IEOR 165 – Lecture 3
Linear Regression

1 Estimating a Linear Model

Recall that the amount of energy consumed in a building on the i-th day E_i heavily depends on the outside temperature on the i-th day T_i. Generally, there are two situations. When the outside temperature is low, decreasing temperature leads to increased energy consumption because the building needs to provider greater amounts of heating. When the outside temperature is high, increasing temperature leads to increased energy consumption because the building needs to provide greater amounts of cooling. This is a general phenomenon, and is empirically observed with the real data shown in Figure 1, which was collected from Sutardja Dai Hall.

Based on the figure, we might guess that when the outside temperature is below 59°F, the relationship between energy consumption and outside temperature is given by

$$E = a \cdot T + b + \epsilon,$$

where a, b are unknown constants, and ϵ is a zero-mean/finite-variance random variable that is independent of T. The method of moments estimator for the unknown constants is given by

$$\hat{a} = \frac{(\frac{1}{n}\sum_{i=1}^n E_i)(\frac{1}{n}\sum_{i=1}^n T_i) - \frac{1}{n}\sum_{i=1}^n E_i T_i}{(\frac{1}{n}\sum_{i=1}^n T_i)^2 - \frac{1}{n}\sum_{i=1}^n T_i^2}$$
$$\hat{b} = \frac{(\frac{1}{n}\sum_{i=1}^n E_i T_i)(\frac{1}{n}\sum_{i=1}^n T_i) - (\frac{1}{n}\sum_{i=1}^n E_i)(\frac{1}{n}\sum_{i=1}^n T_i^2)}{(\frac{1}{n}\sum_{i=1}^n T_i)^2 - \frac{1}{n}\sum_{i=1}^n T_i^2}.$$

However, its derivation is clumsy. The question is whether there is a conceptually simpler approach to estimating parameters of a linear model.

1.1 Abstract Model

We can abstract the problem of estimating parameters of a linear model into the following mathematical setting: Suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are independent and identically distributed (iid) pairs of random variables from some unknown distribution with cdf $F_{(X,Y)}(u)$. We should think of the (x_i, y_i) as n independent measurements from a single unknown joint distribution, meaning the measured random variables x, y are generally dependent. The mathematical issue we are interested in studying is how to determine the parameters a, b if we assume the relationship between x and y is given by

$$y = a \cdot x + b + \epsilon,$$

where ϵ is zero-mean noise with finite variance that is independent of x, y. We will interchangeably refer to the x random variable as either an *input* variable or as a *predictor*. Similarly, we will

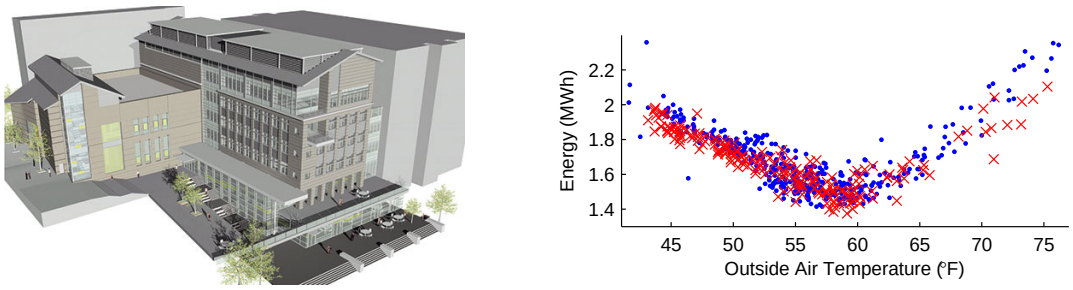


Figure 1: Real data (right) from Sutardja Dai Hall (left) is shown. The data marked with dots and crosses denote measurements made running two different versions of software on the heating, ventilation, and air-conditioning (HVAC) system in Sutardja Dai Hall.

interchangeably refer to the y random variable as either an *output* variable or as a *response* variable.

1.2 Method of Least Squares

One approach with a long history is known as the method of least squares. The intuition of this approach is to pick parameters $\hat{a}, \hat{b}$ which make the estimated response $\hat{y}_i = \hat{a} \cdot x_i + \hat{b}$ "close" to the measured response y_i, because this would be expected if the parameters are an accurate representation of the relationship between the input and output variables. This immediately suggests the question of how can we measure closeness? It turns out that there are many (in fact an infinite) number of ways to measure closeness. However, one popular choice is known as *squared loss function*, which measures the closeness of two numbers u, v using the square of their differences $(u - v)^2$.

In the method of least squares, we estimate $\hat{a}, \hat{b}$ using the above intuition and by measuring closeness using the squared loss function. That is, we determine estimates of our parameters $\hat{a}, \hat{b}$ by ensuring $(y_i - \hat{y}_i)^2$ is small. We can make this intuition more rigorous by considering the following optimization problem

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - a \cdot x_i - b)^2.$$

We include a $\frac{1}{n}$ and sum $(y_i - a \cdot x_i - b)^2$ over all $i = 1, \ldots, n$ measurements because this makes the objective function of the optimization problem converge to $\mathbb{E}((y - a \cdot x - b)^2)$ by the lln as n goes to infinity.

One reason for the initial popularity of this approach is that this optimization problem has a simple solution. Since the objective function is smooth, we can take its gradient with respect to (a, b), set the gradient equal to zero, and solve for the corresponding values of (a, b). If we let

$J(a, b)$ be the objective function of the above optimization problem, then its gradient is equal to

$$\nabla_{(a,b)} J = \begin{bmatrix} \frac{2}{n}\sum_{i=1}^{n} -x_i \cdot (y_i - a \cdot x_i - b) \\ \frac{2}{n}\sum_{i=1}^{n} -(y_i - a \cdot x_i - b) \end{bmatrix} = \begin{bmatrix} \frac{2}{n}\sum_{i=1}^{n} -x_i y_i \\ \frac{2}{n}\sum_{i=1}^{n} -y_i \end{bmatrix} + \begin{bmatrix} \frac{2}{n}\sum_{i=1}^{n} x_i^2 & \frac{2}{n}\sum_{i=1}^{n} x_i \\ \frac{2}{n}\sum_{i=1}^{n} x_i & \frac{2}{n}\sum_{i=1}^{n} 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}.$$

Setting the gradient equal to zero and solving for (a, b) gives

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} n & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \cdot \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

Simplifying the expression for $(\hat{a}, \hat{b})$, we have

$$\hat{a} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{b} = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Interestingly, the expressions for $(\hat{a}, \hat{b})$ derived using the method of least squares are the same as the expressions derived from the method of moments.

Given this equivalence between the equations derived using two approaches, the natural question to ask is why is least squares generally preferred over the method of moments when constructing linear models. There are two reasons. The first is that the method of least squares is conceptually simpler in the sense that we formulate an optimization problem that (i) captures our intuition about what characteristics a good set of model parameters should satisfy, and (ii) can be solved using straightforward calculus and arithmetic. The second reason that the method of least squares is generally preferred is that it is easier to derive the generalization to the case of multivariate linear models, which are linear models with more than one predictor.

1.3 Example: Building Energy Model

Suppose we have $n = 4$ measurements (x_i, y_i) of $(45, 2), (50, 1.8), (55, 1.6), (59, 1.6)$. Then

$$\begin{aligned} \textstyle\sum_{i=1}^{n} x_i &= 45 + 50 + 55 + 59 = 209 \\ \textstyle\sum_{i=1}^{n} y_i &= 2 + 1.8 + 1.6 + 1.6 = 7 \\ \textstyle\sum_{i=1}^{n} x_i^2 &= 45^2 + 50^2 + 55^2 + 59^2 = 11031 \\ \textstyle\sum_{i=1}^{n} x_i y_i &= 45 \cdot 2 + 50 \cdot 1.8 + 55 \cdot 1.6 + 59 \cdot 1.6 = 362.4 \end{aligned}$$

Substituting these into the equations for $(\hat{a}, \hat{b})$ give

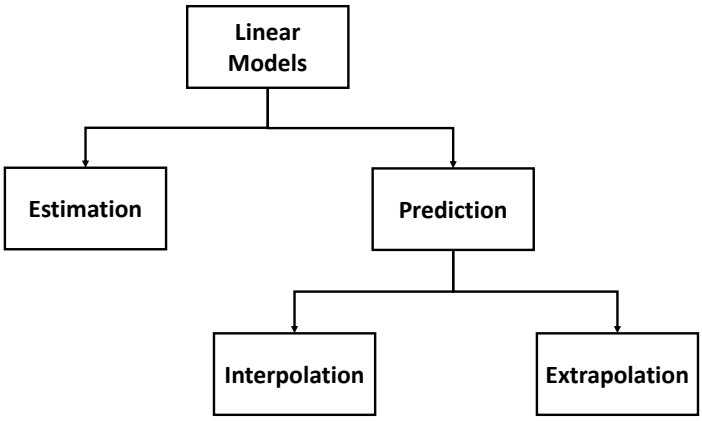
$$\hat{a} = \frac{n\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{4 \cdot 362.4 - 209 \cdot 7}{4 \cdot 11031 - (209)^2} = -0.03$$
$$\hat{b} = \frac{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{11031 \cdot 7 - 209 \cdot 362.4}{4 \cdot 11031 - (209)^2} = 3.33.$$

Thus, the estimated model for energy consumption for temperatures below 59°F is given by

$$E = -0.03 \cdot T + 3.33.$$

2 Purposes of Linear Models

Linear models are perhaps the most widely used statistical model, and so it is useful to discuss some of the different purposes that linear models are used for. A summary is shown below.



2.1 Estimation

The purpose of estimation is to determine the "true" parameters of the linear model. The reason this might be of interest is to determine the direction or magnitude of an effect between the predictor and the response variable. For instance, we might be interested in studying the influence of birth weight of newborns with the corresponding infant mortality rate. In this scenario, the x_i data would consist of birth weight (in units of say kg) of the i-th newborn, and the y_i data would be 0 (or 1) if the i-th newborn did not (did) die within 1 year after birth. If the a parameter was negative, then we would conclude that lower birth weight corresponds to increased infant mortality. Furthermore, if the a parameter was "large", then we would conclude that the impact of birth weight on infant mortality was large.

However, caution must be taken when interpreting the estimated parameter values. It is an oft-repeated maxim of statistics that "Correlation does not equal causation". We must be careful

when interpreting estimated parameters to ensure that we have some prior knowledge before hand regarding the causality of the predictor in impacting the response variable. Otherwise, it is very easy to make incorrect conclusions.

2.2 Prediction

The purpose of prediction is to determine how the response changes as the inputs change. There are two different types of predictions. The first is *interpolation*, and the idea is that we would like to determine how the response changes as the input changes, for a set of inputs that are close to values that have already been observed. One example scenario is if we are identifying a model where the input is the date (in the units of "year") and the output is the amount of newspapers purchased in the United States. Suppose the data we have available for the model are the years 1880–2010, and that we use this data to estimate the parameters of the model. If we would now like to make a prediction of the number of newspapers purchased in 1980, then this would be an example of interpolation. Interpolation can alternatively be thought of as smoothing the measured data.

The second type of prediction is *extrapolation*. In the above newspaper scenario, an example of extrapolation would be trying to predict the number of newspapers that will be purchased in 2020. From a practical standpoint, extrapolation is less accurate than interpolation. The reason is that in extrapolation, we are trying to use our linear model to determine what will happen in situations that have not been seen before. And since we have little data for these novel situations, our model is less likely to produce an accurate prediction.

3 Ordinary Least Squares

We next generalize the method of least squares to the setting of multivariate linear models, where there are multiple predictors for a single response variable. In particular, suppose that we have pairs of iid measurements (x_i, y_i) for $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, and that the system is described by a linear model

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i = x_i' \beta + \epsilon_i,$$

where x_i' denotes the transpose of x_i. Ordinary least squares (OLS) is a method to estimate the unknown parameters $\beta \in \mathbb{R}^p$ given our n measurements. Because the y_i are noisy measurements (whereas the x_i are not noisy in this model), the intuitive idea is to choose an estimate $\hat{\beta} \in \mathbb{R}^p$ which minimizes the difference between the measured y_i and the estimated $\hat{y}_i = x_i' \hat{\beta}$.

There are a number of ways that we could characterize this difference. For mathematical and computational reasons, a popular choice is the *squared loss*: This difference is quantified as

$\sum_i (y_i - \hat{y}_i)^2$, and the resulting problem of choosing $\hat{\beta}$ to minimize this difference can be cast as the following (unconstrained) optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2.$$

For notational convenience, we will define a matrix $X \in \mathbb{R}^{n \times p}$ and a vector $Y \in \mathbb{R}^n$ such that the i-th row of X is x_i' and the i-th row of Y is y_i. With this notation, the OLS problem can be written as

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2,$$

where $\|\cdot\|_2$ is the usual L^2-norm. (Recall that for a vector $v \in \mathbb{R}^k$ the L^2-norm is $\|v\|_2 = \sqrt{(v^1)^2 + \ldots + (v^k)^2}$.)

Now given this notation, we can solve the above defined optimization problem. Because the problem is unconstrained, setting the gradient of the objective to zero and solving the resulting algebraic equation will give the solution. For notational convenience, we will use the function $J(X, Y; \beta)$ to refer to the objective of the above optimization problem. Computing its gradient gives

$$\begin{aligned} \nabla_\beta J = 2X'(Y - X\hat{\beta}) = 0 &\Rightarrow X'X\hat{\beta} = X'Y \\ &\Rightarrow \hat{\beta} = (X'X)^{-1}(X'Y). \end{aligned}$$

This is the OLS estimate of β for the linear model. In some cases, the solution is written as $\hat{\beta} = (\frac{1}{n}X'X)^{-1}(\frac{1}{n}X'Y)$. The reason for this will be discussed in future lectures.

3.1 Geometric Interpretation of OLS

Recall the optimization formulation of OLS,

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2,$$

where the variables are as defined before. The basic tension in the problem above is that in general no exact solution exists to the linear equation

$$Y = X\beta;$$

otherwise we could use linear algebra to compute β, and this value would be a minimizer to the optimization problem written above.

Though no exact solution exists to $Y = X\beta$, an interesting question to ask is whether there is some related linear equation for which an exact solution exists. Because the noise is in Y and

not X, we can imagine that we would like to pick some $\hat{Y}$ such that $\hat{Y} = X\hat{\beta}$ has an exact solution. Recall from linear algebra, that this is equivalent to asking that $\hat{Y} \in \mathcal{R}(X)$ (i.e., $\hat{Y}$ is in the range space of X). Now if we think of $\hat{Y}$ as true signal, then we can decompose Y as

$$Y = \hat{Y} + \Delta Y,$$

where ΔY represents orthogonal noise. Because – from Fredholm's theorem in linear algebra – we know that the range space of X is orthogonal to the null space of X' (i.e., $\mathcal{R}(X) \perp \mathcal{N}(X')$), it must be the case that $\Delta Y \in \mathcal{N}(X')$ since we defined $\hat{Y}$ such that $\hat{Y} \in \mathcal{R}(X)$. As a result, premultiplying $Y = \hat{Y} + \Delta Y$ by X' gives

$$X'Y = X'\hat{Y} + X'\Delta Y = X'\hat{Y}.$$

The intuition is that premultiplying by X' removes the noise component. And because $\hat{Y} \in \mathcal{R}(X)$ and $\hat{Y} = X\hat{\beta}$, we must have that

$$X'Y = X'\hat{Y} = X'X\hat{\beta}.$$

Solving this gives $\hat{\beta} = (X'X)^{-1}(X'Y)$, which is our regular equation for the OLS estimate.

3.2 Example: Building Energy Model

It is known that other variables besides outside temperature affect building energy consumption. For instance, the day of the week significantly impacts consumption. There is reduced consumption on the weekends, and energy usage peaks during the middle of the week. As a result, we may wish to update our building energy model to include these effects as well. Generally speaking, there are other factors that also impact consumption, but for the sake of illustration we will ignore these.

One possible linear model is

$$E = \beta_1 \cdot T + \beta_2 \cdot W + \epsilon,$$

where T is outside temperature, and W is a variable that is equal to 0 (or 1) if the day is a weekend (or weekday). The β parameters and ϵ noise are as before. This is an example of a multivariate linear model. There is a new element in this example: A predictor that belongs to a finite set of classes is known as a *categorical* variable. In this case, the categorical variable is whether a day is a weekday or weekend, and we use a binary input variable to distinguish between these two possibilities. The approach of using a binary variable to distinguish between two possibilities is a common approach to modeling with two categorical variables. When there are three categorical variables, we would actually use two bineary input variables to distinguish the three cases.