ANCOVA

Concomitant Variables (co-variates)

- A variable that has been collected for each subject in advance (usually) of the experimental manipulation
- e.g. Pre-existing differences among subjects on some (continuous) variable

Two ways to use covariates

- covariates can be used BEFORE the experiment to *experimentally* control for pre-existing differences
  - use score on a covariate as a way to assign subjects to groups
  - e.g. match different groups based on score on a covariate measure
  - e.g. match different treatment groups on age, or gender, or IQ
- covariates can be used **AFTER** the experiment to "statistically" adjust for pre-existing differences
  - allow for variation in the concomitant variable both within and between groups
  - during the analysis, statistically account for the relationship between the covariate and the dependent variable

An Example

- let's say in September we assigned a different math textbook to two different grade 10 classrooms
- we want to see if different math texts result in different test scores on the christmas exam
- we select 10 students at random from each class, each takes a 25 item math test
- we also have student's IQ scores

An Example

- the IQ scores allow us to ask the question:
- Is there a difference on the post-test,
  after pre-existing differences in general intelligence are accounted for?

ANCOVA

- ANCOVA is a method to account for the relationship (if it exists) between a covariate and a dependent variable
- it is a way to statistically adjust for differences on the concomitant variable by including it as a predictor variable
- ANCOVA is just like regular ANOVA in terms of the model comparison approach
- we have a restricted model and a full model
- note however one of the variables in our model (i.e. the covariate) is a continuous variable

Logic of ANCOVA

- the question answered by ANCOVA:

★ Would the groups have been different on the post-measure **if they had been equivalent on the covariate?**

- i.e. is the observed difference in the dependent variable due to our experimental manipulation, or simply due to the pre-existing differences in the covariate?

Logic of ANCOVA

- Including a covariate in the model affects the analysis in two ways:
- within-group variability will be reduced
  - by an amount dependent on the strength of the relationship between the dependent variable and the covariate
  - typically a substantial reduction in unexplained variance
  - a smaller error term
  - greater power
- an increase in the estimate of the effect size

ANCOVA is not a substitute for randomization

- ANCOVA will only equate groups "statistically" on a single covarying variable
- randomization (over the long run) guarantees that groups will be equated on ALL relevant dimensions, not just the covariate

Another fictitious example

- the hippocampus is a brain structure thought to be involved in (spatial) memory
- is the volume of the hippocampus greater for taxi drivers than for people who aren't taxi drivers?
- 2 groups of people (taxi drivers, non-taxi drivers) scanned using MRI, volume of hippocampus estimated
- problem: let's say we know that hippocampus volume is also affected by age*
- we can include age as a covariate in the analysis

* *I'm making this up for the purposes of the example, I have no idea if it's true*

Linear Models for ANCOVA

$$\texttt{Restricted:} Y_{ij} = \mu + \beta X_{ij} + \epsilon_{ij}$$

$$\texttt{Full:} Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

- Yij is the score of the ith individual in the jth group on the dependent variable
- mu is a grand mean
- beta is a regression coefficient
- Xij is the score of the ith subject in the jth group on the covariate (e.g. initial weight)
- alpha_j is the treatment effect
- eij is the error term for the ith subject in the jth group

What is the cost?

- we lose a single degree of freedom
- because it's necessary to estimate Beta (slope of line)
- done using a least-squares criterion
- same thing as a linear regression

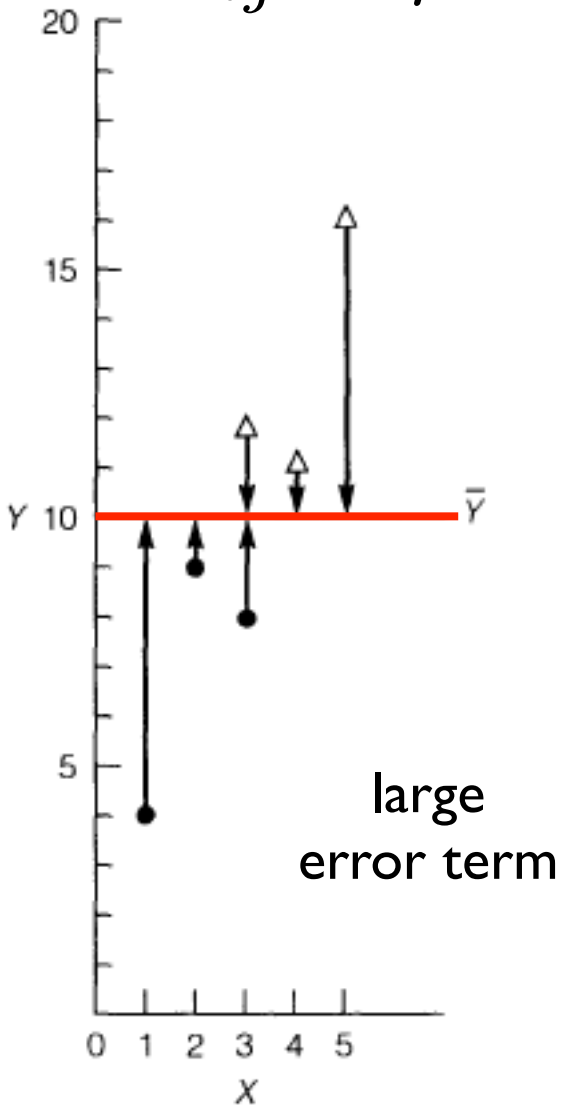
$$\texttt{Restricted:} Y_{ij} = \mu + \beta X_{ij} + \epsilon_{ij}$$

$$\texttt{Full:} Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

Restricted Model

ANOVA

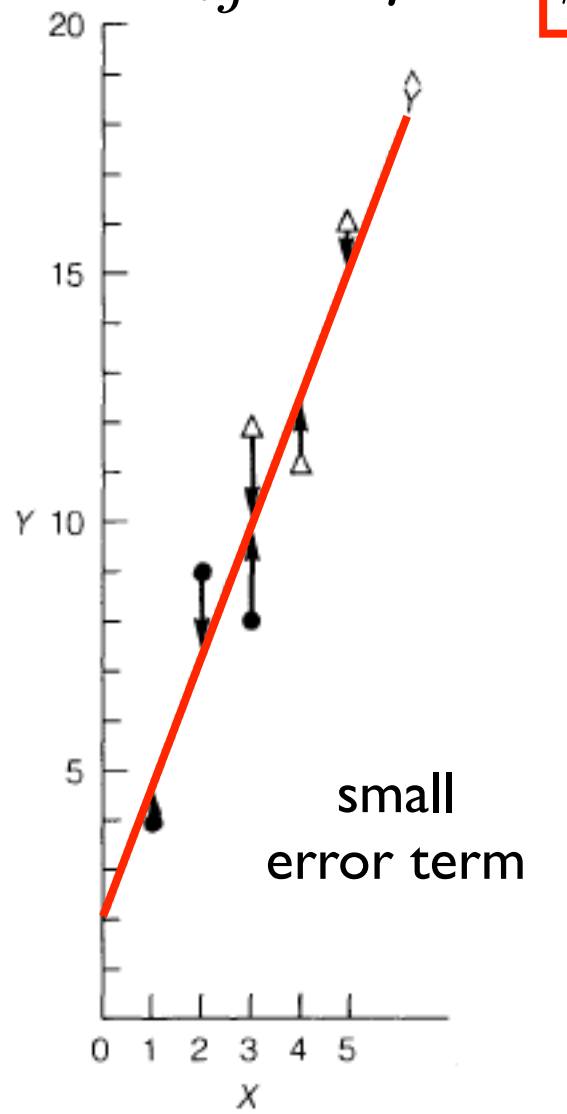
$$Y_{ij} = \mu + \epsilon_{ij}$$



large
error term

ANCOVA

$$Y_{ij} = \mu + \boxed{\beta X_{ij}} + \epsilon_{ij}$$

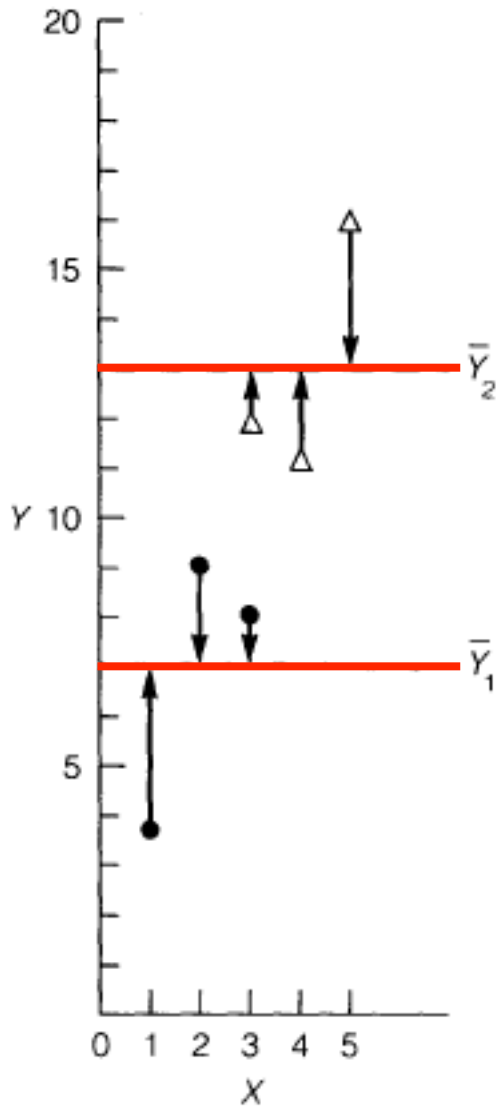


small
error term

Full Model

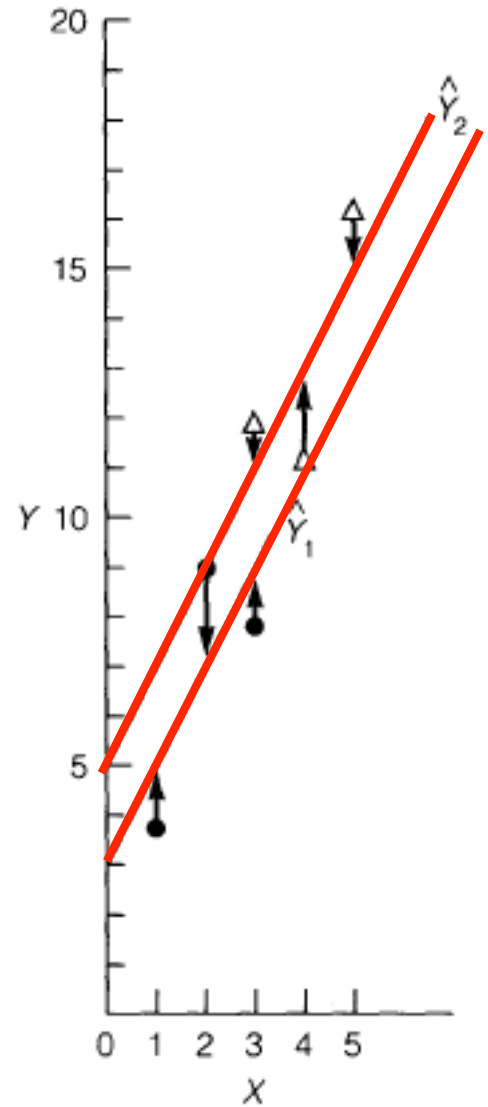
ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$



ANCOVA

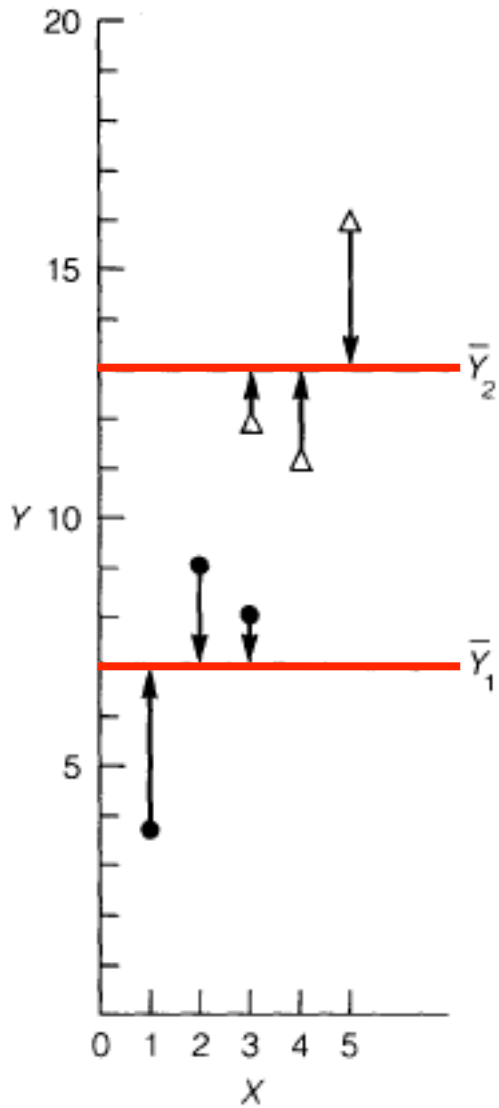
$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$



Full Model

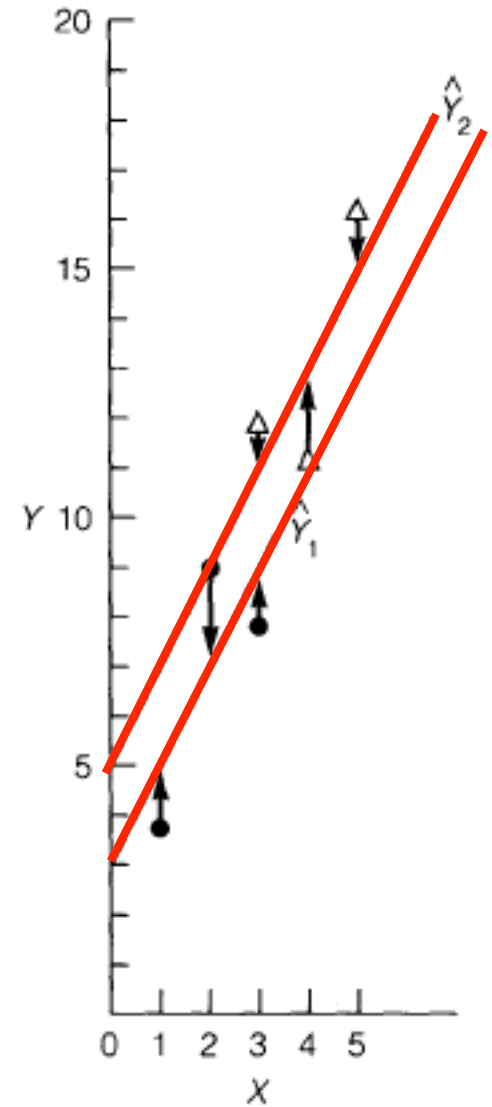
ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$



ANCOVA

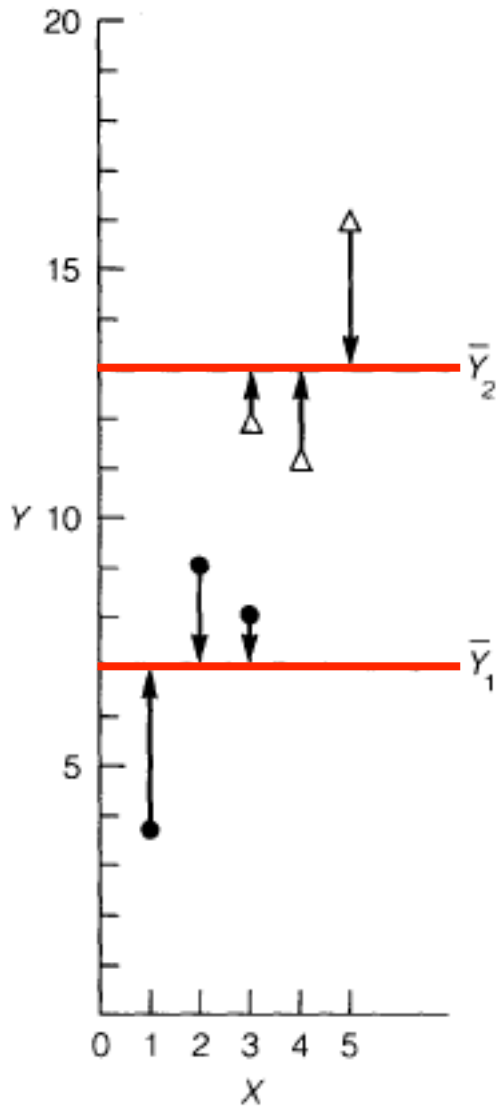
$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$



Full Model

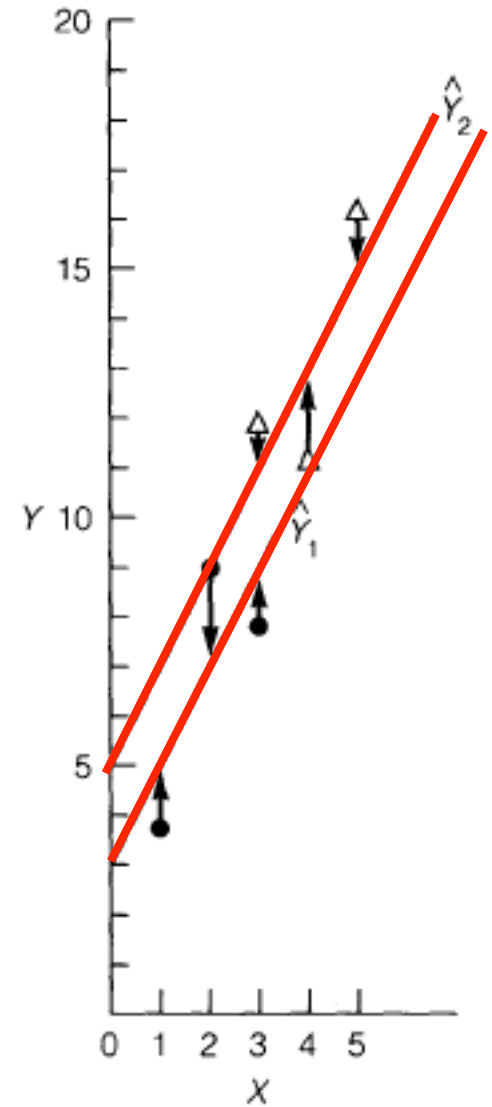
ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$



ANCOVA

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

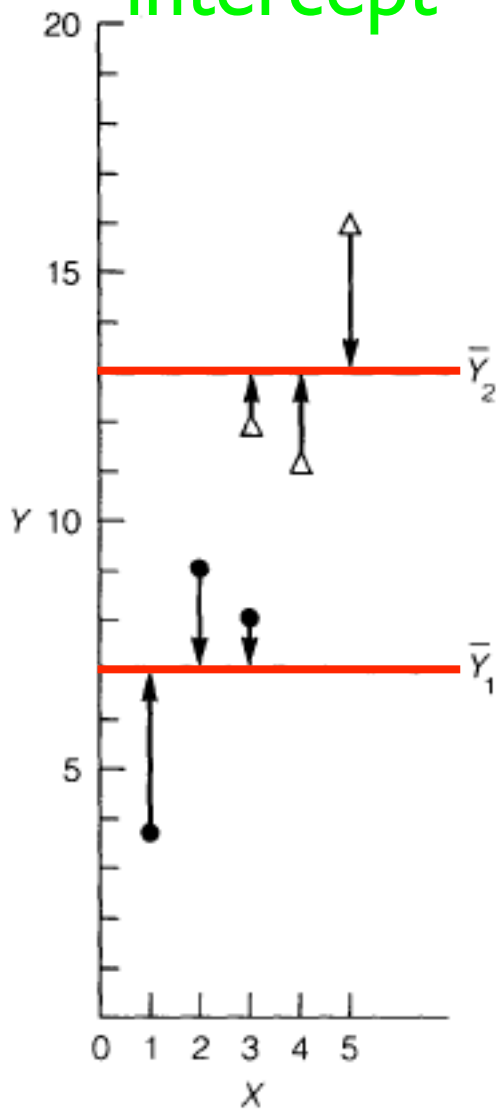


Full Model

ANOVA

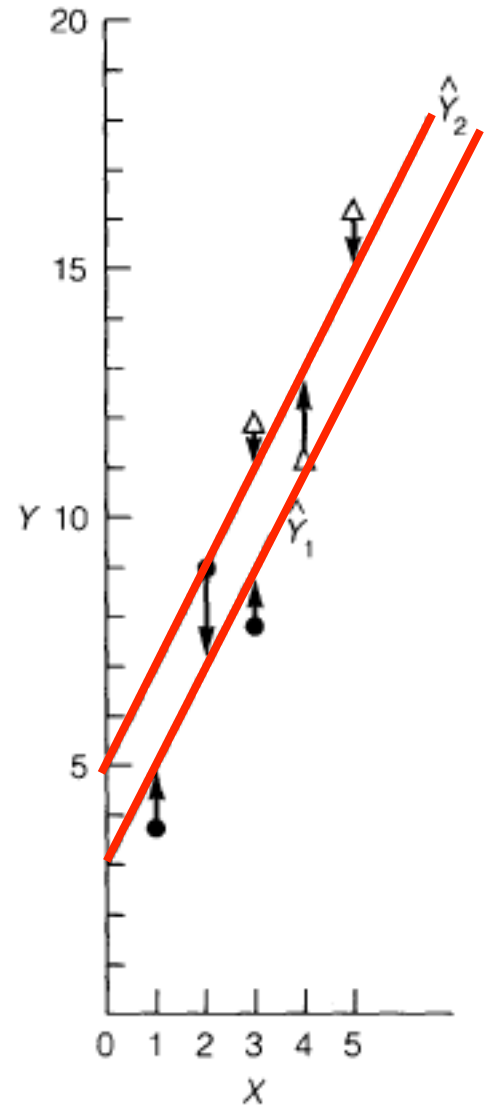
$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

intercept



ANCOVA

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

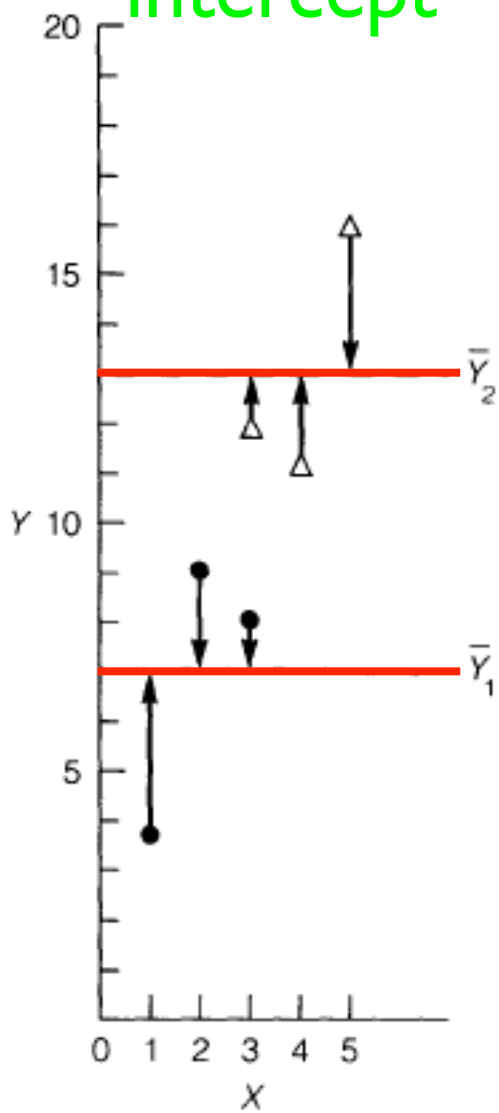


Full Model

ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

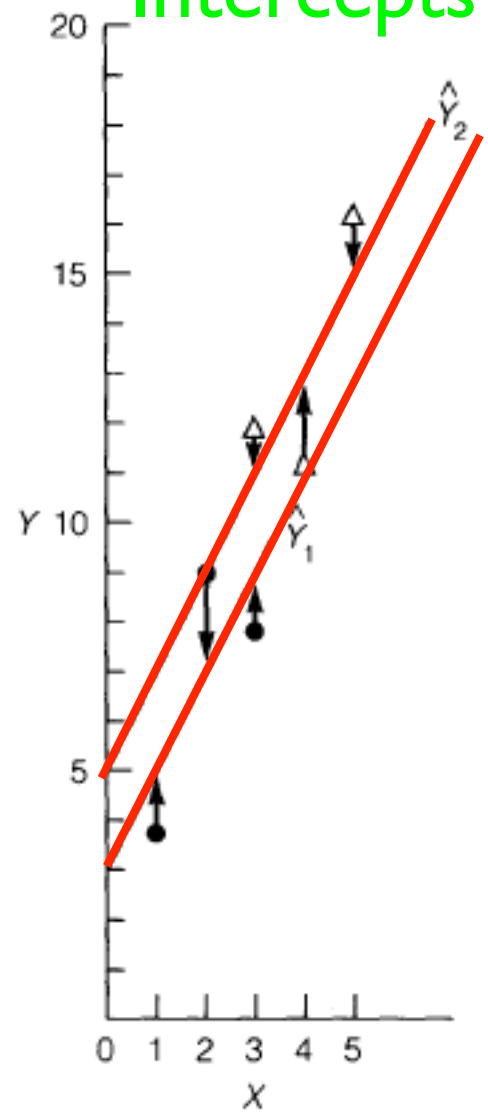
intercept



ANCOVA

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

Intercepts

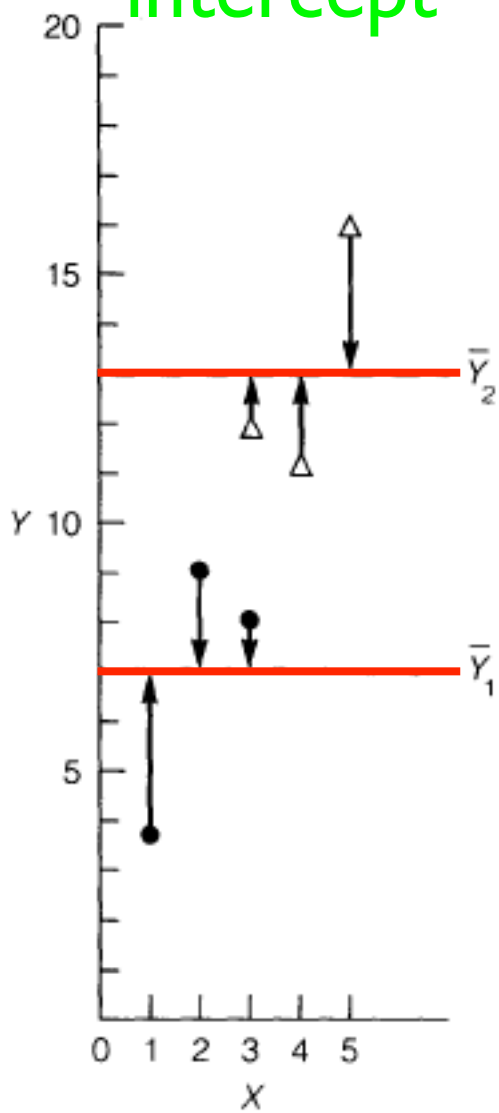


Full Model

ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

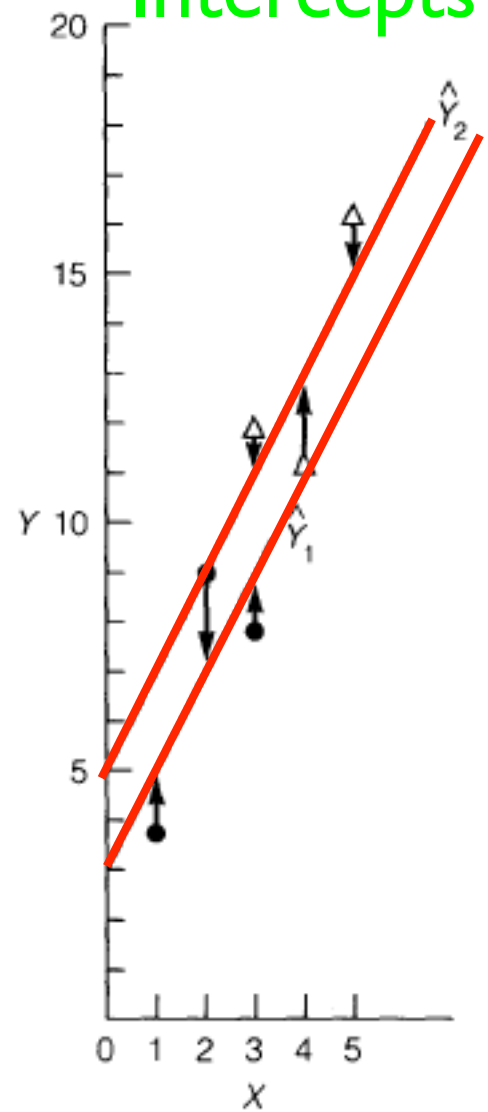
intercept



ANCOVA

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

Intercepts

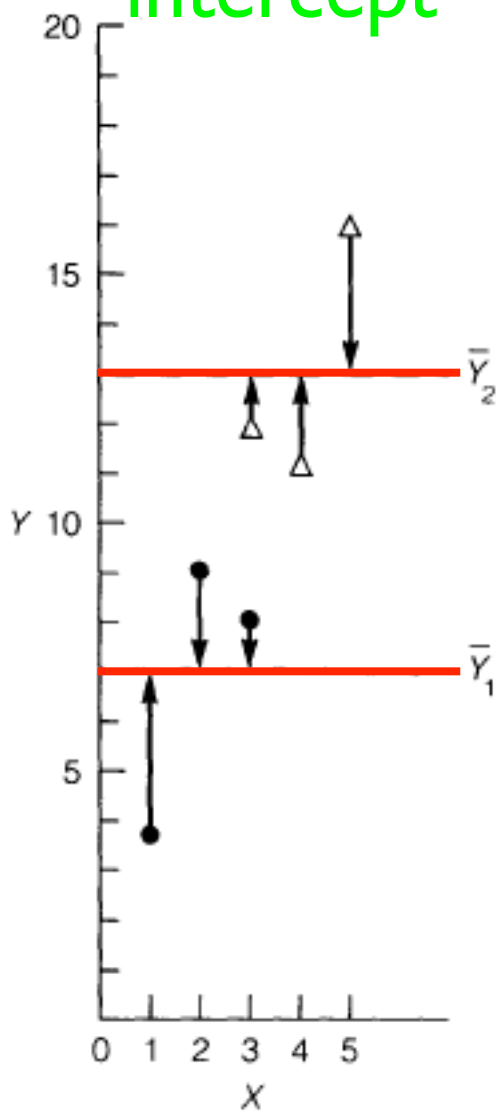


Full Model

ANOVA

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

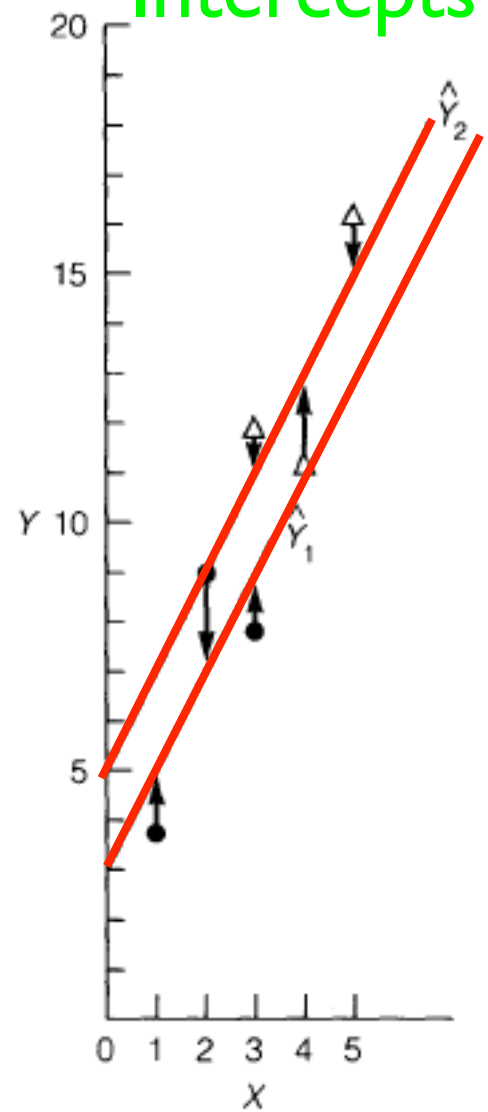
intercept



ANCOVA

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

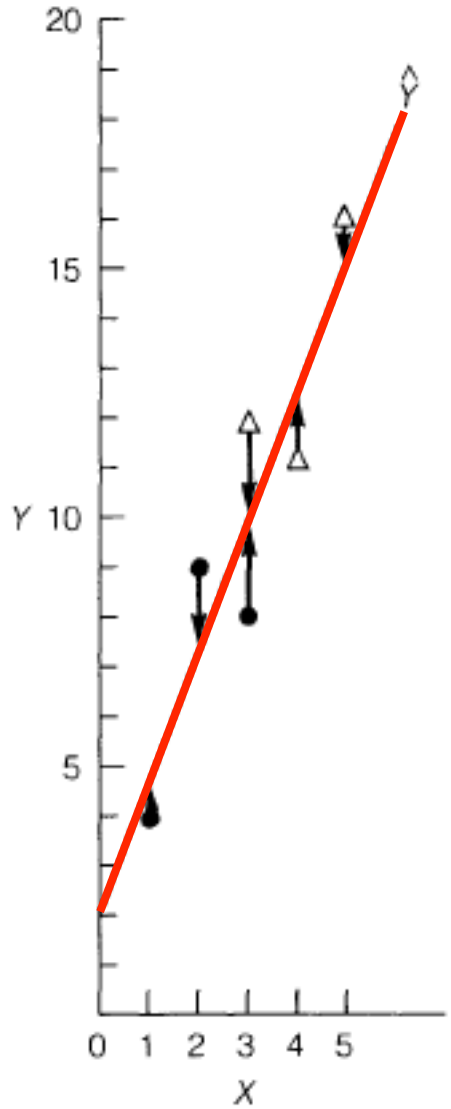
Intercepts slope



ANCOVA

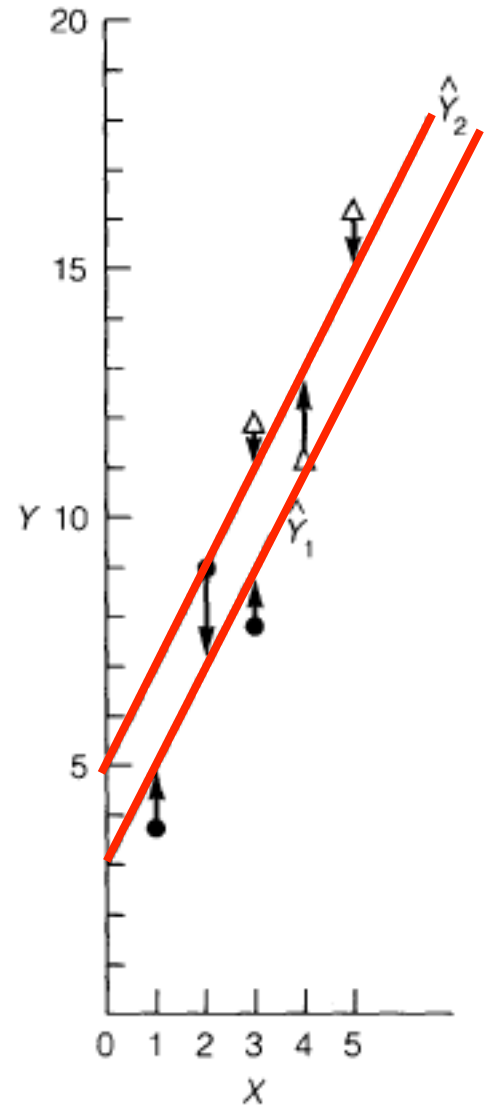
restricted model

$$Y_{ij} = \mu + \beta X_{ij} + \epsilon_{ij}$$



full model

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$



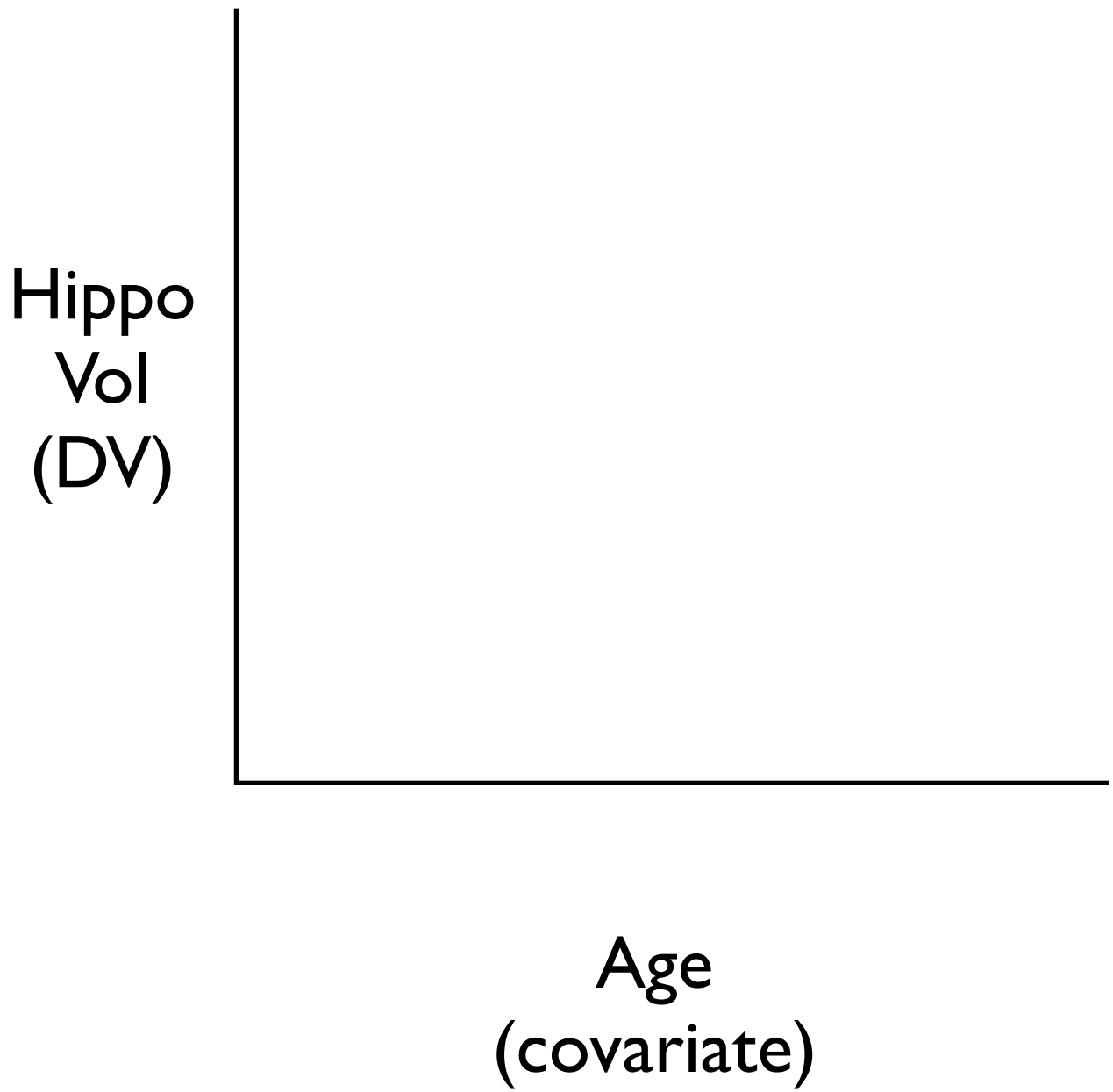
Accounting for Variance

- by including a covariate one can account for more variance in the data
- one can **statistically** adjust for pre-existing differences between groups
- if the covariate is not correlated with the DV there will be no beneficial effect of using ANCOVA
- BUT
- covariate must be independent from the experimental treatment (see later)

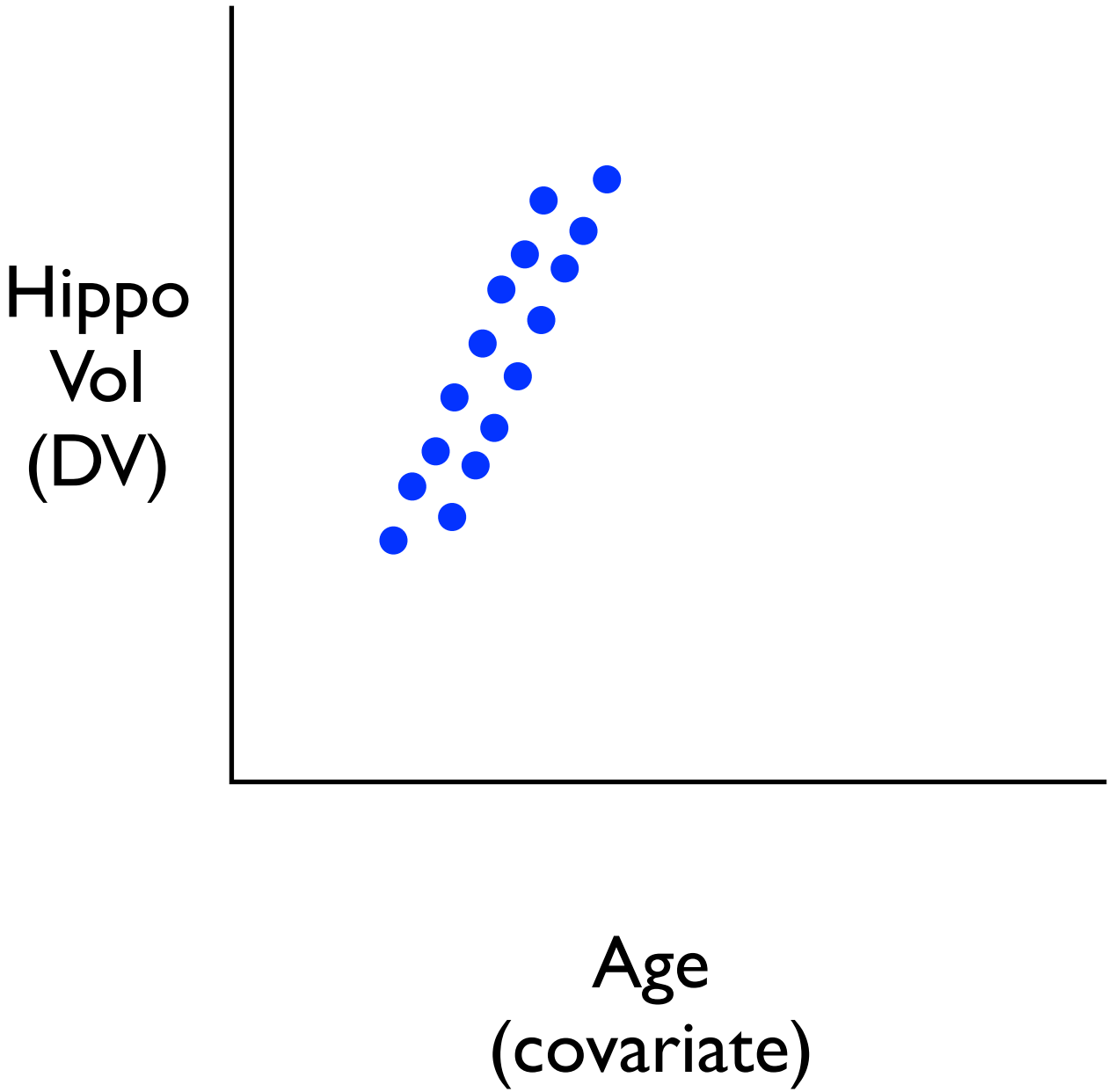
Corrected Means

- once you account for the relationship between a covariate and the dependent variable,
- you can generate "adjusted means" for each group
- ★ What would the means of each group be, ***if the scores on the covariate had been equal?***
- adjust mean of each group in proportion to the relationship between covariate and DV
- let's look at some graphical examples

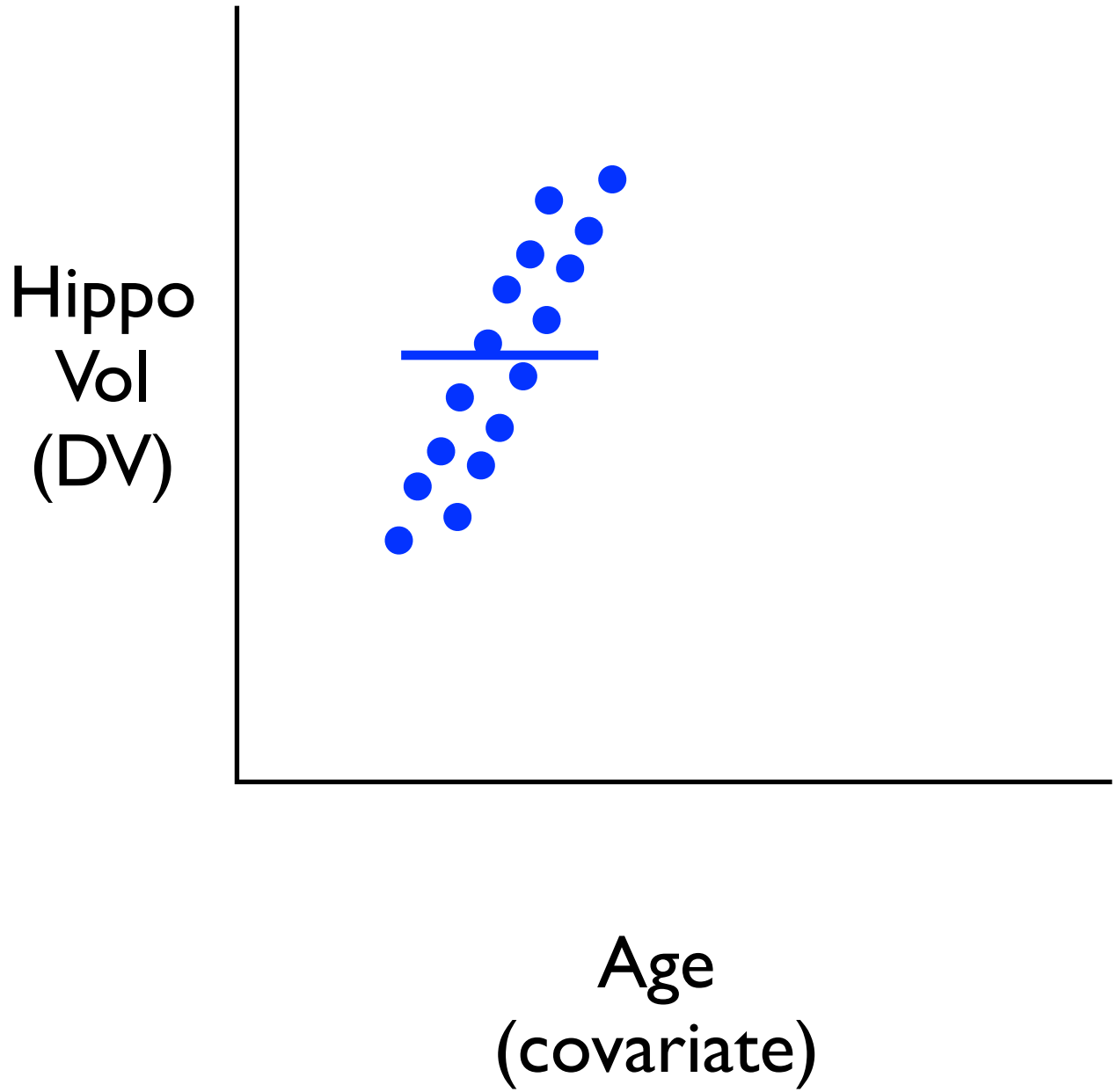
Raw means: Y2 > Y1



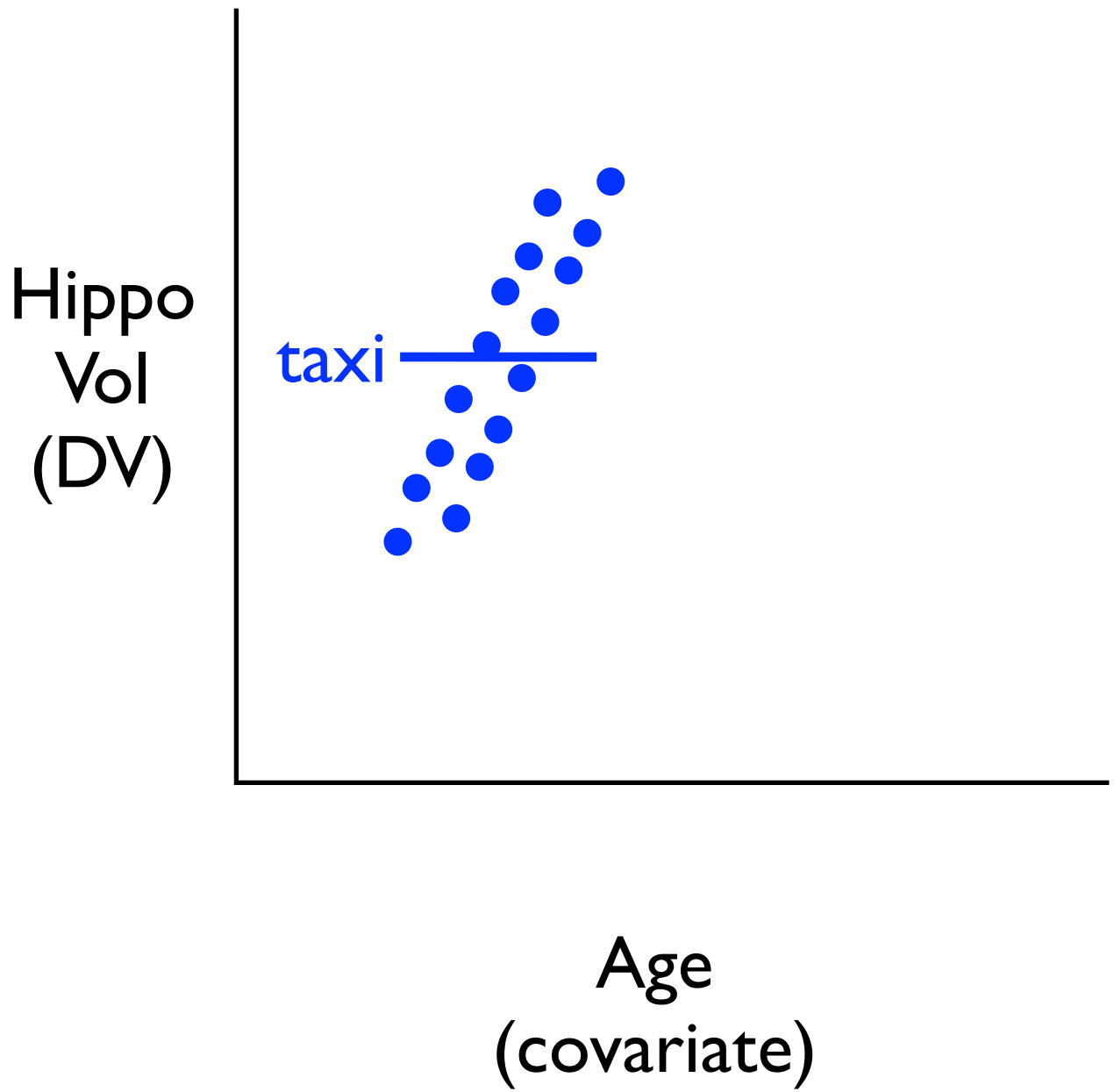
Raw means: Y2 > Y1



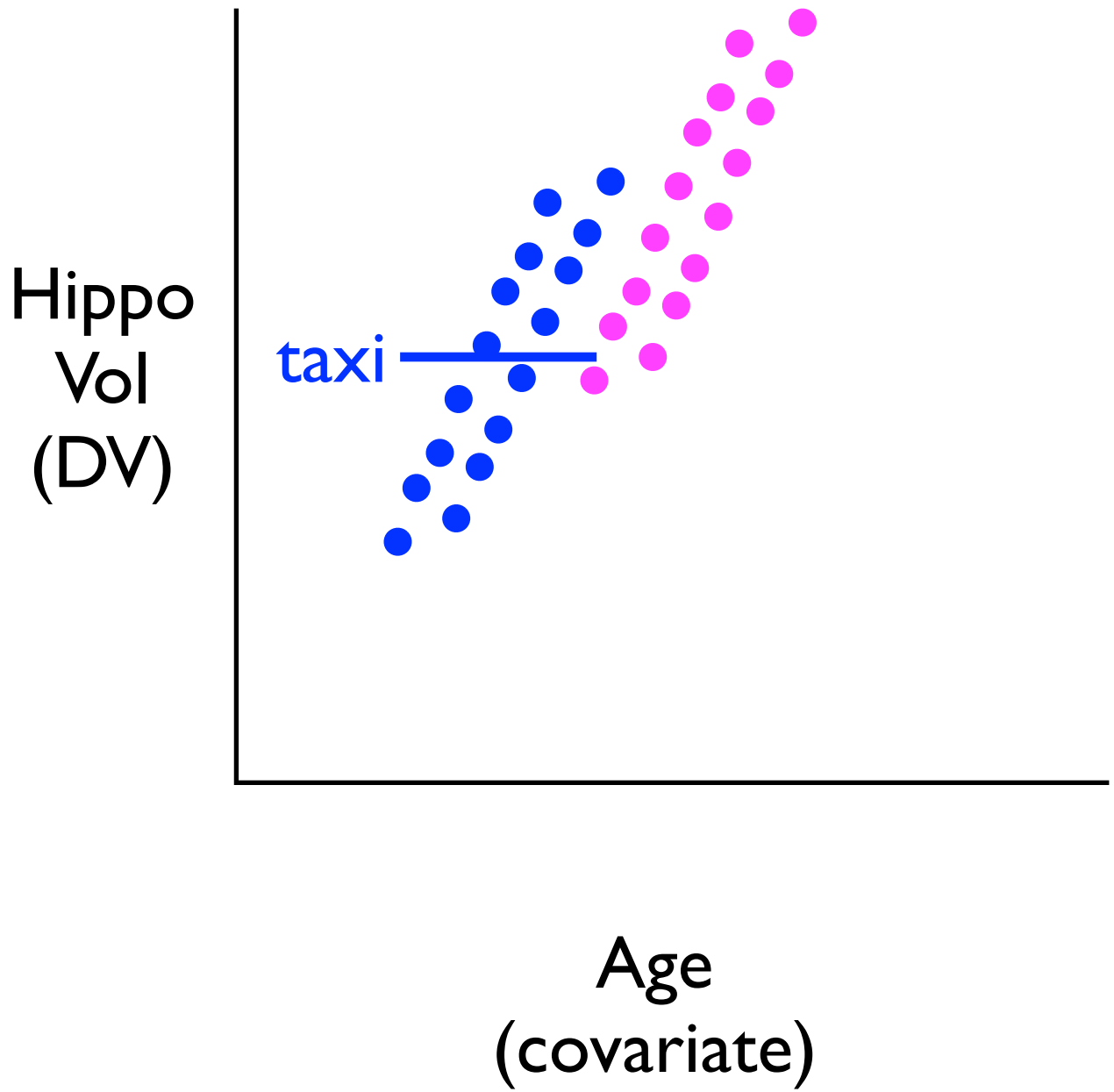
Raw means: Y2 > Y1



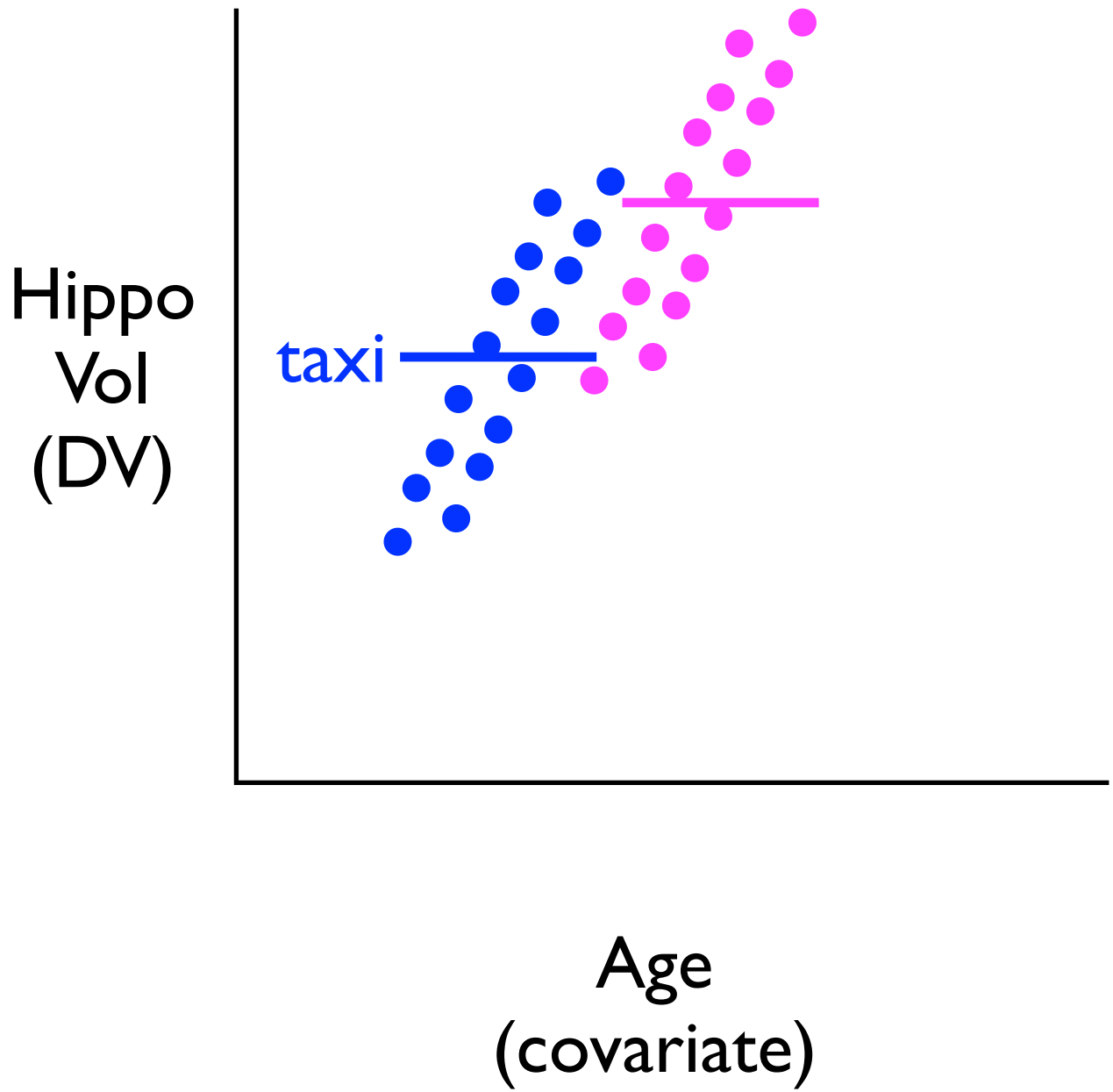
Raw means: Y2 > Y1



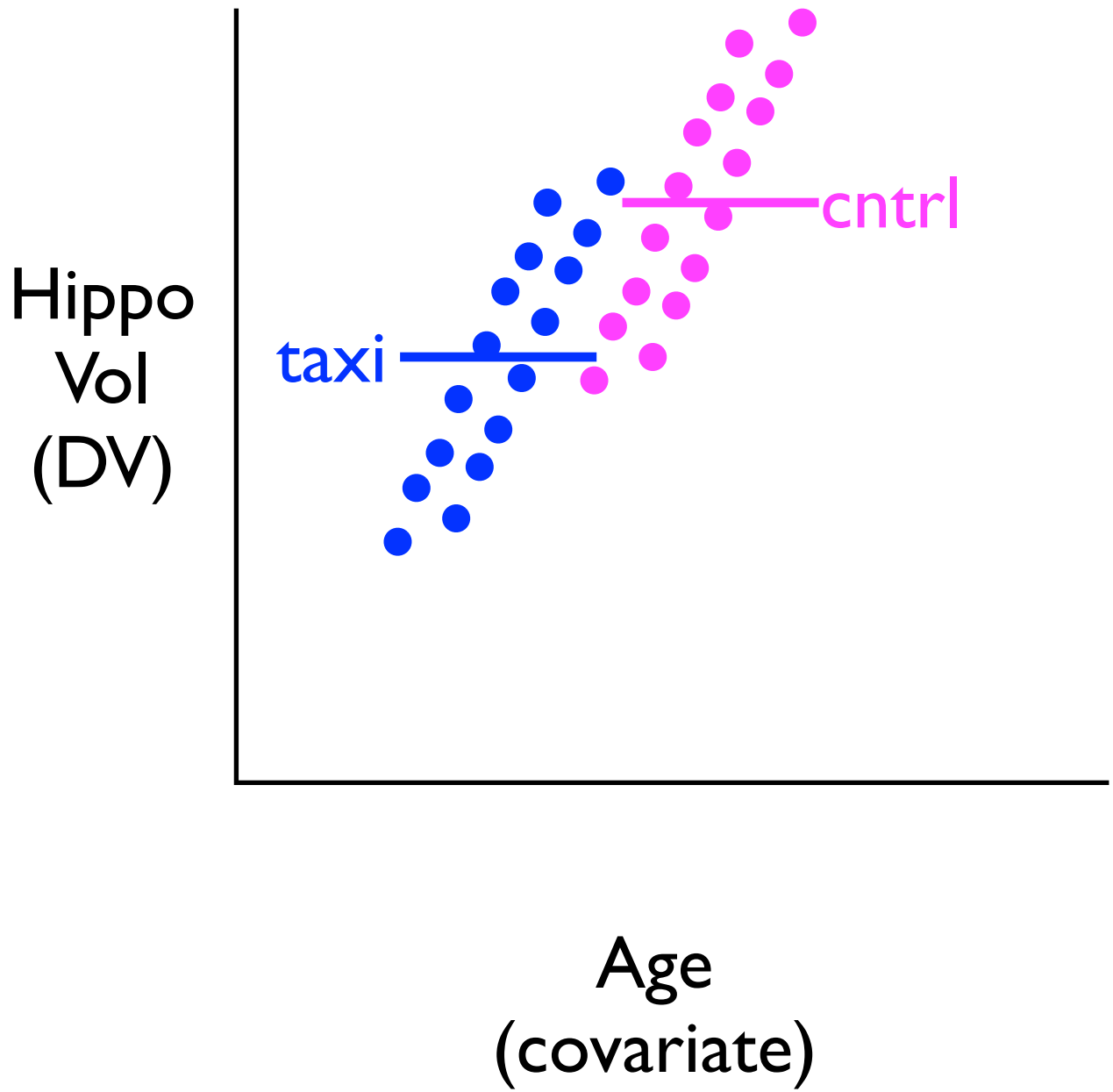
Raw means: Y2 > Y1



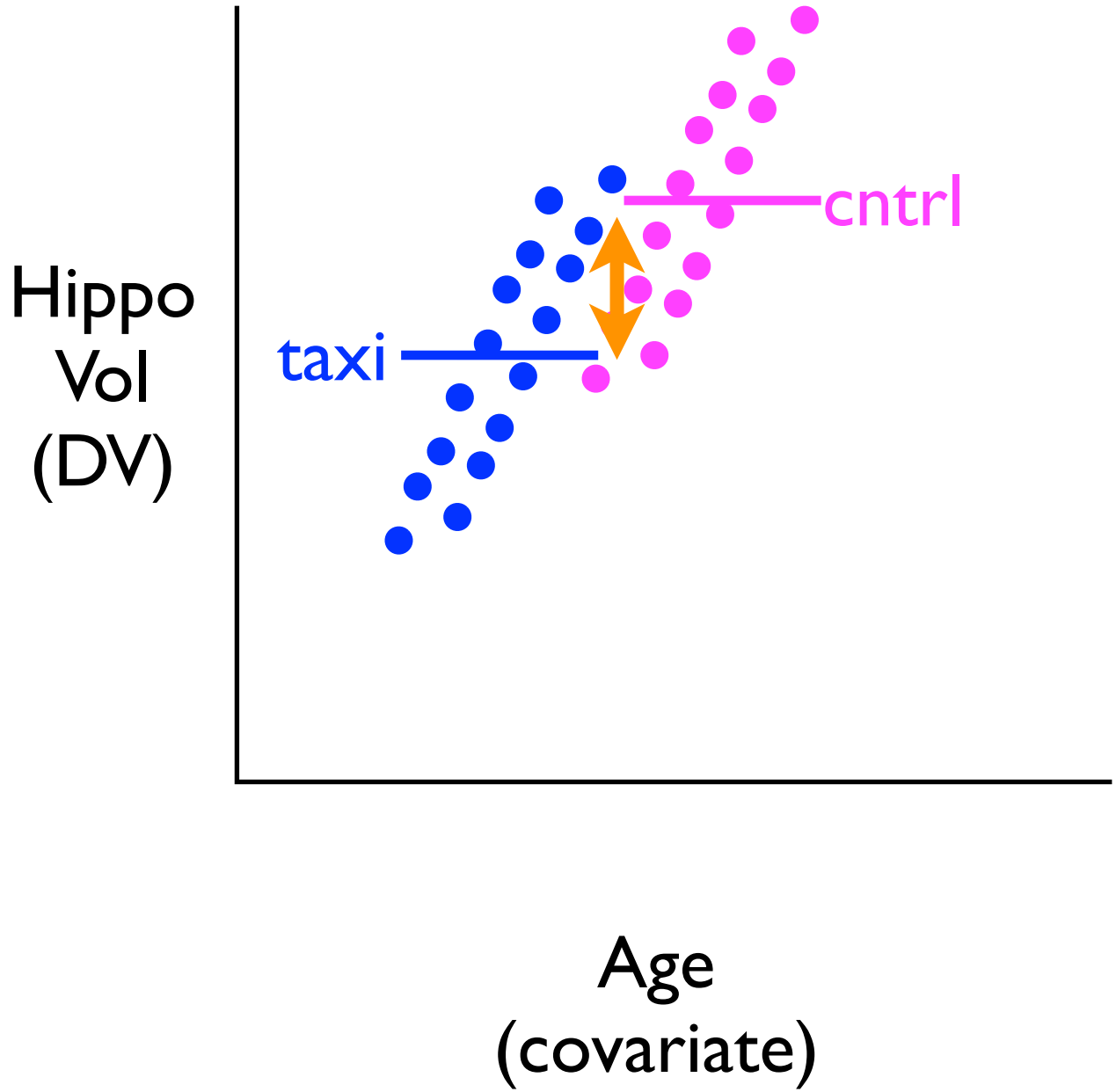
Raw means: Y2 > Y1



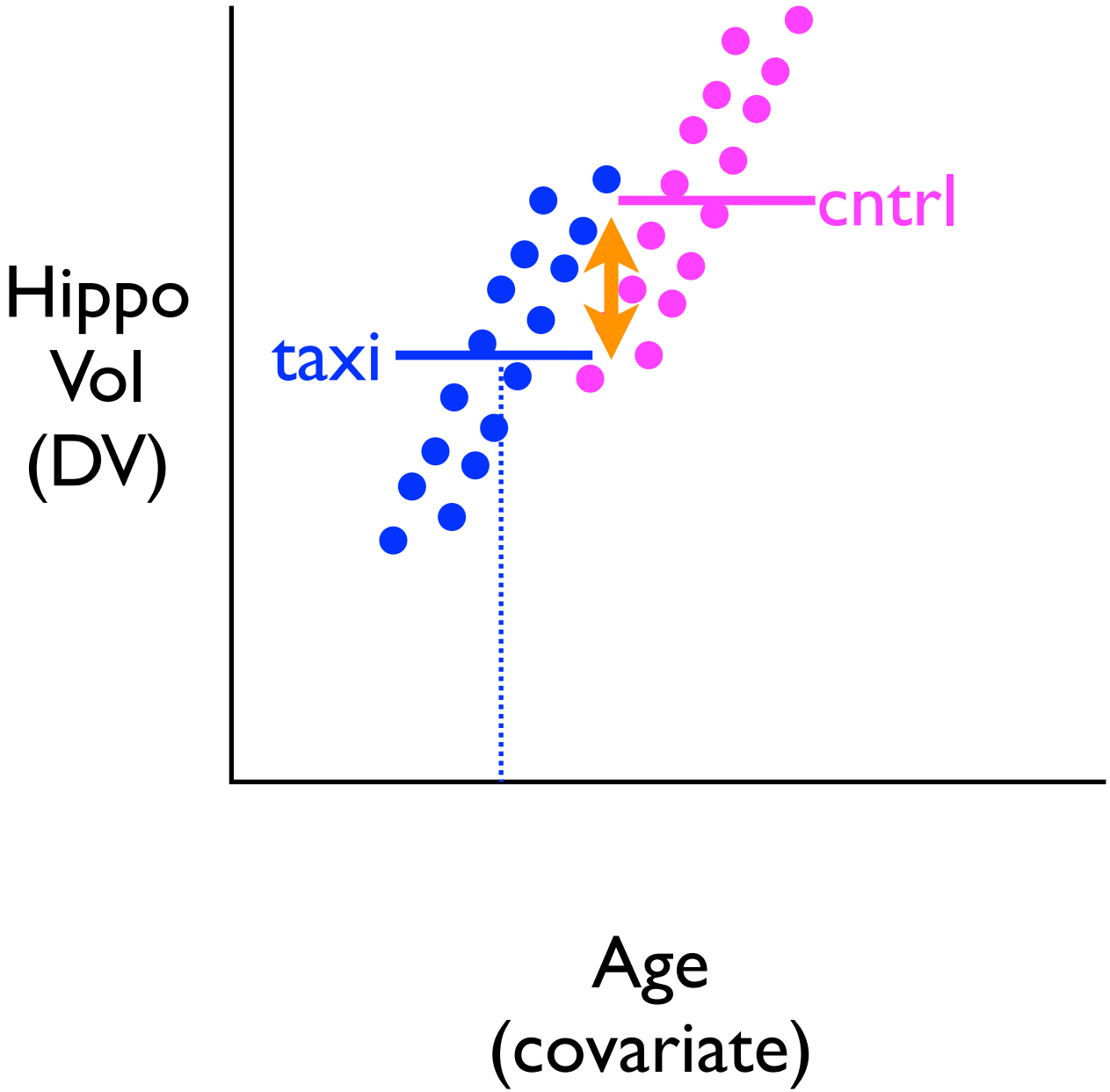
Raw means: Y2 > Y1



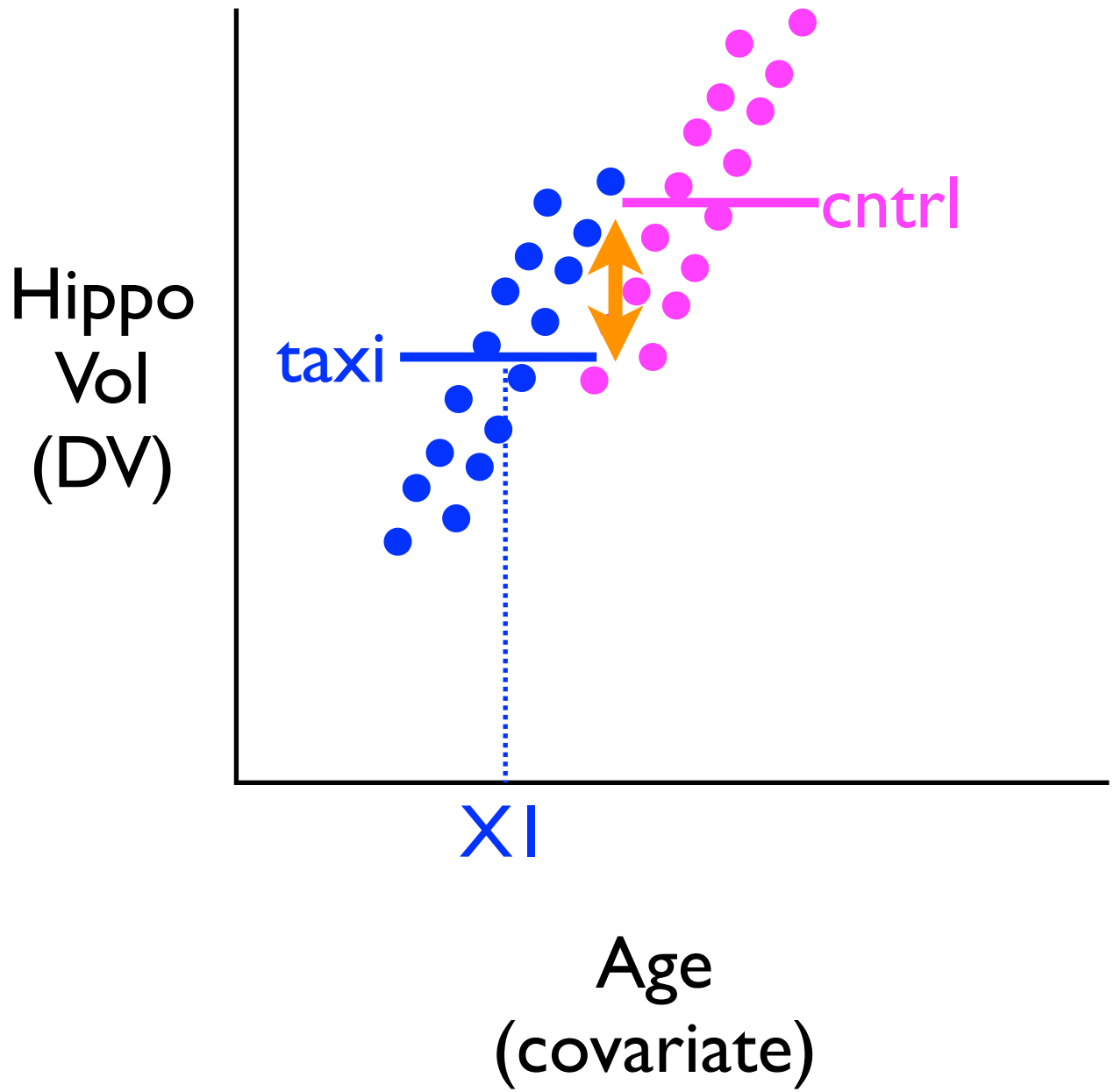
Raw means: Y2 > Y1



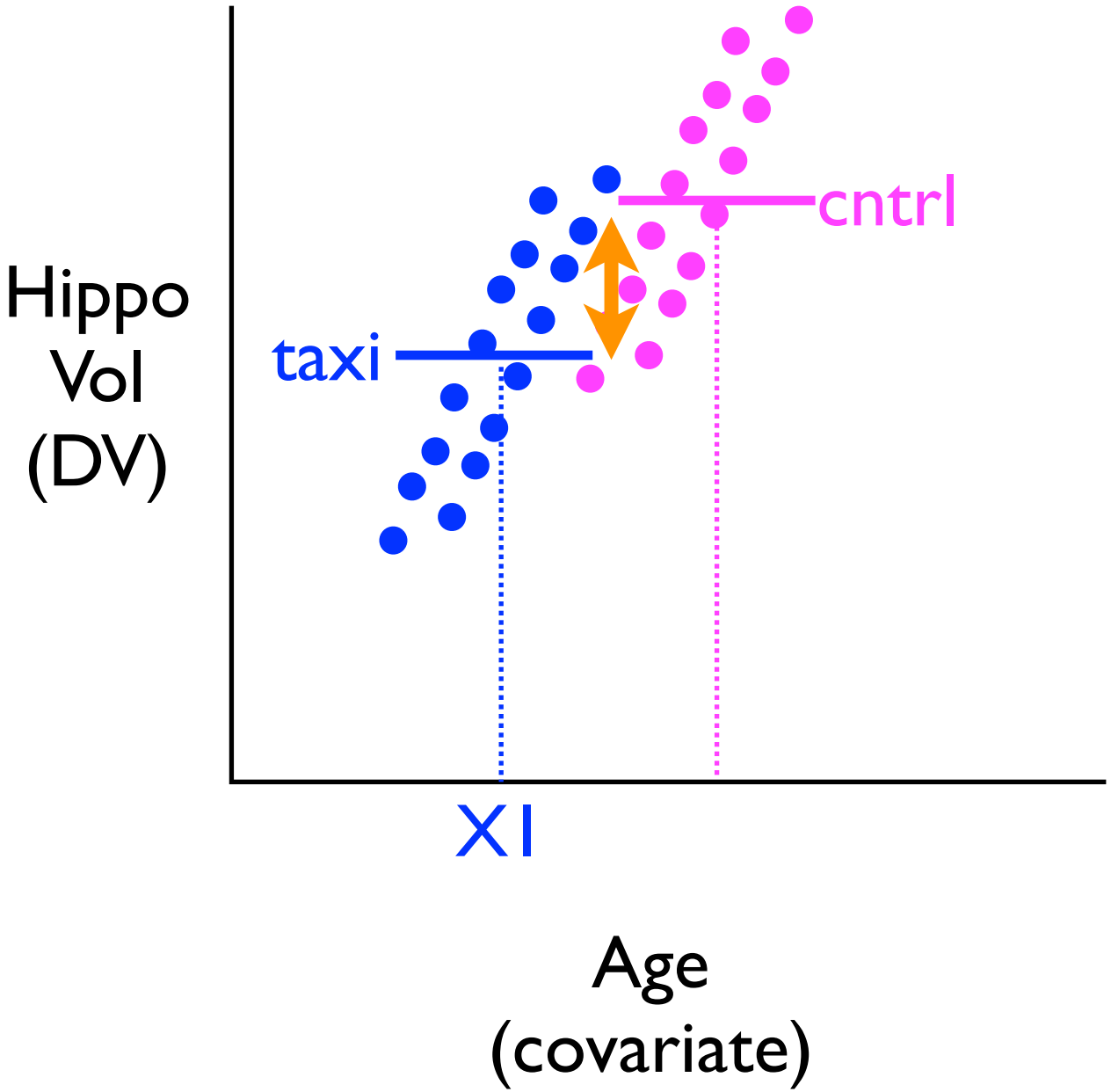
Raw means: Y2 > Y1



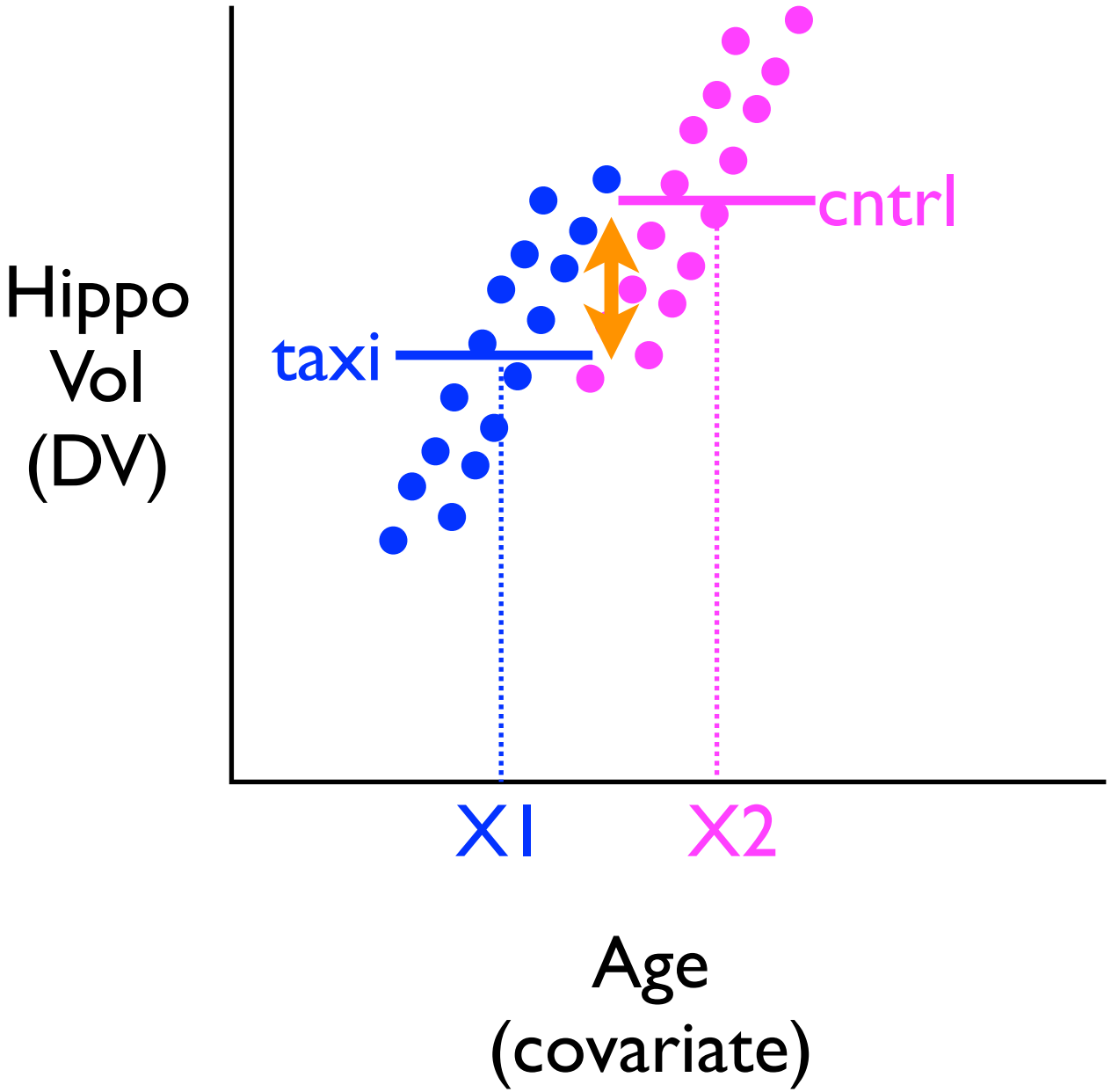
Raw means: Y2 > Y1



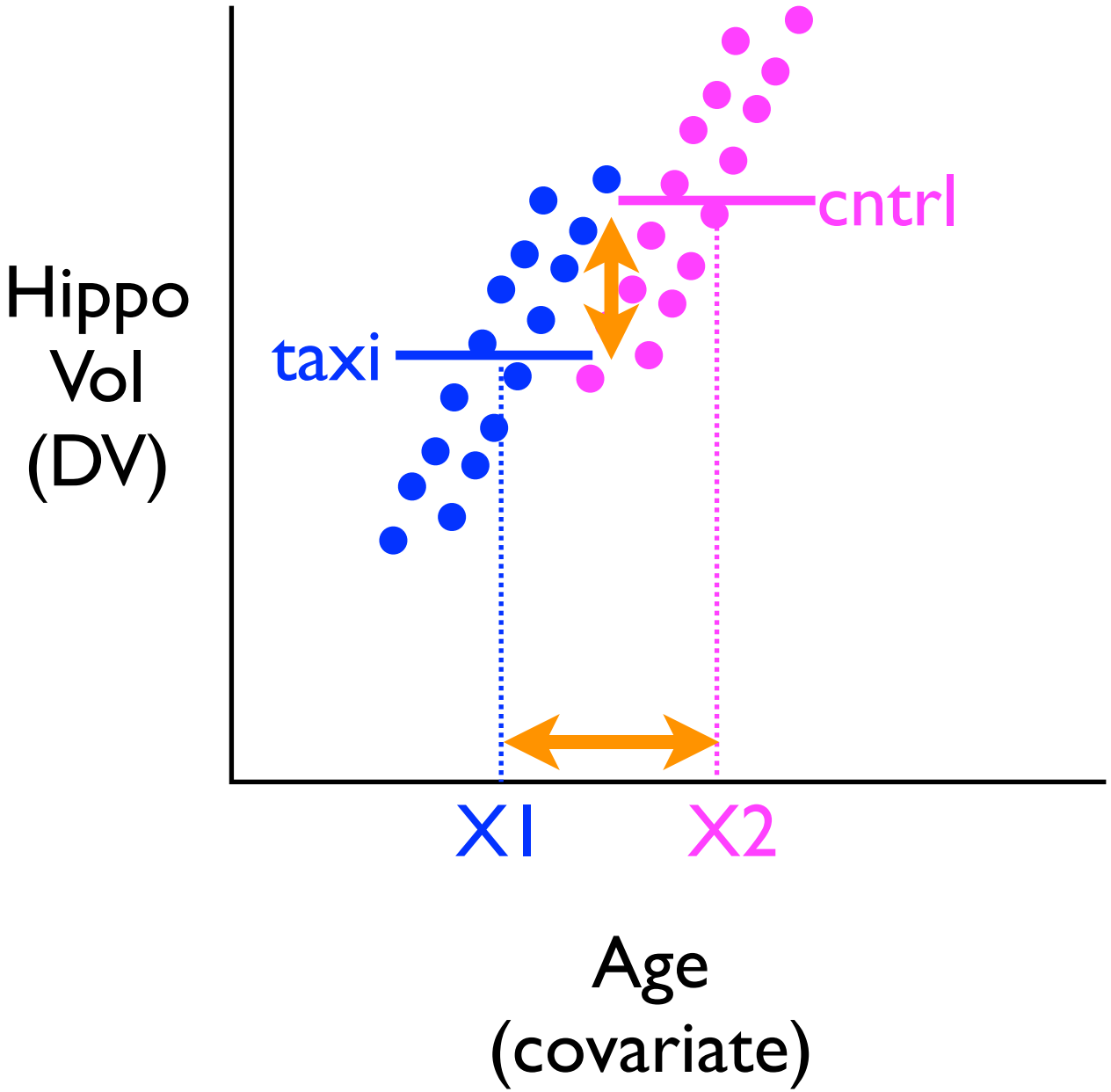
Raw means: Y2 > Y1



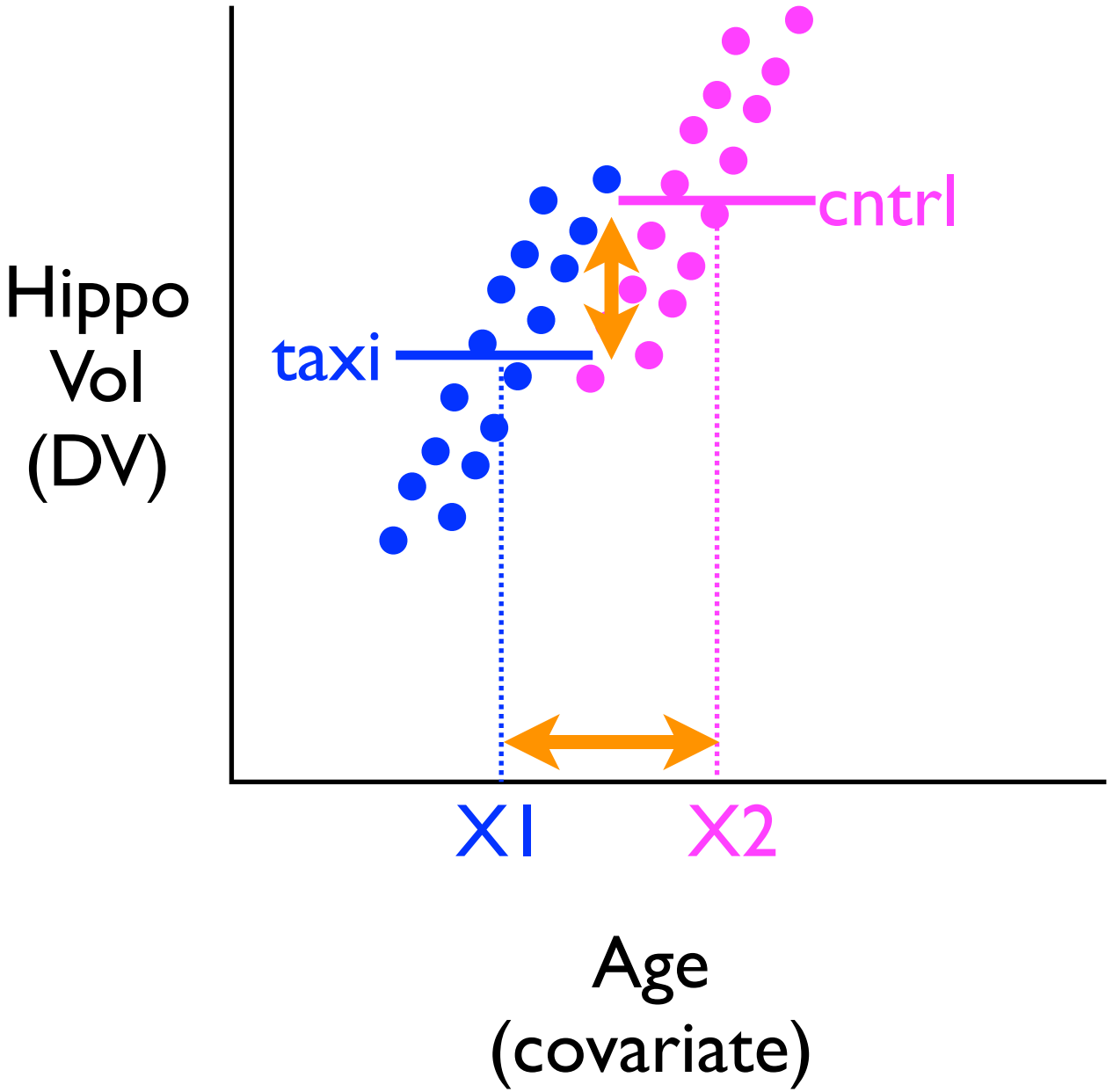
Raw means: Y2 > Y1



Raw means: Y2 > Y1

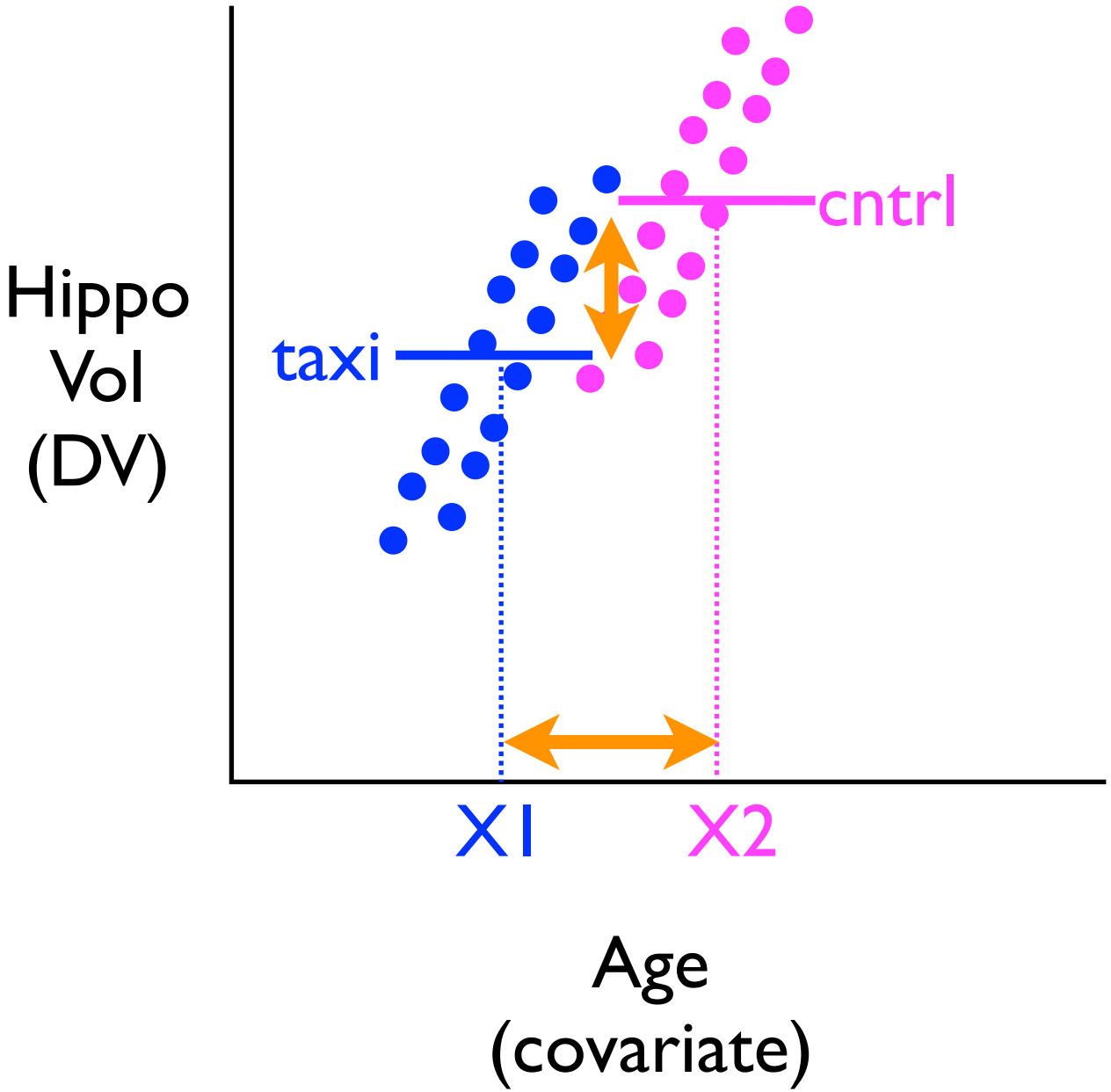


Raw means: Y2 > Y1



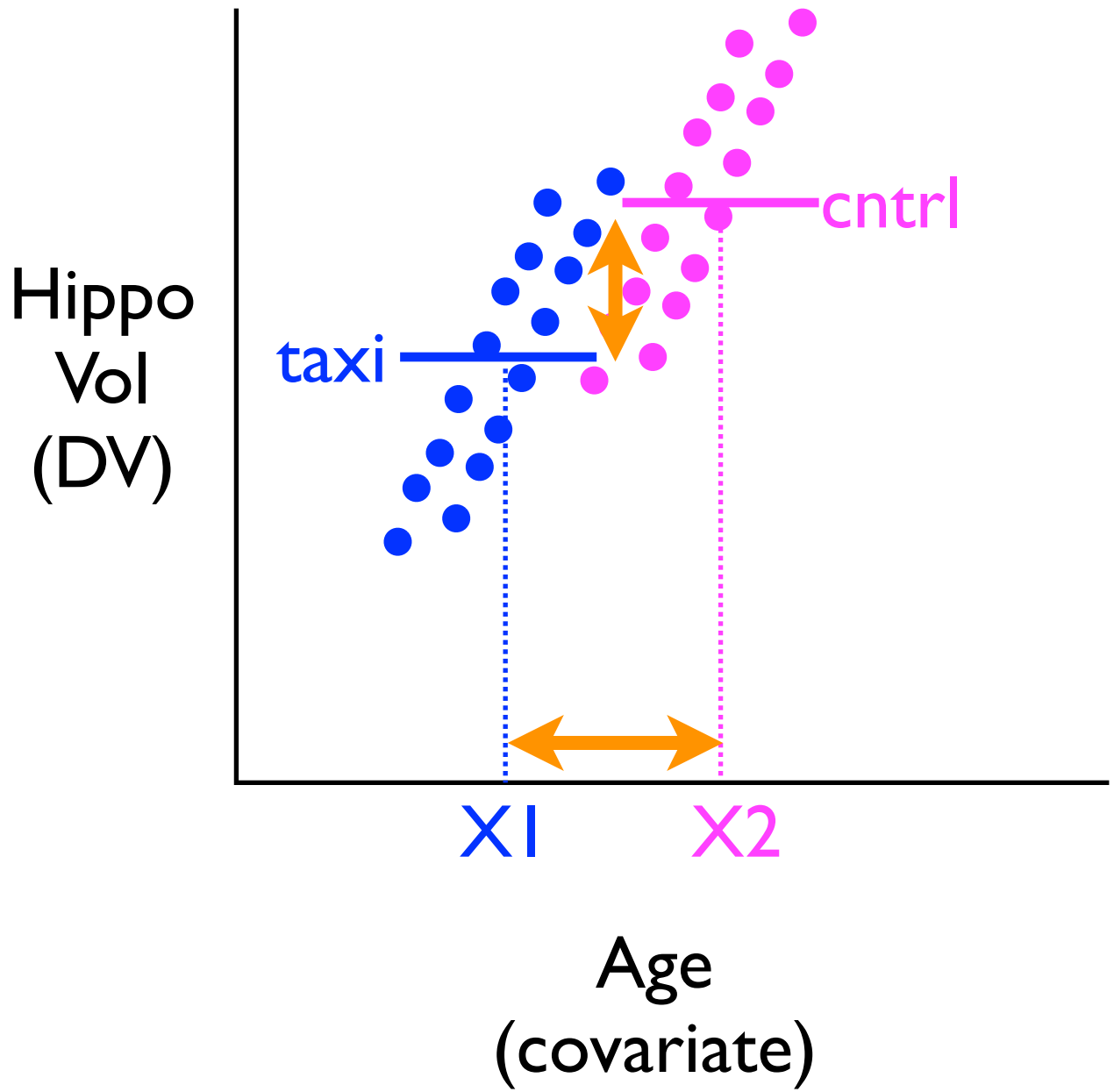
- HippoVol of group 1 (taxiDrvrs) is less than group 2 (controls)

Raw means: Y2 > Y1



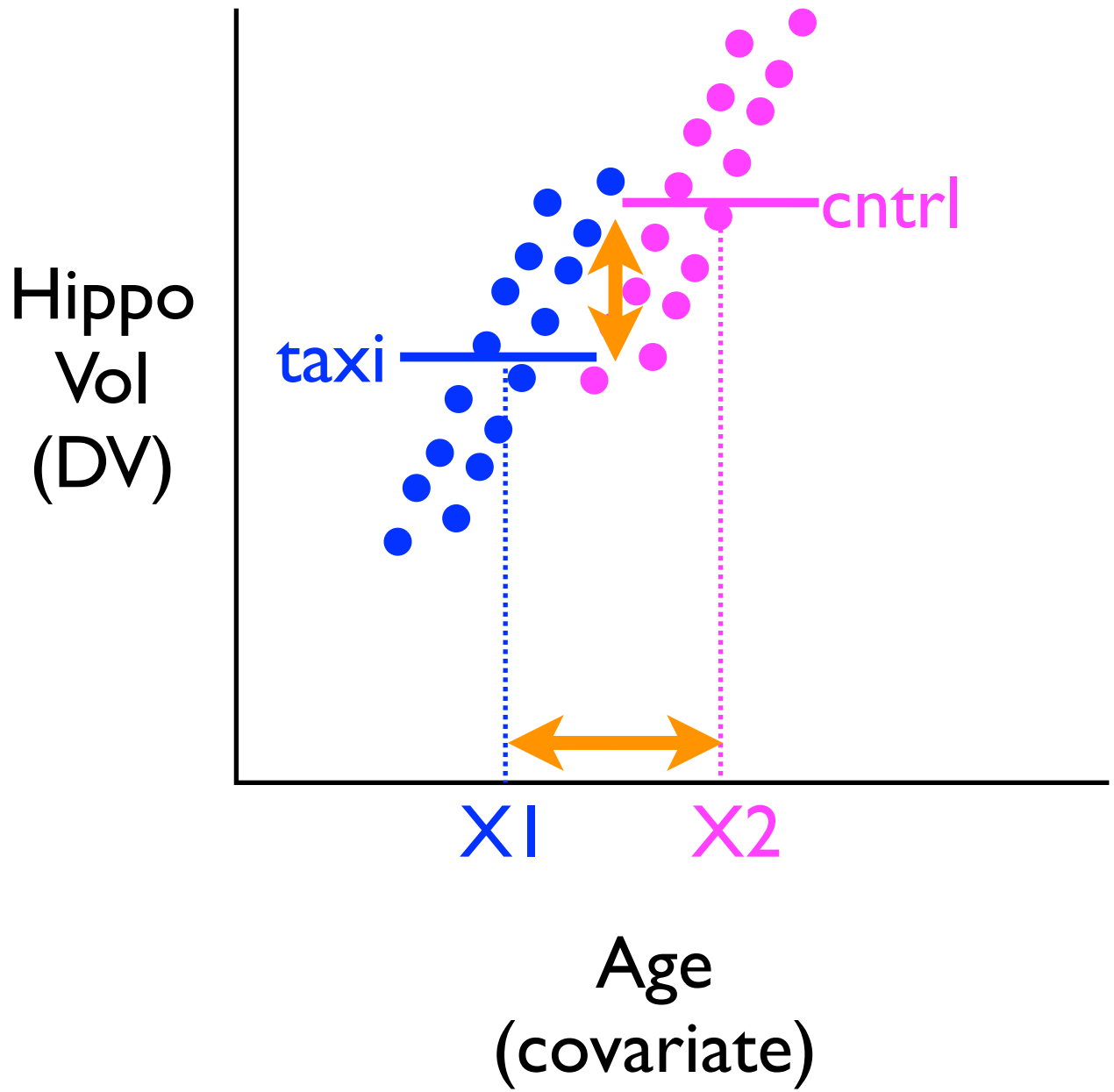
- HippoVol of group 1 (taxiDrvrs) is less than group 2 (controls)
- oops!

Raw means: Y2 > Y1



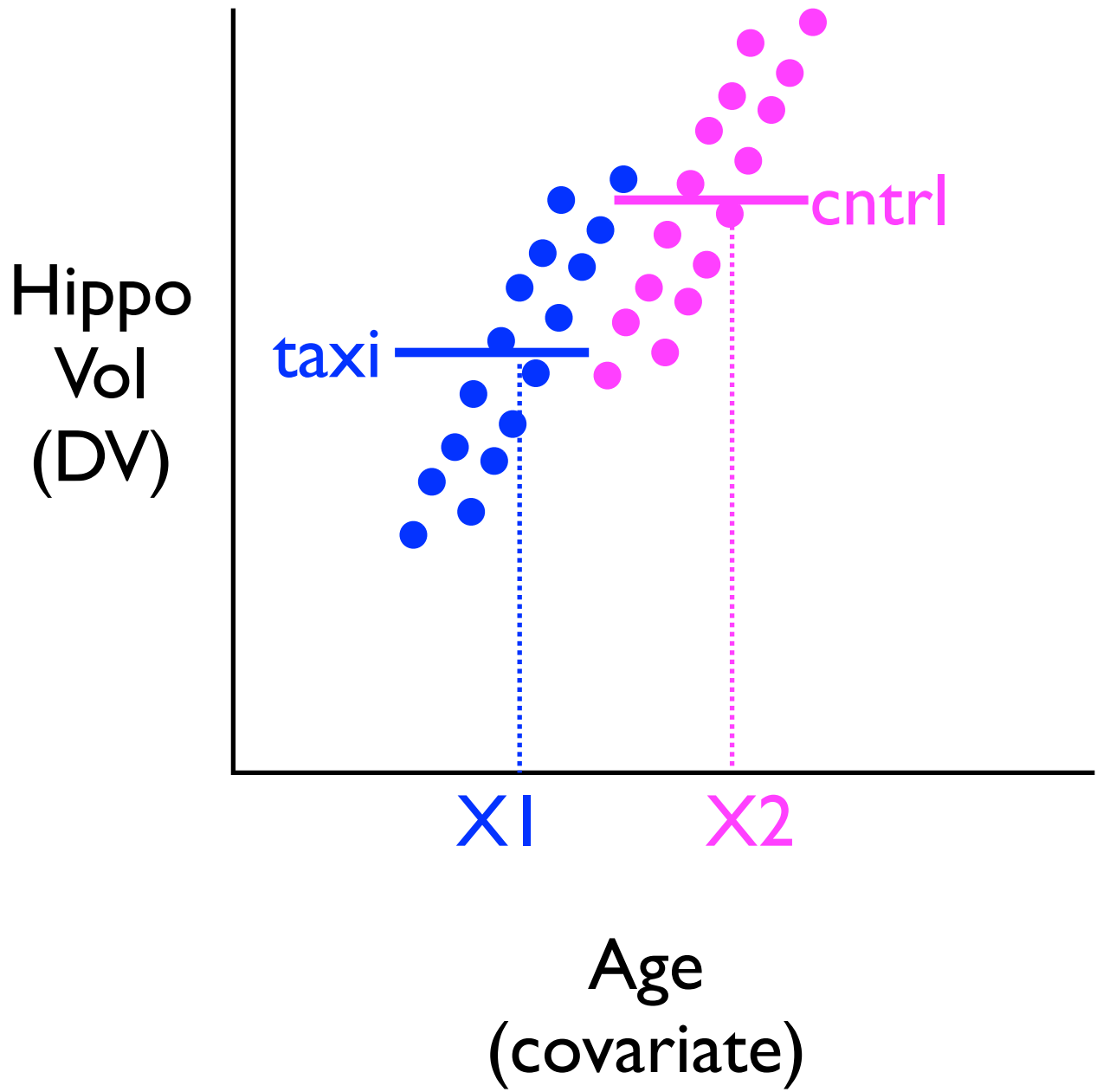
- HippoVol of group 1 (taxiDrvrs) is less than group 2 (controls)
- oops!
- but it happens that our controls tended to have higher age

Raw means: Y2 > Y1



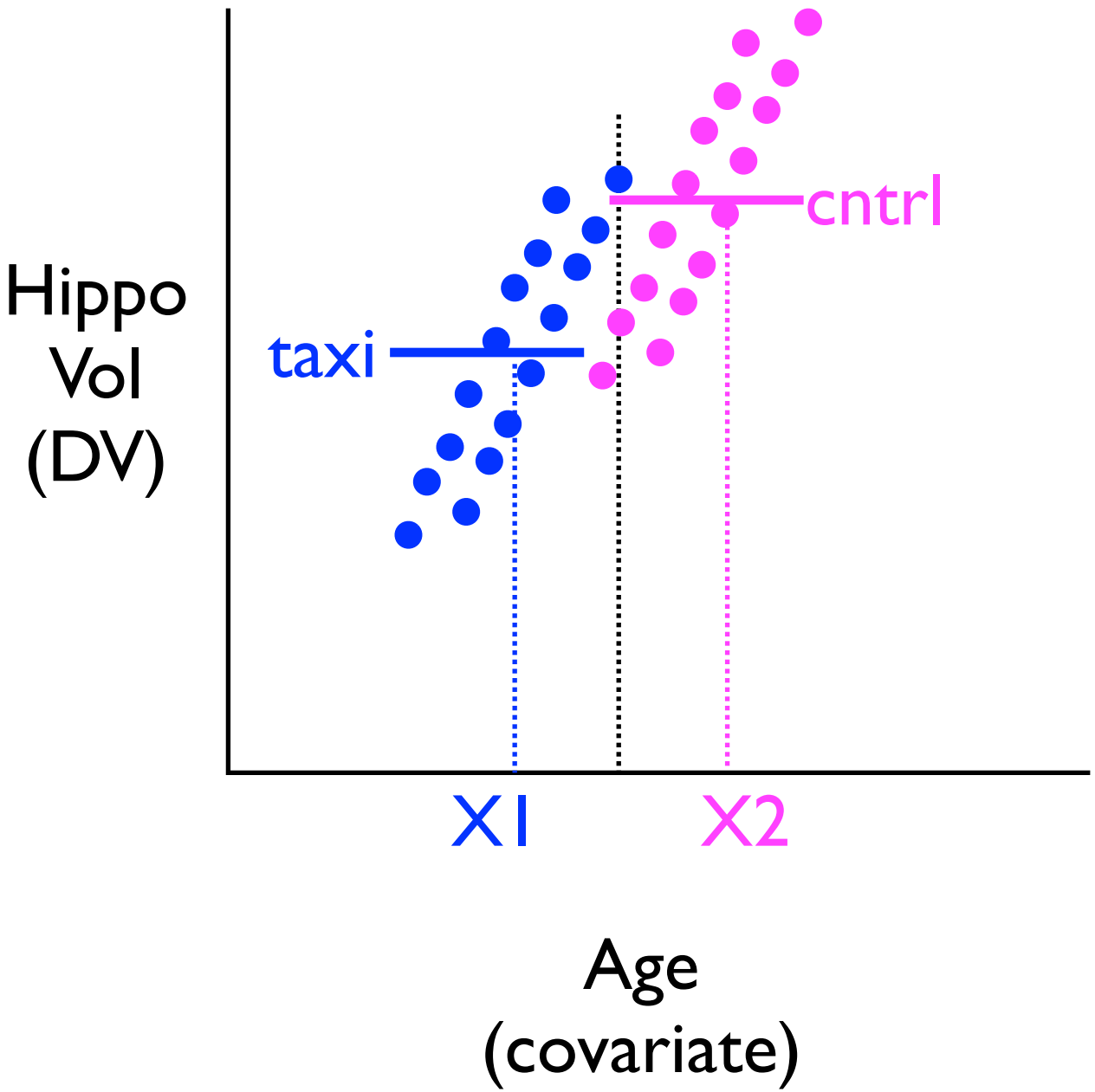
- HippoVol of group 1 (taxiDrvrs) is less than group 2 (controls)
- oops!
- but it happens that our controls tended to have higher age
- what if we "statistically" control for these pre-existing age differences?

Adjusted means: Y2 < Y1



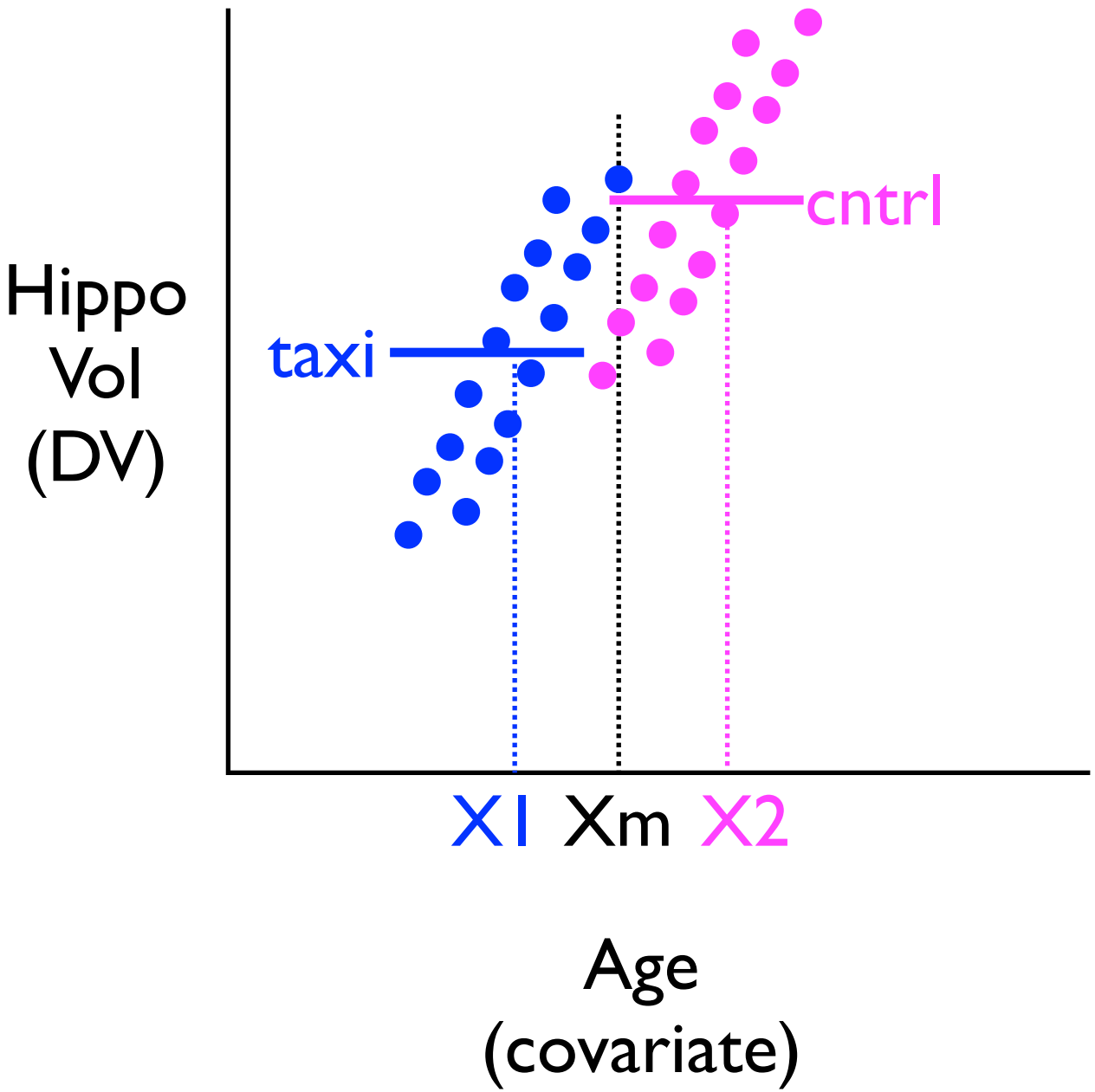
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



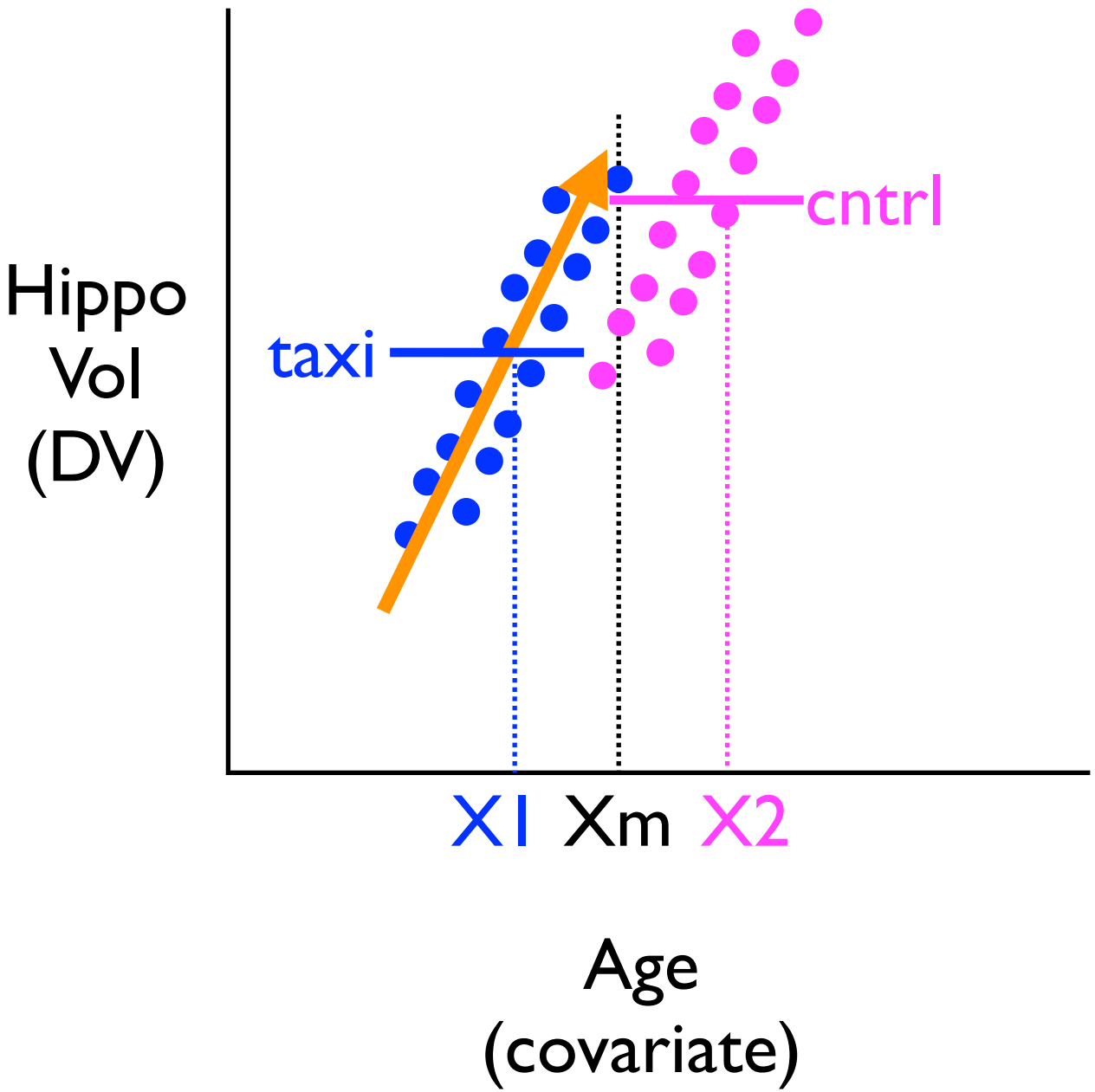
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



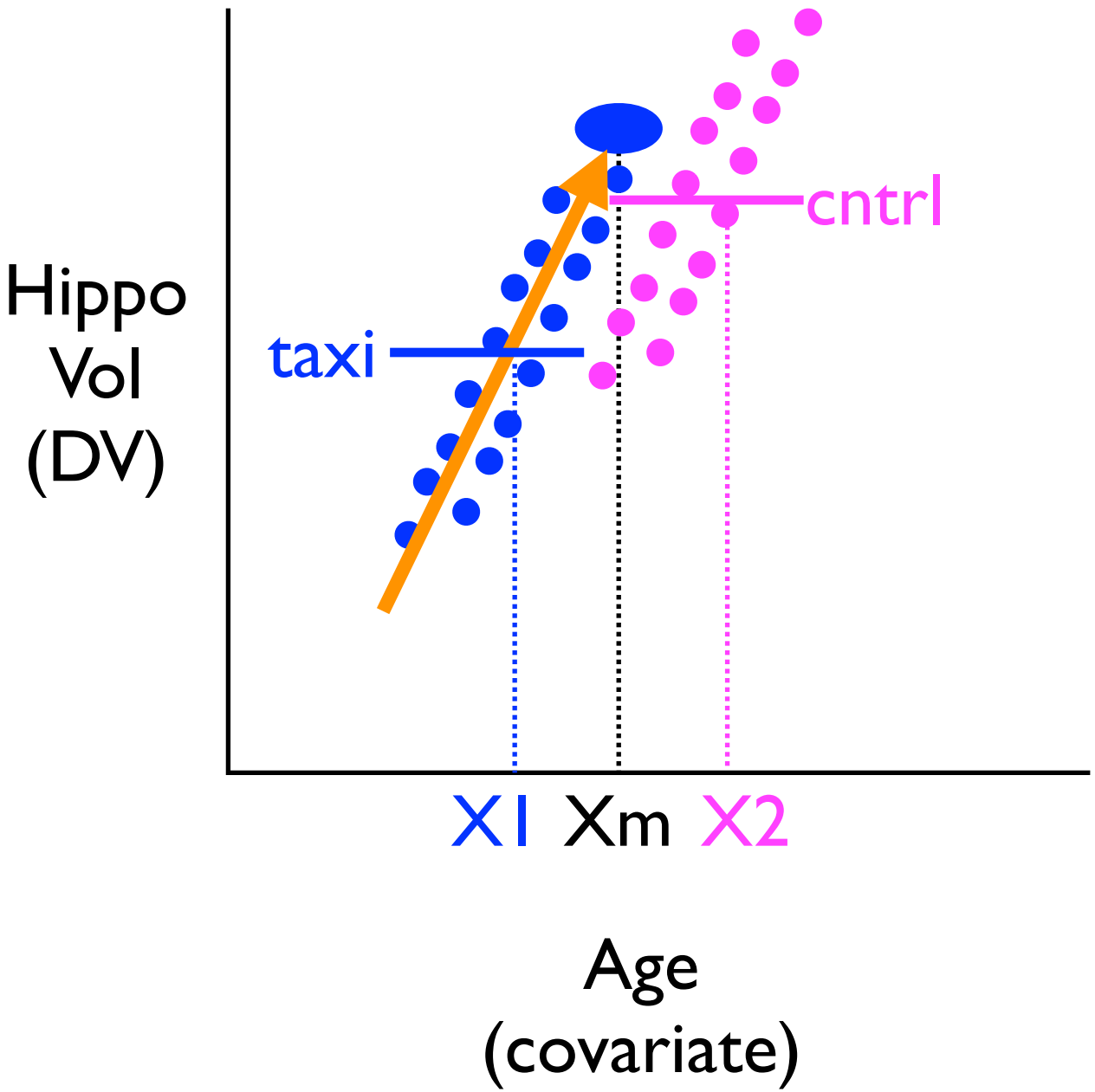
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



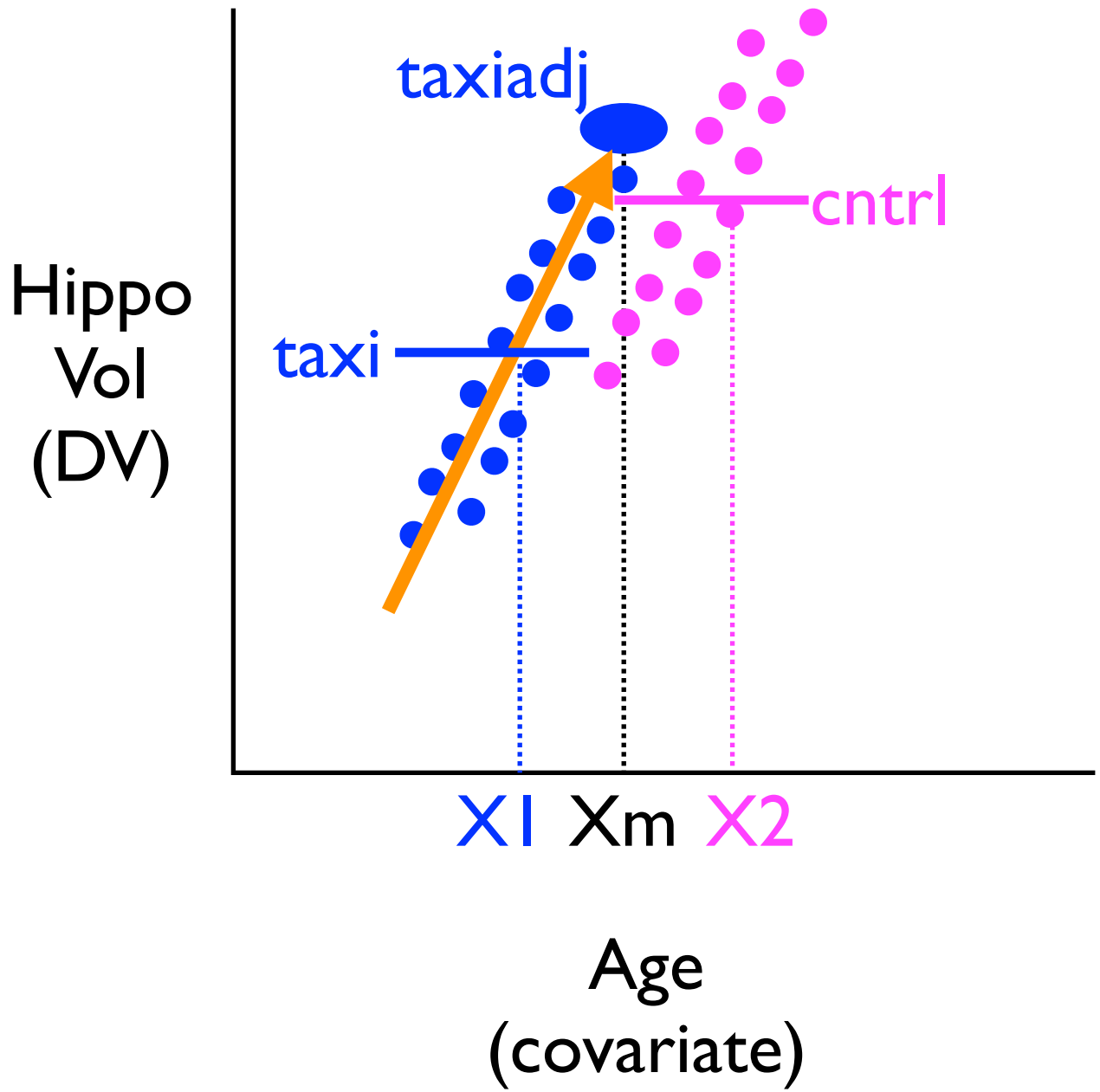
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



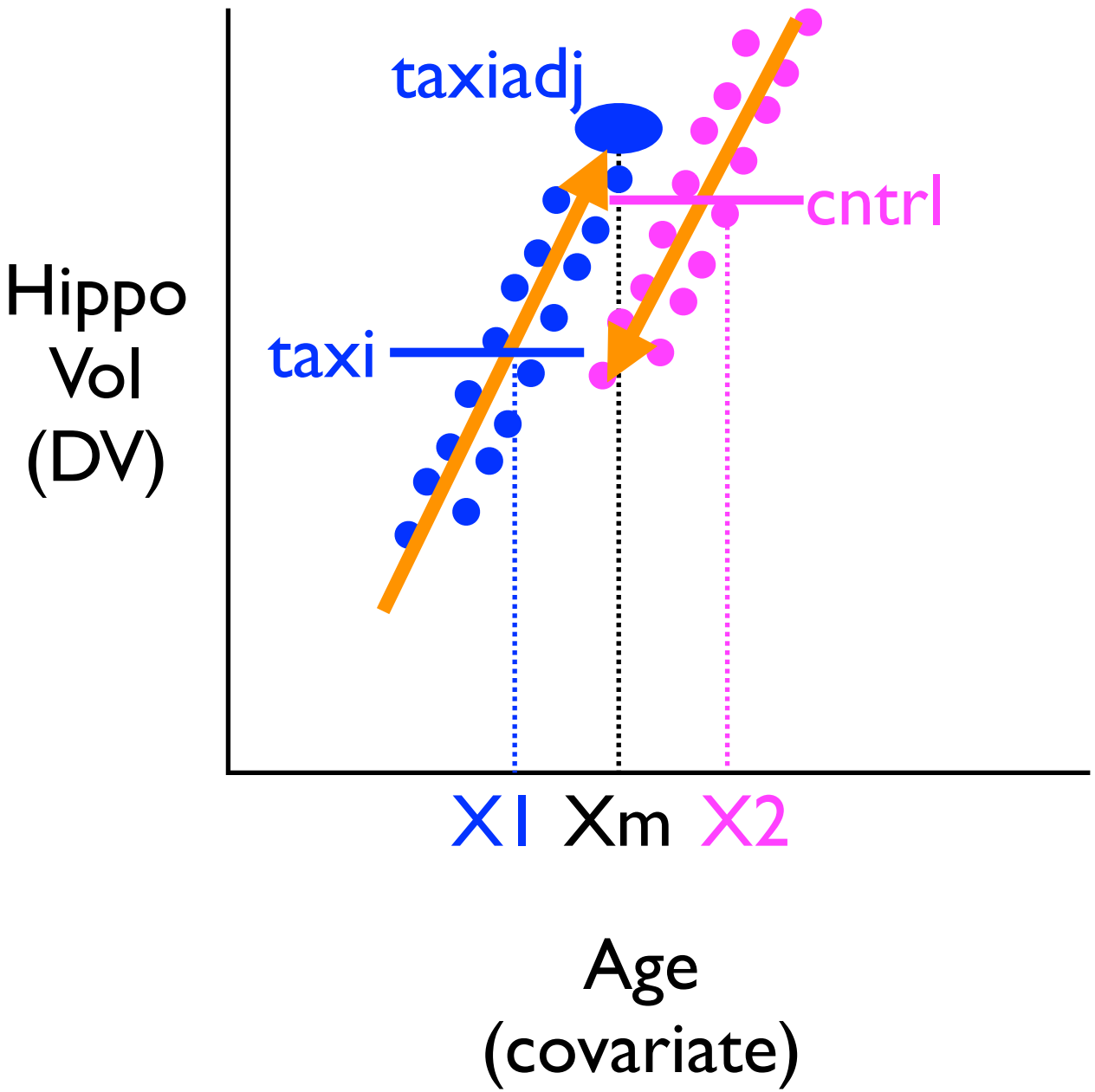
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



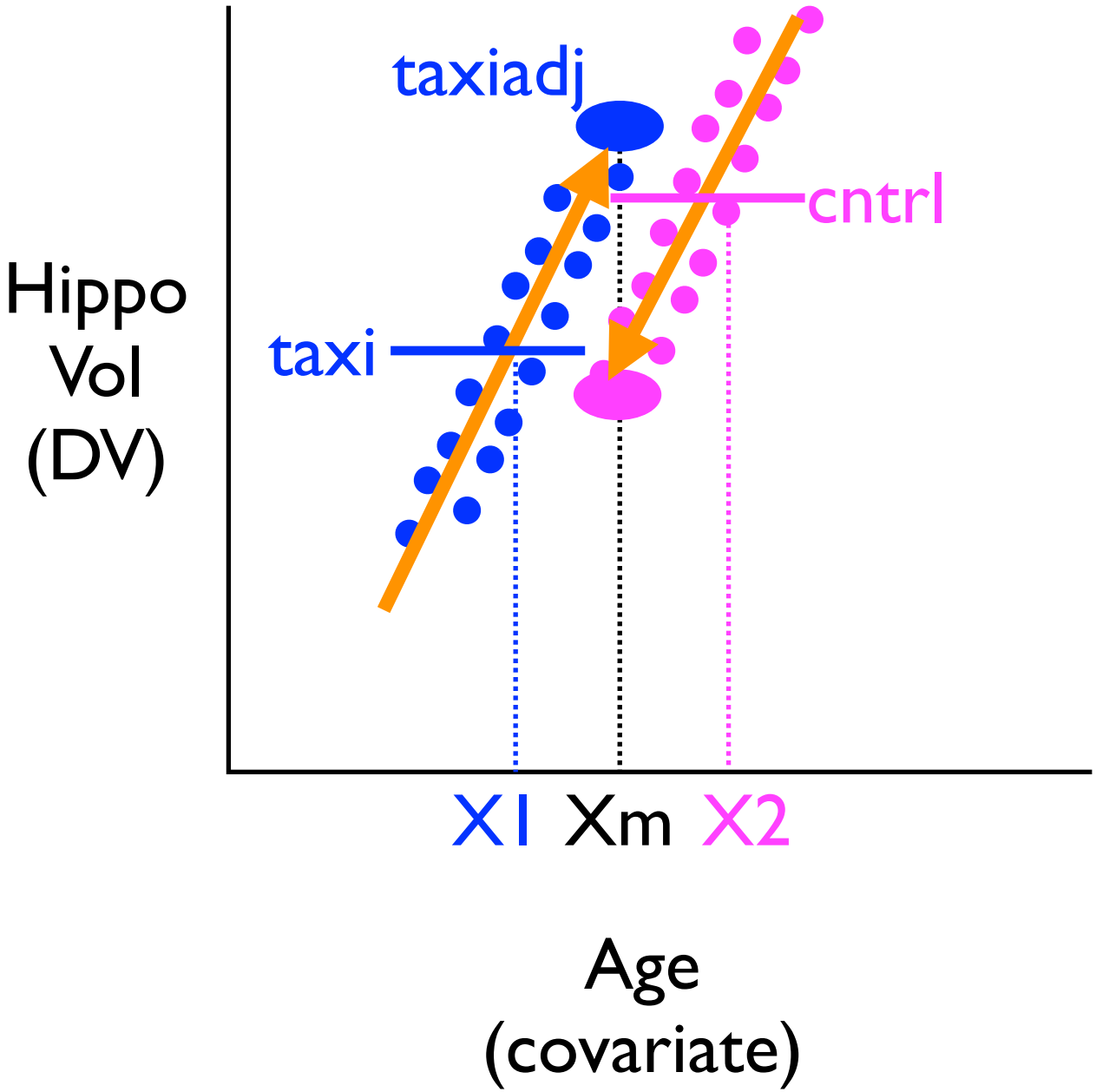
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



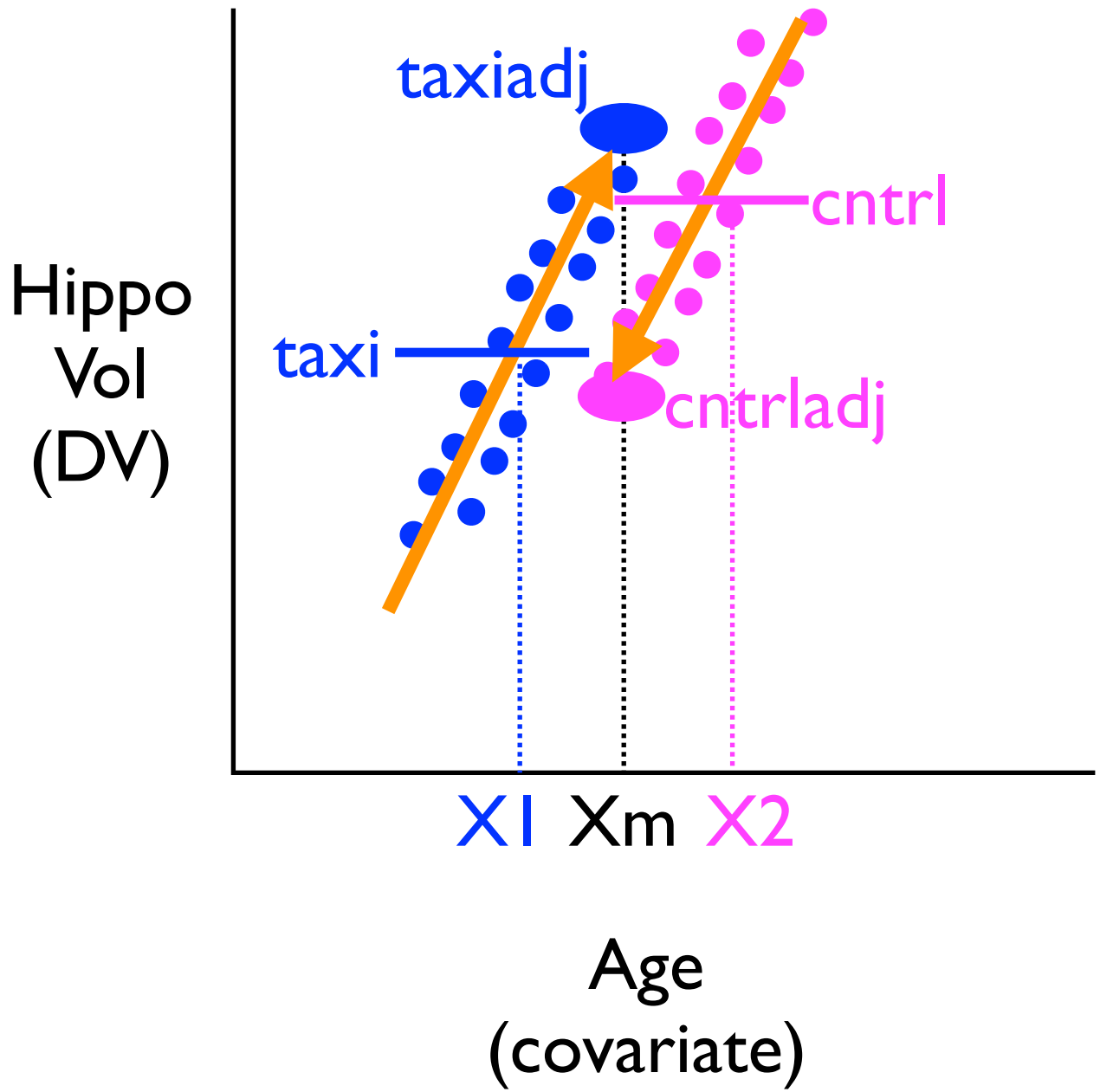
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1



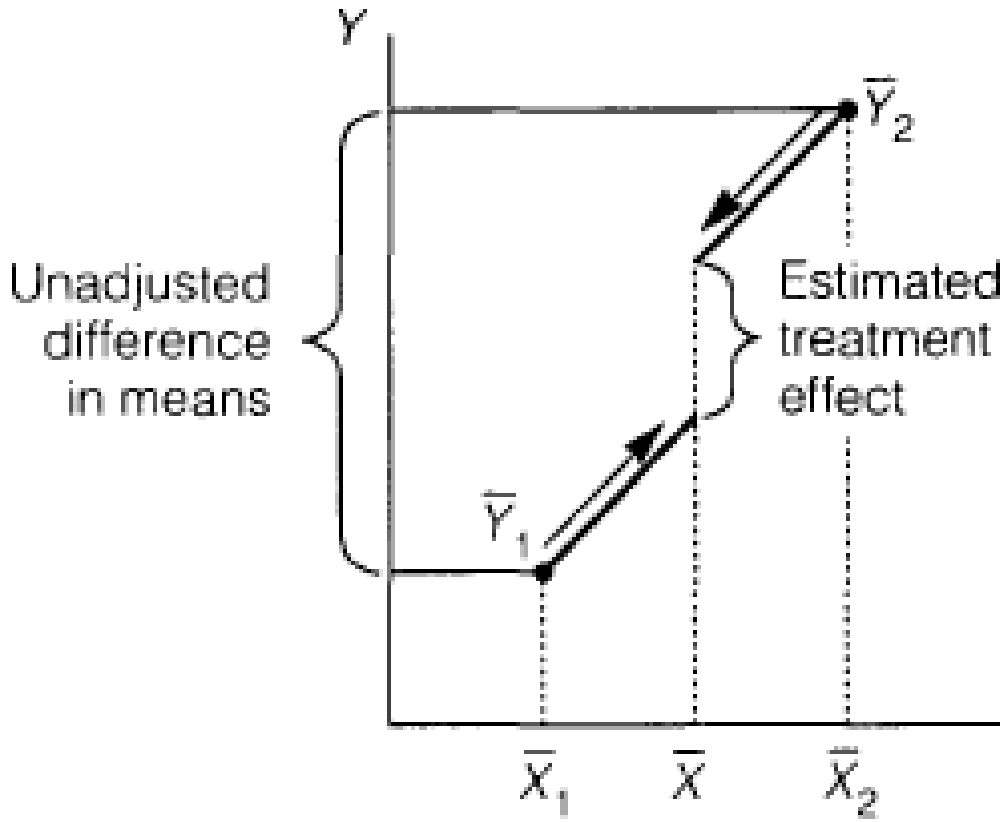
- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

Adjusted means: Y2 < Y1

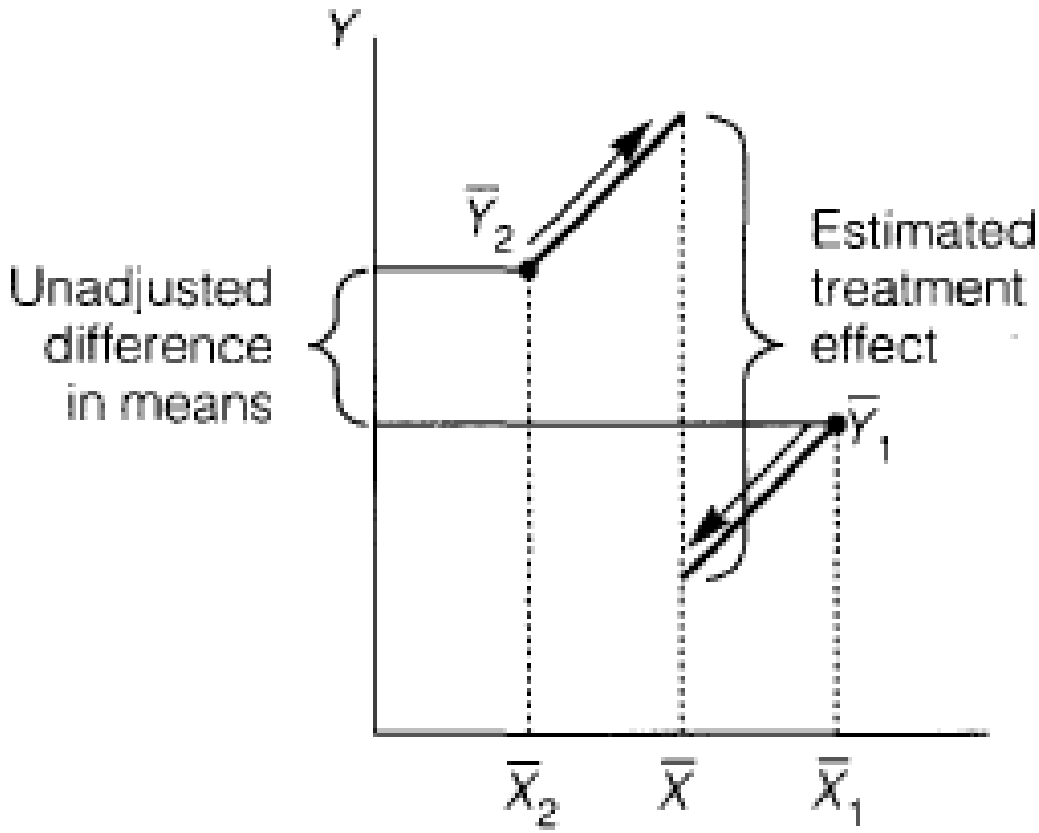


- let's use the known relationship between age (X) and HippoVol (Y) to predict,
- what would HippoVol (Y) have been,
- IF age of both groups were equal?
- IF they were equal to (for example) the grand mean of X (age)
- now we see that on the adjusted means, HippoVol for taxi drivers is > than controls

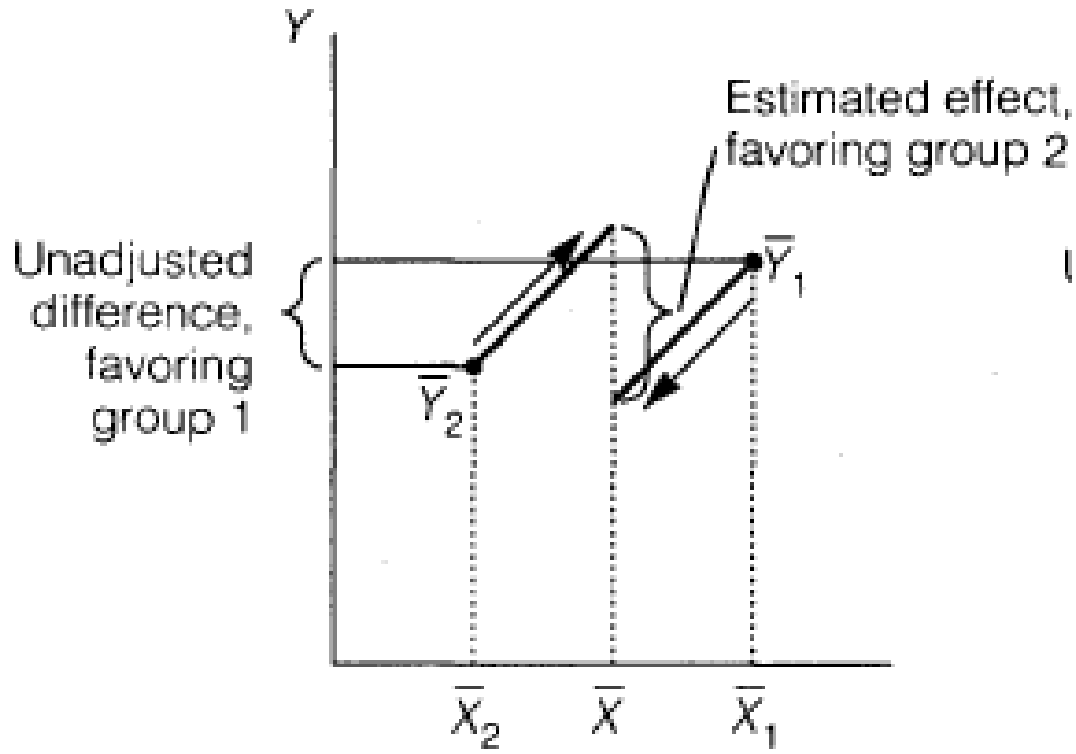
estimate of treatment effect **decreased** by adjusting for pre-existing differences



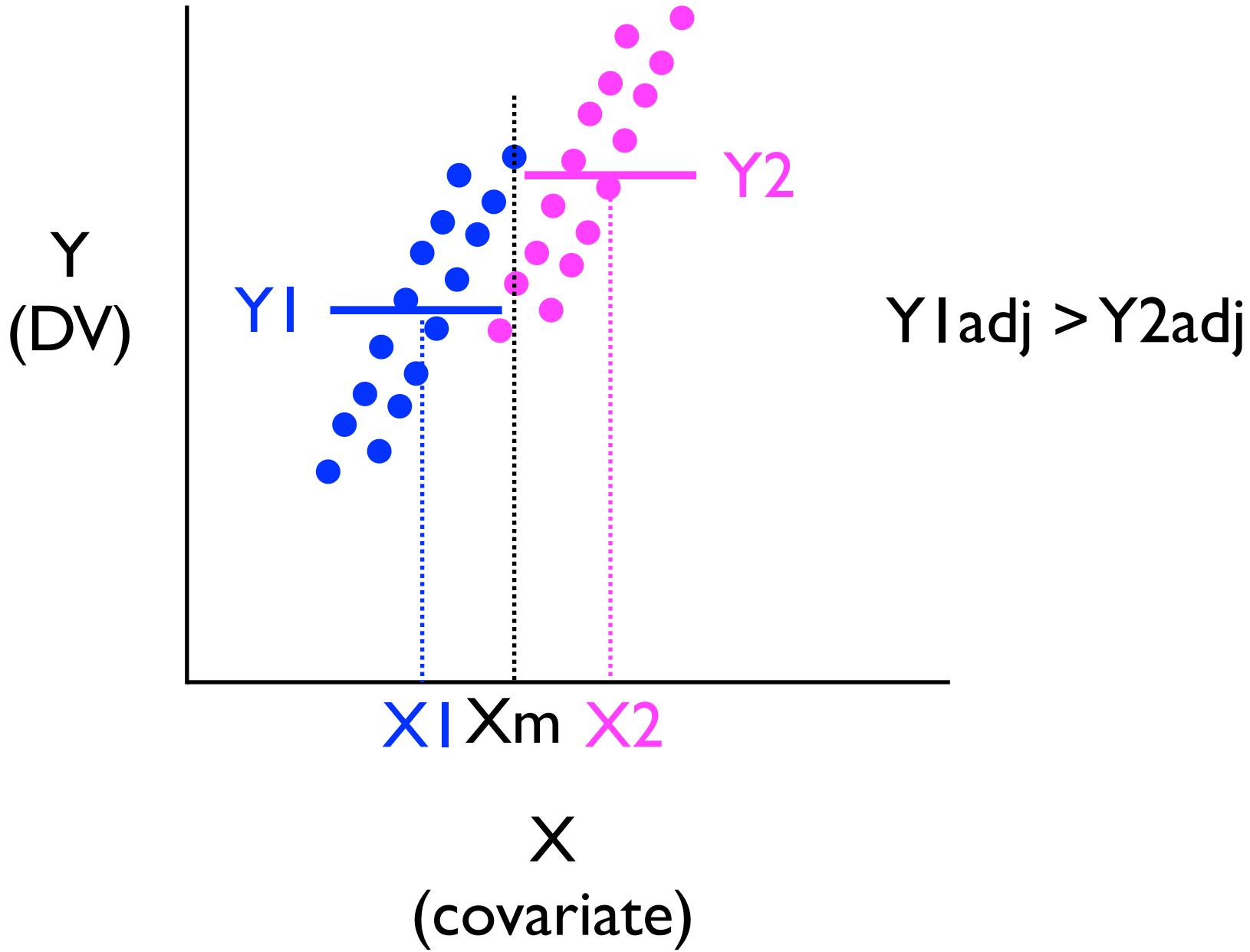
estimate of treatment effect **increased** by adjusting for pre-existing differences



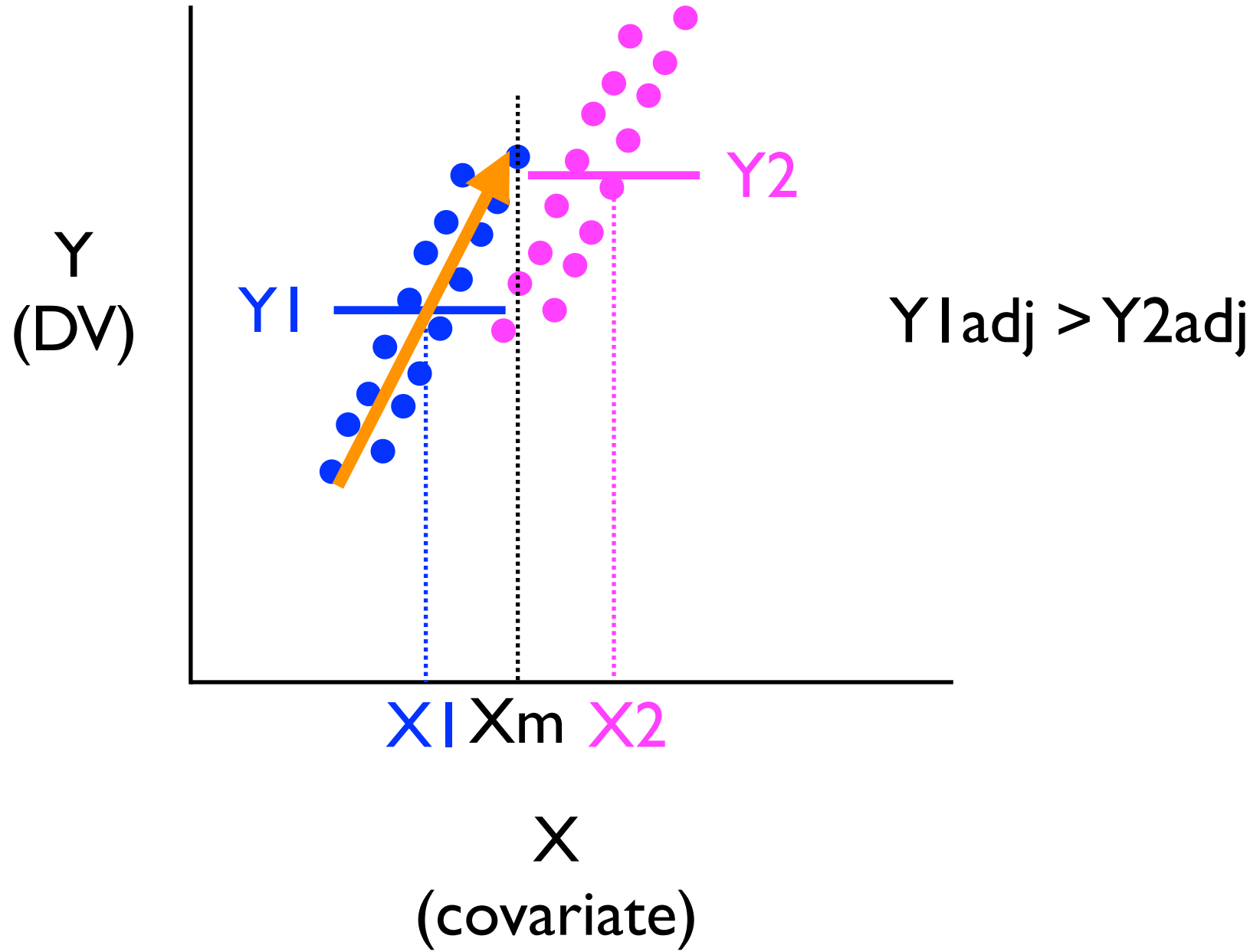
estimate of treatment effect **reversed** by adjusting for pre-existing differences



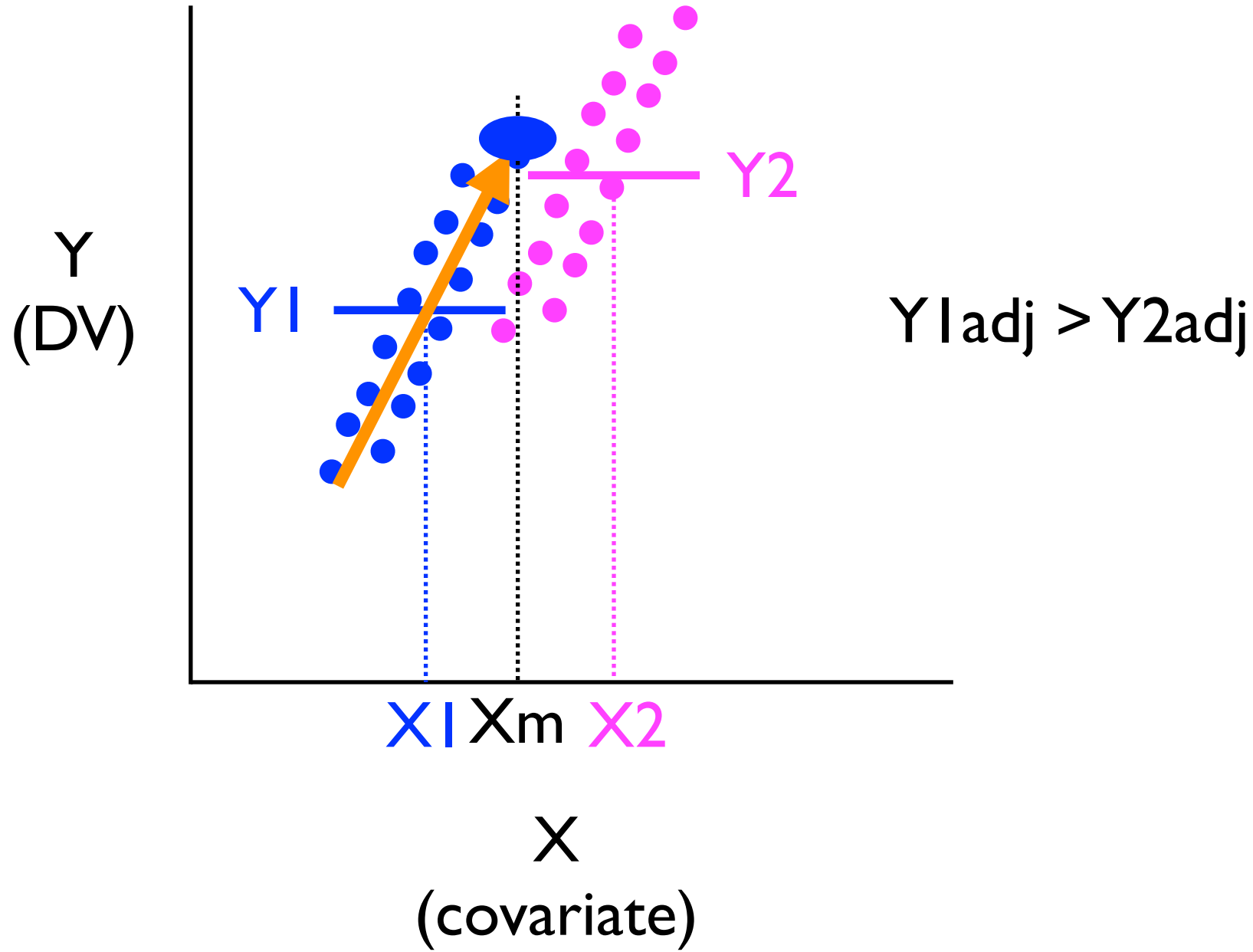
Adjusted Means depend on the relationship between covariate and DV



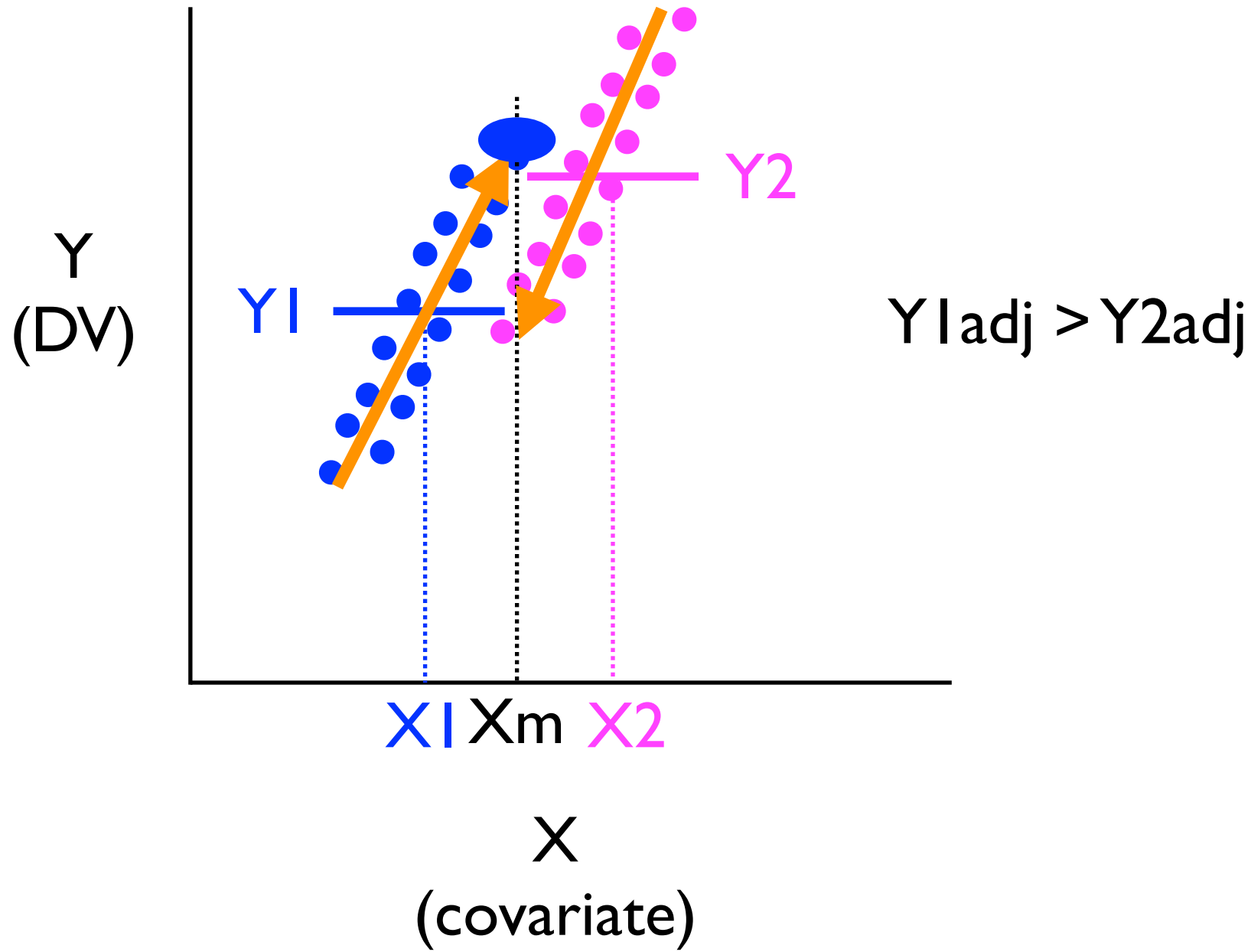
Adjusted Means depend on the relationship between covariate and DV



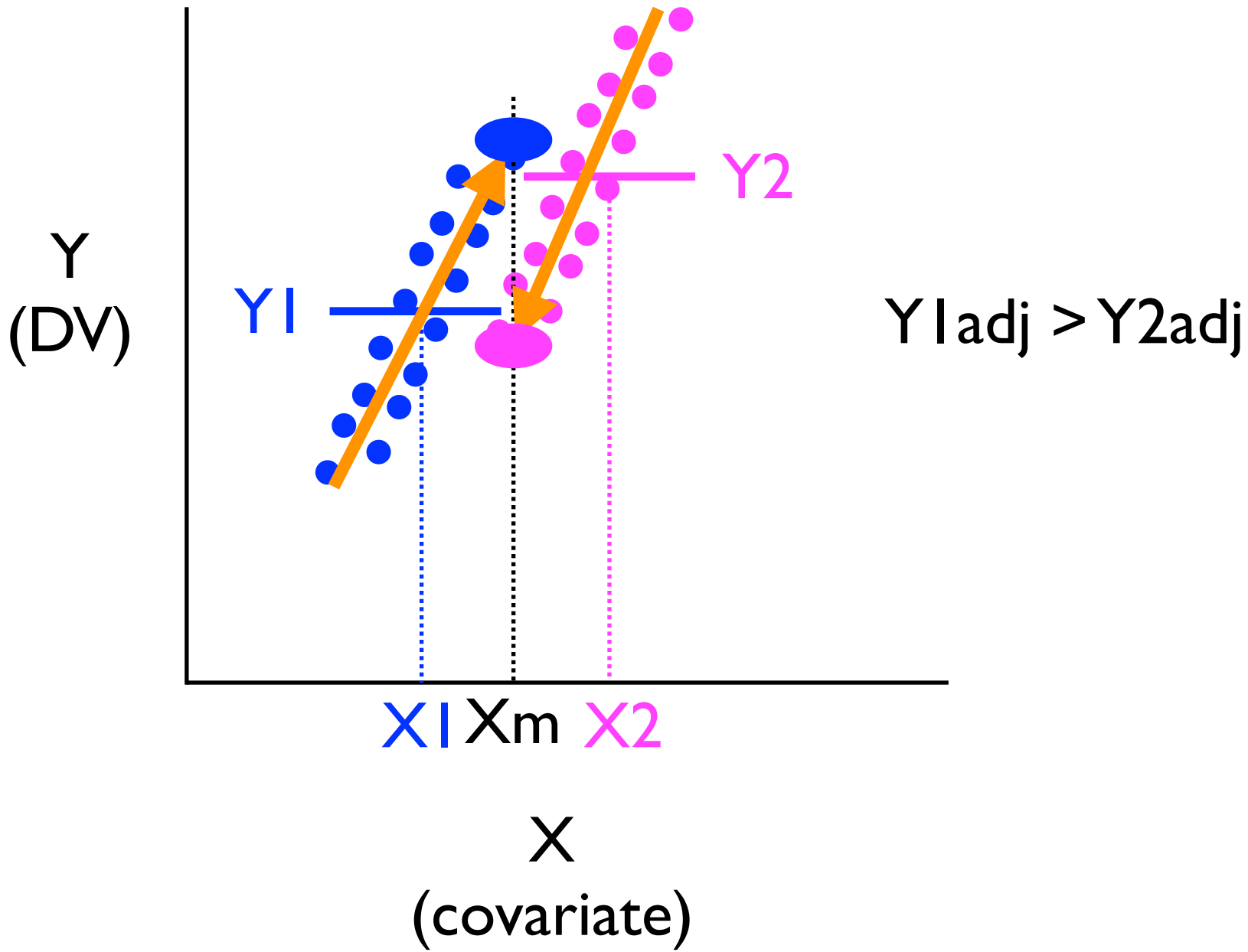
Adjusted Means depend on the relationship between covariate and DV



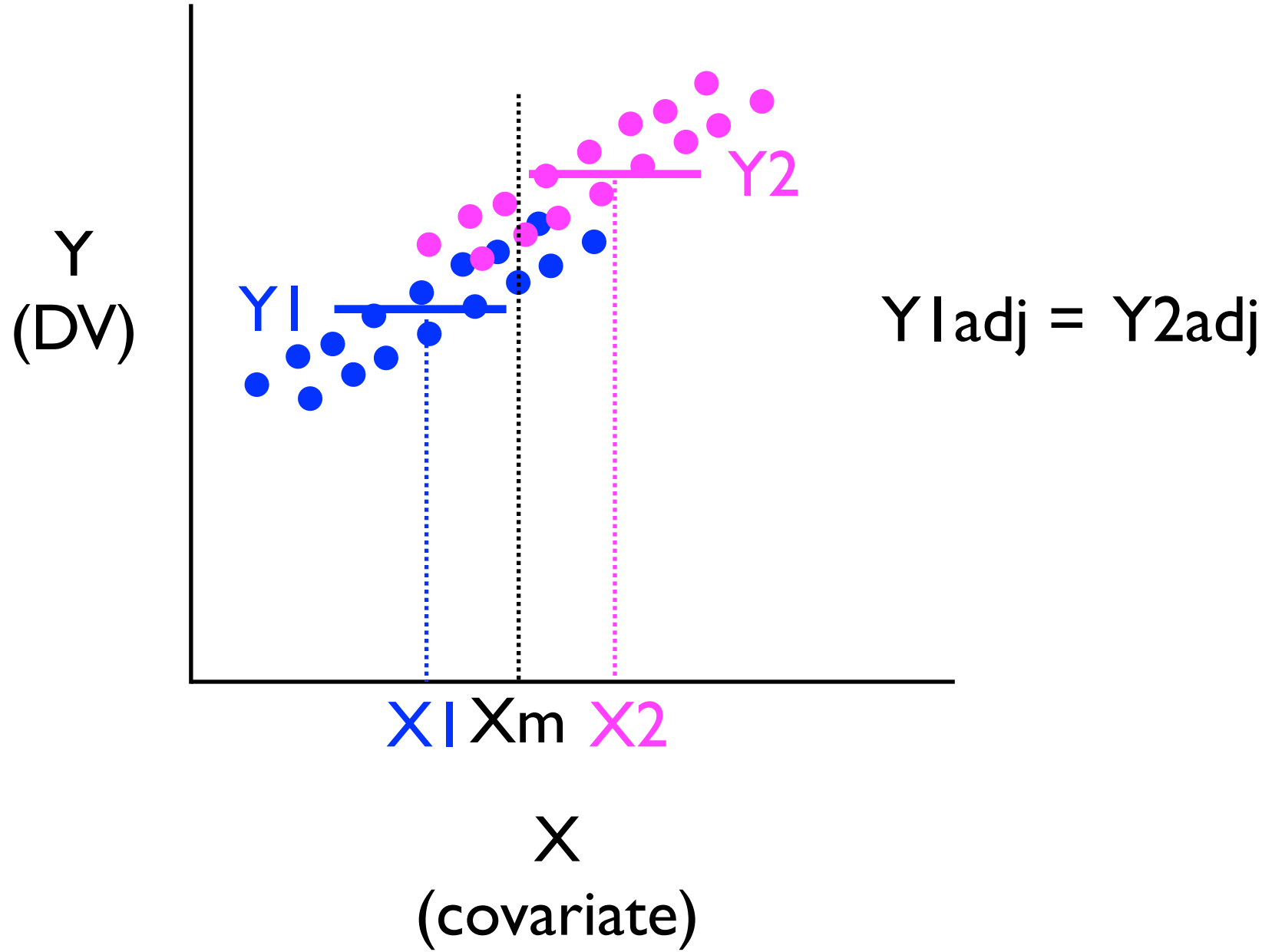
Adjusted Means depend on the relationship between covariate and DV



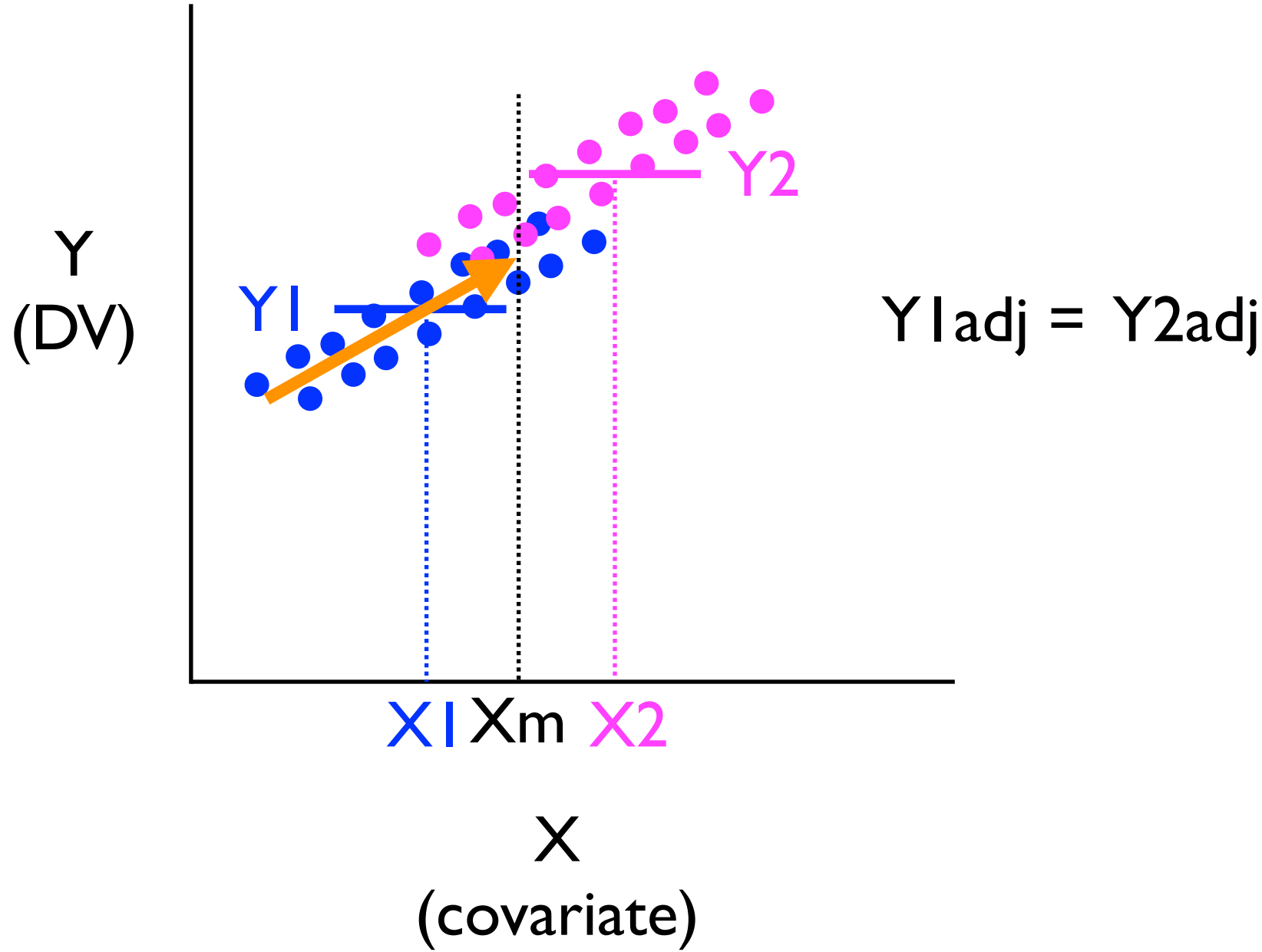
Adjusted Means depend on the relationship between covariate and DV



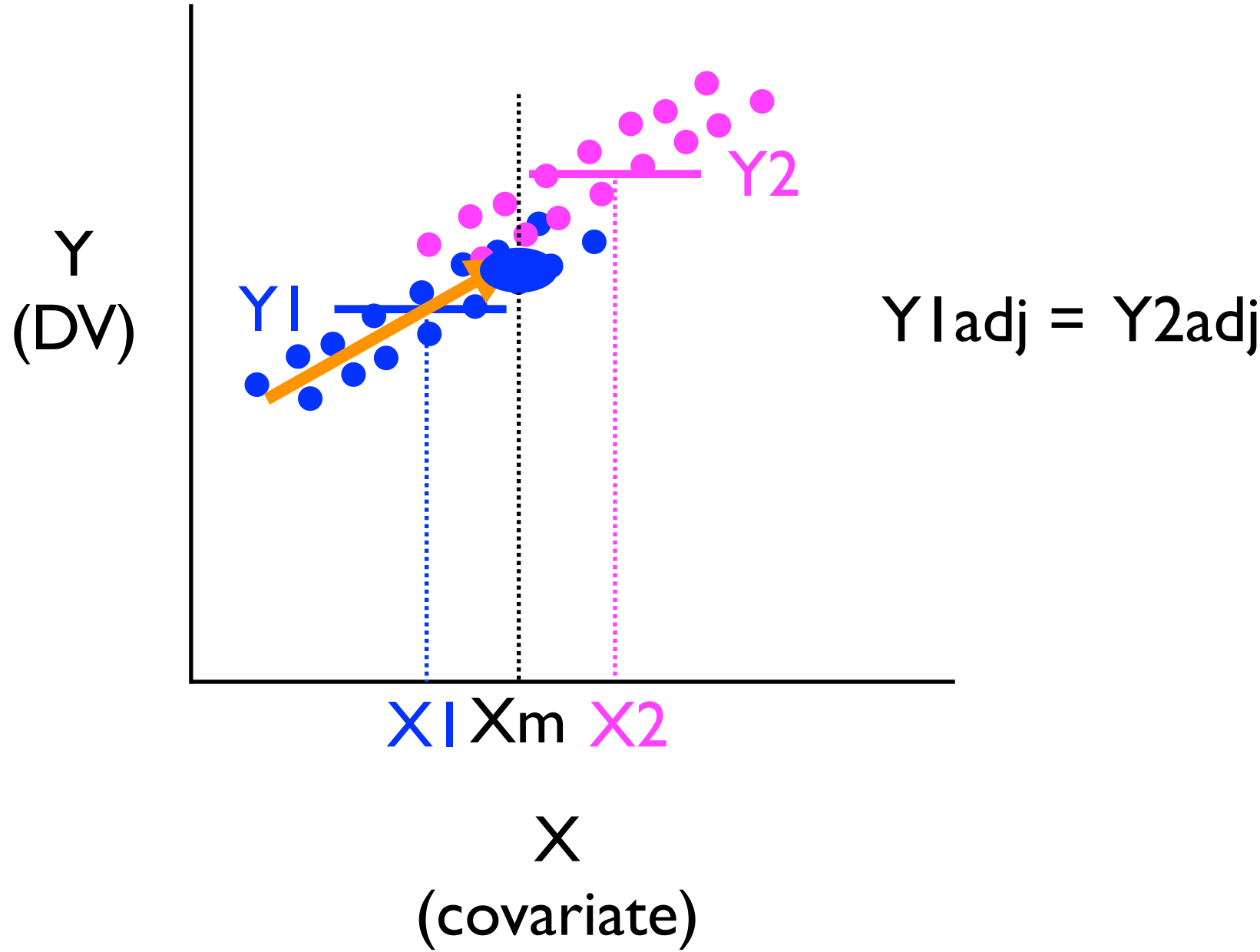
Adjusted Means depend on the relationship between covariate and DV



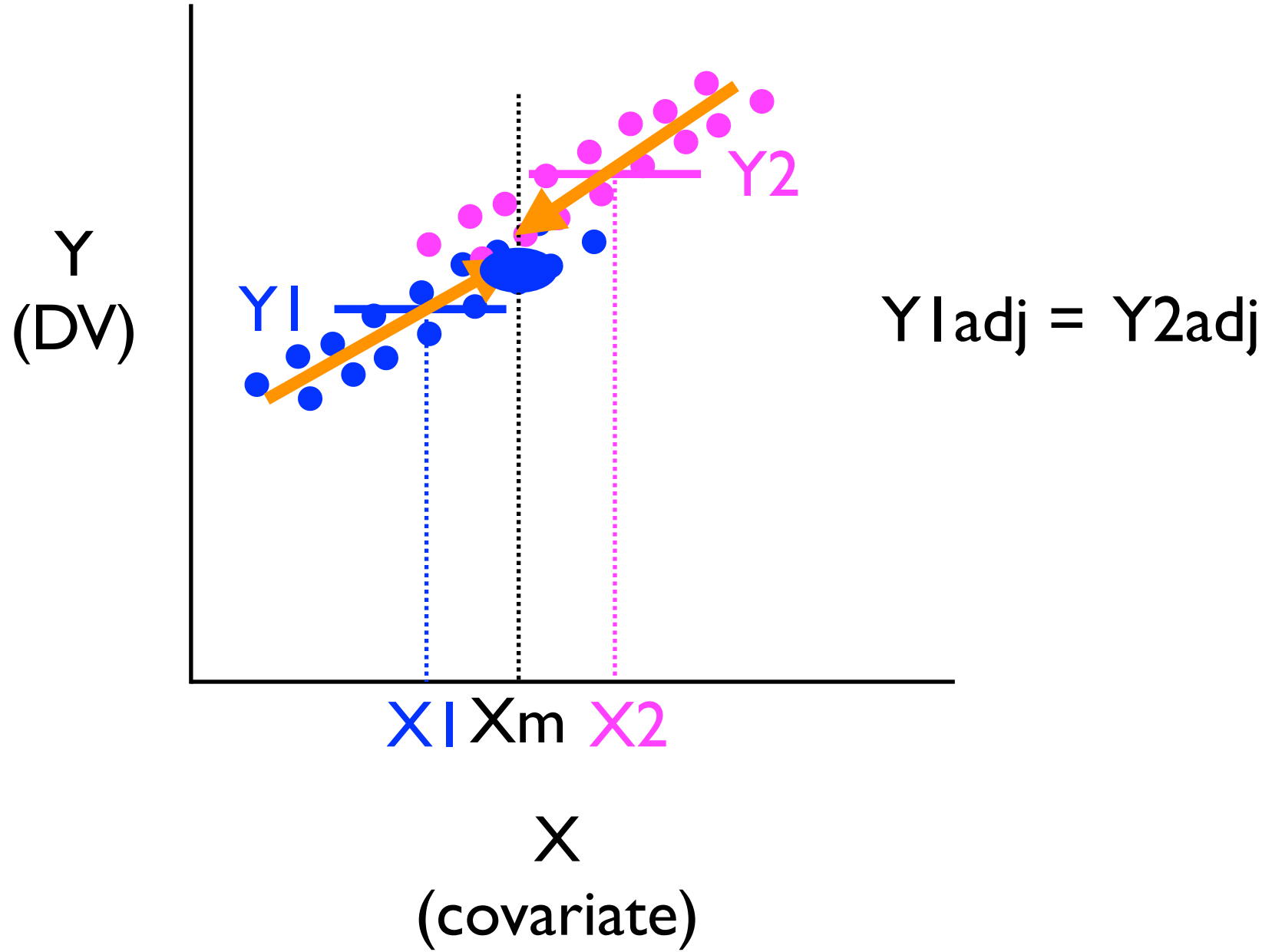
Adjusted Means depend on the relationship between covariate and DV



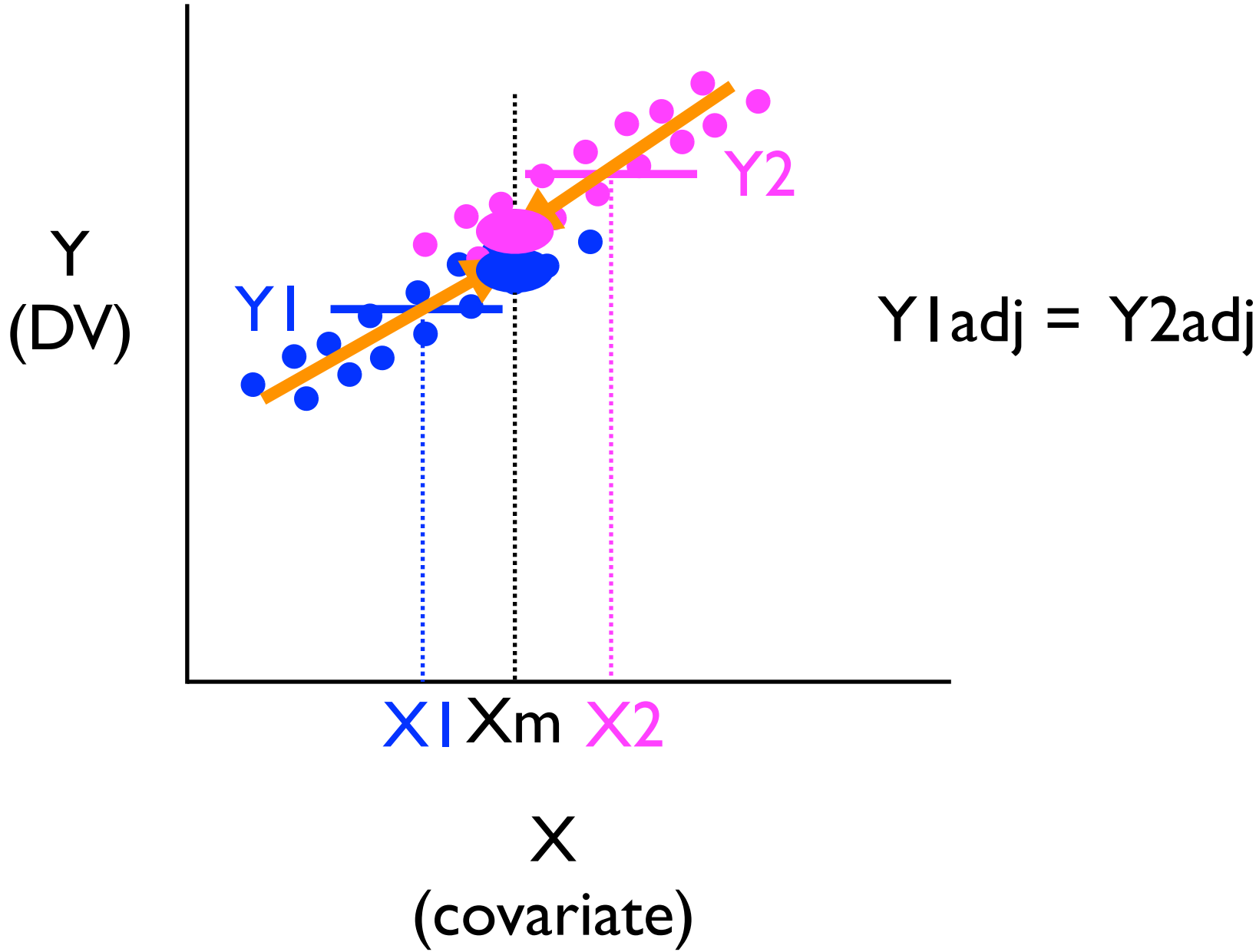
Adjusted Means depend on the relationship between covariate and DV



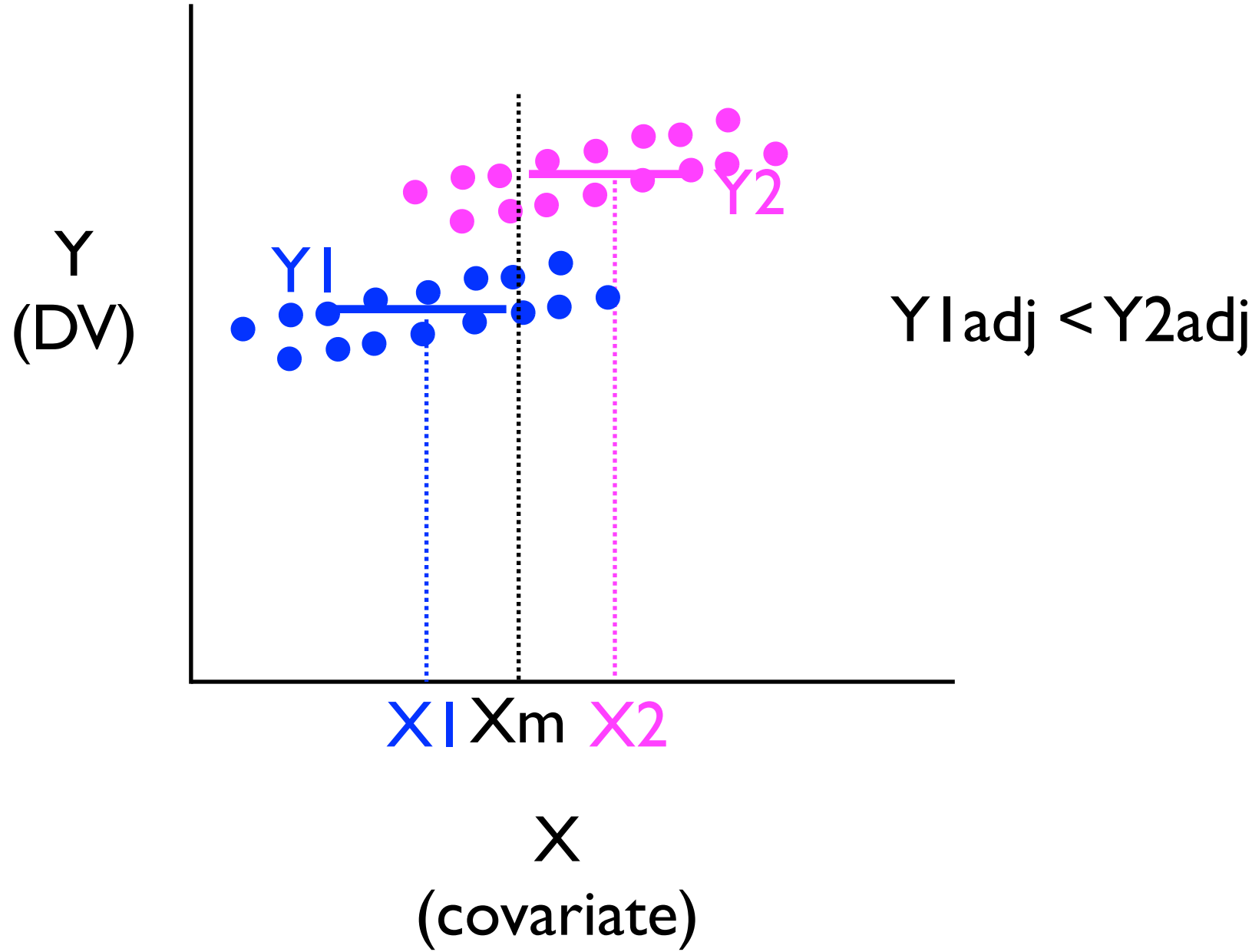
Adjusted Means depend on the relationship between covariate and DV



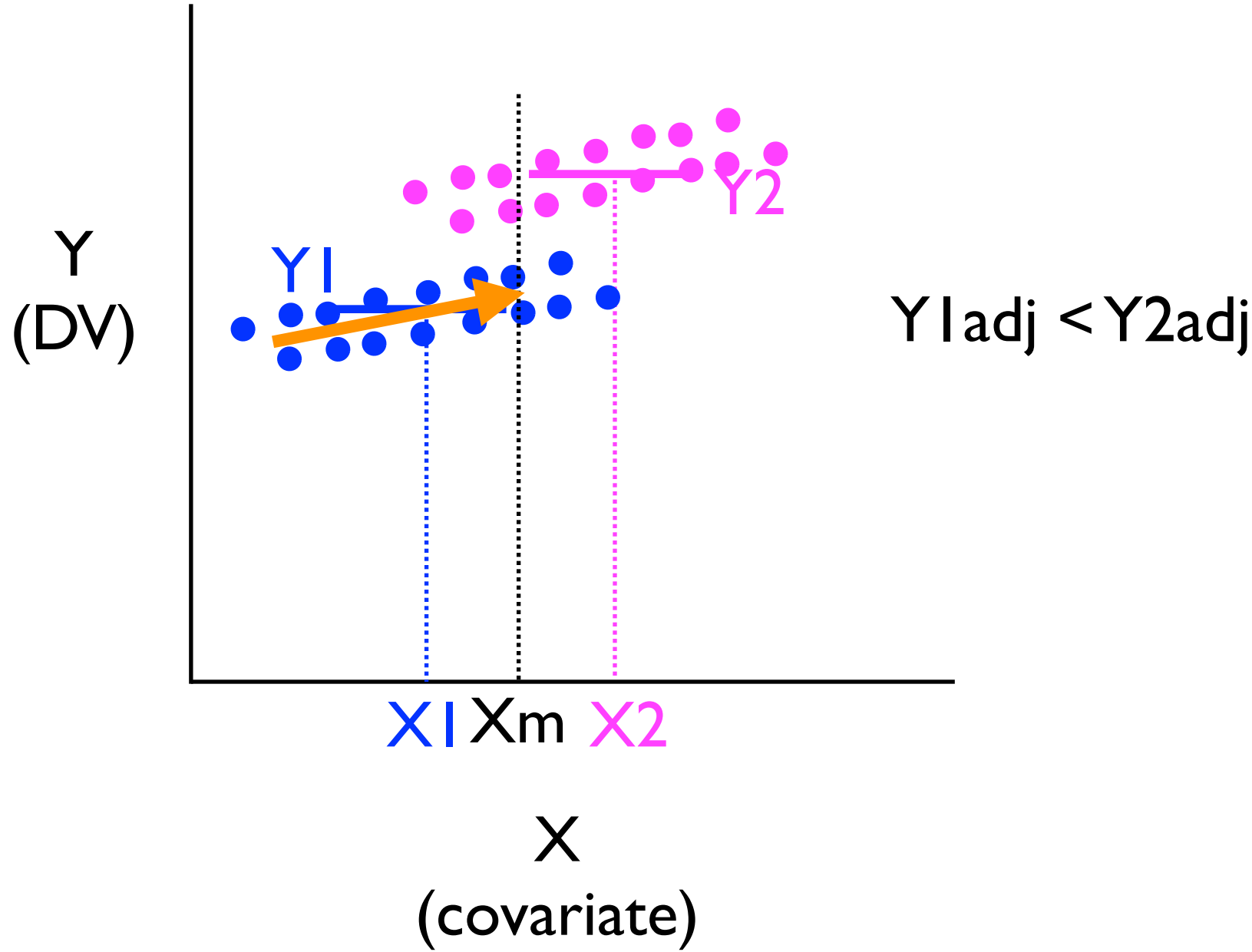
Adjusted Means depend on the relationship between covariate and DV



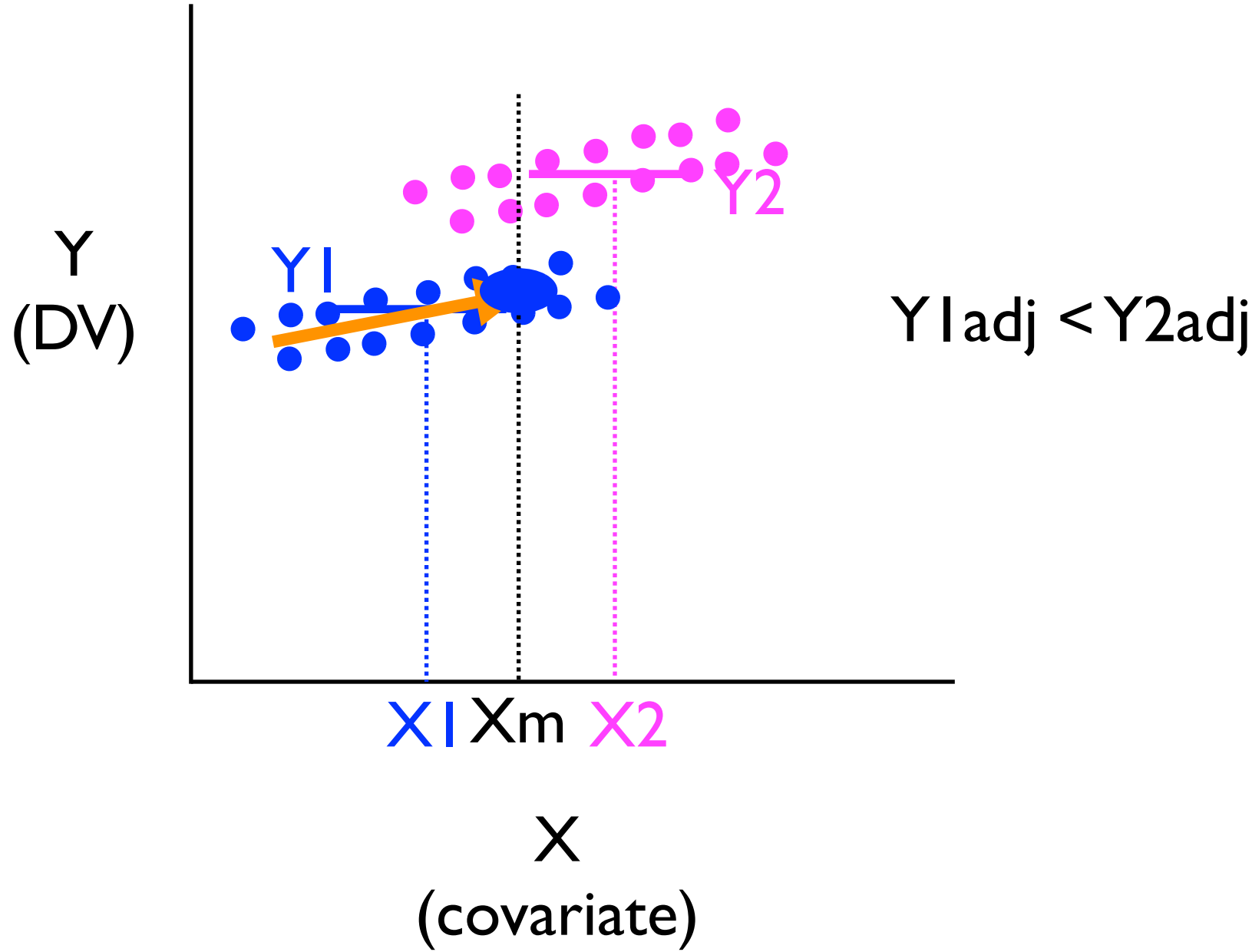
Adjusted Means depend on the relationship between covariate and DV



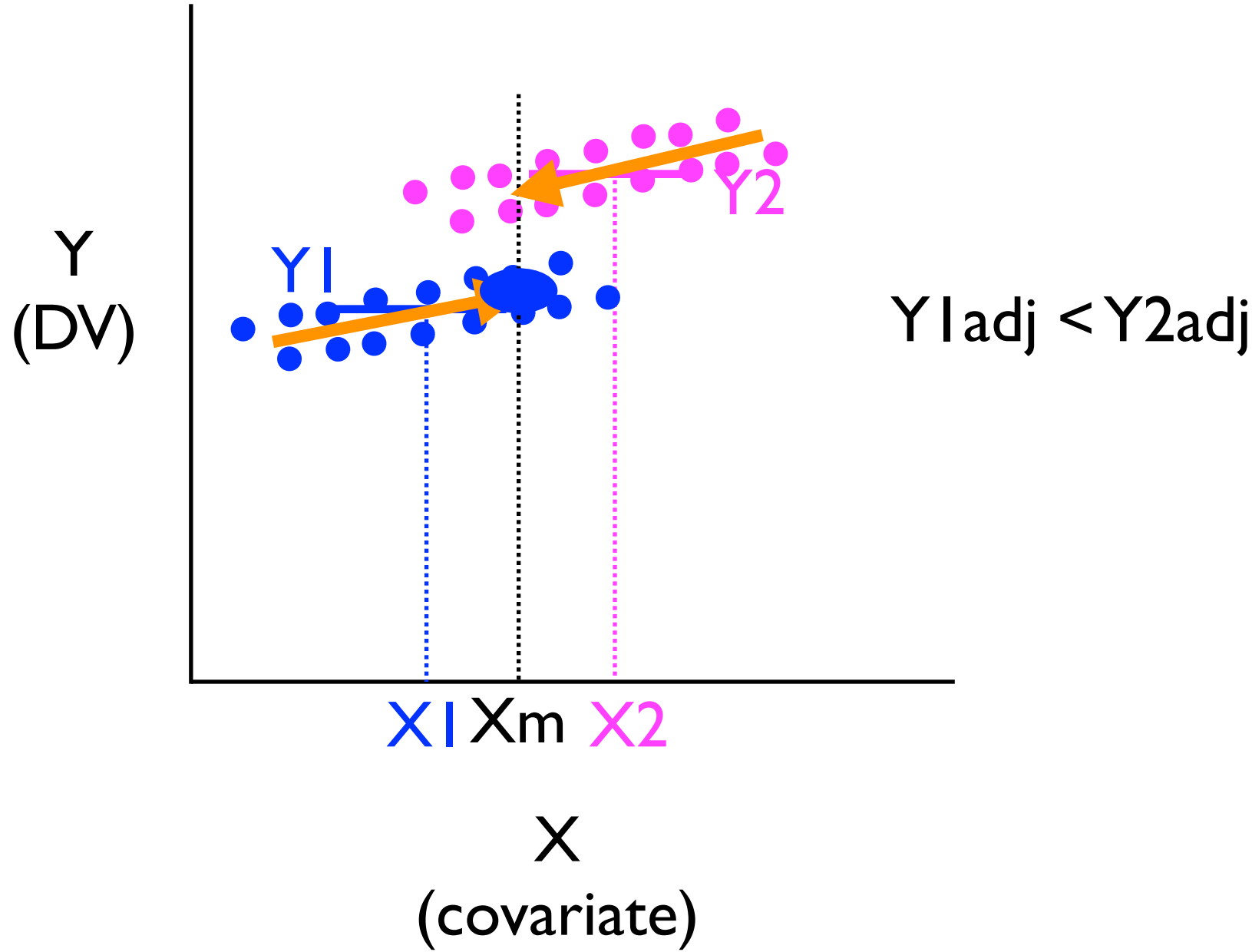
Adjusted Means depend on the relationship between covariate and DV



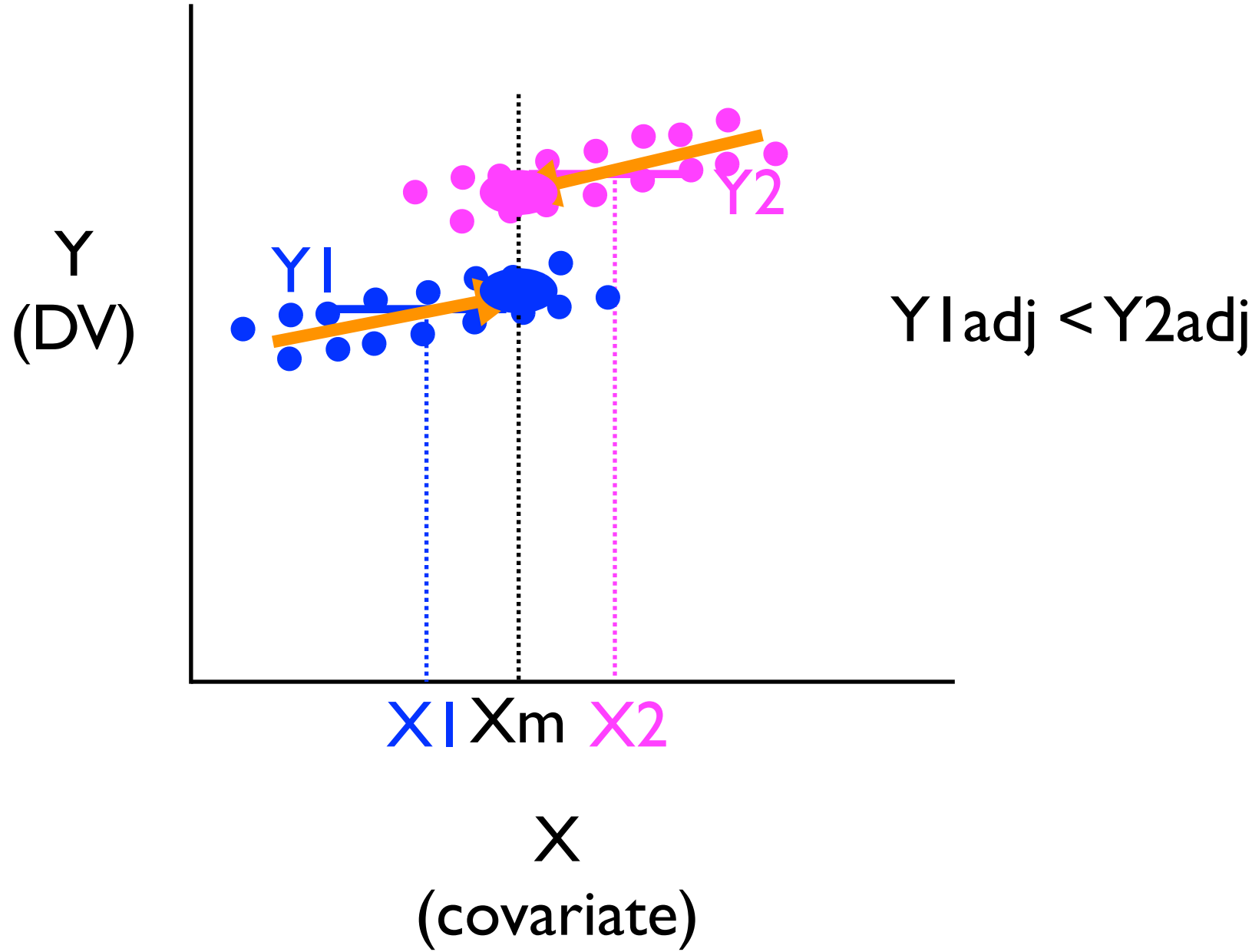
Adjusted Means depend on the relationship between covariate and DV



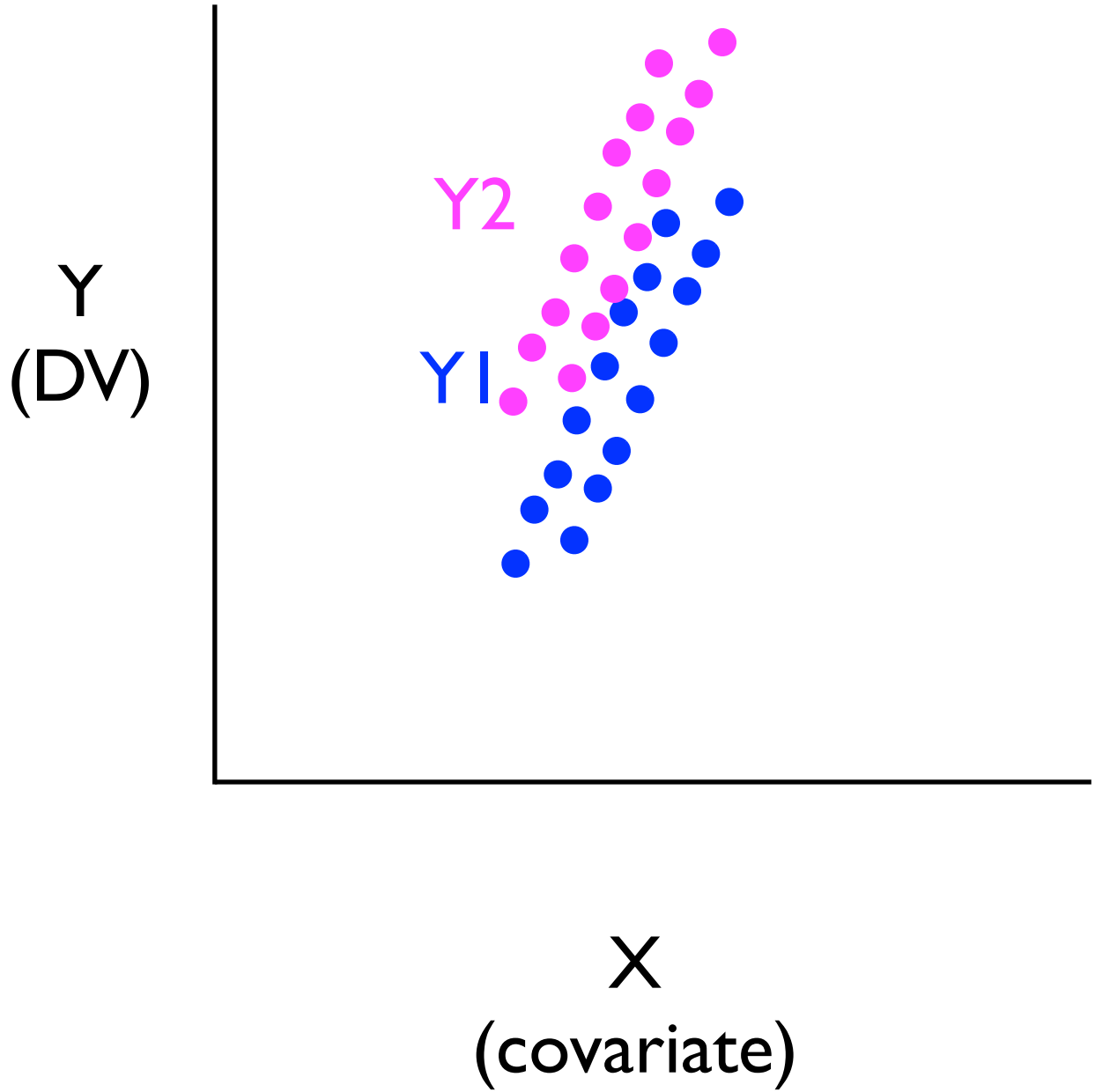
Adjusted Means depend on the relationship between covariate and DV



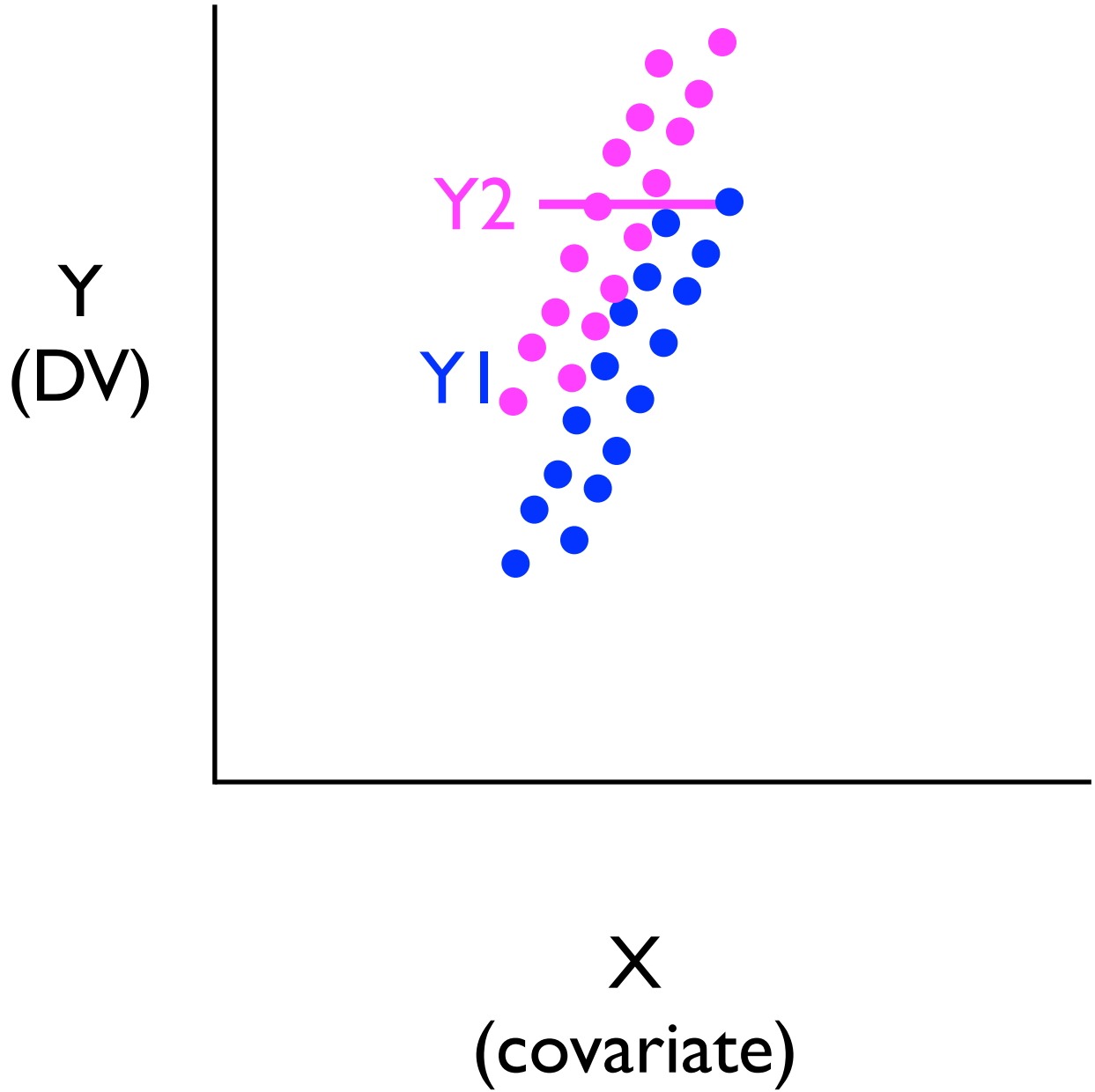
Adjusted Means depend on the relationship between covariate and DV



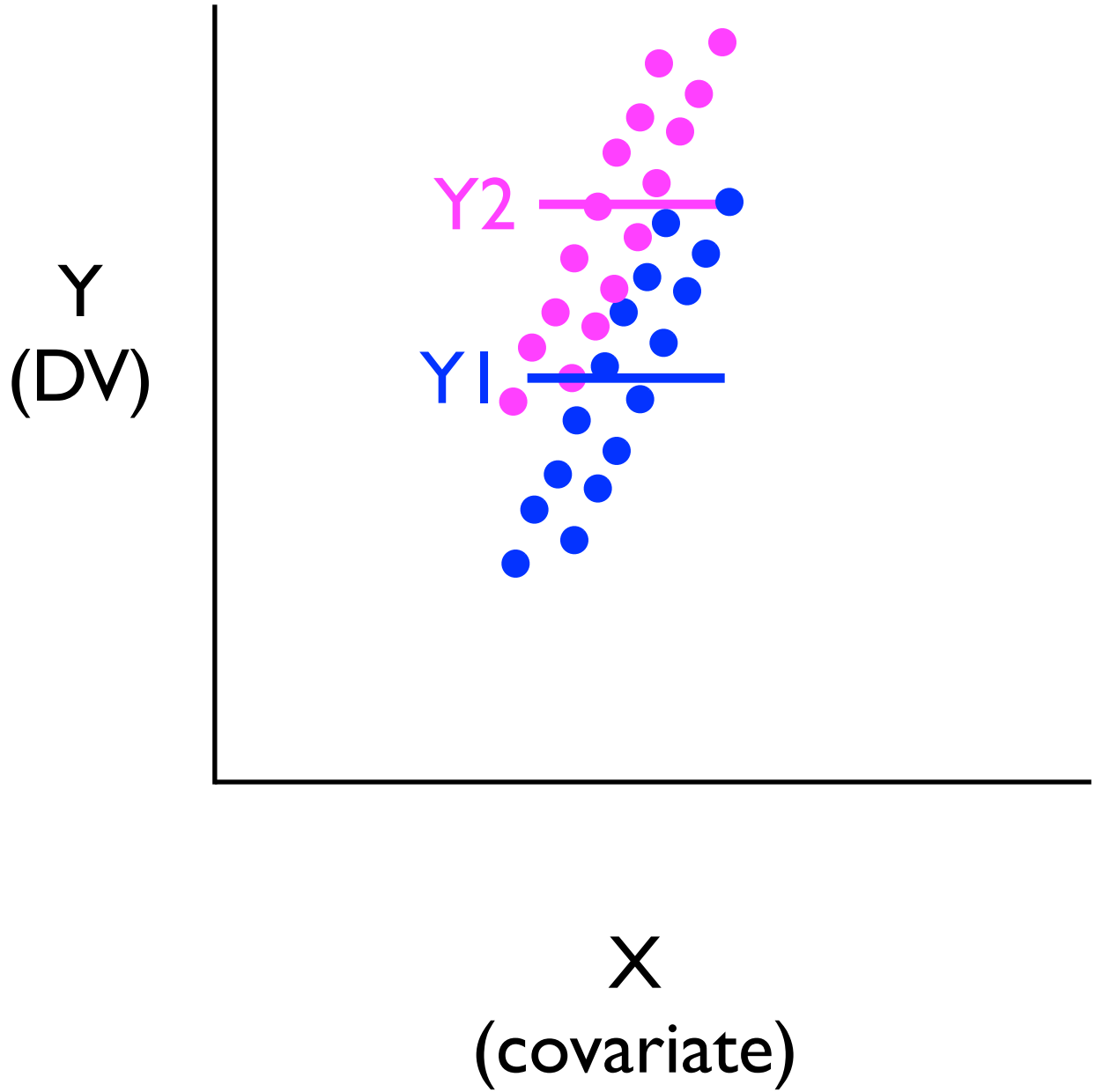
ANCOVA can help even if there are no pre-existing group differences on the covariate



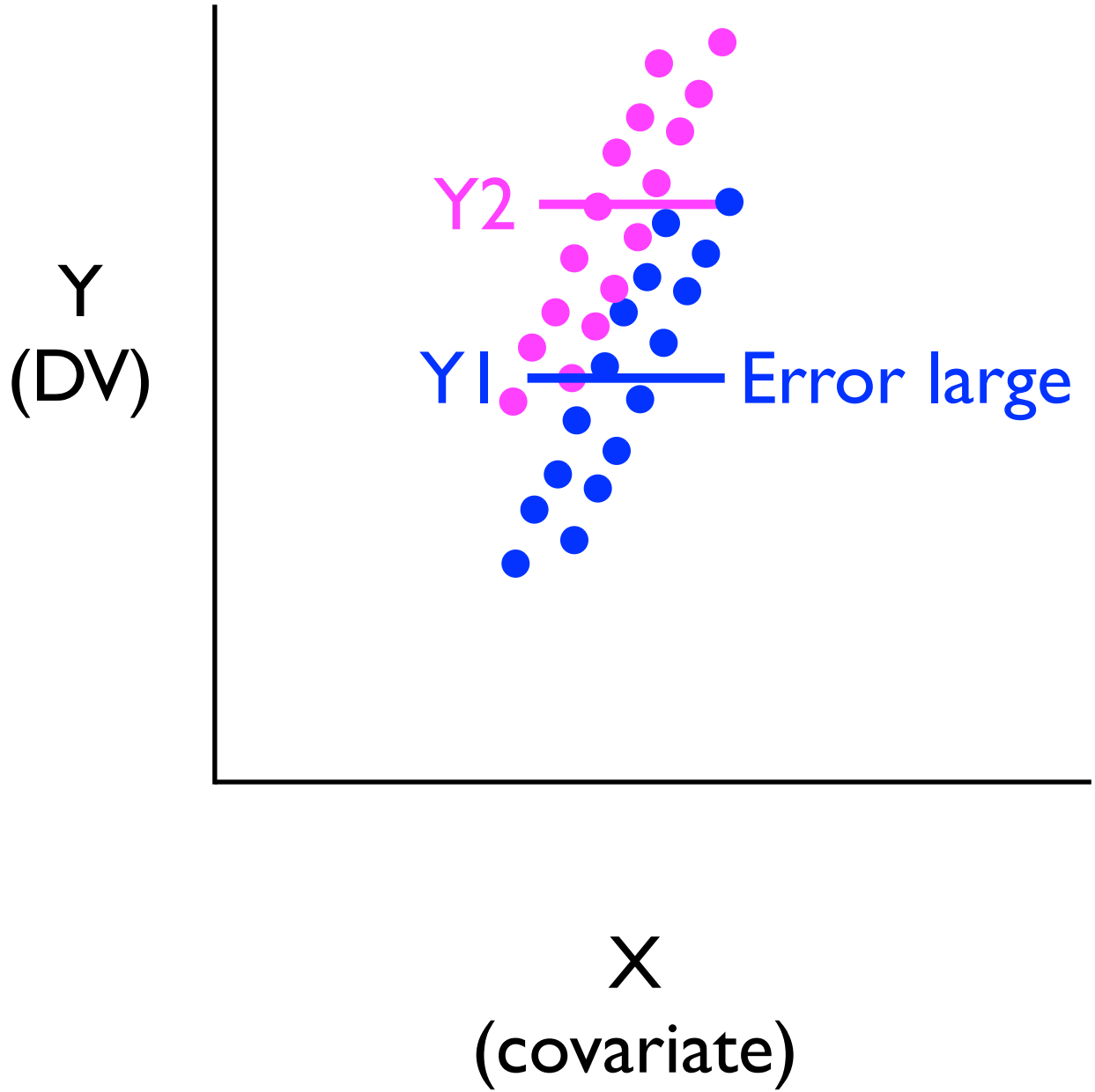
ANCOVA can help even if there are no pre-existing group differences on the covariate



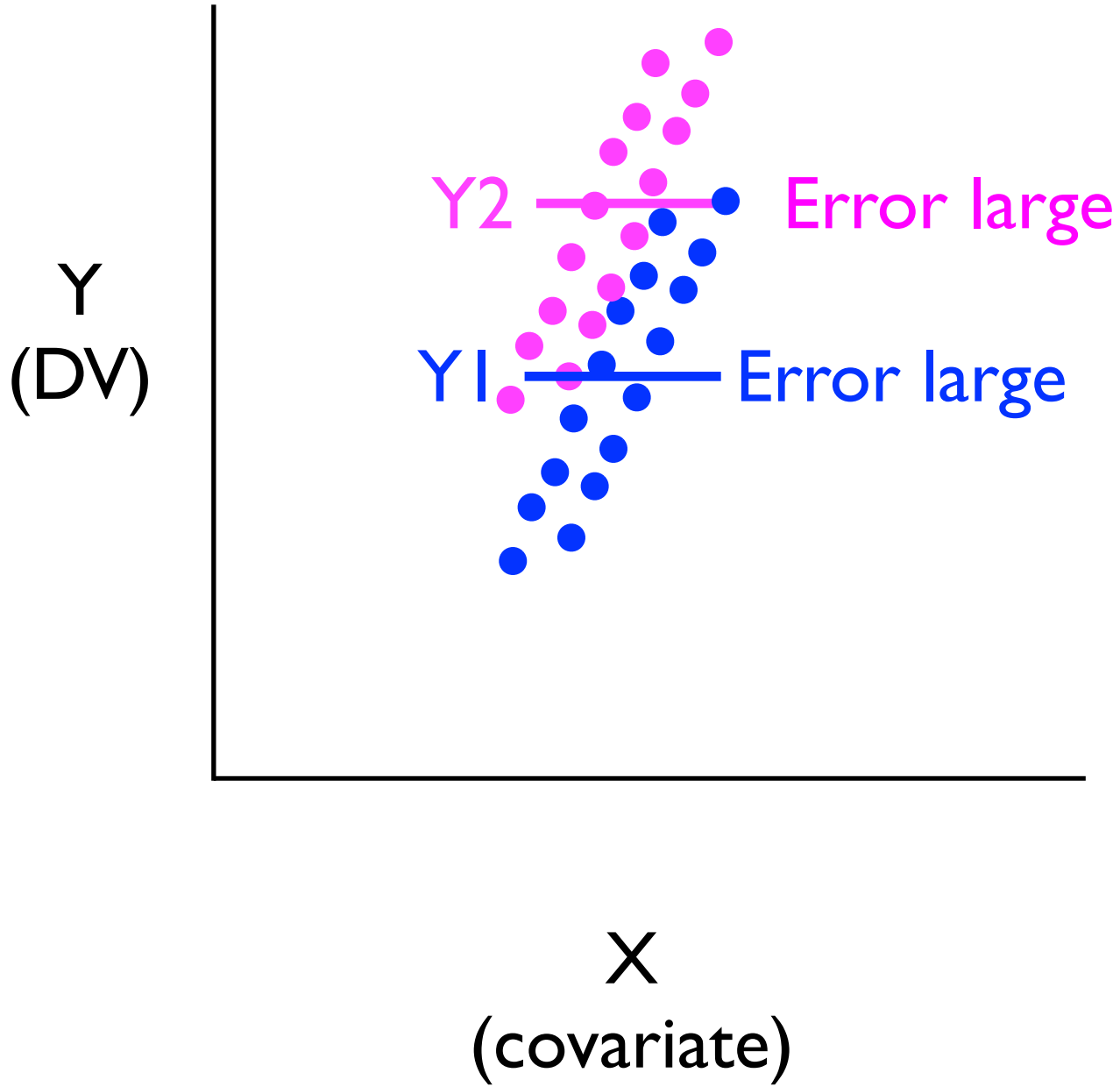
ANCOVA can help even if there are no pre-existing group differences on the covariate



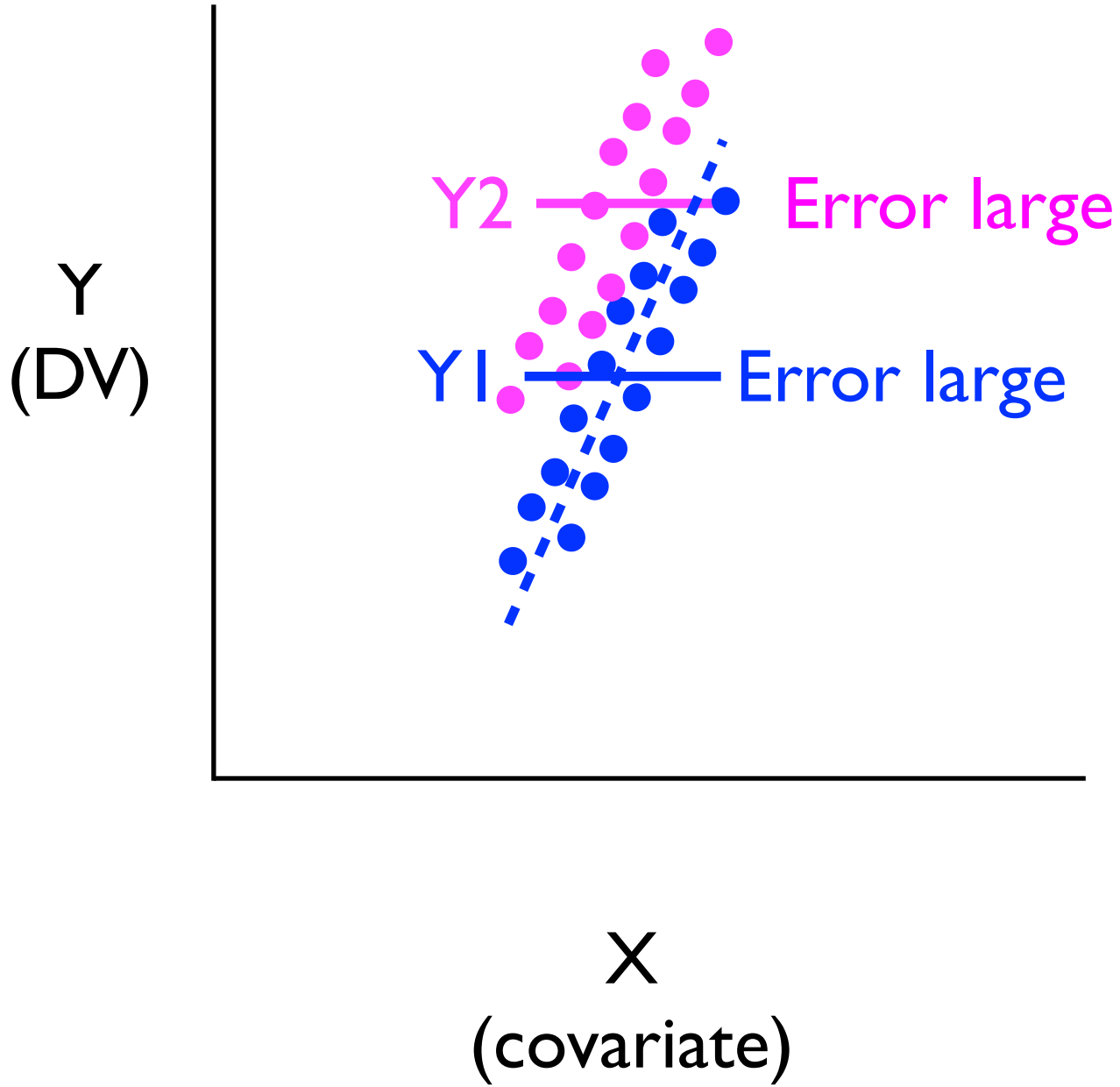
ANCOVA can help even if there are no pre-existing group differences on the covariate



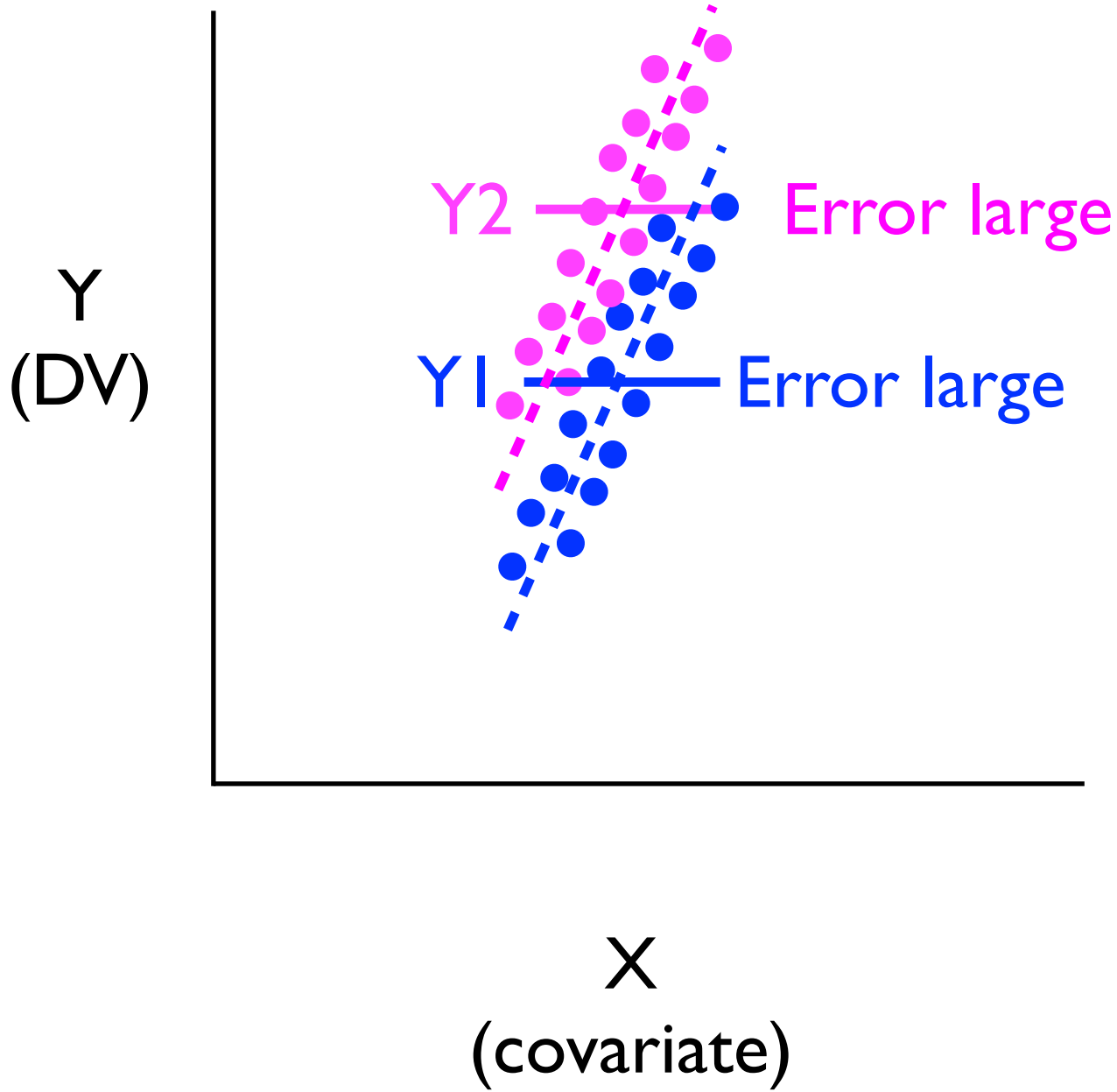
ANCOVA can help even if there are no pre-existing group differences on the covariate



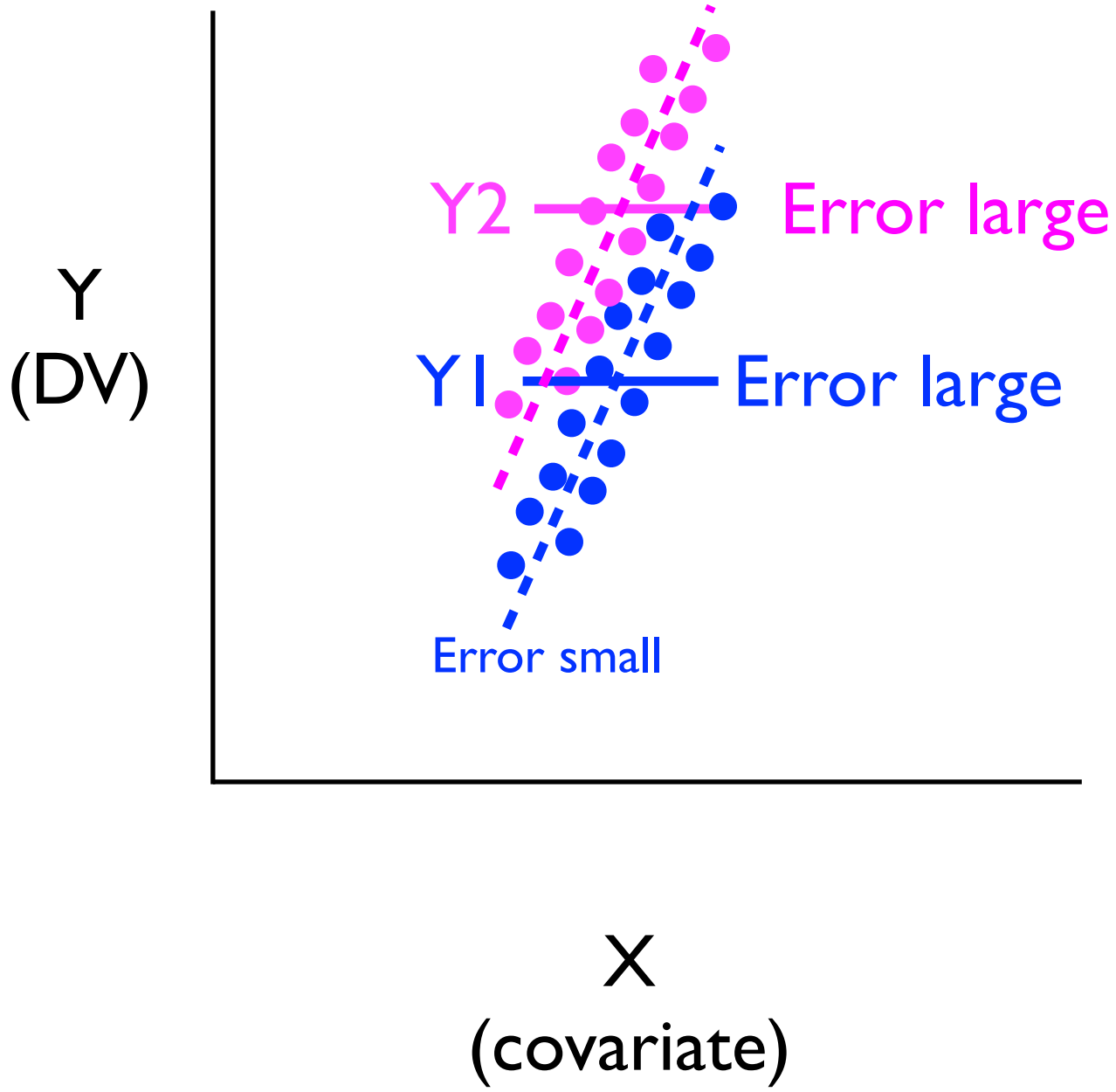
ANCOVA can help even if there are no pre-existing group differences on the covariate



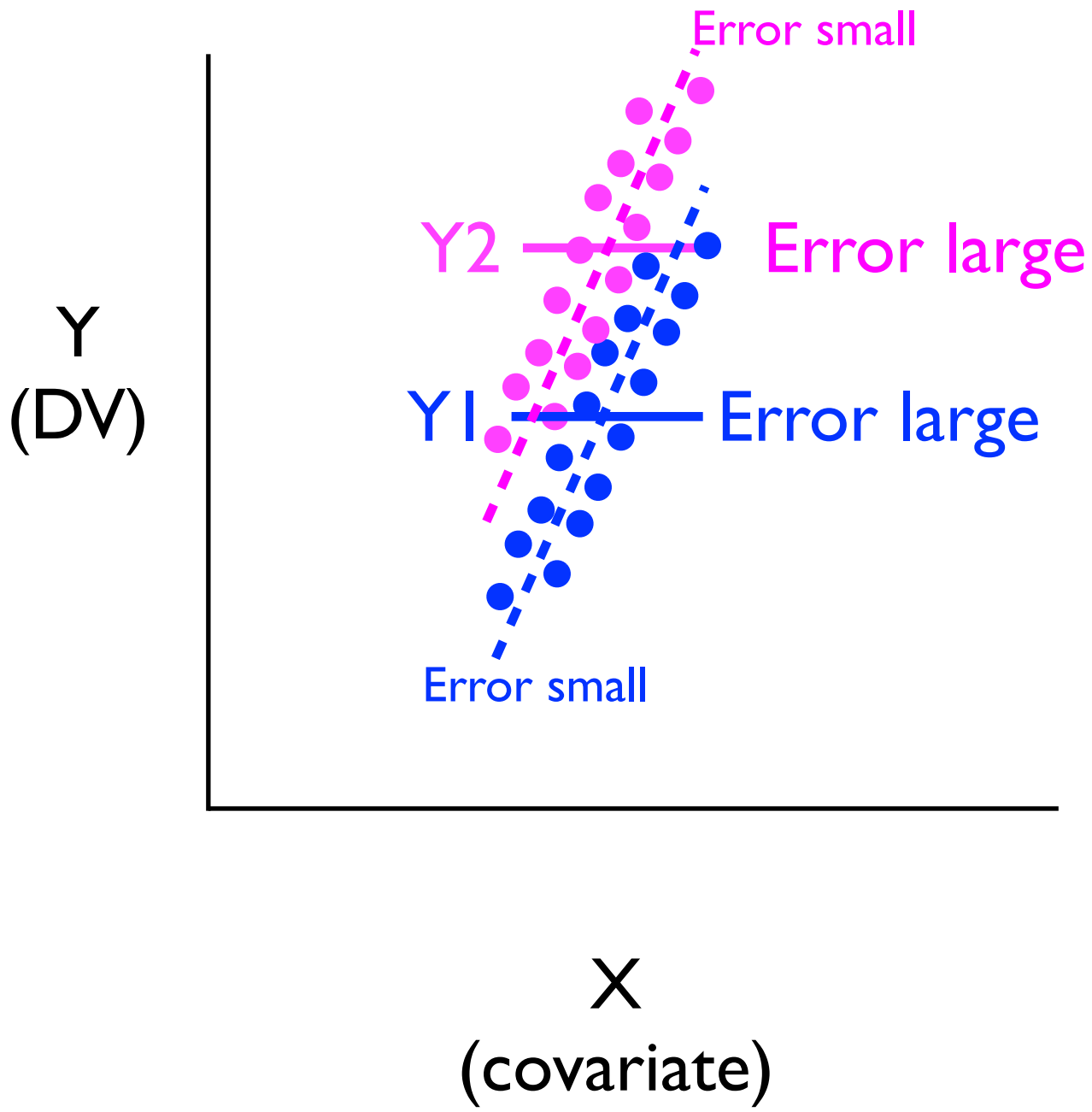
ANCOVA can help even if there are no pre-existing group differences on the covariate



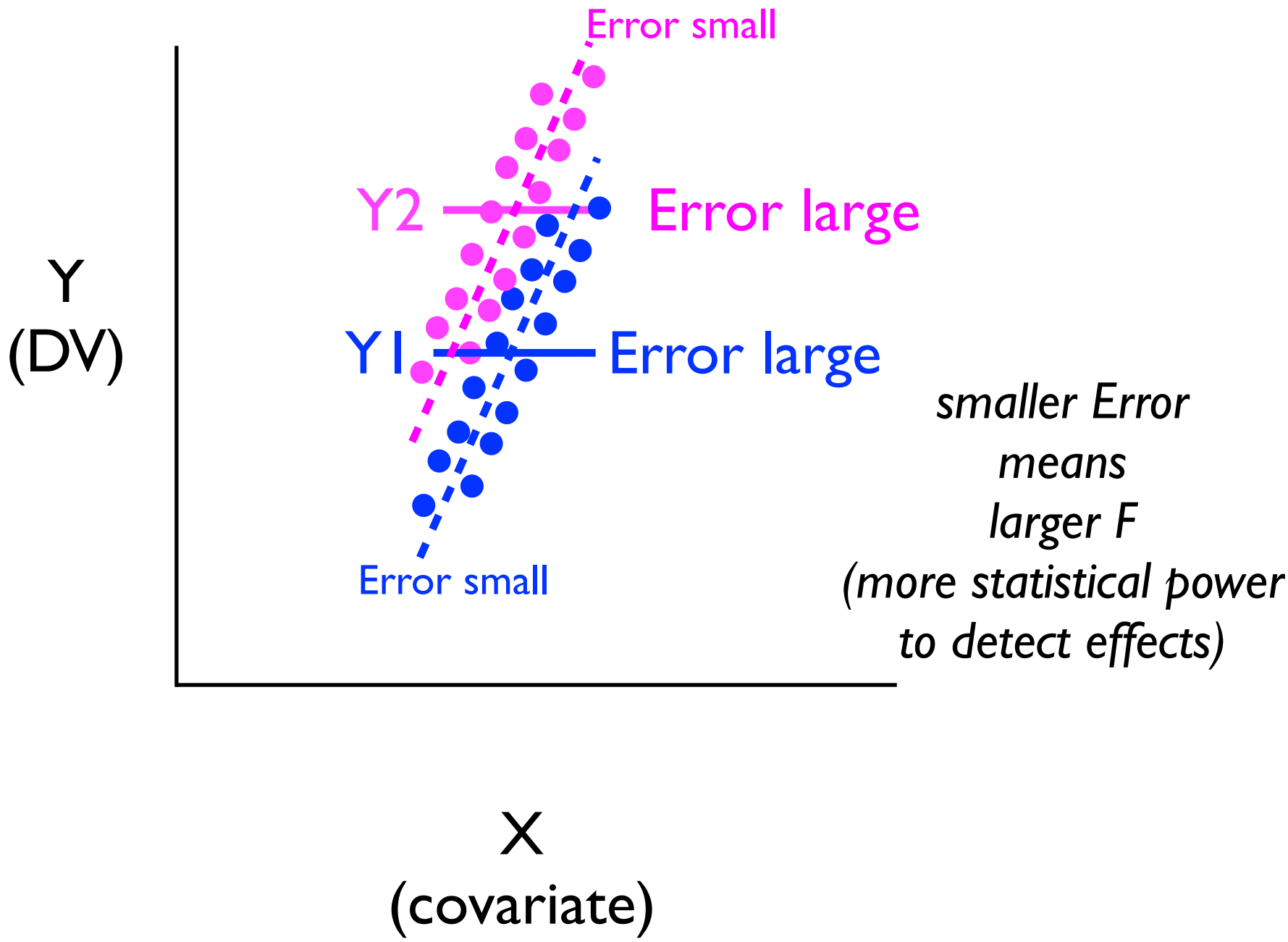
ANCOVA can help even if there are no pre-existing group differences on the covariate



ANCOVA can help even if there are no pre-existing group differences on the covariate



ANCOVA can help even if there are no pre-existing group differences on the covariate

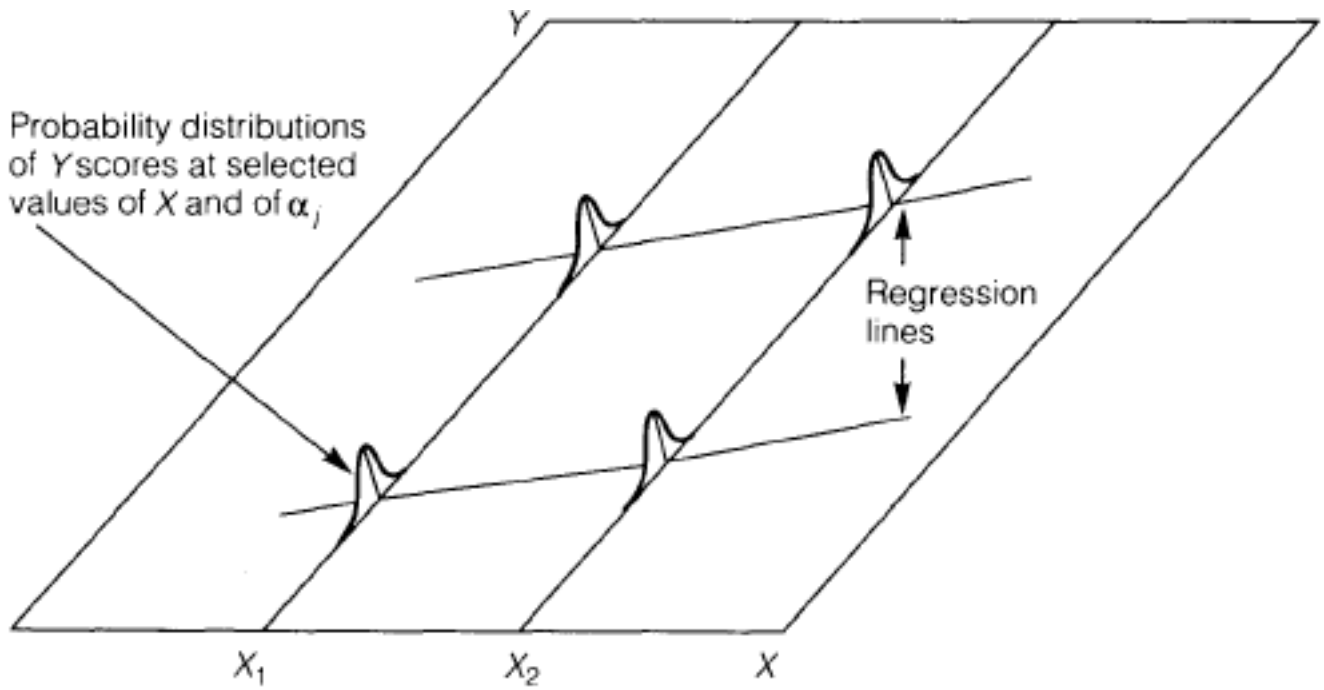


Summary

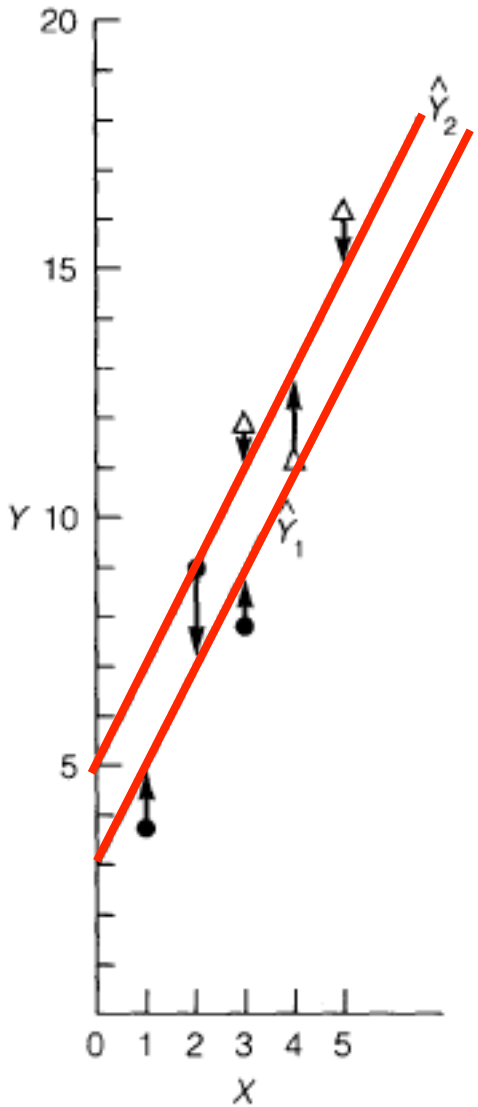
- by including a covariate in the model we bring about a substantial reduction in unaccounted-for variance
- greater power for detection of treatment effects
- adjusted means are thought of as estimates based on the full model of the mean performance that would have been obtained in each treatment group IF there had been no differences between mean scores on the covariate

Assumptions

- DV scores must be normally distributed
- Must be so at all values of the covariate



Assumptions

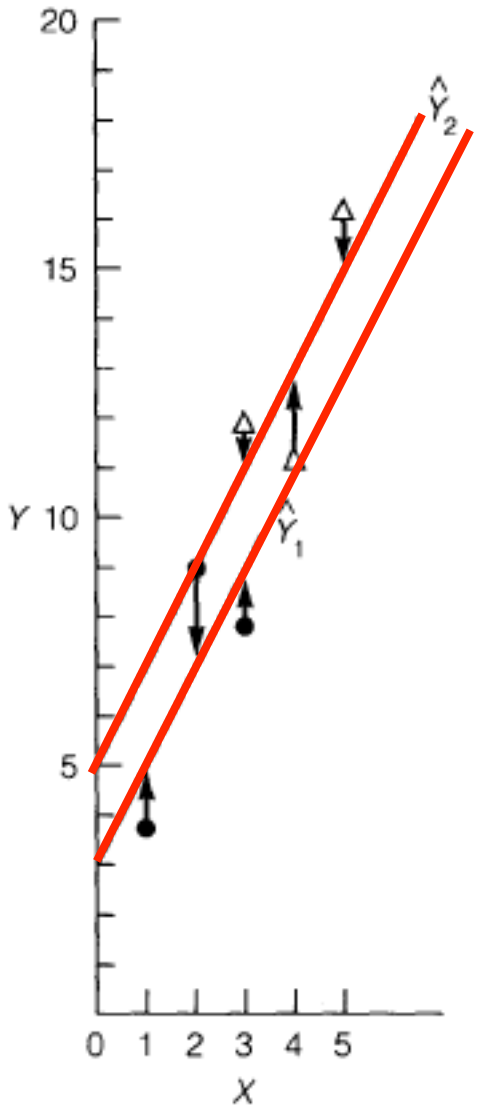


- Homogeneity of Regression
- separate within-group regression lines have the same slope

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

- there is only one Beta coefficient
- (no subscript j)

Assumptions

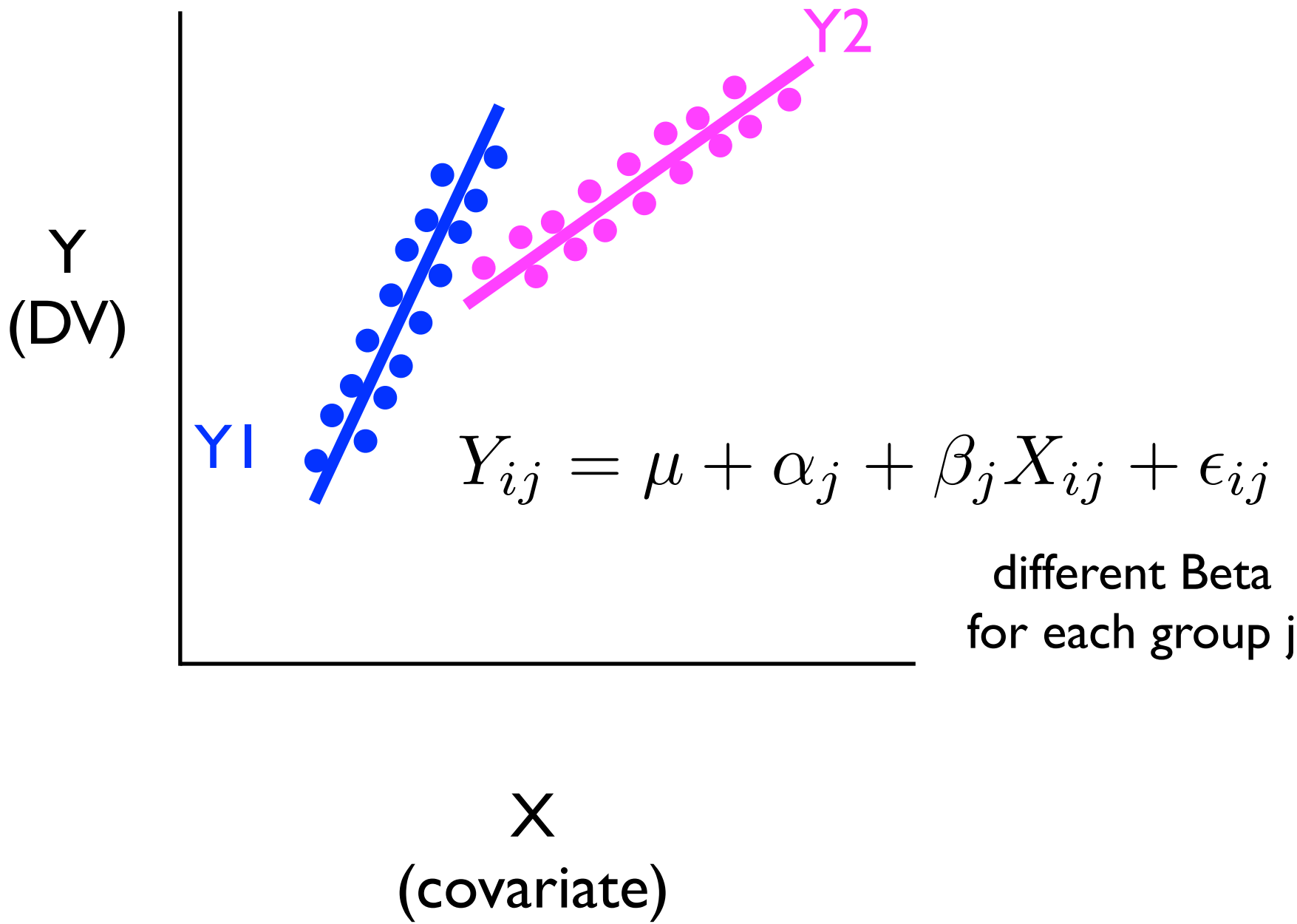


- Homogeneity of Regression
- separate within-group regression lines have the same slope

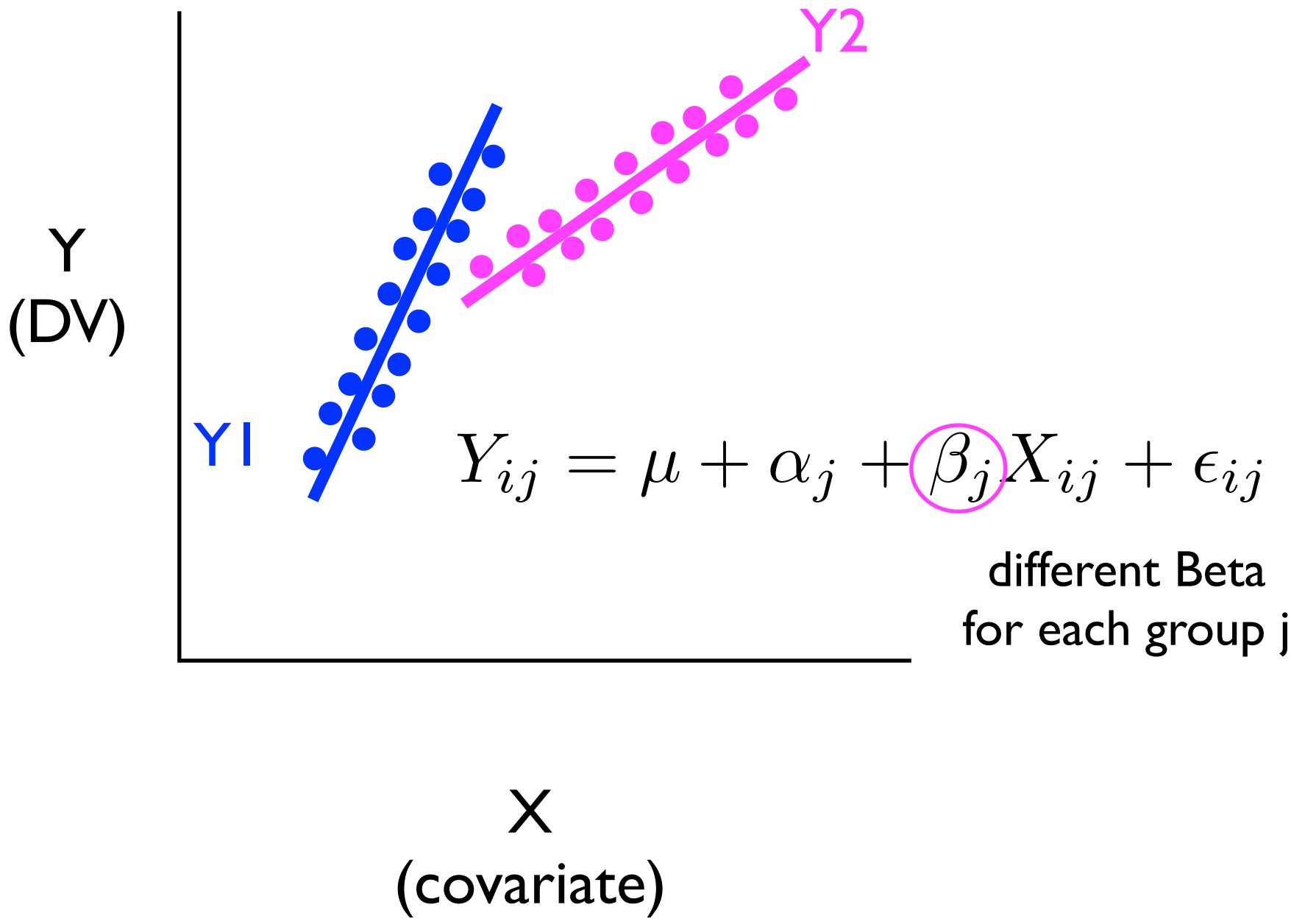
$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

- there is only one Beta coefficient
- (no subscript j)

Heterogeneity of Regression



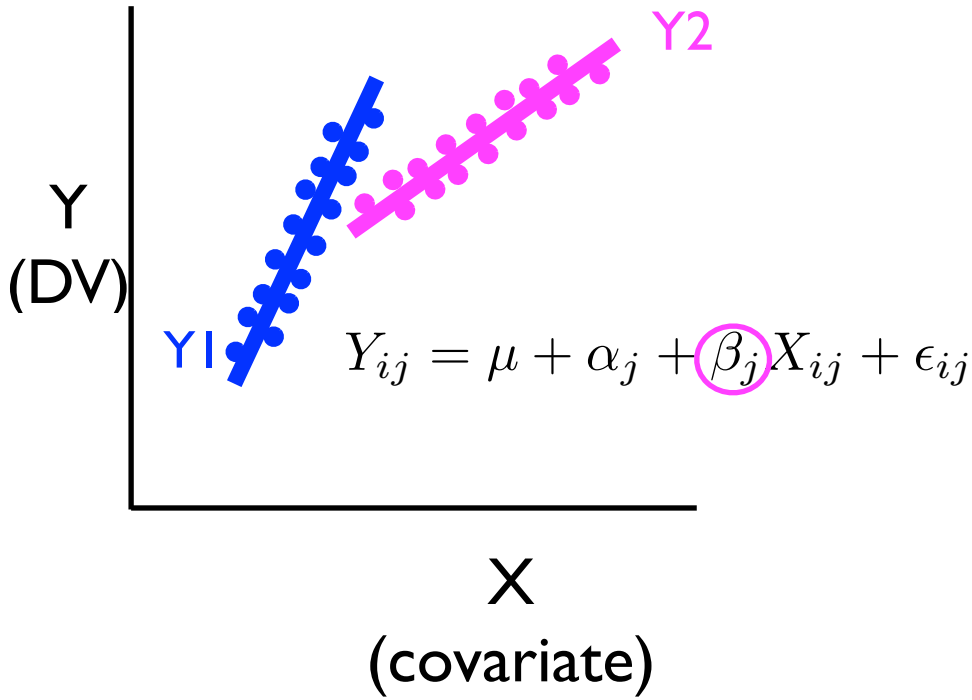
Heterogeneity of Regression



Heterogeneity of Regression

you can do it but it costs one df for every Beta_j you have to estimate

also if Y1 and Y2 overlap in X, but they have different slopes, what does that mean?



Assumptions

- Lack of Independence of Treatment and Covariate
- we can "statistically" adjust for differences in covariate
- however this does not experimentally control for potential interactions between the score on the covariate and the treatment effect

Assumptions

- we can statistically adjust for differences in a covariate (e.g. age) by mathematically shifting each group
- BUT:
- what if an experimental treatment (e.g. a drug) actually has different effects on young vs old people?
- we can't tell; statistically adjusting for the means on the covariate won't solve this problem
- the only way to truly deal with this is to equate groups on age in the first place