Vilniaus
universitetas

MATEMATIKOS IR INFORMATIKOS
FAKULTETAS

MODELIS ĮVYKIŲ SKAIČIUI APRAŠYTI

Laboratorinis darbas

Atliko: Simona Gelžinytė,
Ugnė Kniukštaitė,
Laineda Morkytė,
Austėja Valeikaitė,
duomenų mokslas 3 k. 2gr.

Vilnius, 2023

TURINYS

1. ĮVADAS

1.1 Tyrimo tikslas

Parinkti tinkamą regresijos modelį įvykių skaičiui.

1.2 Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelio prielaidas;
- Sukonstruoti modelį;
- Modelio tinkamumą įvertinti;
- Pateikti gauto modelio interpretacijas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3 Duomenys ir programinė įranga

Pasirinktas duomenų rinkinys apie biochemijos doktorantų mokslinį produktyvumą. Priklausomas kintamasis – per pastaruosius trejus doktorantūros metus paskelbtų straipsnių skaičius, ir 5 kovariantės:

- Lytis (kategorinis);
- Ar vedęs (kategorinis);
- Vaikų skaičius – jaunesnių nei 6 metai;
- Prestižas – studijų programos prestižas;
- Mentorius – doktoranto mentoriaus paskelbtų straipsnių skaičius.

Duomenys paimti iš „R“ paketo „AER“. Iš viso yra 915 stebėjimų, praleistų reikšmių nėra. Tyrimo metu naudota programinė įranga: „SAS“, „R“ ir „Python“.

2. MODELIS ĮVYKIŲ SKAIČIUI APRAŠYTI

Norint turimiems duomenims parinkti tinkamą modelį iš pradžių turime atlikti:

- Pirminę duomenų analizę – vizualiai patikriname ar nėra išskirčių – Kuko matas, standartizuotos liekanos / DFBetų statistika;
- Prielaidų tikrinimą – multikolinearumo problema, modelio parinkimas – dispersijų testas;

Po dispersijų testo nuspręsime taikyti Puasono, neigiamą binominį ar binominį modelį. Tada atliksime:

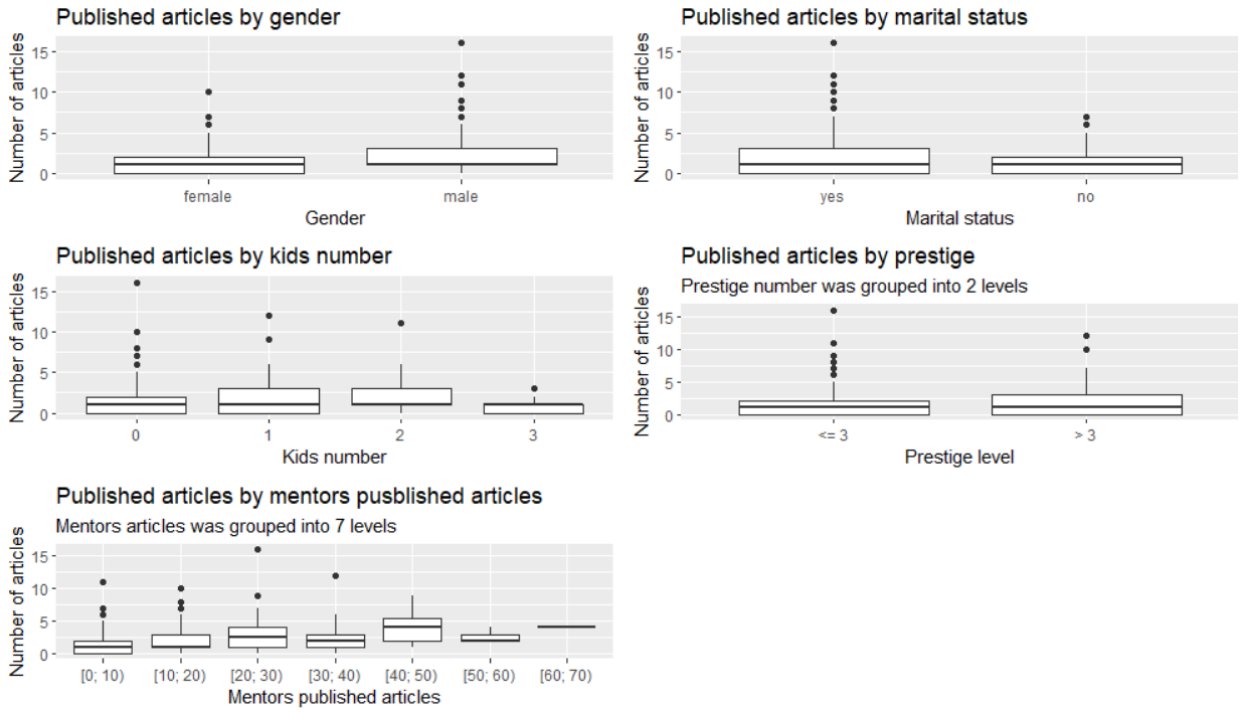
- Reikšmingų kovariančių atranką;
- Modelio tinkamumo analizę;
- Ryšių tarp kintamųjų interpretavimą.

Duomenys buvo padalinti į testavimo ir mokymo aibes 30 : 70 santykiu. Iš pradžių tarsime, kad mūsų duomenis aprašo Puasono modelis, kuris atrodo taip:

$$\ln(publikuoti\ straipsniai) = \beta_0 + \beta_1 \cdot lytis + \beta_2 \cdot vedęs + \beta_3 \cdot vaikai + \beta_4 \cdot prestižas + \beta_6 \cdot mentoriaus\ straipsnių\ skaičius.$$

2.1 Modelis įvykių skaičiui aprašyti naudojant R

Prieš pradedant tikrinti modelio prielaidas, pasižiūrėsime vizualiai į duomenis. Kiekvienam požymiui nubraižysime stačiakampę diagrama, kur y ašyje bus atidėtas priklausomas kintamasis, t. y. pasižiūrėsime kaip kiekvienas požymis yra išsibarstęs priklausomojo kintamojo atžvilgiu. Duomenų rinkinį turime du kategorinius kintamuosius: lytį ir vedybinį statusą, kiekvienai kategorijai atskirai pažiūrėsime duomenų išsibarstymą. Vaikų skaičius šeimoje kinta nuo 0 iki 3 ir kiekvienam skaičiui taip pat atskirai nubraižysime stačiakampes diagramas. Prestižo lygį padalinsime į dvi kategorijas iki 3 imtinai ir daugiau nei 3, o mentorių publikuotus straipsnius suskirstėme į 7 kategorijas, vizualiam pasižiūrėjimui, kaip duomenys yra pasiskirstę.



1 pav. Požymių stačiakampių diagramų panelė

Iš (1 pav.) matome, jog dažniausiai vyrai publikuoja daugiau straipsnių. Taip pat daugiau išleidžia straipsnių vedę vyrai arba ištekėjusios moterys. Mažiausiai straipsnių per 3 metus išleido didžiausią šeimą turintys asmenys. Studijų programos prestižas nedaro didelės įtakos straipsnių publikavimo skaičiui. Daugiausiai straipsnių publikavo tie asmenys, kurių mentorius publikavo nuo 40 imtinai iki 50 straipsnių.

Taip pat dar pasižiūrėsime, kiek imtyje turime moterų / vyrų, vedusių / nevedusių žmonių, vaikų skaičiaus dažnius, prestižo lygio dažnio lentelę, kai prestižo lygis sugrupuotas į 10 intervalų bei mentorių išleistų straipsnių dažnius, kai straipsnių skaičius suskirstytas į 7 intervalus.

1 lentelė. Moterių ir vyrų pasiskirstymas mokymo imtyje

Lytis	**Moteris**	**Vyras**
Asmenų skaičius	298	351

2 lentelė. Vedybinio statuso dažnių lentelė

Vedybinė padėtis	**Vedęs / ištekėjusi**	**Nevedęs / netekėjusi**
Asmenų skaičius	428	221

3 lentelė. Vaikų skaičiaus šeimoje dažniai

Vaikų skaičius šeimoje iki 6 metų	0	1	2	3
Stebėjimų skaičius	430	135	71	13

4 lentelė. Prestižo lygio intervalų dažniai

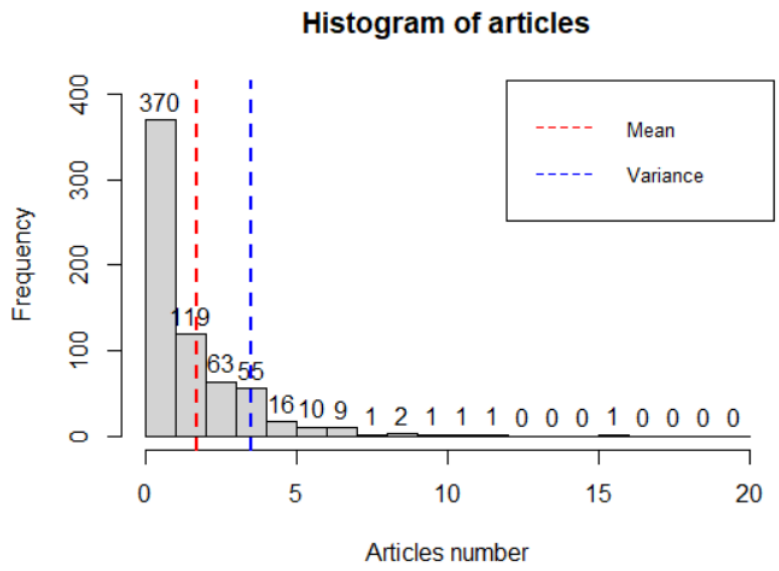
Prestižo lygio intervalai	[0; 0,5]	(0,5; 1]	(1; 1,5]	(1,5; 2]	(2; 2,5]	(2,5; 3]	(3; 3,5]	(3,5; 4]	(4; 4,5]	(4,5; 5]
Stebėjimų skaičius	0	3	28	76	88	117	92	94	93	58

5 lentelė. Mentorių publikuotų straipsnių intervalų dažniai

Mentorių publikuotų straipsnių intervalai	[0; 10]	(10; 20]	(20; 30]	(30; 40]	(40; 50]	(50; 60]	(60; 70]
Stebėjimų skaičius	391	133	39	15	7	3	1

Taigi, matome iš (1 lentelė - 5 lentelė), jog vyrų ir moterų skaičius mokymosi imtyje yra apylygis. Vedę / ištekėję asmenys imtyje sudaro 66 proc. Dažniausias vaikų iki 6 metų skaičius yra 0 arba 1. Dažniausias studijų programos prestižo lygis yra nuo 2,5 iki 3 imtinai, o mažesnio nei 0,5 prestižo lygio studijų programos imtyje nebuvo. Daugiausiai mentoriai buvo išleidę straipsnių nuo 0 iki 20 (imtinai abu galai intervalo).

Iš pradžių tarsime, kad mūsų duomenis aprašo Puasono regresija ir tikrinsime vizualiai, ar dispersija yra lygi vidurkiui.



2 pav. Vizualus tikrinimas, ar priklausomojo kintamojo vidurkis yra lygus dispersijai

Iš (2 pav.) matome, kad mokymosi duomenyse mažiausiais publikuotų straipsnių skaičius yra 0, didžiausias 16. Taip pat matome, jog parašytų straipsnių vidurkis yra netoli 2 (raudona punktyrinė linija), o dispersija yra didesnė apytiksliai 3,5 (mėlyna punktyrinė linija). Išvada iš šio grafiko: vidurkis mažesnis už dispersiją. Statistiškai suskaičiavę vidurkį ir dispersiją gavome reikšmes atitinkamai 1,7 ir 3,5.

Pritaikę duomenims Puasono regresiją, matome, jog šiame modelyje turėtumėme visas reikšmingas kovariantes išskyrus studijų programos prestižo lygį.

```
Call:
glm(formula = articles ~ ., family = poisson, data = train)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6166  -1.5232  -0.3599   0.5973   5.5049
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.174280   0.110053   1.584  0.11329
gendermale   0.275528   0.065397   4.213 2.52e-05 ***
marriedno   -0.153788   0.073412  -2.095  0.03618 *
kids        -0.126844   0.045994  -2.758  0.00582 **
prestige     0.019191   0.032157   0.597  0.55064
```

```
mentor       0.024351   0.002546   9.565  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1250.2  on 648  degrees of freedom
Residual deviance: 1129.5  on 643  degrees of freedom
AIC: 2340.6
```

Vienas iš pagrindinių įvykių skaičiui regresijos modelių tinkamumo įvertinimų yra liekanų deviacija / liekanų laisvės skaičių. Gauta konstanta yra 1,8, o tai parodo, kad Puasono modelis netinka, nes konstanta turėtų patekti į $[0{,}7;\ 1{,}3]$ intervalą.

Norint įsitikinti, ar iš tikrųjų mūsų duomenims netinka Puasono modelis, atliksime dispersijų testą. Tikrinsime hipotezę:

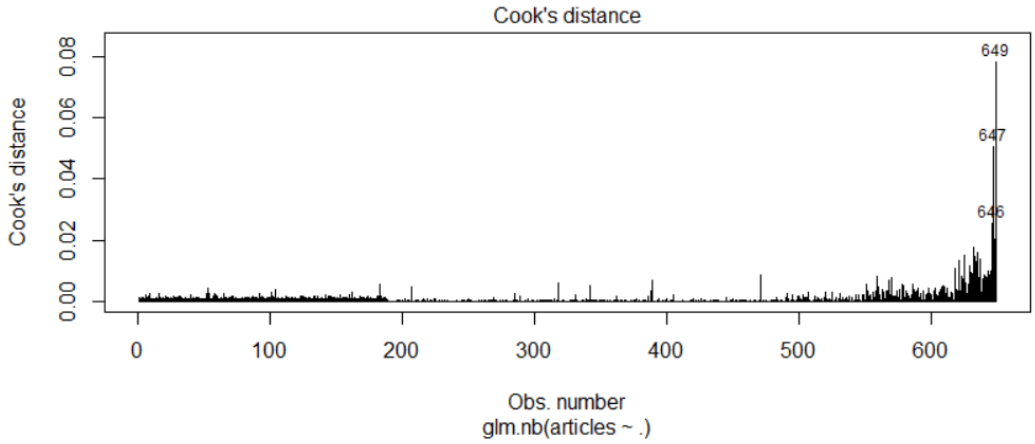
$$\begin{cases} H_0 : \alpha = 0; \\ H_1 :\ \alpha > 0. \end{cases}$$

Jei nulinę hipotezę atmesime, gausime, jog vidurkis nėra lygus dispersijai ir Puasono modelis netinka.

Pritaikę dispersijų testą su argumentu `trafo` = 1, gavome, jog p = $2{,}653 \times 10\text{^}(-7) < 0{,}05$ = reikšmingumo lygmuo, tai nulinę hipotezę atmetame, priimame alternatyvą, t.y. $\alpha > 0$, tai indikuoja, kad mūsų duomenims turėtų tikti neigiamas binominis modelis. Argumentas `trafo` = 1, nurodo, kad alternatyvoje tikrinsime, ar tai bus neigiamas binomis I modelis, kurio dispersija gali būti didesnė už vidurkį, bet jų santykis yra toks pats visiems stebėjimams ir lygus $(1 + \alpha)$.

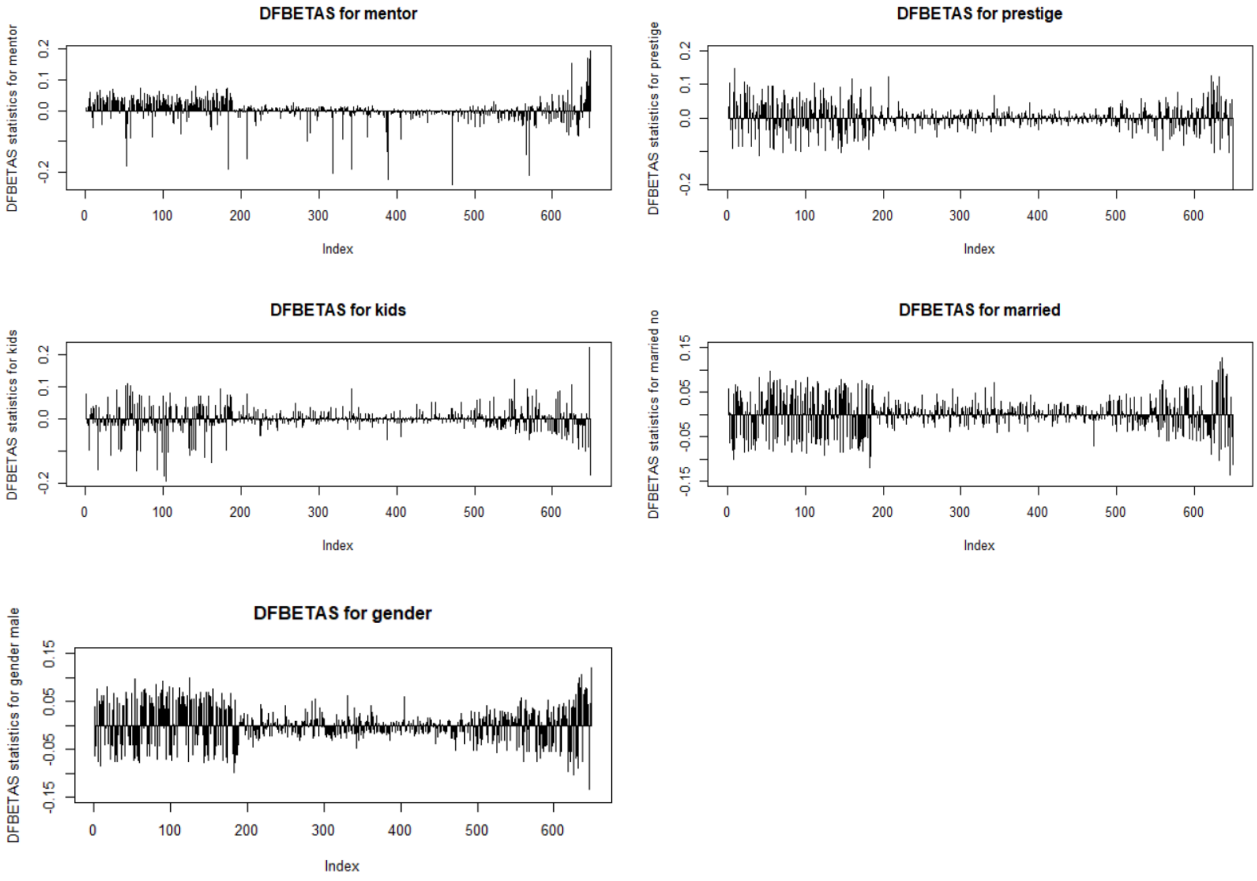
Toliau turime patikrinti, ar yra tenkinamos modelio prielaidos: neturi būti išsiskiriančių stebėjimų bei problemų dėl multikolinearumo.

Norėdami ištirti išskirtis braižėme Cook‘o diagramą, žiūrėjome DFBET‘ų statistika bei standartizuotas liekanas.



3 pav. Cook'o išskirčių grafikas

Iš (3 pav.) matome, jog neturime išsiskiriančių stebėjimų, kadangi Cook‘o koeficientai neviršija 1.



4 pav. DFBet'ų panelė kiekvienam požymiui

Iš (4 pav.) taip pat matome, kad DFBet’ų statistika kiekvienam požymiui neviršija |1|, tai vėlgi parodo, jog išsiskiriančių stebėjimų neturime.

Suskaičiavus standartizuotas liekanas, gavome, jog 2 stebėjimams standartizuotos liekanos yra didesnės už 3.

6 lentelė. Išsiskiriantys stebėjimai pagal standartizuotas liekanas

Asmens nr.	**Straipsnių sk.**	**Lytis**	**Vedybinis statusas**	**Vaikų skaičius iki 6 m.**	**Prestižo lygis**	**Mentoriaus straipsnių skaičius**	**Standartizuotos liekanos reikšmė**
647	11	Vyras	Vedęs	2	2,86	7	3,302
649	16	Vyras	Vedęs	0	1,74	21	3,131

Iš (6 lentelė) matome, išsiskiriančių stebėjimų informaciją. Šie stebiniai taip pat buvo pažymėti Cook'o grafike (3 pav.), tačiau jų Cook'o reikšmės neviršijo 1. Išsiskiriantys asmenys – tai vedę vyrai, kurie per 3 metus vieni iš daugiausiai buvo publikavę straipsnių. Kadangi šių stebėjimų standartizuotos liekanos nedaug viršijo slenkstį bei iš kitų atliktų išskirčių tyrimų šie stebėjimai nebuvo įtarti, tai juos paliksime imtyje.

7 lentelė. VIF'o statistika

Kovariantės	Lytis	Vedybinis statusas	Vaikų skaičius iki 6 m.	Prestižo lygis	Mentoriaus straipsnių skaičius
VIF	1,118	1,291	1,306	1,104	1,104

Iš (7 lentelė) matome, jog visų kovariančių VIF koeficientai neviršija 4, tai galime teigti, jog multikolinearumo problemos neturime.

Kadangi modelio prielaidos patenkintos, nagrinėsime, ar neigiamame binominiame modelyje yra bent viena reikšminga kovariantė, pasinaudodami tikėtinumo santykio kriterijumi. Tikrinsime hipotezę:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \\ H_1: bent\ viena\ lygybė\ neteisinga \end{cases}.$$

Gavome p reikšmę = $4{,}372 \times 10^{-13} < 0{,}05$ = reikšmingumo lygmuo, tai nulinę hipotezę atmetame, t. y. modelyje turime bent vieną reikšmingą kovariantę.

Toliau pasižiūrime gauto modelio santrauką.

```
Call:
glm.nb(formula = articles ~ ., data = train, init.theta = 2.448682202,
    link = log)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1149  -1.3614  -0.2745   0.4626   3.2887
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.125413   0.144292   0.869  0.38476
gendermale   0.267645   0.085153   3.143  0.00167 **
marriedno   -0.162794   0.096261  -1.691  0.09080 .
kids        -0.126604   0.060206  -2.103  0.03548 *
prestige     0.028548   0.042438   0.673  0.50114
mentor       0.026952   0.003827   7.042  1.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Negative Binomial(2.4487) family taken to be 1)
    Null deviance: 785.52  on 648  degrees of freedom
Residual deviance: 714.27  on 643  degrees of freedom
AIC: 2232.2
Number of Fisher Scoring iterations: 1
              Theta:  2.449
          Std. Err.:  0.361
```

Matome, jog beveik visos kovariantės išskyrus prestižo lygmenį yra reikšmingos, kai reikšmingumo lygmuo yra 0,1.

```
Start:  AIC=2230.2
articles ~ gender + married + kids + prestige + mentor
           Df    AIC
- prestige  1 2228.7
<none>        2230.2
- married   1 2231.1
- kids      1 2232.6
- gender    1 2238.0
```

```
- mentor    1 2272.5
Step:  AIC=2228.65
articles ~ gender + married + kids + mentor
           Df    AIC
<none>        2228.7
- married   1 2229.3
+ prestige  1 2230.2
- kids      1 2231.1
- gender    1 2236.6
- mentor    1 2275.8
Call:  glm.nb(formula = articles ~ gender + married + kids + mentor,
    data = train, init.theta = 2.447871346, link = log)
Coefficients:
(Intercept)   gendermale    marriedno         kids       mentor
    0.20452      0.26947     -0.15466     -0.12665      0.02761
Degrees of Freedom: 648 Total (i.e. Null);  644 Residual
Null Deviance:        785.4
Residual Deviance: 714.6      AIC: 2231
```

Pritaikius pažingsninę regresiją, modelyje išmetame prestižo kovariantę.

Gauto modelio AIC ir BIC koeficientai atitinkamai yra 2230,652 ir 2257,504. Liekanų deviacijos ir liekanų laisvės laipsnių santykio rodiklis yra 1,11. Šis rodiklis turi priklausyti intervalui [0,7; 1,3], kad galėtumėme teigti, jog modelis tinka nagrinėjamiems duomenims. Pseudo R kvadrato reikšmė yra:

$$R^2 = 1 - \frac{liekanų\ deviacija}{nulinė\ deviacija} = 0{,}09.$$

Ši reikšmė yra nedidelė, nelabai tinkama prognozavimui. Taip pat buvo gautas α įvertis:

$$\alpha = \frac{1}{\theta} = 0{,}409.$$

Suskaičiavome tikėtinus ir esamus dažnius.

8 lentelė. Tikėtini ir esami dažniai mokymo duomenims

	0	1	2	3	4	5	6	7	8	9	10	11	12	16
Esami dažniai	188	182	119	63	55	16	10	9	1	2	1	1	1	1
Tikėtini dažniai	191	178	120	71	40	22	12	7	4	2	1	1	1	0

Iš (8 lentelė) matome, jog modelis gerai prognozuoja visas galimas reikšmes su nedidele paklaida.

9 lentelė. Koeficientai ir jų eksponentės

Kovariantė	**Lytis vyras**	**Nevedęs / netekėjusi**	**Vaikų skaičius iki 6 m.**	**Mentoriaus išleisti straipsniai**
Koeficientas	0,269	-0,155	-0,127	0,028
$\boldsymbol{Exp(koeficiento)}$	1,309	0,857	0,881	1,028

Iš (9 lentelė) matome, jog būnant vyru didėja išleistų straipsnių vidurkis nei būnant moterimi. Jei vyras nevedęs, moteris netekėjusi, mažėja publikuotų straipsnių vidurkis. Padidėjus vaikų skaičiui šeimoje mažėja išleistų straipsnių vidurkis, o padidėjus mentoriaus išleistų straipsnių skaičiui padidėtų ir doktoranto išleistų straipsnių vidurkis. Norint pateikti detalesnę ir išsamesnę modelio interpretaciją žiūrime į eksponentės reikšmes: būnant vyru apie 31 proc. padidėja išleistų straipsnių vidurkis, o būnant nesusituokus apie 15 proc. mažėja išleistų straipsnių vidurkis bei padidėjus vaikų skaičiui šeimoje vienu vienetu apytiksliai 12 proc. sumažėtų išleistų straipsnių vidurkis, o padidinus vienu vienetu išleistų mentoriaus straipsnių skaičių 3 proc. Padidėtų doktorantų išleistų straipsnių vidurkis.

Taip pat pabandėme į modelį įtraukti sąveikų – radome statistiškai reikšmingą sąveiką tarp prestižo lygmens ir mentoriaus išleistų straipsnių. Pritaikius pažingsninę regresiją naujam modeliui gavome, jog yra paliekamos visos kovariantės.

```
Call:
glm.nb(formula = articles ~ . + prestige * mentor, data = train,
    init.theta = 2.53544063, link = log)
Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-1.9983  -1.3486  -0.2705   0.4520   3.3100
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.200180   0.188499  -1.062 0.288250
gendermale       0.286020   0.084813   3.372 0.000745 ***
marriedno       -0.155262   0.095656  -1.623 0.104562
kids            -0.132220   0.060011  -2.203 0.027575 *
prestige         0.128290   0.056012   2.290 0.021998 *
mentor           0.061240   0.013161   4.653 3.27e-06 ***
prestige:mentor -0.010079   0.003694  -2.728 0.006367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Negative Binomial(2.5354) family taken to be 1)
    Null deviance: 794.37  on 648  degrees of freedom
Residual deviance: 715.27  on 642  degrees of freedom
AIC: 2227.4
Number of Fisher Scoring iterations: 1
              Theta:  2.535
          Std. Err.:  0.382
 2 x log-likelihood:  -2211.419
```

Naujame modelyje visos kovariantės reikšmingos, kai reikšmingumo lygmuo yra ne mažesnis nei 0,1.

Toliau palyginsime du gautus modelius modelius.

10 lentelė. Modelių lyginimas I

Lyginamoji charakteristika	**Modelis be sąveikos**	**Modelis su sąveika**
AIC	2230,652	2227,419
BIC	2257,504	2263,222
Liekanų deviacijos ir liekanų laisvės laipsnių santykis	1,110	1,114
Pseudo R kvadratas	0,090	0,100

Iš (10 lentelė) matome, jog modelių charakteristikos yra labai panašios ir modelių gerumas / tinkamumas neatsiskiria aiškiai.

11 lentelė. Modelių palyginimas II

	0	1	2	3	4	5	6	7	8	9	10	11	12	16
Esami dažniai	188	182	119	63	55	16	10	9	1	2	1	1	1	1
Tikėtini dažniai (modelis be sąveikos)	191	178	120	71	40	22	12	7	4	2	1	1	1	0
Tikėtini dažniai, (modelis su sąveika)	190	178	120	71	40	22	12	7	4	2	1	1	1	0

Taip pat ir iš (11 lentelė) matome, jog abu modeliai gerai prognozuoja visas galimas reikšmes. Modelis su sąveika šiek tiek geriau prognozuoja 0.

Taip pat modelių palyginimui pasižiūrėjome RMSE statistiką testavimo duomenims, kuri yra gaunama pagal formulę:

$$\sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}},$$

$$\text{čia } n - testavimo\ imties\ dydis, P_i - prognozuota\ i - toji\ reikšmė, O_i - esama\ i - toji\ reikšmė.$$

12 lentelė. RMSE statistika

	Modelis be sąveikos	**Modelis su sąveika**
RMSE statistika	2,005	2,704

Iš (12 lentelė) matome, jog mažesnė RMSE statistika yra modelyje be sąveikos. Nors abiejų modelių dauguma modelio tinkamumų statistikų yra panašios, mūsų nuomone, geresnis modelis gavosi be sąveikos, remdamiesi RMSE statistika. Taip pat modelyje su sąveika atsirado multikolinearumo problema, dėl šitų priežasčių modelio su sąveika atsisakėme.

Toliau pateiksime atsitiktinai parinktų 10 prognozuotų reikšmių iš testavimo aibės.

13 lentelė. Prognozuotos reikšmės

Tikroji reikšmė	Prognozuota modeliu be sąveikos
0	2,003
0	1,513
0	1,051
1	1,241
0	1,896
0	1,051
3	2,141
1	1,794
0	1,174
3	3,116

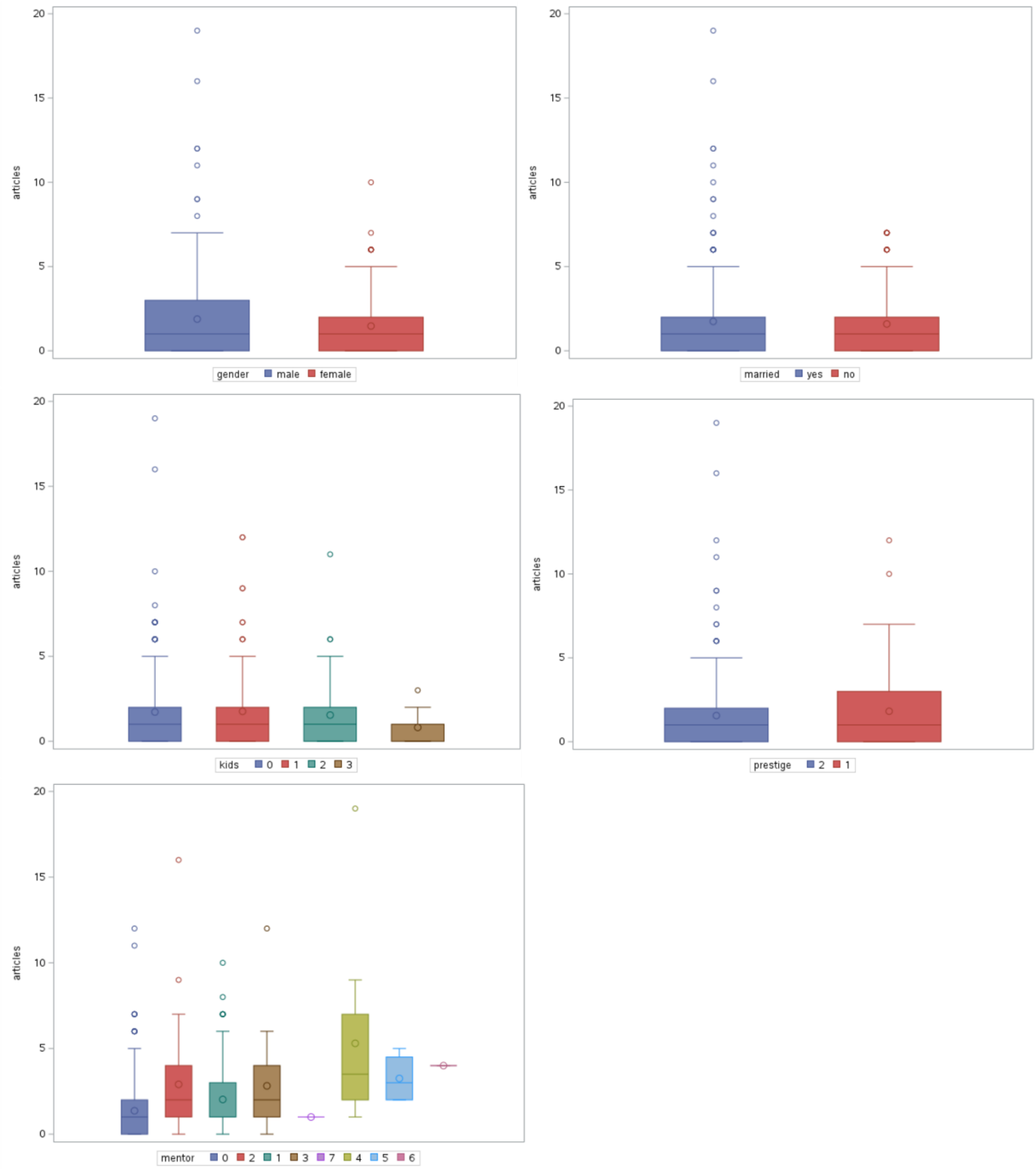
Kaip matome iš (13 lentelė) modelis neprognozuoja tiksliai, tai paaiškina maža pseudo R kvadrato reikšmė.

Taigi, gavome modelį:

$$\ln(publikuoti\ straipsniai) = 0{,}205 + 0{,}269 \cdot lytis\ (vyras) - 0{,}155 \cdot vedybinis\ statusas\ (nevedęs) - 0{,}127 \cdot vaikai + 0{,}028 \cdot mentoriaus\ straipsnių\ skaičius.$$

2.2 Modelis įvykių skaičiui aprašyti naudojant SAS

Analizę pradėsime nusibraižę stačiakampes diagramas. Stulpelį prestižas, padalinome į dvi grupes: 1 grupė – daugiau už 3, 2 grupė – mažiau arba lygų 3. Taip pat stulpelis *mentoriai* buvo padalintas į 8 grupes: 0 grupė – kai straipsnių skaičius buvo 0 iki 9, 1 grupė nuo 10 iki 19, 2 grupė nuo 20 iki 29, 3 grupė nuo 30 iki 39, 4 grupė nuo 40 iki 49, 5 grupė nuo 50 iki 59, 6 grupė nuo 60 iki 69, 7 grupė nuo 70 iki 79 ir galiausiai 8 grupė nuo 80 iki 89 (tik vienas stebėjimas).



5 pav. Stačiakampės diagramos

Iš stačiakampių diagramų galime pamatyti, jog vyrai yra labiau linkę publikuoti straipsnius nei moterys, vertinant pagal vedybinį statusą, tai dažniau daro vedę asmenys, turintys vaikų, prestižo lygis neturi didelės įtakos. Vertinant pagal mentorių publikuotų straipsnių skaičių, daugiausiai straipsnių publikavo asmenys, kurių mentoriai publikavo nuo 40 iki 49 straipsnių.

Taip pat dar pažiūrėkime mokymų aibės dažnių lenteles (stulpelis *prestižas* yra sugrupuotas į 10 grupių (nuo 0 iki 5, kas 0.5; grupės skaičius rodo intervalo galą, pvz. grupė 1 rodo, kad prestižo lygis yra nuo 0.5 iki 1), stulpelis *mentoriai* lieka su tomis pačiomis grupėmis kai ir braižant stačiakampes diagramas).

gender				
gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	298	45.92	298	45.92
male	351	54.08	649	100.00

6 pav. Lyties dažnių lentelė

married				
married	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	221	34.05	221	34.05
yes	428	65.95	649	100.00

7 pav. Vedybinio statuso dažnių lentelė

mentor				
mentor	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	427	65.79	427	65.79
1	153	23.57	580	89.37
2	40	6.16	620	95.53
3	18	2.77	638	98.31
4	7	1.08	645	99.38
5	3	0.46	648	99.85
6	1	0.15	649	100.00

8 pav. Mentorių išleistų straipsnių skaičius

prestige				
prestige	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	0.46	3	0.46
1.5	28	4.31	31	4.78
2	76	11.71	107	16.49
2.5	88	13.56	195	30.05
3	117	18.03	312	48.07
3.5	92	14.18	404	62.25
4	94	14.48	498	76.73
4.5	93	14.33	591	91.06
5	58	8.94	649	100.00

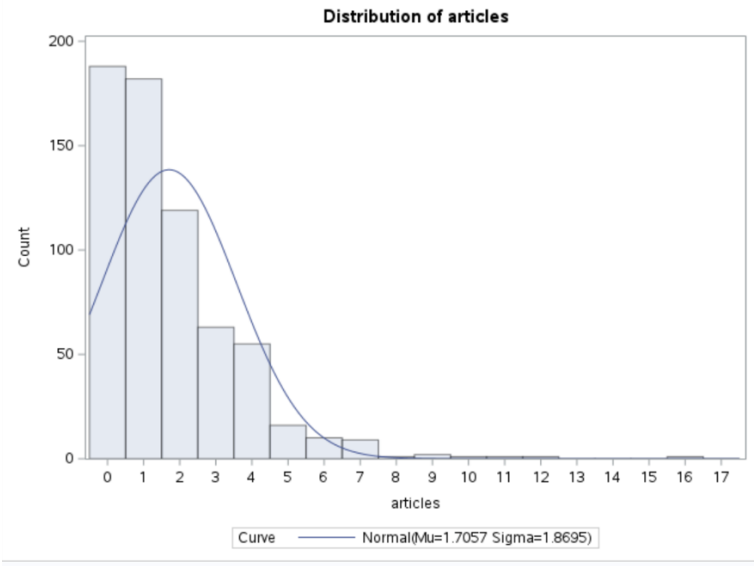
9 pav. Prestižo lygio dažnių lentelė

kids				
kids	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	430	66.26	430	66.26
1	135	20.80	565	87.06
2	71	10.94	636	98.00
3	13	2.00	649	100.00

10 pav. Vaikų skaičiaus dažnių lentelė

Iš dažnių lentelių galime matyti, jog moterų skaičius skiriasi visai nedaug nuo vyrų, vedusių asmenų yra daugiau nei puse, mentorių išleistų straipsnių skaičius didžiausias intervale nuo 0 iki 10, didžiausiais prestižo lygis matomas intervale nuo 2.5 iki 3 (mažesnio nei 0.5 prestižo lygio nebuvo), dažniausias vaikų skaičius yra 0 arba 1.

Dabar patikrinsime vizualiai, ar dispersija lygi vidurkiui.



11 pav. Mokymo aibės histograma

Kaip matome, publikuotų straipsnių skaičius yra intervale nuo 0 iki 16. Dažniausiai buvo nepublikuotas nei vienas straipsnis arba tik vienas. Taip pat histogramoje paskaičiuotas vidurkis (1.7) bei standartinis nuokrypis (1.8), kurie parodo, jog vidurkis yra mažesnis už dispersiją.

Duomenims pritaikome Puasono regresiją.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4498	0.1105	0.2333	0.6663	16.58	<.0001
gender	female	1	-0.2755	0.0654	-0.4037	-0.1474	17.75	<.0001
gender	male	0	0.0000	0.0000	0.0000	0.0000	.	.
married	no	1	-0.1538	0.0734	-0.2977	-0.0099	4.39	0.0362
married	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
kids		1	-0.1268	0.0460	-0.2170	-0.0367	7.61	0.0058
prestige		1	0.0192	0.0322	-0.0438	0.0822	0.36	0.5506
mentor		1	0.0244	0.0025	0.0194	0.0293	91.49	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

12 pav. Kovariančių reikšmingumo lentelė

Kaip matome iš lentelės, visos kovariantės yra reikšmingos išskyrus prestižo lygį. Dabar patikrinime deviacijos ir jos laisvės laipsnių santykius.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	643	1129.5498	1.7567
Scaled Deviance	643	1129.5498	1.7567
Pearson Chi-Square	643	1141.4308	1.7752
Scaled Pearson X2	643	1141.4308	1.7752
Log Likelihood		-455.5619	
Full Log Likelihood		-1164.3131	
AIC (smaller is better)		2340.6261	
AICC (smaller is better)		2340.7570	
BIC (smaller is better)		2367.4787	

13 pav. Suderinamumo matų ir statistikos lentelė

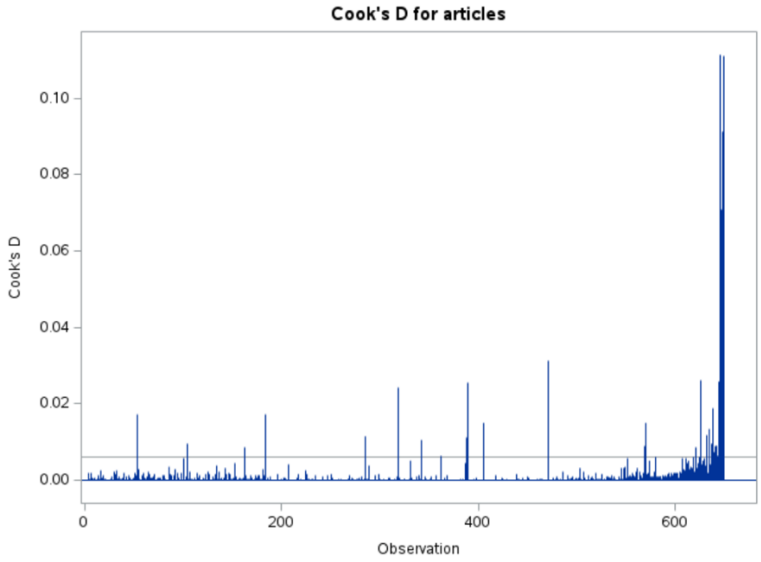
Iš lentelės matome, jog jis lygus 1.75, o tai parodo modelio netinkamumą. Dar kartą tuo įsitikinti galime patikrindami ar deviacijos ir jos laipsnių santykis statistiškai skiriasi nuo 1.

df	chisq	pvalue
643	1129.55	0

14 pav. p-reikšmė

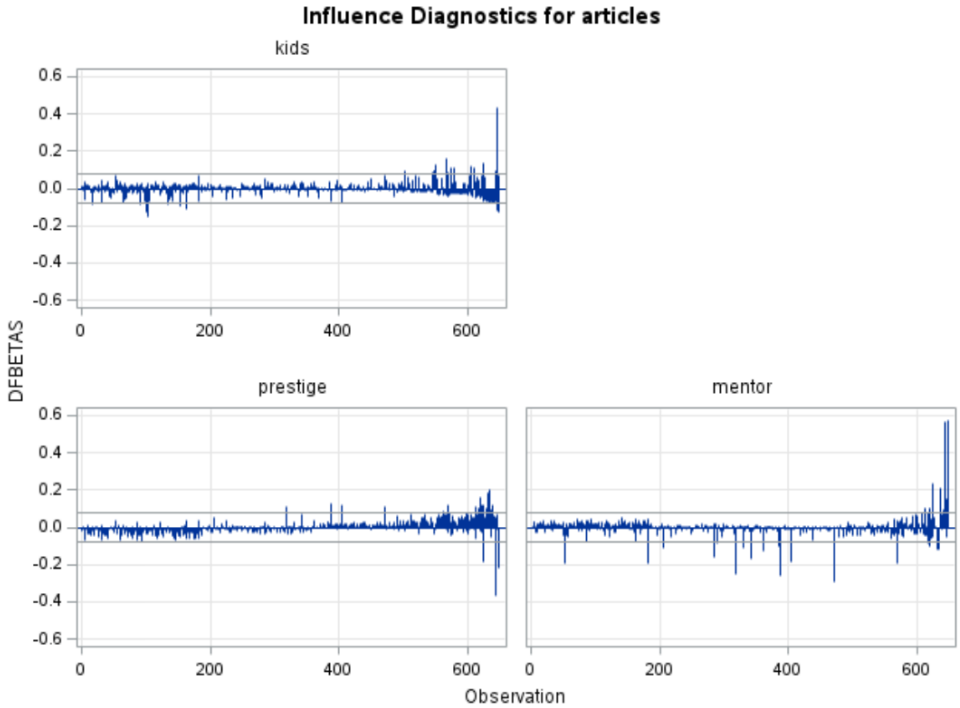
Gauta p-reikšmė lygi 0, o tai atmeta mūsų nulinę hipotezę ir taip dar kartą įsitikiname, jog Puasono modelis netinkamas, tačiau turėtų tikti neigiamas binominis.

Dabar patikrinkime modelio prielaidas: išskirčių bei multikolinearumo nebuvimą. Išskirtims naudosime Kuko diagramą, DFBET‘ų statistiką ir standartizuotąsias liekanas.



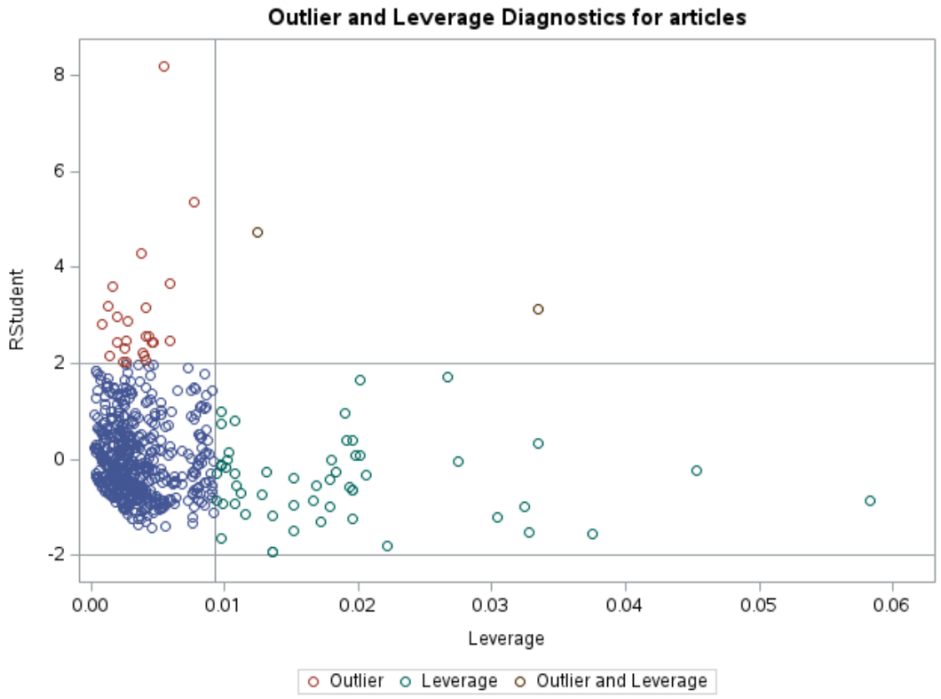
15 pav. Kuko išskirčių grafikas

Kaip matome nei vienas koeficientas neviršija 1, o tai rodo, jog duomenys išskirčių neturi. Patikrinimui galime pasinaudoti DFBET‘ų statistika kiekybiniams duomenims.



16 pav. DFBET’ų statistikos kiekybiniams duomenims

Kaip matome niekas neviršija 1, o tai dar kartą įrodo, jog išskirčių nėra.



17 pav. Standartizuotų liekanų grafikas

Standartizuotų liekanų grafikas rodo, jog vis dėlto išskirčių yra, tačiau kadangi prieš tai atlikti du testai parodė priešingai – duomenis paliksime tokius, kokie yra.

Dabar tikrinsime multikolinearumą.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
kids	kids	1	0.04536	0.09155	0.50	0.6205	1.35182
prestige	prestige	1	0.34144	0.03370	10.13	<.0001	2.33409
mentor	mentor	1	0.05858	0.00787	7.45	<.0001	2.05089

18 pav. Statistikos lentelė

Iš lentelės galime matyti, jog nei viena kiekybinė kovariantė neviršija 4, o tai rodo, jog duomenys neturi multikolinearumo problemos.

Kadangi visi modelio prielaidos patenkintos toliau nagrinėsime ar neigiamame binominiame modelyje yra bent viena reikšminga kovariantė, pasinaudodami tikėtinumo santykio kriterijumi.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3931	0.1473	0.1044	0.6817	7.12	0.0076
gender	female	1	-0.2676	0.0850	-0.4342	-0.1011	9.92	0.0016
gender	male	0	0.0000	0.0000	0.0000	0.0000	.	.
married	no	1	-0.1628	0.0962	-0.3514	0.0258	2.86	0.0907
married	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
kids		1	-0.1266	0.0603	-0.2448	-0.0084	4.40	0.0359
prestige		1	0.0285	0.0425	-0.0548	0.1119	0.45	0.5019
mentor		1	0.0270	0.0041	0.0190	0.0349	44.27	<.0001
Dispersion		1	0.4084	0.0603	0.3057	0.5456		

19 pav. Regresijos parametrų įverčiai

Pritaikę neigiamą binominį modelį gauname regresijos parametrų įverčius su kuriais galime patikrinti, ar bent viena kovariantė yra reikšminga. Kaip galime matyti iš lentelės, tik tokios kovariantės kaip vedybų statusas ir prestižo lygis yra nereikšmingos, o visos kitos – reikšmingos. Dabar pritaikysime pažingsninę regresiją.

Stepwise Selection Summary						
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	F Value	Pr > F
0	Intercept		1	1	0.00	1.0000
1	mentor		2	2	61.45	<.0001
2	gender		3	3	9.50	0.0021
3	kids		4	4	2.16	0.1426
4	married		5	5	2.25	0.1343

20 pav. Kovariančių lentelė po pažingsninės regresijos

Pritaikius stepwise (žingsniuojama pirmyn ir tik pirmyn) regresiją, gaunama lentelė, iš kurios galime matyti, jog kovariantė prestižo lygis buvo pašalinta. AIC reikšmė lygi 1399, pseudo R reikšmė – 0.1.

Dabar skaitiškai įvertinkime regresorių įtaką.

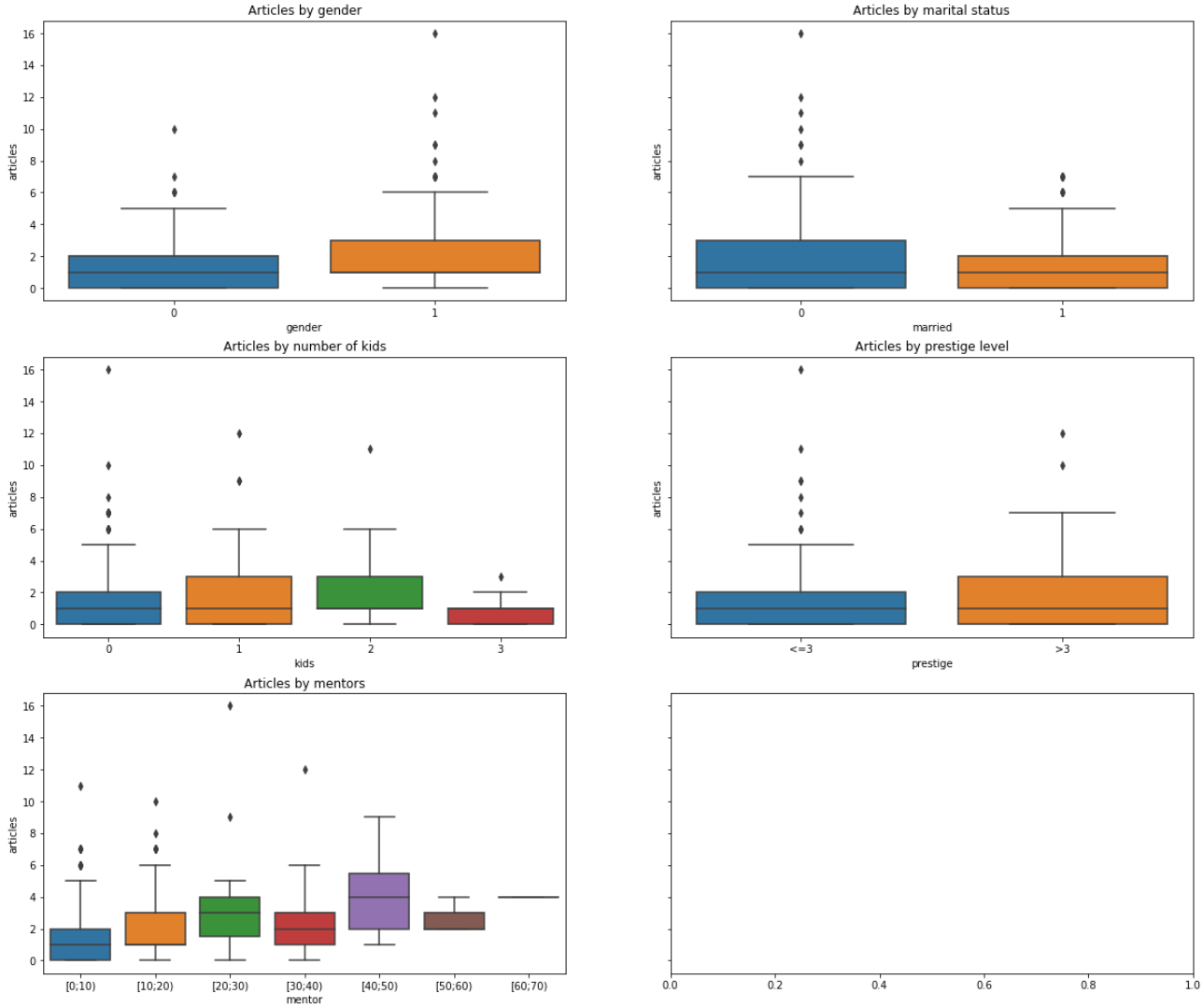
Obs	Parameter	gender	married	Estimate	StdErr	expb
1	gender female	female		-0.2676	0.08499	0.76518
2	gender male	male		0	.	1.00000
3	married no		no	-0.1628	0.09623	0.84977
4	married yes		yes	0	.	1.00000
5	kids			-0.1266	0.06033	0.88108
6	prestige			0.02855	0.04251	1.02896
7	mentor			0.02695	0.004051	1.02732

21 pav. Koeficientai ir jų eksponentės

Iš lentelės matome, jog būnant vyru didėja išleistų straipsnių vidurkis nei būnant moterimi. Jei vyras nevedęs, moteris netekėjusi, mažėja publikuotų straipsnių vidurkis. Padidėjus vaikų skaičiui šeimoje mažėja išleistų straipsnių vidurkis, o padidėjus mentoriaus išleistų straipsnių skaičiui padidėtų ir doktoranto išleistų straipsnių vidurkis.

2.3 Modelis įvykių skaičiui aprašyti naudojant Python

Iš pradžių atliekame pirminę duomenų analizę – kiekvienam požymiui nubrėžiame stačiakampes diagramas. Iš pradžių reikšmės buvo perkoduotos: vyras – 1, moteris – 0, nevedęs/netekėjusi – 1, vedęs/ištekėjusi – 0. Prestižas buvo padalintas į du intervalus: mažiau nei 3 ir daugiau už 3. Taip pat mentorių publikuoti straipsniai buvo padalinti 7 intervalus.



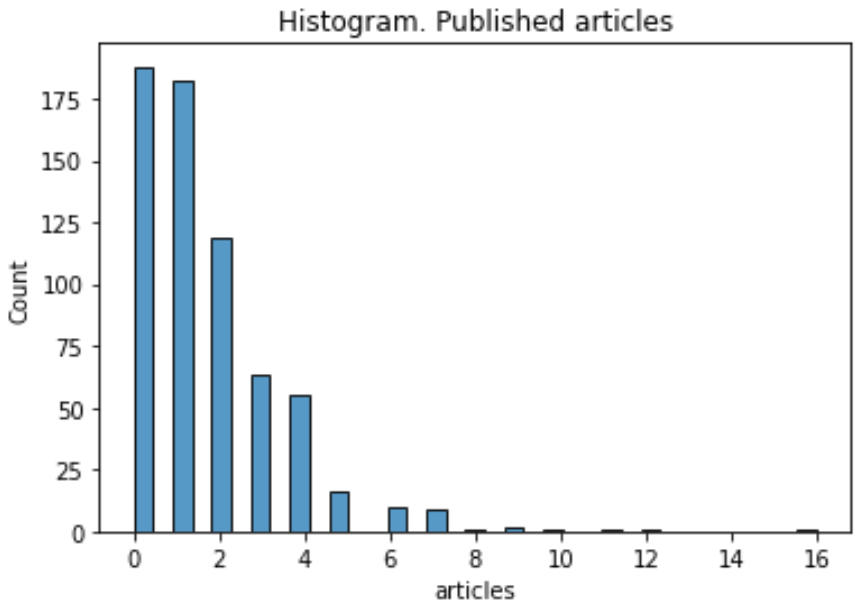
22 pav. Stačiakampės diagramos

Iš stačiakampių diagramų (22 pav.) matome, jog daugiau straipsnių parašo vyrai, vedę vyrai arba ištekėjusios moterys. Mažiausiai straipsnių publikavo daugiausiai vaikų turintys asmenys.

Toliau brėžiame histogramą ir skaičiuojame vidurkį bei dispersiją. Jei vidurkis lygus dispersijai, taikysime Puasono regresiją.

14 lentelė. Vidurkis ir dispersija

Vidurkis	Dispersija
1.71	3.49



23 pav. Publikuotų straipsnių histograma

Matome, jog vidurkis dispersijai nelygus. Taigi, greičiausiai Puasono regresijos modelis netiks.

Pritaikius Puasono regresijos modelį, matome, jog liekanų deviacijos ir laisvės laipsnių skirtumas yra:

$$\frac{1129.5}{643} = 1.75.$$

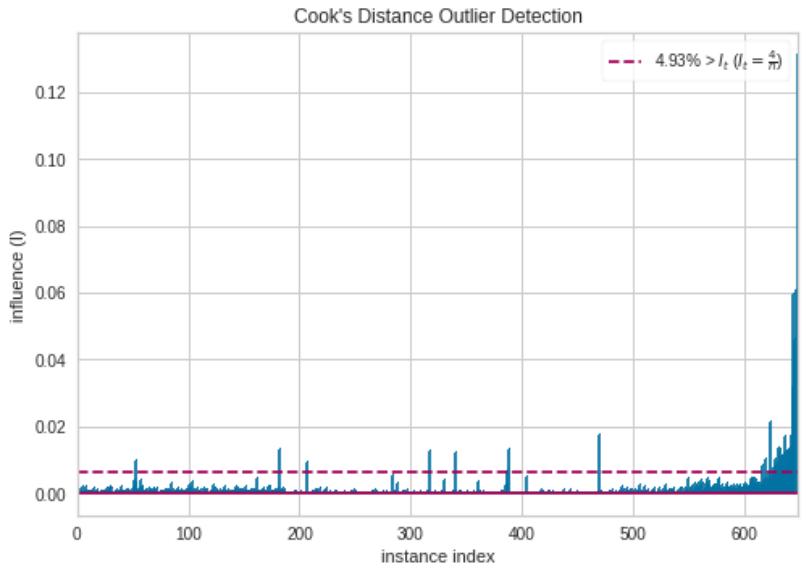
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                articles   No. Observations:                  649
Model:                            GLM   Df Residuals:                      643
Model Family:                 Poisson   Df Model:                            5
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1164.3
Date:                Sat, 18 Mar 2023   Deviance:                       1129.5
Time:                        21:02:32   Pearson chi2:                 1.14e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1696
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1743      0.110      1.584      0.113      -0.041       0.390
gender[T.1]    0.2755      0.065      4.213      0.000       0.147       0.404
married[T.1]  -0.1538      0.073     -2.095      0.036      -0.298      -0.010
kids          -0.1268      0.046     -2.758      0.006      -0.217      -0.037
prestige       0.0192      0.032      0.597      0.551      -0.044       0.082
mentor         0.0244      0.003      9.565      0.000       0.019       0.029
==============================================================================
```

24 pav. Puasono regresijos modelis

Tai dar kartą įrodo, jog Puasono regresijos modelis šiems duomenims nėra tinkamas.

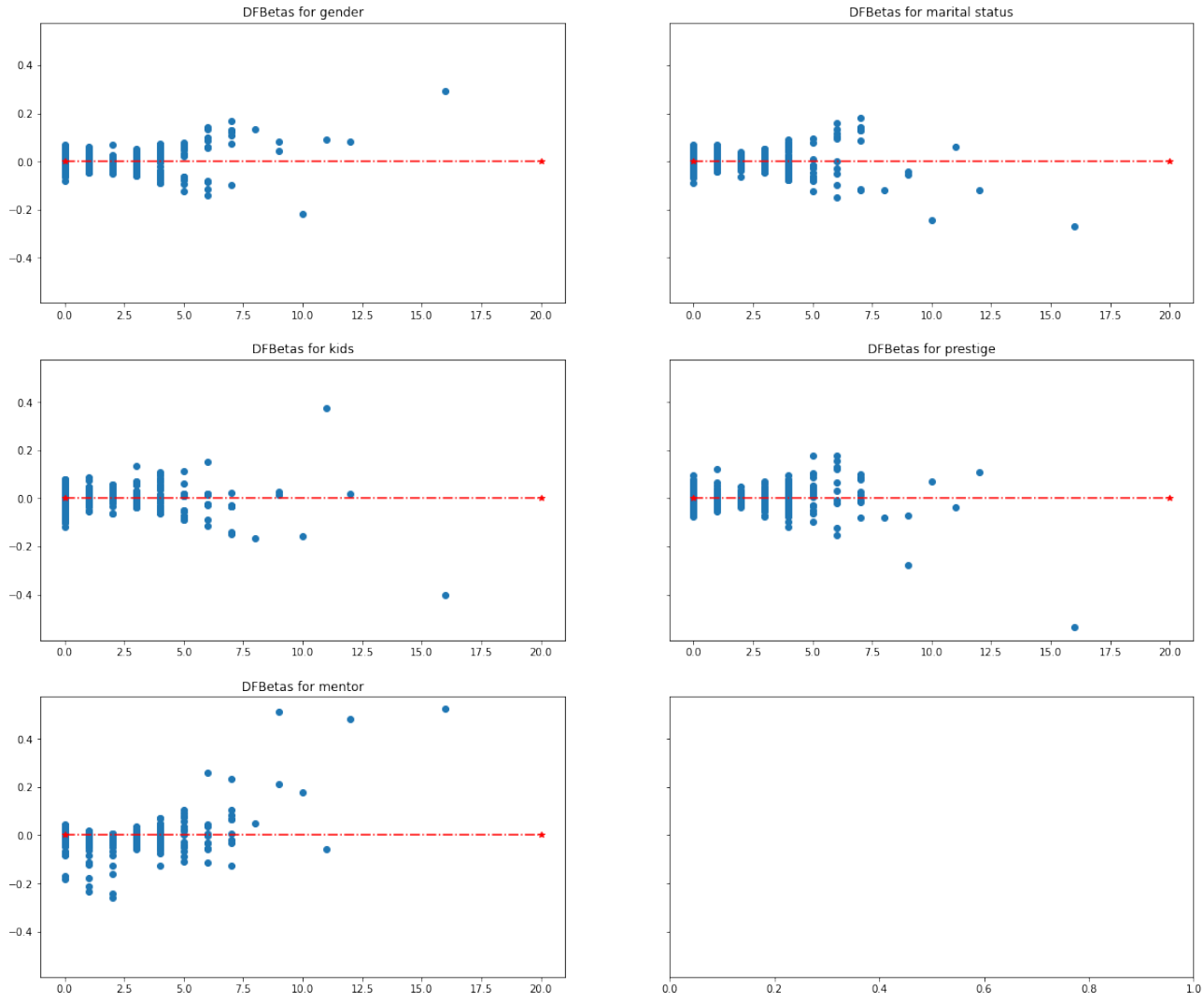
Galiausiai, naudodami dispersijų testą (angl. *overdispersion test*) gauname $\alpha = 1.78$. Kadangi $\alpha > 0$, naudosime neigiamą binominį modelį (NBI).

Toliau turime patikrinti ar nėra išskirčių. Nubrėžus grafiką (25 pav.) matome, jog Kuko matas neviršija vieneto – išskirčių nėra.



25 pav. Kuko matas

Išskirtims taip pat nubrėžiame DFBet‘ų grafikus.



26 pav. DFBet’ų grafikai

Iš grafiko (26 pav.) matome, jog kiekvienam požymiui reikšmės yra intervale $[-1,1]$, vadinasi, išskirčių nėra.

Toliau tikriname multikolinearumą.

15 lentelė. VIF kriterijus

Kovariantės	**VIF**
Lytis	1,13
Ar vedęs	1,31
Vaikų skaičius	1,33
Prestižas	1,09
Mentorius	1,09

Iš lentelės matome, jog nei vienos kovariantės VIF kriterijus neviršija 4, todėl multikolinearumo problemos nėra.

Pritaikius neigiamą binominį modelį matome, jog nereikšmingos kovariantės yra: prestižas ir vedybinis statusas.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                articles   No. Observations:                  649
Model:                            GLM   Df Residuals:                      643
Model Family:        NegativeBinomial   Df Model:                            5
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1134.3
Date:                Sat, 18 Mar 2023   Deviance:                       505.21
Time:                        21:06:35   Pearson chi2:                     427.
No. Iterations:                     7   Pseudo R-squ. (CS):            0.06677
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1038      0.181      0.573      0.567      -0.252       0.459
gender[T.1]    0.2633      0.107      2.460      0.014       0.053       0.473
married[T.1]  -0.1678      0.121     -1.385      0.166      -0.405       0.070
kids          -0.1265      0.076     -1.672      0.094      -0.275       0.022
prestige       0.0330      0.054      0.617      0.537      -0.072       0.138
mentor         0.0281      0.005      5.507      0.000       0.018       0.038
==============================================================================
```

27 pav. Neigiamas binominis modelis

Taigi, pirmiausia pašaliname prestižą.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                articles   No. Observations:                  649
Model:                            GLM   Df Residuals:                      644
Model Family:        NegativeBinomial   Df Model:                            4
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1134.5
Date:                Sun, 19 Mar 2023   Deviance:                       505.59
Time:                        07:50:35   Pearson chi2:                     427.
No. Iterations:                     6   Pseudo R-squ. (CS):            0.06623
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1958      0.107      1.838      0.066      -0.013       0.405
gender[T.1]    0.2649      0.107      2.475      0.013       0.055       0.475
married[T.1]  -0.1585      0.120     -1.316      0.188      -0.394       0.077
kids          -0.1268      0.076     -1.676      0.094      -0.275       0.022
mentor         0.0288      0.005      5.883      0.000       0.019       0.038
==============================================================================
```

28 pav. Neigiamas binominis modelis pašalinus prestižą

Pašalinus prestižą kita nereikšminga kovariante tampa vedybinis statusas, tačiau jį pašalinus padidėja AIC reikšmė, todėl šią kovariantę modelyje paliekame. Taip pat liekanų deviacijos ir laisvės laipsnių santykis yra 0.78. Vis dėlto, gauta maža pseudo R kvdarato reikšmė – 0.06, vadinasi šis modelis nėra tinkamas prognozavimui.

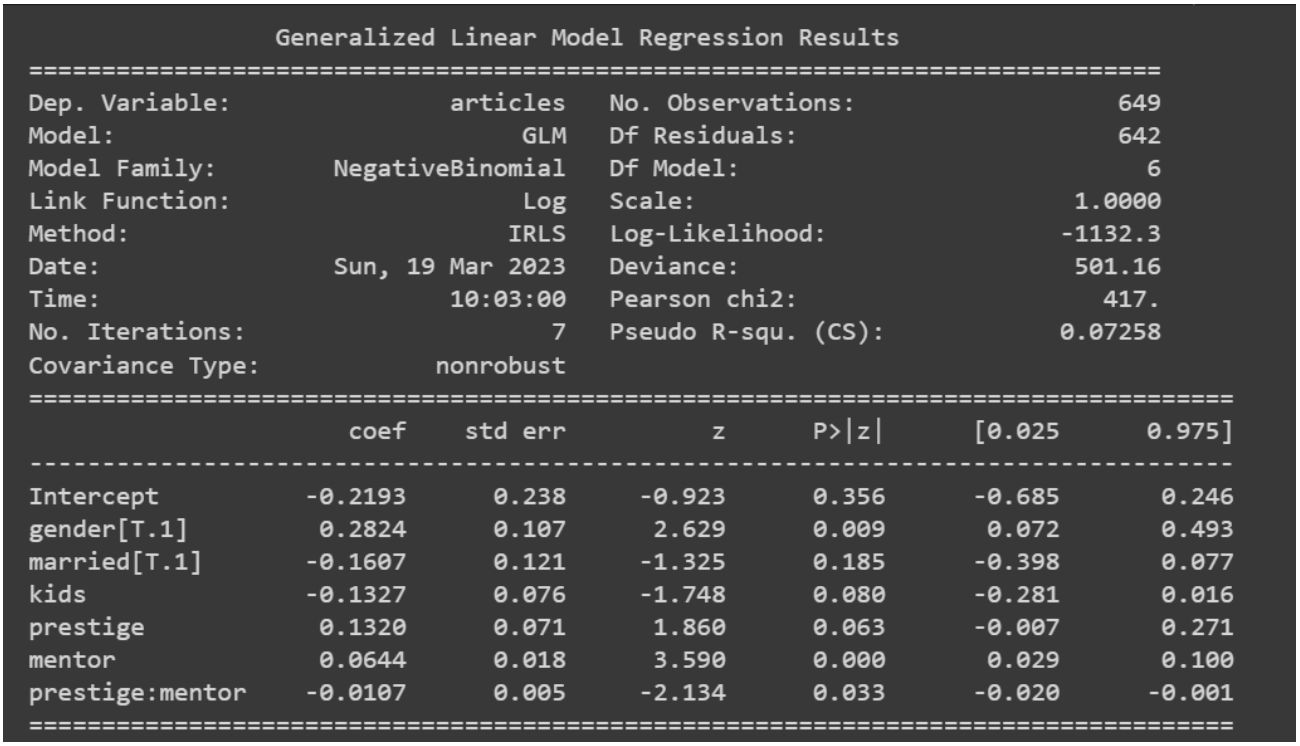
Toliau skaičiavome koeficientų eksponentes.

16 lentelė. Koeficientų eksponentės

Kovariantė	Lytis (vyras)	Nevedęs/netekėjusi	Vaikai	Mentorius
Koeficientas	0.26	- 0.16	- 0.13	0.02
Koeficiento eksponentė	1.3	0.85	0.88	1.03

Iš lentelės (16 lentelė) matome, jog vyrai išleidžia vidutiniškai daugiau straipsnių (30 %). Jei vyras ar moteris yra nevedęs/netekėjusi išleistų straipsnių vidurkis mažėja (15 %). Išleistų straipsniu vidurkis taip pat mažėja esant daugiau vaikų šeimoje (12 %). Mentoriaus išleistų straipsnių kiekis didina doktoranto išleistų straipsnių vidurkį (3 %).

Taip pat patikrinome, ar kuri nors sąveika yra statistiškai reikšminga. Radome prestižas ir mentorius.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                articles   No. Observations:                  649
Model:                            GLM   Df Residuals:                      642
Model Family:        NegativeBinomial   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1132.3
Date:                Sun, 19 Mar 2023   Deviance:                       501.16
Time:                        10:03:00   Pearson chi2:                     417.
No. Iterations:                     7   Pseudo R-squ. (CS):            0.07258
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept         -0.2193      0.238     -0.923      0.356      -0.685       0.246
gender[T.1]        0.2824      0.107      2.629      0.009       0.072       0.493
married[T.1]      -0.1607      0.121     -1.325      0.185      -0.398       0.077
kids              -0.1327      0.076     -1.748      0.080      -0.281       0.016
prestige           0.1320      0.071      1.860      0.063      -0.007       0.271
mentor             0.0644      0.018      3.590      0.000       0.029       0.100
prestige:mentor   -0.0107      0.005     -2.134      0.033      -0.020      -0.001
==============================================================================
```

29 pav. Neigiamas binominis modelis su sąveika

Įtraukiant į modelį šią sąveiką AIC reikšmė mažesnė, pseudo R kvadrato – didesnė, liekanų deviacijos ir laisvės laipsnių santykio reikšmė tokia pati. Vis dėlto įtraukus sąveiką atsirado multikolinearumo problema, todėl jos į modelį neįtraukėme.

Galiausiai, patikriname kaip tiksliai gautas modelis prognozuoja.

17 lentelė. Stebėtos ir prognozuotos reikšmės

Stebėta reikšmė	**Prognozuota reikšmė**
0	1.9
0	1.43
4	1.86
5	1.87

Iš lentelės (17 lentelė) matome, jog modelis prognozuoja netiksliai. Nors liekanų deviacijos ir laisvės laipsnių santykis rodė, jog modelis tinkamas, maža pseudo R kvadrato reikšmė indikuoja, jog modelis nebus tinkamas prognozavimui.

Galutinis modelis atrodo taip:

$$\ln(publikuoti\ straipsniai) = 0{,}195 + 0{,}264 \cdot lytis\ (vyras) - 0{,}158 \cdot vedybinis\ statusas\ (nevedęs) - 0{,}127 \cdot vaikai + 0{,}028 \cdot mentoriaus\ straipsnių\ skaičius.$$

3. IŠVADOS

Iš pirminės duomenų analizės sužinojome, jog vedę / ištekėję asmenys mokymo imtyje sudaro 66 proc. Dažniausias vaikų iki 6 metų skaičius yra 0 arba 1. Dažniausias studijų programos prestižo lygis yra nuo 2,5 iki 3 imtinai, o mažesnio nei 0,5 prestižo lygio studijų programos imtyje nebuvo. Daugiausiai mentoriai buvo išleidę straipsnių nuo 0 iki 20 (imtinai abu galai intervalo). Taip pat pastebėjome, jog dažniausiai vyrai publikuoja daugiau straipsnių. Daugiau išleidžia straipsnių vedę vyrai arba ištekėjusios moterys. Mažiausiai straipsnių per 3 metus išleido didžiausią šeimą turintys asmenys. Studijų programos prestižas nedaro didelės įtakos straipsnių publikavimo skaičiui. Daugiausiai straipsnių publikavo tie asmenys, kurių mentorius publikavo nuo 40 imtinai iki 50 straipsnių.

Turimi duomenys tenkino multikolinearumo ir išskirčių nebuvimo prielaidas, nors mokymo imtyje turėjome du įtartinus stebėjimus. Nustatėme, kad duomenis gerai aprašo neigiamas binominis modelis. Perteklinės nulių problemos neturėjome.

Gavome modelį:

$$\ln(publikuoti\ straipsniai) = 0{,}205 + 0{,}269 \cdot lytis\ (vyras) - 0{,}155 \cdot vedybinis\ statusas\ (nevedęs) - 0{,}127 \cdot vaikai + 0{,}028 \cdot mentoriaus\ straipsnių\ skaičius,$$

kurio AIC 2230,625 (su SAS 1399), BIC 2257,504, liekanų deviacijos ir liekanų laisvės laipsnių santykis 1,110, pseudo R kvadratas 0,090 bei RMSE 2,005. Modelis nėra tinkamas prognozavimui, tačiau yra galima interpretacija: būnant vyru apie 31 proc. padidėja išleistų straipsnių vidurkis, o būnant nesusituokus apie 15 proc. mažėja išleistų straipsnių vidurkis. Padidėjus vaikų skaičiui šeimoje vienu vienetu apytiksliai 12 proc. sumažėtų išleistų straipsnių vidurkis, o padidinus vienu vienetu išleistų mentoriaus straipsnių skaičių 3 proc. padidėtų doktorantų išleistų straipsnių vidurkis.