18.650: Fundamentals of Statistics November 13, 2023

Lecture 23 — Linear Regression

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Definitions and main concepts

In regression, we predict the value of a *response variable* $Y \in \mathbb{R}$ based on a *feature vector* or *predictor* $X \in \mathbb{R}^k$.

Example.

> The feature vector X and response Y could be e.g.
>
> $$X = \begin{pmatrix} \text{weight} \\ \text{age} \\ \text{salary} \\ \text{GPA} \end{pmatrix}, \quad Y = \text{IQ}.$$
>
> Another example of a Y is whether or not the person will default on a credit. In this case Y only takes the value 0 or 1.

Goal: predict Y given X, or understand how Y changes with X. (In particular, understand which of the features in X are particularly relevant for predicting Y.)

Difficulty: For each fixed $X = x$, there is a whole probability distribution $Y|X = x$. For example, we could have $Y|X = x \sim \mathcal{N}(f(x), \sigma^2)$. It is not realistic to assume that there is a function $f(x)$ such that knowing $X = x$ implies $Y = f(x)$ exactly.

Best prediction property: Although we cannot hope to find a function f such that $Y = f(X)$ exactly, we can look for f which minimizes the expected error $\mathbb{E}[(Y - f(X))^2]$. To minimize this expectation, first note that

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}[\mathbb{E}[(Y - f(X))^2 \mid X]]$$

by the tower property of conditional expectation. We can now minimize the inner expectation for each possible $X = x$:

$$\min_{f(x)} \mathbb{E}[(Y - f(x))^2 \mid X = x] = \min_a \mathbb{E}[(Y - a)^2 \mid X = x].$$

In other words, $f(x)$ is just the value a that minimizes $h(a) = \mathbb{E}[(Y - a)^2 \mid X = x]$.

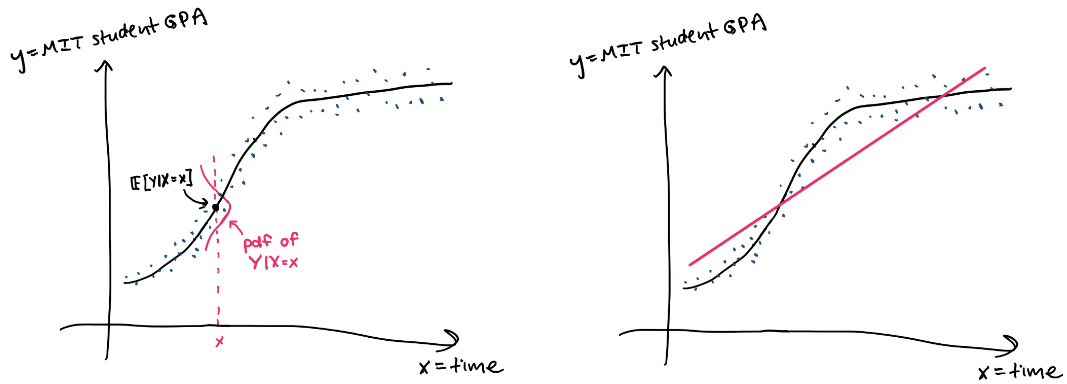

Figure 1: A scatterplot of observations (X_i, Y_i) (blue points). In the lefthand plot, the solid black curve depicts the expectation $\mathbb{E}[Y|X = x]$. In the righthand plot, the pink line denotes a linear fit to the data.

We minimize h by setting the derivative to zero:

$$\begin{aligned} 0 = h'(a) =& 2\mathbb{E}[Y - a \mid X = x] = 2(\mathbb{E}[Y \mid X = x] - a) \\ &\Rightarrow a = \mathbb{E}[Y \mid X = x]. \end{aligned}$$

The minimizing function is $f(x) = \mathbb{E}[Y \mid X = x]$.

Definition 1.1: Regression function

Given a random vector $X \in \mathbb{R}^k$ and a random variable $Y \in \mathbb{R}$, the function $f : \mathbb{R}^k \to \mathbb{R}$ given by

$$f(x) = \mathbb{E}[Y \mid X = x]$$

is called the regression function of Y onto X.

Remark.

> The definition extends to multivariate Y.

In practice, we only get to observe pairs (X_i, Y_i), as in Figure 1. So of course, we cannot perfectly compute the regressor function $f(x) = \mathbb{E}[Y|X = x]$ — there are lots of x's for which we don't get to observe even a single y!

It's also important to keep in mind that even if we *could* perfectly compute $\mathbb{E}[Y|X = x]$, this expectation does not fully capture the distribution $Y \mid X = x$. An alternative to vanilla regression (finding the mean $f(x)$) is quantile regression, which gives a confidence band around the mean.

2 Linear regression.

In linear regression, we make the assumption that $f(x) = \mathbb{E}[Y|X = x]$ is *linear in* x, i.e. of the form

$$f(x) = \mathbb{E}[Y|X = x] = x^\intercal \beta$$

for some unknown ground truth $\beta = \beta^* \in \mathbb{R}^k$. To find an estimator of β^* we'll use the MLE. To compute the MLE, we first need to assume a parametric form for the distribution of $Y|X = x$. We'll use our favorite distribution, the Gaussian.

Assumption G: For some unknown β^* and (usually known) σ^2, it holds

$$Y \mid X = x \sim \mathcal{N}(x^\intercal \beta^*, \sigma^2).$$

Assumption G actually incorporates three separate assumptions:

1. $Y \mid X = x$ is $\mathcal{N}(f(x), \sigma^2(x))$, i.e. some generical Gaussian distribution.
2. The mean function is linear: $f(x) = x^\intercal \beta^*$
3. The variance function is constant: $\sigma^2(x) = \sigma^2$ for all x.

Remark.

> The last assumption, that $\sigma^2(x)$ is constant for all x, is know as homoskedastic regression, as opposed to heteroskedastic regression.

Now that we have a statistical model, we can write down the log likelihood $\ell_n(\beta)$ and use it to compute the MLE $\hat{\beta}^{\mathrm{MLE}}$.

$$\ell_n(\beta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - X_i^\intercal \beta)^2}{2\sigma^2}\right)\right) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - X_i^\intercal \beta)^2 + \text{const.}$$

Since maximizing ℓ_n is equivalent to minimizing $-\ell_n$, we see that the MLE is given by

$$\hat{\beta}^{\mathrm{MLE}} = \mathrm{argmin}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (Y_i - X_i^\intercal \beta)^2. \tag{1}$$

Remark.

> We could get to the same formula for the MLE by minimizing the expected error $\mathbb{E}[(Y - f(X))^2]$ over all linear f:
>
> $$\min_{f \text{ linear}} \mathbb{E}[(Y - f(X))^2] = \min_{\beta} \mathbb{E}[(Y - X^\intercal\beta)^2] \overset{\text{plug-in rule}}{\rightsquigarrow} \min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^\intercal\beta)^2.$$
>
> This has.to do with the fact that minimizing a quadratic loss is equivalent to maximizing a Gaussian probability.

Exercise Suppose that instead of having a constant σ^2, we assumed the variance was given by some known function $\sigma^2(x)$. In this case, what is the MLE $\hat{\beta}^{\text{MLE}}$?

2.1 Closed form solution for least squares

Let's find a closed form solution to $\hat{\beta}^{\text{MLE}}$, the minimizer in (1). To do this we'll need some matrix calculus. First, write all the Y_i's into a vector:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n.$$

Next, put the X_i's as rows in the following matrix:

$$\mathbb{X} = \begin{pmatrix} — & X_1^\intercal & — \\ — & X_2^\intercal & — \\ & \vdots & \\ — & X_n^\intercal & — \end{pmatrix} \in \mathbb{R}^{n \times k}.$$

We can now write the minimization problem (1) as

$$\min_{\beta} \sum (Y_i - X_i^\intercal\beta)^2 = \min_{\beta} \|Y - \mathbb{X}\beta\|^2$$

We set the gradient with respect to β to zero to get

$$2\mathbb{X}^\intercal(Y - \mathbb{X}\beta) = 0 \implies \mathbb{X}^\intercal Y = \mathbb{X}^\intercal X\beta \implies \hat{\beta}^{\text{LS}} = (\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal Y,$$

where the LS stands for least squares (but it's also the MLE).

Interpretation of the LS solution. Note that if we multiply both sides of $\hat{\beta}^{\text{LS}}$

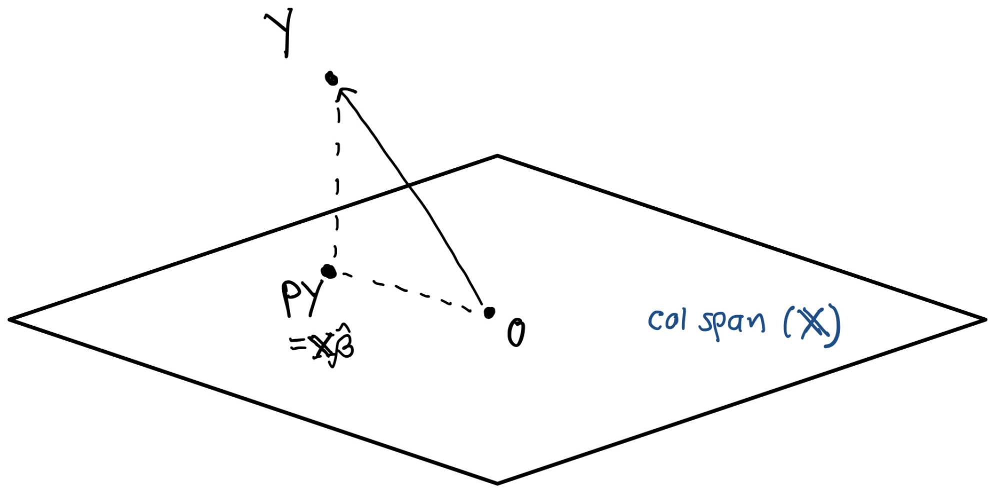

Figure 2: Visualization of the least squares solution

by $\mathbb{X}$, we get

$$\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal Y = PY, \qquad P := \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal.$$

The matrix P is a *projection matrix*: it satisfies $P^\intercal = P$ and $P^2 = P$. Therefore, the fact that $\mathbb{X}\hat{\beta} = PY$ shows that $\mathbb{X}\hat{\beta}$ is the projection of Y onto span($\mathbb{X}$), the linear space spanned by the columns of $\mathbb{X}$.

In other words, the formula for $\hat{\beta}^{\mathrm{LS}}$ is implicitly encoding two steps: first, we find the closest point to Y in the column span of $\mathbb{X}$ (the projection of Y). This is the point PY. Since PY lies in the column span, it is by definition given by some linear combination of the columns of $\mathbb{X}$. Therefore, the second step is to identify the coefficients of this linear combination — these are precisely the entries of $\hat{\beta}^{\mathrm{LS}}$.