Abstract

Genetic association tests for the heaped data

Haewon Choi

Department of Statistics

The Graduate School

Seoul National University

In self-reported surveys, subjects tend to recall the counts of events as particularly multiples of certain numbers. For example, in studies of smoking behavior cigarette counts per day are heaped such as 0, half pack, one pack, one and half packs, two packs and so forth. Because of the error of memory, the frequency of the values ending with 0 or 5 is higher than that of the true distribution. These data is called heaped data. Analysis of such heaped data has been a challenge owing to the reporting bias and the difficulty in estimating the appropriate distribution for the heaped data. Therefore, it is hard to fit a model via the standard maximum likelihood estimation when the interest lies in association studies between the heaped dependent variable and other covariates of interest. In this study, we are

interested in identifying genetic variants such as single nucleotide polymorphism (SNP) for a heaped data such as the cigarette per day (CPD). We first review previously proposed approaches applicable to CPD data in which the heaped data is treated as a dependent variable and the SNP as an ordinal independent variable. We then consider an alternative calibration modelling approach to the association test for heaped data. That is, we consider a reverse model regarding the SNP as an ordinal dependent variable and the heaped data as an independent variable. Unlike the standard modelling approach, this calibration modelling approach becomes robust to the distributional assumption of heaped data. For handling ordinal nature of SNPs, we fit a cumulative logit model in our calibration model. The significant SNPs can be identified from the model. We applied our calibration modelling approach to CPD data from Korean Association Resource project data of 4,183 male samples. Through simulation studies, we investigated performance of the proposed method and compared its performance with other competing approaches.

……

Keywords: Heaped data, Cigarette per day (CPD), Genome-Wide Association Study (GWAS), Self-reported survey

Student number: 2013-20224

Contents

List of Figures

List of Tables

1. Introduction

Over the past few years, genome-wide association studies (GWAS) have identified numerous associations between a specific single nucleotide polymorphism (SNP) and a complex trait of interest [1, 2]. The standard analysis of genome-wide association data consists of single SNP tests based on linear models [3]. For example, quantitative traits are usually analyzed by using linear regression models under the assumption that they are normally distributed with/without an appropriate transformation. For the other types of traits such as binary and count data, they are assumed to follow the binomial distribution and Poisson distribution which are members of the exponential family. Then, the generalized linear models such as logistic regression and Poisson regression models are adopted and maximum likelihood (ML) estimation is employed. However, there are some data that do not follow an exponential family distribution. To these data standard maximum likelihood estimation cannot be applied.

Cigarette per day (CPD) is a typical example to which an appropriate distributional assumption is difficult to be made. CPD tends to have higher frequencies at the values ending with 0 or 5 than standard count data, as shown in Figure 1. CPD is usually measured by self-reported studies, and thus is tended to be rounded off around the exact counts of events to particular multiples of to 0, half pack, one pack, one and half packs, two packs, and so on [4, 5, 6]. This rounding of values makes it difficult to make an appropriate distributional assumption and can result in biased estimation and imprecision of inference. Like CPD, the data that has higher frequency at certain values than the true distribution is called as heaped data. This kind of heaped data is a common type of the data to which the standard

linear modelling approach is difficult to apply. These heaped data are commonly observed in a variety of research [7, 8, 9]. Several statistical models have been proposed to deal with such heaped variables [10, 11].

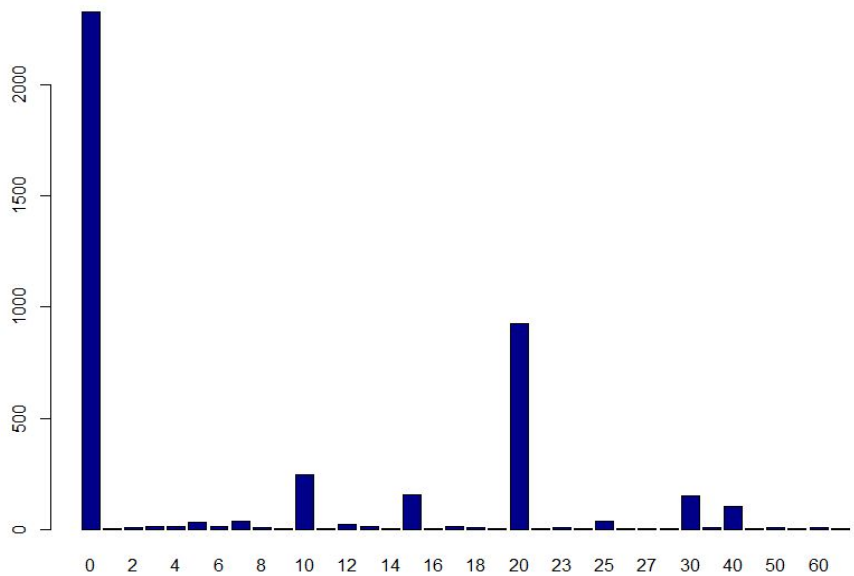

Figure 1. Frequency of the heaped data

It is a barplot of real heaped data of cigarette per day (CPD) of KARE dataset. It shows that heaped data have more frequency at several points than the true distribution of the count data.

Recently, there have been a few approaches to CPD in genetic association studies. For example, the association between each genotype and self-reported CPD was estimated using linear regression, where CPD was categorized and used as a response variable [12]. The cumulative logit model was fit to perform

association analysis for SNPs in targeted region with CPD, where a categorized CPD was used as a response variable [13]. A zero-inflated Poisson model (ZIP) [14] was used to detect significant effects of ADHD symptoms and specific genetic variants on predicting CPD [15]. The ZIP was adopted to simultaneously estimate two analytic features of smoking behaviors: the binary contrast between non-smokers and smokers, and the count construct indicating the CPD among smokers.

However, there are some limitations in these methods. Linear regression assumes that the response variable is normally distributed, but the heaped data does not follow the normal distribution. Thus, the inference based on the normality assumption does not provide a valid inference and may result in loss of power or incur false positive errors. Since the ordinal model, which is a generic term for the cumulative logit model and proportional odds model, is fit to the ordinal categorical variable, the heaped data needs to be categorized before the analysis. However, categorizing counts can not only cause loss of power, but also leads to different results depending on categorization. Zero-inflated model only assumes inflation on zero point and thus might not be suitable for CPD data with many heaping points.

The aim of this paper is to propose a new method for a genetic association test for heaped data. We suggest a new approach to overcome the limitations of the existing methods for heaped data. The unknown distribution of the heaped data makes it difficult to apply the standard association analysis. Motivated by the fact that the distribution of genotypic values is relatively well-known, we propose a new calibration model (CAM) that uses a different paradigm of models. That is, the standard association test always treats the trait of interest as a dependent variable

and the genotype as an independent variable, while our approach treats the genotype as a dependent variable and the trait of interest as an independent variable. The advantage of our CAM is its robustness to the distribution of heaped data in that it does not require a specific distributional assumption for the heaped data. We adopt a cumulative logit model as our CAM because the genotypic values are ordinal.

We first review the previous studies for the heaped data and then propose CAM. Through simulation studies, we investigate the performance of the proposed CAM and show that CAM has higher power than other existing methods. The performance of the CAM is also illustrated by a real GWA dataset of CPD from Korean Association Resource project data of 4,183 male samples.

2. Material and Methods

2.1 KARE data

Our GWA dataset was obtained from Korea Association REsource (KARE) project as a part of Korea Genome Epidemiology Study (KoGES). KARE study undertook a large scale genome-wide association studies for human complex quantitative traits among 10,038 participants [16]. Samples in the dataset were genotyped using Affymetrix Genome-Wide Human SNP Array 5.0. Genotypes were called using the BRLMM algorithm and regular quality control processes were adopted. After the quality control of SNPs and samples, 352,228 SNPs and 8,842 samples remained in the dataset. Additional procedures of the quality control were explained in [16]. Genome-wide analyses were restricted to SNPs with minor allele frequency (MAF) above 0.05 thus only 344,893 SNPs were included in our analysis.

We used cigarette per day (CPD) as a phenotype of interest and searched the SNPs associated with the CPD. Note that CPD was collected through the self-reported survey. In other words, current or former smokers were asked to report the average number of cigarettes they smoked per day. Subjects in such survey tended to round off the exact counts to half pack, one pack, one and half packs, two packs, and so on. Thus the CPD becomes a typical heaped data. Data on CPD were available on 4,183 men and on 4,659 women. Only male samples were used in this analysis because there are few female smokers in the KARE. All subjects in the dataset were 40-69 years old and were recruited from two prospective population-based studies of KoGES; Ansung and Ansan cohorts representing rural and urban communities, respectively. The CPD of KARE dataset was previously analyzed by

[13]. They used the indexed heaped CPD data: defined values for non-smoking, 1 to 10 CPD, 11 to 20 CPD, 21 to 30 CPD, larger than 30 CPD, respectively. These criteria have usually been used in studies for smoking quantity [17, 18]. Basic description about the study subjects and the CPD is shown in Table 1.

Table 1. Demographic characteristics of subject in this study

Category	**Sub-category**	**Values**
Sample size		4183
Mean age ± SD		51.78 ± 8.79
Area	Ansung	1809
	Ansan	2374
Mean cigarette per day (CPD) ± SD		8.66 ± 11.31
Indexed CPD	Non-smoking	2327
	1 to 10	375
	11 to 20	1148
	21 to 30	206
	Larger than 30	127

Values = counts for the sub-category in case of categorical variables or mean ± standard deviation in case of continuous ones.

2.2 Methods

We first briefly review existing models and then describe our proposed calibration model (CAM).

2.2.1 Regression analysis (REG)

One of the natural statistical tools for quantitative traits is a regression model. Tests in the regression model, however, require the response variable to be approximately normally distributed. The normality assumption usually does not hold in case of heaped data. When the sample size is large enough, the coefficient estimates would be approximately normally distributed. Regression models usually assume a linear relationship between the heaped data and a genotype as follows:

$$y_i = \beta_0 + \sum_{l=1}^{p} \beta_l x_{il} + \gamma SNP_{ij} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

where y_i is a heaped count for the *i*th subject, β_l is a coefficient of the *l*th covariate of x_l, SNP_{ij}, coded as 0, 1, and 2, is a genotype for the *j*th SNP in the *i*th subject and γ is the additive effect of the SNP, respectively. The indexed heaped counts can also be used as a dependent variable in (2.1) and the indexed heaped counts for *i*th subject is denoted by y_i^{index}. When y_i is switched by y_i^{index}, (2.1) is a regression model for the indexed heaped data. A genetic association test is performed by a t-test for significance of the coefficient for SNP, γ.

2.2.2 Ordinal model (OM)

For the indexed heaped data, we can fit a cumulative logit model or a proportional odds model by treating it as an ordinal response variable. The cumulative logit model for the indexed heaped count is expressed by

$$log\frac{p(y_i^{index} \le k)}{1-p(y_i^{index} \le k)} = \beta_{0k} + \sum_{l=1}^{p} \beta_{lk} x_{il} + \gamma_k SNP_{ij} + \varepsilon \qquad (2.2)$$
$$, k = 1, 2, \dots, K-1$$

When there are *K* categories, the cumulative logit model is defined for the *K-1* cumulative logits. Each of the *K-1* cumulative logits has its own parameter, γ_k. Therefore, when testing an association between a SNP and a heaped data, the *K-1* parameters need to be considered simultaneously. However, if all γ_k are the same, then one common parameter can be tested in a much more efficient manner and result in an increase of power and simple interpretation in terms of odds ratio. Depending on a result of a test for the homogeneity of parameters, the association test can be performed either by *K-1* parameters or one parameter. This homogeneity test and association tests can be done by the likelihood ratio tests.

2.2.3 Zero-inflated Poisson model (ZIP)

Some heaped data have a distribution especially concentrated on zero. There is a model for count data dealt with the excess of zeroes, which is zero-inflated Poisson model (ZIP) [14]. ZIP assumes that the counts are from two processes. One process is for the zeroes, combining a probability of zero in Poisson distribution and that of extra zero; the other process is for non-zero counts using zero-truncated Poisson model. The zero-inflated Poisson model can be expressed by

$$\mathrm{P}(y_i|\boldsymbol{x_i}, \boldsymbol{z_i}) = \begin{cases} p(\boldsymbol{z_i}) + (1 - p(\boldsymbol{z_i}))e^{-\lambda_i} & if\ y = 0 \\ (1 - p(\boldsymbol{z_i}))\dfrac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!} & if\ y > 0 \end{cases} \quad (2.3)$$

where $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ are covariate vectors of the *i*th subject for the zero-truncated model and the extra zero, respectively. $p(\boldsymbol{z_i})$ gives the extra probability thrust at the zero and λ_i represents the mean of all counts except for the excess zero. We can model $p(\boldsymbol{z_i})$ and λ_i using logit and exponential function as follows:

$$p(\boldsymbol{z_i}) = \frac{\exp(\beta_0' + \sum_{l=1}^{p} \beta_l' z_{il} + \gamma' SNP_{ij})}{1 + \exp(\beta_0' + \sum_{l=1}^{p} \beta_l' z_{il} + \gamma' SNP_{ij})},$$

$$\lambda_i = \exp(\beta_0 + \sum_{l=1}^{p} \beta_l x_{il} + \gamma SNP_{ij})$$

where γ' and γ represent the additive effects of each genetic variant on non-zero and zero counts, respectively. If a genetic variant has a significant association with either the binary contrast between zero and non-zero, or counts of the non-zero, the variant would be considered as significant. Thus, significance of γ' and γ has to be simultaneously tested via LRT which asymptotically follows chi-square distribution with 2 degrees of freedom under the null hypothesis.

2.3 Proposed calibration model (CAM)

The regression model, the ordinal model and the zero-inflated model have been used to test the association between the heaped data and a genetic variant. However, each method has its own drawbacks. The heaped data rarely follows the normal distribution, and thus the regression analysis provides invalid test results. The

ordinal model based on the cumulative logits requires the heaped data to be categorized first and then uses the categorized heaped data as a dependent variable. Depending on categorization, the ordinal model may provide different results. Finally, while ZIP model deal with excess of zeroes well, it is doubtful whether the ZIP model can handle several heaping points.

We propose a new calibration model (CAM) approach which can overcome these drawbacks of the existing methods. In the analysis of heaped count data, it is challenging to find an appropriate distribution upon which the likelihood approach is based. However, if we treat the genotypic value as a dependent variable and a heaped data as an independent variable, we can test the association between a SNP and heaped data without the distributional assumption of heaped data. This can be achieved because distribution of genotypic values is well-known. Therefore we propose a CAM approach which switches an ordinary dependent with an independent variable in the standard regression model.

For handling ordinal nature of SNP, we consider the following cumulative logit model. Then the proposed CAM can be expressed as follows:

$$log\frac{p(SNP_{ij} \le k)}{1-p(SNP_{ij} \le k)} = \beta_{0k} + \sum_{l=1}^{p} \beta_{lk} x_{il} + \gamma_k y_i + \varepsilon, \mathrm{k} = 0, 1 \qquad (2.4)$$

The CAM is defined for two cumulative logits where each cumulative logit has its own association parameter. These two association parameters γ_1 and γ_2 are required to analyze at the same time in order to identify a significant association. When γ_1 and γ_2 are the same, it is much more efficient to test the association using the one common parameter, $\gamma(=\gamma_1=\gamma_2)$. The test for homogeneity of γ_k can be performed by the LRT which compares the fit of (2.4) with and without the

homogeneity assumption, respectively. If the homogeneity can be assumed, the association can be tested by using the common parameter, γ. Otherwise, γ_1 and γ_2 would be tested simultaneously by using the LRT with 2 degrees of freedom. If y_i is switched by y_i^{index}, (2.4) is a CAM for the indexed heaped count. Genetic association test procedure for indexed heaped count is the same as that for the original heaped data.

We might obtain a sparse data with genotype values consisting of mainly 0's and 1's, not 2, even if the dataset filtered using MAF < 0.05, then the estimates or likelihood of the calibration model can be infinite. In that case, we assumed dominant effect of the minor allele, that is, combined 2 with 1 and then fitted a logistic model equal to (2.4) when k = 0. The association tests can be also done by the likelihood ratio tests in the logistic model.

3. Results

3.1 Simulation

Simulation studies were executed to demonstrate validity and performance of our calibration model (CAM) and to compare it with other methods: REG, OM and ZIP. We first assessed whether CAM and other methods have appropriate type 1 error by simulation data set under the null hypothesis of no association. Since the distribution of heaped variable is unknown, the individual heaped counts and covariates were borrowed from a real data set, the CPD of KARE. To make the simulation data set under the null hypothesis, individual genotypic values were generated by random sampling of binomial distribution $B(n, p)$; n is the number of samples in KARE data set and p is MAF. The genotype was independently generated from the real heaped data and thus this simulation data set is under the null hypothesis.

To calculate the type 1 error, the simulation data set contained 1,000 non-causal SNPs with no linkage disequilibrium (LD) as we noted above. After testing associations between heaped data and genotype in the simulation data set, the type 1 error were calculated as a proportion of the false positive. The results of simulation found that REG for the original heaped data, REG for an indexed heaped data, OM, CAM for the original heaped data and CAM for an indexed heaped data successfully controlled the false-positive rate at a significance level, 0.05, because the confidence interval for type 1 error of each method contains the significance level. Figure 2. shows that the type 1 error is well controlled. However, the type 1 error of ZIP was much higher than the significance level and could not control type 1 error. It indicates that the ZIP is not appropriate for data with several

heaping points.

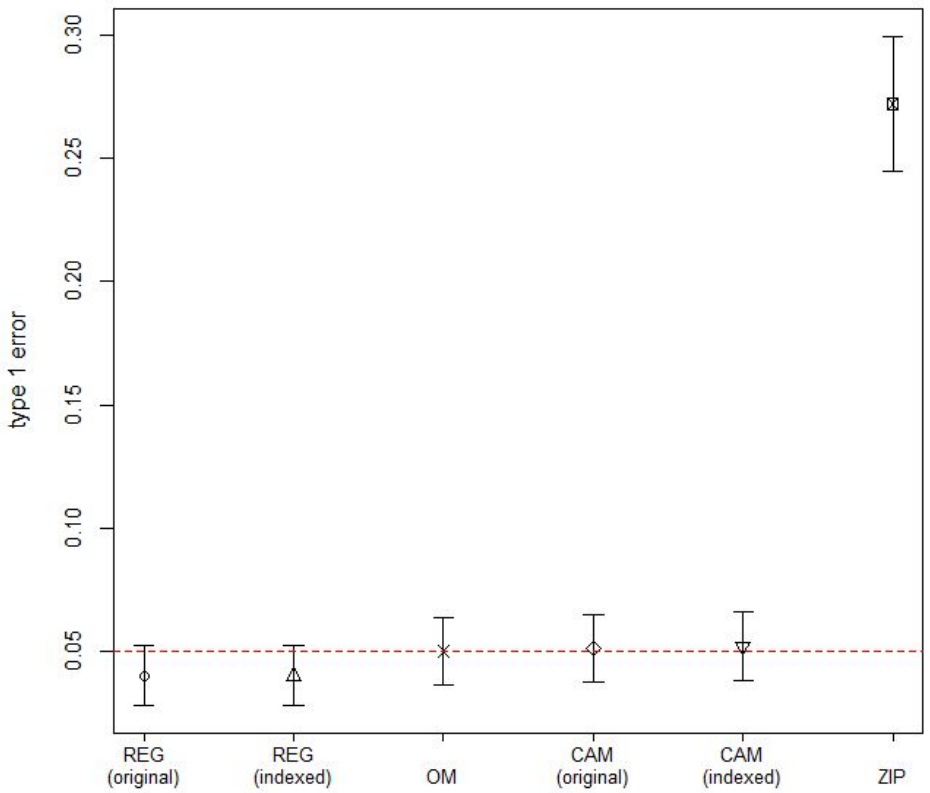

Figure 2. Type 1 error of five methods genetic association test for heaped data

It shows that the zero-inflated model (ZIP) does not control type 1 error and the other methods controlled the type 1 error. Significance level was set at 0.05.

Another simulation study was performed to compare power of our CAM to that of the other methods. The ZIP was excluded from the comparison because it failed to control type 1 error. There was no association between simulation genotype and

heaped data in the simulation for calculating type 1 error, but causal variants are needed to calculate power. Thus, a method was devised to make an association between the independent genotype and heaped data, preserving distribution of the heaped data. The first step of the method is sorting a genotype generated from *B(n, p)* and the heaped data of KARE in ascending order, respectively. Then, there must be an association between the sorted genotype and heaped data. In this setting, however, the sorted genotype would have almost fixed effect size when MAF is the same. In order to simulate various scenarios of effect sizes, some of individual-genotypic values were randomly selected and then were shuffled. On the other hands, the heaped data was left sorted in ascending order. Here, an effect size was estimated by mean correlation between the sorted heaped data and permuted genotype when the number of the permuted elements is the same.

The power of each method was calculated under the several scenarios of its effect size and MAF, by making data set containing 1,000 causal SNPs with no LD as stated above and testing association between the sorted heaped data and permuted genotype. Power was calculated as proportion of the true positive genotypes. The simulation results showed that CAM for indexed heaped data have the highest power regardless of MAF as shown in Figure 3. CAM for the original heaped data was next in power, followed by REG for indexed heaped data, OM and REG for the original heaped data. REG for the original heaped data suffered the largest loss of power and it is attributed to wrong distribution assumption.

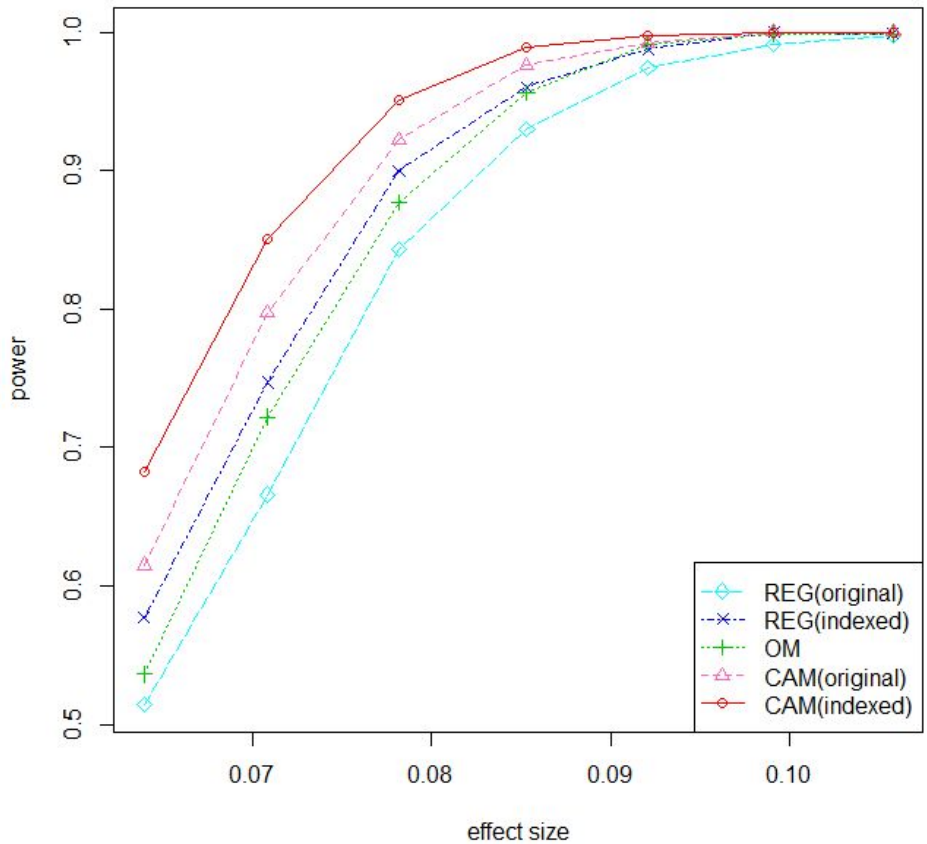

Figure 3. Power of five methods genetic association test for heaped data

The calibration model for indexed heaped data has higher power than any other models. Significance level was set at 0.05.

3.2 Real data analysis

To evaluate the efficiency of method by means of a real data set, CAM and other methods were applied to the KARE data set. Using R software, we performed the real data analysis. The distribution of the CPD of KARE is shown in Figure 1 and its frequency is much higher when CPD has the value of 0, 10, 15, 20, 25 and 30 than that of CPD when it has the other values. It shows that the CPD of KARE is a typical heaped data. Thus, we demonstrated the performance of our CAM for a real genetic data set by applying CAM to the CPD of KARE.

Single SNP tests based REG, OM and CAM were performed for all SNPs in the KARE data set. Individual age and area information were used as covariates in this analysis. The CAM for the data set is given as follows:

$$log\frac{p(SNP_{ij} \le k)}{1-p(SNP_{ij} \le k)} = \beta_{0k} + \beta_{1k}AGE_i + \beta_{2k}AREA_i + \gamma_k CPD_i, \mathrm{k} = 0, 1 \quad (3.1)$$

where AGE_i, $AREA_i$, CPD_i are age, area, CPD for *i*th subject. $AREA_i$ was coded Ansung as 0 and Ansan as 1. Significant associations between a SNP and the CPD was using LRT at significance level 1×10^{-5}. If the CPD is switched by the indexed CPD in (3.1), it becomes CAM for indexed heaped data. REG and OM were also applied to the KARE in the manner described in the Method part.

Table 2. *P* values for SNPs significantly associated with smoking quantity

chrom osome	dbSNP ID	MAF	*P* value				
			REG original	REG indexed	OM	CAM original	CAM. indexed
4	rs3817031	0.1427	**5.80E-06**	**9.75E-06**	1.18E-05	**5.16E-06**	**7.08E-06**
4	rs17515274	0.245	3.54E-02	5.67E-02	**5.01E-06**	1.59E-02	3.03E-02
4	rs9998218	0.2473	2.45E-02	3.92E-02	**6.67E-06**	7.01E-03	1.90E-02
7	rs1404697	0.1283	6.08E-05	1.43E-05	5.22E-04	4.09E-05	**8.36E-06**
8	rs11779225	0.2439	2.36E-05	2.46E-05	**8.90E-06**	1.33E-05	1.30E-05
11	rs12574551	0.1605	6.74E-01	4.41E-01	4.77E-01	**5.79E-06**	**7.57E-06**
11	rs17134231	0.1606	7.18E-01	4.80E-01	5.06E-01	**7.09E-06**	**9.79E-06**
11	rs12282340	0.1608	6.60E-01	4.38E-01	5.14E-01	**5.33E-06**	**7.47E-06**
11	rs12281880	0.1605	6.51E-01	4.33E-01	4.86E-01	**7.87E-06**	1.13E-05
11	rs10160335	0.1607	6.91E-01	4.64E-01	4.95E-01	1.62E-05	**7.89E-06**
13	rs17472526	0.1332	**1.67E-06**	**3.64E-06**	5.86E-05	**2.60E-06**	**5.49E-06**

Significance associations at the 10^{-5} level are expressed in bold. Age and Area are used as covariates in each model.

Each method detected significant associations between a SNP and the CPD. REG for the original CPD and for the indexed CPD detected 2 candidate SNPs. OM found 3 candidate SNPs. CAM for the original CPD detected 6 candidate SNPs and CAM for the indexed CPD found 7 significant SNPs that include 2 SNPs detected by REG and 6 SNPs detected by CAM for the original CPD. These results can also be indicated in Table 2 and Manhattan plots [19] in Fig 4. The variants, rs1404697 that was undetected in other methods has been reported to be significantly associated with smoking behavior, namely nicotine dependence [20].

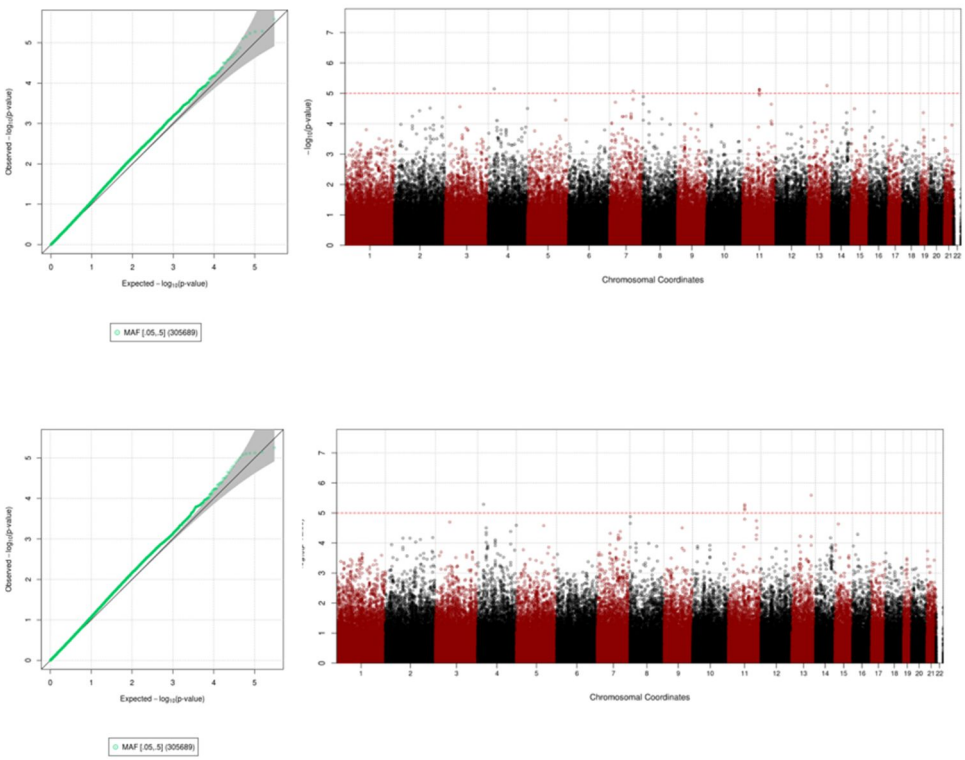

Figure 4. QQ-plot and Manhattan plots for the CPD of KARE dataset when calibration model was used

The first is plots from proposed calibration model with the original CPD value and the second is plots from proposed calibration model with indexed heaped data. Horizontal red line in Manhattan plots represents the threshold 1×10^{-5} for selecting significant SNPs.

Pairwise scatterplots in Figure 5 shows the comparison results of the methods. According to Figure 5, there was little difference between using the original and indexed CPD in the REG and the CAM. Some variants had marked difference between the CAM and the REG, but the difference may has been resulted from low power of REG, because the most of the variants had smaller *P* values in the CAM

than in the REG. However, there were variants that had higher *P* value in the OM than in the CAM. This situation is not resulted from categorization, because in such variants the results of OM differed from that of other methods using the same indexed CPD. There might be a certain type of variants that OM found variants that have the smaller *P* value than other methods.

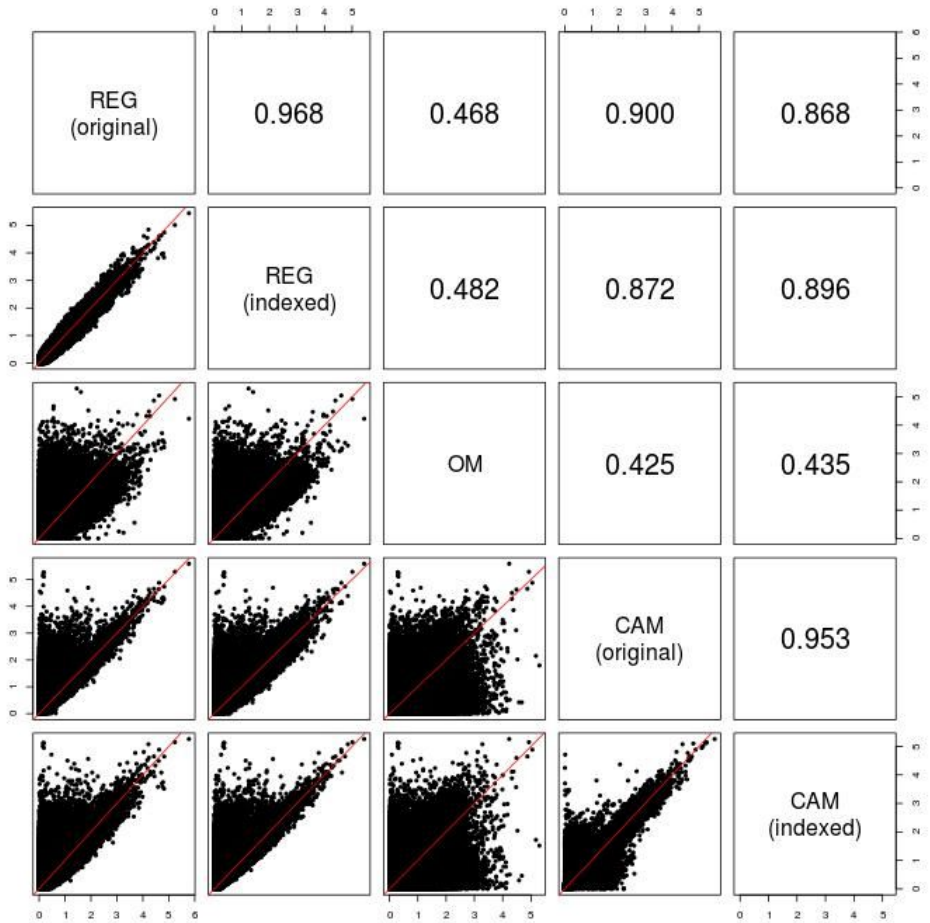

Figure 5. Pairwise scatterplots for each pairs of $-\log_{10}P$

The *P* values are calculated by each method for genetic association test for heaped data. Generally, the CAM has lower *P* value than other methods in most of variants. However, there were variants that had higher *P* value in the OM than in the other methods. The values in upper-triangle of the scatterplot matrix are correlation of between the *P* values.

It is an analyst that determines definition of categorization when an indexed heaped variable is used in analysis of heaped data. Therefore results of analysis can be changed by an analyst's arbitrary decision. Thus, it is a critical issue whether methods using an indexed heaped data are robust to definition of categorization or not. To ascertain the robustness, results from 6 definitions of categorization for an indexed CPD were compared when association tests was performed using REG, OM, and CAM for an indexed CPD, respectively. The six definitions have been used in previously published papers and frequency of indexed CPD when each definition was used is indicated in Table 3. As a result, average pairwise correlations of P values between the definitions of categorization was 0.844 in REG for an indexed CPD, followed by 0.538 in OM and 0.835 in CAM for the indexed CPD. It shows that REG and CAM for an indexed CPD are robust to definition of categorization but OM can alter the analysis results depending on the category.

Table 3. Frequencies of CPD in KARE depending on six definitions of categorization

1st indexed CPD		**2nd indexed CPD**		**3rd indexed CPD**	
Non-smoking	2327	0 to 10	2702	Non-smoking	2327
1 to 10	375	11 to 20	1148	1 to 19	596
11 to 20	1148	21 to 30	206	Larger than 19	1260
21 to 30	206	Larger than 30	127		
Larger than 30	127				
4th indexed CPD		**5th indexed CPD**		**6th indexed CPD**	
Non-smoking	2327	Non-smoking	2327	Non-smoking	2327
1 to 9	129	1 to 10	375	1 to 15	568
10 to 19	467	11 to 15	193	16 to 25	1005
Larger than 19	1260	16 to 19	28	Larger than 25	283
		20 to 25	977		
		Larger than 25	283		

4. Discussion

The purpose of this paper is to propose a method for genetic association test for heaped data. There are some methods used to test an association between a SNP and a heaped data, but they have several drawbacks. To overcome the drawbacks, a calibration model (CAM) is proposed, which treats genotype as a dependent variable and a heaped data as an independent variable. Since a statistical model does not require a specific distributional assumption for the independent variable, an assumption for distribution of a heaped data does not have to be made in using CAM and thus the proposed CAM takes the advantage of robustness to the distribution of heaped data.

Our study demonstrated that CAM achieved substantial increase in statistical power to detect true associations. A simulation study showed that our CAM for an indexed heaped data has higher power than any other methods that control type 1 error. Better performance of our CAM was also validated via our real data analysis. That is, only CAM for the indexed heaped data detected the SNP that has been reported to the association with smoking behavior. In addition, it showed that CAM and REG are robust to definition of categorization unlike OM.

However, our CAM has some limitations. First, the CAM has infinite estimates or likelihood in some cases. It is mainly because a sparse data can be obtained with genotype values consisting of mainly 1's and 2's, not 0, even if the data set is filtered using a criteria for MAF. In that case, a logistic model is fitted after combining 2 with 1, but there might be better methods for the sparseness. Second, there may be a certain type of variants that OM detected better than the CAM. Thus, OM and CAM need to be considered together in testing an association between a

SNP and a heaped data.

In summary, a new modelling approach, CAM, to the association test for heaped data is proposed. Our study confirms that the proposed CAM has better performance than other accessible methods for testing an association between a SNP and a heaped data. Therefore, more candidate SNPs can be found by our CAM. Furthermore, CAM takes another advantage of easy implementation, because CAM can use existing software for a cumulative logit model. As a result, CAM is a simple and efficient method to reveal novel association between a SNP and a heaped data.

Bibliography

1. J. Hardy and A. Singleton (2009). Genomewide Association Studies and Human Disease. N Engl J Med. 2009 Apr 23;360(17):1759-68.
2. TA. Manolio, LD. Brooks, FS. Collins (2008). A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008 May;118(5):1590-605.
3. W.S. Bush and J.H. Moore (2012) Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol. Dec 2012; 8(12): e1002822.
4. B. Means, K. Habina, GE. Swan, L. Jack (1992). Cognitive research on response error in survey questions on smoking. Meryland, USA: National Center for Health Statistics, Vital Health stat 6(5).
5. RC. Klesges, M. Debon, JW. Ray (1995). Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. J Clin Epidemiol, 48, 1225-33.
6. International Agency for Research on Cancer (IARC) Handbooks of Cancer Prevention (2008). Tobacco Control Vol. 12: Methods for Evaluating Tobacco Control Policies. Lyon, France: IARC, World Health Organization.
7. R. J. Myers (1976). An instance of reverse heaping of ages. Demography 13 577-580.
8. H Schneeweiss and J Komlos (2009). Probabilistic rounding and Sheppards correction. Statistical Methodology 6 577–593.
9. J Huttenlocher, L. V. Hedges and N. M. Bradburn (1990). Reports of elapsed time: bounding and rounding processes in estimation. Journal of Experimental Psychology: Learning, Memory, and Cognition 16 196-213.

10. L Farrell, T R. L. Fry, M N. Harris (2011) 'A pack a day for 20 years': smoking and cigarette pack sizes. Applied Economics, 43:21, 2833-2842.
11. H Schneeweiss, J Komlos and A. S. Ahmad (2010). Symmetric and asymmetric rounding: a review and some new results. AStA Advances in Statistical Analysis 94 247-271.
12. M.R. Munafò, M.N. Timofeeva, R.W. Morris, D. Prieto-Merino et al. (2012). Association Between Genetic Variants on Chromosome 15q25 Locus and Objective Measures of Tobacco Exposure. J Natl Cancer Inst. 2012 May 16; 104(10): 740–748.
13. MD. Li, D. Yoon, JY Lee, BG han et al. (2010). Associations of Variants in CHRNA5/A3/B4 Gene Cluster with Smoking Behaviors in a Korean Population. PLoS One. 2010; 5(8): e12183.
14. D. Lambert (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics. 1992;34:1–14.
15. CT Lee, BF Fuemmeler, FJ McClernon, A Ashley-Koch et al. (2013). Nicotinic receptor gene variants interact with attention deficient hyperactive disorder symptoms to predict smoking trajectories from early adolescence to adulthood. Addict Behav. 2013 Nov;38(11):2683-9.
16. YS. Cho, MJ. Go, YJ. Kim, JY Heo et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Human genetics. 2012, 131:1009-1021.
17. JP Rice, SM Hartz, A Agrawal, L Almasy et al. (2012). CHRNB3 is more strongly associated with Fagerström test for cigarette dependence-based nicotine dependence than cigarettes per day: phenotype definition changes

genome-wide association studies results. Addiction. 2012 Nov;107(11):2019-28.

18. L Greenbaum1 and B Lerer (2009). Differential contribution of genetic variation in multiple brain nicotinic cholinergic receptors to nicotine dependence: recent progress and emerging open questions. Mol Psychiatry. 2009 Oct;14(10):912-45.

19. G. Gibson (2010). Hints of hidden heritability in GWAS. Nature Genetics 42 (7): 558–560.

20. D. Yoon, YJ Kim, WY Cui, Van der Vaart A et al. Large-scale genome-wide association study of Asian population reveals genetic factors in FRMD4A and other loci influencing smoking initiation and nicotine dependence. Hum Genet. 2012 Jun;131(6):1009-21.

국문초록

사건의 횟수나 값을 숫자로 기입하는 자기기입식 설문문항에서 응답자들은 실제 횟수나 값을 기억하기 쉬운 어떤 숫자로 기억하는 경향이 있다. 예를 들어, 흡연에 대한 연구에서 하루 평균 흡연량(CPD)의 빈도는 반 갑, 한 갑, 두 갑 등에 해당되는 숫자에서 현저히 높게 나타난다. 이처럼 응답자들의 기억오류로 인해, 0이나 5로 끝나는 값들의 빈도가 다른 값들에 비해 현저히 높게 나타나는 데이터를 heaped data라고 부른다. heaped data를 분석할 때, 잘못된 답변으로 인해 편향이 발생할 수 있기 때문에 heaped data의 적절한 분포를 추정하기가 어렵고 특히 heaped data가 반응변수인 경우에는 일반적인 최대가능도추정법을 이용하여 모형을 적합하는 데 어려움이 있다. 따라서 본 연구에서는 하루 평균 흡연량과 같은 heaped data와 연관성이 있는 유전 변이를 찾는 데에 사용할 수 있는 새로운 calibration 모형화 방법을 제시하고자 한다. 우선 기존의 heaped data의 유전자 단위 연관 분석에 사용되는 방법들에 대해 간략히 서술하고, 우리가 새롭게 제시하는 calibration 모형화 방법을 소개한다. calibration 모형화 방법은 기존의 방법들과 달리 heaped data을 독립변수로, 유전 변이를 종속변수로 하여 연관성을 검정하는 모형화 방법이다. 반응변수인 단일 유전자 변이(SNP)가 순서형 변수이므로 누적로짓모형을 이용한다. 이 모형은 heaped data의 분포에 대한 가정 없이 모형을 적합할 수 있으므로 heaped data의 분포에 대해 로버스트하다는 장

점을 가지고 있다. 또한 모의실험을 통해서 우리의 calibration 방법의 성능이 기존의 다른 방법들에 비해 좋음을 확인하였고 이를 Korean Association Resource(KARE) 프로젝트 데이터의 흡연 데이터에 적용하여 흡연에 유의한 영향을 준다고 알려져 있는 SNP을 찾아낼 수 있었다.

.....

주요어 : heaped data, 하루 평균 흡연량(CPD), 유전체 전장 연관성 연구, 자기기입식 설문문항

학번 : 2013-20224