21. Confounding and Mediation in Linear Regression

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Ch. 10-11

SAS: PROC REG, PROC MIXED

Homework: Homework 9 due by midnight on November 26

Overview

A) Re/Preview of Topics
B) Confounding and Spurious Correlations
C) Mediation

A. Review (Lecture 20)/ Current (Lecture 21)/ Preview (Lecture 22)

Lecture 20:

- Linear Algebra Review (Bonus HW)
- Least Squares Estimation (LSE) using matrices
- Collinearity
  - Columns of X are linearly dependent
    - i.e. BMI, height, weight

Lecture 21:

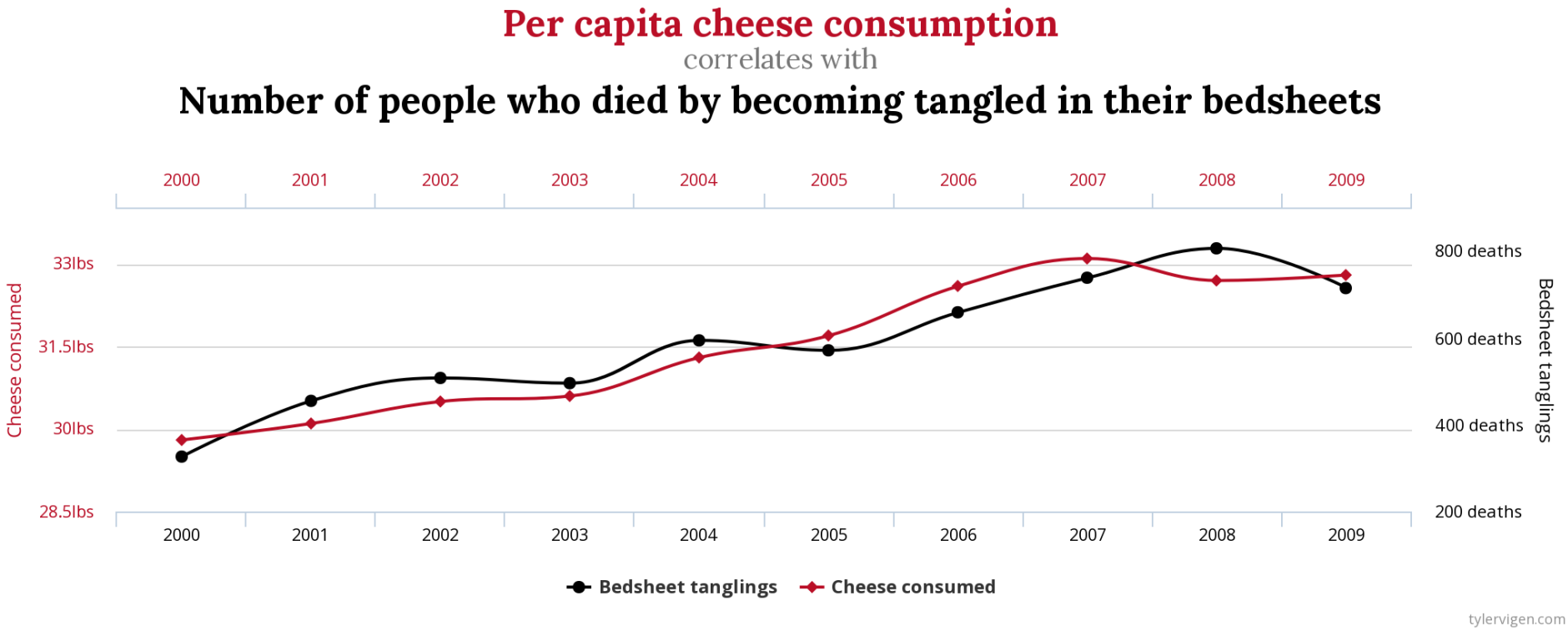
- Confounders (X<-C->Y)
- Mediators (X->C->Y)

Lecture 22:

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

B. Confounding and Spurious Correlations

Spurious correlations involving time (Lecture 19 slide 13): When both variables have time trends, there could be a potentially spurious correlation. From http://www.tylervigen.com/spurious-correlations:



Spurious Correlations and Confounding

FEV is positively correlated with smoking (Lecture 19, slide 18)

Smoking status is associated with FEV (Lecture 19, slides 19-21)

- Smokers have **better** lung function compared to non-smokers
- Smokers tend to be older than non-smokers

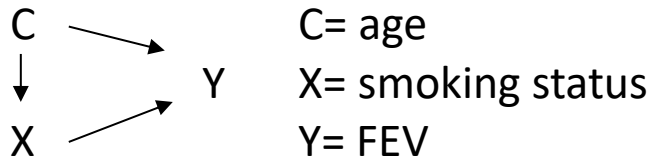
Age is associated with FEV (Lectures 16-17, slides 6 and 30)

- Older children have higher FEV than younger children

Model with age and smoking status (Lecture 19, slides 20-22)

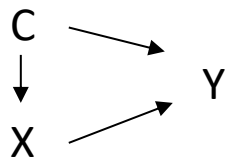
- Smokers have **worse** lung function compared to non-smokers of the same age

Age is a confounder (age 3-19)



C= age
X= smoking status
Y= FEV

Confounding



A very common use of multiple regression models in the health sciences is to adjust an association for the effects of confounding variables.

Confounding is the distortion of an estimated association due to the effect(s) of other variable(s).

The ***Confounder*** is the variable that causes the distortion (C in our plot above).

The ***crude*** (or ***unadjusted***) ***estimate*** is an estimate of the association of primary interest when the potential confounder(s) are ignored (i.e. the potential confounder is NOT included in the model).

The ***adjusted estimate*** is an estimate of the association of primary interest that accounts in some way for the potential confounder(s).

Adjusting for Confounding

During study design:

- Matching: Match cases and controls by known confounders (usually one or two matching factors).

- Restriction: Restrict the study eligibility criteria to include only individuals in specified categories of a confounder (limits generalizability).

- Randomization: Randomly assign individuals to treatment groups. On average this gives groups that are balanced on both measured and unmeasured confounders.

During analysis:

- Stratification: Stratify the analysis by levels of the confounder (e.g., association between smoking and FEV by age). This reduces sample size since analyses are completed on the stratified subgroups instead of using all data.

- Regression models: Adjust for confounders in a statistical model.

Classical Criteria for Confounding

1. A confounding factor must be associated with the exposure (or primary explanatory variable (PEV)) under study.

2. A confounding factor must be a risk factor or a surrogate for a risk factor for the disease.

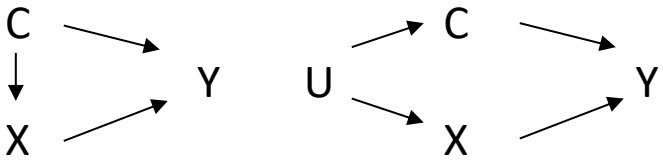
3. A confounding factor must not be affected by the exposure or the disease.

   -It cannot be an intermediate step in the causal path between the exposure and the disease

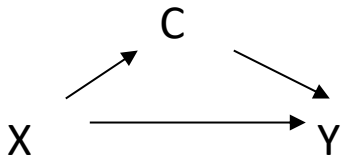
   -This type of variable is sometimes called a ***mediator*** or ***intervening variable***

   -These variables can be used as "surrogate markers" for a clinically meaningful outcome in clinical trials

Confounding:



Mediation:



Operational Criterion for Confounding

The operational criterion for confounding:

> Covariate is a confounding factor if the crude parameter estimate is not equal to the adjusted, or stratum-specific, parameter estimate: $\beta_{crude} \neq \beta_{adj}$

"Operational" method to assess confounding

Confounding is present if there is any meaningful difference between the crude and adjusted estimates.

- Clinically meaningful or relevant change based on clinical judgment

- If clinically meaningful change is uncertain. What change is meaningful? 10% change? 20% change? Two approaches to summarizing % change (depending on whether the crude or adjusted estimate is believed to be the "starting value"):

  $\frac{\beta_{crude}-\beta_{adj}}{\beta_{crude}} \times 100$ (favored by biostatisticians) or $\frac{\beta_{crude}-\beta_{adj}}{\beta_{adj}} \times 100$ (favored by epidemiologists), while the answers will differ slightly, they generally produce similar results.

Crude and Adjusted Estimates

Let X be the primary explanatory variable (PEV), Z be a covariate/potential confounder, and Y be the outcome/response. Then we can define three models of interest:

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

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

3. Covariate Model: $\hat{Z} = \hat{\gamma}_{0} + \hat{\gamma}_{X}X$

The difference between the crude and adjusted estimates ($\hat{\beta}_{crude} - \hat{\beta}_{adj}$) can be calculated as the product between the *Z-X* association ($\hat{\gamma}_X$) and the *Y-Z* association ($\hat{\beta}_z$):

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

operational = classical

Classical Criterion #1 (X and Z association): $\hat{\gamma}_X$
Classical Criterion #2 (Z and Y associated given X): $\hat{\beta}_z$

Crude and Adjusted Estimates cont.

KKNR claim that "a statistical test is neither required nor appropriate" to assess confounding.

However, statistical tests of mediation *are* considered appropriate, particularly in social sciences.

Statistical tests of confounding may also be appropriate if the primary scientific question is to identify "confounders" in the population (rather than to adjust for confounding in a given data set). However, this is not often the case.

One estimator for the standard error of $\hat{\beta}_{crude} - \hat{\beta}_{adj}$ can be found using the delta method as:

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

Positive Confounding

Definition 1: A variable that is positively associated with both exposure and disease or negatively associated with both exposure and disease is called a ***positive confounder***.

Definition 2: Positive confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is **more extreme** – either more positive or more negative – than the true association.

- A positive confounder can create spurious associations.

Negative confounding

Definition 1: A variable that is positively associated with the exposure and negatively associated with the disease (or vice versa) is called a ***negative confounder***.

Definition 2: Negative confounding refers to the situation in which the effect of the confounding factor is to produce an observed estimate of the association between exposure and disease that is an **underestimate** of the true association.

-That is, a negative confounder can mask a true association.

Effect of covariate adjustment in linear regression

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

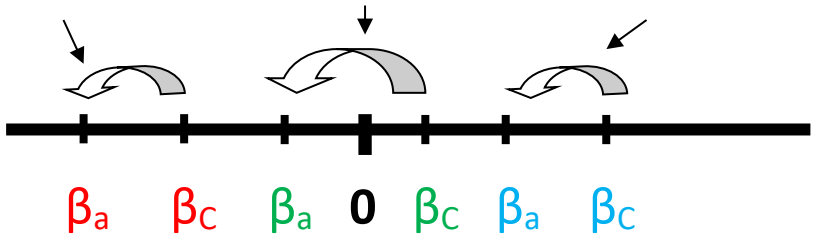
$\hat{\gamma}_X$ *(Exposure-Covariate Association)*	$\hat{\beta}_z$ *(Outcome-Covariate Association, given X)*	Confounding Effect
+	+	$\hat{\beta}_{crude} > \hat{\beta}_{adj}$
-	-	$\hat{\beta}_{crude} > \hat{\beta}_{adj}$
+	-	$\hat{\beta}_{crude} < \hat{\beta}_{adj}$
-	+	$\hat{\beta}_{crude} < \hat{\beta}_{adj}$
0	+, -, or 0	$\hat{\beta}_{crude} = \hat{\beta}_{adj}$
+, -, or 0	0	$\hat{\beta}_{crude} = \hat{\beta}_{adj}$

Below is a representation of three cases of a "positive confounder" using definition #1, but they differ in terms of definition #2. Note in all three cases the adjusted estimate moves to the left:

Adjusted moves away from null

Adjusted crosses null

Adjusted moves closer to null



NOTE: Simpson's Paradox refers to confounding where the point estimate crosses zero after adjusting for the confounding variable.

Confounding in matrix notation

In Lecture 20, we noted that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ if $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] \\ &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] \\ &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})\right] \\ &= E\left[\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})\right] \\ &= \boldsymbol{\beta} \end{aligned}$$

If $E(\boldsymbol{\varepsilon}) \neq 0$, then $\widehat{\boldsymbol{\beta}}$ is a biased estimator of $\boldsymbol{\beta}$. Suppose that $E(\boldsymbol{\varepsilon}) = \boldsymbol{Z\gamma}$ for some $n \times q$ matrix $\boldsymbol{Z}$ of variables different from those in $\boldsymbol{X}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ unknown parameter vector. Then

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma}$$

"Alias" Matrix

and the bias in $\widehat{\boldsymbol{\beta}}$ as an estimate of $\boldsymbol{\beta}$ is given by:

$$bias = \boldsymbol{\beta} - \boldsymbol{E}(\widehat{\boldsymbol{\beta}}) = -(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma}$$

or

$$\widehat{\boldsymbol{\beta}}_{crude} - \widehat{\boldsymbol{\beta}}_{adj} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\gamma}$$

Thus, the bias will be zero (and $\widehat{\boldsymbol{\beta}}_{crude} = \widehat{\boldsymbol{\beta}}_{adj}$) if:

(1) $\boldsymbol{X}'\boldsymbol{Z} = 0$ ($\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent) *or* (2) $\boldsymbol{\gamma} = 0$ ($\boldsymbol{Z}$ is not predictive of Y given X)

Precision

The term ***precision*** refers to the size of an estimator's variance, or equivalently, the narrowness of a confidence interval for the parameter being estimated.

The smaller the variance of the estimator, the higher the precision of the estimator.

$$\beta_z = \rho_{YZ|X}\, sd(y|x)/sd(z|x)$$

$$\frac{Var(\hat{\beta}_{adj})}{Var(\hat{\beta}_{crude})} = \frac{1-\hat{\rho}^2_{YZ|X}}{1-\hat{\rho}^2_{XZ}}\left(\frac{n-2}{n-3}\right)$$

$$\gamma_x = \rho_{XZ}\, sd(z)/sd(x)$$

where $\hat{\rho}^2_{YZ|X}$ is the ***partial correlation*** between *Y* and *Z* that controls for *X* (i.e., the effect of *X* is accounted for in the correlation), and $\hat{\rho}^2_{XZ}$ is the correlation between *X* and *Z*.

- A strong association between *Y* and *Z* has a *beneficial* effect upon the precision of $\hat{\beta}_{adj}$, i.e. it *decreases* the s.e. of $\hat{\beta}_{adj}$.
- A strong association between *X* and *Z* has a *detrimental* effect on the precision of $\hat{\beta}_{adj}$, i.e. it *increases* the s.e. of $\hat{\beta}_{adj}$. (*multicollinearity*)

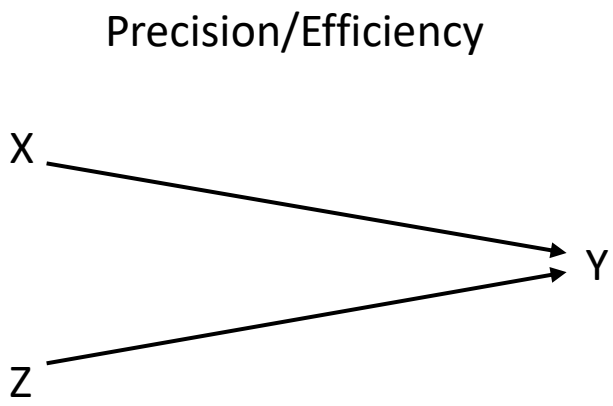
Thus, the precision of $\hat{\beta}_{adj}$ reflects the competing effects of the *Y*-*Z* and *X*-*Z* relationships. A "precision variable" improves the precision of the estimate of the PEV.

Covariate Adjustment in Linear Regression: Precision/Efficiency Covariates

A precision/efficiency covariate is a variable independent of exposure in the source population ($\gamma_X = 0$), but predictive of the outcome ($\beta_Z \neq 0$).

Precision/efficiency covariates ***cannot*** be confounders according to the classical criteria.

Inclusion of a precision/efficiency variable can provide a more efficient test of the exposure-outcome association and a more precise estimate of the exposure-outcome association.



Example of confounding with FEV data: Crude Model

In the FEV data set we were interested in determining if there was a difference in the lung function in children who smoked and children who did not smoke.

We have seen that smokers have higher FEV, but that smokers are older and FEV increases with age (**positive** confounding by age).

```
PROC REG DATA=fev;
    MODEL fev = csmoke;   /*0=Non-smoker 1=Smoker */
RUN;
```

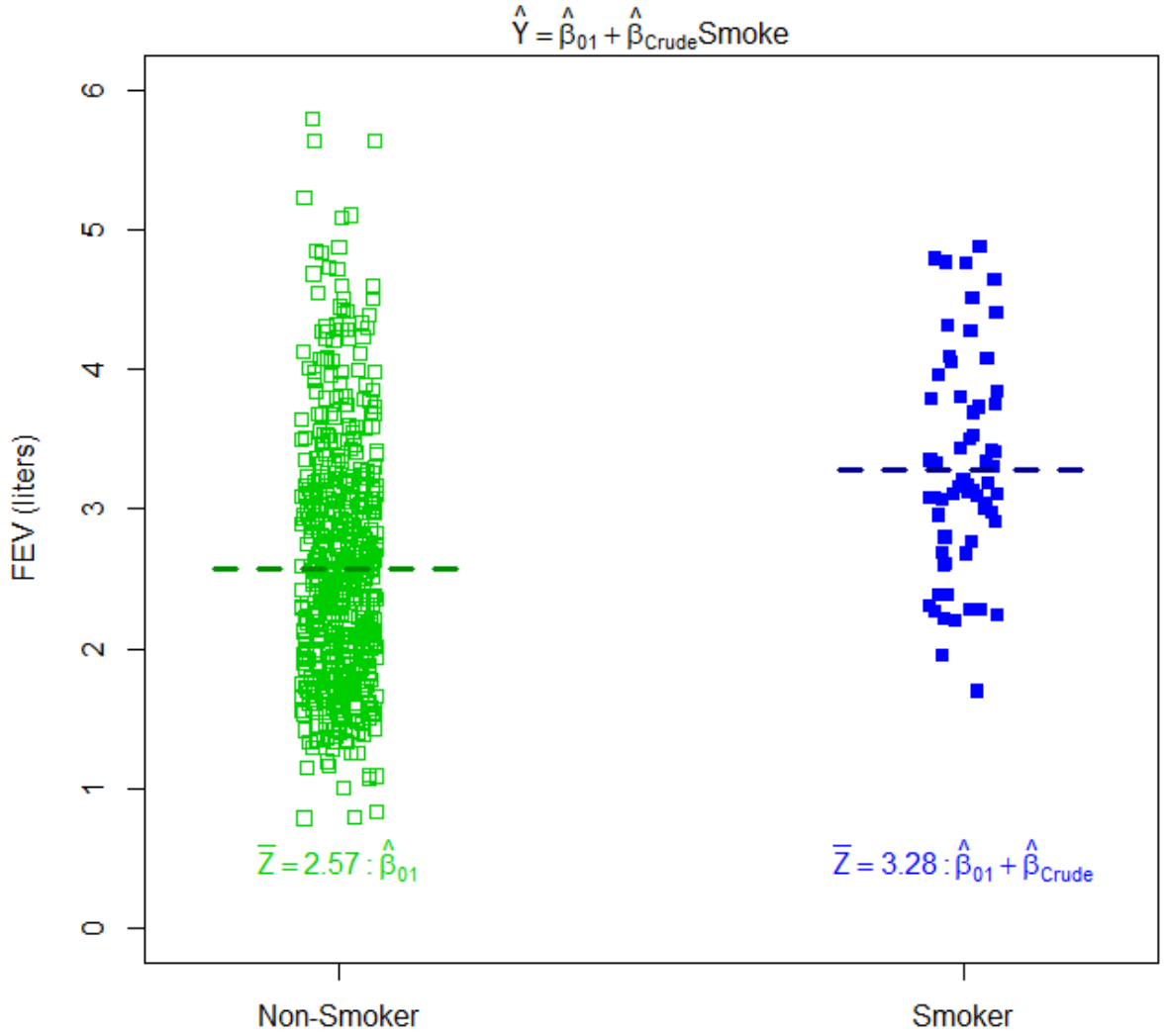
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	29.56968	29.56968	41.79	<.0001
Error	652	461.35015	0.70759		
Corrected Total	653	490.91984			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	2.56614	0.03466	74.04	<.0001
csmoke	1	**0.71072**	0.10994	6.46	<.0001

What is the **crude** estimate for the effect of smoking on FEV?

Crude Model

$$E[Y] = E[FEV] = 2.57 + 0.71 \times smoke$$



Example of confounding with FEV data: Adjusted Model

```
PROC REG DATA=fev;
    MODEL fev = csmoke age;   /*0=Non-smoker 1=Smoker */
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	283.05825	141.52913	443.25	<.0001
Error	651	207.86159	0.31930		
Corrected Total	653	490.91984			

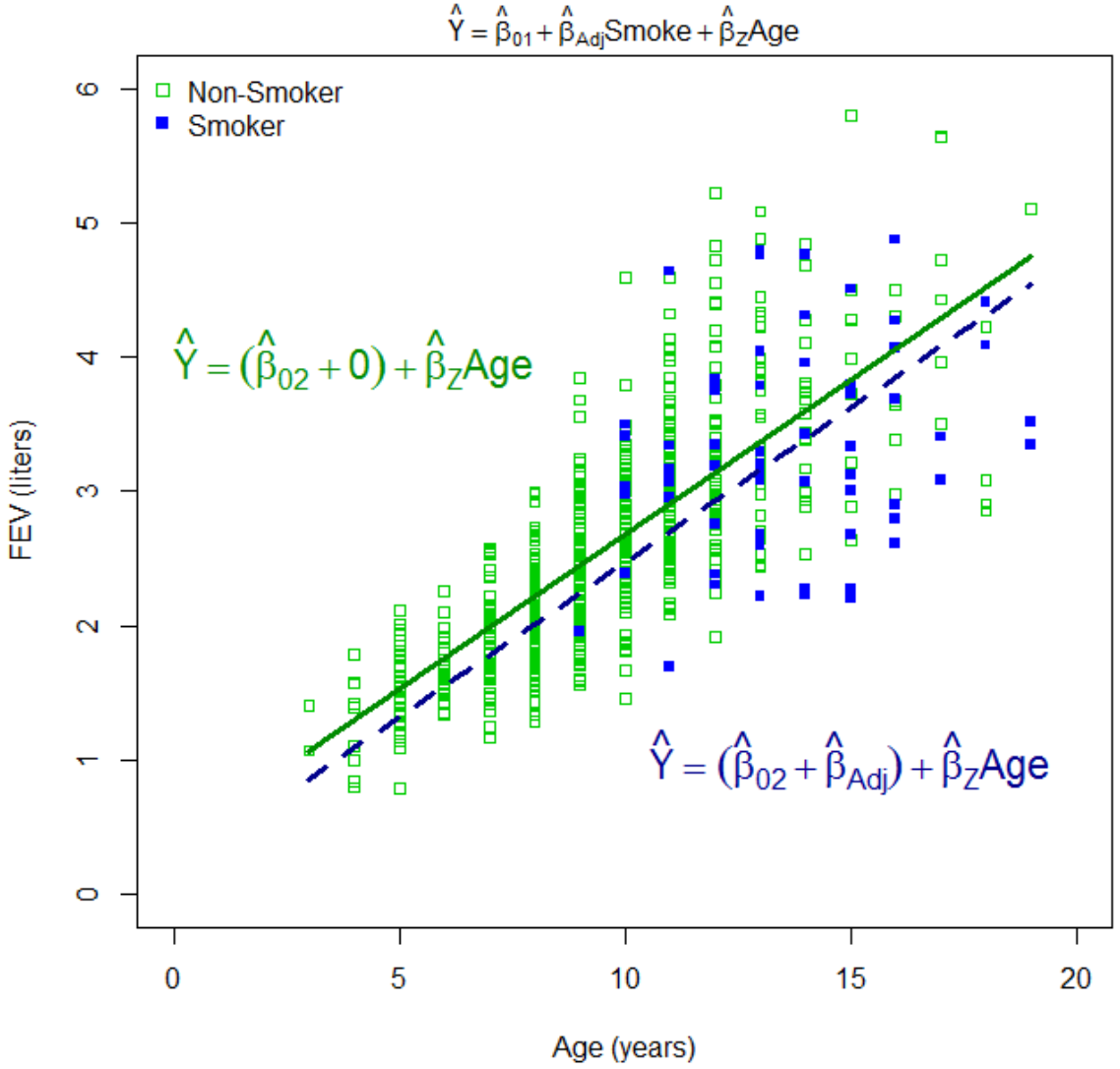
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.36737	0.08144	4.51	<.0001
csmoke	1	**-0.20899**	0.08075	-2.59	0.0099
age	1	0.23060	0.00818	28.18	<.0001

What is the adjusted estimate of the effect of smoking on FEV?

Is age a confounder of the association between smoking and FEV by the operational criterion?

Adjusted Model

$$E[Y] = E[FEV] = 0.37 + (-0.21) \times smoke + 0.23 \times age$$



Example of confounding with FEV data: Covariate Model

```
PROC TTEST DATA=fev;
    CLASS csmoke;  /*0=Non-smoker 1=Smoker */
    VAR age;
RUN;
```

csmoke	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	589	**9.5348**	2.7406	0.1129	3.0000	19.0000
1	65	**13.5231**	2.3393	0.2901	9.0000	19.0000
Diff (1-2)		**-3.9883**	2.7039	0.3534		

csmoke	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		9.5348	9.3130	9.7566	2.7406	2.5926	2.9068
1		13.5231	12.9434	14.1027	2.3393	1.9949	2.8285
Diff (1-2)	**Pooled**	-3.9883	-4.6822	-3.2943	2.7039	2.5648	2.8591
Diff (1-2)	**Satterthwaite**	-3.9883	-4.6074	-3.3692			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	652	-11.29	<.0001
Satterthwaite	Unequal	84.646	-12.81	<.0001

What is the regression model regressing age on smoking status?

Covariate Model: Age = 9.5348 + 3.98827 × Smoke

Is age associated with the primary variable of interest (smoking status)?

Smokers tend to be older (3.99 years older) than non-smokers (13.5 versus 9.5 years).
[**Note**: This association need not be statistically significant for meaningful confounding to occur]

Is age a confounder of the association between smoking and FEV by the classical criteria?

So, age is a confounder (also) by the classical criteria:

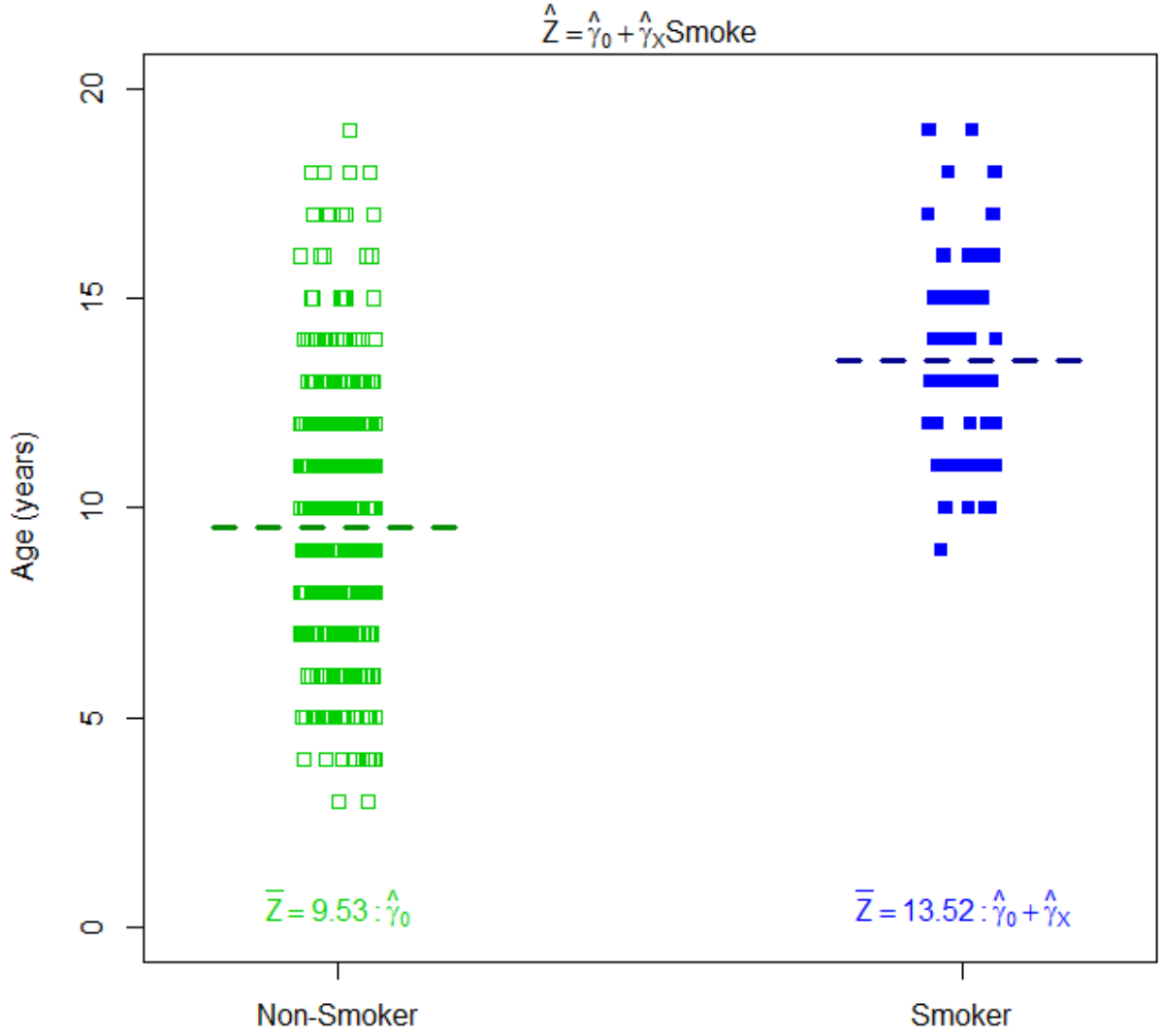
(1) Age is associated with smoking; (comparison of means above)

(2) Age is a predictor of FEV given smoking status; (see slope in regression model with smoking)

(3) Age is not on the causal pathway (smoking doesn't *cause* age). (subject matter consideration)

Covariate Model

$$E[Z] = E[Age] = 9.53 + 3.988 \times smoke$$



Confounding:

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

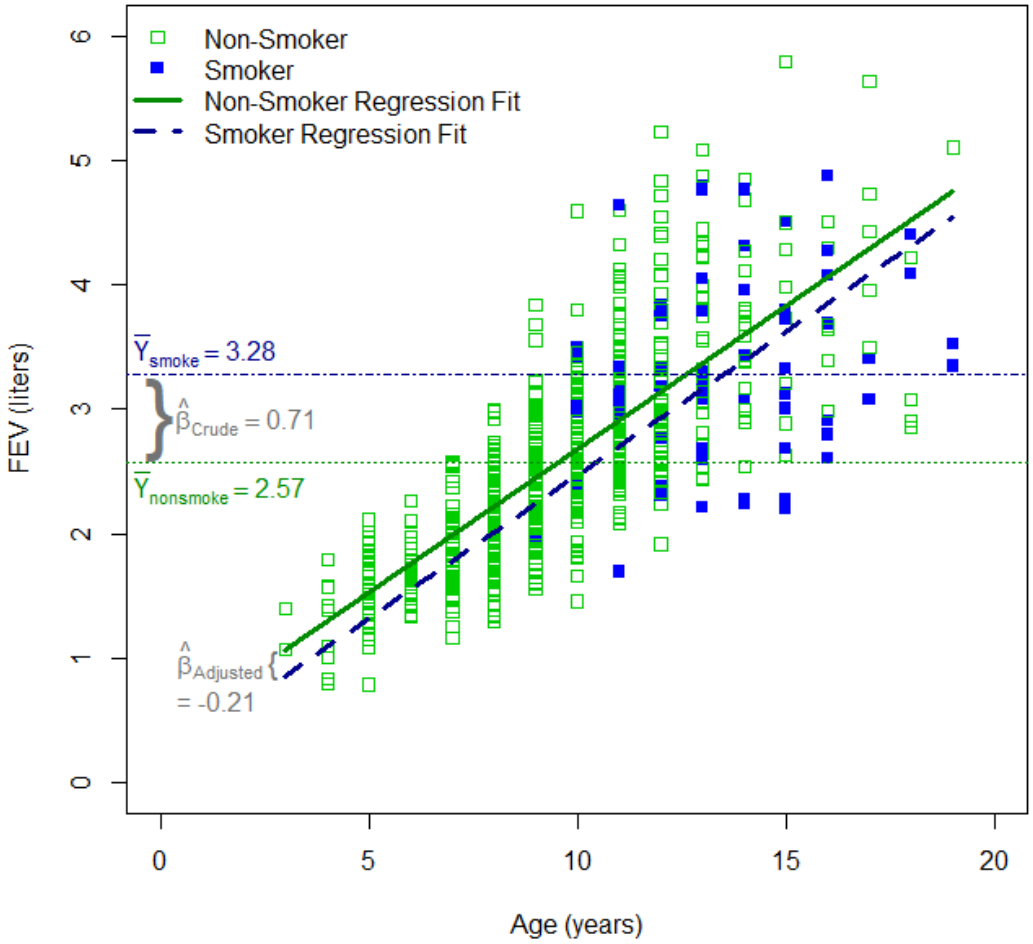
$\hat{\beta}_{crude} - \hat{\beta}_{adj}$ = 0.71072 - (-0.20899) = 0.9197

$\hat{\gamma}_X \times \hat{\beta}_z$ = 3.98827 x 0.23060 = 0.9197

On average, smokers are 3.99 years older than non-smokers ($\hat{\gamma}_X$).

On average, for every one year increase of age, FEV increases by 0.23060L ($\hat{\beta}_z$).

So we'd expect FEV to be 3.99×0.23 = ($\hat{\gamma}_X \times \hat{\beta}_z$) = 0.9197L higher in smokers compared to non-smokers due to age.



Precision:

NOTE: $r_{age,smoke}$ = 0.40425; $r_{age,fev|smoke}$=0.74125 [from SAS PROC CORR, output on next slide]

$$\frac{Var(\hat{\beta}_{adj})}{Var(\hat{\beta}_{crude})} = \frac{(0.08075)^2}{(0.10994)^2} = \frac{(1-0.74125^2)}{(1-0.40425^2)} \times \frac{652}{651} = 0.539$$

```
PROC CORR DATA=fev;
    VAR age csmoke fev;
RUN;
```

```
PROC CORR DATA=fev;
    VAR age fev;
    PARTIAL csmoke;
RUN;
```

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
age	654	9.93119	2.95394	6495	3.00000	19.00000
csmoke	654	0.09939	0.29941	65.00000	0	1.00000
fev	654	2.63678	0.86706	1724	0.79100	5.79300

Pearson Correlation Coefficients, N = 654 Prob > |r| under H0: Rho=0			
	age	**csmoke**	**fev**
age	1.00000	0.40425 <.0001	0.75646 <.0001
csmoke	0.40425 <.0001	1.00000	0.24542 <.0001
fev	0.75646 <.0001	0.24542 <.0001	1.00000

Pearson Partial Correlation Coefficients, N = 654 Prob > |r| under H0: Partial Rho=0		
	age	**fev**
age	1.00000	0.74125 <.0001
fev	0.74125 <.0001	1.00000

C. Mediation

Mediator: A mediator is an intermediate variable on the causal path between an independent variable (e.g., treatment or exposure) and a dependent variable (i.e., outcome).

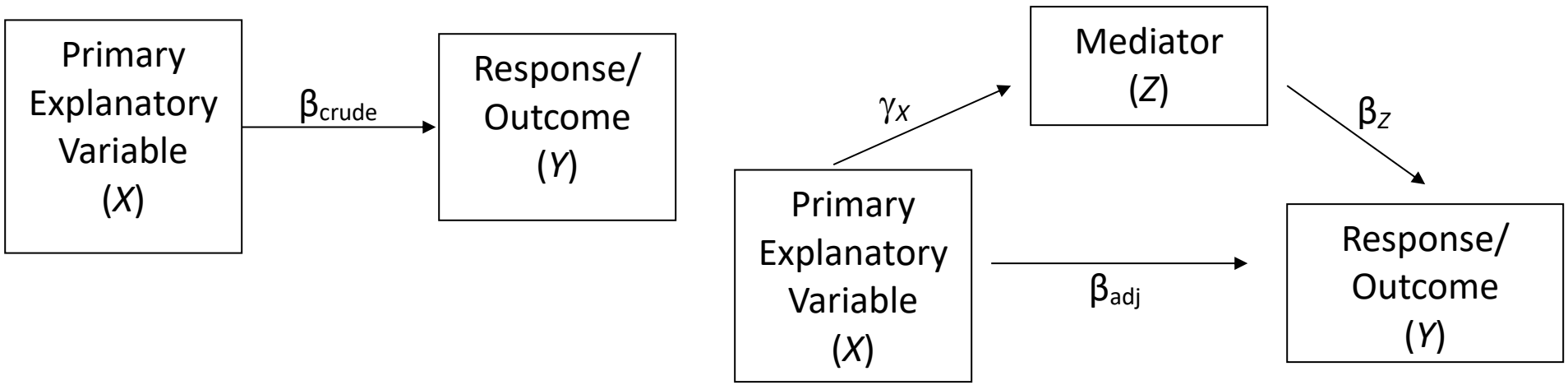
Indirect Effect: Effect of X on Y that works through Z

- Indirect effect: Difference between the crude & adjusted estimates: β_{crude}-$\beta_{\text{adj}} = \gamma_X \times \beta_Z$

Direct Effect: Effect of X on Y that does not work through Z

- It is the 'remaining' effect after adjusting for the mediator (β_{adj}).
- $\beta_{\text{adj}} = \beta_{\text{crude}} - (\gamma_X \times \beta_Z)$

$Y = \beta_{01} + \beta_{crude}X + \epsilon_1$ $Z = \gamma_0 + \gamma_X X + \epsilon_3$ $Y = \beta_{02} + \beta_{adj}X + \beta_Z Z + \epsilon_2$



Fundamental Models for Mediation Analysis (notation consistent with confounding)

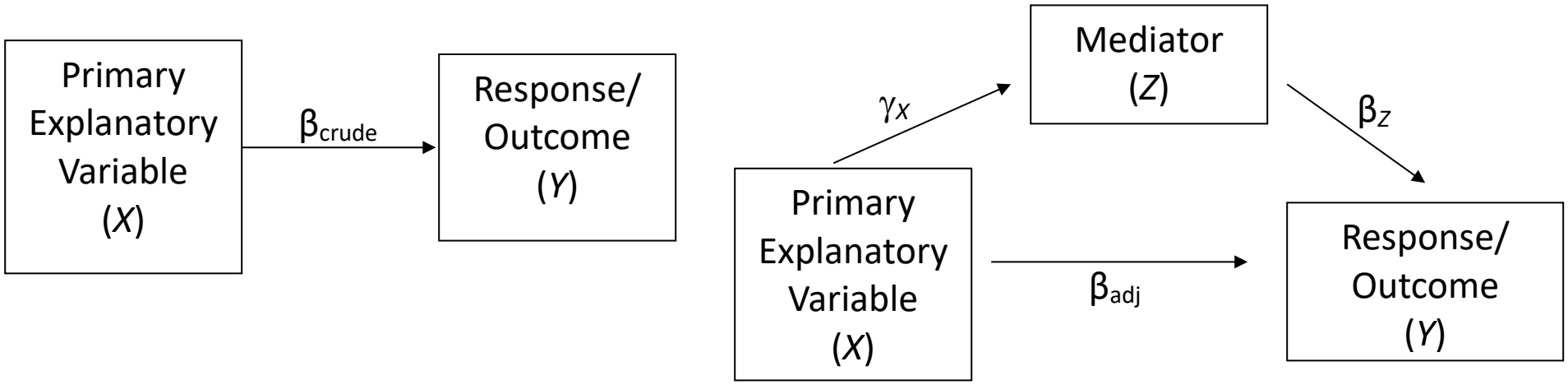
Model 1: Effect of Predictor on Outcome $Y = \beta_{01} + \beta_{crude}X + \epsilon_1$

Model 2: Adjusted effect of the Predictor on Outcome $Y = \beta_{02} + \beta_{adj}X + \beta_Z Z + \epsilon_2$

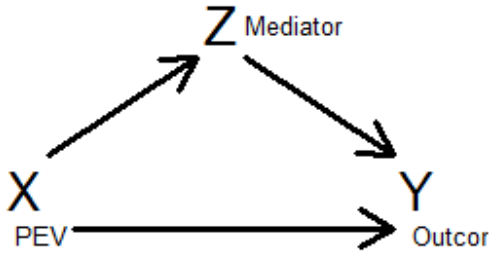
Model 3: Effect of the Predictor on the Mediator $Z = \gamma_0 + \gamma_X X + \epsilon_3$

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

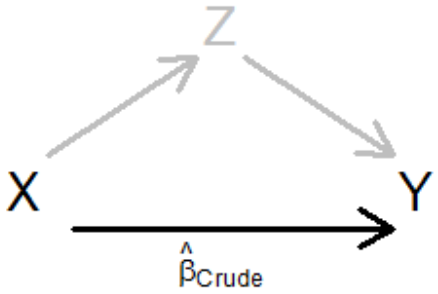
$$SE\left(\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}}\right) = \sqrt{\hat{\beta}_Z^2 Var(\hat{\gamma}_X) + \hat{\gamma}_X^2 Var(\hat{\beta}_Z)}$$



Mediation

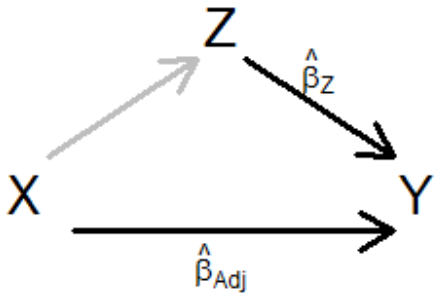


Crude Model



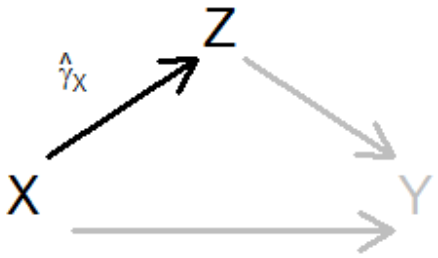
$$\hat{Y} = \hat{\beta}_{01} + \hat{\beta}_{Crude} X$$

Adjusted Model



$$\hat{Y} = \hat{\beta}_{02} + \hat{\beta}_{Adj} X + \hat{\beta}_Z Z$$

Covariate Model



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

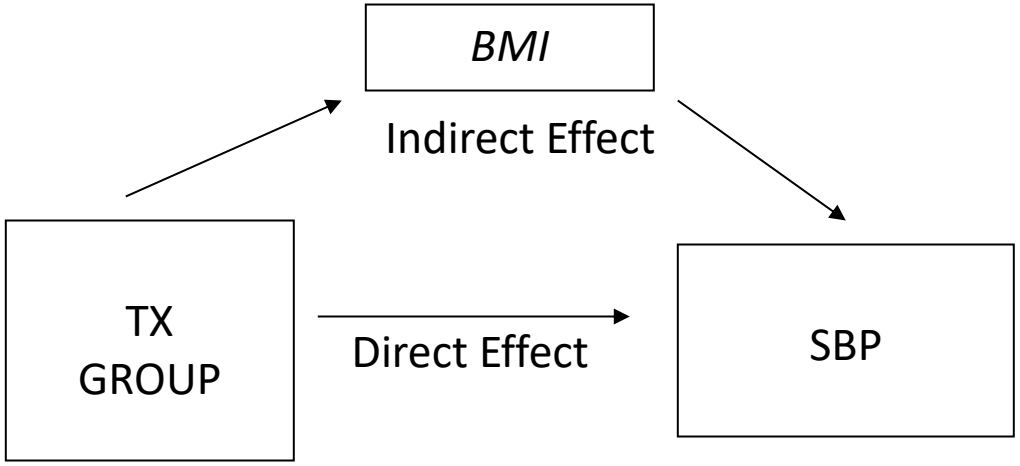
Mediation Example:

A study was performed to examine if a weight loss drug could cause a decrease in systolic blood pressure in adolescents with severe obesity that was already demonstrated in adults.

It was hypothesized that changes in BMI would account for the decrease in blood pressure observed in the subjects on the study drug.

21 adolescents were randomized in total: 9 to the treatment group and 12 to the control group.

The data set includes group (*trtgrp*: 1 = drug, 0 = placebo), change in bmi (*cbmi*), change in systolic blood pressure (*csbp*), and if the participant self-restricted dietary choices to influence shape/weight (*diet_exclude:* 1 = exclude food at least 1 day in the past week, 0 = no exclusion).



Model 1 (Total effect of intervention [i.e. crude model]):

```
proc reg data=top;
    model csbp = trtgrp;
run;
```

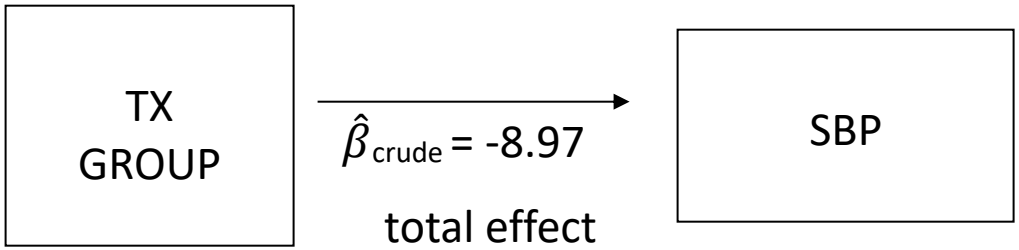
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	414.00397	414.00397	7.11	0.0152
Error	19	1105.80556	58.20029		
Corrected Total	20	1519.80952			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.91667	2.20228	-0.87	0.3950
trtgrp	1	**-8.97222**	3.36403	-2.67	0.0152

Interpretation: There is an 8.97 mmHg greater reduction in systolic blood pressure in the treatment group compared to the control group. This represents a significant difference between the two groups (p = 0.015).

Do you think confounding is likely in this study?

No, since groups were randomized, we'd expect factors to be balanced across groups.



Model 2 (Direct Effect of Intervention [i.e. adjusted model]):

```
proc reg data=top;
    model csbp = trtgrp cbmi;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	456.18881	228.09441	3.86	0.0403
Error	18	1063.62071	59.09004		
Corrected Total	20	1519.80952			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.44570	2.28798	-0.63	0.5354
trtgrp	1	**-4.54517**	6.24040	-0.73	0.4758
cbmi	1	0.80737	0.95554	0.84	0.4092

Interpretation: After adjusting for change in BMI, there is a 4.55 mmHg greater reduction in systolic blood pressure in the treatment group compared to the control group. This difference is not significantly different than zero (p = 0.476).

$$\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}} = -8.972 - (-4.5452) = -4.4268$$

Model 3: Covariate Model

```
proc reg data=top;
    model cbmi = trtgrp;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	154.63000	154.63000	45.40	<.0001
Error	19	64.71667	3.40614		
Corrected Total	20	219.34667			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.58333	0.53277	-1.09	0.2872
trtgrp	1	-5.48333	0.81382	-6.74	<.0001

Calculation of Indirect Effect:

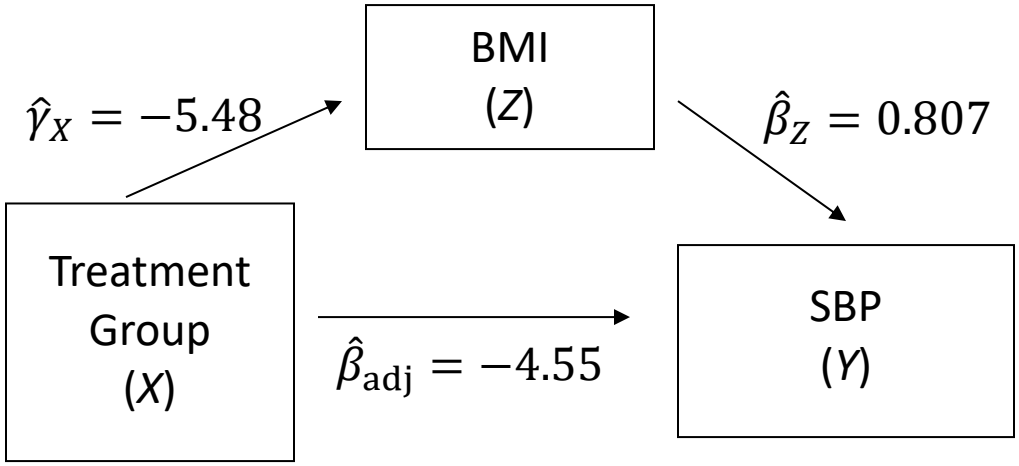
$\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}} = \hat{\gamma}_X \times \hat{\beta}_Z$: -8.972 – (-4.5452) = (-5.4833) x (0.8074) = -4.4268

In our mediation context, the indirect effect is the extent to which the outcome (ΔSBP) changes when the PEV (treatment group) is held fixed and the mediator variable (ΔBMI) changes by the amount it would have changed had the PEV increased by one unit.

Often we are interested in summarizing the proportion of variability explained by mediation…

Proportion Mediated/Percent of Total Effect Mediated

$$\text{Proportion Mediated} = \frac{\text{indirect effect}}{\text{total effect}} = \frac{\hat{\beta}_{\text{crude}} - \hat{\beta}_{\text{adj}}}{\hat{\beta}_{\text{crude}}} = \frac{\hat{\gamma}_X \times \hat{\beta}_Z}{\hat{\beta}_{\text{crude}}} = \frac{-4.4268}{-8.972} \times 100 = 49.34\%$$



Interpretation: 49.3% of the effect of study drug on changes in systolic blood pressure is mediated through (can be explained by) changes in BMI.

Note: The proportion mediated can be negative or greater than 100%. This can occur with *inconsistent (mediation) models* where at least one mediated effect has a different sign than the other mediated or direct effects in a model.

Note: It has been noted that the proportion mediated can be unstable with sample sizes <500.

Freedman LS. Confidence intervals and statistical power of the "validation" ratio for surrogate or intermediate endpoints. J Statist Plan Inference. 2001;96:143–53.
MacKinnon DP, Warsi G, Dwyer JH. A simulation study of mediated effect measures. Multivariate Behav Res. 1995;30:41–62.

Standard Error of the Indirect Effect

We can calculate a confidence interval around our proportion mediated (49.34%), but first we need to calculate the standard error of the indirect effect. To do so we will use the delta method.

Delta Method: The variance of a function of a random variable $f(X)$ is approximated by:

$$\text{Var}[f(X)] = [f'(X)]^2\text{Var}(X)$$

How is this variance derived? Recall the first order Taylor series approximation of a function of X (here let X be a r.v. and the function be expanded about μ, the mean of X):

$$g(X) = g(\mu) + (X - \mu)g'(\mu)$$

The variance of g(X), the function of X, can then be estimated as:

$$\text{var}\{g(X)\} = \text{var}\{g(\mu)\} + \{g'(\mu)\}^2\text{var}\{X - \mu\} \approx \{g'(\mu)\}^2\sigma^2.$$

For example, Var(ln X) $\approx (1/\mu)^2\ \sigma^2$. We substitute the estimates of μ and σ^2 to obtain the *delta method* estimator.

The variance of a function of k random variables $f(\boldsymbol{X})$ is approximated by:

$$\text{Var}[f(\boldsymbol{X})] = \sum_{i=1}^{k} f'_i(X)^2\ \text{Var}(X_i) + 2\sum_{i>j} f'_i(X)\, f'_j(X)\text{Cov}(X_i, X_j)$$

Standard Error of the Indirect Effect (applying the delta method)

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

$$\begin{aligned} SE\left(\hat{\gamma}_X \times \hat{\beta}_Z\right) &= \sqrt{\hat{\gamma}_X^2 \left(SE\left(\hat{\beta}_Z\right)\right)^2 + \hat{\beta}_Z^2 \left(SE(\hat{\gamma}_X)\right)^2} \\ &= \sqrt{(-5.48)^2(0.9555)^2 + (0.807)^2(0.8138)^2} \\ &= 5.277164 \end{aligned}$$

Z = $\frac{\text{indirect effect}}{SE(\hat{\gamma}_X \times \hat{\beta}_Z)} = -\frac{4.4268}{5.277164} = -0.839$, p = 0.401 [from pnorm(-0.839)*2 in R]

95% CI: -4.4268±1.96(5.277164) = (-9.704, 0.850)

95% CI(%): $\left(\frac{-9.704}{-8.972} \times 100, \frac{0.850}{-8.972} \times 100\right) = (-9.5\%, 108.2\%)$

Interpretation: Change in BMI is not a significant mediator of the relationship between treatment and change in systolic blood pressure (p = 0.401), even though it explains 49.3% (95% CI: -9.5% to 108.2%) of the effect of the treatment.

What happens if we also adjust for diet, what is the (indirect) effect of BMI?

```
proc reg data=top;
    model csbp = trtgrp cbmi diet_exclude;
run;
```

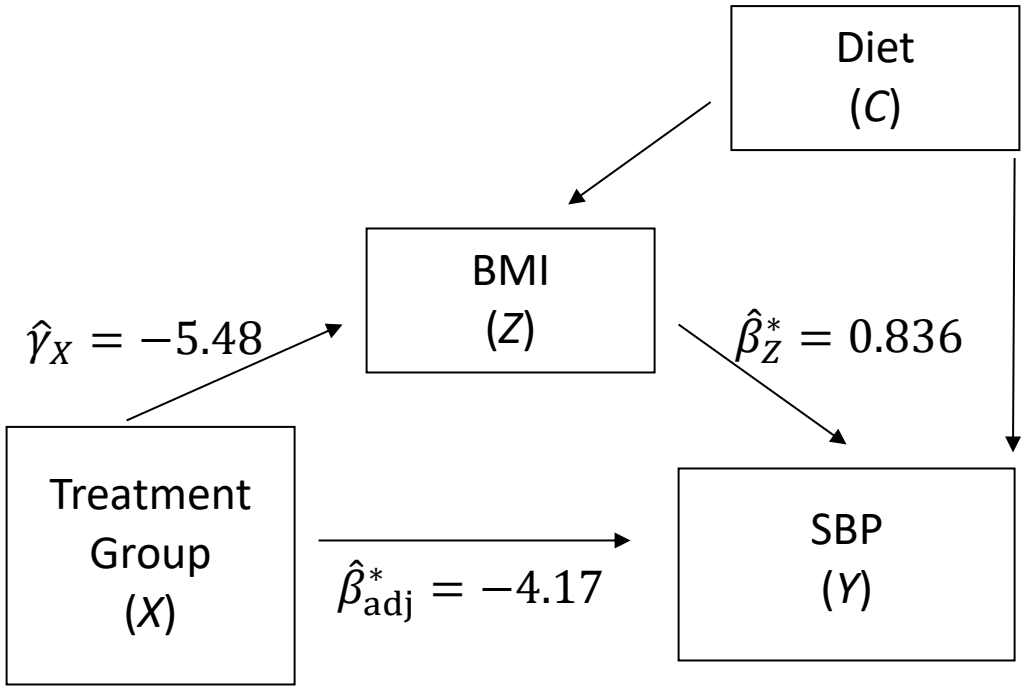
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	457.35282	152.45094	2.44	0.0998
Error	17	1062.45670	62.49745		
Corrected Total	20	1519.80952			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.14910	3.20317	-0.36	0.7242
trtgrp	1	-4.16951	6.98321	-0.60	0.5583
cbmi	1	**0.83619**	1.00514	0.83	0.4170
diet_exclude	1	-0.55959	4.10037	-0.14	0.8931

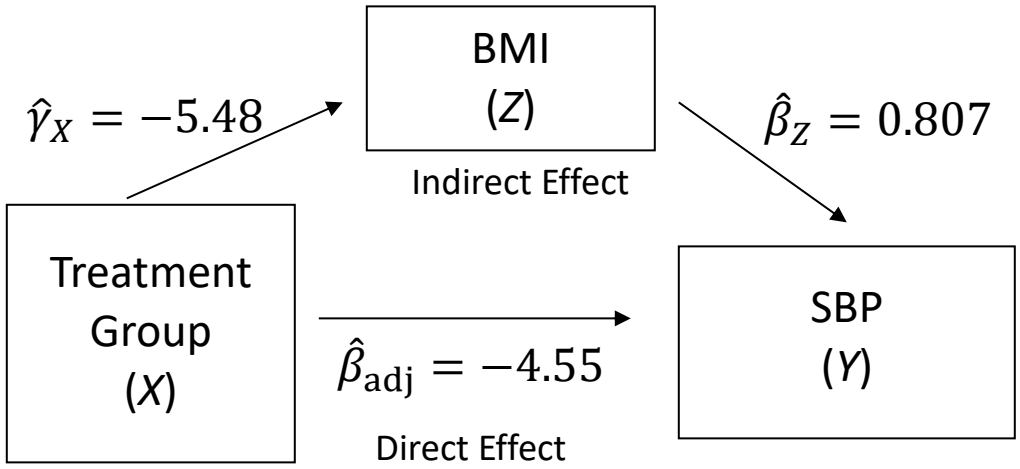
Calculation of Indirect Effect: $\hat{\gamma}_X \times \hat{\beta}_Z$: (-5.4833) x (0.8362) = -4.5851

Proportion Mediated: $\frac{\text{indirect effect}}{\text{total effect}} = \frac{-4.5851}{-8.972} \times 100 = 51.10\%$

True Causal Model:
(Diet is a confounder)



Compare $\hat{\beta}_Z^*$ and $\hat{\beta}_{adj}^*$ to $\hat{\beta}_Z$ and $\hat{\beta}_{adj}$ calculated previously without adjusting for diet:



Summary of Lecture

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