LINEAR MIXED EFFECT MODELS.

1. Motivation.

The objective of a statistical model is to have a mathematical formula that describes the relationship in the data. Using linear regression we assumed that the dependent variable was linearly related to the covariates in an additive way. This assumes that each observation is independant; however they may well be some inter-dependence in the responses in relation to some factor.
To deal with this we add a **random effect** into the model which allows us to *assume* a different *baseline* response value for each factor. We model the individual differences in relation to each factor by assuming different **random intercepts** for each response. Such a model is named a **mixed** model due to the fact that it contains the usual *fixed* effects as seen in linear regression, and one or more *random* effects, essentially giving some structure to the error term characterizing variation due to some factor level.

1.1. **The Data.** The Data being analysed consist of exam results (%) for 91 students, for each student we also have their sex *(Male (n=51), Female (n=40))*, degree program *(Science, Joint, Arts)*, and their previous homework result *(hw1)*.
Shown below are the data for the first 6 students and the command in R to obtain them.

```
> head(data)
```

	exam	sex	hw1	hw2	hw3	hw4	degree	high
1	34	F	18	41	27	10	arts	0
2	59	F	35	35	75	75	joint	0
3	35	F	58	51	37	42	arts	0
4	62	F	87	38	78	33	joint	0
5	86	F	90	98	90	92	joint	1
6	29	F	22	38	0	41	Science	0

2. Model.

The assumptions, for a linear mixed effects model,

- The explanatory variables are related **linearly** to the response.
- The errors have constant variance.
- The errors are independent.
- The errors are Normally distributed.

2.1. Checking the assumptions.

2.1.1. *How to check the assumptions.*

- Plotting the residuals against the explanatory variable will indicate if the wrong model has been fitted (i.e. higher order terms are needed) or if there is some dependence on some other explanatory variable. If this is the case some obvious patterning will be visible in the plot.
- Plotting the residuals in order, any trend visible may indicate *seasonal pattern* or *autocorrelation*.
- Plotting the residuals against the fitted values will indicate if there is non-constant error variance, i.e. if the variance increases with the mean the residuals will fan out as the fitted value increases. Usually transforming the data, or using another distribution will help.
- A Normal probability plot, histogram of the residuals or say a Wilk-Shapiro test will indicate if the normality assumption is valid, however high non-normality should have been picked up from exploring the data initially.

All the relevant code and interpretation to do this is included in the *"Linear Regression"* eample.

2.2. The Equation.

$$Y_i = (\beta_0|factor) + \beta_1 x_{1i} + \beta_{2i} + ... + \beta_{pi} + \epsilon_i$$

Using R notation here we are assuming an intercept that's different for each *factor*. To see if this is the case for our data we could fit two models, to the Male students and Female students data seperately and asses if the slopes of the *"best fitting line"* differ. However just for illistrarive purposes we will go on and fit a linear mixed effects model.

2.2.1. *Interpreting the model.*

```
> library(lme4)
> mod<-lmer(exam ~ (1|sex)+degree*hw1+degree*hw2+degree*hw3+degree*hw4 ,data=data)
> summary(mod)$coefficients
```

```
                     Estimate Std. Error    t value
(Intercept)        8.04526530 6.98388488  1.1519756
degreejoint       -9.34347520 8.38807106 -1.1139003
degreeScience     -5.10644595 9.25251663 -0.5518981
hw1                0.32934647 0.09748766  3.3783403
hw2                0.06594519 0.13997094  0.4711348
hw3                0.31253438 0.13914186  2.2461565
hw4                0.10797152 0.13829986  0.7807059
degreejoint:hw1    0.01372362 0.13012090  0.1054682
degreeScience:hw1 -0.02373421 0.13441718 -0.1765713
degreejoint:hw2    0.10161331 0.16357003  0.6212220
degreeScience:hw2  0.17607780 0.17119481  1.0285230
degreejoint:hw3   -0.06689795 0.16366623 -0.4087462
degreeScience:hw3 -0.14751615 0.17118587 -0.8617309
degreejoint:hw4    0.10868255 0.16621828  0.6538543
degreeScience:hw4  0.05946659 0.17232916  0.3450756
```

The **fixed effects** output, shown above is interpreted in exactly the same way as those in a *"normal"* linear regression model.
The **intercept** estimate of the line, gives the expected value of exam scores when all covariates are zero, i.e. for our model on average the expected exam score increases by 8.04527 if all homework assingments scored zero taking into account the baseline *(i.e (Bachelor degree))* .
The **estimate** column in the **summary** output tells us that there is on average an 0.32935 increase in exam score for an **unit** increase in score for the first homework *(hw1)*, given the other explanatory variables in the model.
The interaction terms can be interpreted as value telling us what the estimated change in relationship is between the students score in each homework and their final exam score dependent on which degree program they are on.

2.2.2. *Testing Hypothesis.* By deafault R doesn't print the associated p-values for each regression coefficient in a mixed effect model, the code below extract the fixed effect regression estimates and performs the usual statistical test which essentially test;

- $H_0 : \beta_0 = 0$ verses $H_1 : \beta_0 \neq 0$ and;
- $H_0 : \beta_1 = 0$ verses $H_1 : \beta_1 \neq 0$

```
> coeffs <- coef(summary(mod))
> p <- pnorm(abs(coeffs[, "t value"]), lower.tail = FALSE) * 2
> cbind(coeffs, "p value" = round(p,3))
```

```
                     Estimate Std. Error    t value p value
(Intercept)        8.04526530 6.98388488  1.1519756   0.249
degreejoint       -9.34347520 8.38807106 -1.1139003   0.265
degreeScience     -5.10644595 9.25251663 -0.5518981   0.581
hw1                0.32934647 0.09748766  3.3783403   0.001
hw2                0.06594519 0.13997094  0.4711348   0.638
hw3                0.31253438 0.13914186  2.2461565   0.025
hw4                0.10797152 0.13829986  0.7807059   0.435
degreejoint:hw1    0.01372362 0.13012090  0.1054682   0.916
degreeScience:hw1 -0.02373421 0.13441718 -0.1765713   0.860
degreejoint:hw2    0.10161331 0.16357003  0.6212220   0.534
degreeScience:hw2  0.17607780 0.17119481  1.0285230   0.304
degreejoint:hw3   -0.06689795 0.16366623 -0.4087462   0.683
degreeScience:hw3 -0.14751615 0.17118587 -0.8617309   0.389
degreejoint:hw4    0.10868255 0.16621828  0.6538543   0.513
degreeScience:hw4  0.05946659 0.17232916  0.3450756   0.730
```

P-values indicate that only the previous homework assignments 1 and 3 are considered important in predeicitng final exam scores *(in our dataset)* as all the associated pvaues are below 0.05, whereas niether degree type of student, intercept, or the interaction terms etc. are considered useful in predicting the response.

2.2.3. *Interpreting the Random effects.* As we model the individual differences in relation to each sex by assuming different **random intercepts** for each response, we interpret the random intercepts in the same way as the fixed effects estimated coefficients. Code below shows how to obtain the coefficient estimates in R for the mixed effect model fitted.

```
> coef(mod)
$sex
  (Intercept) degreejoint degreeScience       hw1        hw2       hw3
F    6.728607   -9.343475     -5.106446 0.3293465 0.06594519 0.3125344
M    9.361924   -9.343475     -5.106446 0.3293465 0.06594519 0.3125344
        hw4 degreejoint:hw1 degreeScience:hw1 degreejoint:hw2 degreeScience:hw2
F 0.1079715      0.01372362       -0.02373421       0.1016133         0.1760778
M 0.1079715      0.01372362       -0.02373421       0.1016133         0.1760778
  degreejoint:hw3 degreeScience:hw3 degreejoint:hw4 degreeScience:hw4
F     -0.06689795        -0.1475162       0.1086825        0.05946659
M     -0.06689795        -0.1475162       0.1086825        0.05946659

attr(,"class")
[1] "coef.mer"
```

From this output we initially note all the *fixed effects* coefficients are the same for the assumed random effects, the only difference is in the estimated intecepts for the random effects. These tell us that for Female students there is an average increase of 6.7286607 in exam score given the other independent varables included in the model, and for Male students theres an average increase of 9.361924 in exam score given the other independent varables included in the model.