Communications for Statistical Applications and Methods
2018, Vol. 25, No. 5, 545–557

https://doi.org/10.29220/CSAM.2018.25.5.545
Print ISSN 2287-7843 / Online ISSN 2383-4757

A sample size calibration approach for the *p*-value problem in huge samples

Yousung Park[a], Saebom Jeon[b], Tae Yeon Kwon[1,c]

[a]Department of Statistics, Korea University, Korea;
[b]Department of Marketing Information Consulting, Mokwon University, Korea;
[c]Department of International Finance, Hankuk University of Foreign Studies University, Korea

Abstract

The inclusion of covariates in the model often affects not only the estimates of meaningful variables of interest but also its statistical significance. Such gap between statistical and subject-matter significance is a critical issue in huge sample studies. A popular huge sample study, the sample cohort data from Korean National Health Insurance Service, showed such gap of significance in the inference for the effect of obesity on cause of mortality, requiring careful consideration. In this regard, this paper proposes a sample size calibration method based on a Monte Carlo t (or z)-test approach without Monte Carlo simulation, and also proposes a test procedure for subject-matter significance using this calibration method in order to complement the deflated p-value in the huge sample size. Our calibration method shows no subject-matter significance of the obesity paradox regardless of race, sex, and age groups, unlike traditional statistical suggestions based on p-values.

Keywords: huge sample, p-value problem, subject-matter significance, Monte Carlo, sample size calibration

1. Introduction

The Korean National Health Insurance Service released a sample cohort database Lee *et al.* (2016) that includes more than one million people from 2002 to 2013. Using this cohort data, we evaluated the effect of waist circumference (WC) and body mass index (BMI) on mortality, respectively. The Cox's proportional hazard regression model, with age as a covariate, showed that the hazard ratio of WC significantly decreased as WC increased; however, the hazard ratio significantly increased in the model without age. The model with age showed a U-shaped effect of BMI on mortality but the model without age revealed a significant decreasing effect of BMI as BMI increased, implying no obesity paradox by the first model, but an obvious obesity paradox in the second model.

The natural questions are as follows. Why do these contradictory results arise from huge samples? How do we identify which model reveals the correct effect of WC or BMI? In this paper, as an answer for the second question, we propose a new testing approach to have a consistent result that is not sensitive to the inclusion or omission of some covariates in the model.

One may answer the first question with the omitted variable bias (lurking variable bias) or multicolinearity problem. However, we may not distinguish whether bias occurs due to omitting a significant covariate or adding an unnecessary variable. Moreover, we cannot filter these biases with the significance test especially in a huge sample, because statistical inference based on p-value depends on

[1] Corresponding author: Department of International Finance, Hankuk University of Foreign Studies, 81 Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: tykwon@hufs.ac.kr

Published 30 September 2018 / journal homepage: http://csam.or.kr

variance, and the variance then decreases as the sample size increases. For a huge sample, the resulting variance and p-value are small enough to reject the null hypothesis even when a consistent estimate is negligibly different from the value specified in the null. Such a deflated p-value associated with the huge sample in statistical inference has been criticized by showing that the p-value approaches to 1 or 0 when the estimate approaches the value specified in the null hypothesis (or not), respectively (Johnson, 1999; Lin *et al.*, 2013; Halsey *et al.*, 2015; Wellek, 2017).

The deflated p-value problem can cause controversy between *statistical significance* and *subject-matter significance*. Different from the statistical significance determined simply by p-value, subject-matter significance is concerned with the interpretation and meaningfulness of the results in the real world. Such controversy has been discussed in several studies using the terms of practical significance (Kirk, 1996; Harlow *et al.*, 2016) and analytical significance (Altman, 2004) which are similar to the subject-matter significance (Johnson, 1999).

Some suggestions for the huge sample problem have been to make the alpha level much smaller than the traditional 0.01 or 0.05 (Leamer, 1978; DeGroot and Schervish, 2002; Greene, 2003) and to replace the single value in the null hypothesis by an interval (DeGroot and Schervish, 2002; Hubbard and Armstrong, 2006). A standardized p-value by a small sample size was also suggested to resolve the deflated p-value (Good, 1980, 1982). Bayesian approaches such as the conditional probability or the p-value calibration (Sellke *et al.*, 2001; Bayarri and Berger, 1999, 2000) were proposed but suffered from the same p-value problem related to the huge sample size (Bayarri and Berger, 2000; Gelman *et al.*, 1996). Some actions in huge sample studies were suggested to mitigate the p-value problem (Ghose *et al.*, 2006; Ghose and Yang, 2009); however, these were not developed for use in practical applications and cannot be a solution for the model selection in our huge sample.

As a solution for the p-value problem of a huge sample, we propose a Monte Carlo simulation approach to calculate sample size calibrated t (or z)-values which are used to provide a test procedure for a subject-matter significant test. The Monte Carlo simulation, however, is not needed because the average of sample size calibrated t (or z)-values converges to a straight line over the square root of the sample size with the slope, called a calibration factor, which is easily obtained from the original huge sample. The calibration factor approaches for resolving the problem in statistical inference due to sample size are also discussed in Tsao (2001, 2004), Emerson (2009).

The sample size calibration test procedure we propose for subject-matter significance is now applied to five sets of real data. The first data set is the Korea sample cohort data from which we estimate the effects of WC and BMI on mortality to show that the results vary in statistical significance depending on inclusion or omission of some covariates in the model. Three sample size calibrated t-values, however, provide consistent significant results regardless of inclusion or omission of covariates in the model.

The remaining four datasets are actually unknown except for the estimation results in literature for the effects of BMI on all causes of mortality. Their sample sizes range from 0.15 to 1.46 million and study subjects are quite different: White adults (Berrington de Gonzalez *et al.*, 2010), 50–71 years aged Americans (Adams *et al.*, 2006), East Asians, Indians and Bangladeshis (Zheng *et al.*, 2011), and Koreans (Kim *et al.*, 2015). Our new test procedure can be applied to these four datasets as the procedure requires only total sample size, total number of deaths, and estimates and their standard error. Our sample size calibration test procedure suggests no subject-matter significance of the obesity paradox regardless of race, sex, and age, unlike traditional statistical suggestions based on p-values.

The rest of the paper is organized as follows. In Section 2, we investigate the reason why contradictory results in a huge sample arise between the regression models that include or exclude relevant covariates and illustrate p-value problems using a simple simulation. In Section 3, we introduce a sample size calibration test procedure with the calibration factor. In Section 4, we apply the new test procedure to five sets of real data to compare the traditional statistical significance with subject-matter significance based on our test procedure. Some concluding remarks are included in Section 5.

2. *P*-value problem in linear models

Consider the following two regression models to investigate the p-value problem arising from huge samples. For $i = 1, \dots, n$,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \text{and} \quad y_i = \alpha_0 + \beta_{10} x_{1i} + \epsilon_{1i}, \tag{2.1}$$

where all errors are assumed to be iid. When there are more explanatory variables other than x_1 and x_2, we assume that they are partialled-out to equally estimate the regression coefficients in the two models given in (2.1) (Qian and Schmidt, 2003).

Denote $r_{z_1 z_2}$ to be the sample correlation of z_1 and z_2. We then have the following relationship between the ordinary least squares (OLS) estimators of β_1 and β_{10}.

Lemma 1. *The OLS estimates of* β_1 *and* β_{10} *in (2.1) satisfy*

$$\hat{\beta}_1 = \left(1 - r_{x_1 x_2}^2\right)^{-1} \hat{\beta}_{10} \left(1 - r_{x_1 x_2} \frac{r_{x_2 y}}{r_{x_1 y}}\right).$$

Lemma 1 shows the well-known fact that the OLS coefficient of x_1 in the multiple regression model is the same as the coefficient from the simple regression model when x_1 and x_2 are uncorrelated. When $r_{x_1 x_2} \neq 0$ and $r_{x_2 y} = 0$ in Lemma 1, the true model is the simple regression model of (2.1) (i.e., $\beta_1 = 0$) and the multiple regression model produces $\hat{\beta}_1$ which is close to zero but might be significant because of large sample size as discussed in Example 1. The multiple regression model produces $\hat{\beta}_1 = (1 - r_{x_1 x_2}^2)^{-1} \hat{\beta}_{10}$ where s_y and s_{x_1} are the standard deviations of y and x_1, respectively. However when $r_{x_1 x_2} \neq 0$ and $r_{x_2 y} \neq 0$, the multiple regression model of (2.1) is true (i.e., $\beta_1 \neq 0$). The simple regression model produces a biased coefficient of x_1 because x_{1i} and ϵ_{1i} are correlated, which violates the basic requirement of OLS estimation (Fomby *et al.*, 1984; Johnston, 1984).

Example 1. (Regression coefficients in simple vs multiple regression)

We generate y_i, x_{1i}, and x_{2i} from a multivariate normal distribution with zero means and unit variances along with correlations, $\rho_{x_1 x_2} = 0.5$, $\rho_{x_1 y} = 0.05$, and $\rho_{x_2 y} = 0.3$. This is equivalent to $y_i = -0.133 x_{1i} + 0.367 x_{2i} + \epsilon_i$ with ϵ_i from iid $N(0, 1)$ where $\rho_{x_1 x_2} = 0.5$. Observe that although $\rho_{x_1 y}$ is close to zero (equivalently, $\beta_{10} = 0.05$ in a simple regression $y_i = \beta_{10} x_{1i} + \epsilon_{1i}$), the coefficient of x_1 is -0.133 which is large in absolute value and the opposite sign to β_{10} because of a much larger correlation between y and x_2 than between y and x_1. We repeat this sample generation 1,000 times for each sample size of 30 to 100,000 and depict p-values along with sample size (Figure 1).

Figure 1 shows that $\hat{\beta}_{10}$ and $\hat{\beta}_1$ are significantly different from zero for the sample size more than 3,000, and their p-values and upper 2.5% percentiles are close to zero from the sample size of 10,000. Due to this deflated p-value, the $\hat{\beta}_{10}$ obtained from the wrong model $y_i = \beta_{10} x_{1i} + \epsilon_{1i}$ is strongly significant and leads to the wrong positive association between x_1 and y as long as the sample size is more than 10,000. However, the significant β_{10} might not be acceptable in real application as the coefficient of determination in $y_i = 0.05 x_{1i} + \epsilon_{1i}$ is only $R^2 = 0.0025$. This requires an evaluation method whether or not $\hat{\beta}_{10}$ estimated by using a huge sample (e.g., $n = 100{,}000$) is still an effective magnitude in a much smaller sample (e.g., $n = 100$) to comply with a subject-matter significance.

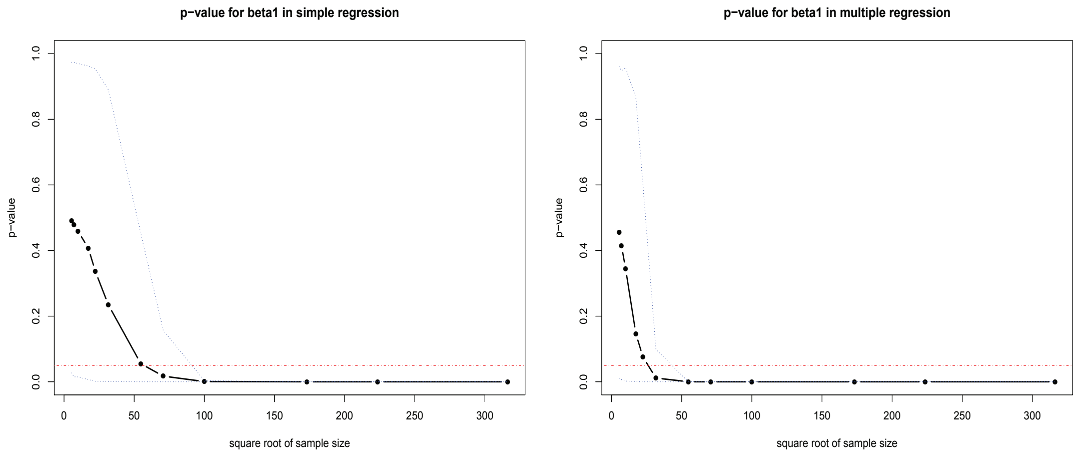

Figure 1: *Mean and upper and lower 2.5% of* p*-values for* β_{10} *and* β_1*.*

All of the discussions related to the deflated p-value problem arising from a huge sample can be easily extended to the generalized regression model with heteroscedastic and autocorrelated disturbance. Both the OLS estimator given in Lemma 2.1 and the general least squares estimator (GLS) for the generalized regression model are consistent under some regular conditions (Greene, 2003).

The deflated p-value problem in the generalized regression model also occurs in the generalized linear model. Let $y_1, y_2, \ldots, y_n$ be independent Poisson observations with respective means μ_i that are linked by $\log \mu_i = \mathbf{x}_i'\beta$ where $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ is explanatory variables for the i^{th} observation. The maximum likelihood (ML) estimates of β (Pawitan, 2001) can be found by iterating the following steps. At step k, update $\mu_i^{(k)} = \exp(x_i'\beta^{(k-1)})$ and then update the parameter β to step k by $\beta^{(k)} = (\sum_{i=1}^n \mu_i^{(k)} x_i x_i')^{-1} \sum_{i=1}^n \mu_i^{(k)} x_i w_i^{(k)}$ where $w_i^{(k)} = x_i'\beta^{(k-1)} + (y_i - \mu_i^{(k)})/\mu_i^{(k)}$. This is the GLS for a generalized regression model of modified w_i with variance μ_i^{-1} on regressors x_i.

The likelihood function for the proportional hazard model and the partial likelihood function for the Cox's proportional hazard model are the same as the Poisson regression model by letting $\mu_i = t_i e^{x_i'\beta}$ for the proportional hazard model (Bolstad, 2009) and $\mu_i = \exp(x_i'\beta)/\sum_{j=1}^n y_j \exp(x_j'\beta)$ for the Cox's proportional hazard model (Whitehead, 1980) where t_i is the survival time of the i^{th} individual, $y_i = 1$ if t_i is the time of death and 0, otherwise, and $y_j = 1$ if $t_j \geq t_i$ and 0 if not. Therefore, the maximum likelihood estimator for β in the generalized linear model suffers from the similar deflated p-value problem for a huge sample.

3. Sample size calibration test

We propose the following Monte Carlo approach to avoid the deflated p-value problem arising from the big data world. First, calculate the t (or z)-value under the null hypothesis $H_0 : \beta_i = 0$ without loss of generality for the i^{th} regression coefficient in a regression model by using a small sample sized n_s that is randomly chosen from the original sample. To reduce the sampling error in such a small sample, repeat the same sampling procedure to have the mean of the t (or z)-values. This is repeated until the full original dataset is used in order to draw the trajectory of the means of the t (or z)-values over the sample size.

The same Monte Carlo approach for the parameter estimate of interest and its p-value was proposed to reveal the deflated p-value problem from the huge sample (Lin *et al.*, 2013). However, this Monte Carlo method is not only computationally intensive but also provides no general rule for the sample size threshold for which the deflated p-value problem becomes an issue.

First of all, the Monte Carlo method for t-values can be conducted out without Monte Carlo simulations, meaning that our Monte Carlo method has no computational burden. To show this, we break down the t-value into two parts by

$$t = \frac{\hat{\beta}_i}{\sqrt{x^{ii}\hat{\sigma}^2}} = \frac{\hat{\beta}_i}{\sqrt{nx^{ii}\hat{\sigma}^2}}\sqrt{n},$$

where n is the original sample size and x^{ii} is the i^{th} diagonal term of $(X'\Omega X)^{-1}$. Note that $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$ is a standardized statistic over both standard deviations of x_i and the error term. It is an indication of sensitivity of the standardized dependent variable for one unit standardized change of x_i. $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$ converges to a constant in probability as $\hat{\beta}_i$ and $\hat{\sigma}^2$ are consistent estimators and nx^{ii} converges a finite constant. We define a sample size as 'huge' when the sample size is as many as $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$ converges. Since $|\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}| < t_{\alpha/2}/\sqrt{n}$ is $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$, one can take $n = 40{,}000$ as a huge sample so that 95% confidence interval of $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$ is ±0.01.

We call such $\hat{\beta}_i/\sqrt{nx^{ii}\hat{\sigma}^2}$ a calibration factor (CF) and denote it by CF(β_i). Therefore, when the original sample is huge, the mean of the t-values calculated from Monte-Carlo simulations, as described above, is close to CF(β_i) $\sqrt{n_s}$ by the law of large numbers where n_s is a sample size used in Monte Carlo simulation. Summarizing this, we have

Theorem 1. *If the original n samples are huge with the calibration factor CF(β_i), the mean $\bar{t}$ of Monte Carlo t-values with sample size n_s approaches CF(β_i) $\sqrt{n_s}$ where $n_s \leq n$.*

This states that the expected t-value is on a straight line of $\sqrt{n_s}$ with slope CF(β_i) passing through the origin. The CF(β_i) is calculated by $t/\sqrt{n}$ where t is the t-statistic for $\hat{\beta}_i$ estimated by using the original huge sample.

CF(β_i) is interpreted as the amount of change in the standardized dependent variable per one unit change of standardized x_i; therefore, the larger the CF(β_i) the higher the possibility to reject $H_0 : \beta_i = 0$. To find a significant level of CF(β_i), we calculate the t-value at sample size n_s by CF(β_i) × $\sqrt{n_s}$, denoted by t_{c,n_c} which increases as n_s increases and is called a sample size calibrated t-value. This means that CF(β_i) becomes stronger evidence to reject H_0 when t_{c,n_c} is significant in a smaller n_s while it becomes stronger evidence not to reject H_0 when t_{c,n_c} is not significant in a larger n_s.

For example, one may take three sample sizes, $n_1 = 50$, $n_2 = 100$, and $n_3 = 300$ as a small, a moderately large, and a large sample size, respectively. We then set a testing criterion as given by: CF(β_i) strongly supports to reject H_0 if $t_{c,50}$ is significant, weakly supports to reject H_0 if only $t_{c,100}$ and $t_{c,300}$ are significant, weakly supports not to reject H_0 if only $t_{c,300}$ is significant, and strongly supports not to reject H_0 if all three t-values are not significant. One may also take more than three sample sizes, different values of n_1, n_2, and n_3, and different testing criteria depending on the accuracy and seriousness of hypothesis testing with regard to the subject-matter.

When a variable of interest is categorical, however, the sample size should be chosen based on cell size with the smallest cell probability. For example, in the Korean National Health Insurance Sample, the death rate is 1.37% (i.e., 6,378 deaths among 466,345 people during 5 years) and the risk of death associated with the three groups of BMI indicated are one of our interests. In this case, we take $n_1 = 10{,}949$ as a small sample to include 150 deaths (i.e., 50 average deaths in each BMI

Table 1: Comparison of statistical and subject-matter significances

	Parameter	Sample size						
		1,000	3,000	5,000	10,000	40,000	50,000	100,000
β_{10}	Estimate	0.050	0.052	0.050	0.051	0.051	0.051	0.050
	t	1.510	2.688	3.395	4.821	9.710	10.781	15.218
	CF(β_{10})	0.048	0.049	0.048	0.048	0.048	0.048	0.048
	t_{c50}	0.338	0.347	0.339	0.341	0.341	0.341	0.340
	t_{c100}	0.478	0.491	0.480	0.482	0.482	0.482	0.481
	t_{c300}	0.827	0.850	0.831	0.835	0.835	0.835	0.834
	Good$_{50}$	0.500	0.426	0.172	0.009	<0.001	<0.001	<0.001
	Good$_{100}$	0.500	0.301	0.121	0.007	<0.001	<0.001	<0.001
	Good$_{300}$	0.427	0.174	0.070	0.004	<0.001	<0.001	<0.001
β_1	Estimate	−0.132	−0.132	−0.133	−0.133	−0.133	−0.133	−0.133
	t	−3.606	−6.272	−8.154	−11.529	−22.946	−25.756	−36.446
	CF(β_1)	−0.114	−0.115	−0.115	−0.115	−0.115	−0.115	−0.115
	t_{c50}	−0.806	−0.810	−0.815	−0.815	−0.813	−0.814	−0.815
	t_{c100}	−1.140	−1.145	−1.153	−1.153	−1.150	−1.152	−1.153
	t_{c300}	−1.975	−1.983	−1.997	−1.997	−1.992	−1.995	−1.996
	Good$_{50}$	0.051	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Good$_{100}$	0.036	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Good$_{300}$	0.021	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
β_2	Estimate	0.364	0.367	0.367	0.367	0.367	0.367	0.367
	t	9.977	17.398	22.478	31.810	63.545	71.102	100.537
	CF(β_2)	0.315	0.318	0.318	0.318	0.318	0.318	0.318
	t_{c50}	2.231	2.246	2.248	2.249	2.246	2.248	2.248
	t_{c100}	3.155	3.176	3.179	3.181	3.177	3.180	3.179
	t_{c300}	5.465	5.502	5.506	5.510	5.503	5.508	5.507
	Good$_{50}$	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Good$_{100}$	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Good$_{300}$	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

CF = calibration factor.

group), $n_2 = 21{,}898$ for 300 deaths (100 deaths in each BMI group) as a moderately large sample, and $n_3 = 65{,}693$ for 900 deaths (300 deaths in each BMI group) as a large sample.

It is similar to Good's p-value (Good, 1980, 1982) in that it calibrates the sample size effect. Note that our method calibrated test statistic t instead of p-value. With following simulation, we compare our method with Good's p-value denoted by $\text{Good}_{n_c} = \min(0.5, p\sqrt{n/n_c})$, where n_c are 50, 100, and 300 and n is original sample size.

Example 1 (continue)

Two regression models, $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and $y = \beta_{10} x_1 + \epsilon$, are applied to the sample size from 1,000 to 100,000 generated from the multivariate normal distribution given in Example 1. In each sample size, we simulate 1,000 hypothetical data sets.

Table 1 shows that Good's p-values give different significances for $\hat{\beta}_{10}$, depending on the sample size as the traditional p-values of t-statistics do. Good's p-value may be works in a moderately sample size, but not in huge samples that are now commonplace. However, CF(β_i) remains the same over different sample sizes as Theorem 1 says and hence the calibrated t_{c,n_c} gives consistent results regardless of sample sizes for significances of $\hat{\beta}_{10}$, $\hat{\beta}_1$, and $\hat{\beta}_2$. That is, $\hat{\beta}_{10}$ is strongly insignificant by insignificant $t_{c,300}$, $\hat{\beta}_1$ is weakly insignificant by insignificant $t_{c,50}$ and $t_{c,100}$, but significant $t_{c,300}$, and $\hat{\beta}_2$ is strongly significant by significant $t_{c,50}$. This implies that x_1 has no subject-matter effect on y in both simple and multiple regression models.

4. Real data analysis

Obesity and overweight are two of the most important risk factors related to health problems and mortality (Renehan *et al.*, 2008; Larsson and Wolk, 2008; Torloni *et al.*, 2009). BMI and WC have been commonly utilized by researchers as measures of individual obesity (Lam *et al.*, 2015; Nyamdorj *et al.*, 2008).

Obesity (a BMI of 30 or more) and overweight (a BMI of 25–29.9) are associated with an increased risk of death among several populations (Katzmarzyk *et al.*, 2003; Ogden *et al.*, 2006; Adams *et al.*, 2006; Berrington de Gonzalez *et al.*, 2010; Zheng *et al.*, 2011). Some studies suggest that overweight is either beneficial or has little effect on mortality (Adams *et al.*, 2006; Flegal *et al.*, 2007; Orpana *et al.*, 2010; Zheng *et al.*, 2011). These inconsistencies are not established; however, they could be due to confounding by a covariate such as smoking and disease related weight loss, different populations, and other BMI related factors (Berrington de Gonzalez *et al.*, 2010). To address these inconsistencies, we evaluate the relationship between BMI and the risk of death by our sample size calibration method applied to five different populations.

We use three sample size calibrated t-values denoted by t_{c,n_1}, t_{c,n_2}, and t_{c,n_3} calculated from small, moderately large, and large samples, respectively. Our interest is in the effects of WC and BMI on mortality; therefore, we define the samples that include an average of 50, 100, and 300 deaths for each WC and BMI group to be small, moderately large, and large. For example, 6,378 deaths occurred in a group of 466,345 Korean people who received a health screening at least one time between 2008 and 2013. Thus, the small, moderately large, and large sample sizes are 10,949, 21,898, and 65,693, respectively when WC or BMI is categorized into three groups, whereas 36,496, 72,992, and 218,978, respectively when categorized into 10 groups. We denote the corresponding sample size calibrated t-values in this section by $t_{c,50}$, $t_{c,100}$, and $t_{c,300}$ to stress the average number of deaths included in each group of WC or BMI.

4.1. The effects of WC and BMI on mortality with different covariates

National health insurance in Korea is mandatory for all citizens and contains all data related to the history of individual medical treatments including health screening data. Since WC has been included in health screening since 2008 and the people who had cancer or CVD in 2006 or 2007 are excluded from the data, our data analysis for the effects of WC and BMI on mortality includes 466,345 Korean people who received a health screening between 2008 and 2013, during which 6,378 (1.37%) deaths were observed.

WC and BMI are divided into three groups: low (WC < 73.9cm, BMI < 21.5), middle (73.9 ≤ WC < 92.2cm, 21.5 ≤ BMI < 28), and high (WC ≥ 92.2, BMI ≥ 28). Five Cox's proportional hazard regression models are considered to investigate the effects of WC and BMI on the risk of death. Model 1 includes only WC without age, and Model 2 includes WC with age. These two models are adjusted by gender with males as the reference. Model 3 and 4 include only BMI without age and with age, respectively where both models are adjusted by gender. Model 5 includes both WC and BMI with gender and age; however, WC and BMI are highly correlated with correlation coefficient 0.8 for both genders. Here, the effects of WC and BMI are estimated based on the middle groups as references.

Table 2 presents five Cox regression models with p-values for statistical significance and CF(β_i) and t_{c,n_c} for subject-matter significance where the convergence of CF(β_i) is confirmed using 90% of the original 466,345 samples, selected at random. Based on the traditional p-value, the effects of WC on mortality are considerably different from model to model. Model 1 shows clear positive association between WC and mortality, whereas, in Model 2 where age is additionally included, the

Table 2: Five Cox's proportional hazard models for the effects of WC and BMI on all causes of mortality

Models	Variables	Estimates	t	p-value	CF(β)	$t_{c,50}$	$t_{c,100}$	$t_{c,300}$
Model 1	WC_{low}	−0.126	−3.64	0.0003	−0.0053	−0.550	−0.780	−1.360
	WC_{high}	0.110	2.74	0.006	0.0040	0.420	0.590	1.030
	$\text{Sex}_{\text{female}}$	−0.495	−17.59	<2e−16	−0.0258	−2.700	−3.820	−6.610
Model 2	WC_{low}	0.414	12.53	<2e−16	0.0183	1.910	2.710	4.690
	WC_{high}	−0.078	−1.93	0.053	−0.0028	−0.290	−0.410	−0.720
	$\text{Sex}_{\text{female}}$	−0.862	−32.39	<2e−16	−0.0474	−4.960	−7.010	−12.100
	Age	0.109	106.13	<2e−16	0.1554	16.260	23.000	39.800
Model 3	BMI_{low}	0.698	26.00	<2e−16	0.0381	3.990	5.640	9.770
	BMI_{high}	−0.204	−3.98	0.000068	−0.0058	−0.610	−0.860	−1.490
	$\text{Sex}_{\text{female}}$	−0.679	−25.62	<2e−16	−0.0375	−3.920	−5.550	−9.610
Model 4	BMI_{low}	0.681	25.60	<2e−16	0.0375	3.920	5.550	9.610
	BMI_{high}	0.010	0.20	0.84	0.0003	0.031	0.044	0.077
	$\text{Sex}_{\text{female}}$	−0.799	−30.60	<2e−16	−0.0448	−4.190	−6.620	−11.500
	Age	0.104	102.60	<2e−16	0.1502	15.700	22.200	38.500
Model 5	BMI_{low}	0.691	22.53	<2e−16	0.0330	3.450	4.880	8.460
	BMI_{high}	−0.048	−0.85	0.395	−0.0012	−0.126	−0.180	−0.310
	WC_{low}	0.007	0.19	0.85	0.0003	0.031	0.044	0.077
	WC_{high}	0.113	2.47	0.013	0.0036	0.377	0.533	0.924
	$\text{Sex}_{\text{female}}$	−0.792	−29.39	<2e−16	−0.0430	−4.500	−6.360	−11.000
	Age	0.104	101.57	<2e−16	0.1487	11.560	22.000	38.100

CF = calibration factor; WC = waist circumference; BMI = body mass index.

situation is completely overturned. In Model 5 where WC are adjusted BMI, sex, and age, WC_{high} is not significant. All of these inconsistent results are caused by big data. A similar problem occurs in the case of BMI as shown in Model 3, 4, and 5. Model 3 and 5 show a negative association between BMI and mortality, addressing the obesity paradox, a phenomenon where mortality decreases as BMI increases in a certain BMI range such as overweight or obesity (Carnethon *et al.*, 2012; Kalantar-Zadeh *et al.*, 2004, 2012; Uretsky *et al.*, 2007), while no such an obesity paradox in Model 4.

However, based on t_{c,n_c}, WC has no effect on mortality in Model 1, 2, and 3 except for WC_{low} in Model 2 because its $t_{c,100}$ and $t_{c,300}$ are greater than 2 (weakly significant). For BMI, BMI_{low}s are strongly significant but BMI_{high}s are strongly insignificant in all Model 3, 4, and 5, showing concordant results of no obesity paradox. Thus, the sample size calibrated test for subject-matter significance provides consistent results for WC and BMI regardless of the adjusting factors included in model. The obesity paradox of BMI is further discussed with four huge data sets in the following section.

4.2. The effect of BMI on mortality in four different populations

We note that an obesity paradox exists when overweight (a BMI of 25–27.9) or moderately obese (29–29.9) is associated with a lower risk of death than normal (a BMI of 22.5–24.9). The study investigated BMI (Berrington de Gonzalez *et al.*, 2010) in relation to the risk of death from any cause by using 1.46 million white (non-Hispanic) adults and 160,087 deaths and their main results are summarized in the first row of Table 3. For women, overweight (a BMI of 25–29.9) and obesity (a BMI of 30 or more) are associated with increased mortality. However, for men, overweight has a significantly lower risk of death than normal, although the hazard ratio of a BMI of 25.0–27.4, 0.97, is slightly lower than 1.

Using 527,265 Americans (313,047 men and 214,218 women) who were 51 to 71 years old in 1995–1996 and among whom 61,317 (42,173 men and 19,144 women) died during a maximum

Table 3: Tests for the effect of BMI on all cause mortality in four populations

Population	Sex	Hazard ratios(> 1: more risky than reference group, < 1: less risky than reference group)									
		15–18.4	18.5–19.9	20-22.4	22.5–24.9	25–27.4	27.5–29.9	30–34.9	35–39.9	40–49.9	
White-	F	2.02*	1.34*	1.06*	reference	1.03*	1.11*	1.25*	1.59*	1.99*	
adults	M	1.98*	1.6*	1.18*	reference	0.97*	1.03*	1.16*	1.44*	1.93*	
50–71		< 18.5	18.5–20.9	21–23.4	23.5–24.9	25–26.4	26.5–27.9	28–29.9	30–34.9	35–39.9	40–
years old	F	2.03*	1.3*	1.07*	reference	1	1.06	1.07*	1.18*	1.49*	1.94*
Americans	M	1.97*	1.54*	1.14*	reference	0.95*	0.95*	1	1.1*	1.35*	1.83*
		≤ 15	15.1–17.5	17.6–20.0	20.1–22.5	22.6–25.0	25.1–27.5	27.6–30.0	30.1–32.5	32.6–35.0	35.1–50
East Asian		2.76*	1.84*	1.35*	1.09*	reference	0.98	1.07*	1.2*	1.5*	1.49*
Indian&Bangla		2.14*	1.59*	1.26*	1.09	reference	0.98	0.94	1.03	0.86	1.27
		< 18.5	18.5–19.9	20–21.4	21.5–22.9	23–24.9	25–26.4	26.5–27.9	28–29.9	30–32.4	32.5–
Koreans		2.31*	1.73*	1.25*	1.23*	reference	0.86*	0.88	0.95	1.12	1.25
Population	Sex	Sample size calibration test significance of BMI hazard ratio									
		15–18.4	18.5–19.9	20–22.4	22.5–24.9	25–27.4	27.5–29.9	30–34.9	35–39.9	40–49.9	
White-	F	+ + +	− + +	− − −	reference	− − −	− − −	− − +	+ + +	+ + +	
adults	M	− − +	− − +	− − +	reference	− − −	− − −	− − +	− − +	− + +	
50–71		< 18.5	18.5–20.9	21–23.4	23.5–24.9	25–26.4	26.5–27.9	28–29.9	30–34.9	35–39.9	40–
years old	F	+ + +	− + +	− − −	reference	− − −	− − −	− − −	− − +	− + +	+ + +
Americans	M	− − +	− + +	− − +	reference	− − −	− − −	− − −	− − −	− − +	− + +
		≤ 15	15.1–17.5	17.6–20.0	20.1–22.5	22.6–25.0	25.1–27.5	27.6–30.0	30.1–32.5	32.6–35.0	35.1–50
East Asian		− − −	− − +	− − −	− − −	reference	− − −	− − −	− − −	− − −	− − −
Indian&Bangla		− + +	− − +	− − −	− − −	reference	− − −	− − −	− − −	− − −	− − −
		< 18.5	18.5–19.9	20–21.4	21.5–22.9	23–24.9	25–26.4	26.5–27.9	28–29.9	30–32.4	32.5–
Koreans		+ + +	+ + +	− − +	− + +	reference	− − +	− − −	− − −	− − −	− − −

*: stands for p-value < 0.05. For sample size calibration test, strong significance is represented by + + +, weak significance by − + +, weak insignificance by − − +, and strong insignificance by − − −.

follow-up of 10 years through 2005 (Adams *et al.*, 2006), the obesity and underweight were shown to be associated with increasing mortality for both men and women but overweight was not in relation to excess risk of death. These are displayed in the second row of Table 3. Overweight is significantly associated with the lowest risk of death for men, showing the obesity paradox.

A total of 19 cohorts from the Asia Cohort Consortium BMI project in 2008 was used to investigate the association of BMI and mortality among 1,141,609 Asian people yielding 120,700 deaths (Zheng *et al.*, 2011). Underweight was associated with an increasing mortality in all Asians, and an excess risk of death was associated with a high BMI in East Asians but not in Indians and Bangladeshis. The third row of Table 3 provides these results with the lowest hazard ratios at overweight with a BMI of 25.1–27.5 among East Asians and overweight with a BMI of 27.5–30.0 among Indians and Bangladeshis, indicating the obesity paradox.

The Korean National Health Insurance Service released a national sample cohort (from 2002–2010) consisting of 1,025,340 Koreans among whom 153,484 received a health screening in 2002 or 2003. During 7.91 ± 0.59 mean years follow-up until 2010, 3,937 deaths occurred among the 153,484 Koreans. The study based on this data (Kim *et al.*, 2015) showed that overweight was associated with a lower risk of death than normal, underweight, and obesity, as presented in the fourth row of Table 3.

For the sample size calibrated test, we use representative symbols as given by a + sign if $|t_{c,n_c}| > 1.96$ and a − sign if not, so that we represent strong significance by + + + for $|t_{c,50}| > 1.96$, $|t_{c,100}| > 1.96$, and $|t_{c,300}| > 1.96$, weak significance by −++ for $|t_{c,50}| \leq 1.96$ but $|t_{c,100}| > 1.96$ and $|t_{c,300}| > 1.96$, weak significance by − − + for only $|t_{c,300}| > 1.96$, and finally strong insignificance by − − − when all of the three values are 1.96 or less.

Note that the sample sizes to calculate t_{c,n_c} are different for different populations. For example, the moderately large sample sizes for $t_{c,100}$ are 9,274 for men among 1.46 million white adults (Berrington de Gonzalez *et al.*, 2010) to include a total of 900 deaths for 9 BMI groups and 18,649 for Indians and Bangladeshis among 1 million Asians (Zheng *et al.*, 2011) to include a total of 1,000 deaths for 10 BMI groups. The subject-matter significance results based on t_{c,n_c} are presented in the second table of Table 3.

In the populations of 1.46 million white adults and 0.53 million 50–71 years old Americans, the sample size calibrated test reveals a wide U-shape with a BMI of 22.5–29.9 or 23.5–34.9, associated with the lowest risk of death for men, whereas, for women, an apparent wide U-shape with a BMI of 20–29.9 or 21–29.9 associated with the lowest risk of death and high hazard ratios are at underweight (a BMI of 15–19.9 or less than 20.9) and obese (35–39.9) or morbidly obese (40–49.9).

However, in the populations of Asian origin (Zheng *et al.*, 2011; Kim *et al.*, 2015), the sample size calibration tests show no effect of BMI on all causes of mortality for East Asians, and a lying L-shape with no higher risk of death associated with overweight and obese for Indians, Bangladeshis, and Koreans.

There is no subject-matter significance of the obesity paradox regardless of race, sex, and age groups; however, the existence of a statistically significant obesity paradox and the range of BMI associated with the statistically lowest risk of death depend on study populations.

5. Conclusion

The deflated p-value problem arising from the large sample world makes the usual statistical testing results unreliable when the null hypothesis is false, even if it is negligible. We decomposed the t-statistic calculated from a huge sample into a calibration factor and the square root of the sample size in general linear models. The calibration factor remains the same over Monte Carlo simulations; therefore, we proposed the sample size calibration test as an evaluation of subject-matter significance to resolve the deflated p-value problem. Simulation and real data analysis showed that the sample size calibration method might be useful when a null hypothesis is rejected and the significance is due to an artifact of the large sample, producing a too narrow confidence interval, and that it is less sensitive to inclusion and omission of some covariates in the model.

The sample size calibration method can be easily extended to an χ^2 test (F test) in the way that we evaluate a calibration χ^2 factor (F factor) by dividing the χ^2 (F) value by the sample size of the huge sample and then calculate the sample size calibrated χ^2 (F) value by the calibration χ^2 (F) factor using appropriate smaller samples than the original huge sample size.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2018R1C1B5043739). This work was supported by Hankuk University of Foreign Studies Research Fund.

Appendix: Proof of Lemma 1

Proof: It is well known that the OLS estimates of regression coefficients are $r_{x_2y}(s_y/s_{x_2})$ for $y = \beta_{20}x_2 + \epsilon_2$, $r_{x_2x_1}(s_{x_1}/s_{x_2})$ for $x_1 = a_1x_2 + \eta_1$, and $r_{x_1x_2}(s_{x_2}/s_{x_1})$ for $x_2 = b_1x_1 + \eta_2$ where we assume that

y, x_1, x_2 are all centered by the respective mean. Thus, the OLS of β_1 is

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^{n} \hat{\eta}_{1i}^2\right)^{-1} \sum_{i=1}^{n} \hat{\eta}_{1i}\hat{\epsilon}_{2i} = \left(\sum_{i=1}^{n} (x_{1i} - \hat{a}_1 x_{2i})^2\right)^{-1} \sum_{i=1}^{n} (x_{1i} - \hat{a}_1 x_{2i})\left(y_i - \hat{\beta}_{20} x_{2i}\right) \\ &= \left(\sum_{i=1}^{n} \left(x_{1i}^2 - 2\hat{a}_1 x_{1i} x_{2i} + \hat{a}_1^2 x_{2i}^2\right)\right)^{-1} \sum_{i=1}^{n} x_{1i}^2 \left(\sum_{i=1}^{n} x_{1i}^2\right)^{-1} \sum_{i=1}^{n} \left(x_{1i} y_i - \hat{\beta}_{20} x_{1i} x_{2i} - \hat{a}_1 x_{2i} y_i + \hat{a}_1 \hat{\beta}_{20} x_{2i}^2\right).\end{aligned}$$

Note that $\hat{a}_1(\sum_{i=1}^{n} x_{1i}^2)^{-1} \sum_{i=1}^{n} x_{1i} x_{2i} = \hat{a}_1 \hat{b}_1 = r_{x_1 x_2}^2$ and $\hat{a}_1^2 \sum_{i=1}^{n} x_{2i}^2 (\sum_{i=1} x_{1i}^2)^{-1} = \hat{a}_1^2 (s_{x_2}/s_{x_1})^2 = r_{x_1 x_2}^2$. Thus, we have

$$\sum_{i=1}^{n} \left(x_{1i}^2 - 2\hat{a}_1 x_{1i} x_{2i} + \hat{a}_1^2 x_{2i}^2\right) \left(\sum_{i=1}^{n} x_{1i}^2\right)^{-1} = 1 - r_{x_1 x_2}^2. \qquad \text{(A.1)}$$

Observe that $\hat{\beta}_{20}(\sum_{i=1}^{n} x_{1i}^2)^{-1} \sum_{i=1}^{2} x_{1i} x_{2i} = \hat{\beta}_{20} \hat{b}_1 = \hat{\beta}_{20} r_{x_1 x_2} (s_{x_2}/s_{x_1})$, $\hat{a}_1(\sum_{i=1}^{n} x_{1i}^2)^{-1} \sum_{i=1}^{n} x_{2i} y_i = r_{x_1 x_2} \hat{\beta}_{20}$ (s_{x_2}/s_{x_1}), and $\hat{a}_1 \hat{\beta}_{20} (\sum_{i=1}^{n} x_{1i}^2)^{-1} \sum_{i=1}^{n} x_{2i}^2 = r_{x_1 x_2} \hat{\beta}_{20} (s_{x_2}/s_{x_1})$. This gives

$$\left(\sum_{i=1}^{n} x_{1i}^2\right)^{-1} \sum_{i=1}^{n} \left(x_{1i} y_i - \hat{\beta}_{20} x_{1i} x_{2i} + \hat{a}_1 x_{2i} y_i + \hat{a}_1 \hat{\beta}_{20} x_{2i}^2\right) = \hat{\beta}_{10} - r_{x_1 x_2} \hat{\beta}_{20} \frac{s_{x_2}}{s_{x_1}}. \qquad \text{(A.2)}$$

The claim is completed by (A.1) and (A.2). □

References

Adams KF, Schatzkin A, Harris TB, Kipnis V, Mouw T, Ballard-Barbash R, Hollenbeck A, and Leitzmann MF (2006). Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old, *New England Journal of Medicine*, **355**, 763–778.

Altman M (2004). Introduction to special issue on statistical significance, *Journal of Socio-Economics*, **33**, 523–675.

Bayarri MJ and Berger JO (1999). Quantifying surprise in the data and model verification. In Bernardo JM, Berger JO, Dawid AP, and Smith AFM (eds) *Bayesian Statistics* (6th ed, pp. 53–82), Oxford University Press, Oxford.

Bayarri MJ and Berger JO (2000). *P*-values for composite null models, *Journal of the American Statistical Association*, **95**, 1127–1142.

Berrington de Gonzalez A, Hartge P, Cerhan JR, *et al.* (2010). Body-mass index and mortality among 1.46 million white adults, *New England Journal of Medicine*, **363**, 2211–2219.

Bolstad WM (2009). *Understanding Computational Bayesian Statistics*, John Wiley & Sons, Boston, MA.

Carnethon MR, De Chavez PJ, Biggs ML, *et al.* (2012). Association of weight status with mortality in adults with incident diabetes, *The Journal of American Medical Association*, **308**, 581–590.

DeGroot MH and Schervish MJ (2002). *Probability and Statistics* (3rd ed), Pearson, Boston.

Emerson S (2009). Small sample performance and calibration of the Empirical Likelihood method (Ph.D. thesis), Stanford University, Stanford.

Flegal KM, Graubard BI, Williamson DF, and Gail MH (2007). Cause-specific excess deaths associated with underweight, overweight, and obesity, *The Journal of Medical Association*, *298*, 2028–2037.

Fomby TB, Hill RC, and Jhonson SR (1984). *Advanced Econometric Methods*, Springer Verlag, New York.

Gelman A, Meng XL, and Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion), *Statistica Sinica*, **6**, 733–807.

Ghose A, Smith M, and Telang R (2006). Internet exchanges for used books: an empirical analysis of product cannibalization and welfare impact, *Information Systems Research*, **17**, 3–19.

Ghose A and Yang S (2009). An empirical analysis of search engine advertising: sponsored search in electronic markets, *Management Science*, **55**, 1605–1622.

Good IJ (1980). C73. The diminishing significance of a p-value as the sample size increase, *Journal of Statistical Computation and Simulation*, **11**, 307–313.

Good IJ (1982). C144. The diminishing significance of a fixed p-value as the sample size increase: a discrete model, *Journal of Statistical Computation and Simulation*, **16**, 312–313.

Greene WH (2003). *Econometric Analysis* (5th ed), Pearson Education, New Jersey.

Halsey LG, Curran-Everett D, Vowler SL, and Drummond GB (2015). The fickle P value generate irreproducible results, *Nature Methods*, **12**, 179–185.

Harlow LL, Mulaik S, and Steiger JH (2016). *What If There were No Significance Tests?: Classic Edition* (Chapter 2–5), Psychology Press, New York.

Hubbard R and Armstrong JS (2006). Why we don't really know what statistical significance means: Implications for educators, *Journal of Marketing Education*, **28**, 114–120.

Johnson DH (1999). The insignificance of statistical significance testing, *Journal of Statistical Computation and Simulation*, **63**, 763–772.

Johnston J (1984). *Econometric Methods* (3rd ed), McGraw-Hill, New York.

Kalantar-Zadeh K, Block G, Horwich T, and Fonarow GC (2004). Reverse epidemiology of conventional cardiovascular risk factors in patients with chronic heart failure, *Journal of the American College of Cardiology*, **43**, 1439–1444.

Kalantar-Zadeh K, Streja E, Molnar MZ, Lukowsky LR, Krishnan M, Kovesdy CP, and Greenland S (2012). Mortality prediction by surrogates of body composition: an examination of the obesity paradox in hemodialysis patients using composite ranking score analysis, *American Journal of Epidemiology*, **175**, 793–803.

Katzmarzyk PT, Janssen I, and Ardern CI (2003). Physical inactivity, excess adiposity and premature mortality, *Obesity Reviews*, **4**, 257–290.

Kim NH, Lee J, Kim TJ, *et al.* (2015). Body mass index and mortality in the general population and in subjects with chronic disease in Korea: a nationwide cohort study (2002–2010), *PloS One*, **10**, e0139924.

Kirk RE (1996). Practical significance: a concept whose time has come, *Educational and Psychological Measurement*, **56**, 746–759.

Lam BCC, Koh GCH, Chen C, Wong MTK, and Fallows SJ (2015). Comparison of Body Mass Index (BMI), Body Adiposity Index (BAI), Waist Circumference (WC), Waist-To-Hip Ratio (WHR) and Waist-To-Height Ratio (WHtR) as Predictors of Cardiovascular Disease Risk Factors in an Adult Population in Singapore, *PLoS One*, **10**, e0122985.

Larsson SC and Wolk A (2008). Excess body fatness: an important cause of most cancers, *The Lancet*, **371**, 536–537.

Leamer EE (1978). *Specification Searches*, John Wiley & Sons, New York.

Lee J, Lee JS, Park SH, Shin SA, and Kim KW (2016). Cohort profile: the national health insurance service–national sample cohort (NHIS-NSC), South Korea, *International Journal of Epidemiology*, **46**, e15–e15.

Lin M, Lucas Jr HC, and Shmueli G (2013). Research commentary–too big to fail: large samples and the p-value problem, *Information Systems Research*, **24**, 906–917.

Nyamdorj R, Qiao Q, Lam TH, *et al.* (2008). BMI compared with central obesity indicators in relation to diabetes and hypertension in Asians, *Obesity (Silver Spring)*, **16**, 1622–1635.

Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, and Flegal KM (2006). Prevalence of overweight and obesity in the United States, 1999–2004, *The Journal of Medical Association*, **295**, 1549–1555.

Orpana HM, Berthelot JM, Kaplan MS, Feeny DH, McFarland B, and Ross NA (2010). BMI and mortality: results from a national longitudinal study of Canadian adults, *Obesity*, **18**, 214–218.

Pawitan Y (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

Qian H and Shmidt P (2003). Partial GLS regression, *Economic letters*, **79**, 385–392.

Renehan AG, Tyson M, Egger M, Heller RF, and Zwahlen M (2008). Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies, *The Lancet*, **371**, 569–578.

Sellke T, Bayarri MJ, and Berger JO (2001). Calibration of p-values for testing precise null hypotheses, *The American Statistician*, **55**, 63–71.

Torloni MR, Betran AP, Horta BL, Nakamura MU, Atallah AN, Moron AF, and Valente O (2009). Prepregnancy BMI and the risk of gestational diabetes: a systematic review of the literature with meta-analysis, *Obesity Reviews*, **10**, 194–203.

Tsao M (2001). A small sample calibration method for the empirical likelihood ratio, *Statistics & Probability Letters*, **54**, 41–45.

Tsao M (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions, *The Annals of Statistics*, **32**, 1215–1221.

Uretsky S, Messerli FH, Bangalore S, Champion A, Cooper-Dehoff RM, Zhou Q, and Pepine CJ (2007). Obesity paradox in patients with hypertension and coronary artery disease, *American Journal of Medicine*, **120**, 1863–1870.

Wellek S (2017). A critical evaluation of the current p-value controversy, *Biometrical Journal*, **59**, 854–872.

Whitehead J (1980). Fitting Cox's Regrssion Model to Survival Data using GLIM, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29**, 268–275.

Zheng W, McLerran DF, Rolland B, Zhang X, Inoue M, Matsuo K, and Irie F (2011). Association between body-mass index and risk of death in more than 1 million Asians, *New England Journal of Medicine*, **364**, 719–729.

Received May 8, 2018; Revised June 15, 2018; Accepted June 25, 2018