KVANTITATIIVNE ANDMEANALÜÜS II

KODUTÖÖ 2: MITMENE LINEAARNE REGRESSIOON

Esitamine: Kodutöö tuleb esitada Moodle's selleks ette nähtud kaustas – Kodutöö 2

Tähtaeg: 03.12.2023

Küsimused: Küsimuste ja probleemide korral kirjutage õppejõule (pauline.kommer@tlu.ee)

Kodutöö eesmärgiks on analüüsida ja tõlgendada tunnuste vahelisi seoseid ühes konkreetses riigis kasutades **mitmest lineaarset regressiooni**. Tulemuste esitamiseks valige R-i väljundist sobivad näitajad tabelis või joonisel ning esitage töös kindlasti ka enda põhitulemused tabelis või joonisel. Esitage ainult seose tõlgendamise seisukohast oluline informatsioon. Kõik statistilised tulemused, mis on töös esitatud, tuleb ka tekstina lahti kirjutada. Samuti tuleb selgitada, mida tulemused uurimisküsimuse seisukohalt tähendavad ja mida leitud seosed sisuliselt tähendavad (stiilis: võrreldes põhiharidusega on kõrgharidusega inimeste sissetulek oluliselt kõrgem vms).

Seoseanalüüsi puudutavat teksti kirjutades **vastake palun järgmistele küsimustele**: kas tunnused on omavahel seotud; kui jah, siis kas on võimalik hinnata seose tugevust, suunda, kas seos on statistiliselt oluline.

Kodutöö põhineb **Euroopa Sotsiaaluuringu** (*European Social Survey*) andmestikul. See on vabakasutusega rahvusvaheline võrdlev uuring, mille kohta saab rohkem lugeda uuringu kodulehelt http://www.europeansocialsurvey.org/. **Iga tudeng analüüsib erineva riigi andmeid (vt tabel viimasel lehel).** Allpool toodud küsimused (ning vastav andmefail **ESS2020_kliima.sav**[1]) pärinevad antud uuringu kümnenda ringi küsitlusest (läbi viidud 2020. aastal), mis muuhulgas käsitles kliima muutust (climate change). Oma töös tuleb tabelite/jooniste all viidata andmestikule järgmiselt: Euroopa Sotsiaaluuring 2020, autori arvutused. Kindlasti vaadake üle ka teised **kodutöö vormistamise nõuded** kolmandal lehel.

ÜLESANNE

Analüüsige Teile jagatud riigi (vt tabel viimasel lehel) kontekstis järgmiseid seoseid:

1. **Millisel määral tunnevad vastajad isiklikku kohustust püüda kliimamuutust vähendada?** ESS2020 ankeedis on selle kohta esitatud küsimus järgmiselt (tunnuse nimetus andmestikus kliima_kohustus):

C31 KAART 30 Mil määral tunnete te isiklikku kohustust püüda kliimamuutust vähendada?

Üldse mitte										Suurel määral	(Keeldus)	(EOÖ)
00	01	02	03	04	05	06	07	08	09	10	77	88

a. Esitage tunnuse ühemõõtmeline kirjeldav analüüs, kasutades vastavalt tunnuse skaala tüübile sobivaid analüüsimeetodeid (tabel ja/või joonis). **Põhjendage** meetodi valikut lähtuvalt tunnuse tüübist.

[1] Tasub tähele panna, et failis võivad kohati oma skaala tüübilt numbrilised tunnused olla R-is faktor-tüüpi tunnused, mistõttu tuleb tunnuse tüüp ära muuta. Kohati aga on mõttekas kategoriaalset tunnust (nt järjestusskaala tunnust enam-vähem võrdsete vahemikega) käsitleda numbrilisena, kuna vastav analüüsimeetod võib sõltuva tunnusena seda eeldada (lineaarne regressioon).

b. Võrrelge lisaks oma analüüsitavale riigile (vt tabel kodutöö lõpus) **veel kahe riigi** vastajate hinnangut selle osas, mil määral nad tunnevad isiklikku kohustust püüda kliimamuutust vähendada. Riigid valige oma huvist lähtuvalt, suurepärane, kui lisate valiku põhjenduse. **Kuidas Teie analüüsitav riik erineb neist, millega võrdlete?**

2. **Kas ja kuidas seletavad erinevad inimese näitajad/tegurid tema isikliku kohustuse tunnetamist kliimamuutuse vähendamise osas?**
   Koostage mitmese lineaarse regressiooni mudelid vastavalt järgnevatele tingimustele/alaküsimustele ning tõlgendage tulemusi. Uute tunnuste lisamisel teostage eelnevalt lühike ühemõõtmeline analüüs (vt all abiks analüüsi tegemisel).

   a. **Mudel 1**: Kas ja kuidas on kliimamuutuse vähendamise kohustuse tunnetamine seotud vastaja **vanuse** ja **sooga**? See tähendab, kas vanuse muutudes muutub kuidagi ka tunnetus kliimamuutuse kohustuse osas ja kas siin esineb soolisi erinevusi?
   b. **Mudel 2**: Kas ja kuidas on kliimamuutuse vähendamise kohustuse tunnetamine lisaks vanusele ja soole seotud vastaja hariduse (eisced_haridus) ja/või sotsiaalse staatusega (staatus)? Uurige mudelit, kas täiendavat seletusvõimet pakuvad mõlemad tunnused või tundub mõttekam piirduda ühega neist? Mõelge, kas/kuidas tunnuste väärtusi kokku kodeerida (abiks pakett forcats), et tulemused oleks ülevaatlikumad.
   c. **Mudel 3**: Kas ja kuidas on kliimamuutuse kohustuse tunnetamine lisaks eelnevatele teguritele seotud leibkonna sissetuleku tasemega (sissetulekutase)?
   d. **Mudel 4**: Kas eelmised mõjud muutuvad, kui mudelisse lisada ka tunnus, mis mõõdab elukohta (elukoht)? Jällegi tasub mõelda, kas mõned vastusevariandid/-väärtused oleks mõttekas kokku kodeerida.
   e. **Mudel 5**: Kuidas on kliimamuutuse kohustuse tunnetamine seotud sellega, kui tähtsaks inimene peab looduse eest hoolt kandmist (loodus_oluline) ja/või kuivõrd mures ta on kliimamuutuse pärast (kliima_mure)? Uurige mudelit, milline täiendav sõltumatu tunnus seostub sõltuva tunnusega selgemini või kas mõlemad on samaväärse tähtsusega? Mudeli headuse hindamisel on abiks mudeli kirjeldusvõime paranemise testimisel, aga ka multikollineaarsuse kontrollimisel (viimane võib (kuigi ei pruugi) näidata, et juba mõni varasemalt mudelisse lisatud tunnus on problemaatiline).
   f. **Mudel 6**: Omal valikul lisage mudelisse ka üks **koosmõju** ja kontrollige selle olulisust (nt kui mudelis on peamõjuna statistiliselt oluline seos sool ja mõnel muul tunnusel, siis saab need tunnused lisada koosmõjuna). Esitage saadud koosmõju kohta ka joonis (selles etapis pole oluline, kas mõju ise on statistiliselt oluline või mitte, küll aga lisage vastav omapoolne tõlgendus seose kohta). Koosmõjude joonisel esitamine on mugav paketiga interactions (eelnevalt tuleks alla laadida ka pakett jtools); nende jooniste korral on eraldi funktsioonid juhtudeks, kui argument pred = on numbriline tunnus (interact_plot()) või kui pred = on kategoriaalne tunnus (cat_plot()) (vt praktikum 3).

Abiks analüüsi tegemisel:

Analüüsi alustage iseenda jaoks iga analüüsitava tunnuse kirjeldava analüüsiga, et veenduda selle sisus ja kasutatavuses (nt kas on palju puuduvaid väärtusi, mis on kategooriate/vastusevariantide väärtused ja sisu, kas kõik kategooriad sobivad analüüsi või tuleks mõni neist määratleda puuduvaks väärtuseks, millised on vastusevariantide koodide suunad, kas nö intuitiivsed/loogilised või mitte). Kas nt lineaarse regressiooni tegemisel on mõned järjestusskaala tunnused mõttekas muuta numbrilisteks (as.numeric) või vastupidi numbrilised tunnuse muuta faktoriks (as.factor(), eriti oluline kategoriaalsete tunnuste korral); kuidas toimida koodidega, mida regressioonis analüüsida ei soovi, kas need tuleb enne analüüsist välja jätta (NA)?

Enne kui hakkate analüüsi tegema, koostage R-is uus andmestik (dataframe), kuhu olete välja **selekteerinud Teile analüüsiks antud riigi andmed (vt nimekirja allpool).** Üks viis sellise osaandmestiku

tegemiseks on näiteks dplyr paketis käsk filter(), aga kasutada võib ka subset() funktsiooni. Küll aga eeldab ülesanne 1 mitme riigi võrdlemist, nii et üks võimalus on selle jaoks teha nt teine osaandmestik, kuhu selekteerite välja kokku kolme riigi andmed.

Tunnuste väärtused on järgmised (PS. siin on ära toodud eestikeelsed väärtuste nimetused, failis on need jäetud ingliskeelseteks, tabelites/joonistel tuleks esitada eestikeelsed nimetused :)):

- kliima_kohustus(ankeedis küsimus C31): 0 üldse mitte … 10 suurel määral; 77 keeldus; 88 ei oska öelda; 99 vastus puudub
- sugu: 1 mees; 2 naine; 9 vastus puudub
- agea (vanus): vt väärtuseid andmestikus, kuna erinevad riigiti
- eisced_haridus: 1 ISCED I; 2 ISCED II, 3 ISCED IIIb, 4 ISCED IIIa, 5 ISCED IV, 6 ISCED V1, 7 ISCED V2, 55 muu, 77 keeldus, 88 ei oska öelda, 99 vastus puudub
- staatus: 1 töötab; 2 õpib; 3 töötu, ei otsi tööd; 4 töötu, otsib tööd; 5 töövõimetu; 6 pensionil; 7 ajateenistus; 8 kodune; 9 muu; 77 keeldus; 88 ei oska öelda; 99 vastus puudub
- sissetulekutase: 1 elame mugavalt; 2 saame hakkama; 3 on raske hakkama saada; 4 on väga raske hakkama saada; 77 keeldus; 88 ei oska öelda; 99 vastus puudub
- elukoht: 1 suur linn; 2 suure linna eeslinn või ääreala; 3 linn või väike linn; 4 küla; 5 talu või kodu maakohas; 7 keeldus; 8 ei oska öelda
- loodus_oluline (ankeedis H1 S): vastajate paluti hinnata, kuivõrd antud kirjeldus on nende moodi: looduse eest hoolt kanda on tema jaoks väga oluline: 1 väga minu moodi; 2 minu moodi; 3 mõnevõrra minu moodi; 4 vaid pisut minu moodi; 5 pole minu moodi; 6 pole üldse minu moodi; 7 keeldus; 8 ei oska öelda
- kliima_mure (ankeedis C32): 1 üldse mitte mures; 2 mitte väga mures; 3 mõnevõrra mures; 4 väga mures; 5 äärmiselt mures; 7 keeldus; 8 ei oska öelda

Nõuded kodutööde vormistusele:

- Igal kodutööl peab olema selle autori nimi.
- Töös peab olema selge, millise riigi andmeid analüüsite ja millist ülesannet lahendate.
- Kodutöö pikkus on umbes 3-4 lehekülge (mh olenevalt tabelite/jooniste suurusest).
- Töö peab olema nii keeleliselt, sisuliselt kui vormiliselt korrektselt teostatud ja esitatud.
- Pange rõhku ka tabelite ja jooniste vormistusele:
    - Tabelid/joonised peavad olema alati peal-/allkirjastatud (s.t tabelitel käib pealkiri tabeli kohale, joonistel joonise alla).
    - Tabeli/joonise all peab olema viide andmeallikale. (Allikas: Euroopa Sotsiaaluuring 2020, autori arvutused.)
    - Kõik tabelid/joonised peavad olema nummerdatud ja tekstis peab leiduma igale ühele neist viide.
    - Tunnuste/kategooriate nimed peavad olema sisukad ja lugejale mõistetavad, mitte lihtsalt andmefailist üle võetud.
    - Täpsustage, millega on tabeli/joonise näol tegu (protsendid, proportsioonid, keskmised vm).
    - Iga tabel/joonis peab moodustama terviku, st olema mõistetav ka tekstist eraldi.
- Esitage oma tulemused konkreetselt, selgelt ja lühidalt.
- Kodutööle (lisaks töö põhiosale) peab olema lisatud R-i skript (skripti osa kopeerituna sama Wordi dokumendi lõppu, kus esitate kodutöö sisulise osa).
    - NB! Skript peab olema „puhas“, s.t selles on esitatud vaid need käsud, mida töös kasutatud tulemuste saamiseks reaalselt vaja läks, üleliigsed (st rakendust mitte-leidnud) käsud tuleb kustutada. Lisage skripti ka kommentaare (nt ülesande nr ja sisu). Palun eristage #-märki kasutades skriptis ülesanded, et oleks aru saada milline osa on millise ülesande lahendamiseks.
- Tabelid/joonised ja tekst peavad olema eesti keeles (tabelite teksti saate muuta kopeerituna nt Wordis).

Kodutöö 2 riikide jaotus

	Nimi		**Riik**
1	Äli	Bergmann	Bulgaaria (BG)
2	Marta	Bogatõr	Tšehhi Vabariik (CZ)
3	Ekaterina	Filippova	Eesti (EE)
4	Kaleria	Frolovskaja	Soome (FI)
5	Artjom	Gnezdilov	Prantsusmaa (FR)
6	Emilia	Heero	Horvaatia (HR)
7	Lisbeth	Ilves	Ungari (HU)
8	Eleonor	Kalamets	Leedu (LT)
9	Karolina Helene	Kaukver	Sloveenia (SI)
10	Kelly	Kohal	Slovakkia (SK)
11	Kaarel	Leedo	Bulgaaria (BG)
12	Claudia Isabel	Lopez Ortiz	Tšehhi Vabariik (CZ)
13	Hosanna Sofia	Mäekallas	Eesti (EE)
14	Kertu	Saks	Soome (FI)
15	Hele	Simson	Prantsusmaa (FR)
16	Daniil	Stõk	Horvaatia (HR)
17	Deborah	Šapovalov	Ungari (HU)
18	Helena	Tihkan	Leedu (LT)
19	Anna Marie	Vasar	Sloveenia (SI)