Example 4: Brain and Weight Data

Table 7 presents the brain weight (in grams) and the body weight (in kilograms) of 28 animals. (This sample was taken from larger data sets in Weisberg 1980 and Jerison 1973.) It is to be investigated whether a larger brain is required to govern a heavier body.

A clear picture of the relationship between the logarithms (to the base 10) of these measurements is shown in Figure 7. This logarithmic

Table 7. Body and Brain Weight for 28 Animals

Index (i)	Species	Body Weight[a] (x_i)	Brain Weight[b] (y_i)
1	Mountain beaver	1.350	8.100
2	Cow	465.000	423.000
3	Gray wolf	36.330	119.500
4	Goat	27.660	115.000
5	Guinea pig	1.040	5.500
6	Diplodocus	11700.000	50.000
7	Asian elephant	2547.000	4603.000
8	Donkey	187.100	419.000
9	Horse	521.000	655.000
10	Potar monkey	10.000	115.000
11	Cat	3.300	25.600
12	Giraffe	529.000	680.000
13	Gorilla	207.000	406.000
14	Human	62.000	1320.000
15	African elephant	6654.000	5712.000
16	Triceratops	9400.000	70.000
17	Rhesus monkey	6.800	179.000
18	Kangaroo	35.000	56.000
19	Hamster	0.120	1.000
20	Mouse	0.023	0.400
21	Rabbit	2.500	12.100
22	Sheep	55.500	175.000
23	Jaguar	100.000	157.000
24	Chimpanzee	52.160	440.000
25	Brachiosaurus	87000.000	154.500
26	Rat	0.280	1.900
27	Mole	0.122	3.000
28	Pig	192.000	180.000

[a]In kilograms.
[b]In grams.
Source: Weisberg (1980) and Jerison (1973).

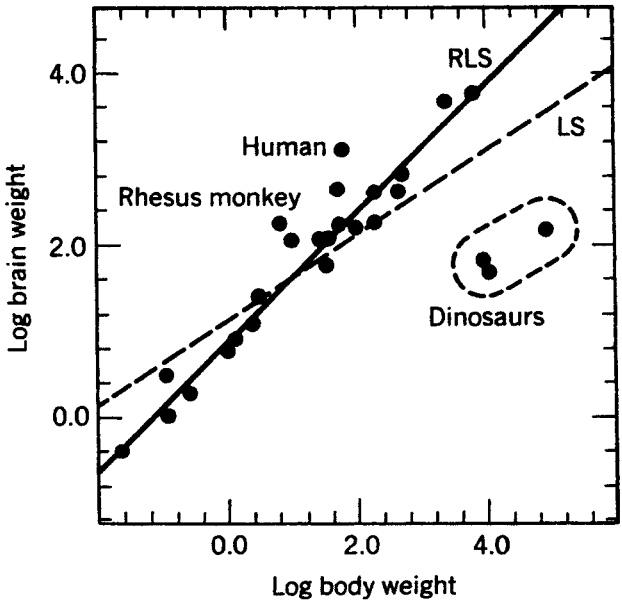

Figure 7. Logarithmic brain weight versus logarithmic body weight for 28 animals with LS (dashed line) and RLS fit (solid line).

transformation was necessary because plotting the original measurements would fail to represent either the smaller or the larger measurements. Indeed, both original variables range over several orders of magnitude. A linear fit to this transformed data would be equivalent to a relationship of the form

$$\hat{y} = \hat{\theta}_2' x^{\hat{\theta}_1}$$

between brain weight (y) and body weight (x). Looking at Figure 7, it seems that this transformation makes things more linear. Another important advantage of the log scale is that the heteroscedasticity disappears.

The LS fit is given by

$$\log \hat{y} = 0.49601 \log x + 1.10957$$

(dashed line in Figure 7). The standard error associated with the slope equals 0.0782, and that of the intercept term is 0.1794. In Section 3, we explained how to construct a confidence interval for the unknown regression parameters. For the present example, $n = 28$ and $p = 2$, so one has to

use the 97.5% quantile of the t-distribution with 26 degrees of freedom, which equals 2.0555. Using the LS results, a 95% confidence interval for the slope is given by [0.3353; 0.6567]. The RLS yields the solid line in Figure 7, which is a fit with a steeper slope:

$$\log \hat{y} = 0.75092 \log x + 0.86914 .$$

The slope estimated by the RLS technique even falls outside the 95% confidence interval associated with the LS fit! The standard error of the regression coefficients in RLS is reduced remarkably as compared with

Table 8. Standardized LS and RLS Residuals for the Brain and Body Weight Data

Index	Species	Standardized LS Residuals	Standardized RLS Residuals	w_i
1	Mountain beaver	−0.40	−0.27	1
2	Cow	0.29	−1.13	1
3	Gray wolf	0.29	0.17	1
4	Goat	0.36	0.50	1
5	Guinea pig	−0.57	−0.65	1
6	Diplodocus	−2.15	−10.19	0
7	Asian elephant	1.30	1.08	1
8	Donkey	0.58	0.21	1
9	Horse	0.54	−0.43	1
10	Potar monkey	0.68	2.02	1
11	Cat	0.06	0.68	1
12	Giraffe	0.56	−0.37	1
13	Gorilla	0.53	0.00	1
14	Human	1.69	4.15	0
15	African elephant	1.13	0.08	1
16	Triceratops	−1.86	−9.20	0
17	Rhesus monkey	1.10	3.47	0
18	Kangaroo	−0.19	−1.29	1
19	Hamster	−0.98	−0.81	1
20	Mouse	−1.04	−0.17	1
21	Rabbit	−0.34	−0.39	1
22	Sheep	0.40	0.29	1
23	Jaguar	0.14	−0.80	1
24	Chimpanzee	1.02	2.22	1
25	Brachiosaurus	−2.06	−10.94	0
26	Rat	−0.84	−0.80	1
27	Mole	−0.27	1.35	1
28	Pig	0.02	−1.50	1

the LS, namely 0.0318 for the slope and 0.0618 for the constant term. A 95% confidence interval for the unknown slope is now given by [0.6848; 0.8171], which is narrower than the interval coming from LS. The t-values associated with the RLS regression coefficients are very large, which implies that the slope and intercept are significantly different from zero. Moreover, the determination coefficient R^2, which is a summary measure for overall goodness of fit, increases from 0.608 for LS to 0.964 for RLS. This example shows that not only the LS regression coefficients, but also the whole LS inference, may become doubtful in the presence of outliers.

Table 8 lists the standardized LS and RLS residuals and the w_i determined on the basis of the LMS. From the RLS, it is easy to detect unusual observations and to give them special consideration. Indeed, looking at the five cases with zero w_i, one can easily understand why they have to be considered as outlying. The most severe (and highly negative) RLS residuals are those of cases 6, 16, and 25, which are responsible for the low slope of the LS fit. These are three dinosaurs, each of which possessed a small brain as compared with a heavy body. In this respect they contrast with the mammals which make up the rest of the data set. The LMS regression also produced a zero weight for cases 14 and 17, namely the human and the rhesus monkey. For them, the actual brain weight is higher than that predicted by the linear model. Unlike the dinosaurs, their residuals are therefore positive. Concluding, one could say that dinosaurs, humans, and rhesus monkeys do not obey the same trend as the one followed by the majority of the data.

5. AN ILLUSTRATION OF THE EXACT FIT PROPERTY

The phrase "exact fit" stands for situations where a large percentage of the observations fits some linear equation exactly. For example, in simple regression this happens when the majority of the data lie exactly on a straight line. In such a case a robust regression method should recover that line. At an Oberwolfach Meeting, Donoho (1984) called this the *exact fit property*. For instance, the repeated median satisfies this property (Siegel 1982), as well as the LMS (Rousseeuw 1984). When at least $n - [n/2] + 1$ of the observations lie on the same line, then the equation of this line will be the LMS solution. More details on the exact fit property and its relation to the breakdown point can be found in Section 4 of Chapter 3.

The data in Table 9 come from Emerson and Hoaglin (1983, p. 139). They were devised by A. Siegel as a counterexample for the resistant line estimator (which will be briefly discussed in Section 7). Looking at the