Introduction to Data Analysis

Nihan Acar-Denizli

12 September 2024

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Vectors

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

$$\mathbf{x}' = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix} \qquad \mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Vectors

- Norm (length):

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{{x_1}^2 + {x_2}^2 + \ldots + {x_n}^2}$$

$$\|\alpha\mathbf{x}\| = \alpha\|\mathbf{x}\|$$

- Scalar Product:

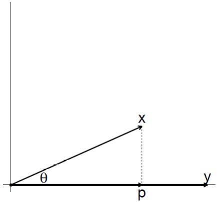
$$\mathbf{x}'\mathbf{y} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

$\mathbf{x}'\mathbf{y} = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are perpendicular.

- Angle Between Two Vectors:

$$cos\theta = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Projection



$$cos\theta = \frac{\|p\|}{\|x\|} \implies \|p\| = \frac{\mathbf{x}'\mathbf{y}}{\|y\|} \qquad (i)$$

$$\|p\| = \alpha\|y\| \implies \alpha = \frac{\|p\|}{\|y\|} \qquad (ii)$$

$$p = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}\mathbf{y}$$

Linear Combination and (In)Dependency

- Linear combination:

$$\mathbf{y} = c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots + c_p\mathbf{x_p}$$

If the equation,

$$c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots + c_n\mathbf{x_p} = 0 \tag{1}$$

  - is satisfied for the scalars $c_1, c_2, \ldots c_p$ are not all zero, the vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ are linearly dependent.
  - is satisfied only for $c_i = 0$, for all $i = 1, \ldots, p$, the vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ are linearly independent.

Matrices

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix}$$

Determinant of a Matrix

The determinant of

$$\mathbf{A}_{2\times 2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is

$$|\mathbf{A}| = ad - bc$$

- $|\mathbf{A}| = 0 \implies$ "linear dependence, $\mathbf{A}$ is singular.

For a $\mathbf{A}_{k\times k}$ matrix,

$$|\mathbf{A}_{k\times k}| = \sum_{j=1}^{k} a_{ij}(-1)^{i+j}\mathbf{A}_{ij}$$

For a $\mathbf{A}_{3\times 3}$ matrix,

$$|\mathbf{A}| = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

Inverse of a Matrix

- Inverse of a 2×2 matrix: If $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

- Inverse of a diagonal matrix:

$$\mathbf{D}^{-1} = \begin{bmatrix} 1/d_1 & 0 & \dots & 0 \\ 0 & 1/d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/d_n \end{bmatrix}$$

Partitioned Matrices

$$\mathbf{A} = \begin{bmatrix} 7 & 2 & 5 & 8 & 4 \\ -3 & 4 & 0 & 2 & 7 \\ 9 & 3 & 6 & 5 & -2 \\ 3 & 1 & 2 & 1 & 6 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where

$\mathbf{A}_{11} = \begin{bmatrix} 7 & 2 & 5 \\ -3 & 4 & 0 \end{bmatrix}$, $\mathbf{A}_{12} = \begin{bmatrix} 8 & 4 \\ 2 & 7 \end{bmatrix}$, $\mathbf{A}_{21} = \begin{bmatrix} 9 & 3 & 6 \\ 3 & 1 & 2 \end{bmatrix}$,

$\mathbf{A}_{22} = \begin{bmatrix} 5 & -2 \\ 1 & 6 \end{bmatrix}$.

Partitioned Matrices

Let $\mathbf{A}_{11} = \begin{bmatrix} 6 & -2 & 3 \\ 2 & 1 & 0 \\ 4 & 3 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 4 \\ 2 \\ -1 \end{bmatrix}$,

$$\mathbf{Ab} = 4 \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} + 2 \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix} - \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}$$

$$\mathbf{Ab} = (A_1, A_2, \ldots, A_k)(b_1 b_2 \ldots b_k)'$$

Trace and Rank of a matrix

- Trace of a matrix:

$$tr(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$$

- Rank: indicates number of linearly independent rows or columns of a matrix.
  - A matrix with a rank equals to the smallest dimension is called a full rank matrix.
  - If all the rows/columns of a matrix are linearly independent, the determinant is zero.
  - The rank of a null matrix is zero.

Quadratic Forms

Let $\mathbf{A}$ be a $n \times n$ symmetric matrix and $\mathbf{x}$ be a vector of length n. The scalar quantity $\mathbf{Q}$,

$$Q = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

is called a quadratic form.

$$Q = 9x_1^2 + 7x_2^2 + 3x_3^2 + 2x_1x_2 + 4x_1x_3 + 6x_2x_3 = \mathbf{x}'\mathbf{A}\mathbf{x}$$

$$\mathbf{A} = \begin{bmatrix} 9 & 1 & 2 \\ 1 & 7 & 3 \\ 2 & 3 & 3 \end{bmatrix}$$

Quadratic Forms

- The squared Euclidean distance from $\mathbf{x}$ to $\mathbf{y}$ can be defined as:

$$(\mathbf{x}-\mathbf{y})'\mathbf{A}(\mathbf{x}-\mathbf{y})$$

- The squared Euclidean distance from the origin in quadratic form:

$$\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x} = x_1^2 + x_2^2 = \mathbf{x}'\mathbf{I}\mathbf{x}$$

Distances

- Euclidean Distance: Straight line distance between two points.

$$d_{\mathbf{x},\mathbf{y}} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}$$

- Mahalanobis Distance:

$$d_{\mathbf{x},\mathbf{y}} = (\mathbf{x} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y})$$

- Minkowski Distance:

$$d_{\mathbf{x},\mathbf{y}} = \Big(\sum_{i=1}^{p} |x_i - y_i|^p\Big)^{1/p}$$

Spectral Decomposition

Let $\mathbf{A}$ be a symmetric matrix. Then, $\mathbf{A}$ can be defined as,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}'$$

- The columns of symmetric matrix $\mathbf{U}$ are eigenvectors of $\mathbf{A}$,
- Diagonal elements of $\mathbf{D}$ are eigenvalues of $\mathbf{A}$

$$\mathbf{A} = \left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\right] \begin{bmatrix} \lambda_1 \mathbf{u}_1' \\ \lambda_2 \mathbf{u}_2' \\ \vdots \\ \lambda_n \mathbf{u}_n' \end{bmatrix} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i'$$

Spectral Decomposition

- The spectral decomposition of $\mathbf{A}^{-1}$,

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i'$$

- Eigenvalues are sorted in descending order $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$
- $tr(\mathbf{A}) = tr(\mathbf{UDU}') = tr(\mathbf{U}'\mathbf{UD}) = tr(\mathbf{ID}) = tr(\mathbf{D}) = \sum_{i=1}^{n} \lambda_i$
- In case that $tr(\mathbf{A})$ is not full rank, there are zero eigenvalues.

Spectral Decomposition

The use of spectral decomposition in Ordinary Least Squares (OLS) Estimation :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Since $\mathbf{X}'\mathbf{X}$ is a square matrix, it can be decomposed as $\mathbf{UDU}'$. Then,

$$\mathbf{b} = (\mathbf{UDU}')^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{U}')^{-1}\mathbf{D}^{-1}\mathbf{U}^{-1}\mathbf{X}'\mathbf{y}$$

Singular Value Decomposition

A rectangular matrix $\mathbf{A}_{n \times p}$ can be decomposed as,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where

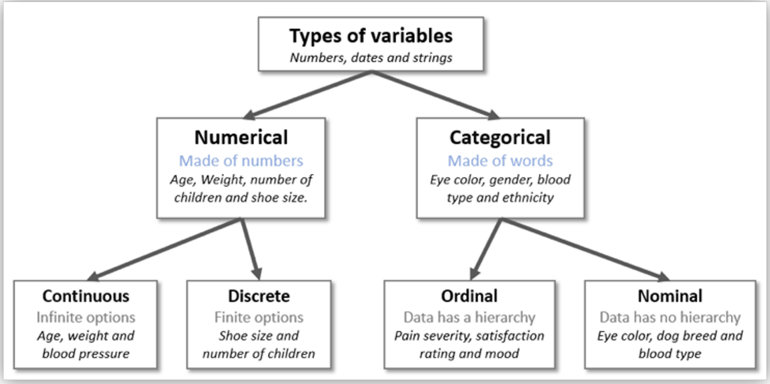
- $\mathbf{U}$ is $n \times k$ orthogonal matrix of left eigenvectors $(\mathbf{U}\mathbf{U}' = \mathbf{I}_k)$
- $\mathbf{D}$ is $k \times k$ diagonal matrix consists of singular values of matrix $\mathbf{A}$ $(d_1 \geq d_2 \geq d_3 \geq \ldots \geq d_k)$
- $\mathbf{V}$ is $p \times k$ orthogonal matrix of right singular vectors $(\mathbf{V}\mathbf{V}' = \mathbf{I}_k)$

$$\mathbf{A} = \sum_{i=1}^{k} d_{ii}\mathbf{u}_i\mathbf{v}_i' = d_1\mathbf{u}_1\mathbf{v}_1' + d_2\mathbf{u}_2\mathbf{v}_2' + \ldots + d_k\mathbf{u}_k\mathbf{v}_k'$$

Singular Value Decomposition

- $\mathbf{A}'\mathbf{A} = \mathbf{VDU}'\mathbf{UDV}' = \mathbf{VD}^2\mathbf{V}'$
- $\mathbf{AA}' = \mathbf{UDV}'\mathbf{VDU}' = \mathbf{UD}^2\mathbf{U}'$
- Eigenvalues of $\mathbf{A}'\mathbf{A}$ and $\mathbf{AA}'$ are squared singular values
- Singular vectors are eigenvectors.
- SVD Song

Types of Variables



Source:https://www.slideshare.net/slideshow/data-distribution-the-probability-distributions

Measurement Scales

		Nominal	Ordinal	Interval	Ratio
Number Meaning		Categories	Order	Equal intervals between characteristic	Equal intervals with true zero point
Arithmetic Operations	Inequality	x	x	x	x
	Ordering / Ranking		x	x	x
	Addition / Subtraction			x	x
	Multiplication / Division				x
Descriptive Statistics	Mode	x	x	x	x
	Median		x	x	x
	Mean			x	x
	Standard Deviation			x	x
Statistical Analysis Techniques Commonly Used	Crosstabs / Chi-Square	x	x		
	Rank Order Correlation		x		
	Analysis of Variance (NP)	x	x		
	Correlation			x	x
	Regression			x	x
	Analysis of Variance			x	x
	Factor Analysis			x	x

BUSINESS BROADWAY

Descriptive Statistics

Suppose $\mathbf{X}' = [\mathbf{X_1}, \mathbf{X_2}, \dots, \mathbf{X_p}]$ where each element of $\mathbf{X}$ is a random variable with a marginal probability distribution with $(j = 1, ..p)$

- Marginal Mean:

$$\mu_j = E[\mathbf{X}_j] = \begin{cases} \sum_j p_j x_j, & \mathsf{x}_j \quad \text{is a discrete variable,} \\ \int x_j f_j(x_j) dx_j, & \mathsf{x}_j \quad \text{is a continuous variable} \end{cases}$$

- Marginal variance:

$$\sigma_j^2 = E[\mathbf{X}_j - \mu_j]^2 = \begin{cases} \sum_j (x_j - \mu_j)^2 p_j(x_j), & x_j \text{ is a discrete variable,} \\ \int (x_j - \mu_j)^2 f_j(x_j) dx_j, & x_j \text{ is a continuous variable} \end{cases}$$

Descriptive Statistics

The association between two random variables $\mathbf{X}_i$ and $\mathbf{X}_k$ is described by joint probability function:

- Covariance:

$$\sigma_{ik} = E[(\mathbf{X}_i - \mu_i)(\mathbf{X}_k - \mu_k)] = \begin{cases} \sum \sum (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) \\ \int \int (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k) dx_i dx_k \end{cases}$$

Population Covariance Matrix

- Population covariance matrix:

$$\sigma = E(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)'$$
$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}$$

Sample Statistics in Matrix Notation

- Sample Mean Vector:

$$\bar{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$$

- Centered matrix:

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}\bar{\mathbf{X}}' = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$$

- Standardized matrix:

$$\mathbf{X}s = \mathbf{X}_c * \mathbf{D}_s^{-1} \quad \text{where} \quad \mathbf{D}_s = diag(s_1, s_2, \ldots, s_p)$$

Sample Covariance Matrix

- Sample covariance matrix:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$$

The association between variable j and k is computed from,

$$s_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j = 1, \dots, p, \quad k = 1, \dots, p$$

Sample Correlation Matrix

- Sample correlation matrix:

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1} = \frac{1}{n-1}\mathbf{X}_s'\mathbf{X}_s$$

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}$$

The correlation coefficient between two variables j and k ($j = 1, \dots, p$, $k = 1, \dots, p$) is found by,

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

Data Preprocessing

1. Data Quality Assesment
   - Mismatched data types
   - Mixed data values
   - Data outliers
   - Missing data
2. Data Cleaning (imputation of missing data, removing irrelevant or incorrect data)
3. Data transformation
4. Data reduction

Data Preprocessing

- Avoiding duplications
- Checking existence of zeros
- Checking existence of outliers
- Checking existence of missing values
- Applying transformations if it is needed.

Missing Values

- Check how the missing values are coded (NA, 99, -9, " ", etc.).
- Determine what percent of the data is missing.
- Do missing values concentrate in some variables or individuals?

Missing Values

- Missing completely at random (MCAR): The probability of missingness is the same for all units.
- Missing at random (MAR): The probability of missingness in a variable depends on other available factors.
- Missing not at random (MNAR)
  - Missingness depends on unobserved predictors (information that has not been recorded)
  - Missingness depends on the variable/missing value itself

Missing Values

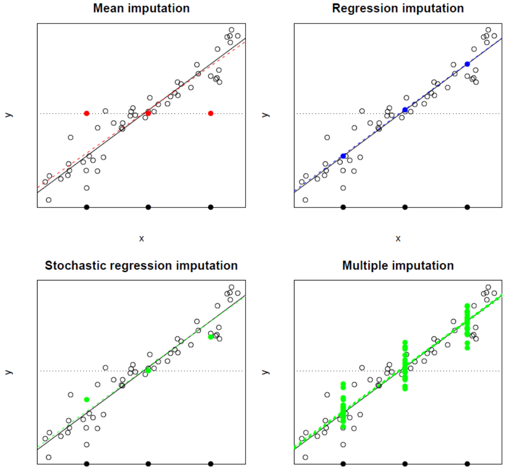
How to deal with missing values?

- Delete missing values.
  - The sample size is reduced and the analysis losses power.
  - Statistical inference can be biased if the missing observations are not MCAR.
- Impute missing values with "a reasonable value" (single imputation)
- Impute missing values many time and do the analysis for each imputed data set (multiple imputation)

Missing Value Imputation

- Mean imputation: Replacing each missing value in a variable with the mean of the observed values for that variable.
- Regression imputation: The missing values of a variable is predicted by using a regression model on other variables.
- Stochastic (random) regression imputation: A normally distributed error term is added to the predicted values.
- Multiple imputation: Missing values are imputed multiple times by using an appropriate model. (Based on averaging the values of parameter estimates to find a single point estimate.)

Missing Value Imputation



Source: Lecture Notes of Jan Graffelman (2021)

How to Deal with Outliers?

- Removing outliers (if the outlier is not part of the studied population)
- Imputing a new value (if the outlier is arised from a mistake in data collection or measurement process)
- Transforming data (if the transformation removes outlier and changes skewness of data, specially in regression)
- Non-parametric statistical analysis (analysis does not require a certain distribution)

Example: Anscombe's Quartet Data

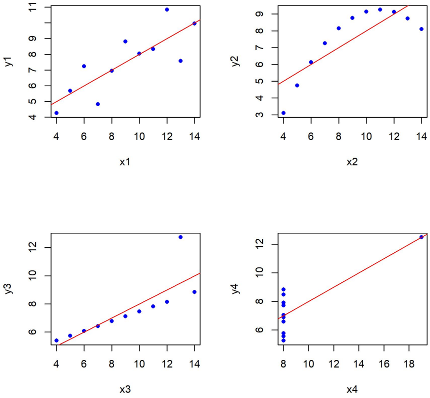


Figure: Anscombe's Quartet Data (Anscombe, F., 1973)

Transformations

Transformations are applied on variables in order to

- reduce effect of outlying observations
- provide homoscedasticity
- obtain a normally distributed variable

Transformations

Main types of transformations:

- Logarithmic transformation (for natural logarithm: $y = ln(x)$)
  to discard or to reduce skewness
- Square root transformation ($y = \sqrt{x}$)
  to transform count or percentage data, can be applied to zero values.
- Logit transformation ($logity = log(p/1 - p)$)
  To fit logistic regression models
- Reciprocal transformation ($y = 1/x$)
  to change shape of the distribution , only for non-zero values

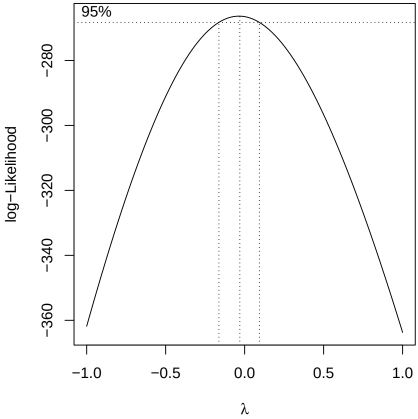
Box-Cox Transformation

- Box-Cox transformations consist of a family of power transformations

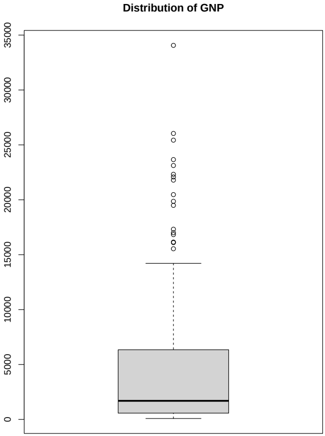
$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda}, & (\lambda \neq 0) \\ ln(Y_i), & (\lambda = 0) \end{cases} \tag{2}$$

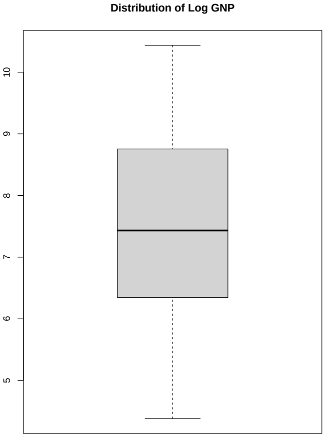
- Generally used to transform non-normal variables to normal variables.
- The optimum λ is found by maximizing log-likelihood function.
- $\lambda = 0$ is equivalent to logarithmic transformation
- $\lambda = 1/2$ is equivalent to square root transformation

Example: Box-Cox transformation



Example: Box-Cox transformation





Example: Logarithmic Transformation

Estimation of Life Expectancy of the Countries based on Their Income per person for Year 2021

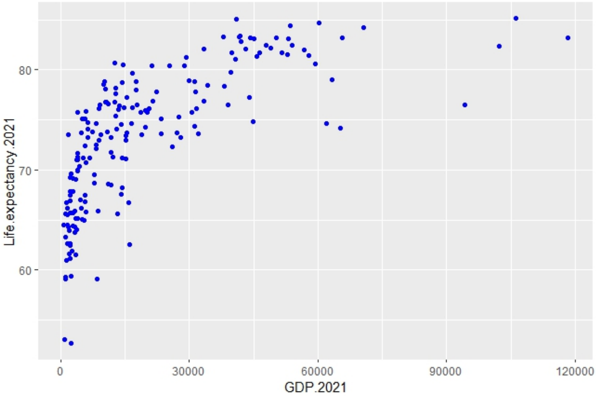


Figure: Income per person vs. Life Expectancy for 2021

Example: Logarithmic Transformation

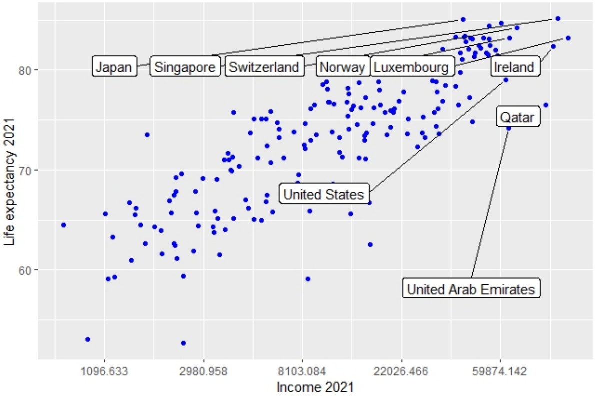
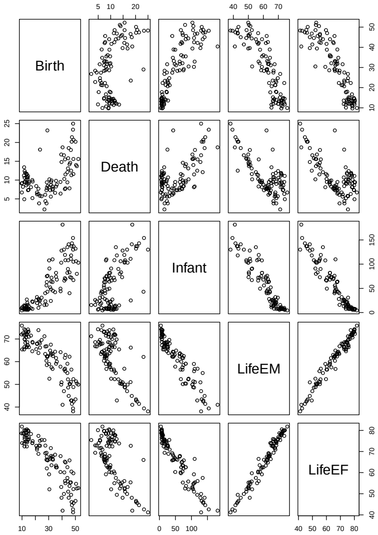


Figure: Income per person (transformed) vs. Life Expectancy for 2021

Multivariate Data Visualization

- To visualize quantitative variables:
  - Scatterplot matrix
  - Chernoff Faces
  - Star Plots
  - Biplots
- To visualize qualitative variables:
  - Joint Bar Charts
  - Stratified Bar Charts
  - Mozaic plots
  - Biplots
- To visualize a qualitative and a quantitative variable:
  - Box Plots

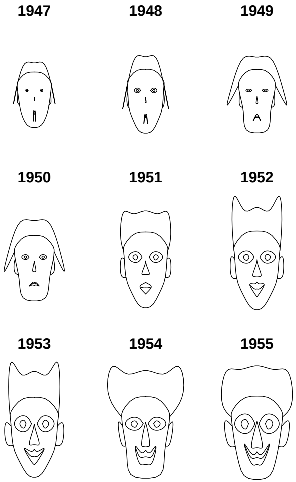
Scatterplot Matrix



Chernoff faces for Longley Data

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year	Employed
1947	83.0	234.289	235.6	159.0	107.608	1947	60.323
1949	88.2	258.054	368.2	161.6	109.773	1949	60.171
1948	88.5	259.426	232.5	145.6	108.632	1948	61.122
1950	89.5	284.599	335.1	165.0	110.929	1950	61.187
1951	96.2	328.975	209.9	309.9	112.075	1951	63.221
1952	98.1	346.999	193.2	359.4	113.270	1952	63.639
1953	99.0	365.385	187.0	354.7	115.094	1953	64.989
1954	100.0	363.112	357.8	335.0	116.219	1954	63.761
1955	101.2	397.469	290.4	304.8	117.388	1955	66.019
1956	104.6	419.180	282.2	285.7	118.734	1956	67.857
1957	108.4	442.769	293.6	279.8	120.445	1957	68.169
1958	110.8	444.546	468.1	263.7	121.950	1958	66.513
1959	112.6	482.704	381.3	255.2	123.366	1959	68.655
1960	114.2	502.601	393.1	251.4	125.368	1960	69.564
1961	115.7	518.173	480.6	257.2	127.852	1961	69.331
1962	116.9	554.894	400.7	282.7	130.081	1962	70.551

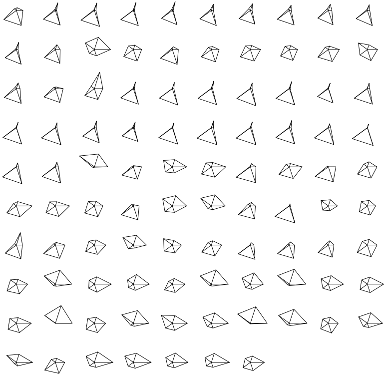
Chernoff Faces



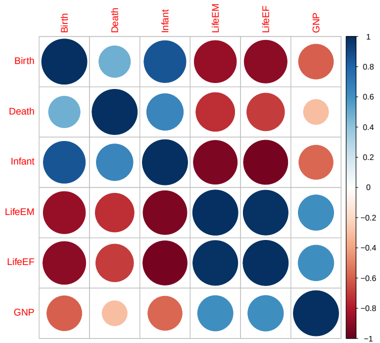

```
> data(longley)
> faces(longley[1:9,],face.type=0)
effect of variables:
 modified item       Var
 "height of face   " "GNP.deflator"
 "width of face    " "GNP"
 "structure of face" "Unemployed"
 "height of mouth  " "Armed.Forces"
 "width of mouth   " "Population"
 "smiling          " "Year"
 "height of eyes   " "Employed"
 "width of eyes    " "GNP.deflator"
 "height of hair   " "GNP"
 "width of hair   "  "Unemployed"
 "style of hair   "  "Armed.Forces"
 "height of nose  "  "Population"
 "width of nose   "  "Year"
 "width of ear    "  "Employed"
 "height of ear   "  "GNP.deflator"
```

Star Plots

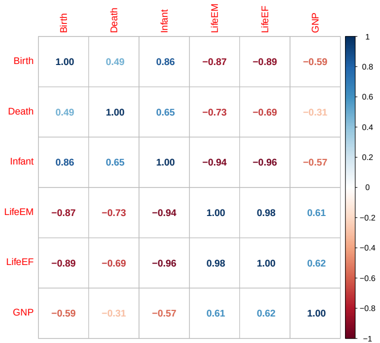
	Birth	Death	Infant	LifeEM	LifeEF	GNP	Country
1	24.7	5.7	30.8	69.6	75.50	600	Albania
2	12.5	11.9	14.4	68.3	74.70	2250	Bulgaria
3	13.4	11.7	11.3	71.8	77.70	2980	Czechoslovakia
4	12.0	12.4	7.6	69.8	75.90	1690	Former_E._Germa
5	11.6	13.4	14.8	65.4	73.80	2780	Hungary
6	14.3	10.2	16.0	67.2	75.70	1690	Poland
7	13.6	10.7	26.9	66.5	72.40	1640	Romania
8	14.0	9.0	20.2	68.6	74.50	1690	Yugoslavia
9	17.7	10.0	23.0	64.6	74.00	2242	USSR
10	15.2	9.5	13.1	66.4	75.90	1880	Byelorussian_SSR
11	13.4	11.6	13.0	66.4	74.80	1320	Ukrainian_SSR
12	20.7	8.4	25.7	65.5	72.70	2370	Argentina
13	46.6	18.0	111.0	51.0	55.40	630	Bolivia
14	28.6	7.9	63.0	62.3	67.60	2680	Brazil
15	23.4	5.8	17.1	68.1	75.10	1940	Chile
16	27.4	6.1	40.0	63.4	69.20	1260	Columbia
17	32.9	7.4	63.0	63.4	67.60	980	Ecuador
18	28.3	7.3	56.0	60.4	66.10	330	Guyana
19	34.8	6.6	42.0	64.4	68.50	1110	Paraguay
20	32.9	8.3	109.9	56.8	66.50	1160	Peru
21	18.0	9.6	21.9	68.4	74.90	2560	Uruguay
22	27.5	4.4	23.3	66.7	72.80	2560	Venezuela
23	29.0	23.2	43.0	62.1	66.00	2490	Mexico



Corplots

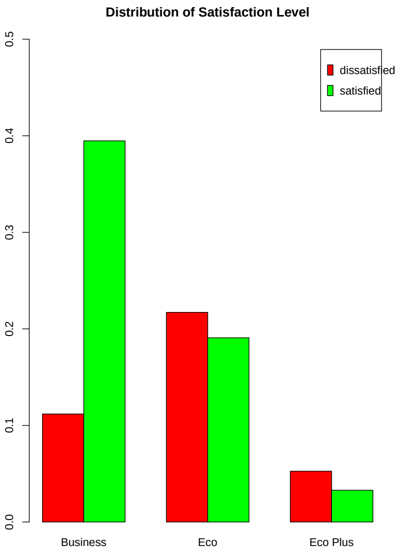


Birth
Death
Infant
LifeEM
LifeEF
GNP
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

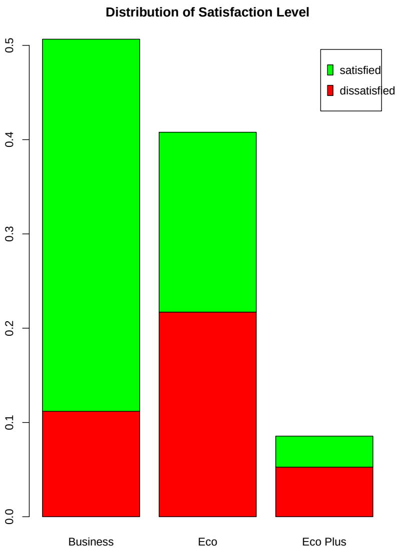


Birth
Death
Infant
LifeEM
LifeEF
GNP
Birth 1.00 0.49 0.86 −0.87 −0.89 −0.59
Death 0.49 1.00 0.65 −0.73 −0.69 −0.31
Infant 0.86 0.65 1.00 −0.94 −0.96 −0.57
LifeEM −0.87 −0.73 −0.94 1.00 0.98 0.61
LifeEF −0.89 −0.69 −0.96 0.98 1.00 0.62
GNP −0.59 −0.31 −0.57 0.61 0.62 1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

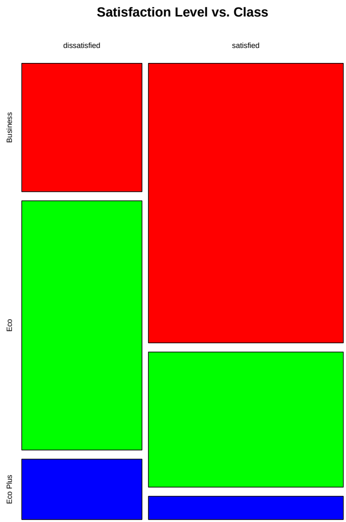
Joint Bar Chart



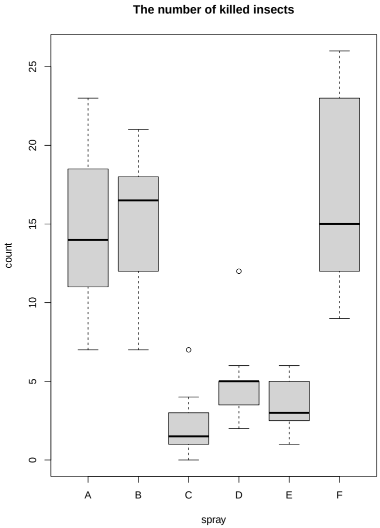
Stacked Bar Chart



Mosaic Plots



Box Plots


The number of killed insects
count
0
5
10
15
20
25
A
B
C
D
E
F
spray

REFERENCES I

- Manly, B.F.J (1989). Multivariate statistical methods: a primer. 3rd edition. Chapman and Hall, London.
- Johnson and Wichern (2002). Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall.
- Peña, D. (2002). Análisis de datos multivariantes, McGraw Hill.
- Rencher,A.C. & Schaalje, G.B. (2007). Linear Models in Statistics, Wiley.
- Missing Data Imputation. `http://www.stat.columbia.edu/~gelman/arm/missing.pdf`, Last Access: 10 March 2022
- `https://mathformachines.com/posts/eigenvalues-and-singular-values/`

REFERENCES II

- https://medium.com/mlearning-ai/4-easy-ways-to-handle-outliers-in-your-data-47f125a3f779
- https://r-graph-gallery.com/stacked-barplot.html
- Grafelman, J. (2021). Lecture Notes of Data Analysis (Bachelor in Data Science,FIB,UPC).
- Anscombe, F. (1973). Graphs in Statistical Analysis, American Statistician, 27(1),17-21.
- https://www.gapminder.org/answers/how-does-income-relate-to-life-expectancy/, Last Access: 18 November 2022

THANK YOU FOR YOUR ATTENTION!