A Tolerance Interval Estimation Problem in Regression Analysis

D. S. Robson

BU-414-M March, 1972

Abstract

The frequency distribution of breeding dates or fertilization dates is oftentimes unobservable in natural populations as well as in domestic populations such as crop plants. If the developmental trajectory can be estimated from data on known-age embryos, however, statistical inferences concerning the frequency distribution of fertilization dates can be drawn from the array of developmental stages observed in embryos collected on a fixed date following the breeding season. Questions of identifiability and statistically efficient methods for this type of back-calculation are still unresolved, even in the case where the developmental trajectory is a simple linear regression on age.

A Tolerance Interval Estimation Problem in Regression Analysis

D. S. Robson

BU-414-M March, 1972

Introduction

Most natural populations follow a synchronized reproduction cycle where breeding occurs over a relatively short season each year. Synchrony is an important aspect of the process and is presumably advantageous to the survival of the population. Statistically, this process is described by the probability distribution of breeding date, which in turn is aptly described by a tolerance interval estimate of the distribution. The direct observation and measurement of breeding dates is oftentimes difficult or impossible to accomplish in practice, however, and indirect methods have therefore been developed to circumvent this difficulty.

If developing embryos are collected on a fixed date shortly after the breeding season, the array of stages of embryonic development observed in this collection is an indirect reflection of the array of dates of fertilization. If each embryo followed the same development trajectory, differing only in the temporal location of starting points, then an embryo of age x is in stage f(x) and on the fixed date t after the breeding season a zygote which was fertilized on date τ is in stage $f(t-\tau)$. In an ideal situation where the function f(x) is known then the array of developmental stages y observed on date t could be transformed:

$$y = f(t-\tau)$$

$$\tau = t - f^{-1}(y)$$

to an array of fertilization dates. In practice there are nuisance variations in rates of development so that $f(x)$ is, at best, a regression function and the state Y on date t is

$$Y = f(t-\tau) + \epsilon$$

and hence the problem becomes more statistical in nature.

The linear regression case

If the physical size of the embryo is taken as a measure of development then typically the size increases exponentially with age during the early phase of growth, and on the logarithmic scale the trajectory is therefore linear. In cases where ambient temperature controls growth rate the age x will be measured in degree days, but with mammalian embryos logarithmic size should be a linear function of calendar age. The slope and intercept of the regression $E(Y|x) = \alpha + \beta x$ as well as the residual variance $\sigma^2_{Y\cdot x} = E\left[(Y-\alpha-\beta x)^2 \middle| x\right]$ can be estimated by sampling known-age embryos, perhaps from captive females or, in the case of plants, from hand pollinated ovaries.

The size Y_{ij} of a randomly selected embryo of unknown age collected on date t_i is then given by

$$Y_{ij} = \alpha+\beta\left(t_i-\tau_{ij}\right) + \epsilon_{ij}$$

where τ_{ij} is the unknown date of fertilization; thus,

$$Y_{ij} = \left[\alpha-\beta\bar{\tau}\right] + \beta t_i + \left[\epsilon_{ij}-\beta(\tau_{ij}-\bar{\tau}\right]$$

$$= \alpha^* + \beta t_i + \epsilon^*_{ij}$$

where $\bar{\tau}$ is the mean date of fertilization.

Given that the experimental design includes regression data for known-age embryos and a random sample of unknown-age embryos on at least one collection date, the statistical problem consists first of specifying conditions under which the distribution of fertilization dates is identifiable and then developing an estimation procedure. It would appear, for example, that homoscedasticity of the ϵ-residuals is a necessary condition for identifiability. Note that if the collection date t_i precedes the last date of fertilization then the τ_{ij} are sampled from a truncated distribution, $\tau_{ij} \leq t_i$; in particular, if the distribution of τ is assumed to be normal then t_i is inevitably a truncation point.