Linear regression of the outputs of a device and optimizing the linear model using linear programming

Dristy Parveg
Mohsin Khan
Finland

March 23, 2017

1 Introduction

It has been observed that the device has two positive outputs $y_1, y_2 \in \mathbb{R}$. We have learnt that y_1 is a function of 5 positive variables $x_1, x_2, x_3, x_4, x_5 \in \mathbb{R}$ and we name this function F. We also have learnt that y_2 is a function of 3 variables x_1, x_2, x_3 and we name this function G. It should be noted here that there are constraints involved about the values of the inputs:

$$\begin{aligned} 2.5 \le x_1 &\le 11 \\ 1.5 \le x_2 &\le 6 \\ 1.25 \le x_3 &\le 2.4 \\ 4 \le x_4 &\le 11 \\ 10 \le x_5 &\le 20 \end{aligned}$$

However, we don't exactly know the definitions of F and G. Fortunately, we can sample these two functions at different input points through experiments as discussed in the previous sections. We have results of 18 experiments. We are interested to optimize the functions F and G. We are interested to optimize the functions in few alternative ways. The candidates are as follows:

1. *Maximize* $F - G$: Finding values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $F(a_1, a_2, a_3, a_4, a_5) - G(a_1, a_2, a_3)$ becomes as small as possible.

2. *Minimize* $|\frac{F}{G} - b|$: Finding values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $\frac{F(a_1,a_2,a_3,a_4,a_5)}{G(a_1,a_2,a_3)}$ becomes as close as possible to a certain constant b

3. *Minimize* $|F - b_1| + |G - b_2|$: Finding values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $F(a_1, a_2, a_3, a_4, a_5)$ and $G(a_1, a_2, a_3)$ goes as close as possible to constant values b_1 and b_2 respectively.

Let $f_i : \mathbb{R} \to \mathbb{R}$ be a function that takes the variable x as input while the other input variables x_j for $j \in \{1, ...5\} \setminus i$ of F are kept at constant values i.e.,

$$f_i(x|a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_5) \quad = \quad F(a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_5)$$

We define the function $g_i : \mathbb{R} \to \mathbb{R}$ in the same way so that

$$g_i(x|a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_5) \quad = \quad G(a_1, ..., a_{i-1}, x, a_{i+1}, ..., a_5)$$

From the result of the experiments, presented in previous section, it is evident that f_i and g_i are almost, if not exactly, linear functions. We are interested to find models $f : \mathbb{R}^5 \to \mathbb{R}$, $g : \mathbb{R}^3 \to \mathbb{R}$ where $f(x_1, x_2, x_3, x_4, x_5) \approx G(x_1, x_2, x_3, x_4, x_5)$ and $g(x_1, x_2, x_3) \approx G(x_1, x_2, x_3)$.

2 Linear Regression:

Motivated by the observation that f_i and g_i for $i, j \in \{1, ...5\}$ are almost linear, we make an educated guess that both F and G are approximately linear. We do not provide an analytical argument to justify the linearity of F and G. Instead, we use linear regression on the 18 experimental results and to the linear models f and g. We have used the statistical programming tool R to do the linear regression.

The experimental results are kept in a file named experiment.txt. Please check the attachment. Then we obtain the following models f and g and their errors by running the source file model.R. The source file is attached. The command to run the file is R is: source("model.R")

$$\begin{aligned} f(x_1, x_2, x_3, x_4, x_5) &= -7.2667 \cdot x_1 + 0.3724 \cdot x_2 + 29.1759 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 78.2264 \\ g(x_1, x_2, x_3) &= -5.477 \cdot x_1 + 4.640 \cdot x_2 + 4.290 \cdot x_3 + 24.343 \end{aligned}$$

For an input $(x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5)$, the error percentages of the models f and g are as follow

$$\frac{|f(a_1, a_2, a_3, a_4, a_5) - F(a_1, a_2, a_3, a_4, a_5)|}{F(a_1, a_2, a_3, a_4, a_5)} \times 100$$
$$\frac{|g(a_1, a_2, a_3) - G(a_1, a_2, a_3)|}{G(a_1, a_2, a_3)} \times 100$$

We compute the mean, standard deviation, minimum and maximum error percentages of the models f and g over all the 18 inputs and present in the below table

		Error percentage		
Model	Mean	Standard Deviation	Min	Max
f	2.209766	2.480924	0.5398369	9.525519
g	1.002286	1.329765	0.008706377	5.115356

We reckon, these error percentages indicate that the underlying functions F and G are fairly linear and the obtained models f and g are fairly good linear approximations of F and G. We optimize F and G by optimzing f and g respectively. As F and G are not not exactly linear and our obtained models f and g are linear approximations, the optimal values obtained from the models f and g will have some errors comparing to the optimal values of F and G. We expect these errors to be similar to the errors of the models.

3 Optimzing the Linear model Using Linear Programming:

As both f and g are linear and all the inputs of both of the functions are positive, the optimization problems are most likely to be representable as a linear programming problem. The linear programming problem would involve 5 variables. A linear programming problem with 5 variables is a small instance of a linear programming problem. There exist efficient algorithms, namely, simplex, ellipsoid, etc. for solving linear programming problems. In practice, a linear programming problem can easily be solved by any standard linear programming solver like lp_solve.

Now we present the optimization problems discussed above as s linear programming problems in the following subsections:

3.1 *Maximize* $f - g$:

We need to find values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $f(a_1, a_2, a_3, a_4, a_5)$ becomes as large as possible while making $g(a_1, a_2, a_3)$ as small as possible, i.e., we need to maximize

$$\begin{aligned}
& f(x_1,x_2,x_3,x_4,x_5) - g(x_1,x_2,x_3) \\
= \; & (-7.2667 \cdot x_1 + 0.3724 \cdot x_2 + 29.1759 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 78.2264) \\
& -(-5.477 \cdot x_1 + 4.640 \cdot x_2 + 4.290 \cdot x_3 + 24.343) \\
= \; & -1.7897 \cdot x_1 - 3.9176 \cdot x_2 + 24.8859 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 53.8834
\end{aligned}$$

Consequently the linear programming problem looks as follows

$$\begin{aligned}
\text{max:} \quad & -1.7897 \cdot x_1 - 3.9176 \cdot x_2 + 24.8859 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 53.8834; \\
\\
& x_1 <= 11; \\
& x_2 <= 6; \\
& x_3 <= 2.04; \\
& x_4 <= 11; \\
& x_5 <= 20; \\
& x_1 >= 2.5; \\
& x_2 >= 1.5; \\
& x_3 >= 1.25; \\
& x_4 >= 4; \\
& x_5 >= 10;
\end{aligned}$$

The above linear programming problem can be solved by saving the above problem in a file named model.lp (the file is attached) and run the command lp_solve model.lp from the shell of an unix or unix like operating system.

Here is the output:

1. We get the maximum value of the objective function to be: 117.648
2. The required values of the variables are $x_1 = 2.5, x_2 = 1.5, x_3 = 2.04, x_4 = 4, x_5 = 20$

Feeding this input to f and g we find

$$\begin{aligned}
f(2.5, 1.5, 2.04, 4, 20) &= -7.2667 \cdot 2.5 + 0.3724 \cdot 1.5 + 29.1759 \cdot 2.04 - 3.2784 \cdot 4 + 1.8231 \cdot 20 + 78.2264 \\
&= 143.4855 \\
g(2.5, 1.5, 2.04) &= -5.477 \cdot 2.5 + 4.640 \cdot 1.5 + 4.290 \cdot 2.04 + 24.343 \\
&= 26.3621
\end{aligned}$$

We can find the value of $F(2.5, 1.5, 2.04, 4, 20)$ and $G(2.5, 1.5, 2.04)$ by running an experiment to see how good the optimization is.

3.2 *Minimize* $|\frac{f}{g} - d|$:

We need to find the values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $\frac{f(a_1,a_2,a_3,a_4,a_5)}{g(a_1,a_2,a_3)}$ becomes as close as possible to a certain constant b. Let us set $b = 6$. Then we need to minimize $\frac{f(x_1,x_2,x_3,x_4,x_5)}{g(x_1,x_2,x_3)} - 6$. In other words we need to minimize

$$\begin{aligned}
& |f(x_1, x_2, x_3, x_4, x_5) - 6g(x_1, x_2, x_3)| \\
= & |(-7.2667 \cdot x_1 + 0.3724 \cdot x_2 + 29.1759 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 78.2264) \\
& -b(-5.477 \cdot x_1 + 4.640 \cdot x_2 + 4.290 \cdot x_3 + 24.343)| \\
= & |25.5953 \cdot x_1 - 27.4675 \cdot x_2 + 3.4358 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 - 67.8316|
\end{aligned}$$

The above simplification can be done easily for a different value of b than 6. However, for $b = 6$, the linear programming problem looks as follows. Note that a guest variable X has been imported in the linear program because the objective function involves absolute value.

$$\begin{array}{lll}
\text{min:} & & X; \\
 & & \\
x_1 & <= & 11; \\
x_2 & <= & 6; \\
x_3 & <= & 2.04; \\
x_4 & <= & 11; \\
x_5 & <= & 20; \\
x_1 & >= & 2.5; \\
x_2 & >= & 1.5; \\
x_3 & >= & 1.25; \\
x_4 & >= & 4; \\
x_5 & >= & 10; \\
 & & 25.5953x_1 - 27.4675x_2 + 3.4358x_3 - 3.2784x_4 + 1.8231x_5 - 67.8316 <= X; \\
 & & -25.5953x_1 + 27.4675x_2 - 3.4358x_3 + 3.2784x_4 - 1.8231x_5 + 67.8316 <= X;
\end{array}$$

The above linear programming problem can be solved by saving the above problem in a file named model1.lp (the file is attached) and run the command lp_solve model1.lp from the shell of an unix or unix like operating system.

Here is the output of the linear program:

1. We get the maximum value of the objective function to be: 0
2. The required values of the variables are $x_1 = 3.89215, x_2 = 1.5, x_3 = 1.25, x_4 = 4, x_5 = 10$

   Feeding this input to f and g we find

$$\begin{aligned}
f(3.89215, 1.5, 1.25, 4, 10) &= -7.2667 \cdot 3.89215 + 0.3724 \cdot 1.5 + 29.1759 \cdot 1.25 - 3.2784 \cdot 4 + 1.8231 \cdot 10 + 78.2264 \\
&= 92.08919 \\
g(3.89215, 1.5, 1.25) &= -5.477 \cdot 3.89215 + 4.640 \cdot 1.5 + 4.290 \cdot 1.25 + 24.343 \\
&= 15.34819
\end{aligned}$$

We can find the value of $F(3.89215, 1.5, 1.25, 4, 10)$ and $G(3.89215, 1.5, 1.25)$ by running an experiment to see how good the optimization is.

3.3 *Minimize* $|f - d_1| + |g - d_2|$:

Finding the values $x_1 = a_1, x_2 = a_2, x_3 = a_3, x_4 = a_4, x_5 = a_5$ so that $f(a_1, a_2, a_3, a_4, a_5)$ and $g(a_1, a_2, a_3)$ goes as close as possible to constant values b_1 and b_2 respectively. Let us set $b_1 = 121$ and $b_2 = 20.7$. We need to minimize

$$\begin{aligned} & |f(x_1, x_2, x_3, x_4, x_5) - 120| + |g(x_1, x_2, x_3) - 24| \\ = & |-7.2667 \cdot x_1 + 0.3724 \cdot x_2 + 29.1759 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 + 78.2264 - 121| \\ & +|-5.477 \cdot x_1 + 4.640 \cdot x_2 + 4.290 \cdot x_3 + 24.343 - 20.7| \\ = & |-7.2667 \cdot x_1 + 0.3724 \cdot x_2 + 29.1759 \cdot x_3 - 3.2784 \cdot x_4 + 1.8231 \cdot x_5 - 42.7736| \\ & +|-5.477 \cdot x_1 + 4.640 \cdot x_2 + 4.290 \cdot x_3 + 3.643| \end{aligned}$$

The above simplification can be done easily for a different value of b_1 and b_2. However, for $b_1 = 121$ and $b_2 = 20.7$, the linear programming problem looks as follows. Note that two guest variables X, Y has been imported in the linear program because the objective function involves absolute value.

min: $X + Y$;

$$\begin{aligned} x_1 &<= 11; \\ x_2 &<= 6; \\ x_3 &<= 2.04; \\ x_4 &<= 11; \\ x_5 &<= 20; \\ x_1 &>= 2.5; \\ x_2 &>= 1.5; \\ x_3 &>= 1.25; \\ x_4 &>= 4; \\ x_5 &>= 10; \end{aligned}$$

$$\begin{aligned} & -7.2667x_1 + 0.3724x_2 + 29.1759x_3 - 3.2784x_4 + 1.8231x_5 - 42.7736 <= X; \\ & 7.2667x_1 - 0.3724x_2 - 29.1759x_3 + 3.2784x_4 - 1.8231x_5 + 42.7736 <= X; \\ & -5.477x_1 + 4.640x_2 + 4.290x_3 + 3.643 <= Y; \\ & 5.477x_1 - 4.640x_2 - 4.290x_3 - 3.643 <= Y; \end{aligned}$$

The above linear programming problem can be solved by saving the above problem in a file named model2.lp (the file is attached) and run the command lp_solve model2.lp from the shell of an unix or unix like operating system.

Here is the output of the linear program:

1. We get the maximum value of the objective function to be: 0
2. The required values of the variables are $x_1 = 3.5338, x_2 = 1.5, x_3 = 2.04, x_4 = 4, x_5 = 11.787$

Feeding this input to f and g we find

$$\begin{aligned} f(3.5338, 1.5, 2.04, 4, 11.787) &= -7.2667 \cdot 3.5338 + 0.3724 \cdot 1.5 + 29.1759 \cdot 2.04 - 3.2784 \cdot 4 + 1.8231 \cdot 11.787 + 78.2264 \\ &= 121.0001 \\ g(3.5338, 1.5, 2.04) &= -5.477 \cdot 3.5338 + 4.640 \cdot 1.5 + 4.290 \cdot 2.04 + 24.343 \\ &= 20.69998 \end{aligned}$$

We can find the value of $F(3.2077, 1, 1.25, 2.16454, 5)$ and $G(3.2077, 1, 1.25)$ by running an experiment to see how good the optimization is.

4 Evaluation:

The underlying functions F and G of the experiments are found to be fairly linear. Keeping this observation in mind, linear approximation f and g can be obtained by linear regression by using 5 experimental values. Because 5 points in a 5 dimensional space completely defines a 4 dimensional hyperplane. So, the linear approximations f and g can be easily obtained. Using the linear approximations, the above method can be used to find the values of the variables x_1, x_2, x_3, x_4, x_5 to obtain the desired values of y_1, y_2.