MLR: Interaction Models

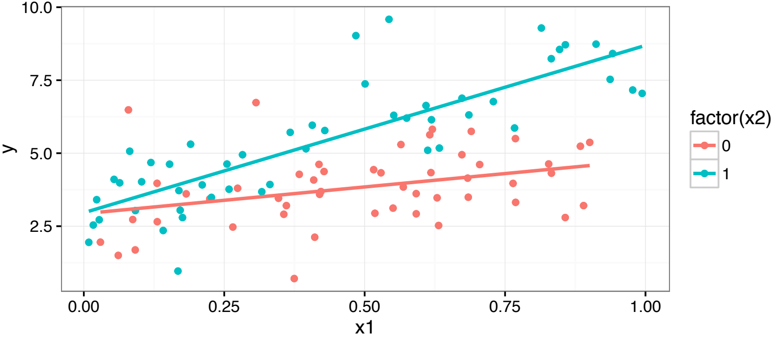
Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

What is interaction?

Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable.



Interaction vs. confounding

Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable. E.g. you could hypothesize that the true relationship between physical activity level and cancer risk may be different for men and women.

Definition of confounding

Confounding occurs when the measurable association between two variables is distorted by the presence of another variable. Confounding can lead to biased estimates of a true relationship between variables.

- It is important to include confounding variables (if possible!) when they may be biasing your results.
- Unmodeled interactions do not lead to "biased" estimates in the same way that confounding does, but it can lead to a richer and more detailed description of the data at hand.

Some real world examples?

How to include interaction in a MLR

Model A: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

Key points

- "easily" conceptualized with 1 continuous, 1 categorical variable
- models possible with other variable combinations, but interpretation/visualization harder
- two variable interactions are considered "first-order" interactions (often used to define a class of models)
- still a **linear** model, but no longer a strictly **additive** model

How to interpret an interaction model

For now, assume x_1 is continuous, x_2 is 0/1 binary.
Model A: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

How to interpret an interaction model

For now, assume x_1 is continuous, x_2 is 0/1 binary.
Model A: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
Model B: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

β_3 is the change in the slope of the line that describes the relationship of $y \sim x_1$ comparing the groups defined by $x_2 = 0$ and $x_2 = 1$.
$\beta_1 + \beta_3$ is the expected change in y for a one-unit increase in x_1 in the group $x_2 = 1$.
$\beta_0 + \beta_2$ is the expected value of y in the group $x_2 = 1$ when $x_1 = 0$.

Example interaction model with FEV data

```
library(Hmisc)
getHdata(FEV)
head(FEV)

##     id age   fev height    sex                smoke
## 1  301   9 1.708   57.0 female non-current smoker
## 2  451   8 1.724   67.5 female non-current smoker
## 3  501   7 1.720   54.5 female non-current smoker
## 4  642   9 1.558   53.0   male non-current smoker
## 5  901   9 1.895   57.0   male non-current smoker
## 6 1701   8 2.336   61.0 female non-current smoker
```

- age: Age in years
- fev: Maximum forced expiratory volume in one second
- height: Height in inchces
- sex: 'male' or 'female'
- smoker: 'current smoker' or 'non-current smoker'

Example interaction model with FEV data

$$fev_i = \beta_0 + \beta_1 age_i + \beta_2 ht_i + \beta_3 sex_{i2} + \beta_4 smoke_i + \beta_5 ht \cdot smoke_i + \epsilon_i$$

```
mi1 <- lm(fev ~ age + height + smoke + sex, data=FEV)
mi2 <- lm(fev ~ age + height*smoke + sex, data=FEV)
c(AIC(mi1), AIC(mi2))
round(summary(mi2)$coef,2)

## [1] 703.7935 700.4992
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -4.35       0.23  -19.12     0.00
## age                           0.07       0.01    7.17     0.00
## height                        0.10       0.00   21.08     0.00
## smokecurrent smoker          -2.61       1.10   -2.37     0.02
## sexmale                       0.15       0.03    4.43     0.00
## height:smokecurrent smoker    0.04       0.02    2.30     0.02
```

Example interaction model with FEV data

$$fev_i = \beta_0 + \beta_1 age_i + \beta_2 ht_i + \beta_3 sex_{i2} + \beta_4 smoke_i + \beta_5 ht \cdot smoke_i + \epsilon_i$$

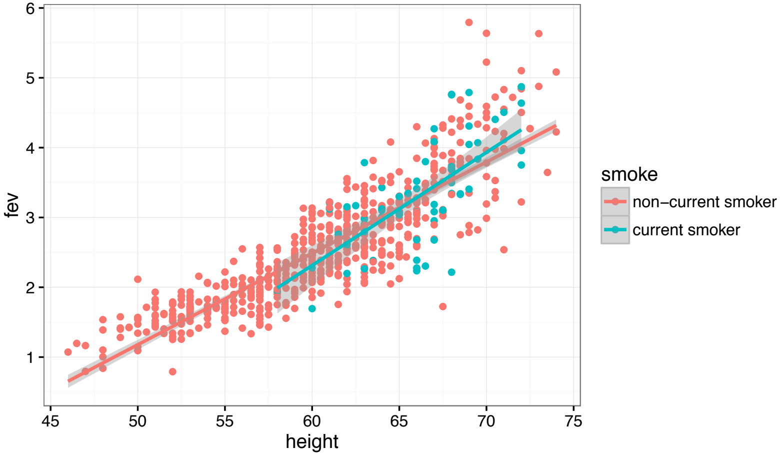
```
mi1 <- lm(fev ~ age + height + smoke + sex, data=FEV)
mi2 <- lm(fev ~ age + height*smoke + sex, data=FEV)
c(AIC(mi1), AIC(mi2))
round(summary(mi2)$coef,2)

## [1] 703.7935 700.4992
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -4.35       0.23  -19.12     0.00
## age                            0.07       0.01    7.17     0.00
## height                         0.10       0.00   21.08     0.00
## smokecurrent smoker           -2.61       1.10   -2.37     0.02
## sexmale                        0.15       0.03    4.43     0.00
## height:smokecurrent smoker     0.04       0.02    2.30     0.02
```

For current smokers, the relationship between height and FEV is stronger than in non-current smokers. In non-current smokers, we observe that a one-unit increase in height is associated with a 0.10 increase in expected FEV. In current smokers, this changes to a 0.14 increase in expected FEV.

Example interaction model with FEV data

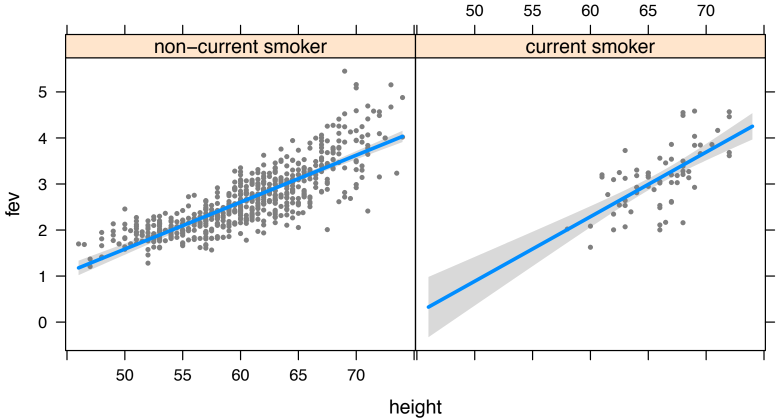
```r
ggplot(FEV, aes(height, fev, color=smoke)) +
    geom_point() + geom_smooth(method="lm")
```



Example interaction model with FEV data

The `visreg` package plots not the data but the partial residuals (a.k.a. the adjusted variable) plot.

```
library(visreg)
visreg(mi2, "height", by="smoke")
```



Example interaction model with FEV data

```
visreg(mi2, "height", by="smoke", overlay=TRUE)
```