Corresemblance – core-resemblance, or deriving of the non-existence of information added-values

László Pitlik, Pitlik Marcell, MY-X team

Abstract: The goodness (e.g., correlation between facts and their estimated values) of the regression models built from more than one independent variable for estimation the values of a given dependent variable can not be estimated in advance with an arbitrary correctness. In order to create estimation methodologies for these goodness values, two different approaches have been realized. The experiment involved 5 independent variables (Xi) and 1 dependent variable (Y) and only the models with two independent variables (Y=f(Xi,Xj)) got derived. Each number of the experiment were a random number – generated between 10 and 99. In frame of the experiment following regression models have been estimated: 5 models with the 5 independent variables (Y=f(Xi)), 10 models with two independent variables (Y=f(Xi,Xj)), where the number 10 is the number of combination for 2 objects (variables) from 5 objects (variables). In each case (5+10), the parameters of the regressions and the correlations between the raw values, estimated values and the facts (Y) got calculated. Quasi the same process could also be executed with similarity analyses instead of regressions – it means there are available 5+10 staircase functions too – with correlations between facts and estimated values. The differences between regression-based and similarity-based modelling are: The similarity analyses produce 5+5 direct and inverse models for each Xi and the models based on two independent variables also use direct and inverse ranked inputs (c.f. double inputs). Based on the above mentioned pre-works, there are one single question to answer: How good can we estimate the correlation values of the cases Y=f(X(i),X(j) (from regressions and staircase functions) based on the descriptive attributes of the cases Y=f(Xi) (also from regression and staircase functions)? The descriptive attributes of the regressions are a&b&correlation for each variable where the symbols can be interpreted based on the following equation: y=a\*x(i)+b. The descriptive attributes of the direct and invers staircase functions are: the number of the different stairs, the maximal and the minimal number of the stairs with the same level, the standard deviation of the number of the stairs with the same level, validity of the estimations based on function symmetries of the staircase functions, correlation between the raw values of Y and their estimated values in case of the two variables (altogether 2\*5\*2+1 = 22 attributes). Conclusions: the regression-oriented inputs about the single-variable-models and the similarity-oriented inputs about the single-variable-models in direct and inverse versions have the same information added-value concerning the correlations based on regressions and similarities in case of Y=f(Xi,Xj). The correlation values of the models based on similarities are massive higher than the correlation values based on linear regressions. Furthermore, the regression models based on estimated values of single and double regression/similarity models produce lower correlation concerning the values of the dependent variable than the similarity models based on estimated values of single and double regression/similarity models. The best estimations concerning the values of the dependent variables delivered a correlation value of 1.00 in case of (5+1)\*20 random numbers… The lack of potential in case of the regressions can also be seen through the correlation value for similarity-based model outputs where the difference between the similarity-based model and the regression-based model is robust.

Keywords: regression, similarity, estimation of correlations

[Introduction 2](#_Toc87595721)

[Literature 3](#_Toc87595722)

[Own approach 3](#_Toc87595723)

[Initializing of the experiment 3](#_Toc87595724)

[Single models 4](#_Toc87595725)

[Regression models 4](#_Toc87595726)

[Staircase models 5](#_Toc87595727)

[Multiple models 5](#_Toc87595728)

[Regression models 5](#_Toc87595729)

[Staircase models 6](#_Toc87595730)

[Modelling of the correlation values from multiple models based on single models 6](#_Toc87595731)

[Regression-based X and Y with staircase functions (reg-reg-sim) 6](#_Toc87595732)

[Similarity-based X and regression-based Y with staircase functions (sim-reg-sim) 7](#_Toc87595733)

[Regression-based X and similarity-based Y with staircase models (reg-sim-sim) 7](#_Toc87595734)

[Similarity-based X and similarity-based Y with staircase models (sim-sim-sim) 8](#_Toc87595735)

[Estimation of the original dependent variable based on variations of the single and multiple (2) models 9](#_Toc87595736)

[Discussion 12](#_Toc87595737)

[Conclusions 12](#_Toc87595738)

[Annex 12](#_Toc87595739)

[References 12](#_Toc87595740)

# Introduction

Multiple (regression) models can be constructed in different ways. The number of the potential models is unlimited – because the length of the models (then number of the integrated variables) is unlimited – even if the number of the variables are limited. However, the best (regression) model could not be derived in an automated way since ever. The paper presents the experiences/results of the challenge: how exact can be derived the fitting of a multiple (regression and/or staircase) models based on the fitting values and descriptive parameters of single (regression and/or staircase) models?! As a quasi side-effect: it will be visible, which fitting levels can be realized based on linear regressions and based on staircase function (especially with direct and inverse – but monotonous views – used them in a parallel way). Parallel, the randomized generated Y values can be derived with arbitrary high correlations based on parallel models and their similarities, but not with regression models in the same logical ways.

# Literature

The connection between the fitting/parameters of single linear regression models and the fitting of multiple linear regression models (in order to estimate the best multiple models based on experiments with the single ones), seems to be irrelevant till now for the scientific community.

The challenge of the derivation of the best (linear) regression models concerning a given data set in a context-free way seems also to be not solved – e.g., the problem of the multiple evaluation (c.f. <https://blog.minitab.com/en/how-to-choose-the-best-regression-model>) is not used in general (c.f. anti-discriminative modelling based on staircase functions: <https://miau.my-x.hu/miau/196/My-X%20Team_A5%20fuzet_EN_jav.pdf>). A problem is just solved, if the solution can be delivered through an algorithm (c.f. KNUTH’s principle about the science/knowledge: [https://miau.my-x.hu/miau2009/index\_tki.php3?\_filterText0=\*knuth](https://miau.my-x.hu/miau2009/index_tki.php3?_filterText0=*knuth)).

The involving of independent variables with different correlation to Y can however lead to specific situations: <https://miau.my-x.hu/miau/274/real_values_of_attributes.docx>

# Own approach

Details: <https://miau.my-x.hu/miau/278/corresemblance.xlsx>

The paper tries to describe the steps (presented in the XLSX-file - see above) in a reproducible way.

## Initializing of the experiment

The first step is an OAM (object-attribute matrix) – in this case with 20 objects and 5+1 attributes (see Fig.Nr.1):



Fig.Nr.1 – The random OAM and its ranking values (source: own presentation)

The stating values have been generated with in a randomized way (see values between 10 and 99). The number of objects and attributes are also a kind of randomized number – where the necessary visualization effect has been accepted (see: not too large for fitting to a page like this).

The green/red cells show the impact of the repeated raw values (concerning Xi).

## Single models

Single (regression and staircase) models can be described with a lot of indicators in MS Excel (c.f. XLS-file: sheet = “info”):



Fig.Nr.0 - [https://support.microsoft.com/hu-hu/office/lin-ill-f%c3%bcggv%c3%a9ny-84d7d0d9-6e50-4101-977a-fa7abf772b6d?ns=excel&version=90&syslcid=1038&uilcid=1038&appver=zxl900&helpid=xlmain11.chm60097&ui=hu-hu&rs=hu-hu&ad=hu](https://support.microsoft.com/hu-hu/office/lin-ill-f%C3%BCggv%C3%A9ny-84d7d0d9-6e50-4101-977a-fa7abf772b6d?ns=excel&version=90&syslcid=1038&uilcid=1038&appver=zxl900&helpid=xlmain11.chm60097&ui=hu-hu&rs=hu-hu&ad=hu)

Staircase functions can be derived with the Solver of MS-Excel and/or based on online tools like: <https://miau.my-x.hu/myx-free/>, <https://miau.my-x.hu/myx-free/coco/index.html>

The description of the direct and inverse staircase functions should be worked out in frame of this paper. The direct models mean each ranking direction is set with the option “the-more-the-more” concerning the connection between Xi and Y. The inverse models mean each ranking direction is set with the option “the-less-the-more” concerning the connection between Xi and Y.

### Regression models

Based on the raw values (see XLS-file: sheet = “raw”, range = A1:AA23), it is already possible to derive the parameters and fitting values of the 5 possible single regression models (see: XLS-file: sheet = “raw”, range = AC1:AF8) – following of the pattern: Y = a\*Xi+b (a=p1, b=p2)



Fig.Nr.2 – Parameters and correlations of the SLR-models (source: own presentation)

Fig.Nr.2 shows: the correlations must always be positive in these cases because of the sing of p1 – but the correlation of the raw values (Xi vs. Y) can lead to negative correlations (see Fig.Nr.1).

### Staircase models

The XLS-file has 5 sheets: X5di\_Y, X4di\_Y, X3di\_Y, X2di\_Y and X1di\_Y where the Xi and Y can be interpreted in a trivial way, di means direct and inverse models.

The 5 sheets present to models in each sheet where the staircase functions for one single Xi can be seen in case of the direct and the inverse ranking inputs.

The staircases are mostly different. The descriptive statistics could be derived based on the pivot-reports (see XLS-file: sheets = “pivot”).

The descriptive attributes of a staircase function are (c.f. XLS-file: sheets “OAM\_sim”, range: A1:X15):



Fig.Nr.3 – Descriptive variables for staircase functions (source: own presentation)

Staircase functions can be characterized e.g., through their count of steps, maximum and minimum values of the stairs with the same level, standard deviation of numbers of stairs with the same level in the stair, validity, and correlations (for direct and inverse staircases). This set of information let expect, that these descriptions have more information added-value than the parameters of the single regression models (a, b, fitting=correlation). It is important to highlight here and now: the paper’s title focuses on this added-values, on the existence of this added-values!

## Multiple models

It is important to clarify here and now: The paper works only with two independent variables (Xi, Xj from the given 5 X-attributes).

### Regression models

The XLS-file (sheet = “raw”, range = AO1:AU13) demonstrates the results of the multiple regression models:



Fig.Nr.4 – The parameters of the multiple regressions (source: own presentation)

Fig.Nr.4 presents the slopes (p1, p2) and the constants (p3). There are also available correlation values between the facts of Y and the estimations concerning the Y). The correlation values of the regression models with 2 independent variables are massive lower than the correlation values based on 2 variables (but direct and inverse views of them) in frame of staircase functions. The correlations based on regressions and based on similarities (staircase functions) produce a reduced parallelism (see: correlation value = 0.451). It means the regressions and the staircases are quite different.

### Staircase models

The XLS-file (sheet = “xixj\_Y\_sim”) presents the similarity analyses with 2+2 (direct+inverse) variables. Their relatively high correlation values could be seen in Fig.Nr.4 – see above).

These staircase models should not be described with specific staircase characteristics because only the correlations of the multiple models will be used in the further steps.

The relatively high correlation values are caused through the higher flexibility of the staircase functions concerning to the regression model parameters. Let alone: the monotonous direct and inverse layers increase this flexibility.

## Modelling of the correlation values from multiple models based on single models

This part is a kind of innovation (see lacks in the literature). The single models can always be described with a lot of characteristics (see Fig.Nr.0 in case of the regression models) and descriptive attributes can be constructed (in case of similarity-oriented staircase functions).

The following sub-chapters belong to a combinatorial space of 2\*2\*2 where the inputs can be defined in 2 ways (inputs are regression-based or staircase-based). Parallel, the output (Y) can also be defined in 2 ways (correlation from regressions or staircase functions), and also the modelling frames can be defined in 2 ways (models are regression models or similarity-based models).

### Regression-based X and Y with staircase functions (reg-reg-sim)

The Fig.Nr.5 demonstrates (see XLS-file: sheet = “OAM\_reg”) a model with its estimation where the inputs are 6 attributes (2\*3 – it means a,b,correlation for each single Xi and Xj):



Fig.Nr.5 – Regression parameters of single regression models for interpretation of the fitting of multiple regression models (source: own presentation)

The primary conclusion is trivial: the regression-based inputs (X) are capable of describing the regression-based Y based on a similarity-oriented model because the correlation is 0.998.

### Similarity-based X and regression-based Y with staircase functions (sim-reg-sim)

Fig.Nr.6 demonstrates (XLS-file: sheet = “oam\_SIM”) the descriptive variables based on single staircase functions (for Xi and Yi) and the correlations from the multiple regression models.



Fig.Nr.6 - Staircase parameters of single staircase models for interpretation of the fitting of multiple regression models (source: own presentation)

The basic conclusion is also simple as before: there is no difference to identify between the information potential of the regression-based and the similarity-based inputs concerning the regression-based outputs. Let alone, the results are the same for sheet = “oam\_SIM2” and sheet = “oam\_SIM3” where the directions are for all variables the same (0) and the correlation are excluded.

### Regression-based X and similarity-based Y with staircase models (reg-sim-sim)

The Fig.Nr.7 presents (XLS-file: sheet = “oam\_reg\_simy”) the challenge where the regression-based inputs try to explain the similarity-based outputs based on similarities.



Fig.Nr.7 - Parameters of single regression models for interpretation of the fitting of multiple staircase models (source: own presentation)

Conclusion: the estimations could explain the outputs with a correlation of 0.901619. It is lower than in case of the regression-based outputs.

### Similarity-based X and similarity-based Y with staircase models (sim-sim-sim)

The Fig.Nr.7 presents (XLS-file: sheet = “oam\_reg\_simy”) the challenge where the regression-based inputs try to explain the similarity-based outputs based on similarities.



Fig.Nr.8 - Parameters of single staircase models for interpretation of the fitting of multiple staircase models (source: own presentation)

Conclusion: the estimations could explain the outputs with a correlation of 0.90698. It is lower than in case of the regression-based outputs, but it is quasi the same as in case before (see Fig.Nr.7).

The similarity based (higher correlations) could therefore be derived with a lower correlation value in both cases (see regression-based X and similarity-based X in frame of staircase models).

(The above mentioned SIM2 and SIM3 variations present the same results.)

### Regression-based X and Y with regression model (reg-reg-reg)

Regression models can also be built based on inputs involving correlation data or not. Correlation, as attribute is the common phenomenon being valid for regression- and similarity-based models. Correlations are measurements for estimated and original Y values.



Fig.Nr.9 – source: own presentation (XLS-file – sheet = “oam\_reg (2)”)

There are no real differences between similarity and regression-based models concerning the coloured correlation values using the same X-Y-pattern for analyses.

### Similarity-based X and regression-based Y with regression model (sim-reg-reg)



Fig.Nr.10 – source: own presentation (XLS-file – sheet = “oam\_sim\_reg1a”)

There are no real differences between similarity and regression-based models concerning the coloured correlation values using the same X-Y-pattern for analyses – with correlations.



Fig.Nr.11 – source: own presentation (XLS-file – sheet = “oam\_sim\_reg1b”)

There are already real differences between similarity and regression-based models concerning the coloured correlation values using quasi the same X-Y-pattern for analyses – without correlations.

### Regression-based X and similarity-based Y with regression model (reg-sim-reg)



Fig.Nr.12 – source: own presentation (XLS-file – sheet = “oam\_reg (3)”)

There are massive differences between similarity and regression-based models concerning the coloured correlation values using the same X-Y-pattern for analyses. It means regression models are not capable to extract the information from the OAM with the effectiveness like similarity based models (see last chapter about estimation based estimations).

### Similarity-based X and similarity-based Y with regression model (sim-sim-reg)



Fig.Nr.13 – source: own presentation (XLS-file – sheet = “oam\_sim\_reg (2)”)

There are real differences between similarity and regression-based models concerning the coloured correlation values using the same X-Y-pattern for analyses – with correlations. This is the only case where a regression-based model could produce better performance than the similarity-based models.



Fig.Nr.14 – source: own presentation (XLS-file – sheet = “oam\_sim\_reg (3)”)

There are no real differences between similarity and regression-based models concerning the coloured correlation values using the same X-Y-pattern for analyses – without correlations.

## Summary of the 8 combinatorial cases

The number of the cases in the combinatorial space could have been bigger if the existence of correlations among the input attributes were involved in a systematic way (see Fig.Nr.15):



Fig.Nr.15 – source: own presentation (source: own presentation)

Summa summarum: The regression-based models are in general less potent than the similarity-based models (see \*\*reg vs. \*\*sim). One single case (using correlations in the OAM) can be seen as a kind of special effect (see simsimreg vs simsimsim) – but without correlation values in the input, the general characteristic is given. Correlations are not staircase-specific descriptors!

The most relevant difference can be seen between regsimreg and regsimsim! The regression-based X and the regression model is less capable for interpreting the similarity-based Y than the similarity models. But the information potential is given in the regression-based inputs for each kind of outputs.

# Estimation of the original dependent variable based on variations of the single and multiple (2) models

The special challenge (see XLS-file: sheet = “estimation\_based”) was to derive models based on estimation values of the single models (regression-based and similarity-based) and also based on estimation values of the multiple models (also regression-based and similarity-based).



Fig.Nr.16 – Estimation-based estimation I. (source: own presentation)

The Y values are the original, randomly generated values (between 10 and 99). In case of the similarity based-models each Y value was transformed (Y’=Y\*1000). The coloured cells are correlation values: in case of Xi, the correlation concerns always the original Y. The so-called naïve columns are always average values of the available inputs. The last row demonstrates the differences between the correlations of the naïve and the similarity-based or regression-based models. Fig.Nr.16 presents similarity-based models with massive differences (quasi double correlation values).

Fig.Nr.17 shows further similarity-based models: the differences are rel. small. It is however to know that the input-estimations are also derived in a similarity-based way. The most important result is the correlation=1.000 position (on the bottom, in the middle). The naïve model could already increase the correlation of the single inputs (0.65---0.84) to 0.92. The optimized model made possible to reproduce 20 randomly generated Y-values based on the combinatorial set of multiple similarity models. This result demonstrates the quasi unlimited potential of the similarity-based knowledge representation concerning the randomization as such.



Fig.Nr.17 - – Estimation-based estimation II. (source: own presentation)

XLS-file: sheet = “reg1of4”



Fig.Nr.18 – Estimation-based estimation III. (source: own presentation)

Fig.Nr.18 presents the same inputs as before in case of Fig.Nr.16-17 – but the models here are regression-based. The correlation values are low and the naïve and the so-called optimized models have quasi the same correlation levels (below the half of the similarity-based models).

Fig.Nr.19 shows the further two scenarios with the same conclusions as in case of Fig.Nr.18 – but on a higher input-correlation level (because of the similarity-based estimation as inputs).

XLS-file: sheet = “reg2of4”



Fig.Nr.19 – Estimation-based estimation IV. (source: own presentation)

XLS-file: sheet = “reg3of4”



Fig.Nr.20 – Estimation-based estimation V. (source: own presentation)

Fig.Nr.20 and Fig.Nr.21 present the last two scenario (inputs are based multiple models). The conclusions seems to be the same as before in case of Fig.Nr.18-19: low and quasi identical correlations where the final models are regression-based (see yellow coloured cells on the top of the figures 18-19-20-21 – calculated with Excel Solver).

XLS-file: sheet = “reg4of4”



Fig.Nr.21 – Estimation-based estimation VI. (source: own presentation)

The regression-based approach (modelling random numbers based on previous model-estimations) demonstrates quasi no chance for information added values contrary to the similarity-based alternative knowledge representation forms.

# Discussion

The above presented combinatorial space of inputs/outputs/models and objectives had a simple initializing expectation in the background: if the similarity-based knowledge representation is capable of estimating random values with a quasi unlimited accuracy, then the potential of the similarities could even be identified concerning the single similarity models. This expectation could not be proved: the regression-based inputs have quasi the same information potential as the descriptive variables about staircase functions. The difference between the similarity- and the regression-based inputs is small, but the regression-based solutions produced always lower correlation values than the similarity-based alternatives.

The similarity-based knowledge representation form has however massive advantages compared to the regression-based alternatives concerning the correlation levels.

# Conclusions

There are no real differences between the single regression-based inputs for modelling correlation values of multiple models and the similarity-based ones.

The randomly generated Y values could be derived with an arbitrary fitting based on similarity analyses.

# Annex

<https://miau.my-x.hu/miau/278/corresemblance.xlsx>

# References

…see in text…