STAT 525

Chapter 25
Random and Mixed Effects Models

Dr. Qifan Song

Single-Factor Studies

Random Effects vs Fixed Effects

- Consider factor with numerous levels
- Want to draw inference **on population of levels**
- Not concerned with any specific levels
- Example of differences
  - **Fixed:** Compare reading ability of 10 2nd grade classes in NY
    - Select $a = 10$ specific classes of interest
    - Randomly choose n students from each classroom
  - **Random:** Compare variability **among all** 2nd grade classes in NY
    - **Randomly choose** $a = 10$ classes from large number of classes
    - Randomly choose n students from each classroom
- Inference broader in random effects case
- Levels chosen randomly $\rightarrow$ inference on population

Data for One-way Random Effects Model

- Y is the response variable
- Factor with levels $i = 1, 2, ..., r$
- Y_{ij} is the j^{th} observation from cell i
- Consider $j = 1, 2, ..., n_i$

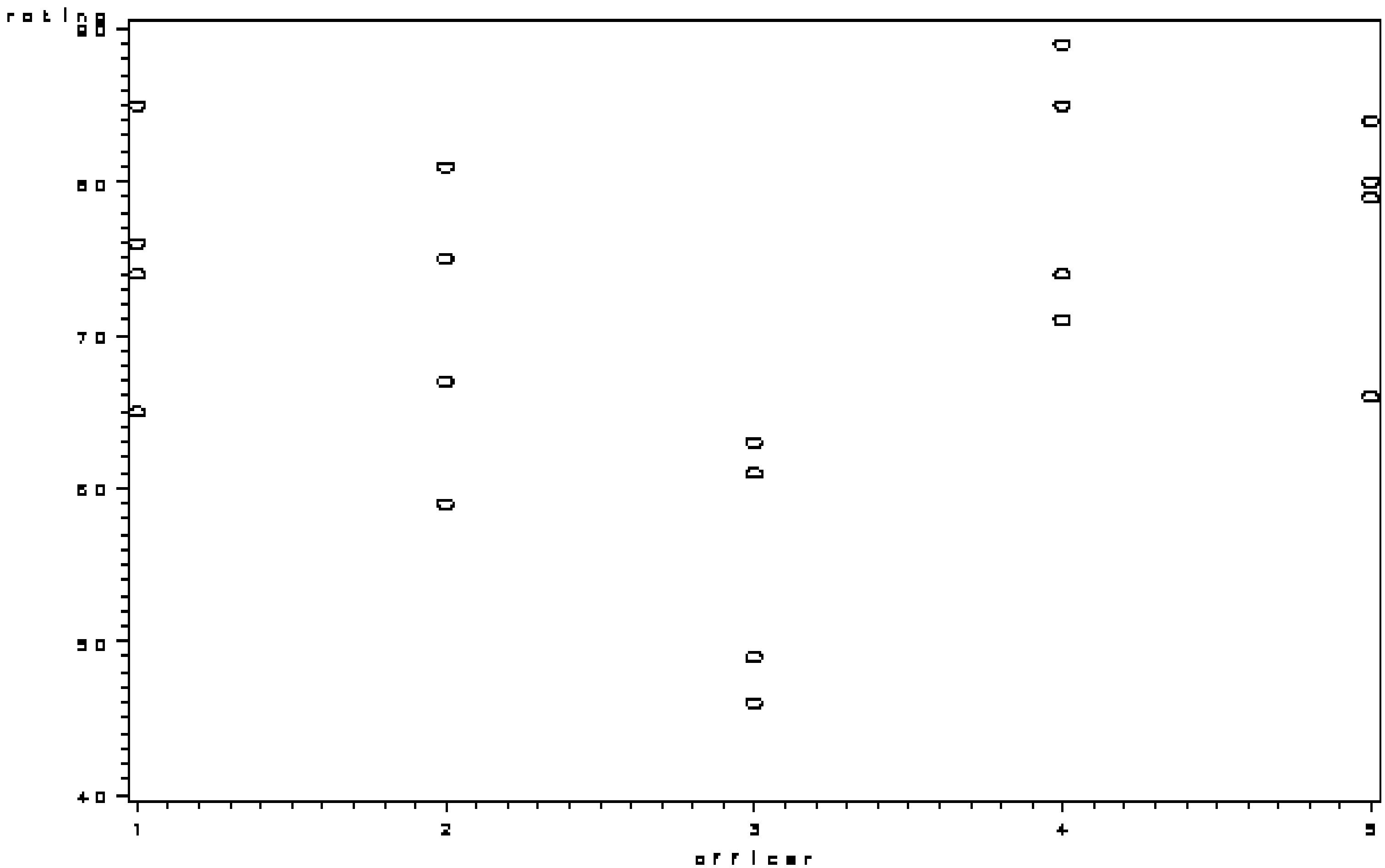
Example (Page 1036)

- Interested in studying the variability in the rating of job applicants
  - Variability among applicants
  - Variability among personnel officers
- Y is the job applicant rating
- Factor: officer/interviewer $(r = 5)$
- Interviewers selected at random from population of personnel officers
- Twenty applicants randomly and equally assigned $(n = 4)$ to officers

```
data a1;
    infile 'u:\.www\datasets525\CH25TA01.txt';
    input rating officer;
proc print data=a1; run; quit;
```

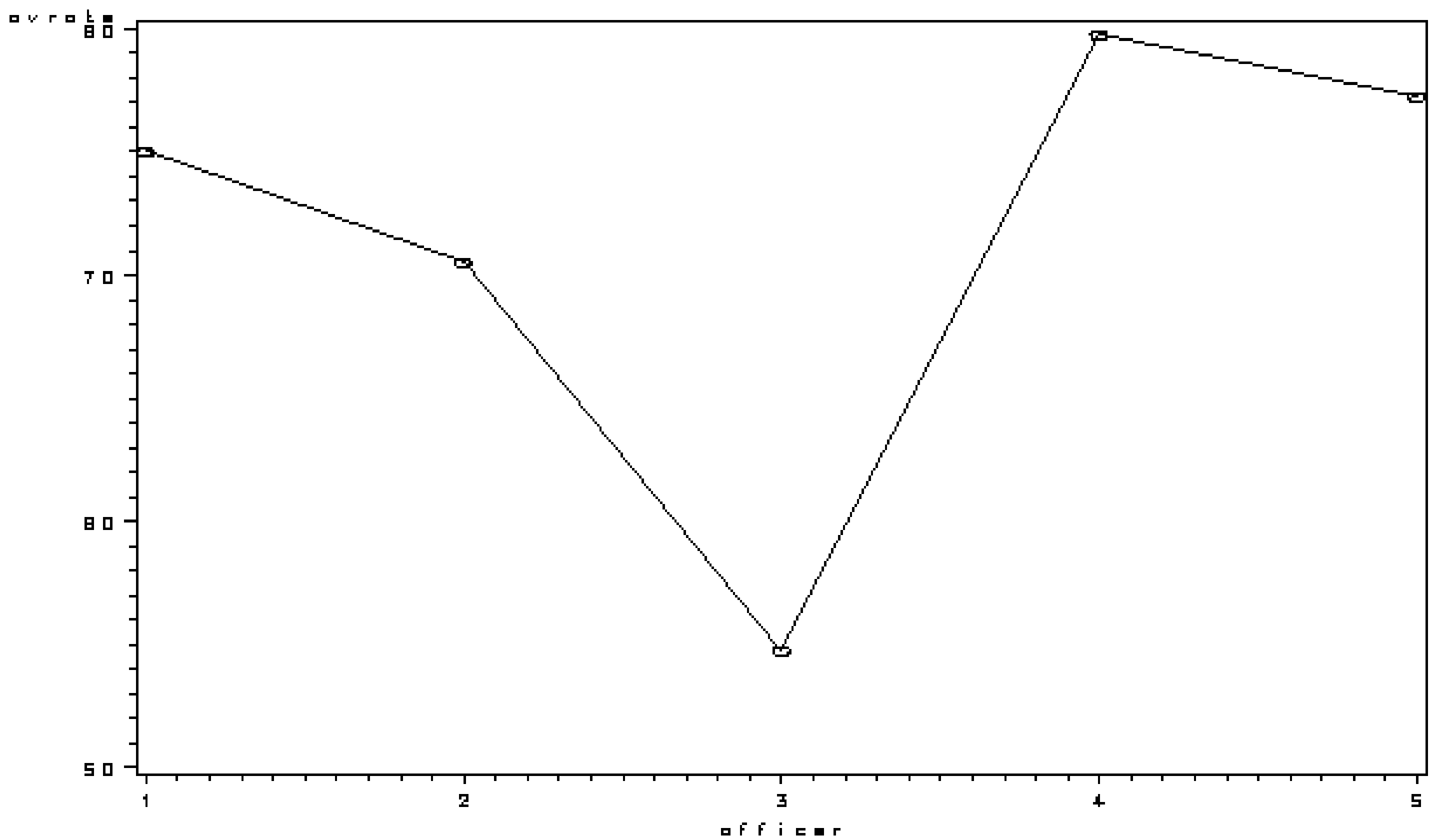
Obs	rating	officer
1	76	1
2	65	1
3	85	1
4	74	1
5	59	2
6	75	2
7	81	2
8	67	2
9	49	3
10	63	3
11	61	3
12	46	3
13	74	4
14	71	4
15	85	4
16	89	4
17	66	5
18	84	5
19	80	5
20	79	5

```
/*-------- Scatterplot --------*/
symbol1 v=circle i=none c=black;
proc gplot data=a1;
    plot rating*officer/frame;
run; quit;
```



```
/*-------- Means Plot --------*/
proc means data=a1;
    var rating;
    by officer;
    output out=a2 mean=avrate;

symbol1 v=circle i=join c=black;
proc gplot data=a2;
    plot avrate*officer/frame;
run; quit;
```



Random Effects Model

- Expressed numerically

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

- $\mu_i \overset{iid}{\sim} N(\mu, \sigma_\mu^2)$
- $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
- μ_i and ε_{ij} independent
- Implies $Y_{ij} \sim N(\mu, \sigma_\mu^2 + \sigma^2)$
- Cov(Y_{ij}, Y_{ik})=σ_μ^2, $j \neq k$
- Also called *Model II*, while the fixed effects model is called *Model I*

Random Factor Effects Model

- Statistical model is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

  - μ - grand mean
  - $\tau_i \stackrel{iid}{\sim} N(0, \sigma_\mu^2)$
  - $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$

- There are THREE parameters (μ, σ_μ and σ^2) in each model
- Cell means are random variables, not parameters

Function of Interest

- Often interested in the quantity

$$\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma^2} = \frac{\sigma_\mu^2}{\sigma_Y^2}$$

  which describes the percentage of TOTAL variability due to the factor.

- Is also called the intraclass correlation coefficient because it describes the correlation between two observations with the same i

$$\rho_{IC} = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ik})}} = \frac{\sigma_\mu^2}{\sigma_Y^2}$$

- May want this value small (as in our example - little variability among officers) or large (measure r items n times each - variability among items dominates the measurement error)

ANOVA Table and EMS

- Terms and layout of ANOVA table are the same as that used in the fixed effects case
- The expected means squares (EMS) are different because of the additional random effects
- This also means the hypotheses being tested are different

Hypothesis Tests

- The hypotheses are:

$$\begin{aligned} H_0 &: \sigma_\mu^2 = 0 \\ H_a &: \sigma_\mu^2 > 0 \end{aligned}$$

- Same break down of Total SS but

  E(MSE)=σ^2

  E(MSR)=$\sigma^2 + n\sigma_\mu^2$

- Under H_0, $F^* = MSR/MSE \sim F_{\alpha,r-1,n_T-r}$
- Same test statistic as before
- Conclusion pertains to entire population of factor effects

Model Estimates

- Interested in estimating variances
- The Type I Method (`METHOD=TYPE1` in SAS)
  - Computes the Type I sum of squares for each effect;
  - Equates each mean square involving only random effects to its expected value;
  - Solves the resulting system of equations.
- For example,
  - $\widehat{\sigma}^2 = \text{MSE}$
  - $\widehat{\sigma}_\mu^2 = (\text{MSR} - \text{MSE})/n$
  - If unbalanced, replace n with

$$n_0 = ((\sum n_i)^2 - \sum n_i^2)/((r-1)\sum n_i)$$

- The Type I Method: Used by `PROC GLM`; available in `PROC VARCOMP` and `PROC MIXED`

- The Type I Method: estimate of σ_μ^2 can be negative
  - Large variability in the data or true H_0?
  - Possible outliers in the data?
  - Inappropriate model?
- Alternative approaches: `MIVQUE0`, `ML`, `REML`
  - `MIVQUE0`: Default method of `PROC VARCOMP`; also available in `PROC MIXED`.
  - `ML`: Available in `PROC VARCOMP` and `PROC MIXED`.
  - `REML`: Default method of `PROC MIXED`; also available in `PROC VARCOMP`.

- `MIVQUE0` (Minimum Variance Quadratic Unbiased Estimation)

  – Produces unbiased estimates that are invariant with respect to the fixed effects of the model and that are locally best quadratic unbiased estimates given that the true ratio of each component to the residual error component is zero;

  – Similar to `TYPE1` except that the random effects are adjusted only for the fixed effects;

  – Its solution is equivalent to the first iterate of REML;

  – Default method of `PROC VARCOMP`; also available in `PROC MIXED`.

- `ML` (Maximum Likelihood)

  – Maximum likelihood estimates of the variance components;

  – Numerically more difficult than ANOVA method but preferred due to well defined estimators (non-negative, well-known large-sample properties, known sampling variances);

  – Available in `PROC VARCOMP` and `PROC MIXED`.

- REML (Restricted Maximum Likelihood Method)

  – Separates the likelihood into two parts: one that contains the fixed effects and one that does not;

  – Simple Example: $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, \cdots, n$

  * ML: $\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$;

  * REML: Maximize the likelihood of $X_i - \bar{X} \sim N(0, \frac{n-1}{n}\sigma^2)$ to obtain $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$;

  * Has taken account of the one degree of freedom required for estimating μ but the ML estimator has not;

  * Unbiased but the ML estimator is not.

  – In the general case of unbalanced data, neither the ML estimators nor the REML estimators are unbiased;

  – Iterated version of MIVQUE0;

  – Default method of `PROC MIXED`; also available in `PROC VARCOMP`.

Example (Page 1036)

```
proc glm data=a1;
    class officer;
    model rating=officer;
    random officer;
run; quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1579.700000	394.925000	5.39	0.0068
Error	15	1099.250000	73.283333		
Corrected Total	19	2678.950000			

R-Square	Coeff Var	Root MSE	rating Mean
0.589671	11.98120	8.560569	71.45000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
officer	4	1579.700000	394.925000	5.39	0.0068

Source	DF	Type III SS	Mean Square	F Value	Pr > F
officer	4	1579.700000	394.925000	5.39	0.0068

Source	Type III Expected Mean Square
officer	Var(Error) + 4 Var(officer)

```
/* VARCOMP: Estimates the contribution of each random effect to
            the variance of the dependent variable              */
/* METHOD: one of TYPE1, MIVQUE0 (default), ML, REML            */
/* TYPE1: compute the Type I sum of squares for each effect     */
proc varcomp data=a1 method=type1;
    class officer;
    model rating=officer;
run; quit;
```

Variance Components Estimation Procedure

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
officer	4	1579.700000	394.925000	Var(Error)+4 Var(officer)
Error	15	1099.250000	73.283333	Var(Error)
Corrected Total	19	2678.950000	.	.

Type 1 Estimates

Variance Component	Estimate
Var(officer)	80.41042
Var(Error)	73.28333

```
/* REML: restricted maximum likelihood  */
proc mixed data=a1 cl;
    class officer;
    model rating=;
    random officer/vcorr;
run; quit;
```

Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Estimated V Correlation Matrix for Subject 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.5232	0.5232	0.5232
2	0.5232	1.0000	0.5232	0.5232
3	0.5232	0.5232	1.0000	0.5232
4	0.5232	0.5232	0.5232	1.0000

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
officer	80.4104	0.05	24.4572	1498.97
Residual	73.2833	0.05	39.9896	175.54

Confidence intervals

- σ^2: Page 1041

$$\frac{r(n-1)\text{MSE}}{\sigma^2} \sim \chi^2_{r(n-1)}$$

$$\frac{r(n-1)\text{MSE}}{\chi^2_{1-\alpha/2;r(n-1)}} \leq \sigma^2 \leq \frac{r(n-1)\text{MSE}}{\chi^2_{\alpha/2;r(n-1)}}$$

- σ^2_μ: Page 1043

$$\frac{(r-1)\text{MSR}}{\sigma^2 + n\sigma^2_\mu} \sim \chi^2_{r-1}$$

so

$$\hat{\sigma}^2_\mu \sim \frac{\sigma^2 + n\sigma^2_\tau}{n(r-1)}\chi^2_{r-1} - \frac{\sigma^2}{nr(n-1)}\chi^2_{r(n-1)}$$

No closed form expression for this distribution

Satterthwaite Procedure page 1043 (`PROC MIXED`)

Modified Large Sample (MLS) Procedure page 1045

- Satterthwaite Procedure for an approximate CI of Linear Combinations of EMS
  - Consider $L = c_1 \times E[MS_1] + \cdots + c_h \times E[MS_h]$
  - An unbiased estimator of L is $\hat{L} = c_1 \times MS_1 + \cdots + c_h \times MS_h$
  - Satterthwaite suggested

$$\frac{df(\hat{L}) \times \hat{L}}{L} \sim \chi^2_{df(\hat{L})}, \quad approximately$$

$$df(\hat{L}) = \frac{(\hat{L})^2}{\sum_{k=1}^{h}(c_k \times MS_k)^2/df(MS_k)}$$

  - The above approximate distribution suggests a CI of L,

$$\frac{df(\hat{L}) \times \hat{L}}{\chi^2(1-\alpha/2; df(\hat{L}))} \leq L \leq \frac{df(\hat{L}) \times \hat{L}}{\chi^2(\alpha/2; df(\hat{L}))}$$

  - For $\sigma^2_\mu = L = \frac{1}{n}E[MSR] - \frac{1}{n}E[MSE]$,

$$\hat{L} = MSR/n - MSE/n,$$

$$df(\hat{L}) = \frac{(MSR - MSE)^2}{\frac{(MSR)^2}{r-1} + \frac{(MSE)^2}{r(n-1)}}$$

- Modified Large Sample (MLS) Procedure
  - Satterthwaite procedure is general and easy to carry out but its accuracy can be quite limited when L has both negative and positive c_k;
  - MLS: improved procedure but computationally complex
  - Only for $L = c_1 \times E[MS_1] + c_2 \times E[MS_2]$, $c_1 > 0, c_2 < 0$ in a balanced study
  - An unbiased estimator of L is $\hat{L} = c_1 \times MS_1 + c_2 \times MS_2$
  - An approximate CI of L: $\hat{L} - H_L \leq L \leq \hat{L} + H_U$

$$\begin{aligned} F_1 &= F(1-\alpha/2; df(MS_1), \infty), & F_2 &= F(1-\alpha/2; df(MS_2), \infty) \\ F_3 &= F(1-\alpha/2; \infty, df(MS_1)), & F_4 &= F(1-\alpha/2; \infty, df(MS_2)) \\ F_5 &= F(1-\alpha/2; df(MS_1), df(MS_2)), & F_6 &= F(1-\alpha/2; df(MS_2), df(MS_1)) \\ G_1 &= 1 - 1/F_1, & G_2 &= 1 - 1/F_2 \\ G_3 &= \frac{(F_5-1)^2 - (G_1F_5)^2 - (F_4-1)^2}{F_5}, & G_4 &= F_6\{(\frac{F_6-1}{F_6})^2 - (\frac{F_3-1}{F_6})^2 - G_2^2\} \end{aligned}$$

$$H_L = \sqrt{\{G_1c_1MS_1\}^2 + \{(F_4-1)c_2MS_2\}^2 - G_3c_1c_2MS_1MS_2}$$

$$H_U = \sqrt{\{(F_3-1)c_1MS_1\}^2 + \{G_2c_2MS_2\}^2 - G_4c_1c_2MS_1MS_2}$$

- Intraclass Correlation Coefficient : Page 1040
  - Uses ratio of two χ^2 distributions (i.e., F dist)

$$\frac{\mathsf{MS_{Trt}}}{n\sigma_\mu^2+\sigma^2}\Big/\frac{\mathsf{MS_E}}{\sigma^2} \sim F_{a-1,N-a}$$

$$\Rightarrow \quad \frac{\mathsf{L}}{\mathsf{L}+1} \le \frac{\sigma_\mu^2}{\sigma^2+\sigma_\mu^2} \le \frac{\mathsf{U}}{\mathsf{U}+1}$$

$$\mathsf{L} = \frac{1}{n}\left(\frac{\mathsf{MS_{Trt}}}{\mathsf{MS_E}\ F_{1-\alpha/2;a-1,N-a}} - 1\right), \mathsf{U} = \frac{1}{n}\left(\frac{\mathsf{MS_{Trt}}}{\mathsf{MS_E}\ F_{\alpha/2;a-1,N-a}} - 1\right)$$

- Grand mean μ : Page 1038-1039

$$\overline{Y}_{..} = \frac{1}{r}(\overline{Y}_{1.} + \overline{Y}_{2.} + ... + \overline{Y}_{r.})$$

$$\overline{Y}_{i.} \sim N\left(\mu, \sigma_\mu^2 + \frac{\sigma^2}{n}\right)$$

Two-Factor Studies

Data for Two-Way Design

- Y is the response variable
- Factor A has levels $i = 1, 2, ..., a$
- Factor B has levels $j = 1, 2, ..., b$
- Y_{ijk} is the k^{th} observation from cell (i, j) with $k = 1, 2, ..., n_{ij}$
- Balanced when $n_{ij} = n$

Example (Page 1080)

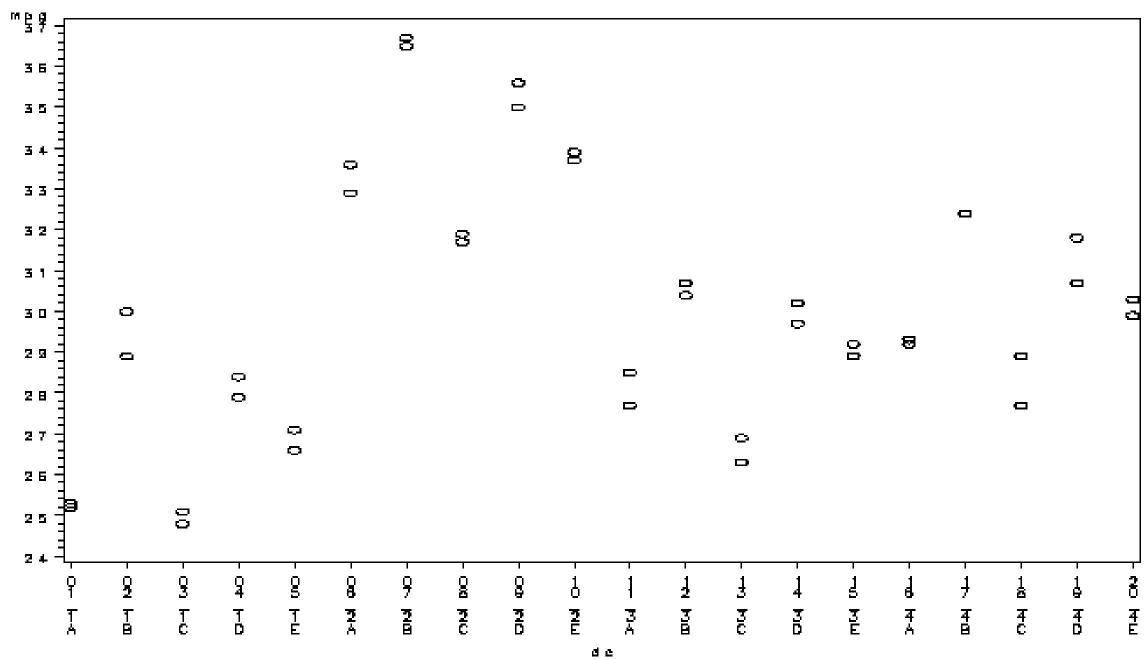
- Interested in the fuel efficiency (mpg)
- Two **random** factors
  - Factor A: Driver
  - Factor B: Car (same model)
- How much of the overall variability is due to driver and/or car?
- Each driver drove each car twice ($n = 2$) over same 40 mile course

```
data a1;
    infile 'u:\.www\datasets525\CH25PR15.txt';
    input mpg driver car;
proc print data=a1; run; quit;
```

```
Obs     mpg    driver    car
  1    25.3       1       1
  2    25.2       1       1
  3    28.9       1       2
  4    30.0       1       2
  5    24.8       1       3
  6    25.1       1       3
  7    28.4       1       4
  8    27.9       1       4
  9    27.1       1       5
 10    26.6       1       5
 11    33.6       2       1
 12    32.9       2       1
 13    36.7       2       2
 14    36.5       2       2
 15    31.7       2       3
 16    31.9       2       3
 17    35.6       2       4
 18    35.0       2       4
 19    33.7       2       5
 20    33.9       2       5
 21    27.7       3       1
 22    28.5       3       1
  .       .       .       .
  .       .       .       .
 37    31.8       4       4
 38    30.7       4       4
 39    30.3       4       5
 40    29.9       4       5
```

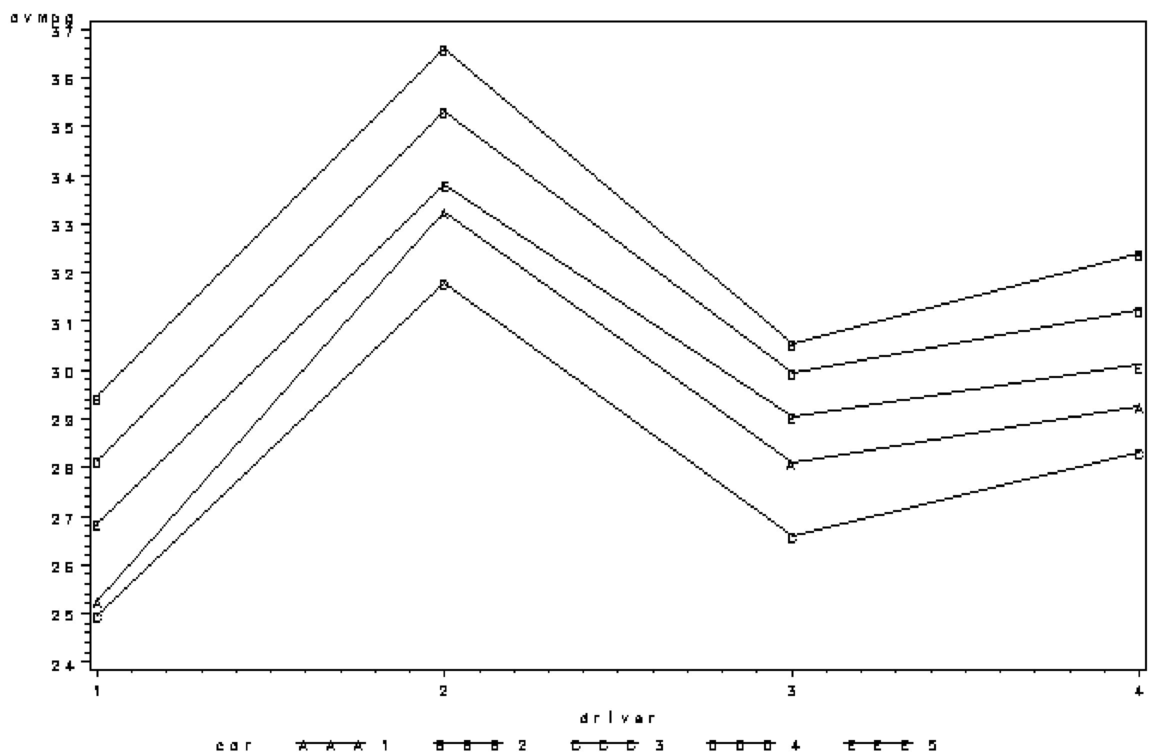
```
data a1; set a1;
   if (driver eq 1)*(car eq 1) then dc='01_1A'; if (driver eq 1)*(car eq 2) then dc='02_1B';
   if (driver eq 1)*(car eq 3) then dc='03_1C'; if (driver eq 1)*(car eq 4) then dc='04_1D';
   if (driver eq 1)*(car eq 5) then dc='05_1E'; if (driver eq 2)*(car eq 1) then dc='06_2A';
   if (driver eq 2)*(car eq 2) then dc='07_2B'; if (driver eq 2)*(car eq 3) then dc='08_2C';
   if (driver eq 2)*(car eq 4) then dc='09_2D'; if (driver eq 2)*(car eq 5) then dc='10_2E';
   if (driver eq 3)*(car eq 1) then dc='11_3A'; if (driver eq 3)*(car eq 2) then dc='12_3B';
   if (driver eq 3)*(car eq 3) then dc='13_3C'; if (driver eq 3)*(car eq 4) then dc='14_3D';
   if (driver eq 3)*(car eq 5) then dc='15_3E'; if (driver eq 4)*(car eq 1) then dc='16_4A';
   if (driver eq 4)*(car eq 2) then dc='17_4B'; if (driver eq 4)*(car eq 3) then dc='18_4C';
   if (driver eq 4)*(car eq 4) then dc='19_4D'; if (driver eq 4)*(car eq 5) then dc='20_4E';

/*------- Scatterplot -------*/
symbol v=circle i= c=black;
proc gplot data=a1;
    plot mpg*dc/frame;
run; quit;
```



```
proc means data=a1;
    output out=a2 mean=avmpg;
    var mpg;
    by driver car;

/*------ Means Plot ------*/
symbol1 v='A' i=join c=black;
symbol2 v='B' i=join c=black; symbol3 v='C' i=join c=black;
symbol4 v='D' i=join c=black; symbol5 v='E' i=join c=black;
proc gplot data=a2;
   plot avmpg*driver=car/frame;
run; quit;
```



Random Effects Model

- Expressed numerically

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

  - $\mu_{ij} \sim N(\mu, \sigma_\mu^2)$
  - $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$
  - μ_{ik} and ε_{ijk} independent
- Not all observations independent
- Will separate means into factor variances

Random Factor Effects Model

- Statistical model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$$

  - μ - grand mean
  - $\alpha_i \stackrel{iid}{\sim} N(0, \sigma^2_\alpha)$
  - $\beta_j \stackrel{iid}{\sim} N(0, \sigma^2_\beta)$
  - $(\alpha\beta)_{ij} \stackrel{iid}{\sim} N(0, \sigma^2_{\alpha\beta})$
  - $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$
  - $\{\alpha_j\}$,$\{\beta_j\}$, $\{(\alpha\beta)_{ij}\}$ and $\{\varepsilon_{ijk}\}$ are pairwise independent

- Decomposition $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
- Decomposition $\sigma^2_\mu = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta}$
- There are FIVE parameters (μ, σ^2_α, σ^2_β, $\sigma^2_{\alpha\beta}$ and σ^2)

Covariance Structure

- Covariances:

$$\begin{aligned}
Cov(Y_{ijk}, Y_{ijk}) &= \sigma^2 + \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta} \\
Cov(Y_{ijk}, Y_{ijk^*}) &= \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_{\alpha\beta}, \quad k \neq k^* \\
Cov(Y_{ijk}, Y_{ij^*k^*}) &= \sigma^2_\alpha, \quad j \neq j^* \\
Cov(Y_{ijk}, Y_{i^*jk^*}) &= \sigma^2_\beta, \quad i \neq i^* \\
Cov(Y_{ijk}, Y_{i^*j^*k^*}) &= 0, \quad i \neq i^*, j \neq j^*
\end{aligned}$$

- Can again look at percentage of variability due to various factors / interaction
- Could look at percentage of total variability or percentage of cell means variability (i.e., ignoring error variance).
- Approach to confidence intervals similar to intraclass correlation coefficient

ANOVA Table

- Terms and layout of ANOVA table the same as that used in the fixed effects case
- The expected means squares (EMS) are different because of the additional random effects
- Results in different F tests
- Use EMS as guide for tests $\longrightarrow$ determine denominator MS

Expected Mean Squares $\Longrightarrow$ Model Estimates

- Same partition of Total Sum of Squares
- Assuming balanced design
  - E(MSE)=$\sigma^2 \Longrightarrow \hat{\sigma}^2 = \text{MSE}$
  - E(MSAB)=$\sigma^2 + n\sigma^2_{\alpha\beta} \Longrightarrow \hat{\sigma}^2_{\alpha\beta} = (\text{MSAB} - \text{MSE})/n$
  - E(MSA)=$\sigma^2 + n\sigma^2_{\alpha\beta} + bn\sigma^2_{\alpha} \Longrightarrow \hat{\sigma}^2_{\alpha} = (\text{MSA} - \text{MSAB})/(bn)$
  - E(MSB)=$\sigma^2 + n\sigma^2_{\alpha\beta} + an\sigma^2_{\beta} \Longrightarrow \hat{\sigma}^2_{\beta} = (\text{MSB} - \text{MSAB})/(an)$
- Estimates can be negative: adjustments may be used
- Alternative estimates of variances: `MIVQUE0`, `ML`, `REML`, etc.

Hypothesis Tests

- Three tests of variance
  - H_{0A}: $\sigma^2_\alpha = 0$ vs H_{1A}: $\sigma^2_\alpha > 0$

    $F^* = MSA/MSAB$
  - H_{0B}: $\sigma^2_\beta = 0$ vs H_{1B}: $\sigma^2_\beta > 0$

    $F^* = MSB/MSAB$
  - H_{0AB}: $\sigma^2_{\alpha\beta} = 0$ vs H_{1AB}: $\sigma^2_{\alpha\beta} > 0$

    $F^* = MSAB/MSE$
- **No hierarchy** in terms of testing
- Not all tests use MSE in denominator
  - To test σ^2_α or σ^2_β use MSAB
  - Will alter denominator DF too

Example (Page 1080)

```
proc glm data=a1;
    class driver car;
    model mpg=driver car driver*car;
    random driver car driver*car/test;
run; quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	377.4447500	19.8655132	113.03	<.0001
Error	20	3.5150000	0.1757500		
Corrected Total	39	380.9597500			

R-Square	Coeff Var	Root MSE	mpg Mean
0.990773	1.395209	0.419225	30.04750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
driver	3	280.2847500	93.4282500	531.60	<.0001
car	4	94.7135000	23.6783750	134.73	<.0001
driver*car	12	2.4465000	0.2038750	1.16	0.3715

Source	DF	Type III SS	Mean Square	F Value	Pr > F
driver	3	280.2847500	93.4282500	531.60	<.0001
car	4	94.7135000	23.6783750	134.73	<.0001
driver*car	12	2.4465000	0.2038750	1.16	0.3715

Source	Type III Expected Mean Square
driver	Var(Error) + 2 Var(driver*car) + 10 Var(driver)
car	Var(Error) + 2 Var(driver*car) + 8 Var(car)
driver*car	Var(Error) + 2 Var(driver*car)

- PROC GLM assumes all factors are fixed effects
- With RANDOM statement and TEST option, PROC GLM will perform correct tests

Source	DF	Type III SS	Mean Square	F Value	Pr > F
driver	3	280.284750	93.428250	458.26	<.0001
car	4	94.713500	23.678375	116.14	<.0001
Error	12	2.446500	0.203875		

Error: MS(driver*car)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
driver*car	12	2.446500	0.203875	1.16	0.3715
Error: MS(Error)	20	3.515000	0.175750		

```
proc mixed data=a1 cl;
    class car driver;
    model mpg=;
    random car driver car*driver/vcorr;
run; quit;
```

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
car	2.9343	0.05	1.0464	24.9038
driver	9.3224	0.05	2.9864	130.79
car*driver	0.01406	0.05	0.001345	3.592E17
Residual	0.1757	0.05	0.1029	0.3665

- `Estimated V Correlation Matrix` has the following entries
  - Same observation: 1.0000
  - Same i and j: 0.9859
  - Same j only: 0.7490
  - Same i only: 0.2358

Two-Factor Mixed Effects Model

- One factor random and one factor fixed (aka Model III)
- Assume A fixed and B random
- Mixed Factor Effects Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

  - $\sum_i \alpha_i = 0$ and $\beta_j \overset{iid}{\sim} \text{N}(0, \sigma^2_\beta)$
  - $(\alpha\beta)_{ij} \sim \text{N}(0, (a-1)\sigma^2_{\alpha\beta}/a)$ subject to the restrictions
    * $\sum_i (\alpha\beta)_{ij} = 0$ for each j
    * $\text{Cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j}) = -\frac{1}{a}\sigma^2_{\alpha\beta}$
  - $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$
  - $\{\beta_j\}$, $\{(\alpha\beta)_{ij}\}$ and $\{\varepsilon_{ijk}\}$ are pairwise independent
- Known as **restricted** mixed effects model

- The $(a-1)/a$ simplifies the EMS
  - E(MSE)=σ^2
  - E(MSA) = $\sigma^2 + bn\sum_i \alpha_i^2/(a-1) + n\sigma^2_{\alpha\beta}$
  - E(MSB) = $\sigma^2 + an\sigma^2_{\beta}$
  - E(MSAB) = $\sigma^2 + n\sigma^2_{\alpha\beta}$
- Not all $(\alpha\beta)_{ij}$ are independent

$$\text{Cov}((\alpha\beta)_{ij}, (\alpha\beta)_{i'j}) = -\frac{1}{a}\sigma^2_{\alpha\beta}, \quad i \neq i'$$

  - If $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ then

$$X_i - \overline{X} \sim N(0, \frac{n-1}{n}\sigma^2)$$
$$\text{Cov}(X_i - \overline{X}, X_j - \overline{X}) = -\frac{1}{n}\sigma^2$$

Hypothesis Tests

- Tests require different MS in denom

$$H_0 : \alpha_1 = \alpha_2 = \ldots = 0 \rightarrow \text{MSA/MSAB}$$

$$H_0 : \sigma^2_\beta = 0 \rightarrow \text{MSB/MSE}$$

$$H_0 : \sigma^2_{\alpha\beta} = 0 \rightarrow \text{MSAB/MSE}$$

- Variance Estimates (Using ANOVA method)

$$\hat{\sigma}^2 = \text{MSE}$$

$$\hat{\sigma}^2_\beta = (\text{MSB} - \text{MSE})/(an)$$

$$\hat{\sigma}^2_{\alpha\beta} = (\text{MSAB} - \text{MSE})/n$$

Multiple Comparisons

- Compare the fixed effects:

$$
\begin{aligned}
Y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \\
\overline{Y}_{i..} &= \mu + \alpha_i + \overline{\beta}_{.} + \overline{(\alpha\beta)}_{i.} + \overline{\varepsilon}_{i..} \\
Var(\overline{Y}_{i..}) &= \sigma^2_{\beta}/b + (a-1)\sigma^2_{\alpha\beta}/(ab) + \sigma^2/(bn)
\end{aligned}
$$

$$
\begin{aligned}
\overline{Y}_{i..} - \overline{Y}_{i'..} &= \alpha_i - \alpha_{i'} + \overline{(\alpha\beta)}_{i.} - \overline{(\alpha\beta)}_{i'.} + \overline{\varepsilon}_{i..} - \overline{\varepsilon}_{i'..} \\
Var(\overline{Y}_{i..} - \overline{Y}_{i'..}) &= 2\sigma^2_{\alpha\beta}/b + 2\sigma^2/(bn) = 2(n\sigma^2_{\alpha\beta} + \sigma^2)/(bn)
\end{aligned}
$$

- For pairwise comparison, use 2MSAB$/(bn)$
  - See (25.59) on p. 1057 for a general contrast
- Note: To construct CIs for marginal means, e.g., $\mu_{i\cdot}$, an approximate method has to be employed for the degrees of freedom. See p. 1059.

Other Two-Way Mixed Model

- SAS uses different mixed model in analysis
- Reduce the restrictions on $(\alpha\beta)_{ij}$
  - $\sum_i \alpha_i = 0$ and $\beta_j \stackrel{iid}{\sim} \mathrm{N}(0, \sigma^2_\beta)$
  - $(\alpha\beta)_{ij} \sim \mathrm{N}(0, \sigma^2_{\alpha\beta})$
  - $\varepsilon_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$
  - $\{\beta_j\}$, $\{(\alpha\beta)_{ij}\}$ and $\{\varepsilon_{ijk}\}$ are pairwise independent
- Known as **unrestricted** mixed model

Unrestricted Mixed Model

- Reduced restrictions alter EMS
  - E(MSE)=σ^2
  - E(MSA) $= \sigma^2 + bn\sum \alpha_i^2/(a-1) + n\sigma^2_{\alpha\beta}$
  - E(MSB) $= \sigma^2 + an\sigma^2_\beta + n\sigma^2_{\alpha\beta}$
  - E(MSAB) $= \sigma^2 + n\sigma^2_{\alpha\beta}$

- Differences
  - Test $H_0 : \sigma^2_\beta = 0$ using MSAB (Note: MSE in Restricted Models)
  - Often more conservative test
  - $\hat{\sigma}^2_\beta =$ (MSB − MSAB)$/(an)$
  - Var($\overline{Y}_{i..}$) $= (n\sigma^2_\beta + n\sigma^2_{\alpha\beta} + \sigma^2)/(bn)$

To decide which model is appropriate, suppose you ran experiment again and sampled some of the same levels of the random effect. Does this mean that the interaction effects for these levels are the same as before? Yes: Restricted No: Unrestricted

Random Effect Model in matrix form

- In terms of linear model

$$Y = X\beta + Z\delta + \varepsilon$$

  - β is a vector of fixed-effect parameters
  - δ is a vector of random-effect parameters
  - ε is the error vector

- δ and ε assumed uncorrelated
- $\delta \sim (0, G(\theta))$
- $\varepsilon \sim (0, \sigma^2 I)$
- Cov$(Y) = ZGZ' + \sigma^2 I$, $E(Y) = X\beta$.
- Parametes: β, θ, σ^2.

Example

- Taken from Montgomery - Design and Analysis of Experiments (§12.3)
- Want to assess variability in a measurement system
- Twenty parts selected from production process
- Gauge used by 3 operators to measure parts
- Will consider operators fixed
- Will investigate both restricted and unrestricted results
- SAS considers unrestricted mixed model

```
options nocenter ls=75;
data randr;
    input part operator resp @@;
    cards;
1 1 21 1 1 20 1 2 20 1 2 20 1 3 19 1 3 21
2 1 24 2 1 23 2 2 24 2 2 24 2 3 23 2 3 24
3 1 20 3 1 21 3 2 19 3 2 21 3 3 20 3 3 22
4 1 27 4 1 27 4 2 28 4 2 26 4 3 27 4 3 28
5 1 19 5 1 18 5 2 19 5 2 18 5 3 18 5 3 21
6 1 23 6 1 21 6 2 24 6 2 21 6 3 23 6 3 22
7 1 22 7 1 21 7 2 22 7 2 24 7 3 22 7 3 20
8 1 19 8 1 17 8 2 18 8 2 20 8 3 19 8 3 18
9 1 24 9 1 23 9 2 25 9 2 23 9 3 24 9 3 24
10 1 25 10 1 23 10 2 26 10 2 25 10 3 24 10 3 25
11 1 21 11 1 20 11 2 20 11 2 20 11 3 21 11 3 20
12 1 18 12 1 19 12 2 17 12 2 19 12 3 18 12 3 19
13 1 23 13 1 25 13 2 25 13 2 25 13 3 25 13 3 25
14 1 24 14 1 24 14 2 23 14 2 25 14 3 24 14 3 25
15 1 29 15 1 30 15 2 30 15 2 28 15 3 31 15 3 30
16 1 26 16 1 26 16 2 25 16 2 26 16 3 25 16 3 27
17 1 20 17 1 20 17 2 19 17 2 20 17 3 20 17 3 20
18 1 19 18 1 21 18 2 19 18 2 19 18 3 21 18 3 23
19 1 25 19 1 26 19 2 25 19 2 24 19 3 25 19 3 25
20 1 19 20 1 19 20 2 18 20 2 17 20 3 19 20 3 17
;
```

```
/* E: specifies the error mean square used in the multiple comparisons,
      the overall MSE is used by default                                */
proc glm data=randr;
    class operator part;
    model resp=operator|part;
    random part operator*part / test;
    means operator / tukey lines E=operator*part;
    lsmeans operator / adjust=tukey E=operator*part tdiff stderr;
run; quit;
```

```
Dependent Variable: resp
                                      Sum of
Source                     DF        Squares   Mean Square  F Value  Pr > F
Model                      59    1215.091667     20.594774    20.77  <.0001
Error                      60      59.500000      0.991667
Corrected Total           119    1274.591667

Source                     DF    Type III SS   Mean Square  F Value  Pr > F
operator                    2       2.616667      1.308333     1.32  0.2750
part                       19    1185.425000     62.390789    62.92  <.0001
operator*part              38      27.050000      0.711842     0.72  0.8614

Source                   Type III Expected Mean Square
operator                 Var(Error) + 2 Var(operator*part) + Q(operator)
part                     Var(Error) + 2 Var(operator*part) + 6 Var(part)
operator*part            Var(Error) + 2 Var(operator*part)
```

Tests of Hypotheses for Mixed Model Analysis of Variance

Source	DF	Type III SS	Mean Square	F Value	Pr > F
operator	2	2.616667	1.308333	1.84	0.1730
part	19	1185.425000	62.390789	87.65	<.0001
Error	38	27.050000	0.711842		

Error: MS(operator*part)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
operator*part	38	27.050000	0.711842	0.72	0.8614
Error: MS(Error)	60	59.500000	0.991667		

Tukey's Studentized Range (HSD) Test for resp

Alpha	0.05
Error Degrees of Freedom	38
Error Mean Square	0.711842
Critical Value of Studentized Range	3.44902
Minimum Significant Difference	0.4601

Means with the same letter are not significantly different.

	Mean	N	operator
A	22.6000	40	3
A	22.3000	40	1
A	22.2750	40	2

```
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

Standard Errors and Probabilities Calculated Using the Type III MS
for operator*part as an Error Term

                                Standard                    LSMEAN
operator     resp LSMEAN           Error    Pr > |t|        Number
1             22.3000000       0.1334018      <.0001             1
2             22.2750000       0.1334018      <.0001             2
3             22.6000000       0.1334018      <.0001             3

   Least Squares Means for Effect operator
   t for H0: LSMean(i)=LSMean(j) / Pr > |t|
          Dependent Variable: resp
i/j              1              2              3
   1                     0.132514       -1.59017
                           0.9904         0.2622
   2      -0.13251                      -1.72269
            0.9904                        0.2100
   3      1.590173       1.722688
            0.2622         0.2100
```

NOTE: SE($\overline{Y}_{i.}$) are incorrect. Should be $\sqrt{(\sigma^2 + n\sigma^2_{\tau\beta} + n\sigma^2_{\beta})/(bn)}$, which can be estimated by $\sqrt{((a-1)MSAB + MSB)/(abn)} = .7292$ (note that $a = 3$, $b = 20$ and $n = 2$). Degrees of freedom would have to be approximated for constructing CIs. Similary, for the restricted mixed model the SE would be that on page 25-35.

```
/* DDFM: specifies the method for computing the denominator df for
         testing fixed effects, KR by Kenward & Roger (1997) */
proc mixed data=randr alpha=.05 cl covtest;
    class operator part;
    model resp=operator / ddfm=kr;
    random part operator*part;
    lsmeans operator / alpha=.05 cl diff adjust=tukey;
run; quit;
```

```
Covariance Structure         Variance Components
Estimation Method            REML
Degrees of Freedom Method    Kenward-Roger
```

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
part	10.2513	3.3738	3.04	0.0012	0.05	5.8888	22.1549
operator*part	0	.	.	.	.	.	.
Residual	0.8832	0.1262	7.00	<.0001	0.05	0.6800	1.1938

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
operator	2	98	1.48	0.2324

- NOTE (Unrestricted Model)
  - $\text{SE}(\overline{y}_{1..}) = \sqrt{(\sigma^2 + n\sigma^2_{\tau\beta} + n\sigma^2_{\beta})/(bn)}$ can be estimated as $\sqrt{(0.8832 + 2(10.2513))/40} =$ 0.7312
    - This estimate is slightly different than GLM because of the zero variance estimate
  - $\text{SE}(\overline{y}_{1..} - \overline{y}_{2..}) = \sqrt{2(\sigma^2 + n\sigma^2_{\tau\beta})/(bn)}$ can be estimated as $\sqrt{0.8832/20} = 0.2101$.

Differences of Least Squares Means

Effect	operator	_operator	Estimate	Standard Error	DF	t Value	Pr > |t|
operator	1	2	0.02500	0.2101	98	0.12	0.9055
operator	1	3	-0.3000	0.2101	98	-1.43	0.1566
operator	2	3	-0.3250	0.2101	98	-1.55	0.1252

Differences of Least Squares Means

Effect	operator	_operator	Adjustment	Adj P	Alpha
operator	1	2	Tukey-Kramer	0.9922	0.05
operator	1	3	Tukey-Kramer	0.3371	0.05
operator	2	3	Tukey-Kramer	0.2811	0.05

Differences of Least Squares Means

Effect	operator	_operator	Lower	Upper	Adj Lower	Adj Upper
operator	1	2	-0.3920	0.4420	-0.4751	0.5251
operator	1	3	-0.7170	0.1170	-0.8001	0.2001
operator	2	3	-0.7420	0.09201	-0.8251	0.1751

- Different p-values/CIs from PROC GLM due to the use of KR

Multi-Factor Models

- 3-Way: Can have 0,1,2, or 3 random effects
- Use EMS to determine tests
- In some cases, no straightforward test exists. In other words, there is no single MS for the denominator/numerator
- Must perform approximate F test
  - `PROC GLM` uses Satterthwaite approximation
  - `PROC MIXED` uses `DDFM=SATTERTH` to specify the use of Satterthwaite approximation
  - `PROC MIXED` uses `DDFM=KR` to specify the use of Kenward & Roger (1997) (working for REML estimation method)

Chapter Review

- One-way Random effects
  - Model
  - Variance component estimation
  - Confidence intervals
- Two-way Random effects
  - Model
  - Variance component estimation
  - F-tests
- Two-way mixed effects
  - Restricted mixed model
  - Unrestricted mixed model
- Computing Expected Mean Squares
  - Helps construct F-statistics