Monte Carlo Regression

This brief paper introduces a new model for running regressions, which will be initially called Monte Carlo Regression (MCR). It is meant to be an alternative to simple linear regressions, only more powerful in the sense that:

- it can find relationships with much higher correlation between one dependent variable and a given set of independent variables;
- with no prior knowledge from the researcher about the relationship between the variables, it can find relationships of any exponenciation; and
- it can even find that some of the independent variables work better together at explaining the dependent variable, either magnifying or compensating for each other, as we will see in the use case presented.

We start with a general overview of what the model is, what it tries to accomplish, and how. In the second section, we will go through one instance of how *f* could be determined, a fundamental aspect of the model.

In the third section, we will have a brief look at one possible use case as an illustration. We will also use it to go into deeper details of the model. In the fourth section, we will learn, using the use case previously presented, how one can analyse the results using differentials. And in the final section, concluding remarks will be made.

The model has been coded in R and is available as open-source at the GitHub repository below. The repository also includes the data used and the results obtained for the use case described in this paper, as well as some brief documentation on how to run the model in R.

https://github.com/f-lungov/MCR

General Overview

We will start with the equation used for simple linear regressions:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

So we will attempt to explain *Y* through *X*. But in our case, *X* will not be just one variable. It will contain the values of (sometimes several) different variables that could explain *Y*. We will call them *Z* variables, and their relationship with *X* will be given by:

$$X_i = f\ (Z_{1,\ i}\ ,\ Z_{2,\ i}\ ,\ Z_{3,\ i}\ ,\ \ldots\ Z_{n\ ,\ i})$$

The novelty of this new model is in using Monte Carlo to determine an *f* that will allow *Z*’s -- passing their values through *X* -- a high explanatory power of *Y*. We will do that by randomizing a set of components of an equation (*f*) with *X* on one side, and the *Z*’s on the other.

MCR will determine that set of components multiple times, and in each time we will have a different *f*, and therefore a different *X*. Then we will run the simple linear regression of *X* on *Y* and assess the explanatory power with R^2.

In other words, MCR will use Monte Carlo to try to find an arrangement of the *Z* variables in such a way that *X* will have a high R^2 in explaining *Y*. Hence for this model, the above equations become:

$$Y_i = \beta_0 + \beta_1 \cdot X_{j\,,\,i} + \varepsilon_i \quad \text{, where:}$$

$$X_{j\,,\,i} = f_j\left(Z_{1,\,i}\ ,\ Z_{2,\,i}\ ,\ Z_{3,\,i}\ ,\ \ldots\ Z_{n\,,\,i}\right)$$

That randomization will occur within preset and customizable parameters. One possibility of which components could be determined by MCR, and within what ranges, will be presented below. This aspect of the model is highly customizable, both per user and per use case, and should certainly be subject of further discussions.

Constructing *f*

We will go through one intance of how MCR can randomly determine *f*. In other words, let us build f_j where j = 1.

The first thing we will have MCR determine is the number of blocks on the *Z* side of the equation. So for this demonstration we set this number to vary between 1 and 4. For j = 1, say that the number randomly picked is 3. This means that we will have 3 blocks, or that:

$$X_{1,\,i} = block_1 + block_2 + block_3$$

Now we need to determine the values of the multipliers of each block. For simplicity, we will have MCR determine the multipliers between -10 and +10, whole numbers only. For j = 1, the numbers +4, -7 and +2 have been randomly picked for blocks 1, 2 and 3 respectively. The equation then becomes:

$$X_{1,\,i} = 4\ \ block_1 - 7\ \ block_2 + 2\ \ block_3$$

Next, we need MCR to determine the size of each block. Let us suppose we want that size to be between 1 and 5 variables. MCR randomly picks numbers 4, 1 and 2, giving us:

$$X_{1,\,i} = 4 \cdot Z_{w\,,\,i} \cdot Z_{w\,,\,i} \cdot Z_{w\,,\,i} \cdot Z_{w\,,\,i} - 7 \cdot Z_{w\,,\,i} + 2 \cdot Z_{w\,,\,i} \cdot Z_{w\,,\,i}$$

where *w* is a number between 1 and *n*, and *n* is the number of *Z* variables we have available.

We now need MCR to determine the power each variable will be taken to. Suppose, for simplicity, that we want each power to be between -5 and +5. So MCR randomly picks +2, -1, +4, +3, -5, +1, -2. That gives us:

$$X_{1,\,i} = 4 \cdot Z_{w\,,\,i}^{2} \cdot Z_{w\,,\,i}^{(-1)} \cdot Z_{w\,,\,i}^{4} \cdot Z_{w\,,\,i}^{3} - 7 \cdot Z_{w\,,\,i}^{(-5)} + 2 \cdot Z_{w\,,\,i}^{1} \cdot Z_{w\,,\,i}^{(-2)}$$

Finally, we need to know which variables are going to be used for this particular instance of *f*. Suppose we have 8 variables to draw from (i.e. *n* = 8). Each variable may only appear once in each block. Say that MCR randomly picked these variables: 3, 5, 8, 2, 5, 8, 1. That gives us X_1:

$$X_{1,\ i} = 4 \cdot Z_{3,\ i}^{2} \cdot Z_{5,\ i}^{(-1)} \cdot Z_{8,\ i}^{4} \cdot Z_{2,\ i}^{3} - 7 \cdot Z_{5,\ i}^{(-5)} + 2 \cdot Z_{8,\ i}^{1} \cdot Z_{1,\ i}^{(-2)}$$

This equation will give us an *X* value for each observation we have in our sample. MCR will then run a simple linear regression of that particular *X* against *Y*, and assess its explanatory power through R^2. Of course chances are that R^2 will be low, but this will be done many millions of times, and we are hoping that some of those R^2's will be much higher.

Moreover, the data scientist could have a closer look at the *X*'s that provided high R^2's and find ways to improve them further, such as adding more variables whose importance was only discovered after analysing the outliers from those regressions.

Use Case: Understanding GDP Per Capita

We will now go over a use case. All the data used, along with the results found, are in the GitHub repository mentioned above.

Y will be the GDP per capita of different countries in 2014 at constant 2011 US dollars. We also have a list of variables that are candidates to help explaining *Y*. We are going to call them *Z* variables:

- Z_1: Gross intake ratio in first grade of primary education as percentage of total
- Z_2: Years of compulsory education
- Z_3: Gross domestic savings as percentage of GDP
- Z_4: Imports of goods and services as percentage of GDP
- Z_5: Exports of goods and services as percentage of GDP
- Z_6: Trade as percentage of GDP
- Z_7: Gross capital formation as percentage of GDP
- Z_8: Value added of services as percentage of GDP
- Z_9: Value added of manufacturing as percentage of GDP
- Z_{10}: Value added of industry as percentage of GDP
- Z_{11}: Urban population as percentage of total
- Z_{12}: Arable land as percentage of land area
- Z_{13}: Real interest rate
- Z_{14}: Account at a financial institution as percentage of population age 15+
- Z_{15}: Infant mortality rate per 1000 live births
- Z_{16}: Years of life expectancy at birth

Although we are working with real-world variables available at the World Bank's database, it is important to note that no attempt was made to run the model as to actually understand the proposed question. Variables were chosen primarily on their availability to a large number of countries. Furthermore, the model was run a very modest number of times -- twenty minutes in a Personnal Computer. The aim of this demonstration is simply to illustrate how the model works.

Now we need to go through one importat transformation to the *Z* variables. We need to fit them in a 1-99 range. That is because MCR may choose some of them to divide other variables, and of course dividing by zero will cause an error. So instead of using Z_i, we will use zz_i:

$$zz_{w,i} = \frac{Z_{w,i} - \text{MIN}(Z_w)}{\text{MAX}(Z_w) - \text{MIN}(Z_w)} \times 99 + 1$$

Of course, the numbers 99 and 1 could be almost any two numbers. And once we find the relationship between *zz*'s and *Y*, it is easy to find the relationship between *Z*'s and *Y*.

During the twenty minutes that the model has been run, MCR produced 571,257 *f* functions using the *Z* variables above, and then the same number of *X* variables. Regression number 41,544 found the highest R^2:

$$X_{41544,\ i} = -66.96 \times zz_{14,\ i}^{0.99} \times zz_{16,\ i}^{2.04} \quad // \ R^2 = 0.7677$$

So when MCR ran the f-constructing routine for the 41544th time, it randomly picked:

- 1 as the number of blocks;
- -66.96 as the multiplier for that block;
- 2 as the number of variables for that block;
- 0.99 and 2.04 as the power the variables should be taken to; and
- zz_{14} and zz_{16} as the variables used.

$X_{41544,i}$ has been reconstructed in Excel (see proof.xlsx), giving the same R^2 as calculated in R.

Very different *f*'s have also provided high R^2's. For instance, $X_{238816,i}$ was constructed using five variables in three blocks (R^2 = 0.7582). And $X_{306908,i}$ was constructed using six variables in two blocks (R^2 = 0.7574). This could suggest that we still have a long way to finding the highest possible R^2, as presumably the *X* equations with the top few highest R^2's should be very similar.

Interestingly, two variables seem to work well together. Z_{14} appears 14 times in the top 10 equations, 9 of which it is in the same block as Z_{16}. And Z_{16} appears 11 times, 9 of which together with Z_{14} in the same block. When looking at what these variables are, Z_{14} is percentage of adults with a bank account, and Z_{16} is life expectancy.

This finding suggests that a high percentage of adults holding bank accounts is even more important for raising GDP in countries where people are able to get older. And a country can only reap the full benefits of high life expectancy if a high percentage of its people have bank accounts. If either of these variables is low for a country, their product will be low, X_i will be low, and Y_i will be low. But if both of them are high, then Y_i will be high.

Of course more study must be done before we can make any real-world conclusion, but this shows that the model allows the researcher to find relationships that would not very easily be known in advance.

Another example where this might be useful is in looking for causes of a disease. Say that we are trying to find the cause for a specific type of cancer. Than according to our sample, only 0,1% of people who reported that they have been taking a specific pain killer developed that type of cancer, and that number is not significantly different from the sample as a whole.

But nearly 100% of the people who both reported taking that pain killer and having a specific species of fish in their regular diet developed the disesase. This would likely form a block with the two variables, and would then help researchers in their work.

Analizing the results with differentials

After running a linear regression, understanding the importance of each independent variable in explaining the dependent variable is very straightforward. For MCR, depending on how complex the chosen *X* equation is, it may be useful to make use of differentials. In this section, we will quickly go over how each *Z* variable impacts US GDP using the results from the use case shown above.

We will use X_{238816} since it employs a higher number of *Z* variables than X_{41544} and it will therefore be a more interesting case to analyse.

$$X_{238816,i}=-\left(96.27\times zz_{14,i}^{6.96}\times zz_{8,i}^{0.79}\right)-\left(60.61\times zz_{7,i}^{0.85}\times zz_{11,i}^{6.86}\right)+\left(61.55\times zz_{5,i}^{(-1.08)}\times zz_{11,i}^{(-2.64)}\times zz_{14,i}^{(-6.71)}\right)$$

It may look like it is hard to draw any meaningful conclusion out of this equation (not to mention that in real use cases, that equation could still be a lot more complex) as to what variables are most relevant to explaining *Y*, and how their relationships are like. But if we use differentials, we will end up with a very simple and informative linear equation.

The analysis based on differentials will be different for every observation in our sample. For this illustration, we will analyse the American economy.

We will shift each *Z* variable, one at a time, up and down one-tenth of a standard deviation. Then we will use the same *f* function to calculate new *X*'s for the observation we are analysing. So when $w = 5$:

$$X_{238816,US,Z5+}=f_{238816}\left(zz_{US,5+},zz_{US,7},zz_{US,8},zz_{US,11},zz_{US,14}\right) \quad \text{, where:}$$

$$zz_{US,5+}=\frac{Z_{US,5}+SD(Z_5)/10-\mathrm{MIN}(Z_5)}{\mathrm{MAX}(Z_5)-\mathrm{MIN}(Z_5)}\times 99+1$$

Note that in the first equation $zz_{US,5+}$ is in the place of $zz_{US,5}$. The same will be done subtracting 1/10 of a standard deviation from $Z_{US,5}$. The difference between *X* calculated with $\mathrm{zz}_{\mathrm{US},5+}$ and *X* calculated with $\mathrm{zz}_{\mathrm{US},5-}$, we divide it by 0.2 (because we shifted 0.1 standard deviations to each side) and call it $\Delta X_{US}/(\Delta Z_{US,5}/SD(Z_5))$.

We do the same for variables Z_7, Z_8, Z_{11} and Z_{14}, both up and down 1/10 of a standard deviation. So more generally:

$$\Delta X_i/(\Delta Z_{i,w}/\mathrm{SD}(\mathrm{Z}_w))=(X_{i,w+}-X_{i,w-})/0.2$$

How much a change in *X* will cause of a change in *Y* is given by β_1 from the MCR equation. In our case, $\Delta Y_i/\Delta X_i$ is -3,1455E-18. Finally, in order to understand how much each of the *Z* variables affect *Y*, we need to multiply:

$$\Delta Y_i/(\Delta Z_{i,w}/\mathrm{SD}(\mathrm{Z}_w))=(\Delta Y_i/\Delta X_i)\,x\,(\Delta X_i/(\Delta Z_{i,w}/\mathrm{SD}(\mathrm{Z}_w))) \quad \text{, or:}$$

$$\Delta Y_i/(\Delta Z_{i,w}/\mathrm{SD}(\mathrm{Z}_w))=(\Delta Y_i/\Delta X_i)\,x\,((X_{i,w+}-X_{i,w-})/0.2)$$

So, for the United States, as per the calculations presented in Excel:

$$\Delta Y/\Delta Z\approx 1.4915\frac{\Delta Z_{14}}{\mathrm{SD}(Z_{14})}+0.1347\frac{\Delta Z_8}{\mathrm{SD}(Z_8)}+0.0951\frac{\Delta Z_{11}}{\mathrm{SD}(Z_{11})}+0.0311\frac{\Delta Z_7}{\mathrm{SD}(Z_7)}+0\frac{\Delta Z_5}{\mathrm{SD}(Z_5)}$$

Z_{14} is by far the *Z* variable that causes the greatest impact in *Y*, while a change in Z_5 should have no impact at all. Had this been a complete study, we would have found that the "easiest" way for the United States to raise its GDP even higher would be by somehow encouraging more people to open bank accounts. These numbers will certainly be different for other countries, and the relative importance of each variable could be different too.

This analysis should help one understand the results given by MCR. It is not yet present in the coding found in the repository, as it requires further discussion concerning its applicability.

Final Remarks

This model is an alternative to linear regressions. The main benefits are higher explanatory power, the possibility of using a much larger set of independent variables, and to find the relationship among them. The main drawback is the need for greater processing power and the results are initially presented in a more complex form.

The model is currently in its first draft version. A lot more needs to be discussed and further developed. Should the differencial analysis be in the model? Or should derivatives be used instead? Should standard deviation be the unit used to imply the “stickiness” of each variable? Are there other parameters that can be customized when determining each f?

How could the researcher use the model to further improve the explanatory power? Should the model after, say, 10,000 attempts average the top 10 results into a new X and assess its explanatory power? Or should it try that with every combination of two equations from the top 10? Os should it take one of the top f functions and randomly add/remove a variable from each? Or add/remove 1 from the power of a randomly picked variable? How often should it try these incremental changes to the best equations at the time?

There are a lot of directions this model could expand to. Input from scholars would certainly help make it more useful to the scientific community, and then to all the people who rely on its findings.

So here is the model presented, open to any changes.