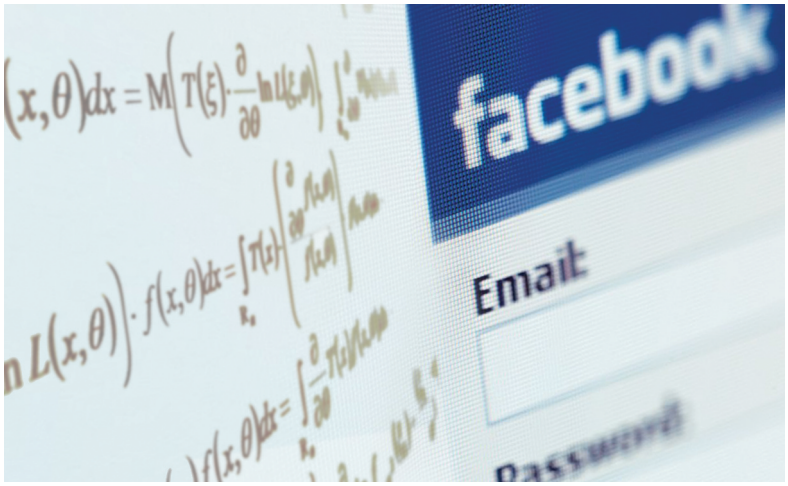
Neki jednostavni
nelinearni regresijski modeli (1. dio)

Kristina Matijević, Požega
Bojan Kovačić, Zagreb



U nastavi opisne (deskriptivne) statistike u srednjim školama i veleučilištima, poglavito na ekonomskim stručnim studijima, prilikom obrade korelacijske i regresijske analize standardno se obrađuje model jednostavne linearne regresije kojim se nastoji protumačiti linearna veza između dviju pojava, od kojih prva predstavlja nezavisnu varijablu (x), a druga zavisnu varijablu (y). No, odnos zavisne i nezavisne varijable često nije linearan jer ravnomjerne promjene nezavisne varijable ne uzrokuju ravnomjerne promjene zavisne varijable i obratno. U takvim je slučajevima potrebno pronaći neku drugu (nelinearnu) funkciju koja najbolje opisuje odnos promatranih varijabli.

U ovom ćemo članku razmotriti neke jednostavne modele jednostruke nelinearne regresije u kojima se primjenjuju eksponencijalna funkcija, logaritamska funkcija i opća potencija. Ti se modeli odgovarajućim zamjenama varijabli uvijek mogu svesti na model jednostavne linearne regresije, ali to ovdje nećemo učiniti jer želimo ukazati kako se svojstva navedenih triju vrsta funkcija mogu primijeniti u regresijskoj analizi.

Sve primjere rješavat ćemo koristeći MS Excel iz dvaju temeljnih razloga: osnove rada s ovim programom poznate su apsolutnoj većini srednjoškolaca i studenata, a dodatno želimo staviti naglasak na analizu i svojstva parametara modela koje ćemo razmatrati, kao i izbjeći linearizaciju svakoga pojedinoga modela nužnu za "klasične" izračune. Za detalje o "klasičnom" radu s razmatranim modelima čitatelje upućujemo na [2], a za detalje o primjeni MS Excela na rješavanje statističkih zadataka upućujemo na [4].

Napomenimo da u svakom regresijskom modelu dogovorno koristimo uobičajene oznake:

— $x =$ empirijski određena (izmjerena) vrijednost nezavisne varijable;
— $y =$ emipirijski određena (izmjerena) vrijednost zavisne varijable;
— $\hat{y} =$ procijenjena (očekivana) vrijednost zavisne varijable izračunana pomoću jednadžbe regresijskoga modela;

— $\hat{x}$ = procijenjena (očekivana) vrijednost nezavisne varijable izračunana korištenjem jednadžbe regresijskoga modela;

1. Model jednostavne eksponencijalne regresije

Standardni oblik regresijske jednadžbe modela jednostavne eksponencijalne regresije je:

$$\hat{y} = b \cdot a^x, \tag{1}$$

pri čemu su $a,\ b \in \mathbf{R}^+$ realne konstante[1] takve da je $a \neq 1$. Zbog tih pretpostavki vrijedi nejednakost $\hat{y} > 0$. Ona potpuno odgovara "praktičnim potrebama" jer se u praksi obično promatraju veze među pojavama čije su numeričke vrijednosti strogo pozitivni realni brojevi.

Primijetimo da se za $a = 1$ dobije konstantna funkcija $\hat{y} = b$. Ta je funkcija trivijalan slučaj modela jednostavne linearne regresije i vrlo se rijetko pojavljuje u promatranju veze među ekonomskim ili društvenim pojavama.

Konstante a i b nazivaju se *osnovni parametri* modela jednostavne eksponencijalne regresije.

Radi što bolje analize regresijskoga modela posebno navodimo statističke interpretacije tih parametara.

Iz jednakosti (1) i uvjeta $a,\ b \in \mathbf{R}^+,\ a \neq 1$, lako slijedi

Tvrdnja 1.

$$\hat{y} = b \text{ ako i samo ako je } x = 0. \tag{2}$$

Stoga se parametar b interpretira kao *očekivana vrijednost varijable y za $x = 0$*. Napomenimo da ova interpretacija nerijetko ne odgovara realnoj situaciji. Npr. u slučaju da promatramo zavisnost ukupnih izdvajanja za hranu (varijabla y) o ukupnim mjesečnim primanjima (varijabla x) neke obitelji, nelogično je izabrati $x = 0$, tj. pretpostaviti da obitelj nema nikakvih mjesečnih primanja.

Statistička interpretacija parametra a iskazuje se posredno. Točnije, vrijedi:

Tvrdnja 2. Ako se vrijednost varijable x promijeni za k jedinica mjere te varijable, pri čemu je $k \in \mathbf{R}$, onda će se očekivana vrijednost varijable y promijeniti za prosječno

$$s_k = 100 \cdot (a^k - 1)\%. \tag{3}$$

Pritom predznak realnoga broja k označava povećanje/smanjenje vrijednosti varijable x, a predznak parametra s_k označava povećanje/smanjenje očekivane vrijednosti varijable y.

Dokaz. Neka je x_1 početna vrijednost varijable x. Prema jednakosti (1), očekivana vrijednost varijable y za $x = x_1$ jednaka je:

$$\hat{y}_1 = b \cdot a^{x_1}. \tag{4}$$

Ako se vrijednost x_1 promijeni za k jedinica mjere varijable x, nova vrijednost varijable x je:

$$x_2 = x_1 + k. \tag{5}$$

Pritom za $k > 0$ imamo povećanje, a za $k < 0$ smanjenje vrijednosti varijable x. Očekivana vrijednost varijable y za $x = x_2$ jednaka je:

$$\begin{aligned} \hat{y}_2 &= b \cdot a^{x_2} = b \cdot a^{x_1 + k} = b \cdot a^{x_1} \cdot a^k \\ &= a^k \cdot (b \cdot a^{x_1}) = a^k \cdot \hat{y}_1. \end{aligned} \tag{6}$$

Stoga je pripadna relativna promjena očekivane vrijednosti varijable y jednaka:

$$\begin{aligned} (\Delta\hat{y})_k &= \frac{\hat{y}_2 - \hat{y}_1}{\hat{y}_1} = \frac{a^k \cdot \hat{y}_1 - \hat{y}_1}{\hat{y}_1} \\ &= \frac{\hat{y}_1 \cdot (a^k - 1)}{\hat{y}_1} = a^k - 1. \end{aligned} \tag{7}$$

Iskažemo li tu promjenu u postocima, dobit ćemo:

$$s_k = \left[100 \cdot (a^k - 1)\right] \%, \tag{8}$$

što je i valjalo pokazati. ■

[1] Standardno je $\mathbf{R}^+ := \langle 0, +\infty \rangle$ skup svih strogo pozitivnih realnih brojeva.

Napomena 1. U iskazu Tvrdnje 1 upotrijebljen je prilog *prosječno* jer se u izvodu formule za izračun vrijednosti parametra a pomoću emipirijskih vrijednosti varijabli x i y koriste prosjeci tih empirijskih vrijednosti, pa je i sâm parametar a pokazatelj prosječne promjene. Detalji se mogu naći u [2].

Napomena 2. Statistički pokazatelj

$$s_1 = 100 \cdot (a - 1) \tag{9}$$

uobičajeno se naziva *prosječna relativna stopa promjene* (očekivane) vrijednosti varijable y. Pomoću te stope najlakše se posredno interpretira parametar a. Njezina "jedinica mjere" je postotak $[\%]$. Ako model jednostavne eksponencijalne regresije kvalitetno opisuje vezu između varijabli x i y, onda približno jednaka prosječna relativna stopa promjene vrijedi i za empirijske vrijednosti varijable y.

Radi potpunosti, zapišimo posebno dvije posljedice Tvrdnje 1 koje koriste činjenicu da eksponencijalna funkcija $f(x) = a^x$ strogo pada za $a \in \langle 0, 1\rangle$, a strogo raste za $a > 1$.

Tvrdnja 3. Ako se vrijednost varijable x *poveća* za k jedinica mjere te varijable, onda će se vrijednost varijable y očekivano *smanjiti* za $s_k\%$ ako i samo ako je $a \in \langle 0, 1\rangle$, a očekivano *povećati* za $s_k\%$ ako i samo ako je $a > 1$. ■

Tvrdnja 4. Ako se vrijednost varijable x *smanji* za k jedinica mjere te varijable, onda će se vrijednost varijable y očekivano *povećati* za $s_k\%$ ako i samo ako je $a \in \langle 0, 1\rangle$, a očekivano *smanjiti* za $s_k\%$ ako i samo ako je $a > 1$. ■

Napomena 3. U svim gornjim tvrdnjama ne smijemo izostaviti riječ *očekivano*. Time naglašavamo da su sve izračunane vrijednosti rezultat naših procjena, a ne npr. istraživanja ili pokusa.

Budući da osnovne parametre regresijskoga modela interpretiramo kao stanovite procijenjene vrijednosti, opravdano se postavlja pitanje: jesu li te naše procjene dovoljno dobre, odnosno koliko dobro neki regresijski model opisuje vezu među promatranim pojavama?

Odgovor na ovo pitanje daje statistički pokazatelj koji se naziva *koeficijent determinacije* i označava s R^2. Koeficijent determinacije je uvijek neki realan broj iz segmenta $[0, 1]$. On se najčešće interpretira posredno pomoću pokazatelja R_1 definiranog formulom

$$R_1 := 100 \cdot R^2. \tag{10}$$

Vrijednost R_1 interpretira se kao *postotak veze među promatranim pojavama koji se može objasniti regresijskim modelom*. Što je R_1 "bliže" $100\ [\%]$, model je bolji i reprezentativniji. Detalje ovdje izostavljamo, a mogu se naći u [1], [2] ili [3].

Ilustrirajmo primjenu ovoga regresijskog modela na primjeru.

Primjer 1. Na uzorku[2] od 15 slučajno odabranih učenika srednje škole "Mirko S. Zlikovski" iz Špičkovine[3] ispituje se veza između prosječnoga dnevnoga vremena provedenoga na društvenoj mreži Facebook i prosjeka brojeva ostvarenih bodova na svim ispitima iz matematike tijekom jednoga polugodišta. Pritom nam je dodatno poznat podatak da je aritmetička sredina najvećih mogućih brojeva bodova na tim ispitima jednaka točno 20. Empirijski podaci su dani u tablici na sljedećoj stranici.

a) Vezu promatranih varijabli grafički prikažite dijagramom rasipanja. Uz grafikon navedite sve potrebne oznake.

b) Odredite jednadžbu modela jednostavne eksponencijalne regresije koji najbolje opisuje zavisnost prosjeka brojeva bodova na ispitima iz matematike o prosječnom dnevnom vremenu provedenu na Facebooku i objasnite značenje parametara dobivenoga modela.

c) Na temelju koeficijenta determinacije procijenite reprezentativnost dobivenoga modela.

[2] Pitanjem reprezentativnosti odabranoga uzorka ovdje se ne bavimo.
[3] Svi nazivi ustanova su potpuno izmišljeni. Korišteni toponimi su stvarni i predstavljaju mjesta u Republici Hrvatskoj. Svi empirijski podaci su također izmišljeni za potrebe ovoga članka i ne predstavljaju rezultat bilo kakvoga znanstvenoga ili stručnoga istraživanja.

Na temelju rezultata **b)** podzadatka *procijenite*:

d) smjer i iznos prosječne relativne promjene prosjeka brojeva bodova na ispitima iz matematike ako se prosječno dnevno vrijeme provedeno na Facebooku poveća za dva sata;

e) prosjek brojeva bodova na ispitima iz matematike učenika koji dnevno na Facebooku provede prosječno 3 sata i 15 minuta;

f) koliko najviše vremena dnevno na Facebooku smije provesti učenik koji želi imati pozitivnu ocjenu iz matematike ako je "prag za prolaz" 10 bodova;

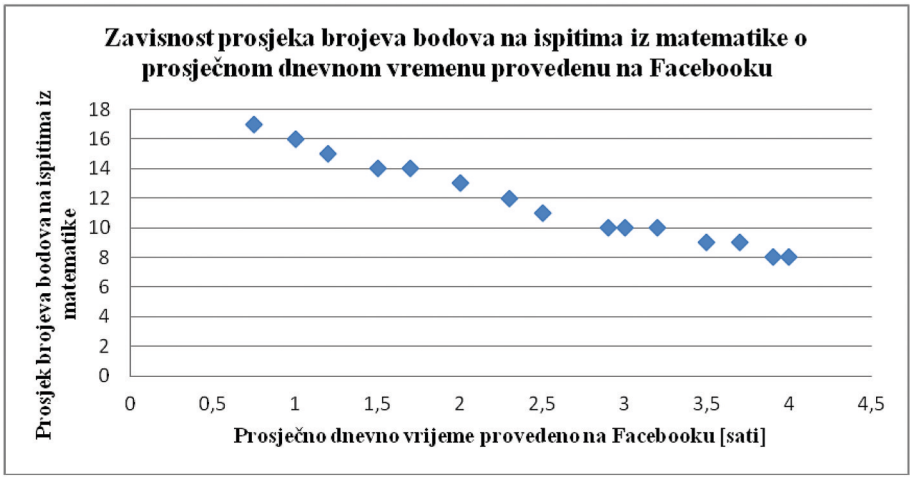
g) prosječno dnevno vrijeme koje na Facebooku provede učenik koji je na ispitima iz matematike ostvario prosječno 2 boda, pa utvrdite realističnost dobivenoga rezultata.

Rješenje primjera 1.

a) Veza promatranih varijabli prikazana je dijagramom rasipanja na slici 1. Dijagram rasipanja tvore točke čije su koordinate uređeni parovi (x, y) dobiveni iz tablice 1. Ucrtane točke *ne* spajamo krivuljom[4] jer pretpostavljamo da veza među promatranim pojavama nije funkcijska, nego stohastička (slučajna). Detalji o konstrukciji dijagrama rasipanja pomoću MS Excela mogu se naći u [4].

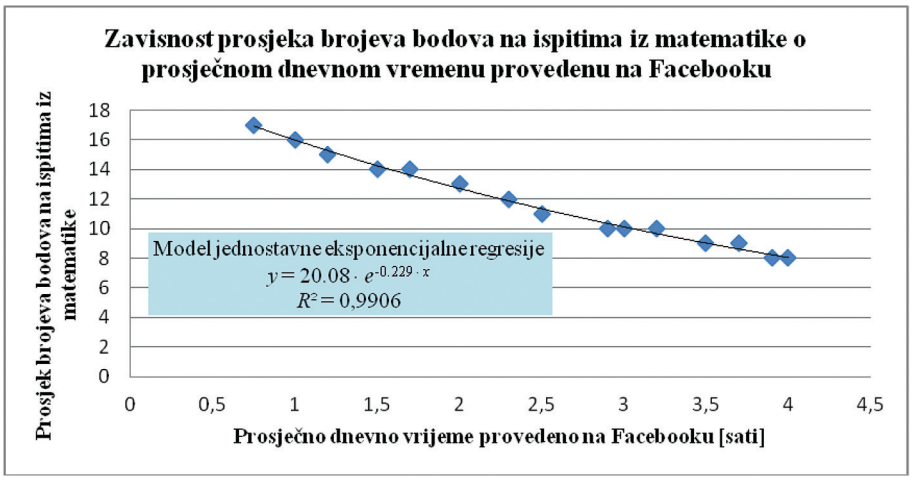
prosječno dnevno vrijeme provedeno na Facebooku [sati]	prosjek brojeva bodova na ispitima iz matematike
0.75	17
1	16
1.2	15
1.5	14
1.7	14
2	13
2.3	12
2.5	11
2.9	10
3	10
3.2	10
3.5	9
3.7	9
3.9	8
4	8

Tablica 1. Empirijski podaci za Primjer 1



Slika 1. Dijagram rasipanja za primjer 1

[4] Čitatelj upoznat s polinomima može zaključiti da postoji barem jedan polinom stupnja najviše 14 (npr. Lagrangeov interpolacijski polinom) čiji graf prolazi svim navedenim točkama.



Slika 2. Model jednostavne eksponencijalne regresije za primjer 1 i pripadna regresijska krivulja

b) Jednadžba modela jednostavne eksponencijalne regresije, pripadna regresijska krivulja i koeficijent determinacije modela prikazani su na slici 2 dobivenoj također pomoću MS Excela. Detalji o postupku kojim se dobiju ti pokazatelji mogu se naći u [4].

Dobivenu regresijsku jednadžbu najprije zapišimo u standardnom obliku $\hat{y} = b \cdot a^x$. U ovome je slučaju $b = 20.08$, dok je vrijednost parametra a jednaka

$$a = e^{-0.229} \approx 0.795329. \qquad (11)$$

Stoga jednadžba dobivenoga modela jednostavne eksponencijalne regresije glasi:

$$\hat{y} = 20.09 \cdot 0.795329^x. \qquad (12)$$

Interpretirajmo parametre dobivenoga modela.

Prema tvrdnji 1, $b = 20.08$ je vrijednost varijable y za $x = 0$. To znači da će učenik koji ne provodi vrijeme na Facebooku (nego vrijeme provodi radeći druge aktivnosti) očekivano postići približno 20.08 bodova po jednom ispitu iz matematike (praktički, najveći mogući broj bodova na pojedinom ispitu).

Prema napomeni 2, parametar $a = 0.795329$ interpretiramo posredno računajući očekivanu prosječnu relativnu stopu promjene prosjeka brojeva bodova iz matematike:

$$s_1 = 100 \cdot (a - 1) = -20.4671. \qquad (13)$$

Dakle, *ako se prosječno dnevno vrijeme provedeno na Facebooku poveća za* 1 *sat, prosjek brojeva bodova na ispitima iz matematike smanjit će se za približno* 20.4671%.

c) Budući da je $R^2 = 0.9906$, prema napomeni 3 zaključujemo da se 99.06% *zavisnosti prosjeka brojeva bodova na ispitima iz matematike o prosječnom dnevnom vremenu provedenu na Facebooku može objasniti modelom jednostavne eksponencijalne regresije*. Ta je vrijednost vrlo blizu 100%, pa je naš model reprezentativan.

d) U ovakvom se zadatku najčešće *pogrešno* zaključuje: *Znamo da ako se prosječno dnevno vrijeme provedeno na Facebooku poveća za* 1 *sat, onda će se prosjek brojeva bodova na ispitima iz matematike očekivano smanjiti za približno* 20.4671%. *Stoga ako se prosječno dnevno vrijeme provedeno na Facebooku poveća za* 2 *sata, onda će se prosjek brojeva bodova na ispitima iz matematike očekivano smanjiti za prosječno* $2 \cdot 20.4671 = 40.9342\%$. Pogreška je u tome što za eksponencijalnu funkciju ne vrijedi upravna razmjernost varijabli, pa npr. dvostruki rast varijable x ne povlači dvostruki rast varijable y i obrnuto.

Podzadatak treba riješiti koristeći jednakost (3). U tu jednakost uvrstimo $k = 2$ i $a = 0.795329$, pa dobijemo:

$$s_2 = 100 \cdot (0.795329^2 - 1) \approx -36.7453\%. \quad (14)$$

Dakle, *prosjek brojeva bodova na ispitu iz matematike će se očekivano smanjiti za prosječno* 36.7453%.

e) Traženi prosjek brojeva bodova procijenit ćemo tako da u jednakost (12) uvrstimo $x = 3.25$ jer je 3 sata i 15 minuta = 3.25 sati. Tako dobijemo:

$$\hat{y} = 20.09 \cdot 0.795329^{3.25} \approx 9.54 \text{ bodova}. \quad (15)$$

f) Treba naći najveću vrijednost varijable x za koju je $\hat{y} \geq 10$. Tako dobivamo eksponencijalnu nejednadžbu

$$20.09 \cdot 0.795329^x \geq 10, \quad (16)$$

iz koje logaritmiranjem (npr. po bazi 10) lagano slijedi $x \leq 3.046$. (Zaokruživanjem vrijednosti varijable x naviše smanjuje se vrijednost varijable y, pa zato moramo zaokruživati naniže.) Stoga traženo vrijeme iznosi 3.046 sati, odnosno (približno) 3 sata i 2 minute.

g) Iz regresijske jednadžbe (12) izrazimo varijablu x. Logaritmiranjem po bazi 10 dobije se:

$$\hat{x} = \frac{\log y - \log 20.09}{\log 0.795329}. \quad (17)$$

U tu jednakost uvrstimo $y = 2$, pa slijedi:

$$\hat{x} = \frac{\log 2 - \log 20.09}{\log 0.795329} \approx 10.07459 \text{ sati}$$
$$\approx 10 \text{ sati i } 4 \text{ minute}. \quad (18)$$

Utvrdimo je li ovakva procjena realistična, tj. provodi li zaista učenik koji postiže vrlo slabe rezultate na ispitima iz matematike toliko vremena dnevno na Facebooku. Iako volimo reći da je "u današnje doba sve moguće", ovaj podzadatak izvrsno ukazuje na tzv. *problem ekstrapolacije* u regresijskim modelima. Okvirno govoreći, regresijski model može davati dobre procjene za vrijednosti varijable x koje se nalaze unutar segmenta omeđenoga najmanjom i najvećom empirijskom vrijednošću te varijable. Međutim, za vrijednosti izvan toga segmenta kvaliteta procjene se smanjuje što je vrijednost varijable x udaljenija od donje ili gornje granice segmenta. Zbog toga kad god je to moguće, treba izbjegavati procjenu vrijednosti varijable y za vrijednosti varijable x koje se nalaze izvan spomenutoga segmenta i obratno.[5] Stoga u ovom zadatku relativno kvalitetne procjene dobivene pomoću regresijske jednadžbe (12) možemo očekivati za $x \in [0.75, 4]$, odnosno, ekvivalentno, pomoću jednadžbe (17) za $y \in [8, 17]$.

LITERATURA

1/ B. Kovačić: *Poslovna statistika*, interna skripta, Visoka poslovna škola PAR, Rijeka, 2012.

2/ I. Šošić, V. Serdar: *Uvod u statistiku*, Školska knjiga, Zagreb, 2000.

3/ I. Šošić: *Primijenjena statistika*, Školska knjiga, Zagreb, 2004.

4/ M. Papić: *Primijenjena statistika u MS Excelu*, Naklada ZORO, Zagreb, 2012.

5/ M. Vukičević, M. Papić: *Matematičko–statistički priručnik za poduzetnike*, Golden-marketing, Zagreb, 2003.

[5] Iz navedenoga modela, međutim, slijedi iskustveno i logički istinita činjenica: što je vrijednost varijable x veća, to je vrijednost varijable y bliža nuli. Iako već za $x \approx 43.5$ slijedi $y \approx 0$, taj podatak nije moguće interpretirati jer prosječan dnevni broj sati provedenih na Facebooku ne može biti strogo veći od 24. Stoga valja biti oprezan i s odabirom vrijednosti varijable x tako da *interpretacija* vrijednosti te varijable ima smisla.