SAMPLE SIZE ESTIMATION IN ANALYTICAL CROSS-SECTIONAL STUDIES FOR TESTING OF LINEAR REGRESSION COEFFICIENT

Jayadevan Sreedharan[1] & Subramanian Chandrasekaran[2] & Aji Gopakumar[3]
[1]Professor of Epidemiology& Biostatistics, Department of Community Medicine, Gulf Medical University, UAE
[2]Assistant Professor, Department of Statistics, Annamalai University, Tamilnadu
[3]PhD Research Scholar, Annamalai University, Annamalai Nagar, Tamilnadu, India

Received: March 02, 2019 Accepted: April 04, 2019

ABSTRACT: : *One of the epidemiological research designs aimed to determine the prevalence, called descriptive cross-sectional study. Cross sectional study also uses to estimate the risk and test of hypothesis, called analytical cross-sectional study. In order to identify the associated factors of outcome variable,an adequate case/exposure group should be reflected on the selected sample. This paper discusses a method for calculating adequate sample size inanalytical cross-sectional study,totest the 'significance of regression coefficient in a simple linear regression model'. The formula proposed in this paperincludes power (1-β), Correlation Coefficient (r), degree of change to be detected (β, unstandardized regression coefficient), margin of error (L),standard deviation (S) and significance level (α). Further, simulation is carried out to evaluate the performance of the sample size formulathat derived.*

Key Words: *Adequate sample size, simple linear regression, analytical crosses sectional study, standardized regression coefficient*

INTRODUCTION

Calculation of sample size is an important step in any type of scientific studies. A valid research conclusion is based on various factors such assuitable choice of study design, determination of adequate sample size, selectionof representative sample, accuracy of the collecteddata,appropriate application ofstatistical methods and correct interpretation of results (Thiese et al., 2015). Objective of this paper is to derive a formula for adequate sample size in one of the observational studies called cross sectional studies, are generally classified as descriptive and analytical studies (CEBM, 2014). In a cross sectional study, subjects are selected at a certain point of time and simultaneously measures exposure and health outcome; rather than selection based on outcome/exposure status as in case-control or a cohort study.Charan andBiswas (2013) reported that sample size formula (n) for cross sectional study includes the indicators such as variance of the variable understudy (S^2), level of significance (α) and the margin of error (L).Generally descriptive cross-sectional studies donot testing any hypothesis about the parameter and thereforeformula of sample size doesn't need to include power term (Z_β). But the study helps to describe the observed facts about the population and suitable for generating the hypothesis. The purpose of conducting descriptive cross-sectional studies is to learn about the characteristics of the population than comparison of groupsforproving any hypothesis (Mann, 2012).

Analytical cross-sectional studies are used to estimate the prevalence as well as the degree of association betweenpotential factorsand the outcome variable. In this type of studies, a single sample is selected at a point of time, which later divides into groups as per exposure and outcome status. Further, a suitable hypothesis is formulated related to the comparison group and conclusion is derived with the help of descriptive statistics and inferential techniques. Hence analytical cross sectional studies not just helping to generate the hypothesis, but it also uses to strengthen the hypothesis about exposure and outcome relationships. In this context, the existing formula is not appropriate for analytical cross-sectional studies, though it is suitable for descriptive studies. This paper aims to introduce a method for calculating sample size for analytical cross sectional study which focuses on testing of hypothesis related to degree of association/relation.

EXISTING METHOD OF SAMPLE SIZE CALCULATION IN DESCRIPTIVE/ANALYTICAL CROSS SECTIONAL STUDY

Since sample size (n)is the ratio of standard deviation (SD) to thestandard error (SE),

Sample size $n = \frac{SD^2}{SE^2}$ ---------- (A)

In reference to Confidence Interval (CI)of the parameter of interest, CI can be expressed as
CI = (estimate) ± (Critical value) (SE of the estimate)
For a normally distributed variable, CI for the estimate mean (μ) is 'mean ± $Z_{\alpha/2}$SE(mean)' where $Z_{\alpha/2}$ is the Z score at a fixed level of α

Since'Margin of Error (L)' denotes the degree of shift from the point estimate,L can be expressed as '$Z_{\alpha/2}$SE(mean)' (Scottish Govt., 2017 &Hazra,2017).

Therefore $SE = \frac{L}{Z_{\alpha/2}}$
Then equation (A) becomes,

$$n = \frac{SD^2}{SE^2} = \frac{SD^2}{\frac{L^2}{Z_{\alpha/2}^2}} = \frac{{Z_{\alpha/2}}^2 SD^2}{L^2} = \frac{Z_{\frac{\alpha}{2}}^2}{L^2}(variance)$$

is the existing sample size formula in cross-sectional studies for estimation of risk and test of its significance (Kasiulevičius et al., 2016;Pourhoseingholi et al., 2013&Wang & Chow, 2007). Herevariance is considered as 'S^2'or 'pq' according to the parameter of interest, where 'S' is the sample standard deviation and 'p' is the expected population proportion that can be identified from previous studies or by conducting pilot studies.The formula also includes L, which is the Margin of Error (absolute measure of error or relative precision) and Z is the standard normal variate at the chosen level of significance (z=1.96 at 5% level of significance or z=2.58 at 1% significance level).

PROPOSED SAMPLE SIZE FOR ANALYTICAL CROSS SECTIONAL STUDY

Analytical cross-sectional studies use inferential testing procedure to test the hypothesis related to the riskassociated and thereby finding the major risk factors. Therefore, power of the test needs to be incorporated in the sample size formula along with other factors that affecting sample size. The risk factors identified from the analytical cross sectional studies help to strengthen the hypothesis. Hypothesis about the relationship or causation which is strengthened based on a descriptive/analytical cross sectional study,can be later proved in extended analytical studies such as case-control or cohort studies;which are the best methods of causation (Grimes& Schulz, 2002).

In the current research paper, calculation of sample size in cross sectional studies is modified by taking both 'inferential techniques and study design' into consideration. As a preliminary study to strengthen the research hypothesis, an analytical cross sectional study is the suitabledesignfor estimation and prediction of the risk. Once the study design is finalized, next foremost step is the calculation of sample size.Sufficient size of the sample needs to be calculated considering the test of significance of 'association/relationship' between the predictor and the response variable.

According to the nature of dependent variable, anyregression modelcan be chosen as the appropriate method to estimate and test the risk. The current study interested to focus on the relationship between a response variable which is continuous type and one predictor variable of any type (quantitative or qualitative). Therefore sample size to be calculated to estimate the change in the response variable for unit change in the explanatory variable. This measure is given by linear regression coefficient (slope, β) from a simple linear regression model. In simple linear regression model Y= α+βX, 'β' is the unstandardized regression coefficient and 'α' is the intercept.Generally, Standardized regression coefficient and correlation coefficient will be same in a simple linear regression. Standardized coefficients ignore scale of units of the independent variablewhich makescomparison easy, but not give précised predictor effect. This may mislead the findingsas a result of sampling error, especially when sample size is small and at skewed variables. Though standardized coefficient is apt in multiple regression that includes variables at different scales, unstandardized regression coefficient in simple regression gives the real effect of the predictor.This is because of unstandardized regression coefficient considersoriginal units of measurement of the variable that involved in the model. Standardized coefficients are estimated in normalized units and it is valid if the variable not skewed and normally distributed. This coefficient explains change of one standard deviation in outcome variable as a result of one standard deviation change in predictor variable. Since there is no need to compare the degree of impact of independent variable in simple linear regression (as it is with only one independent variable), unstandardized regression coefficient place an important role in a simple regression model, especially in interpretation of the results. But unstandardized regression coefficient is not a suitable

statistic to identify the degree of impact in multiple regression. Therefore in simple linear regression,an unstandardized regression coefficient is the parameter to get estimated and testedfor finding the significance of change in the response variable for a unit change in explanatory variable (Kim &Mueller 1976;Greenlandet al.,1986; Bring, 1994;Gelman 2008 and Kim 2011).

Hence the objective here is to estimate the unstandardized regression coefficient (or β) and test the 'significance ofβ'where hypothesis isabout the slope of the regression equation. In order to achieve this objective, hypothesis can be formulated and tested as H_0: β = 0 against H_1: β ≠ 0. In order to test 'the slope is a value specified in the null hypothesis' (predictor variable has no effect on the response variable), sample size should be calculated byconsideringadequate power (1-β).If both variables are continuous in nature and assumption of normality satisfied,prediction of relationship between explanatory and response variable can be performed based on the concept of linear relation, where 'r' is the correlation coefficient that indicates the strength of the relationship (Hsieh et al., 1998 &Thigpen, 1987).

Sample size has greater role in identifying the significance of the slope whether the effect is real or by chance.Since determination of sample size is an important step for accurate reflection of the population in the selected sample,an adequate sample size formula is suggested as follows.

Derivation of sample size is started from type I error, by finding an expression for critical value (cv).

The probabilty of rejecting null hypothesis (H_0) when it is true, called Type I error (α) thatcan be expressed as,

α = $P(X \geq cv/H_0) = 1 - P(X \leq cv/H_0)$

Null hypothesis gets rejected when the value of the test statistic is greater than the critical value (cv). For a normally distributed variable 'X' and the hypothesized meanμ_0,

α $= 1 - P(\frac{X-\mu_0}{\sigma} \leq \frac{cv-\mu_0}{\sigma} / H_0)$

By simplification, critical value cv = ($\mu_0 - Z_\alpha\sigma$) ……eq (1)

Similarly power(1-β)can be expressed as follows,

1-β = $P(X \geq cv/H_1)$

$= (1 - P(\frac{X-\mu}{\sigma} \leq \frac{\mu_0-\mu}{\sigma} - Z_\alpha/H_1)$ (by standerdization&substitution of eq 1)

β = $F(\frac{\mu_0-\mu}{\sigma} - Z_\alpha)$(on simplification)

$Z_\beta = \frac{\mu_0-\mu}{\sigma} - Z_\alpha$

$\frac{Z_\alpha+Z_\beta}{\mu_0-\mu} = \frac{1}{\sigma}$

Sincevariables follow Normal distribution N (μ,σ), σ can be replaced by $\sigma/\sqrt{n}$

$\frac{Z_\alpha+Z_\beta}{\mu_0-\mu} = \frac{1}{\sigma/\sqrt{n}}$……eq (2)

Indescriptive/analytical cross sectional study, only one sample is selected and later segmented to deseased/non-diseased. Then desease rate is observed across the exposed/unexposed to find the effect of risk factor. Therefore, shift of the sample mean from the hypothesized mean can be treated as margin of error (L) and here it is considered as the absoulte error.

Implies, equation (2) in cross sectional study can be expressed as,

$\left(\frac{Z\alpha+Z_\beta}{L}\right) = \frac{1}{\sigma/\sqrt{n}}$……eq (3)

In order to find the required sample size for testing the significance of slope in a simple linear regression model, consider the confidence Intervel (CI) of the regression coefficient as follows.

$\widehat{\beta} \pm Z_{\frac{\alpha}{2}}SE(\widehat{\beta})$ where $L = Z_{\frac{\alpha}{2}}SE(\widehat{\beta})$ is the margin of error.

L $= Z_{\frac{\alpha}{2}}SE(\widehat{\beta})$

$= Z_{\frac{\alpha}{2}}\frac{S_{Res}}{\sqrt{\sum(x_i-\overline{x})^2}} = Z_{\frac{\alpha}{2}}\frac{S_{Res}}{S_x\sqrt{n-1}} = Z_{\frac{\alpha}{2}}\frac{S_y\sqrt{\frac{(1-r^2)(n-1)}{(n-2)}}}{S_x\sqrt{n-1}} = Z_{\frac{\alpha}{2}}\frac{S_y}{S_x}\frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}}$ ……eq (4)

Since$SE(\widehat{\beta}) = \frac{S_{Res}}{\sqrt{\sum(x_i-\overline{x})^2}}$ and unstanderdized regression coefficient (β)= $r\frac{S_y}{S_x}$,(as in BMJ 2019)

substituting the expression of βin 'L' of equaltion (4)

L $= Z_{\frac{\alpha}{2}}\frac{\beta}{r}\frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}}$……eq (5)

substituting the expression of L derived in eq (5) to eq (3)

$$\left(\frac{Z_{\alpha/2} + Z_{\beta}}{Z_{\alpha/2}\frac{\beta}{r}\frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}}}\right) = \frac{1}{\sigma/\sqrt{n}}$$

On simplification,

$$\sqrt{n} = \frac{(Z_{\alpha/2} + Z_{\beta})\sigma\, r}{Z_{\frac{\alpha}{2}}\beta\, \frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}}}$$

Since standard error of sample correlation coefficient is$SE = S_r = \frac{\sqrt{(1-r^2)}}{\sqrt{(n-2)}}$ and margin of error for the correlation coefficient is $L = Z_{\alpha/2}S_r$

$$\sqrt{n} = \frac{(Z_{\alpha/2} + Z_{\beta})\sigma\, r}{Z_{\frac{\alpha}{2}}\beta\, S_r} = \frac{(Z_{\alpha/2} + Z_{\beta})\sigma\, r}{L\, \beta}$$

Squaring both sides,

$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2\sigma^2 r^2}{L^2\beta^2}$ ……..eq (6)

is the number of subjects requiredto test the linear regression coefficeint in an analytical cross sectional study. Adequate sample size can be chosenfor identifying a required degree of change(unstanderdized regression coefficient, β) that to be detected from the predictor. As per the precision required, sample size can be calculated by choosing appropriate level of significance, margin of error, and power. An estimate of the coefficient of determination (r^2) or correlation coefficient (r)can be observedfrom previous studies or pilot studies.Implies, size of the sample can be determined for the required percentage of variation in the response variable (r^2) thatto be detected from the explanatory variable.

SIMULATION RESULTS

Minimum sample size is calculated by the existing method (Table 1),for fixed valaueof significance level (α=0.05) and different values of variance (S^2). Margin of error is considered as 10% of the variance. Calculated sample size is preseneted in the tablebelow for standered deviation 1 to 5.

Table 1: Minimum Sample Sizeby the Existing Method

Sample size for α=0.05& L =10% of variance	Sample SD (S)								
	1	1.5	2	2.5	3	3.5	4	4.5	5
Sample size (n)	784	348	196	125	87	64	49	39	31

Toverify the accuarcy of propsed sample size formula, simulation is carried out for various values of unstanderdized regression coefficient (β) and variance(S^2), both can be identified from previous reliable studies. Adequate sample size is calculated (Table 2-4) forfixed significance level (α=0.05), power (1-β=0.80), margin of error (10% of variance)and coefficient of determination (r^2=0.25, 0.64&1or r=0.5, 0.8 & 1). In a simple linear regression, coefficeint of determination (r^2) is the square of correlation coefficient. Repeated trials were performed for different power values, significance levels, margin of error and r^2 values.

Simulation results by the proposed method of Sample size

Table 2: Adequate Sample size for fixed values of α=0.05, 1-β=0.80, r=0.5,L=10% of variance

β	Sample SD (S)								
	1	1.5	2	2.5	3	3.5	4	4.5	5
0.1	19600	8711	4900	3136	2178	1600	1225	968	784
0.2	4900	2178	1225	784	544	400	306	242	196
0.3	2178	968	544	348	242	178	136	108	87
0.4	1225	544	306	196	136	100	77	60	49
0.5	**784**	**348**	**196**	**125**	**87**	**64**	**49**	**39**	**31**

Table 3: Adequate Sample size for fixed values of α=0.05, 1-β=0.80, r=0.8,L=10% of variance

β	Sample SD (S)								
	1	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
0.1	50176	22300	12544	8028	5575	4096	3136	2478	2007
0.2	12544	5575	3136	2007	1394	1024	784	619	502
0.3	5575	2478	1394	892	619	455	348	275	223
0.4	3136	1394	784	502	348	256	196	155	125
0.5	2007	892	502	321	223	164	125	99	80
0.6	1394	619	348	223	155	114	87	69	56
0.7	1024	455	256	164	114	84	64	51	41
0.8	**784**	**348**	**196**	**125**	**87**	**64**	**49**	**39**	**31**

Table 4: Adequate Sample size for fixed values of α=0.05, 1-β=0.80, r=1.0,L=10% of variance

β	Sample SD (S)								
	1	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
0.1	78400	34844	19600	12544	8711	6400	4900	3872	3136
0.2	19600	8711	4900	3136	2178	1600	1225	968	784
0.3	8711	3872	2178	1394	968	711	544	430	348
0.4	4900	2178	1225	784	544	400	306	242	196
0.5	3136	1394	784	502	348	256	196	155	125
0.6	2178	968	544	348	242	178	136	108	87
0.7	1600	711	400	256	178	131	100	79	64
0.8	1225	544	306	196	136	100	77	60	49
0.9	968	430	242	155	108	79	60	48	39
1	**784**	**348**	**196**	**125**	**87**	**64**	**49**	**39**	**31**

In order to conduct an analytical cross-sectional study with an objective oftesting the significance of regression coefficient, a sample size formula is suggested in this paper and subsequent simulation results indicated few common trends. Requirement of larger samples is observedforattaining the higher power (1-β) and when the Standard deviation of the mean (S) gets decrease. As showed in table 2, sample size gets decreasefor increasing margin of error (L). Theoretically, a smaller sample size is enough to identify a larger shift of the estimate from the specified average. The same finding is observed from the simulation results.

Otherimportant findings from the simulation are as follows,

- Size of the sample increases for increasing value ofr^2(coefficient of determination). It shows the need of collecting larger samples to detect a higher percentage of variation in the outcome variable,which is caused by the predictor variable.
- Small number of samples is enoughto identify a higher rate of change (unstandardized regression coefficient, β) in the dependent variable for the functional relationship with the independent variable. Implies, larger samples are requiredto detect smaller and subtle changes in the response variable,which is resulted from the linear relationship with the explanatory variable.
- Existing method of sample size provides minimum number of subjects that required conductinga cross sectional study, but it is important to choose sufficient sample size that can be done according to β level. In order to test the significance of β in a cross-sectionalstudy, sample size to be collected from a minimum 'n' or above,in view of the degree of effect to be detected (β). Implies, an adequate sample needs to be included in the study for detecting an effect which is small or large, but it is to be ensured that minimum sample size is selected.
- Generally in a simple linear regression model, the value of standardized regression coefficient may vary from -1 to +1. Since there is only single predictor variable in the simple linear regression model,

minimum sample size calculated by the existing method will be same as that of proposed method at the value of standardized regression coefficientβ. But, proposed method suggests sufficient sample size that required for conducting analytical cross sectional study according to various values of β(unstandardized regression coefficient).The existing method only gives a minimum required sample size, and it is given at the value of standardized regression coefficient β (when the independent variable is at scale free unit).but the proposed method additionally suggestsan adequate sample size according to the effect that to be detected (β, unit of change in the outcome variable that resulted from unit change in the potential predictor) when the independent variable is in original units.This method incorporates 'degree of effect of the predictor' or 'unstandardized regression coefficientβ' which can be identified from the recent reliable studies or pilot studies. It helps to calculate 'adequate sample size' than considering the 'required minimum'.

Conclusion

Determination of adequate sample size is an important and essential step in any type of research. An inadequate sample size produces invalid statistical test results. Since application of statistical testing procedures in cross sectional studies is a normal practice, an appropriate formula for calculation ofsample size is suggested considering 'design of the study' and 'type of the test use'. In order to detect specific rate of change in the response variable (unstandardized regression coefficientβ), suitable sample sizes can be identified using the proposed method which is the major advantage in comparison with the existing method of sample size in cross sectional studies. The existing formula lacks adequate 'power'and 'focus of statistical test' to be used;therefore it is recommended to include power term (Z_β) in sample size formula along with other indicators such as significance level, standard deviation, margin of error, and degree of change to be detected.This helps for effective representation of the population in the selected sample that identifiesreal exposure effect and enables generalization of results.

References:

1. Thiese, M.S., Arnold, Z.C.&Walker S.D. (2015). The misuse and abuse of statistics in biomedical research. BiochemiaMedica, 25(1), 5-11.
2. Centre for Evidence-Based Medicine (CEBM). (2014).Study Designs. Retrieved from https://www.cebm.net/2014/04/study-designs/
3. Charan, J. & Biswas T.(2013).How to Calculate Sample Size for Different Study Designs in Medical Research? Indian Journal of Psychological Medicine, 35(2), 121-126.
4. Mann, C.J. (2012). Observational research methods––Cohort studies, cross sectional studies, and case–control studies. African Journal of Emergency Medicine, 2(1), 38-46
5. Scottish Government. (2017). Confidence Intervals. Retrieved fromhttps://www.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey/ConfidenceIntervals
6. Hazra, A. (2017). Using the confidence interval confidently. Journal of Thoracic Disease, 9(10),4125-4130.
7. Kasiulevičius, V., Sapoka V. &Filipavičiūtė R. (2016). Sample size calculation in epidemiological studies. Gerontologija, 7(4), 225–231.
8. Pourhoseingholi, M.A., VahediM. &Rahimzadeh M. (2013). Sample size calculation in medical studies. GastroenterolHepatol Bed Bench, 6(1),14-7.
9. Wang, H. & ChowS.C.(2007).Sample size calculation for comparing proportions. In: Wiley Encyclopedia of Clinical Trials.
10. Grimes, D.A & Schulz K.F. (2002). Descriptive studies: What they can and cannot do. Lancet, 359(9301),145-9.
11. Kim J.O., Mueller C. W. (1976). Standardized and unstandardized coefficients in causal analysis: An expository note. Sociological Methods & Research, 4(4), 423-438.
12. Greenland, S., Schlesselman, J.J., Criqui, M.H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. American Journal of Epidemiology, 123(2), 203–208.
13. Bring, J. (1994). How to Standardize Regression Coefficients. The American Statistician48(3), 209-213.
14. Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine,27,2865–2873.
15. Kim, R.S. (2011). Standardized Regression Coefficients as Indices of Effect Sizes in Meta-Analysis. Electronic Theses. Florida State University Libraries.
16. Hsieh, F.Y., DanielA.B. &Michael D.L. (1998). A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine, 17(14), 1623—1634.
17. Thigpen, C.C. (1987). A Sample-Size Problem in Simple Linear Regression.The American Statistician, 41(3), 214-215.
18. The BMJ. (2019).Correlation and regression. Retrieved from https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression