Better Diagnostics for Linear Mixed-Effects Models Using Visual Inference

Adam Loy
Department of Mathematics
Lawrence University

ISU Graphics Working Group *December 10, 2014*

Overview of my notation

Continuous response LME model for group $i = 1, \ldots, g$ is given by

$$\underset{(n_i \times 1)}{\boldsymbol{y}_i} = \underset{(n_i \times p)}{\boldsymbol{X}_i} \underset{(p \times 1)}{\boldsymbol{\beta}} + \underset{(n_i \times q)}{\boldsymbol{Z}_i} \underset{(q \times 1)}{\boldsymbol{b}_i} + \underset{(n_i \times 1)}{\boldsymbol{\varepsilon}_i}$$

For simplicity assume

- $\boldsymbol{b}_i$ are a random sample from $\mathcal{N}(\mathbf{0}, \boldsymbol{D})$
- $\boldsymbol{\varepsilon}_i$ are a random sample from $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_{n_i})$

Inferential settings

Model selection

- selection of fixed effects
- selection of random effects

Model checking

- homogeneity of variance
- linearity
- distributional assessment

Complications

Model selection

- remembering when you can use REML
- specifying degrees of freedom for F-tests (lots of special cases!)
- covariance parameters on the boundary of the parameter space

Model checking

- artificial structures in residual plots
- issues in testing homogeneity of the error terms with small groups
- distribution of EBLUPs does not match theoretical distribution

Why use visual inference?

- We can worry about choosing a graphic instead of remembering many rules/cases

- Provides a unified framework

- Protects us from over-interpreting artificial structures

- Still applicable when asymptotic results breakdown

Selecting random effects

- Typical strategy: Test $H_0 : \sigma^2_{b_1} = 0$ vs. $H_1 : \sigma^2_{b_1} > 0$ using a likelihood ratio test

- Problem: $\sigma^2_{b_1} = 0$ is on the boundary of the parameter space
  - The likelihood ratio test statistic does not have a χ^2 distribution
  - There is no one-size-fits-all approximation to the sampling distribution of the likelihood ratio test statistic

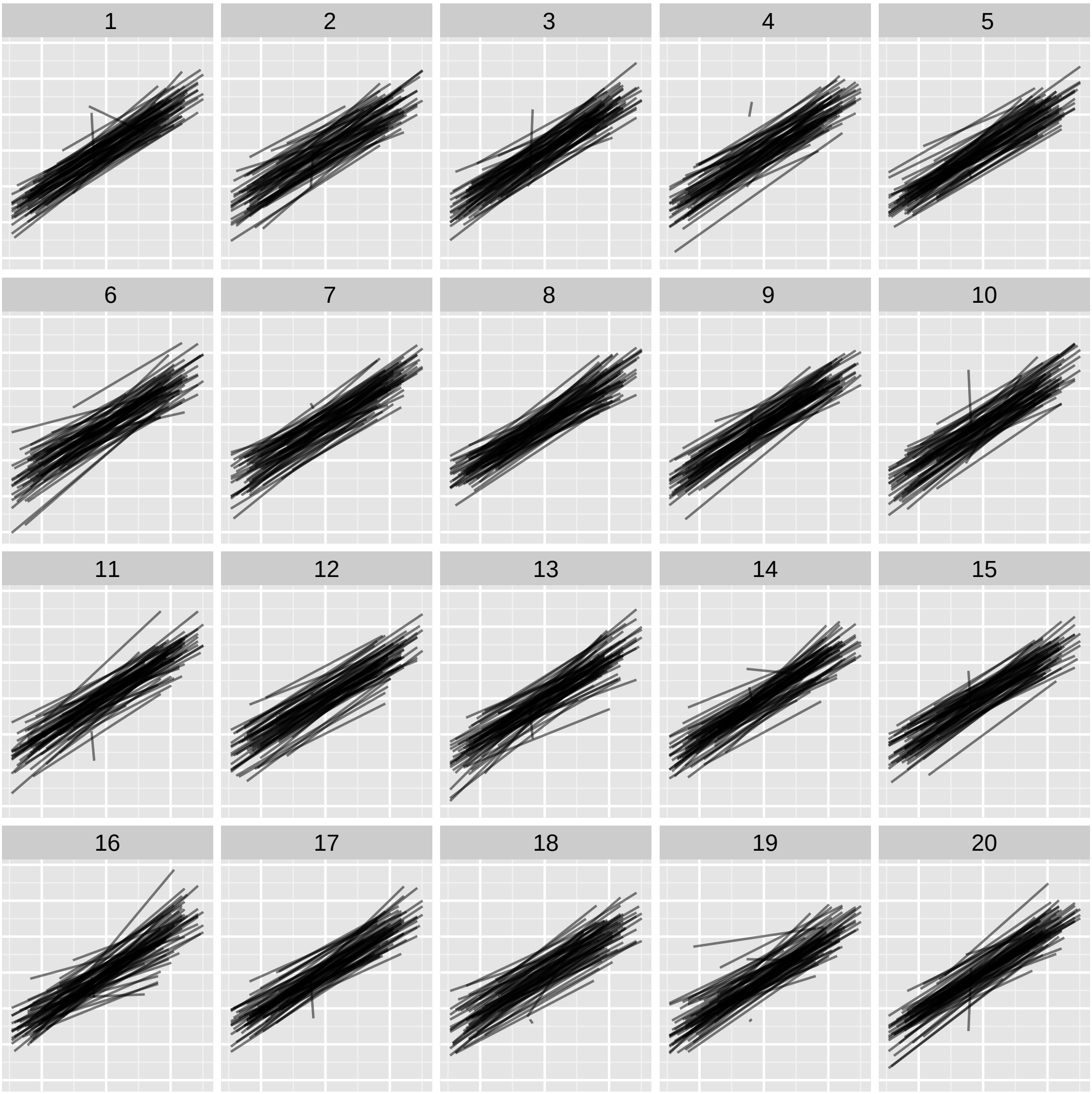
Selecting random effects

- Typical strategy: Test $H_0 : \sigma^2_{b_1} = 0$ vs. $H_1 : \sigma^2_{b_1} > 0$ using a likelihood ratio test

- Problem: $\sigma^2_{b_1} = 0$ is on the boundary of the parameter space
  - The likelihood ratio test statistic does not have a χ^2 distribution
  - There is no one-size-fits-all approximation to the sampling distribution of the likelihood ratio test statistic

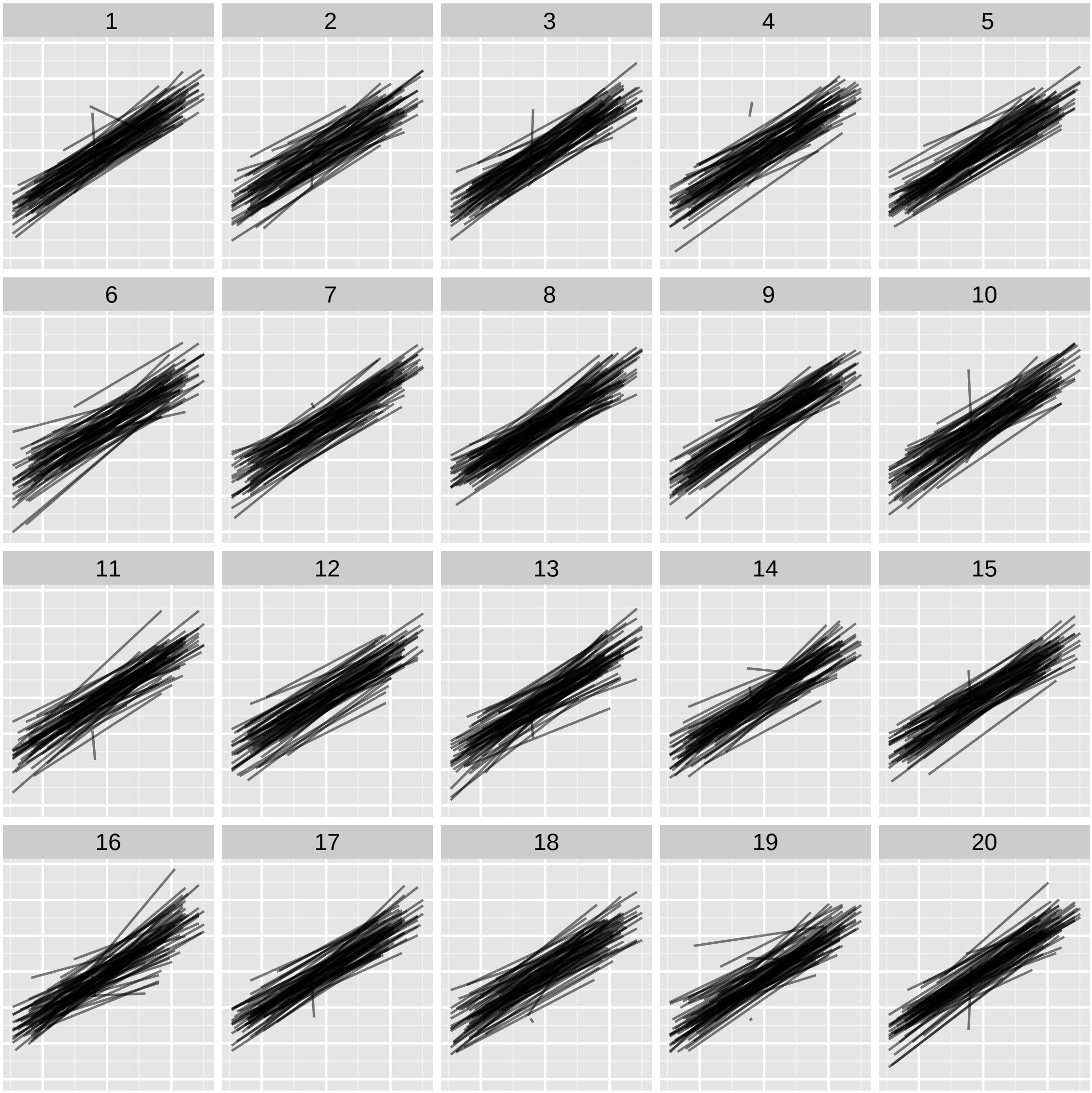
What about a lineup of the group trajectories?

Example: Multilevel data



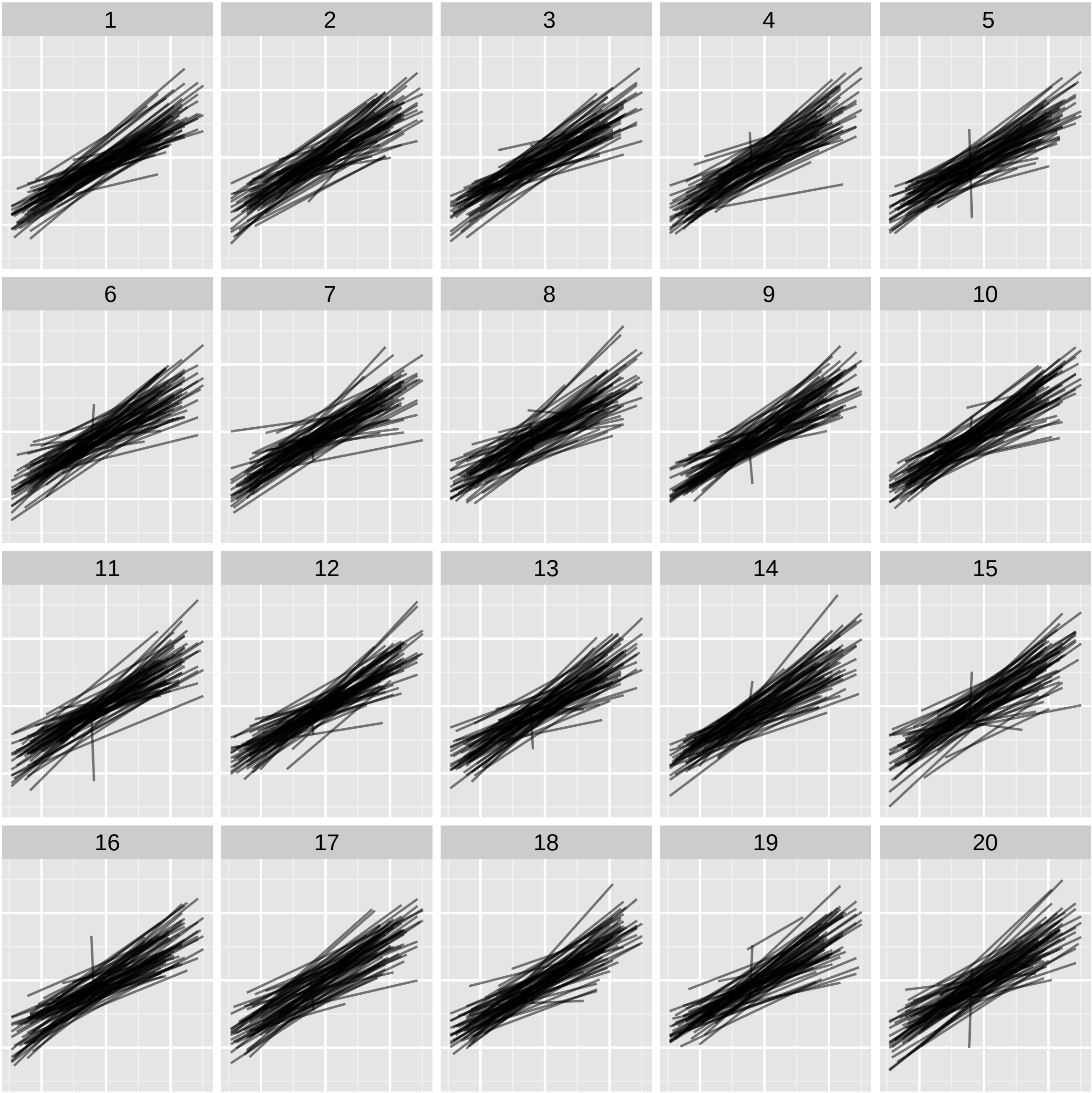
- Exam data, 4065 students in 65 schools

Example: Multilevel data



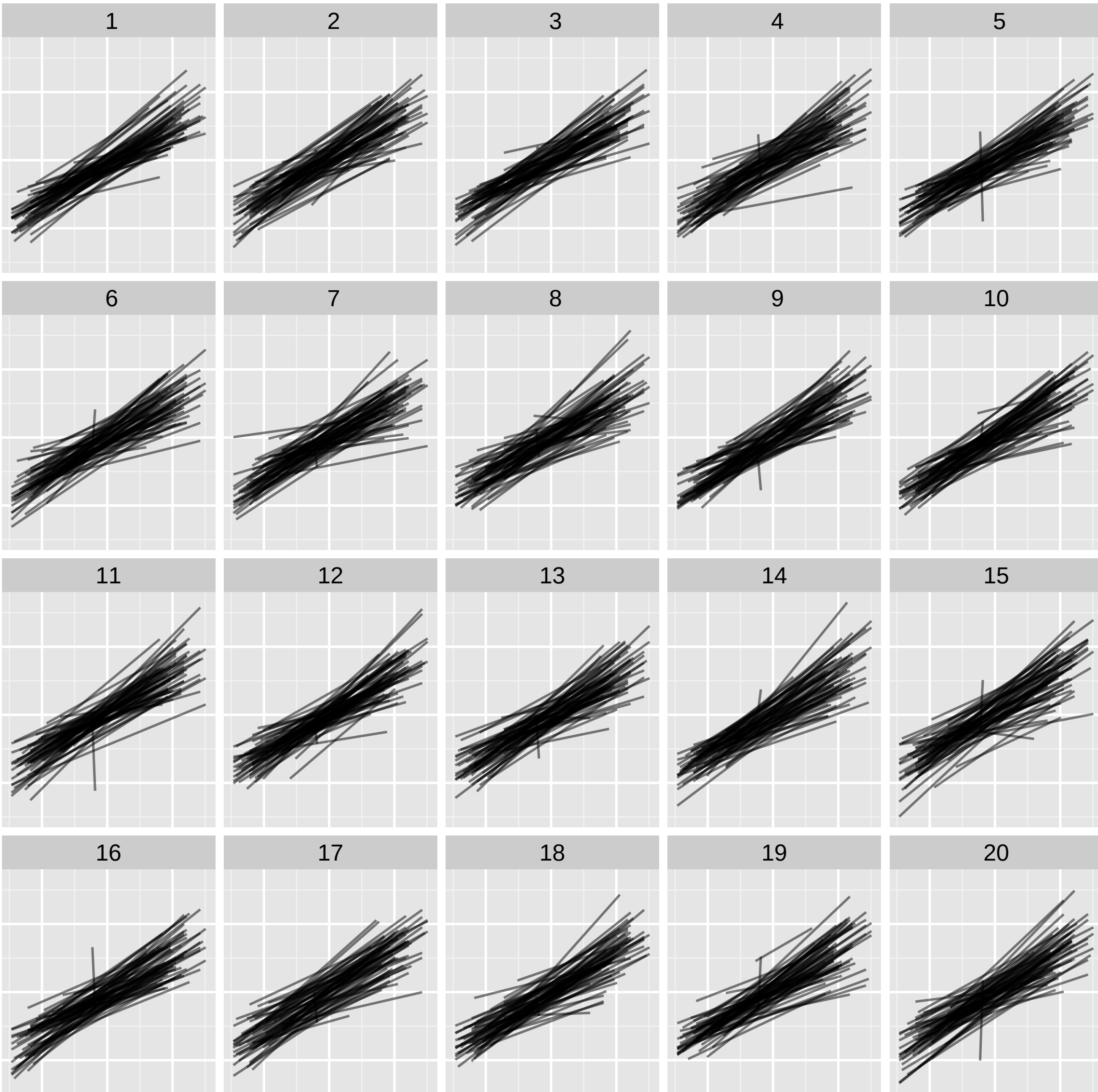
- Exam data, 4065 students in 65 schools
- True plot = 16
- 11 of 73 observers identified true plot (visual p-value of 0.0171)

Example: Multilevel data



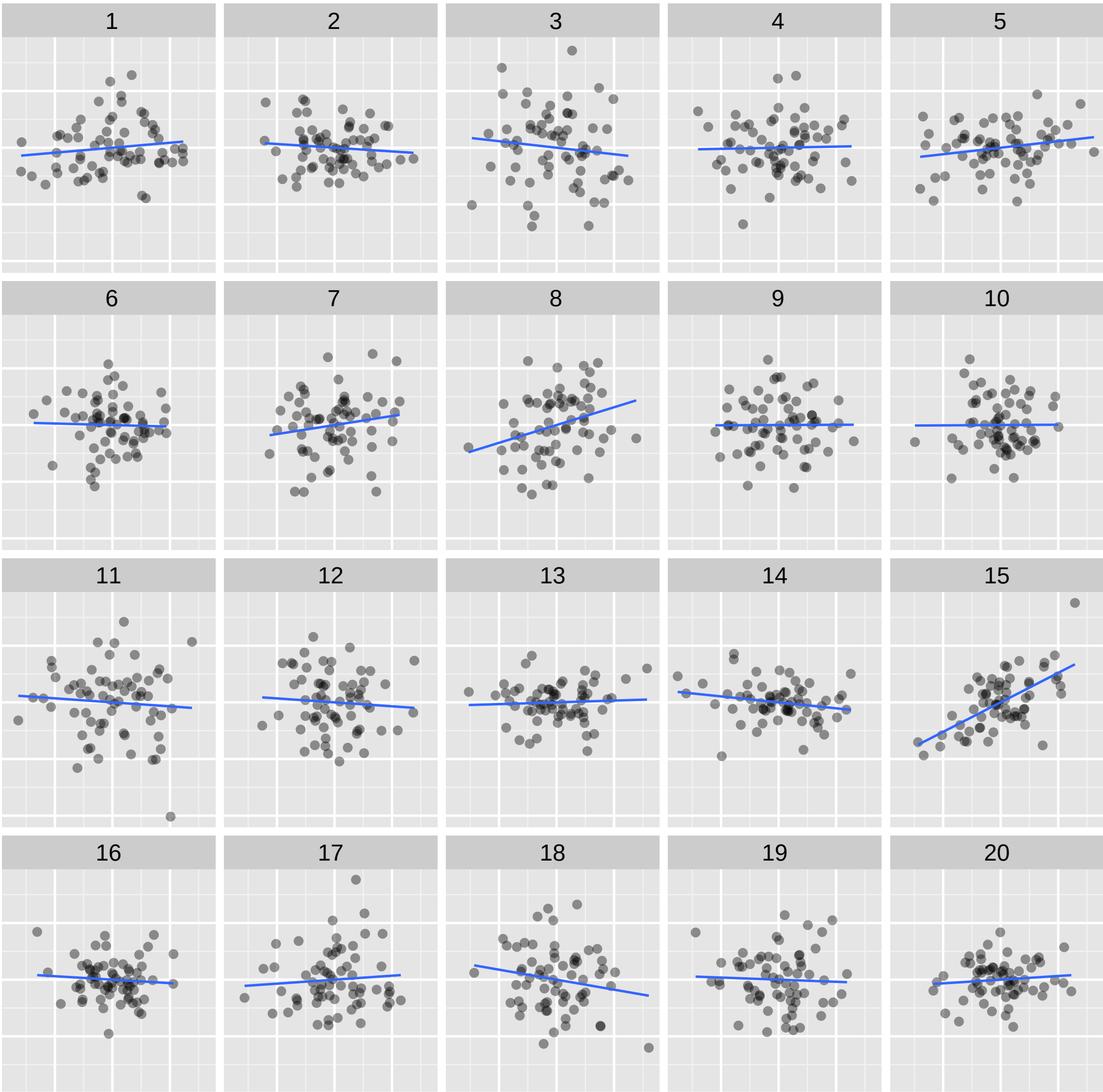
- Follow up: include a random slope for reading test score
- Need different observers

Example: Multilevel data



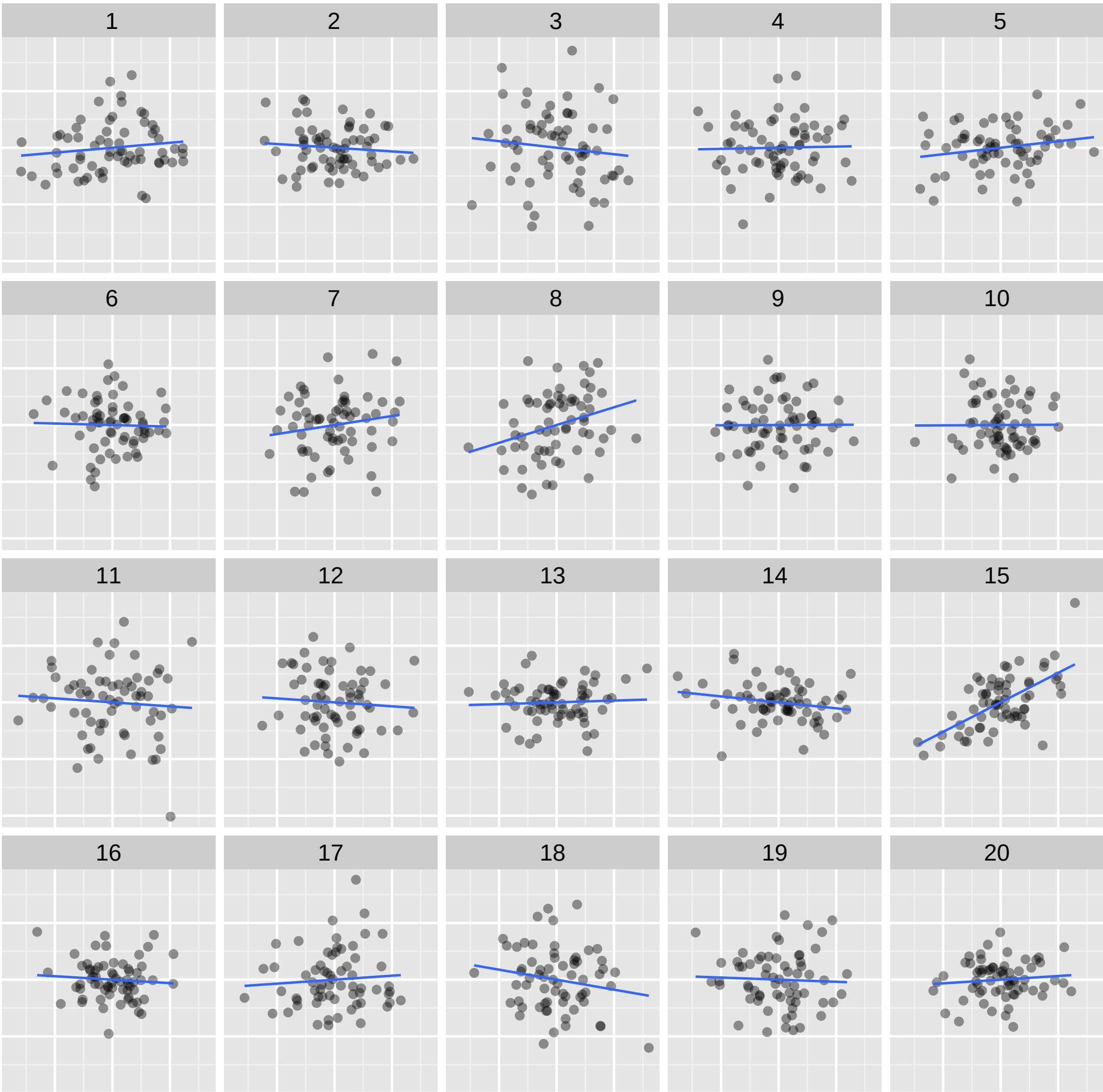
- Follow up including a random slope for reading test score
- Need different observers
- True plot = 18
- 0 of 64 observers identified true plot

What about correlation?



- Do we need to allow the random slope and intercept to be correlated?
- No technical need for visual test

What about correlation?



- Do we need to allow the random slope and intercept to be correlated?
- No technical need for visual test
- True plot = 15
- 41 of 69 observers identified true plot

Assessing homogeneity

- We assume that $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0},\ \sigma_\varepsilon^2)$

- Typical test statistic: $\displaystyle\sum_{i=1}^{g^*} d_i^2$
  - d_i^2 = the standardized measure of dispersion for a regression model fit to each group
  - g^* = number of groups that are "large enough" (often ≥ 10)

- Typical reference distribution: $\chi^2_{g^*-1}$

Assessing homogeneity

- We assume that $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0},\ \sigma_\varepsilon^2)$

- Typical test statistic: $\displaystyle\sum_{i=1}^{g^*} d_i^2$
  - d_i^2 = the standardized measure of dispersion for a regression model fit to each group
  - g^* = number of groups that are "large enough" (often ≥ 10)

- Typical reference distribution: $\chi^2_{g^*-1}$

This will fail when many/all group sizes are small!

Assessing homogeneity

- We assume that $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0},\ \sigma_\varepsilon^2)$

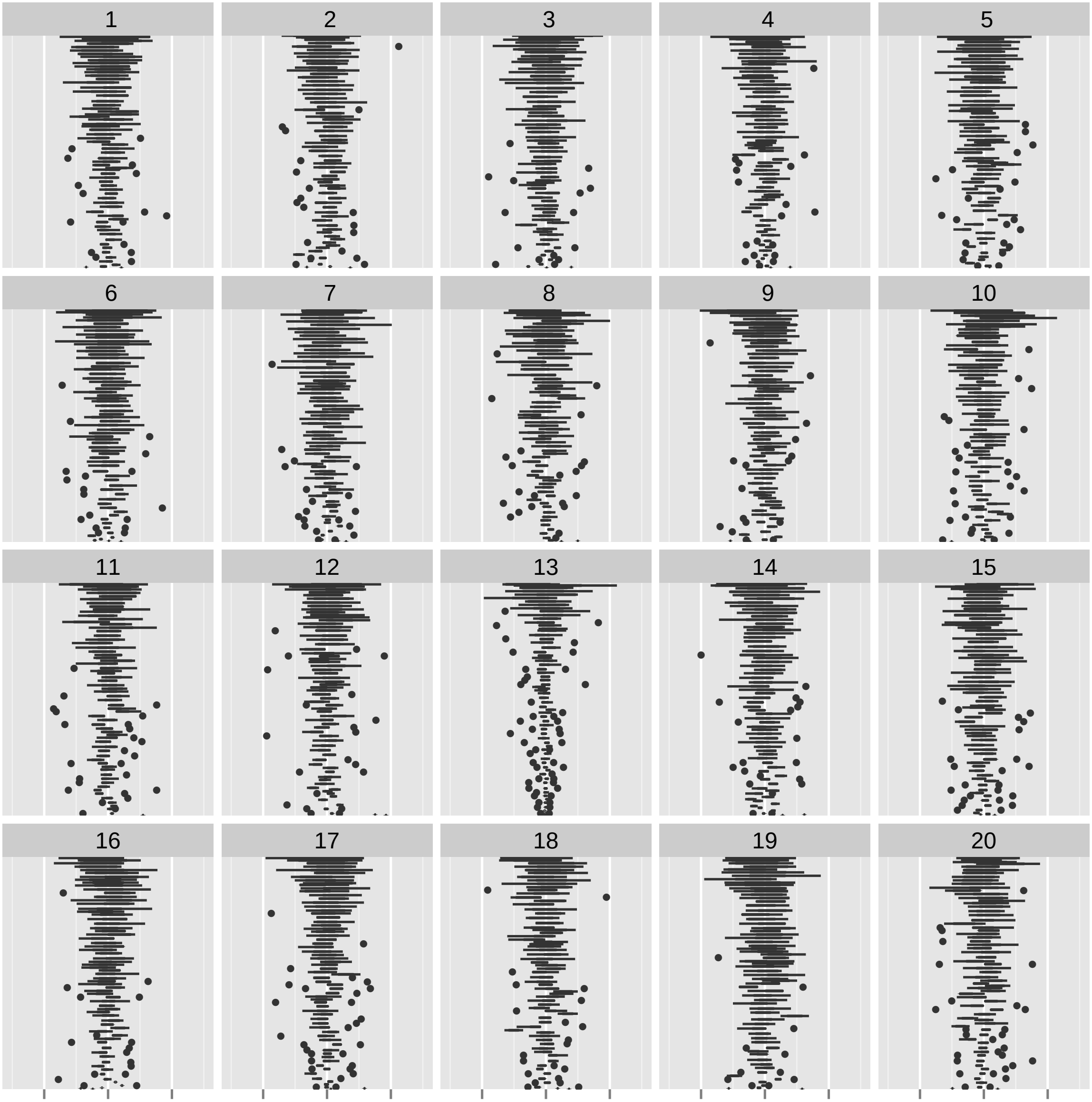
- Typical test statistic: $\displaystyle\sum_{i=1}^{g^*} d_i^2$
  - d_i^2 = the standardized measure of dispersion for a regression model fit to each group
  - g^* = number of groups that are "large enough" (often ≥ 10)

- Typical reference distribution: $\chi^2_{g^*-1}$

This will fail when many/all group sizes are small!

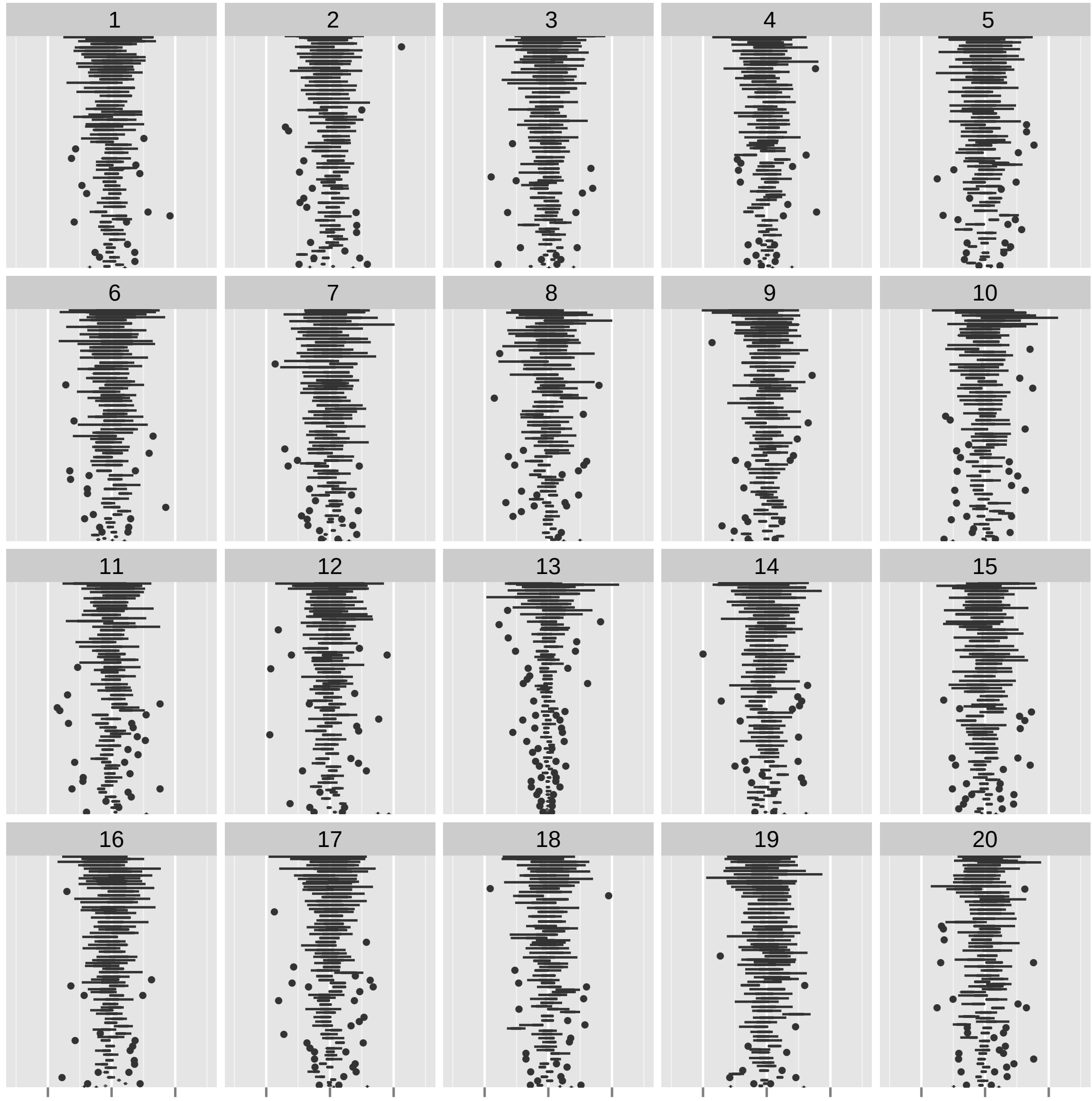
What about a lineup of side-by-side boxplots ordered by IQR?

Example: Longitudinal data



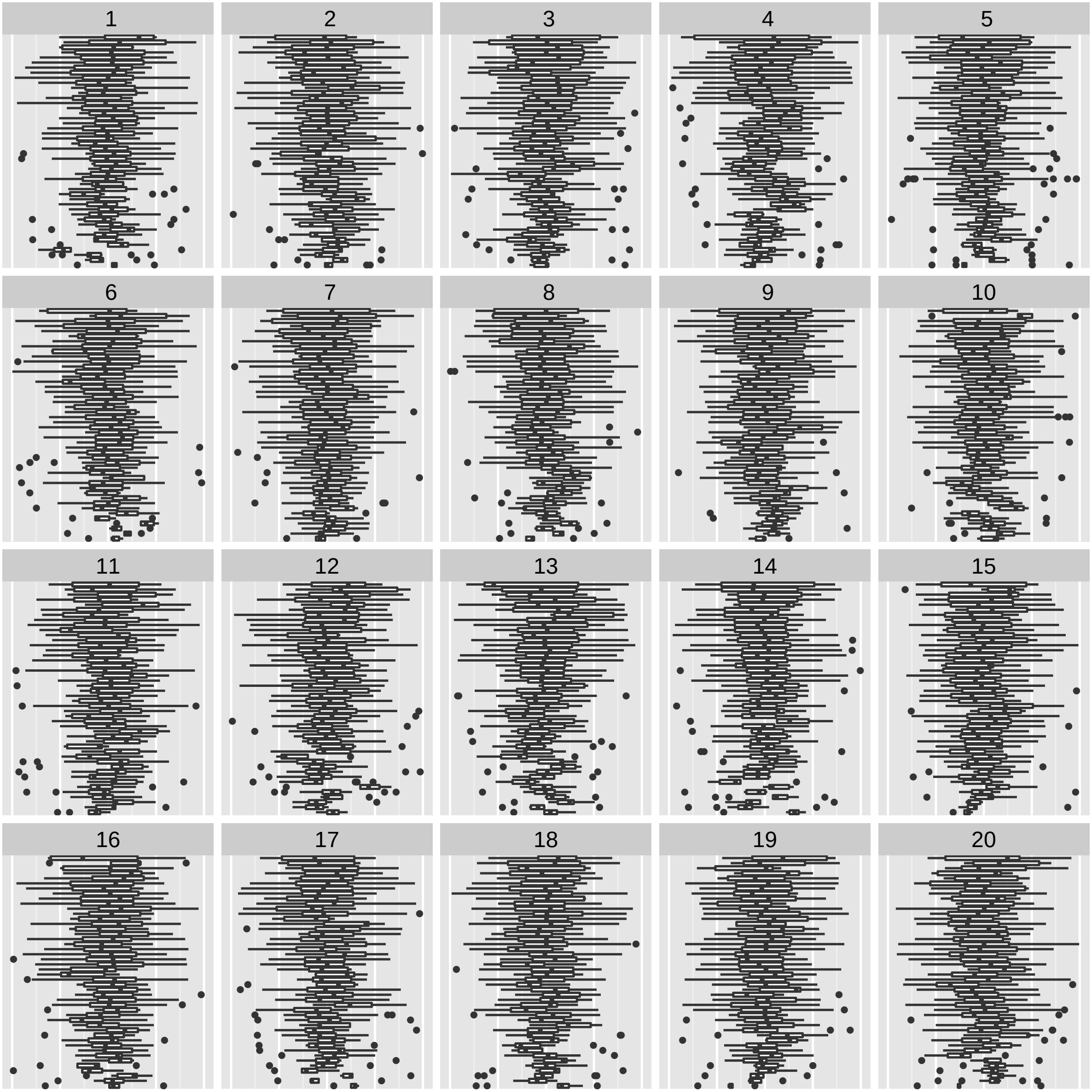
- Longitudinal data, 5 obs. per subject

Example: Longitudinal data



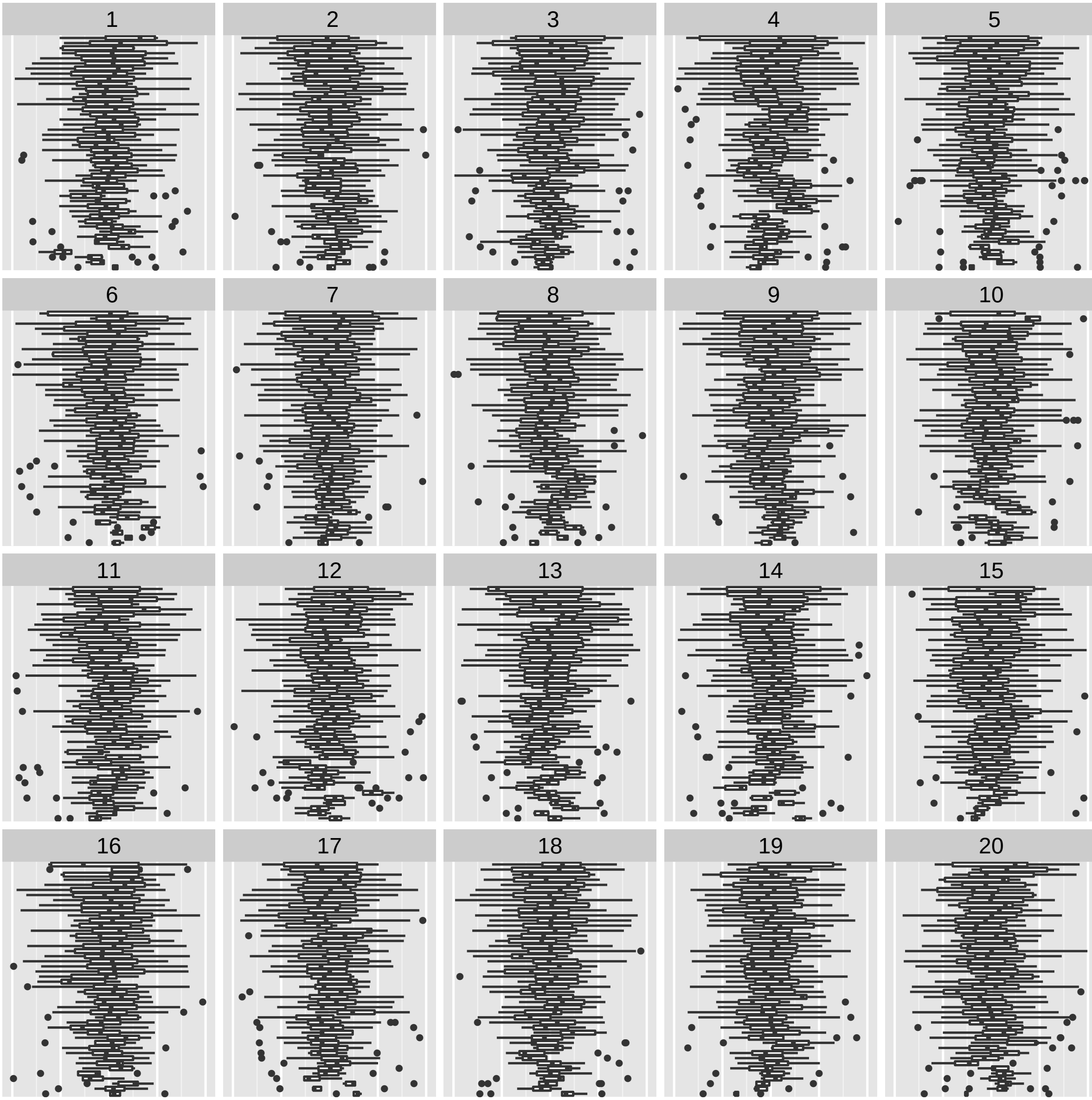
- Longitudinal data, 5 obs. per subject
- True plot = 13
- 50 of 75 observers identified true plot
- typical test p-value = 0.0886

Example: Multilevel data



- 85 counties,
  no. obs. ranges from
  1 to 116

Example: Multilevel data



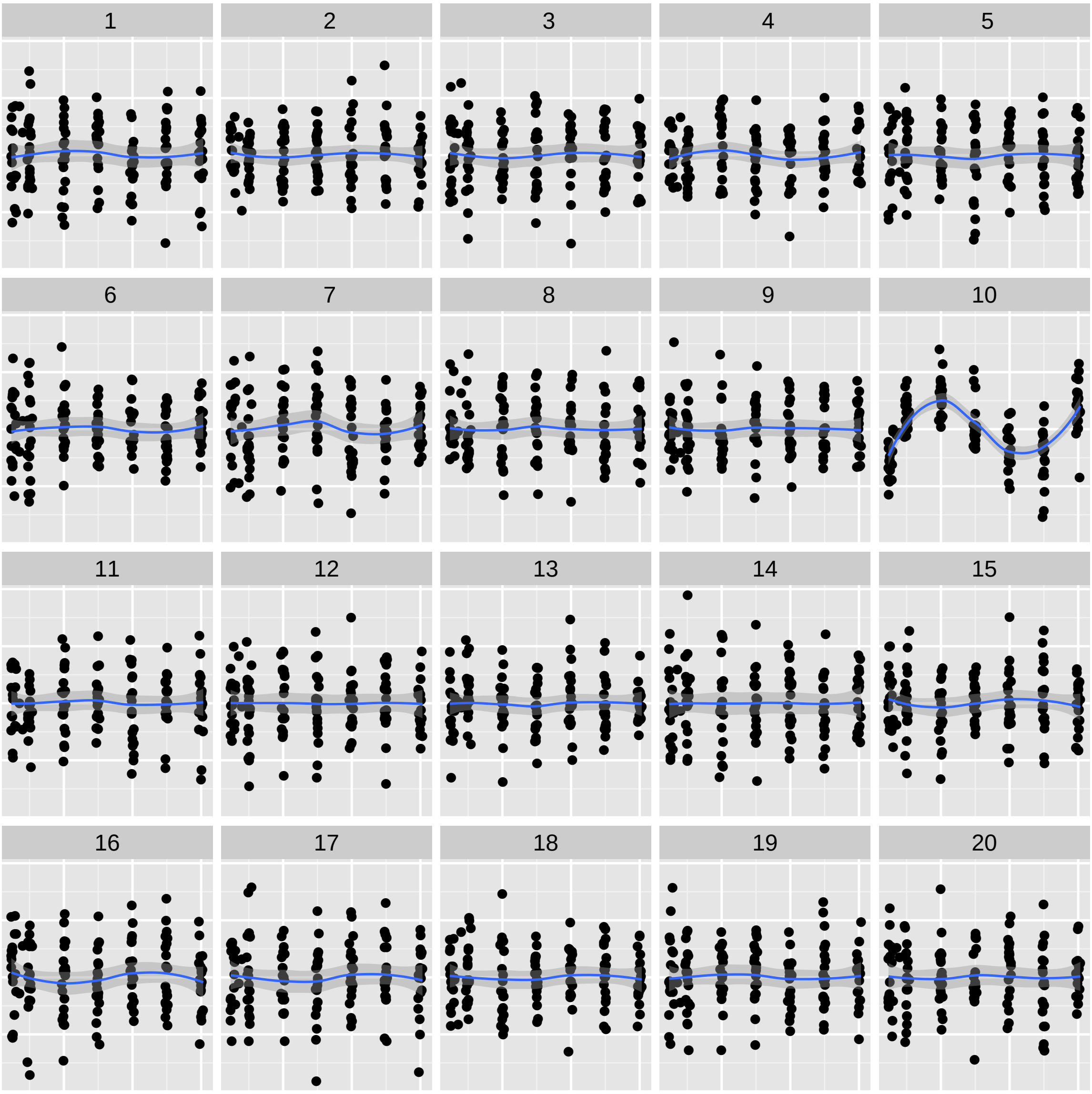
- 85 counties,
  no. obs. ranges from
  1 to 116

- True plot = 10,
  1 of 59 observers
  identified true plot

- typical test
  p-value = 0.24
  if small counties
  excluded

- p-value = 0.0185 if
  small counties are
  retained

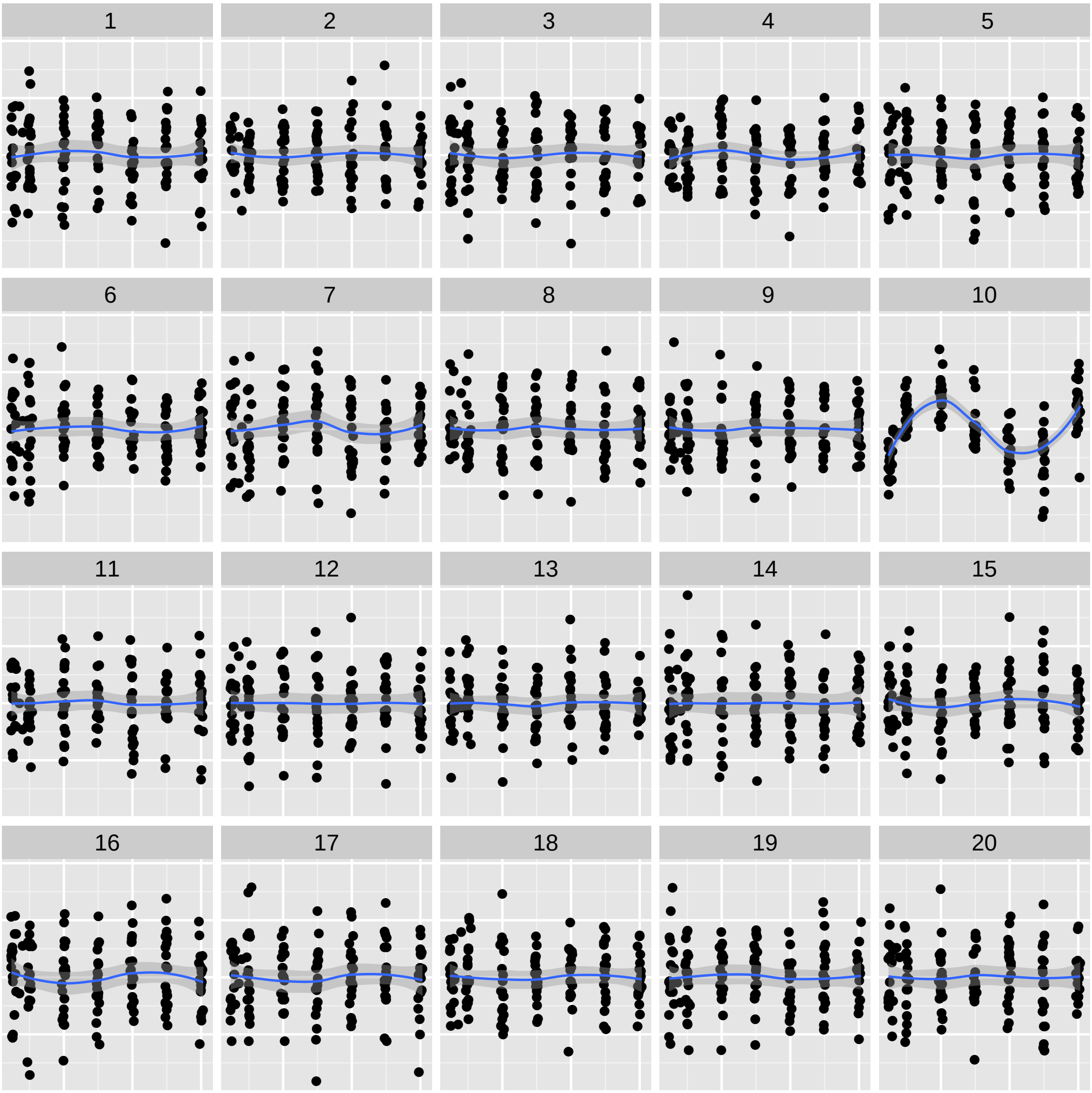
Assessing linearity



- 20 dialyzers,
  7 pressures

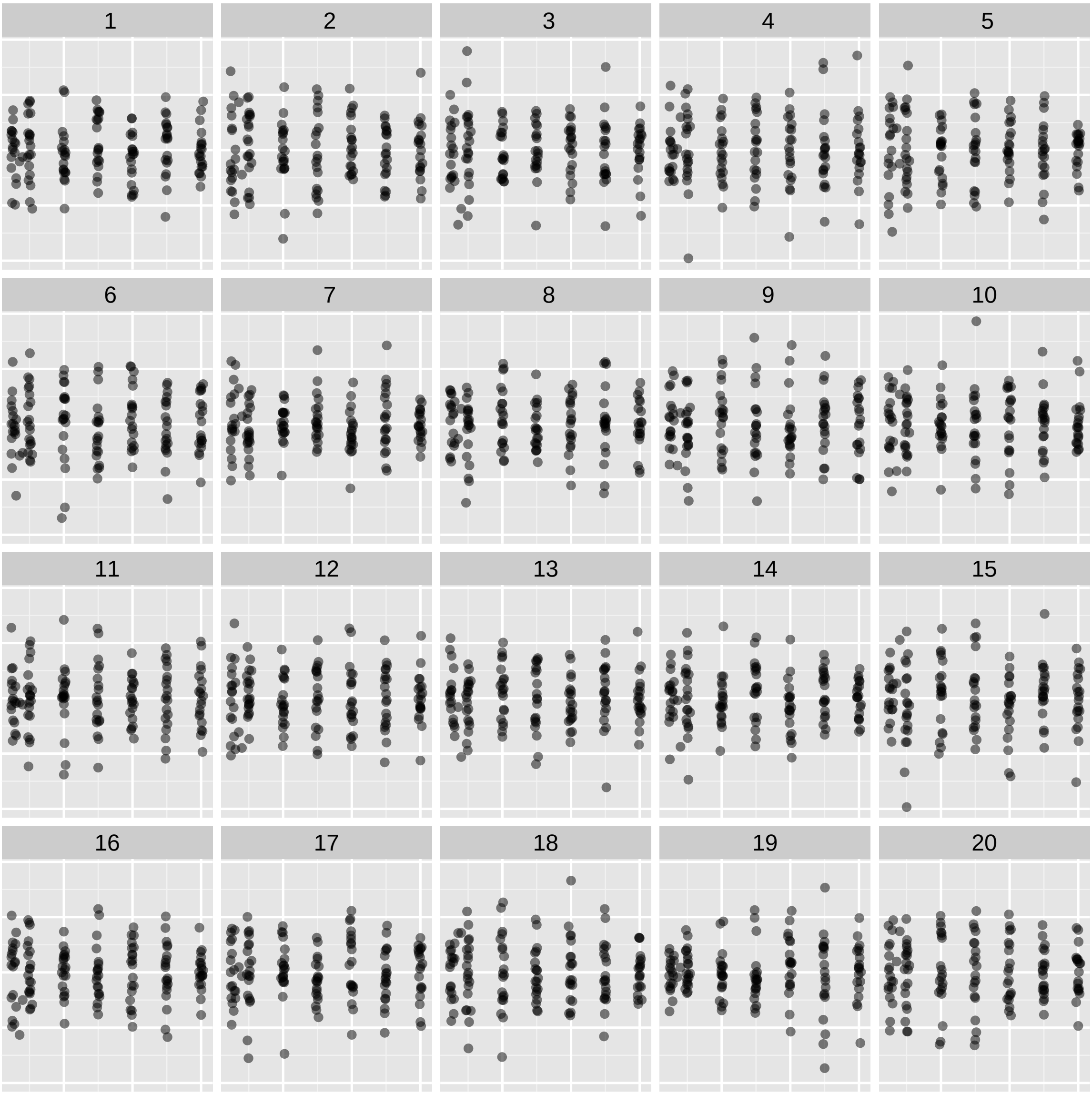
- polynomial of degree
  2 considered

Assessing linearity



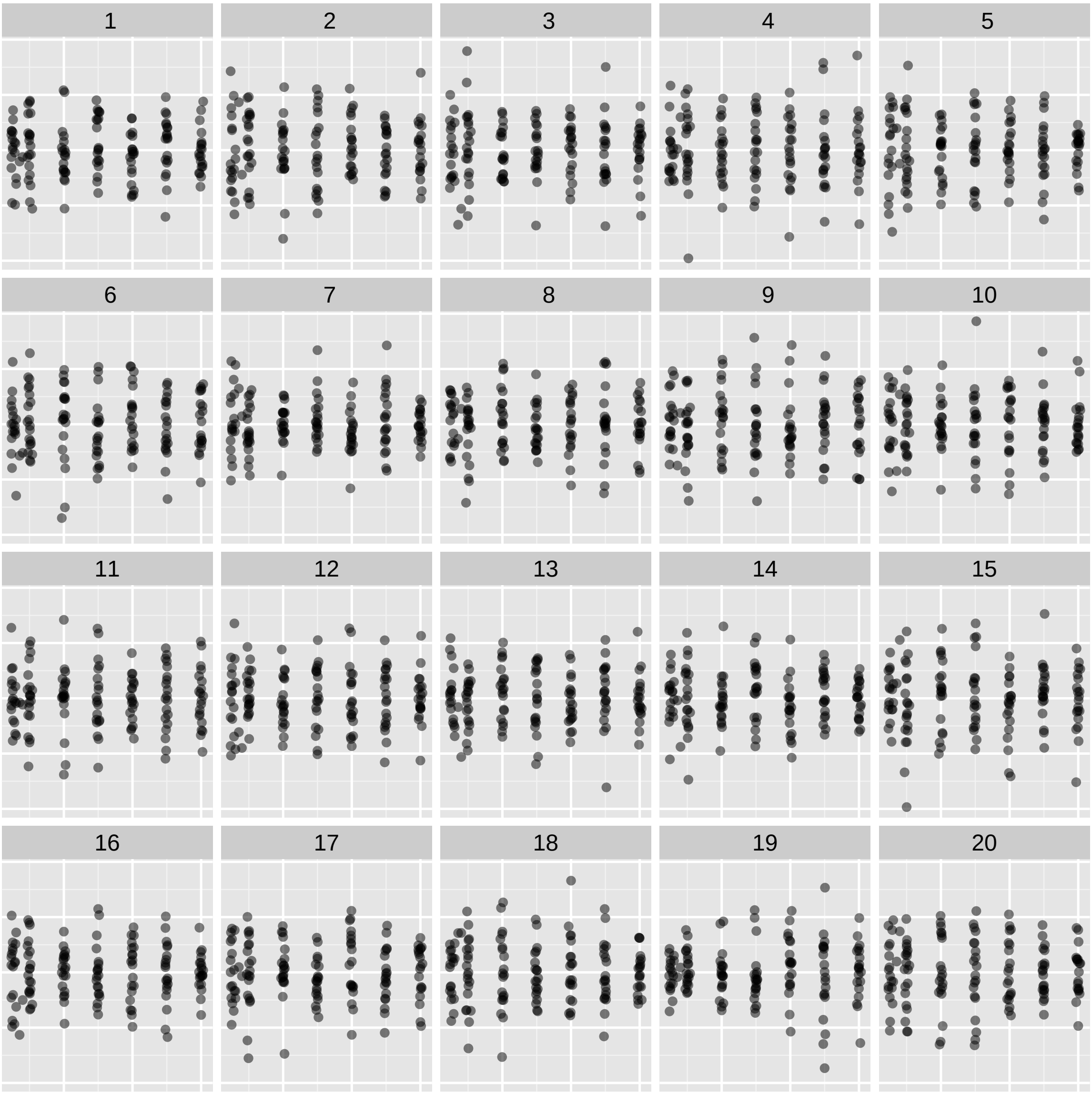
- 20 dialyzers, 7 pressures
- polynomial of degree 2 considered
- True plot = 10, 60 of 63 observers identified true plot

Back to homogeneity of error terms



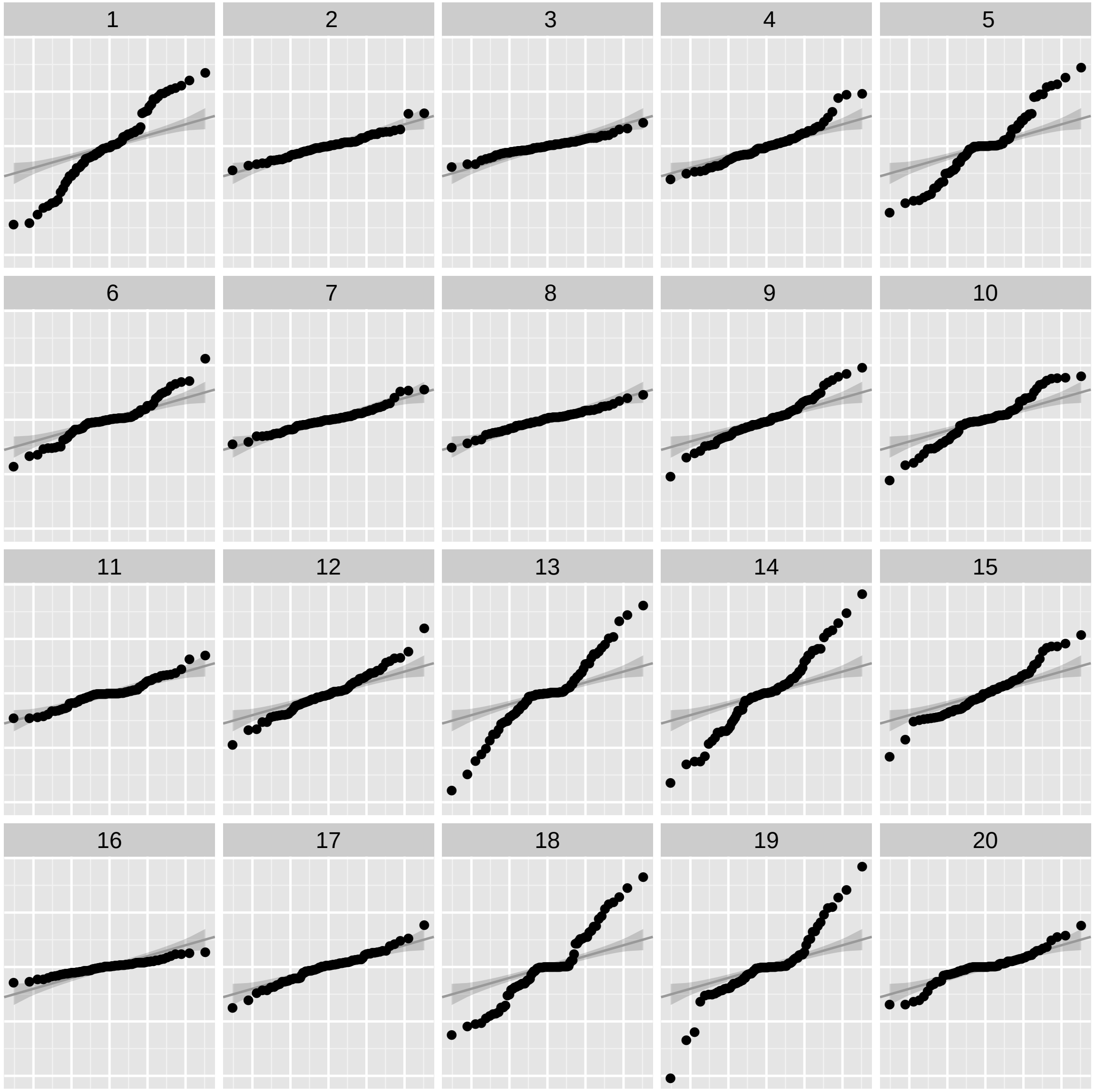
- polynomial of degree 4 considered

Back to homogeneity of error terms



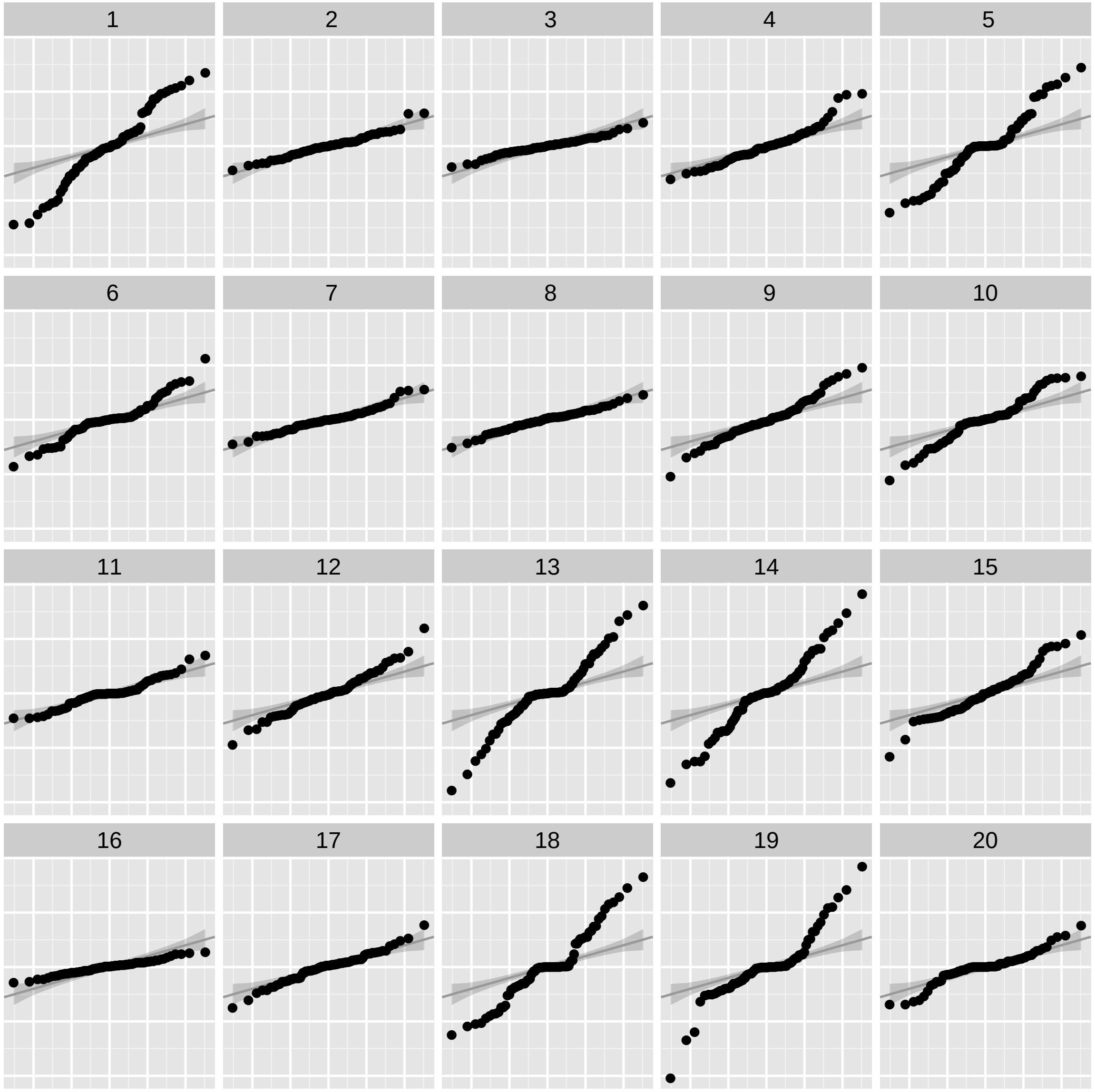
- polynomial of degree 4 considered
- True plot = 19, 29 of 85 observers identified true plot

Assessing normality



- Back to the radon data set

Assessing normality



- Back to the radon data set
- True plot = 10, 0 of 68 observers identified true plot
- Still exploring this visual test

Wrap up

Recap

- Lineups can be used to explore LMEs in situations where asymptotic results breakdown
- Provide a unified testing framework
- Dependent on the simulation process, design of the graphics, and the observers

Future work

- Evaluating the power of this framework to assess normality of random effects
- What impact does the simulation procedure have on the effectiveness of these tests? Can we use the standard guidelines for selecting a bootstrap procedure?