Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 19, part 4: Statistical Interaction and Confounding

Jinko Graham

Overview

- What is the relationship between a response variable, Y, and explanatory variable, X_1, adjusted for other variables?
- Example: Study of low-birthweight babies.
  - Y is head circumference and X_1 is gestational age.
  - Interested in the effect of X_1 on Y as summarized by the slope term in the linear model.
  - However, an extraneous variable X_2, such as birth weight, can ***modify*** (change) the effect of X_1 on Y.
- Earlier, we looked at ***effect modification***, when studying association between a categorical outcome variable, Y, with two levels and a categorical exposure variable, X_1, with two levels.
  - In this context, the effect of X_1 on Y was summarized by the odds ratio (rather than the slope term).
  - When the effect of X_1 on Y was different depending on the value of X_2, then X_2 was a modifier of the effect of X_1 on Y.
  - If X_2 was not an effect modifier, we could go on to consider it as a potential confounder of the association between X_1 and Y.

Steps

- Suppose we're primarily interested in the association between Y and X_1.
- Have also collected data on an extraneous variable, X_2.
- Suggested steps are:

1. First consider whether X_2 ***modifies*** the effect of X_1 on Y
   - Called statistical interaction between X_1 and X_2.
2. If there is no statistical interaction, we can consider X_2 as a potential confounding variable.
   - X_2 could change the association between Y and X_1 when it is included in our MLR model.

- We'll be using the data on low birthweight babies to illustrate ideas.

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
head(lbwt)
```

```
##   headcirc length gestage birthwt momage toxemia
## 1       27     41      29    1360     37       0
## 2       29     40      31    1490     34       0
## 3       30     38      33    1490     32       0
## 4       28     38      31    1180     37       0
## 5       29     38      30    1200     29       1
## 6       23     32      25     680     19       0
```

Statistical Interaction

- Easiest when X_2 is binary; i.e., takes values of 0 or 1.
- X_2 *modifies* the effect of X_1 on Y if the slope of the regression line of Y on X_1 differs in the $X_2 = 0$ and $X_2 = 1$ subgroups.
- Illustrate with the variable `toxemia` in the low birthweight babies dataset.
  - `toxemia=1` if the mother is toxic during pregnancy and 0 otherwise
  - If we stratify the analysis by toxemia and find different slopes for gestational age in the two toxemia groups, there is statistical interaction between gestational age and toxemia.

MLR Model with Statistical Interaction

- Consider the MLR model with *linear predictor*:
  $\mu_{Y|X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, where
  - Y is head circumference, `headcirc`.
  - X_1 is gestational age, `gestage`.
  - X_2 is `toxemia` (1 is toxic, 0 is not)
  - $X_1 \times X_2$ is the statistical interaction between gestational age and toxemia.
  - β_1, β_2, and β_3 are the corresponding regression coefficients:
    - β_1 is the gestational-age ***main effect***
    - β_2 is the toxemia ***main effect***
    - β_3 is the gestational-age-by-toxemia ***interaction effect***

Separate Lines

- Our linear predictor is

$$\mu_{Y|X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

- This model allows separate lines for the two toxemia groups.
  - Line for no-toxemia group ($X_2 = 0$): $\alpha + \beta_1 X_1$
    - intercept α and slope β_1 for `gestage`.
  - Line for toxemia group ($X_2 = 1$): $\alpha + \beta_1 X_1 + \beta_2 + \beta_3 X_1$
    - intercept $\alpha + \beta_2$ and slope $\beta_1 + \beta_3$ for `gestage`.
- Focusing on the slopes, we see that the difference between `gestage` slopes for the two toxemia groups is β_3.
  - $\beta_3 = 0$ implies that the slopes are the ***same*** in the two groups;
  - i.e., toxemia does ***not*** modify the effect of gestational age on head circumference.

- To assess the evidence for statistical interaction between toxemia and gestational age, we test the hypotheses

$$H_0 : \beta_3 = 0 \;\; vs. \;\; H_a : \beta_3 \neq 0.$$

- If H_0 is retained, we conclude that there is insufficient statistical evidence that toxemia modifies the effect of gestational age on head circumference.

- If we retain the no-interaction of hypothesis $H_0 : \beta_3 = 0$, our linear predictor becomes

$$\mu_{Y|X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2$$

- This model allows separate lines for the two toxemia groups but with the same slope for `gestage`:
  - Line for no-toxemia group ($X_2 = 0$): $\alpha + \beta_1 X_1$
    - intercept α and slope β_1 for `gestage`.
  - Line for toxemia group ($X_2 = 1$): $\alpha + \beta_1 X_1 + \beta_2$.
    - intercept $\alpha + \beta_2$ and slope β_1 for `gestage`.
- No interaction between `gestage` and `toxemia` means that each toxemia group has its own line with different intercepts, but with the same slope for `gestage`

Fitted Model

- Let's fit the MLR model with interaction between gestational age and toxemia:

```
lfit <- lm(headcirc ~ gestage+toxemia+gestage:toxemia,data=lbwt)
summary(lfit)$coefficients
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)      1.7629121 2.10225478  0.8385815 4.037874e-01
## gestage          0.8646116 0.07389805 11.7000601 3.529066e-20
## toxemia         -2.8150322 4.98514735 -0.5646839 5.736059e-01
## gestage:toxemia  0.0461658 0.16352127  0.2823229 7.783037e-01
```

- From the row of output for `gestage:toxemia`, we see that the t-test of $H_0 : \beta_3 = 0$ *vs.* $H_a : \beta_3 \neq 0$ retains H_0 at the 5% level ($p = 0.78$).
- No statistical evidence that toxemia modifies the effect of gestational age on head circumference.
- However, toxemia may still be a confounding variable.

Software Notes

- In the model formula
  `headcirc ~ gestage + toxemia + gestage:toxemia`
  - The ***interaction*** term between `gestage` and `toxemia` is indicated by `gestage:toxemia`.
  - The ***main-effect*** terms are indicated by `gestage` and `toxemia`.
- In the model summary:
  - Information about the slope β_3 for the interaction term is in the row labelled `gestage:toxemia`.
  - Information about the slopes β_1 and β_2 for the main effect terms are in the rows labelled `gestage` and `toxemia`, respectively.

Statistical Interaction More Generally

- Interaction terms appear as products of main-effect terms.
  - E.G. in the MLR with linear predictor $\mu_{Y|X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, the interaction term $X_1 X_2$ is a product of the main-effect terms X_1 and X_2.
- Interaction means that the slopes for X_1 can depend on the value of the modifying variable X_2.
- E.G. Say X_2 is a quantitative variable taking on values between 5 and 10.
- Focus on two values $X_2 = 5$ and $X_2 = 6$ one unit apart:
  - Line for $X_2 = 5$ group is $\alpha + \beta_1 X_1 + \beta_2 5 + \beta_3 X_1 5$
    - intercept: $\alpha + \beta_2 x_2 = \alpha + \beta_2 5$, and
    - slope for X_1: $\beta_1 + \beta_3 x_2 = \beta_1 + \beta_3 5$
  - Line for $X_2 = 6$ group is $\alpha + \beta_1 X_1 + \beta_2 6 + \beta_3 X_1 6$
    - intercept: $\alpha + \beta_2 x_2 = \alpha + \beta_2 6$, and
    - slope for X_1: $\beta_1 + \beta_3 x_2 = \beta_1 + \beta_3 6$
- Difference in slopes for X_1 for the two groups is $\beta_3 6 - \beta_3 5 = \beta_3$.

- In general, we interpret the slope β_3 for the interaction term $X_1 X_2$ as the difference between the slopes for X_1 in two groups that are defined by a one-unit change in X_2.
- If $\beta_3 = 0$, the slopes for X_1 are the same.
  - Therefore, X_2 does **not** modify the effect of X_1 on Y.
- To assess the statistical interaction of X_1 and X_2, test the hypotheses that $H_0 : \beta_3 = 0$ *vs.* $H_a : \beta_3 \neq 0$.
- This is equivalent to testing whether or not X_2 modifies the effect of X_1 on Y.

Example: Interaction of Gestational Age and Birthweight

```r
lfit <- lm(headcirc ~ gestage+birthwt+gestage:birthwt,data=lbwt)
summary(lfit)$coefficients
```

```
##                      Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)     -1.2584719300 5.8038051183 -0.2168357 0.8287965594
## gestage          0.7873236476 0.2087286735  3.7719956 0.0002800541
## birthwt          0.0137084745 0.0052930812  2.5898856 0.0110948349
## gestage:birthwt -0.0003137616 0.0001833162 -1.7115869 0.0902018614
```

- $\hat{\beta}_3 = -.000314$ is the estimated difference between the slopes for gestage, in 2 groups defined by a one-unit change in `birthwt`.
- E.G. Define two groups: one for babies with the median `birthwt` of 1155g and another for babies with `birthwt` 1156g.
  - In babies with `birthwt` 1156g, the slope for gestage is estimated to be 0.000314 **less** than in babies with `birthwt` 1155g (since $\hat{\beta}_3$ is negative).

What is the effect of gestage on headcirc in babies with a birthwt of 1156g?

- The linear predictor or population mean is

$$\mu_{Y|X_1,X_2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- In babies with birthwt of $x_2 = 1156$g this simplifies to

$$\begin{aligned} \mu_{Y|X_1,1156} &= \alpha + \beta_1 X_1 + \beta_2\, 1156 + \beta_3 X_1\, 1156 \\ &= \alpha + \beta_2\, 1156 + \beta_1 X_1 + \beta_3 X_1\, 1156 \\ &= \alpha + \beta_2\, 1156 + (\beta_1 + \beta_3\, 1156) X_1 \end{aligned}$$

- The slope for gestage (X_1) in babies with a birthwt of $x_2 = 1156$g is therefore $\beta_1 + \beta_3\, 1156$
- **In babies with a birthwt of 1156g**, we estimate that the effect of gestage on headcirc is

$$\hat{\beta}_1 + \hat{\beta}_3\, 1156 = 0.787 - 0.000314 \times 1156 = 0.424;$$

  i.e., a one-week increase in gestage is associated with an estimated 0.424cm increase in headcirc

Does `birthwt` modify the effect of `gestage` on `headcirc`?

- To address this question, let's test for statistical interaction between `birthwt` and `gestage` at the 5% level.

```
summary(lfit)$coefficients
```

```
##                      Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)     -1.2584719300 5.8038051183 -0.2168357 0.8287965594
## gestage          0.7873236476 0.2087286735  3.7719956 0.0002800541
## birthwt          0.0137084745 0.0052930812  2.5898856 0.0110948349
## gestage:birthwt -0.0003137616 0.0001833162 -1.7115869 0.0902018614
```

- Compare the *p*-value for the interaction term to the level 0.05.
  - Since the *p*-value is 0.09, we retain the null hypothesis of no interaction.
- Conclude that there is insufficient statistical evidence to claim that `birthwt` modifies the effect of `gestage` on `headcirc`.
- Though `birthwt` is not declared an effect modifier, it could still confound the association between `gestage` and `headcirc` ...

Confounding Variables

Changing the role of `birthwt`

- Having declared `birthwt` not to be an effect modifier, we may now consider it as a potential **confounder** of the association between `gestage` and `headcirc`.
- Look at the relationship between head circumference, Y, and gestational age, X_1, both adjusting for and ignoring birth weight, X_2.
- If analyses of an association between Y and X_1 with and without X_2 give "meaningfully different" estimates of the slope for X_1, then X_2 is declared to be a confounder.
- Definition of "meaningfully different" depends on context.
- One rule-of-thumb: If the estimated slope $\hat{\beta}_1$ changes by more than 10% when X_2 is excluded, then X_2 is a confounder (Budtz-Jorgensen et al. 2007, Annals of Epidemiology).
  - **Note**: No statistical test for confounding is involved.

Example: `birthwt` as confounder

```r
coefficients(lm(headcirc ~ gestage + birthwt,data=lbwt))
```

```
## (Intercept)     gestage     birthwt
## 8.308015388 0.448732848 0.004712283
```

```r
coefficients(lm(headcirc ~ gestage,data=lbwt))
```

```
## (Intercept)     gestage
##   3.9142641   0.7800532
```

- Measure change in the estimate of β_1 relative to the fitted model that *includes the confounding variable*, as this is considered the safer estimate of the true effect.
  - Specifically, look at change as % of this estimate.
- The percent change in $\hat{\beta}_1$ is $|0.445 - 0.780| / |0.445| \times 100\% = 75\%$.
  - As this is larger than 10%, `birthwt` would be considered a confounder by the rule-of-thumb.

Interpreting slope for `gestage` when `birthwt` is a confounder

```r
coefficients(lm(headcirc ~ gestage + birthwt,data=lbwt))
```

```
## (Intercept)     gestage     birthwt
## 8.308015388 0.448732848 0.004712283
```

- Interpretations of the slope for `gestage`:
  - "For a given birth weight, a one-week increase in gestational age is associated with an estimated 0.449cm increase in head circumference."
  - Or, "A one-week increase in gestational age is associated with an estimated 0.449cm increase in head circumference, after adjusting for the effect of birth weight."