© Wydawnictwo UR 2016 „Edukacja – Technika – Informatyka” nr 1/15/2016
ISSN 2080-9069 www.eti.rzeszow.pl
ISSN 2450-9221 online DOI: 10.15584/eti.2016.1.27

**ŠTEFAN GUBO**

# Riešenie úloh nelineárnej regresie pomocou tabuľkového kalkulátora

# Solution of nonlinear regression tasks using spredsheet application

RNDr. PhD., Katedra matematiky a informatiky, Ekonomická fakulta, Univerzita J. Selyeho, Slovenska Republika

**Abstrakt**

Nelineárna regresia je druh regresnej analýzy, kde skúmané údaje sú modelované funkciou, ktorá je nelineárnou kombináciou parametrov modelu a je závislá od jeden alebo viacero nezávislých premenných. V príspevku uvádzame riešenie úloh nelineárnej regresie v tabuľkovom kalkulátore MS Excel 2013.

**Kľúčové slová:** nelineárna regresia, tabuľkový kalkulátor, MS Excel.

**Abstract**

Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. In this paper we illustrate how to use the MS Excel 2013 spreadsheet application in solving non-linear regression tasks.

**Key words:** non-linear regression, spreadsheet application, MS Excel.

## Úvod

Predpokladajme, že v experimente chceme zistiť, ako závislá premenná $y$ závisí od nezávislej premennej $x$. Počas merania sme tieto veličiny odmerali $n$-krát s empirickými údajmi $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$. Vyjadrenie hodnoty závislej premennej zo známych hodnôt nezávislých premenných sa nazýva *regresia*. Regresia určuje tvar štatistickej závislosti. Ak závislú premennú vyjadríme pomocou lineárneho vzťahu, hovoríme o *lineárnej regresie*. Ak závislosť nie je lineárna, jej priebeh sa vyjadrí vhodnou nelineárnou (polynomickou, logaritmickou, exponenciálnou, atď.) regresnou funkciou. V tomto prípade ide o *nelineárnej regresie*.

**Nelineárna regresia**

Koeficienty nelineárnej regresnej funkcie je možné určiť priamo pomocou metódy najmenších štvorcov, alebo nepriamo použitím transformácie na lineárnu regresnú funkciu.

Metóda najmenších štvorcov vychádza z požiadavky minimalizovania súčtu štvorcov rozdielov regresných chýb (rozdiel medzi empirickými a teoretickými hodnotami závislej premennej *y*). Túto optimalizačnú úlohu zapíšeme v nasledovnom matematickom tvare:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 w_i \quad \rightarrow \text{MIN},$$

kde

$y_i$ je empirická hodnota *i*-tej závislej premennej,
$\hat{y}_i$ je teoretická hodnota *i*-tej závislej premennej,
$\varepsilon_i$ je hodnota *i*-tej regresnej chyby,
$w_i$ určuje váhu *i*-teho merania, obvykle uvažujeme konštantnú váhu ($w_i = 1$).

Rovnica regresnej krivky je teda vypočítaná tak, aby súčet štvorcov vertikálnych vzdialeností jednotlivých bodov $(x_i, y_i)$ z výberového súboru od nej bol minimálny [Klučka 2009].

**Riešenie úlohy nelineárnej regresie**

**Úloha:** Konateľ firmy pravidelne sleduje týždenné tržby a vynaložené náklady na reklamu. Keďže predpokladá závislosť medzi týmito ukazovateľmi, chce odhadnúť funkciu modelujúcu túto závislosť. Zistené týždenné údaje sú uvedené v tabuľke:

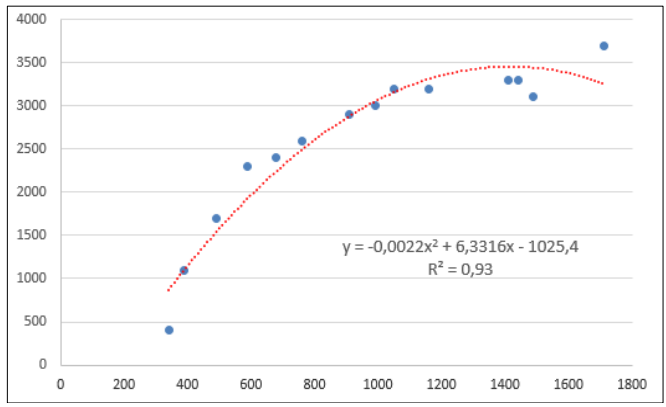
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **tržby** (EUR) | 1100 | 1700 | 2600 | 2400 | 2300 | 2900 | 400 |
| **náklady** (EUR) | 390 | 490 | 760 | 680 | 590 | 910 | 340 |
| **tržby** (EUR) | 3200 | 3300 | 3100 | 3200 | 3000 | 3700 | 3300 |
| **náklady** (EUR) | 1160 | 1410 | 1490 | 1050 | 990 | 1710 | 1440 |

a) modelujte priebeh závislosti týždenných tržieb od nákladov na reklamu kvadratickou (parabola druhého stupňa), exponenciálnou a logaritmickou regresnou funkciou,
b) rozhodnite, ktorý z týchto modelov najlepšie charakterizuje priebeh závislosti tržby od nákladov na reklamu!

**Riešenie 1**:

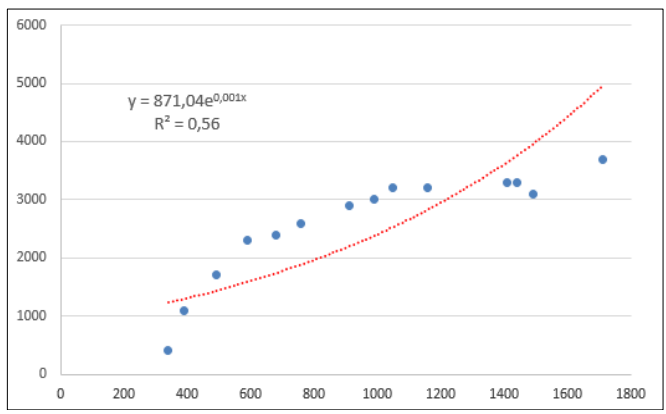
**a)** Spustíme tabuľkový kalkulátor MS Excel 2013, a postupne vložíme hodnoty závislých (tržby – stĺpec *Y*, oblasť **B2:B15**) a nezávislých premenných

(náklady na reklamu – stĺpec $X$, oblasť **C2:C15**). Údaje zobrazíme pomocou bodového grafu, kde na ktoromkoľvek bode klikneme pravým tlačidlom myši a v ponuke vyberieme **Pridať trendovú spojnicu**. V objavenom dialógovom okne najprv zvolíme **polynomickú regresiu s poradím 2** a zaškrtneme *Zobrazovať v grafe rovnicu* a *Zobraziť v grafe rovnicu spoľahlivosti* $R^2$. Na grafe (obrázok 1) sa objaví rovnica regresnej paraboly ($y = -0{,}0022x^2 + 6{,}3316x - 1025{,}4$) a koeficient determinácie ($R^2 = 0{,}93$), pomocou ktorého možno posúdiť ako dobre regresná krivka vysvetľuje variabilitu údajov. Jeho hodnota udáva, že 93,0% vzrastu týždennej tržby je závislý od vzrastu nákladov na reklamu.



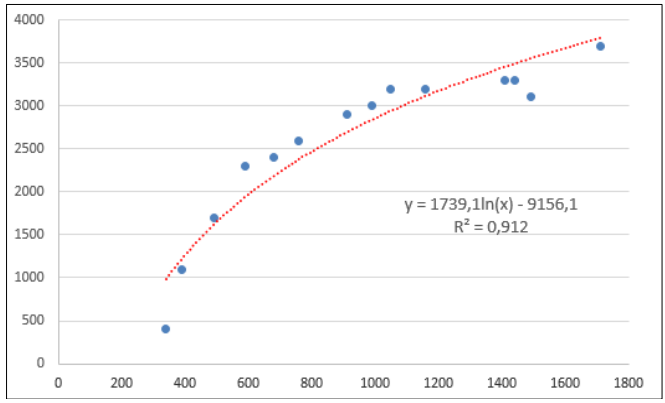

**Obrázok 1. Graf a rovnica regresnej kvadratickej krivky**

V druhom kroku si vytvoríme ďalší bodový graf, kde pridáme exponenciálnu trendovú spojnicu (obrázok 2). Jej rovnica je $y = 871{,}04\ e^{0{,}001x}$; $R^2 = 0{,}56$.




**Obrázok 2. Graf a rovnica regresnej exponenciálnej krivky**

Uvedený postup opakujeme ešte raz a vytvoríme si tretí bodový graf s logaritmickou trendovou spojnicou, na ktorom zobrazujeme rovnicu regresnej logaritmickej krivky (obrázok 3). Jej rovnica je $y = 1739{,}1 \ln x - 9156{,}1$; $R^2 = 0{,}9128$).




**Obrázok 3. Graf a rovnica regresnej logaritmickej krivky**

**b)** Na základe dosiahnutých výsledkov môžeme konštatovať, že z uvedených troch nelineárnych regresných modelov priebeh závislosti tržby od nákladov na reklamu najlepšie charakterizuje kvadratická regresná funkcia.

**Riešenie 2**: Predchádzajúce riešenie má nevýhodu, že pomocou neho nevieme otestovať nulové hypotézy o vhodnosti regresného modelu a o významnosti regresných koeficientov. Aby sme tieto testy mohli urobiť, je potrebné nelineárnu regresnú funkciu transformovať na lineárnu regresnú funkciu zavadením funkčných vzťahov medzi regresnými koeficientmi. Pokladáme za dôležité zdôrazniť, že nie všetky nelineárne funkcie je možné prepočítať, len tie ktoré sú lineárne v koeficientoch.

1) *Kvadratickú regresnú funkciu* zapíšeme v tvare

$$y = a_2 x^2 + a_1 x + a_0,$$

kde $a_0$, $a_1$, $a_2$ sú regresné koeficienty. Po substitúcii $u = x^2$ dostaneme lineárnu regresnú funkciu

$$y = a_2 u + a_1 x + a_0.$$

2) *Exponenciálnu regresnú funkciu* zapíšeme v tvare

$$y = a_0 e^{a_1 x},$$

kde $a_0$, $a_1$ sú regresné koeficienty. Rovnicu na oboch stranách linearizujeme prirodzenými logaritmami a po úpravách dostaneme rovnicu

$$\ln y = \ln a_0 + a_1 x.$$

Po substitúcii $u = \ln y$, dostaneme lineárnu regresnú funkciu

$$u = \ln a_0 + a_1 x.$$

Upozorňujeme, že lineárna regresná funkcia má v tomto prípade nový regresný koeficient, a je potrebné zo zisteného koeficienta $a = \ln a_0$ transformovanej regresnej funkcie spätne prepočítať odhad pôvodného koeficienta ($a_0 = e^a$).

3) *Logaritmickú regresnú funkciu* zapíšeme v tvare

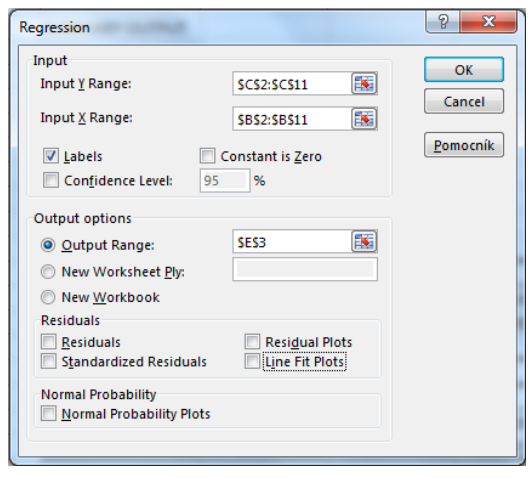
$$y = a_1 \ln x + a_0,$$

kde $a_0$, $a_1$ sú regresné koeficienty. Po substitúcii $u = \ln x$ dostaneme lineárnu regresnú funkciu

$$y = a_1 u + a_0.$$

**a)** Spustíme tabuľkový kalkulátor MS Excel 2013, a vložíme hodnoty závislých (tržby – stĺpec $Y$, oblasť **B2:B15**) a nezávislých premenných (náklady na reklamu – stĺpec $X$, oblasť **C2:C15**). Potom dopočítame stĺpec hodnôt $X^2$ (oblasť **D2:D15**, ktorú budeme potrebovať pre výpočet rovnice paraboly), stĺpec hodnôt $\ln Y$ (oblasť **E2:E15**, ktorú budeme potrebovať pre výpočet rovnice exponenciály) a stĺpec hodnôt $\ln X$ (oblasť **F2:F15**, ktorú budeme potrebovať pre výpočet rovnice logaritmickej krivky).

Regresnú analýzu vo všetkých troch prípadoch realizujeme prostredníctvom voľby **Údaje – Data Analysis**, a v objavenom dialógovom okne zvolíme Regresiu (*Regression*). Po stlačení tlačidla OK sa dostaneme do ďalšieho dialógového okna (obrrazok 4), v ktorom sa definujú vstupné údaje.




**Obrázok 4. Dialógové okno Regression**

1) **Kvadratická regresná funkcia.** Do políčka *Input Y Range* zadávame oblasť závislej premennej *Y* (**B2:B15**) a do políčka *Input X Range* oblasť nezávislej premennej *X* a dopočítaný stĺpec hodnôt $X^2$ (**C2:D15**). Ak údaje zadávame aj s názvami premenných, označíme checkbox *Labels* (Popisky). V tomto okne máme možnosť meniť hladinu spoľahlivosti (*Confidence Level*), MS Excel 2013 štandardne ponúka 95%. Ďalej je dôležité označiť výstupnú oblasť (*Output Range*) a graf regresnej priamky (*Line Fit Plots*). Spracovanie potvrdíme tlačidlom OK a dostávame nasledovný výstup (obrázok 5):

| Parabola | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | | | | |
| | | | | | | | | |
| *Regression Statistics* | | | | | | | | |
| Multiple R | 0.964820282 | | | | | | | |
| R Square | 0.930878176 | | | | | | | |
| Adjusted R Square | 0.918310571 | | | | | | | |
| Standard Error | 268.0834635 | | | | | | | |
| Observations | 14 | | | | | | | |
| | | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| Regression | 2 | 10646586.68 | 5323293 | 74.06966 | 4.14843E-07 | | | |
| Residual | 11 | 790556.1777 | 71868.74 | | | | | |
| Total | 13 | 11437142.86 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -1025.398289 | 404.8880573 | -2.53255 | 0.027847 | -1916.550894 | -134.245683 | -1916.55089 | -134.245683 |
| X Variable 1 | 6.331594933 | 0.91067062 | 6.952673 | 2.41E-05 | 4.327222413 | 8.335967452 | 4.327222413 | 8.335967452 |
| X Variable 2 | -0.002236677 | 0.00044943 | -4.9767 | 0.000418 | -0.003225866 | -0.00124749 | -0.00322587 | -0.00124749 |

**Obrázok 5. Výstup regresnej analýzy 1**

Výstup regresnej analýzy sa skladá z troch častí. V prvej sú výsledky korelačnej analýzy. Hodnota korelačného koeficientu *R* (*Multiple R*) je 0,96, teda sa jedná o vysoký stupeň tesnosti vzťahu medzi týždennými tržbami a nákladmi na reklamu. Hodnota $R^2$ (*R Square*) je hodnota koeficientu determinácie. Upravený koeficient determinácie (*AdjustedR Square*) zohľadňuje aj počet meraní a počet odhadovaných parametrov. Chyba strednej hodnoty (*Standard Error*) je štandardná chyba odhadu regresnej priamky. V poslednom riadku tabuľky je uvedený rozsah súboru (*Observations*).

V časti analýza rozptylu (ANOVA) sa testuje nulová hypotéza (navrhnutý regresný model nie je štatisticky významný) oproti alternatívnej hypotéze (navrhnutý regresný model je štatisticky významný). Na vyhodnotenie tohto tvrdenia slúži F test. Keďže významnosť F (*Significance F*) je v tomto prípade

0,000 < 0,05, testovanú nulovú hypotézu zamietame, čo znamená, že navrhnutý regresný model je vhodný.

V tretej časti výstupu sú uvedené hodnoty koeficientov regresnej funkcie ($a_0 = -1025{,}4$, $a_1 = 6{,}3316$ a $a_2 = -0{,}0022$) a testujú sa nulové hypotézy o významnosti týchto koeficientov, pričom nulová hypotéza tvrdí nevýznamnosť príslušného koeficienta a alternatívna hypotéza jeho významnosť. Keďže pre všetky prípady je hodnota $P < 0{,}05$, testované nulové hypotézy o významnosti regresných koeficientoch zamietame. V posledných dvoch stĺpcoch sú uvedené hranice 95%-ných intervalov spoľahlivosti pre jednotlivé koeficienty.

2) **Exponenciálna regresná funkcia.** Do políčka *Input Y Range* zadávame oblasť dopočítaného stĺpca hodnôt ln $Y$ (**E2:E15**) a do políčka *Input X Range* oblasť nezávislej premennej $X$ (**C2:C15**). Vyhodnotením nulových hypotéz o vhodnosti tohto modelu a významnosti regresných koeficientov sme zistili, že model je vhodný a regresné koeficienty sú štatisticky významné (obrázok 6). Z hodnoty zisteného koeficienta $a = 6{,}7696$ transformovanej regresnej funkcie odhad pôvodného koeficienta prepočítame nasledovne: $a_0 = e^a = 871{,}0355$).

**Exponenciálna krivka**

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.748492818 |
| R Square | 0.560241499 |
| Adjusted R Square | 0.523594957 |
| Standard Error | 0.411920677 |
| Observations | 14 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 2.593996964 | 2.593997 | 15.2877 | 0.002072438 |
| Residual | 12 | 2.036143731 | 0.169679 | | |
| Total | 13 | 4.630140695 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 6.769682723 | 0.272145846 | 24.8752 | 1.08E-11 | 6.176727863 | 7.362637583 | 6.176727863 | 7.362637583 |
| ln(Y) | 0.00101594 | 0.000259835 | 3.909949 | 0.002072 | 0.000449809 | 0.001582071 | 0.000449809 | 0.001582071 |

**Obrázok 6. Výstup regresnej analýzy 2**

3) **Logaritmická regresná funkcia.** Do políčka *Input Y Range* vložíme adresu oblasti závislej premennej $Y$ (**B2:B15**) a do políčka *Input X Range* adresu oblasti dopočítaných hodnôt ln $X$ (**F2:F15**). Na základe výstupu (obrázok 7) môžeme skonštatovať, že aj tento model je vhodný a regresné koeficienty sú štatisticky významné.

| Logaritmická krivka | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | | | | |
| | | | | | | | | |
| *Regression Statistics* | | | | | | | | |
| Multiple R | 0.955429522 | | | | | | | |
| R Square | 0.912845571 | | | | | | | |
| Adjusted R Square | 0.905582702 | | | | | | | |
| Standard Error | 288.2125455 | | | | | | | |
| Observations | 14 | | | | | | | |
| | | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| Regression | 1 | 10440345.2 | 10440345 | 125.6866 | 1.02785E-07 | | | |
| Residual | 12 | 996797.6569 | 83066.47 | | | | | |
| Total | 13 | 11437142.86 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -9156.085757 | 1050.174619 | -8.71863 | 1.54E-06 | -11444.21969 | -6867.95182 | -11444.2197 | -6867.95182 |
| ln(X) | 1739.094835 | 155.1239009 | 11.21101 | 1.03E-07 | 1401.10889 | 2077.080781 | 1401.10889 | 2077.080781 |

**Obrázok 7. Výstup regresnej analýzy 3**

**b)** Výber najvhodnejšieho modelu urobíme porovnaním hodnôt koeficientov determinácie $R^2$. Táto hodnota je najvyššia v prípade kvadratickej regresnej funkcie, preto parabola druhého stupňa je najvhodnejšia na vysvetlenie závislosti medzi týždennými tržbami a nákladov na reklamu.

## Záver

Na základe vyššie spomenutého môžeme skonštatovať, že tabuľkový kalkulátor MS Excel 2013 je vhodným nástrojom na riešenie úloh nelineárnej regresie bez používania hlbších poznatkov z matematickej štatistiky.

## Literatúra

Klučka J. (2009), *Plánovanie a prognostika v aplikáciách*, Žilina.

Ragsdale C.T. (2012), *Spreadsheet Modeling & Decision Analysis*, Mason, OH.

**Poďakovanie**

Tento príspevok vznikol s podporou KEGA Ministerstva školstva, vedy, výskumu a športu SR pre projekt č. 010UJS-4/2014.

**Reviewed by:** Doc. RNDr. Edita Partová, CSc.