Psych 253
Advanced Statistical Modeling

Clustered errors and mixed-effect models

Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford University

Russ Poldrack
Department of Psychology
Stanford University

Assumptions of OLS estimation

$$Y = Xb + \epsilon \qquad \hat{b} = (X'X)^{-1}X'Y$$

For our estimates to be the Best Linear Unbiased Estimate (BLUE), we must assume that:

- observations are linear in parameters **b** (not necessarily in regressors **X**)
- Errors have a zero mean and are independent of regressors (i.e. the expected error for any value of X is zero)
- Errors are uncorrelated and have constant variance (i.e. y's are independent conditional on model)
- Regressors are not perfectly collinear
- Large outliers are uncommon

Hypothesis testing and standard errors require further assumption that errors are Gaussian

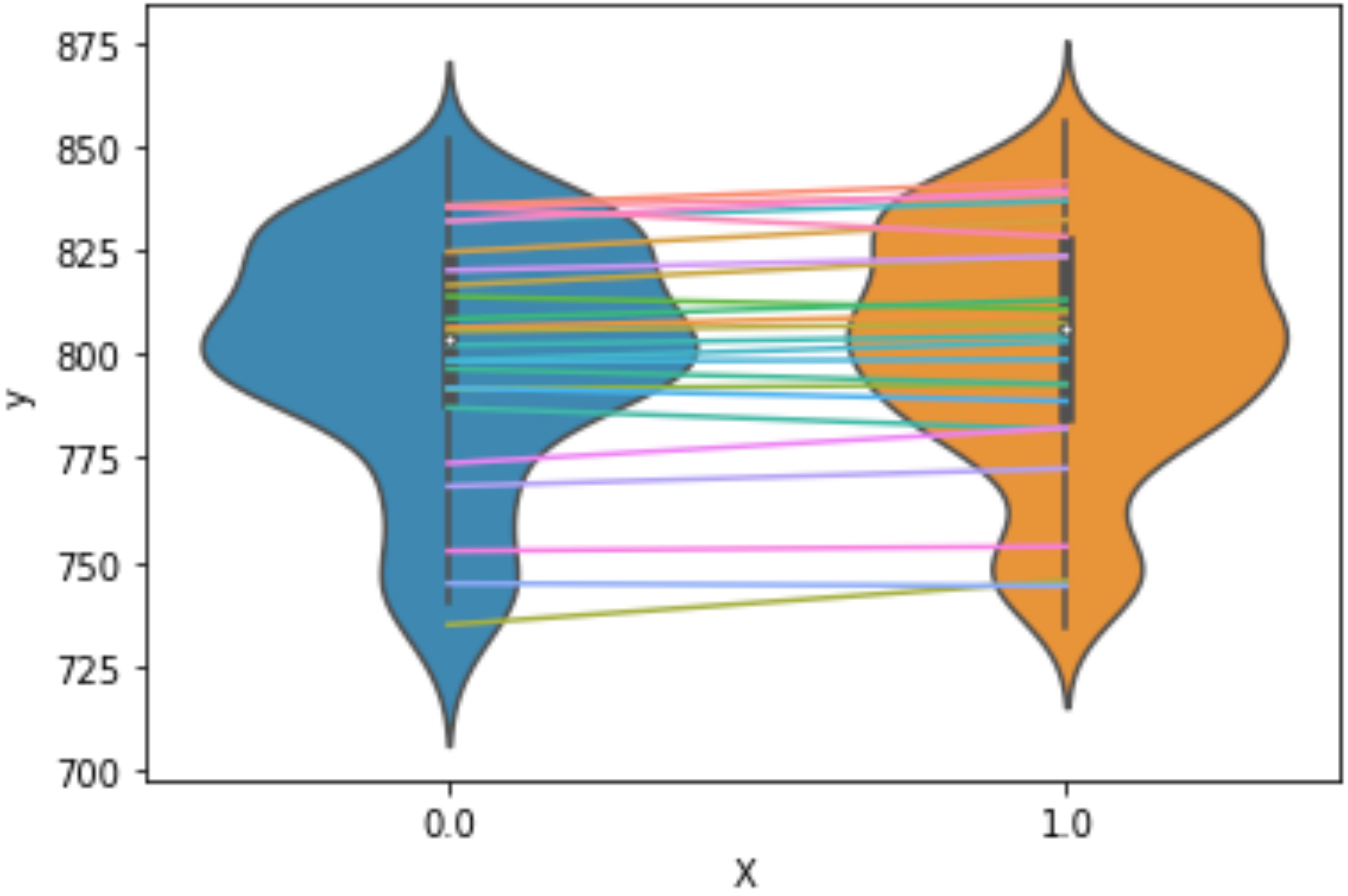
Clustered errors

It is common for errors to be correlated or clustered due to structure in the data

$\hat{\beta}$ remains unbiased under OLS but standard errors are biased and hypothesis tests can be incorrect

An example

24 subjects complete a cognitive test in which they perform 10 trials in each of two experimental conditions



Generative model

$$Y = X\beta + Z\mathbf{u} + \epsilon$$

y: observed data

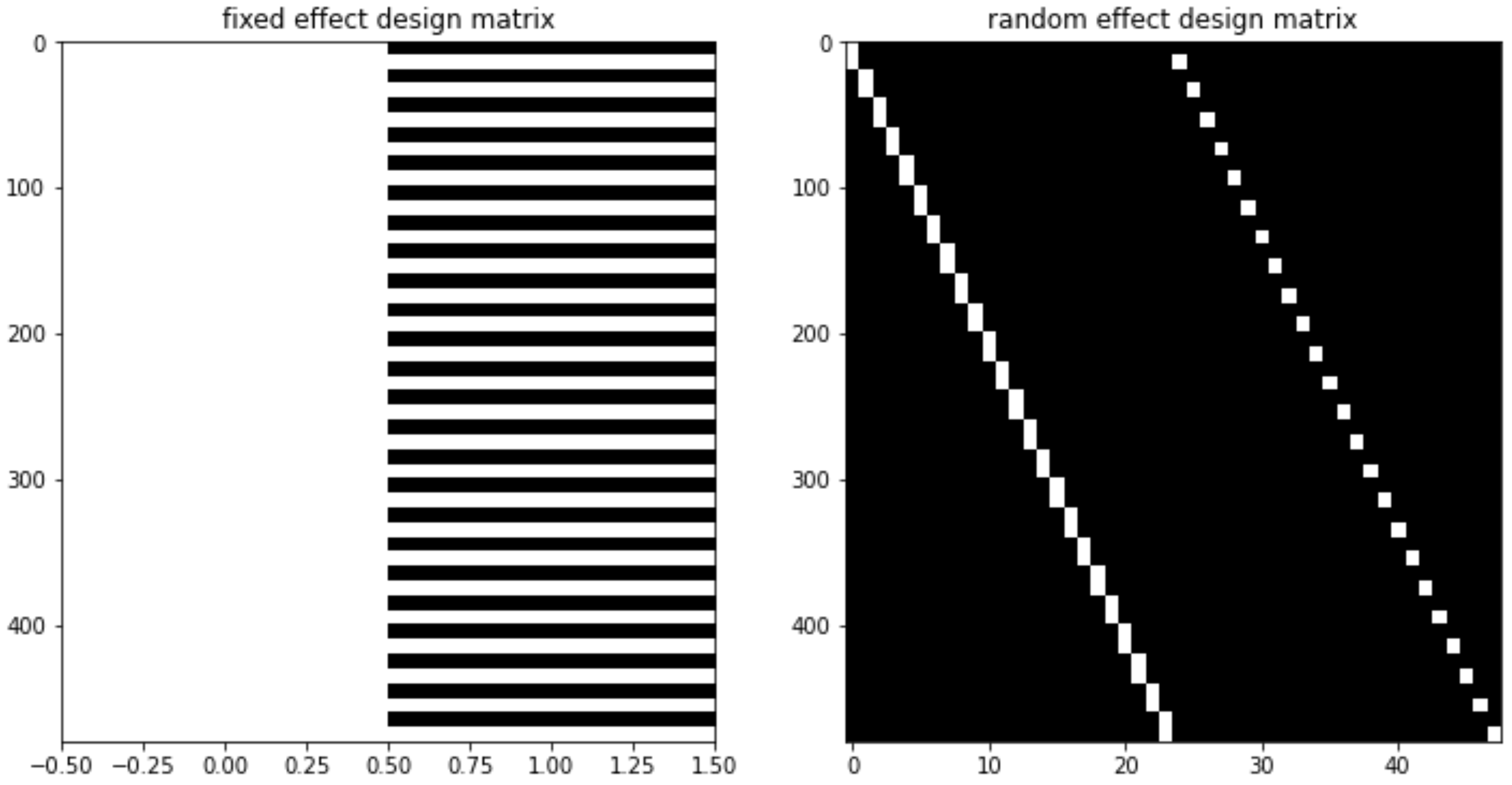
X: design matrix for fixed effects

β: vector of fixed effect parameters [800, 3]

Z: design matrix for random effects

u: vector of random effect parameters $u \sim N(0, \sigma^2 \Sigma)$

ϵ: errors. $\epsilon \sim N(0, \sigma^2 I), u \perp \epsilon$

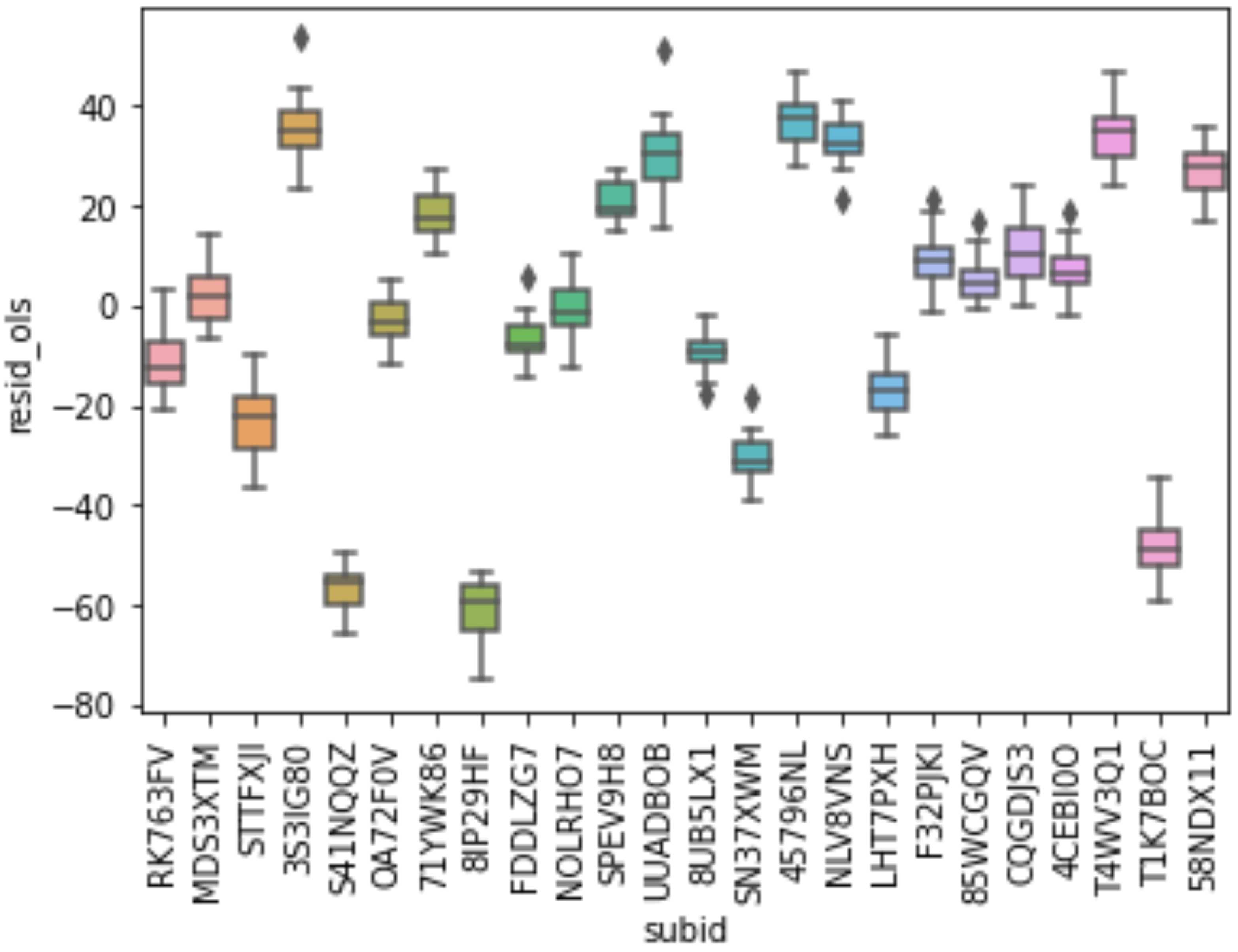


Results from OLS ignoring clustering

```
ols = smf.ols('y ~ X', data_df)
ols_result = ols.fit()
print(ols_result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.8897
Date:                Thu, 15 Apr 2021   Prob (F-statistic):              0.346
Time:                        20:47:48   Log-Likelihood:                -2291.5
No. Observations:                 480   AIC:                             4587.
Df Residuals:                     478   BIC:                             4595.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    800.2692      1.853    431.909      0.000     796.628     803.910
X              2.4716      2.620      0.943      0.346      -2.677       7.620
==============================================================================
Omnibus:                       29.086   Durbin-Watson:                   0.180
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.936
Skew:                          -0.609   Prob(JB):                     1.16e-07
Kurtosis:                       2.667   Cond. No.                         2.62
==============================================================================
```

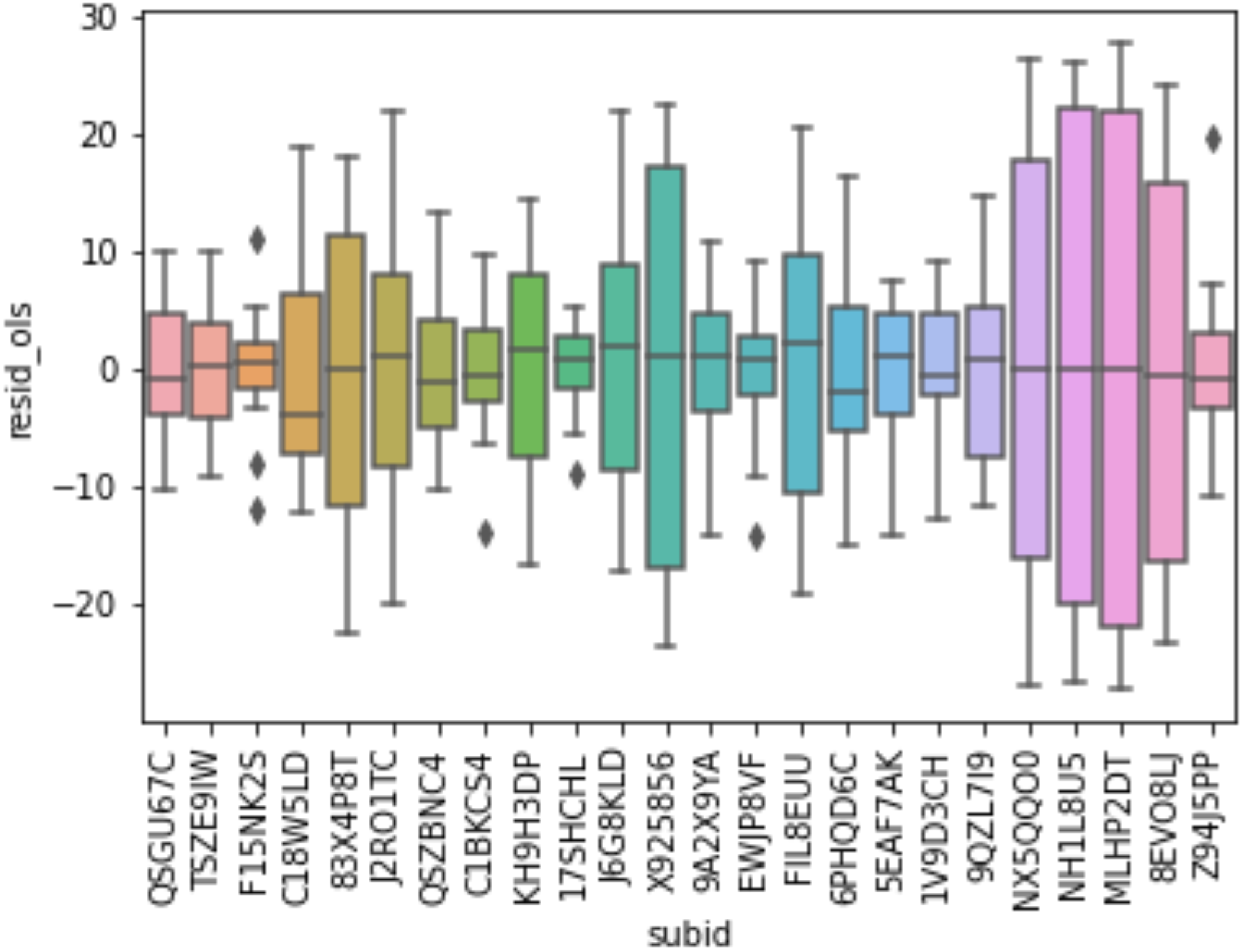
Residual errors are clustered by subject



This clearly violates our assumption of uncorrelated errors!

Heteroskedasticity

Differences in variability between observations



Ways to address clustered errors/heteroskedasticity

- Use a robust estimator of the standard errors
  - Sandwich estimator
- Use a modeling approach that can account for it
  - Mixed effects models
  - Fixed effects models
  - Generalized estimating equations (GEE)

Cluster-robust standard errors

$$y = Xb + \epsilon \qquad E(\epsilon) = 0 \qquad E(\epsilon\epsilon') = \Phi$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1}$$

Called the "sandwich estimator" because the variance of Y is "sandwiched" between the inverses

Under homoskedasticity:

$$\Phi = \sigma^2\mathbf{I}$$

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Cluster-robust standard errors

$$y = Xb + \epsilon \qquad E(\epsilon) = 0 \qquad E(\epsilon\epsilon') = \Phi$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1}$$

When errors are independent but not equal:

$$\Phi = \hat{\Sigma}$$

$$\text{where } \hat{\Sigma}_i = \sigma_i^2$$

(i.e. diagonal with unequal elements)

σ_1^2		0	0	0	0
0	σ_2^2	0	0	0	0
0	0	σ_3^2	0	0	0
0	0	0	σ_4^2	0	0
0	0	0	0	σ_5^2	0
0	0	0	0	0	σ_6^2

Cluster-robust standard errors

$$y = Xb + \epsilon \qquad E(\epsilon) = 0 \qquad E(\epsilon\epsilon') = \Phi$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1}$$

When errors are clustered:

$$\Phi = \hat{\Sigma}$$

$$\text{where } \hat{\Sigma}_j = \hat{\epsilon_j}\hat{\epsilon_j}'$$

(i.e. block diagonal with j clusters)

$\hat{\Sigma}_1$	$\hat{\Sigma}_1$	0	0	0	0
$\hat{\Sigma}_1$	$\hat{\Sigma}_1$	0	0	0	0
0	0	$\hat{\Sigma}_2$	$\hat{\Sigma}_2$	0	0
0	0	$\hat{\Sigma}_2$	$\hat{\Sigma}_2$	0	0
0	0	0	0	$\hat{\Sigma}_3$	$\hat{\Sigma}_3$
0	0	0	0	$\hat{\Sigma}_3$	$\hat{\Sigma}_3$

Cluster-robust standard errors

```
     OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                     7.035
Date:                Thu, 15 Apr 2021   Prob (F-statistic):             0.0142
Time:                        20:47:49   Log-Likelihood:                -2291.5
No. Observations:                 480   AIC:                             4587.
Df Residuals:                     478   BIC:                             4595.
Df Model:                           1
Covariance Type:              cluster
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    800.2692      5.873    136.263      0.000     788.120     812.418
X              2.4716      0.932      2.652      0.014       0.544       4.399
==============================================================================
Omnibus:                       29.086   Durbin-Watson:                   0.180
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.936
Skew:                          -0.609   Prob(JB):                     1.16e-07
Kurtosis:                       2.667   Cond. No.                         2.62
==============================================================================

Warnings:
[1] Standard Errors are robust to cluster correlation (cluster)
```

The mixed effects model

$$Y = X\beta + Z\mathbf{u} + \epsilon$$

y: observed data
X: design matrix for fixed effects (known)
β: vector of fixed effect parameters (unknown)
Z: design matrix for random effects (known)
u: unknown vector of random effect parameters

Goal: Model the structure of the data
so that errors become IID

Note: the values of u are not directly estimated, only
their variance!

Mixed model formula syntax

$$\text{formula} = \text{''}y \sim 1 + x + (1|\text{groupvar})\text{''}$$

Mixed model formula syntax

$$\text{formula} = {}^{\prime\prime}\underset{\substack{\downarrow \\ \text{outcome}}}{\boxed{y}} \sim 1 + x + (1|\text{groupvar})^{\prime\prime}$$

Mixed model formula syntax

$$\text{formula} = \text{''}y \sim \overbrace{\underset{\substack{\downarrow \\ \text{intercept}}}{\boxed{1}} + x}^{\text{fixed effects}} + (1|\text{groupvar})\text{''}$$

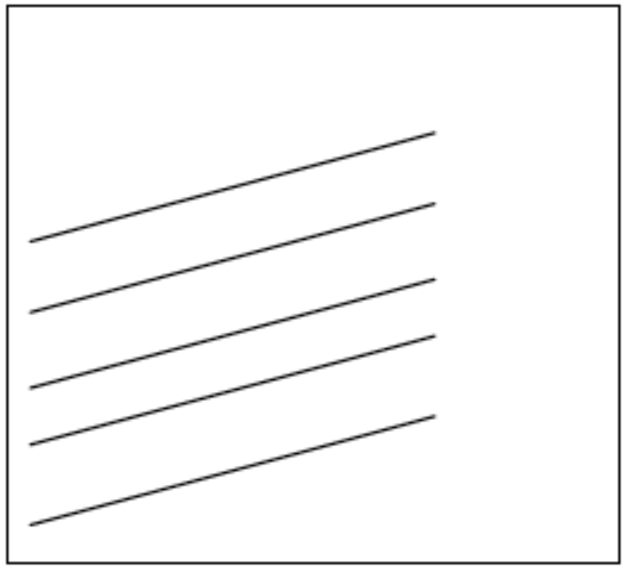
Mixed model formula syntax

$$\text{formula} = ”y \sim \overbrace{1 + \underset{\substack{\downarrow \\ \text{predictor}}}{\boxed{x}}}^{\text{fixed effects}} + (1|\text{groupvar})”$$

Mixed model formula syntax

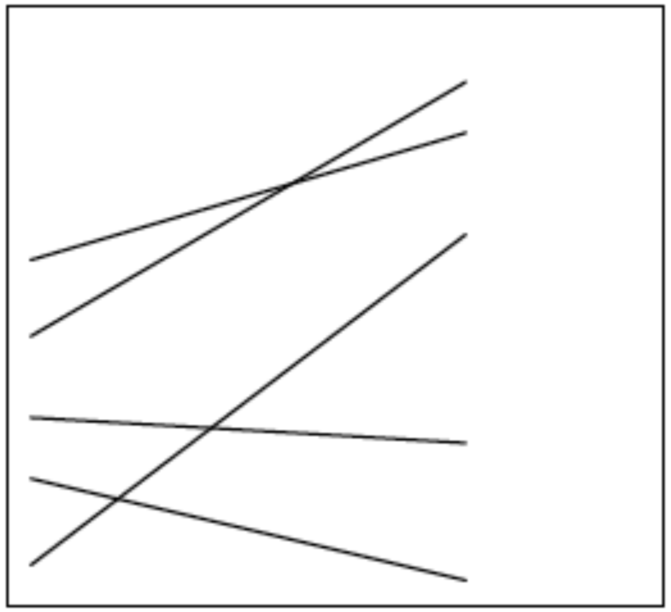
$$\text{formula} = {}''y \sim 1 + x + \overbrace{(\boxed{1|}\text{groupvar})}^{\text{random effects}}''$$

random intercept



Mixed model formula syntax

$$\text{formula} = \text{''}y \sim 1 + x + \overbrace{(\underset{\text{random slope and intercept}}{\boxed{1 + x}}|\text{groupvar})}^{\text{random effects}}\text{''}$$



Random intercept model

```
md = smf.mixedlm("y ~ X", data_df, groups=data_df["subid"])
mdf = md.fit(method=["lbfgs"])
print(mdf.summary())
```

```
   Mixed Linear Model Regression Results
========================================================
Model:            MixedLM Dependent Variable: y
No. Observations: 480     Method:             REML
No. Groups:       24      Scale:              31.2306
Min. group size:  20      Likelihood:         -1578.3310
Max. group size:  20      Converged:          Yes
Mean group size:  20.0
--------------------------------------------------------
           Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
--------------------------------------------------------
Intercept 800.269    5.869 136.353 0.000 788.766 811.772
X           2.472    0.510   4.845 0.000   1.472   3.471
Group Var 823.587   44.619
========================================================
```

Random intercept/slope model

```
md = smf.mixedlm("y ~ X", data_df, groups=data_df["subid"], re_formula='~X')
mdf = md.fit(method=["lbfgs"])
print(mdf.summary())
```

```
 Mixed Linear Model Regression Results
============================================================
Model:              MixedLM  Dependent Variable:  y
No. Observations:   480      Method:              REML
No. Groups:         24       Scale:               27.3563
Min. group size:    20       Likelihood:          -1563.5472
Max. group size:    20       Converged:           Yes
Mean group size:    20.0
------------------------------------------------------------
               Coef.  Std.Err.    z    P>|z|  [0.025  0.975]
------------------------------------------------------------
Intercept     800.269    5.868 136.387 0.000 788.769 811.770
X               2.472    0.931   2.655 0.008   0.647   4.296
Group Var     823.559   47.823
Group x X Cov  -3.239    5.233
X Var          15.327    1.204
============================================================
```

Interpretation of mixed vs. marginal models

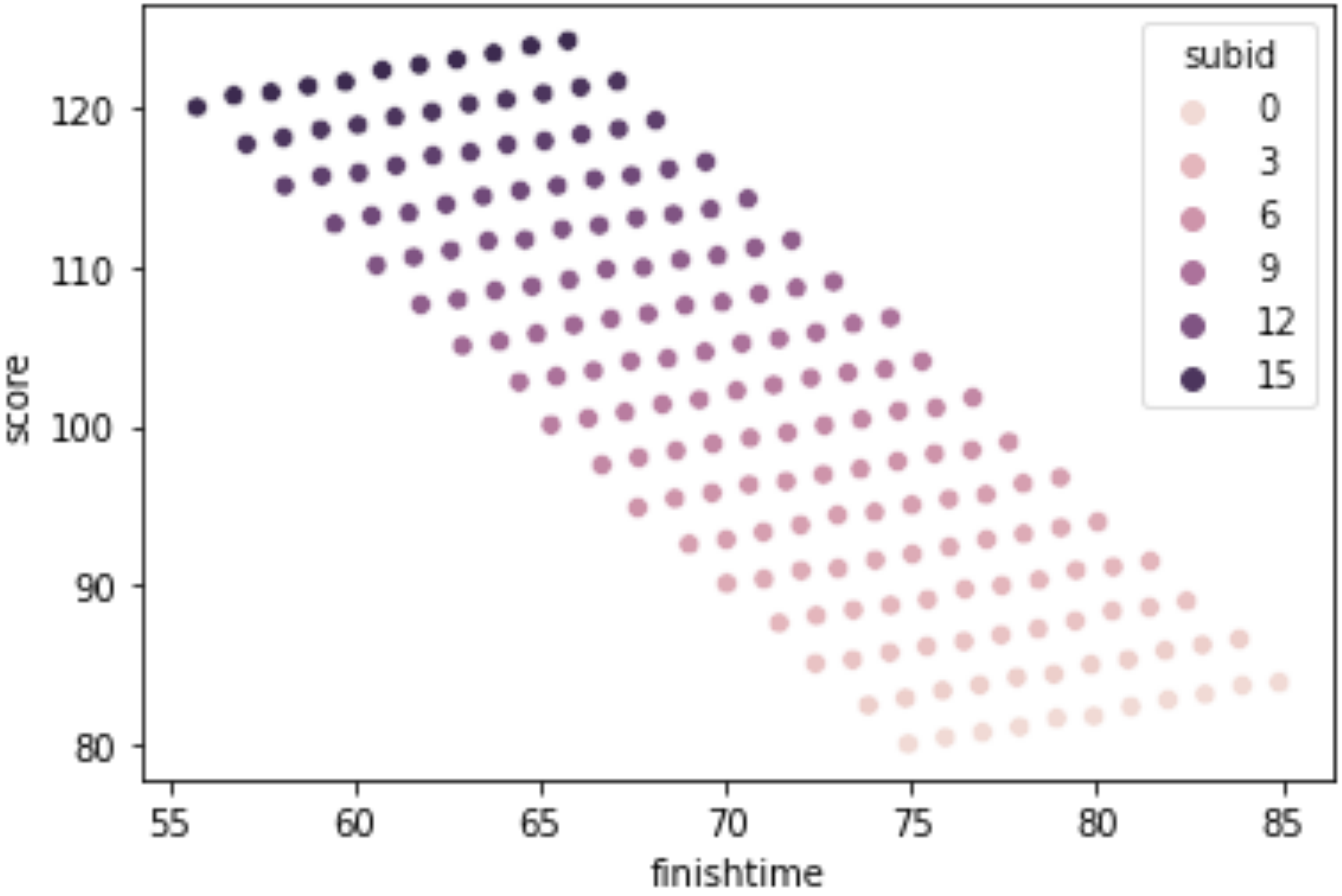
Standard (“marginal”) model interpretation:
What is effect of factor X on average across all observations?

Mixed model interpretation:
What is the average effect of factor X across all combinations of grouping variables?

An example: Simpson's paradox

Test score vs finishing time: The best students usually finish fastest

But for any individual, more test time will lead to a better score



```
OLS Regression Results
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    210.6811      5.463     38.563      0.000     199.903     221.460
finishtime    -1.5422      0.077    -19.955      0.000      -1.695      -1.390
```

```
     Mixed Linear Model Regression Results
=======================================================
           Coef.  Std.Err.    z    P>|z| [0.025 0.975]
-------------------------------------------------------
Intercept   73.823    3.664  20.150 0.000 66.643 81.004
finishtime   0.402    0.002 187.797 0.000  0.398  0.407
Group Var  227.801  901.263
=======================================================
```

The two analyses give opposite results!
Neither is “wrong” but they are relevant to different questions

The “stimulus as fixed effect” fallacy (Clark, 1973)

Our statistical model determines the scope of possible generalization from our results:

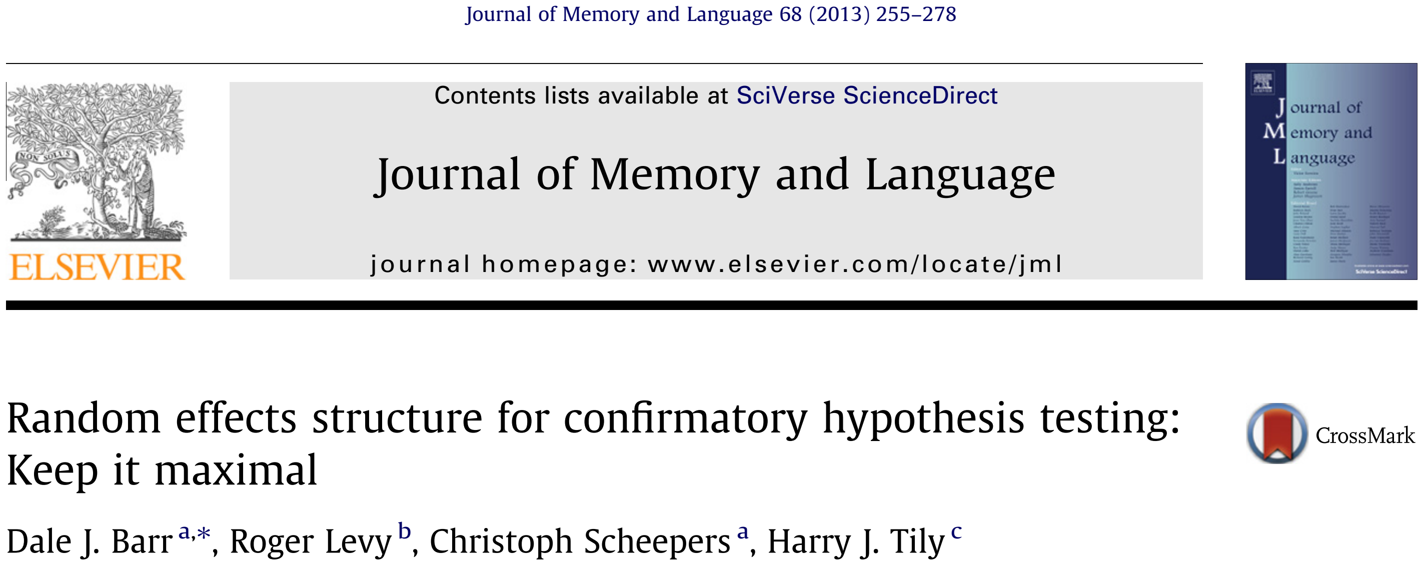
Only when we treat something as a random effect are we licensed to generalize beyond the specifics of our study

Clark, 1973:

- most researchers treat stimuli as a fixed effect
- Treating them as a random effect allow generalization beyond the specific stimulus set used in the study
- But it will generally increase the uncertainty of our estimates (and thus increase p-values)

$$\text{formula} = "y \sim 1 + condition + (1 + condition|\text{subject}) + (1|\text{item})"$$

What effects should I include?



Journal of Memory and Language 68 (2013) 255–278

Contents lists available at SciVerse ScienceDirect

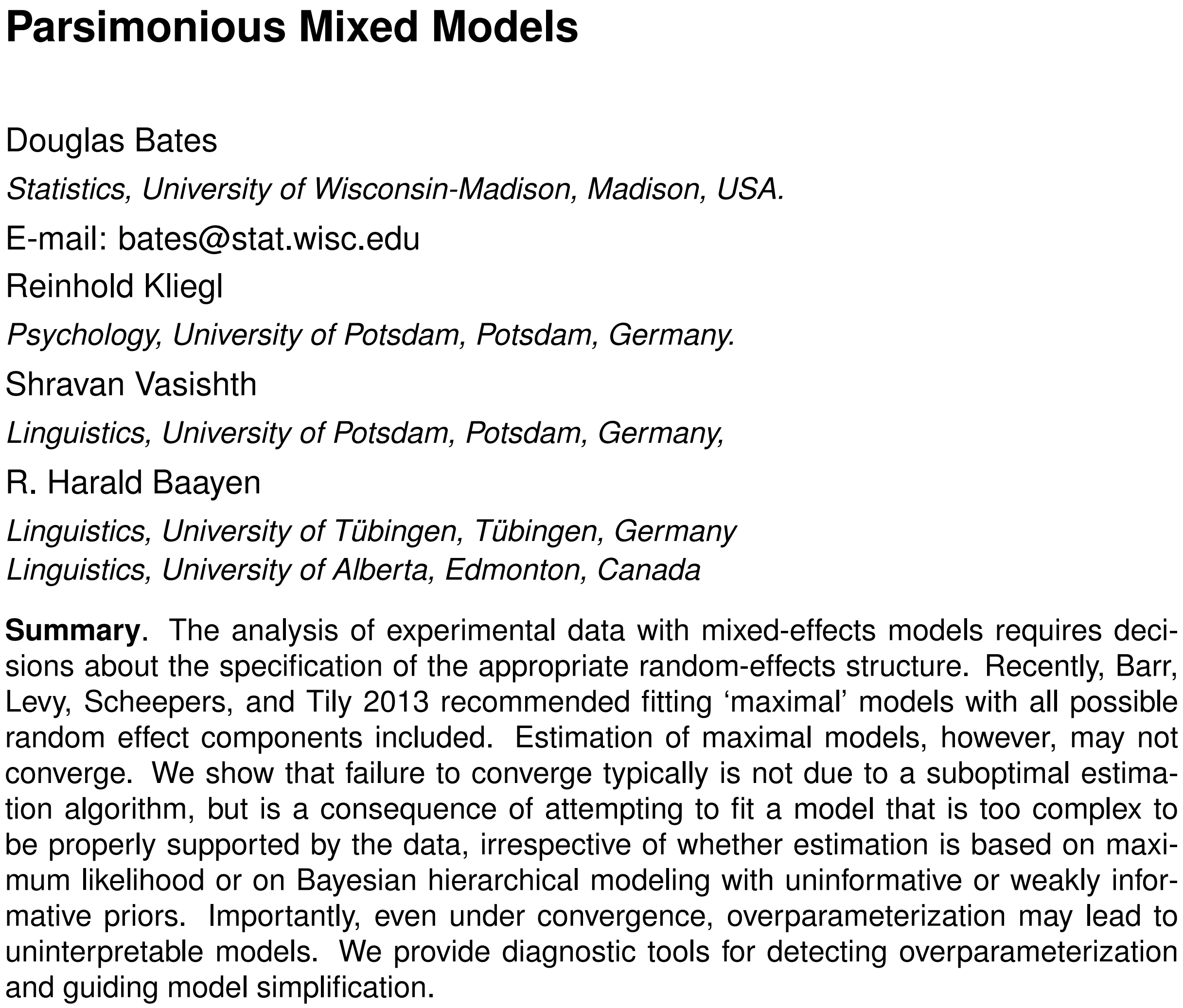
Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Random effects structure for confirmatory hypothesis testing: Keep it maximal

Dale J. Barr[a,*], Roger Levy[b], Christoph Scheepers[a], Harry J. Tily[c]

LMEMs generalize best when they include the maximal random effects structure justified by the design



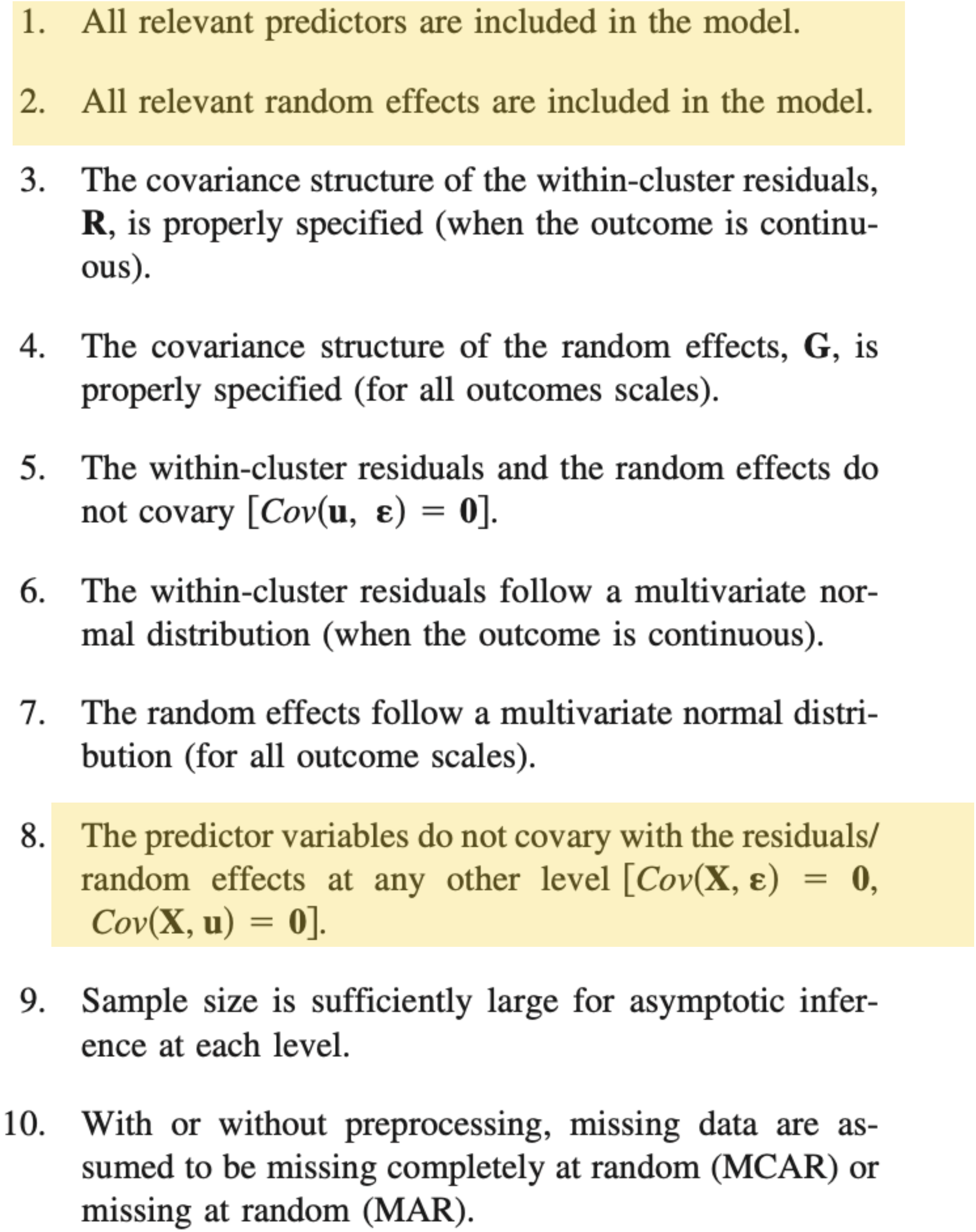
Parsimonious Mixed Models

Douglas Bates
Statistics, University of Wisconsin-Madison, Madison, USA.
E-mail: bates@stat.wisc.edu
Reinhold Kliegl
Psychology, University of Potsdam, Potsdam, Germany.
Shravan Vasishth
Linguistics, University of Potsdam, Potsdam, Germany,
R. Harald Baayen
Linguistics, University of Tübingen, Tübingen, Germany
Linguistics, University of Alberta, Edmonton, Canada

Summary. The analysis of experimental data with mixed-effects models requires decisions about the specification of the appropriate random-effects structure. Recently, Barr, Levy, Scheepers, and Tily 2013 recommended fitting 'maximal' models with all possible random effect components included. Estimation of maximal models, however, may not converge. We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling with uninformative or weakly informative priors. Importantly, even under convergence, overparameterization may lead to uninterpretable models. We provide diagnostic tools for detecting overparameterization and guiding model simplification.

https://arxiv.org/pdf/1506.04967.pdf

Assumptions of hierarchical linear modeling

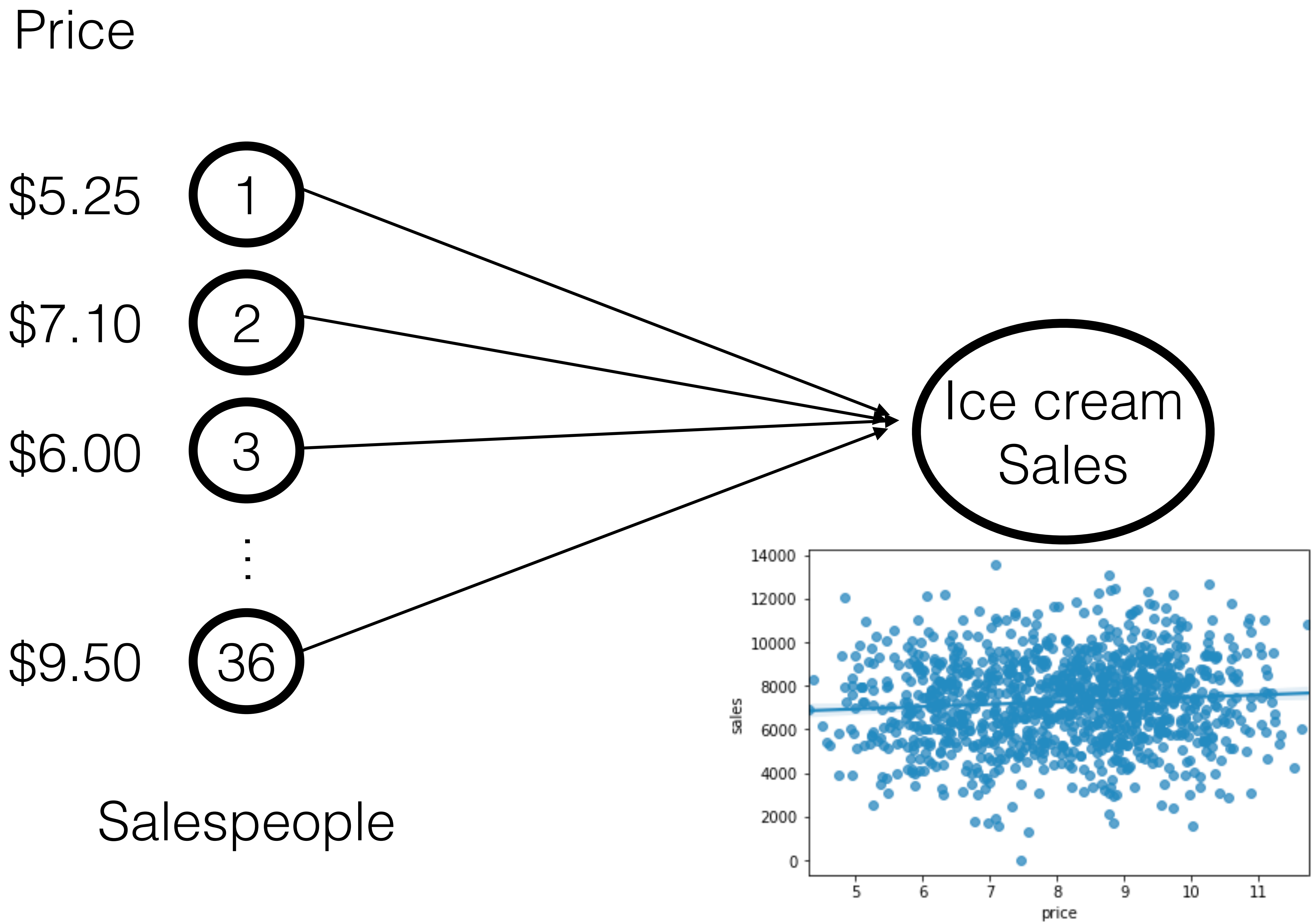


1. All relevant predictors are included in the model.
2. All relevant random effects are included in the model.
3. The covariance structure of the within-cluster residuals, **R**, is properly specified (when the outcome is continuous).
4. The covariance structure of the random effects, **G**, is properly specified (for all outcomes scales).
5. The within-cluster residuals and the random effects do not covary [$Cov(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$].
6. The within-cluster residuals follow a multivariate normal distribution (when the outcome is continuous).
7. The random effects follow a multivariate normal distribution (for all outcome scales).
8. The predictor variables do not covary with the residuals/random effects at any other level [$Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $Cov(\mathbf{X}, \mathbf{u}) = \mathbf{0}$].
9. Sample size is sufficiently large for asymptotic inference at each level.
10. With or without preprocessing, missing data are assumed to be missing completely at random (MCAR) or missing at random (MAR).

“Exogeneity”

McNeish et al., 2017

When exogeneity fails: Omitted variable bias



Price

$5.25

1

$7.10

2

$6.00

3

⋮

$9.50

36

Ice cream Sales

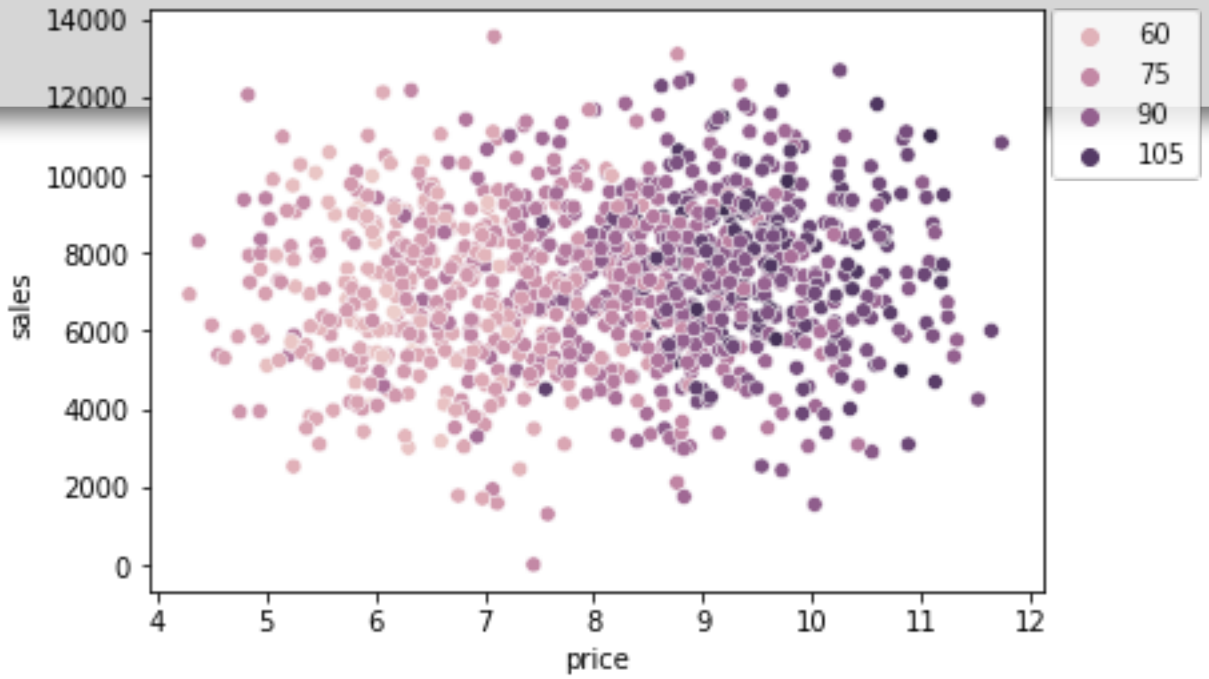
Salespeople

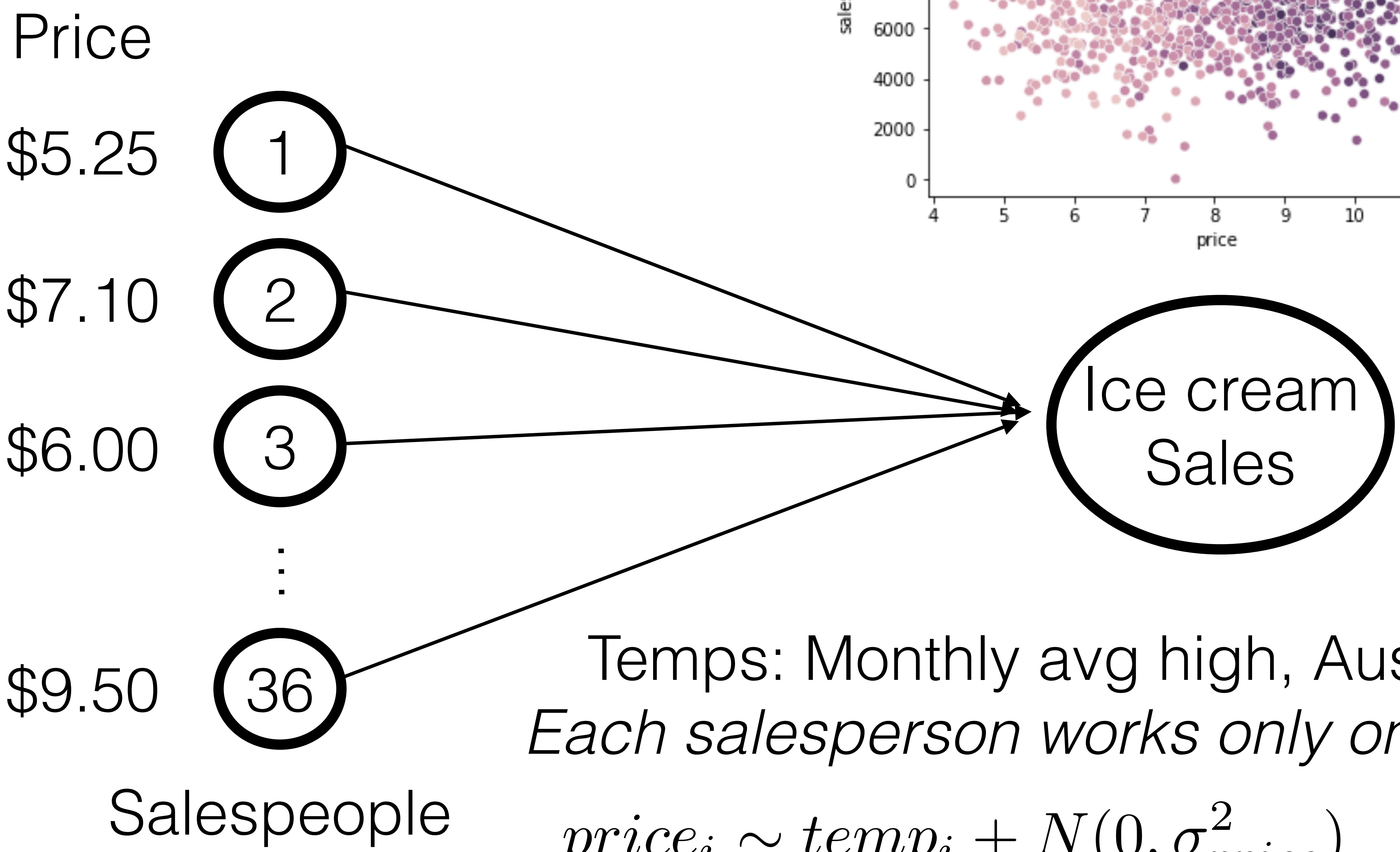
r=.08, p=.006

```
smf.mixedlm('sales ~  price_scaled', ic_df,
        groups=ic_df["salesperson"])
```

```
   Mixed Linear Model Regression Results
=============================================================
Model:                MixedLM  Dependent Variable:  sales
No. Observations:     1080     Method:              REML
No. Groups:           36       Scale:               3929321.8695
Min. group size:      30       Likelihood:          -9730.0794
Max. group size:      30       Converged:           Yes
Mean group size:      30.0
-------------------------------------------------------------
               Coef.   Std.Err.   z    P>|z|  [0.025   0.975]
-------------------------------------------------------------
Intercept     7262.475   79.253 91.637 0.000 7107.142 7417.807
price_scaled   148.609   77.358  1.921 0.055   -3.010  300.228
Group Var    95138.920   28.240
=============================================================
```

Why exogeneity matters





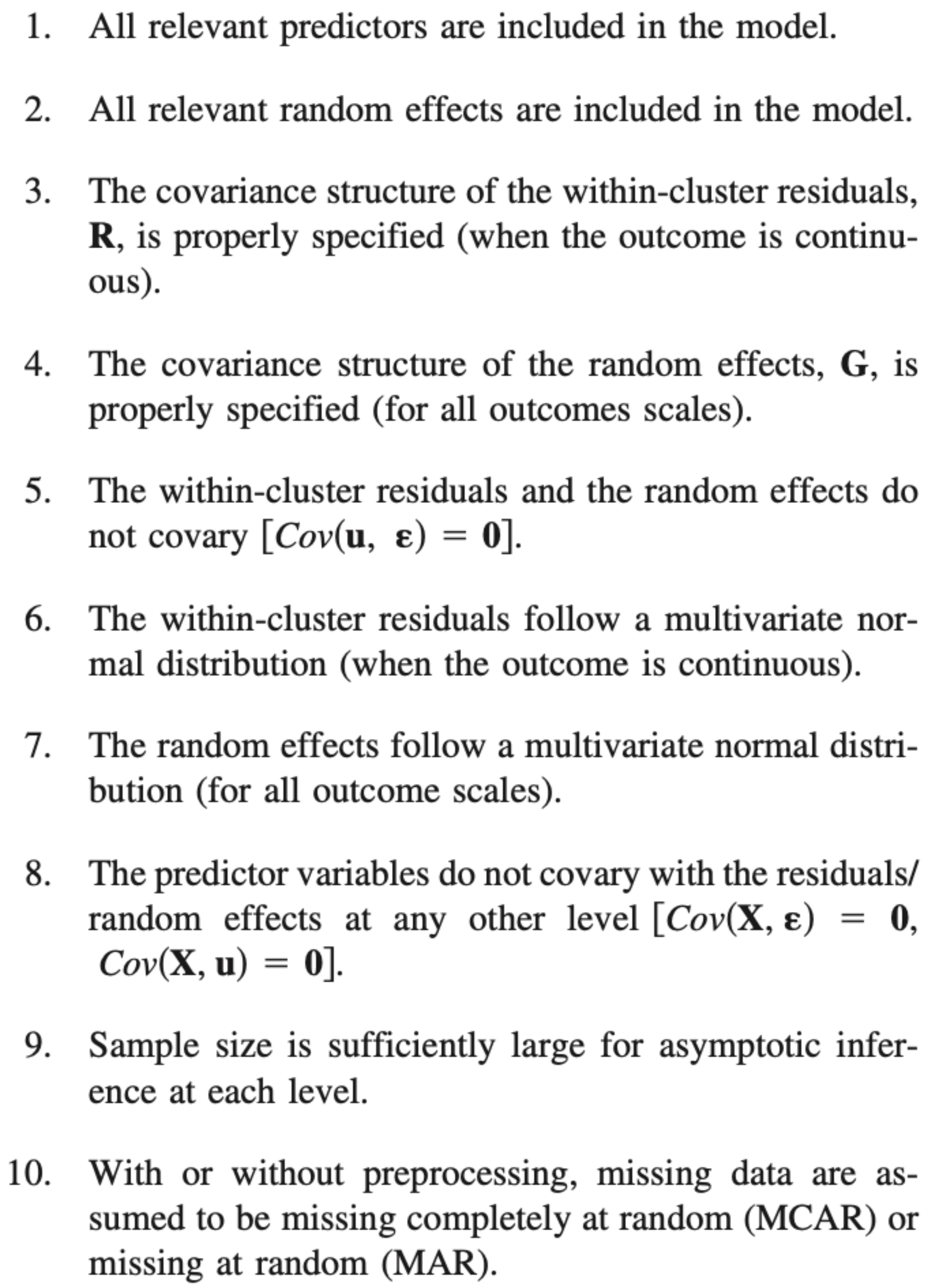
Temps: Monthly avg high, Austin TX

Each salesperson works only one month

$$price_i \sim temp_i + N(0, \sigma^2_{price})$$

$$\text{sales} \sim \beta_{temp} * temp_i + \beta_{price} * price_i + N(0, \sigma^2_{sales})$$

$$\beta_{temp} = 500, \beta_{price} = -200$$



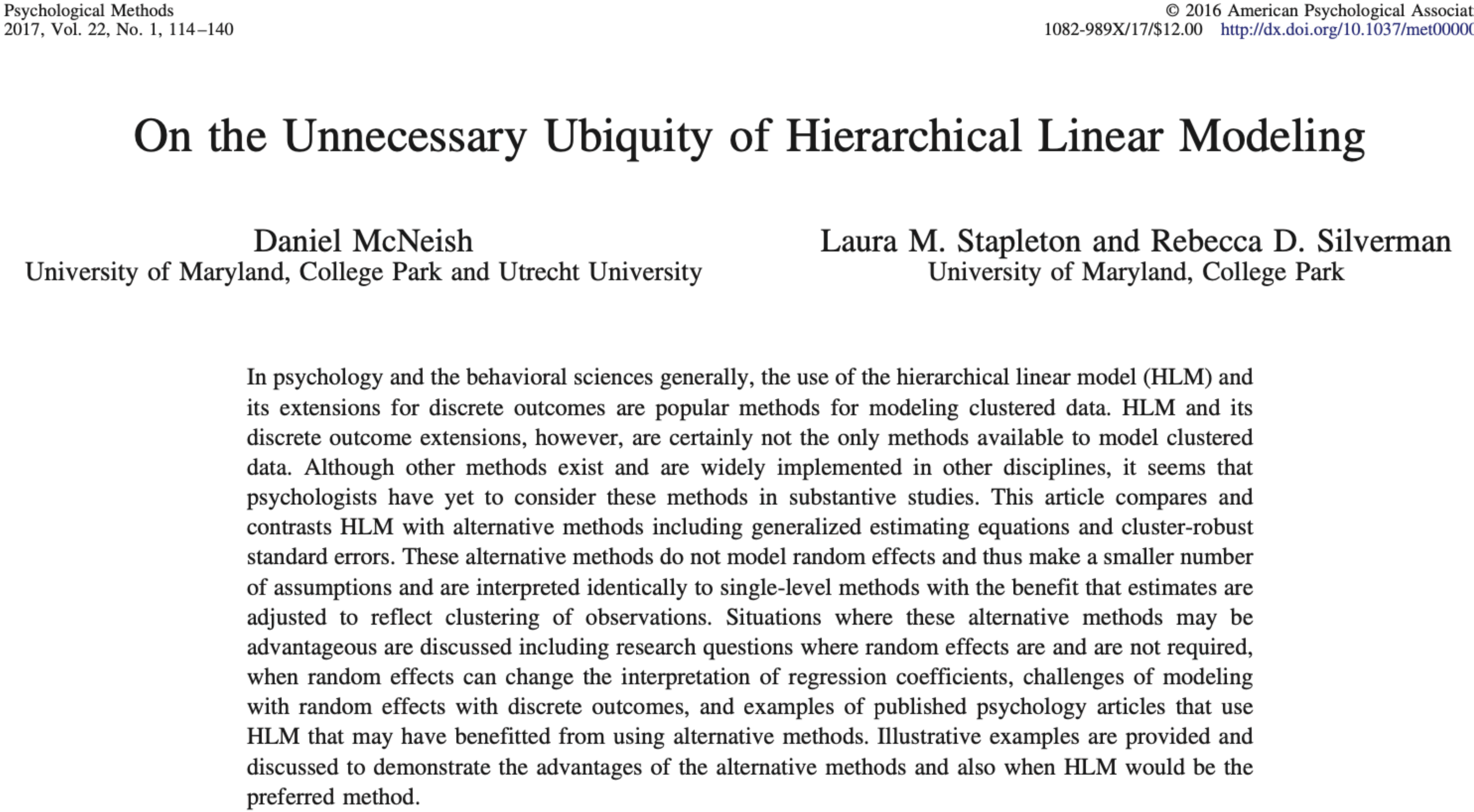
1. All relevant predictors are included in the model.
2. All relevant random effects are included in the model.
3. The covariance structure of the within-cluster residuals, **R**, is properly specified (when the outcome is continuous).
4. The covariance structure of the random effects, **G**, is properly specified (for all outcomes scales).
5. The within-cluster residuals and the random effects do not covary $[Cov(\mathbf{u},\ \boldsymbol{\varepsilon}) = \mathbf{0}]$.
6. The within-cluster residuals follow a multivariate normal distribution (when the outcome is continuous).
7. The random effects follow a multivariate normal distribution (for all outcome scales).
8. The predictor variables do not covary with the residuals/random effects at any other level $[Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0},\ Cov(\mathbf{X}, \mathbf{u}) = \mathbf{0}]$.
9. Sample size is sufficiently large for asymptotic inference at each level.
10. With or without preprocessing, missing data are assumed to be missing completely at random (MCAR) or missing at random (MAR).

```
r(u, e) 0.67
```

```
r(u, price): -0.028
r(u, temp [omitted]): 0.076
```

```
     Mixed Linear Model Regression Results
===============================================================
Model:               MixedLM  Dependent Variable:  sales
No. Observations:    1080     Method:              REML
No. Groups:          36       Scale:               3919345.8437
Min. group size:     30       Likelihood:          -9715.2805
Max. group size:     30       Converged:           Yes
Mean group size:     30.0
---------------------------------------------------------------
               Coef.   Std.Err.    z    P>|z|  [0.025   0.975]
---------------------------------------------------------------
Intercept     7262.475   63.794 113.842 0.000 7137.440 7387.509
price_scaled  -228.107  107.025  -2.131 0.033 -437.873  -18.341
temp_scaled    491.591  107.361   4.579 0.000  281.166  702.016
Group Var    15864.385   18.772
===============================================================
```

Mixed models are not the only game in town



Psychological Methods
2017, Vol. 22, No. 1, 114–140

© 2016 American Psychological Associat
1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000

On the Unnecessary Ubiquity of Hierarchical Linear Modeling

Daniel McNeish
University of Maryland, College Park and Utrecht University

Laura M. Stapleton and Rebecca D. Silverman
University of Maryland, College Park

In psychology and the behavioral sciences generally, the use of the hierarchical linear model (HLM) and its extensions for discrete outcomes are popular methods for modeling clustered data. HLM and its discrete outcome extensions, however, are certainly not the only methods available to model clustered data. Although other methods exist and are widely implemented in other disciplines, it seems that psychologists have yet to consider these methods in substantive studies. This article compares and contrasts HLM with alternative methods including generalized estimating equations and cluster-robust standard errors. These alternative methods do not model random effects and thus make a smaller number of assumptions and are interpreted identically to single-level methods with the benefit that estimates are adjusted to reflect clustering of observations. Situations where these alternative methods may be advantageous are discussed including research questions where random effects are and are not required, when random effects can change the interpretation of regression coefficients, challenges of modeling with random effects with discrete outcomes, and examples of published psychology articles that use HLM that may have benefitted from using alternative methods. Illustrative examples are provided and discussed to demonstrate the advantages of the alternative methods and also when HLM would be the preferred method.

You should decide on your question and then find the right method to ask it, rather than letting the methodological tail wag the conceptual dog