패널자료분석

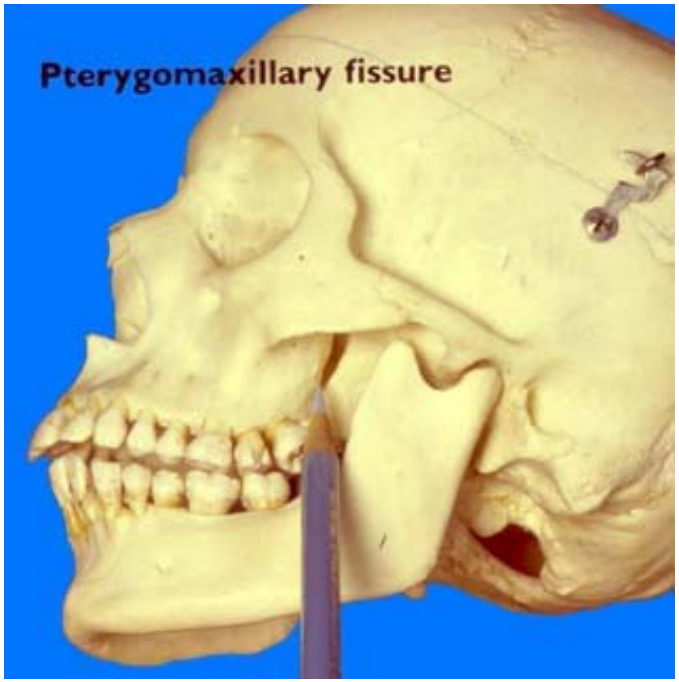
Panel (longitudinal) Data Analysis

패널자료란?

- 동일한 대상 (subject, experimental unit)으로부터 여러 시점에 대해 수집한 자료
  - 인간 혹은 동물에 대한 임상실험
  - 농작물의 성장, 부패

Dental Study

- 27명의 어린이 (16명의 남자, 11명의 여자)
- 각 어린이의 뇌하수체에서 익돌상악열구 (**pterygomaxillary fissure**)까지의 거리(mm)를 8, 10, 12, 14세에 측정
- Questions
  - 시간에 따라 거리가 변화하는가?
  - 변화의 패턴은?
  - 남자와 여자의 패턴 차이는?



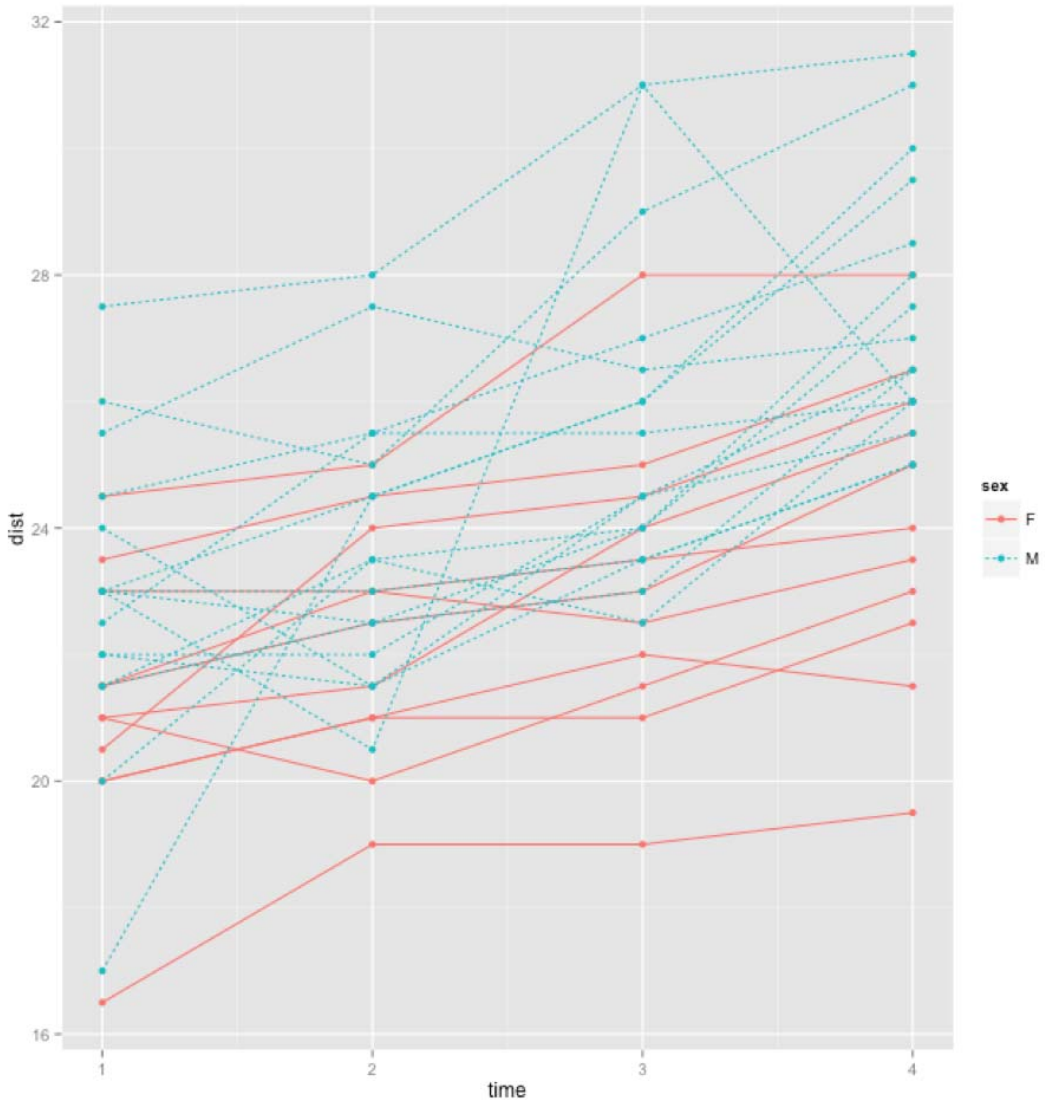
Dental Study: 데이터구조 변형

```
> library(mice)
> potthoffroy
   id sex   d8  d10  d12  d14
1   1   F 21.0 20.0 21.5 23.0
2   2   F 21.0 21.5 24.0 25.5
3   3   F 20.5 24.0 24.5 26.0
4   4   F 23.5 24.5 25.0 26.5
5   5   F 21.5 23.0 22.5 23.5
6   6   F 20.0 21.0 21.0 22.5
7   7   F 21.5 22.5 23.0 25.0
8   8   F 23.0 23.0 23.5 24.0
9   9   F 20.0 21.0 22.0 21.5
10 10   F 16.5 19.0 19.0 19.5
11 11   F 24.5 25.0 28.0 28.0
12 12   M 26.0 25.0 29.0 31.0
13 13   M 21.5 22.5 23.0 26.5
14 14   M 23.0 22.5 24.0 27.5
15 15   M 25.5 27.5 26.5 27.0
16 16   M 20.0 23.5 22.5 26.0
17 17   M 24.5 25.5 27.0 28.5
18 18   M 22.0 22.0 24.5 26.5
19 19   M 24.0 21.5 24.5 25.5
20 20   M 23.0 20.5 31.0 26.0
21 21   M 27.5 28.0 31.0 31.5
22 22   M 23.0 23.0 23.5 25.0
23 23   M 21.5 23.5 24.0 28.0
24 24   M 17.0 24.5 26.0 29.5
25 25   M 22.5 25.5 25.5 26.0
26 26   M 23.0 24.5 26.0 30.0
27 27   M 22.0 21.5 23.5 25.0
```

```
> data=reshape(potthoffroy,idvar="id", varying=list(3:6),v.names="dist",direction="long")
> data$sex.m=1
> data$sex.m[data$sex=="F"]=0
> data
     id sex time dist sex.m
1.1   1   F    1 21.0     0
2.1   2   F    1 21.0     0
3.1   3   F    1 20.5     0
4.1   4   F    1 23.5     0
5.1   5   F    1 21.5     0
6.1   6   F    1 20.0     0
7.1   7   F    1 21.5     0
8.1   8   F    1 23.0     0
9.1   9   F    1 20.0     0
10.1 10   F    1 16.5     0
11.1 11   F    1 24.5     0
12.1 12   M    1 26.0     1
13.1 13   M    1 21.5     1
14.1 14   M    1 23.0     1
15.1 15   M    1 25.5     1
16.1 16   M    1 20.0     1
17.1 17   M    1 24.5     1
18.1 18   M    1 22.0     1
19.1 19   M    1 24.0     1
20.1 20   M    1 23.0     1
21.1 21   M    1 27.5     1
22.1 22   M    1 23.0     1
23.1 23   M    1 21.5     1
24.1 24   M    1 17.0     1
25.1 25   M    1 22.5     1
26.1 26   M    1 23.0     1
27.1 27   M    1 22.0     1
1.2   1   F    2 20.0     0
2.2   2   F    2 21.5     0
3.2   3   F    2 24.0     0
4.2   4   F    2 24.5     0
```

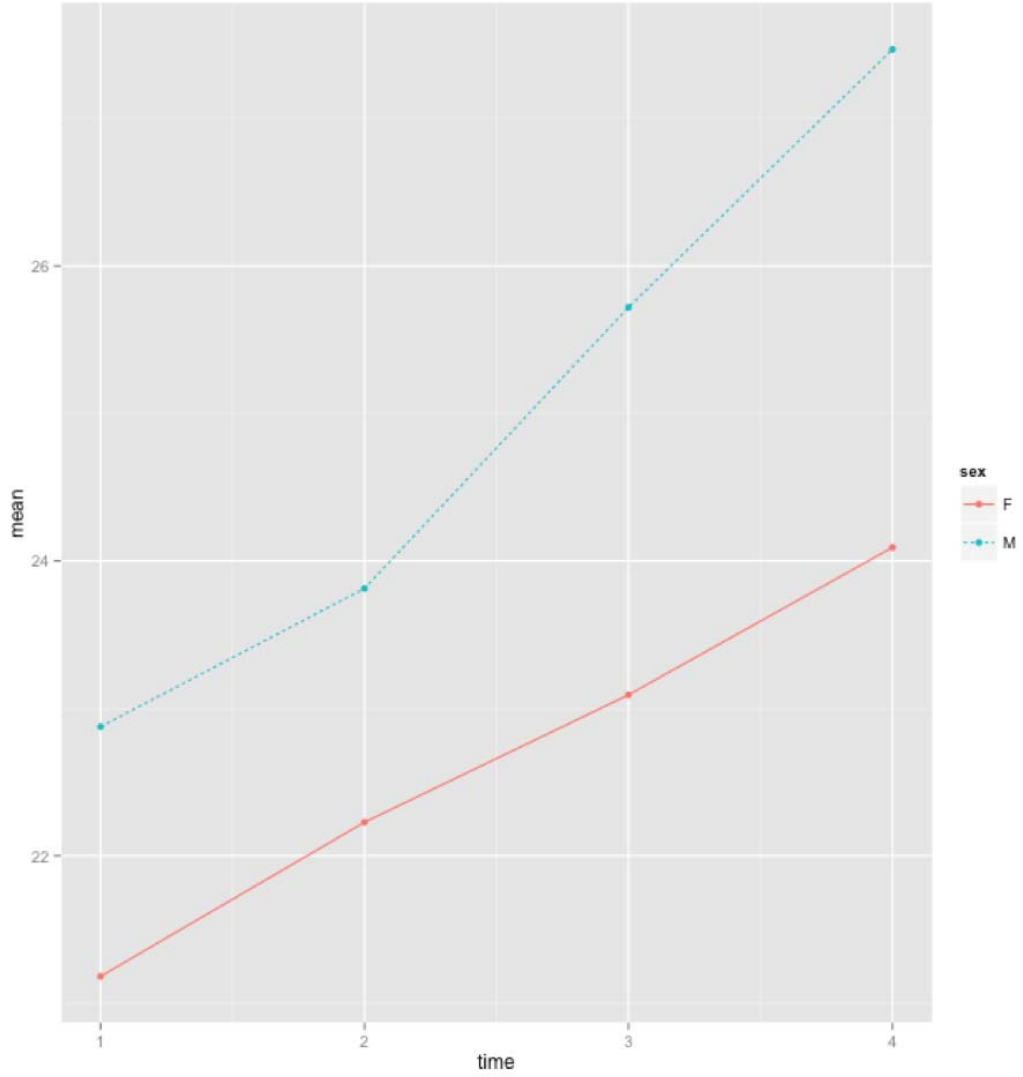
Dental Study: Interaction Plot

```
library(ggplot2)
ggplot(data,aes(y=dist,x=time,group=id,colour=sex))+
  geom_line(aes(linetype=sex),stat="identity")+
  geom_point()
```



Dental Study: 성별 평균

```
library(plyr)
data2=ddply(data,~sex+time,summarize,mean=mean(dist))
ggplot(data2,aes(y=mean,x=time,colour=sex))+
  geom_line(aes(linetype=sex),stat="identity")+
  geom_point()
```



Repeated Measure ANOVA

- 만일 시간의 흐름에 따른 패턴을 파악하는 것이 목적이 아니라 남자와 여자 사이의 평균적인 차이를 파악하는 것이 목적이라면?
  - 각 시점 간의 관측치가 독립이 아님 ➔ 시점 변수를 factor화 한 ANOVA가 적당하지 않음

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \end{bmatrix} \sim N(\mathbf{0}, \sigma^2 V)$$

- Y_{ij}: j번째 그룹의 i번째 subject의 관측치
- μ: 전체 평균
- α_j: 성별 효과 (j번째 그룹 평균과 전체 평균의 차이)

Repeated Measure ANOVA

- Covariance matrix V의 종류
  - Autoregressive(AR)

$$\sigma^2\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

  - Compound symmetry

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix}$$

  - Unstructured

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

library(nlme)사용

AR(1) correlation matrix

```
> model3=gls(dist~sex,data,correlation=corAR1(form=~1|id)) #AR(1)
> anova(model3)
Denom. DF: 106
            numDF  F-value p-value
(Intercept)     1 3280.246  <.0001
sex             1    8.006  0.0056
> summary(model3)
Generalized least squares fit by REML
  Model: dist ~ sex
  Data: data
       AIC      BIC    logLik
  493.0365 503.6903 -242.5183

Correlation Structure: AR(1)
 Formula: ~1 | id
 Parameter estimate(s):
      Phi
0.6354623

Coefficients:
              Value Std.Error  t-value p-value
(Intercept) 22.64243 0.6586210 34.37855  0.0000
sexM         2.42091 0.8555738  2.82957  0.0056

 Correlation:
     (Intr)
sexM -0.77

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.85756288 -0.73122728 -0.03646313  0.50913515  2.28108149

Residual standard error: 2.821756
Degrees of freedom: 108 total; 106 residual
```

Compound Symmetric correlation matrix

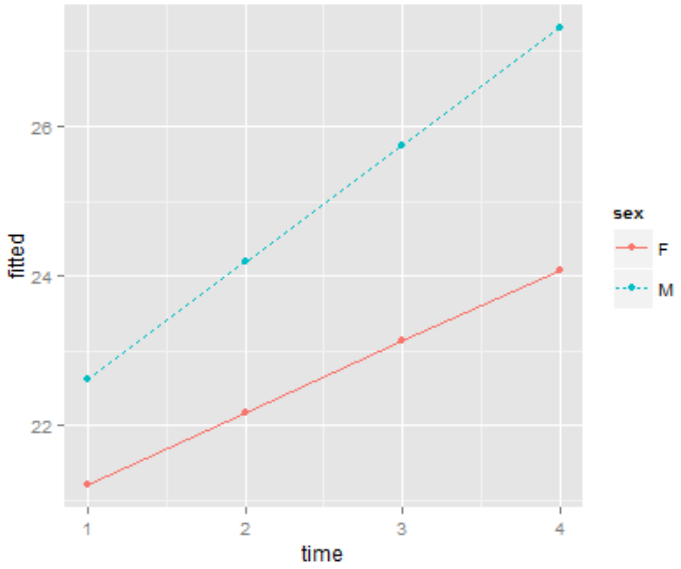
```
> model4=gls(dist~sex,data,correlation=corCompSymm(form=~1|id)) #Compound symmetry
> anova(model4)
Denom. DF: 106
            numDF  F-value p-value
(Intercept)     1 4123.156  <.0001
sex             1    9.292  0.0029
> summary(model4)
Generalized least squares fit by REML
  Model: dist ~ sex
  Data: data
       AIC      BIC    logLik
  513.8718 524.5255 -252.9359

Correlation Structure: Compound symmetry
 Formula: ~1 | id
 Parameter estimate(s):
      Rho
0.3406282

Coefficients:
              Value Std.Error  t-value p-value
(Intercept) 22.647727 0.5861390 38.63884  0.0000
sexM         2.321023 0.7614168  3.04829  0.0029

 Correlation:
     (Intr)
sexM -0.77

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.91434827 -0.72001545 -0.02129916  0.56001202  2.38862270

Residual standard error: 2.734316
Degrees of freedom: 108 total; 106 residual
```

시간의 흐름에 따른 패턴 추정: 선형 회귀모형

- 남녀 간의 차이 뿐 아니라 시간의 흐름에 따른 패턴 추정이 목적이라면?
- 선형 회귀모형
  - 시점 t를 설명변수로 사용 (시점 외에 반복측정에 따라 달라지는 설명변수도 사용 가능)
  - 성별을 구분하는 더미변수($d_i = 1$ for male) 포함
  - id 간의 차이 고려 안함

$$Y_{it} = \beta_0 + (\beta_1 + \beta_3 d_i)t + \beta_2 d_i + \epsilon_{it}$$

```
> model5=lm(dist~sex+time+sex:time,data)
> summary(model5)

Call:
lm(formula = dist ~ sex + time + sex:time, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6156 -1.3219 -0.1682  1.3299  5.2469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.2500     0.8334  24.297  < 2e-16 ***
sexM          0.7969     1.0827   0.736  0.46337
time          0.9591     0.3043   3.152  0.00212 **
sexM:time     0.6097     0.3953   1.542  0.12608
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.257 on 104 degrees of freedom
Multiple R-squared:  0.4227, Adjusted R-squared:  0.4061
F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```



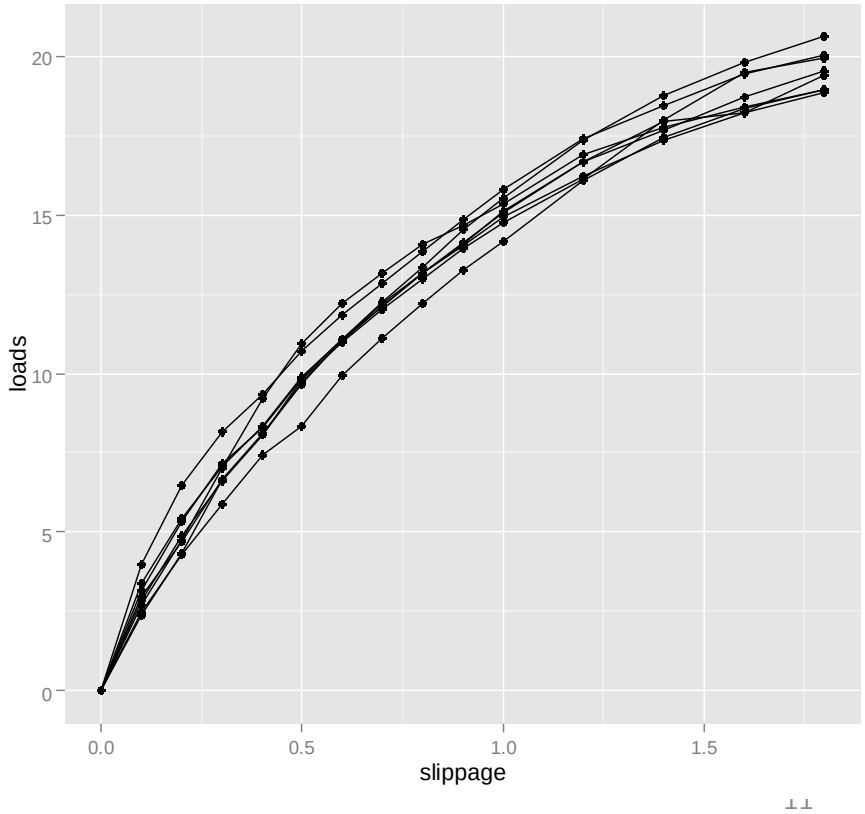

```
data$fitted=model5$fitted.values
ggplot(data,aes(y=fitted,x=time,group=id,colour=sex))+
  geom_line(aes(linetype=sex),stat="identity")+
  geom_point()
```

예: Timber Slippage

- 8개 목재에 대해 클램프에서 미끄러지는데 필요한 무게를 15번씩 반복 측정
- slippage에 따라 load가 어떻게 달라지는가? (각 목재 간의 차이는 큰 관심 없음)

```
> head(timber)
        specimen slippage loads
spec1.1    spec1        0     0
spec2.1    spec2        0     0
spec3.1    spec3        0     0
spec4.1    spec4        0     0
spec5.1    spec5        0     0
spec6.1    spec6        0     0
> tail(timber)
         specimen slippage loads
spec3.15    spec3      1.8 19.40
spec4.15    spec4      1.8 18.93
spec5.15    spec5      1.8 20.62
spec6.15    spec6      1.8 20.05
spec7.15    spec7      1.8 19.54
spec8.15    spec8      1.8 18.87
```

```
ggplot(timber,aes(y=loads,x=slippage,group=specimen))+
  geom_line(stat="identity")+
  geom_point()
```

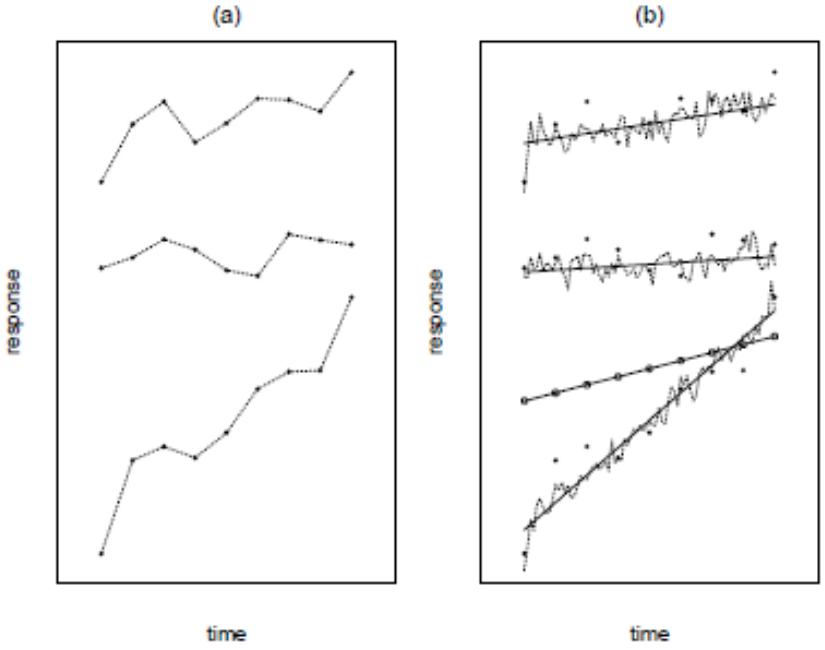


$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon$$

- timber간의 차이 무시
- 한 timber 에서 반복측정된 관측치 사이의 상관관계 무시

```
> fit_timber=lm(loads~slippage+I(slippage^2),data=timber)
> summary(fit_timber)

Call:
lm(formula = loads ~ slippage + I(slippage^2), data = timber)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21305 -0.42879 -0.00969  0.38713  1.75596

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.9434     0.1496   6.305 5.29e-09 ***
slippage      19.8891     0.4038  49.259  < 2e-16 ***
I(slippage^2) -5.4295     0.2209 -24.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6673 on 117 degrees of freedom
Multiple R-squared:  0.9869, Adjusted R-squared:  0.9866
F-statistic:  4394 on 2 and 117 DF,  p-value: < 2.2e-16
```

시간의 흐름에 따른 패턴 추정: 선형혼합모형



IDEA

- 각 subject는 고유한 장기 추세를 가지고 있음
- 실제 관측치는 이 추세를 기반으로 변동성을 가짐
- 측정 오차를 가질 수 있음 (앞의 선형회귀모형은 동일한 성별에 속한 대상들 사이의 측정오차를 고려하지 않음)
- 모든 subject의 추세, 변동성, 측정오차를 각 시점에서 평균을 취하여 전 population에 대한 추세를 얻을 수 있음

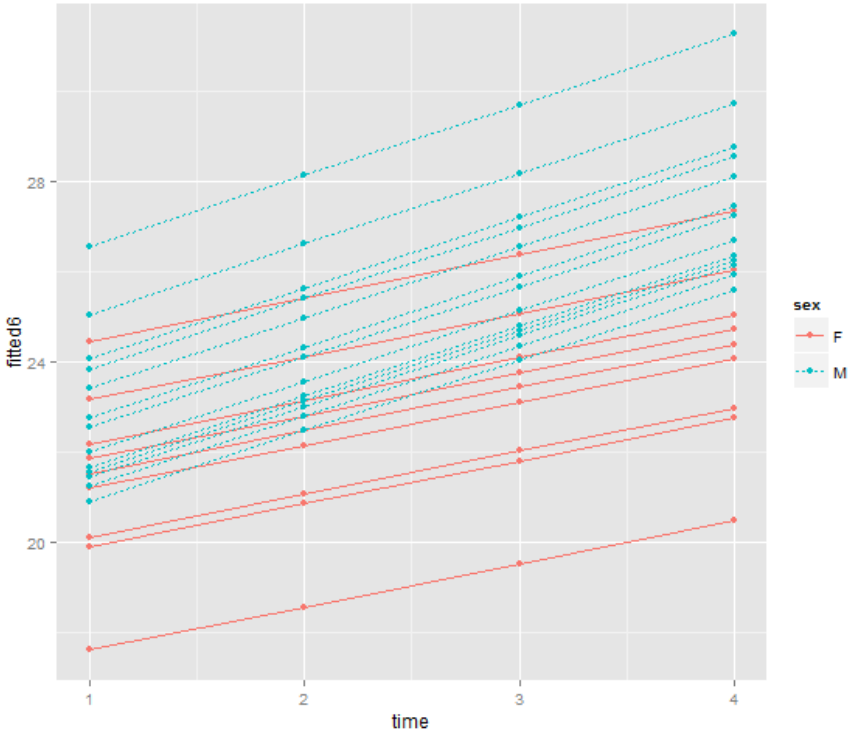
REMARK

- 추세가 반드시 선형일 필요는 없음

- Random intercept model
  - 추세선의 절편에서 subject 간의 오차를 고려

$$Y_{it} = (\beta_0 + u_i) + (\beta_1 + \beta_3 d_i)t + \beta_2 d_i + \epsilon_{it},$$

$$u_i \sim N(0, \sigma_u^2), \epsilon_{it} \sim N(0, \sigma_\epsilon^2)$$



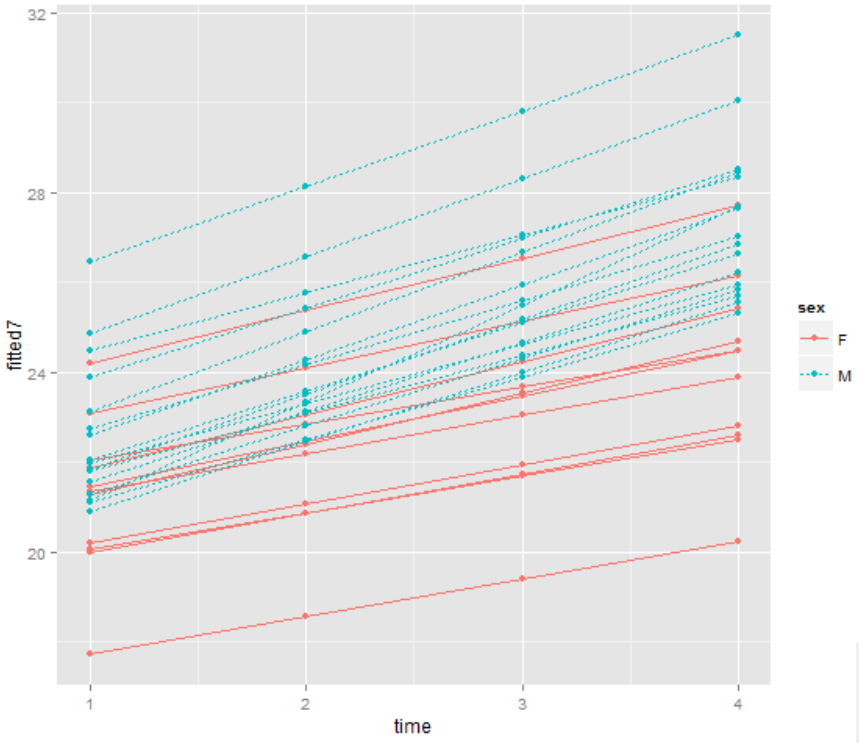

```
> library(lme4)
> library(lmerTest)
> model6=lmer(dist~sex+time+sex:time+(1|id),data=data)
> summary(model6)
Linear mixed model fit by REML
t-tests use  Satterthwaite approximations to degrees of freedom ['merModLmerTest']
Formula: dist ~ sex + time + sex:time + (1 | id)
   Data: data

REML criterion at convergence: 431

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5980 -0.4546  0.0158  0.5024  3.6862

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.299    1.816
 Residual             1.922    1.386
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  20.2500     0.7496 59.3100  27.013  < 2e-16 ***
sexM          0.7969     0.9738 59.3100   0.818   0.4165
time          0.9591     0.1869 79.0000   5.130 2.02e-06 ***
sexM:time     0.6097     0.2428 79.0000   2.511   0.0141 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          (Intr) sexM   time
sexM      -0.770
time      -0.623  0.480
sexM:time  0.480 -0.623 -0.770
```

```
data$fitted6=fitted(model6)
ggplot(data,aes(y=fitted6,x=time,group=id,colour=sex))+
  geom_line(aes(linetype=sex),stat="identity")+
  geom_point()
```

- Random intercept and slope model
  - 추세선의 절편과 기울기에서 subject 간의 오차를 고려

$$Y_{it} = (\beta_0 + u_{1i}) + (\beta_1 + \beta_3 d_i + u_{2i})t + \beta_2 d_i + \epsilon_{it},$$

$$(u_{1i}, u_{2i})^T \sim N(0, \Sigma), \epsilon_{it} \sim N(0, \sigma_\epsilon^2)$$

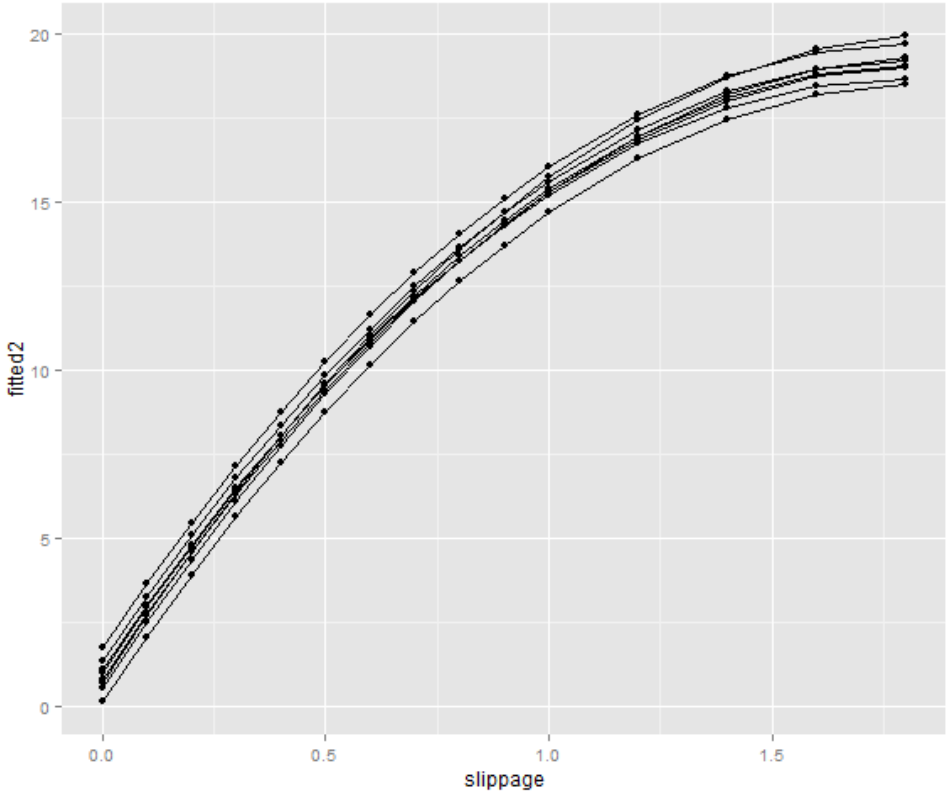



```
> model6=lmer(dist~sex+time+sex:time+(1|id),data=data)
> summary(model6)
Linear mixed model fit by REML
t-tests use  Satterthwaite approximations to degrees of freedom ['merModLmerTest']
Formula: dist ~ sex + time + sex:time + (1 | id)
   Data: data

REML criterion at convergence: 431

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5980 -0.4546  0.0158  0.5024  3.6862

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.299    1.816
 Residual             1.922    1.386
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  20.2500     0.7496 59.3100  27.013  < 2e-16 ***
sexM          0.7969     0.9738 59.3100   0.818   0.4165
time          0.9591     0.1869 79.0000   5.130 2.02e-06 ***
sexM:time     0.6097     0.2428 79.0000   2.511   0.0141 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
          (Intr) sexM   time
sexM      -0.770
time      -0.623  0.480
sexM:time  0.480 -0.623 -0.770
```

```
data$fitted7=fitted(model7)
ggplot(data,aes(y=fitted7,x=time,group=id,colour=sex))+
  geom_line(aes(linetype=sex),stat="identity")+
  geom_point()
```

예: Timber Slippage

- Timber간 절편과 slippage의 기울기 오차 고려




```
> fit_timber2=lmer(loads~slippage+I(slippage^2)+(1+slippage|specimen),data=timber)
> summary(fit_timber2)
Linear mixed model fit by REML
t-tests use  Satterthwaite approximations to degrees of freedom ['merModLmerTest']
Formula: loads ~ slippage + I(slippage^2) + (1 + slippage | specimen)
   Data: timber

REML criterion at convergence: 205.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4946 -0.5045 -0.0281  0.5953  2.1885

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 specimen (Intercept) 0.2824   0.5314
          slippage    0.1270   0.3564   -0.60
 Residual             0.2489   0.4988
Number of obs: 120, groups:  specimen, 8

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)    0.9434     0.2187   9.1900   4.314  0.00186 **
slippage      19.8891     0.3271  78.8600  60.804  < 2e-16 ***
I(slippage^2) -5.4295     0.1651 103.0000 -32.880  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) slippg
slippage    -0.594
I(slippg^2)  0.358 -0.885
```

```
timber$fitted2=fitted(fit_timber2)
ggplot(timber,aes(y=fitted2,x=slippage,group=specimen))+
  geom_line(stat="identity")+
  geom_point()
```