DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 6: Probability Interpretation of Linear Models

Lecturer: Fernando Perez

1 Review of Linear Regression

Before developing a probabilitistic interpretation of linear models, we first review concepts from linear regression. This will help us draw the connection between conditional expectations, least square estimators, and the maximum likelihood estates of linear models clearer in the following lectures. We begin by going over the concept of regression, and then looking at training *linear* models with different loss functions.

1.1 Regression

The problem of regression is one of fitting a parametric model, $f_\theta : X \to Y$ to data to best explain the relationship between two variables x and y. That is, the goal of regression is to find the value of θ that best explains the observed data.

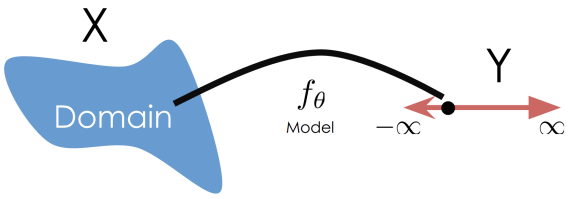


Figure 6.1: The general concept of regression

In the setup we consider, we allow x to be any type of data that can be made into a feature vector, and take y to be a scalar. Therefore, the model $f_\theta : X \to Y$ takes in x and outputs some quantitative value y. The model is learned or fitted using a labeled dataset of pairs $(x_1, y_1), ..., (x_n, y_n)$. This makes regression a form of *supervised learning*. The resulting problem has the form:

$$\min_\theta \sum_{i=1}^{n} L(f_\theta(x_i), y_i) \tag{6.1}$$

where L is the loss function. As we will see, different loss functions promote different properties in the learned model.

1.2 Linear Models

Since the choice of model is a design decision, one of the most common is that of a *linear* model. This is a model that is a linear function in the parameters θ having the form:

$$f_\theta(x) = \sum_{j=1}^{d} \theta_j \phi_j(x)$$

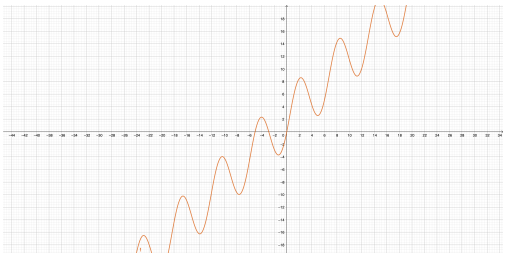
Recall that a linear function must satisfy two properties:

1. scalar multiplication: $f_{a\theta}(x) = a f_\theta(x)$ for all $a \in \mathbb{R}$.
2. addition: $f_{\theta_1+\theta_2}(x) = f_{\theta_1}(x) + f_{\theta_2}(x)$ for all θ_1, θ_2.

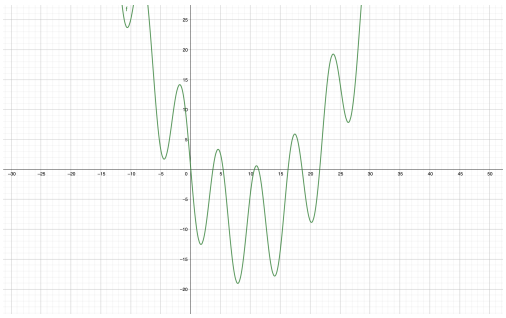
For compactness, we often write the linear model at a dot product between two vectors:

$$f_\theta(x) = \phi(x)^T \theta$$

Where both $\phi(x) \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$. This framework is broad enough to allow us to capture some rather complicated relationships in data, as seen in the functions in Figure 6.2.



(a)



(b)

Figure 6.2: These functions can be represented as linear combinations of features of x.

Example 6.1. Both functions in Figure 6.2 can be seen as linear functions of features of x:

- Figure 6.2a: $f_\theta(x) = x + 4cos(x) + 2sin(x) + 2$.

$$\phi(x) = \begin{bmatrix} x \\ cos(x) \\ sin(x) \\ 1 \end{bmatrix} \quad \theta = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 2 \end{bmatrix}$$

- Figure 6.2b: $f_\theta(x) = -2x - 10sin(x) + 0.1x^2 + 0.1$.

$$\phi(x) = \begin{bmatrix} x \\ sin(x) \\ x^2 \\ 1 \end{bmatrix} \quad \theta = \begin{bmatrix} -2 \\ -10 \\ 0.1 \\ 0.1 \end{bmatrix}$$

For a given set of observations $x_1, ..., x_n$, we can make predictions about $y_1, ..., y_n$, which we write as $\hat{y}_1, ..., \hat{y}_n$ using matrix multiplication:

$$\hat{Y} = \Phi\theta$$

Where:

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \quad , \quad \Phi = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_d(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_d(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \cdots & \phi_d(x_n) \end{bmatrix} = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}$$

Note that $\hat{Y}$ is an $n \times 1$ vector, Φ is an $n \times d$ vector, and θ is a $d \times 1$ vector, making the dimensions amenable to matrix multiplication. Given this setup, the problem of finding the best value of θ posed in 6.1 reduces to:

$$\min_\theta L(\Phi\theta, Y) \tag{6.2}$$

where, Y is the $n \times 1$ vector of labels $y_1, ..., y_n$, and L is the loss function.

Before discussing the effects of different loss functions on the fitted model, we first provide examples of linear models with different features.

Example 6.2. Suppose you would like to give insurance to individuals, and to do so you need to estimate their earning potential. For a given individual have data on:

1. Whether or not an individual is a smoker.
2. Whether or not an individual is has a high school diploma.
3. The number of credit cards they have c.

For individual x_i you can estimate their potential salary $\hat{y}$ as:

$$\hat{y} = \theta_1 + \theta_2\mathbb{I}(x_i \text{ is a smoker}) + \theta_3\mathbb{I}(x_i \text{ has high school diploma}) + \theta_4 c_i$$

Therefore linear models allow us to work even with non-numerical data.

1.3 Loss Functions

Returning to the problem of finding the best parameters for a given loss function, we now discuss various loss functions that we have seen in lecture. The loss function captures, as the name suggests, the error or loss resulting from a particular choice of model parameters. The choice of loss function depends on the estimation task at hand. We now quickly review a three common loss functions. In Figure 6.3 we show the results of fitting a linear model with two different loss functions.

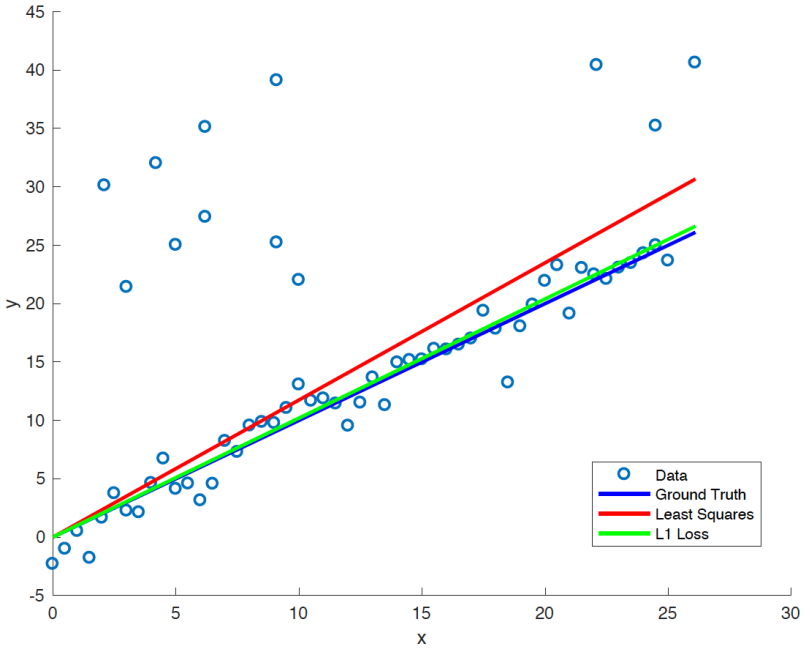


Figure 6.3: Linear Regression with different loss functions

1.3.1 Least squares loss

The most common loss function we encounter when performing linear regression is the least squares or L_2 loss. Given data with labels $(x_1, y_1), ..., (x_n, y_n)$, the least squares loss is given by:

$$L_2(\theta, Y) = \sum_{i=1}^{n} (\phi(x_i)^T \theta - y_i)^2 = ||\Phi\theta - Y||_2^2$$

Note that to simplify the expression we have used the fact that the Euclidean norm on $X \in \mathbb{R}^d$ is given by:

$$||X||_2^2 = X^T X = \sum_{i=1}^{d} x_i^2$$

The least squares loss is of particular interest because the solution to the minimization problem 6.2 with the least squares loss has a closed form solution.

We can derive this solution by either taking the gradient with respect to θ of L (since θ is a vector) and setting it equal to 0 or by a geometric argument. This closed form is called the *Normal Equation*.

Theorem 6.3. *If Φ has full row rank (i.e $rank(\Phi) = d$) The solution to:*

$$\theta^* = arg \min_{\theta \in \mathbb{R}^d} ||\Phi\theta - Y||_2^2$$

, is given by the Normal Equation:

$$\theta^* = (\Phi^T \Phi)^{-1} \Phi^T Y$$

Proof. To prove the result, we first expand out $||\Phi\theta - Y||_2^2$ and take the gradient with respect to θ.

$$\begin{aligned} L(\theta, Y) &= ||\Phi\theta - Y||_2^2 \\ &= (\Phi\theta - Y)^T (\Phi\theta - Y) \\ &= \theta^T \Phi^T \Phi\theta - 2Y^T \Phi\theta + Y^T Y \end{aligned}$$

Now using the identity that $\nabla_\theta A\theta = A^T$, and the product rule, we get that:

$$\begin{aligned} \nabla_\theta L(\theta, Y) &= \nabla_\theta(\theta^T \Phi^T \Phi\theta) - \nabla_\theta 2Y^T \Phi\theta \\ &= 2\Phi^T \Phi\theta - 2\Phi^T Y \end{aligned}$$

Setting this equal to zero, and using the fact that the inverse of $\Phi^T\Phi$ exists since it is full rank by assumption gives us the desired result:

$$\theta^* = (\Phi^T \Phi)^{-1} \Phi^T Y$$

We can also arrive at this result using a geometric argument. To do so, recall that two vectors Y_1 and Y_2 are orthogonal if $Y_1^T Y_2 = 0$.

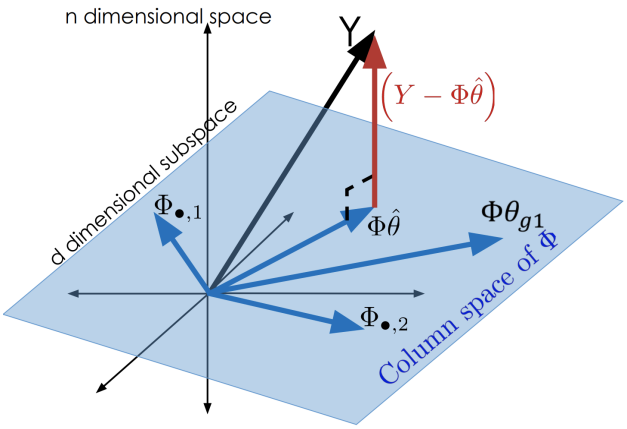


Figure 6.4: Geometric view of least squares regression

Now, consider the column space or range of Φ. If $rank(\Phi) = d$ this is a d-dimensional subspace of $\mathbb{R}^n$. If $n > d$, Y may not be in the column-space of Φ, meaning that there may not exist a θ such that $Y = \Phi\theta$. As such, we settle for a value of θ such that $\Phi\theta$ is as close as possible in the Euclidean distance to Y. This is given by the orthogonal projection of Y onto the column space of Φ, and is achieved when the residual $e = Y - \Phi\theta$ is orthogonal to the column space of Φ. This is illustrated in Figure 6.4. This means that

$$\Phi^T e = \Phi^T(Y - \Phi\theta^*) = 0$$

Solving for θ^* gives the desired result.

□

1.3.2 Regularized least squares loss

A second commonly encountered loss function is the regularized least squares loss given by:

$$L(\theta, Y) = \frac{1}{2}||\Phi\theta - Y||_2^2 + \frac{\lambda}{2}||\theta||^2$$

Where $\lambda > 0$ is the regularization parameter. This loss, as above also has a closed form given by the following theorem.

Theorem 6.4. *The solution to:*

$$\theta^* = arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}||\Phi\theta - Y||_2^2 + \frac{\lambda}{2}||\theta||^2$$

, is given by:

$$\theta^* = (\Phi^T\Phi + \lambda I_d)^{-1}\Phi^T Y$$

This loss is most useful when the matrix Φ does not have full row rank, since it does not require this assumption to have a closed form. However, it biases the solution towards values of θ with smaller norms which can degrade the performance.

1.3.3 L1 Loss

A last commonly encountered loss function is the L_1 loss given by:

$$L(\theta, Y) = \sum_{i=1}^{n} |\phi(x_i)^T\theta - y_i| = ||\Phi\theta - Y||_1$$

There is no closed form for the minimizer of the L_1 loss, so the minimization is performed with gradient descent. The resulting linear model, however, has the benefit of being more robust to outliers in the data than the least squares loss. In Figure 6.3 we can see that the linear model learned from minimizing the L_1 loss is much closer to the ground truth that learned from minimizing the L_2 loss since it is not as sensitive to the outliers in the data.

2 Links between linear least squares and conditional expectations

We now draw links between probability theory and the facts we have just highlighted about linear regression. We first assume that we have two random variables X and Y. In this section we will show that the conditional expectation:

$$\mathbb{E}[Y|X],$$

can be loosely viewed as an orthogonal projection of Y onto X. To make this claim more rigorous requires more advanced mathematics, however, we can show that the conditional expectation has several similar properties to the least squares estimator.

We begin by remarking that the conditional expectation is itself a random variable since it is a deterministic function of the random variable X. To simplify notation, let:

$$b(X) = \mathbb{E}[Y|X],$$

and define the discrepancy D as:

$$D = Y - b(X),$$

In this section we will show that $b(X)$ and D have the following properties:

Theorem 6.5. *The conditional expectation, $b(X) = \mathbb{E}[Y|X]$ satisfies:*

1. *$b(X)$ is an unbiased estimator for Y: $\mathbb{E}[D] = 0$.*
2. *$\mathbb{E}[Dg(X)] = 0$, for any bounded function g.*
3. *$b(X)$ minimizes the mean squared error: $\mathbb{E}[(g(X) - Y)^2]$ over all bounded functions g of X.*

Though the first property is straightforwards to show, the other two are very important. If we think of the discrepancy D as the residual of our estimator $b(X)$, and view the expectation $\mathbb{E}[Dg(X)]$ as a dot product, the second property above means that our residual is orthogonal to all (bounded) functions of X. This is exactly analogous to property that the least squares estimator is orthogonal to the column space of the data. The conditional expectation can therefore be viewed as the closest estimate (in the least squares sense) to Y given the information in X.

Proof. We prove each of the properties of the conditional expectation separately:

1. We first show that $b(X)$ is an unbiased estimator for Y by using the tower property of the conditional expectation.

$$\begin{aligned}\mathbb{E}[D] = \mathbb{E}[Y - \mathbb{E}[Y|X]] &= \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y|X]] \\ &= \mathbb{E}[Y] - \mathbb{E}[Y] = 0\end{aligned}$$

2. To show that $\mathbb{E}[Dg(X)] = 0$, for any bounded function g, we expand out D and condition on X:

$$\begin{aligned}\mathbb{E}[Dg(X)] = \mathbb{E}[g(X)Y - g(X)\mathbb{E}[Y|X]] &= \mathbb{E}[g(X)Y] - \mathbb{E}[g(X)\mathbb{E}[Y|X]] \\ &= \mathbb{E}[\mathbb{E}[g(X)Y|X]] - \mathbb{E}[g(X)\mathbb{E}[Y|X]]\end{aligned}$$

   Since g is bounded and $g(X)$ given X is a constant, we can rearrange the above expression and find that:

$$\begin{aligned}\mathbb{E}[Dg(X)] &= \mathbb{E}[\mathbb{E}[g(X)Y|X]] - \mathbb{E}[g(X)\mathbb{E}[Y|X]] \\ &= \mathbb{E}[g(X)\mathbb{E}[Y|X]] - \mathbb{E}[g(X)\mathbb{E}[Y|X]] = 0\end{aligned}$$

3. Finally to show that $b(X)$ minimizes the mean squared error: $\mathbb{E}[(g(X)-Y)^2]$ over all bounded functions g of X, we first expand the mean square error given an arbitrary function $g(X)$:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[(g(X) - Y)^2]] &= \mathbb{E}[(g(X) - b(X) + b(X) - Y)^2] \\ &= \mathbb{E}[(g(X) - b(X))^2] + \mathbb{E}[(b(X) - Y)^2] - 2\mathbb{E}[(b(X) - Y)(g(X) - b(X)] \\ &= \mathbb{E}[(g(X) - b(X))^2] + \mathbb{E}[(b(X) - Y)^2] - 2\mathbb{E}[D(g(X) - b(X)]\end{aligned}$$

   Now using the property from the previous part, we know that the last term above is always 0 since it is $\mathbb{E}[Dh(X)]$ for some bounded function h. Since $\mathbb{E}[(g(X) - b(X))^2] > 0$ always, we must have that:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[(g(X) - Y)^2]] &= \mathbb{E}[(g(X) - b(X))^2] + \mathbb{E}[(b(X) - Y)^2] \\ &\geq \mathbb{E}[(b(X) - Y)^2]]\end{aligned}$$

   Therefore the conditional expectation achieves the lowest possible mean squared error across all bounded functions of X.

□