Introduction to SPSS and data analysis

Loes Hollestein, PhD and Marlies Wakkee, PhD
EADV Fostering Course 'Clinical Research and Epidemiology', 13-17 June 2016
Erasmus MC University Medical Center, Rotterdam, The Netherlands

Burg. ‘s Jacobplein 51
3015 CA Rotterdam, The Netherlands

Disclaimer
Care has been taken to confirm the accuracy of the information presented and to describe generally accepted practices. However, authors, editor and publishers are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, expressed or implied, with respects to the currency, completeness or accuracy of the contents of the publication. Application of this information in a particular situation remains the professional responsibility of the reader.
The authors have made every effort to trace copyright holders for borrowed material. If they have inadvertently overlooked any, they will be pleased to make the necessary arrangements at the first opportunity.
This book is only intended for personal use. Do not distribute the materials without written permission of the authors. For commercial interest, please contact the Erasmus MC University Medical Center, department of Dermatology: contact person: L.M. Hollestein (l.hollestein@erasmusmc.nl).

Table of Contents

Note: chapters or paragraphs with an asterisk (*) indicate advanced topics in data-analysis.

Introduction

This is an introduction into SPSS and data analyses. You will learn some basic functions for data exploration, data editing and simple inferential statistics. Thereafter an introduction to regression analyses and survival analyses will be presented. Each chapter starts with a combination between theoretical explanation and practical instructions, which you can perform simultaneously. We suggest to read this carefully and perform the accompanied syntax examples. At the end of each chapter you will find assignments on each topic, which you should be able to make with the information from the chapters. Syntax files with the correct solutions will be provided and don't hesitate to ask for help from the tutors.

Italic text indicate where you can find the function in the SPSS menu, e.g.:
Analyze > Descriptive Statistics > Descriptives

Capitals indicate SPSS syntax statements, e.g:
RENAME

In boxes you can find background information on statistical formulas, e.g.:

$$\boxed{\ln(odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k}$$

☹ If you dislike statistical formulas you can skip the information in the box, because it will be described without formulas in the text as well.

Chapters or paragraphs with an asterisk (*) indicate advanced topics in data-analysis.

Loes Hollestein
Marlies Wakkee

Schedule

Day	**Beginners**		**Advanced**	
	Chapters	**Assignment (if time left)**	**Chapters**	**Assignment**
Monday	Chapter 1	1-1	Chapter 2 Chapter 3	1-1 2-1
Tuesday	Chapter 2 Chapter 3.1 - 3.4	2-1 3-1 (questions 1-3)	Chapter 3 Chapter 4	3-1, 3-2 4-1, 4-2
Thursday	Chapter 4.1 - 4.3 Chapter 5	4-1	Chapter 5 Chapter 6	5-1

Chapter 1 Exploring SPSS

Dataset: SCC
Patients with a primary cutaneous invasive squamous cell carcinoma were registered. In this database you will find information about the patients age, sex, year of diagnosis and tumor characteristics, such as stage according to the tumor node metastasis (TNM), body site of the tumor according to the International Classification of Disease for Oncology (ICD-O). All SCC (both first and subsequent SCC) are included for the dataset.

SPSS Screens
SPSS has different screens in which you can work:
Variable view and data view of the data editor, the Syntax Editor and the Output.

Screen	Function	Files
Data editor	Data management and starting statistical procedures Variable view & Data view	.sav
Output	Representation of outcomes of statistical procedures (tables and figures)	.spo
Syntax Editor	Save and edit SPSS syntax or execute SPSS procedures by ‘running’ the syntax	.sps

Copy the data
Make a copy of the dataset before you start working with your data in SPSS. Save the raw datafile and perform the data exploration, editing and analyses in a copy of the original datafile.

1.1 Open a dataset

1a. Import the data from excel

File > Open > Data

Select Excel (.xls) from the picklist, Select SCC and click on open.
Select the worksheet ‘SCC_2011’ from the picklist. The first row of the excel worksheet contains the variable names, which are copied automatically by SPSS. If correctly opened, close the dataset again.

1b. Open a SPSS datafile ‘SCC’

In this file are variable labels and values are already assigned.

File > Open > Data

1.2 Syntax

During this practical you will also make a syntax. It is important to save the syntax, so that you can always look up which analyses you performed in the past. To save the syntax in the syntax editor you should click on ‘Paste’ instead of ‘OK’ or copy the syntax from the output screen (Click on Copy with the right mouse button, Ctrl+C doesn’t work).
Between an asterisk (*) and a dot (.) you can type notes.
To run the syntax: select part of the syntax and click on the ‘Run selection’ button.

▶

2a. Open a new syntax file

File > New > Syntax

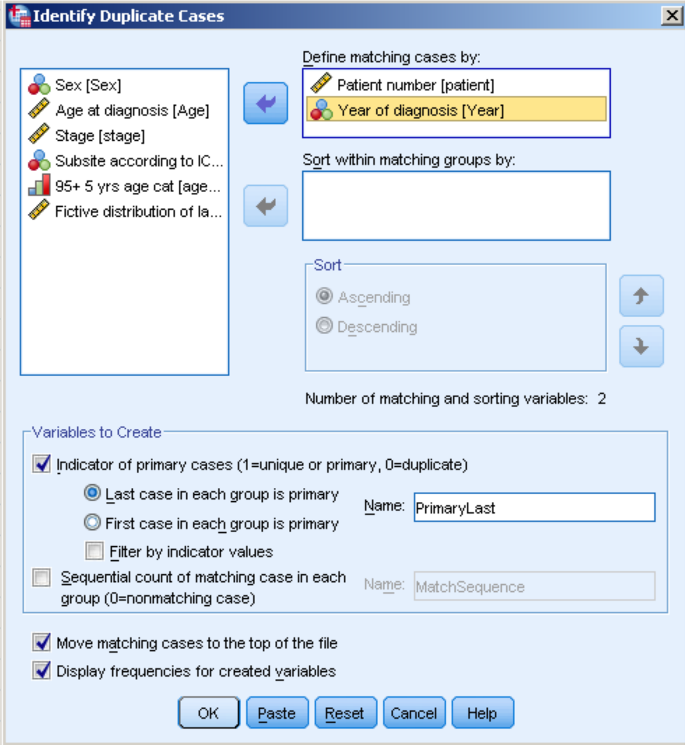
1.3 Unique patient identifier

Mistakes can happen with the construction of a database. It is important to assign a unique number to each patient. In this way you can check if patients are entered multiple times into the database. This could be by mistake, but it could also be that a patient has multiple record in case of multiple tumors. A unique patient identifier can also be used to link a patient database to another database which contains information about the patients.

3a. check for duplicate cases

Use the patient number and year of diagnosis, because patients can have multiple tumors.
Do not forget to save the syntax!

Data > Identify duplicate cases > Define matching cases by Patient and Year



Define matching cases by:
Patient number [patient]
Year of diagnosis [Year]

Sort within matching groups by:

Sort
● Ascending
○ Descending

Number of matching and sorting variables: 2

Variables to Create
☑ Indicator of primary cases (1=unique or primary, 0=duplicate)
● Last case in each group is primary
○ First case in each group is primary
Name: PrimaryLast
☐ Filter by indicator values
☐ Sequential count of matching case in each group (0=nonmatching case)
Name: MatchSequence

☑ Move matching cases to the top of the file
☑ Display frequencies for created variables

In the output you can see that there are 2 duplicate cases. Go to data view and you can see the duplicate cases at the top of your dataset. Duplicate cases have the value 0 for the PrimaryLast variable. Use this variable to delete duplicate cases. Never delete them yourself, because you can make a mistake.

3b. Delete duplicate cases safely.

Data > Select Cases

There are 5 options to select the cases:
-All cases
-Based on a satisfied condition
-A random sample of all cases
-Based on a range
-Using a filter variable.

There are three options for the output:

- Filter out unselected cases (Recommended)
- Copy selected cases to a new data sat
- Delete unselected cases

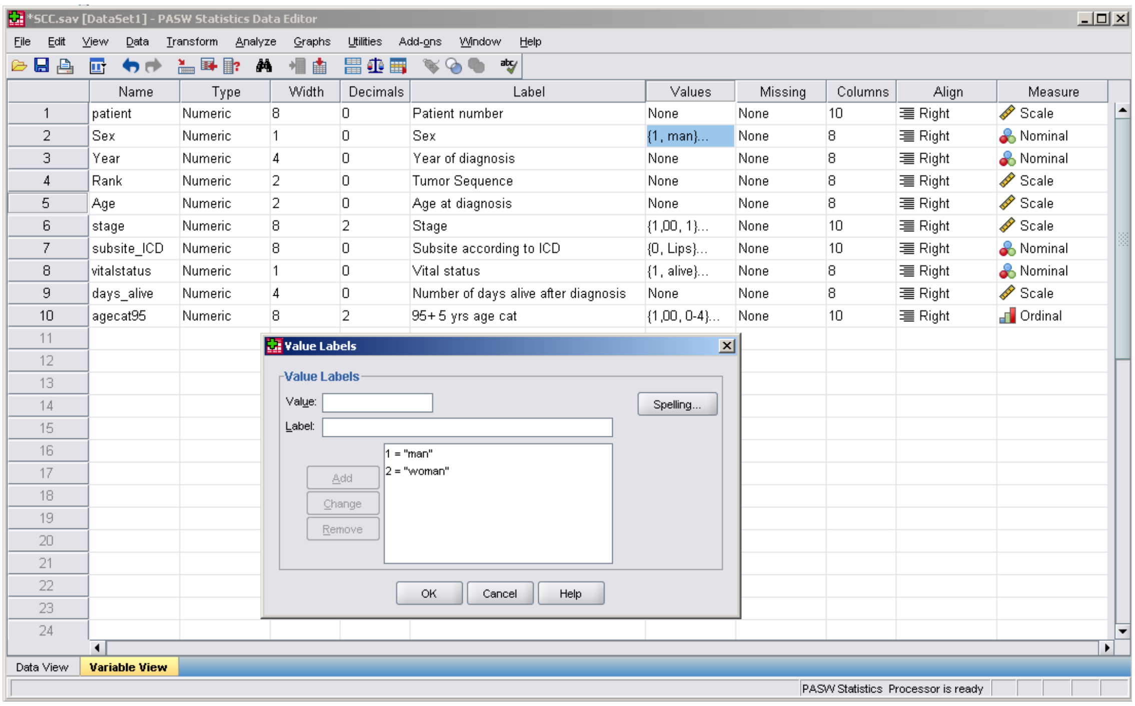
For this assignment: fill in If condition is satisfied PrimaryLast=1 and delete unselected cases. In most situations delete unselected cases is not recommended, because you cannot undo this action. It is better to use ‘filter out unselected cases’, because you can change the selected cases if you made a mistake.

1.4 Data Type

Variables can be numerical or categorical. Numerical data can be subdivided in to discrete variables, which are count data (e.g. like number of seizures) or continuous variable, which can be any value in a range (e.g. age). Categorical data can be subdivided in nominal variables. These categories have no order (e.g. blood type). Ordinal variables are ordered (e.g. disease severity as measured by TNM stage).

In SPSS you can define the variables in Variable View by using data Type and Measure. For example, age can be numeric and Scale (any value from 0 to 100), age category can be numeric and ordinal (1=0-4 yrs, 2=5-9 yrs etc.) or age can be string and ordinal (‘young’, ‘middle aged’, ‘old’). For a categorical variable you can enter a label for each value. See the example for sex on the next page.

Renaming variables, labeling variables and values is not recorded in a syntax if you perform this actions in variable view. Use the RENAME, VARIABLE LABELS and VALUE LABELS statements in the syntax to save these changes. An example of these statements is shown in Chapter 1.6.


*SCC.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Value Labels
Value Labels
Value:
Label:
Spelling...
Add
Change
Remove
OK
Cancel
Help
Data View
Variable View
PASW Statistics Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	patient	Numeric	8	0	Patient number	None	None	10	Right	Scale
2	Sex	Numeric	1	0	Sex	{1, man}...	None	8	Right	Nominal
3	Year	Numeric	4	0	Year of diagnosis	None	None	8	Right	Nominal
4	Rank	Numeric	2	0	Tumor Sequence	None	None	8	Right	Scale
5	Age	Numeric	2	0	Age at diagnosis	None	None	8	Right	Scale
6	stage	Numeric	8	2	Stage	{1,00, 1}...	None	10	Right	Scale
7	subsite_ICD	Numeric	8	0	Subsite according to ICD	{0, Lips}...	None	10	Right	Nominal
8	vitalstatus	Numeric	1	0	Vital status	{1, alive}...	None	8	Right	Nominal
9	days_alive	Numeric	4	0	Number of days alive after diagnosis	None	None	8	Right	Scale
10	agecat95	Numeric	8	2	95+ 5 yrs age cat	{1,00, 0-4}...	None	10	Right	Ordinal

1 = "man"
2 = "woman"

Always fill in the label and the values! This is necessary for your collegues to understand what is in the dataset or for yourself, because if you didn't work with the data for a long period of time you do not remember the meaning of variables and values.

1.5 Explore continuous data

Before you start with data analyses, you explore the dataset. In this way you get a grasp of the distribution of the data, you can check the data for outliers, or you can check assumptions.

1.5a. Explore the variable age.

Analyze > Descriptive Statistics > Descriptives

Obtain standard measures, like mean, standard deviation, median, interquartile range etc.

Analyze > Descriptive Statistics > Explore

More elaborate than descriptives. Obtain measures by category by transferring the category variable to Factor list. Obtain plots and test for normality (for explanation see below)

1.5b. Test if age is normally distributed.

In a linear regression analyses the outcome should be normally distributed. You can check this assumption graphically by plotting a histogram with a normal curve

Graphical check

Analyze > Descriptive Statistics > Frequencies

Transfer age to the 'Variable(s)' box. Go to Charts and click on histogram with normal curve
In the Statistics menu of frequencies you can find handy statistics, like quartiles or percentiles

Statistical test for normality and Quantile-Quantile (Q-Q) plots

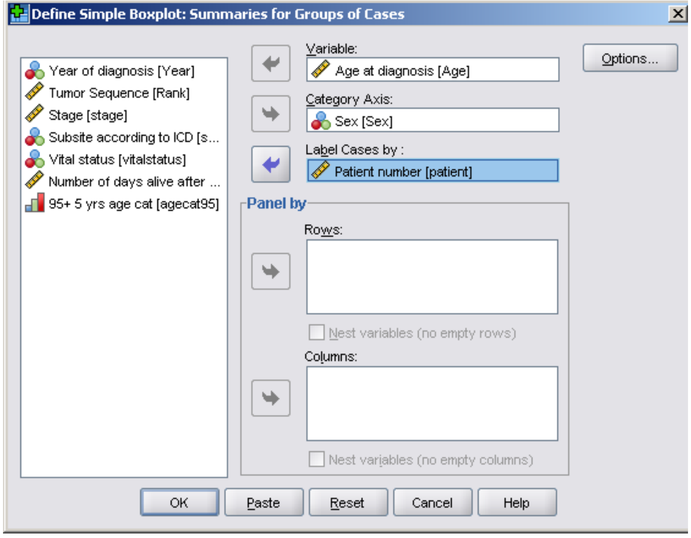
Analyze > Descriptive Statistics > Explore > Plots > Normality plots with tests

The normality assumption can be statistically tested using the Shapiro-Wilk test or visually inspected using a histogram and a Quantile-Quantile (Q-Q) plot. The Shapiro-Wilk test performs better than the Kolmogorov-Smirnov test to test for normality. The null hypothesis is: ‘the variable is normally distributed’. A p-value < 0.05, thus indicates violation of the normality assumption.
The idea of a Q-Q plot is to calculate the expected value for each data point based on the normal distribution. The observed value is plotted against the expected normal distribution. If the datapoints are on the diagonal, this implies normality. Deviation from this straight line indicates that the normality assumption is violated.

1.5c. Make a boxplot for age by sex.

To display the spread of the data you can make a boxplot.

Graphs > Legacy dialogs > Boxplot > Simple

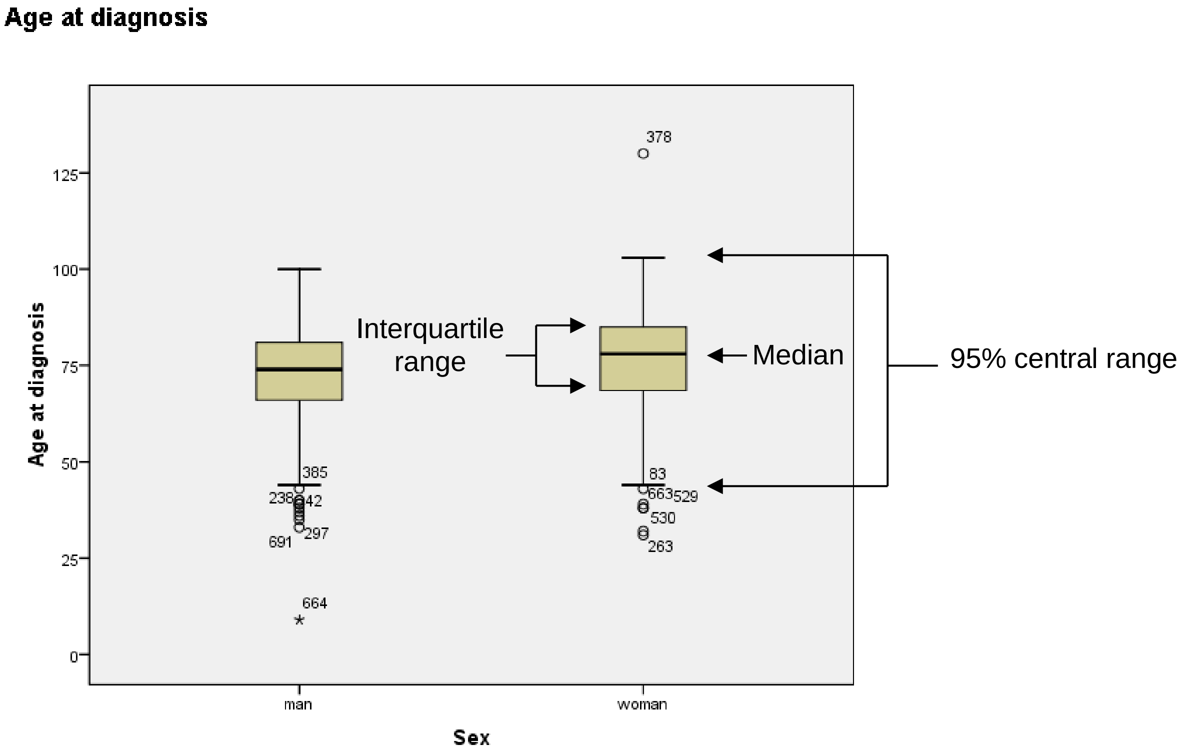

Define Simple Boxplot: Summaries for Groups of Cases
Year of diagnosis [Year]
Tumor Sequence [Rank]
Stage [stage]
Subsite according to ICD [s...
Vital status [vitalstatus]
Number of days alive after ...
95+ 5 yrs age cat [agecat95]
Options...
Panel by
Rows:
☐ Nest variables (no empty rows)
Columns:
☐ Nest variables (no empty columns)
OK
Paste
Reset
Cancel
Help

Variable: Age at diagnosis [Age]

Category Axis: Sex [Sex]

Label Cases by : Patient number [patient]

Fill in the continuous variable and by which variable you would like to group the cases at the category axis. You can label cases by their patientnumber. In this way it is easy to identify, which patient has a strange value for age.

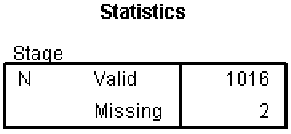


1.6 Explore categorical data

1.6a. Make a frequency table for stage.

Analyze > Descriptive Statistics > Frequencies

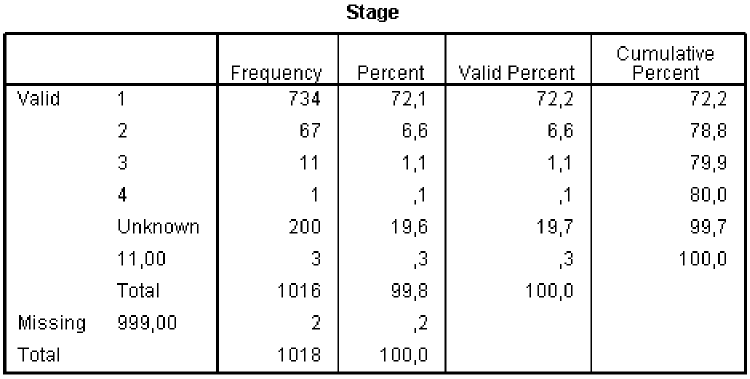
Display the frequencies for each value. You can also make a pie chart. You can check if the distribution among the stages is logical. Indeed, most SCC's are diagnosed in stage I. You can also identify errors in the variables with this function. Stage 11 doesn't exist. This is probably a stage 1 tumor, but you should always try to go back to the patient files to check this before you fill in a new value.



Statistics

Stage

N	Valid	1016
	Missing	2



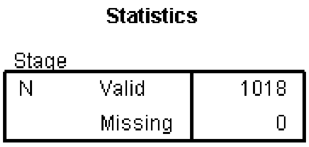
Stage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	734	72,1	72,2	72,2
	2	67	6,6	6,6	78,8
	3	11	1,1	1,1	79,9
	4	1	,1	,1	80,0
	Unknown	200	19,6	19,7	99,7
	11,00	3	,3	,3	100,0
	Total	1016	99,8	100,0	
Missing	999,00	2	,2		
Total		1018	100,0		

1.6b. Discrete Missing values.

Stage information is missing for 2 patients. This was entered into the database as the value 999.

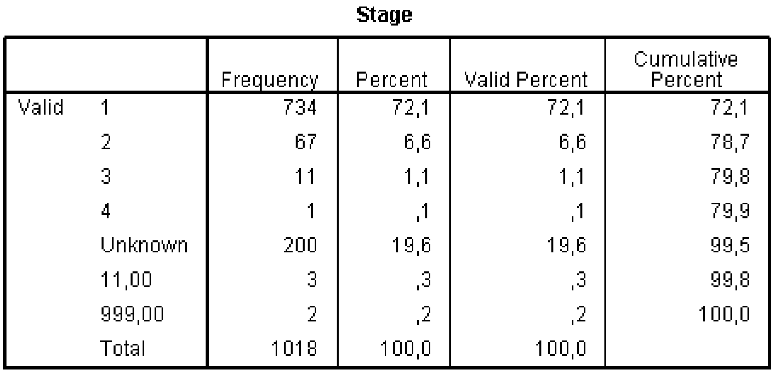
Go to variable View. Go to the ‘Missing’ column on the stage row and check the box ‘No missing values’. Make a frequency table again:
Now you can see, that SPSS doesn’t recognize the missing value. Go back to Missing and fill in ‘999’ in the ‘discrete missing values’ box.



Statistics

Stage

N	Valid	1018
	Missing	0

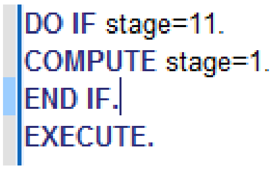


Stage

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	734	72,1	72,1	72,1
	2	67	6,6	6,6	78,7
	3	11	1,1	1,1	79,8
	4	1	,1	,1	79,9
	Unknown	200	19,6	19,6	99,5
	11,00	3	,3	,3	99,8
	999,00	2	,2	,2	100,0
	Total	1018	100,0	100,0	

1.6c. Correct the stage values.

Assuming that you checked the patient files, stage 11, should have been stage 1. To change the values, use the following syntax. In this way, you or your collegues can always retrieve what has been changed.



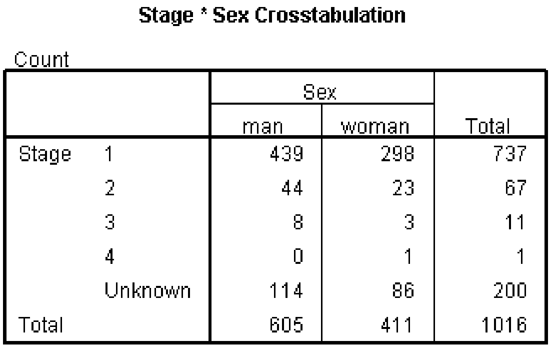
```
DO IF stage=11.
COMPUTE stage=1.
END IF.
EXECUTE.
```

To check if the changes were successful, sort the cases in descending order:

Option 1: Data > Sort Cases
Option 2: Go to data view. Click with the right mouse button on the stage heading and sort descending.

1.6d. Make a Crosstab

To check the stage distribution by sex, you can make a crosstab:
Analyze > Descriptive Statistics > Crosstabs
Fill in Stage as row and Sex as column variables. Go to ‘Statistics’ and tick the box for the Chi-square test.

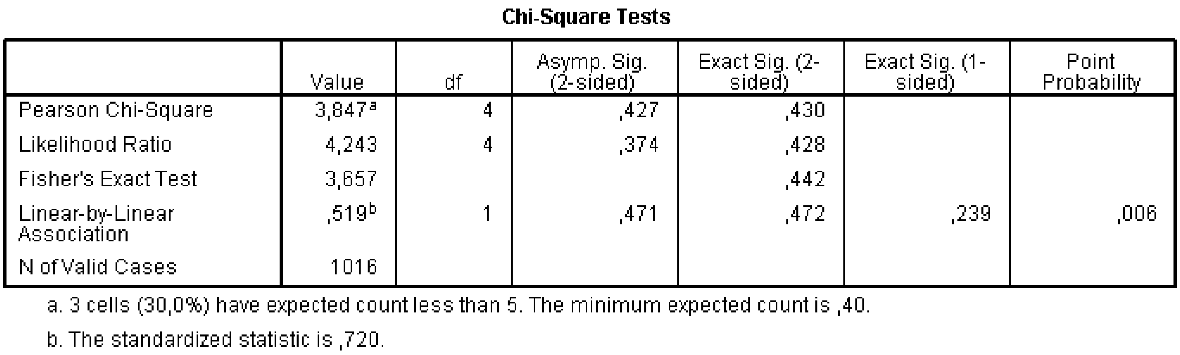


Stage * Sex Crosstabulation

Count

		Sex		Total
		man	woman	
Stage	1	439	298	737
	2	44	23	67
	3	8	3	11
	4	0	1	1
	Unknown	114	86	200
Total		605	411	1016

Three cells contain less then 5 subjects. If you find less then 5 subjects in one cell you should perform the Fisher’s exact test, because the Chi-square test is not reliable with a low number of subjects in one of the cells. Go back to crosstabs and go to ‘Exact..’ to check the option ‘exact’. Do not forget to limit the time (1 minute), because the calculation for a Fisher’s exact test can take a while.



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3,847[a]	4	,427	,430		
Likelihood Ratio	4,243	4	,374	,428		
Fisher's Exact Test	3,657			,442		
Linear-by-Linear Association	,519[b]	1	,471	,472	,239	,006
N of Valid Cases	1016					

a. 3 cells (30,0%) have expected count less than 5. The minimum expected count is ,40.

b. The standardized statistic is ,720.

The p-value obtained by the Fisher’s exact test is 0.442. The null hypothesis for a chisquare test and the Fisher’s exact test is as follows: the proportion of subjects for each stage is equal in the two groups. We can conclude that the proportions are equal.

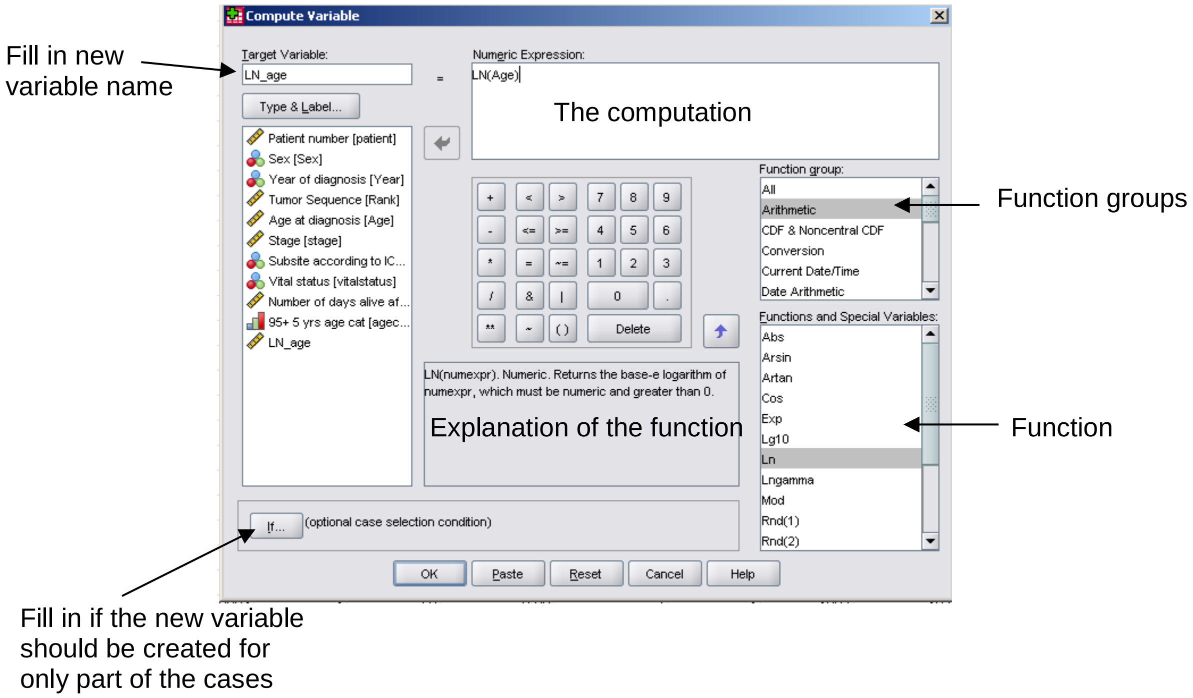
Be aware when you interpret the result of a chi-square test with multiple groups. The null hypothesis is: the proportion of subjects for each category is equal in the three (or more) groups. So from a significant p-value for the chi-square test you can only conclude that the proportion is not equal, but you cannot conclude between which two groups there is a difference.

<u>1.7 Data editing</u>

Compute a new variable

Laboratory values can be right skewed. To normalize right skewed data you can log transform this variable.

Transform > Compute variable



Target Variable: LN_age

Numeric Expression: LN(Age)

LN(numexpr). Numeric. Returns the base-e logarithm of numexpr, which must be numeric and greater than 0.

1.7a. Plot the lab values, log transform the lab values and make a new histogram:

Graphs > Legacy Dialogs > Histogram

Use Lab_value as variable. You can also tick the box ‘Display normal curve’. ‘You can also obtain a histogram by using the chart builder (explained in paragraph 3.1).

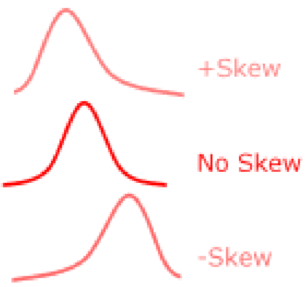
Graphs > Chart Builder

```
GRAPH
 /HISTOGRAM=Lab_value.

COMPUTE LN_lab_value=LG10(Lab_value).
EXECUTE.

GRAPH
 /HISTOGRAM=LN_lab_value.
```

Log transformations are only effective for right skewed data. Age was not normally distributed, but a little left skewed. Log transformations make left skewed data even more left skewed.



Right skewed (positively skewed)

Left skewed (negatively skewed)

1.7b.Repeat the syntax for age

Recode

Transform > Recode

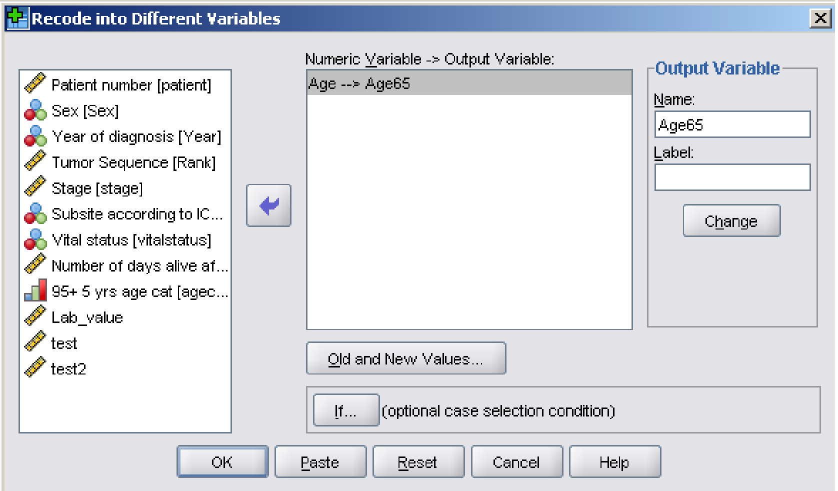
There are three options for recoding data, which you can find in the Transform menu
-Recode into the same variables
-Recode into different variables
-Automatic recode

I advise never to use Recode into the same variables, because you will lose your original data. Automatic recode recodes numerical and string variables into a new numerical variable. The value labels will be automatically displayed in 'Values' in variable view.
You can choose you own new categories with 'recode into different variables'. This is the recommended function for recoding variables.

1.7c. Recode age in two new categories (< 65 years and ≥ 65 yrs)

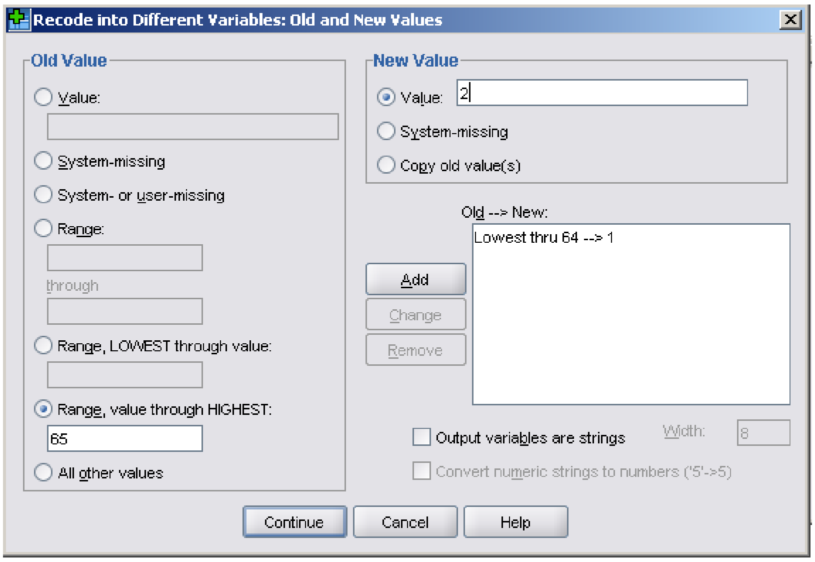
Transform > Recode into different variables

Transfer age at diagnosis to the 'Numerical variable -> output variable' box. Fill in the new variable name and click on 'change. Go to 'Old and New Values'.



Numeric Variable -> Output Variable: Age --> Age65

Name: Age65


Recode into Different Variables: Old and New Values
Old Value
○ Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ All other values
New Value
○ System-missing
○ Copy old value(s)
Add
Change
Remove
☐ Output variables are strings
Width: 8
☐ Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

● Range, value through HIGHEST: 65

● Value: 2

Old --> New: Lowest thru 64 --> 1

Fill in the old values, the new value and click on add. Do not forget to label the new values in the variable view screen or use the following syntax for labelling:

```
Variable labels age65 'Age below or above 65 yrs'.
Value labels Age65  1'<65 yrs'  2'>= 65 yrs'.
```

Sort data

Data > Sort cases

Patients can have multiple tumours, which were diagnosed in different years.

1.7d. Sort cases by patient number and year of diagnosis:

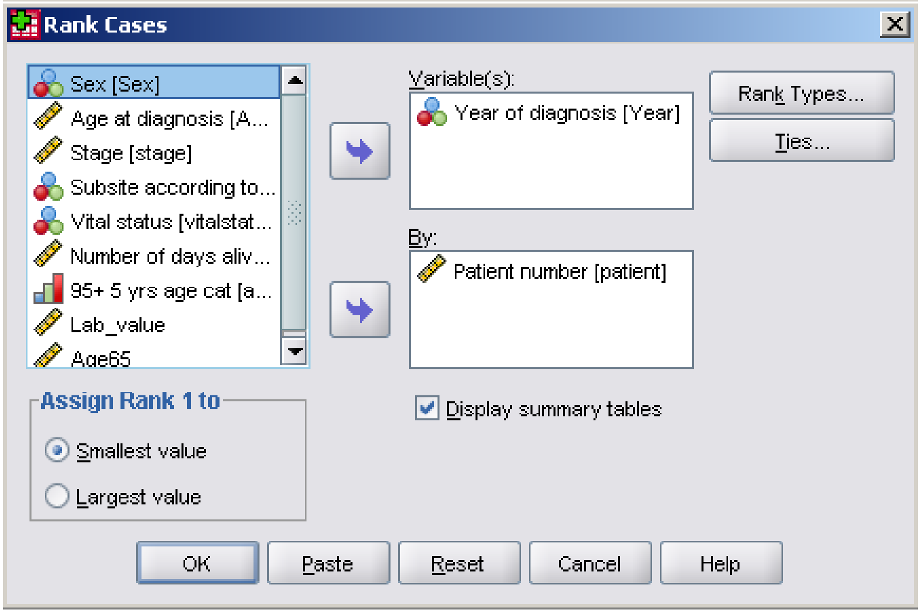
```
SORT CASES BY patient(A) Year(A).
```

Rank cases

1.7e. Make a new variable which indicates the sequence of the tumours

Transform > Rank cases

You would like to rank year of diagnosis by patient number



Variable(s): Year of diagnosis [Year]

By: Patient number [patient]

Assign Rank 1 to
● Smallest value
○ Largest value

☑ Display summary tables

1.7f. Rename the variable which was created by the Rank cases function of SPSS.

Go to Variable view and rename the variable or use the following syntax:

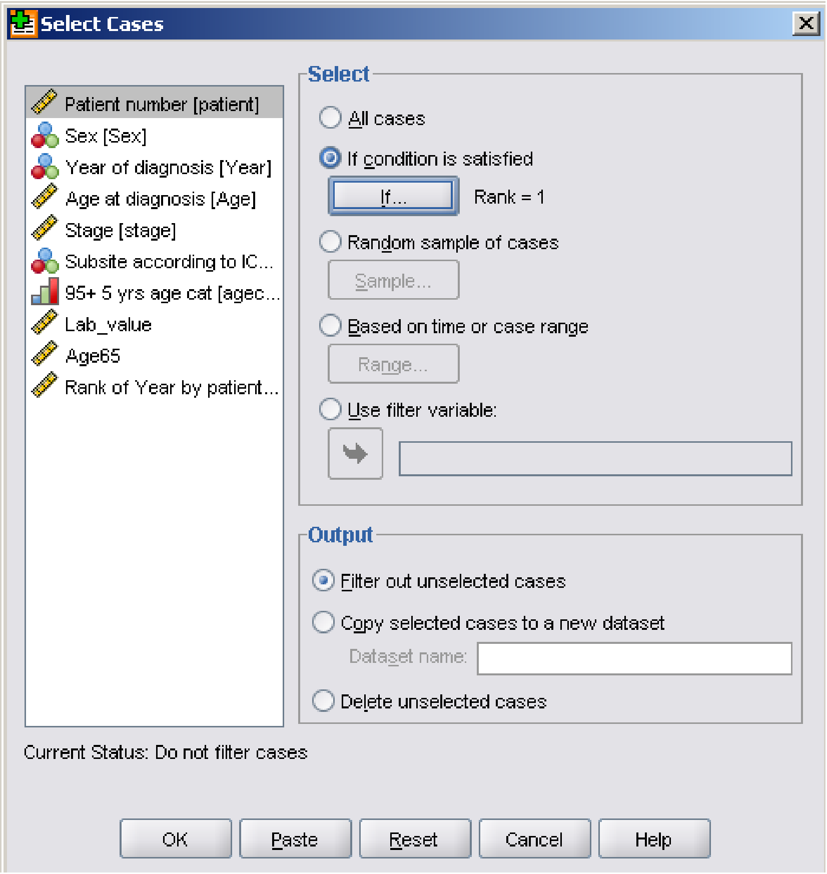
```
RENAME VARIABLES (RYear=Rank).
```

Sort patient number ascending to check the result for patient number 4. Patients with only 1 tumor have only the value 1. Patient number 4 has 4 tumors and should have rank values 1, 2, 3 and 4. It may be convenient to put the variables next to each other.

Select cases

Data > Select cases

1.7g. Select the first tumor of each patient.
You can select the cases if the rank is equal to 1 and filter out unselected cases.



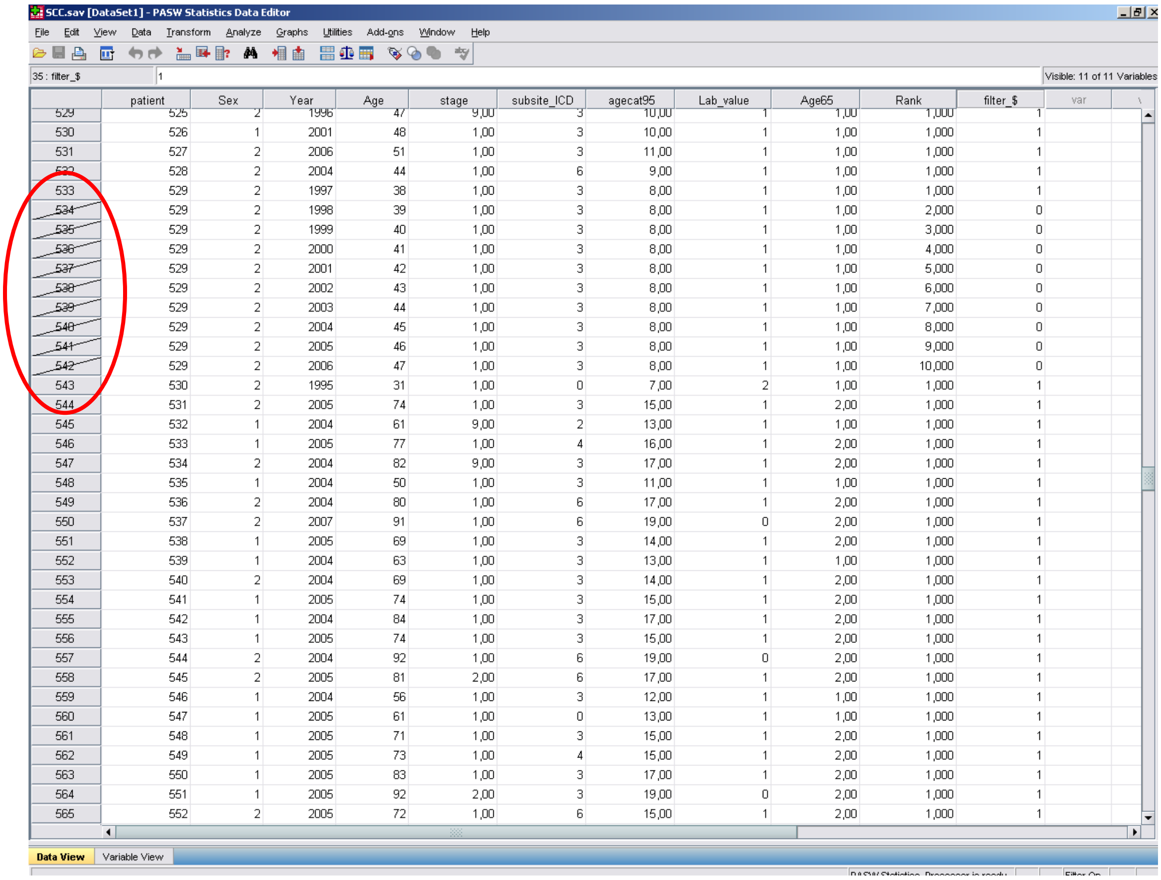
Select
○ All cases
● If condition is satisfied
If... Rank = 1
○ Random sample of cases
○ Based on time or case range
○ Use filter variable:

Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
○ Delete unselected cases

Current Status: Do not filter cases

Result is shown on the next page.

Result of the filter:



	patient	Sex	Year	Age	stage	subsite_ICD	agecat95	Lab_value	Age65	Rank	filter_$
529	525	2	1996	47	9,00	3	10,00	1	1,00	1,000	1
530	526	1	2001	48	1,00	3	10,00	1	1,00	1,000	1
531	527	2	2006	51	1,00	3	11,00	1	1,00	1,000	1
532	528	2	2004	44	1,00	6	9,00	1	1,00	1,000	1
533	529	2	1997	38	1,00	3	8,00	1	1,00	1,000	1
~~534~~	529	2	1998	39	1,00	3	8,00	1	1,00	2,000	0
~~535~~	529	2	1999	40	1,00	3	8,00	1	1,00	3,000	0
~~536~~	529	2	2000	41	1,00	3	8,00	1	1,00	4,000	0
~~537~~	529	2	2001	42	1,00	3	8,00	1	1,00	5,000	0
~~538~~	529	2	2002	43	1,00	3	8,00	1	1,00	6,000	0
~~539~~	529	2	2003	44	1,00	3	8,00	1	1,00	7,000	0
~~540~~	529	2	2004	45	1,00	3	8,00	1	1,00	8,000	0
~~541~~	529	2	2005	46	1,00	3	8,00	1	1,00	9,000	0
~~542~~	529	2	2006	47	1,00	3	8,00	1	1,00	10,000	0
543	530	2	1995	31	1,00	0	7,00	2	1,00	1,000	1
544	531	2	2005	74	1,00	3	15,00	1	2,00	1,000	1
545	532	1	2004	61	9,00	2	13,00	1	1,00	1,000	1
546	533	1	2005	77	1,00	4	16,00	1	2,00	1,000	1
547	534	2	2004	82	9,00	3	17,00	1	2,00	1,000	1
548	535	1	2004	50	1,00	3	11,00	1	1,00	1,000	1
549	536	2	2004	80	1,00	6	17,00	1	2,00	1,000	1
550	537	2	2007	91	1,00	6	19,00	0	2,00	1,000	1
551	538	1	2005	69	1,00	3	14,00	1	2,00	1,000	1
552	539	1	2004	63	1,00	3	13,00	1	1,00	1,000	1
553	540	2	2004	69	1,00	3	14,00	1	2,00	1,000	1
554	541	1	2005	74	1,00	3	15,00	1	2,00	1,000	1
555	542	1	2004	84	1,00	3	17,00	1	2,00	1,000	1
556	543	1	2005	74	1,00	3	15,00	1	2,00	1,000	1
557	544	2	2004	92	1,00	6	19,00	0	2,00	1,000	1
558	545	2	2005	81	2,00	6	17,00	1	2,00	1,000	1
559	546	1	2004	56	1,00	3	12,00	1	1,00	1,000	1
560	547	1	2005	61	1,00	0	13,00	1	1,00	1,000	1
561	548	1	2005	71	1,00	3	15,00	1	2,00	1,000	1
562	549	1	2005	73	1,00	4	15,00	1	2,00	1,000	1
563	550	1	2005	83	1,00	3	17,00	1	2,00	1,000	1
564	551	1	2005	92	2,00	3	19,00	0	2,00	1,000	1
565	552	2	2005	72	1,00	6	15,00	1	2,00	1,000	1

Syntax of the filter:

```
USE ALL.
COMPUTE filter_$=(Rank = 1).
VARIABLE LABEL filter_$ 'Rank = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

In the syntax you can see, that SPSS first computes a filter variable, subsequently labels the values and finally filters the cases by the filter variable. You can save this filter variable or you can create your own filter variable.

```
USE ALL.
COMPUTE New_filter=(Age > 65 AND subsite_ICD=6).
VARIABLE LABEL New_filter 'Filter older then 65 and SCC on arms or shoulder'.
VALUE LABELS New_filter 0 'Not Selected' 1 'Selected'.
FORMAT New_filter (f1.0).
FILTER BY New_filter.
EXECUTE.
```

Don’t forget to turn the filter off!

```
FILTER OFF.
USE ALL.
EXECUTE.
```

1.8 Export a data set

Some analyses can only be done or can be easier performed in other statistical package like STATA, SAS or R. You can export a datafile in the correct format by using the Save As option the the File menu. For SAS use the long extension (*.sas7bdat). If the format for your package is not available in the picklist, save the data as a tab delimited file (*.dat) or a comma delimited file (*.csv). These datafiles are usually easy to import into other statistical packages.

Assignment 1.1 Data exploration

Data: Skin cancer screening
Patients at risk for skin cancer received a full body examination. A questionnaire was filled in by the patients to determine factors which increase skin cancer risk

Open the dataset 'Skin_cancer_screening' and check which variables are included.

1. ID is an unique patient identifier. Check for duplicate cases and delete possible duplicate case by using a syntax (i.e. don't delete duplicate cases by hand). How many duplicated cases were there in the data?

2. Compute age at the date of screening by using birth date. Use a function from the 'Date Extraction' function group.

3. Sort cases by ascending age.

4a. Many patients have missing values on their date of birth and as a consequence on age. Discrete missing values for age with a non-existing age value, for example 999. Do not choose 99 for missings, because patients can also be 99.
4b. Compute 999 for missing values on age. Use the IF option and use a function from the 'Missing Values' function group.

5a. Explore the variable age. Make a boxplot for age and check for outliers.
5b. Write down the casenumbers who have a strange value for age.

6a.How many males and females participated in the screening?
6b. Discrete missing values for sex and recalculate the frequency table

7. How many females older than 40, but younger than 61 participated in the screening? Recode age to answer this question.

8. How many patients were diagnosed with at least one melanoma?

Chapter 2 Simple Inferential Statistics

Data: SCC
The data used is this chapter is equal to the data in chapter 1, but in chapter 2 only the first SCC is included.

The following figure is useful to determine which statistical test is needed to answer the research question:

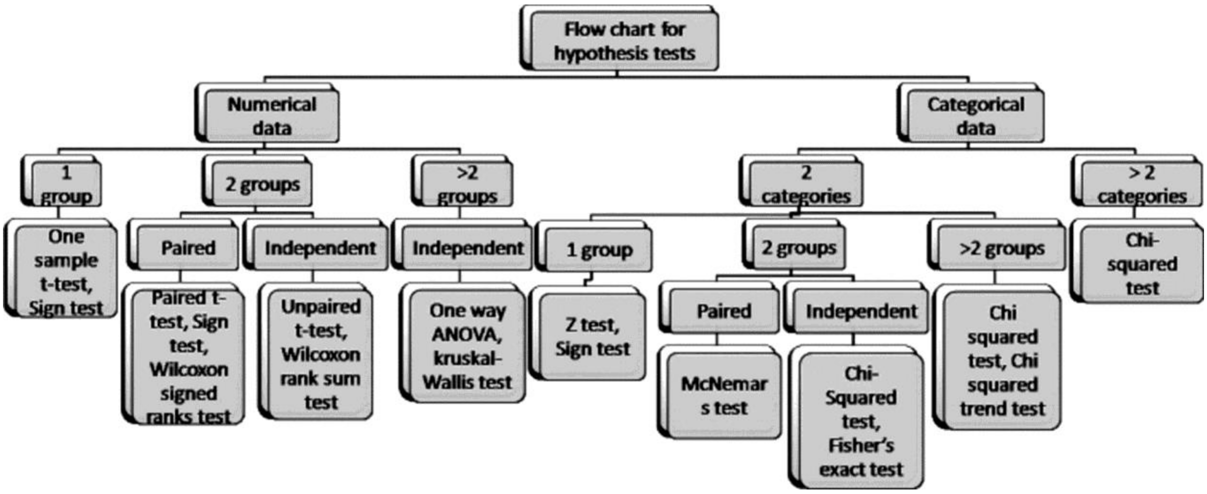

Flow chart for hypothesis tests
Numerical data
Categorical data
1 group
2 groups
>2 groups
2 categories
>2 categories
One sample t-test, Sign test
Paired
Independent
Independent
1 group
2 groups
>2 groups
Chi-squared test
Paired t-test, Sign test, Wilcoxon signed ranks test
Unpaired t-test, Wilcoxon rank sum test
One way ANOVA, kruskal-Wallis test
Z test, Sign test
Paired
Independent
Chi squared test, Chi squared trend test
McNemars test
Chi-Squared test, Fisher's exact test

Figure 2.1: flowchart statistical tests

In the previous chapter you have learned about the different types of numerical and categorical data. To use this figure you need to determine if you would like to compare the mean or the distribution in 1 group, between 2 groups or more than 2 groups. You also need to determine if data are paired (dependent) or unpaired (independent). Paired data means that the outcome variable is measured in the same subjects twice. For example, the blood pressure of a group of patients before *and* after the intervention. Unpaired data means that the outcome variable is measured in different groups of patients. For example: the blood pressure of group A (who received the intervention) and the blood pressure of group B (who received placebo).

Statistical Test	**Location SPSS**
Numerical data	
Parametric tests	
One sample t-test	Analyze > Compare means > One sample t-test
Paired t-test	Analyze > Compare means > Paired samples t-test
Unpaired t-test	Analyze > Compare means > Independent samples t-test
One way ANOVA	Analyze > Compare means > One way ANOVA
Nonparametric tests	
Sign test	Analyze > Nonparametric tests >
Wilcoxon signed rank test	Analyze > Nonparametric tests > Related samples
Wilcoxon rank sum test/ Mann-Whitney U test	Analyze > Nonparametric tests > Independent samples
Kruskal-Wallis test	Analyze > Nonparametric tests > Independent samples

Statistical Test	Location SPSS
Categorical data	
Z-test	Obtain value for p from Frequencies
Sign test	Obtain value for p from Frequencies
McNemar’s test	Analyze > Descriptive Statistics > Crosstabs > Statistics > McNemar
Chisquare test	Analyze > Descriptive Statistics > Crosstabs > Statistics > Chisquare
Fisher’s exact test	Analyze > Descriptive Statistics > Crosstabs > Exact > Exact
Chis-squared trend test	Analyze > Descriptive Statistics > Crosstabs > Statistics > Ordinal

2.1 Student’s t-test

A Student’s t-test is used to compare the means between two groups. For example, if you want to compare the mean height between men and women. The t refers to the t-distribution, which approximates a normal distribution with large sample sizes. There is a t-test for paired and for unpaired data.

2.1a Research Question: is the age of the first tumor different between men and women?
Step 1. Determine which statistical test you need to answer this question and define the null hypothesis (H_0).

H_0: the distribution of age is equal between men and women.

Step 2. Check the assumptions of the t-test:

The assumptions of the t-test are:
- Data is normally distributed
- Variances are equal (separate test is not needed, this is incorporated in the output of the t-test)

You can check this by using frequencies or the Shapiro-Wilk test and Q-Q plots.

Analyze > Descriptive Statistics > Frequencies
Analyze > Descriptive Statistics > Explore > Plots > Normality plots with tests

Age was not completely normally distributed, but based on the histogram age can be regarded as approximately normally distributed. If the sample size is large enough (ca. >30) then the central limit theorem holds, even if the variable is not completely normally distributed. For small sample sizes the Shapiro-Wilk test is recommended.

Central limit theorem:

$$\overline{X} = N(\mu, \frac{\sigma}{\sqrt{n}})$$

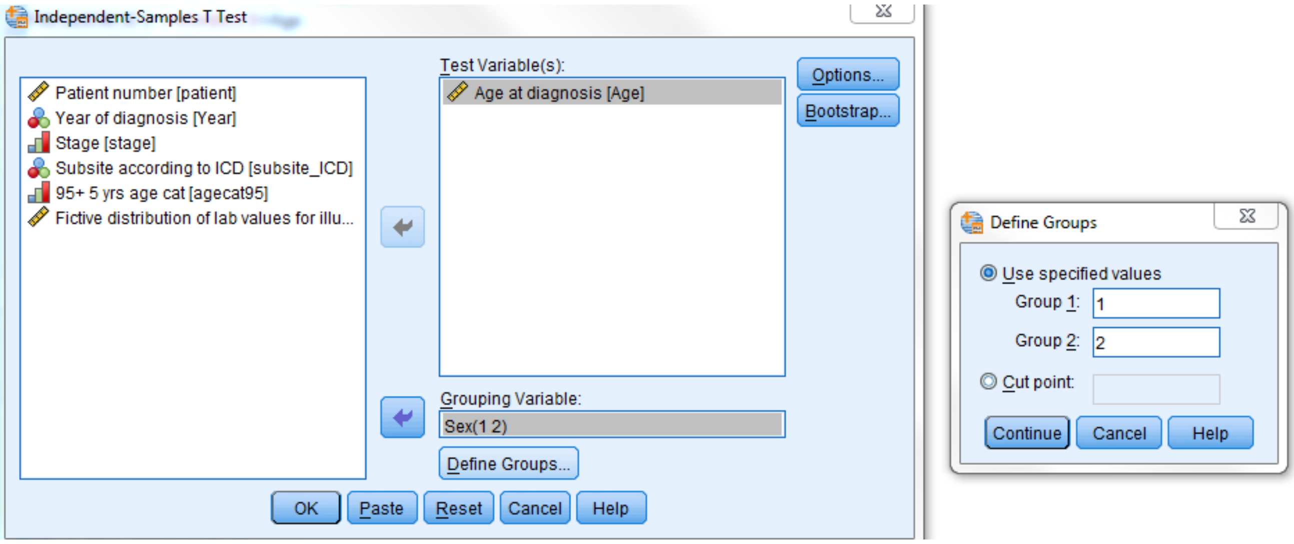
Meaning of this formula:

The mean of a random variable ($\overline{X}$) follows a normal distribution (N), with a mean value (μ), which is equal to the population mean with a standard deviation which is equal to the standard deviation of the population (σ) divided by the square root of the sample size (n).
This approximation becomes better if your sample size (n) is larger.

Step 3. Perform the unpaired t-test.

Analyze > Compare Means > Independent Samples t-test

Age is the test variable and sex is the grouping variable. Define the values of the groups.



Test Variable(s): Age at diagnosis [Age]

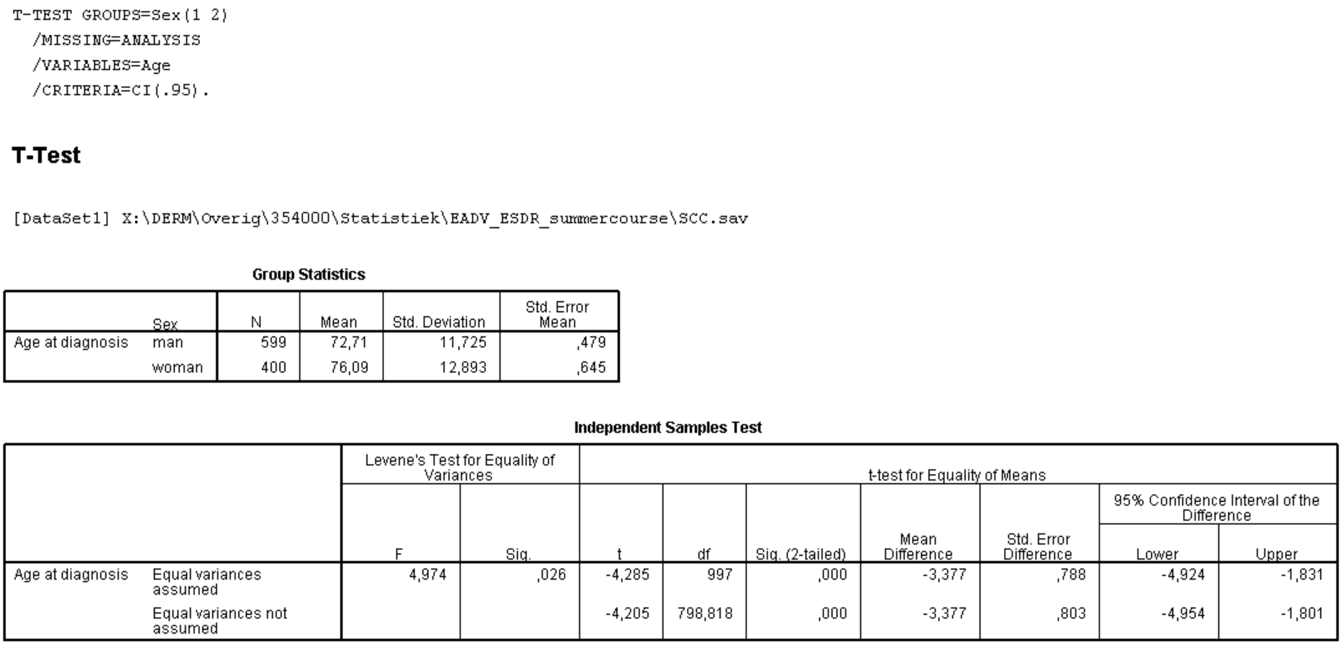
Grouping Variable: Sex(1 2)

● Use specified values

Group 1: 1

Group 2: 2

○ Cut point:



```
T-TEST GROUPS=Sex(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=Age
  /CRITERIA=CI(.95).
```

T-Test

```
[DataSet1] X:\DERM\Overig\354000\Statistiek\EADV_ESDR_summercourse\SCC.sav
```

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Age at diagnosis	man	599	72,71	11,725	,479
	woman	400	76,09	12,893	,645

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Age at diagnosis	Equal variances assumed	4,974	,026	-4,285	997	,000	-3,377	,788	-4,924	-1,831
	Equal variances not assumed			-4,205	798,818	,000	-3,377	,803	-4,954	-1,801

Step 4. Interpretation of the output
In the first table you can see the number of subjects included in the analyses as well as the mean, standard deviation (SD) and the standard error (SE).
SD=$\sqrt{\delta^2}$=Square root of the variance (interpretation: variance in this sample)
SE=$\frac{SD}{\sqrt{N}}$=SD/Square root sample size (interpretation: variance of the mean if you would sample multiple times)

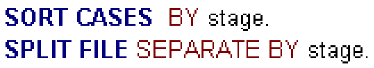
First look at Levene's test for equality of variances. The null hypothesis for this test is: The variances of both groups are equal. A non-significant result means that the variances are equal and that you can use the output for equal variances assumed. The p-value in this analysis is 0.026, so we should look at the results for equal variances NOT assumed. The mean difference is -3.377 years with a 95 % CI of -4.954 to -1.801. From the table with group statistics you can calculate the same mean difference. The value 0 is not in the confidence interval. 0 means no difference. Therefore you already know that the p-value is significant. The significance (2-tailed) is 0.000. A p-value cannot be zero. This means that the p-value < 0.001.

2.1b Is the age of the first tumor different between men and women for each stage?

Perform the same t-test as for assignment 2.1a, but organise the output by stage.

Data > Split File

Check the box ‘Organize output by groups’ and transfer ‘stage’ to ‘Groups based on’. Repeat the unpaired sample t-test.



```
SORT CASES  BY stage.
SPLIT FILE SEPARATE BY stage.
```

Stage I: p-value < 0.001 and mean difference = -3.373 (95% CI -5.254 to -1.492)
Stage II: p-value = 0.026 and mean difference = -5.045 (95% CI -9.481 to -0.608)
Stage III: p-value = 0.473 and mean difference = -6.208 (95% CI -24.948 to 12.531)
Stage IV not calculated

2.2 Non-parametrical test

As you have probably noted, the sample size of stage III is rather small to perform a parametrical test (n=11). The non-parametrical equivalent of an unpaired samples t-test would be more appropriate for this stratum. The non-parametrical test ranks all values and the test statistic is based on the value of the ranks, rather than the true value of the measurements.

2.2 Is the age of the first tumor different between men and women for stage III?

Step 1: determine the appropriate test

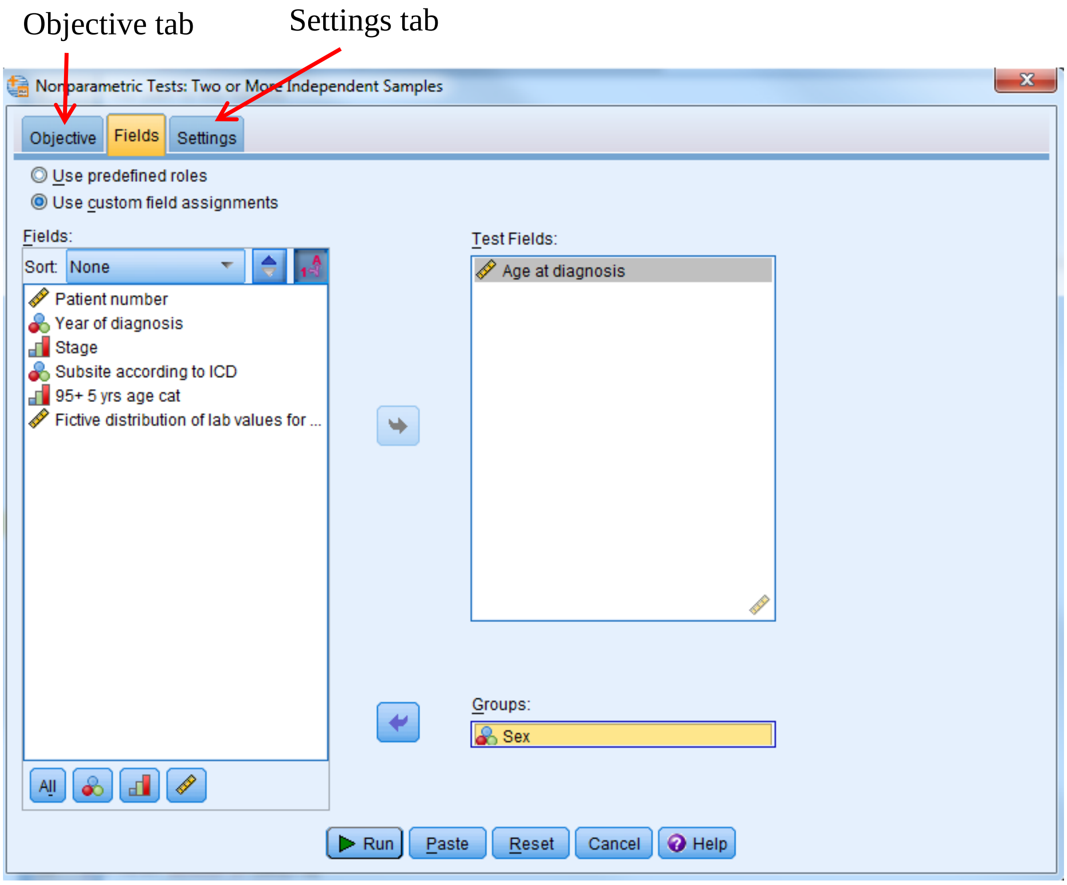
Use Figure 2.1: age is numerical data. We would like to compare 2 groups (men and women). Men and women are independent groups. The sample size is small and age is not normally distributed, thus a non-parametrical test is needed, which is the Wilcoxon signed rank sum test.

Step 2: perform the test and define the null hypothesis.

H_0: The distribution of age is the same between men and women.

Menu > Non-parametrical tests > Independent samples

Choose to ‘customize analysis’ in the ‘Objective’ tab. Go to the ‘Fields’ tab and assign the test variable and the groups in the ‘fields’ tab. Go to the ‘Settings’ tab and choose ‘Customize tests’. Choose ‘Mann-Whitney U test (2 samples)’, which is equivalent to Wilcoxon rank sum test (Figure 2.1).


Objective tab

Settings tab

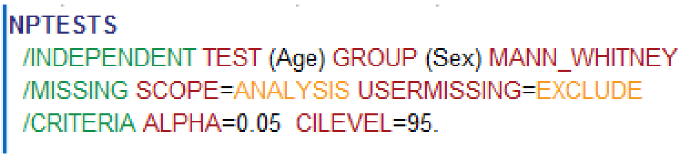
Nonparametric Tests: Two or More Independent Samples

○ Use predefined roles
● Use custom field assignments

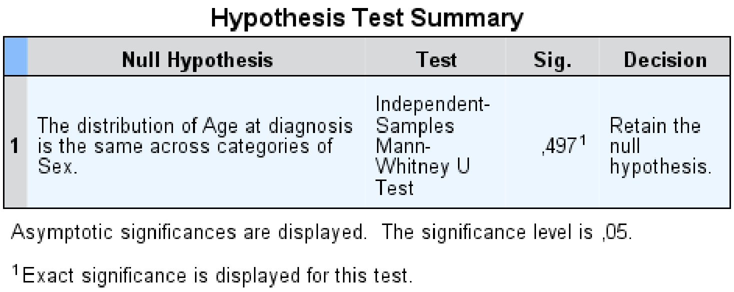
Fields:
Sort: None
Patient number
Year of diagnosis
Stage
Subsite according to ICD
95+ 5 yrs age cat
Fictive distribution of lab values for ...

Test Fields:
Age at diagnosis

Groups:
Sex


```
NPTESTS
  /INDEPENDENT TEST (Age) GROUP (Sex) MANN_WHITNEY
  /MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
  /CRITERIA ALPHA=0.05  CILEVEL=95.
```

Stage = 3


Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Age at diagnosis is the same across categories of Sex.	Independent-Samples Mann-Whitney U Test	,497[1]	Retain the null hypothesis.

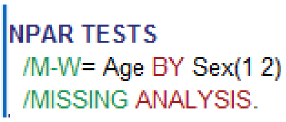
Asymptotic significances are displayed. The significance level is ,05.

[1] Exact significance is displayed for this test.

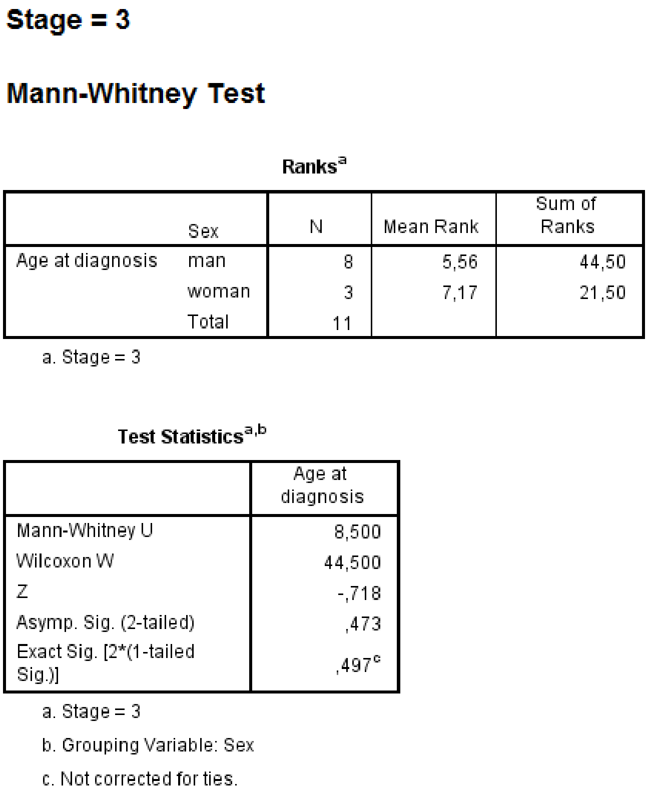
You can also use:

Menu > Non-parametrical tests > Legacy dialogs > 2 Independent samples

The second option via legacy dialogs provides more information on the test statistic, which is based on the rank of age, rather than the true value of each data point.



```
NPAR TESTS
  /M-W= Age BY Sex(1 2)
  /MISSING ANALYSIS.
```



Stage = 3

Mann-Whitney Test

Ranks[a]

	Sex	N	Mean Rank	Sum of Ranks
Age at diagnosis	man	8	5,56	44,50
	woman	3	7,17	21,50
	Total	11		

a. Stage = 3

Test Statistics[a,b]

	Age at diagnosis
Mann-Whitney U	8,500
Wilcoxon W	44,500
Z	-,718
Asymp. Sig. (2-tailed)	,473
Exact Sig. [2*(1-tailed Sig.)]	,497[c]

a. Stage = 3

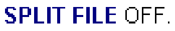
b. Grouping Variable: Sex

c. Not corrected for ties.

Step 3: Interpret the output.

The null hypothesis can be found in the output:
‘The distribution of age at diagnosis is the same across categories of sex’
You will receive the output for all stages, but only stage III is shown in this chapter. The p-value for stage III is 0.497. Thus, the null hypothesis cannot be rejected. There is no statistical evidence that age is different between men and women for stage III SCC.

Do not forget to turn split file off!



```
SPLIT FILE OFF.
```

2.3 ANOVA

ANOVA means ANalyis Of Variance and this type of analysis is used to compare the means of more than 2 groups. A one-way ANOVA compares the means by one grouping factor (categorical variable).

2.3 Is the age distribution different between the different body sites?

Step 1: Define the Null hypothesis

H_0: The mean age is equal across all categories of subsite.

Note: the result of the ANOVA will only tell you if there is a difference in age between the categories, but not between which categories. Therefore multiple post hoc t-tests are needed with a correction for multiple comparisons.

Step 2: Are the model assumptions fulfilled?

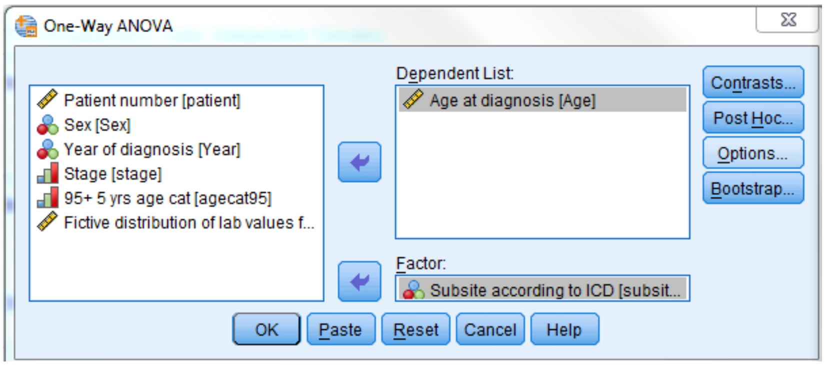
The model assumptions of ANOVA are:
-Normality
-Homogeneity of variances

For larger sample sizes, the normality assumption is not important. The homogeneity of variances assumption is an important assumption of an ANOVA and can be tested using Levene's test of equal variances. The assumption can be tested in the ANOVA procedure, thus a separate check beforehand is not needed.

Step 3: Perform the analysis

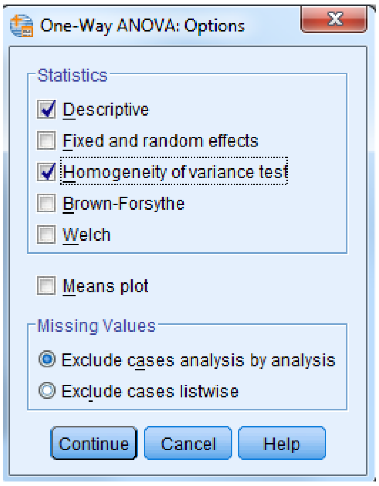
Analyze > Compare Means > one-way ANOVA

Put age in the dependent list and the grouping variable (subsite_ICD) is the factor. Go to options and tick the box 'descriptives' and 'homogeneity of variances test'. Also go to 'post hoc' and tick the box 'Bonferroni'. This is the most widely used correction for multiple comparisons. The α-level should be divided by the number of comparisons to reach statistical significance. For example, in case of 4 comparisons, the p-value should be below 0.05/4=0.0125 to reach statistical significance. This is equal to multiplication of the p-value with the number of comparisons and significance at $p < 0.05$ level.



Dependent List: Age at diagnosis [Age]

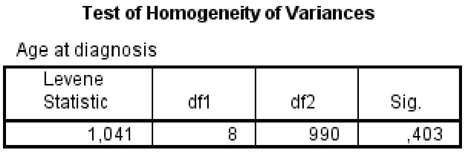
Factor: Subsite according to ICD [subsit...



Statistics
☑ Descriptive
☐ Fixed and random effects
☑ Homogeneity of variance test
☐ Brown-Forsythe
☐ Welch

☐ Means plot

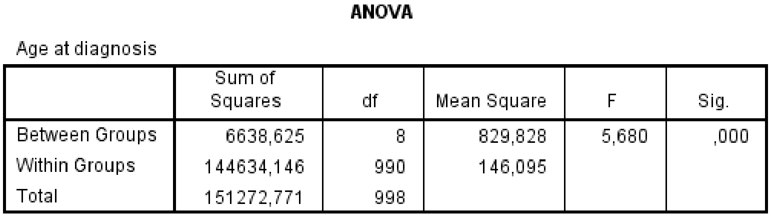
Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise



Test of Homogeneity of Variances

Age at diagnosis

Levene Statistic	df1	df2	Sig.
1,041	8	990	,403



ANOVA

Age at diagnosis

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6638,625	8	829,828	5,680	,000
Within Groups	144634,146	990	146,095		
Total	151272,771	998			

Step 4: interpret the result
First look at the test of homogeneity of variance. The null hypothesis is that the variances are equal across the categories of subsite. A non-significant results thus indicates homogeneity of variances. In this case the p-value is 0.312, which indicates that the model assumption is fulfilled.
The mean age for each subsite is shown in the descriptives table. Double click on the ANOVA table and the p-value. The significance of the ANOVA is $4.45*10^{-7}$, thus the null hypothesis is rejected. In the post hoc test, each category is compared to every other category, using multiple independent samples t-test with a Bonferroni corrected p-value. In this table the p-values of each comparison has been multiplied with the number of comparisons, which means that a significant p-value should be <0.05.

2.4 Chis-square test / Fisher's exact test

To compare the distribution across categorical variables, a chi-square test is used. The chi-square value is not valid, when the number of subjects is less than 5 in one of the cells. In that case a Fisher's exact test should be used.

2.3 Is the distribution between men and women different for different stages?

Step 1: determine which test should be used and define the null hypothesis.
Subjects from different stages are independent groups. From a crosstab between stage and sex can be seen, that the number of subjects in some cell is less than 5, which means that a Fisher's exact test should be used.

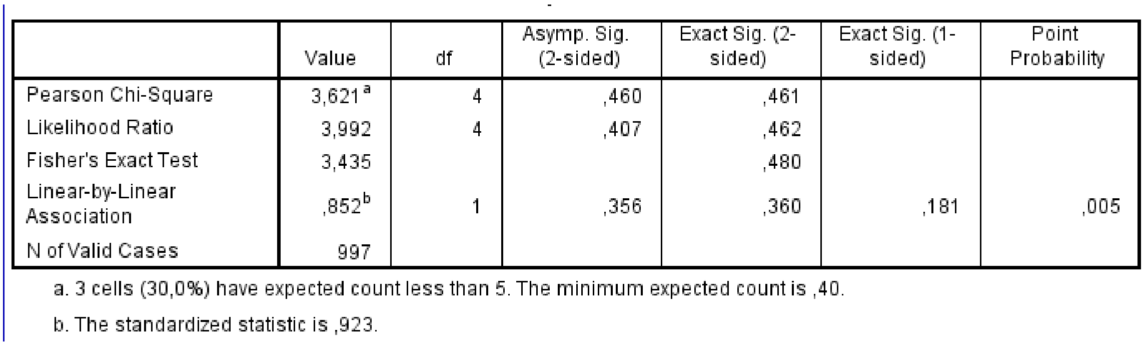
H_0: the proportions of men and women are equal between the groups of stage.

Step 2: Perform the test

Analyze > Descriptive statistics > Crosstabs

Go to statistics and tick the box 'Chi-square' (even if you are going to perform a Fisher's exact test). Go to exact and tick the box 'exact'.

Step 3: Interpret the result



	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3,621[a]	4	,460	,461		
Likelihood Ratio	3,992	4	,407	,462		
Fisher's Exact Test	3,435			,480		
Linear-by-Linear Association	,852[b]	1	,356	,360	,181	,005
N of Valid Cases	997					

a. 3 cells (30,0%) have expected count less than 5. The minimum expected count is ,40.

b. The standardized statistic is ,923.

The 2-sided p-value of the Fisher’s exact test is 0.480, which indicates that the null hypothesis is not rejected.

Assignment 2.1 Simple inferential statistics

Data: skin cancer screening
Patients at risk for skin cancer received a full body examination. A questionnaire was filled in by the patients to determine factors which increase skin cancer risk

Use Figure 2.1 for this assignment. Questions with an asterisk (*) have hints on the next page.

1. Is the age distribution between patient with and without AK different? Define the null and the alternative hypothesis and perform the correct statistical test.

Patients with AK received a 4-week treatment with 5-fluoro-uracil (5-FU) cream. This treatment has side effects, such as redness of the skin, burning sensation and pain. Patients were asked to fill in a questionnaire, which included a visual analogue score for pain score from 0 (no pain) to 10 (severe pain) before treatment (week 0), during treatment (week 2 and 4) and after treatment (week 6).

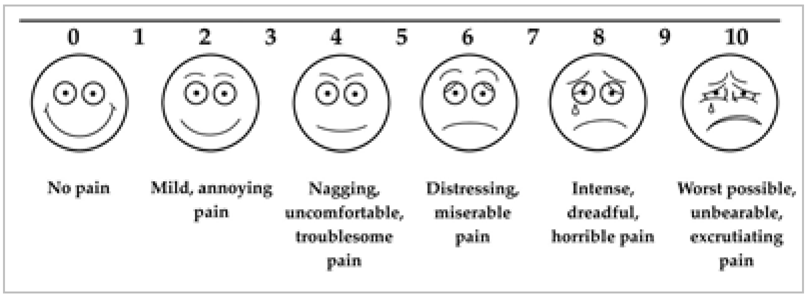

0
1
2
3
4
5
6
7
8
9
10
No pain
Mild, annoying pain
Nagging, uncomfortable, troublesome pain
Distressing, miserable pain
Intense, dreadful, horrible pain
Worst possible, unbearable, excrutiating pain

Figure 1: visual analogue scale for pain

2a. How should the 95% CI for the mean pain score at week 2 be calculated? Choose between option 1 and 2.
Option 1: Mean +/- 1.96*SD (standard deviation) = 3.374 +/- 1.96*1.571 = 0.295 to 6.453
Option 2: Mean +/- 1.96*SE (standard error) = 3.374 +/- 1.96*0.1047= 3.168 to 3.580
2b. We would like to know if there is a difference in pain score between week 2 and week 4? Choose the correct test using Figure 2.1. Define the null hypothesis.
2c. What is the assumption of the chosen test and test the assumption. Don't forget to select only AK patients. Start with calculating a new variable to test the assumption.
2d. Perform the chosen test of question 2b and interpret the output.

3a. We would like to know if there is a difference in pain score between week 4 and week 2 between patients with different skin types. Use the Skin_reaction variable as grouping variable. Explore the skin_reaction variable.
3b. Choose the correct test and define the null hypothesis. You need to make a new variable before you perform the test. Do not forget to test the most important model assumption.

4. Among participants with AK above 75 years old, you would like to know if there is a difference in pain score between week 2 and week 4. Which test would you use? Define the null hypothesis and perform the test. Explore the variables before you perform the test.

5. Is there a difference in family history of melanoma between patients with and without SCC? Define the null hypothesis and perform the correct test. Don't forget to turn the filter for the previous questions off. You need to make a new variable using DO IF/ ELSE IF statements before you can perform the test. Examples of this statement can be found in paragraph 1.6c and 3.4.

Chapter 3 Linear Regression

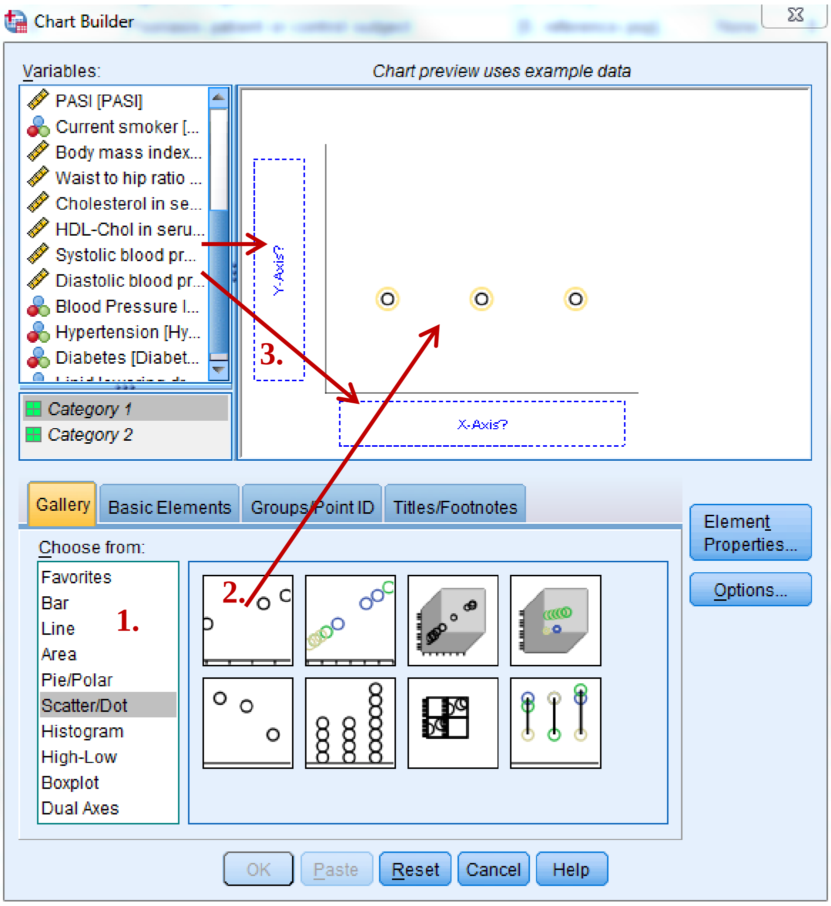
Data: Psoriasis
Psoriasis has been related with a number of comorbidities. Patients with psoriasis may have an increased risk of cardiovascular diseases. For this practical we will use a database of psoriasis patients and control patients of which multiple (sub)clinical measurements of comorbidities have been collected.

3.1 Correlation

Correlation is used to measure the association between two variables.

It can be expected that diastolic and systolic blood pressure are highly correlated. The first step is to visualize the data using a scatterplot. You can use the chart builder.

Graphs > Chart builder



1.
Select the type of graph from the list

2.
Drag the graph to the chart preview

3.
Drag the variables to the axes.

Or use a predefined format:
Graphs > Legacy dialogs > Scatter/dot > Simple scatter

```
GRAPH
 /SCATTERPLOT(BIVAR)=SBP WITH DBP
 /MISSING=LISTWISE.
```

We will calculate Pearson's correlation coefficient, which is used for continuous variables. Pearsons correlation coefficient (r) is a number between -1 and 1.

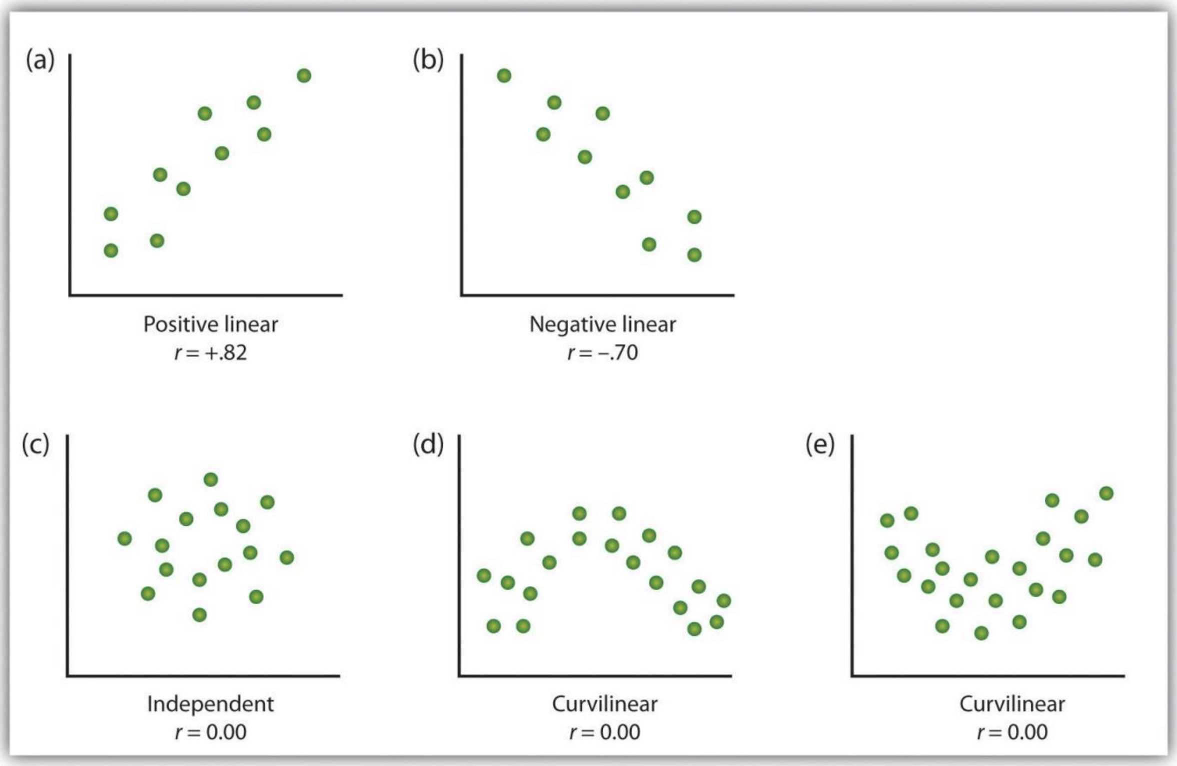


Figure 3.1: Some examples of relationships between two variables as shown in scatter plots. Note that the Pearson correlation coefficient (r) between variables that have curvilinear relationships will likely be close to zero.
Source: Adapted from Stangor, C. (2011). Research methods for the behavioral sciences (4th ed.). Mountain View, CA: Cengage.).

Analyze > *Correlate* > *Bivariate*

The data was plotted first, because it is inappropriate to calculate a Pearsons correlation coefficient when there is no linear relationship. From the output we obtain a pearson's r of 0.607. The Pearson's r^2 indicates how much of the variability in y is explained by x. In this case $0.607^2=0.369$, meaning that 36.9% of the variability in systolic blood pressure is explained by diastolic blood pressure

Spearman rank correlation coefficient is the non-parametrical equivalent and may be calculated when the sample size is very small.

3.2 Simple linear regression

Linear regression is used to predict the outcome of continuous outcome measures based on covariates. Different terminology is used to express the same thing:

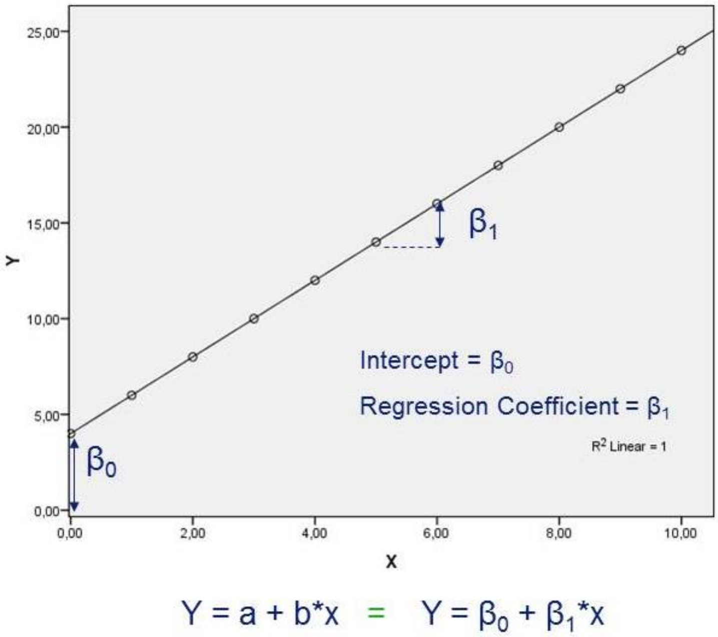
$X_1, X_2, \ldots X_k$ $\longrightarrow$	Y
=	=
'Predictor'	'Outcome'
'Explanatory variables'	'Response'
'Independent variables'	'Dependent variable'

‘Covariates’

Questions, which can be answered using linear regression:

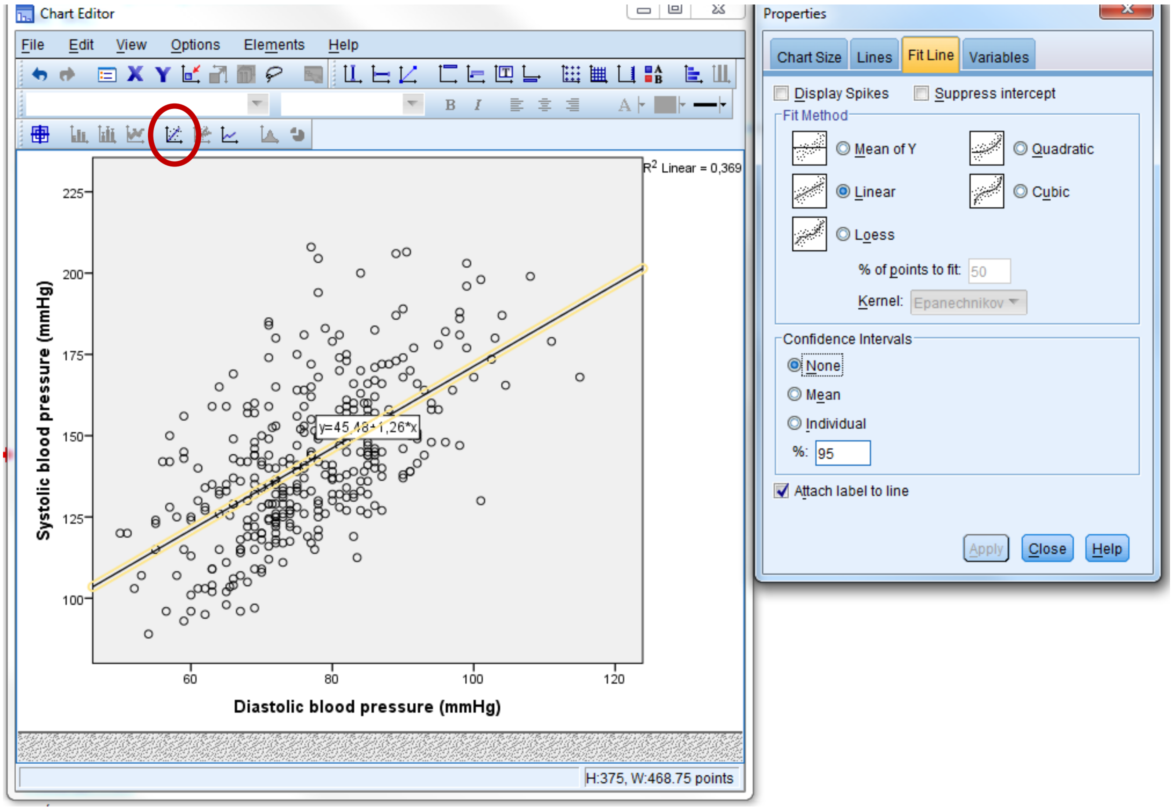
- How can Y be predicted on the basis of the X’s
- Does X_1 have influence on Y, when controlling for the other X’s

We can predict the systolic blood pressure based on the diastolic blood pressure using a simple linear regression model. Recall the following formula from high school mathematics:
$Y = a + b*x \quad = \quad Y = \beta_0 + \beta_1*x$



3.2a a linear regression model for blood pressure

Step 1: Let’s make a similar figure for blood pressure using our data. Double click on the Graph that you have produced in 3.1. The Chart editor is opened. Add a fitted linear line.



☐ Display Spikes ☐ Suppress intercept

Fit Method: ● Linear

Confidence Intervals: ● None

%: 95

☑ Attach label to line

A linear regression line is fitted and the formula is shown.
Y=45.48+1.26*X
Systolic Blood Pressure=45.48+1.26*Diastolic Blood Pressure

45.48 is the intercept (or β_0)
1.26 is the regression coefficient for diastolic blood pressure (or β_1)

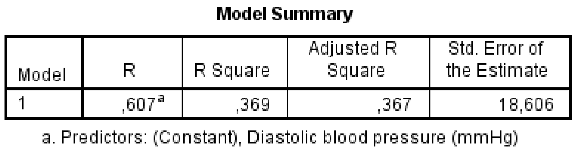
Now we can predict the systolic blood pressure by using diastolic blood pressure. For example: the systolic blood pressure of a person with a diastolic blood pressure of 80 mmHg is on average 45.48+1.26*80=146.28 mmHg.

Step 2: We can obtain the same result by building a linear regression model.

Analyze > Regression > Linear

The dependent variable is systolic blood pressure and the independent variable is diastolic blood pressure. Go to *Statistics* and tick the box for *Confidence Intervals*, *Level 95%* and *Descriptives*.

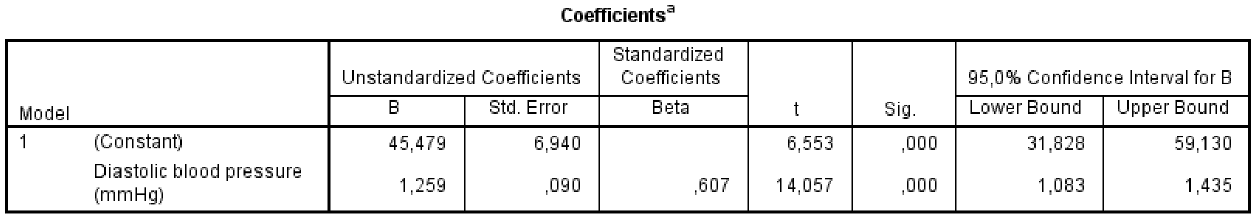
Step 3: Interpret the output. We will focus on only two tables in this assignment.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,607[a]	,369	,367	18,606

a. Predictors: (Constant), Diastolic blood pressure (mmHg)



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	45,479	6,940		6,553	,000	31,828	59,130
	Diastolic blood pressure (mmHg)	1,259	,090	,607	14,057	,000	1,083	1,435

In these tables you can find the r^2, which we calculated before, as well as the regression coefficients. Use the unstandardized coefficients. The interpretation for the β for diastolic blood pressure is: for each unit increase in diastolic blood pressure (1 unit=1 mmHg) the systolic blood pressure increases with 1.259 (95% CI: 1.083-1.435).
The p-value is $1.162*10^{-35}$, which indicates that this is statistically significant. In addition, the 95% CI excludes the value 0. (i.e. 0 indicates no change for each unit increase, thus if the 95% CI includes the value 0 there is no statistical significant difference).

3.3 Model assumptions

The model assumptions of a linear regression are:

1. Independence of observations

All observations should be obtained from different subjects (i.e. patients should not be included twice).

2. Linearity

We assume a linear relation between the outcome and the covariate. Non-linear relationships can be taken into account, which is discussed in paragraph 3.8.

3. Normality (for x=1,x=2, etc.)

This is best illustrated by the figure below. The outcome variable does not have to be normally distributed, but for every value of x we assume a normal distribution. Thus for our example of systolic blood pressure: an overall test of a normal distribution of systolic blood pressure is not required. However we assume, that for a diastolic blood pressure of 80 mmHg the values of systolic blood pressure are normally distributed, and for 81 mmHg diastolic blood pressure, the systolic blood pressure values should be normally distributed, as well as for 82 mmHg, etc.

4. Homoscedasticity (standard deviations are equal, for x=1, x=2, x=3 etc.)

This relates to assumption 3 and is also illustrated by the figure below. Equal standard deviations (SD) for each value of x, result in equal normal distributions for each value of x. This indicates that the variance is equal for each value of x.

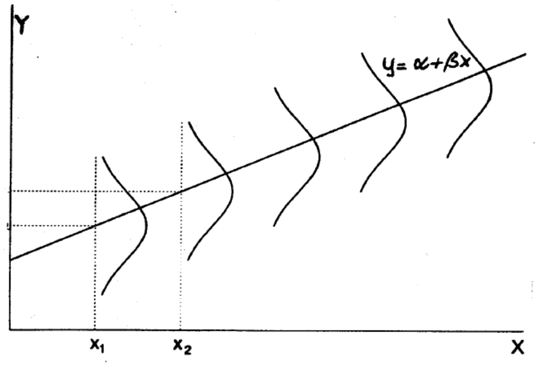


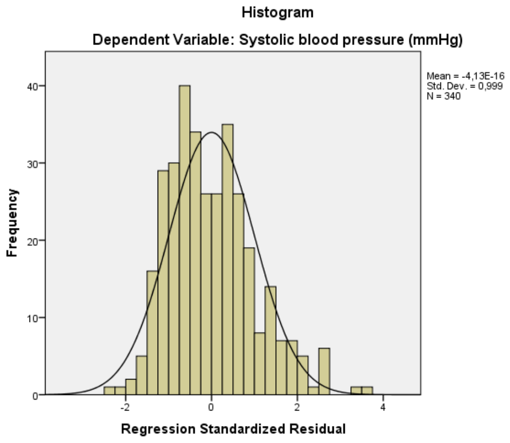
Figure 3.2: a visualization of a linear regression model

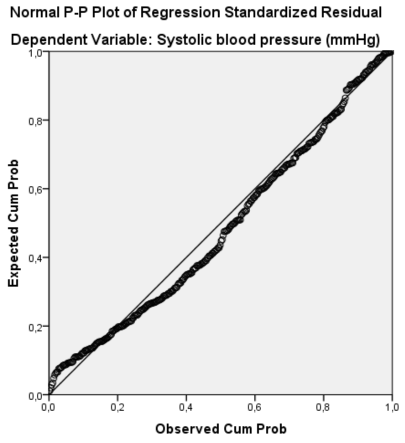
Looking at the model assumptions and figure 3, it may me more clear, that linear regression model may be regarded as an ANOVA (chapter 2.3): each value of x can be regarded as a group and variance of each group should be equal.

The first assumption cannot be statistically tested, but should be included in the design of the study. Assumption 2 can be visualized by using a scatterplot. Further information can be found in paragraph 3.8 . Based on assumption 3, the residuals are expected to be normally distributed. Assumption 3 and 4 (normality and homoscedasticity) can also be tested by saving and plotting the residuals against the predicted values and the covariates:

Step 1: Ask for a histogram of the residuals and a normal probability plot.

Analyze > Regression > Linear > Plots





You obtained a histogram of the residuals with an normal curve and an P-P plot for the standardized residuals. A P-P plot can be interpreted as a Q-Q plot which was described in Chapter 1.4. A deviation from the straight line implies a deviation from normality. The plot indicates that there may be a small deviation from normality.

Step 2: save the predicted values and the unstandardized residuals.

Analyze > Regression > Linear > Save

Go to the ‘Save’ menu of the linear regression model and tick the box unstandardized predicted values and unstandardized residuals. In the variable view and data view you will find two new variables: PRE_1 and RES_1.

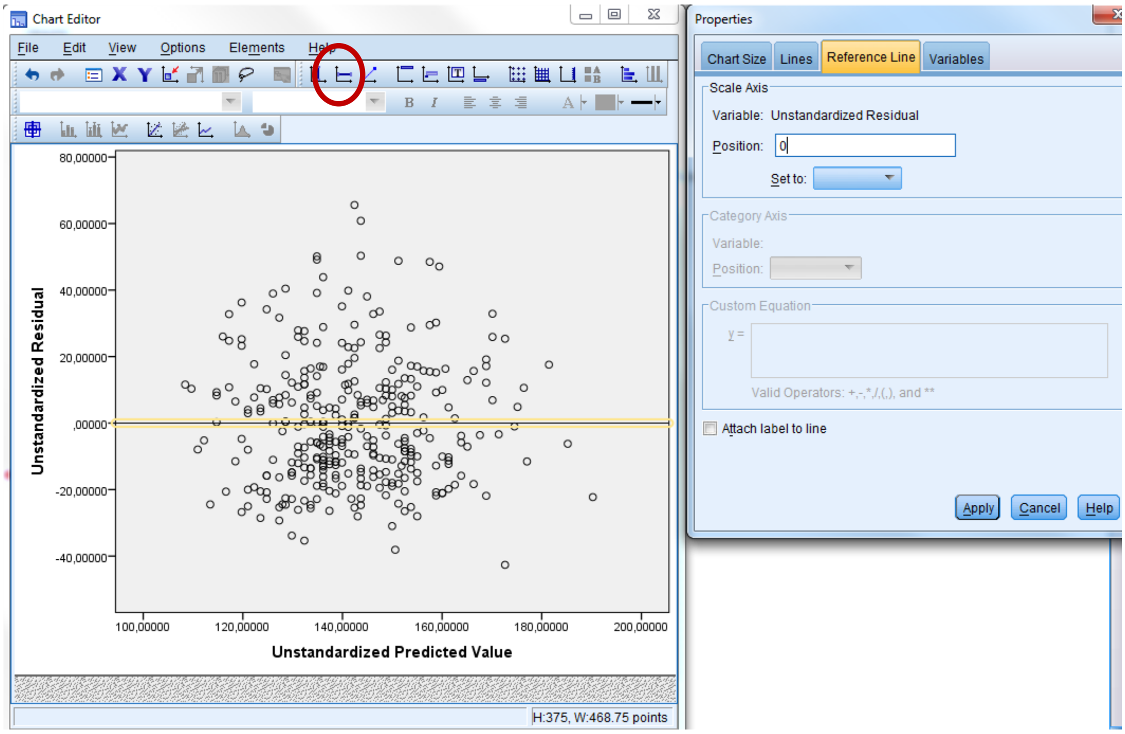
Step 3: plot the diastolic blood pressure against the residuals and the predicted values against the residuals.

```
GRAPH
 /SCATTERPLOT(BIVAR)=DBP WITH RES_1
 /MISSING=LISTWISE.
```

You can easily change the syntax to make the second plot:

```
GRAPH
 /SCATTERPLOT(BIVAR)=PRE_1 WITH RES_1
 /MISSING=LISTWISE.
```

Step 4: Interpret the graph
The residuals should be spread equal around 0 for each value of x (each value of diastolic blood pressure) and each predicted value of systolic blood pressure. You can add a reference line at 0 by opening the chart editor by double clicking on the plot.



Variable: Unstandardized Residual
Position: 0

Both graphs indicate that the model assumption of homoscedasticity is fulfilled. A deviation from normality and homoscedasticity may indicate that a data transformation of the outcome variable (e.g. log transformation) is needed. A common problem is an increase of the variance for larger values of the covariates or predicted values, which is shown in figure 3.3.

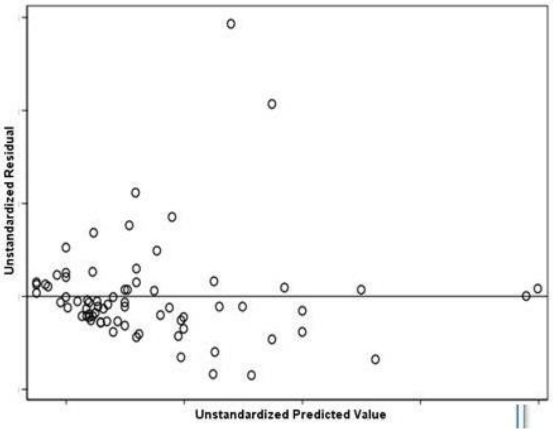


Figure 3.3: indication of heteroscedasticity

3.4 Categorical variable coding

Categorical variables can also be taken into account into a regression model.

Step 1: make a new categorical variable
The following syntax is useful to categorize continuous variables, such as BMI.

```
DO IF MISSING(BMI).
COMPUTE BMI_cat=999999.
ELSE IF BMI <18.5.
COMPUTE BMI_cat=2.
ELSE IF BMI>=18.5 AND BMI<=25.
COMPUTE BMI_cat=1.
ELSE IF BMI>25.
COMPUTE BMI_cat=3.
END IF.
EXECUTE.VARIABLE LABELS BMI_cat 'category of BMI'.
VALUE LABELS BMI_cat 1 'normal weight' 2 'underweight'  3 'overweight'.
```

Note that normal weight is coded as the first category, because SPSS can only take the first or the last category into account as reference category.

Categorizing continuous variables leads to a loss of information. It is generally better to include them as continuous variables, but sometimes or in the final regression model clinically relevant categories may be easier to interpret.

Step 2: create dummy variables for a categorical variable
SPSS uses dummy variable coding with a reference category to include categorical variables in the regression model. The number of dummy variables is always the number of categories minus 1. For 3 categories, 2 dummy variables are sufficient:

BMI_cat	BMI_dummy_1	BMI_dummy_2
1 (normal weight)	0	0
2 (underweight)	1	0
3 (overweight)	0	1

The dummy variables allow the bloodpressure to vary between the categories. In model A, BMI categories (1,2,3) are included in the model as a linear covariate, assuming that the difference in systolic blood pressure between category 2 and 3 is equal to the difference in systolic blood pressure between category 1 and 2 (Model A, Figure 3.4A).
Dummy variable coding allows for differences between categories (model B, Figure 3.4B.

Model A: Systolic blood pressure=β_0+β_1*BMI_cat
Model B: Systolic blood pressure=β_0+β_1*BMI_dummy_1+β_2*BMI_dummy_2

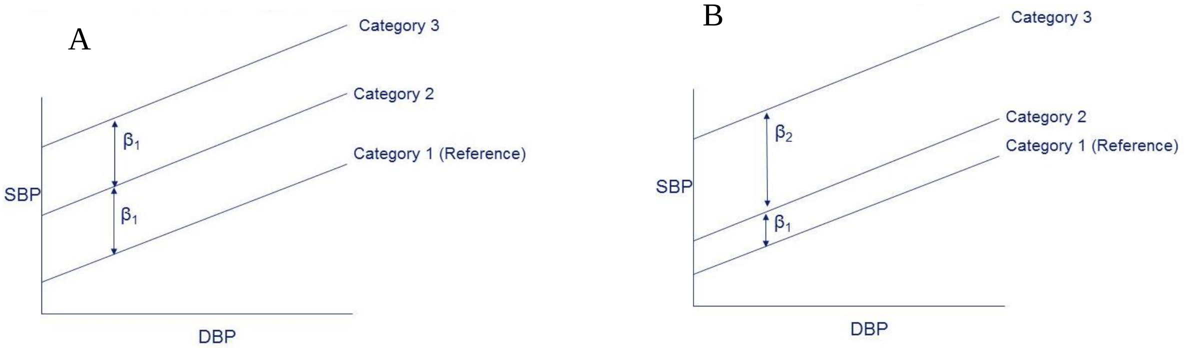
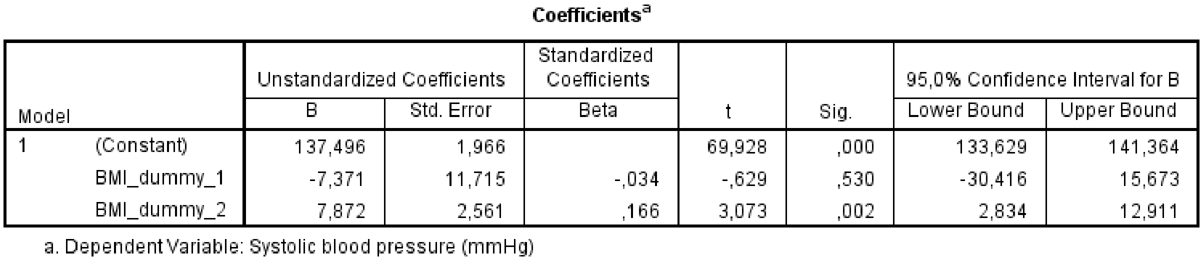


Figure 3.4: Dummy variable coding in model B.

Step 3: Perform the regression analysis with the categorical variable
Perform to regression analyses.

Analyze > Regression > Linear.

Include both BMI_dummy_1 and BMI_dummy_2 as independent variables. This will lead to the following output table:



Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	137,496	1,966		69,928	,000	133,629	141,364
	BMI_dummy_1	-7,371	11,715	-,034	-,629	,530	-30,416	15,673
	BMI_dummy_2	7,872	2,561	,166	3,073	,002	2,834	12,911

a. Dependent Variable: Systolic blood pressure (mmHg)

Step 4: Interpretation of the output:
The output tells you that the mean systolic blood pressure of underweight participants (level 2) is 7.4 mmHg (95% CI: -30 to 16) lower compared to normal weight subjects. The p-value of the regression coefficient is 0.530, meaning that this the difference is not statistically significant. This was to be expected as there were only 4 participants underweight.
The participants who are overweight (level 3) have a mean systolic blood pressure which is 7.8 mmHg higher compared to normal weight participants (level 1) with an 95% between 2.8 and 12.9 and a p-value of 0.002. This difference is statistically significant.

3.5 Multivariable linear regression.

There are two main reasons to perform a multivariable regression model.

1. Predict the outcome based on covariates: confounding is not an issue
2. Investigate the association between exposure and outcome, adjusted for possible confounders.

In this paragraph we will focus on the first option. Confounding will be discussed in paragraph 3.6.

The systolic blood pressure may also be dependent on other variables, such as age. Both age and diastolic blood pressure can be used to predict systolic blood pressure. Note that the following statistical terminology is used:

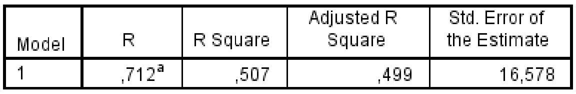
Univariate: one outcome measure
Multivariate: multiple outcome measures
Univariable: one exploratory variable
Multivariable: multiple exploratory variables

3.4 What is the mean systolic blood pressure of a 60-year old non-smoking woman with a BMI of 19 and a diastolic blood pressure of 85 mmHg?

Step 1:
To answer this question we need to include age, current smoking, diastolic blood pressure in one linear regression model.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT SBP
 /METHOD=ENTER sex age BMI current_smoker DBP.
```

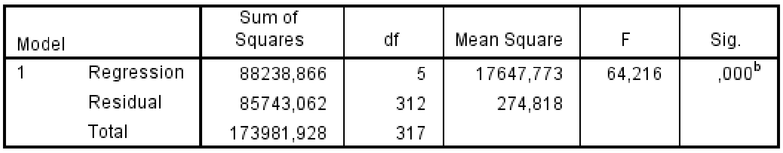
Model Summary



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,712[a]	,507	,499	16,578

a. Predictors: (Constant), Diastolic blood pressure (mmHg), Age, Body mass index (kg/m2), Current smoker, sex

ANOVA[a]

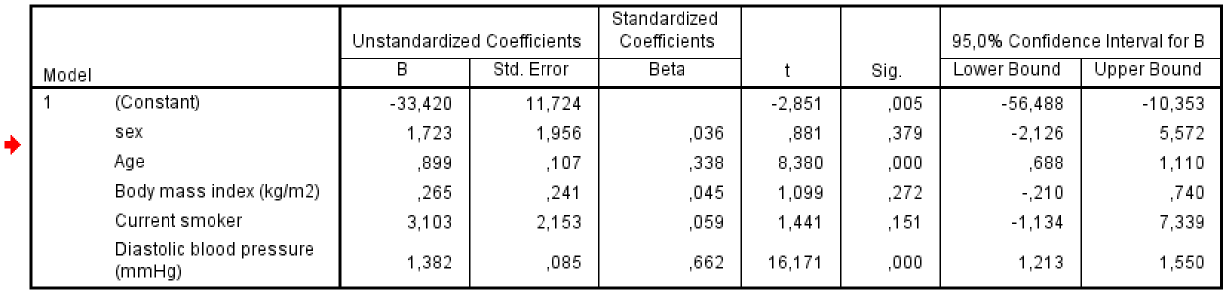


Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	88238,866	5	17647,773	64,216	,000[b]
	Residual	85743,062	312	274,818		
	Total	173981,928	317			

a. Dependent Variable: Systolic blood pressure (mmHg)

b. Predictors: (Constant), Diastolic blood pressure (mmHg), Age, Body mass index (kg/m2), Current smoker, sex

Coefficients[a]



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-33,420	11,724		-2,851	,005	-56,488	-10,353
	sex	1,723	1,956	,036	,881	,379	-2,126	5,572
	Age	,899	,107	,338	8,380	,000	,688	1,110
	Body mass index (kg/m2)	,265	,241	,045	1,099	,272	-,210	,740
	Current smoker	3,103	2,153	,059	1,441	,151	-1,134	7,339
	Diastolic blood pressure (mmHg)	1,382	,085	,662	16,171	,000	1,213	1,550

a. Dependent Variable: Systolic blood pressure (mmHg)

Step 2: Interpret the output

Note: In the descriptive table (not shown above) you can see that only 318 participants are included in the analysis. SPSS performs a complete case analysis, thus participants with any missing value are excluded. There are statistical methods to deal with missing values (e.g. multiple imputation techniques), but that is beyond the scope of this practical.

<u>Model summary</u>

The model summary provides goodness-of-fit measures of the regression model.
R square: equivalent to the simple R square as measured by Pearsons correlation coefficient. The simple R square tends to overestimate the explained variance when there is more than one covariate in the model.
Adjusted R square: R square adjusted for the number of covariates in the model. Use this measure of goodness of fit for multivariable regression models.

Thus, 49.9% of the variance in systolic blood pressure is explained by all variables in the model.

<u>ANOVA table</u>

The ANOVA table of a linear regression model test the overall regression. The null hypothesis of the ANOVA table is:
H_0: all regression coefficients (all β's, except the intercept) are 0.
If we look at the ANOVA table, we can see that the p-value is <0.001 and that the null hypothesis is rejected.

Coefficients

In this table you will find the regression coefficients and the p-values shown for each regression coefficient (β) separately.

The interpretation of the β for diastolic blood pressure is slightly different compared to a univariable model with diastolic blood pressure only.

Interpretation of the β for diastolic blood pressure in a univariable model:
The increase in systolic blood pressure per 1 mmHg increase in diastolic blood pressure.

Interpretation of the β for diastolic blood pressure in a multivariable model:
The increase in systolic blood pressure per 1 mmHg increase in diastolic blood pressure, while holding the value of all other variables in the model constant.

Step 3: Predict the systolic blood pressure based on the model.

The statistical model is:
$SBP = \beta_0 + \beta_1 * sex + \beta_2 * age + \beta_3 * BMI + \beta_4 * current\ smoker + \beta_5 * DBP$
$SBP = -33.420 + 1.723 * sex + 0.899 * age + 0.265 * BMI + 3.103 * current\ smoker + 1.382 * DBP$

The mean systolic blood pressure of a 60-year old non-smoking woman with a BMI of 19 and a diastolic blood pressure of 85 mmHg is:
$-33.420+1.723*1+0.899*60+0.265*19+3.103*0+1.382*85=145$ mmHg

You can repeat the calculation using the estimates of the 95% CI of the β's to obtain the 95% CI for the prediction.

3.6* Confounding

A confounder is a covariate which explains (part of) the association between exposure and outcome, but is not part of the causal pathway. Confounding may be a problem when answering the following research question:

3.6 Is psoriasis associated with a larger intima media thickness (as a subclinical measure of atherosclerosis)?

We suspect that the relation between psoriasis and intima media thickness may be influenced by other variables, such as smoking, which increases risk for psoriasis and is also associated with atherosclerosis. Confounders may be *a priori* selected based on literature or clinical expertise. Confounders may also be selected based on the influence on the association. Including all variables in the multivariable model which are associated with the outcome in the univariable analysis is not necessary, because not all variables will explain (part of) the association between exposure and outcome and may only be related to the outcome.

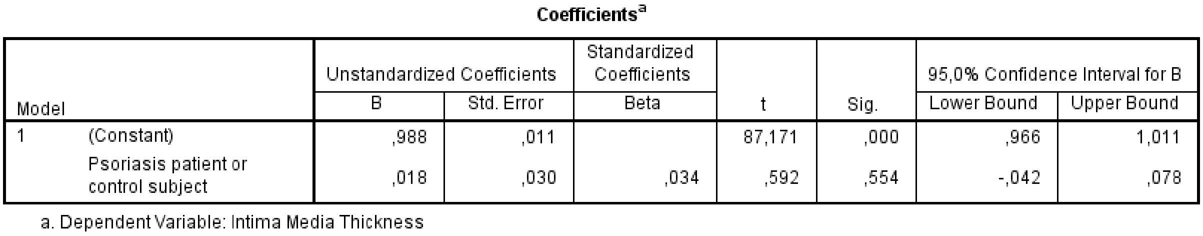
3.6a A priori selection of confounders based on literature

From literature may be known that smoking, BMI, hypertension and serum cholesterol may influence both the risk on psoriasis as well as the risk on atherosclerosis. Therefore can be decided to include all these potential confounders in the multivariable model.

Step 1: Select all patients who do not have missing values in any of the selected variables to prevent analyses of a different samples for different analyses.

Step 2: Perform a univariable model, an age and sex adjusted model and the full multivariable model:

Univariable model:

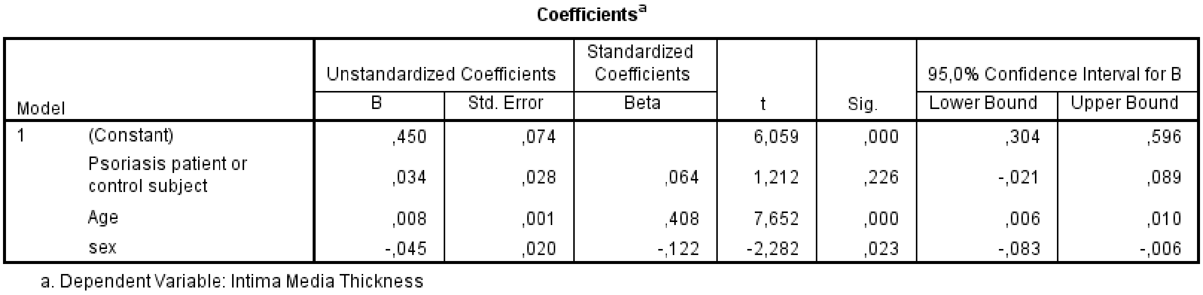


Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,988	,011		87,171	,000	,966	1,011
	Psoriasis patient or control subject	,018	,030	,034	,592	,554	-,042	,078

a. Dependent Variable: Intima Media Thickness

Age and sex adjusted model:

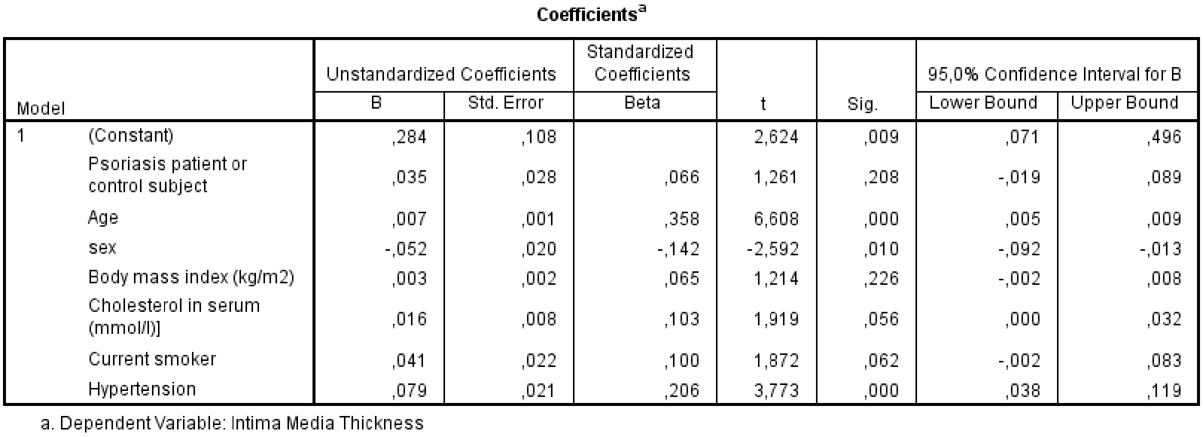


Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,450	,074		6,059	,000	,304	,596
	Psoriasis patient or control subject	,034	,028	,064	1,212	,226	-,021	,089
	Age	,008	,001	,408	7,652	,000	,006	,010
	sex	-,045	,020	-,122	-2,282	,023	-,083	-,006

a. Dependent Variable: Intima Media Thickness

Full multivariable model:



Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,284	,108		2,624	,009	,071	,496
	Psoriasis patient or control subject	,035	,028	,066	1,261	,208	-,019	,089
	Age	,007	,001	,358	6,608	,000	,005	,009
	sex	-,052	,020	-,142	-2,592	,010	-,092	-,013
	Body mass index (kg/m2)	,003	,002	,065	1,214	,226	-,002	,008
	Cholesterol in serum (mmol/l)]	,016	,008	,103	1,919	,056	,000	,032
	Current smoker	,041	,022	,100	1,872	,062	-,002	,083
	Hypertension	,079	,021	,206	3,773	,000	,038	,119

a. Dependent Variable: Intima Media Thickness

3.6b Confounder selection based on influence of risk estimate

Another approach to select confounders could be to include all potential confounders which influence the age and sex adjusted risk estimate by 10%. The β for psoriasis in the age and sex adjusted model was 0.034. An influence of 10% would be 0.034 +/- 0.0034 = 0.0306 and 0.0374. Inclusion of a potential confounder in the age and sex adjusted model which leads to an influence on the regression coefficient of psoriasis outside these boundaries is considered a confounder and included in the multivariable model.

Step 1:
Change the following syntax by changing BMI in the last line into Cholesterol, current_smoker and Hypertension.

```
REGRESSION
   /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT IMT
  /METHOD=ENTER psocase age sex BMI.
```

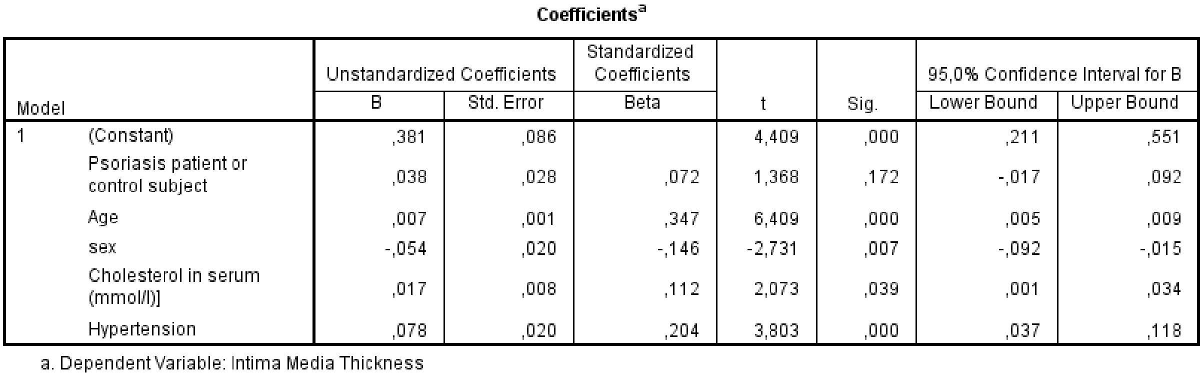
Step 2:
Examine the β for psoriasis for inclusion of each potential confounder.

Potential confounder	β for psoriasis (increase in IMT for psoriasis patients compared to controls)
BMI	0.031
Cholesterol	0.042*
Smoking	0.034
Hypertension	0.029*

* indicates that the β is outside the aforementioned 10% boundaries.

From these analyses we can conclude that only serum cholesterol and presence of hypertension confound the relation between psoriasis and intima media thickness

Step 3: Perform the multivariable model based on the findings in step 2


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,381	,086		4,409	,000	,211	,551
	Psoriasis patient or control subject	,038	,028	,072	1,368	,172	-,017	,092
	Age	,007	,001	,347	6,409	,000	,005	,009
	sex	-,054	,020	-,146	-2,731	,007	-,092	-,015
	Cholesterol in serum (mmol/l)]	,017	,008	,112	2,073	,039	,001	,034
	Hypertension	,078	,020	,204	3,803	,000	,037	,118

a. Dependent Variable: Intima Media Thickness

The conclusion for both methods is that psoriasis is not associated with a significantly increase in intima media thickness adjusted for potential confounders.

3.7* Interaction

In the example above we assumed that the increase in intima media thickness for each year of age is equal for control participants and patients with psoriasis. However, there may be effect modification by psoriasis, meaning that the increase in intima media thickness per year of age may be different between participants with and without psoriasis. This is called interaction and is illustrated by the figure below:

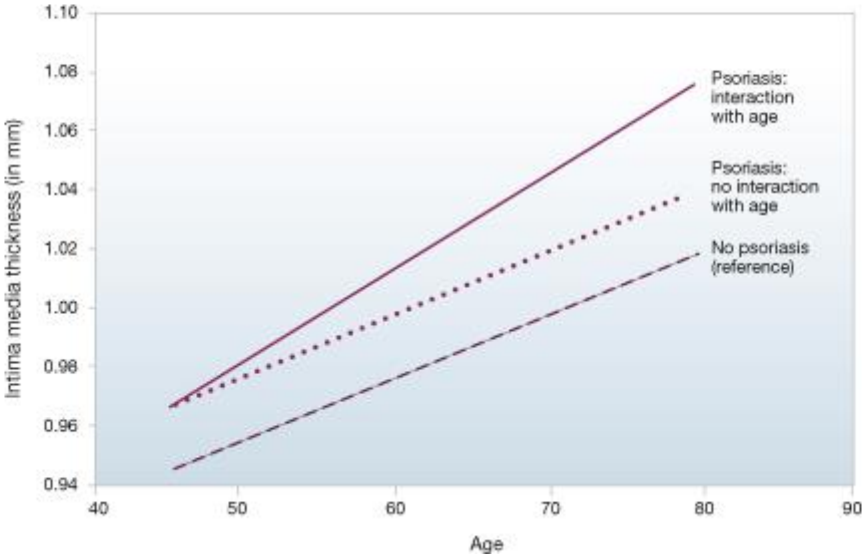


Figure 3.7 An example of interaction in linear regression between psoriasis and age for the difference in carotid intima media thickness. If no interaction is present, the difference in intima media thickness will be equal for all ages. In the case of statistical interaction, the difference in intima media thickness depends on age. Source: Wakkee et al. JID 2014.

There is a difference between biological interaction and statistical interaction.

Biological interaction = effect modification
Statistical interaction = effect measure modification

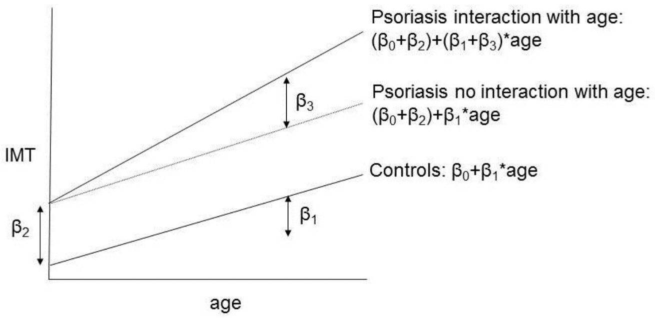
We can test the statistical interaction by including an interaction term in the regression model. However, this does not tell you if there is a biological plausible mechanism for the effect measure modification of the effect of age by psoriasis. Statistical significance is not the same as clinical relevance. Therefore you should specify the relevant interactions that you would like to test for in advance.

Statistical interaction can be tested by using an interaction term in the regression model, which is illustrated in box and figure

Linear regression model without interaction:
IMT = $\beta_0+\beta_1$*age+β_2*psoriasis

Linear regression model with interaction:
IMT = $\beta_0+\beta_1$*age+β_2*psoriasis+β_3*age*psoriasis

To test for interaction it should be tested if β_3 is equal to 0.


Psoriasis interaction with age:
$(\beta_0+\beta_2)+(\beta_1+\beta_3)*age$
β_3
Psoriasis no interaction with age:
$(\beta_0+\beta_2)+\beta_1*age$
IMT
Controls: $\beta_0+\beta_1*age$
β_1
β_2
age

Graphical representation of a regression model with and without interaction between age and psoriasis

3.7 Is there statistical interaction between age and psoriasis in a regression model for IMT?

Step 1: define the null hypothesis;

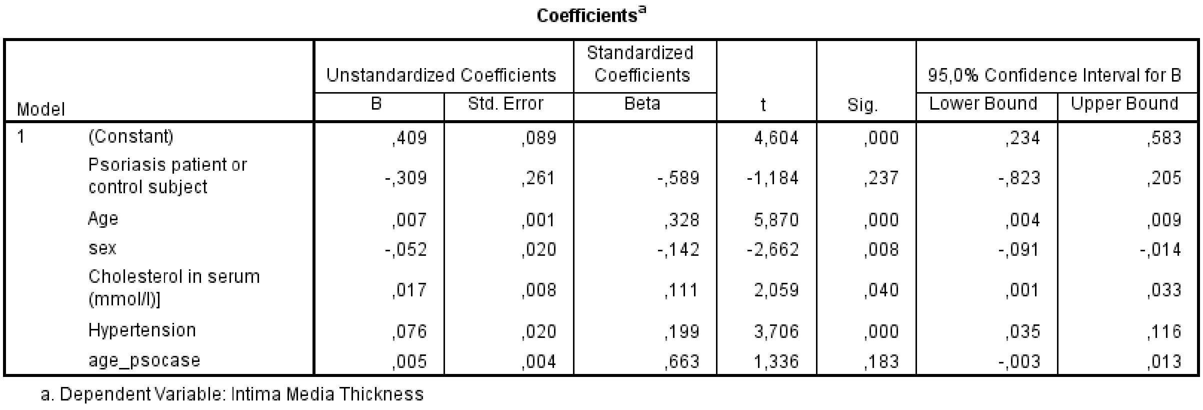
H_0: There is no statistical interaction between age and psoriasis. Thus the β (regression coefficient) of the interaction term is 0.

Step 2: Create a new variable by multiplying age with psoriasis

```
COMPUTE age_psocase=age*psocase.
EXECUTE.
```

Step 3: perform the linear regression model and include the interaction term for age and psoriasis in the final model, which includes all possible confounders.

Step 4: Interpret the output



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	,409	,089		4,604	,000	,234	,583
	Psoriasis patient or control subject	-,309	,261	-,589	-1,184	,237	-,823	,205
	Age	,007	,001	,328	5,870	,000	,004	,009
	sex	-,052	,020	-,142	-2,662	,008	-,091	-,014
	Cholesterol in serum (mmol/l)]	,017	,008	,111	2,059	,040	,001	,033
	Hypertension	,076	,020	,199	3,706	,000	,035	,116
	age_psocase	,005	,004	,663	1,336	,183	-,003	,013

a. Dependent Variable: Intima Media Thickness

The β for the interaction term is 0.005, which means that for each year of increase in age, there is an extra increase of 0.005 mm in IMT on top of the 0.007 mm increase for psoriasis patients compared to controls. However, the p-value of this interaction term is 0.183, which indicates that the null hypothesis should not be rejected and there is no statistical interaction between age and psoriasis. The interaction term should not be included in the final multivariable model.

Don’t forget to put the filter off before the next paragraph.

3.8* Non-linear relationships

From Scatterplots it may be obvious that the predictors are not linearly related to the outcome. This will lead to a non-significant β for a linear predictor. There are various ways to deal with non-linear relationships in regression analysis.

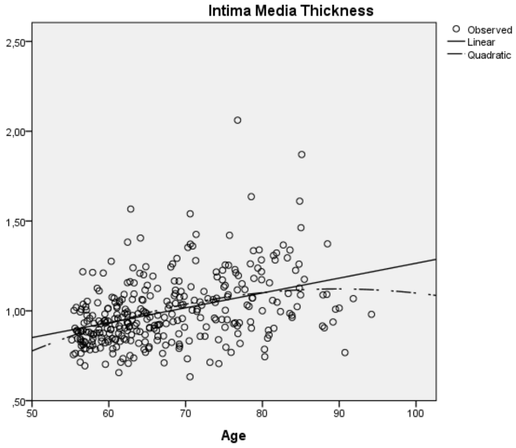
Categorize
This is an easy way of representing non-linear relationships. Clinically relevant categories may be used in the regression model. A disadvantage is that many β’s (and thus degrees of freedom) are spend, cut-off values may be arbitrary and thereby valuable information is lost.

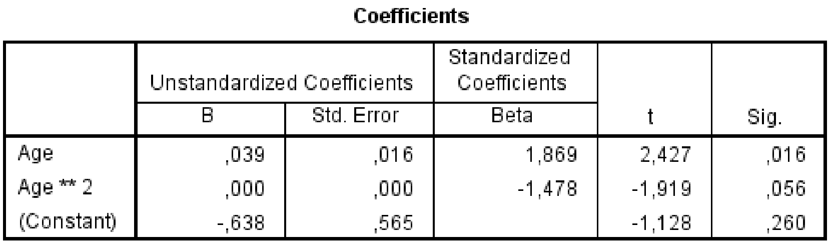
Polynomials
Another easy way of testing for non-linearity may be to include polynomials. The first step is to include a term for X^2 in the regression model. If X is linearly related to the outcome the regression coefficient is 0. However, not all relationships have a parabolic shape. Higher powers of X may be included in the model, but the disadvantage of polynomials is that they may behave ‘wobbly’ in the tails.

Use the SPSS curve estimation to test for a possible quadratic relationship:

Analyze > Regression > Curve estimation

Ask for a linear and a quadratic curve and the plots.





Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Age	,039	,016	1,869	2,427	,016
Age ** 2	,000	,000	-1,478	-1,919	,056
(Constant)	-,638	,565		-1,128	,260

From the table can be obtained that there is no evidence of a quadratic relationship between age and IMT, as the regression coefficient for age^2 is not significant.

Fractional Polynomials

Fractional polynomials have with different powers compared to regular polynomials, e.g.:
$Y=\beta_0 + \beta_1*X^{-1} +\beta_2*X^2$
Fractional polynomials are therefore more flexible, but models may be difficult to interpret. In STATA fracpoly is useful command to select the best degree and powers. In SPSS it is not possible to automatically select the best degree and powers.

Splines

The use of spline functions is a very flexible way of regression modelling with efficient use of the degrees of freedom (meaning that not many β's are needed to model complicated non-linear relationships). There are many sorts of splines; simple linear splines, cubic spline, restricted cubic splines, penalized cubic splines. SPSS can take spline functions into account, but modelling spline functions is also relatively easy in R.

Note: When using logistic regression (Chapter 4), linearity is assumed between the predictors and the logit. A Cox proportional hazards regression (Chapter 6) assumes linearity with the LN(hazard). The same principles and methods of non-linear relationships apply to those types of regression.

Assignment 3-1 Univariable and multivariable linear regression

Data: blood pressure
We want to study with multiple regression the joint relationship of age, body weight and pulse rate with the diastolic blood pressure. The exercise uses the data from BLDPRES.SAV. (Open file 'BLDPRES.SAV'.) Make sure that you Paste each action into your syntax first before you Run it.

1. Obtain the correlation coefficients between age, body weight, pulse rate and diastolic blood pressure and examine these, to learn something about any potential correlation between the variables.

2a. To visualize the data and to see if it is appropriate to calculate a Pearson correlation coefficient, we look at the simple relationships between each of the predictor variables and the diastolic blood pressure. Make a scatter plot with DIAS as Y-variable and AGE as the X-variable. Repeat this for body weight and pulse frequency
2c. Provide the regression coefficients and interpret the r square for age, body weight and pulse frequency using the plot.

3. Fit three univariable models for age, body weight and pulse frequency as covariates and diastolic blood pressure as dependent variable.
Give the regression coefficients, 95% CI and their significance.
What is the interpretation of these regression coefficients?

4. Fit a multivariable model for diastolic blood pressure including age weight and pulse as independent variables. Ask for 95% confidence intervals for the regression coefficients (Statistics) and a Histogram (Plot) and Normal Probability Plot (Plot) and save the predicted values and the unstandardized residuals (Save).

4a. Consider the ANOVA table accompanying the multiple regression analysis. What is the hypothesis that is tested with the F-test and what is the conclusion?
4b. What is the percentage variability in diastolic blood pressure explained by age, weight and pulse together?
4c. Examine the histogram and normal probability plot of the residuals and judge whether the assumption of Normality is reasonably fulfilled.
4d. Plot the fitted values against the residuals and the covariates against the residuals and judge whether the assumption of homoscedasticity is fulfilled.
4e. Look at the estimated regression coefficients. What is their interpretation? Notice that the regression coefficients do not differ much from the simple regression coefficients you have found in part b of this exercise. Can you explain that?

Assignment 3-2 Confounding, Interaction, non-linear relationships

Data: phlebology
Twenty-five patients with varicose veins were examined. We are interested in the relation between severe varicose veins (those with a healed ulcer or an active venous ulcer) and patient reported outcomes. Patients scored their overall health with a number between 0 (very bad health) and 100 (excellent health).

1. The Clinical Etiologic Anatomic Pathophysiologic (CEAP) score indicates the severity. Patients with C5 and C6 were combined into 1 category due to the low patient numbers. Examine the relation between CEAP and overall health score. What is the difference in health score between patients with and without venous ulcers? Is this difference statistically significant?

Other variables may confound the relation between venous ulcer and health score.
2a. Recall the definition of a confounder
2b. Based on the answer on question 2a, think of possible confounders (regardless of statistics).
2c. Perform a model with these a priori determined confounders

It is also possible to statistically test if variables distort the relation between CEAP and health score.
3a. Start with an age and sex adjusted model. What is the point estimate of the age and sex adjusted CEAP score and the interpretation of this estimate?
3b. Use the 10% rule to determine if the diameter of the great saphenous vein (GSV_diameter) and weight confound the age and sex adjusted CEAP estimate.

4. Is the relation between weight and overall health score different between those with and without venous ulcers? Use the final model of 3b.

5a. Examine the linearity assumption using scatterplots for GSV_diameter , weight and heart rate .
5b. Use the curve estimation procedure to find the correct curve for the variable without a linear relationship. Which model would you test?
5c. What is the difference in R^2 between the linear and the non-linear relationship?
5d. What is the univariable model based on the curve-estimation?

Chapter 4 Logistic Regression

Data: actinic keratosis (AK)
More than 800 people were screened on the presence of actinic keratosis. The number of actinic keratosis was registered as well as potential risk factors for the development of actinic keratosis.

Logistic regression is used when the outcome is binary, for example:
0 = no disease
1= disease
These are often case control studies. Logistic regression is used to predict the effect of the covariates on the probability of the outcome.

The probability to have the disease is restricted to the values 0 and 1. To transform these values, the logit link is used (see box below). Logistic regression is used to estimate the odds ratio.

The logit link

π = probability on the disease
odds = $\frac{\pi}{1-\pi}$
π = $\frac{odds}{odds+1}$
logit = ln(odds)

General linear model → linear regression, outcome is continuous
Generalized linear model → e.g. logistic regression, ordinal regression, Poisson regression. The outcome variable can be binary, categorical or count data. A link function for the outcome is used to linearly relate the predictors to this link function.

The logistic regression model:

$$\ln(odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

Recall the formula of a linear regression model en note the similarities:

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

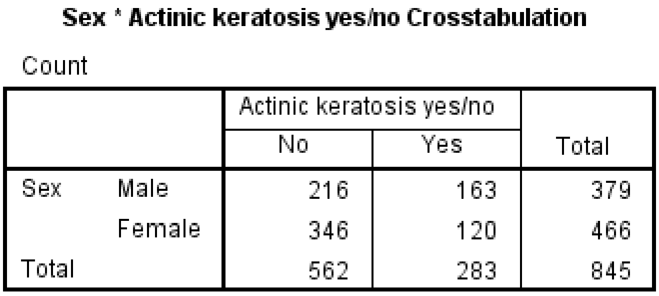
4.1 Odds Ratio

The odds ratio is an approximation of the risk ratio and can be calculated from a simple 2x2 table.

4.1 What is the risk ratio and the odds ratio of females vs males for the development of actinic keratosis?

Analyze > Descriptive Statistics > Crosstabs

Tick the box ‘Risk’ in the ‘Statistics’ menu.



Sex * Actinic keratosis yes/no Crosstabulation

Count

		Actinic keratosis yes/no		Total
		No	Yes	
Sex	Male	216	163	379
	Female	346	120	466
Total		562	283	845

Risk of AK in males=$\frac{163}{379} = 0.43$

Risk of AK in females=$\frac{120}{466} = 0.25$

Risk Ratio $_{\text{females vs males}}$=$\frac{120/466}{163/379} = 0.60$

Odds of AK in males=$\frac{163}{216} = 0.75$

Odds of AK in females=$\frac{120}{346} = 0.35$

Odds Ratio$_{\text{females vs males}}$=$\frac{120/346}{163/216} = 0.46$

The odds ratio is not exactly the same as the risk ratio. The approximation becomes better if the disease is rare, which is not the case for AK. Try to calculate the risk ratio and the odds ratio of females vs males for the following 2x2 table:

		Psoriatic arthritis		Total
		No	Yes	
Sex	Male	370	9	379
	Female	452	14	466
Total		822	23	845

How to calculate a 95% of the OR?

		Disease		Total
		No	Yes	
Exposure	No	a	b	a+b
	Yes	c	d	c+d
Total		a+c	b+d	a+b+c+d

se ln(OR) $= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

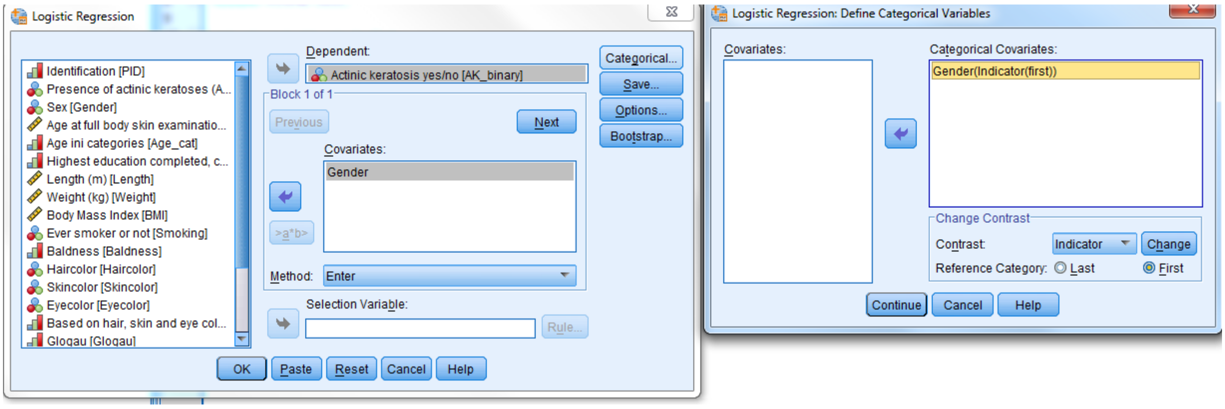
95% CI of OR $= e^{\ln(OR) \pm 1.96 * se(\ln(OR))}$

4.2 Univariable Logistic Regression

We can obtain the same result using logistic regression

Analyze > Regression > Binary Logistic

AK_binary is the dependent variable and sex is the covariate. Sex is a categorical covariate for which the reference category should be specified. Go the ‘Categorical’. Move the covariate to the ‘Categorical Covariates’ box. Use the ‘Change Contrast’ box and choose ‘Indicator’ as contrast and ‘First’ as reference category. Don’t forget to click on ‘Change’ as SPSS takes the last category as the reference by default.



Dependent: Actinic keratosis yes/no [AK_binary]

Covariates: Gender

Method: Enter

Categorical Covariates: Gender(Indicator(first))

Contrast: Indicator

Reference Category: ○ Last ● First

Categorical Variables Codings

		Frequency	Parameter coding (1)
Sex	Male	379	,000
	Female	466	1,000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Gender(1)	-,777	,148	27,487	1	,000	,460
	Constant	-,282	,104	7,363	1	,007	,755

a. Variable(s) entered on step 1: Gender.

Block 0: only includes the intercept. Block:1 includes the covariates. In the ‘Categorical Variables Codings’ table you can see, that the reference category is the category which has the value 0. The OR is the exponent of the regression coefficient (Exp(B)), which is 0.460 as we have calculated before.

Why is Exp(B) the OR?

$$\ln(odds) = \beta_0 + \beta_1 sex$$
$$\ln(odds)_{males} = \beta_0 + \beta_1 * 0 = \beta_0$$
$$\ln(odds)_{females} = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$$
$$odds_{males} = e^{\beta_0}$$
$$odds_{females} = e^{\beta_0 + \beta_1}$$
$$OR_{females\ vs.males} = \frac{odds_{females}}{odds_{males}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

4.3 Multivariable Logistic Regression

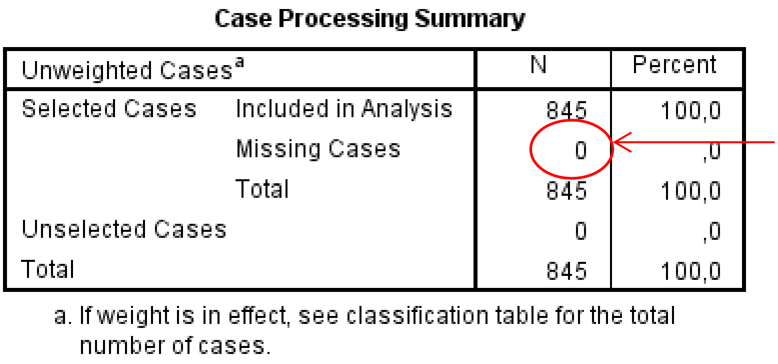
4.3a Are baldness, skin color, sex and age independent predictors of risk on actinic keratosis?

Analyze > Regression > Binary Logistic

Step 1: To answer this question we need to fit a model which includes all covariates. Ask for 95% CI of Exp(B) at ‘options’ and tick the box ‘CI for Exp(B) 95%’. Specify the categorical variables. Use ‘no baldness’ and ‘fair/white’ as reference categories. These are both the last categories.

```
LOGISTIC REGRESSION VARIABLES AK_binary
  /METHOD=ENTER Gender Age Baldness Skincolor
  /CONTRAST (Gender)=Indicator(1)
  /CONTRAST (Baldness)=Indicator
  /CONTRAST (Skincolor)=Indicator
  /PRINT=CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Step 2: Interpret the output:

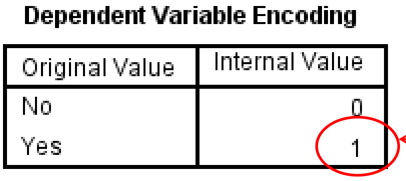


Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	845	100,0
	Missing Cases	0	,0
	Total	845	100,0
Unselected Cases		0	,0
Total		845	100,0

a. If weight is in effect, see classification table for the total number of cases.

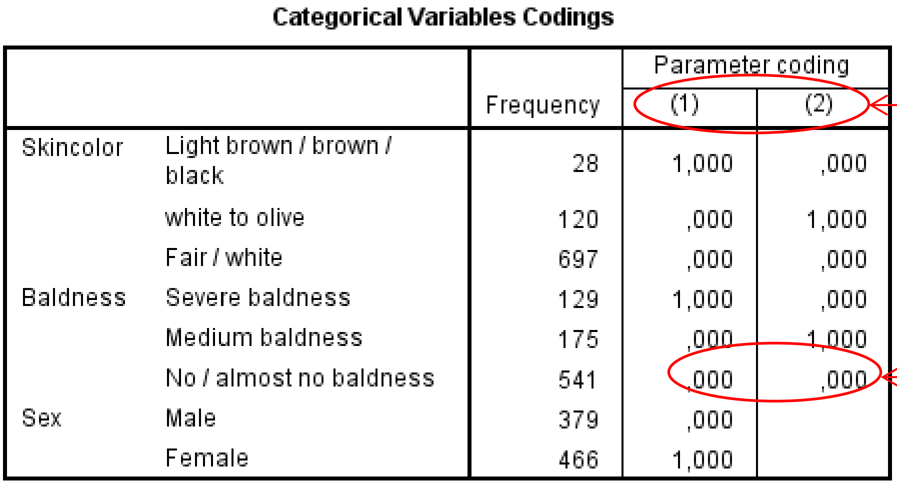
There are no patients with missing values in any of the included variables



Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Coding of AK_binary (No/Yes)
The probability on AK=yes is modeled

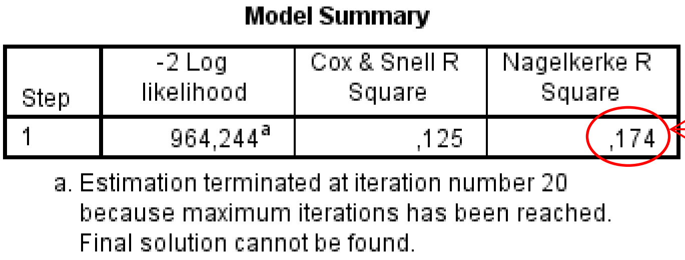


Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
Skincolor	Light brown / brown / black	28	1,000	,000
	white to olive	120	,000	1,000
	Fair / white	697	,000	,000
Baldness	Severe baldness	129	1,000	,000
	Medium baldness	175	,000	1,000
	No / almost no baldness	541	,000	,000
Sex	Male	379	,000	
	Female	466	1,000	

The dummy variables created by SPSS: e.g. Baldness(1) in the 'Variables in the equation' table represent Severe baldness

0 for all dummy variables Indicate the reference category

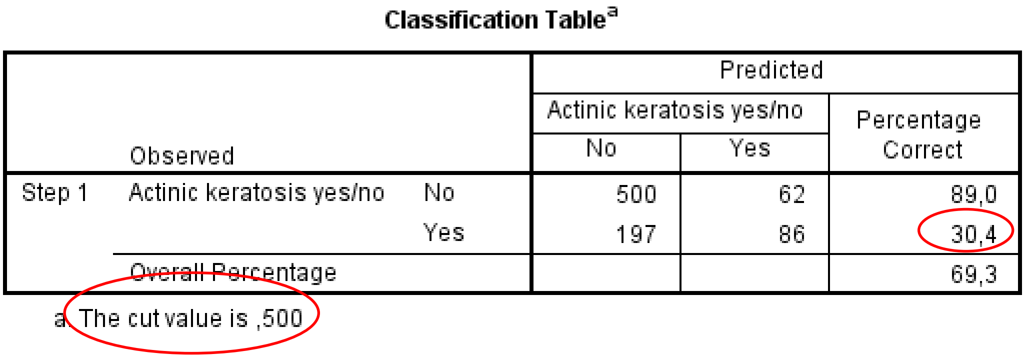


Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	964,244[a]	,125	,174

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The best approximation of the R^2 is Nagelkerke R^2. The interpretation is equal to the interpretation of the R^2 of a linear regression

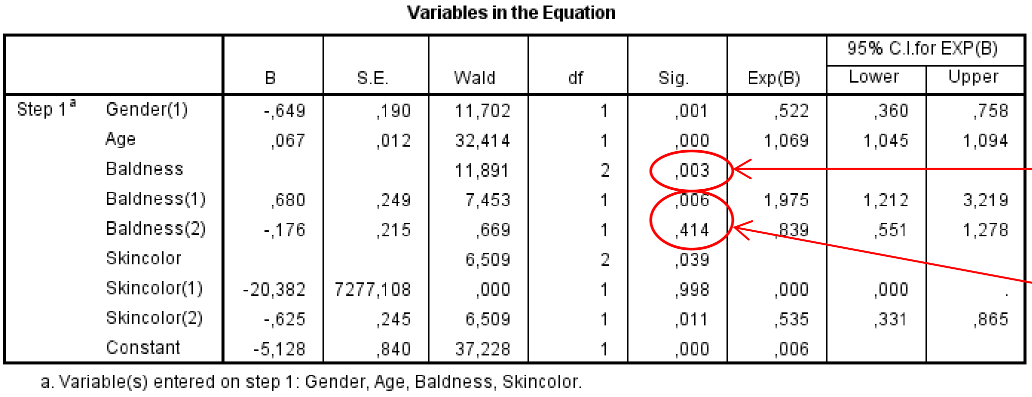


Classification Table[a]

Observed			Predicted: Actinic keratosis yes/no No	Predicted: Actinic keratosis yes/no Yes	Predicted: Percentage Correct
Step 1	Actinic keratosis yes/no	No	500	62	89,0
		Yes	197	86	30,4
	Overall Percentage				69,3

a. The cut value is ,500

If the predicted probability of AK of this regression model is > 0.5 than a patient is classified as having AK. Thus with this model, 30.4% is correctly



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Gender(1)	-,649	,190	11,702	1	,001	,522	,360	,758
	Age	,067	,012	32,414	1	,000	1,069	1,045	1,094
	Baldness			11,891	2	,003			
	Baldness(1)	,680	,249	7,453	1	,006	1,975	1,212	3,219
	Baldness(2)	-,176	,215	,669	1	,414	,839	,551	1,278
	Skincolor			6,509	2	,039			
	Skincolor(1)	-20,382	7277,108	,000	1	,998	,000	,000	.
	Skincolor(2)	-,625	,245	6,509	1	,011	,535	,331	,865
	Constant	-5,128	,840	37,228	1	,000	,006		

a. Variable(s) entered on step 1: Gender, Age, Baldness, Skincolor.

Overall p-value of the variable 'Baldness'

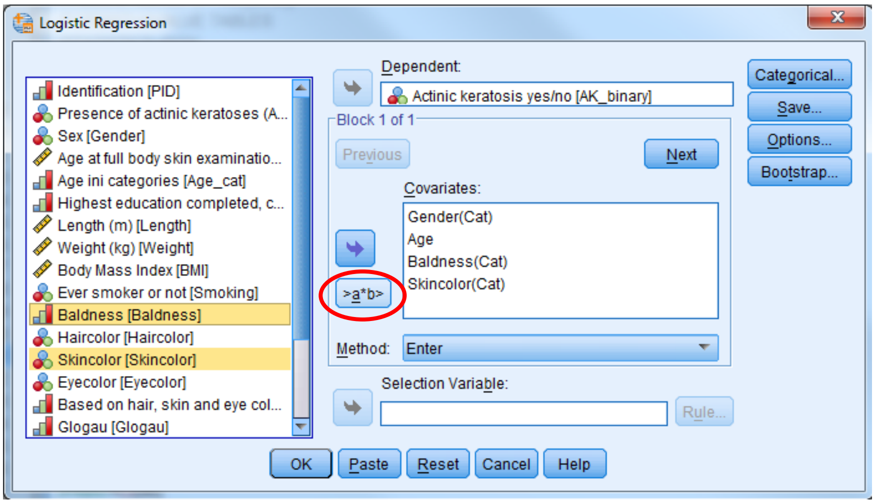
Separate p-values of each category of 'Baldness'

The p-values of all variables are below 0.05, (gender=0.001, age<0.001, baldness=0.003, skin color is 0.039) , thus we can conclude that all variables predict AK risk independent of the other variables in the model.

Interpretation of the OR for a continuous and a categorical variable:
Age: the odds of AK increases with 6.9% per year increase in age adjusted for sex, baldness and skin color
Skin color: patients with a white to olive skin have an decreased odds on AK (OR 0.54 (95% CI: 0.33-0.87) compared to patients with a fair/white skin. The risk of patients with a light brown/ brown or black skin could not be estimated, due to the low number of patients (The actual number of patients cannot be obtained from this output, but should have been calculated in advance by using a frequency table)

4.3b Is there multiplicative interaction between skin color and baldness?

In other words, are patients who are both severely bald and have a fair/white skin at extra increased risk of AK than would have been expected based the risk of baldness and skin color alone?
Include an interaction term in the model by selecting both variables by pressing the Ctrl key and use the '>a*b>' key. Always put both variables in the model as well, otherwise the interaction term will be meaningless.

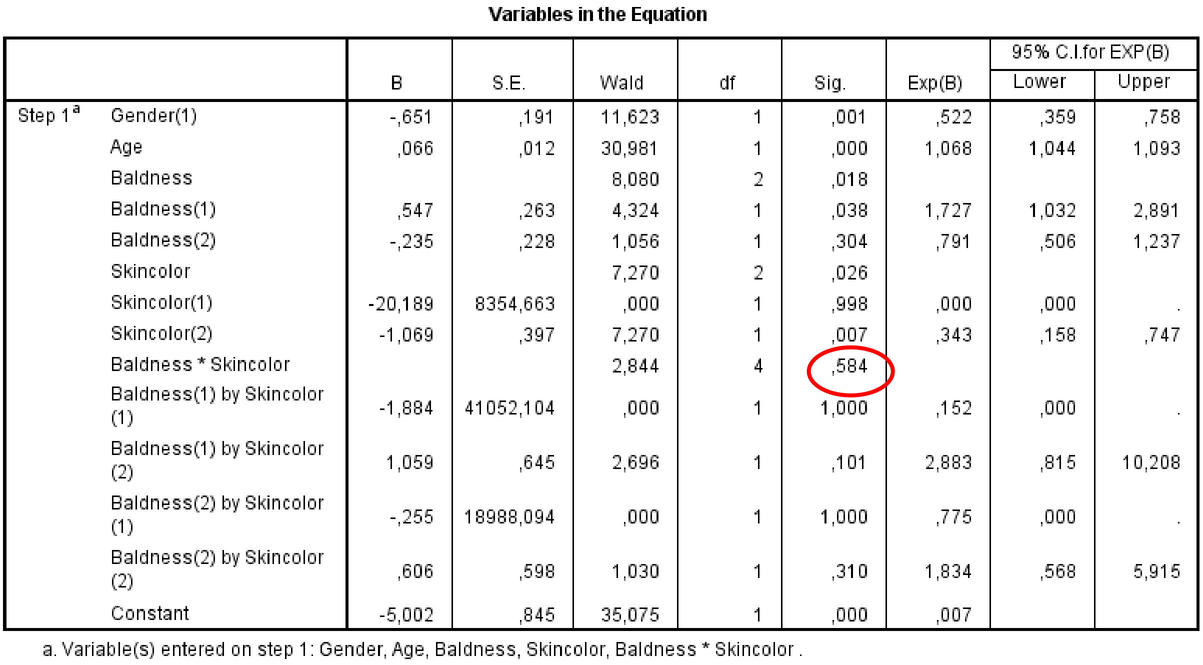


Dependent: Actinic keratosis yes/no [AK_binary]

Covariates:
Gender(Cat)
Age
Baldness(Cat)
Skincolor(Cat)

Method: Enter

Thus the model includes:
Gender
Age
Baldness
Skincolor
Baldness*Skincolor



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender(1)	-,651	,191	11,623	1	,001	,522	,359	,758
	Age	,066	,012	30,981	1	,000	1,068	1,044	1,093
	Baldness			8,080	2	,018			
	Baldness(1)	,547	,263	4,324	1	,038	1,727	1,032	2,891
	Baldness(2)	-,235	,228	1,056	1	,304	,791	,506	1,237
	Skincolor			7,270	2	,026			
	Skincolor(1)	-20,189	8354,663	,000	1	,998	,000	,000	.
	Skincolor(2)	-1,069	,397	7,270	1	,007	,343	,158	,747
	Baldness * Skincolor			2,844	4	,584			
	Baldness(1) by Skincolor (1)	-1,884	41052,104	,000	1	1,000	,152	,000	.
	Baldness(1) by Skincolor (2)	1,059	,645	2,696	1	,101	2,883	,815	10,208
	Baldness(2) by Skincolor (1)	-,255	18988,094	,000	1	1,000	,775	,000	.
	Baldness(2) by Skincolor (2)	,606	,598	1,030	1	,310	1,834	,568	5,915
	Constant	-5,002	,845	35,075	1	,000	,007		

a. Variable(s) entered on step 1: Gender, Age, Baldness, Skincolor, Baldness * Skincolor .

The p-value of the interaction term is 0.584, thus there is no statistical interaction between baldness and skin color. In case of interaction, the ORs are difficult to interpret and the best option would have been to stratify the model on baldness to obtain separate ORs for age, gender and skin color for each category of baldness.

4.4* Selection of variables

Which variables should be included in the multivariable model?

This is an important question. It is dependent on the purpose of the model. Do you want to predict the outcome or adjust for possible confounders? It is also dependent on the sample size of the study (see box below).

In general, there are three possibilities to choose which variable should be included in the model. A combination of these is also possible. Specify the method which you would like to use in advance.

1. based on literature or clinical expertise
2. based on influence on the risk estimate of the exposure variable (has been discussed in paragraph 3.6)
3. based on statistical significance

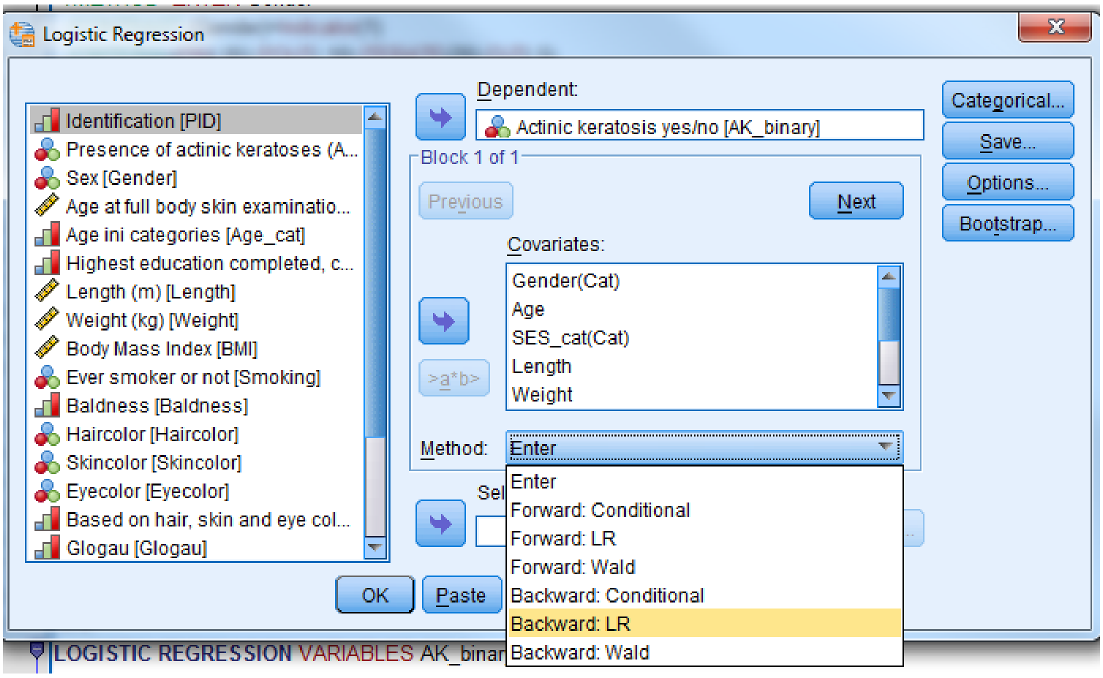
Personally, I prefer option 1 and 2, but option 3 is also widely used and will be discussed in this paragraph. Based on statistical significance you can choose between two stepwise selection methods: backward elimination and forward selection.

Forward selection:	start with an empty model and the most significant variable is included in each step
Backward elimination:	start with a full model and the variable with the highest p-value is excluded in each step.

Backward elimination is to be preferred as this takes all possible associations into account.

Step 1: perform a logistic regression model using backward elimination

Analyze > Regression > Binary Logistics



Dependent: Actinic keratosis yes/no [AK_binary]

Covariates:
Gender(Cat)
Age
SES_cat(Cat)
Length
Weight

Method: Enter

Choose Backward: LR, which means Likelihood Ratio and refers to the method of how the p-value is calculated.
Include the following variables: Gender Age SES_cat Length Weight Smoking Baldness Skincolor Sunburn. Specify all categorical variables and change the reference category for gender into the first category.
Go to 'options' and have a look at the stopping rules (p-values) for a forward selection (entry = PIN(0.05)) and a backward elimination (removal =POUT(0.10)).

```
LOGISTIC REGRESSION VARIABLES AK_binary
  /METHOD=BSTEP(LR)  Gender Age SES_cat Length Weight Smoking Baldness Skincolor
Sunburn
  /CONTRAST (Gender)=Indicator(1)
  /CONTRAST (SES_cat)=Indicator
  /CONTRAST (Smoking)=Indicator
  /CONTRAST (Baldness)=Indicator
  /CONTRAST (Skincolor)=Indicator
  /CONTRAST (Sunburn)=Indicator
  /PRINT=CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Step 2: Interpret the output

In block 1 you see step 1 until step 5 for each table. 'Look at the variables in the equation table'. Step 1 is the full multivariable model. The variable with the highest p-value in this step is weight (p=0.996), thus this term is removed and all OR and p-values are calculated again in step 2. Smoking has the highest p-value in step 2 (p=0.456) and is subsequently removed. The final model is the model as calculated in step 5 where none of the variables has a p-value above 0.10. This stopping rules can be changed. In the output you can see that the p-value of skin color is 0.107. You may wonder, why this variable has not been removed. This p-value is based on the Wald test statistic, while we have based the backward elimination on p-value from the Likelihood Ratio test statistic.

How many variables can be screened for an association based on the sample size to prevent overfitting?

Overfitting occurs if the model is too complex and the sample size too small. This leads to too optimistic regression coefficients in a prediction model and future observations will not be correctly predicted.
Each regression coefficient (β) is also called a parameter and a degree of freedom (df) which is spend. A categorical variable with 3 categories (e.g. baldness) has 2 β's and thus 2 df.
Use the following formulas to calculate the number of df that can be spend on regression modelling. Note that these include all variables screened for association, thus not only those included in the final model.

Linear regression: $\frac{total\ sample\ size}{10}$
Logistic regression: $\frac{\min(cases, controls)}{10}$
Cox proportional hazards regression (survival analysis): $\frac{number\ of\ events}{10}$

Example: a case-control study with 600 cases and 200 controls can spend 20 df for prediction modelling. A more stringent approach would be to divide by 20 instead of 10.
(source: F. Harrell, Regression Modelling Strategies)

4.5* Goodness of fit measures

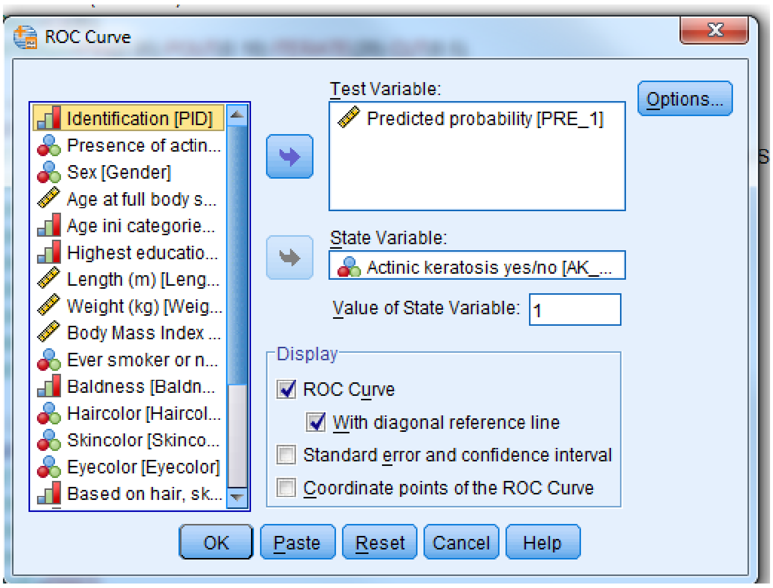
Measures of model performance includes both measures of discrimination (how well can the model discriminate between those with and without the outcome) and calibration (agreement between observed and predicted values). An easy way to interpret measure of overall performance is the R^2. Nagelkerke's R^2 has been mentioned in paragraph 4.3.

A widely used method as a measure of discrimination is the receiver operating curve (ROC) curve. The true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for each cut-off value of the predicted probability.
To use this function in SPSS you need to save the predicted probabilities of the logistic regression model at 'save' and tick the box 'probabilities' of 'Predicted Values. A new variable (PRE_1) will appear in the dataset.

Step 1. Fit a model with Age only and fit a model with Gender Age Baldness Skincolor Sunburn and save the predicted probabilities of both models.

Step 2: Make a ROC curve of both models

Analyze > ROC Curve



Test Variable: Predicted probability [PRE_1]

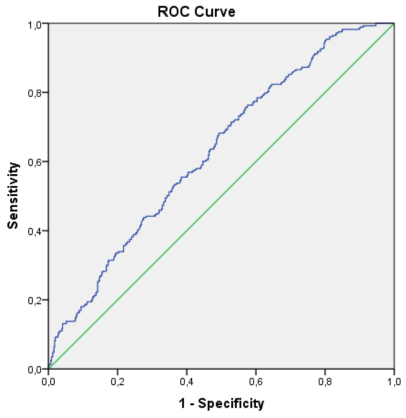
State Variable: Actinic keratosis yes/no [AK_...

Value of State Variable: 1

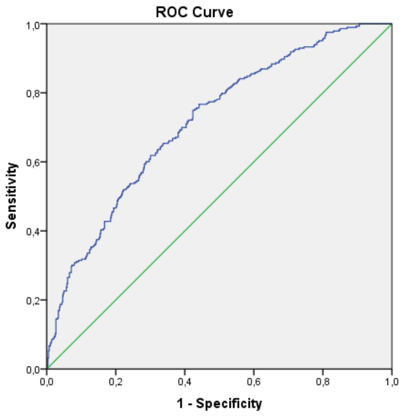
Display
- ☑ ROC Curve
  - ☑ With diagonal reference line
- ☐ Standard error and confidence interval
- ☐ Coordinate points of the ROC Curve

Use PRE_1 (or PRE_2) as test variable and the state variable is AK, indicate that 1 is the value which represents AK. Ask for a diagonal reference line as well.

Step 3: Interpret the output



Diagonal segments are produced by ties.



Diagonal segments are produced by ties.

The Area Under the Curve (AUC) reflects how well the model discriminates between patients with and without AK. The diagonal (green line) represents an AUC of 0.50, which means that you might as well flip a coin. On the left side the ROC curve of the model with age only is shown. This has a AUC of 0.628, which is not very good. The model which includes age, gender, skin color, baldness and sunburn has an AUC of 0.716, which represents acceptable discrimination between patients with and without AK based on the model.
Sometimes the following conventions are used:
AUC>0.7: acceptable discrimination
AUC>0.8: excellent discrimination
AUC>0.9: outstanding discrimination (very unusual)

4.6* Conditional Logistic regression

Case-control studies can be matched on confounding factors. In this dataset each patient with AK was matched to 2 control participants of the same age and sex without AK.
The matching variables (age and gender) should not be included in the regression model. The confounders were taken into account by study design and statistical adjustment is therefore not necessary and incorrect. Possible interaction with the matching variables can be taken into account in the analysis.

Conditional logistic regression should be performed with a trick in SPSS. Each case-control (or case-2 controls) pair is regarded as one stratum. The analysis should be performed using Cox proportional hazards regression

Step 1: select all patients who are included in the nested matched case-control study. (match_id not equal to 0)

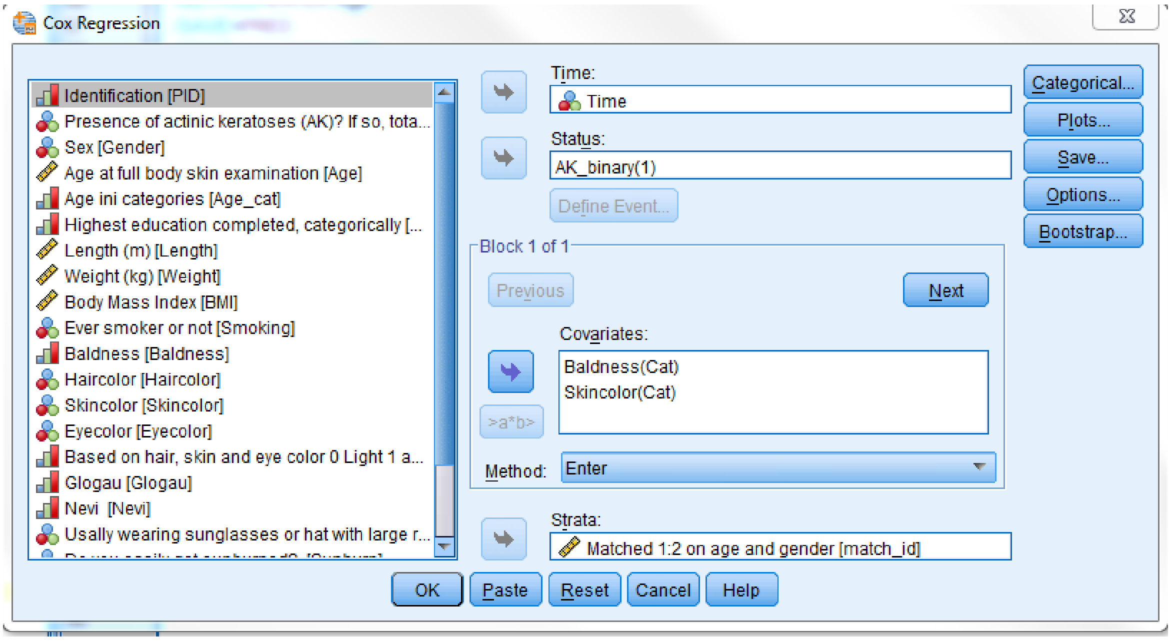
Data > Select Cases

Step 2: calculate a time variable which is 1 for each subject.

Transform > Compute variable

Step 3: Perform a conditional regression model with baldness and skin color

Analyze > Survival > Cox Regression



Time: Time

Status: AK_binary(1)

Block 1 of 1

Covariates:
Baldness(Cat)
Skincolor(Cat)

Method: Enter

Strata: Matched 1:2 on age and gender [match_id]

Ask for 95% CI of Exp(B) at ‘options’ and do not forget to specify the categorical variables.

Step 4: Interpret the output.

In the stratum table you can see the number of pairs (n=48) and the number of cases (event=1) and controls (censored=2). .

Write down the number of pairs with only 1 control and the calculated OR. What is the difference with the OR obtained in paragraph 4.3 for the same variables? And what is the reason for this difference?

4.7* Ordinal and multinomial regression

The number of AK was also assessed in categories (no AK[0], 1-3 AK[1], 4-9 AK[2], >10 AK[3]). A lot of information is lost by reducing this information to a binary variable (AK yes/no). Regression analysis with multiple outcome categories is also possible.

Ordinal regression

Ordinal regression is also called the proportional odds model. Ordinal regression assumes that the odds ratio is equal between every cut-off value of the categories:

OR 0 *vs* 1+2+3
=
OR 0+1 *vs* 2+3
=
OR 0+1+2 *vs* 3

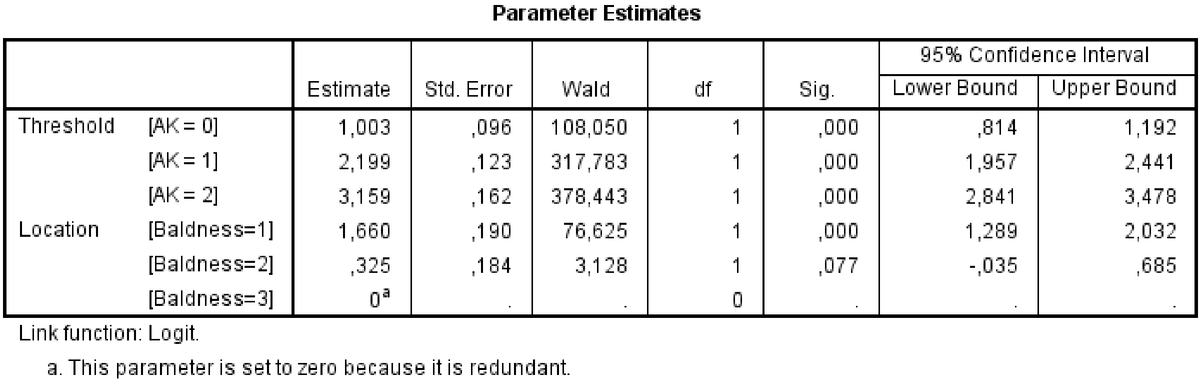
This assumption can be tested using the test of parallel lines. The Null hypothesis of the test of parallel lines is that the regression coefficient (β) is equal across the outcome categories.

Analyze > Regression > Ordinal

Step 1: First select all patients with a value of AK which is not missing (both system and user missing).

Step 2: Perform an ordinal regression. Use the number of AK in categories as outcome variable and baldness as independent variable (Factor). Go to ‘output’ and select the test of parallel lines. Note that SPSS calls continues variables ‘covariates’ and categorical variables ‘factors’.

Step 3: Interpret the output.

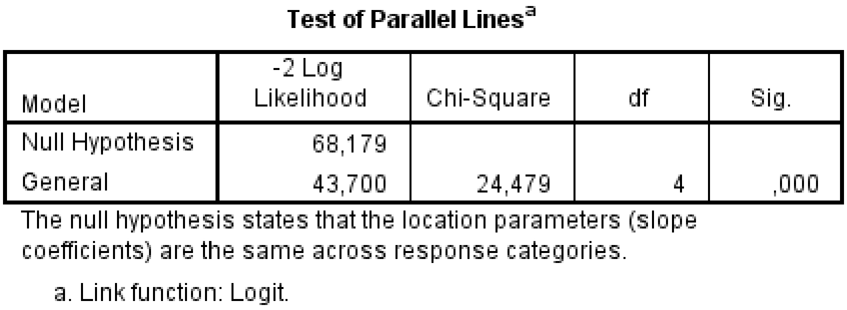


Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[AK = 0]	1,003	,096	108,050	1	,000	,814	1,192
	[AK = 1]	2,199	,123	317,783	1	,000	1,957	2,441
	[AK = 2]	3,159	,162	378,443	1	,000	2,841	3,478
Location	[Baldness=1]	1,660	,190	76,625	1	,000	1,289	2,032
	[Baldness=2]	,325	,184	3,128	1	,077	-,035	,685
	[Baldness=3]	0^a	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.



Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	68,179			
General	43,700	24,479	4	,000

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Note: SPSS always takes the last category as reference category and only LN(odds) estimates are provided. The exponent of the estimates should be taken to calculate the OR.

The OR of medium baldness compared to no baldness is $e^{0.325}$=1.38 between every cut-off value of the categories of number of AK.

However, the p-value of the test of parallel lines is <0.001 and the null hypothesis is rejected. The proportional odds assumption is not fulfilled and an ordinal regression may not be the best model.

Multinomial regression

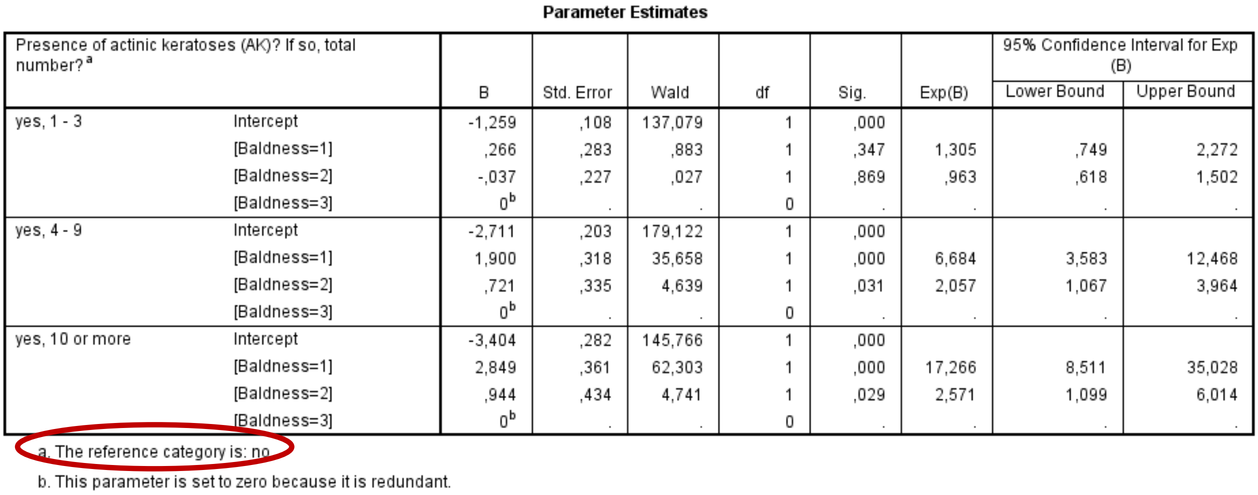
If this assumption does not hold a multinomial regression may be more appropriate. A multinomial regression provides an OR for each category compared to a reference category:

OR 1 *vs* 0
OR 2 *vs* 0
OR 3 *vs* 0

Analyze > Regression > Multinomial Logistic

Step 1: Perform the analysis. Use the number of AK in categories as a dependent variable. Select the first category as the reference category and use baldness as a factor.

Step 2: Interpret the output.



Parameter Estimates

Presence of actinic keratoses (AK)? If so, total number?[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
yes, 1 - 3	Intercept	-1,259	,108	137,079	1	,000			
	[Baldness=1]	,266	,283	,883	1	,347	1,305	,749	2,272
	[Baldness=2]	-,037	,227	,027	1	,869	,963	,618	1,502
	[Baldness=3]	0[b]	.	.	0	.	.	.	.
yes, 4 - 9	Intercept	-2,711	,203	179,122	1	,000			
	[Baldness=1]	1,900	,318	35,658	1	,000	6,684	3,583	12,468
	[Baldness=2]	,721	,335	4,639	1	,031	2,057	1,067	3,964
	[Baldness=3]	0[b]	.	.	0	.	.	.	.
yes, 10 or more	Intercept	-3,404	,282	145,766	1	,000			
	[Baldness=1]	2,849	,361	62,303	1	,000	17,266	8,511	35,028
	[Baldness=2]	,944	,434	4,741	1	,029	2,571	1,099	6,014
	[Baldness=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: no.

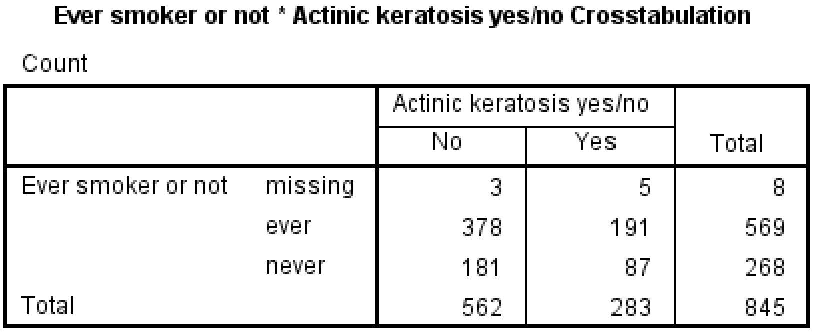
b. This parameter is set to zero because it is redundant.

The reference category is indicated below the table. Odds ratio are directly provided. Note that the last categories of the independent variables (factors) are the reference categories. For example the OR of 4-9 AK vs no AK is 6.7 (95% CI 3.6-12.5) for severe baldness compared to no baldness.

Assignment 4-1 Logistic Regression - basic

Data: actinic keratosis (AK)
More than 800 people were screened on the presence of actinic keratosis. The number of actinic keratosis was registered as well as potential risk factors for the development of actinic keratosis.

1. Calculate the OR of ever smoking vs never smoking from the 2x2 table below.



Ever smoker or not * Actinic keratosis yes/no Crosstabulation

Count

		Actinic keratosis yes/no		Total
		No	Yes	
Ever smoker or not	missing	3	5	8
	ever	378	191	569
	never	181	87	268
Total		562	283	845

2. Perform an univariable binary logistic regression model with smoking as an independent predictor and perform an age and sex adjusted model. Provide the univariable and adjusted OR for ever smoking vs never smoking and their 95% CI.

3a. Are age, sex, socioeconomic status, smoking, baldness and skin color independent risk factors for actinic keratosis? Provide the p-value of each variable.

3b. What is the adjusted OR (95% CI) of a high socioeconomic status vs a low socioeconomic status?

3c. What is the adjusted OR (95% CI) for a 10-year increase in age?

4. Is the risk associated with ageing higher in smokers than in non-smokers?

5. Does the including ‘ever lived in a sunny country for > 1 year’ make a statistical significant contribution to prediction AK risk of the model in question 3a? Provide the p-value and the improvement in explained variance.

Assignment 4-2 Logistic Regression - advanced

1. Calculate the 95% CI for the OR of psoriatic arthritis of females vs males in paragraph 4.1

Data: actinic keratosis (AK)
More than 800 people were screened on the presence of actinic keratosis. The number of actinic keratosis was registered as well as potential risk factors for the development of actinic keratosis.

2. Perform a backward elimination binary logistic regression model using age, sex, baldness, smoking, BMI, skin color, hatglasses and sunburn. Define the categorical covariates correctly. Use a p-value of 0.05 for backward elimination. Which variables are included in the final model?

3. Start with the variables included in the final model obtained in question 2. Is the risk increase per year of age higher among those people who lived in a sunny country for >1 year? Provide the test and the p-value.

4. Provide a measure of discrimination of the final model in question 2. What is your opinion about the discrimination of this model.

5. Include the variable which indicates if people worked outdoors in the final model of question 2. What is the improvement in overall goodness of fit and discrimination?

6. Fit an univariable proportional odds model (ordinal regression) for smoking and number of AK. 6a. Is the proportional odds assumption fulfilled? What is the test and the null hypothesis?
6b.If the assumption is fulfilled, provide the OR and its interpretation.

7. Fit a multivariable multinomial logistic regression model for number of AK including age, sex and baldness.
7a. Is the risk on >10 AK higher among males or females adjusted for the other covariates (both categorical and continuous covariates)? Provide the OR of males vs females and 95% CI.
7b. Is baldness independently related with the risk on the number of AK? Provide the overall p-value.

Chapter 5 Survival Analysis: Kaplan Meier

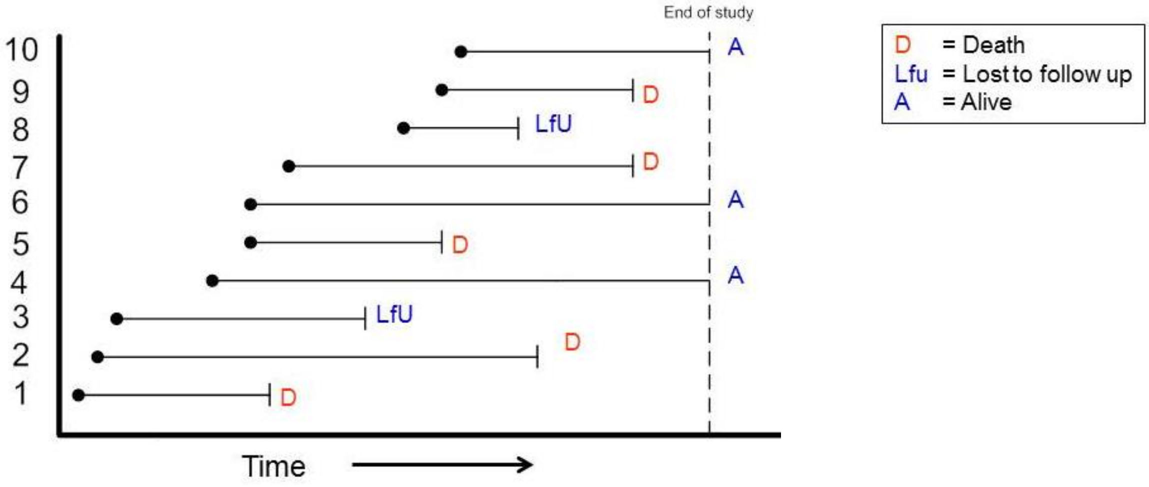
Data: Survival
The survival of 15 cancer patients from a randomized controlled trial was registered. Seven patients received treatment and 8 patients received placebo.

5.1 What is survival analysis?

In survival analyses the time to the **event** (e.g. death, disease, recurrence) is analysed, rather than having the event or not , like in a logistic regression analysis. An event is also called a **failure**. In survival analysis we have to deal with **censored data**, meaning that we have some information about the survival time, but we don't know the survival time exactly.

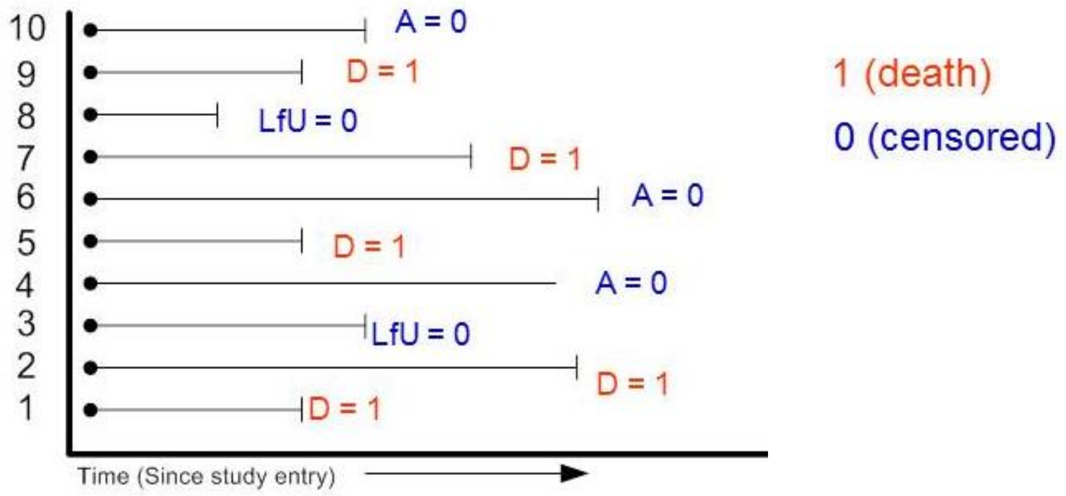
There are several reasons for censoring:

1. the patient has reached the end of the study without experiencing the event
2. the patient is lost to follow up
3. the patient withdraws from the study without experiencing the event.



The model assumption of a Kaplan-Meier curve is 'independent censoring'. This means that we assume that patients who are censored have the same probability of experiencing the event of interest as those who remain in the study. This assumption cannot be tested.

Although patients enter the study throughout the calendar time, we assume that patients who are recruited early have the same survival as those who were recruited later. This can be tested. The study entry is time (t)=0. The event is coded as 1 and the censored observation as 0 (either lost to follow up, withdrawals, or end of study). Thus, we have two outcome variables in survival analysis: the follow-up time and a variable, which indicates failure or censoring.



5.2 Calculate Kaplan-Meier survival estimate

A Kaplan-Meier curve can be easily calculated by hand. Consider the following survival times:

6*,8,15,15,19*,20,22,25,32*,36*,41*,42*,48*,48,52*

An asterisk (*) indicates that the patient has been censored. All other patients died. Finish the following survival table to calculate the Kaplan-Meier (KM) survival estimates

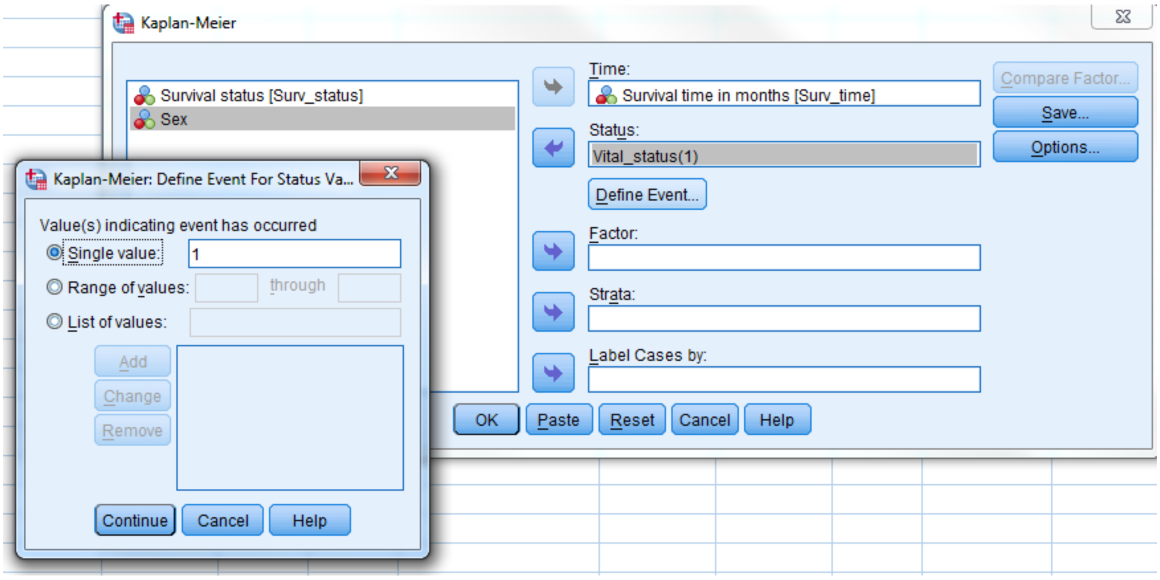
t_i Time	N_i= Patients at risk at t_i	δ_i= Patient with failure at t_i	S_i= Patients who survive after t_i	S_i/N_i	KM survival estimate
t=8	14	1	13	13/14=0.929	0.929
t=15	13	2	11	11/13=0.846	0.846*0.929=0.786
t=20					
t=22					
t=25					
t=48					

5.3 Kaplan – Meier curve in SPSS

The obtained KM survival estimates are usually represented in a curve. The curve starts at t=0, where the survival is 100%. Each time an event happens the curve drops to the next KM survival estimate.

Step 1: Open the survival data of chapter 5 and inspect the variables and go to:

Analyze > Survival > Kaplan Meier



Time: Survival time in months [Surv_time]

Status: Vital_status(1)

Value(s) indicating event has occurred

● Single value: 1

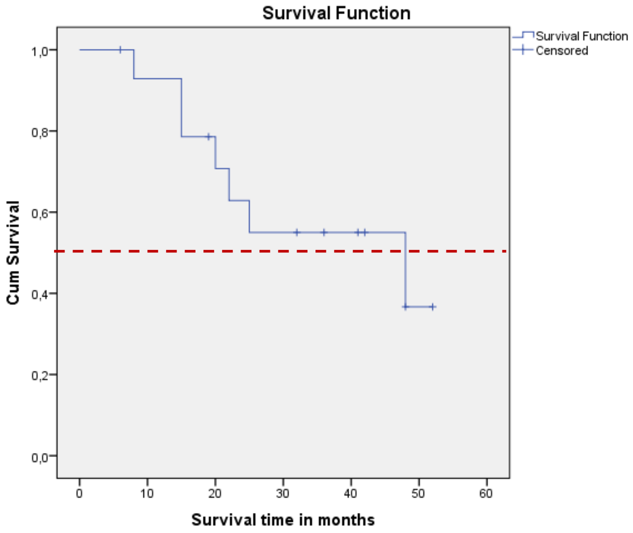
Define the two outcome variables: the survival time and failure/censoring variable. Go to ‘options and ask for a survival table, mean and median survival and a survival plot.

Step 2: Interpret the output.

<u>Survival Table:</u> Look at the survival table. Compare the table to the obtained KM estimates in paragraph 5.2.

<u>Survival Plot:</u> The survival plot is shown below. Compare the plot to the table. The curve drops at the time of each event. SPSS also shows when a patient has been censored.

<u>Mean and Median Survival:</u> The median survival is the time when the survival probability is 50%. The dotted red line indicates 50%. Although the actual survival is 0.367 at t=48, the curve does not cross the red line before this time point and therefore the median survival time is 48 months.
The mean survival is calculated by using the area under the curve. Survival times are frequently highly skewed data and therefore the median survival is a better measure than the mean survival.



5.4 Logrank test

In most questions we are not interested in overall survival, but we would like to compare the survival between groups. Consider the following research question:

5.4. Is the survival of patients who received treatment better compared to those who received placebo?

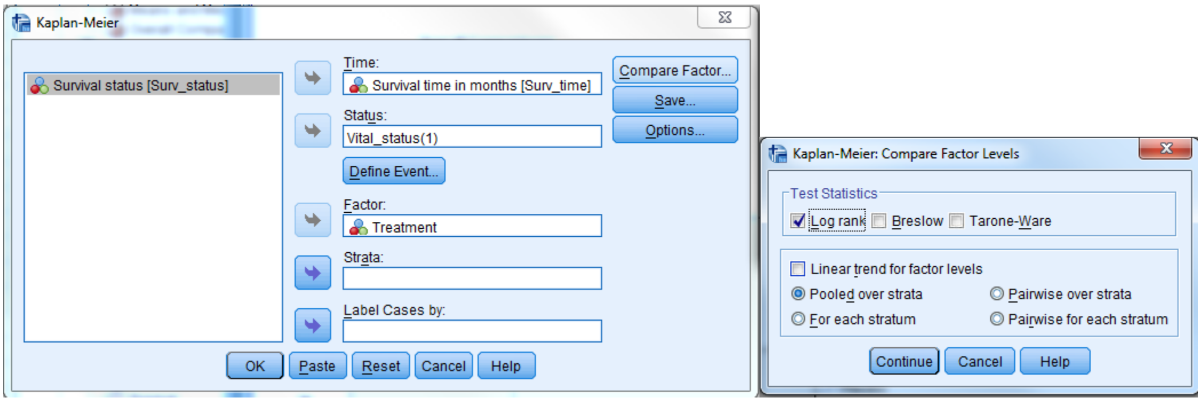
To answer this question we will use the logrank test. The logrank test compares the number of events at every time points of the survival curve. The null hypothesis is as follows:

H_0: there is no difference in survival between the groups

Step 1: Perform the logrank test in SPSS

Analyze > Survival > Kaplan Meier

To perform the logrank test in SPSS, use Factor as grouping variable. Go to 'Compare Factor' and tick the box of Log rank.



Survival status [Surv_status]

Time: Survival time in months [Surv_time]

Status: Vital_status(1)

Factor: Treatment

Strata:

Label Cases by:

Test Statistics

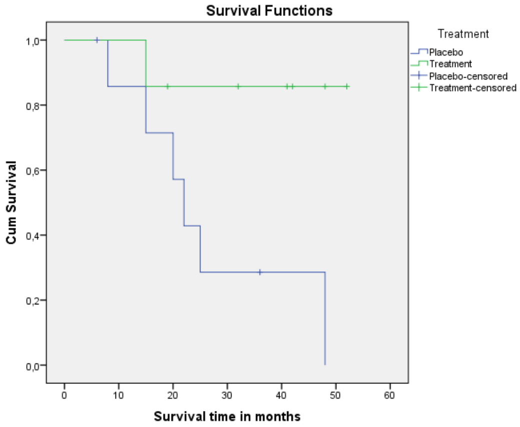
☑ Log rank ☐ Breslow ☐ Tarone-Ware

☐ Linear trend for factor levels

● Pooled over strata ○ Pairwise over strata

○ For each stratum ○ Pairwise for each stratum

Step 2: Interpret the output



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,154	1	,023

Test of equality of survival distributions for the different levels of Treatment.

The p-value of the logrank test is 0.023, thus the null hypothesis can be rejected and the survival of the treatment and the placebo group is not equal. The survival plot shows that the survival of the treatment group is better compared to the placebo group.

Assignment 5-1 Kaplan-Meier curve and Logrank test

Data: melanoma

For this assignment we have created an adapted dataset of melanoma patients selected from a population-based database. There is information on age and date of diagnosis, sex, stage, vital status and survival time.

1. Inspect the data: which variables are in there, how many patients are included? Which are the variables you would need for a survival analysis?

2. Make a Kaplan-Meier curve. If you want to limit the output, you can choose not to display the survival table in the 'Options' menu. What is the mean and median survival time of the melanoma patients?

You have now made the first K-M plot for overall survival, but we might be more interested in some sub-analyses, comparing the survival of melanoma patients by gender, nodal stage and metastatic stage.

3. Compare melanoma survival by sex: who have a better survival? How large is the difference? Is this difference significant? Define the null hypothesis of the logrank test.

4a. Do the same for survival by nodal stage. What can you conclude about patients with stage X? how would you describe them in terms of prognosis? Does this make sense clinically?

4b. Repeat the analysis, but perform also a trend test. Go to 'Compare Factor' and choose 'Logrank' and 'Linear trend for factor levels'. What can you conclude from this test?

5. Compare melanoma survival by metastatic stage: what do you conclude with regards to patient with stage X? Does this make sense clinically?

Chapter 6* Cox proportional hazards regression

Data: melanoma
In this chapter we will continue with the data presented in assignment 5.1. There is information on age and date of diagnosis, sex, stage, vital status and survival time. In addition, data on medication use (beta-blockers) has been collected.

6.1* What is a Cox proportional hazards model?

We have seen, using the Kaplan-Meier method, that survival of melanoma differs by sex and stage. However, using this method, it is not possible to look at more than one variable at the time, unless you make many strata, which is usually practically impossible because you would need very large datasets. Cox proportional hazards method makes it possible to make multivariable models. Cox proportional hazards model provides hazard ratios. The hazard is the probability of experiencing an event at Δt, conditional on being event-free at the beginning of Δt. The hazard is thus a conditional probability per time unit:

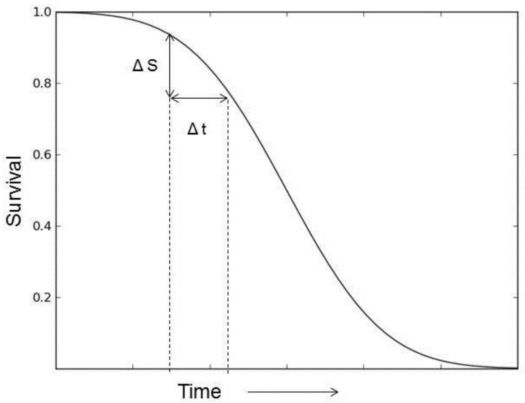


Figure 6.1: the hazard is the probability of experiencing the event at Δt, given that the individual is alive at the beginning of Δt.

Proportional hazards assumption
In a Cox proportional hazards model, we assume that the hazard ratio is constant over time. This means that hazards should be proportional over time. This assumption can be tested, which is shown in paragraph 6.2. The baseline hazard function does not need to be specified and can be any shape as shown in figure 6.2.

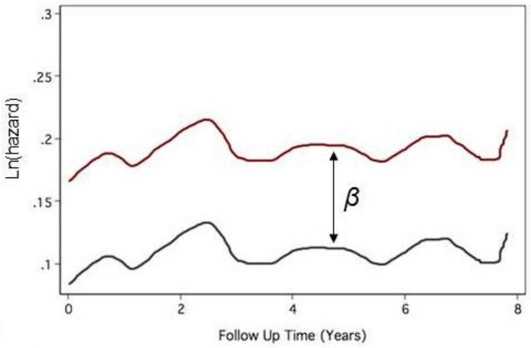


Figure 6.2: The baseline ln(hazard) function and the proportional hazards assumption.

The hazard ratio is calculated by taking the exponent of the β ($HR = e^{\beta}$). As seen in Figure 6.2 the β is independent of the shape of the hazard function and of time.

Non-proportional hazards

A possible problem in Cox PH regression is violation of the proportional hazards assumption. As can be seen in Figure 6.3 the $\mathrm{HR}_{\text{group B vsGroup A}}$ is below 1 at the beginning of follow up and above 1 at the end of follow up. It makes no sense to estimate a HR over the whole time period.

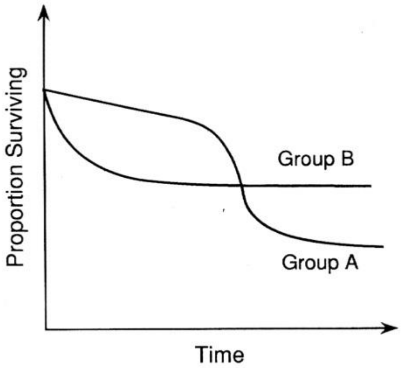


Figure 6.3: Nonproportional hazards Source: Concato et al, Ann.Intern. Med. 1993

There are several options when the proportional hazards assumption is not met:
-fit time-specific models and thereby calculate time-interval specific hazards.
-stratify on the variable with non-proportional hazards
-include time dependent covariates.

The Cox proportional hazards model

The hazard at time=t is a product of the baseline hazard function (*h₀(t))* and the exponential of the covariates (X's) and the regression coefficients (β's):

$$h(t) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k}$$

.

The baseline hazard function ($h_{0)}$ is not specified, which makes the Cox PH model a semi-parametric model. The baseline hazard function does not need to be specified to estimate the hazard ratio, because the baseline hazard cancels out of the equation.

Consider the HR of females vs males:

$$hazard_{males}(t) \quad = h_0(t)e^{\beta_1 X_1} = h_0(t)e^{\beta_1 * 1} = h_0(t)e^{\beta_1}$$
$$hazard_{females}(t) = h_0(t)e^{\beta_1 X_1} = h_0(t)e^{\beta_1 * 0} = h_0(t)e^{0}$$

$$HR = \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{0}} = e^{\beta_1}$$

6.2* The proportional hazards assumption

We would like to estimate the HR of females vs. males for melanoma survival, but we need to check the proportional hazards assumption first. There are multiple ways of checking the assumption:

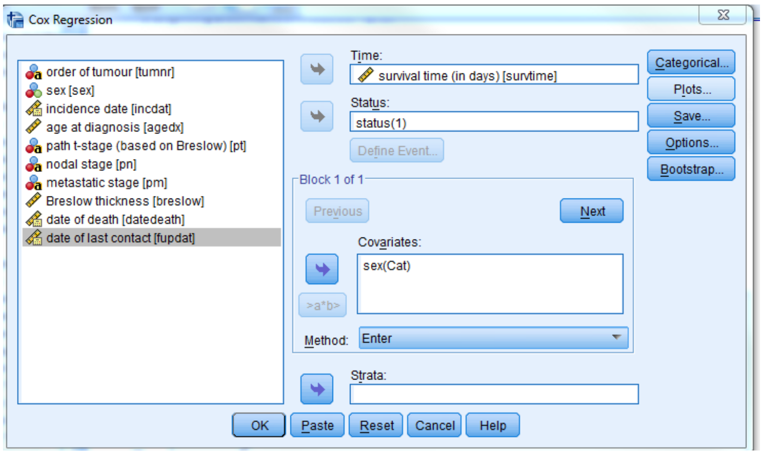
-Visually using logminuslog (LML) survival plots
-Goodness of fit test using Schoenfeld residuals.
-Time-dependent covariates

The easiest way is to use LML survival plots for categorical variables. You don't need to perform all methods. Schoenfeld residuals and time-dependent variables can be used for continuous covariates. Combine an overall goodness of fit test, such as Schoenfeld residuals, with a visual inspection of LML plot.

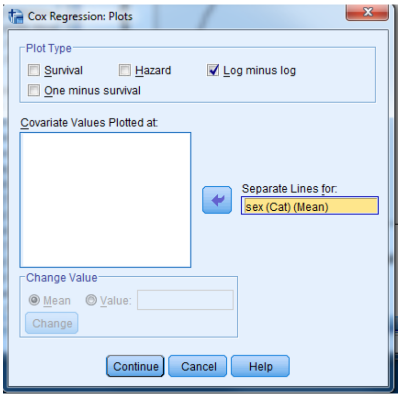
Step 1: Perform the Cox PH model in SPSS

Analyze > Survival > Cox Regression

Specify the survival time variable, the censoring/failure variable. Define the value of the event. Sex is the covariate. Change the reference category to the first value, using 'categorical'. Ask for LogMinusLog (LML) plots by sex. Go to 'Save' and save the partial residuals (Schoenfeld residuals).



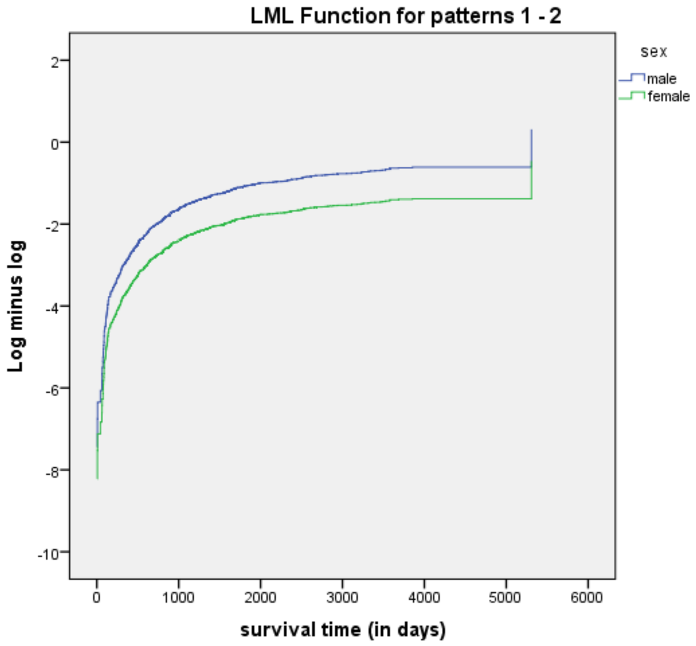
Time: survival time (in days) [survtime]
Status: status(1)
Covariates: sex(Cat)
Method: Enter



☐ Survival ☐ Hazard ☑ Log minus log
☐ One minus survival
Separate Lines for: sex (Cat) (Mean)

Step 2: Inspect the LML survival plot

This is the easiest way to check the assumption. There is no evidence that the LML survival plots cross each other in time. They run perfectly parallel.



Step 3: Test the PH assumption using Schoenfeld residuals.

This is 3-step process:

1. Obtain Schoenfeld residuals
2. Rank failure times
3. Test the correlation between residuals and failure times.

The null hypothesis is that there is no correlation between the residuals and the failure times.

The Schoenfeld residuals have been saved in the variable PR1_1, which should have appeared in the dataset. Select all patients with the event (status=1):

Data > Select cases

Rank the survival time

Transform > Rank Cases

Move survtime to the variable box. By default 1 is already assigned to the smallest value. A new ranking variable appears in the data 'Rsurvtime'.

Correlate the Schoenfeld residuals to the rank of the survival time

Analyze > Correlate > Bivariate

The Pearsons correlation coefficient is 0.043, with a p-value of 0.298. The null hypothesis is not rejected, thus we can conclude that the hazards are proportional.
Don't forget to turn the filter off.

Step 4: Test the PH assumption by using time-dependent covariates.

We assume that the HR is equal over time. Including a variable in the Cox PH model which describes the change of the HR over time should have a regression coefficient of 0.

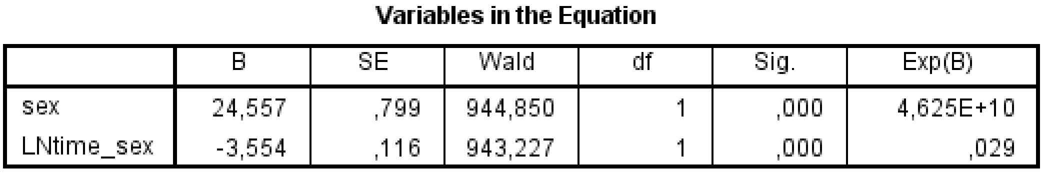
First we need to calculate a new variable of the product of (a function of) time and the covariate. Usually time or LN(time) are commonly used choices.

Transform > Compute variable

Make a new variable ‘LNtime_sex’ which is the product of LN(survtime) and sex. Include this variable in a new Cox PH regression model and test if the β of this variable is equal to 0.

Analyze > Survival > Cox Regression

Specify a Cox PH regression model with sex and the LNtime_sex as covariates



Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
sex	24,557	,799	944,850	1	,000	4,625E+10
LNtime_sex	-3,554	,116	943,227	1	,000	,029

The p-value of the time-dependent variable is <0.001, which indicates that the proportional hazards assumption is violated. However, a disadvantage of this method, is that the significance depends on the choice of the function of time, which may lead to different conclusions. Moreover, a small deviation in a large dataset (n=592) may lead to a statistical significant deviation of the PH assumption, which may not be clinically relevant.

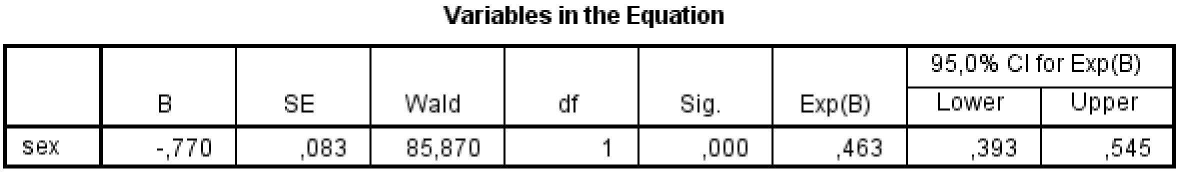
6.3* Univariable Cox proportional hazards regression

Based on the LML survival plot we conclude that the PH assumption is fulfilled.

6.3 What is the hazard ratio (HR) of females vs males of melanoma survival?

Analyze > Survival > Cox Regression

In addition to the model in paragraph 6.3, ask for 95% CI of Exp(B).



Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
sex	-,770	,083	85,870	1	,000	,463	,393	,545

From the output we obtain a HR of females vs males of 0.463 (95% CI: 0.393-0.545). Females are thus at lower risk of death compared to males.

6.4* Multivariable Cox proportional hazard regression

One could reason for the better survival of females, could be because women pay more attention to their skin and present clinically with lower stage tumours. Using our data, we can verify if this is true or not: we can make a multivariable model including sex, but also the stage variables. In this way, the data is ‘corrected’ for stage.

6.4 Is the better survival of females caused by lower stage at diagnosis?

Perform a multivariable Cox PH regression model including sex, Breslow thickness, Nodal stage and Metastatic stage. Do not use pT pN and pM as these are string variables. Breslow thickness is equal to pT and is already a continus variable. Recode the other stage variables.

Transform > Recode into different variables

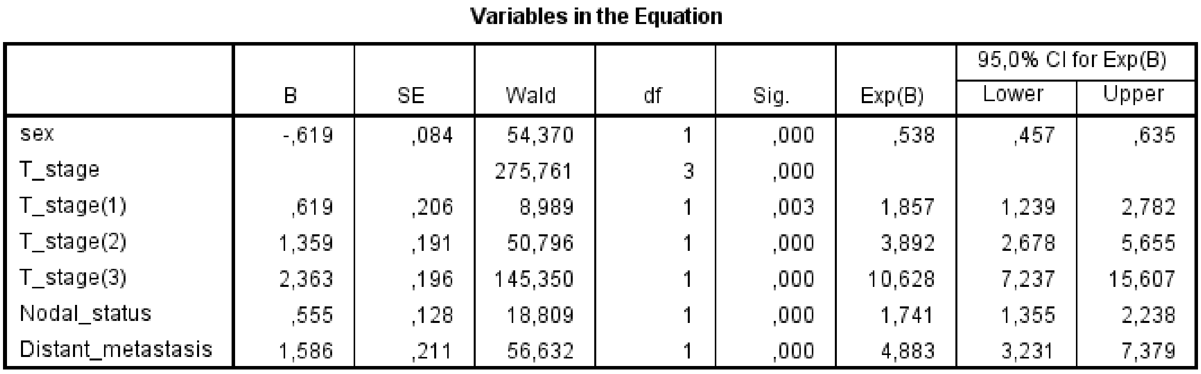
```
RECODE pt ('1'=1) ('2'=2) ('3'=3) ('4'=4) INTO T_stage.
VARIABLE LABELS  T_stage 'T_stage numeric'.
VALUE LABELS T_stage 1 'T1' 2 'T2' 3 'T3' 4 'T4'.
EXECUTE.

RECODE pn ('0'=0) ('1'=1) ('2'=1) ('X'=0) INTO Nodal_status.
VARIABLE LABELS  Nodal_status 'Nodal stage numeric'.
VALUE LABELS Nodal_status 0 'No nodal metastasis' 1 'Nodal metastasis'.
EXECUTE.

RECODE pm ('0'=0) ('1'=1) ('X'=0) INTO Distant_metastasis.
VARIABLE LABELS  Distant_metastasis 'Distant metastasis numeric'.
VALUE LABELS Distant_metastasis 0 'No distant metastasis' 1 'distant metastasis'.
EXECUTE.
```

Analyze > Survival > Cox Regression

Use sex, T-stage, nodal and distant metastasis as covariates. The LML should be checked for T-stage, nodal and distant metastasis.



Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
sex	-,619	,084	54,370	1	,000	,538	,457	,635
T_stage			275,761	3	,000			
T_stage(1)	,619	,206	8,989	1	,003	1,857	1,239	2,782
T_stage(2)	1,359	,191	50,796	1	,000	3,892	2,678	5,655
T_stage(3)	2,363	,196	145,350	1	,000	10,628	7,237	15,607
Nodal_status	,555	,128	18,809	1	,000	1,741	1,355	2,238
Distant_metastasis	1,586	,211	56,632	1	,000	4,883	3,231	7,379

The HR for sex is still statistically significant after adjusting for stage at diagnosis. We can conclude that the survival advantage of females is not caused by lower stage at diagnosis.

6.5* Time-varying covariates

Covariates which are known at time of diagnosis may change during follow up time. Other variable do not change (e.g. sex) and are called time-fixed covariates. Some covariates, such as drug exposure can be analysed as a time-fixed or a time-varying covariate in time-to-event analyses. The use of a time-fixed covariate assumes that exposure is measurable at baseline and remains constant over time (e.g. drug user at diagnosis yes/no). Changes in exposure status during follow up can be taken into account by using a time-varying covariate for exposure (e.g. drug user after diagnosis yes/no). Incorrect use of a time-fixed exposure variable in a Cox proportional hazard (PH) model, or other time-to-event models, can lead to biased estimates, due to immortal time bias. Samy Suissa wrote some excellent papers about this subject.
In this example drug exposure will be analysed correctly as time-fixed covariate (6.5a). In that case only drug use at diagnosis or before diagnosis can be taken into account. Subsequently drug use will be analysed while changing over time (6.5b).

Time-fixed covariate
Betablockers may have a survival advantage by preventing melanoma metastasis. Betablocker use has been assessed at the date of diagnosis.

6.5a Does betablocker use before diagnosis prolong survival among melanoma patients?

Step 1: Perform a Cox PH regression for drug exposure before diagnosis

Analaze > Survival > Cox Regression

Adjust the analysis for age, sex and all TNM variables.
H_0: the regression coefficient for betablocker use before diagnosis is equal to 0.

Step 2: Interpret the output.

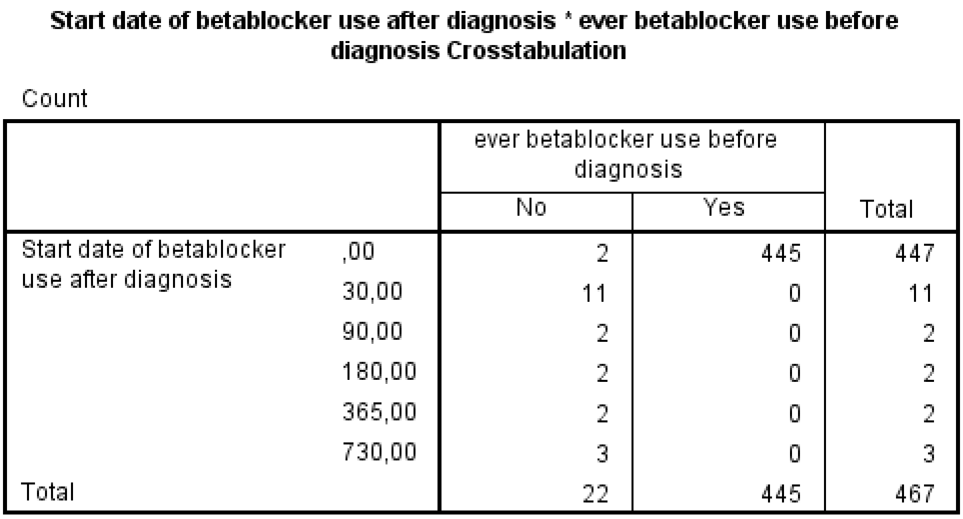
The HR is 1.097 (95% CI: 0.909-1.323) with a p-value of 0.336. The null hypothesis can thus not be rejected. Adjusted for possible confounders betablocker use before diagnosis does not influence survival. This may not surprise you, as patients may start using betablockers after diagnosis as well, who have been analysed as a non-user in this analysis. It would be more close to reality to take them into account as a non-user until they start with betablockers and analyse them as betablocker user afterwards. This will be done in the next analyses.

Time-varying covariate.
The variable BB_start_after_diagnosis indicates when patients started using the drug after diagnosis.

6.5b: Does betablocker use after diagnosis prolong survival among melanoma patients?

Step 1: Make a crosstab between betablocker use before and after diagnosis.

Analyze > Descriptive statistics > Crosstabs



Start date of betablocker use after diagnosis * ever betablocker use before diagnosis Crosstabulation

Count

		ever betablocker use before diagnosis		Total
		No	Yes	
Start date of betablocker use after diagnosis	,00	2	445	447
	30,00	11	0	11
	90,00	2	0	2
	180,00	2	0	2
	365,00	2	0	2
	730,00	3	0	3
Total		22	445	467

This table shows you that some patients who did not use the drug before diagnosis, started using the drug after diagnosis, at day 30, day 90, day 180, day 365 or day 730. All patients who used the drug before diagnosis continued after diagnosis and are users since day 0 (baseline=diagnosis). In the case processing summary you can see that 81.4% have a missing value on this variable. These are the non-users.

SPSS creates an internal time value T_, which can be used to define time-varying covariates. The patient who is a non-user at diagnosis and start using betablockers at day 18. This patient should get a value 0 (non-user) before day 18 and after day 18 the value 1 (user).

Expressed in syntax:

```
DO IF T_ > start_date_after_diagnosis.
COMPUTE BB_user=1.
ELSE.
COMPUTE BB_user=0.
END IF.
EXECUTE.
```

The non-users should not have a missing value, because they will be excluded from the analysis. You can use trick to assign the value 0 to the non-users. They should get an extremely high value, e.g. 1,000,000 days, which is higher than the max. follow up time. The max. follow up time in our data is 5393. T_ will never exceed 5393, thus a patient who has a value of 1,000,000 days will be assigned the value 0 in any case according to the aforementioned syntax.

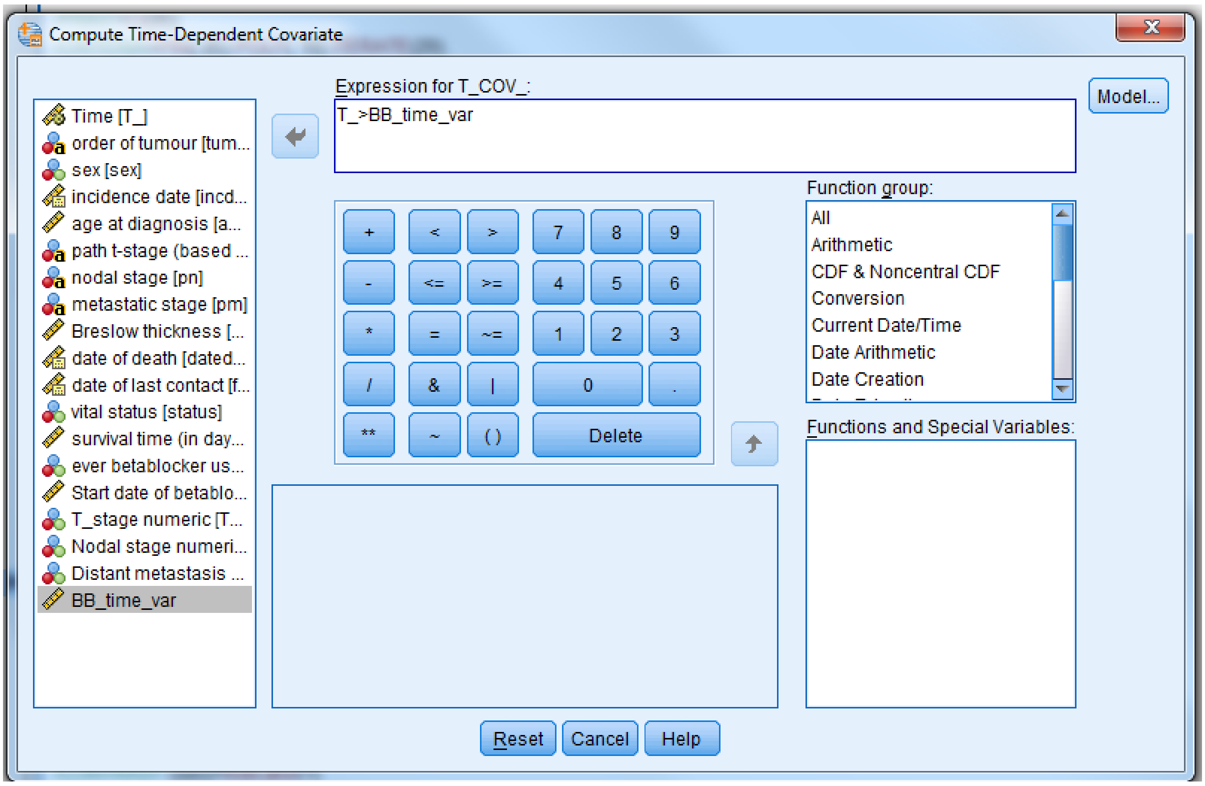
Step 2: Assign the value 1,000,000 to the non-users after diagnosis.

Create a new variable to prevent mistakes:

```
DO IF  MISSING(BB_start_after_diagnosis).
COMPUTE BB_time_var=1000000.
ELSE.
COMPUTE BB_time_var=BB_start_after_diagnosis.
END IF.
EXECUTE.
```

Step 3: Perform the Cox PH regression using time-varying covariate for drug use.

Analyze > Survival > Cox w/Time Dep Cov



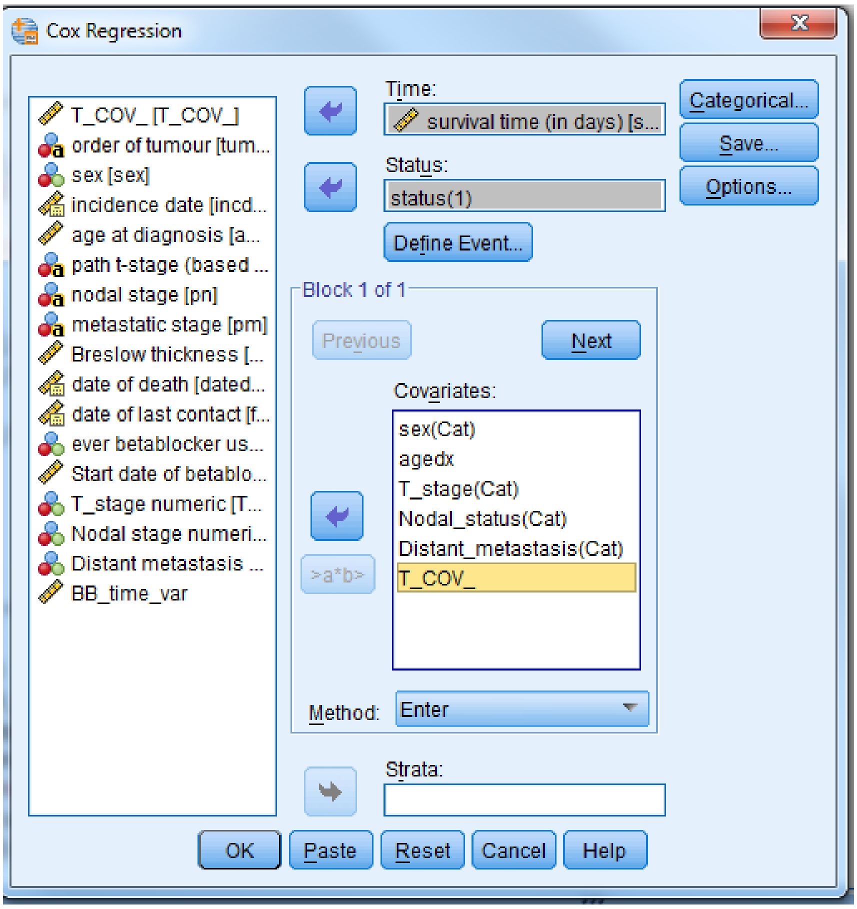
Expression for T_COV_: T_>BB_time_var

First, the time-dependent covariate has to be created.
T_>BB_time_var
is equal to

```
DO IF T_ > start_date_after_diagnosis.
COMPUTE BB_user=1.
ELSE.
COMPUTE BB_user=0.
END IF.
EXECUTE.
```

In words: a value 1 is returned at the moment a patient becomes a betablocker user. Before that time or if the patient doesn't use betablockers at all, a value 0 is returned.

Now go to model. Subsequently you see a familiar Cox Regression screen. T_COV is the time dependent covariate that we just created. Include this in the model, together with age, sex, and the TNM variables.



T_COV_ [T_COV_]
order of tumour [tum...
sex [sex]
incidence date [incd...
age at diagnosis [a...
path t-stage (based ...
nodal stage [pn]
metastatic stage [pm]
Breslow thickness [...
date of death [dated...
date of last contact [f...
ever betablocker us...
Start date of betablo...
T_stage numeric [T...
Nodal stage numeri...
Distant metastasis ...
BB_time_var

Time: survival time (in days) [s...

Status: status(1)

Covariates:
sex(Cat)
agedx
T_stage(Cat)
Nodal_status(Cat)
Distant_metastasis(Cat)
T_COV_

Method: Enter

Strata:

Important! The timescale of the time-varying covariate should be equal to the timescale of the time variable used for the Cox analysis. (eg. both days or both years).

Step 4: Interpret the output.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
sex	-,661	,084	61,714	1	,000	,516	,438	,609
agedx	,039	,003	183,580	1	,000	1,040	1,034	1,046
T_stage			165,348	3	,000			
T_stage(1)	,555	,206	7,228	1	,007	1,742	1,162	2,610
T_stage(2)	1,169	,191	37,377	1	,000	3,218	2,212	4,681
T_stage(3)	1,924	,198	94,367	1	,000	6,846	4,644	10,092
Nodal_status	,722	,129	31,380	1	,000	2,058	1,599	2,649
Distant_metastasis	1,460	,211	47,830	1	,000	4,307	2,848	6,515
T_COV_	,102	,095	1,155	1	,283	1,107	,919	1,334

T_COV is the binary time-dependent covariate for beta-blocker use after diagnosis (yes/no). The HR is 1.107 (95% CI 0.919-1.334), which is almost equal to the results of the time-fixed covariate analysis. The conclusion remains that betablocker use has no effect on survival.

How would you test the proportional hazards assumption for this time-dependent covariate. The solution can be found in the syntax file.

Suggested Readings:

Books

Petri A and Sabin C (2009), Medical Statistics at a Glance (easy, statistics)
Field A (2013), Discovering statistics using IBM SPSS statistics (easy, SPSS)
Katz MW (2011) Multivariable analysis: a practical guide for clinicians. (easy, multivariable analysis)
Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. (advanced, regression analysis)
Steyerberg EW (2009) Clinical prediction models: a practical approach to development, validation, and updating. (advanced, prediction modeling)
Kleinbaum DG and Klein M (2012) Survival analysis, a self-learning text (easy, survival analysis)

Website

Institute for Digital Research and Education. Statistical computing: http://www.ats.ucla.edu/stat (easy, output SPSS, STATA, SAS)

Articles

Wakkee M, Hollestein LM, Nijsten T (2014), Research Techniques Made Simple: Multivariable Analysis, JID (easy, multivariable analysis).

The Statistical Analysis and Methods in the Published Literature (SAMPL) guidelines, available from www.equator-network.org

Hollestein LM, Nijsten T (2015), Guidelines for statistical reporting in the British Journal of Dermatology, BJD 2015 Jul 173(1): 3-5