TECHNICAL REPORT # 59
MAY 2013

Interim sample size recalculation for linear and logistic regression models: a comprehensive Monte-Carlo study

Sergey Tarima, Peng He, Tao Wang, Aniko Szabo
Division of Biostatistics, Medical College of Wisconsin

Abstract

e propose a simple procedure for interim sample size recalculation when testing a hypothesis on a regression coefficient and explore its effects on type I and II errors e consider hypothesis testing in linear and logistic regression models using the ald test e performed a comprehensive Monte Carlo study comprised of experiments with , repetitions each In these experiments we varied the number of predictors or the type of predictors binary and continuous the magnitude of the tested regression coefficient the degree of association between predictors and the lower and upper limits on the total sample size

1 Introduction

The sample size (SS) calculation is complicated by the presence of nuisance parameters. The values of these parameters are often estimated from external or internal pilot data which could substantially decrease the influence of erroneous assumptions on the values of these nuisance parameters.

In this manuscript we explore a "naive" approach to interim sample size re-estimation in linear and logistic regression models. This approach recalculates nuisance parameters at the interim analysis and updates the sample size bounded by the size of the internal pilot, n, and an upper bound, $N_{\max}$, often chosen from budgetary or recruitment considerations. The benefit of sample size recalculation comes with a price, it inflates the type I error and power, and the final sample size becomes a random quantity.

We consider regression models of the form

$$E\left(Y|X_1, ..., X_p\right) = g^{-1}\left(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p\right), \tag{1}$$

where the mean of an outcome Y conditional on $X_1, ..., X_p$ is parameterized by a continuous monotone link function $g(\cdot)$ and a linear combination of regression coefficients $\beta_0, ..., \beta_p$ and $X_1, ..., X_p$. The logistic regression is defined by a LOGIT link function, $g(EY) = \ln\left(\frac{EY}{1-EY}\right.$, for a binary outcome. The gaussian linear regression is defin

The practical use of (6) is complicated by the asymptotic nature of the test and the unknown $\mathcal{I}^{-1}(\beta_1)$. We use internal pilot data to estimate $\mathcal{I}^{-1}(\beta_1)$ and recalculate the sample size. Regular regression model output contains a table of regression coefficients with estimates of their standard errors. From this table, the standard error of the internal pilot based estimate $\hat{\beta}_1$ is $SE(\hat{\beta}_1)$. Then, we use $n \quad SE(\hat{\beta}_1) \quad ^{2}$ to approximate $\mathcal{I}^{-1}(\beta_1)$ and calculate the total sample size. Then, the final formula for the total sample size is

$$N = n \quad SE(\hat{\beta}_1) \quad ^{2} \frac{\left(z_{1-\alpha/2} + z_{1-\beta} \quad ^{2}\right.}{\delta^2}. \tag{7}$$

Multiple linear regression

For a linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon$$

with non-random covariates $X_{i1}, ..., X_{ip}$ and $\epsilon \sim N(0, \sigma^2)$, the Fisher information matrix for a single i^{th} observation $(Y_i, X_{i1}, ..., X_{ip})$ is a symmetric $(p+1) \times (p+1)$ matrix

$$\mathcal{I}(\mathbf{b}|X_{i1}, ..., X_{ip}) = \frac{1}{\sigma^2} \begin{pmatrix} X_{i0}X_{i0} & X_{i1}X_{i0} & \cdots & X_{ip}X_{i0} \\ X_{i0}X_{i1} & X_{i1}X_{i1} & \cdots & X_{ip}X_{i1} \\ \cdots & \cdots & \cdots & \cdots \\ X_{i0}X_{ip} & X_{i1}X_{ip} & \cdots & X_{ip}X_{ip} \end{pmatrix},$$

where $X_{i0} = 1$. Denote $\mathbf{X}_{i\cdot} = (X_{i0}, ..., X_{ip})^T$, then $\mathcal{I}(\mathbf{b}|X_{i1}, ..., X_{ip}) = \sigma^{-2}\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^T$.

If $X_{i1}, ..., X_{ip}$ are random variables, we need to integrate their distribution out,

$$\mathcal{I}(\mathbf{b}) = E_{\mathbf{X}_{i1},...,\mathbf{X}_{ip}}\mathcal{I}(\mathbf{b}|X_{i1}, ..., X_{ip}) = \sigma^{-2}E\left(\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^T\right. . \tag{8}$$

The matrix of second moments $E\left(\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^T\right.$ becomes the variance-covariance matrix of covariates when all covariates are centered, $E(X_{ij}) = 0$ $(j = 1, ..., p)$, except for the constant term $X_{i0} = 1$. The distribution of covariates in observational studies is often unknown. A simple solution is to plug-in internal pilot data based estimates of σ^2 and $E\left(\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^T\right.$ in Equation (8), then

$$\hat{\mathcal{I}}(\mathbf{b}) = \hat{\sigma}^{-2}n^{-1}\mathbf{X}_n\mathbf{X}_n^T,$$

where

$$\mathbf{X}_n = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}.$$

For a sufficiently large n, $\mathcal{I}^{-1}(\mathbf{b}) \approx n\hat{\sigma}^2\left(\mathbf{X}_n^T\mathbf{X}_n \right.^{-1}$, and

$$\sqrt{n} \quad \hat{\beta}_1 - \beta_1 \quad \overset{d}{\approx} N \quad 0, n\hat{\sigma}^2\left(\mathbf{X}_n^T\mathbf{X}_n\right._{(11)}^{-1} \quad ,$$

where $\left(\mathbf{X}_n^T\mathbf{X}_n\right._{(11)}^{-1}$ denotes the second diagonal element of $\left(\mathbf{X}_n^T\mathbf{X}_n\right.^{-1}$. For centered covariates the use of some matrix algebra leads to

$$n\,(\mathbf{X}_\mathbf{n}^\mathbf{T}\mathbf{X}_\mathbf{n}\;{}^{-1}_{(11)} = \hat{\sigma}^{-2}_{\mathbf{X}_1}\;\;\mathbf{1} - \hat{r}^2_{\mathbf{X}_1|\mathbf{X}_2,\ldots,\mathbf{X}_p}\;\;{}^{-1},$$

where $\hat{\sigma}^2_{\mathbf{X}_1}$ is the sample variance of X_1 and $\hat{r}^2_{\mathbf{X}_1|\mathbf{X}_2,\ldots,\mathbf{X}_p}$

ple optimality, however, known Fisher information makes these formulas data independent. Thus, these “gold standard” formulas can be considered as the target quantities in sample size re-estimation.

3 Simulation studies

Each simulation experiment was based on 60000 repetitions, 50000 under β_1 = 0, and 10000 under the alternative β_1 = δ. This secured the Monte-Carlo standard error of 0$.$001 for estimating the type I error and 0$.$

squared errors is not directly applicable to logistic regression models. However, a generalized R^2 can be used instead. We found that the generalized $R^2 \in \{0.02, 0.05, 0.1, 0.2\}$ for logistic regression model with a single continuous predictor is approximately reached at $\beta_1 \in \{0.291, 0.469, 0.702, 1.127\}$ at n = 20. The β_1 at the binary predictor leads to $\beta_1 \in \{0.286, 0.459, 0.667, 1.003\}$ at n = 20. As the sample size increases, the generalized R^2

Table 1: Table numbers classified by model, the number and type of predictors, Pearson correlation (r), B_i = $I(X_i > 0)$

Predictors	r	Linear	Linear	Logistic
		N_{max} = 300	N_{max} = 600	N_{max} = 600
X_1	0	2	28	54
B_1	0	3	29	55
X_1, X_2	0	4	30	56
X_1, X_2	0.4	5	31	57
X_1, X_2	0.8	6	32	58
B_1, B_2	0	7	33	59
B_1, B_2	0.4	8	34	60
B_1, B_2	0.8	9	35	61
B_1, X_2	0	10	36	62
B_1, X_2	0.4	11	37	63
B_1, X_2	0.8	12	38	64
X_1, B_2	0	13	39	65
X_1, B_2	0.4	14	40	66
X_1, B_2	0.8	15	41	67
$X_1, X_2, ..., X_{10}$	0	16	42	68
$X_1, X_2, ..., X_{10}$	0.4	17	43	69
$X_1, X_2, ..., X_{10}$	0.8	18	44	70
$B_1, B_2, ..., B_{10}$	0	19	45	71
$B_1, B_2, ..., B_{10}$	0.4	20	46	72
$B_1, B_2, ..., B_{10}$	0.8	21	47	73
$B_1, X_2, ..., X_{10}$	0	22	48	74
$B_1, X_2, ..., X_{10}$	0.4	23	49	75
$B_1, X_2, ..., X_{10}$	–	24 (r = 0.8)	50 (r = 0.8)	76 (r = 0.72)
$X_1, B_2, ..., B_{10}$	0	25	51	77
$X_1, B_2, ..., B_{10}$	0.4	26	52	78
$X_1, B_2, ..., B_{10}$	–	27 (r = 0.8)	53 (r = 0.8)	79 (r = 0.72)

- The sample size formula is applicable but the calculated sample size does not belong to the sample size range.
  - Solution: If $N < n_1$, then N = n_1. If $N > N_{max}$, then N = N_{max}.

Other settings

We explored three possibilities for the internal pilot sample size, $n \in \{20, 50, 100\}$. All
2Tf 1001 24749 .760Td [(;)0.972873]TJ /R1111.9552-1.8Td [(m)6948(e)38.3333300cmBT /R2311.9553616

4 Empirical Conclusions

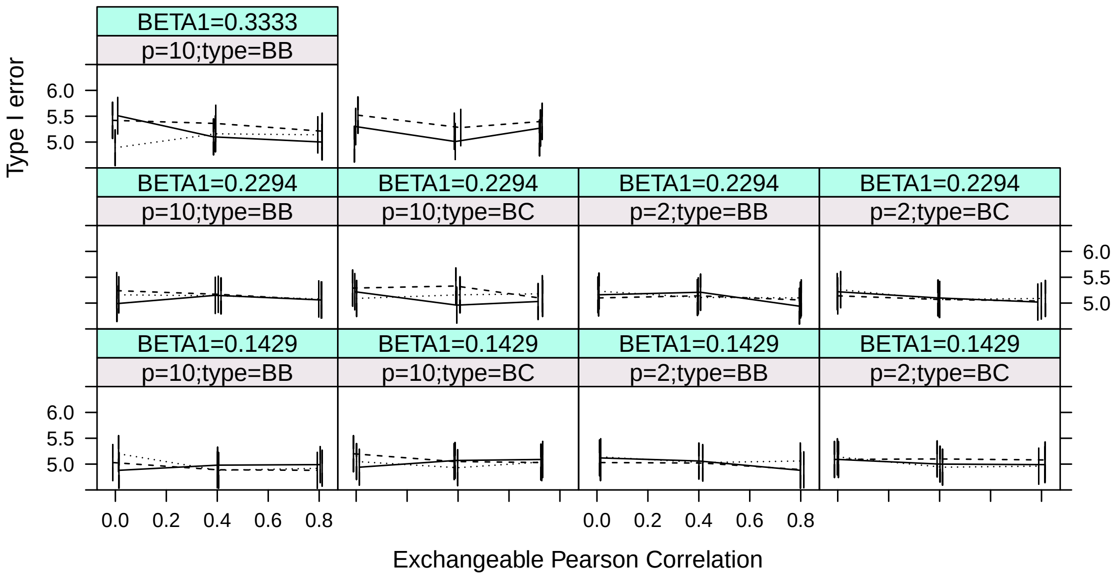
The Monte-Carlo Error for type I error assessments is 0.

is 0.0662, Table 24, at highly correlated predictors, $B_1, X_2, ..., X_{10}$. The binary predictor of interest, B_1, was also associated with an increased type I error when compared to a continuous predictor X_1. As the size of internal pilot increases the inflation of type I error becomes less and less visible.

For illustrative convenience we also reported the type I error and power in eight figures, Figures 1 - 4 (linear regressions) and 7 - 10 (logistic regressions).

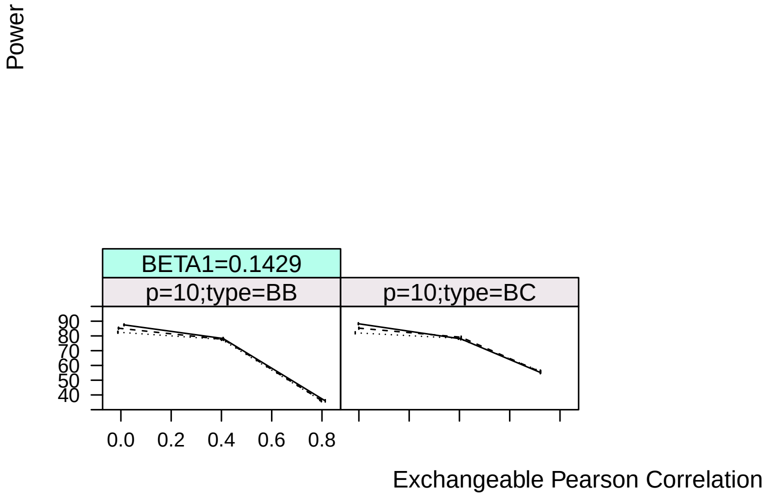
The type I errors for linear regression models with multiple predictors are reported in Figures 1 (X_1 is binary) and 2 (X_1 is continuous).

Linear regression with multiple predictors, X1 is binary



Linear regression with multiple predictors, X1 is continuous

Linear regression with multiple predictors, X1 is binary



Linear regression with multiple predictors, X1 is continuous

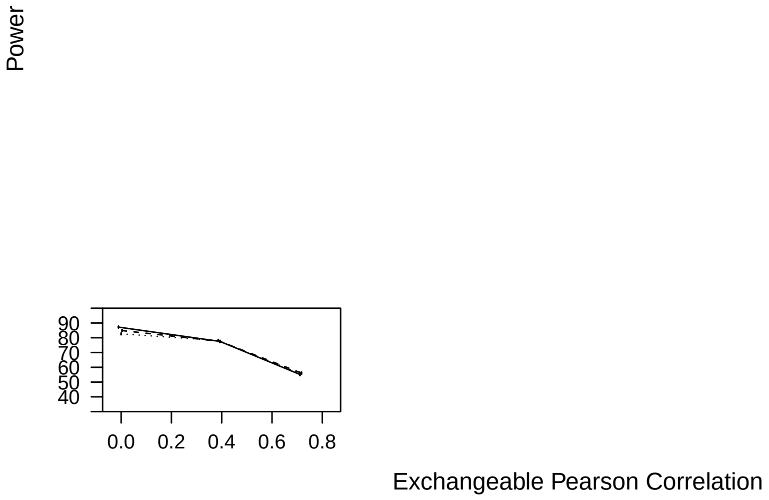


Figure 5: Logistic regression, two binary predictors, $n_1 = 100$, $= 0{:}37$, $r = 0$, type I error = 4.9%, power=81.8%. Left panel presented the final sample size distribution under the null, the right panel is under the alternative hypothesis.

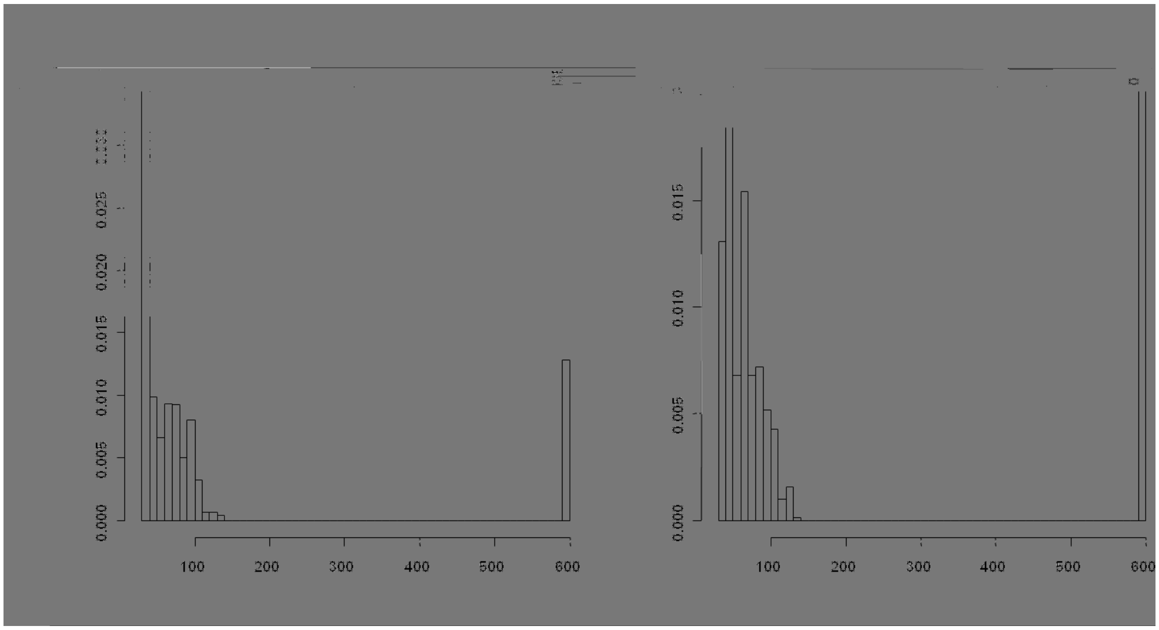
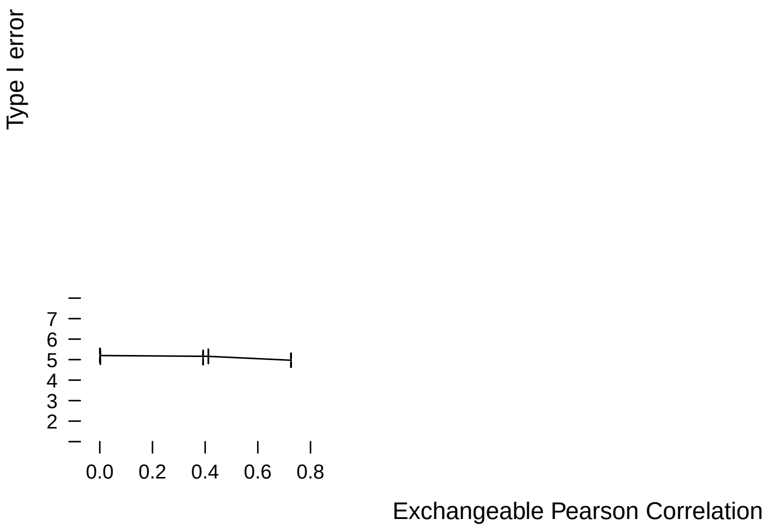


Figure 6: Logistic regression, two binary predictors, $n_1 = 20$, $= 1$, $r = 0{:}8$, type I error = 1.3%, power=54.0%. Left panel presented the final sample size distribution under the null, the right panel is under the alternative hypothesis.

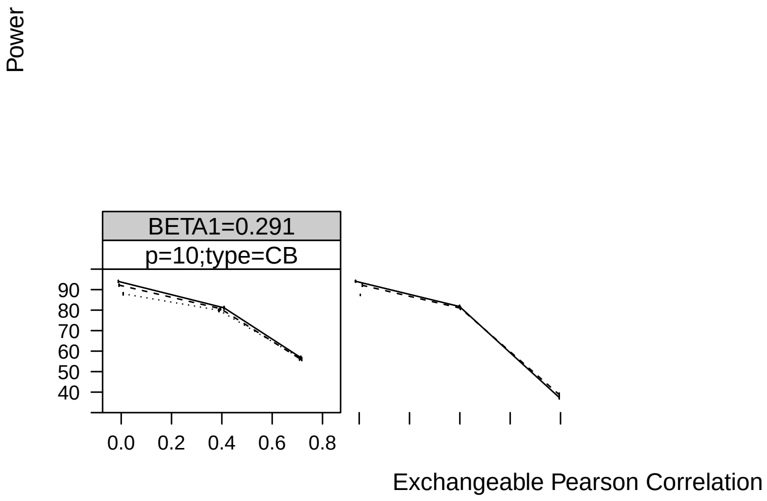
Logistic regression with multiple predictors, X1 is continuous



Logistic regression with multiple predictors, X1 is binary

Exchangeable Pearson Correlation

Logistic regression with multiple predictors, X1 is continuous



Naive internal pilot designs are commonly used in practice. These designs update the total sample size using new values of nuisance parameters at the interim analysis. These methods, however, use unadjusted sample size formulas and test statistics. This naive approach to sample size recalculation is known to inflate the type I and II errors. Meanwhile, a few statistical methods suggest various remedies for a better control of the size of the test and its power properties. We call these adjusted internal pilot designs as non-naive internal pilots designs.

Despite the existence of substantial amount of statistical literature on non-naive internal pilot designs, we did not see enough evidence to justify the use of non-naive internal pilot designs.

In this work, we did not make any modifications to the naive internal pilot designs. We investigated how much the type I and II errors are inflated under various scenarios with the use of the ordinary linear and logistic regression models as the primary-2.26432(i)alysic re tey Tdn

[illegible]

Overall, naive internal pilot designs are useful and legitimate way for sample size recalculation provided that the aforementioned pitfalls are avoided.

Appendix

The generalized R_G^2 (see [8]) is defined by

$$R_G^2 = 1 - \left(\frac{L(0)}{L(\widehat{\beta})}\right)^{2/n}; \qquad (11)$$

where $L(0)$ is the likelihood of the intercept-only model, $L(\hat{\beta})$ is the likelihood based on the estimated model parameters. We consider a simple single predictor case when $\beta = (\beta_0, \beta_1)$ and substitute $L(\hat{\beta})$ with $L(\beta)$ in Equation 11. Then, we consider

$$R_X^2 = 1 - \left(\frac{L(0)}{L(\beta)}\right)^{2/n}; \qquad (12)$$

where the subscript X highlights the dependence on the design matrix. A bit of algebra leads to

$$\ln(1 - R_X^2) = \frac{2}{n}\left\{\ln L(0) - \ln L(\widehat{\beta})\right\} = -\frac{1}{n}\left(2\ln\frac{L(\hat{\beta})}{L(0)}\right):$$

For an i.i.d. sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ the log-likelihood conditional on X_i is

$$\ln L(\beta) = \sum_{i=1}^{n} \ln f(Y_i|\beta, X_i)$$

and

$$\begin{aligned}\ln(1 - R_X^2) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{2\ln\frac{f(Y_i|\hat{\beta}, X_i)}{f(Y_i|0)}\right\} \qquad (13)\\ &= -2\int \ln\frac{f(y|\hat{\beta}, x)}{f(y|0, x)}\,dP_n(x, y);\end{aligned}$$

where $P_n(x, y)$ is the empirical measure with n^{-1} weights on (Y_i, X_i). Under so2 Tf 160.92 9 11.9552 Tf 6

Table 2: Linear regression with a single continuous predictor, N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0505	0.6600	281.50		

Table 4: Linear regression with two independent continuous predictors, N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0521	0.6674	283.91	37.52	384.58
	0.2294	0.05	0.0512	0.8022	176.83	71.56	149.03
	0.3333	0.1	0.0535	0.8094	88.57	45.11	70.69
	0.5	0.2	0.0579	0.8283	40.20	20.00	31.34
	1	0.5	0.0496	0.9596	20.26	1.70	7.86
50	0.1429	0.02	0.0497	0.6799	293.41	19.68	384.58
	0.2294	0.05	0.0516	0.8175	162.30	47.08	149.03
	0.3333	0.1	0.0528	0.8116	77.75	22.19	70.69
	0.5	0.2	0.0506	0.9157	50.56	2.84	31.34
	1	0.5	0.0506	0.9999	50	0	7.86
100	0.1429	0.02	0.0494	0.6810	297.65	9.77	384.58
	0.2294	0.05	0.0498	0.8072	155.91	31.67	149.03
	0.3333	0.1	0.0487	0.9019	100.53	3.07	70.69
	0.5	0.2	0.0491	0.9966	100	0	31.34
	1	0.5	0.0508	1	100	0	7.86

Table 5: Linear regression with two continuous predictors, Pearson correlation between the predictors = 0:4, N_{max} = 300.

n	$_1$	R^2	Type I error	Power

Table 10: Linear regression with two independent predictors (binary and continuous), the predictor of interest is binary, N

Table 12: Linear regression with two predictors (binary andcontinuous), the predictor of interest is binary, Pearson correlation between the predictors = $0{:}8$ (original correlation = $0{:}9991$), $N_{max} = 300$.

Table 14: Linear regression with two predictors (continuous and binary), the predictor of interest is continuous, Pearson correlation between the predictors = 0:4 (original correlation = 0:5037), N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0513	0.6030	291.16	27.89	458.75
	0.2294	0.05	0.0518	0.7942	201.45	72.39	178.24
	0.3333	0.1	0.0537	0.8163	105.01	52.59	84.22
	0.5	0.2	0.0576	0.8183	47.60	24.30	37.45
	1	0.5	0.0506	0.9386	20.57	2.65	9.37
50	0.1429	0.02	0.0512	0.6204	298.10	10.35	458.75
	0.2294	0.05	0.0512	0.8011	191.60	52.09	178.24
	0.3333	0.1	0.0541	0.8115	92.03	27.08	84.22
	0.5	0.2	0.0506	0.8862	52.02	5.82	37.45
	1	0.5	0.0497	0.9997	50	0	9.37
100	0.1429	0.02	0.0481	0.6128	299.72	3.18	458.75
	0.2294	0.05	0.0514	0.8100	185.74	37.36	178.24
	0.3333	0.1	0.0516	0.8646	103.12	8.09	84.22
	0.5	0.2	0.0502	0.9919	100	0	37.45
	1	0.5	0.0507	1	100	0	9.37

Table 15: Linear regression with two predictors (continuous and binary), the predictor of interest is continuous, Pearson correlation between the predictors = 0:8 (original correlation = 0:9991), N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0510	0.3162	299.77	4.37	1052.62
	0.2294	0.05	0.0507	0.6410	285.53	36.52	408.51
	0.3333	0.1	0.0508	0.8045	215.12	73.92	193.89
	0.5	0.2	0.0539	0.8127	111.57	59.11	86.19
	1	0.5	0.0571	0.8481	30.27	14.78	21.53
50	0.1429	0.02	0.0491	0.3146	300.00	0.02	1052.62
	0.2294	0.05	0.0488	0.6516	294.59	18.21	408.51
	0.3333	0.1	0.0523	0.8109	208.59	56.84	193.89
	0.5	0.2	0.0534	0.8180	95.52	31.07	86.19
	1	0.5	0.0505	0.9732	50.04	0.71	21.53
100	0.1429	0.02	0.0495	0.3169	300	0	1052.62
	0.2294	0.05	0.0503	0.6610	298.17	8.90	408.51
	0.3333	0.1	0.0527	0.7971	203.27	43.44	193.89
	0.5	0.2	0.0505	0.8525	104.68	10.62	86.19
	1	0.5	0.0498	0.9999	100	0	21.53

Table 16: Linear regression with 10 independent continuous predictors, N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0503	0.6607	293.49	26.76	384.49
	0.2294						

Table 18: Linear regression with 10 continuous predictors, Pearson correlation among the predictors = 0:8, N_{max} = 300.

n	$_1$	R^2

Table 22: Linear regression with 10 independent predictors, the predictor of interest is

Table 24: Linear regression with 10 predictors, the predictor of interest is binary, other 9 predictors are continuous, pairwise Pearson correlation = 0:72 (original pairwise correlation = 0:9), N_{max} = 300.

n	1	Rrson correlRrson cr26432(r)-0.6T2.2648(s12r)-39242 Q 3(I3924(r)(r)-391.682(a

Table 26: Linear regression with 10 predictors, the predictor of interest is continuous, other 9 predictors are binary, pairwise Pearson correlation = 0:4 (original pairwise correlation = 0:5037), N_{max} = 300.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0515	0.4985	298.19	14.05	587.73
	0.2294	0.05	0.0506	0.8325	274.60	52.62	227.98
	0.3333	0.1	0.0522	0.9054	209.27	83.73	108.06
	0.5	0.2	0.0551	0.8929	119.29	77.78	[redacted]
	1	0.5	0.0601	0.8769	41.13	51.65	11.99
50	0.1429	0.02	0.0506	0.5020	299.93	1.80	587.73
	0.2294	0.05	0.0508	0.8236	260.84	47.26	227.98
	0.3333	0.1	0.0515	0.8345	141.34	45.65	108.06
	0.5	0.2	0.0529	0.8371	65.78	18.34	48.09
	1	0.5	0.0505	0.9955	50.009(9)-2.26309(4)-2.26309]TJ ET Q q 4 0 3-2.		

Table 27: Linear regression with 10 predictors, the predictor of interest is continuous, other 9 predictors are binary, pairwise Pearson correlation = 0.72 (original pairwise correlation = 0.9), N_{max} = 300.

Table 29: Linear regression with a single binary predictor, $N_{max} = 600$.

Table 30: Linear regression with two independent continuous predictors, $N_{max} = 600$.

Table 34: Linear regression with two binary predictors, Pearson correlation between the predictors = 0:4 (tetrachoric correlation = 0:5878), N_{max} = 600.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0506	0.7843	468.85	126.30	457.61
	0.2294	0.05	0.0521	0.8026	210.80	97.72	177.69
	0.3333	0.1	0.0552	0.8011	101.12	53.08	84.18
	0.5	0.2	0.0561	0.8098	45.66	28.21	37.40
	1	0.5	0.0512	0.9490	20.89	15.82	9.35
50	0.1429	0.02	0.0502	0.7947	470.98	94.03	457.61
	0.2294	0.05	0.0514	0.8076	189.06	48.73	177.69
	0.3333	0.1	0.0520	0.8020	90.03	23.23	84.18
	0.5	0.2	0.0499	0.8763	51.22	4.62	37.40
	1	0.5	0.0492	0.9999	50	0	9.35
100	0.1429	0.02	0.0503	0.8040	467.85	73.70	457.61
	0.2294	0.05	0.0511	0.8013	182.93	31.53	177.69
	0.3333	0.1	0.0513	0.8559	101.96	5.92	84.18
	0.5	0.2	0.0502	0.9927	100	0	37.40
	1	0.5	0.0502	1	100	0	9.35

Table 35: Linear regression with two binary predictors, Pearson correlation between the predictors = 0:8 (tetrachoric correlation = 0:9511), N_{max} = 600.

n	$_1$

Table 36: Linear regression with two independent predictors(binary and continuous), the predictor of interest is binary, N_{max} = 600.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0509	0.8013	416.54	125.08	384.36
	0.2294	0.05	0.0522	0.8019	168.55	62.07	149.15
	0.3333						

Table 42: Linear regression with 10 independent continuous predictors, N_{max} = 600.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0484	0.8736	528.34	121.98	384.49
	0.2294	0.05	0.0514	0.9072	327.58	168.39	149.09
	0.3333	0.1	0.0527	0.8976	173.25	118.97	70.71
	0.5	0.2	0.0578	0.8867	79.36	60.32	31.39
	1	0.5	0.0562	0.8956	25.19	12.33	7.86
50	0.1429	0.02	0.0523	0.8466	472.97	109.11	384.49
	0.2294	0.05	0.0529	0.8510	197.32	65.35	149.09
	0.3333	0.1	0.0520	0.8288	93.65	30.40	70.71
	0.5	0.2	0.0515	0.8882	52.59	6.93	31.39
	1	0.5	0.0513	0.9997	50	0	7.86
100	0.1429	0.02	0.0507	0.8287	434.37	85.86	384.49
	0.2294	0.05	0.0540	0.8243	169.94	36.50	149.09
	0.3333	0.1	0.0502	0.8768	101.58	5.72	70.71
	0.5	0.2	0.0516	0.9956	100	0	31.39
	1	0.5	0.0517	1	100	0	7.86

Table 43: Linear regression with 10 continuous predictors, Pearson correlation among the predictors = 0:4, N_{max} = 600.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0495	0.7802	572.13	0.91	584.82
	0.2294	0.05	0.0529	0.9077	429.29	164.76	227.17
	0.3333	0.1	0.0516	0.9082	253.38		

Table 48: Linear regression with 10 independent predictors, the predictor of interest is binary, other 9 predictors are continuous, $N_{max} = 600$.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0494	0.8826	531.53	116.32	384.37
	0.2294	0.05	0.0522	0.9099	313.67	155.33	149.15
	0.3333	0.1	0.0530	0.9019	158.30	97.95	70.65
	0.5	0.2	0.0630	0.8823	71.56	46.96	31.40
	1	0.5	0.0563	0.8939	23.51	8.97	7.85
50	0.1429	0.02	0.0520	0.8526	471.90	96.14	384.37
	0.2294	0.05	0.0529	0.8415	189.08	47.65	149.15
	0.3333	0.1	0.0552	0.8281	89.78	22.39	70.65
	0.5	0.2	0.0485	0.8821	51.07	3.58	31.40
	1	0.5	0.0504	1	50	0	7.85
100	0.1429	0.02	0.0505	0.8212	428.06	66.46	384.37
	0.2294	0.05	0.0509	0.8112	166.60	26.37	149.15
	0.3333	0.1	0.0496	0.8830	100.37	2.11	70.65
	0.5	0.2	0.0509	0.9971	100	0	31.40
	1	0.5	0.0489	1	100	0	7.85

Table 49: Linear regression with 10 predictors, the predictor of interest is binary, other 9 predictors are continuous, pairwise Pearson correlation = 0.4 (original pairwise correlation = 0.5037), $N_{max} = 600$.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0507	0.7801	573.16	76.29	587.90
	0.2294	0.05	0.0496	0.9034	420.60	162.25	228.31
	0.3333	0.1	0.0501	0.8996	240.03	142.39	108.14
	0.5	0.2	0.0572	0.8932	112.35	78.99	48.04
	1	0.5	0.0606	0.8726	30.87	18.65	12.00
50	0.1429	0.02	0.0505	0.7913	575.95	53.68	587.90
	0.2294	0.05	0.0533	0.8498	292.03	85.03	228.31
	0.3333	0.1	0.0528	0.8467	138.21	40.74	108.14
	0.5	0.2	0.0544	0.8337	63.88	15.81	48.04
	1	0.5	0.0497	0.9977	50.00	0.03	12.00
100	0.1429	0.02	0.0493	0.7841	577.05	43.63	587.90
	0.2294	0.05	0.0516	0.8244	255.84	48.56	228.31
	0.3333	0.1	0.0521	0.8156	123.22	20.78	108.14
	0.5	0.2	0.0490	0.9642	100.00	0.22	48.04
	1	0.5	0.0508	1	100	0	12.00

Table 50: Linear regression with 10 predictors, the predictor of interest is binary, other 9 predictors are continuous, pairwise Pearson correlation = 0:72 (original pairwise correlation = 0:9), N_{max} = 600.

n	$_1$	R^2	Type I error	Power	E(N)	SD(N)	Target SS
20	0.1429	0.02	0.0509	0.5523	595.02	32.82	1030.71
	0.2294	0.05	0.0503	0.8642	534.99	115.03	400.25
	0.3333	0.1	0.0527	0.9072	379.71	164.81	189.36
	0.5	0.2	0.0551	0.9016	194.16	122.53	84.07
	1	0.5	0.0648	0.8718	50.51	34.79	21.07
50	0.1429	0.02	0.0503	0.5554	599.75	4.92	1030.71
	0.2294	0.05	0.0509	0.8446	485.32	100.09	400.25
	0.3333	0.1	0.0540	0.8500	243.47	70.03	189.36
	0.5	0.2	0.0529	0.8365	108.56	31.28	84.07
	1	0.5	0.0508	0.9585	50.07	0.88	21.07

Table 52: Linear regression with 10 predictors, the predictor of interest is continuous, other

Table 56: Logistic regression with two independent continuous predictors, $N_{max} = 600$.

Table 58: Logistic regression with two continuous predictors, Pearson correlation between the predictors = 0:8, N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.291	0.0498	0.5715	599.88	3.46	0.0237	0.0117	1062.13
	0.469	0.0471	0.8734	525.05	97.51	0.0532	0.0185	433.57
	0.702	0.0487	0.8924	300.21	129.10	0.1030	0.0313	213.53
	1.127	0.0406	0.8922	124.32	78.21	0.2022	0.0536	103.81
50	0.291	0.0493	0.5533	600.00	0.3191	0.0235	0.0117	1062.13
	0.469	0.0475	0.8321	469.13	88.06	0.0529	0.0192	433.57
	0.702	0.0468	0.8449	215.08	52.58	0.1044	0.0349	213.53
	1.127	0.0397	0.8351	83.98	20.72	0.2040	0.0592	103.81
100	0.291	0.0483	0.5522	600	0	0.0234	0.0118	1062.13
	0.469	0.0485	0.8262	434.66	64.92	0.0532	0.0195	433.57
	0.702	0.0469	0.8203	194.63	30.11	0.1043	0.0352	213.53
	1.127	0.0470	0.8533	100.26	2.00	0.2059	0.0587	103.81

Table 59: Logistic regression with two independent binary predictors, N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0493	0.8905	507.22	67.53	0.0238	0.0127	391.75
	0.459	0.0489	0.8930	214.95	67.46	0.0571	0.0290	157.01
	0.667	0.0477	0.8888	106.20	60.97	0.1098	0.0515	78.71
	1.003	0.0418	0.8782	51.19	61.59	0.2060	0.0844	39.74
50	0.286	0.0497	0.8321	431.31	35.73	0.0242	0.0138	391.75
	0.459	0.0500	0.8355	168.08	15.76	0.0583	0.0315	157.01
	0.667	0.0454	0.8272	79.79	7.44	0.1127	0.0561	78.71
	1.003	0.0490	0.9161	50.04	0.85	0.2163	0.0862	39.74
100	0.286	0.0506	0.8157	405.52	14.85	0.0243	0.0139	391.75
	0.459	0.0483	0.8144	157.76	5.81	0.0601	0.0333	157.01
	0.667	0.0501	0.9092	100.00	0.03	0.1162	0.0557	78.71
	1.003	0.0526	0.9982	100	0	0.2122	0.0702	39.74

Table 60: Logistic regression with two binary predictors, Pearson correlation between the predictors = 0:4 (tetrachoric correlation = 0:5878), N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0479	0.8546	544.52	64.15	0.0236	0.0125	465.48
	0.459	0.0502	0.8858	251.90	83.06	0.0564	0.0269	186.83
	0.667	0.0461	0.8806	123.85	66.24	0.1101	0.0497	93.79
	1.003	0.0416	0.8629	59.81	66.09	0.2075	0.0813	47.28
50	0.286	0.0490	0.8379	513.44	61.26	0.0238	0.0125	465.48
	0.459	0.0480	0.8412	209.22	49.24	0.0574	0.0289	186.83
	0.667	0.0440	0.8320	99.36	23.17	0.1099	0.0521	93.79
	1.003	0.0413	0.8649	51.78	6.87	0025105	005837	47.28
100	0.286	0.0488	0.8268	489.12	48.26	0.0240	0.0128	465.48
	0.459	0.0470	0.8160	191.14	21.86	0.0592	0.0302	186.83
	0.667	0.0465	0.8478	101.45	5.36	0.1122	0.0522	93.79
	1.003	0.0500	0.9923	100.00	0.03	0.2125	0.0703	47.28

Table 61: Logistic regression with two binary predictors, Pearson correlation between the predictors = 0:8 (tetrachoric correlation = 0:9511), N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0484	0.5017	533.25	81.80	0.0239	0.0126	107 Q q 4 0 0 -121 2668.9

Table 62: Logistic regression with two independent predictors (binary and continuous), the predictor of interest is binary, N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0494	0.8866	509.00	68.09	0.0235	0.0126	391.74
	0.459	0.0489	0.8957	219.25	78.61	0.0569	0.0286	157.01
	0.667	0.0474	0.8908	108.95	67.96	0.1091	0.0508	78.71
	1.003	0.0419	0.8812	52.84	65.30	0.2037	0.0807	39.74
50	0.286	0.0509	0.8383	430.88	35.14	0.0240	0.0136	391.74
	0.459	0.0490	0.8350	167.59	14.77	0.0582	0.0313	157.01
	0.667	0.0457	0.8331	79.66	6.97	0.1130	0.0558	78.71
	1.003	0.0502	0.9190	50.03	0.59	0.2137	0.0838	39.74
100	0.286	0.0490	0.8203	405.28	14.65	0.0245	0.0139	391.74
	0.459	0.0489	0.8131	157.69	5.69	0.0596	0.0327	157.01
	0.667	0.0505	0.9093	100.00	0.01	0.1162	0.0557	78.71
	1.003	0.0502	0.9981	100	0	0.2116	0.0698	39.74

Table 63: Logistic regression with two predictors (binary and continuous), the predictor of interest is binary, Pearson correlation between the predictors = 0:4 (original correlation = 0:5037), N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0489	0.8648	554.50	58.52	0.0235	0.0121	467.00
	0.459	0.0476	0.8910	264.90	98.79	0.0561	0.0264	187.54
	0.667	0.0462	0.8834	132.19	78.02	0.1077	0.0473	93.60
	[redacted]	[redacted]	[redacted]					

Table 64: Logistic regression with two predictors (binary and continuous), the predictor of interest is binary, Pearson correlation between the predictors = 0:8 (original correlation = 0:9991), N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0507	0.5537	600.00	0.65	0.0231	0.0116	1073.74
	0.459	0.0509	0.8695	531.41	80.69	0.0535	0.0185	428.96
	0.667	0.0480	0.8919	298.50	108.22	0.1036	0.0321	215.89
	1.003	0.0454	0.8892	141.19	84.00	0.2023	0.0537	108.67
50	0.286	0.0514	0.5563	600	0	0.0233	0.0117	1073.74
	0.459	0.0488	0.8401	469.82	67.86	0.0534	0.0193	428.96
	0.667	0.0488	0.8317	224.60	37.77	0.1053	0.0354	215.89
	1.003							

Table 66: Logistic regression with two predictors (continuous and binary), the predictor of

Table 68: Logistic regression with 10 independent continuous predictors, N_{max} = 600.

Table 71: Logistic regression with 10 independent binary predictors, N_{max} = 600.

n	$_1$	Type I error	Power	E(N)	SD(N)	E(R^2)	SD(R^2)	Target SS
20	0.286	0.0511	0.9369	599.62	5.18	0.0363	0.0132	391.79
	0.459	0.0506	0.9950	515.07	112.61	0.0694	0.0213	157.02
	0.667	0.0554	0.9940	333.89	157.14	0.1310	0.0381	78.72
	1.003	0.0563	0.9917	174.35	130.49	0.2553	0.0686	39.74
50	0.286	0.0539	0.9263	581.64	32.53	0.0366	0.0134	391.79
	0.459	0.0509	0.9469	266.10	66.11	0.0821	0.0276	157.02

Table 75: Logistic regression with 10 predictors, the predictor of interest is binary, other 9 predictors are continuous, pairwise Pearson correlation = 0.4 (original pairwise correlation = 0.5037), N

Table 77: Logistic regression with 10 independent predictors, the predictor of interest is continuous, other 9 predictors are binary, N_{max} = 600.

n	$_1$	Type I error