Chapter 6
Multiple Linear Regression (20 Patients)

1 General Purpose

In the Chap. 5 linear regression was reviewed with one (binary) predictor and one continuous outcome variable. However, not only a binary predictor like treatment modality, but also patient characteristics like age, gender, and comorbidity may be significant predictors of the outcome.

2 Schematic Overview of Type of Data File

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

3 Primary Scientific Question

Can multiple linear regression be applied to simultaneously assess the effects of multiple predictors on one outcome.

T.J. Cleophas, A.H. Zwinderman, *SPSS for Starters and 2nd Levelers*,
DOI 10.1007/978-3-319-20600-4_6

4 Data Example

In a parallel-group study patients are treated with either placebo or sleeping pill. The hours of sleep is the outcome. De concomitant predictors are age, gender, comorbidity.

Outcome	Treatment	Age	Gender	Comorbidity
6,00	,00	65,00	,00	1,00
7,10	,00	75,00	,00	1,00
8,10	,00	86,00	,00	,00
7,50	,00	74,00	,00	,00
6,40	,00	64,00	,00	1,00
7,90	,00	75,00	1,00	1,00
6,80	,00	65,00	1,00	1,00
6,60	,00	64,00	1,00	,00
7,30	,00	75,00	1,00	,00
5,60	,00	56,00	,00	,00
5,10	1,00	55,00	1,00	,00
8,00	1,00	85,00	,00	1,00
3,80	1,00	36,00	1,00	,00
4,40	1,00	47,00	,00	1,00
5,20	1,00	58,00	1,00	,00
5,40	1,00	56,00	,00	1,00
4,30	1,00	46,00	1,00	1,00
6,00	1,00	64,00	1,00	,00
3,70	1,00	33,00	1,00	,00
6,20	1,00	65,00	,00	1,00

Outcome = hours of sleep after treatment
Treatment = treatment modality (0 = placebo, 1 = sleeping pill)

5 Analysis, Multiple Linear Regression

The data file is entitled "chapter6linearregressionmultiple", and is in extras. springer.com. Open the data file in SPSS. For a linear regression 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.

Command:
Analyze....Regression....Linear....Dependent: treatment....Independent(s): group and age....click OK.

Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the estimate
1	,983[a]	,966	,962	,26684

[a]Predictors: (Constant), age, group

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	34,612	2	17,306	243,045	,000[b]
	Residual	1,210	17	,071		
	Total	35,822	19			

[a]Dependent variable: effect treatment
[b]Predictors: (Constant), age, group

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	,989	,366		2,702	,015
	group	−,411	,143	−,154	−2,878	,010
	age	,085	,005	,890	16,684	,000

[a]Dependent variable: effect treatment

In the above multiple regression two predictor variable have been entered: treatment modality and age. The tables resemble strongly the simple linear regression tables. The most important difference is the fact that now the effect of two x-variables is tested simultaneously. The R and the R-square values have gotten much larger, because two predictors, generally, given more information about the y-variable than a single one. R-square $= R^2 = 0.966 = 97\ \%$, meaning that, if you know the treatment modality and age of a subject from this sample, then you can predict the treatment effect (the numbers of sleeping hours) with 97 % certainty, and that you are still uncertain at the amount of 3 %.

The middle table takes into account the sample size, and tests whether this R-square value is significantly different from an R-square value of 0.0. The p-value equals 0.0001, which means it is true. We can conclude that both variables together significantly predict the treatment effect.

The bottom table now shows, instead of a single one, two calculated B-values (the regression coefficients of the two predictors). They behave like means, and can, therefore, be tested for their significance with two t-tests. Both of them are statistically very significant with p-values of 0.010 and 0.0001. This means that both B-values are significantly larger than 0, and that the corresponding predictors are independent determinants of the y-variable. The older you are, the better you will sleep, and the better the treatment, the better you will sleep.

We can now construct a regression equation for the purpose of making predictions for individual future patients.

$$y = a + b_1x_1 + b_2x_2$$

$$\text{Treatment effect} = 0.99 - 0.41{*}\text{group} + 0.085{*}\text{age}$$

with the sign * indicating the sign of multiplication. Thus, a patient of 75 years old with the sleeping pill will sleep for approximately 6.995 h. This is what you can predict with 97 % certainty.

Next we will perform a multiple regression with four predictor variables instead of two.

Command:

Analyze....Regression....Linear....Dependent: treatment....Independent: group, age, gender, comorbidity....click Statistics....mark Collinearity diagnostics....click Continue....click OK.

If you analyze several predictors simultaneously, then multicollinearity has to be tested prior to data analysis. Multicollinearity means that the x-variables correlate too strong with one another. For the assessment of it Tolerance and VIF (variance inflating factor) are convenient. Tolerance = lack of certainty = 1- R-square, where R is the linear correlation coefficient between 1 predictor and the remainder of the predictors. It should not be smaller than 0,20. VIF = 1/Tolerance should correspondingly be larger than 5. The underneath table is in the output sheets. It shows that the Tolerance and VIF values are OK. There is no collinearity, otherwise called multicollinearity, in this data file.

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	,727	,406		1,793	,093		
	Group	−,420	,143	−,157	−2,936	,010	,690	1,449
	Age	,087	,005	,912	16,283	,000	,629	1,591
	Male/female	,202	,138	,075	1,466	,163	,744	1,344
	Comorbidity	,075	,130	,028	,577	,573	,830	1,204

[a]Dependent variable: effect treatment

Also, in the output sheets are the underneath tables.

Model summary

Model	R	R square	Adjusted R square	Std. Error of the estimate
1	,985[a]	,970	,963	,26568

[a]Predictors: (Constant), comorbidity, group, male/female, age

ANOVA[a]

Model		Sum of Squares	df	Mean square	F	Sig.
1	Regression	34,763	4	8,691	123,128	,000[b]
	Residual	1,059	15	,071		
	Total	35,822	19			

[a] Dependent variable: effect treatment
[b] Predictors: (Constant), comorbidity, group, male/female, age

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	,727	,406		1,793	,093
	Group	−,420	,143	−,157	−2,936	,010
	Age	,087	,005	,912	16,283	,000
	Male/female	,202	,138	,075	1,466	,163
	Comorbidity	,075	,130	,028	,577	,573

[a] Dependent variable: effect treatment

They show that the overall r-value has only slightly risen, from 0,983 to 0,985. Obviously, the additional two predictors provided little additional predictive certainty about the predictive model. The overall test statistic (the F-value) even fell from 243,045 to 123,128. The four predictor-variables-model fitted the data less well, than did the two variables-model, probably due to some confounding or interaction (Chaps. 21 and 22). The coefficients table shows that the predictors, gender and comorbidity, were insignificant. They could, therefore, as well be skipped from the analysis without important loss of statistical power of this statistical model. Step down is a term used for skipping afterwards, step up is a term used for entering novel predictor variables one by one and immediately skipping them, if not statistically significant.

6 Conclusion

Linear regression can be used to assess whether predictor variables are closer to the outcome than could happen by chance. Multiple linear regression uses multidimensional modeling which means that multiple predictor variables have a zero correlation, and are, thus, statistically independent of one another.

Multiple linear regression is often used for exploratory purposes. This means, that in a data file of multiple variables the statistically significant independent predictors are searched for. Exploratory research is at risk of bias, because the data are often non-random or post-hoc, which means that the associations found may not be due to chance, but, rather, to real effect not controlled for. Nonetheless, it is interesting and often thought-provoking.

Additional purposes of multiple linear regression are (1) increasing the precision of your data, (2) assessing confounding and interacting mechanisms (Chaps. 21 and 22).

7 Note

More examples of the different purposes of linear regression analyses are given in Statistics applied to clinical studies 5th edition, Chaps. 14 and 15, Springer Heidelberg Germany, 2012, from the same authors. The assessment of exploratory research, enhancing data precision (improving the p-values), and confounding and interaction (Chaps. 22 and 23) are important purposes of linear regression modeling.