Evaluating the Accuracy of Approximate Power and Sample Size Calculations for Logistic Regression

Yezi Yang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2019

Committee:
Kenneth M. Rice
Barbara McKnight

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Evaluating the Accuracy of Approximate Power and Sample Size Calculations for Logistic Regression

Yezi Yang

Chair of the Supervisory Committee:
Kenneth M. Rice
Department of Biostatistics

Abstract

This master's thesis evaluates and implements power, sample size and effect size calculations for logistic regression. The earlier sections set up an ordinary logistic regression model, review the current approaches including those of Whittemore, Hsieh and Schoenfeld & Borenstein, and illustrate comparisons of the existing approaches. Schoenfeld & Borenstein's method exhibits general superiority and, with slight modifications, is implemented in a Shiny web application and an R package. We give examples to demonstrate its use, and make recommendations about when its results can be considered accurate enough for applications.

KEYWORDS: Logistic Regression, Power and Sample Size Calculation, Wald Test, Shiny App, R Package

Contents

1 Introduction

Logistic regression is a widely-used method across the field of biostatistics, with applications in clinical trials and observational studies. The odds ratios it estimates have been suggested as the outright "gold standard" measure of association between binary outcomes and covariates [Senn, 1999], and furthermore have useful properties for understanding prospective implications of covariates from retrospective studies [Prentice and Pyke, 1979]. More pragmatically, the odds ratio is a commonly used measure of association, and investigators need to be able to plan and interpret studies of binary outcomes using odds ratios.

Given the widespread use of logistic regression and the importance of understanding power in applied analysis [Jones et al., 2003], it is surprising that power calculations (outside of situations with a single binary covariate) have limited availability. Specifically, in standard introductory textbooks sample size calculations for multiple logistic regression are not discussed, and the reader is left to seek advice from higher-level texts, or forced to use approximate methods, the relevance of which is not obvious when power or sample size is sought for a logistic regression that adjusts for covariates. Furthermore, the accuracy of the various known approximate methods is not well-understood, and no rules-of-thumb or guidelines are available to indicate when they work well and when they do not. Finally, there is no free and open-source software that offers analytic calculations of power and sample size in general logistic regression settings—outside of writing simulation-based code, which for many investigators is not realistic, being too complex and too slow. Current available software is mostly proprietary and provides limited routines for approximate power and sample size calculation (for example PASS [PASS, 2019], G*Power [Faul et al., 2009] and the `powerlog` routines [Peng et al., 2012] in Stata and SAS). Our closest peer is the R package `powerMediation` that provides power and sample size estimates of a mediator or adjustment variable based on Vittinghoff et al. [2009],

which uses numeric evaluation that can be slower and less scalable than analytic approaches.

In this thesis we aim to address these issues. In the rest of Section 1 we review the currently-available approximate methods, and also illustrate their absence from the introductory literature. In Section 2 we implement the various approximations for 'simple' logistic regression settings, i.e. those with a single covariate, exploring their accuracy relative to exact-but-slow simulation, and the characteristics to which that accuracy might be sensitive. Section 3 extends this exploration to logistic regressions in which we adjust for covariates. In Section 4 we suggest an extension of an existing method where covariate distributions are defined empirically, from a single dataset, and in Section 5 we illustrate our R package and Shiny app that implement power and sample size calculations in user-friendly ways. We conclude with a short discussion in Section 6

1.1 Logistic Regression

Logistic regression, at its most basic, fits a line or surface of the form

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \tag{1}$$

for real-valued coefficients $\beta_0, \beta_1, \ldots, \beta_k$, where *logit* is the logistic function $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, the inverse of the function $\text{expit}(x) = \frac{e^x}{1+e^x}$. Logistic regression fits this line/surface by solving a series of $k+1$ equations

$$\begin{aligned} \sum_{i=1}^{n} Y_i - \text{expit}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k) &= 0 \\ \text{and } \sum_{i=1}^{n} X_{ij}(Y_i - \text{expit}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)) &= 0, \text{for each } j = 1, \ldots k, \end{aligned}$$

where for each of n observations, $Y_i \in \{0, 1\}$ denotes a binary outcome of interest, X_{ij} denotes the value of covariate j.

Typically we assume that the mean response for observations with independent variables $X_1, ...X_k$ follows the form given in Equation (1), so logistic regression estimates the coefficients

that govern the true probability of being a case ($Y_i = 1$), and in the fitted line/surface is for p denotes the probability of being a case, at specific covariate values. Under this assumption, the coefficients have particularly useful interpretations. For example, the assumption legitimizes interpreting β_1 as the log odds ratio, per unit difference in X_1, among observations that have the same values of $X_2, X_3, \ldots, X_k$. This interpretation is helpful for causal inference, when we aim to make statements about the impact on outcome Y of changing a covariate X_1 without changing all other factors.

Whether regarding coefficient β_j as a causal effect or more simply as a measure of association, tests of null hypotheses $H_0 : \beta_j = 0$ are widely-used. Corresponding power calculations can inform investigators planning a study what chance their design would give to reject that null hypothesis at some pre-specified level α, usually $\alpha = 0.05$. Conversely, for a specific level of desired power (usually 80%) and α, the calculation can be inverted, to give the minimum sample size at which that power is achieved.

1.2 Approaches to Power/Sample Size Calculations

Most common approaches to power/sample size calculations use either simulation-based or analytic approximations. Here we review them briefly, focusing on the evolution of three analytic approaches, before summarizing the paucity of their presentation in popular textbooks.

We note here that our primary focus is on power/sample size for analysis using Wald tests, although some of the techniques discussed could be used to evaluate power for Score and Likelihood Ratio Tests. These are, of course, asymptotically equivalent and so in large samples the power from Wald tests alone should suffice in practice, as long as log odds ratios of interest are not too far from zero. We also focus primarily on situations where large-sample approximations work at least reasonably well, and assume that a standard Wald-test analysis will be performed and used. For both reasons we will not address difficulties [Morris et al.,

2019] in making statements about power when Type I error rates are not well-controlled. We also will not address non-monotonicity of the Wald test statistic near the boundary of the parameter space, that can occur when extremely large effects are present [Hauck and Donner, 1977].

1.2.1 Simulation

Given modern computing capacity, simple simulation (also known as the 'Monte Carlo method') is an attractive method to calculate power or sample size. It requires specification of the study design, and the effects (the β_j) that determine the distribution of the outcomes for different observations. Given these, one simply simulates a large number of studies, performs logistic regression on each of them, using it to implement Wald tests of the null hypothesis that $\beta_j = 0$ for the covariate of interest. From this set of simulations, the proportion of datasets for which Wald tests gave significant results (i.e. p-value$<\alpha$) is an approximation of the power; finding the minimum n at which the approximate power exceeds the desired level gives an approximate sample size calculation.

This method is extremely versatile. For example, it requires only trivial modification to give sample sizes for Score and Likelihood Ratio Tests. However, while its basic implementation is easily coded, particularly for small sample sizes or extreme data-generating mechanisms, it may require coding that manages the "edge case" of complete separation (see e.g. Kleinbaum and Klein [2010, Pg 358]) where no finite estimate is returned and standard tests are not calculated and/or give statistically unreliable results. While the impact of such edge cases on power is usually very minor, for completeness the rate at which they occurred, over the simulations, should also be reported, and this may be cumbersome for investigators.

As performing logistic regression with continuous variables requires iterative calculations, the simulation approach may also be slow, particularly if the user wants to plot a smooth

power curve as an aid to interpolating power at different effect sizes.

1.2.2 Analytic approximations

All analytic approximations of power/sample size computation for Wald tests follow the same basic pattern. The Wald test statistic is known to be asymptotically distributed as a non-central chi-squared distribution, where the non-centrality parameter depends on true parameters of the underlying distribution, sample size, and the variance of the parameter estimates. Given the approximate non-centrality parameter, the power can be evaluated as a simple tail area, and inverting the same calculation gives the corresponding sample size calculation. (See e.g. Hanley and Moodie [2011] for a general exposition, including logistic regression but also many other methods.) The variance term above is obtained from the Fisher information per unit observation, and effective approximation of this term is where methods research has focused.

Expressing this formally, to test the null hypothesis $H_0 : \mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = 0$ versus $H_1 : \mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \neq 0$ for coefficients specified in model 1, the Wald statistic T is calculated as

$$T = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \mathbf{A}' (\mathbf{A} \hat{\mathbf{I}}^{-1} \mathbf{A}')^{-1} \mathbf{A} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

with $\boldsymbol{\beta}_a$ as $\boldsymbol{\beta}$ under the alternative hypothesis, $\mathbf{A}$ is a full row rank matrix, and $\hat{\mathbf{I}}$ is the empirical information matrix evaluated at the MLE of $\boldsymbol{\beta}$.

Asymptotically, T has a non-central χ^2 distribution with non-centrality parameter

$$\delta = (\boldsymbol{\beta}_a - \boldsymbol{\beta}_0)' \mathbf{A}' (\mathbf{A} \mathbf{I}^{-1} \mathbf{A}')^{-1} \mathbf{A} (\boldsymbol{\beta}_a - \boldsymbol{\beta}_0), \tag{2}$$

where Fisher information matrix $\mathbf{I}$ is defined below. It can be seen as a squared measure comparing 'signal' $\mathbf{A}(\boldsymbol{\beta}_a - \boldsymbol{\beta}_0)$ to corresponding 'noise' $\mathbf{A}\mathbf{I}^{-1}\mathbf{A}'$, .

For logistic regression, the full Fisher information is most straightforwardly stated when we assume that the covariates $\mathbf{X}$ are random, i.e. that the observations $\{Y, \mathbf{X}\}$ are a sample

of n independent identically-distributed vectors. The Fisher information is then

$$\mathbf{I} = n\mathbb{E}[\,\mathbf{X}\mathbf{X}'f(\boldsymbol{\beta}'\mathbf{X})\,], \tag{3}$$

where for convenience and following the literature, we have defined

$$f(u) = \frac{exp(u)}{(1+exp(u))^2} = \text{expit}(u)\text{expit}(-u).$$

Were we to treat the covariates as fixed, the expectations would be replaced by empiric averages over the covariates determined by whatever n was being considered, but the exact values of every single individual's covariates X_i would need to be specified in order for power to be evaluated, which makes interpretation cumbersome. Moreover, before the data are collected, the values of actual random X are not known, but the distribution can be hypothesized. Under correct-model assumptions, standard model-based inference gives valid inference in large samples regardless of whether the covariates are viewed as fixed or random, so the random-X power statements are valid even when they may not entirely reflect how analyses are actually done. Perhaps more importantly, the differences between random- and fixed-X inferences are usually extremely small in practice, and so the convenience of approximating power under random-X assumptions strongly outweighs the interpretational difficulties of choosing a set of fixed X values and using them to calculate power/sample size.

Direct evaluation of the Fisher information matrix has been suggested in the literature, and will be discussed for important special cases in Section 1.2.5. For general X distributions both Lyles et al. [2007] and Self and Mauritsen [1988] exploit essentially the same idea: by either weighting a logistic regression fit to a large 'exemplary' dataset [Lyles et al., 2007] or constructing a dataset in which every combination of assumed-categorical covariates is enumerated [Self and Mauritsen, 1988], one can approximate the expectation terms in the Fisher information matrix by weighted averages. However, as noted by the authors of the

G*Power software [Faul et al., 2009] that implements Lyles *et al*'s method, the enumeration procedure may thus be rather slow and may need large amounts of computer memory for large sample sizes. Called several times—as when plotting power curves—this slowdown can be severe. Deciding how to construct the exemplary dataset or categorize any quantitative covariates bring further difficulties, particularly for non-experts, and so we will not pursue these methods further.

1.2.3 Whittemore's method

Whittemore [1981] proposed a method for approximating $\mathbf{I}$ for small response probabilities. Arguing that at these small probabilities, the denominator terms in $f(\boldsymbol{\beta}'\mathbf{X})$ are approximately one, we obtain

$$\mathbf{I}_{ij} \approx n e^{\beta_0} \mathbf{E}[\, X_i X_j e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k} \,]$$

for $i, j \in \{1, 2, \ldots, k\}$, which we observe is a second derivative of the moment-generating function for the distribution of the non-intercept terms of $\mathbf{X}$, defined as

$$m(\mathbf{t}) = \mathbb{E}[\, e^{t_1 X_1 + t_2 X_2 + \ldots t_k X_k} \,].$$

Writing $m^{(1)}$ and $m^{(2)}$ as the first and second derivatives of m with respect to $\mathbf{t}$, the full Fisher information matrix can be approximated as

$$H = n e^{\beta_0} \begin{bmatrix} m & m^{(1)} \\ m^{(1)\prime} & m^{(2)} \end{bmatrix},$$

where all terms involving m are evaluated with $\mathbf{t}$ as the non-intercept elements of $\boldsymbol{\beta}$. Writing $v(\boldsymbol{\beta})$ as the 2,2 element of H^{-1}, we obtain the asymptotic variance of $\hat{\beta}_1$ as

$$\mathrm{Var}[\, \hat{\beta}_1] \approx (n e^{\beta_0})^{-1} v(\boldsymbol{\beta}).$$

For tests of the null hypothesis that $\beta_1 = 0$ Whittemore appears to argue that the sample size can be evaluated by equating the upper $\alpha/2$ quantile of the point estimate's distribution

under the null with the lower β quantile under the alternative, and hence that we must have sample size

$$ne^{\beta_0} > \frac{\left(Z_\alpha\sqrt{v(\boldsymbol{\beta}_{\hat{1}})} + Z_\beta\sqrt{v(\boldsymbol{\beta})}\right)^2}{\beta_1^2} \tag{4}$$

where Z_α and Z_β are quantiles of the standard Normal distribution for level α and desired power $1-\beta$, and $\boldsymbol{\beta}_{\hat{1}}$ denotes the full vector $\boldsymbol{\beta}$ but with the β_1 term replaced by zero, i.e. the parameter under the null hypothesis. Inverting Whittemore's sample size formula, we obtain an approximate power estimate.

As noted by Demidenko [2007, 2008], Whittemore's derivation does not entirely reflect the way Wald tests are used in practice. Specifically, use of the standard error of the point estimate under the null is not motivated, as it is never used in practice. Instead the Wald test proceeds by comparing $\hat{\beta}_1^2 / \left(\widehat{\text{StdErr}}[\,\hat{\beta}_1\,]\right)^2$ to χ_1^2, where the standard error is estimated at the fitted value. At no point do we compare point estimate $\hat{\beta}_1$ to its standard error estimated under the null, which is therefore not relevant, and (as Demidenko supports with direct simulation) more accurate power/sample sizes will be obtained if we exploit this fact. However, Whittemore's approach, which is widely used in literature such as in Hosmer et al. [2013] and Hsieh et al. [1998], is usually accurate for rare outcomes unless very large effect sizes are considered [Schoenfeld and Borenstein, 2005].

An attractive feature of Whittemore's approach—particularly given the computational resources available when it was derived—is that it can be used for any $\mathbf{X}$ for which the moment generating function can be obtained. The original paper considers Normal, exponential, Poisson and Bernoulli-distributed covariates.

Whittemore's approach also provides a useful insight into the impact of covariate correlation, in adjusted logistic regression analyses. For multivariate Normal $\{X_1, X_2, \ldots, X_p\}$ with

mean μ and variance Σ, the asymptotic variance of $\hat{\beta}_1$ simplifies to

$$v(\boldsymbol{\beta}) = \left(\text{Var}[\, X_1 \,] e^{\tilde{\boldsymbol{\beta}}' \mu + \frac{1}{2} \tilde{\boldsymbol{\beta}}' \Sigma^{-1} \tilde{\boldsymbol{\beta}}'} (1 - \rho^2_{1 \cdot 2 \cdots k}) \right)^{-1},$$

where $\tilde{\boldsymbol{\beta}}$ denotes the non-intercept coefficients $\{\beta_1, \beta_2, ...\beta_k\}$ and $\rho_{1 \cdot 2 \cdots k}$ is the multiple correlation coefficient of X_1 with all the other non-intercept covariates. (Centering and scaling the covariates X further eliminates μ and sets Σ to be the correlation of the covariates.) We see that the value of the multiple correlation coefficient $\rho_{1 \cdot 2 \cdots k}$ plays a large part in determining how the variance (and hence sample size required) scales compared to an unadjusted analysis with the same effect β_1. For ρ near zero covariate X_1 cannot be explained by linear combinations of the other covariates, and there is almost no impact on sample size (and hence power). In contrast, for high multiple correlations X_1 is essentially a linear combination of the other covariates, and much greater sample sizes are needed to achieve the same power.

A different insight was provided by Hsieh [1989], who noted that Whittemore's approach for simple logistic regression, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ with a standard Normal covariate X gives

$$\begin{aligned} v(\boldsymbol{\beta}_{\hat{1}})^{1/2} &= 1 \\ v(\boldsymbol{\beta})^{1/2} &= e^{-\frac{1}{4}\beta_1^2} \end{aligned}$$

and hence sample size from

$$n e^{\beta_0} \geq \frac{Z_{1-\alpha/2} + Z_{1-\beta} e^{-\frac{1}{4}\beta_1^2}}{\beta_1^2}$$

But using the Whittemore rare-disease approximation, the left hand side here is just the expected number of cases, under the null. This means that, up to the approximations involved, power $1 - \beta$ depends on n only through the expected number of cases, i.e. that power is approximately independent of total sample size n given the number of events. This insight may be particularly useful for studies where the number of cases can be fixed by design,

meaning we only need to specify β_1 (the effect size per standard deviation difference in X) to obtain power, although the accuracy of Whittemore's method should also be considered.

Being alert to the original rare-outcome approximation being used in the denominator of $f(\boldsymbol{\beta}'\mathbf{X})$, Whittemore [1981] also proposes a correction factor. Denoting the sample size obtained from (4) as N_1, and the sample size obtained using the same argument but without the approximation in the denominator of $f(\boldsymbol{\beta}'\mathbf{X})$ as N, Whittemore showed that $(N - N_1)/N_1$ can be expanded in terms of e^{β_0}, and using just the first term in this expansion obtains an approximation of the fractional error in N_1 relative to N. The general form of the correction provides no particular insight, but for small effects (i.e. β_1 close to zero) it simplifies to just

$$N_1 \approx \frac{N}{1 + 2e^{\beta_0}}. \tag{5}$$

As expected from Whittemore's original formulation, for rare outcomes β_0 is low and we find the formula given in (4) is not affected by its use of a simplifying approximation.

While offering less insight, the correction factor can also be written concisely for univariate logistic regression; for standardized X_1 with mean 0 and variance 1 it gives

$$\begin{aligned} N_1 &\approx \frac{N}{1 + 2e^{\beta_0}\delta(\beta_1)}, \\ \text{where } \delta(\beta_1) &= \frac{1 + (1 + \beta_1^2)e^{5\beta_1^2/4}}{1 + e^{-\beta_1^2/4}}. \end{aligned}$$

1.2.4 Hsieh *et al*'s method

Following the work of Hsieh [1989] tabulating and explaining the sample size formula of Whittemore [1981], Hsieh et al. [1998] provide an alternative approach to power/sample size calculations for logistic regression. The key insight uses the equivalence of the score tests under two analyses. The score test statistic from a logistic regression of binary outcome Y on a single covariate X can be shown to be just $n\text{Cor}[X, Y]^2$, i.e. the sample size times the squared

correlation of X and Y. This is exactly the same score test statistic from a "reverse" linear regression of X on binary outcome Y. The precise p-value from logistic regression would usually come from considering test statistic behavior with the variable treated as a fixed covariate, but under correct-model assumptions the difference between doing this instead of considering the covariate as random is small. Consequently, we can expect their powers to be extremely similar.

The near-equivalence of the two tests means that standard sample size calculations for the t-test (i.e. regression of continuous X on binary Y) can be used. We must provide the proportion of outcomes $Y = 0, 1$ the difference in mean covariates between those two groups and the (assumed-common) variance of X within each group, σ^2. In the linear regression case this gives sample size formula

$$n \geq \sigma^2 \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2} \frac{(k+1)^2}{k}$$

where δ denotes the difference in mean X between the $Y = 0$ and $Y = 1$ groups, and k is the sample size ratio, i.e. the relative probability of having $Y = 1$ versus $Y = 0$. After some simplification and rescaling so that $\text{Var}[\, X\,] = 1$, Hsieh et al. [1998] give the sample size formula

$$n \geq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{f(\beta_0 + \beta_1 \mu)\beta_1^2}.$$

where μ is the mean of covariate X.

The assumption of a common variance is somewhat limiting, although for small effects β_1 it seems reasonable that any difference between $\text{Var}[\, X|Y = 1\,]$ and $\text{Var}[\, X|Y = 0\,]$ will not grossly invalidate the analysis. Hsieh et al's formulation also requires that users specify difference in covariate distributions between outcome groups, which may be cumbersome in this case.

A more serious problem with the approach is its lack of an obvious generalization to multiple covariates; regressing multiple X_j on a single binary Y is possible, but methods to do so are not as well understood, in either the power of corresponding tests or their relationship with logistic regression of Y on multiple X_j.

Following Hsieh [1989], an informally-motivated 'workaround' is provided by Hsieh et al. [1998]. They suggest dividing the sample size from a univariate calculation by a factor of $(1 - \rho^2)$, where as in Section 1.2.3 ρ denotes the multiple correlation coefficient of X_1 with all other covariates. This method also draws from power/sample size formulae for linear regression, where it is an exact correction: for the logistic regression setting it is convenient and simple to understand, but does not reflect the true dependence on power/sample size on the data-generating mechanism.

The univariate and multivariate versions of Hsieh's method, as we have described them, are the only ones provided for sample size in the PASS software [PASS, 2019].

1.2.5 Schoenfeld & Borenstein's method

Schoenfeld and Borenstein [2005] (hereafter S-B) developed more reliable analytic arguments than those of Sections 1.2.3 or 1.2.4, in the sense that they directly address the value of the Fisher information discussed in Section 1.2.2, rather than some closely-related but different quantity. To do this S-B use numeric integration, and assume that only either discrete and/or multivariate Normal covariates are present. However, in a major contribution they show that, regardless of the number of covariates, the calculations can be done with only one-dimensional numerical integration.

To illustrate their method, we first consider the case of "simple" logistic regression, with one variable. With $g(t)$ as the probability density function of the variable, with location and

scale parameters (τ, σ), the elements of the Fisher information matrix for a single observation

$$\mathbf{I} = \begin{bmatrix} e_0 & e_1 \\ e_1 & e_2 \end{bmatrix}$$

are calculated as

$$e_i(\tau, \sigma) = \int t^i f(\beta_0 + \beta_1(\sigma t + \tau)) g(t) dt.$$

(When the variable is assumed Normal, $g(t) = \phi(z)$ is used, i.e. the standard Normal's probability density function.) The corresponding large-sample standard error of $\hat{\beta}_1$ is then

$$[v(\beta_0, \beta_1) n^{-1}]^{1/2}, \text{where } v(\beta_0, \beta_1) = \left[\frac{e_0}{e_2 e_0 - e_1^2} \right],$$

and—although stated with typographic errors in Schoenfeld and Borenstein [2005]—the corresponding approximate sample size N required to ensure power $1 - \beta$ at level α is

$$N = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 v(\beta_0, \beta_1)}{\beta_1^2}.$$

Here we have used the arguments of Demidenko [2007], which are justified by simulation as well as theory, to motivate only calculating the standard error under the alternative. For a fixed sample size, we can straightforwardly invert the equation to give power calculations.

The results can be extended to logistic regression with multiple discrete and/or multivariate Normal variables. As mentioned above, the calculations simplify to one-dimensional numerical integrations regardless of the number of variables, as follows:

Consider a logistic regression model with two subsets of variables, some discrete and some continuous. We can think of the discrete variables as defining K distinct strata, and define w_s as the proportion of the population in stratum s, for s in $1, 2, \ldots, K$. We consider the continuous variables to be multivariate Normal within each of the strata, with mean μ_s and Σ_s. Following Schoenfeld and Borenstein [2005], it is also algebraically convenient to define v_s as the vector combination of the discrete variables (i.e. a vector of 0s and 1s) that gives the

value of those variables within each stratum s, which lets us define the overall mean variable in stratum s as $\mathbf{u_s} = (v_s, \mu_s)$. S-B further similarly define the overall variance-covariance in stratum s as matrix $\mathbf{D_s}$, a square matrix with row dimension equal to that of $\mathbf{u_s}$, which has a lower right hand corner elements equal to $\mathbf{\Sigma_s}$, and zero elsewhere. Finally, we define $\sigma_s = \sqrt{\beta' \mathbf{D_s} \beta}$, $\tau_s = \beta' \mathbf{u_s}$, and $\gamma = \frac{\mathbf{D_s}\beta}{\sigma_s}$, where e_i, $f(u)$ and $\phi(t)$ are as defined above. Then after some algebra it can be shown that

$$\begin{aligned}
\mathbb{E}[\,\mathbf{X}\mathbf{X}' f(\beta'\mathbf{X})\,] &= \sum_{s=1}^{K} w_s \left((\mathbf{u_s}\mathbf{u_s}' + \mathbf{D_s} - \gamma_s\gamma_s')e_0(\tau_s,\sigma_s) + (\mathbf{u_s}\gamma_s' + \gamma_s\mathbf{u_s}')e_1(\tau_s,\sigma_s) + \gamma_s\gamma_s' e_2(\tau_s,\sigma_s) \right) \\
\mathbb{E}[\,\mathbf{X} f(\beta'\mathbf{X})\,] &= \sum_{s=1}^{K} w_s \left(\mathbf{u_s} e_0(\tau_s,\sigma_s) + \gamma_s e_1(\tau_s,\sigma_s) \right) \\
\mathbb{E}[\,f(\beta'\mathbf{X})\,] &= \sum_{s=1}^{K} w_s e_0(\tau_s,\sigma_s).
\end{aligned}$$

The basic idea being exploited here is that the common $e_i(\tau_s, \sigma_s)$ terms (for $i = 0, 1, 2$) can be linearly transformed to give the contributions to the overall expectations with respect to all the continuous variables. Summation over the K strata then provides the expectation with respect to the discrete variables. Importantly, within each stratum, the integrals e_0, e_1, e_2 need only be calculated once, and are one-dimensional, which significantly reduces computing time compared to naïvely evaluating multidimensional integrals.

To evaluate the power in these settings, Schoenfeld and Borenstein [2005] argue via use of non-centrality parameters in chi-squared distributions with degrees of freedom equal to the dimension of the parameter component being tested. Using Equation 2 from Section 1.2.2, this again only involves calculation of asymptotic variances under the alternative, as per Demidenko [2007].

The time complexities of computing Hsieh's and Whittemore's methods are negligible. For S-B the slowest step for all situations we consider is the one-dimensional integrations, the number of which grows with the number of strata K. However, unless K is extremely

large – for example several hundred values given by considering multiple discrete covariates — then the computational time required for S-B's method remains negligible. This contrasts sharply with use of direct simulation, for which the time required to fit each logistic regression grows with the number of covariates, and where the number of simulations required is large. However, simulation remains the gold standard, when sufficient simulations can be done.

We note that the numerical evaluation of expectation in the Fisher information terms could also be performed using Monte Carlo methods, as described by Vittinghoff et al. [2009]. While faster than evaluating power directly by simulation, particularly for multiple covariates this approach still requires large numbers of randomly-generated covariates, for each setting considered, and would scale poorly as the number of covariates increased. For these reasons the approach is not considered further here.

1.3 Textbook treatments of power/sample size

It is routine for applied biostatistics textbooks to discuss some form of sample size and power calculations. For example, most texts provide formula for two-group comparisons such as the t-test. A formula for two-groups comparisons of binary outcomes, such as Pearson's χ^2 test for 2×2 tables are present in some texts, although for small samples and with access to modern computing resources, complete enumeration can provide direct statements of power, and hence sample size.

Power and sample size for linear regression are common topics in these texts, and various simple and free online tools such as *OpenEpi* [Dean et al., 2014] or the Free Statistics Calculators index enable their use by investigators. However, beyond the two-groups setting power and sample size for regressions involving binary outcomes—such as the formula in Sections 1.2.3 (Whittemore's), 1.2.4 (Hsieh's) and 1.2.5 (S-B) — are rarely considered, making them little-known to students and researchers who might benefit from the knowledge.

A full literature review of all introductory-level texts is beyond the scope of this thesis, but Table 1 summarizes how power and sample size calculations for logistic regression are covered—if at all—in a sample of popular texts. As it shows, the more advanced approaches are typically omitted, and even in the more advanced texts where they are discussed (specifically Hosmer et al. [2013]) there is very little discussion of their accuracy, nor examples of their application in practice.

Conversely, the methods that are typically included in textbooks at this level are usually only pertinent to the most straightforward scenarios such as approximating sample size formula for binomial proportions. Some texts do suggest simulation as a general-purpose approach to calculating power and sample size, but for example Van Belle et al. [2004, Chap 8] do so without giving example code, of any sort. As noted in Section 1.2.1 the coding involved—including dealing with edge cases—would appear to put it beyond the level of these standard texts.

Given that methods of Whittemore, Hsiesh and S-B from Sections 1.2.3, 1.2.4 and 1.2.5 are straightforward to program, this brief survey suggests that an easy-to-use interface to their function would be of practical value to users who may lack the technical skills and/or time to undertake extensive simulations. Guidance, at the same technical level, for when the approximate methods are reliable, would also be a contribution to the field.

2 Methods evaluation: univariate case

In the univariate case, we assume a "simple" logistic model with $\text{logit}(p) = \beta_0 + \beta_1 X_1$. The effect size of interest is β_1, the log odds ratio per unit difference in variable X_1.

Textbook	Methods discussed
Van Belle et al. [2004]	Presents sample size formula for case-control studies given 2x2 tables.
Rothman et al. [2008]	Discusses power and sample size concepts without formula.
Breslow et al. [1980]	Includes no explicit power formula.
Rosner [2015]	Discusses sample size formula for 2x2 tables in Chapter 10, but no sample size/power formulae in Chapter 11's material on logistic regression.
Hosmer et al. [2013]	Describes some approximation methods without those evaluated or applied.
Kleinbaum et al. [2002]	No explicit mentioning of power/sample size formulas.

Table 1: Summary of power/sample size calculations for binary outcomes in a selection of standard textbooks. With the exception of Hosmer et al. [2013], none of these texts provide formula for evaluation of power or sample size methods in logistic regressions, beyond those for 2×2 tables. Many include only a quick qualitative description, leaving introductory users without help when power or sample size are needed.

2.1 Reproducing S-B's examples

As a first example, we reproduce the power calculations from Table 1 of Schoenfeld and Borenstein [2005], which with $n = 500$ evaluates power by a number of methods over a range of settings where power is approximately 90%. We consider power calculated by direct simulations, by S-B's method, Whittemore's method with sample size correction and Hsieh's method. The original study also included a method in the proprietary nQuery software [nQuery, 2017], that combines ideas from both Whittemore and Hsieh. Given that nQuery is proprietary and was shown to lack advantage over Hsieh's and Whittemore's in the original study, we choose not to include the nQuery method in our project.

Standard Normal covariates are assumed. S-B vary the disease prevalance from approximately 5% (with $\beta_0 = -3, \beta_1 = 0.68$, $e^{\beta_1} = 1.97$) to 73% (with $\beta_0 = 1, \beta_1 = 0.33$, $e^{\beta_1} = 1.39$), presumably to show the impact of Whittemore's rare-disease assumption.

We noticed that, comparing to the stated model coefficients, the last two numbers in the row of "odds ratios (e^{β_1}))" in S-B's Table 1 appear to be typographic errors, and do not follow what S-B wrote in their text. Therefore we present results here for odds ratios of 1.97, 1.57,

1.39, 1.34 and 1.39, i.e. we follow the text, as this gives better agreement with the stated power results. This discrepancy contributes to some minor differences in later sections, which we will discuss as they occur.

Our gold standard measure of power is calculated by simulation, using 10,000 simulated data sets for each setting and calculating the Wald test for each. This is more than the 2,000 used in S-B's simulation results, and provides accuracy up to approximately $\pm 0.5\%$ for power around 90%.

Table 2 gives the results. The agreement with S-B's results is excellent, at least up to the irremovable Monte Carlo error in what S-B present, which is based on only 2,000 simulations for each setting. As S-B found, their method gets closest to the true power in all cases, though Hsieh's method is only slightly further away. (We note our evaluation of Hsieh's method for $\beta_0 = 0$ slightly disagrees with S-B's version, but only for$\beta_0 = 0$, which we interpret as a further typographic error in their manuscript.) Whittemore's method is equivalently good for rare outcomes, but even with Whittemore's correction gives increasingly bad approximations for more common outcomes; for common outcomes the approximate power based on the formula is almost 100% instead of the true value of 90%, which would clearly have important implications for sample size determination. In summary, our calculations support S-B's conclusion that their method gives the best results of all the analytic approaches, although it does not beat Hsieh's method by much.

2.2 Generalizing S-B's example

The accuracy of the asymptotic approximations used in S-B's method (and in turn more crudely approximated by Whittemore and Hsieh's methods) can be expected to increase with growing n, and to perform less well at smaller n. The results in Section 2.1 seem to indicate that S-B's use of asymptotic approximation presents no major difficulties at $n = 500$ for the

β_0	-3	-2	-1	0	1
β_1	0.68	0.45	0.33	0.29	0.33
Odds Ratio e^{β_1}	1.97	1.57	1.39	1.34	1.39
P	0.05	0.12	0.27	0.50	0.73
Simulated Power	0.92	0.90	0.90	0.89	0.90
Corrected S-B	0.92 (-0.00)	0.90 (-0.00)	0.90 (-0.00)	0.88 (-0.01)	0.89 (-0.01)
S-B adding ^{2} only	1.00	1.00	0.99	0.99	0.99
S-B as is	1.00	1.00	1.00	1.00	1.00
Whittemore	0.90 (-0.02)	0.90 (-0.00)	0.92 (+0.02)	0.95 (+0.06)	1.00 (+0.10)
Hsieh	0.90 (-0.02)	0.90 (+0.00)	0.91 (+0.01)	0.90 (+0.01)	0.91 (+0.01)

Table 2: Analytic and simulation-based evaluation of power, based on Schoenfeld and Borenstein's example for a single covariate. Throughout, we take sample size $n = 500$, and use 10,000 simulations for each simulation setting. P denotes the disease prevalence (i.e. probability of being a case) in each setting. "Corrected S-B" denotes the S-B method, after corrections for typos, which are discussed in 1.2.5. "Whittemore" denotes Whittemore's method with the small sample correction using the standard error calculated under the null hypothesis as in 1.2.3. "Hsieh" denotes the reverse linear regression approach of Section 1.2.4. We suspect that the S-B equation 6 has two small typos, namely (1) the omitted ^{2} outside of the parentheses and (2) the misplaced standard error v, which should be outside of the parentheses before squaring. We also present the calculated powers from S-B equation 6 without correction of one or both typographic errors. Obviously the power calculations are problematic without corrections.

parameter values considered, although of course it provides no complete guarantee of accuracy. (Hsieh's method is also almost as reliable here.)

As a first attempt at challenging the methods, we therefore repeat the analysis from Table 2, but at $n = 500/2 = 250$ and $n = 500/4 = 125$, to gain some understanding of where they might start to perform less well. (We also omit results from the typo-afflicted versions of the S-B method, as the intended version was established in Section 2.1.)

The results are given in Table 3. As expected, Whittemore's method still badly overstates power for less-rare disease outcomes, and cannot be recommended for practical use in those settings. With the possible exception of the $\beta_0 = 1$ setting, the S-B method outperforms Hsieh's method, in particular at the rarest-disease setting with $\beta_0 = -3, P = 0.05$, where Hsieh's method understates power by 4% at both sample sizes. This may seem small in practice but as a proportion of the true power (36% for the smaller sample size) it is perhaps

Sample size (N)	250					125				
β_0	-3	-2	-1	0	1	-3	-2	-1	0	1
β_1	0.68	0.45	0.33	0.29	0.33	0.68	0.45	0.33	0.29	0.33
Odds Ratio	1.97	1.57	1.39	1.34	1.39	1.97	1.57	1.39	1.34	1.39
P	0.05	0.12	0.27	0.50	0.73	0.05	0.12	0.27	0.50	0.73
Simulated Power	0.67	0.63	0.64	0.61	0.62	0.36	0.36	0.35	0.34	0.36
Corrected S-B Difference	+0.00	-0.00	-0.02	-0.00	-0.00	+0.00	-0.00	-0.02	-0.00	-0.00
Whittemore Difference	-0.06	-0.01	+0.01	+0.15	+0.28	-0.06	-0.01	+0.01	+0.12	+0.28
Hsieh Difference	-0.04	+0.00	-0.00	+0.02	+0.02	-0.04	+0.01	-0.00	+0.02	+0.02

Table 3: Simulation-based power and differences between analytic and simulation-based evaluation of power for smaller sample sizes, again based on Schoenfeld and Borenstein's example for a single covariate and 10,000 simulations for each simulation setting. 'Corrected S-B', 'Whittemore' and 'Hsieh' denote the same methods referred to in Table 2. The S-B method remains stable for $N = 250$ and $N = 125$. Hsieh's method is also comparable in accuracy, but tends to overestimate for any substantial case proportion P. Whittemore's method cannot be recommended.

less trivial, even though the focus would usually be on power targets around 80% or 90%. Given that Hsieh's method, as used here, makes exactly the same assumptions as S-B but provides less accurate results, we find that its use can only be justified by computational convenience—meaning that if S-B's method were available in a user-friendly form Hsieh's method would not need being considered at all.

To further extend the analysis of Table 3, we repeat the analysis but now for values of n up to 500, and graph the results in Figure 1. The over-statement of power in Whittemore's method is commonly present, but higher severity for the setting with more common diseases i.e. larger P. The close agreement between Hsieh and S-B's method is also apparent at all n on the settings we considered, though with careful inspection we see that S-B appears slightly closer to the gold-standard simulation values, overall.

The differences between the gold-standard and approximate power calculations are displayed directly in Figure 2. This makes it clearer that Hsieh's method appears to systematically but slightly over-estimate power, versus S-B's method which appears to be more unbiased but has a tendency to underestimate power very sligtly. Given that in practice over-estimation

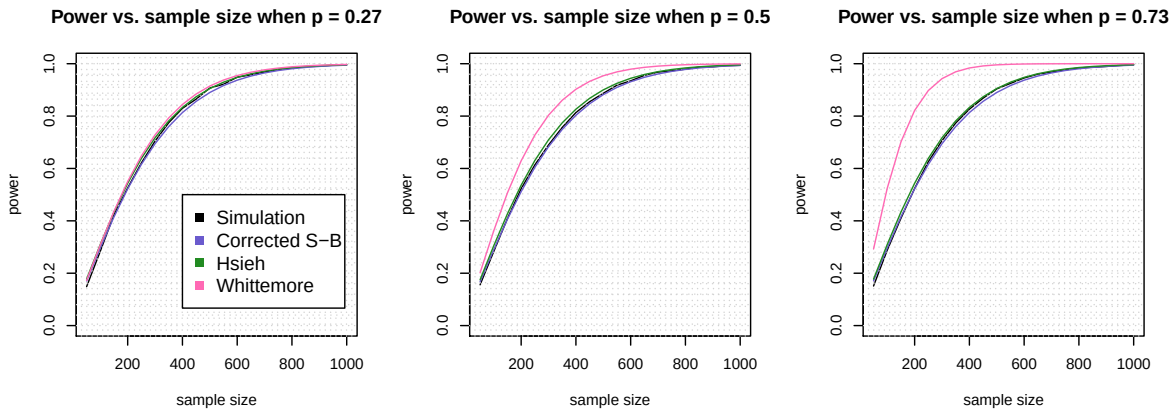


Figure 1: Power approximations by Whittemore (pink), Hsieh (green) and S-B (purple blue) at three fixed proportions $P = 0.27, 0.50$ and 0.73, with $\beta_0 = -1, 0, 1$ and $\beta_1 = 0.33, 0.29, 0.33$, respectively. Sample size is taken between $N = 10$ and $N = 500$. The black line represents power as evaluated by simulation, with 10,000 simulations per setting. The three methods all work fairly well when $P \approx 0.27$. Both Hsieh and S-B retain their accuracy at larger P, yet "S-B" remains the closest to simulated powers. Some over-estimation of power occurs for both S-B and Hsieh at the smallest sample sizes.

of power is probably more serious than under-estimation (as it leads to under-sized studies and under-precision in inferences) this seems to again argue for using S-B's method in place of Hsieh's.

2.3 Impact of the covariate distribution

In the examples so far we have only considered Normally-distributed covariates. However for the univariate case (where S-B's use of 1-dimensional integration instead of integrating out multiple covariates) there is no difficulty in using the theory of Section 1.2.2 to consider general covariate distributions. This is in contrast to Whittemore's approach, which relies on moment generating functions being known – although in practice all software implementations of it use the version where X is Normal. It is also in contrast to Hsieh's method, which uses Normality in small samples, and equal variance of X at different values of Y in large samples – and for which no modifications are available when X is not Normal.

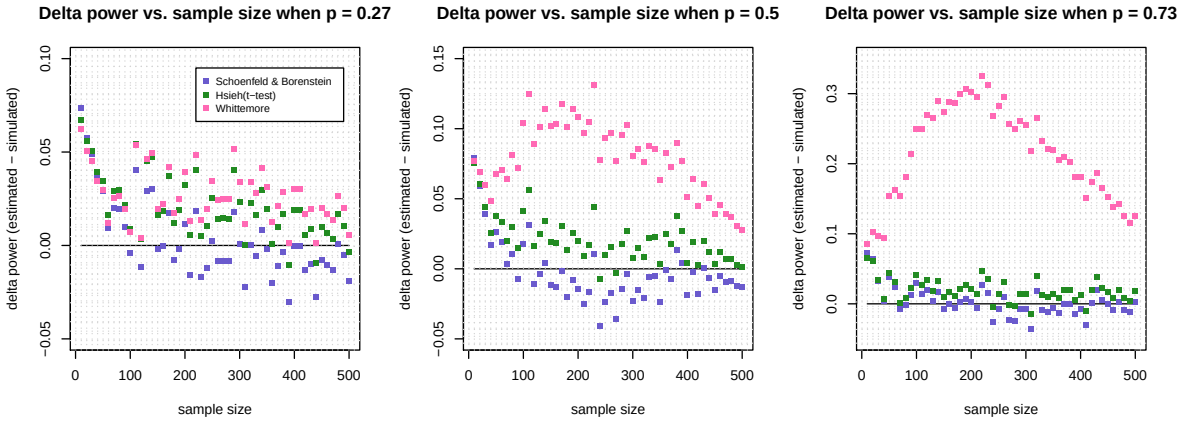


Figure 2: Power differences by "Whittemore", "Hsieh" and "S-B" at three fixed proportions $p = 0.27, 0.50$ and 0.73, with $\beta_0 = -1, 0, 1$ and $\beta_1 = 0.33, 0.29, 0.33$, respectively. Sample size is again taken between $N = 10$ and 500. The S-B method appears to produce powers that are the closest and most unbiased compared to simulated powers. Hseih's is much better than Whittemore's, but tends to overestimate power fairly consistently.

It is therefore of interest to see how well all three approaches perform for non-Normal X. Here we use three covariate distributions to assess the methods' performances, with different tail-weights in each. The distributions are a t-distribution with 10 degrees of freedom, a uniform distribution, and a two-point (Rademacher) distribution, each scaled to have mean 0 and variance 1. For the Whittemore and Hsieh methods we use the same formulas as before, without any modification. For the S-B method we assume knowledge of the covariate distribution, performing the relevant integrations in the Fisher information matrix under the relevant t, uniform or Rademacher distribution. Table 4 shows the results.

With a heavier-tailed covariate distribution, the S-B method clearly outperforms the other methods. Against lighter-tailed uniform and Rademacher covariates, the advantages of S-B's method are less clear cut, but its results appear to be close to gold-standard results everywhere except for rare disease settings, i.e. $P = 0.05$. Here the Hsieh method does at least as well as S-B in all settings. For the light-tailed uniform distribution, this appears to be a modest practical benefit. For the Rademacher covariates — where the data form a 2×2 contingency

β_0	-3	-2	-1	0	1
β_1	0.68	0.45	0.33	0.29	0.33
Odds Ratio	1.47	1.25	1.18	1.16	1.18
P	0.05	0.12	0.27	0.50	0.73
			t-distribution		
Simulated Power	0.97	0.95	0.95	0.94	0.95
Corrected S-B Difference	+0.00	-0.00	-0.00	-0.01	-0.01
Whittemore Difference	-0.07	-0.05	-0.02	+0.02	+0.05
Hsieh Difference	-0.07	-0.05	-0.05	-0.04	-0.05
			Uniform		
Simulated Power	0.91	0.90	0.90	0.89	0.90
Corrected S-B Difference	-0.03	-0.01	-0.01	-0.00	-0.00
Whittemore Difference	-0.01	+0.00	+0.02	+0.07	+0.10
Hsieh Difference	-0.01	+0.00	+0.01	+0.01	+0.01
			Rademacher		
Simulated Power	0.88	0.90	0.90	0.90	0.90
Corrected S-B Difference	-0.05	-0.02	-0.01	-0.00	-0.01
Whittemore Difference	+0.02	+0.01	+0.02	+0.06	+0.10
Hsieh Difference	+0.01	+0.00	+0.00	+0.00	+0.00

Table 4: Simulated power and differences between analytic and simulated powers for the same simulation settings, model coefficients β and sample size ($N = 500$). When X follows the t-distribution, S-B shows clear superiority. Under uniform X both S-B and Hsieh retain reasonable performance. For the Rademacher-distributed X Hsieh's method shows the best performance, though S-B's is almost equally good when the proportion P is not small.

table, it seems likely that a simple Pearson χ^2 or Fisher Exact test might be used, making the comparison to logistic regression Wald tests (used here) less compelling. In these situations exact power calculations are available [Bennett and Hsu, 1960, Fay, 2010], but not compared here.

Essentially the same observations seem to hold over a range of sample sizes, as we show in Figure 3. The superiority of S-B for t-distributed covariates is clear, as is the near-equivalence for uniform X. Given the discussion above of using Fisher exact test and exact power formula for the Rademacher-distributed X, in Figure 4 we focus only on the S-B method, assessing where it is likely to be a poor approximation. The performance is worst with rare outcomes, but for all settings considered with total sample size of $n > 250$ any inaccuracy in the power calculation seems unlikely to matter for practical purposes. Knowing the threshold is itself useful in practice, as with $n \leq 250$ calculating power via complete enumeration—i.e. by brute force—is feasible without any sophisticated algorithm. The results here indicate that the work involved in doing complete enumeration for larger n is unlikely to be useful.

2.4 Impact of the covariate distribution on S-B's method

In Section 2.3 we saw that the S-B method is competitive among the methods we consider, when different covariate distributions are used. A related question is how well S-B does in absolute terms, across different covariate distributions. Here we again consider the accuracy of power calculations when covariates are Normal, t with 10 degrees of freedom, uniform and Rademacher, all scaled to have mean zero and variance one. But our focus is how well S-B's method performs across the covariate distributions; results from this analysis should inform users about how much of how little assumptions of (say) Normal covariates are likely to impact the accuracy of their power/sample size calculations.

As the simulation settings are identical to those of Section 2.3 except for overall sample

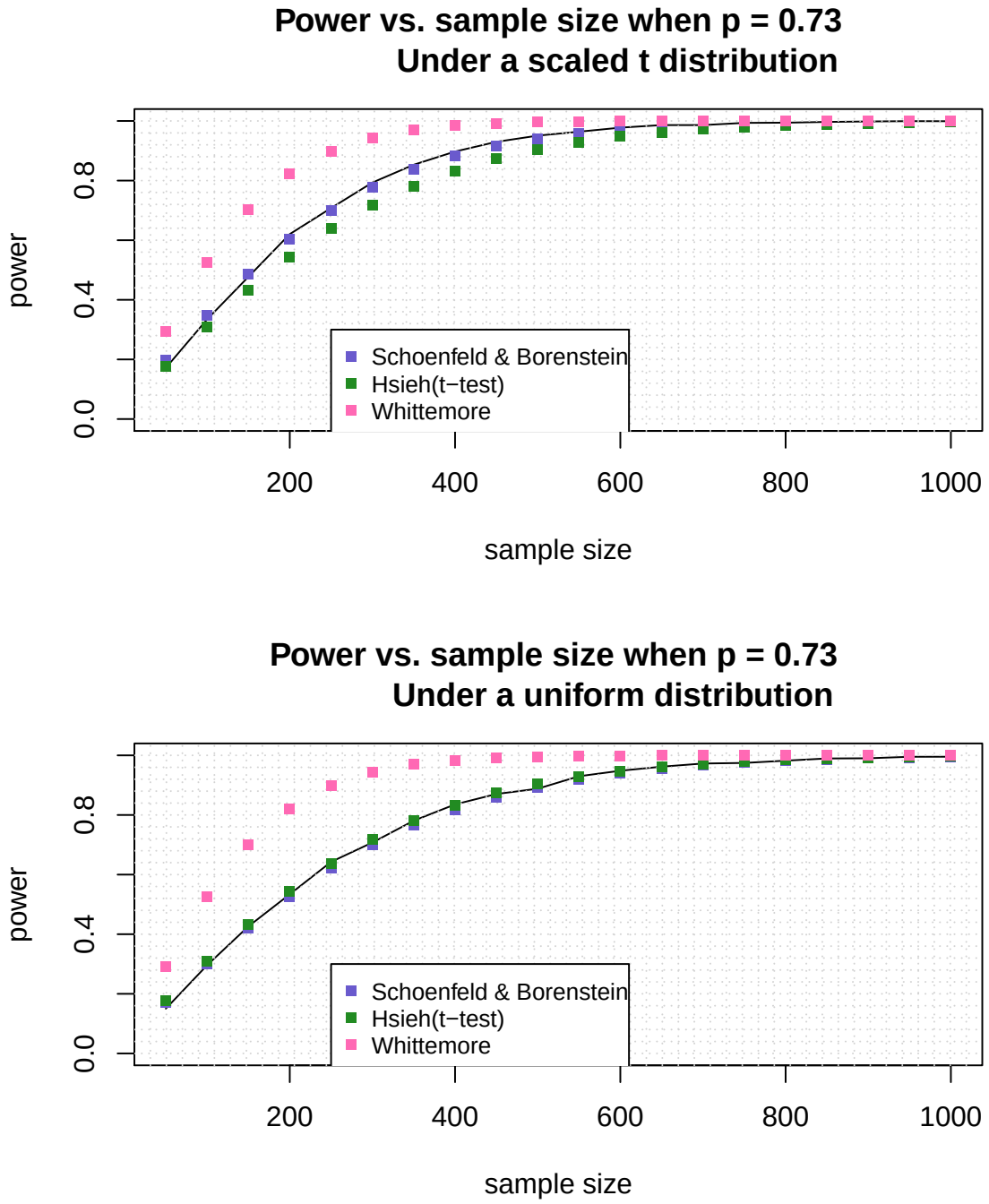


Figure 3: Power approximations of Whittemore, Hsieh and S-B given t-distributed (upper) and uniformly-distributed covariates (lower). The S-B method produces close approximation to true power in both cases. Results based on Whittemore's method consistently overestimate unless n is very large. Hsieh's performs well against a uniform distribution, but less well for the heavy-tailed t distribution.

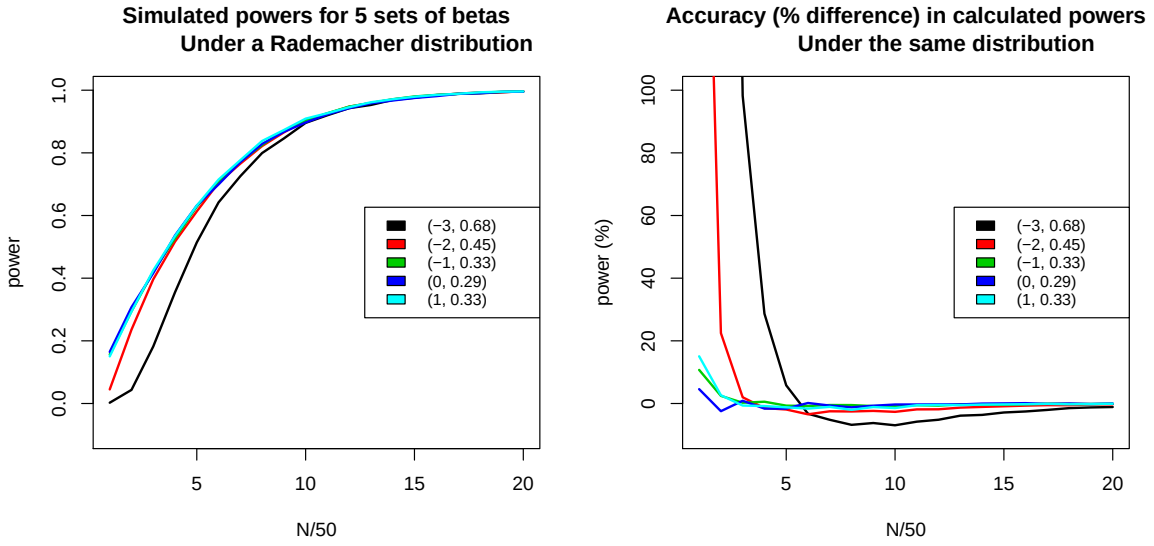


Figure 4: Left panel: powers produced by the S-B method under a Rademacher (binary) distribution under five scenarios each with given coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)$, with sample size between $N = 50$ to 1000. The proportions P are 0.05, 0.12, 0.27, 0.50, 0.73, from top to bottom. Right panel: the same results, but plotted to show accuracy as a percentage difference from the true power, evaluated by simulation. Accuracy generally is higher for pairs of coefficients corresponding to higher proportions P.

size, we do not repeat them again here. The results are given in Figure 5.

The most striking result from these simulations is that the covariate distribution has a modest impact on power, being higher for the heavier-tailed X distributions, albeit only by a few percentage points and then only when the disease is relatively rare. While a full explanation of why this occurs is beyond the scope of this thesis, it is presumably due to the heavier tails generating high-leverage points more often, which in turn provide greater precision with which to assess the underlying effect and hence achieve statistical significance.

The other important feature is the sample size at which accurate power statements are attained. Following Section 2.3's results, it is the Rademacher covariates that require the greatest sample size in order for S-B's power results to be close, in percentage terms, to the true power. In this worst case, and in the settings we consider, samples sizes of $n = 250$ and above would seem to be sufficient to provide accurate-enough knowledge of power.

3 Methods evaluation: multivariate case

Adjustment for covariates is an important part of many analyses using logistic regression, and so it is natural to consider corresponding power and sample size calculations. With many covariates, the relationships between them and their various effects on the outcome present a large space of settings in which we might assess the accuracy of the methods of Section 1.2. Such an extensive evaluation would also likely be too complex to help investigators understand the power their study might reasonably have.

Instead, in this section we focus on evaluating the methods of Section 1.2 when a single adjustment variable is present. Following S-B's example, we update the setting of Table 2 from Section 2 by adding an adjustment variable, X_2, with a standard Normal distribution, effect size $\beta_2 = 0.9$ and correlation $\rho = 0.5$ with X_1 per choice of Schoenfeld and Borenstein

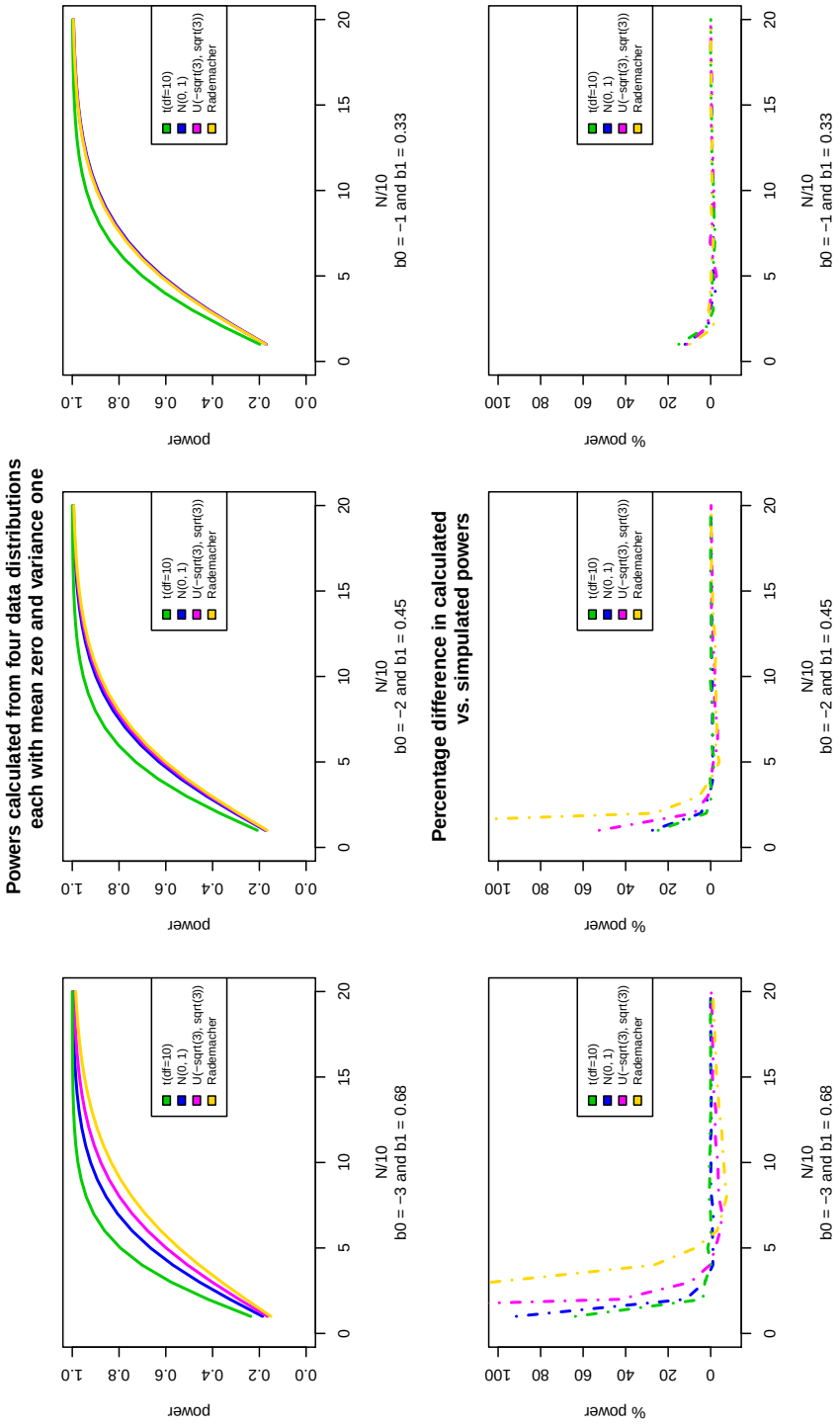


Figure 5: Analytic (S-B) powers under the same four data distributions with same mean (0) and variance (1) with three sets of model coefficients (β_0, β_1). Higher proportions P, which are specified by each set of coefficients, generally result in higher accuracy of power calculation given the same sample size N. The performance of the S-B analytic calculation is again reasonably reliable at moderately large sample size, around or above $N = 50$.

β_0	-3	-2	-1	0	1
β_1	0.68	0.45	0.33	0.29	0.33
Odds Ratio	1.97	1.57	1.39	1.34	1.39
P	0.05	0.12	0.27	0.50	0.73
Simulated Power	0.92	0.83	0.76	0.70	0.75
Corrected S-B Difference	-0.01	-0.00	-0.01	-0.00	-0.01
Whittemore Difference	+0.06	+0.14	+0.21	+0.28	+0.25
Hsieh Difference	-0.22	-0.13	-0.031	+0.10	+0.20

Table 5: Simulated power and differences between analytic and simulated power (at 10000 iterations, $N = 500$) after adjusting for a variable X_2, where $\beta_2 = 0.9$ and $\rho_{X_1,X_2} = 0.5$. The analytic powers (S-B) remain extremely close to the simulated powers, outperforming the Whittemore's and Hsieh's methods that can be quite incorrect.

[2005], where the joint distribution of (X_1, X_2) is multivariate Normal. All other aspects of the study remain as in Table 2. The new results are given in Table 5.

The simulated and approximate powers are similar to those in S-B's original paper, although the implementation of Whittemore and Hsieh's methods in the multivariate case are not clearly documented there. In particular, Whittemore's correction term, described in Section 1.2.3, appears to use the approximation given in Equation 4, where Whittemore's original term δ is set to 1. This seems to be a choice by S-B, as it is not consistent with Whittemore's original paper, but even this is not completely clear in the S-B paper.

The approximations of both Whittemore and Hsieh perform poorly throughout compared to S-B, and even though this is clearly a limited simulation study, the settings cover a range of plausible scenarios. Based on them it does not seem that either Whittemore or Hsieh's method can be reasonably entertained for adjusted analyses, if S-B's approach is available for practical use.

3.1 Methods evaluation: directly varying the case proportion

In the previous subsections, following S-B, we consider the performance of the S-B method at multiple different values of the intercept and slope, which consequently lead to a range of values of P, the proportion of cases. In this section we attempt to evaluate the role of P more

systematically, by selecting simulation settings that fix P and β_1, and use these choices to define intercept β_0. Evaluating the methods in this way is useful as in practice investigators are likely to have the best understanding of P (particularly in case-control studies) and β_1, the effect of the covariate they are primarily studying.

For simplicity we only study values of P up to 50%: beyond this one could evaluate power by switching the 'case' and 'control' labels. We consider sample sizes between 50 and 1000 with increments of 10. All covariates are assumed standard Normal and in bivariate cases are considered multivariate Normal with a specified correlation, as in Section 3.

For simple logistic regression, and fixing $\beta_1 = 0.33$, Figure 6 shows how calculated power varies with sample size and how this relationship depends on P. Specifically, power increases with larger sample size (N) and, independently, also increases as the case proportion P increases to 50%.

Figure 7 again shows how the calculated powers are close to the true power, as approximated by simulation. As well as showing absolute and percentage differences in the calculated and actual power, as an attempt to illustrate the impact of inaccuracy in a more interpretable and generalizable way, we also show the ratios of the non-centrality parameters from the power calculation and the actual power. For the calculation this is just the numeric value from the asymptotic argument of Section 1.2.2. For the empirical version, we calculate the variance of $\hat{\beta}_1$ over the simulations, and then calculate the non-centrality parameter as the average $\hat{\beta}_1^2/\text{Var}[\hat{\beta}_1]$, again over the simulations. The ratio of these two non-centrality parameters can be interpreted as effectively saying how much the power calculation overstates the sample size. For example, a ratio of 1.2 would indicate that the power calculation acts as if we had a 20% larger sample size than is actually available. This ratio can be meaningfully compared across the various settings we consider; it also gives a useful factor by which to adjust sample size

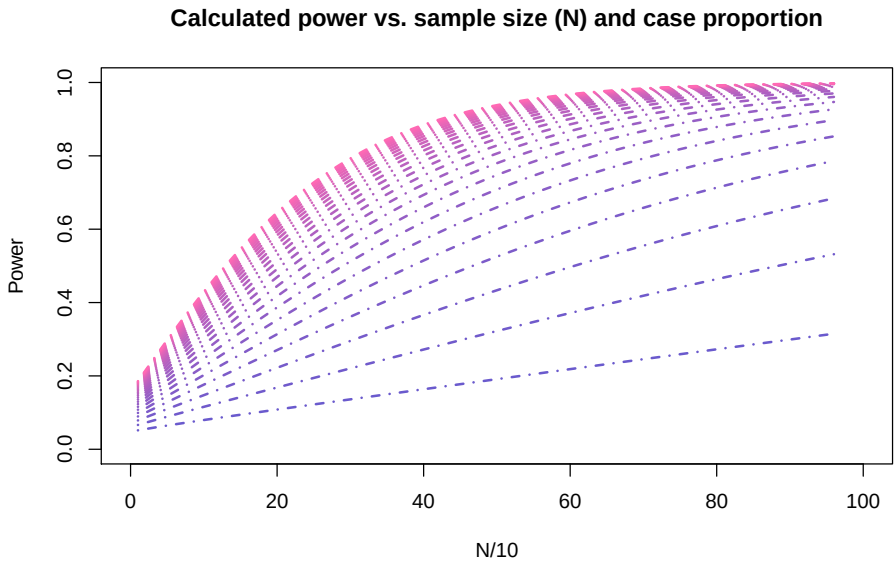


Figure 6: Calculated power given varying case proportions increases for higher sample size (N) for $N = 10$ to 1000 by increments of 10. Case proportion P ranges from 2% (lowest straight line) to 50% (highest pink curve) with 2% increments. Effect size $\beta_1 = 0.33$ throughout. Proportions closer to 50% yield higher power whereas the power curves also exhibit more curvature.

calculations, when these are performed.

From Figure 7 we see that except for very small P, the over-statement of the power seems acceptable in practice in the settings we considered for $n > 200$, and when power only needs to be known to within a factor of 10%, say, then $n > 100$ would probably suffice. The accuracy increases rapidly as sample size increases and as case proportion increases.

While not shown here, for multiple regression on exposure X_1 adjusting for covariate X_2 where the correlation between X_1 and X_2 is $\rho = 0.5$, similar results of trend and accuracy are obtained.

4 Adapting S-B's method for empirically-derived covariates

Throughout the previous sections, and the literature they describe, the assumption of random X from some specified distribution saves us from having to specify a particular design matrix

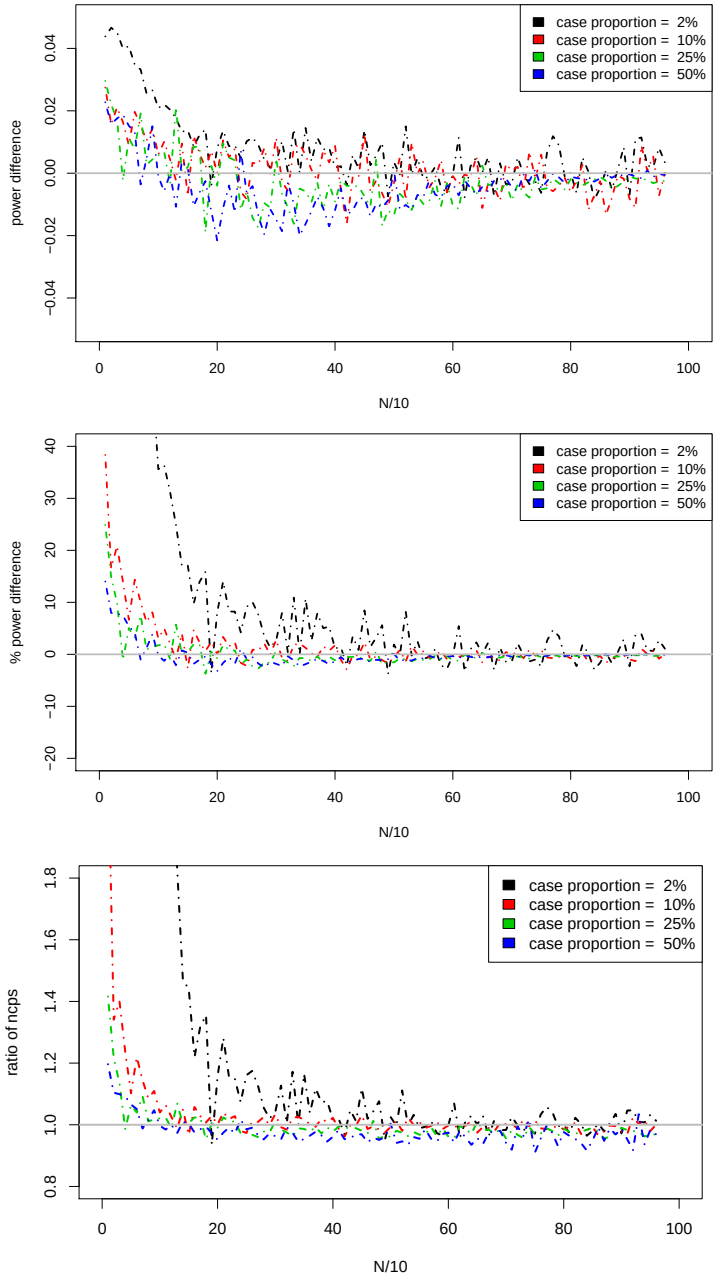


Figure 7: Accuracy of power calculations against sample size (N) on four given case proportions 2% (upper black), 10% (middle red), 25% (lower green) and 50% (lower blue), for simple logistic regression with β_1=0.33. The comparisons are made to power calculated from 5000 simulations per setting, giving accuracy of $\pm 1.5\%$ at power ≈ 0.9 or better. The accuracies are stated as absolute difference (top panel), percentage difference (middle panel) and (bottom panel) as the ratios of the non-centrality parameters of the relevant χ^2_1 distribution to the empirical value obtained by averaging $\hat{\beta}_1^2/\mathrm{Var}[\hat{\beta}_1]$ over the 5000 simulations. Calculations given reasonably large case proportions stabilize quickly against growing sample size, although with very small P the power calculations are less reliable.

and how it may grow with n, leaving us to consider only the regression coefficients β_0, β_1 etc. Particularly for multiple regressions, specifying the joint distribution of many covariates can be challenging in practice: the marginal distribution of each is likely known but their dependencies are usually less clear. In these situations, while still challenging, it may be preferable to allow investigators to supply a design matrix of realistic values, and (with specified β coefficients) extrapolate the power from there. This can be achieved by replacing the expectations in Section 1.2.2's Fisher information matrices by empricial averages over the dataset at hand—essentially replacing the true Fisher information with the observed Fisher information. From there we can obtain non-centrality parameters for the distribution of the corresponding Wald tests, as before, and with them approximate power at any given sample size n, with the implicit assumption that larger datasets would be drawn from the same data-generating mechanism for which the data at hand is a representative sample.

Naturally, this process can be conducted after data are available, using observed values of X as well as beforehand using hypothetical X values. Considerable caution is required when using actually-observed X, however: it is well-known that plugging estimates $\hat{\boldsymbol{\beta}}$ from the observed (Y, X) into power calculations based on the observed X uses the data twice, and can give misleading over-estimates of power. (See e.g. Hoenig and Heisey [2001].) Instead, considering the power that a particular analysis would have at different $\boldsymbol{\beta}$ values, chosen independently of the data at hand, is more useful for analysis. Using the term *retrospective design analysis* to describe it, Gelman and Carlin [2014] recommend this practice, in particular when apparently strong (statistically significant) evidence for non-null effects have been found. In what follows we therefore assume, throughout, that the values of $\boldsymbol{\beta}$ considered are chosen in this way.

4.1 Accuracy of S-B's method with empirical covariates

To illustrate the accuracy of power calculations from our empirical version of S-B's method, we first consider power based on a design where X_1 is a single sample of standard Normal covariates. We fix the effect size (i.e. log odds ratio) at $\beta_1 = 0.33$ and case proportion P at 0.02, 0.10, 0.25 and 0.50. For each setting we vary the sample size from $N = 50$ to 1000, with intervals of 10. Empirically estimated powers (using simulation of outcomes Y under the single simulated X) and the corresponding analytic approximation are shown together in Figure 8. While we cannot be certain that the behavior extends to all possible X samples, for the randomly-chosen ones considered here, we see close agreement between the approximation power calculations and the gold-standard simulation values.

We also evaluate the performance of the S-B method for empirical variables at different log odds ratios with fixed case proportion $p = 0.25$. The top panel in Figure 9 plots power against sample size: the bottom panel plots power against log odds ratio β_1. Sample size is again taken between $N = 50$ and 1000 and log odds ratios between 0.11 and 0.68. Again, there is close agreement between the power calculations and the true values obtained from simulation, though this accuracy does improve notably with larger sample sizes.

Extremely similar results, with good agreement between calculated power and the gold-standard values obtained by simulation. There are shown in Figures 10 and 11, for the situation where (as in Section 3) we generate X_1 and X_2 as multivariate Normal, each with mean zero and variance 1, and correlation $\rho = 0.5$ between them. We consider fixing β_1 at specific values between 0.11 and 0.68, fix $\beta_2 = 0.33$ throughout, fix n and select β_0 by forcing proportion $p = 0.25$ throughout. As before, accuracy is good throughout but improves with larger n.

In conclusion, we have shown proof-of-principle of our empirical extension of S-B's method. For situations when power for a specific design matrix is needed, it provides a simple and ac-

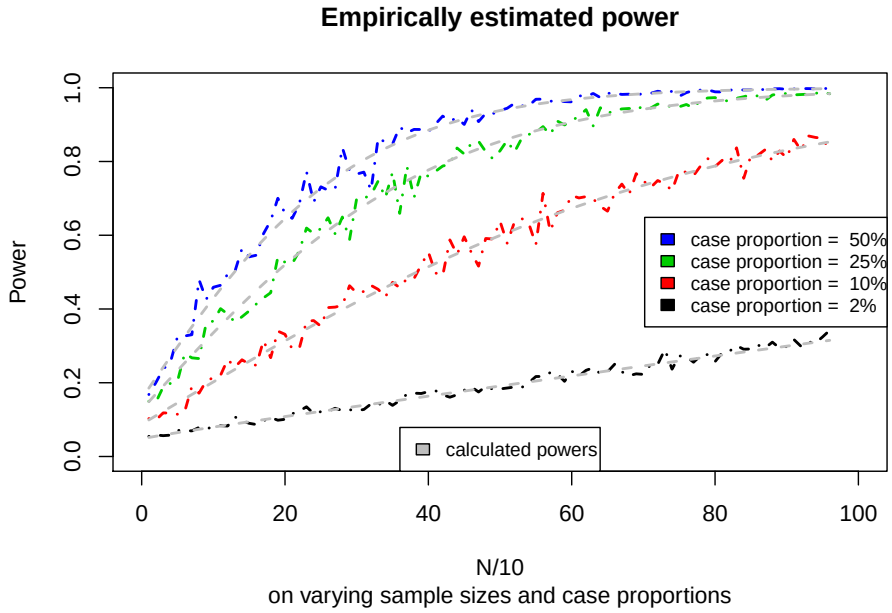


Figure 8: Empirically estimated power matches analytic power against varying sample size between 50 and 1000 given four case proportions between 2% and 50% for $\beta_1 = 0.33$.

curate solution. When covariate distributions are too complex for users to describe, using a representative design matrix may also be a way to help them understand, at least approximately, what power would be available.

5 Software for applications

The availability of user-friendly software is key for the widespread use of statistical methods. While we have seen that the S-B method has considerable advantages over its competitors, it lacks such an implementation. In this section we describe our attempts to fill this gap, presenting examples of our R package and Shiny web application, designed to provide a simple interface to S-B's method, and our empirical extension of it described in Section 4.

5.1 The R Package

In what follows we give output from a vignette, provided in the form of an .RMD file as part of our R package. This vignette covers three specific examples.

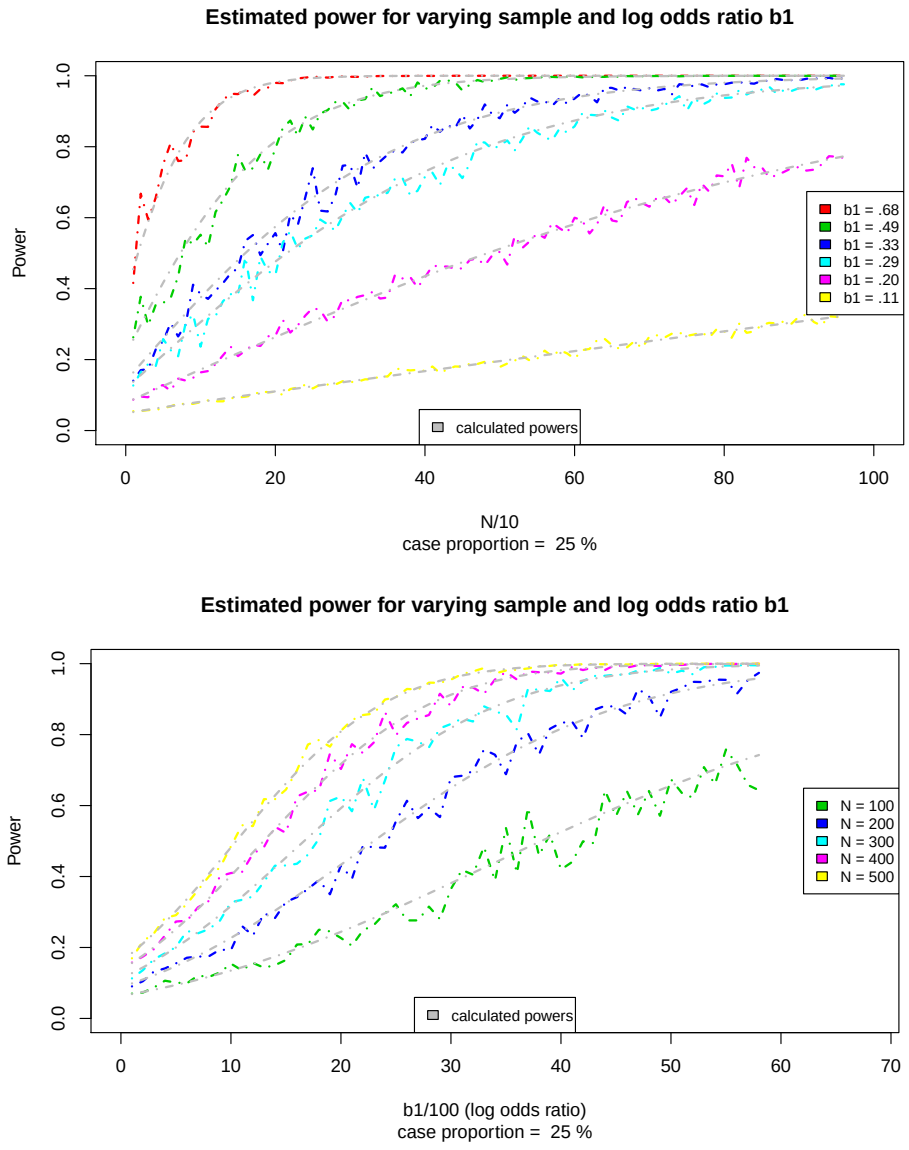


Figure 9: For a fixed case proportion at 25%, the hypothesized β_1 indicates magnitude of association, With six pre-specified hypothesized β_1, we observe empirically estimated powers approximate analytic power closely against varying sample sizes. The same is true as we provide a finer grid for hypothesized β_1 on the horizontal axis. Empirically estimated power matches analytic power are against varying hypothesized β_1 between 0.11 and 0.68. We select sample sizes between $N = 100$ and 500 to illustrate the trend.

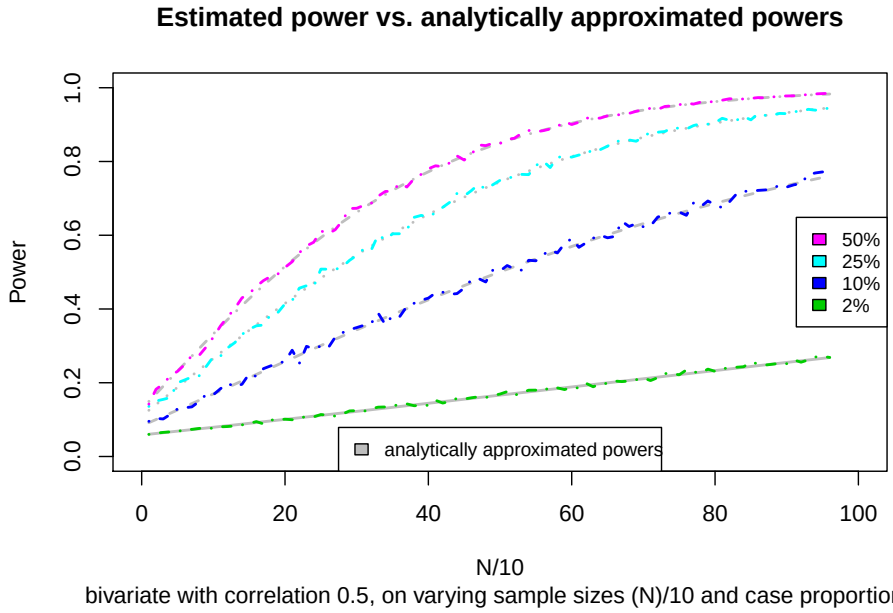
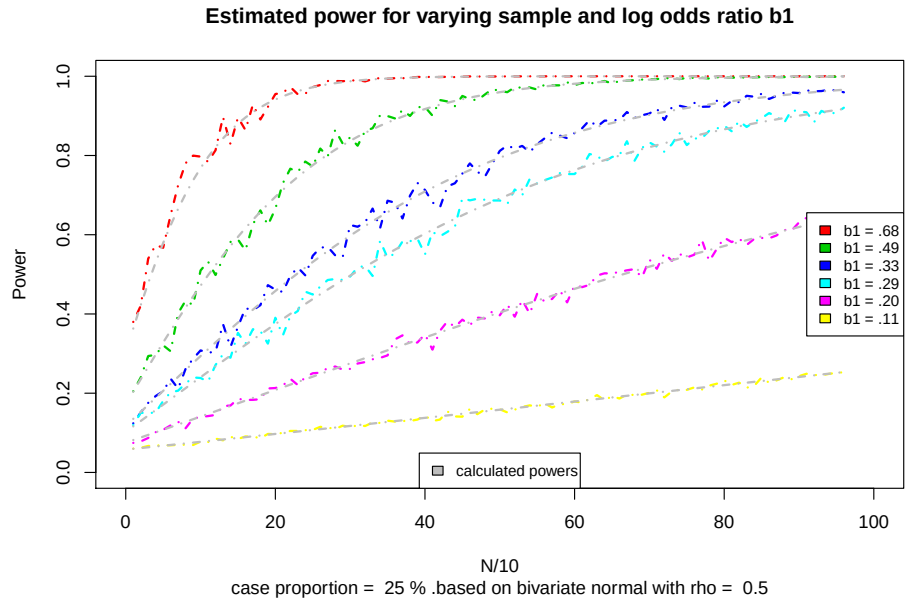


Figure 10: Adjusting for an additional variable X_2 with hypothesized $\beta_1 = \beta_2 = 0.33$ and $\rho_{x_1,x_2} = 0.5$, empirical powers match closely with analytic powers. Sample size again varys between $N = 50$ and 1000, whereas four fixed case proportions between 2% (bottom green) and 50% (top magenta) are included.

As a start, in Figure 12 we reproduce entry (6, 4) from Table 2 and entry (5, 4) from Table 5. The `calc_pwr()` function is used to calculate power for a simple (univariate) logistic regression based on the S-B method, and the `approx_pwr()` function is its bi-variate counterpart.

The second example is the based on the *NHANES* (National Health and Nutrition Examination Survey) dataset. We have a subset of 991 observations included in a prospective logistic regression analysis, of whether higher diastolic blood pressure (DBP) is associated with age. Higher DBP (an indicator) is defined for our purposes as DBP$\geq$70mm Hg. We also consider adjusting the regression for hypothesized confounder body-mass index (BMI). Power approximations from such regression models among a range of hypothesized log odds ratio β_1 are shown in Figure 13.

The package also contains a function `calc_sample_size_emp()` that inverts the power formulae to give minimum samples sizes required for a specified level of power. To illustrate it,



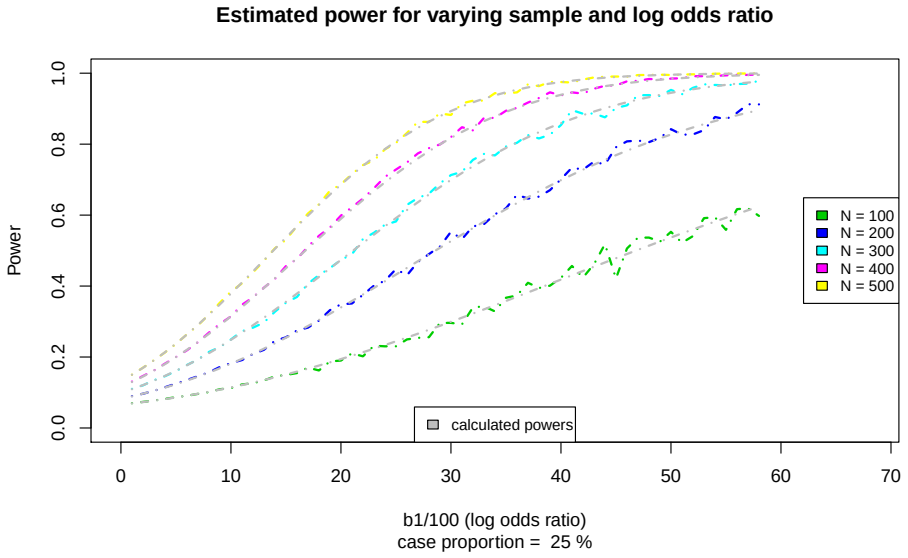


Figure 11: With the same adjusted model where $P = 0.25, \beta_2 = 0.33$ and $\rho = 0.5$, empirical powers coincide with analytic powers for varying sample sizes between $N = 50$ and 1000 (top) and a range of log odds ratio β_1 between 0.11 and 0.68 (bottom).

Example 1: S-B Table 1

```
calc_pwr(b0 = -1, b1 = 0.33, n = 500, dist = "normal", mu = 0, sigma = 1)
```

```
## [1] 0.8901227
```

```
calc_samp_size(b0 = -1, b1 = 0.33, targetpwr = 0.9, alpha = 0.05,
               dist = "normal", mu = 0, sigma = 1)
```

```
## [1] 517.2049
```

```
approx_pwr(b0 = -1, b1 = 0.33, b2 = 0.9, rho = 0.5, n = 500,
           dist = "normal", mu = 0, sigma = 1)
```

```
## [1] 0.7473011
```

```
approx_samp_size(b0 = -1, b1 = 0.33, b2 = 0.9, rho = 0.5, targetpwr = 0.9,
                 alpha = 0.05, dist = "normal", mu = 0, sigma = 1)
```

```
## [1] 761.8756
```

Figure 12: We show two pairs of functions for power and sample size calcluations. For a simple regression $N = 500$ with a standard normal ($\mu = 0, \sigma = 1$) variable of interest, $\beta_0 = -1, \beta_1 = 0.33, P = 0.27$ and $\alpha = 0.05$, using `calc_pwr()`, the analytic power is a little lower than 0.90. On the other hand, to achieve an exact power of at least 0.90, `calc_samp_size()` gives a minimal sample size of 518. Similarly, adding a covariate with hypothesized effect size $\beta_2 = 0.9$ and correlation between variables $\rho = 0.5$, the analytic power calculated using `approx_pwr()` is 0.75, whereas a sample size of 762 is required to achieve a target power of 0.9 based on `approx_samp_size()`.

our third example evaluates the association between functional SNPs associated with muscle size and strength. The outcome measures is a binary indicator of whether the percentage change in muscle strength in non-dominant arm is greater than 60%, comparing measurements before and after training. The SNP of interest is r577x, where we code genotype as 0/1/2, for the number of copies of the minor allele. For a univariate sample size calculation, given a range of hypothesized β_1 for the logistic-linear trend estimated by the regression, a specific power target, and nominal level α, Figure 14 shows how large the sample size would need to be to provide at least 80% power. The intercept term β_0 is calculated using the `findb0emp()` function for a range of hypothesized effect sizes β_1 and a fixed case proportion P.

Finally, while not described here in detail, for the convenience of users we have also included routines that compute minimum detectable effect sizes for a given design as in Figure 15, at given power, sample size, and level. These essentially invert the power function, solving for

Example 2: NHANES The data set

```
names(nhanes)
```

```
##  [1] "BPXSAR"    "BPXDAR"    "BPXDI1"    "BPXDI2"    "race_ethc"
##  [6] "gender"    "DR1TFOLA"  "RIAGENDR"  "BMXBMI"    "RIDAGEYR"
```

```
nhanes <- nhanes[complete.cases(nhanes[,c("BMXBMI","RIDAGEYR","BPXDAR")]),]
nhanes$HighDBP = ifelse(nhanes$BPXDAR>70, 1, 0)
nhanes$BMXBMI = as.numeric(nhanes$BMXBMI)
nhanes$RIDAGEYR = as.numeric(nhanes$RIDAGEYR)

mod = glm(HighDBP ~ RIDAGEYR, data = nhanes)
mod.adj = glm(HighDBP ~ RIDAGEYR + BMXBMI, data = nhanes)

prop_nhanes = sum(nhanes$HighDBP==1)/length(nhanes$HighDBP)
```

Power curves

```
granu = 100
lowerb = -0.1
upperb= 0.15
b1_hypos = seq(lowerb, upperb, length.out = granu)
xmat_nhanes_univ = cbind(rep(1, nrow(nhanes)), nhanes$RIDAGEYR)
b0_hats = sapply(X = b1_hypos, FUN = findb0emp, prop=prop_nhanes,
                 xmat = xmat_nhanes_univ, lower = -1, upper = 1)
powers_univ = vector(length = granu)
for (i in 1:length(powers_univ)){
  pwr_i = calc_pwr_emp(b0 = b0_hats[i], b1 = b1_hypos[i], xmat = xmat_nhanes_univ,
                   alpha =.05, reg = "uni")
  powers_univ[i] = pwr_i
}
plot(x = b1_hypos, y = powers_univ, type = "l", ylim = c(0, 1), xlim = c(-.05, .075),
     xlab = "log odds ratio", ylab = "calculated power",
     main = paste0("Power vs. Log Odds Ratio When Sample Size (N) = ", nrow(nhanes)))
```

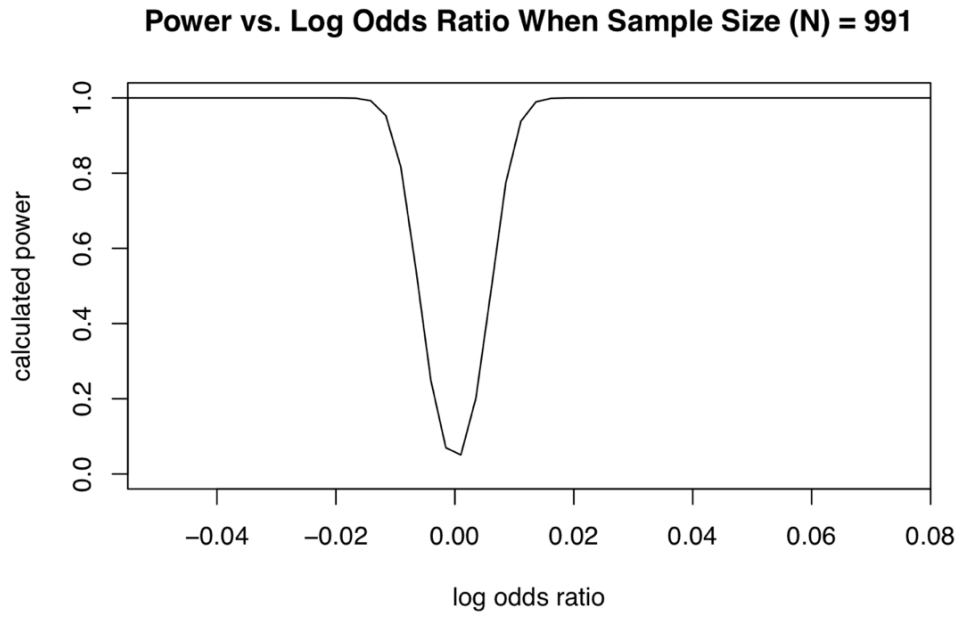


Figure 13: Power calculations based on a subset of the NHANES data set (N=991) for a simple (upper panel) and an adjusted (lower panel) logistic regression models. Powers are approximated empirically using the `cal_pwr_emp()` function where regression is specified as `uni` or `multi`, respectively. The required inputs are a vector of hypothesized `b1` and its corresponding intercept `b0`, the design matrix `xmat` and specified significance level `alpha`.

```r
granu.b = 100
lowerb.b = 0.01
upperb.b = 1
b1.hypos.b = seq(lowerb.b, upperb.b, length.out = granu.b)
xmat.fms.univ = cbind(rep(1, nrow(fms2)), fms2$Geno)
b0.hats.b = sapply(X = b1.hypos.b, FUN = findb0emp, prop=prop.fms2,
                   xmat = xmat.fms.univ, lower = -1, upper = 1)
b2.hypo.b = -0.33
xmat.fms.biv = cbind(rep(1, nrow(fms3)), fms3$Geno, fms3$Male)
powers.univ.b = vector(length = granu.b)
for (i in 1:length(powers.univ.b)){
  pwr.i = calc_pwr_emp(b0 = b0.hats.b[i], b1 = b1.hypos.b[i], xmat = xmat.fms.univ,
                   alpha =.05, reg = "uni")
  powers.univ.b[i] = pwr.i
}
```

Sample size curves

```r
target.power.b = .80
# BinaryTrait ~ Genotype
ss.univ.b = vector(length = granu.b)
for (i in 1:length(powers.univ.b)){
  ss.i = calc_samp_size_emp(b0 = b0.hats.b[i], b1 = b1.hypos.b[i], xmat = xmat.fms.univ,
                   targetpwr = target.power.b, reg = "uni")
  ss.univ.b[i] = ss.i
}
plot(x = b1.hypos.b, y = ss.univ.b, type = "l", xlim = c(0.2, 0.65), ylim = c(0, 1800),
     xlab = "log odds ratio", ylab = "calculated sample size",
     main = paste0("Sample Size vs. Log Odds Ratio When Target Power = ", target.power.b))
```

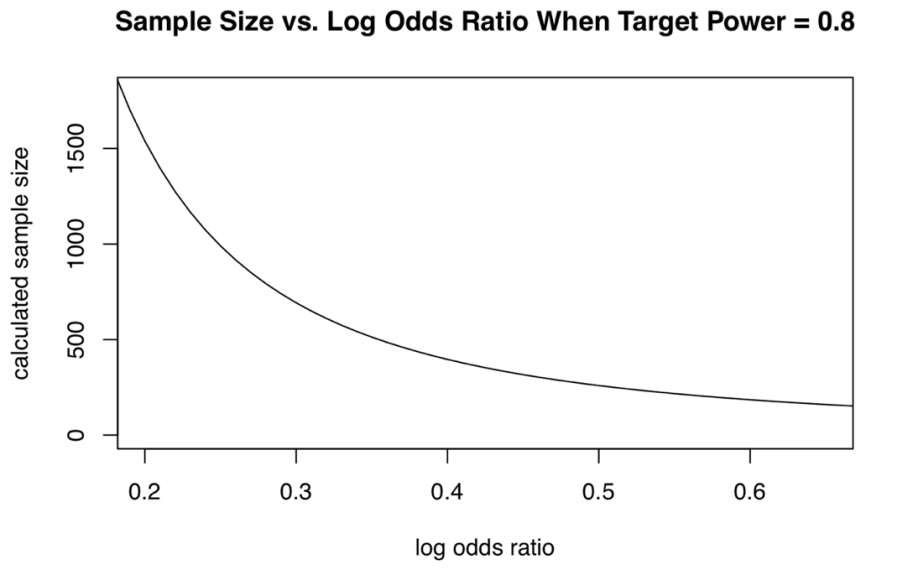


Figure 14: Sample size calculations for a univariate analysis regressing muscle strength improvement to genotype specified by SNP r577x using `cal_samp_size_emp()`. Function parameters are specified in a similar fashion as in Figure 13. For a hypothesized β_1 between 0.2 and 0.65, the required sample size to achieve an 80% power is between $N = 1539$ and 160.

the value of the effect size where power meets the desired value.

5.2 The Shiny App

For users for whom programming in R is a burden, even with pre-packaged functions, we have written a Shiny web application that provides the most important functionality of our R package. This enables power and sample size calculation to be performed with no coding whatsoever; users simply fill in form entries and check boxes indicating the calculation they want to perform. The Shiny Application can be found at https://zoray.shinyapps.io/powerlog/.

Here we provide examples of its use, by which we hope to illustrate its practicality as well as describing the intended use of its options.

We first describe the distribution-based approximations, which implement the original S-B method. The distributions we provide for the univariate case are listed at the top of the screen.

In Figure 16, we estimate power for a sample size $N = 2005$ in a simple logistic regression model assuming a standard Normal distribution for the covariate. The hypothesized log odds ratio per unit difference in X (i.e. effect size) is between $\beta_1 = 0$ and 1. The case proportion is $P = 50\%$, and significance level $\alpha = 0.05$. We note that the two top left boxes are unchecked since we are not adjusting for an additional variable, nor supplying the design matrix from data. With these settings, if the effect size is at least $\beta_1 = 0.126$ (OR $= 1.134$), using S-B's method, the sample size will give power of at least 80%.

In Figure 17 we instead consider sample size, for a for a simple logistic regression with hypothesized log odds ratio β_1 between 0.13 and 0.33. The variable of interest is assumed to follow a t-distribution centered at 0 with 10 degree of freedom, shifted and scaled to have mean zero and variance 1. After setting the desired power target and significance level, the minimum detectable log odds ratio is calculated to be 0.116 (OR $= 1.123$) for a maximum available sample size of 2019.

```r
b2_hypo = 0.013
b2_hypos = rep(b2_hypo, granu)
xmat_nhanes_biv = cbind(rep(1, nrow(nhanes)), nhanes$RIDAGEYR, nhanes$BMXBMI)

powers_biv = vector(length = granu)
for (i in 1:length(powers_biv)){
  b0_hat = findb0emp(b1 = b1_hypos[i], b2 = b2_hypos[i], prop=prop_nhanes,
                     xmat = xmat_nhanes_biv, reg = "multi", lower = -1, upper = 1)
  pwr_i = calc_pwr_emp(b0 = b0_hat, b1 = b1_hypos[i], b2 = b2_hypos[i],
                       xmat = xmat_nhanes_biv, alpha =.05, reg = "multi")
  powers_biv[i] = pwr_i
}

b1hat2u = calc_effect_size_emp(xmat=xmat_nhanes_biv, prop=prop_nhanes,
                               b2=b2_hypo, targetpwr=target_power,
                               alpha=0.05, reg = "multi", lower=0, upper=0.15)
b1hat2l =  calc_effect_size_emp(xmat=xmat_nhanes_biv, prop=prop_nhanes,
                               b2=b2_hypo, targetpwr=target_power,
                               alpha=0.05, reg = "multi", lower=-.10, upper=0)
par(mar = c(5, 4, 5, 2) + 0.1)
plot(x = b1hat2u, y = target_power, type = "p", col = "forestgreen", pch = 16,
     xlab = "log odds ratio", ylab = "calculated power",
     ylim = c(0, 1), xlim = c(lowerb, max(upperb, b1hat2u*1.1)),
     main = paste0("Power vs. Log Odds Ratio When Sample Size (N) = ", nrow(nhanes)))
points(x = b1hat2l,y = target_power, type = "p", col = "forestgreen", pch = 16 )
lines(x = b1_hypos, y = powers_biv)
abline(h = target_power, col = "hotpink", lty = 2)
axis(side = 3, at = b1hat2u, labels = signif(b1hat2u, 3))
axis(side = 3, at = b1hat2l, labels = signif(b1hat2l, 3))
abline(v = b1hat2u, col = "forestgreen", lty = 2)
abline(v = b1hat2l, col = "forestgreen", lty = 2)
```

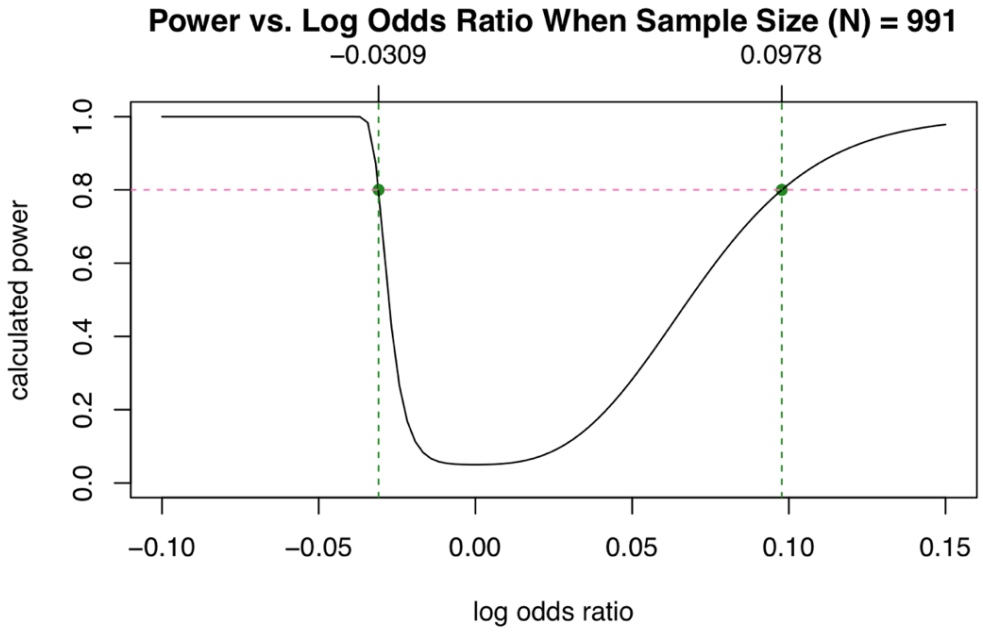


Figure 15: Effect size (log odds ratio) calculation for the association between high DBP and age (in years) adjusted for BMI. For a two-sided test, the range of searchable β_1 is carefully specified from 0 to 0.15 and from -0.1 to 0 according to reasonable guess based on the power curve. We find detectable log odds ratios around or above 0.0978 (OR = 1.10) and around or below −0.0309 (OR = 0.970) for a 80% power target.

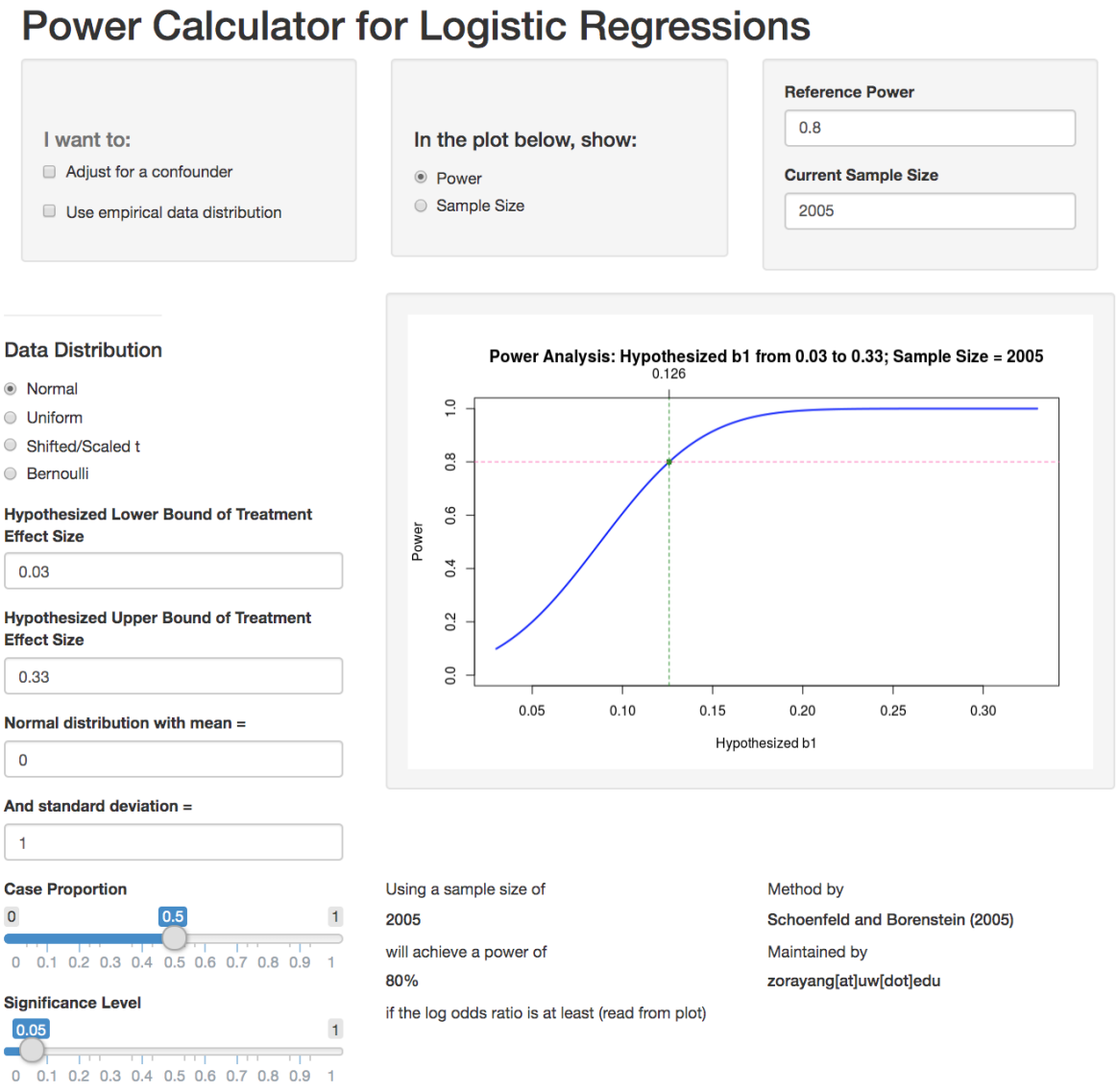


Figure 16: The Shiny app named *powerlog* is an interactive interface for our power and sample size calculator. We recommend starting from the top panels, which offer specification on simple vs. adjusted analysis and power vs. sample size approximations. In this example, we estimate power for a sample size $N = 2005$ in a simple logistic regression model assuming a standard Normal distribution for the covariate. The hypothesized log odds ratio is between $\beta_1 = 0$ and 1. The case proportion is $p = 50\%$, and significance level $\alpha = 0.05$. With these settings, if the log odds ratio is at least $\beta_1 = 0.126$ (OR $= 1.134$), the sample size will give power of at least 80%.

Figure 17: With a maximum available sample size of $N = 2019$ for a univariate logistic regression based on a t-distribution centered at 0 with a degree of freedom of 10, the minimum detectable log odds ratio is $\beta_1 = 0.116$ (OR $= 1.123$) given a range of hypothesized β_1 between 0.03 and 0.33. The left hand side of the sample size curve is rather steep, meaning that a slighter smaller β_1 corresponds to a quite larger required sample size for the same power. The case proportion is set at $p = 50\%$ and level at $\alpha = 0.05$ for a target power of 82%.

Our Shiny web application also permits users to upload their own design matrix, and use it empirically as described in Section 4. In the example of Figure 18, we upload a machine-generated data set where the variable of interest and adjustment variable follow a multivariate-Normal distribution. Given again a range of hypothesized log odds ratios, the power of such a logistic regression is given in the plot on the right .

Due to its menu-based interface, the Shiny web application can only provide a limited set of analyses compared to the full R package, and indeed compared to the full capacity of S-B's method to handle multiple covariates. However, as it gives users the power/sample size calculations and graphs they most often need quickly and without programming or simulation, it should nevertheless provide a useful service in practice.

6 Conclusion

In this thesis we have implemented three approaces for calculating power and sample size for logistic regression. The implementations are based on Schoenfeld and Borenstein [2005] with reference to ideas presented by Whittemore [1981], Hsieh [1989], Hsieh et al. [1998] and Demidenko [2007]. These calculations are applicable to simple ordinary logistic regression in which an outcome is regressed against a single variable of interest, possibly following a distribution in a few families. We included discussion on the Normal, t-, uniform and Rademacher distributions for that variable of interest. These calculations have also been extended to analytic approximations for multiple logistic regression where we present explicit formulations for multivariate Normal covariates. In evaluating the methods, we showed that S-B performs best compared to competitors in almost all practical settings, and does so consistently across covariate distributions of different tail weights. We also found that sample sizes of $n \geq 200$ are sufficient for S-B's method to be highly accurate in all the settings we examined, and in

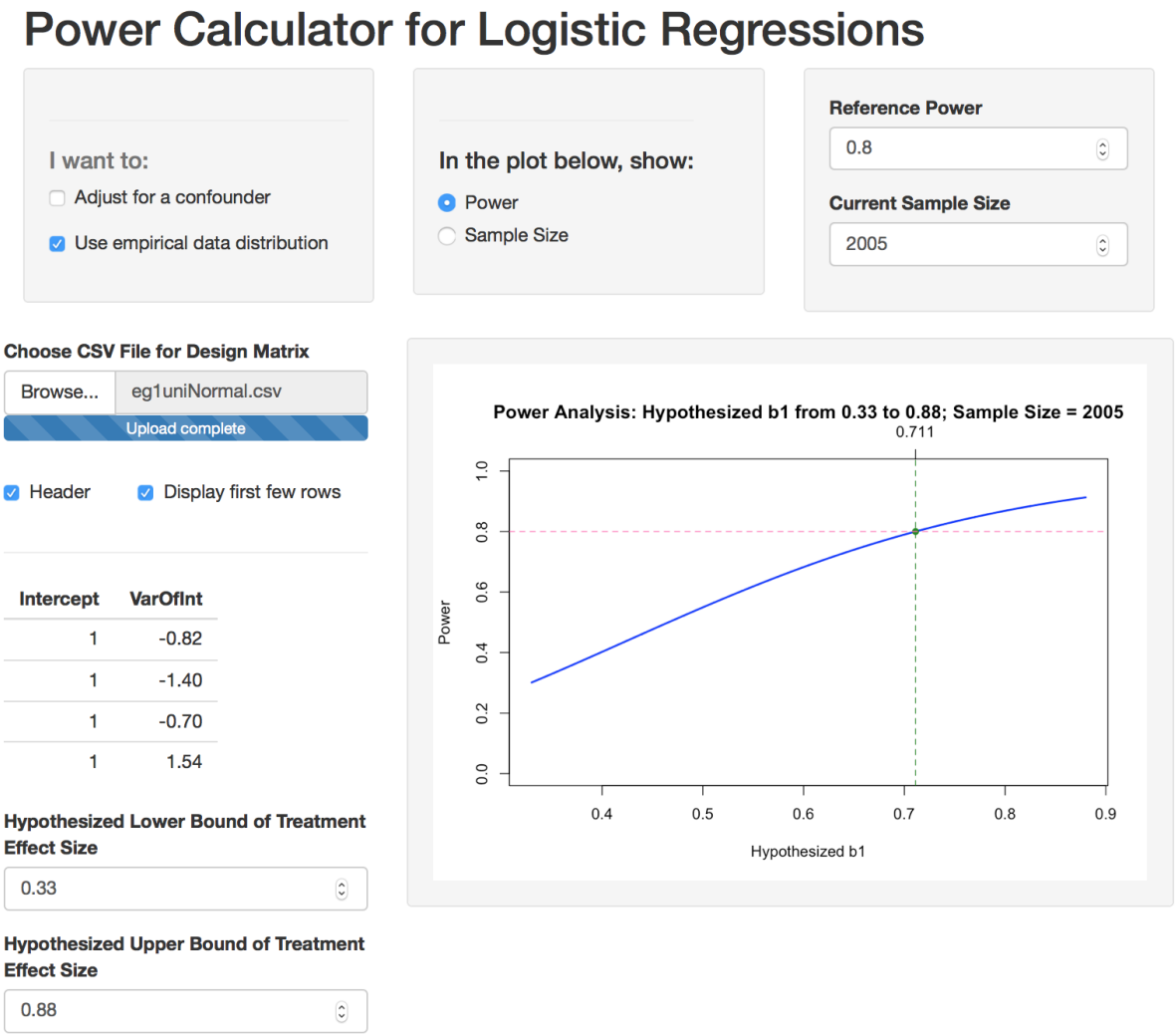


Figure 18: Power curve for a univariate Normal logistic regression from a data set. Design matrix is uploaded and its first few rows displayed on the left. The log odds ratio β_1 is hypothesized to range from 0.33 and 0.88. With a current sample size of 2005, an hypothesized log odds ratio $\beta_1 = 0.711$ corresponds to a reference power of 80% in this adjusted model.

many situations smaller samples will suffice. Rare outcomes and binary covariates were the situations where accuracy was most difficult to attain. We have also shown how the method can be implemented in user-friendly code, as both an R package and a Shiny web application.

For situations where the accuracy of the S-B method is challenged, with further work beyond the scope of this thesis more accurate (or even exact) power calculations could be achieved, either through complete enumeration or some close approximation of it – for example the approach of [Sondhi and Rice, 2018], where extremely rare outcomes are omitted from enumeration. Analytic approximations adapting the S-B evaluation of Fisher information expectations for non-Normal covariates could also be explored, at least for some families of covariate distributions. A further possible extension would be to generalize the S-B method to give power not only for model-based inference via Fisher information, but 'robust' inference that requires fewer assumptions. The required non-centrality parameters involve expectations of information matrices, but are not dramatically more complex than those presented here. However, as switching between model-based and robust inference for logistic regression analyses typically makes little difference in practice unless egregious model violations are present or cases are very rare, the impact on power of choosing model-based or robust inference should be expected to be minor. This can also be interpreted as an indicator that the present calculations will often suffice even when the modeling assumptions do not hold exactly in practice.

References

BM Bennett and P Hsu. On the power function of the exact test for the 2× 2 contingency table. *Biometrika*, 47(3/4):393–398, 1960.

Norman E Breslow, Nicholas E Day, et al. *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.* Number 32 in IARC Scientific Publications. Distributed for IARC by WHO, Geneva, Switzerland, 1980.

AG Dean, KM Sullivan, and MM Soe. Openepi: open source epidemiologic statistics for public health, version, 2014.

Eugene Demidenko. Sample size determination for logistic regression revisited. *Statistics in medicine*, 26(18):3385–3397, 2007.

Eugene Demidenko. Sample size and optimal design for logistic regression with binary interaction. *Statistics in medicine*, 27(1):36–46, 2008.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

Michael P. Fay. Two-sided exact tests and matching confidence intervals for discrete data. *R Journal*, 2(1):53–58, 2010. URL https://journal.r-project.org/.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

JA Hanley and EEM Moodie. Sample size, precision and power calculations: a unified approach. *J Biomet Biostat*, 2(124):2, 2011.

Walter W Hauck and Allan Donner. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360a):851–853, 1977.

John M Hoenig and Dennis M Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Fushing Y Hsieh, Daniel A Bloch, and Michael D Larsen. A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine*, 17(14):1623–1634, 1998.

FY Hsieh. Sample size tables for logistic regression. *Statistics in medicine*, 8(7):795–802, 1989.

S Jones, S Carley, and M Harrison. An introduction to power and sample size estimation. *Emergency medicine journal: EMJ*, 20(5):453, 2003.

David G Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. Springer Science & Business Media, 2010.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

Robert H Lyles, Hung-Mo Lin, and John M Williamson. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine*, 26(7):1632–1648, 2007.

Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

nQuery. *Sample Size and Power Calculation*. "Statsols" (Statistical Solutions Ltd, Cork, Ireland, 2017.

PASS. *Power Analysis and Sample Size Software*. NCSS, LLC, Kayville, Utah, 2019.

Chao-Ying Joanne Peng, Haiying Long, and Serdar Abaci. Power analysis software for educational researchers. *The Journal of Experimental Education*, 80(2):113–136, 2012. doi: 10.1080/00220973.2011.647115. URL https://doi.org/10.1080/00220973.2011.647115.

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

Bernard Rosner. *Fundamentals of biostatistics*. Nelson Education, 2015.

Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

David A Schoenfeld and Michael Borenstein. Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation*, 75(10):771–785, 2005.

Steven G Self and Robert H Mauritsen. Power/sample size calculations for generalized linear models. *Biometrics*, pages 79–86, 1988.

S Senn. Rare distinction and common fallacy. *British Journal of Medicine*, 1999.

Arjun Sondhi and Kenneth Rice. Fast permutation tests and related methods, for association between rare variants and binary outcomes. *Annals of Human Genetics*, 82(2):93–101, 2018.

Gerald Van Belle, Lloyd D Fisher, Patrick J Heagerty, and Thomas Lumley. *Biostatistics: a methodology for the health sciences*, volume 519. John Wiley & Sons, 2004.

Eric Vittinghoff, Ś Sen, and CE McCulloch. Sample size calculations for evaluating mediation. *Statistics in medicine*, 28(4):541–557, 2009.

Alice S Whittemore. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, 76(373):27–32, 1981.