VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Tiesiniai modeliai

Laboratorinis darbas

Atliko: 3 kurso 2 grupės studentai:

Matas Amšiejus

Salvija Račkauskaitė

Sandra Macijauskaitė

Darbo vadovė: doc. dr. Rūta Levulienė

Vilnius, 2021

TURINYS

ĮVADAS

Šiame laboratoriniame darbe analizuosime 2014 ir 2015 metų filmų duomenis. Tikslas – nustatyti ar skiriasi filmų reitingai pagal žanrus taikant kovariacinę analizę. Laboratorinio darbo uždavinį įgyvendinti pasitelksime R ir SAS programavimo kalbas.

1. DUOMENYS

1.1.Duomenys

Duomenų rinkinį pasirinkome iš viešai prieinamo duomenų šaltinio „UCI Machine Learning Repository“ (nuoroda šaltiniuose). Duomenyse yra surinkta informacija apie 2014 – 2015 metų filmus.

1.2.Duomenų aprašymas

Tyrimo imtį sudarė 223 stebėjimai. Duomenų stulpeliai:

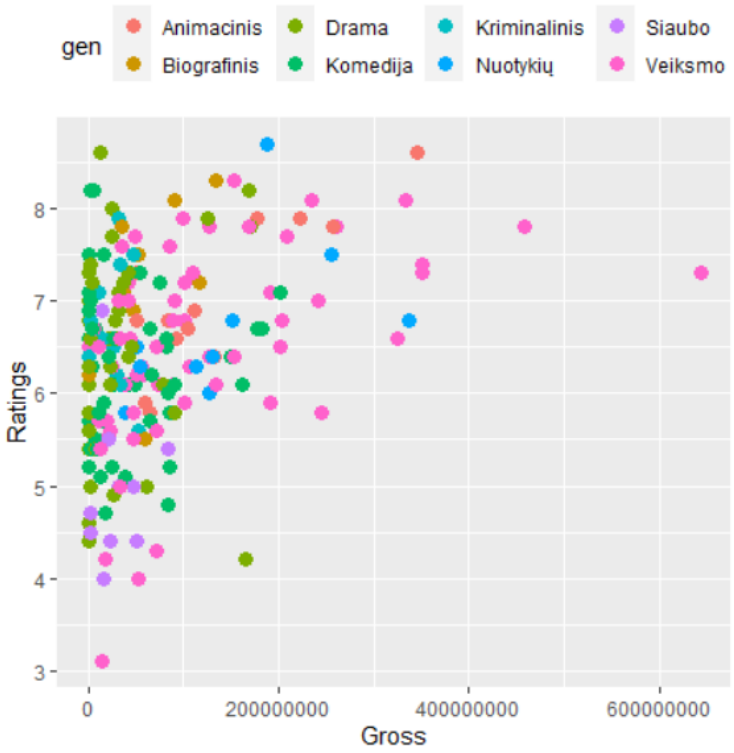
1. Ratings – įvertinimai (IMDB platformoje);
2. Gross – bendras filmo uždarbis;
3. Genre – filmo žanras;
4. Budget – biudžetas;
5. Screens – kino seansai
6. Views – peržiūros;
7. Likes – patikimai;
8. Dislikes – nepatikimai;
9. Comments – komentarai;
10. Movie – filmo pavadinimas;
11. Year – metai.

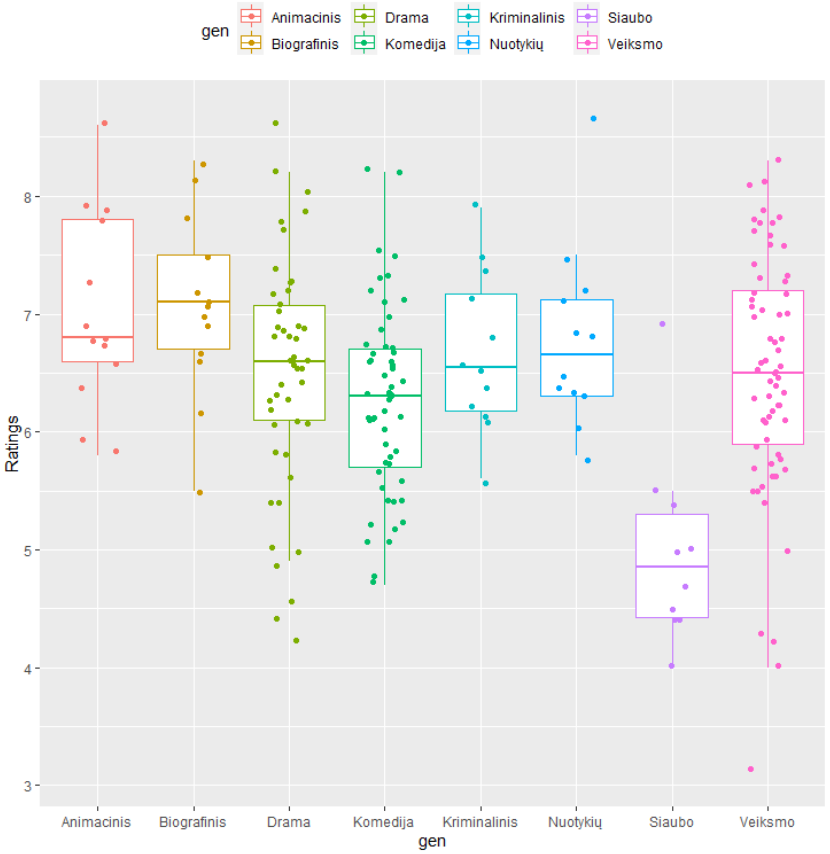
2. ATLIKTAS TYRIMAS

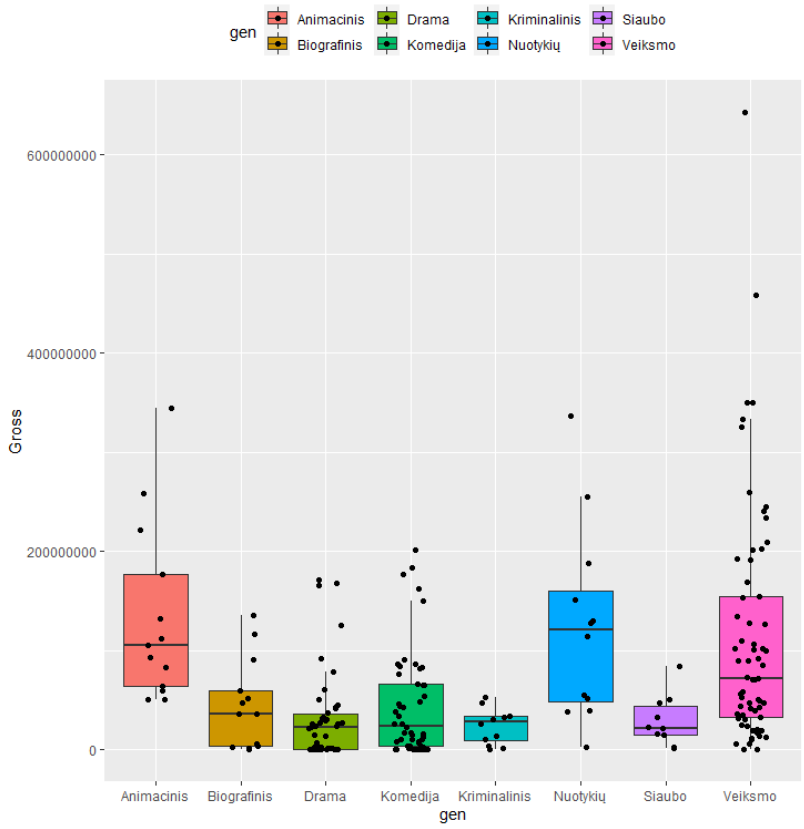
Atlikome kovariacinę analizę su R ir SAS programavimo kalbomis. Priklausomą kintamąjį pasirinkome filmo įvertinimą (Ratings), faktorių pasirinkome žanrą (Genre) ir kovariantę – filmo uždarbį (Gross). Tyrime naudosime reikšmingumo lygmenį alpha = 0,05.

2.1.Bendra kovariacinės analizės eiga

Pirmiausia nuskaitome duomenis iš *xlsx* failo, atsirenkame reikiamus stulpelius, filmo skaitinius žanrus pakeičiame į kategorinius. Braižome sklaidos grafiką bei stačiakampes diagramas.


gen
Animacinis
Biografinis
Drama
Komedija
Kriminalinis
Nuotykių
Siaubo
Veiksmo
Ratings
Gross
0
200000000
400000000
600000000
3
4
5
6
7
8


gen
Animacinis
Biografinis
Drama
Komedija
Kriminalinis
Nuotykių
Siaubo
Veiksmo
Ratings
gen
3
4
5
6
7
8



Kadangi gross reikšmės labai didelės, keičiame matavimo vienetus į milijonus (t. y. Padaliname iš milijono). Išmetame dvi didžiausias išskirtis su Gross kovariante. Pereiname prie prielaidų tikrinimo.

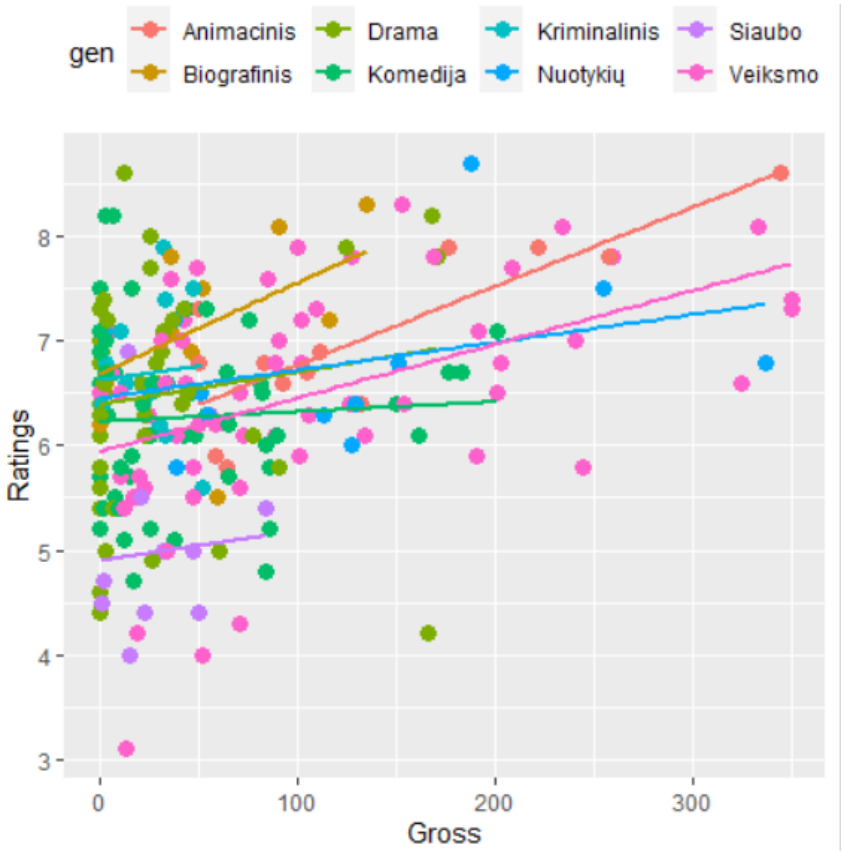
Pirmiausia tikriname hipotezę dėl krypties koeficientų lygybės.

```
ANOVA Table (type III tests)

       Effect      SSn     SSd DFn DFd       F         p p<.05   ges
1 (Intercept) 1953.459 153.564   1 207 2633.202 1.10e-119     * 0.927
2       Gross    2.072 153.564   1 207    2.794  9.60e-02       0.013
3         gen   12.495 153.564   7 207    2.406  2.20e-02     * 0.075
4   Gross:gen    3.928 153.564   7 207    0.756  6.25e-01       0.025
```

Matome, kad Gross::gen p reikšmė = 0,625 yra daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Krypties koeficientai yra lygūs.

Nubraižome grafiką norėdami patikrinti, kad yra tiesiniai sąryšiai tarp Ratings (įvertinimai) ir Gross (uždarbis).



Matome tiesinius sąryšius tarp Ratings ir Gross. Toliau tikriname, kad liekanos pasiskirsčiusios pagal normalųjį skirstinį. Tam naudosime Shapiro – Wilk normalumo testą.

```
        Shapiro-Wilk normality test

data:  resid(aov(Ratings ~ gen + Gross, data = movies))
W = 0.9857, p-value = 0.02428
```

Gauname, kad p reikšmė mažiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinę hipotezę atmetame. Liekanos netenkina normalumo prielaidos. Dėl to naudosime Box-Cox transformaciją.

Pirmiausia ieškome optimalios korekcijos modeliui. Gauname, kad lambda = 1,636364, dėl to ją suapvalinsime iki sveikojo skaičiaus. Sukuriame naują kintamąjį rt (įvertinimai kvadratu). Tikriname su nauju kintamuoju, kad liekanos pasiskirsčiusios pagal normalųjį skirstinį.

```
        Shapiro-Wilk normality test

data:  resid(aov(rt ~ gen + Gross, data = movies))
W = 0.99228, p-value = 0.2927
```

Pritaikius optimalią korekciją modeliui gauname, kad p reikšmė yra daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Liekanos tenkina normalumo prielaidą.

Dabar tikrinsime, kad dispersijos tarp žanrų grupių yra lygios. Taikysime Leveni testą.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   7  0.8145 0.5761
      215
```

Gauname, kad p reikšmė daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Vadinasi, dispersijos tarp žanrų grupių yra lygios.

Dar kartą patikrinsime dispersijų lygybę naudodami koreguotas Ratings reikšmes. Modifikuosime reikšmes pagal pilną mūsų modelį (t. y. įtraukiant ir Gross (kovariantę)). Tam reikės krypties koeficiento β. Ją gauname iš įvertinio prie Gross.

Pastaba: tam, kad rezultatai sutaptų su SAS, mes pakeičiame intercept į veiksmo filmus.

Pirma sukuriame modelį.

```
Call:
lm(formula = rt ~ Gross + gen1, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-33.342  -6.452  -1.114   7.147  31.994

Coefficients:
                Estimate Std. Error t value    Pr(>|t|)
(Intercept)     36.81293    1.74628  21.081     < 2e-16 ***
Gross            0.05897    0.01078   5.468 0.000000126 ***
gen1Animacinis   5.30623    3.32787   1.594    0.112304
gen1Biografinis 11.13953    3.36120   3.314    0.001079 **
gen1Drama        4.38007    2.23396   1.961    0.051214 .
gen1Komedija     0.50520    2.10566   0.240    0.810617
gen1Kriminalinis 6.87812    3.51909   1.955    0.051942 .
gen1Nuotykių     2.44728    3.42910   0.714    0.476202
gen1Siaubo     -13.13608    3.77507  -3.480    0.000608 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.86 on 214 degrees of freedom
Multiple R-squared:  0.2567,    Adjusted R-squared:  0.2289
F-statistic: 9.238 on 8 and 214 DF,  p-value: 0.00000000006488
```

Gauname, kad β reikšmė prie Gross yra 0.05897. Tada kuriame koreguotas įvertinimų reikšmes naudodami formulę $Z = Y - \beta(X - \bar{X})$ ir atliekame Leveni testą.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   7  0.7441 0.6348
      215
```

Gauname, kad p reikšmė padidėjo, kai pakoregavome modelį. Gauta p = 0,6348 reikšmė yra daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Vadinasi, pagal koreguotą modelį dispersijos tarp žanrų grupių yra lygios.

Toliau pažiūrime koreguotus įvertinimų vidurkius pagal modelį.

	Gross	gen	emmean	se	df	conf.low	conf.high	method
	<dbl>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	65.2	Animacinis	46.0	3.10	214	39.8	52.1	Emmeans test
2	65.2	Biografinis	51.8	3.02	214	45.8	57.7	Emmeans test
3	65.2	Drama	45.0	1.64	214	41.8	48.3	Emmeans test
4	65.2	Komedija	41.2	1.50	214	38.2	44.1	Emmeans test
5	65.2	Kriminalinis	47.5	3.17	214	41.3	53.8	Emmeans test
6	65.2	Nuotykių	43.1	3.20	214	36.8	49.4	Emmeans test
7	65.2	Siaubo	27.5	3.46	214	20.7	34.3	Emmeans test
8	65.2	Veiksmo	40.7	1.42	214	37.9	43.5	Emmeans test

Iš lentelės matome, kad prasčiausiai yra vertinami siaubo filmai, o geriausiai – biografiniai. Darome porinius žanrų vertinimų vidurkių palyginimus. Atrenkame tik tas poras, kurios tarpusavyje statistiškai reikšmingai skiriasi.

	term	.y.	group1	group2	df	statistic	p	p.adj	p.adj.signif
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Gross*gen	rt	Animacinis	Siaubo	214	3.92	0.000120	0.00335	**
2	Gross*gen	rt	Biografinis	Drama	214	1.98	0.0490	1	ns
3	Gross*gen	rt	Biografinis	Komedija	214	3.17	0.00175	0.0489	*
4	Gross*gen	rt	Biografinis	Siaubo	214	5.31	0.000000271	0.000000760	****
5	Gross*gen	rt	Biografinis	Veiksmo	214	3.31	0.00108	0.0302	*
6	Gross*gen	rt	Drama	Siaubo	214	4.62	0.00000652	0.000183	***
7	Gross*gen	rt	Komedija	Siaubo	214	3.65	0.000334	0.00935	**
8	Gross*gen	rt	Kriminalinis	Siaubo	214	4.30	0.0000254	0.000713	***
9	Gross*gen	rt	Nuotykių	Siaubo	214	3.27	0.00124	0.0346	*
10	Gross*gen	rt	Siaubo	Veiksmo	214	-3.48	0.000608	0.0170	*

Matome, kad beveik visi filmų žanrų vertinimai reikšmingai skiriasi nuo biografinių ir siaubo filmų.

Galutinis kovariacinės analizės modelis.

```
Call:
lm(formula = rt ~ Gross + gen1, data = movies)

Residuals:
    Min      1Q  Median      3Q     Max
-33.342  -6.452  -1.114   7.147  31.994

Coefficients:
                 Estimate Std. Error t value    Pr(>|t|)
(Intercept)      36.81293    1.74628  21.081     < 2e-16 ***
Gross             0.05897    0.01078   5.468 0.000000126 ***
gen1Animacinis    5.30623    3.32787   1.594    0.112304
gen1Biografinis  11.13953    3.36120   3.314    0.001079 **
gen1Drama         4.38007    2.23396   1.961    0.051214 .
gen1Komedija      0.50520    2.10566   0.240    0.810617
gen1Kriminalinis  6.87812    3.51909   1.955    0.051942 .
gen1Nuotykių      2.44728    3.42910   0.714    0.476202
gen1Siaubo      -13.13608    3.77507  -3.480    0.000608 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.86 on 214 degrees of freedom
Multiple R-squared:  0.2567,    Adjusted R-squared:  0.2289
F-statistic: 9.238 on 8 and 214 DF,  p-value: 0.00000000006488
```

Iš modelio matome, kad R-squared (determinacijos koeficientas) = 0,2567, t. y. apie 26 % duomenų sklaidos galima nusakyti mūsų modeliu (kokią dalį duomenų sklaidos lemia skirtumai tarp grupių su skirtumais grupių viduje).

2.2Palyginimas su ANOVA

```
             Df Sum Sq Mean Sq F value     Pr(>F)
gen1          7   5208   743.9   5.533 0.00000713 ***
Residuals   217  29179   134.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
 Siaubo-Veiksmo   Siaubo-Animacinis  Siaubo-Biografinis        Siaubo-Drama     Siaubo-Komedija Siaubo-Kriminalinis
  0.00027599716       0.00002485128       0.00001465533       0.00052491721       0.00844638277       0.00250363675
Siaubo-Nuotykių
  0.00077911684
```

Matome, kad skirtumai tarp žanrų yra reikšmingi, tačiau atlikus porinius palyginimus gauname, kad reikšmingai skiriasi tik siaubo žanras.

IŠVADOS

Atlikus kovariacinę analizę nustatėme, kad filmų reitingai statistiškai reikšmingai skiriasi tarp žanrų. Geriausiai vertinami filmai yra biografiniai, o prasčiausiai – siaubo. Atlikus porinius palyginimus gauname, kad tik šie žanrai reikšmingai skiriasi nuo likusių.

ŠALTINIAI

[1] „UCI Machine Learning Repository“ tinklapis. Tema: CSM (Conventional and Social Media Movies) Dataset 2014 and 2015 Data Set. Prieiga per internetą: https://archive.ics.uci.edu/ml/datasets/CSM+%28Conventional+and+Social+Media+Movies%29+Dataset+2014+and+2015?fbclid=IwAR39maXqXXEQzXygR17yJNNwdITR_NlwfESJApThmlR4sI7j86MCsa53XR0