Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Logistic Regression

- The general linear model assumes continuous response variables
- For binary (or categorical) responses, the model needs to be generalized
- Instead of the observed values, their probability of occurrence (their expected value) is modeled

$$\pi_i = P(Y_i = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta}) = F(\eta_i)$$

- With $F = \frac{\exp(\eta)}{1+\exp(\eta)}$ (*logistic function*), we obtain a model for logistic regression

$$\pi_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})} = \frac{1}{1 + \exp[-(\mathbf{x}_i'\boldsymbol{\beta})]}$$

- The choice of the function F ensures that the model's prediction always lies between 0 and 1

Predictors

- As in multiple regression, logistic regression may include metric and categorical predictors
- Categorical predictors need to be coded appropriately
- As in the general linear model, we have
  - Dummy coding: Interpretation w.r.t. the reference category
  - Effect coding: Interpretation w.r.t. the total mean
- Example: Consider one metric and one 3-level categorical predictor

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 0 & 1 \\ 1 & x_{31} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & 0 & 0 \end{pmatrix}$$

Probability, Odds, Logit

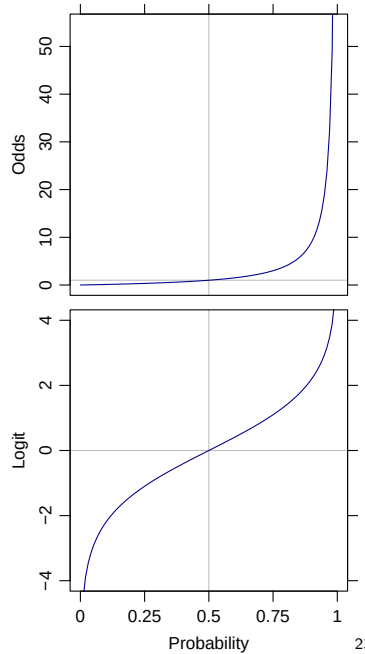
- odds $= \dfrac{\pi}{1-\pi}$

  expected count of events per complementary event

- logit $= \log \dfrac{\pi}{1-\pi}$

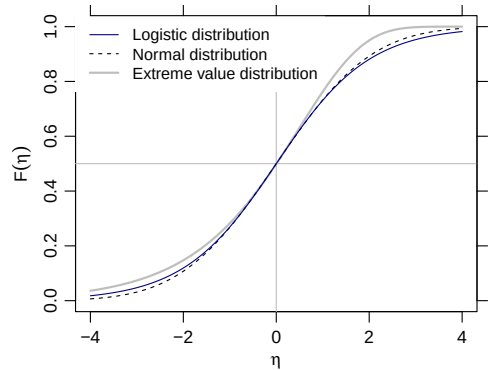
  log-odds, inverse of logistic function

π	0.01	0.05	0.33	0.50	0.66	0.95	0.99
odds	$1/99$	$1/19$	0.49	1	1.94	19	99
logit π	-4.60	-2.94	-0.71	0	0.66	2.94	4.60



Response Function and Link Function

- The *response function* maps the linear predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ to a value in the $[0, 1]$ interval
- Alternatives to the logistic distribution are less common
- The inverse of the response function is called *link function*
- In logistic regression we have



$$\operatorname{logit} \pi_i = \log \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- As the logit function serves as link function (i. e., $\operatorname{logit}(x) = F^{-1}(x)$), this model is called a *logit model*

Interpretation of Parameters

- The odds ratio expresses how much the odds change between two conditions
- The parameters in the logit model can be seen as logarithmic odds ratios

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0) \cdot \exp(\beta_1 X_{i1}) \cdot \exp(\beta_2 X_{i2}) \cdot \dots \cdot \exp(\beta_p X_{ip})$$

- Metric predictors
  - The odds ratio when increasing the predictor value X_{ij} by one unit is $\exp(\beta_j)$
  - The odds of the event change by a factor of $\exp(\beta_j)$
- Categorical predictor with dummy coding
  - The odds ratio to the reference category is $\exp(\beta_j)$
  - The odds of the event change by a factor of $\exp(\beta_j)$

Example: Interpretation of Parameters

- For elementary school students, the probability of solving a math problem depends on the metric predictor age X_1 and the categorical predictor sex X_2 (dummy coded, boys forming the reference class)

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0) \cdot \exp(\beta_1 X_{i1}) \cdot \exp(\beta_2 X_{i2})$$

- Age
  - The factor $\exp(\beta_1)$ describes how the odds of solving the problem change by increasing the age by 1 year
- Sex
  - The factor $\exp(\beta_2)$ describes how the odds of solving the problem changes for girls relative to the boys

Example: Preference for Women's Magazines

- In a study on "Informationsgewohnheiten von Frauen" (women's information habits), the influence of socio-demographic factors on the preference to purchase a particular magazine M was tested on a sample of $N = 941$ women

	Working (W)	yes		no	
	Magazine (M)	yes	no	yes	no
Age (A)	Education (E)				
18–29	lower secondary school, no apprenticeship	1	14	3	24
	lower secondary school with apprenticeship	32	49	12	41
	secondary school	20	34	19	20
	Abitur/university	8	3	14	13
30–39	lower secondary school, no apprenticeship	9	23	1	37
	lower secondary school with apprenticeship	31	57	12	68
	secondary school	11	26	14	43
	Abitur/university	5	7	4	7
40–49	lower secondary school, no apprenticeship	1	33	11	54
	lower secondary school with apprenticeship	12	50	14	53
	secondary school	5	11	8	15
	Abitur/university	1	7	1	3

Example: Preference for Women's Magazines

- When all predictors are categorical (grouped data), we can represent the data in a multi-dimensional contingency table, which would not be possible with metric (ungrouped) predictors
- The logit model

$$\begin{aligned}\text{logit}\,\pi = \beta_0 &+ \beta_1 W + \beta_2 A2 + \beta_3 A3 + \beta_4 E2 + \beta_5 E3 + \beta_6 E4 \\ &+ \beta_7 WA2 + \beta_8 WA3\end{aligned}$$

  was chosen, with dummy coded indicator variables, and the last category forming the reference class
- The model includes main effects of working (W), age (A) and eduction (E), and the interaction between W and A

Example: Preference for Women's Magazines

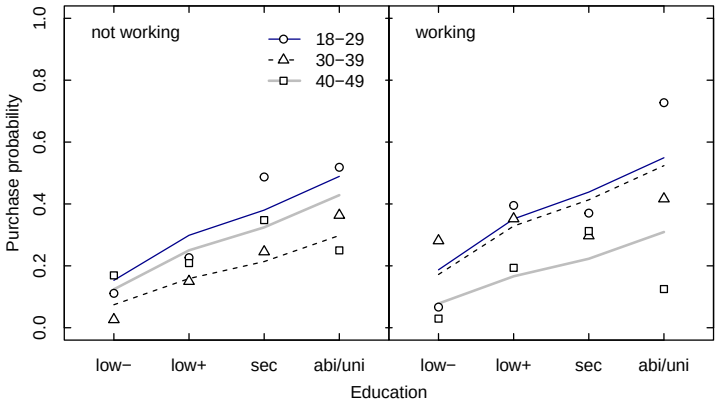
- The following parameter estimates were obtained

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−1.71	0.28	−6.21	.00
W	0.24	0.25	0.97	.33
A2	−0.81	0.27	−3.00	.00
A3	−0.24	0.27	−0.89	.37
E2	0.86	0.24	3.54	.00
E3	1.22	0.26	4.67	.00
E4	1.67	0.33	5.10	.00
WA2	0.71	0.36	1.97	.05
WA3	−0.76	0.41	−1.85	.06

- Predictors without interaction can be interpreted independently
  - Compared to lower secondary school without apprenticeship, the odds of buying the magazine for Abitur/university are $\exp(\hat{\beta}_{E4}) = 5.29$ times higher
  - The preference for buying the magazine increases with education level

Example: Preference for Women's Magazines

- Predictors with interaction need to be interpreted jointly
- Transition from "not working" to "working" in
  - A1: $\hat{\beta}_W = 0.24$, probability increases
  - A2: $\hat{\beta}_W + \hat{\beta}_{WA2} = 0.95$, probability increases
  - A3: $\hat{\beta}_W + \hat{\beta}_{WA3} = -0.51$, probability decreases



Parameter Estimation

- The parameters in the logit model (and comparable models) are estimated using maximum likelihood estimation (MLE)
- When there is no analytical solution, then parameter estimation needs to proceed by iterative methods
- Often the *Newton-Raphson method* is used
  - It aims at finding the zeroes of a (continuously differentiable) non-linear function
  - The basic idea is to approximate the function by its tangent line (differential of the function)

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0)$$

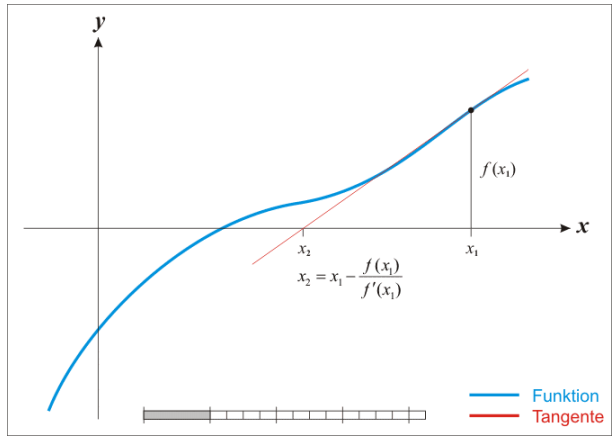
  - Starting from an initial value, zeroes are determined iteratively

Parameter Estimation

- Newton-Raphson method

General form of the iteration for a function f of one variable

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$



Prediction

- The probability predicted by a logit model is computed in two steps
  1. Calculate the value of the linear predictor $\hat{\eta}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ based on the predictor vector $\mathbf{x}_0$ and the parameter estimates $\hat{\boldsymbol{\beta}}$
  2. Input the linear predictor into the response function to obtain the probability $\hat{\pi}_0 = F(\hat{\eta}_0)$
- Example
  - For a working 30-years old ($A2$) in the lowest education category ($E1$) we obtain

$$\begin{aligned}\hat{\eta}_0 &= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_7 \\ &= -1.71 + 0.24 - 0.81 + 0.71 = -1.57\end{aligned}$$

  - The probability for buying magazine M then is

$$\hat{\pi}_0 = (1 + \exp(-1.57))^{-1} = 0.17$$

Likelihood-Ratio Test

- Two nested models can be compared by means of their likelihoods
  - Model M_0 is nested in M_1 if the former results from restricting parameters in the latter
- The test statistic

$$G^2 = 2 \log \frac{L_{M_1}}{L_{M_0}}$$

  is based on the likelihood ratio and is approximately χ^2 distributed with

$$(\text{number of parameters in } M_1) - (\text{number of parameters in } M_0)$$

  degrees of freedom
- The null hypothesis that M_0 describes the data as good as M_1 is rejected if the test is significant

Overall Test

- We want to test if in a model M with parameters $\beta_0, \beta_1, \ldots, \beta_p$ at least one of the predictors contributes to predicting the response variable
- Here, the constrained model M_0 contains the parameter β_0 only
- We test the null hypothesis

$$H_0\colon\ \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{or} \quad H_0\colon\ \eta_j = \beta_0 \quad (\eta_j = \mathbf{x}_j'\boldsymbol{\beta})$$

- Under the null hypothesis, the test statistic

$$G^2 = 2 \log \frac{L_M}{L_{M_0}}$$

  is approximately χ^2 distributed with p degrees of freedom
- H_0 is rejected if $G^2 > \chi^2_{1-\alpha}(p)$

Deviance

- In general, we call a model *saturated* if it contains as many independent parameters as observations (or observation categories)
- The *saturated model* is used for goodness-of-fit tests of constrained models M
- The deviation of a model M from the saturated model M_{sat} is described by

$$G^2 = 2 \log \frac{L_{M_{sat}}}{L_M}$$

  which is called the *deviance*
- In the case of grouped data the deviance is approximately χ^2 distributed

Goodness-of-Fit Test

- Let I denote the number of observed categories or conditions in grouped data and

$$Y_i \text{ the number of events in condition } i$$
$$n_i \text{ the number of observations in condition } i$$

- Goodness-of-fit test of model M
  - Likelihood-ratio statistic (deviance)

$$G^2 = 2\sum_{i=1}^{I}\left[y_i \log \frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right]$$

    with $\hat{\pi}_i$ the probabilities predicted by M
  - Pearson chi-squared statistic

$$X^2 = \sum_{i=1}^{I}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

Goodness-of-Fit Test

- Both test statistics G^2 and X^2 are approximately χ^2 distributed with $I - p - 1$ degrees of freedom
- If significant, then we reject the null hypothesis that the logit model M describes the data adequately (or rather that M is among the potential data generators)

Testing Individual Coefficients (Wald Test)

- Null hypothesis: $H_0 \colon \beta_j = 0$
- Test statistic

$$T = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \quad \left(\text{or } T^2 = \frac{\hat{\beta}_j^2}{\widehat{Var}(\hat{\beta}_j)} \right)$$

  where $\widehat{Var}(\hat{\beta}_j)$ is the respective diagonal element of the estimated covariance matrix $\widehat{Cov}(\hat{\boldsymbol{\beta}})$
- Under the null hypothesis, the test statistic is approximately $N(0, 1)$ distributed (or approximately χ^2 distributed with $df = 1$)
- H_0 is rejected if

$$|T| > z_{1-\alpha/2} \quad \left(\text{or } T^2 > \chi^2_{1-\alpha}(1) \right)$$

Testing Linear Hypotheses (Wald Test)

- Arbitrary linear hypotheses regarding the parameter vector $\boldsymbol{\beta}$ can be tested using another Wald-like statistic
- In analogy to the general linear model, the null hypothesis is written as

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$$

- where $\mathbf{C}$ is an appropriate contrast matrix
- Under the null hypothesis, the test statistic

$$W^2 = (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})' \left[\mathbf{C}\widehat{Cov}(\hat{\boldsymbol{\beta}})\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi})$$

is approximately $\chi^2(r)$ distributed with degrees of freedom $r = rk(C)$ (usually its number of rows)

Generalized Linear Models

- A generalized linear model is defined by

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

  with the link function g linking the expected value of the response variable to the linear combination of the predictors
- We assume that the response variable Y is independent (independent cases) and follows a distribution from the exponential family
- Every target distribution allows for a selection of link functions
  - This makes the generalized linear model a flexible model class for analyzing a variety of data types

Exponential Family

- Overview of target distributions and link functions

distribution	link function
Binomial	logit, probit, log, cloglog
Normal	identity, log, inverse
Gamma	identity, inverse, log
Inverse normal	$1/\mu^2$, identity, inverse, log
Poisson	identity, log, sqrt
Quasi	logit, probit, cloglog, identity, inverse, log, $1/\mu^2$, sqrt

- A generalized linear model is formed by a special combination of target distribution, an appropriate link function, and a linear combination of predictors

Example: Preference for Women's Magazines

Data

```
dat <- read.table(header = TRUE, text = "
  yes no      W     A       E
    1 14   work young    low-
   32 49   work young    low+
   20 34   work young     sec
  ...
")
```

Reorder factor levels

```
dat$W <- factor(dat$W, levels = c("nowork", "work"))
dat$A <- factor(dat$A, levels = c("young", "mid", "old"))
dat$E <- factor(dat$E, levels = c("low-", "low+", "sec", "abi/uni"))
```

Example: Preference for Women's Magazines

Estimate logit models

```
m1 <- glm(cbind(yes, no) ~ 1,         binomial, dat)  # no effects
m2 <- glm(cbind(yes, no) ~ W + A + E, binomial, dat)  # main effects
m3 <- glm(cbind(yes, no) ~ W*A + E,   binomial, dat)  # interaction
m4 <- glm(cbind(yes, no) ~ W*A*E,     binomial, dat)  # saturated
```

Likelihood-ratio test

```
anova(m1, m2, m3, m4)
```

Parameter estimation, individual Wald tests

```
summary(m3)
```

Prediction

```
predict(m3, data.frame(W = "work", A = "young", E = "abi/uni"),
        type = "response")
predict(m3, data.frame(W = "nowork", A = "mid", E = "low+"),
        type = "response")
```

References

Agresti, A. (2002). Categorical data analysis. New York: Wiley.

Agresti, A. (2007). An introduction to categorical data analysis. New York: Wiley.

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). Multivariate statistische Verfahren. Berlin: de Gruyter. Kapitel 6

Faraway, J. J. (2006). Extending the linear model with R. Boca Raton, FL: Chapman & Hall/CRC.

Wickens, T. D. (1989). Multiway contingency tables analysis for the social sciences. Hillsdale: Lawrence Erlbaum.

Principal Component Analysis

Objectives

- Principal component analysis (PCA) was invented by Pearson (1901) and later independently developed by Hotelling (1933)
- Its goal is to reduce a larger number of p intercorrelated variables to a few, which account for as much of the total variance as possible
- The observed variables are linearly transformed to new variables, the uncorrelated principal components, arranged in decreasing order of explained variance

Principal Component and Factor Analysis

Objectives

- Principal component and factor analysis have similar goals but are not the same
  - Principal component analysis is essentially a descriptive procedure, with no explicit model (and respective distributional assumptions) behind
  - Factor analysis is based on an explicit model, which introduces latent variables (*factors*) that try to explain the relationships between observed variables
  - Explicit modeling of latent variables is developed further in linear structure equation models

Data

- Let there be n observations of the random variables $X_1, \ldots, X_p$ (n dimensional column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$), combined to form data matrix $\mathbf{X}$

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)' = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}$$

- Alternative starting points for the principal component analysis are
  - The covariance matrix $\mathbf{S}$ of $\mathbf{X}$
  - The correlation matrix $\mathbf{R}$ of $\mathbf{X}$ (that is the covariance matrix of the standardized variables collected in the matrix $\tilde{\mathbf{X}}$)

Covariance and Correlation Matrix

- Since the results of PCA are not invariant to linear transformations of the variables, different results are obtained for **S** vs. **R**
- Standardizing variables, or considering the correlation matrix, which is often justified by the argument that psychological measurements come without unit, obscures this problem
- Nevertheless, various textbooks recommend using the correlation matrix (but see Stevens, 2001, p. 388)

Principal Components

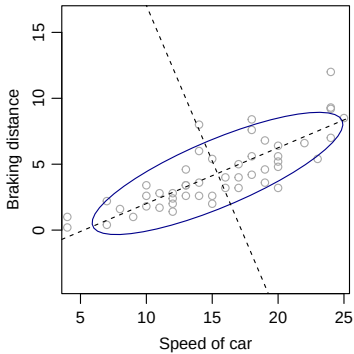
- The *principal components* $Y_1, \ldots, Y_p$ result from linear combinations of the observed variables $X_1, \ldots, X_p$
- The variables $Y_1, \ldots, Y_p$ are uncorrelated and successively explain as much of the remaining variance as possible
- Thus, we are looking for an orthogonal transformation matrix $\mathbf{G}$ (i. e., a square matrix with $\mathbf{G}^{-1} = \mathbf{G}'$) such that

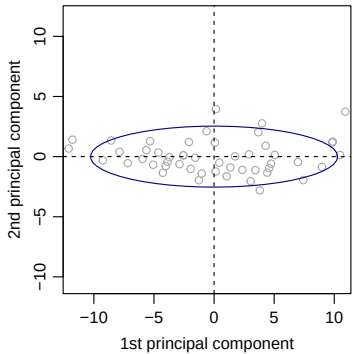
$$\mathbf{Y} = \mathbf{XG}$$

- Calculation of the principal components is based on the eigenvalues of the covariance matrix $\mathbf{S}$ (having full rank p) and its normalized eigenvectors

Principal Axis Transformation

- In two dimensions, the covariance is represented by an ellipse that is characterized by the location of the centroid and the length and direction of its orthogonal *principal axes*

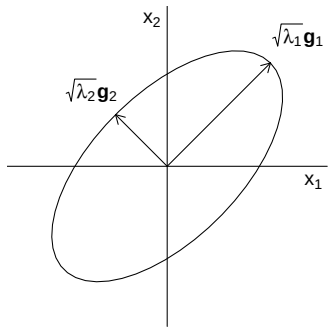




- The principal axes can be conceived as constituting a new coordinate system for the observations

Principal Axes

- The principal axes of the ellipse corresponding to the covariance or correlation matrix are determined by solving an *eigenvalue problem*
  - The lengths of the principal axes are proportional to the square root of the respective eigenvalues
  - The direction of the principal components is defined by the respective eigenvectors



Principal Component Analysis

- We consider the eigenvalues (in decreasing order)

$$\lambda_1, \ldots, \lambda_p$$

of $\mathbf{S}$, and the respective normalized eigenvectors

$$\mathbf{g}_1, \ldots, \mathbf{g}_p,$$

- For the normalized vectors $\mathbf{g}_j$ $(j = 1, \ldots, p)$ we have

$$\mathbf{S}\,\mathbf{g}_j = \lambda_j\,\mathbf{g}_j$$

Principal Components

- With the eigenvector
$$\mathbf{g}_1 = (g_{11}, g_{12}, \ldots, g_{1p})'$$
corresponding to the largest eigenvalue λ_1 we obtain the first principal component by
$$Y_1 = g_{11} \cdot X_1 + g_{12} \cdot X_2 + \ldots + g_{1p} \cdot X_p$$
- Here, the X_j are usually conceived as standardized variables
- Because λ_1 can be interpreted as the variance of Y_1, the variable Y_1 has the largest variance of all principal components

Principal Components

- With the eigenvector
$$\mathbf{g}_2 = (g_{21}, g_{22}, \ldots, g_{2p})'$$
corresponding to the second largest eigenvalue λ_2 we obtain the second principal component by
$$Y_2 = g_{21} \cdot X_1 + g_{22} \cdot X_2 + \ldots + g_{2p} \cdot X_p$$
- The variable Y_2 is uncorrelated to Y_1 since eigenvectors to different eigenvalues are orthogonal, and Y_2 has the largest variance of all variables uncorrelated to Y_1
- Accordingly, the $(j+1)$th principal component Y_{j+1} $(j = 1, \ldots, p-1)$ has the largest variance of all (normalized) linear combinations that are uncorrelated to $Y_1, \ldots, Y_j$

Principal Components Transformation

- The normalized eigenvectors can be combined to a matrix

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_1 & \dots & \mathbf{g}_p \end{pmatrix}$$

- Then the vector

$$\mathbf{y} = \mathbf{G}'\mathbf{x}$$

  corresponding to the data point $\mathbf{x}$ can geometrically be interpreted as the coordinate vector regarding the coordinate system where the axes are constituted by the eigenvectors $\mathbf{g}_1, \dots, \mathbf{g}_p$ (principal components)
- The result of the principal component transformation is $\mathbf{Y} = \mathbf{XG}$
- The covariance matrix of $\mathbf{Y}$ is the $(p \times p)$ diagonal matrix

$$\mathbf{\Lambda} = \mathbf{G}'\mathbf{S}\mathbf{G},$$

  with the eigenvalues $\lambda_1, \dots, \lambda_p$ as diagonal elements

Principal Components Transformation

- Computations may also be based on the correlation matrix $\mathbf{R}$, which is the covariance matrix of the standardized variables
- For this, replace the data matrix $\mathbf{X}$ by the matrix $\tilde{\mathbf{X}}$ of the standardized variables (i. e., divide each of the column vectors of $\mathbf{X}$ by the standard deviation of its components)
- Computing the matrix $\tilde{\mathbf{G}}$ of eigenvectors with respect to $\mathbf{R}$, the principal component transformation then results in

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\mathbf{G}},$$

and the covariance matrix of $\tilde{\mathbf{Y}}$ is the $(p \times p)$ diagonal matrix

$$\tilde{\mathbf{\Lambda}} = \tilde{\mathbf{G}}'\mathbf{R}\tilde{\mathbf{G}},$$

with the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p$ of $\mathbf{R}$ as diagonal elements

Dimensionality Reduction

- Referring to the principal components $Y_1, \ldots, Y_p$ we obtain a complete description of the data

$$\mathbf{X} = \mathbf{Y}\mathbf{G}'$$

- Often, however, we look for a parsimonious description of the data, using $k < p$ variables, with an as-small-as-possible loss of information

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{Y}^{(k)}\mathbf{G}^{(k)\prime}$$

- Selecting the first k principal components $Y_1, \ldots, Y_k$ results in such an approximate reconstruction of the data
  - The average squared deviation of the original to the reconstructed points equals the sum of the last $p - k$ eigenvalues

Example: Reasons for Studying Psychology

- At LMU Munich, 114 first-year psychology students were surveyed in WS 1979/80 about their reasons for choosing psychology as their field of study
- Respondents were asked to rate the influence that particular reasons had on their choice (on a seven-point Likert scale)

Example: Reasons for Studying Psychology

I study psychology,

1. because I am generally interested in the subject matter of psychology
2. to be able to analyze interpersonal relationships
3. in order to be able to work scientifically in this field later on
4. in order to better understand social norms and constraints
5. to learn to better respond to the needs and views of others
6. because I pursue a specific career goal that requires psychological knowledge
7. to learn how to discover and work independently on new problems in the scientific field
8. to better see through and cope with some difficulties
9. in order to be able to solve concrete social problems in professional practice later on
10. because I see a wide field of professional options for psychological knowledge
11. to be able to work towards improvements in society
12. to learn to help people with psychological problems
13. Recommendations from others (parents, friends, career counseling) played a role in my choice of psychology

Example: Reasons for Studying Psychology

Empirical correlation matrix

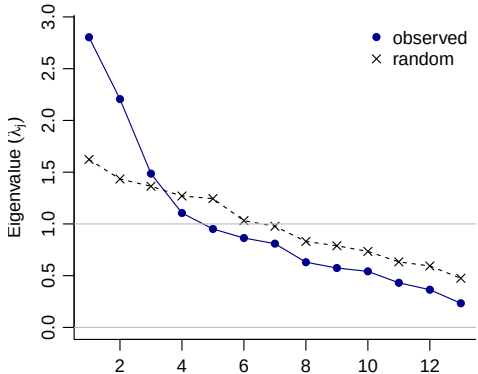
$$\mathbf{R} = \begin{pmatrix}
1.000 & -.020 & .091 & .145 & .145 & -.122 & .152 & -.003 & .048 & .091 & .095 & .141 & -.074 \\
-.020 & 1.000 & -.110 & .393 & .459 & -.023 & -.076 & .375 & .025 & -.072 & .208 & .267 & .072 \\
.091 & -.110 & 1.000 & .043 & -.134 & .149 & .696 & -.085 & .063 & .102 & -.027 & -.025 & -.021 \\
.145 & .393 & .043 & 1.000 & .316 & -.077 & .049 & .296 & .121 & -.071 & .380 & .003 & -.029 \\
.145 & .459 & -.134 & .316 & 1.000 & .063 & .045 & .406 & .216 & .077 & .312 & .298 & .025 \\
-.122 & -.023 & .149 & -.077 & .063 & 1.000 & .238 & -.018 & .475 & .297 & .215 & .316 & .124 \\
.152 & -.076 & .696 & .049 & .045 & .238 & 1.000 & -.094 & .240 & .233 & .218 & .111 & -.010 \\
-.003 & .375 & -.085 & .296 & .406 & -.018 & -.094 & 1.000 & .050 & -.030 & .154 & .135 & .031 \\
.048 & .025 & .063 & .121 & .216 & .475 & .240 & .050 & 1.00 & .380 & .374 & .284 & .113 \\
.091 & -.072 & .102 & -.071 & .077 & .297 & .233 & -.030 & .380 & 1.000 & .170 & .164 & -.001 \\
.095 & .208 & -.027 & .380 & .312 & .215 & .218 & .154 & .374 & .170 & 1.000 & .349 & .001 \\
.141 & .267 & -.025 & .003 & .298 & .316 & .111 & .135 & .284 & .164 & .349 & 1.000 & .099 \\
-.074 & .072 & -.021 & -.029 & .025 & .124 & -.010 & .031 & .113 & -.001 & .001 & .099 & 1.000
\end{pmatrix}$$

Number of Principal Components

Criteria for determining the number of principal components to be retained

- Kaiser-Guttman criterion
  - Leave out principal components whose (standardized) eigenvalues are less than 1 because their variance is smaller than that of the initial variables
- Scree test
  - Plotting the eigenvalues in decreasing order produces a decaying curve, which usually remains almost constant at a low level from an index $k+1$ on, so that only the first k principal components are considered
- Parallel analysis according to Horn (1965)
  - Include principal components with eigenvalues larger than the eigenvalues resulting from analyzing random data

Example: Parallel Analysis



- Parallel analysis suggests a three-component solution
  - The first three eigenvalues are larger than the eigenvalues obtained by random draws from a discrete uniform distribution
- The Kaiser-Guttman criterion extracts too many factors

Factor Loadings

- For describing the contributions of the variables $X_1, \ldots, X_p$ to each of the principal components, the bivariate correlations between the variables and the principal components are considered
- These *factor loadings* are usually collected in a *factor structure matrix*

$$\mathbf{L} = \mathbf{G}\boldsymbol{\Lambda}^{1/2}$$

- If a principle component shows substantial correlations with many of the variables $X_1, \ldots, X_p$, then an interpretation will be difficult (general factor), whereas interpretation is facilitated by high factor loadings (in absolute value) in only a few variables (group factor)

Example: Factor Loadings

With the first three eigenvalues $\lambda_1 = 2.80$, $\lambda_2 = 2.21$ and $\lambda_3 = 1.49$, the following factor loadings result

$$\mathbf{L} = \begin{pmatrix} 0.20 & -0.04 & 0.42 \\ 0.48 & 0.56 & 0.08 \\ 0.13 & -0.61 & 0.59 \\ 0.47 & 0.39 & 0.44 \\ 0.64 & 0.41 & 0.05 \\ 0.46 & -0.47 & -0.43 \\ 0.36 & -0.64 & 0.49 \\ 0.41 & 0.50 & 0.08 \\ 0.63 & -0.34 & -0.30 \\ 0.37 & -0.45 & -0.21 \\ 0.69 & 0.02 & 0.02 \\ 0.60 & -0.03 & -0.27 \\ 0.12 & -0.02 & -0.34 \end{pmatrix}$$

- The column sums of squares of the factor structure matrix reflect the variance explained by the principal components and coincide with the eigenvalues
- Proportions of explained variance

Component 1: $2.80/13$	$= 0.22$
Component 2: $2.21/13$	$= 0.17$
Component 3: $1.49/13$	$= 0.11$
Total:	0.50

Rotation of Principal Components

- To increase interpretability, the pattern of factor loadings may be changed by rotating the principal components
- The factor structure matrix is multiplied by an appropriately specified rotation matrix $\mathbf{M}$

$$\tilde{\mathbf{L}} = \mathbf{LM}$$

- There are two broad classes of rotations
  - Orthogonal rotations
  - Oblique(-angled) rotations

Rotation of Principal Components

- Orthogonal rotations
  - The components remain uncorrelated (but lose the property of variance maximization)
  - Example: The *varimax* method looks for components with high absolute loadings in some variables and low absolute loadings on the remaining variables (it tries to maximize the variance of the loadings)
- Oblique rotations
  - The components will not remain uncorrelated
  - Example: The *promax* method searches for squared loadings close to 0 or 1

Example: Rotating Factor Loadings

- The varimax rotation matrix is

$$\mathbf{M} = \begin{pmatrix} 0.678 & -0.694 & -0.243 \\ 0.509 & 0.682 & -0.525 \\ 0.530 & 0.232 & 0.816 \end{pmatrix}$$

- The column sums of squares of the rotated loading matrix specify the explained variance

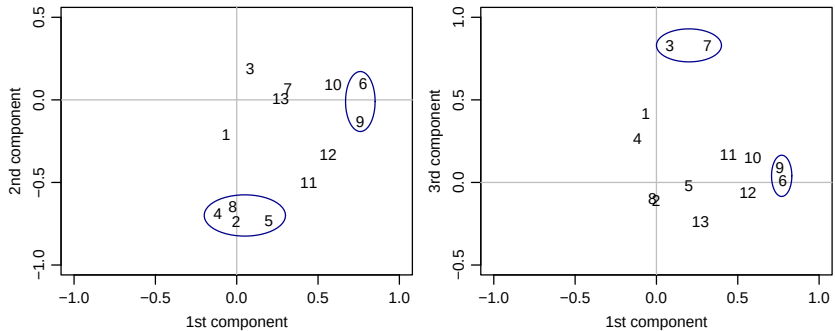
Component 1: $2.28/13$	$= 0.17$
Component 2: $2.45/13$	$= 0.19$
Component 3: $1.76/13$	$= 0.14$
Total:	0.50

- The rotated factor loadings are

$$\tilde{\mathbf{L}} = \begin{pmatrix} 0.07 & 0.21 & 0.42 \\ 0.00 & 0.74 & -0.11 \\ -0.08 & -0.19 & 0.83 \\ 0.12 & 0.69 & 0.27 \\ -0.20 & 0.73 & -0.02 \\ -0.78 & -0.10 & 0.01 \\ -0.31 & -0.07 & 0.83 \\ 0.02 & 0.65 & -0.10 \\ -0.76 & 0.13 & 0.09 \\ -0.59 & -0.09 & 0.15 \\ -0.44 & 0.50 & 0.17 \\ -0.56 & 0.33 & -0.06 \\ -0.27 & -0.01 & -0.24 \end{pmatrix}$$

Example: Rotating Factor Loadings

- If the rotated factor loadings are plotted against each other, the particularly high or low loadings may provide clues for the interpretation of the components



- Rotated principal components
  1. Career goal (17%)
  2. Socio-analytic factor (19%)
  3. Science factor (14%)

Factor Scores

- The values assigned to subjects on the extracted components (factor scores) may be obtained from the principal components by

$$\mathbf{F} = \mathbf{Y}\mathbf{\Lambda}^{-1/2} = \mathbf{X}\mathbf{G}\mathbf{\Lambda}^{-1/2}$$

- This allows for predicting factor scores for a new person with observations $\mathbf{x}$
- Rotated factor scores are generally obtained by means of

$$\tilde{\mathbf{F}} = \mathbf{F}\mathbf{M}^{-1\prime}$$

- For orthogonal rotations we have $\mathbf{M}' = \mathbf{M}^{-1}$, and therefore

$$\tilde{\mathbf{F}} = \mathbf{F}\mathbf{M}$$

Principal Component and Factor Analysis

- Principal component and factor analysis rely on the same basic equation

$$\mathbf{X} = \mathbf{F}\mathbf{L}' + \mathbf{E}$$

  i. e., the data are "modeled" as a linear combination of the extracted components and their loadings as well as an error term
- In principal component analysis, however, no distributional assumptions are imposed on the error term
- Principal components
  - are manifest variables that can be calculated from the data using affine transformations
  - are descriptive in nature and should retain "as much information as possible" of the data

Principal Component and Factor Analysis

- In factor analysis, on the other hand, the basic equation is explicitly formulated as a statistical model whose parameters may be estimated (e. g., by maximum likelihood estimation)
- The factors
  - are hypothetical variables, which in principle cannot be observed (latent variables)
  - can be estimated under the (generally untested) assumption that they exist

Example: Reasons for Studying Psychology

Correlation matrix

```
R <- as.matrix(read.table(text = "
  1.000 -.020  .091  .145  .145 -.122  .152 -.003 .048  .091  .095  .141 -.074
  -.020 1.000 -.110  .393  .459 -.023 -.076  .375 .025 -.072  .208  .267  .072
   .091 -.110 1.000  .043 -.134  .149  .696 -.085 .063  .102 -.027 -.025 -.021
   .145  .393  .043 1.000  .316 -.077  .049  .296 .121 -.071  .380  .003 -.029
   .145  .459 -.134  .316 1.000  .063  .045  .406 .216  .077  .312  .298  .025
  -.122 -.023  .149 -.077  .063 1.000  .238 -.018 .475  .297  .215  .316  .124
   .152 -.076  .696  .049  .045  .238 1.000 -.094 .240  .233  .218  .111 -.010
  -.003  .375 -.085  .296  .406 -.018 -.094 1.000 .050 -.030  .154  .135  .031
   .048  .025  .063  .121  .216  .475  .240  .050 1.00  .380  .374  .284  .113
   .091 -.072  .102 -.071  .077  .297  .233 -.030 .380 1.000  .170  .164 -.001
   .095  .208 -.027  .380  .312  .215  .218  .154 .374  .170 1.000  .349  .001
   .141  .267 -.025  .003  .298  .316  .111  .135 .284  .164  .349 1.000  .099
  -.074  .072 -.021 -.029  .025  .124 -.010  .031 .113 -.001  .001  .099 1.000
")))
```

Eigenvalues and eigenvectors

```
eigen(R)
```

R Demo

Scree test and parallel analysis

```
plot(eigen(R)$values, type = "b", col = "darkblue")
sample(1:7, 114*13, replace = TRUE) |>  # random ratings
  matrix(114, 13) |> cor() |> eigen() |> _$values |>
  points(type = "b", pch = 4, lty = 3)
abline(h = 1, lty = 2)
```

Principal component analysis

```
m <- princomp(covmat = R)
summary(m)
```

Factor loadings

```
L <- loadings(m) %*% diag(m$sdev)  # L = GΛ^{1/2}
```

Varimax rotation of the first three principal components

```
Lt <- varimax(L[, 1:3])$loadings
print(Lt, cutoff = 0.2)
```

References

Bortz, J. (1999). Statistik für Sozialwissenschaftler. Berlin: Springer. Kapitel 15

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). Multivariate statistische Verfahren. Berlin: Walter de Gruyter. Kapitel 11

Schönemann, P.H. & Borg, I. (1996). Von der Faktorenanalyse zu den Strukturgleichungsmodellen. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), Handbuch Quantitative Methoden. Weinheim: Psychologie Verlags Union.

Stevens, J. (2001). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. Chapter 11