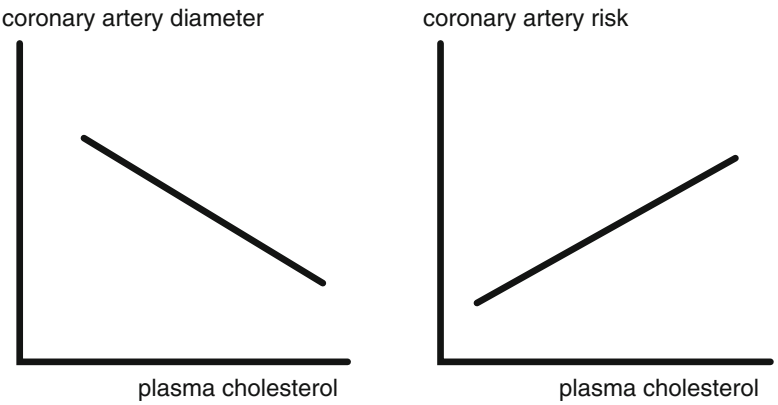
Chapter 5
Linear Regression (20 Patients)

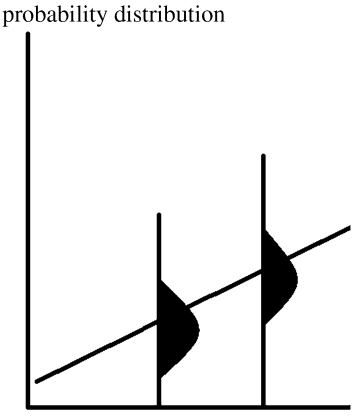
1 General Purpose



Similarly to unpaired t-tests and Mann-Whitney tests (Chap. 4), linear regression can be used to test whether there is a significant difference between two treatment modalities. To see how it works, picture the above linear regression of cholesterol levels and diameters of coronary arteries. It shows that the higher the cholesterol, the narrower the coronary arteries. Cholesterol levels are drawn on the x-axis, coronary diameters on the y-axis, and the best fit regression line about the data can be calculated. If coronary artery diameter coronary artery risk is measured for the y-axis, a positive correlation will be observed (right graph).

T.J. Cleophas, A.H. Zwinderman, *SPSS for Starters and 2nd Levelers*,
DOI 10.1007/978-3-319-20600-4_5

Instead of a continuous variable on the x-axis, a binary variable can be adequately used, such as two treatment modalities, e.g. a worse and better treatment. With hours of sleep on the y-axis, a nice linear regression analysis can be performed: the better the sleeping treatment, the larger the numbers of sleeping hours. The treatment modality is called the x-variable. Other terms for the x-variable are independent variable, exposure variable, and predictor variable. The hours of sleep is called the y-variable, otherwise called dependent or outcome variable. A limitation of linear regression is, that the outcomes of the parallel-groups are assumed to be normally distributed.



The above graph gives the assumed data patterns of a linear regression: the measured y-values are assumed to follow normal probability distributions around y-values

2 Schematic Overview of Type of Data File

Outcome	binary predictor
.	.
.	.
.	.
.	.
.	.
.	.
.	.
.	.
.	.

3 Primary Scientific Question

Is one treatment significantly more efficaceous than the other.

4 Data Example

In a parallel-group study of 20 patients 10 are treated with a sleeping pill, 10 with a placebo. The first 11 patients of the 20 patient data file is given underneath.

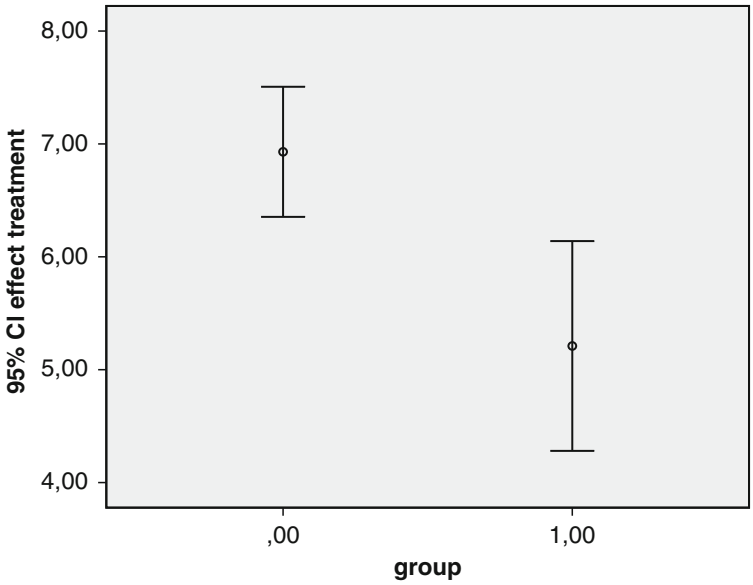
Outcome	Group
6,00	,00
7,10	,00
8,10	,00
7,50	,00
6,40	,00
7,90	,00
6,80	,00
6,60	,00
7,30	,00
5,60	,00
5,10	1,00

Group variable has 0 for placebo group, 1 for sleeping pill group
Outcome variable = hours of sleep after treatment

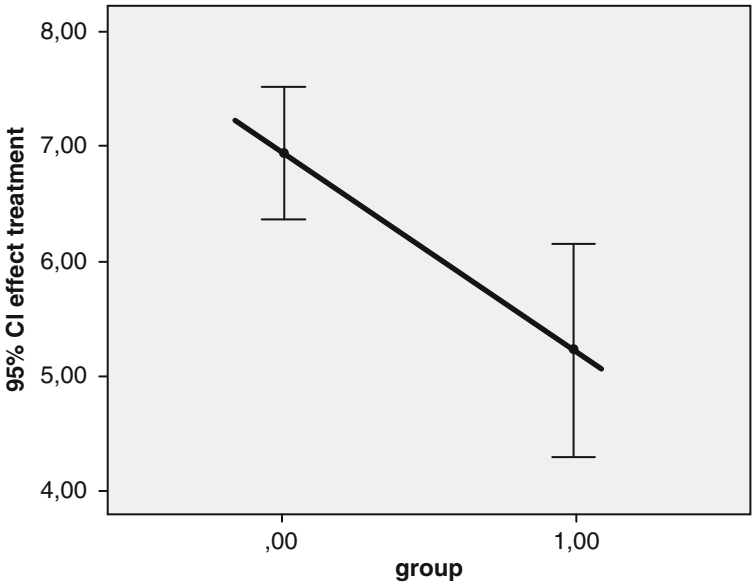
We will start with a graph of the data. The data file is entitled “chapter5linearregression”, and is in extras.springer.com. Start by opening the data file in SPSS.

Command:

Graphs....Legacy Dialogs....Error Bar....click Simple....mark Summaries for groups of cases....click Define....Variable: enter "effect treatment"....Category Axis: enter "group"....Bars Represent: choose "Confidence interval for means".... Level: choose 95%....click OK.



We used Google's Paint program to draw a regression line.



We will now try and statistically test, whether the data are closer to the regression line than could happen by chance. If so, that would mean that the treatment modalities are significantly different from one another, and that one treatment is significantly better than the other.

5 Analysis: Linear Regression

For a linear regression the module Regression is required. It consists of at least ten 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; enter treatment....Independent: enter group....click OK.

Model summary

Model	R	R square	Adjusted R square	Std. Error of the estimate
1	,643[a]	,413	,380	1,08089

[a]Predictors: (Constant), group

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	14,792	1	14,792	12,661	,002[b]
	Residual	21,030	18	1,168		
	Total	35,822	19			

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

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6,930	,342		20,274	,000
	group	−1,720	,483	−,643	−3,558	,002

[a]Dependent variable: effect treatment

The upper table shows the correlation coefficient ($R = 0.643 = 64\ \%$). The true r-value should not be 0,643, but rather −0,643. However, SPSS only reports positive r-values, as a measure for the strength of correlation. R-square $= R^2 = 0.413 = 41\ \%$, meaning that, if you know the treatment modality, you will be able to predict the treatment effect (hours of sleep) with 41 % certainty. You will, then, be uncertain with 59 % uncertainty.

The magnitude of R-square is important for making predictions. However, the size of the study sample is also important: with a sample of say three subjects little prediction is possible. This is, particularly, assessed in the middle table. It tests with analysis of variance (ANOVA) whether there is a significant correlation between the x and y-variables.

It does so by assessing whether the calculated R-square value is significantly different from an R-square value of 0. The answer is yes. The p-value equals 0.002, and, so, the treatment modality is a significant predictor of the treatment modality.

The bottom table shows the calculated B-value (the regression coefficient). The B-value is obtained by counting/ multiplying the individual data values, and it behaves in the regression model as a kind of mean result. Like many mean values from random data samples, this also means, that the B-value can be assumed to follow a Gaussian distribution, and that it can, therefore, be assessed with a t-test. The calculated t-value from these data is smaller than -1.96, namely -3.558, and, therefore, the p-value is <0.05. The interpretation of this finding is, approximately, the same as the interpretation of the R-square value: a significant B-value means that B is significantly smaller (or larger) than 0, and, thus, that the x-variable is a significant predictor of the y-variable. If you square the t-value, and compare it with the F-value of the ANOVA table, then you will observe that the values are identical. The two tests are, indeed, largely similar. One of the two tests is somewhat redundant.

6 Conclusion

The above figure shows that the sleeping scores after the placebo are generally larger than after the sleeping pill. The significant correlation between the treatment modality and the numbers of sleeping hours can be interpreted as a significant difference in treatment efficacy of the two treatment modalities.

7 Note

More examples 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.