MS4S08

Applied Statistics for Data Science

Dr. Angelica Pachon

angelica.pachon@southwales.ac.uk

University of South Wales

2020/21

Lecturer details

Dr Angelica Pachon

angelica.pachon@souhwales.ac.uk

Room number: J414

Telephone number: 01443 654730

Proposed syllabus 2020/21

Lecture	Content
1 (Week 13)	Introduction to the module and recap on SAS UE
2 (Week 13)	Multivariate Regression
3 (Week 14)	Logistic Regression
4 (Week 15)	Principal Component Analysis
6 (Week 16)	Factor Analysis
7 (Week 16)	Cluster Analysis

Lecture notes and solutions

A copy of the completed lecture notes will be provided on Blackboard at the beginning of the module.

Solutions to tutorial exercises will be provided on Blackboard after sufficient time for these to be completed has passed.

The Timetable

This module will be block taught over 4 weeks. Therefore there are 6 hours each week allocated to the module per student.

Coursework

There is one piece of coursework associated with this half of the module and this will count as 50% towards the overall module mark. The coursework will be available on Blackboard and the required hand-in date is **Tuesday 12th January 2021**. The statistical techniques required to complete the tasks will have been taught in the lectures and the computer laboratory tutorials. The coursework will be marked by the Lecturer and returned to you where individual feedback will be provided.

Reading List

Der, G. and Everitt, B. (2016) Essential statistics using SAS university edition. University edition. Cary, NC : SAS Institute 2016

O'Rourke, Norm, Larry Hatcher, and Edward J. Stepanski (2005). A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics, Second Edition. SAS Institute.

Alvin C. Rencher (2002) Methods of Multivariate Analysis, Second Edition. Wiley-Inter-Science.

M.S. Srivastava (2002) Methods of Multivariate Statistics. Wiley-Inter-Science.

Support with SAS University Edition/SAS Studio

General information: https://www.sas.com/en_us/software/studio.html

SAS Analytics U support videos:

https://www.youtube.com/playlist?list=PLVBcK_IpFVi9cajJtRel2uBLbtcLz-WIN

Introduction to the module

The purpose of this course is to demonstrate the applications of various statistical methods when applied to real-life data through the use of a statistical software: SAS. SAS is a global analytics, business intelligence and data management software utilised by companies across industry sectors.

The primary software tool that will be used for analysis is the **SAS University Edition** package. This software is feely available for students to download on their personal machines, see https://www.sas.com/en_gb/software/university-edition.html. On campus it is also available in J109, J202, J204, J208 and J448.

The principal motivation of the course is to demonstrate the practical application of key statistical concepts through the use of a menu-driven program. An appreciation of the statistical code driving the menus is also required.

The processes by which the analyses are performed in SAS University Edition (SAS UE) prove to be very similar to those used in other popular statistical software packages, namely: SPSS and R. Thus familiarity with this package will provide a very good foundation for applying the same techniques in other statistical software packages.

In essence the primary purpose of statistics is to make sense out of the data at hand. A variety of tools are available to achieve this goal. One of the most difficult decisions for the inexperienced analyst is to determine the most adequate techniques to study the data that they are presented with.

Before determining which approach to utilise, one must first consider the motivation of the analysis.

In many circumstances the purpose of the analysis is to explore the data – to try to look for any patterns/relationships, and allow the data to determine the form of the analysis. This idea is commonly known as **exploratory data analysis (EDA).**

On the other hand, the study can be motivated by interest in specific relationships that are assumed prior to the analysis; this approach is known as **confirmatory analysis**.

1. Recap

Statistical Inference is the branch of statistics concerned with using sample data to make inferences about the population. Populations are characterised by numerical descriptive measures, referred to as *parameters*.

Note: A *parameter* is a value, usually unknown (and which therefore has to be estimated), that is used to represent a certain population characteristic.

Hypothesis Testing

The mechanism for hypothesis testing follows a series of general steps, in this course we will:

1. **Determine the appropriate statistical test;**
2. **State the null and alternative hypotheses of interest;**
3. **Check the assumptions regarding the variables of interest are satisfied;**
4. **Perform the analysis using SAS;**
5. **Document conclusions on the basis of the analysis performed.**

The conclusions that are derived in step 5 are dictated by the significance level (α). This is decided upon prior to the analysis and is the level at which the analyst decides to reject the null hypothesis. The outcomes of hypothesis testing can be expressed as follows:

Hypothesis	**True outcome**	
	H_0	**H_1**
H_0 (Null hypothesis)	Correct decision $(1 - \alpha)$	False negative decision (Type II Error (β))
H_1 (Alternative hypothesis)	False positive decision	Correct decision

	(Type I error α (or p-value)	(1 – β)

A general rule of thumb that is applied by statisticians is to reject the null hypothesis if the p-value (α) that is obtained in the output is less than 0.05, and is referred to as ***rejecting the null hypothesis at 5% level of significance***.

The table below provides a summary of the relevant statistical tests for alternative situations where hypothesis testing is required (the methods used will be demonstrated later in the notes).

Hypothesis	**Normally distributed**	**Non-normal/rank/scores**
Compare a variable to a hypothetical value	One-sample t-test	Wilcoxon-test
Compare two independent variables	Two-sample t-test	Wilcoxon Mann-Whitney test
Compare two variables with same response for the same subjects	Paired t-test	Wilcoxon-test (Signed rank test)
Compare three or more variables	One-way ANOVA	Kruskal-Wallis test
Relationship between two variables	Pearson correlation coefficient	Spearman correlation coefficient
Linear relationship between two variables	Simple linear regression	Non-parametric linear regression (outside of scope for this course)

Linear relationship between a dependent variable and two or more independent variables	Multiple linear regression	

2. Correlation

The correlation coefficient (*r)* measures the strength of association between two variables. A correlation coefficient near to 1 implies a strong positive relationship, near 0 implies the variables are independent and near –1 implies a negative relationship.

If we were asked to study the relationships between the scores obtained in the different IQ domains and the MRI count (found in the IQ data studied in a previous lecture), an appropriate method of analysis would be to construct a correlation matrix. In essence this involves recording all the correlations in a single matrix as opposed to individually.

Pearson's correlation coefficient is the most commonly used on continuous variables and requires that the variables follow a Normal distribution. However, if the data is skewed or if one of the variables is on an ordinal scale and the other is not on an ordinal scale, then a more appropriate measurement is Spearman's rho.

How to test if a variable is normally distributed in SAS?

In order to determine which correlation coefficient should be used within the analysis we first need to determine if the variables we wish to analyse are Normally distributed.

One way we can carry out a **test for normality** is as follows:

- Click on **Tasks** then **Distribution Analysis.**
- Select the variables for analysis.
- Under the **Options tab** select to **Add normal curve**.
- Under this section, you can select **Histogram and Probability Plots.**

The hypotheses in this case are:

H_o: there is no difference between the distribution of the variable and that of the normal distribution;
H_1: there is a difference between the distribution of the variable and that of the normal distribution.

The following table is achieved for the results of the fitted Normal Distribution for FISQ.

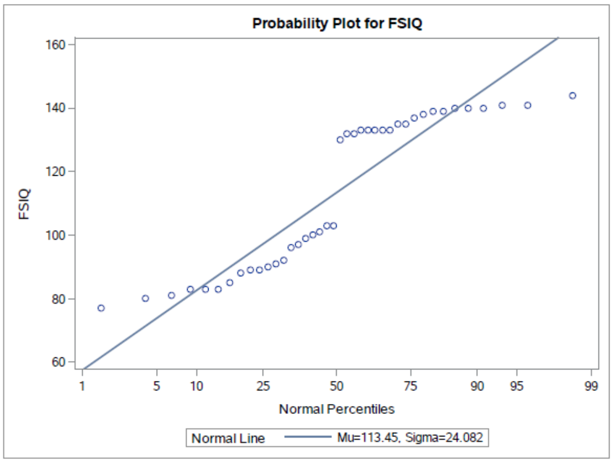
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.25443386	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.51127276	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.89955516	Pr > A-Sq	<0.005

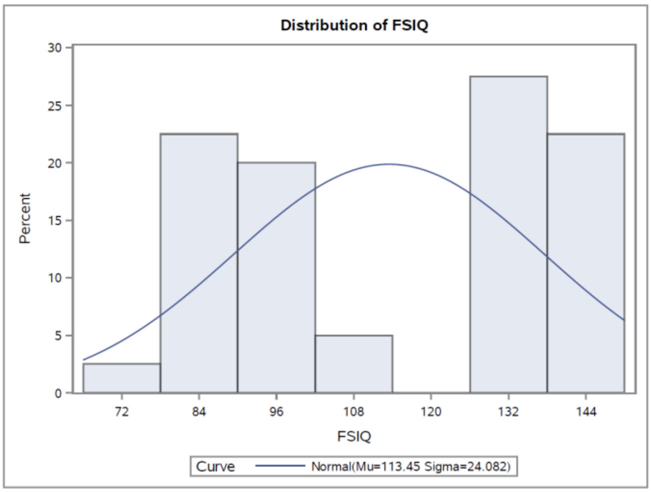
When you fit a parametric distribution, **PROC UNIVARIATE** provides a series of goodness-of-fit tests based on the empirical distribution function (EDF), these are the Kolmogorov-Smirnov statistic, the Anderson-Darling statistic, and the Cramér-von Mises statistic.

The EDF tests offer advantages over a traditional chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints.

By studying the output for all three tests we can conclude that all 3 p-valuies are significant at the 5% level, therefore the distribution of the FISQ variable is signficantly different to that of a Normal distribution.

This is also supported by the probability plot and the histogram which both differ from the plots that would be expected if the data followed the desired distribution.





Example 2

As our data is not normally distributed, we therefore need to use the Spearman's rho correlation coefficient to determine if a significant relationship (correlation) exists between the IQ variables present in the IQ data.

We perform a **correlation** as follows:

- Select **Tasks** then **Correlation Analysis.**
- Select all variables for analysis.
- Under **Options, STATISTICS, Nonparametric Correlations**, select Spearman's rank**.**
- You can also select a scatter plot under **PLOTS** if you wish.

The hypotheses in this case are:

H_o: there is no linear relationship between the IQ variables;
H_1: there is a linear relationship between the IQ variables.

The following output table is achieved by generating a correlation matrix on FSIQ, VIQ, PIQ and MRI_Count using the Spearman Correlation Coefficient.

Spearman Correlation Coefficients, N = 40 Prob > |r| under H0: Rho=0

	FSIQ	**VIQ**	**PIQ**	**MRI_Count**
FSIQ	1.00000	0.9180	0.87869	0.47214
		<.000	<.0001	0.0021
VIQ	0.91805	1.00000	0.71498	0.39981
	<.0001		<.0001	0.0106
PIQ	0.87869	0.71498	1.00000	0.41246
	<.0001	<.0001		0.0082

Spearman Correlation Coefficients, N = 40 Prob > |r| under H0: Rho=0				
	FSIQ	**VIQ**	**PIQ**	**MRI_Count**
MRI_Count	0.47214	0.39981	0.41246	1.00000
	0.0021	0.0106	0.0082	

By studying the correlation matrix we find that all of the measurements are positively correlated and are significant at the 5% level. However, the strength of the association varies considerably.

Determining the strength of the association is subjective, and can vary from analyst to analyst, however for the purpose of this course, we shall apply the following rules:

- -1.0 to -0.6 strong negative association
- -0.6 to -0.3 weak negative association
- -0.3 to +0.3 little or no association
- +0.3 to +0.6 weak positive association
- +0.6 to +1.0 strong positive association.

The results obtained are expected in practise due to the responses of interest all measuring the same characteristic (intelligence), consequently we would expect the measurements to increase simultaneously.

Comment on the outcomes of these tests.

It appears as though there is a significant relationship between all pairs of variables.

All p-value are significant at the 5% level.

The strongest correlation is between FSIQ and VIQ with the weakest correlation being MRI Count with VIQ.

3. Multivariate Regression

Multivariate regression analysis is an extension of simple regression (which involves one independent variable) to the situation where more than one variable must be considered. This type of probabilistic model relating y to various independent variables, say, $x_1, x_2, \dots, x_k$ is called a **general linear statistical model** (or a linear model) and is expressed as:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \, .$$

In performing a multiple linear regression, one must follow a series of steps. Standard assumptions include:

- **Suggested regression model is appropriate;**
- **Traditional model fitting assumptions requires an error term in the suggested regression model;**
  - **errors have zero mean,**
  - **errors have constant variance for fixed values of the predictors,**
  - **errors are normally distributed,**
  - **errors associated with any two observations are independent;**
- **responses are independent for fixed values of the predictors.**

If the **Normality assumption of errors** is breached then one should look at transforming *Y*. Transforming the dependent (or independent) variables is a method that is commonly applied to real life data; the principal motivation for applying this methodology is:

- to stabilise the variance of the dependent variable, thus attaining homoscedasticity (constant variance);
- to normalise the dependent variable if it violates the normality assumption;
- to make the regression model linear if the original data suggests non-linear model.

For the purpose of the course we shall restrict ourselves to the most commonly known transformations on the dependent variable.

***log y* (*log transformation*)** – this is used when the variance of the residuals markedly increases with the dependent variable, for normalisation purposes and to linearise the regression model if the relationship between dependent and independent variable has an increasing slope.

***y^2* (*square transformation*)** – is used to stabilise the variance of the residuals, if the variance decreases with *y,* to normalise the dependent variable and to linearise the model if the relationship with the independent variable is curvilinear downward.

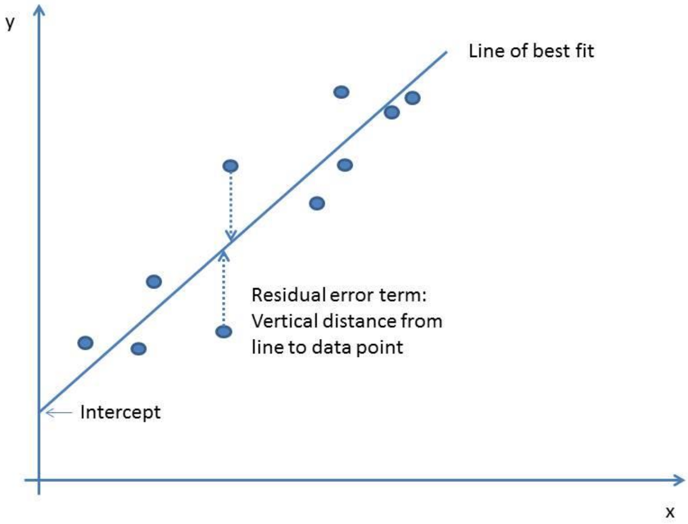
$\sqrt{y}$* *(square root transformation) – this is used when the variance of the residuals are proportional to the mean of *y*. This transformation is the natural choice when the dependent variable is known to have a Poisson distribution.

4.1 Residual Analysis

To test the assumptions in multivariate regression analysis one can use the **residual plots** that are produced in the regression analysis; these come in the form of scatterplot representation of the residuals versus the dependent variable.

The regression residual (r_i) is defined as the difference between an observed value and its corresponding predicted value, thus in multivariate regression this can be expressed as:

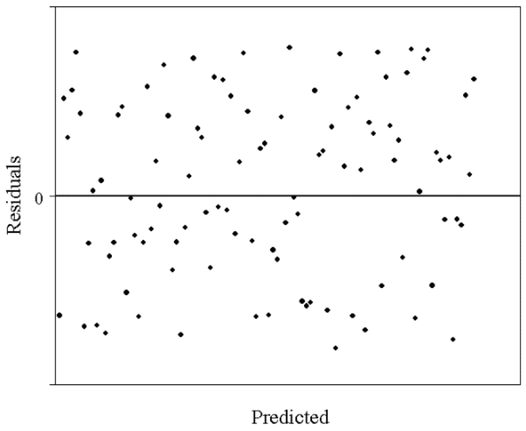
$$r_i = y_i - \hat{y}_i = y_i - (\alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \dots \beta_k x_{ki}) .$$



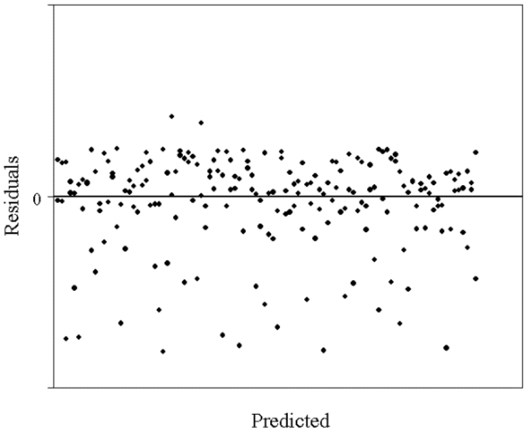
Graphically observing the residuals versus the predicted response provides a convenient way for examining lack of fit. The residual scatterplots below provide typical examples of output obtained from statistical packages after performing a multivariate linear regression analysis.

Figure (i) demonstrates a scenario where all of the assumptions of the regression analysis are satisfied; implying that the analysis is valid and interpretations can be made directly without further consideration as to the appropriateness of the model.

Figure (ii) is an example of where the errors are not normally distributed, this is evident through the fact that the residual plot has a large number of residuals near 0. In this case the distribution of the residuals are skewed, implying that the model is consistently over estimating the value of the dependent variable (as the values are < 0).



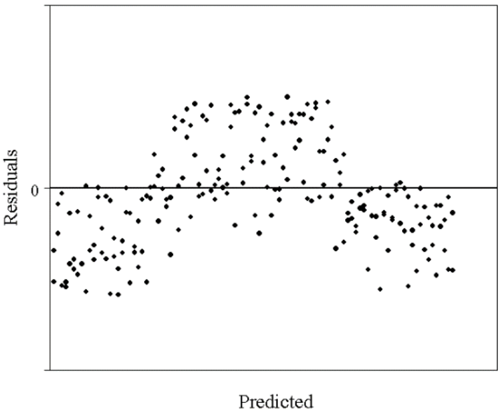
(i) Assumptions satisfied.



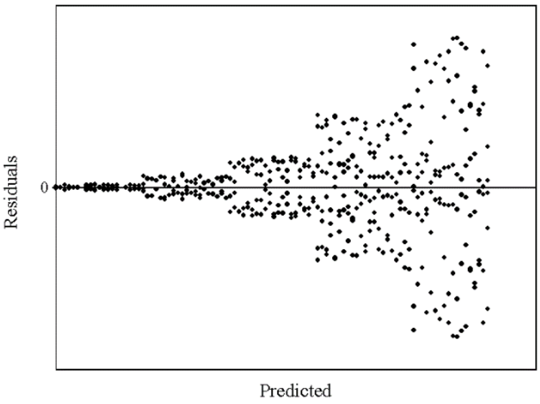
(ii) Non-normal errors.

In Figure (iii) we observe a curvature of the residuals, which reveals the inappropriateness of a linear model in predicting the response. Ideally one would expect a rectangular distribution as seen in (i).

Heteroscedasticity (non-constant variance) of residuals is demonstrated in Figure (iv) with the errors becoming exponentially larger as the predicted values increase. Transforming the response can be applied to obtain homogeneous errors, additionally in these circumstances a weighted least squares regression could be deemed more appropriate.



(iii) Non-linear response.



(iv) Non-homogeneous errors.

In practice, seldom does a regression analysis satisfy all of the assumptions after one run, with the analyst requiring several runs until they can obtain a model with which they are fully content.

4.2 Interpreting the analysis

After hypothesising the linear model of interest and verifying that the assumptions for a multivariate linear regression model have been satisfied, the analyst must interpret the output obtained from the analysis to determine if any significant relationships are present and in general make inferences about the model and its parameters.

The pivotal statistic provided by the analysis is the *p–value* found in the analysis of variance (**ANOVA**) table, which reveals whether there is a significant overall regression.

Thus, for our model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

we test the hypotheses:

$$H_0: \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

H_1: At least one of the β parameters in H_0 is nonzero.

This statistically tests if the independent variables are collectively significant in explaining the variation in the response, or whether a horizontal model (constant) would be adequate.

If the p-value is < 0.05, we accept that the model is accountable for a significant amount of the variation in the analysis and then concentrate on determining the roles that the independent variables have in the model.

Another method for assessing the reliability of the model is to assess the ***coefficient of determination*** (***R^2***), this represents the amount of variability in the data that is purely attributable to the model fitted.

Statistical tests on the significance of the β_i's and the formation of confidence intervals are provided by default with the linear regression facility in SAS UE.

Example 4

An experiment was performed to investigate the relationship between Systolic Blood Pressure (SBP), weight in kilograms and age in years of a cohort of 12 children who attend a certain medical practise. The data can be found in **Multivariate_Regression**.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
SBP (mm/Hg)	136	152	102	112	114	154	116	110	134	106	142	128
Weight (kg)	57	61	42	42	48	55	50	51	62	49	59	57
Age (years)	9	12	6	10	9	10	7	8	11	6	10	8

To carry out a **Linear Regression** we need to:

- Select **Tasks** then **Linear Regression.**
- Select the **Dependent variable** and **Continuous variables.**
- Under **Model** select the predictor variables as single effects**.**
- Under **Options**, **STATISTICS,** select the **Tolerance values for estimates.**
- Under **Plots,** select **Residual plots.**
- Click **Run.**

The results of fitting a multivariate linear regression model for the dependent variable **SBP**, using the independent variables **Weight** and **Age** were as follows:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	2771.29043	1385.64521	15.95	0.0011
Error	9	781.70957	86.85662		
Corrected Total	11	3553.00000			

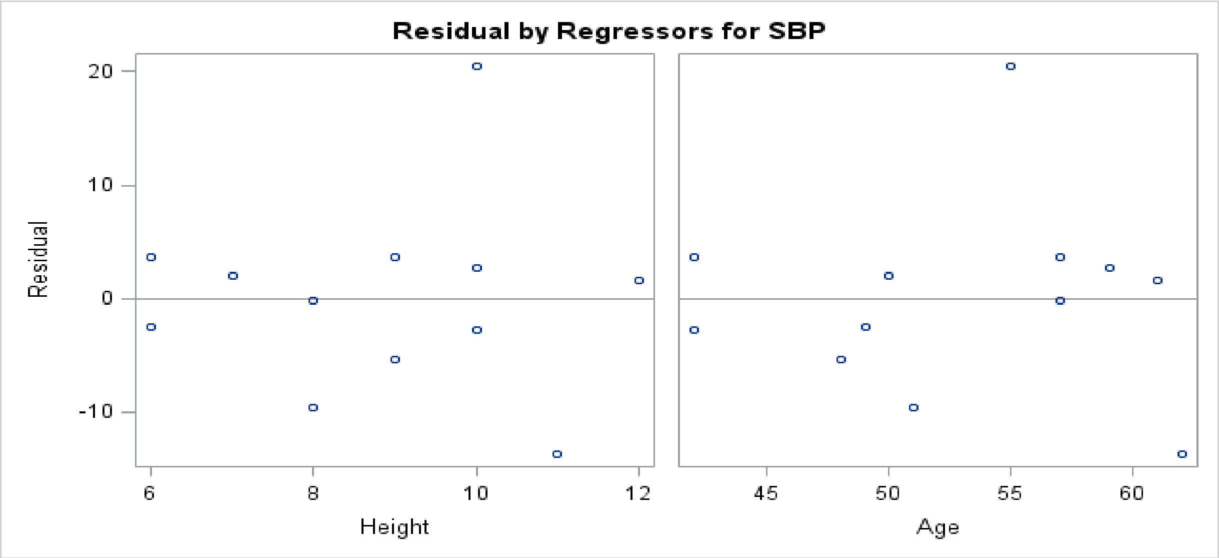
From the results we discover that the *p*-value for the model is 0.0011, which means that we reject the null hypothesis at the 5% level. This means that we accept that a linear model is appropriate. However, we need to check our assumptions to see if they are valid.

To examine the independence of errors we will need to add a further statistic to our code. Open the code for the linear regression and select to edit. On the model line after the "/" we need to add **"dw"**. This performs a Durbin-Watson test for correlated residuals.

There are many interpretations of what D should be, by rule of thumb, **if 1 < D < 3** then the errors are reasonable independent.

Durbin-Watson D	1.760
Number of Observations	12
1st Order Autocorrelation	0.111

Based on the residual plots below we can also assume that the errors are normally distributed, have a constant variance and a mean of zero.



We also need to check for multicollinearity (**models within models**), we do not want our predictor variables to be highly correlated with each other. It is undesirable to have more than one copy of a variable in a model (double counting).

This can be checked with collinearity statistics, in particular the tolerance for each variable should be **more than 0.2**, which it is in our case.

Now we have determined that the assumption have been met, we can interpret the output.

[**Definition** - In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy.]

Root MSE	9.31969	**R-Square**	0.7800
Dependent Mean	125.50000	**Adj R-Sq**	0.7311
Coeff Var	7.42605		

On inspection of the significance of the terms included in the model we find that only **Age** is significant (*p* = 0.0218).

The β_i value for **Age** tells us that for each year older that a subject gets causes the SBP to increase by 1.4408 (assuming that height remains constant).

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**
Intercept	1	13.10610	21.88965	0.60	0.5641	.
Age	1	4.10025	1.87445	2.19	0.0565	0.62320
Weight	1	1.44408	0.52161	2.77	0.0218	0.62320

The R-square value of 0.7800 tells us that the regression model accounts for 78% of the variation in the original data. So the model can be seen as a good fit.

As **Weight** is the only significant variable in the model it would be natural for the analyst to include only this term in the model and base interpretations on the output of the subsequent analysis. We will therefore re-run the model with weight as the only predictor variable.

The model is still significant and the assumptions are still met. The final output is as follows.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**
Intercept	1	12.37970	25.69749	0.48	0.6404	.
Weight	1	2.14446	0.48346	4.44	0.0013	1.00000

The β_i value for **Weight** as the only predictor value tells us that for each increase of 1 in weight causes the SBP to increase by 2.14446.

5. Logistic Regression

Logistic regression is foremost used to model a binary (0,1) variable Y, based on one or more other variables, called predictors. The binary variable being modelled is generally referred to as the response variable, or the dependent variable. We shall use the term "response" for the variable being modelled since it has now become the preferred way of designating it. For a model to fit the data well, it is assumed that

- The predictors are uncorrelated with one another.
- That they are significantly related to the response.
- The observations or data elements of a model are also uncorrelated.

At the center of logistic regression analysis is the task estimating the log odds of an event. Mathematically logistic regression estimates

$$\text{Logit(p)=log(odds)=}log\left(\frac{P(Y=1\ |x_1,\dots x_r)}{P(Y=0\ |\ x_1,\dots,x_r)}\right)=b_0+b_1x_1+\cdots+b_rx_r$$

The logit function or the log-odds is the inverse of the sigmoidal "logistic" function or logistic transform used in mathematics, (take exponential to both sides of the previous equation and find p) especially in statistics, from which we obtain

$$p=\frac{1}{1+e^{-(b_0+b_1x_1+\cdots+b_rx_r)}}$$

In general, we can see that the parameter b_j of a continuous random variable x_j, j=1,…,r, refers to the effect of x_j on the log-odds, adjusting for the other x_k, k≠ j. Thus, we can say that for every one unit change in x_j, the log odds of Y=1 (versus Y=0) increases by b_j.

The coefficients for the categories of categorical variables have a slightly different interpretation. We will see that better later with an example.

An odds ratio is the exponentiated coefficient, exp(b_j), and can be interpreted as the multiplicative effect in the odds of a 1-unit increase in x_j, when we can keep fixed the levels of the other x_k.

For example, for a one unit increase in x_j, the odds of Y=1 (versus Y=0) increase by a factor of e^{b_j}.

The probabilities p and the regression coefficients are unobserved, and the means of determining them is not part of the model itself. They are typically determined by some sort of optimization procedure, e.g. maximum likelihood estimation, that finds values that best fit the observed data (i.e. that give the most accurate predictions for the data already observed), usually subject to regularization conditions that seek to exclude unlikely values, e.g. extremely large values for any of the regression coefficients.

Evaluating goodness of fit

After fitting the model, it is likely that researchers will want to examine the contribution of individual predictors. To do so, they will want to examine the regression coefficients.

In linear regression, the significance of a regression coefficient is assessed by computing a t-test. In logistic regression, there are several different tests designed to assess the significance of an individual predictor, most notably the **likelihood ratio test and the Wald test**.

In the Wald test the hypothesis are: H_0: b_j=0 vs. H_1: b_j≠ 0, j=1,…,r

The Likelihood ratio test is used to compare two models. It is a measure of the lack of fit to the data in a logistic regression model. For example to test the null hypothesis that an arbitrary group of r coefficients from the model is set equal to zero (e.g. no relationship with the response), we need to fit two models: the reduced model which omits the p predictors in question (only takes into account b_0), and the current model which includes them (b_0,b_1,…,b_r).

To perform the Likelihood ratio test, we must look in the SAS output at the "Model Fit Statistics" section and examine the value of "-2 Log L" for "Intercept and Covariates." Here, the reduced model is the "intercept-only" model (e.g. no predictors) and "intercept and covariates" is the current model we fitted.

An alternative statistic for measuring overall goodness-of-fit is **Hosmer-Lemeshow** test. It is more useful when there is more than one predictor and/or continuous predictors in the model too, and test the hypothesis,

H_0 : the current model fits well, vs. H_1 : the current model does not fit well.

Pseudo R^2: In linear regression the squared multiple correlation, R^2 is used to assess goodness of fit as it represents the proportion of variance in the criterion that is explained by the predictors. In logistic regression analysis, there is no agreed upon analogous measure, but there are several competing measures each with limitations. One of the most commonly used indices used is the Likelihood ratio R^2_L

If the model has no predictive ability, R^2_L will be close to zero, otherwise it would be close to 1.

Akaike Information Criterion: The Akaike information criterion (AIC) test, named after Japanese statistician Hirotsugu Akaike (1927-2009), is perhaps the most well-known and well used information statistic in current research. The Akaike Information Criterion provides a method for assessing the quality of your model through comparison of related models.

Unlike adjusted R-squared, the number itself is not meaningful. If you have more than one similar candidate models (where all of the variables of the simpler model occur in the more complex models), then you should select the model that has the ”smallest” AIC.

So its useful for comparing models, but isn't interpretable on its own.

To carry a logistic regression model we need to do:

- Select **Task**, then **Linear Models** then **Binary Logistic Regression**
- Select the data set
- Select the Response variable . This variable is a binary variable.
- Select the event of interest, that is the event that corresponds with p (that is 0 or 1). In this case select 1.
- Select the classification and the continuous variables.
- In **Model** select all the variables and in *Single effects* select **Add**
- In **Options** select Generalized R Square and in **Goodness-of-fit and Overdispersion**
- Select Hosmer and Lemeshow goodness-of-fit.
- Run the model.

Example 5

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The outcome variable, admit/do not admit, is binary. The data is on Black-Board and call binary.sas7bdat.

Description of the data: This data set has a binary response (outcome, dependent) variable called **admit**, which is equal to 1 if the individual was admitted to graduate school, and 0 otherwise**.** There are three predictor variables: **gre**, **gpa**, and **rank**. We will treat the variables **gre** and **gpa** as continuous. The variable **rank** takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.

We run the logistic regression model. To model 1s rather than 0s, we use in the **event of interest** option 1. We do this because by default, **proc logistic** models 0s rather than 1s, in this case that would mean predicting the probability of not getting into graduate school (**admit**=0) versus getting in (**admit**=1). Mathematically, the models are equivalent, but conceptually, it probably makes more sense to model the probability of getting into graduate school versus not getting in. We put **rank** as a categorical variable and the others as continuous variables.

Data Set	MS3S30.BINARY	Written by SAS
Response Variable	ADMIT	
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	400
Number of Observations Used	400

Response Profile

Ordered Value	ADMIT	Total Frequency
1	0	273
2	1	127

Probability modeled is ADMIT=1.

Class Level Information

Class	Value	Design Variables			
RANK	**1**	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

- The first part of the above output tells us the file being analysed (library MS3S30 and data Binary) and the number of observations used. We see that all 400 observations in our data set were used in the analysis (fewer observations would have been used if any of our variables had missing values).

- We also see that SAS is modelling **admit** using a binary logit model and that the probability that of **admit** = 1 is being modelled.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	501.977	470.517
SC	505.968	494.466
-2 Log L	499.977	458.517

R-Square	0.0985	**Max-rescaled R-Square**	0.1380

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	41.4590	5	<.0001
Score	40.1603	5	<.0001
Wald	36.1390	5	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
RANK	3	20.8949	0.0001
GRE	1	4.2842	0.0385
GPA	1	5.8714	0.0154

- The portion of the output labelled Model Fit Statistics describes and tests the overall fit of the model. The -2 Log L (499.977) can be used in comparisons of nested models, but we won’t show an example of that here.
- In the next section of output, the likelihood ratio chi-square of 41.4590 with a p-value of 0.0001 tells us that our model as a whole fits significantly better than an empty model. The Score and Wald tests are asymptotically equivalent tests of the same hypothesis tested by the likelihood ratio test, not surprisingly, these tests also indicate that the model is statistically significant.

- The section labeled Type 3 Analysis of Effects, shows the hypothesis tests for each of the variables in the model individually. The chi-square test statistics and associated p-values shown in the table indicate that each of the three variables in the model significantly improve the model fit. For **gre** and **gpa**, this test duplicates the test of the coefficients shown below. However, for class variables (e.g., **rank**), this table gives the multiple degree of freedom test for the overall effect of the variable.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-5.5414	1.1381	23.7081	<.0001
RANK	**1**	1	1.5514	0.4178	13.7870	0.0002
RANK	**2**	1	0.8760	0.3667	5.7056	0.0169
RANK	**3**	1	0.2112	0.3929	0.2891	0.5908
RANK	**4**	0	0	.	.	.
GRE		1	0.00226	0.00109	4.2842	0.0385
GPA		1	0.8040	0.3318	5.8714	0.0154

- The above table shows the coefficients (labelled Estimate), their standard errors (error), the Wald Chi-Square statistic, and associated p-values. The coefficients for **gre**, and **gpa** are statistically significant, as are the terms for **rank**=1 and **rank**=2
- (versus the omitted category **rank**=4). The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.
  - For every one unit change in **gre**, the log odds of admission (versus non-admission) increases by 0.002.
  - For a one unit increase in **gpa**, the log odds of being admitted to graduate school increases by 0.804.

- The coefficients for the categories of rank have a slightly different interpretation. For example, having attended an undergraduate institution with a **rank** of 1, versus an institution with a **rank** of 4, increases the log odds of admission by 1.55.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
RANK 1 vs 4	4.718	2.080	10.701
RANK 2 vs 4	2.401	1.170	4.927
RANK 3 vs 4	1.235	0.572	2.668
GRE	1.002	1.000	1.004
GPA	2.235	1.166	4.282

The first table above gives the coefficients as odds ratios. An odds ratio is the exponentiated coefficient and can be interpreted as the multiplicative change in the odds for a one unit change in the predictor variable. For example, for a one unit increase in **gpa**, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.24. The output gives a test for the overall effect of **rank,** as well as coefficients that describe the difference between the reference group (**rank**=4) and each of the other three groups. We can also test for differences between the other levels of **rank**. For example, we might want to test for a difference in coefficients for **rank**=2 and **rank**=3, that is, to compare the odds of admission for students who attended a university with a rank of 2, to students who attended a university with a rank of 3. We can test this type of hypothesis by adding a **contrast** statement to the code for **proc logistic.**

```
proc logistic data=MS3S30.BINARY plots=(roc);
    class RANK / param=glm;
    model ADMIT(event='1')=RANK GRE GPA / link=logit lackfit rsquare
            technique=fisher;
            contrast 'rank 2 vs 3' rank 0 1 -1/ estimate=parm;
run;
```

Following the word **contrast**, is the label that will appear in the output, enclosed in single quotes (i.e., **'rank 2 vs. rank 3'**). This is followed by the name of the variable we wish to test hypotheses about (i.e., **rank**), and a vector that describes the desired comparison (i.e., **0 1 -1**). In this case the value computed is the difference between the coefficients for **rank**=2 and **rank**=3. After the slash (i.e., **/**) we use the **estimate = parm** option to request that the estimate be the difference in coefficients.

Contrast Test Results			
Contrast	DF	Wald Chi-Square	Pr > ChiSq
rank 2 vs 3	1	5.5052	0.0190

Contrast Estimation and Testing Results by Row									
Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
rank 2 vs 3	PARM	1	0.6648	0.2833	0.05	0.1095	1.2200	5.5052	0.0190

Because the models are the same, most of the output produced by the above **proc logistic** command is the same as before. The only difference is the additional output produced by the **contrast** statement. Under the heading Contrast Test Results we see the label for the contrast (rank 2 versus 3) along with its degrees of freedom, Wald chi-square statistic, and p-value. Based on the p-value in this table we know that the coefficient for **rank**=2 is significantly different from the coefficient for **rank**=3. The second table, shows more detailed information, including the actual estimate of the difference (under Estimate), it’s standard error, confidence limits, test statistic, and p-value. We can see that the estimated difference was 0.6648, indicating that having attended an undergraduate institution with a **rank** of 2, versus an institution with a rank of 3, increases the log odds of admission by 0.67.

In the syntax below we use multiple contrast statements to estimate the predicted probability of admission as **gre** changes from 200 to 800 (in increments of 100). When estimating the predicted probabilities we hold **gpa** constant at 3.39 (its mean), and **rank** at 2. The term **intercept** followed by a **1** indicates that the intercept for the model is to be included in estimate.

```
proc logistic data=MS3S30.BINARY plots=(roc);

        class RANK / param=glm;

        model ADMIT(event='1')=RANK GRE GPA / link=logit lackfit rsquare

                technique=fisher;

                contrast 'gre=200' intercept 1 gre 200 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=300' intercept 1 gre 300 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=400' intercept 1 gre 400 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=500' intercept 1 gre 500 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=600' intercept 1 gre 600 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=700' intercept 1 gre 700 gpa 3.3899 rank 0 1 0  / estimate=prob;

  contrast 'gre=800' intercept 1 gre 800 gpa 3.3899 rank 0 1 0  / estimate=prob;
run;
```

As with the previous example, we have omitted most of the **proc logistic** output, because it is the same as before. The predicted probabilities are included in the column labeled Estimate in the second table shown above. Looking at the estimates, we can see that the predicted probability of being admitted is only 0.18 if one's **gre** score is 200, but increases to 0.47 if one's gre score is 800, holding **gpa** at its mean (3.39), and **rank** at 2.

Thing to consider:

- Empty cells or small cells: You should check for empty or small cells by doing a crosstab between categorical predictors and the outcome variable. If a cell has very few cases (a small cell), the model may become unstable or it might not run at all.
- Sample size: It is sometimes possible to estimate models for binary outcomes in datasets with only a small number of cases using exact logistic regression (available with the **exact** option in **proc logistic**
- See the Pseudo-R-squared and Hosmer-Lemeshow Statistic.

Classification Statistics:

Logistic regression is many times used as a classification tool. However, it should be noted that the ability to classify or discriminate between the two levels of the response variable is due more to the degree of separation between the levels and size of the regression coefficients than it is to the logistic model itself. Discriminate analysis and other classification schemes can also do a good job in classifying and are not logistic models. On the other hand, logistic models are easy to work with and are robust in the classification results they provide the analyst.

There are basic or standard types of classification tools used with logistic regression, two of them are the sensitivity-specificity (S-S) plot and the receiver operator characteristic (ROC) curve. We will address each of these in this section. Each of these tests is based on a cut point, which determines the optimal probability value with which to separate predicted versus

observed successes (1) or failures (0).

Sensitivity and Specificity

In a binary set up, the dependent variable or the target variable in a logistic regression is the probability of the event that a customer is likely to respond or not likely to respond. We have to evaluate these probabilities on the real set of data. In short, what actually happened vs. what the model is giving us. Based on the fit of the model, we can then apply the results for predictions and identify the best set of customers.

There are number of methods of evaluating whether a logistic model is a good model. One such way is sensitivity and specificity. In theory this is how both these terms are defined.

We call True negatives values (TN) (False posives (FP) values) when the real values are 0 and are equal to the predicted values (are different to the predicted values). In a similar way we define False negatives (FN) and True positives (TP).

The sensitivity and the specificity are defined by these values.

	Predicted=0	**Predicted=1**
Actual=0	True Negative (TN)	False Positive (FP)
Actual=1	False Negative (FN)	True Positive (TP)

Sensitivity (also called the true positive rate, or the recall in some fields) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positive}$$

To define what are positive or negative we need to calculate the estimated probabilities,

$$p_i = \frac{1}{1 + e^{-(b_0 + b_1 x_{i1} + \cdots + b_r x_{ir})}} = \frac{e^{b_0 + b_1 x_{i1} + \dots + b_r x_{ir}}}{1 + e^{b_0 + \dots + b_r x_{ir}}}$$

i=1…n, where n is the number of individuals.

Then we need to define a boundary c, for instance c=0.5. Thus the decision is:

If $p_i > c$ then $Y_i = 1$ otherwise $Y_i = 0$.

We would like to maximise both these quantities, Sensitivity and Specificity, but there is often a trade-off: as we decrease the threshold probability, c, we tend to increase the rate of true positives, but decrease the rate of true negatives. This trade-off can be visualised using a ROC curve, allowing us to pick an optimal threshold.

ROC Analysis

The position of the cut-off, c, determines the number of true positives, true negatives, false positives, and false negatives. As we increase our sensitivity (true positives) and can identify more cases with a certain condition, we also sacrifice accuracy on identifying those without the condition (specificity).

Receiver Operating Characteristic (ROC) Curve

A Receiver Operating Characteristic (ROC) curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut off. By tradition, the plot shows the false positive rate (1-specificity) on the X axis and the true positive rate (sensitivity or 1 - the false negative rate) on the Y axis.

The curve starts at (0,0) corresponding to c = 1 and stops at (1,1) corresponding to c= 0. If we want to choose an optimal cut-off point for the purposes of classification, one might select a cut-off point that maximizes both sensitivity and specificity. This choice is facilitated by the use of the ROC curve area.

The best choice for the cut-off point is approximately where the curve starts bending.

The accuracy of a test (i.e. the ability of the test to correctly classify cases with a certain condition and cases without the condition) is measured by the area under the ROC curve (AUC). An area of 1 represents a perfect test, while an area of .5 represents a worthless test. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test; the true positive rate is high and the false positive rate is low.

Statistically, more area under the curve means that it is identifying more true positives while minimizing the number/percent of false positives.

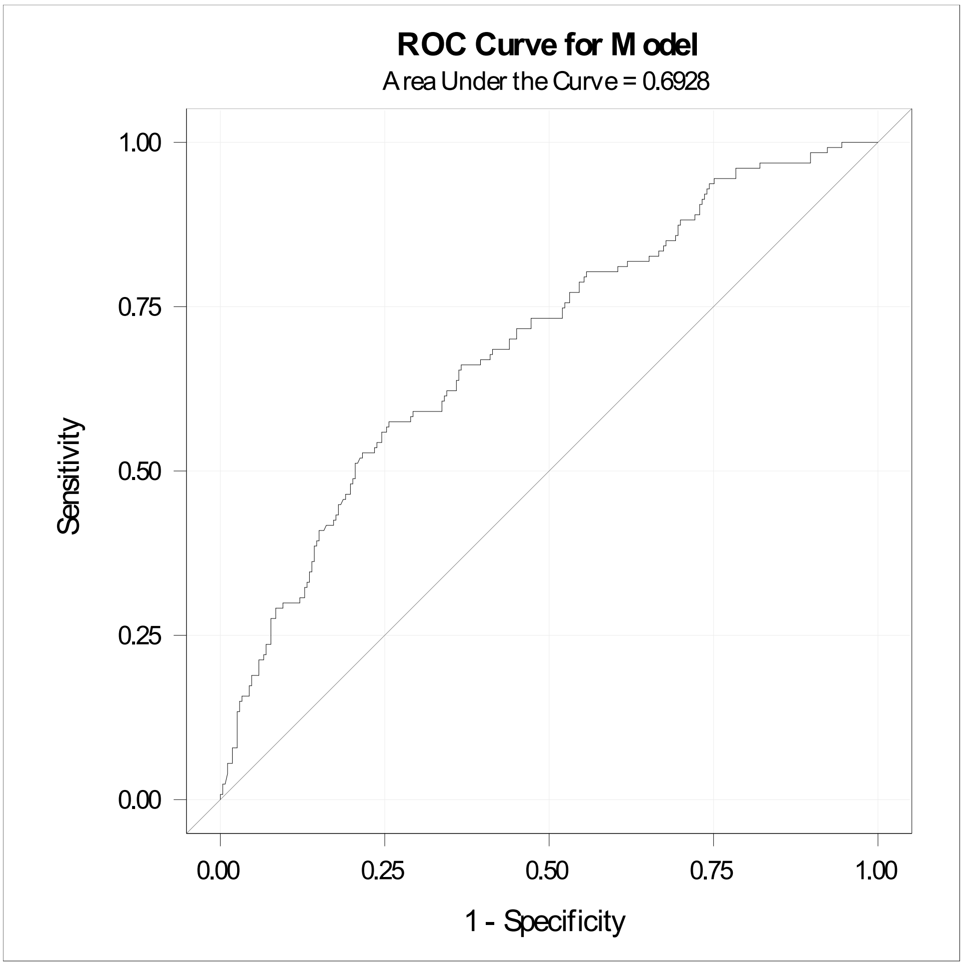
The Area Under the Curve, also referred to as index of accuracy (A), or *concordance index, c, in the "Association of Predicted Probabilities and Observed Responses in SAS*, and it is an accepted traditional performance metric for a ROC curve. c = 0.8 can be interpreted to mean that a randomly selected individual from the positive group has a test value larger than that for a randomly chosen individual from the negative group 80 percent of the time.

How to do that in SAS?

Do the same as before but in **Options**, Plots, add ROC plots. In **Statistics** add Classification table.

For the model with gre, gpa and rank as simple effects we can see by the ROC curve that a possible selection for the cut-off is, a c ≈ 0.314 such that sensitivity ≈ 0.66 and 1-Specificity ≈ 0.36, so that means that 66% of true admitted were correctly identified.

The following is the Roc curve for the logistic regression model in Example 5.



Since the AUC is closer to 0.5 than 1, we can say that the current model do not fit very good the data. Recall that more area under the curve means that it is identifying more true positive while minimizing the number of false positive. Looking at this curve we cannot see clearly what should be the cut-off “c”. For c=0.5, the sensitivity is very low 22.8 and the specificity high, 99.3 (see the classification table). Recall we would like increase both, however that is not possible. Usually we need to sacrifice one, but in this case, 22.8 is very low, so that means that only 22.8% of true admitted were correctly identified.

For a better comprehension see also:

https://video.sas.com/detail/video/4363855630001/introduction-to-roc-curves-and-proc-logistic

https://www.graphpad.com/guides/prism/8/curve-fitting/reg_logistic_roc_curves.htm

and

https://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/

6. Multivariate Data Analysis

Multivariate data analysis is concerned with the study of associations amongst sets of measurements. The techniques that will be studied in this section of the course are concerned with looking at the **interrelationships** that exist between variables.

This differs from the methods used in the previous sections of multiple regression and ANOVA where we were primarily concerned with predicting or explaining the role of **one dependent variable** on the basis of other independent variables/covariates.

The techniques we will consider in the following chapters are:

- **Principal Component Analysis (PCA)**
- **Factor Analysis (FA)**
- **Cluster Analysis (CA)**

A theoretical description of these methods is given in the **Appendices**, which is based in Linear Algebra. Due to the sheer mass of data that these techniques can be applied to, the role of statistical software has become very important in this field. Some of the main objectives of

multivariate techniques in the field of scientific investigation are:

- **Data reduction/simplification** – the aim of this is to collapse the dimensionality of a data set whilst not sacrificing important information contained in the system.

- **Sorting and grouping** – groups of similar variables are created based upon measurement characteristics.

- **Investigating the interrelationships** – the nature of the relationships between variables is of interest, determining the independence and dependences present.

5.1 Principal Component Analysis (PCA)

In the modern working environment it is not unfamiliar for researchers to find themselves confronted with hundreds of different variables potentially entering an analysis. With this mass of data it is virtually impossible to comprehend or visualise the associations that exist amongst the variables, this is further complicated by the redundancy that can exist between the dimensions, leading to high levels of **correlation** and **multicollinearity**.

Due to this redundancy, it should be possible to reduce the observed variables into a smaller number of **Principal Components (artificial variables)** that will account for the majority of the variance in the observed variables.

A Principal Component Analysis is therefore concerned with **explaining the variance-covariance structure of a set of variables** through the **construction of new variables** that are **linear combinations** of the initial variables.

Principal Component Analysis (PCA) seeks a linear combination of all the original variables such that the maximum variance is extracted from the data. It then removes this variance and seeks a second linear combination, which explains the maximum proportion of the remaining variance in the data; this process is continued until all the variability in the system (data) is accounted for.

If we have ***m*** variables in the original data set (say, $X_1, X_2, \dots , X_m$), the PCA will generate a maximum of ***m*** new linear combination variables (called ***principal components***), with the first few principal components, on the majority of occasions, accounting for a large quantity of the variation inherent in the data.

The **researcher must decide** how many principal components to retain for subsequent analysis (explained in more detail later), trading off simplicity (that is, a small number of dimensions is easier to manage) against completeness (that is, a larger amount of dimensions captures a greater amount of the available information).

The calculations used in deriving the principal components are beyond the scope of this course and will thus be omitted (these can be obtained from a variety of books specialising in multivariate statistics).

Example 6

Imagine a six-item questionnaire is used to assess job satisfaction of 12 employees at a small company.

The output obtained is a number for each question ranging from 1 (strongly disagree) to 7 (strongly agree), and the questions are be as follows:

1. My employer treats me with consideration;
2. My employer consults me concerning decisions that affect my work;
3. My employer gives me the support necessary for me to do my job well;
4. My wage is good;
5. My pay is appropriate, given the amount of responsibility that comes with my job;
6. My pay is comparable to those obtained by employees in similar positions at other companies.

It is evident through reading the questions that there will be **redundant variables** in the analysis.

The first 3 questions are linked to the relationship with the employer, if this is a positive relationship you would think that the answers to these questions would score quite highly, however if this was negative, you would think that these would generally score low.

The last 3 questions are all linked to feelings around pay and again you would expect there to be a relationship of the scores achieved for each of these.

Principal components can be derived from either the **correlation** (ρ) or **covariance** matrix. By using the correlation matrix we are basing the results on standardised output, that is the original variables are transformed such that they all have an equivalent variance and mean (that is Mean = 0 and Variance = 1).

This ensures that the data are expressed in comparable units. The total variance in the data is simply the sum of the variances of the observed variables, thus it is equal to m.

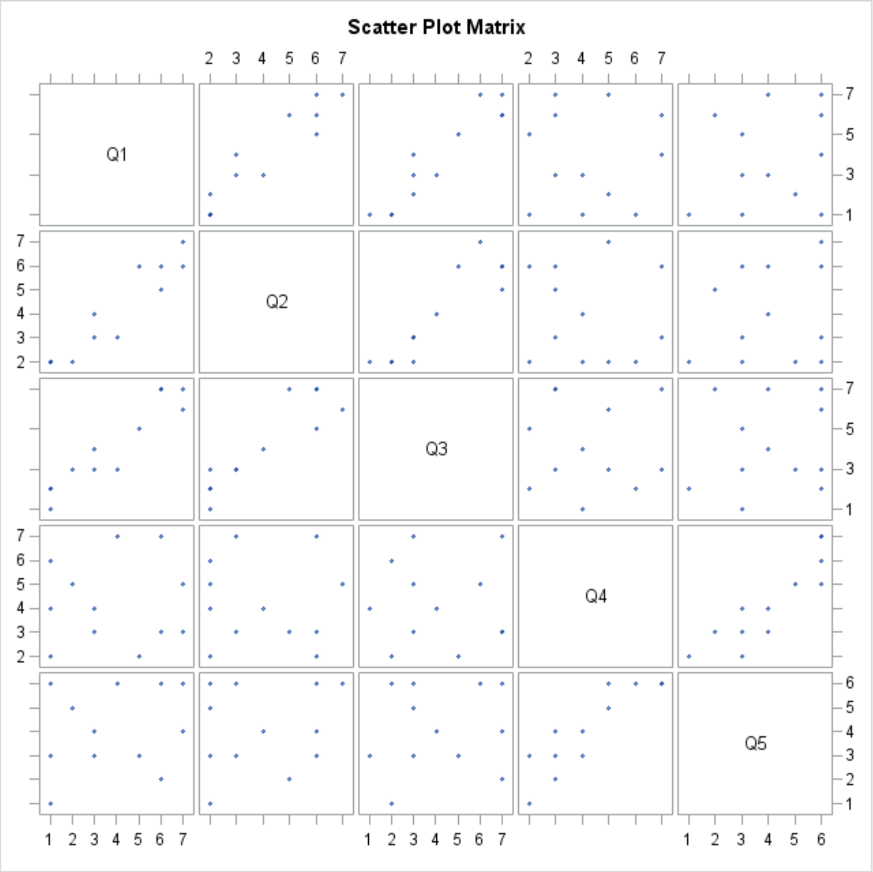
The covariance matrix is **unstandardised**, and due to the PCA being a technique that seeks to maximise variance, it can be sensitive to scale differences across the variables. Situations where the covariance matrix would be appropriate include surveys, which have the same scoring mechanism for each of the variables (such as the example above).

In this course we will be using the **correlation matrix**, for the above example this is represented as follows:

Spearman Correlation Coefficients, N = 12 Prob > |r| under H0: Rho=0

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	1.00000					
Q2	0.94010 <.0001	1.00000				
Q3	0.92806 <.0001	0.87961 0.0002	1.00000			
Q4	0.03052 0.9250	-0.04372 0.8927	-0.02342 0.9424	1.00000		
Q5	0.24366 0.4454	0.21402 0.5042	0.14052 0.6631	0.89260 <.0001	1.00000	
Q6	-0.09892 0.7597	-0.17154 0.5940	-0.16606 0.6060	0.93153 <.0001	0.89056 0.0001	1.00000

As expected, questions 1, 2, 3 have a significant strong positive correlation and similarly so do questions 4, 5 and 6. However, the correlations across these two groups of questions are low and not significant.



5.1.1 Factor Loadings

The factor loadings matrix is the name given to representation of the correlations that exist between the original variables used in the analysis and the principal components.

The factor loadings are also useful in telling us how much of the variance in each of the original variables is accounted for by the principal component.

It is highly desirable to have at least ***three*** (and preferably more) variables **loading** on each component used in the PCA. The criteria for which we can define a variable as being a defining part of a principal component **(loading on a factor)** is purely arbitrary, however cut-off points do exist.

For the purpose of this course we shall apply the following rule of thumb:

Loadings < 0.4 are **weak**

0.4 ≤ Loadings ≤ 0.6 are **moderate**

Loadings > 0.6 are **strong**

You can compute the initial factor loadings for the questionnaire data in SAS UE as follows:

- Select **Tasks**, then **Multivariate Analysis**, then **Factor Analysis**
- Under the **Data,** select the **Questionnaire** data, then select all variables for analysis
- Select the **OPTIONS** tab ensure the factor extraction method is set to PCA.
- As the variables in this data are all measure on the same scale, we could have selected the covariance matrix, however for comparison with later examples we will not select this here, the default is the correlations matrix.
- We will discuss the other options later, for now, select **Run**.

Ideally we do not want variables to be highly correlated with more than one of the principal components (this is commonly referred to as **cross loading**), as this would make the interpretation of the analysis particularly difficult.

5.1.2 How many Factors should we extract?

As a principal component analysis is driven by the necessity to reduce the dimensionality of the data, a natural question would be **"how many components should be retained"**.

Specific criteria have been created for the researcher to aid in this process, as it is essential that the interpretability is as clear as possible for the analyst.

Three of the most popular rules are outlined below, however in each case the application of the rule must be accompanied by a substantial amount of judgement.

1. **Kaisers Rule**

Kaiser (1959) recommended only including components in the analysis that had an **eigenvalue greater than one**. This dictates that a principal component must account for at least as much variation as one of the original variables used in the analysis.

As the purpose of the PCA is to reduce the dimensionality of the data this ensures that no components are retained which are of less value than the original variables.

This criteria is popular due to its simplicity, and the fact that very often it leads to the correct decision in the number of components that should be retained.

However, there are also a number of pitfalls associated with it: it is not always as accurate with large numbers of variables and it can lead to the omission of components when the actual difference in eigenvalues are trivial, for example component 3 could have an eigenvalue of 1.0001 and component 4 of 0.99999.

The eigenvalue obtained from the questionnaire analysis were as follows:

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	2.93631554	0.14523156	0.4894	0.4894
2	2.79108398	2.64866956	0.4652	0.9546
3	0.14241442	0.07366042	0.0237	0.9783
4	0.06875400	0.02764307	0.0115	0.9898

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
5	0.04111093	0.02078980	0.0069	0.9966
6	0.02032113		0.0034	1.0000

Consequently, Kaiser's rule would suggest keeping only the first two components, as the eigenvalues for the subsequent components are all < 1.

2. Proportion of Variance

The number of components to be retained in the analysis can be decided by choosing the number that account for a pre-specified amount of variation. The percentage of variance that is accounted for by say *k* of the *m* components are calculated using the following simple formula:

$$\%\text{ of variation} = \frac{\sum_{i=1}^{k} \lambda_i}{m}.$$

The computation of this value is not necessary though as it is provided in tandem with the analysis. It is possible to set minimum criteria for either the percentage of variation that a variable contributes to the analysis, or alternatively the cumulative percentage of variation.

Usually researchers retain enough components to account for 70% of the variation. This approach ensures that minimum criteria are set for the amount of variation represented, however the cut-off points are arbitrary, consequently it has been criticised for its subjectivity.

For our example above we would include two component in the analysis if we wished to account for a minimum of 70% of the variation.

Note: Communalities is the term that is used to describe the portion of the variance that the variables contribute to the retained principal components. This measurement is more appropriate in factor analysis, and will be described in further detail later.

The priors communality estimates should always be set to one for a principal component analysis.

3. Scree Plot

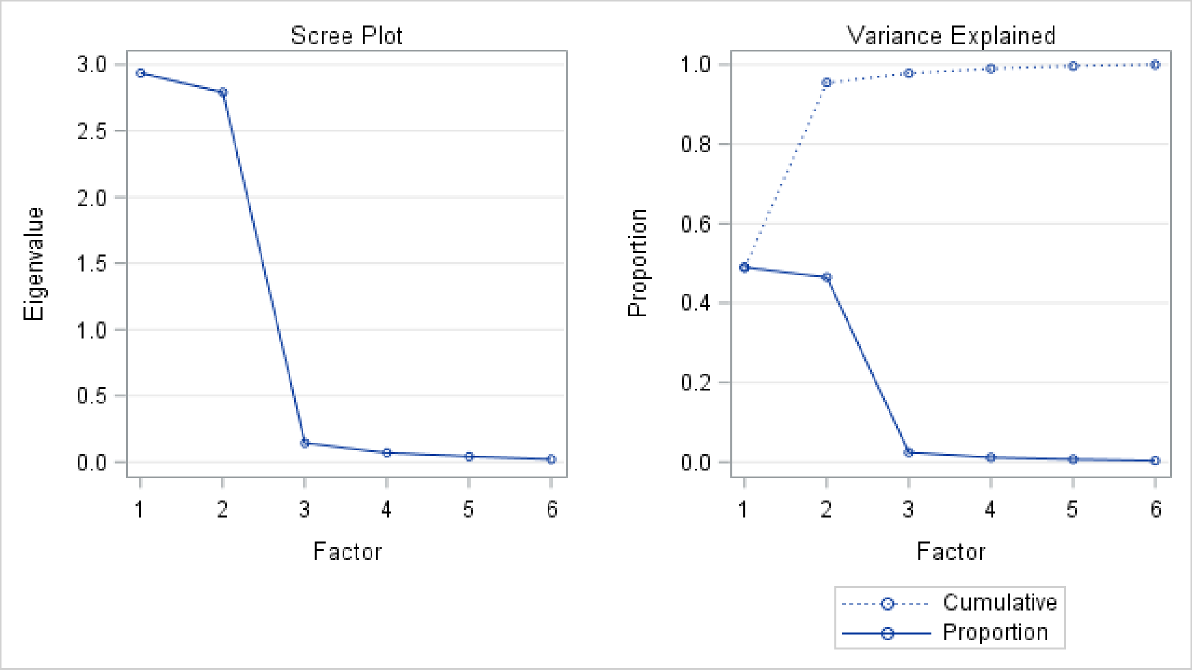
This graphical approach was proposed by Cattell (1966), and involves plotting the eigenvalue associated with each component (y-axis) versus the corresponding component number.

The researcher must look at the curve and try to find an "elbow", that is a point in which the eigenvalues decrease in an approximately linear fashion**.**

Only the components that lie above this point should be retained for further analysis.

Unfortunately the identification of the elbow seldom proves to be a straightforward exercise and the researcher can be confronted with instances of there being several breaks (which would mean maintaining all components until the final break), or alternatively the appearance of no breaks at all.

Below is the scree plot for the satisfaction questionnaire, which suggests keeping 2 components in the analysis.



We will therefore determine that extracting 2 factors is the most appropriate for our questionnaire data, which is what we would have expected.

We will therefore re-run our analysis but this time under the OPTIONS tab, select 2 as the number of factors.

The following output is achieved.

Factor Pattern		
	Factor1	**Factor2**
Q1	0.91024	-0.37907
Q2	0.87143	-0.43412
Q3	0.86895	-0.42936
Q4	0.39860	0.88116
Q5	0.58444	0.78579
Q6	0.30472	0.93844

From the factor loadings matrix we can see that Q1, Q2 and Q3 are all highly loaded on Factor 1. Similarly Q4, Q5, Q6 are all highly loaded with Factor 2.

We are also shown the variance explained by each factor, which in total is approximately 5.73 and how this is dispersed by variable.

Remember that the total variance in the original model is equal to the number of variables, which in this case is 6 (so the 2 factors account for 95.5% of the variation from the original data).

Variance Explained by Each Factor	
Factor1	**Factor2**
2.9363155	2.7910840

Final Communality Estimates: Total = 5.727400					
Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
0.97223365	0.94785494	0.93942324	0.93532009	0.95903906	0.97352855

Examining the factors more closely we can see that Q5 is also moderate to highly correlated with Factor 1. If a variable is highly correlated with more than one factor then **the rotation of the principal components** is necessary.

After extracting the principal components, rotation is applied to maximise high correlations and minimise the low correlations. By default the analysis is run with no rotation, which essentially creates a set of principal components that explains as much of the variance in the original set of variables as possible.

However, the solutions can prove difficult to interpret because variables can be highly correlated on several principal components, and we commonly find that the first component is an average of the original variables.

Thus we rotate the components, the rotation involves moving the axis without changing the relative locations of the points to each other. Two alternatives are available: **Oblique rotation or Orthogonal rotation.** Orthogonal rotation involves the rotation of the principal components such that the principal components remain uncorrelated producing factors which are easier to interpret, hence will be the method used within this course. Several types of orthogonal rotation exist, including: ***Varimax***: this is by far the most popular form of rotation; it simplifies the principal components by maximising the variance of the loadings within principal components and across variables (operates on the **columns** of the loadings matrix).

We can rotate the initial factors produced in our model as follows:

- In the **OPTIONS** tab, under **Rotation**.
- Select the **Rotation and Plots** tab.
- Select for now **Varimax.**
- Select **Run.**

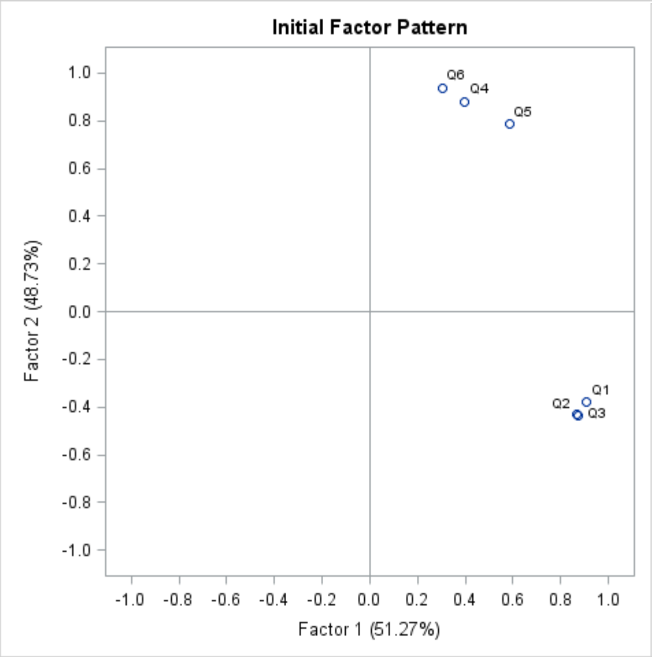
The factor loading matrix is now as follows.

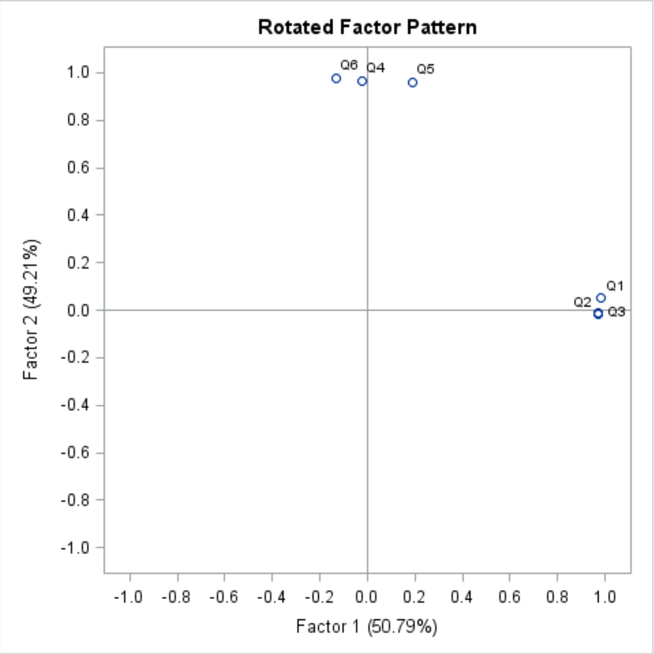
Rotated Factor Pattern		
	Factor1	**Factor2**
Q1	0.98466	0.05184
Q2	0.97347	-0.01458
Q3	0.96917	-0.01136
Q4	-0.02165	0.96688
Q5	0.18715	0.96126
Q6	-0.13107	0.97793

It can clearly be seen that each variable is highly correlated with only one factor. The variance explained by the factors has however shifted slightly as can be observed from the following table. The total variance explained is still the same, however it is now more evenly spread over the 2 factors.

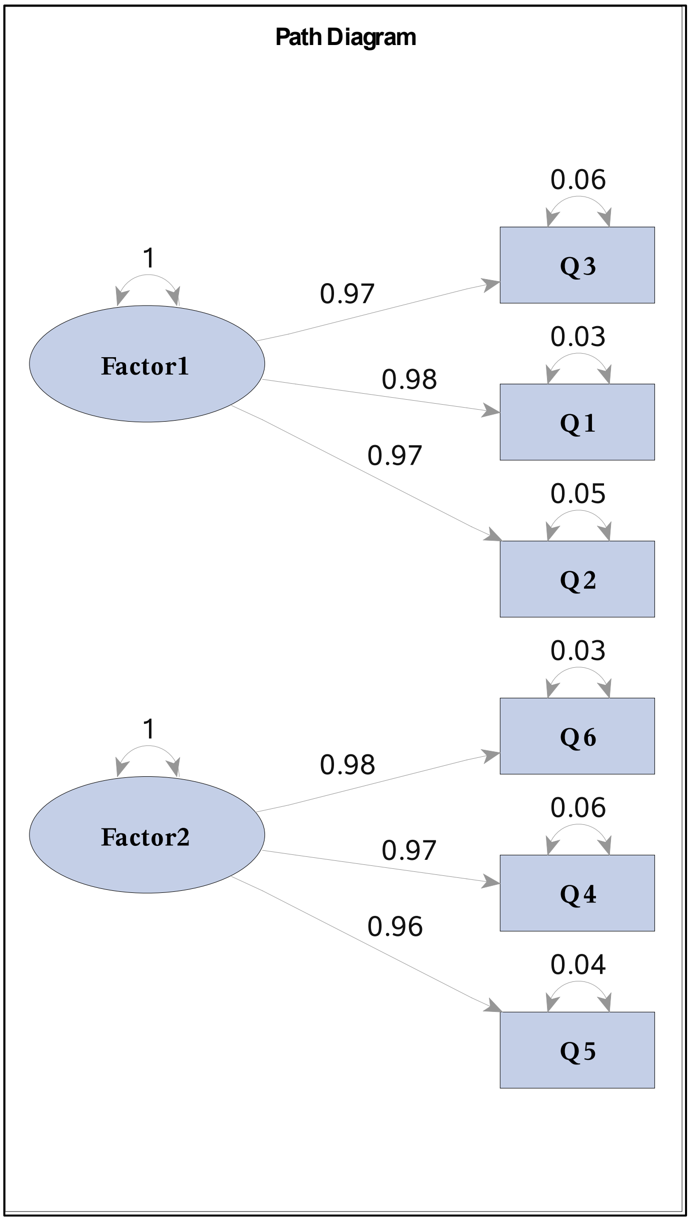
Variance Explained by Each Factor	
Factor1	**Factor2**
2.9091559	2.8182436

To help us understand what is happening when we rotate our factors we can examine the factor patterns matrix both before and after rotation, these can be found under the plot options in the OPTIONS tab.





Along with the following Path diagram. This explain the amount of variation account for in each question by the new Factors.



We will now continue with this topic in more detail including the communalities, naming the components and consider a further example.

5.1.3 Name the Components

Assigning a label for the newly formed principal component can prove to be complex, and if a name is assigned it must have face validity and/or be rooted in theory. It is notoriously difficult to give valid meaning to the factors.

A general rule is that if a component is to be given a name, at least 80% of the variables highly loaded with it would be correctly assigned.

Two important rules applied in assigning these labels are as follows:

- Be succinct: use one or two words;
- Communicate the nature of the construct.

Thus, we need to look at the variables that are marker variables (that is load highly on a component), and determine characteristics that could be deemed as making them similar.

Additionally it is worth looking at which variables **do not** load on a principal component, to determine what features that the component is not representing.

From our example we could label the first component as being representative of "***Employer Satisfaction***", and the second principal component as "***Wage Satisfaction***".

However, as stated previously assigning a name to a component can frequently prove to be extremely complicated.

Frequently, a principal component analysis is not an end to itself, but an intermediate step on the way to further analysis of the data. For subsequent analysis we require the location of each of the original variables in the newly constructed factor space.

These are the factor (component) scores. These give us a value for the observations (rows) on each of the principal components (columns) retained in the analysis.

This output can be automatically generated in conjunction with the analysis as follows:

- Under the **OUTPUTS** tab, select to **create factor scores data set** and Run.
- You may wish to select the exact number of factors you which to generate, these can then be viewed in the **OUTPUT DATA** window of the results.

Numerous alternative arbitrary "rules of thumb" exist dictating how many cases/subjects (***rows***) are needed in order to perform a principal component analysis such as:

- the subjects to variable ratio should be no lower than 5 (Bryant and Yarnold, 1995);
- the rule of 100: the number of subjects should be the larger of 5 times the number of variables, or 100.

However, the criteria differ greatly depending on the field of research.

5.2 Factor Analysis

Factor Analysis is similar to principal component analysis (PCA) in that it is a technique that is applied to data in order to gain an understanding of the interrelationships between variables in the system.

However, PCA and factor analysis should not be confused as being the same processes, even though the steps involved in the analyses are similar.

In principal component analysis the major objective is to select a number of components that will express as much of the total variance in the data as possible.

The factors formed in the factor analysis are generated to identify the latent (hidden) variables that are contributing to the common variance in the data.

The essential purpose of factor analysis is to describe, if possible, the **covariance relationships among many variables in terms of a few underlying (but unobservable), random quantities called factors.**

Examples of variables which cannot be directly measured (*latent variables):*

- intelligence
- social class
- happiness

However, we can measure these concepts indirectly:

- IQ test, performance in school
- Occupation, salary, value of home

Factor analysis assumes that the old variables can be represented as a linear combination of some unknown variables common to all.

Data reduction is achieved as each original variable can be reconstructed using the m common factors $F_1, F_2, \ldots, F_m$, where m is less than the number of original variables.

A factor analysis can be used for several purposes, some of which are:

- to reduce the dimensionality of a large data set for modelling purposes;
- to select a subset of variables from the original data based on which of the original variables have the highest correlations with the newly formed principal factors;
- to create a set of factors to be treated as uncorrelated (orthogonal) variables as one approach to handling multicollinearity in such procedures as multiple regression.

As with a principal component analysis the researcher is confronted with numerous decisions when performing a factor analysis, many of which are **subjective**.

In this course we will be specifically concentrating on **Exploratory Factor Analysis**, here a priori assumption is that any indicator may be associated with any factor. This is the most common form of factor analysis. There are no prior assumptions and one uses the factor loadings to understand the factor structure of the data.

The other form of factor analysis is referred to as **Confirmatory Factor Analysis**; allows you to test very specific hypotheses regarding the number of factors, factor loadings, and factor inter-correlations. However this is beyond the scope of the course.

As with the principal component analysis, the application of the factor analysis is demonstrated best through example.

Consider the following hypothetical situation; psychological testing was performed on 200 primary school children.

Five tests were administered, where the children were tested on the following:

- paragraph comprehension (PARA)
- sentence completion (SENT)
- word meaning (WORD)
- addition (ADD)
- counting dots (DOTS)

The factor model assumes that we can write these five variables as a linear combination of $\boldsymbol{m}$ **latent variables**.

For the factor loadings matrix (say, L), we have:

$$PARA = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1$$

$$SENT = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2$$

$$WORD = l_{31}F_1 + l_{32}F_2 + \ldots + l_{3m}F_m + \varepsilon_3$$

$$ADD = l_{41}F_1 + l_{42}F_2 + \ldots + l_{4m}F_m + \varepsilon_4$$

$$DOTS = l_{51}F_1 + l_{52}F_2 + \ldots + l_{5m}F_m + \varepsilon_5$$

Where l_{ij} are the factor loadings of the ith variable on the jth factor and ε_i, are the errors.

Several alternatives are available in **SAS** for the extraction method. We have already studied the principal component analysis technique, however other alternatives are available.

Iterated principal factor analysis (IPFA) is the form of extraction that is most commonly used, and for the purpose of this course we shall restrict ourselves to using this.

IPFA a form of factor analysis which seeks the least number of factors which can account for the common variance (correlation) of a set of variables, whereas the more common principal components analysis (PCA) in its full form seeks the set of factors which can account for all the common and unique variance in a set of variables.

5.2.1 Communalities

The amount of variation in X_i (original variable i) that is accounted for by the common factors is also referred to as the **communality** of the original variable X_i.

The nearer that value is to 1, the less error variance there is for the original variable and the more perfectly X_i is represented by the underlying common factors.

However, on the other hand if this sum was nearer to 0, then this would suggest that nearly all of the variation in X_i would be explained by the specific factor.

Thus, if the factor loading matrix for our example (two factors extracted as decided by PCA) was:

	Factor 1	**Factor 2**
PARA	0.81	0.06
SENT	0.72	0.08
WORD	0.91	0.01
ADD	0.02	0.69
DOTS	0.11	0.92

This implies that:

$$\text{PARA} = 0.81F_1 + 0.06F_2 + \varepsilon_1$$

$$\text{SENT} = 0.72F_1 + 0.08F_2 + \varepsilon_2$$

$$\text{WORD} = 0.91F_1 + 0.01F_2 + \varepsilon_3$$

$$\text{ADD} = 0.02F_1 + 0.69F_2 + \varepsilon_4$$

$$\text{DOTS} = 0.11F_1 + 0.92F_2 + \varepsilon_5$$

Looking firstly at the variance of the variable PARA. Since we used the correlation matrix in the extraction of factors, the variable PARA has been standardised, and thus has unit variance. This means F1 and F2 are uncorrelated, with zero mean and unit variance.

$$\mathrm{VAR[PARA]} = \mathrm{VAR}[0.81F_1 + 0.06F_2 + \varepsilon_1] = 0.81^2 + 0.06^2 + \mathrm{VAR}[\varepsilon_1] = 1.$$

The specific variance for the variable PARA is therefore given by:

$\mathrm{VAR}[\varepsilon_1] = 1 - (0.81^2 + 0.06^2) = 0.3403$

Communality of the variable PARA is $0.81^2 + 0.06^2 = 0.6597$ (65.97%).

Remember the nearer the communality is to 1, the more variation can be explained by the common factors.

In conclusions, for the total variation in the variable PARA, 66% of the variation in the variable is accounted for by common factors in the model and 34% of the variation is accounted for by a unique variable only associated to PARA.

The resulting communalities for all variables are:

PARA	SENT	WORD	ADD	DOTS
0.6597	0.5248	0.8282	0.4765	0.8585

A key issue on interpreting communality is the interpretability of the factors.

- A communality of, say, 0.75 for a variable seems high, but is meaningless unless the factor is interpretable.
- A communality of, say, 0.25 seems low but may be meaningful if the item is contributing to a well-defined factor.

That is, what is critical is not the communality coefficient, but rather the extent to which the item plays a role in the interpretation of the factor, though often this role is greater when communality is high.

Through studying the factor loading matrix we find that PARA, SENT and WORD are highly loaded on the first principal factor. This would suggest that the first factor could represent the "Verbal ability" of the student.

The second factor is highly correlated with ADD and DOTS, which suggests that this could be representative of the "Numerical ability" of the student.

We will be using the IPFA method with the prior communality estimates set to be the squared multiple correlation with all other columns.

5.2.2 Anomalies

The eigenvalues that result from the analysis are not interpreted in the same was as for the PCA. This is attributable to the fact that the sum of the eigenvalues does not amount to the number of original variables used in the analysis, which is explained by the fact that it is only the **common variance** that is included in the analysis.

Negative eigenvalues can occur with a factor analysis, which is not intuitively appealing just as a negative variance is not.

With data that do not fit the common factor model perfectly, you can expect some of the eigenvalues to be negative. If an IPFA converges properly, the sum of the eigenvalues corresponding to rejected factors should be 0; hence, some eigenvalues are positive and some negative.

If an IPFA fails to yield any negative eigenvalues, the prior communality estimates are probably too large. Negative eigenvalues cause the cumulative proportion of variance explained to exceed 1 for a sufficiently large number of factors.

Another anomaly is when the final communality estimates exceed 1. This situation is referred to as the **ultra-Heywood case** (when the communality equals 1, it is referred to as the Heywood case). An ultra-Heywood case is a clear indication that some unique factor has negative variance and invalidates the solution, possible causes include:

- bad prior communality estimates;
- too many common factors;
- too few common factors;
- not enough data to provide stable estimates;
- the common factor model is not an appropriate model for the data.

Example 7

In a summer school camp for American 15 – 18 year olds, the ability of the 202 students in different athletic events was assessed to see if there were any relationships in performance over the events.

The times and distances achieved by the students in thirteen different disciplines was recorded, namely (with measurement units in brackets):

- Shot Putt (metres);
- High Jump (metres);
- 100m (seconds);
- 1500m (min);
- 200m (seconds);
- Javelin (metres);
- 800m (minutes);
- Pole Vault (metres);
- Discuss (metres);
- Long Jump (metres);
- 400m (seconds);
- 50m hurdles (seconds);
- Cross Country (minutes).

A PCA was performed to determine the number of factors to interpret; the correlation matrix was used as the variables are measured differently.

By Kaiser’s criteria it was seen that four factors should be extracted. These accounted for 82.77% of the overall variation. An Orthogonal Varimax rotation was applied.

An IPFA was performed on the correlation matrix, using the squared multiple correlation with the other columns as the prior communality estimate.

The preliminary eigenvalue total was 9.502 with an average of 0.731, after iterations were performed that converged to the communality values the eigenvalue total changed to 9.883 and an average of 0.7602.

This tells us that (9.883/13 x 100) 76.02% of the variation in the system is shared amongst the new principal factors.

The breakdown of the eigenvalues for the principal factors were as follows:

Eigenvalues of the Reduced Correlation Matrix: Total = 9.88318387 Average = 0.76024491				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.71909769	2.24340095	0.4775	0.4775
2	2.47569674	0.82880942	0.2505	0.7280
3	1.64688732	0.60498488	0.1666	0.8946
4	1.04190244	0.77859765	0.1054	1.0000
5	0.26330478	0.17795375	0.0266	1.0267
6	0.08535104	0.07478411	0.0086	1.0353
7	0.01056693	0.01010549	0.0011	1.0364
8	0.00046144	0.01487368	0.0000	1.0364
9	-.01441224	0.02812100	-0.0015	1.0350
10	-.04253324	0.00422145	-0.0043	1.0307
11	-.04675469	0.02781701	-0.0047	1.0259
12	-.07457170	0.10724092	-0.0075	1.0184
13	-.18181262		-0.0184	1.0000

We base our interpretations on the first four components, so need not concern ourselves with the negative eigenvalues.

The variance explained by each factor was as follows.

Variance Explained by Each Factor			
Factor1	**Factor2**	**Factor3**	**Factor4**
2.8353706	2.6327654	2.4210357	1.9944125

The factor pattern (loadings) matrix is as follows.

Rotated Factor Pattern				
	Factor1	**Factor2**	**Factor3**	**Factor4**
Shot Putt	0.07399	0.99633	0.03084	0.00423
High Jump	0.38509	0.10146	0.69146	0.11510
100m	0.77037	0.10018	0.18858	0.34519
1500m	0.21198	-0.04447	0.08957	0.91167
200m	0.91991	0.06619	0.16093	0.14803
Javelin	0.13987	0.91131	0.05199	0.00343
800m	0.37056	0.00284	0.08702	0.73620
Pole Vault	0.14523	0.02344	0.94382	0.06440
Discuss	0.03802	0.86410	0.09385	0.00353
Long Jump	0.10671	0.04611	0.85688	0.06276
400m	0.82653	0.05510	0.15058	0.16012
50m Hurdles	0.56557	0.17217	0.45056	0.14118
Cross Country	0.05597	0.02990	0.05170	0.64286

Through studying the matrix we can make the following interpretations.

Factor 2 is highly loaded with the Shot Putt, Javelin and Discuss. Thus this factor could be considered as being representative of “Explosive arm strength”, it accounts for 26.6% of the common variation, and 21.3% of the total variation in the data.

Factor 3 is highly loaded with high jump, pole vault, long jump and is moderately loaded on 50m hurdles. Thus this factor could be considered as being representative of “Explosive leg strength”, it accounts for 24.5% of the common variation, and 18.6% of the total variation in the data.

It is evident from the matrix that 50m hurdles is a marker variable on more than one of the factors, that is, it is cross loading on factors 1 and 3. However, it does not contradict our names for the new factors, as 50m hurdles is an amalgamation of sprinting ability and explosive leg strength.

Factor 4 is highly loaded with 1500m, 800m, and cross country. Thus this factor could be considered as being representative of “Running endurance”, which accounts for 20.2% of the common variation, and 15.3% of the total variation in the data.

The final communalities for the original variables used in the analysis were as follows:

Final Communality Estimates: Total = 9.883584												
Shot Putt	**High Jump**	**100m**	**1500m**	**200m**	**Javelin**	**800m**	**Pole Vault**	**Discuss**	**Long Jump**	**400m**	**50m Hurdles**	**Cross Country**
0.999	0.650	0.758	0.886	0.898	0.853	0.687	0.917	0.757	0.752	0.735	0.572	0.420

This table tells us that the first four principal factors are accountable for as much as 99.9% of the variation in Shot Putt, and going down to as little as 42.0% in Cross Country.

These new factors describe the underlying structure of the original data in four dimensions, and factor scores can be made for them so that they can be used in subsequent analysis.

5.3 Cluster Analysis

Exploring data for structure or natural subgroupings is a very important statistical technique. The newly formed subgroups can provide a means for assessing dimensionality, identifying outliers, and identifying possible interesting hypotheses about relationships in the data.

Cluster analysis is a primitive technique, as no assumptions are made prior to analysis regarding group size or structure. The newly formed groups are formed on the basis of distances between observations.

Simply speaking, cluster analysis entails categorisation; the division of a large set of observations into a number of smaller distinct groups, such that observations within the same group/*cluster* are similar (that is they possess similar characteristics), and the observations in different groups are dissimilar.

We shall be concentrating on the clustering of cases (observations) in this course, however it is also possible to cluster the variables to determine which are close to each other in terms of the individual's response. This could be considered as an alternative to factor analysis.

5.3.1 Measure of Distance

In order to perform a cluster analysis we must define some measure to represent the closeness/similarity of two observations; the ***distance*** between the observations. This representation of proximity can be represented in numerous ways.

For data that has metric properties (measured on interval or ratio scales), the most natural form of representation would be via a distance measure. The most commonly used form being **Euclidean distance**, which numerically is defined as:

$$d_{ij} = \sqrt{\left[\sum_k (x_{ik} - x_{jk})^2\right]}$$

With $\boldsymbol{d_{ij}}$ representing the Euclidean distance between observations *i* and *j*.

It is common for this calculation to be applied to standardised data. Consequently, this has the effect of assigning equal weights to the variables used in the analysis, and thus for the purpose of this course we will be using **squared Euclidean distances.**

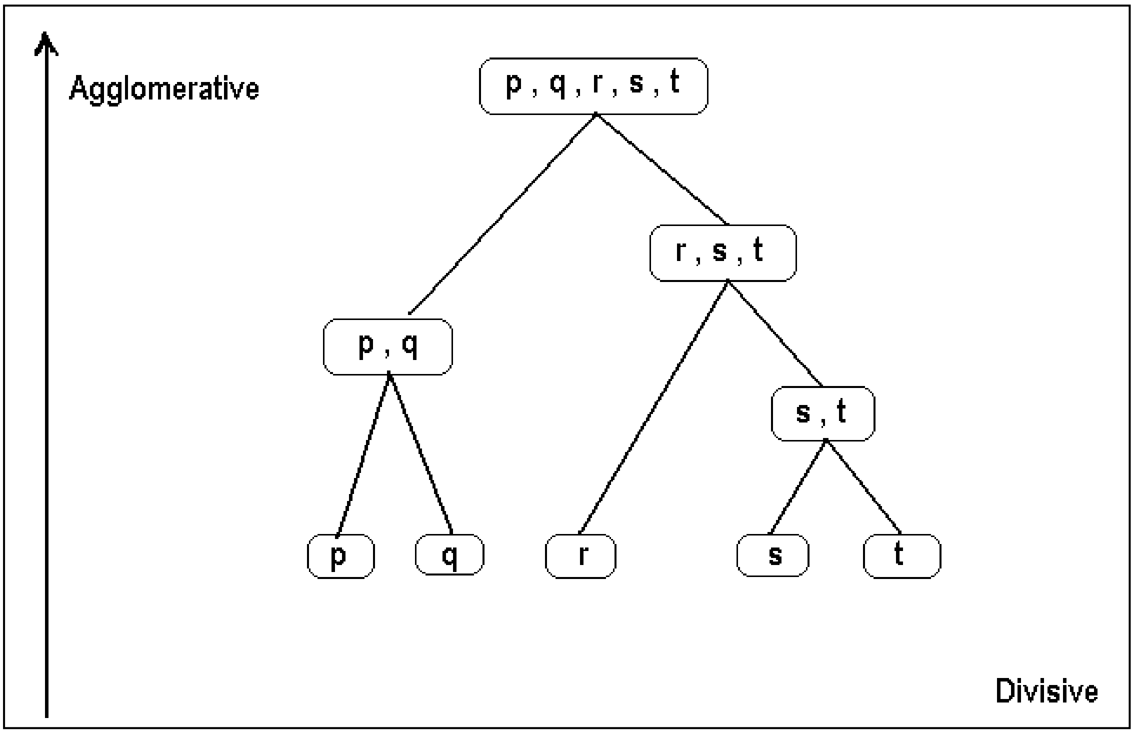
Clustering techniques can be considered as falling into two general categories, **hierarchical** and **non – hierarchical**.

5.3.2 Hierarchical Clustering

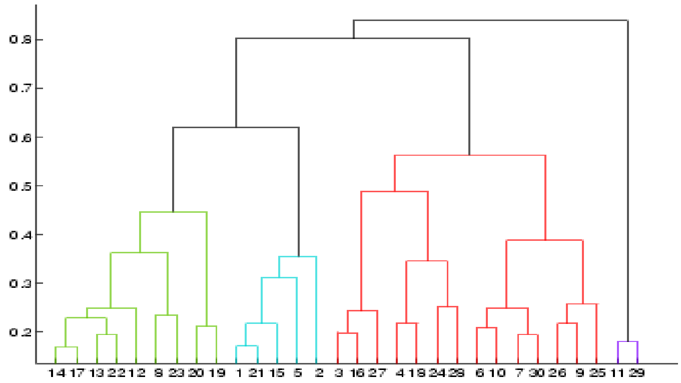
Hierarchical clusters are arranged such that one cluster can be entirely contained within another, but no other kind of overlap between the clusters can exist.

This type of clustering can be ***agglomerative*** or ***divisive***. In the ***agglomerative method*** we begin with N clusters, that is, each observation is considered as being its own cluster. In successive steps the two closest clusters are combined, consequently reducing the number of clusters by one in each step. Until eventually at the last step all observations are fused into one cluster.

In ***divisive methods*** the opposite methodology is applied. We begin with one cluster and in successive steps we split off the cases (observations) that are most dissimilar to the remaining ones, i.e. objects in one subgroup are "far from" the objects in the other.

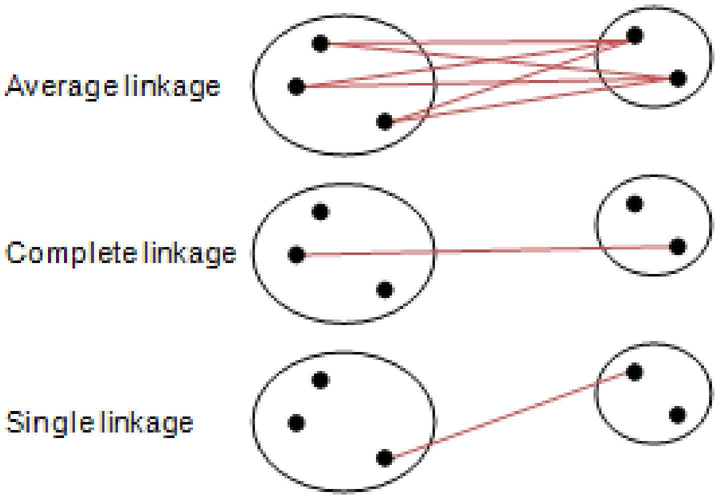


The results of both the agglomerative and divisive methods can be represented in the form of a ***dendogram***, a hierarchical tree diagram generated from the iterative procedure**.** This plot can be used to illustrate the mergers or divisions that have been made in each successive step, and can be used to determine the number of clusters to create, and the respective members of each cluster (this will be demonstrated by example later).



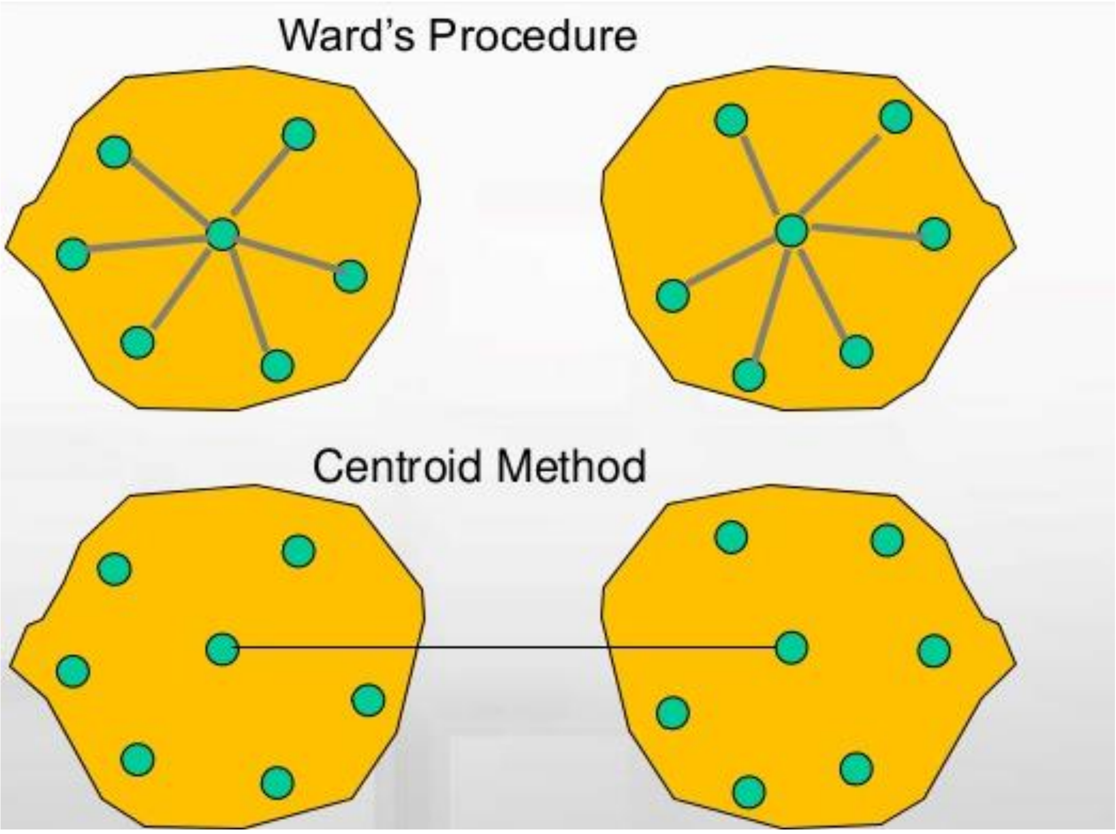
In most statistical programs the agglomerative method is applied, and this is the case with ***SAS***. Several different measurements are available to define the **distance between clusters**, including: single linkage, complete linkage, average linkage, centroid method, Ward's minimum variance method.

- Single linkage - the distance between clusters is defined as that between the nearest neighbour in each cluster.
- Complete linkage - the distance between the clusters is determined between the two elements (one from each cluster) that are most distant. Thus, this ensures that are within some maximum distance of one another.
- Average linkage - the distance between two clusters is considered to be the average distance between all pairs of items where one member of a pair belongs to each cluster.



- Centroid method - the objects in each cluster are “averaged” (calculating the cluster centroids), with the difference between the two centroids representing the distance between the clusters.

- Wards minimum variance method - this method seeks to join clusters that give the smallest within-group variance. This has the consequence of giving clusters that are similar in size, convex and compact.



5.3.3 Non – Hierarchical Clustering

A popular alternative is the partitioning clustering technique, which entails the separation of the sample into a predetermined number, say *K,* of non-overlapping groups.

The objects in the same group are relatively similar, and those in different groups are dissimilar. In order to achieve this, we must measure the within-group similarity and between group-difference. Then construct a way of finding the best way to partition these groups.

The ***K-means clustering*** is a popular non-hierarchical clustering technique.

For a pre-specified number of clusters (***K***), the algorithm can be described as follows:

1. Divide the data into ***K*** initial clusters. This can be pre-specified, however is determined by an arbitrary process according to ***SAS.***
2. Calculate the means or centroids of the ***K*** clusters.
3. For an observation, calculate its distance to each centroid. If the case is closest to the centroids of its own cluster, leave it in that cluster; otherwise, reassign it to the cluster with the closest centroid.
4. If required, recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
5. Repeat Step 3 for each case.
6. Repeat Steps 2 – 4 until no cases are reassigned.

Options available in ***SAS*** for this clustering method include: specifying a maximum number of clusters, specifying a maximum number of iterations, and select a seed replacement method from the **Seed replacement** drop-down list. If you select **random**, you can also select the **Specify random seed** check box and specify a random seed.

Other methods of clustering include (not studied in this course): **Overlapping clusters** – the number of observations belonging simultaneously to two clusters can be constrained. Alternatively, they can be unconstrained allowing any degree of overlap in cluster membership. **Fuzzy clusters** – these are defined by a probability or grade of membership of each object in each cluster. These can be disjoint, hierarchical or overlapping.

5.3.4 Determining the Number of Clusters – Agglomerative

Determining the number of clusters to use in an analysis is not a straightforward exercise. This is achieved through a variety of mechanisms, the principal methods are detailed below.

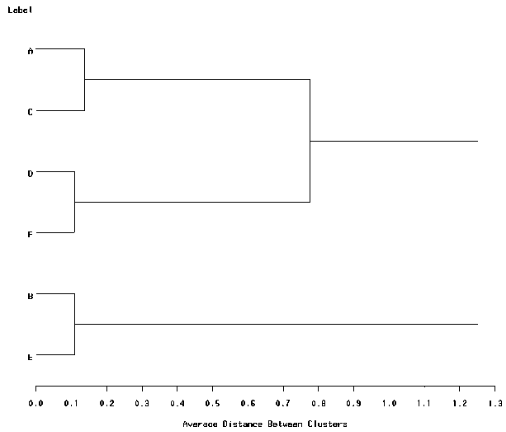
1. Dendrogram

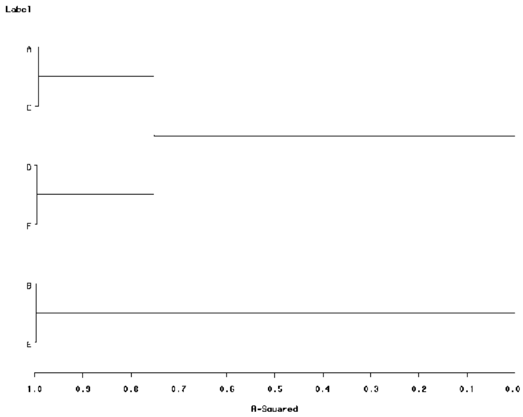
This is a graphical representation that is used in agglomerative methods to denote the hierarchy of nested cluster solutions: from the one-cluster solution to the *n*-cluster solution. This can be represented vertically or horizontally in ***SAS***.

For example, consider the following (unstandardised) dataset:

Observation	x coordinate	y coordinate
A	2.1	3.1
B	5.1	2.4
C	1.9	2.9
D	3.5	3.2
E	5.0	2.2
F	3.6	3.4

The dendrograms from the subsequent ***Average Linkage*** cluster analysis would be as follows:





The dendrogram on the left based on minimum distance between clusters is provided by default. However to obtain the clusters against the corresponding R squared values requires the following to be added to the PROC TREE code: **Height _RSQ_;** .

The dendrograms can be used to determine whether one of the nested solutions provides a better representation of the data through looking for a wide range of distances on the plot for which the number of clusters in the solution do not change.

In the above plots it is evident that 3 clusters would be best, namely: {AC}; {DF} and {BE}. However, in reality the choice of the number of clusters is not as clear-cut, requiring a considerable amount of subjectivity and judgement on the part of the analyst.

2. Statistical Output

A variety of statistics are generated in the output of a cluster analysis. When one performs a cluster analysis in SAS with ***average linkage, centroid method or Wards minimum variance method*** a “Cluster History” table is generated:

Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**	**Pseudo t-Squared**	**Tie**
30	CL68	CL37	16	0.0040	.908	.788	25.3	40.8	5.3	-
29	CL55	OB98	5	0.0017	.906	.783	25.8	41.8	2.7	-
28	CL59	OB21	3	0.0014	.905	.777	26.4	43.0	2.0	-
27	CL50	CL40	8	0.0033	.902	.771	26.4	43.3	5.0	-

This table contains several important statistics, including three key statistics that are used to determine the number of clusters to generate.

Cubic Clustering Criterion (*CCC*) – values greater than 2 indicate good clusters; values between 0 and 2 indicate potential clusters but should be treated with caution, and large negative values indicate outliers.

Pseudo *F*-statistic (PSF) – relatively large values of this indicate a stopping point. This can be determined by reading down the PSF column in the output and comparing neighbouring values for clusters.

Pseudo t^2-statistic (PST2) – a rule of thumb for interpreting the pseudo t^2 statistic is to move down the column until a value is discovered that is markedly higher than the previous value and move back up the column by one cluster.

It is preferable to have a consensus among the three statistics outlined above, that is, local peaks of the *CCC* and pseudo *F*-statistic, combined with a small value of the pseudo t^2-statistic and a larger pseudo t^2-statistic for the next cluster fusion.

It should also be noted that the pseudo *F* and t^2-statistics might be useful indicators for the number of clusters, however they are **NOT** distributed as *F* and t^2 random variables.

The other statistics in the table are defined as follows:

- **Clusters Joined –** this details the clusters (CL*i*)/observations (OB*j*) that were joined;
- **Freq –** the number of observations in the current cluster;
- **Semipartial R-Square –** semi-partial R^2 value, this represents the decrease in the proportion of variance accounted for by joining two clusters;
- **R-Square –** squared multiple correlation R^2, which is the proportion of variance accounted for by the clusters;
- **Approximate Expected R-Square –** the approximate estimated value of R^2 under the uniform null hypothesis;
- **Tie –** this lists the ties for minimum distance, with a blank value denoting the absence of a tie.

5.3.5 Determining the Number of Clusters – K–Means Clustering

The *K*–means algorithm finds a cluster solution for a pre-specified value *k.* The decision as to what the optimal value for *K* should be is made on the basis of performing the analysis several times for different values of *K*, this will involve some trade-off between the simplicity of the solution (where a smaller number of clusters is better) and its adequacy (if the reduction of the within group heterogeneity is the objective, then more clusters are favourable).

A statistic that captures the trade-off between simplicity and adequacy is the pseudo-*F* statistic. **The larger the pseudo-*F*, the more effective the partitioning is in reducing within group heterogeneity.**

The pseudo-*F* does not increase monotonically, but will attain a maximum value for some *K.* An increase in the number of clusters that results in a decrease in the pseudo-*F*, is an indication that the additional complexity of the solution is probably not worthwhile.

Simulations have demonstrated that the pseudo-*F* statistic proved to be the best of 30 different criteria for determining the number of clusters.

One of the best methods for interpreting the clusters is to examine the cluster centroids, which involves calculating the mean value of each variable across the variables assigned to the clusters. This indicates which clusters are relatively high or low on each of the variables.

Cluster Means		
Variable 1	**Variable 2**	**Variable 3**
81.4	70.4	65.9
41.2	82.1	32.2
69.2	50.4	49.3

However, the centroids give no indication of overlap between the clusters for the variables. Consequently in the output we also have a decomposition of the total sum of squares for the variable into the within-cluster sum of squares (that is the sum of the squared deviations between each observation and its cluster mean), and the between cluster sum of squares. Compact separate clusters would have a small within group's sum of squares and large between group sums of squares.

Two summary measures can be used to represent these measures: the ratio of the between-cluster sum of squares to the total squares (similar to the R^2 value), or we can look at the ratio of the between-cluster sum of squares to the within-cluster sum of squares (similar to R^2/ (*1* – R^2).

Comparisons of these measures across the variables make it evident which are the most important in determining the differences among clusters.

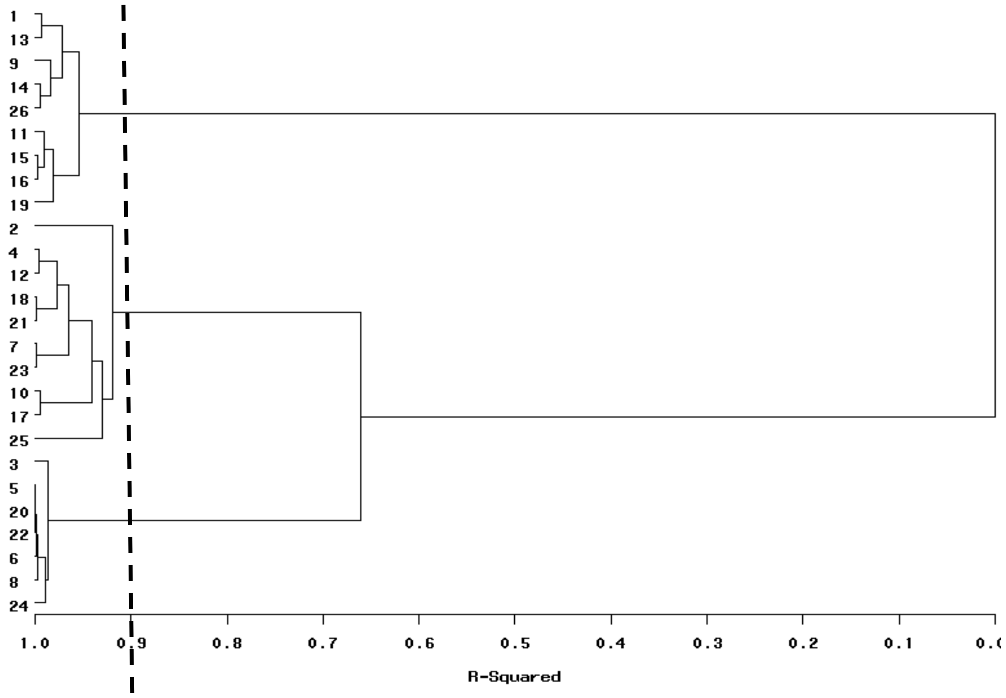
Example 8

Psychological tests were performed on 26 people in a manufacturing company, the tests involved assessing three characteristics of the workforce.

Leadership	Leadership ability of individual (0 – 100);
Team	Ability to work as part of a team (0 – 100);
Individual	Ability to work individually (0 – 100).

A cluster analysis was performed on the data to determine if the workforce could be clustered in to groups with similar characteristics.

The hierarchical agglomerative method of cluster analysis was used with Wards minimum variance method used as the cluster method. The dendrogram from the analysis is as follows.



The dashed line drawn on the plot indicates that 3 clusters should be extracted; as there is wide range of distances on the plot for which the number of clusters in the solution do not change (R^2 0.65 to 0.9).

Further confirmation that three clusters should be formed is given by studying the CCC, pseudo – F, and the Pseudo –t^2 statistics that is given in the output, as outlined below.

Number of Clusters	Freq	Semipartial R-square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
5	8	0.0129	.941	.795	10.2	83.2	5.8	
4	9	0.0106	.930	.737	11.9	97.5	2.8	
3	10	0.0112	.919	.646	12.7	130	2.4	
2	17	0.2583	.660	.492	3.39	**46.7**	**71.4**	
1	26	0.6605	.000	.000	0.00	.	46.7	

We can obtain details of the cluster membership by re-running the cluster analysis under the same conditions, but with the **Clusters** set to **3** under the **Results tab** and in the **Save output data** section.

This has the consequence of creating a new data set from the resulting analysis, which is stored under the name "**Tree data**". By opening this dataset we obtain details about the observations that form the clusters.

For our example we find:

Cluster 1 – Employee: 3, 5, 6, 8, 20, 22 and 24 (9 employees);

Cluster 2 – Employee: 2, 4, 7, 10, 12, 17, 18, 21, 23 and 25 (7 employees);

Cluster 3 – Employee: 1, 9, 11, 13, 14, 15, 16, 19 and 20 (10 employees).

TUTORIALS

Tutorial Sheet 1

Save the following data files from Blackboard – ***Concrete,*** **lumberdata and appraisal**. Each of the files contains data that requires analysis: use your notes to determine the appropriate methodology, apply the test, perform post –hoc tests (if necessary) and interpret the output. You must write a **short summary of your findings** for each question, including your reasoning for the test and its assumptions and the output achieved. This may be completed in a word document or a LaTex file. Attempt to also run the code produced for each test you perform in SAS. This will help you become familiar with the SAS environment and the SAS language

1. CONCRETE

Using the concrete data discussed in lecture this week, determine if a relationship exists between the measured strength of concrete by operator 1 and operator 2.

2. APRRAISAL

A property appraiser wants to model the value on the market (£'s) of a property in a rural area. He is of the opinion that this can be modeled efficiently by using three key variables, namely:

- **Land Value:** appraised land value (£'s);
- **Improvement Value:** appraised value of improvement made to the property (£'s);
- **Property Area:** area of the property (square feet).

Do you recommend that the property appraiser would be wise to use such a model?

3. LUMBER DATA

The weight of lumber is to be estimated from external measurements of a sample of pine trees. The variables used in the analysis were as follows:

- **Weight**: Weight of lumber from a tree (kg);
- **Height**: Height of the tree (in feet);
- **Age**: Age of the tree (in years).

Construct a linear model for **Weight** using the **Height** and **Age** variables. Is this model appropriate, is a transformation necessary?

4. Bacterial Growth

Fourteen samples of bacteria were grown under different laboratory conditions to determine if there was a relationship between the humidity in the laboratory, and the amount of light that the bacteria received. The variables to be used in the analysis are:

- **Density:** Density of the bacteria;
- **Humidity:** humidity in laboratory;
- **Light:** intensity of light.

The density of the bacteria was the response of interest, and it was aimed to model the data via multiple linear regression with a transformation if necessary. Generate the most appropriate model and interpret the output from the analysis.

Tutorial Sheet 2

Download the following two SAS Enterprise data files from the learning schedule on blackboard ***donner and binary..***

1. SURVIVAL

In 1846, the Donner party (Donner and Reed families) left Springfield, Illinois for California in covered wagons. After reaching Fort Bridger, Wyoming, the leaders decided to find a new route to Sacramento. They became stranded in the eastern Sierra Nevada mountains at a place now called Donner Pass (right) when the region was hit by heavy snows in late October. By the time the survivors were rescued on April 21, 1847, 40 out of 87 had died.

Three variables: Survivor (1 if person i survived, 0 otherwise), Age and Sex (1 if person i is male). The main objective of this task is to study the relationship between survival and gender.

Use the data set: donner.sas7bdat which is your Black Board. Then

1. Predict the probability of survival as a function of age.
2. After taking into account age, are women more likely to survive harsh conditions than men? Does age affect the survival rate of men and women differently?
3. Is the model good for classification? Use the ROC curve.

2. BINARY

Use binary.sas7bdat. Make a ROC analysis for the model in Example 5. Calculate the specificity and the sensitivity for the chosen cut-off. Do that for different cut-offs and compare the results.

Tutorial Sheet 3

Download the data file from blackboard that was used in lecture: ***Questionnaire.*** Attempt to recreate the analysis that was performed in lecture and write a **short summary of your findings.** This should include your reasoning for the analysis and the output achieved. This may be completed in a word document or a LaTex file.

Download the following three SAS Enterprise data files from the learning schedule on blackboard ***Leadership, Cereals and Psychological.***

Perform a principal component analysis on the data, determine the number of principal components to retain, name the components and compare the factor scores of the new components. Use your notes to aid in the interpretation of the output. You must write a **short summary of your findings** for each question. This may be completed in a word document or a LaTex file.

1. LEADERSHIP

A survey was conducted to assess the function of leaders in different sectors of the economy namely: manufacturing, agriculture and banking. Eight questions were asked relating to different aspects of being a team leader. You are required to compare the factor scores of the new components for each of the different sectors.

2. CEREALS

A survey was conducted on 203 subjects consisting of children, teenagers and adults to determine what they value the most in cereals. The subjects were asked to assign a percentage to the importance that they would give to thirteen different properties of cereals. Perform a principal component analysis on the data, determine the number of factors, name the new components and compare the scores obtained for the different groups on the new components.

3. PSYCHOLOGICAL

Psychological tests that consisted of nine different test were performed on four different occupations in the public sector, namely: civil servants, firemen, nurses and teachers. Determine if the structure can be simplified.

Tutorial Sheet 4

Download the following three data files from Blackboard: **carsurvey**, **snack_bar_survey, travelling** and **athletics**. Perform a factor analysis on the data, determine the number of principal factors to retain, interpret the loadings and communalities, name the new factors and look at plots of the newly generated factor scores. You must write a **short summary of your findings** for each question.

1. Car Survey

An automobile retailer wished to gain a greater understanding of the motivations that their potential purchasers had prior to buying. A random sample of 278 visiting customers were surveyed over a six months period, and were asked to rate on a scale of 0 – 100 sixteen different attributes related to their motives for purchasing, these were:

Exciting, Dependable, Luxurious, Outdoorsy, Powerful, Stylish, Comfortable, Rugged, Fun, Safe, Performance, Family, Versatile, Sports, Status, and finally **Practical.**

2. Snack Bar Survey

A company that was developing a new snack bar, developed a questionnaire to gain an understanding of what potential customers desired out of their product. The questionnaire was randomly given to customers attending a local supermarket, 332 were completed. The participants were asked to score between 0 – 100, 11 different characteristics associated with a snack bar including: **Taste, Value for money, Filling, Sweet, Suitable as snack, Provides energy, long lasting, fun** and **convenient.**

3.Travelling

A large chain of estate agents provided questionnaires to 43 customers to see if there were any relationships evident between people based on how they scored three different questions (0 – 100) concerning how much it influenced their choice of holiday. The questions used were: **Expense; Weather, and Facilities.**

4. ATHLETICS

A sporting institute conducted an analysis on the times and distances achieved by thirty-seven 18 year old male athletes that represented the institute in club meetings, to determine whether they fell into groups of similar performances. Four events were observed namely: **100m (seconds); 1500m (seconds), Javelin (m) and Shot Putt (m).**