1 Linear Regression

1.1 Introduction

Suppose that we are trying to study two quantities x and y that we suspect are related – at least approximately – by a linear equation $y = ax + b$. Sometimes this linear relationship is predicted by theoretical considerations, and sometimes it is just an empirical hypothesis.

For example, if we are trying to determine the velocity of an object travelling towards us at constant speed, and we measure measure the distances $d_1, d_2, \ldots, d_n$ between us and the object at a series of times $t_1, t_2, \ldots, t_n$, then since "distance equals rate times time" we have a theoretical foundation for the assumption that $d = rt + b$ for some constants m and b. On the other hand, because of unavoidable experimental errors, we can't expect that this relationship will hold exactly for the observed data; instead, we likely get a graph like that shown in fig. 1.

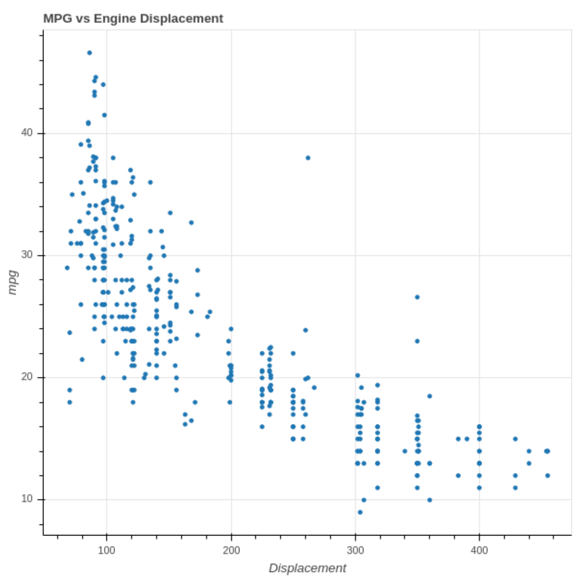

Figure 1: Physics Experiment

On the other hand, we might look at a graph such as fig. 2, which plots the gas mileage of various car models against their engine size (displacement), and observe a general trend in which bigger engines get lower mileage. In this situation we could ask for the best line of the form $y = mx + b$ that captures this relationship and use that to make general conclusions without necessarily having an underlying theory.

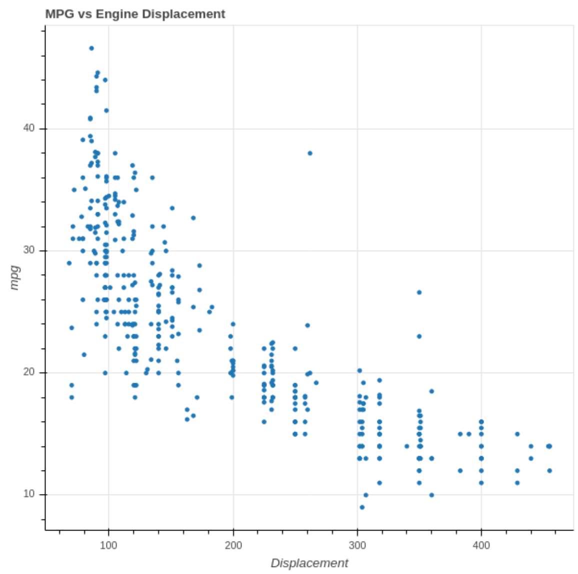

Figure 2: MPG vs Displacement ([1])

1.2 Least Squares (via Calculus)

In either of the two cases above, the question we face is to determine the line $y = mx + b$ that "best fits" the data $\{(x_i, y_i)_{i=1}^N\}$. The classic approach is to determine the equation of a line $y = mx + b$ that minimizes the "mean squared error":

$$MSE(m, b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - mx_i - b)^2$$

It's worth emphasizing that the MSE is a function of two variables – the slope m and the intercept b – and that the data points $\{(x_i, y_i)\}$ are constants for these purposes. Furthermore, it's a quadratic function in those two variables. Since our goal is to find m and b that minimize the MSE, we have a Calculus problem that we can solve by taking partial derivatives and setting them to zero.

To simplify the notation, let's abbreviate MSE by E.

$$\frac{\partial E}{\partial m} = \frac{1}{N}\sum_{1}^{N} -2x_i(y_i - mx_i - b)$$
$$\frac{\partial E}{\partial b} = \frac{1}{N}\sum_{1}^{N} -2(y_i - mx_i - b)$$

We set these two partial derivatives to zero, so we can drop the -2 and regroup the sums to obtain two equations in two unknowns (we keep the $\frac{1}{N}$ because it is illuminating in the final result):

$$\begin{aligned}
\frac{1}{N}(\sum_{i=1}^{N} x_i^2)m + \quad \frac{1}{N}(\sum_{i=1}^{N} x_i)b &= \frac{1}{N}\sum_{i=1}^{N} x_i y_i \\
\frac{1}{N}(\sum_{i=1}^{N} x_i)m + \quad b &= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned} \tag{1}$$

In these equations, notice that $\frac{1}{N}\sum_{i=1}^{N} x_i$ is the average (or mean) value of the x_i. Let's call this $\overline{x}$. Similarly, $\frac{1}{N}\sum_{i=1}^{N} y_i$ is the mean of the y_i, and we'll call it $\overline{y}$. If we further simplify the notation and write S_{xx} for $\frac{1}{N}\sum_{i=1}^{N} x_i^2$ and S_{xy} for $\frac{1}{N}\sum_{i=1}^{N} x_i y_i$ then we can write down a solution to this system using Cramer's rule:

$$\begin{aligned}
m &= \frac{S_{xy} - \overline{xy}}{S_{xx} - \overline{x}^2} \\
b &= \frac{S_{xx}\overline{y} - S_{xy}\overline{x}}{S_{xx} - \overline{x}^2}
\end{aligned} \tag{2}$$

where we must have $S_{xx} - \overline{x}^2 \neq 0$.

1.2.1 Exercises

1. Verify that eq. 2 is in fact the solution to the system in eq. 1.
2. Suppose that $S_{xx} - \overline{x}^2 = 0$. What does that mean about the x_i? Does it make sense that the problem of finding the "line of best fit" fails in this case?

1.3 Least Squares (via Geometry)

In our discussion above, we thought about our data as consisting of N pairs (x_i, y_i) corresponding to n points in the xy-plane $\mathbf{R}^2$. Now let's turn that picture "on its side", and instead think of our data as consisting of *two* points in $\mathbf{R}^n$:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Let's also introduce one other vector

$$E = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

First, let's assume that E and X are linearly independent. If not, then X is a constant vector (why?) which we already know is a problem from section 1.2, Exercise 2. Therefore E and X span a plane in $\mathbf{R}^n$.

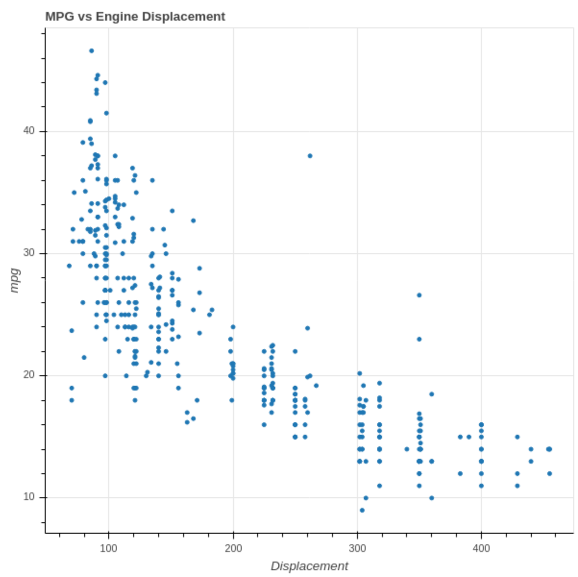

Figure 3: Distance to A Plane

Now if our data points (x_i, y_i) all *did* lie on a line $y = mx + b$, then the three vectors X, Y, and E would be linearly dependent:

$$Y = mX + bE.$$

Since our data is only approximately linear, that's not the case. So instead we look for an approximate solution. One way to phrase that is to ask:

What is the point $\hat{Y}$ in the plane H spanned by X and E in $\mathbf{R}^n$ which is closest to Y?

If we knew this point $\hat{Y}$, then since it lies in H we would have $\hat{Y} = mX + bE$ and the coefficients m and b would be a candidate for defining a line of best fit

$y = mx + b$. Finding the point in a plane closest to another point in $\mathbf{R}^n$ is a geometry problem that we can solve.

Proposition: The point $\hat{Y}$ in the plane spanned by X and E is the point such that the vector $Y - \hat{Y}$ is perpendicular to H.

Proof: See fig. 3 for an illustration – perhaps you are already convinced by this, but let's be careful. $\hat{Y} = mX + bE$ such that

$$D = \|Y - \hat{Y}\|^2 = \|Y - mX - bE\|^2$$

is minimal. Using some vector calculus, we have

$$\frac{\partial D}{\partial m} = \frac{\partial}{\partial m}(Y - mX - bE) \cdot (Y - mX - bE) = -2(Y - mX - bE) \cdot X$$

and

$$\frac{\partial D}{\partial b} = \frac{\partial}{\partial b}(Y - mX - bE) \cdot (Y - mX - bE) = -2(Y - mX - bE) \cdot E.$$

So both derivatives are zero exactly when $\hat{Y} = (Y - mX - bE)$ is orthogonal to both X and E, and therefore every vector in H.

We also obtain equations for m and b just as in our first look at this problem.

$$\begin{aligned} m(X \cdot E) + b(E \cdot E) &= (Y \cdot E) \\ m(X \cdot X) + b(E \cdot X) &= (Y \cdot X) \end{aligned} \tag{3}$$

We leave it is an exercise below to check that these are the same equations that we obtained in eq. 2.

1.3.1 Exercises

1. Verify that eq. 2 and eq. 3 are equivalent.

1.4 The Multivariate Case (Calculus)

Having worked through the problem of finding a "line of best fit" from two points of view, let's look at a more general problem. We looked above at a scatterplot showing the relationship between gas mileage and engine size (displacement). There are other factors that might contribute to gas mileage that we want to consider as well – for example: - a car that is heavy compared to its engine size may get worse mileage - a sports car with a drive train that gives fast acceleration as compared to a car with a transmission designed for long trips may have different mileage for the same engine size.

Suppose we wish to use engine displacement, vehicle weight, and acceleration all together to predict mileage. Instead of looking points (x_i, y_i) where x_i is the displacement of the i^{th} car model and we try to predict a value y from a corresponding x as $y = mx + b$ – let's look at a situation in which our measured value y depends on multiple variables – say displacement d, weight w, and acceleration a with $k = 3$ – and we are trying to find the best linear equation

$$y = m_1 d + m_2 w + m_3 a + b \tag{4}$$

But to handle this situation more generally we need to adopt a convention that will allow us to use indexed variables instead of d, w, and a. We will use the *tidy* data convention.

Tidy Data: A dataset is tidy if it consists of values x_{ij} for $i = 1, \ldots, N$ and $j = 1, \ldots, k$ so that:

- the row index corresponds to a *sample* – a set of measurements from a single event or item;
- the column index corresponds to a *feature* – a particular property measured for all of the events or items.

In our case,

- the *samples* are the different types of car models,
- the *features* are the properties of those car models.

For us, N is the number of different types of cars, and k is the number of properties we are considering. Since we are looking at displacement, weight, and acceleration, we have $k = 3$.

So the "independent variables" for a set of data that consists of N samples, and k measurements for each sample, can be represented by a $N \times k$ matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{k2} & \cdots & x_{Nk} \end{pmatrix}$$

and the measured dependent variables Y are a column vector

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

If $m_1, \ldots, m_k$ are "slopes" associated with these properties in eq. 4, and b is the "intercept", then the predicted value $\hat{Y}$ is given by a matrix equation

$$\hat{Y} = X \begin{bmatrix} m_1 \\ m_2 \\ \cdots \\ m_k \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} b$$

and our goal is to choose these parameters m_i and b to make the mean squared error:

$$MSE(m_1, \ldots, m_k, b) = \|Y - \hat{Y}\|^2 = \sum_{i=1}^{N} (y_i - \sum_{j=1}^{k} x_{ij} m_j - b)^2.$$

Here we are summing over the N different car models, and for each model taking the squared difference between the true mileage y_i and the "predicted" mileage $\sum_{j=1}^{k} x_{ij} m_j + b$. We wish to minimize this MSE.

Let's make one more simplification. The intercept variable b is annoying because it requires separate treatment from the m_i. But we can use a trick to eliminate the need for special treatment. Let's add a new feature to our data matrix (a new column) that has the constant value 1.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & 1 \\ x_{21} & x_{22} & \cdots & x_{2k} & 1 \\ \vdots & \vdots & \ddots & \vdots & 1 \\ x_{N1} & x_{k2} & \cdots & x_{Nk} & 1 \end{pmatrix}$$

Now our data matrix X is $N \times (k+1)$ and we can put our "intercept" $b = m_{k+1}$ into our vector of "slopes" $m_1, \ldots, m_k, m_{k+1}$:

$$\hat{Y} = X \begin{bmatrix} m_1 \\ m_2 \\ \cdots \\ m_k \\ m_{k+1} \end{bmatrix}.$$

and our MSE becomes

$$MSE(M) = \|Y - XM\|^2$$

where

$$M = \begin{bmatrix} m_1 \\ m_2 \\ \cdots \\ m_k \\ m_{k+1} \end{bmatrix}.$$

The Calculus approach to minimizing the MSE is to take its partial derivatives with respect to the m_i and set them to zero. Let's first work out the derivatives in a nice form for later.

Proposition: The gradient of $MSE(M) = E$ is given by

$$\nabla E = \begin{bmatrix} \frac{\partial}{\partial m_1} E \\ \frac{\partial}{\partial m_2} E \\ \vdots \\ \frac{\partial}{\partial m_{k+1}} E \end{bmatrix} = -2X^\intercal Y + 2X^\intercal X M \tag{5}$$

where $X^\intercal$ is the transpose of X.

Proof: First, remember that the ij entry of $X^\intercal$ is the ji entry of X. Also, we will use the notation $X[j,:]$ to mean the j^{th} row of X and $X[:,i]$ to mean the i^{th} column of X. (This is copied from the Python programming language; the ':' means that index runs over all possibilities).

Since

$$E = \sum_{j=1}^{N} (Y_j - \sum_{s=1}^{k+1} X_{js} M_s)^2$$

we compute:

$$\begin{aligned} \frac{\partial}{\partial m_t} E &= -2 \sum_{j=1}^{N} X_{jt} (Y_j - \sum_{s=1}^{k+1} X_{js} M_s) \\ &= -2 (\sum_{j=1}^{N} Y_j X_{jt} - \sum_{j=1}^{N} \sum_{s=1}^{k+1} X_{jt} X_{js} M_s) \\ &= -2 (\sum_{j=1}^{N} X^\intercal_{tj} Y_j - \sum_{j=1}^{N} \sum_{s=1}^{k+1} X^\intercal_{tj} X_{js} M_s) \\ &= -2 (X^\intercal[t,:] Y - \sum_{s=1}^{k+1} \sum_{j=1}^{N} X^\intercal_{tj} X_{js} M_s) \\ &= -2 (X^\intercal[t,:] Y - \sum_{s=1}^{k+1} (X^\intercal X)_{ts} M_s) \\ &= -2 (X^\intercal[t,:] Y - (X^\intercal X)[t,:] M) \end{aligned} \tag{6}$$

Stacking up the different rows to make E yields the desired formula.

Proposition: Assume that $D = X^{\intercal}X$ is invertible (notice that it is a $(k+1) \times (k+1)$ square matrix so this makes sense). The solution M to the multivariate least squares problem is

$$M = D^{-1}X^{\intercal}Y$$

and the "predicted value" $\hat{Y}$ for Y is

$$\hat{Y} = XD^{-1}X^{\intercal}Y. \tag{7}$$

1.5 The Multivariate Case (Geometry)

Let's look more closely at the equation obtained by setting the gradient of the error, eq. 6, to zero. Remember that M is the unknown vector in this equation, everything else is known:

$$X^{\intercal}Y = X^{\intercal}XM$$

Here is how to think about this:

1. As M varies, the $N \times 1$ matrix XM varies over the space spanned by the columns of the matrix X. So as M varies XM is a general element of the subspace H of R^N spanned by the $k+1$ columns of X.
2. The product $X^{\intercal}XM$ is a $(k+1) \times 1$ matrix. Each entry is the dot product of the general element of H with one of the $k+1$ basis vectors of H.
3. The product $X^{\intercal}Y$ is a $(k+1) \times 1$ matrix whose entries are the dot product of the basis vectors of H with Y.

Therefore, this equation asks for us to find M so that the vector XM in H has the same dot products with the basis vectors of H as Y does. The condition

$$X^{\intercal} \cdot (Y - XM) = 0$$

says that $Y - XM$ is orthogonal to H. This argument establishes the following proposition.

Proposition: Just as in the simple one-dimensional case, the predicted value $\hat{Y}$ of the least squares problem is the point in H closest to Y – or in other words the point $\hat{Y}$ in H such that $Y - \hat{Y}$ is perpendicular to H.

1.5.1 Orthogonal Projection

Recall that we introduced the notation $D = X^\intercal X$, and let's assume, for now, that D is an invertible matrix. We have the formula (see eq. 7):

$$\hat{Y} = XD^{-1}X^\intercal Y.$$

Proposition: The matrix $P = XD^{-1}X^\intercal$ is an $N \times N$ matrix called the orthogonal projection operator onto the subspace H spanned by the columns of X. It has the following properties:

- PY belongs to the subspace H for any $Y \in \mathbf{R}^N$.
- $(Y - PY)$ is orthogonal to H.
- $P * P = P$.

Proof: First of all, $PY = XD^{-1}X^\intercal Y$ so PY is a linear combination of the columns of X and is therefore an element of H. Next, we can compute the dot product of PY against a basis of H by computing

$$X^\intercal PY = X^\intercal XD^{-1}X^\intercal Y = X^\intercal Y$$

since $X^\intercal X = D$. This equation means that $X^\intercal(Y - PY) = 0$ which tells us that $Y - PY$ has dot product zero with a basis for H. Finally,

$$PP = XD^{-1}X^\intercal XD^{-1}X^\intercal = XD^{-1}X^\intercal = P.$$

It should be clear from the above discussion that the matrix $D = X^\intercal X$ plays an important role in the study of this problem. In particular it must be invertible or our analysis above breaks down. In the next section we will look more closely at this matrix and what information it encodes about our data.

1.5.2 Variance, Covariance and the Covariance Matrix

Recall from last section that the matrix $D = X^\intercal X$ is of central importance to the study of the multivariate least squares problem. Let's look at it more closely.

Lemma: The i, j entry of D is the dot product

$$D_{ij} = X[:, i] \cdot X[:, j]$$

of the i^{th} and j^{th} columns of X.

Proof: In the matrix multiplication $X^\intercal X$, the i^{th} row of $X^\intercal$ gets "dotted" with the j^{th} column of X to product the i, j entry. But the i^{th} row of $X^\intercal$ is the i^{th} column of X, as asserted in the statement of the lemma.

A crucial point in our construction above relied on the matrix D being invertible. The following Lemma shows that D fails to be invertible only when the different features (the columns of X) are linearly dependent.

Lemma: D is not invertible if and only if the columns of X are linearly dependent.

Proof: If the columns of X are linearly dependent, then there is a nonzero vector m so that $Xm = 0$. In that case clearly $Dm = X^{\intercal}Xm = 0$ so D is not invertible. Suppose D is not invertible. Then there is a nonzero vector m with $Dm = X^{\intercal}Xm = 0$. This means that the vector Xm is orthogonal to all of the columns of X. That in turn means that $Xm = 0$ so the columns of X are linearly dependent.

In fact, the matrix D captures some important statistical measures of our data, but to see this clearly we need to make a slight change of basis. First recall that $X[:, k+1]$ is our column of all 1, added to handle the intercept. As a result, the dot product $X[:, i] \cdot X[:, k+1]$ is the sum of the entries in the i^{th} column, and so if we let μ_i denote the average value of the entries in column i, we have

$$\mu_i = \frac{1}{N}(X[:, i] \cdot X[:, k+1])$$

Now change the matrix X by elementary column operations to obtain a new data matrix X_0 by setting

$$X_0[:, i] = X[:, i] - \frac{1}{N}(X[:, i] \cdot X[:, k+1])X[:, k+1] = X[:, i] - \mu_i X[:, k+1]$$

for $i = 1, \ldots, k$.

In terms of the original data, we are changing the measurement scale of the data so that each feature has average value zero, and the subspace H spanned by the columns of X_0 is the same as that spanned by the columns of X. Using X_0 instead of X for our least squares problem, we get

$$\hat{Y} = X_0 D_0^{-1} X_0^{\intercal} Y$$

and

$$M_0 = D_0^{-1} X_0^{\intercal} Y$$

where $D_0 = X_0^{\intercal} X_0$.

Proposition: The matrix D_0 has a block form. Its upper left block is $k \times k$ symmetric block with entries

$$(D_0)_{ij} = (X[:, i] - \mu_i X[:, k+1]) \cdot (X[:, j] - \mu_j X[:, k+1]).$$

Its $(k+1)^{st}$ row and column are all zero, except for the $(k+1),(k+1)$ entry, which is N.

Proof: This follows from the fact that the last row and column entries are (for $i \neq k+1$):

$$(X[:,i] - \mu_i X[:,k+1]) \cdot X[:,k+1] = (X[:,i] \cdot X[:,k+1]) - N\mu_i = 0$$

and for $i = k+1$ we have $X[:,k+1] \cdot X[:,k+1] = N$ since that column is just N 1's.

1.5.3 Exercises

1. When proving that D is invertible if and only if the columns of X are linearly independent, we argued that if $X^{\intercal}Xm = 0$ for a nonzero vector m, it must be the case that Xm is orthogonal to every column of X and so $Xm = 0$. Flesh out this argument. For example, why is it the case that if $v \in \mathbf{R}^N$ is a linear combination of vectors $x_1, \ldots, x_k$ in $\mathbf{R}^N$, and $v \cdot x_i = 0$ for $i = 1, \ldots, k$, then $v = 0$?

References

[1] U.C. Irvine ML Repository. Auto MPG Dataset.Available at http://https://archive.ics.uci.edu/ml/datasets/Auto+MPG.