R-seminar 5: Løsningsforslag

STV1020 Vår 2021

Uke 16

Du skal bruke datasettet wvs_us17.csv og lineær regresjon for å undersøke relasjonen mellom (avhengig) AV og uavhengig (UV) variabel. Datasettet er en survey gjennomfør i USA i 2017 for World Value Survey. Individer svarer på surveyen. Tilsvarende surveyer er gjort for mange andre land og år (se World Value survey sine nettsider for mer informasjon), men i dag bruker vi et datasett som bare inneholder observasjoenen fra USA i 2017.

Løsningsforslag til oppgaver

Husk at det finnes flere måter å løse disse på. Dersom du har fått det til, men ikke brukt akkurat samme kode som oss så kan du likevel ha gjort helt riktig. Det kommer an på hvilke datasett du velger å gå videre med i resten av oppgavesettet, f.eks. det med eller uten missing, det med eller uten 7 variabler.

1. Last inn datasettet og oppgi antall enheter og variabler i datasettet. Datasett: https://raw.githubusercontent.com/liserodland/STV1020/main/data/wvs_us17.csv
   Tips: Lim inn lenken, slik som i seminar 4.

```
# Laster inn pakker
library(tidyverse)
library(stargazer)

# Laster inn datasettet
data <- read.csv("https://raw.githubusercontent.com/liserodland/STV1020/main/data/wvs_us17.

# Finner antall enheter og variabler
dim(data)


## [1] 2596    7


# Kan også lese samme informasjon i environment
```

2. Finn navn på variablene i datasettet.

```r
# Henter ut variabelnavn
names(data)
```

```
## [1] "year"          "country"       "age"           "gender"
## [5] "corruption"    "imp_democracy" "income_group"
```

3. Opprett en nytt datasett med kun variablene imp_democracy, age og income_group. Pass på at klassen til variablene er numeric.

```r
# Oppretter nytt datasett data2
data2 <- data %>%
  select(imp_democracy, age, income_group)

# Sjekker klassen
str(data2)
```

```
## 'data.frame': 2596 obs. of  3 variables:
##  $ imp_democracy: int  2 5 1 10 5 1 10 6 5 1 ...
##  $ age          : int  43 35 48 49 20 27 41 52 24 57 ...
##  $ income_group : int  3 5 1 1 5 1 6 10 NA 1 ...
```

```r
# Alle er integer, altså heltall og således numeriske

# Evt:
class(data2$imp_democracy)
```

```
## [1] "integer"
```

```r
# ... for hver variabel
```

4. Vis hvordan du fjerner enheter som mangler opplysninger fra datasettet. Oppgi antall enheter i datasettet etter at du har fjernet enhetene.

```r
# Fjerner enheter som mangler opplysninger
data2_nona <- data2 %>%
  drop_na()

dim(data2_nona)
```

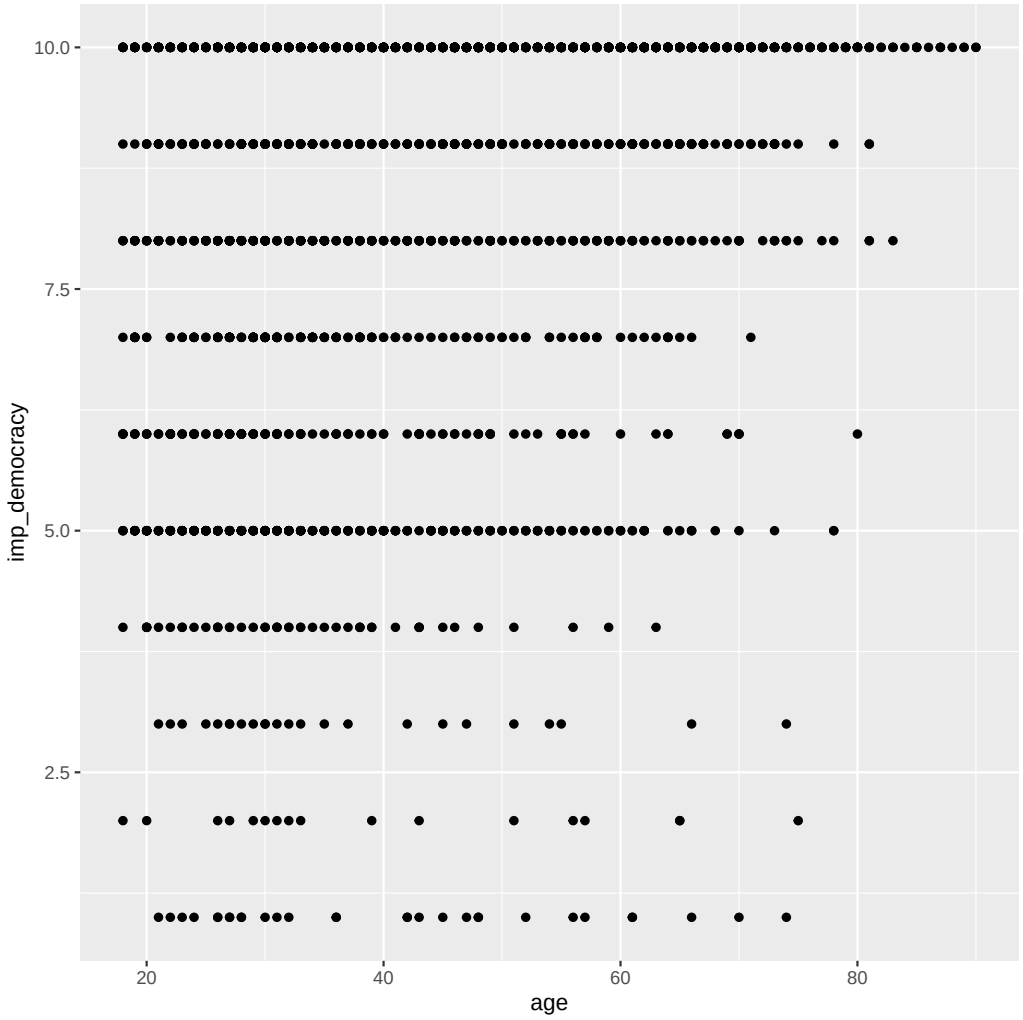
```
## [1] 2515    3
```

```r
# Antall observasjoner er nå 2515
```

5. Lag et spredningsplott av sammenhengen mellom imp_democracy (AV) og age (UV) Selv om imp_democracy egentlig har ordinalt målenivå kan du tenke på den som tilnærmet kontinerlig i de følgende oppgavene.

```r
ggplot(data, aes(x = age, y = imp_democracy)) +
  geom_point()

## Warning:  Removed 44 rows containing missing values (geom_point).
```



6. Kjør en lineær regresjonsmodell med imp_democracy som AV og age som UV.

```r
# Kjører lineær model
model1 <- lm(imp_democracy ~ age,
             data = data)
```

7. Tolk koeffisienten til age

```r
summary(model1)


##
## Call:
```

```
## lm(formula = imp_democracy ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4620 -1.0444  0.6499  1.5080  2.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.701380   0.116739   57.41   <2e-16 ***
## age         0.037305   0.002514   14.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.068 on 2550 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.07948,Adjusted R-squared:  0.07912
## F-statistic: 220.2 on 1 and 2550 DF,  p-value: < 2.2e-16


# Når age (UV) øker med en skalaenhet (her ett år),
# øker imp_democracy (AV) med 0.037 skalaenheter.
# Resultatet er statistisk signifikant på 1 prosentsnivå ***.
```

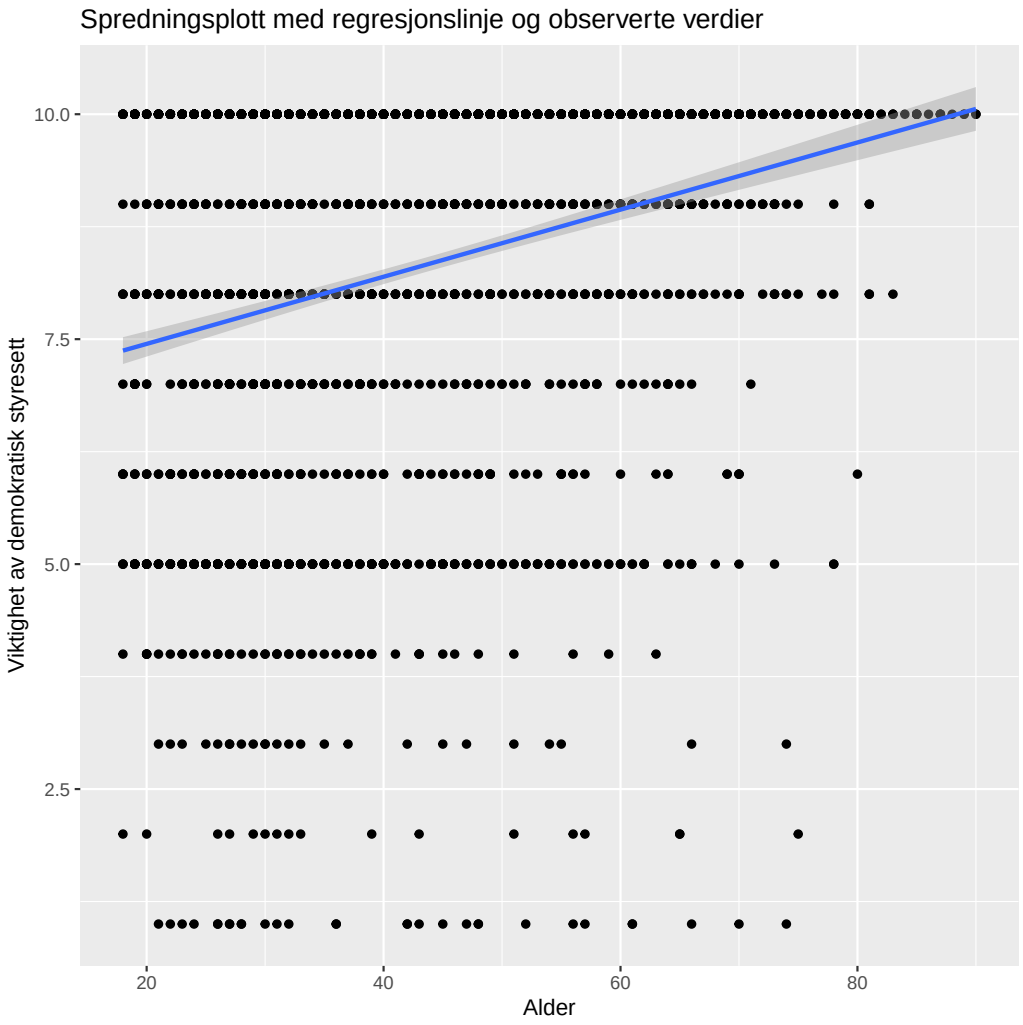
8. Print resultatet av regresjonen i en tabell ved hjelp av stargazer(). Lagre tabellen lokalt på pc-en din og åpne den i f.eks. word eller en nettleser.

```
# Lager tabell med stargazer og lagrer lokalt
stargazer(model1,
          type = "html",
          out = "regresjonsmodell.html")
```

9. Lag et plott med observerte verdier av imp_democracy på y-aksen og observerte verdier av age på x-aksen. Legg til en regresjonslinje, endre aksetitlene og legg til en tittel. Lagre plottet lokalt på pc-en din.

```
# Plot med regresjonsline
ggplot(data, aes(x = age, y = imp_democracy)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Alder",
       y = "Viktighet av demokratisk styresett",
       title = "Spredningsplott med regresjonslinje og observerte verdier")


## Warning:  Removed 44 rows containing non-finite values (stat_smooth).
## Warning:  Removed 44 rows containing missing values (geom_point).
```




```
# Lagrer
ggsave(filename = "plot.jpg")

## Warning:  Removed 44 rows containing non-finite values (stat_smooth).
## Warning:  Removed 44 rows containing missing values (geom_point).
```

10. Lag en ny variabel gender2 som tar verdien "Female" når observasjonen har verdien "2" på gender og "Male" når observasjonen har verdien "1" på gender. Sjekk at det ble riktig.

```
# Alternativ 1
data$gender2 <- ifelse(data$gender == 1, "Male", "Female")

# Alternativ 2
data <- data %>%
  mutate(gender3 = recode(gender, "1" = "Male", "2" = "Female"))

# Sjekker at det ble riktig
table(data$gender2, data$gender)
```

```
##
##             1    2
##   Female    0 1206
##   Male   1390    0
```

```
table(data$gender3, data$gender)
```

```
##
##             1    2
##   Female    0 1206
##   Male   1390    0
```

11. Estimer en ny multivariat regresjonsmodell med imp_democracy som AV og age og gender 2 som UVs (dette er tema for seminar 6)

```
# Estimerer en multivariat regresjonsmodell
model2 <- lm(imp_democracy ~ age + gender2,
             data = data)

summary(model2)
```

```
##
## Call:
## lm(formula = imp_democracy ~ age + gender2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4575 -1.1459  0.6098  1.5035  2.8150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.579658   0.117497  55.998  < 2e-16 ***
## age         0.033629   0.002564  13.117  < 2e-16 ***
## gender2Male 0.523782   0.083711   6.257 4.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.052 on 2549 degrees of freedom
##   (44 observations deleted due to missingness)
## Multiple R-squared:  0.09341,Adjusted R-squared:  0.0927
## F-statistic: 131.3 on 2 and 2549 DF,  p-value: < 2.2e-16
```

```
stargazer(model2,
          type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                             imp_democracy
## ------------------------------------------------
```

```
## age                             0.034***
##                                  (0.003)
##
## gender2Male                     0.524***
##                                  (0.084)
##
## Constant                        6.580***
##                                  (0.117)
##
## ------------------------------------------------
## Observations                      2,552
## R2                                0.093
## Adjusted R2                       0.093
## Residual Std. Error       2.052 (df = 2549)
## F Statistic           131.313*** (df = 2; 2549)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```