3 - regression

~~11.1.~~ The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

$b_1 = (0.86)\left(\frac{0.01}{35}\right)$ $\quad G(x) = b_1 x + b_0$

$b_1 = 0.000245714\overline{28}$ $\quad G(400) = 0.107326$

$b_0 = 0.04 - (b_1)(126)$

$b_0 = 0.00904$

11.2. The following statistics were obtained from a sample of size $n = 75$:
- the predictor variable X has mean 32.2, variance 6.4;
- the response variable Y has mean 8.4, variance 2.8; and
- the sample covariance between X and Y is 3.6.

(a) Estimate the linear regression equation predicting Y based on X.

correlation $= \frac{3.6}{\sqrt{(64)(28)}} = 0.85$

$b_1 = 0.85\left(\frac{2.8}{\sqrt{6.4}}\right)$

$b_1 = 0.56$

$b_0 = (8.4) - (0.56)(32.2)$

$b_0 = -9.7$

$G(x) = 0.56x - 9.7$

11.3. At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$.

(a) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage. What do the obtained slope and intercept mean in this situation?

$b_1 = -0.17\left(\frac{3.4}{14,634}\right)$

$b_1 = -0.0395$

$b_0 = 23.8 - b_1(24,598)$

$b_0 = 24.775$

$G(x) = b_0 + b_1 x$

11.8. Anton wants to know if there is a relation between the number of hours he spends preparing for his weekly quiz and the grade he receives on it. He keeps records for 10 weeks.

It turns out that on the average, he spends 3.6 hours a week preparing for the quiz, with the standard deviation of 0.5 hours. His average grade is 82 (out of 100), with the standard deviation of 14. The correlation between the two variables is $r = 0.62$.

(a) Find the equation of the regression line predicting the quiz grade based on the time spent on preparation.

$b_1 = 0.62\left(\frac{14}{0.5}\right) = 17.36$

$b_0 = 82 - b_1(3.6) = 19.504$

$G(x) = 19.504 + (17.36)x$

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287	2015	?
1970	3712	1995	5700	2020	?

Example 11.3 (WORLD POPULATION). In Example 11.1, x_i is the year, and y_i is the world population during that year. To estimate the regression line in Figure 11.1, we compute

$$\bar{x} = 1980; \quad \bar{y} = 4558.1;$$

$$\begin{aligned} S_{xx} &= (1950 - \bar{x})^2 + \ldots + (2010 - \bar{x})^2 = 4550; \\ S_{xy} &= (1950 - \bar{x})(2558 - \bar{y}) + \ldots + (2010 - \bar{x})(6864 - \bar{y}) = 337250. \end{aligned}$$

Then

$$\begin{aligned} b_1 &= S_{xy}/S_{xx} = 74.1 \\ b_0 &= \bar{y} - b_1\bar{x} = -142201. \end{aligned}$$

if correlation coefficient r is given

$b_1 = r\left(\frac{s_y}{s_x}\right)$

$b_0 = \bar{y} - b_1\bar{x}$

The estimated regression line is

$$\widehat{G}(x) = b_0 + b_1 x = \underline{-142201 + 74.1x}.$$

We conclude that the world population grows at the average rate of 74.1 million every year.

We can use the obtained equation to predict the future growth of the world population. Regression predictions for years 2015 and 2020 are

$$\widehat{G}(2015) = b_0 + 2015\, b_1 = \underline{7152 \text{ million people}}$$
$$\widehat{G}(2020) = b_0 + 2020\, b_1 = \underline{7523 \text{ million people}}$$

◊

if correlation coefficient is given

$b_1 = (\text{corr. coef.})\left(\frac{s_y}{s_x}\right)$ ← std dev

$b_0 = \bar{y} - b_1 \bar{x}$ $\quad b_0 + b_1 x$

!! if covariance is given $\quad$ correlation $= \frac{\text{covariance}}{(\sigma x)(\sigma y)}$ → std dev of x, y

predicting y based on x (like y time)

The time it takes to copy a file from one disk to another always depends on the file size. Assume that you have copied 50 files, with the average size of 10 MB and the standard deviation of 5 MB. The average time to carry out the copying was 200 ms with the standard deviation of 100 ms. The correlation coefficient between the time and the size was 0.8.

Based on this data, fit a linear regression model and predict the time it will take to transmit a 90 MB file (in ms).

$b_1 = 0.8\left(\frac{100}{5}\right) = 16$

$b_0 = 200 - 16 \cdot 10 = 40$

$G(x) = 40 + 16x$

$G(90) = 1480$

Answer: 1480 ✔