Advanced Regression: 4 Random effects and hierarchical models

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

7th February 2019

Motivation

All methods presented so far assume that the observations are iid.

iid: Independent and identically distributed

- **Independent**: The observations are independent from each other

$$cor(x_i, x_{i'}) = 0 \text{ for all } i, i' \in 1, ..., n$$

- **Identically**: All observations have the same distribution. For example when assuming a Normal distribution they all have the same mean and variance.

PS: Exchangeability: Allows for dependence between observations and only states that future observations behave like past ones.

Motivation: How realistic is iid?

- Often our data contains structure depending on how our data was sampled.
  - Within K boroughs in London we select n participants ...
  - From K schools we sample n students ...
  - From K hospitals we select n patients ...
  - At K stores we sampled n costumers ...
- $k \in 1, ..., K$ group index

Grouping creates dependence

Observations within a group are likely to be more similar to each other than to observations from other groups.

Motivation: GP data

- We are interested in the relationship of cholesterol and age and how age impacts cholesterol.
- Sampling: We take measurements of patients from certain GPs.
  - Group-level: GPs $K = 12$ `table(data.chol[["doctor"]])`

```
 1  2  3  4  5  6  7  8  9 10 11 12
36 36 36 39 36 36 39 36 36 39 36 36
```

  - Individual-level: Patients $n = 441$

```
head(data.chol)
```

	chol	doctor	age	bmi	agedoc	sex
1	7.13	1	54	27.39	55	0
2	7.70	1	55	29.10	55	0
3	7.30	1	56	27.90	55	0
4	6.89	1	71	26.67	55	1
5	6.90	1	72	26.70	55	1
6	7.90	1	73	29.70	55	1

Pooled analysis

Linear model using all $i = 1, ..., n$ observations ignoring the grouping

$$y_i = \alpha_0 + \beta x_i + \epsilon_i$$

Assumptions

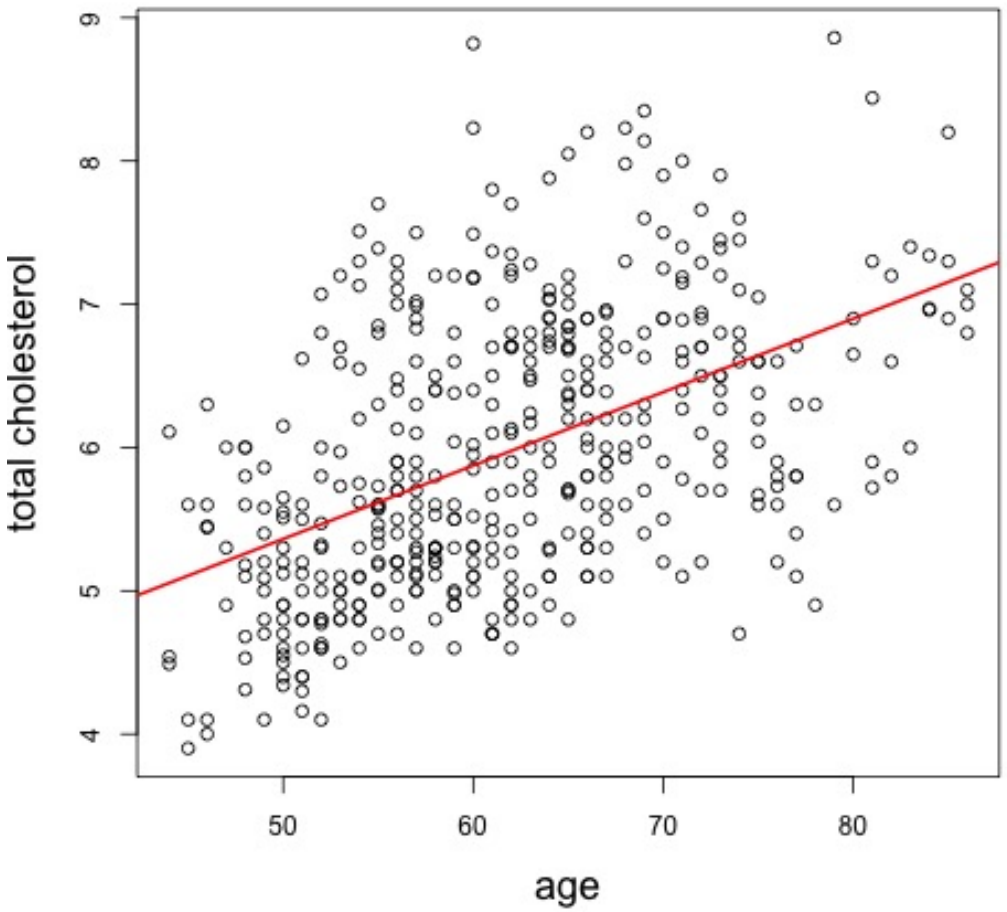
- All observations independent (incorrect).

Consequences

- Estimated errors on regression coefficients are too small.
- Overstate significance of association.

GP data: Pooled analysis

```
Pooled.Model = lm(chol ~ age, data=data.chol)
```



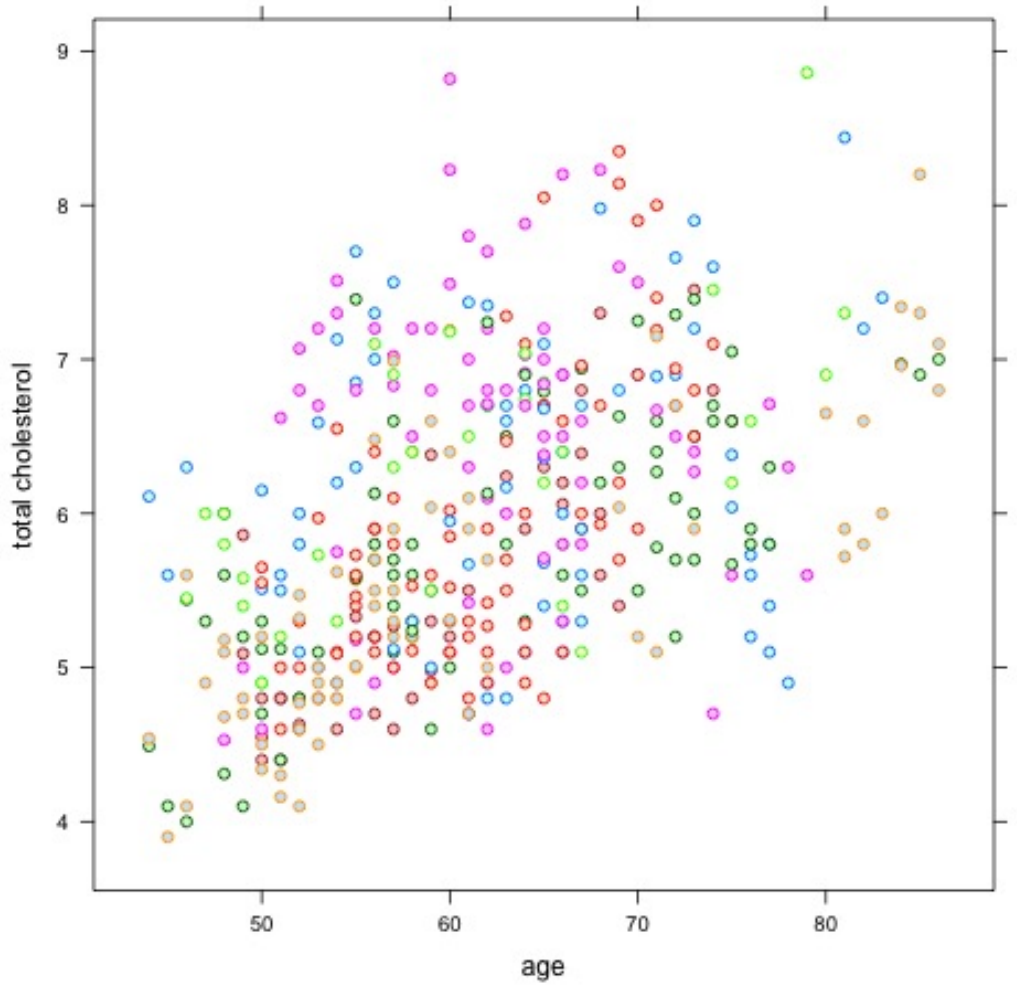
GP data: Pooled analysis

```
Pooled.Model = lm(chol ~ age, data=data.chol)
summary(Pooled.Model)
```

```
Call:
lm(formula = chol ~ age, data = data.chol)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8971 -0.6206 -0.1105  0.5693  2.9456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.798691   0.268571   10.42   <2e-16 ***
age         0.051262   0.004301   11.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8362 on 439 degrees of freedom
Multiple R-squared:  0.2445,    Adjusted R-squared:  0.2428
F-statistic: 142.1 on 1 and 439 DF,  p-value: < 2.2e-16
```

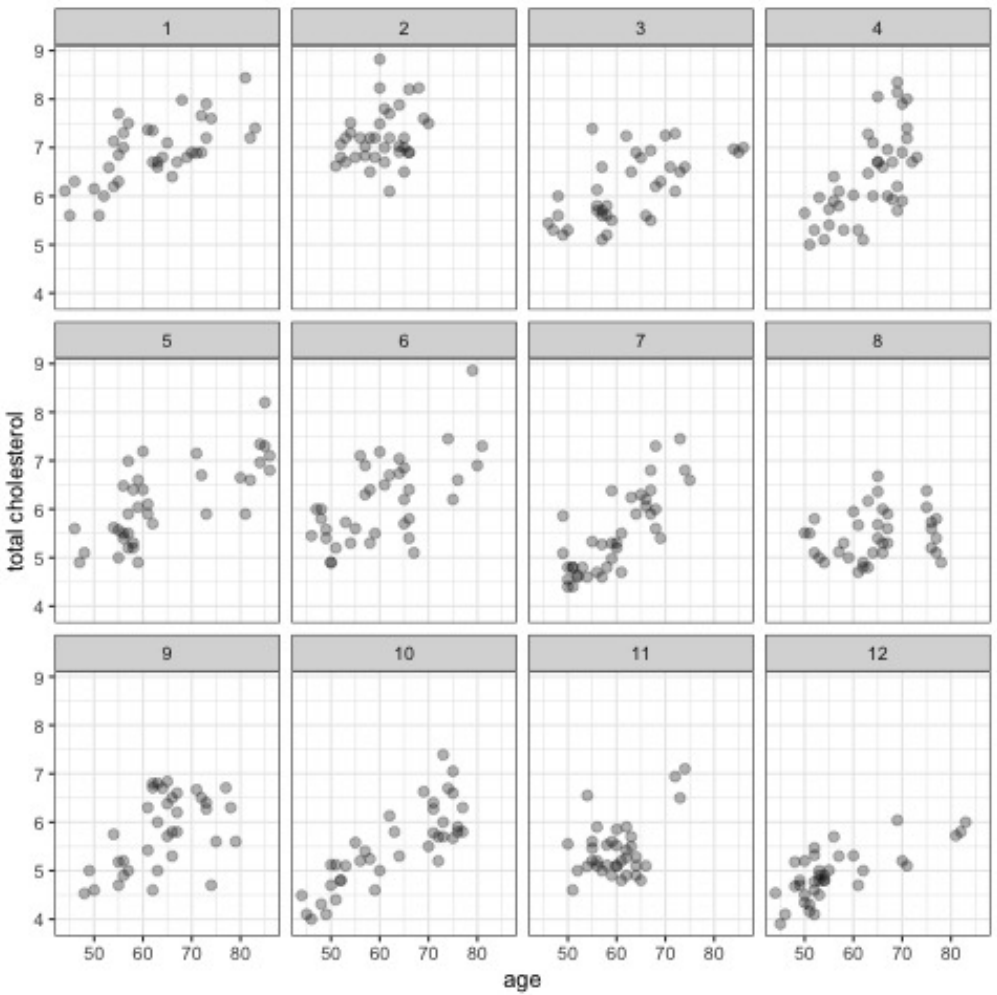
GP data: Pooled analysis

```
xyplot(chol~age, groups = doctor, data=data.chol,
pch = 21)
```



GP data: Pooled analysis

```
ggplot(data.chol, aes(x = age, y = chol, group =
doctor)) + facet_wrap(~doctor)
```



Ignoring dependence

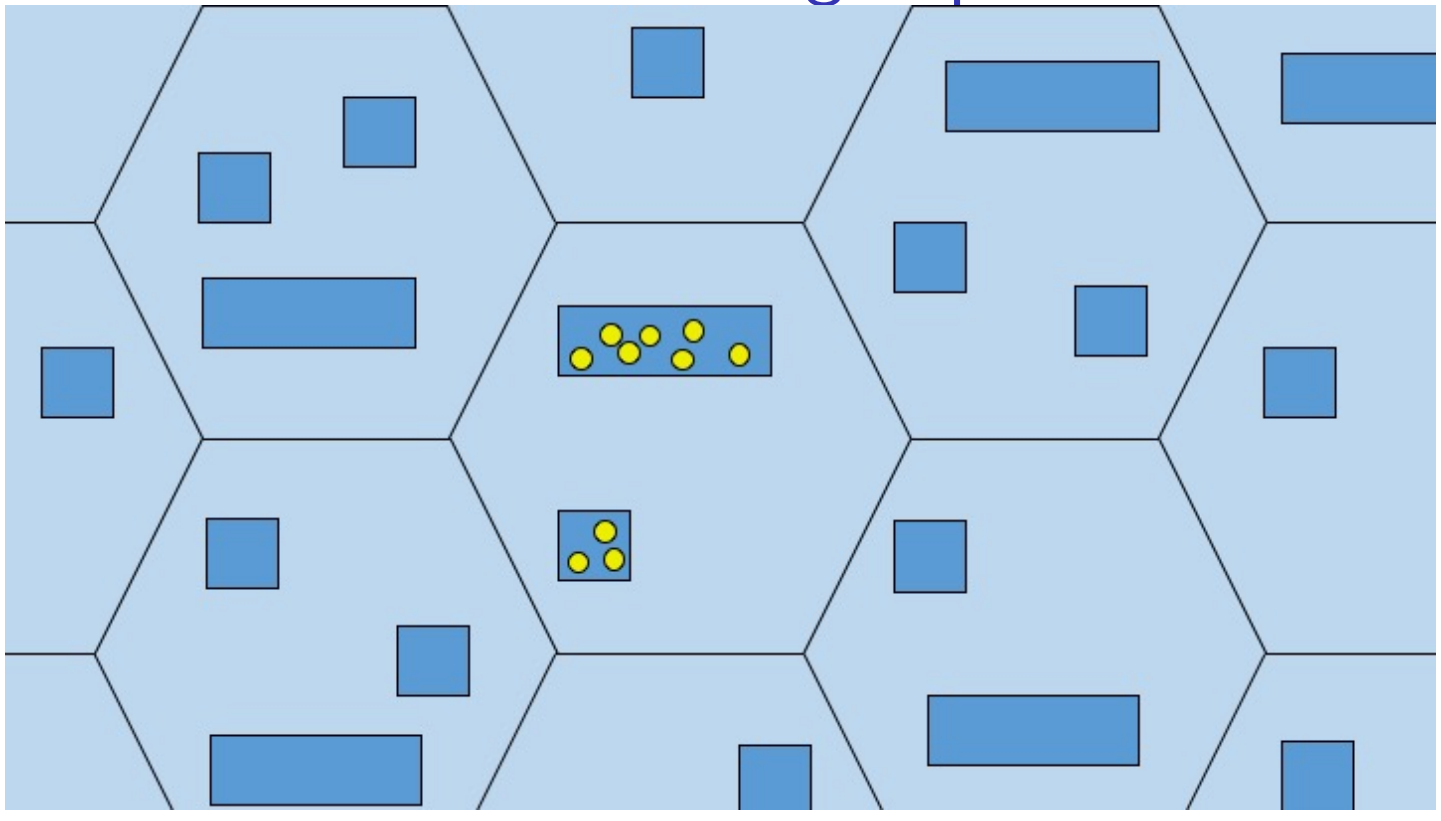
- standard errors too small
- $p-$values too small / confidence intervals too narrow
- over-estimate significance

Intuitively, there is less information in the data than an independent sample.

This has to be taken into account in our models:

1. Perform analysis for each group separately.
2. Calculate summary measures for each group and use standard analysis (Group-level analysis).
3. Fixed effects model to account for group structures.
4. Use random effects models that explicitly model the similarity of observations in a group.

Motivation: Individual-level and group-level



- Observations are grouped with grouping information known.
- Multi-level: Multiple levels of groupings, e.g. classrooms within schools within districts.
- Variables can be measured on the individual and group level.

1. Separate analysis

How to?

- Estimate separate regression coefficients for each group.

Assumptions

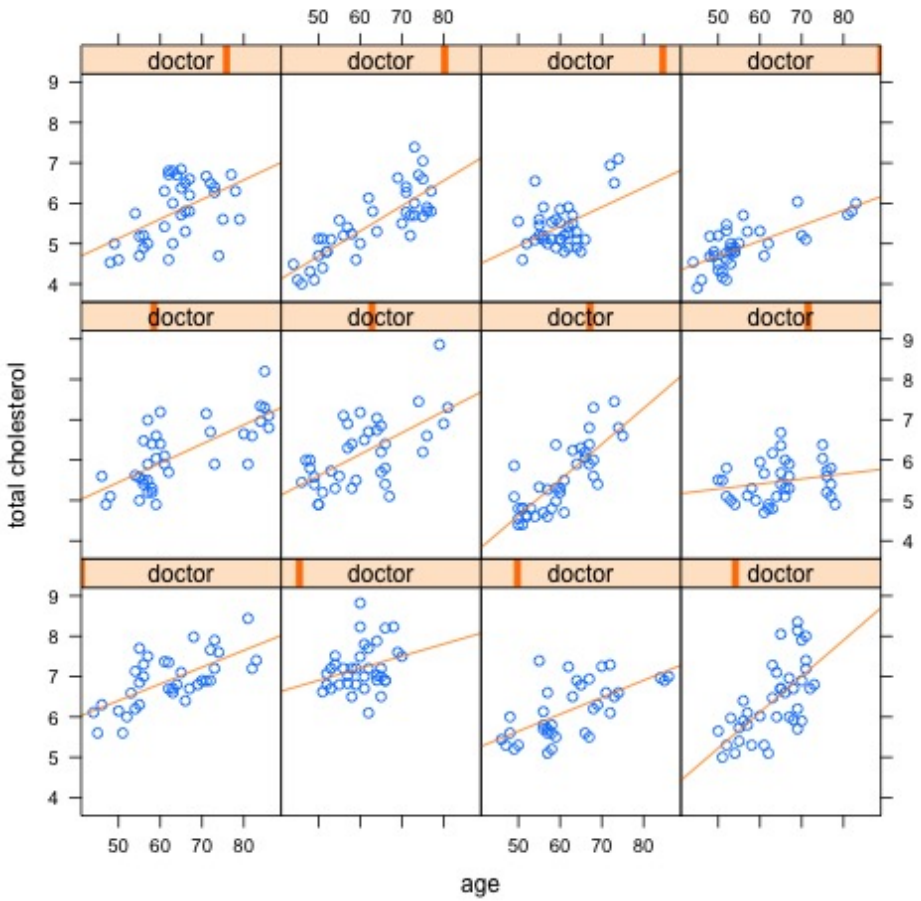
- Independence between groups.

Consequences

- This is a reasonable approach to exploratory analysis.
- If the number of individuals in each group is small, we will get imprecise estimates.
- Multiple testing is an issue.

GP data: Separate analysis

```
xyplot(chol ~ age | doctor, data=data.chol)
```



2. Group-level analysis

How to?

- Summarise outcome and predictors for each group k, e.g. using mean or median.

```
chol.group =
tapply(data.chol$chol,INDEX=data.chol$doctor,FUN=mean)
age.group =
tapply(data.chol$age,INDEX=data.chol$doctor,FUN=mean)
```

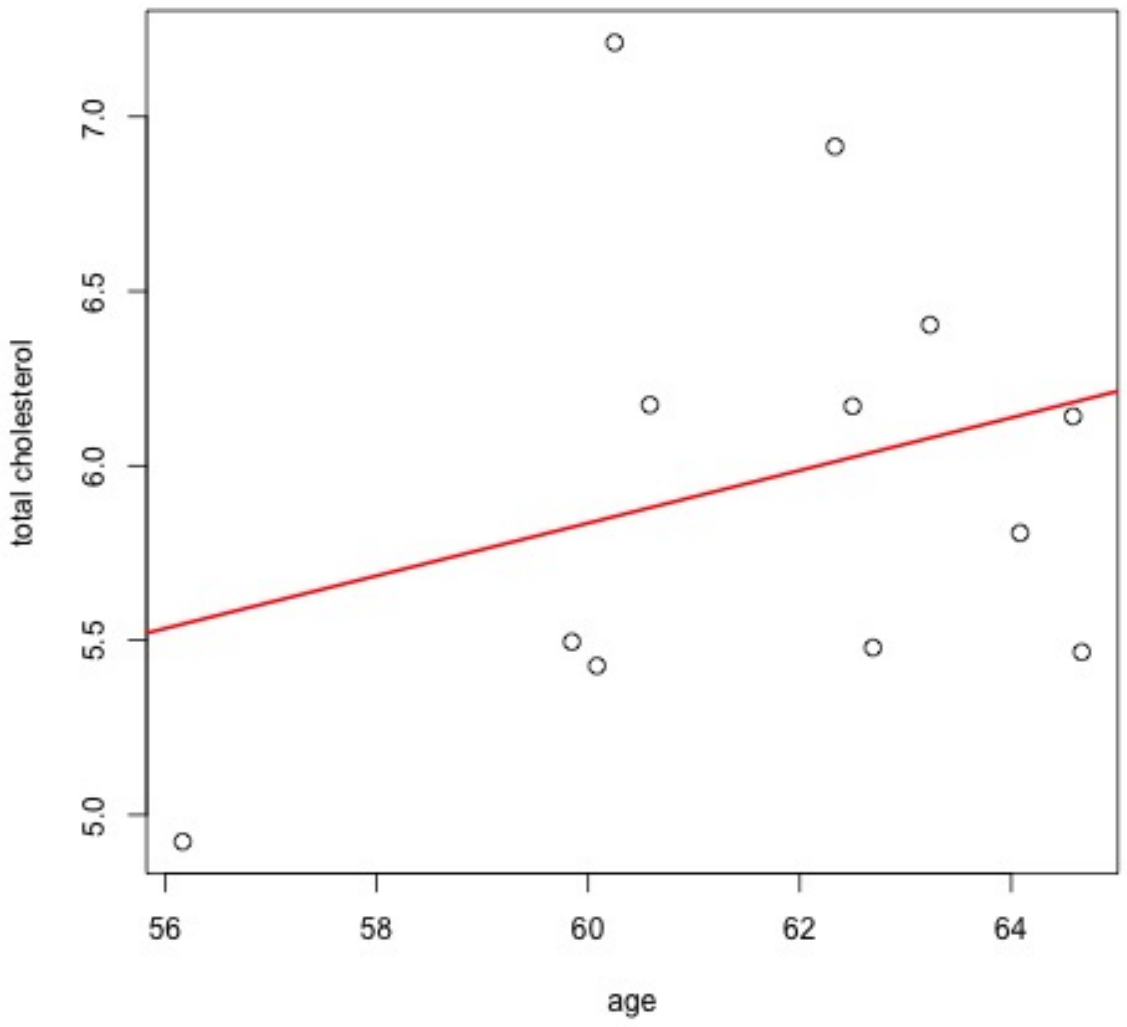
- Treat the group summaries as observations.

```
Group.Model = lm(chol.group ~ age.group)
summary(Group.Model)
```

Assumptions

- One regression line fit: Associations between outcome and predictors are the same for each group.
- Independence between groups.
- All groups are treated equal, irrespective of size.

GP data: Group level analysis



GP data: Group level analysis

```
Call:
lm(formula = chol.group ~ age.group)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7216 -0.4513 -0.1844  0.3020  1.3576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30687    5.05233   0.259    0.801
age.group    0.07548    0.08176   0.923    0.378

Residual standard error: 0.67 on 10 degrees of freedom
Multiple R-squared:  0.07854,   Adjusted R-squared:  -0.0136
F-statistic: 0.8524 on 1 and 10 DF,  p-value: 0.3776
```

Consequences

- This model lacks power as the number of data points used is the number of groups ($k < n$)
- Regression coefficients will be averaged over all groups, $\rightarrow$ real within-group effects may be diluted.

Inverse variance weighted (IVW) meta-analysis

Each random variable is weighted in inverse proportion to its variance.

Assume we have independent observations y_k with variance σ_k. Then the IVW estimate is defined as

$$\hat{y}_{\text{IVW}} = \frac{\sum_{k=1}^{K} y_k/\sigma_k}{\sum_{k=1}^{K} 1/\sigma_k}$$

Weighted regression over groups

Assume y_k is a vector of group summaries, x_k is a $k \times p$ matrix of group summaries. Assume w is a diagonal matrix with $w[k,k] = \frac{1}{\sigma_k^2}$, then the weighted least squares estimate is defined as

$$\hat{\beta}_w = (x_k^t w x_k)^{-1} x_k^t w y_k$$

3. Fixed effects

Motivation:

- Keep the idea of modelling within groups: Allow associations to differ across groups.
- But now we model all the data (n observations) together: Maximise the power to detect associations.

Joint model with group-specific intercept

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

where α_k is a fixed effect.

- α_k captures the effect of unobserved group specific confounders.
- Residual errors $\epsilon_i, i \in 1, ..., n$ are assumed independent.

Fixed effects

How to?

- A fixed effects model is fit in the same way as the simple linear model including the group as a covariate.

Assumptions

- Information on α_k comes from observations in group k only.

Consequences

- By including group effects we have controlled for group characteristics.
- But introduced a large number of parameters (one for each group).
- May be a problem if there are few observations in some groups.

R: Fixed effects in lm()

- Fixed effects in `R` can be computed using the `lm()` model.
- Fixed effects are essentially categorical covariates (`as.factor()`).
- There are two different types of fixed effect:
  1. Group-specific intercept α_k

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

  2. Group-specific slope β_k

$$y_i = \alpha_0 + \beta_k x_i + \epsilon_i$$

R: Group-specific intercept in lm()

1. Group-specific intercept

$$y_i = \alpha_k + \beta x_i + \epsilon_i$$

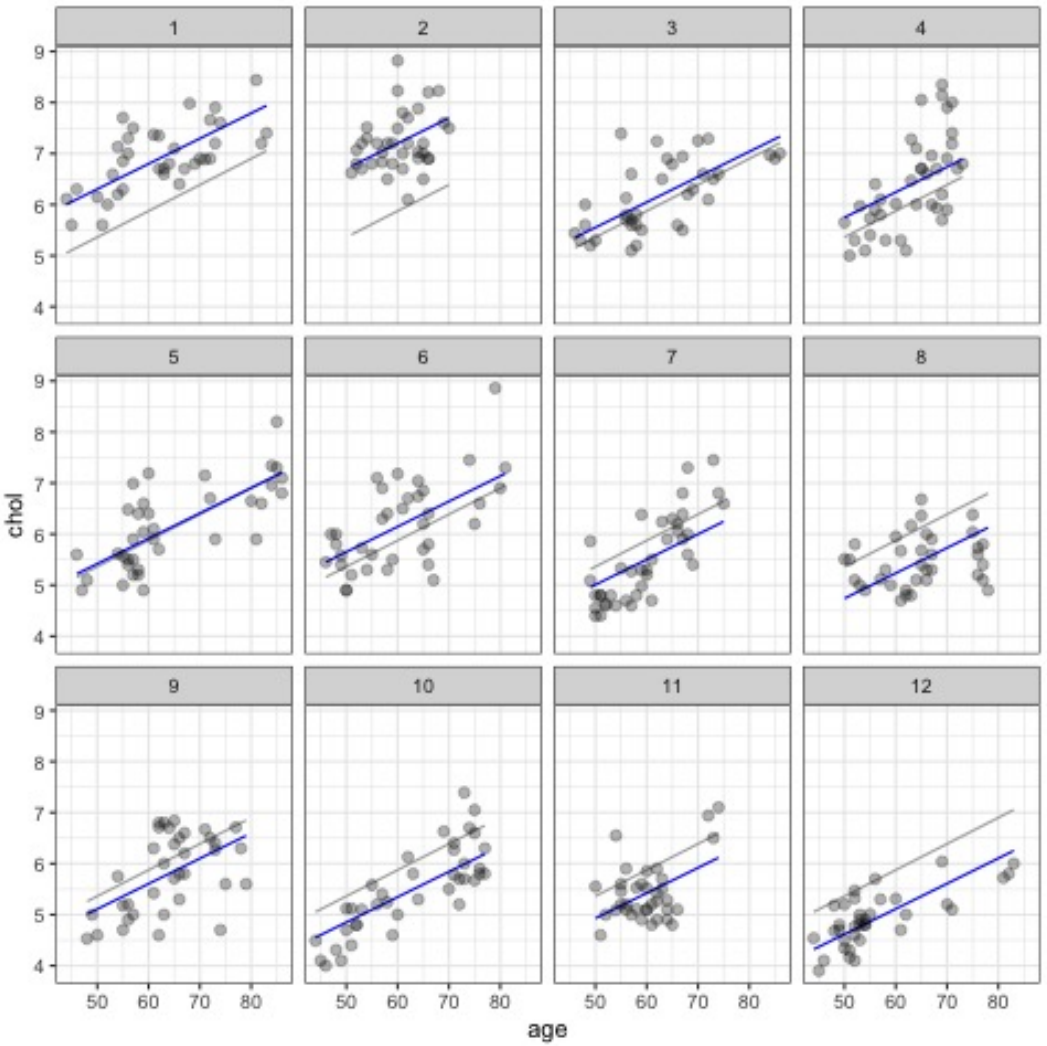
- Add the group variable as additional categorical (`as.factor()`) covariate.
- `Varying.Intercept.Model = lm(chol ~ age + as.factor(doctor), data=data.chol)`

R: Group-specific intercept in lm()

`summary(Varying.Intercept.Model)`

```
Call:
lm(formula = chol ~ age + as.factor(doctor), data = data.chol)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59881 -0.40321 -0.08463  0.37929  1.77313

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.826236   0.213854  17.892  < 2e-16 ***
age                 0.049543   0.003065  16.164  < 2e-16 ***
as.factor(doctor)2  0.400993   0.136014   2.948  0.00337 **
as.factor(doctor)3 -0.752146   0.135865  -5.536 5.41e-08 ***
as.factor(doctor)4 -0.555317   0.133254  -4.167 3.73e-05 ***
as.factor(doctor)5 -0.884528   0.136039  -6.502 2.21e-10 ***
as.factor(doctor)6 -0.653299   0.135970  -4.805 2.15e-06 ***
as.factor(doctor)7 -1.295580   0.133444  -9.709  < 2e-16 ***
as.factor(doctor)8 -1.563657   0.136053 -11.493  < 2e-16 ***
as.factor(doctor)9 -1.193645   0.135970  -8.779  < 2e-16 ***
as.factor(doctor)10 -1.453255   0.133231 -10.908  < 2e-16 ***
as.factor(doctor)11 -1.376027   0.136039 -10.115  < 2e-16 ***
as.factor(doctor)12 -1.685593   0.137173 -12.288  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5764 on 428 degrees of freedom
Multiple R-squared:   0.65,     Adjusted R-squared:  0.6402
F-statistic: 66.24 on 12 and 428 DF,  p-value: < 2.2e-16
```

R: Group-specific intercept in lm()



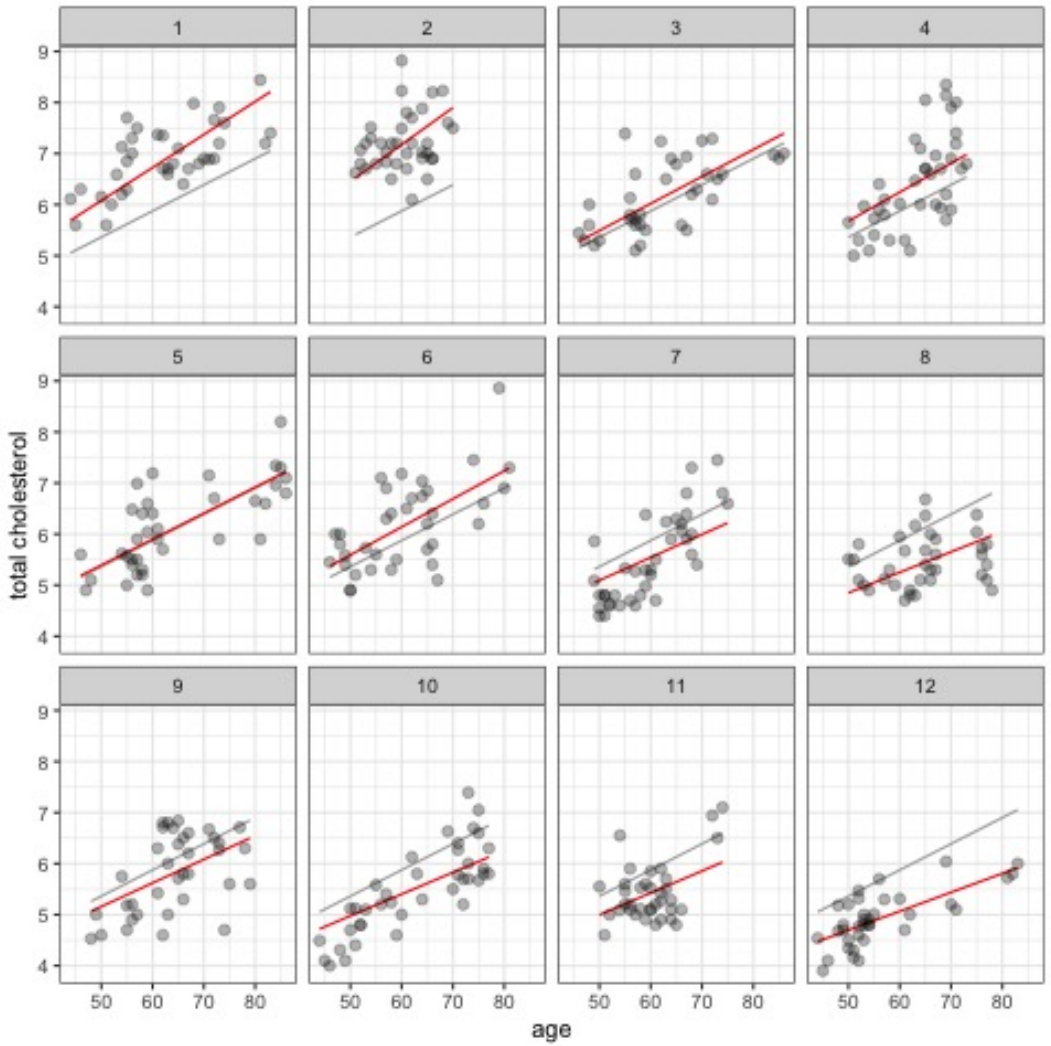
R: Group-specific slope in lm()

2. Group-specific slope

$$y_i = \alpha_0 + \beta_k x_i + \epsilon_i$$

- Add the group variable as an interaction with the predictor of interest.
- `lm(chol ~ age : as.factor(doctor), data=data.chol)`
- : only adds the interaction.

R: Group-specific slope in lm()



R: Fixed effects in lm()

How to specify formulas in the `lm()` function?

- Main formula: `y` $\sim$ `x`, where y is the outcome and x the predictor(s)
- Predictors can be added as:

$+$	main effect
$:$	interaction only
$\star$	main effect and intercept

Values:

- summary()
- coef()
- fitted()

Fixed effects: Disadvantages

- Fixed effects account for **any** unobserved group-specific confounders $\rightarrow$ Including both a group-specific intercept and slope is not identifiable.
  - When the intercept α_k is group-specific, then the slope is assumed to be the same for all groups.
  - When slope β_k is group-specific, then the intercept is assumed to be the same for all groups.
- If we add new groups to the dataset we may not consistently estimate α_k:
  - Consider α_1, the intercept for the first group.
  - When we add new groups, the slope may vary.
  - Changing slope will change the intercept, also α_1.
- Information on α_k or β_k comes only from observations in group k and we need to estimate one parameter per group.

4. Random effects

1. Random effect model with random intercept

$$y_i = (\alpha_0 + u_k) + \beta x_i + \epsilon_i,$$

where $u_k \sim N(0, \sigma_u^2)$

2. Random effects model on both, the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i$$

where $w_k \sim N(0, \sigma_w^2)$

Group effects are random variables, also called random effects.

1. Random effect for the intercept $u_k \sim N(0, \sigma_u^2)$
2. Random effect for the slope $w_k \sim N(0, \sigma_w^2)$

Random intercept

1. Random effect model with random intercept

$$\begin{aligned} y_i &= (\alpha_0 + u_k) + \beta x_i + \epsilon_i, \\ &= \alpha_0 + \beta x_i + u_k + \epsilon_i, \end{aligned}$$

- Where α_0 is the intercept and β the regression coefficient.
- There are two distinct error terms
  1. Group-specific error
  $$u_k \sim N(0, \sigma_u^2)$$
  2. Individual-specific error
  $$\epsilon_i \sim N(0, \sigma^2)$$
- Note that u_k and ϵ_i are independent of each other.

Random effect model with random intercept

Interpretation of random intercept α_k:

$$\alpha_k = (\alpha_0 + u_k)$$

- α_0 is the global intercept
- u_k group-level variations around the global intercept

This is equivalent to assuming α_k is a **random variable** that follows a Normal distribution

$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Multi-level interpretation (two levels of variability):

1. First level
   Defined on the individual level for observation $i = 1, ..., n$, similar to a standard linear regression

$$y_i = \alpha_k + \beta x_i + \epsilon_i,$$

2. Second level
   But the intercept is not fixed, it is a random variable

$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Assumptions

- Slope of regression line is the same across all groups. Each group has a different intercept (α_k).
- But $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ has now a common distribution which is estimated from **all observations**, and not just from the observations in a specific group as in fixed effects.
- We pool information across groups.

Consequences

- We control for group characteristics by including the group-specific intercept.
- Number of group-specific parameters to estimate is much smaller than in the fixed effect models (σ_u^2 vs k intercepts).

(Restricted) Maximum Likelihood estimation of random effect

$$y_i = \alpha_0 + \beta x_i + u_k + \epsilon_i,$$

Parameters to estimate are

- α_0, β intercept and regression coefficient
- σ_u^2, σ^2 variance components

Maximum Likelihood estimation is based on the Normal distribution of u_k and ϵ_i

- ML estimate for σ_u^2 requires subtracting 2 empirical estimates of variance $\rightarrow$ ML estimates for σ_u^2 can be negative.
- Restricted Maxium Likelihood (REML): Imposes positivity constraints on the variance estimates.

Random intercept in R

Implementations of Restricted Maximum Likelihood (REML) in R

- `lmer` function in the `lme4` package
- `lme` function in the `nlme` package

Focus here is the `lme` function in the `nlme` package.

`lme(fixed, data, random)`

- `fixed`: Formula $y \sim x$
- `random`: Formula $\sim 1 \mid factor$
- `data`: Dataset to use

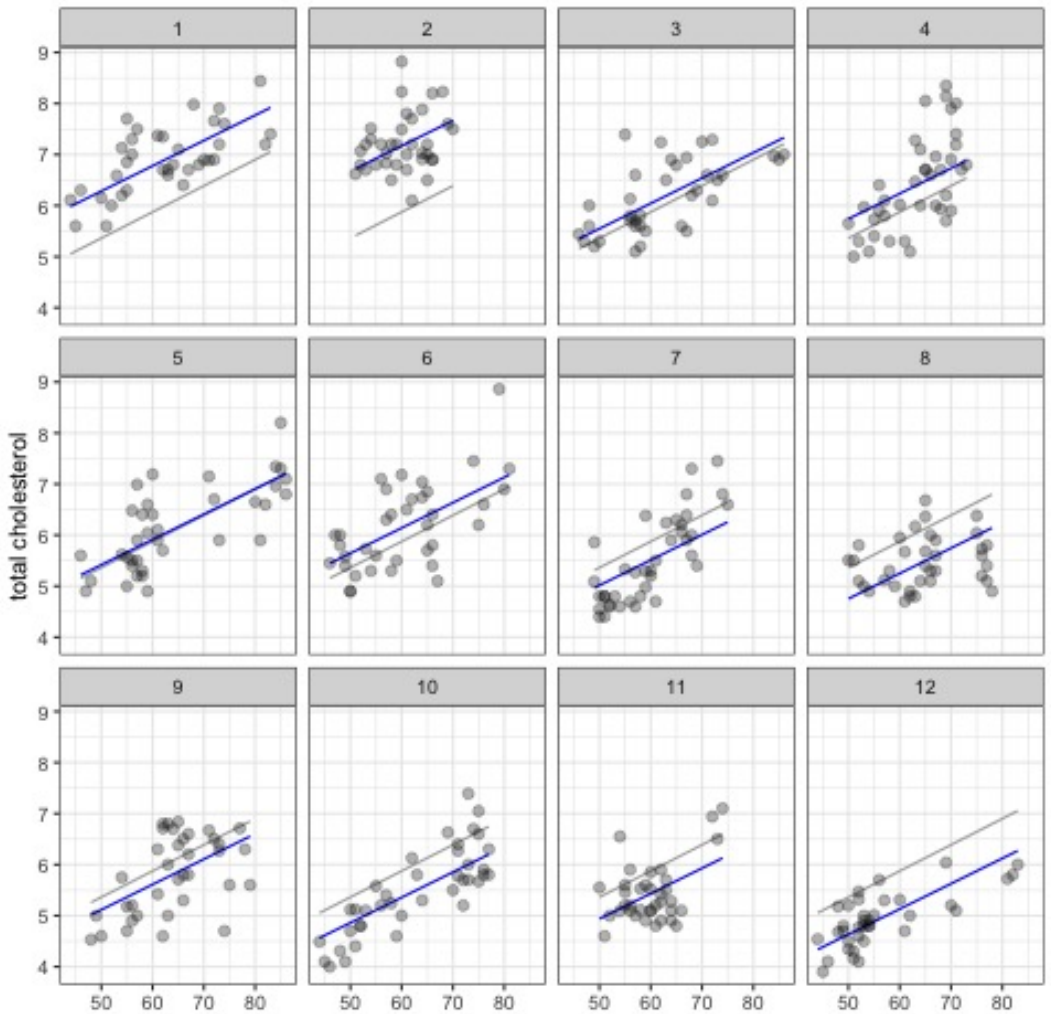
R: Random intercept using lme

```
RandomIntercept = lme( chol ~ age, random = ~ 1 |
doctor, data = data.chol)
summary(RandomIntercept)
```

```
Linear mixed-effects model fit by REML
 Data: data.chol
      AIC      BIC    logLik
  828.697 845.035 -410.3485

Random effects:
 Formula: ~1 | doctor
        (Intercept)  Residual
StdDev:   0.6347908 0.5764246

Fixed effects: chol ~ age
                Value  Std.Error  DF  t-value p-value
(Intercept) 2.9060357 0.26477408 428 10.97553       0
age         0.0495831 0.00306279 428 16.18888       0
```

R: Random intercept using `lme`

```
RandomInterceptPredictions = fitted(RandomIntercept)
```



Random effect model and variance partition

Variance decomposition for observation i in group k

$$\begin{aligned} var(y_i) &= var(u_k + \epsilon_i) \\ &= var(u_k) + var(\epsilon_i) + 2cov(u_k, \epsilon_i) \\ &= \sigma_u^2 + \sigma^2 + 0 \end{aligned}$$

Further we can look at the covariance of observations

- i and i' within group k

$$cov(y_i, y_{i'}) = cov(u_k + \epsilon_i, u_k + \epsilon_{i'}) = \sigma_u^2$$

- i and i' from different groups k and k'

$$cov(y_i, y_{i'}) = cov(u_k + \epsilon_i, u_{k'} + \epsilon_{i'}) = 0$$

Random effect model and variance partition

Variance decomposition of variability between and within groups

Intra-class correlation coefficient ρ

$$\rho = cor(y_i, y_{i'}) = \frac{cov(y_i, y_{i'})}{\sqrt{var(y_i)var(y_{i'})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

Interpretation:

- Intra-class correlation coefficient ρ is the correlation between two observations i and i' in the same group.
- It is the ratio of between-group variance σ_u^2 over the total variance.
- If $\rho \to 0$ there is little variation explained by the grouping and we might consider a model without the random effect.

Variance partition in R

```
summary(RandomIntercept)
```

```
Random effects:
 Formula: ~1 | doctor
        (Intercept)  Residual
StdDev:   0.6347908 0.5764246
```

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = \frac{0.6347908^2}{0.6347908^2 + 0.5764246^2} \approx 0.54$$

Interpretation:

- There is substantial evidence for between-group heterogeneity.
- More than half of the total variance can be explained by the between-group variance.
- It is beneficial to include the random effects on the intercept.

Random effect model with random intercept and random slope

2. Random effects model on both, the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i$$

- There are three distinct error terms
  1. Group-specific error of the intercept

$$u_k \sim N(0, \sigma_u^2)$$

  2. Group-specific error of the regression slope

$$w_k \sim N(0, \sigma_w^2)$$

  3. Individual-specific error

$$\epsilon_i \sim N(0, \sigma^2)$$

- Note that u_k and w_k are correlated and independent of ϵ_i.

Random effect model with random intercept and random slope

Assumptions

- Each group has a different intercept ($\alpha_k = \alpha_0 + u_k$) and a different regression slope ($\beta_k = \beta + w_k$).
- We allow for correlation between α_k and β_k.
- Both, $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ and $\beta_k \sim N(\beta, \sigma_w^2)$ have a common distribution which is estimated from **all observations**, and not just from the observations in a given group as in fixed effects.
- We pool information across groups.

Consequences

- Including a random slope can be interpreted as creating an interaction between the group and the strength of association.
- We only have three additional parameters in the model: σ_u^2, σ_w^2 and $cor(\sigma_u, \sigma_w)$.

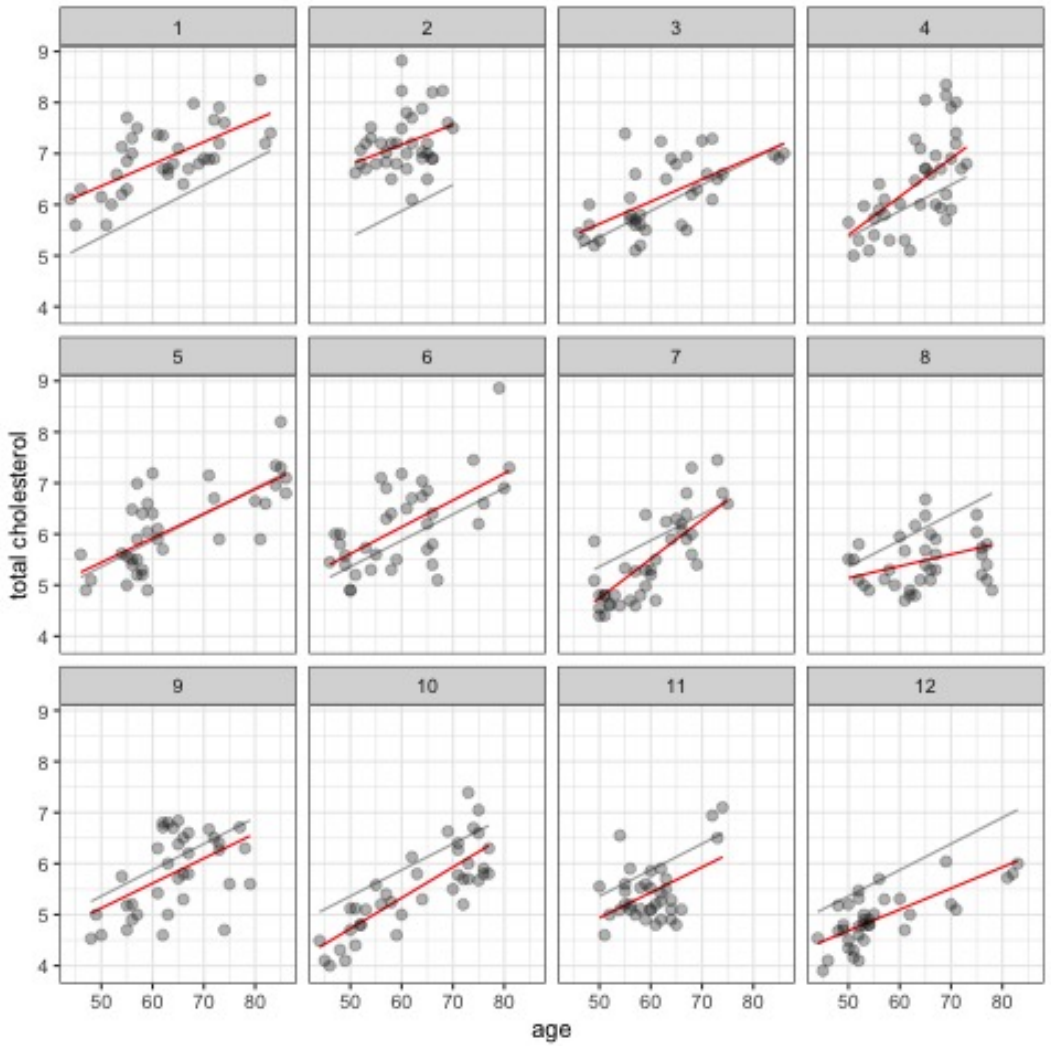
R: Random intercept and slope using lme

```
RandomSlope = lme( chol ~ age, random = ~ 1+age |
doctor, data = data.chol)
summary(RandomSlope)
```

```
Linear mixed-effects model fit by REML
 Data: data.chol
       AIC      BIC    logLik
  821.9886 846.4956 -404.9943

Random effects:
 Formula: ~1 + age | doctor
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.28163791 (Intr)
age         0.01771585 -0.872
Residual    0.55997509

Fixed effects: chol ~ age
               Value Std.Error  DF  t-value p-value
(Intercept) 2.8791744 0.4215200 428 6.830458       0
age         0.0500704 0.0060597 428 8.262837       0
```

R: Random intercept and slope using `lmer`

```
RandomSlopePredictions = fitted(RandomSlope)
```



Variables on individual level and group level

When considering variables or predictors we need to distinguish:

- Individual-level variables
- Group-level variables, that are the same for all observations in a group

GP example:

- Individual-level variables: Age and sex of patient
- Group-level variables: Age of doctor

	chol	doctor	age	bmi	agedoc	sex
1	7.13	1	54	27.39	55	0
2	7.70	1	55	29.10	55	0
3	7.30	1	56	27.90	55	0
4	6.89	1	71	26.67	55	1
5	6.90	1	72	26.70	55	1
6	7.90	1	73	29.70	55	1

Variables on individual level and group level

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \gamma x_g + \epsilon_i$$

Example: GP data

```
RandomCov = lme( chol ~ age + agedoc, random = ~
1+age | doctor, data = data.chol)
summary(RandomCov)
```

```
Fixed effects: chol ~ age + agedoc
                 Value Std.Error  DF   t-value p-value
(Intercept) -2.7897788 1.1824050 428 -2.359411  0.0188
age          0.0501492 0.0060673 428  8.265423  0.0000
agedoc       0.1280030 0.0253576  10  5.047908  0.0005
```

Model comparison

- Likelihood-ratio test for nested models:
  Models must have the same fixed effects. Does not work with group-level covariates.
- Akaike information criterion (AIC)

GP example:

- Model A (Random intercept)
  `modelA = lme( chol ~ age, random = ~1 | doctor, data = data.chol)`
- Model B (Random intercept and slope)
  `modelB = lme( chol ~ age, random = ~ 1+age | doctor, data = data.chol)`
- Model C (Random intercept and slope and group covariate)
  `modelC = lme( chol ~ age + agedoc, random = ~ 1+age | doctor, data = data.chol)`

Model comparison

- Likelihood-ratio test for nested models
  (Model A is nested in Model B)
  `anova(modelA,modelB)`

```
> anova(modelA, modelB)
       Model df      AIC      BIC    logLik   Test L.Ratio p-value
modelA     1  4 828.6970 845.0350 -410.3485
modelB     2  6 821.9886 846.4956 -404.9943 1 vs 2 10.7084  0.0047
```

- AIC for non-nested models
  `anova(modelB,modelC)`

```
> anova(modelB, modelC)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
modelB     1  6 821.9886 846.4956 -404.9943
modelC     2  7 815.6956 844.2712 -400.8478 1 vs 2 8.292926   0.004
Warning message:
In anova.lme(modelB, modelC) :
  fitted objects with different fixed effects. REML comparisons are not
meaningful.
```

Generalised linear mixed models

- Generalised Linear Mixed models (GLMM) can be used to adapt linear mixed models to outcomes that do not follow a Normal distribution.
- The package `lme4` includes the function `glmer` that can fit GLMMs.

```
glmer(formula, family = gaussian)
```

Formula:

- `y~x` to specify outcome and predictors
- `+ (1 | factor)` add random intercept depending on factor
- `+ x + (x | factor)` add random slope depending on factor

Take away: Structured Data

- Most statistical methods are developed for independent and identically distributed (iid) data.
- But often in practice we observe structured data, where there is an intrinsic group structure.
- Grouping creates dependence: Observations within a group are likely to be more similar to each other than to observations from other groups.
- Ignoring the group structure can lead to over-confident results or even false positives.
- Analysing each group separately, we do not assume any shared mechanisms and need to fit a model on the samples within a group only.
- Aggregating and working only on the group-level drastically reduces the sample size k.

Take away: Fixed and random effect

- Fixed effect models can account for group structure but many parameters need to be estimated and no information is shared between groups.
- Random effect models treat group-specific parameters as random variables.
- Instead of estimating one parameter for each group, random effect models only estimate the distribution parameter of the random variable.
- Thus, they pool information across groups.
- The intra-class coefficient gives a measure of how relevant the group structure is.
- Implementation in `R`: `lme()` function in the `nlme` package.
- Models including both, fixed and random effects, are often called linear mixed models.

Outlook:

Lectures today by Deborah Schneider Luftman:

- Non-linear models (lowess, spline, GAM)
- Non-parametric models (decision trees and random forests)

Practical next week: The epigentic clock

- Hierachical and non-linear models in R.
- Build a non-parametric prediction rule using random forests.
- Evaluate your prediction rules of the epigenetic clock on a new data set and decide if mice exposed to nitrogen dioxide have a reduced biological age than control mice.