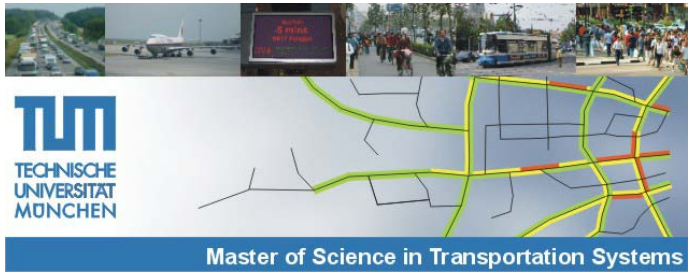
M.Sc. in ‚Transportation Systems'




# Applied Statistics in Transport
# Statistical Tests, Models

Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 17/01/2012

# Last Week: Statistical Tests on the Mean

- One sample problem
- Comparison of two independent samples
- Comparison of two paired samples
- Variance known/unknown
- Normal/non-normal data

- Checking the assumptions:
- Comparison of variances
- Test for normality of the data

# Plan for Today's Lecture: Statistical Models

- Statistical Models – Overview
- Statistical models in R

# Statistical Modelling - Introduction

- Choosing the right kind of statistical analysis is an important step of your analysis
- Choice depends on the nature of your data and on the particular question you are trying to answer

Key:

- Type of response variable = variable whose variation you are attempting to understand, y-axis of the graph
- Type of explanatory variables = which variation in the response variable is associated with variation in the explanatory variable

Questions:

- Which of your variables is the response variable?
- Which are the explanatory variables?
- Are the explanatory variables continuous or categorical, or a mixture of both?
- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, a category?

- These simple keys will lead you to the appropriate statistical model.

# Statistical Modelling - Introduction

- Objective: determine the values of the parameters in a specific model that lead to the best fit of the model to the data
- The model is fitted to the data, not the other way round
- Best model: produces the least unexplained variation (the minimal residual deviance), subject to the constraint that all parameters in the model should be statistically significant.
- There is not one model.

- The model should be minimal: principle of parsimony (Occam's Razor):
- Given a set of equally good explanations for a given phenomenon, the correct explanation is the simplest explanation.
- Einstein: The model should be as simple as possible. But not simpler.
- Oscar Wilde: Truth is rarely pure, and never simple.

# Principle of Parsimony for Statistical Modelling

- Models should have as few parameters as possible.
- Linear models should be preferred to non-linear models.
- Experiments relying on few assumptions should be preferred to those relying on many.
- Models should be simplified until they are minimal adequate.
- Simple explanations should be preferred to complex explanations.
- Prefer explanatory variables that are easy to measure.
- Prefer models that are based on a sound mechanistic understanding of the process over purely empirical functions.

# Model Formulae in R

- Response variable ~ explanatory variable(s)
- The right hand side of the model formulae shows:
- The number of explanatory variables and their identities
- The interactions between the explanatory variables
- (Non-linear terms in the explanatory variables)

- As with the response variable, the explanatory variables can appear as transformations, or as powers, or polynomials.

# Model Formulae in R

- Important: symbols are used differently in model formulae than in arithmetic expressions:
- + indicates inclusion of an explanatory variable in the model (not addition)
- - indicates deletion of an explanatory variable from the model (not subtraction)
- * indicates inclusion of explanatory variables and interactions (not multiplication)
- | indicates conditioning (not 'or'),
  so that y~x|z is read as 'y as a function of x given z‘
- The colon denotes an interaction.

# Model Formulae in R

- Important: symbols are used differently in model formulae than in arithmetic expressions:
- A*B*C is the same as A+B+C+A:B+A:C+B:C+A:B:C
- (A+B+C)^3 is the same as A*B*C
- (A+B+C)^2 is the same as A*B*C-A:B:C

- Interactions between explanatory variables
- Number of interaction effects: (a-1)(b-1)
  with a and b being the number of levels of the two factors
- Interaction between a categorical and a continuous variable are interpreted as an analysis of covariance: a separate slope and intercept are fitted for each level of the categorical variable.

# Choosing the Appropriate Statistical Model

The explanatory variables:

- All explanatory variables continuous: **Regression**
- All explanatory variables categorial: **Analysis of Variance (ANOVA)**
- Explanatory variables both continuous and categorial: Analysis of covariance (ANCOVA)

The response variable:

- Continuous Normal regression, ANOVA, ANCOVA
- Proportion Logistic regression (glm(y~x,family=binomial)
- Count Log-linear models (glm(y~x,family=poisson)
- Binary Binary logistic analysis (glm(y~x,family=binomial)

Multivariate models: More than one response variable

- Multivariate ANalysis Of VAriance - MANOVA
- Principal Component Analysis - PCA
- Cluster analysis

# Regression

- Regression and correlation analysis:
  describe and analyse the relationship between random variables

- Regression: describes the type of directional relationship between mainly ratio/interval scaled variables (the more … the more/less …)

- Correlation: describes the intensity of the non-directional relationship

- Example:
- Relation between the cubic capacity and the fuel consumption of a car
- Regression of the response variable PS as a function of household type, income, spatial location of the household, usage of the car

# ANOVA

- Response variable: ratio/interval scaled
- Explanatory variables: categorical (ordinal, nominal, grouped interval data), are called factors
- Each factor has two or more levels
- For one single factor we can use the t-test

- ANOVA can be classified by the number of explanatory variables
- One-way ANOVA: one explanatory variable
- Two- and more way ANOVA: interaction effects can be included.

- We test whether the differences between the means of the different groups are high enough to conclude on differences in the populations; is the variation between the groups higher than within the groups?

# Analysis of Covariance – ANCOVA

- Combines elements from regression and ANOVA
- Response variable is continuous
- At least one continuous and at least one categorical explanatory variable

- Approach:
- Fit two or more linear regressions of y against x (one for each level of the factor)
- Estimate different slopes and intercepts for each level
- Use model simplification (deletion tests) to eliminate unnecessary parameters

# Analysis of Covariance – ANCOVA, Example

- Medical experiment: response variable: days to recovery; explanatory variables: smoker or not (categorical) and blood cell count (continuous)
- Economics: response variable: local unemployment rate; explanatory variables: country (categorical) and population size (continuous)

- Weight: response variable: weight; explanatory variables: sex (categorical) and age (continuous)
- Maximal model: four parameters: two slopes (one for males and one for females) and two intercepts (one for males and one for females):
- $Weight_{male} = a_{male} + b_{male} * age$
- $Weight_{female} = a_{female} + b_{female} * age$
- Try to simplify the model (principle of parsimony):
- Possible models: two intercepts and a common slope, one intercept and two slopes, ...

# Factor Analysis / Principal Component Analysis

- Factor Analysis is a familiy of approaches, PCA is one of them
- Basic aim of PCA: describe variation in a set of correlated variables, $x_1$, $x_2$, ..., $x_n$, in terms of a new set of uncorrelated variables, $y_1$, $y_2$, ..., $y_n$, each of which is a linear combination of the x variables.
- The new variables are derived in decreasing order of “importance” in the sense that 1 accounts for as much of the variation in the original data amongst all linear combinations of $x_1$, $x_2$, ..., $x_n$.
- Then $y_2$ is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with $y_1$, etc.
- The new variables defined by this process, $y_1$, $y_2$, ..., $y_n$, are the principal components.
- PCA is often done as a first step of cluster analysis.

# Cluster Analysis

- Cluster analysis is a generic term for a range of numerical methods for examining multivariate data
- Goal: uncover groups/clusters of observations that are homogeneous and separated from other groups
- Clusters are identified by the relative distances between points
- Examples for methods for finding the clusters:
- Agglomerative hierarchical clustering: start with each object being a cluster, compute the distances, merge the two objects with the smallest distance, compare the remaining clusters, merge the two with the smallest distance, etc.
- k-means: consider every possible partition of the n objects into k groups, select the one with the lowest within-sum-of-squares, search algorithms are necessary, initial partition can be found with hierarchical clustering techniques
- Example: Clustering of test-drive-data for finding typical clusters of kinematic characteristics (e.g. average speed, acceleration, share stop-and-go)

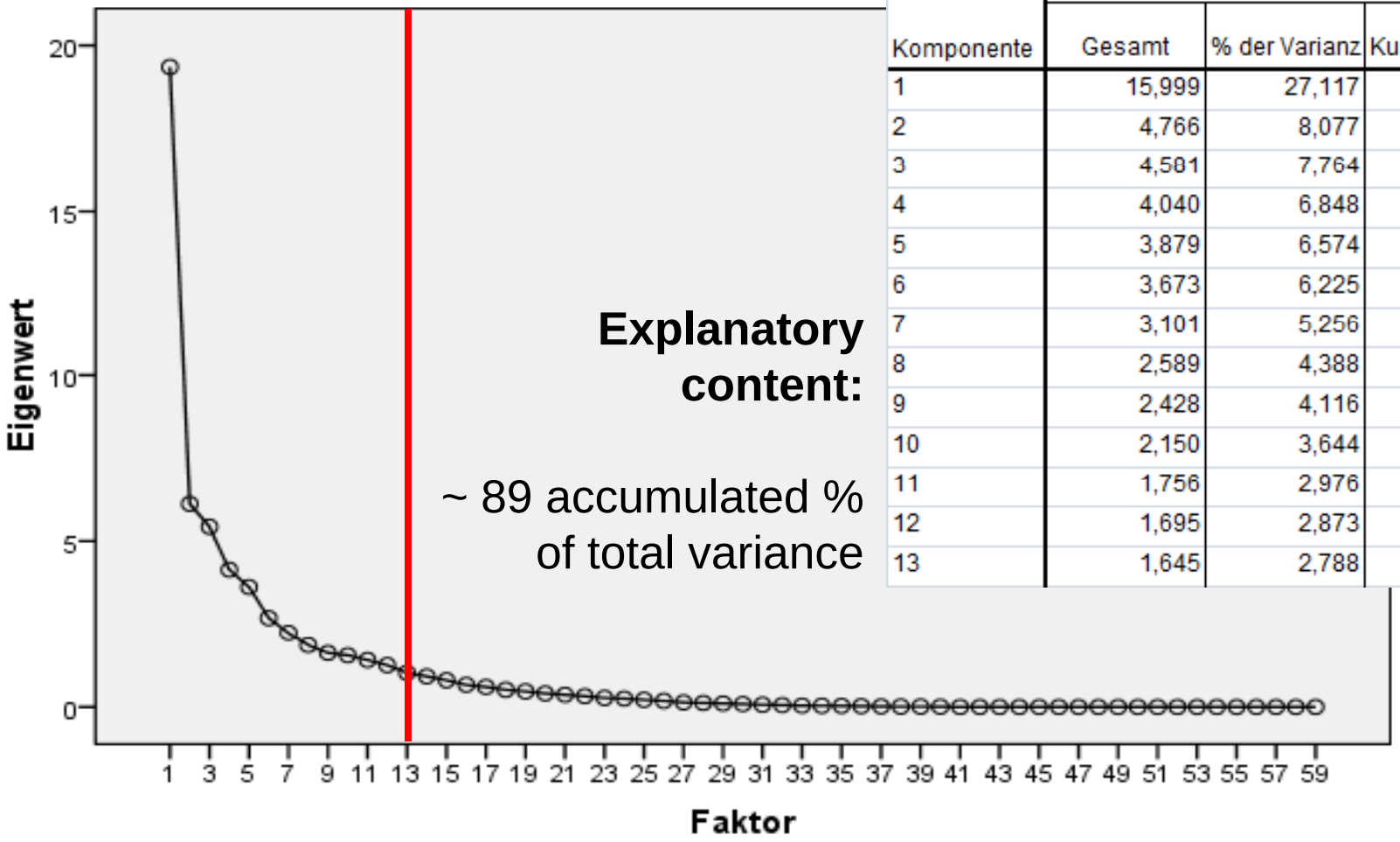
# Factor / Cluster Analysis - Example

- Megacities Project (Chair Prof. Wulfhorst)
- Aim: Find typical clusters of Mega-Cities in terms of transport indicators
- Approach: Combination of PCA and cluster analysis

- Start with 59 relevant key indicators:
- General characteristics
  (e.g. urban density, job density, GDP per capita)
- Transport supply indicators
  (e.g. length of road network, length of PT network, motorization rate)
- Mobility indicators
  (e.g. no. of daily trips (per mode), trip distances, …)
- Investment in transport (per mode)
- Transport Externalities (e.g. energy use, emissions, fatalities)

# First Step: Factor Analysis, Finding the Factors

## Result: 13 relevant factors




**Explanatory content:**

~ 89 accumulated % of total variance

| Erklärte Gesamtvarianz nach Rotation | | | |
|---|---|---|---|
| | Ladungen | | |
| Komponente | Gesamt | % der Varianz | Kumulierte % |
| 1 | 15,999 | 27,117 | 27,117 |
| 2 | 4,766 | 8,077 | 35,195 |
| 3 | 4,581 | 7,764 | 42,958 |
| 4 | 4,040 | 6,848 | 49,806 |
| 5 | 3,879 | 6,574 | 56,380 |
| 6 | 3,673 | 6,225 | 62,605 |
| 7 | 3,101 | 5,256 | 67,861 |
| 8 | 2,589 | 4,388 | 72,249 |
| 9 | 2,428 | 4,116 | 76,365 |
| 10 | 2,150 | 3,644 | 80,009 |
| 11 | 1,756 | 2,976 | 82,985 |
| 12 | 1,695 | 2,873 | 85,858 |
| 13 | 1,645 | 2,788 | 88,646 |

# First Step: Factor Analysis, Interpretation of the Factors

| Factor | |
|---|---|
| 1 | Urban sprawl and automobile dependance |
| 2 | Taxi traffic |
| 3 | PT vehicle intensity |
| 4 | Shared taxi traffic |
| 5 | Transport deaths and scattering PT supply |
| 6 | Congestion |
| 7 | Scarce PT supply |
| 8 | PT usage |
| 9 | Preconditions for non-motorized transport |
| 10 | Private traffic trip length |
| 11 | PT energy use |
| 12 | Parking charges |
| 13 | Parking management/restrictions |

# Second Step: Cluster Analysis
# World Map of Megacities Clusters

Hybrid Cities

Traffic saturated Cities

Transit cities

Non-motorized Cities

Auto cities

Paratransit Cities

*Manila*