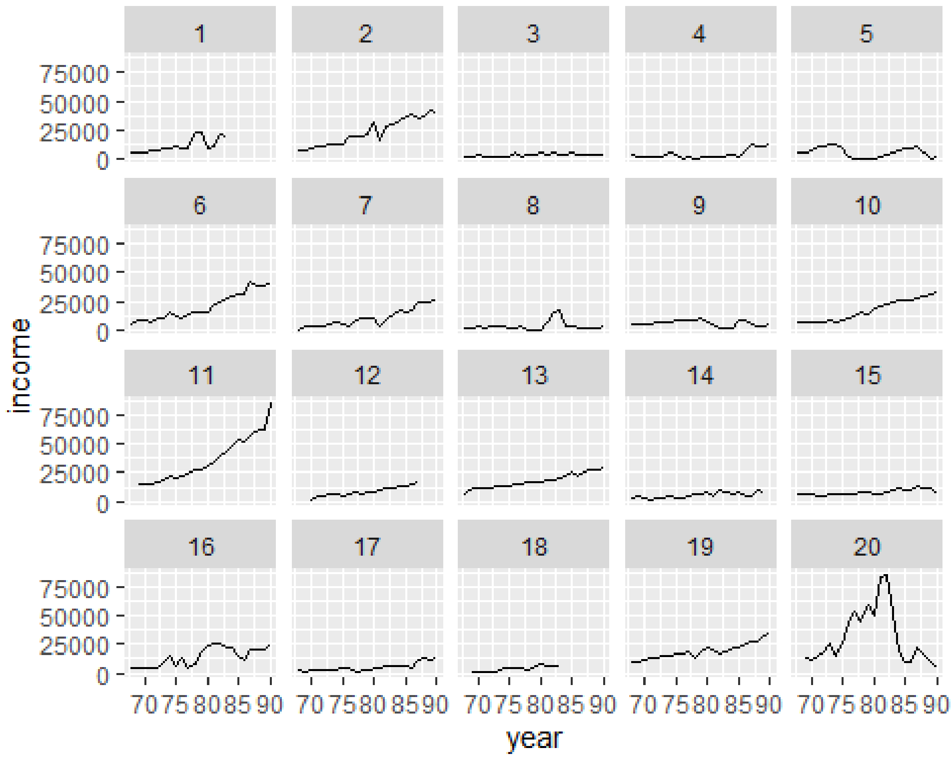
R_HW7_Longitudinal Data

Eom SangJun

2020 11 12

어떤 사람들에 대해 age, years of education, sex 와 20 년 동안(20 년을 다 채우지 못하는 데이터들이 있지만 최소 11 년은 채웠다) income 의 변화를 조사한 데이터를 분석해보자.

```
data(psid, package = 'faraway')
head(psid)

##   age educ sex income year person
## 1  31   12   M   6000   68      1
## 2  31   12   M   5300   69      1
## 3  31   12   M   5200   70      1
## 4  31   12   M   6900   71      1
## 5  31   12   M   7500   72      1
## 6  31   12   M   8000   73      1

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

우선 20 명의 데이터만 살펴보자.

psid20 <- filter(psid, person <=20)
library(ggplot2)
ggplot(psid20, aes(x=year, y=income)) + geom_line() + facet_wrap(~ person)
```



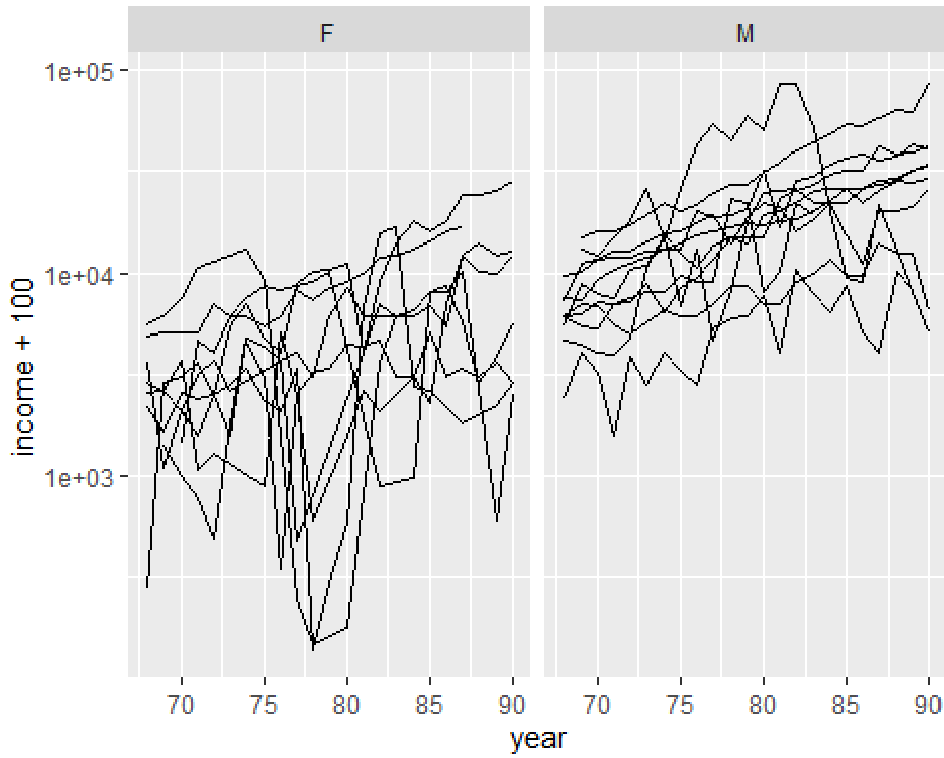
- ➔ 어떤 사람들은 income 이 꾸준히 오르는 반면, 어떤 사람은 변화가 거의 없기도 하고 20 번의 경우 굉장히 dramatic 한 변화가 있었다.

- ➔ 중간에 그래프가 끊긴 사람은 조사가 중간에 끊겼음을 의미.

년도의 흐름에 따른 income 변화를 성별로 구별하여 살펴보자.

```
ggplot(psid20, aes(x=year, y=income+100, group = person)) +
  geom_line() +
  facet_wrap(~ sex) +
  scale_y_log10()
```

여기서 income 에 100 을 더해준 이유는 짧은 기간동안 매우 낮은 income 을 받은 사람들의 effect 를 없애 주기 위함이다.



➔ 일반적으로 여성보다 남성의 income 이 높으며 더 stable 한 모습을 보인다.

➔ 다만 여성의 income 증가가 더 가파른 형태를 보인다.

각 line 을 각 개인에게 fitting 할 수 있다.

예시로 첫 번째 사람의 line 을 fitting 해보자.

그 전에 앞서 year 에서 78(median of year)를 빼 주어야 한다. 왜냐하면 모든 사람들이 조사를 끝까지 마치지는 않았는데 만약 year 를 그대로 둘 경우 intercept 가 1978 년이 아니라 1990 년의 predicted income 을 나타낼 것이기 때문이다. 하지만 1990 년의 income 이 없는 사람들이 있으니 1978 년을 기준으로 하기로 한다.

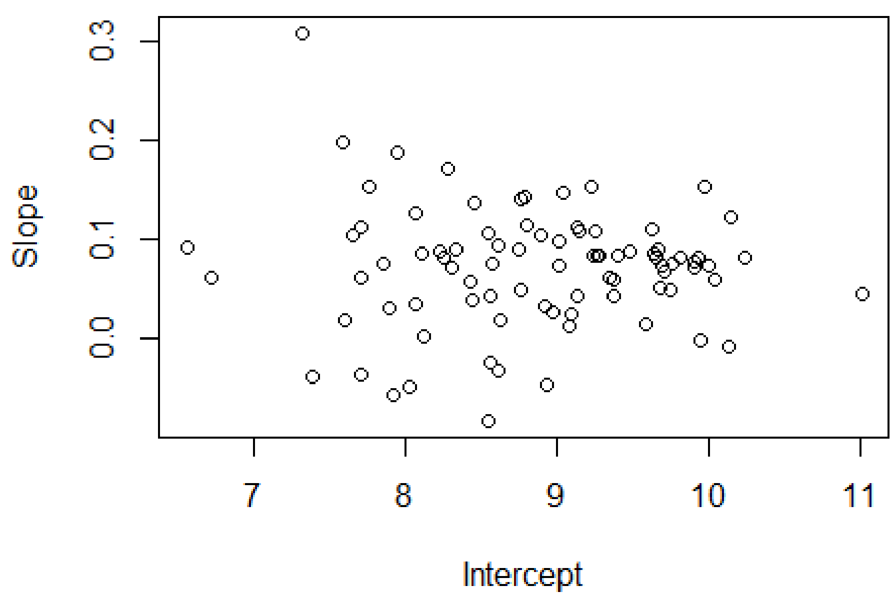
```
lmod <- lm(log(income) ~ I(year-78), subset=(person==1), psid)
coef(lmod)
```

```
##  (Intercept) I(year - 78)
##    9.3999568    0.0842667
```

```
library(lme4)
```

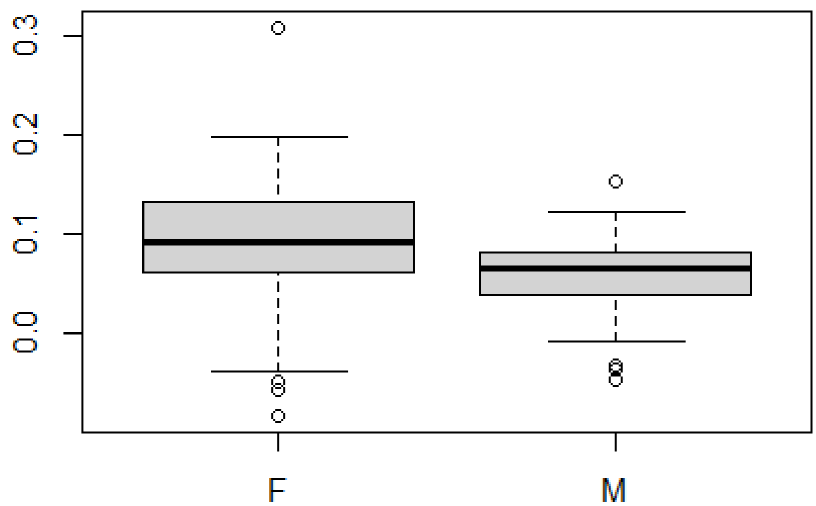
```
## Loading required package: Matrix
```

```
이번에는 모든 개인에게 대해 line 을 fitting 하고 결과를 그려보자.

lmList command 는 data 내에서 각 group 에 대해 linear model 을 fit 한 결과를 구해준다. 여기서 group 은 person 으로 지정해준다.

각 linear model 에서 slope 와 intercept 들을 구해서 살펴보자.

ml <- lmList(log(income) ~ I(year-78) | person, psid)
intercepts <- sapply(ml, coef)[1,]
slopes <- sapply(ml, coef)[2,]

plot(intercepts, slopes, xlab='Intercept', ylab='Slope')
```



- ➔ Slope 와 intercept 간에는 상관관계가 거의 없는 것으로 보인다.
- ➔ 이는 income 과 income growth 를 각각 따로 test 할 수 있음을 알려준다.

```
성별에 따른 income growth 차이를 살펴보자.

psex <- psid$sex[match(1:85, psid$person)]
boxplot(split(slopes, psex))
```



→ 여성이 남성보다 income growth 가 더 커 보인다.

실제로 그러한 지를 t-test 를 통해 살펴보자.

```
t.test(slopes[psex=='M'], slopes[psex=='F'])

##
##  Welch Two Sample t-test
##
## data:  slopes[psex == "M"] and slopes[psex == "F"]
## t = -2.3786, df = 56.736, p-value = 0.02077
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05916871 -0.00507729
## sample estimates:
##  mean of x  mean of y
## 0.05691046 0.08903346
```

→ p-value 가 0.05 보다 낮기 때문에 여성의 income growth 는 남성의 것보다 유의미하게 크다는 것을 알 수 있다.

Income 자체는 어떨까?

```r
t.test(intercepts[psex=='M'], intercepts[psex=='F'])

##
##  Welch Two Sample t-test
##
## data:  intercepts[psex == "M"] and intercepts[psex == "F"]
## t = 8.2199, df = 79.719, p-value = 3.065e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8738792 1.4322218
## sample estimates:
## mean of x mean of y
##  9.382325  8.229275
```

→ p-value 가 매우 낮으므로 남성이 여성보다 income 자체는 유의미하게 크다는 것을 알 수 있다.

이러한 분석들을 우리는 response feature analysis 라고 부른다.

이러한 분석은 사용하기는 쉽지만 data 의 중요한 특징을 고르기를 요한다.

그런데 문제는 이 것이 쉽지 않을 뿐만 아니라 선택된 특징 말고 다른 추가적인 정보들은 무시당할 가능성이 있다는 점이다. 즉 정보 손실이 있는 것.

일단 우리는 현재 가지고 있는 데이터, 즉 age, sex, edu 등을 활용하여 year 에 따른 income 을 어느 정도 예측할 수는 있다. 그러나 이는 부분적인 것일 뿐 완벽한 예측은 아니라는 것을 예상할 수 있다. 왜냐하면 income 에 관련한 모든 데이터를 가지고 있는 것은 아니기 때문이다. 즉, subject's income 에 영향을 미치는 factor 들은 분명히 존재하는데, 이러한 factor 들은 income 의 일반적인 크기에 영향을 끼칠 수도 있고 또는 growth 속도에 영향을 미칠 수도 있다. 다시 말하면 income 의 intercept 에 영향을 미칠 수도, slope 에 미칠 수도 있다. 즉, variation 이 두 부분으로 나뉘며 variation 을 모델링할 때, random intercept 와 slope 로 반영해줄 수 있다.

우리는 또한 어느 정도 year-to-year variation 이 subject 내에서 존재할 것이라고 예상할 수 있다. 따라서 결론적으로 다음과 같은 식의 모델을 세워보자. (이 때 random effect 의 random 들은 homogeneous 이며 uncorrelated 되어 있다고 가정한다)

$$\log(income)_{ij} = \mu + \beta_y year_i + \beta_s sex_j + \beta_{ys} sex_j * year_i + \beta_e educ_j + \beta_a age_j + \gamma_j^0 + \gamma_j^1 year_i + \varepsilon_{ij}$$

Where i indexes the year and j indexes the individual

In addition

$$\begin{pmatrix}\gamma_k^0\\ \gamma_k^1\end{pmatrix} \sim N(0, \sigma^2 D)$$

```
library(lme4)
psid$cyear <- psid$year - 78
mmod <- lmer(log(income) ~ cyear*sex + age + educ + (cyear|person), psid)

library(faraway)

##
## Attaching package: 'faraway'

## The following object is masked _by_ '.GlobalEnv':
##
##     psid

sumary(mmod, digits=3)

## Fixed Effects:
##               coef.est coef.se
## (Intercept)  6.674    0.543
## cyear        0.085    0.009
## sexM         1.150    0.121
## age          0.011    0.014
## educ         0.104    0.021
## cyear:sexM  -0.026    0.012
##
## Random Effects:
##  Groups   Name        Std.Dev. Corr
##  person   (Intercept) 0.531
##           cyear       0.049    0.187
##  Residual             0.684
## ---
## number of obs: 1661, groups: person, 85
## AIC = 3839.8, DIC = 3751.2
## deviance = 3785.5
```

→ 우선 fixed effect 부터 살펴보자.

한 단위의 추가적인 educational year 에 대해 income 은 약 10% 정도 상승한다.

또한 age 는 통계적으로 별로 유의하지 않은 것으로 보인다.

여성(reference group)의 경우 매년 8.5%의 임금 상승이 나타나며, 남성의 경우는 8.5-2.6 = 5.9%이다.

또한 이 데이터에서 남성의 income 은 여성에 비해 exp(1.15) (=3.16) 배 높다.

우리는 남성과 여성의 평균적인 값은 알지만, 개개인은 변동이 있을 것이다.

Intercept 와 slope 의 standard deviation 값($\sigma\sqrt{D_{11}}$,$\sigma\sqrt{D_{22}}$)은 0.531, 0.049 이다. 또한 둘의 correlation 값($cor(\gamma^0,\gamma^1)$)은 0.189 이다. 최종적으로 설명되지 않은 variation 의 값(ε_{ijk})은 0.684 이다.

해석해보면, 임금 상승에서의 variation 은 크지 않지만, 임금 자체의 variation 은 상대적으로 큰 편이다. 더 나아가, 큰 residual variation 값을 보았을 때 year-to-year variation in income 이 크다는 것을 알 수 있다.

Kenward-Roger adjusted F-test 를 이용하여 fixed effect term 의 significance 를 test 하자. 우선 그 중 interaction term 을 test 하자.

```
library(pbkrtest)
```

```
## Warning: package 'pbkrtest' was built under R version 4.0.3
```

```
mmod <- lmer(log(income) ~ cyear*sex + age + educ +(cyear|person), psid, REML
=FALSE)
mmodr <- lmer(log(income) ~ cyear + sex + age + educ + (cyear|person), psid,
REML = FALSE)
KRmodcomp(mmod, mmodr)
```

```
## F-test with Kenward-Roger approximation; time: 4.03 sec
## large : log(income) ~ cyear + sex + age + educ + (cyear | person) + cyear:
sex
## small : log(income) ~ cyear + sex + age + educ + (cyear | person)
##          stat     ndf     ddf F.scaling p.value
## Ftest  4.6142  1.0000 81.3279         1 0.03468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ test 결과, interaction term 이 통계적으로 유의미하다고 나온다. 따라서 여성이 남성에 비해 income 이 빠르게 상승한다고 해석할 수 있다.

우리는 random effect term 에 대해 parametric bootstrap 방법을 이용하여 test 할 수 있다. 이 방식을 이용하면 모든 파라미터에 대해 confidence interval 을 비교적 안정적으로 구해줄 수 있다.

```
confint(mmod, method = 'boot')
```

```
## Computing bootstrap confidence intervals ...
```

```
##
## 1 warning(s): Model failed to converge with max|grad| = 0.00314273 (tol =
0.002, component 1)
```

```
##                   2.5 %        97.5 %
## .sig01       0.41928373  0.5838344888
## .sig02      -0.07988384  0.4516737179
## .sig03       0.03719285  0.0562105540
## .sigma       0.66223769  0.7066608900
## (Intercept)  5.69646308  7.7902829774
## cyear        0.06682020  0.1037156239
## sexM         0.94780207  1.3962677896
## age         -0.01351970  0.0347559833
## educ         0.06052459  0.1488210376
## cyear:sexM  -0.05239976 -0.0005482126
```

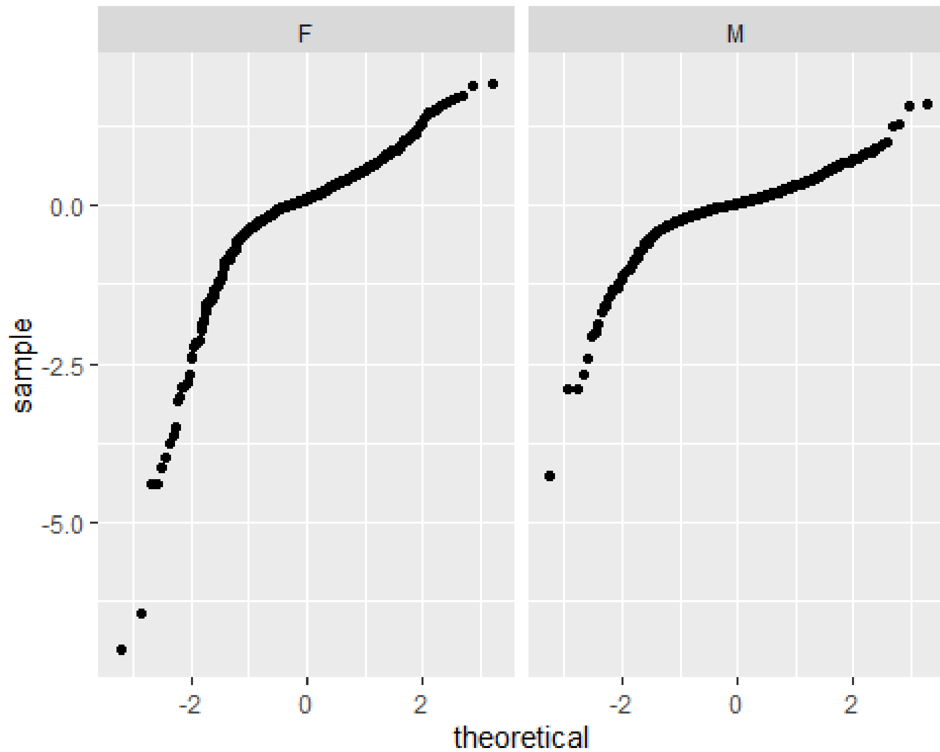
→ .sig01 이 intercept 의 standard deviation, .sig03 이 slope 의 sd 인데 두 term 모두 confidence interval 이 명확히 0 이상에 있음을 알 수 있다.

따라서 두 term 을 남겨두는 것이 좋은데, .sig02 에 해당하는 correlation 값은 구간이 0 을 포함하고 있음을 알 수 있다. 하지만 이 term 은 해석하기도 어려울뿐더러, 이를 없앰으로써 얻는 이익이 거의 없다. 따라서 그냥 두자.

일반적인 linear model 보다 longitudinal data 의 경우 더 넓은 범위의 사용 가능한 diagnostic plot 들이 있다. 일반적인 residual 에 더해 random effects 를 조사해보아야 한다.

여기서는 우선 residual 을 성별로 쪼개서 보자.

```
diagd <- fortify.merMod(mmod)
ggplot(diagd, aes(sample=.resid)) + stat_qq() + facet_grid(~sex)
```



➔ 두 plot 모두 residual 이 normally distributed 하지 않다는 것을 알 수 있다.

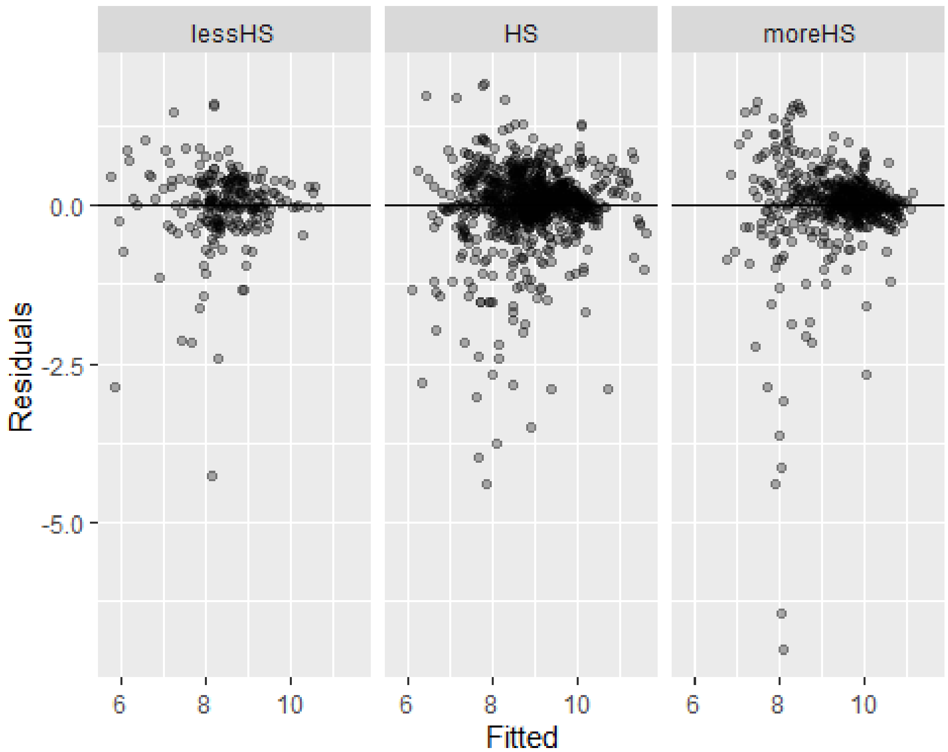
➔ Lower income 에 대해 long tail 을 가진다.

➔ 이는 response 에 대해 log transformation 을 해주어야 한다는 것을 알 수 있다.

➔ 또한 여성의 경우 남성에 비해 variance 가 더 크다는 것을 알 수 있다. 이는 model 에 수정이 필요함을 보여준다.

마지막으로 education level 을 세 개로 구분해서 residuals vs fitted value plot 을 그려보자.

```
diagd$edulevel <- cut(psid$educ, c(0, 8.5, 12.5, 20), labels=c('lessHS', 'HS',
 'moreHS'))
ggplot(diagd, aes(x=.fitted, y=.resid)) +
  geom_point(alpha=0.3) +
  geom_hline(yintercept = 0) +
  facet_grid(~ edulevel) +
  xlab('Fitted') +
  ylab('Residuals')
```



➔ 마찬가지로 constant variance assumption 이 깨진다는 것을 알 수 있다.

➔ 따라서 또 다시 response transformation 에 대해 고려해볼 수 있고 random effects 의 plot 이 유용할 수 있다.