



# Multivariable analysis: A brief introduction

**Chihaya Koriyama**

Department of Epidemiology & Preventive Medicine

Kagoshima University

January 7<sup>th</sup>, 2021





# Why do we need multivariable analysis?

“Treatment “ for the **confounding effects** at analytical level

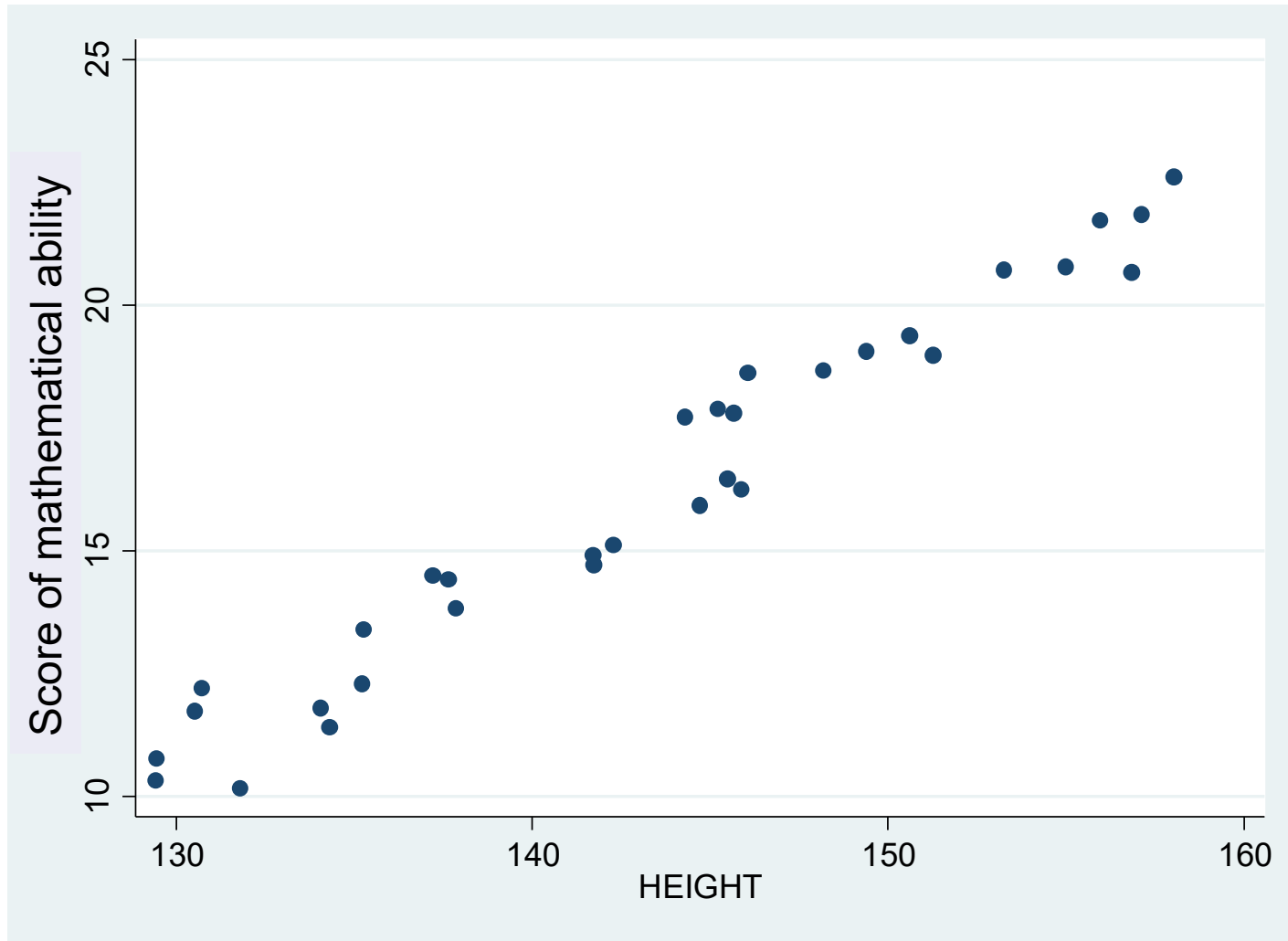
- ✓ Stratification by confounder(s)
- ✓ Multivariable / multiple analysis

**Prediction of individual risk**

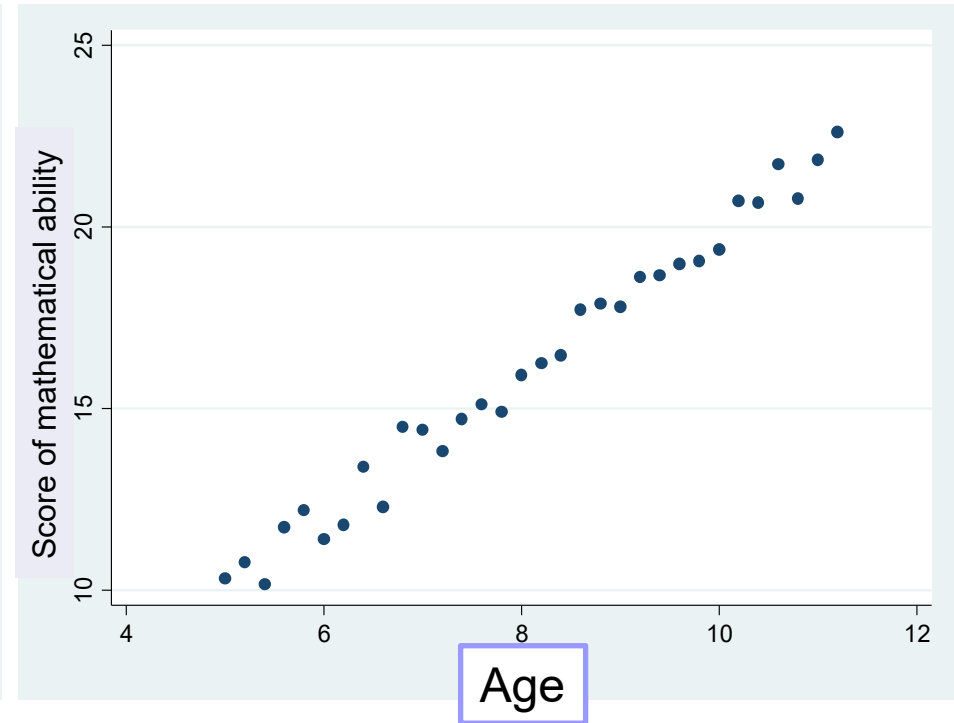
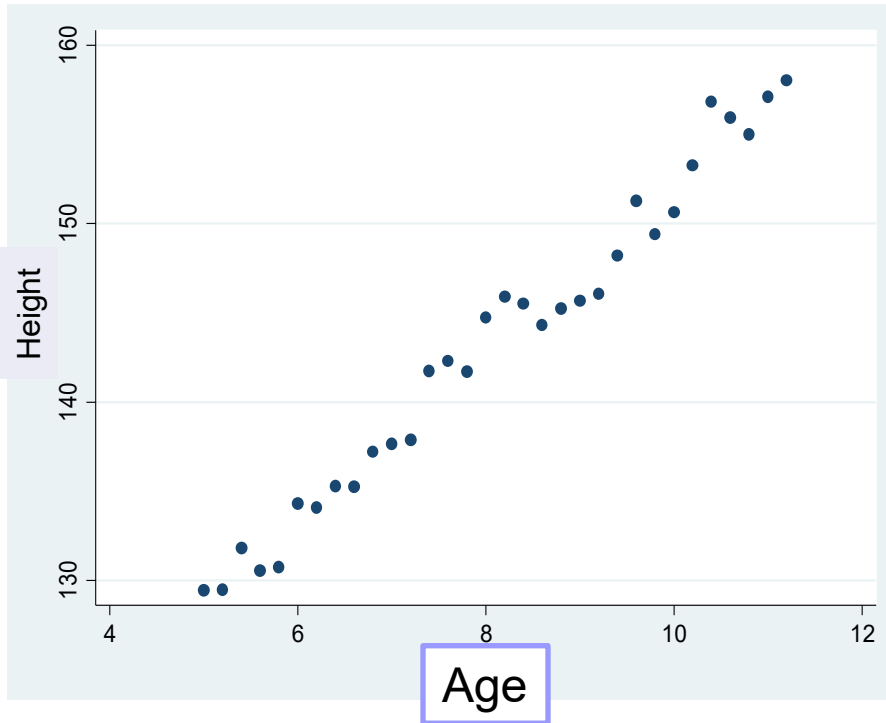


# **CONFOUNDING EFFECTS**

# Association between height and score of maths



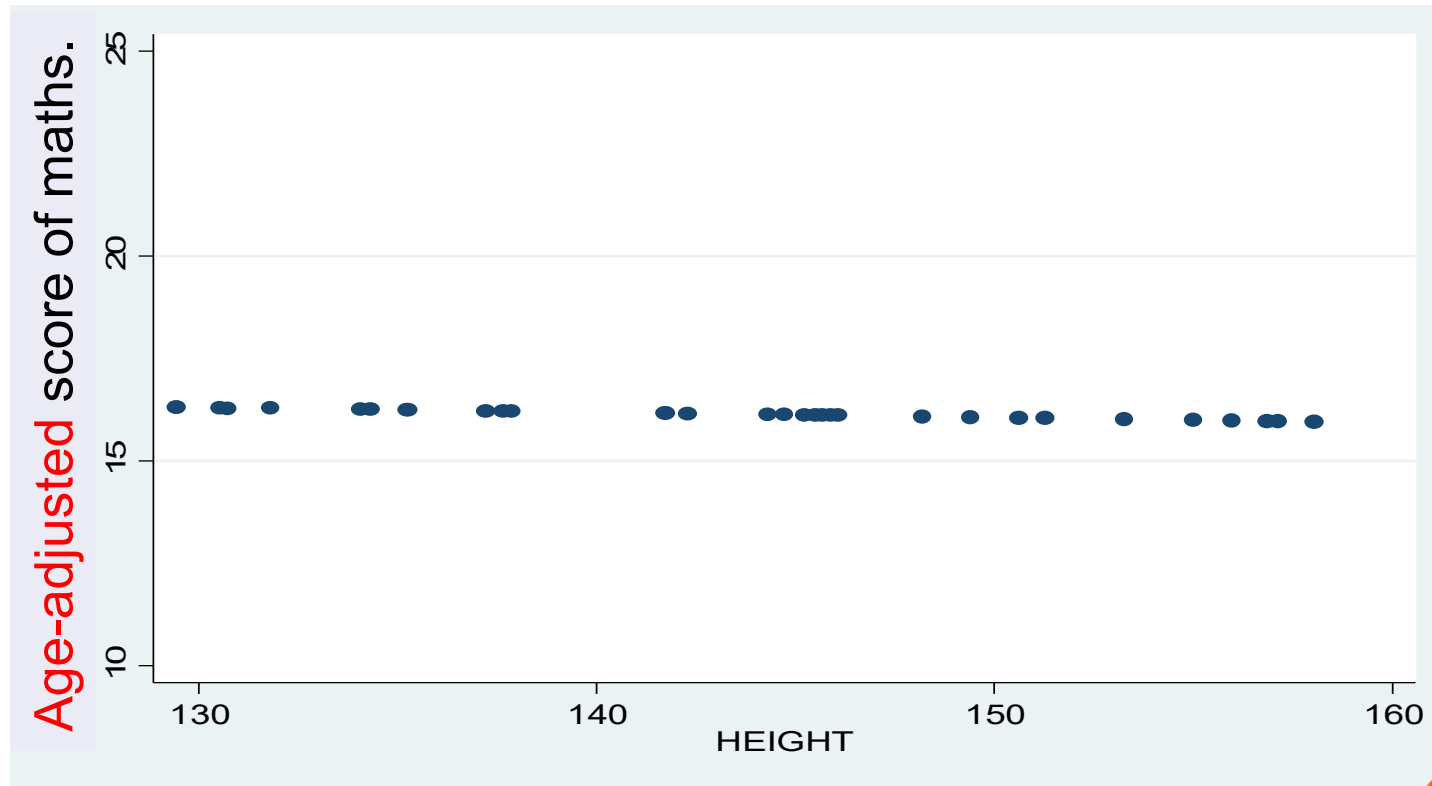
# Both height and ability of maths increase with age



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



After age-adjustment, there is no association between height and score of maths



We use **multivariable analysis** to adjust the effect of confounding factor(s), age in this case.



# Regression models for multivariable analysis

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



# **LINEAR REGRESSION ANALYSIS**



# Results of regression analysis before adjusting the effect of age

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

Ability score of maths

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

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

Source	SS	df	MS	
-----+-----				Number of obs = 32
Model	422.6119	2	211.305972	F(2, 29) = 851.23
Residual	7.19885	29	.248236138	Prob > F = 0.0000
-----+-----				R-squared = 0.9833
Total	429.81079	31	13.8648643	Adj R-squared = 0.9821
				Root MSE = .49823

The coefficient was **0.411** before adjustment

ama	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
<b>height</b>	<b>-.0121303</b>	<b>.0680948</b>	<b>-0.18</b>	<b>0.860</b>	<b>-.1513998</b>	<b>.1271393</b>
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
-----+-----						

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

Let's see the equation

$$\text{Ability of maths (AM)} = \alpha + \beta_1(\text{Height})$$

$$\rightarrow \text{AM} = -42.8 + 0.41(\text{Height})$$

0.41 points increase by 1cm increase of height

$$\text{AM} = \alpha + \beta_1(\text{Height}) + \beta_2(\text{Age})$$

$$\rightarrow \text{AM} = 1.48 - 0.01(\text{Height}) + 2.02(\text{Age})$$

0.01 points decrease by 1cm increase of height

Confounding effect: magnitude and direction of the association

ANOVA table

Sum of Squares  
 Degrees of freedom  
 Mean sum of squares (SS/df)

F statistic (df<sub>m</sub>, df<sub>r</sub>)

P value of F test

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

t = Coef. / SE

P value (H<sub>0</sub>: coef.=0)

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

CI of Coef.



## Interpretation of coefficients in general

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

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

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

**Difference:**  $Y_e - Y_u = \beta$

- **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

$$\text{Exposed: } \ln(Y_e) = \alpha + \beta(\text{Exp}=1) = \alpha + \beta$$

$$\text{Unexposed: } \ln(Y_u) = \alpha + \beta(\text{Exp}=0) = \alpha$$

$$\text{Difference: } \ln(Y_e) - \ln(Y_u) = \beta$$

$$\text{Ratio: } Y_e / Y_u = e^\beta$$

log (Ye/Yu)

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



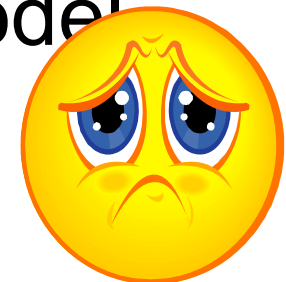
# **NOTES ON PERFORMING A REGRESSION ANALYSIS**

# Control of confounding with regression model

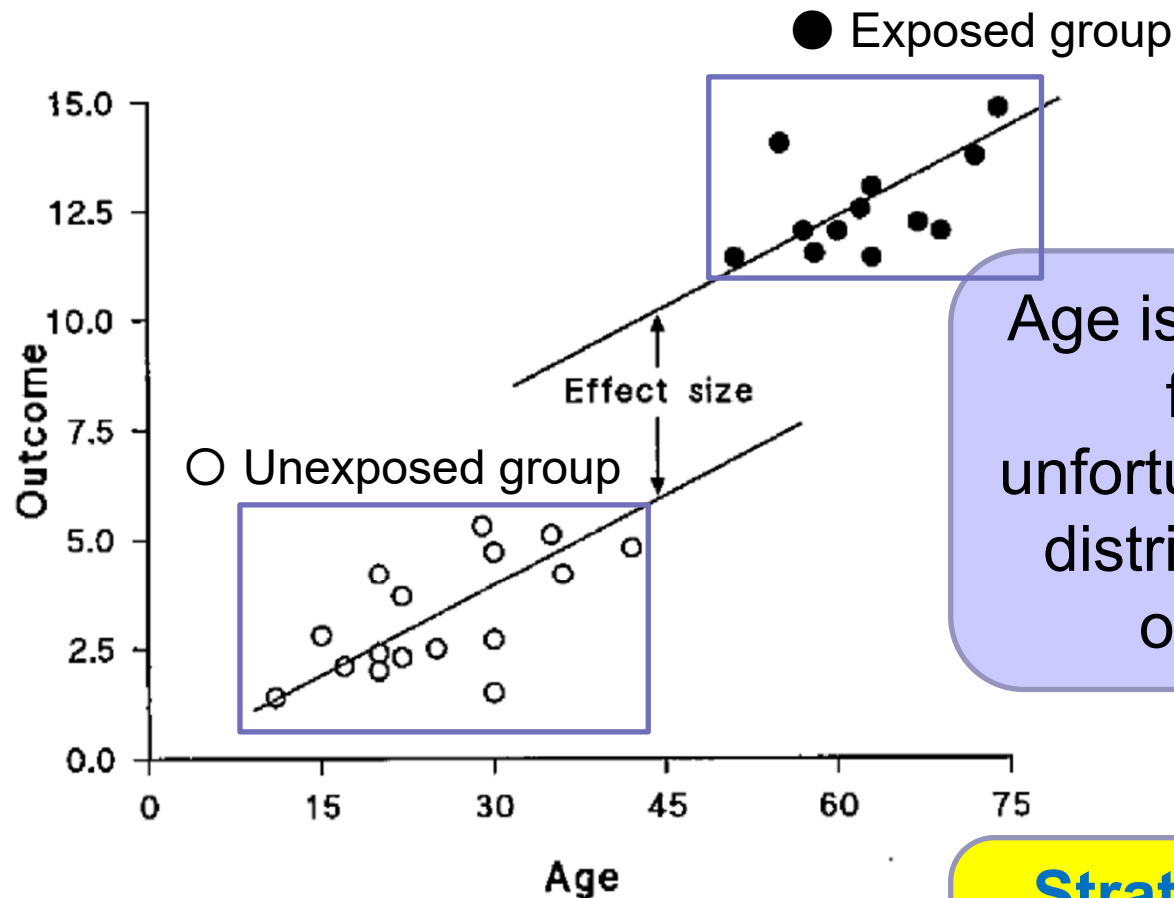
- Compared to **stratified analysis**, several confounding variables can be easily controlled simultaneously using a multivariable regression model.



- 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 the effect.

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

In a regression analysis, however,

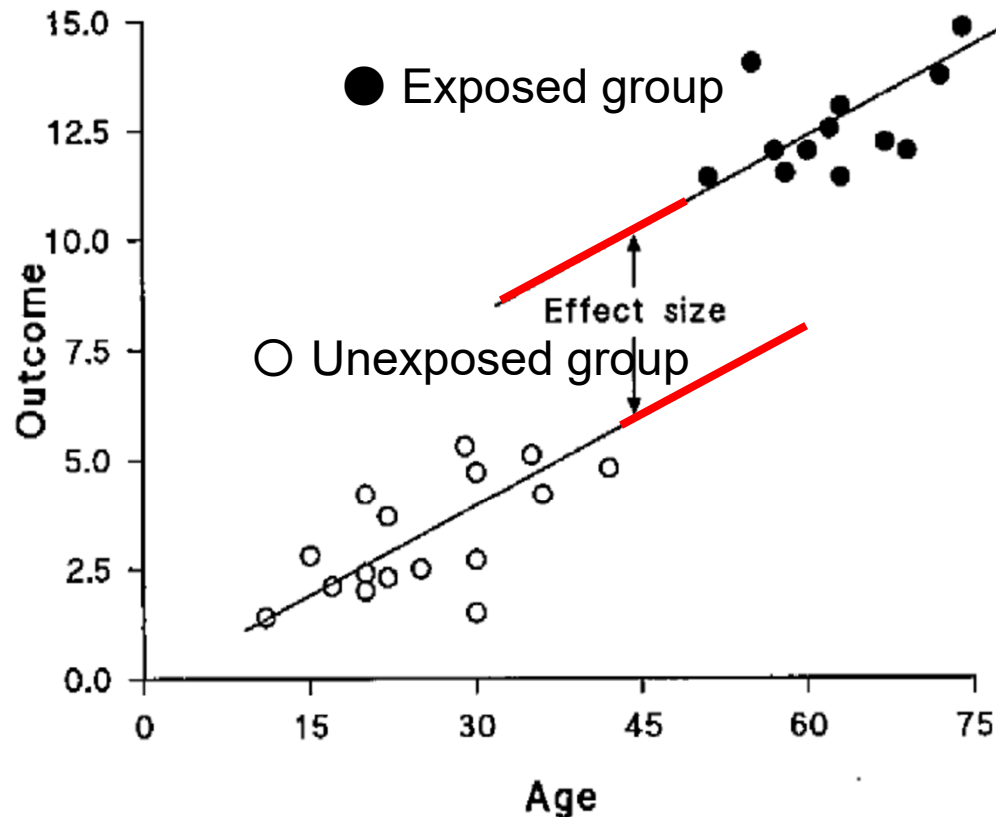
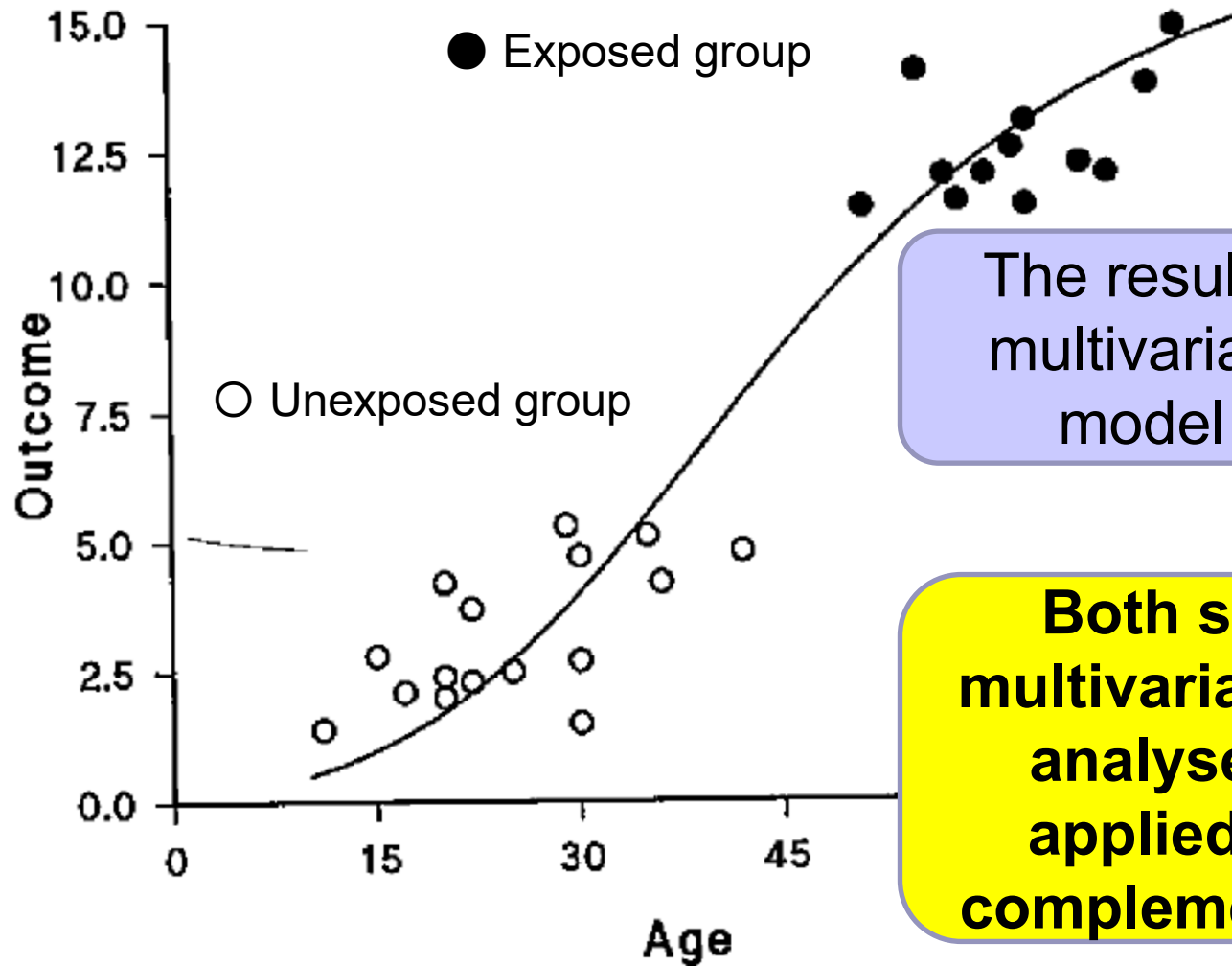


Figure 12-4 Hypothetical example of a multivariable linear regression of outcome data involving a dichotomous exposure variable (exposed = solid circles, unexposed = open circles) and age.

Two parallel lines are extrapolated, even for non-overlapping age groups. As a result, the regression analysis would result that there is a difference in the outcome between the exposed and non-exposed groups.

If the truth was like in this graph, no linear association between age and outcome



The result obtained by a multivariable regression model was biased.



**Both stratified and multivariable regression analyses should be applied as mutually complemented methods**



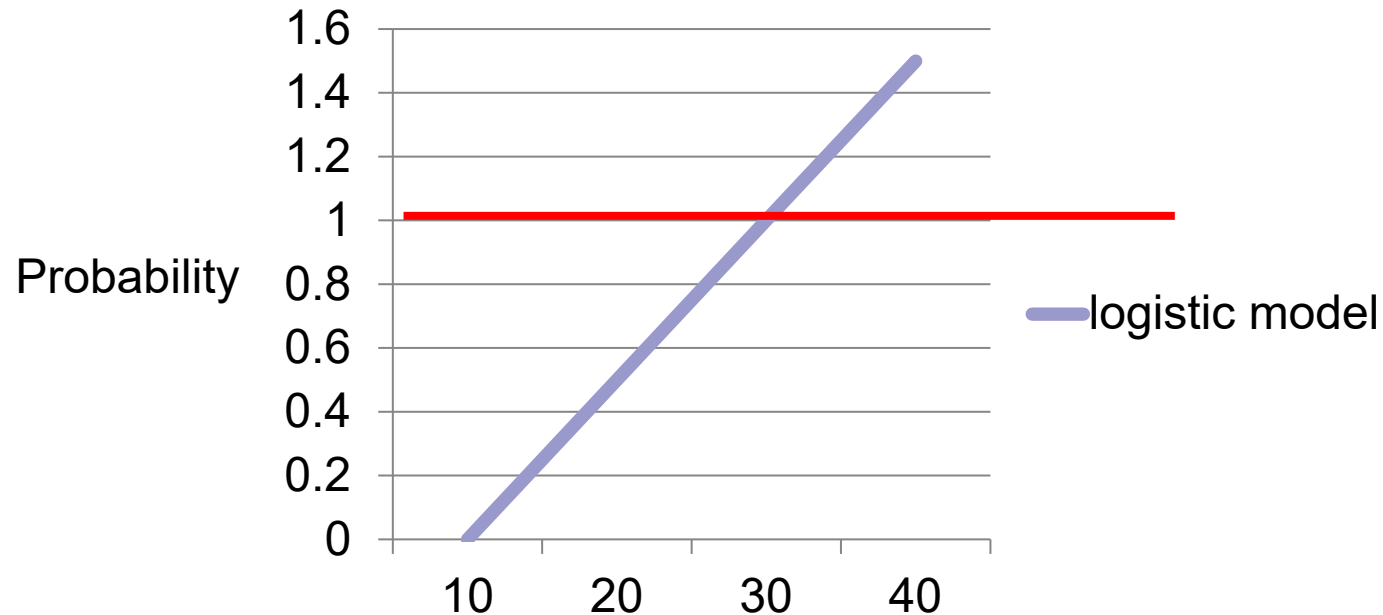
# **LOGISTIC REGRESSION ANALYSIS**

# Logistic regression analysis

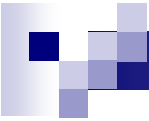
- Logistic regression is used to model the probability of a binary response as a function of a set of variables thought to possibly affect the response (called covariates).

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

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 regression 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 ( $0 \sim 1$ )

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

This transformation allows us to use a linear model.

As always, we use the natural log.

The logit is therefore **the log odds**, since  $\text{odds} = p / (1-p)$

# 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  $\alpha$  and  $\beta$  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:**  $O_e / O_u = e^\beta$

$\log(O_e/O_u)$


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



# **STRATEGY FOR CONSTRUCTING REGRESSION MODELS**



# Basic principles

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

## **Dose-response relation**

4. Evaluate **interaction(s)**

# 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) =  $\alpha + \beta_1(\text{Height})$

→ AM =  $-42.8 + 0.41(\text{Height})$

> 10% change

AM =  $\alpha + \beta_1(\text{Height}) + \beta_2(\text{Age})$

→ AM =  $1.48 - 0.01(\text{Height}) + 2.02(\text{Age})$

# 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?**





# **MAJOR PROBLEMS OF REGRESSION MODELS**



# Overfitting

- The phenomenon of overfitting in regression models is caused by trying to estimate **too many parameters from too few samples**.
- An overfit model result in **misleading** regression coefficients, p-values, and R-squared statistics.





# Solution of overfitting

- The best solution to an overfitting problem is **avoidance**.
- Identify the important variables carefully, and think about the model that you are likely to specify, then, plan ahead to collect a sample large enough handle all predictors, interactions, and polynomial terms your response variable might require.

# How many explanatory variables (predictors) can we use in a model?

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



# **MULTICOLLINEARITY**



# Multicollinearity

- A state of very high **intercorrelations** or **inter-associations** among the independent variables.
- This is a kind of disorder of the data, and statistical inferences about the data may NOT be reliable if multicollinearity exists.



## The reasons why multicollinearity occurs

- An inaccurate use of dummy variables.
- The inclusion of a variable which is computed from other variables in the data set.
- The repetition of the same kind of variable.



# How to detect multicollinearity

- Simple addition or removal of a variable to or from the regression model
  - If you observe a dramatic change in the model, it indicates the presence of multicollinearity in the data.
- Variance Inflation Factor (VIF)
  - If the value of VIF **10 and above**, then the multicollinearity is problematic.



# **PROPENSITY SCORE**



# 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 (94)	2167 (56)	<.001
Clinical history			
Diabetes, No. (%)	432 (19)	432 (11)	<.001
Hypertension, No. (%)	1569 (68)	1569 (41)	<.001
Tobacco use, No. (%)	500 (22)	500 (13)	.001
Prior coronary artery disease, No. (%)	178 (8)	178 (5)	<.001
Prior coronary artery bypass grafting, No. (%)	27 (1)	27 (1)	<.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
$\beta$ -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/m <sup>2</sup>	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	130 (15)	130 (15)	.92

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.

**Table 3.** Selected Baseline and Exercise Characteristics According to Aspirin Use in Propensity-Matched Patients\*

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P Value
<b>Demographics</b>			
Age, mean (SD), y	60 (11)	61 (11)	.16
Men, No. (%)	951 (70)	974 (72)	.33
<b>Clinical history</b>			
Diabetes, No. (%)	203 (15)	207 (15)	.83
Hypertension, No. (%)	679 (50)	698 (52)	.46
Tobacco use, No. (%)	161 (12)	162 (12)	.95
<b>Cardiac variables</b>			
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 (†).

Usually, you do not need to adjust any variables after matching by propensity score

Same results by adjusting for PS → indicating the robust result in this study