Multilevel Models

Philippe Rast

PSC 204B
UC Davis, Winter 2018

Topics

Single-level Regression:

- Week 1 Linear Regression (G&H: 3,4)
- Week 2 Multiple Regression
- Week 3 Violation of Assumptions
- Week 4 Logistic Regression and GLM (G&H: 5, 6)
- Week 5 Over-fitting, Information Criteria and Model comparison (McE: 6)
- Week 6 Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

- Week 7 Multilevel Linear Models (G&H: 11–13)
- Week 8 Multilevel Models (G&H: 14, 15)
- Week 9 Bayesian Inference (G&H: 18 / McE: 1, 2, 3)
- Week 10 Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Overview

1. Sources of Variation
2. Inflated Correlations
   - Layout
   - Simulation

- Define Population Model
- Estimate correlation
- influence of Vart_i
- Narrow Age Cohort Design

Sources of Variation

- Where is variation coming from?
- Between-person?
- Within-person?
- Within-person between measurement occasions?
  - e.g. Intensive measurement designs

Side Topic:
Mean induced associations in Cross-Sectional Data

Hofer & Sliwinski 2001: “mean induced associations”

- Correlations may reflect differences in means:
- The passage of time leads to “fake“ correlations

Problem Mean differences at the population level are confounded with changes at the individual level

e.g.: Speed Hypothesis, common cause hypotheses (sensory functioning, etc.)

We always knew it...

There has been no source more fruitful of fallacious statistical argument than the common influence of the time factor.
Cave and Pearson (1914, p. 354)

Biased correlations

Indices about rates of change based on cross-sectional age in age-heterogeneous samples are *always* biased. Why?
Individual values of two variables x and y of person i at a given age t corresponds to the sum of fixed and random effects:

$$\begin{aligned} x_{it} &= L_x + L_{xi} + S_x t_i + S_{xi} t_i + e_{xi} \\ y_{it} &= L_y + L_{yi} + S_y t_i + S_{yi} t_i + e_{yi} \end{aligned} \tag{1}$$

- L is the average level
- L_i is the individual departure from the fixed effect
- S is the average slope
- S_i is the individual departure from the fixed effect
- t_i age of person i at occasion t

Biased correlations

What is the relation among x_{it} and y_{it}?
Sample covariance:

$$cov_{xy} = \frac{1}{n-1}\sum_{1=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Population covariance:

$$COV(X,Y) = E[(X - \mu_x)(Y - \mu_y)] \tag{2}$$

Biased correlations

Inserting our linear model (1) into equation (2). Note that $E(X) = \mu_x$, and $E(Y) = \mu_y$. Hence:

$$
\begin{aligned}
&COV(X,Y) \\
&= E[(L_x + L_{xi} + S_x t_i + S_{xi} t_i + e_{xi} - E(L_x + L_{xi} + S_x t_i + S_{xi} t_i + e_{xi})) \\
&\quad (L_y + L_{yi} + S_y t_i + S_{yi} t_i + e_{yi} - E(L_y + L_{yi} + S_y t_i + S_{yi} t_i + e_{yi}))] \\
&= E[(L_x + L_{xi} + S_x t_i + S_{xi} t_i + e_{xi} - (L_x + 0 + S_x E(t_i) + 0 + 0)) \\
&\quad (L_y + L_{yi} + S_y t_i + S_{yi} t_i + e_{yi} - (L_y + 0 + S_y E(t_i) + 0 + 0))] \ \mid E(t_i) = \bar{t} \\
&= E[(L_{xi} + S_x t_i + S_{xi} t_i + e_{xi} - S_x \bar{t}) \\
&\quad (L_{yi} + S_y t_i + S_{yi} t_i + e_{yi} - S_y \bar{t})] \ \mid \text{multiply all elements} \\
&= E[L_{xi} L_{yi} + L_{xi} S_y t_i + L_{xi} S_{yi} t_i \\
&\quad + S_x t_i L_{yi} + S_x t_i S_y t_i + S_x t_i S_{yi} t_i - S_x t_i S_y \bar{t} \\
&\quad + S_{xi} t_i L_{yi} + S_{xi} t_i S_y t_i + S_{xi} t_i S_{yi} t_i - S_{xi} t_i S_y \bar{t} \\
&\quad - S_x \bar{t} S_y t_i - S_x \bar{t} S_{yi} t_i + S_x \bar{t} S_y \bar{t}]
\end{aligned}
\tag{3}
$$

Biased correlations

The covariance of the fixed effects does not contain any random effects as their expected value is 0. Hence, (3) can be reduced to

$$\begin{aligned} &COV(X,Y) \\ &= E[S_x t_i S_y t_i - S_x t_i S_y \bar{t} - S_x \bar{t} S_y t_i + S_x \bar{t} S_y \bar{t}] \\ &= S_x S_y E(t_i^2 - 2\bar{t}t_i + \bar{t}^2) = S_x S_y E(t_i - \bar{t})^2 \\ &= S_x S_y Var(t_i) \end{aligned} \tag{4}$$

Conclusion

The covariance among x and y depends on the slope parameter and the variance in t_i. With increasing variance in t_i the correlation can take almost any value – even if there is no relation among both variables.

Simulation

Simulate Data with R which correspond to the models in (1).
Aim: Create dataset which contains three variables:

- Age
- x
- y

These values are to be generated given the linear models in (1)
Procedure:

1. Define own function which generates Data
    - Population model
2. Estimate correlation among x and y
3. manipulate t_i to understand its effect on the correlation

Simulation

Generate a function in R with `function()` and pass it to the object `score` which contains our function

```
score <- function(N,L, sdL, S, sdS, t.range){
  # N= number of subj, L= fixed Level; sdL= sd L;
  # t.range= e.g. agerange
  e <- rnorm(N, sd=abs(S)/2)
  ranL <- rnorm(N, sd=sdL)
  ranS <- rnorm(N, sd=sdS)
  vart <- runif(N, min=-(t.range/2), max=t.range/2)
  val <- matrix(ncol=2, nrow=N)
  for(i in 1:N){
  val[i,1] <- vart[i]
  val[i,2]<- L+ranL[i] + S*vart[i] + ranS[i]*vart[i] + e[i]}
  as.data.frame(val)
}
```

Simulation

Our object is `score`

Objects are assigned values and contents:

- Numbers, strings, datasets, functions, etc.

`R` uses the arrow `<-` to assign content to objects.
Here, our function is assigned to the object `score`:

```
score <- function(N,L, sdL, S, sdS, t.range)
```

The arguments `N, L, sdL, S, sdS` and `t.range` are called in our function.

Here: The arguments in the function are the population and simulation values for our model

After the call of `function()` its definition follows in curly brackets `{}`.

Simulation

- # is used to comment the code
- `rnorm(n, mean=0, sd=sdL)`; `n` random draws from a normal distribution $\sim N(0, sdL)$
- `runif(n, min=-(t.range/2), max=t.range/20`; `n` random draw from the uniform distribution
- `matrix()`; a Matrix with `N` rows and 2 columns is assigned to the object `val`
- `for(i in 1:N){...}`; A loop is defined
- In the first line, the i'th value of the vector `vart` is written into the first column and i'th row of `val`

Simulation

- `L+ranL[i] + S*vart[i] + ranS[i]*vart[i] + e[i]` is our model (cf. Equation 1)

  Second line of the loop
  - `L` and `S` are the fixed values
  - `ranL[i]`, `ranS[i]`, `vart[i]` and `e[i]` are random effects
- The index `i` increments after each loop
- `as.data.frame` transforms the matrix into a data set – not necessary but easier to handle later.

Simulation

Population values are taken from Lindenberger and Ghisletta (2009): *Digit Letter*.

- N = 500
- L= 48.29
- sdL= 0.43
- S= -0.81
- sdS= 0.05
- t.range= 10

Our function returns a dataset with 500 observations:

```
> score(500, 48.29, 0.43, -0.81, 0.05, 10)
             V1       V2
1   -4.44477040 53.25156
2   -3.18142497 51.20585
3    2.45730549 45.31237
.     .           .
    < omitted >
.     .           .
498  4.47916833 45.17909
499 -3.71434784 51.15925
500  2.75847019 45.88259
```

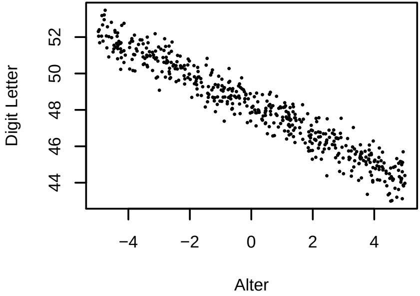
Simulation

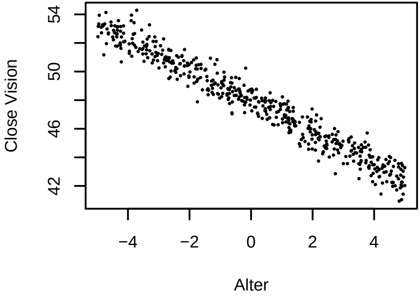
Two variables x and y are generated covering an age-range of 10 years (values from Lindenberger & Ghisletta, 2009):

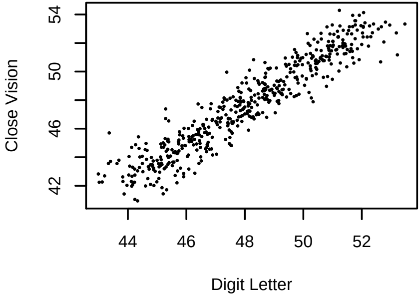
```
# Digit Letter
x <- score(1000, 48.29, 0.43, -0.81, 0.05, 10)
# Close Vision
y <- score(1000, 47.75, 0.37, -1.10, 0.09, 10)
```

These variables are sorted according to age and a new dataset `sim` is created.

```
# Data set with simulated data
sim <- data.frame(x[order(x[,1]),],y[order(y[,1]),2])
names(sim) <- c("Age", "x", "y")
> sim[1:3,]
          Age        x        y
851 -4.971680 51.70464 52.81497
187 -4.964126 52.19422 52.89674
520 -4.963927 51.52887 52.53314
```








```
plot(sim$Age, sim$x, xlab="Age",
     ylab="Digit Letter")
plot(sim$Age, sim$y, xlab="Age",
     ylab="Close Vision")
plot(sim$x,sim$y, xlab="Digit Letter",
     ylab="Close Vision")
```

Simulation

The correlation among x and y is:

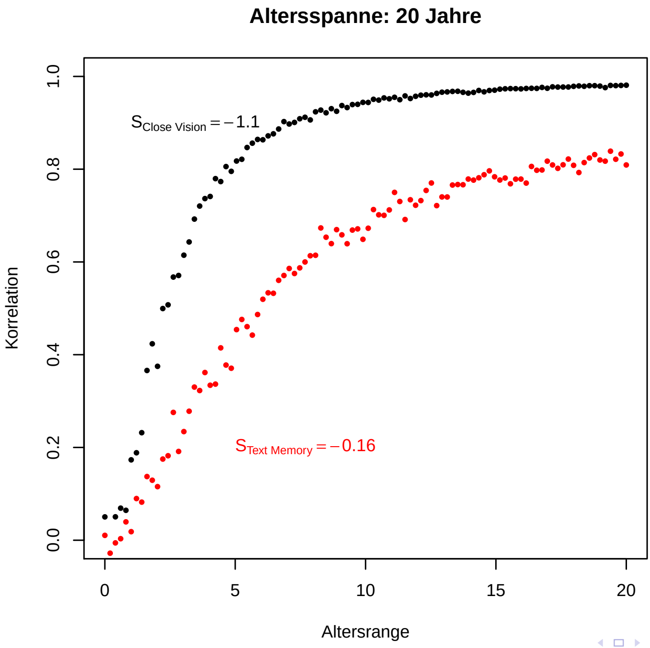
```
> cor.test(sim$x, sim$y)

Pearson's product-moment correlation

data:  sim$x and sim$y
t = 86.1402, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9310502 0.9457991
sample estimates:
      cor
0.9388538
```

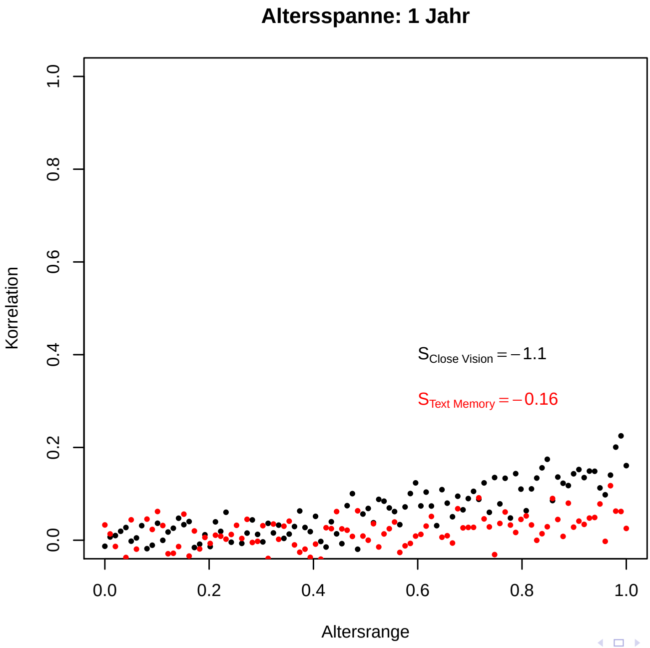
Simulation

Illustration of inflated correlation as function of $var(t_i)$.
Agerange: 0 to 20 years.

```
agerange <- 20
plot(0:agerange, seq(0,1, length.out=agerange+1), xlab="Agerange",
    ylab="Correlation", type="n")
for(a in seq(0,agerange, length.out=100)){
  x <- score(1000, 48.29, 0.43, -0.81, 0.05, a) # Digit Letter
  y <- score(1000, 47.75, 0.37, -1.10, 0.09, a) # close Vision
    # Dataframe with simulated data
  sim <- data.frame(x[order(x[,1]),],y[order(y[,1]),2])
  names(sim) <- c("Age", "x", "y")
  points(a, cor(sim$x, sim$y), cex=.8)
  }
```

Altersspanne: 20 Jahre




Altersspanne: 1 Jahr
Korrelation
Altersrange
0.0
0.2
0.4
0.6
0.8
1.0
$S_{\text{Close Vision}} = -1.1$
$S_{\text{Text Memory}} = -0.16$

Narrow Age Cohort Design
A Solution

Hofer and Sliwinski (2001) suggest the use of "narrow age cohorts" (NAC) designs:

- Narrow age bands
- Bias due to age-heterogeneity is reduced
- Remaining correlations are probably reflecting "true" interrelations
- NAC can be used post-hoc on existing data
- Poblem: Power issues with small age bands

Summary

- Different perspectives on same data
- Transformations can yield different interpretations
- Fundamental: Variance, covariance, correlation
- Correlation is index of consistency of individual differences
- Bias from multiple sources

- Why is this relevant?
- Mixing up sources of variation *is* relevant any application
- Also in longitudinal models

Confounding Sources of Variation

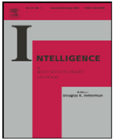
Recent Example

Intelligence 40 (2012) 260–268

Contents lists available at SciVerse ScienceDirect

Intelligence

journal homepage:



Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared

Paolo Ghisletta [a,b,*], Patrick Rabbitt [c,d], Mary Lunn [e], Ulman Lindenberger [f]

Confounding Sources of Variation

Recent Example

sents the MMLM, were Y_{aik} is the cognitive score at age a for individual i on the cognitive task k; I, lS, and qS are the Intercept and the linear and quadratic Slopes, respectively; $\beta_{1,2,3}$ are three retest effects; β_4 is the city effect; β_5 estimates sex' effects; $\beta_{6,7,8,9,10,11}$ are the socio-economic class' effects; and ε_{aik} is the error component.

$$Y_{aik} = I_{ik} + lS_{ik} \cdot A_{ai} + qS_k \cdot A_{ai}^2 + \beta_{(1,2,3)k} \cdot r_{(1,2,3)k} + \beta_{4k} \cdot city_i + \beta_{5k} \cdot sex_i + \beta_{(6,7,8,10,11)k} \cdot soc_{(1,2,3,4,5,6)ik} + \varepsilon_{aik} \quad (1)$$