Topic materials:
Dr Raquel Iniesta

Narration and contribution:
Zahra Abdula

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience
Department of Biostatistics and Health Informatics

KING'S College LONDON

Module Title: Introduction to Statistics

Session Title: Confounding

Topic title: Multiple regression with several explanatory variables: Adjusting for confounders

Learning Outcomes

After listening to this session you should be able to:-

- Understand what is meant by “confounding” or “confounders”.
- Be aware that confounding is a theory and always involves at least three variables, an exposure-outcome relationship of interest and a third variable that is thought to be cause of both the exposure and the outcome.
- Understand the statistical problem caused by the existence of confounding.
- Know how to use multiple linear regression models to adjust for potential confounding variables.

Confounding Variables

Confounding: A situation in which the association between an explanatory variable (e.g exercise $\boldsymbol{x_1}$) and outcome (e.g. weight $\boldsymbol{y}$) is distorted by the presence of another variable (e.g. hours of free time $\mathbf{x_2}$).

Theory:

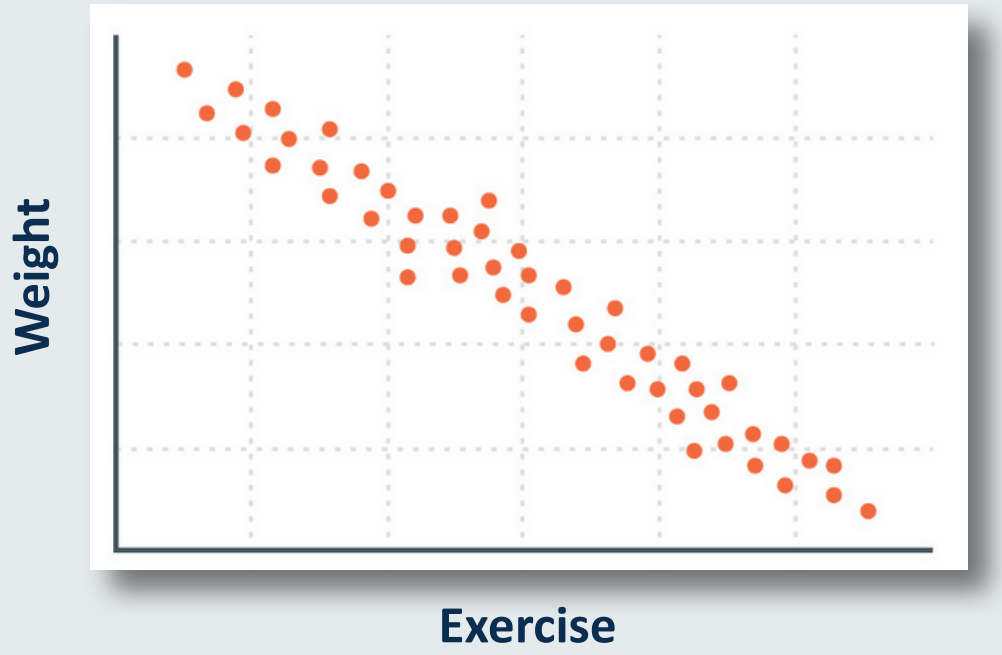
hours of free time causes **more exercise**
hours of free time also affects **weight**

i.e. hours of free time is a **common cause** of the exposure (exercise) and outcome (weight) of interest.

What happens if we only test the relationship between exercise ($\boldsymbol{x_1}$) and weight ($\boldsymbol{y}$) in a simple linear regression model with only exercise as an independent variable?

Simple linear regression result: unadjusted relationship

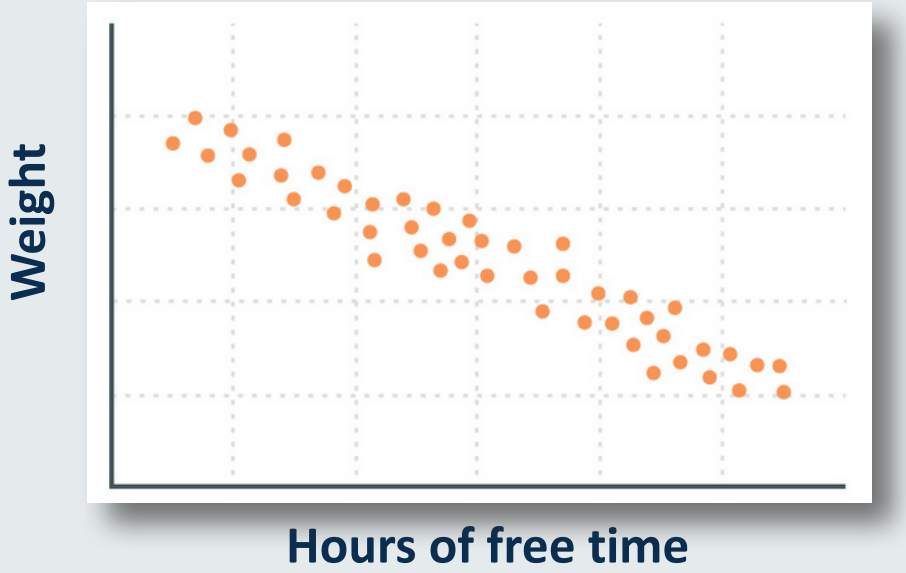
What happens if we only assess the relationship between exercise (x_1) and weight (y) in a simple linear regression model with only exercise as an independent variable?



$y = 70 - 5x_1 + \varepsilon$; p=0.01 for β_1

Simple linear regression results: possible relationship between hours of free time and outcome of interest

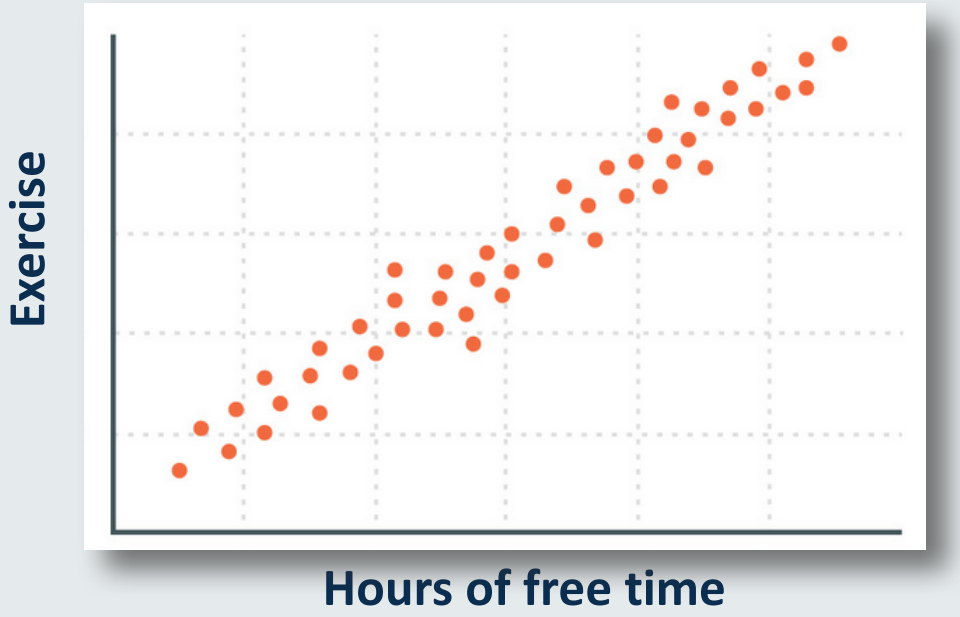
What happens if we only assess the relationship between hours for free time (x_2) and weight (y) in a simple linear regression model with only hours free time as an independent variable?



$y = 69 - x_2 + \varepsilon$; p=0.001 for β_2

Simple linear regression: possible relationship between hours of free time and exposure of interest

Note that exercise is also associated with hours of free time ... so our two potential independent variables are also associated



$Y = 0.5x + \varepsilon$; p=0.015

Confounding Problem

It appears:

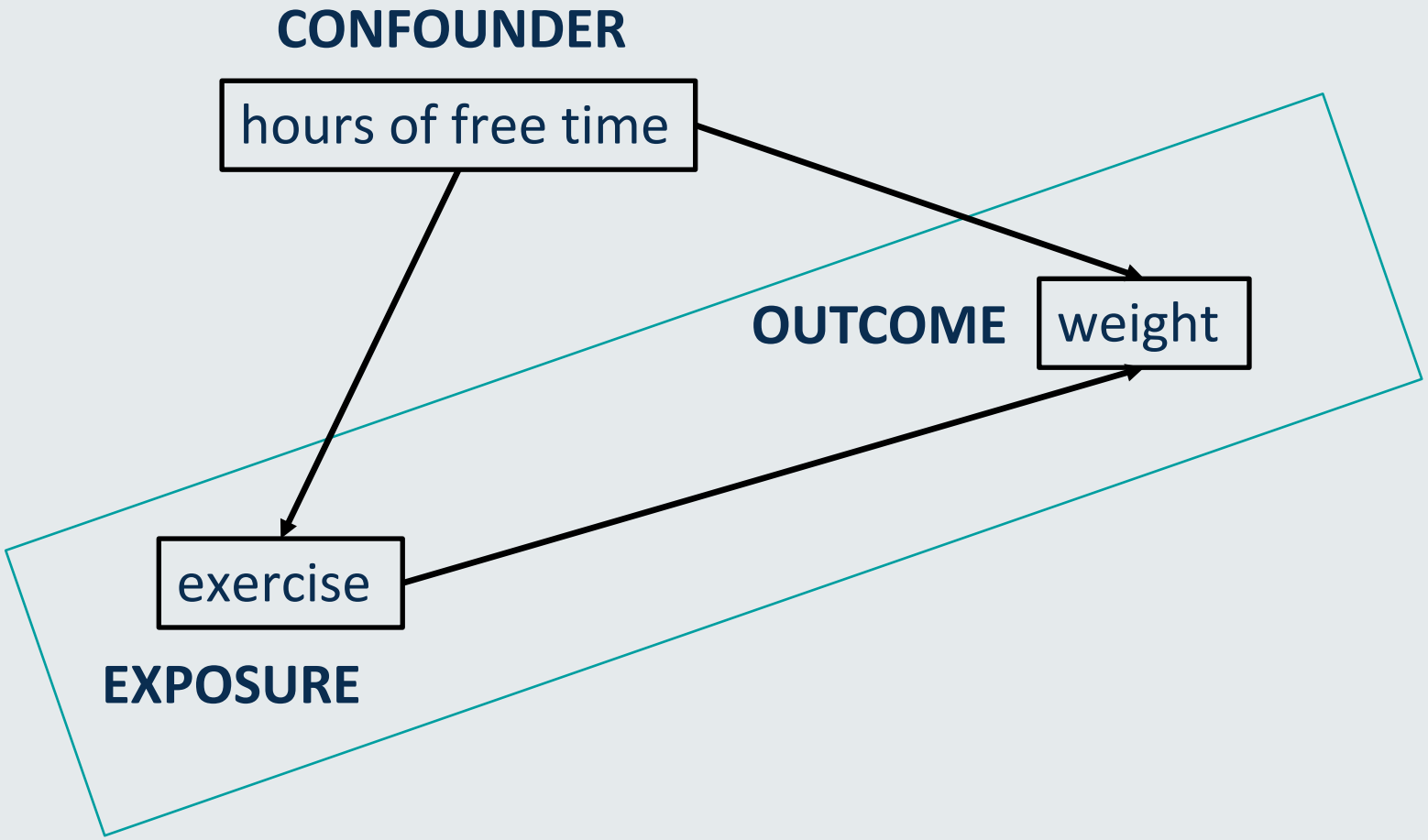
- Weight might go down with more free time.
- Those with more free time might exercise more.

So even if more exercise did not cause any weight loss, we might still observe an association between exercise and weight.

In the presence of such a common cause, we **cannot attribute all the observed association** between the exposure and the outcome **to the exposure causing the outcome**.

This is known as **confounding**; or in other words free time is a **confounder** of the effect of exercise on weight.

How Does Confounding Work?



Confounding Variables: Explanation

In an experiment, the **independent variable** is typically thought to **cause** your dependent variable.

Example: If you are researching whether lack of exercise leads to weight gain:
Exercise is your independent variable and weight gain is your dependent variable.

Confounding variables are any other variables that cause both your dependent and your main independent variable of interest.

They are like extra independent variables that are having a **hidden** effect on your dependent variables while being related to the independent variable of interest.

If not taken into account, confounding variables will **introduce bias** in the estimation of β_1.

Multiple Linear Regression Model: Confounding

We can formulate the model in terms of confounding

The researcher's ultimate goal is to be able to estimate the effect of an independent variable (or exposure) on a dependent variable (or outcome) while adjusting for other variables that distort this relationship (the confounders).

If we have an independent variable we are interested in – we want to get an **adjusted estimate** of the association between this independent and the dependent variable.

Using multiple linear regression allows us to hold all other independent variables constant allowing us to get an estimate of the effect of the independent variable of interest while adjusting for other variables in the model which are hypothesized to be confounders.

Multiple Regression Model: Adjust for Confounding

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

- Multiple regression framework is a **natural** and **practical** way of **adjusting for confounders**.

- To **adjust the association** between exercise (x_1) and weight (y) for hours of free time (x_2), **all we need to do is include the confounder** hours of free time (x_2) in the regression model as an **additional predictor**, which will automatically adjust the y and x_1 association for x_2.

- The coefficient of x_1 (i.e., β_1) in this case will represent the **adjusted association** between weight and exercise, controlling for the effect of the number of hours of free time.

Adjusting for One Confounder

$y = 72 - 3x_1 - x_2 + \varepsilon$

		p-value
Slope for x_1 (β_1)	-3	0.04
Slope for x_2 (β_2)	-1	0.01

Where:

y=weight;
x_1=frequency of exercise per week;
x_2=hours of free time per week;

The effect of exercise on weight **adjusted for** hours of free time is β_1=-3.
This effect is statistically significant (p=0.04).
We can infer the association for the whole population.

As I increase number of weekly exercise sessions by 1, I am decreasing my weight by 3kg, keeping the hours of free time per week fixed.

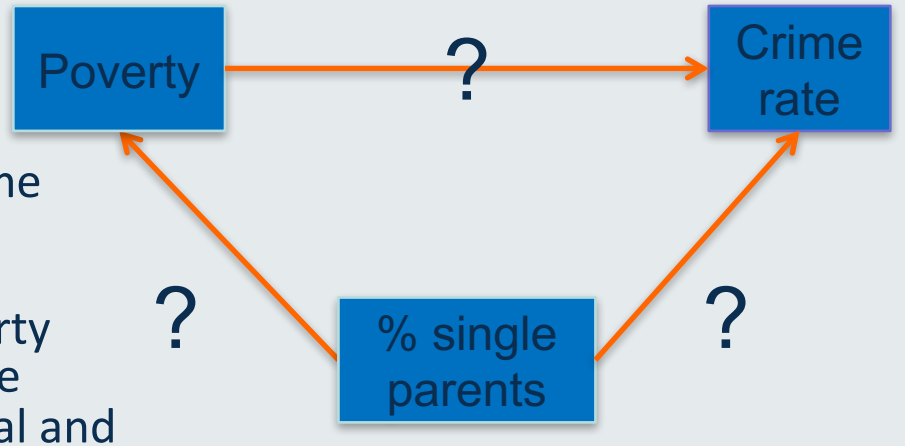
Dealing with Multiple Confounders

- Multiple regression model can deal with any number of confounders (within reason).
- All confounders are adjusted for by including them simultaneously as additional predictors.
- However, the sample size can restrict the number of variables that can be included in a regression model.
- A rule of thumb is to ensure that there are more than 10 observations (data points) per each independent variable.
- A sample of size 100, for example, will allow us to consider up to 10 independent variables.

Knowledge Check – Confounding

Use the lecture_7_data.sav dataset on Keats

We want to look at the relationship between poverty and crime rate and consider % single parents in the State as a confounder of the poverty – crime rate relationship.



Q1: Fit a simple linear regression model for the relationship between poverty (independent) and crime rate (dependent). Write down the estimate of the regression coefficient, β_1, for poverty, as well as the 95% confidence interval and p-value for β_1.

Q2: To check whether the data is consistent with the theory that % single parents is a confounder, we assess whether it is associated with both our independent and dependent variables of interest. Fit two simple linear regression equations: poverty (dependent) and % single parents (independent), and crime rate (dependent) and % single parents (independent). Write down the estimates of the regression coefficients, the β_1 (and 95% confidence intervals and p-values) for % single parents from both equations.

Q3: Does it seem that % single parents is likely to be a confounder of the poverty – crime rate association? Why or why not? What would you need to do to account for confounding by % single parents in assessing the poverty – crime rate association?

Knowledge Check Solutions

Q1:

B_1 for poverty and crime rate is 25.452, 95% CI 6.833 to 44.072, $p = 0.008$

Q2:

B_1 for % single parents and poverty is 1.250, 95% CI 0.489 to 2.012, $p = 0.02$

B_1 for % single parents and crime rate is 130.110, 95% CI 85.809 to 174.411, $p < 0.001$

Q3:

As % single parents is associated both with our independent (poverty) and dependent (crime rate) variables of interest, the data set confirms that it acts as a confounder of the poverty – crime rate relationship.

To deal with this, we would want to include this variable in a multiple regression model where both poverty and % single parents were independent variables to get an estimate of this relationship **adjusted for confounding**.

References

Agresti, A., & Finlay, B. (2009).
Statistical Methods for the Social Sciences (4th ed.). New Jersey, NJ: Prentice Hall Inc.

Douglas, C., Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006).
Introduction to Linear Regression Analysis. New York, NY: Wiley.

Thank you

Please contact your module leader or the course lecturer of your programme, or visit the module's forum for any questions you may have.

If you have comments on the materials (spotted typos or missing points) please contact Dr Iniesta:

Raquel Iniesta, PhD
Department of Biostatistics and Health Informatics
IoPPN, King's College London, SE5 8AF, London, UK
raquel.iniesta@kcl.ac.uk

For any other comments or remarks on the module structure, please contact one of the three module leaders of the Biostatistics and Health Informatics department:

Zahra Abdula: **zahra.abdulla@kcl.ac.uk**

Raquel Iniesta: **raquel.iniesta@kcl.ac.uk**

Silia Vitoratou: **silia.vitoratou@kcl.ac.uk**