Journal of the Korean Data & Information Science Society
2015, **26**(3), 773–779

http://dx.doi.org/10.7465/jkdi.2015.26.3.773
한국데이터정보과학회지

Statistical analysis of KNHANES data with measurement error models

Jinseub Hwang[1]

[1]National Evidence-based Healthcare Collaborating Agency

Received 10 April 2015, revised 22 May 2015, accepted 22 May 2015

Abstract

We study a statistical analysis about the fifth wave data of the Korea National Health and Nutrition Examination Survey based on linear regression models with measurement errors. The data is obtained from a national population-based complex survey. To demonstrate the availability of measurement error models, two results between the general linear regression model and measurement error model are compared based on the model selection criteria which are Akaike information criterion and Bayesian information criterion. For our study, we use the simulation extrapolation algorithm for measurement error model and the jackknife method for the estimation of standard errors.

Keywords: Blood pressure, body mass index, Korea National Health and Nutrition Examination Survey, measurement error model, simulation extrapolation.

1. Introduction

In many fields of application, practice variables are often contaminated with measurement error. This may be the case due to bad measurement tools or impossibility of direct measurement. We provide a few examples to illustrate this. First, we want to predict the yield of corn in several counties in Iowa and the covariate used is available nitrogen in the soil. To estimate the available soil nitrogen, it is necessary to sample the soil of the experimental plot and to perform a laboratory analysis of the sample. As a result of the sampling and of the laboratory analysis, the true available nitrogen was not observed but only its estimate was obtained. This example is taken from Fuller (1987). Second, we are looking at patients undergoing a rare surgical procedure in different hospitals. As this is a rare surgery, the number of observed cases in each hospital will be very small. Now, suppose, we are interested in modeling the time to recovery by blood pressure, heart rate and other such measurements; then, again, it seems likely that the measured covariates are affected by some error. Lastly, we are interested in estimating the volume of trees for several areas. We have data on volume of stem wood for the current year and measures of the diameter and height of stem wood from a previous year. Here, we may model the volume of trees by their diameter and height. However, the latter two are likely to be measured with error.

[1] Associate research fellow, Ph.D, National Evidence-based Healthcare Collaborating Agency, Seoul 100-705, Korea. E-mail: hjs0409@neca.re.kr

In this paper, we want to show the requirement of measurement error models based on a real dataset. Specifically, we consider systolic blood pressure (SBP) and diastolic blood pressure (DBP) as outcome variables and the associated factors with blood pressure like age and body mass index (BMI) as covariates. It seems like that BMI would be measured with error. We use the general linear regression model and the measurement error model. Especially, we have used simulation extrapolation (SIMEX) algorithm by Cook and Stefanski (1994). In this paper, we have used the last year data on the fifth wave of the Korea National Health and Nutrition Examination Survey, which is a nationally representative cross-sectional survey on the health and nutrition of the non-institutionalized civilian population of South Korea (Korea Centers for Disease Control & Prevention, 2013). We give a brief overview of the method including the data source, study population and statistical models in Section 2. Section 3 presents the results and we discuss possible refinements in Section 4.

2. Materials and methods

2.1. Korean National Health and Nutrition Examination Survey

This study considers 2012 data from the fifth wave of the Korean National Health and Nutrition Examination Survey (KNHANES). This survey has been systematically conducted since 1998 by the Korean Center for Disease Control and Prevention (KCDC) and composed of a Health Interview Survey, a Health Behavior Survey, a Health Examination Survey, and a Nutrition Survey. Since 2007, every year and any combination of consecutive years comprise a nationally representative sample. A three-stage sample designing is used for the KNHANES. The primary sample units (PSUs) are selected through proportional allocation based on the Korean Population and Housing Census through a stratified, multistage, probability sampling design that was based on the sex, age, and geographical area using household registries. Each PSU consists of 20 households. Following the selection of PSUs, all residence units in the PSU are listed and 20 household are selected through the field survey for household screening. The final stage of selection occurs in the household, where all members aged 1 year and over are selected to participate. Approximately 10,000 persons are sampled in total in all 192 PSUs per year. Sampling units were households from which the data were collected. The subjects were informed that they had been randomly selected as a sample household and were asked to voluntarily participate in the survey (Korea Centers for Disease Control and Prevention, 2013).

The KNHANES survey consisted of four stages: (1) selection and recruitment of a representative sample of civilians, (2) performance of the Health Interview Survey and Health Examination Survey at mobile examination centers (MECs), (3) performance of the Health Nutrition Survey within 2 weeks after the completion of the Health Examination Survey, and (4) mailing the test results to the subjects within 2 weeks after the end of the survey. The KNHANES Health Interview Survey was conducted through face-to-face interviews at MECs by trained interviewers and the Health Nutrition Survey was also conducted by trained nutritionists using face-to-face interviews at the homes of the subjects. Informed consent was given by each participant before inclusion in the study (Korea Centers for Disease Control and Prevention, 2009). The unit non-response rate of the KNHANES survey was 22.2% in 2008 and 17.2% in 2009 (Korea Centers for Disease Control and Prevention, 2013). Age, marital status, education level, occupation, household income, and comorbidity

were collected to identify the demographic characteristics of the sample. Health behaviors were measured by questions about disease management, health choices, and functional and psychological status. The disease management category included questions about current treatment status, medication compliance, the status of symptoms, and asthma attacks. The health choices category assessed the use of tobacco or alcohol, and the frequency of exercise, health examinations, and influenza vaccinations. Functional health was assessed by activity limitations, measured using a dichotomous scale (yes/no), and perceived health status, which was represented by the choice of good, fair, or poor. Psychological health was assessed by symptoms of depression in the last two weeks, measured with a dichotomous scale (yes/no), and the level of stress, which was measured with a 4-point Likert scale (1= very much to 4 = very little) (Jang and Yoo, 2013).

2.2. Variables and study population

The factors associated with blood pressure are known as follows; age, gender, race, family history, obesity, physical activity, tobacco, stress, consumption level of alcohol and sodium, and intake of potassium and vitamin D (Mayo Clinic, http://www.mayoclinic.org/diseases-conditions/high-blood-pressure/basics/risk-factors/con-20019580). The standardized questionnaire include demographic variables (age, gender) and lifestyle variables (physical activity level). Personal medical histories, family history, stress, and alcohol and tobacco consumption habits were determined by a self-administered questionnaire. Blood pressure was measured three times on the right arm while the individual was in a seated position after at least 5 min of rest using a mercury sphygmomanometer (Baumanometer; Baum, Copiague, NY, USA). The final blood pressure value was obtained by averaging the values of the second and third measurements. Weight was measured using a medical balance (GL-6000-20; CAS, Seoul, Korea), and height using a wall-mounted stadiometer (Seca 220; Seca, Hamburg, Germany). BMI was calculated by dividing weight (kg) by square of height (m^2). Also, the amount of sodium (mg/day) and potassium (mg/day) intake were calculated through the food frequency survey and vitamin D (ng/mL) was measured from blood test. In this paper, we use SBP and DBP as outcome variables, separately and other variables are used as covariates. Specifically, BMI was used as obesity information with the measurement error.

The number of total individuals for 2012 from the fifth waves of the KNHANES was 8,058 subjects aged 1-98 years. First, we selected 4,831 participants after removing those who aged below 19 years, who had hypertension based on the criteria of SBP above 140 $mmHg$, DBP above 90 $mmHg$ or taking medication. Finally, 405 subjects were included except of the participants with missing data.

2.3. Statistical analysis

We have carried out two types of regression analyses. First, the general linear regression analysis (Model1) was conducted without measurement error concept as below

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{11} x_{11} + \epsilon$$

where $\epsilon \sim N(0, \sigma_e^2)$. Here y are SBP and DBP, x_1 is BMI, and $(x_2, \cdots, x_{11})$ are other covariates like age and gender.

Second model considered measurement error in covariate (BMI) based on SIMEX algorithm (Model2). We will start with the simplest regression model with one independent variable. Suppose we wish to estimate the population relationship

$$y = \alpha + \beta x + \epsilon \tag{2.1}$$

Regrettably, we only have a data on

$$x^* = x + \eta \tag{2.2}$$

where $\eta \sim N(0, \sigma_\eta^2)$. i.e. our observed covariate is measured with an additive error. Then, the regression of y on x can be obtained by inserting (2.2) into (2.1);

$$y = \beta_0 + \beta_1 x^* + u, u = \epsilon - \beta\eta$$

To check the size of the bias we consider the ordinary least squares estimator for β

$$\hat{\beta} = \frac{cov(x^*, y)}{var(x^*)} = \frac{cov(x + u, \beta x + \epsilon)}{var(x + u)} = \frac{\beta\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \lambda\beta$$

The quantity λ is referred to as reliability or signal-to-total variance ratio. Since $0 < \lambda < 1$ the coefficient $\hat{\beta}$ will be biased towards zero. Also, we can consider what happens to the bias as we add more variables to the model (Hyslop and Imbens, 2001).

SIMEX algorithm has become a useful tool for correcting effect estimates in the presences of additive measurement error. This is a simulation-based method aimed at reducing bias caused by the inclusion of measurement error covariate. SIMEX-method uses the relationship between the size of the measurement error described of the effect estimator when ignoring the measurement error. So we can define the function

$$\sigma_u^2 \to \beta^*(\sigma_u^2) := g(\sigma_u^2)$$

where β^* is the limit to which the naive estimator converges as the sample size $n \to \infty$. Consistency implies that $g(0) = \beta$. The idea of the SIMEX method is to approximate the function $g(\sigma_u^2)$ by a parametric approach $g(\sigma_u^2, \Gamma)$, for example with a quadratic approximation $g_{quad}(\sigma_u^2, \Gamma) = \gamma_0 + \gamma_1\sigma_u^2 + \gamma_2(\sigma_u^2)^2$.

Simulation step

To estimate Γ, the method adds measurement error with variance $\lambda\sigma_u^2$ to the contaminated variable, where $\lambda > 0$ quantifies the additional amount of measurement error that is added. The resulting measurement error variance is then $(1 + \lambda)\sigma_u^2$. The naive estimator for this increased measurement error is calculated. This simulation procedure is repeated B times. The average over the B estimators estimates $g((1 + \lambda)\sigma_u^2)$. Performing these simulations for a fixed grid of λs, leads to an estimator for $\hat{\Gamma}$ of the parameters $g(\sigma_u^2, \Gamma)$, for example by least squares. Simulation results indicate that $\lambda \in (0.5, 1, 1.5, 2)$ is a good choice in most cases.

Extrapolation step

The approximated function $g(\sigma_u^2, \hat{\Gamma})$ is extrapolated back to the case of no measurement error and so the SIMEX estimator is defined by $\hat{\beta}_{simex} := g(0, \hat{\Gamma})$, which corresponds to $\lambda = -1$.

Variance estimation

The ease of getting corrected parameter estimates is somewhat offset by the complexity of the calculation of the parameter's standard error. With its simulation character it is a natural candidate for the bootstrap. Although this is a valid method for obtaining standard errors, it is rather time consuming and for complexs not feasible. There are two methods for the estimation of standard errors which have a smaller computational burden. The jackknife method was developed for the SIMEX method by Stefanski and Cook (1995) and an asymptotic approach based on estimation equations was developed by Carroll et al. (1996) for the SIMEX method. We implemented the SIMEX method by Stenfanski and Cook.

3. Results

The computations are performed with "lm" and "simex" functions in R version 3.1.3. Table 3.1 shows the characteristics of subjects. Mean age (s.d.) is 41.15 (11.73) and female (65.40%) has higher proportion than male (34.60%). Mean SBP and DBP are 110.27 and 73.36, respectively and mean BMI is 22.63. Majority of subjects have drinks (85.60%), feel stress (97.70%) and have physical activity (94.60%).

Table 3.1 Characteristics of subjects (N=410)

Variables	Mean	S.D.	Median	Variables	n (%)
Age (yr)	41.15	11.73	41.00	Male	142 (34.60)
SBP ($mmHG$)	110.27	11.50	109.50	Family history (Yes)	159 (38.80)
DBP ($mmHG$)	73.36	8.61	73.00	Alcohol (Yes)	351 (85.60)
BMI (kg/m^2)	22.63	3.18	22.34	Stress (Yes)	380 (97.70)
Sodium (g/day)	3.66	1.63	3.43	Physical activity (Yes)	388 (94.60)
Potassium (g/day)	3.16	1.20	3.01	*Alcohol (Yes): 1 per week or more	
Vitamin D (ng/mL)	14.78	5.01	41.03	*Stress (Yes): feel a little bit or more	
Smoking ($sticks/day$)	2.28	5.58	0.00	*Physical activity (Yes): 1 per week or more	

Fonseca *et al.* (2010) showed the standard deviation of measurement errors for BMI. From this study, we have used $\sigma_\eta = 2.0$ for BMI in Model 2. Table 3.2 and 3.3 indicate the estimated coefficients $\hat{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_{11})^T$ and 95% confidence intervals (C.I.) of $\hat{\boldsymbol{\beta}}$ for each model. The estimated coefficients are very similar and except for BMI that was considered to measurement error covariate. The estimated coefficient of BMI for model 2 is larger than model 1. In previous paper for Korean, over the BMI range 25-31 each BMI unit was associated with a difference of 1.0 mmHg in diastolic blood pressure. Over the BMI range 16-25 each BMI unit was associated with a difference of 0.89 mmHg in diastolic blood pressure (Jones *et al.*, 1994).

Also, we calculate AIC (Akaike information criterion) and BIC (Bayesian information criterion) which are a criterion for model selection among models. The model with the lowest AIC and BIC are preferred and BIC is closely related to the AIC based on the likelihood function. Both AIC and BIC of Model 2 is smaller than Model 1 for SBP and DBP. Figure 3.1 is the scatter plot of outcome variable and covariate, and the real line (——) and dashed line (- - -) are the fitted lines by Model 1 and 2, respectively.

Table 3.2 Results for SBP

Covariate	Model 1 (without measurement error)			Model 2 (with measurement error)		
	$\hat{\beta}$	s.e.	95% C.I.	$\hat{\beta}$	s.e.	95% C.I.
BMI	0.722	0.167	(0.395) - (1.048)	1.024	0.225	(0.583) - (1.465)
Age	0.266	0.045	(0.177) - (0.355)	0.254	0.046	(0.164) - (0.344)
Sodium	0.001	0.001	(0.000) - (0.002)	0.001	0.001	(0.000) - (0.002)
Potassium	-0.002	0.001	(-0.003) - (-0.001)	-0.002	0.001	(-0.003) - (-0.001)
Vitamin D	-0.091	0.103	(-0.294) - (0.111)	-0.086	0.103	(-0.289) - (0.117)
Smoking	-0.106	0.101	(-0.304) - (0.092)	-0.114	0.102	(-0.313) - (0.086)
Sex(Male)	-5.489	1.225	(-7.890) - (-3.088)	-5.049	1.246	(-7.491) - (-2.608)
Alcohol(Yes)	-1.105	1.451	(-3.949) - (1.739)	-1.158	1.445	(-3.990) - (1.674)
Stress(Yes)	-0.357	1.951	(-4.181) - (3.468)	-0.298	1.954	(-4.127) - (3.531)
Family history (Yes)	3.711	1.041	(1.670) - (5.752)	3.590	1.044	(1.543) - (5.636)
Physical activity (Yes)	0.065	2.245	(-4.335) - (4.465)	0.095	2.245	(-4.304) - (4.495)
AIC	3097.530			3095.570		
BIC	2892.020			2888.980		

Table 3.3 Results for DBP

Covariate	Model 1 (without measurement error)			Model 2 (with measurement error)		
	$\hat{\beta}$	s.e.	95% C.I.	$\hat{\beta}$	s.e.	95% C.I.
BMI	0.575	0.127	(0.326) - (0.825)	0.824	0.170	(0.491) - (1.158)
Age	0.114	0.035	(0.046) - (0.181)	0.104	0.035	(0.036) - (0.172)
Sodium	0.001	0.000	(0.000) - (0.002)	0.001	0.000	(0.000) - (0.002)
Potassium	-0.002	0.001	(-0.003) - (-0.001)	-0.002	0.001	(-0.003) - (0.000)
Vitamin D	-0.023	0.079	(-0.178) - (0.132)	-0.026	0.079	(-0.181) - (0.128)
Smoking	-0.197	0.077	(-0.348) - (-0.046)	-0.209	0.077	(-0.360) - (-0.058)
Sex(Male)	-5.556	0.935	(-7.388) - (-3.723)	-5.220	0.950	(-7.083) - (-3.357)
Alcohol(Yes)	-1.423	1.107	(-3.594) - (0.748)	-1.378	1.108	(-3.549) - (0.794)
Stress(Yes)	-1.260	1.489	(-4.179) - (1.659)	-1.166	1.487	(-4.080) - (1.748)
Family history (Yes)	0.956	0.795	(-0.602) - (2.514)	0.810	0.794	(-0.746) - (2.365)
Physical activity (Yes)	2.148	1.714	(-1.211) - (5.506)	2.155	1.711	(-1.199) - (5.509)
AIC	3149.740			3147.780		
BIC	2944.230			2941.190		

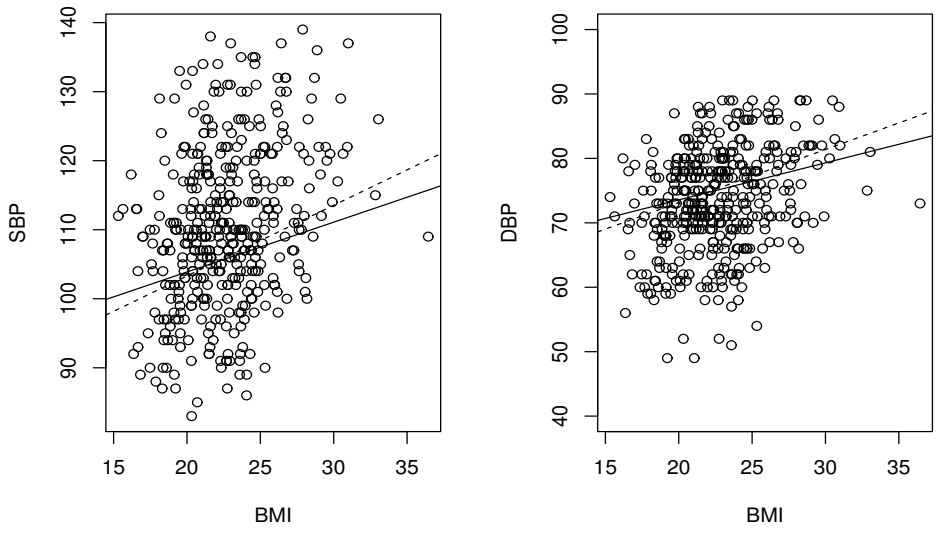

Figure 3.1 scatter plot; Model 1 : ——, Model 2 : - - - -,

4. Discussion

When the variable has measurement error, the estimated coefficient from model without measurement error concept should has a bias. We have compared the results of two models that the general linear regression model and measurement error models in covariate. When we consider measurement error covariate, the estimated coefficient of the measurement error covariate are different although confidence intervals are overlap each other. And AIC and BIC of measurement error model are smaller than the general linear regression model. So, if the covariate has measurement error we need to consider measurement error models.

In this paper, we only consider BMI as measurement error covariate but others covariates like the amount of Vitamin D, Sodium and Potassium also could be measurement error covariates. Also, outcome variables (SBP and DBP) could have measurement error. Measurement error model in outcome and covariate is developed by Deming (1943). But we have not conducted that because Deming regression was not allowed to perform multiple regression yet. So, we need to develop multiple measurement error models in outcome and covariate.

And, the KNHANES used a stratified, multistage probability sampling design to select household units. We need to use weights, strata, and primary sampling unit information provided in the KNHANES public use dataset in order to compute descriptive statistics in terms of general Korean population. But there are no models with measurement error considering the complex survey design. Therefore, we can also develop measurement error models with the complex survey data.

References

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of American Statistical Association*, **89**, 1314-1328.

Deming, W. E. (1943). *Statistical adjustment of data.* Wiley, New York.

Fuller, W. A. (1987). *Measurement Error Models*. Dekker, New York.

Fonseca, H., Silva, A. M., Matos, M. G., Esteves, I., Costa, P., Guerra, A. and Gomes-Pedro, J. (2010). Validity of BMI based on self-reported weight and height in adolescents. *Acta Paediatrica*, **99**, 83-88.

Hyslop, D. R. and Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *American Statistical Association Journal of Business and Economic Statistics*, **19**, 475-481.

Jones, D. W., Kim, J. S., Andrew, M. E., Kim, S. J. and Hong, Y. P. (1994). Body mass index and blood pressure in Korean men and women: the Korean National Blood Pressure Survey. *Journal of Hypertnesion*, **12**, 1433-1437.

Korea Centers for Disease Control and Prevention. (2009). *The Fourth Korea National Health and Nutrition Examination Survey IV*, Korea Centers for Disease Control and Prevention, Seoul, Korea.

Korea Centers for Disease Control and Prevention. (2013). *Statistics on the Sixth Korea National Health and Nutrition Examination Survey (KNHANES IV)*, Osong, Korea.

Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, **91**, 242-250.

Stefanski, L. A. and Cook, J. R. (1995). The measurement error jackknife. *Journal of the American Statistical Association*, **90**, 1247-1256.

Jang, Y. and Yoo, H. (2013). Gender differences of health behaviors and quality of life of Koreans with asthma. *Journal of Nursing*, **3**, 420-425.