CSM 70 Lance Mathias
Spring 2022

Extra Note: Linear Least Squares

1 Estimating a Sample

Consider a sample over 1 dimension. We can represent this as an n-dimensional vector of points x. Suppose we want to find the value that best represents this set. This is simply the mean of the set, which we will call $\hat{x}$. More formally, this means our estimator $\hat{x}$ is defined as

$$\hat{x} = \mathbb{E}[x] \tag{1}$$

How good is this estimator? One way to measure the "quality" of the estimator is to find the mean of the squared differences of each point in the sample from our estimator. This is simply definition of variance, since we have

$$var(x) = \mathbb{E}[(x-\hat{x})^2] = \frac{1}{n}\sum^{n}(x_i-\hat{x})^2 \tag{2}$$

which is exactly what we wanted to show, since each point in the sample represents $\frac{1}{n}$ of the total sample.
We can prove that the optimal value of $\hat{x}$ that minimizes variance is the mean, or $\mathbb{E}[x]$:
We can rewrite the variance as

$$var(x) = \frac{1}{n}\sum^{n}(x_i-\hat{x})^2 = \frac{1}{n}\sum^{n}x_i^2 - 2x_i\hat{x}+\hat{x}^2 \tag{3}$$

We can differentiate with respect to $\hat{x}$ and can find the minimum, where the derivative is equal to 0. Another perspective is that we are shifting $\hat{x}$ until we find the value that minimizes the variance:

$$\frac{\partial}{\partial\hat{x}}\frac{1}{n}\sum^{n}x_i^2 - 2x_i\hat{x}+\hat{x}^2 = \frac{1}{n}\sum^{n}-2x_i+2\hat{x} = 2\hat{x}-\frac{1}{n}\sum^{n}2x_i = 0 \tag{4}$$

since $\hat{x}$ is a scalar and we can simply move it outside the summation. This means that

$$2\hat{x}-\frac{1}{n}\sum^{n}2x_i = 0 \implies \hat{x} = \frac{1}{n}\sum^{n}x_i = \mathbb{E}[x] \tag{5}$$

which is what we wanted to show.

2 Multiple Dimensions

Now let's extend this to multiple dimensions. Say we have multi-dimensional data points, and we want to predict a final data point (for instance, say we want to predict someone's CS70 final grade based on their midterm score). We can do this using least squares, where we want to minimize

the squared distance between our prediction and reality. Given data points x, we want to find an estimator A such that our prediction is as close as possible to reality, y. More formally, we want to minimize the quantity

$$\min_A ||Ax - y||^2 \tag{6}$$

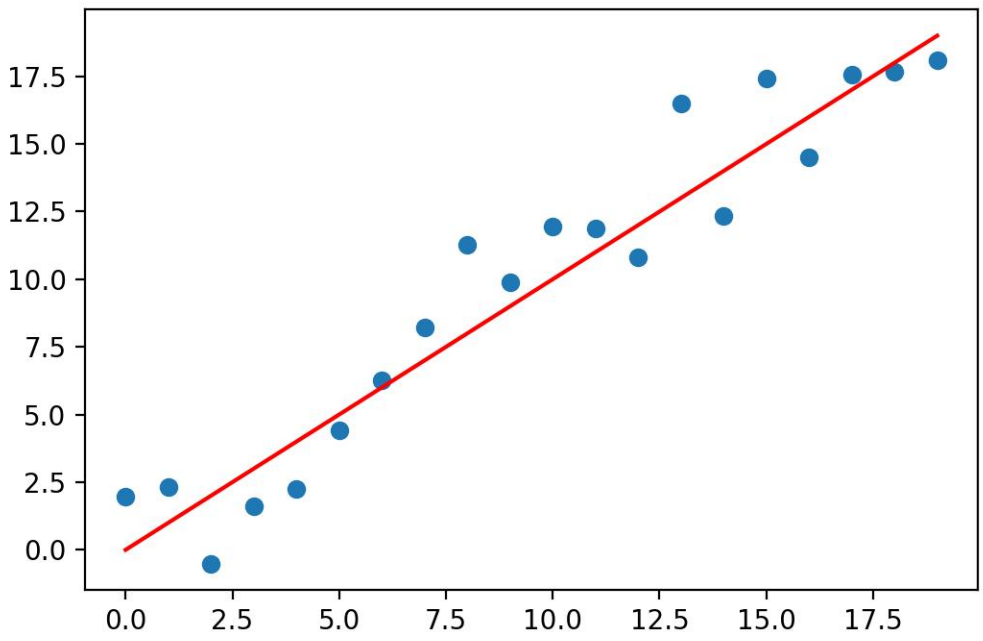
If we normalize this over the number of points as follows, then we will get what's called Mean Squared Error. Intuitively, this makes sense, as we're taking the average of each value of the "error," which is the difference between our prediction and reality.

$$MSE = \min_A \frac{1}{n}||Ax - y||^2 \tag{7}$$

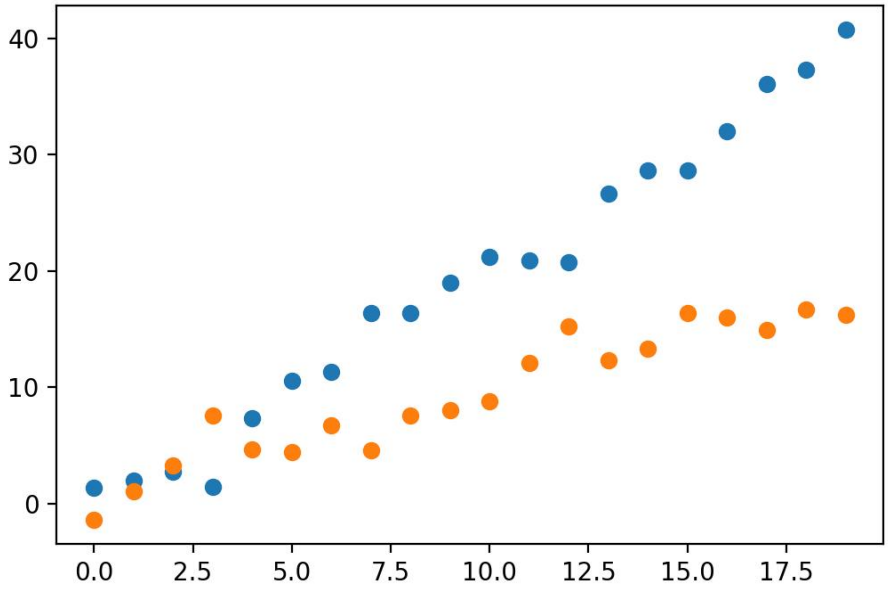
We can also represent this using expected value. This time, instead of representing reality as a vector with n components, we represent it as a random variable Y, where each data point has a probability of $\frac{1}{n}$. Similarly, we represent our domain x as a random variable X, and our estimator A becomes a function $\hat{f}$ of random variable X, since $\hat{f}$ is a line.

$$MSE = \mathbb{E}[(Y - \hat{f}(X))^2] \tag{8}$$

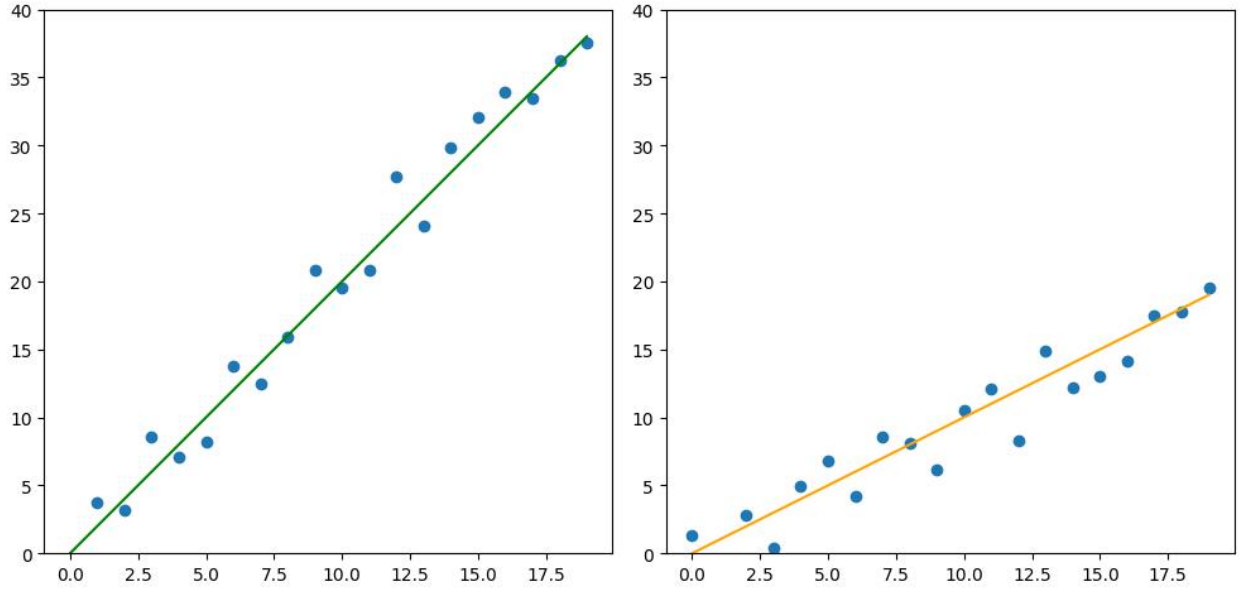
This is identical to (6), where instead of the line Ax, our line is now $\hat{f}(X)$, a linear function on a random variable. Our linear estimator would look something like this:



Makes sense, but let's say our data is not so nice, and doesn't follow a single line so well:



Going off of our previous example, suppose that the blue points are students who go to CSM and the orange points represent students who don't. Clearly, there's correlation between CSM attendance and final grade, so let's construct separate estimators for students who did and did not attend CSM:

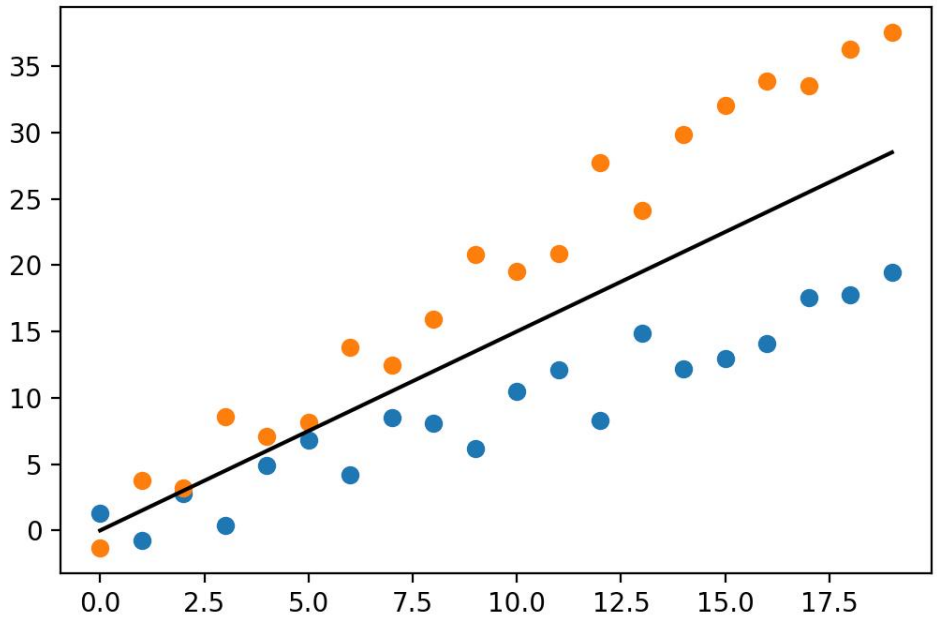


This is simply taking the best estimator for a certain subset of the original set of points, which is the same thing as conditioning the estimator on a certain characteristic of the sample. Again, our best estimator is the expectation of the set of points we are predicting over. In this example, the

left plot is final grades of students given that they went to CSM, while the right plot is final grades of students given that they did not go to CSM. We can relate this back to our notion of expectation as follows, where $\hat{f}_i$ is our conditioned estimator and $H = h$ defines a characteristic of the dataset (e.g. students who go to CSM):

$$\hat{f}_i(X|H) = \mathbb{E}[X|H = h] \tag{9}$$

Furthermore, we can average the estimators for each subset of the dataset in order to get an estimator for the dataset as a whole:



We can relate this back to our notion of expectation as follows:

$$\hat{f}(X) = \sum_H \mathbb{E}[\hat{f}_i(X)|H = h]\mathbb{P}[H = h] \tag{10}$$

Intuitively, this is just averaging each estimator of a certain subset, weighting it relative to the number of points in that subset.
Recall that estimator $\hat{f}_i$ is actually an expected value of the dataset as well. Thus, we can express the above equation as follows:

$$\hat{f}(X) = \sum_H \mathbb{E}[\hat{f}_i(X)|H = h]\mathbb{P}[H = h] = \mathbb{E}[\mathbb{E}[X|H = h]] = \mathbb{E}[X] \tag{11}$$

Note that we are taking the expectation of an expectation, which may seem confusing, but is actually the same thing we did before! This is known as the law of total expectation or the law of iterated expectation. We can expand out the inner expectation using the definition of expectation to get

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|H = h]] = \sum_H \mathbb{E}[\hat{f}_i(X)|H = h]\mathbb{P}[H = h] = \sum_H \sum_X x\mathbb{P}[X = x|H = h] \tag{12}$$

Which follows from the definition of expectation.
Hopefully this note helped motivate and visualize how least-squares estimation relates to our notion of expected value and iterated expectation! If these concepts seem interesting, I would recommend checking out EECS126 (probability) or EECS127 (optimiation) to dig deeper into these topics.