5
Exploring "Data Science" via Linear Regression

So far, our treatment of machine learning has been from the perspective of a computer scientist. It is important to note, however, that models such as linear regression are useful in a variety of other fields including the physical sciences, social sciences, etc. In this chapter, we present case studies from different fields. Here, the inputs x_i are considered to be *explanatory variables*, the output y is considered to be the *effect variable*, and the weights w_i quantify the causal significance of the associated inputs x_i on the output y. The interpretation of weights as a type of causality is crucial; often, the ideal method of determining causality through a set of rigorous randomized control trials is too expensive.

5.1 Boston Housing: Machine Learning in Economics

Our first case study comes from the field of economics. In 1978, Harrison and Rubinfeld released a classic study on the willingness to pay for clean air in the Boston metropolitan area. Their methodology involved an economic model called *hedonic pricing*, [1] which essentially estimates the value of a good by breaking it down into "constituent characteristics." It turns out we can use linear regression can help determine which of these attributes are most important. Specifically, suppose we have a dataset of house sales where y represents the price of the house and $\vec{\mathbf{x}} \in \mathbb{R}^{15}$ represents a set of house attributes. [2] Then, we aim to find an optimum set of weights $\vec{\mathbf{w}}$ for the linear model:

$$y \approx \sum_{i=0}^{14} w_i x_i \tag{5.1}$$

Table 5.1 lists all 14 attributes that were used in the linear regression model. Before fitting the model with these attributes, it is useful to intuitively reason about some of the attributes. For instance, we expect the weight w_5 corresponding to RM, the number of bedrooms,

[1] This definition is paraphrased from the following Wikipedia article: https://en.wikipedia.org/wiki/Hedonic_regression

[2] x_0 is a dummy variable, and the remaining 14 coordinates $x_1, \ldots, x_{14}$ each correspond to an attribute.

Index	Code	Description
1	ZN	proportion of residential land zoned for lots over $25,000$ ft^2
2	INDUS	proportion of non-retail business acres per town
3	CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
4	NOX	nitric oxides concenteration (parts per 10 million)
5	RM	average number of rooms per dwelling
6	AGE	proportion of owner-occupied units built prior to 1940
7	DIS	weighted distances to five Boston employment centres
8	RAD	index of accessibility to radial highways
9	TAX	full-value property-tax rate per \$10,000
10	MEDV	Median value of owner-occupied homes (in \$1,000s)
11	CRIM	per capita crime rate in town
12	PTRATIO	pupil-teacher ratio by town
13	LSTAT	% lower status of the population
14	B	$1000(Bk - 0.63)^2$ where Bk is the proportion of black population in town

Table 5.1: 14 attributes used in the Boston housing regression model. The attributes are presented in a different order from the paper.

to be positive because larger houses typically sell for more. Conversely, we expect the weight w_4 corresponding to NOX, the amount of air pollution, to be negative as people would prefer not to live in a polluted environment. After running the regression, it indeed turns out that these intuitions are correct. [3] In general, it can be useful to double-check that the calculated weights align with intuition: if they do not, it could be a sign that a modeling assumption is incorrect.

[3] The regression weights can be found on page 100 of the original paper. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/22636/0000186.pdf?sequence=1&isAllowed=y.

5.1.1 *The Strange Math of Feature B*

The headline result of the paper is that the willingness to pay for cleaner air increases both when income level is higher and when the current pollution level is higher. However, if you read the paper closely, you may notice the presence of a curious parameter B, which is defined in terms of Bk, the proportion of black population in the neighborhood. This parameter is meant to represent a social segregation effect present within the Boston housing market. The authors of the paper speculated that (1) at a lower level of Bk, the housing price will decrease as Bk increases since the white population tend to avoid black population, but (2) at a very high level of Bk, the housing price will increase as Bk increases because black population prefer predom-

inantly black neighborhoods. To capture this intuition, they defined the attribute B in the parabolic expression $B = 1000(Bk - 0.63)^2$. It indeed turns out that the weight w_{14} corresponding to B is positive as shown in Figure 5.1.

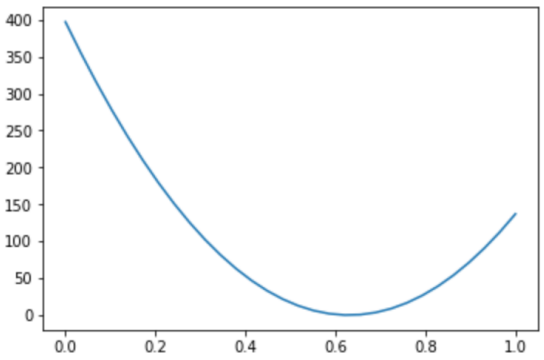

Figure 5.1: The graph of $B = 1000(Bk - 0.63)^2$. This is an example of featurization as discussed in Chapter 1. It encodes prevailing discrimination of that period. The term "black" is not favored today either.

5.1.2 *Ethnic Concerns Behind a Model*

It seems strange to have such a sensitive attribute B have an influence on the model. We might wonder about the social harm that could arise if the model was used by real-life sellers or buyers (*e.g.*, the buyers could demand a house for a lower price based on the proportion of black population in the neighborhood). On the other hand, the fitted model confirms that there is an underlying segregation effect already present in the society. Also, we cannot guarantee that the model would be race-neutral even if we eliminated the parameter B. For instance, maybe one or more of the other variables (*e.g.*, air quality variables) is highly correlated with B. [4]

[4] We will revisit such issues of bias in Chapter 16.

Ultimately, the primary takeaway from this case study is that implementing machine learning models in real life is a challenge itself. At a technical level, the model may make sense and make good predictions of house prices. But one has to consider the social effects of an ML model on the phenomenon being studied: in particular, whether it supports or extends prevailing inequities. The following are some important pointers to keep in mind:

1. If the world has a problem, the data will reflect it and *so will our models*
2. If a problematic model later gets used in real life, it can *exacerbate the existing problem*
3. The choices of attributes when making a model might *bias the outcome*

4. Carelessly using data can later *lead to modeling issues*

5.2 *fMRI Analysis: Machine Learning in Neuroscience*

We next consider an application of ML in a vastly different field. One of the most important tools in contemporary neuroscience is Functional Magnetic Resonance Imaging (fMRI). fMRI has been used successfully to map human functionality (*e.g.*, speech, memory) to brain regions. In a more active role, it can assist with tumor surgery or "decoding" thoughts and emotions.

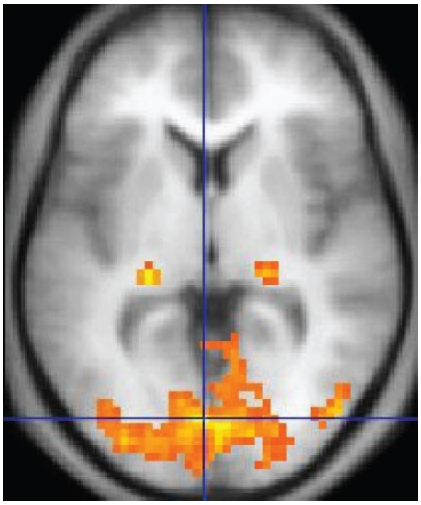

Figure 5.2: A sample image of a fMRI reading. Source: https://en.wikipedia.org/wiki/Functional_magnetic_resonance_imaging

fMRI experiments often involve presenting a set of stimuli (*e.g.*, images of human face) to the subject in order to elicit a neurological response, which is then captured through a fMRI reading. Each reading reveals the concentration of oxygen in the blood stream throughout the brain, which is used as a proxy for brain activity. [5] Through the result of the reading, we are able to conclude if a particular voxel responds to a particular stimulus. The naive way of conducting these experiments is to present one stimulus at a time and wait until we get a reading of the brain response before we move on to the next stimulus.

[5] Formally, this is referred to as the blood-oxygen-level-dependent (BOLD) signal

But if you have previously taken a course in neuroscience, you may recall that fMRI is unfortunately a double-edged sword. It features excellent spatial resolution, with each voxel as small as 1 mm^3. However, it has poor temporal resolution: often, readings require several seconds for blood flow to stabilize! Coupled with the fact that regulations limit the amount of time human subjects can spend in the scanner, it becomes clear that methodologies based on sequential presentation of stimuli are too inefficient. In this section, we explore

how to leverage techniques from linear regression in order to solve this problem.

5.2.1 *Linear Superposition*

The key intuition involves a concept called *linear superposition*: if a subject is shown multiple stimuli in quick succession, the strength of the voxel's response is the sum of the strength of its response to each of the individual stimuli. [6] Instead of waiting until we have the image of one stimulus to move on, considering showing a new stimulus every 1 or 2 seconds. Each fMRI reading will now capture the *composite* brain response to the stimuli from the past few seconds. We will use linear regression to *disentangle* the information, and extract which voxel responded to which stimulus. [7]

Consider the following example.

Example 5.2.1. *See Figure 5.3. The graph on the top left represents a voxel's response when the subject is shown the image of a face. The graph on the top right represents the response when the subject is shown the image of a flower. The bottom graph represents the response when the subject is shown the image of a flower* 1 *second after the image of a face. Notice that the first two graphs have been* ***superposed*** *to create the third graph. In practice, we are interested in the problem of extracting the individual graphs when given the superposed graph.*

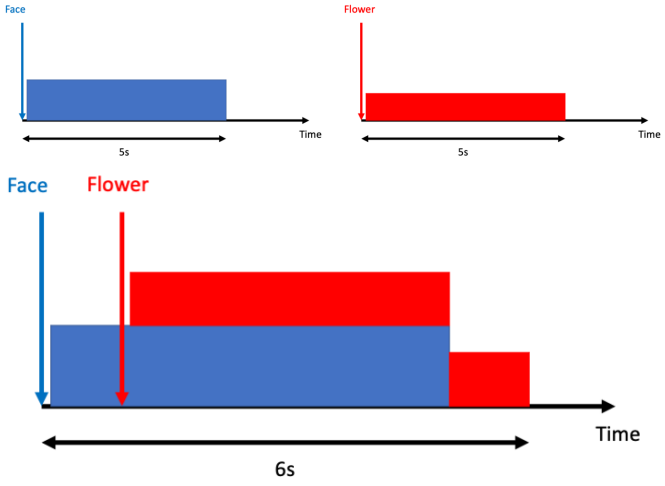

Figure 5.3: Three graphs explaining the effect of linear superposition.

[6] This is exactly like the linear superposition of wave functions in physics.

[7] Note that this is a very simplified version. The actual process is much more complicated.

5.2.2 *Linear Regression*

Now let us describe how to formulate this problem in terms of linear regression. First assume that the subject is shown one of k types of stimuli at each time step t where $t \in \{1, 2, \ldots, T\}$. Let y_t be the

response of a particular voxel at step t. The main assumption is that y_t is the linear superposition of the responses to stimuli from the steps in $[t-10, t]$. We also define a $T \times k$ matrix X with 0/1 entries, where $X_{ts} = 1$ if stimulus type s is shown during $[t-10, t]$ and 0 otherwise. Then we can set up the following linear regression model:

$$y_t \approx \sum_{s=1}^{k} w_s X_{ts}$$

When we find the optimal values of w_s via least squares, $w_s = 1$ means that the particular voxel responds to the stimulus type s.

5.2.3 *Neural Correlates of Thought*

Now we know how to find the values of w_s for a specific voxel. That is, we can test if a particular voxel responds to a particular stimulus. Combining this method with a *spatial smoothing* (*i.e.*, applying the principle that nearby voxels behave similarly), [8] we are able to identify which region of a brain is associated to which stimulus. So far, more than 1,000 regions of the brain have been identified and mapped.

[8] The simplest smoothing method is to take the w_s values for one voxel and replace them with the average of the neighboring voxels.

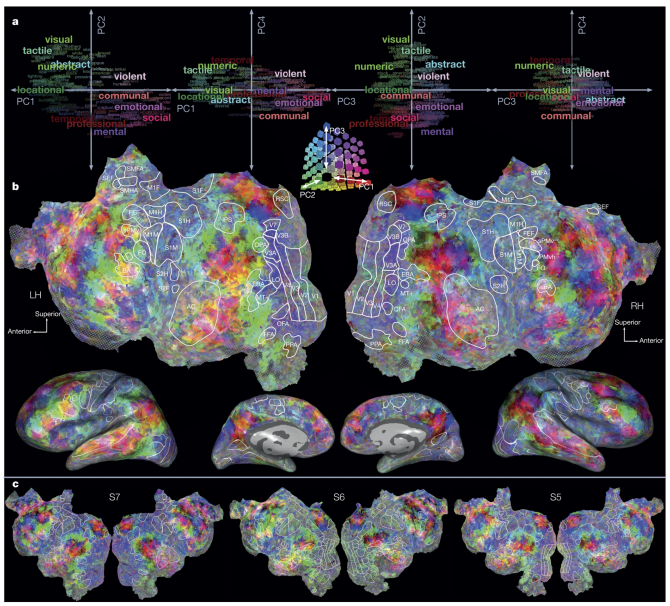

Figure 5.4: A detailed map labeling areas of the brain with corresponding stimuli. https://www.nature.com/articles/nature17637

5.2.4 *Brain-Computer Interface (BCI)*

We finish off with a tangible example of how our studies can help people. Patients who are suffering from Locked-in Syndrome (LIS) are aware of their surroundings and have normal reasoning capacities

but have *no way* of communicating with others through speech or facial movements. Using a combination of a technology called Brain-Computer Interface and a linear regression model, we are able to communicate with these patients.

Brain-Computer Interface is an electode sensor implanted near the motor cortex that can detect the electric signal that LIS patients are trying to send to the motor cortex. We can teach the patients to *visualize* writing with their dominant hand if they want to answer "no" and visualize writing with their non-dominant hand if they want to answer "yes." Since the neural correlates of the two movements are very different, BCI will pick up essentially disjoint signals, and we can use linear regression model to distinguish between them. [9]

[9] Note: training also requires labeled data, which can be produced by asking the patient questions about known facts (*e.g.*, birth date, marital status, etc.). This technique has been used to communicate with patients in deep coma and presumed to be in a vegetative state. See *Science of Mind Reading*, New Yorker, December 6 2021, which also profiles several Princeton researchers.