Multivariable analysis in clinical epidemiology

Chihaya Koriyama

August 5th, 2017

Why do we need multivariable analysis?

"**Treatment (control)**" for the confounding effects at analytical level

- ✓Stratification by confounder(s)
- ✓Multivariable / multiple analysis

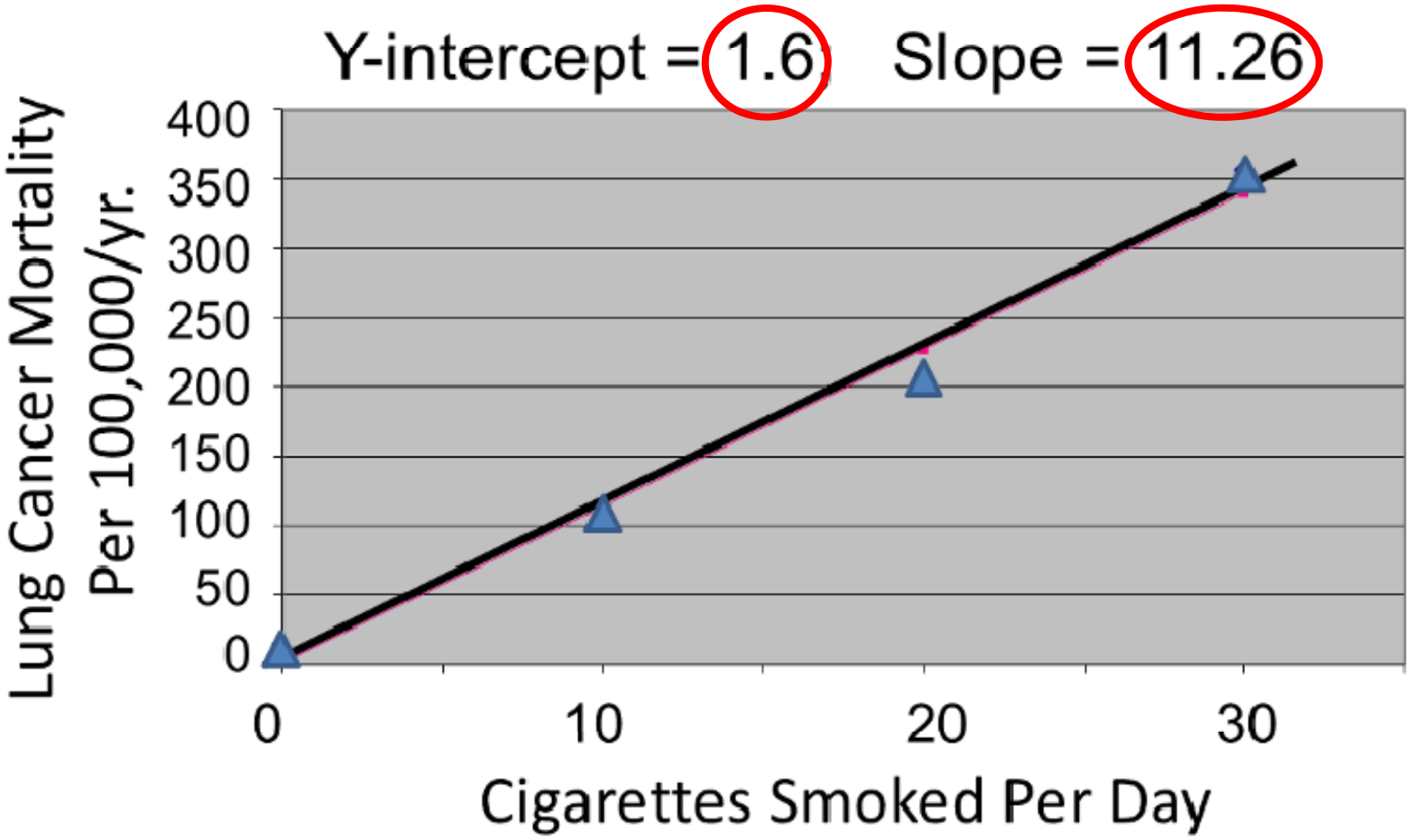
Prediction of individual risk

Regression models for multivariable analysis

Paired?	Outcome variable	Proper model
No	Continuous	Linear regression model
	Binomial	Logistic regression model
	Categorical (≥3)	Multinomial (polytomous) logistic regression model
	Binomial (event) with censoring	Cox proportional hazard model
Yes	Continuous	Mixed effect model, Generalized estimating equation
	Categorical (≥3)	Generalized estimating equation

LINEAR REGRESSION ANALYSIS

Lung cancer mortality by daily cigarettes smoked



Original data: Doll and Hill Br Med J 1956

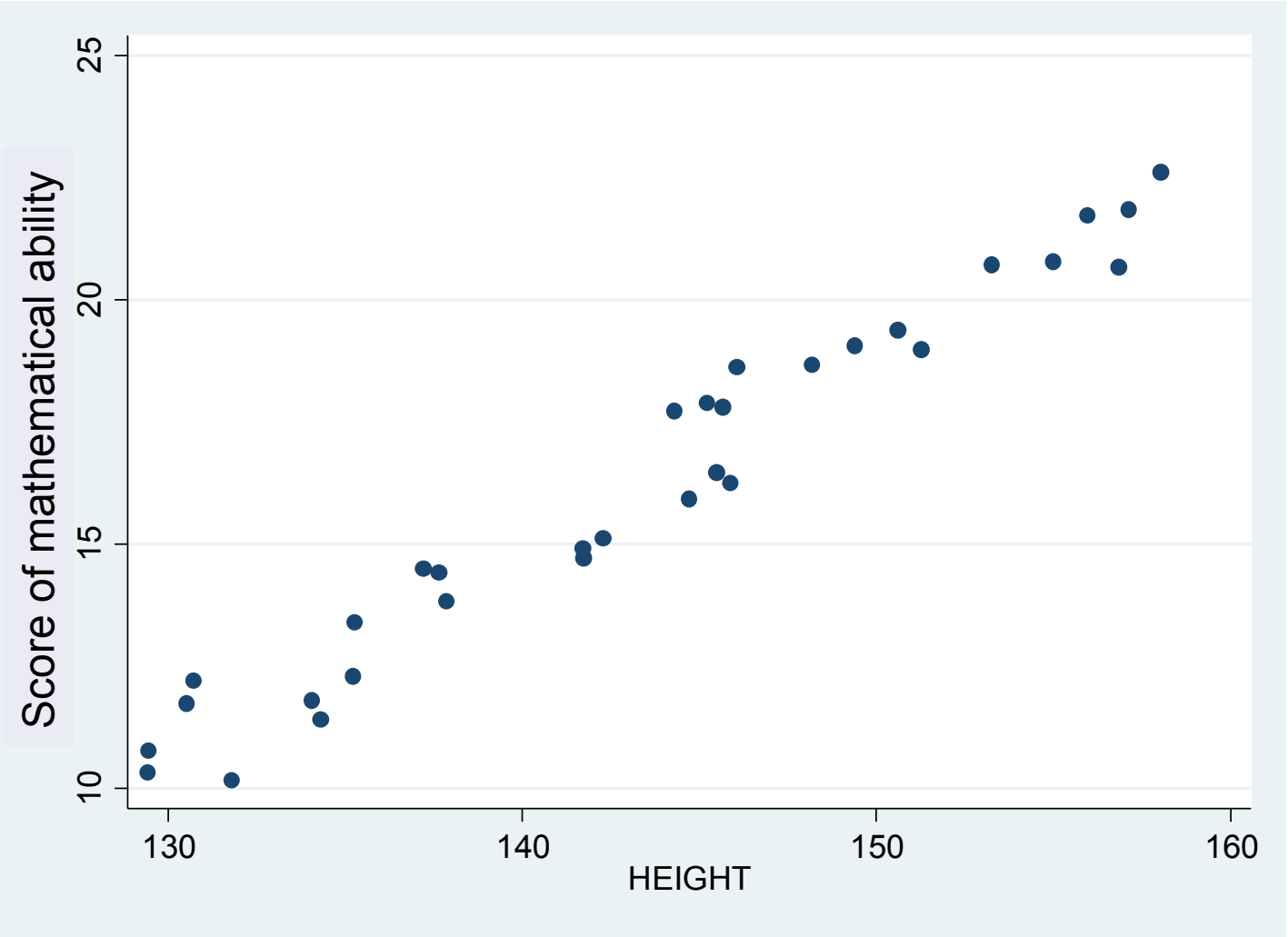
Height explaining mathematical ability!!??

```
      Source |       SS       df       MS              Number of obs   =        32
-------------+------------------------------           F(1, 30)        =    726.87
       Model |  412.7743       1  412.774322           Prob > F        =    0.0000
    Residual |  17.0365       30  .567882354           R-squared       =    0.9604
-------------+------------------------------           Adj R-squared   =    0.9590
       Total |  429.8108       31  13.8648643          Root MSE        =    .75358
```

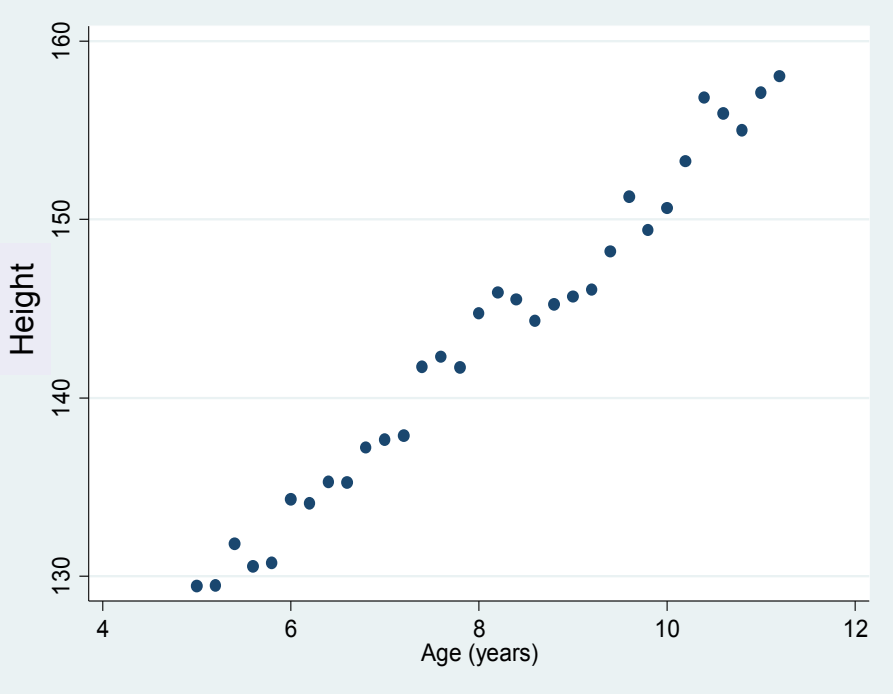
Ability score of maths

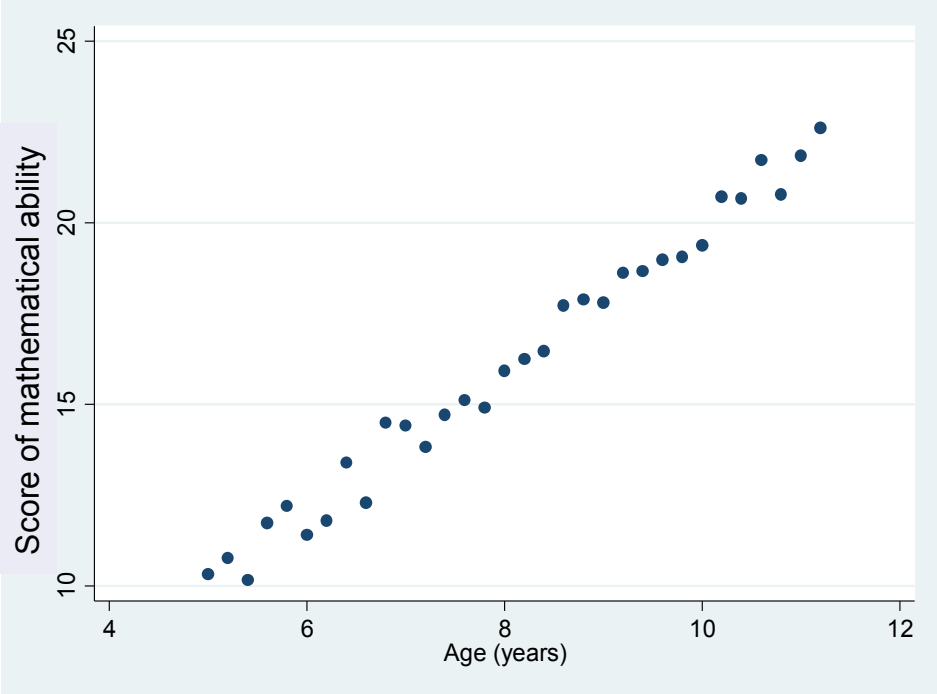
ama	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	**.4118029**	**.0152743**	**26.96**	**0.000**	**.3806086**	**.4429973**
_cons	-42.82525	2.191352	-19.54	0.000	-47.30059	-38.34992

Association between height and score of maths



Both height and ability of maths increase with age





Age is a confounding factor in the association between height and ability of maths.

How age itself influences the association between height and the ability of maths?

Let's see the equation

Ability of maths (AM) = α + β1(Height)

$\rightarrow$ AM = -42.8 + **0.41**(Height)

AM = α + β1(Height) + β2(Age)

$\rightarrow$ AM = 1.48 **- 0.01**(Height) + 2.02 (Age)

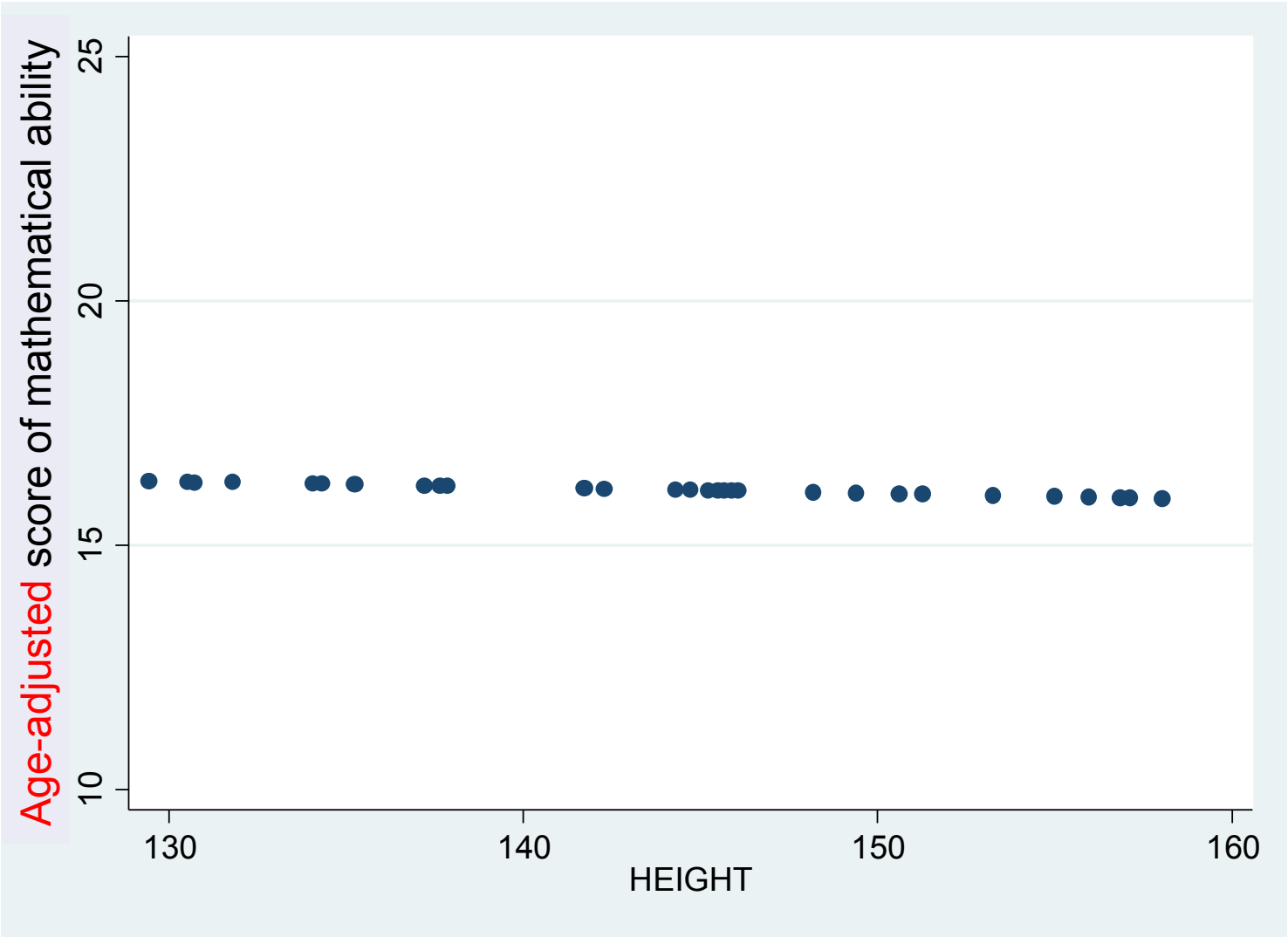
Significant association between height and the ability of maths was gone after adjusting for the effect of age

Source	SS	df	MS
Model	422.6119	2	211.305972
Residual	7.19885	29	.248236138
Total	429.81079	31	13.8648643

Number of obs	= 32
F(2, 29)	= 851.23
Prob > F	= 0.0000
R-squared	= 0.9833
Adj R-squared	= 0.9821
Root MSE	= .49823

ama	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	**-.0121303**	**.0680948**	**-0.18**	**0.860**	**-.1513998**	**.1271393**
age	2.02461	.3216095	6.30	0.000	1.366845	2.682375
_cons	1.483038	7.185946	0.21	0.838	-13.21387	16.17995

No association between height and score of maths



Interpretation of coefficients

Let's see the equation

Ability of maths (AM) = α + β1(Height)

$\rightarrow$ AM = -42.8 + **0.41**(Height)

0.41 points increase by 1cm increase of height

AM = α + β1(Height) + β2(Age)

$\rightarrow$ AM = 1.48 - **0.01**(Height) + 2.02 (Age)

0.01 points decrease by 1cm increase of height

Confounding effect: magnitude and direction of the association

Interpretation of coefficients in general

To simplify, the explanatory variable is binomial one: 1=exposed or 0=unexposed

Exposed: $Ye = \alpha + \beta(Exp=1) = \alpha + \beta$

Unexposed: $Yu = \alpha + \beta(Exp=0) = \alpha$

Difference: $Ye - Yu = \beta$

- **Coefficient estimate: difference in dependent value**

Interpretation of coefficients after log-transformation of dependent variable

The explanatory variable is binomial one:
1=exposed or 0=unexposed

Exposed: $\ln(Ye) = \alpha + \beta(Exp=1) = \alpha + \beta$

Unexposed: $\ln(Yu) = \alpha + \beta(Exp=0) = \alpha$

Difference: $\ln(Ye) - \ln(Yu) = \beta$

Ratio: $Ye / Yu = e^{\beta}$

- **Coefficient estimate: ratio of dependent value (after exponentiating)**

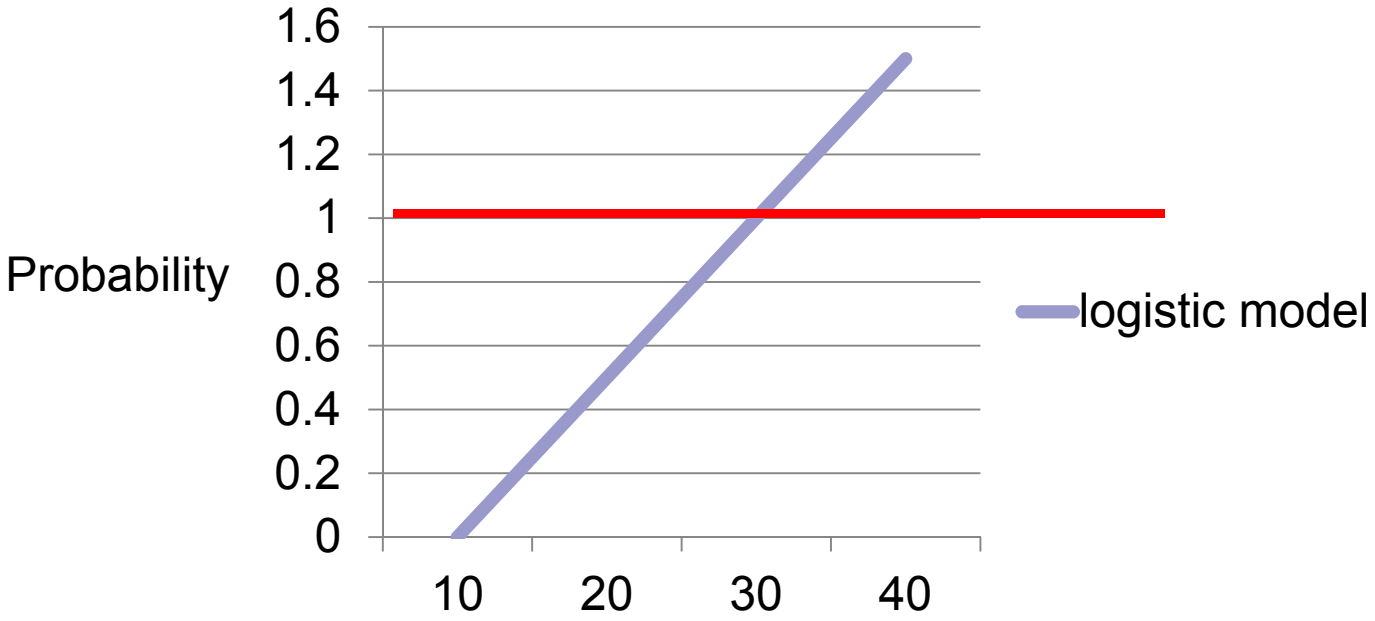
LOGISTIC REGRESSION ANALYSIS

Logistic regression analysis

- Logistic regression is used to model the probability of a binary response as a function of a set of variables thought to possibly affect the response (called covariates).

$$Y = \begin{cases} 1: \text{case (with the disease)} \\ 0: \text{control (no disease)} \end{cases}$$

One could imagine trying to fit a linear model (since this is the simplest model !) for the probabilities, but often this leads to problems:



In a linear model, fitted probabilities can fall outside of 0 to 1. Because of this, linear models are seldom used to fit probabilities.

In a logistic regression analysis, the **logit** of the probability is modeled, rather than the probability itself.

P = probability of getting disease (0～1)

$$\text{logit}(p) = \textbf{log}\left[\frac{p}{1-p}\right]$$

This transformation allows us to use a linear model.

As always, we use the natural log.

The logit is therefore **the log odds**, since odds = p / (1-p)

Logistic regression model

Now, we have the same function with linear regression model in the right side.

$$\text{logit}(p_x) = \log\left[\frac{p_x}{1 - p_x}\right] = \alpha + \beta x$$

where p_x = probability of event for a given value x, and α and β are unknown parameters to be estimated from the data.

→ **Multivariable analysis** is applicable to adjust the effect of confounding factor.

Interpretation of coefficients of logistic regression model

The explanatory variable is binomial one: 1=exposed or 0=unexposed

Exposed: $\log (O_e) = \alpha + \beta(\text{Exp}=1) = \alpha + \beta$

Unexposed: $\log (O_u) = \alpha + \beta(\text{Exp}=0) = \alpha$

Difference: $\log(O_e) - \log (O_u) = \beta$

Odds ratio: $O_e / O_u = e^{\beta}$

- **Coefficient estimate: Odds ratio (after exponentiating)**

STRATEGY FOR CONSTRUCTING REGRESSION MODELS

Basic principles

1. **Stratified analysis** should be first.
2. Determine which **confounders to include** in the model.
3. Estimate the shape of the exposure-disease relation.

Dose-response relation

1. Evaluate **interaction**

How to determine confounders: data-dependent manner

1. Start with a set of predictors of outcome based on the strength of their relation to the outcome.
2. Build a model by introducing predictor variables one at a time: check the amount of change in the coefficient of the exposure term

> 10% change: include it as a confounder

Example of a confounder (age)

Ability of maths (AM) = α + β1(Height)

$\rightarrow$ AM = -42.8 + **0.41**(Height)

AM = α + β1(Height) + β2(Age)

$\rightarrow$ AM = 1.48 **- 0.01**(Height) + 2.02 (Age)

How to determine confounders: data-independent manner

Some researchers argue that

"Without data analysis, decide confounders, important risk factors of the outcome, based on the previous studies."

How can we pick-up "important risk factors"? If there are few studies, how can we know confounders?

How many explanatory variables can we use in a model?

Model	Number of explanatory variables	Example
Linear regression model	**Sample size / 15**	Up to around 6-7 variables in **100 subjects**
Logistic regression model	**Smaller sample size of outcome / 10**	Up to 10 variables if the numbers of cases and controls are **100** and 300, respectively.
Cox proportional hazard model	**The number of event / 10**	Up to 9 variables if you have **90 events** out of 150 subjects

ATTENTION!

- When you include a categorical variable in your model, you have to count that as "the number of categories – 1".
  - For example, the variable of age group used in the previous practice, we have to count it as "two" (=3 categories -1) variables.

If you cannot recruit enough sample size

- Calculate "propensity score" which can be used for adjustment of confounding effects.

Example

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease

A Propensity Analysis

Patricia A. Gum, MD
Maran Thamilarasan, MD
Junko Watanabe, MD
Eugene H. Blackstone, MD
Michael S. Lauer, MD

Context Although aspirin has been shown to reduce cardiovascular morbidity and short-term mortality following acute myocardial infarction, the association between its use and long-term all-cause mortality has not been well defined.

Objectives To determine whether aspirin is associated with a mortality benefit in stable patients with known or suspected coronary disease and to identify patient characteristics that predict the maximum absolute mortality benefit from aspirin.

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	*P* Value
Demographics			
Age, mean (SD), y	62 (11)	56 (12)	<.001
Men, No. (%)		2167 (56)	<.001
Clinical history			
Diabetes, No. (%)		432 (11)	<.001
Hypertension, No. (%)		1569 (41)	<.001
Tobacco use, No. (%)		500 (13)	.001
Prior coronary artery c		778 (20)	<.001
Prior coronary artery by			<.001
Prior percutaneous coronary intervention, No. (%)	667 (29)	148 (4)	<.001
Prior Q-wave MI, No. (%)	369 (16)	285 (7)	<.001
Atrial fibrillation, No. (%)	27 (1)	55 (1)	.04
Congestive heart failure, No. (%)	127 (6)	178 (5)	.12
Medication use			
Digoxin use, No. (%)	171 (7)	216 (6)	.004
β-Blocker use, No. (%)	811 (35)	550 (14)	<.001
Diltiazem/verapamil use, No. (%)	452 (20)	405 (10)	<.001
Nifedipine use, No. (%)	261 (11)	283 (7)	<.001
Lipid-lowering therapy, No. (%)	775 (34)	380 (10)	<.001
ACE inhibitor use, No. (%)	349 (15)	441 (11)	<.001
Cardiovascular assessment and exercise capacity			
Body mass index, mean (SD), kg/m²	29 (5)	30 (7)	<.001
Ejection fraction, mean (SD), %	50 (9)	53 (7)	<.001
Resting heart rate, mean (SD), beats/min	74 (13)	79 (14)	<.001
Resting blood pressure, mean (SD), mm Hg			

Almost all prognostic factors (n=28) are related to aspirin use!

After matching by propensity score, the distribution of prognostic factors are similar between aspirin users and non-users.

It is just like a RCT! (pseud RCT)

Table 3. Selected Baseline and Exercise Characteristics According to Aspi
Use In Propensity

	Aspirin (n = 1351)	No Aspirin (n = 1351)	*P* Value
Demographics			
Age, mean (SD), y	60 (11)	61 (11)	.16
Men, No. (%)	951 (70)	974 (72)	.33
Clinical history			
Diabetes, No. (%)	203 (15)	207 (15)	.83
Hypertension, No. (%)	679 (50)	698 (52)	.46
Tobacco use, No. (%)	161 (12)	162 (12)	.95
Cardiac variables			
Prior coronary artery disease, No. (%)	652 (48)	659 (49)	.79
Prior coronary artery bypass graft, No. (%)	251 (19)	235 (17)	.42
Prior percutaneous coronary intervention, No. (%)	166 (12)	147 (11)	.25
Prior Q-wave MI, No. (%)	194 (14)	206 (15)	.52
Atrial fibrillation, No. (%)	21 (2)	24 (2)	.65
Congestive heart failure, No. (%)	79 (6)	89 (7)	.43

Table 4. Cox Proportional Hazards Analyses of Aspirin Use and Mortality Among Propensity-Matched Patients (n = 2702)*

Model	Hazard Ratio (95% CI)	*P* Value
Unadjusted	0.53 (0.38-0.74)	.002
Adjusted for propensity	0.53 (0.38-0.74)	<.001
Adjusted for propensity and selected variables†	0.59 (0.42-0.83)	.002
Adjusted for propensity and all covariates‡	0.56 (0.40-0.78)	<.001

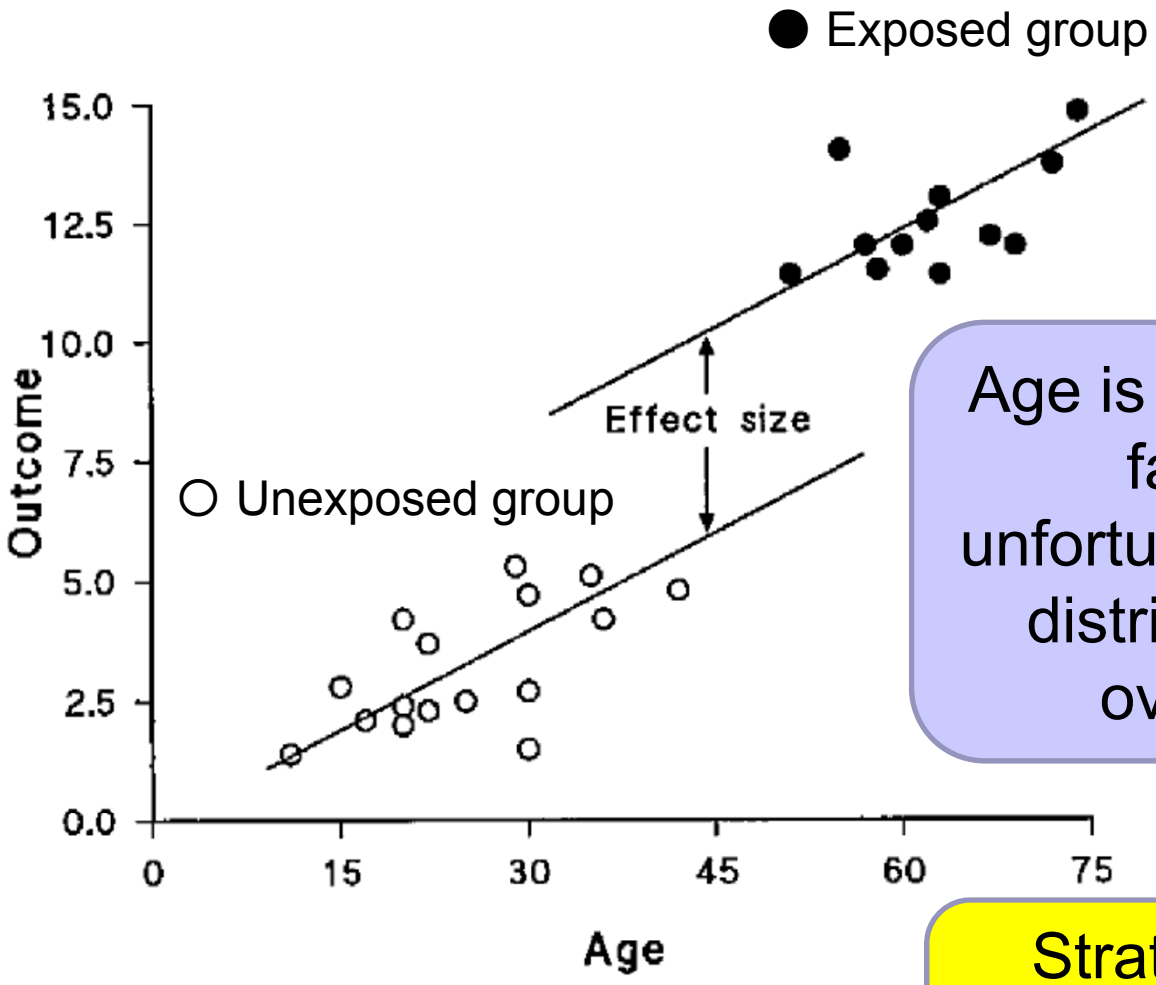
*CI indicates confidence interval.
†Selected variables included prior coronary artery disease, prior coronary artery bypass grafting, prior percutaneous intervention, and ejection fraction ≤40%.
‡For a list of covariates, see Table 2 footnote (†).

You need to include only propensity score in the model.

Control of confounding with regression model

- Compared to stratified analysis, several confounding variables can be easily controlled simultaneously using a multivariable regression model.

- Results from the regression model are readily susceptible to bias if the model is not a good fit to the data.

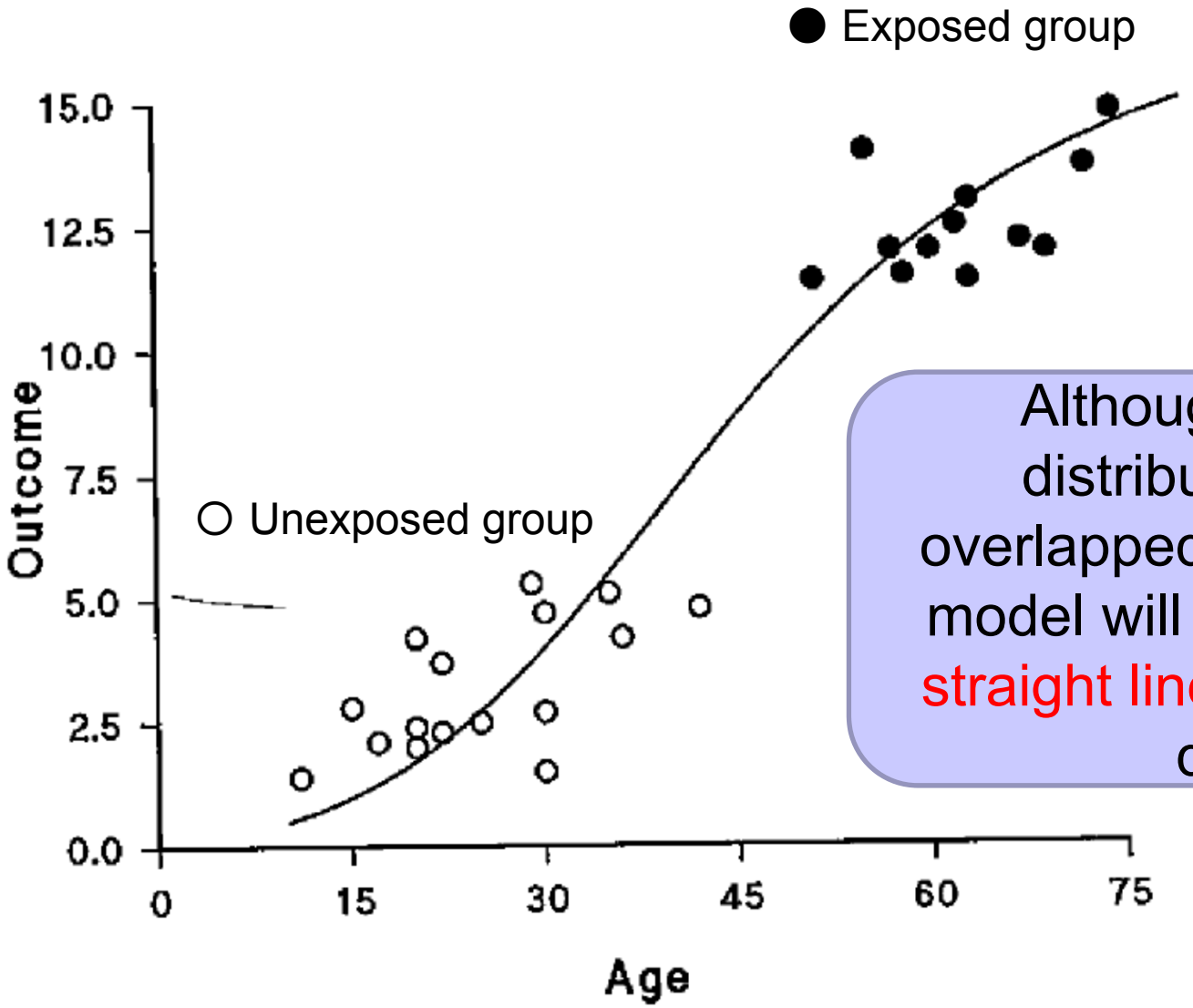


Age is a confounding factor, but unfortunately, the age distribution is not overlapped.

Stratified analysis would produce no estimate of effect.

Figure 12–4 Hypothetical example of a multivariable linear reg
data involving a dichotomous exposure variable (exposed = *solid*
open circles) and age.

Epidemiology (Rothman KJ, Oxford University Press)



Although the age distribution is not overlapped, a regression model will fit two parallel straight lines through the data.

Epidemiology (Rothman KJ, Oxford University Press)

Statistic significance vs. Clinical significance

Statistic significance ≠ Clinical significance

- P value(s) do NOT tell us the significance in clinical practice / biological importance.
- If your sample size is quite large, you may obtain a result with statistic significance. So what?

RCT of donepezil for Alzheimer's disease

Lancet. 2004 Jun 26;363(9427):2105-15.

Long-term donepezil treatment in 565 patients with Alzheimer's disease (AD2000): randomised double-blind trial.

Courtney C1, Farrell D, Gray R, Hills R, Lynch L, Sellwood E, Edwards S, Hardyman W, Raftery J, Crome P, Lendon C, Shaw H, Bentham P; AD2000 Collaborative Group.

Auth

Abstrac

BACKGI ...eimer's disease. We aimed to determin ...ogical symptoms, carers' psycholo ...g?

METHOI ...eriod in which they were randomly allocated donepezil (5 mg/ ...no completed this period were rerandomised to either donepezil (5 or 10 mg/day) or placebo, with double-blind treatment co ... defined by loss of either ... bas assessments were soug ... all patier

FINDINGS: Cognition averaged 0.8 MM ... points better (0.5-1.6; p<0.0001) with d ... institutionalisation (42% vs 44% at 3 ye ...onal care in the donepezil group compared ... institutional care was 0.96 (95% CI 0.7 ...l and psychological symptoms, carer psy ... mg donepezil.

INTERPRETATION: Donepezil is not cost effective, with benefits below minimally relevant thresholds. More effective treatments than cholinesterase inhibitors are needed for Alzheimer's disease.

Cognition averaged 0.8 MMSE points better (95%CI 0.5-1.2; p<0.0001) and functionality 1.0 BADLS points (0.5-1.6; p<0.0001) with donepezil over the first 2 years.

Donepezil is not cost effective, with benefits below minimally relevant thresholds. More effective treatments than cholinesterase inhibitors are needed for AD.