FinSote-analyysit: R-ohje

7.9.2021

Terveyden ja hyvinvoinnin laitos

Sisältö

- Otanta-asetelman määrittäminen R:n *survey*-kirjastolla
  - Aineiston lataus ja muokkaus
  - Otanta-asetelman muodostus
- Perustunnuslukujen laskeminen
  - Perustunnusluvut
  - Tunnusluvut osajoukottain
- Regressioanalyysit
  - Lineaarinen regressiomalli
  - Logistinen regressiomalli

FinSote-aineiston lataus ja muokkaus, 1/2

- Aineiston lataus SAS-muodossa
  - F201 <- haven::read_sas(”*tiedostopolku*”)

- Aineiston muokkaus tapahtuu kätevimmin *dplyr*-kirjaston komennoilla:
  - *%>%: pipeline-operaattori, jatkaa edeltävästä funktiosta saadun objektin seuraavan funktion ensimmäiseksi argumentiksi*
  - *mutate(): lisää/muokkaa muuttujia*

Aineiston lataus ja muokkaus, 2/2

- Esim: lisätään ikäryhmäjaottelu 20v-34v, 35v-44v, 45v-54v, 55-64v ja +65v:

```
- library(dplyr)
- dat <- F201 %>% mutate(age = case_when(IKA2 %in%c(20:34) ~ 20,
                  IKA2 %in% c(35:44) ~ 35,
                  IKA2 %in% c(45:54) ~ 45,
                  IKA2 %in% c(55:64) ~ 55,
                  IKA2 > 64 ~ 65))
```

Otanta-asetelman muodostus

- Otanta-asetelman muodostukseen tarvittavat muuttujat FinSote 2020 -aineistossa:
  - Painokerroinmuuttuja: **w_analysis_suomi**, ositusmuuttuja: **rg_stratum_suomi**, perusjoukon koko osittessa: **rg_N_suomi**
  - Huom! Muuttujien nimet eivät välttämättä ole samat muissa aineistoissa

  - **library(survey)**
  - **dat_dsgn <- svydesign(id=~1, fpc=~rg_N_suomi, weights=~w_analysis_suomi, strata=~rg_stratum_suomi, data= dat)**

  - Jos tarkastelua tehdään hyvinvointialueiden tasolla, korvataan muuttujista loppuliitteet ”_suomi” loppuliitteillä ”_hyvinvointialue”

Perustunnuslukujen laskeminen

- survey-kirjastossa määritellyt, otanta-asetelman huomioivat funktiot alkavat usein **svy**-etuliitteellä
- Esim: terveysliikuntasuosituksen saavuttavien osuus
  - kyseessä 0/1-indikaattorimuuttuja, joten keskiarvo vastaa suosituksen saavuttavien osuutta
  - **svymean(~fs_phexcer_guidel_enough, design=dat_dsgn, na.rm=T)**

```
> svymean(~fs_phexcer_guidel_enough, design=dat_dsgn, na.rm=T)
                              mean     SE
fs_phexcer_guidel_enough 0.39077 0.0048
```

Perustunnusluvut osajoukottain, 1/2

- Esim: terveysliikuntasuosituksen saavuttavien osuus sukupuolen ja ikäryhmän mukaan
  - **svyby(~fs_phexcer_guidel_enough, ~sukupuoli+age, design=dat_dsgn, FUN=svymean, na.rm=T)**

```
> svyby(~fs_phexcer_guidel_enough, ~sukupuoli+age, design=dat_dsgn, FUN=svymean, na.rm=T)
     sukupuoli age fs_phexcer_guidel_enough          se
1.20         1  20                0.4682257 0.018969301
2.20         2  20                0.4629281 0.016356280
1.35         1  35                0.4153866 0.020045247
2.35         2  35                0.3755192 0.018213039
1.45         1  45                0.4038928 0.018044699
2.45         2  45                0.3652696 0.015592280
1.55         1  55                0.3510848 0.013211341
2.55         2  55                0.3720847 0.012102154
1.65         1  65                0.3487256 0.008795046
2.65         2  65                0.3341483 0.008014107
```

Perustunnusluvut osajoukottain, 2/2

- Osajoukkokohtaisia tuloksia voi laskea myös yksitellen käyttäen esim. subset-funktiota svydesign-objektille:
  - **svymean(~fs_phexcer_guidel_enough, design=subset(dat_dsgn, age==20&sukupuoli==1), na.rm=T)**

```
> svymean(~fs_phexcer_guidel_enough, design=subset(dat_dsgn, age==20&sukupuoli==1), na.rm=T)
                            mean    SE
fs_phexcer_guidel_enough 0.46823 0.019
```

Luokitellut muuttujat: jakaumat

- Luokkaesiintyvyydet saa esim. **svymean**-komennolla:
  - **svymean(~interaction(fs_phexcer_guidel_enough, age), design=dat_dsgn, na.rm=T)**

```
> svymean(~interaction(fs_phexcer_guidel_enough, age), design=dat_dsgn, na.rm=T)
                                                      mean     SE
interaction(fs_phexcer_guidel_enough, age)0.20 0.126249 0.0039
interaction(fs_phexcer_guidel_enough, age)1.20 0.110016 0.0038
interaction(fs_phexcer_guidel_enough, age)0.35 0.104012 0.0035
interaction(fs_phexcer_guidel_enough, age)1.35 0.068123 0.0029
interaction(fs_phexcer_guidel_enough, age)0.45 0.100564 0.0030
interaction(fs_phexcer_guidel_enough, age)1.45 0.062955 0.0024
interaction(fs_phexcer_guidel_enough, age)0.55 0.115119 0.0026
interaction(fs_phexcer_guidel_enough, age)1.55 0.065279 0.0020
interaction(fs_phexcer_guidel_enough, age)0.65 0.163283 0.0026
interaction(fs_phexcer_guidel_enough, age)1.65 0.084400 0.0018
```

Lineaarinen regressioanalyysi

- Luokitellut muuttujat määritellään käyttämällä factor()-funktiota
- Esim:
  - **mod1 <- svyglm(weight_kg ~ height_cm + factor(age), design=dat_dsgn)**
  - **summary(mod1)**

```
Survey design:
survey::svydesign(id = ~1, fpc = ~rg_N_suomi, weights = ~w_analysis_suomi,
    strata = ~rg_stratum_suomi, data = dat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -81.6662     3.9205 -20.831  < 2e-16 ***
height_cm      0.9216     0.0227  40.610  < 2e-16 ***
factor(age)35  3.5278     0.6003   5.877 4.23e-09 ***
factor(age)45  6.5630     0.5454  12.034  < 2e-16 ***
factor(age)55  5.6937     0.4712  12.083  < 2e-16 ***
factor(age)65  2.8334     0.4193   6.757 1.43e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lineaarinen regressio: Waldin testi

- Testataan, onko ikäryhmien välillä eroa:
  - **regTermTest(mod1, ~factor(age))**

```
> regTermTest(mod1, ~factor(age))
Wald test for factor(age)
 in svyglm(formula = weight_kg ~ height_cm + factor(age), design = dat_dsgn)
F =  56.94265  on  4  and  27312  df: p= < 2.22e-16
```

  - regTermTest-funktion nollahypoteesi on, että kaikkien annettujen selittäjien (tässä siis pelkkä ikäryhmä) vaikutukset ovat nollia, eli tässä tapauksessa, että ikäryhmien välillä ei olisi eroa

  - Pieni p-arvo viittaa, että ikäryhmien välillä on eroa

Lineaarinen regressio: yhdysvaikutukset

- Yhdysvaikutus kuvataan :- tai *-merkinnällä. Jälkimmäinen muodostaa myös päävaikutustermit
  - **mod2 <- svyglm(weight_kg ~ height_cm*factor(age), design=dat_dsgn)**
  - **summary(mod2)**

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -95.446402   9.424028 -10.128   <2e-16 ***
height_cm                 1.001226   0.054649  18.321   <2e-16 ***
factor(age)35            35.475989  16.700796   2.124   0.0337 *
factor(age)45             6.187914  12.699962   0.487   0.6261
factor(age)55            21.926504  13.170330   1.665   0.0960 .
factor(age)65            23.799146   9.772866   2.435   0.0149 *
height_cm:factor(age)35  -0.184916   0.096464  -1.917   0.0553 .
height_cm:factor(age)45   0.002462   0.073706   0.033   0.9733
height_cm:factor(age)55  -0.093891   0.076550  -1.227   0.2200
height_cm:factor(age)65  -0.122334   0.056795  -2.154   0.0313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yhdysvaikutusten testaaminen

- Testataan, onko yhdysvaikutus merkitsevä
  - **regTermTest(mod2, ~height_cm:factor(age))**

```
> regTermTest(mod2, ~height_cm:factor(age))
Wald test for height_cm:factor(age)
 in svyglm(formula = weight_kg ~ height_cm * factor(age), design = dat_dsgn)
F =  2.680463  on  4  and  27308  df: p= 0.029896
```

- Huom! :-merkki tarkoittaa, että testataan ainoastaan yhdysvaikutustermiä. Jos olisikin käytetty ~height_cm*factor(age), niin testattaisiin, ovatko yhdysvaikutus JA päävaikutukset nollia.

- Vastaavasti ~height_cm+factor(age) testaisi, ovatko molemmat päävaikutukset nollia.

Logistinen regressiomalli, 1/2

- Logistinen regressiomalli saadaan antamalla svyglm-funktioon argumentiksi ”family = quasibinomial()”:
  - **mod3 <- svyglm(fs_phexcer_guidel_enough ~ factor(sukupuoli)*factor(age), design=dat_dsgn, family = quasibinomial())**
  - **summary(mod3)**

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -0.12727    0.07618  -1.671   0.0948 .
factor(sukupuoli)2              -0.02129    0.10066  -0.212   0.8325
factor(age)35                   -0.21447    0.11233  -1.909   0.0562 .
factor(age)45                   -0.26200    0.10687  -2.452   0.0142 *
factor(age)55                   -0.48701    0.09574  -5.087 3.68e-07 ***
factor(age)65                   -0.49738    0.08546  -5.820 5.98e-09 ***
factor(sukupuoli)2:factor(age)35 -0.14558    0.15158  -0.960   0.3369
factor(sukupuoli)2:factor(age)45 -0.14200    0.14238  -0.997   0.3186
factor(sukupuoli)2:factor(age)55  0.11228    0.12719   0.883   0.3774
factor(sukupuoli)2:factor(age)65 -0.04355    0.11371  -0.383   0.7018
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistinen regressiomalli, 2/2

- Logistisen regressiomallin tulosteessa vaikutusten estimaatit ovat logit-skaalalla, joten negatiiviset estimaatin arvot tarkoittavat pienempää todennäköisyyttä havaita vasteen arvo 1 (olettaen, että vaste on koodattu 0/1) ja positiiviset estimaatit suurempaa todennäköisyyttä.