

Evaluation of Probability Forecasts

Tze Leung Lai

Stanford University

Joint work with
Shulamith Gross, Bo Shen,
and Kevin Sun

Joint Statistical Meetings

August 2010

Outline

- Prediction of binary outcomes in epidemiology: risk models, cohort studies, PPV and NPV, predictiveness curve, AUC and variants to compare predictors
- Evaluation of probability forecasts in meteorology: scores, MOS of US National Weather Service
- A new unified approach
- Illustrative applications of the approach

Prediction of Binary Outcomes in Epidemiology

- Logistic regression models for probability of disease (“risk”) given covariate vector (biomarkers); Cox regression for censored data
 - Prentice & Pyke (1979): Asymptotics can be applied to case-control data as if the data had been obtained in a prospective study
 - Marker considered useful if it has strong effect on risk as measured by a logistic regression model

- Classification performance measures for biomarkers
 - Fraction of diseased subjects detected by marker (i.e., sensitivity) and fraction of non-diseased subjects falsely identified as diseased (i.e., 1–specificity)
 - **Pepe et al. (2004)**: A marker that is strongly related to risk may be a poor classifier
 - **Pencina, D'Agostino et al. (2008)**: Area under ROC curve (AUC), net reclassification improvement (NRI), integrated discrimination improvement (IDI); NRI and IDI show HDL cholesterol offers significant improvement in the performance of a coronary heart disease model in the Framingham Heart Study.

- Predictiveness curve (Huang, Pepe & Feng, 2007; Pepe et al., 2007; Gu & Pepe, 2009)
 - Predictive accuracy of single marker X with cdf F :

$$R(v) = P\{D = 1|X = F^{-1}(v)\}$$

$$P(D = 1|X) = e^{\beta_0 + \beta_1 X} / (1 + e^{\beta_0 + \beta_1 X})$$
 - Cystic Fibrosis Registry: annually updated data on over 20,000 people diagnosed with cystic fibrosis in the US. Data from 11,960 patients, whose weight and measure FEV₁ of pulmonary function in 1995 are used to predict pulmonary exacerbations in 1996. Use a semiparametric location-scale model for the distribution of X given Y to obtain $R_y(v) = P\{D = 1|X = F_y^{-1}(v), Y = y\}$ after stratifying Y into 3 classes.

- Predictiveness curve plots the disease risk versus normalized risk rank
- Prostate Cancer Prevention Trial, 1993–2003: 5,519 men on the placebo arm had prostate biopsy and PSA measurements in the 3 years prior to biopsy. Fit logistic regression model relating risk of high-grade disease (4.7% of the 5,519 men) to PSA and other covariates. Predictiveness curve for PSA alone is found to be almost identical to that for PSA and other risk factors.
- Positive and negative predictive values of biomarker (Moskowitz & Pepe, 2004):
$$\text{PPV}(v) = P\{D = 1|F(X) > v\}$$
$$\text{NPV}(v) = P\{D = 0|F(X) \leq v\}$$

- Performance of clinical decision rules based on a risk model can be evaluated by loss functions (Gail and Pfeiffer, 2005) and model's predictiveness curve (Pepe et al., 2007).
- Wacholder, Prentice, Gail et al. (2010) used information on traditional risk factors and 10 common genetic variants associated with breast cancer in 5,590 case subjects and 5,998 control subjects, from 4 US cohort studies and 1 case-control study in Poland, to fit models of absolute risk of breast cancer. The results based on AUC showed that the inclusion of newly discovered genetic factors modestly improved the performance of risk models for breast cancer.

Evaluation of Probability Forecasts in Meteorology

- US National Weather Service
 - Transition from non-probabilistic to probability prediction (Murphy & Winkler, 1984)
 - Model Output Statistics (MOS)
- Reliability and accuracy measures for probability forecasts: Direct accuracy assessment is difficult because it requires comparing a forecaster's predicted probabilities with the actual but unknown probabilities of the events under study.

- Reliability is measured using “scoring rules”, which are empirical distance measures between repeated predicted probabilities of an event, such as having no rain the next day, and indicator variables that take on the value 1 if the predicted event actually occurs, and 0 otherwise.
- A scoring rule for a sequence of n probability forecasts \hat{p}_i , $i = 1, \dots, n$, is the average score $n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$, where $Y_i = 1$ or 0 according to whether the i th event actually occurs or not. In particular, the popular score

$$L(y, \hat{p}) = (y - \hat{p})^2$$

was proposed by [Brier \(1950\)](#).

- Skill score: Widely used to evaluate weather forecasts is the percentage improvement in average score over that provided by climatology, denoted by \hat{p}_i^c and considered as an “unskilled” forecaster, i.e.,

$$S_n = \frac{n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i^c) - n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)}{n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i^c)}.$$

Climatology refers to the historic relative frequency, also called the base rate, of precipitation; we can take it to be

$$\hat{p}_i^c = (M + 1)^{-1} \sum_{t=-M}^0 Y_t.$$

It is not a “proper” score, and Winkler (1994) proposed to replace the average climatology score in the denominator by individual weights $l(\hat{p}_i, \hat{p}_i^c)$, i.e.,

$$W_n = n^{-1} \sum_{i=1}^n \frac{L(Y_i, \hat{p}_i) - L(Y_i, \hat{p}_i^c)}{l(\hat{p}_i, \hat{p}_i^c)}$$

where

$$l(p, c) = \{L(1, p) - L(1, c)\} I_{\{p \geq c\}} + \{L(0, p) - L(0, c)\} I_{\{p < c\}}.$$

- Reliability diagram groups probability forecasts in bins (buckets) and plots the average Y_i over each bin.

A New Unified Approach

- The function L in the scoring rule

$$n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i)$$

measures the reliability of the probability forecast \hat{p}_i of event i before the indicator variable Y_i of the event is observed. We can also use L as a loss function in measuring the accuracy of \hat{p}_i as an estimate of the probability p_i of event i :

$$L_n = n^{-1} \sum_{i=1}^n L(p_i, \hat{p}_i).$$

- Besides the squared error loss

$$L(p, \hat{p}) = (p - \hat{p})^2$$

in Brier's score, another widely used loss function is the Kullback–Leibler divergence (Good, 1952):

$$L(p, \hat{p}) = p \log(p/\hat{p}) + (1 - p) \log[(1 - p)/(1 - \hat{p})]$$

- A loss function $\tilde{L}(p, \hat{p})$ is a *linear equivalent* of the loss function $L(p, \hat{p})$ if $\tilde{L}(p, \hat{p})$ is a linear function of p and

$$L(p, \hat{p}) - \tilde{L}(p, \hat{p}) \text{ does not depend on } \hat{p}.$$

For example,

$$\tilde{L}(p, \hat{p}) = -2p\hat{p} + \hat{p}^2$$

is a linear equivalent of the squared error loss $(p - \hat{p})^2$.

A linear equivalent \tilde{L} of the Kullback–Leibler divergence is given by

$$-\tilde{L}(p, \hat{p}) = p \log(\hat{p}) + (1 - p) \log(1 - \hat{p}).$$

- Allowing forecast \hat{p}_k to depend on an information set \mathcal{F}_{k-1} consisting of the event and forecast histories and other covariates before Y_k is observed, the conditional distribution of Y_i given \mathcal{F}_{i-1} is Bernoulli(p_i), and thus

$$P(Y_i = 1 | \mathcal{F}_{i-1}) = p_i.$$

- Suppose $L(p, \hat{p})$ is linear in p , as in the case of linear equivalents of general loss functions. Then

$$E \{L(Y_i, \hat{p}_i) | \mathcal{F}_{i-1}\} = L(p_i, \hat{p}_i),$$

so $L(Y_i, \hat{p}_i) - L(p_i, \hat{p}_i)$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}$. Martingale theory can then be used to prove the following.

Theorem 1. *Suppose $L(p, \hat{p})$ is linear in p . Letting*

$$\sigma_n^2 = n^{-1} \sum_{i=1}^n \{L(1, \hat{p}_i) - L(0, \hat{p}_i)\}^2 p_i(1 - p_i),$$

assume that $\sigma_n^2 = O(1)$ with probability 1. Then $\hat{L}_n - L_n$ converges to 0 with probability 1. If σ_n^2 converges in probability to some non-random positive constant, then $\sqrt{n}(\hat{L}_n - L_n)/\sigma_n$ has a limiting standard normal distribution.

- To apply **Theorem 1** to construct confidence intervals for L_n , one needs to estimate the unknown p_i in σ_n^2 . A conservative confidence interval for L_n can be obtained by replacing $p_i(1 - p_i)$ by its upper bound $1/4$.
- Consider two sequences of probability forecasts

$$\hat{\boldsymbol{p}}' = (\hat{p}'_1, \dots, \hat{p}'_n) \quad \text{and} \quad \hat{\boldsymbol{p}}'' = (\hat{p}''_1, \dots, \hat{p}''_n)$$

of $\boldsymbol{p} = (p_1, \dots, p_n)$. Suppose a loss function $L(\boldsymbol{p}, \boldsymbol{q})$ is used to evaluate each forecast, and let $\tilde{L}(\boldsymbol{p}, \boldsymbol{q})$ be its linear equivalent. Then

$$L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i) = \tilde{L}(p_i, \hat{p}'_i) - \tilde{L}(p_i, \hat{p}''_i)$$

is a linear function of p_i , so we can estimate

$$\Delta_n = n^{-1} \sum_{i=1}^n \{L(p_i, \hat{p}'_i) - L(p_i, \hat{p}''_i)\}$$

by the difference

$$n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}_i) - n^{-1} \sum_{i=1}^n L(Y_i, \hat{p}'_i)$$

of scores of the two forecasts. Application of **Theorem 1** then yields the following theorem.

Theorem 2. *Let*

$$\hat{\Delta}_n = n^{-1} \sum_{i=1}^n \{L(Y_i, \hat{p}'_i) - L(Y_i, \hat{p}''_i)\},$$

$$\delta_i = \{L(1, \hat{p}'_i) - L(0, \hat{p}'_i)\} - \{L(1, \hat{p}''_i) - L(0, \hat{p}''_i)\},$$

$$s_n^2 = n^{-1} \sum_{i=1}^n \delta_i^2 p_i (1 - p_i).$$

Assume that $s_n^2 = O(1)$ with probability 1. Then $\hat{\Delta}_n - \Delta_n$ converges to 0 with probability 1. If furthermore s_n converges in probability to some non-random positive constant, then $\sqrt{n}(\hat{\Delta}_n - \Delta_n)/s_n$ has a limiting standard normal distribution.

Illustrative Application: Precipitation Prob.

- As an application of **Theorem 2**, we compare the Brier scores B_k for the k -day ahead forecasts $\hat{p}_t^{(k)}$, $1 \leq k \leq 7$, for Queens, NY, and Jefferson City, MO, provided by US National Weather Service from June 8, 2007, to March 31, 2009. The values of B_1 and $\Delta(k) \doteq B_k - B_{k-1}$ for $2 \leq k \leq 7$ appear in the following table.
- Using $1/4$ to replace $p_i(1 - p_i)$, we can use **Theorem 2** to construct conservative 95% confidence intervals for

$$\Delta(k) = n^{-1} \left\{ \sum_{t=1}^n (p_t - \hat{p}_t^{(k)})^2 - \sum_{t=1}^n (p_t - \hat{p}_t^{(k-1)})^2 \right\},$$

in which p_t is the actual probability of precipitation on day t . These confidence intervals, which are centered at $B_k - B_{k-1}$, are given in the table below.

	B_1	$\Delta(2)$	$\Delta(3)$	$\Delta(4)$	$\Delta(5)$	$\Delta(6)$	$\Delta(7)$
Queens, NY	.125	.021 $\pm .010$.012 $\pm .011$.020 $\pm .012$.010 $\pm .011$.015 $\pm .011$.007 $\pm .010$
Jefferson City, MO	.159	.005 $\pm .010$	-.005 $\pm .011$.007 $\pm .011$.024 $\pm .010$	-.000 $\pm .010$.008 $\pm .008$

Brier scores B_1 and 95% confidence intervals for $\Delta(k)$.

Illustrative Application: Default Probability

- Multilogit mixed model for loan default (Lai & Sun):
 - A retail loan (e.g., mortgage) has competing risks of default ($r = 1$) and prepayment ($r = 2$); the case $r = 0$ corresponds to the loan still “surviving.”
 - Loans are divided into 5 classes according to obligors’ FICO; Y_{cms} ($= 0, 1, 2$): response of m th loan in class c at age s .
 - Staggered entry: $s = (t - \tau)_+$, where t = calendar time, τ ($= \tau_m$) = origination date of m th loan.

- Thus we can label the n loans at calendar time t by (c, m, s) , where $s = 0$ denotes that the loan has not been originated. Let

$$\eta_{cms}^{(r)} = \log \left(\frac{P\{Y_{cms} = r | Y_{cm,s-1} = 0\}}{P\{Y_{cms} = 0 | Y_{cm,s-1} = 0\}} \right)$$

and $\bar{Y}_{s-1,t}^{(r)}$ the cross-sectional mean of $I_{\{Y_{\cdot\cdot,s-1}=r\}}$ at t .

- Dynamic empirical Bayes model for panel data:

$$\eta_{cms,t}^{(r)} = \rho^{(r)} \log \left(\frac{\bar{Y}_{s-1,t}^{(r)}}{\bar{Y}_{s-1,t}^{(0)}} \right) + a_c^{(r)} + \boldsymbol{\beta}^{(r)T} \mathbf{X}_{cms} + \mathbf{b}_c^{(r)T} \mathbf{Z}_{t-1}$$

- Subject-specific covariates in \mathbf{X}_{cms} include loan size and
 - Purpose (0 for purchase, 1 for refinance)
 - Occupancy (0 for owner occupied, 1 for investment)
 - Documentation (0 for full documentation, 1 for otherwise)
- Macroeconomic covariates in \mathbf{Z}_{t-1} (Calhoun & Deng):

$$\text{Mortgage premium value} = 1 - \frac{\text{1-year maturity Treasury rate}}{\text{mortgage contract rate}}$$

$$\text{Equity position} = \frac{\text{current market value of property}}{\text{unpaid balance}} - 1$$

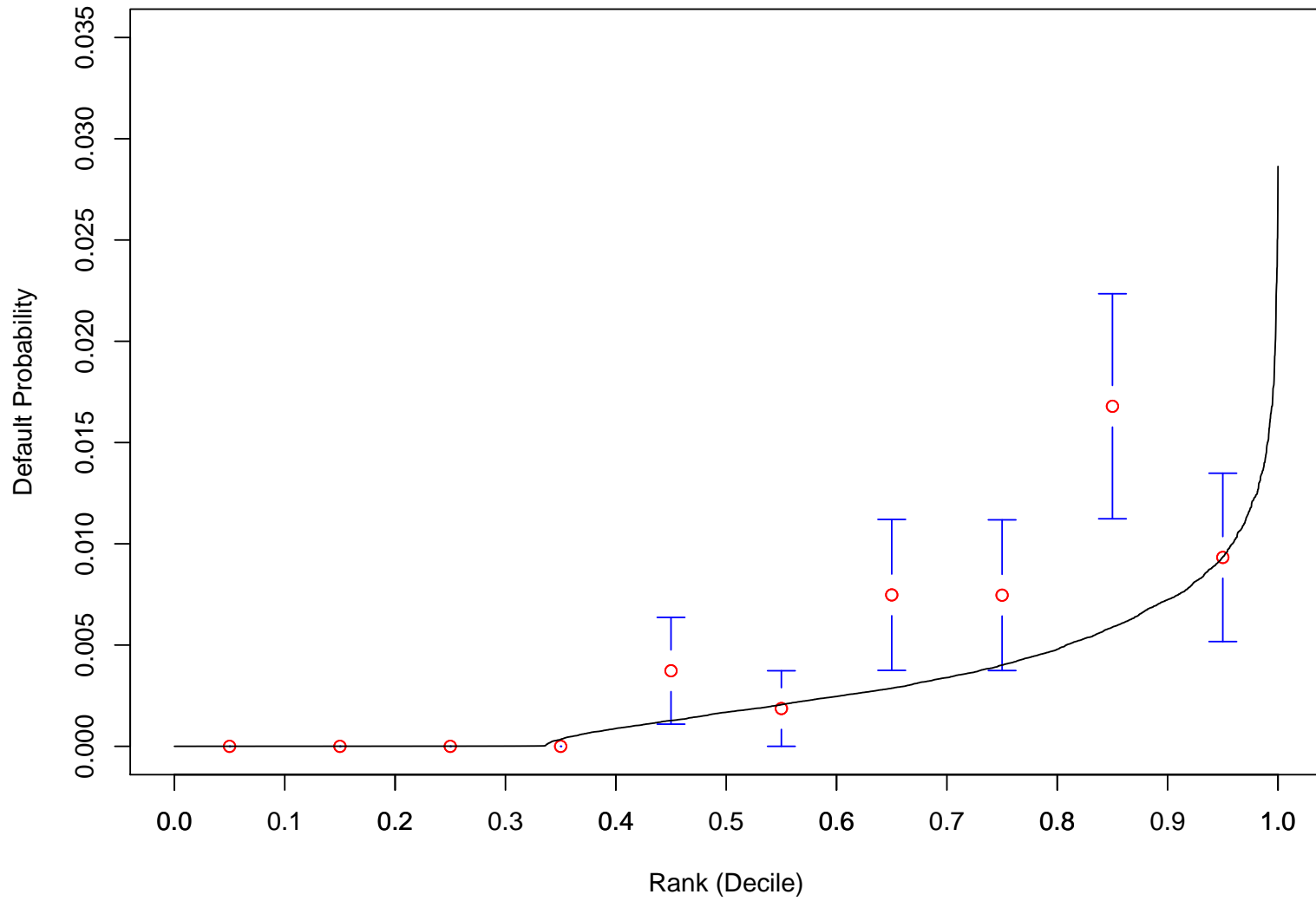
- 10,000 subprime 2-28 ARM loans originated in 2004–06:
 - Prediction of default of loan i in the next month at calendar time t : $\hat{p}_{t,i}$ = predicted default probability.
 - Predictiveness curve: Divide the $\hat{p}_{t,i}$ into deciles D_1^t, \dots, D_{10}^t .
Let

$$I_{j,t} = \{i : \hat{p}_{t,i} \in D_j^t\}, \quad n_{j,t} = \text{card}(I_{j,t}),$$

$$\bar{Y}_t(j) = \frac{\sum_{i \in I_{j,t}} Y_i}{n_{j,t}}, \quad \hat{\sigma}_t(j) = \frac{n_{j,t}}{n_{j,t} - 1} \bar{Y}_t(j) (1 - \bar{Y}_t(j)),$$

$$\bar{Y}_t(j) \pm \sqrt{\hat{\sigma}_t(j)/n_{j,t}}.$$

Predictiveness Curve of Default Probability (t = 20)



- Reliability diagram (from meteorology):
 - Group the $\{\hat{p}_{t,i} : 1 \leq t \leq T, i \leq n_t\}$ into bins (“risk buckets”) B_1, \dots, B_J
 - Letting $I_{j,t} = \{i : \hat{p}_{t,i} \in B_j\}$, define $n_{j,t}$, $\bar{Y}_t(j)$, and $\hat{v}_t(j)$ with B_j in place of D_j^t . Let

$$n_j = \sum_{t=1}^T n_{j,t}, \quad \bar{Y}(j) = \frac{\sum_{t=1}^T \sum_{i \in I_{j,t}} Y_i}{n_j},$$

$$\hat{v}(j) = \frac{\sum_{t=1}^T n_{j,t} \hat{v}_t(j)}{n_j}, \quad \text{se} = \sqrt{\frac{\hat{v}(j)}{n_j}}.$$

	0	(0,.1]	(.1,.2]	(.2,.3]	(.3,.4]	(.4,.5]	(.5,.6]	(.6,.8]	(.8,1]	(1,2]	(2,4.1]
100 \hat{p}	0	(0,.1]	(.1,.2]	(.2,.3]	(.3,.4]	(.4,.5]	(.5,.6]	(.6,.8]	(.8,1]	(1,2]	(2,4.1]
100 (\bar{Y}	0	.03	.20	.34	.56	.51	.73	1.64	1.21	1.57	1.32
±se)	0	.02	.06	.08	.11	.11	.16	.20	.23	.22	.65
Rel. freq.	.26	.12	.11	.11	.09	.07	.05	.07	.04	.06	.01

Tabulated reliability diagram for risk buckets of default probability (%).