Managing Large Sets Of Models

R. Seger A. Unwin

Department of Computer Oriented Statistics and Data Analysis
University of Augsburg

UseR 2006

Outline

Model And Model Space

- `Models` may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Introduction
Managing Models
Summary

Model And Model Space

- `Models` may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Model And Model Space

- Models may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Model And Model Space

- Models may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Model And Model Space

- Models may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Model And Model Space

- Models may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Model And Model Space

- `Models` may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Introduction
Managing Models
Summary

Model And Model Space

- `Models` may be fit from a variety of different classes
  - e.g. lm, gam, rpart
- and to variations of the original dataset
  - the (unmodified) dataset
  - subsets of the data
  - versions using transformed variables
  - datasets from sampling (e.g. bootstrapping)

Conclusion

Model space size is affected by the range of models, by transformations and by sampling

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin Managing Large Sets Of Models

Model Analysis

- model statistics (for model selection/comparison)
  - global statistics for comparing overall fit
  - measures of variable importance
  - residuals for comparing local fit

Conclusion

Models yield much useful data for further analysis and you need a tool to organise that flood of data

Introduction
Managing Models
Summary

Model Analysis

- model statistics (for model selection/comparison)
  - global statistics for comparing overall fit
  - measures of variable importance
  - residuals for comparing local fit

Conclusion

Models yield much useful data for further analysis and you need a tool to organise that flood of data

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin Managing Large Sets Of Models

Model Analysis

- model statistics (for model selection/comparison)
  - global statistics for comparing overall fit
  - measures of variable importance
  - residuals for comparing local fit

Conclusion

Models yield much useful data for further analysis and you need a tool to organise that flood of data

Introduction
Managing Models
Summary

Model Analysis

- model statistics (for model selection/comparison)
  - global statistics for comparing overall fit
  - measures of variable importance
  - residuals for comparing local fit

Conclusion

Models yield much useful data for further analysis and you need a tool to organise that flood of data

Introduction
Managing Models
Summary

Model Analysis

- model statistics (for model selection/comparison)
  - global statistics for comparing overall fit
  - measures of variable importance
  - residuals for comparing local fit

Conclusion

Models yield much useful data for further analysis and you need a tool to organise that flood of data

Outline

Technical Requirements

- store model forms and results
- commands in R ⇔ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Technical Requirements

- store model forms and results
- commands in R ⇔ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis

R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Technical Requirements

- store model forms and results
- commands in R ⇔ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis

R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Technical Requirements

- store model forms and results
- commands in R ⇔ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis

R. Seger, A. Unwin
Managing Large Sets Of Models

Technical Requirements

- store model forms and results
- commands in R $\Leftrightarrow$ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Technical Requirements

- store model forms and results
- commands in R $\Leftrightarrow$ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Technical Requirements

- store model forms and results
- commands in R $\Leftrightarrow$ series of textual input
- model information includes
  - formula
  - global statistics (e.g. deviance)
  - coefficient estimates
  - residuals

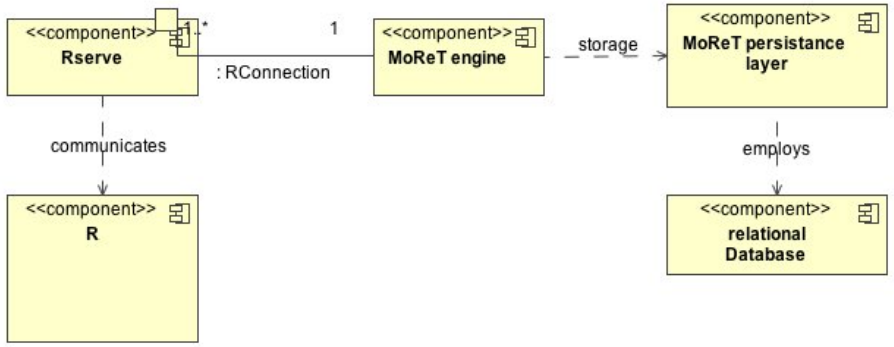
MORET

In order to manage and compare model results a lot of data has to be stored. The software "MORET" uses a relational database for this (see next slide).

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Used Components



Outline

1. Introduction
2. Managing Models
   - Software Requirements
   - Model Repository
   - Model Comparison

Working With Moret

- load datasets with the MORET-GUI
- fit models to the data
- use MORET to manage the database of the models.
  - export model data
  - use the Model Explorer

Working With Moret

- load datasets with the MORET-GUI
- fit models to the data
- use MORET to manage the database of the models.
  - export model data
  - use the Model Explorer

Working With Moret

- load datasets with the MORET-GUI
- fit models to the data
- use MORET to manage the database of the models.
  - export model data
  - use the Model Explorer

Working With Moret

- load datasets with the MORET-GUI
- fit models to the data
- use MORET to manage the database of the models.
  - export model data
  - use the Model Explorer

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

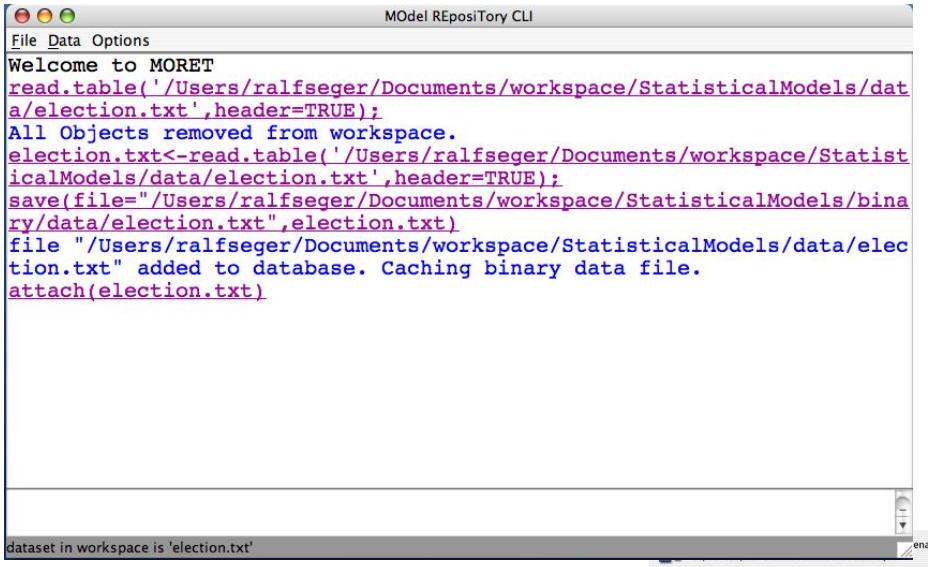
Working With Moret

- load datasets with the MORET-GUI
- fit models to the data
- use MORET to manage the database of the models.
  - export model data
  - use the Model Explorer

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Working With Moret/2



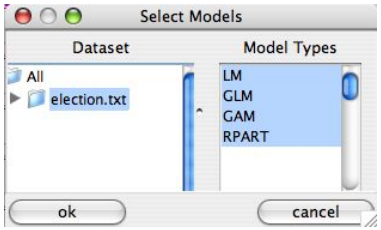
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin
Managing Large Sets Of Models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Outline

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin
Managing Large Sets Of Models

Model Selection

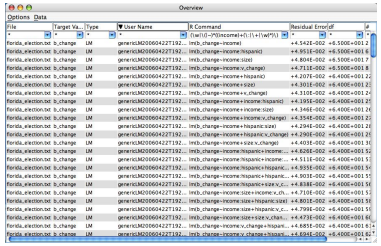


Preselect Models

- Models space is huge.
- Work with selected subsets of all models.

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Model Selection

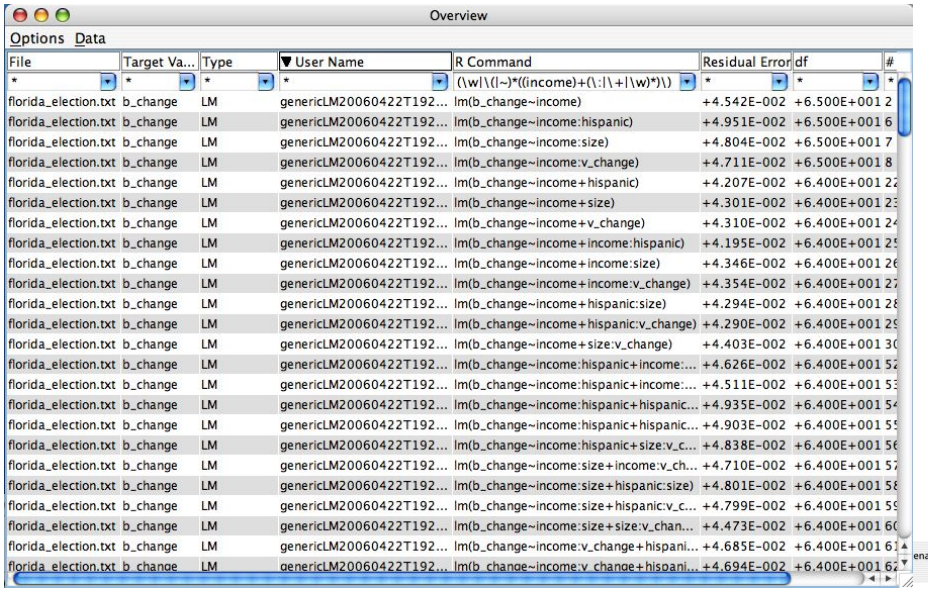


Preselect Models

- Models space is huge.
- Work with selected subsets of all models.

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

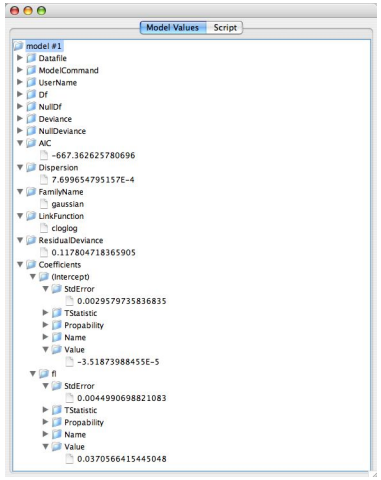
Model Overview



File	Target Va...	Type	User Name	R Command	Residual Error	df	#
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income)	+4.542E-002	+6.500E+001	2
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic)	+4.951E-002	+6.500E+001	6
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:size)	+4.804E-002	+6.500E+001	7
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:v_change)	+4.711E-002	+6.500E+001	8
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+hispanic)	+4.207E-002	+6.400E+001	22
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+size)	+4.301E-002	+6.400E+001	23
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+v_change)	+4.310E-002	+6.400E+001	24
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+income:hispanic)	+4.195E-002	+6.400E+001	25
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+income:size)	+4.346E-002	+6.400E+001	26
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+income:v_change)	+4.354E-002	+6.400E+001	27
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+hispanic:size)	+4.294E-002	+6.400E+001	28
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+hispanic:v_change)	+4.290E-002	+6.400E+001	29
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income+size:v_change)	+4.403E-002	+6.400E+001	30
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic+income:...	+4.626E-002	+6.400E+001	52
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic+income:...	+4.511E-002	+6.400E+001	53
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic+hispanic...	+4.935E-002	+6.400E+001	54
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic+hispanic...	+4.903E-002	+6.400E+001	55
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:hispanic+size:v_c...	+4.838E-002	+6.400E+001	56
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:size+income:v_ch...	+4.710E-002	+6.400E+001	57
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:size+hispanic:size)	+4.801E-002	+6.400E+001	58
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:size+hispanic:v_c...	+4.799E-002	+6.400E+001	59
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:size+size:v_chan...	+4.473E-002	+6.400E+001	60
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:v_change+hispani...	+4.685E-002	+6.400E+001	61
florida_election.txt	b_change	LM	genericLM20060422T192...	lm(b_change~income:v_change+hispani...	+4.694E-002	+6.400E+001	62

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

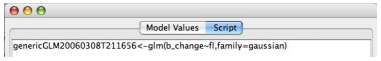
Model Overview Options



Working with the Model Table

- inspect a single model
- view the creation script
- export models

Model Overview Options

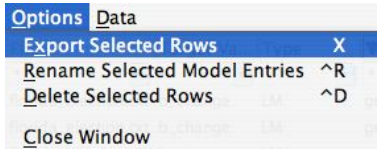


Working with the Model Table

- inspect a single model
- view the creation script
- export models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Model Overview Options

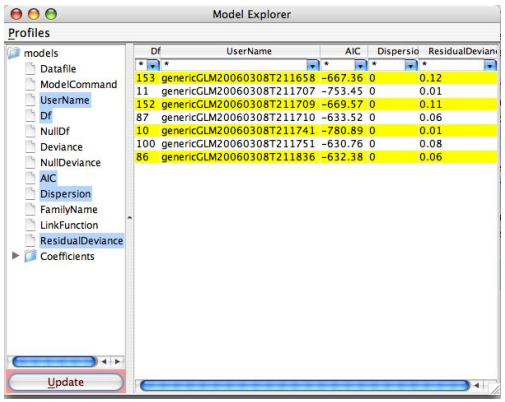


Working with the Model Table

- inspect a single model
- view the creation script
- export models

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Model Explorer

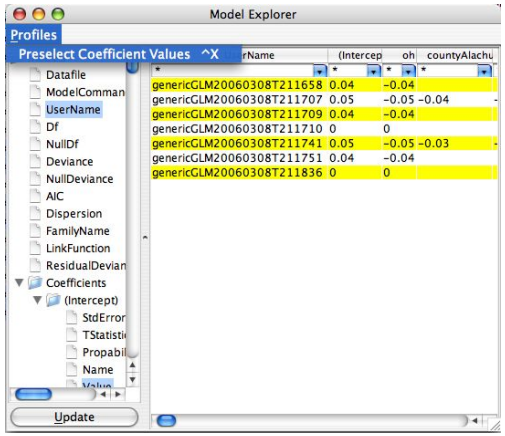


Actions

- dynamically select attributes of the stored models
- use predifined profiles (e.g. for all coefficient values)

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Model Explorer



Actions

- dynamically select attributes of the stored models
- use predifined profiles (e.g. for all coefficient values)

Using External Applications

External Interfaces

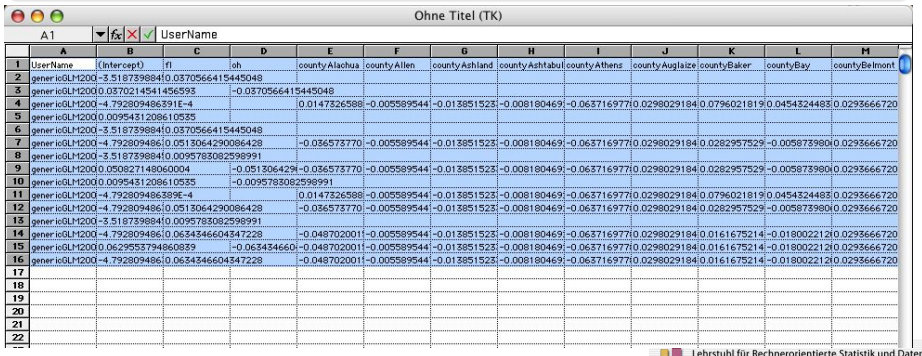
- export models to XML or CSV
- copy and paste to a spreadsheet

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Using External Applications

External Interfaces

- export models to XML or CSV
- copy and paste to a spreadsheet



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	UserName	(Intercept)	fl	oh	countyAlachua	countyAllen	countyAshland	countyAshtabul	countyAthens	countyAuglaize	countyBaker	countyBay	countyBelmont
2	generic	GLM200	-3.518739884	0.037056641	5445048								
3	generic	GLM200	0.037021454	1456593	-0.037056641	5445048							
4	generic	GLM200	-4.792809486	391E-4		0.014732658	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.079602181	9 0.045432448	3 0.029366672	0
5	generic	GLM200	0.009543120	8610535											
6	generic	GLM200	-3.518739884	0.037056641	5445048										
7	generic	GLM200	-4.792809486	0.051306429	0086428	-0.036573770	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.028295752	9 -0.005873980	0.029366672	0
8	generic	GLM200	-3.518739884	0.009578308	2598991										
9	generic	GLM200	0.050827148	060004	-0.051306429	-0.036573770	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.028295752	9 -0.005873980	0.029366672	0
10	generic	GLM200	0.009543120	8610535	-0.009578308	2598991									
11	generic	GLM200	-4.792809486	389E-4		0.014732658	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.079602181	9 0.045432448	3 0.029366672	0
12	generic	GLM200	-4.792809486	0.051306429	0086428	-0.036573770	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.028295752	9 -0.005873980	0.029366672	0
13	generic	GLM200	-3.518739884	0.009578308	2598991										
14	generic	GLM200	-4.792809486	0.063434660	4347228	-0.048702001	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.016167521	4 -0.018002212	0 0.029366672	0
15	generic	GLM200	0.062955379	4860839	-0.063434660	-0.048702001	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.016167521	4 -0.018002212	0 0.029366672	0
16	generic	GLM200	-4.792809486	0.063434660	4347228	-0.048702001	-0.005589544	-0.013851523	-0.008180469	-0.063716977	0.029802918	4 0.016167521	4 -0.018002212	0 0.029366672	0
17															
18															
19															
20															
21															
22															

Introduction
Managing Models
Summary
Software Requirements
Model Repository
Model Comparison

Summary

- Managing large sets of models is possible using MORET
- MORET can be used as link to other applications for further analysis

- Further developments
  - Integration of more model alternatives
  - Better control of transformed variables

Summary

- Managing large sets of models is possible using MORET
- MORET can be used as link to other applications for further analysis

- Further developments
  - Integration of more model alternatives
  - Better control of transformed variables

Additional Sources I

R. Seger
project homepage.
http://stats.math.uni-augsburg.de/software/.

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Dept. of Computer Oriented Statistics and Data Analysis
R. Seger, A. Unwin Managing Large Sets Of Models