Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 23

11-1-2007

A Simple Method for Finding Emperical Liklihood Type Intervals for the ROC Curve

Ayman Baklizi
Qatar University, a.baklizi@qu.edu.qa

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 589-595

1538 – 9472/07/$95.00

A Simple Method For Finding Empirical Likelihood Type Intervals For The ROC Curve

Ayman Baklizi
Qatar University

Interval estimation of the ROC curve is considered using the empirical likelihood techniques. Suggested is a procedure that is very simple computationally and avoids the constrained optimization problems usually faced with empirical likelihood methods. Various modifications are suggested and the performance of the intervals is evaluated in terms of their coverage probability. The results show tat some of the suggested intervals compete well with other intervals known in the literature.

Key words: ROC curve, empirical likelihood, kernel estimators, bootstrap

Introduction

The Receiver Operating Characteristic (ROC) curve is used to assess the accuracy of a diagnostic test in discriminating between healthy and diseased individuals. A threshold value c is determined, and people with test measurements greater than c are classified as diseased, otherwise as healthy. Let X be a random variable representing the test score of a healthy individual and let Y be the score of a diseased patients. Let F and G be the distribution functions of X and Y respectively. The sensitivity of the test is defined as $1-G(c)$. It is the probability that the test score of diseased patient is greater than c. The specificity of the test is defined as $F(c)$, it is the probability of correctly classifying a healthy individual. The receiver operating characteristic curve is defined as the plot of $1-F(c)$ against $1-G(c)$ as c varies from $-\infty$ to ∞ or equivalently as the plot of $1-G\left(F^{-1}(1-t)\right)$ where $0 \le t \le 1$, (Hsieh and Turnbull, 1996).

Ayman Baklizi is an Associate Professor of statistics, and is at the department of mathematics and physics at Qatar University. He is an elected-member at the ISI. Email: bakliziI@yahoo.com

The estimation of the ROC curve has received considerable attention. The problem has been considered in parametric, nonparametric and semi-parametric situations. For example, see Hsieh and Turnbull (1996), Li et al. (1999), Hall et al., (2003).

Claeskens et al. (2003) developed empirical likelihood confidence regions for the ROC curve. Let $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ be two random samples from the distributions F and G respectively. Define the ROC curve as $R(t)=1-G\left(F^{-1}(1-t)\right)$ where $0 \le t \le 1$ and let $\theta = R(t)$, Claeskens et al. (2003) constructed confidence intervals for θ using the smoothed empirical likelihood function

$$L(\theta)=\sup_{(p,q,\eta)}\left(\prod_{i=1}^{n} p_i\right)\left(\prod_{j=1}^{m} q_j\right),$$

where $p' = (p_1,\ldots,p_n)$ and $q' = (q_1,\ldots,q_m)$ are probability vectors each summing to one and subject to certain constraints on the smoothed versions of the empirical distributions of X and Y. They showed that the asymptotic distribution of the log-likelihood ratio $l(\theta) = -2\log L(\theta)$ is chi square with one degree of freedom and conducted some simulations to investigate the performance of their intervals and show that it performs better than some other asymptotic and bootstrap intervals.

Purpose

An alternative procedure is suggested here based on the empirical likelihood which is very simple computationally, does not need numerical constrained optimization, and produces interval estimates that are, in some cases, about as accurate as those of Claeskens et al. (2003). This procedure and some modifications are described. A simulation experiment was conducted to investigate and compare the suggested procedure with other well known procedures.

Empirical Likelihood Based Intervals

Assume that an interval estimator of $R(t^*) = 1 - G(F^{-1}(1-t^*))$ is desired where t^* is some specific point in the unit interval. Proceed in two stages as follows; in the first stage obtain a point estimator for $F^{-1}(1-t^*)$. This is equivalent to estimating x_{1-t^*}: $(1-t^*)^{th}$ quantile of F denote this estimator by $\hat{x}_{1-t^*}$. In the second stage obtain an interval estimator of $\overline{G}(\hat{x}_{1-t^*}) = 1 - G(\hat{x}_{1-t^*})$ which is the right tail probability of the random variable Y having distribution function G.

In an empirical likelihood setup, the first stage amounts to estimating x_{1-t^*} which may be done using interpolation between the values of the ordered statistics of the sample of the distribution of X. In the second stage consider the empirical likelihood function for quantiles (Owen, 2001) given by

$$R(p,q) = \max\left\{\prod_{i=1}^{m} nw_i \mid \sum_{i=0}^{m+1} w_i Z_i(p,q) = 0, w_i \geq 0, \sum_{i=0}^{m+1} w_i = 1\right\}$$

where $0 \leq p \leq 1$, $-\infty < q < \infty$ is the p^{th} quantile and $Z_i(p,q) = I_{(X_i \leq q)} - p$. Substituting $\hat{x}_{1-t^*}$ for q and $G(\hat{x}_{1-t^*})$ for p and, conditional on $\hat{x}_{1-t^*}$ using the empirical likelihood function $R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*})$ one can construct confidence interval for $G(\hat{x}_{1-t^*})$ as

$$\left\{G(\hat{x}_{1-t^*}) \mid -2\log R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > \chi^2_{\alpha,1}\right\}$$

and then transform it to a confidence interval for $1 - G(\hat{x}_{1-t^*})$ this results in a confidence interval for $R(t^*)$ Call this interval the (EL) interval.

The chi square calibration used in the empirical likelihood interval may be replaced by the E-Calibration of Tsao (2004). This calibration is based on the quantiles of a new family of distributions arising from the normal distribution. It is derived using the finite sample similarity between the empirical and parametric likelihoods. Some quantiles $e_{\alpha,1,m}$ of that distribution are given in Tsao (2004). The E – calibration corrects for under coverage resulting from using the chi square calibration. The new interval (EC interval) based on this calibration is given by

$$\left\{G(\hat{x}_{1-t^*}) \mid -2\log R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > e_{\alpha,1,m}\right\}$$

Another modification may be obtained by using the "smoother" version of the empirical likelihood function for quantiles introduced by Adimari (1998). In this modification the empirical likelihood is replaced by a smoother version which, when considered as a function of $G(\hat{x}_{1-t^*})$, may be written as

$$-2\log \tilde{R}\left(G\left(\hat{x}_{1-t^*}\right), \hat{x}_{1-t^*}\right) = 2m\left[\tilde{G}\left(\hat{x}_{1-t^*}\right)\log\left(\frac{\tilde{G}\left(\hat{x}_{1-t^*}\right)}{G\left(\hat{x}_{1-t^*}\right)}\right) + \left(1-\tilde{G}\left(\hat{x}_{1-t^*}\right)\log\left(\frac{1-\tilde{G}\left(\hat{x}_{1-t^*}\right)}{1-G\left(\hat{x}_{1-t^*}\right)}\right)\right)\right]$$

where

$$\tilde{G}(\hat{x}_{1-t^*}) = \begin{cases} G^*(\hat{x}_{1-t^*}) & \text{if } [Y_{(1)}, Y_{(m)}] \text{ contains } \hat{x}_{1-t^*} \\ \hat{G}(\hat{x}_{1-t^*}) & \text{otherwise} \end{cases}$$

where $G^* = \dfrac{2i-1}{2n}$ on each $Y_{(i)}$ and is linear in each $\left[Y_{(i)}, Y_{(i+1)}\right]$, and where $Y_{(1)}, \ldots, Y_{(n)}$ are the order statistics of the sample of Y values. Adimari showed that the limiting distribution is also χ_1^2. A $(1-\alpha)\%$ confidence interval for $G(\hat{x}_{1-t^*})$ (AD interval) is given by

$$\left\{G(\hat{x}_{1-t^*}) \mid -2\log \widetilde{R}(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > \chi_{1,\alpha}^2\right\}$$

Simulation

Simulation studies were conducted to assess the performance of the interval estimates based in the empirical likelihood. Also considered were the bootstrapped version of the empirical likelihood interval (BEL), and the bootstrapped version of the (AD) interval, the (BTAD) interval. A Bartlett type correction factor is obtained as the mean of the B bootstrap empirical log-likelihood ratios which in turn used to find the (BRT) interval. The simulation design used similar to those used by Claeskens et al.(2003) and Hall et. al. (2003). The coverage probability were investigated at values of t = 0.1, 0.3, 0.5, 0.7 and 0.9 with sample size (n,m) = (30,30), (50,50), (70, 70), (100,100), (50,70) and (70,50). In each case 2000 pair of samples is generated from

1- $X \sim N(0,1)$, $Y \sim N(1,1)$

2- $X \sim \Gamma(2)$, $Y \sim \Gamma(3)$

3- $X \sim t(5)$,

$Y \sim 0.2(t(5)-1)+0.8(t(5)-1)$

B = 500 is used in bootstrap calculations. The coverage probabilities of the intervals with nominal confidence levels $(1-\alpha)=0.90$ and 0.95 are given in Tables 1-3.

Result

The results are given in tables 1 – 3 where the following abbreviations are used EL: The empirical likelihood interval based on the asymptotic χ^2 approximation. BEL: The empirical likelihood interval based on bootstrap cririical values.

EC: The empirical likelihood interval based on Tsao's E-Calibration. BRT: The empirical likelihood interval with the bootstrap Bartlett type correction. AD: The empirical likelihood interval based on Adimari's modification.

BTAD: The empirical likelihood interval based on Adimari's modification and bootstrap critical values.

Conclusion

It appears that the coverage probabilities of the intervals are close to the nominals for small values of t. For larger values of t most intervals tend to have an undercoverage problem. Exceptions are the bootstrapped empirical likelihood interval (BEL) and the corrected interval (BRT). These two intervals tend to be conservative for larger values of t. A drawback of the (BEL) interval is that it has a very low coverage probability for small values of t when the sample sizes differ.

This is not the case with the (BRT) interval. These observations are also applicable to the results given in tables 2 and 3. The BRT in most cases have the closest coverage probability to nominal. Comparison of these results with Hall et. al. (2003) and Claeskens et al. (2003) shows that the (BRT) interval considered in this article competes very well with theirs in terms of its coverage probability. The simplicity of the methods discussed in this article and the avoidance of complicated restricted optimization problems or sophisticated bandwidth rules used for the construction of kernel based intervals may balance the slightly better performance of the Hall et al. (2003) or Claeskens et. al. (2003) intervals.

References

Adimari, G. (1998). An empirical likelihood statistic for quantiles. *Journal of Statistical Computation and Simulation*, *60*, 85 – 95.

Hall, P., Hyndman, R. & Fan, Y. (2003). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, *91*, 3, 743 – 750.

Claeskens, G., Jing, B. Peng, L., & Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics*, *31*, 173 – 190.

Table 1. Coverage Probabilities of the Intervals, The Normal Distribution

n	m	t	EL	BEL	EC	BRT	AD	BTAD	EL	BEL	EC	BRT	AD	BTAD
					$\alpha = 0.10$						$\alpha = 0.05$			
30	30	0.1	887	182	887	791	926	932	887	228	967	900	951	961
		0.3	743	741	743	779	761	768	794	791	946	803	789	947
		0.5	816	852	816	872	805	817	864	903	864	906	877	899
		0.7	719	881	743	903	767	789	840	928	840	931	848	863
		0.9	600	839	600	876	676	695	623	921	705	917	768	788
50	50	0.1	927	378	927	878	920	925	927	411	927	917	951	954
		0.3	852	809	852	824	848	855	905	864	905	872	905	906
		0.5	757	906	819	926	803	811	868	951	868	953	874	884
		0.7	729	914	729	950	746	756	811	965	811	966	832	841
		0.9	651	890	651	948	692	700	731	965	731	971	783	791
70	70	0.1	905	531	905	927	936	937	966	561	966	956	960	961
		0.3	855	874	855	890	836	840	855	909	892	919	922	925
		0.5	783	924	783	952	790	797	860	961	860	971	866	870
		0.7	756	932	756	971	761	765	838	978	838	984	838	844
		0.9	662	907	662	972	692	694	732	974	732	984	773	777
100	100	0.1	941	658	941	945	938	938	941	677	941	970	967	967
		0.3	830	895	830	918	833	833	880	940	880	941	906	906
		0.5	800	946	800	978	803	803	868	978	868	986	882	882
		0.7	719	945	719	986	743	743	832	988	832	992	832	832
		0.9	646	929	646	980	661	661	722	980	722	991	752	752
50	70	0.1	878	506	878	901	918	922	947	540	947	939	946	948
		0.3	815	847	815	867	812	813	815	889	859	899	902	907
		0.5	747	910	747	943	763	768	833	959	833	959	836	842
		0.7	708	911	708	958	715	724	791	973	791	978	800	800
		0.9	587	885	587	946	610	620	646	962	646	971	694	701
70	50	0.1	958	397	958	924	952	954	958	435	958	953	972	974
		0.3	869	815	869	823	864	870	912	869	912	875	907	911
		0.5	780	909	845	925	823	835	881	954	881	951	890	899
		0.7	791	928	791	962	789	797	871	976	871	979	867	874
		0.9	710	910	710	957	735	743	786	965	786	976	812	818

Table 2. Coverage Probabilities of the Intervals, Asymmetric Distributions Case

n	m	t	EL	BEL	EC	BRT	AD	BTAD	EL	BEL	EC	BRT	AD	BTAD
					$\alpha = 0.10$						$\alpha = 0.05$			
30	30	0.1	922	292	922	844	919	925	922	346	922	884	950	955
		0.3	761	767	840	779	822	833	840	828	840	825	867	878
		0.5	769	871	769	892	754	769	820	918	868	920	844	863
		0.7	704	878	704	913	725	748	735	933	813	937	812	834
		0.9	561	828	561	853	698	711	649	905	649	895	776	802
50	50	0.1	857	516	857	885	914	918	943	546	943	932	950	951
		0.3	796	840	796	874	807	816	850	898	850	899	875	881
		0.5	749	905	749	949	764	775	842	967	842	964	841	850
		0.7	691	928	691	959	741	749	767	970	780	978	820	826
		0.9	601	881	675	920	713	721	740	945	757	949	787	796
70	70	0.1	906	648	906	921	912	915	906	681	906	950	946	949
		0.3	769	878	769	904	790	799	873	926	873	928	874	884
		0.5	723	929	762	973	760	765	804	980	832	986	840	849
		0.7	689	934	736	978	732	737	820	979	820	990	810	816
		0.9	638	900	638	947	677	684	723	965	739	965	765	770
100	100	0.1	733	703	733	752	729	729	923	762	923	969	948	948
		0.3	805	924	805	952	806	806	875	961	875	967	875	875
		0.5	758	943	758	984	761	761	850	984	850	991	839	839
		0.7	729	945	729	988	715	715	774	988	774	994	807	807
		0.9	631	926	631	971	680	680	732	978	732	985	768	767
50	70	0.1	874	659	874	895	900	901	874	687	874	924	939	943
		0.3	738	872	738	898	761	766	844	924	844	927	837	841
		0.5	690	920	727	958	726	732	769	967	805	976	814	819
		0.7	647	924	697	971	711	717	772	973	772	981	794	799
		0.9	566	860	566	922	641	643	632	940	645	949	719	728
70	50	0.1	880	532	880	898	927	930	960	566	960	938	958	960
		0.3	824	854	824	866	817	824	870	894	870	898	887	898
		0.5	795	921	795	956	797	806	876	961	876	968	875	885
		0.7	746	939	746	980	784	792	830	981	843	988	866	874
		0.9	662	899	729	937	737	747	789	957	811	961	818	827

Table 3: Coverage Probabilities of the Intervals, Mixture Distributions Case

n	m	t	EL	BEL	EC	BRT	AD	BTAD	EL	BEL	EC	BRT	AD	BTAD
					$\alpha = 0.10$						$\alpha = 0.05$			
30	30	0.1	808	792	808	804	844	852	861	850	861	844	873	879
		0.3	820	899	877	921	846	858	877	938	877	949	912	930
		0.5	828	927	828	956	854	874	914	964	914	972	916	928
		0.7	755	903	755	937	792	809	825	960	842	961	864	884
		0.9	590	824	590	851	678	692	699	906	699	906	764	783
50	50	0.1	838	866	838	886	826	834	838	902	838	917	907	913
		0.3	863	934	863	962	842	848	895	968	927	971	907	914
		0.5	811	952	851	984	846	853	911	985	911	990	909	916
		0.7	813	937	813	977	798	807	879	978	879	986	869	876
		0.9	595	880	595	919	668	679	685	946	685	951	749	760
70	70	0.1	828	887	828	906	815	820	867	928	867	939	891	895
		0.3	841	951	881	984	852	857	908	982	908	988	917	921
		0.5	837	956	837	988	830	836	893	987	893	994	901	905
		0.7	807	962	807	990	803	808	872	991	885	995	880	882
		0.9	604	907	662	957	663	672	734	968	734	974	752	762
100	100	0.1	787	919	787	941	811	811	873	959	873	959	889	889
		0.3	843	965	843	991	853	853	924	989	924	997	923	923
		0.5	834	970	834	996	843	843	899	995	899	998	907	907
		0.7	775	967	775	996	799	799	873	993	873	998	868	868
		0.9	634	908	634	968	662	662	686	977	686	986	746	746
50	70	0.1	780	858	780	887	781	786	827	908	827	922	858	861
		0.3	832	949	871	977	840	846	894	981	894	986	900	904
		0.5	820	964	820	992	810	815	875	990	875	997	882	886
		0.7	766	943	766	979	760	767	819	981	836	990	841	846
		0.9	521	870	570	929	600	605	638	946	638	956	696	702
70	50	0.1	859	869	859	884	839	851	859	903	859	911	918	924
		0.3	886	940	886	965	867	871	912	975	946	979	931	938
		0.5	818	962	853	980	847	853	920	985	920	990	915	922
		0.7	834	942	834	983	824	835	907	983	907	990	897	904
		0.9	656	889	656	949	715	724	733	958	733	971	793	801

Hsieh, F., & Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, *24*, *1*, 25 – 40.

Li, G., Tiwari, R., & Wells, M. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curve. *Biometrika, 86, 3*, 487 – 502.

Owen, A. B. (2001). *Empirical Likelihood*. NY: Chapman and Hall.

Tsao, M. (2004). A new method of calibration for the empirical loglikelihood ratio. *Statistics and Probability Letters*, *68*, 305 – 314.