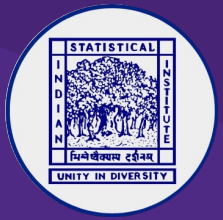


UNFOLDING THE NUMBERS:
School Bullying and Violence

Payal Rajora,
Sayan Bhadra,
Subhojit Maji,
Yash Saraiwala

April 16, 2025

— Foreword —

School-related violence in all of its forms is an infringement of children's and adolescents' rights to education and to health and well-being. No state can achieve inclusive and equitable quality education for all if learners experience violence and bullying in school.

Addressing school violence and bullying is essential in order to achieve the Sustainable Development Goals (SDG's). Indian Penal Code, The Juvenile Justice Act are laws in India which discuss topics regarding child protection and bullying. Right of children to free and compulsory education, Section 16(2) and 17(1) also discuss topics regarding protection of children regarding mental and physical abuse.

Table of Contents

Introduction

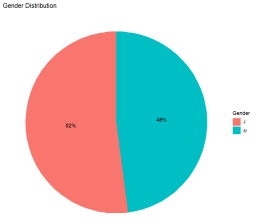
• OBJECTIVE

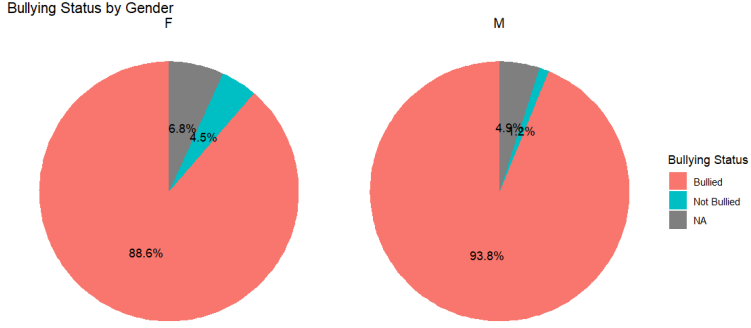
The objective of this report is to analyze the Data set on bullying among school going students obtained from a survey given by Christ University , Bangalore, India and to study the behavior of the bullying involvement, in schools, grades six to twelfth, across the developmental category of early, middle, and late adolescence.

• DATA DESCRIPTION

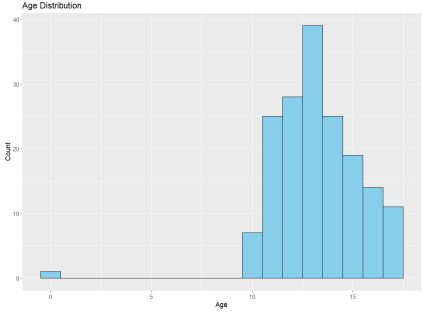
The data was collected from 169 school-going adolescents of grades sixth to twelfth from two cities in South India. The data set contains information of the frequency and type of bullying involvement (perpetration and victimization; physical, verbal, and social) among the participants in traditional and virtual classrooms.

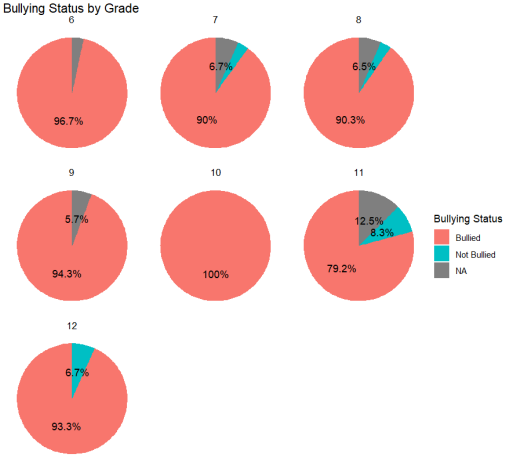
Gender Based Pie-Charts





Age Based Pie-Charts





— Definitions —

- **Bullying —**

Bullying is defined as abuse and mistreatment of someone vulnerable by someone stronger, more powerful. It is characterized by aggressive behavior that involves unwanted , negative actions, is repeated over the time and an imbalance of power and strength between the perpetrators and the victim.

- **Physical Bullying —**

Physical bullying includes repeated aggression such as being hit, hurt, kicked, pushed, shoved around, locked indoors, having things stolen , etc. It is different from other sorts of physical violence such as physical attacks or fights.

— Definitions —

- **Verbal Bullying —**

Verbal bullying is a form of aggression that involves using words to harm, intimidate, or control another person. It can manifest through insults, teasing, name-calling, and other forms of harmful language.

- **Cyber Bullying —**

It includes being bullied by messages like texting something mean instant messages, postings, emails or creating a website that makes fun of a student by pictures or remarks.

- **Social bullying —**

Social bullying, also known as relational bullying or aggression, is a form of bullying used to intentionally damage someone's social reputation or relationships with others.

— Definitions —

- Psychological Bullying —

Includes verbal abuse , Social abuse and social exclusion. For example being called mean names or being teased in an unpleasant way.

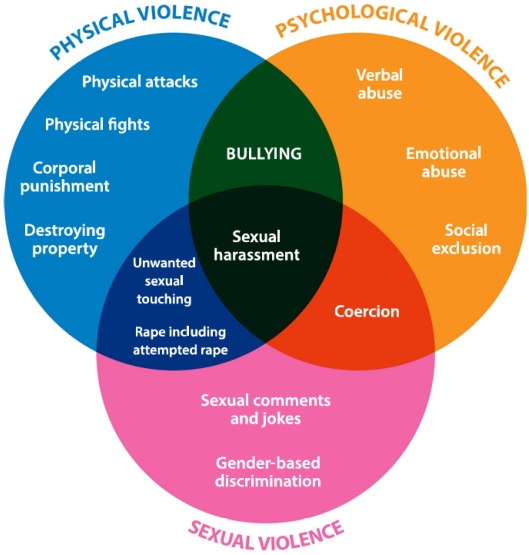
- Sexual Bullying —

Refers to being made fun of with sexual jokes, comments and gestures.

- Victimization —

Victimization refers to the process or state of being harmed, exploited, or oppressed by another person or institution. Victimization is a specific type of discrimination under the law . It means ‘suffering a detriment’ because you’ve done or intend to do a ‘protected act’. victimization has quite specific meaning - while ‘bullying’ doesn’t feature as a legal term at all.

Figure 1. Conceptual framework of school violence and bullying



Methods Used

— Logistic Regression —

Logistic Regression

What is Logistic Regression

- Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not.
- Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.
- For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0.

Understanding Logistic Regression

Logistic regression is a statistical method used for binary classification (predicting one of two possible outcomes, e.g., Yes/No, 1/0). Here's how to perform it:

Goal

Predict the probability that an observation belongs to a particular category.

Model Output

A probability between 0 and 1 (transformed using the logistic function also called as the Sigmoid function). here the logistic function is a function which converts continuous data into probabilities between 0 and 1.

Prepare the data

Prepare the Data

- **Dependent Variable (Y)**: Binary (0 or 1).
- **Independent Variables (X)**: Can be continuous, categorical, or ordinal.
- **Check Assumption**: that there is no multi co-linearity between the predictor variables.

Reason We don't use Linear Regression in our model

Assume we have a dataset that is linearly separable and has the output that is discrete in two classes (0, 1).
The equation for SLR is $y = \beta_0, +\beta_1 x + \epsilon$, where Y is the dependent variable, X is the predictor, β_0, β_1 are coefficients/parameters of the model, and Epsilon(ϵ) is a random variable called error term. In linear regression, we draw a straight line (the best-fit line) L such that the sum of distances of all the data points to the line is minimal. The equation of the line L is $y = mx + c$, where m is the slope and c is the y-intercept.

The two limitations of using a linear regression model for classification problems are:

- the predicted value may exceed the range (0,1)
- error rate increases if the data has outliers

How does Logistic Regression work ?

The logistic regression equation is quite similar to the linear regression model. Consider we have a model with one predictor “X” and one Bernoulli response variable “ŷ” and p is the probability of ŷ = 1. The linear equation can be written as:

- $p = \beta_0 + \beta_1 x$ ————-> eq 1

The right-hand side of the equation $\beta_0 + \beta_1 x$ is a linear equation and can hold values that exceed the range (0,1). But we know probability will always be in the range of (0,1).

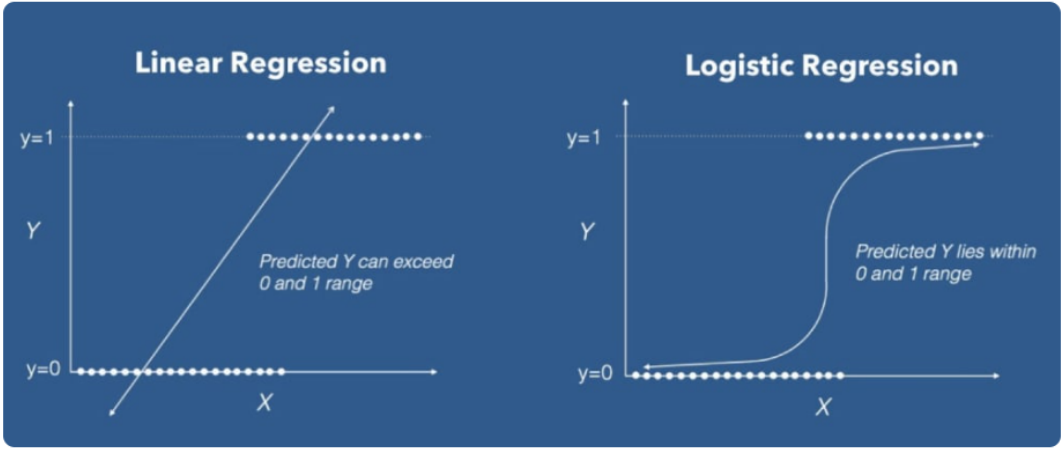
To overcome that, we predict odds instead of probability.

Odds :

- The ratio of the probability of an event occurring to the probability of an event not occurring.
- odds = $\frac{p}{1-p}$
- Odds can only be a positive value, to tackle the negative numbers, we predict the logarithm of odds.

How does Logistic Regression work ?

So, unlike linear regression, we get an ‘S’ shaped curve in logistic regression.

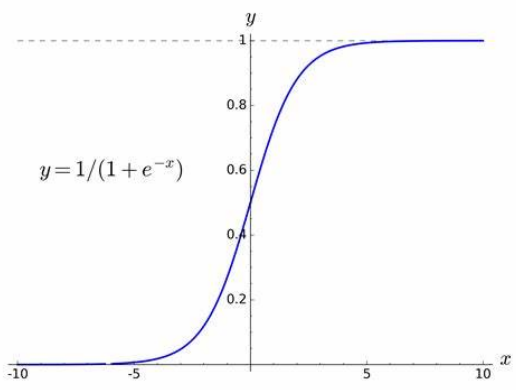


Understanding Sigmoid function

Sigmoid Function

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.

Understanding Sigmoid Function



- As shown above, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.
- $\sigma(z)$tends to 1 as$z \to \infty$
- $\sigma(z)$ tends to 0 as $z \to -\infty$
- $\sigma(z)$ is always bounded between 0 and 1.
- Here, we can measure the probabilities y = 1 or 0 as , $\mathbb{P}(y = 1) = \sigma(z)$ and $\mathbb{P}(y = 0) = 1 - \sigma(z)$

Logistic Regression Equation

Let $X_1, X_2, ..., X_n$ be the predictor variables and let Y be our dependent variable. then,

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Logistic regression computes the probability of some outcome given the predictor variables as,

Logistic Regression Equation

- $\mathbb{P}(Y) = \frac{1}{1+e^{-Y}}$
- $\mathbb{P}(Y) = \frac{1}{1+e^{-(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$
- When we fit a logistic regression model, the coefficients in the model output represent the average change in the log odds of the response variable associated with a one unit increase in the predictor variable.

Remark

the average change in the odds of the response variable associated with a one unit increase in the predictor variable, can be found by using the formula e^{β}.

Multi-collinearity between predictor variables

We can figure out the multicollinearity between predictors by using VIF. i.e, Variance Inflation Factor.

Variance Inflation Factor

- VIF measures multicollinearity (how much one predictor variable is linearly explained by others in the model).
- VIF = 1: No multicollinearity.
- VIF > 5 (or 10): High multicollinearity (consider removing the variable).

Calculating VIF

- For a predictor X_j, $VIF_j = \frac{1}{1-R_j^2}$

Variance Inflation Factor

R squared

- R-squared, also known as the coefficient of determination, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s). Essentially, it indicates how well the data fit a regression line or curve
- The formular for R squared is r^2 . where, r is the correlation coefficient.

interpretation of R squared

- The R-squared value ranges from 0 to 1
- 0 indicates that the model does not explain any of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.

Calculating VIF between gender and age in our data set

Calculating the correlation coefficients

- Correlation coefficient can be found using the formulae, $r = \frac{Cov(X,Y)}{(\sigma(X)\sigma(Y))}$. We calculate and get the correlation coefficient to be 0.045.

Calculating VIF for age and gender

- Here , after calculating, we get the value of VIF . Which is approximately 1. Hence, we can say there is negligible correlation between age and gender.

Predicting Traditional Bullying from gender and age

Here, Our dependent variable is tv (traditional bullying) and independent variable or the predictors are age and gender.
so, we would like to predict using gender and age if someone has faced traditional bullying or not.
we fit the model using statistical software(R) and receive the following output:

	Estimate	Std. Error	z- value	P(> |z|)
(Intercept)	-2.34567	0.45678	-5.135	2.83e-07
gender F	-0.5108	0.2002	-2.551	0.0107
age	0.05678	0.01234	4.601	4.21e-06

Observations

- Here, we can see that the coefficient estimate for gender is negative . That means that being female decreases the chances of experiencing traditional bullying.

Predicting Traditional Bullying from Gender and Age

To understand exactly how gender and age affects whether or not an individual has experienced traditional bullying, we can use the formula e^{β}.

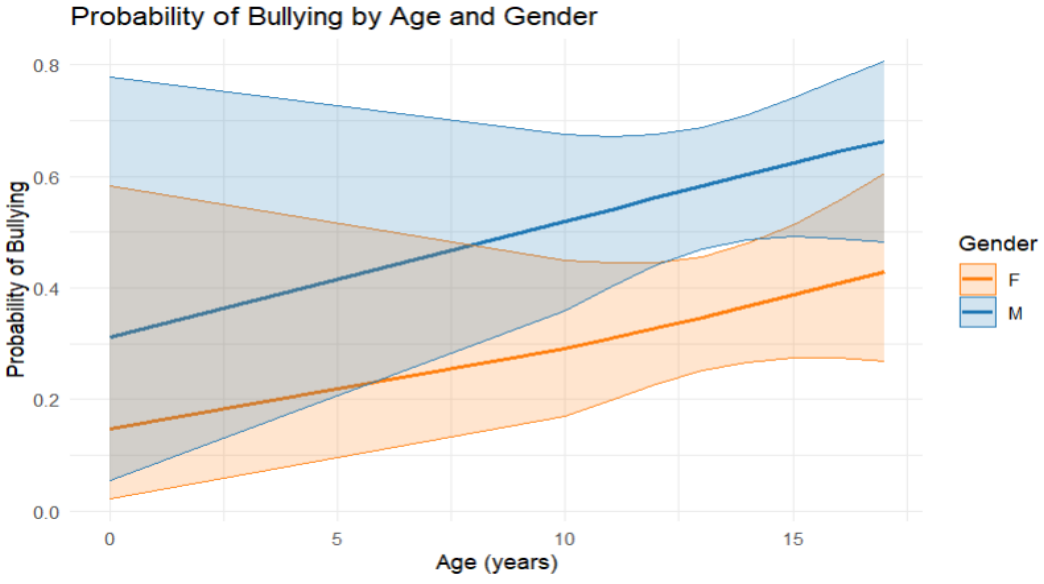
$$\text{Odds ratio (Gender F)} = e^{-0.5108} \approx 0.60$$

$$\text{Odds ratio (Age)} = e^{0.056} = 1.058$$

Observations

- For each 1-year increase in age, the odds of bullying increase by 5.8% (OR = 1.058, $p < 0.001$).
- We interpret this to mean that according to the data given, Females (genderF) have 40% lower odds of traditional bullying compared to males

Predicted probabilities by gender and age



— Principal Component Analysis —

Principal Component Analysis —

- PCA is a statistical technique introduced by mathematician Karl Pearson in 1901.
- Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends.
- PCA is an unsupervised learning algorithm, meaning it doesn't require prior knowledge of target variables. It's commonly used in exploratory data analysis and machine learning to simplify datasets without losing critical information.

Mathematical Foundation of PCA: Goal

Goal: Find a lower-dimensional representation of data while preserving variance. Given a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ (mean-centered), PCA follows these steps:

Understanding the Goal

- **Lower-dimensional representation:** Imagine you have data with many features (columns). PCA aims to reduce this number, making the data easier to visualize, understand, and process. Think of summarizing a long document into a few key sentences.
- **Preserving variance:** Variance in the data represents the spread or variability of the data points. This variability often contains the important information and patterns we want to analyze. PCA tries to find a lower-dimensional representation that retains as much of this original variance as possible. We don't want to lose the crucial information when reducing the dimensions.
- **Mean-centered:** Before applying PCA, the data is typically"mean-centered." This means that for each feature (column), the average value is subtracted from all the data points in that column. This step is important because it centers the data around the origin, which helps PCA to identify the directions of maximum variance more accurately. It ensures that the principal components we find are truly capturing the variance in the data, rather than just the magnitude of the values.

Mathematical Foundation of PCA: Step 1

Step 1: Compute Covariance Matrix

$$\mathcal{A} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

Measures relationships between features.

Mathematical Foundation of PCA: Step 2

Step 2: Eigenvalue Decomposition

Solve:

$$\mathcal{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

where λ_i are eigenvalues and $\mathbf{v}_i$ are eigenvectors.

Mathematical Foundation of PCA: Step 3

Step 3: Select Principal Components

Choose the top k eigenvectors $\mathbf{V}_k$ corresponding to the largest eigenvalues.

Mathematical Foundation of PCA: Step 4

Step 4: Transform Data

Project $\mathbf{X}$ onto new basis:

$$\mathbf{Z} = \mathbf{X}\mathbf{V}_k$$

where $\mathbf{Z}$ is the reduced representation.

Step 4 Explained: Analogy

Think of a Shadow

Imagine shining a light on a 3D object. The shadow cast on a 3D surface is a lower-dimensional representation of the object. PCA is somewhat similar. The principal components are like the best directions to "shine the light" so that the "shadow" (the projection, $\mathbf{Z}$) retains the most important features and variations of the original "3D object" (the data, $\mathbf{X}$).

Key Property: Principal components maximize variance and are uncorrelated.

Our goal is to find a lower-dimensional representation of data while preserving variance. Hence, according to the steps explained before, we first calculate the covariance matrix.

Component-wise:

$$S_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

- Diagonal: Variances (S_{jj})
- Off-diagonal: Covariances ($S_{jk}, j \neq k$)

2D Example:

$$\begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

Where:

- $\sigma_{xy} = \rho \sigma_x \sigma_y$
- $-1 \leq \rho \leq 1$ (correlation)

Key Properties

Symmetric, positive semi-definite, captures linear relationships.

Heat Map of Covariance Matrix

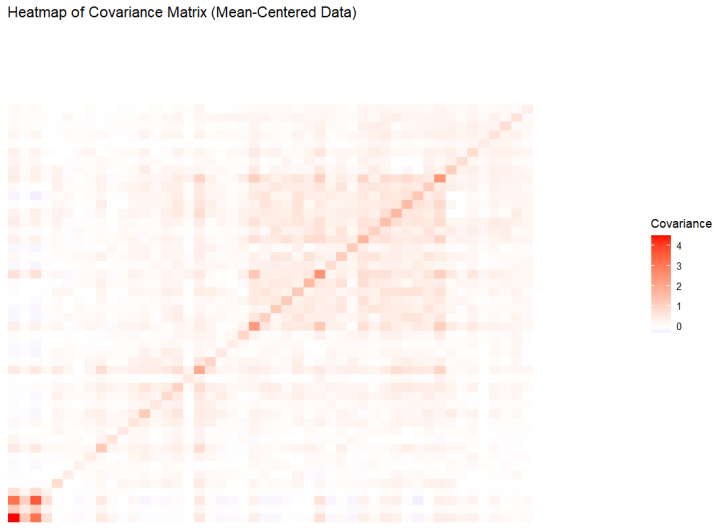


Figure: HeatMap of the Covariance Matrix

Geometric Explanation

Geometric Explanation of PCA

- Principal component analysis works by rotating the axes to produce a new coordinate system. It projects the data along the directions where the data varies the most.
- The first direction is decided by $\mathbf{v_1}$ corresponding to the largest eigenvalue λ_1. The second direction is decided by $\mathbf{v_2}$ corresponding to the second largest eigenvalue λ_2 and so on.
- The variance of the data along the principal component directions is associated with the magnitude of the eigenvalues.
- The choice of how many components to extract is fairly arbitrary.

Interaction of PCA with correlation and collinearity

PCA and uncorrelated components

- The primary function of PCA is to transform a set of potentially correlated variables into a new set of uncorrelated variables, the principal components.
- Therefore, even if the original dataset has "very less components that are correlated to each other," PCA will still produce perfectly uncorrelated principal components.

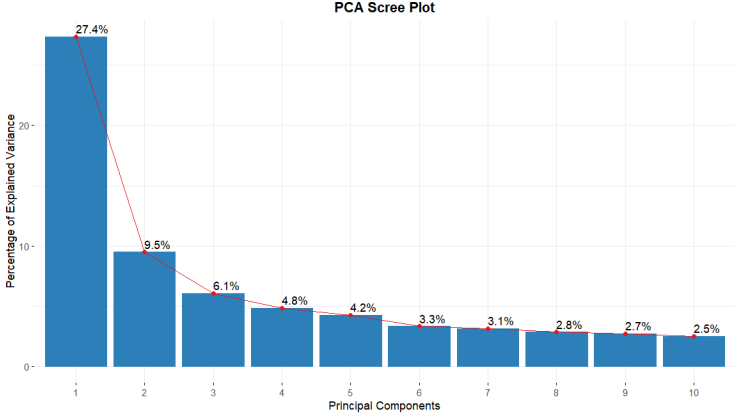
Collinearity Mitigation

- Collinearity can cause problems in linear models, such as unstable coefficient estimates and difficulty in interpreting the effects of individual variables.
- By creating uncorrelated components, PCA effectively eliminates collinearity among the transformed variables.

Visualizing Eigenvalues: The Scree Plot

Purpose of the Scree Plot

The scree plot is a line plot of the eigenvalues (or the percentage of variance explained by each principal component) against the component number. It helps in determining the number of principal components to retain for further analysis.



PCA Scree Plot Results

Variance Explained by Principal Components

Principal Component	Eigenvalue	Variance Explained (%)	Cumulative (%)
PC1	14.2	27.4	27.4
PC2	8.6	19.5	46.9
PC3	4.3	6.1	53.0
PC4	3.1	4.8	57.8
PC5	2.7	4.2	62.0
PC6	2.4	3.3	65.3

Variance explained by principal components

Interpreting the Scree Plot

Component Selection Rules

- **Kaiser's Rule**: Retain components with eigenvalues >1 (if using standardized data)
- **Elbow Method**: Retain components before the "elbow" (PC3 in our case)
- **Variance Threshold**: Keep enough components to explain 70-90% of variance

PC1: General Bullying Involvement

- **Variance explained**: 27.4%
- **Key loadings**:
  - Traditional bullying (tb): 0.92
  - Traditional victimization (tv): 0.89
  - Physical bullying (pb): 0.85
  - Physical victimization (pv): 0.83
- **Interpretation**: Represents overall involvement in bullying (both perpetration and victimization)

PC2: Online vs Traditional Bullying

- **Variance explained**: 19.5%
- **Key loadings**:
  - Online victimization (ov): 0.91
  - Traditional bullying (tb): -0.72
  - Traditional victimization (tv): -0.68
- **Interpretation**: Contrasts online victimization with traditional bullying behaviors

PC3: Verbal/Social vs Physical Bullying

- **Variance explained**: 6.1%
- **Key loadings**:
  - Verbal bullying (vb): 0.85
  - Social bullying (sb): 0.82
  - Physical bullying (pb): -0.79
  - Physical victimization (pv): -0.76
- **Interpretation**: Distinguishes verbal/social forms from physical forms of bullying

Demographic Patterns in Bullying

	PC1 (General)	PC2 (Online)	PC3 (Verbal/Social)
Gender			
Male	0.42*	-0.18	0.31*
Female	-0.38*	0.21	-0.28*
Grade Level			
Middle School	0.25*	-0.12	0.19
High School	-0.22*	0.15	-0.17
SES			
Low	0.31*	-0.05	0.12
High	-0.24*	0.02	-0.08

PCA Recap

Principal Component Analysis (PCA):

- Transforms correlated variables into uncorrelated **principal components (PCs)**.
- Each PC is a weighted combination of original variables:

$$\text{PC}_k = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p$$

- **Eigenvectors**: Weights (w_i) for each variable.
  Eigenvalues: Variance explained by each PC.

Demographic Scores in PCA

What are demographic scores?
Mean component scores for groups (e.g., males/females) on each PC.

Steps to Calculate:

1. Compute individual PC scores for all students.
2. Group scores by demographics (gender, SES, etc.).
3. Average scores within each group.
4. Standardize (mean=0, SD=1) for comparability.

Example Calculation: PC1 Scores for Males

Given:

- PC1 loadings:

$$\text{tb} = 0.92,\ \text{tv} = 0.89,\ \text{pb} = 0.85,\ \text{pv} = 0.83$$

- A male student's standardized responses:

$$\text{tb} = 1.2,\ \text{tv} = 0.8,\ \text{pb} = 1.1,\ \text{pv} = 0.9$$

PC1 Score for this student:

$$(0.92 \times 1.2) + (0.89 \times 0.8) + (0.85 \times 1.1) + (0.83 \times 0.9) = 3.29$$

Average PC1 for all males = 0.42* (standardized).
*Significant at $p < 0.05$ (t-test vs. overall mean).

Interpreting Demographic Scores

Group	PC1 (General)	PC2 (Online)	PC3 (Verbal/Social)
Male	0.42*	-0.18	0.31*
Female	-0.38*	0.21	-0.28*
Low-SES	0.31*	-0.05	0.12

Key Insights:

- Males score higher on PC1: More general bullying involvement.
- Females score lower on PC3: Less verbal/social bullying.
- Low-SES groups higher on PC1: Socioeconomic risk factor.

Why Demographic Scores Matter

Applications:

- **Targeted Interventions**: Prioritize high-risk groups (e.g., low-SES students).
- **Behavioral Patterns**: Address gender-specific trends (e.g., male verbal aggression).
- **Policy Design**: Tailor anti-bullying programs to PC dimensions (general vs. online).

Limitations:

- Self-report biases may affect scores.
- Cross-sectional data $\rightarrow$ no causality.

Summary

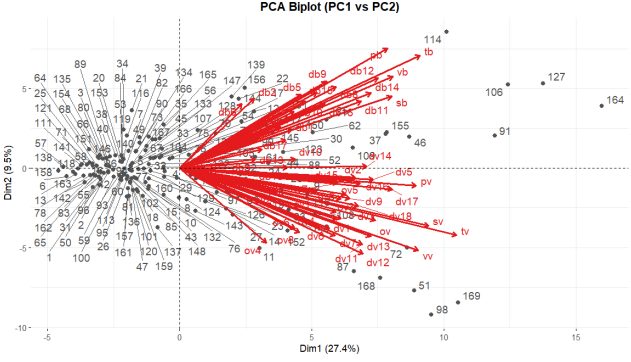
- Demographic scores = Mean PC scores for groups.
- Calculated by:
  1. Weighting responses by PCA loadings.
  2. Averaging within demographics.
- Reveal actionable patterns (e.g., males $\rightarrow$ PC1).

Understanding the Biplot in PCA

Purpose of Biplots

- A **biplot** is a graphical representation of the results of Principal Component Analysis (PCA).
- It simultaneously displays:
  - **Data Points:** Represent individuals (observations) in the reduced-dimensional space.
  - **Variable Arrows:** Show how original variables relate to the principal components.

PCA Biplot Interpretation



Key to Elements

- Gray points: Survey respondents
- Red arrows: Bullying variables
- Axis labels: % variance explained

Most Influential Variables

- Longest arrows contribute most to PCs
- Angles show variable correlations

Interpreting Variable Relationships

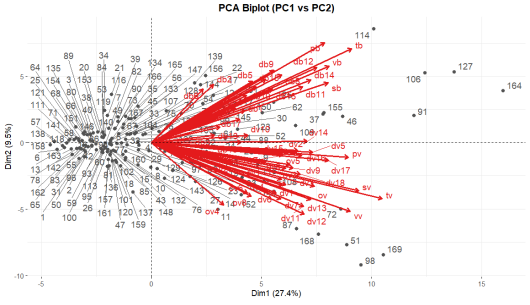
Arrow Direction Interpretation

- **Right-pointing**: High values on PC1 (27.4%)
  - Likely general bullying severity factors
- **Upward-pointing**: High values on PC2 (19.5%)
  - May represent specific bullying types

Example Correlations

- Variables pointing similarly: Positive correlation
- Opposite directions: Negative correlation
- 90° angle: Uncorrelated

Respondent Clustering Patterns



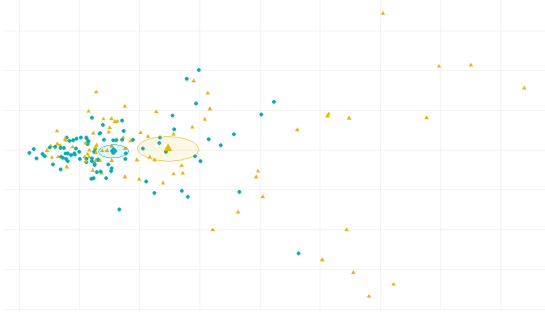
Cluster Interpretation

- **Top-right**: High on both PC1 & PC2
  - Frequent bullying involvement
- **Bottom-left**: Low on both PCs
  - Minimal bullying experiences

Example —

Points 118-132 cluster near the physical bullying arrow, suggesting these students experience more physical aggression. Their position far right on PC1 indicates high overall bullying severity.

Respondent Groups in Bullying Space



How to Interpret

- Each point represents one student
- Blue circles = male students
- Yellow triangles = female students
- Ellipses show 95% confidence intervals

Key Observation

Partial gender separation with substantial overlap suggests some gender-specific patterns in bullying behaviors.

Gender Differences in Bullying Behaviors

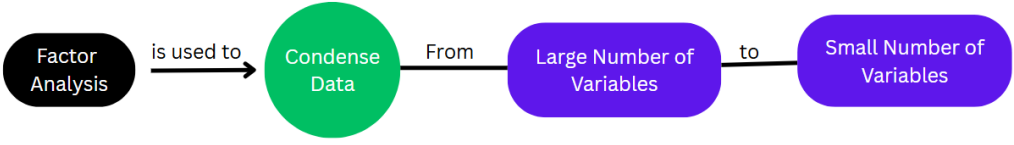
- Male students (blue) cluster more densely on the left side
  - Suggests more homogeneous bullying behavior patterns
  - Less variability in reported behaviors
- Female students (yellow) show wider dispersion
  - Indicates more heterogeneous bullying behavior patterns
  - Greater variability in how female students engage with or experience bullying
- Overlap between groups indicates shared behaviors across genders

— Factor Analysis —

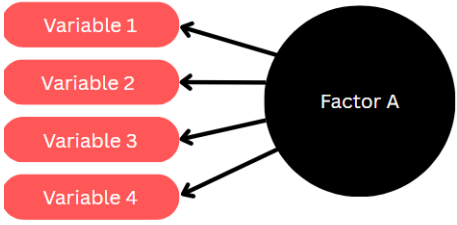
Factor Analysis

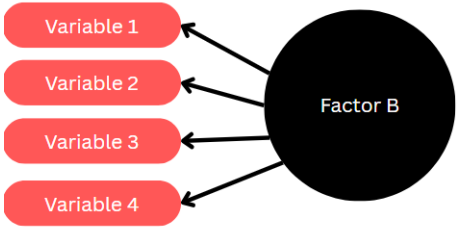
Factor Analysis

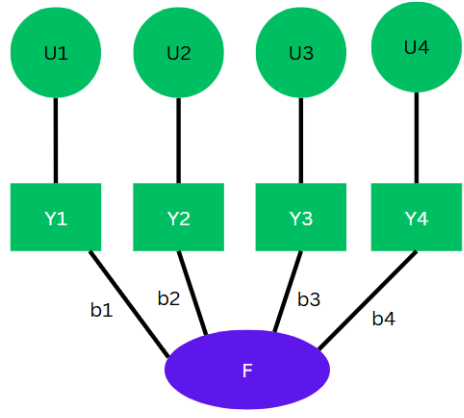
- Factor Analysis is a statistical method used to describe variability among observed , correlated variables in terms of a potentially lower number of unobserved variables called factors .
- Unobserved (Latent) variables: These are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly measured.



Factor Analysis







Key Features

Key features of factor Analysis

- Dimensionality Reduction : Summarizes numerous variables into fewer factors.
- Latent Variables: Identifies underlying constructs not directly observable.
- Exploratory or Confirmatory: Can be used to explore data or test hypotheses.

Importance

Importance of Factor Analysis

- Simplifies Complex Data: Reduces redundancy among correlated variables.
- Construct Validity: Assesses whether variables measure the intended constructs.
- Improves Interpretation: Helps researchers identify patterns and relationships.
- Data Reduction: Optimizes datasets for further analysis or model development.
- Supports Theory Development: Identifies structures or dimensions underlying a phenomenon.

Types of Factor Analysis

Exploratory Factor Analysis

- Purpose: Discovers the underlying structure of a dataset without prior assumptions.
- Use : Initial stages of research when relationships between variables are unknown.

Confirmatory Factor Analysis

- Purpose: Tests whether the data fit a predefined factor structure based on theoretical expectations.
- Use : Later stages of research to validate hypothesized relationships.

Exploratory Factor Analysis

We perform exploratory factor analysis by following these steps:

- Find the correlation between the variables.
- Conduct the Kaiser-Meyer-Olkin (KMO) test for sampling adequacy . If selected variables satisfy criteria for KMO Test then we proceed with next step.
- Perform Bartlett's test of sphericity to check if factor analysis is appropriate.
- Conduct factor analysis on the selected variables(if appropriate).

Correlation matrix

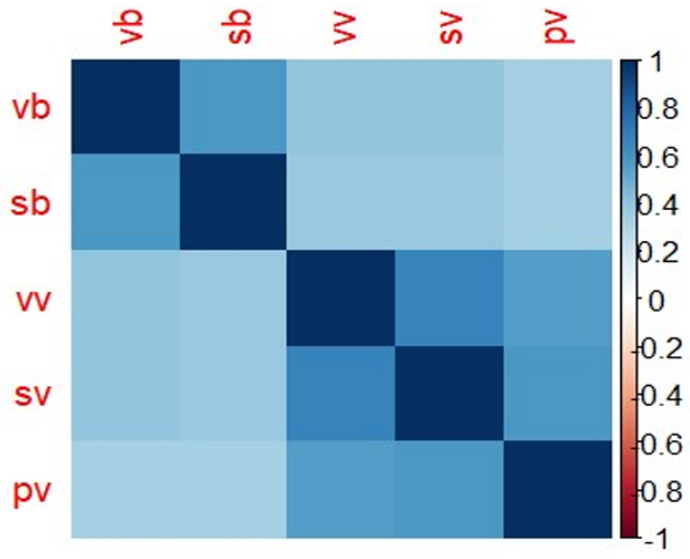
R code

```
library(psych)
library(GPArotation)
install.packages("corrplot")
library(corrplot)
.
data <- read.csv("Survey Data2.csv", header = true)
bully_vars <- data[, grep("^b|sb|vv|sv|pv", names(data))]
cor_matrix <- cor(bully_vars, use = "complete.obs")
print(cor_matrix)
corrplot(cor_matrix , method = 'color')
```

Output (Correlation Matrix)

	vb	sb	vv	sv	pv
vb	1.000000	0.579157	0.390472	0.3983518	0.3303640
sb	05791517	1.000000	0.3736662	0.377375	0.3359319
vv	0.390472	0.373662	1.000000	0.663334	0.5586405
sv	0.3983518	0.377375	0.6633341	1.000000	0.5726202
pv	0.3303640	0.3359319	0.5586405	0.5726202	1.000000

Visualization



Observations

Observation 1

- Strong Correlations (r > 0.6)
- vv & sv(r = 0.663).
- **Interpretation** : Victims of one bullying are also likely to experience another.

Observation 2

- Moderate Correlations (0.4 - 0.6).
- sv & pv(r = 0.573)
- vv & pv(r = 0.559)
- vb & sb(r = 0.579)
- **Interpretation** :Those who engaged in a bullying may have experienced another

Observations

Observation 3

- Weak Correlations ($r < 0.4$)
- vb & vv($r = 0.390$)
- vb & sv($r = 0.398$)
- sb & sv($r = 0.377$)
- vb & pv($r = 0.330$)
- sb & pv($r = 0.335$)
- **Interpretation**:The two variables have a slight relationship, but it is not strong or consistent . This suggests that bullies may occur independently rather than together

Kaiser-Meyer-Olkin (KMO) Test

The KMO test is a statistical test that is used to determine whether factor analysis is appropriate for a given data set. It measures the adequacy of the data for exploratory factor analysis (EFA).It checks whether the correlations between variables are large enough for meaningful factor analysis. It evaluates the proportion of common variance among variables, helping to determine whether factor analysis is useful.

Interpretation of KMO values

KMO Value	Interpretation
0.90 – 1.00	Excellent for factor analysis
0.80 – 0.89	Good for factor analysis
0.70 – 0.79	Acceptable for factor analysis
0.60 – 0.69	Could be used but not ideal
0.50 – 0.59	Factor analysis is questionable
Below 0.50	Unacceptable

KMO test

R Code

```
library(mice)
bullyvars <- mice::completer(mice(bullyvars, method = "pmm" , m=1))
KMO(bullyvars)                KMO test
```

Output

Kaiser-Meyer-Olkin factor adequacy call: KMO(r = bullyvars)
overall MSA = 0.79
MSA for each item =

vb	sb	vv	sv	pv
0.76	0.77	0.79	0.79	0.83

KMO test

A KMO value above 0.70 generally indicates that the data is suitable for factor analysis.
MSA for Individual Variables:
vb = 0.76 (Good)
sb = 0.77(Good)
sv = 0.79 (Good)
pv = 0.83 (Strong)
vv = 0.79 (Good)
As the overall KMO value is 0.79 which is above 0.70 and all individual variables have MSA values above the acceptable threshold , the dataset is suitable for factor analysis.
Hence, we proceed with Bartlett's test of sphericity to further confirm whether factor analysis is appropriate.

Bartlett's Test of Sphericity

Bartlett's test of sphericity is used to determine whether a correlation matrix is significantly different from an identity matrix. It is commonly used in factor analysis to check if there are significant relationships between variables.

Hypothesis of Bartlett's test

- Null Hypothesis (H□): The correlation matrix is an identity matrix (no significant correlations among variables)
- Alternative Hypothesis (Ha): The correlation matrix is not an identity matrix, meaning that at least some variables are correlated and suitable for factor analysis.
- Interpretating the Bartlett's Test Results , we get:Chi-Square (X^2) Value = 269.7701 . The chi-square test statistic measures how much the correlation matrix deviates from an identity matrix.

R code

```
n <- nrow(bullyvars)
coetest.bartlett(cor(bullyvars, use = "pairwise.complete.obs",n = n).
```

Bartlett's Test

output

- $chisq
  269.7701
- $pvalue
  3.739243e-52
- $df
  10

Degree of Freedom

- Degrees of Freedom (df) = 10
- The degrees of freedom are determined by the number of variables in the dataset.
- The formula for degrees of freedom in Bartlett's test is:df=p(p−1)/2where p is the number of variables.
- In this case, with 5 variables (vb, sb, sv, pv, vv), so df=10.The degrees of freedom help define the expected chi-square distribution for comparison.
- p-value =3.74e-52
- Since the p-value < 0.05, we reject the null hypothesis
- This means that the correlation matrix significantly differs from an identity matrix , confirming that factor analysis is appropriate

conclusion

Since Bartlett's test is statistically significant (p-value < 0.05), the dataset is suitable for factor analysis.

Factor Analysis

```
fa.parallel(bullyvars, fa = "both")
fa.result <- fa(bullyvars, nfactors = 2, fm = "pa", rotate = "varimax", scores =
"Bartlett")              # Adjust 'nfactors'
print(fa_result, cut = 0.3)              #Displayloadingsgreaterthan0.3
fa.diagram(fa_result)
```

output

```
> fa.parallel(bullyvars, fa = "both")
Parallel analysis suggests that the number of factors = 2 and the number of
components = 1.
```

Output

Output

```
     PA1   PA2    h2    u2 com
vb        0.70 0.56 0.44 1.3
sb 0.31 0.69 0.58 0.42 1.4
vv 0.73      0.59 0.41 1.2
sv 0.73      0.61 0.39 1.3
pv 0.65      0.50 0.50 1.4

                       PA1  PA2
SS loadings           1.65 1.18
Proportion Var        0.33 0.24
Cumulative Var        0.33 0.57
Proportion Explained  0.58 0.42
Cumulative Proportion 0.58 1.00
```

Understanding each factor

Factor (PA 1) - ”Victimization”

- High loadings on: Social Victimization (0.73), Physical Victimization (0.65), and Verbal Victimization (0.73).
- Interpretation: This factor primarily represents victimization experiences—students who experience different forms of bullying.

Factor 2 (PA 2) - ”Bullying Aggression”

- High loading on: Verbal Bullying (0.70) and Social Bullying(0.69)
- Interpretation: This factor is mainly associated with bullying behavior

Variance Explained

Factor analysis aims to reduce a large number of variables into fewer meaningful components while still retaining most of the information.

Statistic	Factor 1 (PA1)	Factor 2 (PA2)
SS Loadings	1.65	1.18
Proportion of Variance Explained	33.0%	24.0%
Cumulative Variance Explained	33.0%	57.0%

What does this mean ?

- Factor 1 explains 33.0% of the variance in the dataset.
- Factor 2 explains 24% of the variance
- Together, they explain 57% of the total variance, meaning these two factors capture most of the patterns in bullying and victimization data.

Factor Loading

A factor loading is a numerical value that shows how strongly an observed variable (e.g., ”verbal bullying”)is related to a latent factor (e.g., ”bullying aggression”). It represents the correlation between a variable and a factor. Factor loadings range from -1 to +1:

- High positive loading (close to +1) → The variable is strongly related to the factor.
- Low loading (close to 0) → The variable is weakly related to the factor.
- High negative loading (close to -1) → The variable is strongly related but in the opposite direction.

Formula for proportion of Variance

For each Factor:
Proportion of Variance = SS Loadings of Factor/Total number of variables where SS Loadings is the sum of squared factor loadings.
Total Number of Variables (p) is the number of observed variables used in the factor analysis.

Factor	SS Loadings	Total Variables (p)	Proportion of Variance
Factor 1 (PA1 - Victimization)	1.65	5	1.65/5=0.33(33.0%)
Factor 2 (PA2 - Bullying Aggression)	1.18	5	1.18/5=0.24(24%)

Cumulative variance

The cumulative variance is simply the sum of the variance explained by each factor:
In this case:
Factor 1 explains 33.0%
Factor 2 explains 24.0%
Total (Cumulative Variance) = 33.0% + 24.0% = 57.0%
This means that 57.0% of the total variability in the dataset is explained by these two factors, which is considered a good result in factor analysis.

Communalities (h^2):

Shows the proportion of variance in each variable explained by the extracted factors.
Higher values mean the factors explain more variance in that variable.

Unique variance

Represents the unexplained variance of each variable.

Complexity

Indicates how many factors each variable is substantially loading on.
Lower values (closer to 1) suggest the variable mainly loads on one factor

Communalities (h^2) and Unique Variance (u^2)

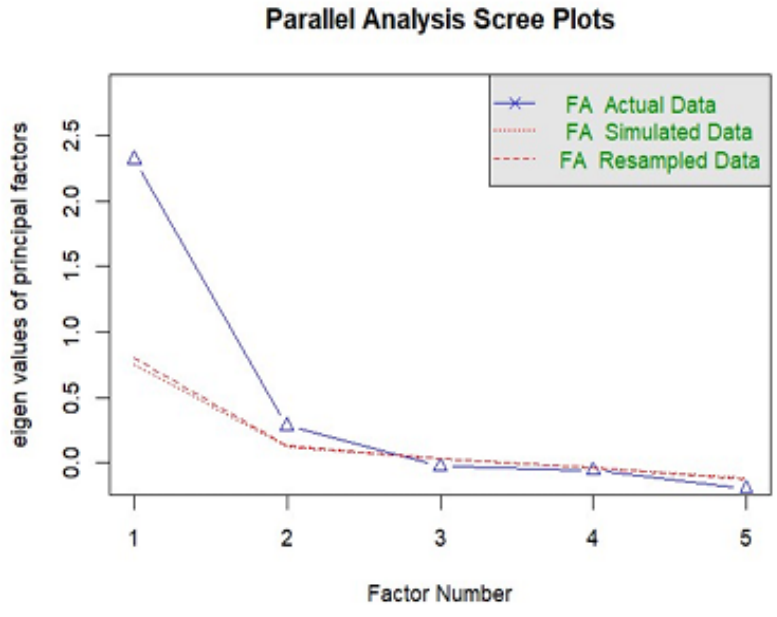
h^2=(Factor loading on PA1)^2+ (Factor loading on PA2)^2 and u^2=1- h^2

Variable	h^2 (Communality)	u^2=1- h^2 (Unique Variance)	Interpretation
vb	0.56	0.44	Well explained by factors.
sb	0.58	0.42	Well explained by factors.
vv	0.59	0.41	Well explained by PA1.
sv	0.61	0.39	Strong factor representation.
pv	0.50	0.50	Moderately explained.

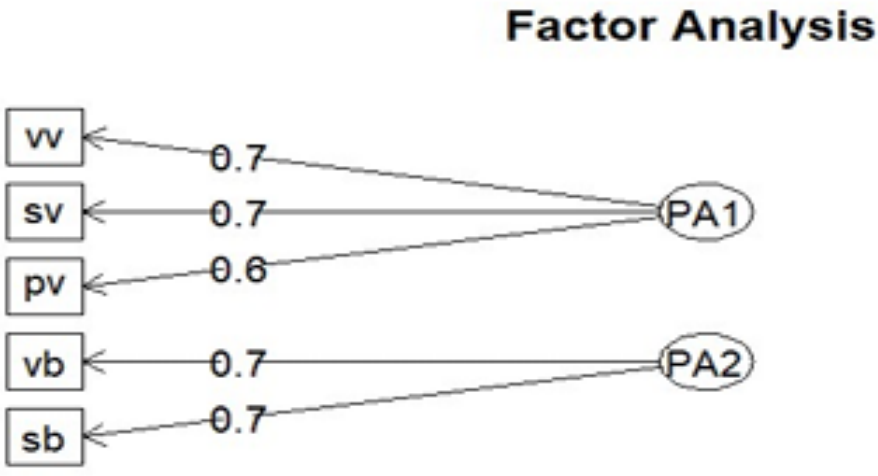
Complexity(com)

Variable	Complexity (com)	Interpretation
vb	1.3	Slight cross-loading but mainly on one factor.
sb	1.3	Similar to vb, loads mostly on one factor.
vv	1.2	Simple structure, loads mostly on one factor.
sv	1.3	Mostly loads on one factor.
pv	1.4	Mostly loads on one factor.

R Plot


Parallel Analysis Scree Plots
FA Actual Data
FA Simulated Data
FA Resampled Data
eigen values of principal factors
0.0
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
Factor Number

R Plot



Final Interpretation

The analysis suggests two main dimensions of bullying involvement:
Factor 1 (PA1) –Victimization: Captures students who experience social, verbal, and physical victimization.
Factor 2 (PA2) –Bullying Aggression: Captures students who engage in verbal bullying(vb) and social bullying(sb) with verbal bullying as a slightly stronger indicator of bullying behavior.

Confirmatory Factor Analysis (CFA)

Now, we use Confirmatory Factor analysis (CFA) on our dataset.
To Confirm a Factor Structure : CFA checks if observed variables correctly belong to their expected factors based on theory.
To Ensure Measurement Validity : CFA verifies that each variable strongly relates to its factor and not to unrelated factors.

R code

```r
# Install and load necessary libraries
install.packages("lavaan")
install.packages("semPlot")
library(lavaan)
library(semPlot)

# Load the data
data <- read.csv("Survey Data2.csv")

# Define the CFA model
cfa_model <- '
  Victimization =~ sv + pv + vv
  Bullying_Aggression =~ vb + sb
'

# Fit the CFA model
fit <- cfa(cfa_model, data = data)

# Print summary of results with fit measures
summary(fit, fit.measures = TRUE, standardized = TRUE)

# Plot the CFA model
semPaths(fit, whatLabels = "std", layout = "tree",
         edge.label.cex = 1.2, title = TRUE,
         style = "lisrel", curve = 2, nCharNodes = 0)
```

Output

```
  Estimator                                      ML
  Optimization method                        NLMINB
  Number of model parameters                     11

  Number of observations                        169

Model Test User Model:

  Test statistic                              1.683
  Degrees of freedom                              4
  P-value (Chi-square)                        0.794

Model Test Baseline Model:

  Test statistic                            275.475
  Degrees of freedom                             10
  P-value                                     0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                 1.000
  Tucker-Lewis Index (TLI)                    1.022
```

Output

```
Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -2243.516
  Loglikelihood unrestricted model (H1)      -2242.675

  Akaike (AIC)                                4509.032
  Bayesian (BIC)                              4543.461
  Sample-size adjusted Bayesian (SABIC)       4508.632

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.075
  P-value H_0: RMSEA <= 0.050                    0.890
  P-value H_0: RMSEA >= 0.080                    0.040

Standardized Root Mean Square Residual:

  SRMR                                           0.012

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured
```

Output

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
Victimization =~						
sv	1.000				3.730	0.783
pv	0.860	0.104	8.302	0.000	3.206	0.712
vv	1.187	0.137	8.641	0.000	4.426	0.757
Bullying_Aggression =~						
vb	1.000				2.422	0.732
sb	0.840	0.130	6.481	0.000	2.035	0.777

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
Victimization ~~						
Bllyng_Aggrssn	6.216	1.224	5.076	0.000	0.688	0.688

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.sv	8.789	1.538	5.713	0.000	8.789	0.387
.pv	9.984	1.436	6.951	0.000	9.984	0.493
.vv	14.634	2.347	6.234	0.000	14.634	0.428
.vb	5.076	0.965	5.262	0.000	5.076	0.464
.sb	2.710	0.631	4.294	0.000	2.710	0.396
Victimization	13.912	2.576	5.400	0.000	1.000	1.000
Bllyng_Aggrssn	5.868	1.318	4.451	0.000	1.000	1.000

Explanation

- Estimator: ML (Maximum Likelihood) – a common estimation method in CFA that assumes normally distributed data.
- Optimization method: NLMINB, a numerical optimization algorithm.
- Number of parameters: 11, indicating the number of estimated values (factor loadings, variances, covariances)
- Parameter : Factor, Latent , Factor , Residual Type Loadings , Variances, Covariance, Variances Count 5,2,1,3
  Total 11
- Number of observations: 169, meaning the analysis is based on 169 participants

Explanation

- Test statistic : 1.683
- Degrees of freedom : 4
- P-value (Chi-square): 0.794
- Here it is Chi-Square Test (χ^2):χ^2 = 1.683, df = 4, p = 0.794
- A high p-value (>0.05) suggests that the model fits well (i.e., the difference between the model and data is not significant)

Explanation

Test statistic: 275.475
Degrees of freedom: 10
P-value :0.000
The baseline model assumes no relationships between variables.
The high chi-square value (275.475, $p < 0.001$) suggests that this model fits poorly compared to our CFA model

Explanation

- Comparative Fit Index (CFI):1.000
- Tucker-Lewis Index (TLI):1.023

Fit Index	Value	Threshold	Interpretation
CFI	1.000	≥ 0.90	**Perfect Fit**
TLI	1.022	≥ 0.90	**Overfitting slightly**

Explanation

Loglikelihood user model (H0)	-2243.516
Loglikelihood unrestricted model (H1)	-2242.675

Likelihood Ratio Test (LRT):

- The **Likelihood Ratio Test** (LRT) uses the difference between the log-likelihoods of the two models to assess whether the more constrained model (H0) fits the data significantly worse than the more complex, unconstrained model (H1).
- The Test Statistic Formula:

$$\text{Chi-Square} = -2 \times (\text{Loglikelihood of H0} - \text{Loglikelihood of H1})$$

The test statistic follows a **Chi-Square distribution**.

Null Hypothesis (H0): The restricted (user) model fits the data as well as the unrestricted model.

Alternative Hypothesis (H1): The unrestricted model fits the data significantly better than the restricted model

Likelihood ratio test

We have

- Log-likelihood of H0 (user model): -2243.516
- Log-likelihood of H1 (unrestricted model): -2242.675

Thus **Chi-Square**=−2×(−2243.516−(−2242.675))=−2×(−0.841)=1.682

- Degrees of Freedom (df): Here df=1
- From R, using a **Chi-Square distribution** with df = 1, the calculated p-value for the test statistic (1.682) is 0.19
- Since p-value>0.05, we fail to reject the null hypothesis.

- Thus, there is no significant difference between the restricted model (H0) and the unrestricted model (H1).

Explanation

- Akaike (AIC) 4509.032
- Bayesian (BIC) 4543.461
- Sample-size adjusted Bayesian (SABIC) 4508.632

AIC/BIC: Used for model comparison. Lower values indicate better model fit.

SABIC: A modified BIC adjusted for small sample sizes.

Explanation

- **AIC** and **SABIC** are very close in value (4509.032 vs. 4508.632), suggesting that the model fits the data reasonably well, with only a marginal difference between the two.
- Since **SABIC** adjusts for smaller sample sizes and is only slightly lower than AIC, it suggests that this model provides a good balance between fit and complexity.

Conclusion

The model shows a reasonable fit to the data, and the difference between the indices is small enough to suggest that no drastic changes are needed

Explanation

- RMSEA 0.000
- 90 Percent confidence interval - lower 0.000
- 90 Percent confidence interval - upper 0.075
- P-value H_0: RMSEA <= 0.050 0.890
- P-value H_0: RMSEA >= 0.080 0.040

Root Mean Square Error of Approximation (RMSEA): 0.000 (perfect fit)

90% CI: [0.000, 0.075]

The upper bound (0.075) is slightly above 0.06, but since the p-value (0.890) is high, the model is still acceptable.

Conclusion: The model has minimal approximation error.

Explanation

- SRMR(Standardized Root Mean Square Residual) 0.012
  SRMR ≤ 0.08 is considered good.

  SRMR = 0.012 indicates a nearly perfect fit

Overall Conclusion: The model fits the data very well.

Factor Loadings

Observed Variable	Factor	Unstandardized Estimate	Standardized Estimate	Interpretation
sv (Social Victimization)	Victimization	1.000	**0.783**	Strong relationship
pv (Physical Victimization)	Victimization	0.860	0.712	Moderate relationship
vv (Verbal Victimization)	Victimization	1.187	**0.757**	Strong relationship
vb (Verbal Bullying)	Bullying Aggression	1.000	0.732	Strong relationship
sb (Social Bullying)	Bullying Aggression	0.840	**0.777**	Strong relationship

Interpretation

- All loadings are above 0.7, meaning that the variables strongly define their latent factors.
- Social victimization (sv) and verbal victimization (vv) have the highest loadings, suggesting they are the most representative indicators of Victimization.
- Social bullying (sb) has the highest loading in the Bullying Aggression factor.

Factor Correlation

Bllyng_Aggrssn	6.216	1.224	5.076	0.000	0.688	0.688

Correlation = 0.688, $p < 0.001$

Interpretation: A moderate to strong positive correlation (0.688) between Victimization and Bullying Aggression.

This suggests that students who experience victimization are also more likely to engage in bullying behaviors.

Variances and Residuals

Residual variance represents unexplained variance in observed variables.

Lower values = better.

Variance Explained (%):

sv = (1 - 0.387) × 100 = 61.3%

pv = (1 - 0.493) × 100 = 50.7%

vv = (1 - 0.428) × 100 = 57.2%

vb = (1 - 0.464) × 100 = 53.6%

sb = (1 - 0.396) × 100 = 60.4%

Conclusion: sv, vv, sb have the highest explained variance (~60%).

pv has the highest residual variance, meaning it is the least strongly predicted.

Visualization

Final Summary

- Model Fit is Excellent (CFI = 1.000, RMSEA = 0.000, SRMR = 0.012).
- All Factor Loadings are Strong (>0.7).
- Victimization & Bullying Aggression are Positively Correlated (0.688, $p < 0.001$).
- Most variance explained (~60%) for sv, vv, sb.

Final Conclusion of Factor Analysis

The factor analysis on the five observed variables (**vb, sb, sv, pv, vv**) successfully identified **two distinct latent factors**, confirming the theoretical structure of the data.

Two factors were extracted, explaining a total of 57% of the variance in the dataset.

Factor 1 ("Victimization") includes high loadings from sv (social victimization), pv (physical victimization), vv (verbal victimization).

Factor 2 ("Bullying/Aggression") includes high loadings from vb (verbal bullying) and sb (social bullying).

This suggests that victimization and bullying aggression are distinct but related psychological constructs.

Intervention Strategies for victimization

- **Emotional Support & Counseling:** Schools should provide psychological support and safe spaces for victims to share their experiences.
- **Social Skills Training:** Teaching assertiveness and coping skills can help students respond to victimization.
- **Anti-Bullying Awareness Programs:** Educating students about the negative impact of bullying and encouraging a culture of kindness.

Intervention Strategies for Bullying/aggression

- **Parental Involvement:** Engaging parents in discussions about **bullying behavior** and encouraging positive disciplinary strategies at home.
- **School Policy Enforcement:** Ensuring strict anti-bullying policies with clear consequences for aggressive behaviors.
- **Behavioral Therapy & Conflict Resolution Training:** Teach empathy, anger management, and social problem-solving to aggressive students.

Conclusion

conclusion

The factor analysis revealed two major dimensions:

- Victimization (sv, pv, vv) – Support and protection needed.
- Bullying/Aggression (vb, sb) – Behavioral intervention needed.

References (Factor Analysis)

- Factor Analysis - Steps, Methods and Examples - Research Method.
- Factor analysis - Wikipedia

Questions?

Thank You!