WHY MODEL?
AN INTRODUCTION TO MODELING

Barbra Dickerman, Joy Shi, Miguel Hernán
DEPARTMENT OF EPIDEMIOLOGY

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Learning objectives
At the end of this lecture you will be able to

- Explain what a model is
- Explain why models are used in research
- Describe the most commonly used models

- Key concepts
  - Estimand, estimator, estimate
  - Consistent estimator
  - Parametric and nonparametric estimators
  - Linear and logistic regression

Plan for today

A. The need for models: A motivation

B. Types of models frequently used in epidemiology

C. Linear regression and logistic regression

A. The need for models: A noncausal motivation

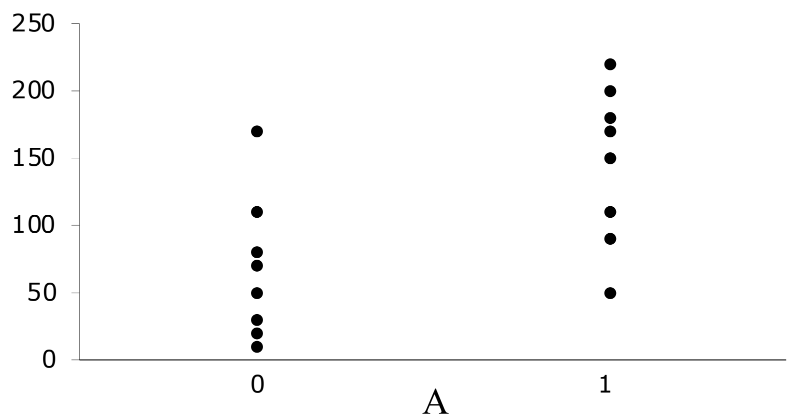
- Consider the following study
- Study population: 16 individuals living with HIV
  - Not 16,000 or 16 million
- Predictor: antiretroviral therapy A
  - Each individual receives certain level a
- Outcome: CD4 cell count at the end of follow-up Y
  - A continuous variable

Goal of our predictive analysis

- To estimate the mean of Y among individuals with treatment level $A=a$ in the population from which these individuals were randomly sampled

- This conditional population mean is represented as $E[Y|A=a]$
  - Expected value of Y given (among those with) treatment A equal to a

Dichotomous predictor



Your best estimate of the mean outcome in those receiving A=1 in the population?

0

(A) 200 — 0%

(B) 150 — 0%

(C) 100 — 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

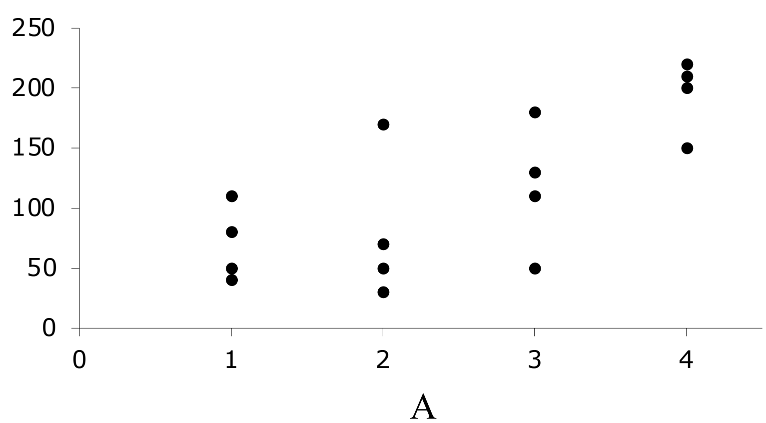
Estimand, Estimator, Estimate

- The estimand is the unknown population parameter
  - The mean of Y among those with $A=1$ in the population
- An estimator is some function of the data that is used to estimate the estimand
  - The sample average of Y among those with $A=1$
- An estimate is the result of applying the estimator to a particular data set
  - 146.25

A consistent estimator

- "The larger the sample size, the closer the sample estimate to the population estimand"
  - Formally: an estimator provides a consistent estimate $\hat{\mathrm{E}}[Y|A=a]$ of the estimand $\mathrm{E}[Y|A=a]$ if the difference $\hat{\mathrm{E}}[Y|A=a] - \mathrm{E}[Y|A=a]$ approaches zero as the sample size increases towards infinity
  - The hat ^ commonly used to refer to estimates
- Examples:
  - Consistent estimator of the population mean: the sample average
  - Inconsistent estimator of the population mean: the value of the first observation in the data
- We require that estimators be consistent

Polytomous predictor



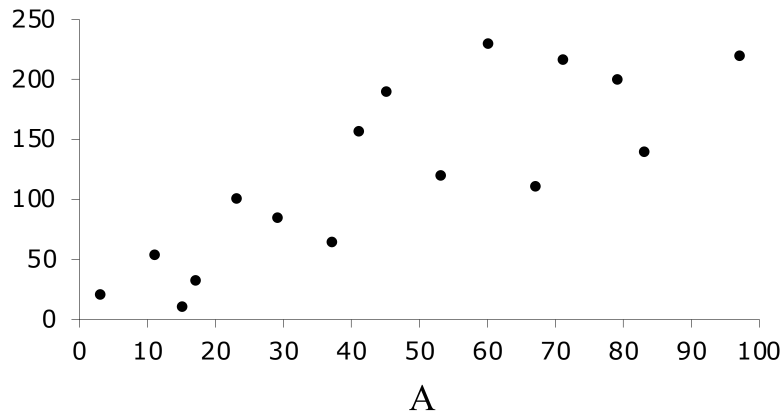
Your best estimate of the mean outcome in those receiving A=4 in the population? 0

200	**0%**
150	**0%**
100	**0%**

Discrete predictor (dichotomous or polytomous)

- As the number of categories increase, the number of individuals per category decreases
  - Variance increases
- But the sample average is still a consistent estimator of the population mean
  - The average of Y in each level of A in our sample consistently estimates the mean of Y in each level of A in the population

Continuous predictor



Your best estimate of the mean outcome in those receiving A=50 in the population? 0

200 **0%**

150 **0%**

100 **0%**

Continuous predictor

- Conceptually, a categorical variable with an infinite number of categories
  - What if there are no individuals with treatment value $A=a$?
  - Cannot use the average of Y in each level of A
- In general, it is impossible to consistently estimate $\mathrm{E}[Y|A=a]$ by using the data only
- We need to supply additional information
  - A priori knowledge that is not in the data

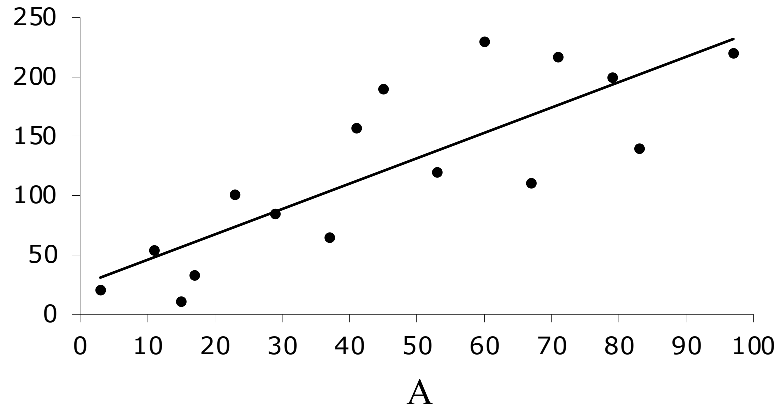
An example of a priori information

- The mean of Y follows a straight line
  - the mean of Y is directly proportional to the value of A
  - The mean of Y is θ_0 when $A=0$, and increases (or decreases) by θ_1 units per unit of A
- Or, more compactly,

  $\mathrm{E}[Y|A]= \mathrm{E}[Y|A=0] + \theta_1 A = \theta_0 + \theta_1 A$
  - θ_0 is known as the intercept
  - θ_1 is known as the slope

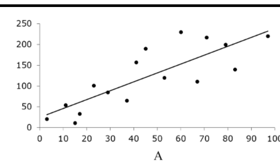
A linear model

$E[Y|A] = \theta_0 + \theta_1 A$



A linear model

$E[Y|A] = \theta_0 + \theta_1 A$



- The parameters θ_0 and θ_1 are consistently estimated by ordinary least squares estimation
- Find the line that results in the minimum sum of squared differences between each point and the straight line
  - θ_0 is estimated as the point at which the line crosses (intercepts) the vertical axis
  - θ_1 is estimated as the slope of the line

Smoothing with a linear model

$E[Y|A] = \theta_0 + \theta_1 A$

- One can use the estimates of θ_0 and θ_1 to predict the mean of Y for any possible value $A=a$, including those values not present in the data

- The mean of Y in those with $A=a$, i.e., $E[Y|A=a]$ is estimated by borrowing information from individuals with A not equal to a
  - Because ordinary least squares estimation uses all data points to find the best line

Definition of Model: a restriction on the possible values of the quantity of interest

- Consider our linear model for the conditional mean

$$E[Y|A] = \theta_0 + \theta_1 A$$

  - the mean of Y for $A=50$ cannot take any value
  - It is restricted to be in between the mean of Y for $A=40$ and the mean of Y for $A=60$
  - The restriction is encoded by parameters like θ_0, θ_1

- How do we choose the restrictions of the model?
  - Using a priori knowledge, if available, or
  - Making unverifiable (modeling) assumptions

Parametric and nonparametric estimators

- Nonparametric estimators
  - Use ONLY the data
  - Do not impose a priori restrictions on the value of the estimate
  - Example: the sample average
- Parametric estimators
  - Use the data plus a priori restrictions on the value of the estimate
  - Example: the above linear model

Nonparametric models: not really models

- No a priori restrictions because they have
  - as many parameters as quantities the model can estimate
  - also known as saturated models
- Example: for a dichotomous treatment

$$\mathrm{E}[Y|A] = \theta_0 + \theta_1 A$$

  is not really a model
  - Just says that $\mathrm{E}[Y|A=1]$ is equal to $\mathrm{E}[Y|A=0]$ plus a quantity θ_1, which is of course always true, so there is no restriction

High dimensionality: multiple covariates

- Estimate the mean $E[Y|A=a, L=l]$ in the presence of a vector of 20 covariates L
  - Number of strata grows exponentially, much faster than number of individuals
  - e.g., if 20 dichotomous covariates, there are 2^{20} (approx 1 million) possible combinations per treatment value
  - The curse of dimensionality
- Nonparametric estimators unfeasible
- Models are our only hope in high-dimensional settings

The price of modeling

- Models allow us to estimate quantities that cannot be nonparametrically consistently estimated
- But not a free lunch
  - Parametric inference correct only if the model is correctly specified
  - Model specification can be empirically checked only to some extent
- Causal inference with models requires the condition of no model misspecification

Plan for today

A. The need for models: A motivation

B. Types of models frequently used in epidemiology

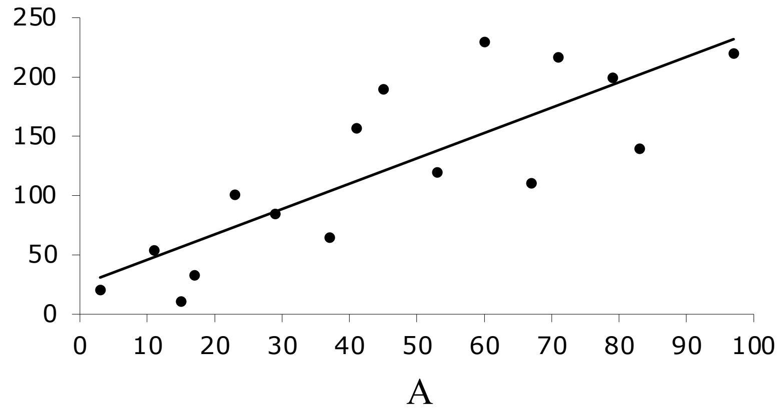
C. Linear regression and logistic regression

B. Models frequently used in epidemiology

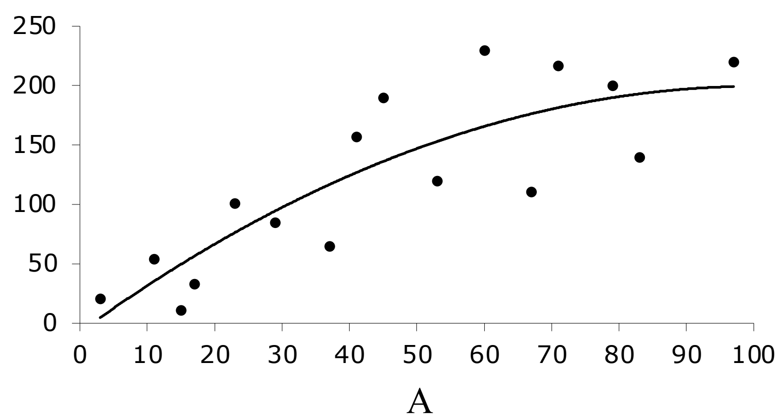
- ☐ Linear models
- ☐ Generalized linear models
- ☐ Generalized additive models
- ☐ Models for survival analysis

➢ Known as regression models

A linear model
$E[Y|A]= \theta_0 + \theta_1 A$

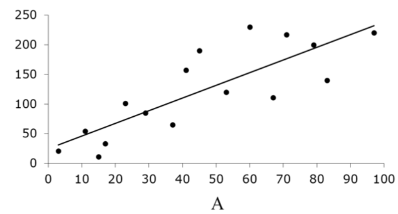


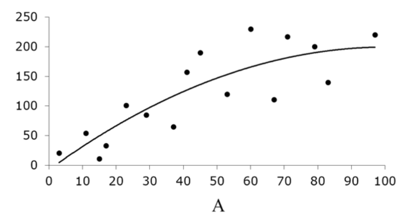
Another linear model
$E[Y|A]= \theta_0 + \theta_1 A+ \theta_2 A^2$



We need to decide which model to use to estimate $\mathrm{E}[Y|A]$

- Problem: we don't know what the true relation between A and Y is
  - a straight line or a curve?
- If we knew the shape of the true relation, it'd be easy to decide. Right?





If the true relation is a straight line, will the model $E[Y|A] = \theta_0 + \theta_1 A$ estimate $E[Y|A]$ correctly?

Yes — **0%**

No — **0%**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

If the true relation is a straight line, will the model $E[Y|A] = \theta_0 + \theta_1 A + \theta_2 A^2$ estimate $E[Y|A]$ correctly?

Yes 0%

No 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

If the true relation is a quadratic curve, will the model $E[Y|A] = \theta_0 + \theta_1 A$ estimate $E[Y|A]$ correctly?

0

Yes 0%

No 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Which linear model imposes more restrictions on (makes more assumptions about) the true relation?

The straight line

0%

The quadratic curve

0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Generalized linear models Examples

- Linear
  - Means

$$E[Y|A] = \theta_0 + \theta_1 A$$

- Log-linear
  - Counts *N*
  - Rates N/PT (PT: person-time)
    - $\ln(N/\mathrm{PT}|A) = \theta_0 + \theta_1 A$
    - $\ln(N|A) = \theta_0 + \theta_1 A + \ln(\mathrm{PT}|A)$

$$\ln(E[Y|A]) = \theta_0 + \theta_1 A$$

- Logistic
  - Proportions
  - $E[Y|A=a]=\Pr[Y=1|A=a]$

$$\mathrm{logit}(E[Y|A]) = \theta_0 + \theta_1 A$$

Extensions of generalized linear models

- Generalized linear models assume
  - a functional form
    - e.g., a linear combination of parameters and covariates
  - a statistical distribution
    - e.g., errors are independent and normally distributed with mean zero and constant variance
- Generalized linear models may relax the statistical assumptions for longitudinal data
  - GEE (generalized estimating equations) models
  - Random effects/Mixed effects models

Generalized additive models

- Like generalized linear models but they replace the linear function of covariates by a sum of functions of the covariates
  - Examples of functions: moving average, locally-weighted running mean
- $E[Y|A=a]$ may be estimated by borrowing information from some, but not all, individuals with A not equal to a
  - More "nonparametric", varying degrees of smoothing

Models for survival (failure time) data

- Need to accommodate censoring
- Parametric
  - Exponential
  - Weibull
- Semiparametric
  - Cox proportional hazards model
  - Accelerated failure time model
  - Baseline hazard is unspecified (not restricted a priori)

Plan for today

A. The need for models: A motivation

B. Types of models frequently used in epidemiology

C. Linear regression and logistic regression

C. Linear and logistic regression

- Two types of general linear models

- Linear regression for continuous outcomes
  - e.g., blood pressure

- Logistic regression for dichotomous outcomes
  - e.g., death (1: yes, 0: no)

Linear regression

- Can be used to estimate the mean Y conditional on treatment A and covariates L
- For example
  - Y is weight gain
  - A smoking cessation (1: yes, 0: no)
  - L is age (in years)

- Consider the model $\mathrm{E}[Y|A,L]= \theta_0 + \theta_1 A + \theta_2 L$
  - Parameter estimates for θ_0, θ_1, θ_2 are obtained by ordinary least squares or maximum likelihood (see Biostatistics courses)

Interpretation of θ_0 (intercept) in $E[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L$?

Mean outcome when both A and L take value zero — **0%**

Difference in mean outcome between the strata A=1 and A=0, within levels of L — **0%**

Change in mean outcome per unit of L (‘year’ here), within levels of A — **0%**

Interpretation of θ_1 in $E[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L$?

Mean outcome when both A and L take value zero — **0%**

Difference in mean outcome between the strata A=1 and A=0, within levels of L — **0%**

Change in mean outcome per unit of L (‘year’ here), within levels of A — **0%**

Interpretation of θ_2 in $E[Y|A,L] = \theta_0 + \theta_1 A + \theta_2 L$?

Mean outcome when both A and L take value zero

0%

Difference in mean outcome between the strata A=1 and A=0, within levels of L

0%

Change in mean outcome per unit of L ('year' here), within levels of A

0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Predicted values
$\hat{E}[Y|A=a, L=l]$

- The estimates of $E[Y|A=a, L=l]$ for each combination of values of treatment $A=a$ and covariates $L=l$
  - Obtained by replacing the parameters θ_0, θ_1, θ_2 by their estimates $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$
  - Example:
    - for treated individuals aged 30 years, the predicted value is
    - $\hat{E}[Y|A=a, L=l] = \hat{\theta}_0 + \hat{\theta}_1 \times 1 + \hat{\theta}_2 \times 30$
  - Residual: the difference between an individual's value of Y and the predicted value $\hat{E}[Y|A=a, L=l]$ for their combination of values of A and L

Logistic regression

- Can be used to estimate the probability of an event D conditional on treatment A and covariates L
- Consider the logistic model

$$\text{logit}\ \Pr[D=1|A,L]= \theta_0 + \theta_1 A+ \theta_2 L$$

  - D is death (1: yes, 0: no)
  - A is smoking cessation (1: yes, 0: no)
  - L is age (in years)

Interpretation of θ_1 in $logitPr[D=1|A,L]=\theta_0+\theta_1 A+\theta_2 L$?

log risk ratio of D for A=1 compared with A=0, within levels of L — 0%

log odds ratio of D for A=1 compared with A=0, within levels of L — 0%

Neither of the above — 0%

See Homework #1

- For a detailed interpretation of the parameters of logistic models and a description of the logit function

- See Biostatistics courses
  - For a description of maximum likelihood estimation to obtain parameters estimates for θ_0, θ_1, θ_2

Predicted values

- The estimates of $\text{logit}\Pr[D=1|A=a, L=l]$ for each combination of values of treatment $A=a$ and covariates $L=l$
  - Obtained by replacing the parameters θ_0, θ_1, θ_2 by their estimates $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$

- To get the probability $\Pr[D=1|A=a, L=l]$ rather than the logit of the probability, we need to do some algebra
  - In practice, computers do it for us
  - You will do it yourself in Homework #1

Why logistic regression for dichotomous outcomes?

- Because the logit transformation ensures that the predicted values will always be between 0 and 1
  - regardless of the values of the parameter estimates and the covariates
- Other transformations (e.g., probit) also have this property
  - but the logit transformation is by far the most widely used in epidemiologic research

Readings

- Chapter 11
  - Hernán MA, Robins JM. *Causal Inference: What If*.

Progress report

1. Introduction to modeling

2. Stratified analysis: Outcome regression