22. Effect Modification (Interaction) in Regression and Maximum Likelihood Estimators

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Ch. 11-13, 21

SAS: PROC REG, PROC GLM, PROC MIXED

Homework: Homework 9 due by 11:59 pm on November 28 (Thanksgiving)
Homework 10 due by 11:59 pm on December 5

Overview

A) Re/Preview of Topics
B) Effect Modification (Statistical Interaction)
C) Interpretation of the Beta Coefficients and Inference
D) Maximum Likelihood Estimation
E) Example of MLE in Regression

A. Review (Lecture 21)

Set up is the same for Confounders or Mediators:

Crude Model: $\hat{Y} = \hat{\beta}_{01} + \hat{\beta}_{crude}X$

Adjusted Model: $\hat{Y} = \hat{\beta}_{02} + \hat{\beta}_{adj}X + \hat{\beta}_z Z$

Covariate Model: $\hat{Z} = \hat{\gamma}_0 + \hat{\gamma}_X X$

X = PEV
Y = Outcome/Response
Z = Covariate/Potential Confounder

$$\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}} = \hat{\gamma}_X \times \hat{\beta}_Z$$

$$SE\left(\hat{\beta}_{crude} - \hat{\beta}_{adj}\right) = SE\left(\hat{\gamma}_X \times \hat{\beta}_Z\right) = \sqrt{\hat{\beta}_Z^2 Var(\hat{\gamma}_X) + \hat{\gamma}_X^2 Var(\hat{\beta}_Z)}$$

Confounders (X<-C->Y)

- Operational Criterion: $\beta_{crude} \neq \beta_{adj}$ or a 10% or 20% change in $\frac{\beta_{crude} - \beta_{adj}}{\beta_{adj}}$
- Classical Criterion #1 (X and Z association): $\hat{\gamma}_X$
- Classical criterion #2 (Z and Y associated given X): $\hat{\beta}_Z$
- Classical criterion #3: Not an intermediate step in the causal path X->C->Y

Mediators (X->C->Y)

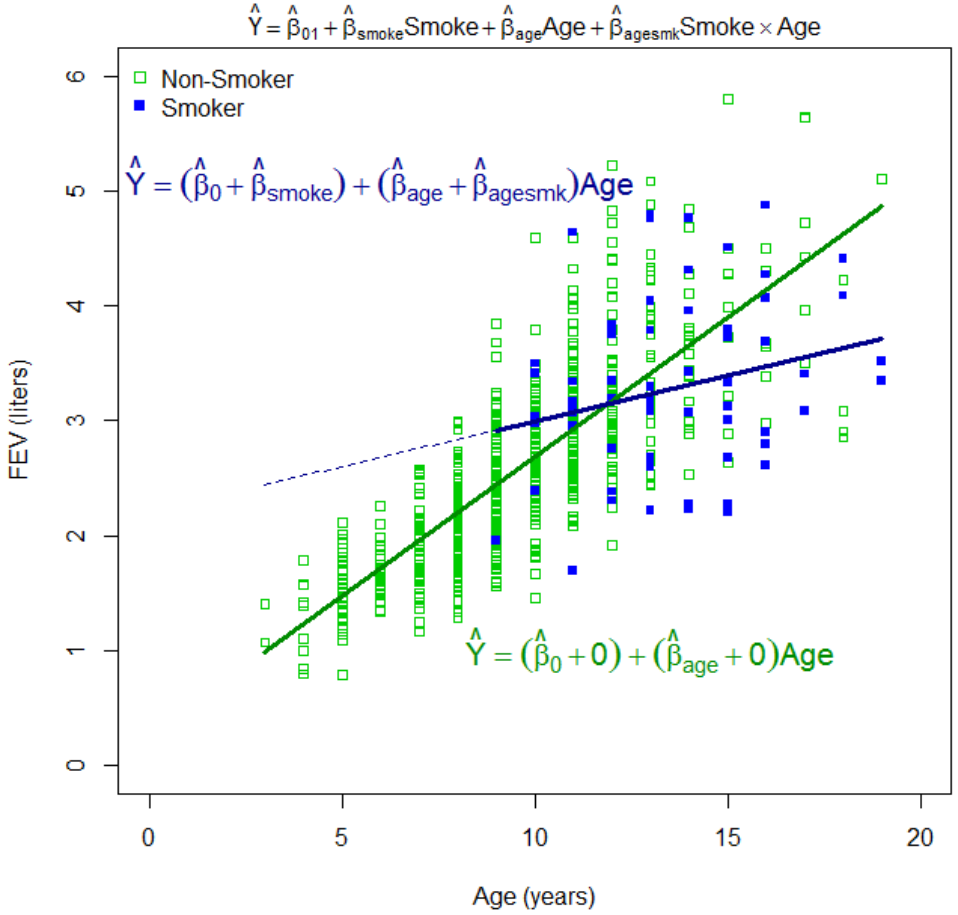
- Indirect effect: $\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}} = \hat{\gamma}_X \times \hat{\beta}_Z$
- Direct effect: $\hat{\beta}_{\text{adj}} = \hat{\beta}_{\text{crude}} - \left(\hat{\gamma}_X \times \hat{\beta}_Z\right)$
- Proportion Mediated = Indirect Effect / Total Effect $= \frac{\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}}}{\hat{\beta}_{\text{crude}}} = \frac{\hat{\gamma}_X \times \hat{\beta}_Z}{\hat{\beta}_{\text{crude}}}$

Current (Lecture 22)/ Preview (Lecture 23)

Lecture 22:

- Effect Modification (Interactions):
  - $E[FEV_i] = \beta_0 + \beta_{age} Age_i + \beta_{smoke} Smoke_i + \beta_{agesmk} Age_i \times Smoke_i$
  - Allows for different slopes (FEV vs. age) for smokers and non-smokers

- Maximum Likelihood Estimation (MLE)
  - MLE for the βs are identical to the OLS estimates
  - MLE of σ^2 is actually a biased estimate, OLS estimate is not



Lecture 23:

- Categorical Predictors
  - Indicator variables

- Test of general linear hypothesis
- Linear contrasts
- Orthogonal polynomials

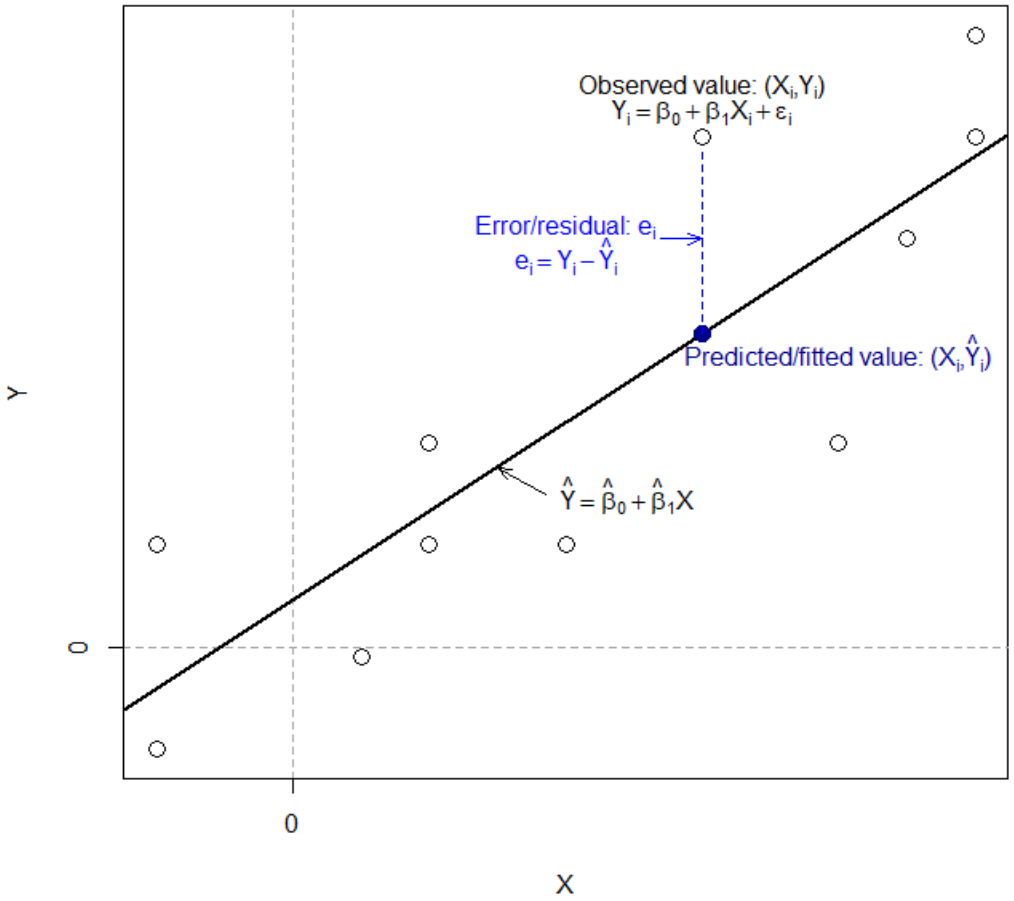
Notation Review

Right Notation:

Truth: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Expected: $E[Y_i] = \beta_0 + \beta_1 X_i$ because $E[\varepsilon_i] = 0$

Estimate: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$



Wrong Notation:

$Y_i \neq \beta_0 + \beta_1 X_i$

Implies Y vs X is a perfect line

$E[Y_i] \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$

$E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i$

$E[\hat{Y}_i] = E[\hat{\beta}_0 + \hat{\beta}_1 X_i] = \beta_0 + \beta_1 X_i$

because $E[\hat{\beta}] = \beta$

Truth vs. estimate:

$\hat{Y}_i \neq \beta_0 + \beta_1 X_i$

$Y_i \neq \hat{\beta}_0 + \hat{\beta}_1 X_i$

$\varepsilon_i \neq e_i = Y_i - \hat{Y}_i$

B. Effect Modification (Statistical Interaction)

Effect Modification or interaction: relationship between X and Y varies by C

- **Effect modification:** non-quantitative clinical or biological attribute of <u>population</u>.
- **Interaction:** Quantitative attribute of a <u>dataset</u>. May be scale dependent.

Effect modifier or moderator: variable 'causes' effect modification.

In general terms, an effect modifier:

- Affects direction and/or strength of relationship between X variable and Y
- Qualitative (gender, race, smoker, etc.) or Quantitative (age, height, etc.) variable

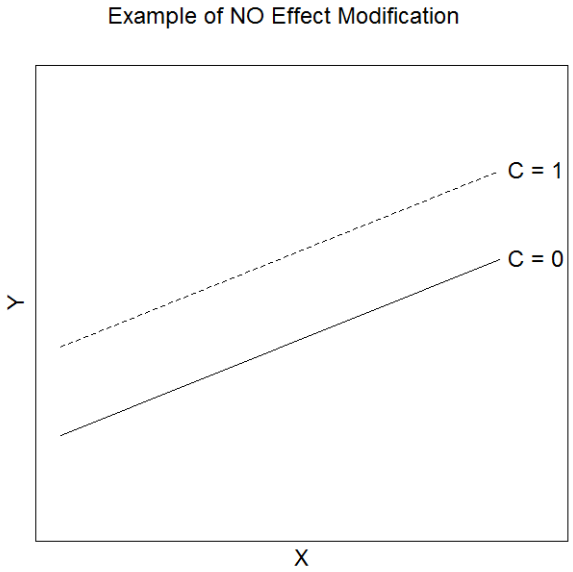
When a variable is an effect modifier

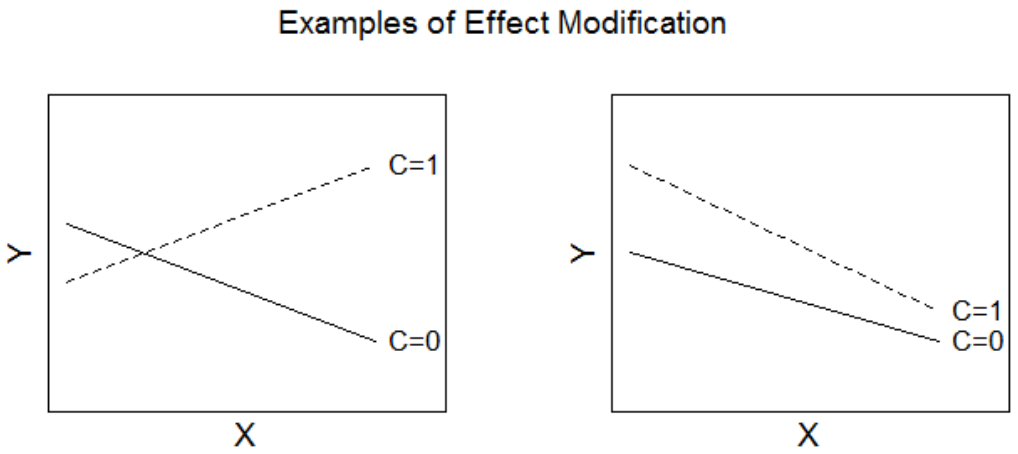
- Role as a possible confounder is secondary
- Model and interpret the interaction

<u>Application of Linear Regression:</u> Assess interactive effects of 2 (or more) independent variables with regard to the dependent variable.

Effect Modification (Statistical Interaction)

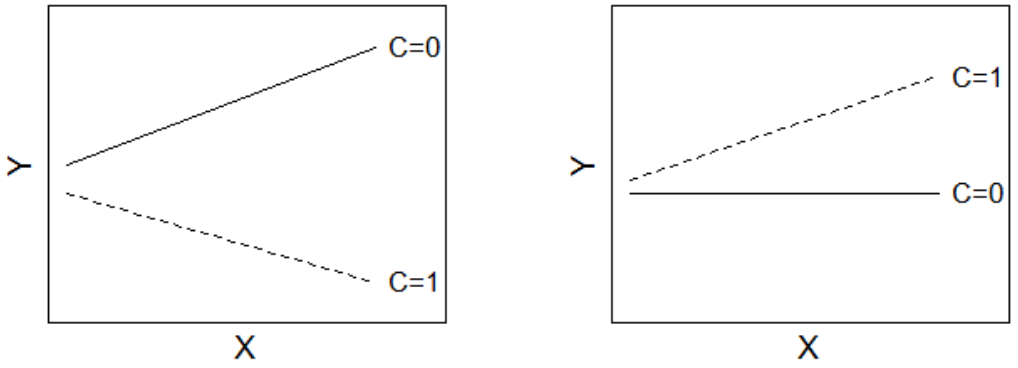
Relationship between variable (X) and outcome (Y) differs by level of third variable (C).





Interaction Interpretation:

- Slopes differ by level of factor C
- Difference between C=0 & C=1 depends on X



Example 1A: FEV data set with no 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:

```
PROC REG data=fev;
    MODEL fev = csmoke age;          /* csmoke: 0=Non-smoker; 1=Smoker */
    OUTPUT OUT=pred1 p=yhat;
RUN;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.36737	0.08144	4.51	<.0001
csmoke	1	-0.20899	0.08075	-2.59	0.0099
age	1	0.23060	0.00818	28.18	<.0001

Regression equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\widehat{FEV} = 0.37 + (-0.21) \times Smoke + 0.23 \times Age$$

FEV Example 1A: 1 equation, 2 parallel lines (no interaction) cont.

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

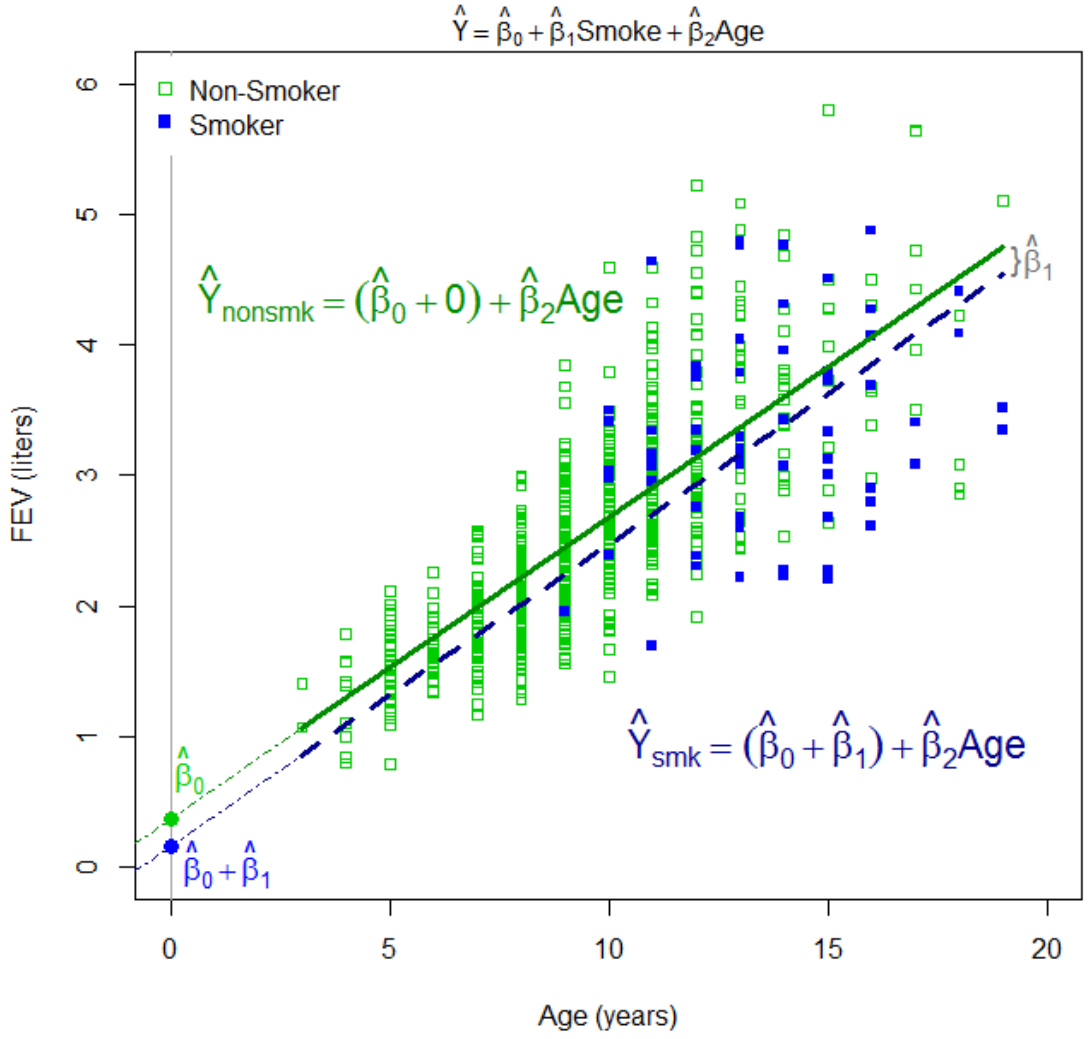
For non-smokers (Smoke=0):

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

For smokers (Smoke=1):

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

This model allows for <u>different intercepts</u> but the <u>same slope</u>.



FEV Example 1A: 1 equation, 2 parallel lines (no interaction) cont.

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

For non-smokers (Smoke=0):

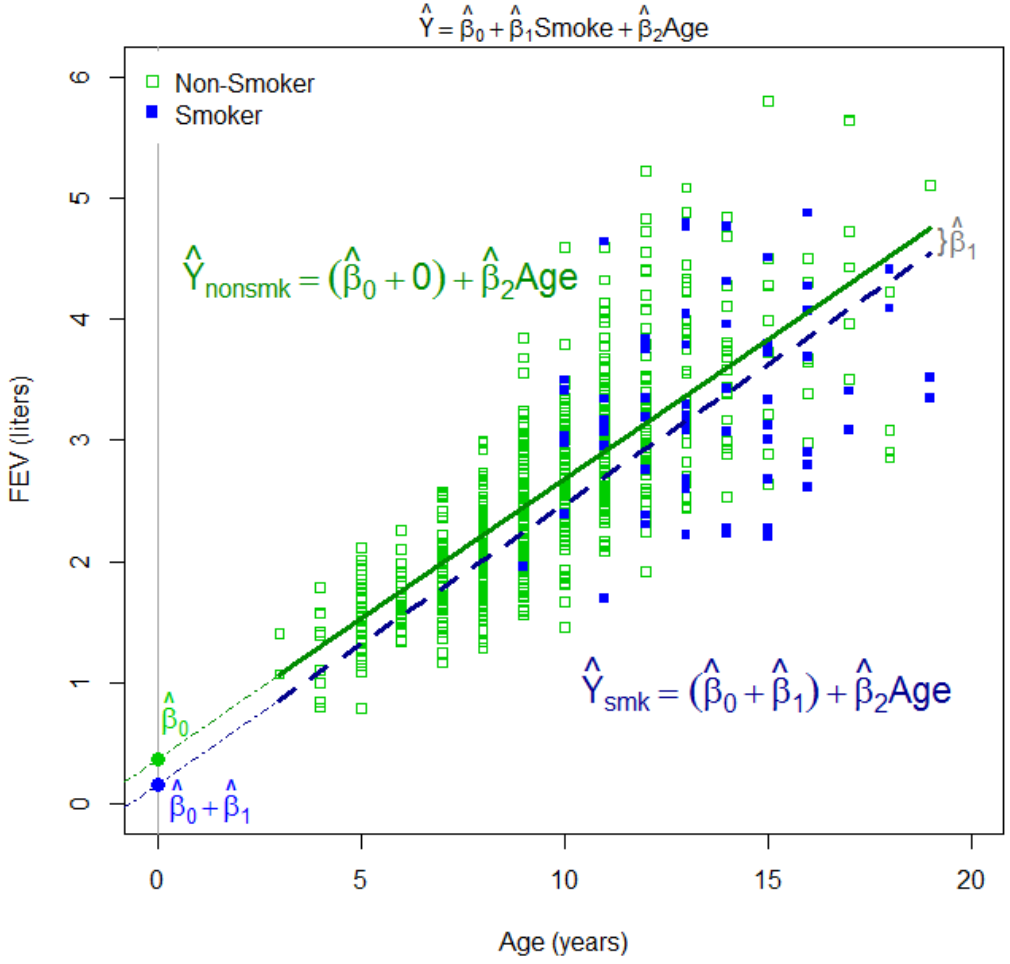
$$\hat{Y}_{nonsmk} = \hat{\beta}_0 + \hat{\beta}_2 \times Age$$

For smokers (Smoke=1):

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

This model allows for different intercepts but the same slope:

$$\hat{Y}_{smk} - \hat{Y}_{nonsmk}$$
$$= [(\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 \times Age] - [\hat{\beta}_0 + \hat{\beta}_2 \times Age]$$
$$= \hat{\beta}_1$$ (the "distance" between the two groups when 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: $\beta_1 = 0$).

Example 1B: FEV data set with an interaction term

Let's now create an interaction variable and incorporate it into our model to examine how it changes the associations between smoking status and age with FEV:

```
DATA fev;
    SET fev;

    agesmk = age*csmoke;
    sexsmk = sex*csmoke;

    LABEL   agesmk = "Age x smoke"
            sexsmk = "Sex x smoke:";
RUN;


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

FEV Example 1B: 1 equation, 2 nonparallel lines (interaction) cont.

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

For non-smokers (Smoke=0):

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

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

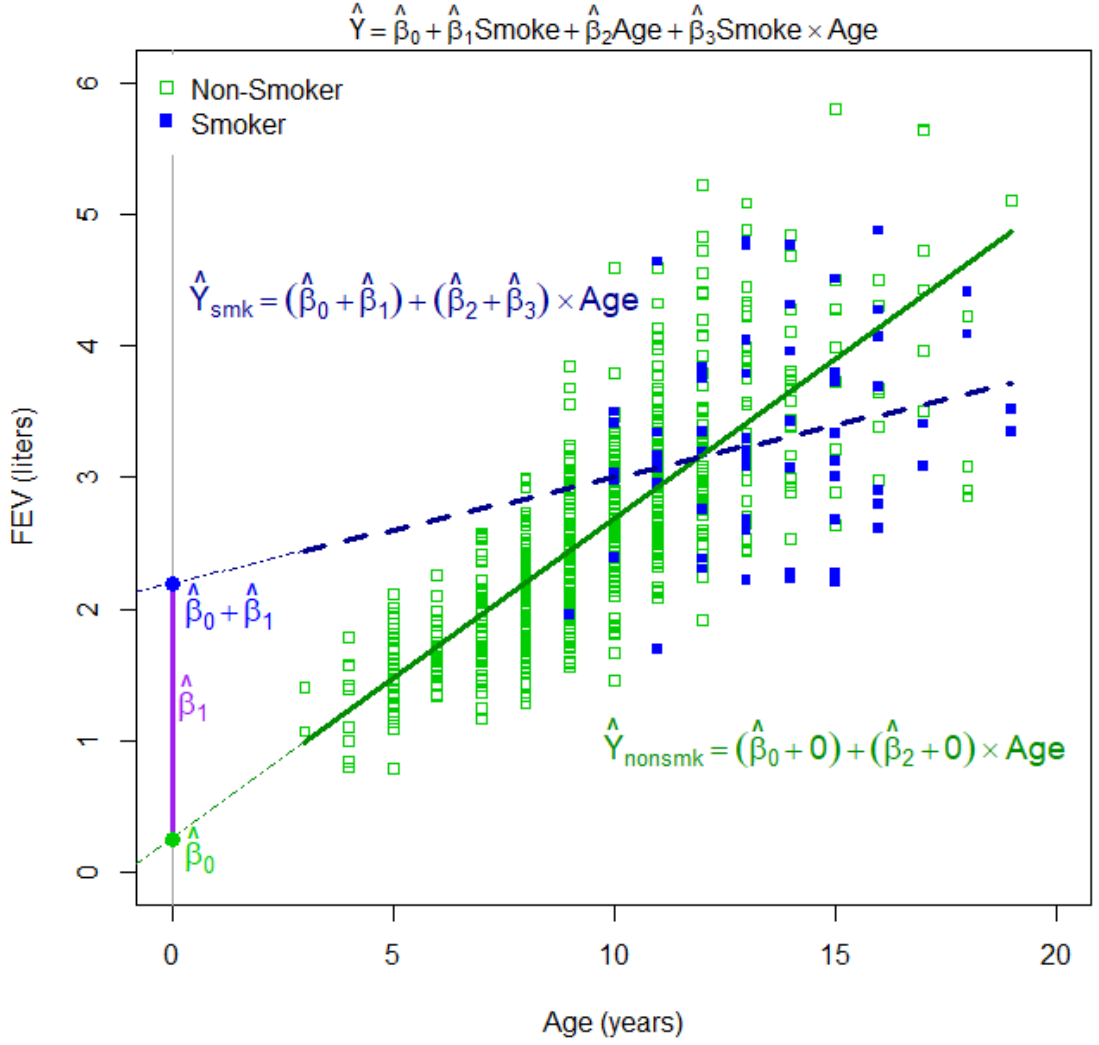
- $H_0: \beta_2 = 0$

For smokers (Smoke=1):

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

H_0: No association between age and FEV for smokers

- $H_0: \beta_2 + \beta_3 = 0$



FEV Example 1B: 1 equation, 2 nonparallel lines (interaction) cont.

Non-smokers (Smoke=0)

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

Smokers (Smoke=1)

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

Intercept for smokers minus intercept for non-smokers =

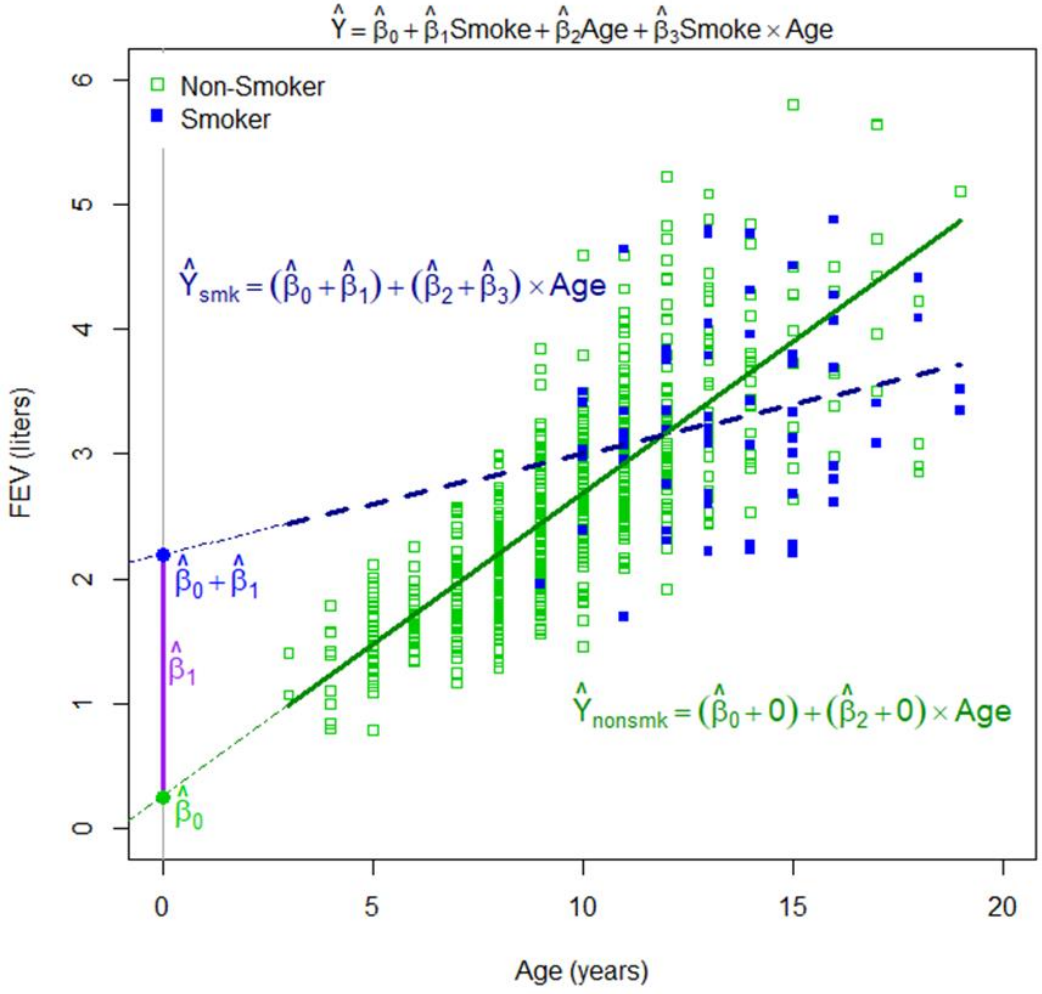
$$\hat{Y}_{smk\ \&\ Age=0} - \hat{Y}_{nonsmk\ \&\ Age=0}$$
$$= [(\hat{\beta}_0 + \hat{\beta}_1)] - [\hat{\beta}_0] = \hat{\beta}_1$$

Slope for smokers minus slope for non-smokers =

$$(\hat{\beta}_2 + \hat{\beta}_3) - \hat{\beta}_2 = \hat{\beta}_3$$

H_0: The relationship between FEV and age does not differ for non-smokers compared to smokers

- $H_0: \beta_3 = 0$



FEV Example 1B: 1 equation, 2 nonparallel lines (interaction) cont.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	291.64807	97.21602	317.11	<.0001
Error	650	199.27177	0.30657		
Corrected Total	653	490.91984			

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	0.25340	0.08265	3.07	0.0023
csmoke		1	1.94357	0.41428	4.69	<.0001
age		1	0.24256	0.00833	29.11	<.0001
agesmk	Age x smoke	1	-0.16270	0.03074	-5.29	<.0001

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?)

C. Interpretation of the Beta Coefficients

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

Regression Equation:

$$\widehat{FEV} = 0.25 + 1.94 \times Smoke + 0.24 \times Age + (-0.16 \times Smoke \times Age)$$

β_0 Average FEV for non-smokers at age 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

$$\hat{Y}_{smk} - \hat{Y}_{nonsmk} = \left[\left(\hat{\beta}_0 + \hat{\beta}_1\right) + \left(\hat{\beta}_2 + \hat{\beta}_3\right) \times Age\right] - \left[\hat{\beta}_0 + \hat{\beta}_2 \times Age\right] = \hat{\beta}_1 + \hat{\beta}_3 \times Age$$

Interpretation

Scientific interpretation: The interpretation of an interaction depends on which variable is being considered the PEV (primary explanatory variable) and which is the effect modifier.

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

$$\widehat{FEV} = 0.25 + 1.94 \times Smoke + 0.24 \times Age + (-0.16 \times Smoke \times Age)$$

Regression Equation for non-smokers (smoke = 0):

$$\begin{aligned} \widehat{FEV} &= 0.25 + 1.94 \times 0 + 0.24 \times Age + (-0.16 \times 0 \times Age) \\ &= 0.25 + 0.24 \times Age \end{aligned}$$

Regression Equation for smokers (smoke = 1):

$$\begin{aligned} \widehat{FEV} &= 0.25 + 1.94 \times 1 + 0.24 \times Age + (-0.16 \times 1 \times Age) \\ &= (0.25 + 1.94) + (0.24 - 0.16) \times Age \\ &= 2.19 + 0.08 \times Age \end{aligned}$$

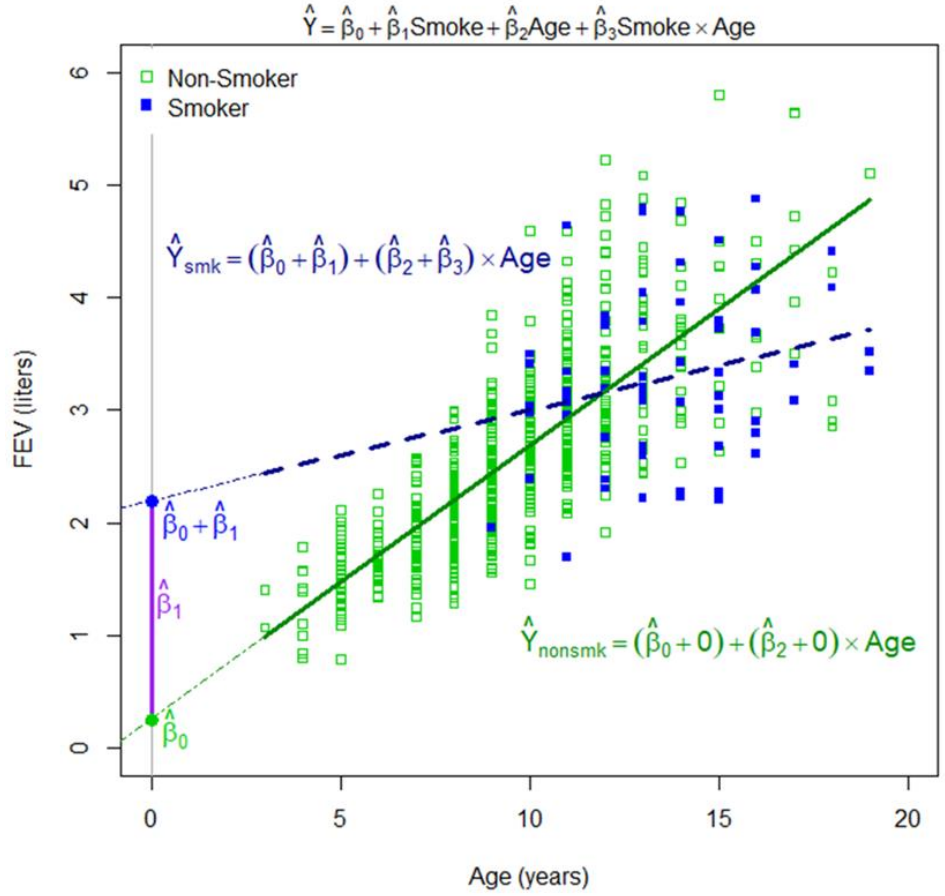
FEV Example 1B cont.: Testing the Null Hypotheses

For non-smokers (Smoke=0):

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

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

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



For smokers (Smoke=1):

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

H_0: No association between age and FEV for smokers

- $H_0: \beta_2 + \beta_3 = 0 \Rightarrow t = \frac{\hat{\beta}_2 + \hat{\beta}_3}{SE(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{\hat{\beta}_2 + \hat{\beta}_3}{\sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2Cov(\hat{\beta}_2, \hat{\beta}_3)}}$

Model output: the regression parameters and covariance matrix (from the COVB command)

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	0.25340	0.08265	3.07	0.0023
csmoke		1	1.94357	0.41428	4.69	<.0001
age		1	0.24256	0.00833	29.11	<.0001
agesmk	Age x smoke	1	-0.16270	0.03074	-5.29	<.0001

√(0.0068311473) = **0.08265** = SE(Intercept)

$\text{Cov}(\hat{\beta}_2, \hat{\beta}_3) = \text{Cov}(\hat{\beta}_{age}, \hat{\beta}_{agesmk})$

Covariance of Estimates					
Variable	**Label**	**Intercept**	**csmoke**	**age**	**agesmk**
Intercept	Intercept	**0.0068311473**	-0.006831147	-0.000661854	0.0006618543
csmoke		-0.006831147	0.1716317529	0.0006618543	-0.012499701
age		-0.000661854	0.0006618543	**0.0000694146**	**-0.000069415**
agesmk	Age x smoke	0.0006618543	-0.012499701	-0.000069415	**0.0009447957**

Covariance of Estimates

Variances: **diagonal elements**

Covariances: **off-diagonal elements**

$\text{Var}(\hat{\beta}_2) = \text{Var}(\hat{\beta}_{age})$

$\text{Var}(\hat{\beta}_3) = \text{Var}(\hat{\beta}_{agesmk})$

The association between age and FEV by smoking status

Point Estimate, Interval Estimate, and uncertainty for the association between FEV and age for **non-smokers**:

Point Estimate: 0.24256 liters/year

95% CI: $0.24256 \pm 1.96(0.00833) = (0.2262, 0.2589)$

NOTE: $t_{650,0.975} = 1.96$ (df: n-p-1)

Uncertainty: $t=29.11$, $p<.0001$

The association between age and FEV by smoking status (cont.)

Point Estimate, Interval Estimate, and uncertainty for the association between FEV and age for **smokers**:

Point Estimate: 0.24256-0.16270 = 0.07986 liters/year

Standard Error:

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

= 0.0000694146 + 0.0009447957 + 2(-0.000069415)

= 0.0008753803

SE = $\sqrt{0.0008753803}$ = 0.029587

95% CI: 0.07986 ± 1.96(0.029587) = (0.0219, 0.1378)

Uncertainty: t=0.07986/0.029587 = 2.699, p=0.007

Recode Smoking Group

Note: We can get parameter estimate, SE, and CI for smokers by recoding model to make smokers our reference group!

```
DATA fev;
     SET data.fev;

     agesmk = age*csmoke;
     sexsmk = sex*csmoke;

     nsmoke = (csmoke = 0);
     agensmk = nsmoke*age;

     LABEL   id = "ID Number"
             age = "Age (years)"
             fev = "FEV (Liters)"
             height = "Height"
             sex = "Sex"
             csmoke = "Current Smoker"
             agesmk = "Age x smoke"
             sexsmk = "Sex x smoke"
             nsmoke = "Non Smoker"
             agensmk = "Age x Non-smoke";

RUN;
```

IF 'true' then assign
nsmoke = 1;

csmoke	0 = non-smoker 1= smoker
nsmoke	0 = smoker 1 = non-smoker

Equivalent to:

IF csmoke = 0 THEN nsmoke = 1;
IF csmoke = 1 THEN nsmoke = 0;

Interpretation of the beta coefficients with reversing the coding of the smoking variable

Equation: $\hat{Y} = \hat{\beta}_0^* + \hat{\beta}_1^* \times Nsmoke + \hat{\beta}_2^* \times Age + \hat{\beta}_3^* \times (Age \times Nsmoke)$

$\hat{Y}_{smk} = \hat{\beta}_0^* + \hat{\beta}_2^* \times Age$

- H_0: No association between age and FEV for smokers $(H_0: \beta_2^* = 0)$

$\hat{Y}_{nonsmk} = (\hat{\beta}_0^* + \hat{\beta}_1^*) + (\hat{\beta}_2^* + \hat{\beta}_3^*) \times Age$

- H_0: No association between age and FEV for non-smokers $(H_0: \hat{\beta}_2^* + \hat{\beta}_3^* = 0)$

β_0^* Average FEV for **smokers**, age 0

β_1^* Difference in FEV at age 0 between non-smokers and smokers (i.e., difference in intercepts)

β_2^* Slope for **<u>smokers</u>** (increase in FEV per year of age for non-smokers).

β_3^* <u>Difference</u> in slope between non-smokers and smokers.

FEV Example: with Interaction (Smoking reference group recoded)

```
PROC REG DATA=fev;
    MODEL fev = age nsmoke agensmk / CLB;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	291.64807	97.21602	317.11	<.0001
Error	650	199.27177	0.30657		
Corrected Total	653	490.91984			

Changing the coding does not alter the overall fit of the model

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	1	2.19697	0.40596	5.41	<.0001	1.39982	2.99411
age	Age (years)	**1**	**0.07986**	**0.02959**	**2.70**	**0.0071**	**0.02176**	**0.13795**
nsmoke	Non Smoker	1	-1.94357	0.41428	-4.69	<.0001	-2.75707	-1.13007
agensmk	Age x Non-smoke	1	0.16270	0.03074	5.29	<.0001	0.10235	0.22306

β_{age} is now the slope for smokers!

We also have the 95% CI for slope for smokers

Use PROC GLM and ESTIMATE statements for the parameter of interest

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	291.64807	97.21602	317.11	<.0001
Error	650	199.27177	0.30657		
Corrected Total	653	490.91984			

R-Square	Coeff Var	Root MSE	fev Mean
0.594085	20.99870	0.553689	2.636780

Source	DF	Type I SS	Mean Square	F Value	Pr > F
csmoke	1	29.5696825	29.5696825	96.45	<.0001
age	1	253.4885681	253.4885681	826.85	<.0001
agesmk	1	8.5898163	8.5898163	28.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
csmoke	1	6.7473849	6.7473849	22.01	<.0001
age	1	259.8450824	259.8450824	847.58	<.0001
agesmk	1	8.5898163	8.5898163	28.02	<.0001

Output from ESTIMATE statement

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

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	0.253395524	0.08265075	3.07	0.0023	0.091100822	0.415690226
csmoke	1.943570739	0.41428463	4.69	<.0001	1.130073024	2.757068453
age	0.242558411	0.00833154	29.11	<.0001	0.226198429	0.258918393
agesmk	-0.162702674	0.03073753	-5.29	<.0001	-0.223059512	-0.102345836

The relationship between FEV and age differs significantly for non-smokers compared to smokers ($p<0.0001$). 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, $p<0.0001$).

For non-smokers, FEV increases an average of 0.24 liters per year (95% CI: 0.23 to 0.26 L/yr, $p<0.0001$).

For smokers, FEV only increases an average of 0.08 liters per year (95% CI: 0.02 to 0.14 L, $p=0.0071$).

D. Maximum Likelihood

The term maximum likelihood (ML) refers to a very general algorithm for obtaining estimators of population parameters.

ML estimators have large-sample statistical properties that are very useful for practical applications of regression modeling.

- The ML method is applicable to a wide variety of statistical models (linear, generalized linear, and non-linear models).

- ML estimators of regression coefficients are approximately normally distributed when computed from large samples (convenient for making statistical inferences).

Maximum Likelihood Estimates (MLE)

- Maximizes the likelihood function

Least Squares Estimates (LSE)

- Minimizes the sum of squares error

Maximum Likelihood Estimation for Linear Regression Parameters

Consider the multiple linear regression model (Model 1) fit using a random sample of *n* individuals:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

The observed data for the *i*th subject is given by $(Y_i, X_{i1}, \ldots, X_{ik})$, $i = 1, 2, \ldots, n$.

We will assume that the Y_i are normally distributed with variance $\text{Var}(Y_i) = \sigma^2$ not varying with *i* and that the ***X*** are measured without error.

We must also assume that the n random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.

> This allows the precise description of the joint distribution of the variables (i.e., the likelihood function) solely on the basis of knowledge of the separate behavior (i.e., the so-called marginal distribution) of each variable in the set (the product of the marginal distributions).

Ultimately we wish to estimate $\theta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$.

Maximum Likelihood Estimation for Linear Regression Parameters (cont.)

The expression for the distribution (density function) of the normally distributed random variable Y_i under Model 1 is given by:

$$f(Y_i;\beta_0,\beta_1,..,\beta_k,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(-\frac{1}{2\sigma^2}\left(Y_i-(\beta_0+\beta_1X_{i1},+..,\beta_kX_{ik})\right)^2\right)$$

The joint probability of the data (the likelihood of the data) is given by

$$\begin{aligned}L(\boldsymbol{Y};\beta_0,\beta_1,..,\beta_k,\sigma^2) &= \prod_{i=1}^{n} f(Y_i;\beta_0,\beta_1,..,\beta_k,\sigma^2)\\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-(\beta_0+\beta_1X_{i1},+..,\beta_kX_{ik})\right)^2\right)\end{aligned}$$

This is the probability that $Y_1=y_1$, $Y_2=y_2$,…, $Y_n=y_n$.

Maximum Likelihood Estimation for Linear Regression Parameters (cont.)

Note: log(abc) = log(a) + log(b) + log(c)
Note: $\log(z^a) = a*\log(z)$

The log likelihood is given by

$$log\ L\ (\boldsymbol{Y}; \beta_0, \beta_1, .., \beta_k, \sigma^2) = log\left(\frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_{i1}, +.., \beta_k X_{ik})\right)^2\right)\right)$$

$$= \frac{-n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_{i1}, +.., \beta_k X_{ik})\right)^2$$

$$= \frac{-n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$$

Then, $\frac{\partial LL(\boldsymbol{Y};\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} \propto \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) = \frac{\partial SS_{Error}}{\partial \boldsymbol{\beta}}$

The MLE for the βs are identical to the least squares estimators (LSE)!

Maximum Likelihood Estimation for Linear Regression Parameters (cont.)

However, the ML estimator $\hat{\sigma}^2$ of σ^2 is actually a biased estimate of σ^2:

$$log\, L\,(\boldsymbol{Y}; \beta_0, \beta_1, .., \beta_k, \sigma^2) = \frac{-n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$$

$$\frac{\partial LL(\boldsymbol{Y}; \boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2(\sigma^2)^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta}) = 0$$

$$\Rightarrow \frac{n}{2}\left(\frac{1}{\hat{\sigma}^2{}_{MLE}}\right) = \frac{1}{2(\hat{\sigma}^2{}_{MLE})^2}(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})$$

$$\Rightarrow n\hat{\sigma}^2{}_{MLE} = (\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})$$

$$\Rightarrow \hat{\sigma}^2{}_{MLE} = \frac{(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X\hat{\beta}})}{n} = \frac{SS_{Error}}{n} = \left(\frac{n-p-1}{n}\right)\left(\frac{SS_{Error}}{n-p-1}\right)$$

$$= \left(\frac{n-p-1}{n}\right)\hat{\sigma}^2{}_{Y|X} = \left(\frac{n-p-1}{n}\right) MSE$$

$$\Rightarrow E[\hat{\sigma}^2{}_{MLE}] = \left(\frac{n-p-1}{n}\right)\sigma^2{}_{Y|X} \neq \sigma^2{}_{Y|X}$$

General Principle of Maximum Likelihood Estimation

(1) Construct the Likelihood Function:

- Write down an expression for the probability of the data as a function of the *q* unknown parameters, $\boldsymbol{\vartheta}$.

(2) Maximization:

- Find the values of the unknown parameters that make the value of this expression as large as possible.

- That is, the ML estimator of $\boldsymbol{\vartheta}$, is the set of estimators $\hat{\theta}$ for which L(Y; $\hat{\theta}$) > L(Y; $\boldsymbol{\vartheta^*}$) where $\boldsymbol{\vartheta^*}$ denotes any other set of estimators of the elements of $\boldsymbol{\vartheta}$.

- For some special cases, the MLE can be found algebraically.

- Typically, finding the MLE requires an iterative numerical solution to solve the system of *q* equations obtained by setting the partial derivative of LL with respect to each θ_j to 0.
  - The Newton-Raphson algorithm is one such technique.
    https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization

Another important function that can be derived from the likelihood is the **Fisher information** about the unknown parameter(s). The Fisher information function is the negative of the curvature in *LL* = *logL,* i.e.

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} log\, L\,(Y;\theta)|\theta\right]$$

Hypothesis Testing

(1) Wald (Chi-Square) Test Statistic (reported as Z or $Z^2=\chi^2$)

- The Wald Chi-Square Statistic is calculated as the ratio between a coefficient and its standard error, squared, where the standard error is calculated by the inverse of the information matrix.
- This is the p-value reported in many parameter estimate tables.

(2) Likelihood Ratio Test / Likelihood Ratio Statistic

- The Likelihood Ratio Test Statistic is based on the change in value of the log-likelihood (Log L) between two nested models (a full model with and reduced model without the parameter(s) of interest).

-2LogL(reduced) – -2LogL(full) $\sim \chi^2_{p(full)-p(reduced)}$

p is the number of parameters

You can use AIC or SC/BIC for non-nested models.

- This test is similar to the multiple partial *F* test.
- The Likelihood Ratio Test should be used to test the significance of *K*-level categorical variables.

Hypothesis Testing cont.

(3) Score Test / Score Statistic (not covered in BIOS 6612)

- Equivalent to χ^2 test for 2×2 table and the standard z-test for the difference between two proportions.

Note: For large n, all three tests are (asymptotically) equivalent.

E. An Example of MLE in Linear Regression

(KKNR, p. 666) A laboratory study is undertaken to determine the relationship between the dosage (X) of a certain drug and weight gain (Y) after two weeks. 8 laboratory animals of the same sex, age, and size are selected and 1 animal is randomly assigned to each dose.

KKNR Table 21.2

Dosage Level (X)	1	2	3	4	5	6	7	8
Weight gain (Y)	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Consider the following three models:

Model 0: $E(Y) = \beta_0$

Model 1: $E(Y|X) = \beta_0 + \beta_1 X$

Model 2: $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$

Since n=8 (small), $\hat{\sigma}^2{}_{Y|X}(MLE) = \left(\frac{n-p-1}{n}\right)\hat{\sigma}^2{}_{Y|X}(LSE) \neq \hat{\sigma}^2{}_{Y|X}(LSE)$

```
DATA wtgain;
INPUT wgtgain dose dosesq;
DATALINES;
1   1.0 1.0
1.2 2.0 4.0
1.8 3.0 9.0
2.5 4.0 16.0
3.6 5.0 25.0
4.7 6.0 36.0
6.6 7.0 49.0
9.1 8.0 64.0
;

*-------------------------------------------------------------------
*-------------------------------------------------------------------
Fit models with no predictor, just dose, and dose+dose^2
Estimate the LSE (least squares equation) estimates;

PROC REG DATA=wtgain;
    MODEL wgtgain =;
    MODEL wgtgain = dose;
    MODEL wgtgain = dose dosesq;
RUN;
```

Least Squares Estimation (LSE)

Model 0: $E(Y) = \beta_0$, LSE

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	0	0	.	.	.
Error	7	57.06875	8.15268		
Corrected Total	7	57.06875			

$\hat{\sigma}^2_Y(LSE)$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.81250	1.00950	3.78	0.0069

Model 1: $E(Y|X) = \beta_0 + \beta_1 X$, LSE

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	52.03720	52.03720	62.05	0.0002
Error	6	5.03155	0.83859		
Corrected Total	7	57.06875			

$\hat{\sigma}^2_{Y|X}(LSE)$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-1.19643	0.71354	-1.68	0.1446
dose	1	1.11310	0.14130	7.88	0.0002

Model 2: $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$, LSE

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	56.87202	28.43601	722.73	<.0001
Error	5	0.19673	0.03935		
Corrected Total	7	57.06875			

$\hat{\sigma}^2_{Y|X}(LSE)$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	1.34821	0.27674	4.87	0.0046
dose	1	-0.41369	0.14109	-2.93	0.0326
dosesq	1	0.16964	0.01530	11.09	0.0001

Partial F Model 2 versus 0?

$H_0: \beta_{dose} = \beta_{dose^2} = 0 \Rightarrow$

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} = \frac{(56.87202 - 0)/2}{0.03935} = 722.73 \sim F_{k,n-p-k-1} = F_{2,8-0-2-1} = F_{2,5}$$

Partial F Model 2 versus 1?

$H_0: \beta_{dose^2} = 0 \Rightarrow$

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} = \frac{(56.87202 - 52.03720)/1}{0.03935} = 122.867 \sim F_{k,n-p-k-1} = F_{1,8-1-1-1} = F_{1,5}$$

Maximum Likelihood (ML) Estimation

Using PROC MIXED we can specify that the MLE estimators are used. Note, in 6612 you will use more of the capabilities of PROC MIXED to fit mixed effect models.

```
*-------------------------------------------------------------------
*-------------------------------------------------------------------
Fit mixed model for wtgain;

PROC MIXED DATA=wtgain METHOD=ML;
    MODEL wgtgain = /SOLUTION;
RUN;

PROC MIXED DATA=wtgain METHOD=ML;
    MODEL wgtgain = dose /SOLUTION;
RUN;

PROC MIXED DATA=wtgain METHOD=ML;
    MODEL wgtgain = dose dosesq /SOLUTION;
RUN;
```

Maximum Likelihood Estimation (MLE)

Model 0: $E(Y) = \beta_0$, MLE

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	7.1336

$\hat{\sigma}^2_Y(MLE)$

Fit Statistics	
-2 Log Likelihood	38.4
AIC (Smaller is Better)	42.4
AICC (Smaller is Better)	44.8
BIC (Smaller is Better)	42.6

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	3.8125	0.9443	7	4.04	0.0049

Recall that $\hat{\sigma}^2_Y(MLE)$ is a biased estimator of σ^2_Y.

$$\hat{\sigma}^2{}_{Y|X}(MLE) = \left(\frac{n-p-1}{n}\right)\hat{\sigma}^2{}_{Y|X}(LSE) \Rightarrow \hat{\sigma}^2{}_{Y|X}(LSE) = \left(\frac{n}{n-p-1}\right)\hat{\sigma}^2{}_{Y|X}(MLE)$$

If we multiply by n/(n-1), we have an unbiased estimator: 7.1336 * (8/7) = 8.15268 = $\hat{\sigma}^2_Y(LSE)$

Model 1: $E(Y|X) = \beta_0 + \beta_1 X$, MLE

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.6289

Fit Statistics	
-2 Log Likelihood	19.0 (18.9916)
AIC (Smaller is Better)	25.0
AICC (Smaller is Better)	31.0
BIC (Smaller is Better)	25.2

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-1.1964	0.6179	6	-1.94	0.1010
dose	1.1131	0.1224	6	9.10	<.0001

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
dose	1	6	82.74	<.0001

Likelihood ratio test of Model 1 versus Model 0 (H_0: $\beta_{dose} = 0$):

38.442 - 18.9916 = 19.43 $\sim \chi^2_1$ and p<0.001.

Model 2: $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$, MLE

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.02459

Fit Statistics	
-2 Log Likelihood	-6.9
AIC (Smaller is Better)	1.1
AICC (Smaller is Better)	14.4
BIC (Smaller is Better)	1.4

Solution for Fixed Effects					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	1.3482	0.2188	5	6.16	0.0016
dose	-0.4137	0.1115	5	-3.71	0.0139
dosesq	0.1696	0.01210	5	14.02	<.0001

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
dose	1	5	13.76	0.0139
dosesq	1	5	196.61	<.0001

Likelihood ratio test of Model 2 versus Model 0 (H_0: $\beta_{dose} = \beta_{dose^2} = 0$):

38.442 – (-6.9) = 43.342 $\sim \chi^2_2$ and p<0.001.

Conclusions

- The maximum likelihood estimators of the linear regression coefficients are identical to the least squares estimators.

- The maximum likelihood estimators of variance (and corresponding standard errors) are biased while the least squares estimators provide unbiased variance estimators when all assumptions are satisfied.

- For small n, inference-making conclusions may differ depending on whether least squares or maximum likelihood estimation are used.

- Least squares estimation (PROC REG; PROC GLM; or PROC MIXED with (METHOD=)REML estimation) are preferred to maximum likelihood estimation (PROC MIXED with ML estimation or PROC GENMOD), when fitting multiple linear regression models *assuming normally distributed and mutually independent responses with homogeneous variance*.

- For large *n*, however, the bias in the variance estimators from maximum likelihood estimation will typically be small, so that inference-making conclusions using least square and maximum likelihood estimation approaches typically will not differ very much (they are asymptotically equivalent).