Chapter 7
Automatic Linear Regression (35 Patients)

1 General Purpose

Automatic linear regression is in the Statistics Base add-on module SPSS version 19 and up. X-variables are automatically transformed in order to provide an improved data fit, and SPSS uses rescaling of time and other measurement values, outlier trimming, category merging and other methods for the purpose.

2 Schematic Overview of Type of Data File

Outcome	binary predictor	additional predictors…..
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.

This chapter was previously partly published in "Machine learning in medicine a complete overview" in the Chap. 31, 2015.

T.J. Cleophas, A.H. Zwinderman, *SPSS for Starters and 2nd Levelers*,
DOI 10.1007/978-3-319-20600-4_7

3 Specific Scientific Question

Can automatic rescaling and outlier trimming as available in SPSS be used to maximize linear relationships in multiple linear regression models.

4 Data Example

In a clinical crossover trial an old laxative is tested against a new one. Numbers of stools per month is the outcome. The old laxative and the patients' age are the predictor variables. Does automatic linear regression provide better statistics of these data than traditional multiple linear regression does.

Outcome	Predictor	Age category	Patient id	Predicted values
24,00	8,00	2,00	1,00	26,41
30,00	13,00	2,00	2,00	27,46
25,00	15,00	2,00	3,00	27,87
35,00	10,00	3,00	4,00	38,02
39,00	9,00	3,00	5,00	37,81
30,00	10,00	3,00	6,00	38,02
27,00	8,00	1,00	7,00	26,41
14,00	5,00	1,00	8,00	25,78
39,00	13,00	1,00	9,00	27,46
42,00	15,00	1,00	10,00	27,87

Outcome = new laxative
Predictor = old laxative

Only the first 10 patients of the 35 patients are shown above. The entire file is in extras.springer.com and is entitled "chapter7automaticlinreg". We will first perform a standard multiple linear regression. For analysis the module Regression is required. It consists of at least 10 different statistical models, such as linear modeling, curve estimation, binary logistic regression, ordinal regression etc. Here we will simply use the linear model.

5 Standard Multiple Linear Regression

Command:
Analyze....Regression....Linear....Dependent: enter newtreat....Independent: enter oldtreat and agecategories....click OK.

Model summary

Model	R	R square	Adjusted R square	Std. Error of the estimate
1	,429[a]	,184	,133	9,28255

[a]Predictors: (Constant), oldtreat, agecategories

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	622,869	2	311,435	3,614	,038[b]
	Residual	2757,302	32	86,166		
	Total	3380,171	34			

[a]Dependent variable: newtreat
[b]Predictors: (Constant), oldtreat, agecategories

Coefficients[a]

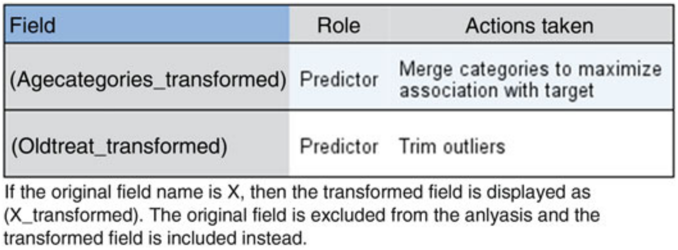
		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	20,513	5,137		3,993	,000
	Agecategories	3,908	2,329	,268	1,678	,103
	Oldtreat	,135	,065	,331	2,070	,047

[a]Dependent variable: newtreat

6 Automatic Linear Modeling

The same commands are given, but, instead of the model Linear, click the model Automatic Linear Modeling. The underneath interactive output sheets are given.

Automatic data preparation
Target:newtreat

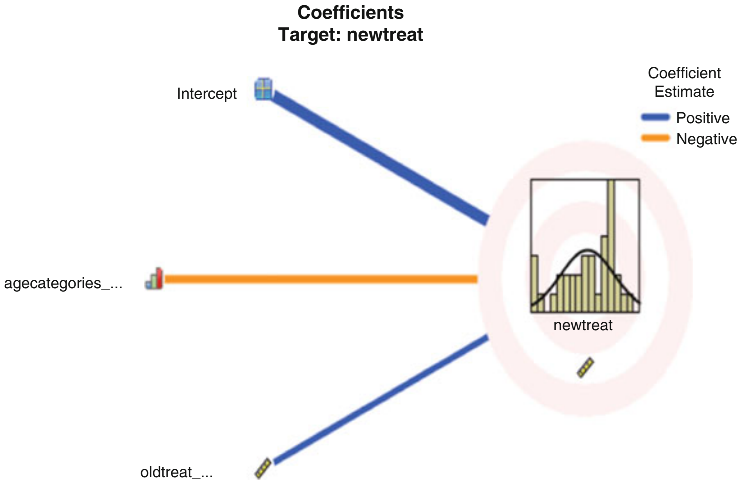


Field	Role	Actions taken
(Agecategories_transformed)	Predictor	Merge categories to maximize association with target
(Oldtreat_transformed)	Predictor	Trim outliers

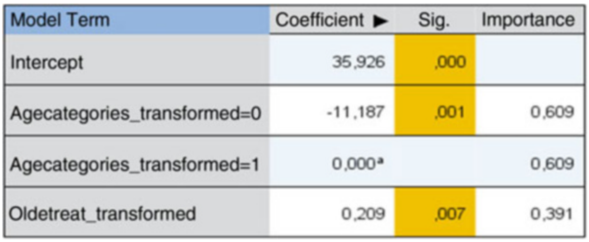
If the original field name is X, then the transformed field is displayed as (X_transformed). The original field is excluded from the anlyasis and the transformed field is included instead.

An interactive graph shows the predictors as lines with thicknesses corresponding to their predictive power and the outcome in the form of a histogram with its best fit Gaussian pattern. Both of the predictors are now statistically very significant with a correlation coefficient at $p < 0{,}0001$, and regression coefficients at p-values of respectively 0,001 and 0,007.

Coefficients
Target: newtreat

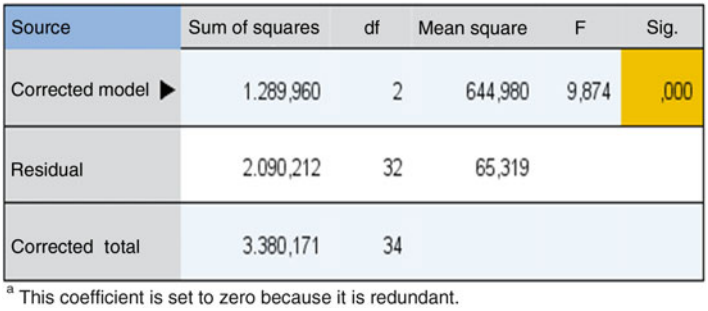


Coefficients
Target: newtreat



Model Term	Coefficient ▶	Sig.	Importance
Intercept	35,926	,000	
Agecategories_transformed=0	-11,187	,001	0,609
Agecategories_transformed=1	0,000[a]		0,609
Oldetreat_transformed	0,209	,007	0,391

Effects
Target: newtreat



Source	Sum of squares	df	Mean square	F	Sig.
Corrected model ▶	1.289,960	2	644,980	9,874	,000
Residual	2.090,212	32	65,319		
Corrected total	3.380,171	34			

[a] This coefficient is set to zero because it is redundant.

Returning to the data view of the original data file, we now observe that SPSS has provided a novel variable with values for the new treatment as predicted from statistical model employed. They are pretty close to the real outcome values.

Outcome	Predictor	Age category	Patient id	Predicted values
24,00	8,00	2,00	1,00	26,41
30,00	13,00	2,00	2,00	27,46
25,00	15,00	2,00	3,00	27,87
35,00	10,00	3,00	4,00	38,02
39,00	9,00	3,00	5,00	37,81
30,00	10,00	3,00	6,00	38,02
27,00	8,00	1,00	7,00	26,41
14,00	5,00	1,00	8,00	25,78
39,00	13,00	1,00	9,00	27,46
42,00	15,00	1,00	10,00	27,87

Outcome = new laxative
Predictor = old laxative

7 The Computer Teaches Itself to Make Predictions

The modeled regression coefficients are used to make predictions about future data using the Scoring Wizard and an XML (eXtended Markup Language) file (winRAR ZIP file) of the data file. Like random intercept models (see Chap. 45) and other generalized mixed linear models (see Chap. 12), automatic linear regression includes the possibility to make XML files from the analysis, that can subsequently be used for making outcome predictions in future patients. SPSS uses here software called winRAR ZIP files that are "shareware". This means that you pay a small fee and be registered if you wish to use it. Note that winRAR ZIP files have an archive file format consistent of compressed data used by Microsoft since 2006 for the purpose of filing XML files. They are only employable for a limited period of time like e.g. 40 days. Below the data of 9 future patients are given.

Newtreat	Oldtreat	Agecategory
	4,00	1,00
	13,00	1,00
	15,00	1,00
	15,00	1,00
	11,00	2,00
	80,00	2,00
	10,00	3,00
	18,00	2,00
	13,00	2,00

Enter the above data in a novel data file and command:

Utilities....click Scoring Wizard....click Browse....Open the appropriate folder with the XML file entitled "exportautomaticlinreg"....click on the latter and click Select....in Scoring Wizard double-click Next....mark Predicted Value....click Finish.

Newtreat	Oldtreat	Agecategory	Predicted new treat
	4,00	1,00	25,58
	13,00	1,00	27,46
	15,00	1,00	27,87
	15,00	1,00	27,87
	11,00	2,00	27,04
	80,00	2,00	41,46
	10,00	3,00	38,02
	18,00	2,00	28,50
	13,00	2,00	27,46

In the data file SPSS has provided the novel variable as requested. The first patient with only 4 stools per month on the old laxative and young of age will have over 25 stools on the new laxative.

8 Conclusion

SPSS' automatic linear regression can be helpful to obtain an improved precision of analysis of clinical trials and provided in the example given better statistics than traditional multiple linear regression did.

9 Note

More background theoretical and mathematical information of linear regression is available in Statistics applied to clinical studies 5th edition, Chap. 14, Linear regression basic approach, and Chap. 15, Linear regression for assessing precision confounding interaction, Chap. 18, Regression modeling for improved precision, pp 161–176, 177–185, 219–225, Springer Heidelberg Germany, 2012, from the same authors.