

Interaction(s) among variables

Jianrong Wu PhD

Cardiovascular disease is the leading cause of death in the United States. The US Preventive Services Task Force provides detailed recommendations on preventive aspirin use based on age groups for both men and women.¹ Since aspirin has been shown to be cardioprotective, isn't it a good idea to provide a more generalized recommendation?

Although the recommendation suggests a calculated balance between benefit (cardioprotection) and harm (e.g., gastrointestinal bleeding and hemorrhagic stroke) of using aspirin for preventing heart disease (mainly reducing myocardial infarction), it is clear that men and women have different responses, and there is an interaction between gender (male vs. female) and aspirin use (yes vs. no) for cardioprotection.

In statistics, if the effects of one variable (factor) in a statistical model depend in some way on the presence or absence of other variable(s), then we say that an interaction(s) is present. In general, evaluating the effect of interaction is more difficult than evaluating the main effect of individual variables.²

1. Two-way interaction

As the simplest type of interaction, two-way interaction occurs, when the effects of one independent variable (e.g., aspirin use) on the dependent variable (myocardial infarction) differ at different levels of another independent variable (gender).

(A) Dependent variable:

Regression models are commonly used for evaluating interaction. In a regression model, the dependent (outcome) variable (Y) can have either a con-

tinuous or discrete (e.g., categorical or count data) distribution. For example, if the dependent variable is continuous and approximately follows a normal distribution, simple linear regression can be used in data analysis. If the dependent variable is dichotomous, or of a count nature, then logistic or log-linear regression can be used to meet the statistical requirements.

(B) Independent variables

For two-way interaction, the two independent variables (X_1 and X_2) interacting with each other can be either continuous or categorical, i.e., both are continuous, both are categorical, or one continuous and one categorical. In addition, you can have many additional independent variables in your model, as long as they do not interact with the former two variables.

(C) Statistical test of interaction

One of the two commonly used strategies for dealing with interaction was adding an interaction term in the statistical model, and performing a formal statistical test. Assuming Y is the dependent variable, and α and β are the main effects of the two independent variables X_1 and X_2 respectively, then, without considering the interaction effect between the two variables, the (reduced) regression model is,

$$Y = \mu + \alpha X_1 + \beta X_2 + \varepsilon,$$

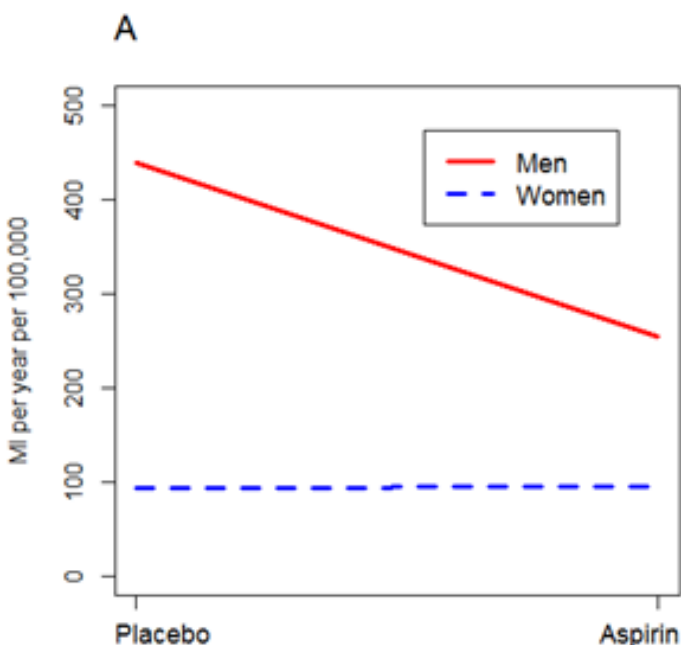
where ε is a random error term.

Corresponding author: Jianrong Wu PhD
Contact Information: Jianrong.Wu@STJUDE.ORG
DOI: 10.12746/swrccc2016.0414.189

By adding the interaction term, the (full) regression model is,

$$Y = \mu + \alpha X_1 + \beta X_2 + \gamma X_1 X_2 + \varepsilon.$$

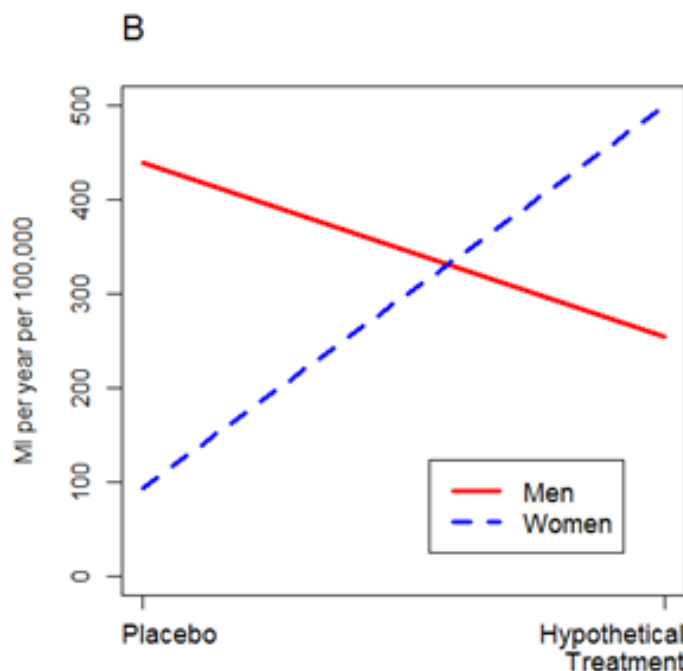
The effect of interaction (γ) can be tested by using a likelihood ratio test (LRT),³ which asymptotically follows a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters of the full and the reduced models.⁴



In the aspirin use study, aspirin has been shown to be effective in reducing myocardial infarction (MI) in men but not in women (Figure above; A). In other words, there was an interaction between gender and aspirin use for preventing myocardial infarction, and thus it makes sense to investigate the effects of aspirin use for men and women separately. The recommendation made by the US Preventive Services Task Force was not only gender specific, but age specific, indicating that besides gender, people at different ages might have different responses (both benefit and harm) to aspirin use as well. Note that in panel A of the figure above, the two lines did not cross, and thus besides the interaction effect, it makes sense to eval-

(D) Interpreting interactions

In general, models with interaction effect should include the main effects of the variables used to compute the interaction term even if the main effects are not significant. In situations in which both variables are categorical variables, the problem translates into interpreting interactions in a two-way ANOVA.



uate the main difference between men and women as well. However, in the hypothetical case presented in panel B, the two lines crossed, i.e., men who received the placebo had higher MI rate than women, and men who received the hypothetical treatment had lower MI rate than women. As a result, the difference between men and women is reversed for the placebo and hypothetical treatment, and thus comparing the difference between men and women would not be necessary.

Evaluating the interaction between a categorical and a continuous variable is basically to test the equality of regression slopes of the continuous variable across the categorical variable levels.

For the interaction between two continuous variables, it means that the relationship (regression slope) between one continuous independent variable and the dependent variable changes, with the changes in value of a second continuous independent variable. To make the interpretation more straightforward, one approach is to hold one independent variable (X_1) constant at different combinations of high/low/mean values, e.g., one standard deviation above/below the mean, as well as the mean value. Specifically, we can center X_1 to the above specified values by adding or subtracting a constant to make the value 0 meaningful. By subtracting the mean from X_1 forces the mean of the new X_1 to be 0. Subsequently, the slope of the dependent variable on the other independent variable (X_2) can be interpreted as the relationship between them when the X_1 is at its mean. In fact, by including a combination of high/low/mean values, we can present the interaction between two continuous independent variables in a similar way as we present models with at least one categorical variable.

Many statistical software packages can be used for evaluating interactions among variables. Although the fundamentals are the same, the choice of the test procedure depends largely on the distribution of the dependent variable. We will use SAS as an example to demonstrate these differences.

i. Binary dependent variable

If the dependent variable is binary, then the SAS proc logistic or proc glm procedures can be used. Let X_1 be a categorical independent variable, X_2 be a continuous independent variable, and Y be the dependent variable; where the data is stored in dat, then we can use,

```
proc logistic data=dat descending;
class X1; /*other categorical covariates that do not
interact with X1 and X2*/
model Y = X1 X2 X1*X2; /*other covariates that do
not interact with X1 and X2*/
run;
```

The class statement tells SAS that X_1 is a categorical variable. For situations where both X_1 and X_2 are categorical variables, then we need to include both in the class statement, while noting that the option descending is used by default to be consistent with how the outcome variable is coded.

ii. Continuous dependent variable

If the dependent variable has a normal distribution, then the SAS proc reg, proc glm, proc mixed procedures can be used. Again, let X_1 be a categorical independent variable, X_2 be a continuous independent variable, and Y be the dependent variable; then we can use proc glm,

```
proc glm data=dat;
class X1; /*other categorical covariates that do not
interact with X1 and X2*/
model Y = X1 X2 X1*X2; /* other covariates that
do not interact with X1 and X2*/
run;
```

Note that the coding for the class and model statements is quite similar to the proc logistic procedure.

2. High-order interactions

Although high-order interaction is a natural extension of the two-way interaction, in reality, epidemiological/biomedical studies are very often designed to avoid having high-order interactions; not only is the effect difficult to detect and interpret, but also the result is hard to generalize. For example, in most epidemiological studies, the main goals usually focus on identifying single risk factors that contribute to the study outcome, with possible interaction (e.g. two-way interaction) among risk factors ignored. On the other hand, should a high-order interaction be a concern; the statistical model used for data analysis is supposed to include all main effects, as well as lower-order interactions. If the high-order interaction is significant and theoretically possible, then the in-

terpretation should in general focus on reporting the highest-order interaction and its simple effects.

3. Additional issues

While formally testing interaction in a regression model is one of the strategies for dealing with interaction, stratified analysis is another option, when one or both of the independent variables are categorical. However, by stratification, the number of subjects in each stratum (or certain stratum) can be small (and unequal), which results in reduced (inconsistent) statistical power. In fact, studies investigating interactions in general require considerably larger sample size than those investigating the main effects. For example, a 2×2 factorial fixed effects ANOVA with equal cell sizes requires four times of observations for detecting an interaction comparing to those required for detecting a main effect.⁵

Author affiliation : Jianrong Wu is an associated member in the Biostatistics Department at the St. Jude Children's Research Hospital in Memphis, TN.

Submitted: 3/14/2016

Accepted: 4/12/2016

Published electronically: 4/15/2016

REFERENCES

1. U.S. Preventive Services Task Force, Aspirin for the Prevention of Cardiovascular Disease: U.S. Preventive Services Task Force Recommendation Statement. *Annals of Internal Medicine*, 2009; 150(6):396-404.
2. Schwartz S. Modern Epidemiologic Approaches to Interaction: Applications to the Study of Genetic Interactions. Institute of Medicine (US) Committee on Assessing Interactions Among Social, Behavioral, and Genetic Factors in Health; Hernandez LM, Blazer DG, editors. Washington (DC): National Academies Press (US); (2006)
3. Casella G, Berger RL. *Statistical Inference* (2nd ed.) Duxbury Press. (2001).
4. Wilks SS. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* 1938; 9: 60-62.
5. Fleiss JL. *The Design and Analysis of Clinical Experiments*. NY: Wiley and Sons; (1986).