

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 4

1. DUOMENYS 5

1.1.Duomenys 5

1.2.Duomenų aprašymas 5

2. ATLIKTAS TYRIMAS 5

2.1.Bendra kovariacinės analizės eiga 5

IŠVADOS 12

ŠALTINIAI 13

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.
12. 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.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

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.

Text

Description automatically generated with medium confidencePirmiausia tikriname hipotezę dėl krypties koeficientų lygybės.

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.

Chart, scatter chart

Description automatically generated 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.

Text, letter

Description automatically generated

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.

Text

Description automatically generated

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.

A picture containing text

Description automatically generated

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.

A screenshot of a computer

Description automatically generated with low confidence

Gauname, kad β reikšmė prie Gross yra 0.05897. Tada kuriame koreguotas įvertinimų reikšmes naudodami formulę ir atliekame Leveni testą.

Text

Description automatically generated with medium confidence

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.

A screenshot of a computer

Description automatically generated with low confidence

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.

A screenshot of a computer

Description automatically generated with low confidence

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

Galutinis kovariacinės analizės modelis.

Table

Description automatically generated

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).

* 1. Palyginimas su ANOVA

FOOFOOFOFOF

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>