STAT 528 - Advanced Regression Analysis II

Linear Mixed Models (part II)

Daniel J. Eck
Department of Statistics
University of Illinois

Learning Objectives Today

- Linear Mixed Models examples

Split plots

The design arises as a result of restrictions on full randomization.

For example, consider an agricultural experiment that studies the effects of plant varieties and irrigation techniques. It may be difficult to vary irrigation techniques over a field while it is easy to plant different plant varieties across a field.

To overcome this challenge we consider a split plot design in which a sample of main plots (fields) each receive one irrigation technique at random.

Each main plot is then divided into a number of split plots equal to the number of levels of plant varieties under study. Then each plant variety is randomly assigned to each split plot at random.

Generally speaking, split plot designs arise when one factor is easy to change and another factor takes much more effort to manipulate.

Here is a simple conceptual visual for split plot designs..

R example irrigation split plot design

In this example we want to determine which irrigation and plant variety combination produces the highest yield.

There are four irrigation methods and two plant varieties under consideration. Eight fields were available, but only one type of irrigation may be applied to each field.

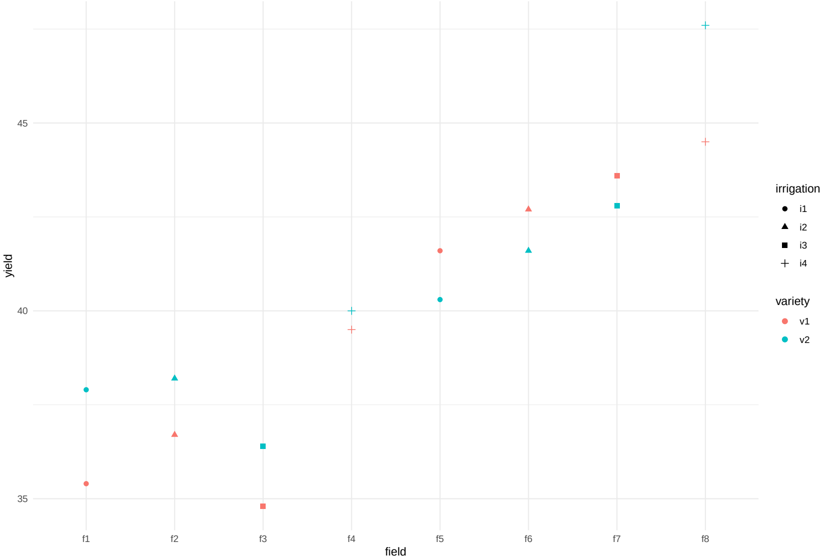
The whole plot factor is the irrigation method which should be randomly assigned to each of the fields. The fields may be divided into two halves with each plant variety randomly assigned to each of the halves.

```
library(lme4)
library(faraway)

head(irrigation, 3)
```

```
##   field irrigation variety yield
## 1    f1         i1      v1  35.4
## 2    f1         i1      v2  37.9
## 3    f2         i2      v1  36.7
```

```
ggplot(irrigation, aes(y=yield, x=field, shape=irrigation, color= variety)) +
  geom_point(size = 2) +
  theme_minimal()
```



Both irrigation and plant variety are fixed effects, but the field is clearly a random effect (**Why so?**).

We must also consider the interaction between field and variety, which is necessarily also a random effect because one of the two components is random.

The fullest model that we might consider is:

$$y_{ijk} = \mu + r_i + v_j + (rv)_{ij} + f_k + (vf)_{jk} + \varepsilon_{ijk},$$

where

- μ is a fixed model intercept,
- r_i is the fixed effect for irrigation that ranges over $i \in \{1, \ldots, 4\}$ levels,
- v_j is the fixed effect for plant variety that ranges over $j \in \{1, 2\}$ levels,
- $(rv)_{ij}$ is the fixed effect interaction between irrigation and plant variety with an index ranging over i and j,
- f_k is the random effect for field for which there are 8 realizations each indexed by k,
- $(vf)_{jk}$ is the random effect interaction between plant variety and field with an index ranging over j and k, and
- ε is an error term for each observation.

We consider normal distributions for the random effect terms that are indexed by single variance parameters σ_f^2, σ_{vf}^2, and σ_ε^2.

Note that there is no irrigation and field interaction terms in this model. This is because it would not be possible to estimate this effect since only one type of irrigation method is assigned to a field.

We try to fit this model as follows:

```r
lmod <- lmer(yield ~ irrigation * variety + (1|field) +
             (1|field:variety), data = irrigation)
```

```
Error: number of levels of each grouping factor must
be < number of observations (problems: field:variety)
```

This failed because it is not possible to distinguish the variety within field variation.

We resort to a simpler model that omits the variety by field interaction random effect:

$$y_{ijk} = \mu + r_i + v_j + (rv)_{ij} + f_k + \varepsilon_{ijk}$$

```r
lmod <- lmer(yield ~ irrigation * variety + (1|field),
             data = irrigation)
```

```r
sumary(lmod)
```

```
## Fixed Effects:
##                           coef.est coef.se
## (Intercept)               38.50     3.03  
## irrigationi2               1.20     4.28  
## irrigationi3               0.70     4.28  
## irrigationi4               3.50     4.28  
## varietyv2                  0.60     1.45  
## irrigationi2:varietyv2 -0.40     2.05  
## irrigationi3:varietyv2 -0.20     2.05  
## irrigationi4:varietyv2  1.20     2.05  
##
## Random Effects:
##  Groups   Name        Std.Dev.
##  field    (Intercept) 4.02    
##  Residual             1.45    
## ---
## number of obs: 16, groups: field, 8
## AIC = 65.4, DIC = 91.8
## deviance = 68.6
```

We can see that the largest variance component is that due to the field effect, with $\hat{\sigma}_f = 4.03$ compared to $\hat{\sigma}_\varepsilon = 1.45$.

The relatively large standard errors compared to the fixed effect estimates suggest that there may be no significant fixed effects.

Wait, no p-values? See Ben Bolker's GLMM FAQ document and Douglas Bates's commentary on this issue.

We can check this sequentially using a backwards model selection procedure with F-tests with adjusted degrees of freedom:

```
library(pbkrtest)
lmoda <- lmer(yield ~ irrigation + variety + (1|field), data=irrigation)
KRmodcomp(lmod, lmoda)
```

```
## large : yield ~ irrigation * variety + (1 | field)
## small : yield ~ irrigation + variety + (1 | field)
##         stat    ndf    ddf F.scaling p.value
## Ftest 0.2452 3.0000 4.0000         1  0.8612
```

We find there is no significant interaction term.

We can now test each of the main effects starting with the fixed effect for plant variety:

```
lmodi <- lmer(yield ~ irrigation + (1|field), irrigation)
KRmodcomp(lmoda, lmodi)
```

```
## large : yield ~ irrigation + variety + (1 | field)
## small : yield ~ irrigation + (1 | field)
##         stat    ndf    ddf F.scaling p.value
## Ftest 1.5782 1.0000 7.0000         1  0.2493
```

Dropping variety from the model seems reasonable since the p-value of 0.25 is large. We can test irrigation in a similar manner:

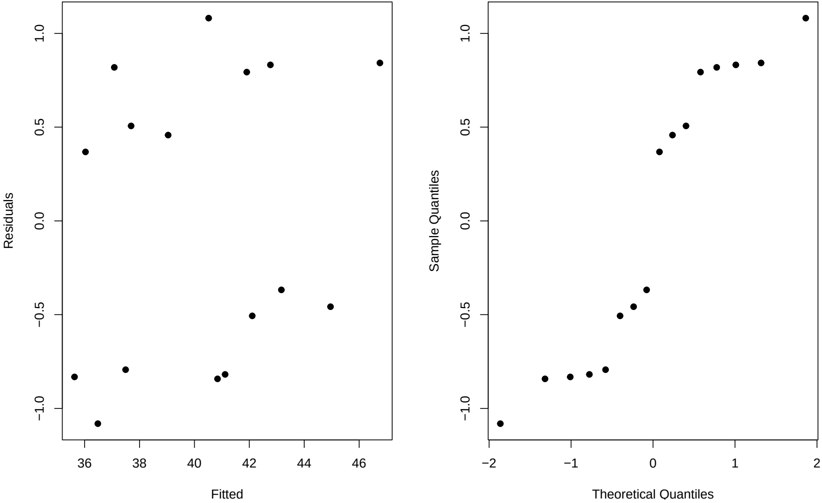
```
lmodv <- lmer(yield ~ variety + (1|field), irrigation)
KRmodcomp(lmoda, lmodv)
```

```
## large : yield ~ irrigation + variety + (1 | field)
## small : yield ~ variety + (1 | field)
##         stat    ndf    ddf F.scaling p.value
## Ftest 0.3882 3.0000 4.0000         1  0.7685
```

Irrigation also fails to be significant.

We should check the diagnostic plots to make sure there is nothing amiss:

```
par(mfrow = c(1,2))
plot(fitted(lmod), residuals(lmod), xlab="Fitted", ylab="Residuals", pch = 19)
qqnorm(residuals(lmod), main="", pch = 19)
```



We can see that there is no problem with the nonconstant variance, but that the residuals indicate a bimodal distribution caused by the pairs of observations in each field. This type of divergence from normality is unlikely to cause any major problems with the estimation and inference.

We can test the random effects like this:

```
library(RLRsim)
exactRLRT(lmod)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 6.1118, p-value = 0.0095
```

We see that the fields do seem to vary as the result is clearly significant.

Example of nested random effects

Consistency between laboratory tests is important and yet the results may depend on who did the test and where the test was performed.

In an experiment to test levels of consistency, a large jar of dried egg powder was divided up into a number of samples.

Because the powder was homogenized, the fat content of the samples is the same, but this fact is withheld from the laboratories.

Follow the randomness

Four samples were sent to each of six laboratories.

Two of the samples were labeled as G and two as H, although in fact they were identical.

The laboratories were instructed to give two samples to two different technicians.

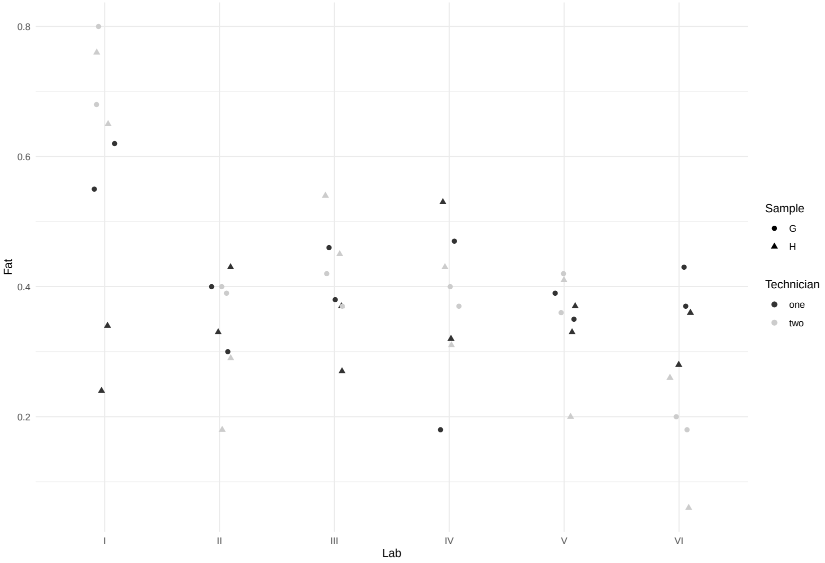
The technicians were then instructed to divide their samples into two parts and measure the fat content of each.

So each laboratory reported eight measures, each technician four measures, that is, two replicated measures on each of two samples. The data comes from Bliss (1967):

```r
data(eggs)
head(eggs, 3)
```

```
##    Fat Lab Technician Sample
## 1 0.62   I        one      G
## 2 0.55   I        one      G
## 3 0.34   I        one      H
```

```
ggplot(eggs, aes(y=Fat, x=Lab, color=Technician, shape=Sample)) +
  geom_point(size = 2, position = position_jitter(width=0.1, height=0.0)) +
  scale_color_grey() +
  theme_minimal()
```



Our model is

$$y_{ijkl} = \mu + L_i + T_{ij} + S_{ijk} + \varepsilon_{ijkl}$$

where,

- μ is a fixed model intercept,
- L_i is the random effect for lab for which there are 6 realizations each indexed by i,
- T_{ij} is the random effect for technician nested within lab for which there are 2 realizations each indexed by j within each lab i,
- S_{ijk} is the random effect for sample nested within technician nested within lab. There are four realizations each indexed by k by each technician j within each lab i, and
- ε is an error term for each observation.

This can be fit with:

```
cmod <- lmer(Fat ~ 1 + (1|Lab) + (1|Lab:Technician) +  (1|Lab:Technician:Sample),
             data=eggs)
sumary(cmod)
```

```
## Fixed Effects:
## coef.est  coef.se
##     0.39     0.04
##
## Random Effects:
##  Groups                Name        Std.Dev.
##  Lab:Technician:Sample (Intercept) 0.06
##  Lab:Technician        (Intercept) 0.08
##  Lab                   (Intercept) 0.08
##  Residual                          0.08
## ---
## number of obs: 48, groups: Lab:Technician:Sample, 24; Lab:Technician, 12; Lab, 6
## AIC = -54.2, DIC = -73.3
## deviance = -68.8
```

So we have that the estimated variance components are all similar in magnitude.

The lack of consistency in measures of fat content can be ascribed to variance between labs, technicians, measurement due to different labeling, and measurement error.

Although the data has a natural hierarchical structure which suggests a particular order of testing, we might reasonably wonder which of the components contribute substantially to the overall variation.

A look at the confidence intervals reveals the problem:

```
confint(cmod, method="boot")
```

```
##                   2.5 %     97.5 %
## .sig01      0.00000000 0.09628667
## .sig02      0.00000000 0.13598404
## .sig03      0.00000000 0.14699799
## .sigma      0.05966933 0.10614784
## (Intercept) 0.31175286 0.47064824
```

We might drop any of the three random effect terms but it is not possible to be sure which is best to go.

It is safest to conclude there is some variation in the fat measurement coming from all three sources.

Repeated measures and longitudinal data

In repeated measures designs there are several individuals (or units) under study and multiple measurements are taken repeatedly on each individual.

When these repeated measurements are taken over time, it is called a longitudinal study or, in some applications, a panel study.

Typically there are various covariates concerning the individual that are also recorded and interest centers on how the response depends on these covariates over time. 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;
- a random effect, which expresses the variation between individuals;
- and an error, which is due to measurement or unrecorded variables.

In a repeated measures design we will suppose that each individual has a response y_i which is now a vector of length n_i and is modeled conditionally on the random effect b_i as

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

Note that this is a very similar model used in the linear mixed-effects model construction above, with the exception that we now allow for the individuals to have a more general covariance structure Λ_i.

As before, we will assume that $b_i \sim N(0, \sigma^2 D)$ so that

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

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

Now suppose that we have N individuals and assume that the measurement errors and random effects between individuals are uncorrelated. Then we can combine the data as:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \qquad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix},$$

and

- $\tilde{D} = \text{diag}(D, D, \ldots, D)$,
- $Z = \text{diag}(Z_1, Z_2, \ldots, Z_N)$,
- $\Sigma = \text{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_N)$, and
- $\Lambda = \text{diag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_N)$, where

$y_i \in \mathbb{R}^{n_i}$, $X_i \in \mathbb{R}^{n_i \times p}$, $\beta \in \mathbb{R}^p$, $b_i \in \mathbb{R}^q$, $Z_i \in \mathbb{R}^{n_i \times q}$, and the rest of the quantities in the above follow from these specifications.

With this setup we can write the model as

$$Y \sim N(X\beta, \Sigma) \qquad \text{where} \qquad \Sigma = \sigma^2(\Lambda + Z\tilde{D}Z^T).$$

The log-likelihood for the data is then computed as previously and estimation, testing, standard errors and confidence intervals all follow using standard likelihood theory as before.

There is no strong distinction between repeated measures methodology and the methodology of linear mixed-effects models.

Example: Panel Study of Income Dynamics

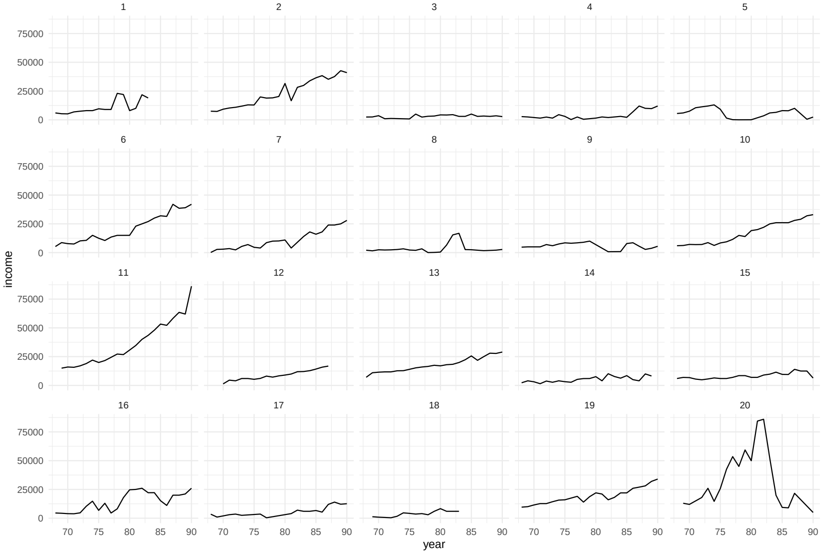
The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals described in Hill (1992).

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:

```
library(tidyverse)
head(psid)
```

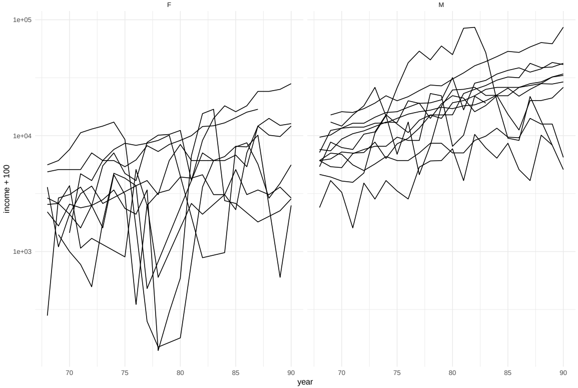
```
##   age educ sex income year person
## 1  31   12   M   6000   68      1
## 2  31   12   M   5300   69      1
## 3  31   12   M   5200   70      1
## 4  31   12   M   6900   71      1
## 5  31   12   M   7500   72      1
## 6  31   12   M   8000   73      1
## first 20 observations
psid20 <- filter(psid, person <= 20)
ggplot(psid20, aes(x=year, y=income)) +
  geom_line() + facet_wrap(~ person) + theme_minimal()
```



We see that some individuals have a slowly increasing income, typical of someone in steady employment in the same job. Other individuals have more erratic incomes. We can also show how the incomes vary by sex.

Income is more naturally considered on a log-scale:

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



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

We see that men's incomes are generally higher and less variable while women's incomes are more variable, but are perhaps increasing more quickly.

We could fit a line to each subject starting with the first:

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

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

Note that we have centered the predictor year at its median value so that the intercept will represent the predicted log income in 1978 and not the year 1900 which is nonsense (any sensible location transformation would be appropriate, but centering with respect to the median is perhaps most appropriate).

We now fit a line for all the subjects and plot the results:

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

The `lmList` command fits a linear model to each group within the data, here specified by person.

We can test the difference in income growth rates for men and women:

```
psex <- psid$sex[match(1:85,psid$person)]
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
## 95 percent confidence interval:
##  -0.05916871 -0.00507729
## sample estimates:
##  mean of x  mean of y
## 0.05691046 0.08903346
```

We see that women have a significantly higher growth rate than men.

We can also compare the incomes at the intercept (which is 1978):

```
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
## 95 percent confidence interval:
##  0.8738792 1.4322218
## sample estimates:
## mean of x mean of y
##  9.382325  8.229275
```

We see that men have significantly higher incomes.

Response feature analysis

This is an example of a response feature analysis.

It requires choosing an important characteristic. We have chosen two here: the slope and the intercept. For many datasets, this is not an easy choice and at least some information is lost by doing this.

Response feature analysis is attractive because of its simplicity. By extracting a univariate response for each individual, we are able to use a wide array of well-known statistical techniques.

However, it is not the most efficient use of the data as all the additional information besides the chosen response feature is discarded. Notice that having additional data on each subject would be of limited value.

Suppose that the income change over time can be partly predicted by the subject's age, sex and educational level.

We do not expect a perfect fit. 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, let us initially assume that this error is homogeneous and uncorrelated, that is, $\Lambda_i = I$.

We also center the year to aid interpretation as before. We may express these notions in the model:

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

This model can be written as

$$\begin{aligned}\log(\text{income})_{ij} = \mu + \beta_y \text{year}_j + \beta_g \text{sex}_i + \beta_{yg} \text{sex}_i * \text{year}_j + \beta_e \text{education}_i \\ + \beta_a \text{age}_i + \gamma_i^0 + \gamma_i^1 \text{year}_j + \varepsilon_{ij}\end{aligned}$$

where i indexes the individuals and j indexes the years, and log is the natural logarithm. We have:

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

```r
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
```

We now analyze this summary table. Lets start with the fixed effects. We see that:

- income increases about 10% ($\exp(0.104) \approx 0.104$) for each additional year of education. We see that age does not appear to be significant.
- For females, the reference level in this example, income increases about 8.5% a year, while for men, it increases about 8.5% - 2.6% = 5.9% a year.
- We see that, for this data, the incomes of men are exp(1.150) = 3.16 times higher (far higher!).

Now the random effects. We see that:

- 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 0.187 ($\text{cor}(\gamma^0, \gamma^1)$).
- There is some additional variation in the measurement not so far accounted for having standard deviation of 0.684 ($\text{sd}(\varepsilon_{ij})$).
- We see that the variation in increase in income is relatively small while the variation in overall income between individuals is quite large.
- Furthermore, given the large residual variation, there is a large year-to-year variation in incomes.