Power of Logistic Model with Surrogate Measures for Both Outcome and Covariate in Genetic-Disease Association Study

Munechika Misumi, Tomomi Yamada, Yoshiaki Nose, and Tsuyoshi Nakamura

Abstract— This study deals with the combined effects of misclassifications incidental to diagnostics, or a binary response, and genotyping, or a discrete covariate, on the statistical power of a logistic model testing for a treatment effect. The loss of power due to differential and nondifferential misclassifications in a response and a covariate, respectively, has not been well documented. This paper first obtained a general expression for the loss of statistical power due to those misclassifications based on the Pitman asymptotic relative efficiency (ARE). Numerical studies confirmed the validity of the general expression for a reasonable sample size. It revealed that the effect of even low misclassification rates is not negligible. Misclassifications in both response and covariates should be taken into account when determining the sample size.

***Index Terms*—Pitman Asymptotic Relative Efficiency, Sample Size, Statistical Power, Genetic-Disease Association Study.**

I. Introduction

A binary logistic regression is one of the most popular models for genetic disease association studies, biomedical data analysis, and epidemiological studies. The loss of statistical power due to nondifferential misclassifications in a response or covariates has been well documented [1-2], however, the loss due to differential misclassifications in a response with a discrete covariate subject to nondifferential misclassifications has not been reported yet. As for an example of misclassification in a response, consider an ordinary cancer clinical trial where the effect of a treatment is diagnosed based solely on diagnostic imaging. The response is inevitably subject to differential errors, since the diagnosis is subjective rather than objective. We investigated the power of a logistic model with a binary response subject to differential misclassifications and a covariate subject to nondifferential misclassifications using the Pitman ARE [3]. An application to a genetic disease association study is demonstrated.

Manuscript received on August 16th, 2010.

M Misumi is with Graduate school of Science and Technology, Nagasaki University, Japan (corresponding author to provide phone: +81-95-819-2747; fax: +81-95-843-1782;e-mail: mmisumi@med.kyushu-u.ac.jp).

T Yamada is with Department of Social and Environmental Medicine, Mie University, Japan (e-mail:t-yamada@doc.medic.mie-u.ac.jp).

Y Nose is with Graduate School of Health Sciences, Kumamoto Health Sciences University, Japan (e-mail: nose@kumamoto-hsu.ac.jp).

T Nakamura is with graduate school of sciences and technology, Nagasaki University, Japan (e-mail: naka@nagasaki-u.ac.jp).

II. Pitman Asymptotic Relative Efficiency

To compare two statistical testing procedures A and B, if A requires n_1 observations and B requires n_2 observations to achieve the same power, the efficacy of A relative to that of B is often compared by the ratio n_1/ n_2. Especially, for large samples, asymptotic relative efficiency (ARE) is used for the comparison. The earliest and most popular approach to ARE was introduced by the lecture notes of Pitman in 1978 [4]. It is practical that the measure of ARE does not depend on a particular alternative. Also, the relationship between the Pitman ARE and the asymptotic correlation of test statistics is widely examined [3] [5].

Definition (Asymptotic Relative Efficiency)

If two consistent estimators W_n and V_m for $\tau(\theta)$ are such that

$$\sqrt{n}[W_n - \tau(\theta)] \to N(0, {\sigma_W}^2)$$

$$\sqrt{m}[V_m - \tau(\theta)] \to N(0, {\sigma_V}^2)$$

then, Pitman asymptotic relative efficiency (ARE) of V_m with respect to W_n is defined as

$$ARE(V_m, W_n) = \frac{{\sigma_W}^2}{{\sigma_V}^2}.$$

To understand the usefulness of ARE, let us suppose that the asymptotic variances of W_n and V_m are equal for particular values of n and m; that is, ${\sigma_W}^2 / n = {\sigma_V}^2 / m$ holds. It then follows that the power of the test based on W_n is equal to the power of the test based on V_m. In other words, in order to retain the same power when replacing W_n by V_m, it is necessary and sufficient that $n/m= ARE(V_m, W_n)$. Thus ARE directly tells us the sample size necessary to attain the same power as a standard test.

III. Method

Let Y and Z denote a binary response (0=non-responder, 1=responder) and an explanatory variable, respectively. We consider a test for association between Z and Y when Y is binary and related to Z by the logistic relationship:

$$\ln\{\Pr(Y = 1 \mid Z) / \Pr(Y = 0 \mid Z)\} = \alpha + \beta z .$$

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The null hypothesis of no association is given by $\beta = 0$, and can be assessed by one of three asymptotically equivalent tests: the Wald test, the likelihood ratio test, or the score test [6]. The score test is of the form:

$$Q(Y,Z) = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})Y_i}{\left\{\bar{Y}(1-\bar{Y})\sum_{i=1}^{n}(Z_i - \bar{Z})^2\right\}} \quad (1)$$

where Y_i and Z_i are the values of *Y* and *Z* for the i th of n subjects, $\bar{Y} = \sum Y_i / n$, and $\bar{Z} = \sum Z_i / n$. When $\beta = 0$, $Q(Y,Z)$ is asymptotically *N(0,1)*. Then, suppose that we can only observe Z^*, the surrogate of *Z*, and that the values of Y^* observed for *Y* are subject to misclassification and we test for association with a statistic $Q(Y^*,Z^*)$. That is, we use a test with (1), but replacing *Z* and Y by Z^* and Y^*, respectively. Also, we assume that Z and Z^* have finite second moments, and that usual regularity conditions apply to ensure the asymptotic distribution of N(0.1) of (1). When $\beta = 0$, Y^* is not associated with Z^* and hence $Q(Y^*,Z^*)$ is also asymptotically N(0,1). However, $Q(Y^*,Z^*)$ is less efficient than $Q(Y,Z)$ when the null hypothesis does not hold. [1] described this loss of efficiency using the ARE of $Q(Y^*,Z)$ to $Q(Y,Z)$ for local alternatives to $\beta = 0$, denoted ARE(Y^*:*Y*). We extend these results to the case where *Z* is not observed, and only a surrogate Z^* is available in place of *Z*.

For practical use, we derived following proposition.

Proposition

The Pitman ARE of (Y^*, Z^*) to (*Y*, *Z*), denoted as *ARE(Y^*,Z^*|Y,Z)* hereafter, is equal to the product of the squared correlation coefficient $\rho(Y,Y^*)$ between the true response and the error-prone response and the squared correlation coefficient $\rho(Z,Z^*)$ between the true and surrogate covariates (See APPENDIX for the proof).

This result indicates that in a logistic model, when a binary response is subject to misclassification, and a covariate is subject to measurement error, the power of a test of the model using a sample of size *N* is approximately equal to the power of the corresponding test with the exact response and covariate values that uses a sample of size $N \cdot \rho(Y,Y^*)^2 \cdot \rho(Z,Z^*)^2$. Once a sample size N_0 is obtained for a test with the exact response and covariate values, the sample size required to attain the same power for a test using error-prone response and covariate values is then obtained by dividing N_0 by $\rho(Y,Y^*)^2 \cdot \rho(Z,Z^*)^2$, which is usually obtained by analytical calculations.

Definition (Nondifferential and differential misclassification)

When the misclassifications of a disease to a category occur independently from the other classifications to categories, the misclassification is called nondifferential. Misclassification is called differential otherwise.

It is well-known that nondifferential misclassifications cause "attenuation", or the bias to the null, in the estimation of a regression coefficient. In contrast, the effect of differential misclassification on the parameter estimate is not so simple. It can cause either underestimation or overestimation of the parameter depending on the bias of classification. We assessed the effects of differential misclassifications on the power of the test.

IV. APPLICATION

We applied the Pitman's ARE to calculate the sample size required to obtain enough power for the association study of genetic polymorphisms for response to interferon (IFN) –alpha therapy [7]. In the study, we followed the patients with metastatic renal cancer and known a certain SNP, and the "reduction of tumor mass" would be observed as the response variable. Then, the diagnostic of the "reduction of tumor mass" with a misclassification rate of 4% would normally be regarded as fairly accurate in cancer clinical trials [8] [9]. Also, although recently the genotyping error rate is getting quite small, the impact of genotyping error could easily be taken into consideration for the sample size calculation by using the Pitman's ARE.

Parameter Specification

For genetic-disease association studies, prospective logistic regression is the standard method of analysis. The program is designed for a cohort study to explore the association between single nucleotide polymorphisms (SNPs) and responses of treatments. The expected response rate, hereafter denoted by *s*, is usually determined by medical specialists with knowledge and experience of the disease of interest. Also, the genetic model, or the type of effect of the genotype on patient response, or diagnosis, is specified as a parameter of the program. This is of intermediate, dominant or recessive type. Suppose that C and T denote low and high response rate alleles, respectively, which may be replaced by any such allele. Then, the genetic model can be coded as Table 1.

TABLE 1. NUMERIC CODING FOR GENOTYPES BY GENETIC MODEL WHERE C AND T DENOTE LOW AND HIGH RESPONSE RATE ALLELES, RESPECTIVELY.

Genetic model	Genotypes		
	CC	CT	TT
Intermediate	0	1	2
Dominant	0	1	1
Recessive	0	0	1

Next, expected misclassification rates for the response, the false negative rate, the false positive rate, denoted by ε_0, ε_1 respectively, must be specified. Table 2 shows the joint distribution of *Y* and Y^* where *s*=Pr(*Y*=1) denote the probability of a responder. The $ARE(Y^* \mid Y)$ can be obtained as:

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$ARE(Y \mid Y^*) = \rho(Y, Y^*)^2 = \frac{(1-\varepsilon_0-\varepsilon_1)^2 s\{1-s\}}{\{(1-\varepsilon_0-\varepsilon_1)s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)s-\varepsilon_0\}}$$

TABLE 2. THE JOINT PROBABILITY WHEN THE FALSE POSITIVE RATE AND THE FALSE NEGATIVE RATE IS ε_0 AND ε_1, RESPECTIVELY, AND THE TRUE PROBABILITY OF A RESPONDER IS s.

Diagnosis	TRUE	
	Responder (Y=1)	Non-responder (Y=0)
Responder (Y^*=1)	$(1-\varepsilon_1)s$	$\varepsilon_0(1-s)$
Non-responder (Y^*=0)	$\varepsilon_1(1-s)$	$(1-\varepsilon_0)s$

Next, the non-differential genotyping errors with an error rate δ, as given in Table 2, should be specified. When the genotyping error is not assumed, it can be specified as $\delta = 0$.

TABLE 3. ERROR MATRIX FOR GENOTYPING ERROR.

Observed	TRUE	
	C	T
C	$1-\delta$	δ
T	δ	$1-\delta$

Let Z specify the number of allele C in the genotype. Then, Z =0, 1, or 2 when the genotype is CC, CT, or TT, respectively. Let Z^* denote the observed genotype, then
$p^*_C = \Pr(Z^* = C) = (1-\delta)p_C + \delta\cdot(1-p_C)$.
It is straightforward to obtain the following equation:

$$ARE(Z \mid Z^*) = \rho(Z, Z^*)^2 = (1-2\delta)^2 \frac{Var(Z)}{Var(Z^*)}$$

$$= (1-2\delta)^2 \frac{p_C(1-p_C)}{p^*_C(1-p^*_C)}.$$

The odds ratio

The codes assigned to the genotypes by the genetic models are shown in Table 1. If the genetic model is of intermediate type, then (CC, CT, TT)=(0,1,2), and if it is of dominant or recessive type, then (CC, CT, TT)=(0,1,1), (0,0,1), respectively. When logistic regression is conducted, the odds ratio for genotypes CT and TT compared to CC is $e^{a\beta}$ and $e^{b\beta}$, respectively, where (*0, a, b*) denote the code for (CC, CT, TT) determined by the genetic model as depicted in Table 1.

The odds of a response for CC

Let w denote the odds of genotype CC, that is,

$$w = \frac{\Pr(responder \mid CC)}{\Pr(non-responder \mid CC)}.$$

Because the odds of CT and TT are $e^{a\beta}w$ and $e^{b\beta}w$, respectively, and the Hardy-Weinberg equilibrium that the frequency of CC,CT,TT is proportional to $p_C^2, 2p_Cp_T, p_T^2$, where $p_C, p_T = 1-p_C$ is the proportion of allele C, T, respectively, is assumed to be hold, it follows from the definition of *s*, the overall response rate, that

$$s = \left(\frac{w}{1+w}\right)\cdot p_C^2 + \left(\frac{e^{a\beta}\cdot w}{1+e^{a\beta}\cdot w}\right)\cdot 2p_Cp_T + \left(\frac{e^{b\beta}\cdot w}{1+e^{b\beta}\cdot w}\right)\cdot p_T^2.$$

It is straightforward to show that the equation has a unique solution for w in the range $w > 0$ and is easily attained using the Newton-Raphson iteration algorithm with initial value $w = 0$.

Calculation

The program calculates the sample size and power using values of the following items:

1) Expected proportion of allele C in a population, or p_C
2) Effect size: Odds ratio based on genetic models
3) The overall response rate *s* (<0.5)
4) The genetic model, or the value of *a* and *b*
5) A false negative rate of response misclassification ε_0
6) A false positive rate of response misclassification ε_1
7) Sample size *N*
8) A two-sided significance level *P*, with a default value of 0.05
9) The number of iterations *M*, with a default value of 10,000
10) A non-differential genotyping error rate δ.

The power of a test for the null hypothesis of no association between the SNPs and the treatment is obtained according to the following steps.

Step 1: For each subject *i*, two independent (0,1) uniform random variables U_{i1} and U_{i2} are generated. If $U_{i1} < p_C$, then one allele is C, otherwise T. In the same way, the other allele is determined using U_{i2}. Genotypes CC, CT and TT are coded as 0, a and b, respectively, where a=1 and b=2, a=b=1, or a=0 and b=1, depending on whether the genetic model is intermediate, dominant or recessive. Next, a (0,1) uniform random variable U_{i3} is generated, so that the subject is determined as a responder when $U_{i3} < w/(1+w)$, $U_{i3} < e^{a\beta}w/(1+e^{a\beta}w)$ or $U_{i3} < e^{b\beta}w/(1+e^{b\beta}w)$, depending on whether the genotype is CC, CT or TT, respectively. The subject is determined as a non-responder otherwise.

Step2: Logistic regression with responder/non-responder as the dependent variable and the genotype as the covariate is performed. Then, the maximum likelihood estimate $\hat{\beta}$ and the p-value for the null hypothesis that the population regression coefficient $\beta = 0$, i.e., no association between the genotype and treatment.

Step3: Power is obtained as the proportion of simulations in which we rejected the null hypothesis. That is, repeating the simulation and regression analysis M times, power is calculated as the proportion of the significant results among M testing. Then, average of estimates $\hat{\beta}$ is also obtained,

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

which should be approximately equal to β .

Step4: The power obtained in Step3, say F, assumes no misclassification. In the presence of misclassification in the response/non-response diagnosis and/or the genotyping of the covariate, the power of the sample size N is asymptotically equal to that of the sample size $N \cdot ARE(Y^*, Z^* | Y,Z)$ without misclassification. Therefore, when misclassifications exist, the sample size required to attain the same power as F is increased to $N \cdot ARE(Y^*, Z^* | Y,Z)^{-1}$. To determine the sample size for power required, the program should be run several times with different sample size. When a sample size produces a power close enough to that required, this is the sample size that should be used when no misclassification is expected.

V. RESULTS

As a power of 80% is normally specified by protocols of clinical trial, we first obtained the sample sizes required to attain this power without misclassification for expected response rates of 15%, 20%, 30%, and 40%. Since the response used in the study was "tumor mass reduction" diagnosed from a two-dimensional image, misclassification in response was inevitable as well as the genotyping error expected in genomic data. Based on the Pitman ARE, we also obtained the sample sizes necessary to attain the power of 80% for several possible misclassification rates.
When the intermediate genetic model is used, and the population allele frequency of C, p_C , and the odds ratio of CT to CC are assumed as 0.6 and 3.22 ($=e^{1.2}$), respectively, the proportion of the genotypes CC, CT and TT are 0.36, 0.48 and 0.16, respectively. Then, the odds ratio of TT to CC is calculated as 11.0 ($=e^{2.4}$). Table 4 and table 5 list the sample sizes required to attain a power of 80% by response rate, diagnostic error rate of response, and genotyping error rate.

TABLE 4. SAMPLE SIZE TO ATTAIN THE POWER OF 80% WHEN $p_C = 0.63$ AND THE FALSE POSITIVE RESPONSE RATE IS $e_1 = e_0 / 2$.

		False negative rate ε_0 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	95	106	110	126	143	161	182
	0.005	98	108	112	128	146	165	186
	0.01	100	110	115	131	149	168	190
	0.02	104	116	120	137	156	176	198
	0.03	109	121	125	143	163	184	207
	0.04	114	127	131	150	170	193	217
	0.05	119	133	137	157	178	202	227
0.3	0	61	65	67	72	78	85	93
	0.005	63	67	68	74	80	87	95
	0.01	64	68	70	75	82	89	97
	0.02	67	71	73	79	86	93	101
	0.03	70	74	76	82	89	97	106
	0.04	73	78	80	86	94	102	110
	0.05	77	82	83	90	98	106	116

With regarding the diagnostic error rate, we assumed in each of the simulations that the false positive rate ε_1 and false negative rate ε_0 were set equal to the half of the other, that is either $\varepsilon_1 = \varepsilon_0/2$ or $\varepsilon_0 = \varepsilon_1/2$. In Table 4-9, Rate[a] is the rate of response, and Error[b] is the value of genotyping error rate. The values of both categories are specified for the calculation.
A misclassification rate of 4% would normally be regarded as a fairly accurate diagnosis [7], and the response rate of IFN-α therapy is known to be around 0.15. When the response rate is 15%, the false negative diagnostic error rate is 4%, the false positive diagnostic error rate is 2%, and the genotyping error rate is 3%, the sample size must be increased from 95 to 143 (a 50% increase) to attain the power of 80% (Table 4).
On the other hand, when the false positive diagnostic error rate is 4% and the false negative rate is half of it, and the other settings are the same, the sample size increase is from 95 to 130 (37%).

TABLE 5. SAMPLE SIZE TO ATTAIN THE POWER OF 80% WHEN $p_C = 0.63$ AND THE FALSE NEGATIVE RESPONSE RATE IS $e_0 = e_1 / 2$.

		False positive rate ε_1 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	95	102	111	114	125	138	151
	0.005	98	104	107	117	128	141	154
	0.01	100	107	109	119	131	144	158
	0.02	104	111	114	125	137	150	165
	0.03	109	117	119	130	143	157	172
	0.04	114	122	125	137	150	164	180
	0.05	119	128	131	143	157	172	189
0.3	0	61	64	68	70	75	80	86
	0.005	63	66	67	71	76	82	88
	0.01	64	67	68	73	78	83	89
	0.02	67	70	71	76	81	87	93
	0.03	70	73	75	80	85	91	98
	0.04	73	77	78	83	89	95	102
	0.05	77	81	82	87	93	100	107

The actual loss due to the error is tabulated in Table 6 and Table 7. When the false negative rate is greater than the false positive rate, the power tends to be lower.

TABLE 6. POWER WITH THE SAMPLE SIZE OBTAINED BY $N_0 \cdot ARE$ WHEN $p_C = 0.63$ AND THE FALSE POSITIVE RESPONSE RATE IS $e_1 = e_0 / 2$.

		False negative rate ε_0 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	0.802	0.749	0.733	0.672	0.598	0.547	0.480
	0.005	0.790	0.746	0.726	0.652	0.599	0.532	0.469
	0.01	0.783	0.744	0.720	0.659	0.585	0.512	0.463
	0.02	0.765	0.720	0.693	0.631	0.563	0.501	0.430
	0.03	0.743	0.687	0.670	0.612	0.530	0.482	0.415
	0.04	0.714	0.662	0.653	0.593	0.505	0.458	0.381
	0.05	0.693	0.651	0.622	0.550	0.492	0.415	0.351
0.3	0	0.804	0.770	0.767	0.717	0.669	0.623	0.577
	0.005	0.786	0.763	0.754	0.702	0.655	0.626	0.567
	0.01	0.777	0.746	0.734	0.692	0.651	0.614	0.551
	0.02	0.744	0.724	0.720	0.671	0.636	0.574	0.537
	0.03	0.741	0.706	0.690	0.652	0.590	0.551	0.497
	0.04	0.708	0.684	0.672	0.630	0.571	0.516	0.467
	0.05	0.684	0.646	0.640	0.598	0.559	0.500	0.446

TABLE 7. POWER WITH THE SAMPLE SIZE OBTAINED BY $N_0 \cdot ARE$ WHEN $p_C = 0.63$ AND THE FALSE NEGATIVE RESPONSE RATE IS $e_0 = e_1 / 2$.

		False positive rate ε_1 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	0.792	0.770	0.756	0.715	0.680	0.622	0.574
	0.005	0.790	0.762	0.746	0.699	0.656	0.619	0.564
	0.01	0.783	0.755	0.735	0.697	0.649	0.599	0.561
	0.02	0.765	0.730	0.721	0.678	0.642	0.582	0.521
	0.03	0.748	0.711	0.700	0.657	0.607	0.565	0.512
	0.04	0.721	0.693	0.679	0.634	0.585	0.541	0.486
	0.05	0.699	0.657	0.651	0.604	0.564	0.508	0.458
0.3	0	0.801	0.773	0.764	0.735	0.707	0.660	0.619
	0.005	0.779	0.761	0.757	0.728	0.696	0.655	0.614
	0.01	0.777	0.752	0.745	0.718	0.675	0.638	0.603
	0.02	0.744	0.725	0.721	0.691	0.664	0.620	0.579
	0.03	0.741	0.708	0.700	0.679	0.638	0.601	0.549
	0.04	0.713	0.680	0.677	0.639	0.606	0.566	0.520
	0.05	0.686	0.667	0.653	0.610	0.587	0.537	0.489

Table 8 and Table 9 present the power calculated with

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

samples generated by a simulation. The sample was randomly generated as the response variable and genotype covariate contained the misclassification error specified. The response variable and covariate were generated in the process of Step1 in the Calculation section. Then, the misclassification was made using other uniform random variables U_ε. U_δ. If $U_\varepsilon < \varepsilon$ (if the subject is a responder, $\varepsilon=\varepsilon_0$, and if the subject is a nonresponder, $\varepsilon=\varepsilon_1$) then the subjects who were responders changed to non-responders, and vice versa. Similarly, the genotype classification was changed by switching the allele using the value of the uniform random variable for the subject. Then, the power was calculated with the simulated sample of size N_0.

The corresponding figures in Table 6 and Table 7 agreed well to Table 8 and Table 9, respectively, indicating that the asymptotic formula is highly accurate for the small sample sizes.

TABLE 8. POWER WITH THE SAMPLE OF SIZE N_0 OBTAINED BY SIMULATION WITH WHEN $p_C = 0.63$ AND THE FALSE POSITIVE RESPONSE RATE IS $e_1 = e_0 / 2$.

		False negative rate ε_0 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	0.802	0.755	0.739	0.686	0.632	0.579	0.536
	0.005	0.792	0.747	0.732	0.676	0.623	0.570	0.523
	0.01	0.778	0.755	0.721	0.662	0.602	0.560	0.510
	0.02	0.756	0.722	0.712	0.641	0.603	0.547	0.490
	0.03	0.751	0.691	0.689	0.624	0.578	0.522	0.489
	0.04	0.730	0.678	0.678	0.610	0.561	0.506	0.460
	0.05	0.710	0.658	0.647	0.586	0.535	0.483	0.446
0.3	0	0.804	0.777	0.767	0.735	0.699	0.648	0.621
	0.005	0.794	0.766	0.757	0.729	0.691	0.639	0.610
	0.01	0.773	0.755	0.740	0.708	0.681	0.629	0.585
	0.02	0.757	0.738	0.725	0.683	0.655	0.613	0.589
	0.03	0.741	0.716	0.700	0.671	0.632	0.595	0.556
	0.04	0.718	0.702	0.685	0.640	0.609	0.573	0.539
	0.05	0.704	0.672	0.665	0.624	0.588	0.552	0.520

TABLE 9. POWER WITH THE SAMPLE OF SIZE N_0 OBTAINED BY SIMULATION WITH WHEN $p_C = 0.63$ AND THE FALSE NEGATIVE RESPONSE RATE IS $e_0 = e_1 / 2$.

		False positive rate ε_1 of response						
Rate[a]	Error[b]	0	0.015	0.02	0.04	0.06	0.08	0.1
0.15	0	0.792	0.761	0.761	0.721	0.690	0.641	0.606
	0.005	0.784	0.754	0.752	0.711	0.682	0.636	0.596
	0.01	0.783	0.760	0.738	0.711	0.658	0.634	0.586
	0.02	0.757	0.737	0.730	0.685	0.647	0.608	0.567
	0.03	0.742	0.715	0.708	0.661	0.629	0.587	0.550
	0.04	0.730	0.697	0.697	0.651	0.609	0.567	0.531
	0.05	0.705	0.684	0.664	0.626	0.584	0.552	0.513
0.3	0	0.801	0.783	0.773	0.738	0.720	0.679	0.650
	0.005	0.795	0.777	0.763	0.731	0.710	0.671	0.641
	0.01	0.783	0.765	0.754	0.722	0.686	0.649	0.635
	0.02	0.763	0.740	0.737	0.709	0.678	0.642	0.619
	0.03	0.744	0.724	0.712	0.684	0.650	0.617	0.594
	0.04	0.722	0.699	0.693	0.663	0.637	0.602	0.570
	0.05	0.701	0.681	0.680	0.648	0.609	0.582	0.545

VI. DISCUSSION

We investigated the impact of misclassifications of both the response variable and a covariate on the power of the test based on the logistic regression model. The impact was demonstrated with an application to a genetic disease association study. Even with the realistic error rate assumptions, dramatic increases of sample sizes should be taken into account when designing genetic disease association studies. As far as the authors are concerned, this is the first time to obtain the Pitman asymptotic relative efficiency to demonstrate the impact of misclassifications incidental to both a response variable and a covariate simultaneously. We observed a trend that the differences between the powers obtained by sample size calculated using ARE and those obtained by simulation got larger as the misclassification errors became larger. This might suggest that larger sample size were preferred to support the asymptotic property we showed when data were contaminated by large errors. Although we dealt with just a locus, an extension to multiple loci will be accomplished using the approach described in [10] for the sample size calculation of multivariate logistic regressions. Since SNPs tend to be highly correlated, sample size inflation should be taken into account, too. Also, as the Pitman ARE can be applied to continuous variables, our approach can be extended to other multivariate regression models. We will be able to see similar trends in other regression analysis.

APPENDIX

Let (Z_i, Z_i^*, Y_i, Y_i^*) denote independent observations following the logistic regression model in Section III. Further, we set a local alternative $\beta = \beta_0 / \sqrt{n}$, where β_0 is a constant not depending on n.
Define $p(Z) = \Pr(Y = 1 \mid Z) = [1 + \exp\{-(\alpha + \beta Z)\}]^{-1}$. Since Z^* is a surrogate for Z, it follows that

$$\Pr(Y_i = 1 \mid Z_i, Z^*_i) = p(Z_i)$$

and

$$p^*(Z_i) = \Pr(Y_i^* = 1 \mid Z_i, Z_i^*) = (1-\varepsilon_1)p(Z_i) + \varepsilon_0\{1 - p(Z_i)\},$$

where ε_0, ε_1 is the false negative rate, the false positive rate, respectively, of Y and

Let $G(z)$ and $H(z^* \mid z)$ denote the distribution function of Z and the distribution function of Z^* given Z, respectively. The overall response rate is obtained as

$$E(Y = 1) = \int E(Y = 1 \mid Z = z)dG(z) = \int p(z)dG(z)$$

Denote the numerator and denominator of (1) by $U(Y,Z)$ and $I(Y,Z)$, respectively. Then, the conditional expectation of $U(Y^*, Z^*)$ given Z and Z^* is

$$E[U(Y^*, Z^*) \mid Z, Z^*] = \sum_{i=1}^{n} (Z^*_i - \bar{Z})\{(1-\varepsilon_1)p(Z_i) + \varepsilon_0(1 - p(Z_i))\}$$

Therefore, the unconditional expectation is

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$E[U(Y^*,Z^*)] = (1-\varepsilon_0-\varepsilon_1)(n-1)\{E[Z^* p(Z)] - E(Z^*)E[p(Z)]\}$

since $E[Z^* p(Z)] = \iint z^* p(z) dH(z^* \mid z) dG(z)$.

Taking the derivative with respect to β at β =0, we have

$$\frac{d}{d\beta} E[Z^* p(Z)]\Big|_{\beta=0} = s(1-s)E(Z^* Z), \text{ and}$$

$$\frac{d}{d\beta} E[Z^*]E[p(Z)]\Big|_{\beta=0} = s(1-s)E(Z^*)E(Z),$$

where we put $s = \{1+\exp(-\alpha)\}^{-1}$.
Thus,

$$\frac{d}{d\beta} E[U(Y^*,Z^*)]\Big|_{\beta=0} = (1-\varepsilon_0-\varepsilon_1)(n-1)s(1-s)Cov(Z^*,Z).$$

Similar calculation show that $E[U(Y,Z)] = \sum_{i=1}^{n}(Z_i - \bar{Z})p(Z_i)$
from which it follows that

$$\frac{d}{d\beta} E[U(Y,Z)]\Big|_{\beta=0} = (n-1)s(1-s)Var(Z).$$

On the other hand,

$$\frac{I(Y^*,Z^*)}{n} = \frac{\bar{Y}^*(1-\bar{Y}^*)\sum_{i=1}^{n}(Z^*_i - \bar{Z}^*)^2}{n}$$

and

$$\frac{I(Y,Z)}{n} = \frac{\bar{Y}(1-\bar{Y})\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n}.$$

Then, since β is $O(1/\sqrt{n})$, $I(Y,X)/n \to_p s(1-s)Var(Z)$ and $I(Y^*,Z^*)/n \to_p \{(1-\varepsilon_0-\varepsilon_1)^2 s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)^2 s-\varepsilon_0\}Var(Z^*)$, as $n \to \infty$. Therefore, the Pitman asymptotic relative efficiency of Y^* to Y is

$$\frac{\{(1-\varepsilon_0-\varepsilon_1)(n-1)s(1-s)Cov(Z,Z^*)\}^2}{\{(1-\varepsilon_0-\varepsilon_1)s(1-s)+\varepsilon_0\}\{(1-(1-\varepsilon_0-\varepsilon_1)s(1-s)-\varepsilon_0\}Var(Z^*)\}} \cdot \frac{s(1-s)Var(Z)}{\{(n-1)s(1-s)Var(Z)\}^2}$$

$$= \frac{(1-\varepsilon_0-\varepsilon_1)^2 s(1-s)}{\{(1-\varepsilon_0-\varepsilon_1)s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)s-\varepsilon_0\}} \cdot \frac{Cov(Z,Z^*)^2}{Var(Z)Var(Z^*)}$$

$$= \frac{(1-\varepsilon_0-\varepsilon_1)^2 s(1-s)}{\{(1-\varepsilon_0-\varepsilon_1)s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)s-\varepsilon_0\}} \cdot \rho(Z,Z^*)^2,$$

where $\rho(Z,Z^*)$ is the correlation coefficient between Z and Z^*.

Aside from this, since $Var(Y \mid Z) = p(Z)\{1-P(Z)\}$, $Var(Y^* \mid Z) = \{(1-\varepsilon_0-\varepsilon_1)p(Z)+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)p(Z)-\varepsilon_0\}$, and $Cov(Y,Y^*) = (1-\varepsilon_0-\varepsilon_1)p(Z)\{1-p(Z)\}$, the squared correlation coefficient between Y and Y^* given Z, Z^* is $[correlation(Y,Y^* \mid Z,Z^*)]^2$

$$= \frac{(1-\varepsilon_0-\varepsilon_1)^2 p(Z)\{1-p(Z)\}}{\{(1-\varepsilon_0-\varepsilon_1)p(Z)+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)p(Z)-\varepsilon_0\}}.$$

Again, since β is $O(1/\sqrt{n})$, as $n \to \infty$
$[correlation(Y,Y^* \mid Z,Z^*)]^2 \to_p$

$$\frac{(1-\varepsilon_0-\varepsilon_1)^2 s\{1-s\}}{\{(1-\varepsilon_0-\varepsilon_1)s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)s-\varepsilon_0\}}$$

,or

$$\rho(Y,Y^*)^2 \to \frac{(1-\varepsilon_0-\varepsilon_1)^2 s\{1-s\}}{\{(1-\varepsilon_0-\varepsilon_1)s+\varepsilon_0\}\{1-(1-\varepsilon_0-\varepsilon_1)s-\varepsilon_0\}}$$

in probability.

Thus, it follows that

$$ARE(Y^*,Z^* \mid Y,Z) = \rho(Y,Y^*)^2 \cdot \rho(Z,Z^*)^2.$$

REFERENCES

[1] T. Yamada, N. Kinukawa, T. Nakamura, Y. Nose, "Simulation program for power and sample size determination in logistic analysis of single nucleotide polymorphisms when the response variable is subject to misclassification." *Computer Methods and Programs in Biomedicine* Volume 96, Issue 1, 42-48, 2009.

[2] R.J. Carroll, D. Ruppert, L.A. Stefanski, C.M. Crainiceanu, "Measurement Error in Nonlinear Models", Chapman and Hall/CRC, 2006.

[3] S.W.Lagakos," Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable", *Stat. Med.* 7: 257–274. 1988.

[4] S. Zacks, "Pitman efficiency". In: Armitage, P. and Colton, T., editors, Encyclopedia of biostatistics, vol. 4, Wiley, New York (1998), pp. 3380–3384.

[5] T.D. Tosteson, A.A. Tsiatis, "The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates " *Biometrika* 75(3):507-514. 1988.

[6] C.R.Rao, "Linear Statistical Inference and Its Applications", 2nd ed., Wiley Interscience, Canada, 1973, pp. 417–420.

[7] N. Ito, M. Eto, E.Nakamura, et al., "STAT3 polymorphism predicts the interferon-α response in patients with metastatic renal cell carcinoma", *J. Clin. Oncol.* 25 (19): 2785–2791, 2007.

[8] A.M. Ekström, L.B. Signorello, et al., *Evaluating gastric cancer misclassification: a potential explanation for the rise in cardiac cancer incidence*, J. Natl. Cancer Inst. 91: 786–790, 1999.

[9] K. James, E. Eisenhauer, M. Christian, M. Terenziani, D.Vena, A. Muldal, P. Therasse,, "Measuring response in solid tumors: unidimensional versus bidimensional measurement", *J. Natl. Cancer Inst.* 91: 523–528, 1999.

[10] F.Y. Hsieh, D.A. Bloch, M.D. Larson, "A simple method of sample size calculation for linear and logistic regression", *Statist. Med.*17: 1623-1634, 1998.

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)