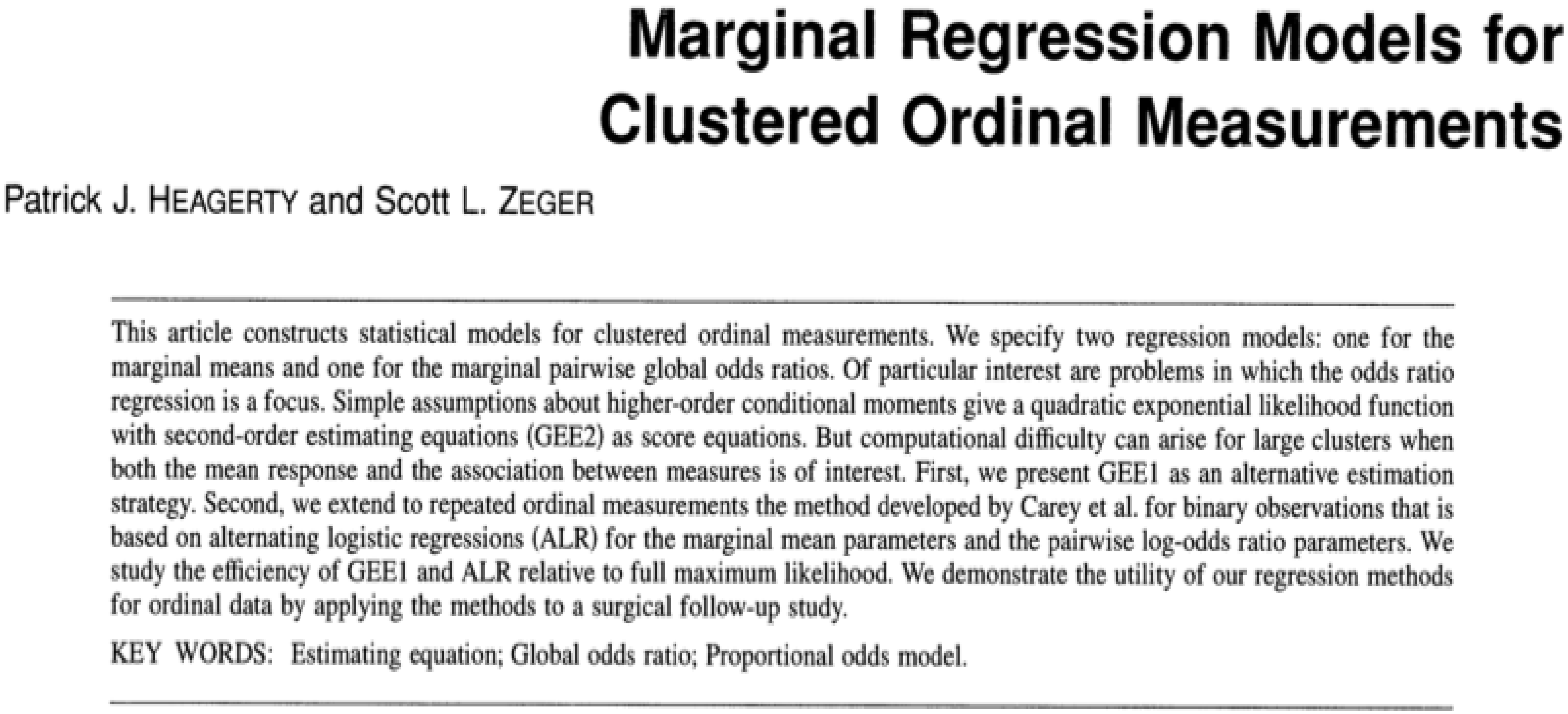
Anatomy of a Statistics Paper
(with examples)

- PJ Heagerty
- Department of Biostatistics
- University of Washington

Anatomy of a Statistics Paper

- **Q**: How to write an introduction?
- Comments on common elements
  - ▷ Methods
  - ▷ Evaluation
  - ▷ Illustration
- Steps in the writing process



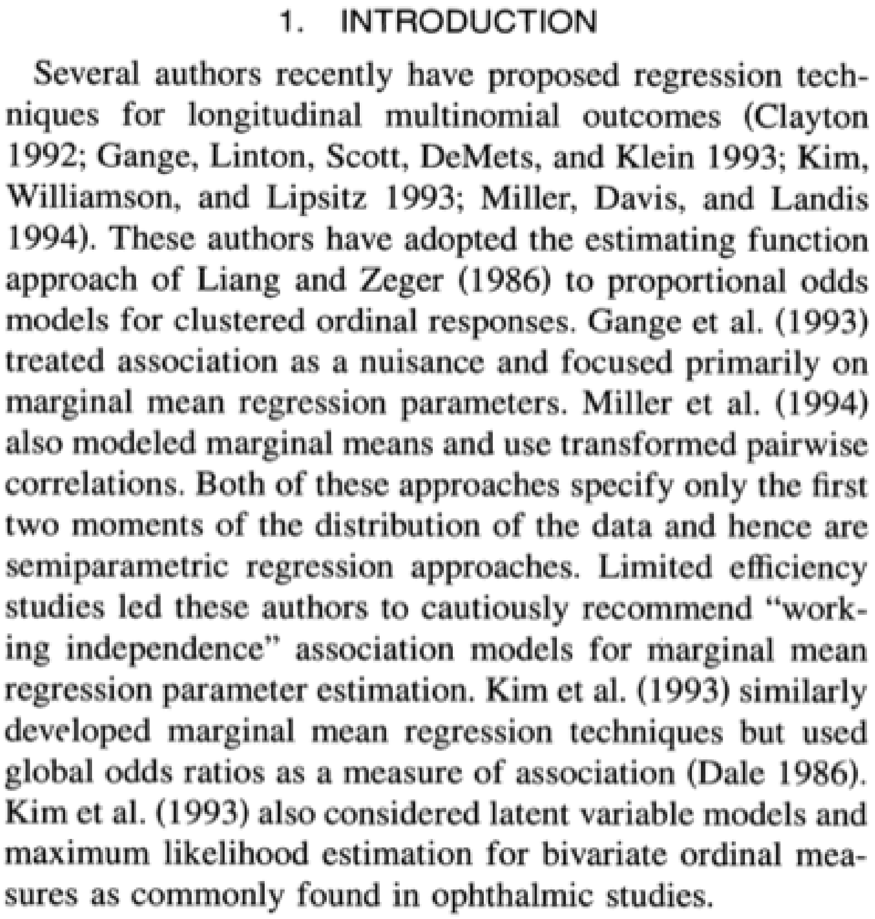
Marginal Regression Models for Clustered Ordinal Measurements

Patrick J. HEAGERTY and Scott L. ZEGER

This article constructs statistical models for clustered ordinal measurements. We specify two regression models: one for the marginal means and one for the marginal pairwise global odds ratios. Of particular interest are problems in which the odds ratio regression is a focus. Simple assumptions about higher-order conditional moments give a quadratic exponential likelihood function with second-order estimating equations (GEE2) as score equations. But computational difficulty can arise for large clusters when both the mean response and the association between measures is of interest. First, we present GEE1 as an alternative estimation strategy. Second, we extend to repeated ordinal measurements the method developed by Carey et al. for binary observations that is based on alternating logistic regressions (ALR) for the marginal mean parameters and the pairwise log-odds ratio parameters. We study the efficiency of GEE1 and ALR relative to full maximum likelihood. We demonstrate the utility of our regression methods for ordinal data by applying the methods to a surgical follow-up study.

KEY WORDS: Estimating equation; Global odds ratio; Proportional odds model.

Heagerty and Zeger (1996)

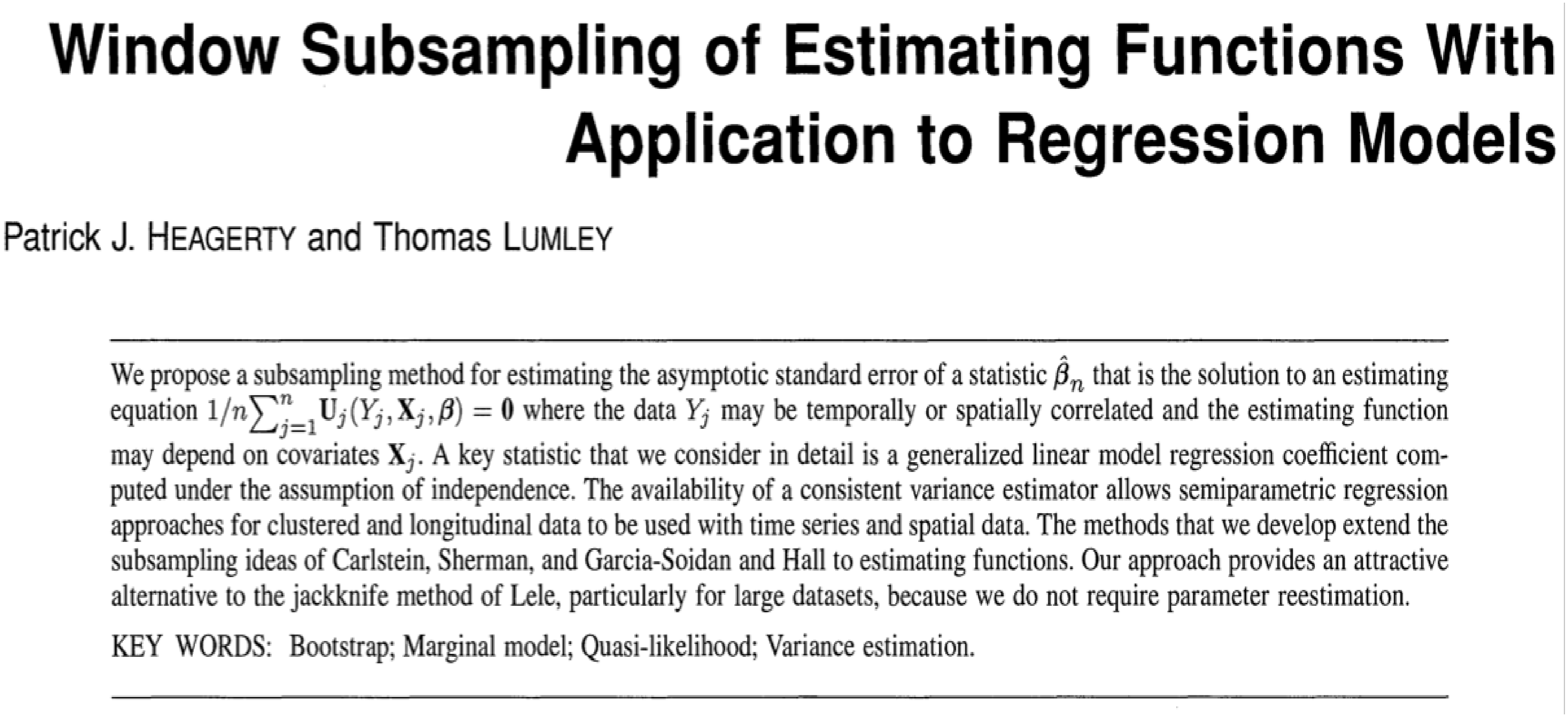


1. INTRODUCTION

Several authors recently have proposed regression techniques for longitudinal multinomial outcomes (Clayton 1992; Gange, Linton, Scott, DeMets, and Klein 1993; Kim, Williamson, and Lipsitz 1993; Miller, Davis, and Landis 1994). These authors have adopted the estimating function approach of Liang and Zeger (1986) to proportional odds models for clustered ordinal responses. Gange et al. (1993) treated association as a nuisance and focused primarily on marginal mean regression parameters. Miller et al. (1994) also modeled marginal means and use transformed pairwise correlations. Both of these approaches specify only the first two moments of the distribution of the data and hence are semiparametric regression approaches. Limited efficiency studies led these authors to cautiously recommend "working independence" association models for marginal mean regression parameter estimation. Kim et al. (1993) similarly developed marginal mean regression techniques but used global odds ratios as a measure of association (Dale 1986). Kim et al. (1993) also considered latent variable models and maximum likelihood estimation for bivariate ordinal measures as commonly found in ophthalmic studies.

Abstract / Introduction

- Introduction:
  - ▷ If possible the introduction should include clear **scientific motivation** for the new methods.
  - ▷ The introduction should clearly state the **statistical motivation** for the work.
    - ∗ **Q**: Why are existing methods not sufficient?
    - ∗ **Q**: What are elements of an attractive solution?
- Abstract: summarize and motivate.



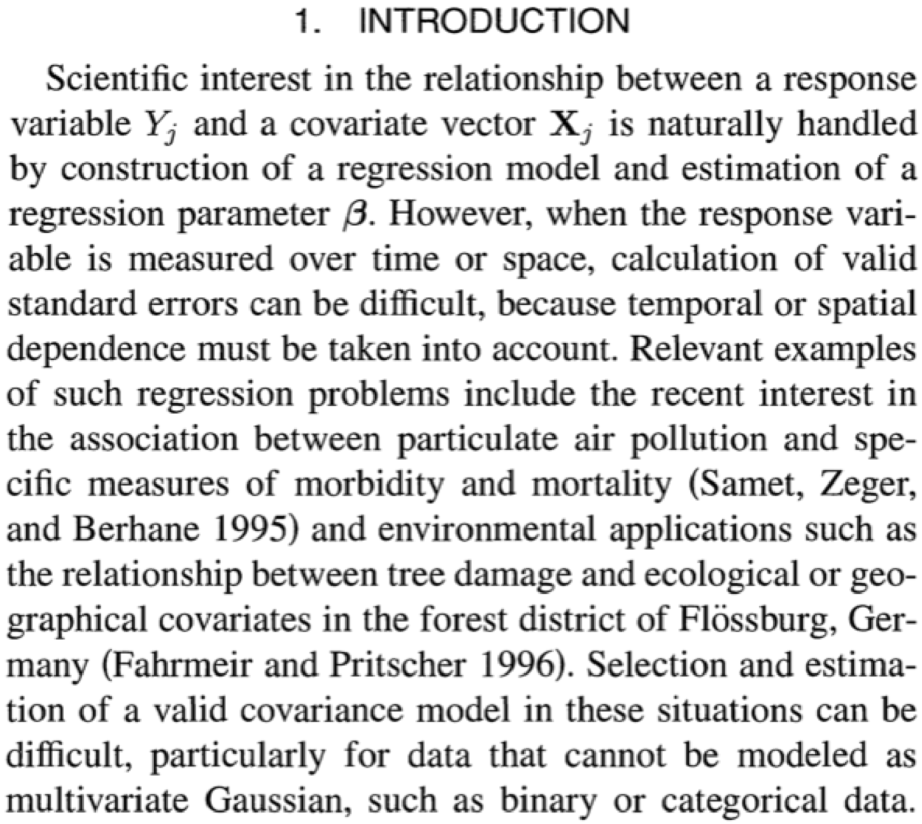
Window Subsampling of Estimating Functions With Application to Regression Models

Patrick J. HEAGERTY and Thomas LUMLEY

We propose a subsampling method for estimating the asymptotic standard error of a statistic $\hat{\beta}_n$ that is the solution to an estimating equation $1/n\sum_{j=1}^{n}\mathbf{U}_j(Y_j,\mathbf{X}_j,\beta)=\mathbf{0}$ where the data Y_j may be temporally or spatially correlated and the estimating function may depend on covariates $\mathbf{X}_j$. A key statistic that we consider in detail is a generalized linear model regression coefficient computed under the assumption of independence. The availability of a consistent variance estimator allows semiparametric regression approaches for clustered and longitudinal data to be used with time series and spatial data. The methods that we develop extend the subsampling ideas of Carlstein, Sherman, and Garcia-Soidan and Hall to estimating functions. Our approach provides an attractive alternative to the jackknife method of Lele, particularly for large datasets, because we do not require parameter reestimation.

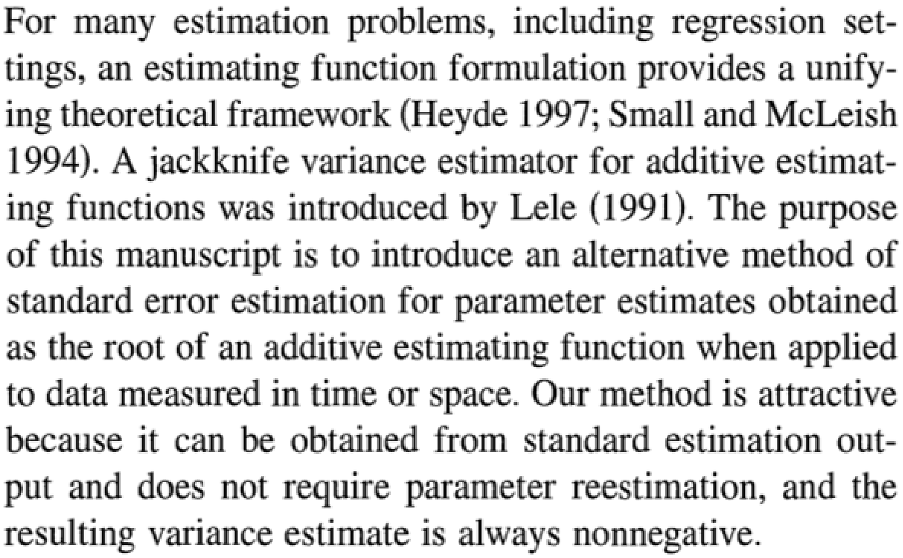
KEY WORDS: Bootstrap; Marginal model; Quasi-likelihood; Variance estimation.

Heagerty and Lumley (2000)



1. INTRODUCTION

Scientific interest in the relationship between a response variable Y_j and a covariate vector $\mathbf{X}_j$ is naturally handled by construction of a regression model and estimation of a regression parameter $\boldsymbol{\beta}$. However, when the response variable is measured over time or space, calculation of valid standard errors can be difficult, because temporal or spatial dependence must be taken into account. Relevant examples of such regression problems include the recent interest in the association between particulate air pollution and specific measures of morbidity and mortality (Samet, Zeger, and Berhane 1995) and environmental applications such as the relationship between tree damage and ecological or geographical covariates in the forest district of Flössburg, Germany (Fahrmeir and Pritscher 1996). Selection and estimation of a valid covariance model in these situations can be difficult, particularly for data that cannot be modeled as multivariate Gaussian, such as binary or categorical data.

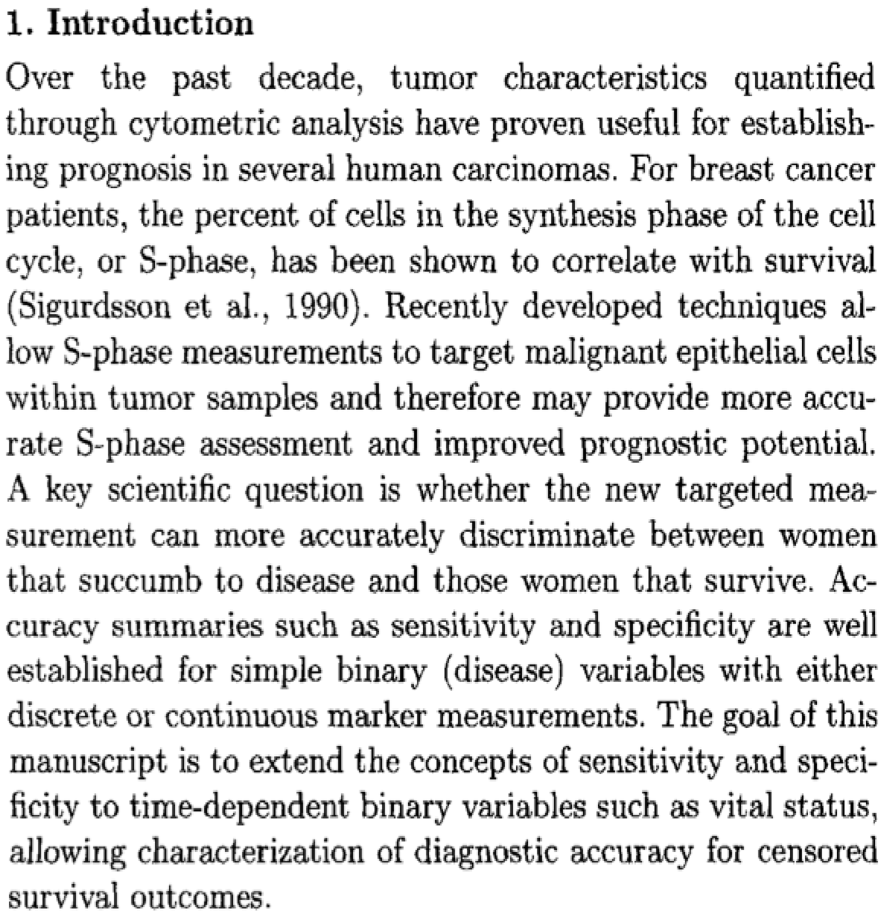


For many estimation problems, including regression settings, an estimating function formulation provides a unifying theoretical framework (Heyde 1997; Small and McLeish 1994). A jackknife variance estimator for additive estimating functions was introduced by Lele (1991). The purpose of this manuscript is to introduce an alternative method of standard error estimation for parameter estimates obtained as the root of an additive estimating function when applied to data measured in time or space. Our method is attractive because it can be obtained from standard estimation output and does not require parameter reestimation, and the resulting variance estimate is always nonnegative.

Introduction

- IMHO the introduction is ultimately the most important **writing** you will do for the paper.
- IMHO your reader will either be interested and continuing on with your paper, or...
- A scholarly introduction is respectful of the literature.
- In my experience, the introduction is part of a paper that I will outline relatively early in the process, but will finish and repeatedly edit at the end of the process.

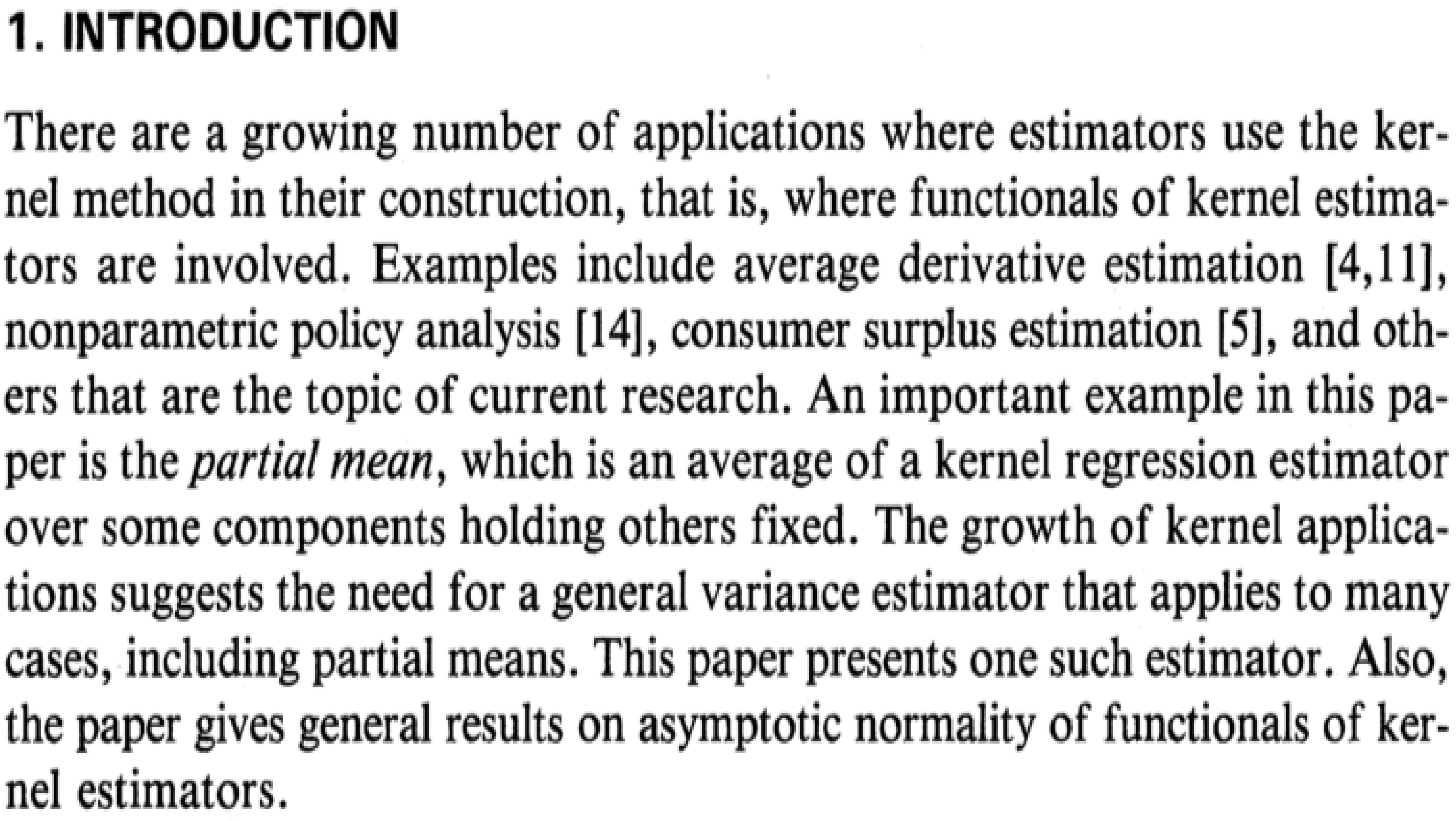
Heagerty, Lumley and Pepe (2000)



1. Introduction

Over the past decade, tumor characteristics quantified through cytometric analysis have proven useful for establishing prognosis in several human carcinomas. For breast cancer patients, the percent of cells in the synthesis phase of the cell cycle, or S-phase, has been shown to correlate with survival (Sigurdsson et al., 1990). Recently developed techniques allow S-phase measurements to target malignant epithelial cells within tumor samples and therefore may provide more accurate S-phase assessment and improved prognostic potential. A key scientific question is whether the new targeted measurement can more accurately discriminate between women that succumb to disease and those women that survive. Accuracy summaries such as sensitivity and specificity are well established for simple binary (disease) variables with either discrete or continuous marker measurements. The goal of this manuscript is to extend the concepts of sensitivity and specificity to time-dependent binary variables such as vital status, allowing characterization of diagnostic accuracy for censored survival outcomes.

Newey (1994) *Econometric Theory*



1. INTRODUCTION

There are a growing number of applications where estimators use the kernel method in their construction, that is, where functionals of kernel estimators are involved. Examples include average derivative estimation [4,11], nonparametric policy analysis [14], consumer surplus estimation [5], and others that are the topic of current research. An important example in this paper is the *partial mean*, which is an average of a kernel regression estimator over some components holding others fixed. The growth of kernel applications suggests the need for a general variance estimator that applies to many cases, including partial means. This paper presents one such estimator. Also, the paper gives general results on asymptotic normality of functionals of kernel estimators.

Paper Anatomy

- Methods:
  - ▷ Establish notation
  - ▷ Outline methods/results (use appendix for details)
  - ▷ The methods section is often the first part of a paper that I write.
- Evaluation:
  - ▷ Simulations (validity)
  - ▷ Relative efficiency (comparison)
  - ▷ Simulations can be non-informative (so your method seem to work...)

Paper Anatomy

- Illustration:
  - ▷ Genuine application can reinforce the theory/results through demonstration.
  - ▷ Purpose of statistics...
- Discussion:
  - ▷ I always write this last (and usually fast!)
  - ▷ Summarize, and “mark territory”

Summary

- As you read papers also notice the **construction** of the papers (learn from the good and bad examples).
- Abstract and Introduction – **keys** for getting readers engaged.
- Be **gentle** with your audience.
- Tell them your **story**.
- Writing is **work** – but ultimately **rewarding**!