
VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

DUOMENŲ MOKSLAS. BAKALAURAS

2 Laboratorinis darbas

# Modeliai įvykių skaičiui aprašyti

Ataskaita

Užduotį atliko: Ugnius Vilimas

Rytis Baltaduonis

Justinus Pipiras

Aivaras Varkalis

Vilnius 2024

# Turinys

# Įvadas

Modelis įvykių skaičiui aprašyti yra toks modelis, kuris spėja kiek įvykių įvyks per tam tikrą laiką. Šiame darbe mes palyginsime puasono ir neigiamą binominį modelį, jų veikimą ir tikslumą bei pasirinksime tinkamiausią, pagal mūsų pasirinktus duomenis. Darbas atliktas su „R“, „SAS“ ir „Python“ programavimo kalbomis.

# Tikslas

Tikslas – Atlikti regresijos modelio įvykių skaičiui analizę mūsų pasirinktam duomenų paketui.

# Uždaviniai

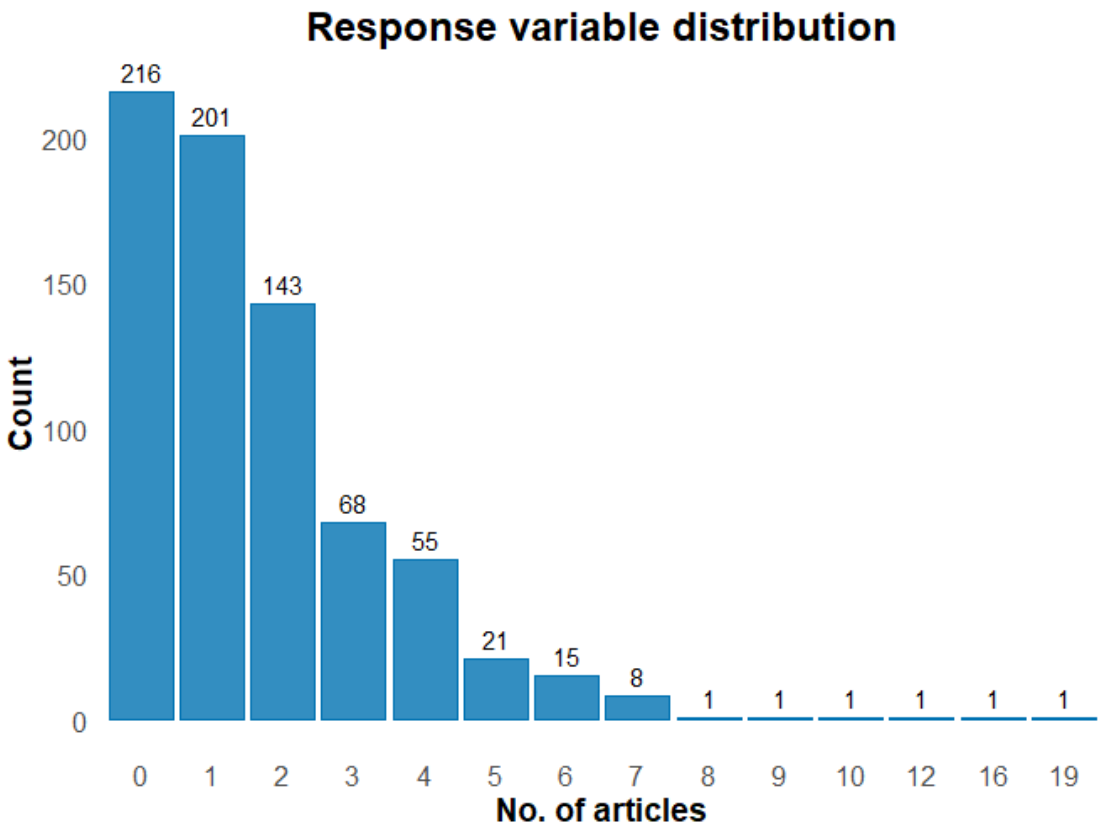
- Pasirinkti tinkamus duomenis būtent šiai analizei atlikti.
- Juos susitvarkyti ir paruošti naudoti.
- Susipažinti su pasirinkto modelio etapais.
- Patikrinti modelį ir palyginti rezultatus su kitomis programavimo kalbomis.
- Padaryti išvadas.

# Duomenys

Mes pasirinkome duomenis iš R paketo (,,bioChemists“), kuriuos sudaro 915 daktaro laipsnių studentų su 6 kintamaisiais: art – išleistų mokslinių straipsnių skaičius, fem – kategorinis kintamasis lytis, kuris įgyja dvi reikšmes (Male, Female), mar – vedybinis statusas, ar susituokęs ar ne, kid5 – vaikų, kuriem 5 metai arba jaunesni skaičius, phd – daktaro laipsnio departamento įvertinimas (angl. ,,prestige‘‘), ment – mentoriaus išleistų straipsnių skaičius per paskutinius 3 metus. Mūsų priklausomas kintamasis yra „art“. Duomenys buvo padalinti į „train_data“ ir „test_data“ santykiu 80/20 atsitiktiniu būdu.

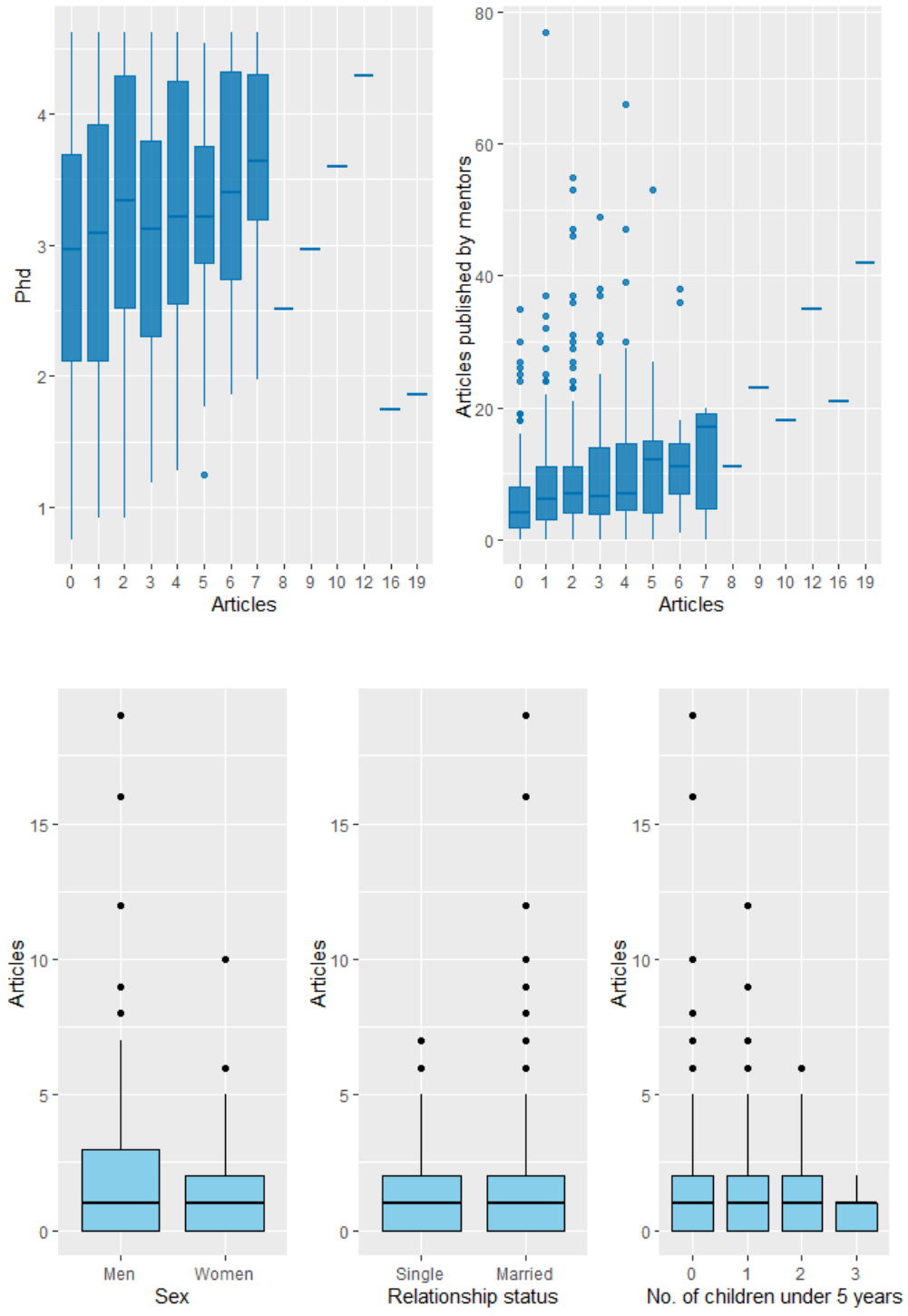
# Analizė su „R“

Pirmiausia atliekame pradinę duomenų analizę. Nubraižome priklausomojo kintamojo („art“) histogramą. Matome, kad dažniausiai pasikartojantis skaičius yra 0, toliau eina 1, 2 ir taip toliau. Galime matyti, kad turime puasono pasiskirstymą. Taip pat matome, kad nėra perteklinių nulių problemos.




*1 pav. Kintamojo „art“ histograma*

Matome nubraižytas stačiakampes diagramas, kuriose pavaizduoti visi nepriklausomi kintamieji. Labiausiai galime pastebėti priklausomybę tarp nepriklausomojo ir priklausomojo kintamojo matome diagramoje, kurioje vaizduojama straipsnių kiekis ir mentoriaus straipsnių kiekis. Taip pat pastebime, kad lytis ir vaikų skaičius taip pat gali turėti įtakos parašytų straipsnių skaičiui.




*2 pav. Visų kintamųjų stačiakampės diagramos*

Čia galime matyti priklausomo kintamojo („art“) vidurkį ir dispersiją. Matome, kad dispersija didesnė negu vidurkis daugiau negu 2 kartus. Tokiu atveju iškarto galime daryti prielaidą, kad puasono modelis nebus tinkamas, tačiau vis tiek jį patikrinsime, kad įsitikintumėme.

*1 Lentelė kintamojo „art“ vidurkis ir dispersija*

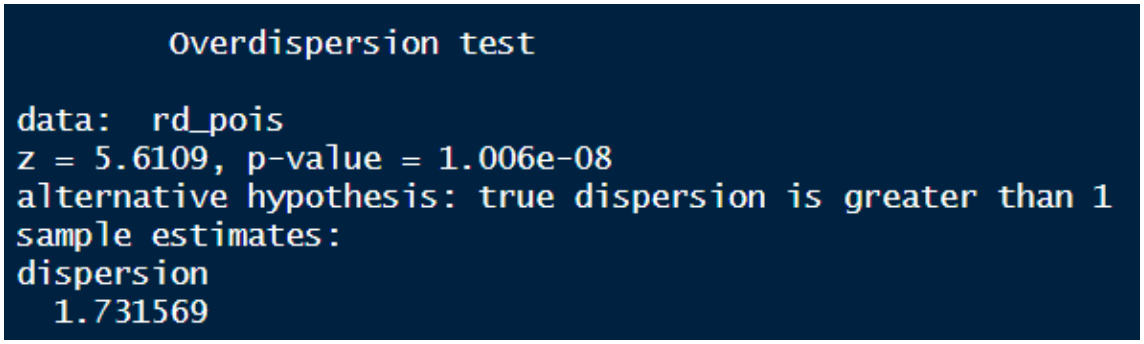
| | |
|---|---|
| Vidurkis | 1,686221 |
| Dispersija | 3,611791 |

Taigi dabar tikriname puasono modelio tinkamumą. Tam sudarėme lentelę.

*2 Lentelė puasono modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,481620 | 0,052096 | 9,245 | ~0 |
| femWomen | -0,273906 | 0,060124 | -4,556 | ~0 |
| Kid5 | -0,186970 | 0,041652 | -4,489 | ~0 |
| ment | 0,025014 | 0,002234 | 11,195 | ~0 |

Iš lentelės daryti išvadų negalime, tačiau sužinome, kad deviacija yra lygi 1284,5, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 1,762. Tai reiškia, kad puasono modelis yra netinkamas naudoti, nes šis skaičius nepatenka į 0,7 ir 1,3 intervalą. Taip pat patikrinome šio modelio AIC (Akaikės informacinis indeksas), kuris buvo – 2635,6.




*3 pav. dispersijos testas*

Matome, kad atlikus dispersijos testą, gauname, kad mūsų p reikšmė yra mažesnė už mūsų nustatytą alfą (0,05). Tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą, kad puasono modelis yra netinkamas naudoti, nes vidurkis ir dispersija statistiškai reikšmingai skiriasi. Toliau žiūrėsime ar tinka neigiamas binominis modelis.

Štai čia matome sudarytą neigiamą binominį modelį pagal pažingsninę regresiją ir tikrinsime jo tinkamumą.

*3 Lentelė neigiamas binominis modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,4344990 | 0,072398 | 5,835 | ~0 |
| femWomen | -0,258973 | 0,079255 | -3,203 | 0,001062 |
| Kid5 | -0,179044 | 0,053687 | -3,335 | 0,000853 |
| ment | 0,028619 | 0,003505 | 8,166 | ~0 |

Šiuo atveju deviacija yra lygi 805,69, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 1,105. Tai reiškia, neigiamas binominis modelis yra tinkamas naudoti ir taikyti, nes gautas skaičius patenka į jau minėtą intervalą. Šio modelio AIC gavome – 2504,6. Palyginus puasono ir neigiamo binominio modelių AIC, matome, kad neigiamo binominio modelio AIC mažesnis, todėl modelis yra labiau tinkamas naudoti.

Toliau atliekame tikėtinumo santykio kriterijaus testą.

```
Likelihood ratio tests of Negative Binomial Models

Response: art
              Model    theta Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
1                 1 1.816703       732       -2569.632
2 fem + ment + kid5 2.399803       729       -2494.622 1 vs 2     3 75.00967 3.330669e-16
```

*4 pav. tikėtinumų santykio kriterijaus testas*

Rezultatų išklotinėje yra tikėtinumų santykio kriterijaus chi kvadrato statistika (75,01) ir p reiškmė. Kadangi p reikšmė yra mažesnė negu alfa (0,05), galime teigti, kad tikėtinumų santykio kriterijus patvirtina neigiamos binominės regresijos modelio tinkamumą.

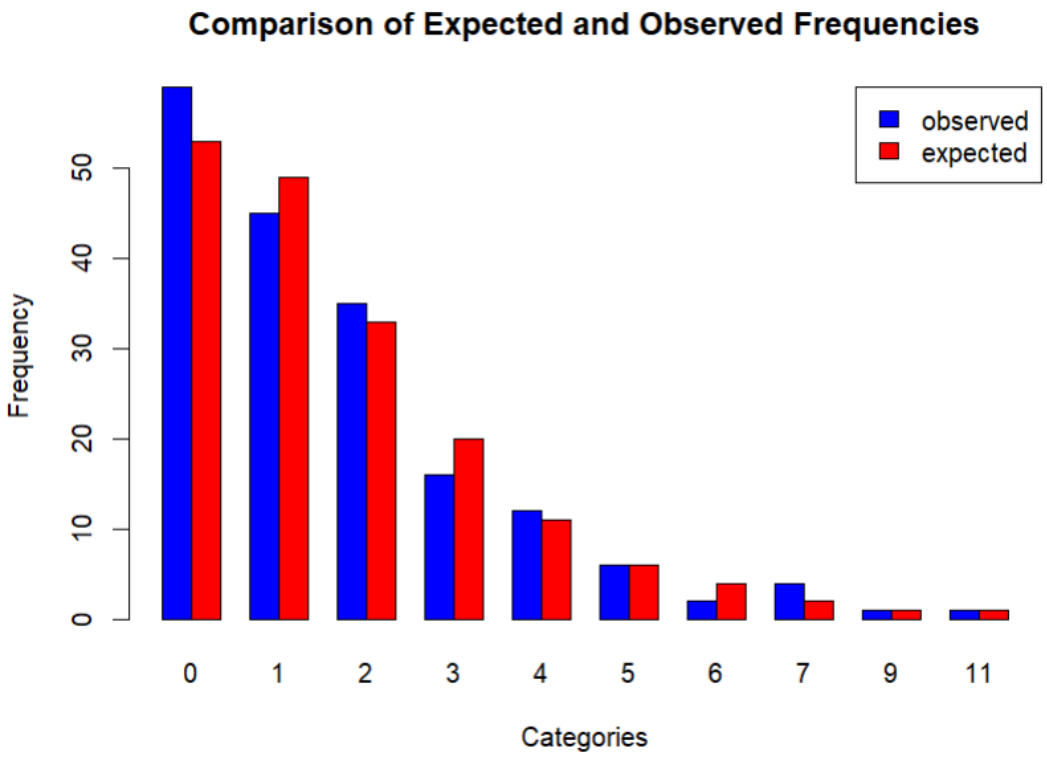
Norėdami staitiškai įvertinti regresorių įtaką, sudarėme tokią lentelę.

*4 Lentelė koeficientų įverčiai ir pasikliovimo intervalai*

| | Estimate | 2,5% | 97,5% |
|---|---|---|---|
| (Intercept) | 1,5449469 | 1,3396419 | 1,7801384 |
| femWomen | 0,7718438 | 0,6611478 | 0,9007478 |
| Ment | 1,0290321 | 1,0214609 | 1,0368064 |
| kid5 | 0,8360688 | 0,7518220 | 0,9286203 |

Lentelėje peteiktos koeficientų įverčių eksponentės ir pasikliovimo intervalai, t.y. daugikliai, rodantys, kiek kartų padidės arba sumažės modeliuojamo kintamojo vidurkis, regresoriui padidėjus vienetu. Matome, kad padidinus mentorių parašytų straipsnių skaičių vienetu, priklausomojo kintamojo reikšmė padidės 2,9%.

Toliau patestavome mūsų modelį su testavimo duomenimis ir pažiūrėjome kokie spėjimų rezultatai. Rezultatai nenuvylė!




*5 pav. testavimo rezultatai*

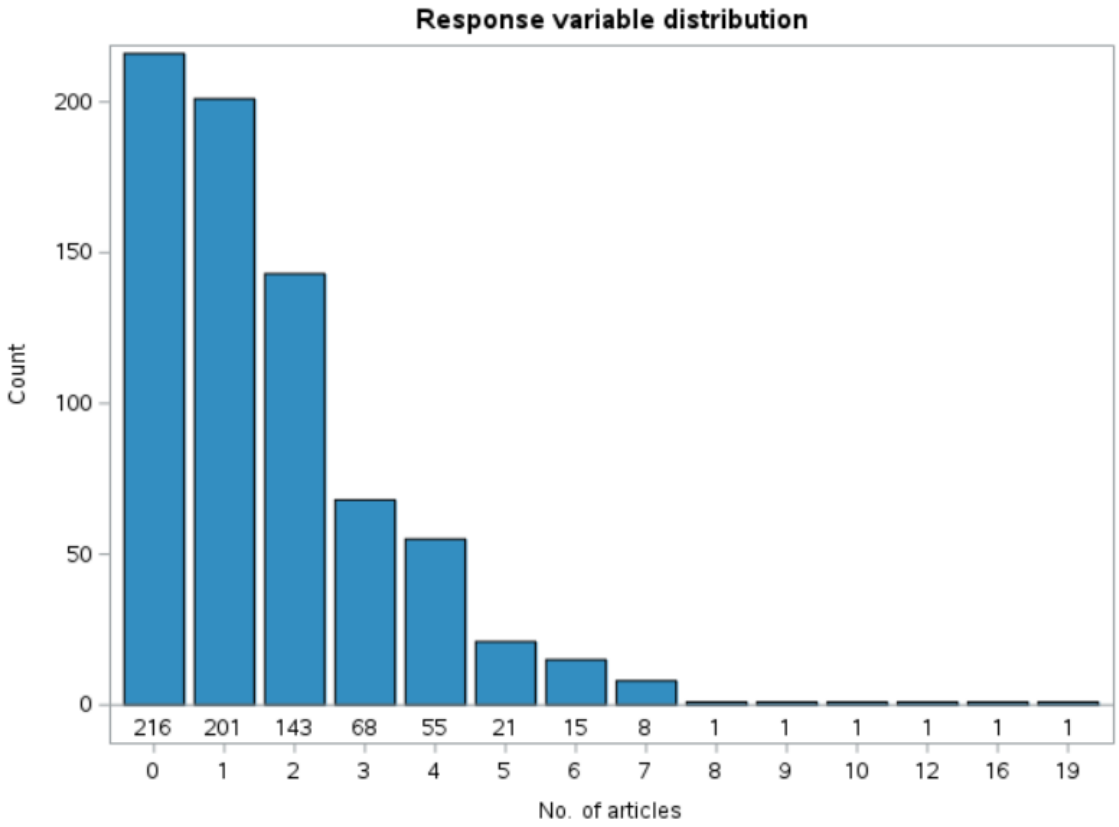
Rezultatus taip pat pateikėme ir lentelėje. Modelio tikslumas - 94%.

*5 Lentelė testavimo duomenų rezultatai*

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| obs | 59 | 45 | 35 | 16 | 12 | 6 | 2 | 4 | 1 | 1 |
| exp | 53 | 49 | 33 | 20 | 11 | 6 | 4 | 2 | 1 | 1 |

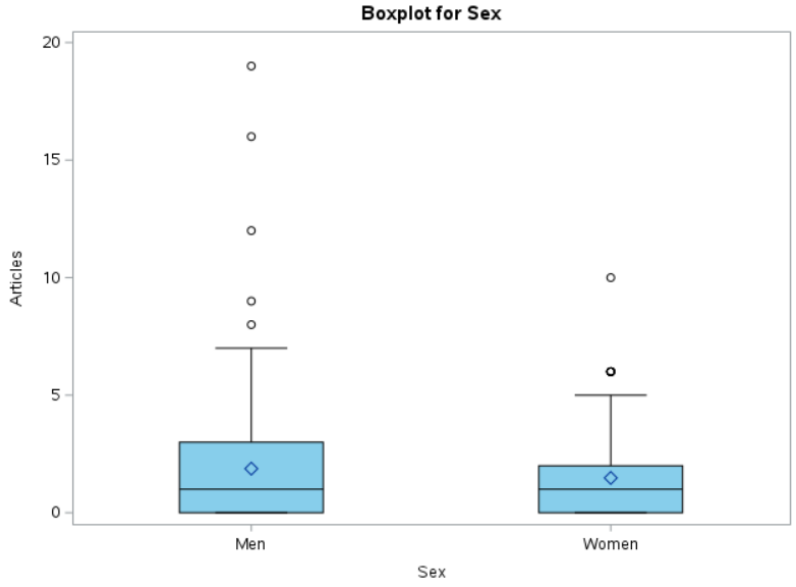
# Analizė su „SAS“

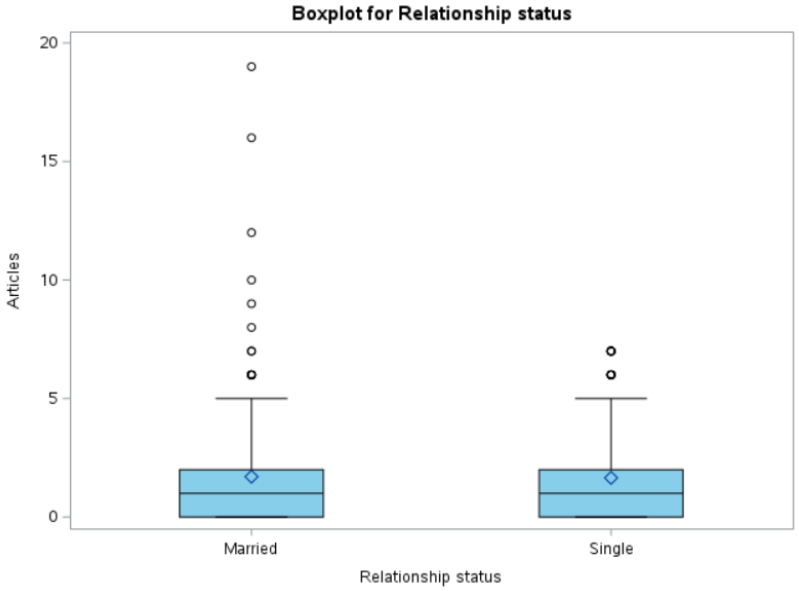
Pirmiausia atliekame pradinę duomenų analizę. Nubraižome priklausomojo kintamojo („art“) histogramą. Matome, kad dažniausiai pasikartojantis skaičius yra 0, toliau eina 1, 2 ir taip toliau. Galime matyti, kad turime puasono pasiskirstymą. Taip pat matome, kad nėra perteklinių nulių problemos.

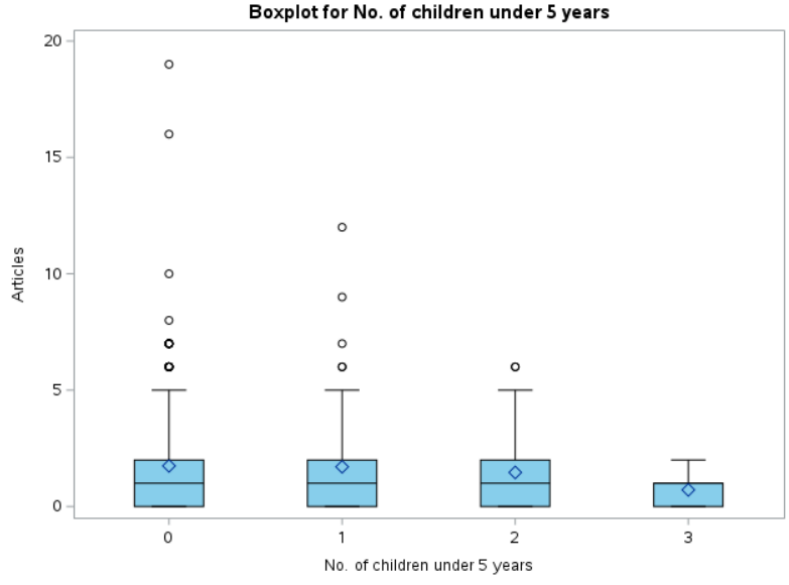


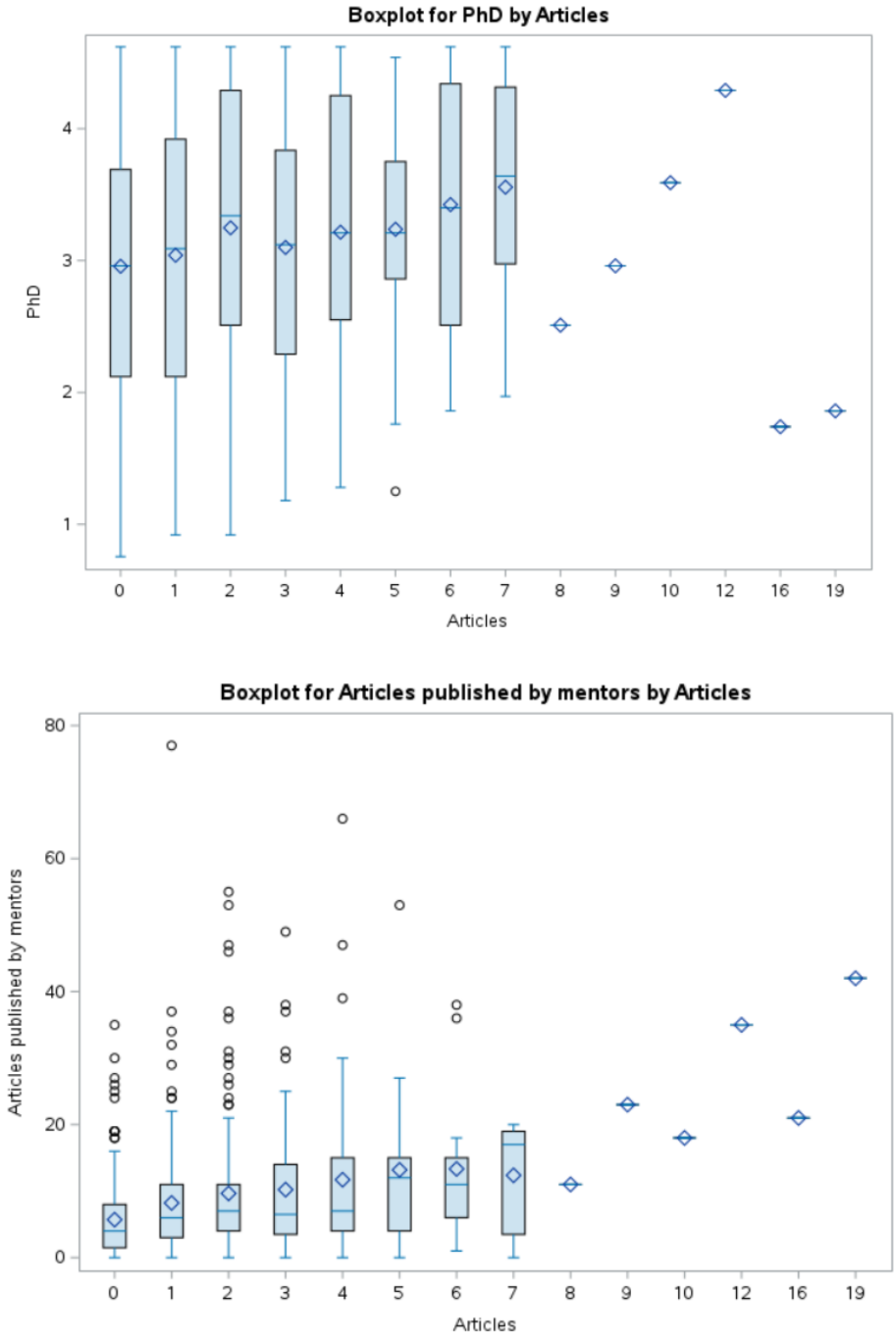

*6 pav. kintamojo „art“ pasiskirstymas*

Matome nubraižytas stačiakampes diagramas, kuriose pavaizduoti visi nepriklausomi kintamieji. Labiausiai galime pastebėti priklausomybę tarp nepriklausomojo ir priklausomojo kintamojo matome diagramoje, kurioje vaizduojama straipsnių kiekis ir mentoriaus straipsnių kiekis. Taip pat pastebime, kad lytis ir vaikų skaičius taip pat gali turėti įtakos parašytų straipsnių skaičiui.


Boxplot for Sex
Articles
20
15
10
5
0
Men
Women
Sex


Boxplot for Relationship status
Articles
20
15
10
5
0
Married
Single
Relationship status


Boxplot for No. of children under 5 years
Articles
20
15
10
5
0
0
1
2
3
No. of children under 5 years




*7 pav. visų kintamųjų stačiakampės diagramos*

Čia galime matyti priklausomo kintamojo („art“) vidurkį ir dispersiją. Matome, kad dispersija didesnė negu vidurkis daugiau negu 2 kartus. Tokiu atveju iškarto galime daryti prielaidą, kad puasono modelis nebus tinkamas, tačiau vis tiek jį patikrinsime, kad įsitikintumėme.

*6 Lentelė kintamojo „art“ vidurkis ir dispersija*

| | |
|---|---|
| Vidurkis | 1,686221 |
| Dispersija | 3,611791 |

Taigi dabar tikriname puasono modelio tinkamumą. Tam sudarėme lentelę.

*7 Lentelė puasono modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,4816 | 0,0521 | 9,245 | ~0 |
| femWomen | -0,2739 | 0,0601 | -4,556 | ~0 |
| Kid5 | -0,1870 | 0,0417 | -4,489 | ~0 |
| ment | 0,0250 | 0,0022 | 11,195 | ~0 |

Iš lentelės daryti išvadų negalime, tačiau sužinome, kad deviacija yra lygi 1284,5, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 1,762. Tai reiškia, kad puasono modelis yra netinkamas naudoti, nes šis skaičius nepatenka į 0,7 ir 1,3 intervalą. Taip pat patikrinome šio modelio AIC (Akaikės informacinis indeksas), kuris buvo – 2635,6.

Štai čia matome sudarytą neigiamą binominį modelį pagal pažingsninę regresiją ir tikrinsime jo tinkamumą.

*8 Lentelė neigiamas binominis modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,4350 | 0,0726 | 5,835 | ~0 |
| femWomen | -0,2590 | 0,0789 | -3,203 | 0,0010 |
| Kid5 | -0,1790 | 0,0539 | -3,335 | 0,0009 |
| ment | 0,0286 | 0,0038 | 8,166 | ~0 |

Šiuo atveju deviacija yra lygi 805,69, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 1,105. Tai reiškia, neigiamas binominis modelis yra tinkamas naudoti ir taikyti, nes gautas skaičius patenka į jau minėtą intervalą. Šio modelio AIC gavome – 2504,6. Palyginus puasono ir neigiamo binominio modelių AIC, matome, kad neigiamo binominio modelio AIC mažesnis, todėl modelis yra labiau tinkamas naudoti.

Norėdami staitiškai įvertinti regresorių įtaką, sudarėme tokią lentelę.

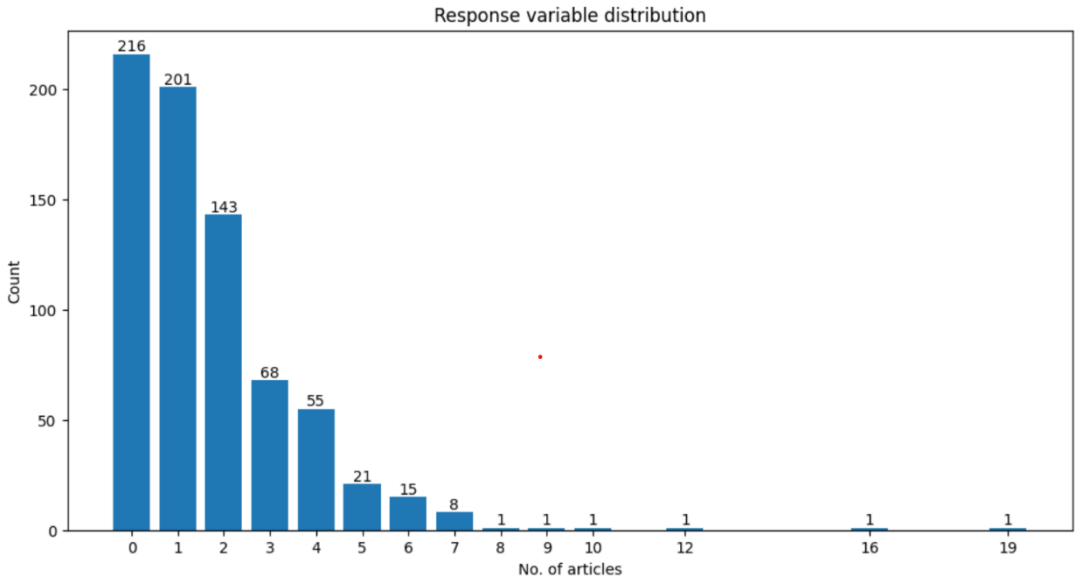
*9 Lentelė koeficientų įverčiai ir pasikliovimo intervalai*

| | Estimate | 2,5% | 97,5% |
|---|---|---|---|
| (Intercept) | 1,5449469 | 1,3396419 | 1,7801384 |
| femWomen | 0,7718438 | 0,6611478 | 0,9007478 |
| Ment | 1,0290321 | 1,0214609 | 1,0368064 |
| kid5 | 0,8360688 | 0,7518220 | 0,9286203 |

Lentelėje peteiktos koeficientų įverčių eksponentės ir pasikliovimo intervalai, t.y. daugikliai, rodantys, kiek kartų padidės arba sumažės modeliuojamo kintamojo vidurkis, regresoriui padidėjus vienetu. Matome, kad padidinus mentorių parašytų straipsnių skaičių vienetu, priklausomojo kintamojo reikšmė padidės 2,9%.

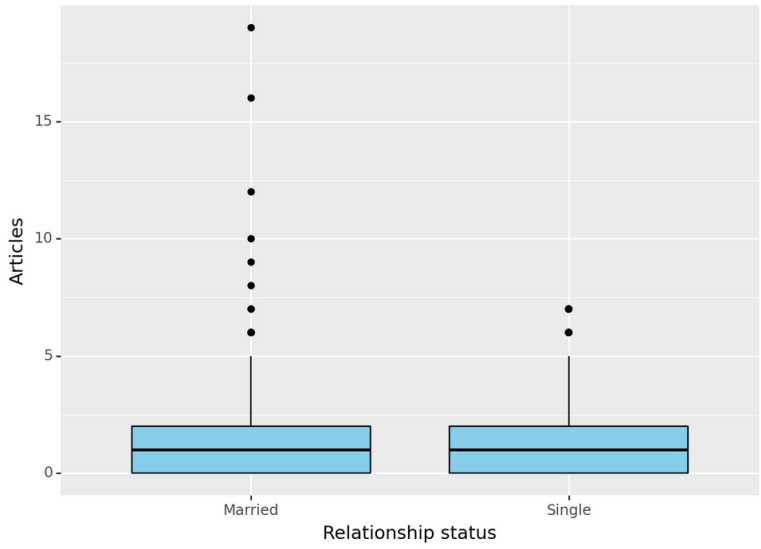
# Analizė su „Python“

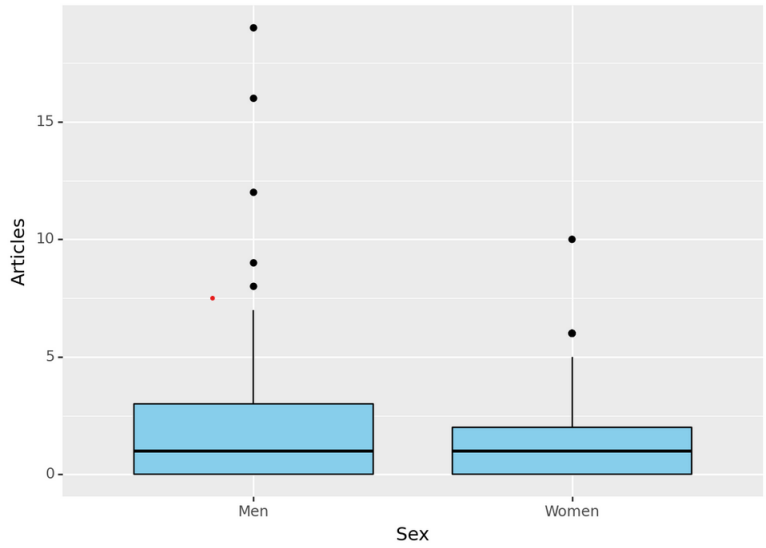
Pirmiausia atliekame pradinę duomenų analizę. Nubraižome priklausomojo kintamojo („art“) histogramą. Matome, kad dažniausiai pasikartojantis skaičius yra 0, toliau eina 1, 2 ir taip toliau. Galime matyti, kad turime puasono pasiskirstymą. Taip pat matome, kad nėra perteklinių nulių problemos.

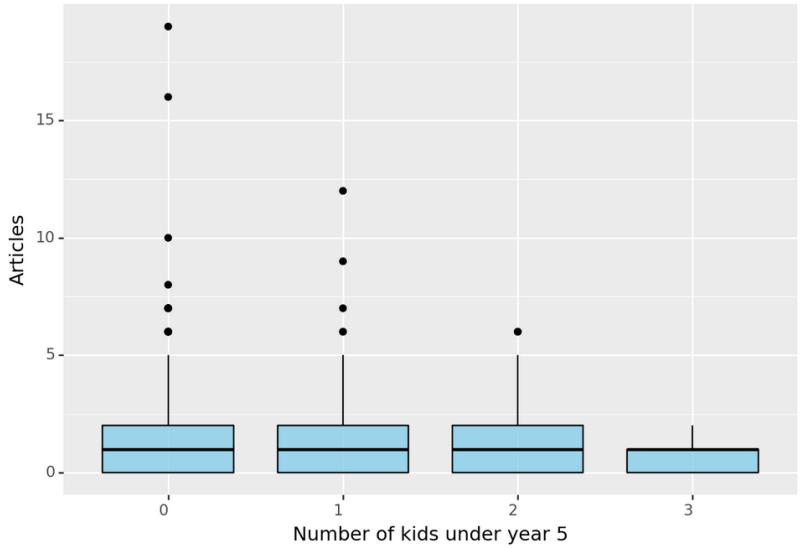


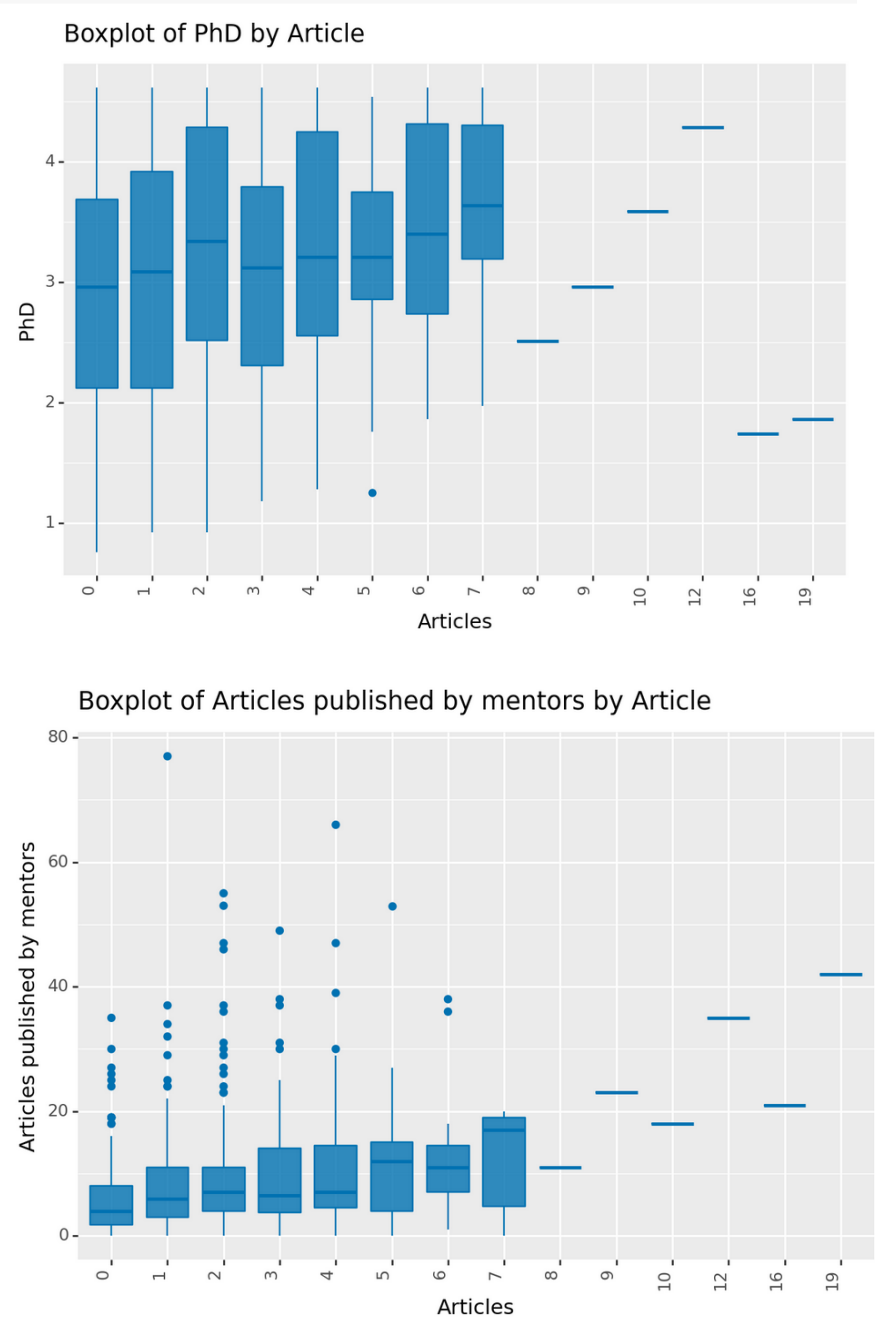

*8 pav. kintamojo „art“ pasiskirstymas*

Matome nubraižytas stačiakampes diagramas, kuriose pavaizduoti visi nepriklausomi kintamieji. Labiausiai galime pastebėti priklausomybę tarp nepriklausomojo ir priklausomojo kintamojo matome diagramoje, kurioje vaizduojama straipsnių kiekis ir mentoriaus straipsnių kiekis. Taip pat pastebime, kad lytis ir vaikų skaičius taip pat gali turėti įtakos parašytų straipsnių skaičiui.


Articles
15
10
5
0
Married
Single
Relationship status


Articles
15
10
5
0
Men
Women
Sex


Articles
15
10
5
0
0
1
2
3
Number of kids under year 5




*9 pav. visų kintamųjų stačiakampės diagramos*

Čia galime matyti priklausomo kintamojo („art“) vidurkį ir dispersiją. Matome, kad dispersija didesnė negu vidurkis daugiau negu 2 kartus. Tokiu atveju iškarto galime daryti prielaidą, kad puasono modelis nebus tinkamas, tačiau vis tiek jį patikrinsime, kad įsitikintumėme.

*10 Lentelė kintamojo „art“ vidurkis ir dispersija*

| | |
|---|---|
| Vidurkis | 1,686221 |
| Dispersija | 3,611791 |

Taigi dabar tikriname puasono modelio tinkamumą. Tam sudarėme lentelę.

*11 Lentelė puasono modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,4816 | 0,0521 | 9,245 | ~0 |
| femWomen | -0,2739 | 0,0601 | -4,556 | ~0 |
| Kid5 | -0,1870 | 0,0417 | -4,489 | ~0 |
| ment | 0,0250 | 0,0022 | 11,195 | ~0 |

Iš lentelės daryti išvadų negalime, tačiau sužinome, kad deviacija yra lygi 1284,5, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 1,762. Tai reiškia, kad puasono modelis yra netinkamas naudoti, nes šis skaičius nepatenka į 0,7 ir 1,3 intervalą. Taip pat patikrinome šio modelio AIC (Akaikės informacinis indeksas), kuris buvo – 2633,6.

Štai čia matome sudarytą neigiamą binominį modelį pagal pažingsninę regresiją ir tikrinsime jo tinkamumą.

*12 Lentelė neigiamas binominis modelis*

| | β | SE | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0,4198 | 0,090 | 4,655 | ~0 |
| femWomen | -0,2518 | 0,099 | -2,538 | 0,011 |
| Kid5 | -0,1773 | 0,066 | -2,667 | 0,008 |
| ment | 0,0297 | 0,005 | 6,354 | ~0 |

Šiuo atveju deviacija yra lygi 574,69, o laisvės laipsnis – 729. Padalinus deviacija iš laisvės laipsnio gauname 0,788. Tai reiškia, neigiamas binominis modelis yra tinkamas naudoti ir taikyti, nes gautas skaičius patenka į jau minėtą intervalą. Šio modelio AIC gavome – 2555,2. Palyginus puasono ir neigiamo binominio modelių AIC, matome, kad neigiamo binominio modelio AIC mažesnis, todėl modelis yra labiau tinkamas naudoti.

Norėdami staitiškai įvertinti regresorių įtaką, sudarėme tokią lentelę.

*13 Lentelė koeficientų įverčiai ir pasikliovimo intervalai*

| | Estimate | 2,5% | 97,5% |
|---|---|---|---|
| (Intercept) | 1,521663 | 1,275142 | 1,815844 |
| femWomen | 0,777371 | 0,639966 | 0,944278 |
| Ment | 1,030182 | 1,020777 | 1,039674 |
| kid5 | 0,837522 | 0,735213 | 0,954068 |

Lentelėje peteiktos koeficientų įverčių eksponentės ir pasikliovimo intervalai, t.y. daugikliai, rodantys, kiek kartų padidės arba sumažės modeliuojamo kintamojo vidurkis, regresoriui padidėjus vienetu. Matome, kad padidinus mentorių parašytų straipsnių skaičių vienetu, priklausomojo kintamojo reikšmė padidės 2,9%.

# Išvados

- Su mūsų duomenimis, išbandėme 2 modelius, puasono ir neigiamą binominį.
- Pagal tam tikrus kriterijus, pasirinkome naudoti neigiamą binominį modelį, dėl jo geresnio tinkamumo.
- Su skirtingomis programavimo kalbomis, gavome šiek tiek kitokius rezultatus, tačiau iš esmės, niekas nesikeičia. „R“ ir „SAS“ programavimo kalbose gaunami identiški rezultatai, o „Python“ programavimo kalboje, sudarius neigiamą binominį modelį, gauname šiek tiek kitas reikšmes, tačiau vis tiek išvada tokia, kad neigiamas binominis modelis, mūsų duomenims buvo tinkamesnis naudoti.
- Modelio spėjimai visose programavimo kalbose buvo panašūs. Modelis spėja gana tiksliai, tai yra – 94%.

# Šaltiniai


- https://stats.stackexchange.com/questions/576282/how-do-you-test-if-the-average-of-a-population-is-the-same-as-the-variance-of-th
- https://en.wikipedia.org/wiki/Akaike_information_criterion
- https://en.wikipedia.org/wiki/Likelihood-ratio_test
- V.Čekanavičius, G. Murauskas, Taikomoji regresinė analizė socialiniuose tyrimuose, 2014, Vilniaus universiteto leidykla, 561 p., ISBN 978-609-459-300-0. http://www.statistika.mif.vu.lt/wp-content/uploads/2014/04/regresine-analize.pdf