Article:

Mansournia, M.A., Nazemipour, M., Naimi, A.I. et al. (2 more authors) (2021) Reflections on modern methods: demystifying robust standard errors for epidemiologists. International Journal of Epidemiology, 50 (1). pp. 346-351. ISSN 0300-5771

https://doi.org/10.1093/ije/dyaa260

Demystifying robust standard errors

Abstract

A standard error is a measure of uncertainty around a sample estimate (such as the mean of a set of observations, or a regression coefficient). All estimates have uncertainty due to sampling variability. Standard errors are usually calculated based on assumptions underpinning the statistical model used in the estimation. However, there are situations where some assumptions of the statistical model including the variance or covariance of the outcome across observations are violated which leads to incorrect 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), however 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 where model assumptions involving the variance or covariance of the outcome is misspecified.

All statistical analyses are based on a statistical model 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., independent t test) are based on 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—that is, to assume sampling variation in the parameter estimates is fully captured by the statistical model. However, when certain model assumptions 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 certain violations of the statistical model.

Robust standard errors

Robust standard errors, also known as *Huber-White standard errors*, essentially adjust the model-based standard errors using the empirical variability of the model *residuals* which 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 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 where 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 the Appendix 1.

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 outcome's variance is not constant across observations. In the case of an independent t test (a special case of linear regression the independent variable simply taking values 0 and 1 for the two groups), the model-based standard error for the mean difference exploits the assumptions of independence of the observations and homogeneity of variance.

The model-based standard error equals $\surd(S^2{}_p(1/n_1 + 1/n_2))$ where $S^2{}_p$ is the pooled variance estimate i.e., the weighted average of sample variances $S^2{}_1$ and $S^2{}_2$: $(S^2{}_1(n_1 – 1) + S^2{}_2(n_2 – 1)) /(n – 2)$ where $n=n_1+n_2$. However, under a heteroscedastic outcome distribution, the model-based standard error would be biased, resulting in incorrect test statistics, p-values, and confidence intervals. This bias is especially important when the sample sizes in the two groups are substantially different, as pooling the variances suggests that the smaller sample contributes less to the standard error, though it really contributes more to the variation of the mean difference. One remedy is to use following robust standard error referred to as, HC0 (HC=Heteroscedasticity consistent), which is based on the square of the residuals (as the mean of the residuals is zero one estimate of the variance, σ^2 for an observation is its squared residual):

$$HC0 = \sqrt{S_1^2\left[\frac{n_1 - 1}{{n_1}^2}\right] + S_2^2\left[\frac{n_2 - 1}{{n_2}^2}\right]} \qquad (eq\ 1)$$

The robust standard error HC0 is downwardly biased in small to moderately large samples. A simple bias adjustment multiplier is n/(n-1), which gives HC0m:

$$HC0m = \sqrt{\frac{n}{n-1}\left(S_1^2\left[\frac{n_1 - 1}{{n_1}^2}\right] + S_2^2\left[\frac{n_2 - 1}{{n_2}^2}\right]\right)} \qquad (eq\ 2)$$

Under the assumption of homoscedasticity and Normality and for more than one parameter estimated, n/(n-k) is a better multiplier than n/(n-1) where k is the number of parameters estimated.

The robust standard error using the multiplier n/(n-k) is known as HC1 (k=2 in our example because the mean in each of two groups were estimated):

$$HC1 = \sqrt{\frac{n}{n-2}\left(S_1^2\left[\frac{n_1-1}{{n_1}^2}\right] + S_2^2\left[\frac{n_2-1}{{n_2}^2}\right]\right)} \qquad (eq\ 3)$$

Another problem with HC0 which also holds for its modifications including HC0m and HC1 is that it underestimate the true variance when there are observations with high leverage i.e., with an extreme values of the predictor variables (X). The reason is that the squared residuals for high leverage observations tend to be small and hence the true variance using HC0, HC0m and HC1 which are based on squared residuals will be underestimated. A remedy is to inflate the squared residuals by (1-h) where h is the leverage. Using this modification and noting that the leverage in our example equals 1/n$_1$ for the first sample and 1/n$_2$ for the second, the robust standard error will be the following:

$$HC2 = \sqrt{\left(\frac{S_1^2\left[\frac{n_1-1}{{n_1}^2}\right]}{1-\frac{1}{n_1}}\right) + \left(\frac{S_2^2\left[\frac{n_2-1}{{n_2}^2}\right]}{1-\frac{1}{n_2}}\right)} \qquad (eq\ 4)$$

HC2 is unbiased under the assumption of homoscedasticity because expected value of the squared residuals equals σ^2(1-h). Note that the average value of h is k/n so that 1-h has an average value of (n-k)/n .Thus HC0m divides the squared residuals by the average of 1-h. Another robust standard error formula reduces the impact of leverage data on the inference even more than HC2 using the squared residuals divided by $(1\text{-h})^2$:

$$HC3 = \sqrt{\left(\frac{S_1^2\left[\frac{n_1-1}{{n_1}^2}\right]}{(1-\frac{1}{n_1})^2}\right) + \left(\frac{S_2^2\left[\frac{n_2-1}{{n_2}^2}\right]}{(1-\frac{1}{n_2})^2}\right)} \qquad (eq\ 5)$$

HC3 is the most conservative, and is recommended when the sample size is less than 250.(3) However, with large sample sizes in both groups, all five formulas simplify to HC2 which is the same as the standard error used in the Normal approximation i.e., $S^2{}_1/n_1 + S^2{}_2/n_2$. Recently several other robust standard error estimators, HC4, HC4m and HC5, have been proposed which are based on scaling the squared residuals by specific powers of (1-h) and have been shown to have better performance than HC3 using leverage data. The formulas for the robust standard error HC4, HC4m and HC5 are presented in the Appendix 2. We note that using a robust standard error is not a remedy for the violation of Normal distribution assumption for the outcome which sometimes produces unequal variances.(4)

To illustrate the impact of different standard errors, we use the data on child asthma to compare the mean dead space (ml) between asthmatics and non-asthmatics.(4) 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 ($n_2 = 7$) were 31, 78, 79, 88, 92, 101, and 112. The mean dead space in asthmatic and non-asthmatic groups are 52.9 ml and 83.0 ml, respectively and the mean difference is 30.1 ml. The standard deviations of the two groups are S_1= 7.8 ml and S_2 = 25.9 ml, however S_2/S_1 is much larger than 1, suggesting unequal variances (see Figure 1). The model-based standard error for the difference in mean dead space between asthmatics and non-asthmatics (based on S_p) is 9.6, and the two-sample t test assuming equal variances gives P=0.008 and 95% CI: (9.5, 50.8). The robust standard errors using HC0, HC0m, HC1, HC2, HC3, HC4, HC4m and HC5 are 9.41, 9.74, 10.11, 10.16, 10.96, 10.21, 11.01, and 9.80. Using the robust standard error HC3 gives P=0.017 and 95% CI: (6.4, 53.8) suggesting that P-value using two-sample t test assuming equal variances is too small and the resulting 95% confidence interval is too narrow.

Robust standard errors for incorrect variance function

Robust standard errors can also be used when the variance function is misspecified. Usually with a binary exposure on 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. If the outcome is not rare we cannot assume the odss ratio is a good estimate of the relative riskThe natural model to use is the log-binomial regression model. While the left-side of the log-binomial regression model is the logarithm of risk (π) which takes a negative value, the right-side of the model includes the sum of coefficient(s)×predictor(s) which is unbounded. When there are many predictors, or when the sum of the coefficients×predictors is large, the sum of these values can yield a risk that is greater than one. Therefore, the log-binomial regression suffers from a structural problem which results in nonconvergence of the model and failure to estimate the adjusted risk ratio.(5-7)

One remedy is to use a log-Poisson regression model. Hence, it is possible to so the left-side which is the logarithm of mean is unbounded.(8) The point estimate of the adjusted risk ratio from the log-Poisson regression model is valid as the latter model resembles log-binomial regression. But because the mean of Poisson distribution can take any positive number (i.e., is not bounded between 0 and 1), the standard error of the estimate will be overestimated. To see why, note that the variance of a binomial outcome is $\pi(1-\pi)$. However, using the Poisson distribution, we assume that the variance is equal to the mean i.e., π. Instead, 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 log-Poisson regression model instead of log-binomial regression model along with robust standard error.(8)

We illustrate the application of robust standard error using unadjusted (crude) risk ratio for the study of the association between printers' versus farmers' wives and breastfeeding less than 3 months versus more than 3 months.(4) The data have been displayed in the BMJ statistics at square one.(9) In this study, the exposure $X = 1$ for wives of printers and 0 for wives of farmers. The outcome $Y = 1$ for breastfeeding for less than 3 months and 0 for breastfeeding for more than 3 months. There were $n_1 = 50$ printers' wives, of whom $a = 36$ breastfed for less than 3 months. There were $n_2 = 55$ farmers' wives, of whom $c = 30$ breastfed for less than 3 months. The total sample size $n=n_1+n_2$ was 105. The risk ratio for breastfeeding less than 3 months is 1.32 for printers' wives relative to farmers' wives. A log-binomial model regressing Y against X gives an estimated risk ratio of 1.32 (95% CI: 0.98, 1.78; P=0.07). The model-based standard error is 0.151. Furthermore, using the square root of $(1/a + 1/c - 1/n_1 – 1/n_2)$ provides an approximate large-sample estimate of the standard error of the logarithm of risk ratio.

Note that a log-binomial model converges in this example, as there is just one binary predictor (X) in the model. A log-Poisson regression model yields a similar risk ratio (1.32 with 95% CI of 0.81, 2.14; P=0.26). However, the reported model-based standard error for the log-risk ratio (based on the Poisson distribution), 0.247, is calculated as the square root of $(1/a + 1/c)$. This estimate of the standard error is clearly an overestimate leading to a P-value which is too large and 95% confidence interval which is too wide. We can use the same robust standard error formulas for heteroscedasticity mentioned above to correct the standard (see Appendix 2 for the formulas of HC4, HC4m and HC5):

$$HC0 = \sqrt{\left(\frac{1}{a} - \frac{1}{n_1}\right) + \left(\frac{1}{c} - \frac{1}{n_2}\right)} \qquad (eq\ 6)$$

$$HC0m = \sqrt{\frac{n}{n-1}\left(\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}\right)} \qquad (eq\ 7)$$

$$HC1 = \sqrt{\frac{n}{n-2}\left(\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}\right)} \qquad (eq\ 8)$$

$$HC2 = \sqrt{\left(\frac{\frac{1}{a} - \frac{1}{n_1}}{1 - \frac{1}{n_1}}\right) + \left(\frac{\frac{1}{c} - \frac{1}{n_2}}{1 - \frac{1}{n_2}}\right)} \qquad (eq\ 9)$$

$$HC3 = \sqrt{\left(\frac{\frac{1}{a} - \frac{1}{n_1}}{(1 - \frac{1}{n_1})^2}\right) + \left(\frac{\frac{1}{c} - \frac{1}{n_2}}{(1 - \frac{1}{n_2})^2}\right)} \qquad (eq\ 10)$$

Note that HC0 is the same as the model-based standard error of the logarithm of risk ratio. In this example all HCs estimate are equal in two decimals, 0.15 which is not supervising given the moderate sample size. For example fitting a log-Poisson regression model with robust standard error HC0m gives the estimate of 0.28 (95% CI: -0.02, 0.58; P=0.07) for log-risk ratio (1.32 with 95% CI of (0.98, 1.78) for risk ratio).

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),(10) or repeated outcome measurements in longitudinal data whereby each unit is a group of observations (cluster), where the outcomes within each unit are correlated.(11) A cluster-robust standard error does not make any assumptions about independence within cluster but does assume between-cluster independence and so is appropriate for analysis of clustered data. A cluster-robust standard error is based on the cluster-level residuals which are simply the sum of individual-residuals in each cluster. The cluster-robust standard error uses the empirical variability of the cluster-level residuals to adjust the biased ordinary standard error ignoring clustering. Using the ordinary standard error ignoring clustering will lead to confidence intervals which are too narrow and P-values which are too small.

We illustrate cluster-robust standard errors using the following cluster randomized trial example in which 10 practices were randomly assigned to two treatment groups (patient centred care and normal care) and BMI at year 1 was assessed as the outcome(4):

Subject	BMI (kg/m2)	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

A measure of clustering is the intra-cluster correlation(12) 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(13) of BMI over practice, we can verify that the intracluster correlation coefficient estimate is 0.87 indicating high levels of clustering by practice.

The mean BMI (kg/m^2) in treatment groups 1 and 2 are 28.81 kg/m^2 and 28.31 kg/m^2, respectively. A two-sample t test (assuming equal variances) which ignores clustering by practice gives a mean difference estimate of 0.42 kg/m^2 with 95% CI of (-3.6, 4.4) and P=0.83 based on the ordinary standard error estimate of 1.9. To estimate the cluster-robust standard error which accounts for clustering we should first calculate the cluster-level residuals. For example, the residuals for the first and second subjects in the first practice equals 26.2 – 28.81 = -2.61 and 27.1 – 28.81 = -1.71, respectively. The cluster-level residual for the first practice is the sum of individual-residuals: -2.61 + (-1.71) = -4.32. We can verify that the variance of cluster-level residuals in treatment groups 1 and 2 (S^2_1 and S^2_2) is 18.07 and 135.01, respectively. Assuming that there are n_1 subjects in m_1 clusters in the treatment group 1 and n_2 subjects in m_2 clusters in the treatment group 2, $n=n_1+n_2$ and $m=m_1+m_2$, and as said before, S^2_1 and S^2_2 represent the variance of cluster-level residuals, the cluster-robust standard error is simply the square root of $S^2_1[(m_1 – 1)/n_1^2] + S^2_2[(m_2 – 1)/n_2^2]$ with one of two possible small-sample adjustments, (asymptotic-like and regression-like formulas, respectively):

$$\sqrt{\frac{m}{m-1}\left(S_1^2\left[\frac{m_1-1}{{n_1}^2}\right]+S_2^2\left[\frac{m_2-1}{{n_2}^2}\right]\right)} \qquad (eq\ 11)$$

$$\sqrt{\frac{n-1}{n-2}\times\frac{m}{m-1}\times S_1^2\left[\frac{m_1-1}{{n_1}^2}\right]+S_2^2\left[\frac{m_2-1}{{n_2}^2}\right]} \qquad (eq\ 12)$$

The cluster-robust standard error using the equation 12 for the BMI data is 2.7 which gives 95% CI of (-5.6, 6.5) and P=0.88. It is noteworthy that in the absence of clustering $m_1=n_1$ and $m_2=n_2$ so that the equation 12 will reduce to HC1 in equation 3 for heteroscedasticity.

You might like to reference Zou and Donner (which I will attach) which nicely combines clustering and Poisson

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 use of robust standard errors in simple cases where there is one binary predictor, but 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 for example if there are sampling weights or inverse probability-of-exposure weights.(11, 14, 15)

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 estimate and wider confidence intervals than using the valid model based standard error. Second, robust standard errors perform poorly in small sample sizes (where sample size refers to the number of clusters for cluster-robust standard error) than the model based standard errors, especially with non-linear models such as logistic regression as they are then only approximations. In such cases, the bootstrap(4, 16, 17) may be a preferable alternative. Third, applying robust standard errors is not the only method to take into account for violations of statistical model assumptions. One can derive valid standard errors using more elaborate models which 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)(18) 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.

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.

Summary points

- 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 where assumptions of the statistical model are violated leading to incorrect standard errors. One simple remedy is using 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.

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

The Figure looks like there are three groups. I would plot the repeated points closer together.

References

1. Altman DG, Bland JM. Statistics notes Variables and parameters. *Bmj* 1999;318(7199):1667.
2. Altman DG, Bland JM. Standard deviations and standard errors. *Bmj* 2005;331(7521):903.
3. Long JS, Ervin LH. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* 2000;54(3):217-24.
4. Campbell MJ. *Statistics at square two: understanding modern statistical applications in medicine*. Blackwell; 2006.

5. Naimi A, Whitcomb B. Estimating Risk Ratios and Risk Differences Using Regression. *Am J Epidemiol* In Press.
6. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American journal of epidemiology* 2004;160(4):301-5.
7. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerging themes in epidemiology* 2013;10(1):14.
8. Zou G. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology* 2004;159(7):702-6.
9. Campbell MJ Swinscow . *Statistics at square one 11th ed*. Bmj London; 2009.
10. Mansournia MA, Altman DG. Some methodological issues in the design and analysis of cluster randomised trials. *Br J Sports Med* 2019;53(9):573-5.
11. Mansournia MA, Etminan M, Danaei G, et al. Handling time varying confounding in observational research. *Bmj* 2017;359:j4587.
12. Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *Bmj* 1998;316(7142):1455-60.
13. Altman DG, Bland JM. Statistics Notes: Comparing several groups using analysis of variance. *Bmj* 1996;312(7044):1472-3.
14. Mansournia MA, Altman DG. Inverse probability weighting. *Bmj* 2016;352:i189.
15. 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* 2012:631-40.
16. Bland JM, Altman DG. Statistics notes: bootstrap resampling methods. *bmj* 2015;350:h2622.
17. Naimi A, Zhong Y, Rudolph J. Bootstrap methods for confidence interval estimation with parametric and machine-learning based estimators. *Epidemiology* Under Review.
18. Hanley JA, Negassa A, Edwardes MDd, et al. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology* 2003;157(4):364-75.