SUCCESS AT STATISTICAL RESEARCH: SOME THINGS YOU SHOULD KNOW

Jerry Lawless
University of Waterloo
jlawless@uwaterloo.ca

Talk for Southern Ontario Statistical Graduate Student Seminar Days
(SOSGSSD)
McMaster University
May 11-12, 2006

OUTLINE

General Remarks About Research
Statistical Activities and Research in Statistics
Some Tools You Should Have or Acquire
Some Broad Guidelines
Some Personal Illustrations
Conclusion

General Remarks About Research

- Research is about asking (the right) questions.
- So, to begin:

  (a) Is ability at research learned, or innate?

  (b) Can research skills be taught and learned?

  (c) Assuming the answer to (b) is yes, what are some of the key tools that a researcher should acquire?

- Learning how to ask questions is important, as well as developing sufficient technical skill to formulate "solutions".
- Enthusiasm for one's research topics is important, especially since good research is hard to do.

- Breadth of exposure seems helpful in promoting "lateral" thinking and creativity, as well as helping us to develop an overview.
  - Attending seminars, conference talks
  - Talking with colleagues, scientists, others
  - Surveys and relatively non-technical papers
- For research in the statistical sciences: exposure to real problems in science, business and industry, government etc.
- Communication of your research results is crucial; this requires speaking, writing and presentation skills.
- But remember: researchers have different backgrounds, strengths, interests and styles, so my viewpoint here is necessarily personal.

Statistical Activities and Research in Statistics

- You need to learn to ask interesting questions (which you can then try to address or answer). Some guidelines based on "real problems" are
  - (a) look at the problems and ask what the needs or objectives are, in general language
  - (b) ask what resources are available (existing data, collection of new data, auxiliary information etc.)
  - (c) think about ways to address the objectives with statistical methods
  - (d) ask what are the strengths and weaknesses of the available data, and of different statistical methods, for the topic in question
- Similarly, when reading papers, reports etc, or listening to talks: ask what the objectives are, and whether they are really being addressed
- Technical (mathematical, computational) problems: go through a similar process; be critical.

Types of statistical activity include

- Design of studies and schemes for collecting data
- Exploration, description and summarization of data
- Statistical inference (study design, estimation, testing) regarding parameters or models
  - The parameters should be relevant and interpretable in terms of scientific or other objectives
- Classification and other decision procedures
  - On the basis of available data, make a decision.
- Statistical (Stochastic) modelling of random processes
- Prediction

Some Tools You Should Have or Acquire

The big picture: general concepts that a statistical scientist should understand and keep in mind.

- the nature, sources and effects of variation in populations and processes
- stochastic models, their types and roles
- the nature of inductive inference and of causation
- types of studies, measurement, observation and data collection
- sources of uncertainty and the roles of probability in inference
- frameworks for statistical decision-making
- methods of computation

Technical skills: acquire as much as you can

Some Technical Background and Tools

STATISTICAL MODELS

- Probability distributions $F(y)$, $F(y|x)$ for random variables representing responses Y and covariates X in some population or process
- Stochastic processes $\{Y(t), X(t) : t \epsilon T\}$
- Be familiar with the structure and properties of as many models as you can.
  e.g. Multivariate distributions, point processes, diffusion processes, counting processes, models with unobservable random effects.
- Good books (mostly older): CR Rao, Wilks, Feller, Karlin and Taylor, Cox and Isham ...
- Simulation: a good way to develop and extend understanding is to consider how to simulate data from a given model. (Also crucial in many applications.)

ESTIMATION AND HYPOTHESIS TESTING

- Likelihood and Bayesian methods for parametric models
- Estimating function methods (e.g. pseudo likelihood, M-estimation, GEE's)
- Nonparametric and semi-parametric methods
- Roles of conditioning and marginalization with certain models: if θ is the vector of parameters and if the potential data D is partitioned somehow as (D_1, D_2), then we often base inference on a conditional distribution $Pr(D_1 = d_1 | D_2 = d_2; \theta)$ or a marginal distribution $Pr(D_1 = d_1; \theta)$.
  - Useful for removing "nuisance" parameters.
  - Need to understand the "geometry" of models to see what can be done.
- Identifiability, estimability and information concerning parameters (how to quantify or assess?)
- Model assessment methods.

PREDICTION, DECISION-MAKING, CLASSIFICATION

- Basic framework: make a prediction or decision concerning a "future" random variable Y, on the basis of available data X.
- Measures of performance for prediction or decision procedures.

COMPUTATION

- Understanding of optimization and simulation tools.
- Structure of algorithms. (Many branches of statistics are increasingly algorithmic.)

Some Broad Guidelines

- Expose yourself to "real" problems and methodology.
- Read journal articles on topics you find interesting; they don't have to be highly technical. Go to lots of talks and talk with others. You can zero in on technical background as needed, once a topic is identified.
- Ask questions but make them relevant by thinking hard about the objectives of any scientific or statistical problem.
- Remember that "all models are wrong but some are useful". Try to focus on questions and methods that are useful.
- Look for "gaps" in methodology. Tackle problems or topics that are not overworked by others. Think laterally.
- Papers by "top" people in general methodology or applied areas often identify or suggest gaps in theory or methods.

Some Personal Illustrations

LINEAR MODELS: RIDGE REGRESSION AND SHRINKAGE

- Consider $Y_i = \beta_0 + \beta' x_i + e_i \qquad$ i=1,. . . , n
  with x_i a $p \times 1$ vector and centred so that $\bar{x} = 0$, and $e_i \sim N(0, \sigma^2)$
- Least squares (normal ML) estimator of β is $(X'X)^{-1}X'Y$, where Y is $n \times 1$ and X is $n \times p$.
- Ridge estimators: $\hat{\beta}_\lambda = (X'X + \lambda I_p)^{-1}X'Y$
  - useful when $X'X$ is badly conditioned (max eigenvalue/ min eigenvalue is large), i.e. when near multicollinearity exists.
- Other shrinkage estimators: $\hat{\beta}_\lambda$ closer to 0 then LS estimator $\hat{\beta}_0$, e.g. James-Stein.
  Lawless (Commun. Statist. 1978, JASA 1981)

- Studies and discussion around 1970-74: comparison of estimators using squared error loss function $(\hat{\beta}-\beta)'(\hat{\beta}-\beta) = \sum_{j=1}^{p}(\hat{\beta}_j-\beta_j)^2 = L_1(\hat{\beta},\beta)$
  - Big potential gains in mean square error $E\{(\hat{\beta}-\beta)'(\hat{\beta}-\beta)\}$ for shrinkage estimators.
- But when X is multicollinear ($X'X$ near-singular) then $L_1(\hat{\beta},\beta)$ is not very appropriate, because
  - interpretation of individual β_j's is tenuous (depends on what other variables are in the model)
  - concept of a "true" β is not very relevant
- So focus instead on prediction: look at $X\hat{\beta} - X\beta$, and consider

$$L_2(\hat{\beta},\beta) = (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta) = (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)$$

  - Gives a much different picture than $L_1(\hat{\beta},\beta)$.
- Bayes: prior information is really about $X\beta$, not β

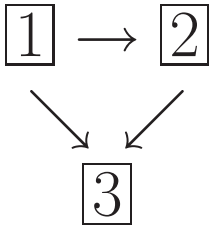
GENERALIZED LINEAR MODELS: ALL-SUBSETS REGRESSION

- About 1974-75: Fast algorithms for fitting all 2^p possible linear regression models involving p covariates $x_1, \ldots, x_p$.
  - Idea is to run through the models in a sequence where each new model has only one covariate changed from the last model.
    e.g. $p = 3 : (-) \to (1) \to (12) \to (2) \to (23) \to (3) \to (13) \to (123)$
  - Simple updates for the sequence of least squares estimates, plus use of "bounding " to reduce models that have to be searched
    (SAS: selection of "best" k models of each size)
- Question: how to do this for generalized linear models?
  - Nonlinear ML equations for $\hat{\beta}$, but asymptotically linear
  - Can define a computationally efficient procedure

Lawless and Singhal (1978 Biometrics)
SAS Enterprise Miner: can find top k models (e.g. $k = 10$) in terms of AIC for p as big as 40.

MULTISTATE MODELS

- About 1980, lots of discussion of multistate models in probability and in papers on modelling, but not much methodology for estimation, inference.
  1981- Kalbfleisch and Lawless conference survey paper
- Individual data: follow individuals over time and see what states they are in at different times
  e.g. Illness - death model: 1 - Healthy, 2 - Sick, 3 - Dead



- Aggregate data: only observe the total numbers of individuals in each state at different times (no tracking of individuals)

Insect Life Cycles (Kalbfleisch, Lawless and Vollmer, 1983 Biometrics)

$$\boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{3} \rightarrow \boxed{4} \rightarrow \boxed{5} \rightarrow \boxed{6}$$

State 1: egg stage State 6: adult insect
States 2-5: developmental stages

- Field data (Steve Smith, UW Biology): every 2-3 days, count the total number of individuals in each stage on a given habitat (tree)

  - aggregate data: $y(t) = (y_1(t), \ldots, y_6(t))$ for $t = 2, 5, 7$ etc.

- How to model development process (amount of time spent in various states) and how to fit models to the data?

  - early use of longitudinal estimating functions (marginal, conditional)

  Later (Lawless and McLeish, 1984 Biometrika): amount of information in aggregate data, relative to individual longitudinal data?

  - implications for study design

INFORMATION IN SELECTIVE SAMPLES: AIDS INCUBATION TIMES

- In many observational studies the individuals selected have to satisfy certain conditions.
- Transfusion - associated HIV infections (Kalbfleisch and Lawless, JASA 1989) early in the HIV-AIDS epidemic
  T = time from HIV infection to AIDS
  - Data from persons who were infected by blood transfusions and who got AIDS by 1986; for each person the time of HIV infection x_i and time of AIDS $x_i + T_i$ can be determined. However, T_i must satisfy $T_i \leq R_i$, where R_i = December 31, 1986 - x_i.
  - Estimation of $F(t) = Pr(T \leq t)$ is based on $Pr(T_i | T_i \leq R_i)$.
  - Parametric Weibull models: .95 confidence interval for 63rd percentile of T is about (.7 months, 73000 months)!
- So - need to incorporate auxiliary data

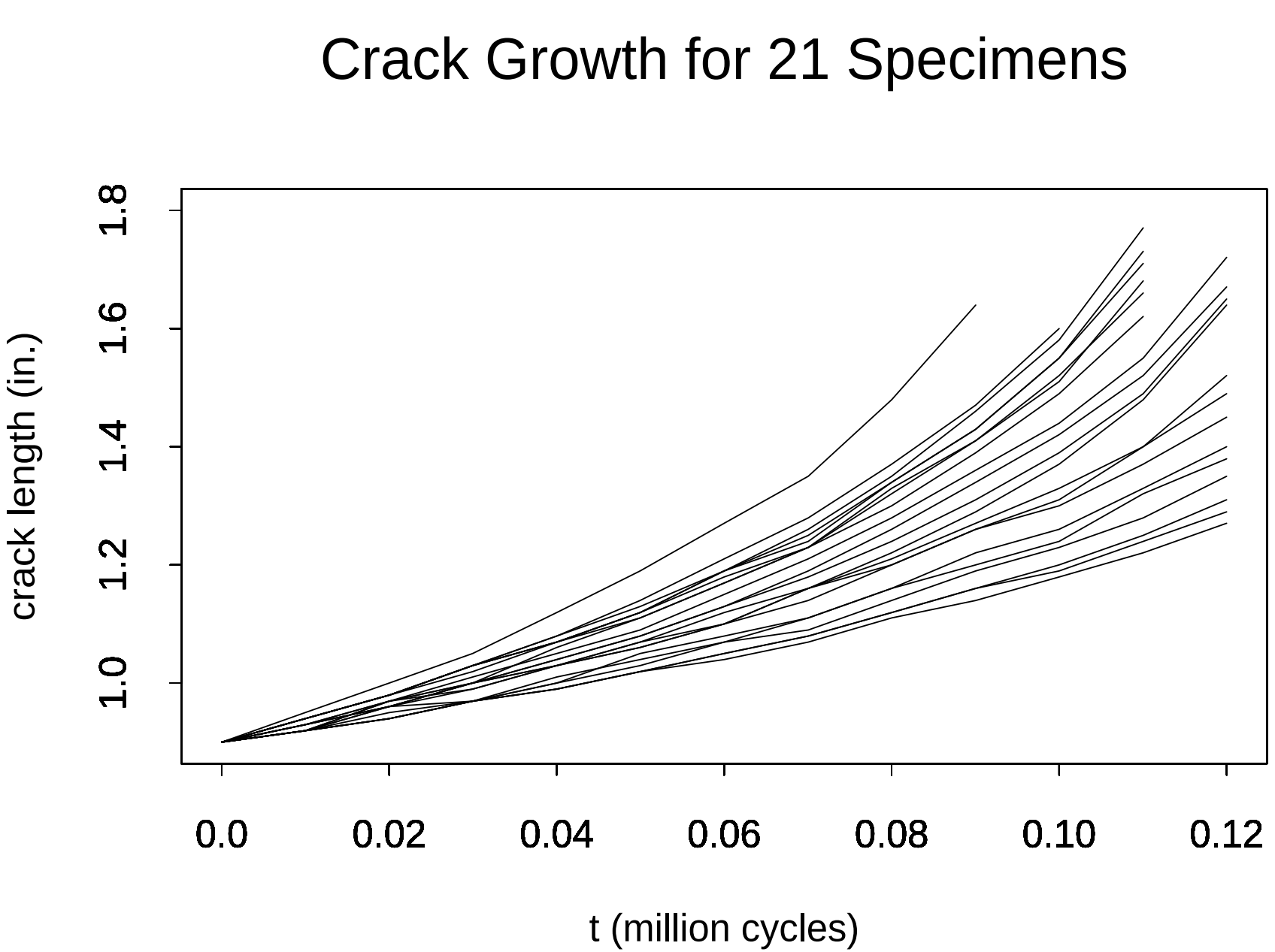
- Similar situations occur in many other settings (e.g. case-control studies)
  - How to incorporate auxiliary data?
  - Best type of auxiliary data (re added information)?
  - In general: combination of data from different sources.

- Hu and Lawless (JASA 1996): failure rates from car warranty data
  - missing information on cars with no warranty claims

STOCHASTIC PROCESSES WITH RANDOM EFFECTS: DEGRADATION AND FAILURE

- In many settings involving a failure time T, there is a related process $\{X(t), t > 0\}$ that gives a measure of the deterioration or degradation of an individual.

- Example - crack growth in metal components: $X(t)$ = diameter of crack after time t (see plot)


Crack Growth for 21 Specimens
crack length (in.)
1.0
1.2
1.4
1.6
1.8
0.0
0.02
0.04
0.06
0.08
0.10
0.12
t (million cycles)

So: Need individual-level random effects in order to represent the large degree of unit-to-unit variability in the sample paths

- stochastic processes with random effects
(Lawless and Crowder, 2004 Lifetime Data Analysis)

- Earlier work: Poisson processes with random effects (Lawless, 1987 JASA)

  - tractable models plus methods of estimation

- Recent problem: treatment (anti-retroviral therapy) for persons with HIV.

  - Monitor, give treatment based on {CD4 (t), ViralLoad (t)}, where t = time.

CONCLUSION: A FEW GUIDELINES, REITERATED

- Expose yourself to a wide variety of topics and problems, time permitting.
- Acquire at least a basic overview of a broad range of tools (stochastic modelling, modes of design, inference and decision, computational tools).
- Think critically about the objectives in specific problems, and how well proposed or existing methods address them.
- Find topics that you are enthusiastic about pursuing, whether in methodology, theory, or a specific field of application.
- Work just as hard on seeing "gaps" in methodology or theory as in trying to generalize or extend existing work.