



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: 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

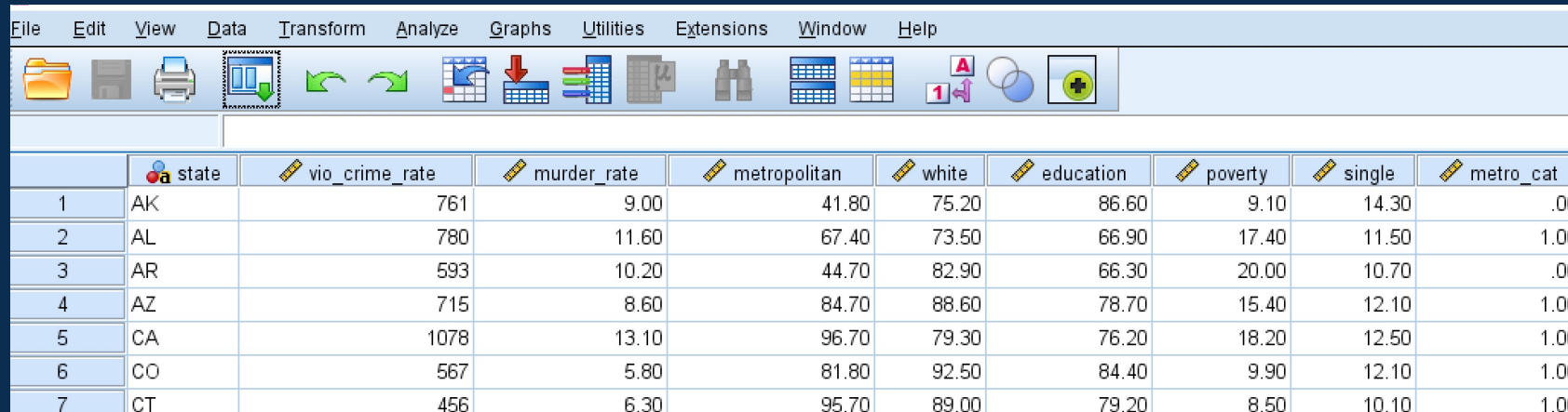
- An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.
- Outliers can be problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.
- Finding outliers depends on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



The screenshot shows the SPSS Data Editor window. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The toolbar contains icons for file operations, data manipulation, and analysis. The data grid displays the following variables and values:

	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

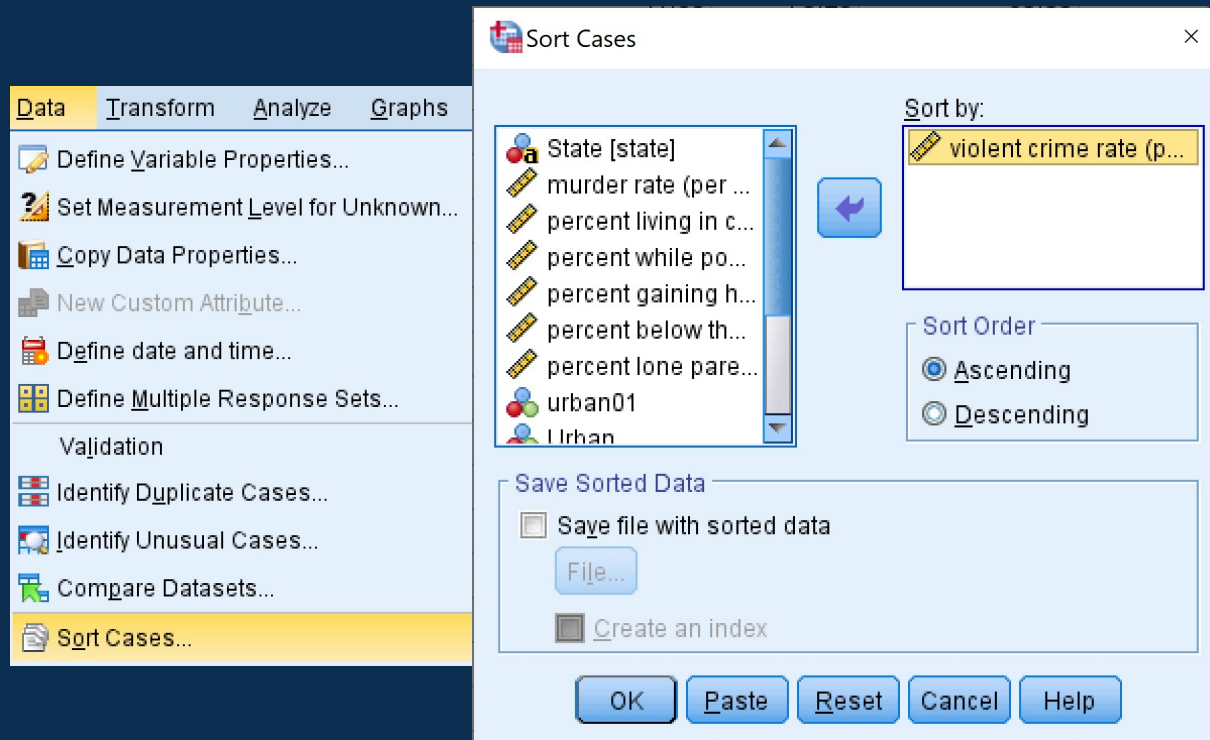
- To demonstrate how much a single outlier can affect the results, let's examine the effect of a potential outlier in the lecture_9b_data.sav.
- The table above shows the mean and standard deviation for violent crime and murder rate with and without the potential outlier.
- From the table, it's easy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

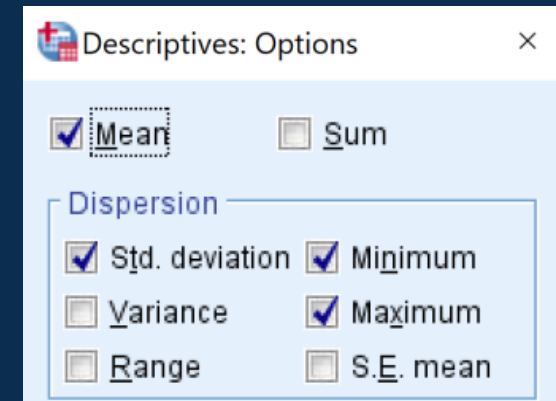
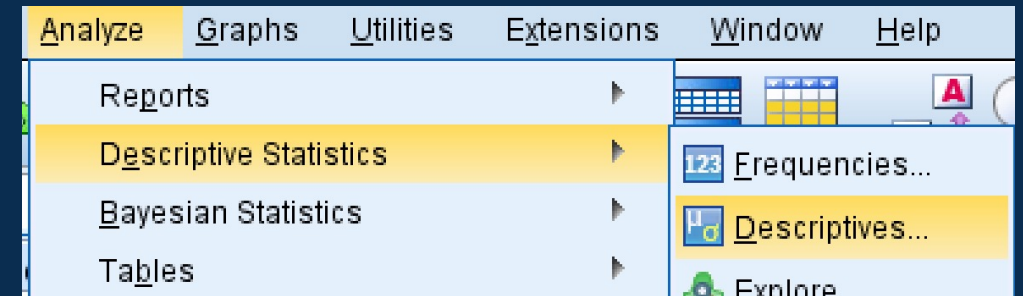
Sorting Your Datasheet to Find Outliers

- 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



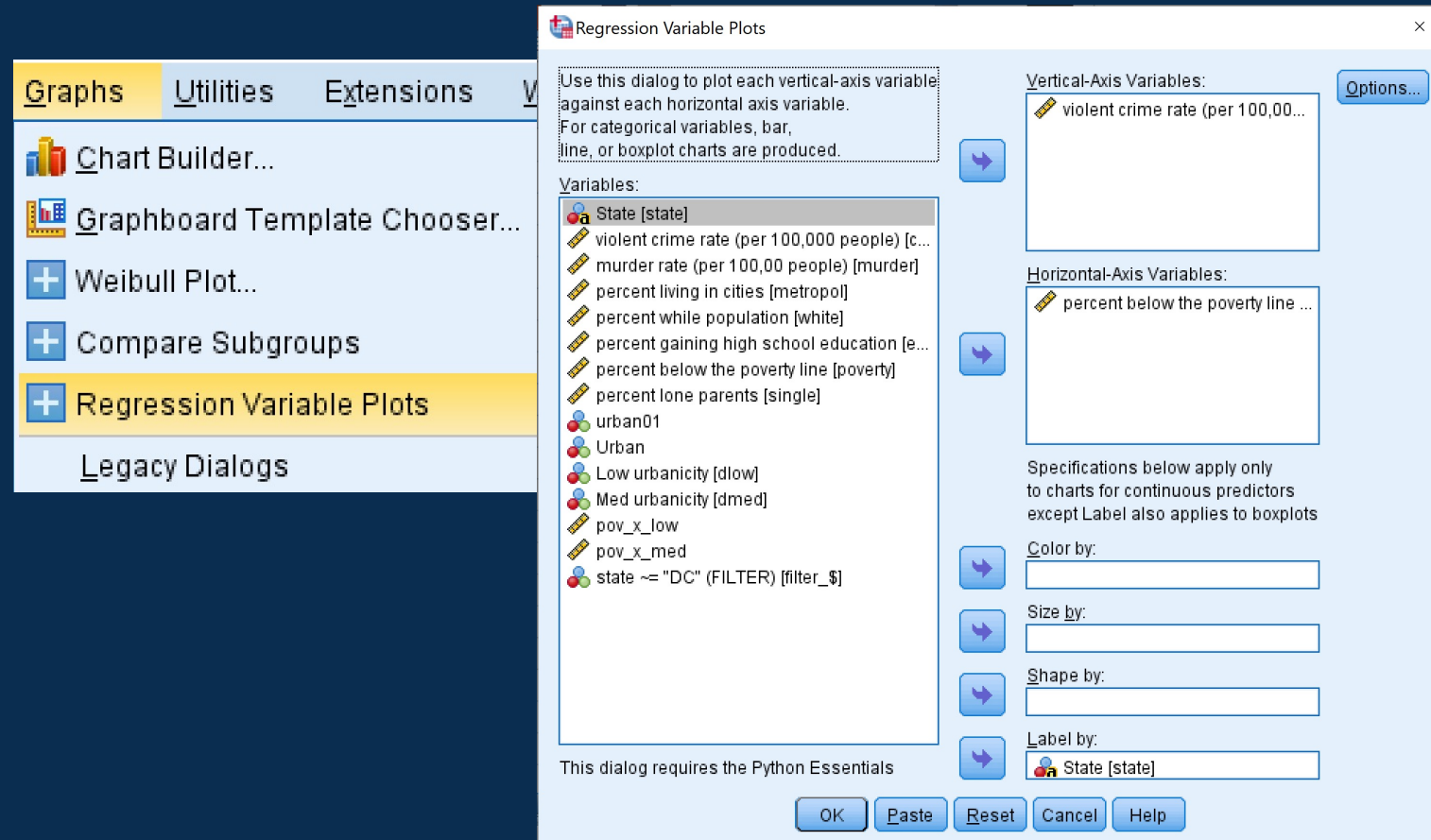
2



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.

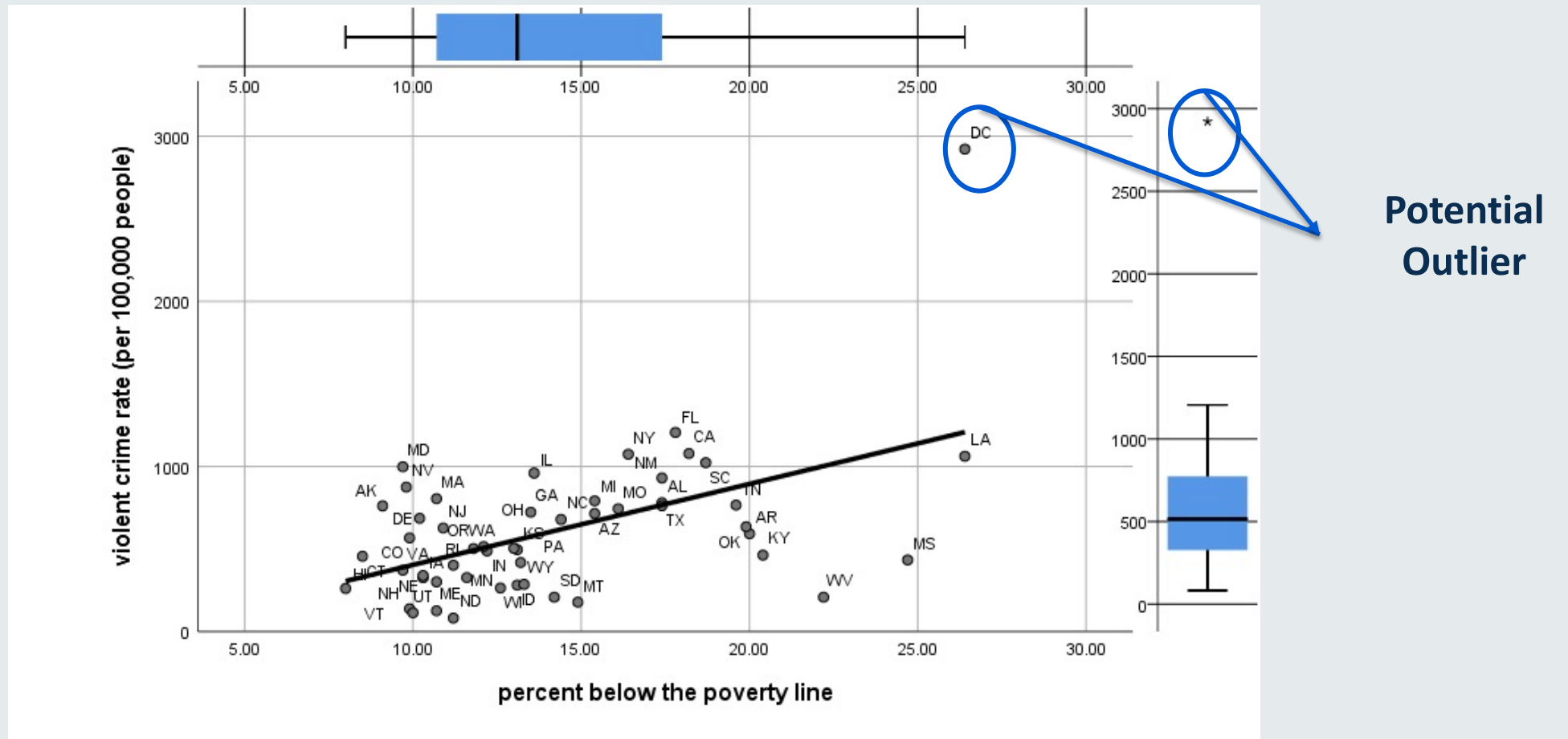
Graphs->Regression Variable Plot - > put dependent variable in the vertical axis, and the independent variable in the horizontal axis

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



Finding Outliers and Influential Points

Tukey's Method: Using the Interquartile Range

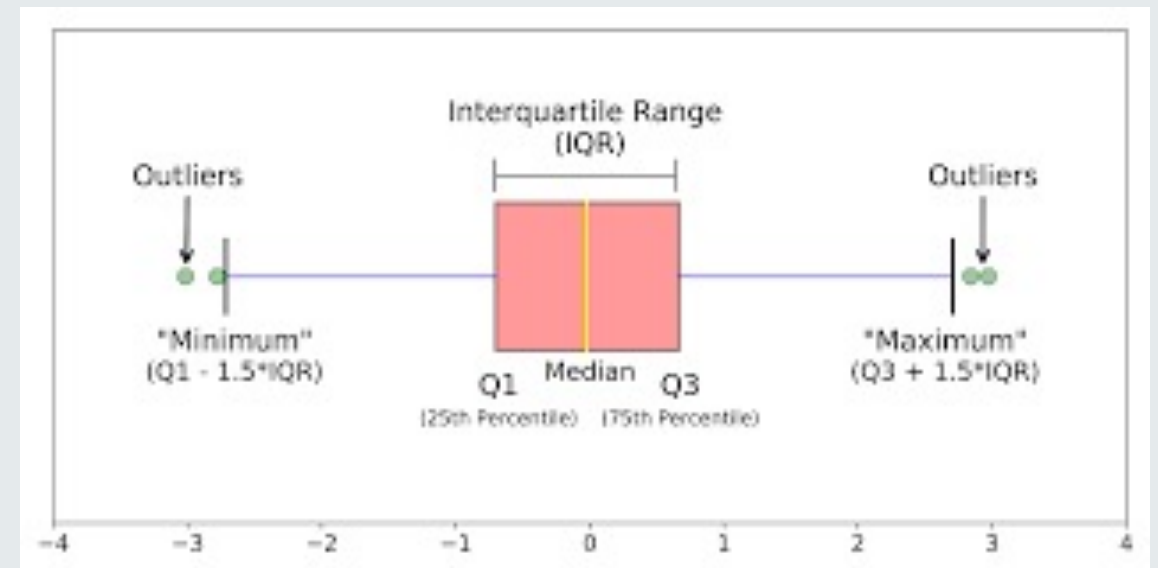
The **IQR** is the middle 50% of the dataset. It's the range of values between the third quartile and the first quartile ($Q3 - Q1$).

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

$$Q1 = 3.80$$

$$Q3 = 10.30$$

$$IQR = Q3 - Q1$$

$$IQR = 6.5$$

$$\text{Lower Outer} = Q1 - 3 \times IQR = -15.7$$

$$\text{Lower Inner} = Q1 - 1.5 \times IQR = -5.95$$

$$\text{Upper Inner} = Q3 + 1.5 \times IQR = 20.5$$

$$\text{Upper Outer} = Q3 + 3 \times IQR = 29.8$$

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

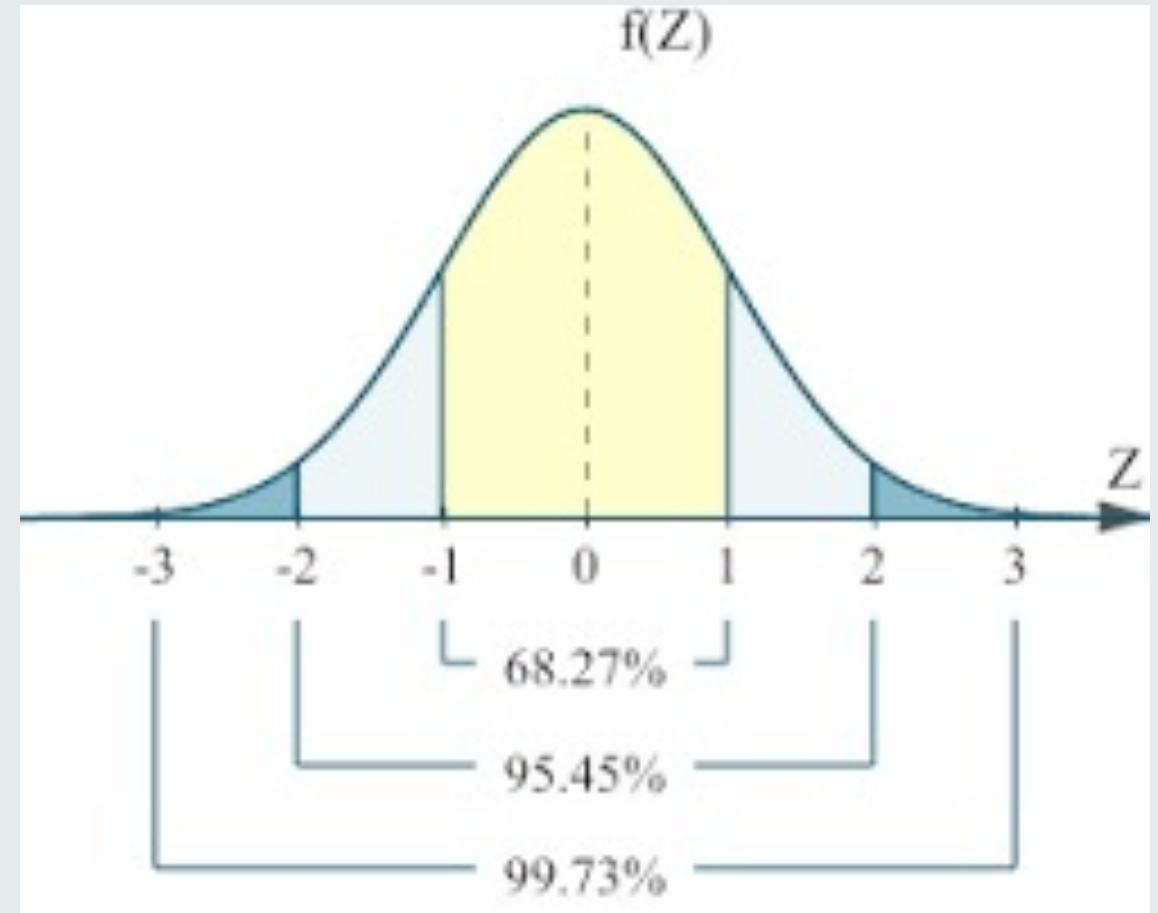
Finding Outliers and Influential Points

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.



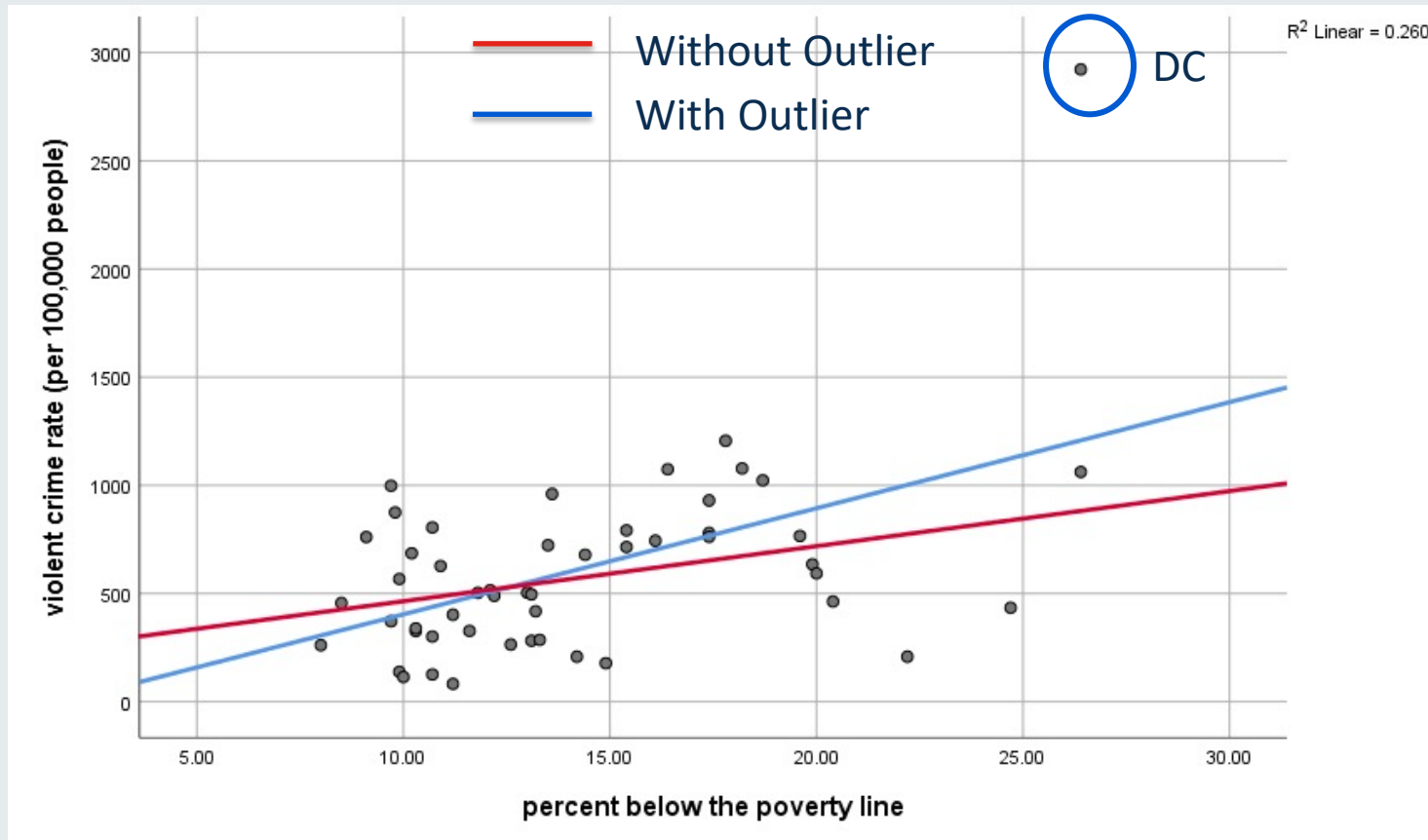
Outliers and Influential Observations

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',

The image displays two screenshots from the IBM SPSS Statistics software interface. The left screenshot shows the 'Linear Regression' dialog box with 'State [state]' selected as the dependent variable and 'percent below the poverty line' as the independent variable. The right screenshot shows the 'Linear Regression: Statistics' sub-dialog box with 'Estimates', 'Confidence intervals', and 'Model fit' selected.

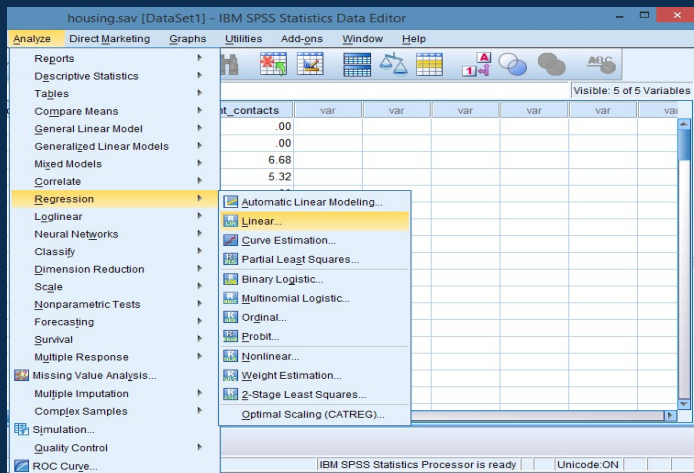
1

2

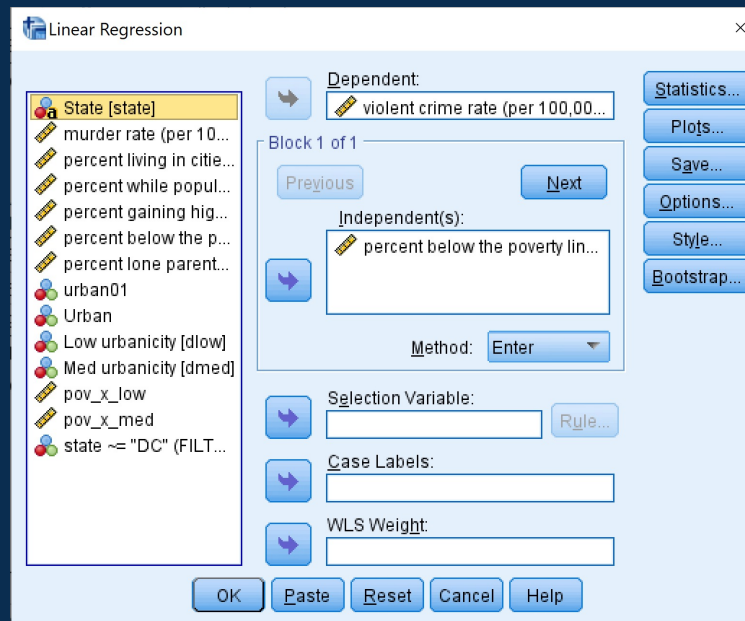


SPSS Slide: 'How to' Steps

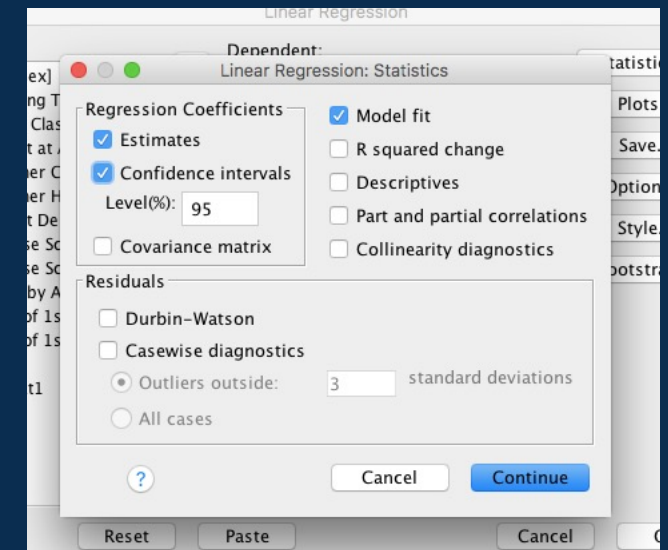
- Researchers believe that the state of DC is giving a distorted understanding of the Crime – poverty relationship. They have decided to run an analysis including this potential outlier and without it to check the level of influence. Use **'Select Cases'** option under the **'Data'** Menu to remove the 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								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		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

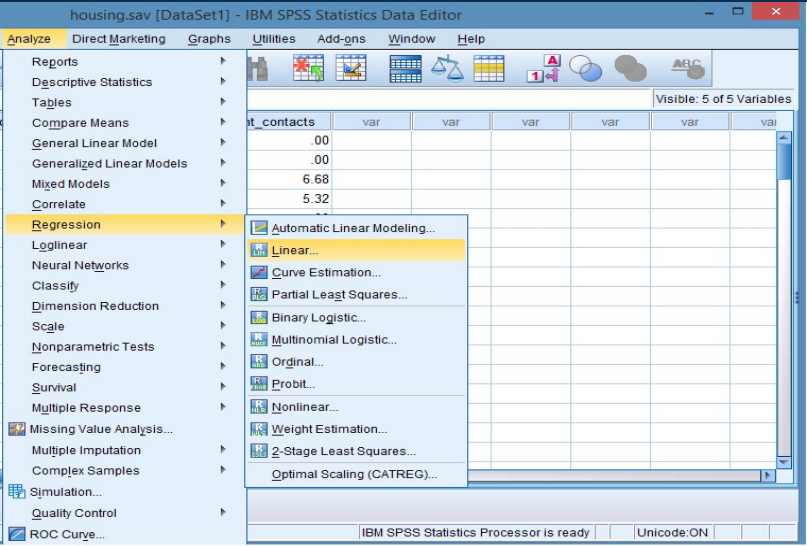
Coefficients ^a								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	209.920	135.613		1.548	.128	-62.748	482.588
	percent below the poverty line	25.452	9.260	.369	2.749	.008	6.833	44.072

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

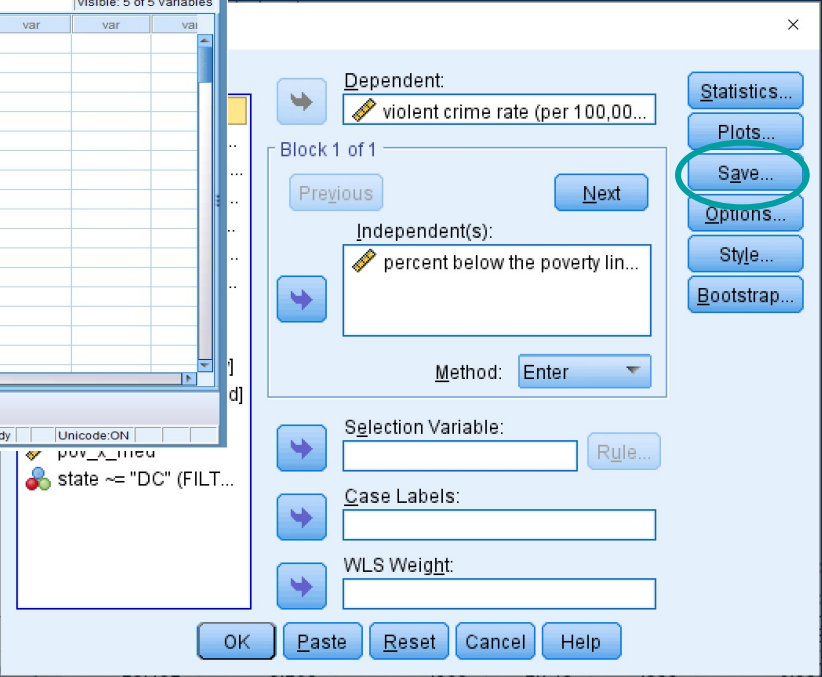
- ➡ DFBETA for the poverty variable can be calculated as the difference of the coefficient from the full model to the adjusted models ($49.025 - 25.452 = 23.573$) standardized by accounting for the full model and adjusted model error and covariance (not covered in this course).
- ➡ To estimate the standardized 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

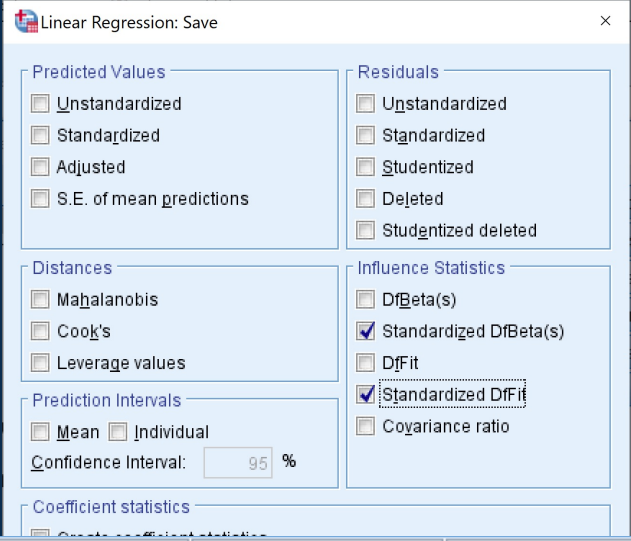
- Dfbeta and Dffit in SPSS
- 1) Use 'Analyse' -> 'Regression' -> 'Linear'
- 2) In **dependent** put 'crime' and in **independent** put 'poverty',
- 3) Click on 'Save'



1



2



3

SDF_1	SDB0_1	SDB1_1
.22778	.22784	.71003
.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

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 or 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 or 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

$$Q1 = 48.5$$

$$Q3 = 84$$

$$IQR = Q3 - Q1$$

$$IQR = 35.5$$

$$\text{Lower Outer} = Q1 - 3 \times IQR = -58$$

$$\text{Lower Inner} = Q1 - 1.5 \times IQR = -4.75$$

$$\text{Upper Inner} = Q3 + 1.5 \times IQR = 137.25$$

$$\text{Upper Outer} = Q3 + 3 \times IQR = 190.5$$

-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