ELSEVIER

Journal of Clinical Epidemiology 79 (2016) 112−119

Journal of Clinical Epidemiology

Simple and multiple linear regression: sample size considerations

James A. Hanley*

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada

Accepted 6 May 2016; Published online 5 July 2016

Abstract

Objective: The suggested "two subjects per variable" (2SPV) rule of thumb in the Austin and Steyerberg article is a chance to bring out some long-established and quite intuitive sample size considerations for both simple and multiple linear regression.

Study Design and Setting: This article distinguishes two of the major uses of regression models that imply very different sample size considerations, neither served well by the 2SPV rule. The first is etiological research, which contrasts mean Y levels at differing "exposure" (X) values and thus tends to focus on a single regression coefficient, possibly adjusted for confounders. The second research genre guides clinical practice. It addresses Y levels for individuals with different covariate patterns or "profiles." It focuses on the profile-specific (mean) Y levels themselves, estimating them via linear compounds of regression coefficients and covariates.

Results and Conclusion: By drawing on long-established closed-form variance formulae that lie beneath the standard errors in multiple regression, and by rearranging them for heuristic purposes, one arrives at quite intuitive sample size considerations for both research genres.

Keywords: Precision; Power; Prediction; Confounding; Degrees of freedom

1. Introduction and background

The suggested "two subjects per variable" (2SPV) rule of thumb in the Austin and Steyerberg [1] article is a chance to bring out some long-established and quite intuitive sample size considerations for both simple and multiple linear regression. The basis for these considerations is becoming increasingly obscured by the use of specialized black-box power-and-sample size software, by reliance on rules of thumb based on very specific and not always informative numerical simulations, and by limited coverage of the structure of the variance formulae behind the regression outputs.

By way of orientation, it is important to distinguish two major uses of regression models; they imply very different sample size considerations, neither served well by the 2SPV rule. The first is etiological research, which contrasts mean Y levels at differing "exposure" (X) values and thus tends to focus on a single regression coefficient; I will deal later with the sample size issues for this genre, particularly in (nonexperimental) etiological research involving adjustment for confounders. I will begin with statistical considerations for a second research genre, one that guides clinical practice. This type of research addresses Y levels for individuals with different covariate patterns or "profiles." It focuses on the profile-specific (mean) Y levels themselves, estimating them via linear compounds, that is, combinations of regression coefficients and covariate values.

2. Sample size issues in fitting "clinical prediction" models

In the "clinical prediction" models used in Steyerberg's 2012 book [2] to estimate diagnostic and prognostic probabilities, the "Y" is binary. The antilogit of the fitted linear compound yields the fitted mean Y at any specific profile (covariate pattern) and serves as the estimated probability for that profile. Assuming that the statistical model is appropriate and that the setting remains the same, a profile-specific estimate of say 76% probability, with a (say 95%) "margin of error" of 10% conveys the entire statistical uncertainty concerning the Y of a new (i.e., unstudied) individual with that same profile. Of course, the interval could be narrowed, to say 74% plus or minus 5%, by using a sample size four times larger. (If the issue is the probability that a cancer in a particular type of patient is confined to the prostate, or that therapy will be successful, or that it will rain tomorrow, it is not clear how much is gained by the increased precision.)

Conflict of interest: None.

This work was supported by the Canadian Institutes of Health Research.

* Corresponding author. Tel.: +1 514 398 6270; fax: +1 514 398 4503.

E-mail address: james.hanley@mcgill.ca

http://dx.doi.org/10.1016/j.jclinepi.2016.05.014

0895-4356/© 2016 Elsevier Inc. All rights reserved.

What is new?

Key findings, What this adds to what was known

- Variance formulae in multiple regression can be rearranged and used heuristically to plan the sizes of studies that use linear regression models for clinical prediction and for confounder adjustment.

What is the implication and what should change now?

- These two different research genres demand different sample size approaches, focusing on either the value of one specific coefficient in a multiple regression, or a linear compound of the regression coefficients and the variates formed from a patient-specific covariate profile.
- Formulae derived from first principles are more instructive than rules of thumb derived from simulations.

Many of the principles in the textbook apply equally to situations where Y is "continuous" (e.g., the length of catheter [3] or breathing tube [4] required, or body surface area estimate for a drug dose calculation) in a patient with a specific anthropometric or clinical profile. However, although "regular" (i.e., quantitative Y) regression is considered simpler to understand than, and usually taught before, its logistic regression counterpart, there is one important aspect in which it is more complex. The single parameter—the probability or proportion—that governs a "Bernoulli" random variable Y allows us to fully describe the distribution of Y. But (ever and ever more precise estimates of) the mean of the distribution of a quantitative random variable Y tell(s) little else about the distribution: its center and spread are usually governed by separate parameters. A profile-specific estimate of say 40 cm, with a (say 95%) margin of error of 1 cm, for the mean catheter length required for children of a given height, conveys no information about where, in relation to this 39- to 41-cm interval, the required length might be in a future child of that same height.

2.1. Simple linear regression

Many of the sample size/precision/power issues for multiple linear regression are best understood by first considering the simple linear regression context. Thus, I will begin with the linear regression of Y on a single X and limit attention to situations where functions of this X, or other X's, are not necessary. As an illustration, I will use a genuine "prediction" problem. (Some clinical "pre"-diction problems, including diagnostic ones, and the quantitative examples I cite and use, do not involve the future but the present. They might be more suitably described as "post"-diction problems. The Y already exists, and the uncertainty refers to what it would be if it were measured now, rather than allowed to develop and be observed in the future.) Although it erupts much more frequently than others, the Old Faithful geyser in Yellowstone Park is not nearly as regular as its name suggests: the mean of the intervals (Y) between eruptions is approximately 75 minutes, but the standard deviation is more than 15 minutes. So that tourists to the (quite remote and not easily accessed) Park can plan their few hours onsite, officials (and now the live webcam [5] and special app [6]) provide them with an estimate of when the next eruption will occur. Rather than providing the overall mean and SD, they use the duration of the previous eruption (X, lasting 1–5 minutes) to considerably narrow the uncertainty concerning the wait until the next one.

Panels A–D in Fig. 1 show the prediction intervals derived from nonoverlapping samples of size $n = 16$, 32, 64, and 128 daytime observations from November 1995. (Subsequent earthquakes in the region have lengthened the mean interval and altered the prediction equation.) For illustration, we show the (estimated) prediction intervals at three specific X values (X = 2, 3, and 4 minutes). Each prediction interval reflects the statistical uncertainty involved. Its half width is calculated as a Student-t multiple of an X-specific standard error (SE). The SE, in turn, is a multiple of the root mean squared error, or RMSE, an n-2 degrees-of-freedom estimate of the standard deviation (σ), obtained from the n squared residuals.

As shown in the Fig. 1A inset, the SE has three components. The first is related to how precisely the point of departure—the mean Y level at the mean X of the studied observations—is estimated. This precision, reflected by the narrowest part of the inner shaded region, involves just (the RMSE estimate of) σ, and n. The second, related to the estimated mean Y level at the X value of interest, is governed by the precision of the estimated slope (this precision is a function of the RMSE, n, and the spread of the X's in the sample) and how far the X value associated with the "new" Y is from the mean of the X's in the sample. The X factor can be simplified to a z-value, one that governs the bow shape of the inner region. The first and second components involve the RMSE and n in the same way, and so, as Fig. 1 shows, the width of the inner region can be narrowed indefinitely by increasing n. However, the inner region only refers to the center of the X-specific distribution of Ys, not to the possible individual Y values. For this, one must add the third variance component (σ^2 itself) reflecting the variation of a future individual.

A number of lessons can be illustrated with this simple example. First, the research "deliverable," and thus the statistical focus, is not a regression coefficient or an R-square value. For every X value that might arise, it is a pair of numbers, both measured in minutes. Assuming that the distribution has a Gaussian form (In scientific contrasts involving means, the Central Limit Theorem helps statistics