~~Mean~~ Regression

One of the first estimation problems we considered was that of estimating the mean of an unknown distribution, given iid samples from the distribution. We learnt two approaches to this problem:

1. Compute the sample mean, & use it as the estimator.
2. If the distribution is known (or assumed) to belong to a parametric family, use the MLE to estimate the parameter from the data. Then compute the mean of the distribution with the estimated parameter.

Sometimes, though not always, the two methods yield the same answer.

We now remind ourselves that the first method can also be viewed as the solution of an optimisation problem.

MMSE estimators

MMSE stands for minimum mean-squared error. That is the error metric these estimators minimise.

Fact: The sample mean $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ of data $x_1, x_2, \ldots, x_n$ is the unique minimiser of the mean-squared error

$$MSE(y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - y)^2$$

Proof: The proof is straightforward. We want to minimise the function $MSE(y)$ over all $y \in \mathbb{R}$. We do this by setting its derivative to zero, and checking if any solution is a minimum or maximum or neither.

$$\frac{d}{dy} MSE(y) = \frac{2}{n} \sum_{i=1}^{n} (y - x_i)$$

$$= 0 \iff \sum_{i=1}^{n} (y - x_i) = 0$$

$$\iff ny = \sum_{i=1}^{n} x_i \iff y = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Thus, the derivative is zero at $y^* = \frac{1}{n} \sum_{i=1}^{n} x_i$

Also $\frac{d^2}{dy^2} MSE(y)\Big|_{y=y^*} = \frac{2}{n} \sum_{i=1}^{n} 1 \Big|_{y=y^*}$

$$= 2\Big|_{y=y^*} = 2.$$

As $\frac{d^2}{dy^2} MSE(y) > 0$ at $y = y^*$, therefore

y^* is a local ~~minimum~~ minimiser.

As the function $MSE(y)$ is continuously differentiable and has no other critical points (where the derivative is zero), y^* is also the unique global minimiser.

Alternative proof: There is an alternative proof by completion of squares which is also instructive / insightful.

Define $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$

$$MSE(y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu + \mu - y)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[(x_i - \mu)^2 + 2(x_i - \mu)(\mu - y) + (\mu - y)^2 \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{n} \cdot n(\mu - y)^2 + \frac{2(\mu - y)}{n} \sum_{i=1}^{n} (x_i - \mu)$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 + (\mu - y)^2 + 0$$

since $\sum_{i=1}^{n} (x_i - \mu) = 0$ by definition of μ

$\therefore MSE(y) \geq \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$, with equality if and only if $\mu = y$.

Linear Regression

Quite often, we are interested not just in estimating the mean of a random variable, but the conditional mean of one "response variable" given a bunch of other "explanatory variables".

For example, we might want to know the mean received power at a communication receiver conditional on certain choices of transmit antenna parameters, carrier frequency, modulation scheme, etc. Or the conditional yield of a plot of land given the amount of different fertilisers applied.

Usually, we denote the response variable by Y, and the explanatory variables by $X_1, X_2, \ldots, X_k$. These are random variables with some joint distribution $F_{Y,X_1,\ldots,X_k}(y,x_1 \ldots x_k)$

We don't want to learn the whole joint distribution, just the conditional expectation

$$\mathbb{E}(Y \mid X_1, X_2, \ldots, X_k)$$

The most common assumption is that the conditional expectation is a linear function of the explanatory variables (plus a constant):

$$\mathbb{E}(Y \mid X_1, X_2, \cdots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where the β_i are unknown coefficients; β_0 is called the intercept.

Let's rewrite the above equation in vector notation as

$$\mathbb{E}(Y \mid \underline{X}) = \underline{X}\,\underline{\beta}$$

where $\underline{X} = (1 \; X_1 \; X_2 \cdots X_k)$ is a row vector

$\underline{\beta} = (\beta_0 \; \beta_1 \cdots \beta_k)^T$ is a column vector

Typically, we have some number n of observations of the joint random variables $(Y, X_1, \cdots, X_k)$, and wish to estimate $\underline{\beta}$.

In other words, we are given

$(y(1), x_1(1), \cdots, x_k(1))$, $(y(2), x_1(2), \cdots, x_k(2))$, $\cdots$ $(y(n), x_1(n), \cdots x_k(n))$.