Health, 2014, 6, 2973-2998
Published Online December 2014 in SciRes. http://www.scirp.org/journal/health
http://dx.doi.org/10.4236/health.2014.621336

Simulation Program to Determine Sample Size and Power for a Multiple Logistic Regression Model with Unspecified Covariate Distributions

Naoko Kumagai[1,2*], Kohei Akazawa[3], Hiromi Kataoka[1], Yutaka Hatakeyama[1], Yoshiyasu Okuhara[1]

[1]Center of Medical Information Science, Kochi Medical School, Kochi University, Kochi, Japan
[2]Integrated Center for Advanced Medical Technologies, Kochi Medical School, Kochi University, Kochi, Japan
[3]Department of Medical Informatics, Niigata University Medical and Dental Hospital, Niigata, Japan
Email: *poosant@gmail.com, akazawa@med.niigata-u.ac.jp, kataokah@kochi-u.ac.jp, hatake@kochi-u.ac.jp, okuharay@kochi-u.ac.jp

Received 14 October 2014; revised 30 November 2014; accepted 13 December 2014

Academic Editor: Dr. Rodrigo Vale, Coordinator of the Laboratory of Exercise Physiology (LAFIEX), Estacio de Sá University, Rio de Janeiro, Brazil

Abstract

Binary logistic regression models are commonly used to assess the association between outcomes and covariates. Many covariates are inherently continuous, and have a variety of distributions, including those that are heavily skewed to the left or right. Existing theoretical formulas, criteria, and simulation programs cannot accurately estimate the sample size and power of non-standard distributions. Therefore, we have developed a simulation program that uses Monte Carlo methods to estimate the exact power of a binary logistic regression model. This power calculation can be used for distributions of any shape and covariates of any type (continuous, ordinal, and nominal), and can account for nonlinear relationships between covariates and outcomes. For illustrative purposes, this simulation program is applied to real data obtained from a study on the influence of smoking on 90-day outcomes after acute atherothrombotic stroke. Our program is applicable to all effect sizes and makes it possible to apply various statistical methods, logistic regression and related simulations such as Bayesian inference with some modifications.

[*]Corresponding author.

How to cite this paper: Kumagai, N., Akazawa, K., Kataoka, H., Hatakeyama, Y. and Okuhara, Y. (2014) Simulation Program to Determine Sample Size and Power for a Multiple Logistic Regression Model with Unspecified Covariate Distributions. *Health*, **6**, 2973-2998. http://dx.doi.org/10.4236/health.2014.621336

Keywords

Logistic Regression Model, Monte Carlo Simulation, Non-Standard Distributions, Nonlinear, Power, Sample Size, Skewed Distribution

1. Introduction

Logistic regression models have been used to determine the association between risk factors and outcomes in various fields, including medical and epidemiological research [1] [2]. However, they sometimes produce contradictory conclusions for the same hypothesis. For example, some studies have indicated that cigarette smoking enhances the risk of Barrett's Esophagus, whereas other studies have concluded that there is no association between the two because of a lack of power [3]. The robustness of such inferences is dependent on the relationship between sample size and power [4]. It is clearly important to calculate the sample size and estimate the power of observational studies, as well as randomized control studies, while accounting for the effects of other covariates.

Theoretical formulas, criteria, and software applications have been developed to enable the accurate determination of sample size and statistical power in a binary logistic regression model [5]-[11]. However, these tend to consider only specific, well-known probability distributions, even though it is clear that the power differs according to the shape of the covariate distribution. In practice, many covariates are inherently continuous, and their distributions take a variety of shapes (e.g., being heavily skewed to the left or right). Another problem is that the size of the effect can sometimes differ between outcomes and covariate. For example, J-shaped relationships are sometimes found in medical and epidemiological studies and an inverse relationship between diastolic pressure and adverse cardiac ischemic events (*i.e.*, the lower the diastolic pressure the greater the risk of coronary heart disease and adverse outcomes) has been observed in numerous studies [12]. The distribution shape and effect size of covariates must be carefully considered. Therefore we have developed a software program that uses Monte Carlo simulations to estimate the exact power of a logistic regression model corresponding to the actual data structure. This program has numerous advantages. It can handle any distribution shape and effect size and enables the application of various statistical methods, logistic regression, and other simulations such as Bayesian inference with some modifications. In this paper, we report the application of our simulation program to real data obtained from a study on the influence of smoking on 90-day outcomes after acute atherothrombotic stroke in 292 Japanese men [13].

2. Theoretical Background

2.1. Standard Binary Linear Logistic Regression Model

We consider a case-control study in which the binary response variable y denotes each patient's disease status (y = 1 for cases and y = 0 for controls). For each subject, we have a set of p covariates X_1, X_2 , $\cdots$, X_p. Let the conditional probability that an outcome is present be denoted by. The logit of the multiple logistic regression model is

$$g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \cdots, +\beta_p X_p,$$

in which case the logistic regression model is

$$\pi(X) = \frac{e^{g(X)}}{1+e^{g(X)}},$$

where β is an unknown parameter.

2.2. Two-Segment Logistic Regression Model for Nonlinear Association between a Logit Outcome and a Covariate

We replace the linear term associated with the covariate X_1 in the standard binary logistic regression model with a two-segment function containing a change-point. The relationship between the logit outcome and X_1 is different either side of this change-point. The two-segment logistic regression model shown in **Figure 1** can be expressed as follows

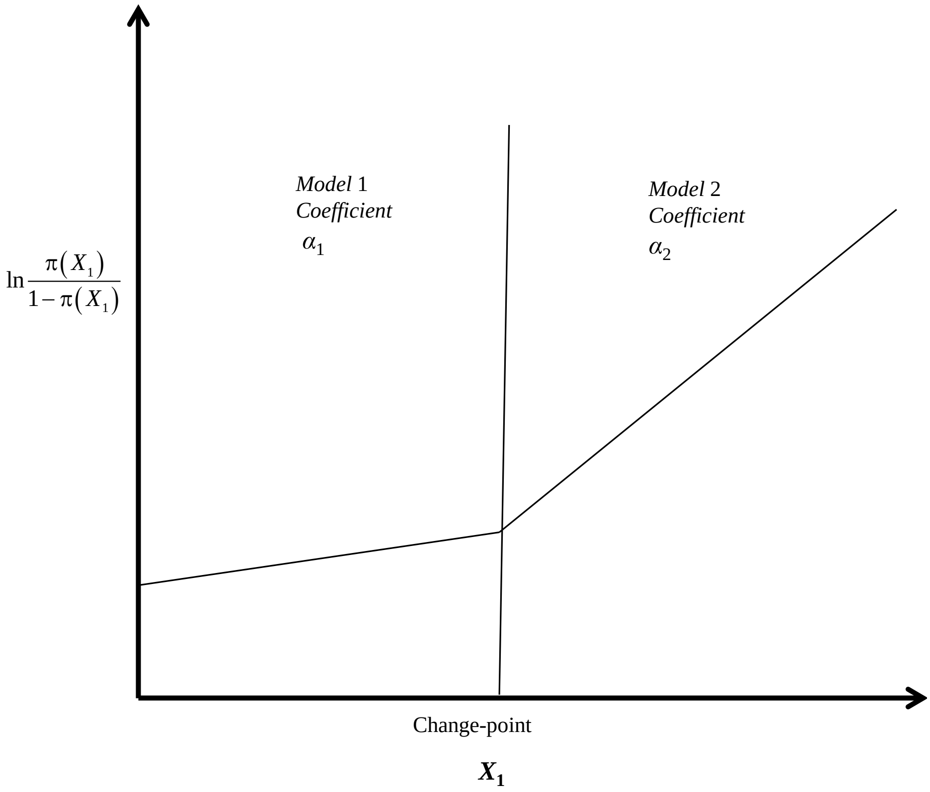


Figure 1. Two-segment logistic regression model for non-linear association between logit outcomes and covariates.

$$\text{If } X_1 \le \tau,\ g(X_1) = \beta_0 + \alpha_1 X_1$$

$$\text{If } X_1 > \tau,\ g(X_1) = \beta_0 + \alpha_1 \tau + \alpha_2 (X_1 - \tau)$$

where represents the value of the change-point and; α_1, α_2 are the unknown regression coefficients of X_1.

3. Methods

3.1. Outline of Simulation Program

Our Monte Carlo simulation program is written in the SAS/STAT/IML language; the source code is given in the appendix. The program consists of three parts: data generation, parameter estimations, and statistical power calculation. Users should modify and add to these conditions according to their specific purposes and interests. **Table 1** describes the input parameters required to run the program. Users should assign suitable values, as determined by the relevant test problem. **Table 2** describes some macro modules for modifying this program.

Continuous distributions are generated by specifying the mean, standard deviation, skewness, kurtosis, and correlation, or by assigning frequencies in each designated interval of a continuous variable (see **Figure 2**). The nonlinear relationship between the continuous covariates and the logit outcome can be specified by varying the regression coefficients on either side of a change point, as shown in **Figure 1**.

In the proposed program, users assign values for the proportion of events, sample size, Type I error, regression coefficients, distribution type (dichotomous, polytomous, or continuous), and distribution shape, as well as the quantile number for the categorization approach. The output of this program shows the average and standard error of each coefficient, as well as the power. A flowchart describing this program is shown in **Figure 3**.

The validity of the program is confirmed by comparing its results with those given by Hsieh's program. The output of our program is almost the same as that from Hsieh's program. For example, when the event proportion, sample size, and regression coefficient were set to 0.01, 12,580, and −0.223, respectively, our program estimated a power of 0.81, whereas Hsieh's gave a result of 0.8. When the event proportion, sample size, and regression

Table 1. Description of input parameters.

Input parameter	Explanation
SEED	Random number seed (should be a positive integer)
ALEVEL	Significance level of the statistical test (Type I error)
P	Event proportion (response probability)
NITER	Number of iterations performed
N_REPEAT	Number of iterations performed
	*NITER and N_REPEAT should be the same number
PATH	Directory in which results are saved
TABLE	Table name for saved results
R	Number of categorized groups
	Example: continuous = 1, median = 2, tertile = 3, quantile = 4
	*If model includes nominal variables, R should be >1
CHANGE_POINT	Change point (see **Figure 1**)
Regression coefficients for the covariates in the full model, except for predictors and intercept, specified as:	
MODEL_1	%NRSTR(α_1^{*}X1+, $\cdots$, + $\beta_i X_i$)
MODEL_2	%NRSTR(α_1^{*}(the value of change_point) + α_2^{*}(X1 − (the value of change_point) +, $\cdots$, + $\beta_i X_i$)
	α and β are the given regression coefficient values
*If model is linear, the regression coefficients α_1 and α_2 are the same	
Sample size, mean, standard deviation, skewness, kurtosis, and correlation are specified as:	
	Example DATA a (type=CORR); LENGTH _TYPE_ $40; INPUT _NAME_ $_TYPE_$ X1 X2 ; IF TRIM(LEFT(_TYPE_))=’N’ THEN call symput(‘NSP’, X1); CARDS; . MEAN 70 50 . STD 4 5 . N 300 300 X1 CORR 1 0 X2 CORR 0 1 ; RUN;
*If only one covariate is defined, the correlation should be set to 1. The sample size of all covariates should be the same.	
SKW_KRT	%NRSTR ({skewness 1 kurtosis 1, skewness 2 kurtosis 2, $\cdots$ })
*If covariates are normally distributed, both skewness and kurtosis are set to 0.	
LIST_VARNAME	%NRSTR (X1, X2, $\cdots$, Xi); list of variable names in A of above dataset
MIN	Minimum value of a continuous variable
MAX	Maximum value of a continuous variable
SUB_GROUP	Number of subgroups
CATEGORIZATION	%NRSTR (list of covariates to be categorized)
CATEGORIZATION_R	%NRSTR (list of new covariate names after categorization)
CONTI_MODEL	%NRSTR (list of covariates in a continuous logistic regression model)
*Even if some parameters are not needed, please assign all parameters and specify necessary variables in a logistic regression model.	

Table 2. Description of the macro module for modifications.

To assign the desired number of observations to each subgroup, as shown **Figure 1**, part of the %Ratio module must be modified.

For sample size, n_k, observation values of the k^{th} subgroup are extracted from a randomly generated U(a_k, b_k). U1 corresponds to the lowest interval subgroup, and U2 corresponds to the next lowest interval subgroup. For example, the minimum value, maximum value and number of subgroups are set to 1, 21 and 5, respectively. Therefore, the subgroups are (1, 5), (5, 9), (9, 13), (13, 17), (17, 21). The subgroups are assigned frequencies of 0.55, 0.05, 0.2, 0.15 and 0.05, respectively. &NSP denotes the total sample size; ID is the observation identification.

```
%MACRO RATIO;
IF 1=< ID <&NSP.*0.55 THEN _H1=U1;
ELSE IF &NSP.*0.55 =<ID <&NSP.*0.6 THEN _H1=U2;
ELSE IF &NSP.*0.6 =< ID <&NSP.*0.8 THEN _H1=U3;
ELSE IF &NSP.*0.8 =< ID <&NSP.*0.95 THEN _H1=U4;
ELSE _H1=U5;
%MEND RATIO;
```

If model includes discrete variables, then specify the model in part of PROC LOGISTIC in %MODEL_CATEGORICAL, and ensure the input parameter R is greater than 1.

```
%MACRO MODEL_CATEGORICAL;
ODS OUTPUT PARAMETERESTIMATES=PARAM_&R CONVERGENCESTATUS=STATUS_&R TYPE3=TYPE3_&R;
PROC LOGISTIC DATA=G&R;
/*******modification**********************/
CLASS C1(PARAM=REF REF="0")    D1(PARAM=REF REF="0") ;
MODEL Y(EVENT='1')= C1 X1 X2 X3 D1
/**********************************/
/TECH=NR MAXITER=8 XCONV=0.01;
BY STRATA; RUN;
%MEND;
```

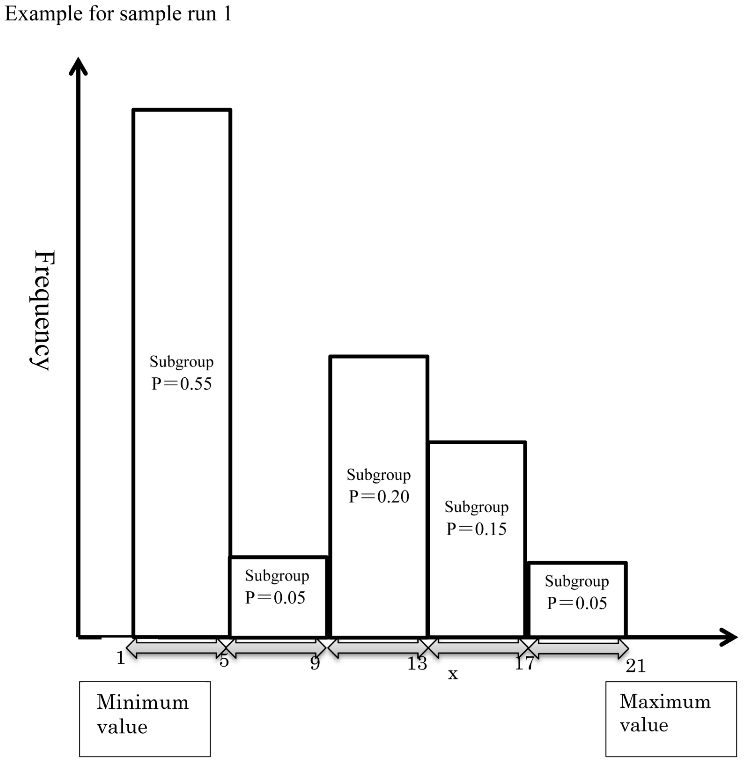


Figure 2. Algorism of generating a continuous covariate which has a unique distribution.

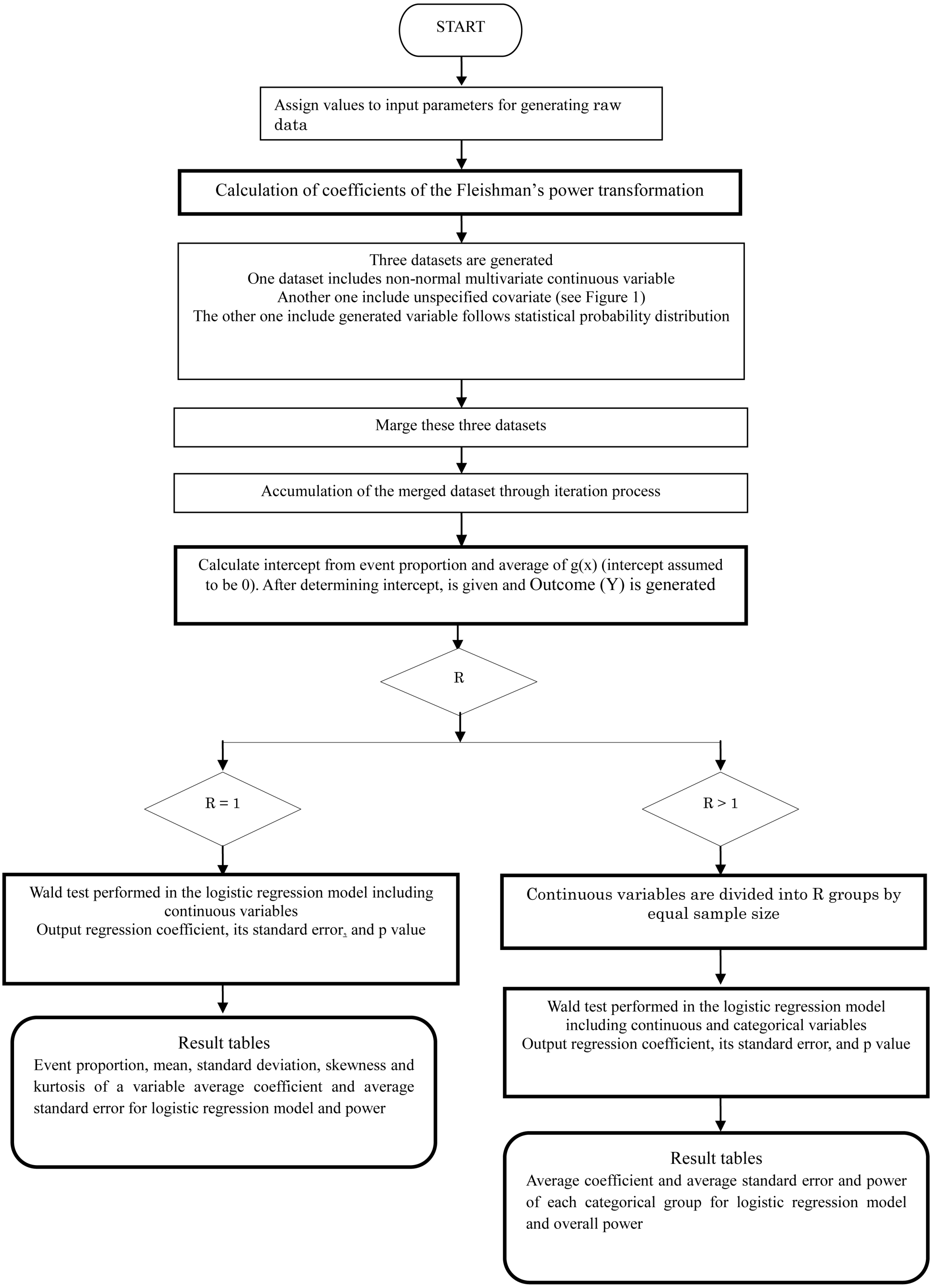

START
Assign values to input parameters for generating raw data
Calculation of coefficients of the Fleishman's power transformation
Three datasets are generated
One dataset includes non-normal multivariate continuous variable
Another one include unspecified covariate (see Figure 1)
The other one include generated variable follows statistical probability distribution
Marge these three datasets
Accumulation of the merged dataset through iteration process
Calculate intercept from event proportion and average of g(x) (intercept assumed to be 0). After determining intercept, is given and Outcome (Y) is generated
R
R = 1
R > 1
Wald test performed in the logistic regression model including continuous variables
Output regression coefficient, its standard error, and p value
Result tables
Event proportion, mean, standard deviation, skewness and kurtosis of a variable average coefficient and average standard error for logistic regression model and power
Continuous variables are divided into R groups by equal sample size
Wald test performed in the logistic regression model including continuous and categorical variables
Output regression coefficient, its standard error, and p value
Result tables
Average coefficient and average standard error and power of each categorical group for logistic regression model and overall power

Figure 3. Flow chart of simulation program.

coefficient were set to 0.5, 225, and 0.405, respectively, our program estimated a power of 0.89, which compares well with Hsieh's result of 0.9 [14].

3.2. Construction of Raw Simulation Data

3.2.1. Continuous Covariates

Non-normal or normal multivariate continuous variables are generated by specifying the mean, standard deviation, kurtosis, skewness, and correlation through a procedure in the %COEFF and %CONTINOUS SAS modules. A detailed explanation can be found in a book on SAS® for Monte Carlo Studies [15].

3.2.2. Continuous Covariates That Are Uniquely Distributed (Figure 2)

A continuous variable is divided into l subgroups of equal intervals as shown in **Figure 2**. The minimum value of the original covariate is assumed to be Min and the maximum value is assumed to be Max.

The length of the interval of each subgroup is $S = \frac{(\text{Max} - \text{Min})}{l}$.

The k^{th} subgroup ranges from a_k to b_k ($k = 1, \cdots, l$), where $k = 1$ indicates the lowest subgroup and l indicates the highest. a_k and b_k can be expressed as

$$a_k = \text{Min} + \text{s} \times (k - 1)$$

$$b_k = \text{Min} + \text{s} \times (k)$$

Random numbers from a uniform distribution on the interval (0, 1) are converted to a uniform distribution on the interval (a_k, b_k) with the equation $a_k + (b_k - a_k) \times$ (generated number). The k^{th} subgroup, consisting of n_k observations in the interval (a_k, b_k), is denoted by the variable H_1.

3.2.3. Statistical Probability Distribution

If the covariate is assumed to follow a probability distribution, the RAND function can be inserted into a macro PDF module. In the example given for this program, nominal variables are generated using the SAS TABLE function.

3.2.4. Determination of Binary Outcome

The individual probability of event occurrences is calculated from the assigned parameters and generated covariates using a logistic regression model. The initial intercept value is set to zero, and then the average is calculated. The intercept is determined from $\overline{\pi(X)}$ and p by the following equation:

$$\text{Intercept} = \ln \frac{p}{1-p} - \overline{\pi(X)}$$

After determining the intercept, the individual probability $\pi(\text{X}_i)$ (for $i = 1, \cdots, n$ observations) is calculated by a logistic regression model. The binary outcome Y is generated from the individual $\pi(\text{x}_i)$ and random numbers from a uniform distribution on the interval (0, 1). If $\pi(\text{x}_i)$ is less than the corresponding random number, $\text{Y}_i = 1$ (denoting that the event occurred); otherwise, $\text{Y}_i = 0$. Finally, we have a dataset consisting of a covariate and response variable (Y). For skewed distributions, the event proportion of the generated dataset might not be the same as the input value. However, our program outputs the event proportion of this dataset. This difference can be adjusted by changing the input parameters of the event proportion.

3.2.5. Estimation of Regression Coefficients and Standard Errors

We conducted a logistic regression analysis in a model including continuous and/or design variables to obtain maximum likelihood estimates of and the significance level, or p-value, for the null hypothesis with population regression coefficient $\beta = 0$.

There is a possibility of non-convergence if the data are completely or partially separated. This is because one or more parameters in the model become theoretically infinite, and it may not be possible to obtain reliable maximum likelihood estimates [16]. These instances of non-convergence must be appropriately handled. Our simulation program overcomes this problem by neglecting samples that lead to non-convergence.

3.2.6. Test Module

The test module outputs the mean of the asymptotic standard error, and the statistical power. The proportion of tests in which the *p*-value is less than the Type I error level is defined as the power.

4. Sample Runs

4.1. Sample Run 1

Variable H_1 was set to be the National Institute of Health Stroke Scale (NIHSS, a tool used by healthcare providers to objectively quantify the impairment caused by a stroke). The minimum, maximum, and the number of subgroups were set as 1, 21, and 5, respectively. Therefore, the subgroups ranged from (1, 5), (5, 9), (9, 13), (13, 17), and (17, 21). The frequency of each subgroup was assumed to be 0.55, 0.05, 0.2, 0.15, and 0.05, respectively. The generated numbers were rounded off, and the event proportion was set to 0.2. Regression coefficient parameters (,) were taken as (0.00, 0.00), (0.06, 0.06), and (0.06, 0.15), and the change point was set to 4. H_1 was set to be either continuous, divided at median or tertile points, or categorized into three groups: 1 - 4, 5 - 15, and ≥16. We executed the logistic model for these values of H_1, and present the results in **Table 3**. When and were set to 0.06 and 0.06, the average coefficient value was correctly estimated to be 0.062. When these parameters were set to 0.06 and 0.15, the categorization using the change point produced higher coefficient values than that using the tertile points. Moreover, when and were set to 0.0 and 0.0, the power was approximately 0.05, the same as the Type I error.

4.2. Sample Run 2

We used age and systolic arterial pressure as continuous variables X_1 and X_2, respectively. The mean and standard deviation of X_1 were 70 and 8, and the skewness and kurtosis were set to combinations of 0 and 0, −0.5 and 0.5, and −1.0 and 1.0. The regression coefficient of X_1 was 0.05 under a linear relationship. The mean and standard deviation of X_2 were 160 and 25, and the skewness and kurtosis were set to combinations of 0 and 0, 0.4 and 0.3, and 0.8 and 0.6. The regression coefficient of X_2 was set to 0.02 as a linear relationship. The correlation between the variables was set to 0, 0.3, and 0.6. The binary variable D_1 denotes smoking or non-smoking. The proportion of non-smokers and smokers was 0.5 and 0.5, and the regression coefficient was 0.83. The sample size was set to 300 and 500. We executed the logistic model for X_1, X_2, and D_1. The results are shown in **Table 4**.

Table 3. Sample run 1: estimated power of the Wald test in two-segment logistic regression model with an event proportion of 0.2.

α_1	α_2	Categorization	Coefficient	SE	Power
0.00	0.00	Continuous	0.001	0.027	0.061
		Median	0.006	0.294	0.055
		Tertile	0.013	0.367	0.049
		1 - 4	0.013	0.368	
		5 - 15	−0.001	0.314	0.043
		≥16	−0.033	0.546	
0.06	0.06	Continuous	0.062	0.026	0.682
		Median	0.616	0.293	0.545
		Tertile	0.238	0.397	0.459
		1 - 4	0.768	0.374	
		5 - 15	0.530	0.313	0.516
		≥16	0.904	0.477	
0.06	0.15	Continuous	0.153	0.027	1.000
		Median	1.527	0.309	0.999
		Quantile	0.588	0.456	1.000
		1 - 4	1.896	0.417	
		5 - 15	1.323	0.326	1.000
		≥16	2.287	0.468	

Table 4. (a) Sample run 2: estimated power of the Wald test for left- and right-skewed distributions with an event proportion of 0.2 and N = 300; (b) Sample run 2: estimated power of the Wald test for left- and right-skewed distributions with an event proportion of 0.2 and N = 500.

(a)

X_1 X_2			X_1	Left-skewed covariate		X_2	Right-skewed covariate					D_1 binary		
(skewness, kurtosis)		Mean	SD	(S, K) Coefficient	SE	Power	Mean	SD	(S, K) Coefficient	SE	Power	Coefficient SE	SE	Power
		70	8	$\beta_1 = 0.05$			160	25	$\beta_2 = 0.02$			$\beta_2 = 0.83$		
Correlation of (X_1, X_2) = 0.0														
(0.0, 0.0)	(0.0, 0.0)	70.0	8.0	(0.0, 0.0) 0.052	0.019	0.774	160.0	24.9	(0.0, 0.0) 0.020	0.006	0.931	0.860	0.304	0.801
(−0.5, 0.5)	(0.0, 0.0)	70.0	8.0	(−0.5, 0.5) 0.052	0.020	0.754	160.0	24.9	(0.0, 0.0) 0.020	0.006	0.933	0.864	0.304	0.815
(−1.0, 1.0)	(0.0, 0.0)	70.0	8.0	(−1.0, 1.0) 0.053	0.021	0.727	160.0	24.9	(0.0, 0.0) 0.020	0.006	0.933	0.864	0.304	0.818
(0.0, 0.0)	(0.4, 0.3)	70.0	8.0	(0.0, 0.0) 0.052	0.019	0.764	160.0	24.9	(0.4, 0.3) 0.020	0.006	0.948	0.861	0.305	0.805
(0.0, 0.0)	(0.8, 0.6)	70.0	8.0	(0.0, 0.0) 0.052	0.019	0.766	160.0	24.9	(0.8, 0.6) 0.020	0.006	0.957	0.861	0.306	0.805
(−0.5, 0.5)	(0.4, 0.3)	70.0	8.0	(−0.5, 0.5) 0.052	0.020	0.753	160.0	24.9	(0.4, 0.3) 0.020	0.006	0.948	0.863	0.305	0.812
(−1.0, 1.0)	(0.8, 0.6)	70.0	8.0	(−1.0, 1.0) 0.053	0.021	0.723	160.0	24.9	(0.8, 0.6) 0.020	0.006	0.955	0.861	0.305	0.816
Correlation of (X_1, X_2) = 0.3														
(0.0, 0.0)	(0.0, 0.0)	70.0	8.0	(0.0, 0.0) 0.051	0.020	0.742	160.0	25.0	(0.0, 0.0) 0.021	0.007	0.910	0.871	0.305	0.820
(−0.5, 0.5)	(0.0, 0.0)	70.0	8.0	(−0.5, 0.5) 0.052	0.021	0.704	160.0	25.0	(0.0, 0.0) 0.021	0.007	0.910	0.871	0.304	0.830
(−1.0, 1.0)	(0.0, 0.0)	70.0	8.0	(−1.0, 1.0) 0.052	0.022	0.665	160.0	25.0	(0.0, 0.0) 0.021	0.007	0.906	0.875	0.306	0.826
(0.0, 0.0)	(0.4, 0.3)	70.0	8.0	(0.0, 0.0) 0.052	0.020	0.747	160.0	25.0	(0.4, 0.3) 0.020	0.006	0.922	0.873	0.303	0.827
(0.0, 0.0)	(0.8, 0.6)	70.0	8.0	(0.0, 0.0) 0.052	0.020	0.755	160.0	25.0	(0.8, 0.6) 0.020	0.006	0.935	0.874	0.305	0.823
(−0.5, 0.5)	(0.4, 0.3)	70.0	8.0	(−0.5, 0.5) 0.052	0.021	0.704	160.0	25.0	(0.4, 0.3) 0.020	0.006	0.926	0.874	0.305	0.829
(−1.0, 1.0)	(0.8, 0.6)	70.0	8.0	(−1.0, 1.0) 0.052	0.023	0.662	160.0	25.0	(0.8, 0.6) 0.020	0.006	0.938	0.874	0.305	0.829
Correlation of (X_1, X_2) = 0.6														
(0.0, 0.0)	(0.0, 0.0)	70.0	8.0	(0.0, 0.0) 0.052	0.024	0.587	160.0	25.0	(0.0, 0.0) 0.021	0.008	0.779	0.873	0.305	0.817
(−0.5, 0.5)	(0.0, 0.0)	70.0	8.0	(−0.5, 0.5) 0.052	0.025	0.542	160.0	25.0	(0.0, 0.0) 0.020	0.008	0.780	0.875	0.304	0.823
(−0.1, 1.0)	(0.0, 0.0)	70.0	8.0	(−1.0, 1.0) 0.052	0.027	0.508	160.0	25.0	(0.0, 0.0) 0.021	0.008	0.805	0.875	0.303	0.819
(0.0, 0.0)	(0.4, 0.3)	70.0	8.0	(0.0, 0.0) 0.052	0.024	0.595	160.0	25.0	(0.4, 0.3) 0.020	0.007	0.812	0.873	0.307	0.822
(0.0, 0.0)	(0.8, 0.6)	70.0	8.0	(0.0, 0.0) 0.052	0.024	0.608	160.0	25.0	(0.8, 0.6) 0.020	0.007	0.844	0.873	0.308	0.815
(−0.5, 0.5)	(0.4, 0.3)	70.0	8.0	(−0.5, 0.5) 0.052	0.025	0.557	160.0	25.0	(0.4, 0.3) 0.020	0.007	0.815	0.872	0.305	0.818
(−1.0, 1.0)	(0.8, 0.6)	70.0	8.0	(−1.0, 1.0) 0.052	0.026	0.530	160.0	25.0	(0.8, 0.6) 0.020	0.007	0.878	0.869	0.305	0.812

S, skewness; K, kurtosis.

(b)

X_1 X_2		X_1	Left-skewed covariate		X_2	Right-skewed covariate					D_1 binary		
(skewness, kurtosis)	Mean	SD	(S, K) Coefficient	SE	Power	Mean	SD	(S, K) Coefficient	SE	Power	Coefficient SE	SE	Power
	70	8	$\beta_1 = 0.05$			160	25	$\beta_2 = 0.02$			$\beta_2 = 0.83$		
Correlation of (X_1, X_2) = 0.0													
(0.0 0.0, 0.0 0.0)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.943	160.1	25.0	(0.0, 0.0) 0.020	0.005	0.997	0.846	0.233	0.947
(−0.5 0.5, 0.0 0.0)	70.0	8.0	(−0.5, 0.5) 0.051	0.015	0.930	160.1	25.0	(0.0, 0.0) 0.020	0.005	0.995	0.846	0.233	0.945
(−1.0 1.0, 0.0 0.0)	70.0	8.0	(−1.0, 1.0) 0.051	0.016	0.919	160.1	25.0	(0.0, 0.0) 0.020	0.005	0.996	0.843	0.233	0.942
(0.0 0.0, 0.4 0.3)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.945	160.0	25.0	(0.4, 0.3) 0.020	0.005	0.996	0.846	0.234	0.947
(0.0 0.0, 0.8 0.6)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.938	160.0	25.0	(0.8, 0.6) 0.020	0.005	0.997	0.843	0.234	0.950
(−0.5 0.5, 0.4 0.3)	70.0	8.0	(−0.5, 0.5) 0.051	0.015	0.929	160.0	25.0	(0.4, 0.3) 0.020	0.005	0.996	0.845	0.234	0.949
(−1.0 1.0, 0.8 0.6)	70.0	8.0	(−1.0, 1.0) 0.051	0.016	0.922	160.0	25.0	(0.8, 0.6) 0.020	0.004	0.999	0.841	0.234	0.945
Correlation of (X_1, X_2) = 0.3													
(0.0, 0.0) (0.0, 0.0)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.919	160.0	25.0	(0.0, 0.0) 0.020	0.005	0.984	0.852	0.234	0.957
(−0.5, 0.5) (0.0, 0.0)	70.0	8.0	(−0.5, 0.5) 0.051	0.016	0.903	160.0	25.0	(0.0, 0.0) 0.020	0.005	0.988	0.852	0.233	0.960
(−1.0, 1.0) (0.0, 0.0)	70.0	8.0	(−1.0, 1.0) 0.052	0.017	0.880	160.0	25.0	(0.0, 0.0) 0.020	0.005	0.987	0.853	0.233	0.965
(0.0, 0.0) (0.4, 0.3)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.920	160.0	25.0	(0.4, 0.3) 0.020	0.005	0.990	0.850	0.234	0.953
(0.0, 0.0) (0.8, 0.6)	70.0	8.0	(0.0, 0.0) 0.051	0.015	0.920	160.0	25.0	(0.8, 0.6) 0.020	0.005	0.991	0.849	0.235	0.952
(−0.5, 0.5) (0.4, 0.3)	70.0	8.0	(−0.5, 0.5) 0.051	0.016	0.902	160.0	25.0	(0.4, 0.3) 0.020	0.005	0.992	0.852	0.234	0.954
(−1.0, 1.0) (0.8, 0.6)	70.0	8.0	(−1.0, 1.0) 0.052	0.017	0.879	160.0	25.0	(0.8, 0.6) 0.020	0.005	0.991	0.850	0.234	0.957
Correlation of (X_1, X_2) = 0.6													
(0.0, 0.0) (0.0, 0.0)	70.0	8.0	(0.0, 0.0) 0.051	0.018	0.811	160.0	25.0	(0.0, 0.0) 0.021	0.006	0.949	0.853	0.235	0.962
(−0.5, 0.5) (0.0, 0.0)	70.0	8.0	(−0.5, 0.5) 0.051	0.019	0.772	160.0	25.0	(0.0, 0.0) 0.020	0.006	0.945	0.851	0.234	0.960
(−1.0, 1.0) (0.0, 0.0)	70.0	8.0	(−1.0, 1.0) 0.052	0.020	0.753	160.0	25.0	(0.0, 0.0) 0.020	0.006	0.951	0.853	0.233	0.966
(0.0, 0.0) (0.4, 0.3)	70.0	8.0	(0.0, 0.0) 0.051	0.018	0.815	160.0	25.0	(0.4, 0.3) 0.020	0.006	0.959	0.852	0.235	0.946
(0.0, 0.0) (0.8, 0.6)	70.0	8.0	(0.0, 0.0) 0.051	0.018	0.819	160.0	25.0	(0.8, 0.6) 0.020	0.005	0.970	0.850	0.236	0.947
(−0.5, 0.5) (0.4, 0.3)	70.0	8.0	(−0.5, 0.5) 0.051	0.019	0.785	160.0	25.0	(0.4, 0.3) 0.020	0.006	0.959	0.853	0.235	0.956
(−1.0, 1.0) (0.8, 0.6)	70.0	8.0	(−1.0, 1.0) 0.052	0.020	0.757	160.0	25.0	(0.8, 0.6) 0.020	0.005	0.979	0.850	0.234	0.966

S, skewness; K, kurtosis.

The mean, standard deviation, skewness, and kurtosis in the generated variables were almost equal to those of their input parameters. An increase in sample size leads to higher power, whereas a higher correlation produces a lower power. Our results clearly illustrate that the power will differ depending on the shape of the distribution. Negatively skewed distributions exhibit low power, whereas a positive skew results in high power. There is an inverse relationship between the logit outcome and the covariates. For example, when the skewness of X_1 was changed from −0.5 to −1.0, the power decreased from 0.75 to 0.73. When the skewness of X_2 was changed from 0.4 to 0.8, the power increased from 0.81 to 0.84 or 0.88 for a sample size of 300 and correlation of 0.0.

4.3. Sample Run 3: Epidemiological Studies

It is important to establish that the results observed in the above simulations hold for real data. For this purpose, we used data from a study of the influence of smoking on 90-day outcomes after acute atherothrombotic stroke in 292 Japanese men [14]. In this study, body temperature, age, NIHSS score at admission, systolic blood pressure, and smoking status were included in the logistic model. Detailed input parameter information is given in **Table 5(a)**, and the estimated results are listed in **Table 5(b)**. The event proportion of this real study was 0.2. We obtained an event proportion of 0.206 in the generated dataset by setting an input value of 0.15 for the event proportion. The estimated coefficients were similar to the results of the epidemiological study. Real data analysis showed that all factors, *i.e.* body temperature, age, NIHSS score at admission, systolic blood pressure, and smoking status, were significantly associated with the outcome ($p < 0.05$), and our results also exhibited high power (minimum to maximum of 0.686 to 1.000).

Table 5. (a) Assigned input parameters for sample run 3; (b) Sample run 3: estimated power of the Wald test for an epidemiological study with a sample size of 292.

(a)

Input parameter	Explanation
SEED	9
ALEVEL	0.05
P	0.15
NITER	1000
PATH	C:\
TABLE	Table_samplerun_3
R	2
CHANGE_POINT	4
MODEL_1	%NRSTR(0.04*H1+0.8*D1+ 0.06*X1 + 0.02*X2 + 1.1*X3)
MODEL_2	%NRSTR(0.04*4+0.15*(H1-4)+0.8*D1+0.06*X1+0.02*X2 +1.1*X3)
	DATA a (type=CORR); LENGTH _TYPE_ $40; INPUT _NAME_ $_TYPE_$ X1 X2 X3 ; IF TRIM(LEFT(_TYPE_))=’N’ THEN call symput(‘NSP’,X1); CARDS; . MEAN 70 160 36.4 . STD 8 25 0.5 . N 292 292 292 X1 CORR 1 -0.1 -0.1 X2 CORR -0.1 1 0.1 X3 CORR -0.1 0.1 1 ; RUN;
SKW_KRT	%NRSTR({-0.5 0.5, 0.4 0.3, -0.08 0.7})
LIST_VARNAME	%NRSTR(X1 X2 X3)
CONTI_MODEL	%NRSTR(X1 X2 X3 C1)
Min=	1
Max=	21
SUB_GROUP=	5
CATEGORIZATION	%NRSTR(H1)
CATEGORIZATION_	%NRSTR(H1_R)

(b)

Risk factor	Results of epidemiological study		Results of simulation	
	Coefficient	*p* value	Coefficient	Power
Smoker	0.82	0.019	0.83	0.686
Age (years)	0.06	0.014	0.06	0.792
Systolic arterial pressure	0.02	0.0096	0.02	0.847
Body temperature	1.18	0.0013	1.14	0.893
NIHSS score at admission				
5 - 15	1.40	0.001	1.33	1.000
≥16	2.25	0.001	2.30	

Event proportion = 0.206.

5. Conclusions and Discussion

Estimating the sample size or inference of statistical power is critical. If the sample size is too low, the experiment will lack the precision needed to provide reliable answers to the questions it is investigating. If the sample size is too large, time and resources will be wasted, often for minimal gain [17]. In this study, we developed a Monte-Carlo simulation program that estimates the powers of covariates in the binary logistic regression model. Users can evaluate the relationship between sample size and covariates, in observational and power randomized studies. In this situation, our simulation results clearly indicated the relationship between statistical power and covariate distribution shape, as shown by the data in **Table 4**. Right- and left-skewed distributions exhibit different powers. This phenomenon has clarified that the shape of a distribution affects its statistical power [18] [19]. The advantage of using a theoretical equation to estimate the power is that it is quick and easy to implement using existing software. For this reason, power equations are used to inform most studies. However, in practical analysis, we must often compute the power with a relatively complex distribution.

Our program is flexible enough to accommodate any number or type (continuous or discrete) of covariate and categorization, continuous distribution shapes and correlations, and the association level between logit outcome and covariates, although some modifications may be necessary. This program can also be applied to other statistical methods, logistic regression and Bayesian inference. The SAS/STAT/IML program written for the simulations and a user manual are available upon request [20] [21].

Acknowledgements

The authors wish to thank Dr. Motonori Hatta for helpful advice on the study.

Conflict of Interest

The authors have no conflicts of interest to declare.

Contribution

K.A. and N.K. were responsible for the study conception. N.K. wrote the program and drafted this manuscript. N.K., K.A., Y.O., H.K., and Y.H. made revisions to the manuscript. All authors approved the final version of the manuscript.

References

[1] Ottenbacher, K.J., Ottenbacher, H.R., Tooth, L. and Ostir, G.V. (2004) A Review of Two Journals Found That Articles Using Multivariable Logistic Regression Frequently Did Not Report Commonly Recommended Assumptions. *Journal of Clinical Epidemiology*, **57**, 1147-1152. http://dx.doi.org/10.1016/j.jclinepi.2003.05.003

[2] Brenner, H. and Blettner, M. (1997) Controlling for Continuous Confounders in Epidemiologic Research. *Epidemiology*, **8**, 429-434. http://dx.doi.org/10.1097/00001648-199707000-00014

[3] Andrici, J., Cox, M.R. and Eslick, G.D. (2013) Cigarette Smoking and the Risk of Barrett's Esophagus. A Systematic Review and Meta-Analysis. *Journal of Gastroenterology and Hepatology*, **28**, 1258-1273. http://dx.doi.org/10.1111/jgh.12230

[4] Bergtold, J., Yeager, E. and Featherstone, A. (2011) Sample Size and Robustness of Inferences from Logistic Regression in the Presence of Nonlinearity and Multicollinearity. *The Agricultural & Applied Economics Association's* 2011 *AAEA & NAREA Joint Annual Meeting*, Pittsburgh, Pennsylvania, 24-26 July 2011.

[5] Demidenko, E. (2007) Sample Size Determination for Logistic Regression Revisited. *Statistics in Medicine*, **26**, 3385-3397. http://dx.doi.org/10.1002/sim.2771

[6] Whittemore, A.S. (1981) Sample Size for Logistic Regression with Small Response Probability. *Journal of the American Statistical Association*, **76**, 27-32. http://dx.doi.org/10.1080/01621459.1981.10477597

[7] Hsieh, F.Y., Bloch, D.A. and Larsen, M.D. (1998) A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Statistics in Medicine*, **17**, 1623-1634. http://dx.doi.org/10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S

[8] Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996) A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, **49**, 1373-1379. http://dx.doi.org/10.1016/S0895-4356(96)00236-3

[9] Vittinghoff, E. and McCulloch, C.E. (2007) Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, **165**, 710-718. http://dx.doi.org/10.1093/aje/kwk052

[10] SAS/STAT(R) 9.2 User's Guide, Second Edition. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_power_sect003.htm

[11] Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, John Wiley & Sons, New York. http://dx.doi.org/10.1002/0471722146

[12] Messerli, F.H. and Panjrath, G.S. (2009) The J-Curve between Blood Pressure and Coronary Artery Disease or Essential Hypertension: Exactly How Essential? *Journal of the American College of Cardiology*, **54**, 1827-1834. http://dx.doi.org/10.1016/j.jacc.2009.05.073

[13] Kumagai, N., Okuhara, Y., Iiyama, T., Fujimoto, Y., Takekawa, H., Origasa, H., Kawanishi, Y. and Yamaguchi, T. (2013) Effects of Smoking on Outcomes after Acute Atherothrombotic Stroke in Japanese Men. *Journal of the Neurological Sciences*, **335**, 164-1168. http://dx.doi.org/10.1016/j.jns.2013.09.023

[14] Hsieh, F.Y. (1989) Sample Size Tables for Logistic Regression. *Statistics in Medicine*, **8**, 795-802. http://dx.doi.org/10.1002/sim.4780080704

[15] Fan, X., Felsovalyi, A., Sivo, S.A. and Keenan, S.C. (2003) SAS® for Monte Carlo Studies: A Guide for Quantitative Researchers. SAS Institute, Cary.

[16] Webb, M.C., Wilson, J.R. and Chong, J. (2004) An Analysis of Quasi-Complete Binary Data with Logistic Models: Applications to Alcohol Abuse Data. *Journal of Data Science*, **2**, 273-285.

[17] Arnold, B.F., Hogan, D.R., Colford Jr., J.M. and Hubbard, A.E. (2011) Simulation Methods to Estimate Design Power: An Overview for Applied Research. *BMC Medical Research Methodology*, **11**, 94. http://dx.doi.org/10.1186/1471-2288-11-94

[18] Royston, P. and Sauerbrei, W. (2005) Building Multivariable Regression Models with Continuous Covariates in Clinical Epidemiology—With an Emphasis on Fractional Polynomials. *Methods of Information in Medicine*, **44**, 561-571.

[19] Grund, B. and Sabin, C. (2010) Analysis of Biomarker Data: Logs, Odds Ratios, and Receiver Operating Characteristic Curves. *Current Opinion in HIV & AIDS*, **5**, 473-479. http://dx.doi.org/10.1097/COH.0b013e32833ed742

[20] Li, A. (2013) Handbook of SAS® DATA Step Programming. Chapman and Hall & CRC, London.

[21] Burlew, M.M. (2007) SAS Macro Programming Made Easy. SAS Institute, Cary.

Appendix: Simulation Program for Estimating the Statistical Power of Logistic Regression Model

```
%LET SEED=;
%LET ALEVEL=;
%LET PATH=;
%LET TABLE=;
%LET NITER=;
%LET SKW_KRT=%NRSTR({});
%LET LIST_VARNAME=%NRSTR();
%LET MODEL_1=%NRSTR();
%LET MODEL_2=%NRSTR();
%LET CATEGORIZATION=%NRSTR();
%LET CATEGORIZATION_R=%NRSTR();
%LET CONTI_MODEL=%NRSTR();
%LET CHANGE_POINT=;
%LET MAX=;
%LET MIN=;
%LET SUB_GROUP=;
%LET P=;
/*EXAMPLE*/
DATA A (TYPE=CORR);
LENGTH _TYPE_    $40;
    INPUT    _NAME_ $ _TYPE_$     X1 X2 ;
    IF TRIM(LEFT(_TYPE_))='N'THENCALL SYMPUT('NSP', X1);
    CARDS;
.               MEAN            70        160
.               STD             8         25
.               N               300       300
X1              CORR            1         0
X2              CORR            0         1
;
RUN;
%DATASET(N_REPEAT=);
%LR(R=);
/*************************%COEFF and %CONTINUOUS **************/
/*%COEFF and %CONTINUOUS generate random variables following a Multivariate /*Normal distribution
with given means, standard deviations, and correlation matrix, /*and then transform each variable to the desired
distributional shape with specified /*population univariate skewness and kurtosis
/*%COEFF
/*Macro COEFF calculates coefficients of the Fleishman’s power transformation
/*Equation X= A + B*C1 + C*C2^2 +D*C3^3 where A=-C
/*Parameters
/*SKW_KRT; %NRSTR({skewness1 kurtosis1, skewness2 kurtosis2,…, });
/*LIST_VARNAME; list of variable names that define the skewness and kurtosis.
/*OUT the name of the output file (name of COEFF) that has thecoefficient values (A B C) of each variable.
/**************************************************************/
%MACROCOEFF;
PROC IML;
/* COEFFICIENTS OF B, C, D FOR FLEISHMAN'S POWER TRANSFORMATION*/
SKEWKURT=&SKW_KRT;
MAXITER=25;
CONVERGE=.000001;
```

```
START FUN;
    C1=COEF [1];
    C2=COEF [2];
    C3=COEF [3];
F=(C1**2+6*C1*C3+2*C2**2+15*C3**2-1)//
    (2*C2*(C1**2+24*C1*C3 +105*C3**2+2)-SKEWNESS)//
    (24*(C1*C3+C2**2*(1+C1**2+28*C1*C3)+C3**2*
    (12+48*C1*C3+141*C2**2+225*C3**2))-KURTOSIS);
FINISH FUN;

START DERIV;
J=         ((2*C1+6*C3) || (4*C2) || (6*C1+30*C3))//
    ((4*C2*(C1+12*C3)) || (2*(C1**2+24*C1*C3+105*C3**2+2))||
    (4*C2*(12*C1+05*C3)))//((24*(C3+C2**2*(2*C1+28*C3)+48*C3**3))||
    (48*C2*(1+C1**2+28*C1*C3+141*C3**2))||
    (24*(C1+28*C1*C2**2+2*C3*(12+48*C1*C3+141*C2**2+225*C3**2)
    +C3**2*(48*C1+450*C3))));
FINISH DERIV;

START NEWTON;
RUN FUN;
    DO ITER = 1 TO MAXITER
    WHILE (MAX(ABS(F))> CONVERGE);
           RUN DERIV;
           DELTA=-SOLVE(J,F);
           COEF=COEF+DELTA;
           RUN FUN;
    END;
FINISH NEWTON;

DO;
    NUM=NROW (SKEWKURT);
    DO VAR = 1 TO NUM;
           SKEWNESS = SKEWKURT [VAR,1];
           KURTOSIS = SKEWKURT [VAR, 2];
           COEF = {1.0, 0.0, 0.0};
           RUN NEWTON;
           COEF = COEF`;
           SK_KUR= SKEWKURT [VAR,];
           COMBINE=SK_KUR||COEF;
           IF VAR = 1 THEN RESULT=COMBINE;
           IF VAR >1 THEN RESULT=RESULT//COMBINE;
    END;
END;

RESULT=RESULT`;
CREATE _COEF_ FROM    RESULT [COLNAME={&LIST_VARNAME}];
APPEND FROM RESULT;

DATA _COEF;
    SET _COEF_;
    LENGTH _TYPE_ $40;
    MARK = _N_; _TYPE_="COEFF";
    FORMAT MARK;
RUN;

DATA COEFF (DROP= MARK);
    SET _COEFF;
    IF MARK >2 THEN OUTPUT COEFF;
RUN;
%MEND COEFF;
/***********************%CONTINUOUS********************************/
/*This program generates random variables following a Multivariate Normal /*distribution with given name,
standard deviation, and correlation matrix, and then /*transforms each variable to the desired distributional
shape with Fleishman’s /*coefficient.
/*Parameter
/*N_Repeat; the number of iterations
/*SEED; seed of the random number generator
/*DATA the name, A, of the input file that determines the characteristics of the random /*numbers to be generated.
The file specifies the mean, standard deviation, number of /*observations of each random number, and the
correlation coefficients between the /*variables. It must be a TYPE=CORR file, and its structure must comply
with that of /*such files. The file has _Type_=MEAN, STD, N, CORR. Its variables are _TYPE_, /*_NAME_
and the variables to be generated. The number of observations should be /*the same value. In this file, the sample
size 'NSP' should be specified as a parameter, /*using IF TRIM(LEFT(_TYPE_))='N' THEN CALL
SYMPUT('NSP', X1 (one of the /*variable names)).
/*
/*Example
/*DATA A (TYPE=CORR);
/*LENGTH _TYPE_    $40;
/*         INPUT    _NAME_ $    _TYPE_$      X1              X2 ;
/*         IF TRIM(LEFT(_TYPE_))='N' THEN CALL SYMPUT('NSP', X1);
/*         CARDS;
/*  .                MEAN            70           160
/*  .                STD             8            25
/*  .                N               300          300
/*         X1  CORR                  1            0
/*         X2  CORR                  0            1
/*         ;
/*         RUN;
/*OUT random variables generated according to the file given in parameter DATA and observation identification
number (ID)
/*****************************************************************/

%MACROCONTINUOUS;

PROC CONTENTS DATA=A (DROP=_TYPE_ _NAME_)
OUT=_DATA_ (KEEP=NAME) NOPRINT;
RUN;

/*SUPPOSE WE HAVE X1,......, X_P VARIABLE IN DATASET A WHICH IS AN INPUT DATASET.
WE ASSIGN THESE VARIABLES AS NAME OF V1,..., V_P MACRO REFERENCE OF &NV IS ASSIGNED
THE NUMBER OF TOTAL VARIABLES*/

DATA _DATA_;
SET _LAST_    END=END;
          RETAIN N 0;
```

```
        N=N+1;
        V=COMPRESS('V' || COMPRESS(PUT (N, 6.0)));
        CALL SYMPUT(V, NAME);
        IF END THEN CALL SYMPUT('NV', LEFT(PUT (N, 6.0)));
RUN;

%LET VNAMES=&V1;
%DO I=2%TO&NV.;
    %LET VNAMES=&VNAMES &&V&I;
%END;

/*OBTAIN THE MATRIX OF FACTOR PATTERNS AND OTHER STATISTICS.*/
PROC FACTOR DATA=A NFACT=&NV NOPRINT
    OUTSTAT=PATTERN_(WHERE=(_TYPE_ IN('MEAN','STD','N','PATTERN')));
RUN;

DATA _PATTERN_;
    SET COEFF PATTERN_;
RUN;

/*GENERATE THE RANDOM NUMBERS.*/
%LET NV2=%EVAL(&NV.*&NV.);
%LET NV3=%EVAL(3*&NV.);

DATA B&REPEAT. (KEEP=&VNAMES);
    SET _PATTERN_ (KEEP=&VNAMES _TYPE_ RENAME=(
    %DO I=1%TO&NV;
        &&V&I = V&I
    %END;
    )) END=LASTFACT;
    RETAIN;
/*SET UP ARRAYS TO STORE THE NESSESARRY STATISTICS.*/
    ARRAY VCOEFF(3,&NV)     C1-C&NV3;
    ARRAY FPATTERN(&NV,&NV) F1-F&NV2;
    ARRAY VSTD(&NV) S1-S&NV;
    ARRAY VMEAN(&NV) M1-M&NV;
    ARRAY V(&NV)V1-V&NV;
    ARRAY VTEMP(&NV)VT1-VT&NV;
    LENGTH LBL $40;
/* READ AND STORE THE MATORIX OF FACTOR PATTERNS.  */
    IF _TYPE_='PATTERN' THEN DO;
        DO I=1 TO &NV;
            FPATTERN(_N_ -6, I)=V(I);
        END;
    END;

    IF _TYPE_='COEFF' THEN DO;
        DO I=1 TO &NV;
        VCOEFF(_N_,I) =V(I);
        END;
    END;

/* READ AND STORE THE MEANS  */
```

```
    IF _TYPE_ = 'MEAN' THEN DO;
        DO I=1 TO &NV;
        VMEAN(I)=V(I);                                     END;
    END;

/* READ AND STORE THE STD.      */
    IF _TYPE_ = 'STD' THEN DO;
        DO I=1 TO &NV;
            VSTD(I) =V(I);                                 END;
    END;

/* READ AND STORE THE NUMBER OF OBSERVATIONS.*/
    IF _TYPE_ = 'N' THEN NNUMBERS=V(1);
    IF LASTFACT THEN DO;
/* SET UP LABELS FOR THE RANDOM VARIABLES. THE LABELSARE STORED IN MACRO
VARIABLES LBL1, LBL2,.... AND USED IN THE SUBSEQUENT PROC DATASETS.*/
        %DO I=1%TO&NV;
LBL="ST.NORMAL                VAR.                ,M-"||COMPRESS(PUT(VMEAN(&I),
BEST8.))||",STD="||COMPRESS(PUT(VSTD(&I),BEST8.));
            CALL SYMPUT("LBL&I",LBL);
        %END;
    DO K=1 TO NNUMBERS;
        DO I =1 TO &NV;
        SEED=(&SEED.+&REPEAT.+1);
        VTEMP(I)=RANNOR(SEED);
        END;
/* IMPOSE THE INTERCORRELATION ON EACH VARIABLE. THE
TRANSFORMED VARIABLES ARE STORED ARRAY 'V'.*/
        DO I=1 TO &NV;
            V(I)=0;
            DO J=1 TO &NV;
            V(I) = V(I) + VTEMP(J)*FPATTERN(J, I);
            END;
        END;

/* TRANSFORM THE RANDOM VARIABLES SO THEY HAVE
MEANS AND STANDARD DEVIATIONS AS REQUESTED. */

    DO I=1 TO &NV;
V(I)= VCOEFF(2, I)*(-1)+V(I)*VCOEFF(1, I) +VCOEFF(2,I)*V(I)*V(I)+VCOEFF(3, I)*V(I)*V(I)*V(I);
        V(I) = VSTD(I) *V(I) + VMEAN(I);
    END;
OUTPUT;
END;
END;

RENAME
    %DO I=1%TO&NV;
        V&I = &&V&I
    %END;
    ;
RUN;
```

```
DATA BB&REPEAT.;
    SET B&REPEAT.;
    ID=_N_;
    FORMAT ID;
RUN;
%MEND CONTINUOUS;

/***************** HISTGRAM and RATIO *************************/
/* This program generate random variables as in Figure 1 based on a Uniform /*distribution. % macro RATIO
is executed in % macro HISTGRAM.
/*%Macro HISTGRAM
/*Parameters
/*SEED= seed of the random number
/*N_REPEAT= the number of iterations
/*NSP= total sample size, which is already defined as an input parameter of the DATA   /*file used to ex-
ecute %macro CONTINUOUS.
/*MAX=Maximum value of an original variable
/*MIN=Minimum value of an original variable
/*SUB_GROUP= the number of subgroups
/*OUT
/*The name of the output file (name of HH&REPEAT) containing the random variable, /*H1 and ID number.
/*
/*%RATIO assigns the frequencies of each subgroup, using an IF function.
/*&NSP.*0.55 = (sample size × accumulated percentage)=frequencies of subgroup.
/*Example IF 1=< ID <&NSP.*0.55 THEN _H1=U1;
/*U1 indicates random number for the lowest subgroup.
/****************************************************************/
%MACROHISTGRAM;
DATA HH&REPEAT.;
    DO ID=1 TO &NSP.;
    CALL STREAMINIT(&REPEAT. +    &NSP. + &SEED. + 2.);
           SCALE=%EVAL((&MAX.-&MIN.)/&SUB_GROUP.);
                  %DO I=1%TO&SUB_GROUP.;
                         _U&I.=RAND("UNIFORM");
                           U&I.=&MIN.+(&I.-1)*SCALE+SCALE*_U&I.;
                  %END;
%RATIO;
OUTPUT;
END;
RUN;
%MEND HISTGRAM;

%MACRORATIO;
IF 1=<ID<&NSP.*0.55 THEN _H1=U1;
ELSE IF &NSP.*0.55 =<ID<&NSP.*0.6 THEN _H1=U2;
ELSE IF &NSP.*0.6 =<ID<&NSP.*0.8 THEN _H1=U3;
ELSE IF &NSP.*0.8=<ID<&NSP.*0.95 THEN _H1=U4;
ELSE _H1=U5;

H1=INT(_H1);

IF H1=<4 THEN C1=0;
ELSE IF 4< H1 =<15 THEN C1=1;
ELSE IF H1 >15 THEN C1=2;
%MEND RATIO;
/*************************** PDF ************************************/
/*This macro generates random variables from the RAND function.
/*RAND function generates random numbers with certain probability distributions.
/*Parameter
/*SEED= seed of the random number
/*N_REPEAT= the number of iterations
/*NSP= total sample size, which is already defined as an input parameter of the DATA file used to execute %macro CONTINUOUS.
/*RAND function.
/*OUT
/*The name of the output file (name of CC&REPEAT) containing the random variable /*defined by the probability distributions given by the RAND function and ID number.
/*****************************************************************/
%MACROPDF;
DATA CC&REPEAT.;
    DO ID=1 TO &NSP.;
           CALL STREAMINIT( &REPEAT. + &NSP. + &SEED. + 3);
           STRATA=&REPEAT.;
           UN=RAND("UNIFORM");
/*INSERT    RAND    FUNCTION TO GENERATE RANDOM NUMBER USING RAND FUNCTION*/
           X=RAND("NORMAL", 0,1);
           _D1=RAND("TABLE", 0.5 , 0.5);

           IF _D1=1 THEN D1=0;
           ELSE IF _D1=2 THEN D1=1;
    OUTPUT;
    END;
%MEND PDF;
/*********************** Merge ****************************/
/*This program merges all datasets including randomly generated variables specified /*in %macro CONTINUOUS (BB&REPEAT), %HISTGRAM (HH&REPEAT) /*and %macro PDF(CC&REPEAT) by ID number.
/*OUT file name of _D&REPEAT.
/*****************************************************************/
%MACRO  MERGE;
PROC SORT DATA=BB&REPEAT.; BY ID; RUN;
PROC SORT DATA=HH&REPEAT.; BY ID; RUN;
PROC SORT DATA=CC&REPEAT.; BY ID; RUN;

DATA _D&REPEAT.;
    MERGE BB&REPEAT. CC&REPEAT. HH&REPEAT.;
    BY ID;
RUN;

DATA DATASET0;
    SET DATASET0;

DATA DATASET&REPEAT.;
    SET DATASET%EVAL(&REPEAT. -1) _D&REPEAT.;
RUN;
%MEND MERGE;
```

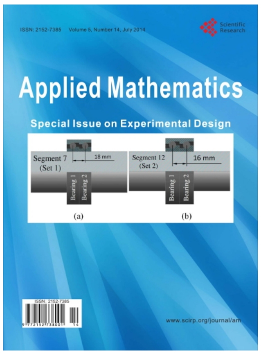
```
/**************************** OUTCOME ***************/
/*This program generates outcome variable, y, from the individual probability of event /*occurrence. Individual
probability is calculated using two segment logistic regression /*model.
/*Parameter
/*CHANGE_POINT= flexion point of two segment logistic regression model
/*MODEL_1 = logistic regression model when values of covariates ≦values of /*CHANGE_POINT
/*MODEL_2 = logistic regression model when values of covariates > values of /*CHANGE_POINT
/*NITER= number of final datasets
/*P = event proportion
/*OUT DATASET For logistic regression model
/************************************************************************/
%MACROOUTCOME;
DATA _D_&NITER.;
    SET    DATASET&NITER.;
    %IF H1 =<&CHANGE_POINT.%THEN%DO;
           G=&MODEL_1;
    %END;
    %IF H1 >&CHANGE_POINT.%THEN%DO;
           G=&MODEL_2;
    %END;
RUN;

PROC SUMMARY DATA=_D_&NITER.;
    VAR G;
    OUTPUT OUT= PROCMEAN&NITER. MEAN=;
RUN;

DATA M&NITER. (KEEP=INT ID NITER);
    SET PROCMEAN&NITER.;
    DO ID=1 TO &NSP;
           MEAN=%SCAN(G, 1);
           INT=LOG(&P/(1-&P))-MEAN;
           NITER=&NITER.;
           OUTPUT;
    END;
RUN;

PROC SORT DATA=M&NITER.;BY ID;RUN;
PROC SORT DATA=_D_&NITER.;BY ID;RUN;

DATA D_&NITER.;
    MERGE M&NITER. _D_&NITER.;
    BY ID;
RUN;

DATA D&NITER. ;
    SET D_&NITER.;
    PRO=EXP(INT+ G)/(1 +EXP(INT+ G));
    IF 0=<UN< PRO THEN Y=1;
ELSE Y=0;
RUN;

PROC SORT DATA=D&NITER. ;BY STRATA;RUN;
```

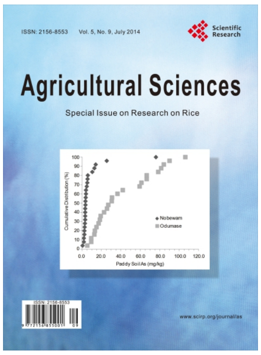
```
%MEND OUTCOME;
/**********************************************************************/
/*This program performs a stratified continuous logistic regression model and produces a repeated number of parameters (coefficient, its standard error, and p value), then calculates the average coefficient value, average standard error, and power.
/*Parameter
/*NITER=number of final datasets
/*CONTI_MODEL=a continuous logistic regression model
/*ALEVEL=significance level of the statistical test (Type I error)
/*NITER=specify final dataset
/*PATH=directory in which results are saved
/*TABLE=table name for saved results
/*OUT=Result (excel format)
/*Results include event proportion, mean, standard deviation, skewness, and kurtosis of a variable average coefficient and average standard error of logistic regression model and power
/**********************************************************************/
%MACROCONTINUOUSLR;
/****MODEL***********************************************************/
ODS OUTPUT PARAMETERESTIMATES=PARAM CONVERGENCESTATUS=STATUS;
    PROC LOGISTIC DATA=D&NITER. ;
           MODEL Y (EVENT='1')=&CONTI_MODEL
/TECH=NR MAXITER=8 XCONV=0.01 ;
           BY STRATA;
    RUN;
/**********************************************************************/
    PROC SORT DATA=PARAM
           OUT=PARAM2
(RENAME=(ESTIMATE=ESTIMATION STDERR=STANDARDERRORS));
           BY STRATA;
    RUN;

    PROC SORT DATA=STATUS;
           BY STRATA;
    RUN;

    DATA RESULT;
           MERGE PARAM2 STATUS ;
           BY STRATA;
    RUN;

DATA RESULT_CONTINUOUS E;
    SET RESULT;
    IF 0=< PROBCHISQ<&ALEVEL. THEN POWER=1;
    ELSE POWER=0;
    IF STATUS=0 THEN OUTPUT RESULT_CONTINUOUS;
    ELSE OUTPUT E;
RUN;

ODS HTML PATH="&PATH" BODY="&TABLE..XLS";
    PROC TABULATE DATA=D&NITER.    OUT=J;
           VAR Y &CONTI_MODEL;
TABLE (&CONTI_MODEL)*(MEAN STD SKEWNESS KURTOSIS)/MISSTEXT = 'NO DATA';
    RUN;
    PROC FREQ DATA=D&NITER;
        TITLE 'PROPORTION';
        TABLE Y/NOCOL NOROW;
    RUN;

    PROC FREQ DATA=RESULT_CONTINUOUS;
        TITLE 'POWER';
        TABLE VARIABLE*(POWER)/NOCOL NOPERCENT;
    RUN;

    PROC TABULATE DATA=RESULT_CONTINUOUS;
    TITLE 'MEAN OF COEFFICIENT AND THEIR MEAN OF STANDARD ERROR';
        CLASS VARIABLE ;
        VAR ESTIMATION STANDARDERRORS;
TABLE VARIABLE,(ESTIMATION STANDARDERRORS)*(N MEAN*F=8.4)/MISSTEXT = 'NO DATA';
    RUN;
ODS HTML CLOSE;
%MEND CONTINUOUSLR;
/************************* CONTINUOUSLR ***************************/
/*Continuous variables are divided into categorical groups by quantile, and then a stratified logistic regression model is executed. Users specify the model in %MACRO CATEGORICAL_MODEL. Then, parameters (coefficient, standard error, and p value) and average coefficient values, average standard error of each group of a variable are calculated, and the power is calculated.
/*Parameter
/*R=Number of categorized groups
/*Example: Continuous=1, median=2, tertile=3, quantile=4,
/*CATEGORIZATION=%NRSTR(List of covariates to be categorized)
/*CATEGORIZATION_R=%NRSTR(List of new covariate names after /*categorization)
/*NITER=number of final datasets
/*CONTI_MODEL=a continuous logistic regression model
/*ALEVEL=significance level of the statistical test (Type I error)
/*NITER= specify final dataset
/*PATH=directory in which results are saved
/*TABLE=table name for saved results
/*OUT= Result (excel format)
/*Results include average coefficient and average standard error of logistic regression model and power for each categorized group and overall power of a variable.
/******************************************************************/

%MACROCATEGORICAL_MODEL;
ODS OUTPUT PARAMETERESTIMATES=PARAM_&R CONVERGENCESTATUS=STATUS_&R TYPE3=TYPE3_&R.;
    PROC LOGISTIC DATA=G&R.;
    CLASS C1(PARAM=REF REF="0") X1_R(PARAM=REF REF="0")    ;
        MODEL Y(EVENT='1')= C1 X1_R X2 /TECH=NR MAXITER=8 XCONV=0.01;
        BY STRATA;
        RUN;
%MEND CATEGORICAL_MODEL;

%MACROCATEGORICALLR;

PROC RANK DATA= D&NITER. GROUPS=&R. OUT=G&R.;
    VAR &CATEGORIZATION;
    RANKS &CATEGORIZATION_R ;
    BY STRATA ;
RUN;

%CATEGORICAL_MODEL;

PROC SORT DATA=PARAM_&R
OUT=P_&R (RENAME=( ESTIMATE=ESTIMATION STDERR=STANDARDERRORS));
    BY STRATA;
RUN;

PROC SORT DATA=STATUS_&R OUT =S_&R(KEEP = STRATA STATUS) ;
    BY STRATA;
RUN;

PROC SORT DATA=TYPE3_&R OUT =_TYPE3_&R;
    BY STRATA;
RUN;

DATA _POWER_&R;
    MERGE _TYPE3_&R S_&R;
    BY STRATA;
RUN;

DATA POWER_&R E_&R;
    SET _POWER_&R;
    RENAME;
TYPE3_WALDCHISQ=WALDCHISQ;
           TYPE3_PROBCHISQ=PROBCHISQ;
           LABEL TYPE3_PROBCHISQ="P VALUE OF TYPE3";
           TYPE3_WALDCHISQ="CHISQ OF TYPE 3";
    IF 0 =< TYPE3_PROBCHISQ <&ALEVEL. THEN POWER=1;
    ELSE IF TYPE3_PROBCHISQ >= &ALEVEL. THEN POWER=0;

    IF STATUS=0 THEN OUTPUT POWER_&R;
    ELSE OUTPUT E_&R;

    KEEP POWER STATUS TYPE3_WALDCHISQ TYPE3_PROBCHISQ EFFECT;
RUN;

DATA _RESULT_CATEGORICAL_&R;
    MERGE P_&R S_&R;
    BY STRATA;
RUN;

DATA RESULT_CATEGORICAL_&R E_&R;
    SET _RESULT_CATEGORICAL_&R;

    IF CLASSVAL0=. THEN CLASSLEVEL=0;
    ELSE CLASSLEVEL=CLASSVAL0;

    IF 0=< PROBCHISQ<&ALEVEL. THEN GROUP_POWER=1;
    ELSE IF PROBCHISQ>= &ALEVEL. THEN GROUP_POWER=0;
    DROP CLASSVAL0;
    IF STATUS=0 THEN OUTPUT RESULT_CATEGORICAL_&R;
    ELSE OUTPUT E_&R;
RUN;

ODS HTML PATH="&PATH" BODY="GROUP_&R._&TABLE..XLS";
    PROC FREQ DATA=POWER_&R;
        TITLE 'POWER OF B';
        TABLE EFFECT*(POWER)/NOCOL NOPERCENT;
    RUN;

    PROC TABULATE DATA= RESULT_CATEGORICAL_&R;
        CLASS VARIABLE CLASSLEVEL;
        VAR ESTIMATION STANDARDERRORS;
TABLE VARIABLE*CLASSLEVEL,(ESTIMATION STANDARDERRORS ) *(N MEAN*F=8.4) /RTS=20
MISSTEXT = 'NO DATA';
    RUN;

    PROC TABULATE DATA= RESULT_CATEGORICAL_&R;
        CLASS VARIABLE CLASSLEVEL GROUP_POWER;
TABLE VARIABLE*CLASSLEVEL,( GROUP_POWER )*(N ROWPCTN) /RTS=20  MISSTEXT = 'NO
DATA';
    RUN;
ODS HTML CLOSE;

%MEND CATEGORICALLR;
/*********************************************************************/
Dataset generation
One dataset is created from each iteration of %COEFF, %CONTINUOUS, %HISTGRAM, %PDF,
and %MERGE.
This dataset is accumulated until the iterations are complete and the iteration time is identified as strata.
/*********************************************************************/

%MACRO DATASET(N_REPEAT=);
    %COEFF;
    %DO REPEAT=1%TO&N_REPEAT.;
        %CONTINUOUS;
        %HISTGRAM
        %PDF;
        %MERGE;
    %END;
    %OUTCOME;
%MEND DATASET;
/*********************************************************************/
The parameter estimations and statistical power calculation.
/*********************************************************************/
%MACRO LR(R=);
%DO R=1%TO&R;
%IF&R=1%THEN%DO;
        %CONTINUOUSLR;
    %END;
    %ELSE%DO;
        %CATEGORICALLR;
        %END;
    %END;QUIT;
    %MEND LR;
```

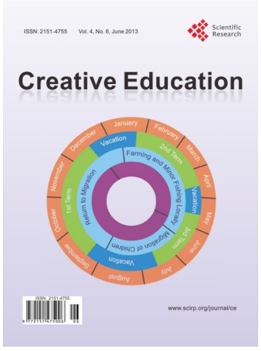
Advances in Bioscience and Biotechnology

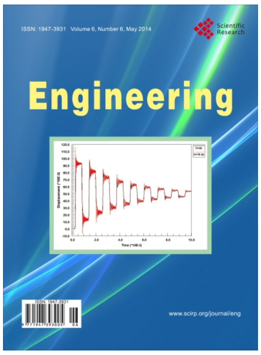
American Journal of Analytical Chemistry

American Journal of Plant Sciences
Special Issue on Photosynthesis

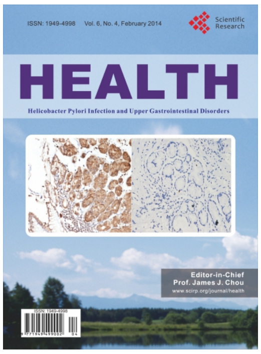

Applied Mathematics
Special Issue on Experimental Design


Agricultural Sciences
Special Issue on Research on Rice


Creative Education

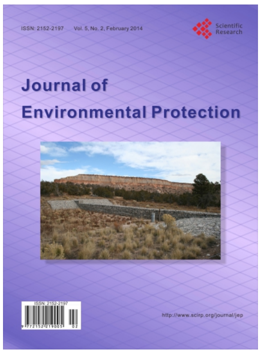

Engineering

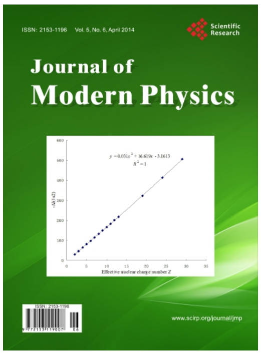
Food and Nutrition Sciences


HEALTH
Helicobacter Pylori Infection and Upper Gastrointestinal Disorders
Editor-in-Chief
Prof. James J. Chou

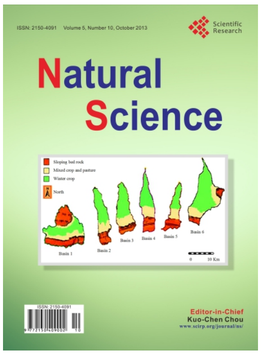
Journal of Computer and Communications

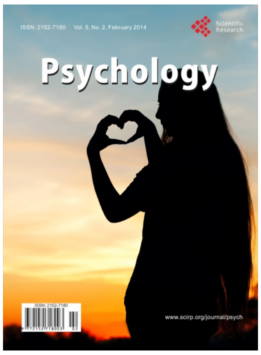
Journal of Cancer Therapy
Special Issue on Bone Tumor Research


Journal of Environmental Protection


Journal of Modern Physics

Modern Economy


Natural Science
Editor-in-Chief
Kuo-Chen Chou


Psychology