Sample Size Issues for Categorical Analyses and Logistic Regression

Chi-square, Likelihood Ratio, and Loglinear Models

The effect size (Cohen's w) for the Pearson chi-square test illustrates that, in addition to overall sample size, the power of the chi-square test will be a function of the discrepancy between the observed proportion and the expected proportions (e.g., $w = \sum\left[\left(p_{Ei} - p_{O_i}\right)^2 / p_{Oi}\right]$. Power analyses, of course, are the best way to estimate the power of a given study or design, but Cohen (1992) provides a range of sample sizes needed for adequate power of the chi-square test. For example, for a 1-*df* test, such as a 2 × 2 chi-square analysis, Cohen indicates that small, medium, and large effect sizes would require sample sizes of 785, 87, and 26 cases, respectively. As discussed earlier in the course, the standard Pearson chi-square test with no adjustments appears to perform accurately with small samples (e.g., n = 25; e.g., Koehler, 1986). This appears to be the case for loglinear modeling approach as well, which makes sense given the equivalences among these approaches. The oft-cited Cochran rule of a minimum expected count of 5 appears to be overly cautious (Haberman, 1977) and the Fisher's exact test and the Yates continuity correction are overly conservative. The rule may be better stated as requiring expected counts above 1 but this is an oversimplification. The performance of these tests with small n not only depends on the the presence of cells with low expected frequencies but also in large disparities among the cells in the expected frequencies (Koehler, 1986).

Logistic Regression

There are two issues that researchers should be concerned with when considering sample size for a logistic regression. One concerns statistical power and the other concerns bias and trustworthiness of standard errors and model fit tests.

*Sample Size***.** The first issue concerns understanding the sample size that is required for attaining adequate statistical power. As with any other statistical analysis, power, the probability of finding significance when the alternative hypothesis is true in the population, depends on sample size, variance of the independent and dependent variable, and effect size (e.g., odds ratio, proportional difference), among a few other things (e.g., number of predictors, the magnitude of the correlations among them, alpha level). Because all of these factors vary from sample to sample and model to model, it is difficult to give a simple answer to the question "How many cases do I need?" When planning a study, the best thing to do is to conduct a power analysis in which you can specify the factors specific to your study and analysis. Power analyses can be conducted for logistic regression using dedicated software, free (e.g., G*power; see Faul, Erdfelder, Buchner, & Lang, 2009) or otherwise (SPSS Sample Power; NQuery Advisor; PASS), or by setting up a simulation routine in standard statistical software (e.g., Aberson, 2019; and see Online Power Analysis Resources Below). Bush (2015) reviews power and sample size estimation methods.

With this in mind, it can be useful to know about some general guidelines and conventional recommendations. These guidelines are the most often cited but should not be taken as universal laws of nature—they are simply some general suggestions to consider, are not precise, and they do not apply in all circumstances. Based on his experience, Long (1997) suggests that maximum likelihood estimation including logistic regression with less 100 cases is "risky," that 500 cases is generally "adequate," and there should be at least 10 cases per predictor. 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 for example, then a total of 20,000 cases may be sufficient, 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 α = .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 π = .5, 225 cases are needed, whereas for an odds ratio of 1.5 and π = .1, 628 cases are needed.

The other sample size issue to consider involves the validity of coefficient and odds ratio estimates, standard errors, and model fit statistics for small sample sizes or sparse data. Maximum likelihood estimation is known to have a "small sample bias" and produces odds ratio that are too large for small samples (Nemes, Jonasson, Genell, & Steineck, 2009). Odds ratios tend to be farther away from 1.0 (higher for positive relationships, lower for negative relationship) for smaller samples. Roughly speaking, based on their Figure 3, the bias appears to be about 10-15% for the log odds ratio when n = 100, and nearly entirely disappears as n = 1000. Smaller samples can be expected to have a larger bias. With 100 cases, this degree of bias is not ideal but also may not be terrible—if the true odds ratio is 2.0, a sample estimate with n = 100 might be 2.1 to 2.2 (note exponential conversion needed).

Standard errors and significance tests require caution for smaller sample sizes, say less than 100 under ideal circumstances. The Wald test and likelihood ratio test of individual parameters (comparing nested models with and without one of the predictors) test the same hypothesis and are asymptotically equivalent, but the Wald test performs much more poorly for small samples (e.g., Hauck & Donner, 1977; Vaeth, 1985). With about 100 cases, there is very good agreement between the two tests, but with fewer, the Wald test has wider intervals, is more likely to include the null value when the null is false (Type II error), and has an inappropriately symmetric distribution when the alternative hypothesis is true. Sample sizes of 100 also appear to be adequate for accurate tests with ordered logistic models (Holtbrugge & Schumacher, 1991).

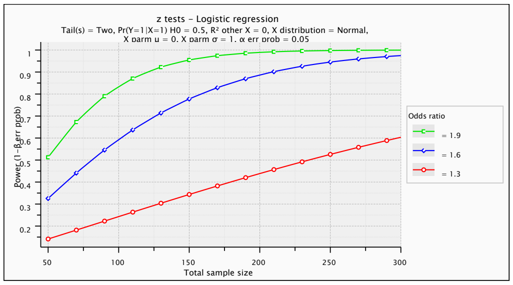
Caution is warranted in interpreting model fit statistics when data are sparse. Sparseness occurs when the number of expected cases for particular pattern of X values is small, which becomes more likely with a small sample size. In a simple example with two binary predictors, a cell count of success on Y close to zero in a 2 × 2 table formed from the two predictors less than 5 can be problematic for fit as measured by the likelihood ratio test or the Pearson chi-squared test (McCullagh,1985). For many predictors and particularly when they have skewed distributions, this can become a more likely problem if the sample size is small. This is just another reason for keeping a minimal sample size of 100 or more. There are a variety of alternative tests that have been suggested (e.g., Copas, 1989; Farrington, 1996; Stukel, 1988), which some simulation work has suggested can perform better than the often-reported Hosmer and Lemeshow test (e.g., Katsaragakis et al., 2005), and penalized likelihood estimation (Firth, 1993) is an alternative estimation approach that seems to perform better than logistic when data are sparse (Heinze & Schemper, 2002).

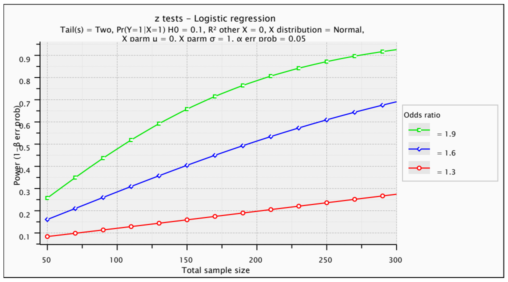
Estimation Problems. For most data sets and most situations, logistic regression models have no estimation difficulties. One particular problem that can arise is *separation* (Albert and Anderson 1984). Separation occurs when the predictor or set of predictors has a perfect relationship to Y. It is an extreme case of the sparseness issue mentioned above, and the term *quasi-complete separate* is used when the relationship is very high but less than perfect. Ironically, the logistic regression coefficient can be 0 sometimes when this occurs. The other possibility is that it is equal to infinity. Consider a simple logistic regression with a binary predictor. The coefficient can be expressed in terms of the frequencies for a 2 × 2 table as

$$\beta = \ln\left(\frac{n_{11}n_{22}}{n_{21}n_{22}}\right)$$

If any of these cells is equal to 0, the coefficient can be equal to 0 (if occurring in the numerator) or infinity (if occurring in the denominator). Wald tests will not be printed or are problematic when separation or quasi-separation occur. The likelihood ratio tests will be ok, and can be used to test individual predictors when separation issues arise. Software may or may not print informative messages when there are separation issues, so one needs to be on the lookout and careful visual inspection of diagnostics, such a residual plots or fit and parameter change statistics are valuable initial steps that should not be skipped. Penalized likelihood (Firth, 1993) is a good alternative when there are separation problems. Allison (2008) has an excellent brief discussion of separation and its solutions.

Power Analysis Illustration. I used G*Power[1] to illustrate the effect of sample size on power for several odds ratio values to get some idea of power (with α = .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.





Survival Analysis

Survival analysis (here, I am assuming mostly Cox regression, which overall has equal or better power than other survival models; Li et al., 1996) has similar considerations to logistic regression plus a few additional considerations. Generally, power of survival models is a function of the number of events experienced rather than the total sample size, and power declines with fewer events (Allison, 2010; Li et al., 1991). Some authors therefore recommend planning for a length of the study to have approximately half of the sample experience the event during the observation period (e.g., Singer & Willett, 1991). The number of predictor variables relative to the number of events also seems to be important, in which a larger number of predictors leads to coefficient bias in particularly small sample sizes (e.g., ratio of events to covariates is less than 10), is detrimental to power to detect significance, convergence, and accurate confidence interval coverage (Peduzzi et al., 1995). For Cox models, the Wald and likelihood ratio tests generally perform similarly in terms of power and both outperform the score test (Li et al., 1996).

Latent Class Analysis

Nylund and colleagues (2007) show that the standard likelihood ratio tests should not be used to statistically compare the fit of the number of classes and that the bootstrap likelihood ratio p-values perform best across conditions for this purpose. Among several information criteria evaluated, the Bayesian Information criteria is preferable to the other examined (e.g., AIC, CAIC) although has weaknesses for few classes of unequal sizes. These results generally held for their minimum sample size condition of 200 cases. Power for Wald tests for response probability parameters (the relationship between the latent class factor and the indicators) appears to depend on several factors (Gudicha et al.,

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

2016), with greater power resulting from larger sample size, better class separation (entropy), a smaller number of classes, and a larger number of indicators. From their simulation results, Gudicha and colleagues concluded that a minimum of 500 cases are needed for small sample sizes and weak class separation, a minimum of 100 cases for medium effect sizes and moderate class separation, and only 75 cases are needed for large effect sizes with high class separation. In another simulation paper, Gudicha and colleagues (2017) examined the tests for the association between covariates and a latent class factor, also finding that the power of these tests depend on sample size, the magnitude of the response probabilities, and class separation. The Wald and likelihood ratio tests of these parameters had acceptable and similar type I error rates with large samples and moderate or high class separation, but both had unacceptable type I error rates for smaller samples and weaker class separation (minimum n of 200 for strong separation, 500 for moderate separation, and 1000 for weak separation). These papers also describe computation of noncentrality parameters and power analysis approaches. The Latent Gold package has some built in power analysis functions and Monte Carlo simulation can be used to determine sample size in Mplus.

References and Further Reading

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences, second edition*. New York: Routledge.
Albert, A., & J. A. Anderson (1984) On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika, 71,* 1-10.
Allison, P. D. (2008, March). Convergence failures in logistic regression. In *SAS Global Forum* (Vol. 360, pp. 1-11).
Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. SAS Institute.
Allison, P. (2012). *Logistic Regression for Rare Events*, website post at http://statisticalhorizons.com/logistic-regression-for-rare-events.
Allison, P. D. (2014). Measures of fit for logistic regression. *SAS Global Forum, Washington, DC*.
Bush, S. (2015). Sample size determination for logistic regression: A simulation study. *Communications in Statistics-Simulation and Computation*, *44*(2), 360-373.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155.
Copas, J.B. (1989) "Unweighted sum of squares test for proportions." Applied Statistics 38:71 –80.
Farrington, C. P. (1996) On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society, Series B* **58**: 344–366.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.
Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika, 80,* 27-38.
Gudicha, D. W., Tekle, F. B., & Vermunt, J. K. (2016). Power and sample size computation for Wald tests in latent class models. *Journal of Classification, 33*(1), 30-51.
Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2017). Statistical power of likelihood ratio and Wald tests in latent class models with covariates. Behavior research methods, 49(5), 1824-1837.
Haberman, S.J. (1977). Log-Linear Models and Frequency Tables With Small Ex- pected Cell Counts. *The Annals of Statistics, 5*, 1148-1169.
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.
Heinze, G., & Schemper, M. (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine, 21*:,2409-2419.
Holtbrügge, W., & Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 40(2), 249-259.
Hsieh, F. Y. (1989). Sample size tables for logistic regression. Statistics in medicine, 8(7), 795-802.
Katsaragakis, S., Koukouvinos, C., Stylianou, S., & Theodoraki, E. M. (2005). Comparison of statistical tests in logistic regression: The case of hypernatreamia. *Journal of Modern Applied Statistical Methods, 4*(2), 16.
Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association, 81*(394), 483-493.
Li, Y. H., Klein, J. P., & Moeschberger, M. L. (1996). Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. *Computational Statistics & Data Analysis, 22(*2), 177-192.
McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistics in linear exponential family models. *International Statistical Review 53*, 61–67.
Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology, 9*, 56-60.
Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology, 48(*12), 1503-1510.
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.
Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. psychological Bulletin, 110(2), 268.
Stukel, T. A. (1988) Generalized logistic models. *Journal of the American Statistical Association* 83: 426–431.
Vaeth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review/Revue Internationale de Statistique*, 199-214.

Online Power Analysis Resources

Power/Sample Size Calculation for Logistic Regression with Binary Covariate(shttps://www.dartmouth.edu/~eugened/power-samplesize.php
SOCR Java Applets
http://www.socr.ucla.edu/htmls/ana/
R Programs

https://rpubs.com/candrea/ssizelogreg
https://www.r-bloggers.com/logistic-regression-simulation-for-a-power-calculation/