Power Analysis for Logistic Regression Models Fit to Clustered Data: Choosing the Right Rho

Steve Gregorich
May 16, 2014

Abstract

Context

Power analyses for logistic regression models fit to clustered data

Approach

. estimate *effective* sample size (N_{eff}: cluster-adjusted total sample size)

. input N_{eff} into standard power analysis routines for independent obs.

Wrinkle

. in the context of logistic regression there are two general approaches to estimating the intra-cluster correlation of *Y*:

. phi-type coefficient and

. tetrachoric-type coefficient.

Resolution

. The phi-type coefficient should be used when calculating N_{eff}

I will present background on this topic as well as some simulation results

Simple random sampling (SRS)

. Fully random selection of participants

e.g., start with a list, select *N* units at random

. Some key features wrt statistical inference:

representativeness

all units have equal probability of selection

all sampled units can be considered to be independent of one another

. SRS with replacement versus without replacement

Clustered sampling

. Rnd sample of *m* clusters; rnd sample of *n* units w/in each cluster
multi-stage area sampling
patients within clinics

. Repeated measures
Random sample of *m* respondents; *n* repeated measures are taken
repeated measures are clustered within respondents

. Typically, elements within the same cluster are more similar to each other than elements from different clusters

. The *n* units w/in a cluster usually do not contain the same amount of info wrt some parameter, θ, as the same number of units in an SRS sample
…the concept of effective sample size, N_{eff}…

Therefore, it is usually true that $\sigma^2_{\text{clus}}\left(\hat{\theta}\right) \neq \sigma^2_{\text{srs}}\left(\hat{\theta}\right)$

Two-stage clustered sampling design

Unless otherwise noted, I assume

. Clustered sampling of m clusters, each with n units:

$N = m \times n$

. Normally distributed unit-standardized x, binary y

exchangeable / compound symmetric correlation structure

$\rho_y > 0$: intra-cluster correlation of y (outcome) response

$\rho_x = 0$ or 1: intra-cluster correlation of x (explanatory var) response

. Regression of y onto x via

. a mixed logistic model with random cluster intercepts or

. a GEE logistic model

. Common effects of x across clusters, i.e., no random slopes for x

. Common between- and within-cluster effects of x

The design effect, *deff*

. *deff* can be thought of as a design-attributable multiplicative change in variation that results from choice of a clustered sampling versus an SRS design

$\widehat{deff} = \frac{\sigma^2_{\text{clus}}(\hat{\theta})}{\sigma^2_{\text{srs}}(\hat{\theta})}$ and $\widehat{N_{\text{eff}}} = \frac{N}{\widehat{deff}}$, where

$\sigma^2_{\text{clus}}\left(\hat{\theta}\right)$ is the estimated parameter variation given a clustered sampling design;

$\sigma^2_{\text{srs}}\left(\hat{\theta}\right)$ is the estimated parameter variation given a SRS design;

N is the common size of the SRS and clustered ($N=m\times n$) samples;

$\hat{N}_{\text{eff}}$ estimated effective size of the clustered sample wrt information about $\hat{\theta}$, relative to what would have been obtained with a SRS of size N

Assumes compound symmetric covariance structure of the response

The misspecification effect, *meff*

Conceptually similar to *deff* except that the multiplicative change corresponds to the effect of correctly modeling the clustering of observations versus ignoring the cluster structure

$\widehat{meff} = \frac{\sigma^2_{\text{clus}}(\hat{\theta})}{\sigma^2_{\cancel{\text{clus}}}(\hat{\theta})}$ and $\widehat{N_{\text{eff}}} = \frac{N}{\widehat{meff}}$, where

$\sigma^2_{\text{clus}}\left(\hat{\theta}\right)$ is the estimated parameter variation given clustered responses;

$\sigma^2_{\cancel{\text{clus}}}\left(\hat{\theta}\right)$ is the estimated parameter variation ignoring clustering of responses;

N is the total size of the clustered sample;

$\hat{N}_{\text{eff}}$ is the effective size of the clustered sample wrt information about $\hat{\theta}$, relative to what would have been obtained with a SRS of the same size

Assumes compound symmetric covariance structure of the response

deff, *meff*, and the sample size ratio

A 'context free' label for *deff* and *meff* is the sample size ratio, SSR

$$\text{SSR}=\frac{N}{\hat{N}_{\text{eff}}}$$

. *deff*, *meff*, and SSR have equivalent meaning wrt power analysis,
but *deff* and *meff* are conceptually distinct

. *deff* assumes that you are considering SRS versus clustered sampling

. *meff* assumes that you have chosen a clustered sampling design and
want to make adjustments to an analysis that assumed SRS

. I will use *meff* for this talk

Estimating *meff* via the intra-cluster correlation

. Given positive intra-cluster correlation of *y*: $\rho_y > 0$,

the *meff* estimator depends on ρ_x

#1. Level-2 (cluster-level) *x* variables will have zero *within*-cluster variation and $\rho_x = 1$

$$\rho = \frac{\sigma^2_{\text{btw}}}{\left(\sigma^2_{\text{btw}} + \sigma^2_{w/in}\right)}$$

. In this case

$$\widehat{meff} = \frac{\sigma^2_{\text{clus}}(\hat{\theta})}{\sigma^2_{\text{clus}}(\hat{\theta})} = \frac{N}{\widehat{N_{\text{eff}}}} = 1 + (n-1)\rho_y,$$

. note: when estimating $\bar{y}$, assume $\rho_x = 1$

Estimating *meff* via the intra-cluster correlation

#2. Consider a level-1 stochastic *x* variable with positive within-cluster variation and zero between-cluster variation: $\rho_x = 0$:

$$\rho = \frac{\sigma^2_{\text{btw}}}{\left(\sigma^2_{\text{btw}} + \sigma^2_{w/in}\right)}$$

. In this case

$$\widehat{meff} = \frac{\sigma^2_{\text{clus}}(\hat{\theta})}{\sigma^2_{\text{clus}}(\hat{\theta})} = \frac{N}{\widehat{N_{\text{eff}}}} \approx 1 - \rho_y^{(n/(n-1))}$$

note: $n/(n-1) \to 1$ as $n \to \infty$

(In this talk, I do not cover Level-1 *x* variables with $0 < \rho_x < 1$)

Power analysis for clustered sampling designs using *meff*: Option 1

Option 1. Given a chosen model, power, and alpha level, plus a proposed clustered sample of size N=m×n, and a meff estimate

. $\widehat{N_{\text{eff}}} = \frac{N}{\widehat{meff}}$

. Use standard power analysis software, plug in $\widehat{N_{\text{eff}}}$ (instead of *N*), and estimate

Power analysis for clustered sampling designs using *meff*: Option 1 Example

Estimate Power by Simulation

. Simulate data from a CRT with 100 clusters (j) and 30 individuals/cluster (i)

$$y_{ij} = \text{group}_j \mathbf{0.5} + u_j + e_{ij}$$

where, VAR(u_j) = VAR(e_{ij}) = 1,
VAR(u_j) + VAR(e_{ij}) = **2**, and
$\rho_y = \sigma_u^2/(\sigma_u^2 + \sigma_e^2) =$ **0.50**

needed later for PASS

. Linear mixed model results from analysis of 2000 replicate samples

. $\rho_y =$ **0.501**
. residual std dev = **1.416** $\approx \sqrt{2}$
. $\hat{b}_{\text{group}} =$ **0.495**

all relatively unbiased

. simulated power for group effect: 67.7%

Power analysis for clustered sampling designs using *meff*: Option 1 Example

. Simulation result: power = 67.7%

. Use PASS Linear Regression routine to solve for power

. $\widehat{meff} = 1 + (30 - 1) \times \mathbf{0.501} = 15.529$

. $\widehat{N_{\text{eff}}} = 100 \times 30 \div 15.529 \approx 193$

.specify 193 as *N* in PASS

. specify H_1 slope = **0.495**

. specify Residual Std Dev = **1.416** (resid. @ level-1 plus level-2)

. PASS result: power = 67.6%

Summary

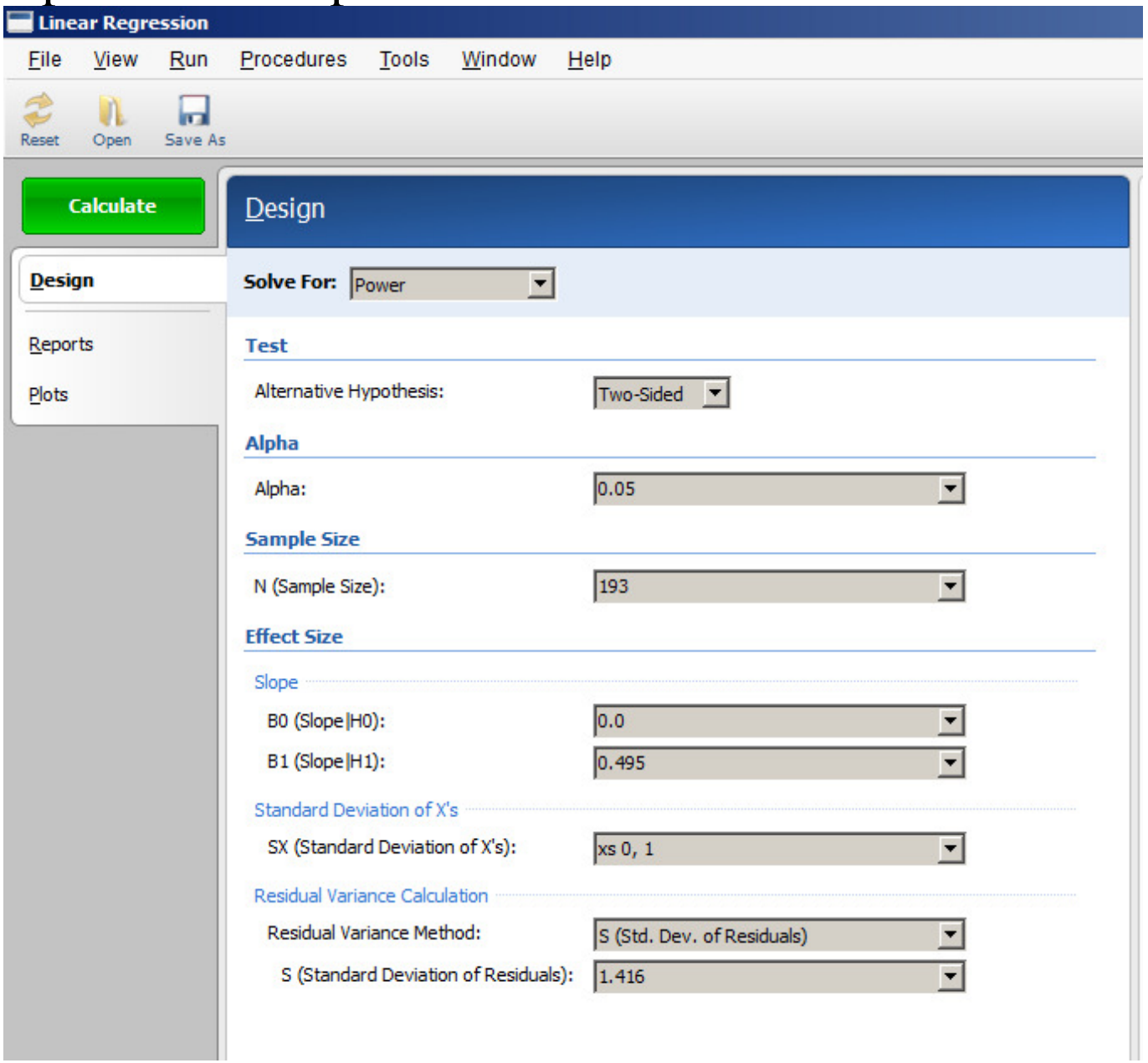
. choose *meff* estimator and estimate *meff*

. estimate N_{eff}

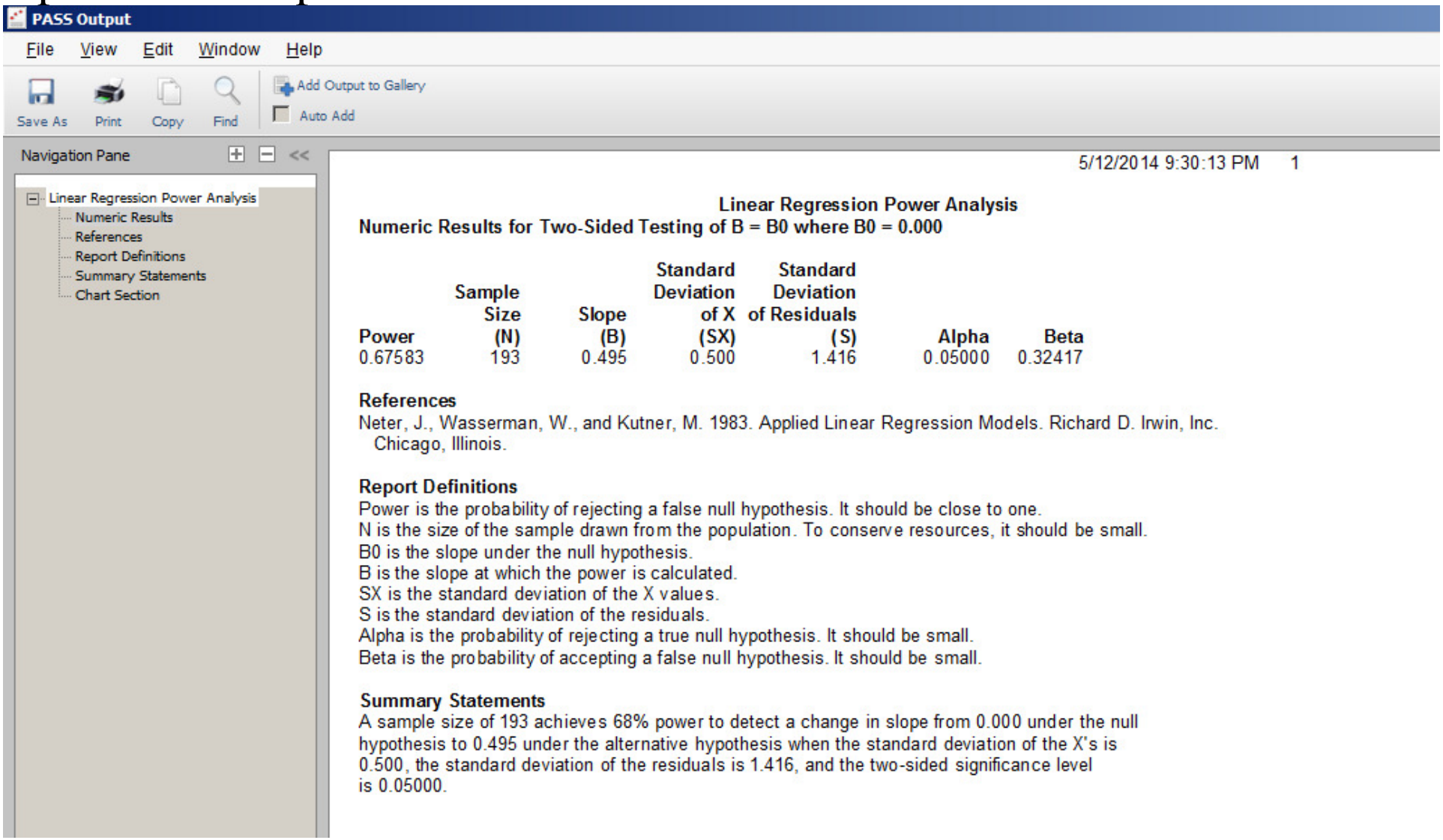
. plug N_{eff} into power analysis software (w/ other parameters)

. estimate power

Power analysis for clustered sampling designs using *meff*: Option 1 Example



Power analysis for clustered sampling designs using *meff*: Option 1 Example



PASS: power = 67.6%

Simulation: power = 67.7%

Power analysis for clustered sampling designs using *meff*: Option 2 example

Option 2. Given a clustered sample design, chosen model, power, and alpha level, plus an effect size estimate and a meff estimate

. Use standard power analysis software to estimate required sample size assuming independent observations, i.e., N_{eff}. Then estimate required N

. $\widehat{N} = \widehat{N_{\text{eff}}} \times \widehat{meff}$

Option 2: Step 1

Start with…

. the group effect (b=**0.495**),

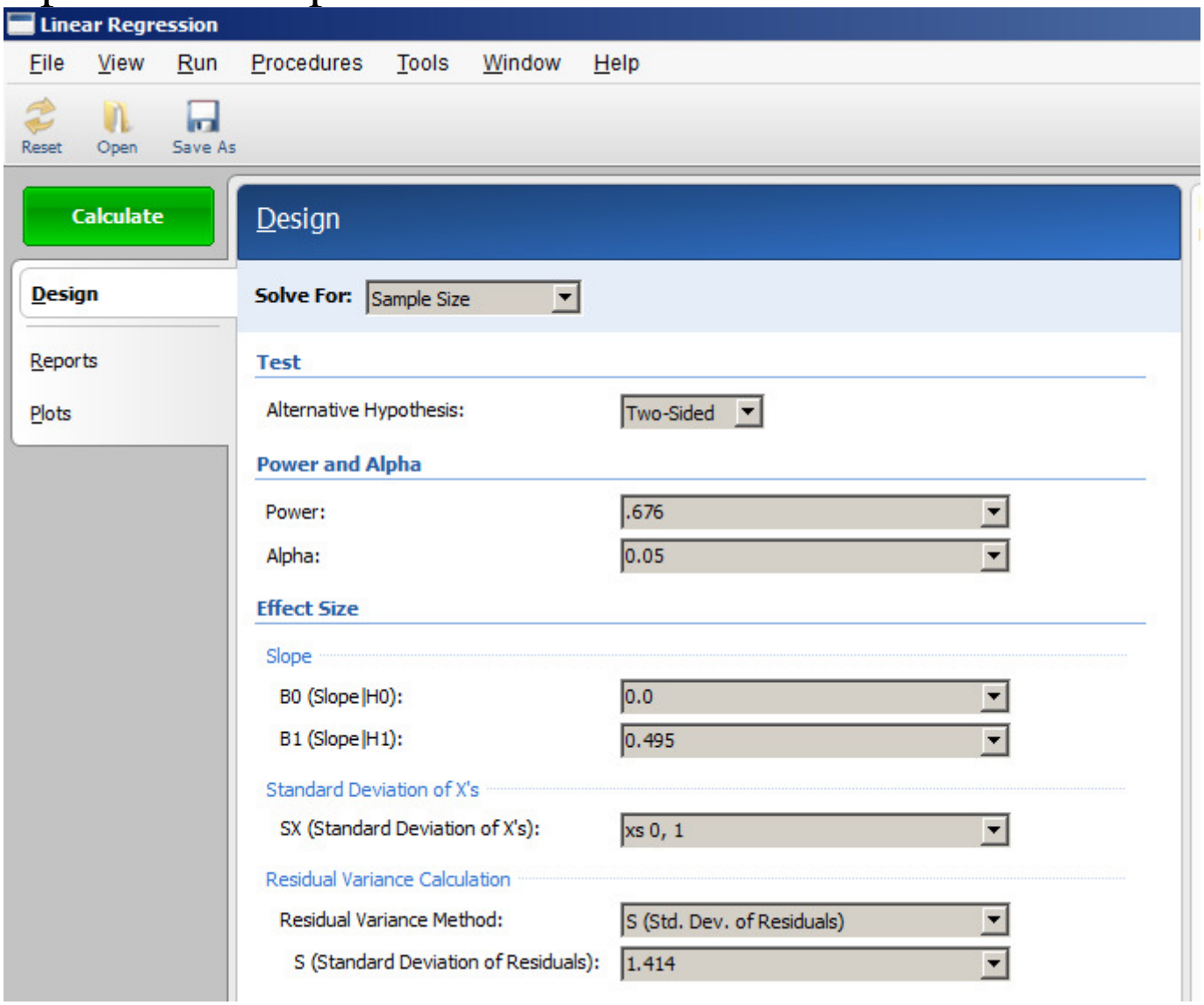
. a residual standard deviation of **1.416**,

. and power equal to 67.6%,

. Use PASS to estimate the required effective sample size, $\widehat{N_{\text{eff}}} = 193$

Power analysis for clustered sampling designs using *meff*:

Option 2 example



Result: $\widehat{N_{\text{eff}}} = 193$

Power analysis for clustered sampling designs using *meff*: Option 2 example

Option 2: Step 2

. Given $\widehat{N_{\text{eff}}}$ = 193, clusters of size *n*=30, and ρ_y = 0.501,
adjust $\widehat{N_{\text{eff}}}$ = 193 to obtain the required needed sample size

. for a CRT, $\rho_x = 1$ and $\widehat{meff} = 1 + (n-1)\rho_y$

. $\widehat{N} = 193 \times [1 + (30 - 1) \times 0.501] \approx 3000$

. Given clusters of size *n*=30, $\widehat{N}$=3000 suggests that
100 clusters need to be sampled and randomized (i.e., 3000 ÷ 30)

This example used the linear mixed models framework.
Now onto the models for clustered data with binary outcomes.

Logistic Regression Models Fit to Clustered Data

misspecification effects

. Consider a logistic model fit to 2-level clustered data
    . e.g., primary care clinics, patients within clinics
    . exchangeable correlation

. Assume the GEE or GLMM (not the survey sampling) modeling framework

. With binary outcomes, there is more than one type of ρ_y estimate
    . a phi-type estimate
    . a tetrachoric-type estimate
    . note that for linear models, there is no corresponding distinction

. Which estimate of ρ_y should be used when estimating *meff* ?
    . Answer: the phi-type coefficient, whether modeling via GEE or GLMM
    . Investigate via Monte Carlo simulation.

Simulated data: Mixed Logistic Model

. m=100 clusters, each with n=50 units: $N = m \times n = 5000$ per replicate sample

. Generate binary y values with exchangeable correlation structure
via a mixed logistic model with random intercepts,

$$y_{ij}^{*} = 0.5 + 0.1x1_{ij} + 0.5x2_{j} + u_{j} + e_{ij}; \quad \text{if } y^{*} > 0 \text{ then } y = 1, \text{ else } y = 0$$

where

. $u_j \sim N(0, \pi^2/3)$; the level-2 residuals; between-cluster variation

. $e_{ij} \sim LOGISTIC(0, \pi^2/3)$; the level-1 residuals; within-cluster variation

. $\rho_y = 0.5$ and $\hat{r}_{\text{tet}.y} = 0.54$

. $x1_{ij} \sim N(0,1)$; a stochastic level-1 x variable with ρ_x=0; $\boldsymbol{meff_{x1} \approx 1\text{-}\rho_y}$

. $x2_j \sim N(0,1)$; a stochastic level-2 x variable: ρ_x=1; $\boldsymbol{meff_{x2} = 1+(n\text{-}1)\rho_y}$

. $r_{x1,x2} = 0$

. 500 replicate samples

Simulation: Logistic Regression Models Fit to Clustered Data

Fit two models to each replicate sample:

GEE logistic and mixed logistic with random intercepts (Laplace)

. Save parameter and standard error estimates, $\hat{\rho}_y$, simulated power

Simulation: Logistic Regression Models Fit to Clustered Data
Results: Intra-cluster correlation of outcome response

	intra-cluster correlation
$\rho_{y(\text{GEE})}$	0.348
phi†	0.365
$\rho_{y(\text{GLMM})}$	0.493
tetrachoric†	0.543

† estimated from first two units of each cluster

As you would expect, GEE working correlations are phi-like,
whereas mixed logistic model intra-cluster correlations are tetrachoric-like

Simulation: Logistic Regression Models Fit to Clustered Data
Results: Parameter and Standard error estimates

	GEE	GLMM
Intercept		
parameter (std dev)	0.330 (.123)	0.509 (.189)
standard error	.124	.186
*x*1		
parameter (std dev)	0.064 (.024)	0.099 (.036)
standard error	.024	.036
*x*2		
parameter (std dev)	0.327 (.128)	0.501 (.190)
standard error	.126	.187

Summary

. GLMM parameter estimates are relatively unbiased (green highlight)

. GEE and GLMM standard error estimates relatively unbiased (yellow highlight)

Simulation: Logistic Regression Models Fit to Clustered Data
Results: GEE Parameter Estimates Relatively Unbiased

	GEE	GLMM	ratio
Intercept			
parameter est.	0.330	0.509	.648
*x*1			
parameter est.	0.064	0.099	.651
*x*2			
parameter est.	0.327	0.501	.652

GEE parameter estimates are relatively unbiased

. $\rho_{y(\text{GEE})} = 0.348$

. Scaling factor: $1 - \rho_{y(\text{GEE})} = .652$ (equal to $meff_{x1(\text{GEE})}$ in this example)

. $b_{\text{GEE}} \approx b_{\text{GLMM}} \times (1 - \rho_{y(\text{GEE})})$

The same scaling factor applies to standard error estimates

Neuhaus and Jewel (1990); Neuhaus, Kalbfleisch, and Hauck (1991); Neuhaus 1992 report #21, Eq. 14

Using PASS to estimate power (compare to simulated power)

. For the GEE and GLMM results, calculate

a. $\Pr(y_{ij}=1 \mid x1 = x2 = 0)$ (intercept)

b. $\Pr(y_{ij}=1 \mid x1 = 1)$

c. $meff_{x1} \approx 1 - \rho_y$ (because ρ_x=0 and *n* is large)

d. $\Pr(y_{ij}=1 \mid x2 = 1)$

e. $meff_{x2} = 1 + (n-1)\rho_y$ (because ρ_x=1)

.I estimated $meff_{x1}$ and $meff_{x2}$ using both $\rho_{y(\text{GEE})}$ and $\rho_{y(\text{GLMM})}$

. To solve for power for logistic regression, PASS requests

. specification of alpha: 0.05, two-tailed

. sample size: $5000 \div meff_{x1}$ or $5000 \div meff_{x2}$, as appropriate

. baseline probability: *a*

. alternative probability: *b* or *d*, as appropriate

. distribution of *x*: unit-standardized normal

PASS: estimate power for int., *x*1, *x*2, using both GEE- and GLMM-based *meff*s

Simulation: Logistic Regression Models Fit to Clustered Data
Results: Power

	GEE $\rho_{y(\text{GEE})} = 0.348$	GLMM $\rho_{y(\text{GLMM})} = 0.493$
Intercept		
power: simulated [PASS]	**.742 [.760]**	**.762 [.942]**
meff = 1+(*n*-1)ρ_y (N_{eff})	0.652 (277)	0.507 (199)
x1		
power: simulated [PASS]	**.788 [.787]**	**.778 [.997]**
meff ≈ 1-ρ_y (N_{eff})	18.032 (7,664)	25.172 (9,868)
x2		
power: simulated [PASS]	**.726 [.734]**	**.756 [.942]**
meff = 1+(*n*-1)ρ_y (N_{eff})	0.652 (277)	0.507 (199)

. *meff*-based estimates of N_{eff} in combination with PASS provided power estimates that were roughly equivalent to simulated values.

. Clearly, when $\rho_{y(\text{GLMM})}$ is used to estimate *meff*s, the result is not correct.

Implications: Power for 2-level logistic models with exchangeable response correlation.

. If you have $\hat{\rho}_{y(\mathrm{GEE})}$ or $\hat{\varphi}$ as an estimate of intra-cluster correlation of binary response, then you can estimate power via *meff*s and standard software (PASS)

. When using *meff*-derived N_{eff} to help estimate power for logistic models, the regression parameters input into (or estimated by) the standard power analysis software will represent population average parameter estimates,
i.e., the type of parameter estimates produced by GEE logistic regression

After completing a *meff*-driven power analysis, you can approximate the minimum detectable unit-specific parameter estimates from their population average counterparts using the scaling factor described by John Neuhaus

Implications: Power for 2-level logistic models with exchangeable response correlation.

. If you only have $\hat{\rho}_{y(\mathrm{GLMM})}$ or $\hat{r}_{tet}$ as an intra-cluster correlation estimate of binary response, then you should not use them to estimate power via *meff*s

Instead…

(i) estimate power by simulation using a GLMM data-generating model

When using a GLMM data-generating model, you subsequently can estimate power via GLMM or GEE logistic regression

It is your call, because given exchangeable response correlation GEE and GLMM models provide equivalent power

or

(ii) use the GLMM-generated data to estimate $\hat{\rho}_{y(\mathrm{GEE})}$ by simulation and then proceed with *meff*-based methods

Limitations

Very limited simulation

. 'large' number of clusters and 'large' clusters considered

. *meff*-based approximations may not work as well with smaller *m* or *n*

. simple two-level model

. balanced cluster size

. limited values of ρ_y and ρ_x considered.

. limited replicate samples

When in doubt, estimate power by simulation

Thank you