Sample Size and Power for Regression

Statistical power for regression analysis is the probability of a significant finding (i.e., a relationship different from 0 typically) when in the population there is a significant relationship. By convention, .80, which represents an expectation that 80% of random samples from the same population would find significance if there is a relationship in the population (i.e., H_1 is true), is often used as a minimum acceptable level of power when estimating the sample size needed in a planned study. In general, power is dependent on the significance criteria used (nearly always α = .05), sample size, and effect size. Sufficient power is not only critical for ensuring that we do not miss important significant effects, but it is also important because power may play a major role in failures to replicate findings and even in a greater chance that a given finding may be a false positive (Fraley & Vazire, 2014).

In regression analysis, we may be interested in the significance of all of the predictors together, which is the F test of significance of R^2, or the significance of the partial regression coefficient, B. There are a number of "rules of thumb" that have been proposed for what should be an adequate sample size for regression analysis (Maxwell, 2000). Sometimes these are based on a ratio of the number of cases to predictors or other conventions. These suggestions are nearly always overly general.

Power to Detect a Significant R^2

The effect size estimate (which is sometimes abbreviated ES) for R^2 is Cohen's f^2 which is a simple ratio of the proportion of variance accounted for relative to the proportion of variance unaccounted for.

$$f^2 = \frac{R^2}{1 - R^2}$$

Cohen (1988) suggested an f^2 value of .02, .15, and .35 be used for small, medium, and large effect sizes, respectively. Because we know that R^2 depends on n (sample size) and the k (number of predictors), it is easy to see what factors contribute to effect size in addition to the correlations of the predictors with the outcome. The f^2 can be computed using the same equation for incremental R^2 (or R^2-change). Taking a look at the F-statistic equation can be instructive about how n, k and R^2 affect power.

$$F = \frac{R^2 / k}{1 - R^2 / (n - k - 1)} = \frac{R^2 (n - k - 1)}{(1 - R^2) k}$$

First, from the numerator of the form of the equation on the farthest right above, it is clear that increasing sample size, n, would increase F and therefore power. Second, the denominator from that same equation also indicates that adding predictors, k, would decrease F. The numerator of that same form of the equation, which contains $df = n - k - 1$, also suggests there is an additional small penalty for having more predictors, k.

The quantity L, the non-centrality parameter,[1] is a simple computation from f^2 and the degrees of freedom.

$$L = f^2 (n - k - 1)$$

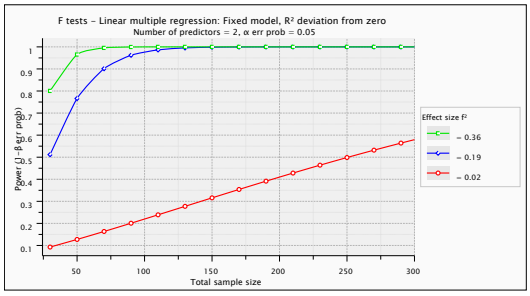
With a little algebra we can find an equation that would help us estimate the sample size, n*, given some value of f^2 and the number of predictors.

$$n^* = \frac{L}{f^2} + k + 1$$

[1] L is referred to as λ ("lambda") by Cohen (1988) and many other sources. See my "Power" handout from the univariate quant class http://web.pdx.edu/~newsomj/uvclass/.

Or, we can use L to estimate power with Appendix Table E.2 from the text (Cohen, Cohen, West, & Aiken, 2003). Effect sizes can be taken from relevant literature or it is often convenient to use some range of effect sizes (such as Cohen's conventional values for small, medium, and large effects).

Power analyses and plots can also be obtained using a computer program such as G*Power,[2] a freeware program. I used G*Power to create the plot below showing the relationship between sample size and power for a range of small to large sample sizes for R^2 (assuming α = .05, two tailed).



The plot suggests we need fewer than about 25 cases to detect a large effect, about 60-70 or so to detect a medium effect size (my f^2 of .19 is a little higher than Cohen's .15 medium, because the program constrains the values), and something over 300 cases for a small effect.

Power to Detect a Significant Regression Coefficient

For simple regression, the test of the regression coefficient is the same as the test of r, so one way to estimate power in this case is to use a power table for r, such as Appendix Table F.2 of your text (Cohen et al., 2003) where you will find a power table for r values. This can also serve as a rough guide for statistical power and sample size requirements for β for any model.

Because β^2 is not really a very exact estimate of unique variance accounted for by a variable, it is better to use R^2-change for a single variable added to the model or equivalently sr^2. The equations for f^2 and L are readily adaptable for that. Cohen (1988) suggested (where R^2 on the denominator refers to the total R^2).

$$f^2 = \frac{sr^2}{1 - R^2}$$

It is worth looking at the standard error equation for a partial regression coefficient (one form of it) to get a sense of how various factors affect power. Remember that because $t = B/SE$, a smaller standard error leads to greater probability of finding significance.

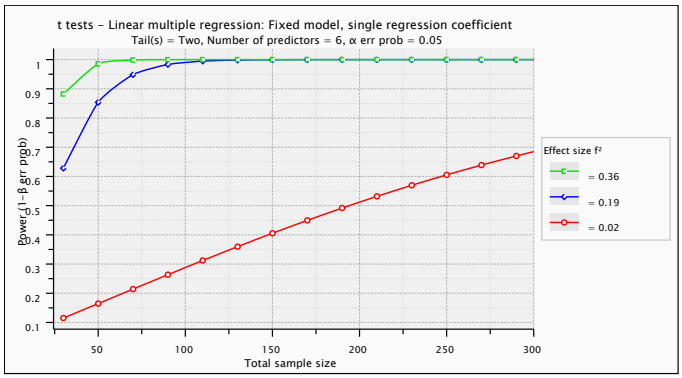
$$SE_B = \frac{sd_Y}{sd_X}\sqrt{\frac{1 - R^2}{n - k - 1}}\sqrt{\frac{1}{1 - R_i^2}}$$

The R_i^2 on the denominator in the third quantity on the right is the multiple regression of the predictor as predicted by all of the other predictors, and so represents the amount of correlation (collinearity) among the predictors (as used in VIF and tolerance). And, as the equation suggests, a larger sd_X, a smaller sd_Y,

[2] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

a larger total R^2, and smaller intercorrelations among the predictors (R_i^2) all will also lead to smaller standard errors and greater power.

I used G*Power to plot power by sample size for a range of effect sizes (again α = .05, two-tailed) for a partial regression coefficient. To gauge what we might expect for β values, the small, medium, and large f^2 values that I used below (.02, .19, .36), corresponding to β values of about .14, .36, and .51.



From the figure, it looks like n of about >300, 40, about 25 are needed to power of .80 with small, medium, and large effect sizes.

Recent versions of G*Power have added power for tests of regression coefficients. For simple regressions, standardized coefficient values can be used for effect sizes (perhaps using the correlation values of .1, .3, and .5 as small, medium, and large) as long as you make sure the standard deviations of x and y are set to 1. Power for regression coefficients in multiple regressions also can be estimated under “Multiple Regression; Fixed Model; Single Regression Coefficient.” This option can take into account the number of predictors but does not incorporate information about the correlation among predictors or the total R^2 value.

Power of Tests of Interactions and Indirect Effects

Both interaction tests and tests of indirect effects are notoriously lacking in power. For interactions, the reduced power comes from the fact that a product variable has a reliability equal to the product of the reliabilities of the two variables (Aiken & West, 1991) and the tendency for the product variable to have a nonnormal distribution (McClelland & Judd, 1993; O'Connor, 2006; Shieh, 2009). The effect size for the interaction also depends on the shape of the interaction, with magnitude (ordinal) patterns having smaller effect sizes than cross-over (disordinal) patterns (Champoux & Peters, 1987; Blake & Gangstead, 2020). All things combined, the power to test the interaction can be considerably lower compared to the power to detect significance for "main effects."

Indirect effect tests to investigate mediation also suffer from lower power (e.g., MacKinnon et al., 2002). One of the reasons is that the sampling distribution for the indirect effect is nonnormal. This has been addressed to some extent with improved tests, using bootstrapping, for example. Power to detect significance for the indirect depends on the value of the a (X predicting M) and b (M predicting Y) effects and has some complex results. For small b effects, moderately-sized a coefficients, more than smaller- and larger-sized a coefficients, may lead to a stagnation of power, in which power is worse than expected given the size of the coefficient (Fritz, Taylor, & MacKinnon, 2012). Indirect coefficients also pose some problems for how to conceptualize an appropriate effect size. Calculating the proportion of the total effect due to the indirect effect (or “proportion mediated”) has been proposed (Preacher & Kelley, 2011), but it

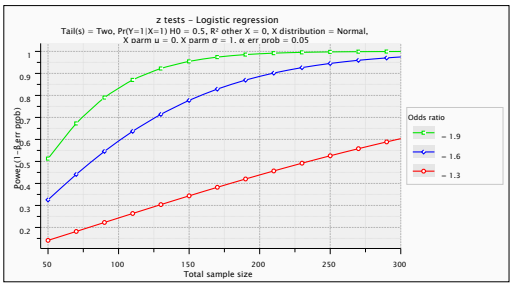
has some potential pitfalls (MacKinnon, Kisbu-Sakarya, & Gottschall, 2013). Alternatives include standardized or partially standardized values (Fairchild, MacKinnon, Taborga, & Taylor, 2009). Based on standardized indirect coefficient values, Kenny suggests small, medium, and large effects size for indirect effects should be .01, .09, and .25, respectively (http://davidakenny.net/cm/mediate.htm#DI). Kenny has created an online calculator for estimating power with indirect effect tests, https://davidakenny.shinyapps.io/MedPower/.

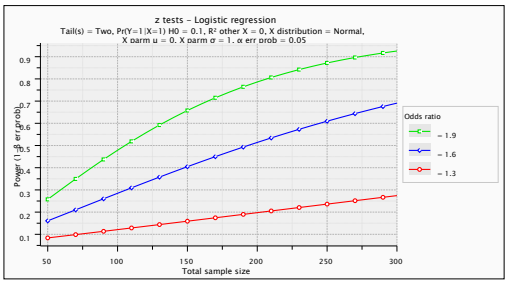
Logistic Regression

Power analysis and sample size recommendations for logistic regression are more complicated by the fact that there is not really a clearly accepted effect size measure that works with all applications, given that there is no well-defined R^2 and odds ratios are scale dependent in the case of a continuous predictor. No doubt that researchers should plan for larger sample sizes—some have suggested two or three times larger for logistic regression than for OLS (Taylor, West, & Aiken, 2006). In addition, there are a number of precautions about significance testing for small n, rare events, and sparse data.

For sufficient power, a number of "rules of thumb" have been suggested, but are likely to be oversimplifications—power analysis is better able to take more specific circumstances into account. Many authors have recommended a 10:1 ratio of cases to predictors. Based on simulations, Peduzzi and colleagues (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996) refine the 10:1 recommendation, stating that ten times the number of predictors, k, should take into account the proportion, p, of successes, $n = 10k/p$. The proportion of successes should be formulated as a proportion between 0 and .5, so that when the proportion is close to .5, fewer cases are needed (always using a minimum of 100). When modeling rare events, one should consider the absolute frequency of the event rather than the proportion, according to Allison (2012). If the overall probability of disease is .01 (1 in a 1000) for example, then one may need a total of 20,000 cases for sufficient power, because the number of events is 200. Recall that the Wald test can behave erratically with smaller sample sizes (e.g., Hauck & Donner, 1977), so, for smaller samples, it is wise to also examine likelihood ratio (or perhaps score) tests for individual predictors. Finally, Hsieh (1989) published tables of required sample sizes for various odds ratios × event proportion which are widely cited. These tables can be difficult to use because all of the values are based on one-tailed tests, a more liberal standard (equal to $\alpha = .10$ two-tailed). To give a very general idea of what sample size might be required for the usual power = .8 with a two-tailed test using the Hsieh tables, consider two fairly arbitrary examples from the table using a more conservative power value than usual (.9 instead of the usual .8): for an odds ratio of 1.5 when the outcome $\pi = .5$, 225 cases are needed, whereas for an odds ratio of 1.5 and $\pi = .1$, 628 cases are needed. Power and adequate sample sizes for logistic regression is a fairly complex issue, where sparseness and the size of odds ratios have some biasing effects on fit and odds ratios (see my handout "Sample Size and Estimation Problems with Logistic Regression" from my categorical data analysis class for a brief synopsis and further references, http://web.pdx.edu/~newsomj/cdaclass/)

I used G*Power to illustrate the effect of sample size on power for several odds ratio values to get some idea of power (with $\alpha = .05$). These results assume a binary predictor in a simple regression. On the left the H_0 probability for $Y = 1$ is $p = .5$ when power is likely to be greatest and on the right the H_0 probability for $Y = 1$ is $p = .1$ when power is considerably lower.





References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences, second edition*. Routledge.

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage.

Allison, P. (2012). *Logistic Regression for Rare Events*, website post at http://statisticalhorizons.com/logistic-regression-for-rare-events.

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin, 46*(12), 1702-1711.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size and power in moderated regression analysis. *Journal of Occupational Psychology, 60*(3), 243-255.

Cheung, M. W.-L. (2009). Comparison of methods for constructing confidence intervals of standardized indirect effects. *Behavior Research Methods, 41*, 425– 438. doi:10.3758/BRM.41.2.425

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences, 2nd edition*. New York: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple correlation/regression analysis for the behavioral sciences.* New York: Erlbaum.

Fairchild, A. J., MacKinnon, D. P., Taborga, M., & Taylor, A. B. (2009). R2 Effect-size Measures for Mediation Analysis. *Behavioral Research Methods, 41*, 486–498.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019. doi:10.1371/journal.pone.0109019.

Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research, 47*(1), 61-87.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association, 72*(360a), 851-853.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. Statistics in medicine, 8(7), 795-802.

MacKinnon D.P., Lockwood C.M., Hoffman J.M., West S.G., Sheets V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104

MacKinnon, D.,P, Kisbu-Sakarya, Y., & Gottschall, A.C. (2013). In Little, T. D. (Ed.). *The Oxford handbook of quantitative methods, volume 1: Foundations* (pp. 338-360). Oxford University Press.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*(4), 434-458.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. Psychological bulletin, 114(2), 376-390.

O'Connor, B. P. (2006). Programs for problems created by continuous variable distributions in moderated multiple regression. *Organizational Research Methods, 9*, 554-567.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, *49*(12), 1373-1379.

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables power and sample size considerations. *Organizational Research Methods, 12*(3), 510-528.

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. Educational and psychological measurement, 66(2), 228-239.