Confounding and interaction

Biometry 755

Spring 2009

What is confounding?

Confounding is a distortion of the true relationship between exposure and disease by the influence of one or more other factors. These “other factors” are known as *confounders*. Confounding variables are *nuisance* variables, in that they “get in the way” of the relationship of interest. It is therefore desirable to remove their effects.

Identifying potential confounders

Under what circumstances does a variable confound an exposure-disease relationship?

1. When the factor is associated with the exposure, but is not believed to be a result of the exposure.
2. When the factor is a risk-factor for the disease, in that it is either
   (a) a cause of the disease,
   (b) a correlate of the disease, or
   (c) influential in the recognition or diagnosis of the disease.

Example

A woman's risk of breast cancer is directly correlated with her age at the time of birth of her first child. When evaluating the association between total number of births and breast cancer risk, should maternal age at first birth be controlled?

Potential confounder Age at first birth

Exposure Total number of births

Disease Breast cancer

Age at first birth (POT. CONF) is associated with total number of births (EXP), but age at first birth (POT. CONF) is not a result of total number of births (EXP). Age at first birth (POT. CONF) is a known risk-factor for BrCa (DIS). Therefore, we need to control for age at first birth (POT. CONF).

Controlling for confounders

We can control for confounders in the design stage or at the time of analysis. In other courses, you've learned about controlling for confounders by design ... matching, randomization, etc. In this course, we will focus on controlling for confounders at the time of analysis. This is done by including the confounder as an independent variable in the model.

If a factor is a known confounder, it should be included as a covariate, and kept in the model regardless of the level of significance of the corresponding partial F test, because the confounder's inclusion in the model provides a better understanding of the true disease-exposure relationship.

Identifying confounders

But what if you don't know *a priori* that a variable is a confounder? You can identify a covariate as a confounder based on the results of two regression analyses: one with and one without the potential confounder.
Suppose you fit the following models:

Model 1 $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Model 2 $Y = \beta_0^* + \beta_1^* X_1 + \beta_2 X_2 + \varepsilon$

If β_1 is "appreciably" different from β_1^*, then X_2 is said to confound the relationship between Y and X_1. However, what constitutes an "appreciable" difference varies by context. In practice, we examine the difference between $\hat{\beta}_1$ and $\hat{\beta}_1^*$.

Example

```
Regression of INFRISK on LOS
Variable      Parameter Estimate

Intercept     0.7443
LOS           0.3742
```

```
Regression of INFRISK on LOS, controlling for REGION
Variable      Parameter Estimate

Intercept     1.0339
LOS           0.4125
REGION    1  -0.7479
REGION    2  -0.6350
REGION    3  -0.8987
REGION    4   0.0000
```

Example (cont.)

From the results of the regression of INFRISK on LOS, $\beta_{\text{LOS}} \doteq 0.37$, with the interpretation that risk of nosocomial infection increases by 0.37 % for every 1-day increase in length of stay.

From the results of the regression of INFRISK on LOS, controlling for REGION, $\beta^{*}_{\text{LOS}} \doteq 0.41$, with the interpretation that risk of nosocomial infection increases by 0.41 % for every 1-day increase in length of stay, after controlling for the effects of region of the U.S.

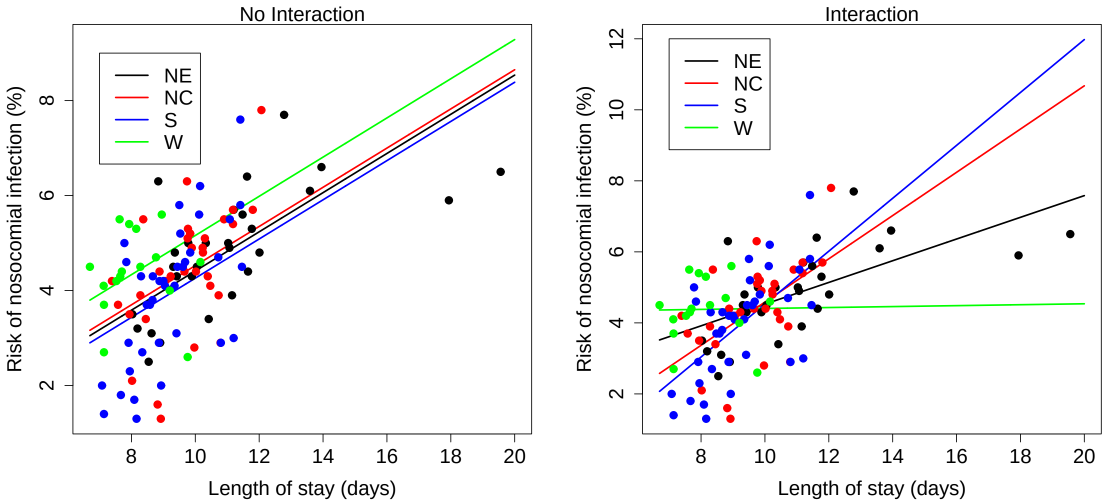
Are these appreciably different results? Does REGION confound the relationship between INFRISK and LOS?

What constitutes an "appreciable" change?

- *BEST:* Use your clinical expertise to identify a meaningful change in the parameter estimate.
- Always consider including variables identified as a confounders from previous studies. You could be criticized for not doing so.
- The "usual suspects" (age, race, gender) make their way into many analyses involving person-level data.
- *A rule of thumb:* Greenland et al. report that in simulations a "...10 per cent change-in-estimate method produced the most valid point and interval estimates." (Sander Greenland, Modeling and variable selection in epidemiologic analysis, *American Journal of Public Health* 1989; 79:340-349)

What is interaction?

Statistical interaction describes the relationship in which the effect of an explanatory variable on the response differs significantly across levels of a second explanatory variable.



Modeling interaction in SAS

No interaction

```
proc glm data = one;
    class region;
    model los = region/solution ss3;
run;
quit;
```

Interaction

```
proc glm data = one;
    class region;
    model infrisk = los region los*region/
                    solution ss3;
run;
quit;
```

SAS output: No Interaction

```
Parameter     Estimate

Intercept      1.0339
LOS            0.4125
REGION    1   -0.7479
REGION    2   -0.6350
REGION    3   -0.8987
REGION    4    0.0000
```

No Interaction

How do we get the parallel lines from this output?

SAS output: Interaction

```
Parameter                Estimate

Intercept             4.274209152 B
LOS                   0.013192525 B
REGION     1         -2.801855473 B
REGION     2         -5.777007350 B
REGION     3         -7.193493691 B
REGION     4          0.000000000 B
LOS*REGION 1          0.292369964 B
LOS*REGION 2          0.595738777 B
LOS*REGION 3          0.731672001 B
LOS*REGION 4          0.000000000 B
```

Interaction

How do we get the intersecting lines from this output?

How would we have done this in PROC REG?

The “*” operator does NOT work in PROC REG, so to consider this interaction using PROC REG, you must:

1. Make three indicator variables for REGION in a DATA step.
2. Make three indicator variables for REGION*LOS in a DATA step.
3. Run PROC REG with 7 variables in the model.

CONCLUSION: If interaction is in the model, it is always a little less work if you use PROC GLM rather than PROC REG.

Assessing the significance of the interaction

```
Source          DF      Type III SS      Mean Square     F Value     Pr > F

LOS             1      24.83776648      24.83776648       21.19    <.0001
REGION          3      13.61064193       4.53688064        3.87    0.0114
LOS*REGION      3      12.27692104       4.09230701        3.49    0.0183
```

Note: If an interaction term is retained in a model, then you must include the main effect terms as well, regardless of their level of significance.

What about subsetting the data?

Q: If we believe there is meaningful interaction between two variables, wouldn't it be equivalent to simply subset the data?
A: Let's see ...

Confounding and interaction redux

When significant interaction is present, it is misleading and incorrect to report an overall adjusted summary index of the relationship of interest (i.e. the coefficient from the MLR corresponding to the covariate of primary interest). To do so masks the interaction effects that are present. For example, since the INFRISK-LOS relationship differs meaningfully for different values of REGION, then using an overall (REGION-adjusted) summary index of the relationship between INFRISK and LOS will hide the interesting interaction finding.

Therefore, if you are going to consider interaction terms in your model, *interaction is always assessed before confounding. Using an overall adjusted summary estimate is advised only if no significant interaction is present.*