SAMPLE SIZE CONSIDERATION IN MULTIPLE REGRESSIONS:

APPLICATION TO LINEAR, LOGISTIC AND COX REGRESSION

Using R.

Diklah Geva

October 2014

Manuscript for publication

Table of Contents

Essay Objectives 3

Introduction 3

1. Sample Size considerations for Linear Regression: 5

2. Sample Size Considerations For Multiple Logistic Regressions: 10

3. Sample Size considerations for Cox Regression 16

Discussion 24

Summary and Concluding Remarks 27

Essay Objectives

The essay objectives are a) to review the methods to obtain sample size in multiple regression and b) to evaluate the impact of the degree of association among the explanatory variables on the sample size, N. R code was developed to realize it for linear, logistics and cox regression.

Introduction

Obtaining sample size is part any experiment design. It should help investigators to establish how large a sample to select from a population based on statistical information and practical considerations. Geva(2004)[1] provided a friendly discussion on the practical and statistical considerations for selecting sample size. Practical considerations includes; budget constraints, patients availability and clinical—important-effect-size, while the statistical considerations includes: power, significance level and effect size. Sample and size and power analysis are closely related and often, in the literature the two terms are mixed. The main elements impacting sample size are:

1. Sample Size --- noted by ***N***
2. Effect Size --- noted by ES and it is usually fraction such as $\text{ES} = \dfrac{\text{Means difference}}{\text{Pooled Standard Error}}$
3. Significance level = P(Type I error) = probability of finding an effect that is not there noted by α
4. Power = 1 - P(Type II error) = probability of finding an effect that is there, noted by 1-β

The following four quantities have close relationship, so, given any three, we can determine the fourth. Often time, in the planning stage of a study, the researcher is seeking for the statistical power given the other three factors (i.e. sample size, significance level, and effect size). Therefore, from historical point of view, this was also referred as “power analysis”.

Two groups comparison in randomized clinical trials is considered the simplest case. In this case, sample size is computed by first specifying the significance level and power, customary set at 5% significance and 80% power, and then specifying the effect size. Effect size reflects the anticipated clinical change and it is usually driven from publications or a pilot study. The assumed effect size can varied by the different sources and this leads researchers to conduct sensitivity analysis to explore sample-size estimate as a function of the assumed effect-size.

Sometimes, when a study failed to show significance, it is argued that sample size was too small or power too low. This argument has been challenged by statisticians recognizing that the retrospective power of a study is 1 if the test was significant, 0 if the tests failed. Since retrospective power calculations are defined based on the observed effect size from the test just performed, thus it becomes a useless exercise, see "The Abuse of Power: The Pervasive Fallacy of Power Calculations" by Hoenig and Heisey (2001)[2].

Nevertheless, in a well design two groups clinical trial using 2-groups t-test for means or proportions the issue of sample size is straight forward, and sample size tables and formulas are available in most basic statistical text books, for example se Chow or ZAR[3,4] .

For power analysis or sample sizing the web offers numerous sample size calculators, many of which are free. Shiboski[5] provides a comprehensive listing of power and sample size programs updated to 2006. Piface, WINPEPI and PASS are just a few of the popular calculators that we have described below:

Piface[6] is a very simple and easy to use such program is, which is a Java applet for power and sample size. It is intended to be useful in planning studies and selection of the plan statistical test by the user. Each selection has a nice graphical interface for studying the power or sample size of that test. In Piface, each dialog window also offers a useful help menu.

WINPEPI[7] (PEPI-for-Windows) is a simple programs, easy to use calculators for basic statistical tools for computing effect-size, p-values, confidence intervals, power and sample size. It is most popular among epidemiologist and it is often used at the study design stage.

PASS[8] *s*oftware is a commercialize program design to provide a research tool for determining the number of subjects that should be used in a study. It is one of the leaders in sample size technology; PASS performs power analysis and calculates sample sizes for over 230 statistical tests and confidence intervals, and as such it is also suitable also for regulatory submissions.

In epidemiological studies, usually number of covariates are considered in a multi-variable statistical model. Sample size and power analysis in this situation becomes more difficult because a) it is necessary to accommodate for the interplay (or correlation) among the variables and also b) there are a multitude of hypotheses being tested regarding the effect-sizes of each variable and the interaction between variables. This complexity led to the establishment of "workable" rules-of-thumb in the selection of sample size for epidemiological studies using regression analysis. For example, select 10 or 30 cases per explanatory variable or per event[9-12] .

Moreover, multicolinearity defined as high correlation among the regression explanatory variables (covariates) and cause numerical deviations in regressions' parameter estimates and therefore O'Brian(2007) [13] noted that due to Variance Inflated Factor(VIF) or Tolerance, many researchers restricts the regression model with 10 explanatory variables. However the issue of VIF do not relates to sample size or power per se, and will not be addressed in this essay.

In linear multiple regression analysis the effect size depends on the multiple correlation among the explanatory variables were higher correlation requires smaller sample size to achieve the same statistical power. The case is different in logistic or Cox's regression because sample size depends on the rate of events in the population, in addition to the multiple correlation among the explanatory variables, resulting in a complex data structure. In this cases, sample size calculations are often accompanied by sensitivity analysis to understand the implication of data structure on the required sample size. Given this complexity in obtaining sample size for multiple regressions, several simplified formulas have been proposed in the literature.

The aims of this paper is to provide a review of sample size determination formulas for multiple regression models including:
i)Linear, ii)Logistic and iii) Cox regressions and to demonstrate how these can be realized to obtain tabulated and graphical sensitivity analysis to a range of assumptions, using R program.
R provides a valuable platform to conduct simulation for sensitivity analysis in the multivariate setting, and thus the models presented in this essay could be customized to variety of generalized linear models.

1. Sample Size considerations for Linear Regression:

In multiple linear regression the set of explanatory variables are regressed on a single continues outcome variable represented by the following formula:

$$y_i = \beta 0 + \beta 1 \cdot X1_i + \beta 2 \cdot X2_i + \cdots + \beta k \cdot Xk_i + \varepsilon_i ,$$

Where y_i is the outcome variable and $x1_i, x2_i \ldots xk_i$ are the k explanatory variables which may be categorical or continues, not necessarily from a known distribution. i represents the index of the i^{th} individual out of a sample size of N cases and ε_i represents the individual deviation from the regression curve. y_i is the outcome variable and it is assumed that the expectation of the outcome variable given all the explanatory variables is normally distributed such that:

$$E(Y|x_1, x_2, \cdots x_k) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Sample size in multiple regression analysis is usually derived by power analysis using formula suggested by Cohen (1988)[14] . For this purpose it is required to specify the anticipated effect size as defined by the regression multiple R^2.

The R package **pwr** developed by Stéphane Champely[15] provides power calculations for t-test and for linear models -multiple regressions with the use of the function:

```
pwr.f2.test(u =, v = , f2 = , sig.level = , power = )
```

Where u and v are the numerator and denominator degrees of freedom and f2 is the effect size measured. The sample size is derived from the total degree of freedoms that is v+u.

There are 3 ways to define the effect size:

1- $f^2 = \frac{R^2}{1 - R^2}$ where R^2 is the population squared multiple correlation in the regression.

2- $f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}$ where R^2_A is the population variance accounted by variable set A, and R^2_{AB} is the population variance account by variable sets A and set B together. Use this to test for variable set A control for variables in set B.

3- $f^2 = \frac{R^2_K - R^2}{1 - R^2_K}$ where R^2 is the population variance accounted by variable of interest-X_1, and R^2_K is the variance of of X_1 regressed over variable set K .

The first formula is appropriate when we are evaluating the impact of a set of predictors on an outcome. The second formula is appropriate when we are evaluating the impact of one set of predictors above and beyond a second set of predictors (or covariates). Finally, the third option is for the interest of 1 variable with additional k cofactors. Cohen[14] suggests that f^2 values of 0.02, 0.15, and 0.35 represent small, medium, and large effect sizes. Notice that the total number of variables in the regression is only specified through the numerator degrees of freedom u and the sample size is the total of u+v. Below is an example of sample size table and graph using the pwr R package under the following specifications; R^2 range of (0.05-0.95) with 5-25 variables; significance and statistical-power are set to customary values of α=0.05 and 1-β=0.80.

We have used the following code chunk 1 to obtain sample size tables and graphs presented below.

Code 1. Sample size for multiple regressions according to Cohen 1988 formula with a range of R^2 and number of variables using R.

```
####################################################################
# Sample Size and Power analysis Multiple Regression
# Cohen 1988 Statistical power analysis for the behavioral sciences
#####################################################################
install.packages("pwr")
library(pwr)

## Exercise 9.1 P. 424 from Cohen (1988)
pwr.f2.test(u=1,v=NULL,f2=0.1/(1-0.1),sig.level=0.05,power=0.8)$v+1

## sample size for 10 variables in a range of R2
R2<-seq(0.05,0.95,0.05)
Nx<-seq(1,19,1)
nvars<-10
 for(i in 1:19){
    N10[i]<-pwr.f2.test(u=nvars ,v=NULL,f2=R2[i]/(1-R2[i]),
            sig.level=0.05,power=0.8)$v+nvars   }
cbind(R2,N210=ceiling(N10))

## sample size for 5-25 variables in a range of R2
## using function Nx
## obtaining table 1.a and figure 1.a

Nx<-function(nvars,R2,alpha=0.05,pwr=0.8){
  ceiling(
  pwr.f2.test(u=nvars-1,v=NULL,f2=R2/(1-R2),sig.level=alpha,power=pwr)$v+nvars
  )}
## applying function Nx
R2<-seq(0.05,0.95,0.05)
N25<-N20<-N15<-N10<-N5<-seq(1,19,1)

for(i in 1:19){
  N25[i]<-Nx(nvars=25,R2=R2[i],alpha=0.05,pwr=0.8)
  N20[i]<-Nx(nvars=20,R2=R2[i],alpha=0.05,pwr=0.8)
  N15[i]<-Nx(nvars=15,R2=R2[i],alpha=0.05,pwr=0.8)
  N10[i]<-Nx(nvars=10,R2=R2[i],alpha=0.05,pwr=0.8)
  N5[i]<-Nx(nvars=5,R2=R2[i],alpha=0.05,pwr=0.8)
}
## generating table and plots
Tab1.a<-as.data.frame(cbind(R2,N25,N20,N15,N10,N5))
```

```
colnames(Tab1.a)<-c("   R2","Nvars=25","Nvars=20","Nvars=15","Nvars=10","Nvars=5")
Tab1.a

      plot(R2,N25,ty="l",col=4,ylab="Sample Size",
           main="Multiple linear Regression sample size by R2 for K Variables",
           xlim=c(0,1), ylim=c(0,450),cex=0.8,cex.main=0.8,cex.sub=0.6,
           cex.axis=0.8,cex.lab=0.8)
      lines(R2,N20,ty="l",col=3)
      lines(R2,N15,ty="l",col=5)
      lines(R2,N10,ty="l",col=8)
      lines(R2,N5,ty="l",col=6)
      legend(x=0.75,y=420,legend=c("25","20","15","10","5"),lty=c(1,1,1,1,1),
             cex=0.6,col=c(4,3,5,8,6), title="K Variables")
      ##################################END OF CODE ########################################
```

Table 1. Sample size for multiple linear regressions formula with a range of R2 and number of variables; according to Cohen 1988 using R package pwr.

R2	K=25	K=20	K=15	K=10	K=5
0.05	453	414	370	317	249
0.1	225	204	182	155	121
0.15	149	135	119	101	78
0.2	111	100	88	74	57
0.25	89	79	69	58	44
0.3	74	66	57	47	35
0.35	63	56	48	40	29
0.4	56	49	42	34	25
0.45	50	43	37	30	21
0.5	45	39	33	26	19
0.55	41	36	30	24	16
0.6	39	33	27	21	15
0.65	36	31	25	19	13
0.7	34	29	23	18	12
0.75	32	27	22	17	11
0.8	31	26	21	15	10
0.85	30	25	20	14	9
0.9	29	24	19	14	8
0.95	28	23	18	13	8

Figure 1. Sample size for multiple linear regressions according to Cohen 1988 formula for a range of R^2 and number of variables using R.

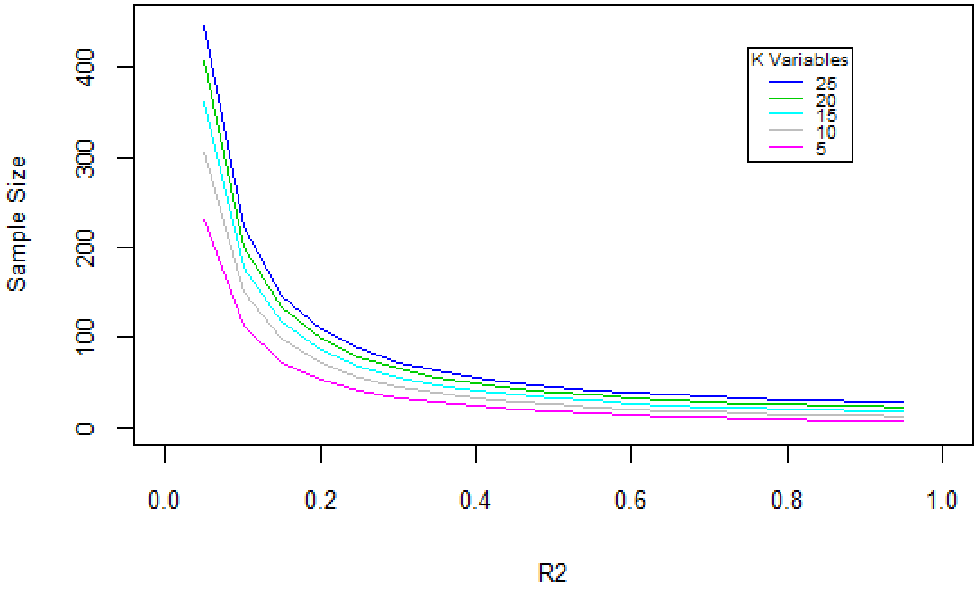


Table 1 shows that sample size has a wide range according to the assumptions. The larger the number of variables in the regression the larger the required sample size, the within correlation plays an important role where lower R^2 requires larger sample size, in other wards; if the variable set is independent, larger sample size is required. For 10 variables with R^2=0.1 the required sample size is N=155, while for R^2=0.8 the required sample size considerably decreases to N=15.

One important comment, that the example calculations presented above assumes that all covariates are continuous and linearly associated with the dependent variable; and thus each variable is contributing 1 degree of freedom in the regression. Should categorical variable included in the regression, larger sample size will be required.

Using only power analytic point in of viewfor selecting appropriate sample size , in epidemiological setting, hosts many difficulties because the focus is at (1) the statistical power of a model rather than the effect size accuracy (as defined by confidence interval) and also (2) no account for the further completion due to interactions[2].

Keley et al[16] argued that sample size for multiple regression can be approached from at least four different perspectives: (a) power for the overall fit of the model, (b) power for a specific predictor, (c) precision of the estimate for the overall fit of the model, and (d) precision of the estimate for a specific predictor. The goal of the first perspective is to estimate the necessary sample size such that the null hypothesis of the population multiple correlation coefficient equaling zero can be correctly rejected with some specified probability (e.g., Cohen[14], 1988, chapter 13 and others[17,18]).

B requires that sample size is computed for the power to test a specific predictor rather than the desired power for the test of the overall goodness of fit for the model[14,19].

A serious problem in regression is the uncertainty regarding the degree of interdependence in the multivariate model which can influence effect size and sample size conciderably[20]

Kellye[16] in his paper encourage researchers to think about effect size measures in multiple regression analysis and presents guidelines for appropriate sample size in multiple regression considering the accuracy in parameter estimation (AIPE) along with power. He argues that the necessary sample size should let the confidence interval around a regression coefficient be reasonably narrow. This is to avoid confidence intervals beeing computed at the conclusion of a study, and only then to realized that the sample size used was too small to yield precise estimates. The AIPE approach to sample size planning allows researchers to plan necessary sample size, a priori, such that the computed confidence interval is likely to be as narrow as specified.

Simulation analysis offers an additional approach for sample size evaluation, however, it is not feasible when prior knowledge regarding the final set of variables included in the model is absent. Some authors have presented simulation studies which may be useful in a sensitivity analysis of targeted hypothesis regarding moderating effects or interaction terms in multiple regression[9,21].

2. Sample Size Considerations For Multiple Logistic Regressions:

In multiple logistic regression the set of explanatory variables are regressed on a single binary outcome variable represented by the following formula:

$$\ln\left(\frac{y_i}{1-y_i}\right) = \beta 0 + \beta 1 \cdot X1_i + \beta 2 \cdot X2_i + \cdots + \beta k \cdot Xk_i ,$$

where y_i is the binary outcome variable and $x1_i$,$x2_i$ …xk_i are the k explanatory variables which may be categorical or continues, not necessarily from a known distribution. i represents the index of the ith individual out of a sample size of N cases. y_i is the outcome variable and it is assumed that the expectation of the outcome variable given the all explanatory variables is normally distributed such that:

$$E(Y|x_1, x_2, \cdots x_k) = \frac{EXP(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k)}{1 + EXP(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k)}$$

The key source to size studies with logistic regression is the paper by Hsieh 1989[22]. The paper presents sample size tables for epidemiologic studies which extend the use of Whittemore's formula[23]. The tables are easy to use for both simple and multiple logistic regressions. Monte Carlo simulations are performed which show three important results. Firstly, the sample size tables are suitable for studies with either high or low event proportions. Secondly, although the tables can be inaccurate for risk factors having double exponential distributions, they are reasonably adequate for normal distributions and exponential distributions. Finally, the power of a study varies both with the number of events.

No R code to obtain sample size for logistic regression according to Whittemore's formula[23] was found to date which follows the formulas in the appendix of Hsieh 1989[22]. Therefore dedicated R function NLR() following was developed to obtain power and sample size in logistic regression.
NLR depends on *Event Frequency* (P), *Odds Ratio* and *Correlation* ρ of the exposure X with the other explanatory variables X2,…,Xk. In this formulation it is not required to speciy the number of covariates, and this information is only accounted by the correlation parameter ρ.

Code 2. Sample size for multiple logistic regressions

```
#######################################################################
# 2. Sample Size and Power analysis Multiple logistic regression
Regression
# HSIEH 1989. SAMPLE SIZE TABLES FOR LOGISTIC REGRESSION
# For multiple logistic regression for one binary outcome
# predicted by continues exposure X and K covariates.
# the correlation of the K covariates with X is given by roh
#######################################################################
setwd("~/100BGU/Reading")

#Set NLR function to obtain N of multiple logistic regression
#with X1 and covariates X2-Xk
#NLR parameters are:
#alpha=0.05, beta=0.80 - significance level and power
#Roh=0 multiple correlation of X1 with X2..XK
#OR=1.5 is the expected OR for 1 SD in X1 (X1 has normal
distribution)
# P is the average frequency of events P=0.5 in the defalt
NLR <- function(alphaw=0.05,Powerw=0.80,Roh=0,OR=1.5,P=0.5){
  theta<-log(OR)     theta2<-theta^2
  theta2m<- -1*theta2
  Zalpha<-qnorm(1-alphaw)
  Zbeta<- -1*qnorm(1-Powerw)

    lam<-(1+(1+theta2)*exp(5*theta2/4)) * (1+exp(theta2m/4))^-1
    n<-round(((Zalpha+exp(theta2m/4)*Zbeta)^2) * (
(1+2*P*lam)/(P*theta2) ))

    NM<-round(n/(1-Roh^2))
  return(NM)
  }

NLR(alphaw=0.05,Powerw=0.80,Roh=0,OR=1.5,P=0.5)

## Obtain sample size table for logistic regression
#with Roh(0-1) OR=or P=p
NLRtab<-function(p=0.5){
Roh<-seq(0,0.9,0.1)
or1.1<-or1.5<-or2.0<-or2.5<-seq(1,10,1)
for(i in 1:10){
  or1.1[i]<-NLR(alphaw=0.05,Powerw=0.80,Roh=Roh[i],OR=1.1,P=p)
  or1.5[i]<-NLR(alphaw=0.05,Powerw=0.80,Roh=Roh[i],OR=1.5,P=p)
  or2.0[i]<-NLR(alphaw=0.05,Powerw=0.80,Roh=Roh[i],OR=2.0,P=p)
  or2.5[i]<-NLR(alphaw=0.05,Powerw=0.80,Roh=Roh[i],OR=2.5,P=p)
}
P<-rep(p,10)
return(cbind(P,Roh,or1.1,or1.5,or2.0,or2.5))
}

tab2<-
rbind(NLRtab(0.01),NLRtab(0.05),NLRtab(0.1),NLRtab(0.3),NLRtab(0.5))
tab2

plotdatP1<-cbind(
  rbind(NLRtab(0.05)[,c(1,2,3)] ,NLRtab(0.05)[,c(1,2,4)] ,
        NLRtab(0.05)[,c(1,2,5)] ,NLRtab(0.05)[,c(1,2,6)] )
  ,rep(c(1.1,1.5,2.0,2.5),each=10))
plotdatP2<-cbind(
                  rbind(NLRtab(0.1)[,c(1,2,3)]
,NLRtab(0.1)[,c(1,2,4)] ,
                        NLRtab(0.1)[,c(1,2,5)]
,NLRtab(0.1)[,c(1,2,6)] )
                  ,rep(c(1.1,1.5,2.0,2.5),each=10))
plotdatP3<-cbind(
  rbind(NLRtab(0.3)[,c(1,2,3)] ,NLRtab(0.3)[,c(1,2,4)] ,
        NLRtab(0.3)[,c(1,2,5)] ,NLRtab(0.3)[,c(1,2,6)] )
  ,rep(c(1.1,1.5,2.0,2.5),each=10))
plotdatP4<-cbind(
  rbind(NLRtab(0.5)[,c(1,2,3)] ,NLRtab(0.5)[,c(1,2,4)] ,
        NLRtab(0.5)[,c(1,2,5)] ,NLRtab(0.5)[,c(1,2,6)] )
  ,rep(c(1.1,1.5,2.0,2.5),each=10))

colnames(plotdatP1)<-colnames(plotdatP2)<-colnames(plotdatP3)<-
colnames(plotdatP4)<-c("P","Roh","N","OR")

########### plot N for logisitc regression ##########

pdat1<-as.data.frame(plotdatP1[11:40,])
pdat2<-as.data.frame(plotdatP2[11:40,])
pdat3<-as.data.frame(plotdatP3[11:40,])
pdat4<-as.data.frame(plotdatP4[11:40,])

par(mfrow=c(2,2),cex=0.6, mar=c(4,4,3,1))
interaction.plot(pdat1$Roh,as.factor(pdat1$OR),
                 pdat1$N,main="Event Rate P=0.05", cex.main=0.9,
                 xlab="",ylab="N",col=c(1,1,1),lty=c(1,2,3),
                 legend=FALSE ,xaxt="n")

interaction.plot(pdat2$Roh,as.factor(pdat2$OR),
                 pdat2$N,main="Event Rate P=0.1", cex.main=0.8,

xlab="",ylab="",cex.ylab=0.8,col=c(1,1,1),lty=c(1,2,3),
                 legend=FALSE,xaxt="n")

interaction.plot(pdat3$Roh,as.factor(pdat3$OR),
                 pdat3$N,main="Event Rate P=0.3",
cex.main=0.8,cex.xlab=0.5,
                 xlab="Covariates
RSQ",ylab="N",cex.ylab=0.6,col=c(1,1,1),lty=c(1,2,3),
                 legend=FALSE)

interaction.plot(pdat4$Roh,as.factor(pdat4$OR),
                 pdat4$N,main="Event Rate P=0.5",
cex.main=0.8,cex.xlab=0.5,
                 xlab="Covariates RSQ",ylab="",cex.ylab=0.8,
                 col=c(1,1,1),lty=c(1,2,3), legend=FALSE)
legend(x=1.5,y=800,legend=c("1.5","2.0","2.5"),lty=c(1,2,3),
       cex=0.8,col=c(1,1,1), title="OR")
##################### END OF CODE ###########################
```

Using the developed R function NLR() sample size for logistic regression was calculated for event rate (P) between 0.01-0.5, OR=1.1-2.5 and for a given multiple correlation (Roh), see table 2.

On the contrary to the linear regression, it appears that as the multiple correlation increase a larger sample size is required. For example; for OR=1.5 and P=0.3 if ρ=0 the required sample size is N=213, but if ρ=0.5 the sample size increases to N=284. And in the case that ρ=0.9, the required sample size N=1121 cases. The case of ρ=0 is essentially the single logistic regression with no cofactors in the model. In addition table 2 shows that smaller effect size indicated by smaller OR, requires larger sample size, also shown on figure 2 and this is parallel to the trend showed for linear regression above.

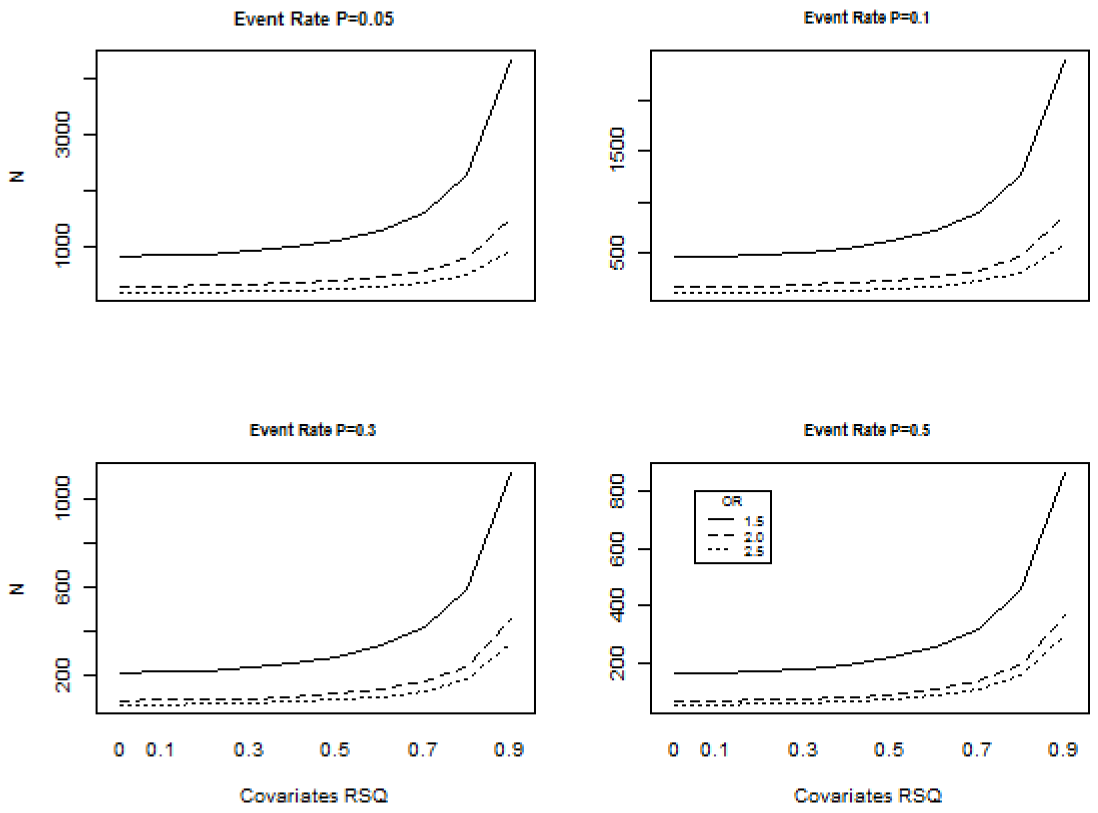
Table 2. Sample size for multiple logistic regressions with a range of event rate P, multiple correlation R^2 and odds ratios OR values for customary α=0.05 and 1-β=0.8.

Roh	OR=1.1	OR=1.5	OR=2.0	R=2.5	Roh	OR=1.1	OR=1.5	OR=2.0	R=2.5
		P=0.01					**P=0.1**		
0	69330	3750	1237	690	0	8170	457	166	109
0.1	70030	3788	1249	697	0.1	8253	462	168	110
0.2	72219	3906	1289	719	0.2	8510	476	173	114
0.3	76187	4121	1359	758	0.3	8978	502	182	120
0.4	82536	4464	1473	821	0.4	9726	544	198	130
0.5	92440	5000	1649	920	0.5	10893	609	221	145
0.6	108328	5859	1933	1078	0.6	12766	714	259	170
0.7	135941	7353	2425	1353	0.7	16020	896	325	214
0.8	192583	10417	3436	1917	0.8	22694	1269	461	303
0.9	364895	19737	6511	3632	0.9	43000	2405	874	574
		P=0.05					**P=0.3**		
0	14966	823	285	174	0	3640	213	86	66
0.1	15117	831	288	176	0.1	3677	215	87	67
0.2	15590	857	297	181	0.2	3792	222	90	69
0.3	16446	904	313	191	0.3	4000	234	95	73
0.4	17817	980	339	207	0.4	4333	254	102	79
0.5	19955	1097	380	232	0.5	4853	284	115	88
0.6	23384	1286	445	272	0.6	5688	333	134	103
0.7	29345	1614	559	341	0.7	7137	418	169	129
0.8	41572	2286	792	483	0.8	10111	592	239	183
0.9	78768	4332	1500	916	0.9	19158	1121	453	347
		P=0.075					**P=0.5**		
0	10435	579	205	131	0	2734	164	70	57

Roh	OR=1.1	OR=1.5	OR=2.0	R=2.5	Roh	OR=1.1	OR=1.5	OR=2.0	R=2.5
0.1	10540	585	207	132	0.1	2762	166	71	58
0.2	10870	603	214	136	0.2	2848	171	73	59
0.3	11467	636	225	144	0.3	3004	180	77	63
0.4	12423	689	244	156	0.4	3255	195	83	68
0.5	13913	772	273	175	0.5	3645	219	93	76
0.6	16305	905	320	205	0.6	4272	256	109	89
0.7	20461	1135	402	257	0.7	5361	322	137	112
0.8	28986	1608	569	364	0.8	7594	456	194	158
0.9	54921	3047	1079	689	0.9	14389	863	368	300

***Using R package pwr and according to Hsieh 1989.**

Figure 2. Sample size for multiple logistic regressions for a range of covariates correlation R^2, event rate P and Odds-ratios OR values.



The above code provides a useful and quick tool to obtain a sample size for multiple regression and to perform a brief sensitivity analysis. However, many have raised the concern that Hsieh 1989[24] formula does not account for the number of variables per event. Others had concern that the formula also fails to account for multitude of factors important for study design. For example Peduzzi et al. 1996[11] argue that it is not only the total sample size that matters but the number of variables per event (EPV) also should be considered. He performs a simulation study of the number of events per variable in a logistic regression analysis based on a data from a cardiac trial of 673 patients in which 252 deaths occurred and seven variables as cogent predictor. They found that for EPV values of 10 or greater, no major problems occurred. But for EPV less than 10, the regression coefficients were biased in both positive and negative directions; the large sample variance estimates from the logistic model both overestimated and underestimated the sample variance of the regression coefficient with paradoxical associations i.e. significance in the wrong direction.

In contrast, Vittinghoff et al 2007[12] have shown that in some circumstances, study design may relax the rule of ten events per variable in a logistic regression. They found a range of conditions, in which coverage and bias were within acceptable levels despite EPV being less than 10, as well as other factors that were influential in addition to EPV.

In a recent publication Courvoisier et al 2011[10] argue that there are much more beyond EPV. One should consider, e.g. bias and precision beyond power alone. This is closely related to the role of data structure requiring a simulation of unique design for a given dataset.

This leads to the approach shared by researchers recommending to conduct a resampling-simulation on a pilot study to derive sample size meeting precision, bias and EPV requirements. We next shall see, how this can be achieved in a cox regression setting.

3. Sample Size considerations for Cox Regression

Cox proportional hazards regression in the epidemiological studies follows the following formula:

$$\boldsymbol{h}(\boldsymbol{t} \,|\, \mathrm{x\_1}, \boldsymbol{x\_2}) = \boldsymbol{h\_0}(\mathrm{t}) \cdot \exp(\beta_1 x_1 + \beta_2 x_2),$$

Where in this example x_1 is a nonbinary explanatory variable of interest and x_2 is a vector of other covariables. E.g. the aim here is to assess a hazard ratio of x_1 =1 to x_1 =0 for a given significance level $\alpha,$ and 1-β power the n is the function so that n=f($\psi, \rho, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}$) , where Ψ is proportion of deaths, estimated by P and exp(β_1) is hazard rate of X_1

and ρ^2 is multiple correlation coefficient from a regression of x_1 on the other covariates x_2 :

$$x_1 = b_0 + b^T 2_2$$

and ρ^2 is estimated by R^2.

PowerSurvEpi[25] is an R package which includes a set of functions to calculate power and sample size for testing a main or an interaction effect in the survival analysis of epidemiological studies (non-randomized studies), taking into account the correlation between the covariate of interest and other covariates. Some functions also take into account the competing risks and stratified analysis.

numDEpi - calculates Number of Deaths Required for Cox Proportional Hazards Regression with Two Covariates for Epidemiological Studies with binary variables of interest. numDEpi follows Schoenfeld (1983)[26] formulations and also allow for competing risk as presented by Latouche(2004)[27].

powerEpiInt and powerEpiInt2- provides further platform for the case of interaction between the two binary covariates according to Schmoor(2000)[28] .

ssize.stratify should be used for stratified sample as formulated by Palta (1985)[29] . This also gives the sample size calculation for survival analysis with binary predictor and exponential survival function. ssizeEpiCont is the most general sample size procedure it allow the calculation of sample size for cox proportional hazard regression with continues predictors it follows Hsieh (2000)[30] formulations

For the epidemiologist working in a set of covariables ssizeEpiCount is most relevant because it is the most flexible to numerous parameters.

The result of selecting several parameters and performing a sensitivity analysis on the other parameter does no longer holds in the cox regression. It is required to use a given pilot study to address the requirements of precision, bias, number of variables in the model, event rate, significance and power. This, typically, depends on the data structure and thus the sample size will be derived based on the given data-table of the pilot study.

A small simulation study, describing the required sample size for customary α and 1-β. In an artificial pilot study of n=100 from bivariate normal distribution with a correlation range between 0.01 - 0.9 for the two explanatory variables X1 and X2. The code, tables and graphs are given below:

Code 3. Sample size for multiple Cox regression

```
####################################################################
#
# Sample Size and Power analysis  for Multiple cox Regression
# According to Hsieh and Lavori (2000)
####################################################################
#
install.packages("powerSurvEpi")
install.packages("MASS")
library(powerSurvEpi)
library(MASS)

# simulate a pilot dataset size ns with correlation of roh between
x1 and x2
# X1 variable of interest, X2 variable to control value,
# where X1,X2 comes from continues MVnorm

NCOX<-function(npilot=25,P=0.1,r=0.6,HRs=1.5,seed=123456) {
set.seed(seed)
Sigma <- matrix(c(1,r,r,1),2,2)
tmp<-mvrnorm(n=npilot, mu=rep(0, 2), Sigma,empirical=TRUE)
X1<-tmp[,1]
X2<-tmp[,2]
failureFlag <- sample(c(0, 1), npilot, prob = c(1-P, P), replace =
TRUE)
dat <- data.frame(X1 = X1, X2 = X2, failureFlag = failureFlag)
retN<-ssizeEpiCont(formula = X1 ~ X2, dat = dat, X1 = X1,
failureFlag = failureFlag,
             power = 0.80, theta = HRs, alpha = 0.05)$n
return(retN)
}

NCOX(npilot=10,P=0.3,r=0.5,HRs=1.5)

## Obtain sample size table for COX regression
#with Roh(0-1) HR=HR P=p
NCOXtab<-function(p=0.075){
  Roh<-seq(0,0.9,0.1)
  HR1.1<-HR1.5<-HR2.0<-HR2.5<-seq(1,10,1)
  for(i in 1:10){
    HR1.1[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=1.1)
    HR1.5[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=1.5)
    HR2.0[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=2.0)
    HR2.5[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=2.5)
  }
  P<-rep(p,10)
  return(cbind(P,Roh,HR1.1,HR1.5,HR2.0,HR2.5))
```

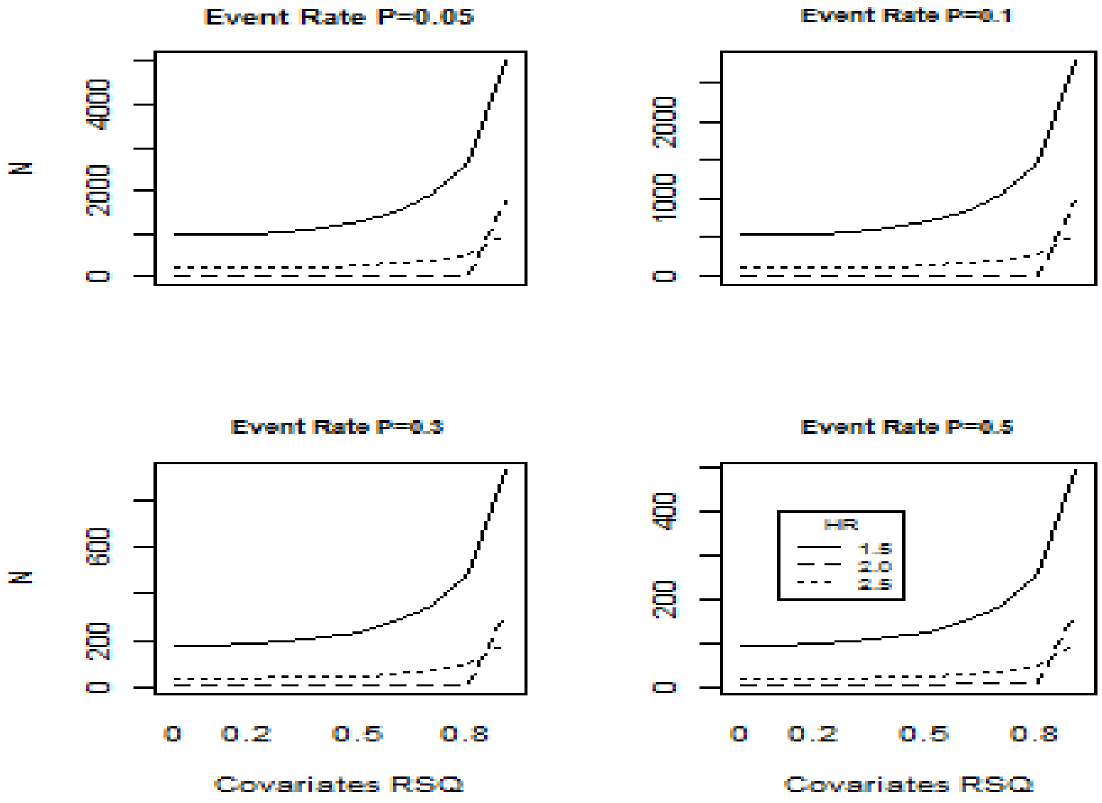
```
}

tab3a<-rbind(NCOXtab(0.01),NCOXtab(0.05),NCOXtab(0.075))
tab3a

tab3b<-rbind(NCOXtab(0.1),NCOXtab(0.3),NCOXtab(0.5))
tab3b


#### data for plot
NCOXdat<-function(p=0.075){
  Roh<-seq(0,0.9,0.1)
  HR1.5<-HR2.0<-HR2.5<-seq(1,10,1)
  for(i in 1:10){
    HR1.5[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=1.5)
    HR2.0[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=2.0)
    HR2.5[i] <- NCOX(npilot=100,P=p,r=Roh[i],HRs=2.5)
  }
    ret <- rbind(cbind(P=rep(p,10),Roh,N=HR1.5,
HR=rep(1.5,10)),
                 cbind(P=rep(p,10),Roh,N=HR2.0,
HR=rep(2.0,10)),
                 cbind(P=rep(p,10),Roh,N=HR2.5,
HR=rep(2.5,10)) )
return(ret)
}
plotdatP1<-NCOXdat(p=0.05)
plotdatP2<-NCOXdat(p=0.1)
plotdatP3<-NCOXdat(p=0.3)
plotdatP4<-NCOXdat(p=0.5)

pdat1<-as.data.frame(plotdatP1)
pdat2<-as.data.frame(plotdatP2)
pdat3<-as.data.frame(plotdatP3)
pdat4<-as.data.frame(plotdatP4)

########### plot N for Cox regression ##########

par(mfrow=c(2,2),cex=0.6, mar=c(4,4,3,1))

interaction.plot(pdat1$Roh,as.factor(pdat1$HR),
                 pdat1$N,main="Event Rate P=0.05",
cex.main=0.9,
                 xlab="",ylab="N",col=c(1,1,1),lty=c(1,2,3),
                 legend=FALSE ,xaxt="n")

interaction.plot(pdat2$Roh,as.factor(pdat2$HR),
                 pdat2$N,main="Event Rate P=0.1",
       cex.main=0.8,
                        xlab="",ylab="",col=c(1,1,1),lty=c(1,2,3),
                        legend=FALSE,xaxt="n")

       interaction.plot(pdat3$Roh,as.factor(pdat3$HR),
                 pdat3$N,main="Event Rate P=0.3", cex.main=0.8,
                 xlab="Covariates
RSQ",ylab="N",col=c(1,1,1),lty=c(1,2,3),
                 legend=FALSE)

interaction.plot(pdat4$Roh,as.factor(pdat4$HR),
                 pdat4$N,main="Event Rate P=0.5", cex.main=0.8,
                 xlab="Covariates RSQ",ylab="",
                 col=c(1,1,1),lty=c(1,2,3), legend=FALSE)
legend(x=2,y=400,legend=c("1.5","2.0","2.5"),lty=c(1,2,3),
       cex=0.7,col=c(1,1,1), title="HR")



#sensitivity to npilot size- table4

n.pilot<-seq(7,500,by=5)
n.estimated<-rep(1,length(n.pilot))

for(i in 1:length(n.pilot)){
  n.estimated[i]<-
NCOX(npilot=n.pilot[i],P=0.3,r=0.5,HRs=1.5,seed=12)}

hm<-mean(n.estimated)
plot(n.estimated~n.pilot,ylim=c(50,500),pch=20,cex=0.8)
abline(h=hm)
segments(n.pilot,rep(hm,length(n.pilot)),x1=n.pilot,y1=n.estimated)
################################# End of CODE ###################
```

Table 3. Sample size for Cox regressions with a range of R^2 and Hazard Ratios, event rate P for the customary α=0.05 and 1-β=0.8

Roh	HR=1.1	HR=1.5	HR=2.0	HR=2.5	Roh	HR=1.1	HR=1.5	HR=2.0	HR=2.5
		P=0.01					P=0.1		
0	43202	2388	817	468	0	9601	531	182	104
0.1	43638	2412	826	473	0.1	9698	536	184	105
0.2	45002	2487	851	487	0.2	10001	553	190	109
0.3	47475	2624	898	514	0.3	10550	583	200	115
0.4	51431	2842	973	557	0.4	11429	632	217	124
0.5	57603	3183	1090	624	0.5	12801	708	243	139
0.6	67503	3730	1277	731	0.6	15001	829	284	163
0.7	84709	4681	1602	917	0.7	18825	1041	356	204
0.8	120005	6631	2269	1299	0.8	26668	1474	505	289
0.9	227377	12564	4300	2461	0.9	50529	2792	956	547
		P=0.05					P=0.3		
0	17281	955	327	187	0	3201	177	61	35
0.1	17456	965	331	189	0.1	3233	179	62	35
0.2	18001	995	341	195	0.2	3334	185	64	37
0.3	18990	1050	360	206	0.3	3517	195	67	39
0.4	20573	1137	389	223	0.4	3810	211	73	42
0.5	23041	1274	436	250	0.5	4267	236	81	47
0.6	27001	1492	511	293	0.6	5001	277	95	55
0.7	33884	1873	641	367	0.7	6275	347	119	68
0.8	48002	2653	908	520	0.8	8890	492	169	97
0.9	90951	5026	1720	985	0.9	16843	931	319	183
		P=0.075					P=0.5		
0	12344	683	234	134	A	1695	94	33	19
0.1	12468	689	236	135	0.1	1712	95	33	19
0.2	12858	711	244	140	0.2	1765	98	34	20
0.3	13565	750	257	147	0.3	1862	103	36	21
0.4	14695	812	278	159	0.4	2017	112	39	22
0.5	16458	910	312	179	0.5	2259	125	43	25
0.6	19287	1066	365	209	0.6	2648	147	51	29
0.7	24203	1338	458	262	0.7	3322	184	63	36
0.8	34287	1895	649	371	0.8	4707	261	89	51
0.9	64965	3590	1229	703	0.9	8917	493	169	97

- Using ssizeEpiCont function from powerSurvEpi Rpackage with a simulated multivariate normal pilot sample of size 100. According to Hsieh and Lavori (2000) implemented with 2- sided alpha.

Figure 3 Sample size for Cox regressions with a range of R^2 and Hazard Ratios, event rate P with customary α=0.05 and 1-β=0.8

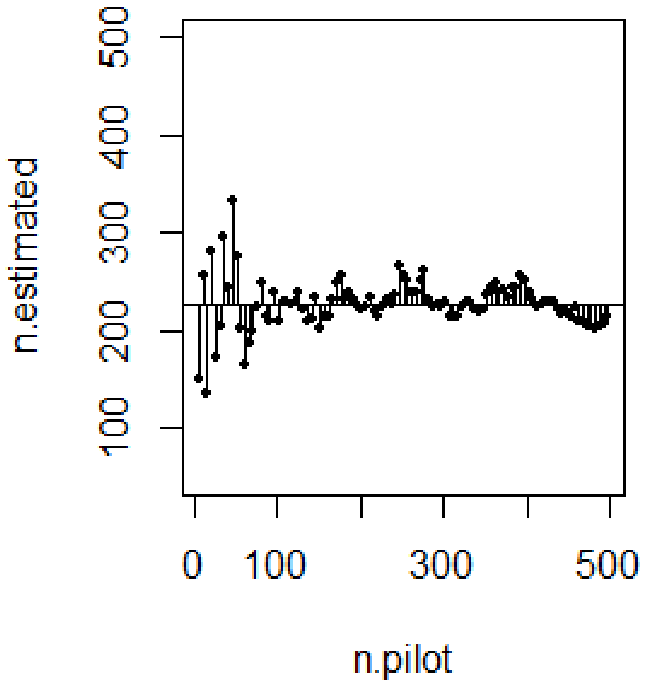


Footnote: from- ssizeEpiCont function from powerSurvEpi R package with a simulated multivariate normal pilot sample of size 100. According to Hsieh and Lavori (2000) implemented with 2- sided alpha.

Similarly to the logistic regression results, in the cox regression - it appears that as the variable correlation is increasing, larger sample size is required. For example; a HR=1.5 and P=0.3 if ρ=0 the required sample size is N=177 but if ρ=0.5 the N is increase to N=236. In the case that ρ=0.9 the requires sample size is of N=931. The case of ρ=0 is actually the single hazard regression with no cofactors in the model. In addition table 2 shows that smaller effect size indicated by smaller HR, requires larger sample size, also shown on figure 2. It appears that in these cases the required sample size for HR is comparable to the size generated by the formula for logistic regression, essentially replicating the trends with respect to the correlation ρ^2 and event rates P.

In a sepate simulation we have noted that the sample size estimates is converging as the number of pilot size (n-pilot) is increasing. For illustration see figure 4 wee see that a pilot of size n will converge for the rtifitial by-variate normal covariates in a cox regression.
On one hand this results are reassuring, because sample size estimates are shown to be consistent with larger pilot size. However, on the other hand the convergence occurs with a considerably large pilot size of about 100 cases. Such large pilot can be feasible with artificial simulated pilot data, but can be of a unreasonable burden in field studies.

Figure 4. Estimated sample size as function of pilot sample size for Cox regression with two explanatory variables from binormal distribution, α=0.05 and 1-β=0.8 HR=1.5 P=0.3 and Roh=0.5.



Discussion

This article reviewed methods for calculating sample size in multiple regressions frame works including: linear, logistics and Cox proportional hazard regression.

The results presented by many authors and replicated here show that in the multi-dimensional-studies as in epidemiology, the design is impacted by many factors that can influence sample size determination. Therefore, no single factor is deemed superior and sensitivity analyses should be conducted in this process.

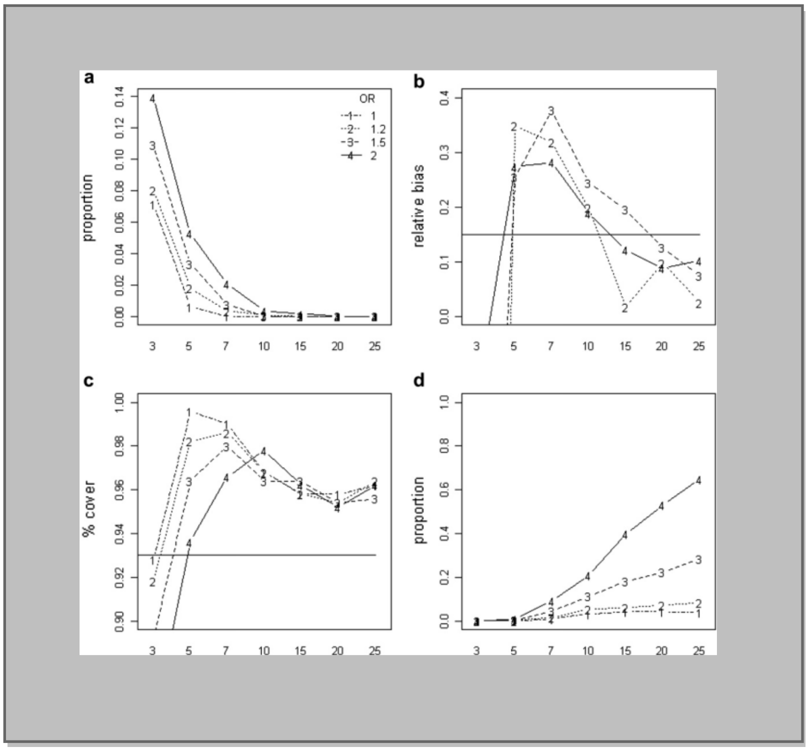
Impact of R^2 on sample size N was examined in this essay in Linear, logistics and Cox regressions. The results were conflicting; for linear regression- increasing R^2 leads to smaller N, while for logistics and Cox regression opposite results were found, increasing R^2 leads to increase in N. Two possible explanations to this unexpected results: I) Cohen(1988)[14] formula was used to derive sample size which do not accounts for the possible effect size reduction in the present of other covariates and II) The mean and variance of the outcome are dependent for event rate such as in logistics/cox regression but are assumed independent for normal outcome variable such as in the linera regression. This dependency forces the sample size estimates (N) to follow the correlation in logistics/cox regression but are not bounded in the linear regression.

Rules-of-thumb have been suggested for determining the minimal number of subjects required to conduct multiple regression analyses. Despite the development of procedures for calculating sample size as a function of relevant effect size parameters, rules of thumb tend to persist in designs of multiple regressions studies common in epidemiology. One explanation for their common usage may be the difficulty in formulating a reasonable a priori value of an effect size to be detected. The example we have conducted above provides an explanation of why rules of thumb for choosing sample size have been used and also shows that the outcome of it may results in too large (or too small) sample size, also showed by others for example see Maxwell[19].

These rules-of-thumb are evaluated by comparing their results against those based on power analyses for tests of hypotheses of multiple and partial correlations. Green[18] have argued that the exact power analysis simulations did not support the use of rules-of-thumb that simply specify some constant (e.g., 100 subjects) as the minimum number of subjects or a minimum ratio of number of subjects (N) to number of predictors (k).

Not only power ought to be factor in sample size considerations. Other authors[16] have raised concern regarding the effect size particularly in interaction terms and in term of accuracy of parameters confidence interval coverage. Failing to account for the accuracy of effect size may result in biased estimates and incorrect inference. This bias is of particular concern when many explanatory variables appears in the regression. In contrast, within a cox regression simulations performed by Hsieh 2000[30] , the censored observations did not contribute to the power of the test of the proportional hazards. This paper also provides a variance inflation factor together with simulations for adjustment of sample size when additional covariates are included in the model. Courvoisier et al 2011[10] provide simulation results for four important issues to account in the study design see the figure below.

Figure 5. Example of simulations for (a) Percentage of nonconverged replications, (b) Median relative bias of the estimate, (c) Percentage of cover and (d) Percentage of significant coefficients plotted against the number of events per variable (EPV) in a logistics regression.



Taken from Courvoisier et al[10] (2011) to illustrate sensitivity analysis to several parameters

This sensitivity analysis shows large variation with respect to the assumptions and thus it supports the necessity of using methods to determine sample size that incorporate multitude of considerations.

Variable selection problem is one of the most general model selections. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. In large epidemiological data sets with many variables the issue is not only pf how many variables to include in a model but also which variable to include in the model. Nevertheless this is not the focus of this essay and for a review of the key developments which have led to the wide variety of approaches to select variables to the model sees George 2000[31] and also an introductory updated tutorial is given by Cassotti & Grisoni[32]. This issue is relevant for "fishing expedition" more common in the machine learning settings.

Generalized Linear Models(GzLM) Sample Sizing – The issue of sample size in multi-variable setting may be extended to the GzLM framework and thus generalized to include also the case of count or categorical outcome, beyond the normal, binary or time to event presented here. This can also be further generalized for the cluster data setting. In such setting, the close-form formula can fail in determining a sample size due to ignorance of within covariance structure. Method similar to that presented in this essay for Cox regression should be applied. By using data from a pilot study to establish the required data structure and asses by simulation number of sample-size considerations, including: power, coverage, mediation, goodness of fit, number of events per variable etc. This is currently an active area of research and people have approached it from different angles. For example, Moineddin et al[33] provided a simulation to assess the effect of a varying sample size, at both the individual and group levels, on the accuracy of the estimates of the parameters and variance components of multilevel logistic regression models. In addition, the influence of prevalence of the outcome and the intra-class correlation coefficient (ICC) is examined. Their result indicated that the estimates of the fixed effect parameters are unbiased for 100 cluster with size of 50 or higher. The estimates of the covariance components may be biased even with this large size. However, the random effect is baised only when cluster size below 5.

Qianyu [34] gives sample size and power calculations for GLIMMIX which are affected by prior information about random effects, within-subject correlations. The SAS program resample from a pilot data in order to compute addequate sample sizes for correlated binary outcomes.

Bayesian, Resampling, and MCMC are becoming the method of choice for obtaining a sample size in an epidemiological study[35,36] . R programs has strong capabilities in all three aspects and therefore can provide a useful platform for developing the desired study plan and sample size.

Summary and Concluding Remarks

We have provided a review of different approaches for sample size calculations and power analysis in the multiple-variables setting shared by many epidemiological studies. During the 80th and 90th of the previous century methods were developed using a single formula. Those were in turn simulated over a range of conditions and bunch of rule-of-thumbs were developed. It appears that none of these rules really covers all important issues required for sample size estimation: i.e. power, bias, coverage, moderation and goodness of fit. This multitude of issues highly depends on data structure and only can be fully assesed in simulation studies. During the last decade more and more studies have been designed based on Bayesian approach, i.e. MCMC and resampling methods to better address the factors impacting the sample size. The availability of software like R with a set of flexible tools enables more researchers to plan multi-variable and multi-level complex epidemiological studies.

Bibliography

1. Geva D. Dose the sample size calculation begins and ends by a placement in a formula?<br />(האם חישוב גודל מדגם מתחיל ונגמר בהצבה לנוסחה?). e-med.co.il publication. http://www.e-med.co.il/emed/new/usersite/content.asp?CatID=31&ContentID=45371. Updated 2004.

2. Hoenig JM, Heisey DM. The abuse of power. *The American Statistician*. 2001;55(1).

3. Chow S, Wang H, Shao J. *Sample size calculations in clinical research.* CRC press; 2007.

4. Zar JH. *Biostatistical analysis.* Pearson Education India; 1999.

5. Shiboski SC. List of power and sample size programs. divition of biostatistics UCSF. http://www.epibiostat.ucsf.edu/biostat/sampsize.html. Updated 2006.

6. Lenth RV. Java applets for power and sample size [computer software]. http://www.stat.uiowa.edu/~rlenth/Power. Updated 2006-9.

7. Abramson JH. WINPEPI (PEPI-for-windows): Computer programs for epidemiologists. *Epidemiol Perspect Innov*. 2004;1(1):6. doi: 1742-5573-1-6 [pii].

8. Hintze J.   Sample size & power analysis software - PASS 12. NCSS LLC. kaysville, utah, USA<br /><br />. www.ncss.com. Updated 2013.

9. Aguinis H, Harden EE. Sample size rules of thumb. *Statistical and methodological myths and urban legends*. 2009:267-286.

10. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: Beyond the number of events per variable, the role of data structure. *J Clin Epidemiol*. 2011;64(9):993-1000.

11. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379.

12. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox

regression. *Am J Epidemiol*. 2007;165(6):710-718. doi: kwk052 [pii].

13. O'brien RM. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*. 2007;41(5):673-690.

14. Cohen J. *Statistical power analysis for the behavioral sciences.* Routledge Academic; 1988.

15. Champely S. CRAN pwr-package: Basic power calculations pwr. http://cran.r-project.org/web/packages/pwr/pwr.pdf. Updated 2009.

16. Kelley K, Maxwell SE. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychol Methods*. 2003;8(3):305.

17. Gatsonis C, Sampson AR. Multiple correlation: Exact power and sample size calculations. *Psychol Bull*. 1989;106(3):516.

18. Green SB. How many subjects does it take to do a regression analysis. *Multivariate behavioral research*. 1991;26(3):499-510.

19. Maxwell SE. Sample size and multiple regression analysis. *Psychol Methods*. 2000;5(4):434.

20. Green BF. Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research*. 1977;12(3):263-287.

21. Shieh G. Sample size determination for confidence intervals of interaction effects in moderated multiple regression with continuous predictor and moderator variables. *Behavior research methods*. 2010;42(3):824-835.

22. Hsieh F,Y. Sample size tables for logistic regression. *Stat Med*. 1989;8(7):795-802.

23. Whittemore AS. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*. 1981;76(373):27-32.

24. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med*. 1998;17(14):1623-1634.

25. Qiu W, Chavarro J, Lazarus R, Rosner B, Ma J. **powerSurvEpi: Power and sample size calculation for survival analysis of epidemiological studies**. *CRAN-PackagepowerSurvEpi*. 2012;V6 2014.

26. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983:499-503.

27. Latouche A, Porcher R, Chevret S. Sample size formula for proportional hazards modelling of competing risks. *Stat Med*. 2004;23(21):3263-3274.

28. Schmoor C, Sauerbrei W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Stat Med*. 2000;19(4):441-452.

29. Palta M, Amini SB. Consideration of covariates and stratification in sample size determination for survival time studies. *J Chronic Dis*. 1985;38(9):801-809.

30. Hsieh F, Lavori PW. Sample-size calculations for the cox proportional hazards regression model with nonbinary covariates. *Control Clin Trials*. 2000;21(6):552-560.

31. George EI. The variable selection problem. *Journal of the American Statistical Association*. 2000;95(452):1304-1308.

32. Cassotti M, Grisoni F. Variable selection methods: An introduction . http://www.moleculardescriptors.eu/tutorials/T6_moleculardescriptors_variable_selection.pdf.

33. Moineddin R, Matheson FI, Glazier RH. A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol*. 2007;7:34. doi: 1471-2288-7-34 [pii].

34. Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Comput Methods Programs Biomed*. 2008;91(2):122-127.

35. Wang F, Gelfand AE. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*. 2002:193-208.

36. Bolker BM, Brooks ME, Clark CJ, et al. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in ecology & evolution*. 2009;24(3):127-135.