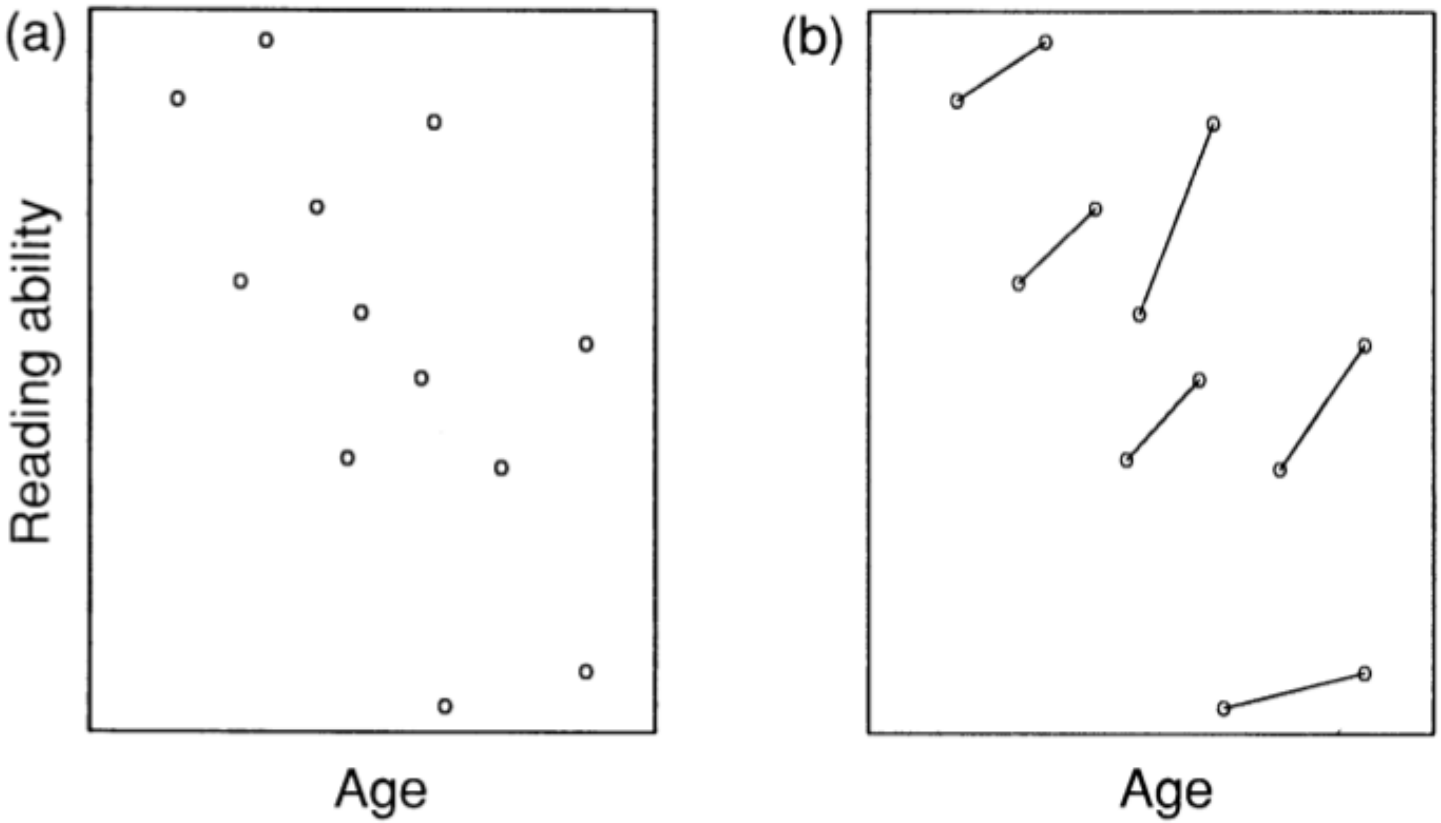
BIST P8110: Applied Regression II

23. Random Intercept Model

Qixuan Chen

Department of Biostatistics
Columbia University

Some hypothetical data



Random intercept model

- A random intercept model with one covariate for continuous responses is given by

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

where

- $b_i \overset{iid}{\sim} N(0, \tau^2)$
- $\epsilon_{ij} \overset{iid}{\sim} N(0, \nu^2)$
- b_i and ϵ_{ij} are independent

Fixed and random part

- The random intercept model has two parts:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij}$$

  - “Fixed" part: parameters that we estimate are the coefficients

$$\beta_0, \beta_1, \ldots$$

  - “random" part: parameters that we estimate are the variances

$$\tau^2 \text{ and } \nu^2$$

Variance

- Under the random intercept model with one covariate:

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

  where
  - $b_i \overset{iid}{\sim} N(0, \tau^2)$
  - $\epsilon_{ij} \overset{iid}{\sim} N(0, \nu^2)$
- The variance of y_{ij} can be derived as
  - $var(y_{ij}) =$

With-subject covariance

- Under the random intercept model with one covariate:

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

  where
  - $b_i \overset{iid}{\sim} N(0, \tau^2)$
  - $\epsilon_{ij} \overset{iid}{\sim} N(0, \nu^2)$
- Within-subject covariance
  - $cov(y_{ij}, y_{ij'}) =$

Within-subject correlation

- Under the random intercept model with one covariate:

$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}$$

  where
  - $b_i \overset{iid}{\sim} N(0, \tau^2)$
  - $\epsilon_{ij} \overset{iid}{\sim} N(0, \nu^2)$
- The correlation of any two observations from the same subject can be derived as
  - $cor(y_{ij}, y_{ij'}) = \rho$

Correlation structure

- The random intercept model implies a correlation structure
  - $$\begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \\ \rho & \rho & & 1 \end{bmatrix}$$

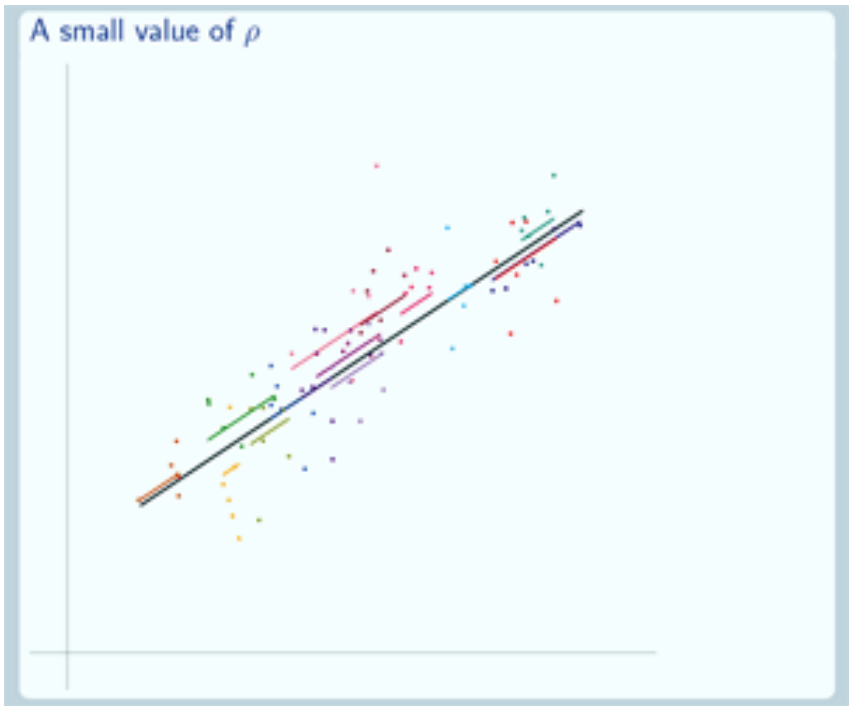
Correlation ρ

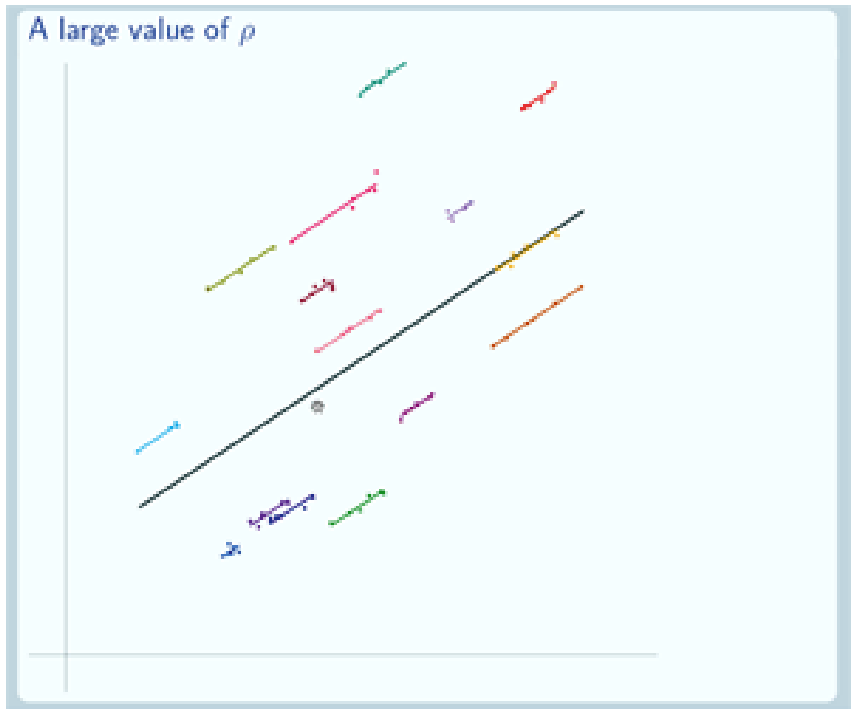
- The quantity ρ
  - tells how strongly observations within a subject are correlated
  - represents proportion of total variance that is "explained" by subjects
- When there are no covariates X in the model
  - ρ is also called intracluster correlation coefficient (ICC)

ρ and clustering

- Large ρ
  - a lot of the variance is at subject level
  - observations within each subject are more similar
  - more variation between subjects
  - values of the response are largely determined by which subject the unit belongs to
  - the data are clustered
- What about small ρ?

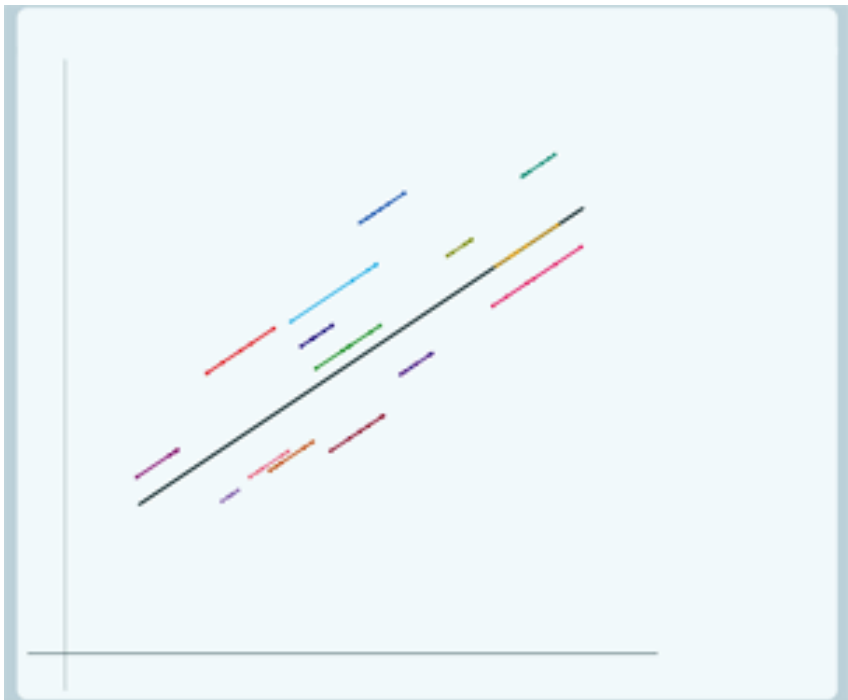
Large vs. small ρ



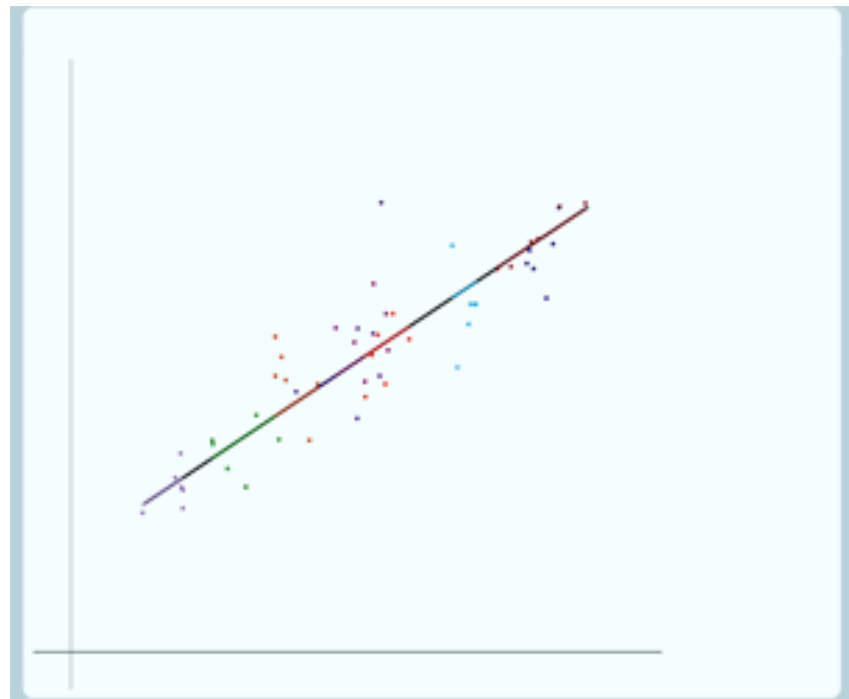


ρ=0 or 1?

ρ =



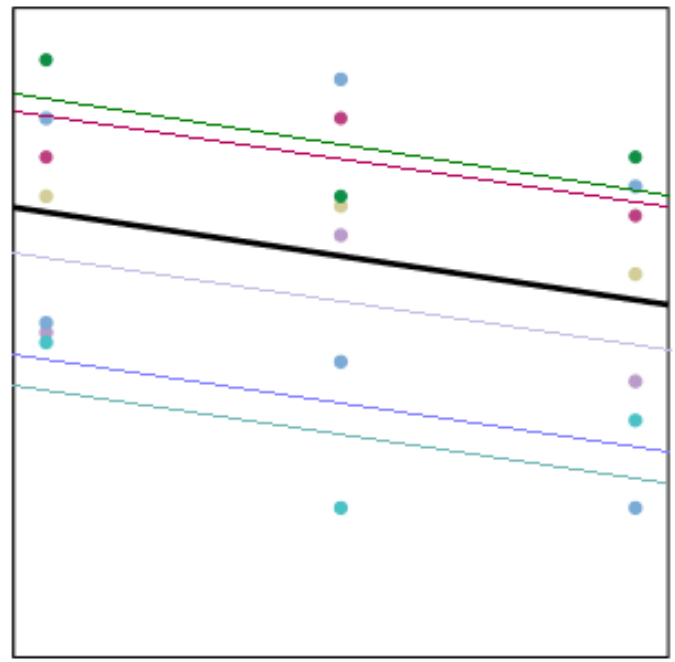
ρ =



Regression lines

- For a random intercept model
  - The intercept for the overall regression line is still β_0
  - For each subject the intercept is $\beta_0 + b_i$

- How the fitted regression lines look like?
  - The overall average line has equation $\beta_0 + \beta_1 x_{ij}$
  - Each subject has its own line, parallel to the overall average line $(\beta_0 + b_i) + \beta_1 x_{ij}$



Interpretation

- "Fixed" part
  - β_1 is the mean increase in the response for each unit increase in X
  - the same as that in linear regression models
- "Random" part
  - τ^2 is the unexplained variance at subject level after controlling for the covariates
  - ν^2 is the unexplained variation at observation level after controlling for the covariates

Adding more covariates

- The random intercept model with one covariate can be easily extended to allow multiple covariates
  - covariates can be continuous or categorical
  - can be at observation level or subject level
  - can include interaction terms

Hypothesis testing

- "Fixed" part
  - $H_0 : \beta_1 = 0$ vs. $H_\alpha : \beta_1 \neq 0$
  - t-test or F-test
- "Random" part
  - $H_0 : \tau^2 = 0$ vs. $H_\alpha : \tau^2 \neq 0$
  - Likelihood ratio test: $G = -2(l_{\text{without RI}} - l_{\text{with RI}})$
  - p-value: $Pr(\chi^2_1 \geq G)$

Residuals

- Random intercept model has two residuals
  - subject level residuals $\hat{b}_i$
  - observation level residual $\hat{\epsilon}_{ij}$

PROC MIXED

Sample SAS code:

```
proc mixed data=A;
   class ID;
   model Y = X /s;
   random int / subject=ID s;
run;
```

Case Study: Blood Flow Data

- Revisit the blood flow example in GEE
  - with the goal of assessing the effect of race and dose of isoproterenol on blood flow
  - the repeated measures data for the first subject:

id	dose	race	fbf
1	0	1	1.0000
1	10	1	1.4000
1	20	1	6.4000
1	60	1	19.1000
1	150	1	25.0000
1	300	1	24.6000
1	400	1	28.0000

Research Questions

- The key research questions to address in this study are
  1. whether forearm blood flow increases as the dose of isoproterenol increases,
  2. and whether the effect of dose of isoproterenol on blood flow is different between normotensive black and white men.

The Random Intercept Model

- The random intercept model for the forearm blood flow is

$$y_{ij} = \beta_0 + \beta_1 \times dose_{ij} + \beta_2 \times race_i + \beta_3 \times dose_{ij} \times race_i + b_i + \epsilon_{ij}$$

where,

  - Y_{ij} = forearm blood flow (ml/min/dl)
  - race = 1 if white, = 2 if black (0 if white, 1 if black, after using the CLASS statement)
  - dose = $0, 10, 20, 60, 150, 300,$ and 400 (ng/min)
  - b_i is the deviation from the mean intercept for subject i, with $b_i \sim N(0, \tau^2)$
  - ϵ_{ij} is the observation-level error term with $\epsilon_{ij} \sim N(0, \sigma^2)$

SAS Codes

```
proc mixed data=isoproterenol;
class race id;
model fbf = race dose race*dose /s;
random int / subject=id s;
run;
```

NOTE: This is similar to the GEE model with "type=cs" assuming compound symmetry correlation matrix.

SAS Output

Dimensions

Covariance Parameters	2
Columns in X	6
Columns in Z Per Subject	1
Subjects	22
Max Obs Per Subject	7

Number of Observations

Number of Observations Read	154
Number of Observations Used	150
Number of Observations Not Used	4

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	961.03396866	
1	2	910.45133653	0.00000007
2	1	910.45131262	0.00000000

Convergence criteria met.

SAS Output

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	id	15.9302
Residual		17.7394

Fit Statistics

-2 Res Log Likelihood	910.5
AIC (smaller is better)	914.5
AICC (smaller is better)	914.5
BIC (smaller is better)	916.6

Solution for Fixed Effects

Effect	race	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		3.1463	1.5126	20	2.08	0.0506
race	1	3.0360	1.9741	126	1.54	0.1266
race	2	0	.	.	.	.
dose		0.01522	0.003620	126	4.20	<.0001
dose*race	1	0.03431	0.004739	126	7.24	<.0001
dose*race	2	0	.	.	.	.

SAS Output

Solution for Random Effects

Effect	id	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	1.9310	1.8045	126	1.07	0.2866
Intercept	2	1.5735	1.8045	126	0.87	0.3849
Intercept	3	-3.6030	1.8045	126	-2.00	0.0480
Intercept	4	-1.5274	1.8045	126	-0.85	0.3989
Intercept	5	-1.5200	1.8045	126	-0.84	0.4012
Intercept	6	1.5242	1.8045	126	0.84	0.3999
Intercept	7	-6.4944	1.8045	126	-3.60	0.0005
Intercept	8	2.0712	2.2536	126	0.92	0.3598
Intercept	9	11.4459	1.8045	126	6.34	<.0001
Intercept	10	3.8907	1.8045	126	2.16	0.0330
Intercept	11	-3.6141	1.8045	126	-2.00	0.0473
Intercept	12	-4.3548	1.8045	126	-2.41	0.0173
Intercept	13	-1.3228	1.8045	126	-0.73	0.4649
Intercept	14	-1.3879	1.9270	126	-0.72	0.4727
Intercept	15	0.3117	1.9270	126	0.16	0.8718
Intercept	16	-1.4730	1.9270	126	-0.76	0.4461
Intercept	17	-0.2220	1.9270	126	-0.12	0.9085

Residuals

- Raw residuals
  - marginal residuals: $r_{mij} = y_{ij} - (\hat{\beta}_0 + X^T\hat{\beta})$
  - conditional residuals: $r_{cij} = r_{mij} - \hat{b}_i$
- Studentized residuals
  - marginal studentized residuals: $r_{mij}^{student} = \frac{r_{mij}}{\sqrt{\widehat{Var}(r_{mij})}}$
  - conditional studentized residuals: $r_{cij}^{student} = \frac{r_{cij}}{\sqrt{\widehat{Var}(r_{cij})}}$
- Residuals plots using PROC MIXED

ODS Graph Name	Plot Description	Statement
ResidualBoxplot	Box plot of (raw) residuals	PLOTS=RESIDUALPANEL(UNPACK BOX)
ResidualByPredicted	Residuals vs. predicted	PLOTS=RESIDUALPANEL(UNPACK)
ResidualHistogram	Histogram of raw residuals	PLOTS=RESIDUALPANEL(UNPACK)
ResidualPanel	Panel of (raw) residuals	MODEL / RESIDUAL
ResidualQQplot	Q-Q plot of raw residuals	PLOTS=RESIDUALPANEL(UNPACK)
StudentBoxplot	Box plot of studentized residuals	PLOTS=STUDENTPANEL(UNPACK BOX)
StudentByPredicted	Studentized residuals vs. predicted	PLOTS=STUDENTPANEL(UNPACK)
StudentHistogram	Histogram of studentized residuals	PLOTS=STUDENTPANEL(UNPACK)
StudentPanel	Panel of studentized residuals	MODEL / RESIDUAL
StudentQQplot	Q-Q plot of studentized residuals	PLOTS=STUDENTPANEL(UNPACK)

Model Diagnostics

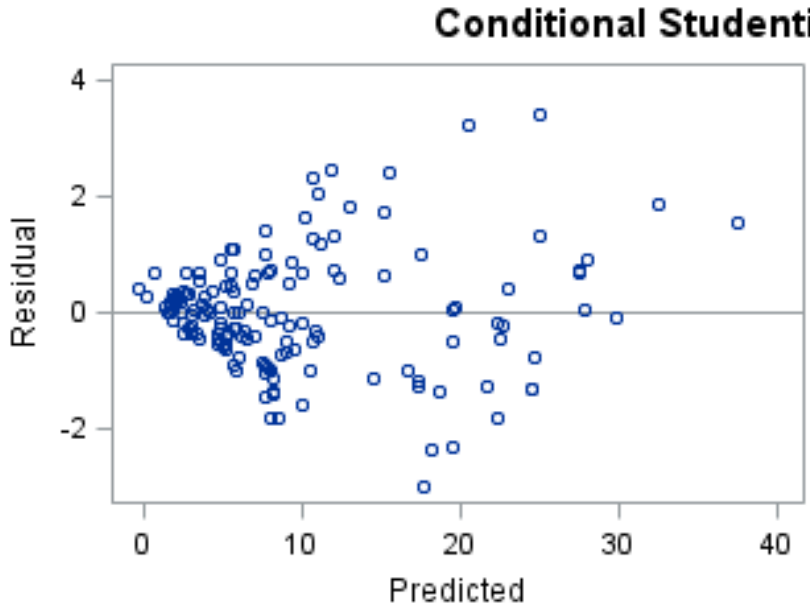
- Influence diagnostics
  - inluence on parameter estimates: Cook's D (< 1)
  - influence on precision of estimates: CovRatio (> 1 for precision loss and < 1 for precision gain by excluding the obs)
- Deletion estimates
- Syntax in PROC MIXED

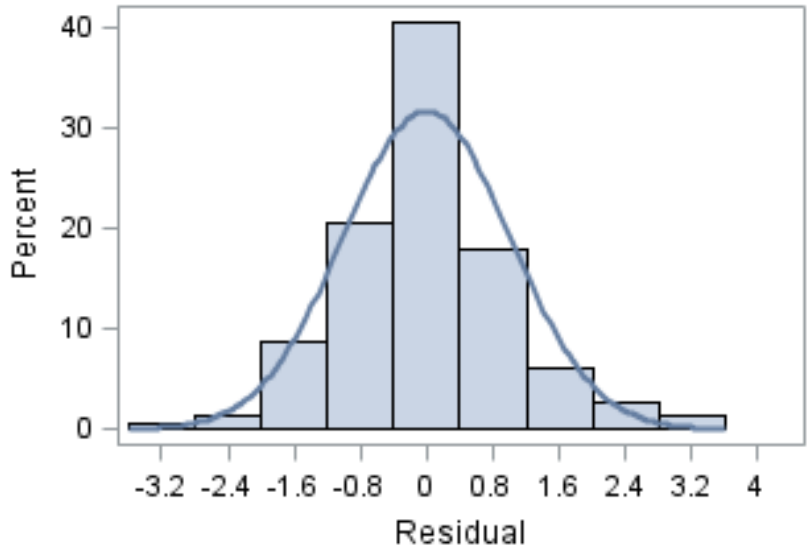
ODS Graph Name	Plot Description	Statement
InfluenceEstPlot	Panel of deletion estimates	MODEL / INFLUENCE(EST)
		or PLOTS=INFLUENCEESTPLOT and MODEL / INFLUENCE
InfluenceStatPanel	Panel of influence statistics	MODEL / INFLUENCE

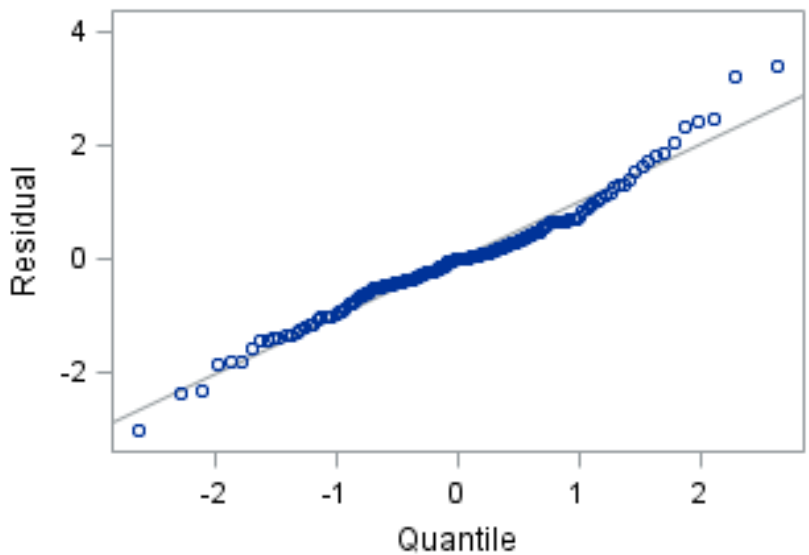
SAS Codes: residuals and diagnostics

```
ods graphics on;
proc mixed data=isoproterenol;
 class race id;
 model fbf = race dose race*dose /s residual
                          influence(iter=5 effect=id est);
 random int /subject=id s;
run;
ods graphics off;
```

SAS output: residuals





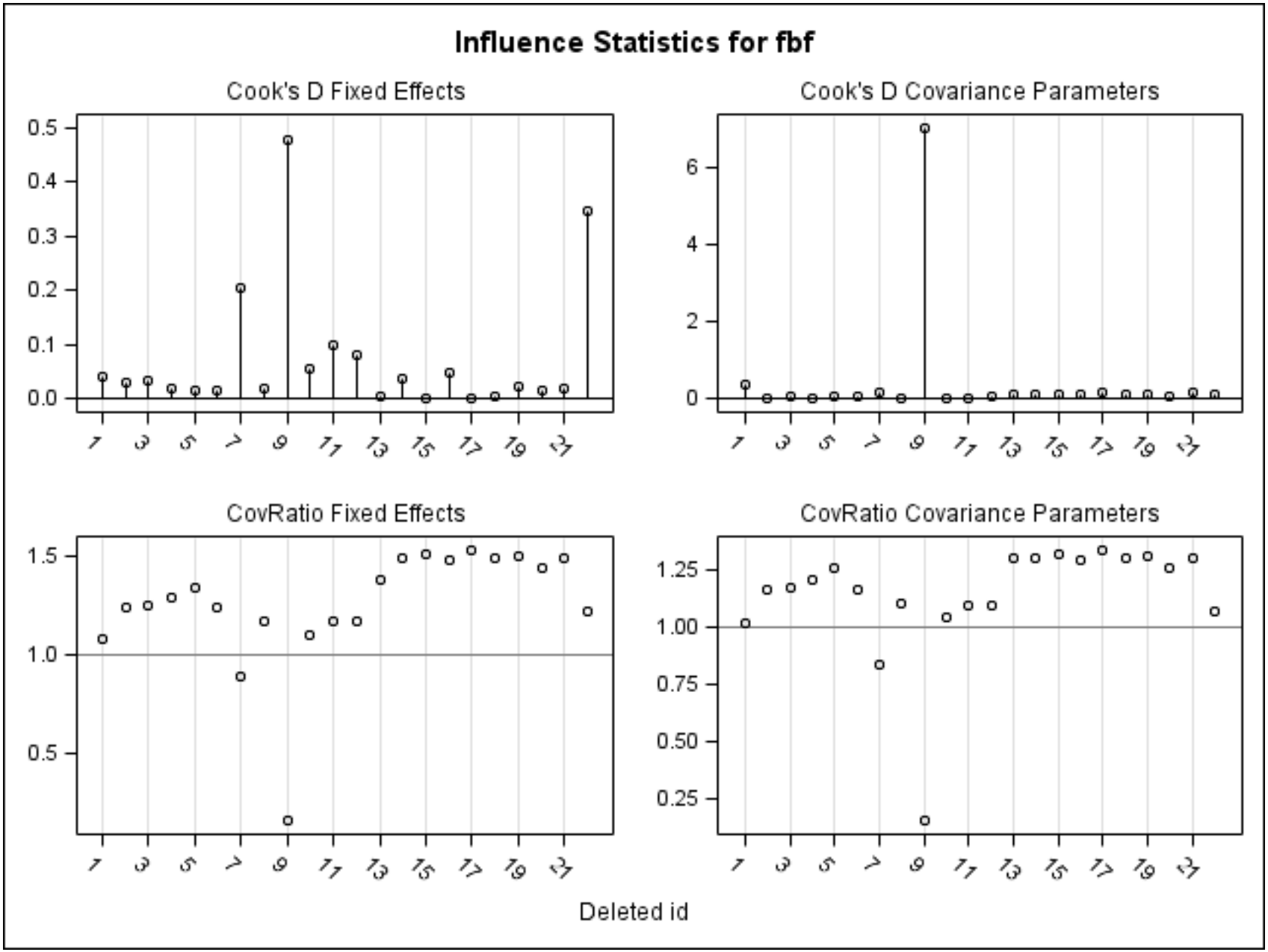


Conditional Studentized Residuals for fbf

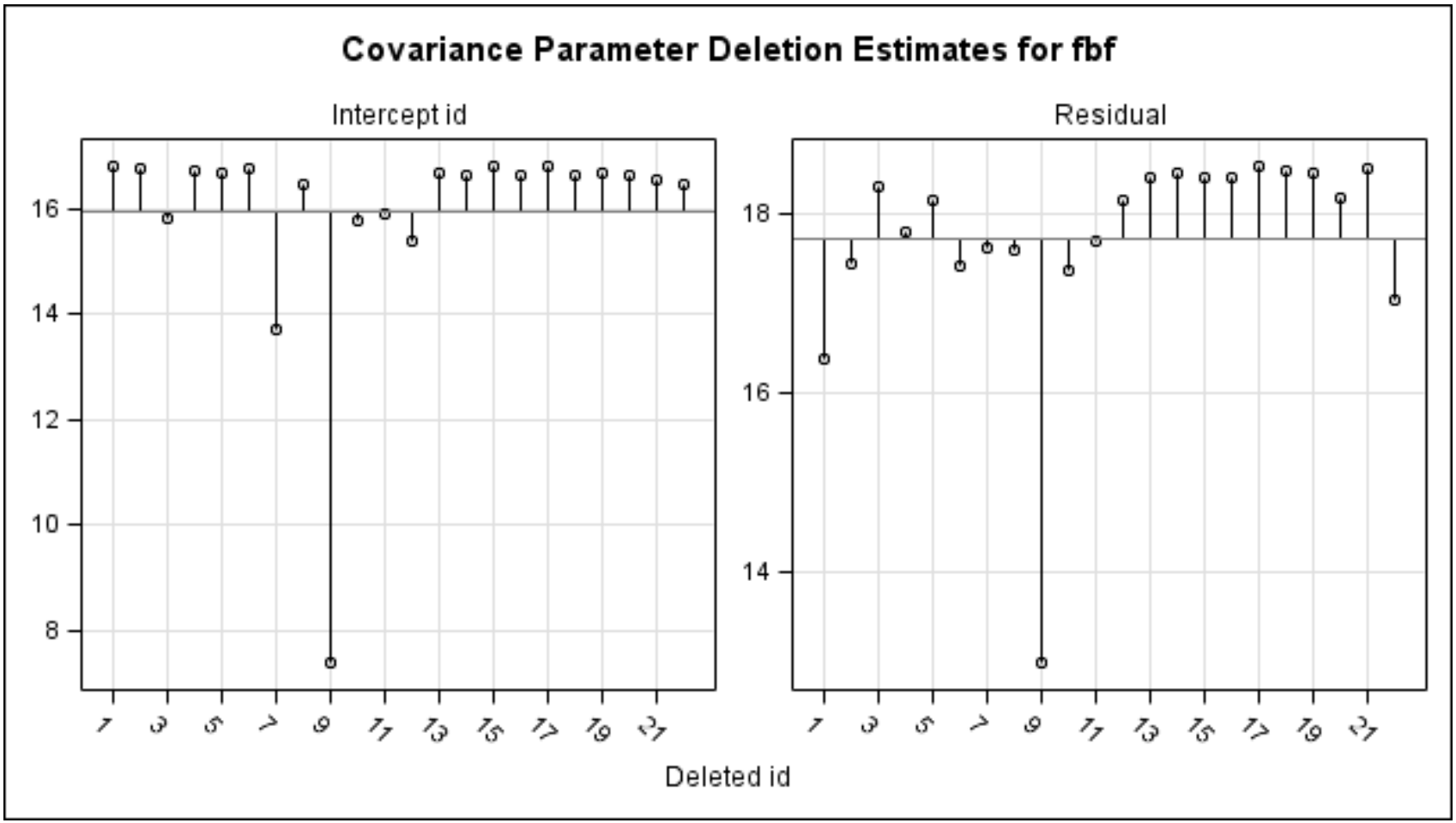
Residual Statistics	
Observations	150
Minimum	-3.018
Mean	-0.002
Maximum	3.4118
Std Dev	1.0039

Fit Statistics	
Objective	910.45
AIC	914.45
AICC	914.54
BIC	916.63

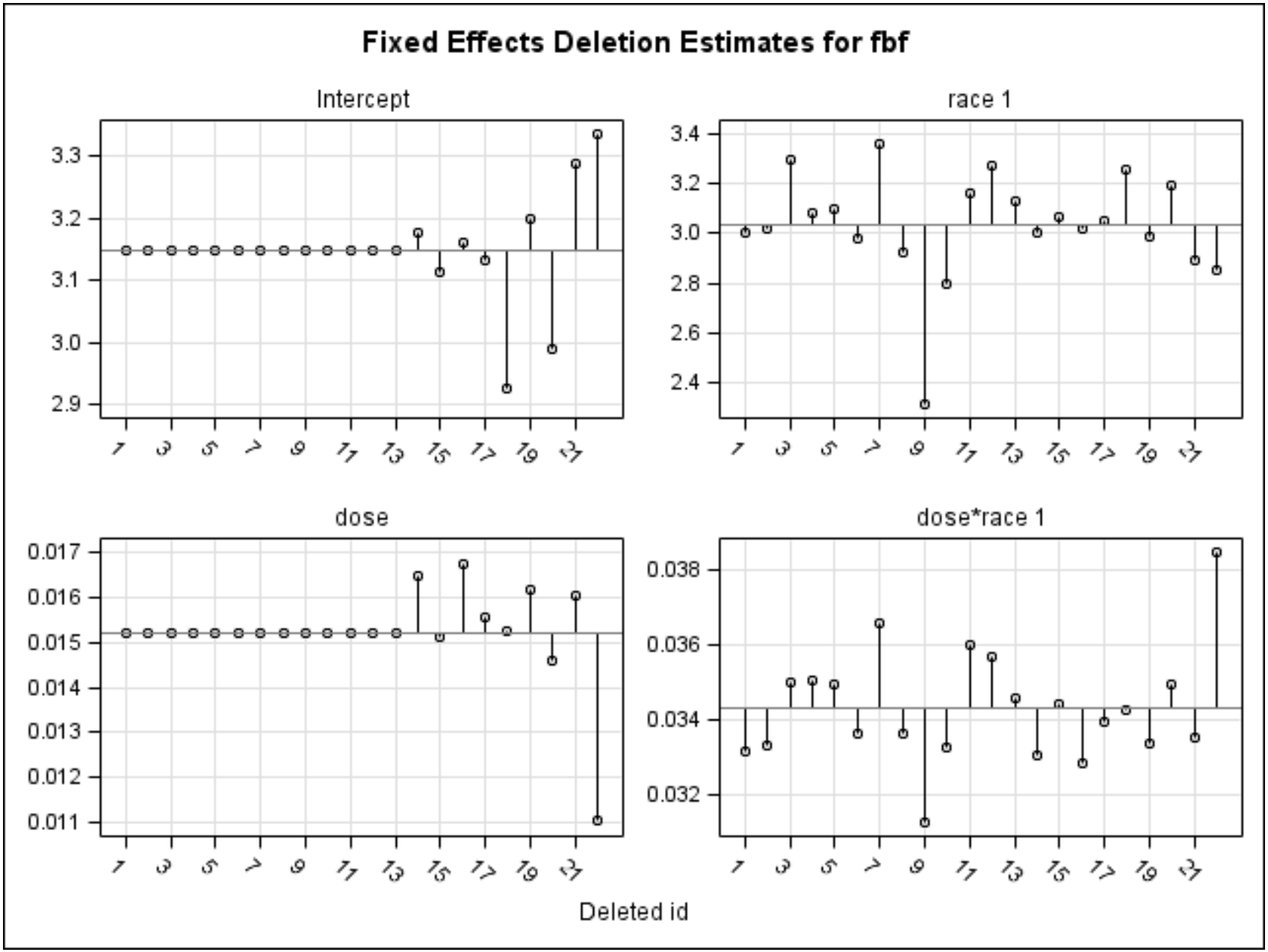
SAS output: Influence



SAS output: Influence



SAS output: Influence



Question

- Why do we use random intercept model rather than creating subject-level indicator variable and estimating fixed effects?

Summary: Key Points

- What is random intercept model?
- What are the fixed and random parts in random intercept model?
- What is ρ and how to estimate?
- What do small and large ρs imply?
- How to interpret parameters in random intercept model?
- How to decide if random intercept is needed (hypothesis testing)?
- What are subject level and observation level residuals?

Suggested Readings

- Chapter 9 (Davidian)