Introduktion til R III

12. maj 2023

Kristian G. Kjelmann (kgk@socsci.aau.dk)
&
Rolf L. Lund (rolfll@socsci.aau.dk)

Institut for Sociologi og Socialt Arbejde

AALBORG UNIVERSITY
DENMARK

Dagens program

1. Opsamling på datahåndtering
2. Eksport af statistiske modeller (med stargazer)
3. Opsætning af modeller i tekstpublikationer
4. Øvelser

Dagens læringsmål

I kan lave publicérbare modeller med R

I kan løse forskelligartede datahåndteringsudfordringer i det samme datasæt

Datahåndteringsbegreber

Har I styr på jeres datahåndteringsbegreber?

- Subsetting
- Filtrering
- Variabelændringer
- Rekodning
- Missingværdier

Statistiske modeller i R

Der er to dele i at få resultaterne af en statistisk model frem i R:

1. Specificér modellen (fx med en funktion som `lm()` for lineære modeller)
2. Få koefficienter og resultater fra modellen (med brug af funktionen `summary()`)

Statistiske modeller i R - Specificer model

Tre ting krævet for at specificere en model:

1. Et datasæt
2. Funktion for typen af model, man vil lave
3. Formel, der specificerer det sammenhæng, som man vil modellere

Al datahåndtering i datasæt skal ske *inden* man laver modellen.

Man specificerer en formel med R's formelsyntax, fx `y ~ x1 + x2 + x3`.

Eksempel:

```
grsp_model <- lm(grspnum ~ eduyrs + wkhtot, data = ess18)
```

Statistiske modeller i R - Resultater

En model i R er blot endnu en type objekt.

For at se resultatet af modellen, skal man derfor spørge R korrekt om det.

Ved blot at kalde modellen får man begrænsede resultater:

```
grsp_model
```

```
Call:
lm(formula = grspnum ~ eduyrs + wkhtot, data = ess18)

Coefficients:
(Intercept)       eduyrs       wkhtot
    -3668.7        891.1        888.4
```

Statistiske modeller i R - Resultater

Ved at bruge `summary()` funktionen gives de relevante resultater:

```
summary(grsp_model)
```

```
Call:
lm(formula = grspnum ~ eduyrs + wkhtot, data = ess18)

Residuals:
    Min      1Q  Median      3Q     Max
 -64614  -16639  -10324   -1448 2955649

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3668.7    20645.3  -0.178   0.8590
eduyrs         891.1      856.4   1.041   0.2984
wkhtot         888.4      458.2   1.939   0.0529 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 121600 on 707 degrees of freedom
  (575 observations deleted due to missingness)
Multiple R-squared:  0.007698,  Adjusted R-squared:  0.004891
F-statistic: 2.742 on 2 and 707 DF,  p-value: 0.06511
```

Statistiske modeller i R - factors

Som standard vil R behandle tekstvariable i statistiske modeller som *unordered factors*; altså nominalt skalerede variable.

R vil desuden tage kategorien, der kommer først i alfabetisk rækkefølge, som referencekategori.

Dette kan lede til uhensigtsmæssige resultater. Man bør derfor altid tage aktiv stilling til, hvordan den kategoriske variabel skal behandles, inden man laver modellen.

Statistiske modeller i R - factors

Overvej følgende:

Skal variablen behandles som *ordinal* eller *nominal*? (`ordered` eller `unordered`)

Skal variablen behandles som *intervalskaleret*? (konvertér til numerisk)

Hvis variablen skal behandles som nominal, hvilken kategori skal så være *referencekategorien*? (kan fx ændres med `relevel` i `mutate` funktionen)

HUSK: Der er forskel på hvad variablen *er*, og hvordan vi behandler den i en model!

Eksport af statistiske modeller

```
Call:
lm(formula = grspnum ~ eduyrs + wkhtot, data = ess18)

Residuals:
    Min      1Q  Median      3Q     Max
-64614 -16639 -10324  -1448 2955649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3668.7    20645.3  -0.178   0.8590
eduyrs        891.1      856.4   1.041   0.2984
wkhtot        888.4      458.2   1.939   0.0529 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121600 on 707 degrees of freedom
  (575 observations deleted due to missingness)
Multiple R-squared:  0.007698,    Adjusted R-squared:  0.004891
F-statistic: 2.742 on 2 and 707 DF,  p-value: 0.06511
```

Effekt af års uddannelse og arbejdstid på løn

Antal års uddannelse	891,15
	(856,41)
Arbejdstimer om ugen	888,39
	(458,22)
Constant	-3.668,75
	(20.645,31)
Observations	710
R^2	0,01
Adjusted R^2	0,005
Residual Std. Error	121.581,20 (df = 707)
F Statistic	2,74 (df = 2; 707)
Note:	*p<0,05; **p<0,01; ***p<0,001