Topic materials:
Dr Raquel Iniesta
Department of Biostatistics
and Health Informatics

Narration and contribution:
Zahra Abdulla

Improvements:
Nick Beckley-Hoelscher
Kim Goldsmith
Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

Module Title: Introduction to Statistics

Session Title: Estimating interaction effects

Topic title: Effect Modification (Interaction)

Learning Outcomes

After working through this session you should be able to:

- understand the meaning of the effect modification
- understand how to estimate effects in presence of interaction
- understand how to interpret the effect modification

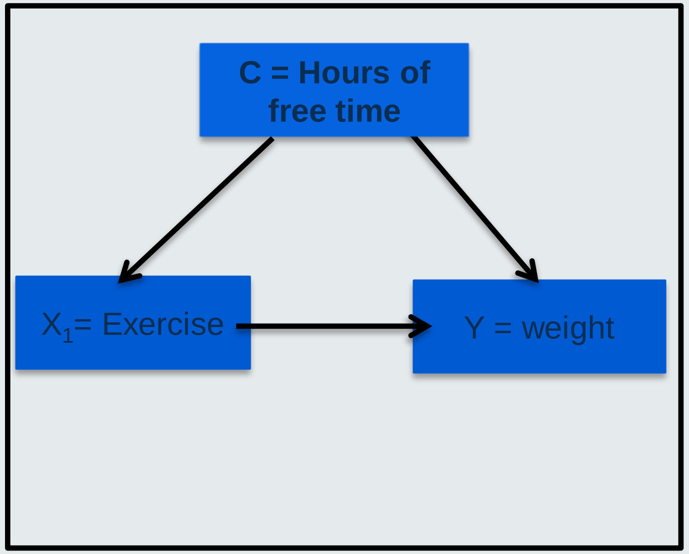
Previously on ‘Introduction to Statistics’

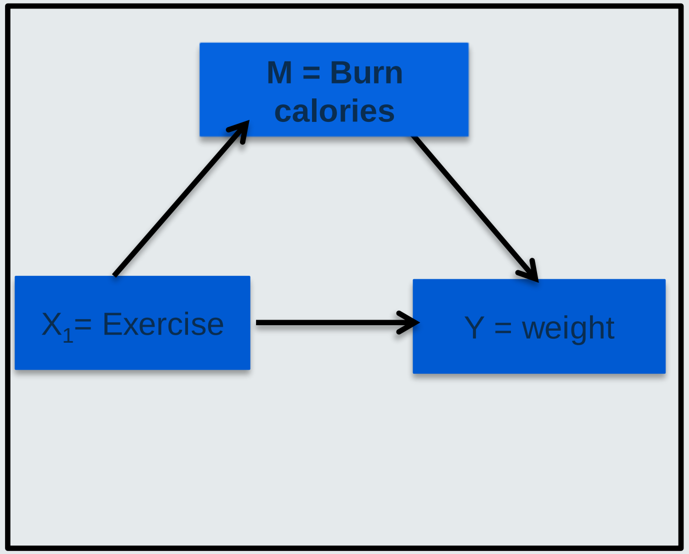
Before, we focused on the 3 variables (Y, X_1 and X_2) case.

We discussed the different roles that a third variable $\mathbf{X_2}$ can have while investigating the association between an independent X_1 and a dependent variable Y.

X_2 could be a **confounder** (C) or a **mediator** (M)

Previously on ‘Introduction to Statistics’



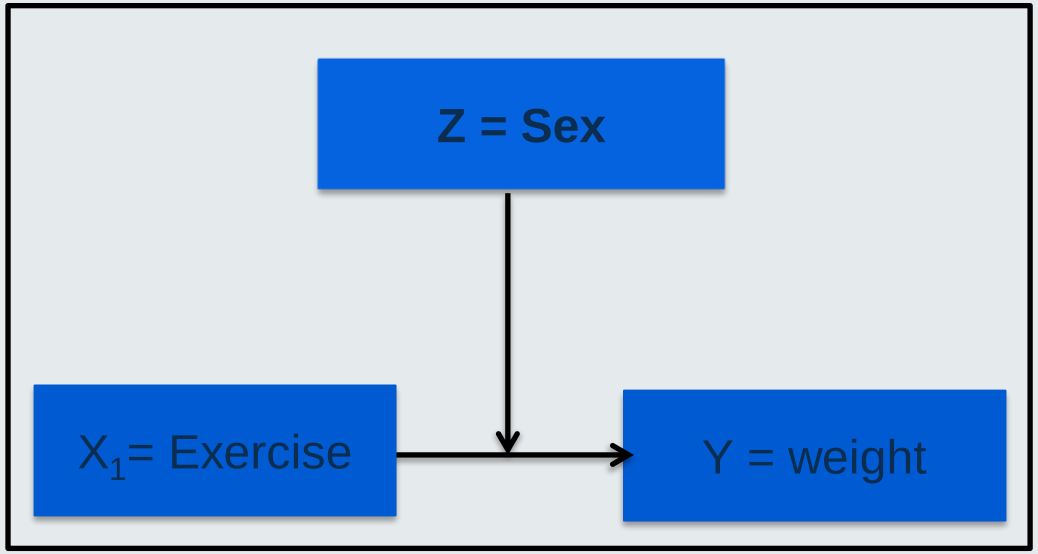


A **confounder** (C) has a common effect on the independent and dependent variables. A confounder **is extrinsic to the causal** pathway.

A **mediator** (M) is caused by the independent variable which in turn causes the dependent variable. A mediator **is in the causal** pathway

Effect Modification (Interaction)

- The third variable X_2 can have another role.
- X_2 can be a **modifier** (or moderator or have an interaction effect) on the association between Y and X1:



We will denote the moderator with letter **Z**

Key point: The association between X_1 and Y is not the same at different values or levels of **Z**.
In other words, a **modifier** is a variable that **alters** the relationship between the independent X_1 and dependent Y variables.

Example: If a man and a woman do the same exercise, the effect on weight is different. **Sex modifies the effect of Exercise on Weight.**

Establishing Effect Modification

To assess the **significance** of an effect modification or interaction:

Step 1:

- A new variable needs to be considered
- This new term is the **cross-product** between X_1 and the modifier Z. This is called the **Interaction Term.**
- The new term is noted like $\mathbf{X_1 \times Z}$

Step 2:

Consider the original linear regression to test the effect between Y , X_1 and Z
Add the new variable $X_1 \times Z$ to the regression model: $\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 Z + \beta_3 x_1 \times Z + \varepsilon}$

In the context of regression analysis, assessing effect modification is the same as assessing interaction effect.

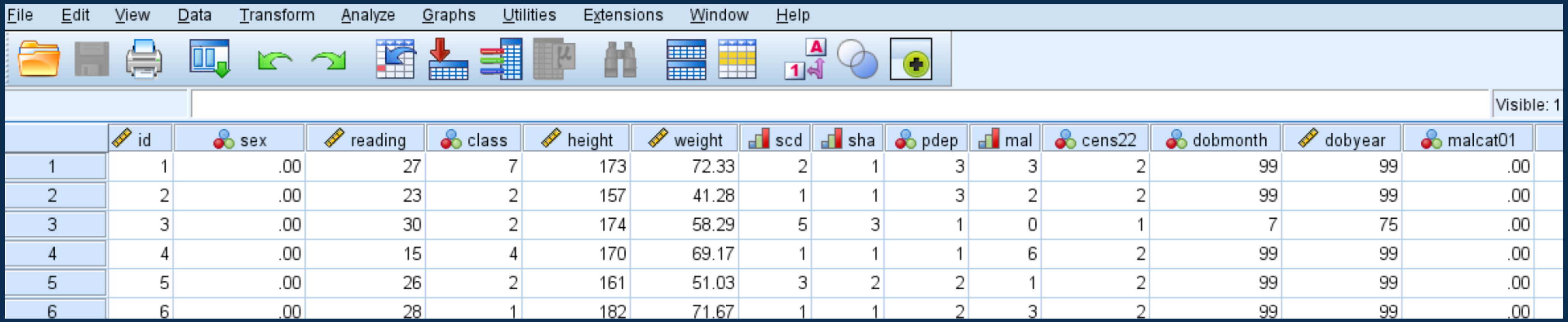
Step 3:

Primary focus: Test coefficient β_3; $\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$ If p value < 0.05 then there is a **significant effect modification**.

If p value < 0.05 we will conclude there is a significant interaction between X_1 and the modifier Z .

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_9a_data.sav**.



	id	sex	reading	class	height	weight	scd	sha	pdep	mal	cens22	dobmonth	dobyear	malcat01
1	1	.00	27	7	173	72.33	2	1	3	3	2	99	99	.00
2	2	.00	23	2	157	41.28	1	1	3	2	2	99	99	.00
3	3	.00	30	2	174	58.29	5	3	1	0	1	7	75	.00
4	4	.00	15	4	170	69.17	1	1	1	6	2	99	99	.00
5	5	.00	26	2	161	51.03	3	2	2	1	2	99	99	.00
6	6	.00	28	1	182	71.67	1	1	2	3	2	99	99	.00

The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **sex**: gender of child defined at birth (0=male, 1=female)
- **height**: height in cm at age 16
- **weight**: weight in kg at age 16
- **reading**: reading score
- **mal**: malaise (a feeling of general discomfort/uneasiness) score
- **class**: general classification of social class (7 Categories)

Example: The NCDS Data

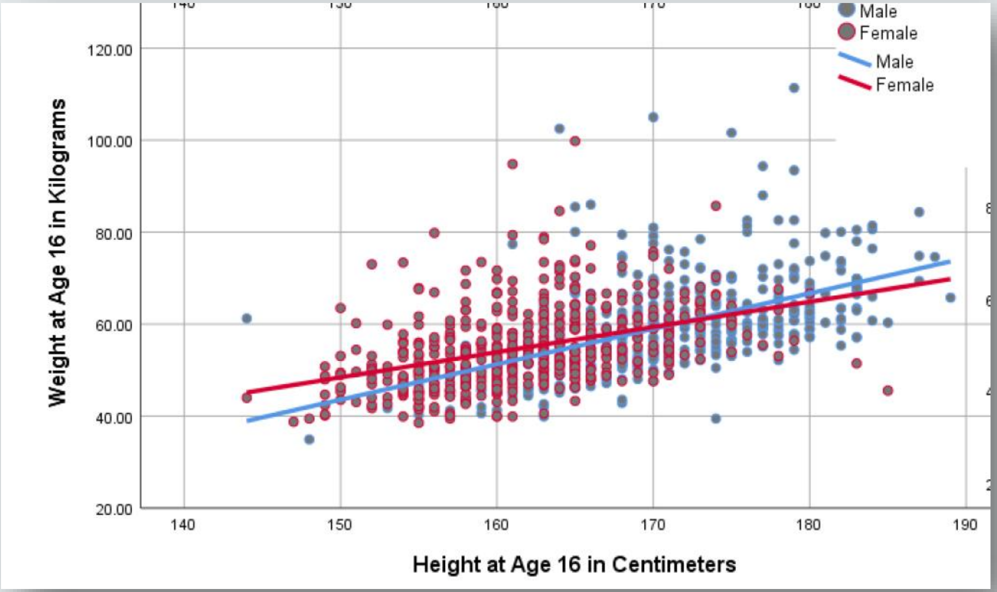
Consider estimating the linear relationship between x_1 = height and Y= weight:

$$Y = \beta_0 + \beta_1 x_1$$

We estimate two models, separately, in boys and girls:

Boys: $Y = -72.01 + 0.77x_1$

Girls: $Y = -33.75 + 0.55x_1$



Interpretation:

- For boys: 1 cm increase in **height** leads to **0.77 kg** increase in **weight**
- For girls: 1 cm increase in **height** leads to **0.55 kg** increase in **weight**
- **There might be interaction between height and sex: the height-weight relationship differs between sex categories**

Estimating the Interaction Effect

Step 1:

- A new variable needs to be considered.
- This new term is the **cross-product** between height and the modifier sex .
- It is noted like height $\times$ sex.
- We create a new variable that is the product between height and sex.

Step 2:

- Consider the original linear regression to test the effect between weight, height and sex.
- Add the new variable from step 1 'height $\times$ sex' to the regression model.

$$\boldsymbol{weight = \beta_0 + \beta_1 height + \beta_2 sex + \beta_3 height \times sex}$$

Where β_3 is the interaction term and height x sex is the cross-product term

- Estimate β coefficients.

Step 3:

- Primary focus: Test coefficient β_3; $\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$ Is there a significant interaction between the variables?

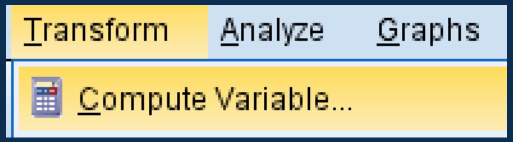
SPSS Slide: ‘How to’ Steps

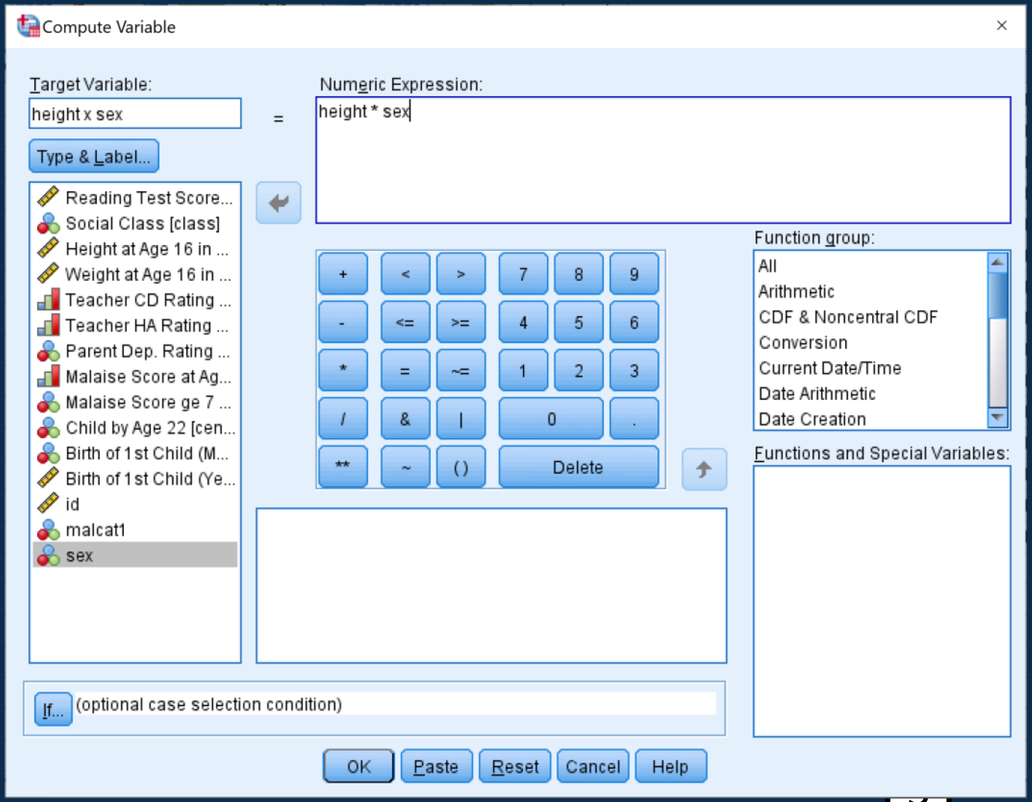
Create an **interaction term height_x_sex** from lecture_9a_data.sav

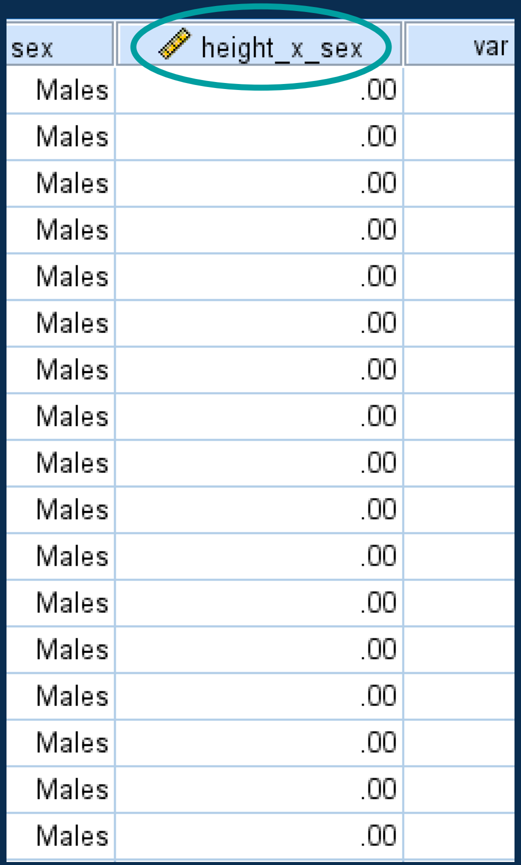
1) Use ‘**Transform**’ -> ‘**Compute variable**’

2) In “**Target variable**” write the name of your interaction term: “**height_x_sex**”

In “Numeric Expression” drag ‘height’ and ‘sex’ separated by a ‘*’





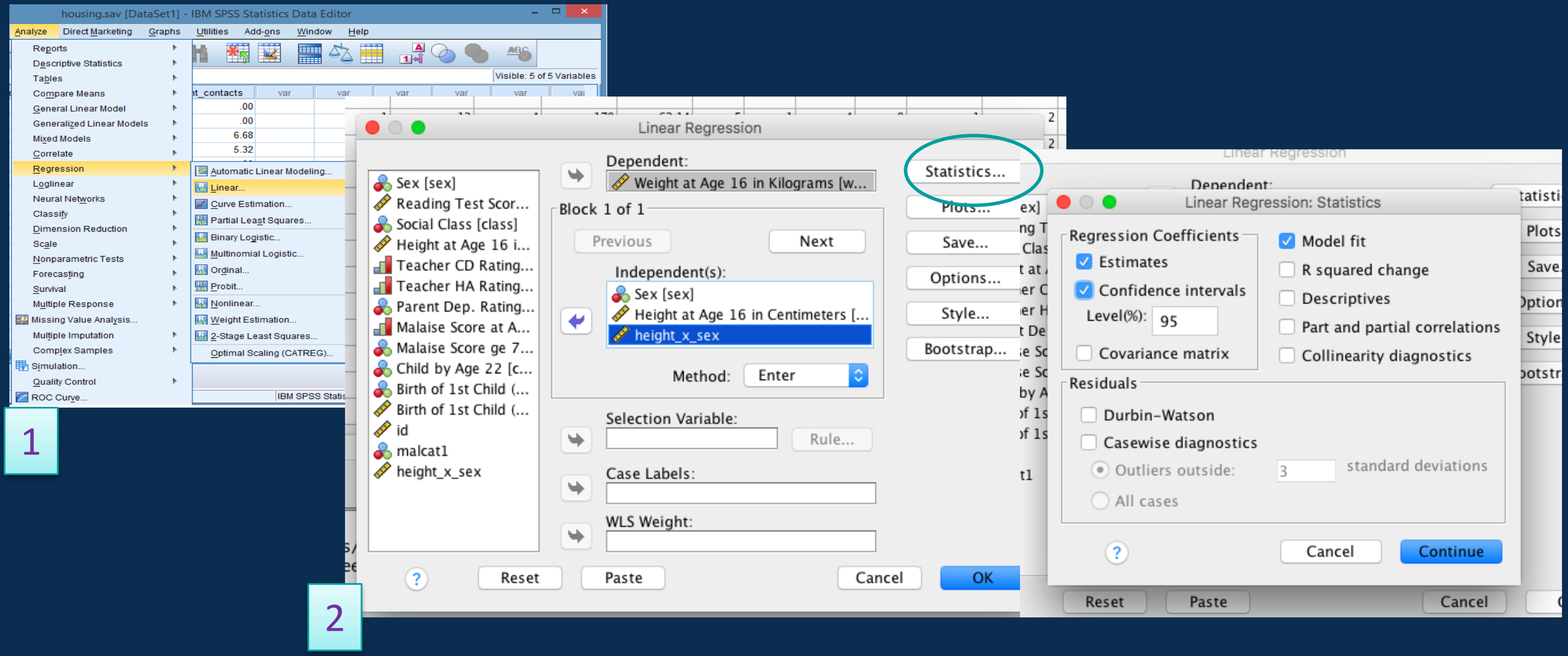


New variable in data set

SPSS Slide: 'How to' Steps

Estimating the interaction effect height_x_sex in a multiple linear regression model for weight, height and sex from lecture_9a_data.sav data

1) Use 'Analyse' -> 'Regression' -> 'Linear'
2) In dependent put 'weight' and in independent put 'height', 'sex', 'height_x_sex'



Output and Interpretation

$$weight = b_0 + b_1 height + b_2 sex + b_3 height \times sex$$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-72.014	8.893		-8.098	.000	-89.464	-54.563
	Height at Age 16 in Centimeters	.771	.052	.640	14.804	.000	.669	.873
	Gender	38.263	12.963	1.982	2.952	.003	12.825	63.702
	hxs	-.223	.078	-1.870	-2.851	.004	-.376	-.069

a. Dependent Variable: Weight at Age 16 in Kilograms

- β_1 = 0.77 is interpreted as the effect of height on weight when sex=0 (boys)
- β_2 = 38.26 represents the effect of sex on weight when height=0 (not meaningful! because a person's height can not be zero)
- $\beta_3 = -0.22$ represents the difference of the effect of height on weight between girls (sex=1) and boys (sex=0)
- The p-value of the Interaction effect (β_3= -0.22) is 0.004, we conclude that height × sex interaction effect is statistically significant
- The height-weight relationship significantly differs between boys and girls

Interpretation of $\beta' s$

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_1 + \boldsymbol{\beta}_2 \mathbf{Z} + \boldsymbol{\beta}_3 \mathbf{x}_1 \times \mathbf{Z} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\beta}_1$ is interpreted as the effect of $\mathbf{x}_1$ on **Y** when **Z = 0**
- Similarly, $\boldsymbol{\beta}_2$ represents the effect of $\mathbf{Z}$ on **Y** when $\mathbf{x}_1$ **= 0**
- $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are no longer useful unless zero values of the respective predictors are of particular interest.
- Both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are called **main effects**
- $\boldsymbol{\beta}_3$ is interpreted as the difference of the effect of $\mathbf{x}_1$ on **Y** by levels of $\mathbf{Z}$ variable.
- If the hypothesis test for $\boldsymbol{\beta}_3$ concludes that it significantly differs from 0, that will imply:
  - Both $\mathbf{x}_1$ and $\mathbf{Z}$ are associated with **Y**
  - The effect of $\mathbf{x}_1$ on Y will depend on $\mathbf{Z}$ and vice versa
  - The $\mathbf{x}_1 \times \mathbf{Z}$ interaction effect is interpreted as the difference of the effect of $\mathbf{x}_1$ between different levels of $\mathbf{Z}$

Estimating Effects in the Presence of an Interaction

$$\text{weight} = -72.014 + 0.771\text{height} + 38.263\text{sex} - 0.223\text{height} * \text{sex}$$

Given the above equation:

- What is the effect of height on weight?
- What is the effect of sex on weight?
- In a multiple linear regression model with no interaction:
  - the effect of height on weight would be β_1=0.771
  - the effect of sex on weight would be β_2=38.263.

  - What about β_3 ?

Estimating Effects in the Presence of an Interaction

- In presence of interaction, to estimate the effects we cannot just consider the main effects represented by the β_1 and β_2 coefficients.
- The above fitted equation, general formulae for the effects of height and sex on weight:

- Effect of height = $\beta_1 + \beta_3 \times$ sex = 0.771 - 0.223 * sex
  - Effect of height for boys = 0.771 – 0.223 x 0 **= 0.77kg**
  - Effect of height for girls = 0.771- 0.223 x 1 **= 0.55kg**
- Effect of sex = $\beta_2 + \beta_3 \times$ height = 38.263 -0.223 * height
- In this case it makes more sense to use average height
  - Effect of sex at average height =38.263 – 0.223 x 166.16 = **1.21kg**

Knowledge Check

- Consider the model:

  hours _of_sleep=7 + 2tiredness + 1.1 go_bed_before11pm + 0.3 tiredness x go_bed_before11pm

- The p value for the interaction term is 0.002.

- Please select the correct interpretation:

a) 0.3 represents the difference of the effect of tiredness on hours_of_sleep between those who go to sleep before 11pm and those who do not
b) For each unit increase of tiredness, hours_of_sleep increases by 2 hours
c) Those who go bed before 11pm are less tired

- Write out the calculation for the effect of tiredness and the effect of going to bed before 11pm.

Knowledge Check Solutions

- (a) is the correct solution

$\boldsymbol{\beta_3}$ = 0.3 is interpreted as the difference of the effect of $\mathbf{x_1}$=tiredness on **Y**=hours of sleep by levels of $\mathbf{Z}$ = go_bed_before11pm variable.

- Effect of tiredness = $\beta_1 + \beta_3 \times$ go to bed before 11pm

  = 2 +0.3* go to bed by 11pm

- Effect of going to bed before 11 pm

  = $\beta_2 + \beta_3 \times$ tiredness

  = 1.1 +0.3* tiredness

References

Agresti, A. and Finlay, B. (2009). *Statistical Methods for the Social Sciences* (4th Edition), Prentice Hall Inc.

- Chapter 10: Introduction to Multivariate Relationships
- Chapter 11: Multiple Regression and Correlation

Hayes, A .F. (2013). Introduction to Mediation, Moderation, and Conditional Process Analysis, Guildford Press.

- Chapter 7: Fundamentals of Moderation Analysis
- Chapter 8: Extending Moderation Analysis Principles

Frazer, Baron and Tix (2004) Testing Moderator and Mediator Effects in Counselling Psychology
Journal of Counselling Psychology Copyright 2004 by the American Psychological Association, Inc.
2004, Vol. 51, No. 1, 115–134 0022-0167/04/$12.00 DOI: 10.1037/0022-0167.51.1.115

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**