EPIB-609: Assignment based on material listed under "Controlling Confounding"

The learning aims for the readings and the exercise are to (i) review some regression-based and classical (not regression-based) statistical procedures for controlling confounding (ii) appreciate the connections between, and the essence of, the methods and (iii) acquire ways to describe these methods to lay people.

Although there are no ‘statistical free lunches,’ multivariable regression methods have increasingly replaced the classical standardization, stratification and matching (the finest stratification) approaches that were formerly used to control confounding. These classical methods are easier to explain to lay people than the ‘black boxes’ we rely on today.

Fortunately, when we are dealing with ‘regular’ rather than logistic or Poisson or Cox regression, i.e., when the ‘y’ is measured on an interval (continuous) scale, it is possible and indeed easy to show what control-by-regression involves. It involves the same (handicapping) strategy that is used to ensure a ‘fair’ game of (i) golf between two unequal players, or (ii) table-tennis between you and your niece or nephew, or (iii) a 100 metre dash between you and JH to measure the effect of steroids (JH’s religion doesn’t permit him to take steroids, so the race is between you on steroids and JH not on steroids!).

This ‘transmuting’ of data idea is a very old one: Francis Galton wanted to determine the correlation between the heights of parents and their (adult) offspring. He did not want to calculate a separate correlation for daughters and parents (about 0.5 in his data) and sons and parents (also about 0.5); he knew that if he ignored the sex of the offspring and calculated a correlation between all offspring and parents he would have obtained a very different (confused? confounded?) answer (Q: which direction?). So what could he do to have 1 combined-sexes dataset that would give one meaningful correlation? Today you could ‘put sex in the model’ (be careful to explain that phrase to your parents, who would wonder what you are learning at McGill). But Galton didn’t have Stata or R, or ‘multivariable’ regression models for that matter, and his friend Karl Pearson hadn’t yet come up with the product moment correlation; so what did Galton do? Think about it for a while, then if interested, consult JH’s article, poster and talk on the topic (see links).

Incidentally, when the ‘y’ is measured on an interval (continuous) scale, the use of analysis of covariance [a classical name for using multiple regression models with 1 binary variate (x) of interest, and 1 or more covariates (z’s) not of scientific interest] has a another advantage, beyond that of controlling confounding or making comparison *fairer*: it can make comparisons *sharper*. Sadly, the same is not necessarily true of the corresponding uses of multivariable logistic, Poisson and Cox regression. Worse still, the criteria for ‘confounding’ aren’t the same either.

Q1 on 681 Assignment effect of sexual activity on male longevity.

1. Use your favourite software package to replicate the numerical answers to parts a-e.

Article (and JH data-analyses) on Women and Math

1. Write a 1-paragraph explanation of regression-based control of confounding that would be understandable to a journalist who (a) majored in languages and the humanities (b) has a biology background (c) has a physics background. Use the Women and Math study as the example.

2. Write a 1-paragraph explanation of ‘regression-based reduction of statistical noise’ i.e., ‘regression-based sharpening of contrasts’ that would be understandable to these same journalists. Again, use the Women and Math study as the example.

3. Write a 1-sentence summary, using the ideas in the ‘anatomy of an adjustment’ graphic, to link the usual requirements for confounding with their equivalents in the regression approach.

4. The ‘no-confounding⇔collapsibility’ check works for models with a measured y and an identity link. It does not work for logistic regression, as Mantel illustrated using the two 2x2 tables, {a=90, b=10, c=50, d=50, or=9} when c=0 and {a=50, b=50, c=10, d=90, or=9} when c=1. Put these 8 frequencies into your logistic regression software, and verify that you get the same ‘anomaly’ (or = 9 within each level of c, but or << 9 when the data are collapsed over the two levels of c). While you are at it, also fit a ‘constant-risk-difference’ model to these frequencies, to see if the problem is with the nature of the *data* (continuous vs. binary y) or the *link* [identity (risk difference) vs logit(odds ratio)].

Article (and other links) on Galton and the family data on human stature

1. Write a 1-paragraph explanation of how Galton addressed the fact that sons tend to be taller than daughters. Was his approach inferior/superior to the typical approach students would use today? What ‘transmuting’ does this contemporary approach involve?

2. A different topic, but that (or a variant of it) could very easily be asked on the comprehensive examination. A propos the 2 pp excerpt from Pearson’s bio of Galton: why did Galton get lower and differential pair-specific correlations in the heights of m/f parents with m/f offspring than Pearson (and his grad students) did?

Data from Danish study of possible link between MMR vaccination and Autism

1. Use two classical (non-regression) approaches to ‘control’ (just for) for age.

2. Use a regression-based approach to ‘control’ (just for) for age. Compare.

3. In the case of a ‘y’ measured on an interval (continuous) scale, we saw that it is easy to give the ‘adjusted’ difference a very concrete representation. Try to come up with a representation in the ‘count data’ case. {*Hint*: don’t use this paper, found by Google when JH searched for “Breslow standardization regression”; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1060387/ instead you might get some inspiration from Breslow, N. E. and Day. N. E. (1975). Indirect standardization and multivariate models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. Journal of Chronic Diseases 28, 289-303.}

4. If you were writing the abstract, would you have reported the (crude) rate ratio of 1.45? Why/why not?

2012.08.23