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

King's College London

Module Title: Introduction to Statistics

Session Title: Interactions and types of data

Topic title: Effect Modification (Interaction)

Learning Outcomes

After working through this session you should be able to:

- understand how to estimate interactions for different types of variables
- understand how to present interactions using the tabular format
- understand how to present interactions using the graphical format

Previously on “Introduction to Statistics”

- The example we discussed in the session before, illustrates the interaction between a **continuous** variable (height) and a **binary categorical** variable (sex)

- Categorical independent variables with **more than two levels** (for example ‘urbanicity’: Low, Medium, High) need to be recoded into **dummy** variables before defining cross-product terms.

- A **“dummy variable**” is a numerical variable used in regression analysis to represent subgroups of the sample in your study.

- Interaction terms should be considered for each dummy variable

Dummy Variables

Example: Y = Income; X_1 = job; Z = born city ; (London, Manchester, Leicester).

Z is converted into two binary dummy variables:

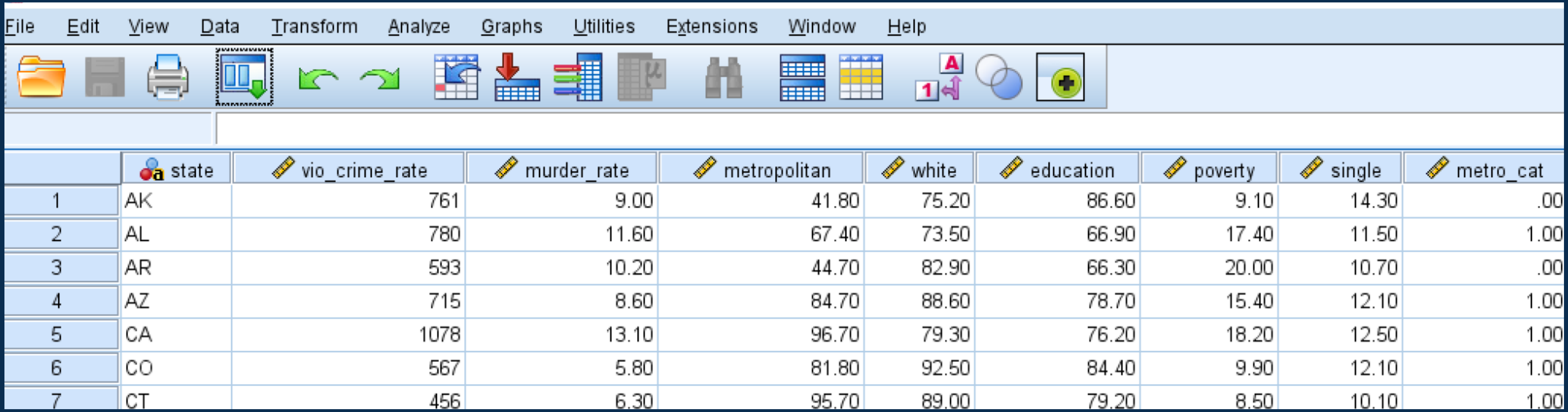
d_{London} = 1, 0, 0
$d_{Manchester}$ = 0, 1, 0

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 d_{London} + \beta_3 d_{Manchester} + \beta_4 x_1 \times d_{London} + \beta_5 x_1 \times d_{Manchester} + \varepsilon$$

Test coefficients β_4; $\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases}$ and β_5; $\begin{cases} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{cases}$

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_9b_data.sav**.



	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

The dataset contains data from 51 US states, measuring the crime rates and background measures for each State with respect to their

- **violent crime**: per 100,000 population
- **murder** : per 100,000 population
- **poverty**: percent below the poverty line
- **single**: percentage of lone parents
- **urban**: level of urbanicity

Dummy Variables

Example: Y = crime rate; X_1 = poverty; Z = Urban; (Low, Medium, High)

Only 2 dummy variables (e.g. d_{Low} and d_{Medium}) are needed to represent a variable with 3 levels.

Z is converted into two binary dummy variables:

d_{Low} = 1, 0, 0
d_{Medium} = 0, 1, 0

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 d_{Low} + \beta_3 d_{Medium} + \beta_4 x_1 \times d_{Low} + \beta_5 x_1 \times d_{Medium} + \varepsilon$$

Test coefficients β_4; $\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases}$ and β_5; $\begin{cases} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{cases}$

Dummy Variables

US crime data. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High”

state	urban
AK	Low
AR	Low
IA	Low
ID	Low
KY	Low
ME	Low
AL	Medium
GA	Medium
KS	Medium
MN	Medium
MO	Medium
NC	Medium
AZ	High
CA	High
CO	High
CT	High
DE	High

The variable **urban** is a categorical variable with three levels “**Low**”, “**Medium**” and “**High**”

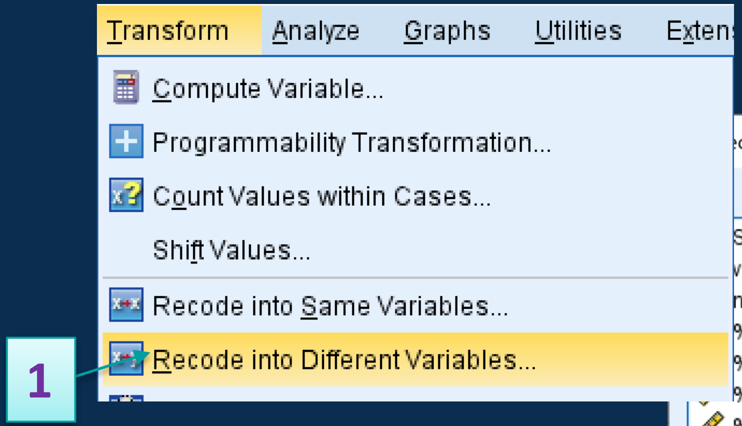
Dummy coding of **urban** ($k=3$)

d1	d2	d3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

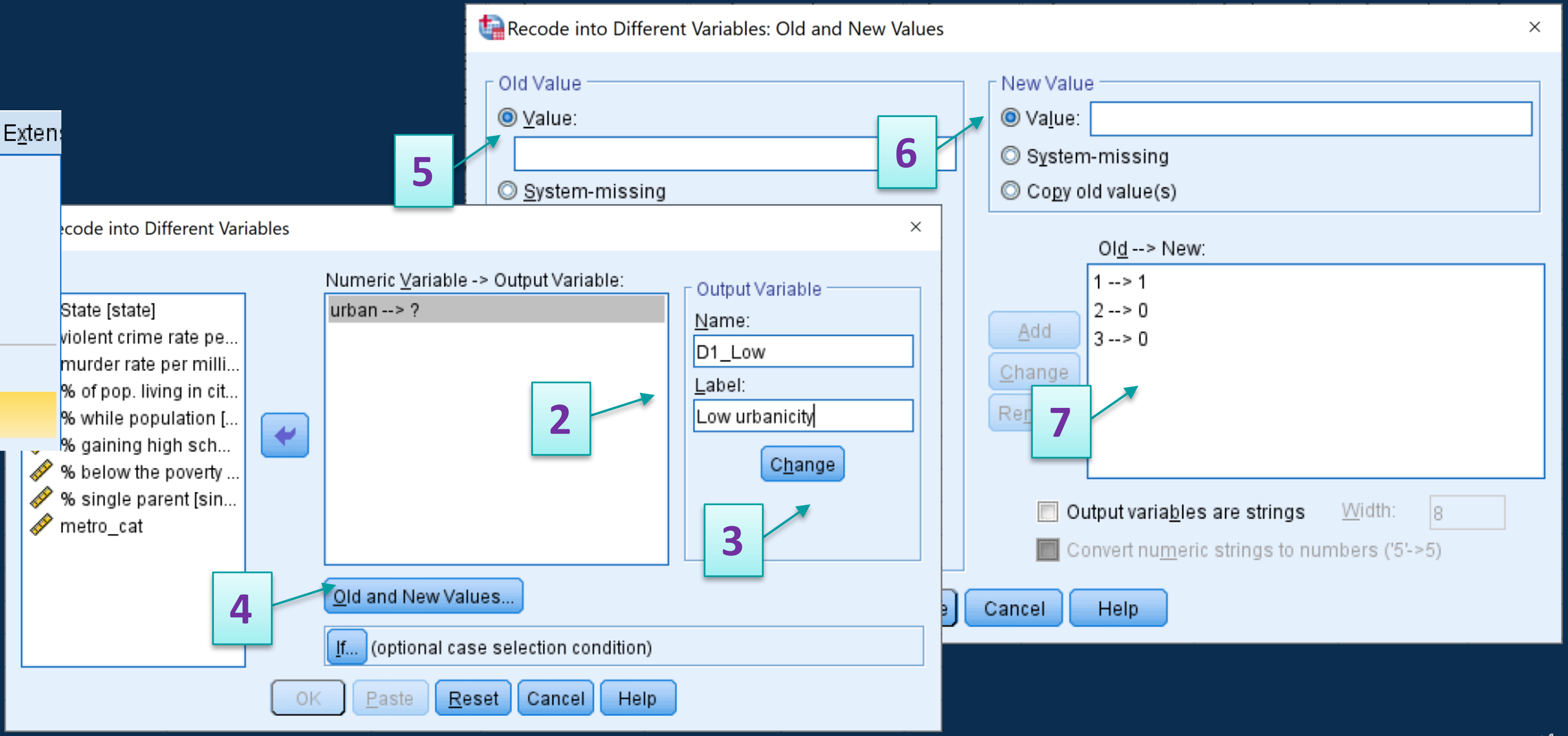
SPSS Slide: ‘how to’

Researchers believe there is a relationship between Violent Crime and poverty and the level of urbanicity in an area modifies this effect. The variable urban is a categorical variable with three levels “Low”, “Medium” and “High” and needs to be converted to dummy variables to include in the regression.

Step 1: Generating dummy variables for ‘urban’ variable from US crime

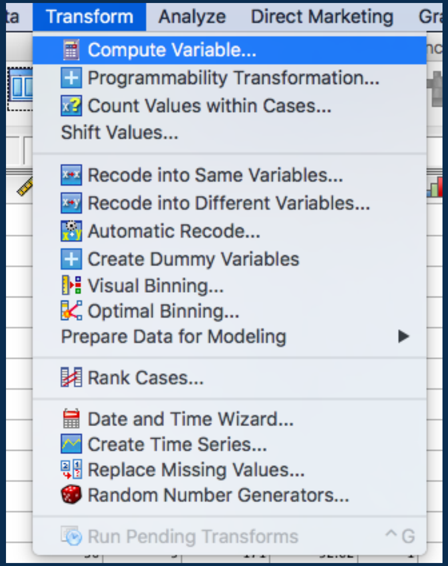


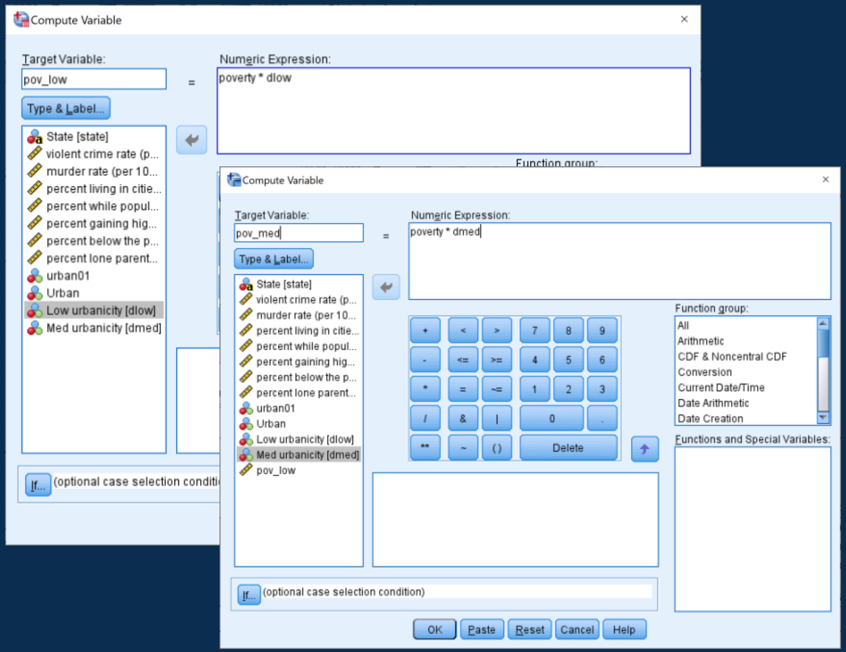
Use **‘Transform’ ->’ Recode into Different Variables’**

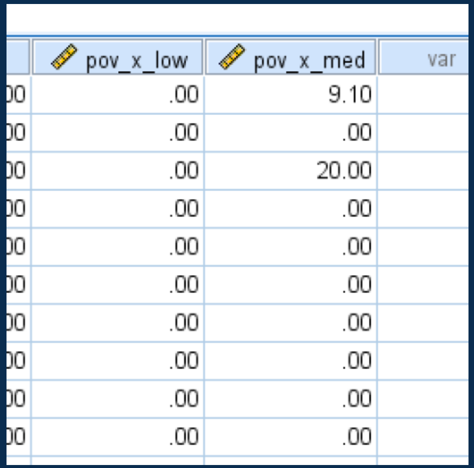


SPSS Slide: ‘How to’ Steps

- Create an interaction term poverty_x_dlow and poverty_x_dmed where high urbanicity is the reference from Lecture_9b_data.sav
- Use ‘Transform’ -> ‘Compute variable’
- In ‘Target variable’ write the name of your interaction term: “pov_x_low”
- In ‘Numeric Expression’ drag ‘poverty’ times (*) ‘low’ and accept.
- Repeat for “pov_x_med”





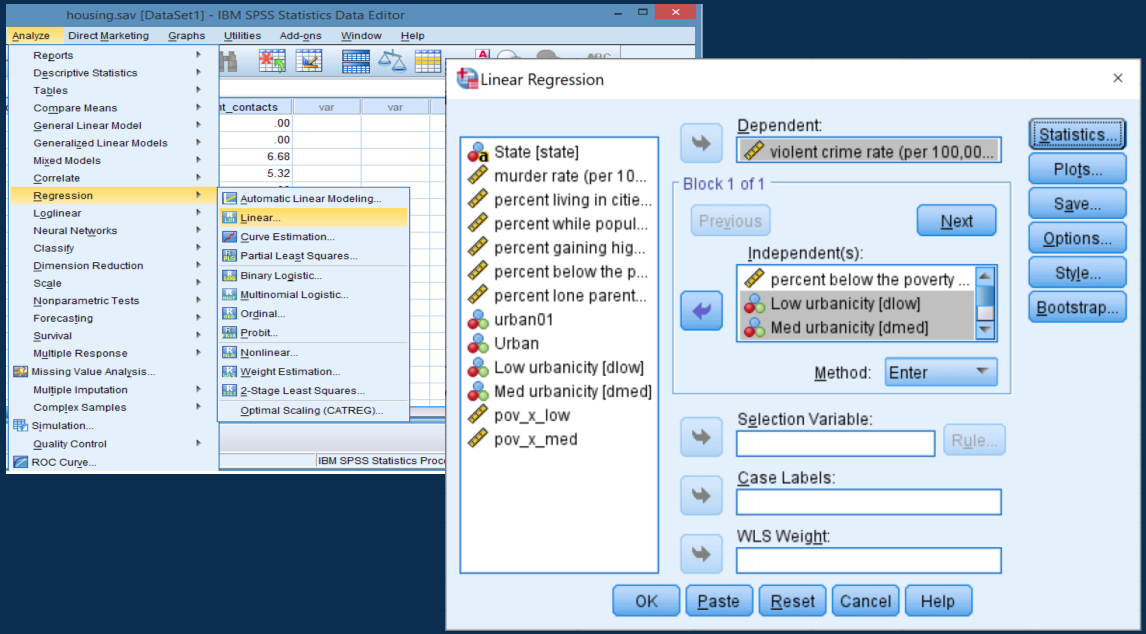


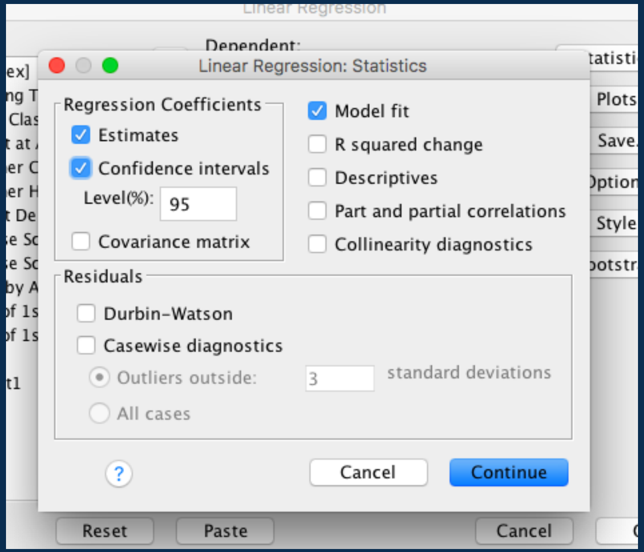
New variables in data set

SPSS Slide: ‘How to’ Steps

- Estimating the interaction effect **pov_x_low and pov_x_med** in a multiple linear regression model for crime rate, poverty, dlow and dmed from lecture_9b_data.sav data

- 1) Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’
- 2) In dependent put ‘crime’ and in independent put ‘poverty’, ‘dlow’, ‘dmed’, ‘pov_x_low’ and ‘pov_x_med’





Output and Interpretation

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	Upper Bound
1	(Constant)	-296.662	179.306		-1.655	.105	-658.029	64.705
	percent below the poverty line	74.694	12.195	.776	6.125	.000	50.116	99.271
	Low urbanicity	263.137	468.308	.194	.562	.577	-680.676	1206.949
	Med urbanicity	481.824	337.218	.467	1.429	.160	-197.795	1161.442
	pov_x_low	-55.812	30.105	-.650	-1.854	.070	-116.485	4.862
	pov_x_med	-58.218	22.288	-.876	-2.612	.012	-103.136	-13.299

a. Dependent Variable: violent crime rate (per 100,000 people)

$$\boldsymbol{crime = -296.662 + 74.694 poverty + 263.137 low + 481.824 med - 55.812 pov * low - 58.218 pov * med}$$

The Coefficient of **pov** $\times$ **low** interaction is – 55.812, p=0.070
The Coefficient of **pov** $\times$ **med** interaction is – 58.218, p=0.012

Effect of poverty on crime decreases in low and medium urbanised areas compared to high urbanised areas

The mean crime rate at average poverty level (mean = 14.2588) for low urbanised states = -296.662 + 74.694*14.2588 +263.137-55.812*14.2588 = 235.71 per 100,000 people.
The mean crime rate at average poverty level (mean = 14.2588) for med urbanised states = -296.662 + 74.694*14.2588 +481.824-58.218*14.2588 = 420.09 per 100,000 people, only the interaction between poverty and med urbanised areas showed a significant effect.

Interaction & Type of Variables

Interaction between variables where both independent variables (x_1 and Z) are either categorical or continuous is handled in the same way, i.e., by creating cross-product terms:

- **continuous × continuous**
- **categorical × categorical**

Example: Continuous × Continuous Interaction

- In Lecture_9a_data.sav, The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS), both height and reading scores are **continuous** variables

- There is **no reason** to believe that reading score will affect weight, but let's see an example involving reading score to demonstrate how we can investigate interactions when the two independent variables are continuous.

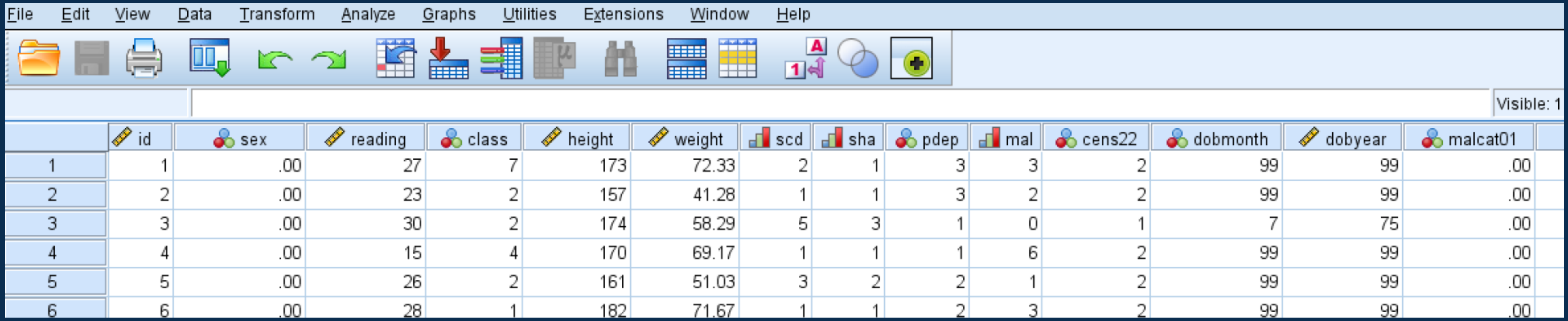
- We are interested in testing if reading score modifies the effect of height on weight

- This will require **computing a new variable** – the cross-product of height and reading score (as we did before for height x sex) **height × reading**

- And then **including the product term** as an additional predictor in a regression model

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

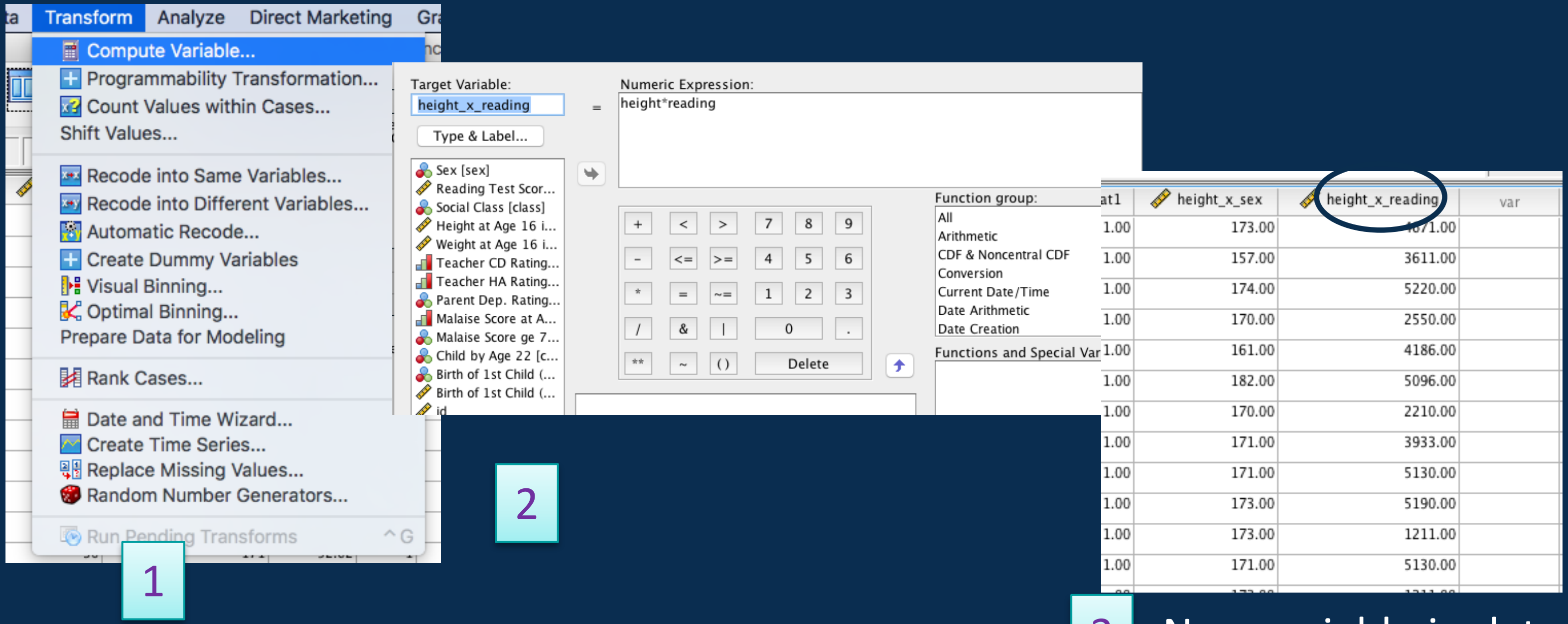
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **sex**: gender of child (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)

SPSS Slide: ‘How to’ Steps

- Create an interaction term height_x_reading from ncds.sav data

- Use ‘Transform’ -> ‘Compute variable’
- In ‘Target variable’ write the name of your interaction term: “height_x_reading”
- In ‘Numeric Expression’ drag ‘height’ times (*) ‘reading’ and accept.

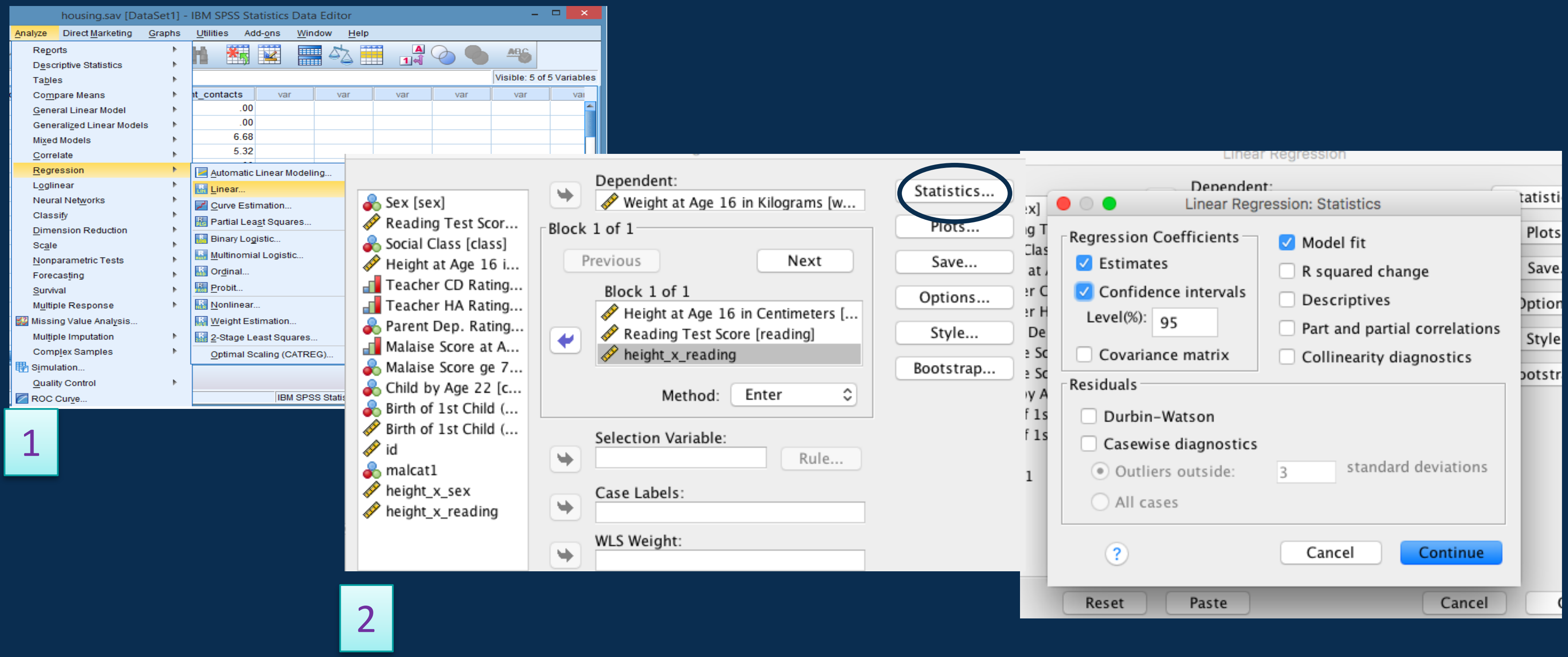


3 New variable in data set

SPSS Slide: 'How to' Steps

- Estimating the interaction effect **height_x_reading** in a multiple linear regression model for weight, height and reading from lecture_9_a_data.sav data

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



Output and Interpretation

- The Coefficient of **height × reading** interaction is - 0.005
- **Negative interaction** effect means that:
  - Effect of height decreases as reading scores increases, and
  - Effect of reading scores decreases as height increases
- However, the height × reading interaction is **not significant** (p=0.286) The height-weight relationship does not significantly differ by reading score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-67.302	19.900		-3.382	.001	-106.352	-28.251
	Height at Age 16 in Centimeters	.748	.119	.622	6.265	.000	.514	.983
	Reading Test Score	.858	.800	.582	1.072	.284	-.712	2.429
	hxr	-.005	.005	-.588	-1.068	.286	-.015	.004

a. Dependent Variable: Weight at Age 16 in Kilograms

$$\boldsymbol{weight = -67.302 + 0.748height + 0.858reading - 0.005height * reading}$$

Presenting Continuous × Continuous Interactions: Tabular Format

- The **effect for height on weight** is: $\beta 1 + \beta 3 \times$ reading
- The **effect for reading on weight** is: $\beta 2 + \beta 3 \times$ height

- For example, in the model for NCDS data, effect of height can be calculated at different values (e.g., quartiles) of reading scores, and vice-versa:

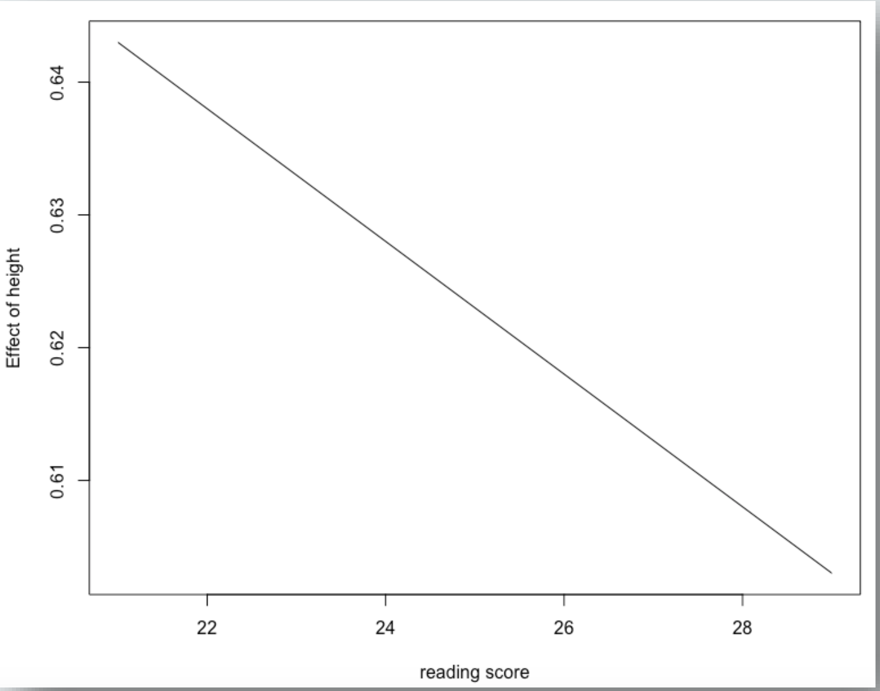
$$\mathbf{weight} = -\mathbf{67.302} + \mathbf{0.748height} + \mathbf{0.858reading} - \mathbf{0.005height} * \mathbf{reading}$$

Reading score (quartiles)	Effect for height: 0.748 – 0.005 × reading
reading= 21 (first quartile)	0.748 – 0.005 × 21 = 0.643 kg/cm
reading= 27 (median)	0.748 – 0.005 × 27 = 0.613 kg/cm
reading= 29 (3rd quartile)	0.748 – 0.005 × 29 = 0.603 kg/cm

*Similar table can be created for the effect of reading scores at varying values of height

Presenting Continuous × Continuous Interactions: Graphical Format

Reading score (quartiles)	Effect for height:
21	0.643
27	0.613
29	0.603



- The plot shows the effect of **height** as a function of reading scores
- Effect of height **decreases** as reading scores **increases**
- Similar plot can be created for the effect of reading scores as a function of height

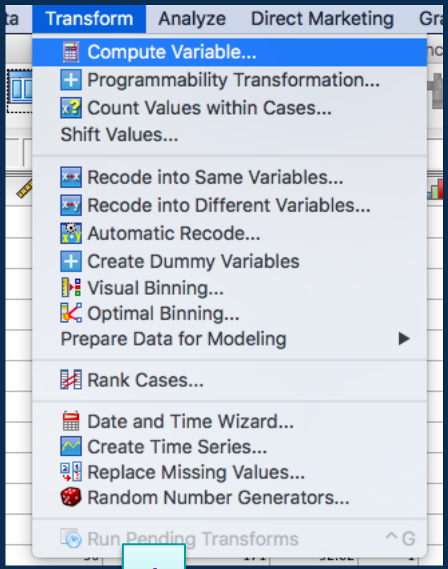
Example: Categorical × Categorical Interaction

- In NCDS data, **sex** and **malcat** are two **categorical** (binary) variables
- The variable **malcat (0=low, 1=high)** represents a categorised version (median split) of the continuous variable malaise scores (mal) (a feeling of general discomfort/uneasiness)
- Suppose we are interested in testing the **sex** × **malcat** interaction
- As before, this will require computing a new variable – the **cross-product of sex and malcat**, and including it as an additional predictor in the regression model.

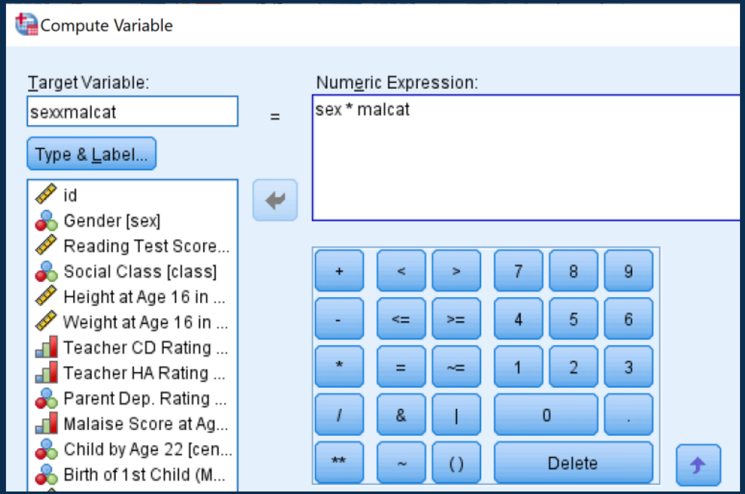
SPSS Slide: ‘How to’ Steps

- Create an interaction term sex_x_malcat from lecture_9a_data.sav.

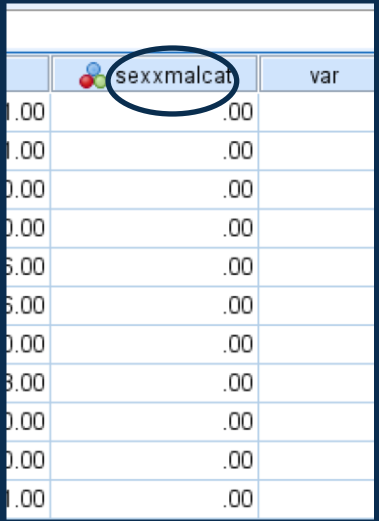
- Use ‘Transform’ -> ‘Compute variable’
- In ‘Target variable’ write the name of your interaction term: “sex_x_malcat”
- In ‘Numeric Expression’ drag ‘Gender’ times (*) ‘malcat’ and ‘Ok’.



1



2

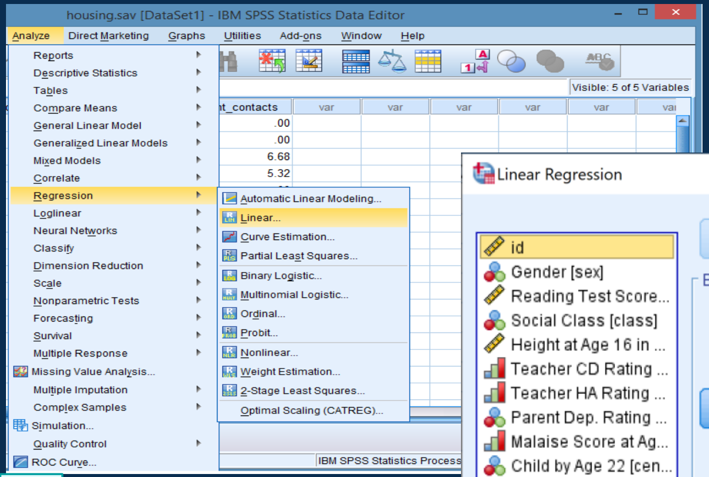


3 New variable in data set

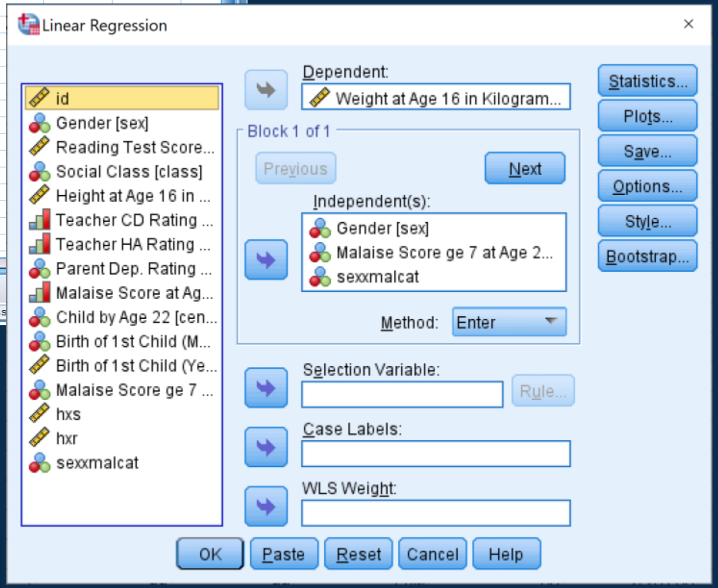
SPSS Slide: ‘How to’ Steps

- Estimating the interaction effect **sex_x_malcat** in a multiple linear regression model for weight, sex and malcat from lecture_9_a_data.sav data

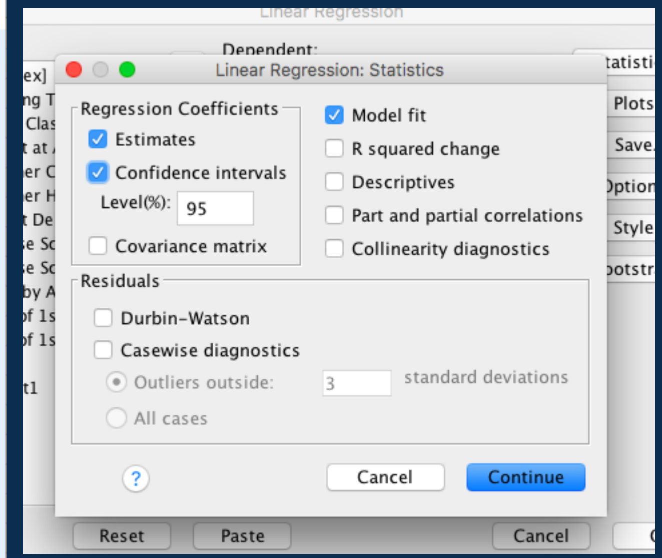
- 1) Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’
- 2) In dependent put ‘weight’ and in independent put ‘sex’, ’malcat’, ‘sex_x_malcat



1



2



Output and Interpretation

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)	59.534	.435		136.706	.000	58.680	60.389
	Gender	-4.676	.626	-.242	-7.474	.000	-5.904	-3.448
	Malaise Score ge 7 at Age 22 0 = No, 1=Yes	-.324	1.892	-.010	-.171	.864	-4.038	3.390
	sexxmalcat	.690	2.238	.018	.309	.758	-3.701	5.082

a. Dependent Variable: Weight at Age 16 in Kilograms

$$\mathbf{weight = 59.534 - 4.676\ sex\ - 0.324 malcat + 0.690\ sex * malcat}$$

- Coefficient of **sex** × **malcat** interaction = 0.690
- Positive interaction effect means that:
  - Effect of gender **is higher** for high (=1) category of malaise score, and
  - Effect of malaise score **is higher** for girls (sex=1) than for boys (sex=0)
- The sex × malcat interaction is **not significant** (p=0.758)

Presenting Categorical × Categorical Interactions

- Effect of each variable can be estimated at each level of the other variable
- For example, effect of gender can be calculated at low and high levels of malaise scores
- Effect of sex on weight = $\beta_1 + \beta_3 \times$ malcat = -4.676 + 0.690 x malcat

$$\mathbf{weight = 59.534 - 4.676\ sex\ - 0.324 malcat + 0.690\ sex * malcat}$$

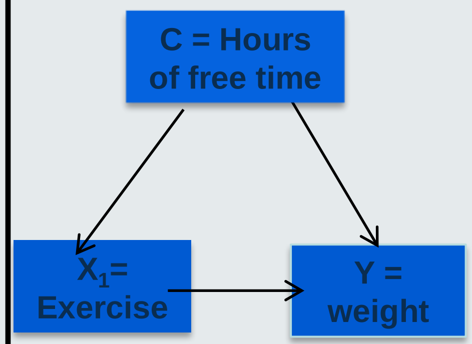
malcat	= -4.676 + 0.690×malcat
Low (=0)	-4.676 kg
High (=1)	-3.986 kg

Difference in mean weight between girls (sex=1) and boys (sex=0) at low malaise scores

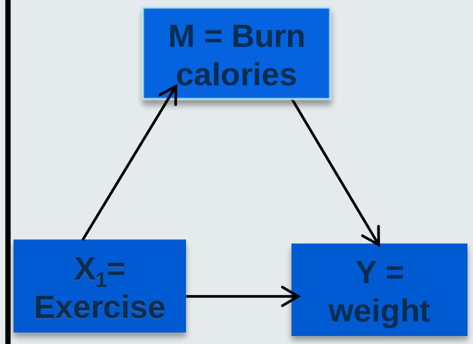
Difference in mean weight between girls (sex=1) and boys (sex=0) at high malaise scores

Confounding vs Mediation vs Interaction

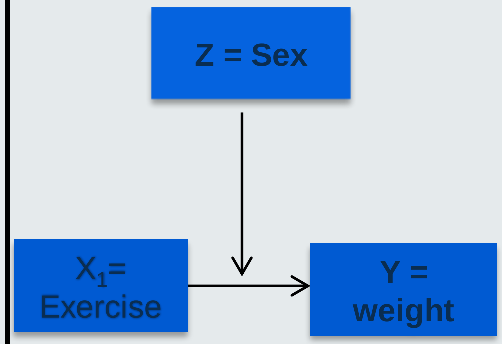
- Both confounder, mediator and moderator, are third variables that explain a part (or most) of the association between an independent and dependent variable.



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

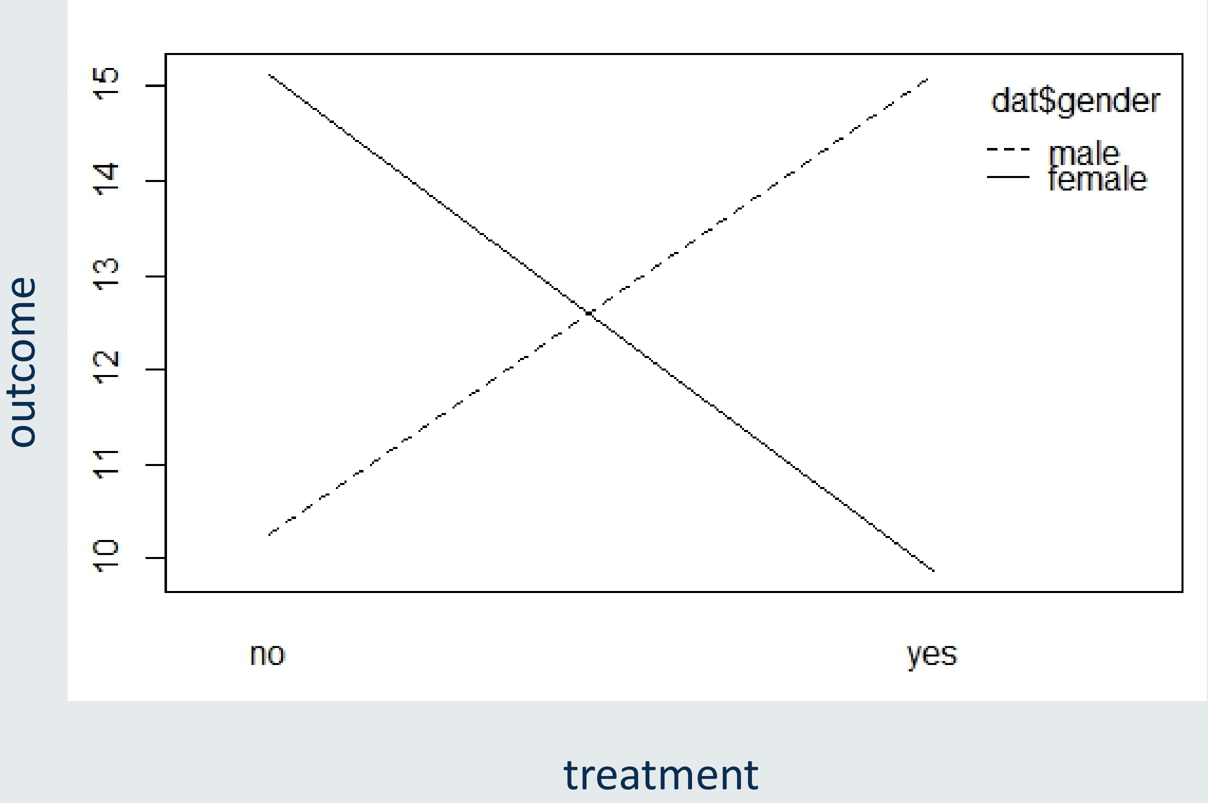


A **moderator (Z)** modifies the effect of an independent variable on a dependent variable. The association varies depending on the values/levels of Z

Knowledge Check

Q1.

- The next plot shows the interaction effect between **treatment** and **gender** variables (two categorical variables) on a continuous outcome.
- The **P value** for treatment*gender term = 0.02



Interpret the interaction.

Knowledge Check

Q2.

- A study is investigating the effect of maternal deprivation on lowbirthweight. There are other 3 factors that have a role on this association:

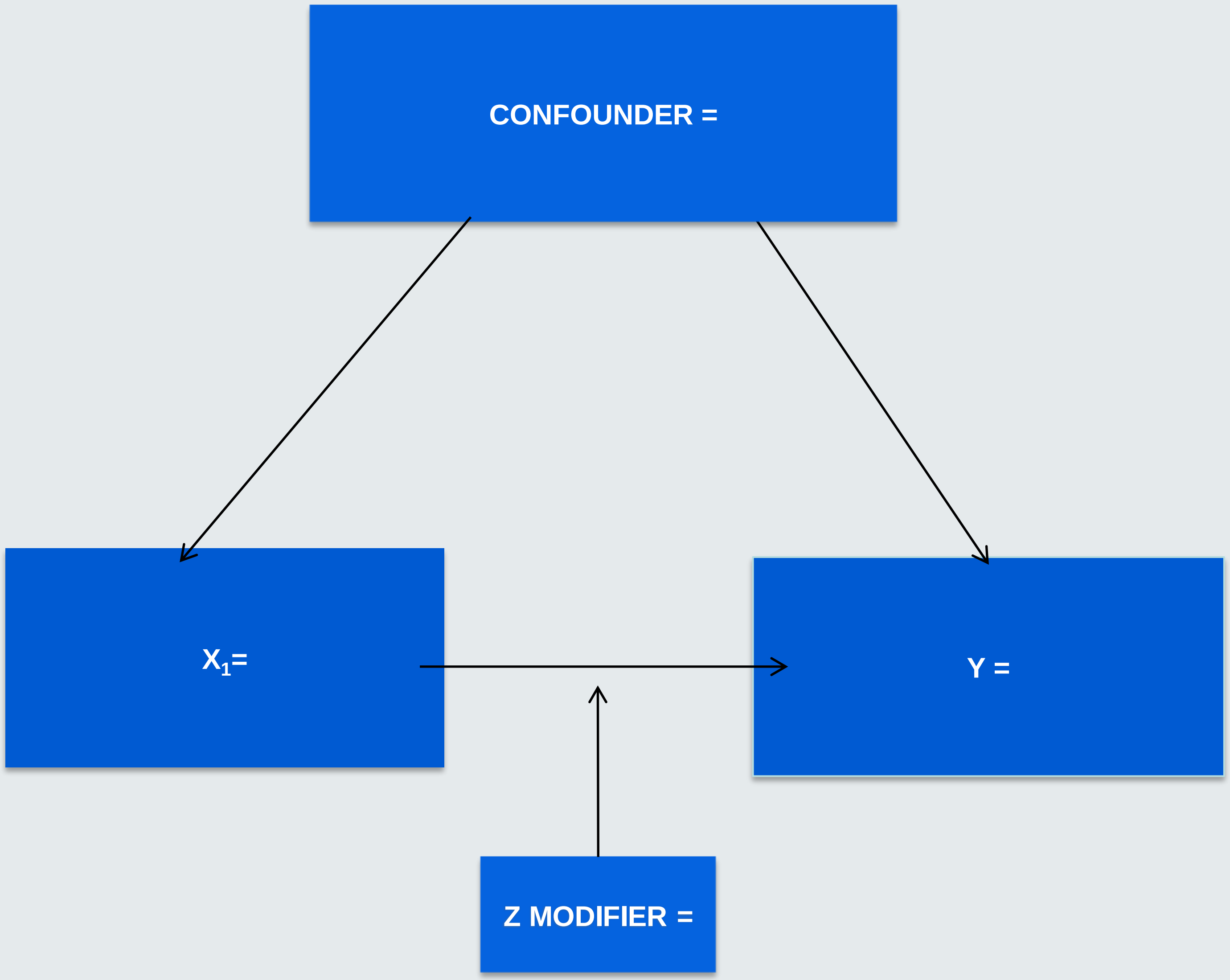
  - Diet
  - Smoking
  - Age

We know that:

1. Diet is on the causal pathway through which deprivation might act on low birth weight.
2. The association between Maternal deprivation and lowbirthweight differs between those that are smokers and not smokers
3. Age affects maternal deprivation and lowbirthweight

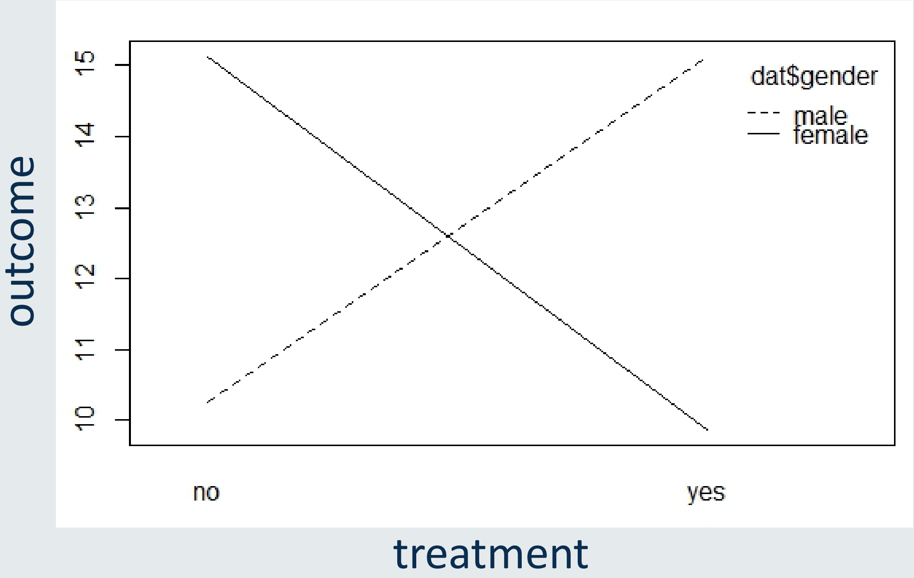
Please fill the boxes in the next diagram with the variable names.

Knowledge Check



Knowledge Check Solutions

Q1. Interpret the interaction.



P value for treatment*gender term= 0.02

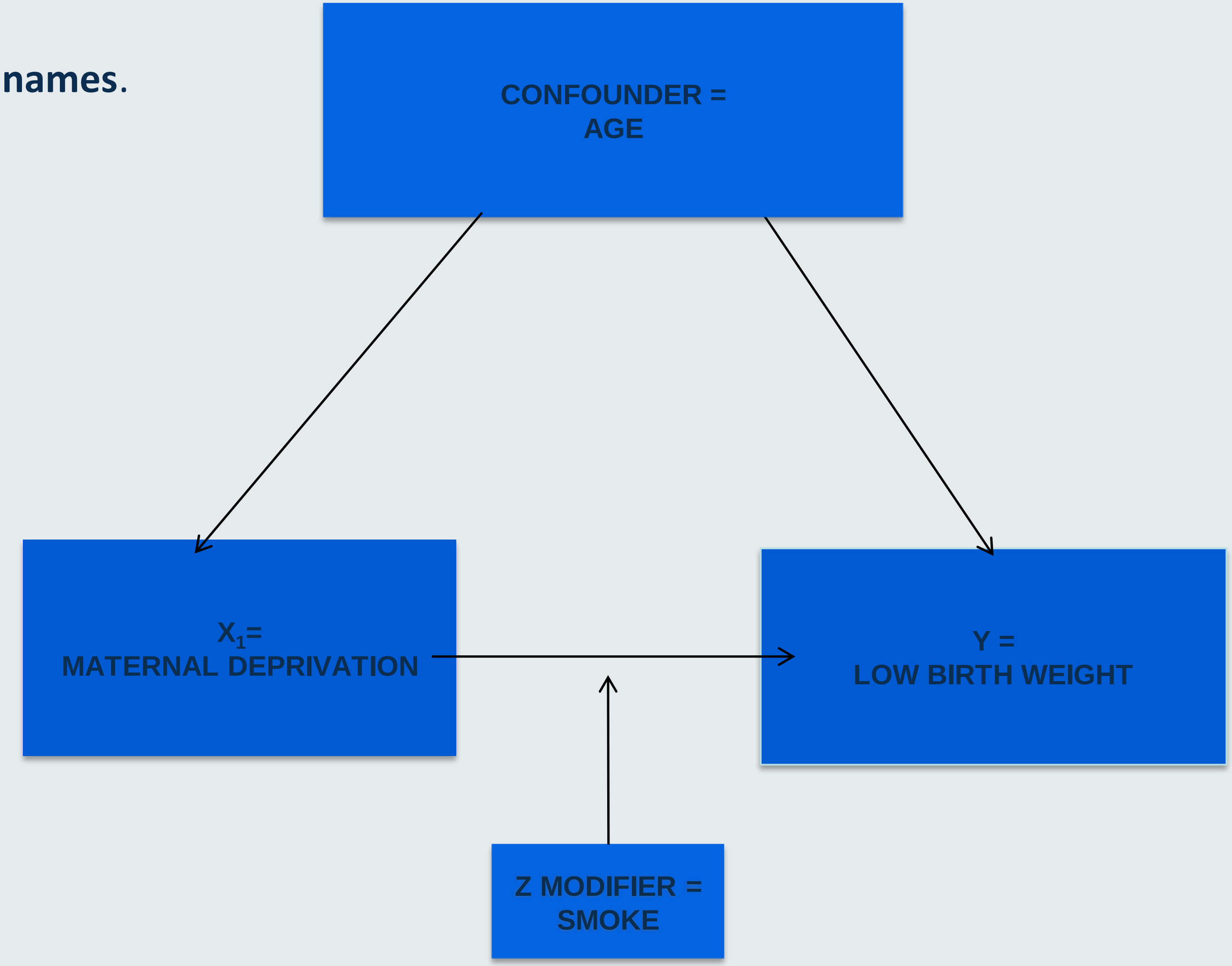
- The effect of treatment*gender term on the outcome is **significantly** different from 0.

- Males under no treatment show a **lower outcome** than females under no treatment
- Males under treatment show a **higher outcome** than women under treatment
- Females under no treatment show a **higher outcome** than males under no treatment.
- Females under treatment, females show a **lower outcome** than male under treatment

- Treatment has the **opposite effect** on men than in women

Knowledge Check Solutions

Q2. Please fill the boxes with the variable names.



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**