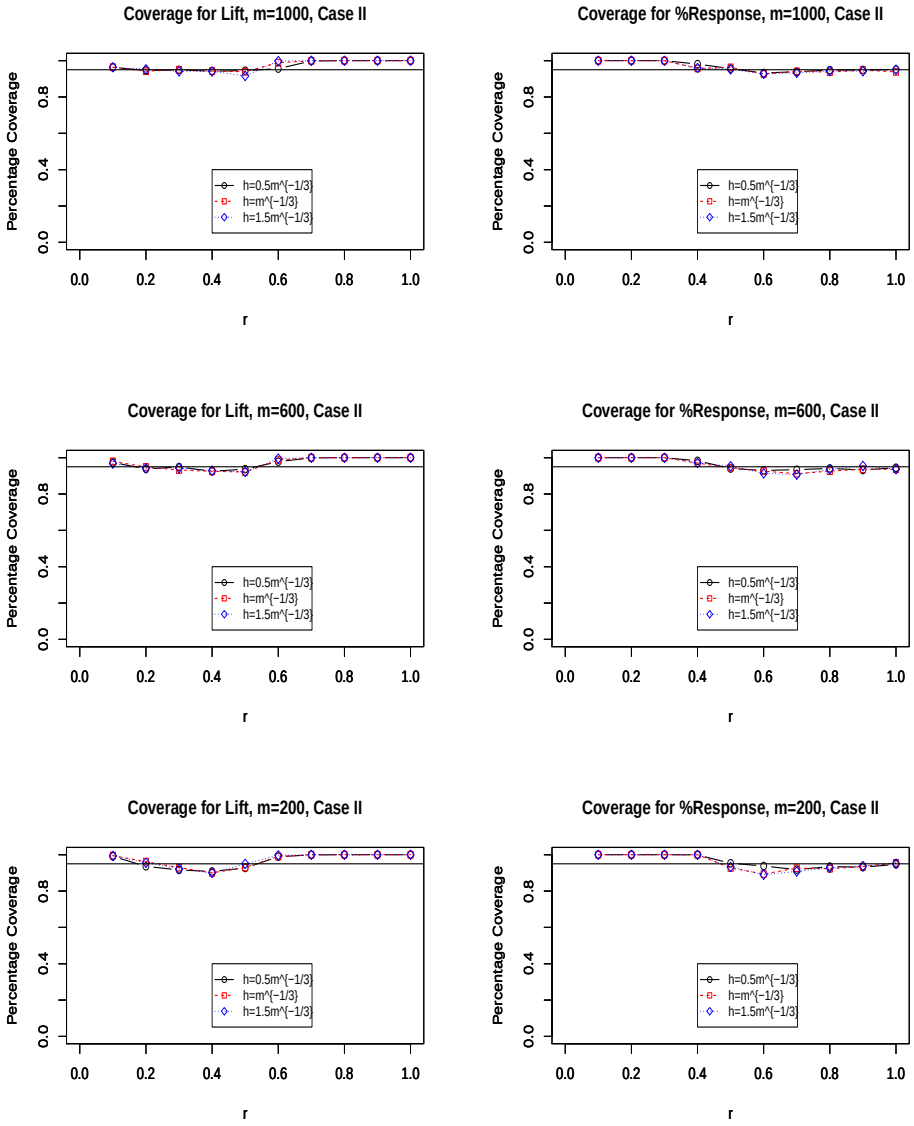


Figure 7: Different q for subsampling method, with approximate plus 4 correction, for Case I.

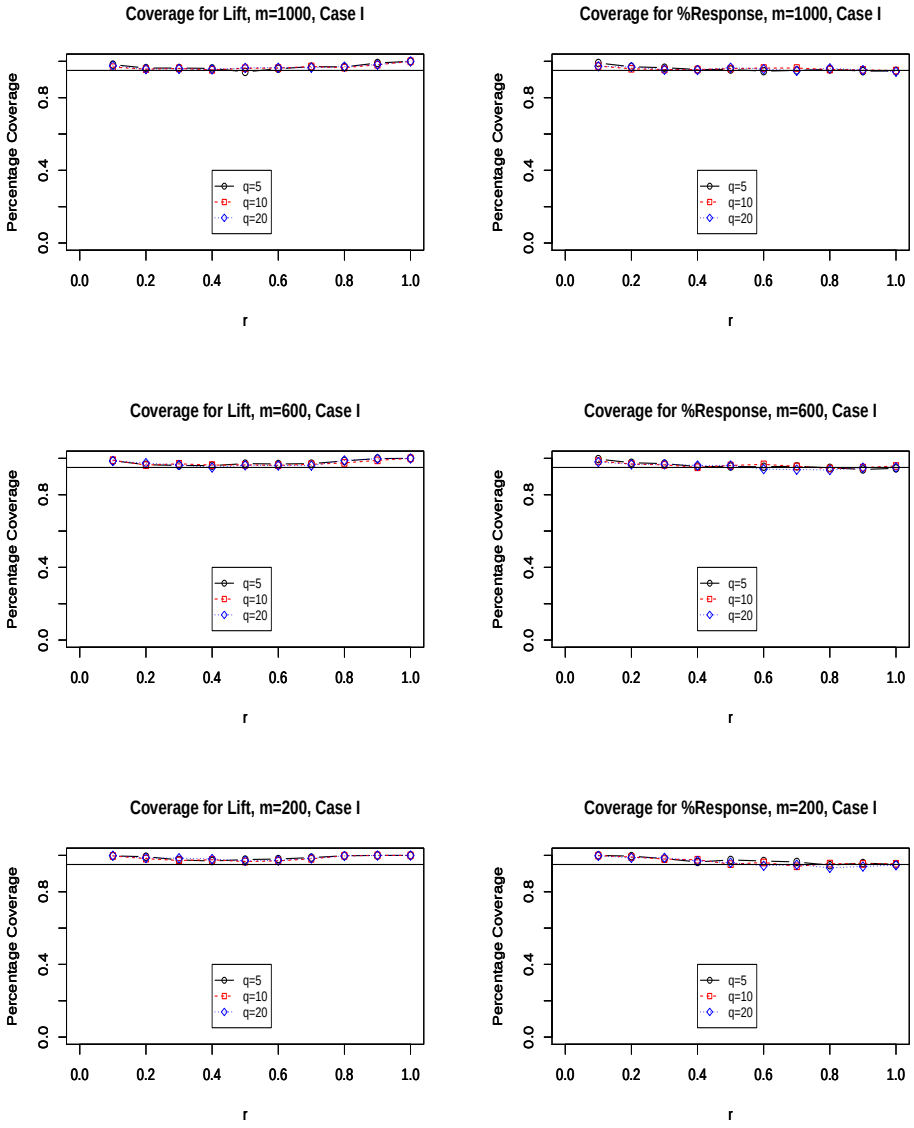
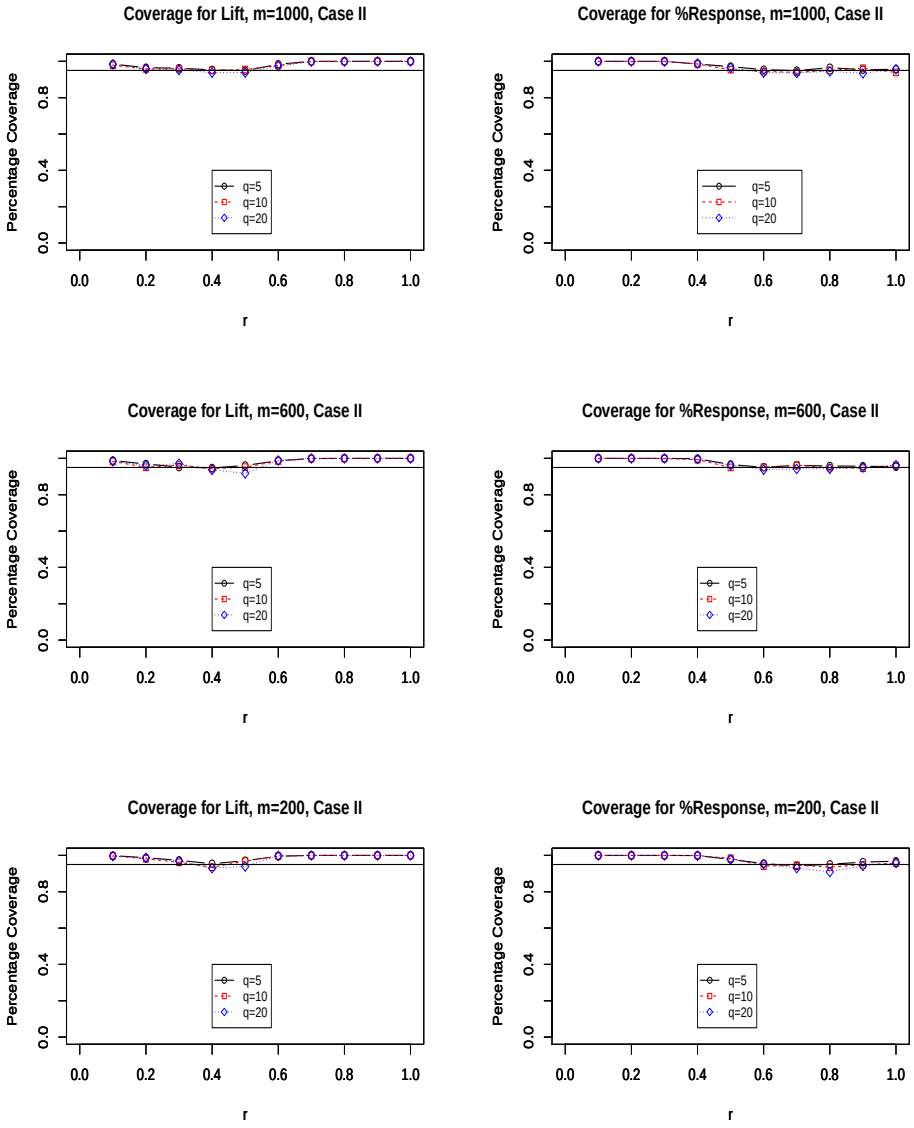


Figure 8: Different q for subsampling method, with approximate plus 4 correction, for Case II.



it is easy to adapt these to have *simultaneous* coverage probabilities at least $1-\alpha$ (asymptotically for large m), by using a Bonferoni coefficient $z_{\alpha/p}$ instead of z_α, since a union bound implies that asymptotically for large m, $P(\theta_k \in \tilde{\theta}_k \pm z_{\alpha/p} s_k \ \forall k = 1, ..., p) \gtrapprox 1-\alpha$. This inequality reflects *conservativeness*, which may lead to wider confidence intervals than needed. Improvement can be obtained as below, based on accounting for the the correlations between the pairs of different $\tilde{\theta}_k$'s.

In this part of the paper we consider how to construct simultaneous confidence intervals with the *correct* asymptotic coverage probability, of the form $\tilde{\theta}_k \pm q_\alpha s_k$, $k = 1, ..., p$, so that $P(\theta_k \in \tilde{\theta}_k \pm q_\alpha s_k \ \forall k) \approx 1-\alpha$ for large m.

Denote $\zeta_k = (\tilde{\theta}_k - \theta_k)/s_k$ for $k = 1, ..., p$. Then $(\zeta_1, ..., \zeta_p)'$ is asymptotically normal with $N(\mathbf{0}, \mathbf{R})$, where $\mathbf{0}$ is a $p \times 1$ vector and $\mathbf{R}_{jk} = corr(\tilde{\theta}_j, \tilde{\theta}_k)$, $j, k \in \{1, ..., p\}$, are matrix elements of correlation coefficients. Let $Z = (Z_1, ..., Z_p)$ be $N(\mathbf{0}, \mathbf{R})$. Suppose we find $q_\alpha(\mathbf{R})$ such that $P[\max_{k=1}^p |Z_k| \leq q_\alpha(\mathbf{R})] = 1-\alpha$ (†). (This equation can be solved by using Monte Carlo simulations of Z's from $N(\mathbf{0}, \mathbf{R})$.) Then $P(\theta_k \in \tilde{\theta}_k \pm q_\alpha(\mathbf{R}) s_k \ \forall k) = P[\max_{k=1}^p |\zeta_k| \leq q_\alpha(\mathbf{R})] \to P[\max_{k=1}^p |Z_k| \leq q_\alpha(\mathbf{R})] = 1-\alpha$ as $m \to \infty$. This means that the confidence coefficients should be $q_\alpha(\mathbf{R})$ now, instead of z_α.

In practice, $\mathbf{R}$ can be replaced by a consistent estimator $\tilde{\mathbf{R}}$ based on the validation sample, where $\tilde{\mathbf{R}}_{jk} = \tilde{cov}(\tilde{\theta}_j, \tilde{\theta}_k)/\sqrt{\tilde{cov}(\tilde{\theta}_j, \tilde{\theta}_j)\tilde{cov}(\tilde{\theta}_k, \tilde{\theta}_k)}$. To estimate the covariances $\tilde{cov}$, we use $\tilde{cov}(\tilde{\theta}_j, \tilde{\theta}_k) = C(r_j, r_k | P_m, \tilde{\Lambda}, \tilde{a}, \tilde{b})$ as explained in Remark 3 of Appendix B, for the selected deciles r_j, r_k in $D = \{0.1, 0.2, ..., 0.9, 1.0\}$.

We have run additional simulations and have confirmed that the method proposed here indeed provides satisfactory *simultaneous* empirical coverage probabilities for both setups described in Section 5. For example, in the gradual case, for ($\%response$, $lift$, $\%captured.response$), the coverage probability of 1000 nominal 95% CIs (based on samples of size $m = 1000$) from the local average method is $(0.942, 0.940, 0.941)$,

simultaneouly for all the deciles from 0.1 to 0.9. The corresponding result in the steep case becomes $(0.963, 0.945, 0.944)$.[5]

The above method for simultaneous confidence intervals is based on the local estimation method. It is not obvious to us how to adapt the other methods (binomial, subsampling, or nonparametric bootstrap) to form simultaneous confidence intervals which are asymptotically correct.

Appendix B: Technical proofs

To prove Proposition 2, we first show a set of more general asymptotic distribution results, in a functional sense, when the LIFT measures are treated as functions of r (the percentage of contacted population) in $[p, q] \subset (0, 1)$, rather than just taking a value at one point of r. Proposition 2 will then be proved as a corollary.

Weak convergence to Gaussian processes:

Denote $W = (Y, S)$ and $W_i = (Y_i, S_i)$. Let $W, W_1, ..., W_n$ be a random sample from a probability distribution P on a measurable space $(\mathcal{W}, \mathcal{A})$, where $\mathcal{W} = \{0, 1\} \times \Re$.

In general, for a probability distribution Q on $(\mathcal{W}, \mathcal{A})$, we denote $Qf = \int f dQ$, $cov_Q(f_1, f_2) = Q(f_1 f_2) - (Qf_1)(Qf_2)$, and $var_Q(f) = cov_Q(f, f)$, where f_1, f_2 and f are measurable function from $\mathcal{W} \mapsto \Re$. In particular, $Pf = \int f dP$, and $P_m f = m^{-1} \sum_{i=1}^{m} f(W_i)$ (corresponding to the empirical distribution P_m on the validation

[5] In these simulations, we noticed that when the population being contacted can have a very high percentage of responses at some decile values of r, a commonly used plus 4 method (see, e.g., a popular textbook Moore 2010, p.508) is needed to achieve a good finite sample performance for the proposed confidence intervals, which involves adding 2 subjects each with $Y = 1$ and $Y = 0$ to the validation data set.

sample $W_1, ..., W_m$).

Let $[p, q] \subset (0, 1)$. In general, for any probability distribution Q on $(\mathcal{W}, \mathcal{A})$, and any continuously differentiable mapping $(\lambda, a, b) : [p, q] \mapsto R^3$ define on $[p, q] \subset (0, 1)$, we denote

$C(r, t|Q, \lambda, a, b) \equiv cov_Q[(Y-\lambda(r))(a(r)I(S > F_Q^{-1}(1-r))+b(r)), (Y-\lambda(t))(a(t)I(S > F_Q^{-1}(1-t))+b(t))]$,

where $F_Q(\cdot) = QI(S \leq (\cdot))$, and $F_Q^{-1}(\cdot) = \inf\{s : F_Q(s) \geq (\cdot)\}$, are respectively the cdf and the quantile function of Q.

For each continuously differentiable $(a, b) : [p, q] \mapsto R^2$, define on $[p, q] \subset (0, 1)$, a zero mean Gaussian process $\{\mathbb{G}_{a,b}(t), t \in [p, q]\}$ with covariance functions, for each $r, t \in [p, q]$:

$E\mathbb{G}_{a,b}(r)\mathbb{G}_{a,b}(t) = cov_P[(Y-\Lambda(r))(a(r)I(S > F_P^{-1}(1-r))+b(r)), (Y-\Lambda(t))(a(t)I(S > F_P^{-1}(1-t))+b(t))] \equiv C(r, t|P, \Lambda, a, b)$,

where $\Lambda(r) = P(Y|S = F_P^{-1}(1-r))$,

and $F_P(\cdot) = PI(S \leq (\cdot))$ is the cdf of S.

The process $\mathbb{G}_{a,b}$ has almost surely continuous sample paths if $\Lambda(r)$ satisfies a Hölder condition, according to a corollary of the Kolmogorov-Chenstov theorem (see, e.g., Exercise 2.3, Lalley 2011). Condition 2 below implies that Λ is continuosly differentiable and that the Hölder condition is satisfied.

Condition 1: $PY > 0$.

Condition 2: The conditional probability density function $P_{S|Y}(s|y)$ is positive and continuously differentiable on $s \in [\bar{p}, \bar{q}] \equiv [F_P^{-1}(p) - \epsilon, F_P^{-1}(q) + \epsilon]$ for some $\epsilon > 0$, for $y = 0, 1$.

For each Q being a probability distribution (such as P or P_m) on the measurable space $(\mathcal{W}, \mathcal{A})$, Let

$F_Q(s) = Q(I(S \leq s)$ and

$\%response_Q(r) = r^{-1}QYI(S > F_Q^{-1}(1-r))$,

$\%captured.response_Q(r) = r * lift_Q(r)$,

$lift_Q(r) = \%response_Q(r)/QY$,

$expected.profit_Q(r) = Qg(Y, I(S > F_Q^{-1}(1-r)))$.

where $r \in [p,q] \subset (0,1)$.

Proposition 4 *Under Conditions 1 and 2, in* $\ell^\infty([p,q])$*, as* $m \to \infty$*, we have:*

(i) $\sqrt{m}[\%response_{P_m} - \%response_P] \rightsquigarrow \mathbb{G}_{a,b}$,
where $(a,b)(r) = (r^{-1}, 0)$*;*

(ii) $\sqrt{m}[\%captured.response_{P_m} - \%captured.response_P] \rightsquigarrow \mathbb{G}_{a,b}$,
where $(a,b)(r) = ((PY)^{-1}, -(PY)^{-1}\%captured.response)$*;*

(iii) $\sqrt{m}[lift_{P_m} - lift_P] \rightsquigarrow \mathbb{G}_{a,b}$,
where $(a,b)(r) = ((rPY)^{-1}, -(rPY)^{-1}\%captured.response)$*;*

(iv) $\sqrt{m}[expected.profit_{P_m} - expected.profit_P] \rightsquigarrow \mathbb{G}_{a,b}$,
where $(a,b)(r) = (g(1,1) + g(0,0) - g(1,0) - g(0,1),\ g(1,0) - g(0,0))$.

Proof :

For each Q being a probability distribution (such as P or P_m) on the measurable space $(\mathcal{W}, \mathcal{A})$, let its bivariate cumulative distribution function (cdf) be denoted as $F_{W,Q}(y,s) = QI[(Y,S) \in (-\infty, y) \times (-\infty, s)]$. The bivariate empirical distribution function of (Y,S) follows a functional central limit theorem : $\sqrt{m}(F_{W,P_m} - F_{W,P})$ converges weakly to a zero mean Gaussian process in $\ell^\infty(\bar{\Re}^2)$, according to Example 2.1.3 on P.82 of van der Vaart and Wellner (1996), who use $\bar{\Re}$ to denote the extended real numbers $[-\infty, +\infty]$.

For Q a bivariate distribution of $W = (Y, S) \in \{0, 1\} \times \Re$, consider (QY, F_Q, G_Q), where $QY = QI(Y > 0.5)$ $G_Q(s) = QYI(S > s) = QI(Y > 0.5, S > s)$, $F_Q(t) = QI(S \leq t)$. They all can be reformulated as a linear transformation of the bivariate cdf $F_{W,Q}$. Therefore, $\sqrt{m}((P_m - P)Y, F_{P_m} - F_P, G_{P_m} - G_P)$ also converges weakly to a limiting Gaussian process on $\Re \times \ell^\infty(\bar{\Re}) \times \ell^\infty(\bar{\Re})$.

Now we relate the four LIFT measures to (QY, G_Q, F_Q):

$\%response_Q(r) = r^{-1}G_Q \circ F_Q^{-1}(1 - r)$,

$\%captured.response_Q(r) = G_Q \circ F_Q^{-1}(1 - r)/QY$,

$lift_Q(r) = r^{-1}G_Q \circ F_Q^{-1}(1 - r)/QY$,

$expected.profit_Q(r) = a_1 G_Q \circ F_Q^{-1}(1 - r) + a_2 QY + a_3 r + a_4$,

where $a_1 = g(1, 1) + g(0, 0) - g(1, 0) - g(0, 1)$ and $a_2 = g(1, 0) - g(0, 0)$, $a_3 = g(0, 1) - g(0, 0)$, $a_4 = g(0, 0)$.

Under Conditions 1 and 2, all the four lift measures $\%response_P$, $\%captured.response_P$, $lift_P$, $expected.profit_P$ are Hadarmard differentiable with respect to (PY, F_P, G_P) on appropriate domains and tangential sets. In particular, the Hadamard differentiability of $G_P \circ F_P^{-1}$ at (F_P, G_P) is established by using the chain rule, the differentiablity of the composite map, and the differentiablility of the inverse map, given respectively in Lemma 3.9.3, Lemma 3.9.27, and Lemma 3.9.23 of van der Vaart and Wellner (1996).

Applying the functional delta method Theorem 3.9.4 of van der Vaart and Wellner (1996) then leads to the corresponding functional Gaussian limiting distributions.

Q.E.D.

Proof of Proposition 2:

Proposition 2 is a straightforward corollary to Proposition 4 at one point r. Q.E.D.

Proof of Corollary 1:

According to Proposition 2, in the case of the estimated $\%response$ $\tilde{\pi}$, the asymptotic variance is $var(\tilde{\pi}) = var(H)/m$, where $H = (Y - \Lambda)A/r$. We have $var[(Y - \Lambda)A] = Evar[(Y - \Lambda)A|A] + varE[(Y - \Lambda)A|A]$, where the first term is equal to $E[\pi(1-\pi)A^2] = \pi(1-\pi)r$, and the second term is equal to $var[(\pi - \Lambda)A] = (\pi - \Lambda)^2 r(1 - r)$. Therefore, the asymptotic variance $var(\tilde{\pi}) = m^{-1}r^{-2}Var[(Y - \Lambda)A] = (mr)^{-1}\pi(1 - \pi) + (mr)^{-1}(\pi - \Lambda)^2(1 - r)$. This leads to (2).

For (3) for $var(\tilde{\kappa}) = var(H)/m$, where κ is the $\%captured.response$, we use $H = (Y - \Lambda)(A - \kappa)/\pi_0$, where $\pi_0 = EY$. Next, we write $var[(A - \kappa)(Y - \Lambda)] = Evar[(A - \kappa)(Y - \Lambda)|Y] + varE[(A - \kappa)(Y - \Lambda)|Y] = \pi_0 var[(A - \kappa)(Y - \Lambda)|Y = 1] + (1 - \pi_0)var[(A - \kappa)(Y - \Lambda)|Y = 0] + varE[(A - \kappa)(Y - \Lambda)|Y] = (1 - \Lambda)^2\kappa(1 - \kappa)\pi_0 + \Lambda^2\kappa'(1 - \kappa')(1 - \pi_0) + \Lambda^2(\kappa' - \kappa)^2\pi_0(1 - \pi_0)$, (*) where $\kappa' = E(A|Y = 0) = E[(1 - Y)A]/E(1 - Y) = r(1 - \pi)/(1 - \pi_0)$. For the last step (*) note that the distribution of $A|Y = 1$ is $Bernoulli(\kappa)$ and $A|Y = 0$ is $Bernoulli(\kappa')$. Then the conditional mean and variance of A given Y can be evaluated using κ and κ', to obtain the three terms of (*). Now expand (*) as a second order polynomial of Λ to obtain $\pi_0\kappa(1-\kappa)(1-2\Lambda)+r(1-r)\Lambda^2$ which leads to the final expression . [The coefficient of Λ^2 was originally complicated but can be identified as $Evar(A|Y)+varE(A|Y) = var(A) = r(1 - r)$.]

Q.E.D.

Remark 3 Let θ_P denote any of the four LIFT parameters corresponding to the true distribution P, and θ_{P_m} denote its sample version corresponding to the sample distribution P_m. Let D be the set of deciles $\{0.1, 0.2, 0.3, ..., 0.9\}$. (The last point 1.0 is omitted here, but it can be considered separately and incorporated in D without changing the covariance formulas, since the parameters involved at $r = 1$ are very

simple, e.g., $\%response(1) = PY$, and $lift(1) = \%captured.response(1) = 1$.)

Then Proposition 4 implies that $\sqrt{m}\{\theta_{P_m}(s)) - \theta_P(s))\}_{s\in D}$ converges in distribution to a zero mean multivariate normal distribution with covariance function $C(r, t|P, \Lambda, a, b)$, with the choice of a, b corresponding to the choice of the LIFT measure as specified in the Proposition.

The covariance function $C(r, t|P, \Lambda, a, b)$ at any $r, t \in [p, q] \subset (0, 1)$ can be consistently estimated by $C(r, t|P_m, \tilde{\Lambda}, \tilde{a}, \tilde{b})$, where $(\tilde{\Lambda}, \tilde{a}, \tilde{b})$ are consistent estimates of (Λ, a, b) at r, t. These can be used for constructing the pointwise confidence intervals as well as the simultaneous confidence intervals of $\theta_P(r)$ on $r \in D$. (See Appendix A.)

References

Haas, P. J. (2006). Lecture Notes #9 on Quantile Estimation, Spring Quarter 2005-06, MS&E 223 - Simulation - Stanford University. http://www.stanford.edu/class/msande223/handouts/lecturenotes09.pdf

Hall, P. G., Hyndman, R. J., and Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika 91, 743-750.

Horváth, L., Horváth, Z., and Zhou, W. (2008). Confidence bands for ROC curves. Journal of Statistical Planning and Inference 138, 1894-1904

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Annals of Statistics 24, 25-40.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. Journal of Business and Economic Statistics 28, 453-468.

Lalley, S. P. (2011). Gaussian processes; Kolmogorov-Chenstov theorem. http://galton.uchicago.edu/∼lalley/Courses/385/GaussianProcesses.pdf

Ma, G. and Hall, W. J. (1993). Confidence bands for receiver operating characteristic curves. Medical Decision Making, 13, 191-197.

Macskassy, S., Provost, F., and Rosset, S. (2005a). Pointwise ROC Confidence Bounds: An Empirical Evaluation. Proceedings of the Workshop on ROC Analysis in Machine Learning (ROCML-2005) at ICML-2005.

Macskassy, S., Provost, F., and Rosset, S. (2005b). ROC confidence bands: an empirical evaluation. Proceedings of the 22nd International Conference on Machine Learning (ICML). Bonn, Germany.

Moore, D. S. (2010). The Basic Practice of Statistics. Fifth Edition. W. H. Freeman, New York.

Moro, S., Laureano , R. , and Cortez., P. (2011). Using data mining for bank direct marketing: an application of the CRISP-DM methodology. In Novais, P. et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimar$\tilde{A}$es, Portugal, October, 2011. EUROSIS.

Rosset, S., Neumann, E., Eick, U., Vatnik, N., and Idan, I. (2001). Evaluation of prediction models for marketing campaigns. KDD-01. p.456-461, ACM Press. http://www.tau.ac.il/∼saharon/papers/Evaluation%20of%20Prediction%20Models.pdf

Stine, R. A., Foster, D. P., and Waterman, R. P. (1998). Business Analysis Using Regression: A Casebook . Springer, New York.

Su, H., Qin, Y. and Liang, H. (2009). Empirical likelihood-based confidence interval of ROC curves. Statistics in Biopharmaceutical Research 1, 407-414.

SAS Institute Inc. (2003). Data Mining Using SAS^R Enterprise Miner^{TM}: A Case Study Approach, Second Edition. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf

van der Vaart, A. W. (2000). Asymptotic Statistics. Cambridge University Press, Cambridge.

van der Vaart, A. W., and Wellner, J. (1996). Weak Convergence and Empirical Processes, Springer, New York.

Zhao, Y. (2014). On Asymptotic Distributions and Confidence Intervals for LIFT Measures in Data Mining. Ph.D. Dissertation, Northwestern University.