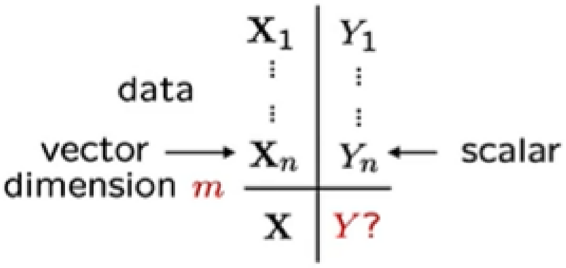
Linear Regression

Let's start with a dataset,



This is our dataset and it takes the form of a spreadsheet. So our dataset consists of several records and these records are in this form, the typical record is a vector of X's together with a corresponding label Y. Each one of those X's has a certain dimension, if we need to refer to it, that dimension is m and we have n such data records. Then a new person is going to come in, they will show up and they will have some characteristics or attributes associated with a data record X and we want to make a guess of what Y is going to be.

$$\textbf{Regressor/predictor: } \hat{Y} = g(\mathbf{X})$$

So we're trying to build the function g that takes an input X and outputs a prediction Y. That's the business of regression. We're trying to construct a good predictor based on the dataset that we're working with. We want to learn a such a good g from the training data that we have.

How do we go about it ?

It's useful to have a criteria about what would be a good g. Think of having a population and we would like our predictor to be good for individuals in that population. If you take a random individual in that population, our predictor is going to make a prediction.

$$\begin{array}{c}\text{objective:}\\ \text{(risk)}\end{array} \quad \mathbb{E}\left[\left(g(\mathbf{X}) - Y\right)^2\right]$$

It's going to be compared to the Y value of that person and we're interested in the error that we're making. Now, in order to have positive errors, we take the square of that. So this quantity is the squared error on a particular individual. We want to keep the error small as possible so that our predictions are good. But we want our predictor to work well on the average throughout the population and that's why we take the expected value of that quantity. So that expected value corresponds to a population average. We're interested in a predictor which on the average over the population of interest, results in prediction errors that are small in the mean squared sense.

Why do we take the square?

We could have used different criteria, for example, we would have used the absolute value of the error or something else, which would be a sensible choice as well. However, this squared error criterion is the most popular one and that's for many different reasons. Some of the reasons are,

- It's historical
- Convenience mathematically - because of the mean squared error you get very nice analytical solutions that are computationally retractable.
- It works well in practice and it reflects a desire that says the following "If errors are very big, we want to really penalize them, but when errors are small, then the quadratic doesn't really give a big penalty". So we care more about the big errors compared to the small errors.

How to find the predictor that keeps the objective as small as possible ?

We cannot calculate the expectation as we do not have access to it. If we had the full probabilistic model of how X's is related to Y's, then probability textbooks give us the formula. The optimal predictor is just the conditional expectation of Y when you're given the value of X. This is an abstract formula. In reality, it means that if you have access to the conditional probability of Y's, given the X's then you could use those conditional probabilities and come up with the prediction but we have not assumed that we have a probabilistic model. So we just need to work with data. If we only work with data there is no way that we can play with this expectation here but we can play with a surrogate or the proxy of that expectation. Intuitively, an expected value of a random variable is essentially an average of that random variable if we repeat an experiment multiple times. So this expectation can be approximated by an average of this kind.

What is this average?

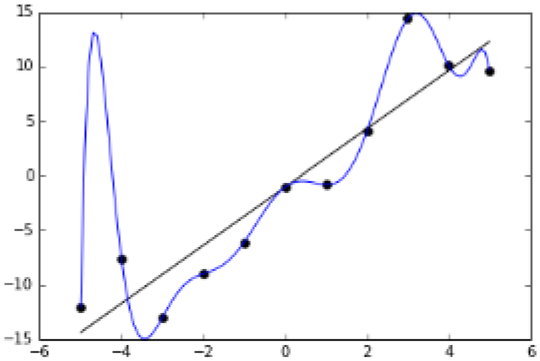
We draw at random an individual from our population, that individual will have a certain vector X and a certain Y and $(g(X_i) - Y_i)^2$ is the error of our predictor. We draw at random n individuals and so this is the mean squared prediction error. The average prediction error on the n individuals that we have drawn.

$$\text{proxy:} \quad \frac{1}{n}\sum_{i=1}^{n}\left(g(\mathbf{X}_i) - Y_i\right)^2$$

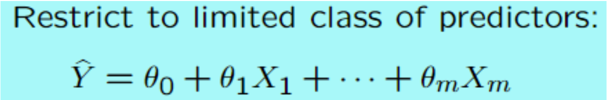
So the above quantity is the average squared error on the sample on the data set that we have, whereas objective(risk) is the one for the entire population. If we have a big enough data set, the mean squared error on a sample should be representative of the mean squared error on the overall population. So we wish to optimize this Expectation (Risk), since it's not accessible, we'll use the proxy and try to minimize over all possible predictors g. It's called empirical risk minimization. Instead of optimizing the true risk we optimize this approximation of the risk, which is the empirical risk.

What's the best g?

Let's pretend we've given you some data points and asked you to find the g function that produces very small errors. I can create a curve that runs through all of the points, like this blue curve.



The above curve perfectly predicts all of the Y’s for the X's in my dataset. It has zero error on my data set. We cannot really trust the predictions and they look kind of fake. We're dealing with **overfitting** in this scenario. Here we have too many free parameters when we build our predictor and using that freedom in our predictor, we fit all the data in our dataset exactly but this creates a weird artifact which means that we should not really trust our predictor in more general ways. So, overfitting is something that is not desirable. We want to avoid it and the way to avoid it is instead of allowing an arbitrary prediction curve, like this blue curve that we have here, we want to restrict it to a limited class of predictors. And in linear regression what we do is restrict it to the class of linear predictors.



Restrict to limited class of predictors:

$$\widehat{Y} = \theta_0 + \theta_1 X_1 + \cdots + \theta_m X_m$$

So, for a typical individual who comes in and has some characteristics that are these Xs, we're going to make a prediction for that individual by taking a linear combination of those components X together with a constant term. That's what linear regression is. Instead of looking at arbitrary ways of making predictions, we just use predictors that have a special structure.

For some concise notation, it helps to denote a vector X as,

$$\mathbf{X} = (1, X_1, \ldots, X_m)$$

Where $X_{1,}$..... X_m are different attributes of an individual and the constant at the beginning. Now we can write our predictor in the form,

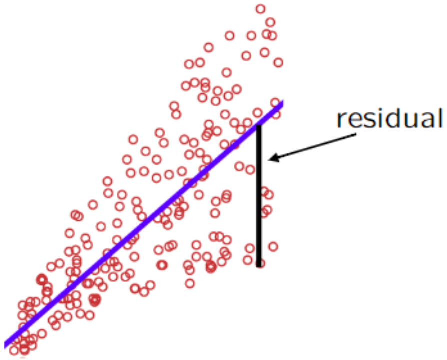
$$\hat{Y} = g(\mathbf{X}) = \theta^T \mathbf{X}$$

We have a vector theta, which is a vector of theta that would be the vector θ_0, θ_1, θ_2.... θ_m. We take the inner product of these vectors together with the vector X.This is a compact notation for the linear predictors. So, we want to build linear predictors of this particular type. We are going to minimize this squared prediction error on our population. Since we are using linear predictors, the formula for the predictor will look in the form,

$$\sum_{i=1}^{n} (\boldsymbol{\theta}^T \mathbf{X}_i - Y_i)^2$$

Any particular choice of theta corresponds to a different predictor and we want to choose theta so that this quantity is as small as possible.

Pictorially, this corresponds to the following,



For each individual in our data set, for each individual who is building a line. The line is given by the formula $\theta^T X$ and Theta determines the slope of that line. For a particular individual who has a certain X and a particular Y, our predictor predicts this value of Y. The difference between true value and predicted value of Y is called the prediction error or another name that goes for it is the residual. So, the vertical distance of a typical point in our dataset from the prediction line is the residual. Mathematically, it corresponds to this term shown below,

$$residual\ =\ (\theta^T X\ -\ Y)\ =\ predicted\ value\ \ -\ \ observed\ value$$

We take the square of that residual and then we sum those squares over all the data points in our data set. Linear regression tries to find a line for which the sum of the squared residuals is as small as possible.

So, to summarize, the regression problem consists of finding a vector of theta for which the sum of the squared residuals is as small as possible. It corresponds to this mathematical optimization problem provided,

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\boldsymbol{\theta}^T \mathbf{X}_i - Y_i)^2 \qquad \begin{array}{l} n \text{ data points} \\ \mathbf{X}_i \text{ and } \boldsymbol{\theta} \text{ have dimension } m+1 \end{array}$$

It turns out that this optimization problem has an easy solution. It has a closed-form solution.

$$\hat{\boldsymbol{\theta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$$

We can get the formula from linear algebra. Here the X symbol in this formula denotes the following matrix,

$$\begin{array}{cc} n & \begin{bmatrix} 1 & \mathbf{X}_1^T \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & \mathbf{X}_n^T \end{bmatrix} \\ & m+1 \end{array} = \mathbb{X} \qquad \begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \mathbf{Y}$$

The big matrix really corresponds to the entry in our spreadsheet. The first row corresponds to the first record in our dataset and the second row corresponds to the second record in our data set and so on. These are all the records, we stack them together and form a big matrix. In the formula, we take the transpose of that matrix and multiplied with itself, then we invert it, then we do a little extra work, and that gives us a formula for the optimal choice of the parameter vector theta.

How is this possible?

The following observation is what makes this formula possible,

The objective that we are optimizing here lets call it as $H(\theta)$, it's some function of theta, and it has the property that theta shows up quadratically in this formula.

$\min_{\theta} H(\theta)$ quadratic in θ

optimality conditions: $\nabla H(\theta) = 0$ $\quad \frac{\partial H}{\partial \theta_j} = 0, \quad j = 0, 1, \ldots, m$

linear system of $m + 1$ equations

When you minimize, you take the derivatives of that expression with respect to theta and set it to zero. That's the way to optimize the function. The derivative of something quadratic is linear. The function we are dealing with is quadratic and when you take the derivative, the equations that you get are linear. So, by setting those derivatives to zero, we end up with a linear system of equations on theta's, and linear systems are easy and fast to solve and they also admit formulas. So, this is where the formula is derived from, and there is very fast and efficient software that solves these linear systems and so these days, regression problems are essentially solved in an instant, even if data is of huge dimension, thousands, even millions. One can solve linear regression problems very fast and that's one of the reasons for the popularity of this method.

Let's implement it in our particular example,

In the marketing example, we have 200 data records, so n is 200. We have three advertising (TV/Radio/Newspaper) channels, so m is equal to three. We have three X variables for each market, and there is one additional coefficient, the theta zero which is constant in our linear predictor. So, there are a total of four theta parameters that have to be chosen. We run the software and got the following results,

$$\hat{\theta} = \begin{bmatrix} 2.94 \\ 0.046 \\ 0.19 \\ -0.001 \end{bmatrix}$$

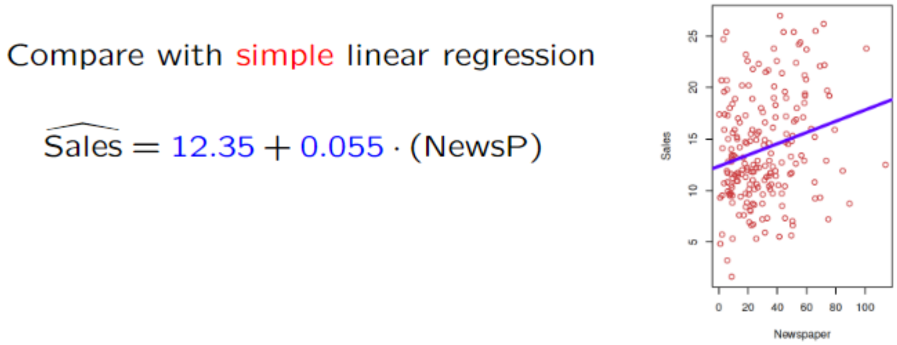
What does the above vector mean?

It means that we have a predictor of the following kind,

$$\widehat{\text{Sales}} = 2.94 + 0.046 \cdot (\text{TV}) + 0.19 \cdot (\text{Radio}) - 0.001 \cdot (\text{NewsP})$$

Our prediction of sales is a certain constant plus TV advertising multiplied by a certain coefficient, and so on for the other channels as well. So the linear regression software produces those coefficients that multiply the advertising expenditures over the different channels. Notice that we got a minus sign here. So, our predictor almost seems to say that the more you spend on newspaper advertising, the lower your sales will be. We will have to explore it in a little more depth to understand it. For now, let's just take it that the software has given us the particular predictor. It remains to assess the quality of that predictor and to say things about how much we can trust it. Here, we built a predictor using all available variables, both TV, radio, and newspaper.

We could also look just at one of the variables. For example, let's look at the relationship between newspapers and sales.



Compare with simple linear regression

$$\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP})$$

What does this correspond to?

Our dataset is a big spreadsheet that has X1, X2, and X3 expenditures in the different channels, and we try to predict Y using all three of those variables. We might throw away two of the columns of the spreadsheet and then do a prediction using just this part of the data. So we are throwing TV, radio and newspaper spending to try to predict the sales. Whenever we are using just one variable in our prediction it's called simple linear regression. We run our simple linear regression and we get a predictor similar to the above example. When you look at these results, these coefficients in front of the newspaper sales actually look a little more substantial. So, when we do linear regression, it appears that newspaper advertising has a strong relationship with sales. But when we look at everything together, it appears that the newspaper doesn't. What does that mean? Which one is the truth? Is newspaper related to sales or not? So in order to address questions like that, we need to do a little more conceptual work.