Notes For MScFE610 M1 - Written By Khalid Yaqub

Simple Linear Regression vs Multiple Linear Regression

Here's a table comparing Simple Linear Regression and Multiple Linear Regression:

	Simple Linear Regression	
Number of variables	2 (1 dependent variable, 1 independent variable)	3 or more
Equation	$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1$
Purpose	Predicting the dependent variable based on a single independent variable	Predicting
Relationship	Linear relationship	Linear rela
Coefficients	β_0, β_1	$\beta_0, \beta_1, \beta_2, ..$
Method	Ordinary Least Squares (OLS)	Ordinary L
Assumptions	Linearity, Independence, Homoscedasticity, Normality, No multicollinearity	Linearity, I
Interpretation	The change in the independent variable is associated with a change in the dependent variable	The chang
Applications	Simple relationships between two variables	Complex r

Note that this table provides a general comparison between the two regression techniques. It's important to understand that the choice between simple linear regression and multiple linear regression depends on the specific research question, the availability of data, and the nature of the relationship between the variables being studied.

Ordinary Least Squares (OLS)

Imagine you want to understand how one thing (let's say height) affects another thing (let's say weight). You collect data from different people, including their heights and weights. You want to find a line that best predicts weight based on height.

OLS is a method that helps us find that line. It does this by minimizing the errors between the actual weights and the predicted weights based on the line. The line is determined by two things: the starting point (called the intercept) and how much the line goes up or down for each increase in height (called the slope).

Here are the main steps of OLS:

1. Collect data: Get information about heights and weights from different people.
2. Build the line: Determine the line that best fits the data. The line starts at a certain point (intercept) and goes up or down at a certain rate (slope).
3. Minimize errors: Adjust the intercept and slope of the line to minimize the differences between the predicted weights and the actual weights from the data.

4. Evaluate the line: Check how well the line fits the data by using measures like R-squared. A higher R-squared means the line is a better fit.
5. Use the line: Once we have the line, we can use it to make predictions about weight based on height. We can also draw conclusions about how height and weight are related in the population.

OLS is a popular method because it's simple and helps us understand the relationship between variables. However, we need to make sure our data meets certain assumptions for OLS to work properly.

That's the basic idea of Ordinary Least Squares (OLS)! It helps us find the best line that predicts one thing based on another, like weight based on height.

Sure! Let's go through a simple example of Ordinary Least Squares (OLS) using a hypothetical dataset.

Suppose we are interested in understanding how the number of hours studied (independent variable) affects a student's test score (dependent variable). We collected data from 10 students and recorded their hours studied and corresponding test scores.

Here is the dataset:

Hours Studied	Test Score
2	65
3	70
4	75
5	80
6	85
7	90
8	95
9	100
10	105
11	110

To apply OLS, we want to find the line that best predicts the test score based on the hours studied.

Step 1: Data Collection: We have collected the hours studied and test scores for 10 students.

Step 2: Model Specification: We assume a linear relationship between the hours studied and the test score. So, our model is: Test Score = β_0 + β_1 * Hours Studied, where β_0 is the intercept and β_1 is the slope.

Step 3: Parameter Estimation: We estimate the parameters β_0 and β_1 that minimize the sum of squared differences between the observed test scores and the predicted test scores.

To do this, we can use statistical software or equations. In this example, we'll use equations.

We calculate the following:

$$\bar{x} = \frac{\sum \{\text{Hours Studied}\}}{n} = \frac{2 + 3 + 4 + \cdots + 11}{10} = 6.5$$

$$\bar{y} = \frac{\sum \{\text{Test Score}\}}{n} = \frac{65 + 70 + 75 + \cdots + 110}{10} = 87.5$$

$$SS_{xy} = \sum(\{\text{Hours Studied}\} - \bar{x})(\{\text{Test Score}\} - \bar{y})$$

$$SS_{xx} = \sum(\{\text{Hours Studied}\} - \bar{x})^2$$

Using these calculations, we find:

$$SS_{xy} = 825$$

$$SS_{xx} = 55$$

Then, we can calculate the slope (β_1) as:

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{825}{55} = 15$$

To find the intercept (β_0), we use the formula:

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x} = 87.5 - 15 \cdot 6.5 = -17.5$$

So, our estimated linear regression model is:

Test Score = -17.5 + 15 * Hours Studied

Step 4: Model Evaluation: We can evaluate the model's fit using measures like R-squared or by visually inspecting the scatterplot and the line of best fit.

Step 5: Use the Model: With the estimated model, we can now make predictions for test scores based on the number of hours studied.

For example, if a student studies for 8 hours, we can predict their test score using the equation:

Test Score = -17.5 + 15 * Hours Studied

If a student studies for 8 hours, we can substitute the value into the equation:

Test Score = -17.5 + 15 * 8

Calculating this, we find:

Test Score = -17.5 + 120 = 102.5

Therefore, based on our estimated model, we predict that a student who studies for 8 hours is likely to score 102.5 on the test.

```
import numpy as np
import statsmodels.api as sm

# Define the independent variable (hours studied)
hours_studied = np.array([2, 3, 4, 5, 6, 7, 8, 9, 10, 11])

# Define the dependent variable (test scores)
```

```
test_scores = np.array([65, 70, 75, 80, 85, 90, 95, 100, 105, 110])

# Add a constant column to the independent variable
X = sm.add_constant(hours_studied)

# Fit the OLS model
model = sm.OLS(test_scores, X)
results = model.fit()

# Print the model summary
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 1.486e+30
Date:                Sun, 21 May 2023   Prob (F-statistic):          2.30e-118
Time:                        23:17:52   Log-Likelihood:                 296.13
No. Observations:                  10   AIC:                            -588.3
Df Residuals:                       8   BIC:                            -587.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         55.0000   2.92e-14   1.89e+15      0.000      55.000      55.000
x1             5.0000    4.1e-15   1.22e+15      0.000       5.000       5.000
==============================================================================
Omnibus:                        0.571   Durbin-Watson:                   0.091
Prob(Omnibus):                  0.752   Jarque-Bera (JB):                0.521
Skew:                           0.077   Prob(JB):                        0.771
Kurtosis:                       1.893   Cond. No.                         17.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correct
/usr/local/lib/python3.10/dist-packages/scipy/stats/_stats_py.py:1736: UserWar
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

```
# Predict the test score for 13 hours of study
new_hours = np.array([1, 13])  # Include the constant term (1) for prediction
new_X = sm.add_constant(new_hours)

predicted_scores = results.predict(new_X)

# Print the predicted test score
print("Predicted test score for 13 hours of study:", predicted_scores[1])
```

```
Predicted test score for 13 hours of study: 120.00000000000007
```

Covariance and Correlation

Here's a detailed comparison of Covariance and Correlation in a table format:

Aspect	
Definition	Covariance measures the extent and direction of the linear relationship between two variables.
Range of Values	Can take any real value, positive or negative.
Scale Dependency	Dependent on the scale of the variables being measured. The magnitude of covariance depends on the
Interpretation	Covariance values alone are difficult to interpret due to the scale dependency. A positive value indicates
Unit of Measurement	The unit of covariance is the product of the units of the variables being measured (e.g., square units, cu
Sensitivity to Outliers	Covariance is sensitive to the presence of outliers, as it is influenced by extreme values.
Mathematical Formula	Cov(X, Y) = Σ((X - μX) * (Y - μY)) / (n - 1), where X and Y are the variables, μX and μY are their respective
Relationship to Slope	The sign of the covariance indicates the direction of the relationship between variables. The magnitude

Principal Components Analysis

Principal Components Analysis is a statistical method used to reduce the dimensionality of a dataset while preserving most of the information. It transforms a set of possibly correlated variables into a new set of uncorrelated variables called principal components. These principal components are linear combinations of the original variables.

The goal of PCA is to find a lower-dimensional representation of the data that captures the maximum amount of variation in the dataset. This is achieved by identifying the directions in the data along which the variance is maximized. These directions are called principal axes or eigenvectors, and the associated variances are called eigenvalues.

The steps involved in PCA are as follows:

1. Standardize the data: PCA works best when the variables are on the same scale. Therefore, it is common practice to standardize the variables to have zero mean and unit variance.
2. Compute the covariance matrix: Calculate the covariance matrix of the standardized variables. The covariance matrix describes the relationships and variances between variables.
3. Compute the eigenvectors and eigenvalues: Determine the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors represent the principal axes, and the eigenvalues represent the amount of variance explained by each principal component.
4. Select the principal components: Sort the eigenvectors based on their corresponding eigenvalues in descending order. The eigenvectors with the highest eigenvalues capture the most variance in the data and are selected as the principal components.

5. Project the data: Project the original data onto the selected principal components to obtain the lower-dimensional representation of the data.

PCA has various applications, including dimensionality reduction, data visualization, noise filtering, and feature extraction. It is commonly used in fields such as data science, machine learning, image processing, and signal processing to analyze and understand high-dimensional datasets.

Please note that the specific content covered in the mentioned course may include additional details or variations of PCA. For a comprehensive understanding, it would be best to refer to the original course material or resources provided in that course.

Box-Cox Transformation

The Box-Cox transformation is a mathematical technique used to transform non-normal data into a more normally distributed form. It is named after statisticians George Box and Sir David Cox. The transformation is defined by a power parameter, denoted as λ (lambda), which determines the type of transformation applied.

The general form of the Box-Cox transformation can be expressed using LaTeX notation as follows:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(y), & \text{if } \lambda = 0 \end{cases}$$

Here, y is the original data, and

$$y^{(\lambda)}$$

represents the transformed data.

In the equation, when λ = 0, the natural logarithm (ln) of y is taken. This is known as a log transformation. Log transformations are useful for dealing with data that exhibits multiplicative relationships or exponential growth.

For λ ≠ 0, the transformation formula involves subtracting 1 from y raised to the power of λ and then dividing by λ. This form allows for a range of transformations, including square root (λ = 0.5), reciprocal (λ = -1), and other power transformations.

The choice of λ depends on the data and the desired transformation. It is often determined by maximizing the log-likelihood function or optimizing another objective criterion. The goal is to find the λ value that maximizes the normality of the transformed data.

By applying the Box-Cox transformation, skewed or non-normal data can be made more suitable for statistical analyses that assume normality, such as regression models. It helps to stabilize variance, improve the validity of statistical tests, and meet the assumptions of the analysis.

Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) is a statistical technique used to identify underlying factors or dimensions in a set of observed variables. It is a data reduction and dimensionality reduction method that helps uncover the latent structure or patterns in the data.

EFA is commonly used in fields such as psychology, social sciences, market research, and data analysis. The primary goal of EFA is to determine how many factors are needed to explain the relationships among a set of observed variables and to understand the nature of these factors.

Here's a simplified explanation of the steps involved in EFA:

1. **Data Collection:** Researchers collect data on a set of variables believed to be related or indicative of an underlying construct or concept.
2. **Factor Extraction:** EFA begins by extracting factors from the observed variables. The goal is to find a smaller number of unobserved factors that explain the common variance among the variables.
3. **Factor Rotation:** After factor extraction, researchers can perform rotation to simplify and enhance the interpretability of the factors. Rotation transforms the factors to make them easier to understand and interpret. Orthogonal rotation methods (e.g., Varimax) create factors that are unrelated or uncorrelated, while oblique rotation methods (e.g., Promax) allow for correlation between factors.
4. **Factor Interpretation:** Researchers examine the factor loadings, which indicate the strength and direction of the relationship between each variable and the factors. They interpret the factors based on the variables with high loadings, identifying the underlying themes or dimensions represented by each factor.
5. **Assessing Model Fit:** Various statistical measures, such as eigenvalues, scree plots, and fit indices, are used to evaluate the adequacy of the factor solution and determine the optimal number of factors to retain.

The output of EFA provides insights into the underlying structure of the data, revealing the relationships between observed variables and the latent factors. These factors can then be further analyzed and interpreted to gain a better understanding of the underlying constructs or concepts being measured.

Linear Regression, Exploratory Factor Analysis (EFA), Principal Component Analysis (PCA), and Correlation Matrix

Here's a comparison table summarizing the key aspects of Linear Regression, Exploratory Factor Analysis (EFA), Principal Component Analysis (PCA), and Correlation Matrix:

Technique	Purpose	Type of Analysis	Output
Linear Regression	Predicting a dependent variable	Predictive modeling	Regression coefficients
Exploratory Factor Analysis	Identifying underlying factors	Dimensionality reduction	Factor loadings
Principal Component Analysis	Dimensionality reduction	Dimensionality reduction	Principal components
Correlation Matrix	Examining relationships between vars	Descriptive analysis	Correlation coefficients

Imagine you have a bunch of different toys, like cars, dolls, and blocks. You want to find out if there are any groups or categories among these toys. For example, you want to see if there is a group of toys that are all cars or if there's a group of toys that are all dolls.

To do this, we can use a special tool called **Exploratory Factor Analysis**. This tool helps us find the different groups or categories among the toys. It looks at how the toys are similar to each other and tries to group them together based on their similarities.

On the other hand, tools like **Linear Regression** and **Correlation Matrix** are used for different things. Linear Regression helps us predict something, like how fast a car can go based on its size or color. Correlation Matrix helps us see how toys are related to each other, like if playing with blocks makes you more likely to also play with cars.

Lastly, **Principal Component Analysis (PCA)** is another tool that is similar to Exploratory Factor Analysis. It also helps us group the toys based on their similarities. But for our specific goal of finding the different categories or groups among the toys, Exploratory Factor Analysis is the best tool to use.

So, to summarize, when we want to find out if there are different groups or categories among toys, we use Exploratory Factor Analysis. It helps us see which toys are similar to each other and belong to the same group.

Assumptions when running a Regression Model

It's worth noting that violating these assumptions may lead to biased or inefficient estimates, invalid hypothesis tests, or unreliable predictions. Therefore, it is important to assess and address these assumptions when running a regression model.

When running a regression model, there are several key assumptions that need to be considered. These assumptions help ensure the validity and reliability of the regression analysis. Here are the common assumptions:

1. **Linearity**: The relationship between the dependent variable and the independent variables is assumed to be linear. This means that the effect of the independent variables on the dependent variable is additive and constant. This rule says that when things change, they

do so in a straight line. It means that if one thing goes up or down, the other thing also changes by the same amount.

2. **Independence**: The observations in the dataset should be independent of each other. This assumption implies that the value of one observation does not depend on or influence the value of another observation. This rule means that we look at things one at a time and don't let one thing influence the other. We want to focus on each thing separately.
3. **Homoscedasticity**: Homoscedasticity assumes that the variability of the errors (residuals) is constant across all levels of the independent variables. In simpler terms, it means that the spread of the residuals should be consistent throughout the range of the dependent variable. This rule tells us that the amount of change or difference between things should be about the same no matter what. We want things to be fair and equal.
4. **No multicollinearity**: This assumption states that there should be no high correlation between the independent variables in the regression model. High multicollinearity can lead to instability in the estimates and make it difficult to discern the individual effects of each independent variable. This rule says that we don't want things to be too similar to each other. It's like having too many toys that are all the same. We want to have different toys to play with.
5. **Normality**: The residuals (the differences between the observed and predicted values) are assumed to be normally distributed. This assumption is important for hypothesis testing, confidence intervals, and obtaining reliable p-values. This rule means that things should follow a nice, smooth pattern. We want things to be like a rainbow or a pretty shape, not all jumbled or messy.
6. **No endogeneity**: Endogeneity refers to the presence of a relationship between the error term and one or more independent variables. It can arise when there are omitted variables or measurement errors, which can bias the estimated coefficients. This rule is about not having any hidden secrets. We want to make sure that everything we look at is what it seems and there are no surprises hiding behind the scenes.
7. **No autocorrelation**: Autocorrelation, or serial correlation, assumes that there is no correlation between the residuals at different points in time or across observations. In other words, the errors should be independent and not exhibit a systematic pattern. This rule means that things should not depend on what happened before. We want each thing to be its own special thing and not influenced by what happened earlier.

It's worth noting that violating these assumptions may lead to biased or inefficient estimates, invalid hypothesis tests, or unreliable predictions. Therefore, it is important to assess and address these assumptions when running a regression model.

p-value

The p-value is a number that tells us how likely it is to get our observed results by chance alone, assuming that there is no real effect. It helps us decide whether our results are statistically significant or not. If the p-value is small (usually less than 0.05), it means that our results are unlikely to have happened just by chance. So, we reject the idea that there is no real effect and conclude that there is likely a meaningful relationship or difference. If the p-value is large (greater than 0.05), it means that our results could easily happen by chance, even if there is no real effect. In this case, we don't have enough evidence to say there is a significant relationship or difference.

Remember, the p-value doesn't tell us how big or important the effect is, it just helps us determine if our results are likely due to a real effect or simply random variation.

The p-value is a statistical measure that is commonly used in hypothesis testing to assess the strength of evidence against a null hypothesis. It quantifies the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the data, assuming that the null hypothesis is true.

In hypothesis testing, the null hypothesis (H0) represents the default or initial assumption, often stating that there is no significant effect or relationship between variables. The alternative hypothesis (H1 or Ha) contradicts the null hypothesis, suggesting that there is a significant effect or relationship.

The p-value helps determine whether the observed data provide enough evidence to reject the null hypothesis in favor of the alternative hypothesis. The general guideline is as follows:

- If the p-value is less than a pre-determined significance level (often denoted as α, commonly set at 0.05 or 0.01), it is considered statistically significant. In this case, the evidence suggests that the observed data are unlikely to have occurred by chance alone, leading to the rejection of the null hypothesis in favor of the alternative hypothesis.
- If the p-value is greater than the significance level, it is not considered statistically significant. This implies that the observed data are reasonably likely to occur even if the null hypothesis is true, and there is not enough evidence to reject the null hypothesis.

It is important to note that the p-value does not measure the magnitude of an effect or the importance of a finding. It only assesses the strength of evidence against the null hypothesis. Additionally, the p-value is influenced by factors such as the sample size, the magnitude of the effect being tested, and the variability in the data.

It is crucial to interpret p-values correctly and in conjunction with other measures and considerations, such as effect size, confidence intervals, and domain-specific knowledge. The p-value should not be the sole determining factor in drawing conclusions but rather one piece of evidence to support decision-making in statistical hypothesis testing.

Null Hypothesis

In statistical hypothesis testing, the null hypothesis (denoted as H0) is a statement that assumes there is no significant relationship, effect, or difference between variables or groups being compared. It represents the default position or initial assumption in a statistical analysis.

The null hypothesis is typically a statement of no effect or no difference. For example:

- In a study comparing the effectiveness of two treatments, the null hypothesis would state that there is no difference in effectiveness between the two treatments.
- In a study examining the relationship between two variables, the null hypothesis would state that there is no relationship or correlation between the variables.

The purpose of the null hypothesis is to be tested against an alternative hypothesis (H1 or Ha), which represents the opposite of the null hypothesis and suggests that there is a significant relationship, effect, or difference.

Through statistical analysis, researchers gather data and evaluate whether the evidence from the data supports rejecting the null hypothesis in favor of the alternative hypothesis. The goal is to determine if the observed results are statistically significant, indicating that there is enough evidence to conclude that the null hypothesis is unlikely to be true.

It is important to note that failing to reject the null hypothesis does not prove that the null hypothesis is true; it simply means that there is insufficient evidence to conclude otherwise. The null hypothesis is assumed to be true unless there is enough evidence to suggest otherwise based on the data and statistical analysis.

Nonparametric Regression and Measurement Error

The key difference between **nonparametric regression** and **measurement error** is their focus and purpose. Nonparametric regression aims to estimate relationships between variables without assuming a specific functional form, while measurement error focuses on correcting for inaccuracies in the observed values of variables caused by measurement errors.

Here's a comparison table between Nonparametric Regression and Measurement Error:

Aspect	Nonparametric Regression	Measurement Error
Approach	Uses flexible models to estimate relationships	Accounts for inaccuracies in measured varia
Assumptions	Does not assume a specific functional form	Assumes measurement errors are present in
Model Complexity	Can handle complex and nonlinear relationships	Primarily focuses on measurement error corr
Data Requirements	Sufficient data points for accurate estimation	Information on measurement error character
Interpretability	May be less straightforward due to flexible modeling	Focuses on adjusting observed values for me
Bias and Efficiency	Can trade off bias and efficiency for better fit	Corrects for bias introduced by measuremen
Applications	Suitable for exploring complex relationships	Important when measurement accuracy is a
Computational Complexity	May require more computational resources	May involve additional calculations for error

Please note that this table provides a general overview and the characteristics may vary depending on specific methods and techniques used within each approach.

Factor Analysis

In statistics, we usually work with observable variables. However, there are situations where we're interested in variables that can't be directly observed. For example, when studying the past environmental conditions in a remote region, we may not have direct measurements. Instead, we rely on proxy data like tree rings, pollen records, or lake sediments. These proxy data give us indirect information about the past environment.

Variables that can't be directly observed but are inferred from observed variables are called latent variables. We use mathematical models called latent variable models to connect these latent variables with the observed variables.

Factor analysis is a type of latent variable model. It helps us understand the relationships between observed variables by identifying underlying, unobserved variables called factors. Factors represent common patterns or themes that explain the correlations between the observed variables.

There are two main types of factor analysis:

1. Exploratory factor analysis (EFA): In EFA, we explore which factors are related to which observed variables. We don't have preconceived notions about how the factors and observed variables are connected.
2. Confirmatory factor analysis (CFA): In CFA, we have a predefined hypothesis about the relationships between specific observed variables and factors. We aim to confirm if a set of observed variables is indeed related to a particular factor.

In this tutorial, we focus on exploratory factor analysis (EFA), which helps us uncover the underlying factors that explain the correlations between observed variables.

The Exploratory Factor Model (EFM)

The basic idea behind factor analysis is to express each observed variable as a linear combination of the latent factors. This can be simplified as:

$X = \mu + \Lambda F + e,$

where X is the matrix of observed variables, μ is the vector of means, Λ is the matrix of factor loadings, F is the vector of latent factors, and e is the vector of errors.

In summary, factor analysis aims to model the relationships between observed variables and latent factors by expressing observed variables as a combination of latent factors, along with some error term.

Explained and Unexplained Variability

Explained variability refers to the proportion of the total variability in the observed variables that can be accounted for by the extracted factors. It represents the portion of the variation in the data that is explained by the underlying factors. On the other hand, unexplained variability, also known as error variance or unique variance, refers to the portion of variability that cannot be explained by the factors and is specific to each observed variable. It represents the random or unique component of each variable that is not accounted for by the shared factors.

Parameter Estimation

Parameter estimation in factor analysis involves estimating the factor loadings, which quantify the relationship between the observed variables and the underlying factors. The factor loadings represent the degree of influence of each factor on each observed variable. There are several estimation methods used in factor analysis, including maximum likelihood estimation (MLE), principal axis factoring (PAF), and others. These methods aim to find the best-fitting values for the factor loadings that minimize the discrepancy between the observed data and the estimated factor model.

Factor Rotation

Factor rotation is a technique used to simplify and interpret the factor structure obtained from factor analysis. It aims to achieve a more interpretable solution by applying an orthogonal or oblique rotation to the factors. Orthogonal rotation methods, such as Varimax or Quartimax, aim to create uncorrelated factors, which can facilitate a clearer interpretation of the underlying structure. On the other hand, oblique rotation methods, such as Promax, allow for correlated factors, which can better reflect the real-world relationships among the latent factors.

It is important to realize that the linear factor model is not identifiable. This means that there are two or more parametrizations which are observationally equivalent, or in other words, there exist an infinite number of different matrices Λ which may generate the same x values.

That is why factor analysis usually proceeds in two stages. In the first, one set of loadings, Λ , is calculated which yields theoretical variances and covariances that fit the observed ones as closely as possible. These loadings, however, may not provide a reasonable interpretation. Thus, in the second stage, the loadings, Λ , are transformed in an effort to arrive at another set that fit equally well the observed variances and covariances, but are easier to interpret.

The process of transforming a factor pattern is generally referred to as rotation. There are two basic types of transformations: orthogonal (uncorrelated factors) and oblique (correlated factors). Using orthogonal rotation, we preserve the independence of the factors. With oblique rotation factors are allowed to correlate. Two poplar methods are the varimax method for

orthogonal factor transformation, and the promax method for oblique factors rotation. Both methods are implemented in R.

The varimax method maximizes the variance of the squared loadings for each factor, thus making some of these loadings as large as possible, and the rest as small as possible in absolute value. Consequently, the variables become divisible into groups such that the loadings within each group are high on a single factor, moderate to low on a few factors and negligible on the remaining factors. The varimax method encourages the detection of factors each of which is related to few variables. It discourages the detection of factors influencing all variables (Tryfos 1997).

Confusion with Principal Component Analysis

There is sometimes confusion between principal component analysis (PCA) and factor analysis (FA). Both methods have the aim of reducing the dimensionality of a vector of random variables. However, the most fundamental difference is that factor analysis explicitly specifies a model relating the observed variables to a smaller set of underlying unobservable factors. This assumed model may fit the data or not. In contrast PCA is just a data transformation method. Furthermore while Factor Analysis aims at explaining (covariances) or correlations, PCA concentrates on variances.

A simple example of factor analysis in R

Please Visit https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/A-simple-example-of-FA/index.html