Approximating Cohen's d from Regression Results

Maya B. Mathur[1,2*], *Peng Ding*[3], *Corinne A. Riddell*[4], *and Tyler J. VanderWeele*[1,5]

[1] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[2]Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA

[3]Department of Statistics, University of California at Berkeley, Berkeley, CA, USA

[4]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Quebec, CA

[5]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

∗: Corresponding author:
mmathur@stanford.edu
Quantitative Sciences Unit (c/o Inna Sayfer)
1070 Arastradero Road
Palo Alto, CA
94305

Type of manuscript: Research Letter

Keywords: software, code, sensitivity analysis, confounding

Running head: "Package and Website for E-Values"

Conflicts of interest: The authors declare that they have no conflicts of interest.

Reproducibility: No data analyses were conducted.

Source of funding: MM was supported by National Defense Science and Engineering Graduate Fellowship 32 CFR 168a. PD was supported by IES Grant R305D150040 from the Institute for Education Science and DMS grant 1713152 from the National Science Foundation. CAR received salary support from McGill University's Department of Epidemiology, Biostatistics, and Occupational Health. TVW was supported by NIH grant ES017876. The funders had no role in the design, conduct, or reporting of this research.

Main text

1. Intro paragraph
   (a) Describe E-value and why extend to regression
   (b) If you have raw data, could choose somewhere to dichotomize and use E-value for SMD or RR
   (c) We provide E-value for cases where only standard regression info is reported
   (d) Approach involves converting regression results to SMD with well-defined effect size (say why critical for E-value): show table with RR and lower and upper RR
2. β-amyloid example
   (a) .
3. Caveats
   (a) What to do with preventative
   (b) Multiple regression
   (c) Be careful with interpretation: depends on Δ
   (d) Interpretation: E-value for a Δ-unit increase in X and any dichotomization of Y

In general, we would recommend using $\Delta = 1$ in order to assess the E-value for the effect size corresponding directly to the regression coefficient, which represents a 1-unit contrast in X. However, if the units of X are very fine-grained (e.g., if X is blood pressure in mmHg), then a 1-unit increase may not be considered clinically meaningful, and a different choice of Δ may be used (e.g., $\Delta = 10$ to represent an increase in blood pressure of 10 mmHg), which is equivalent to rescaling the regression coefficient. It is imperative to report the choice of Δ if it is not taken to be 1, since it directly impacts the size and interpretation of the E-value analog.

β-amyloid example

```
##               point     lower     upper
## RR        0.3783315 0.2887908 0.4956347
## E-values 4.7272287        NA 3.4504987
```

```
##              point     lower    upper
## RR        1.037101 0.7406488 1.452211
## E-values 1.233257 1.0000000       NA
```

```
##              point     lower     upper
## RR        0.505352 0.3366446 0.7586062
```

```
## E-values 3.370546        NA 1.9658663

##              point     lower     upper
## RR       0.6880847 0.4813375 0.9836354
## E-values 2.2649736        NA 1.1466895
```

Appendix

Cover somewhere:

- Assumptions inherited from Chinn's conversion from d to log-OR: distrubtion of Y within each outcome group is logistic, but according to Chinn, this is basically the same as normality
- Describe what to do for preventive: just set delta to be negative?
- Say that, for univariable case, if SD of X isn't available, can set Delta = SD of X

Univariable regression

Lemma 1. *Under the standard OLS framework, suppose that* $Y = \beta_0 + \beta X + \epsilon$ *with* $X \amalg \epsilon$ *and* $E[\epsilon] = 0$*. Then* $\sigma^2_{Y|X} = \left(1 - \rho^2_{YX}\right) \sigma^2_Y$.

Proof.

$$\begin{aligned}
\sigma^2_Y &= E[\sigma^2_{Y|X}] + Var\left(E[Y|X]\right) \\
&= \sigma^2_{Y|X} + Var\left(\beta_0 + \beta X\right) && \text{(homoskedasticity)} \\
&= \sigma^2_{Y|X} + \beta^2 \sigma^2_X \\
&= \sigma^2_{Y|X} + \rho^2_{YX} \sigma^2_Y \\
\sigma^2_{Y|X} &= \left(1 - \rho^2_{YX}\right) \sigma^2_Y
\end{aligned}$$

□

Suppose that the effect size of interest is the increase in Y caused by a Δ-unit increase in X, and consider the Cohen's d associated with an increase of Δ units in X:

$$\begin{aligned}
d &= \frac{E[Y \mid X = c + \Delta] - E[Y \mid X = c]}{\sigma_{Y|X}} \\
&= \frac{\Delta \beta}{\sigma_Y \sqrt{1 - \rho^2_{YX}}} \\
&= \frac{\Delta \beta}{\sigma_Y \sqrt{1 - \frac{\beta^2 \sigma^2_X}{\sigma^2_Y}}} \\
&= \frac{\Delta \rho_{YX}}{\sigma_X \sqrt{1 - \rho^2_{YX}}}
\end{aligned}$$

An approximate standard error can be derived using the delta method, treating σ_X as known. Let $z^f = \text{arctanh}(\rho)$ be the Fisher-transformed correlation, which is approximately normal with variance $\frac{1}{N-3}$. Define the transformation:

$$
\begin{aligned}
g(z^f) = d &= \frac{\Delta \tanh\left(z^f\right)}{\sigma_X \sqrt{1-\tanh^2\left(z^f\right)}} \\
SE_d &\approx \sqrt{\text{Var}(z^f)}\left(g'\left(z^f\right)\right) \\
&= \frac{1}{\sqrt{N-3}} \times \frac{\Delta}{\sigma_X \sqrt{\text{sech}^2(z^f)}} \\
&= \frac{\Delta}{\sigma_X \sqrt{(N-3)\left(1-\rho_{XY}^2\right)}} \\
&= \frac{\Delta}{\sigma_X \sqrt{(N-3)\left(1-\beta^2 \frac{\sigma_X^2}{\sigma_Y^2}\right)}} \\
&= \frac{d}{\rho_{YX}\sqrt{N-3}}
\end{aligned}
$$

To obtain an approximate E-value, we can approximately convert the point estimate to a relative risk (VanderWeele and Ding 2017):

$$
RR \approx \exp\left(0.91 \times \frac{\Delta \rho_{XY}}{\sigma_X \sqrt{1-\rho_{XY}^2}}\right)
$$

Approximate confidence interval limits are (VanderWeele and Ding 2017):

$$
\begin{aligned}
RR_{lb} &\approx \exp\left(0.91 \times \frac{\Delta \rho_{XY}}{\sigma_X \sqrt{1-\rho_{XY}^2}} - 1.78 \times SE_d\right) \\
&= RR_{ub} \approx \exp\left(0.91 \times \frac{\Delta \rho_{XY}}{\sigma_X \sqrt{1-\rho_{XY}^2}} + 1.78 \times SE_d\right)
\end{aligned}
$$

Multivariable regression

Extend the regression model to include arbitrary measured covariates $\mathbf{Z}$:

$$
E\left[Y \mid X, \mathbf{Z}\right] = \beta_0 + \beta_X X + \boldsymbol{\beta}_{\mathbf{Z}}' \mathbf{Z}
$$

where $\boldsymbol{\beta}_Z$ denotes a p-vector of estimated coefficients for $\mathbf{Z}$. Let $R^2_{Y\sim X|Z}$ be the coefficient of partial determination of Y on X, controlling for $\mathbf{Z}$ (equivalently, the squared partial correlation). Then:

$$\begin{aligned}
R^2_{Y\sim X|Z} &= 1 - \frac{SSE_{full}}{SSE_{red}} \\
&\approx 1 - \frac{(N-p-2)\cdot \sigma^2_{Y|X,\mathbf{Z}}}{(N-2)\cdot \sigma^2_{Y|\mathbf{Z}}} \\
&\approx 1 - \frac{\sigma^2_{Y|X,\mathbf{Z}}}{\sigma^2_{Y|\mathbf{Z}}} && \text{(n » p)} \\
\sigma^2_{Y|X,\mathbf{Z}} &= \sigma^2_{Y|\mathbf{Z}}\left(1 - R^2_{Y\sim X|Z}\right)
\end{aligned}$$

where the second line follows from unbiasedness of the mean squared error for the error variance. Then, an approximate Cohen's d is:

$$\begin{aligned}
d &= \frac{E[Y \mid X = c + \Delta, \mathbf{Z}] - E[Y \mid X = c, \mathbf{Z}]}{\sigma_{Y|X,\mathbf{Z}}} \\
&= \frac{\Delta\beta}{\sigma_{Y|\mathbf{Z}}\sqrt{1 - R^2_{Y\sim X|Z}}} \\
&\geq \frac{\Delta\beta}{\sigma_Y\sqrt{1 - R^2_{Y\sim X|Z}}}
\end{aligned}$$

Because $\sigma_{Y|\mathbf{Z}}$ is not commonly reported, the final line provides a conservative lower bound on d using the more commonly reported σ_Y. Unlike in the univariable case, a simple relationship between β and $R^2_{Y\sim X|Z}$ is not available with additional distributional assumptions, so both quantities are needed to approximate d.

To estimate the standard error, we first assume the following approximate analogs to exact relationships in the univariable setting:

$$\begin{aligned}
\left(\beta\frac{\sigma_X}{\sigma_Y}\right)^2 &\approx R^2_{Y\sim X|Z} \\
\text{Var}(z^f_{YX\bullet Z}) &\approx \text{Var}(z^f_{YX}) = \frac{1}{N-3}
\end{aligned}$$

where $z^f_{YX\bullet Z} = \text{arctanh}\left(\sqrt{R^2_{Y\sim X|Z}}\right)$ is the Fisher-transformed partial correlation. Then, proceeding algebraically and applying the delta method as in the univariable case, we obtain:

$$SE_d = \frac{d}{\sqrt{R^2_{Y\sim X|Z}(N-3)}}$$

This approximate standard error appears to perform very well in simulations across a number of scenarios, including those with high correlations between **Z** and X and between **Z** and Y.

References

VanderWeele, Tyler J, and Peng Ding. 2017. "Sensitivity Analysis in Observational Research: Introducing the E-Value." *Annals of Internal Medicine* 167 (4). Am Coll Physicians: 268–74.