**Az attribútumok valódi értékének speciális összefüggései parallel termelési függvények generálásakor, (avagy regresszió helyett lépcsős függvények parciális korrelációi)**

(Special rules for the value of attributes concerning parallel derivation of production functions – or staircase function with partial correlation instead of regression)

Pitlik László, Pitlik Marcell (MY-X team)

Kivonat: A laikus modellező azt vélelmezi, hogy adott következményváltozó (Y) modellezése annál sikeresebb, minél (az Y kapcsán) több magas pozitív korrelációt mutató független változó (Xi) épül be egy modellbe. Vagyis a nagy negatív korrelációjú attribútumok quasi a legértéktelenebbek. A hajósok is azt hitték évszázadokig, hogy csak hátszéllel lehet előre haladni – míg nem rájöttek a megfelelő technikákra. A regressziós modellekre, melyek évtizedek óta a modellezés gerincét képezték, sajnos igaznak tűnik a gyanú, hogy kevésbé tudnak mit kezdeni a nagyobb negatív korrelációjú építőelemekkel, mint a pozitív korrelációjúakkal. A hasonlóságelemzésre azonban ez már nem igaz. Sőt, a hasonlóságelemzés dupla-attribútumkészletű modelljei képesek a káros polinomizációt is domesztikálni. A hasonlóságelemzés irány-preferenciákat tartó termelési függvényei kapcsán pedig bizonyítható, hogy a primer termelési függvény, vagyis a rendelkezési álló attribútumok látszólag legjavát kiválogató első termelési függvény tény vs. becslés korrelációja meghaladható, ha lépésről lépésre minden attribútumot bevonunk több párhuzamos termelési függvényvariánsba, a már bevont attribútumok fokozatos kizárása mellett. Az így előálló alternatív becslések, mint független változók alapján nem csak magasabb korreláció érhető el, mint a primer termelési függvény esetén, hanem a primer termelési függvény attribútumainak elhagyásával, vagyis a primer becslés elhagyásával is magasabb lehet még mindig a korreláció, mint a primer termelési függvény ún. csak kékvérű attribútumait felhasználva. Mindemellett a primer termelési függvény előállítása során legkevésbé értékesnek tűnő attribútumokból a kékvérű attribútumok számánál akár kevesebb attribútummal is el lehet érni a primer termelési függvény korrelációs szintjét, ill. ennek tetszőleges közelségét a mindenkori párhuzamos becslések dupla-attribútumos modelljeivel. Következésképpen: az attribútumok értékét ezek felhasználásának módja határozza meg, s nem a korreláció Xi és Y között: vagyis grafitporból is lehet tehát gyémántot préselni.

Mindezen additív modellezést feltételező gondolatok ellenpontjaként értelmezendők a multiplikatív regressziós modellek, mely kapcsán, ha egyetlen egy regressziós együttható is nulla lenne, akkor a szorzáslánc eredménye mindenkor nulla lenne, ami olyan Y esetén, ahol nincs nulla érték, téves becslés lenne. Így a multiplikatív modellek esetén formálisan általában véve nincs is kieső attribútum, azaz nincs fontosabb és kevésbé fontos attribútum – mert minden mindennel összefügg.

A modellezésről alkotott naiv hitek elsődleges cáfolatának bizonyítása egyetlen egy véletlen-jellegű OAM alapján végzett modellezési lépéssor esetén azonnal bizonyítékként értelmezendő, ha egyetlen egyszer is a naiv hiedelmekkel ellentétes trendek, arányok, viszonyszámok érhetők tetten – már pedig ez az egyediség egyetlen egy kísérletből, azaz azonnal előállítható volt. A benchmark tehát nem más, mint a naiv hiedelmekre jellemző trendek, arányok, viszonyszámok egyedi értéke, vagy ezek halmaza. A benchmarktól való eltérés hermeneutikája triviális: ha nem a benchmark a szélsőérték, akkor a naiv hiedelmek per definitionem tévesek.

A naiv hiedelmek tévességének/korlátozottságának bizonyítása kikényszeríti, hogy a jövőben semmilyen attribútum ne legyen elhanyagolható csak úgy. A primer termelési függvényekre alapozó hermeneutikák abban az esetben, ha más módszertannal a tény vs. becslés korreláció quasi tetszőleges magasra emelhető, vagyis előállhat a tény=becslés állapot minden objektumra, a STEP-IX módszertan bevonásával újra értelmezhetővé teszi a tényadatok becsülhetőségének (fuzzy-jellegű) árnyalatait.

S végül a parciális korreláció fogalma az a logikai magyarázat, mely a naiv elvárások és a látszólag ennek ellentmondó bizonyítékok között megteremti a kapcsolatot. A parciális korreláció, melyet lépcsős függvények képesek kezelni, a regressziók nem, lehetővé teszi, hogy egy-egy attribútumnak csak részletei hassanak a modellekben, ami a legnagyobb (teljes) negatív korrelációt is átfordíthatja akár a legnagyobb pozitív (parciális) korrelációba.

Kulcsszavak: regresszió, additivitás, multiplikativitás, hatás, hasonlóságelemzés, parciális korreláció

Abstract: The naïve modelling experts think, models will become better if more attributes (Xi) with (as far as possible higher) positive correlation values (compared to Y) can be involved into a model. Therefore, it seems to be that the attributes (Xi) with the most negative correlations are the less useful attributes to derive a production function for a Y-attribute. It is the same effect as in case of ships, which could only be used for centuries if the wind comes from behind – and suddenly the appropriate technique as such got identified. Unfortunately, the regression models (being since ever the fundamental solution in modelling challenges) can really intagrate attributes (Xi) with a higher probability when the correlation values (compared to Y) are positive and as higher as possible. All this is but not valid for the similarity analyses (for the staircase functions). The models with double attribute-sets are even capable of domesticating useless polynomial effects. In case of staircase functions, it is to derive, that the primary correlation value of a production function between the facts and estimations can be increased if (in case of a lot of Xi) the attributes of the first selection are hidden and based on this step-by-step logic more parallel production functions can be built – while each single attribute belongs to a production function AND a new production function will be created based on the estimations (Xi) of the parallel models. Even, the primary layer (estimation) and therefore each of its attributes can be excluded from the modelling process and the correlation value is still higher than in case of the primary model with the first selection of (“full-blood”-) attributes. The number of the full-blood-attributes can also be higher than the number of such (half-blood) attributes which are used for the less robust production functions in a direct way – and these quasi useless attributes lead to the quasi same high correlation values as the full-blood-attributes after the first selection phase – in form of more alternative solutions. Therefore, the value of an attribute depends on the involved modelling methods/techniques and not directly from the correlation between Xi and Y (c.f. from coal powder can diamond be pressed).

All these are valid for additive models. The alternatives, the multiplicative models can not exclude attributes in a way where the parameters will be set to zero because one single multiplication with null leads to null independent from each further attribute. So, the multiplicative models always reflect the basic rule: each attribute should always have a connection to each other attribute.

The invalidity of the naïve rule about correlation and usefulness can be proven based on one single case study where a random-like input OAM can demonstrate the above-mentioned results. The paper presents this successful case study – where the first analysis led to the needed evidence at once.

Based on this evidence, the attributes may not be interpreted from now on as before. If the new higher correlation level is arbitrary near to 1.000, then a new aspect of the modelling should be focused (c.f. STEP-IX-logic as an alternative approximation of the fuzzy-logic) – in order to be capable of talking about robustness of the modelling.

Finally, the bridge between the naïve logic and the explored evidence is the term of the partial correlation. This is only valid in case of staircase functions and not in case of regression models. The partial correlation makes possible that an attribute with highly negative (full) correlation value may deliver a highly positive (partial) correlation value if only parts of the known values will be interpreted.

Keywords: regression, additivity, multiplicativity, impact, similarity analysis, partial correlation

# Bevezetés

Előzmények:

* <https://arago.elte.hu/sites/default/files/Mindennapok-fizikaja-07.pdf> (hajózási analógia)
* <https://tanuljhajotvezetni.ewk.hu/vitorlazaselmelet/> (hajózási ismeretek)
* <https://miau.my-x.hu/myx-free/index_e8.php3?x=stp4> (STEP-IX)
* <https://miau.my-x.hu/miau/270/multikollinearitas_vs_hasonlosagelemzes.docx> (multikollinearitás)
* <https://miau.my-x.hu/miau/274/solver_driven_alternatives.docx> (validitás fuzzy-jellege)
* <https://miau.my-x.hu/miau/274/csak_invalid_objektum.xlsx> (csak invalid objektumok egy hasonlóságelemzés vigaszágán – önálló OAM-ként)
* <https://www.google.com/search?q=%22dupla+attrib%C3%BAtum%22+site%3Amiau.my-x.hu> (dupla attribútumkészlettel dolgozó modellek)
* <https://miau.my-x.hu/miau/272/Roo_Boo_Mooc_2.docx> (genetikai potenciál-alapú keresésvezérlés a genetikus algoritmusok alternatívájaként a modellezéstől elvárható jóság tervezését támogatandó)

A mesterséges intelligenciák lényege a matematikai statisztikai kánontól való másság akarása és képessége. A genetikus algoritmusok logikái ékes példáit jelentik ennek a másságnak, hiszen az összeadás, kivonás, szorzás, osztás, stb. műveletek helyett (mellett) új műveletek kerülnek előtérbe: pl. mutáció, keresztezés, szelekció. S ez még csak a kezdete annak a gondolkodásbeli sokszínűségnek, amire az emberi agy úgy képes, hogy mindez megfelel a knuth-i elvárásoknak – azaz automatizálható – sőt, innentől már optimalizálható (vö. robotszem vs. naiv emberi látás/értelmezés buborék modellekben/aranykor-modellekben: <https://miau.my-x.hu/miau/273/adidas_story.docx>, ill. <https://miau.my-x.hu/miau2009/index.php3?x=e0&string=eye>).

A regressziós modellek a korreláció fogalmán keresztül értelmezik a kapcsolatot a nyers attribútumok (Xi), ill. a becslések és a következményváltozó (Y) között. A korreláció kapcsán minden objektum adata értelmezésre került és eleve az objektumok nyers adatai hatnak a modellben/képletben is (vö: Y=a\*x+b). Ezzel szemben a lépcsős függvényekre alapozó termelési függvények képesek a parciális korreláció fogalmát kiaknázni, vagyis egy-egy független változó csak azon tartalmát értelmezni, melyek garantálják a többi változóval való együttműködés keretében a jobb tény vs. becslés korrelációkat. Ezt egészíti ki a dupla-attribútumos hasonlóságelemzések logikája, mely modellek egyszerre kétféle „regressziós” együttható-előjelet képesek kezelni (vö. káros polinomizáció kizárása mellett) és egyben mindkét előjel kapcsán a parciális korreláció előnyeit kiaknázva.

# Adatok

Részletek: <https://miau.my-x.hu/miau/274/production_function_chains.xlsx>

Vegyünk egy véletlenszerű primer OAM-állapotot:

* Objektumok száma = 34 (pl. gyógyszerek azonos beteg által bevehető azonos betegségre)
* Attribútumok száma = 94 (mellékhatás-gyakoriságok)
* Cellaértékek értelmezési intervalluma (rangsor-számok): 1., 2., 3., 4., 5., 6., 7.
* Y = valós jelenség (pl. gyógyszer-ár)
* Feladat: irány-preferenciatartó termelési függvények levezetése (optimalizáltan, solver-rel) – Milyen korrelációs benchmark áll elő a már modellbe bevont attribútumok kizárása esetén?

Állítsunk elő egy másodlagos OAM-ot:

* Objektumok száma = 34
* Attribútumok száma = 8 (párhuzamos modellek becslései), ill. dupla esetben 16 (valamint ennek szűkített nézetei – 14;12;10;8;6;4;2)
* Cellaértékek: az Y(tény) körül ingadozó értékek (tetszőleges távolságot felvéve a ténytől)
* Y = valós jelenség (pl. gyógyszer-ár – a primer OAM-mal értelemszerűen azonos)
* Feladat: termelési függvények levezetése (optimalizáltan, solver-rel) – Javul-e a korreláció a párhuzamos termelési függvények becsléseit, azaz egyre több attribútumot felhasználva?

Állítsunk elő egy kiegészítő OAM-ot:

* Objektumok száma = 8 (step1-2-3-4-5-6-7-8)
* Attribútumok száma = 4 (ill. dupla esetben 8) – a nyers Xi és a primer Y közötti korrelációs értékek statisztikái (csoportátlag, szórás, max, min)
* Cellaértékek = korrelációk (-1, …, +1)
* Y (step1-2-3-4-5-6-7-8): modell-korreláció
* Feladat: termelési függvények levezetése (optimalizáltan, solver-rel) – magyarázható-e a korreláció minden step esetén?

# Elemzések

A részleteket bemutató XLS az alábbi modelleket

* Regressziós modellek
	+ Additív
	+ Multikplikatív
* Hasonlóságelemzés (COCO-STD)
	+ Irány-preferenciát tartó
	+ Dupla-attribútumkészlettel dolgozó

tartalmazza:

* M1 (alias step1): 34 objektum, 94 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 27 db modellbe integrált attribútum
* M2 (alias step2): 34 objektum, 94-27=67 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 24 db modellbe integrált attribútum
* M3 (alias step3): 34 objektum, 94-27-24=43 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 16 db modellbe integrált attribútum
* M4 (alias step4): 34 objektum, 94-27-24-16=27 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 11 db modellbe integrált attribútum
* M5 (alias step5): 34 objektum, 94-27-24-16-11=16 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 8 db modellbe integrált attribútum
* M6 (alias step6): 34 objektum, 94-27-24-16-11-8=8 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 5 db modellbe integrált attribútum
* M7 (alias step7): 34 objektum, 94-27-24-16-11-8-5=3 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 2 db modellbe integrált attribútum
* M8 (alias step8): 34 objektum, 94-27-24-16-11-8-5-2=1 attribútum, Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – 1 db modellbe integrált attribútum
* M9: (steps – OAM1): 34 objektum, 8 attribútum (a step1-2-3-4-5-6-7-8 becsült értékei – irány=0), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – csak 4 db modellbe integrált attribútum (step1-2-3-5 – kihagyva: step4-6-7-8)
* M10: (steps – OAM2): 34 objektum, 8 attribútum (a step1-2-3-4-5-6-7-8 becsült értékei és a tényárak közötti abszolút eltérések – irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – mind a 8 db integrálva
* M11: (steps – OAM3): 34 objektum, 16 attribútum (a step1-2-3-4-5-6-7-8 becsült értékei – irány=0 ÉS a tényárakhoz képest mért abszolút hiba – irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD, direkt – csak 15 db modellbe integrált attribútum (kihagyva: step6nyers)
* M12: (steps – OAM4): 34 objektum, 16 attribútum (a step1-2-3-4-5-6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 10 db attribútum integrálva (vö. részletek xls)
* M13: (steps – OAM5): 34 objektum, 16 attribútum (a step1-2-3-4-5-6-7-8 becsült értékei és a tényárak közötti abszolút távolság – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 15 db attribútum integrálva (vö. részletek xls)
* M14: (steps – OAM6): 34 objektum, 14 attribútum (a step2-3-4-5-6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 13 db attribútum integrálva (vö. részletek xls)
* M15: (steps – OAM7): 34 objektum, 12 attribútum (a step3-4-5-6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 9 db attribútum integrálva (vö. részletek xls)
* M16: (steps – OAM8): 34 objektum, 10 attribútum (a step4-5-6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 6 db attribútum integrálva (vö. részletek xls)
* M17: (steps – OAM9): 34 objektum, 8 attribútum (a step5-6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 5 db attribútum integrálva (vö. részletek xls)
* M18: (steps – OAM10): 34 objektum, 6 attribútum (a step6-7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 5 db attribútum integrálva (vö. részletek xls)
* M19: (steps – OAM11): 34 objektum, 4 attribútum (a step7-8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 3 db attribútum integrálva (vö. részletek xls)
* M20(=M8): (steps – OAM12): 34 objektum, 2 attribútum (a step8 becsült értékei – dupla attribútumos modell irány=0&irány=1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), COCO STD (direkt/inverz-futtatás itt nem értelmezhető) – csak 1 db attribútum integrálva (vö. részletek xls)
* M21: regressziós additív modell 94 db Xi (1-2-3-4-5-6-7-es rangsorok a mellékhatások gyakorisága szerint – irány = 1 – vö. step1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), integrált attribútum csak 22
* M22: regressziós multiplikatív modell 94 db Xi (1-2-3-4-5-6-7-es rangsorok a mellékhatások gyakorisága szerint – irány = 1 – vö. step1), Y=transzformált ár (transzformáció=nyersadat\*1000+100000), integrált attribútum mind a 94
* M23: COCO STD, objektum = 8 (steps), attribútum = 5 (bevont attribútumszám, bevont attribútumok korrelációinak átlaga, maximuma, minimuma, szórása, Y = modellkorreláció) - direkt
* M24: COCO STD, objektum = 8 (steps), attribútum = 4 (bevont attribútumok korrelációinak átlaga, maximuma, minimuma, szórása, Y = modellkorreláció) - direkt
* M25: COCO STD, objektum = 8 (steps), attribútum = 4+4=8 (bevont attribútumok korrelációinak átlaga, maximuma, minimuma, szórása – irány=0&irány=1, Y = modellkorreláció) – (nincs értelmezve a direkt/inverz függvényszimmetria dupla-attribútumkészlettel dolgozó modellek esetén)

Megjegyzések:

* Az M8 és az M20 kényszerűen azonos.
* Mivel nem volt 0 az Y 34 elemű halmazában, így az M22 kényszerűen minden attribútumot bevont.
* A dupla-attribútumkészletes modellek esetén nincs direkt és inverz értelmezés.
* A csak direkt modellek esetén egyelőre nem készült inverz modell és így validitás vizsgálat sem.
* Az M25 is dupla attribútumkészlettel dolgozó modell.

Kimutatásvarászlással támogatott riportok:

* Minden egyes step(i) modell esetén: a bevont és a be nem vont (inkl. már egyszer felhasznált) X-attribútumok Y-nal kapcsolatos korrelációinak: átlaga, maximuma, minimuma, szórása
* Additív regressziós és step1 modell esetén: a bevont és a be nem vont (inkl. már egyszer felhasznált) X-attribútumok Y-nal kapcsolatos korrelációinak: átlaga, maximuma, minimuma, szórása csak regresszióra, csak step1-re (vö. előző riportok), közös attribútumokra, azonos szerepű attribútumokra
* Becsléseken belüli átlagos határmértékek átlaga step(i) esetben pozitív és negatív korrelációt mutató nyers attribútumok szerinti csoportokra

Diagrammok:

* Polinomizálódási hajlam feltárását támogató ceteris paribus alakzatok grafikus nézete OAM3 esetén
* Modell-korrelációk
	+ Step1-2-3-4-5-6-7-8
	+ OAM6-7-8-9-10-11-12
* Modell-hibák
	+ Step1-2-3-4-5-6-7-8
	+ OAM6-7-8-9-10-11-12

Egyéb elemzések:

* Parciális korreláció-számítások
* Átlagos hatásmérték számítások attribútumonként adott modell esetén

# Eredmények



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

Az 1. ábra értelmezése:

* A már bevont attribútumok kizárása mellett képződő újabb és újabb modellek korrelációja a tények és a becslések között egyre romlik.
* A már bevont attribútumok kizárása mellett képződő újabb és újabb modellek tény vs. becslés eltérése egyre romlik, azaz a modell-hiba egyre nő.
* A legjobbnak választott 27 változó alapján készülő modell (step1) korrelációs értéke (0.966) számos modellel meghaladható: pl.
	+ Step1-2-3-4-5-6-7-8 becslései alapján készülő modellel (OAM1 = 0.979)
	+ OAM3
	+ OAM4
	+ OAM5
	+ OAM6
	+ OAM7
	+ Az OAM8 korrelációja és a benchmark (step1) korreláció quasi azonos (0.966 vs. 0965)
* A legjobbnak választott 27 változó alapján készülő modell (step1) korrelációs értéke (0.966) számos modellel jól közelíthető: pl.
	+ Additív regresszió (22 változó bevonásával): 0.938
	+ Csak step5-6-7-8 becslésekből kiinduló modell (OAM9) esetén (ahol a teoretikus attribútum szám már csak 15+1=16=8+5+2+1) – hiszen step8 eleve semmilyen információ értékkel nem bír, ill. a valós attribútum-szám és a teoretikus (16) attribútum szám nem tér el egymástól kihagyott (mert teljes step5-6-8 modell nem kerül soha kihagyásra, csak egy-egy inverz/direkt réteg: 7d, ill. 8d/i) – korreláció: 0.943



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

A 2. ábra értelmezése:

* A minimum = 0.499-es érték pirossal kiemelve arra utal, hogy elvileg illett volna egy 3. sorcímkét is definiálni, vagyis a már felhasználat attribútumok részhalmazát, de ez nem hat ki a mellékletben látható modell-korreláció-értelmezésre.
* S nem hat ki a step-modellekbe bevont attribútum-csoportok korrelációinak értelmezésére:
	+ A step1 esetén az átlag, max, min korreláció kisebb a bevont attribútumoknál, mint a be nem vontak esetén – ahol a korreláció alatt itt a teljes 34 elemű adatsorra érvényes korrelációk értendők.
	+ A step1 átlagos korrelációs értéke a bevont attribútumokra nézve eleve negatív.
	+ A step1 nem vonta be a maximális korrelációjú változót.
	+ A step1 bevonta a minimális korrelációjú változót.
	+ A step2 és a step1 értelmezése azonos.
	+ A step3-tól indulva az átlagos korreláció a bevont változók esetén már pozitív ugyan, de ennek oka csak a negatívak fokozatos integrálódásában keresendő.
	+ Csak a step6 vonja be a maximális Xi vs. Y korrelációt mutató attribútumot.



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

A 3. ábra értelmezése:

* A step1 (piros) értékei azonosak a 2. ábra step1 rétegének értékeivel.
* Az additív regresszió azonnal bevonja a maximális korrelációt mutató független változót.
* A regressziós aktív változók korrelációjának átlaga pozitív, azaz értelemszerűen magasabb, mint az eleve negatív átlagot felmutató step1 esetén.
* A regressziós modell is von be negatív korrelációjú attribútumot (noha itt nem értelmezhető a parciális korreláció csak indirekt módon – vö. nyers inputértékek nominális kicsinysége hasonlóan hat, mint egyes inputérték-opciók nulla helyettesítési értéke).
* A step1 és a regresszió közös (7 db) attribútumának átlagos korrelációja is negatív.
* A step1 és a regresszió azonos státuszú (59=7+52 db) attribútumának átlagos korrelációja is negatív.



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

A 4. ábra értelmezése:

* A step3 kivételével a becsült értékek mögötti additív módon aggregálandó lépcsőértékek átlagos mértéke (hatásmérték/változó – mértékegység = %) a többi step-modellben a negatív korrelációval rendelkező változók esetén nagyobb (vö. arány > 100%).
* A step3 esetén az arány 93%-os értéke közel van a 100%-os küszöbértékhez.

Az 5. ábra értelmezése:

* A step1 modell 27 aktív attribútumából 16 esetben volt értelmezhető a parciális korreláció fogalma, vagyis az az eset, amikor az X oldalon a lépcsős függvény kapcsán 0 feletti csereérték nem monoton módon szerepel (hiszen a korreláció-számítás nem értelmeződik, ha az egyik vagy mindkét számsor csak azonos értékeket tartalmaz).
* A sárgával jelzett csökkentett input-mennyiségű mátrix alapján kerültek számításra a parciális korrelációk, míg minden adat alapján számítódott a teljes korreláció.
* A parciális korrelációk átlaga pozitív (+0.601), míg a teljes korrelációk átlaga -0.747.
* A 16 érintett attribútum esetén 9 esetben volt növekedés a parciális adatok alapján a teljes adathalmazhoz képest.
* Ezen növekedések aggregált mértéke is pozitív: 1.348
* A 16 érintett attribútum 9 növekvő esete közül 4 esetben a növekedés negatív korrelációból váltott át pozitívba (kék cellák).
* A 16 érintett attribútum 7 csökkenő esete közül 2 esetben a növekedés pozitív korrelációból váltott át negatívba (narancssárga cellák).
* A 16 érintett attribútum közül 6+4 esetben a változás nem váltott át előjelet.
* A teljes korrelációk minimuma (-0.499) a parciális nézetben +0.621 lett, ami megközelíti a teljes korrelációk maximumát (0.689).



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

# Konklúziók

* Az additív regressziók keretében a magasabb korrelációt mutató független változók modellbe kerülési valószínűsége nagyobb, mint az alacsonyabb korrelációt mutató változók modellbe kerülési valószínűsége.
* Ez a logikus elvárás csak látszólag nem igaz a lépcsős függvények esetén, ugyanis a parciális korrelációk itt is átlagosan magasabbak, mint a teljes attribútum-értékkészletre vonatkozó korrelációk. Sőt, a parciális korreláció negatív tartományból átcsaphat pozitívba, de előfordulhat ennek ellentettje is. Így a lépcsős függvények a parciális korrelációkon keresztül alakítanak ki hatásmechanizmusokat a hasonlóságelemzések során – ráadásul mindezen parciális korrelációk akár mind két rangsor-irányítottság esetén párhuzamosan is képesek egymást eltérő parcialitások mellett hatni, mintha egy sejtfalba épülő fehérjének két aktív zónája is lenne a sejtfallal való együttműködés során és soha nem az egész komplex fehérje értelmezendő, csak ennek egy/két aktív zónája, ahol az aktivitás mértéke (mélysége) is eltérő lehet.
* A lényeg tehát a látszatok, hermeneutikai érzékcsalódások feloldásában rejlik: hiszen
	+ igaz, hogy jó építőanyagból (vö. magas korreláció) lehet jó épületet (modellt) építeni
	+ de a jó építőanyag nem akkor jó, ha minden (vö. teljes korreláció) szempontból jó,
	+ hanem már akkor is, ha valamilyen (vö. parciális korreláció) szempontból jó...
	+ ahol a lépcsős függvények adják a parcialitást, a dupla-attribútumkészlet pedig a valamilyenség esélyének megkettőzését…
* A neurális hálók domesztikálása, vagyis hermeneutikáik egyre finomabb hangolása hasonló értelmezési kísérleteket vár el, mely témakörben az anti-diszkriminatív modellezés ceteris paribus alakzatainak tudatos befolyásolásáról szóló kísérletek eredményei hamarosan bemutatásra kerülnek egy önálló cikkben…
* A tudományos vita kapcsán indokolni kell minden paraméter-beállítást (vö. miért pont ez, miért pont ennyi, stb.).
* Ez a cikk a korrelációk (a jóság) modellezését, alakulásuk megértését igyekszik elősegíteni...
* Vagyis a modelljóság tervezése nem eretnekség - vö. genetikai potenciál alapú keresés (vö.

# Jövőkép

A bemutatott vizsgálatok nem használták ki a hasonlóságelemzés elsődleges minőségbiztosítását, a függvényszimmetriát, mely hatása a hermeneutikára masszív is lehet a jövőben.

A COCO STD futtatások mindegyike COCO Y0-ként is értelmezhető, ahol a modellbe be nem vont attribútumnak az számít, ahol az S1 értéke az objektumok számánál egy egységgel kisebb, amennyiben a modellgravitáció = 1 (vagyis két szomszédos lépcső közötti távolság minimum 1 kell, hogy legyen).

Az egyes (best of közeli) hasonlóságelemzési modellek összehasoníthatók a jövőben a kapott lépcsősfüggvények polinomizáltsági mutatói mentén (vö. Occam borotvája elv).

A TDK esetén a T azt jelenti, minél több mindenről tud a szerző logikusan nyilatkozni (vö. önkényes véleményekkel szemben objektivitásra törekvés), ami a PhD-fokozatszerzés kapcsán még inkább releváns): pl.

* milyen modellezési paraméter miért az és miért nem más?
* miért lett annyi egy eredmény, amennyi?
* mi lett volna, ha más lett volna a bemenő jel? (vö. modellezés JÓSÁGÁNAK tervezése/előrejelzése)
* mikor tekinthető valami bizonyítottnak?
* két bizonyítás közül melyik a jobb és miért? (a jobbság bizonyítását bizonyító bizonyítás bizonyosságának bizonyítása - azaz a végtelen komplexitás kezelése)

# Referenciák

…szövegközben…

# Mellékletek



1. Ábra: Az eltérő adatmennyiséggel dolgozó modellek jellemzői (forrás: saját számítások)

A 6. ábra értelmezése:

* A 2. ábra step1-2-3-4-5-6-7-8 modellekbe bevont változóinak korrelációs statisztikái alapján OAM építhető.
* Az OAM képes magyarázni dupla-attribútumkészletű (zöld) esetben a modell-korrelációkat káros polinomizálódás nélkül (vö. részletek – xls).
* A modell-jóság tehát tervezhető, hiszen ezek a modellek szimulátorok, melyek valamilyen pontossággal tetszőleges inputhelyzet várható sikerességét képesek „előrejelezni”.