Chapter 5

Descriptive linear regression analysis

For strongly correlating sample data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ for some metrically scaled 2-D variable (X, Y), i.e., when $0.6 < |r| \leq 1.0$, it is meaningful to construct a mathematical model of the linear quantitative statistical association so diagnosed. The standard method to realise such a model is due to the German mathematician and astronomer Carl Friedrich Gauss (1777-1855) and is known by the name of descriptive linear regression analysis, cf. Gauss (1809). We here restrict our attention to the case of simple linear regression which involves data for two variables only. To be determined is a best-fit linear model to given bivariate metrical data $\{(x_i, y_i)\}_{i=1,\ldots,n}$. The linear model in question can be expressed in mathematical terms by

$$\hat{y} = a + bx, \tag{5.1}$$

with unknown regression coefficients y-intercept a and slope b. Gauss' method works as follows.

5.1 Method of least squares

At first, one has to make a choice: assign X the status of an independent variable, and Y the status of a dependent variable (or vice versa; usually this freedom of choice does exist, unless one is testing a specific functional relationship $y = f(x)$). Then, considering the measured values x_i for X as fixed, to be minimised for the Y-data is the sum of the squared vertical deviations of the measured values y_i from the model values $\hat{y}_i = a + bx_i$ associated with an arbitrary straight line through the cloud of data points $\{(x_i; y_i)\}_{i=1,\ldots,n}$ in a scatter plot, i.e., the sum

$$S(a, b) := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{5.2}$$

$S(a, b)$ constitutes a (non-negative) real-valued function of two variables a and b. Hence, determining its (local) minimum values entails satisfying

(i) the necessary condition of simultaneously vanishing first partial derivatives

$$0 = \frac{\partial S(a, b)}{\partial a};\ 0 = \frac{\partial S(a, b)}{\partial b}, \tag{5.3}$$

this yields a well-determined (2×2) system of linear equations for the unknowns a and b, and

(ii) the sufficient condition of a positive definite Hessian matrix $H(a, b)$ of second partial derivatives

$$H(a, b) := \begin{pmatrix} \dfrac{\partial^2 S(a, b)}{\partial a^2} & \dfrac{\partial^2 S(a, b)}{\partial a \partial b} \\ \dfrac{\partial^2 S(a, b)}{\partial b \partial a} & \dfrac{\partial^2 S(a, b)}{\partial b^2} \end{pmatrix} \tag{5.4}$$

$H(a, b)$ is referred to as positive definite when all of its eigenvalues are positive.

Exercise 1

1. Find all second derivatives $\frac{\partial^2 z}{\partial x^2}$, $\frac{\partial^2 z}{\partial y \partial x}$, $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y^2}$ of the function

$$z = f(x, y) = 2x^2 - 2xy - 16x + 5y^2 + 2y + 34.$$

2. Calculate the Hessian matrix and its determinant. Is it positive or negative?

3. Compute the Hessian matrix at $(x, y) = (0, 0)$ and $(x, y) = (-1, 1)$ and check your results with `R`.

Exercise 2 A T-shirt shop carries two competing shirts, one endorsed by Michael Jordan and the other by Shaq O'Neal. The owner of the store can obtain both at a cost of \$2 per shirt and estimates that if Jordan shirts are sold for x dollars apiece and O'Neal shirts for y dollars apiece, consumers will buy approximately $40 - 50x + 40y$ Jordan shirts and $20 + 60x - 70y$ O'Neil shirts each day.

1. Express as functions of x and y:
   (a) the revenue from selling Jordan shirts,
   (b) the revenue from selling O'Neal shirts
   (c) the costs for shirts and
   (d) the overall profit.
2. Find the critical point of the profit function.
3. How should the owner price the shirts in order to generate the largest possible profit?
4. Calculate the Hessian matrix for this problem and its determinant. Is the solution found in 2. indeed an absolute maximum?

5.2 Empirical regression line

It is a fairly straightforward algebraic exercise to show that the values of the unknowns a and b which determine a unique global minimum of $S(a, b)$ amount to

$$b = \frac{s_Y}{s_X} r \quad ; \quad a = \overline{y} - b\overline{x} \tag{5.5}$$

These values are referred to as the least square estimators for a and b. Note that they are exclusively expressible in terms of familiar univariate and bivariate measures characterising the joint distribution of X and Y.
With the solutions a and b of Eq. (5.5), the resultant best-fit linear model is thus

$$\hat{y} = \overline{y} + \frac{s_Y}{s_X} r(x - \overline{x}) \tag{5.6}$$

It may be employed for the purpose of generating intrapolating predictions of the kind $x \mapsto \hat{y}$ for x-values confined to the interval $[x_{(1)}, x_{(n)}]$.

Exercise 3 Use the method of least squares to fit a straight line to the $n = 5$ data points given in Table below.

x	-2	-1	0	1	2
y	0	0	1	1	3

Use first formulas (5.5) and (5.6) and next, the appropriate commands in `R`.

Exercise 4 Auditors are often required to compare the audited (or current) value of an inventory item with the book (or listed) value. If a company is keeping its inventory and books up to date, there should be a strong linear relationship between the audited and book values. A company sampled ten inventory items and obtained the audited and book values given in the accompanying table. Fit the model $Y = \beta_0 + \beta_1 x + \varepsilon$ to these data.

Item	Audit Value(y_i)	Book Value(x_i)
1	9	10
2	14	12
3	7	9
4	29	27
5	45	47
6	109	112
7	40	36
8	238	241
9	60	59
10	170	167

1. What is your estimate for the expected change in audited value for a one-unit change in book value?
2. If the book value is $x = 100$, what would you use to estimate the audited value?

5.3 Coefficient of determination

The quality of any particular simple linear regression model, its goodness-of-the-fit, can be quantified by means of the coefficient of determination B (metr). This measure is derived starting from the algebraic identity

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5.7}$$

which, upon conveniently re-arranging, leads to defining a quantity

$$B := \frac{\sum_{i=1}^{n}(yi - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5.8}$$

with range $0 \leq B \leq 1$. For a perfect fit $B = 1$, while for no fit $B = 0$. The coefficient of determination provides a descriptive measure for the proportion of variability of Y in a data set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ that can be accounted for as due to the association with X via the simple linear regression model. Note that in simple linear regression it holds that

$$B = r^2 \tag{5.9}$$

Exercise 5 Find the coefficient of determination for the simple linear regression model of the data set `faithful`.

This concludes Part I of these lecture notes: the discussion on descriptive statistical methods of data analysis. To set the stage for the application of inferential statistical methods in Part III, we now turn to review the elementary concepts underlying probability theory.