

Power Calculations for Multiple Linear Regression

Clint Moore, USDI National Biological Survey

Many problems in biological studies rely on the estimation of parameters of a multiple regression model. In a typical example, one might estimate the increase or decline in abundance of a species by regressing an index of abundance on time. In most cases, repeated measurements of the index at a fixed time point are not expected to yield identical values because exogenous factors, either measured, unmeasured, or unmeasurable, all influence the outcome. Those measured factors that are believed to be linearly related to the response may serve as model *covariates*. For example, many abundance indices depend on the level of effort of the observer. Therefore, effort may enter the model as a nuisance covariate to provide greater accuracy and precision for the estimation of the covariate of primary interest, time. Covariates may take on continuous values, as in ordinary regression analysis, or they may appear as groups of one or more 0/1 *dummy* values, as in ANOVA.

Statistical power, the likelihood of rejecting a false null hypothesis, is an issue traditionally overlooked by researchers in reporting data analyses. The statement of "no statistical significance" for a hypothesis test is rather hollow without an accompanying statement of estimated power of the test. For example, failing to find a decrease in abundance of a species has more relevance for a study design having a 90% chance of detecting an annual decline of $\geq 3\%$ than for a design having only a 10% chance of accomplishing the same objective. For the first design, failing to detect a significant decline means that we have high confidence that abundance is stable, as long as we are willing to define "stability" as any annual decline less than 3%. On the other hand, failing to detect a decline under the second design is inconclusive with regard to population stability: the design's lower power implies that annual declines of 3% or greater are likely to go undetected.

We will need to introduce notation to further discuss multiple regression and power analysis. Suppose the following multiple regression model describes a set of data:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi} + \epsilon_i$$

for $i = 1, \dots, N$ observations and $p = 1, \dots, P$ independent variables. We would like to find (A) the precision of the partial regression coefficient ("slope") b_p for a test with a stated power, or, (B) the power of the test required to detect a slope of a specific magnitude. We denote power, the probability of rejecting the null hypothesis $b_p = 0$ when the hypothesis is false, as $1 - \beta$. We will need to make 3 assumptions about the ϵ_i , the model prediction errors, before starting. These assumptions apply to any regression analysis problem:

1. the ϵ_i are independent of one another,
2. the ϵ_i vary homogeneously (often, distribution of the ϵ_i varies across values of x_p , invalidating the assumption), and
3. the ϵ_i come from a normal distribution.

It is possible to express all the necessary calculations for power analysis in algebraic form and perform the computations on a pocket calculator, but it becomes impractical when $P \geq 2$. Therefore, we will resort to matrix algebra to carry out the calculations. A personal computer spreadsheet package with matrix transpose, multiplication, and inverse operators should be able to provide the calculations. Automating the calculations through the use of macros makes it possible to quickly experiment with variations of the design and reduces the likelihood of math errors. Rawlings (1988) provides a good introduction to matrix algebra calculations for multiple regression.

A word of caution: matrix inversion operators in many computer packages are notoriously prone to producing garbage when x variables are nearly collinear (that is, when one variable is almost redundant with another variable or with a linear combination of several variables) or when values for an x variable are excessively great ($> 10^2$) or small ($< 10^{-2}$) in magnitude. Unfortunately, there is little that can be done in the first case short of dropping offending x variables, but scaling data is a safe remedy in the second case. For example, if x_p is measured in grams and takes on values $> 10^5$ g, one may want to convert the values to kilograms, then center the values by subtracting the mean from each value.

Multiple regression is accomplished in a series of operations on the matrix X , created by placing the columns x_1, x_2, \dots, x_p side by side after a leading column of ones, and the column vector y . An important quantity is the symmetric matrix $C = (X'X)^{-1}$. We are particularly interested in c_{pp} , the $p+1$ -th diagonal element of C (the first diagonal element, which corresponds to the leading column of ones in X , is indexed $0,0$; the last one, the $P+1$ -th, is indexed P,P), because it is used in calculating the estimated variance of \hat{b}_p : $\hat{\sigma}_b^2 = c_{pp} \hat{\sigma}_{y|X}^2$, where $\hat{\sigma}_{y|X}^2$ is the estimated residual variance of y given X (i.e., MSE). Better software packages may be able to supply C and $\hat{\sigma}_{y|X}^2$ explicitly, as a part of a linear regression routine.

In non-regression situations, e.g., the one-sample location problem or the problem of comparing 2 means, sample size N represents the sole design component of the power relationship. Therefore, it is simple to calculate a sample size for a given value of power for these designs. In contrast, components of the multiple linear regression design include P , distribution of any single x_p , the joint distribution of any pair of (x_j, x_k) , as well as N . In this situation, it generally becomes impossible to associate a single value of sample size, or any other design component, with a single value of power. For example, 2 designs yielding the same power may differ

appreciably in sample size; the design with the smaller sample size may have its observations distributed over a wider span of the independent variable, as just one possibility. What one typically does is invoke a trial-and-error process of achieving a desired power or level of precision under likely variations of the design.

Each design variation implies a unique matrix X . Element c_{pp} derived from X is inversely related to sample size and range of x_p and is directly related to degree of correlation between x_p and any other x_k .

The general procedure will be laid out below, and specific examples will follow. The first step is to obtain an estimate of $\hat{\sigma}_{y|X}^2$, either through an analysis of pilot data or by guessing its extreme values. If you can only supply a guess, err on the side of conservatism by choosing the largest value that seems reasonable to you. Otherwise, if you have pilot data comprising N_0 observations of y_i on P_0 independent variables in the pilot design X_0 , calculate

$$\begin{aligned} C_0 &= (X_0'X_0)^{-1} \\ df_0 &= N_0 - (P_0 + 1) \\ Z &= X_0'y_0 \\ B_0 &= C_0Z \\ \hat{\sigma}_{y|X}^2 &= (y_0'y_0 - B_0'Z) / df_0 \end{aligned}$$

where y_0 is the column of y_i values. For purposes of planning a study, you may wish to consider a range of values around $\hat{\sigma}_{y|X}^2$.

The next step is the most difficult, and it requires specifying one or more X matrices corresponding to alternative study designs. The difficulty depends on how well the values of each x_p can be predicted in a study yet to be conducted. For some x_p , values either change fairly systematically or can be controlled by the experimenter. For example, values representing successive periods of time are an example of how values of x_p change in a predictable way. Values of x_p representing settings of a machine are an example of experimental control. In either case, design values of x_p are not difficult to propose. Specification of X is most difficult if one or more of the x_p vary randomly or in a way not controlled or anticipated by the experimenter. If air temperature, for example, influences y_i , it is important to include the variable in the design, but it is impossible to predict what future values the variable may take. However, it should be possible to predict the range of values for the variable, and to specify various degrees of correlation with other variables. Specific combinations of these predictions produce different X matrices and different estimates of power.

The third step is to calculate components of the power analysis. For each design, construct matrix X and determine N and P . Calculate residual degrees of freedom for X as $df = N - (P + 1)$. Calculate C and extract c_{pp} . Calculate

$\hat{\sigma}_b = (c_{pp} \hat{\sigma}_{y|x}^2)^{1/2}$. Decide the highest acceptable rate of falsely rejecting the null hypothesis (type I error rate α), and decide if the rejection region for the alternative hypothesis should be one or two-sided. Locate the Student's t -value for df degrees of freedom that corresponds to the lower tail probability $1 - \alpha$ for one-sided alternatives or $1 - \alpha/2$ for two-sided alternatives. We will call this value t^* .

The last step is the calculation of power or precision. If finding precision with stated power is the objective (objective A), then we must locate $t_{df, 1-\beta}$, the t -value that corresponds to the lower tail probability $1 - \beta$ and df degrees of freedom. Calculate $\tilde{b}_p = \hat{\sigma}_b (t^* + t_{df, 1-\beta})$. Quantity \tilde{b}_p (or $-\tilde{b}_p$ if the test is one-sided and the rejection region is negative) is the estimate of the smallest detectable slope given power, type I error, design, and residual variance.

If finding power with stated precision is the objective (objective B), then we must specify the smallest value of b_p we wish to be able to detect. Calculate $\tilde{t}_{df} = (|b_p| / \hat{\sigma}_b) - t^*$. Power given precision, type I error, design, and residual variance is the lower tail probability $1 - \beta$ corresponding to \tilde{t}_{df} for df degrees of freedom. Power increases monotonically with \tilde{t}_{df} , so larger values of $|b_p|$ and smaller values of c_{pp} , $\hat{\sigma}_{y|x}^2$ and t^* increase power. In other words, more power is available for testing the null hypothesis if a) the minimum detectable size of the slope is large, b) a large sample is collected, c) data are collected over a wide range of the covariate, d) the covariate is independent of other covariates, e) residual variability is small, f) the alternative hypothesis is one-sided, and g) type I error is large.

These concepts and calculations are applicable to any b_p of interest and will be illustrated through several examples. All quantities except P -values were calculated in Quattro Pro 4.0 for DOS. However, Cox (1991) provided algorithms that permit spreadsheet calculation of P -values.

Example 1. Simple linear regression on equally-spaced covariate.

A single covariate with equally-spaced values is one of the simplest cases of linear regression. Suppose the following pilot data were collected on x_1 and y :

x_1	10.25	10.40	10.55	10.70	10.85	11.00	11.15	11.30
y	245	240	239	241	232	226	229	235

Note that each value of x_1 is separated by 0.15. If we subtract 10.10 from each value and divide the result by 0.15, then $x_{11}, x_{12}, \dots, x_{18}$ take on the integer values 1, 2, ..., 8. Calculations for the pilot data yield

$$C_0 = \begin{matrix} 0.60714 & -0.10714 \\ -0.10714 & 0.02381 \end{matrix}, \quad Z = \begin{matrix} 1887 \\ 8405 \end{matrix}, \quad B_0 = \begin{matrix} 245.143 \\ -2.05952 \end{matrix}$$

$df_0 = 6$, and $\hat{\sigma}_{y|x}^2 = 19.7877$. Note that $c_{11} = 12/(N^3 - N)$ which is always the case for simple linear regression when x is scaled into the integers $1, 2, \dots, N$. The estimated regression slope is -2.06 , and it applies to the scaled values of x_1 . Let us investigate several questions about power and precision for a series of possible designs.

Question 1A. What is the minimum detectable slope size for this design, given $\alpha = 0.05$, H_A : slope $\neq 0$, and power = 0.90?

From a t -table, values $t^* = t_6(1 - 0.05/2) = 2.447$ and $t_6(0.90) = 1.440$. For the pilot design, $\hat{\sigma}_b = (0.02381 \cdot 19.7877)^{1/2} = 0.68640$. From the formula above, $\tilde{b}_1 = 0.68640(2.447 + 1.440) = 2.67$. If data with similar variability were obtained from a future study replicated at the same 8 design points, chances are 90% of detecting a true slope ≥ 2.67 or ≤ -2.67 with no greater than a 5% chance of a type I error.

Question 1B. Is detection ability improved if H_A : slope < 0 ?

Now $t^* = t_6(1 - 0.05) = 1.943$, therefore, $\tilde{b}_1 = -2.32$ (negative, because the one-sided rejection region is negative). Because the range over which the null hypothesis may be rejected has been made less restrictive (≤ -2.32 rather than ≤ -2.67), detection ability for negative slopes has been improved.

Question 1C. What is the power for detecting a slope ≤ -1.5 ?

If the design and α remain the same, and the test remains one-sided, we calculate $\tilde{t}_6 = (|-1.5| / 0.68640) - 1.943 = 0.242$. A statistical package that provides P -values calculated 0.592 as the lower tail probability for this t -value (linear interpolation of values from a t -table provided 0.591). Thus, under this design, type I error rate, and level of variability, chances of detecting a slope of -1.5 or smaller are no greater than 60%.

Question 1D. What is the best approach to doubling sample size to increase power: selecting 8 additional design points, or doubling sampling effort at each existing design point?

We will calculate power for 2 designs under the same variance and hypothesis conditions for the preceding question. For each design, we will double the sample size, so $df = 16 - 2 = 14$ and $t^* = t_{14}(1 - 0.05) = 1.761$. For the first design, we will add the 8 integer-scaled design points 9, 10, ..., 16. Value c_{11} for this design becomes $c_{11} = 12/(16^3 - 16) = 2.9412 \cdot 10^{-3}$, and $\tilde{t}_{14} = 4.457$. Thus, power increases to > 0.999 . For the second design, we will also collect 16 total samples, but we will do this by collecting 2 samples at each original design point. In general,

we would recalculate C , but in this special case, c_{11} is just the pilot design value of c_{11} divided by 2. Thus, we have $c_{11} = 0.02381/2 = 0.01190$, and $\tilde{t}_{14} = 1.3295$. Power is then 0.898. Though alternate design 2 has less power than alternate design 1, the second design is not necessarily inferior to the first. Through replication at each design point, design 2 allows one to directly estimate $\sigma_{y|x}^2$ regardless of how well the linear model describes the pattern of the data; design 1 does not.

Example 2. Linear regression on 2 covariates.

Suppose now that y responds linearly to a pair of covariates, x_1 and x_2 , and that we wish to investigate questions of power and precision for b_1 . We will assume that x_2 takes on values not controlled or predicted by the observer, but that the mean and range of x_2 values can be anticipated. We will also assume that x_1 takes on values either controlled or predicted by the observer, and, to simplify matters, that the values are successive integers. The only pilot data available are a series of observations on y taken at a single combination of x_1 and x_2 settings. The estimate of variance yielded by the pilot study was $\hat{\sigma}_{y|x}^2 = 1.30$. Of course, we must assume that $\hat{\sigma}_{y|x}^2$ would be expected at all other combinations of x_1 and x_2 . Our objective is to detect a slope greater than 0.4 in magnitude and ≤ 0.05 probability of type I error. We will determine power for several designs.

Question 2A. Outcomes for x_2 may be uniformly distributed between 0 and 1. What is the power of a design with $N = 10$, $x_1 = 1, 2, \dots, 10$?

Here is a possible outcome for x_2 :

x_1	1	2	3	4	5	6	7	8	9	10
x_2	0.75	0.73	0.24	0.58	0.89	0.95	0.84	0.27	0.99	0.60

For this design, $c_{11} = 0.012248$, $\hat{\sigma}_b = (0.012248 \cdot 1.30)^{1/2} = 0.126184$, and $t^* = t_8(1 - 0.05/2) = 2.306$. Thus, $\tilde{t}_8 = (0.4 / 0.126184) - 2.306 = 0.864$, and estimated power is 0.794. Note that $12/(N^3 - N) = 0.012121$, which is relatively close but is not equal to c_{11} . When other covariates appear in the model, $c_{11} = 12/(N^3 - N)$ only when the correlation between x_1 and every other covariate is zero. Also note that the design for the pilot study in no way has to resemble the proposed regression design. The only requirement is that $\hat{\sigma}_{y|x}^2$ measures average response variability at fixed levels of X , regardless of whether one fixes X through experimental control or through analysis.

Question 2B. Suppose new random values are supplied for x_2 ; otherwise the design remains the same. For the same objectives, does power change?

You bet. Here is a new set of values for x_2 :

x_1	1	2	3	4	5	6	7	8	9	10
x_2	0.67	0.95	0.14	0.70	0.45	0.69	0.90	0.04	0.14	0.12

Now $c_{11} = 0.017222$, $\hat{\sigma}_b = 0.149628$, and $\tilde{t}_8 = 0.367$. Estimated power then is reduced to 0.638. What has happened is that the subtle change to x_2 essentially varied the predictive strength of x_1 through a change in the correlation structure of the design. Another set of values for x_2 could yield higher power for x_1 , up to a limit of 0.798 (when x_1 is uncorrelated with x_2). Although the uncertainty of the actual outcome of x_2 makes designing any study difficult, an idea of the likely range of power could be accomplished through simulation of several sets of x_2 . If the simulation could be automated to run hundreds of times, even more certainty about power could be achieved.

Question 2C. Suppose values drawn for x_2 correlate strongly with x_1 , either by chance or because x_2 and x_1 are causally related. What is the effect on power?

To consider this situation, we will order the values in the question above from smallest to largest:

x_1	1	2	3	4	5	6	7	8	9	10
x_2	0.04	0.12	0.14	0.14	0.45	0.67	0.69	0.70	0.90	0.95

The values we need are $c_{11} = 0.204254$, $\hat{\sigma}_b = 0.515296$, and $\tilde{t}_8 = -1.530$. Estimated power is now only 0.082. Through these scenarios, we see that power drops consistently with increasing strength of correlation between x_1 and x_2 :

	Correlation r			
	0.0 (Q. 2B)	0.101 (Q. 2A)	-0.544 (Q. 2B)	0.970 (Q. 2C)
power	0.798	0.794	0.638	0.082

In planning any design in which strong correlation between the covariate of interest and any other covariate is a possibility, it is important to distinguish whether the phenomenon is due to chance or to a causative or associative relationship. If chance is the principal reason, then the simulation exercise described above should provide an indication of the likelihood of drawing an unlucky sample. Otherwise, one must try to weigh the merit of keeping x_2 for total variance reduction against its cost in power for x_1 .

Example 3. Regression on 2 linear covariates and a dummy covariate.

This final example illustrates a more complex situation. Suppose the following pilot data were collected:

y	51	59	58	60	64	55	61	73	68	81
x_1	0	2	4	6	8	0	2	4	6	8
x_2	1.3	1.9	1.0	0.9	0.8	1.3	1.3	2.1	1.4	2.0
x_3	A	A	A	A	A	B	B	B	B	B

Variable x_3 is a dummy variable, and it represents a qualitative factor with discrete, usually nominal, levels. In this example, x_3 could represent sex, study area, or any other of a number of effects. Variable x_1 may be an experimentally-controlled variable; x_2 appears to take on random values. If the linear model for the pilot study is just the sum of the x effects, the X matrix would appear as follows:

1	0	1.3	1
1	2	1.9	1
1	4	1.0	1
1	6	0.9	1
1	8	0.8	1
1	0	1.3	-1
1	2	1.3	-1
1	4	2.1	-1
1	6	1.4	-1
1	8	2.0	-1

The fourth column is a typical way of expressing a dummy variable in a design matrix; more detail may be found in Rawlings (1988). For the questions below, we will let $\alpha = 0.05$ and let alternative hypotheses be two-sided.

Question 3A. What is the power for $|b_1| \geq 0.5$?

For this design, $N_0 = 10$, and $df_0 = 6$; therefore, $t^* = t_6(1 - 0.05/2) = 2.447$. We calculate C_0 and all other statistics as before to obtain $c_{11} = 0.012611$ and $\hat{\sigma}_{y|x}^2 = 1.913336$. We calculate $\hat{\sigma}_b = 0.155335$ and $\tilde{t}_6 = 0.772$. Thus, estimated power = 0.765.

Question 3B. What is the smallest detectable effect difference for b_3 with power = 0.90?

We need to determine $t_6(0.90) = 1.440$ and $\hat{\sigma}_b = (1.913336 \cdot c_{33})^{1/2} = 0.507262$. Then $b_3 = 1.97$, the magnitude of the smallest detectable effect size for the dummy effect b_3 .

If you need more help

O'Brien and Muller (1992) present a more general framework for power estimation in fixed-effects linear models, but they also point out that the methods are still useful when effects yield normally distributed values. In the setting of trend estimation, specifically for the case of heterogeneous response variance, refer to papers by Gerrodette (1987) and Link and Hatfield (1990). For a general introduction to linear models, see Rawlings (1988).

I thank W. L. Kendall for comments and corrections to this document.

References

- Cox, M. A. A. 1991. The implementation of functions to evaluate percentage points of the normal and Student's *t* distributions on a spreadsheet. *The Statistician* 40:87-94.
- Gerrodette, T. 1987. A power analysis for detecting trends. *Ecology* 68:1364-1372.
- Link, W. A., and J. S. Hatfield. 1990. Power calculations and model selection for trend analysis: a comment. *Ecology* 71:1217-1220.
- O'Brien, R. G., and K. E. Muller. 1992. Unified power analysis for *t* tests through multivariate hypotheses. Ch. 8 *in* L. K. Edwards, ed. *Applied analysis of variance in the behavioral sciences*. Marcel Dekker, New York.
- Rawlings, J. O. 1988. *Applied regression analysis*. Wadsworth and Brooks/Cole, Pacific Grove, Calif. 553pp.