Monte Carlo Regression

This 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 current methods in understanding what causes the values of a particular variable *Y*.

A data scientist might be trying to understand what are the causes of a certain illness, for instance. Given a sample of observations, he or she could test how much a variety of conditions contribute to the probability of any person having that illness, conditions such as genetics, diet, habits, etc. In trying to find such a relationship, the scientist may want to employ MCR as an alternative to standard linear or non-linear regressions.

MCR may find that a certain set of variables has much higher explanatory power than would be found by other methods because it will try to combine the variables in functions in nearly every possible way and test one by one. There is also no constrain to adding more variables to be tested as possible explanatory variables, which means that statistically significant relationships that would be unexpected from a theoretic point of view are more likely to be found.

We start with a general overview of what the model is, what it tries to accomplish, and how. We also revisit two key concepts of the model: simple linear regressions and Monte Carlo. In the second section, we will go through what we refer to as an iteration, a fundamental aspect of the model. In the third section, we will have a brief look at one possible use case as an illustration. 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

The Monte Carlo Regression model is built upon both simple linear regressions and Monte Carlo. So let us begin by quickly revisiting the aspects of simple linear regressions that will be most important in MCR.

Simple Linear Regressions

A simple linear regression is a linear approach to modelling the relationship between a dependent variable and one explanatory or independent variable. Say we have a sample of observations and two variables for each observation in the data: *X* and *Y*. The relationship between these two variables may be modeled as:

$$\text{(Eq. 1.A)} \quad Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$$

, or rearranging it:

$$\text{(Eq. 1.B)} \quad \varepsilon_i = Y_i - \alpha + \beta \cdot X_i$$

This means that if we assign values for α and β, we will have a value for ε corresponding to each observation in our sample. Notice how ε is whatever of Y is left unexplained by X. If ε_i were equal to 0 for every i, this would mean that there is a value for α and a value for β that would allow us to use Eq. 1.A to accurately calculate ("explain") every Y_i from its corresponding X_i. In other words, every time we knew X_i under such conditions, we would also know Y_i.

Of course, depending on the values for X and Y, that will not always be possible. But it is mathematically possible, given X and Y, to find the values for α and β that will cause the least total ε_i^2. It is beyond the scope of this paper to explain how this can be done, but understanding that those values can be found is enough to comprehend MCR.

A standard measure of how well a particular set of values for α and β work to reveal the relationship between the values of X and Y is R^2, as defined in (Eq. 2).

$$\text{(Eq. 2)} \quad R^2 = 1 - \frac{\sum \varepsilon_i^2}{\sum (Y_i - \overline{Y})^2} = 1 - \frac{unexplained\ \ variation\ \ of\ \ Y}{total\ \ variation\ \ of\ \ Y}$$

So R^2 is 1 minus how much of the variation of Y is left unexplained by X. Or in other words how much of the variation is actually *explained* by X. Given a sample of data and α and β that minimize ε_i^2, the higher R^2 is, the better X is at, roughly speaking, explaining Y.

In MCR, R^2 is what will be considered as a measure of success in finding the best variables to explain Y.

Monte Carlo

The other technique used by MCR is Monte Carlo, which is a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Their essential idea is using randomness to solve problems that might be deterministic in principle[1].

We are assuming that there are deterministic causes to Y, and we want to use Monte Carlo to randomly create X variables that could explain the value of Y. These X variables will be created by combining other observed variables called Z variables.

$$\text{(Eq. 3.A)} \quad X_i = f\ (Z_{1,i}\ ,\ Z_{2,i}\ ,\ Z_{3,i}\ ,\ \ldots\ Z_{n,i})$$

Therefore, the data scientist will start with a sample of observations with data for variables Y and Z_1 to Z_n. Then MCR will repeatedly go through the steps below:

1. Randomly pick a subset of the Z variables.[2]
2. Place these variables in a randomly-generated function (this is where Monte Carlo is used) that will result in X.
3. Run a linear regression of X on Y, and calculate its R^2 value.
4. If R^2 is low, discard the attempt.

[1] From Wikipedia.

[2] Steps 1 and 2 are presented here in reverse order for clarity. Please find below a more accurate description of how MCR will go through all these steps.

5. If R^2 is high, record the *Z* variables used, the function generated, and the R^2 found.

We will refer to each run through the steps above as an iteration. At the end, the data scientist will have the best-performing functions as measured by the R^2 they achieved. The next section will go through the iterations in more details.

Monte Carlo Iterations

The aim of steps 1 and 2 is to turn Eq. 3.B into a numeric equation.

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

Each *j* is a different iteration, and each *i* is a different observation in our sample.

For each Monte Carlo iteration, MCR will randomly create a different *f* function such that, although we keep all *Z* variables unchanged, the final *X* value will be different each time. The general form of the *f* function is that $X_{j,i}$ will be equal to:

(i) the sum of 1 to 4 blocks;
(ii) each block consisting of 1 to 5 *Z* variables multiplying each other, and also multiplying a number from -10 to +10; and
(iii) each variable taken to a power ranging from -10 to +10.

What MCR will do at each iteration is to randomly pick numbers within those ranges (1 to 4 for number of blocks, -10 to +10 for each block multiplier, etc), and then finally the *Z* variables themselves. In other words, MCR will:

(a) Randomly pick a number between 1 and 4 for the number of blocks.
(b) For each block:
    (b.1) randomly pick a number between 1 to 5 for the number of *Z* variables in the block.
    (b.2) randomly pick the variables used in the block.
    (b.3) randomly pick a number between -10 and +10 to multiply the block.
(c) For each variable, randomly pick a number between -10 and +10 as the power it will be taken to.

To make things clearer, let us go through one hypothetical such iteration. In other words, we are determining f_j when $j = 1$. The first thing MCR will do is to determine the number of blocks on the function. This is related to component (a) from the list above. For f_1, say that the number randomly picked is 3. This means that f_1 will have 3 blocks, or that:

$$\text{(Eq. 4.A)} \quad X_{1\ ,\ i} = block_1 + block_2 + block_3$$

Each block is a set of variables multiplying each other. These variables will be determined further below. Next we need to determine the values of the multipliers of each block, as described in component (b.3) in the list above. These multipliers will range between -10 and +10. For $j = 1$, let us assume that the numbers +4, -7 and +2 have been randomly picked for blocks 1, 2 and 3 respectively.

Next, we need MCR to determine the size of each block, as described in component (b.1) in the list above. Let us suppose that MCR randomly picked numbers 4, 1 and 2. Combining components (b.3) and (b.1) we have that:

- $block_1$ will be the product of 4 different Z variables and its multiplier +4;

- block$_2$ will be the product of 1 Z variable and its multiplier -7; and
- block$_3$ will be the product of 2 different Z variables and its multiplier +2.

Our function then becomes:

$$\text{(Eq. 4.B)} \quad X_{1,i} = 4 \cdot Z_{a,i} \cdot Z_{b,i} \cdot Z_{c,i} \cdot Z_{d,i} - 7 \cdot Z_{e,i} + 2 \cdot Z_{f,i} \cdot Z_{g,i}$$

We now need MCR to determine the power each variable will be taken to as described in component (c) in the list above. Since we have seven variables in total, MCR randomly picks +2, -1, +4, +3, -5, +1, -2. That gives us:

$$\text{(Eq. 4.C)} \quad X_{1,i} = 4 \cdot Z_{a,i}^{2} \cdot Z_{b,i}^{(-1)} \cdot Z_{c,i}^{4} \cdot Z_{d,i}^{3} - 7 \cdot Z_{e,i}^{(-5)} + 2 \cdot Z_{f,i}^{1} \cdot Z_{g,i}^{(-2)}$$

Finally, we need to know which variables are going to be used for this particular instance of *f*. This is component (b.2) from the list above. Suppose we have 8 variables to draw from. Each variable may only appear once in each block. Let us suppose that MCR randomly picked these variables: Z_3, Z_5, Z_8, Z_2, Z_5, Z_8, Z_1. That gives us $X_{1,i}$:

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

Note how each variable randomly picked was placed in the function in the same order that they were drawn. This equation will give us an *X* value for each observation we have in our sample. But before we do that, we still need to go through one last transformation to the *Z* variables. We need to fit them in a 1-99 range.

$$\text{(Eq. 5)} \quad zz_{w,i} = \frac{Z_{w,i} - \text{MIN}(Z_w)}{\text{MAX}(Z_w) - \text{MIN}(Z_w)} \times 99 + 1$$

That is because in the way that the functions are created, some of the variables divide other variables (such as Z_5 above). So we need to make sure that no variable put into the equation has a value of zero for any observation. We will do that by fitting every value of each variable in a 1 to 99 range. So instead of using Z_i, we will use zz_i. Once we find the relationship between *zz's* and *Y*, it is easy to find the relationship between *Z's* and *Y*.

So what we have at the end of the iteration when *j* = 1 is that:

$$\text{(Eq. 6.A, from Eq. 2)} \quad R_1^2 = 1 - \frac{\sum \varepsilon_{1,i}^2}{\sum (Y_i - \bar{Y})^2}$$

where:

$$\text{(Eq. 6.B, from Eq. 1.B)} \quad \varepsilon_{1,i} = Y_i - \alpha + \beta \cdot X_{1,i}$$

where:

$$\text{(Eq. 6.C)} \quad X_{1,i} = 4 \cdot zz_{3,i}^{2} \cdot zz_{5,i}^{(-1)} \cdot zz_{8,i}^{4} \cdot zz_{2,i}^{3} - 7 \cdot zz_{5,i}^{(-5)} + 2 \cdot zz_{8,i}^{1} \cdot zz_{1,i}^{(-2)}$$

where:

$$\text{(Eq. 6.D, from Eq. 5)} \quad zz_{w,i} = \frac{Z_{w,i} - \text{MIN}(Z_w)}{\text{MAX}(Z_w) - \text{MIN}(Z_w)} \times 99 + 1$$

Eq. 6.A will give us how well *X* explains *Y*, which is what we are trying to understand. MCR will then repeat the steps for another iteration millions of times again. Each time we will build a different function and find a different R^2. Of course chances are that R^2 will be low for most of the iterations, but we are hoping that some of those R^2's will be much higher than we would find through other methods.

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 analysis from some initial results.

Use Case: Understanding GDP Per Capita

We will now go over a sample use case for illustration. All the data used, along with the results found, are in the GitHub repository mentioned above under the folder "gdp".

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* – the *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 – 60 minutes in a Personal Computer. The aim of this demonstration is simply to illustrate how the model works.

Best Result

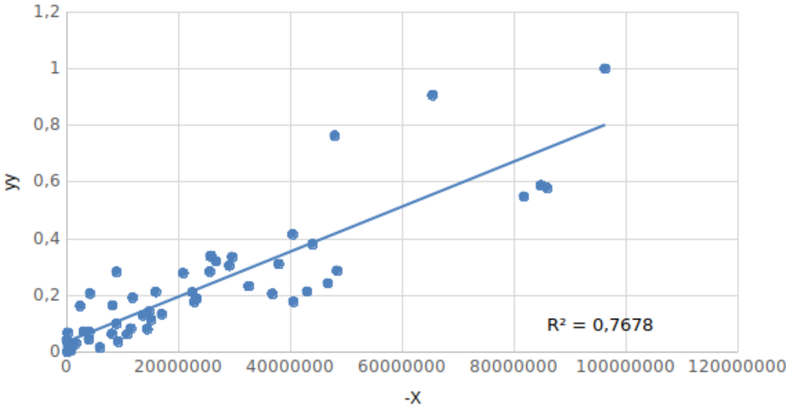
During the time period that the model has been run, MCR produced 1,983,739 *f* functions using the *Z* variables above, and therefore the same number of *X* variables. Regression number 990,246 found the highest R^2:

$$\text{(Eq. 7)}\quad X_{990246,\ i} = -\ 52.91 \cdot zz_{14,\ i}^{0.95} \cdot zz_{16,\ i}^{2.18} \quad // \ R^2 = 0.7678$$

This means that when MCR ran the iteration for the 990,246th time, the MCR model randomly picked:

(a) 1 as the number of blocks;
(b.3) -52.91 as the multiplier for that block;
(b.1) 2 as the number of variables for that block;
(c) 0.95 and 2.18 as the power those variables should be taken to; and
(b.2) zz_{14} and zz_{16} as the variables used.

$X_{990246,i}$ has been reconstructed in Excel (see proof.xlsx), providing the same R^2 as calculated in R. Below is the graph of -*X* plotted against *yy*. *yy* is *Y* passed through the same transformation made on each *Z* variable to become *zz*. -*X* has been used instead of *X* to highlight the positive relationship between zz_{14} and zz_{16} with *yy*.



The relationship between *X* and *yy* is clear. Since *yy* is just a linear transformation of *Y*, and since *X* is easily calculated from zz_{14} and zz_{16}, which in turn are linear transformations of Z_{14} and Z_{16}, we can say that we have found a relationship of *Y* with Z_{14} and Z_{16}.

Top 10 Results

Find below the top 10 results as R will display them. In the first column we find the function's R^2 ranking. The top one [1] is the function we saw above with an R^2 of 0.7678. In the other rows we find the other functions that made it to the top 10 list, from [2] to [10].

In the second column we find the R^2 of each function. And in the third column we find a human readable representation of the function. Note that what R lists as Z_w is actually what we have described as zz_w above, as MCR does not change the variable name during the transformation presented in this paper.

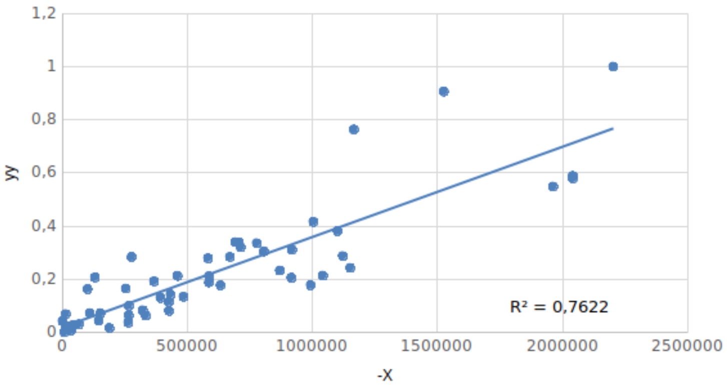
```
$eq
 [1] "0.7678 => ( -52.91 x Z14^0.95 x Z16^2.18 )"
 [2] "0.7676 => ( 3.25 x Z14^0.96 x Z16^2.25 ) + ( 60.01 x Z1^-7.41 x Z9^-2.15 )"
 [3] "0.7665 => ( 34.48 x Z14^1.09 x Z16^2.29 )"
 [4] "0.7658 => ( 87 x Z14^1.27 x Z3^0.93 x Z8^0.94 )"
 [5] "0.7639 => ( -77.23 x Z16^1.62 x Z14^1.04 x Z9^-0.21 ) + ( 78.05 x Z12^0.53 x Z5^-5.92 )"
 [6] "0.7636 => ( 95.72 x Z3^-0.11 x Z14^2.82 ) + ( 49.7 x Z16^2.94 )"
 [7] "0.7634 => ( -93.85 x Z16^2.28 x Z14^0.72 ) + ( -29.74 x Z2^1.86 ) + ( -11.2 x Z10^-7.91 )"
 [8] "0.7632 => ( 49.98 x Z14^0.92 x Z16^2.62 ) + ( -47.42 x Z10^-5.49 )"
 [9] "0.7622 => ( 24.94 x Z2^-9.27 x Z1^-0.1 x Z8^-8.2 x Z5^-4.86 ) + ( -10.7 x Z16^1.91 x Z14^0.75 x Z15^-0.04 )"
[10] "0.7613 => ( 17.32 x Z14^0.74 x Z16^2.58 ) + ( -58.24 x Z15^-6.58 x Z2^-7.6 x Z13^-4.87 )"

$r2
 reg-1-990246   reg-1-62551  reg-1-524914 reg-1-1001613  reg-1-512030 reg-1-1127337 reg-1-1132634  reg-1-199977
    0.7678103     0.7675925     0.7665208     0.7658461     0.7638750     0.7635670     0.7633877     0.7631705
reg-1-407567  reg-1-779795
   0.7621508     0.7613416
```

At the bottom, we can see a list of each regression above its R^2 value. Each regression is in the format reg-session-iteration. Since only one session was run, all sessions are equal to 1. And “reg-1-990246” is once more the function we saw above.

One thing to note is that all functions have their R^2 within the very tight range of 0.7613 – 0.7678. This suggests that we could be nearing the highest possible R^2 with these variables. The fact that some of the functions are very different to the others (such as [1] and [9]), on the other hand, indicate that we would *not* be so close to the highest R^2 possible since the top few functions would presumably be just micro-variations from the top one.

Function “reg-1-407567” (ranked 9th in the list above) has also been replicated in Excel in order to show how a more complex 7-variable function would look like:



Relationship Between Independent Variables

Furthermore, Z_{14} and Z_{16} seem to work very well together. Out of the ten functions, they are together in the same block in eight of them, with the power of Z_{16} usually being around twice as high as that of Z_{14}. This could indicate that these variables help each other in explaining Y, and this relationship would hardly be found using other models, unless theory provided us a hint to actively investigate.

When 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 when both of them are high, then Y_i will be high.

When $\mathbf{Z_{14,i}}$ *is*	*and* $\mathbf{Z_{16,i}}$ *is*	*then* $\mathbf{Z_{14,i} \times Z_{16,i}}$ *is*	$\mathbf{X_i = -52.91 \times Z_{14,1}^{0.95} \times Z_{16,1}^{2.18}}$ *is*	*and* $\mathbf{Y_i}$ *is*
low	low	low	low	low.
low	high	low	low	low.
high	low	low	low	low.
high	high	high	high	high.

So a country will only have a high GDP per capita (Y) if it has a high percentage of adults holding bank accounts (Z_{14}) *and* its life expectancy (Z_{16}) is also high. If either of those variables is low, GDP per capita will be low because it is highly correlated to the product of those variables.

A possible deterministic explanation would be 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.

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.

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 relationships among them. The main drawback is the need for greater processing power.

The model is currently in its second draft version. A lot more needs to be discussed and further developed. Once a function with high explanatory power is found, it will likely be very complex and therefore very difficult to understand at first any causality between the variables.

In order to simplify its results, would differentials be a valid approach? If we calculate the change in Y for a small change in each Z, would the result be meaningful? Should we use derivatives instead?

Are there other parameters of the f function that could be randomly picked by MCR? Would the results be better without the multipliers?

Should the model after, say, 10,000 attempts average the top 10 results into a new *X* and assess its explanatory power? The reason behind possibly doing this is that it may be a better attempt at finding a higher R^2 than simply starting another *f* function from random.

Or should it instead try that with every combination of two equations from the top 10? Os should it take one of the top *f* functions and add/remove 0.01 from the power of one of the variables picked at random? 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.