**Kvantitatiivsed meetodid II**

**SHSS.01.007, 3 EAP, Kadri Rootalu**

**Regressioonanalüüsi kasutusvaldkonnad suurandmete ajastul**

Senistes loengutes oleme regressioonanalüüsi õppinud kasutama peamiselt küsitlusandmete analüüsimisel. Meie eesmärgiks sotsioloogias on pigem olnud saada aru inimeste käitumismustreid ja arvamusi kujundavatest teguritest. Kuid regressioonanalüüsi kasutatakse laialdaselt ka muudes valdkondades ja kuigi tehnika ise ei ole erinev, võib erinev olla mudeli valiku strateegia ning soovitav järelduste tase. Käesoleva õppematerjali eesmärgiks ongi tutvustada regressioonanalüüsi kasutamise peamisi ideid teistes valdkondades (majandusteaduses, ettevõtluses, masinõppimise vallas) ning selgitada erinevusi ja sarnasusi sotsioloogiaga.

**Regressioonimudeli koostamise eesmärk**

Nagu öeldud, on sotsioloogias sageli regressioonimudelite koostamisel peamiseks eesmärgiks saada võimalikult hästi arusaadavad kirjeldus sõltuvale tunnusele ehk võimalikult hästi tõlgendatavad kordajad mudelisse.

Järgnevas näiteandmestikus on Eesti tuludeklaratsioonide andmed aastast 2015 (saadaval Eesti avaandmete portaalis https://opendata.riik.ee/dataset/fidek2015). Andmestikus on järgmised tunnused:

yhisdeklar (T/F). Kas tegu on ühisdeklaratsiooniga

tulud\_kokku. Kõik deklaratsiooni tulud kokku

mahaarvamised. Kõik mahaarvamised summeerituna

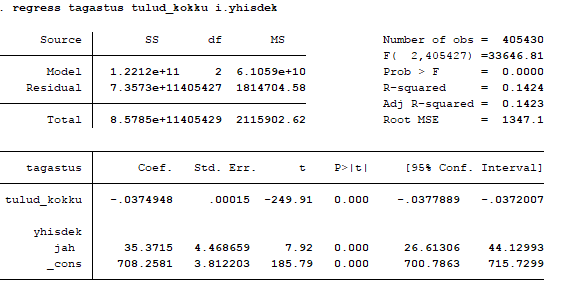
tagastus. Mis summas tulumaksu kas tagastati (positiivne number) või juurde tuli maksta (negatiivne)

Sotsioloogi huviks võiks siin olla kirjeldada, kui suur on Eesti elanike saadud tulumaksutagastus ja kuidas seda mõjutavad järgmised näitajad:

1. Kui suur on inimese kogutulu (näiteks: kas suuremaid maksutagastusi saavad kõrgepalgalised või madalapalgalised)
2. Kas inimene esitas ühisdeklaratsiooni või mitte

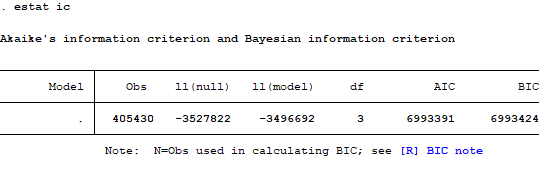
Selle ülesande lahendus lineaarse regressioonimudeli kaudu (koos Stata tellimiskäsuga) on toodud tabelis 1. Tulemustest selgub, et suuremat tulu saavad inimesed on saanud tulumaksutagastust väiksemas summas (regressioonikordaja -0,037) ning ühisdeklaratsiooni esitajad saavad keskmiselt 35 euro ja 37 sendi võrra suurema tagastuse.

Tabel 1. Lineaarne regressioonimudel tulumaksutagastuse summa kirjeldamiseks



Mudeli kirjeldusvõime osas võiks antud juhul vaadata determinatsioonikordajat (R-squared), mis on 0,14 või keskmist ruutviga (Root MSE) väärtusega 1347,1. Mõlemad vihjavad, et antud argumenttunnuste abiga oleme suutnud ära kirjeldada vaid väikese osa tulumaksutagastuse tunnuse hajuvusest ning palju on jäänud kirjeldamata. Mudeli AIC ja BIC on näidatud tabelis 2.

Tabel 2. Lineaarse regressioonimudeli AIC ja BIC

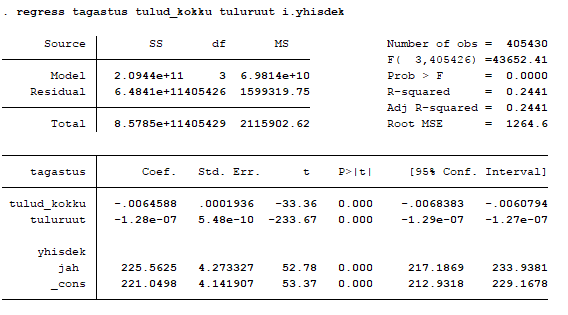


Majandusteaduses ja ka masinõppimise valdkonnas on analüütikute ees aga tihti olukorrad, kus peamiseks eesmärgiks on pigem leida mudel, millel on võimalikult hea sobitusaste andmetega ja mille abiga saaks teha võimalikult hea prognoosi uuritava nähtuse kohta. Näiteks: pole oluline kirjeldada väga täpselt konkreetse tarbija tarbimismustreid, pigem on oluline osata prognoosida konkreetse toote läbimüüki.

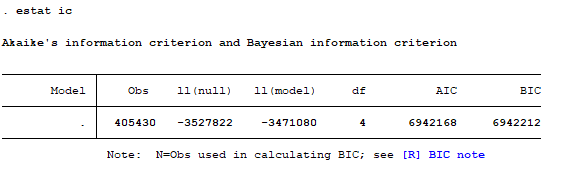
Käesoleva ülesande puhul võiksime siis mõelda, kas mudeli parandamiseks võiksime näiteks lubada mudelis mittelineaarseid liikmeid või ka koosmõjusid erinevate tunnuste vahel.

Järgnevas mudelis (tabel 3) ongi võetud arvesse ka kogutulu ruudus, mille kordaja tuleb negatiivse märgiga. Nii determinatsioonikordaja (0,24), keskmise ruutviga (1264,6) kui AIC ja BIC (tabel 4) annavad märku, et saadud mudel kirjeldab tulumaksutagastuse suurust paremini kui eelmine mudel (tabel 1 ja 2).

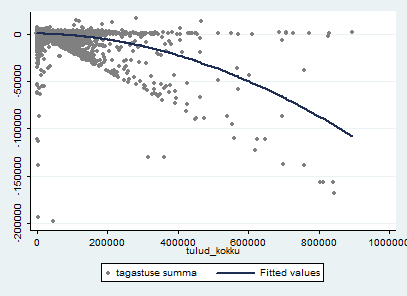
Tabel 3. Lineaarne regressioonimudel, kuhu on lisatud argumendina teenitud tulu ruudus



Tabel 4. Lineaarne regressioonimudel tulu ruuduga, AIC ja BIC



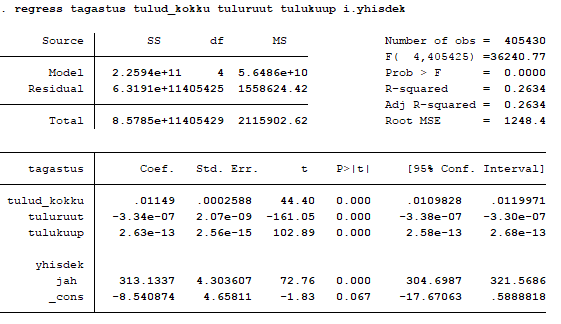
Kuid kuidas jääb tõlgendusega? Kas samas mudelis esinevad tulu suuruse liige ja ruutliige on intuitiivselt hästi mõistetavad? Selleks võiksime vaadata, milliseid prognoose meie mudel erinevate tulude puhul andis. Neid saaks salvestada käsuga predict ja lisades salvestatava tunnuse nime. Vaatame saadud prognoose joonisel 1. Sellelt selgub, et negatiivse tagastustrendi taga on mõned väga suure sissetulekuga vastajad, kes pidid deklaratsiooni esitamise järel palju juurde maksma.



Joonis 1. Lineaarse regressioonimudeli prognoosid ja tegelikud väärtused (mudel tabelist 3).

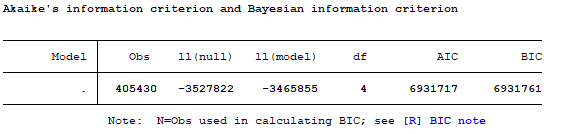
Järgnevalt proovime mudelisse olemasolevatele liikmetele lisada veel kuupliikme. Tulemused on esitatud tabelis 5. Selgub, et ka tulu kuubis omab positiivset mõju tagastusele.

Tabel 5. Lineaarse regressiooni mudel tulu, selle ruut- ja kuupliikmega.

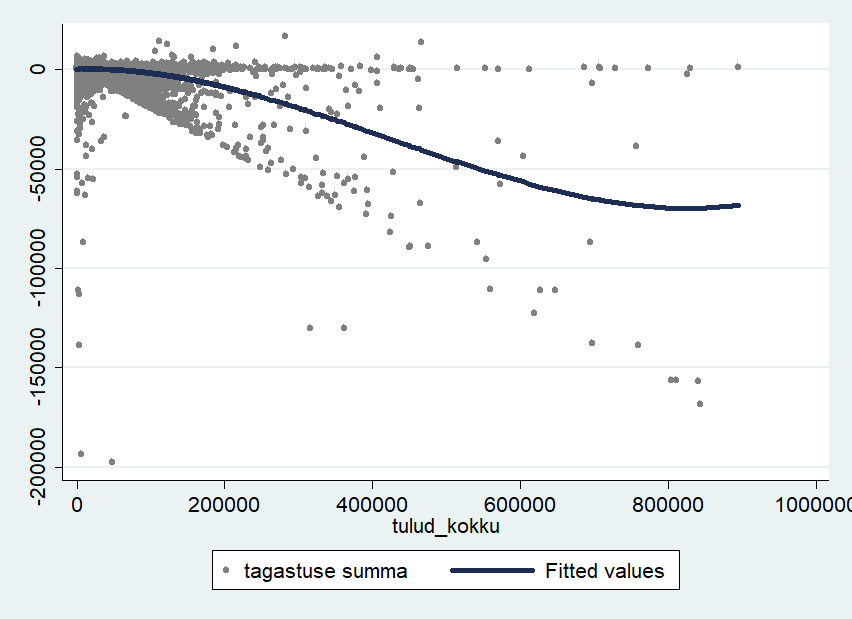


Saadud mudel on pisut parem eelmisest ruutliikmega mudelist (tabel 6), AIC ja BIC väärtused on tabelis 4 nähtust väiksemad.

Tabel 6. Lineaarne regressioonimudel ruut- ja kuupliikmega, AIC ja BIC.

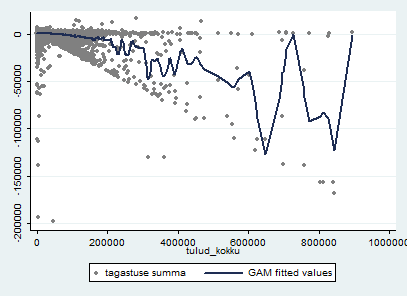


Kuidas näeks prognoositud tagastus välja graafiliselt, võib vaadata jooniselt 2.



Joonis 2. Lineaarse regressioonimudeli prognoosid ja tegelikud väärtused (mudel tabelist 5).

Eeltoodud tabelitest ja joonistelt selgub, et mudelit keerulisemaks muutes on ka mudeli kirjeldusvõime paranenud. Tihtipeale on nii, et mida komplekssemat mudelit me kasutame, seda parem mudeli kirjeldusvõime on. Lõppkokkuvõttes võiksime koostada ka funktsiooni, mis läbib praktiliselt kõiki punkte joonisel ja oleks praktiliselt 100%lise kirjeldusvõimega. Seda saaks teha näiteks GAM mudelite abiga (vt joonis 3). Samas aga ei ütleks funktsiooni kulgemise jälgimine meile enam eriti palju tunnustevaheliste seoste sisuliste seaduspärade kohta (kui need oleks olemas).



Joonis 3. GAM (General Additive Models) lahenduse prognoosid.

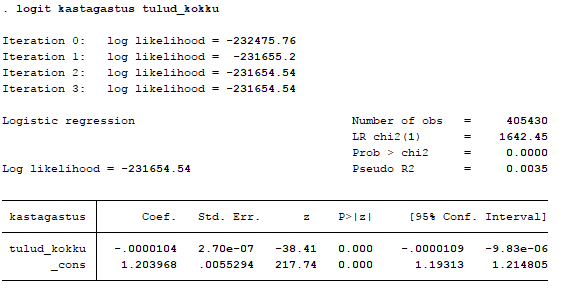
Eelnevates näidetes vaatasime lahendusi probleemidele, mida masinõppimise valdkonnas nimetatakse regressiooniprobleemideks. Mudeli headust hindasime determinatsioonikordaja, vea ning AIC ja BIC abiga.

Järgnevalt aga vaatame analoogilist näidet kaheväärtuselise sõltuva tunnuse jaoks. Selleks on andmestikus moodustatud uus sõltuv tunnus: kas on saanud tulumaksutagastuse (0 ei, 1 jah), mudeliga prognoositakse tagastuse saamise suhtelist tõenäosust. Tegemist on analüüsipüstitusega, mida masinõppimise valdkonnas nimetatakse klassifikatsiooniprobleemiks: ülesandeks on klassifitseerida objektid ning prognoosida, kas nad said tulumaksutagastuse või mitte. Antud probleemi lahendamiseks kasutame logistilise regressiooni mudelit.

Tabelis 7 ongi antud lahendus tagastuse prognoosimiseks kogutulu kaudu. Vastav Stata käsk on toodud tabeli ees.

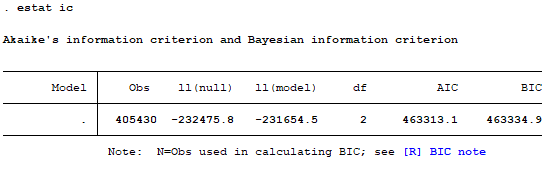
Selgub, et mida suurem on vastaja sissetulek, seda väiksem on tema suhteline tõenäosus saada tulumaksutagastust. Mõju ei ole küll suur ja ka mudeli kirjeldusvõime on väike.

Tabel 7. Logistilise regressiooni mudel tulumaksutagastuse prognoosimiseks



Ka logistilise regressiooni mudeli puhul võib leida AIC ja BIC, need on esitatud tabelis 8.

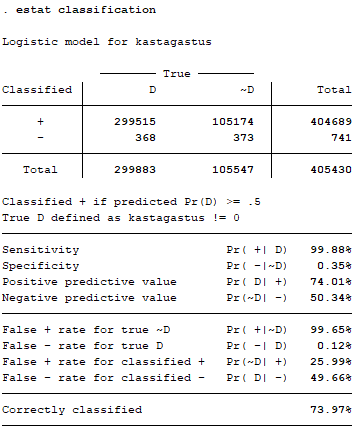
Tabel 8. Logistilise regressiooni mudel (tabelist 7), AIC ja BIC



Lisaks mudeli üldise kirjeldusvõime, AIC ja BIC esitamisele võib logistilise regressiooni mudeli headuse kriteeriumina välja tuua ka klassifikatsioonitabeli (tabel 9). Selles on esitatud objektide tegelikud kuuluvused ja prognoositud kuuluvused tulumaksutagastuse saamise osas. Vastav Stata käsk on antud tabeli ees.

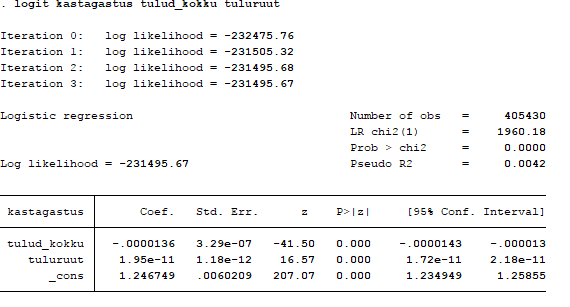
Selgub, et 299883st tegelikult tagastuse saanud inimesest klassifitseeris meie mudel õigesti 299515. Ning 105547st tegelikult tagastust mitte saanud inimesest klassifitseeris meie mudel õigesti 373. Seega, meie mudel kipub olema üleliia optimistlik tulumaksutagastuse saamise osas. Õigesti klassifitseeritud inimeste osakaal oli 73,97%.

Tabel 9. Klassifikatsioonitabel logistilise regressioonimudeli põhjal (tabel 7).



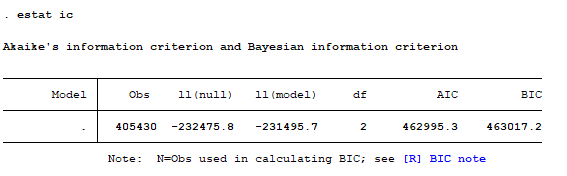
Järgnevalt proovime saadud mudelit parandada lisades argumenttunnuseks ka tulu ruutliikme. Tulemused on esitatud tabelis 10.

Tabel 10. Logistilise regressiooni mudel tulumaksutagastuse prognoosimiseks tulu ruuduga



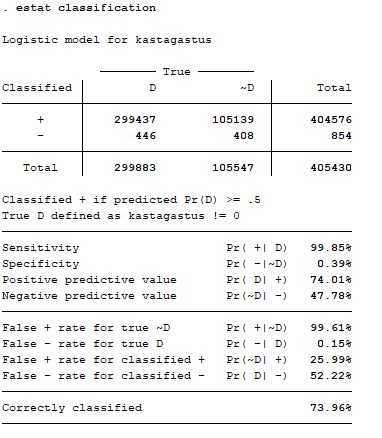
Selgub, et tulu ruutliikme kordaja mudelis on küll positiivne, kuid mudeli kirjeldusvõimet selle lisamine väga palju ei paranda (tabel 11).

Tabel 11. Logistilise regressiooni mudel (tabelist 10), AIC ja BIC



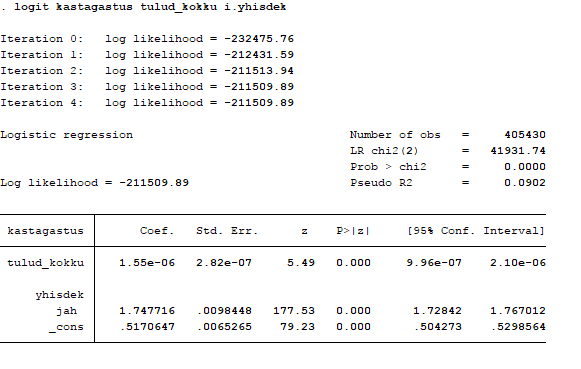
Samuti on näha, et mudeli abiga õigesti klassifitseeritud inimeste osakaal ei ole suurenenud (tabel 12).

Tabel 12. Klassifikatsioonitabel logistilise regressioonimudeli põhjal (tabel 10).



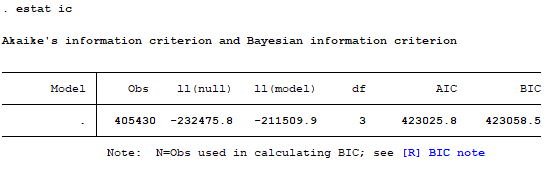
Seega võiks teha järelduse, et ruutliikme võtmine mudelisse meile kuigi palju võitu ei anna. Selle asemel proovime mudelisse võtta ka info selle kohta, kas oli tegemist ühisdeklaratsiooniga (tabel 13)

Tabel 13. Logistilise regressiooni mudel tulumaksutagastuse prognoosimiseks tulu ja ühisdeklaratsiooni tunnustega.



Mudeli kirjeldusvõime on eelnevatest esitatud mudelitest parem (Pseudo R2=0,0902) ja ka AIC ja BIC väärtuste võrdlus eelmiste mudelitega kinnitab seda järeldust (tabel 14).

Tabel 14. Logistilise regressiooni mudel (tabelist 13), AIC ja BIC



Sama mudeli klassifikatsioonitabel on esitatud tabelis 15. Sellest selgub, et õigeid prognoose oli jällegi 73,97%, kuid seejuures ei prognoosinud meie mudel ühelegi inimesele olukorda, kus tulumaksutagastust ei oleks. Võib arvata, et sedalaadi mudelist maksuametis märkimisväärset kasu poleks.

Tabel 15. Klassifikatsioonitabel logistilise regressioonimudeli põhjal (tabel 13).

