ФОРМИРАНЕ НА КОМПЕТЕНЦИИ ЗА ОТКРИВАНЕ НА ЗАВИСИМОСТИ ЧРЕЗ РЕГРЕСИОНЕН АНАЛИЗ

Веска Нончева, Джемиле Сюлейман

Пловдивски университет "Паисий Хилендарски", ул. Цар Асен 24, 4000 Пловдив
wesnon@uni-plovdiv.bg

***Резюме**: Тази статия представя практика за изграждане на компетенции за визуализиране на информацията и търсене на зависимости, извлечени от данните.*

***Ключови думи**: линеен регресионен анализ, стохастично моделиране*

1. Въведение

Компетенциите са съвкупност от знания, умения и поведение за постигане на резултати. Компетенциите са способности, но не вродени способности, а такива, които са развити чрез натрупване на знания и формиране на умения. Компетенциите не могат да бъдат постигнати завинаги, защото те отразяват динамиката на средата.

Съвременното образование е ориентирано към усвояване на компетенции чрез използване на нови научни методи и информационни ресурси [1].

Една от задачите на статистическия анализ е да получи информация за зависимости от данните. Методите на регресионния анализ предоставят съвременни средства за откриване на знания, носени от данните. Но един предварителен познавателен процес е необходим за да можем да виждаме информацията представена чрез графиките на регресионния анализ.

Целта на настоящата статия е да подпомогне формирането на компетенции за получаване на информация от данните като представи процеса на построяване на адекватен регресионен модел. Притежаването на такива компетенции е добра основа за пълноценна професионална реализация.

2. Линеен регресионен анализ

Математическият модел на регресионния анализ е следният:

$Y= f(X_1, X_2, \ldots, X_k, \varepsilon)$,

където Y е зависимата променлива, $X_1, X_2, \ldots, X_k$ са обясняващи променливи, ε е грешката. На езика на регресионния анализ независимите променливи $X_1, X_2, \ldots, X_k$ се наричат предиктори, а зависимата променлива Y се нарича отклик. Грешките от наблюденията са независими и еднакво разпределени гаусови случайни величини с нулево очакване. В случая на линеен регресионен модел функцията f е линейна [4].

3. Изследване на зависимостти между външни белези на насекоми с R

Малки бръмбари, подобни на бълхи (flea beetles), се срещат в Съединените щати. Те са вид малки насекоми, които често се срещат в градините в началото на вегетационния период на растенията.

Фигура 1. Бръмбар бълха

Тези малки насекоми се хранят с голямо разнообразие от растения, включително боб, зеле, царевица, патладжан, картофи, чушки, домати, маруля, и повечето разсад. Някои видове от тези малки насекоми могат да повредят растенията. Когато популациите от бръмбари са големи, те могат бързо да обезлистят и дори да убият цели растения. Тези малки насекоми могат да предават вирусни и бактериални заболявания.

Възрастните са малки - 2-3 мм по дължина, лъскави, тъмно кафяви или черни насекоми с големи задни крака, които им позволяват да скачат от растенията, когато са обезпокоени (виж Фигура 1).

Ще изследваме тези насекоми. За тази цел е направена случайна извадка от 74 екземпляра [2]. Измерени са шест характеристики, описани в Таблица 1.

Таблица 1. Описание на измерените характеристики на насекомите

Променлива	Обяснение
tars1	ширина на първата става на първите тазови крайници
tars2	ширина на втората връзка на тазовите крайници
head	максималната ширина на главата измерена между външните ръбове на очите
aede1	максималната ширина на aedeagus (размножителен орган на насекомото) в предната част
aede2	предния ъгъл на aedeagus (размножителен орган на насекомото)
aede3	ширината от страни на aedeagus (размножителен орган на насекомото)

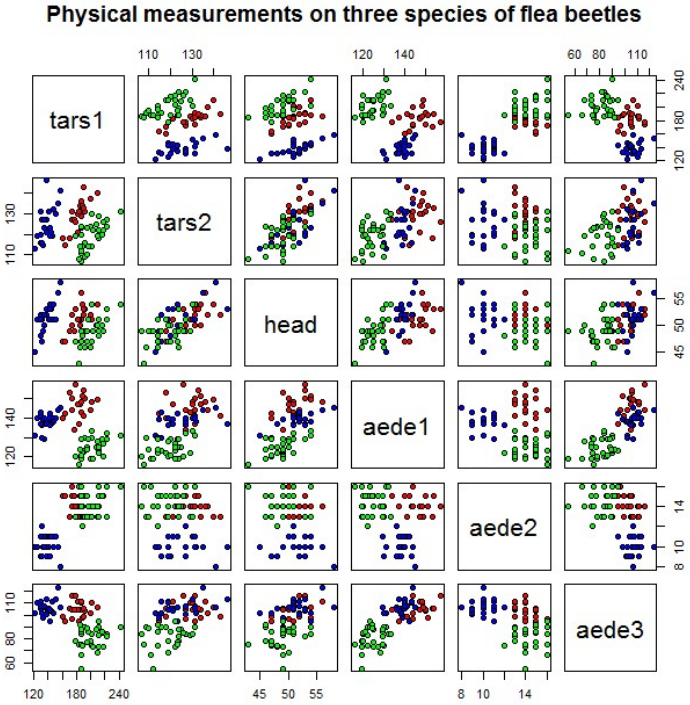
Нашата цел е да получим информация за зависимости между външните белези на насекомите, скрита в данните. За постигане на целта ще използваме линеен регресионен анализ. Като средство за анализа ще използваме функции на средата за анализ на данни R ([3]).

Данните са записани във файл „flea_beetles.csv". Ще прочетем всичките данни:

> flea <- read.csv ("C:\\Users\\userl\\Desktop\\flea_beetles.csv")

Начертаваме графика на всичките променливи с функцията pairs:

> pairs (flea[2:7], main = "Physical measurements on three species of flea beetles", pch = 21, bg = c("red", "green", "blue")[unclass (flea$species)])



Фигура 2. Визуализиране на данните за бръмбари бълхи

Графиката от *Фигура 2* показва линейни зависимости на променливата tars2 от другите променливи. Затова ще построим линеен модел за tars2, в който като предиктори участват всички останали променливи.

```
> y <- flea$tars2; x1 <- flea$tars1; x2 <- flea$head; x3 <- flea$aede1;
> x4 <- flea$aede2; x5 <- flea$aede3;
```

С функцията lm() ще намерим оценки на коефициентите на модела.

```
> lm (y~x1+x2+x3+x4+x5)
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
Coefficients:
(Intercept)          x1          x2          x3          x4          x5
    7.9733      0.1058      1.2957      0.1469     -0.3564      0.1762
```

С функцията summary (lm (y~x1+x2+x3+x4+x5)) ще намерим оценките на коефициентите на модела и ще проверим дали са значими при ниво на значимост α = 0.05.

```
> summary (lm (y~x1+x2+x3+x4+x5))
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
Residuals:
    Min      1Q  Median      3Q     Max
-11.5908  -3.7362   0.2842   3.7439  18.3762
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.97333   15.62098   0.510 0.611407
x1           0.10581    0.04655   2.273 0.026203 *
x2           1.29573    0.36646   3.536 0.000737 ***
x3           0.14693    0.12066   1.218 0.227537
x4          -0.35639    0.59306  -0.601 0.549884
x5           0.17625    0.09373   1.880 0.064351 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.898 on 68 degrees of freedom
Multiple R-squared:  0.5495,    Adjusted R-squared:  0.5164
F-statistic: 16.59 on 5 and 68 DF,  p-value: 1.112e-10
```

Извод: Само коефициентите пред предикторите x_1 и x_2 са значими. Коефициентите пред останалите променливи и свободният член не са значими защото *p*-стойността Pr(>|t|) е по-голяма от 0.05.

Затова ще построим нов модел за y (tars2 - ширината на втората връзка на тазовите крайници), в който участват само предикторите x1 (tars1 - ширина на първата става на първите тазови крайници) и x2 (head - максималната ширина на главата измерена между външните ръбове на очите).

```
> lm (y~x1+x2-1)
Call:
lm(formula = y ~ x1 + x2 - 1)
Coefficients:
   x1     x2
1.8742  0.2192
```

Получихме следния модел:

$y = 1.8742\ x_1 + 0.2192\ x_2$.

Ще проверим хипотезата $H_0 : \beta_i = 0$ срещу $H_1 : \beta_i \neq 0$, i = 1,2.

С функцията summary (lm (y ~ модел)) проверяваме хипотезите при ниво на значимост $\alpha = 0.05$.

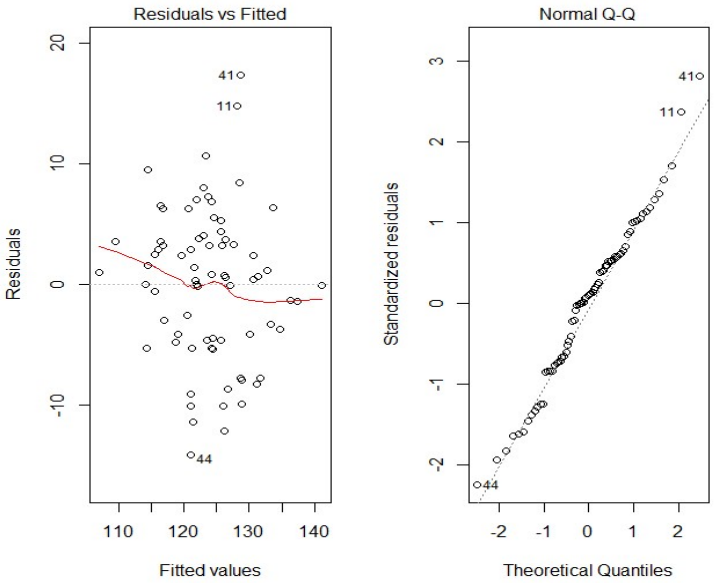
```
> summary (lm (y~x1+x2-1))
Call:
lm(formula = y ~ x1 + x2 - 1)
Residuals:
    Min      1Q  Median      3Q     Max
-12.8016  -4.0604   0.2627   4.0343  16.6394
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x1  1.87417    0.22828   8.210 6.22e-12 ***
x2  0.21919    0.08514   2.575   0.0121 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.129 on 72 degrees of freedom
Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9976
F-statistic: 1.517e+04 on 2 and 72 DF,  p-value: < 2.2e-16
```

Основната хипотеза $H_0 : \beta_i = 0$ се отхвърля в полза на алтернативната хипотеза H_1: $\beta_i \neq 0$, i = 1,2 при което се прави извод, че коефициентите са значими.

Правим две общоприети проверки за адекватност на модела, които се реализират със следните две графики: графика на остатъците и нормална графика.

> op <- par (mfrow = c(1, 2)) ; plot (lm (y~x1+x2-1), which=1);

> plot (lm (y~x1+x2-1), which=2); par(op)



Фигура 3. Визуализиране на остатъците

От получената графика виждаме, че остатъците се визуализират като облак от случайно разпръснати точки. В нормалната графика почти всички остатъци лежат на диагоналната права. Има само три рязко отклоняващи се наблюдения. Изводът от графичното представяне на остатъците, който в този случай може да направим, е: Предположенията на модела са изпълнени.

Коефициентът на детерминация (Adjusted R-squared) е 0.9976, което показва, че 99.76% от всичките данни, които сме използвали, могат да се опишат с този модел. Построеният модел е добър и може да се използва за прогнозиране.

Извод: Намерихме адекватен линеен регресионен модел на зависимостта и той има вида $y = 1.8742\ x_1 + 0.2192\ x_2$, където y е ширината на втората връзка на тазовите крайници, x_1 е ширина на първата става на първите тазови

крайници, x_2 е максималната ширина на главата, измерена между външните ръбове на очите.

Заключение

Методите на регресионния анализ успешно се използват за извличане на информация за зависимости от данните и за прогнозиране. Средата за анализ на данни *R* ни предоставя удобни средства за постигане на тези цели.

Благодарности

Авторите считат за свой приятен дълг да отбележат благодарността си към Фонд "Научни изследвания" при ПУ "Паисий Хилендарски" за финансовата подкрепа при реализацията на проект ФП17-ФМИ-008.

Литература

1. Кендеров П., Чехларова Т., Сендова Е. ЕВРОПЕЙСКИЯТ ПРОЕКТ KeyCoMath И ОРИЕНТИРАНОТО КЪМ УСВОЯВАНЕ НА КЛЮЧОВИТЕ КОМПЕТЕНТНОСТИ ОБРАЗОВАНИЕ ПО МАТЕМАТИКА, МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2015.
2. Cook Dianne, Deborah F. Swayne. Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and Ggobi, 2007.
3. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016, https://www.R-project.org
4. Seber George A. F., Alan J. Lee. Linear Regression Analysis, 2nd Edition, 2003.

MATHEMATICAL COMPETENCE FOR DETECTION OF DEPENDENCIES BY REGRESSION ANALYSIS

Veska Noncheva, Djemile Syuleiman

Abstract: *This paper attempts to outline teaching practice for building mathematical competency for data-driven modeling.*