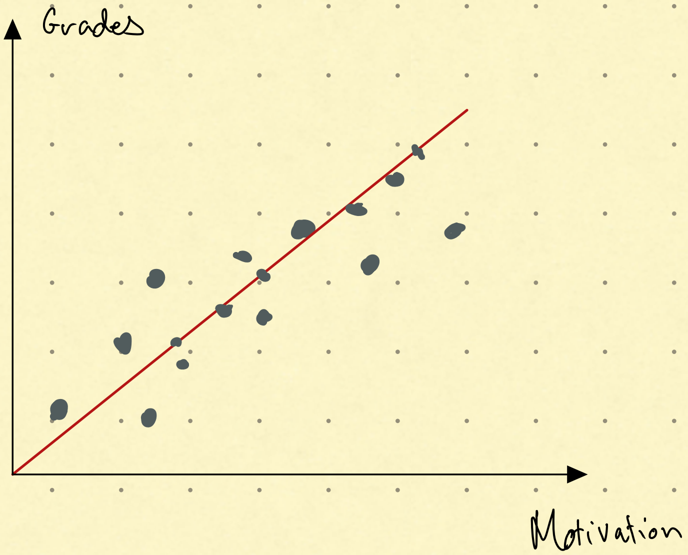
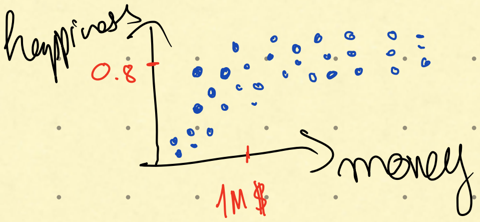
How can we make Regressions Better ?

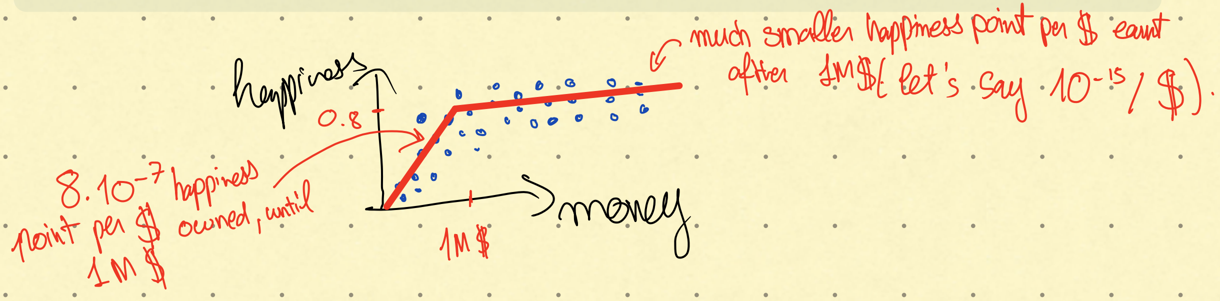


Linear regression is a great way to link together two mathematical, empirically observed, quantities, and understand how they interact with each other. Everybody knows the classical formula to infere this connection: y = a*x + b.

Now, isn't this formula a bit too basic? What if we want to reveal more complex, intricate, interactions between two quantities? Let's say, what if we wanted to infer the relationship between money and happiness (on a [0,1] scale) on individuals, from the following graph?



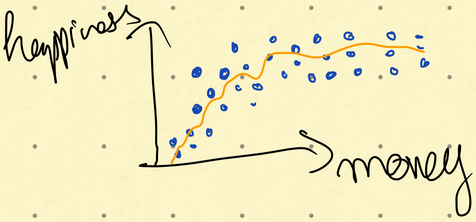
At a first glance, it doesn't look to be very linear! Intuitively, we see that there are two effects: one for smaller x, and one for greater x. We'd like, if we could, infer something like this:



8.10^{-7} happiness point per $ owned, until 1M $

much smaller happiness point per $ earnt after 1M$ (let's say 10^{-15} / $).

We cannot do it with a simple linear regression. How about Machine learning (which is nothing but a ‘big’ regression)? We can maybe obtain the following inference curve between the two quantities:



Indeed, but now we have lost the simplicity of the previous model, and the fact that we could actually interpret the coefficients in the frame of linear regressions: 8e-7 happiness/$ for regular people, e-15 happiness/$ for rich people, which are all the advantages of linear regressions. Now, can we still find a way to resort to the latter? I will try to explain several enhancements we can make to regular regressions, in order to take account of broader, more complex, phenomena.

.

1- Use more regressors.

What if, instead of just using x to predicting y, we also used a lot of other parameters that also have an impact on y?

In the previous example, what if we used not only x, but also x^2 to take into account the dynamic of money and happiness for very rich people?
Now, this is called multivariate regression, and instead of looking for scalar a and b, we look for the following beta for resolving this problem:

$$X = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}, \quad y \in \mathbb{R}$$

We try to find the best $\beta \in \mathbb{R}^{d+1}$ and $\alpha \in \mathbb{R}$ such that $\hat{y} = X^t \beta + \alpha$ is a good estimator of y.

(We usually resolve this by OLS).

We call X the vector of regressors, and each xi (for i in [1,d]) is called a regressor; it also has a constant regressor: **1** that holds for constant effects (same role as ‘a’ previously).

We can think of any regressors that we want for explaining y: in the case of happiness, we can for example **use the number of relatives and friends, the number of sunny days each individual from our set experience each year, age,**… Any parameter that we can think of. Still, one warning: you cannot have a linear dependence in your collection of regressors; e.g. you cannot have a constant regressor, a ‘Woman’ regressor (equal to 1 if the individual is a woman) and a ‘Man’ regressor (equal to 1 if the individual is a man), because
Man = Woman - 1 in this case and you end up with a spurious regression. Usually , we want to have regressors weakly correlated to each other and highly correlated to y. Also, and it is a general rule for regressions, all your observations must be i.i.d.

Then, once we’ve made our regression, we can directly interpret the coefficients of beta.

If y = grades,
↳(on a $[0,100]$ scale)

$$X = \begin{pmatrix} 1 \\ \text{\# hours of study / day} \\ \text{time of sleep / night} \\ \text{\# of Uber eats / week} \end{pmatrix}$$

and we find
$$\beta = \begin{pmatrix} 20 \\ 5 \\ 6 \\ 3e^{-40} \end{pmatrix}.$$

It means that → 1 extra hour of study / day will get you + 5 pt on average on your grades.

→ 1 extra hour of sleep/night will get you + 6 pt on average.

→ the number of Ubereats order you make each week has no significant impact.

2 - Use control variables

Now, we can even refine the prediction of y by X, by using control variables.

For example if we want to regress the salary of employees on their experience (number of years worked), the result differ given the social group of these workers: the ramp of salary would be much greater for white collar salarymen than for construction worker. This is why we introduce a new vector, G, called the vector of control variables.

Now, without going too much into the details, instead of predicting the random variable y using X, we want to predict y|G (y “given” G, same meaning as in Probability) given X|G.

If we come back to the previous example, if we find a control variable G that accounts for the social background of employees (e.g. their parents’ salary), we can then predict the ‘real’ impact of experience on wages. It’s as if you only compare each individual with only individuals of the same social group to infer the impact of experience on their wage.

3 - Afterwords / to go beyond...

The two above techniques are the major changes we make to linear regressions when we want to extend to the multivariate case. I will not go into further enhancements we can do to improve linear regression and generalize it to broader cases (e.g. when we take a weaker hypothesis than the independence of observations), as these need a great deal of heavy mathematical definitions to be explained (and to explain why these are useful).

I also wanted to talk a bit about the case where y is not continuous but binary: in this case, we can also make use of regressions. Now, we will use Logistic Regression, and for a given observation X we will try to predict the probability that it corresponds to y=1.

We can even extend beyond this case, when y is discrete but not binary: it can take more than 2 values. Then, we perform what we all 'Polytomic regression', to try and predict the probability of each individual Xi to correspond to each outcome of y (y=0,1,⋯, or n).

Finally, we can also extend regression to cases where part of the data are censored or pre-selected (for example, if you try to predict the life expectancy, you only know the total life span of dead people in your dataset which can add a bias to the results by underestimating y).

This is it for this (small) bestiary of regressions. It is hard to be exhaustive about these numerous techniques, and I wanted to keep the things simple and light to read, but I hope that I was still able to give you a glimpse of what wonders we can do with this tool, and maybe make you as passionate about these as I am!

Waël Boukhobza

Source : Introductory Econometrics:
A Modern approach
by Jeffrey Wooldridge