Lab 1: (Yet another) Intro to Regression

Info 251 — Applied Machine Learning
University of California, Berkeley ∘ Spring 2026
authors: Simón Ramírez Amaya
{simonra}@berkeley.edu

Objectives

1. Become familiar with the general setup of regression problems.
2. Derive and implement canonical linear regression algorithms.
3. Use a bit of linear algebra to develop graphical intuition.

Setup

In AML, we will spend most of the time studying the relationship between random variables. In linear regression problems we distinguish two types:

- Y is the outcome that we are interested in explaining/predicting. In the regression problems that we will study, it is always a real-valued scalar ($Y \in \mathbb{R}$). Also commonly referred to as dependent variable, response variable, target and LHS.
- $\mathbf{X}$ is the predictors that we control/observe and that we hypothesize are informative about Y. In the regression problems that we will study, it is always a real-valued vector of dimension k ($\mathbf{X} \in \mathbb{R}^k$). Also commonly referred to as regressors, independent variables, covariates, features and RHS.

We will assume that the relationship between $\mathbf{X}$ and Y is governed by a stable *joint probability distribution* that we will denote as D.[1] In regression problems, the main object of interest is the conditional mean function:

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]. \tag{CMF}$$

Exercise 1

As a working example, let X denote the years of completed schooling (k=1) and Y the annual income in US dollars of a random draw from some hypothetical adult population. In the companion notebook, we have included code to generate and visualize three different joint distributions of schooling and income. Execute the code cells and take a look at the resulting figures.

- In plain english, explain the main differences between these job markets.
- Are the conditional mean functions linear?
- Which of the three would you say is a better depiction of a real-life job market?

The problem

In Exercise 1, we used knowledge of the joint distribution to derive the CMF. This is a benefit that we never have, since in any real-life application D is unknown. We will assume that we have access to an *independent and identically distributed* (i.i.d.) sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ drawn from D. What can we say about the CMF using only the information contained in the sample?

[1] Not an innocuous assumption by any means. See the argmin blog by Ben Recht for recent work suggesting that it is also not necessary.

Visual inspection and binscatters

A reasonable first step in cases in which you are interested in a single predictor –or at most two– is to create a scatter plot. By plotting every data point in the sample, one obtains a visualization of D. However scatter plots have several limitations and have generally fallen out of favor in regression analysis. In larger datasets the cloud of points becomes increasingly dense and results in uninformative plots. Even for moderately sized but noisy samples it can be difficult to visually asses the shape of the CMF.

A helpful alternative is to construct a binned scatter plot, also known as binscatter.[2] In this type of scatter plots, the support of $\mathbf{X}$ is partitioned into a modest number of bins and only one point is plotted for each bin, usually showing the average outcome for all observations falling within the bin.

Exercise 2

In the companion notebook, implement a function that creates a binscatter plot and use it to inspect a random sample of size $n = 10,000$ from each of the three distributions (A, B and C). You'll find that the code to draw the random samples is already implemented. You only need to complete the body of the plotting function (we provide the signature and some scaffolding). We suggest:

- Create uniformly spaced bins that are closed to the left and open to the right.
- Use a width 0.5 years, starting at 8 years and ending at 20 years.
- For the plot, mantain the scale of the y-axis that we used for the previous plots [20k,100k].

Simple linear regression

Let's try to move beyond simple visual inspection and instead estimate a function $f : X \mapsto Y$ that is a *good approximation* of the CMF using the sample information. We will use a deceptively simple but powerful method called *simple linear regression*. The name carries a lot of information. It is *simple* because there is only one predictor variable (as in our schooling example). It is a *regression* because we are trying to estimate the CMF. Last but definitely not least, it is *linear* because we will only consider functions that are linear in two scalar parameters β_0 (intercept) and β_1 (slope):

$$\mathcal{L} = \{\beta_0 + \beta_1 x \mid \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}\}.$$

To choose among all the possible functions in $\mathcal{L}$, we will find the parameter estimates ($\hat{\beta}_0$,$\hat{\beta}_1$) that minimize the Mean Squared Error (MSE). That is:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0 \in \mathbb{R},\, \beta_1 \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \tag{1}$$

Exercise 3

Show that the solution to this problem is given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

[2] If you find binscatters interesting, consider looking at the binsreg package by Matias Cattaneo and coauthors.

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ We suggest that you begin by establishing the first order conditions (partial derivatives equal to zero) and then solve the two unknowns - two equations system. For the last stretch of the proof, it is useful to realize that:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Exercise 4

In the companion notebook, implement the simple linear regression solution. Then, for each distribution (A,B,C), display the regression line and the cmf together in a single plot. As usual, we provide the function signature and some scaffolding, you just need to fill in the blanks.

A bit of geometric intuition

Some deeper questions motivated by the simple linear regression results:

- There a unique solution (if you are not convinced verify the second order condition). How come? What is the intuition behind this result?
- Is this solution specific to simple linear regression? Can we generalize it to other problems? For instance, can we recycle this solution and use it to find linear parameters when we have more than one predictor ($k > 1$, also called *multiple linear regression*)?
- Can we avoid calculus altogether and find a straightforward solution?

The key to answering all this questions is realizing that optimization and approximation problems can be solved using a bit of linear algebra.[3] Here is an informal statement of the landmark result that underpins linear regression.

Informal Statement of the Projection Theorem
Given a point y and a plane $\mathcal{L}$, (i) there is a unique point $\hat{y}$ in $\mathcal{L}$ that minimizes the distance to y. The error vector $y - \hat{y}$ is perpendicular to any vector in $\mathcal{L}$.

Exercise 5

Consider the following examples:

- Think of the 3D space in this room as a subset of $\mathbb{R}^3$. We will locate the origin in the door, with the increasing x direction towards the front of South Hall (east), the increasing y direction towards Doe Library (north) and the increasing z direction upwards. Suppose that the floor of the room is completely flat. What is the point in the floor that is closer to the tip of your right-hand index finger?
- Also in $\mathbb{R}^3$, what is the closest point in the horizontal plane that goes through the origin ($z = 0$) to the point $\mathbf{y} = (1, 1, 1)$?

[3] You can find all the technical detail in this econometrics course notes by Bryan Graham

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x_2} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \hat{\mathbf{y}} = \beta_1 x_1 + \beta_2 x_2$$

Propose a solution $\hat{\boldsymbol{\beta}}$. Can you justify it using the projection theorem?

The OLS estimator

It turns out that linear regression is nothing more than a projection and that we can use the same tricks to derive the general OLS estimator (also called the matrix form estimator). The key thing to realize is that in linear regression problems we are projecting the column of sample values for the outcomes onto the plane spawned by the columns of the predictors (augmented with a constant column so that the regression function doesn't need to go through the origin).

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \hat{\mathbf{Y}} = \mathbb{X}\boldsymbol{\beta}$$

The orthogonality condition that $\hat{\boldsymbol{\beta}}$ must meet is:

$$(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}})^T \mathbb{X}\boldsymbol{\beta} = 0 \quad \text{for any } \boldsymbol{\beta} \in \mathbb{R}^{k+1}$$

Choosing the betas smartly yields a system of k+1 equations with k+1 unknowns. Under some regularity conditions, we can solve for $\hat{\boldsymbol{\beta}}$:

$$\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$$

Exercise 6

In the companion notebook, complete the function implementing the OLS estimator. Then, test your function using the same samples that you used in exercise 4. Do you obtain the same results?

Notation

a	scalar
$\mathbf{a}$	vector
$\mathbb{A}$	matrix
X	random variable in $\mathbb{R}$
$\mathbf{X}$	random vector in $\mathbb{R}^k$
$\mathbb{X}$	random matrix in $\mathbb{R}^n \times \mathbb{R}^k$
$\mathbb{X}^T$	transpose of matrix X
$\mathbb{X}^{-1}$	inverse of (square) matrix $\mathbb{X}$
$\mathbb{E}[X]$	expectation of X
$\mathbb{E}[Y\|X = x]$	conditional expectation of Y given X