Diagnostic, Etiognostic & Prognostic Regression Models in Medicine

Markku Nurminen

Department of Public Health
Faculty of Medicine
University of Helsinki
Finland

Seminar Presentation:

School of Population Health
The University of Western Australia
Perth, 17 November 2009
THE KNOWLEDGE BASE (Conceptual)

In clinical medicine, a doctor’s essential work pertains to three fundamental subtypes of medical **knowing**:

- **diagnosis** – knowing whether a particular illness was/is present;
- **etiology** – knowing whether a particular antecedent was causal to the patient’s illness; and
- **prognosis** – knowing about the future course of the patient’s health in respect to a particular event or state, including how the prospects for this would depend on the choice of intervention or treatment.
MEDICAL ’GNOSIS’

These 3 elements constitute the first-order subtypes of the medical knowing genus, the medical *gnosis*.

Olli S. Miettinen
Prof. of Epidemiology & Biostatistics
McGill University, Montreal, Canada

References:
- Miettinen OS. Up from Clinical Epidemiology and Evidence-Based Medicine. Text 30 June 2009.
CHALLENGE IN MEDICAL STATISTICS: ESTIMATION OF 'INDIVIDUALIZED' RISKS OF ILLNESS SPECIFIC TO

- Location and severity of illness
- Patient characteristics (profiles)
- Treatments or interventions
- Points in follow-up time course
MODERN STATISTICAL METHODOLOGIES

- Trellis Graphics: Scatterplot Matrix
- Generalized Linear Models
  - Smooth-in-Time Logistic Regression
  - Time-Sliced Log-Linear Regression
- Survival Time Analysis: Cox Regression
- Smooth Regression
  - Local Regression
  - Spline Functions
- Tree-Based Methods:
  - Regression Tree
  - Classification Tree
The Occurrence of Annular Tears and their Relation to Lifetime Back Pain History

Tapio Videman, Prof. University of Alberta, Edmonton, Canada

Markku Nurmininen, Doc. Finnish Institute of Occupational Health, Helsinki, Finland

Vol. 29 2004
CAUSAL PATH FOR SPINAL PATHOLOGY

Heavy physical work load in adolescence → Lumbar disc degeneration & annular tears → Frequent lower back pain

Risk factor profile → Manifestation profile → Symptom profile

← Etiognosis
← Diagnosis

Prognosis →
Annular BaSO₄ Discograph

- L1/L2  II Outer
- L2/L3  I  Inner
- L3/L4  0  None
- L4/L5  II Outer
- L5/S1  III Leaking
WHAT CAN STATISTICS AT BEST OFFER EPIDEMIOLOGISTS?

- Can detect difficult-to-distinguish interactions and subtleties in highly complex and voluminous data sets.

- Can supply informative, compact data summaries and graphics (undervalued functions in medicine in lieu of P values and confidences).
The **probability of annular tear** in the discs of the lumbar spine by age and the level of intervertebral disc, with a fitted nonparametric polynomial regression curve using a locally weighted scatterplot smoother (statistical S system function LOWESS). The data points are based on 785 discs.
RISK FUNCTION MODELING
APPROACHES

1. **Fully-parametric** regression; any generalized linear hazard function, e.g. log-linear and logistic regression

2. **Semi-parametric** (Cox) regression with proportional hazard functions

3. **Non-parametric** regression, such as local regression, spline functions, classification and regression trees
Diagnostic Probability Model

For diagnosis, an appropriate, fully parametric form of the model for the presence of an illness is the \textit{logistic} one:

$$
\log \left[ \frac{P}{1-P} \right] = B_0 + \sum_i B_i X_i ,
$$

where $P$ is the diagnostic probability; the $X_i$ are statistical variates representing the diagnostic indicators; the $B_i$ are the model parameters that constitute the object of diagnostic-relevant knowledge regarding the illness.
Etiognostic Probability Model

For *etiognosis*, the model is properly given by a 'generalized linear' form; i.e., a linear function of the *rate* (*R*):

\[ f(R) = B_0 + \sum_i B_i X_i = L \]

Specifically, for an incidence density that uses a logarithmic transformation \( f(R) = \log(R) \), the *causal rate ratio* (\( RR \)) is:

\[ RR = \frac{R_1}{R_0} = \exp(L_1 - L_0) \]

where \( L_0 \) is evaluated at the referent category of the *etiogenetic* determinant of the rate that represents the alternative to the causal one \( L_1 \).
Prognostic Probability Model

For *prognosis* about an event, the model for incidence density (ID) is tractably of the *log-linear* form

\[ -\log(\text{ID}) = L; \text{ID} = \exp (L) \]

Thus the cumulative incidence (CI), of the event’s occurrence in the prognostic time interval from \( t_1 \) to \( t_2 \), is as follows:

\[ \text{CI} = 1 - \exp \left[ - \int_{t_1}^{t_2} \text{ID}_t \, dt \right]. \]


Let the follow-up of patients’ illness outcome, $y$, constitute the study base of B person-years.

Any given person-moment in the study base is characterized by the point in time $t$; and the person's value $x$ for the variates based on prognostic indicators and intervention.

The dataset consists of the information $(y; x; t)$ for the case series of size $c$ and for a (random or) representative sample of size $b$ in the study base.

The incidence density or hazard function is $h(x; t)$.
The Cox proportional hazards model (1972)

The semi-parametric hazard function is of the form:

\[ h(x; t) = h_0(t) \exp (\beta' x) \]

In this famous and influential model, \( h_0(t) \) is a baseline hazard function that is modified by covariates \( X \) and the interest is in the proportional factors.

The \( \beta \) parameter vector is estimated by maximizing a partial likelihood.
The Hanley-Miettinen 'individualized' risk function model (2009)

The proposed smooth-in-time hazard function is:

\[ h(x; t) = \left( \frac{b}{B} \right) \exp[L(x; t)] \]

where \( L \) is the linear compound from the fitting of the logistic regression model. One can obtain \( \log[h(x; t)] \) directly by fitting a logistic regression model to the \((c+b)\) data vectors \([y; x; t; \log(B=b)]\) and specifying \( \log(B=b) \) as an 'offset'.
Figure 3: Estimated cumulative incidence (risk) of stroke for patients with higher-risk (a.0 if untreated, a.1 if treated) and lower-risk (b.0 and b.1) profiles, fitted by the proposed fully-parametric approach, and by the semi-parametric Cox regression. Data are from the SHEP (1991).
Despite their wide availability, the patient’s profile-specific estimates of risk available from Cox's semi-parametric model are seldom used.

It may be that end-users are averse to the fitted risk function's 'steps-in-time' — and thus 'raw' and 'unsophisticated-looking' — form.

In the Cox model, the baseline hazard is treated as a high-dimensional nuisance parameter, and thus it has been estimated informatively, i.e., smoothly by cubic spline functions.
The Hanley-Miettinen method is a very different way of fitting the hazard function. It is based on bringing to the context of survival analysis the data structure of an etiologic study in epidemiology (Mantel 1973): 

a case series coupled with

a base series (sample of study population-time) together with logistic regression analysis of these data.

The Hanley-Miettinen approach provides for estimation of the hazard function per se, and thereby estimation of cumulative incidence and risk.
The Nurminen likelihood score-based rate ratio function model (1992)

The piecewise exponential failure time, $T$, model is:

$$R_{1j}(x; t) = R_{0j} \exp(\alpha + \beta x_j)$$

The time stratum ($j$) -specific rate ratios $RR_j = R_{1j}/R_{0j}$ are estimated under a Poisson regression model for the case series $c_{ij} \sim \text{Poisson}(R_{ij}, T_{ib}b_{ij}/b_j)$, $i=1,0$.

ML estimates were obtained by the iteratively re-weighted least squares fitting algorithm using GLIM.

A TIME-SLICING APPROACH USED TO FIT THE SMOOTH MODEL

The parametric, smooth-in-time, log-linear form that Hanley & Miettinen adopted for the hazard function is certainly not new. It is a natural extension of the form proposed by Gompertz (1825) for human mortality.

But, with more complicated terms in time $t$, such as a treatment and time product term or any other time-varying covariate, computing difficulties appear to have prevented this very natural form from being used in a broader regression framework.
Tree-Based Models

A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. The construction of medical decision-making trees may be seen as a type of variable selection.

In a regression tree each terminal node gives a predicted value for the probability of illness. This serves as the basis for:

A classification tree that has as its endpoint a factor giving the subtypes of patients or illnesses.

Issue of interaction is handled automatically, so is monotonic transformation of the X variates.
Diagnosis of epithelial mesothelioma using tree-based regression analysis & a minimal panel of antibodies

Sonja Klebe, Markku Nurminnen, James Leigh, Douglas W. Henderson

Department of Anatomical Pathology, Flinders Medical Centre, Adelaide, South Australia

Finnish Institute of Occupational Health, Helsinki & Department of Public Health, University of Helsinki, Finland

Centre for Occupational and Environmental Health, School of Public Health, University of Sydney, NSW

01 Feb 2009
Aims

- **Immuno-histo-chemistry** with panels of antibodies is a standard procedure to distinguish between malignant mesothelioma (MM) and metastatic adenocarcinoma (AC).

- Most studies assess only the **sensitivity** and **specificity** for single antibodies, even when an antibody panel is recommended.

- This study used a novel statistical approach – tree-based modeling – to identify a **minimal set of markers**, which would reliably make this distinction in the majority of cases, and would correlate with a diagnosis of MM.
**Cases and Antibodies**

**200 consecutive cases** of pleural malignancy:

173 pleural mesotheliomas of epithelial type and 27 cases of secondary adenocarcinoma were investigated using a standard panel of

**12 antibodies**, which included mesothelial-related antibodies and carcinoma-related antibodies, and 2 general epithelial antibodies:

CAM5.2, CK5/6, calretinin, thrombomodulin, HBME-1, WT-1, EMA, CEA, CD15, B72.3, BG8, and TTF-1
Previous Approaches

Meta-analysis: due to heterogeneity in the raw data in the individual studies, the validity of this approach is limited.

Sensitivities and specificities of antibodies (ROC analysis): old-fashioned method which does not yield a clear answer.

Stepwise (backwards variate reduction) logistic regression: the model includes all the markers simultaneously; interactions between markers is difficult to interpret.
Adopted Study Approach

- Multivariate statistical analysis — tree-based modeling, which involves observation and analysis of more than one variable at a time while taking simultaneously into account the effects of all the diagnostic variates on the endpoint of interest.

- This automatic construction of decision trees provides a modern way to encapsulate and structure the knowledge of experts to be used by less-experienced practitioners.
Applied Statistical Methods

Regression and classification tree-based methods were applied to select the best, minimal-sized combination of markers.

The modeling procedures employ successive, hierarchical predictions computed for individual cases to sort them into homogeneous classes.
Regression Tree Model

Probability
Number of Meso. cases
Marker value
Split node (Oval)
Terminal node [Rectangular]
Tree Construction

The tree is constructed in phases. First an overly complex tree was fitted, and then 'prune' the tree was 'pruned' down to a suitable size.

If the sample numbers are sufficiently large, the study yields unbiased estimates of the disease classification probabilities and misclassification rates.

Any split which did not improve the model fit by a default factor of the complexity parameter was pruned off by cross-validations employed to ensure the numerical stability of the terminal nodes or *leaves* of the tree.
Classification Tree Model

Ac/ Ac+MM

Probability of MM
Discussion Issues

- The action taken for handling missing values
- Stability of the statistical model
- Problem of overfitting predictive models
- Comparison to other statistical models:
  - Tree-based modeling is hierarchical, with successive predictions being applied to cases.
  - Traditional methods use simultaneous techniques to make one and only one prediction for each and every case.
Conclusions

A panel of three antibodies is sufficient in most cases to diagnose, or to exclude, epithelial mesothelioma.

Calretinin exhibits the strongest correlative power of the antibodies tested.

The tree-based approach deserves wider application for the immunohistochemical diagnosis of a variety of tumors.
Prognostic Models for Predicting Delayed Onset of Renal Allograft Function

Markku Nurminen
MA (Math.), PhD (Stat), DrPH(Epid.)
Finnish Institute of Occupational Health, and University of Helsinki

The Internet Journal of Epidemiology
2003: Volume 1 Number 1 Pages 1-20
Abstract

- Prognostic models were developed and evaluated to determine important risk factors predicting *delayed graft function* as well as factors that prolong the delay.

- The data consisted of 1,215 patients transplanted with *renal allografts* at a single center in Finland in 1986-1995.

- Most important predictors related to delayed graft function were cold ischemia time, type of dialysis, and time in dialysis.
Graft Function Delay vs. Cold Ischemia Time

- Probability of delayed graft function
- Cold ischemia time, h

Graph showing the relationship between graft function delay and cold ischemia time.
Cox Regression Analysis

Probability of onset of graft function since time of transplantation for subgroups by duration of dialysis.
Regression Tree Model

Y-var: Time, days

Split (Oval) Node

Terminal [Rectangular] Node
Probability of delayed renal allograft function

Number of patients
Closing Remarks

- In epidemiology and in other prognostic research contexts involving complex data with many potential risk factors, the tree-based methods fall short of outright excelling consistently those used commonly.

- Yet it is premature to say if tree-based analyses will be accepted as the preferred methods. Nevertheless, they have the potential power of providing complementary information and contributing to the insightful interpretation of prognostication.
Thank you; tomorrow I’ll fly away!
Meet Me at MarkStat.net