

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 tiesinės regresijos eiga 5

2.2Tiesinės regresijos veiksmai ir rezultatai 5

IŠVADOS 6

ŠALTINIAI 7

ĮVADAS

Šiame laboratoriniame darbe analizuosime drabužių siuvyklos duomenis. Taikant tiesinės regresijos modelį bandysime nustatyti kaip priklauso gamyklos produktyvumas nuo įvairių faktorių. Laboratorinio darbo uždavinį įgyvendinti pasitelksime R, SAS ir Python programavimo kalbas.

1. DUOMENYS

1.1.Duomenys

Duomenų rinkinį pasirinkome iš viešai prieinamo duomenų šaltinio „UCI Machine Learning Repository“ (nuoroda šaltiniuose). Duomenyse yra žymimi įvairūs drabužių gamyklos ir jos darbuotojų rodikliai.

1.2.Duomenų aprašymas

Duomenų rinkinį sudarė 1197 stebėjimai su 15 atributų. Modelyje naudosime šiuos:

1. Targeted\_productivity – gamyklos valdžios nustatyti produktyvumo tikslai ;
2. Smv – užduočiai skiriamas laikas minutėmis;
3. Wip – kiekis trukstamų dalių produktui;
4. Over\_time – viršvalandžiai;
5. Incentive – piniginė paskata (BDT (Bangladesh taka) valiuta);
6. No\_of\_workers – darbuotojų skaičius;
7. Day – savaitės diena;
8. Actual\_productivity – tikrasis produktyvumas tą dieną.
9. ATLIKTAS TYRIMAS

Atlikome daugelio kintamųjų tiesinę regresiją su SAS, R ir Python programavimo kalbomis. Priklausomą kintamąjį pasirenkame Actual\_productivity (darbuotojų produktyvumas). Tyrime naudosime reikšmingumo lygmenį alpha = 0,05.

2.1.Bendra tiesinės regresijos eiga

Pirmiausia nuskaitome duomenis ir atrenkame mus dominančius stulpelius. Tada tikriname visų kintamųjų sklaidos diagramas. Pagal jas atsifiltruojame dalį duomenų. Taip pat, kad išvengtume daugiau išskirčių, pasirenkame tik tuos įrašus, kur darbuotojų produktyvumas stipriai nesiskyrė nuo vadovų iškelto tikslo (tenkino bent 90 %).

Scatter chart

Description automatically generatedScatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Table

Description automatically generatedTikriname, kaip priklausomas kintamasis (darbuotojų produktyvumas) koreliuoja su kovariantėmis.

Matome, kad koreliacija pakankamai stipri su parinktomis kovariantėmis, todėl galime bandyti nustatyti tiesinės regresijos modelį.

Sukuriame tiesinės regresijos modelį. Tikriname išskirtis pagal Kuko (Cook‘s D) ir stjudentizuotų paklaidų (R student) kriterijus.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Matome, kad pagal Kuko kriterijų išskirčių nėra. Tačiau pagal Rstudent jų turime nemažai. Šaliname išskirtis. Pastaba: šalinsime ne visas iš karto, atsižvelgiant į tai, kaip modelis kis po individualios išskirties pašalinimo.

Pašalinus visas išskirtis, gauname labai gerą R square reikšmę.

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

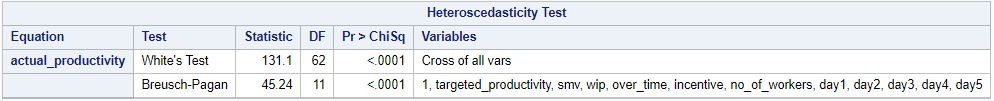
Tikriname, kad paklaidos pasiskirsčiusios pagal normalųjį skirstinį. Tam naudosime Shapiro – Wilk normalumo testą.

Table

Description automatically generated

Gauname, kad p reikšmė daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Paklaidos tenkina normalumo prielaidą.

Dabar tikrinsime homoskedastiškumo prielaidą (paklaidų dispersijos lygios). Tam naudosime Breusch – Pagan homoskedastiškumo testą.



Gauname, kad p reikšmė yra mažiau už reikšmingumo lygmenį, todėl nulinę hipotezę atmetame. Paklaidų dispersijos yra nevienodos (heteroskedastiškos), todėl naudosime HC0 korekciją. Gauname naujas pataisytas standartines paklaidas bei stulpelių reikšmingumo p reikšmes.

A screenshot of a computer

Description automatically generated with medium confidence

Iš lentelės matome, kad ne visos kovariantės yra reikšmingos. Svarbiausios yra targeted\_productivity, wip, over\_time ir incentive. Modelio R square reikšmė gaunasi labai gera (apie 95 %).

Chart, scatter chart

Description automatically generated

Pagal Rstudent kriterijų išskirčių nebeliko.

Diagram

Description automatically generated with medium confidence

Paklaidų pasiskirtymas primena normalųjį skirstinį, modelio nuspėtos reikšmės išsidėsčiusios ant tiesės.

Tobuliname modelį. Atrenkame tik reikšmingas kovariantes. Tam naudosime pažingsninę regresiją (reikšmingumo lygmuo 0,05).

Table

Description automatically generated

Nustatėme, kad smv yra nereikšminga. Išmetus smv matome, kad savaitės dienos irgi tampa nereikšmingos, todėl pašaliname jas iš modelio.

A screenshot of a computer

Description automatically generated with medium confidence

Pašalinus dienų faktorių (su visais pseudokintamaisiais) turime galutinį tiesinės regresijos modelį. VIF niekur nesiekia 4, todėl daugiau nieko nekeičiame. Galutinė Adjusted R Square reikšmė yra 95,3 % (tokia dalis duomenų paaiškinama tiesinės regresijos modeliu).

Table

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Gauti galutiniai parametrų įvertiniai: β0 0,0432, β1  0,753, β2 0,000017, β3 -0,000002, β4 0,00258, β5 0,00054. Didžiausią įtaką gamyklos produktyvumui daro vadovų tikslai ir piniginė paskata.

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

Diagram

Description automatically generated with medium confidence

KOMENTARAS?? Kvk

IŠVADOS

Atlikus pilną regresinę analizę sukūrėme gana tikslų modelį, pritaikytą nustatyti gamyklos produktyvumą pagal svarbiausias kovariantes. Didžiausią įtaką produktyvumui turėjo valdžios iškelti reikalavimai bei piniginė paskata. Įtakos nedarė savaitės diena ir užduočiai skiriamas laikas minutėmis.

ŠALTINIAI

1. ,,UCI Machine Learning Repository“ tinklapis. Tema: Productivity Prediction of Garment Employees. Prieiga per internetą: <https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees>
2. Heteroskedastiškumo pavyzdžiai su R. Prieiga per internetą: [Dealing with heteroskedasticity; regression with robust standard errors using R (brodrigues.co)](https://www.brodrigues.co/blog/2018-07-08-rob_stderr/)
3. „SAS Help Center“ tinklapis. Prieiga per internetą: [SAS Help Center: SAS Help Center: Welcome](https://documentation.sas.com/doc/en/helpcenterwlcm/1.0/home.htm)