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/

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

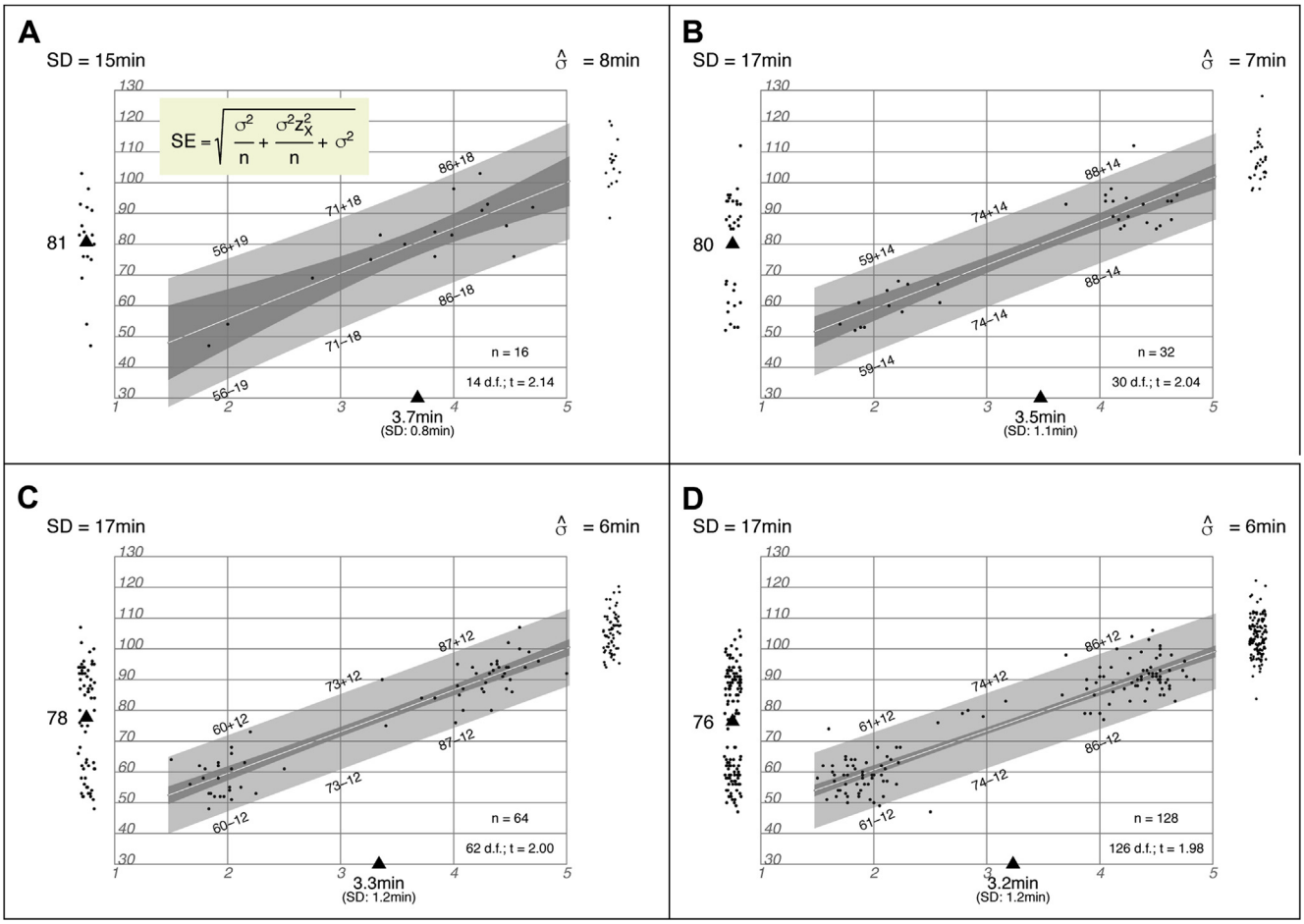

Fig. 1. (A–D) Prediction intervals for time to next Old Faithful eruption (vertical axis), based only on duration of previous eruption, derived from four different size samples. Estimated prediction intervals at three specific X values (2, 3, and 4 minutes) are shown. The darker and lighter shaded regions reflect the uncertainties associated with the mean and the individual, respectively. The half width of each interval is the product of the t-value and the SE (formula shown in inset). The σ in the latter is estimated by the root mean squared error, or RMSE, the standard deviation of the residuals from the simple linear model. These residuals are repeated at the right side of the panel. Shown on the left of each panel are the residuals from the null (intercept only) model, and above them their standard deviation. The means of the X and Y values are shown as triangles.

such as regression coefficients, based on Ys from non-Gaussian distributions, to have a close-to-Gaussian sampling distribution. This law-of-cancellation-of-extremes does not apply to individual Y values, and so the assumptions concerning the shape of the X-specific Y distribution are important.), they are the best estimates of the boundaries that enclose some central percentage (usually 95%) of the distribution of future Y values at that X.

Second, because the first two components of the SE involve *n*, larger sample sizes can narrow the statistical uncertainty about the center of this distribution, but they cannot alter σ itself, even if the greater number of degrees of freedom ensure that it is estimated more precisely. In the Old Faithful context, additional powers of X, and additional easily measured variables (e.g., the height of the previous eruption, the duration of and intervals between even earlier ones) did not substantially narrow σ. In the clinical contexts, the smallest (and honestly estimated) σ achievable is very much a function of the anthropometric or physiological or conceptual proximity of the Y and one of just a few determinants and is seldom reduced by increasing additional more distant ones.

Third, in many situations, only one of the two boundaries will be of interest: to avoid injury to the target organ, the specialist blindly introducing the catheter will stop short of, and maybe use fluoroscopy to guide the tip to, its final location; thus, the lower boundary is more relevant. Park officials also are probably more legally concerned about the statistical correctness of statements concerning the earliest the next eruption will occur, whereas in a somewhat related context [7], concern is with the statistical correctness of statements concerning an upper boundary of a reference distribution.

Fourth, most textbooks limit their coverage of ‘‘predictions for a new individual’’ to point-estimates of the boundaries, just as the panels in Fig. 1 do, and ignore the estimation error involved in these. Nowadays, with resampling methods, it is possible, as we did [7] to widen the boundaries to allow for this additional uncertainty.

Finally, both the Old Faithful and the anthropometric examples show the limitations of focusing on R-square as a

measure of "fit" or goodness of the predictions. How useful they are is better measured (or, rather, judged) in the same units as Y is, namely by the RMSE, and by the narrowness of the X-specific Y intervals. By contrast, the R-square measure is range dependent: it will have a higher value if calculated over a wider X range, although the σ of the X-specific Ys does not necessarily change over that wider X range. In addition, even if we ignore this arbitrariness, the fact that R-square is based on reductions in variance (rather than SD) leads to exaggerated measures of performance. Narrowing the standard deviation from 15 to 5 minutes narrows it by 2/3rds, or 67%, not by 8/9ths or $R^2 = 89\%$. Variance (the square of the SD) is indeed the more useful entity in mathematical statistics: as is evident inside the square root sign in the inset, uncertainties add "in quadrature"; moreover, when they occur together in a term, it is σ^2, not σ that opposes (is counteracted by) the sample size, *n*. But the variance (squared SD) is not a useful unit for Park officials or clinicians, or their clients. (A former colleague—a physician and statistician—liked to point out that if the average is 1.4 children per mother, and the standard deviation is 1.5, then the variance is 2.25 square children per square mother.)

Finally, there are two technical statistical comments. They concern the estimation of σ, and the numbers of subjects per variable that Austin and Steyerberg focused on. First, the various sample sizes used in Fig. 1 give an informal sense of the (in)stability of the estimates of the dominant parameter σ. The margins of error in estimating σ follow a predictable pattern (percentages derived theoretically [MSE ~ $\sigma^2 \times$ Chi-Sq/df; exact lower limit. Upper limits different; approx.] but rounded to nearest 5 for simplicity). As Table 1 summarizes, the precision depends directly on the number of degrees of freedom, and thus (across situations where *p* may vary widely) only indirectly on the SPV.

Second, the instability of the RMSE at lower sample sizes can be compensated for by using t multiples rather than z multiples of the SE, but the two multiples are practically equal from 30 degrees of freedom onwards. Ultimately, however, the concern is not so much with the (reducible by sample size) uncertainty in estimating σ, but with—even if σ were known perfectly—the uncertainty that σ itself implies about the Y value for a future individual. How small σ needs to be to be of practical use is a subject-matter judgment, not something that is settled with a larger or smaller *n*.

2.2. Multiple linear regression

What changes, as for sample sizes, as one moves up from predictions based on multiple, rather than simple, linear regression? The prediction of adult height—an early application of linear regression to human data—is instructive. In Pearson and Lee's carefully collected late 1880s family data set [9], the overall standard deviation (the RMSE in the null, intercept-only regression model) of the $n > 2{,}000$ adult heights was approximately 3.7 inches. In their fitted regression models, the RMSE was 2.5 inches when the model was limited to gender, 1.5 inches when one or other parental height was added, and 1.3 inches when both heights were added. In the $n \approx 60$ Berkeley data set of children, born in 1928/9, and used in Weisberg' classic regression textbook [10], the RMSEs obtained by beginning with a null model, and sequentially adding gender, and height at 2 or 9 years were 3.6, 2.5, 1.8, and 1.4 inches, respectively. Current online calculators [11] use a combination of the (half a dozen or so) parental heights and child gender height and age variables. Genomics companies will likely soon offer predictions based on several thousand. However, although they may be able to further reduce the "nature" component of variance, the "nurture" component will not be tamed.

The 2SPV rule does not help plan the *n* for studies of how much the overall standard deviation (here more than 3.5 inches) can be reduced by including *p* variates (variate: a term in the model; several variates might be a derived from one variable). When $n > p$, the main determinant of the precision with which the SD can be estimated is the number of degrees of freedom, $(n - 1) - p$, not (n/p) per se. Table 1 continues to apply, as long as *p* is small relative to *n*. If it is not, then the RMSE multiplied (inflated) by the square root of $n/(n - p)$ provides a more realistic measure of future performance. $[(n - 1)/[(n - 1) - p]$ is used in computing an adjusted R-square, a quantity introduced to econometrics by Theil in 1961. It is based on the same theory that governed the behaviors studied, via extensive simulations, by Austin and Steyerberg.]. The fact that one needs to consider both $n - p$ and *p*, and not just the n/p ratio, explains why the SPV-only rules cited in the first paragraph of section 2 of Austin and Steyerberg's article, as well as their own rule, vary as much as they do.

Clearly, when $n < p$, so that NPV < 1, as it often is in genomic studies, there are no degrees of freedom to provide an internal estimate of σ. Even if $n > p$, but the "*p*" used in the final model is a "best" subset of the much larger set of *p* variates searched, sample size guidelines based only on an *n*:*p* ratio are difficult to specify. The honest way to assess the performance in future subjects is to use an entirely separate test sample.

Table 1. Margin of error as a function of degrees of freedom

Degrees of freedom used to estimate σ	10	20	40	80	160
95% margin of error in estimating σ	30%	25%	20%	15%	10%

See chapter 4.4 of Harrell [8] for additional details and on why a higher SPV is needed if predictability is low.

3. Etiological research: the sample size cost of adjusting for confounding

Vittinghoff and McCulloch [12] used simulations to study binary (as well as failure time) end points (Ys). To mimic analyses of causal influences in observational data,

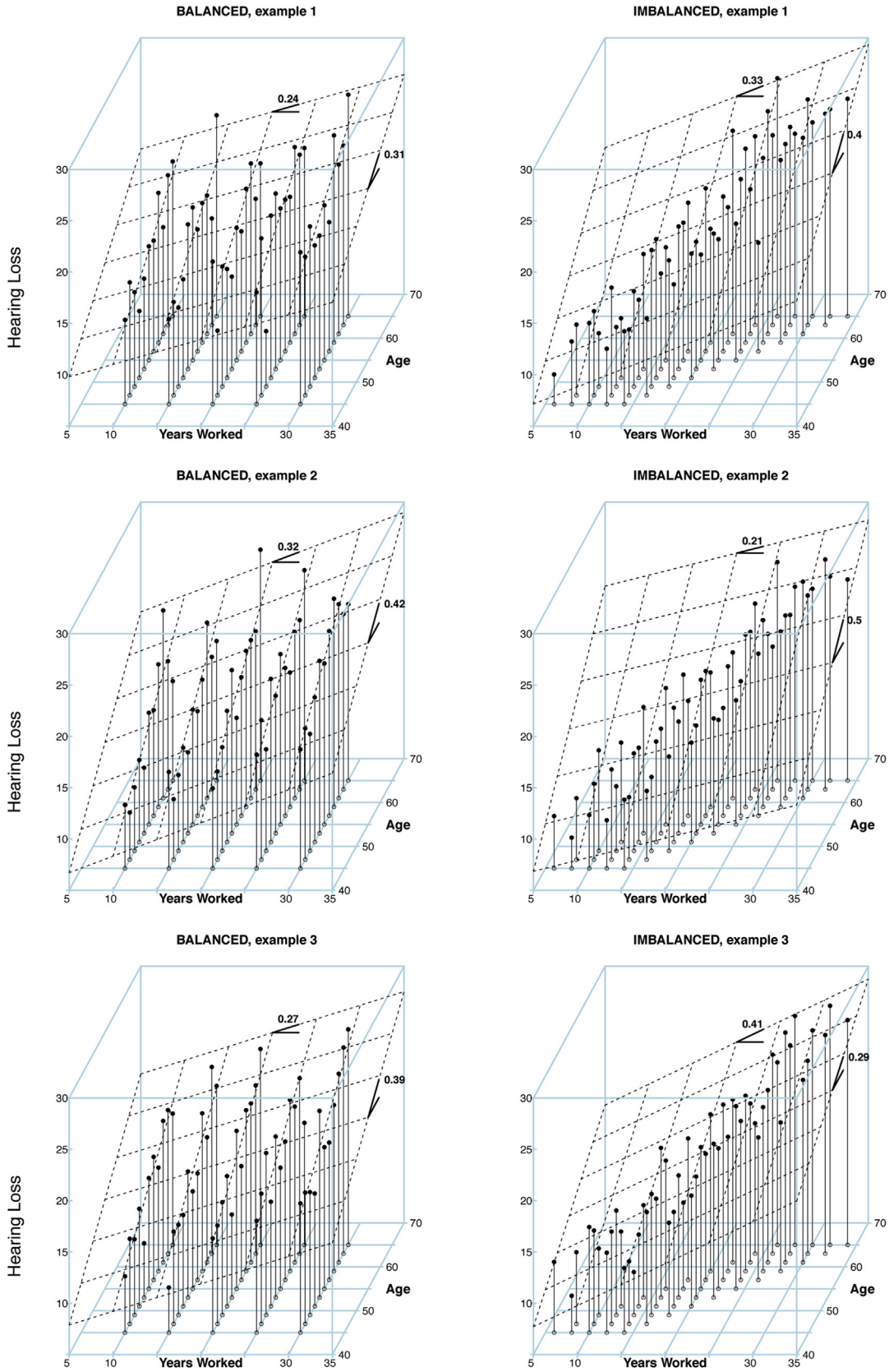

Fig. 2. (In)stability of fitted hearing loss regression equations (planes) when the correlation between age and duration of work (shown as open circles on the ''floor'' of each panel) is minimal (3 leftmost panels) and very high (3 rightmost panels) (squared correlations: 0 and 0.83 = 5/6). Hearing loss data (filled circles) were generated from true regression coefficients of 0.3 (work) and 0.4 (age). The fitted regression coefficients

they focused on a primary X, either binary or continuous. They regarded the other Xs (taken to be multivariate normal with pairwise correlations of 0.25; and having a multiple correlation of 0, 0.25, 0.5, or 0.75 with the binary X, and 0, 0.1, 0.25, 0.5, or 0.9 with the continuous X) as adjustment variables.

For continuous Ys, sample size calculations can again be based on closed-form formulae, derived from mathematical statistics and matrix algebra. These show directly and explicitly what determines the precision and power with which the primary regression coefficient can be investigated. These and other formulae have already been set out for a larger number of settings [13] and so only those relevant to the present context will now be summarized.

3.1. Simple linear regression

Central in the precision and power considerations is the SE of the estimated primary regression coefficient. Again, its structure is best understood in the simple regression context, where it equals the RMSE multiplied by the inverse of the square root of the number, n, studied, and by the inverse of the standard deviation, SD_X, of the Xs studied: $SE = RMSE \times [1/SD_X] \times [1/sqrt(n)]$. (As explained in Hanley and Moodie [13], there is no need to consider a binary and a continuous X separately.) Sadly, this structure is rarely used and often goes without comment, although it can be taken as a very valuable point of departure for heuristic purposes [13]. Some prefer the standardized regression coefficient, that is, $B' = B \times SD_X$; its SE is RMSE/sqrt(n), a helpful form JH has not seen given explicitly elsewhere.

Null and alternative values (For planning purposes, they would usually be considered equal under the null and the alternative.) of the SE can be used in the universal sample size and power formula $Z_{\alpha/2} \times SE_{null} + Z_\beta \times SE_{alt} = \Delta$, and algebraically rearranged as needed to project the precision or power achievable with a given n, or the n required for a given precision or power [13].

A related issue needs to be addressed before considering multiple regression. The fact that the SE of the estimated regression coefficient is inversely related to the SD of the X values used makes explicit what researchers instinctively know: it is difficult to measure a slope (e.g., fuel consumption of a car) over a short X range (distance). Even if the range cannot be widened, the slope is more precisely estimated if (as in the Old Faithful example in Fig. 1) the X values are spread more evenly over, or even more toward the extremities of, that range. Sadly, some researchers insist that their trainees check both the Ys and the Xs for normality. Neither check makes sense. If normality is in fact relevant for the "Y" variable (it is in an individual prediction setting, but not in an etiological setting), it is the normality of the residuals (not the Ys themselves) that matters. But normality of the X's is not a good thing; it would be better if, over the range of interest, the Xs had a closer to upside-down-normal distribution, with maybe some additional values from the center of the X range so as to check for linearity. Indeed, this author has heard a well known teacher make this point using a (hypothetical, Framingham-like) study where the sole focus was the slope of the relationship between heart disease and serum cholesterol concentrations (X). A random sample of subjects would lead to Xs concentrated near the center. It would have been far more statistically efficient to use say a three-point design, with equal numbers sampled from the bottom, middle, and top of the cholesterol range. If possible, unless the naturalistic X distribution is already favorable, one should choose the X values at which to measure Y. If one could be assured of linearity, the ideal is an X distribution where all observations are 1 standard deviation for the mean, that is, half are at each extreme.

3.2. Adjustment via multiple linear regression

Perhaps surprisingly, the SE of the estimated primary regression coefficient from a model that also includes $p - 1$ adjustment variables has a closed form that, when presented in a suitable didactic form, is again both concise and intuitive. It contains just one additional multiplier, involving a squared multiple correlation $R^2_{X\text{-other}X's}$, that reflects the correlation between the primary X and these adjustment variables:

$$SE = RMSE \times [1/SD_X] \times [1/sqrt(n)] \times \left[1/sqrt\left(1 - R^2_{X-other X's}\right)\right].$$

Hanley and Moodie [13] have rearranged this formula to link n with the precision and to estimate the power which the primary regression coefficient can be studied.

To understand why and how this additional term comes into it, and why the 2SPV rule is too limiting, consider two researchers who are interested in estimating the effect of working in a noisy workplace on hearing loss. They measure it as loss per year worked, that is, as a regression "slope," taking care to separate their estimate from the (also substantial) effect of age. Each has a budget to measure loss in n workers who have been exposed to a noisy work environment for different numbers of years.

for years worked and age (which fluctuate more in the imbalanced instances) are shown in bold along the edges of the fitted regression planes. For both designs to achieve the same precision of the estimated work coefficient, the unbalanced sample would need to be $1/(1 - 5/6) = 6$ times larger than the balanced one. The instability (the fluctuation in fitted regression coefficients) is easier to appreciate using the animated versions (using many different samples) available on the author's web site (http://www.biostat.mcgill.ca/hanley/software). Many versions of this figure were randomly generated, each one showing a different amount of variation among the three regression coefficients for years worked. The version whose variation lay at the 67th percentile was selected for this figure.

The two use different sampling schemes, illustrated in the two columns in Fig. 2. One simply randomly selects n workers with a range of ages, hoping to obtain a sample with a sufficiently wide spread in the numbers of years worked in a noisy environment. However, because many workers began working at around the same age, these n ages may, to a considerable extent, determine the numbers of years worked, and thus may lead to a very "unbalanced" sample, such as those in the rightmost panels of Fig. 2. Age and duration of work will be highly correlated, and it will be very difficult to isolate the effect of one from that of the other, even if it will be easy to obtain a precise estimate of the combined effect of the two.

The other researcher purposefully selects workers from each of several age slices, not randomly, but on the basis of years worked. In doing so, she tries to ensure, within each slice, the widest possible spread of numbers of years worked, and thus the greatest possible degree of "balance" (the lowest possible correlation) between age and work duration (leftmost panels of Fig. 2).

The mean age and the mean numbers of years worked are the same in both designs; the variance in the years worked is similar in both, whereas the variance in age is identical. Yet, the estimates of the work effect are much less variable in the second design because they fluctuate independently of those of age and because they are estimated across a wider range of work duration.

In the unbalanced design, the spread of the work durations within each age slice is much smaller and thus makes it more difficult to estimate the slope. This instability is easily visualized if, as J.H. does, one thinks of the fitted surface (regression model) as a "hammock" that is only secure at the bottom left and top right corners: but it can easily tip sideways so that the duration and age slopes (coefficients) are negative and positive, respectively, or vice versa. (R and Excel files that produce animated versions of Fig. 2 are available on the author's web site [http://www.biostat.mcgill.ca/hanley/software].) Only their sum (true value 0.3 + 0.4 = 0.7) is reliably estimated. Other teachers [14] have likened this situation to resting a flat surface (e.g., a rectangular sheet of cardboard) on a narrow base, or in the extreme, on a knife edge. Yet others [15,16] have used the "picket fence" analogy, where "responses resemble the pickets along a not-so-straight fence row" and "fitting a regression surface to these data is analogous to balancing a sheet of plywood on these pickets." (Yet others [17] make the task even more arduous by imagining that the picket fence runs uphill!) By contrast, because the fitted model in the balanced cases is secured/supported by a wide "base," its fitted coefficients are much more stable.

The increased instability with the unbalanced (collinear X) design is reflected in the multiplier 1/sqrt(1 − $R^2_{\text{X-otherX's}}$). In Austin and Steyerberg's investigation, these "SE inflation factors" for each of the 13 predictor variables were all less than 5%, a negligible degree of multicollinearity, seldom found in etiological studies. So as to guide the design of such studies, Table 2, adapted from that in Hanley and Moodie [13] shows the inflation in variance, and thus in sample size, to offset researchers' inability to study an etiology factor using an ideal (balanced) sample with no confounding.

Table 2. Variance (sample size) inflation as a function of the multiple R^2 of X with the remaining Xs

Multiple R^2 of X with remaining Xs	0.4	0.5	0.6	0.7	0.8
Variance (sample size) inflation[a]	1.7	2.0	2.5	3.3	5.0

[a] VIF = 1/(1 − multiple R^2).

4. Concluding remarks

In these two genres of research, sample size considerations are dominated by rules/algebra other than those that led Austin and Steyerberg to the 2SPV rule. The findings that lead to their "rule" could have been predicted from the same long-established mathematical-statistics theory relied on here. The absence of bias in Fig. 1 of Austin and Steyerberg is to be expected, and the negligible bias for a few variables at the lower subjects/variable end of that same figure may well stem from the fact that the full 13-term model could not always be fitted. The correct mean coverage proportions in Fig. 2 of Austin and Steyerberg are again a vindication, if such were needed, of the z- and t-based CIs worked out by Fisher in the 1920s; the tightness and size of empirical variation in the individual proportions are no surprise to (but surely the subject of envy by) pollsters who work with margins of error from samples of a thousand persons rather than a million simulated data sets (again, at the left end of the SPV scale, the extra amplitude probably reflects the attrition when the full model could not be fitted).

Austin and Steyerberg chose to plot the ratio of the mean estimated standard error to the standard deviation of the estimated coefficients. In their Figure 3, it starts at about 0.95 for the lowest SPV values, and increases to, but never quite reaches unity. Had they reported the mean estimated variance to the variance of the estimated coefficients, the pattern would have been simpler—and might have led to different conclusions. The complex behavior they reported was to be expected, given their choice of a quantity characterized by the chi rather than the chi-squared distribution. In a nonregression context, s^2 is a mean-unbiased estimator of σ^2 but s is not a mean-unbiased estimator of σ. In the regression context of their Figure 3, the estimated variance is a mean-unbiased estimator of the true variance, no matter the SPV value, whereas the estimated standard error is never an unbiased estimator of the square root of the true variance. In the regression setting that Austin and Steyerberg studied, the choice of scale was not an issue because

both patterns can be predicted using closed-form expressions derived from mathematical statistics. In more complex simulations, where we cannot rely on such laws, it is not obvious which reporting scale makes more sense. The broader issue of the difference between median-bias vs. mean-bias deserves to be better appreciated by the increasing number of investigators who rely on simulations to study the performance of statistical estimators.

We, like others, were impressed by Austin and Steyerberg's use of a real data set to simulate a million data sets of each of 50 different sample sizes from $13 \times 1 = 13$ to $15 \times 50 = 6{,}500$. Unfortunately, the statistical criteria they used are not the most relevant ones, so that the resulting 2SPV rule—which could have been derived directly from mathematical statistics, and for any number of variables or sample size—is of limited value. They did warn that such a ''rule'' should not be used to justify a proposed sample size to a peer review committee, where adequate statistical power and precision are more relevant.

It was the gaps in these latter and more important aspects that this note attempted to fill. An important first step was to draw a clear distinction between studies focusing on etiology, group prediction, and individual prediction, so that corresponding differences in sample size considerations for these different genres become more obvious. The second was to rely on relevant results from mathematical statistics as they apply to the reliability of results from fitted regression models. By some simple manipulation of the closed-form variance formulae found there, considerable sample size guidance can be found for a wide range of research scenarios, both etiologic and clinical.

Acknowledgments

Jay Kaufman and Ernest Lo provided valuable feedback on the first draft.

References

[1] Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol 2015;68:627−36.
[2] Steyerberg EW. Clinical prediction models. A practical approach to development, validating and updating. New York, NY: Springer-Verlag; 2009.
[3] Andropoulos DB, Bent ST, Skjonsby B, Stayer SA. The optimal length of insertion of central venous catheters for pediatric patients. Anesth Analg 2001;93:883−6.
[4] Szeto C, Kost K, Hanley JA, Roy A, Christou N. A simple method to predict pretracheal tissue thickness to prevent accidental decannulation in the obese. Otolaryngol Head Neck Surg 2010;143:223−9.
[5] Old Faithful Geyser Streaming Webcam, 2016. Available at http://www.nps.gov/features/yell/webcam/oldFaithfulStreaming.html. Accessed July 22, 2016.
[6] National Park Service. NPS Yellowstone Geysers: app for smart phones, 2016. Available at https://www.nps.gov/yell/learn/news/14089.htm. Accessed July 22, 2016.
[7] Hanley JA, Saarela O, Stephens DA, Thalabard JC. hGH isoform differential immunoassays applied to blood samples from athletes: decision limits for anti-doping testing. Growth Horm IGF Res 2014;24:205−15.
[8] Harrell FE. Regression modeling strategies. With applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Cham: Springer; Cham; 2015.
[9] Pearson K, Lee A. On the laws of inheritance in man: I. inheritance of physical characteristics. Biometrika 1903;2:357−462.
[10] Weisberg S. Applied linear regression. New York: Wiley; 1980.
[11] Healthy calculators. Available at http://www.healthycalculators.com/childrens-height-predictor.php. Accessed July 22, 2016.
[12] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007;165:710−8. Epub 2006 Dec 20.
[13] Hanley JA, Moodie EEM. Sample size, precision and power calculations: a unified approach. J Biomet Biostat 2011;5:2.
[14] Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med 2004;11:94−102.
[15] Hocking RR, Pendleton OJ. The regression dilemma. Commun Stat Theory Methods 1983;12:497−527.
[16] Hocking RR. Methods and applications of linear models: regression and the analysis of variance. 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2013.
[17] Cook RD, Weisberg S. Applied regression including computing and graphics. New York: Wiley; 1999.