171:290 Model Selection
Lecture I: Introductory Principles, Concepts, and Procedures

Joseph E. Cavanaugh

Department of Biostatistics
Department of Statistics and Actuarial Science
The University of Iowa

August 21, 2012

Introduction

- What is a *statistical model*?

- The Oxford Dictionary defines a **model** as *a simplified or idealized description of a particular system, situation, or process, often in mathematical terms, that is put forth as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.*

- In a statistical application, the "particular system, situation or process" is some random phenomenon of interest.

- The phenomenon cannot be described or predicted with complete accuracy.

- However, we believe that the phenomenon consists of a *deterministic* component which *is* predictable, and that this component can ultimately be characterized and represented.

Introduction

- One of the primary goals of statistical modeling is to obtain a quantitative description of the deterministic component.

- This description should be realistic, yet also interpretable and concise.

- Many random phenomena, such as those arising in biological and ecological applications, are extremely complex, potentially involving an endless assortment of variables and interactions.

- Any interpretable, concise description of a complex phenomenon will invariably be simplified or idealized.

Introduction

- The non-deterministic component of a random phenomenon, the *stochastic component*, represents aspects of the phenomenon that are not predictable.

- In the formulation of a statistical model, one usually attempts to specify a probabilistic mechanism that provides an adequate reflection of how the stochastic component behaves.

- This aspect of statistical modeling implies that the model description is cast in probabilistic terms.

Introduction

- Taking the preceding considerations into account, we might propose the following modification of the Oxford definition to define a statistical model: "a simplified or idealized description of a random phenomenon, generally in probabilistic terms, that is put forth as a basis for theoretical or empirical understanding, or for calculations, predictions, etc."

Introduction

- What properties are associated with a good statistical model?
- An optimal statistical model is characterized by three fundamental attributes.
  - Parsimony (model simplicity).
  - Goodness-of-fit (conformity of the fitted model to the data at hand).
  - Generalizability (applicability of the fitted model to describe or predict new data).
- What statistical procedures can guide us towards a model that possesses these properties?

Introduction

Outline:

- Philosophical Basis for Model Selection: Occam's Razor
- Bias and Variability
- Statistical Procedures for Model Selection
  - Hypothesis Testing
  - Automatic Variable Selection Algorithms
  - Model Selection Criteria

Occam's Razor

- Occam's Razor is a philosophical principle credited to the medieval English philosopher and Franciscan monk William of Ockham (1285–1349).
- Quote of William of Ockham: "Plurality should not be posited without necessity."
- Occam's Razor recommends that we "shave off" extraneous ideas to better reveal the truth.

Occam's Razor

- The principle of Occam's Razor is reflected in the writings of many renowned philosophers and scientists throughout history.

- "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices." – Thomas Aquinas (1225 – 1274).

- "We are to admit no more causes of natural things than such are both true and sufficient to explain their appearances." – Isaac Newton (1643 – 1727).

- "Everything should be made as simple as possible, but not simpler." – Albert Einstein (1879–1955).

Occam's Razor

- "When you hear hoofbeats, think horses, not zebras." –
  popular adage from medical schools and residency programs.
- Classic illustration of Occam's Razor: The story of Clever
  Hans, the German horse who could do arithmetic!

Occam's Razor

Clever Hans

Occam's Razor

- Translated to statistical modeling, Occam's Razor is sometimes referred to as the *Law of Parsimony*.

- **Law of Parsimony:** No more causes should be assumed than those that will account for the effect.

- Selecting a model that complies with the Law of Parsimony amounts to choosing the simplest model in the candidate collection that adequately accommodates the data.

- From a statistical perspective, what is the advantage of adherence to the Law of Parsimony?
  - Is the main advantage based on interpretability; i.e., a simple model is more easily understood and explained than a complex one?
  - Are any advantages based on inferential objectives?

Occam's Razor

- A model is **underfit** if its structure is too simplistic: i.e., key explanatory variables or effects are excluded, the functional form of the model is too rudimentary.

- A model is **overfit** if its structure is unnecessarily complex: i.e., extraneous explanatory variables or effects are included, the functional form of the model is too complicated.

- An important concept in statistical modeling:

    underfitting induces bias

  whereas

    overfitting induces increased variability.

Occam's Razor

- Thus, the statistical advantage of adherence to Occam's Razor is an improvement in the accuracy of inferential results: e.g., estimators of parameters, predictors of response variables.

- This improvement results from controlling the variability associated with overfitting while protecting against the bias associated with underfitting.
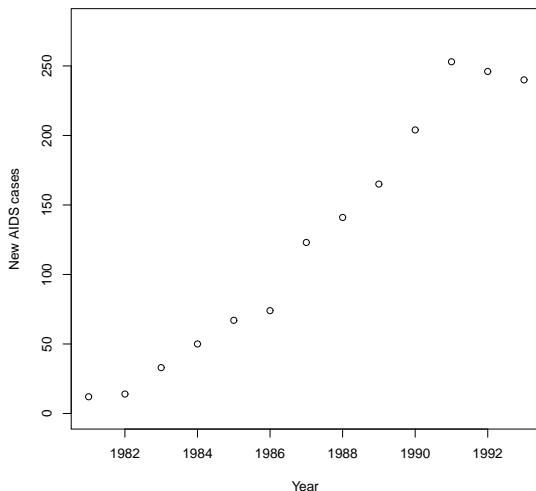
Bias and Variability

Overfitting / Underfitting Example

Incidence of AIDS in Belgium, 1981-1993

- $x =$ year (1981 - 1993)
- $y =$ number of new AIDS cases reported in Belgium in year x

Figure: Plot of new Belgium AIDS cases versus the year (1981-1993).

Bias and Variability

Models:

$$y \sim \text{Poisson}(\lambda)$$
$$\log \lambda = \text{polynomial of degree } d \text{ in year index}$$

We fit four models to the data based on various polynomial degrees: $d = 1$ (linear), $d = 2$ (quadratic), $d = 3$ (cubic), $d = 12$.

Figure: Fit of polynomial of degree 1 (linear). The model is simple, yet may exhibit some lack-of-fit due to bias.
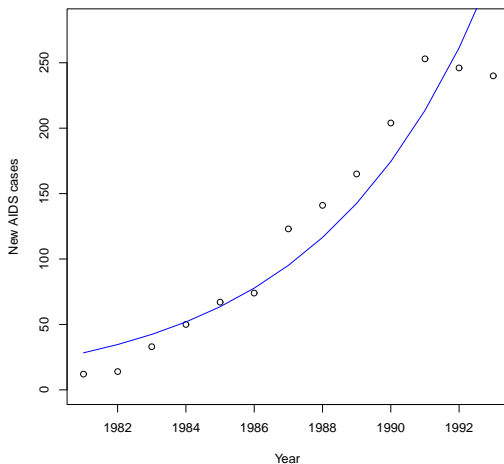
Figure: Fit of polynomial of degree 2 (quadratic). The model conforms well to the data; is not overly complex.
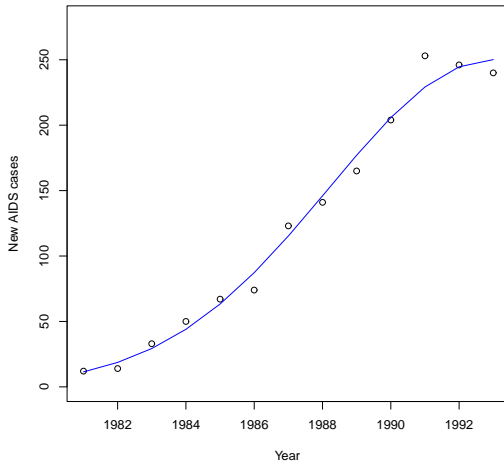
Figure: Fit of polynomial of degree 3 (cubic). The model conforms well to the data, but does the added complexity improve the fit?
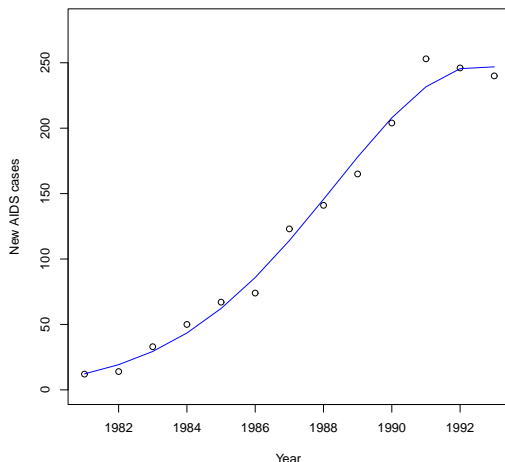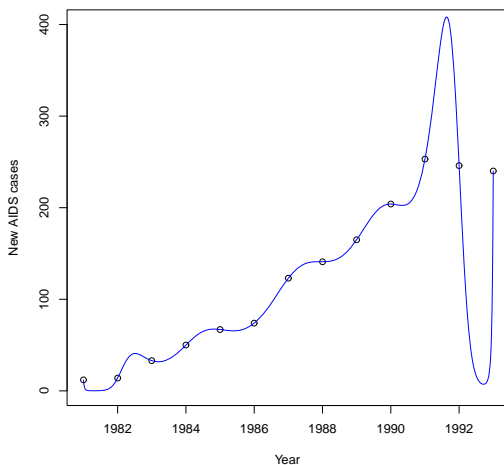
Figure: Fit of polynomial of degree 12. The model conforms perfectly to the data, yet is overly complex, unstable, and plagued by variability.

Model Selection Procedures

- What statistical procedures are available to search for a fitted model that complies with Occam's Razor?
  - Hypothesis Testing
  - Automatic Variable Selection Algorithms
  - Model Selection Criteria

Hypothesis Testing

- In practice, hypothesis testing is often used for model selection.
- For variable determination in multivariable models, a two-step procedure is often employed.
  - Single variable models are fit to the response. Those covariates that exhibit statistical significance are included in the fit of a multivariable model.
  - In the multivariable model, backwards elimination is employed by systematically eliminating covariates with large p-values.
- Does hypothesis testing provide an appropriate paradigm for selecting a statistical model?

Hypothesis Testing

- Classical hypothesis testing procedures are generally based on the assumption of nested models, with the larger model (represented under the alternative hypothesis) being initially regarded as correct.

- In many statistical modeling applications, especially those in the biomedical and health sciences, the notion of *any* model being correct is difficult to defend.

- "All models are wrong, some are useful." – George Box.

Hypothesis Testing

- In hypothesis testing, the interpretation of p-values, power, and statistical significance relies upon the notion of one or the other hypothesis representing truth.

- Akaike (1974): "In the Neyman–Pearson theory of statistical hypothesis testing only the probabilities of rejecting and accepting the correct and incorrect hypotheses, respectively, are considered to define the loss caused by the decision. In practical situations the assumed null hypotheses are only approximations and they are almost always different from reality. Thus the choice of the loss function in test theory makes its practical application logically contradictory." (See also Rao and Yu, 2001.)

Hypothesis Testing

Hypothesis Testing for Model Selection

- Suppose $\mathcal{M}_S$ and $\mathcal{M}_L$ are candidate models, where $\mathcal{M}_S$ is nested within $\mathcal{M}_L$.

- According to the Box philosophy, $\mathcal{M}_S$ and $\mathcal{M}_L$ are both *wrong*.

- Our hypotheses are as follows:

$$H_0 : \quad \mathcal{M}_S \text{ is the correct model}$$
$$H_A : \quad \mathcal{M}_L \text{ is the correct model}$$

- We use the data to determine which hypothesis represents truth.

- However, if we follow the Box philosophy, we do not believe that either $\mathcal{M}_S$ or $\mathcal{M}_L$ is correct!

Hypothesis Testing

- How do we interpret the power of the test, which is a probability calculated under the assumption that model $\mathcal{M}_L$ is correct?

- How do we interpret the p-value for the test, which is a probability calculated under the assumption that model $\mathcal{M}_S$ is correct?

- If we obtain a large p-value, why do we then treat $\mathcal{M}_S$ as the correct model in any subsequent test?

- Akaike (1973, 1974) argued that model selection should not be viewed as a *testing* problem, but rather as a *decision* problem: given a collection of fitted models, each of which represents an *approximation* to reality, which approximation is best?

Automatic Variable Selection Algorithms

- Automatic variable selection algorithms have been extensively used for model selection, especially in applications where a large number of potential explanatory variables must be considered.

- The most popular procedures are backwards elimination and forward selection.

Backwards Elimination

The backwards elimination algorithm:

- Fit the model containing all of the explanatory variables of interest.
- Based on the size of the partial-test p-values, systematically remove explanatory variables from the model until all remaining variables have p-values beneath a pre-defined threshold.

Forward Selection

The forward selection algorithm:

- For each explanatory variable, fit a model containing only this variable. Choose a single-variable model based on the variable with the minimum p-value.

- Based on the size of the partial-test p-values, systematically add explanatory variables to the model. Stop when the p-value associated with any remaining variable inclusion exceeds a pre-defined threshold.

Automatic Variable Selection Algorithms

- The main advantages of automatic variable selection
  algorithms are that they are simple to apply and they are
  computationally efficient.

Automatic Variable Selection Algorithms

- The main disadvantage of such algorithms is that they exclude consideration of many candidate models based on many different possible subsets of explanatory variables, and may lead one to a final fitted model based on an inferior subset.
    - Example: Suppose that 10 candidate variables are considered for inclusion in a multivariable model.
        - In this setting, there are $2^{10} = 1028$ possible candidate models.
        - Backwards elimination or forward selection will consider at most $10 + 9 + 8 + \ldots + 1 = 55$ of these models.

- A second disadvantage is that the steps are generally based on hypothesis testing.

- A third disadvantage is that an automatic algorithm cannot take into account scientific or clinical considerations that may be important from a practical perspective.

Model Selection Criteria

- A model selection criterion is a measure that assesses the propriety of a fitted model by gauging how well the model balances the competing objectives of conformity to the data and parsimony.

- The smaller the value of the criterion, the better the fitted model balances these objectives.

- Given a set of fitted candidate models, the model corresponding to the minimum value of the criterion is preferred.

- A parsimonious fitted model that adequately conforms to the data (i.e., a fitted model that is as simple as possible, but not simpler) generally has good predictive properties.

Akaike Information Criterion

- The *Akaike Information Criterion*, AIC, is the most widely known and used model selection criterion.

- AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of maximum likelihood estimators.

- Other popular model selection criteria include the corrected Akaike information criterion (AICc), the Takeuchi information criterion (TIC), Mallows' conceptual predictive statistic (C_p), Akaike's final prediction error (FPE), the predictive sum of squares statistic (PRESS), and the Bayesian information criterion (BIC).

Akaike Information Criterion

- For a candidate model of interest, let $f(y|\theta)$ denote the likelihood (i.e., the joint density of y), and let k denote the number of model parameters.

- Definition:

$$\text{AIC} = -2\ln f(y|\hat{\theta}) + 2k$$

- $-2\ln f(y|\hat{\theta})$ is called the "goodness-of-fit" term. This term decreases as the fit of the model improves.

- $2k$ is called the "penalty" term. This term increases as the complexity of the model grows.

Akaike Information Criterion

AIC Example

- Consider again the fitted Poisson polynomial regression models fit to the Belgium AIDS data.
- Which model is favored by AIC?

d	k	$-2\ln f(y|\hat{\theta})$	AIC
1	2	162.4	166.4
2	3	90.9	96.9
3	4	90.7	98.7
12	13	81.7	107.7

- AIC favors the quadratic model.

Akaike Information Criterion

- Advantages to the use of AIC
  - The application of the criterion does not require the assumption that one of the candidate models is the "true" or "correct" model.
  - AIC can be used to compare non-nested models.
  - AIC can be used to compare models based on different probability distributions.
- Disadvantages to the use of AIC
  - If the class of candidate models is large, the AIC values for several fitted models may be close to the minimum AIC value, meaning that an "optimal" fitted model is not clearly identified.
  - The successful application of AIC requires large samples, especially in complex modeling frameworks.

Final Thoughts

- If hypothesis testing is a problematic paradigm for model selection, why is it used so pervasively for this purpose?
- Nester (1996): "Tests of hypotheses are seemingly performed because . . .
  - (a) they appear to be objective and exact,
  - (b) they are readily available and easily invoked in many commercial statistics packages,
  - (c) everyone else seems to use them,
  - (d) students, statisticians and scientists are taught to use them,
  - (e) some journal editors and thesis supervisors demand them."
- As an alternative to hypothesis testing for model selection, the use of model selection criteria warrant consideration, especially in applications arising in the biomedical and health sciences.

Final Thoughts

- Model simplicity is defensible!
- "Simplicity is the ultimate sophistication."
  – Leonardo da Vinci (1452–1519).