2 Simple Linear Regression (Exercise 1)

A way to analyze an available data set is to perform linear regression. The most common form of linear regression is known as least squares fitting, whose aim is to fit a polynomial curve to the data such that the sum of the squares of the distance from the data points to the line is minimized.

2.1 Short Theory

Simple linear regression is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible. The adjective simple refers to the fact that this regression is one of the simplest in statistics. The slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass $(\bar{x}, \bar{y})$ of the data points.

Suppose there are n data points $(x_i, y_i), i = 1, 2, \ldots, n$. The function that describes $x$ and $y$ is $y_i = a + \beta x_i + \epsilon_i$. The goal is to find the equation of the straight line $y = a + \beta x$ that would provide a “best” fit for the data points. Here the ”best” will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals of the linear regression model. In other words, $a$(the y-intercept) and $\beta$ (the slope) solve the following minimization problem:

$$\hat{a}, \hat{\beta} = \arg\min_{a, \beta} \sum_{i=1}^{N} (y_i - a - \beta x_i)^2$$

(1)

By using either calculus, the geometry of inner product spaces, or simply expanding to get a quadratic expression in $a$ and $\beta$, it can be shown that the values of $a$ and $\beta$ that minimize the expression above are:

$$\hat{\beta} = \frac{\text{cov}[x, y]}{\text{var}[x]}$$

(2)

$$\hat{a} = \bar{y} - \hat{\beta} \bar{x}$$

(3)

A more thorough description of the simple linear regression theory can be found in any of the following sources:

- The lecture slides.
- Section 3.1, Ch. 3 from the book “Pattern recognition and machine learning” of Bishop and Nasrabadi.

2.2 Matlab simulation

Often mathematical models of experimental or physiological systems are developed to form the basis of a measurement technique. For example, it is not possible to measure cardiac output directly, but we may be able to infer it from analysis of a mathematical model of respiratory gas exchange. For
such models we are often solving an inverse problem of parameter estimation (the cardiac output),
where the parameter of interest is embedded (somewhere) within the mathematical model. In these
circumstances, it is often advisable to introduce simulated experimental error into the system to test
the robustness of the recovery procedure. The following is one of the simplest possible examples,
where the parameters form part of a linear model.

The following activity will lead you through generating some experimental data adding artifi-
cicial noise and then performing least squares estimation to try to elucidate the underlying model
parameters.

1. Generate and plot the data (x and y values) for a simple straight line of the form \( y = a + \beta x \)
where \( a = 2 \) and \( \beta = 1 \) are constants and \( x \in [0, 1] \). You should calculate the value of \( y \) at 21
evenly spaced points between 0 and 1.

2. Generate random errors from a normal distribution, with zero mean \( \mu = 0 \) and standard devi-
ation \( \sigma = 0.1 \), using Matlab’s built-in commands, and add these to each of the \( y \) coordinates
in step 1. Plot these values as points on the same graph as in step 1. What happens if \( \mu \) is
non-zero?

3. Find in the Matlab help the syntax and the use of the function \texttt{polyfit}. Then, using the
command \texttt{polyfit} perform a linear regression through the data generated in step 2, and plot
the regression line on the same plot.

4. Repeat steps 2 and 3 for differ-
ing values of the standard deviation, from \( \sigma = 0 \) to \( \sigma = 1 \), in
steps of 0.2, and plot the regression lines on the same graph. \textit{You should use a loop to do
this.}

5. For a suitable value of the standard deviation (\( \sigma = 0.1 \)), repeat steps 2 and 3 1000 times
to investigate the statistical properties of the regression coefficients. To do this you should
calculate the mean and standard deviation of the regression coefficients and compare them to
the theoretical values of \( \mu = 1, \sigma \approx 0.042 \) for \( a \), and \( \mu = 2, \sigma \approx 0.071 \) for \( \beta \). The estimators
should also be normally distributed. You do not need to know how these were calculated, but
note that the estimators are unbiased, as the mean values of the estimators of the parameters
are the parameters themselves.

6. Explain why the above can be used to simulate the effects of random experimental error.