Biostatistics 140.654
Fourth Term, 2019
Scott L. Zeger

Note on Confounding and Effect Modification – Definitions and Measures

Confounding means confusion. The definition of "confounder" in the Oxford English Dictionary is: "One who causes confusion or disorder".

The hallmark strategy to estimate the causal effect of a treatment (risk factor) on an outcome is to make all factors besides the treatment the same between comparison groups so that differences in the distribution of outcomes can be reasonably attributed to the treatment. In non-human experimentation, we make groups equivalence by controlling experimental conditions. For example, we use mice of the same genetic strain, housed in identical cages, fed the same food and so on. Only the treatment is different.

In human research, we cannot control genetic and environmental factors that influence outcomes. Instead, when possible, we make groups equivalent "in distribution" by randomizing participants to receive the treatment or not. Randomization produces comparison groups that are equivalent in expectation for all other factors, whether observed or not.

In observational research, treatments or risk factors are rarely randomly assigned to people. Therefore, we must understand that the groups are likely not to be comparable for variables other than the risk factor or treatment of interest. At a minimum, we must only compare groups with and without the risk factor that have similar distributions for those **measured variables** that themselves can cause differences in the outcome or that proxy for other variables that do. If we fail to compare "like to like", we may incorrectly ascribe to the risk factor of interest a difference that has been caused by other factors.

Epidemiologists have traditionally called these other factors "confounders" as they confuse our assessment of the effects of the risk factor on the health outcome (Morabia A. History of the modern epidemiological concept

of confounding. Journal of Epidemiology and Community Health. 2011;65:297–300.)

For example, when studying the effects of smoking on disease incidence, we compare smokers to non-smokers who are similar for factors, for example age, that cause these diseases. The comparison groups must have similar age distributions because the incidence of many diseases caused by smoking, for example lung cancer and coronary heart disease, increases with age. If we were to compare younger smokers with older non-smokers, we may incorrectly attribute to smoking the benefits to health of being younger and thereby understate the effects of smoking. Here age would "confound" our assessment of the smoking-disease relationship.

Measuring confounding: In the regression context, we can measure how much variables Z confound our estimate of the X-effect on Y by comparing the regression coefficient, B_x, for X in a regression model of Y on X excluding Z to its value in a model including Z. That is, for linear regression, we fit two models:

$$Y = B_0 + B_X X + e$$
$$Y = A_0 + A_{X|Z} X + A_Z Z + e.$$

The difference $\delta = A_{X|Z} - B_X$ is a measure of the confounding of the X->Y relationship by Z. It is the change in the linear association of Y with X when we control for Z relative to when we do not. Note X and Z can be single variables or vectors of variables.

The population value, δ, will be non-zero if Z is actually a confounder or if X causes Z which in turn causes Y. In the latter case, epidemiologists call Z a "mediator" of the effect of X on Y, rather than a "confounder". The term "confounder" is restricted for variables Z that are not in the causal pathway of X to Y.

Given this definition, it follows that:

- there are degrees of confounding as measured by estimates of the continuous parameter δ . Confounding is not a dichotomy.

- the extent of confounding depends on the joint distribution of Z, X and Y in the population. The estimator, d, of δ is a random variable; it will vary from sample to sample; The "true" confounding is the value of δ when we use the whole population in the two regressions.
- if X and Z are uncorrelated in a particular sample, d=0 in that sample.
- if Y and Z are uncorrelated, Z can still confound, contrary to what many elementary texts say. What is important is that Y adjusted for other covariates and Z adjusted for other covariates are uncorrelated in which case d=0.

The simplest way to display the extent of confounding is to plot side by side on a single set of axes, the estimate and confidence intervals for $A_{X|Z}$ and B_X.

Extra enjoyment problem. In the linear regression model, derive an expression for d and its variance. Propose a test for confounding using these results.

Example - Medical expenditures and major smoking caused diseases. When studying the medical costs associated with smoking, we are interested in the fraction of people who have sizable expenditures in a particular year and whether this rate is greater for persons with a major smoking caused disease (MSCD) such as lung cancer or coronary heart disease. We define *bigexp* to be 1 if a person has expenditures greater than $1,000 in a year, 0 otherwise and *mscd* to be 1 if a person has an MSCD, 0 otherwise. The variable *lastage* is the person's age in years.

Below find output for two logistic regressions using data from the National Medical Expenditure Survey (NMES) for 1987.

Model A

summary(lm(bigexp~ mscd))

```
------------------------------------------------------------------------------
      bigexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mscd |   1.825045   .0667707    27.33   0.000     1.694177    1.955914
       _cons |  -.7395315    .021003   -35.21   0.000    -.7806966   -.6983663
------------------------------------------------------------------------------
```

Model B

bigexp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
mscd	1.603332	.0682724	23.48	0.000	1.46952	1.737143
agem65	.028163	.0027894	10.10	0.000	.0226959	.0336301
age_sp65	-.0063276	.0059568	-1.06	0.288	-.0180027	.0053476
_cons	-.5743886	.0378007	-15.20	0.000	-.6484766	-.5003006

In Model A, the logistic regression coefficient for *mscd* is 1.83 corresponding to an odds ratio of exp(1.83)= 6.2. The data indicate that the odds of having an expenditure greater than $1,000 is more than six times as large for persons with an MSCD than with person who do not have an MSCD.

However, older persons tend to have more diseases and may also have higher medical expenditures than younger persons even absent an MSCD. We do not want to attribute to MSCD what may be caused by age alone.

In Model B, we control for age using two variables: agem65 = age –65; and age_sp65 = age–65 when age is greater than 65 and 0 otherwise. These two variables define a linear spline or "broken line" function of age with one slope (agem65 coefficient) prior to 65 years and another (agem65+age_sp65 coefficients) after age 65.

The coefficient for MSCD in Model B is 1.60 corresponding to an odds ratio of 4.97. When we compare groups with roughly the same age distribution, the odds of a big expenditure for cases (*mscd*=1) is about 5 times as great as for controls, in contrast to the estimate 6 when we did not control for age.

Is this difference between the two MSCD coefficients an indication of modest confounding? Unfortunately, with logistic regression and other non-linear links other than the log-link, the question is a bit more complicated. Even if age is independent of mscd so that age is not a confounder, the two regressions would give different coefficients because of the non-linearity of the logistic function. That they are not the same has been termed "noncollapsibility" by Rothman (e.g. Rothman and Greenland, Modern Epidemiology, pp 52-55; Greenland and Robins, 2009). When you average over even an uncorrelated independent covariate, you attenuate the

relationship of MSCD with the outcome. As a first approximation when the potential confounder's contribution to the linear predictor is approximnately Gaussian, you can address the confounding question by comparing the Z-statistics rather than the coefficients. Because the attenuation tends to be the same for the standard error as it is for the coefficient in logistic regression, the Z-ratio is less effected by averaging over the potential confounder.

Extra enjoyment. To read a statistical paper about measuring confounding, check out Janes, Holly, Francesca Dominici, and Scott Zeger. "On quantifying the magnitude of confounding." *Biostatistics* 11.3 (2010): 572-582.

Effect modification: When the effect of X on Y is different at different values of Z, epidemiologists say "Z modifies the effect of X on Y" or "Z is an effect modifier of the X-Y relationship." Statisticians say that: "X interacts with Z in its association with Y." For example, suppose that the odds ratio of a large medical expenditure for persons with and without a major smoking caused disease (MSCD) varies with age. We would say: "age modifies the effect of having an MSCD on medical expenditures".

In statistical terms, we measure effect modification using interaction terms in regression models. The simplest case is when Z is dichotomous. We can estimate a separate regression coefficient for X when Z=0 and Z=1 using the model

$$Y = (B0 + B1\ X)*(1- Z) + (B2 + B3\ X)*Z + e$$
$$= B0 + (B2-B0)\ Z + B1\ X + (B3- B1)\ X*Z + e.$$

Here, B1 is the coefficient for X when Z=0 and B3 is its coefficient when Z=1. The difference B3-B1 is a measure of the effect modification of the X-Y relationship by Z.

The same general approach can be used when Z is continuous. However, this specific model imposes a strong assumption that the degree of effect modification is a linear function of Z. A better approach is to create a few basis functions of Z and to include their interaction with X. The simplest is

to stratify Z into quartiles and create four indicator variables Z1, Z2, Z3, and Z4 where Zj = 1 if Z is in the jth quartile and 0 if not. We then include the interaction of each Zj with X in the model for Y. The coefficient for (Zj*X) is the linear dependence of Y on X when Z in in the jth quartile. A limitation of this approach is the inherent assumption that the effect of X on Y changes abruptly as a step function of Z. We can allow for a continuous interaction by using linear or cubic spline basis functions of Z, call them Z1, Z2, …, Zp, rather than the quartile indicators. We can then summarize the effect of X on Y by plotting Z1 A1 + …+ Zp Ap against Z where A1,…, Ap are the estimated regression coefficients for the interactions of X with Z1,…,Zp respectively.

Returning to the medical expenditure example, an important question is whether the odds ratio of having a large expenditure (>$1,000) for persons with and without a MSCD varies by age. We examine this question, we have fit an interaction of MSCD with a LASTAGE in Model E. We assume a natural cubic spline form for the interaction. The graph below shows the predicted values from the two models.

```
#
lrE.1=glm(data=data1,bigexp~mscd + ns(LASTAGE,3),family=binomial(link="logit"))
summary.glm(lrE.1)
pE.1=fitted.values(lrE.1,se.fit=TRUE)
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.29160    0.06412 -20.142  < 2e-16 ***
mscd         1.60386    0.06832  23.476  < 2e-16 ***
ns(age, 3)1  0.72762    0.09346   7.785 6.95e-15 ***
ns(age, 3)2  1.51049    0.17569   8.598  < 2e-16 ***
ns(age, 3)3  1.13785    0.13415   8.482  < 2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15411  on 11683  degrees of freedom
Residual deviance: 14269  on 11679  degrees of freedom
AIC: 14279

Number of Fisher Scoring iterations: 4
```

```
lrE.2=glm(data=data1,bigexp~mscd * ns(LASTAGE,3),family=binomial(link="logit"))
summary.glm(lrE.2)
pE.2=fitted.values(lrE.2,se.fit=TRUE)
```

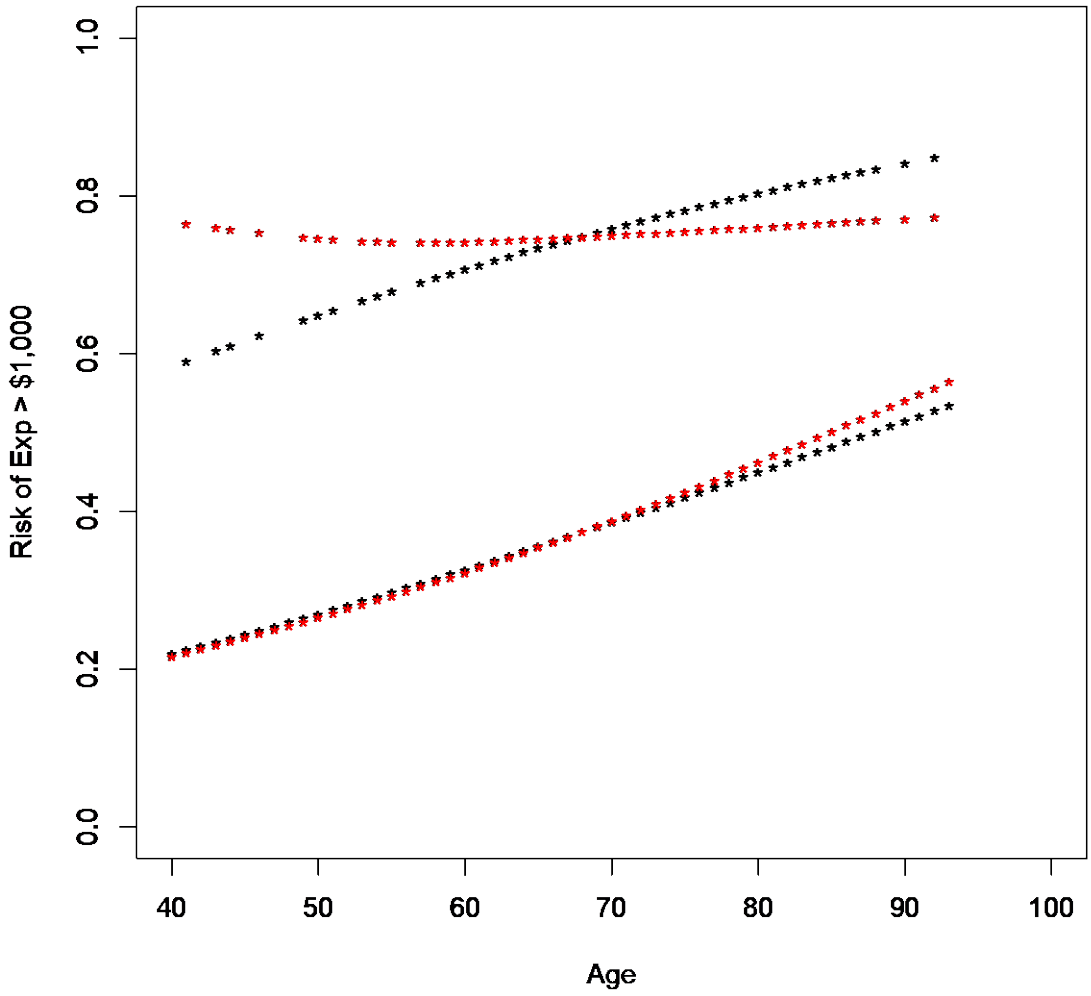
```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.31135    0.06553 -20.011  < 2e-16 ***
mscd             2.61985    0.43727   5.991 2.08e-09 ***
```

```
ns(age, 3)1       0.75677    0.09905   7.640 2.17e-14 ***
ns(age, 3)2       1.58797    0.18265   8.694  < 2e-16 ***
ns(age, 3)3       1.25306    0.14549   8.613  < 2e-16 ***
mscd:ns(age, 3)1 -0.84668    0.34698  -2.440  0.01468 *
mscd:ns(age, 3)2 -2.03558    0.97037  -2.098  0.03593 *
mscd:ns(age, 3)3 -1.10697    0.38985  -2.839  0.00452 **

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15411  on 11683  degrees of freedom
Residual deviance: 14253  on 11676  degrees of freedom
AIC: 14269
```

sub=sample(1:length(data1$LASTAGE),1000)
plot(data1$LASTAGE[sub],pD[sub],xlim=c(40,100),ylim=c(0,1),xlab="Age",ylab="Risk of Exp > $1,000",pch="d",col="black")
points(data1$LASTAGE[sub],pE[sub],col="red",pch="e")



We can ask whether the interaction terms improve the fit to the observed data using a likelihood ratio test. The deviance available in the R output is -2*(log likelihood of fitted model - smallest possible value). So the difference in the deviance between Models E.1 and E.2 is the difference in -2* log likelihood between these two models. Under the null hypothesis that the interaction term coefficients are all 0, the change in deviance follows a chi-squared distribution with degrees of freedom equal to the number of new regressors added.

The observed change in deviance is (`14269 – 14253)=16 on 3 degrees of freedom`. The p-value for a chi-square with 3 df is given by `1-pchisq(16,df=3) = 0.0011`. Hence, we can conclude that the mscd effect does change with age.

Extra enjoyment. Estimate the difference in the predicted risk of a big expenditure for persons with and without an mscd as a function of age and its confidence interval.