

APPLICATION OF THE REGRESSION ANALYSIS TO THE WORK WAGE

Assoc. prof. Trufka Dimitrova, PhD

Department of Mathematics and informatics

Shumen University "Konstantin Preslavski

Tel.: "+359 892 603018

E-mail: trfktdt@abv.bg

Desislava Ivanova, BSc Student

Department of Mathematics and informatics

Shumen University "Konstantin Preslavski

Tel.: "+359 892 603018

E-mail: nikolaeva_desi@abv.bg

Abstract: *A multidimensional regression analysis has been considered, based on an online survey of 83 people living and working in Bulgaria, a model with three independent variables and cut out highly distinctive values. The aim is to show the dependence of a dependent variable, such as the average monthly salary, on the independent variables: Years of training / self-study, Years of service, Level of cognitive abilities. The programming language R, was used in the grouping and in constructing the graphical images of the data as well as in realizing the regression analysis and evaluating the model parameters.*

Keywords: *Regression, Salary, Internship, Education, Cognitive abilities.*

ВЪВЕДЕНИЕ

Техниките на регресионен анализ в тази статия се прилагат към данните получени от онлайн анкетиране на 83 случайно избрани лица работещи в България. Основната цел е да се определи кой от факторите "Години трудов стаж"(W), "Години образование и самообразование"(S) и "Ниво на когнитивни способности"(C) оказва най-силно, статистически значимо влияние върху средната брутна работна заплата (Earn) и при какви условия.

ИЗЛОЖЕНИЕ

Модел с три независими променливи и изрязани силно отличаващи се стойности

За да се реализира регресионния анализ и да се оценят параметрите на модела се използва средата и възможностите на програмния език R. Нормално разпределение на грешката на модела се осигурява като се изрязват екстремалните аутлайри. Използвайки възможностите на R намираме, че първият квартил е 637,5 лв., т. е. 25% от анкетираните имат средна месечна работна заплата до 637,5 лв. Вторият и третият квартил са съответно 825 лв. и 1000 лв. Т.е. 25% от наблюдаваните имат средна месечна работна заплата в интервала от 637,5 до 825 лв. Други 25% от наблюдаваните имат средна месечна работна заплата в интервала от 825 до 1000 лв. и последните 25% от анкетираните имат средна месечна работна заплата над 1000 лв.

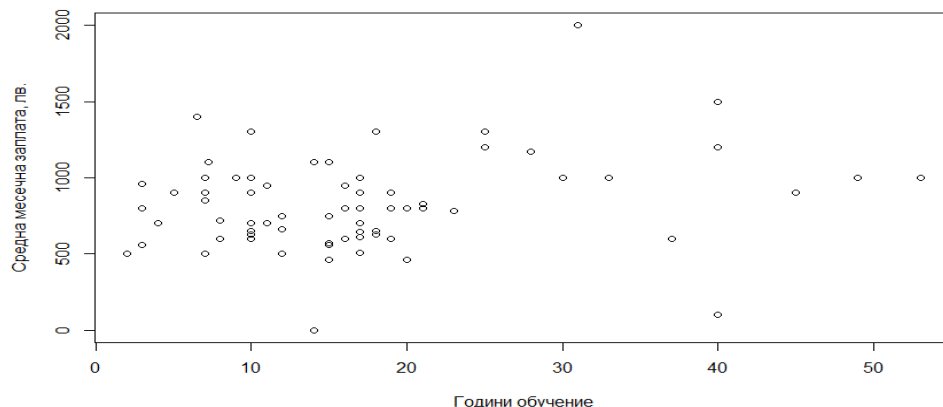
От изразите

$$R = 1000 + 3(1000 - 637.5) = 2087.5 \quad (1)$$

$$L = 637.5 - 3(1000 - 637.5) = -450 \quad (2)$$

се определят границите, извън които се намират екстремалните аутлайъри. Това показва, че ако се анализират данните в интервала $[0; 2087,5]$ има по-голям шанс да се получи регресионен модел с нормална грешка. За да има нормален модел е необходимо да се изключат всички лица със средна брутна работна заплата над 2087,5 лева. Като се обединят данните за всички други лица се получава двумерното съвместно емпирично разпределение между S - годините,

изкарани в обучение и самообучение и Earn - средната месечна работна заплата в лева. Получената графика е дадена на Фигура 1.



Фиг. 1. Двумерно разпределение на анкетираните според годините обучение и самообучение и средната месечна работна заплата за последната година

От фиг. 1 личи, че три от наблюденията биха могли да нарушат хомоскедастичността на модела. Това са хората с най-висока и най-ниска заплата, но те не оказват влияние за нормалността на тествания модел.

Проверяваме хипотезите, чрез командата `MyModel=lm(df$Earn ~ df$S+df$C+df$W)` в R и използваме функцията `summary`, чиито изход е

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	696.96956	178.69039	3.900	0.000214 ***
df\$S	6.38993	3.14416	2.032	0.045811 *
df\$C	0.28930	2.02432	0.143	0.886759
df\$W	0.06611	3.45915	0.019	0.984806

Multiple R-squared: 0.05482, Adjusted R-squared: 0.01544

F-statistic: 1.392 on 3 and 72 DF, p-value: 0.2522

Видно е, че статистически значими са само свободният член и коефициентът пред S. Т.е. за лицата със заплата под 2087,5 лв. за средната месечна работна заплата статистически най-важна променлива измежду тестваните е S, т. е. годините, изкарани в образование и самообразование. За да се установи, чрез прилагане на съставения модел, влиянието върху размера на работната заплата на годините обучение/самообучение, годините трудов стаж и когнитивните способности (измерени в проценти, според броя правилни отговори в направената анкета) се построява уравнението на регресия. Коефициентите в него са такива, че равнината

$$\text{Earn} = 696,9695 + 6,3899*S + 0,2893*C + 0,06611*W \quad (3)$$

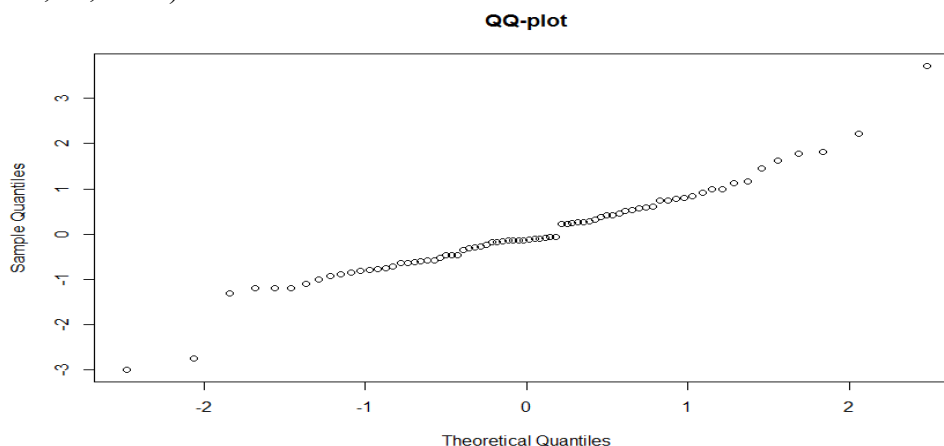
минимизира сумите от квадратите на грешките ε в регресионния модел

$$\text{EarnRM} = 696,9695 + 6,3899*SRM + 0,2893*CRM + 0,06611*WRM + \varepsilon \quad (4)$$

Според този модел за всеки допълнителен процент от резултатите от когнитивния тест за памет очакваната заплата се увеличава с 0,2893 лв. Този, коефициент, обаче не е статистически значим. Статистически значим за тази група е резултата, че при увеличаване броя на годините прекарани в обучение (самообучение) с една, очакваната заплата се увеличава средно с 6,3899 лв. Според направеното наблюдение, отново не е статистически значимо твърдението, че при увеличаване на годините трудов стаж с една, очакваната заплата на лицата със средна месечна работна заплата под 2087,5 лв. се увеличава с 0,0661 лв. Т.е. в тази група отново това не е определяща независима променлива за резултативната величина.

Използваме графичната техника qq-plot за да построим доверителни интервали на оценките на зависимата променлива (фиг. 2.). С изключение на три наблюдения, точките от тази графика са близо до ъглополовящата на I и III квадрант, т.е. можем да приемем, че

разпределението на грешките е нормално. По-точно и при наличие на повече наблюдения тази проверка може да бъде направена с χ^2 -критериите на Пирсън или с тестовите на Колмогоров-Смирнов (Hansen, B., 2019).



Фиг. 2. Нормална qq-plot на остатъците от Модел 1

Построяваме регресионен модел само на зависимостта на средната работната заплата от годините обучение (самообучение), след което чрез сравнение свойствата на грешките ε на моделите и техните стандартни стохастични грешки и коефициенти на детерминация се определя кой модел е по-добър. Функциите `lm` и `summary` в R връщат следния резултат:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 717.409    62.102  11.552 <2e-16 ***
df$S         6.405     3.100   2.066  0.0423 *
```

F-statistic: 4.269 on 1 and 74 DF, p-value: 0.04231

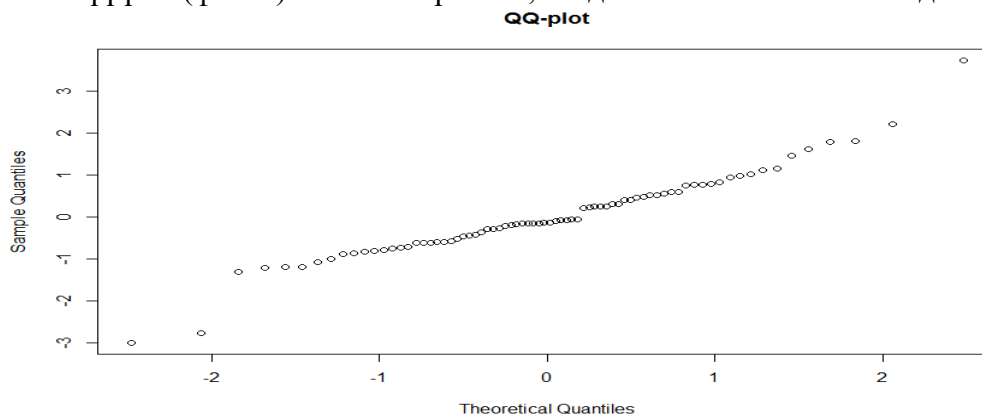
Той показва, че макар и със сравнително голяма стойност на p-value коефициентът пред S е статистически значим и според данните, с които разполагаме се получава следното регресионно уравнение

$$\text{Earn} = 717.409 + 6.405 \cdot S + \varepsilon \quad (5)$$

и съответно регресионният модел е

$$\text{Earn} = \alpha + \beta S + \varepsilon \quad (6)$$

За да се провери дали грешките ε в този модел са нормално разпределени отново се изчертава тяхната qq-plot (фиг. 3). По този признак, Модел 2 не отстъпва на Модел 1.



Фиг. 3. Нормална qq-plot на остатъците от Модел 2
Модел с 3 независими променливи, който не е с нормална грешка.

Разглежда се същия пример, но с включени всички анкетирани, т.е. и тези, които надвишават 2087, 5 лева заплата.

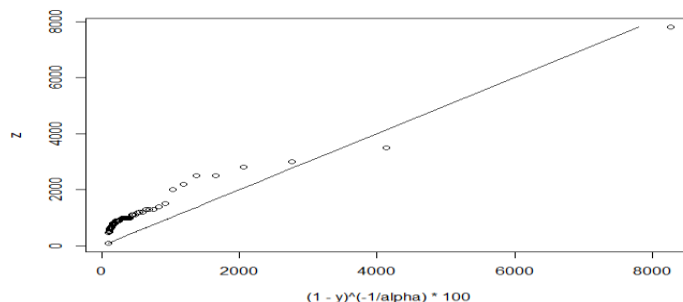
Равнината, която минимизира сумите от квадратите на грешките има уравнение:

$$\text{Earn} = 974,188 - 8,886 \cdot S + 2,882 \cdot C + 1,255 \cdot W \quad (7)$$

Този път обаче, коефициентите в това уравнение не са статистически значими.

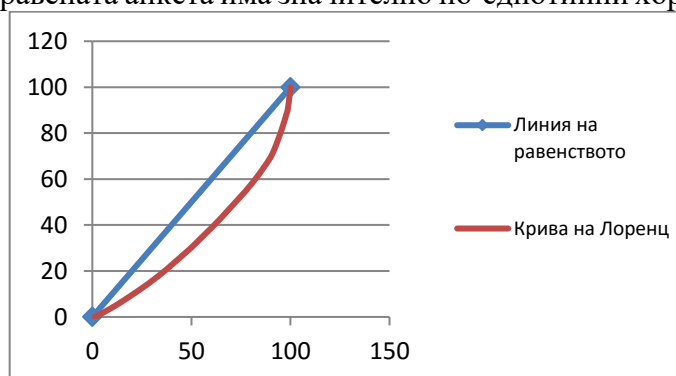
Като се приложи критерият на Пирсън се показва, че заплата не се различава статистически значимо от разпределението на Парето.

Използвайки възможностите на R се изчертава qq-plot, която съдържа точките с координати $(F_{Pareto}^{-1}(\frac{i}{n+1}), x_{(i)})$, където $x_{(1)} \leq \dots, \leq x_{(n)}$, са подредените по-големина наблюдения.



Фиг. 4. Парето qq-plot на доходите

Ако се изчертае кривата на Лоренц, може да се види, че нашият анализ е доста по оптимистичен от заключението на Вилфредо Парето, направено в началото на миналия век, че 80 % от капитала е в ръцете на 20% от най-богатите хора, защото според направената анкета само около 40 % от капитала е в ръцете на 20% от най-богатите хора. Това, може би се дължи на факта, че в направената анкета има значително по-еднотипни хора от хората по света.



Фиг. 5. Крива на Лоренц на доходите

ИЗВОДИ

При модела с нормална грешка, т.е. ако се наблюдават само лицата със средна месечна заплата за предходната година до 2087,5 лв., годините на обучение/самообучение оказват статистически значимо влияние върху размера на работната заплата, т. е. тя ще се увеличи средно с 6,3899 лв. за всяка година учене.

Ако се разгледа цялата генерална съвкупност се оказва, че нито образованието, нито трудовият стаж, нито когнитивните способности оказват статистически значимо влияние на средната месечна работна заплата. Т.е. в този случай статистически значими са някои неизследвани от нас фактори.

Когнитивните способности оказват много слабо, почти незабележимо влияние върху размера на заплата. То не е статистически значимо. Точният анализ показва, че едва с 0,2893 лв ще се увеличи очакваната средна работна заплата при подобряване на когнитивните способности с една единица нагоре според приложените в тази работа критерии.

Трудовия стаж също оказва слабо влияние върху размера на заплатата, което не е статистически значимо. Работната заплата се увеличава едва с 0, 0661 лв. за всяка година стаж. Трудовият стаж оказва по-слабо влияние дори от когнитивните способности.

Предложеният модел може да се приложи към други фактори на работната заплата, за да се установят въздействията им. Това позволява насочване на мениджмънта за оптимизиране управлението на основния мотиватор за труд както и ориентир за работната сила за увеличаване на работната заплата.

Статията се публикува с финансовата подкрепа на проект РД 08-79 от 30.01.2019 г. "Финансирането като фактор на ефективността и растежа".

REFERENCES

- Dougherty Cr. (2006). Introduction to Econometrics, Oxford
- Garet, J., Witten, D., Hastie, Tr. & Tibshirani, R. (2013). An Introduction to Statistical Learning with Application in R
- Hansen, B. (2019). Econometrics, University of Wisconsin
- Jordanova, P. Econometrics Lectures, University of Shumen, URL: <http://shu.bg/faculties/fmi/prepodavateli?faculty=fmi&teacherId=166#files> (Accessed on 16.01.2019)
- Kleiber, Chr., Zeileis & Achim (2008). Applied Econometrics with R
- Rencher A.C (2002). Methods of multivariate analysis & A John Willey and Sons
- Salvatore D. & Reagle D. (2001) Theory and Problems of Statistics and econometrics, McGraw Hill