Strategies to adjust for confounding: a friendly introduction

Zagreb 2019

CHARITÉ
BERLIN SCHOOL OF PUBLIC HEALTH

Quick review of confounding

- Mixing of effects between the association under study and a third variable
- Properties of confounding variables
    - A ‘confounder’ is a common cause (direct or indirect via another variable) of both the exposure and the outcome
    - Once a ‘confounder’, not always a ‘confounder’! - DAG-dependent
    - Important: NOT on the causal pathway (intermediate) & not collider

Methods to control for confounding

In the design of the study:

- Restriction
- Matching
- Randomization

In the analysis of the data:

- Stratification (& pooling, weighting, standardization)
- Regression modeling

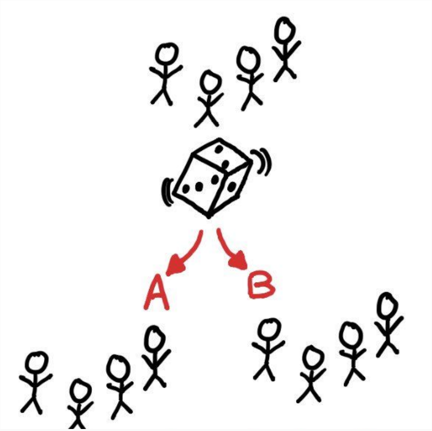
In the design: Restriction

- E.g. analysis only in females
  - Removes confounding by sex
- Often seen for certain age groups (e.g. we studied Exposure X in healthy individuals over 65).
- Careful! Lots of restriction = low generalizability of results

In the design: Matching

- In cohort studies:
  - Match exposed and not exposed persons on confounding factors
    - e.g. age and sex matching
- What if many confounders relative to number of outcomes?
  - Propensity score matching
  - Collapse many confounding variables into a single ‘score’
    - 0 to 1 = prediction of exposure based on these confounders
    - Match participants with similar scores (e.g., 0.41 and 0.42)

In the design: Randomization



Credit: @epiellie

- Goal: groups on average at same risk for outcome before the treatment is assigned/given
- Confounding factors balanced on average between arms

Methods to control for confounding

In the design of the study:

- Restriction
- Matching
- Randomization

In the analysis of the data:

- Stratification (pooling, standardization)
- Regression modeling

In the analysis: stratification and pooling

- Divide the data into strata according to categories of the confounder
- Within each stratum, calculate stratum-specific measures of association
- If appropriate, pool information over all strata by calculating a weighted average of the stratum-specific measures of association
  - (e.g. Mantel-Haenszel formula)
- Assumption: Constant effect across all strata

Problem with pooling?

What if the effect is NOT constant across all strata?

When the overall magnitude of the relationship between the exposure and disease depends (differs, is modified) by the level of a third variable (called the effect modifier) in size or even direction.

Effect (measure) modification

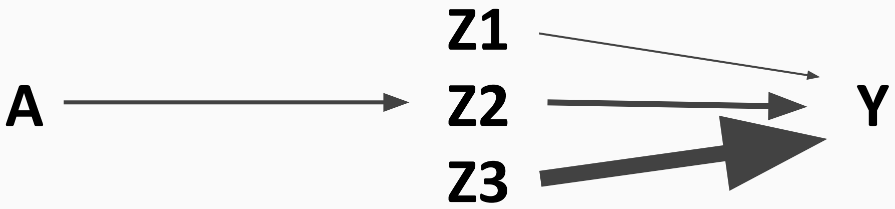
Difference: ‘effect modifier’ vs. ‘confounder’

- Effect modifier is a factor that modifies (alters) the relationship between the exposure and disease
- Provides insight into the **nature of the biologic relationship** between exposure and disease
- Thus, w**e do not want to control/adjust** for effect modification – want to explore and report
- Not a nuisance, not a threat to validity

Difference: ‘effect modifier’ vs. ‘confounder’

- A confounding factor **distorts** the measure of association relating exposure to disease because of its relationship with the exposure and outcome of interest in the population under study
- Confounding is a **nuisance** factor, does not provide biologic insight into the relationship
- It is a threat to the validity of the study
- Need to remove the effect of confounding to understand the exposure/disease relationship – we **want** to control/adjust for it

Effect measure modification: visualized

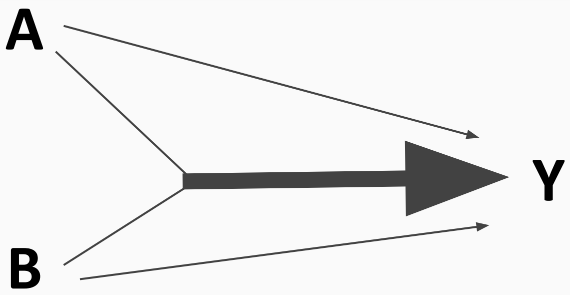


The effect of ***A*** on ***Y*** is modified by levels of ***Z***

THIS IS NOT A DAG!!

Note: EMM cannot easily be shown in a DAG

Interaction: visualized



THIS IS NOT A DAG!!

- The effect of ***A*** and ***B*** on ***Y*** interact
- Given A or B alone has less of an effect than given both together

Note: Interactions cannot easily be shown in a DAG

Effect measure modification and confounding

- Confounding is a nuisance effect that we want to ideally remove completely to isolate causal effects
- Effect measure modification is describing important variation of the exposure - outcome effect in levels of a third variable
  - We should report this
- If a variable is modifying the exposure effect on the outcome, it cannot be part of confounding based on causal structures!

Confounding and effect measure modification

- The causal conception of **confounding** must happen **before** the exposure (open *backdoor path…*)
    - Temporality is crucial
- **Effect measure modification** can only happen **after** exposure
    - To evaluate whether confounding or effect measure modification is present ***cannot*** be decided solely based on inference from the data!
    - Cannot test for this

Wrap up:
In the analysis: stratification and pooling

- Divide the data into strata according to categories of the confounder
- Within each stratum, calculate stratum-specific measures of association
- If appropriate, pool information over all strata by calculating a weighted average of the stratum-specific measures of association
  - (e.g. Mantel-Haenszel formula)
- Assumption: Constant effect across all strata

In the analysis: regression modeling

- Model relationship of exposure, outcome and other covariates
- Estimates the dependent variable based on a function of the explanatory variable(s)
- Type of regression model depends on type of data & form of dependent variable
  - Linear, logistic, Cox proportional hazards, Poisson, etc.
- Many use regression, few understand what it means… → excursion.

What is a statistical model?

- Mathematical description of relationship between variables
- Relationship between:
  - **Dependent variable** (our outcome, disease)
  - **Independent (explanatory) variable(s)** (our exposure, treatment)

Regression model

- Estimates (predicts) the **dependent variable** based on a function of the **explanatory variable(s)**
- Type of regression model depends on
  - Type of data
    - Count data, person-time data, repeated measures, etc.
  - Form of dependent variable
    - Binary, linear, ordinal, etc.

Regression model

Multifunctional tool used to:

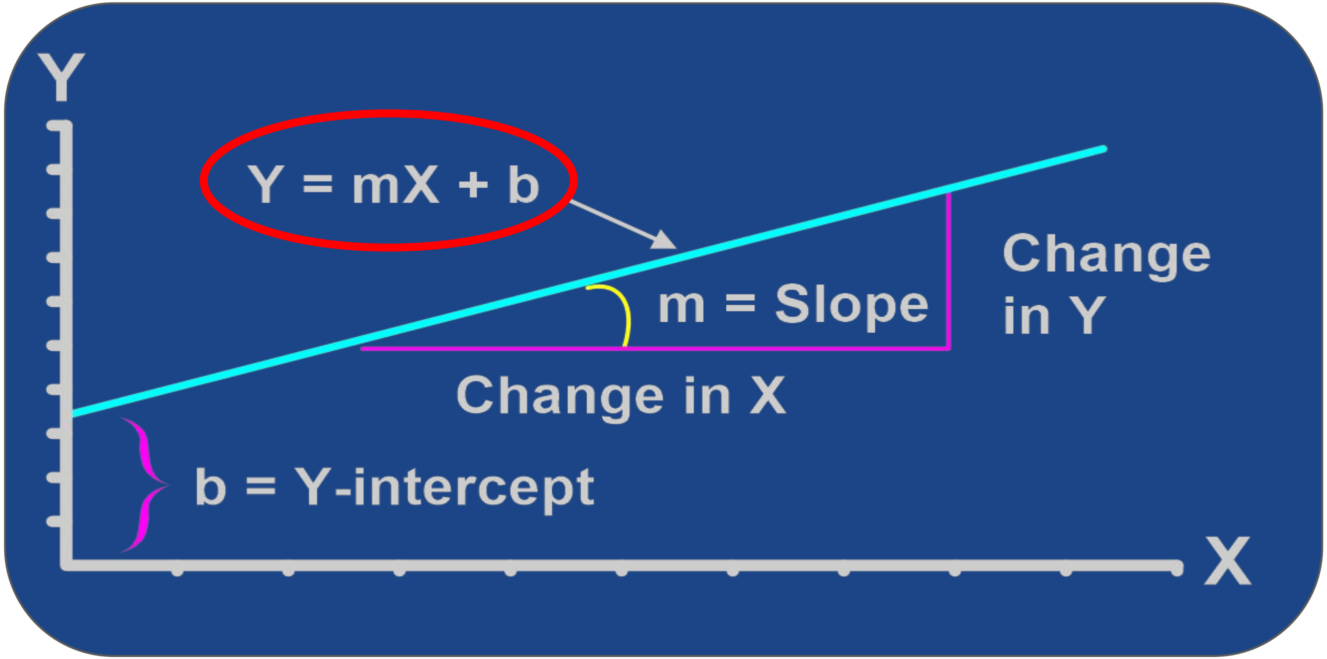
1. Estimate (causal) effects
    - Requires pre-defined underlying causal structure (DAGs)
2. Predict outcome
3. Learn from the given data
    - hypotheses generation, exploratory, “descriptive”

→ **How to use a regression model solely depends on the scientific question!**

Regression model: examples

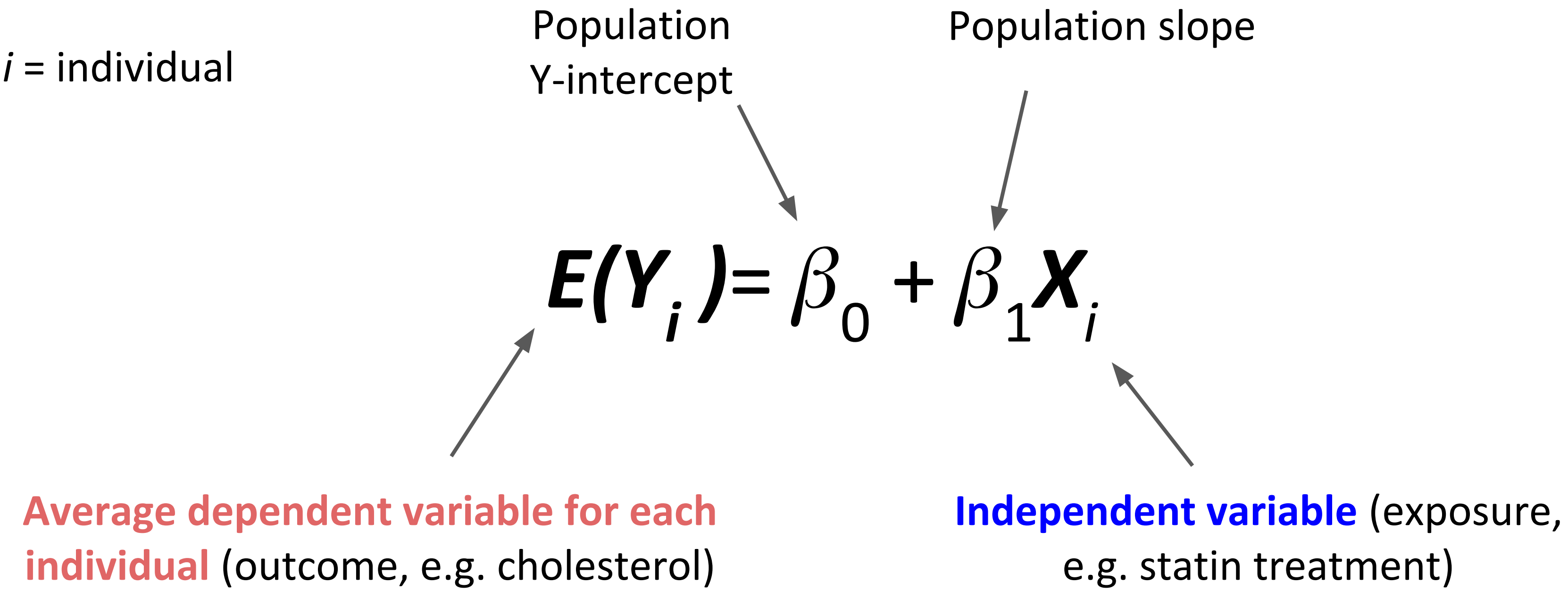
- Continuous outcome = linear regression
  - Example: systolic blood pressure values, weight
- Dichotomous outcome (yes/no) = logistic regression
  - Example: myocardial infarction, death
- Time-to-event = Cox proportional hazards model
  - Example: survival after treatment, time to death

Simple linear regression model



Relationship between variables is a linear function

Simple linear regression model: Deterministic part



i = individual

Population Y-intercept

Population slope

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Average dependent variable for each individual (outcome, e.g. cholesterol)

Independent variable (exposure, e.g. statin treatment)

Simple linear regression model: Stochastic part

i = individual

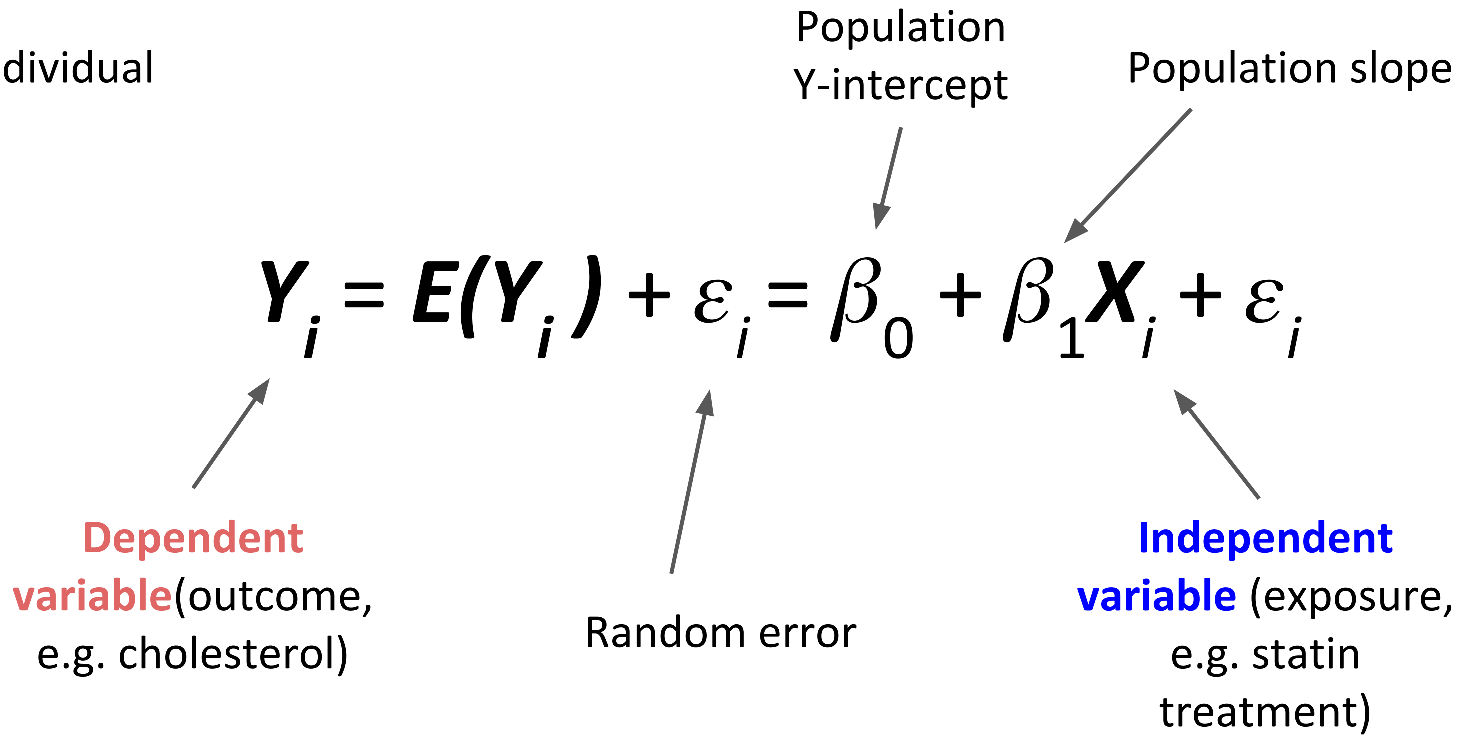
$$Y_i = E(Y_i) + \varepsilon_i$$

Dependent variable(outcome,
e.g. cholesterol)

Random error

Simple linear regression model

i = individual



Population Y-intercept

Population slope

$$\boldsymbol{Y_i} = \boldsymbol{E(Y_i)} + \varepsilon_i = \beta_0 + \beta_1 \boldsymbol{X}_i + \varepsilon_i$$

Dependent variable(outcome, e.g. cholesterol)

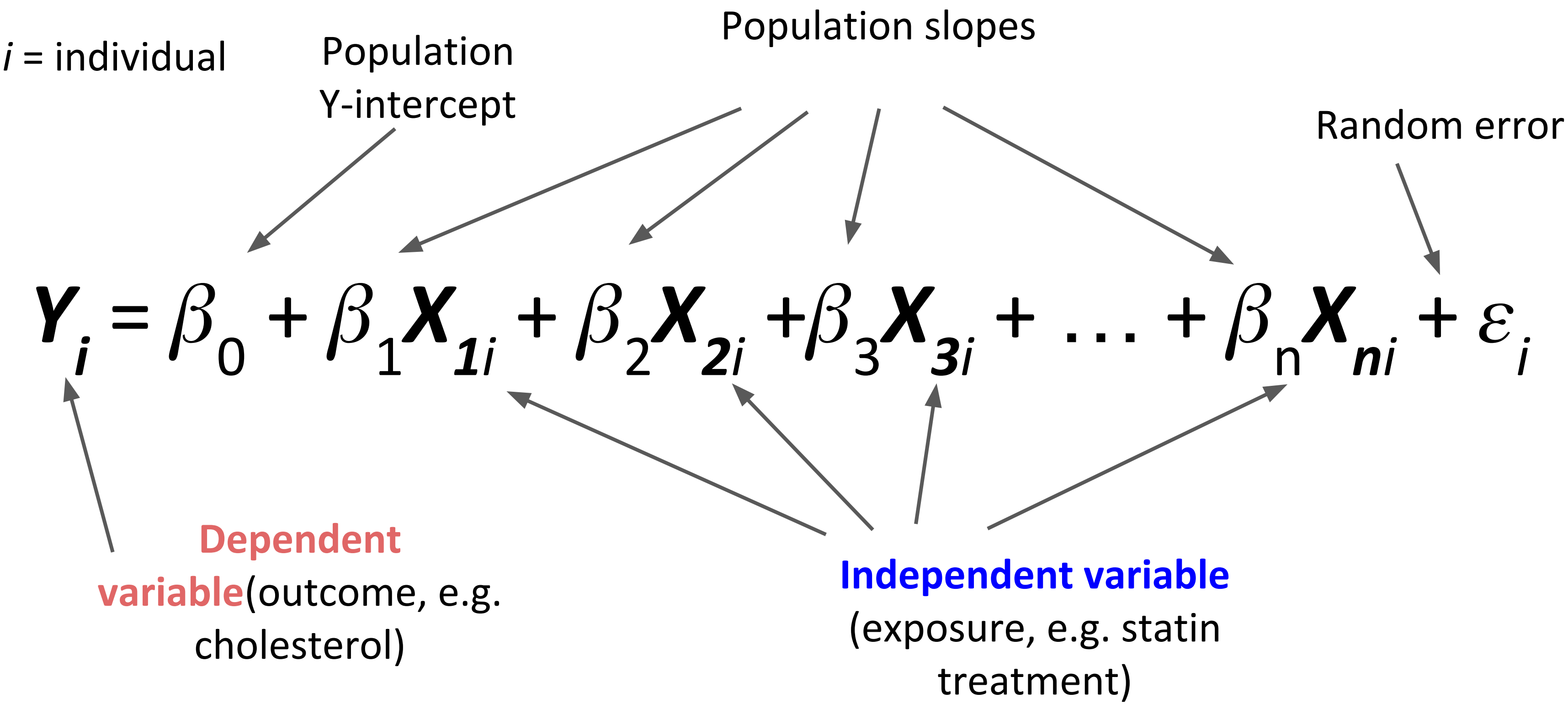
Random error

Independent variable (exposure, e.g. statin treatment)

Multiple or *multivariable* linear regression

$$\boldsymbol{Y_i} = \beta_0 + \beta_1 \boldsymbol{X}_{\boldsymbol{1}i} + \beta_2 \boldsymbol{X}_{\boldsymbol{2}i} + \beta_3 \boldsymbol{X}_{\boldsymbol{3}i} + \ldots + \beta_{\mathrm{n}} \boldsymbol{X}_{\boldsymbol{n}i} + \varepsilon_i$$

Multiple or *multivariable* linear regression



i = individual

Population Y-intercept

Population slopes

Random error

$$\boldsymbol{Y_i} = \beta_0 + \beta_1 \boldsymbol{X_{1i}} + \beta_2 \boldsymbol{X_{2i}} + \beta_3 \boldsymbol{X_{3i}} + \ldots + \beta_n \boldsymbol{X_{ni}} + \varepsilon_i$$

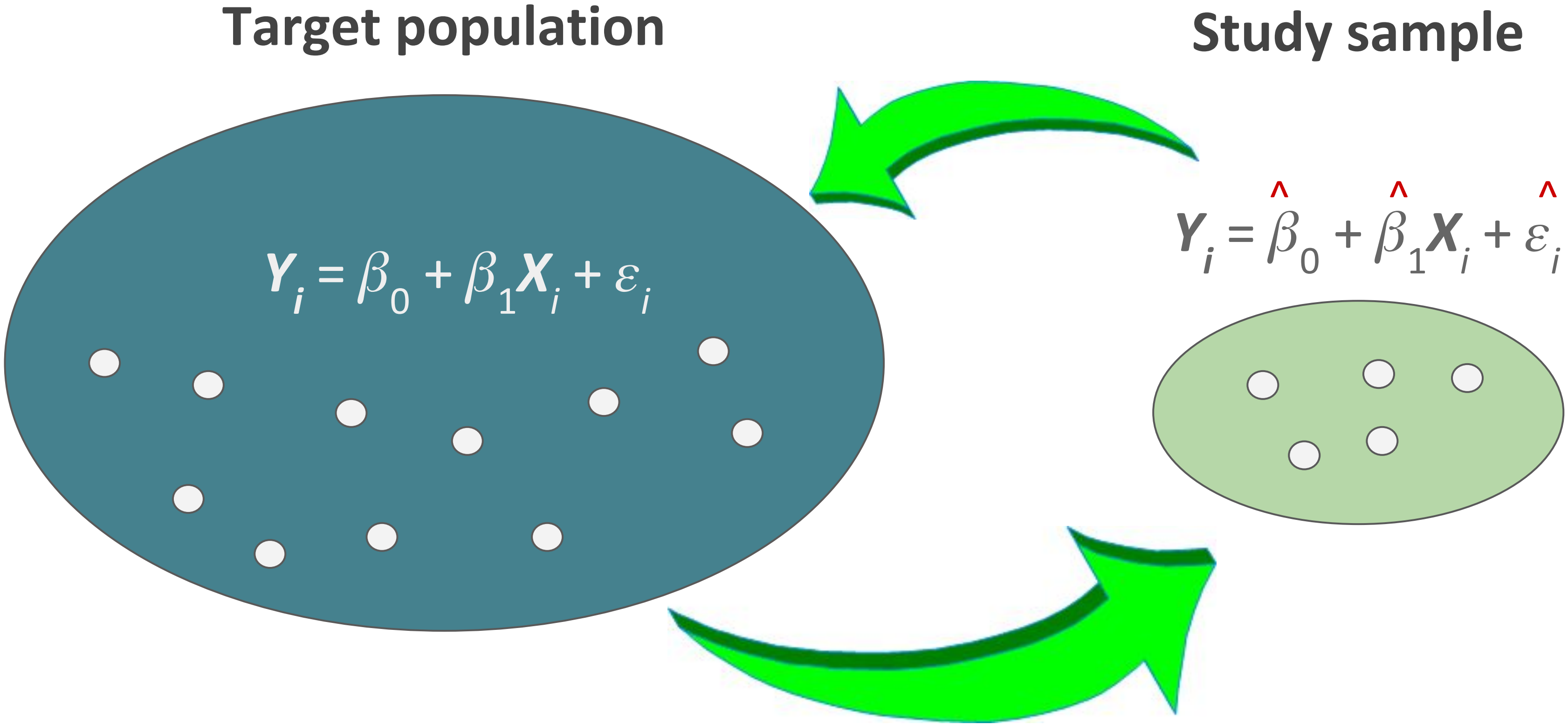
Dependent variable(outcome, e.g. cholesterol)

Independent variable (exposure, e.g. statin treatment)

Multivariate vs. multiple or multivariable regression

- **Multiple or multivariable:**
  - A model with ***multiple independent variables*** (=multivariable) that predicts a single outcome
- **Multivariate:**
  - Modeling of data wherein an ***outcome*** is measured for the same individual at ***multiple time points*** (repeated measures), or
  - Modeling of ***more than one outcome event*** (nested, clustered data)

Population sample and regression



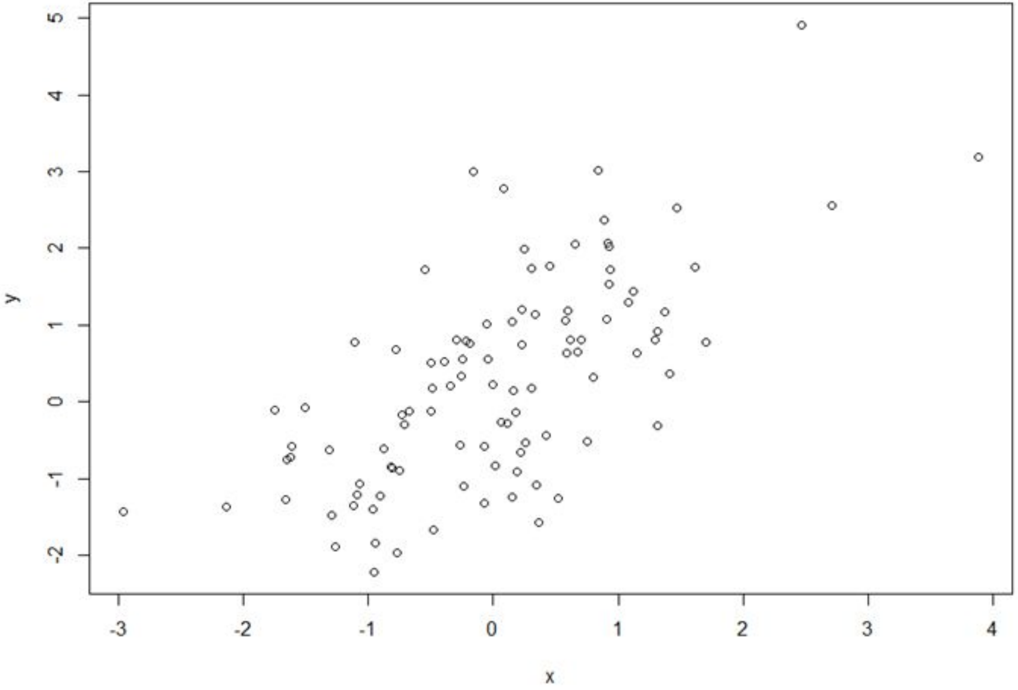
Target population

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

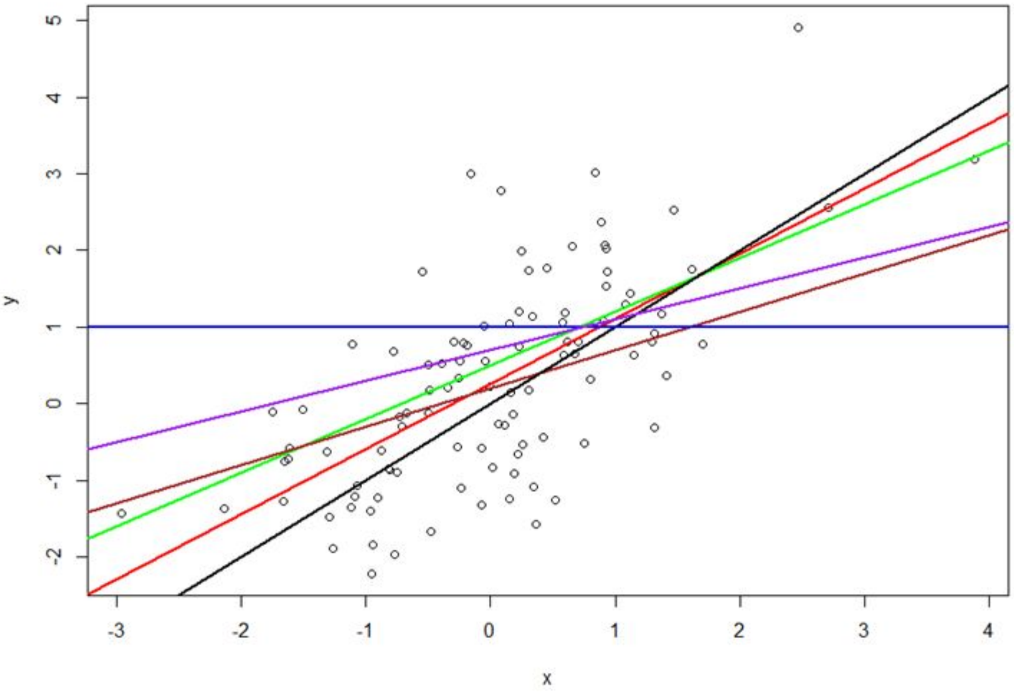
Study sample

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$

How can we estimate the best line?



How can we estimate the best line?

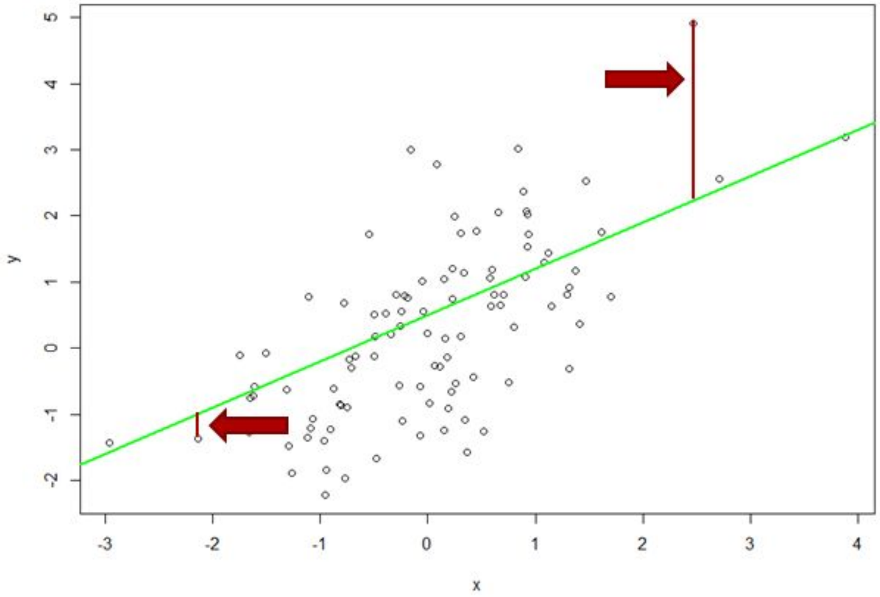


Infinite possible lines
We need the “best”
line

Sum of least squares

$$y_i = f(x_i) + \varepsilon_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$$

$$\varepsilon_i = y_i - f(x_i) = y_i - (\beta_0 + \beta_1 * x_i)$$



Find the y-intercept and slope that minimize this quantity:

$$\sum_i^n \varepsilon_i^2 = \sum_i^n (y_i - f(x_i))^2 = \sum_i^n (y_i - (\beta_0 + \beta_1 * x_i))^2$$

Coefficient interpretation

Slope ($\hat{\beta}_1$):

- Expected change in the average of ***Y*** for each one unit increase in ***X***
  - ○ If $\hat{\beta}_1$= 0.85, then ***Y*** is expected to increase by 0.85 on average for each one unit increase in ***X***

Y-intercept ($\hat{\beta}_0$):

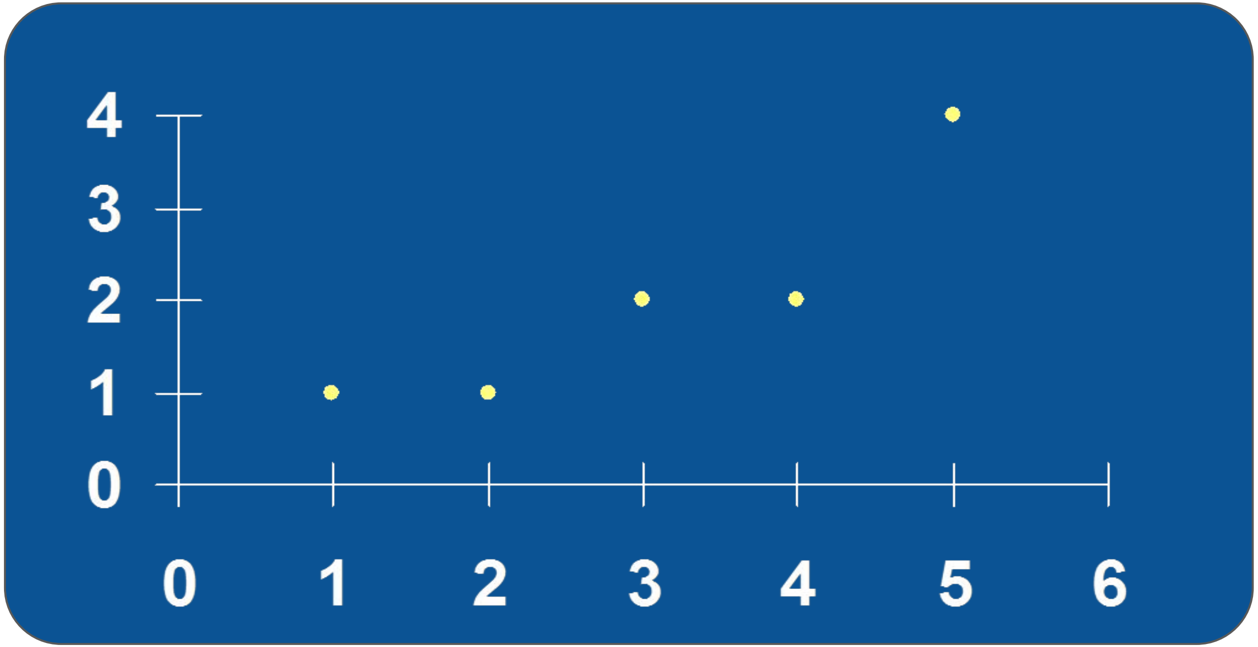
- Average value of ***Y*** when ***X*** = 0
  - ○ If $\hat{\beta}_0$= 0.25, then the average ***Y*** is expected to be 0.25 when ***X*** is 0

Parameter estimation example

- What is the relationship between mothers’ estriol level and the birthweight of their children?

Estriol (mg/24h)	Birthweight (g/1000)
1	1
2	1
3	2
4	2
5	4

Scatterplot: birthweight by estriol levels



Coefficient interpretation

Slope ($\hat{\beta}_1$):

- Birthweight (***Y)*** is expected to increase on average by 0.7 ($\hat{\beta}_1$) units for each one unit increase in estriol (***X)***

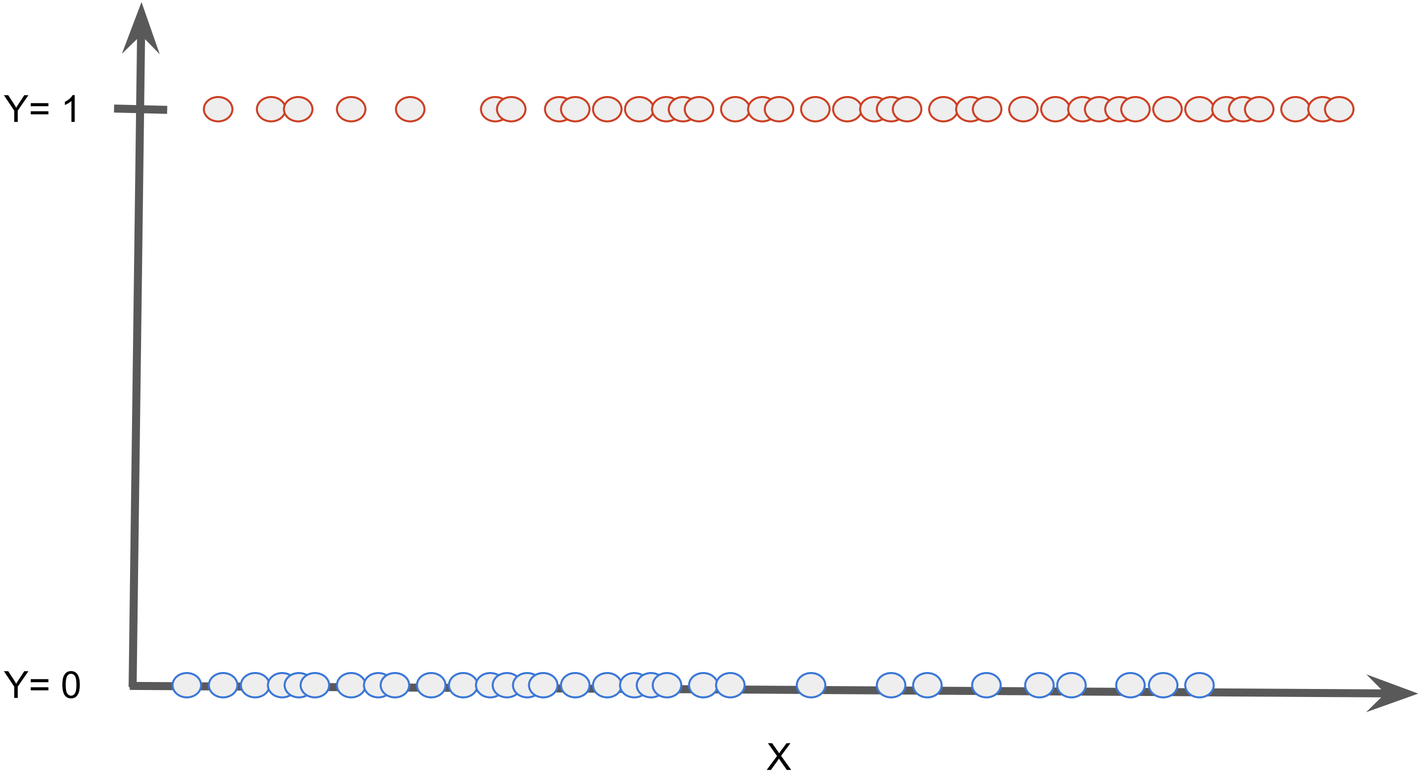
Y-intercept ($\hat{\beta}_0$):

- The average Birthweight (***Y***) is expected to be - 0.10 ($\hat{\beta}_0$) units when estirol (***X***) = 0
  - Difficult to explain as we extrapolating in areas with no biological plausibility (i.e., an estriol level in women of 0 is not plausible)

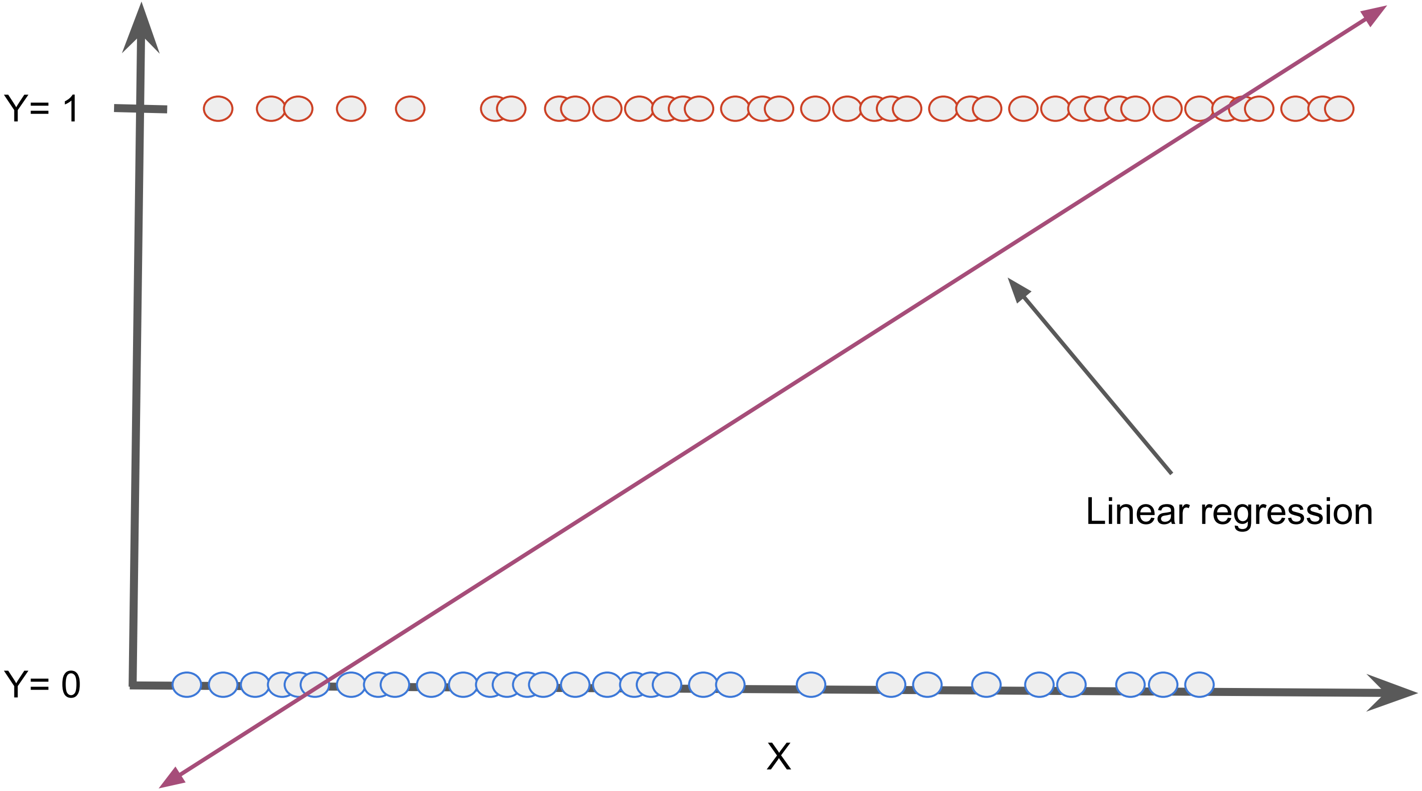
Logistic regression model

- Generalized linear model
- Regression model able to describe the relationship between a dichotomous dependent variable and one/more than one independent variables
- Why we can’t use a linear regression model?

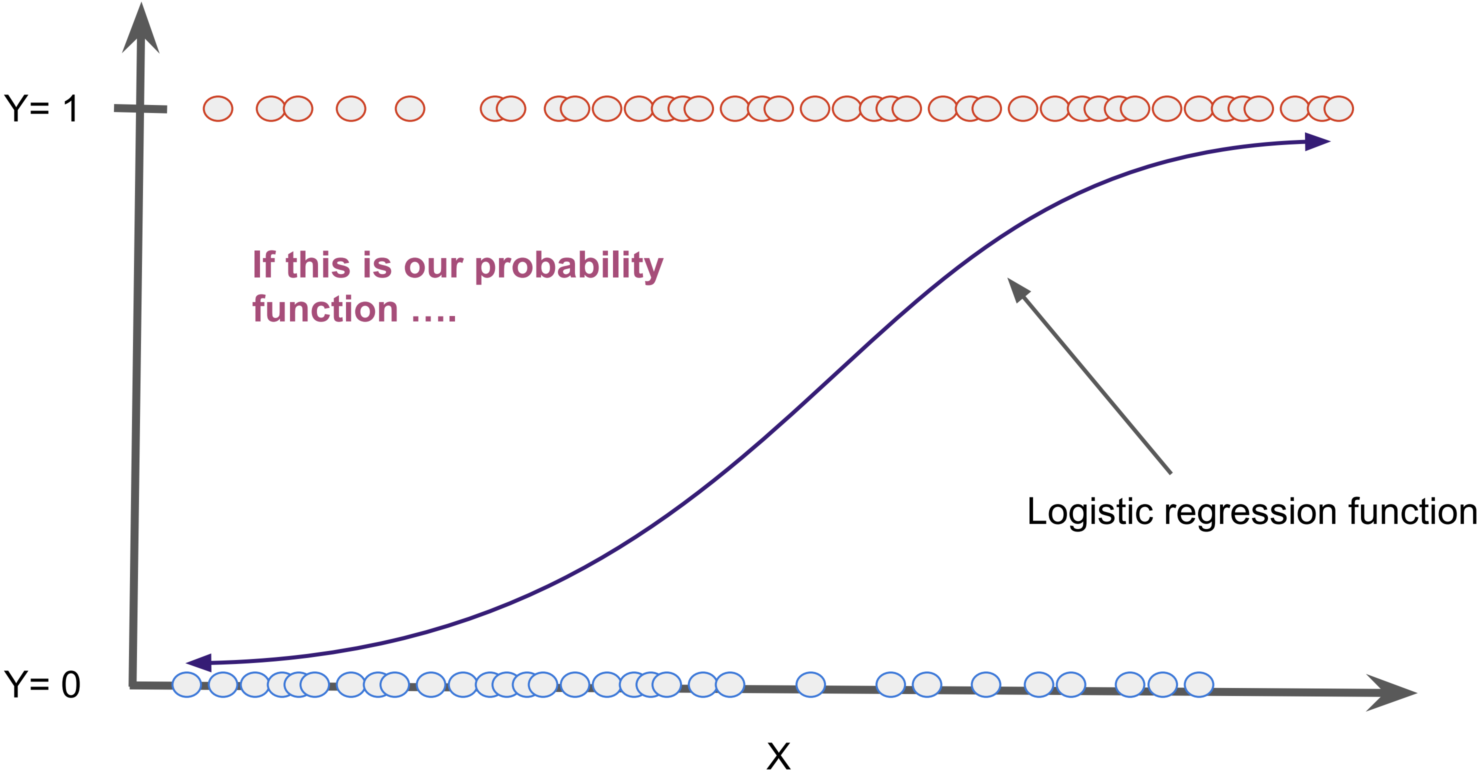
Dichotomous outcome *Y(1,0)*, predictor *X*



Dichotomous outcome *Y(1,0)*, predictor *X*



Dichotomous outcome *Y(1,0)*, linear predictor *X(x)*



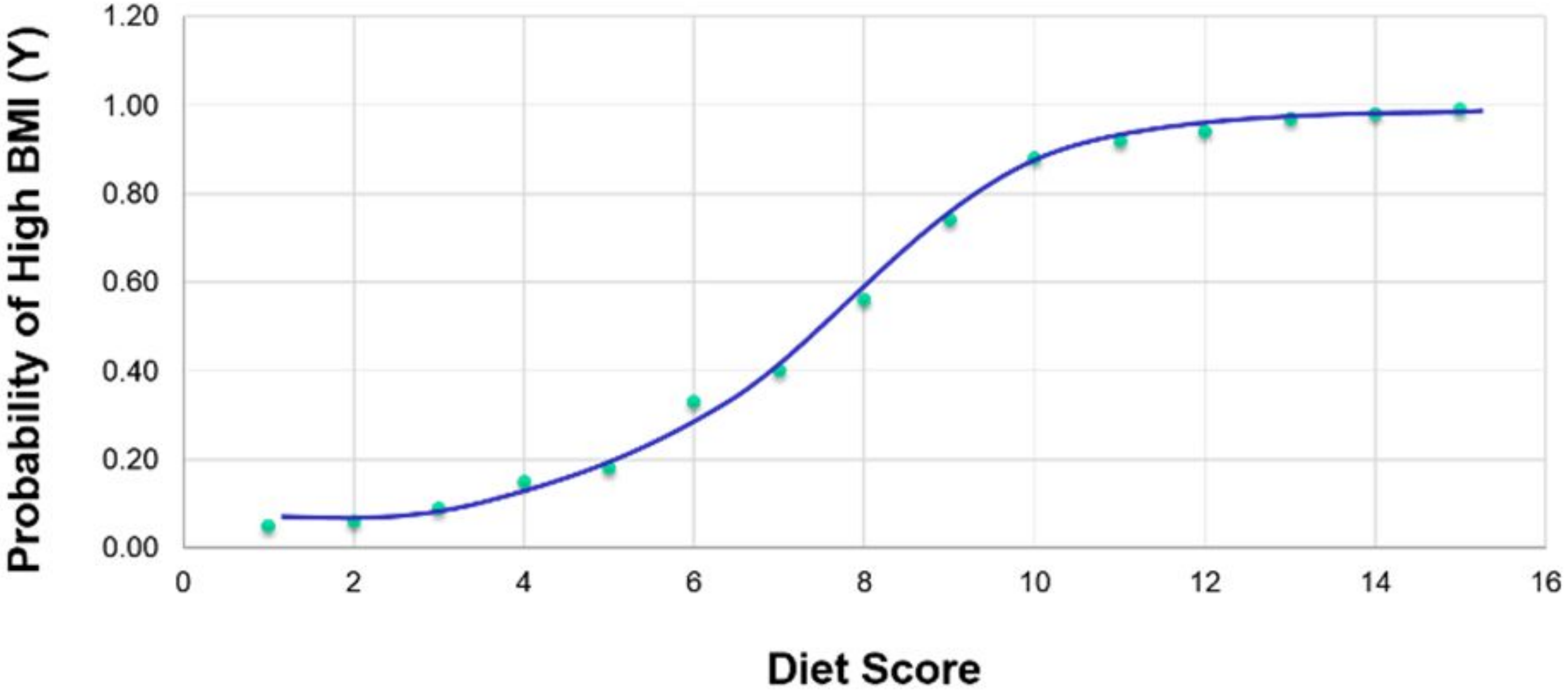
If this is our probability function ….

Logistic regression function

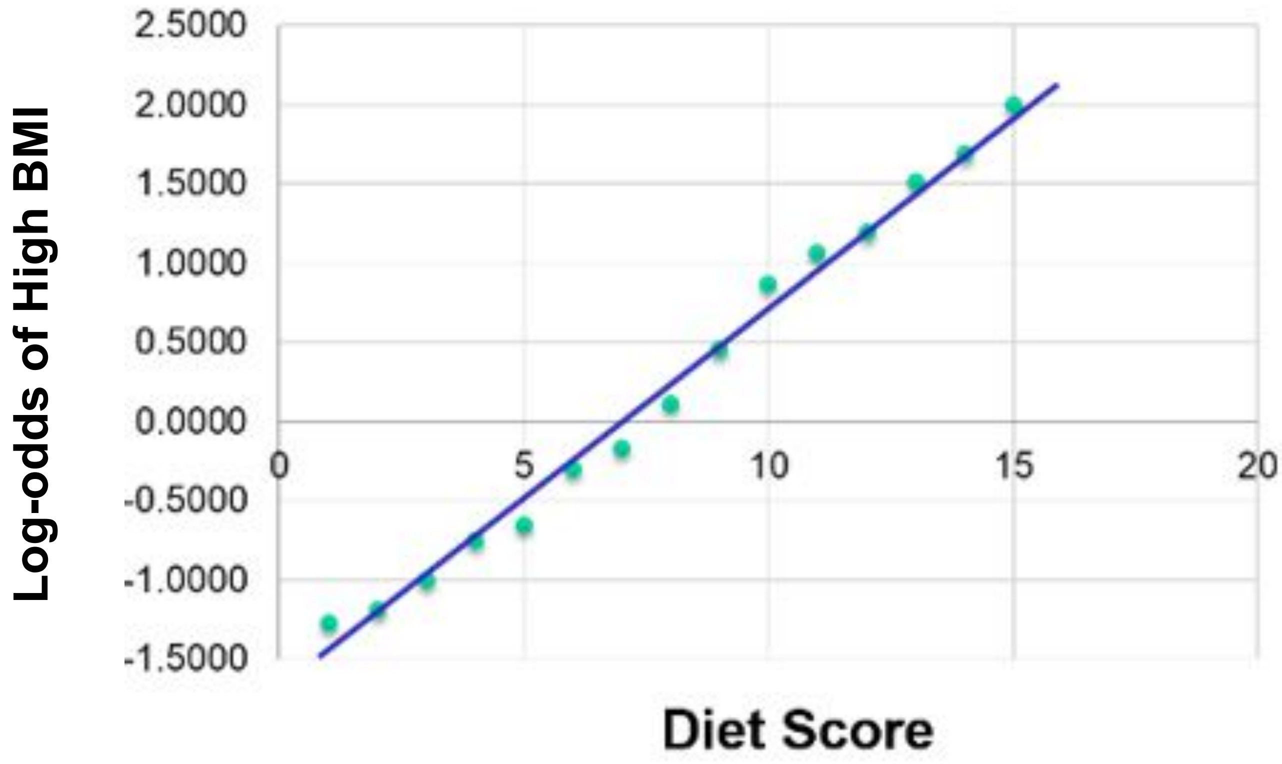
Logistic regression model

$$ln\left(\frac{Prob(Y=1)}{1-Prob(Y=1)}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k$$

Logistic regression model: example



Logistic regression model: example



Coefficient interpretation - binary X_1 (simple case)

- If X_1 is binary, e.g. drug use yes, no (coded as 1/0):
  - **ß$_0$** corresponds to the **log odds** of outcome for X_1=0, i.e. non-drug user
    - Because if X_1=0: $\text{Logit}(P(Y_i = 1)) = \beta_0 + \beta_1 * 0 = \beta_0$
  - **ß$_1$** corresponds to the **log odds ratio** between X_1=1 and X_1=0
    - Because if X_1=1: $\text{Logit}(P(Y_i = 1)) = \beta_0 + \beta_1$
    - β_1=logit drug user - logit non-drug user = log(OR)
    - Thus, **e^{ß1}** corresponds to the **odds ratio** between X_1=1 and X_1=0

Coefficient interpretation - continuous X_1

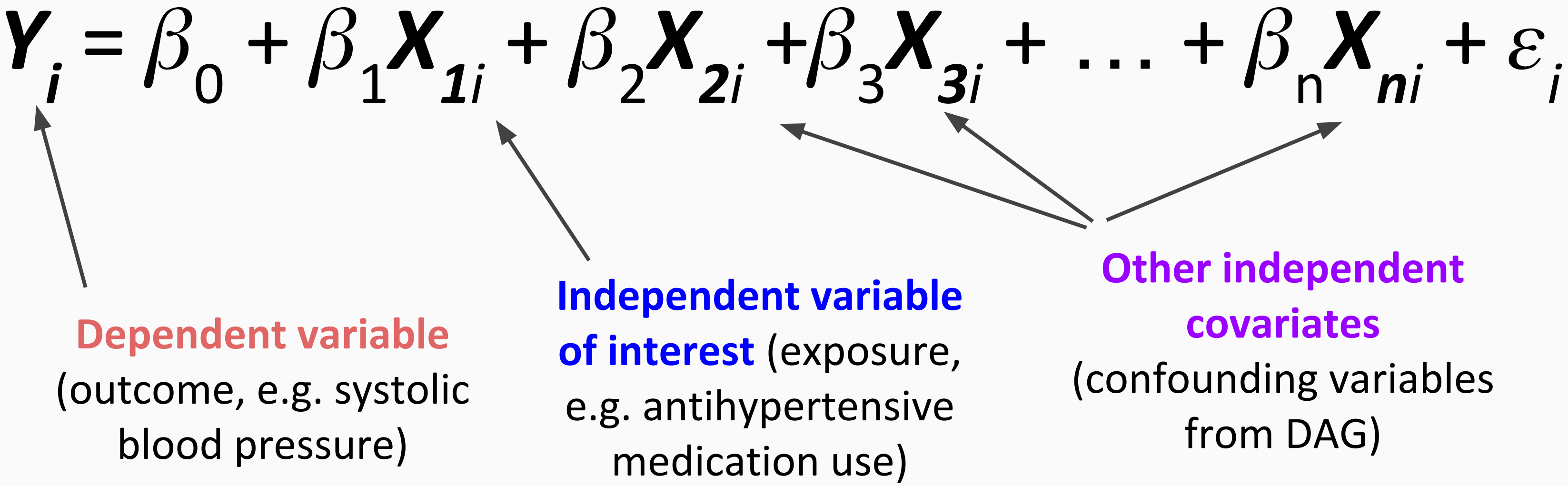
- If X_1 is continuous, e.g. diet score (0 to 50)
  - $ß_0$ corresponds to the **log odds** of outcome for X_1=0, i.e. score=0
  - $ß_1$ corresponds to the **log odds ratio** for 1 unit increase in X_1
    - Thus, $e^{ß1}$ corresponds to the **odds ratio** per 1 unit increase in X_1 (i.e. 1 unit increase in diet score)

Remember: it is a model!

“All models are wrong, but some are useful”

(George E. P. Box)

Back to confounding control...



$$\boldsymbol{Y_i} = \beta_0 + \beta_1 \boldsymbol{X_{1i}} + \beta_2 \boldsymbol{X_{2i}} + \beta_3 \boldsymbol{X_{3i}} + \ldots + \beta_n \boldsymbol{X_{ni}} + \varepsilon_i$$

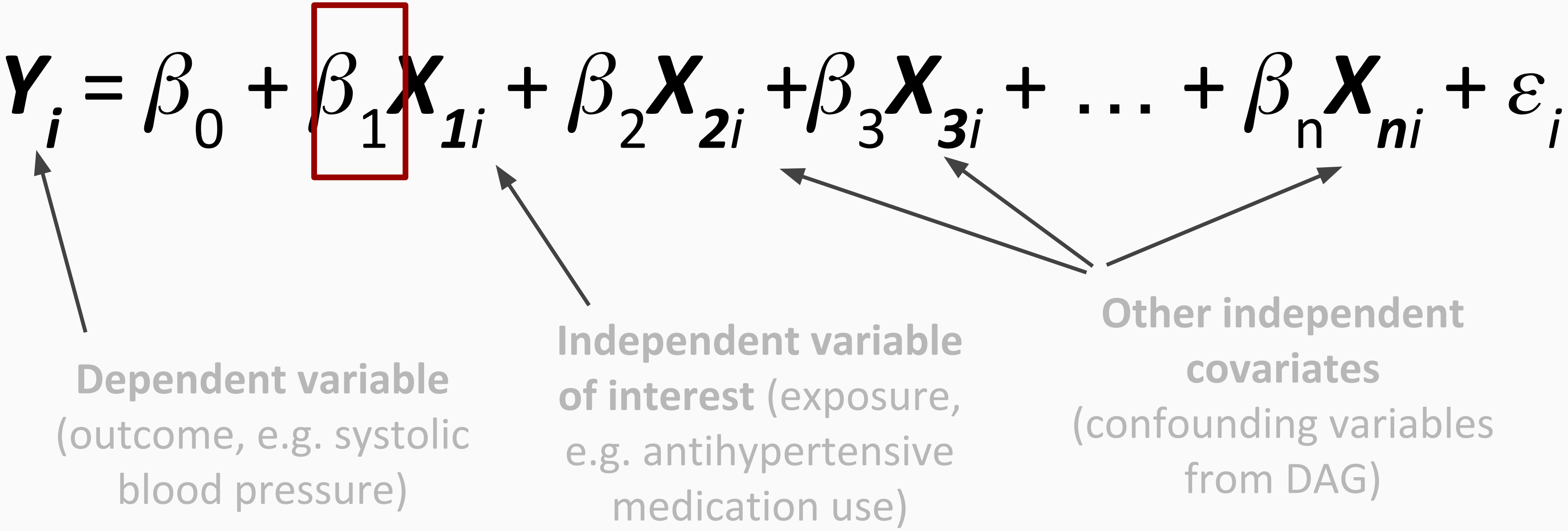
Dependent variable (outcome, e.g. systolic blood pressure)

Independent variable of interest (exposure, e.g. antihypertensive medication use)

Other independent covariates (confounding variables from DAG)

β_1: Coefficient of interest for interpretation of results

(e.g. if a person from study sample uses antihypertensives, after adjustment for potential confounders, their systolic BP is decreased on average by 20 mmHg)



$$\boldsymbol{Y_i} = \beta_0 + \boxed{\beta_1} \boldsymbol{X_{1i}} + \beta_2 \boldsymbol{X_{2i}} + \beta_3 \boldsymbol{X_{3i}} + \ldots + \beta_n \boldsymbol{X_{ni}} + \varepsilon_i$$

Dependent variable (outcome, e.g. systolic blood pressure)

Independent variable of interest (exposure, e.g. antihypertensive medication use)

Other independent covariates (confounding variables from DAG)

Wrap up:
In the analysis: regression modeling

- Type of regression model depends on type of data & form of dependent variable
  - We have shown linear and logistic, but Cox proportional hazards, ordinal logistic, Poisson, etc. work in analogous way
- Discussion questions:
  - How many confounding variables can be put in the regression model?
  - How do you choose these variables?

Methods to control for confounding

In the design of the study:

- Restriction
- Matching
- Randomization

In the analysis of the data:

- Stratification (pooling, standardization)
- Regression modeling

Final thoughts

- Makes sense to think about confounding control already in the design phase and not first when analysing data at end of study
    - This is not always possible (secondary data analysis)
- Research question should drive design (DAG), analysis and interpretation of results
- When done well, observational studies are just as credible as trials and fill in important knowledge gaps
- For statistical modeling questions, don’t hesitate to consult a biostatistician!

Thanks!

CHARITÉ

BERLIN SCHOOL OF PUBLIC HEALTH