Chapter Three

THEORETICAL FRAME WORK

3.1: Introduction

A critical review of statistical methods; namely binary logistic regression; suitable for modeling the probability of dying before age five leading to the determinants of such probability is presented in this chapter. In addition, factor analysis is used for defining under-five mortality indicators.

3.2: Logistic Regression

Logistic regression statistical method analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting logistic curve to data. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. Many distributions have been proposed for use in analysis of dichotomous outcome variable. (Cox snel ,1989) discuss some of these. Logistic regression calculates the probability of an event occurring over the an event not occurring, the impact of independent variables is usually explained in terms of odds (Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur). If the probability of an event occurring is *p,* the probability of the event not occurring is (*1-p*). Then the corresponding odds is a value given by

$$\text{Odds of Event} = \frac{p}{1-p} \quad (3.1)$$

Unfortunately, this is not a good model because extreme values of x will give values of α+βx that does not fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural logarithm (Peng, Lee & Ingersoll, 2002). With logistic regression we model the natural log odds as a linear function of the explanatory variable:

$$\text{logit}(y) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta x \quad (3.2)$$

Where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model.

Taking the antilog of equation (2) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as p=P (Y=interested outcome/X= χ, a specific value)

3.2.1: Fitting The Logistic Regression Model

The parameters are usually estimated using the method of maximum likelihood of the sample values (Menard, 2001). Maximum likelihood will provide values of $\beta'(\beta 0, \beta 1, \ldots . \beta p)$ which maximize the probability of obtaining the data set. It requires iterative computing with computer software.

A "likelihood" is a probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. The likelihood varies from 0 to 1 like any other probabilities. Practically, it is easier to work with the logarithm of the likelihood function. This function is known as the log-likelihood. Log-likelihood will be used for inference testing when comparing several models. The log likelihood

varies from 0 to -∞ (it is negative because the natural log of any number less than 1 is negative).In logistic regression, we observe binary outcome and predictors, and we wish to draw inferences about the probability of an event in the population.

$$L=\prod_{i=1}^{n} P(Y_i|X_{i1},\ldots,X_{ip})=\prod_{i=1}^{n}\left[\left(\frac{e^{\alpha+\sum_{j=1}^{p}\beta_j X_j}}{1+e^{\alpha+\sum_{j=1}^{p}\beta_j X_j}}\right)^{Y_i}\times\left(\frac{1}{1+e^{\alpha+\sum_{j=1}^{p}\beta_j X_j}}\right)^{1-Y_i}\right] \quad (3.3)$$

Note that Y is the 0/1 outcome for the i^{th} case and, $X_{i1},\ldots,X_{ip}$ are the values of the predictor variables for the$c$ case based on a sample of n cases.The use of Yi and 1-Yi as exponents in the equation above includes in the likelihood the appropriate probability term dependent upon whether Yi =1 or Yi =0 (note that F^0 =1 for any expression, F). Using the methods of calculus, a set of values for a and the β_j can be calculated that maximize Land these resulting values are known as maximum likelihood estimates (MLE's). This iterative solution procedure is performed with the aid of SPSS.

3.2.2: Assessing The Fit of The Model

We begin our discussion of methods for assessing the fit an estimated logistic regression model with the assumption that we are at least preliminarily satisfied with our efforts at the model building stage. By this we mean that, to the best of our knowledge, the model contains those variables (main effects as well as interactions) that should be in the model and that variables have been entered in the correct functional form.

Suppose we denote the observed sample values of the outcome variable in vector form as y where $\hat{y}^1 = (y_1, y_2, \ldots, \hat{y}_n)$. We conclude that the model fits if:

a. Summary measures of the distance between y and $\hat{y}$ are small.
b. The contribution of each pair $(y_i, \hat{y}_i); i = 1, 2, \ldots, n$ to these summary measures is unsystematic and is small relative to the error structure of the model.

When the model building stage has been completed, a series of logical steps may be used to assess the lit of the model. The components of the proposed approach are:

a. Computation and evaluation of overall measures of fit.
b. Examination of the individual components of the summary statistics, often graphically.
c. Examination of other measures of the difference or distance between the components of y and $\hat{y}$.

In many statistical inference procedures, we have used chi-square distribution based on the *likelihood ratio,* Score, or Wald test statistics. The global chi-square addresses the question "*Is this model better than nothing*?" The answer "yes" suggests the acceptance of the model. After we arrived the final model, we may want to ask the question "Is these a better model than this one?" Now the answer "yes" will lead to rejection of the model. Goodness-of-fit tests are methods to determine the suitability of the fitted model, and the goal of a logistic regression analysis is to find the best fitting model to describe the relationship between independent variables and an outcome.

1. R^2 Statistics for Logistic Regression

Logistic regression uses R^2 as measures of goodness-of-fit, there are many different ways to calculate R^2 for logistic regression, and (Menard, 2000) considered several others. The Cox-Snell R^2 and McFadden (both corrected and uncorrected) was actually discussed earlier by (Maddala, 1983) and by (Cragg et al,1970). Cox-Snell measures for binary logistic regression but McFadden's measure for multinomial. Also for goodness of fit logistic regression has likelihood ratio test and Nagelkerke R Square statistic was developed to ensure statistic was developed to ensure Cox-Snell R^2statisticto have values between 0 and 1. The likelihood ratio test for a particular parameter compares the likelihood of obtaining the data Logistic regression is estimated by maximizing the likelihood function. Let L_0 be the value of the likelihood function for a model with no predictors, and let LM be the likelihood for the model being estimated. McFadden's $R^2$is defined as

$$R^2 \text{ McFadden's} = 1 - \frac{\ln(LM)}{\ln(L0)} \quad (3.4)$$

The statistic was developed to ensure CS-R^2 statistic to have values between 0 and 1. R^2 is

$$R^2 \text{ C\&S} = 1 - \frac{(L0)}{(LM)} \quad (3.5)$$

The likelihood ratio test

$$-2 \times \ln(L_0/L_1) = -2 \times (\ln L_0 - \ln L_1) \quad (3.6)$$

$$\text{Nagelkerke R Square} = \frac{1-\left(\frac{1-(MIntercept)2/N}{L(Mfull)}\right)}{1-L(MIntercept)2/N} \quad (3.7)$$

2.The Hosmer-Lemeshow Goodness Of Fit Test

After a logistic regression model has been fitted, a global test of goodness of fit of the resulting model should be performed. A test that is commonly

used to assess model fit is the Hosmer–Lemeshow test, which are available in statistical software programs. The purpose of any overall goodness-of-fit test is to determine whether the fitted model adequately describes the observed outcome experience in the data (Hosmer and Lemeshow, 2000).

The validity of inferences drawn from modern statistical modeling techniques depends on the assumptions of the statistical model being satisfied. A critical step in assessing the appropriateness of the model is to examine its fit, or how well the model describes the observed data. Very few articles in subject-matter journals make any mention of having carried out this important step in model development. Much of what will be presented here is described in detail in texts by (Hosmer and Lemeshow, 1989;Kleinbaum et al, 1988)

The Hosmer-Lemeshow goodness of fit test statistic is given by:

$$\sum_{j=1}^{g} \frac{(Oj - Nj - \pi j)}{(Nj - \pi j - (1 - \pi j))} \qquad (3.8)$$

Where :

g=number of groups

Nj= number of observation in j^{th} group

Oj=number of response in j^{th} group

Πj=average of predicted probability for j^{th} group

3.2.3. Statistical tests and confidence interval

Inferences about individual regression coefficients, groups of regression coefficients, goodness-of-fit, mean responses, and predictions of group membership of new observations are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence

intervals. The inference procedures in logistic regression rely on large sample sizes for accuracy.

Two procedures are available for testing the significance of one or more independent variables in a logistic regression: likelihood ratio tests and Wald tests. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

3.2.4: Testing the significance of the coefficients

A logistic regression model allows us to establish a relationship between a binary outcome variable and a group of predictor variables. A regression coefficient describes the size and direction of the relationship between a predictor and the response variable. Use the coefficient to determine whether a change in a predictor variable makes the event more likely or less likely. The estimated coefficient for a predictor represents the change in the link function for each unit change in the predictor, while the other predictors in the model are held constant. The relationship between the coefficient and the probability depends on several aspects of the analysis, including the reference event for the response and the reference levels for categorical predictors. Generally, positive coefficients make the event more likely and negative coefficients make the event less likely. An estimated coefficient near 0 implies that the effect of the predictor is small.

1.The Standard Error of the Coefficient

The standard error of the coefficient estimates the variability between coefficient estimates that you would obtain if you took samples from the same population again and again. Use the standard error of the coefficient to

measure the precision of the estimate of the coefficient. The smaller the standard error, the more precise the estimate.

2. Wald Test

The Wald test is the test of significance for individual regression coefficients in logistic regression. For maximum likelihood estimates, the ratio

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \qquad (3.9)$$

Where SE($\hat{\beta}_i$) is an estimate of the standard error of $\hat{\beta}_i$ provided by the square root of the corresponding diagonal element of the covariance matrix, V($\hat{\beta}_i$).

3. Confidence interval for coefficient (95% CI)

These confidence intervals (CI) are ranges of values that are likely to contain the true value of the coefficient for each term in the model. The confidence interval is accurate if the sample size is large enough that the distribution of the sample coefficient follows a normal distribution. The confidence interval ($\hat{\beta}_i \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_i)$) will be used to assess the estimate of the population coefficient for each term in the model.

3.2.5: Odds Ratio

The odds ratio (OR) is a comparative measure of two odds relative to different events. For two events A and B, the corresponding odds of A occurring relative to B occurring is

$$\text{Odds ratio } \{\text{A vs B}\} = \frac{\text{odds } \{\text{A}\}}{\text{odds } \{\text{B}\}} \qquad (3.10)$$

An OR is a measure of association between an exposure and an outcome The OR represents the odds that an outcome (e.g. disease or disorder) will occur given a particular exposure (e.g. health behavior, medical history), compared to the odds of the outcome occurring in the absence of that exposure. When a logistic regression is calculated, the regression coefficient (b_1) is the estimated increase in the logged odds of the outcome per unit increase in the value of the independent variable. In other words, the exponential function of the regression coefficient (e^b) is the OR associated with a one unit increase in the independent variable. The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome. OR=1 indicates exposure does not affect odds of outcome. OR>1 indicates exposure associated with higher odds of outcome. OR<1 indicates exposure associated with lower odds of outcome.

3.3: Factor Analysis

Factor analysis is a multivariate statistical approach commonly used in psychology, education, and more recently in the health-related professions.

Factor analysis represents the variables y_1, y_2 ... y_p as linear combinations of a few random variables f_1, f_2. ... $f_m (m<p)$ called factors. The factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The factors are underlying constructs or latent variables that "generate" the y's.

3.3.1: Orthogonal Factor Model

Factor analysis is basically a one-sample procedure which assumes a random sample $y_1, y_2. . . y_n$ selected from a homogeneous population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sum$.The factor analysis model expresses each variable as a linear combination of underlying common factors $f_1, f_2, . . . , f_m$, with an accompanying error term to account for that part of the variable that is unique (not in common with the other variables). For $y_1, y_2. . . y_p$ in any observation vector y, the model is as follows:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \ldots + \lambda_{1m}f_m \\ y_1 - \mu_1 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \ldots + \lambda_{2m}f_m \\ y_p - \mu_1 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \ldots + \lambda_{2m}f_m \end{aligned} \qquad (3.12)$$

Ideally, m should be substantially smaller than p; *f*'s in (3.12) as random variables that engender the *y*'s. The coefficients λ_{ij} are called loadings and serve as weights. The emphasis in factor analysis is on modeling the covariances or a correlation among the y's and could result communality and specific variance,

$$\begin{aligned} \sigma_{ii} &= (y_i) \\ &= (\lambda^2_{i1} + \lambda^2_{i2} \ldots\ldots + \lambda^2_{im}) + \psi_i \\ &= h^2_i + \psi_i \\ &= \text{communality} + \text{specific variance} \end{aligned} \qquad (3.13)$$

The communality h^2_i is also referred to as common variance, and the specific variance ψ_i has been called specificity, unique variance, or residual variance.

3.3.2: Estimation of Loading and Communalities

Factor analysis is based on the 'common factor model' which is a theoretical model. This model postulates that observed measures are affected by underlying common factors and unique factors, and the

correlation patterns need to be determined. There is an array of extraction methods available, but we will briefly touch on a few commonly used techniques.

1. Principal Factor Method

The first technique that study considers is commonly called the principal component method, this method actually does not calculate any principal components, and it is the most widely used method for determining a set of loadings is the principal component method (Johnson and Wichern, 1992) (Rencher, 1995). This method seeks values of the loadings that bring the estimate of the total communality as close as possible to the total of the observed variances. A method of extracting factors from the original correlation matrix, with squared multiple correlation coefficients placed in the diagonal as initial estimates of the communalities. These factor loadings are used to estimate new communalities that replace the old communality estimates in the diagonal. Iterations continue until the changes in the communalities from one iteration to the next satisfy the convergence criterion for extraction.

$$\mathbf{S} \overset{\text{def}}{=} \hat{\Lambda}\hat{\Lambda}' + \hat{\psi} \qquad (3.14)$$

In the principal component approach, we neglect $\hat{\psi}$ and factor S into S = $\hat{\Lambda}\hat{\Lambda}'$. In order to factor S, the method uses the spectral decomposition.

$$S=CDC' \qquad (3.15)$$

Where:

C is an orthogonal matrix constructed with normalized eigenvectors (c_ i ci = 1) of S as columns and D is a diagonal matrix with the eigenvalues θ_1, θ_2, . . . , θ_p of S on the diagonal:

$$D = \begin{matrix} \theta_1 & 0 & .... & 0 \\ 0 & \theta_2 & .... & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & .... & \theta_p \end{matrix} \qquad (3.16)$$

Here θI for eigenvalues instead of the usual $\lambda_{i\text{in}}$ order to avoid confusion with the notation λ_{ij} used for the loadings.

To factoring $CDC^{'}$ in (3.21) to $\hat{\Lambda}\hat{\Lambda}'$ the observation that of the eigenvalues θi of the positive semi definite matrix **S** are all positive or zero, D can factor to $D = D^{\frac{1}{2}} D^{\frac{1}{2}}$ *where:*

$$D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\theta_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\theta_2} & \ldots & 0 \\ \ldots & & .... & \\ .... & & .... & \ldots \\ 0 & 0 & .... & \sqrt{\theta_p} \end{bmatrix}$$

With this factoring of D, (3.17)

$$S = CDC'$$

$$= C\, D^{1'2} D^{1'2} C'$$

$$= (C\, D^{1'2})(C\, D^{1'2})' \qquad (3.18)$$

This is of the form s = $\hat{\Lambda}\hat{\Lambda}'$ but $\hat{\Lambda}$ do not define to be C $D^{1'2}$ because C $D^{1'2}$ is $p \times p$ and $\hat{\Lambda}$ is $p \times m$ therefore the definition of Di as diag ($\theta 1, \theta 2 \ldots\ldots \theta m$

) with the m largest eigenvalues $\theta 1 > \theta 2 > \ldots > \theta m$ and C1 = (c1, c2. . . cm) containing the corresponding eigenvectors. We then estimate Λ by the first m columns of C $D^{1'2}$,

$$\hat{\Lambda} = C1\ D1^{1'2} = (\sqrt{\theta 1}\ c1, \sqrt{\theta 2} c2 \ldots\ldots, \sqrt{\theta m}\ cm) \qquad (3.19)$$

So

$$\begin{pmatrix} \lambda ij & \lambda 12 \\ \lambda 21 & \lambda 22 \\ \lambda 31 & \lambda 32 \\ \lambda 41 & \lambda 41 \\ \lambda 51 & \lambda 51 \end{pmatrix} = \begin{pmatrix} c11 & c12 \\ c21 & c22 \\ c31 & c32 \\ c41 & c42 \\ c51 & c52 \end{pmatrix} \begin{pmatrix} \sqrt{\theta 1} & 0 \\ 0 & \sqrt{\theta 2} \end{pmatrix} = \qquad (3.20)$$

$$\begin{pmatrix} c11\ \sqrt{\theta 2} & c12\sqrt{\theta 1} \\ c21\ \sqrt{\theta 2} & c22\sqrt{\theta 1} \\ c31\ \sqrt{\theta 2} & c32\sqrt{\theta 1} \\ c41\ \sqrt{\theta 2} & c41\sqrt{\theta 1} \\ c51\ \sqrt{\theta 2} & c51\sqrt{\theta 1} \end{pmatrix}$$

In (3.20) the source of the term principal component solution. The columns of $\hat{\Lambda}$ are proportional to the eigenvectors of S, so that the loadings on the j^{th} factor are proportional to coefficients in the j^{th} principal component. The factors are thus related to the first m principal components, and it would seem that interpretation would be the same as for principal components. But after rotation of the loadings, the interpretation of the factors is usually different. Hence to complete the approximation of S in (3.14) and it defines

$$\hat{\psi}=s_{ii}-\sum_{i=1}^{m}\hat{\lambda}_{ij}^2 \qquad (3.21)$$

And write

$$\mathbf{S} \cong \hat{\Lambda}\hat{\Lambda}' +\hat{\psi} \qquad (3.22)$$

sums of squares of the rows and columns of $\hat{\Lambda}$ are equal to communalities and where $\widehat{\psi}$ = diag $(\widehat{\psi 1},\widehat{\psi 2},\ldots.,\widehat{\psi p})$. Thus in (3.22) the variances on the diagonal of S are modeled exactly, but the off-diagonal covariance's are only approximate. In this method of estimation, the eigenvalues, respectively. This is easily shown. By (3.21) and the i^{th} communality is estimated by

$$\hat{h}i^2=\sum_{j=1}^{m}\lambda^2_{ij} \qquad (3.23)$$

This is sum of squire of i^{th} row of $\hat{\Lambda}$ .The sum of squares of the j^{th} column of $\hat{\Lambda}$ is the jth eigenvalue of S:

$$\sum_{i=1}^{p}\hat{\lambda}^2ij=\sum_{i=1}^{p}(\sqrt{\emptyset}jC_{ij})^2$$

$$=\emptyset j\sum_{i=1}^{p}c^2ij$$

$$=\emptyset_j \qquad (3.24)$$

Since the normalized eigenvectors (columns of C) have length 1. By (3.22) and (3.23), the variance of the i^{th} variable is partitioned into a part due to the factors and a part due uniquely to the variable:

$$S_i = \hat{h}_i^2 + \hat{\psi}_i$$

$$= \hat{\lambda}^2 + \hat{\lambda}_2^2 + ... + \hat{\lambda}_m^2 + \hat{\psi}_i \quad (3.25)$$

Thus the j^{th} factor contributes $\hat{\lambda}^2 ij$ to s_{ij} The contribution of the *jth* factor to the total sample variance, $tr(\mathbf{S}) = s11 + s12 + .. + spp$

$$\text{Variance due to jth} \sum_{i=1}^{p} \hat{\lambda}^2 ij = = \hat{\lambda}^2 + \hat{\lambda} 2^2 + \ldots + \hat{\lambda} m^2 \quad (3.26)$$

Which is sum of square of loadings in the j^{th} column of $\hat{\Lambda}$ this is equal to the j^{th} eigenvalue; θ_j .The proportion of total sample variance due to the j^{th} factor is, therefore,

$$\frac{\sum_{i=1}^{p} \hat{\lambda}}{\text{tr}(\mathbf{S})} = \frac{\theta j}{\text{tr}(\mathbf{S})} \quad (3.27)$$

If the variables are not commensurate, we can use standardized variables and work with the correlation matrix R. The eigenvalues and eigenvectors of R are then used in place of those of S in (3.19) to obtain estimates of the loadings. In practice, R is used more often than S and is the default in most software packages. Since the emphasis in factor analysis is on reproducing the covariances or correlations rather than the variances, use of R is more appropriate in factor analysis than in principal components. In applications, R often gives better results than S. If we are factoring R, the proportion corresponding to (3.27) is

$$\frac{\sum_{i=1}^{p} \hat{\lambda}}{\mathrm{tr}(\mathbf{S})} = \frac{\theta\, j}{\mathrm{p}} \qquad (3.28)$$

The assessing of the fitting the factor analysis model by comparing the left and right sides of (3.22) .The error matrix

$$\mathrm{E} = S - (\hat{\wedge}\,\hat{\wedge}' + \hat{\Psi}) \qquad (3.29)$$

Has zeros on the diagonal but nonzero off-diagonal elements. The following inequality gives a bound on the size of the elements in E:

$$ij \leq \theta^2\ \mathrm{m+1} + \theta^2\ \mathrm{m+2} + .. + \theta^2\ \mathrm{p} \qquad (3.30)$$

$$\sum_{ij} e^{\,2}$$

2.The Principal Axis Factoring

The Principal Axis Factoring method is based on the notion that all variables belong to the first group and when the factor is extracted, a residual matrix is calculated. Factors are then extracted successively until there is a large enough of variance accounted for in the correlation matrix (Tucker & McCollum, 1997). Principal Axis Factor is recommended when the data violate the assumption of multivariate normality (Costello & Osborne, 2005).

Principal axis factor analysis and principal components analysis are computationally similar (Stevens, 2002). In the case of principal components analysis (PCA), the linear combination of variables results in components that account for all of the variance in the original data. Principal axis factor analysis yields factors that account for the common variance in the original data (Cureton & D'Agostino, 1983; Harman, 1976; Stevens, 2002). Principal components analysis employs a correlation matrix

as the matrix of association. When conducting a principal axis factor analysis, researchers focus on a reduced correlation matrix as the matrix of association. This reduced correlation matrix contains communality estimates on the main diagonal as opposed to ones (Cureton &D'Agostino, 1983; Harman, 1976; Stevens, 2002). the principal axis factor method includes determining the coefficients for the second factor. These coefficients are selected to maximize the factor's contribution to the remaining, or residual, communality. In a fashion similar to finding the coefficients for the first factor, the second largest root is equivalent to the root, or eigenvalue, of the first factor's residual communality matrix. This root of this residual matrix is equivalent to the second largest eigenvalues of the original, reduced correlation matrix (Cureton&D'Agostino, 1983; Harman, 1976). This procedure proceeds until the entire matrix of factor coefficients is developed.

the principal axis method uses an initial estimate $\hat{\psi}$ and factors $S - \hat{\psi}$ or $R - \hat{\psi}$ to obtain

$$S - \hat{\psi} \cong \widehat{(\Lambda \hat{\Lambda}')} \quad (3.31)$$

$$R - \hat{\psi} \cong \widehat{(\Lambda \hat{\Lambda}')} \quad (3.32)$$

Where $\hat{\Lambda}$ is p×m and is calculated as in (3.19) using eigenvalues and eigenvectors of $S - \hat{\psi}$ or $R - \hat{\psi}$,The i^{th} diagonal element of $S - \hat{\psi}$ is given by $s_{ij} - \hat{\psi}$ which is the i^{th} communality $\hat{h}_i^2 = s_{ij} - \hat{\psi}$ .A popular initial estimate for communality in $R - \hat{\psi}$ is $\hat{h}^2 i = R^2 i$

$$\hat{h}^2 i = R^2 i = 1 - \frac{1}{r^{ii}} \quad (3.33)$$

$\hat{h}^2$i=sii -$\frac{1}{s^{ii}}$(3.34)

Where r^{ii} is the i^{th} diagonal element of R^{-1}. For S –$\hat{\psi}$ an initial estimate of communality analogous to (3.33) is

$\hat{h}i^2$= sii -$\frac{1}{s^{ii}}$= sii R²i (3.35)

Which is a reasonable estimate of the amount of variance that *yi*has in common with the other *y*'s. To use (3.33) or (3.34), **R** or **S** must be nonsingular. If **R** is singular, we can use the absolute value or the square of the largest correlation in the *i*th row of **R** as an estimate of communality. After obtaining communality estimates, we calculate eigenvalues and eigenvectors of S $-\hat{\psi}$ or R $-\hat{\psi}$ and use (3.19) to obtain estimates of factor loadings, $\hat{\Lambda}$.Then the columns and rows of $\hat{\Lambda}$ can be used to obtain new eigenvalues (variance explained) and communalities, respectively. The sum of squares of the jth column of $\hat{\Lambda}$ is the jth eigenvalue of S $-\hat{\psi}$ or R $-\hat{\psi}$, and the sum of squares of the ith row of ˆ_ is the communality of yi . The proportion of variance explained by the jth factor is

$$\frac{\theta\, j}{\mathrm{tr}(\mathrm{S}-\hat{\psi})}=\frac{\theta\, j}{\sum_{i=1}^{p}\theta\, i}$$

Or

$$\frac{\theta\, j}{\mathrm{tr}(\mathrm{R}-\hat{\psi})}=\frac{\theta\, j}{\sum_{i=1}^{p}\theta\, i}$$

where θ_j is the j^{th} eigenvalue of $S-\hat{\psi}$or $R-\hat{\psi}$ .The matrices $S-\hat{\psi}$ and $R-\hat{\psi}$ are not necessarily positive semidefinite and will often have some small negative eigenvalues. In such a case, the cumulative proportion of variance will exceed 1 and then decline to 1 as the negative eigenvalues are

added. [Note that loadings cannot be obtained by (3.19) for the negative eigenvalues.

3. Maximum Likelihood

Maximum Likelihood attempts to analyze the maximum likelihood of sampling the observed correlation matrix (Tabachnick&Fidell, 2007). Based on the assumption that a specified number offactors exists in a population, maximum likelihood factor analysis yields estimates of factor loadings for a given sample size and number of observed variables (Harman, 1976). When the observed variables exhibit multivariate normality and the sample size is large, maximum likelihood strategies facilitate the calculation of confidence intervals for the estimated loadings (Chen, 2003). Maximum likelihood strategies are dependent on the assumptions that, in addition to the observed variables, the common factors exhibit multivariate normality. Maximum likelihood techniques provide researchers with both parameter estimates and statistical indicators of model adequacy (Conway &Huffcutt, 2003; Harman, 1976; Mislevy, 1986).

Maximum Likelihood is more useful for confirmatory factor analysis and is used to estimate the factor loadings for a population. But this method for estimating factor models can yield distorted results when observed data are not multivariate normal (Costello & Osborne, 2005; Fabrigar et al., 1999).

Assume that the observations y1, y2. . . yn constitute a random sample from Np(μ, ∑),then Λ and ψ can can be estimated by the method of maximum likelihood. It can be shown that the estimates $\widehat{\Lambda}$ and $\widehat{\psi}$ satisfy the following

$$S\widehat{\psi}^{-1}\widehat{\Lambda} = \widehat{\Lambda}(1+\widehat{\Lambda}'\widehat{\psi}^{-1}\widehat{\Lambda}) \qquad (3.36)$$

$$\widehat{\psi} = \text{diag}(S-\widehat{\Lambda}\widehat{\Lambda}') \qquad (3.37)$$

$\hat{\Lambda}'\hat{\psi}^{-1}\hat{\Lambda}$ is diagonal (3.38)

3.3.3:Determining The Number of Factors

As mentioned previously, one of the main objectives of factor analysis is to reduce the number of parameters. The number of parameters in the original model is equal to the number of unique elements in the covariance matrix. A good place to start in understanding conventional methods used to determine the number of factors to retain is with an examination of the *k* eigenvalues for an item correlation matrix, where *k* is the number of items. Each eigenvalue represents the share of the total item variance that can be captured using one linear combination of the items, and the sum of the eigenvalues equals k. To the extent that a small number of eigenvalues are relatively large and most are relatively small, this pattern suggests that a small number of linear combinations can be used to capture much of the total item variance. These linear combinations correspond to the latent factors underlying the item responses. Thus, many methods for determining the number of factors to retain in FA are based on an examination of the pattern of eigenvalues.

(Cattell's ,1966) scree test is a graphical method in which the k eigenvalues are plotted in descending order, and a graph constructed in this way is called a scree plot. The scree test is performed by searching for an "elbow" in the plot, or an abrupt transition from large to small eigenvalues. However, there is not always a visual elbow on the scree plot, in which case the test requires a difficult subjective judgment of where a line should be drawn to determine the number of factors to retain. (Gorsuch, 1983) noted

that scree plots can be ambiguous due to the presence of more than one discontinuity in the graph or the lack of a visible.

The Kaiser criterion is the most commonly used stopping rule (Henson & Roberts, 2006). The rule sets the threshold between large and small values at an eigenvalue of 1, the arithmetic mean of the eigenvalues (recall that these sum to *k* for *k* items). Each eigenvalue greater than 1 is interpreted as representing a factor, and each value below 1 is not.

3.3.4: Rotation Methods

In the literature, definitions of rotation abound. For example, (McDonald,1985,) defines *rotation* as "performing arithmetic to obtain a new set of factor loadings (*v-f* regression weights) from a given set," and (Bryant and Yarnold ,1995) define it as "a procedure in which the eigenvectors (factors) are rotated in an attempt to achieve simple structure." Perhaps a bit more helpful is the definition supplied in (Vogt ,1993,): "Any of several methods in factor analysis by which the researcher attempts to relate the calculated factors to theoretical entities. This is done differently depending upon whether the factors are believed to be correlated (oblique) or uncorrelated (orthogonal)." And even more helpful is (Yaremko, Harari, Harrison, and Lynn, 1986), who define factor rotation as follows: "In factor or principal-components analysis, rotation of the factor axes (dimensions) identified in the initial extraction of factors, in order to obtain simple and interpretable factors." They then go on to explain and list some of the types of orthogonal and oblique procedures.

The goal of rotation is to attain an optimal simple structure which attempts to have each variable load on as few factors as possible, but maximizes the number of high loadings on each variable (Rummel, 1970). Ultimately, the

simple structure attempts to have each factor define a distinct cluster of interrelated variables so that interpretation is easier (Cattell, 1973). For example, variables that relate to language should load highly on language ability factors but should have close to zero loadings on mathematical ability.

As mentioned earlier, rotation methods are either orthogonal or oblique. Simply put.

1. Orthogonal rotation methods

Orthogonal rotation methods assume that the factors in the analysis are uncorrelated. (Gorsuch, 1983) lists four different orthogonal methods: equamax, orthomax, quartimax, and varimax. In contrast, Two common orthogonal techniques are Quartimax and Varimax rotation.

Quartimax involves the minimization of the number of factors needed to explain each variable (Gorsuch, 1983).

Varimax is the graphical approach to rotation is generally limited to $m = 2$. For $m > 2$, various analytical methods have been proposed for minimizing the number of variables that have high loadings on each factor and works to make small loadings even smaller.

2.Oblique rotation:

Oblique rotation methods assume that the factors are correlated (Gorsuch, 1983) lists 15 different oblique methods.

Oblique rotation is when the factors are not rotated 90° from each other, and the factors are considered to be correlated. Oblique rotation is more complex than orthogonal rotation, since it can involve one of two coordinate systems: a system of primary axes or a system of reference axes

(Rummel, 1970). Additionally, oblique rotation produces a pattern matrix that contains the factor or item loadings and factor correlation matrix that includes the correlations between the factors. The common oblique rotation techniques are Direct Oblimin and Promax.

Direct Oblimin attempts to simplify the structure and the mathematics of the output, while Promax is expedient because of its speed in larger datasets. Promax involves raising the loadings to a power of four which ultimately results in greater correlations among the factors and achieves a simple structure (Gorsuch, 1983). When interpreting the factors, you need to look at the loadings to determine the strength of the relationships.

Factors can be identified by the largest loadings, but it is also important to examine the zero and low loadings in order to confirm the identification of the factors (Gorsuch, 1983). A cross loading is when an item loads at .32 or higher on two or more factors (Costello & Osborne, 2005). Depending on the design of the study, a complex variable (i.e., an item that is in the situation of cross loading) can be retained with the assumption that it is the latent nature of the variable, or the complex variable can be dropped when the interpretation is difficult. Another option is to choose a significant loading cut-off to make interpretation easier. The signs of the loadings show the direction of the correlation and do not affect the interpretation of the magnitude of the factor loading or the number of factors to retain (Kline, 1994). A general rule to determine the reliability of the factor is to look at the relationship between the individual rotated factor loading and the magnitude of the absolute sample size. That is, the larger the sample size, smaller loadings are allowed for a factor to be considered significant (Stevens, 2002).

3.3.5 Interpretation of Factor Analysis

The main goal is to achieve a simple structure in which each variable loads highly on only one factor, with small loadings on all other factors. In practice, we often fail to achieve this goal, but rotation usually produces loadings that are closer to the desired simple structure.

The guidelines for interpreting the factors by examination of the matrix of rotated factor loadings. Moving horizontally from left to right across the m loadings in each row, identify the highest loading (in absolute value). If the highest loading is of a significant size. circle or underline it. This is done for each of the *p* variables. On the other hand, there may be variables with such small communalities that no significant loading appears on any factor. In this case, the researcher may wish to increase the number of factors and run the program again so that these variables might associate with a new factor. To assess significance of factor loadings $\hat{\lambda}ij$. A threshold value of .3 has been advocated by many writers. For most successful applications, however, a critical value of .3 is too low and will result in variables of complexity greater than 1. A target value of .5 or .6 is typically more useful. The .3 criterion is loosely based on the critical value for significance of an ordinary correlation coefficient.

After identifying potentially significant loadings, the experimenter then attempts to discover some meaning in the factors and, ideally, to label or name them. This can readily be done if the group of variables associated with each factor makes sense to the researcher. But in many situations, the groupings are not so logical, and a revision can be tried, such as adjusting the size of loading deemed to be important, changing m, using a different method of estimating the loadings, or employing another type of rotation.

3.3.6: Factor Scores

In many applications, the researcher wishes only to ascertain whether a factor analysis model fits the data and to identify the factors. There are two potential uses for such scores:

a- The behavior of the observations in terms of the factors may be of interest.

b-The use the factor scores as input to another analysis, such as MANOVA.