Statistical Methods with SPSS

(Statistics session for Staffs and PhD students)

Graduate School, Staffordshire University

(Asad) Md Asaduzzaman
Department of Engineering
md.asaduzzaman@staffs.ac.uk
www.mdasad.com

11th November, 2022

Preliminaries ...

Session plan

- SPSS introduction and some data manipulation techniques in SPSS
- Basic stats and tests of significance (one sample and two sample tests) in SPSS
- Correlation and multiple linear regression with SPSS
- Break (10 -15 mins)
- Analysis of variance (ANOVA): one-way, two-way and ANCOVA with SPSS
- Logistic regression with SPSS (if time permits!!)
- Hop-On, Hop-Off (... feel free to leave/join ...).

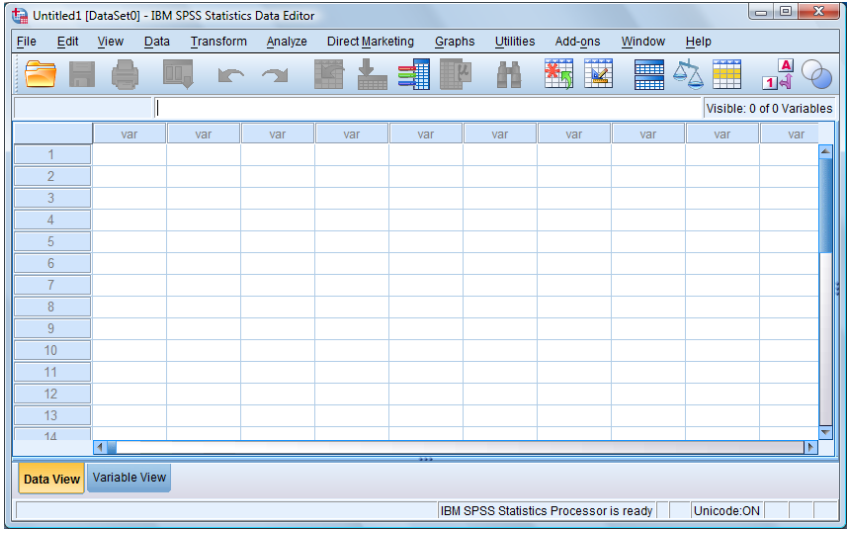
Expectation: Statistics knowledge (tests, correlation, regression, one-way/two-way ANOVA and logistic regression).

What is SPSS?

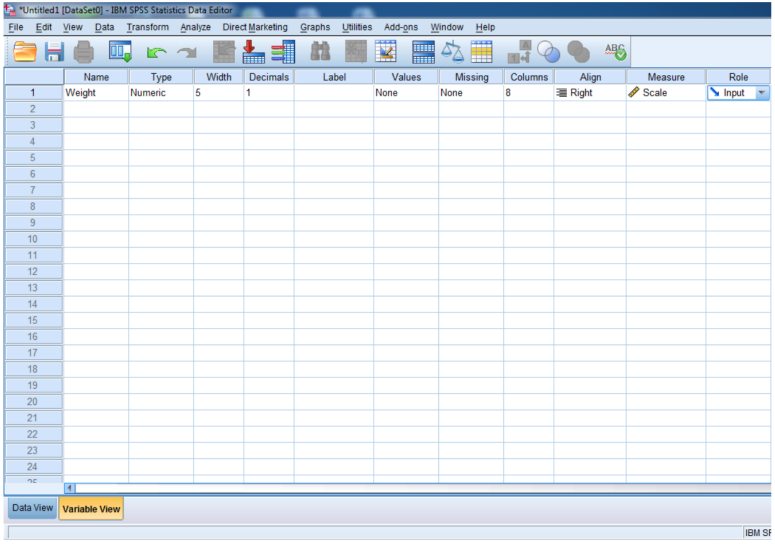
SPSS is short for Statistical Package for the Social Sciences, and it is used by various kinds of researchers for complex statistical data analysis. The SPSS software package was created for the management and statistical analysis of data.

- Data entry, reading or import and handling are very easy (Text, CSV, Excel files can be imported easily)
- Many built-in data manipulation tools such as computing, recoding or transforming variables, split files
- Advanced statistical analysis, model fitting and model diagnostics can be performed easily.
- Output can be imported or transferred easily into word or word-processing softwares.
- Staffordshire University has the full version of SPSS, and the software licence is updated every year.

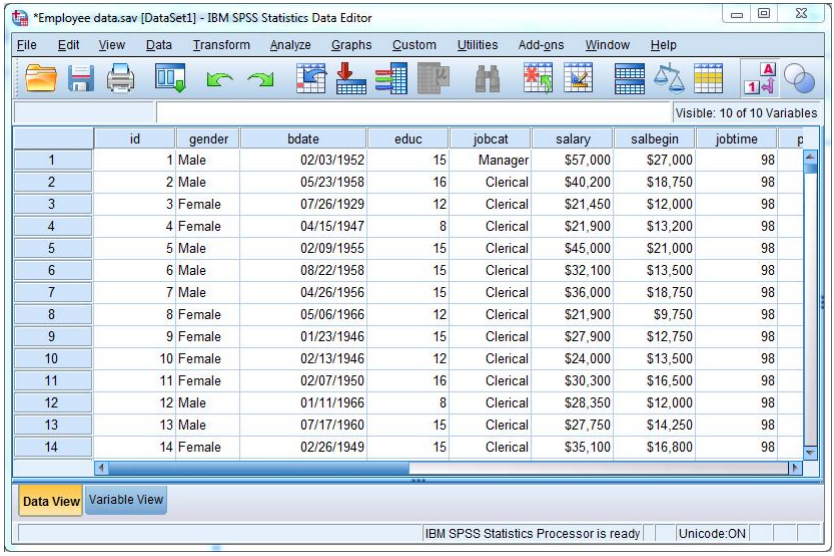
SPSS blank data editor - Data view



SPSS blank data editor - Variable view



SPSS data file view



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98	
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98	

A quick demo on SPSS.

Some basic operations in SPSS

- Other two important windows are: Syntax window and Output window. Many prefers to use menu bars rather than syntax.
- Data manipulation can be performed easily with lots of options. Some frequently used options are: Compute, Recode, Select Cases, Split File, etc. These options can be found under menu: Data, Transform. (a quick demo)
- Analysis menus can be found under the tab: Analyze
- Our focus today
  - Basic stats and tests: Analyze → Descriptive Statistics/Compare Means
  - Correlation: Analyze → Correlate → Bivariate
  - Linear regression: Analyze → Regression → Linear
  - ANOVA (1-way): Analyze → Compare Means → One-Way ANOVA
  - ANOVA (2-way & ANCOVA): Analyze → General Linear Model → Univariate
  - Logistic regression: Analyze → Regression → Binary Logistic

Descriptive stats and tests of significance with `Employee.sav` data

`Employee.sav` datafile contains information on (474 employees):

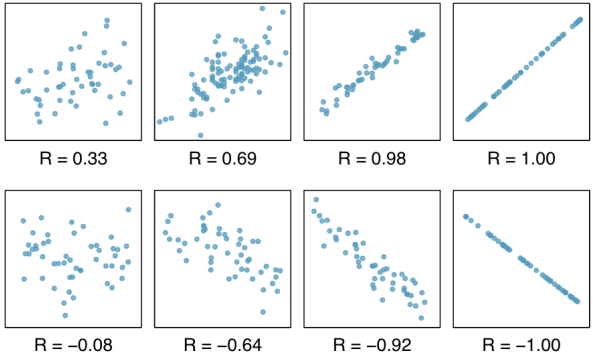
- `id, gender, birth date, education level (in single years), job category (managerial, clerical, custodial), current salary, beginning salary, months since hire, previous experience in months, minority classification (yes/no)`

Basic stats and tests

- Frequency tables
- Cross tables and Chi-square test
- One-sample t-test
- Two independent sample t-test
- Paired sample t-test

Correlation & Regression

Correlation: Simply measures the strength of association between variables. In statistical terms, correlation (r) denotes linear relationship between two quantitative variables. If one increases the other will also increase or decrease or vice-versa. Varies between -1 and +1. Scatter diagram is a useful visual tool to explore correlation.



Correlation ... continued

Some basic examples of correlation:

- age and height
- advertisement spending and product sell
- amount of fertiliser use and crop yield
- IQ score and exam mark
- car mileage and car price
- item price and their demand

Correlation ... continued

A quick demo of correlation into SPSS with `Employee.sav` dataset.

The datafile contains information on (474 employees):

- `id, gender, birth date, education level (in single years), job category (managerial, clerical, custodial), current salary, beginning salary, months since hire, previous experience in months, minority classification (yes/no)`

Scatter plot: Graphs → Legacy Dialogs → Scatter/Dot

Correlation: Analyze → Correlate → Bivariate

Correlation ... Scatter plot

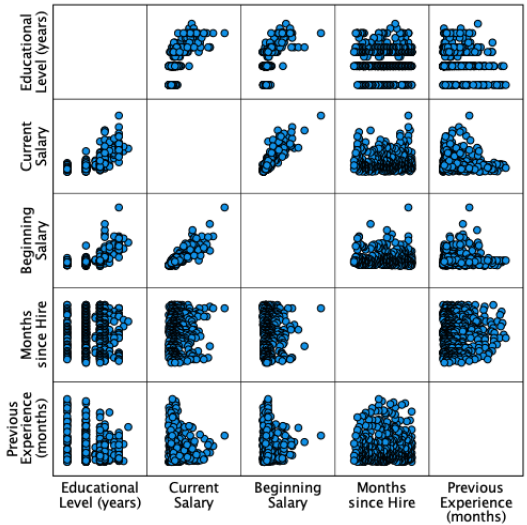
- Explore the relationship between variables: `education level, current salary, beginning salary, months since hire, previous experience in months`
- Plot scatter diagrams
- Obtain the correlation coefficients
- Check whether correlation coefficients are significant

SPSS demo on scatter plot and correlation:

Scatter plot: Graphs → Legacy Dialogs → Scatter/Dot

Correlation: Analyze → Correlate → Bivariate

Correlation matrix



Correlation ... test of significance

Correlations

		Educational Level (years)	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Educational Level (years)	Pearson Correlation	1	.661**	.633**	.047	-.252**
	Sig. (2-tailed)		.000	.000	.303	.000
	N	474	474	474	474	474
Current Salary	Pearson Correlation	.661**	1	.880**	.084	-.097*
	Sig. (2-tailed)	.000		.000	.067	.034
	N	474	474	474	474	474
Beginning Salary	Pearson Correlation	.633**	.880**	1	-.020	.045
	Sig. (2-tailed)	.000	.000		.668	.327
	N	474	474	474	474	474
Months since Hire	Pearson Correlation	.047	.084	-.020	1	.003
	Sig. (2-tailed)	.303	.067	.668		.948
	N	474	474	474	474	474
Previous Experience (months)	Pearson Correlation	-.252**	-.097*	.045	.003	1
	Sig. (2-tailed)	.000	.034	.327	.948	
	N	474	474	474	474	474

Regression analysis ... introduction

Regression analysis is a statistical technique for investigating and modelling the relationship between variables; more specifically, estimating the effect of a set of variables (explanatory or independent variables) on the response variable (dependent variable).

Applications of regression are numerous and occur in almost every field, including:

- engineering
- physical sciences
- economics
- business & management
- biological sciences
- social sciences

In fact, regression analysis is one of the most widely used statistical techniques.

Typical regression analysis examples

One may be interested in estimating the effect of

- age on height (plants, human being)
- advertisement spending on product sell
- amount of fertiliser use on crop yield
- IQ score on exam mark
- car mileage on car price
- item price on their demand

In multiple linear regression, you may want to estimate the effect of several variables simultaneously.

Typical regression analysis examples ...

However, the analysis, particularly, the choice of variables (dependent and the set of explanatory variables) will depend on your specific research objectives. For example, age is a common explanatory variable in medical studies.

However, age may be the dependent variable in many cases. For instance,

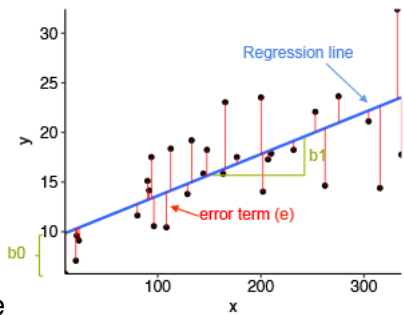
- botanist may be interested in predicting age of trees based on their heights and other factors
- archaeologist may want to determine the age of a historic site based on a number of explanatory variables

Regression model ... setup

A multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

$Y \rightarrow$ dependent/response variable
$X_1, X_2, \ldots x_n \rightarrow$ independent/explanatory variables
$\beta_0 \rightarrow$ intercept
$\beta_1, \beta_2, \ldots, \beta_n \rightarrow$ slopes or effect of the variables
$\epsilon \rightarrow$ error term



Assumptions:

- liner relationship
- $\epsilon \rightarrow$ error terms are independent, normally distributed with mean 0 and constant variance

Regression analysis strategy

First step:

- investigate the relationship among the variable, particularly, the response variable with other variables using correlation analysis
- if you find a reasonable indication that a linear regression of the response variable with other variables is suitable then perform the analysis

Second step:

- Perform the regression analysis selecting the variables (response and explanatory) appropriately
- Check the results whether the error assumptions are met (independent or scattered, constant variance)
- Normality of the errors
- Any multicollinearity among explanatory variables? Any outlier? Any influential observation?
- Find the best set of explanatory variables (only significant variables - final model)

Regression example ... `Employee.sav` data

Information on:

- `id, gender, birth date, education level (in single years), job category (managerial, clerical, custodial), current salary, beginning salary, months since hire, previous experience in months, minority classification (yes/no)`

Our research question is:

- Are the factors: gender, education level, job category, beginning salary, months since hire, previous experience in months, minority classification significant for **salary change**?
- If so, can we predict the salary change for a person based on the set of explanatory variable values for that person.

Regression analysis with `Employee.sav` data

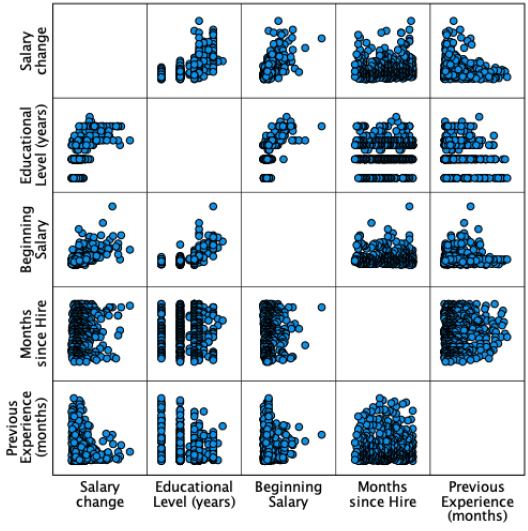
- First, compute salary change (`salchange`) = Current salary - Beginning salary

Descriptive statistics of `salchange`:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Salary change	474	5550.00	76240.00	17403.4810	10814.6200
Valid N (listwise)	474				

Explore the relation between variables



Testing significance of correlation coefficients

Correlations

		Salary change	Educational Level (years)	Beginning Salary	Months since Hire	Previous Experience (months)
Salary change	Pearson Correlation	1	.582**	.662**	.147**	-.187**
	Sig. (2-tailed)		.000	.000	.001	.000
	N	474	474	474	474	474
Educational Level (years)	Pearson Correlation	.582**	1	.633**	.047	-.252**
	Sig. (2-tailed)	.000		.000	.303	.000
	N	474	474	474	474	474
Beginning Salary	Pearson Correlation	.662**	.633**	1	-.020	.045
	Sig. (2-tailed)	.000	.000		.668	.327
	N	474	474	474	474	474
Months since Hire	Pearson Correlation	.147**	.047	-.020	1	.003
	Sig. (2-tailed)	.001	.303	.668		.948
	N	474	474	474	474	474
Previous Experience (months)	Pearson Correlation	-.187**	-.252**	.045	.003	1
	Sig. (2-tailed)	.000	.000	.327	.948	
	N	474	474	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

Tests for other categorical variables

- Perform t-test whether salary change is significant for gender and minority (two category variable) (some people use point biserial correlation)
- Perform one-way ANOVA for to test whether salary change is significant for job category

Performing ML regression analysis - variable setup in SPSS

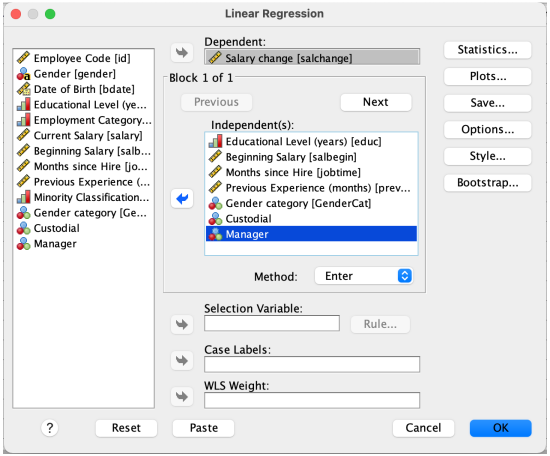
Categorical variables need to re-generated as dummy variables:

- Variables with two categories like gender and minority can be easily by coded as 1 and 0.
- Job category variable has three categories. Therefore, two dummy variables need to be created:

	Dummy Variables	
	Custodial	Managerial
Clerical (baseline category)	0	0
Custodial	1	0
Managerial	0	1

Performing ML regression analysis– a demo with `Employee.sav`

Linear regression: Analyze → Regression → Linear



Regression SPSS Demo ... with `Employee.sav`

Look at:

- Model Summary
- ANOVA table
- Coefficients

Further checks on:

- error assumptions: independent or scattered, constant variance
- normality of the errors
- any multicollinearity
- any outlier or influential observation
- variable selection

Regression SPSS Demo ... with `Employee.sav`

Issues:

- Errors are not scattered
- Variance is not constant
- Error distribution is not normal

Remedial measures:

- There are several ways to solve these issues
- One simple way is to make a transformation of the response variable (salary change)
- we will perform a natural logarithm transformation and re-perform the analysis

Note: No multicollinearity is observed as VIF for all variables found to be between 1 and 10. VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Regression SPSS Demo ... with `Employee.sav`

Identifying outliers: if the error is too high for an observation

Identifying influential observations:

- Difference in Fits (DFF): An observation is deemed influential if the absolute value of its DFF value is greater than:

$$2\sqrt{\frac{k+2}{n-k-2}} = 2\sqrt{\frac{7+2}{474-7-2}} = 0.0387$$

  $k \rightarrow$ no. of explanatory variables and $n \rightarrow$ no. of total observations
- Cook's distance: if greater than 0.5, then it may be influential, if greater than 1 or far apart from other values, then it quite likely to be influential.
- Leverage: A common rule is to flag any observation whose leverage value is 3 times larger than the mean leverage value: p/n = 7/474 = 0.0148 ($\times 3 = 0.0444$).

Choosing the best set of explanatory variables

SPSS options:

- Enter: forces all variables to be in the model
- Stepwise: removing the weakest correlated variable
- Remove: all variables in a block are removed in a single step.
- Backward: all variables are entered into the equation and then sequentially removed based on the smallest partial correlation
- Forward: adding variables based on the highest correlation/partial correlation

A demo with different selection method for `Employee.sav` data.

ANOVA

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of three or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. In one-way ANOVA, we consider only one factor (with three or more categories).

Some examples:

- whether different variety of crops give different amount of production
- whether different levels of factors affect plants and wildlife
- whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales
- whether or not different medications affect patients differently

ANOVA assumptions

Assumptions:

- Independence of observations
- Normally-distributed response variable
- Homogeneity of variance

If the assumptions are not satisfied, we can perform non-parametric approaches.

One-way ANOVA - `Diet.sav` dataset

Variables

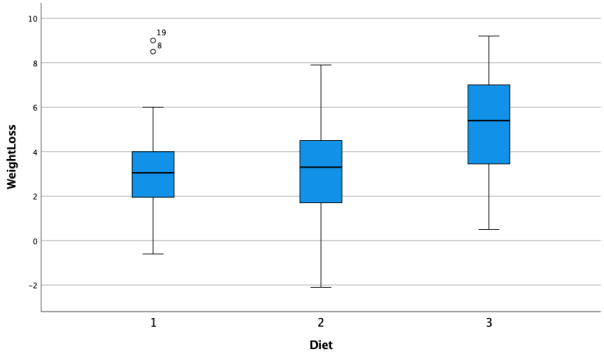
- `Person, Gender, Age, Height, Preweight, Diet, Weight6weeks`

One-way ANOVA: we shall now consider only diet and weight loss (= Preweight - Weight6weeks). Also, we will think that experiment was conduct with a homogeneous cohort of people (no other extraneous source of variation involved).

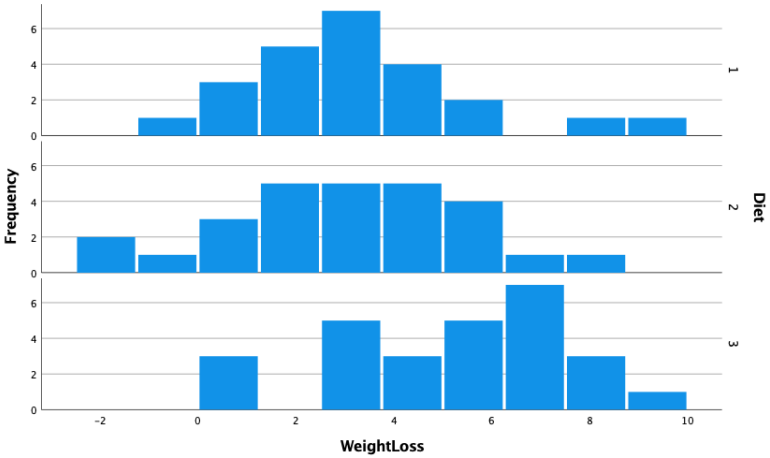
Research interest:

- Our main goal is to find whether the different diets have impacted weight loss. More specifically, whether the different diets impacted differently
- If so, then which diet is different than others

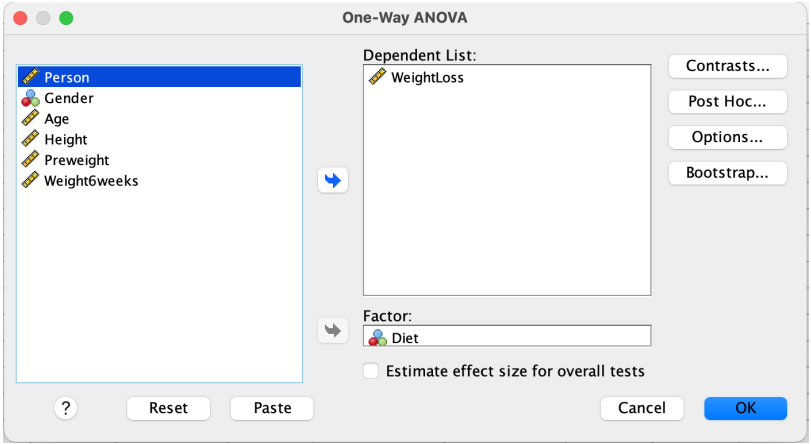
Pre-analysis descriptives and assumption checks



Pre-analysis descriptives and assumption checks



ANOVA (1-way): Analyze → Compare Means → One-Way ANOVA



Demo of ANOVA with `Diet.sav` data

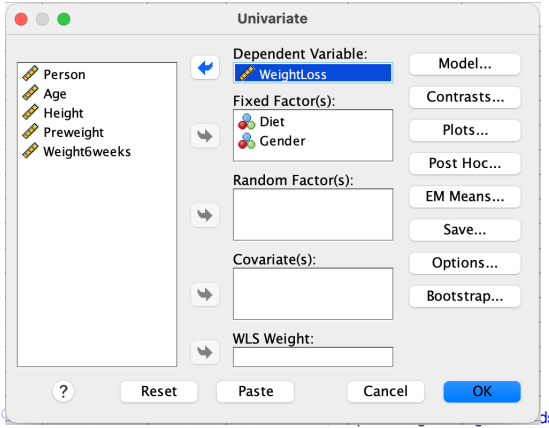
ANOVA outputs:

- Tests of homogeneity of variances
- ANOVA table
- Post Hoc test table (multiple comparison)

Two-way ANOVA with `Diet.sav` data ...

In two-way ANOVA, you add (believe) an extra source of variation, which is referred to as “blocking” effect (gender is added as block)

ANOVA (2-way): Analyze → General Linear Model → Univariate



Two-way ANOVA ...

Options:

- Model (selection)
- Plots (interaction plot)
- Post Hoc tests (between different levels of the factor)
- Options (residual plot)

Note: You still need to check the assumption (normality and homogeneity of variance) like you have done for one-way ANOVA.

Two-way ANOVA with `Diet.sav` data ...

Outputs:

- ANOVA table (Tests of between-subjects effects)
- Estimated marginal means
- Post Hoc Test
- Profile plot (for interaction checking)

ANCOVA

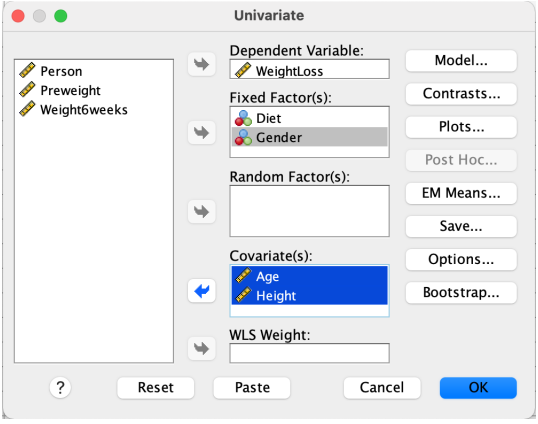
ANCOVA is similar to traditional ANOVA but is used to detect a difference in means of three or more independent groups, whilst controlling for scale covariates.

Difference with MLR: the research objective and data collection (in MLR concentration is on all explanatory factors and data collected come from many ways, whereas in ANCOVA still you look for the effect of the study factor and data are generated through experiments).

- Performed exactly the same way as you have seen in two-way ANOVA ANOVA (2-way): Analyze $\rightarrow$ General Linear Model $\rightarrow$ Univariate
- model assumptions, selection, SPSS options exactly the same, other than selecting the covariables

ANCOVA ...

ANCOVA: Analyze → General Linear Model → Univariate



ANCOVA ... demo

A quick demo of ANCOVA with `Diet.sav` data ...

Logistic regression - introduction

In MLR, we have seen:

- the response variable (dependent) is continuous and
- takes values between $-\infty$ and $+\infty$

Now consider that you want to find the significant factors associated with

- developing lung cancer (Yes/No) – age, gender, ethnicity, occupation, smoking status, family history, etc.
- customers would default (Yes/No) – age, gender, ethnicity, occupation, income group, number of family members, credit history, etc.
- preference of apple's iPhone (Yes/No) – age, gender, ethnicity, occupation, income group, region, etc.

Logistic regression ...

In each example, the response variables has two outcomes: Yes and No. Therefore, the MLR cannot be applied.

- We can apply the logistic regression model
- The model is, more specifically, referred to as "binary logistic regression model"
- The functional form of the model is given by:

$$\mathbb{P}(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$$

- We don't have to understand the complex form, but it is worth noting that the "Yes" and "No" are modelling through some probabilistic mechanism

Logistic regression ...

Some good aspects

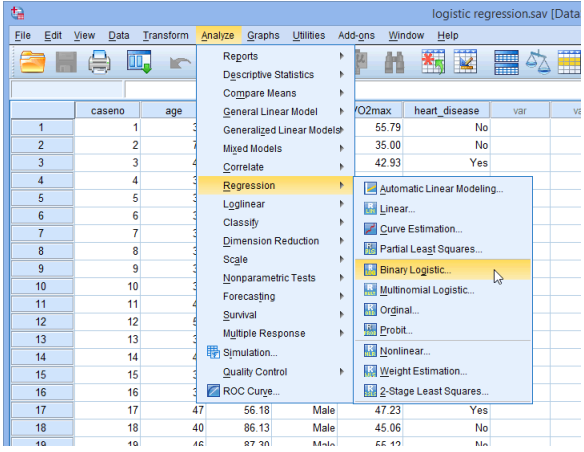
- minimal assumption unlike multiple linear regression (MLR)
- easy way of interpretation of parameters using odds ratios
- SPSS implementation is much easier, even easier than MLR
- significance testing of factors and model selection options in SPSS are similar to MLR (though mathematical setup are different)

A quick example ... heart disease incidence

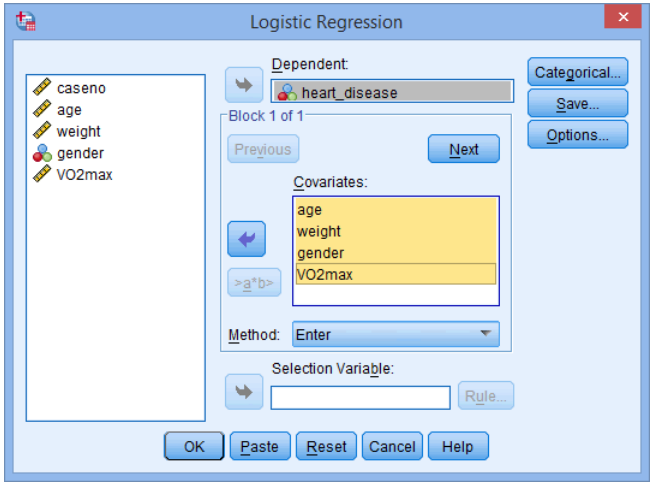
- Response variable: incidence of heart disease – Yes/No
- Explanatory variables: age (in years), weight (in Kg), gender (male - 1/female - 0), VO2max (maximal aerobic capacity)

A quick example ... heart disease incidence

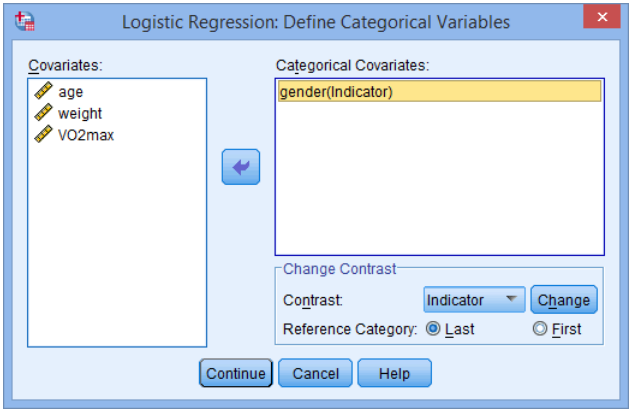
Logistic regression: Analyze → Regression → Binary Logistic



A quick example ... heart disease incidence



A quick example ... heart disease incidence



A quick example ... heart disease incidence

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

Logistic regression - SPSS demo with `CHD.sav` dataset

Logistic regression: Analyze → Regression → Binary Logistic

- Run the model
- See the model result with “Enter method”
- Find the most suitable model using forward/backward - conditional/LR/Wald

Learning summary and practical

- Basics on SPSS (brief)
- Basic stats, tests of significance, correlation and ML regression
- ANOVA (one and two-way) and ANCOVA
- Logistic regression

Practice:

- Three datasets: `Employee.sav`, `Diet.sav` and `CHD.sav` are available through graduate school
- Have some practice the same you've learned today
- Perform basic stats, tests, correlation analysis and develop a regression model for salary increase with the relevant explanatory variables
- Perform one and two-way ANOVA and ANCOVA with the research questions we have discussed for diet data
- Perform a logistic regression analysis to identify factors for heart disease and find the most suitable model with CHD data.

References

- Arbuckle, J. L. (2020). IBM SPSS Amos 27 User's Guide. Amos Development Corporation.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.
- Montgomery, D. C. (2017). Design and analysis of experiments. John Wiley & sons.
- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.
- Many online lecture notes, websites and resources (from where images, texts and datasets are taken).

Many thanks for attending the session

Please feel free to question/comment

md.asaduzzaman@staffs.ac.uk
www.mdasad.com
www.staffs.ac.uk/people/md-asaduzzaman