Statistical Databases and Registers with some datamining

a course in
Survey Methodology and Official Statistics

Departement of Statistics
Stockholm University

October 2010

What this course is about

The course consists of mainly three parts: Databases, Registers and Data mining and the syllabus is

	Day	Time	Comment	
Nov	1	14:15-15	Introduction	MM
	1	15:15-17	Construction of a database	PW
	3	15:15-17	Database-theory	BS
	8	15:15-17	Database-theory	BS
	10	15:15-17	Database-theory	BS
	11	15:15-17	Statistical data editing	AN
	15	15:15-17	Register-theory	BS
	17	15:15-17	Register-theory	BS
	22	10-16	Astra Zeneca, Södertälje	
			Registerbaserad statistik	

What this course is about (forts)

	Day	Time	Comment	
	24	15:15-17	Register-theory	BS
	29	15:15-17	Analysis of a commersial register	PW
Dec	1	15:15-17	Statistics from databases, visualisation and new media	AF
	8	15:15-17	Datamining-theory	MM
	13	15:15-17	Datamining-theory	MM
	15	15:15-17	Datamining-theory	MM
	20	15:15-17	Datamining with SAS	ML
Jan	10	15:15-17	Reserve	

Place: Room B705

Literature

1. In database theory notes by Bo Sundgren.
2. In register theory the book by Wallgren, A. och Wallgren, B. *Register-based Statstics – Administrative Data for Statistical Purposes*. Chichester, Wiley.
3. In statistical learning the book by Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning*. This book may also be found as a free pdf.
4. In data editing the following articles by Anders Norberg: Swedish Editing Methods and a short article.
5. Lecture notes to be found at the course home page.

Examination

The exam consists of a total of 10 tasks and each task give at most 10 credits.

Eight of the tasks shall be solved at the exam date which is January the 13:th, 2012 (answers in either svenska or English) or at reexam which is February the 8:th, 2012.

There is one assignment that consist of two tasks. This assignment should be finished and published in 2011.

Since each task give at most 10 credits a maximum of 100 credits is possible.

Course must be finished February 2012. Next available examination time is next time the course is given.

The tasks

Database theory: The exam will consist of 3 questions.

Register theory: The exam will consist of 3 questions.

Statistical learning: The exam will consist of 2 questions.

Assignment: Consists of 2 tasks.

Final grade

Final grading is according to the following table

Betyg		**Poäng**
A	Excellent	$90-100$
B	Very good	$80-89$
C	Good	$70-79$
D	Satisfactory	$60-69$
E	Adequate	$50-59$
Fx	Insufficient	$30-49$
F	Fail	≤ 29

Examiner and coordinator: Mikael Möller

Database and register

The subject register is under development and hence there are confusing notations. The following is my personal view.

- A **database** is a collection of tables that obeys certain parsimonious criteria. Once a data is entered it is (almost) never changed. A database consists of a collection of linked tables.
- A **register** is a collection of tables where updating is common. A register is usually not parsimonious.
- Registers are split into two different types
  1. **Administrative** registers: primarily used in administrative information systems
  2. **Statistical** registers: primarily used for statistical information as sums, means, deviations and so on. They are usually based on data from the administrative registers

Specialized registers

There is also another division of the registers into **base** and **specialized** registers

- **Administrative base** registers are kept as a basic resource for public administration
- **Statistical base** registers are registers for statistics and they are based on administrative base registers
- **Administrative specialized** registers such as the vehicle register
- **Statistical specialized** registers are statistical registers based on several administrative registers

Seminarium i Södertälje

Anders och Britt Wallgren, SCB/Örebro univ

Har skrivit en bok om registerbaserad statistik och kommer nog att prata allmänt om metodskillnader mellan urvalsundersökningar och registerundersökningar.

Lotta Persson, SCB

Berättar om en studie om tredjebarns-födslar, där flera register använts och man tillämpat överlevnadsanalys.

Stefan Franzén, Astra Zeneca

Tror att det kommer att handla om undersökningar mot patientjournaler och liknande

Ytterligare 1-2 talare: MEB vid KI, epidemiologiska tillämpningar

Example of registers

In Sweden we have four official administrative **base** registers

- Register on persons
- Register on property
- Register on business
- Register on business activity

But each company have registers of their own.

- Register of customers
- Register of transactions
- Register on products
- and so on

Datamining

When we start to study the subject datamining we will realize that it is a question of supervised and unsupervised learning.

And with learning we mean estimation of parameters.

Hence statistical learning would be a better name for the statistical procedures in datamining.

Our intention is to study

1. Supervised learning
   1. Linear Regression and Nearest Neighbour
   2. Neural Networks
2. Unsupervised learning
   1. Cluster Analysis
   2. The Google PageRank Algorithm