Regression and correlation examples

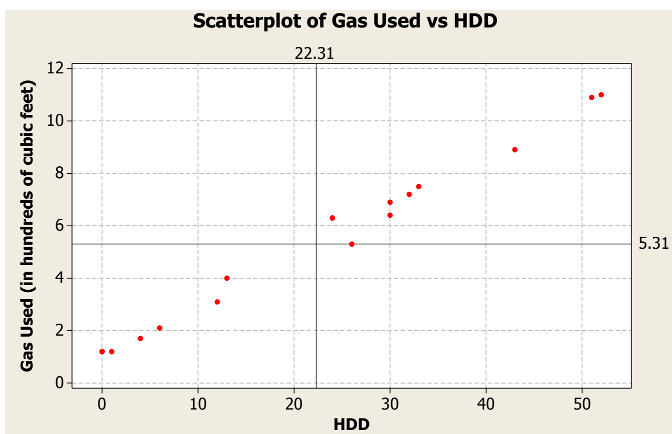
Example 1 Data is gathered to explore the relationship between outside temperature and the amount of gas used to heat a specific house. A standard measure of outside temperature used for this purpose is the *heating degree day (HDD)*. For a given day, the value of HDD is the difference between 65°F and the average outside temperature for that day. So, for a day on which the average outside temperature is 49°F , we have 16 heating degree-days. (The reference temperature of 65°F is used because a typical house needs no heating when the average outside temperature is 65°F .) The language here is a bit awkward since "heating degree-day" refers to both the variable and the unit used for the variable. We'll denote the variable HDD and the unit hdd. So, for the example we have HDD=16 hdd.

The table below gives data for HDD and gas usage (in hundreds of cubic feet) for a specific house. Here are summary statistics for the individual distributions:

H=HDD	$\bar{h} = 22.31$ hdd	$s_h = 17.74$ hdd	
G=Gas Used	$\bar{g} = 5.306$	$s_g = 3.368$	(both in hundred cubic feet)

The scatterplot below includes a vertical line for the HDD mean and a horizontal line for the Gas Used mean. For these two variables, the correlation is $r = 0.995$.

HDD	Gas Used
24	6.3
51	10.9
43	8.9
33	7.5
26	5.3
13	4.0
4	1.7
0	1.2
0	1.2
1	1.2
6	2.1
12	3.1
30	6.4
32	7.2
52	11.0
30	6.9



1. Compute the slope and intercept of the least-squares regression line for this data. Write down a formula for the least-squares regression line. Use this to plot the least-squares regression line on the scatterplot given above.

2. Use the least-squares regression line to predict the amount of gas used on a day when the average outside temperature is 45 °F .

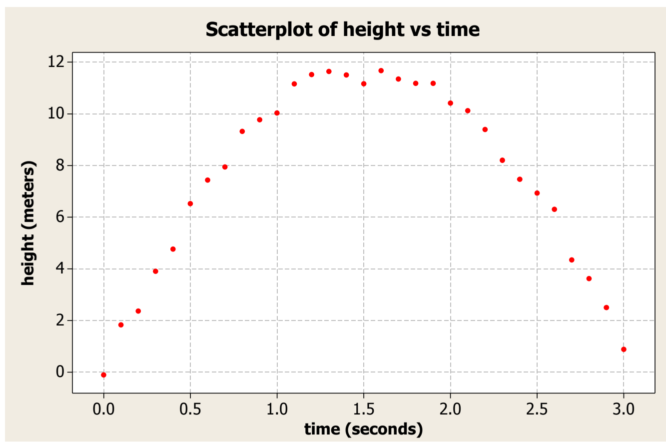
Example 2 A physics student does an experiment that involves launching a ball straight up and then measuring the height of the ball every tenth of a second. The table below shows the data with time t given in seconds and height h given in meters. For the time data distribution, the mean is $\bar{t} = 1.50$ inches and the standard deviation is $s_t = 0.909$ seconds. For the height data distribution, the mean is $\bar{h} = 7.626$ meters and the standard deviation is $s_h = 3.663$ meters. The correlation for these two variables is $r = 0.072$. With these values, we can calculate the slope and intercept values for the least-squares regression line as

$$b = r\frac{s_h}{s_t} = 0.072 \times \frac{3.663 \text{ m}}{0.909 \text{ s}} = 0.290 \text{ m/s}$$

and

$$a = \bar{h} - b\bar{t} = 7.626 \text{ m} - 0.290 \text{ m/s} \times 1.50 \text{ s} = 7.191 \text{ m}.$$

time (s)	height (m)
0.0	-0.10
0.1	1.83
0.2	2.37
0.3	3.91
0.4	4.77
0.5	6.52
0.6	7.44
0.7	7.95
0.8	9.32
0.9	9.77
1.0	10.04
1.1	11.16
1.2	11.52
1.3	11.64
1.4	11.50
1.5	11.16
1.6	11.67
1.7	11.35
1.8	11.18
1.9	11.18
2.0	10.42
2.1	10.13
2.2	9.40
2.3	8.20
2.4	7.47
2.5	6.93
2.6	6.30
2.7	4.35
2.8	3.62
2.9	2.51
3.0	0.88



1. Describe the association (form, direction if relevant, strength) between time and height seen in this scatterplot.

2. What does the correlation value of $r = 0.072$ tell us about this association?

3. Write down the formula for the least-squares regression line and plot this line on the scatterplot. How useful is the regression line as a predictor for heights?