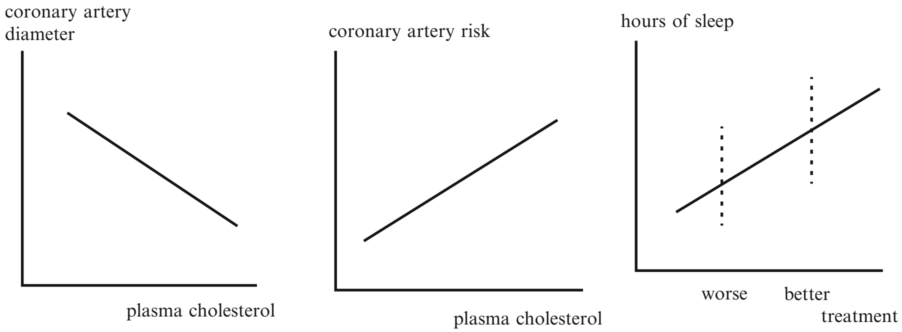
Chapter 8
Linear Regression (Regression Coefficient, Correlation Coefficient and Their Standard Errors)

1 General Purpose



Similarly to unpaired t-tests and Mann-Whitney tests (Chap. 7), linear regression can be used to test whether there is a significant difference between two unpaired 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 risk is measured on the y-axis instead of coronary artery diameter, then a positive correlation will be observed (right graph). 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

T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_8

the y-variable, otherwise called the dependent or outcome variable. This chapter is to show how a linear simple linear analysis works.

2 Schematic Overview of Type of Data File

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

3 Primary Scientific Question

Can linear regression be applied to demonstrate, whether, in an unpaired two group study, one treatment is significantly more efficaceous than the other treatment.

4 Data Example

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

Outcome	group (1, 2)
6.0	1
7.1	1
8.1	1
7.5	1
6.4	1
7.9	1
6.8	1

6.6	1
7.3	1
5.6	1
5.1	2
8.0	2
3.8	2
4.7	2
5.2	2
5.4	2
4.3	2
6.0	2
3.7	2
6.2	2

The group variable has 1 for sleeping pill group, 2 for the placebo.
The outcome variable is hours of sleep after treatment.

5 Analysis: Linear Regression

The equation of a linear regression model is given by

$$y = a + bx,$$

with y named the dependent variable and x the independent variable.

The line drawn from this linear function provides the best fit line for the data given, where y = socalled dependent, and x = independent variable, b = regression coefficient, a = intercept:

a and b from the equation y = a+bx can be calculated.

$$b = \text{regression coefficient} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$a = \text{intercept} = \bar{y}\text{-}b\bar{x}$$

r = correlation coefficient = another important determinant and looks a lot like b.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

r = measure for the strength of association between the y and x-data. The stronger the association, the better y predicts x, with +1 and -1 as respectively maximal and minimal r-values.

If b and r are statistically significantly larger than 0, then x is a significant predictor of y, and in the example given, this would mean, that there is a significant difference between the groups 1 and 2. One group performs better than the other, and, so, one treatment is better than the other.

6 Electronic Calculator for Linear Regression

We will use Electronic Calculator (see Chap. 1) for computations. First, we will calculate the b and r values.

```
Command:
click ON....click MODE....press 3....press 1....press SHIFT, MODE,
and again 1....press =....start entering the data.... [1,  6,0]....[1,
7,1]....[1,  8,1] etc....
```

In order to obtain the b value, press: shift, S-VAR, ►, ►, 2, = .
In order to obtain the r value, press: shift, S-VAR, ►, ►, 3, = .
The b value equals 1.70, the r value equals -0.643.
We wish to assess whether these two values are significantly larger than 0.

The standard error of $r = (1 - r^2) / \surd\ (n - 2)$

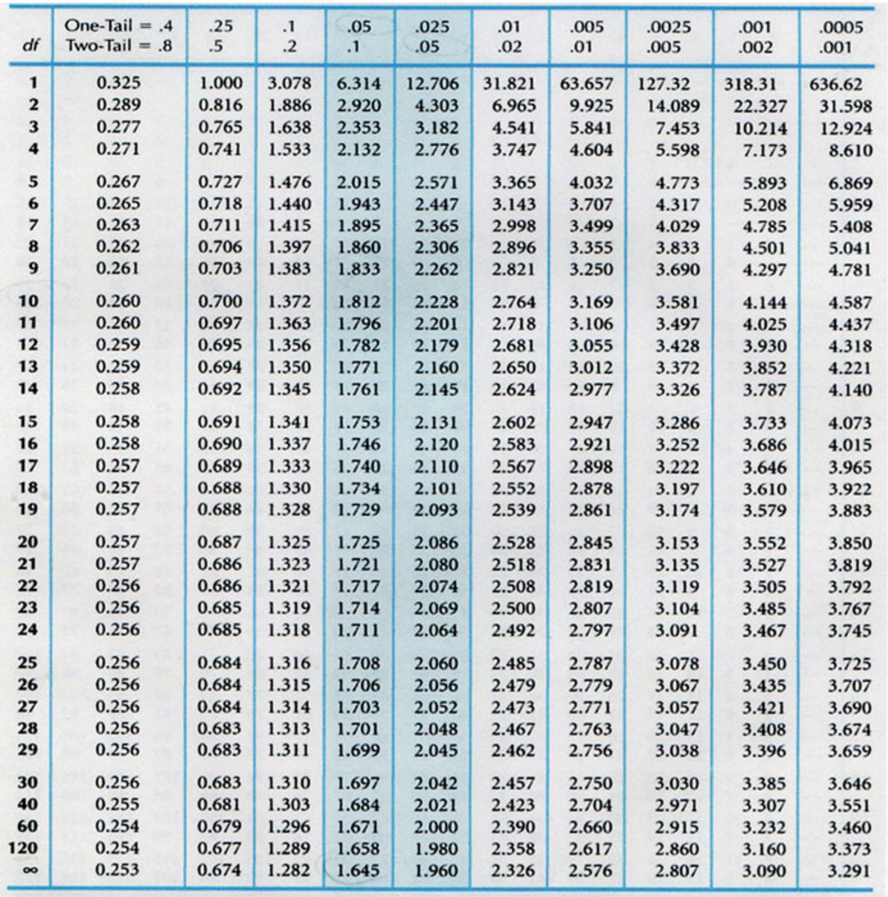
The r-value is a kind of summary-value of data, and follows a t-distribution, and can, thus, be tested with a *t*-test.

$t = \mid r / (\text{its standard error}) \mid$
$t = 0.643 \times 5.539$
$t = 3.56$

This value is much larger than 1.96, and, thus, r is significantly larger / smaller than 0. The t-value of b can be demonstrated to be equally 3.56.

7 T-Table

In the above study we have 20 outcome values and 2 groups. According to the underneath t-table, with 20-2 degrees of freedom (see 18th row of t-values), a t-value of 3.56 will be close to 3.610. This means, that the treatment 1 is better than the treatment 0 at a p-value close to 0.002. The t-table is briefly explained in the legends underneath the t-table. It is more fully explained in the Chaps. 4, 5, 6 and 7.



df	One-Tail = .4	.25	.1	.05	.025	.01	.005	.0025	.001	.0005
	Two-Tail = .8	.5	.2	.1	.05	.02	.01	.005	.002	.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The t-table has a left-end column giving degrees of freedom (≈ sample sizes), and two top rows with p-values (areas under the curve = p - values), one-tail meaning that only one end of the curve, two-tail meaning that both ends are assessed simulataneously. The t-table is, furthermore, full of t-values, that, with ∞ degrees of freedom, are equal to z-values (Chap. 36). The t-values are to be understood as mean results of studies, but not expressed in mmol/l, kilograms, but in so-called SEM-units (Standard error of the mean units), that are obtained by dividing your mean result by its own standard error. With many degrees of freedom (large samples) the curve will be a little bit narrower, and more in agreement with nature.

8 Conclusion

We can conclude that the correlation and regression coefficients, r and b, are very significant with p-values close to 0.002. This demonstrates that the sleeping scores after active treatment are generally larger than after placebo treatment. 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. An interesting thing about linear regression is that the linear regression equation can be used for estimating from future x-values the best fit predictions of y-values. In our example we only have two x-values, but if you have more of them, the size of your dependent variable can pretty well be predicted from measured x-values, particularly, if your level of statistical significance is very high (r values close to +1 or -1). R values > 95 % are, actually, applied for validating quantitative diagnostic tests (see also Chap. 25).

9 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.