

Computer Exercises Day 4 – Health Register Epidemiology – September 2018

1- Consider `dta.mfr`. This is a data set that was made to look like data from the Medical Birth Registry. The variables are...

- `lopenr`: ID number. Unique for all subjects
- `fdato`: Date of birth for subject
- `ffdato`: Date of birth for the father of subject
- `kjonn`: Sex of subjects
- `vekt`: Birth weight of subjects

We want to see if there are sex differences in birth weight, and if paternal age at birth affects birth weight.

i. Clean up the data

- a) How many records are there?
- b) How many unique "lopenr" are there?
- c) Why is there a difference between your findings in i) and ii)? Remove redundant records.
- d) Make a histogram of `fdato`. Does everything look OK? Explain.
- e) How many birth dates are missing?
- f) How do we handle the missing data?
- g) How many paternal birth dates are missing?
- h) Make a histogram of paternal birth date. Does everything look OK? Explain.
- i) Create a variable, `agedad`, which is paternal age at birth, and make a histogram.

Does everything look OK?

- j) Drop records where paternal age at birth is unrealistic. Which cutoff(s) do you use?
- k) Re-draw the histogram from h). Comment on the differences.
- l) What are the minimum and maximum birth weights? Do they look realistic?

If not, drop records with unrealistic birth weights. Which cutoff(s) do you use?

m) Make a histogram of birth weight. Does everything look OK?

Now that we are happy with the data, we may move on to the analyses.

ii. Run analyses

- a) Are there sex differences in birth weight? If yes, how big?
- b) Does paternal age affect birth weight? If yes, how much?
- c) Repeat a) and b) on the original data set. Comment on the results.

2- Consider `edu.dta`. This is a data set that was made to look like data from the Educational Database. The variables are...

- `lopenr`: ID number. Same as in the first data set.

- faar: Birth year for subject.
- ffaar: Birth year for father of subject
- utdaar: Education year
- utd: Education in education year
 - 0- No elementary school (barneskolen)
 - 1- Elementary school (barneskolen)
 - 2- Lower secondary school (ungdomsskolen)
 - 3- Upper secondary school, first two years (VGS, grunnutdanning)
 - 4- Upper secondary school, third year (VGS, avsluttende utdanning)
 - 5- Upper secondary school, additional year (VGS, påbygging)
 - 6- Lower level university (e.g., bachelor)
 - 7- Upper level university (e.g., master)
 - 8- PHD
 - 9- Not given

We want to see if paternal age at birth affects education

- i. Prepare data
 - a) How many records are there?
 - b) How many unique "lopenr" are there?
 - c) Is the large number of rows a problem? Why (not)?
 - d) Convert from long to wide format, using education each year as a time-varying variable
 - e) How many records are there now?
 - f) How many have missing values on education in 1967?
 - g) How many have missing values on education in 2016?
 - h) How do we handle missing data in the education variable?
 - i) Tabulate education in 2016. Why are there so many zeros?

We are happy about the education data set, and want to create our final data set.

- ii. Merge with former data set
 - a) Merge the two data sets. How many records do not match?
 - b) Handle non-matching records. What did you do?

We are finally ready to run the final analyses.

- iii. Run analyses
 - a) Does paternal age affect education?
 - b) Explain how you performed the analyses, and what the results were.