

Chapter 16 Solutions

Making Sense of Social Data

Mark Paradis, Ph.D.

September 2, 2026

These are the worked solutions to the *end-of-chapter* exercises for Chapter 16 of *Making Sense of Social Data*. Solutions to the exercises inside each section are in the book itself, printed immediately after the exercises they answer.

Do the exercise before you read the solution. A solution read first is just more text to read; a solution read second is a check on your reasoning, and only the second one teaches you anything.

The complete set for every chapter is at <https://www.markparadispolitics.com/books/making-sense-of-social-data/online/solutions/>.

Chapter 16 Simple Linear Regression

Solution to Exercise 16.E1. (a) The statistic is the slope over its standard error:

$$t = \frac{0.42}{0.15} = 2.80$$

with $df = n - 2 = 1,198$. Two degrees of freedom are spent, one on the intercept and one on the slope. The result is significant, $p = .005$.

(b) The critical value at this many degrees of freedom is essentially 1.960, so

$$0.42 \pm 1.960 \times 0.15 = 0.42 \pm 0.29 = [0.13, 0.71]$$

(c) Across these 1,200 adults, each additional year of education goes with 0.42 more hours of volunteering per month, with a 95 per cent interval from 0.13 to 0.71 hours.

The four parts are the amount (0.42 hours), the outcome (volunteering per month), the unit of the predictor (one year of education), and the population (these 1,200 adults). Dropping the last invites the reader to apply the finding to anyone.

(d) Without r^2 you cannot say how much of the variation in volunteering the model accounts for, and you cannot recover the standard error of the estimate — so you cannot say how wrong the model's predictions typically are. The slope and its interval describe the relationship; neither describes the fit.

Solution to Exercise 16.E2. (a) Study B. The standard error is $s_e / \sqrt{\sum (x_i - \bar{x})^2}$, and a single neighbourhood has far less variation in density than a whole province. With the same n and similar scatter, the smaller denominator gives the larger standard error — the effect measured in §16.3.1, where narrowing the age range multiplied the standard error by 3.5.

(b) Homogeneity helps with confounding and hurts precision, and the two have to be weighed rather than one asserted. Study B's cases are alike on the things a neighbourhood determines, which does remove some alternative explanations. But a slope is estimated from variation in the predictor, and Study B has removed most of it. Their sample is better placed to answer a question it has made itself unable to answer precisely.

(c) It confuses failing to detect a relationship with establishing there is none. Study B's design guarantees a large standard error, so a wide interval around zero was the likely outcome whether or not a relationship exists. The reportable finding is the interval: if it is narrow, the study has bounded the relationship; if wide, it has established nothing in either direction.

Solution to Exercise 16.E3. (a)

$$s_e \approx 12.0 \times \sqrt{1 - .09} = 12.0 \times 0.954 = 11.45 \text{ points}$$

(b) Guessing the mean for every student misses by about 12.0 points; using the regression and a student's sleep misses by 11.45. The improvement is 0.55 points, or 4.6 per cent.

(c) It depends on the claim. If “substantial” means the relationship is real and worth a sentence, an r^2 of .09 at a decent sample size can support that. If it means the model predicts a student's score usefully, no: the typical prediction is still wrong by 11.45 points.

What would settle it is the slope in the units of the two variables, with its interval. How many points is an hour of sleep worth? That is the question “substantial” is gesturing at, and r^2 cannot answer it.

Solution to Exercise 16.E4. (a) *Linearity*. The true relationship curves and a straight line has been fitted through it. This damages the *estimate*: the slope describes neither end of the range. Remedy: transform the predictor (§16.5.3) or fit a curve.

(b) *Constant variance*. The scatter grows with the fitted value. This damages the *standard error*, which uses one number for a quantity that varies. The slope itself is unaffected. Remedy: robust standard errors, or a transformation of the outcome.

(c) *Independence*. Consecutive cases resembling one another means the errors are related, which usually happens when the data are ordered in time or clustered in groups. This damages the *standard error*, and almost always downward, so the analysis reports more confidence than it has earned. Remedy: a method that models the grouping — beyond this book, and the reason `cma_panel` is absent from this chapter.

(d) *Nothing is at risk*. Right skew in the residuals violates normality, and at $n = 900$ the central limit theorem makes the sampling distribution of b_1 close to normal regardless. Neither the estimate nor the standard error is damaged. Remedy: none needed. The skew may still be worth a sentence, because it says something about the outcome — a floor and no ceiling — even though it says nothing about the arithmetic.

Solution to Exercise 16.E5. (a) A rise of one in a natural logarithm is a multiplication by e , about 2.7 — so a city roughly two-and-three-quarter times larger, not one person larger.

(b)

$$3.8 \times \ln 2 = 3.8 \times 0.693 = 2.63$$

Doubling a city's population goes with about 2.63 more cultural venues per hundred thousand residents.

(c) A city of one million against one of ten thousand is a one hundred-fold difference, which is 2 tenfold steps:

$$3.8 \times \ln 100 = 2 \times 8.75 = 17.50$$

venues per hundred thousand.

(d) Because city populations have a floor, no ceiling and a very long right tail, so a handful of large cities would otherwise carry all the leverage (§16.5.2) and the relationship is unlikely to be straight on the raw scale. The substantive reason is stronger than either: what plausibly matters for cultural provision is a city being *twice* the size of another, not ten thousand people larger, and a log model is the one that asks that question.

Solution to Exercise 16.E6. (a) They must be substantially correlated. A control can only take from a slope what it shares with the predictor (§16.6.1), so a fall of 66 per cent means screen time and family income overlap heavily — lower-income households report more screen time.

(b) Screen time and family income are two indicators of much the same underlying circumstance, and income is the one doing the work. What looked like an effect of screen time was a reflection of household income all along, and the control revealed it.

(c) Low family income causes higher screen time — fewer alternatives, less supervised activity, smaller homes — and screen time then affects wellbeing. Income is a step on the path from circumstance to outcome, so controlling for it removes part of the mechanism by which screen time operates. Screen time mattered, and the analysis subtracted the reason it mattered.

(d) Nothing in the data. The two readings produce identical coefficients, and a larger sample would produce the same identical coefficients more precisely. Distinguishing them requires an argument about causal order — ideally a design in which screen time varies for reasons unrelated to income (Chapter 3).

Solution to Exercise 16.E7. (a) Answers will depend on the download, but the regression of `infant_mort` on `educ_years` across the 52 countries of wave 1 gives a clearly negative slope: more schooling goes with fewer infant deaths. Report it in deaths per thousand live births per year of schooling, with its interval, r^2 , and the standard error of the estimate in deaths per thousand.

(b) Yes. Infant mortality behaves like GDP per capita in reverse — it falls steeply among the poorest countries and flattens among the richest — so a straight line through either shows the same arc in its residuals: the line above the data at both ends and below it in the middle.

(c) On the residual plot and the shape of the variable, not on r^2 . If the log

model's residuals are formless where the raw model's arc, that is evidence about specification rather than a larger number. The substantive argument is the stronger one: infant mortality plausibly responds to proportional differences in resources rather than absolute ones, which is what a log predictor encodes.

(d) The highest-leverage country will be at one end of the predictor's range — the poorest or the most-schooled, depending on the model. Whether it is *influential* is a separate question, and the way to answer it is to refit without it and compare the slopes. In this dataset the answer is almost certainly no: §16.5.2 reports the largest Cook's distance across all 52 countries as 0.11.

Solution to Exercise 16.E8. (a)–(b) Answers depend on the census year. A regression of median after-tax income on the degree share across census metropolitan areas gives a positive slope of a few hundred dollars per percentage point, with a fairly narrow interval: there are few enough areas that the sample is small, but the relationship is strong enough that the interval still excludes zero comfortably.

(c) Your r^2 will be far higher than .017 — values above .5 are usual for this pair at the area level.

The reason is aggregation. A metropolitan area's median income is an average over hundreds of thousands of people, and averaging cancels the individual variation that makes .017 small. What is left is the systematic part, so area-level relationships look much tidier than the individual-level relationships underlying them. Nothing about the world got stronger; the noise was averaged away.

(d) That an area-level r^2 says almost nothing about how well the same variables predict a person. Concluding from a strong area-level relationship that educated individuals earn proportionally more is the ecological fallacy (§10.7.5), and this pair of variables is one of its standard illustrations.