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: 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

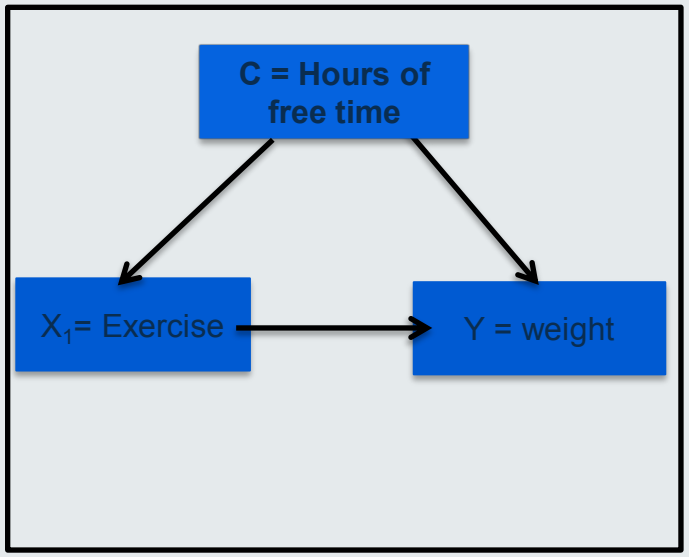
Previously on ‘Introduction to Statistics’

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

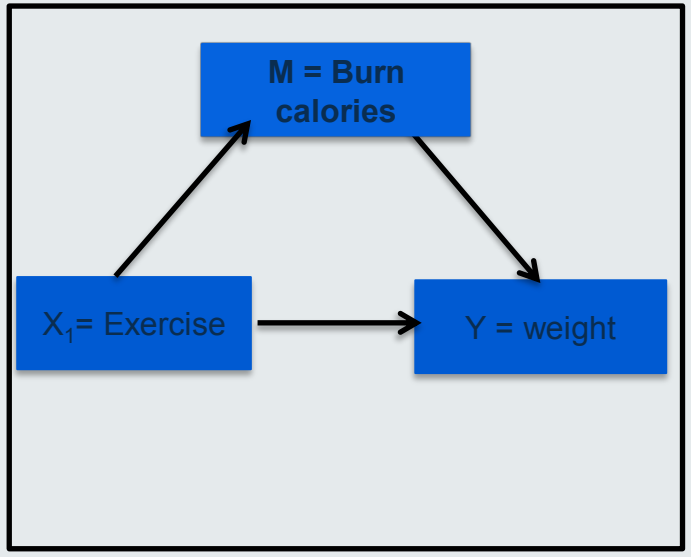
We discussed the different roles that a third variable **X_2** can have while investigating the association between an independent X_1 and a dependent variable Y.

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

Previously on ‘Introduction to Statistics’



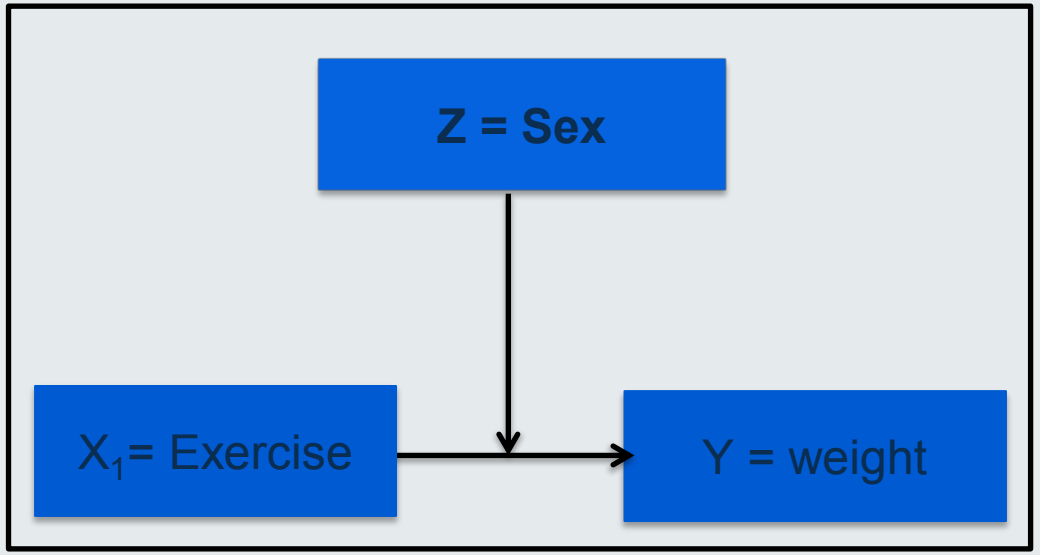
A **confounder** (C) has a common effect on the independent and dependent variables. A confounder **is extrinsic to the causal** pathway.



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.

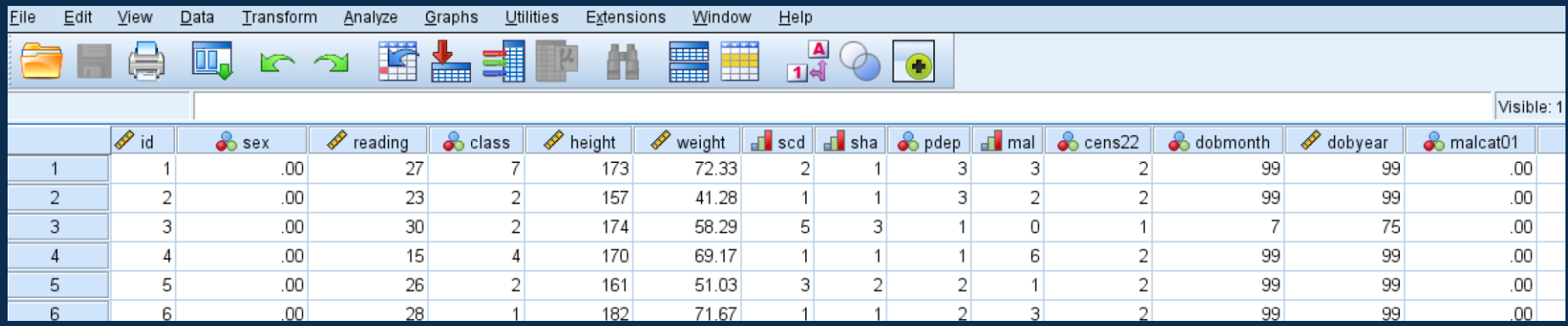
Step 3:

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

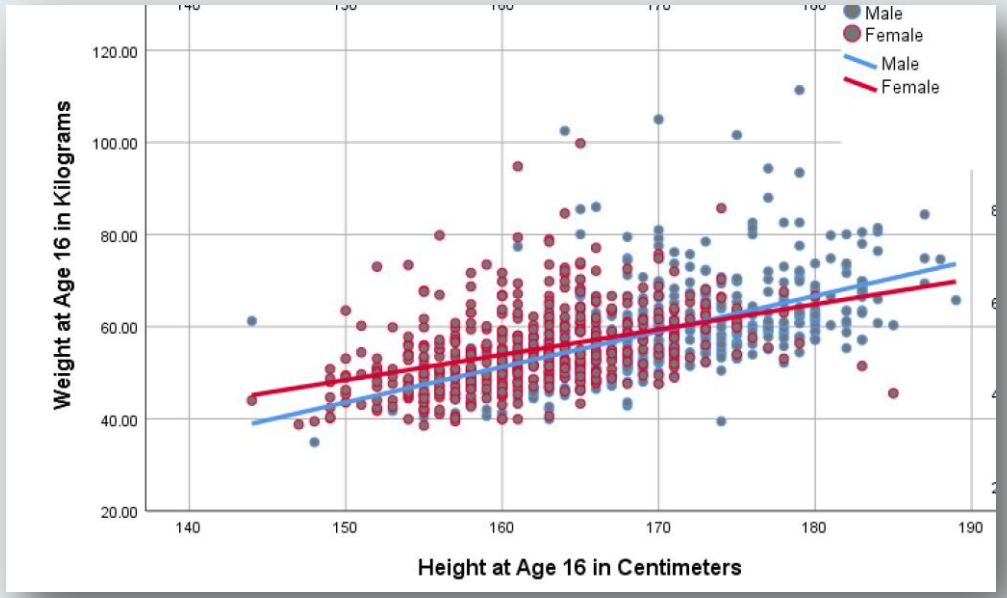
Consider estimating the linear relationship between x_1 = height and Y= weight:

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

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 × 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 × 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?

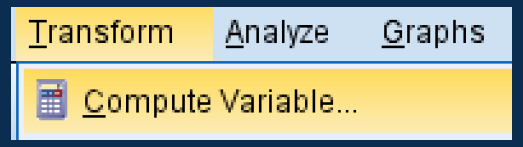
SPSS Slide: ‘How to’ Steps

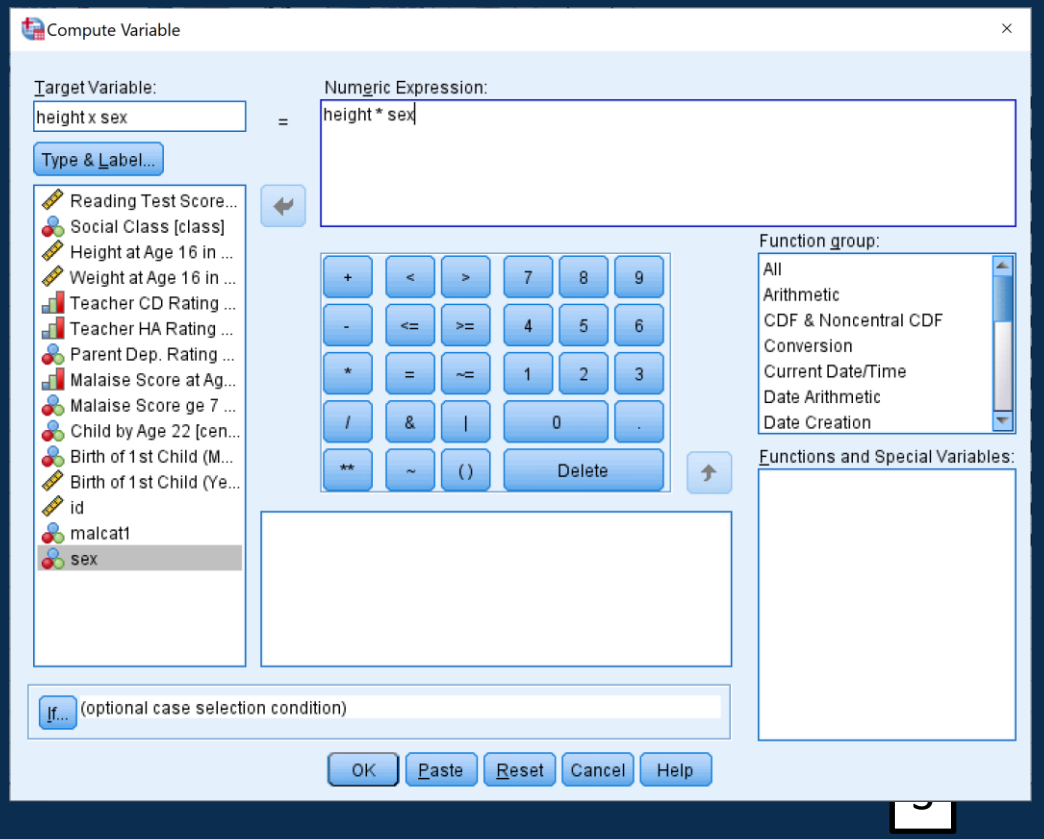
Create an **interaction term height_x_sex** from lecture_9a_data.sav

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 ‘*’





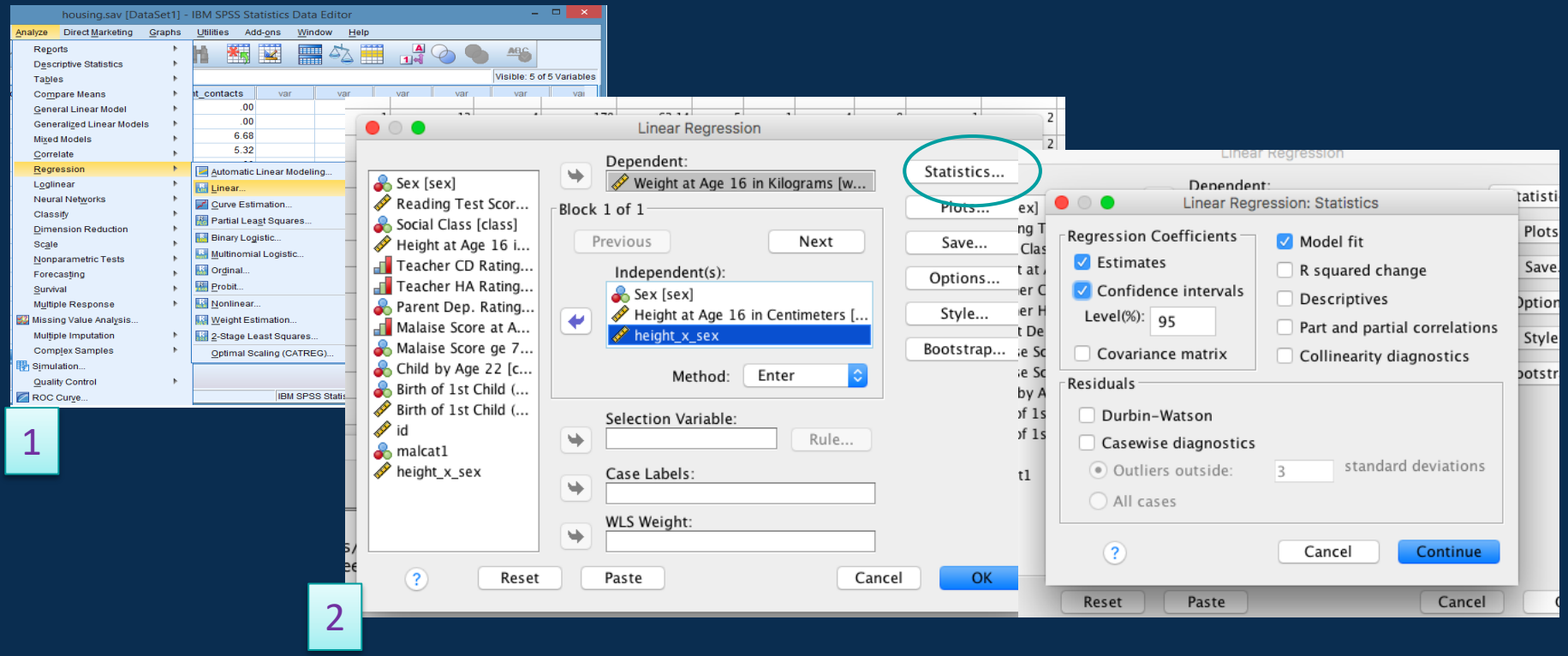
sex	height_x_sex	var
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	
Males	.00	

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

$$weight = b_0 + b_1 height + b_2 sex + b_3 height \text{ ´ } sex$$

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)	-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 $\beta's$

$$\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 Z + \beta_3\, x_1 \times Z + \varepsilon}$$

- $\boldsymbol{\beta_1}$ is interpreted as the effect of $\mathbf{x_1}$ on **Y** when $\mathbf{Z = 0}$
- 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 $\mathbf{x_1 \times Z}$ 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**

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:

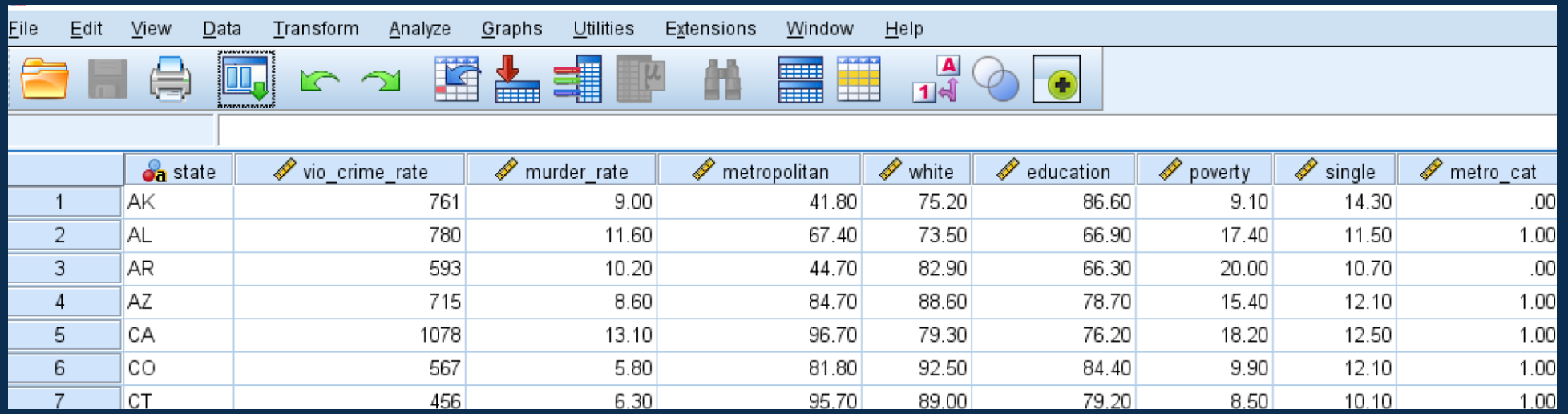
$d_{London} = 1, 0, 0$
$d_{Manchester} = 0, 1, 0$

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

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**”

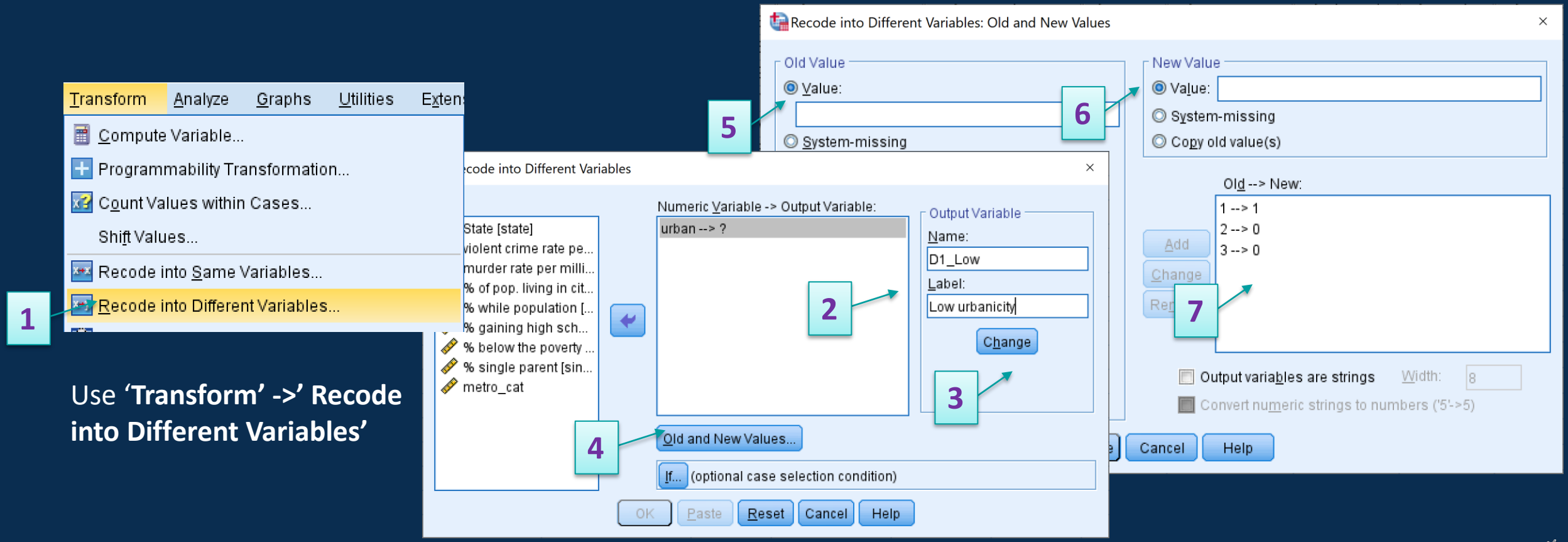
Dummy coding of **urban** (*k=3*)

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.

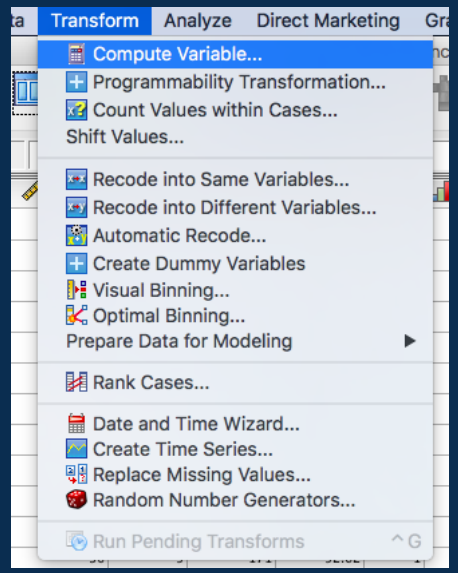
Step 1: Generating dummy variables for 'urban' variable from US crime

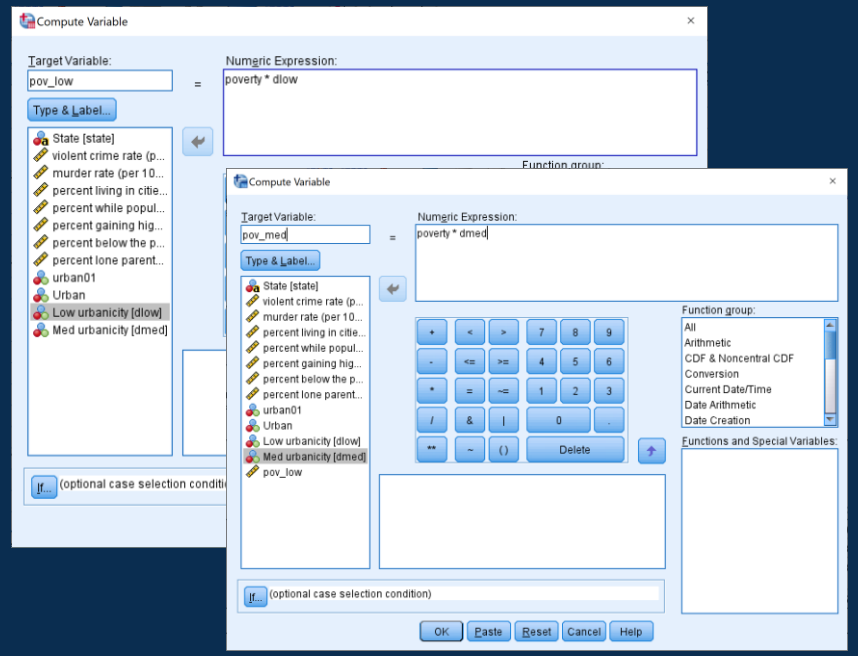


Use **'Transform' ->' Recode into Different Variables'**

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”





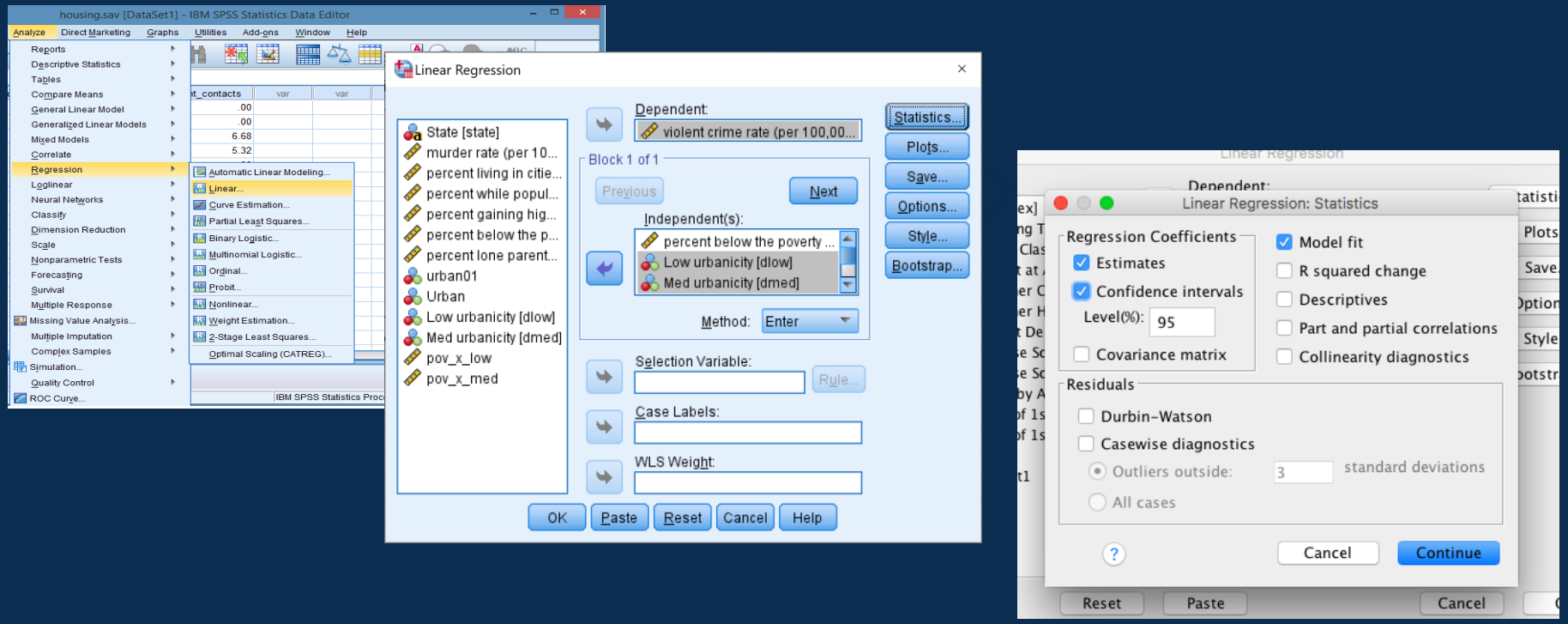
	pov_x_low	pov_x_med	var
00	.00	9.10	
00	.00	.00	
00	.00	20.00	
00	.00	.00	
00	.00	.00	
00	.00	.00	
00	.00	.00	
00	.00	.00	
00	.00	.00	
00	.00	.00	

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		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	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)

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

The Coefficient of **pov** × **low** interaction is – 55.812, p=0.070
The Coefficient of **pov** × **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:

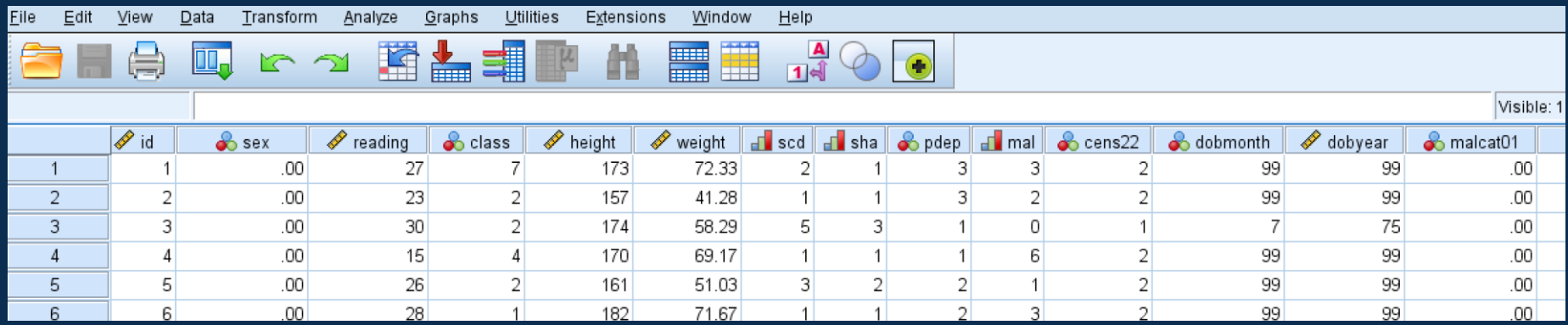
- **continuous × continuous**
- **categorical × categorical**

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

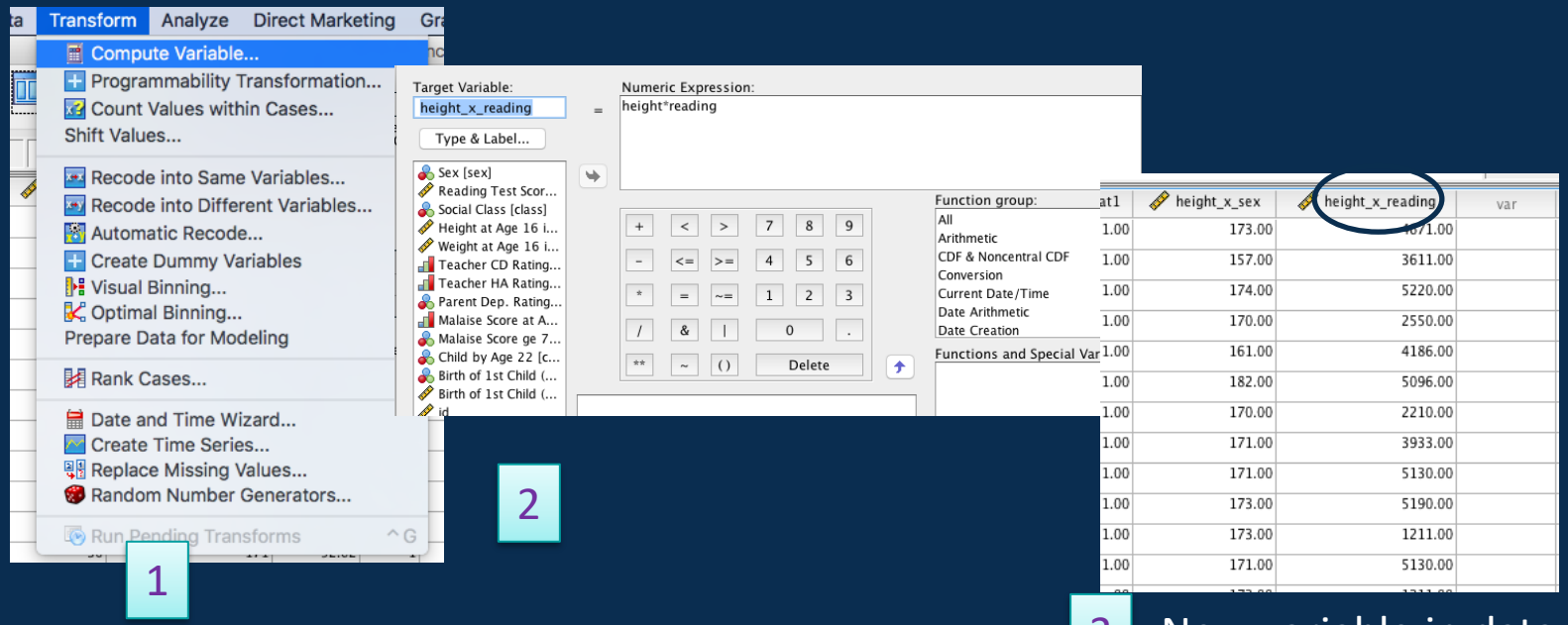
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **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

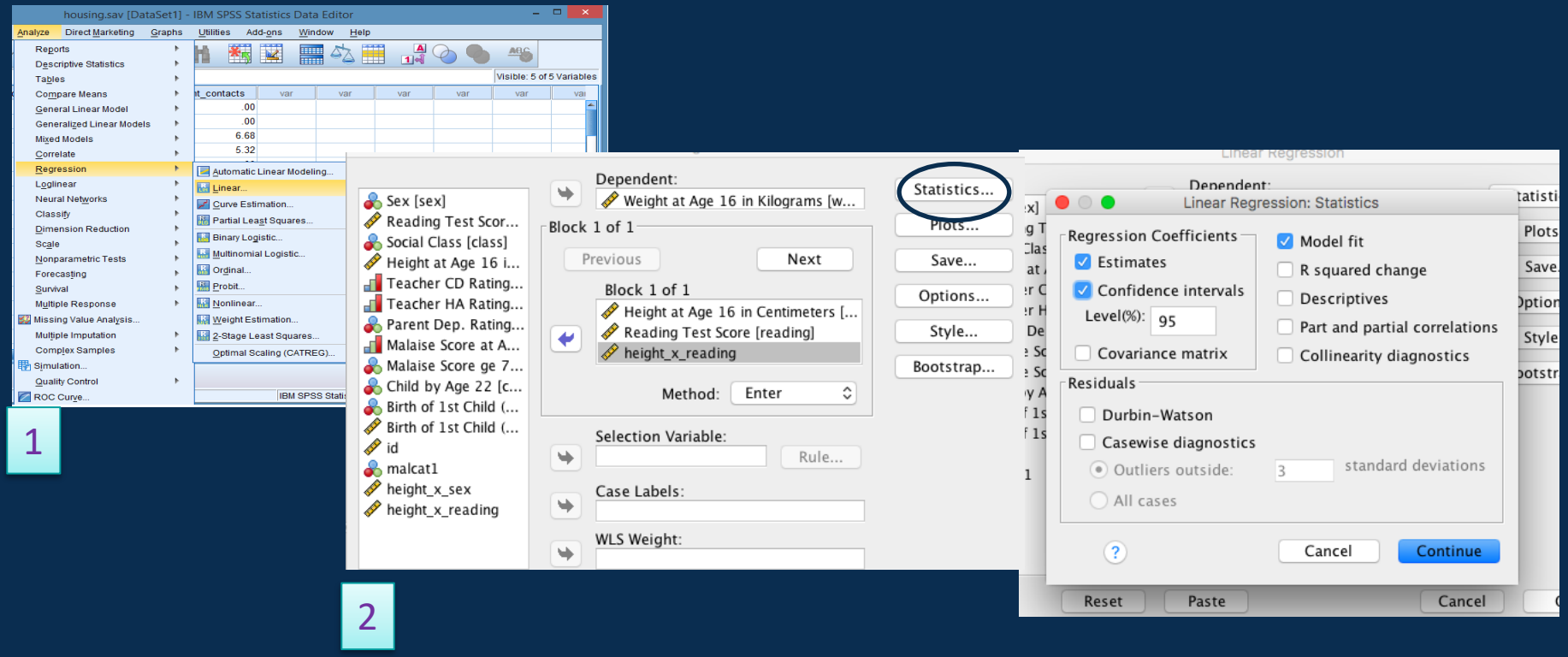
- Use ‘Transform’ -> ‘Compute variable’
- In ‘Target variable’ write the name of your interaction term: “height_x_reading”
- In ‘Numeric Expression’ drag ‘height’ times (*) ‘reading’ and accept.



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

$$weight = -67.302 + 0.748height + 0.858reading - 0.005height * reading$$

Presenting Continuous × Continuous Interactions: Tabular Format

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

- 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:

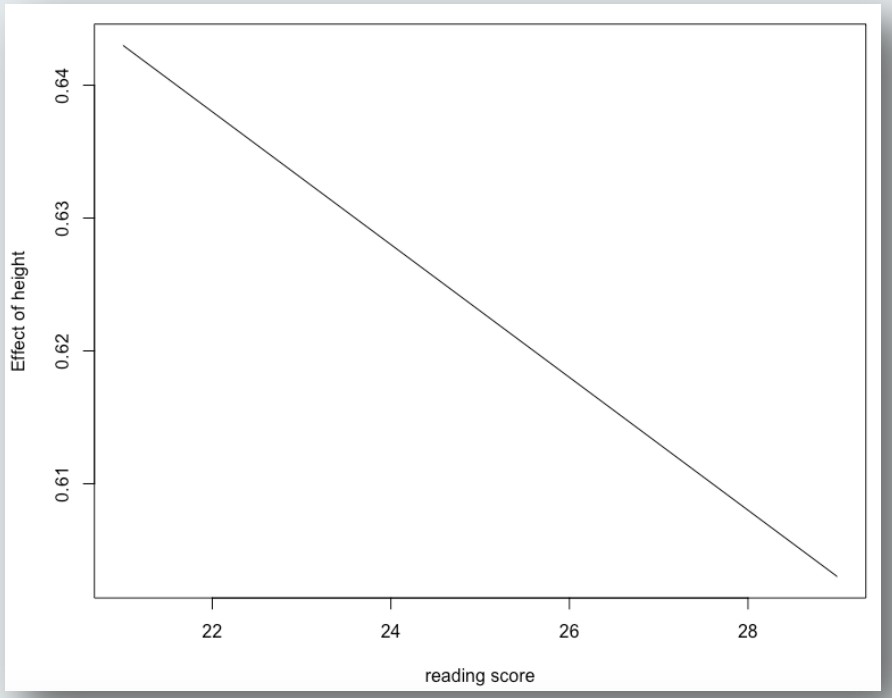
$$\mathbf{weight} = -\mathbf{67.302} + \mathbf{0.748height} + \mathbf{0.858reading} - \mathbf{0.005height} * \mathbf{reading}$$

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

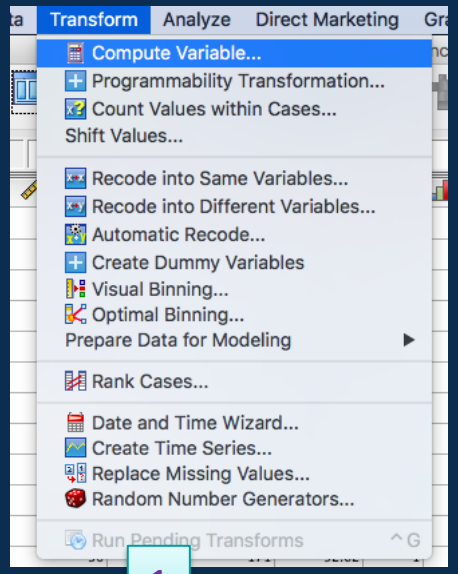
Example: Categorical × Categorical Interaction

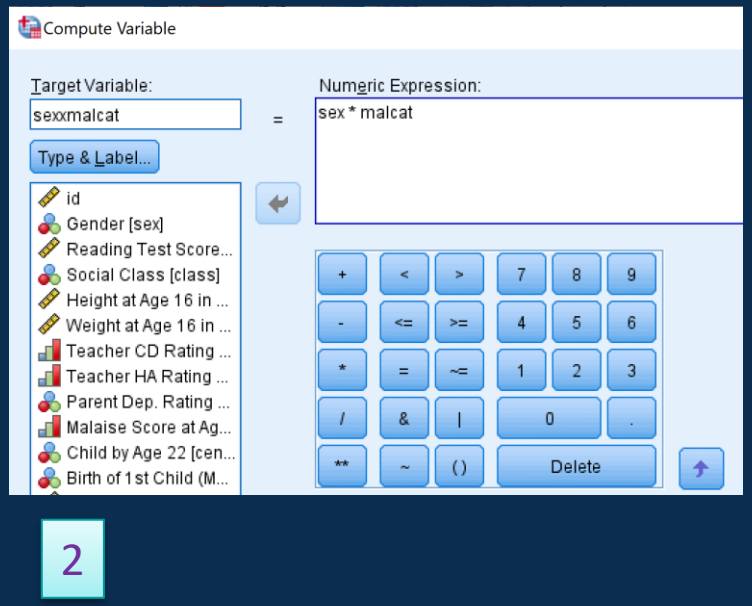
- In NCDS data, **sex** and **malcat** are two **categorical** (binary) variables
- The variable **malcat (0=low, 1=high)** represents a categorised version (median split) of the continuous variable malaise scores (mal) (a feeling of general discomfort/uneasiness)
- Suppose we are interested in testing 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.

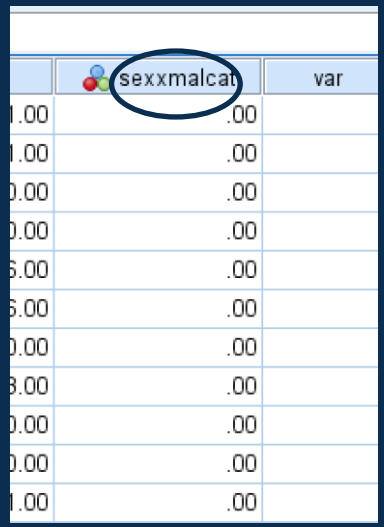
SPSS Slide: ‘How to’ Steps

- 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’.



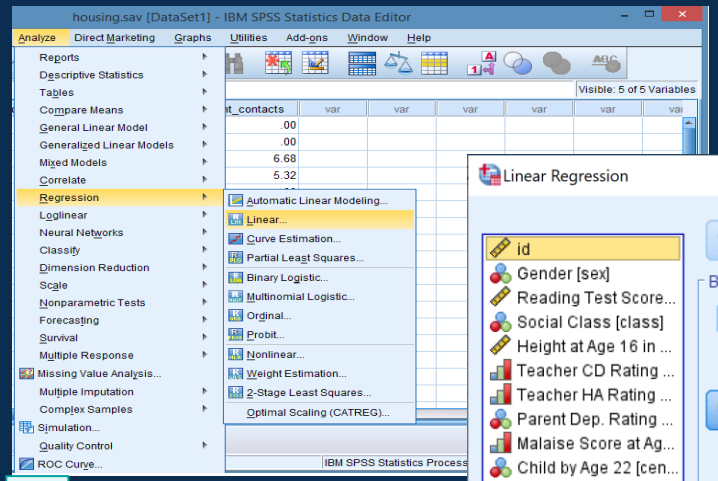


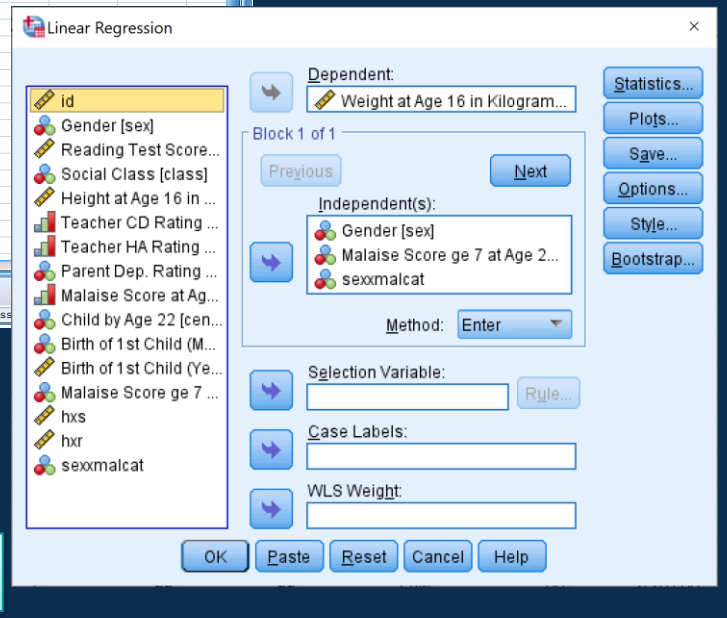


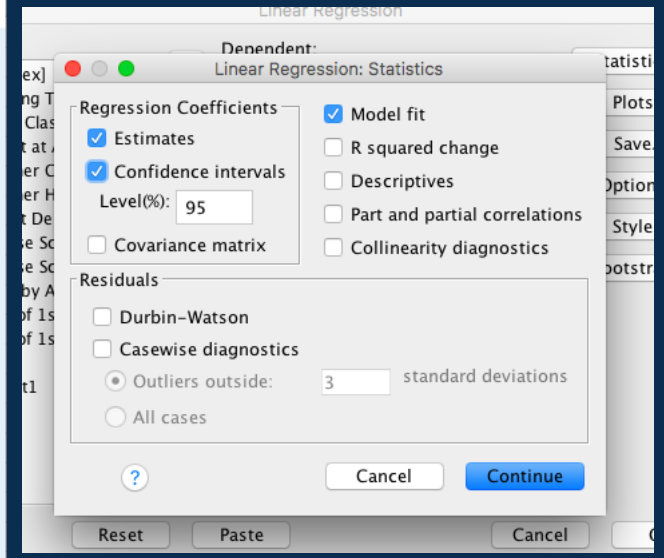
SPSS Slide: ‘How to’ Steps

- 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







Output and Interpretation

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)	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

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

- 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

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

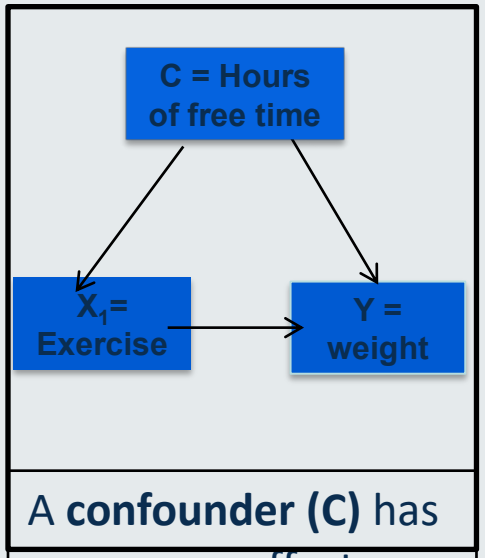
malcat	= -4.676 + 0.690×malcat
Low (=0)	-4.676 kg
High (=1)	-3.986 kg

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

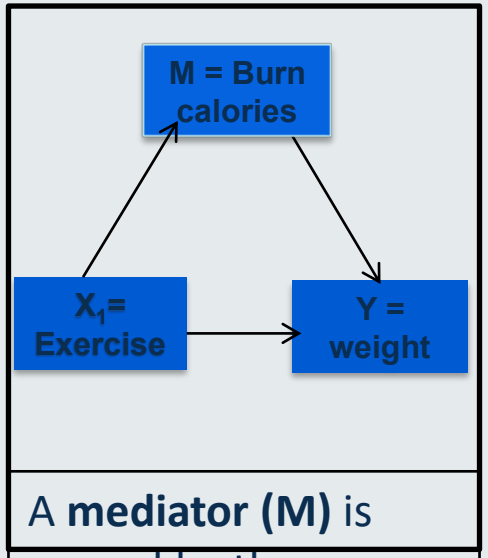
Difference in mean weight between girls (sex=1) and boys (sex=0) at high malaise scores

Confounding vs Mediation vs Interaction

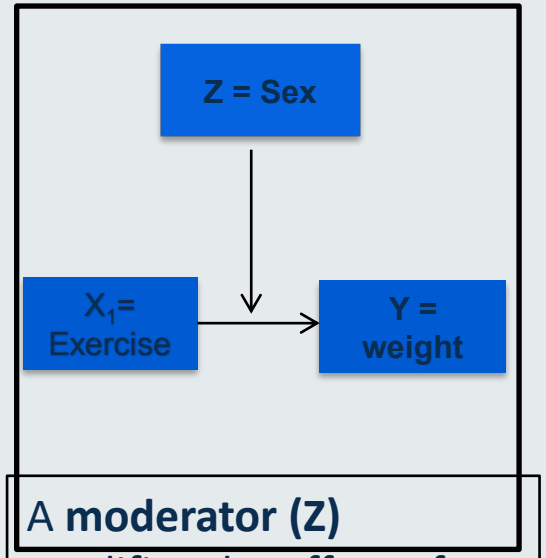
- Both confounder, mediator and moderator, are third variables that explain a part (or most) of the association between an independent and dependent variable.



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

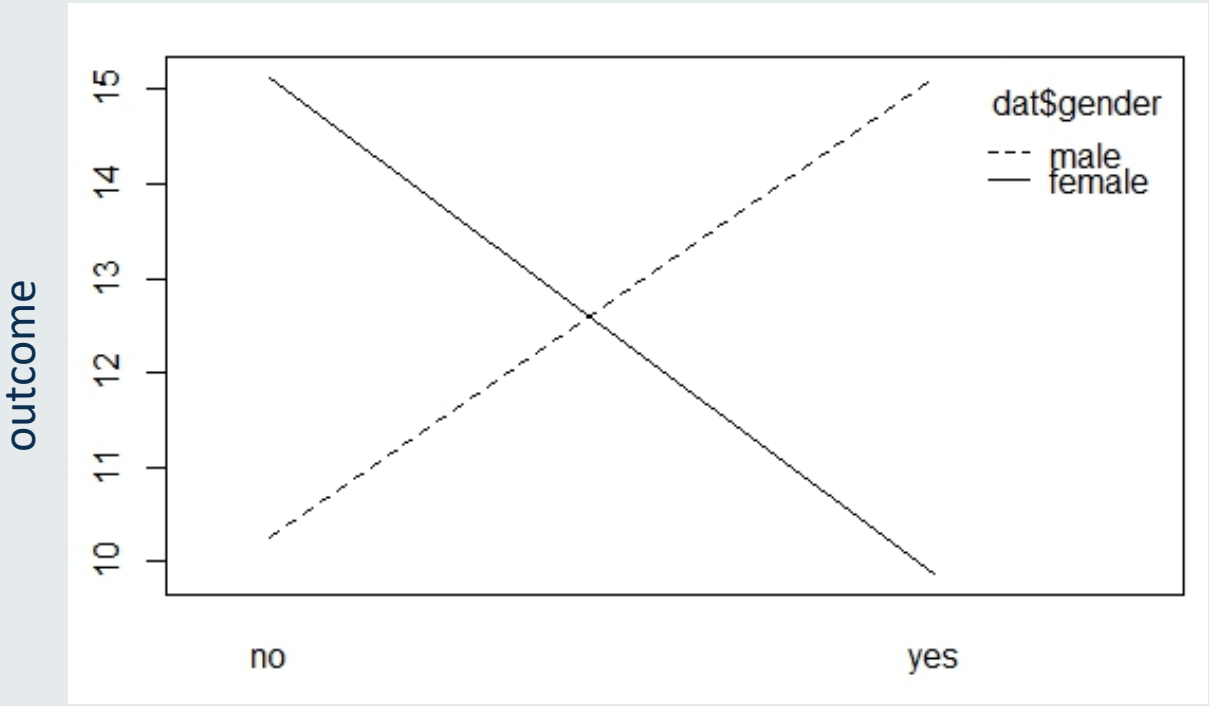


A **moderator (Z)** modifies the effect of an independent variable on a dependent variable. The association varies depending on the values/levels of Z

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:

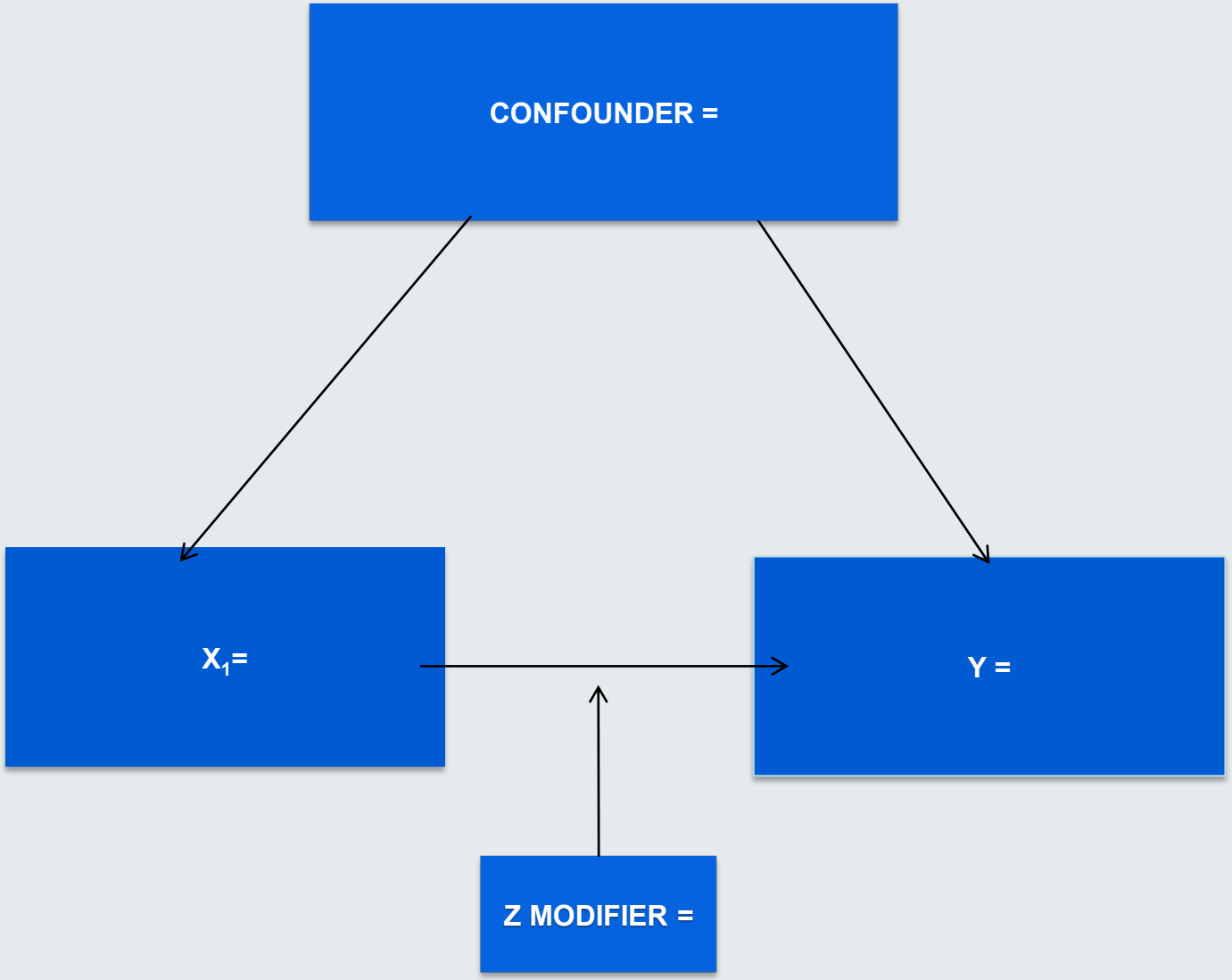
  - Diet
  - Smoking
  - Age

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

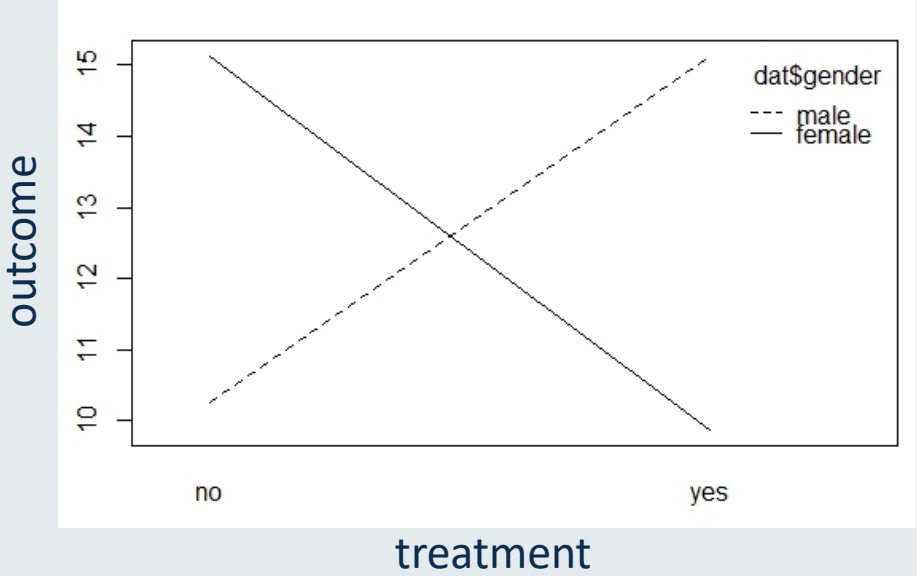
Please fill the boxes in the next diagram with the variable names.

Knowledge Check



Knowledge Check Solutions

Q1. Interpret the interaction.



P value for treatment*gender term= 0.02

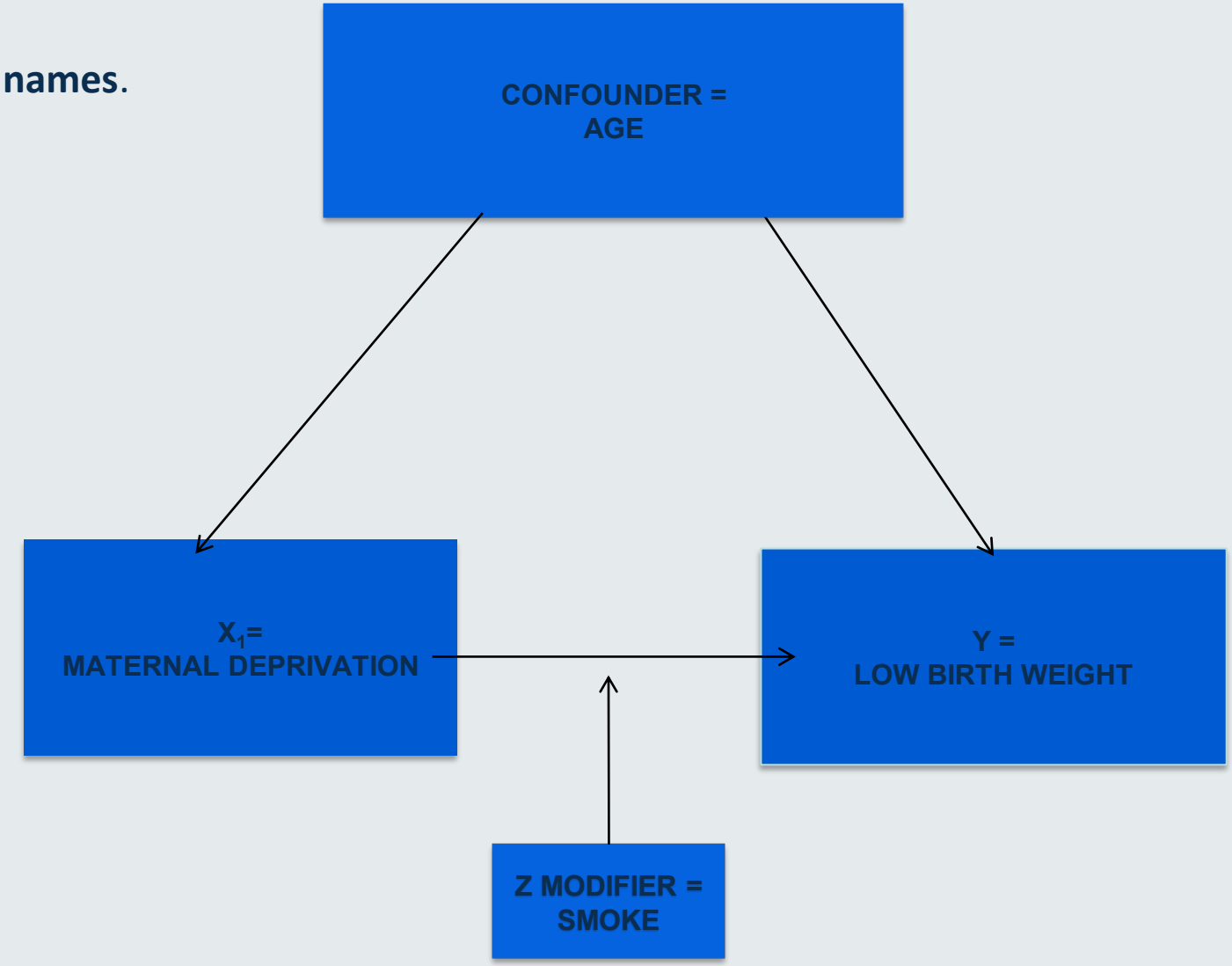
- The effect of treatment*gender term on the outcome is **significantly** different from 0.

- 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**

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: Outliers and Influential Points

Topic title: Effect Modification (Interaction)

Learning Outcomes

After working through this session you should be able to:

- understand what outliers and influential data points are
- understand how to flag outliers and influential data points

Outliers and Influential Points

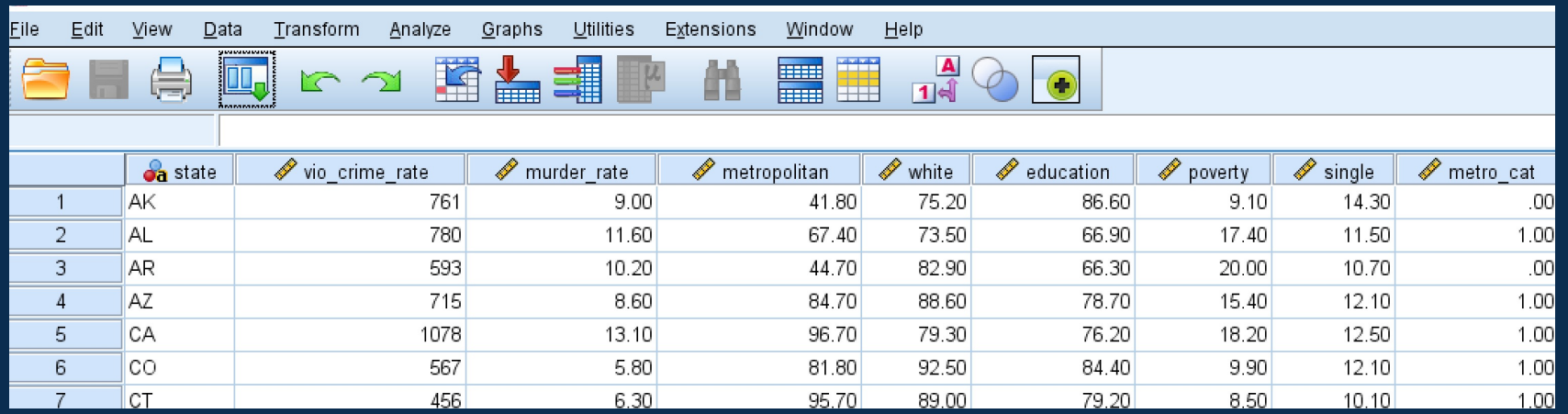
- An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.
- Outliers can be problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.
- Finding outliers depends on subject-area knowledge and an understanding of the data collection process.

Outliers and Influential Points in Regression

- All outliers are not harmful. Some outliers influence the regression model more than the others
- Outliers with large influence on the fitted regression model are called **influential observations**
- Influential observations need special attention as they may distort the actual relationship

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

Outliers

Variable	With Outlier		Without Outlier		Mean Difference	Std Difference
	Mean	Std. Deviation	Mean	Std. Deviation		
violent crime rate (per 100,000 people)	612.84	441.1	566.66	295.9	46.18	145.2
murder rate (per 100,00 people)	8.33	11.0	6.92	4.6	1.40	6.4

- To demonstrate how much a single outlier can affect the results, let's examine the effect of a potential outlier in the lecture_9b_data.sav.
- The table above shows the mean and standard deviation for violent crime and murder rate with and without the potential outlier.

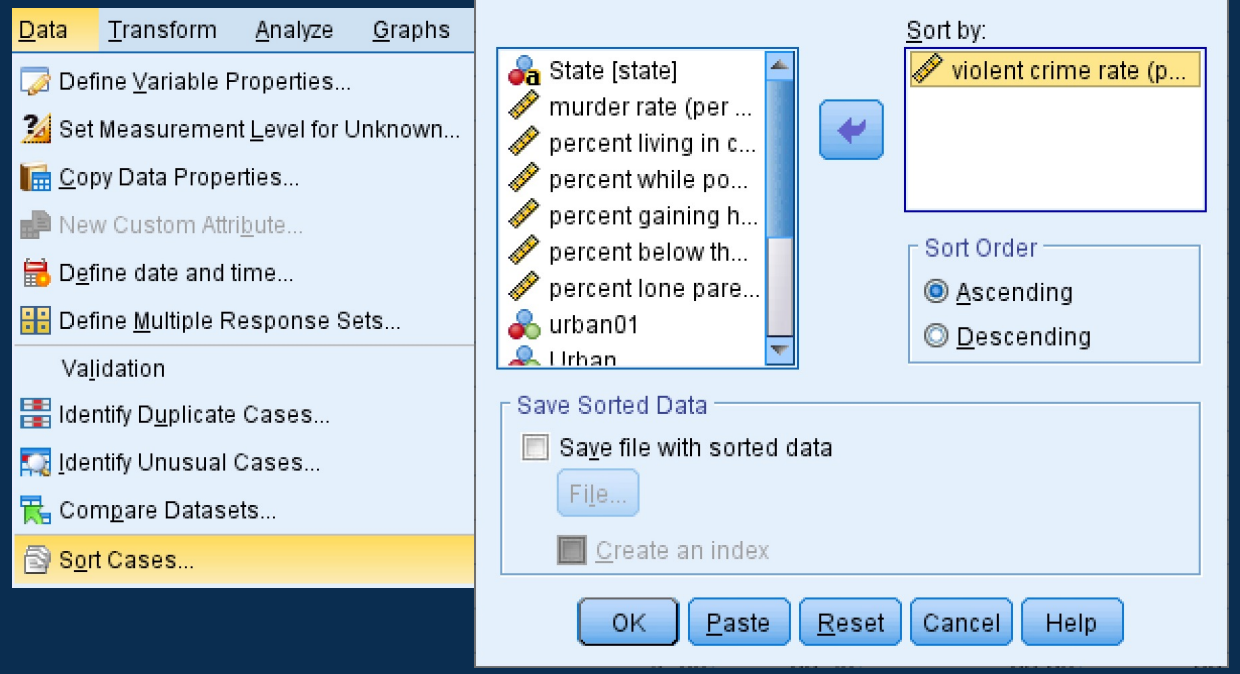
- From the table, it's easy to see how a single outlier can distort the data summaries. A single value changes the mean crime rate by 46.18 (per 100 000) and the standard deviation by a large amount 145.2.

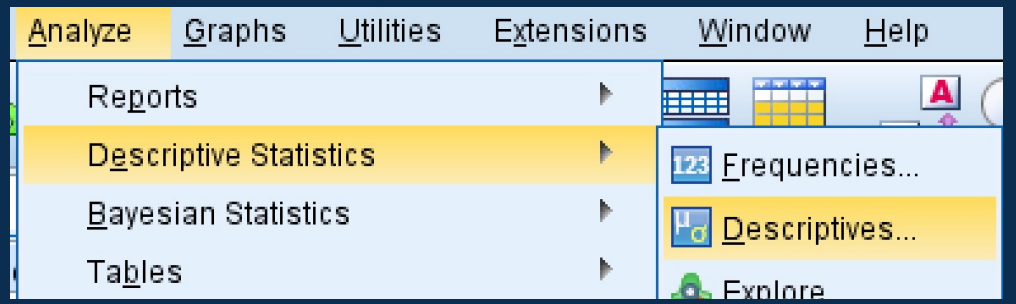
SPSS Slide: Finding Outliers and Influential Points

Sorting Your Datasheet to Find Outliers

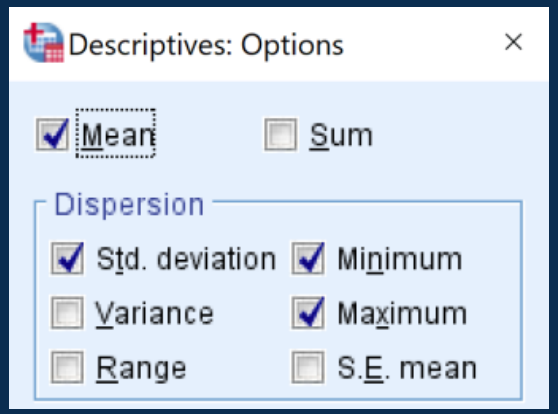
- Sorting your datasheet is a simple but effective way to highlight unusual values. Simply sort your data sheet for each variable and then look for unusually high or low values.
- Alternatively, when asking for “Descriptives” ask for the minimum and maximum to be included in the output

1





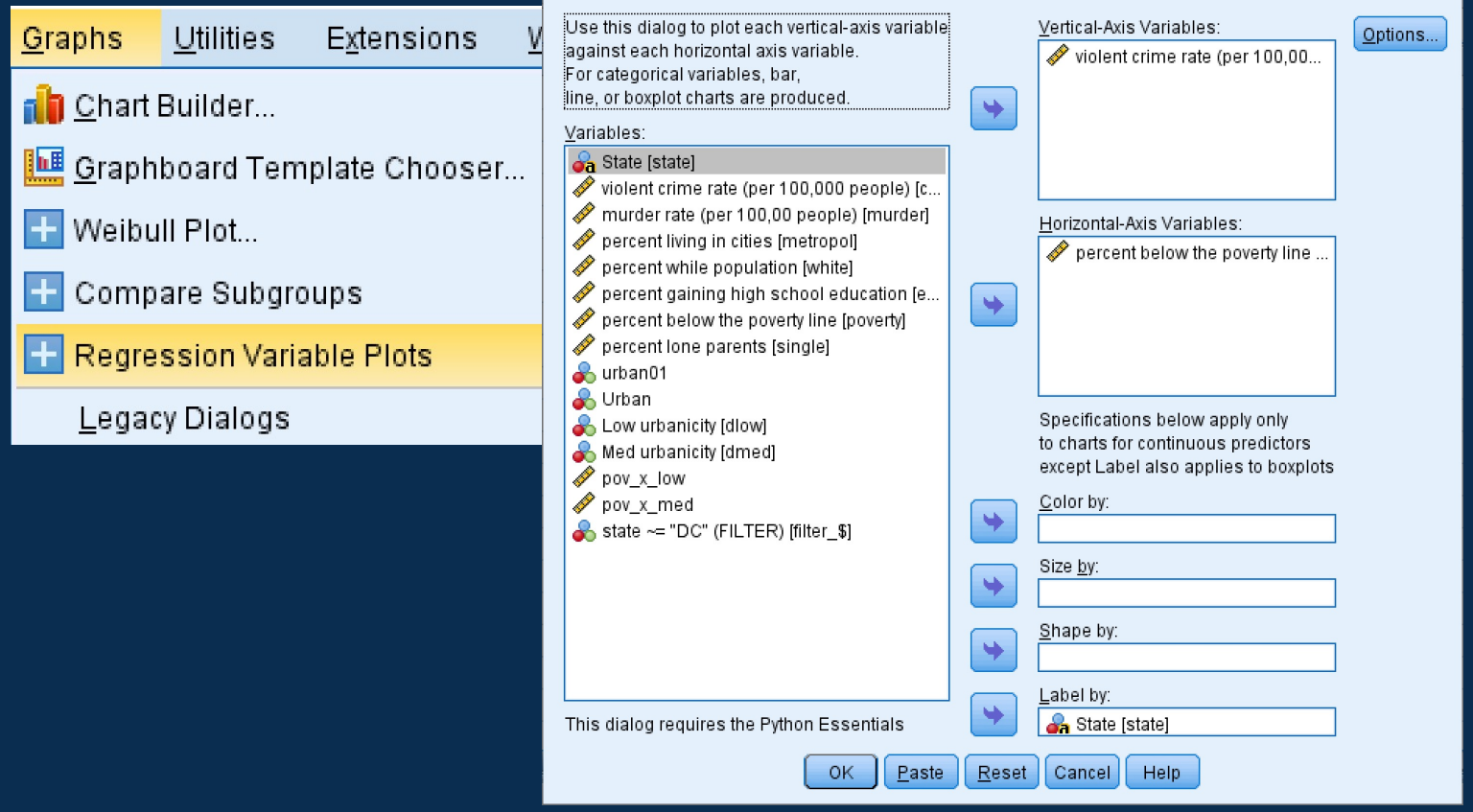
2



SPSS Slide: Finding Outliers and Influential Points

Graphing Your Data to Identify Outliers

- Boxplots, histograms, and scatterplots can highlight outliers



In SPSS you are able to now create a Regression variable plot which shows a scattergraph of two variables and a box plot of their data.

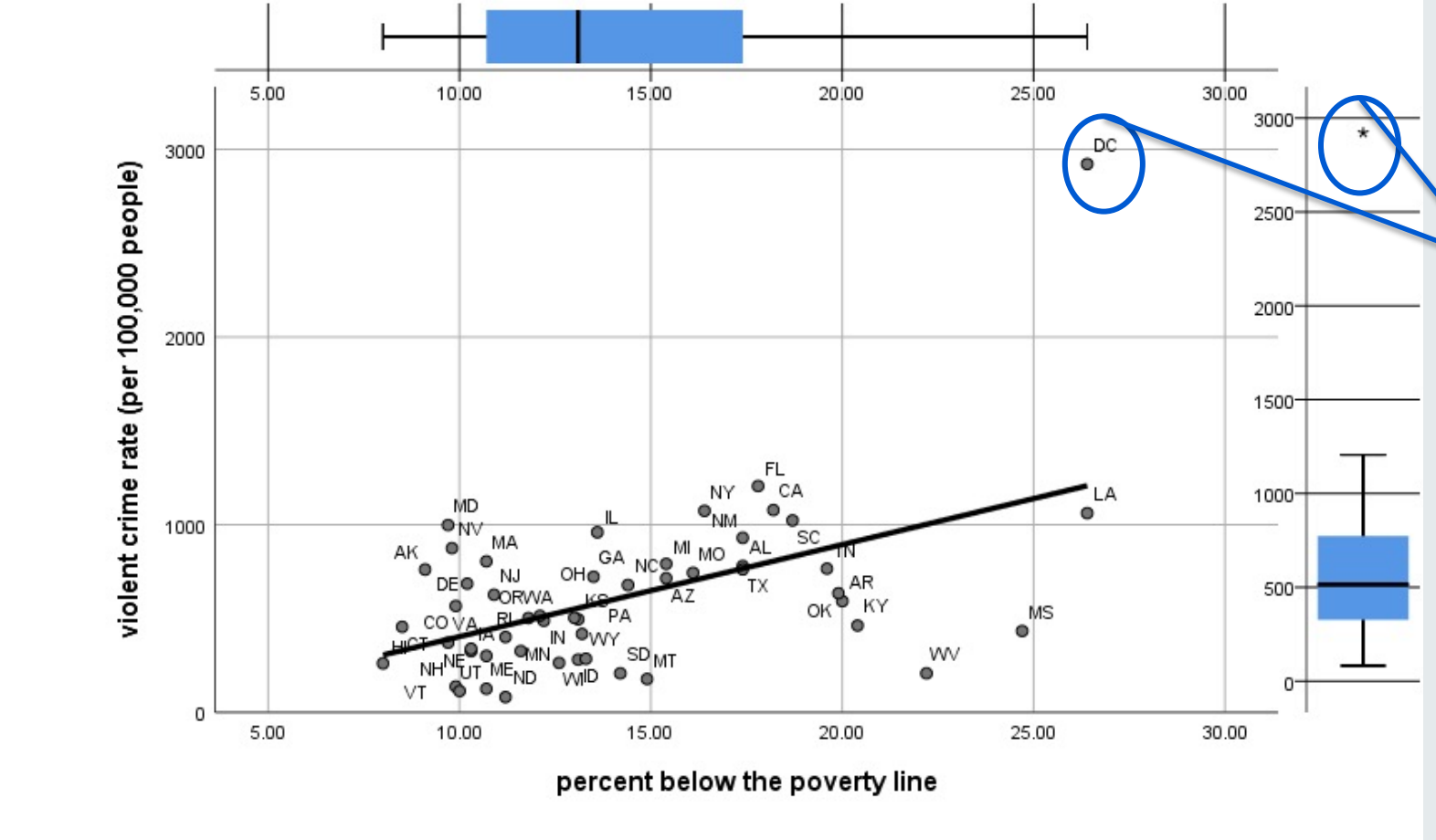
Graphs->Regression Variable Plot - > put dependent variable in the vertical axis, and the independent variable in the horizontal axis

Label by “state” so you can identify any outliers.

Output: Finding Outliers and Influential Points

Graphing Your Data to Identify Outliers

- Boxplots, histograms, and scatterplots can highlight outliers



Potential Outlier

Finding Outliers and Influential Points

Tukey's Method: Using the Interquartile Range

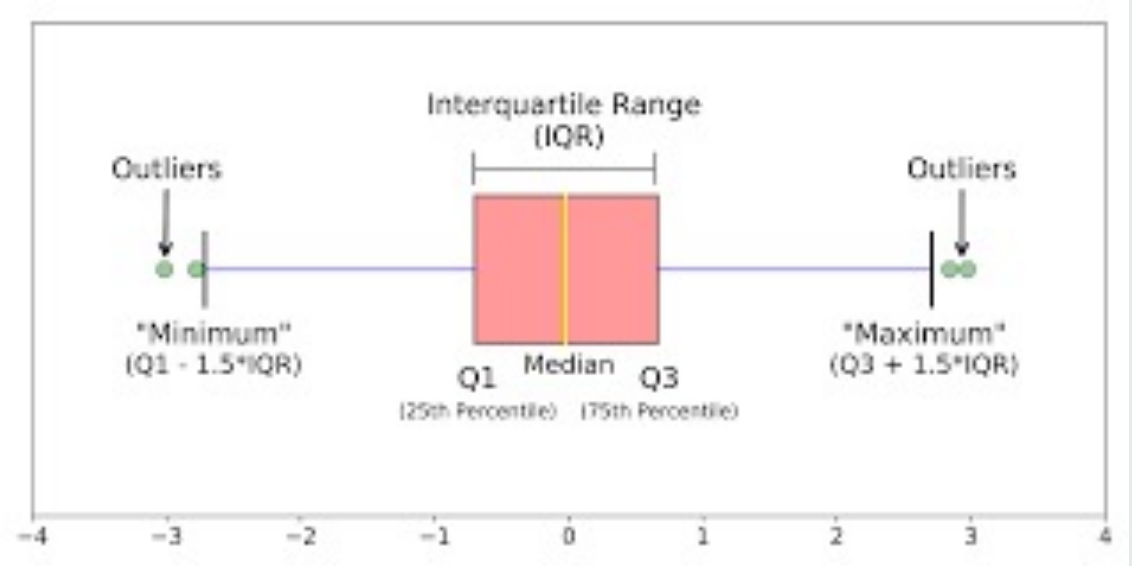
The **IQR** is the middle 50% of the dataset. It's the range of values between the third quartile and the first quartile (Q3 – Q1).

We can take the IQR, Q1, and Q3 values to calculate the following outlier fences for our dataset: lower outer, lower inner, upper inner, and upper outer.

These fences determine whether data points are outliers and whether they are **mild** or **extreme**.

Extreme outliers tend to lie more than **3** times the interquartile range (below the first quartile or above the third quartile), and

Mild outliers lie between **1.5** and three times the interquartile range (below the first quartile or above the third quartile).



Finding Outliers and Influential Points

Example:

Statistics

murder rate (per 100,00 people)

N	Valid	51
	Missing	0
Mean		8.3275
Median		6.6000
Minimum		-9.00
Maximum		78.50
Percentiles	25	3.8000
	50	6.6000
	75	10.3000

Q1 = 3.80
Q3 = 10.30
IQR = Q3 – Q1
IQR = 6.5

Lower Outer = Q1 – 3 x IQR = -15.7
Lower Inner = Q1 – 1.5 x IQR = -5.95
Upper Inner = Q3 + 1.5 x IQR = 20.5
Upper Outer = Q3 + 3 x IQR = 29.8

Order your data:
-9 murder rate for 'IL' is a mild outlier as it lies between the lower inner and outer limits

78.50 murder rate for "DC" is a extreme outlier as it lies outside of the upper outer limit

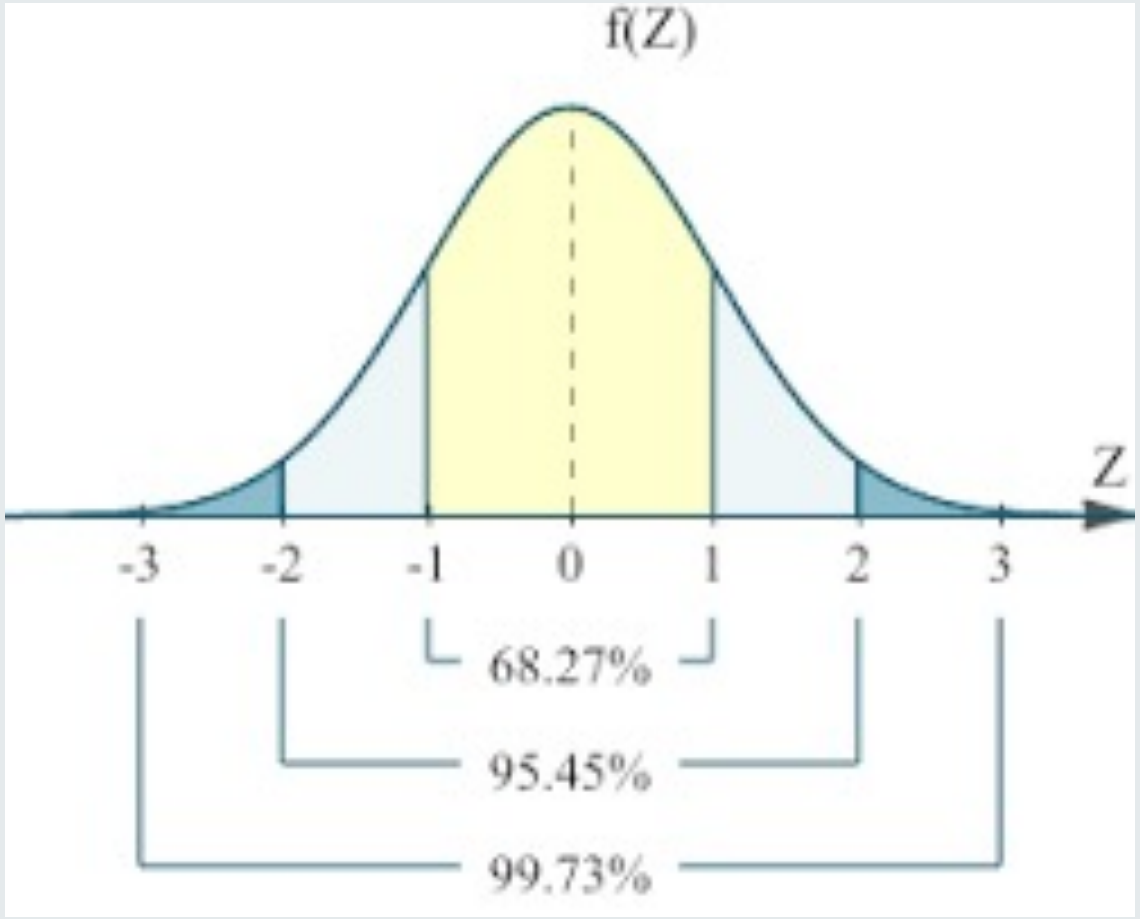
Finding Outliers and Influential Points

Using the Standard Deviation

The **standard deviation (SD)** is a reasonable method to detect outliers when the data distribution is symmetric such as the normal distribution.

68%, 95%, and 99.7% of the data from a normal distribution are within 1, 2, and 3 standard deviations of the mean, respectively.

If data follows a normal distribution, this helps to estimate the likelihood of having extreme values in the data , so that the observation **two or three standard deviations** away from the mean may be considered as an outlier in the data.



Outliers and Influential Observations

Using Standardised Residuals

The good thing about standardized residuals is that they quantify how large the residuals are in standard deviation units, and therefore can be easily used to identify outliers: An observation with an **Absolute standardized residual** that is **larger than 3** (in absolute value) is deemed by some to be an outlier.

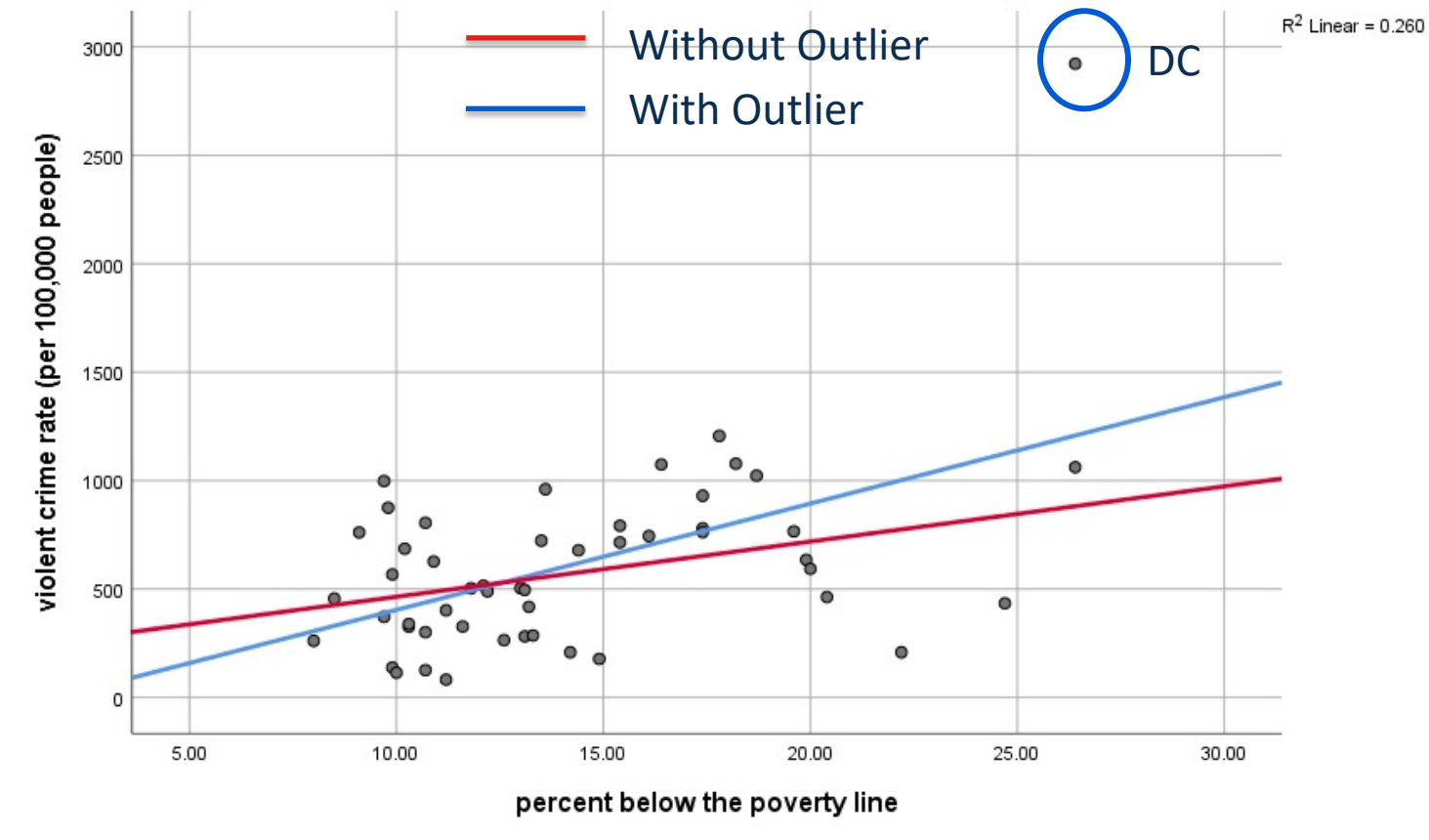
Outliers and Influential Observations

DFBETA and DFFIT

- **DFBETA** and **DFFIT** are two diagnostic measures for flagging influential observations
- For a given observation, **DFBETA** measures the **change in the estimated coefficient β_j** due to deleting that observation
  - Standardised **DFBETA** is defined as **DFBETA divided by the SE (est β_j)**for the adjusted dataset
- For a given observation, **DFFIT** measures the **change in the predicted value ($\hat{y}$)** due to deleting that observation
  - Standardised **DFFIT** is defined as **DFFIT divided by SE($\hat{y}$)** for the adjusted data
- A general guideline:
  - Absolute **standardised DFBETA > 1** suggests **influential** observations
  - Absolute **standardised DFFIT > 1** suggests **influential** observations

Influential Observations

Consider the following Scatterplot from Lecture_9b_data.sav showing the US crime rate. The figure shows that the state DC is an outlier. Crime rate is very high in DC compared to the other states

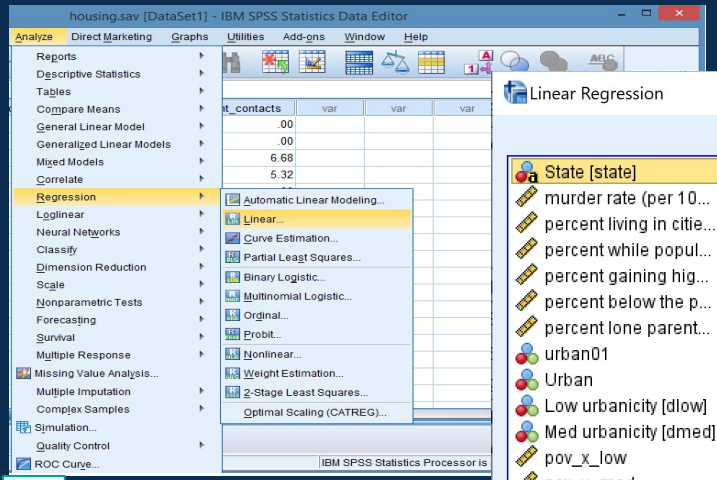


The slope estimated by including the outlier is much higher than the slope estimated with the outlier removed.

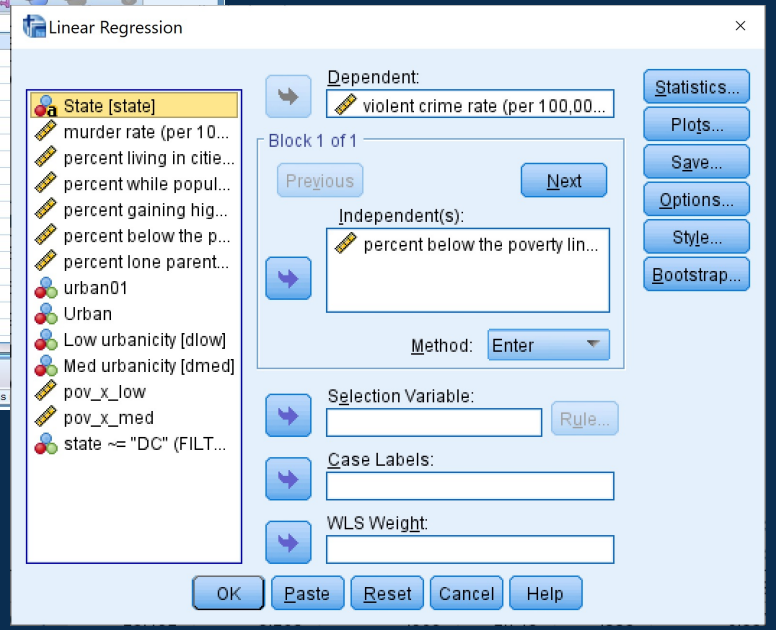
The slope difference is the **influence** of the outlier state DC

SPSS Slide: ‘How to’ Steps

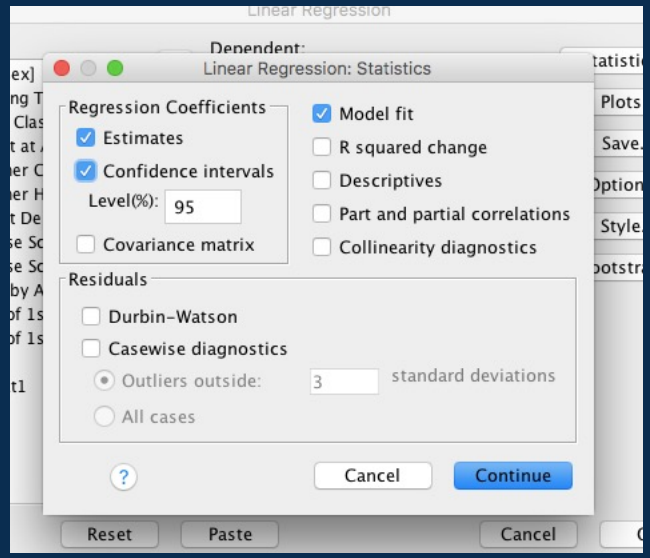
- Researchers believe that the state of DC is giving a distorted understanding of the Crime – poverty relationship. They have decided to run an analysis including this potential outlier and without it to check the level of influence
- 1) Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’
- 2) In dependent put ‘crime’ and in independent put ‘poverty’,



1

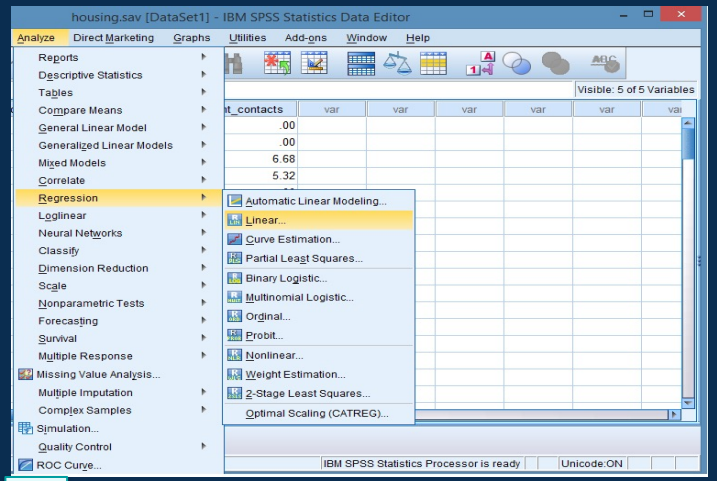


2

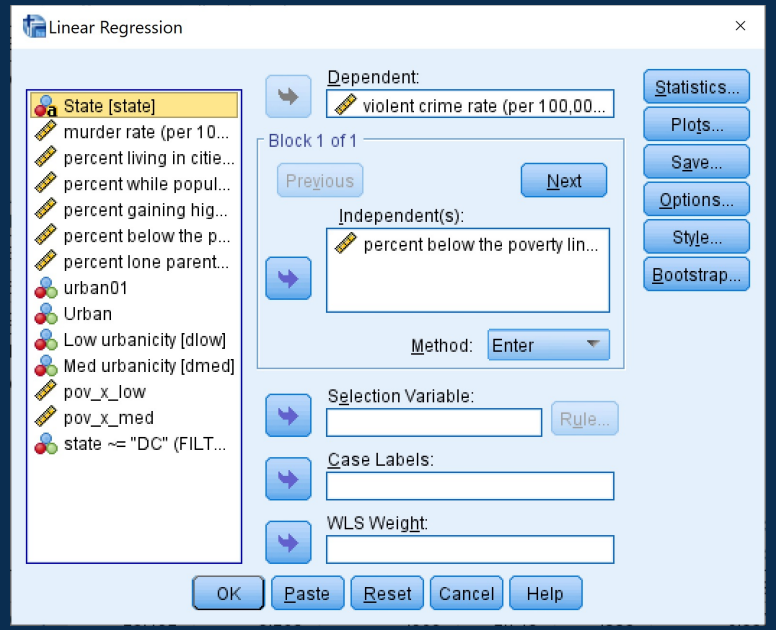


SPSS Slide: ‘How to’ Steps

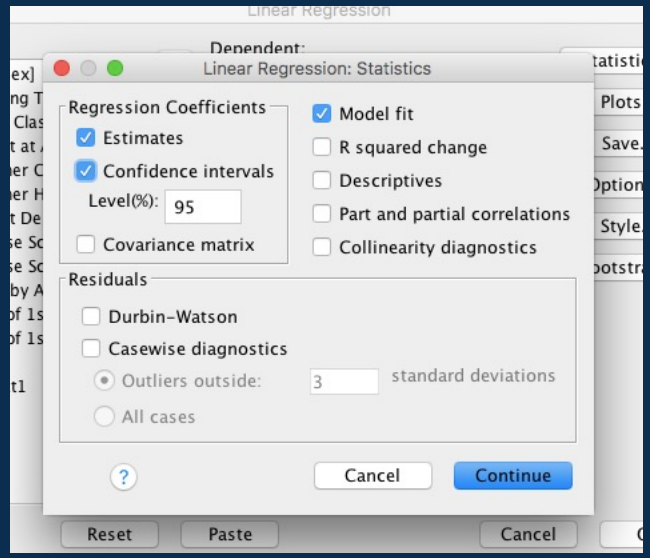
- Researchers believe that the state of DC is giving a distorted understanding of the Crime – poverty relationship. They have decided to run an analysis including this potential outlier and without it to check the level of influence. Use **‘Select Cases’** option under the **‘Data’** Menu to remove the outlier from the analysis. Re-run the regression
- 1) Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’
- 2) In dependent put ‘crime’ and in independent put ‘poverty’,



1



2



Output and Interpretation

The first table were generated for data in all US states, whilst the second table was generated excluding DC state.

All data

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)	-86.201	176.990		-.487	.628	-441.876	269.474
	percent below the poverty line	49.025	11.828	.510	4.145	.000	25.256	72.794

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

Excluding state DC

Coefficients[a]

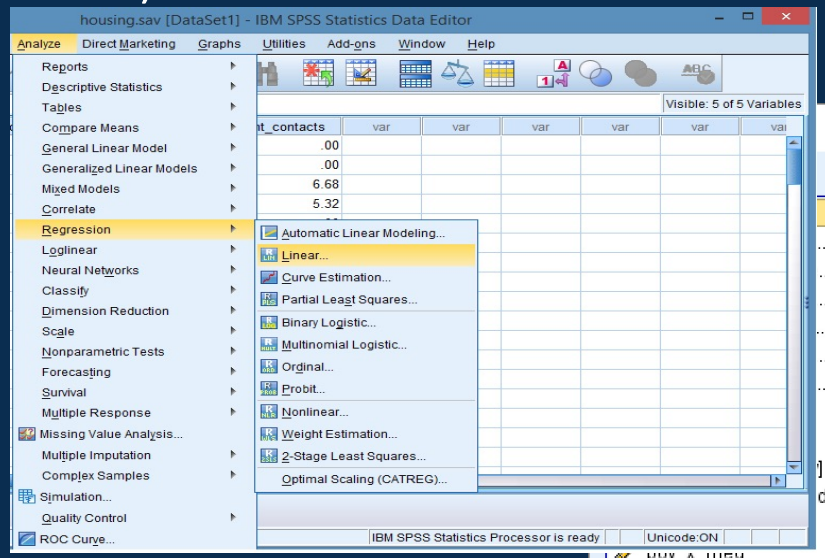
Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	209.920	135.613		1.548	.128	-62.748	482.588
	percent below the poverty line	25.452	9.260	.369	2.749	.008	6.833	44.072

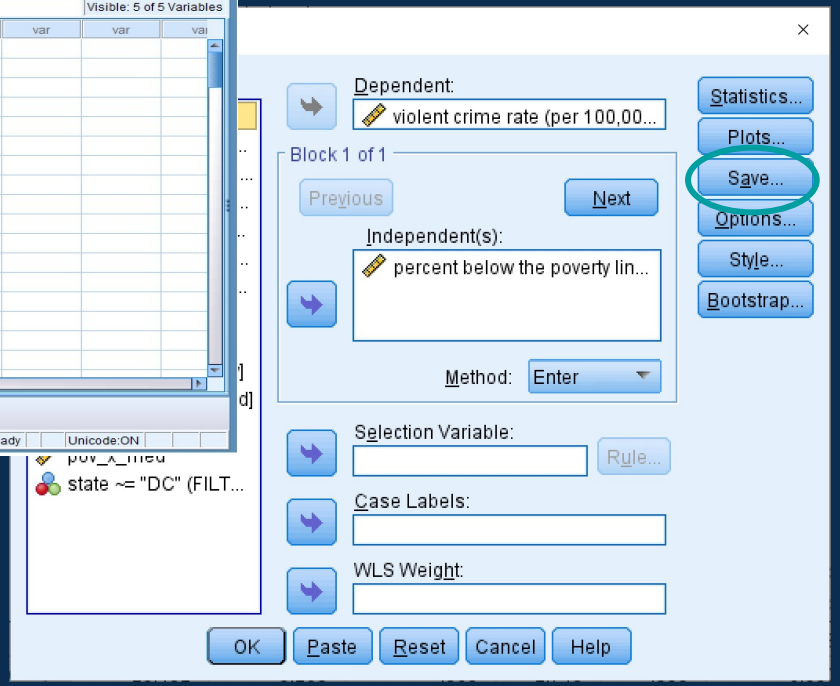
a. Dependent Variable: violent crime rate (per 100,000 people)

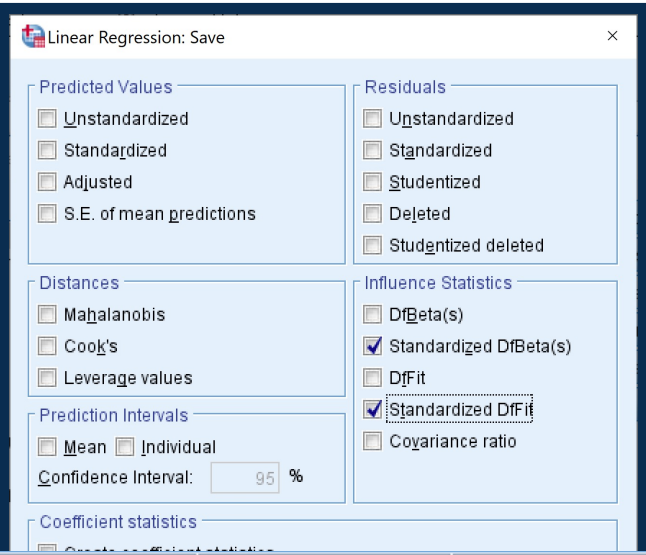
- DFBETA for the poverty variable can be calculated as the difference of the coefficient from the full model to the adjusted models (49.025 – 25.452 = 23.573) standardized by accounting for the full model and adjusted model error and covariance (not covered in this course).
- To estimate the standardized DFBETA and standardized DFFIT, we select this option from SPSS as it is shown in the next slide
- As per SPSS, standardized DFBETA for the coefficient for poverty is = 2.75

SPSS Slide: ‘How to’ Steps

- Dfbeta and Dffit in SPSS
- 1) Use ‘Analyse’ -> ‘Regression’ -> ‘Linear’
- 2) In **dependent** put **‘crime’** and in **independent** put **‘poverty’,**
- 3) Click on **‘Save’**







SDF_1	SDB0_1	SDB1_1
.14719	-.01946	.06280
.33367	.29672	-.23645
.20117	-.06980	.12374
-.02732	-.02142	.01527
.07732	.07238	-.06072
.17561	.14517	-.10835
.13554	-.05363	.08886
.08312	.06738	-.04944
2.93579	-2.30848	2.74990

1

2

3

4

Why Should Outliers and Influential Points be Considered?

- Outliers generally serve to increase error variance and reduce the power of statistical tests.

- If non-randomly distributed, they can decrease normality (and in multivariate analyses, violate assumptions of sphericity and multivariate normality), altering the odds of making both Type I and Type II errors.

- They can seriously bias or influence estimates that may be of substantive interest

What Should I do with Outliers?

When considering whether to remove an outlier, you'll need to evaluate

- if it appropriately reflects your target population, subject-area, research question, and research methodology.
- Did anything unusual happen while measuring these observations, such as power failures, abnormal experimental conditions, or anything else out of the norm?
- Is there anything substantially different about an observation, whether it's a person, item, or transaction?
- Did measurement or data entry errors occur?

What Should I do with Outliers?

If the outlier in question is:

- A measurement error or data entry error, correct the error if possible. If you can't fix it, remove that observation because you know it's incorrect.

- Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.

- A natural part of the population you are studying, you should **not** remove it.

What Should I do with Outliers?

When you decide to remove outliers

- document the excluded data points and explain your reasoning.
- You must be able to attribute a specific cause for removing outliers.
- Another approach is to perform the analysis with and without these observations and discuss the differences.
  - Comparing results in this manner is particularly useful when you're unsure about removing an outlier and when there is substantial disagreement within a group over this question.

Knowledge Check

Q1: In the Metropol Data when ordered the researcher sees that the state of “MS” has a percentage living in cities as -30.7 and the states of “NJ” and “DC” has a percentage of 100. The researcher wants to identify if any of these points are outliers in the data and whether they are mild of extreme outliers. Using the summary below determine if the researcher is correct.

Statistics

percent living in cities

N	Valid	51
	Missing	0
Mean		66.1863
Median		69.8000
Mode		41.80[a]
Std. Deviation		25.41943
Minimum		-30.70
Maximum		100.00
Percentiles	25	48.5000
	50	69.8000
	75	84.0000

a. Multiple modes exist. The smallest value is shown

Knowledge Check Solutions

Q1: In the Metropol Data when ordered the researcher sees that the state of "MS" has a percentage living in cities as -30.7 and the states of "NJ" and "DC" has a percentage of 100. The researcher wants to identify if any of these points are outliers in the data and whether they are mild of extreme outliers. Using the summary below determine if the researcher is correct.

Statistics

percent living in cities

N	Valid	51
	Missing	0
Mean		66.1863
Median		69.8000
Mode		41.80[a]
Std. Deviation		25.41943
Minimum		-30.70
Maximum		100.00
Percentiles	25	48.5000
	50	69.8000
	75	84.0000

a. Multiple modes exist. The smallest value is shown

Q1 = 48.5
Q3 = 84
IQR = Q3 – Q1
IQR = 35.5

Lower Outer	= Q1 – 3 x IQR	= -58
Lower Inner	= Q1 – 1.5 x IQR	= -4.75
Upper Inner	= Q3 + 1.5 x IQR	= 137.25
Upper Outer	= Q3 + 3 x IQR	= 190.5

-30.7 percentage living in cities for 'MS' is a mild outlier as it lies between the lower inner and outer limits

References

Grubbs, F. E. (February 1969). "Procedures for detecting outlying observations in samples". Technometrics. 11 (1): 1–21. doi:10.1080/00401706.1969.10490657

Tukey, John W (1977). Exploratory Data Analysis. Addison-Wesley. ISBN 978-0-201-07616-5.

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**