Effect Modification (Statistical Interaction)

BIOS 6611

CU Anschutz

Week 12

Effect Modification (Statistical Interaction)

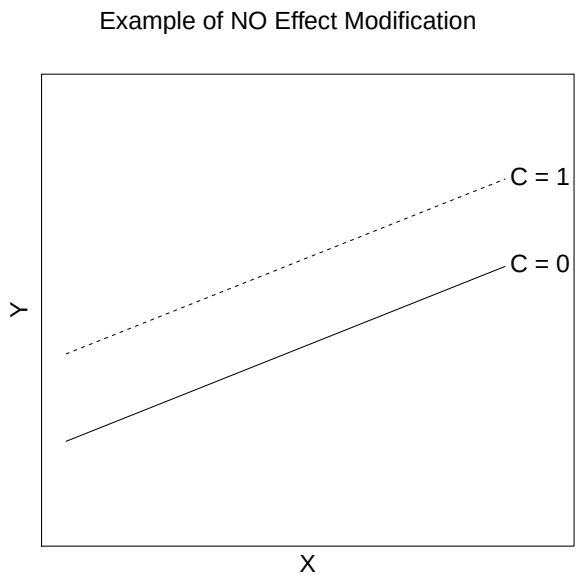
Effect Modification (Statistical Interaction)

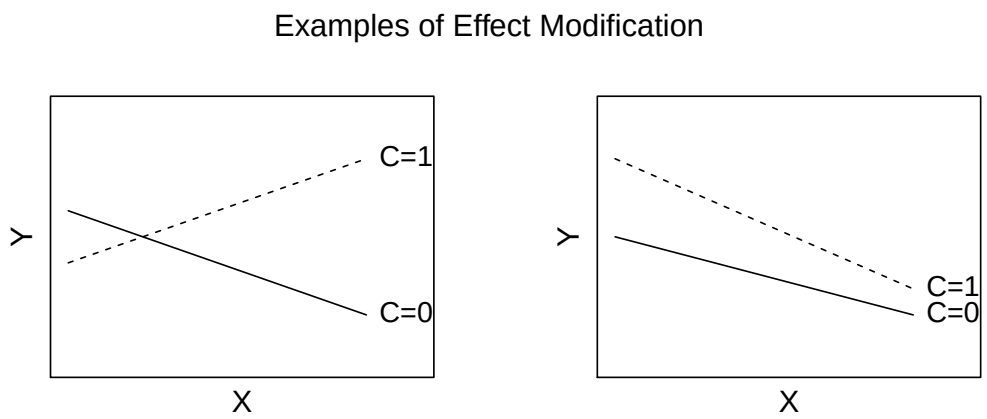
Effect modification or interaction is when the relationship between our outcome, Y, and primary explanatory variable (PEV), X, varies by some other covariate, C. Specifically the terms are defined as:

- **Effect Modification:** non-quantitative clinical or biological attribute of *population*
- **Interaction:** quantitative attribute of a *dataset*, may be scale dependent

An **effect modifier** or **moderator** is the variable that "causes" the effect modification. If a variable is an effect modifier, the role as a possible confounder is secondary.

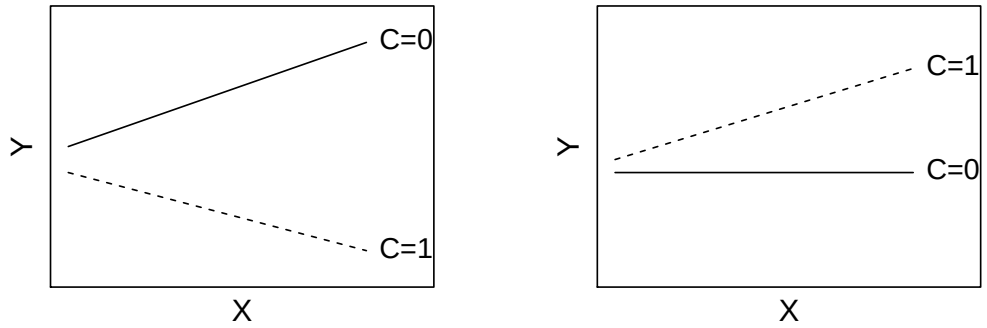
Effect Modification (Statistical Interaction)





Interaction interpretation:

- Slopes differ by level of C
- Difference between $C = 0$ and $C = 1$ depends on X



FEV Example without Interaction

Before incorporating an interaction term to our FEV data example, let's revisit the multiple linear regression results for the model including smoking status and age:

```
dat <- read.csv('FEV_rosner.csv')
mod1 <- glm(fev ~ smoke + age, data=dat)
summary(mod1)$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
```

From our regression table, the fitted regression equation is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = 0.37 + (-0.21) \times \text{Smoker} + 0.23 \times \text{Age}$$

FEV Example without Interaction

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Smoker} + \hat{\beta}_2 \times \text{Age}$$

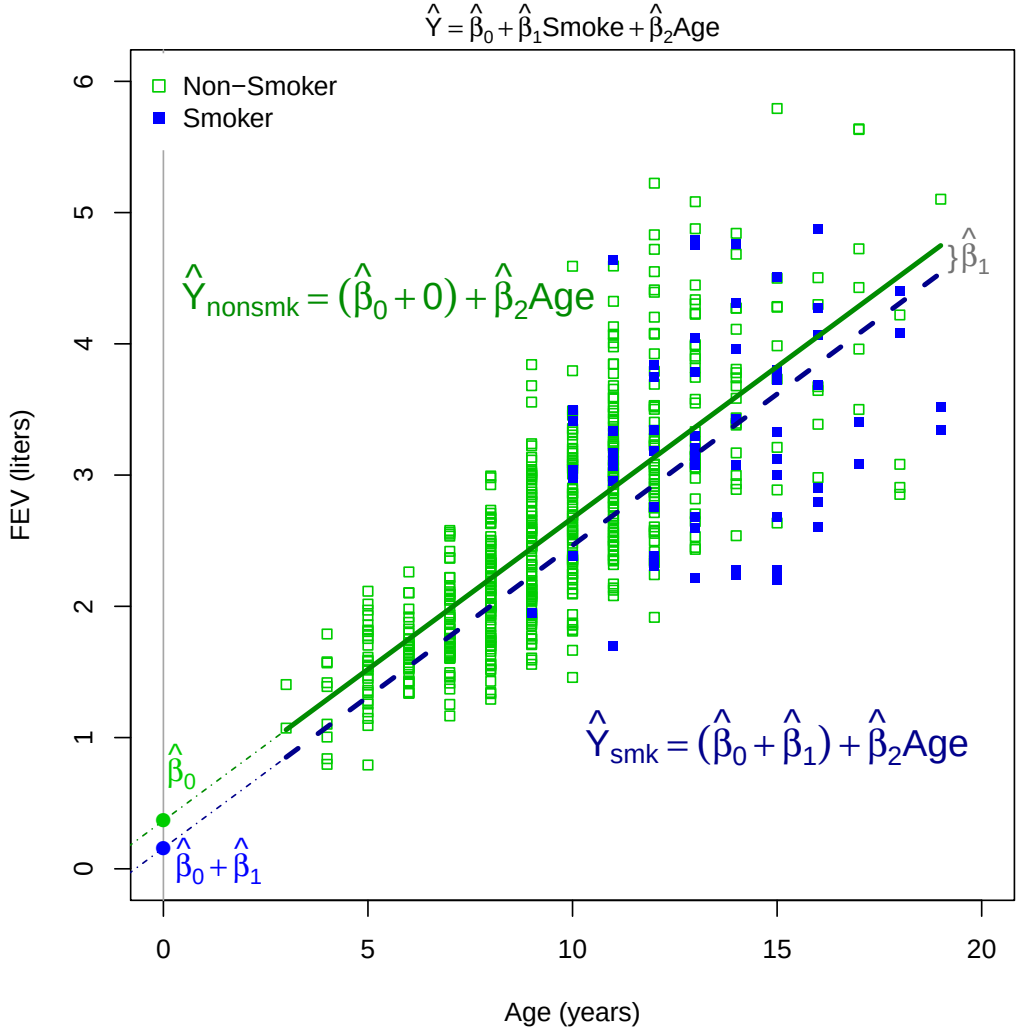
For non-smokers (Smoker $= 0$):

- $\hat{Y}_{nonsmk} = \hat{\beta}_0 + \hat{\beta}_2 \times \text{Age}$
- Intercept $= \hat{\beta}_0$
- Slope $= \hat{\beta}_2$

For smokers (Smoker $= 1$):

- $\hat{Y}_{smk} = (\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 \times \text{Age}$
- Intercept $= \hat{\beta}_0 + \hat{\beta}_1$
- Slope $= \hat{\beta}_2$

This model allows for *different intercepts* but the *same slope* in our smoking status groups.



FEV Example without Interaction

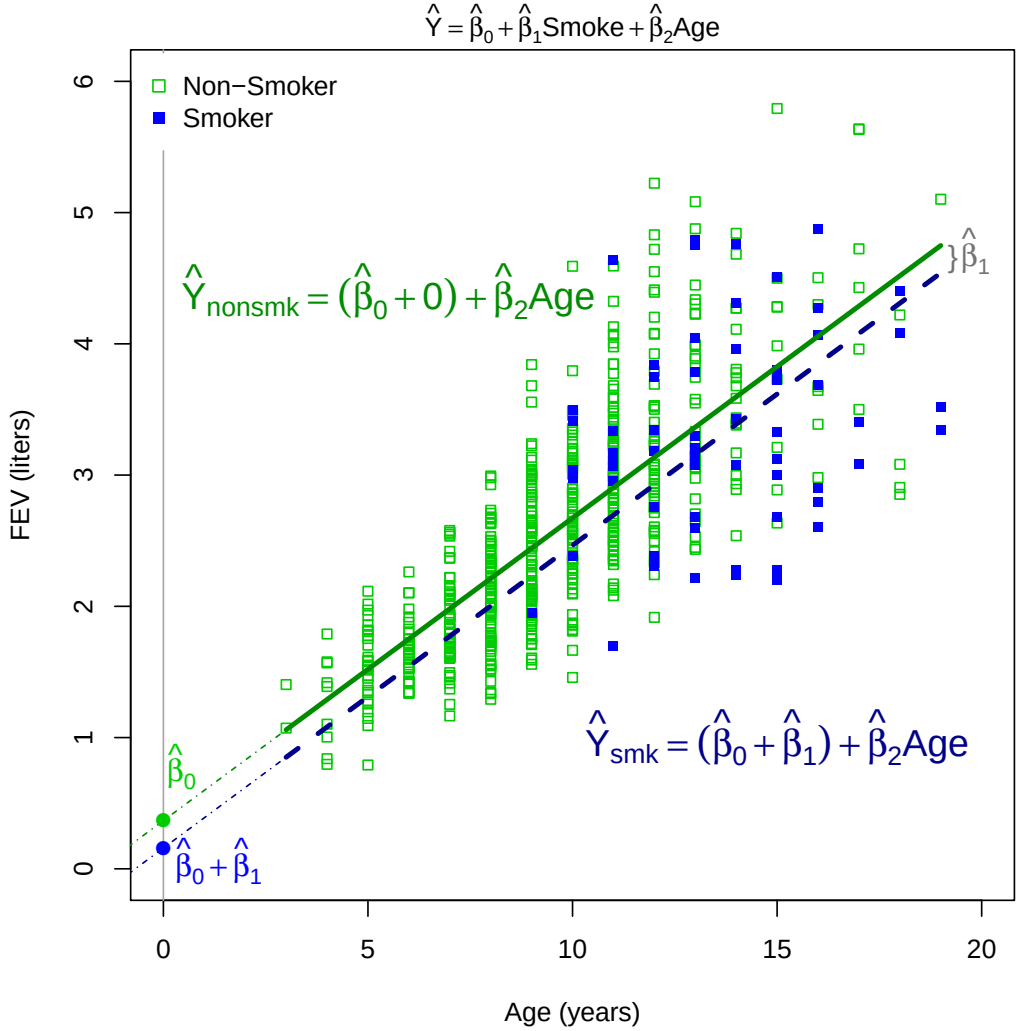
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Smoker} + \hat{\beta}_2 \times \text{Age}$$

This model allows for *different intercepts* but the *same slope*:

$$\begin{aligned}
&\hat{Y}_{smk} - \hat{Y}_{nonsmk} \\
&= [(\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 \times \text{Age}] - \\
&\quad [\hat{\beta}_0 + \hat{\beta}_2 \times \text{Age}] \\
&= \hat{\beta}_1
\end{aligned}$$

$\hat{\beta}_1$ is the "distance" between the two groups for both Age=0 and over the entire range since there is no interaction term.

H_0: the relationship between age and FEV is the exact same for smokers compared to non-smokers (i.e., H_0: $\beta_1 = 0$).



FEV Example with Interaction

Now let's explore a model that includes an interaction between age and smoking status. But first, a few coding comments.

In R, we can either create a new variable for the interaction or define the interaction directly within `glm` or `lm`:

```
dat$agesmoke <- dat$age * (dat$smoke=='smoker') # create manually

### All four models are equivalent:
# specify interaction with "*":
mod2a <- glm(fev ~ smoke + age + smoke*age, data=dat)

# specify interaction with ":":
mod2b <- glm(fev ~ smoke + age + smoke:age, data=dat)

# automatically includes non-interaction pieces:
mod2c <- glm(fev ~ smoke*age, data=dat)

# use our created interaction variable:
mod2d <- glm(fev ~ smoke + age + agesmoke, data=dat)
```

FEV Example with Interaction - SAS Code

If you are using PROC REG in SAS, we will need to create new variables for the interaction to use:

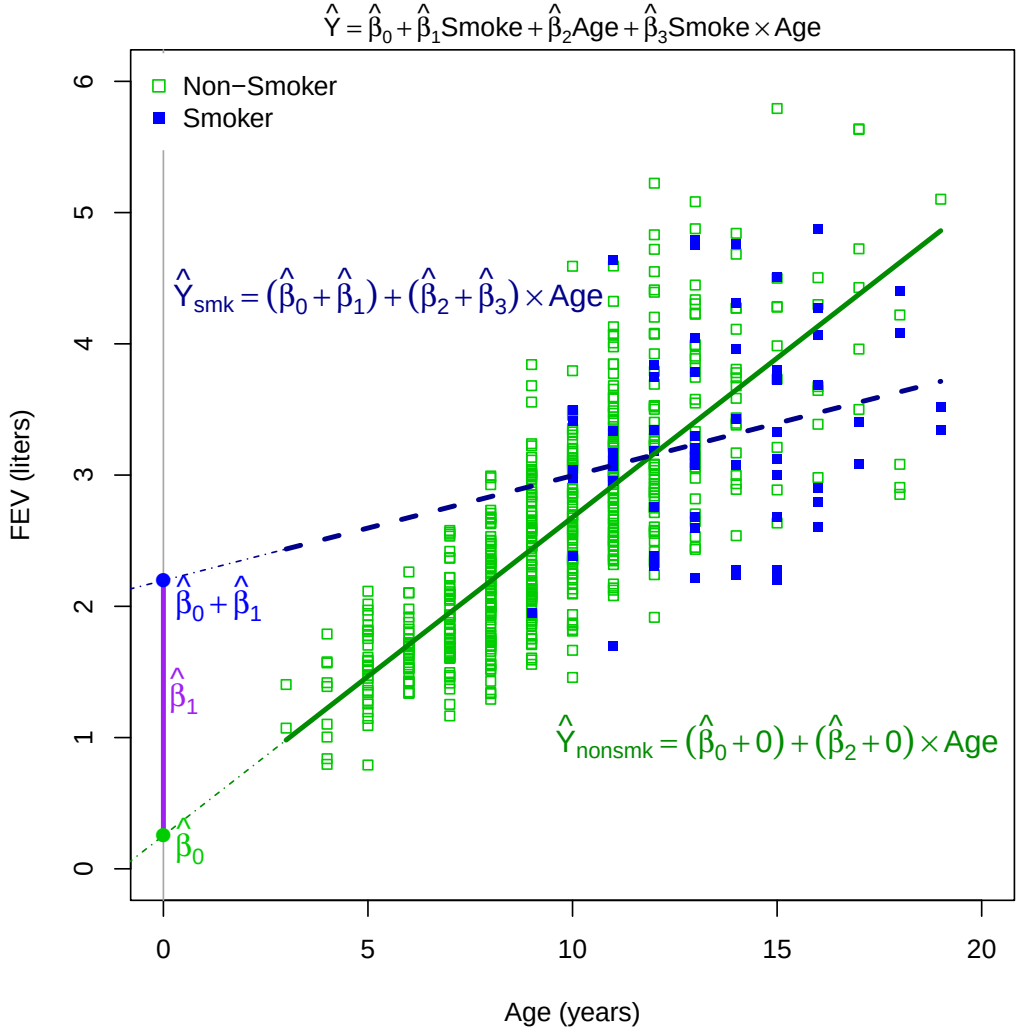
```
DATA fev;
    SET fev;

    if smoke='nonsmoker' then csmoke=0; /* smoker=1; nonsmoker=0 */
    else if smoke='smoker' then csmoke=1;

    agesmk = age*csmoke;

    LABEL   agesmk = "Age x Smoke:";
RUN;

PROC REG data=fev;
    MODEL fev = csmoke age agesmk / COVB;
RUN;
```

FEV Example with Interaction

Let $X_1 = 1$ for smokers and X_2 be for age, then
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$

For non-smokers ($X_1 = 0$):

- $\hat{Y}_{nonsmk} = \hat{\beta}_0 + \hat{\beta}_2 \times X_2$
- Intercept $= \hat{\beta}_0$
- Slope $= \hat{\beta}_2$
- H_0: no association between age and FEV for non-smokers (i.e., $\beta_2 = 0$)



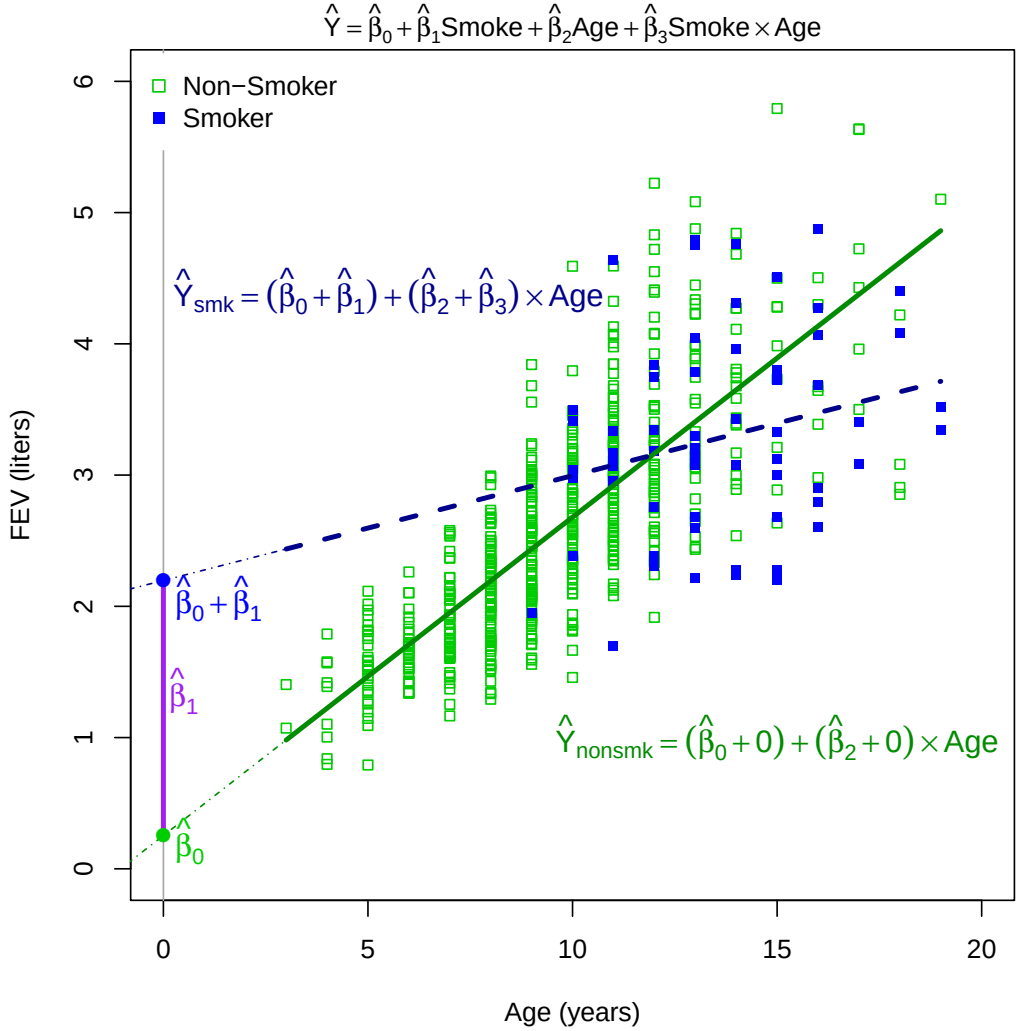
FEV Example with Interaction

Let $X_1 = 1$ for smokers and X_2 be for age, then
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$

For smokers ($X_1 = 1$):

- $\hat{Y}_{smk} = (\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_3) X_2$
- Intercept $= \hat{\beta}_0 + \hat{\beta}_1$
- Slope $= \hat{\beta}_2 + \hat{\beta}_3$
- H_0: no association between age and FEV for smokers (i.e., $\beta_2 + \beta_3 = 0$)

This model allows for *different intercepts* and *different slopes*.



FEV Example with Interaction

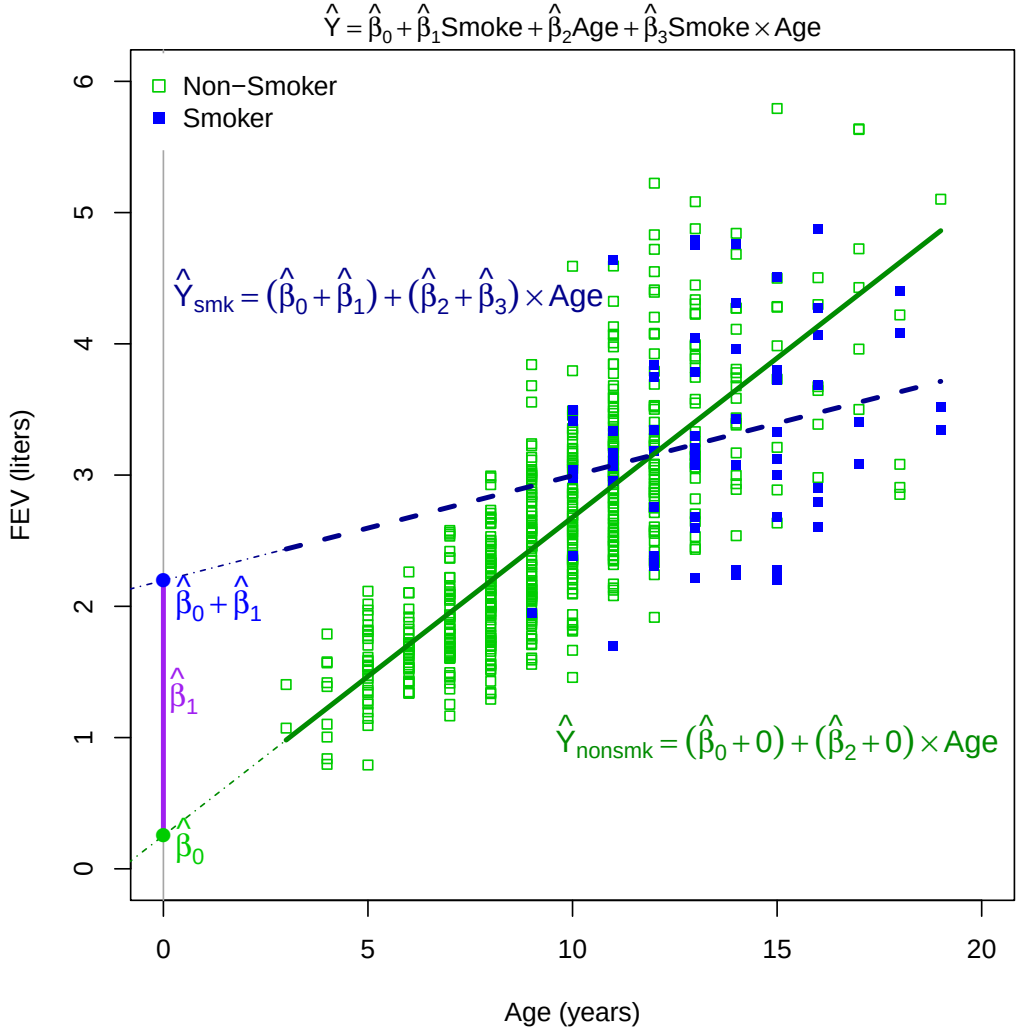
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

This model allows for different slopes and intercepts for each group:

Intercept Difference:
$\hat{Y}_{smk\&age=0} - \hat{Y}_{nonsmk\&age=0}$
$= [(\hat{\beta}_0 + \hat{\beta}_1)] - [\hat{\beta}_0] = \hat{\beta}_1$

Slope Difference for Smokers and Non-smokers:
$= [(\hat{\beta}_2 + \hat{\beta}_3)] - [\hat{\beta}_2] = \hat{\beta}_3$

H_0: the relationship between age and FEV does not differ for non-smokers compared to smokers (i.e., H_0: $\beta_3 = 0$).



FEV Example with Interaction

```
summary(mod2a)$coefficients
```

```
##                    Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)       0.2533955 0.08265075  3.065859  2.260474e-03
## smokesmoker       1.9435707 0.41428463  4.691390  3.309624e-06
## age               0.2425584 0.00833154 29.113274 6.504859e-120
## smokesmoker:age -0.1627027 0.03073753 -5.293290  1.645078e-07
```

From our regression table, the fitted regression equation is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 = 0.25 + 1.94 X_1 + 0.24 X_2 + (-0.16) X_1 X_2$$

Does the effect of smoking depend on age? (i.e., Does the effect of smoking differ for different ages?)

Does the effect of age depend on smoking status? (i.e., Does the effect of age differ for smokers and non-smokers?)

Interpretation of the Beta Coefficients

Interpretation of the Beta Coefficients

Our fitted regression equation for our FEV example is

$$\hat{Y} = 0.25 + 1.94 \times \text{Smoker} + 0.24 \times \text{Age} + (-0.16 \times \text{Smoker} \times \text{Age})$$

The meaning of each β is:

- β_0: average FEV for non-smokers at age 0 (i.e., all $X = 0$)
- β_1: difference in FEV at age 0 between smokers and non-smokers (i.e., difference in intercepts)
- β_2: slope for *non-smokers* (increase in FEV per year of age for non-smokers)
- β_3: *difference* in slope between smokers and non-smokers

Scientific Interpretation

The interpretation of an interaction depends on which variable is being considered the PEV and which is the effect modifier.

In our example, we are interested in whether smoking modifies the relationship between FEV and age:

$$\hat{Y} = 0.25 + 1.94 \times \text{Smoker} + 0.24 \times \text{Age} + (-0.16 \times \text{Smoker} \times \text{Age})$$

The regression equation for non-smokers:

$$\begin{aligned}\hat{Y}_{nonsmk} =& 0.25 + 1.94 \times 0 + 0.24 \times \text{Age} + (-0.16 \times 0 \times \text{Age}) \\ =& 0.25 + 0.24 \times \text{Age}\end{aligned}$$

The regression equation for smokers:

$$\begin{aligned}\hat{Y}_{smk} =& 0.25 + 1.94 \times 1 + 0.24 \times \text{Age} + (-0.16 \times 1 \times \text{Age}) \\ =& (0.25 + 1.94) + (0.24 - 0.16) \times \text{Age} \\ =& 2.19 + 0.08 \times \text{Age}\end{aligned}$$

Testing the Null Hypothesis

For *non-smokers*, our fitted regression is:

$$\hat{Y}_{nonsmk} = \hat{\beta}_0 + \hat{\beta}_2 \times \text{Age}$$

H_0: no association between age and FEV for non-smokers:

$$H_0 : \beta_2 = 0 \implies t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

For *smokers*, our fitted regression is:

$$\hat{Y}_{smk} = (\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_3) \times \text{Age}$$

H_0: no association between age and FEV for smokers:

$$H_0 : \beta_2 + \beta_3 = 0 \implies t = \frac{\hat{\beta}_2 + \hat{\beta}_3}{SE(\hat{\beta}_2 + \hat{\beta}_3)}$$

where $SE(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2 \times Cov(\hat{\beta}_2, \hat{\beta}_3)}$

Testing the Null Hypothesis: Non-Smokers

```
summary(mod2a)$coefficients
```

```
##                     Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)       0.2533955 0.08265075  3.065859  2.260474e-03
## smokesmoker       1.9435707 0.41428463  4.691390  3.309624e-06
## age               0.2425584 0.00833154 29.113274 6.504859e-120
## smokesmoker:age -0.1627027 0.03073753 -5.293290  1.645078e-07
```

The point estimate, interval estimate, and uncertainty for the association between FEV and age for *non-smokers* is:

- Point Estimate: $\hat{\beta}_2 = 0.24256$ liters/year
- Interval Estimate (95% CI):
  $\hat{\beta}_2 \pm t_{n-p-1,\alpha/2} SE(\hat{\beta}_2) = 0.24256 \pm 1.96(0.00833) = (0.2262, 0.2589)$, where $t_{650,0.975} = 1.96$
- Uncertainty: $t = 29.11$ with $p < 0.001$

On average, FEV increases by 0.24 liters for a one year increase in age (95% CI: 0.23, 0.26 L) for non-smokers, which is significantly different from 0 (p<0.001).

Testing the Null Hypothesis: Smokers

In order to calculate the interval estimate and uncertainty we need to either use a function in R to conduct a general linear hypothesis test or extract the relevant information to do the calculation "by hand".

By hand, we'd need to extract the variance-covariance matrix from our `glm` (or `lm`) model:

```
vcov(mod2a) # extract variance-covariance matrix for fitted model object
```

```
##                     (Intercept)   smokesmoker           age smokesmoker:age
## (Intercept)       0.0068311473 -0.0068311473 -6.618543e-04    6.618543e-04
## smokesmoker      -0.0068311473  0.1716317529  6.618543e-04   -1.249970e-02
## age              -0.0006618543  0.0006618543  6.941456e-05   -6.941456e-05
## smokesmoker:age   0.0006618543 -0.0124997012 -6.941456e-05    9.447957e-04
```

$$
\begin{aligned}
SE(\hat{\beta}_2 + \hat{\beta}_3) &= \sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2 \times Cov(\hat{\beta}_2, \hat{\beta}_3)} \\
&= \sqrt{0.0000694146 + 0.0009447957 + 2(-0.000069415)} \\
&= \sqrt{0.0008753803} \\
&= 0.029587
\end{aligned}
$$

Testing the Null Hypothesis: Smokers

```
summary(mod2a)$coefficients
```

```
##                    Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)       0.2533955 0.08265075  3.065859  2.260474e-03
## smokesmoker       1.9435707 0.41428463  4.691390  3.309624e-06
## age               0.2425584 0.00833154 29.113274 6.504859e-120
## smokesmoker:age -0.1627027 0.03073753 -5.293290  1.645078e-07
```

The point estimate, interval estimate, and uncertainty for the association between FEV and age for *smokers* is:

- Point Estimate: $\hat{\beta}_2 + \hat{\beta}_3 = 0.24256 + (-0.16270) = 0.07986$ liters/year
- Interval Estimate (95% CI): $0.07986 \pm 1.96(0.029587) = (0.0219, 0.1378)$
- Uncertainty: $t = \frac{0.07986}{0.029587} = 2.699$ with
  p =2*pt(2.699,650,lower.tail=F)= 0.007

On average, FEV increases by 0.08 liters for a one year increase in age (95% CI: 0.02, 0.14 L) for smokers, which is significantly different from 0 (p=0.007).

Testing the Null Hypothesis: Smokers (R Code)

We can also use the `glht` function in the `multcomp` package in `R` to calculate the interval estimate and uncertainty:

```
library(multcomp)
K <- matrix(c(0,0,1,1),nrow=1) #beta2+beta3
summary(glht(mod2a, linfct=K))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = fev ~ smoke + age + smoke * age, data = dat)
##
## Linear Hypotheses:
##        Estimate Std. Error z value Pr(>|z|)
## 1 == 0  0.07986    0.02959   2.699  0.00695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

For `glm` models the normal approximation is used, whereas for `lm` models the t-distribution is used.

Testing the Null Hypothesis: Smokers (SAS Code)

In SAS we can very easily get this information in PROC GLM using the ESTIMATE statement:

```
PROC GLM;
    MODEL fev = csmoke age agesmk /CLPARM;
  /* Estimate 1 yr increase in age */
    ESTIMATE 'Age Effect: Non-Smokers' age 1;
  /* Estimate 1 yr increase in age + being a smoker */
  ESTIMATE 'Age Effect: Smokers' age 1 agesmk 1;
RUN;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Age Effect: Non-Smokers	0.24255841	0.00833154	29.11	<.0001	0.22619843	0.25891839
Age Effect: Smokers	0.07985574	0.02958684	2.70	0.0071	0.02175841	0.13795306

Testing the Null Hypothesis: Difference in Slopes

```
summary(mod2a)$coefficients
```

```
##                     Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)        0.2533955 0.08265075  3.065859  2.260474e-03
## smokesmoker        1.9435707 0.41428463  4.691390  3.309624e-06
## age                0.2425584 0.00833154 29.113274 6.504859e-120
## smokesmoker:age   -0.1627027 0.03073753 -5.293290  1.645078e-07
```

- Point Estimate: $\hat{\beta}_3$ = -0.1627 liters/year
- Interval Estimate (95% CI):
  $\hat{\beta}_3 \pm t_{n-p-1,\alpha/2} SE(\hat{\beta}_3) = -0.1627 \pm 1.96(0.0307) = (-0.2229, -0.1025)$,
  where $t_{650,0.975} = 1.96$
- Uncertainty: $t = -5.29$ with $p < 0.001$

The relationship between FEV and age differs significantly for non-smokers compared to smokers (p<0.001). Thus, on average, FEV decreases an average of 0.16 liters more per year in smokers compared to non-smokers (95% CI: -0.10 to -0.22 L/yr).

Closing Comments

We examined interactions between a categorical effect modifier and a continuous PEV. Interactions can also be between two categorical or two continuous variables. The same concepts apply for any combination of variables in an interaction.

In practice, we need a larger sample size to detect most interaction effects. Therefore, if we are planning a study and expect an interaction effect, we should power for the interaction.

It is possible to have interactions between more than two predictors, but it makes interpretation more challenging and we may not be powered to detect differences if they truly exist.