**Aštuntas žingsnis. Modelio patikimumo tikrinimas Paklaidų analizė.**

Prisimename, kad regresinį modelį sudaro regresijos lygtis – tai sisteminė modelio dalis ir paklaidos -atsitiktinė modelio dalis. t.y.,

 Yi =f(X1; X2; …X1k) + ɛi

 Sisteminė dalis Atsitiktinė dalis

Kaip ir buvo minėta, adekvatus ir patikimas modelis bus tuomet, kai modelyje dominuos sisteminė dalis, t.y., kai ji bus statistiškai reikšminga. Taip pat modelyje yra labai svarbu patikrinti ir modelio paklaidas. Sudarytas modelis bus adekvatus ir patikimas tuomet, kai o paklaidos bus atsitiktinis dydis, kurio kitimas nepriklauso nuo jokių dėsningumų. Kitais žodžiais tariant modelio paklaidos turi tenkinti klasikines modelio prielaidas, kurios yra nurodytos konspekto 19psl.o, šioje lentelėje pateikiamos tik tos, kurios susijusios su modelio paklaidomis.

|  |  |
| --- | --- |
| **Prielaida** | **Prielaidos simbolinė išraiška** |
| II. Paklaidų vidurkis lygus nuliui (nulinis vidurkis) | E(εi) = 0 |
| III. Paklaidos neautokoreliuoja (likučių ne autokoreliacijos) , t.y, paklaidos tarpusavyje nėra susijusios ir nestebimi sklaidos dėsningumai. | Cov(εi εj) = 0, ∀i,j / i≠j |
| IV. Paklaidų dispersija yra pastovi (ne heteroskedastiškumas) Didėjant nepriklausomų kintamųjų reikšmėms, priklausomojo kintamojo sklaidos intervalas išlieka pastovus.  | σ2(εi) = konstanta |
| VI. Paklaidos pasiskirsčiusios pagal normalųjį skirstinį (normalumas).  | εi ~ N (0, σ2) |

II prielaida yra visuomet tenkinama, jeigu į modelį yra įtrauktas laisvasis narys β0. Šio nario įtraukimas į modelį užtikrina, kad bus tenkinama II modelio prielaida. Todėl atradus, kad regresijos lygtyje nereikšmingas laisvasis narys, būtina įsitinti, ar jį pašalinus, bus tenkinama II prielaida.

Tai galima padaryti, suskaičiavus EXCEL skaičiuokle regresinį modelį ir apskaičiavus išklotinės 4 lentelėje pateiktų paklaidų vidurkį. Jeigu modelyje be laisvojo nario paklaidų vidurkio reikšmė nėra lygi 0, tuomet verta palikti modelyje laisvąjį narį, nors jis statistiškai ir nėra reikšmingas.

III prielaida reikalauja, kad paklaidos nebūtų koreliuotos, t.y., nepriklausytų vienos nuo kitų. Šios prielaidos netenkinimas nurodo, kad paklaidos nėra atsitiktinis dydis, o jose galima įžvelgti dėsningumus, o tai reiškia, kad modelis nėra sudarytas adekvačiai ir jo paklaidose atsispindi sisteminės dalies elementas.

**Paklaidų autokoreliacija**

Autokoreliacijos problema yra tada, kai modelio paklaidos yra susijusios tarpusavyje, t.y. jų koreliacijos koeficientas nėra lygi 0.

**Kodėl blogai?** Kai sudarytame modelyje yra autokoreliacija, tai:

1. Mažiausių kvadratų metodu (MKM) apskaičiuotas determinacijos koeficientas R2 yra didesnis už tikrąjį.
2. Mažiausių kvadratų metodu (MKM) apskaičiuotas standartinės paklaidos SEbj yra mažesnės.
3. Tikrinant hipotezes negalima naudoti nei t-stjudento nei F kriterijaus, nes gaunamos didesnės negu iš tikrųjų statistikų reikšmės ir galima suklysti, darant išvadas apie veiksnių statistinį reikšmingumą.

**Kokiu būdu nustatyti?** Paklaidų autokoreliacijai nustatyti yra žinomi specialūs statistiniai testai, iš kurių šiame paskaitų konspekte bus pateikti du:

1. Durbin-Watson testas
2. Ženklų sekos kriterijus.

Durbin-Watson testas – tai autokoreliacijos nebuvimo modelio paklaidose hipotezės tikrinimo procedūra:

H0: autokoreliacijos nėra. H1: autokoreliacija yra.

Šiam testui atlikti reikia suskaičiuoti DW d statistiką: kurioje ei – tai paklaida, o ei-1 – tai ankstesnio stebėjimo arba vėluojanti paklaida.

 ,

Apskaičiuota statistika d gali įgauti reikšmes nuo 0 iki 4. Autokoreliacijos nėra, kai d statistika lygi arba artima 2. Norint tiksliai atsakyti į klausimą, ar yra ar nėra paklaidų autokoreliacija, reikia Durbin\_Watson d reikšmių lentelėse (pateiktos dėst. tnklalapyje) rasti kritines :apatinio rėžio - dL irviršutinio rėžio dU reikšmes, priklausomai nuo stebėjimų skaičiaus n (lentelės pirmas stulpelis) ir į modelų įtrauktų veiksnių skaičiaus k (viršutinė lentelės eilutė)

* Autokoreliacijos nėra arba H0 neatmetama , kaidU ≤ d ≤ 4 - dU
* Autokoreliacija yra, kai d ≤ dL (teigiama autokoreliacija) arba d ≥ 4 - dL (neigiama autokoreliacija). Šiuo atveju hipotezė H0 yra atmetama.⇒ H1
* Šiame teste gali būti dvi neapibrėžtumo sritys, dL ≤ d ≤ dU arba 4- dU ≤ d ≤ 4 - dL kuriose griežto atsakymo apie paklaidų autokoreliacija nėra.

Grafinis testo rezultatų pavidalas pateiktas žemiau.

 

Ženklų sekų kriterijus.

Šiam kriterijui reikalinga turėti modelio paklaidas, kurias galima rasti Excel Regression skaičiuoklės Residual Output lentelėje. Šalia paklaidų reikia surašyti ženklus: jei paklaida didesnė už 0, tai “+”, ir jei paklaida mažesnė už 0, tai “–”. Tada reikia suskaičiuoti sekas, o viena seką sudaro vienodi ženklai. Kai ženklas pasikeičia, tai prasideda antra seka. Reikia suskaičiuoti sekų vidurkį ir dispersiją. :

  ir ,

kur n- stebėjimų skaičius, n1 – paklaidų su ”+” ženklu skaičius, n2 –paklaidų su ”-”, skaičius k- sekų skaičius. Tada reikia atlikti hipotezių tikrinimo procedūrą:

H0:  Autokoreliacijos nėra. Sekų skaičius k atsitiktinis, nepriklausomas ir pasiskirstęs pagal normalųjį skirstinį dydis.

HA:  Autokoreliacija yra. Sekų skaičius k nėra atsitiktinis, nepriklausomas ir pasiskirstęs pagal normalųjį skirstinį,

Jei apskaičiuota k (sekų skaičiaus reikšmė) patenka į intervalą:

 ,

 tai H0 atmesti nėra pagrindo, tad su 95 % tikimybe galima teigti, jog autokoreliacijos nėra.

**Kaip išspręsti?** Autokoreliacijos sprendimo būdai:

* Įtraukti naujus veiksnius. arba laiko veiksnį, kuris rodo stebėjimo periodo numerį (1,2,3..it t.t.) Šie veiksniai gali išimti inerciją, jeigu ji yra būdinga nagrinėjamam reiškiniui.
* Į modelį įtraukti vėluojantį priklausomąjį knitamąjį yt-1, , bet šiuo atveju būtina patikrinti ar neatsirado stipri koreliacija tarp yt-1 ir modelio paklaidų ei.
* Peržiūrėti modelio matematinę išraišką. Gal būt šį reiškinį gali geriau atspindėti ne tiesinė, o kita, pvz, logaritminė, eksponentinė ar kita, regresijos lygties forma.
* Transformuoti duomenis. Pavyzdžiui, skaičiuoti ne absoliučius dydžių, o pokyčių regresiją: 

**Autokoreliacijos tikrinimas duonos kainos modelyje**

Pirmiausia patikrinsime ar egzistuoja paklaidų autokoreliacija Durbin Watson testu. Šiam tikslui mums reikia turėti modelio paklaidas. Jos yra pateikiamos EXCEL\_Regression išklotinės ketvirtoje lentelėje RESIDUALS OUTPUT. Pateiktoje lentelėje apačioje trys pirmieji stulpeliai – tai tiesinio modelio Residuals Output išklotinės fragmentas.

|  |  |  |
| --- | --- | --- |
| Autokoreliacijos tikrinimas DW -testu | Ženklų kriterijus |  |
| *Stebėjimo nr.*  | *Y-apskaičiuota Duonos kaina, Lt/kg*  | *Paklaidos ei (Residuals)* | *vėluojančios paklaidos ei-1*  | (ei -e i-1)^2 | (ei )^2 |   | *Paklaidos ei (Residuals)* | Ženklai  |  |
| 1 | 4,4288 | 0,0712 |  |  |  |   | 0,0712 |  + |  |
| 2 | 4,4346 | 0,0754 | 0,0712 | 0,0000 | 0,0057 |   | 0,0754 |  + |  |
| 3 | 4,4065 | 0,0735 | 0,0754 | 0,0000 | 0,0054 |   | 0,0735 |  + |  |
| 4 | 4,4201 | -0,0201 | 0,0735 | 0,0088 | 0,0004 |   | -0,0201 |  - |  |
| 5 | 4,4163 | -0,0863 | -0,0201 | 0,0044 | 0,0075 |   | -0,0863 |  - |  |
| 6 | 4,4018 | -0,0918 | -0,0863 | 0,0000 | 0,0084 |   | -0,0918 |  - |  |
| 7 | 4,4042 | -0,0442 | -0,0918 | 0,0023 | 0,0020 |   | -0,0442 |  - |  |
| 8 | 4,3660 | -0,0260 | -0,0442 | 0,0003 | 0,0007 |   | -0,0260 |  - |  |
| 9 | 4,3461 | 0,0339 | -0,0260 | 0,0036 | 0,0012 |   | 0,0339 |  + |  |
| 10 | 4,3226 | -0,0026 | 0,0339 | 0,0013 | 0,0000 |   | -0,0026 |  - |  |
| 11 | 4,3458 | -0,0058 | -0,0026 | 0,0000 | 0,0000 |   | -0,0058 |  - |  |
| 12 | 4,3568 | -0,0268 | -0,0058 | 0,0004 | 0,0007 |   | -0,0268 |  - |  |
| 13 | 4,2788 | -0,0088 | -0,0268 | 0,0003 | 0,0001 |   | -0,0088 |  - |  |
| 14 | 4,2984 | -0,0084 | -0,0088 | 0,0000 | 0,0001 |   | -0,0084 |  - |  |
| 15 | 4,3022 | 0,0678 | -0,0084 | 0,0058 | 0,0046 |   | 0,0678 |  + |  |
| 16 | 4,3295 | 0,0305 | 0,0678 | 0,0014 | 0,0009 |   | 0,0305 |  + |  |
| 17 | 4,3344 | -0,0544 | 0,0305 | 0,0072 | 0,0030 |   | -0,0544 |  - |  |
| 18 | 4,3384 | 0,0216 | -0,0544 | 0,0058 | 0,0005 |   | 0,0216 |  + |  |
| 19 | 4,4533 | -0,0133 | 0,0216 | 0,0012 | 0,0002 |   | -0,0133 |  - |  |
| 20 | 4,5346 | -0,0046 | -0,0133 | 0,0001 | 0,0000 |   | -0,0046 |  - |  |
| 21 | 4,5932 | -0,0632 | -0,0046 | 0,0034 | 0,0040 |   | -0,0632 |  - |  |
| 22 | 4,5974 | 0,0226 | -0,0632 | 0,0074 | 0,0005 |   | 0,0226 |  + |  |
| 23 | 4,6077 | 0,0523 | 0,0226 | 0,0009 | 0,0027 |   | 0,0523 |  + |  |
| 24 | 4,8126 | -0,0626 | 0,0523 | 0,0132 | 0,0039 |   | -0,0626 |  - |  |
| 25 | 4,8742 | -0,1142 | -0,0626 | 0,0027 | 0,0130 |   | -0,1142 |  - |  |
| 26 | 4,9314 | -0,0814 | -0,1142 | 0,0011 | 0,0066 |   | -0,0814 |  - |  |
| 27 | 4,8981 | 0,0719 | -0,0814 | 0,0235 | 0,0052 |   | 0,0719 |  + |  |
| 28 | 4,8792 | 0,1208 | 0,0719 | 0,0024 | 0,0146 |   | 0,1208 |  + |  |
| 29 | 4,8342 | 0,1058 | 0,1208 | 0,0002 | 0,0112 |   | 0,1058 |  + |  |
| 30 | 4,9549 | -0,0249 | 0,1058 | 0,0171 | 0,0006 |   | -0,0249 |  - |  |
| 31 | 4,9483 | 0,0117 | -0,0249 | 0,0013 | 0,0001 |   | 0,0117 |  + |  |
| 32 | 4,9316 | -0,0516 | 0,0117 | 0,0040 | 0,0027 |   | -0,0516 |  - |  |
| 33 | 4,9358 | 0,0242 | -0,0516 | 0,0057 | 0,0006 |   | 0,0242 |  + |  |
| 34 | 4,9757 | 0,0043 | 0,0242 | 0,0004 | 0,0000 |   | 0,0043 |  + |  |
| 35 | 4,9687 | 0,0313 | 0,0043 | 0,0007 | 0,0010 |   | 0,0313 |  + |  |
| 36 | 4,9675 | -0,0275 | 0,0313 | 0,0035 | 0,0008 |   | -0,0275 |  - |  |
|  |  |  | Suma | 0,1305 | 0,1087 |  |  |  |  |
|  |  |  | **DW-d 0,1305/0,1087** |
|  |  | **dL** | **1,098** |  |  |  |  |  |  |
|  |  | **dU** | **1,442** | **Išvada apie autokoreliaciją : Rezultatas neapibržtas** |

Atliekamas DW testas. **Pirmas žingsnis**: Formuluojamos hipotezės:

H0: autokoreliacijos nėra. H1: autokoreliacija yra.

**Antras žingsnis:** Skaičiuojama DW-d statistika pagal formulę

 

 Šiam tikslui reikia turėti modelio paklaidas ei ir vėluojančias vienu periodu paklaidas ei-1. Paklaidos ei yra pateikiamos trečiajame lentelės stulpelyje. Vėluojančios paklaidos , tai tos pačios paklaidos tik paslinktos viena eilute žemiau, kaip parodyta ketvirtajame lentelės stulpelyje. Tuomet yra skaičiuojamas DW-d statistikos skaitikliui reikalingos paklaidų ir vėluojančių paklaidų skirtumo kvadratų reikšmės, t.y., (ei- ei-1)2 reikšmės, kurios pateiktos penktajame lentelės stulpelyje. Apskaičiavus reikšmes, jos susumuojamos ir gaunama formulės skaitiklyje esanti suma lygi 0,1305. Po to skaičiuojama vardiklyje esanti paklaidų kvadratų reikšmė, pradedant nuo antrosios paklaidos reikšmės. Šios reikšmės yra pateiktos lentelės šeštajame stulpelyje. Taip pat suskaičiuojama jų suma, kurios reikšmė yra 0,1087. Tuomet belieka pirmąją sumą padalinti iš antrosios ir gauname DW –d reikšmę 1,199817.

Trečias žingsnis:

Palyginti apskaičiuotą DW-d statistikos reikšmę su teorinėmis d reikšmėmis, kurios yra pateikiamos specialiose lentelėse. Tokia lentelė yra pateikta dėstytojos tinklalapyje (V.Karpuskiene rubrikoje KVST ) Kritines reikšmes rasime tokiu būdu: Pirmame DW statistinių lentelių stulpelyje yra nurodoma, kiek sudarytame modelyje yra įtraukta stebėjimų. Mūsų pavyzdyje yra n36 stebėjimai. Tuomet eilutėje surandame k=3. Šis skaičius k parodo, kiek yra į modelį įtraukta nepriklausomų kintamųjų. Mūsų pavyzdyje yra 3 kintamieji : rugiai, dyzelinas ir darbo užmokestis.

Eilutės su 36 stebėjimais ir stulpelių ksusikirtime randame dL ir du reikšmes . Jos yra lygios
dL ir dU

Apskaičiuota statistika d gali įgauti reikšmes nuo iš intervalo nuo 0 iki 4, kuris padalintas į:

* Autokoreliacijos nėra arba H0 neatmetama , kaidU ≤ d ≤ 4 - du,.t.y.,1,442≤d ≤2,558
* Autokoreliacija yra, kai d ≤ dL (teigiama autokoreliacija) arba d ≥ 4 - dL (neigiama autokoreliacija). Šiuo atveju hipotezė H0 yra atmetama.⇒
* Šiame teste gali būti dvi neapibrėžtumo sritys, dL ≤ d ≤ dU arba 4- dU ≤ d ≤ 4 - dL, kuriose griežto atsakymo apie paklaidų autokoreliacija nėra. Mūsų pavyzdyje apskaičiuota d reikšmė patenka į neapibrėžtumo intervalą: 1,098≤ d ≤ 1,442. Todėl išvada , kad DW testas tiksliai atsakyti apie autokoreliacija negali.

Kai gaunamas neapibrėžtas atsakymas, tuomet verta atlikti Ženklų sekų testą.

Ženklų sekų kriterijus.

Šiam kriterijui reikalinga Residual Output lentelė, kurioje būtų paklaidų reikšmės. Šalia paklaidų reikia surašyti ženklus: jei paklaida > 0, tai “+”, ir jei paklaida < 0, tai “–”.

Po to, reikia suskaičiuoti ženklų sekas, o viena seką sudaro vienodi ženklai. Tai pavaizduota lentelėje viršuje 8 ir 9 stulpelyje. Kaip matyti pirmos trys paklaidos yra teigiamos, todėl prie jų rašome “+” ženklą. Tai pirmoji ženklų seka. Ketvirtoji ir dar kelios paklaidos žemiau yra neigiamos, todėl prie jų pažymime “-“ ženklą. Tai yra antroji ženklų seka. Kai ženklas pasikeičia, tai prasideda kita seka. Mūsų pavyzdyje yra 16 sekų. Reikia suskaičiuoti kiek yra n1 - teigiamų su ”+” paklaidų : Mūsų pavyzdyje n1 16 ir kiek neigiamų su ”-” paklaidų n220

Po to suskaičiuojame

ir =8.52

H0:  Autokoreliacijos nėra.

HA:  Autokoreliacijos yra.

Jei apskaičiuota k (sekų skaičiaus reikšmė) patenka į intervalą:

 ,

 gauname kad apskaičiuotas sekų skaičius patenka į intervalą: 13,07<k=16<24.52

 H0 atmesti nėra pagrindo, tad su 95 % tikimybe galima teigti, jog autokoreliacijos nėra.

Taigi pagal šį kriterijų negalime atmesti H0, kuri teigia, kadpaklaidų autokoreliacijos nėra. Galima konstatuoti faktą, kad abu sudaryti modeliai analizei yra tinkami.