International Journal of Epidemiology, 2021, 346–351
doi: 10.1093/ije/dyaa260
Advance Access Publication Date: 22 December 2020
Education Corner

Education Corner

Reflection on modern methods: demystifying robust standard errors for epidemiologists

Mohammad Ali Mansournia,[1]*† Maryam Nazemipour,[2†] Ashley I Naimi ,[3] Gary S Collins[4,5] and Michael J Campbell[6]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, [2]Psychosocial Health Research Institute, Iran University of Medical Sciences, Tehran, Iran, [3]Department of Epidemiology, Emory University, Atlanta, GA, USA, [4]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK, [5]Oxford University Hospitals NHS Foundation Trust, Oxford, UK and [6]ScHARR, University of Sheffield, Sheffield, UK

†These authors contributed equally to this work.

*Corresponding author. Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO Box 14155–6446, Tehran, Iran. E-mail: mansournia_ma@yahoo.com, mansournia_ma@sina.tums.ac.ir

Accepted 27 November 2020; editorial decision 22 November 2020

Abstract

All statistical estimates from data have uncertainty due to sampling variability. A standard error is one measure of uncertainty of a sample estimate (such as the mean of a set of observations or a regression coefficient). Standard errors are usually calculated based on assumptions underpinning the statistical model used in the estimation. However, there are situations in which some assumptions of the statistical model including the variance or covariance of the outcome across observations are violated, which leads to biased standard errors. One simple remedy is to use *robust standard errors*, which are robust to violations of certain assumptions of the statistical model. Robust standard errors are frequently used in clinical papers (e.g. to account for clustering of observations), although the underlying concepts behind robust standard errors and when to use them are often not well understood. In this paper, we demystify robust standard errors using several worked examples in simple situations in which model assumptions involving the variance or covariance of the outcome are misspecified. These are: (i) when the observed variances are different, (ii) when the variance specified in the model is wrong and (iii) when the assumption of independence is wrong.

Key words: Robust standard error, model-based standard error, heteroscedasticity, clustering

Introduction

All statistical analyses are based on a statistical model often involving one or more quantities in the population, known as *parameters*.[1] The model may not always be explicit, but it is always present. As a simple example, most statistical tests (e.g. the independent *t*-test) are based on

Key Messages

- The standard error of an estimate can be derived using various methods. The most common approach is based on assumptions underpinning the statistical model used in the estimation.
- There are situations in which assumptions of the statistical model are violated leading to biased standard errors. One simple remedy is to use robust standard errors.
- Robust standard errors can be used when certain model assumptions involving the variance or covariance of the observations are misspecified. Common examples include unequal variances across observations, using a Poisson distribution instead of a binomial distribution, and clustered data.

models that assume independent and identically distributed observations. In practice, parameter estimates (e.g. mean differences) will vary from one sample to the next. The variation in estimates across multiple samples is quantified by the *standard error*, which is simply the standard deviation of the estimates in hypothetical repeated samples of the population.[2] Standard errors can be derived using various methods. The most common approach is based on the underlying model—i.e. to assume that sampling variation in the parameter estimates is fully captured by the statistical model. However, when the assumptions of independence and identically distributed observations are violated, the model-based standard errors can be incorrect because they are calculated based on the assumptions intrinsic to the model being used. One simple remedy is to use *robust standard errors*, which are robust to violations of the statistical-model assumptions involving the variance or covariance of the outcome. The aim of this paper was to explain robust standard errors and their applications at an introductory level for epidemiologists.

Robust standard errors

Robust standard errors, also known as *Huber–White standard errors*,[3,4] essentially adjust the model-based standard errors using the empirical variability of the model *residuals* that are the difference between observed outcome and the outcome predicted by the statistical model. For example, in estimating the mean difference between two groups, the residuals are simply the difference between the observed outcome and the mean in each group.

The robust standard error is sometimes called the *sandwich standard error* due to its mathematical formulation: the 'bread' of the sandwich is the variance based on the statistical model and the 'meat' is the empirical variance based on the residuals. By adjusting the model-based standard errors, the robust standard errors can sometimes give a better assessment of the sample-to-sample variability of the estimates when the statistical-model assumptions are violated. We will discuss their use in three situations in which (i) the assumption of equal variances is wrong, (ii) the assumptions about the variance function is wrong and (iii) the assumption of the independence of the outcomes is wrong. Stata and R code for all analyses are presented in Supplementary Appendix 1, available as Supplementary data at *IJE* online.

Robust standard errors for heteroscedasticity

Robust standard errors can be used when the assumption of uniformity of variance, also known as *homoscedasticity*, in a linear-regression model is violated. This situation, known as *heteroscedasticity*, implies that the variance of the outcome is not constant across observations. Under the assumption of independence of observations, one remedy is using a robust standard error that is based on the square of the individual residuals.[4] However, this simple robust standard error underestimates the true variance in small samples or with leveraged data (which occurs when there are extreme values for the predictor variables). A modified conservative robust standard error that adjusts for the impact of small samples or leveraged data, known as HC3 (homoscedasticity consistent), is recommended.[5] Supplementary Appendix 2A, available as Supplementary data at *IJE* online, provides further details for various robust standard errors for heteroscedasticity in the simplest case of linear regression equivalent to an independent *t*-test.

To illustrate a robust standard error for heteroscedasticity, we use the data on child asthma to compare the mean dead space (ml) between asthmatics and non-asthmatics.[6] The dead spaces (ml) in people with asthma ($n_1 = 8$) were 43, 44, 45, 56, 56, 57, 58 and 64, and in people without asthma ($n_0 = 7$) were 31, 78, 79, 88, 92, 101 and 112. The mean dead space in asthmatic and non-asthmatic groups were 52.9 and 83.0 ml, respectively, and the mean

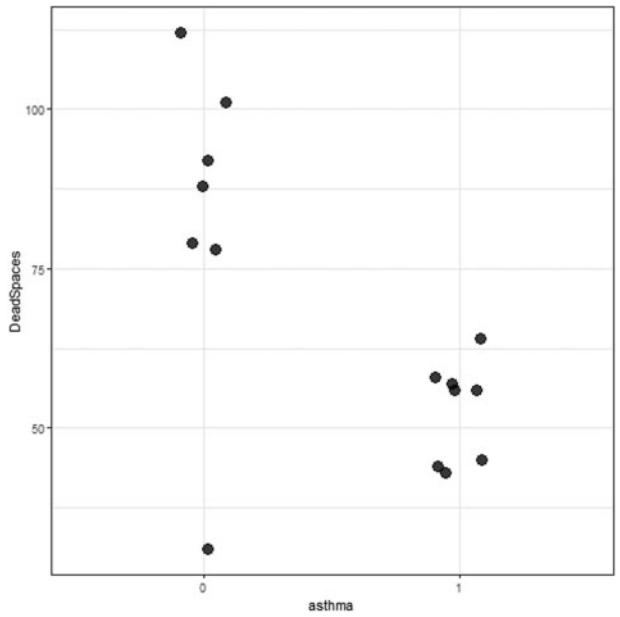

Figure 1 The scatter plot of dead space (ml), separately for people with asthma and people without asthma

difference was –30.1 ml. The standard deviations of the two groups were $S_1 = 7.8$ ml and $S_0 = 25.9$ ml, although S_0/S_1 was >2, suggesting unequal variances[7] (see Figure 1). The model-based standard error for the difference in mean dead space between asthmatics and non-asthmatics (based on S_p) was 9.6 and the two-sample *t*-test assuming equal variances gives $P = 0.008$ and 95% confidence interval (CI) of (–9.5, –50.8). The robust standard error HC3 was 11.0, which gives $P = 0.017$ and 95% CI of (–6.4, –53.8), suggesting that the *P*-value using a two-sample *t*-test assuming equal variances is too small and the resulting 95% CI is too narrow. In fact, the model-based standard error was 13% smaller than the robust standard error. It is important to note that the main concern in this example was heteroscedasticity and a robust standard error was used to handle this problem but not the violation-of-normality assumption (the *P*-value of the Shapiro–Wilk normality test was 0.54).

Robust standard errors for an incorrect variance function

Robust standard errors can also be used when the variance function is misspecified. Usually with a binary outcome, one would use logistic regression. However, this results in an odds ratio and one may wish to estimate a risk ratio, as its interpretation is easier. One effect-size measure should not be interpreted as if it is another one and, in particular, the odds ratio is not generally a valid estimate of the risk ratio. The natural model for estimating the risk ratio is the log-binomial regression model.

This is given by $\log(\pi) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$, where the predictor variables are X_1 to X_p with coefficients β_1 to β_p. Whereas the left-hand side of the log-binomial regression model is the logarithm of risk (π), which takes a negative value, the right-hand side of the model is the linear predictor, which is unbounded. When the linear predictor is large (which can occur with many predictors), it can yield a risk that is greater than one. Therefore, the log-binomial regression suffers from a structural problem that can result in non-convergence of the model and failure to estimate the adjusted risk ratio.[8–10]

One remedy is using the log-Poisson regression model whose left-hand side is the logarithm of mean and unbounded.[11] The beta-coefficient from the log-Poisson regression model is a valid estimate of the adjusted log-risk ratio due to the resemblance of the functional form of log-binomial and log-Poisson regression. However, the standard error of the estimate will be overestimated. To see why, note that, whereas the variance of a binomial outcome is $\pi(1-\pi)$, the variance is equal to the mean, i.e. π in the Poisson distribution. The robust standard error can be used to correct the standard error obtained from the Poisson model. In summary, to estimate the adjusted risk ratio, we use a log-Poisson regression model instead of a log-binomial regression model along with robust standard error.[11]

We illustrate the application of robust standard error using an unadjusted (crude) risk ratio for the study of the association between printers' vs farmers' wives and breastfeeding for <3 vs >3 months.[6] The data have been displayed in the BMJ statistics at square one.[7] In this study, the wives of printers were considered as the exposed and the wives of farmers as the unexposed. The outcome was breastfeeding for <3 months (the reference level was breastfeeding for >3 months). There were 50 printers' wives, of whom 36 breastfed for <3 months. There were 55 farmers' wives, of whom 30 breastfed for <3 months. The risk ratio for breastfeeding for <3 months is $\frac{36/50}{30/55} = 1.32$ for printers' wives relative to farmers' wives. A log-binomial model gives an estimated risk ratio of 1.32 (95% CI: 0.98, 1.78; $P = 0.07$). The model-based standard error, an approximate large-sample estimate of the standard error of the logarithm of risk ratio, was 0.15.

Note that a log-binomial model converges in this example, as there is just one binary predictor in the model. A log-Poisson regression model yields a similar risk ratio but wider CI and larger *P*-value (1.32 with 95% CI of 0.81, 2.14; $P = 0.26$). In fact, the reported model-based standard error for the log-risk ratio (based on the Poisson distribution) was increased to be 0.247. We can use an approximate robust standard error taking into account biases due to small samples and leveraged data known as HC3 to correct the overestimation.

Fitting a log-Poisson regression model with robust standard error HC3 gives the estimate of 1.32 (95% CI: 0.98, 1.78; $P = 0.07$) for the risk ratio. The robust standard error estimate was 0.152, suggesting a 62% overestimation for the model-based standard error. See Supplementary Appendix 2B, available as Supplementary data at *IJE* online, for the formulae for HC3 and several robust standard errors for the log-Poisson regression model with just one binary predictor.

Robust standard errors for clustering

The so-called *cluster-robust standard error* is a generalization of the robust standard error for clustered data, e.g. cluster-randomized-trial data in which treatments are randomly assigned to clusters of participants (e.g. hospitals)[12,13] or repeated outcome measurements in longitudinal data in which each unit is a cluster of observations over time.[14] It does not make any assumptions about independence within a cluster but does assume between-cluster independence and so is appropriate for the analysis of clustered data. The cluster-robust standard error is based on the cluster-level residuals, which are simply the linear combination of individual residuals in each cluster and use the empirical variability of the cluster-level residuals to adjust the biased model-based standard error ignoring clustering.[15] Using the latter will lead to CIs that are too narrow and *P*-values that are too small.

We illustrate cluster-robust standard errors using the following cluster-randomized-trial example in which 10 practices were randomly assigned to 2 treatment groups (patient-centred care and normal care) and body mass index (BMI) at Year 1 was assessed as the outcome (**Table 1**).[6]

A measure of clustering is the intra-cluster correlation[16] coefficient, which is the proportion of the total variance explained by cluster membership, i.e. the between-cluster variance divided by the sum of the between-cluster variance and the within-cluster variance. Using a one-way analysis of variance[17] of BMI over practice, we can verify that the intra-cluster correlation-coefficient estimate is 0.87, indicating high levels of clustering by practice.

The mean BMI (kg/m^2) in treatment Groups 1 and 0 were 28.81 and 28.39 kg/m^2, respectively. A two-sample *t*-test (assuming equal variances) that ignores clustering by practice gives a mean difference estimate of 0.42 kg/m^2 with a 95% CI of (–3.58, 4.42) and $P = 0.83$ based on the ordinary standard error estimate of 1.90.

The cluster-robust standard error with a small-sample adjustment for both cluster and individual (Supplementary Appendix 2C, available as Supplementary data at *IJE* online)[18] was 2.68, which gives 95% CI of (–5.64, 6.48) and $P = 0.88$. Compared with a cluster-robust standard error, the model-based standard error was underestimated by 29%.

Table 1 A data example of a cluster randomized trial

Subject	BMI (kg/m^2)	Treatment	Practice
1	26.2	1	1
2	27.1	1	1
3	25.0	1	2
4	28.3	1	2
5	30.5	1	3
6	28.8	1	4
7	31.0	1	4
8	32.1	1	4
9	28.2	1	5
10	30.9	1	5
11	37.0	0	6
12	38.1	0	6
13	22.1	0	7
14	23.0	0	7
15	23.2	0	8
16	25.7	0	8
17	27.8	0	9
18	28.0	0	9
19	28.0	0	10
20	31.0	0	10

Discussion

Robust standard errors can be used to adjust model-based standard errors to allow for certain violations of the model assumptions. We have illustrated a few examples of using robust standard errors in simple cases in which there is one binary predictor, although they can be used in regression models with many covariates, as well as models not considered here such as logistic regression or Cox regression. Robust standard errors can also be used when the mechanism of data generation does not follow a theoretical distribution, e.g. if there are sampling weights or inverse probability-of-treatment weights.[14,19,20]

Some caution is warranted when using robust standard errors. First, using the robust standard error when the model assumption is not violated will lead to less precise estimates and wider confidence intervals than when using the valid-model-based standard error. Second, robust standard errors perform poorly in small sample sizes (where the sample size refers to the number of clusters for cluster-robust standard errors) than the model-based standard errors, especially with non-linear models such as

log-Poisson and logistic regression, as they are then only approximations. Third, applying robust standard errors is not the only method to take into account violations of statistical-model assumptions. One can derive valid standard errors using more elaborate models that account for heteroscedasticity or clustering. For example, one can use inverse-variance (precision) weighting to accommodate unequal variances or random-effect models to account for clustering. Generalized estimating equations (GEEs)[21] use not only a working correlation structure to account for clustering, but also a cluster-robust standard error to adjust for errors in the working correlation structure used. In clustered data, GEE and random-effect models are more efficient than ordinary regression models with robust standard errors (such as illustrated above) if the model correlation assumption is correct. An alternative to robust and cluster-robust standard errors is the bootstrap, which may be preferred in small sample sizes.[6,22] Forth, it is important to note that, depending on the method of adjustment for small samples and leveraged data, the same robust variance estimators in the same data set may not return the same results in different statistical software programs (see Supplementary Appendices 1 and 2, available as Supplementary data at *IJE* online). Finally, we warn that using a robust standard error does not make an analysis 'robust' to all modelling assumptions.[6] When a regression model is used to estimate marginally adjusted treatment effects,[23] robust standard errors can improve variance estimation over model-based approaches.[24,25] However, robust standard errors are less useful when estimating conditionally adjusted effects with a seriously misspecified regression model, where one would obtain accurate standard errors to a mostly meaningless parameter.[26] Robust standard errors (also referred to as sandwich or Huber–White standard errors) are commonly encountered in modern epidemiologic analyses. However, their precise form, strengths and limitations are not well understood by the broader epidemiologic community. We have provided an overview of what robust standard errors are and how they can be used to overcome problems encountered with more traditional model-based approaches. Researchers should carefully consider when robust standard errors can be useful and when they should be avoided. Simulation studies are still needed to compare the different robust standard errors presented in Supplementary Appendix 2, available as Supplementary data at *IJE* online, considering several factors such as sample size and variance ratio, but they are beyond the scope of this educational paper.

Supplementary data

Supplementary data are available at *IJE* online.

Conflict of interest

None declared.

References

1. Altman DG, Bland JM. Statistics notes variables and parameters. *BMJ* 1999;**318**:1667.
2. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ* 2005;**331**:903.
3. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967*. University of California Press, 1967; 221–33.
4. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980;**48**:817–38.
5. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *Am Stat* 2000;**54**:217–24.
6. Campbell MJ. *Statistics at Square Two: understanding Modern Statistical Applications in Medicine*. Oxford, Blackwell, 2006, p. 113.
7. Campbell MJ, Swinscow TDV. *Statistics at Square One*. Chichester, John Wiley & Sons, 2011.
8. Naimi A, Whitcomb B. Estimating risk ratios and risk differences using regression. *Am J Epidemiol* 2020;**189**:508–10.
9. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 2004;**160**:301–05.
10. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerg Themes Epidemiol* 2013; **10**:14.
11. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;**159**:702–06.
12. Mansournia MA, Altman DG. Some methodological issues in the design and analysis of cluster randomised trials. *Br J Sports Med* 2019;**53**:573–75.
13. Campbell MJ, Walters SJ. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester, Wiley, 2014.
14. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. *BMJ* 2017;**359**:j4587.
15. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000;**56**:645–46.
16. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998;**316**:1455–60.
17. Altman DG, Bland JM. Statistics Notes: Comparing several groups using analysis of variance. *BMJ* 1996;**312**:1472–73.
18. StataCorp L. *Stata 16 Base Reference Manual*. College Station, TX: Stata Press, 2019.
19. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;**352**:i189.
20. Mansournia MA, Danaei G, Forouzanfar MH *et al.* Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: analysis with marginal structural models. *Epidemiology* 2012;**23**:631–40.
21. Hanley JA, Negassa A, MDd E, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;**157**:364–75.

22. Bland JM, Altman DG. Statistics notes: bootstrap resampling methods. *BMJ* 2015;**350**:h2622.
23. Robins JM, Hernán MA, Brumback B. *Marginal structural models and causal inference in epidemiology*. Epidemiology 2000;**11**: 550–60.
24. Maldonado G, Greenland S. Interpreting model coefficients when the true model form is unknown. *Epidemiology* 1993;**4**:310–18.
25. Greenland S, Maldonado G. The interpretation of multiplicative-model parameters as standardized parameters. *Stat Med* 1994;**13**: 989–99.
26. Freedman DA. On the so-called 'Huber sandwich estimator' and 'robust standard errors'. *Am Stat* 2006;**60**:299–302.