Introduction

Welcome to Advanced Statistics course. This is a major course for the Doctor of Philosophy in Nursing program aimed. The course reviews important concepts in statistics and presents their applications in nursing situations. This course equips you with the necessary skills and knowledge on data processing, data analysis, and data presentation. The focus is on advanced statistics that doctoral students would need in preparation for their dissertation.

The course is divided into three units. Unit 1, Review of Basic Statistics, is about descriptive statistics – how to organize and present data, such that they can be summarized and interpreted more meaningfully and it includes basic principles in inferential statistics. Unit 2, Common Inferential Statistical Methods, showcase statistical applications common in nursing, which includes tests of differences, and tests of relationships. Unit 3, Multivariate Analysis, is about statistical techniques that are useful in modeling techniques especially when testing or validating theories or models, which are required in the PhD in nursing program.

Course outcomes

At the end of the course, students should be able to:

1. Explain basic concepts and principles in statistics;
2. Demonstrate application of common inferential statistics used in nursing;
3. Perform advanced statistical methods and analysis given a data set; and
4. Critique data analysis and interpretations of select statistical applications in nursing studies.

Content outline

I. Review of Basic Statistics
   1. Basic statistical concepts
   2. Frequency tables
   3. Measures of central tendency and variability
   4. Probability and sampling
   5. Estimation
   6. Testing of hypothesis

II. Common Inferential Statistical Methods
   1. Test of independence
   2. Analysis of variance
   3. Correlation
   4. Regression

III. Multivariate Analysis
   1. Multivariate analysis of variance
   2. Multiple regression
   3. Logistic regression
   4. Factor analysis
   5. Path analysis
   6. Structural equation modeling

UNIT 3. Advanced Statistics

Organizing Principle

Advanced statistics includes statistical tools and techniques that are needed for multivariate analysis whether finding or predicting associations / relationships, determining significant factors or model building considering different levels of measurement of data and sampling distributions.

This Unit explores advanced statistics used in nursing theory and model development. This includes multivariate analysis of variance, multiple regression, logistic regression, factor analysis, path analysis and structural equation modeling.

Learning Objectives

After working on this module, you should be able to:

1. Apply tests of independence given parametric and statistic data;
2. Compare more than two groups (or characteristics of a group) using analysis of variance;
3. Infer significant relationships between variables using correlational techniques; and
4. Predict relationships between variables using linear regression;

Underlying Concepts

- Multivariate analysis of variance
- Multiple regression
- Logistic regression
- Factor analysis
- Path analysis
- Structural equation modeling

Module 11. Multivariate analysis of variance

Multivariate analysis of variance (MANOVA) is a procedure for comparing multivariate sample means. As a multivariate procedure, it is used when there are two or more dependent variables. and is often followed by significance tests involving individual dependent variables separately.

This module also includes other types of ANOVA involving different conditions between the dependent and independent variables.

Learning Objectives

After working on this module, you should be able to:

1. Describe the different types of analysis of variance;
2. State the assumptions required in testing multivariate analysis of variance and other types of ANOVA;
3. Compute for the statistics in multivariate analysis of variance and other types of ANOVA
4. Interpret the results of multivariate analysis of variance and other types of ANOVA

11.1 MANOVA

MANOVA used when there are two or more dependent variables, uses the covariance between outcome variables in testing the statistical significance of the mean differences.

For example:

Nurse Betty became interested whether exercise can not only lead to weight loss but also improved self-esteem.

- What are her hypotheses?
- What statistics to use?
- How to interpret the results?

- Do changes in the independent variable(s) have significant effects on the dependent variables?
- What are the relationships among the dependent variables?
- What are the relationships among the independent variables?

11.2 Two-way ANOVA

- Two-way ANOVA - when the dependent variable is affected by two independent variables/factors
- Not only assessing the main effect of each independent variable but also if there is any interaction between them.

For example:

Nurse Betty wanted to find out whether age group is a factor for patients enrolled in the exercise program. What if the exercise weight loss program affect different age groups?

- What are her hypotheses?
- What statistics to use?
- How to interpret the results?

Possible questions that can be answered:

- Is exercise the main factor affecting weight loss? In other words, do groups subjected to different exercise differ significantly in their weight loss?
- Is age the main factor affecting weight loss? In other words, do patients of different age differ significantly in their weight loss?
- Is there a significant interaction between the factors? In other words, how do age and exercise interact with regard to a patient's weight loss? For example, it might be that older people reacted differently to exercise.
- Can any differences in one factor be found within another factor? In other words, can any differences in exercise and weight loss be found in different age groups?

11.3 Repeated Measures ANOVA

- Repeated Measures ANOVA - compares means across one or more variables that are based on repeated observations.

For example:

- Nurse Betty wanted to find out whether doing the weight loss program over and over would achieve the same results. What if the weight loss program is more effective given a particular time duration?
    - What are her hypotheses?
    - What statistics to use?
    - How to interpret the results?

Assumptions

- Normality—For each level of the within-subjects factor, the dependent variable must have a normal distribution.
- Sphericity— Difference scores computed between two levels of a within-subjects factor must have the same variance for the comparison of any two levels. (This assumption only applies if there are more than 2 levels of the independent variable.)
- Randomness—Cases should be derived from a random sample, and scores from different participants should be independent of each other.
- Multivariate normality—The difference scores are multivariately normally distributed in the population.
- Randomness—Individual cases should be derived from a random sample, and the difference scores for each participant are independent from those of another participant.

11.4 ANCOVA

- ANCOVA - evaluates whether the means of a dependent variable (DV) are equal across levels of a categorical independent variables (IV) often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates (CV).

For example:

- Nurse Betty wanted to find out whether AGE is a factor for patients enrolled in the weight loss programs. What if the weight loss program affect different age groups?
    - What are her hypotheses?
    - What statistics to use?
    - How to interpret the results?

- **Assumptions:**
    - Independent variables (minimum of two) should be categorical variables.
    - The dependent variable and covariate should be continuous variables.
    - Make sure observations are independent.
    - The dependent variable should be roughly normal for each of category of I.V.
    - Data should show homogeneity of variance.
    - The covariate and dependent variable (at each level of independent variable) should be linearly related.
    - Your data should be homoscedastic of Y for each value of X.
    - The covariate and the independent variable shouldn't interact. In other words, there should be homogeneity of regression slopes.

Module 12. Multiple regression

Finding predictors of certain outcome variables in nursing and health studies usually involve more than one independent variable. This would entail a regression analysis capable of handling more than one predictor variable. This module discusses the uses of multiple linear regression analysis, the assumptions in using it, the computations and the testing of statistical significance of the results.

Learning Objectives

After working on this module, you should be able to:

1. Describe the use of Multiple Linear Regression Analysis;
2. Discuss the assumptions in using Multiple Linear Regression Analysis;
3. Compute for estimates of Multiple Linear Regression Analysis;
4. Test the statistical significance of regression equation;
5. Interpret results of Multiple Linear Regression Analysis; and
6. Discuss issues related to use of Multiple Linear Regression Analysis

12.1 Uses of Multiple Linear Regression Analysis

There are several uses of multiple linear regression analysis. Read "Applications of Multiple Regression Analysis" from the following resource:

Kuan, L.G., & Bonito, S.R. & UPOU (2001). *N298: Statistical methods applied in nursing*. University of the Philippines Open University, Manila. pp 267-286.

Activity 12-1

Give examples of the following uses of multiple linear regression analysis

Use of Multiple Linear Regression	***Example***
1. Determine correlation of variables	
2. Determine the relationship of one or more independent variable (Xs) to Y, controlling for confounder	
3. Predict the value of Y given the values of independent variables (Xs)	
4. Select the important variables (Xs) in predicting Y	

12.2 Assumptions in Multiple Linear Regression Analysis

Determine the assumptions in doing inferential statistics in multiple linear regression analysis by reading "Assumptions in Multiple Regression Analysis" from the following resource:

Kuan, L.G., & Bonito, S.R. & UPOU (2001). *N298: Statistical methods applied in nursing*. University of the Philippines Open University, Manila. pp 267-286.

Activity 12-2

Identify the assumption being addressed by the following questions:

Questions	***Assumption/s***
1. Are the data points independent?	
2. Do the combinations of data have finite mean and variance?	
3. Is the variability between Y and Xs the same?	
4. Does the scatterplot produced by the model follow the Gaussian distribution?	
5. Does the relationship between each X variables and Y follow a straight line?	

12.3 Computing Estimates in Multiple Linear Regression Analysis

Read "Multiple Regression Models", "Approaches to Estimating Parameters" and "Assessing Multiple Regression Equation" from the following resource:

Kuan, L.G., & Bonito, S.R. & UPOU (2001). *N298: Statistical methods applied in nursing*. University of the Philippines Open University, Manila. pp 267-286.

Study Guide Questions

1. What is the meaning of F ratio in interpreting results in regression equation?
2. What is the meaning of R^2 in interpreting results in regression equation?

Activity 12-3

A study was conducted to find the predictors of infant birthweight. The variables of interest were mother's age, mother's number of pregnancies, mother's weight, and mother's smoking status.

Given the above variables, present the following models:

1. A model showing mother's age as confounder for the relationship between mother's number of pregnancies and infant birthweight
2. A model showing an interaction between mother's weight and mother's smoking status and infant birthweight
3. A model showing all the above variables as predictors of infant birthweight

12.4 Finding Statistical Significance in Multiple Linear Regression Analysis

Read "Testing Hypothesis in Multiple Regression Models" from the following resource:

Kuan, L.G., & Bonito, S.R. & UPOU (2001). *N298: Statistical methods applied in nursing*. University of the Philippines Open University, Manila. pp 267-286.

Study Guide Questions

1. *What is the use of overall F-test in multiple linear regression analysis?*
2. *What is the use of partial F-test in multiple linear regression analysis?*

12.5 Issues Related to Multiple Linear Regression Analysis

Read "Issues Related to Multiple Regression Models" from the following resource:

Kuan, L.G., & Bonito, S.R. & UPOU (2001). *N298: Statistical methods applied in nursing*. University of the Philippines Open University, Manila. pp 267-286.

Study Guide Questions

1. What is the issue in coding nominal variables in multiple linear regression analysis?
2. What are the different methods in selecting variables in multiple linear regression?
3. What is the implication of having too many independent variables in small sample size?

Module 13. Logistic regression

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, *Y,* from one or more response variables, *X*. However, unlike linear regression the response variables can be categorical *or* continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative **maximum likelihood** method rather than a least squares to fit the final model. This means the researcher has more freedom when using LR and the method may be more appropriate for nonnormally distributed data or when the samples have unequal covariance matrices. Logistic regression assumes independence among variables.

Learning Objectives

- Describe the uses of logistic regression
- Identify assumptions needed for logistic regression
- Perform logistic regression
- Interpret the results of logistic regression

13.1 Use of logistic regression

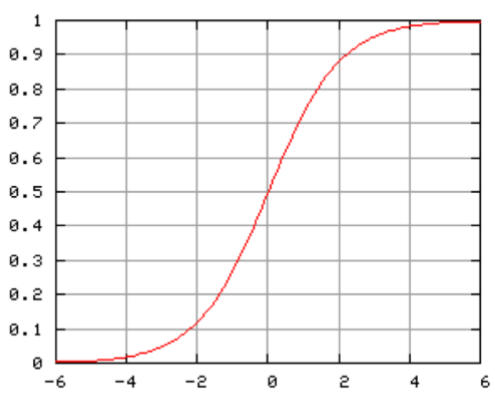
- Used to model dichotomous or binary outcomes (but not limited to) using predictor variables.
- Used when the research method is focused on whether or not an event occurred.
- Instead of modeling the outcome (Y) directly, the method models the log odds(Y) using the logistic function
- Estimate adjusted prevalence rates, adjusted for potential confounders (sociodemographic or clinical characteristics)
- Estimate the effect of a treatment on a dichotomous outcome, adjusted for other covariates
- Explore how well characteristics predict a categorical outcome

13.2 Type of data required

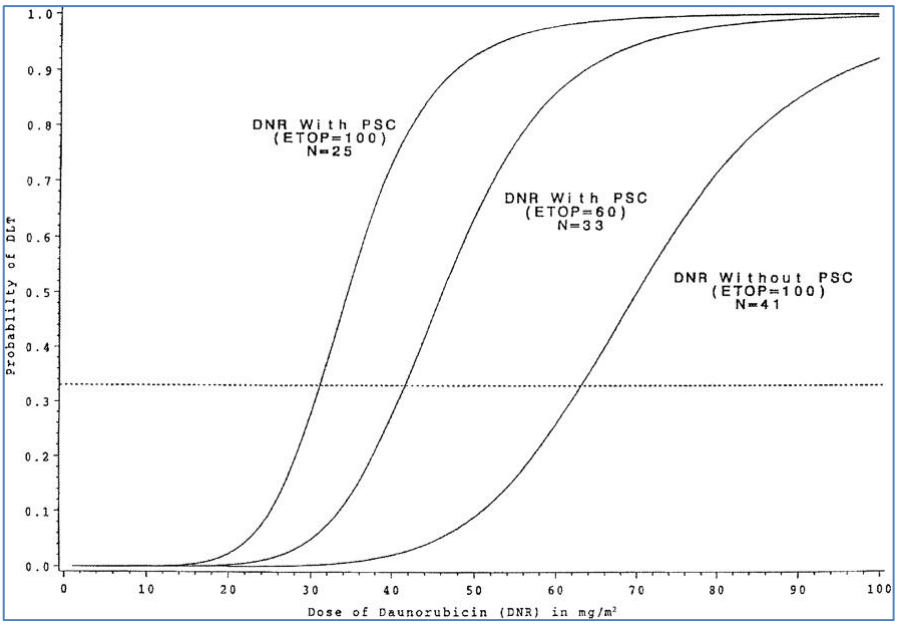
- Dependent variable
  - Dichotomous variable, may only assume two discrete values (0 and 1)
- Independent Variable
  - One or more
  - Can be any scale
  -

13.3 The Logistic Curve

$$LOGIT(p) = \ln\left(\frac{p}{(1-p)}\right) = z \quad \Leftrightarrow \quad p = \frac{\exp(z)}{1+\exp(z)}$$



An example:



Logistic regression curves for the three drug combinations. The dashed reference line represents the probability of DLT of .33. The estimated MTD can be obtained as the value on the horizontal axis that coincides with a vertical line drawn through the point where the dashed line intersects the logistic curve.

13.4 Assumptions

- Y_i are from Bernoulli or binomial (n_i, m_i) distribution
- Y_i are independent
- Log odds $P(Y_i = 1)$ or logit $P(Y_i = 1)$ is a linear function of covariates

13.5 The Logistic Regression Model

- **Simple** logistic regression = with one predictor variable
- **Multiple** logistic regression = with multiple predictor variables

Logistic Regression:

$$\ln\left(\frac{\mathrm{P(Y)}}{\mathrm{1\text{-}P(Y)}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

Linear Regression:

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$$

Using this estimation gives model coefficient estimates that are asymptotically consistent, efficient, and normally distributed.

Thus, a 95% Confidence Interval for b_K is given by:

$$\hat{b}_K \pm z_{a/2}\left(SE_{\hat{b}_K}\right)$$

$$= (L\ ,\ U)$$

- The Odds Ratio for the k[th] model coefficient is:

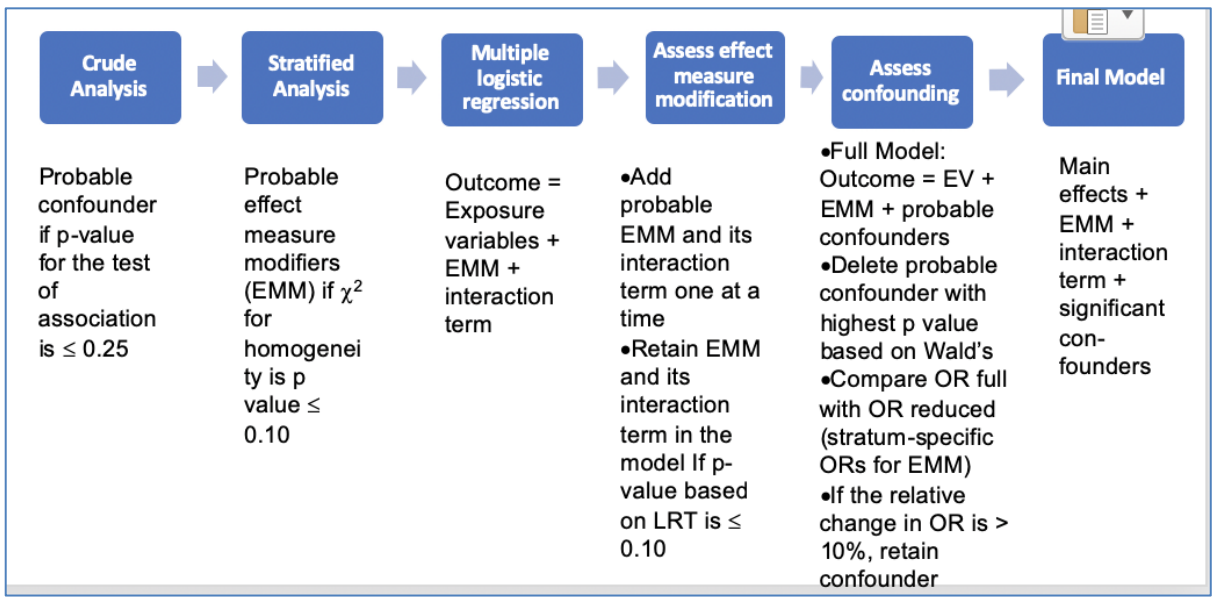
$$\hat{\mathrm{OR}} = \exp\left(\hat{b}_K\right)$$

- We can also get a 95% CI for the OR from:

$$= \left(e^{L},\ e^{U}\right)$$

$$\text{where } (L\ ,\ U) \text{ is a 95\% CI for } b_K$$

13.6 Performing Logistic Regression



Example:

A study examined the effect of the mother's age, along with clinical characteristics, on the odds of pregnancy success on the first Assisted Reproductive Technology (ART) attempt

The Logistic Regression Model

$$\ln\left(\frac{\Pr(\text{pregnancy})}{1-\Pr(\text{pregnancy})}\right) = 2.67 - 0.13 * \text{Age}$$

$$\Updownarrow$$

$$\Pr(\text{pregnancy}) = \frac{\exp(2.67 - 0.13 * \text{Age})}{1 + \exp(2.67 - 0.13 * \text{Age})}$$

Question

1. What is the effect of Age on Pregnancy?

   Answer:

   $$\hat{\text{OR}}_{\text{Age}} = \exp(-0.13) = 0.88$$

   This implies that for every 1 yr. increase in age, the odds of pregnancy decrease by 12%.

2. What is the predicted probability of a 25 yr. old having pregnancy success with first ART attempt?

   $$\hat{\Pr}(\text{pregnancy}) = \frac{\exp(2.67 - 0.13 * 25)}{1 + \exp(2.67 - 0.13 * 25)} = 0.359$$

   From this model, a 25 yr. old has about a 36% chance of pregnancy success.

13.7 Hypothesis testing

- Usually interested in testing $H_0 : \beta_K = 0$
- Two types of tests we'll discuss:
  1. Likelihood Ratio test
  2. Wald test

Likelihood Ratio Test

- Idea is to compare the (log) Likelihood of two models to test

- Two models:
  - Full model = with predictor included
  - Reduced model = without predictor
  - Then,

$$-2\ln\left(\frac{\hat{L}_{\text{Reduced}}}{\hat{L}_{\text{Full}}}\right) = -2\ln\hat{L}_{\text{Reduced}} - \left(-2\ln\hat{L}_{\text{Full}}\right)$$

$$\sim \chi^2 \text{ with df} = \#\text{ of extra parameters in full model}$$

$$(\text{here df} = 1;\ \text{Critical } \chi_1^2 = 3.84 \text{ for } \alpha = 0.05)$$

Wald Test

- Idea is to use large sample *Z* statistic from a single model to test: $H_0 : \beta_K = 0$

$$\text{Here, } Z = \frac{\hat{\beta}_K}{SE_{\hat{\beta}_K}} \text{ where } Z \sim N(0,1)$$

- Critical *Z* value for α=0.05 is 1.96 (two-sided)

As the sample size gets larger and larger, the Wald test will approximate the Likelihood ratio test.

- The LR test is preferred but Wald test is common

A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. It models the logit-transformed probability as a linear relationship with the predictor variables.

Module 14. Factor analysis

Factor analysis is used to examine variables that are not easily measured. Factor analysis involves grouping similar variables into dimensions. This process is used to identify latent variables or constructs. The purpose of factor analysis is to reduce many individual items into a fewer number of dimensions. The key concept is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable.

Learning Objectives

- Describe the uses of factor analysis
- Perform factor analysis
- Interpret the results of factor analysis

14.1 Uses of factor analysis

- Factor analysis can be used to simplify data, such as reducing the number of variables in regression models.

- Factor analysis is also used to verify scale construction.

- Factor analysis can also be used to construct indices.

14.2 Performing factor analysis

- In reducing number of variables in regression models: factors are rotated after extraction. Factor analysis has several different rotation methods, and some of them ensure that the factors are orthogonal (i.e., uncorrelated), which eliminates problems of multicollinearity in regression analysis.

- In verifying scale construction, the items that make up each dimension are specified upfront. This form of factor analysis is most often used in the context of structural equation modeling and is referred to as confirmatory factor analysis. For example, a confirmatory factor analysis could be performed if a researcher wanted to validate the factors in a behavior scale.

- In constructing an index, the easiest way is to simply sum up all the items in an index. However, some variables that make up the index might have a greater explanatory power than others. A factor analysis could be used to justify dropping questions to shorten questionnaires.

14.3 Interpreting factor analysis

For example, questions on income, occupation, and education are all associated with socio economic status. In a factor analysis. each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.

Factor loadings

Here is an output of a simple factor analysis looking at indicators of wealth, with just six variables and two resulting factors. The relationship of each variable to the underlying factor is expressed by the so-called factor loading.

Variables	Factor 1	Factor 2
Income	0.67	0.11
Education	0.59	0.25
Occupation	0.48	0.19
House value	0.38	0.60
Number of public parks in community	0.13	0.57
Number of violent crimes per year in community	0.23	0.55

The variable with the strongest association to the underlying latent variable.

Factor 1, is income, with a factor loading of 0.67. Since factor loadings can be interpreted like standardized regression coefficients, one could also say that the variable income has a correlation of 0.67 with Factor 1. This would be considered a strong association for a factor analysis. Two other variables, education and occupation, are also associated with Factor 1. Based on the variables loading highly onto Factor 1, this could be "individual socioeconomic status."

The other variables: house value, number of public parks, and number of violent crimes per year, however, have high factor loadings on the other factor, Factor 2. They seem to indicate the overall wealth within the community, so we may want to call Factor 2 "community socioeconomic status."

Eigenvalues

The eigenvalue is a measure of how much of the variance of the observed variables a factor explains. Any factor with an eigenvalue ≥1 explains more variance than a single observed variable.

So if the factor for socioeconomic status had an eigenvalue of 2.3 it would explain as much variance as 2.3 of the three variables. This factor, which captures most of the variance in those three variables, could then be used in other analyses. The factors that explain the least amount of variance are generally discarded.

Module 15. Path Analysis

Path analysis is a form of multiple regression statistical analysis that is used to evaluate causal models by examining the relationships between a dependent variable and two or more independent variables. Path analysis can very well evaluate, test or compute two or more than two types of causal hypotheses (although) it cannot establish the direction of causality.

Learning Objectives

- Describe the uses of path analysis
- Discuss the assumptions in path analysis
- Perform path analysis
- Interpret the results of path analysis

15.1 Uses of Path analysis

- Evaluate causal models by examining the relationships between a dependent variable and two or more independent variables.

- Estimate both the magnitude and significance of causal connections between variables.

15.2 Assumptions in Path Analysis

- The association among the model should be linear in nature.

- The associations among the models should be additive in nature.

- The association among the model should be causal in nature.

- The data that is used should follow an interval type of scale.

- All the error terms are not correlated among the various variables.

- Errors are not correlated among themselves.

- There is only one way causal flow. All causal relationships between variables must go in one direction only (you cannot have a pair of variables that cause each other)

- The variables must have a clear time-ordering since one variable cannot be said to cause another unless it precedes it in time.

15.3 Performing Path Analysis

Path analysis is usually conducted with the help of an added module called the analysis of moment structures (AMOS) in SPSS. Other statistical software like SAS, LISREL, etc. can also be used to conduct path analysis.

The ultimate goal is to predict the regression weight. The regression weight is predicted during path analysis, and then compared to the observed correlation matrix. This type of analysis is also applicable in cases where the researcher wants to perform a goodness of fit test.

Path analysis involves the construction of a path diagram in which the relationships between all variables and the causal direction between them are specifically laid out. When conducting a path analysis, one might first construct an input path diagram, which illustrates the hypothesized relationships. In a path diagram, researchers use arrows to show how different variables relate to each other. An arrow pointing from, say, Variable A to Variable B, shows that Variable A is hypothesized to influence Variable B.

After the statistical analysis has been completed, a researcher would then construct an output path diagram, which illustrates the relationships as they actually exist, according to the analysis conducted. If the researcher's hypothesis is correct, the input path diagram and output path diagram will show the same relationships between variables.

15.4 Interpretation

- The exogenous variables in path analysis are variables whose causes are outside of the model.
- The endogenous variables are variables whose causes are inside the model.
- The recursive model in is a causal model that is unidirectional. In other words, they have one way causal flow. This model in has neither feedback loops nor any reciprocal effects. In this type of model in path analysis, the variables cannot be both cause and affect at the same time.
- The non-recursive model in path analysis is a causal model with feedback loops and reciprocal effects.
- The path coefficient is the standardized regression coefficient that predicts one variable from another.

Module 16. Structural equation modeling

Structural equation modeling is a multivariate statistical analysis technique that is used to analyze structural relationships. This technique is the combination of factor analysis and multiple regression analysis, and it is used to analyze the structural relationship between measured variables and latent constructs.

Learning Objectives

- Describe the use of structural equation modeling
- Discuss the assumptions in structural equation modeling
- Perform structural equation modeling
- Interpret the results of structural equation modeling

16.1 Uses of structural equation modeling

Structural equation modeling is also called causal modeling because it tests the proposed causal relationships. It estimates the multiple and interrelated dependence in a single analysis.

It involves a theory which is a set of relationships providing consistency and comprehensive explanations of the actual phenomena and two types of models:

- **Measurement model:** Represents the theory that specifies how measured variables come together to represent the theory.

- **Structural model:** Represents the theory that shows how constructs are related to other constructs.

16.2 Assumptions in structural equation modeling

- **Data:** Interval data is used.
- **Outlier:** Data should be free of outliers. Outliers affect the model significance.
- **Multivariate normal distribution:** The maximum likelihood method is used and assumed for multivariate normal distribution. Small changes in multivariate normality can lead to a large difference in the chi-square test.
- **Linearity:** A linear relationship is assumed between endogenous and exogenous variables.
- **Sequence:** There should be a cause and effect relationship between endogenous and exogenous variables, and a cause has to occur before the event.
- **Non-spurious relationship:** Observed covariance must be true.
- **Model identification:** Equations must be greater than the estimated parameters or models should be over identified or exact identified. Under identified models are not considered.
- **Sample size:** Most of the researchers prefer a 200 to 400 sample size with 10 to 15 indicators. As a rule of thumb, that is 10 to 20 times as many cases as variables.
- **Uncorrelated error terms:** Error terms are assumed uncorrelated with other variable error terms.

16.3 Performing structural equation modeling

1. **Defining individual constructs:** The first step is to define the constructs theoretically. Conduct a pretest to evaluate the item. A confirmatory test of the measurement model is conducted using CFA.

2. **Developing the overall measurement model:** The measurement model is also known as path analysis. Path analysis is a set of relationships between exogenous and endogens variables. This is shown by the use of an arrow. The measurement model follows the assumption of unidimensionality. Measurement theory is based on the idea that latent constructs cause the measured variable and that the error term is uncorrelated within measured variables. In a measurement model, an arrow is drawn from the measured variable to the constructs.

3. **Design the study to produce the empirical results:** In this step, the researcher must specify the model. The researcher should design the study to minimize the likelihood of an identification problem. Order condition and rank condition methods are used to minimize the identification problem.

4. **Assessing the measurement model validity:** Assessing the measurement model is also called CFA. In CFA, a researcher compares the theoretical measurement against the reality model. The result of the CFA must be associated with the constructs' validity.

5. **Specifying the structural model:** In this step, structural paths are drawn between constructs. In the structural model, no arrow can enter an exogenous construct. A single-headed arrow is used to represent a hypothesized structural relationship between one construct and another. This shows the cause and effect relationship. Each hypothesized relationship uses one degree of freedom. The model can be recursive or non-recursive.

6. **Examine the structural model validity:** In the last step, a researcher examines the structural model validity. A model is considered a good fit if the value of the chi-square test is insignificant, and at least one incremental fit index (like CFI, GFI, TLI, AGFI, etc.) and one badness of fit index (like RMR, RMSEA, SRMR, etc.) meet the predetermined criteria.

Source: Statistics solutions (2020). Available at: https://www.statisticssolutions.com/structural-equation-modeling/#:~:text=Structural%20equation%20modeling%20is%20a,measured%20variables%20and%20latent%20constructs.

16.4 Interpreting results in structural equation modeling

The most important part of SEM analysis is the causal model to draw before attempting an analysis. Some basic rules when drawing a model includes:

- *Rule 1*. Latent variables/factors are represented with circles and measured/manifest variables are represented with squares.
- *Rule 2*. Lines with an arrow in one direction show a hypothesized direct relationship between the two variables. It should originate at the causal variable and point to the variable that is caused. Absence of a line indicates there is no causal relationship between the variables.

- *Rule 3.* Lines with an arrow in both directions should be curved and this demonstrates a bi-directional relationship (i.e., a covariance).

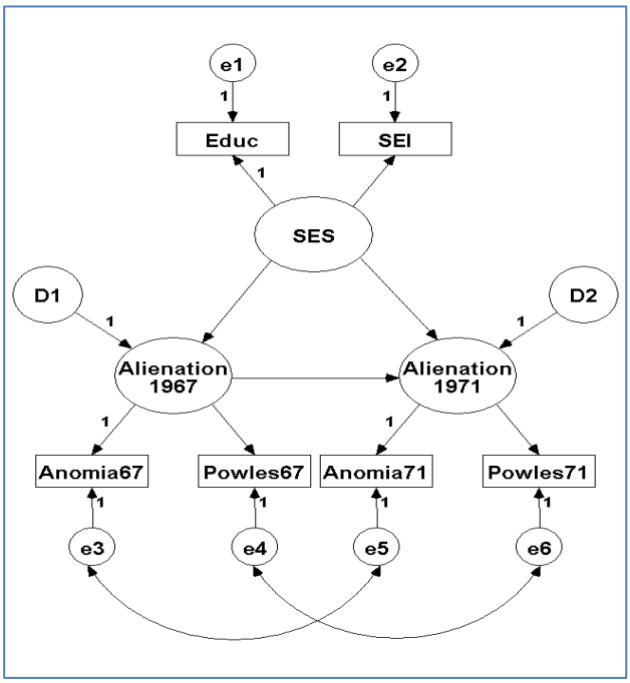
- *Rule 3a.* Covariance arrows should only be allowed for exogenous variables.

- *Rule 4.* For every endogenous variable, a residual term should be added in the model. Generally, a residual term is a circle with the letter E written in it, which stands for error.

- *Rule 4a.* For latent variables that are also endogenous, a residual term is not called error in the lingo of SEM. It is called a disturbance, and therefore the "error term" here would be a circle with a D written in it, standing for disturbance.

SEM example from AMOS:

The model is built in AMOS and the diagram is shown below, please see the SAS example for the explanation on the variables. The standardized parameter estimates are shown in the graph. The squares represent the observed variables and the circles are for the error terms. Three latent variables are assumed with 3 confirmatory factor analyses used to derive them. Ovals are used to indicate these latent variables. The correlation structure between error terms of the confirmatory factor analysis are suggested by AMOS after the initial model fitting without any correlated error terms. This helps improve the overall model fitting.



The goodness-of-fit test statistics are displayed below. Please note the Chi-square test statistic is not significant at 0.05, which suggest that the model fitting is only acceptable. Root mean square error of approximation (RMSEA) is 0.03202 and since it is less than 0.05, it indicates a good fit. Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI) are larger than 0.9 which again reflect a good fit although GFI and AGFI may not be as informative as Chi-square test statistics and RMSEA.

Result (Default model)

Minimum was achieved
Chi-square = 7.81724
Degrees of freedom = 4
Probability level (p-value) = .09851

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.03202	.00000	.06532	.78202
Independence model	.39072	.37687	.40475	.00001

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	.08052	.99724	.98553	.18995
Saturated model	.00000	1.00000		
Independence model	3.98590	.48216	.27503	.34440

Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
Alienation1967	<---	SES	-.64495	.05350	-12.05418	***	
Alienation1971	<---	SES	-.22497	.05509	-4.08390	***	
Alienation1971	<---	Alienation1967	.58916	.05580	10.55811	***	
Educ	<---	SES	1.00000				
SEI	<---	SES	.58409	.04264	13.69760	***	
Powles67	<---	Alienation1967	1.00000				
Anomia67	<---	Alienation1967	1.12575	.06772	16.62422	***	
Powles71	<---	Alienation1971	1.00000				
Anomia71	<---	Alienation1971	1.13332	.07111	15.93816	***	

Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
e3	<-->	e5	1.61074	.32703	4.92541	***	
e4	<-->	e6	.53090	.24851	2.13634	.03265	

All the parameter estimates are high significant. In other words, all of them are significantly differently from 0. The interpretations on the parameter estimates are straight forward. For example, Alienation in 1967 decreases -.726 for each 1.00 increase in SES. The correlation structure between e3 and e5, e4 and e6 is also estimated by AMOS with significant results.

Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
Alienation1967	<---	SES	-.62795
Alienation1971	<---	SES	-.21489
Alienation1971	<---	Alienation1967	.57797
Educ	<---	SES	.78908
SEI	<---	SES	.67865
Powles67	<---	Alienation1967	.82379
Anomia67	<---	Alienation1967	.78535
Powles71	<---	Alienation1971	.80590
Anomia71	<---	Alienation1971	.81502

The standardized the regression estimates are comparable, which may assist us to pick up more important factors and relationships.