Confounding

- "Confounded" in ordinary English means confused or perplexed. by another factor

- The statistical use is essentially the same: Ex/ Our attempt to estimate the true effect of weight on heart disease was unsuccessful because the effect was mixed-up with the effect of age. Confounding is intrinsically about cause and effect, and depends on which effect one is interested in estimating.

- Confounding differs from **effect modification** (what we call an interaction effect in statsitics), but is not mutually exclusive from it, and thus much difficulty results in trying understand the two concepts together.

Confounding

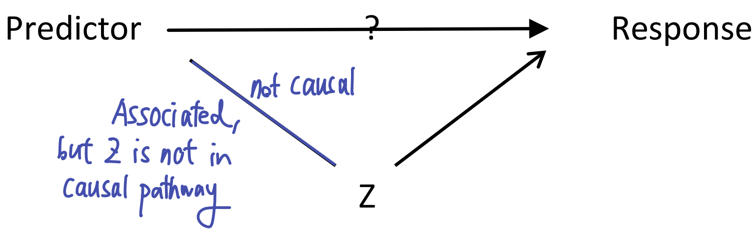
- **How much does increased weight place individuals at increased risk of dying from heart disease?**
  - People gain weight as they age. But as they get older, their risk of cardiac death is also greater due to many other factors. Age as confounder
  - Thus, unless we control for the effects of age, weight will appear to be more influential than it actually is.

  Here, age is said to **confound** the relationship between cardiac death risk and weight. "Exagerate or hidden the effect of..."

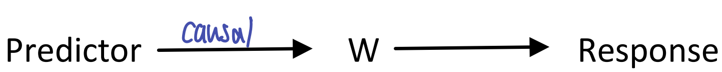
- A **confounder** is a variable related to the predictor of interest and causally related to outcome, but which is not in the causal pathway (i.e., in the relationship with the predictor, but not causally). One could argue individuals gain weight *because* they age, but not strictly, systematically in some mechanistic way according to their chronological age → Not in the causal pathway: "Association"

Confounding - Pathways

- Here, arrows indicate causal pathways, connecting bars without arrows indicate associations



(a) Z is a confounder



(b) W is not a confounder b/c in the causal pathway

Confounding and Model Variables

- One must consider actual (not statistical) relationships among variables when assessing confounding. For example, we might assume the causal pathway below, plausibly identifying nicotine as the *relevant exposure* for the outcome here:

not confounding.

$$\text{Smoking} \longrightarrow \text{Nicotine exposure increase} \longrightarrow \text{birth defect}$$

Nicotine is not caused solely by one's smoking

Given *this* pathway, would you adjust for (ie include) nicotine level in your regression predicting birth defect rates by smoking?

- **Probably not:** The meaning of this is unclear at best. Effect of smoking holding "the nicotine level fixed" when evaluating smoking doesn't make sense, because one cannot hold the nicotine level fixed at some value when it changes as a direct consequence of smoking (we assume here the only way to incur nicotine exposure is by smoking) but it's not.

Nicotine is also caused by:

- Second hand smoke
- Pollution
- Nicotine patch

Confounding and Model Variables

- However, note that 'adjusting for smoking' could make sense if nicotine is also present because of heavy environmental smoke exposure (second-hand smoke) or nicotine patch use, but that is a different causal pathway/question

- **Temporal ordering** is necessary in causal relationships. Example:

$$\text{Maternal Smoking} \longrightarrow \text{birth defects} \longrightarrow \text{increased infant hospital stays}$$

- Maternal smoking causes increased number of hospital stays after birth and birth defect for infants. When analyzing whether maternal smoking causes birth defects, is number of hospital stays a confounder? Hospital Stay causes with birth defect?
  No. Temporality Wrong.

  No. In addition to being nonsensical, increased number of hospital stays occurs after birth defect occurrence. Thus it cannot be causally related to the response variable - birth defect.

Confounding vs. Effect Modification

- As mentioned earlier, the distinction between confounding and effect modification can be elusive. It can be conceptualized as follows:
  - Stratums = **Confounding factors** are third factors that create an apparent relationship between two other factors (a predictor and response) that is actually absent. This third factors is associated with both of the other factors. It may be causally associated with the response, may have reverse causality with the predictor.
    - **Ex:** smoking is related to lung cancer, consuming alcohol is related to smoking (somewhat higher co-prevalence of the behaviors). Consuming alcohol is not meaningfully related to lung cancer, but might appear so if one did not consider smoking as a **confounder**

Confounding vs. Effect Modification

- Stratums ≠
  Effect modifiers are those factors that are related to both the predictor and the response, and modify the strength of the association between the predictor and response

- **Ex:** smoking is related to lung cancer, radon exposure is also related to lung cancer.
  * Smoking may intensify the effect of radon on lung cancer, or alternatively . . .
  * prevalence variations in lung cancer by radon exposure may be trivial once we account for smoking.
  * When analyzing one predictor (say, radon exposure), we should definitely *adjust for smoking*. (i.e., have it in our model, stratify by it (more on this soon), etc).

Confounding vs. Effect Modification

- Consider now this third role (after main effects and effect modification (interaction) that variables can play in a regression model: we have dealt with response variable Y, predictor variable X, but up until now we have not made a distinction between the predictor variable and a confounder **What is the difference between a predictor and a confounder?**

- We are usually interested in saying something about the effect of the predictor, while we are not particularly interested in confounder effect estimates, although they affect the response.

- A confounder can be thought of as a *nuisance factor* we have to adjust for so that our estimate of predictor effect are unconfounded (or at least, less so)

Detect EM ; $\underbrace{\text{Remove Confounder}}_{\text{Adjust}}$

Confounding vs. Effect Modification

Consider two models for assessing the effect of X (exposure) on Y:

1. **The marginal model:** $Y = \beta_0 + \beta_1 X + \epsilon$
2. **The adjusted model:** $Y = \beta_0' + \beta_1' X + \underbrace{\beta_2' Z}_{\text{Confounder}} + \epsilon$

- What if β_1 is different from β_1'? (not as a statistical question, but substantively)

- Here, "different" means that when you take the values of $\overset{\text{Confounder}}{Z}$ into account, you would change your opinion about the effect of X on Y in a practically important sense. Z is then said to be a "confounder" and it should be included in the regression model.

- What if β_2' is not significantly different from 0?
  Answer: statistical significance of β_2' has much less bearing on whether Z should be considered a confounder, and whether we include it.

Modeling Goals and Purposes

There are two goals of regression models, both equally important:

1. **Prediction**
   - want the model to fit data well
   - want replicability
   - mechanism is not (as) important v.s. explanatory
2. **Explanation**
   - need accurate estimates of coefficients
   - the "correct" form of the model is one goal in itself
   - fitted model may be used for important policy decisions

Thus far, we have focused on models with explanation in mind, proposing and testing models that made sense and had plausibility we could justify on the basis of biological, ecological, or socio-economic theories.

Control of confounding is key in explanatory modeling.

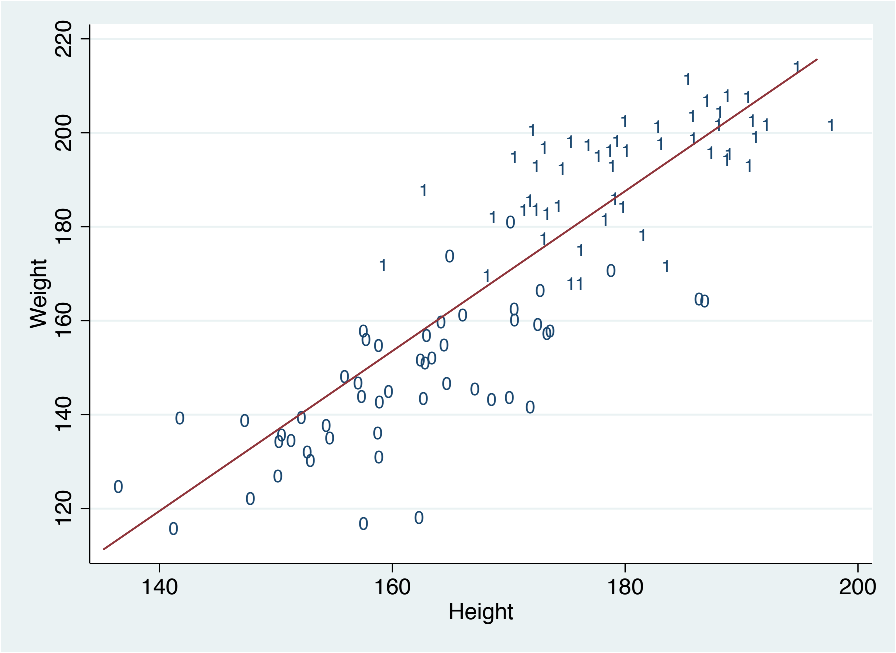
Confounding

- Again, a confounder is a variable that is related to the predictor as well as the response (even in the absence of the predictor). It is not 'caused' by the predictor

  For example, examine the weight-height relationship in a dataset: the marginal relationship is given with the following regression line:

```
. regress w h
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  264.06
       Model |  51652.0816     1  51652.0816           Prob > F      =  0.0000
    Residual |   19169.393    98  195.606051           R-squared     =  0.7293
-------------+------------------------------           Adj R-squared =  0.7266
       Total |  70821.4746    99   715.36843           Root MSE      =  13.986
------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   1.701886   .1047316    16.25   0.000      1.49405    1.909723
       _cons |  -118.7635   17.78881    -6.68   0.000    -154.0649   -83.46223
------------------------------------------------------------------------------
.  twoway (scatter w h, mlabel(sex) msymbol(none))  (lfit w h), xtitle("Height")
```

```
ytitle("Weight") legend(off)
```



- males (symbol=1) and females (symbol=0) form somewhat distinct groups with respect to height. What do we know about the relationship between sex and weight? Males tend to be heavier (for any given height) than women. Males also tend to be taller.

Confounding

- The slopes for these subgroups might also be different. So, is sex is a potential confounder?

\- Add sex to the model:

```
. regress w h sex

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  311.78
       Model |  61287.6579     2  30643.829            Prob > F      =  0.0000
    Residual |   9533.8167    97  98.2867701           R-squared     =  0.8654
-------------+------------------------------           Adj R-squared =  0.8626
       Total |  70821.4746    99  715.36843            Root MSE      =   9.914

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   .9639984   .1051921     9.16   0.000     .7552211    1.172776
         sex |   27.81749   2.809484     9.90   0.000     22.24144    33.39354
       _cons |  -7.728952   16.87486    -0.46   0.648    -41.22088    25.76298
------------------------------------------------------------------------------
```

\- How much does slope change? Less effect at 0.963 now vs. 1.702

Correction of Confounding

- We can also examine the relationship of height to weight separately by sex, performing a *stratified* analysis

Stratify

```
. sort sex
. by sex: reg w h

--------------------------------------------------------------------------------
-> sex = 0

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   50.56
       Model |  5692.98528     1  5692.98528           Prob > F      =  0.0000
    Residual |  5404.60962    48  112.596034           R-squared     =  0.5130
-------------+------------------------------           Adj R-squared =  0.5028
       Total |  11097.5949    49  226.481529           Root MSE      =  10.611

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   1.024336   .1440569     7.11   0.000     .7346905    1.313982
       _cons |  -17.37486   23.07847    -0.75   0.455    -63.77722    29.02751
------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
-> sex = 1

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   30.74
       Model |  2612.09064     1  2612.09064           Prob > F      =  0.0000
    Residual |  4078.44897    48  84.9676868           R-squared     =  0.3904
-------------+------------------------------           Adj R-squared =  0.3777
       Total |   6690.5396    49  136.541625           Root MSE      =  9.2178

------------------------------------------------------------------------------
           w |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           h |   .8692904   .1567825     5.54   0.000      .554058    1.184523
       _cons |   37.02104   28.06089     1.32   0.193    -19.39915    93.44124
------------------------------------------------------------------------------
```

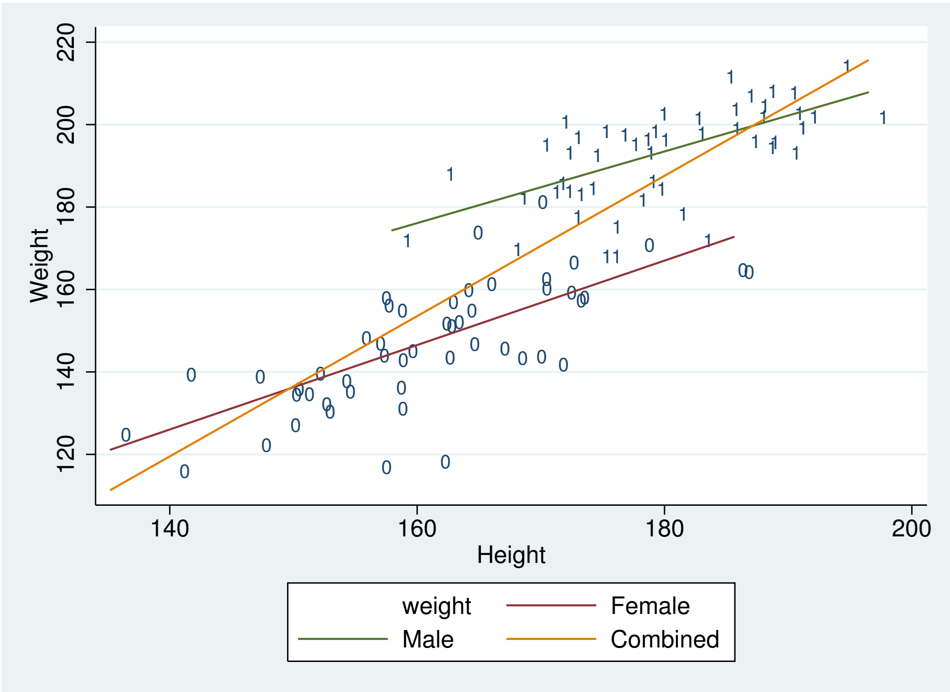
- Note that the slope overall (adjusting for sex) of 0.963/cm is approximately an average of the sex-specific slopes of 1.024 (females) and 0.869 (males)

Correction of Confounding

- The similar but less steep slopes in men and women separately:

Separate regression for each line.

```
. twoway (scatter w h, mlabel(sex) msymbol(none)) (lfit w h if (sex==0))
 (lfit w h if (sex==1)) (lfit w h), xtitle("Height")
ytitle("Weight")  legend(order(1 "Weight" 2 "Female" 3 "Male" 4 "Combined"))
```



· Slope appears similar, but distinct β_0 and different lines.

Correction of Confounding

- We can formally test whether the slopes are different (how?).

- Turns out that the slopes are not different (by statistical or material criteria) as the plot seems to indicate. Adjusting for sex lets us examine the true relationship between weight and height more accurately.

- Note that age and sex are the confounding usual suspects in medical and epidemiologic studies, and so we often adjust for them in analyses.)

• Question: Why is sex not considered an effect modifier?

Sex is not magnifying / shrinking weight

Both gender (Stratum) are ≈ slope ; v.s. EM have different slopes across.

Confounding vs. Effect Modification: Again

- Here, while sex is an important predictor of weight . . .
  - There is clearly no differential effect of height on weight according to sex. The slopes for height within males and within females are about the same. There is no interaction effect
  - Both of these slopes (Stratums) are different from the marginal slope or unadjusted (Crude) effect for height (e.g., ignoring sex)
- Thus, the effect of height on weight is said to be *confounded* by sex

Confounding in Observational Studies

- Framework for many observational studies: three types of variables:

  a) Response (outcome, dependent variable) Y

  b) Predictor variable X - exposure of interest

  c) Covariates that may be confounders, sometimes called control variable(s), Z

- Distinction between X and Z is that we CARE about predictors X while the covariates Z are considered nuisance variables we must control to avoid biased effects, leading to wrong conclusions

- To address confounders in studies.

  a. Must carefully consider context, conceptual model, and **collect suspected confounding factors**

  b. Practically, might analyze with and without adjustment for a suspected confounder

Confounding in Observational Studies

- Models to contrast (often presented in epidemiologic studies)
  unadjusted model: $Y = \beta_0 + \beta_1 X + \epsilon$
  adjusted model: $Y = \beta_0' + \beta_1' X + \beta_2' Z + \epsilon$
- Check to see whether β_1 and β_1' are different from each other (not strictly a statistical question, but materially)
- If yes: Z could be a confounder. What if β_2' is not statistically significant? Does not mean that Z is not a confounder. We may nonetheless retain to maximally control bias
- So, what variable should we consider the "usual suspects"?
  a. Factors known or generally thought to influence Y ("the risk factors" for the response)
  b. Factors thought to be important for interpretability, credibility of findings

Consequences of Ignoring Confounders

- When confounding is present, the contributing effect of X is the same for each value of Z (i.e., in a linear main effects model), but not taking Z into account distorts the true effect.

  Same across stratums

- Ignoring **positive confounders** (those directly related (positively or negatively) related to both outcome and predictor) may Causing **overestimate** the effect.

  Example: Radon exposure is associated with lung diseases, including cancer. So is working in mines (where radon is present) and related behavioral factors associated with miners

- Ignoring **negative confounders** (those positively related to one, and negatively to the other) may Causing **underestimate** the effect.

  Example: exercise may reduce risk of several cancers, while aging increases the risk. Without controlling for age, exercise effect may be underestimated

Can we Eliminate Confounding?

- In studies where we are sampling outcomes and observing covariates that may be explanatory or predictive, confounding can not be completely ruled out. These observational studies (a bit of a misnomer) are always subject to measured and more importantly <u>unmeasured</u> factors that influence both predictors and outcomes
- Confounding can be addressed via **prospective study design** that assigns exposure to the key variable(s) of interest and uses **randomization** to make the assignment. This is the design used historically in agricultural and industrial testing, and ubiquitously in medicine since the mid 20th century
- Needless to say, we cannot randomize to deleterious exposures, which are most often of interest in epidemiology. Sometimes, exposure mimicking randomization may occur
- Thus, we cannot completely 'model our way' out of confounding

Can we Eliminate Confounding?

- **Why not?** With the advent of Big Data, machine learning, neural networks, etc, it would seem that we can account for all or nearly all confounders. We can get far with this approach
- However, we cannot account for unmeasured or unknown potential confounders. Randomization equally allocates these to the exposure groups
- **Effect modification** of the observed effects persist, and can be explored and largely validated via modeling

A Randomized Experiment

We have two treatments (exposures) that we apply to experimental units. We randomize to Trt 1, and then randomize to Trt 2, or equivalently randomize to the 4 groups

	Exposure 1		Total
Exposure 2	level 1	level 2	
level 1	a	b	a+b
level 2	c	d	c+d
Total	a+c	b+d	a+b+c+d=N

Then we can evaluate different models:

$$\mathrm{E}(Y|exp) = \beta_0 + \beta_1 exp1 + \beta_2 exp2 + \beta_3 exp1xexp2$$

$$\mathrm{E}(Y|exp) = \beta_0 + \beta_1 exp1 + \beta_2 exp2$$

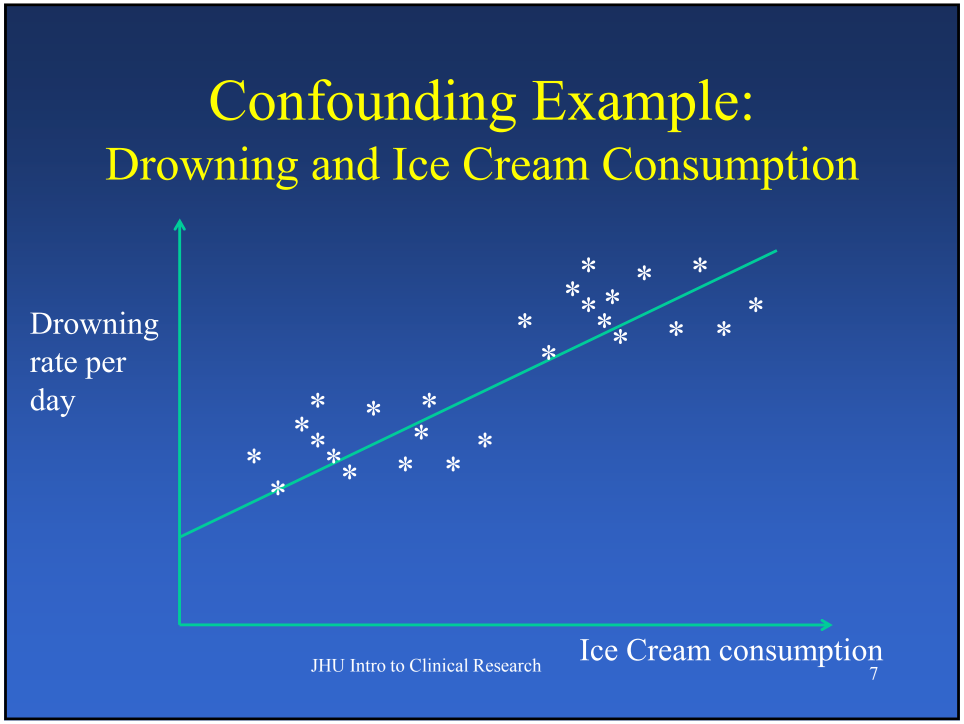
$$\mathrm{E}(Y|exp) = \beta_0 + \beta_1 exp1$$

$$\mathrm{E}(Y|exp) = \beta_0 + \beta_2 exp2$$

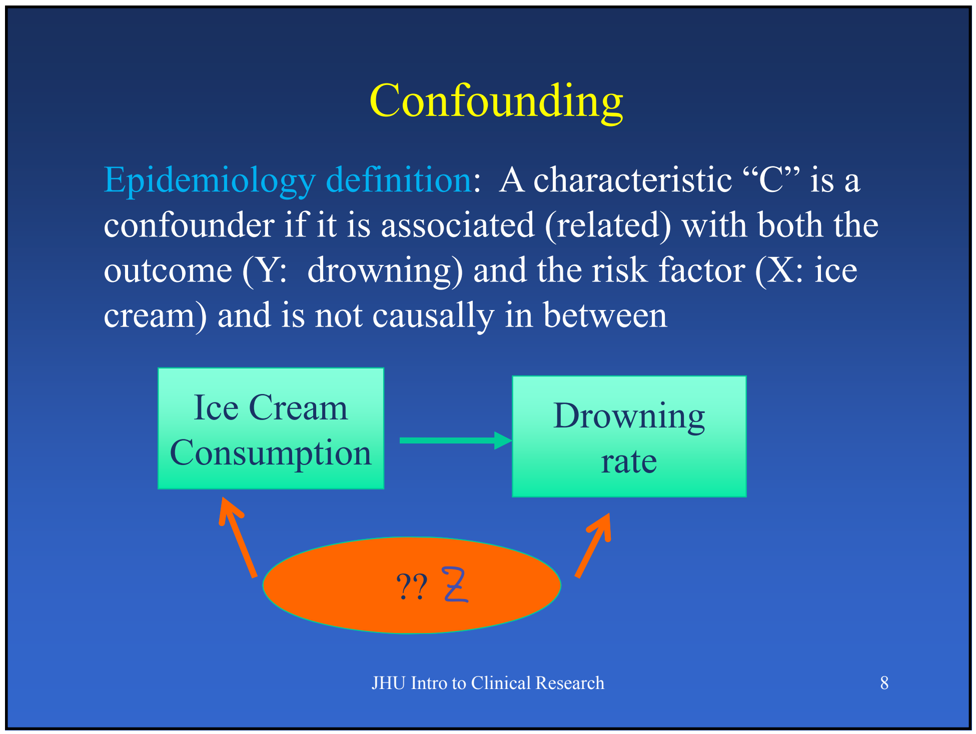
Effect Modification (Interaction) and Confounding - Summary

- Confounding is a bias that we hope to prevent or control - makes X seem related to Y but it is not
- Effect modification is a real effect - differential "Strength" effect on Y of X_1 in presence/absence/at value of X_2 Without EM: "Average" effect
- Confounding is something to avoid and so confounders need to be included in the analysis, but can stratify to adjust to compare to crude.
- Effect modification, if not accounted for, provides an 'average effect' ignoring the third variable, may not be wrong but is much less informative. With qualitative interactions, conclusion may be wrong

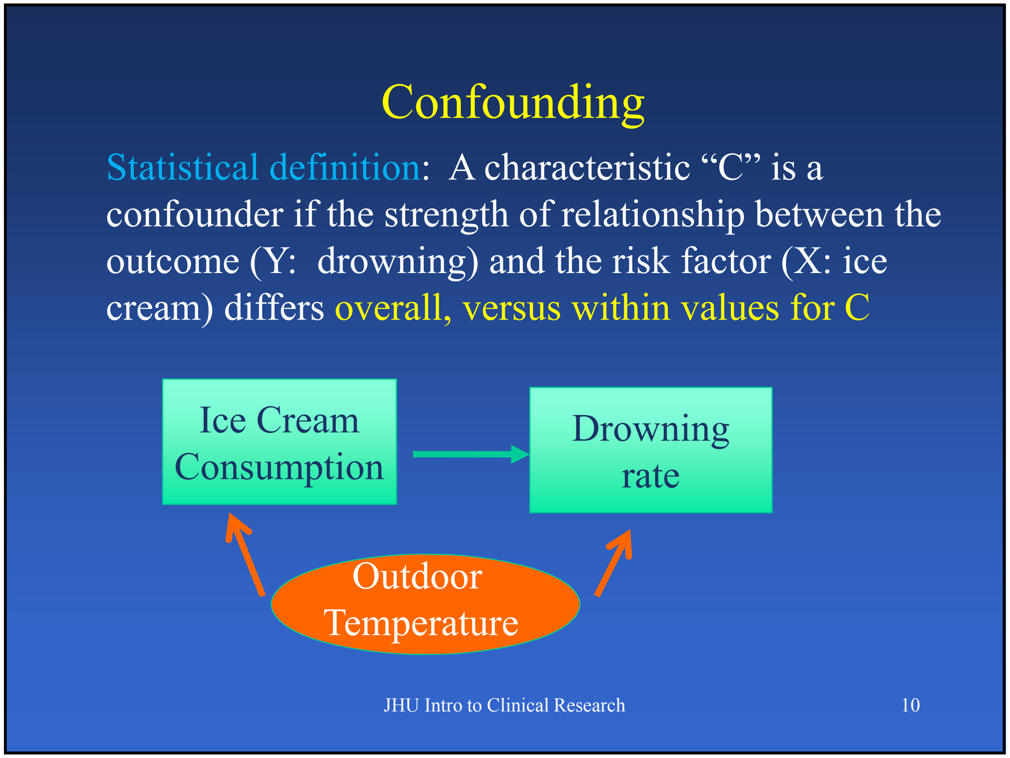
Thinking About Confounders
(w/out permission (i.e. stolen) from K. Bandeen-Roche JHU)



Thinking About Confounders (cont.)

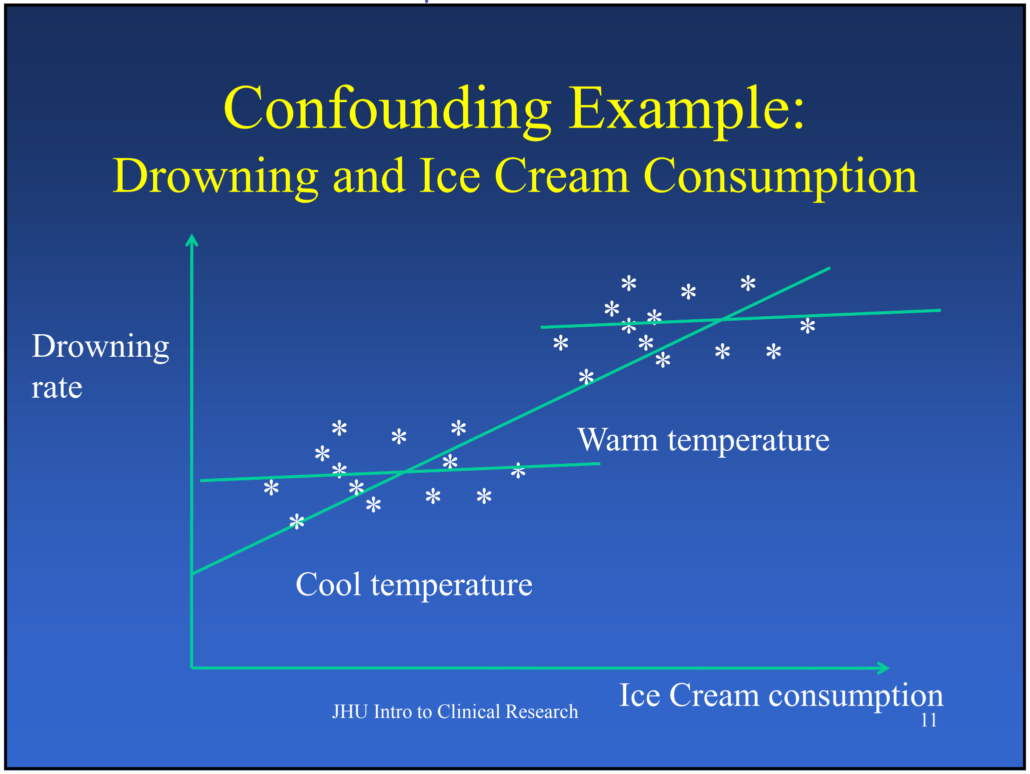


Thinking About Confounders (cont.)



Thinking About Confounders (cont.)

Same slope



Effect Modification? (hypothetical)

Different slope