Confounders and Precision Variables

BIOS 6611

CU Anschutz

Week 12

1. Confounding

2. Precision Variables

Confounding

Confounding

A common use of multiple regression models in the health sciences is to adjust an association for the effects of confounding variables.

Confounding is the distortion of an estimated association due to the effect(s) of other variable(s).

The **confounder** (C) is the variable that causes the distortion.

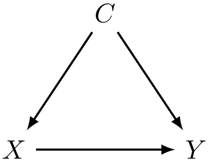
We will discuss *two* criteria for confounding after introducing some terminology.

Some Terminology for Confounding

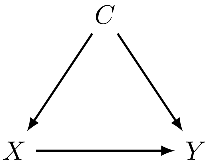
A **crude** (or **unadjusted**) **estimate** is of the association of the **primary explanatory variable (PEV)** (X) with our outcome (Y) when potential confounder(s) are excluded from the model.

The **adjusted estimate** is of the association of X with Y when we account for C in the model.

We can represent this relationship using a **directed acyclic graph** (DAG):



Three Models of Interest



From our DAG, we can define three models of interest:

1. Crude Model: $\hat{Y} = \hat{\beta}_{01} + \hat{\beta}_{crude}X$
2. Adjusted Model: $\hat{Y} = \hat{\beta}_{02} + \hat{\beta}_{adj}X + \hat{\beta}_C C$
3. Covariate Model: $\hat{C} = \hat{\gamma}_0 + \hat{\gamma}_X X$

We will use these three models to evaluate potential confounders.

Classical Criteria for Confounding

1. A confounding factor must be associated with the exposure (or PEV) under study. From our three models, this is the association of X and C represented by $\hat{\gamma}_X$.
2. A confounding factor must be a risk factor or a surrogate for a risk factor for the disease. From our three models, this is the association of C and Y given X represented by $\hat{\beta}_C$.
3. A confounding factor must not be affected by the exposure or the disease.

Note, the confounding factor **cannot** be an intermediate step in the causal path between the exposure and the disease. In this case, we would have a *mediator*.

Operational Criterion for Confounding

The operational criterion for confounding states a covariate is a confounding factor if the crude parameter estimate is not equal to the adjusted parameter estimate: $\hat{\beta}_{crude} \neq \hat{\beta}_{adj}$.

Confounding is present if there is a “meaningful” difference between these estimates, which will depend on the context and what would represent a clinically relevant change.

If clinically meaningful change is uncertain, we might calculate the percent change in one of two ways:

- $\frac{\hat{\beta}_{crude} - \hat{\beta}_{adj}}{\hat{\beta}_{crude}} \times 100$ (favored by biostatisticians)
- $\frac{\hat{\beta}_{crude} - \hat{\beta}_{adj}}{\hat{\beta}_{adj}} \times 100$ (favored by epidemiologists)

While the answers will differ slightly, they generally produce similar results. In practice we may be looking for a 10% or 20% change as meaningful (*although it will depend on context!*).

Connection between Classical and Operational

Based on our three models, there is a direct connection between the operational and classical definitions of confounding:

$$\hat{\beta}_{crude} - \hat{\beta}_{adj} = \hat{\gamma}_X \times \hat{\beta}_C$$
$$\text{operational} = \text{classical}$$

Statistical tests for confounding are generally not used (with some authors claiming they are "neither required nor appropriate"[1]). Indeed, the classical and operational definitions are largely built on context of a given problem to evaluate if they are meaningful.

[1] Kleinbaum, D.G., Kupper, L.L., Nizam, A., Rosenberg, E.S. Applied Regression Analysis and Other Multivariable Methods. Boston, MA: Brooks/Cole Cengage Learning, 2014.

Positive Confounding

Definition 1: A variable that is positively associated with both exposure and disease or negatively associated with both exposure and disease is called a positive confounder.

Definition 2: Positive confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is more extreme – either more positive or more negative – than the true association.

A positive confounder can create spurious associations.

Negative Confounding

Definition 1: A variable that is positively associated with the exposure and negatively associated with the disease (or vice versa) is called a negative confounder.

Definition 2: Negative confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is an underestimate of the true association.

A negative confounder can mask a true association.

Accounting for Confounding

During study design:

- Matching: match cases and controls by known confounders
- Restriction: restrict the study eligibility criteria to include only individuals in specified categories of a confounder (limits generalizability)
- Randomization: randomly assign individuals to treatment groups. On average, this gives groups that are balanced for both measured and unmeasured confounders.

During analysis:

- Stratification: stratify the analysis by levels of a confounder (reduces sample size since analyses are completed on stratified subgroups instead of using all data)
- Regression: adjust for confounders in a statistical model

Confounding Example

In the FEV data set we were interested in determining if there was a difference in the lung function of children who smoked and children who did not smoke.

We can see that smokers have higher FEV, but that smokers are also older and FEV generally increases with age (*positive* confounding by age):

```
dat <- read.csv('FEV_rosner.csv')
doBy::summaryBy( fev + age ~ smoke, data=dat, FUN=mean)
```

```
##       smoke fev.mean  age.mean
## 1 nonsmoker 2.566143  9.534805
## 2    smoker 3.276862 13.523077
```

```
cor( dat$fev,  dat$age )
```

```
## [1] 0.756459
```

Confounding Example

Let's fit our three models to use the estimated coefficients to evaluate confounding:

```
crude <- lm(fev ~ smoke, data=dat)
summary(crude)$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 2.5661426 0.03466043 74.036674 1.487335e-319
## smokesmoker 0.7107189 0.10994262  6.464453  1.992846e-10
```

```
adjusted <- lm(fev ~ smoke + age, data=dat)
summary(adjusted)$coefficients
```

```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)  0.3673730 0.081435716  4.511203  7.647680e-06
## smokesmoker -0.2089949 0.080745337 -2.588321  9.859773e-03
## age          0.2306046 0.008184372 28.176209 8.279537e-115
```

```
covariate <- lm(age ~ smoke, data=dat)
summary(covariate)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 9.534805  0.1114115 85.58191 0.000000e+00
## smokesmoker 3.988272  0.3533963 11.28555 4.188788e-27
```

Confounding Example

Operational Criterion: $\frac{\hat{\beta}_{crude}-\hat{\beta}_{adj}}{\hat{\beta}_{crude}} \times 100 = \frac{0.711-(-0.209)}{0.711} \times 100 = 129.4.$
Yes, age is a confounder since $\hat{\beta}_{crude} \neq \hat{\beta}_{adj}$ and their difference is much greater than 20%.

Classical Criteria:

1. Age is associated with smoking (our exposure/PEV) (e.g., comparison of the mean age, could do a t-test, etc.)
2. Age is associated with FEV given smoking ($\hat{\beta}_C = 0.231$, $p < 0.001$)
3. Age is not on the causal pathway (smoking doesn't *cause* age) (subject matter consideration)

Connection:

$$\hat{\beta}_{crude} - \hat{\beta}_{adj} = \hat{\gamma}_X \times \hat{\beta}_C$$
$$0.711 - (-0.209) = 3.988 \times 0.231$$
$$0.92 = 0.92$$

Confounding Example

Putting it all together:

On average, smokers are 3.99 years older than non-smokers ($\hat{\gamma}_X$).

On average, for every one year increase of age, FEV increases by 0.23060 liters ($\hat{\beta}_C$).

So we'd expect FEV to be $3.99 \times 0.23 = (\hat{\gamma}_X \times \hat{\beta}_C) = 0.9197$ L higher in smokers compared to non-smokers due to age.

Precision Variables

Precision Variables

The term **precision** refers to the size of an estimator's variance, or equivalently, the narrowness of a confidence interval for the parameter being estimated.

The smaller the variance of the estimator, the higher the precision of the estimator:

$$\frac{Var(\hat{\beta}_{adj})}{Var(\hat{\beta}_{crude})} = \frac{1 - \hat{\rho}^2_{YZ|X}}{n-3}\left(\frac{n-2}{1-\hat{\rho}^2_{XZ}}\right)$$

where Z is another independent variable, $\hat{\rho}_{YZ|X}$ is the **partial correlation** between Y and Z that controls for X, and $\hat{\rho}_{XZ}$ is the correlation between X and Z.

Precision Variables

A strong association between Y and Z has a *beneficial* effect upon the precision of $\hat{\beta}_{adj}$ (i.e., it *decreases* $SE(\hat{\beta}_{adj})$).

A strong association between X and Z has a *detrimental* effect on the precision of $\hat{\beta}_{adj}$ (i.e., it *increases* $SE(\hat{\beta}_{adj})$).

Thus, the precision of $\hat{\beta}_{adj}$ reflects the competing effects of the Y-Z and X-Z relationships. A **precision variable** improves the precision of the estimate of the PEV.

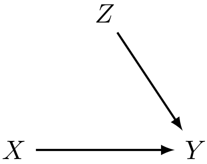
Covariate Adjustment in Linear Regression

A precision covariate is a variable independent of exposure in the source population ($\gamma_X = 0$), but predictive of the outcome ($\beta_Z \neq 0$).

Precision covariates **cannot** be confounders according to the classical criteria.

Inclusion of a precision variable can provide a

- more efficient test of the exposure-outcome association
- more precise estimate of the exposure-outcome association



Precision Variable Example

Let's return to our earlier example with FEV and smoking status, with age (as a confounder).

Although the classical criteria of confounding indicates a confounder cannot be a precision variable, we can still evaluate the change in precision for smoking status by including age:

$$\frac{Var(\hat{\beta}_{adj})}{Var(\hat{\beta}_{crude})} = \frac{(0.08075)^2}{(0.10994)^2} = 0.539$$

Since our ratio is < 1, we have much better precision around $\hat{\beta}_{adj}$ by including age, and since it is a confounder we accounted for some of the potential bias. However, since age is associated with our PEV, it is a confounder instead of a precision variable.