

Hiányzó adatok kezelése, következményei	Olvasási idő: 30 perc 	Készítette: Dr. habil. Kiss Gábor Dávid 
--	--	---

Hiányzás oka

A többváltozós idősorok elemzése során szinkronizált és folytonos adatokra van szükségünk – azaz minden kereskedési nap esetében valamennyi vizsgált piaci eszköz esetében rendelkezünk kell egy árfolyammal. Mindazonáltal lehetnek olyan speciális alkalmak, ahol egy vagy több adat hiányzik a kereskedési aktivitás hiánya miatt. Ebben az alfejezetben a hiányzó adatok pótlására alkalmazott főbb módszereket hasonlítom össze. A hiányzó adat (vagy hiányzó érték) definíciója az alábbi: „azon adatok összessége, melyek nem találhatók meg a vizsgálatra kiválasztott mintában” (Kang 2013), idősorok esetében az adathiány az adatgeneráló folyamat átmeneti felfüggesztése miatt jön létre (Graham 2012). Az idősorok esetében komoly elvárásokat szokás megfogalmazni: a hiányzó adatok kezelése során sem illik torzítani az ARIMA és GARCH paramétereket, rontani a reprezentativitást vagy az átlag, szórás értékét és autokorreláltság fokát (Juan Carlos et al. 2010, Kang 2013).

A pénzügyi idősorok, különösen a napi záróárfolyamok esetében előfordul adathiány a kereskedési adatokban, hiszen egyes piacok zárva tarthatnak, míg a többi piacon javában zajlik a kereskedés. Ez egy érdekes többváltozós jelenség. A piacokon a kereskedés hiánya egyrészt a nemzeti sajátosságokra vezethető vissza, mint például az ünnepek, hétvégék, másrészt a piaci erők okozta illikvid helyzetekre (ez leggyakrabban a small-cap részvényekkel⁶ fordul elő), továbbá az árak hirtelen összeomlása miatt felfüggesztésre. Óriási irodalma van az árképzést és a piaci hatékonyságot aláásó jelenségeknek, amelyek közül a legtöbbet idézett a “hétvége hatás” (*weekend effect*) (Keim – Stambaugh 1984, Robins – Smith 2015, Shahid – Mehmood 2015).

A szakirodalom három típusát különbözteti meg az adathiánynak (Graham 2012, Junger – Leon 2015, Oravecz 2008):

- teljesen véletlenszerű adathiányról (*missing completely at random*, röviden: MCAR) akkor beszélhetünk, amikor az adathiány nem függ az adatok értékétől vagy más, különösen a megfigyelt változó értékeitől. A kihagyásuk nem torzítja a becslésünket a homogenitásuk miatt (Enders 2010, Junger – Leon 2015, Kang 2013).
- véletlenszerű adathiány (*missing at random*, röviden: MAR) esetén a hiányzás attól a változótól független amely adatsorában felmerül (Kang 2013), de valamilyen mechanizmust feltételezhetünk a hiányzás mögött (Graham 2012). Kihagyásuk károsíthatja az olyan időbeli struktúrákat, mint az autokorreláció, trendek vagy szezonális (Junger – Leon 2015).

⁶ Erről a jelenségről bővebben ír például Giovanni (2005).

- nem véletlenszerű adathiány (*missing no at random*, röviden: MNAR) akkor következik be, ha a hiány nem független az azt tartalmazó változótól (Oravecz 2008), de amikor lehetséges, akkor torzítatlan becslést illeszthetünk a hiányzó adatokra (Graham 2012).

Előfordul, hogy az adathiányt a kutató nem tudja kezelni, mert az eloszlása ismeretlen, a MAR esetén csak feltételezéssel lehet élni (Graham 2012)⁷.

Módszertan

Tegyük fel, hogy a vizsgált n külföldi deviza közül kiválasztott i -dik deviza ($1 \leq i \leq n$) árfolyamából vett P_i minta (17) minden y kereskedési napon v mintamérettel az alábbiak szerint írható le:

$$P_i = \begin{bmatrix} y_1 & p_{i,1} \\ \dots & \dots \\ y_v & p_{i,v} \end{bmatrix}. \quad (17)$$

Emellett kiválasztunk még egy másik k -adik ($1 \leq k \leq n$, és $k \neq i$) devizát is (18) w minta és z ($z \neq y$) időindexszel.

$$P_k = \begin{bmatrix} z_1 & p_{k,1} \\ \dots & \dots \\ z_w & p_{k,w} \end{bmatrix}. \quad (18)$$

A minta mérete rendre megegyezik a kereskedési napok számával. Ha a fentiek mentén leírható $P_{1,...,i,k,...,n}$ mátrixokat egyesíteni kell egy többváltozós elemzés céljából, akkor az időindexek összehangolására (szinkronizációjára) van szükségünk. A következőkben három különféle adathiány kezelésére használható Baraldi et al. (2015) által is leírt gyakorlatot vázolok.

Az első, az adathiányt tartalmazó esetek listászerű vagy páronkénti törlése (*listwise* vagy *pairwise deletion*), amikor eltávolítjuk azokat az intervallumokat, ahol legalább egy hiányzó adatunk van. A törlés az idősor töredezettségét okozhatja vagy torzítottá válhat a paraméter becslése, ezért csak MCAR adathiány esetén alkalmazható (Kang 2013). A listászerű adattörlés (19) esetén kizárunk minden olyan esetet, amikor legalább egy érték hiányzik:

$$T = Y \cap Z. \quad (19)$$

A második, az adathiány átlaggal vagy a rendelkezésre álló historikus mediánnal történő pótlása, amit Junger – Leon (2015) is javasol aszimmetrikus adatok esetén. Ennek a módszernek is hasonló hatása lehet, mint az előzőnek a tőkepiaci logaritmikus hozamokra, melynek időszora zéró átlagú és móduszú⁸. Ezt a megoldást Graham (2012) nem javasolja az MCAR adathiány esetén, az eloszlások magasabb átlag körüli koncentrációja és a standard hibák és variancia alulbecslése miatt (Junger – Leon 2015, Enders 2010). Az átlaggal való pótlás (20) csak akkor alkalmazható, ha a logaritmikus hozamok nulla közeli átlaggal és módusszal rendelkeznek. Az utolsó elvégzett megfigyelés érték behelyettesítése (*Last observation carried forward* – LSCF) módszer alkalmazása szintén erre az eredményre vezet – nulla logaritmikus hozamot produkálva:

$$T = (Y \cup Z) \text{ és } p_{i,0} = p_{i,0-1} \text{ és } T \notin (Y \cap Z). \quad (20)$$

A harmadik és egyben legmodernebb eljárástípusok a hiányzó adatokat megpróbálják rekonstruálni, minimalizálva a hibafüggvényt (Baraldi et al. 2015, Ceylan et al. 2013, Juan Carlos 2010). A várakozás maximalizációs (*expectation maximization*, röviden: EM) modellek közül a leggyakrabban alkalmazott a maximum likelihood becslés, de neurális hálózatokra építő és genetikai struktúrára

⁷ További érdekességek: <https://cxl.com/blog/outliers/>

⁸ Feltételezve, hogy P_t and P_{t-1} árak megegyeznek és logaritmikus hozamuk nulla: $r_t = \log(P_t / P_{t-1}) = (P_t - P_{t-1})/P_{t-1} = 0$.

alapuló megközelítései is elterjedtek (Ceylan et al. 2013, Juan Carlos 2010). A várakozás maximalizációs eljárások hátránya, hogy számítási módja több időt vehet igénybe, mert az EM algoritmus és a likelihood függvény nehezen számítható ki (Ruud 1991) valamint szükség lehet további adatgeneráló modellre is (Horari et al. 2013). Nem okoznak problémát MCAR adathiány esetén, ilyenkor jól használhatók az EM eljárások. A torzítatlan MAR és magasabb hatékonyságú MCAR esetén a legjobb választási lehetőség a maximum likelihood a hiányos multinormális adatok kezelésére, ugyanis ezek az eljárások kevésbé torzítanak, mint az adathiány törlése vagy az átlaggal való pótlás. A maximum likelihood előnye függ a hiányzó adatok arányától, a mintanagyságtól és adatstruktúra kovarianciájától (Wothke 1998). Az általános várakozás maximalizációs algoritmus alapja egy iterált lineáris regressziós elemzés, de ezt helyettesítettük egy gauss eloszlású regressziós paraméter feltételes maximum likelihood becslésével (21) Schneider (2001) tanulmánya⁹ alapján. Egyes $p_{t,i} \in P$ hiányzó értékkel rendelkező mátrix esetén, a hiányzó értékkel (kereskedési napok) rendelkező árak és a rendelkezésre álló értékekkel megadott árak közötti kapcsolatra lineáris regressziós modell írható fel:

$$p_{NaN} = \mu_{NaN} + (p_a - \mu_a)B + \varepsilon \quad (21)$$

Ahol a a meglévő adatot jelenti, $B \in \mathbb{R}^{n_a \times n_{NaN}}$ pedig a regressziós együtthatók mátrixa a hiányzó és meglévő értékek kovariancia mátrixával, az n számú összes mintából. Az $\varepsilon \in \mathbb{R}^{1 \times n_{NaN}}$ reziduumból feltételezzük, hogy nulla-átlagú, és hogy $C \in \mathbb{R}^{n_{NaN} \times n_{NaN}}$ egy ismeretlen kovariancia mátrix vektor. Az EM algoritmus iterációiban a $\mu \in \mathbb{R}^{1 \times n}$ átlagát és a $\Sigma \in \mathbb{R}^{n \times n}$ kovariancia mátrix becslését adottnak tekintjük, és ezek becsléséből számoljuk ki a B mátrix regressziós együtthatóira és a C kovariancia mátrix hibatagjaira vonatkozó feltételes maximum likelihood becsléseket minden hiányzó értéket tartalmazó bejegyzésre – hogy az algoritmus minden hiányzó érték helyére imputált értéket helyettesítse be a teljes μ vektor és Σ mátrix újraszámítását megelőzően.

Matlabos alkalmazás

```
% 1. Hiányzó adatot (NaN) tartalmazó sorok törlése
T=size(raw)
for i=1:length(raw)
    if sum(raw(i,:)>0)==T(1,2)
        good_raw(i,:)=i;
        good_dates(i,1)=raw(i,T(1,2));
    else
        good_raw(i,:)=0;
        good_dates(i,2)=raw(i,T(1,2));
    end
end
nice_raw=nonzeros(good_raw);
for i=1:length(nice_raw)
    Excluded(i,:)=raw(nice_raw(i,1),:);
end
% 2. Nulla körüli várható érték biztosítása
x=10^-6;
Zeroer=raw;
for j=1:T(1,2)
    for i=1:length(raw)
        if isnan(Zeroer(i,j))==1
            Zeroer(i,j)=Zeroer(i-1,j)+x;
        end
    end
end
% 3. Regularized Expectation Maximization (EM) eljárás
for i=1:length(raw)
```

⁹ A számoláshoz használt Matlab-kód is ezen alapul, amely letölthető: <http://climate-dynamics.org/software/#regem>

```
for j=1:T(1,2)
    if raw(i,j)==0
        raw_nan(i,j)=NaN;
    else
        raw_nan(i,j)=raw(i,j);
    end
end
end
cd 'c:\documents\matlab\Imputation'
[RegEM, M, C, Xerr, B, peff, kavlr, kmisr, iptn] = regem(raw_nan);
```

Források

- Sávai, Marianna ; Kiss, Gábor Dávid (2016): Pénzügyi idősorok hiányzó adatainak kezelése – afrikai devizaárfolyamok példáján. *STATISZTIKAI SZEMLE* 94 : 7 pp. 736-756. , 21 p.
- Baraldi, P., Di Maio, F., Genini, D., & Zio, E. (2015). Reconstruction of Missing Data in Multidimensional Time Series by Fuzzy Similarity. *Applied Soft Computing Journal*, 26, 1–9. <http://dx.doi.org/10.1016/j.asoc.2014.09.038>
- Cappeiello, L., Engle, R. F., & Sheppard, K., (2006). Asymmetric Dynamics in the Correlations of Global Equity and Bond Returns. *Journal of Financial Econometrics*, 4 (4), 537–572. <http://dx.doi.org/10.1093/jffinec/nbl005>
- Figueroa García, J. C., Kalenatic, D., & López Bello, C. A. (2010). An Evolutionary Approach for Imputing Missing Data in Time Series. *Journal of Circuits, Systems & Computers*, 19(1), 107–121. <http://dx.doi.org/10.1142/S0218126610006050>
- Graham, J. W. (2012). *Missing Data Analysis and Design*. New York: Springer
- Houari, R., Bounceur, A., Kechadi, T., & Euler, R. (2013). *A New Method for Estimation of Missing Data Based on Sampling Methods for Data Mining*. CCSEIT http://dx.doi.org/10.1007/978-3-319-00951-3_9
- Junger, W.L., & Ponce de Leon, A. (2015). Imputation of Missing Data in Time Series for Air Pollutants. *Atmospheric Environment*, 102, 96–104. <http://dx.doi.org/10.1016/j.atmosenv.2014.11.049>
- Kang, H. (2013). The Prevention and Handling of the Missing Data. *Korean J Anesthesiol*, 64(5), 402–406 <http://dx.doi.org/10.4097/kjae.2013.64.5.402>
- Keim, D. B., & Stambaugh, R. F. (1984). A Further Investigation of the Weekend Effect in Stock Returns. *The Journal of Finance*, 39(3), 819–835. <http://dx.doi.org/10.1111/j.1540-6261.1984.tb03675.x>
- Robins, R. P., Smith, G. P. (2016). No More Weekend Effect. *Critical Finance Review*, 5(2), Forthcoming.
- Ruud, P. A. (1991). Extensions of Estimation Methods Using the EM Algorithm. *Journal of Econometrics*, 49(3), 305–341. [http://dx.doi.org/10.1016/0304-4076\(91\)90001-T](http://dx.doi.org/10.1016/0304-4076(91)90001-T)
- Schneider, T. (2001). Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *Journal of Climate*, 14, 853–871. [http://dx.doi.org/10.1175/1520-0442\(2001\)014%3C0853:AOICDE%3E2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(2001)014%3C0853:AOICDE%3E2.0.CO;2)
- Zafar, N., Urooj, S. F., Chughtai, S., Amjad, S. (2012). Calendar anomalies: Case of Karachi Stock Exchange. *African Journal of Business Management*, 6 (24), 7261-7271. <http://dx.doi.org/10.5897/AJBM11.1847>
- Wothke, W. (1998). *Longitudinal and Multi-group Modelling with Missing Data*. Mahwah. NJ: Lawrence Erlbaum Associates.

- Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz I. (2013). Comparison of Missing Value Imputation Methods in Time Series: the Case of Turkish Meteorological Data. *Theoretical & Applied Climatology*, 112(1/2), 143–167. <http://dx.doi.org/10.1007/s00704-012-0723-x>

Önellenőrző kérdések

1. Milyen torzításokat eredményezhetnek a hiányzó adatok egy idősor esetében?
2. Milyen koncepcióink vannak a hiányzás megfogalmazására?
3. Milyen kezelési eljárásokat ismerünk?
4. Milyen torzításokat eredményezhet a hiányzó adatok kezelése negyedéves illetve napi idősoron?

Jelen tananyag a Szegedi Tudományegyetemen készült az Európai Unió támogatásával. Projekt azonosító: EFOP-3.4.3-16-2016-00014

