Applied statistics

SM-4337

Chapter 4: Longitudinal Data Analysis

Dr Elvynna Leong

Longitudinal data analysis

- In repeated measures designs, there are several individuals and measurements are taken **repeatedly** on each individual.
- When these repeated measurements are taken **over time**, it is called a longitudinal study (or panel study).
- Typically various covariates concerning the individual are recorded and the interest centers on how the response depends on the covariates over time.
- This module will cover the analysis and interpretation of results from longitudinal studies.
- Emphasis will be on model development, use of R software, and interpretation of results.
- Theoretical basis for results mentioned but not developed.

Longitudinal data analysis

Features of longitudinal data

- Defining feature of longitudinal studies is that measurements of the same individuals are taken repeatedly through time.
- Longitudinal studies allow direct study of change over time.
- Objective: primary goal is to characterize the change in response over time and the factors that influence change.

Longitudinal data analysis

Complications

(since they are the same person)

1. Repeated measures on individuals are correlated.
2. Variability is often heterogeneous across measurement occasions.

- Longitudinal data require somewhat more sophisticated statistical techniques because the repeated observations are usually correlated.
- Correlation arises due to repeated measures on the same individuals.
- Correlations must be accounted for in order to obtain valid inferences.

Longitudinal data analysis

Typically, various covariates concerning the individual are recorded and the interest centers on how the response depends on the covariates over time.

(Longitudinal study)

Often, it is reasonable to believe that the response of each individual has several components:

- a fixed effect, which is a function of the covariates; (independent variables)
- a random effect, which expresses the variation between individuals; (extra)
- an error, which is due to measurement or unrecorded variables.

Longitudinal data analysis

Suppose each individual has response y_i, a vector of length n_i which is modeled conditionally on the random effects γ_i as:

$$y_i|\gamma_i \sim N(X_i\beta + Z_i\gamma_i, \sigma^2\Lambda_i)$$

— Not in syllabus

We assume that the random effects $\gamma_i \sim N(0, \sigma^2 D)$ so that

response

$$y_i \sim N(X_i\beta, \Sigma_i)$$

where $\Sigma_i = \sigma^2(\lambda_i + Z_i D Z_i^T)$.

Longitudinal data analysis

Now suppose we have M individuals and we can assume the errors and the random effects between individuals are uncorrelated, then we can combine the data as:

response $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}$ x-axis $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix}$ random effects $\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_M \end{bmatrix}$

variance of γ

and $\bar{D} = diag(D, D, \cdots, D), Z = diag(Z_1, Z_2, \cdots, Z_M), \Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_M)$, and $\Lambda = diag(\Lambda_1, \Lambda_2, \cdots, \Lambda_M)$. Now we can write the model simply as

$$y \sim N(X\beta, \Sigma),\ \Sigma = \sigma^2(\Lambda + Z\bar{D}Z^T)$$

Longitudinal data analysis

Example

(Longitudinal Study)

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals described in Hill (1992). There are currently 8700 households in the study and many variables are measured. We chose to analyze a random subset of this data consisting of *85 heads of household* who were aged 25-39 in 1968 and had complete data for at least 11 of the years between 1968 and 1990. The variables included were *annual income*, *gender*, *years of education* and *age* in 1968.

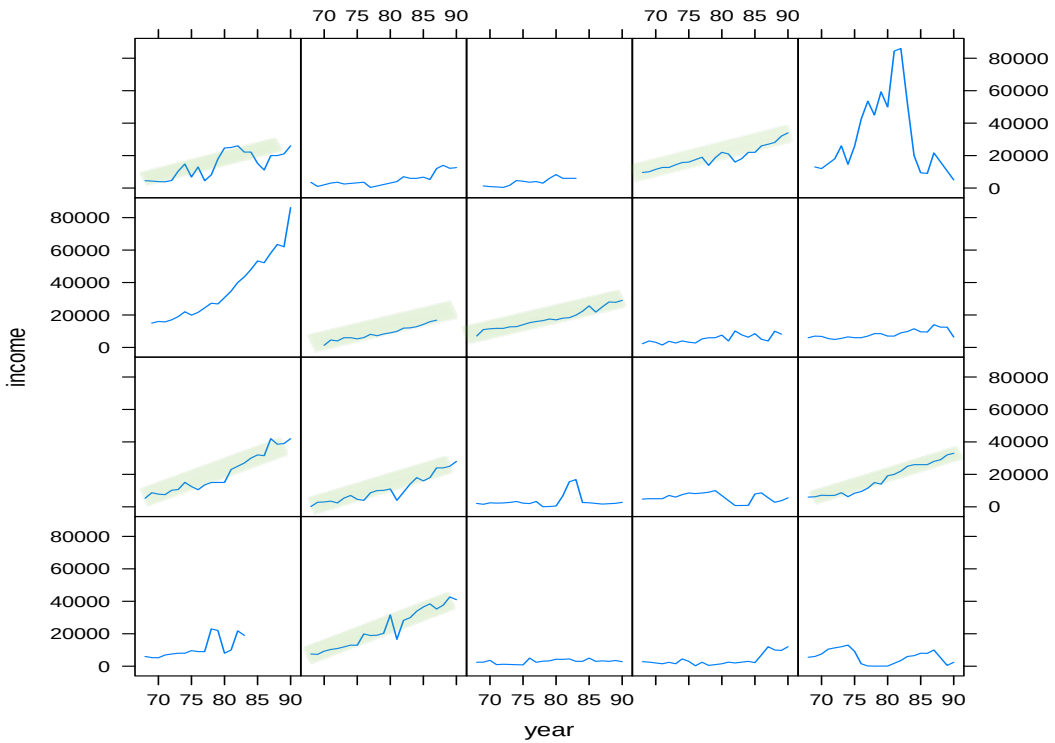
Longitudinal data analysis


```
> library(faraway)
> data(psid)
> 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(lattice)
> pdf(file="LA1.pdf")
> xyplot(income~year|person, psid, type="l", subset=(person<21),strip=FALSE)
> dev.off()
```

(given) "year by person"

- The first 20 subjects are shown in the figure below.

Longitudinal data analysis



- Some individuals have a slowly increasing income, typical of someone in steady employment in the same job.
- Other individuals have more erratic incomes.

Longitudinal data analysis

- We can show how the incomes vary by sex. Income is more naturally considered on a log-scale.

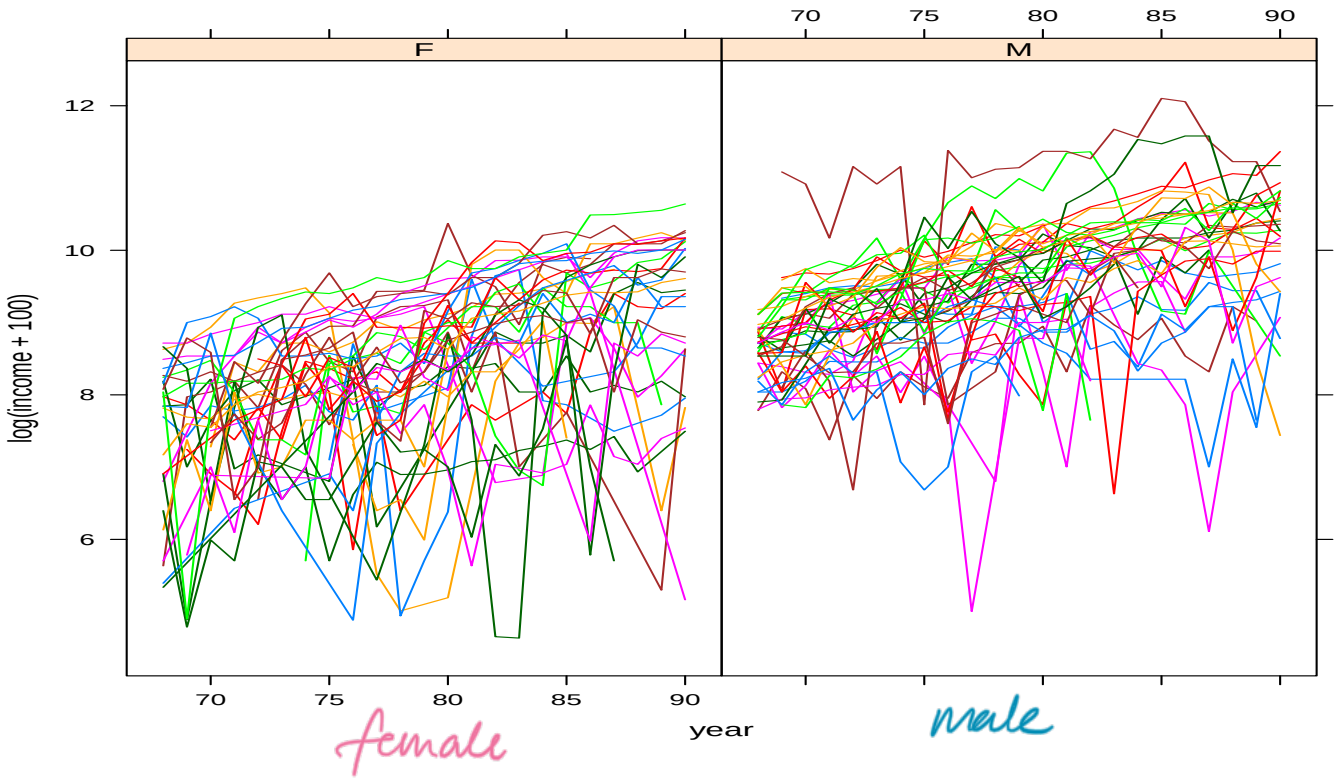
```
> pdf(file="LA4.pdf")
> xyplot(log(income+100)~year|sex,psid,type="l",groups=person)
> dev.off()
```

"year by gender"

(transform)

- We added \$100 to the income of each subject to remove the effect of some subjects having very low incomes for short period of time.
- These cases distorted the plots without the adjustment.

Longitudinal data analysis



- The men's incomes are generally higher and less variable while women's incomes are more variable, but are perhaps increasing more quickly.

Longitudinal data analysis

- We have centered the predictor at the median value so that the intercept will represent the predicted log income in 1978 and not the year 1900 (which would be nonsense).
- We fit a line for all the subjects and plot the results:

```
> slopes<-numeric(85)
> intercepts<-numeric(85)
> for(i in 1:85){
+   lmod<-lm(log(income)~I(year-78),subset=(person==i),psid)
+   intercepts[i]<-coef(lmod)[1]
+   slopes[i]<-coef(lmod)[2]
+   }
>
> pdf(file="LA6.pdf")
> plot(intercepts, slopes, xlab="Intercept",ylab="Slope")
> dev.off()
```

"empty spaces for 85 numbers"

gradient

y-intercept

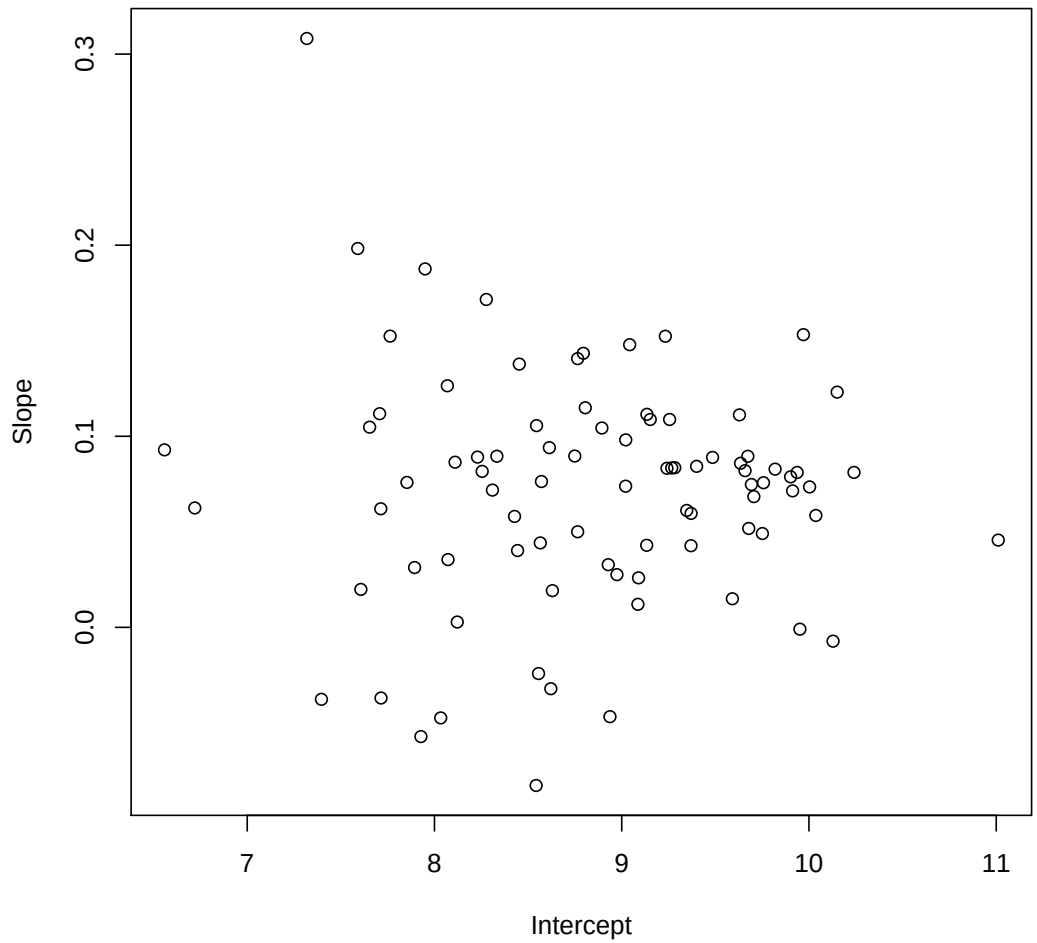
$y = \beta_0 + \beta_1 1(year-78)$

$= 0$ when year 1978 ✓ (median)

$y = \beta_0 + \beta_1 1(year)$

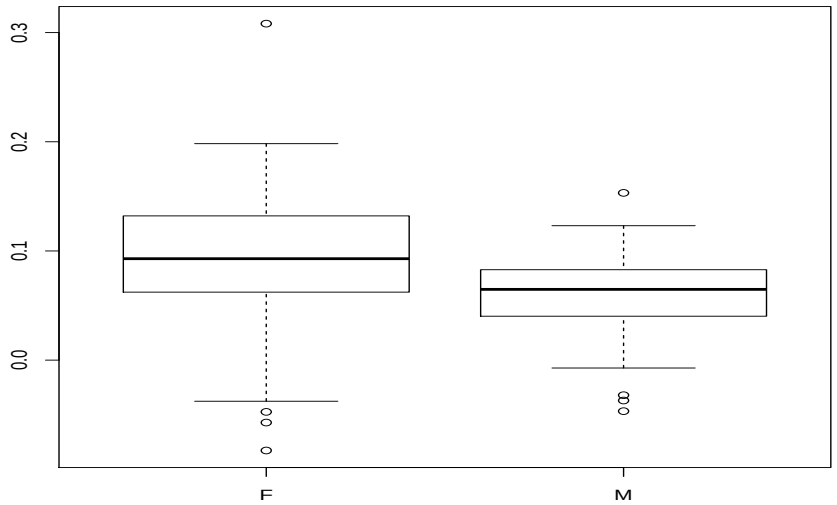
$= 0$ when year 1900 ✗

Longitudinal data analysis



Longitudinal data analysis


```
> pdf(file="LA8.pdf")
> psex<-psid$sex[match(1:85,psid$person)]
> boxplot(split(slopes,psex))
> dev.off()
```



growth rates (slopes): female is higher

We can simply compare the income growth rates for men and women.

Longitudinal data analysis

t-test


```
> 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
```

growth rates

(M) (F)

$H_0: \mu_M = \mu_F$

$H_1: \mu_M \neq \mu_F$

We see that women have significantly higher growth rate then men.

```
> 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
```

income

(M) (F)

We see that men have significantly higher incomes.

Longitudinal data analysis

- Suppose that the income change over time can be partly predicted by the subject's age, sex and educational level.
- Clearly there are other factors that will affect a subject's income.
- These factors may cause the income to be generally higher or lower or they may cause the income to grow at a faster or slower rate.
- We can model this variation with a random intercept and slope, respectively, for each subject.
- We also expect that there will be some year-to-year variation within each subject.
- For simplicity, assume that this error is homogeneous and uncorrelated.
- We also center the year to aid interpretation as before.

Longitudinal data analysis

two-way interaction (with higher order variable)

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

longitudinal study

random effects

This model can be written as:

$$log(\text{income})_{ij} = \mu + \beta_y \text{year}_i + \beta_s \text{sex}_j + \beta_{ys} \text{sex}_j \times \text{year}_i + \beta_e \text{educ}_j + \beta_a \text{age}_j + \gamma_j^0 + \gamma_j^1 \text{year}_i + \varepsilon_{ij}$$

error

where i indexes the year and j indexes the individual. We have

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

Longitudinal data analysis

The model summary is:

```
> summary(mmod)
Linear mixed model fit by REML ['lmerMod']
Formula: log(income) ~ cyear * sex + age + educ + (cyear | person)
   Data: psid

REML criterion at convergence: 3819.8

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.2310  -0.2134   0.0795   0.4147   2.8254

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 person   (Intercept) 0.2817   0.53071
          cyear       0.0024   0.04899  0.19
 Residual             0.4673   0.68357
Number of obs: 1661, groups:  person, 85
```

Longitudinal data analysis


```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.67420    0.54332  12.284
cyear        0.08531    0.00900   9.480
sexM         1.15031    0.12129   9.484
age          0.01093    0.01352   0.808
educ         0.10421    0.02144   4.861
cyear:sexM  -0.02631    0.01224  -2.150
```

another way to find sig.:

95% C.I. — C.I.: $\bar{X} \pm 1.96 s.e.$

(0.068, 0.103)
(0.913, 1.388)
(−0.016, 0.037)
(0.062, 0.146) *
(−0.050, −0.002) *

(interaction) cyear:sexM — F=0 — cyear : sex F

if 0 is within 95% C.I. ✗ sig.
if 0 is not within 95% C.I. ✓ sig.

Interpretation of fixed effects:

Holding all other variables constant,

- We see that income increases about 10% for each additional year of education.
- Age does not appear to be significant.
- For females, the reference level in this example, income increase about 8.5% a year, while for men, it increases about 8.5-2.6=5.9% a year.
- For this data, the incomes of men are $exp(1.15) = 3.16$ times higher.

Longitudinal data analysis


```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 person   (Intercept) 0.2817   0.53071
          cyear       0.0024   0.04899  0.19
 Residual             0.4673   0.68357
Number of obs: 1661, groups:  person, 85
```

Interpretation of random effects:

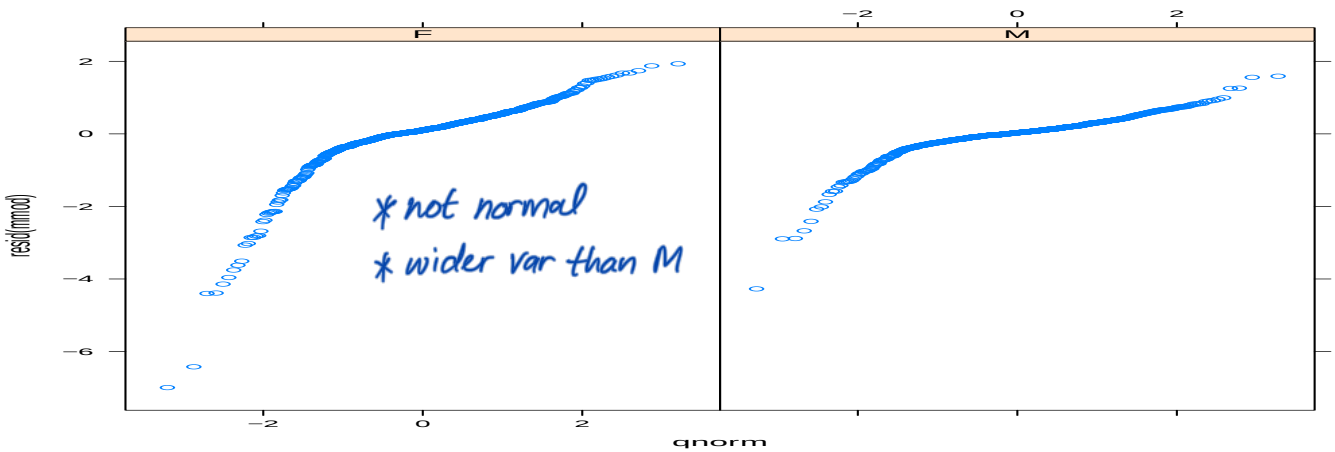
- We know the mean for males and females, but individuals will vary about this.
- The standard deviation for the intercept and slope are 0.531 and 0.049 ($\sigma\sqrt{D_{11}}$ and $\sigma\sqrt{D_{22}}$), respectively.
- These have a correlation of 0.19 ($\text{corr}(\gamma^0, \gamma^1)$).
- There is some additional variation in the measurement not so far accounted for having standard deviation of 0.68($sd(\varepsilon_{ijk})$).

Longitudinal data analysis

- We see that the variation increase in income is relatively small while the variation in overall income between individuals is quite large. Furthermore, given the large residual variance, there is a large year-to-year variation in incomes.
- There is a wider range of possible diagnostic plots that can be made with longitudinal data than with a standard linear model.
- In addition to the usual residuals, there are random effects to be examined.
- We may wish to break the residuals down by sex as shown in the QQ plots below.

```
> pdf(file="LA12.pdf")
> qqmath(~resid(mmod)|sex,psid)
> dev.off()
```

Longitudinal data analysis



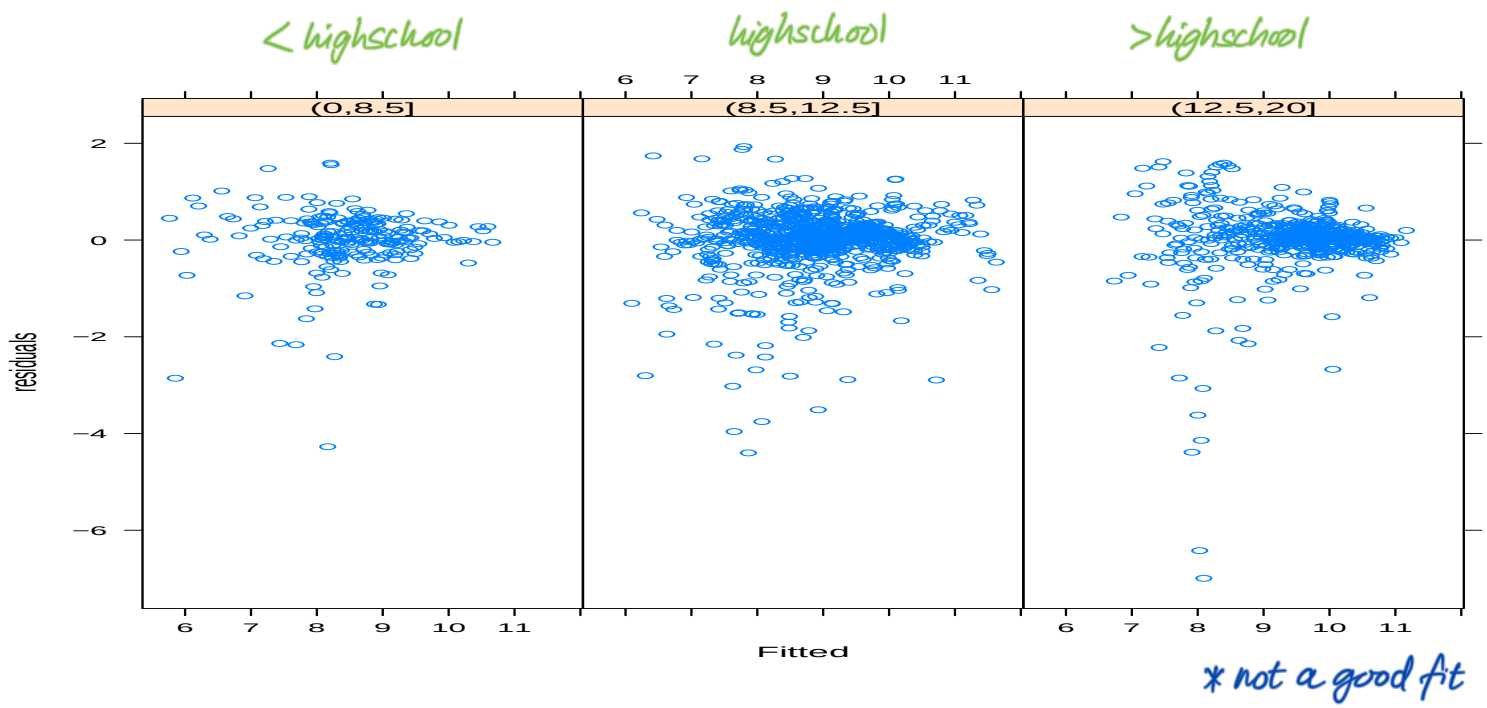
- We see that the residuals are not normally distributed, but have a long tail for the lower incomes.
- We should consider changing the log transformation on the response.
- Furthermore, we see that there is greater variance in the female incomes.
- This suggests a modification to the model.
- We can make the same plot broken down by subject although there will be rather too many plots to be useful.

Longitudinal data analysis

- Plots of residuals and fitted values are also valuable.
- We have broken education into three levels: less than high school, high school or more than high school.

```
> pdf(file="LA13.pdf")
> xyplot(resid(mmod)~fitted(mmod)|
cut(educ,c(0,8.5,12.5,20)),psid,layout=c(3,1),xlab="Fitted",ylab="residuals")
> dev.off()
```

Longitudinal data analysis



Again, we can see evidence that a different response transformation should be considered.