De-mystifying random effects models

Peter J Diggle

Lecture 4, Leahurst, October 2012

Linear regression

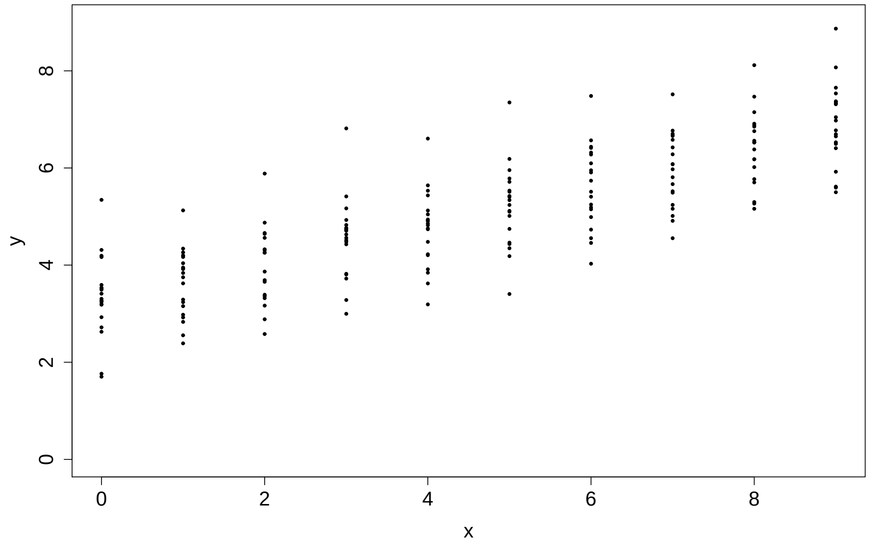
- **input variable x**

  factor, covariate, explanatory variable, ...

- **output variable y**

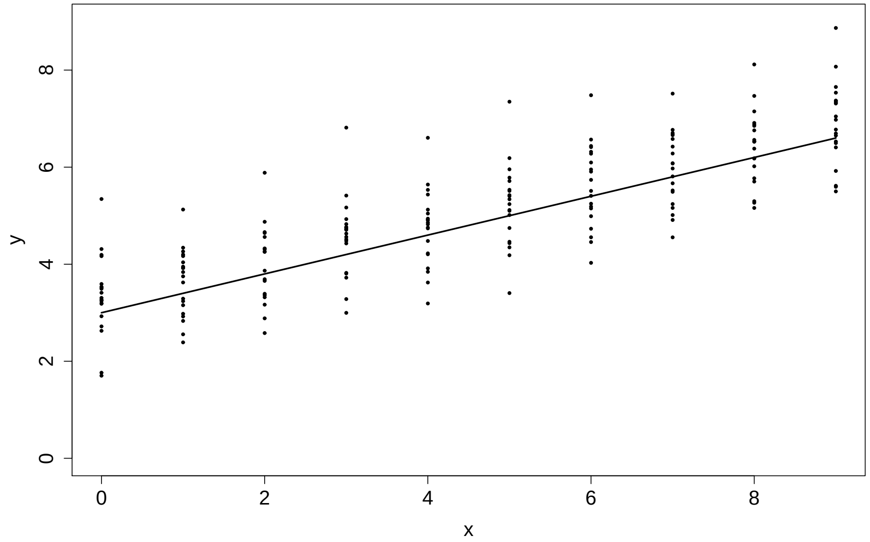
  response, end-point, primary outcome,...

A synthetic example



- relationship between x and y can be captured approximately by a straight line
- scatter about the line is approximately the same at all values of x

Interpreting the linear regression model



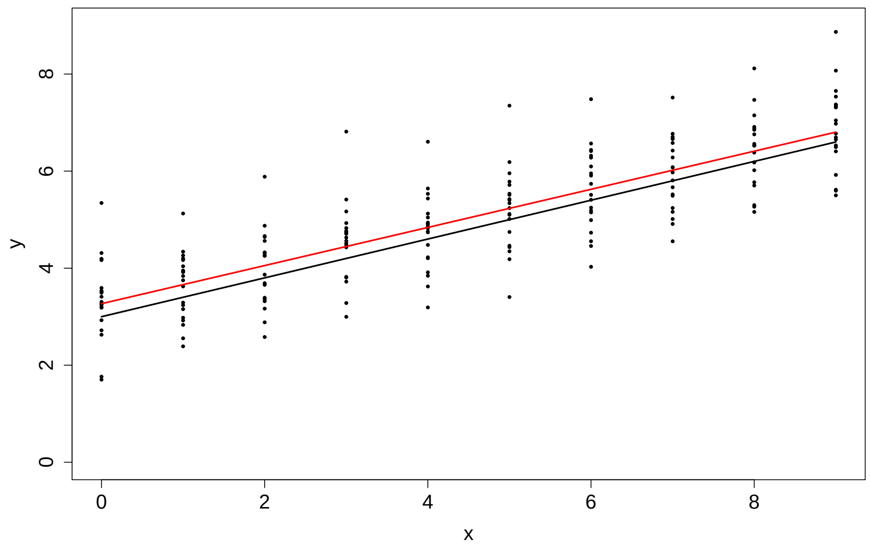
$$y = \alpha + \beta x + z$$

- **intercept α: predicted value of y when $x = 0$**
- **slope β: predicted change in y for unit change in x**
- **residual z: difference between actual and predicted y**

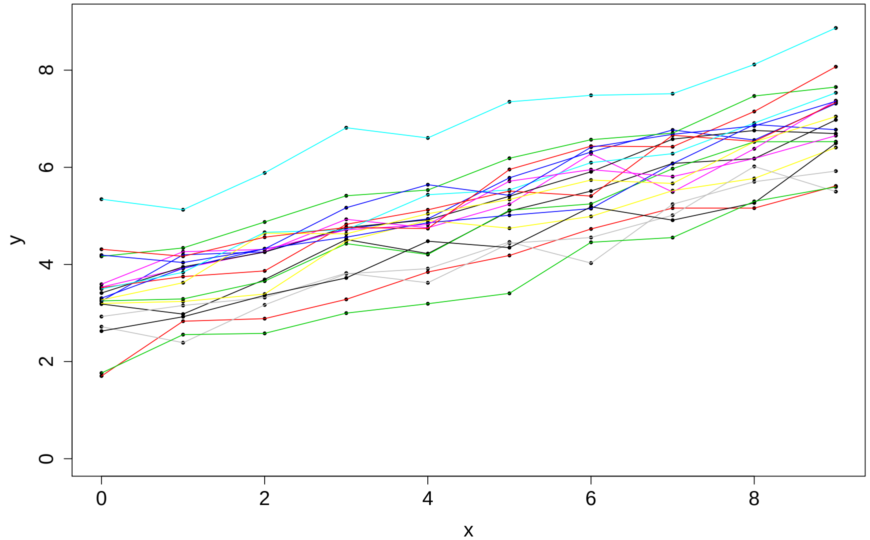
Precision

- **how precisely can we estimate the straight-line relationship?**

- **how precisely can we predict a future value of y?**



The sting in the tail: longitudinal studies



- simple linear regression software assumes that data are uncorrelated
- in longitudinal studies, with repeated measurements on each subject, this is rarely true
- as a consequence, nominal standard errors, p-values,... are WRONG

```
> fit1<-lm(y~1+x)
> summary(fit1)

... plus lots of stuff you don't want to know

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.26789    0.10274   31.81   <2e-16 ***
x            0.39286    0.01924   20.41   <2e-16 ***

Residual standard error: 0.7817 on 198 degrees of freedom
Multiple R-squared: 0.6779,     Adjusted R-squared: 0.6763
F-statistic: 416.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
> library(nlme)
> fit2<-lme(y~1+x,random=~1|id)
> summary(fit2)
Linear mixed-effects model fit by REML

... plus lots of stuff ...

Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:   0.7477531 0.2730349

Fixed effects: y ~ 1+x
               Value  Std.Error  DF  t-value p-value
(Intercept) 3.267887 0.17100989 179 19.10935       0
x           0.392856 0.00672165 179 58.44637       0

Number of Observations: 200
Number of Groups: 20
```

Random effects

- random effects can be thought of as missing information on individual subjects that, were it available, would be included in the statistical model
- to reflect our not knowing what values to use for the random effects, we model them as a random sample from a distribution
- this induces correlation amongst repeated measurements on the same subjects

Example: some subjects are intrinsically high responders, others intrinsically low responders

Correlation doesn't always hurt you

Model	$\hat{\alpha}$	SE	$\hat{\beta}$	SE
fixed effects	3.268	0.103	0.393	0.019
random effects	3.268	0.171	0.393	0.007

Random effects or fixed effects?

For our synthetic example, write:

$$\begin{array}{rcl} Y_{ij} & = & j^{th} \text{ response from } i^{th} \text{ subject: } i = 1, ..., n \\ x_{ij} & = & \text{corresponding value of explanatory variable} \end{array}$$

A random effects model

1. $Y_{ij} = \alpha + \beta x_{ij} + U_i + Z_{ij}$
2. U_i = random effect for subject i
3. Z_{ij} = residual
4. all U_i and all Z_{ij} mutually independent
5. different responses on same subject positively correlated:

$$\rho = \frac{\mathrm{Var}(U)}{\mathrm{Var}(U) + \mathrm{Var}(Z)}$$

Random effects or fixed effects?

For our synthetic example, write:

$$\begin{aligned} Y_{ij} &= j^{th} \text{ response from } i^{th} \text{ subject: } i = 1, ..., n \\ x_{ij} &= \text{corresponding value of explanatory variable} \end{aligned}$$

A fixed effects model

1. $Y_{ij} = \alpha_i + \beta x_{ij} + Z_{ij}$
2. α_i = intercept for subject i
3. Z_{ij} = residual
4. all Z_{ij} mutually independent
5. different responses on same subject uncorrelated

Which model is correct?

1. they both are

2. the choice between them depends primarily on why you are analysing the data

   (a) to find out about the particular subjects in your data

   (b) to find out about the population from which your subjects were drawn

   Cases (a) and (b) call for fixed effects and random effects models, respectively

3. and secondarily on considerations of statistical efficiency: for large n, fixed effects model has many more parameters

A philosophical objection to fixed effects?

Y_{ij} = mark for student i on exam paper $j = 1, \ldots, p$

Question: what overall mark should you give to student i?

Fixed effects model: $Y_{ij} = \alpha_i + Z_{ij}$

- $Z_{ij} \sim \mathrm{N}(0, \tau^2)$, independent

Answer: $\hat{\alpha}_i = \bar{Y}_i$ (observed average mark for student i))

Random effects model: $Y_{ij} = A_i + Z_{ij}$

- $Z_{ij} \sim \mathrm{N}(0, \tau^2)$, independent
- $A_i \sim \mathrm{N}(\alpha, \sigma^2)$, independent

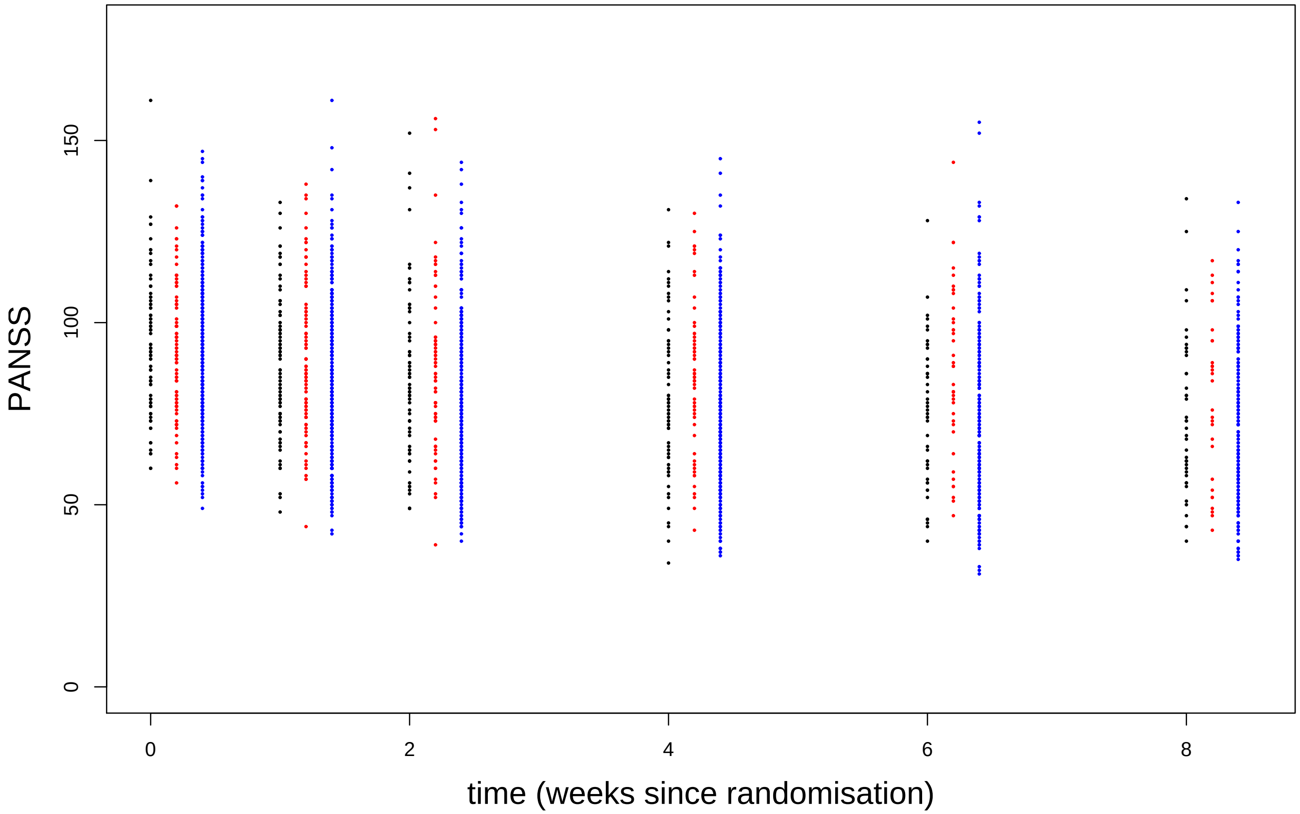
Answer: $\hat{A}_i = c \times \bar{y}_i + (1 - c) \times \bar{y} \qquad c = p/(p + \tau^2/\sigma^2)$

An RCT of drug therapies for schizophrenia

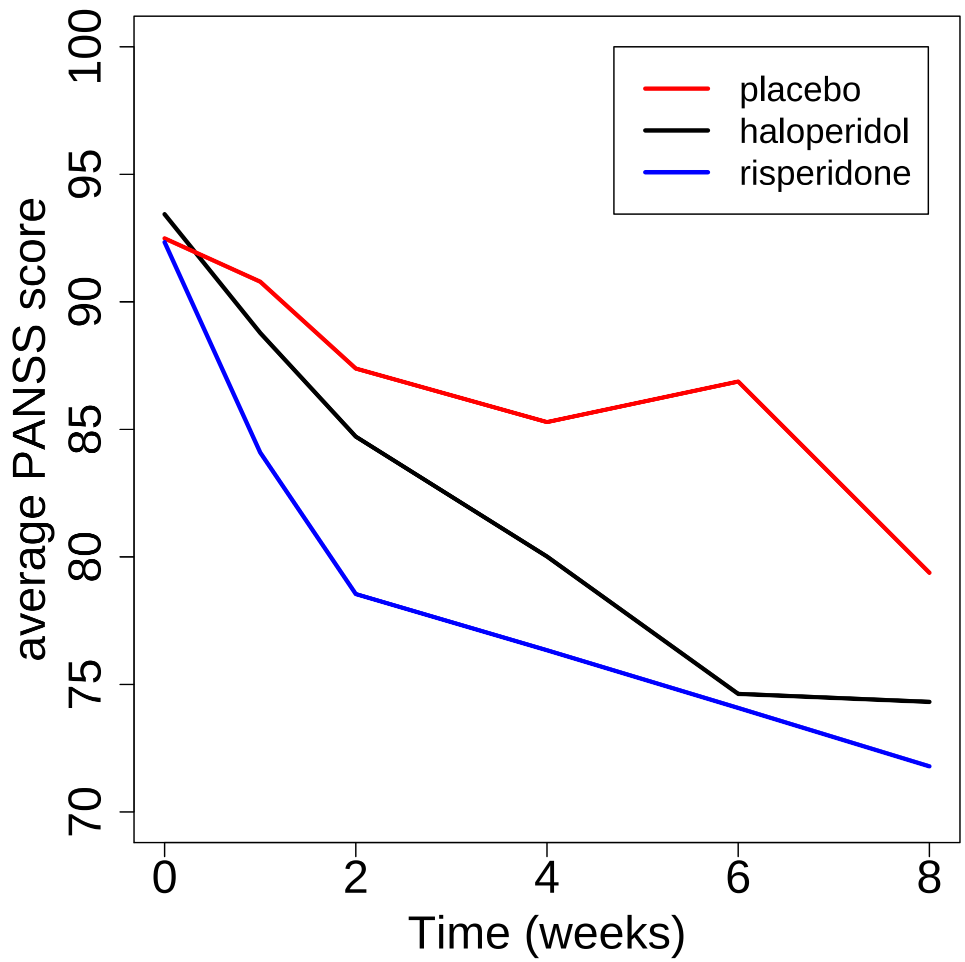
- randomised clinical trial of drug therapies
- three treatments:
  - haloperidol (standard)
  - placebo
  - risperidone (novel)
- dropout due to "inadequate response to treatment"

Treatment	Number of non-dropouts at week					
	0	1	2	4	6	8
haloperidol	85	83	74	64	46	41
placebo	88	86	70	56	40	29
risperidone	345	340	307	276	229	199
total	518	509	451	396	315	269

The schizophrenia trial data



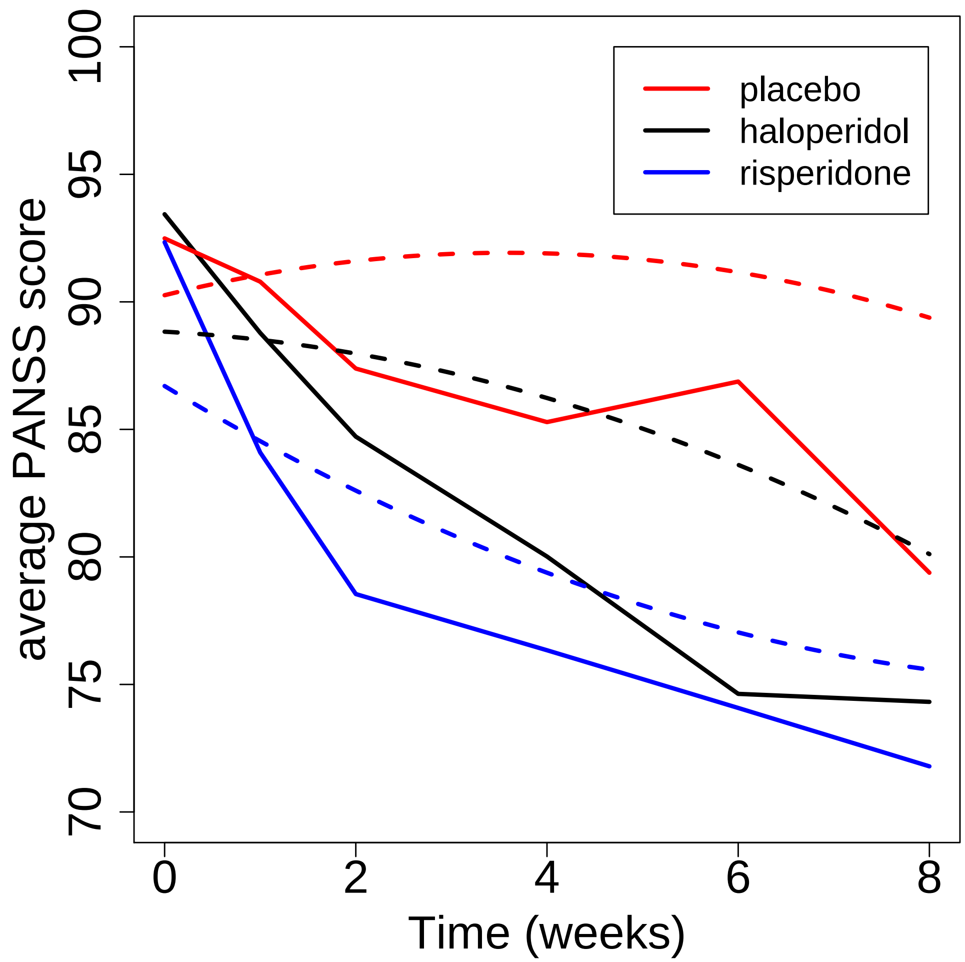
A summary of the schizophrenia trial data



A model for the schizophrenia trial data

- mean response depends on treatment and time
- two random effects:
  - between subjects (high or low responders)
  - between times within subjects (good and bad days)
- method of analysis allows for dropouts

PANSS mean response profiles



What's going on?

- **dropout is selective (high responders more likely to leave)**
- **but the data are correlated**
- **and this allows the model to infer what you would have seen, had there not been any dropouts**
- **which may or may not be what you want**

Closing remarks

- **fixed effects describe the variation in average responses of groups of subjects according to their measured characteristics (age, sex, treatment,...)**
- **random effects describe variation in subject-specific responses according to their unmeasured characteristics**
- **both kinds of model can easily be fitted using open-source software (`R`), or in various proprietary packages**
- **dropout in longitudinal studies can have surprising consequences**
- **random effects and parameters are different things:**
  - **parameters don't change if you re-run an experiment**
  - **random effects do**