EXPERIMENT 7

CO4: Analyzing and transforming data for predictive modeling through various data transformation techniques.

1. Aim: To perform Confirmatory Factor Analysis (CFA) on a set of observed variables to confirm the hypothesized relationship between these variables and their underlying latent factors using Origin Pro 2024 software.

2. Objective:

- To test a predefined factor model based on theory or prior research.
- To evaluate the fit between observed data and the hypothesized model of relationships between latent factors and observed variables.
- To determine the factor loadings, error variances, and model fit indices, which help in validating the factor structure of the dataset.

3. Theory:

Discriminant Analysis is a statistical technique used to classify a set of observations into predefined groups based on several continuous or binary predictor variables. It creates a discriminant function, which is a linear combination of the predictor variables that best separates the groups.

3.1 Factor Analysis Overview

Factor Analysis is a multivariate statistical technique used to reduce data dimensions by grouping correlated variables into underlying factors. It comes in two main types:

- **Exploratory Factor Analysis (EFA):** Identifies the underlying factor structure without any preconceived notion about how variables relate to factors.
- **Confirmatory Factor Analysis (CFA):** Tests whether the data fits a predefined factor structure, often based on prior theoretical models or research.

3.2 Confirmatory Factor Analysis (CFA)

CFA differs from EFA in that the researcher specifies the number of factors and their relationship to observed variables before conducting the analysis. CFA is typically used to confirm or reject a hypothesized factor structure, making it a **theory-driven** approach.

3.3 Latent Variables and Observed Variables

- **Latent Variables (Factors):** Variables that are not directly observed but are inferred from observed variables. For example, "intelligence" is a latent variable that could be inferred from various observed measures such as problem-solving scores or memory tests.
- **Observed Variables (Manifest Variables):** The actual data points or measurements collected in a study, which are used to infer the latent variables.

3.4 The CFA Model

The relationship between observed variables and latent factors can be represented using the following model:

$$X = \Lambda F + \epsilon$$

Where:

- X is the vector of observed variables.
- Λ is the matrix of factor loadings (the degree to which each observed variable is associated with each factor).
- F is the vector of latent factors.
- ϵ is the vector of error terms (specific variances of each observed variable not explained by the latent factors).

3.5 Steps in CFA

1. **Specify the Model:**
    - Define how many latent factors are expected.
    - Assign each observed variable to one or more latent factors based on theoretical reasoning.
2. **Estimate the Model:**
    - Use software (e.g., Origin Pro 2024) to estimate the factor loadings, error variances, and covariances between factors.
3. **Assess Model Fit:**
    - Various indices are used to evaluate how well the model fits the data, such as the **Chi-square statistic**, **Root Mean Square Error of Approximation (RMSEA)**, and **Comparative Fit Index (CFI)**.
4. **Refine the Model:**
    - Based on model fit statistics, modify the model if necessary by re-specifying factor loadings or allowing certain error terms to correlate.

3.6 CFA Mathematical Representation

For an example with two latent factors (F_1 and F_2) and five observed variables (X_1, X_2, X_3, X_4, X_5), the measurement model can be written as:

$$
\begin{aligned}
X_1 &= \lambda_{11} F_1 + \epsilon_1 \\
X_2 &= \lambda_{21} F_1 + \epsilon_2 \\
X_3 &= \lambda_{31} F_1 + \epsilon_3 \\
X_4 &= \lambda_{42} F_2 + \epsilon_4 \\
X_5 &= \lambda_{52} F_2 + \epsilon_5
\end{aligned}
$$

Where:

- λ_{ij} are the factor loadings (strength of the relationship between latent factors and observed variables).
- ϵ_i are the error terms or unique variances.

3.7 Model Fit Indices in CFA

- **Chi-Square Test (χ^2):** Compares the expected covariance matrix (from the model) to the observed covariance matrix (from the data). A non-significant chi-square indicates a good fit.
- **RMSEA (Root Mean Square Error of Approximation):** Measures how well the model fits the data per degree of freedom. Values below 0.06 indicate a good fit.
- **CFI (Comparative Fit Index):** Compares the fit of the target model to a baseline model. Values above 0.90 indicate a good fit.
- **SRMR (Standardized Root Mean Square Residual):** Measures the difference between observed and predicted correlations. A value less than 0.08 suggests a good fit.

4. Procedure

4.1 Step 1: Input the Data into Origin Pro 2024

- Open Origin Pro 2024.
- Input your data matrix of 10 samples with 5 observed variables into the software.

Example Data (10 Samples with 5 Variables):

Sample No.	X_1	X_2	X_3	X_4	X_5
1	3.1	2.9	3.5	4.1	3.9
2	3.0	2.8	3.7	4.0	3.8
3	3.2	3.0	3.6	4.2	4.0
4	2.9	2.7	3.4	4.3	3.7
5	3.3	3.1	3.8	4.4	4.1
6	2.8	2.6	3.3	4.0	3.6
7	3.5	3.2	3.9	4.5	4.2
8	2.7	2.5	3.2	3.9	3.5
9	3.4	3.0	3.7	4.3	4.0
10	3.1	2.9	3.6	4.2	3.9

4.2 Step 2: Define the Hypothesized Factor Model

- In Origin Pro, define your latent variables (factors) and assign the observed variables to each factor based on theoretical assumptions.

Example Model:

- X_1, X_2, X_3 are associated with Factor 1.
- X_4, X_5 are associated with Factor 2.

4.3 Step 3: Perform CFA

- Use the "Factor Analysis" tool in Origin Pro 2024 to run CFA.
- Select the option to estimate the parameters using the maximum likelihood method.
- Define the number of factors and the relationship between the observed variables and the latent factors as per the hypothesized model.

4.4 Step 4: Evaluate the Model Fit

- After running the analysis, assess the model fit by examining fit indices such as:
  - Chi-Square test statistic.
  - RMSEA.
  - CFI.
  - SRMR.
- Ensure that the fit indices meet acceptable thresholds (e.g., RMSEA < 0.06, CFI > 0.90).

4.5 Step 5: Interpret Factor Loadings

- Check the factor loadings for each observed variable. Ideally, loadings should be above 0.30 or 0.40 to indicate a strong relationship between the latent factor and observed variable.

Example Results:

Variable	Factor 1 Loading	Factor 2 Loading
X_1	0.78	0.05
X_2	0.82	0.07
X_3	0.75	0.09
X_4	0.06	0.88
X_5	0.05	0.84

5. Results

- **Model Fit:** The model shows an acceptable fit with the following fit indices:
  - **Chi-Square Test:** p > 0.05 (indicating a good fit).
  - **RMSEA:** 0.045 (good fit).
  - **CFI:** 0.93 (acceptable fit).
- **Factor Loadings:**
  - Factor 1 strongly correlates with X_1, X_2, X_3.
  - Factor 2 strongly correlates with X_4, X_5.
- **Error Variance:** Each observed variable has a small amount of error variance, indicating that the latent factors explain most of the variance.

6. Learning Outcomes

- Understand the theory and principles behind Confirmatory Factor Analysis (CFA).
- Gain practical experience in using CFA to test predefined factor structures.
- Learn how to assess model fit using various fit indices (Chi-Square, RMSEA, CFI, SRMR).
- Understand the importance of factor loadings and error variances in interpreting CFA results.
- Acquire skills in using Origin Pro 2024 to perform CFA on real-world datasets.