

# Avoiding Data Snooping in Multilevel and Mixed Effects Models

David Afshartous

School of Business  
University of Miami

Joint with Michael Wolf, University of Zurich

# Outline

- 1 Multilevel and Mixed Effects Models
- 2 Data Snooping
- 3 Avoiding Data Snooping
- 4 Applications
- 5 Conclusions

# Outline

- 1 **Multilevel and Mixed Effects Models**
- 2 Data Snooping
- 3 Avoiding Data Snooping
- 4 Applications
- 5 Conclusions

# Background and Notation

Multilevel and Mixed Effects models used in many disciplines.

- Many names
- Basic problem: Hierarchical or clustered data, e.g., students within schools
- Longitudinal data, e.g, growth curves
- Motivations: better estimates of uncertainty; borrowing strength; modeling variability between and within groups

# Background and Notation

Usual Regression framework:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Hierarchical structure and random coefficients:

- $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$ 
  - $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$
  - $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$
- The intercept and slope coefficients are random variables!

Combined model by substituting accordingly

# Inference on Random Effects

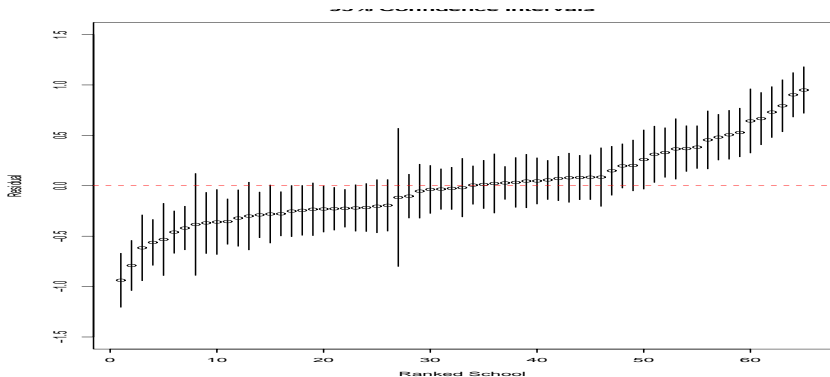
We focus on inference for the random effects

Substantive interest: represent the effect or departure of  $j$ th group from the grand mean

Diagnostic interest

Common to see rankings of random effects

# Inference on Random Effects



**Figure:** The ‘caterpillar plot’ of Rasbash et al. (2004, page 39): the level-2 residuals of the 65 schools in ascending order together with their respective 95% confidence intervals.

# Inference on Random Effects

Previous figure likely to be used for different purposes by different individuals:

- School head may want to know if his/her school differs from the average
- A governing board may be interested in finding out which are the schools that differ from average
- A parent may be comparing several schools under consideration

# Outline

- 1 Multilevel and Mixed Effects Models
- 2 Data Snooping**
- 3 Avoiding Data Snooping
- 4 Applications
- 5 Conclusions

## Basic Problem: Data Snooping

Whenever several hypotheses are tested at the same time, a *multiple testing* scenario arises:

- Naive approach: decisions based on individual p-values, overly liberal results.
- Example: 100 independent tests of true null hypotheses at  $\alpha = .05$ . The expected number of false rejects is 5; Probability of at least one false rejection is  $1 - 0.95^{100} = 0.994$ .
- Classical approach: control of *familywise error rate* (FWE) defined as the probability of making at least one false rejection
- E.g., Bonferroni (**single-step**) uses  $\alpha/S$ , where  $S$  = number of tests

# Basic Problem

Holm (1979) method (**stepwise**):

- Order  $p$ -values from smallest to largest
- Reject  $H_{(s)}$  if  $\hat{p}_{(j)} \leq \alpha / (S - j + 1)$  for all  $j = 1, \dots, s$
- More powerful than Bonferroni
- But can still be overly conservative

# Outline

- 1 Multilevel and Mixed Effects Models
- 2 Data Snooping
- 3 Avoiding Data Snooping**
- 4 Applications
- 5 Conclusions

# Unified Framework

Unknown probability distribution generating the data:  $P$   
Parameter vector  $\theta = \theta(P)$ :  $\theta = (\theta_1, \dots, \theta_S)'$ . Individual hypotheses about the elements of  $\theta$ :

$$H_S: \theta_S = 0 \quad \text{vs.} \quad H'_S: \theta_S \neq 0. \quad (1)$$

## Example (Absolute Comparisons)

The values of the level-2 residuals  $u_j$  are under test:  $S = J$  and  $\theta_S = u_S$ .

## Example (Pairwise Comparisons)

All pairwise comparisons of the level-2 residuals are of interest:  $S = \binom{J}{2}$ ; an element  $\theta_S$  is of the form  $\theta_S = u_j - u_k$ , where  $S$  can be taken as referring to the ordered pair  $(j, k)$ .

# Problem solution based on the familywise error rate

$$\text{FWE}_P = P\{\text{Reject at least one } H_s: \theta_s = 0\}.$$

Focus on asymptotic control of FWE:

$$\limsup_{\min_{1 \leq j \leq J} n_j \rightarrow \infty} \text{FWE}_P \leq \alpha \quad \text{for all } P.$$

i.e., control max FWE as smallest  $n_j$  increases.

Stepwise method of Holm (1979) improves power of Bonferonni

Romano & Wolf (2005) develop a novel stepwise multiple testing procedure that accounts for the dependence structure of the test statistics; more powerful than the Holm method.

Afshartous & Wolf (2007) extend procedure to the 2-sided case

# StepM Method

## Basic Setup:

- The test statistic for the null hypothesis  $H_s$  is of the form  $|z_s| = |w_s|/\hat{\sigma}_s$
- where  $w_s$  is a (consistent) estimator of the parameter  $\theta_s$  and  $\hat{\sigma}_s$  is a standard error of  $w_s$
- Sort the test statistics from largest to smallest
- Label  $r_1$  corresponds to the largest test statistic and label  $r_S$  to the smallest one, so that  $|z_{r_1}| \geq |z_{r_2}| \geq \dots \geq |z_{r_S}|$ .

# StepM Method

## Step 1:

- compute  $1 - \alpha$  (asymptotic) joint confidence region for the parameter vector  $(\theta_{r_1}, \dots, \theta_{r_S})'$

$$[w_{r_1} \pm \hat{\sigma}_{r_1} \hat{d}_1] \times \dots \times [w_{r_S} \pm \hat{\sigma}_{r_S} \hat{d}_1] \quad (2)$$

- $\hat{d}_1$  is chosen for asymptotic  $1 - \alpha$  joint coverage
- Reject  $H_{r_s}$  if zero not contained in  $[w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_1]$ .
- Denote by  $R_1$  the number of hypotheses rejected in this first step. If  $R_1 = 0$ , we stop.

# StepM Method

## Step 2:

- construct  $1 - \alpha$  (asymptotic) joint confidence region for the 'remaining' parameter vector

$$[w_{r_{R_1+1}} \pm \hat{\sigma}_{r_{R_1+1}} \hat{d}_2] \times \dots \times [w_{r_S} \pm \hat{\sigma}_{r_S} \hat{d}_2] \quad (3)$$

- Then, for  $s = R_1 + 1, \dots, S$ , the hypothesis  $H_{r_s}$  is rejected if zero is not contained in the interval  $[w_{r_s} \pm \hat{\sigma}_{r_s} \hat{d}_2]$ .
- Denote by  $R_2$  the number of hypotheses rejected in this second step. If  $R_2 = 0$ , we stop.
- and so on ...

## StepM Method: calculation of $d$

- Which ideal value  $d_1$  would result in a finite sample joint coverage of exactly  $1 - \alpha$ ?
- One can show that this ideal value is the  $1 - \alpha$  quantile of the sampling distribution under  $P$  of  $\max_{1 \leq s \leq S} |w_{r_s} - \theta_{r_s}| / \hat{\sigma}_{r_s}$ .
- The ideal constant, called  $d_2$ , in the second step is the  $1 - \alpha$  quantile of the sampling distribution under  $P$  of  $\max_{R_1+1 \leq s \leq S} |w_{r_s} - \theta_{r_s}| / \hat{\sigma}_{r_s}$ .
- Thus  $d_2 \leq d_1$  and it is possible that some more hypotheses will be rejected in the second step

## StepM Method: calculation of $d$

- **Problem:**  $d_1, d_2, \dots$  depend on the unknown probability distribution  $P$ .
- **Solution:** bootstrap approach to replace  $P$  by an estimator  $\hat{P}$ ; the quantiles are computed under  $\hat{P}$ .

## Solution based on false discovery proportion

### False discovery proportion

- $F = \#$  false rejections;  $R = \#$  total rejections

$$\text{FDP} = \frac{F}{R}$$

- Control  $P\{\text{FDP} > \gamma\}$  for  $\gamma \in [0, 1)$

As a stepping stone towards FDP control we first need to control the *generalized familywise error rate* ( $k$ -FWE):

$$k\text{-FWE}_P = P\{\text{Reject at least } k \text{ of the } H_s: \theta_s = 0\}.$$

Benjamini & Hochberg (1995) propose stepwise method for controlling the  $E(\text{FDP})$ , coined the *false discovery rate* (FDR).

# Solution based on false discovery proportion (FDP-StepM Method)

Illustrate the method using  $\gamma = 0.1$ :

- Start with 1-FWE control (at level  $\alpha$ )
- Suppose  $N_1$  hypotheses are rejected, and consider rejecting the next most significant hypothesis. If falsely rejected the FDP will be  $\frac{1}{N_1+1}$ .
- Thus, if less than 9 hypotheses are rejected, stop
- Otherwise, move on to 2-FWE control (at level  $\alpha$ )
- Suppose  $N_2$  hypotheses are rejected, and consider rejecting the next most significant hypothesis. If falsely rejected the FDP will be  $\frac{1+1}{N_2+1}$ .
- If less than 19 hypotheses are rejected, stop
- Otherwise, move on to 3-FWE control (at level  $\alpha$ )
- And so on . . .

# Solution based on false discovery proportion (FDP-StepM Method)

General stopping rule at any given step  $j$ :

- Stop if # rejections  $< j/\gamma - 1$

Comments:

- Works for any 'underlying'  $k$ -FWE controlling method
- But in the interest of power, use the  $k$ -StepM method

## Alternative ways to improve power

Consider controlling the FWE at level  $\alpha = 0.05$ . This corresponds to controlling the FDP at level  $\alpha = 0.05$  and choosing  $\gamma = 0$ .

Option 1:

Stick with FWE control, but increase  $\alpha$ , say to 0.1.

Option 2:

Switch to 'actual' FDP control with positive  $\gamma$ , say  $\gamma = 0.1$ .

Philosophically different choices: Is it better to be 90% confident that all rejections are true rejections or it better to be 95% confident that the realized FDP is at most 0.1?

# Outline

- 1 Multilevel and Mixed Effects Models
- 2 Data Snooping
- 3 Avoiding Data Snooping
- 4 Applications**
- 5 Conclusions

# General Setup

Compare the various multiple testing methods for three data sets

Random effects models were estimated via the `nlme` package of Pinheiro & Bates (2000) which is contained in the statistical software R.

R extensions were written for the standard errors of the random effects estimates, the covariances between random effects estimates, the bootstrapping of the data, as well as the StepM, k-StepM, and FDP-StepM methods.

Code available at

<http://moya.bus.miami.edu/~dafshartous/>.

Significance level  $\alpha = 0.05$  and the value  $\gamma = 0.1$  (for the FDP-StepM and the FDR methods).

# Data Snooping When Making Absolute Comparisons

## Rasbash et al. (2004) Local Education Authority Data

- predict exam score achieved by 16 year old students with their London Reading Test score obtained just before they entered secondary school at age of 11 years.
- Data is from an English Local Education Authority; 4,059 students in 65 schools.
- random effects model with random intercept and constant slope across schools.
- $S = 65$  absolute comparisons, examining whether the school's average exam score differs from the grand mean after accounting for LRT score.
- individual p-values: 28 null hypotheses are rejected
- The application of the StepM, FDP-StepM, and FDR methods yield 17, 27, and 27 rejections, respectively.

# Data Snooping When Making Absolute Comparisons

## National Educational Longitudinal Study of 1988 (NELS)

- base-year sample consists of 24,599 eighth grade students, distributed amongst 1,052 schools nationwide.
- response variable is student mathematics score and the predictor is the socio-economic status (SES) of the student.
- As above, we fit a multilevel or random effects model with random intercept and constant slope across schools.
- $S = 1,052$  absolute comparisons.
- individual p-values: 289 hypotheses are rejected
- The application of the StepM, FDP-StepM, and FDR methods yield 38, 249, and 244 rejections, respectively.

# Data Snooping When Making Pairwise Comparisons

Wafer data presented in Pinheiro & Bates (2000)

- The data was collected to study the variability in the manufacturing of analog MOS circuits
- 40 observations on 10 wafers each; response variable = intensity of current; predictor variable = voltage.
- Given 10 wafers:  $S = 45$  possible pairwise comparisons.
- individual p-values: 30 hypotheses are rejected.
- The application of the StepM, FDP-StepM, and FDR methods yield 26, 30, and 32 rejections, respectively.

Education data set of Rasbash et al. (2004).

- 65 schools:  $S = 2,080$  possible pairwise comparisons.
- individual p-values: 1,027 rejections.
- The application of the StepM, FDP-StepM, and FDR methods yield 348, 966, and 1,026 rejections, respectively.

## Alternative ways to improve power

LEA data, absolute comparisons, $S = 65$	
StepM with $\alpha = 0.05$	17
StepM with $\alpha = 0.1$	17
FDP-StepM with $\alpha = 0.05$ and $\gamma = 0.1$	27
NELS data, absolute comparisons, $S = 981$	
StepM with $\alpha = 0.05$	38
StepM with $\alpha = 0.1$	42
FDP-StepM $\alpha = 0.05$ and $\gamma = 0.1$	249
Wafer data, pairwise comparisons, $S = 45$	
StepM with $\alpha = 0.05$	26
StepM with $\alpha = 0.1$	27
FDP-StepM $\alpha = 0.05$ and $\gamma = 0.1$	30
LEA data, pairwise comparisons, $S = 2,080$	
StepM with $\alpha = 0.05$	348
StepM with $\alpha = 0.1$	411
FDP-StepM $\alpha = 0.05$ and $\gamma = 0.1$	966

# Outline

- 1 Multilevel and Mixed Effects Models
- 2 Data Snooping
- 3 Avoiding Data Snooping
- 4 Applications
- 5 Conclusions**

# Conclusions

## Methodology:

- Two novel multiple testing methods which account for data snooping: **StepM** and **FDP-StepM**
- Inference on the 'absolute' level-2 residuals to determine which are significantly different from zero
- Inference on pairwise comparisons of level-2 residuals.
- **Bootstrap** methods that account for the dependence structure improve upon existing methods based on the individual  $p$ -values
- **Stepwise** methods provide a 'free lunch' compared to their single-step counterparts

## $k$ -StepM Method: $k$ -FWE Control

Define:

$$k\text{-FWE}_P = P\{\text{Reject at least } k \text{ of the } H_s: \theta_s = 0\}.$$

Modify StepM to achieve  $k$ -FWE Control

- $d_1$ : The  $(1 - \alpha)$  quantile under  $P$  of  $k\text{-max}\{|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s}\}$
- $R_1 \geq k$  hypotheses rejected in the first step.  $d_2 = ?$
- Let  $K$  be an index set corresponding to  $k - 1$  of the rejected hypotheses and all remaining hypotheses
- $d_K$  is the  $(1 - \alpha)$  quantile of  $k\text{-max}_{s \in K}\{|w_{r_s} - \theta_{r_s}|/\hat{\sigma}_{r_s}\}$
- Then  $d_2 = \max\{d_K\}$  [there are  $\binom{R_1}{k-1}$  such  $d_K$ ]
- In case  $\binom{R_1}{k-1}$  is too large, the method can be made operative by maximizing over a feasible subset
- Continue with such steps until no further rejections occur

## References

- Goldstein, H. and Healy, M.J. (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society: Series A*, 158:175-177.
- Goldstein, H. and Spiegelhalter, D.J. (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A*, 159: 385-443.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6: 65-70.
- Rasbash, J., Steele, F., Browne, W., and Prosser, B. (2004) A User's guide to MLwiN Version 2.0. *Institute of Education, London*, available at <http://multilevel.ioe.ac.uk/download/userman20.pdf>.
- Romano, J.P. and Wolf, M. (2005) Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4): 1237:1282.