Statistical Analysis: Linear Regression and ANOVA

Fotios Siannis

Department of Mathematics
National and Kapodistrian University of Athens

CODEJAM / 1st Workshop on Computational Biology
6-11 April 2025, Athens, Greece

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Presentation Outline

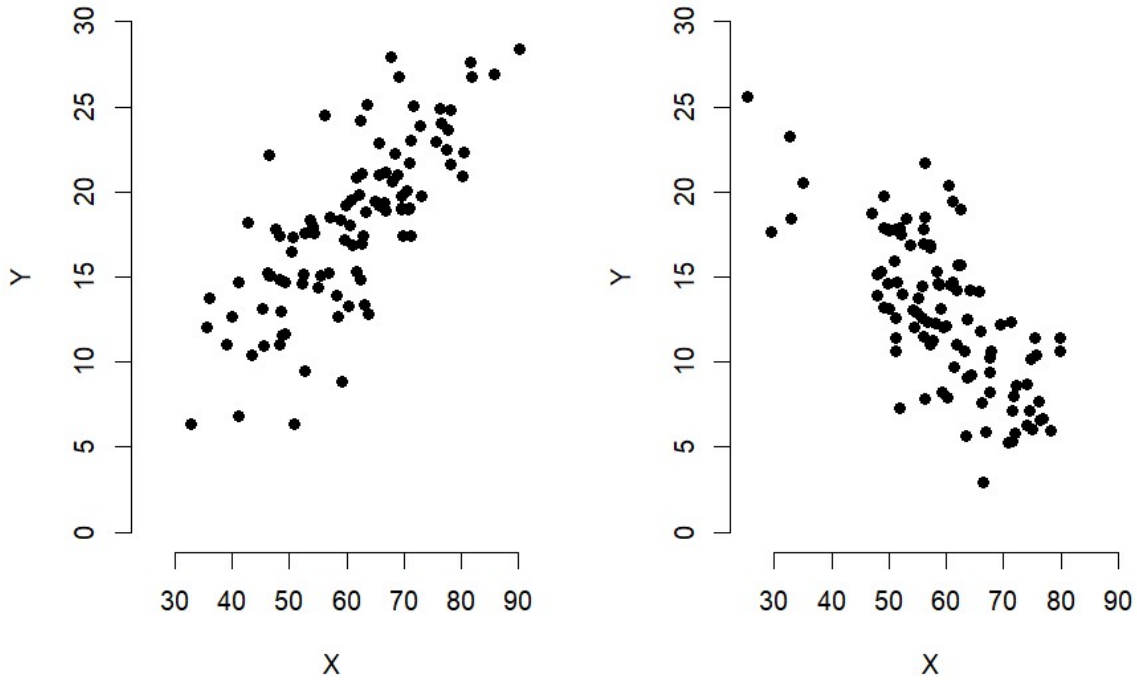
Information

- All information related to this course can be found at **https://eclass.uoa.gr/courses/MATH861/**
- Data used for the examples are part of several R packages, from Kaggle and from our own analyses
- Introductory notes on R can be found on the above site
- For further questions email at **fsiannis@math.uoa.gr** or contact through the above site

Presentation Outline

Data

Assume we have data that look like



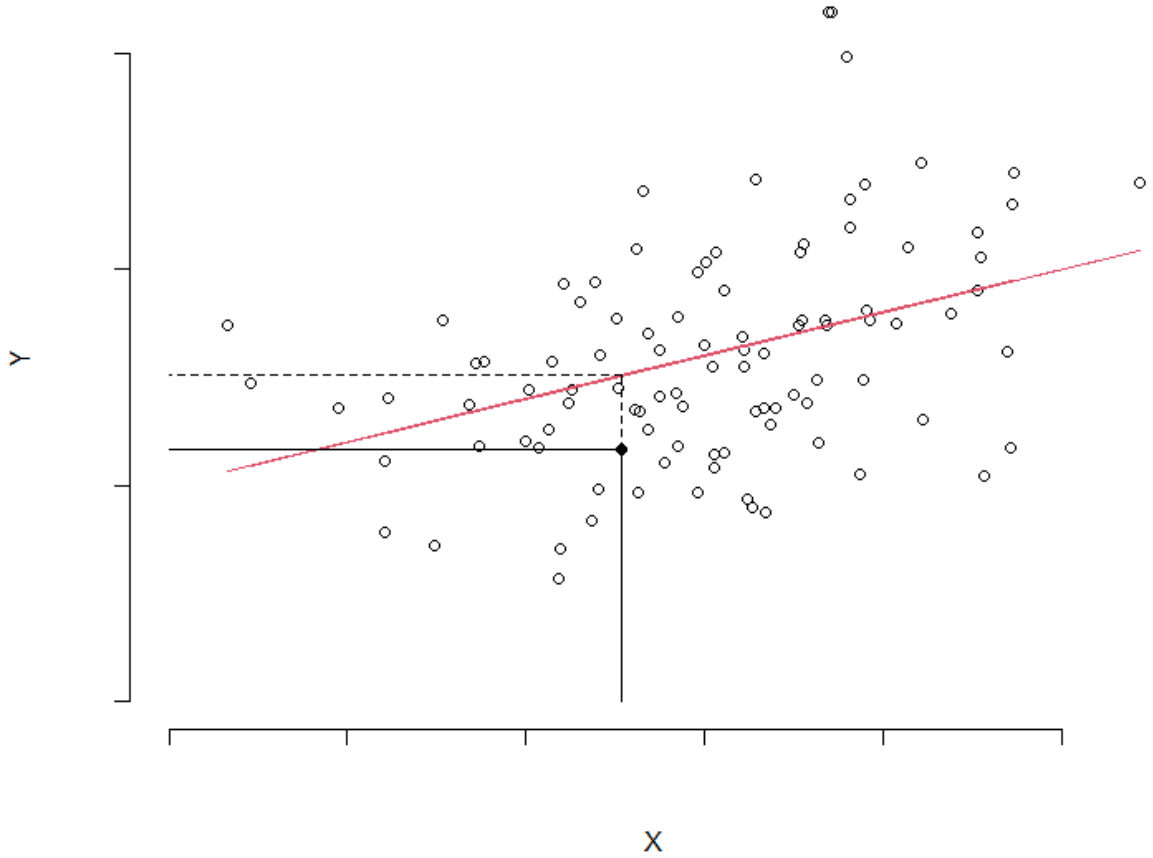
Simple Linear Model

- It is the simplest version of the linear models, with one dependent variable (Y) and only one independent (prognostic) variable (X)
- The model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where:

- β_0: the *intercept*
- β_1: the *slope*
- $i = 1, 2, \ldots, n$ where n the size of the sample



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Main Assumptions

We have pairs of observations (x_i, y_i) for $i = 1, 2, ..., n$, where

- X: fixed variable (not random), called covariate or explanatory variable
- Y: random variable, from population with mean

$$E(Y) = \beta_0 + \beta_1 X,$$

called the response, and

- ϵ: random errors, for which we assume
  (a) $E(\epsilon_i) = 0$
  (b) $Var(\epsilon_i) = \sigma^2$, ie. common variance for all i
  (c) $Cov(\epsilon_i, \epsilon_j) = 0$, ie. independence for i, j
  (d) Usual assumption:

$$\epsilon_i \sim N(0, \sigma^2)$$

Ordinary Least Squares Estimation (OLS)

- In simple linear model we have to estimate β_0, β_1 (and σ^2)
- We take $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the parameters estimates. Then:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  is an estimate of the mean for Y for every value of x_i
- The OLS priciple is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the square distance between y_i and $\hat{y}_i$ to become minimum
- Therefore, we want to minimize the following quantity

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SSE$$

(Residual Sum of Square or Sum of Square Error)

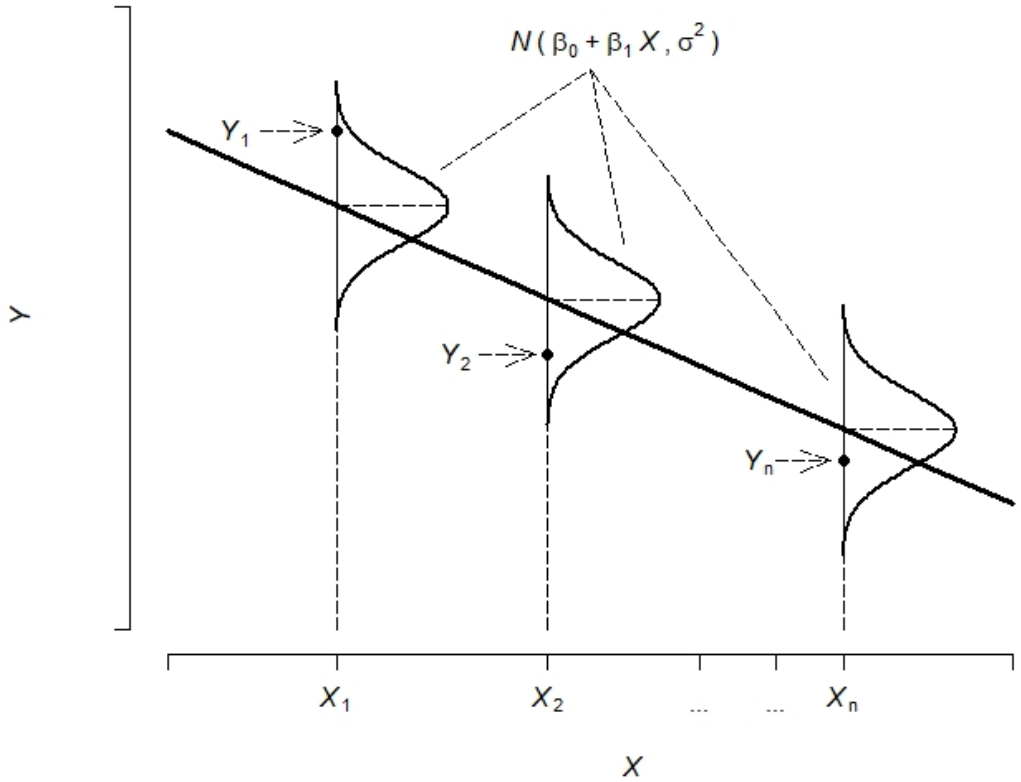
Estimation - Prediction

- We obtain the fitted regression line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X,$$

- The above equation helps us estimate $E(Y)$ for various values of X
- The power (quality) of our estimates depends on the assumption that our model is appropriate (correct) or at least a good approximation of the true model
- "*All models are wrong, but some are useful*" George Box (1976)

- Every $\hat{Y}_i$ obtained from the fitted regression line can be used for:
  (a) Estimation of population mean $E(Y)$ for given value of X
  (b) Prediction of Y that may be obtained in the future for specific value of X
- The point estimates of (a) and (b) are the same, however they differ in how much we believe them to be the "correct" values
- Therefore, is the uncertainty (variability) that makes the difference
- This is reflected on the confidence intervals (CI), where the CI for prediction is greater


$N(\beta_0 + \beta_1 X, \sigma^2)$
Y_1
Y_2
Y_n
Y
X_1
X_2
...
...
X_n
X

Statistical Inference

Assume the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$

Then:

- We have
$$\hat{\beta}_0 \sim N\left(\beta_0, V(\hat{\beta}_0)\right)$$
- and
$$\hat{\beta}_1 \sim N\left(\beta_1, V(\hat{\beta}_1)\right).$$
- We know that:
$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Confidence Interval for β_1

Knowing the distribution of $\hat{\beta}_1$, we can construct a (1-α)100% confidence interval for β_1.

- We know that:

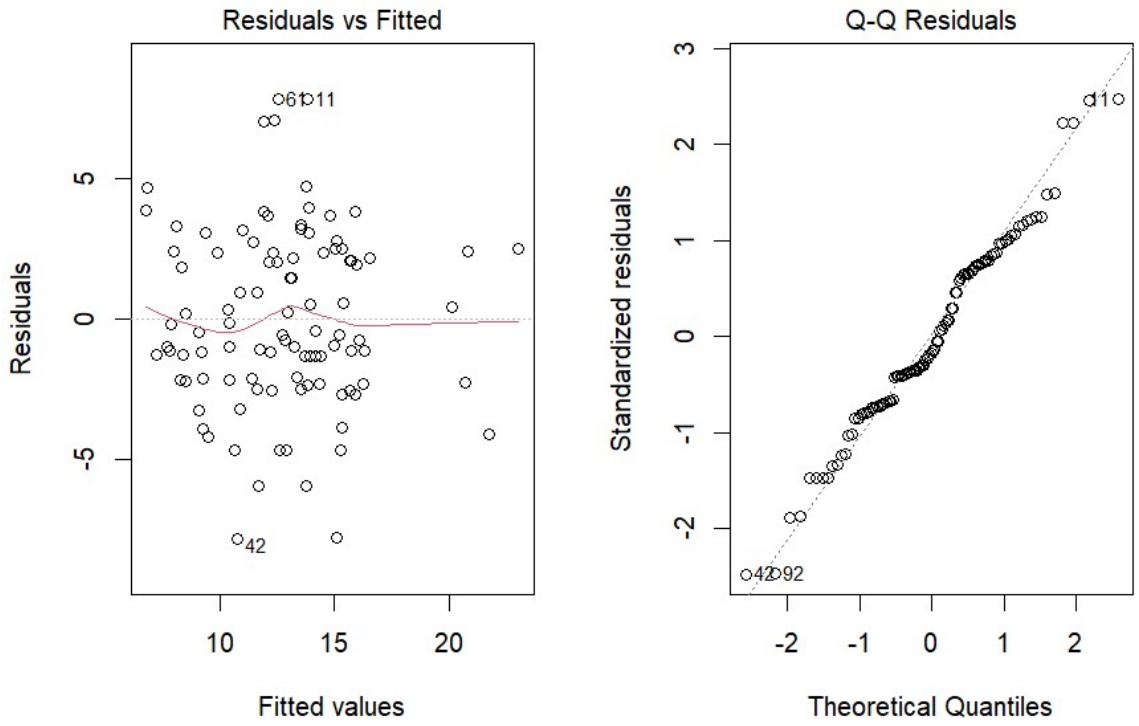
$$\mathbf{t} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{V(\hat{\beta}_1)}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}$$

- This leads to the (1-α)100% confidence interval

$$\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2,\alpha/2}\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Residuals

After fitting the model, it is always wise to check the residuals $(y_i - \hat{y}_i)$ to see if the initial model assumptions hold.



Multivariate Linear Model

- The simple model can be extended to incorporate more than one explanatory variables
- If we have k-variables, the model takes the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, ..., n.$$

  - β_0: the *intercept*
  - β_j: the j-th predictor's regression *slope* $(j = 1, 2, ..., k)$
  - $\epsilon_i \sim N(0, \sigma^2)$

Parameter interpretation: β_j represents the amount by which y_i will change if x_{ji} increase by one unit and all other covariates remain unchanged

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Analysis of Variance (ANOVA)

In the linear model, total variability is expressed:

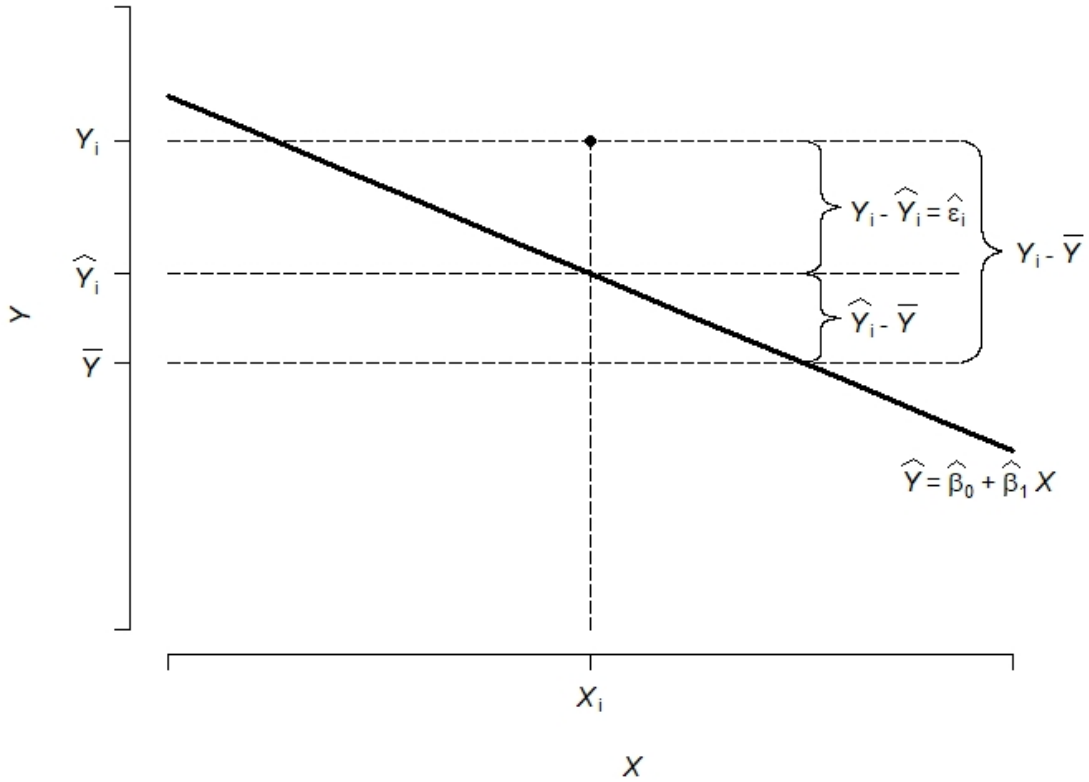
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

We can show that it can be broken down to two components:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 = SSR + SSE$$

- SSR is the sum-of-squares explained by the model (regression sum-of-squares)
- SSE is the sum-of-squares that cannot be explained by the model (sum-of-squares error)

In the simple model the analysis of variance can be represented as follow:



ANOVA TABLE

Source	*DF*	*SS*	MS	$F-test$
Model	k	SSR	$MSR = \frac{SSR}{k}$	$\mathbf{F} = MSR/MSE$
Residuals	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total	$n-1$	SST		

- We can show that MSE is an unbiased estimate of σ^2
- The ratio in the last column of ANOVA

$$\mathbf{F} = \frac{MSR}{MSE}$$

serves as an overall test for the model

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- Quantity $\mathbf{F}$ follows an $\mathcal{F}$ distribution, thus

$$\mathbf{F} \sim \mathcal{F}_{k,n-k-1}$$

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Dummy (Binary) Variables

- A dummy variable d is a 0/1 variable
- Basically, a dummy variable represents the existence of a binary (qualitative) characteristic
- Typical examples: (a) male/female, (b) treatment/no-treatment, (c) age $< 50/ > 50$, where $d = 1$ represents one level and $d = 0$ the other
- Notice that this type of categorization is not unique. A dummy variable can be expressed in many other ways
- A set of dummy variables can be used to express a qualitative characteristic with more than 2 levels

Regression using Dummy Variables

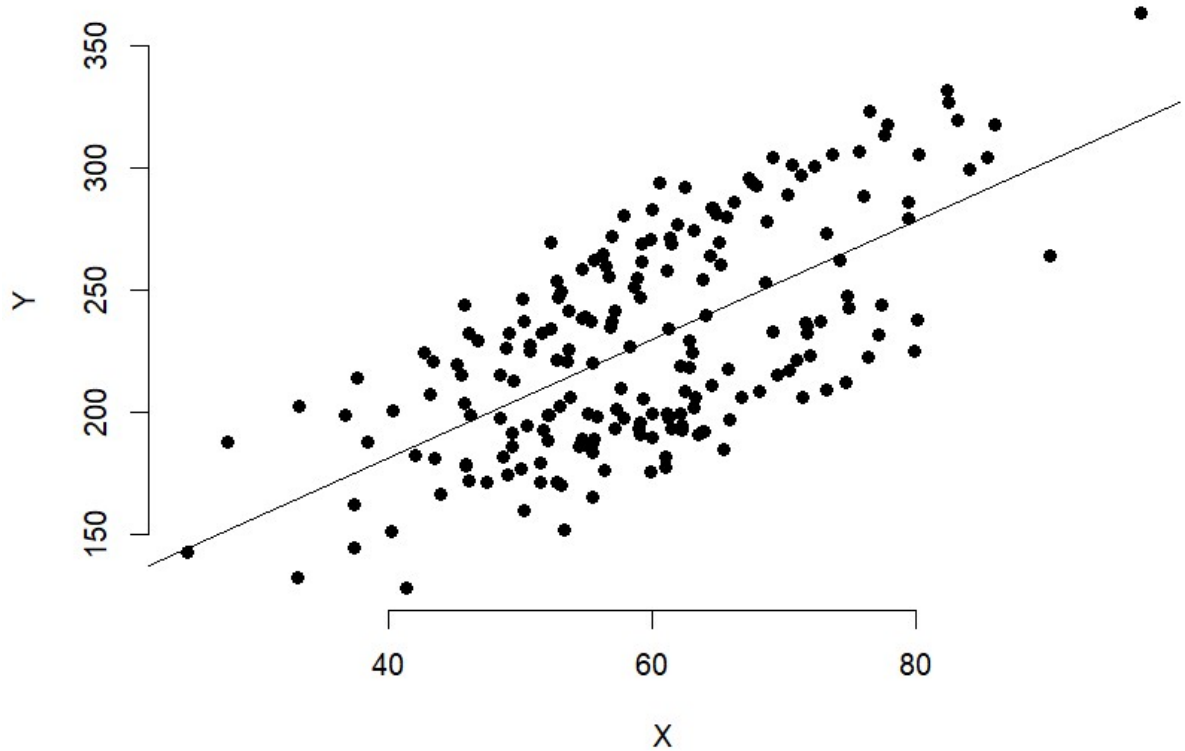
Assume the following model with a binary (dummy) variable

$$y_i = \beta_0 + \beta_1 d_i + \beta_2 x_i + \epsilon_i.$$

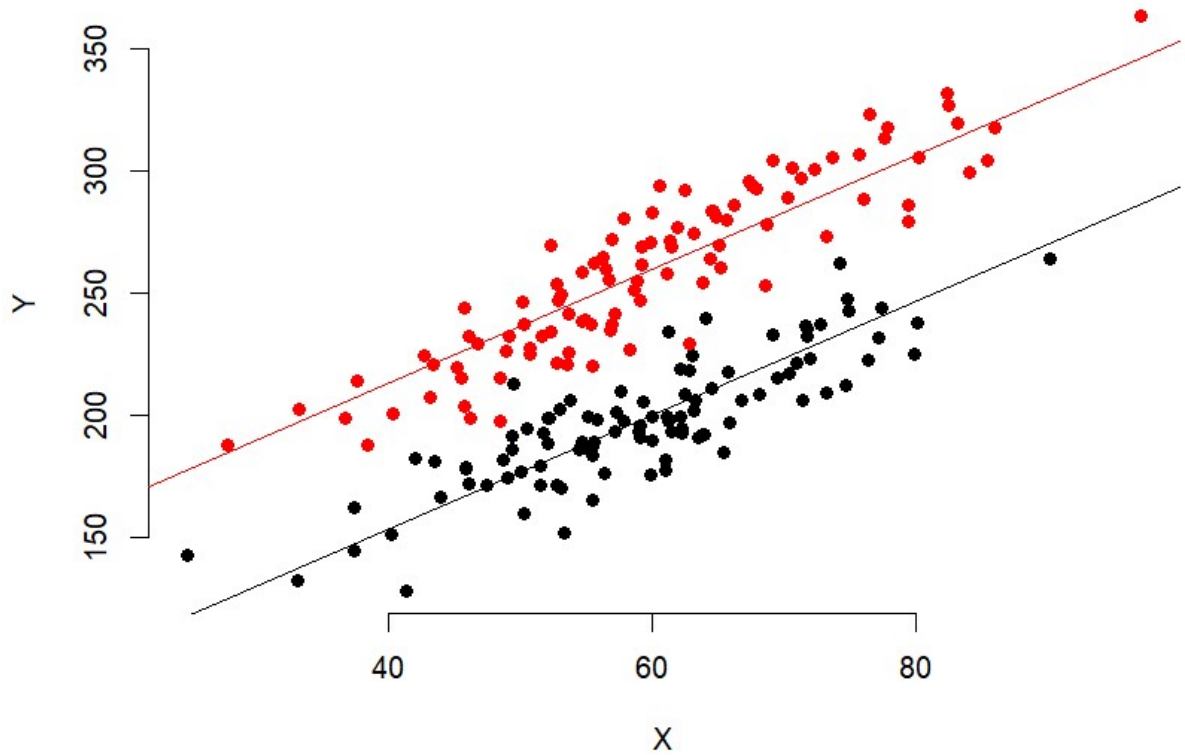
The model can take the following forms:

- If $\underline{d = 0}$, then:
$$y_i = \beta_0 + \beta_2 x_i + \epsilon_i$$
- If $\underline{d = 1}$, then:
$$y_i = (\beta_0 + \beta_1) + \beta_2 x_i + \epsilon_i$$

For example, consider the following case:



In reality we have two populations:



Interpretation

- Observations are split in two groups, according to variable d
- The group with $d = 0$ is called *baseline* group
- Parameter β_1 reflects the expected impact of group with $d = 1$ vs group with $d = 0$, keeping all other variables fixed
- Therefore, a statistical test of the form

$$H_0 : \beta_1 = 0$$

practically tests whether the expected value of y is the same in the two groups

Categorical Variables

- Dummy variables can be used to describe a categorical variable with m-levels, where $m > 2$
- For a variable with m-levels we need $m - 1$ dummy variable
- For example, 'Education Level' (edu) can be a variable with the following levels

	edu
1	High School
2	University Degree
3	Post-Graduate Degree

and the question is about the impact of edu on income (Y)

- Introducing *edu* straight in the model (continuous variable)

$$y = \beta_0 + \beta_1 x_{edu} + \epsilon$$

means that the impact of a 'University Degree' compared to 'High-School' would be exactly the same with the impact of a 'Post-Graduate Degree' compared to 'University Degree'

- To include *edu* in the model, we need 2 dummy variables

	edu	d_1	d_2
1	High School	0	0
2	University Degree	1	0
3	Post-Graduate Degree	0	1

- Baseline group has all dummy variables equal to zero (High School)
- Value $d_1 = 1$ then 'University Degree', else 0
- Value $d_2 = 1$ then 'Post-Graduate Degree', else 0
- Clearly, we cannot have $d_1 = 1$ and $d_2 = 1$ at the same time
- The model takes the form

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \epsilon$$

- Since *edu*=High School (i.e. $d_1 = d_2 = 0$), then

$$y = \beta_0 + \epsilon,$$

where β_0 represents baseline

- β_1 is the effect of 'University Degree' compared to 'High School'

$$y = \beta_0 + \beta_1 + \epsilon,$$

- β_2 is the effect of 'Post-Graduate Degree' compared to 'High School'

$$y = \beta_0 + \beta_2 + \epsilon,$$

- Various tests regarding β_1 and β_2 can test various hypothesis
- Interpretation remains the same in the presence of other explanatory variables

Coeficient of Determination

- The Coeficient of Determination is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

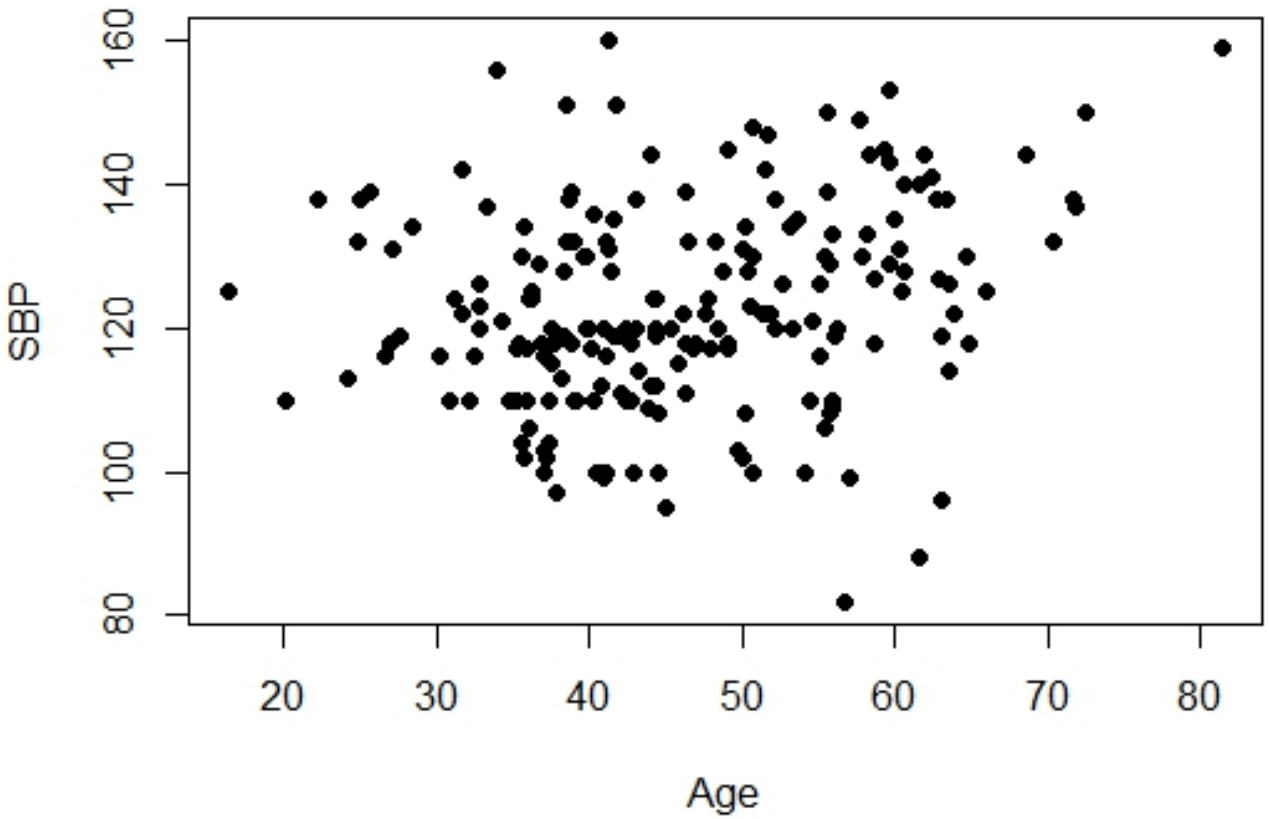
- Represents the percentage of total variability of y_i's explained by the model, and serves as a 'quality' measure of the model
- The Adjusted Coeficient of Determination takes the form

$$R^2_{Adj} = 1 - \frac{MSE}{MST} < R^2$$

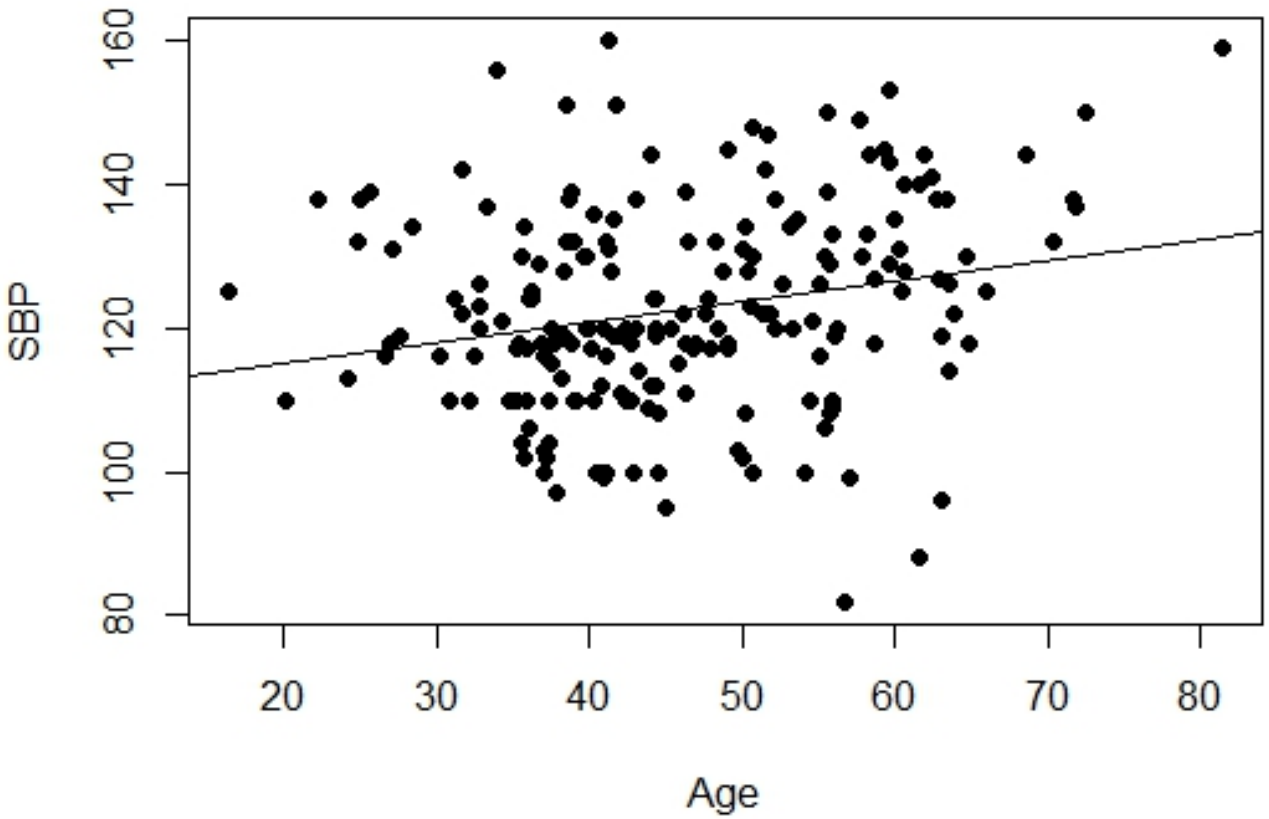
- R^2_{Adj} takes into account the number of observations as well as the number of explanatory variables.
- If sample size is large then $R^2_{Adj} \simeq R^2$

Example

Systolic Blood Pressure (SBP) $\sim$ Age



Fit Regression Line

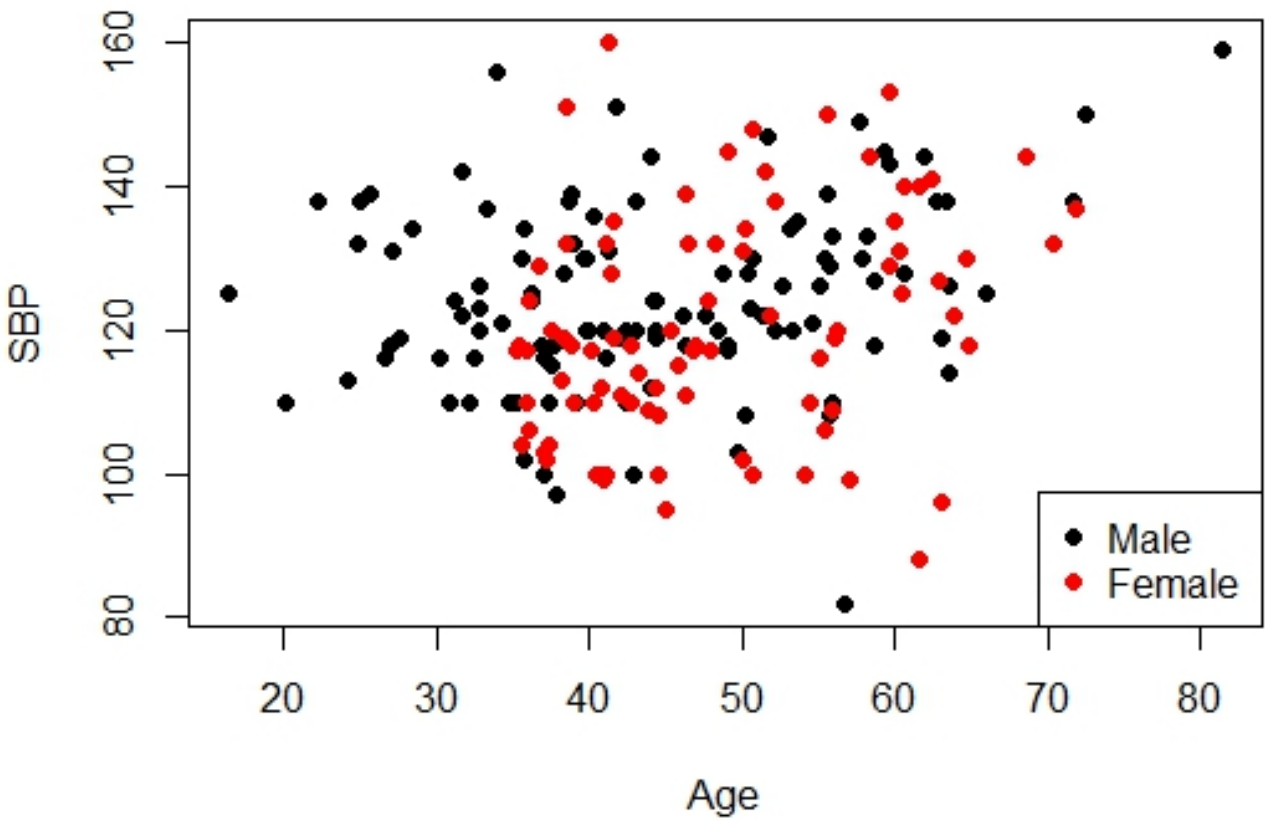


HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

R output:

```
> lm1 = lm(systolic~age3,data=smokew)
> summary(lm1)

Call:
lm(formula = systolic ~ age3, data = smokew)

Residuals:
    Min      1Q  Median      3Q     Max
-43.769  -8.854  -1.211   9.178  38.637

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.58306    4.09653  26.750  < 2e-16 ***
age3          0.28510    0.08652   3.295  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.95 on 198 degrees of freedom
Multiple R-squared:  0.05199,   Adjusted R-squared:  0.0472
F-statistic: 10.86 on 1 and 198 DF,  p-value: 0.001165
```

Systolic Blood Pressure (SBP) $\sim$ AGE + Gender

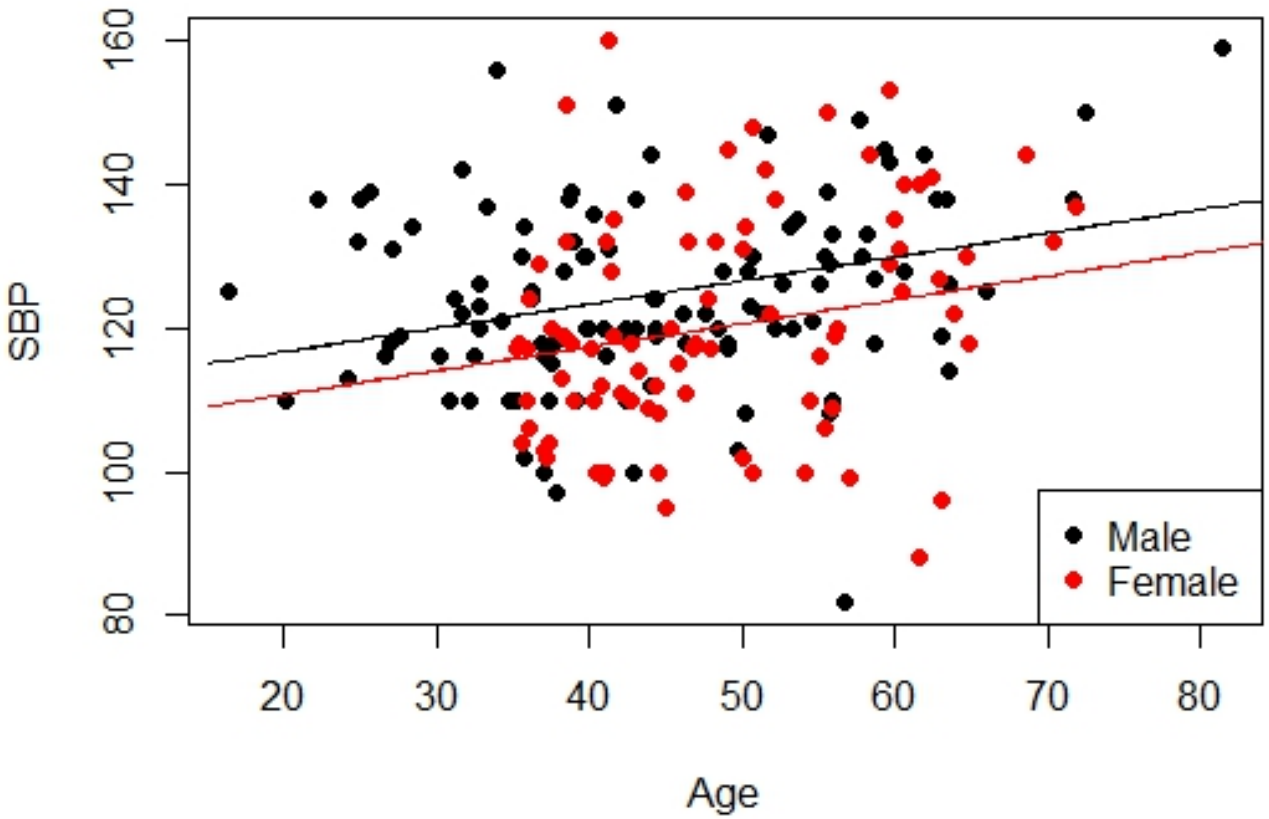


HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

R output:

```
> lm2 = lm(systolic~age3+gender,data=smokew)
> summary(lm2)

Call:
lm(formula = systolic ~ age3 + gender, data = smokew)

Residuals:
    Min      1Q  Median      3Q     Max
-46.852  -7.571  -0.994   9.053  42.220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.15318    4.39971  23.673  < 2e-16 ***
age3          0.32979    0.08608   3.831 0.000171 ***
genderM       5.97539    1.97969   3.018 0.002878 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.67 on 197 degrees of freedom
Multiple R-squared:  0.09389,   Adjusted R-squared:  0.0847
F-statistic: 10.21 on 2 and 197 DF,  p-value: 6.055e-05
```

Fit Regression Lines:



Model Selection

- Many times we have several explanatory variables to consider
- Not all of them are important (explanatory for Y)
- We can adopt model selection strategies so we can come up with the optimal model
  - Forward Selection
  - Backward Elimination
  - Stepwise Selection
- We can select the criterion of our choice (AIC, BIC,p-value, R^2,etc)
- If the number of covariates is not big, we can run a grid search (fit all the models) and choose the optimal one

Presentation Outline

Random Effects

- We need to understand (at least qualitatively) what are the likely sources of random variation
- One possible source is `Random Effects`, when units are sampled at random from a population and various aspects of their behavior may show stochastic variation between units
- We introduce Linear Random Effects model where
  - the response is assumed to be a linear function of exploratory variables with regression coefficients that vary from one individual to the next
  - variability reflects natural heterogeneity due to unmeasured factors

Example: Children birth weight and growth rate

- A RE model is reasonable if the set of coefficients of children can be thought of as a sample from a population
- Association (correlation) arises because we cannot observe the underlying growth curve but we have only imperfect measurements of weight on each infant
- RE models allow for this association on the observations of the same infant over time
- So a (simple) model with random intercept takes the form

$$E(Y_{ij}|U_i) = (\beta_0 + b_i) + \beta_1(\text{time})_{ij}$$

where $(\text{time})_{ij}$ are the times where measurements were taken.

- Typically, b_i follows $N(0, \sigma_b^2)$

Linear Mixed Models

- The Usual Linear Model

$$y = X\beta + \epsilon,$$

where

- $y = (y_1, ..., y_n)'$ is an $n \times 1$ vector of independent observations
- β is a $p \times 1$ vector of unknown parameters
- X an $n \times p$ design (model) matrix
- $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ is an $n \times 1$ vector of independent errors

- The linear mixed model has the following (general) form

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

  - Y_i, β and e as before with
    - $E(\epsilon_i) = 0_n$
    - $Var(\epsilon_i) = W$
  - Matrix Z is a given $n \times q$ matrix (columns of Z a subset of X)
  - b_i is an unobservable random vector of dimensions $q \times 1$, following any multivariate distribution (usually MVN) with
    - $E(b_i) = 0$
    - $Var(b_i) = B$
  - In addition, vectors b_i and e_i are assumed uncorrelated.
  - $E(Y_i) = X_i\beta$
  - $Var(Y_i) = Var(X\beta + Zb + \epsilon) = ZBZ' + W$.

Random Intercept Model

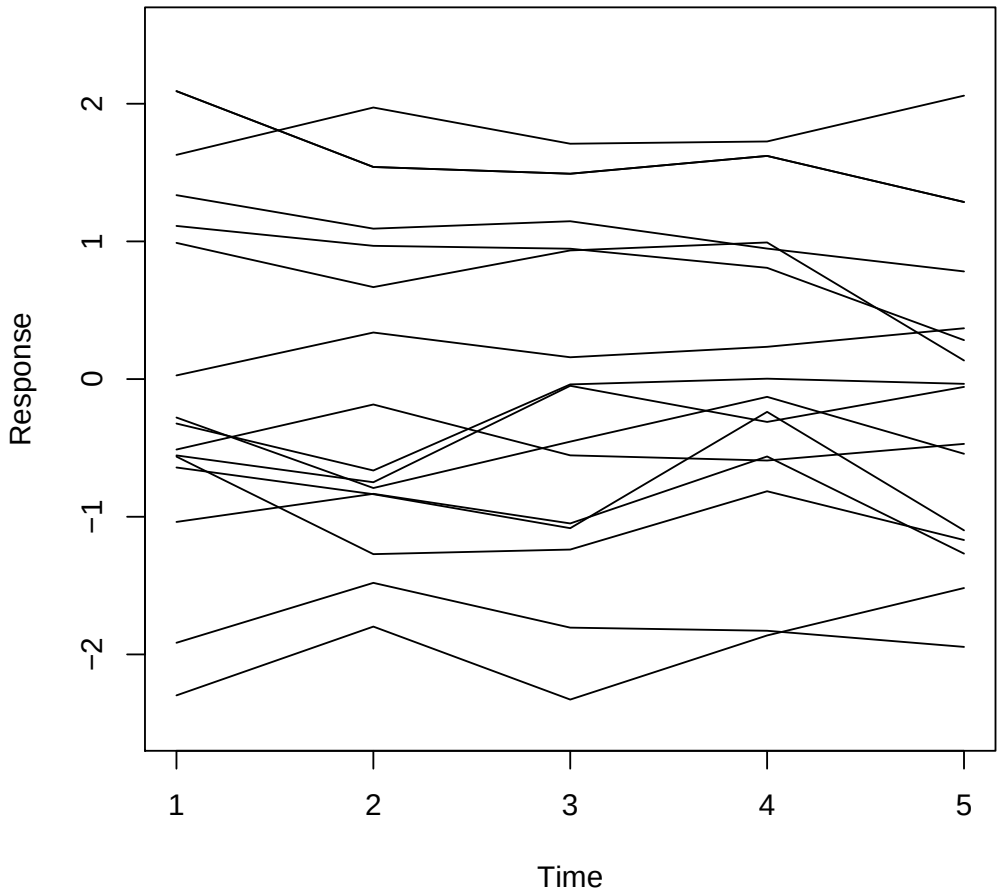
Consider the model

$$Y_{ij} = (\beta_1 + b_i) + X_{ij2}\beta_2 + ... + X_{ijp}\beta_p + \epsilon_{ij}$$

- Each subject's profile appears flat or parallel (over time)
- Observations Y_{ij} vary around a different value for each subject
- These values are the intercepts of the lines for each subject's responses vary around
- b_i represents the deviations of subject's i intercept from the population one (β_1)
- The set of intercepts are sample from the population of intercepts
- This implies that there is *between-subject variability*

Example

Consider observations measured in subjects over time.
Such profiles support the assumption of a model with random intercept.



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

- Furthermore, the variance of Y_{ij} takes the form

$$Var(Y_{ij}) = Var(b_i) + Var(e_{ij}) = \sigma_b^2 + \sigma^2$$

- The covariance between any pair of observations of the same subject

$$Cov(Y_{ij}, Y_{ik}) = Cov(b_i, b_i) = \sigma_b^2.$$

- The correlation between two observations becomes

$$\rho = Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}.$$

- The presence of random effect induce correlation among repeated measurements. This is also known as *intra-class correlation*.

Note: In statistics, the intraclass correlation is a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups. It describes how strongly units in the same group resemble each other. While it is viewed as a type of correlation, unlike most other correlation measures it operates on data structured as groups, rather than data structured as paired observations.

- The model

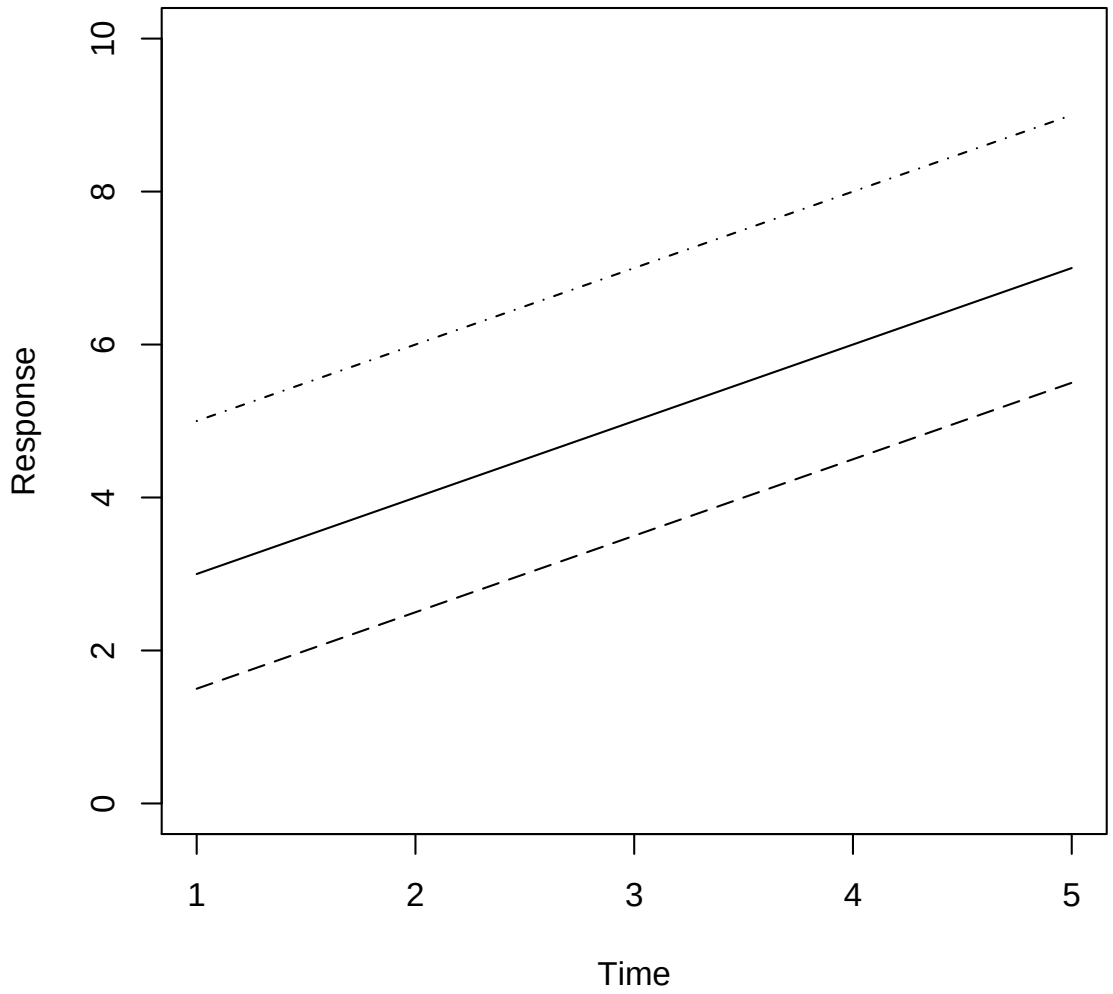
$$E(Y_{ij}|b_i) = X_{ij}'\beta + b_i$$

  is referred to as the *conditional* or *subject specific* mean model

- The model

$$E(Y_{ij}) = X_{ij}'\beta$$

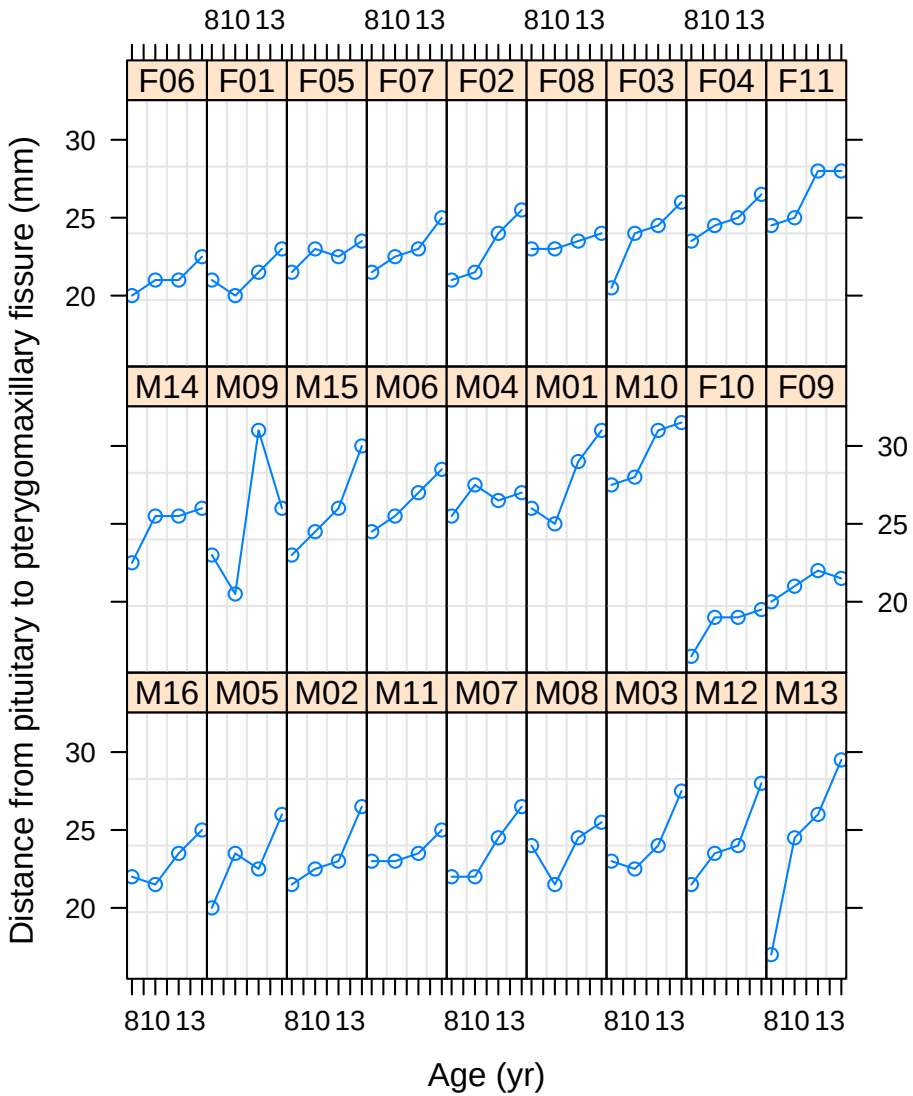
  is referred to as the *marginal* or *population averaged* mean model



Example: `Orthodont` Data [included in `nlme` package]

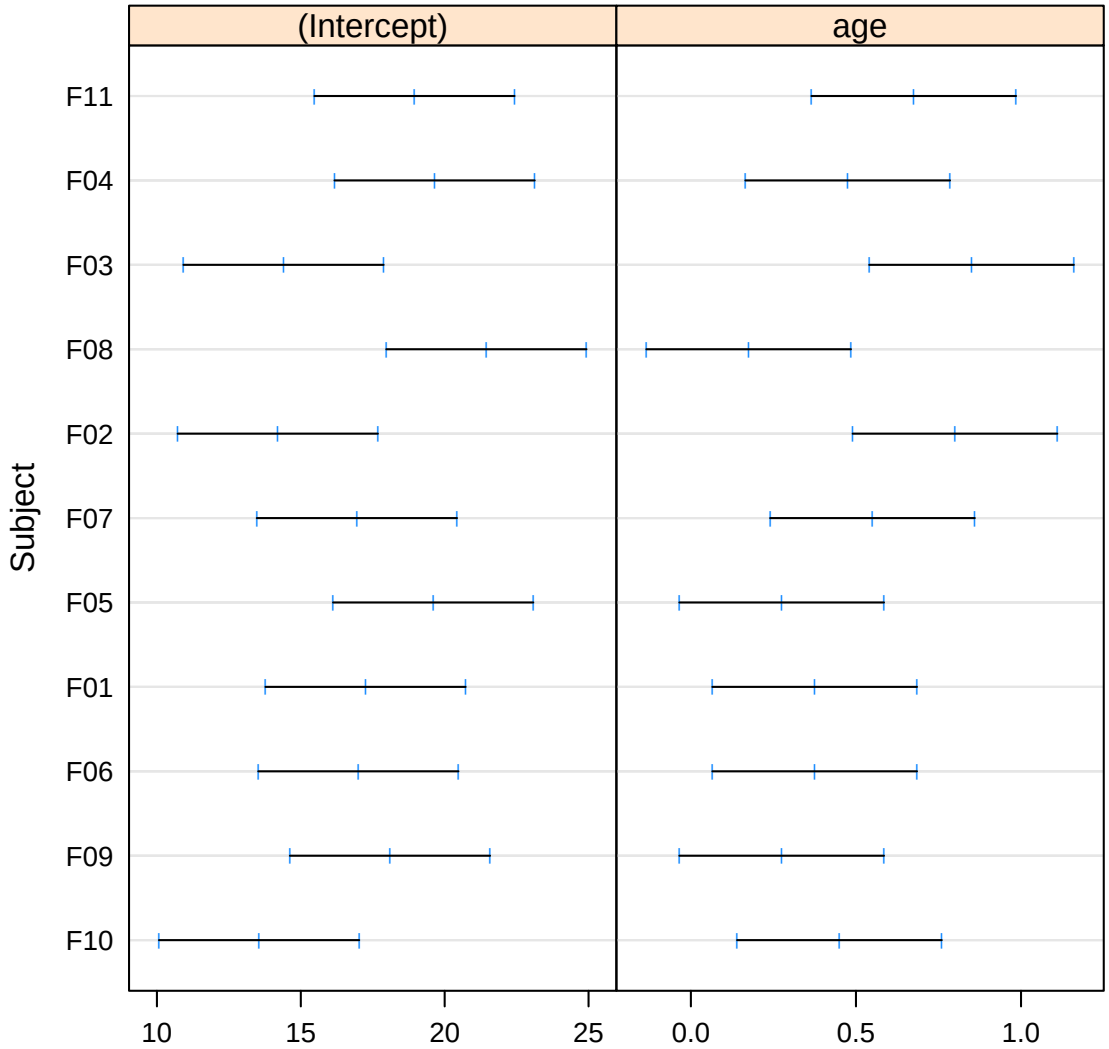
- A set of measurements of the distance from the pituitary gland to the pterygomaxillary fissure taken every 2 years.
- Measurements taken from 8 till 14 years of age.
- We have 27 children: 16 males - 11 females
- Data collected from x-rays.

Plot the data:



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

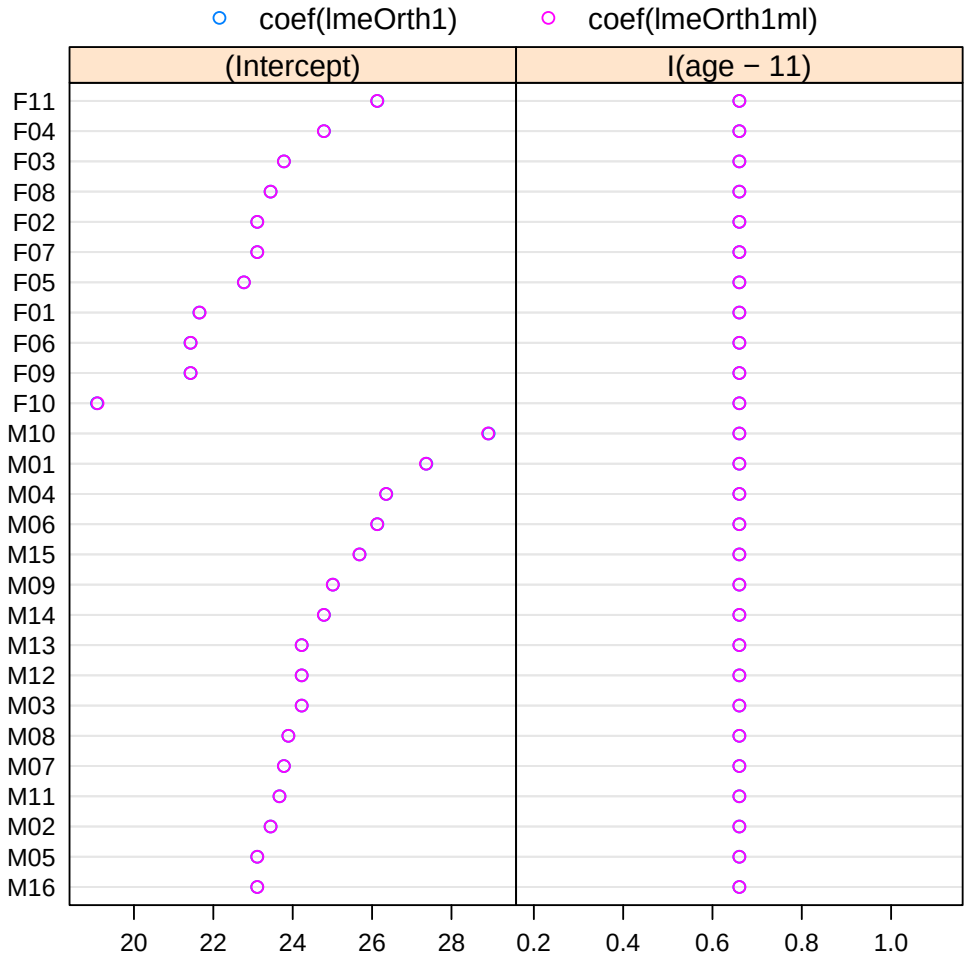
Parameter estimates:

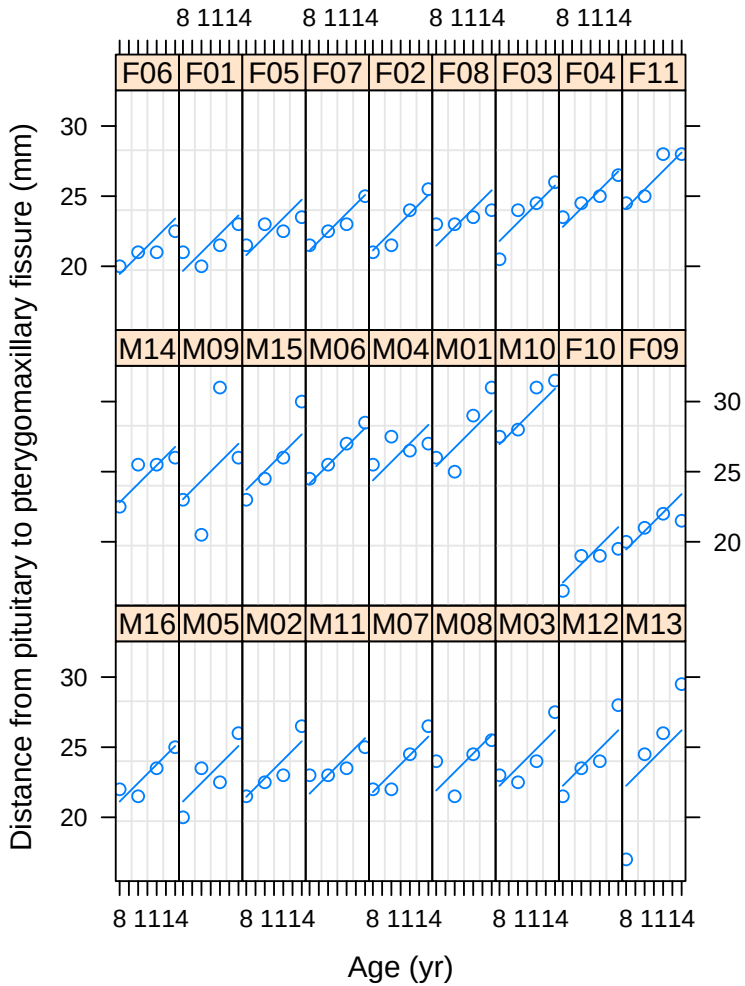


HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Run a random intercept model and get:

```
R Console                                                                Page 1

> coef(lmeOrth1)#subject specific coefficients (random intercept only)
    (Intercept) I(age - 11)
M16    23.10517   0.6601852
M05    23.10517   0.6601852
M02    23.44163   0.6601852
M11    23.66593   0.6601852
M07    23.77808   0.6601852
M08    23.89023   0.6601852
M03    24.22668   0.6601852
M12    24.22668   0.6601852
M13    24.22668   0.6601852
M14    24.78744   0.6601852
M09    25.01174   0.6601852
M15    25.68464   0.6601852
M06    26.13325   0.6601852
M04    26.35755   0.6601852
M01    27.36691   0.6601852
M10    28.93702   0.6601852
F10    19.06774   0.6601852
F09    21.42291   0.6601852
F06    21.42291   0.6601852
F01    21.64721   0.6601852
F05    22.76872   0.6601852
F07    23.10517   0.6601852
F02    23.10517   0.6601852
F08    23.44163   0.6601852
F03    23.77808   0.6601852
F04    24.78744   0.6601852
F11    26.13325   0.6601852
```




8 1114
F06 F01 F05 F07 F02 F08 F03 F04 F11
M14 M09 M15 M06 M04 M01 M10 F10 F09
M16 M05 M02 M11 M07 M08 M03 M12 M13
30
25
20
Distance from pituitary to pterygomaxillary fissure (mm)
Age (yr)

Random Intercept and Slope Model

Consider the model

$$Y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + e_{ij}.$$

- Each subject varies with respect
  (i) baseline level when $t_{i1} = 0$ and
  (ii) rate of change of response over time
- In this case we have the same fixed and random terms
- The variance is a function of time
  $Var(Y_{ij}) = Var(b_{1i}) + 2t_{ij}Cov(b_{1i}, b_{2i}) + t_{ij}^2 Var(b_{2i}) + Var(e_{ij})$ and the covariance too
  $Cov(Y_{ij}, Y_{ik}) = Var(b_{1i}) + (t_{ij} + t_{ik})Cov(b_{1i}, b_{2i}) + t_{ij}t_{ik}Var(b_{2i})$

Covariance Structure

In the linear mixed model

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

the matrix $W_i = Cov(e_i)$ introduces the covariance between the repeated observations when focusing on the conditional mean response profile of a specific individual. In other words, it is the covariance of the i^{th} individual's deviations from the response profile

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

- The usual assumption is $W = \sigma^2 I_n$. This is referred as the *conditional independence* assumption.
- The conditional covariance becomes

$$Cov(Y_i|b_i) = Cov(e_i) = W_i$$

- The marginal then takes the form

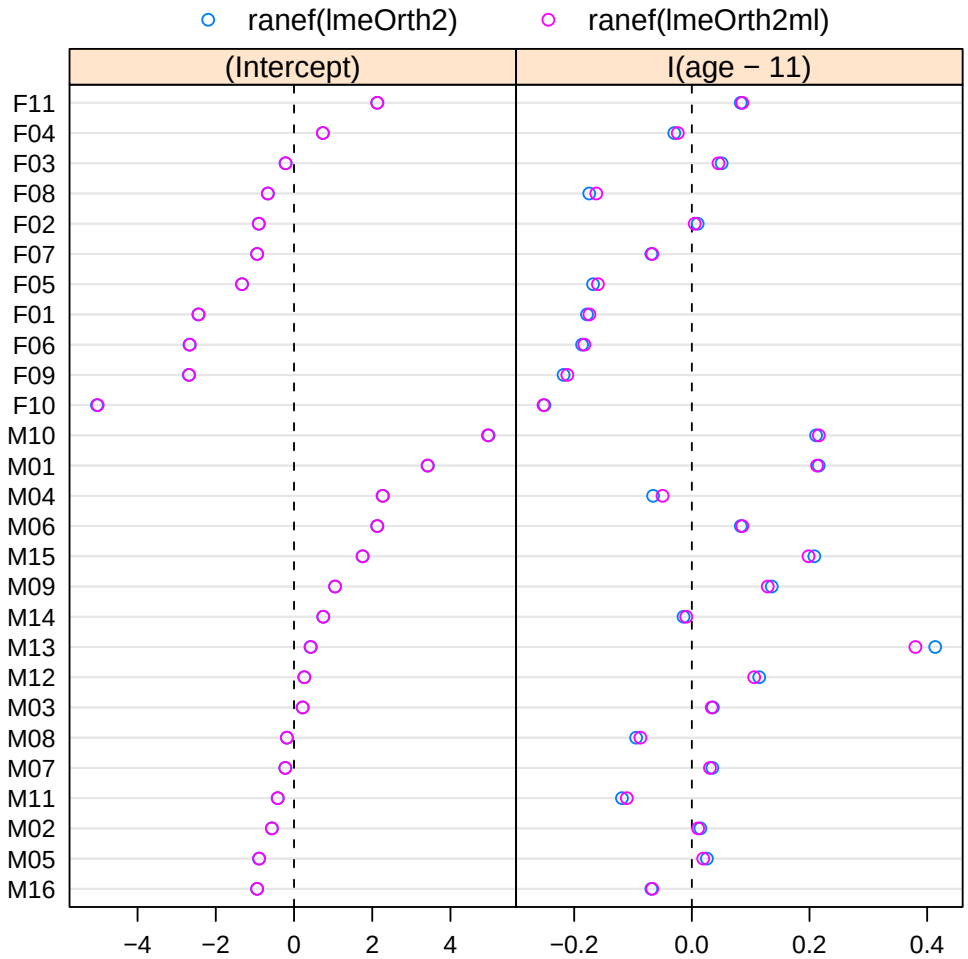
$$Cov(Y_i) = Z_i B Z_i' + W_i$$

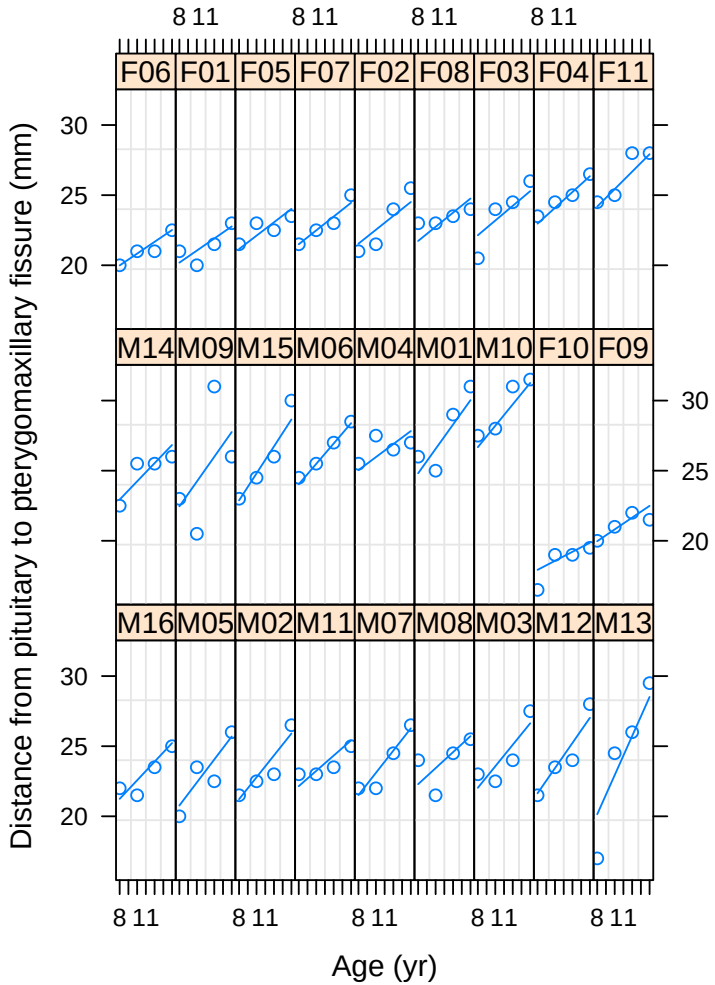
.

Some Characteristics

- There is no need of balanced data.
- The covariances are functions of time. As a result, if time is included in Z_i, each patient can have his own sequence of measurement times. This property makes these models suitable for the analysis of *real life* longitudinal data.
- The number of covariance parameters that need to be estimated remains unchanged regardless of the number of measurements.
- The random effects covariance structure allows the variances and covariances to change (increase or decrease) as a function of measurement times, without introducing restrictive structures as the covariance pattern models do.

Example: Ortodont data (cont)




F06 F01 F05 F07 F02 F08 F03 F04 F11
M14 M09 M15 M06 M04 M01 M10 F10 F09
M16 M05 M02 M11 M07 M08 M03 M12 M13
8 11
30
25
20
Distance from pituitary to pterygomaxillary fissure (mm)
Age (yr)

Presentation Outline

Overview of Analysis of Variance

- In general ANOVA serves as a generalization of t-tests (comparison between 2 groups)
- It can be seen as a special case of linear models, separately developed with great use in experimental (designs) studies
- One-way ANOVA involves one dependent variable (DV) and one independent (IV), while two-way ANOVA involves two IVs
- The DV must be a continuous variable
- The IV is the grouping/categorical variable, called 'factor'
- In a linear model it is expressed through a collection of dummy variables
- ANOVA compares the within classes variances to overall one

- A main effect is the direct effect of an IV on the DV
- ANOVA uncovers the main (and interaction) effects of IVs
- An interaction effect is the joint effect of two IVs on the DV
- Key statistic in ANOVA is the F-test for difference of means
- It tests if the means of groups (formed by factor levels or combinations of them) are different enough not to have occurred by chance
- If the group means do not differ significantly then it is inferred that the IV(s) do not have an effect on the DV
- If the F-test shows that overall the IV(s) is (are) related to the DV, then multiple comparison tests will follow to explore which levels of the IV(s) have the most to do with the DV

One-way ANOVA

The usual equation presented for ANOVA is

$$Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where Y_{ij} is the response for the j-individual in the i-group/level

- Parameter μ represent the grand/overall mean (average of the population means)
- The effect of level i to the total overall mean is expressed as

$$\alpha_i = \mu_i - \mu$$

The question is whether the values of Y differ across groups or not

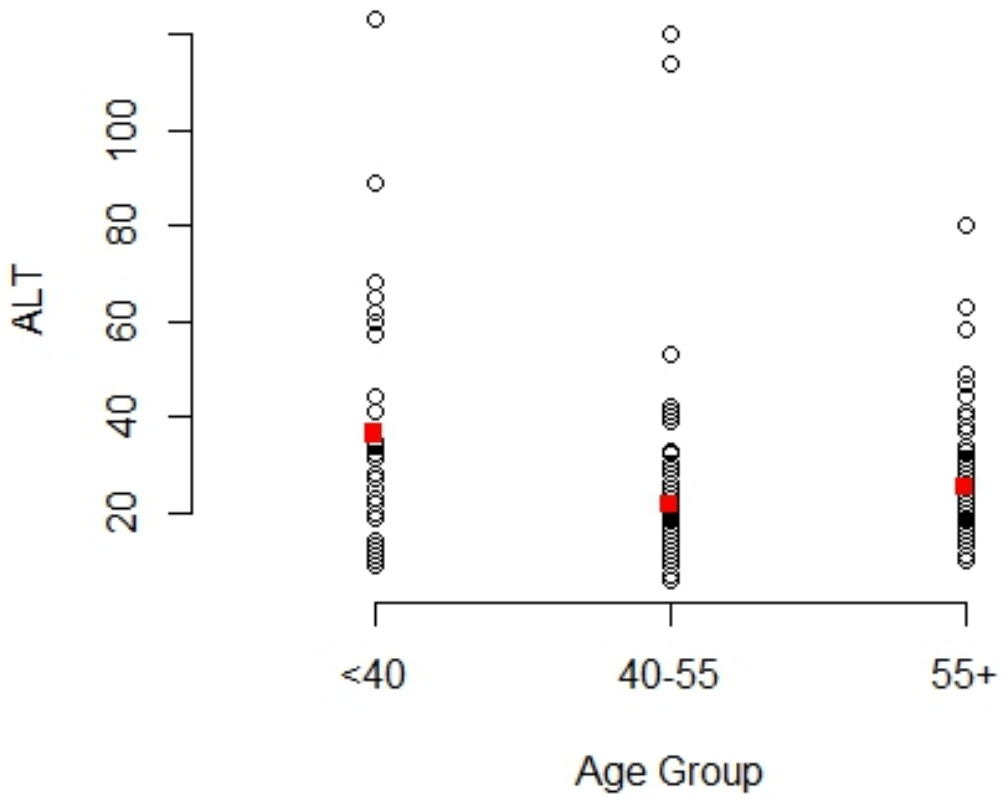
$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$
$$H_1 : \ \text{at least one } \mu_i \text{ different from the others}$$

Data

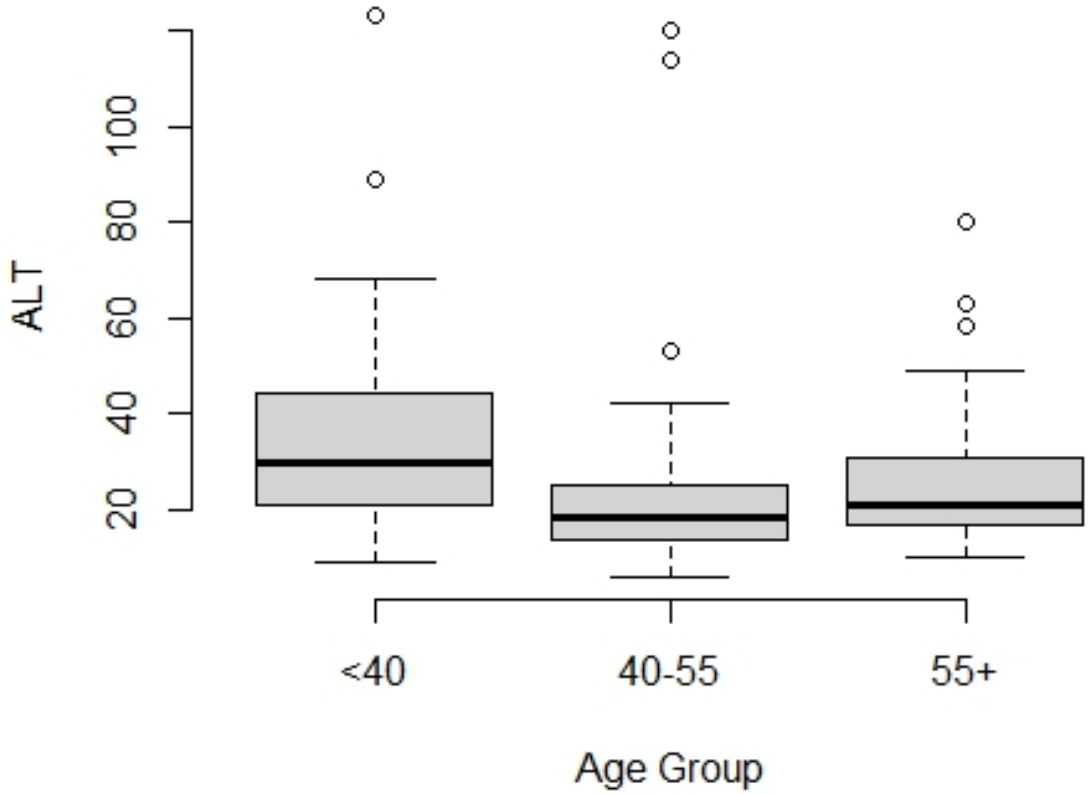
Factor

$Level-1$	$Level-2$	$\ldots$	$Level-k$
x_{11}	x_{21}	$\ldots$	x_{k1}
x_{12}	x_{21}	$\ldots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n_1}	x_{2n_2}	$\ldots$	x_{kn_k}

Example: ALT (alanine transaminase) from blood test for three Age Groups



HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837


ALT
20
40
60
80
100
<40
40-55
55+
Age Group

Notation

Complete data

- n: total number of subjects
- k: number of groups/levels
- $\bar{x}$: overall mean

Group i $(i = 1, 2, \ldots, k)$

- n_i: number of subjects in group i
- x_{ij}: value of subject i in group j
- $\bar{x}_i$: mean for group i

How it Works

ANOVA measures two sources of variation in the data and compares them

- BETWEEN groups variation

$$SSB = \sum_i (\bar{x}_i - \bar{x})^2$$

- WITHIN groups variation

$$SSW = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

- Test:

$$F = \frac{MSB}{MSW} = \frac{SSB/(k-1)}{SSW/(n-k)}$$

ANOVA Table

Source of Variation	DF	SS	MS	$F-test$
Between	$k-1$	SSB	$MSB = \frac{SSB}{k-1}$	$\mathrm{F} = \frac{MSB}{MSW}$
Within	$n-k$	SSW	$MSW = \frac{SSW}{n-k}$	
Total	$n-1$			

ALT example (R output):

```
> summary( aov(ALT ~ as.factor(age2),data=smoke[w,]) )
                  Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(age2)    2   5391  2695.4   9.406 0.000125 ***
Residuals        197  56453   286.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Assumptions

- The dependent variable has a normal distribution
- The F-test is robust to this assumption if the sample sizes are large
- Identical variances in each group (homoscedasticity assumption)
- Independence between groups and random sampling
  (in general these assumptions cannot be tested and the best thing we can do is to obtain the data in such a way that we are certain about these assumptions)

Two-way ANOVA

- A natural extension of the one-way ANOVA
- Now we have two independent variables, with k and r levels respectively
- We have three sets of hypotheses, so we have to test:
  - The means of the first factor are equal
    (like one-way ANOVA for the factor A - row factor)
  - The means of the second factor are equal
    (like one-way ANOVA for the factor B - column factor)
  - There is no interaction between the two factors

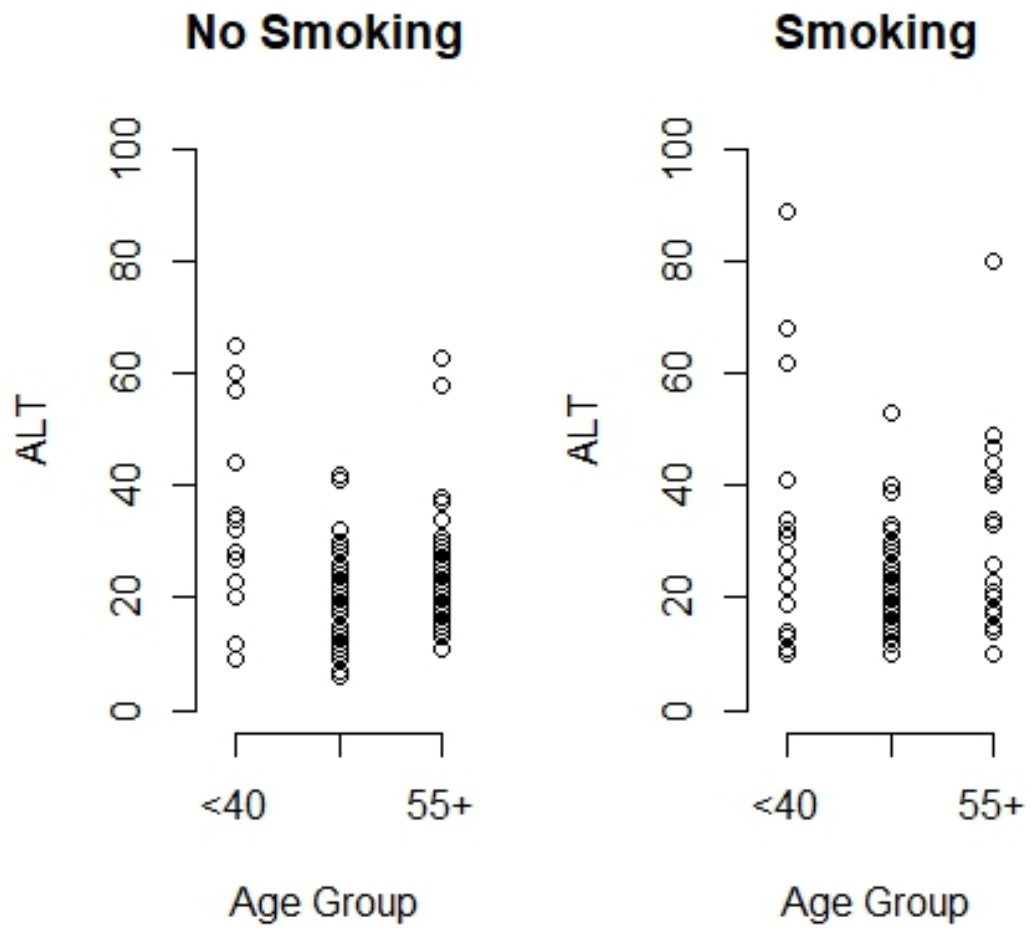
Treatment Groups: these are formed by making all possible combinations of the two factors
(if one has 3 levels and the other has 2, then $3 \times 2 = 6$ groups)

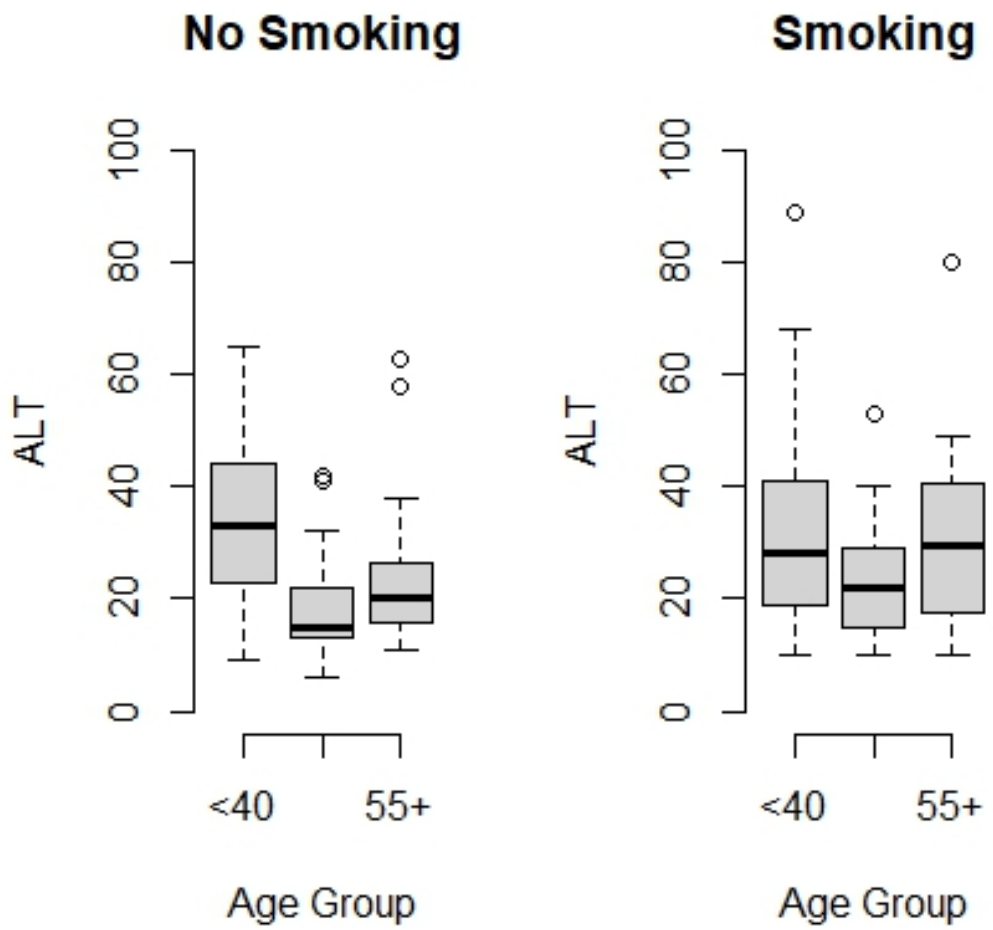
Main Effect: Is the effect of one variable at a time

Interaction: Is the combined effect of both variables

Source of Variation	DF	SS	MS	$F - test$
Main effect A Main effect B Interaction Within				
Total				

ALT example with Age Group and Smoking Status as factors:




No Smoking
Smoking
ALT
0
20
40
60
80
100
<40
55+
Age Group

ALT example with Age Group and Smoking Status and Interaction:

```
> summary( aov(ALT ~ as.factor(age2) * as.factor(smoking),data=smoke[w,]) )
                                 Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(age2)                   2   5391  2695.4   9.808 8.76e-05 ***
as.factor(smoking)                1   2859  2859.4  10.405  0.00147 **
as.factor(age2):as.factor(smoking)  2    280   139.8   0.509  0.60201
Residuals                       194  53314   274.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ALT example with Age Group and Smoking Status main effects:

```
> summary( aov(ALT ~ as.factor(age2) + as.factor(smoking),data=smoke[w,]) )
                     Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(age2)       2   5391  2695.4   9.857 8.33e-05 ***
as.factor(smoking)    1   2859  2859.4  10.457  0.00143 **
Residuals           196  53593   273.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Multivariate ANOVA (MANOVA)

- What if we are interested in two or more DVs (outcomes)?
- Why do MANOVA, when one can get much more information by doing a series of ANOVAs?
- Even if all our DVs are completely independent of one another, when we do lots of tests like that, (Type I) error inflates
- Nevertheless, in many biological or ecological studies, the variables are not independent at all
- Many times they have strong interactions, inflating the (Type I) error even more
- In many cases where multiple ANOVAs are done, MANOVA is actually the more appropriate test

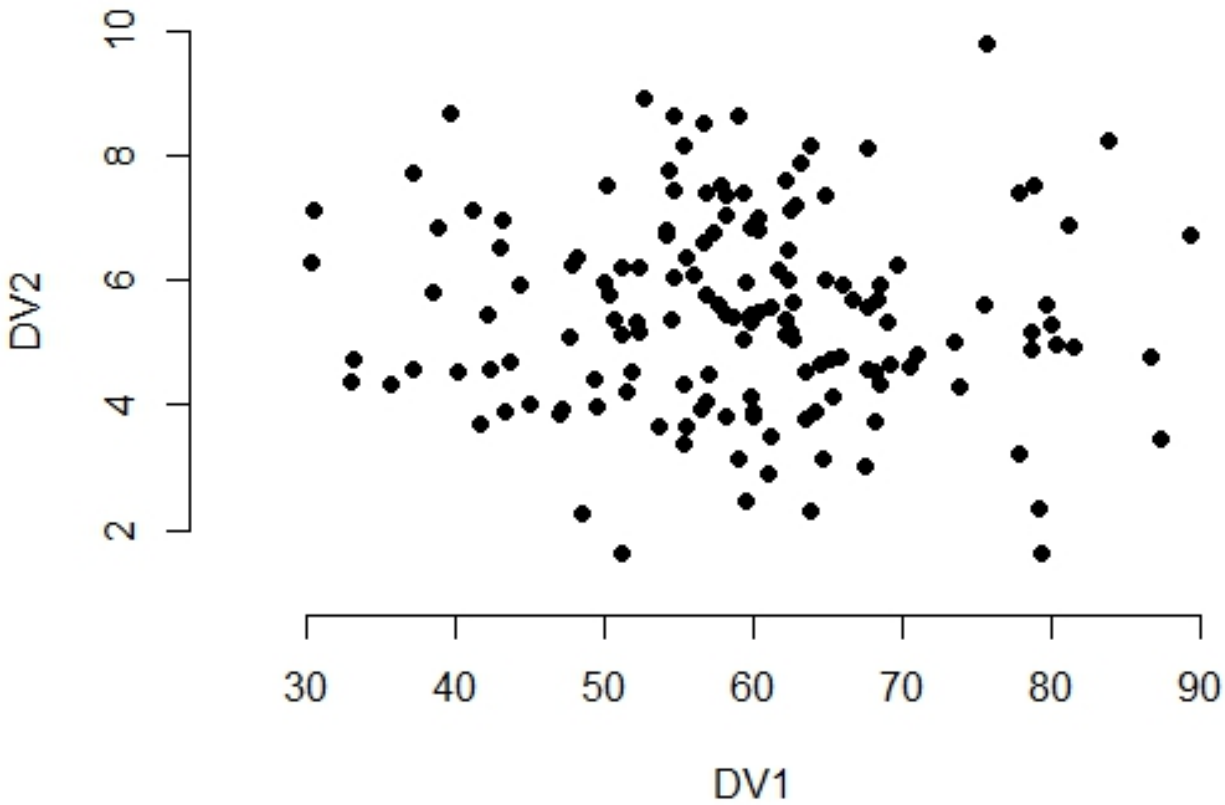
- Essentially MANOVA is the ANOVA application to a vector (list) of DVs, rather than just one
- Therefore, instead of looking into different means across groups we look into different locations in the DVs space across groups
- The null hypothesis (H_0) is that the different groups all have the same centroid (like the center of mass or center of gravity in purely geometrical terms) in the DV-space
- The alternative hypothesis (H_1) is that at least one group has a distinct centroid in the DV-space

A Hypothetical Example

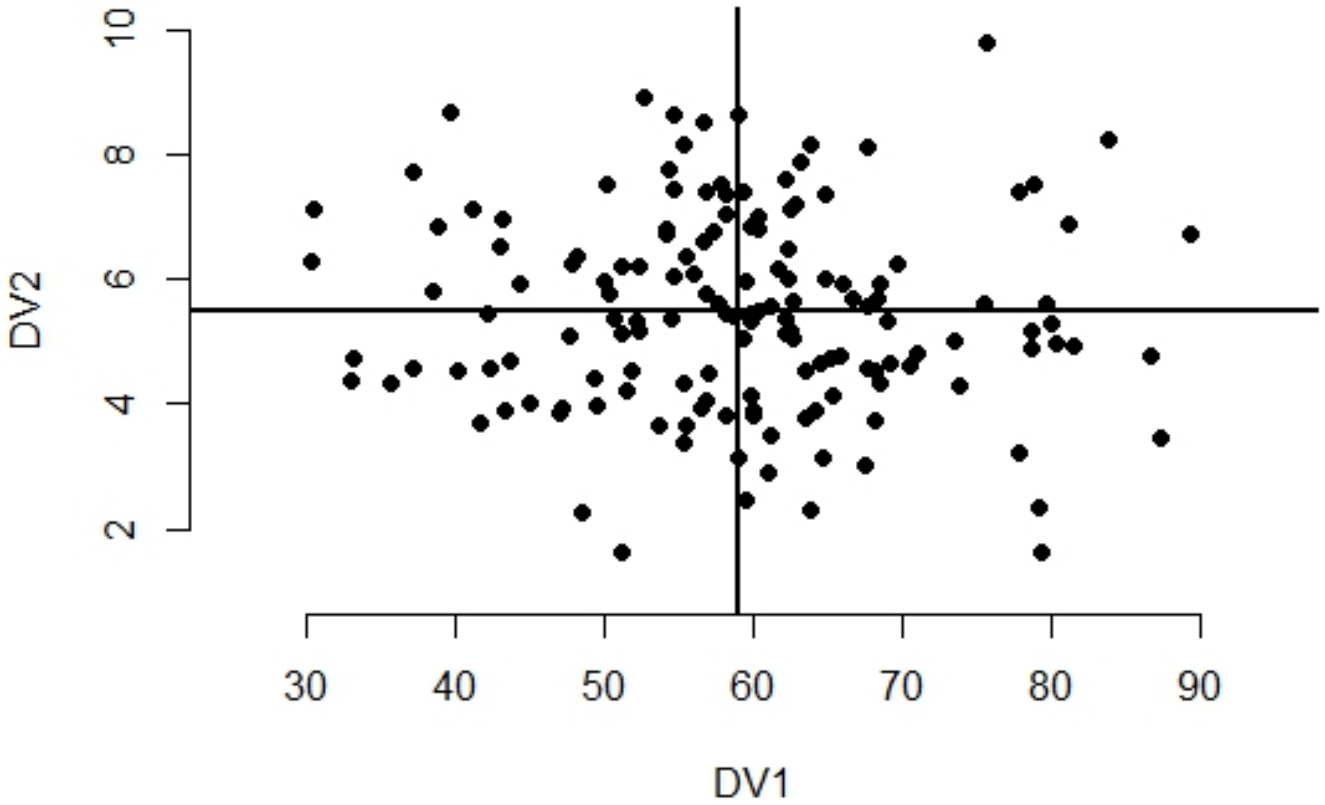
We will go through the MANOVA with the help of a simple hypothetical example.

- Assume we are trying to compare the performance of three cars
- We have 150 drivers and we equally allocate them ($n = 50$) to three different groups
- The drivers in each group drive the same car and they rate "Performance" (scale 0-100) and "Enjoyment" (scale 0-10)
- Our job is to investigate if there are differences in performance and/or enjoyment for the three cars
- We have: DV1=Performance, DV2=Enjoyment and IV=Car (3 levels)
- So our dependent variable is not a scalar quantity, but a collection of points in the (Performance, Enjoyment) space

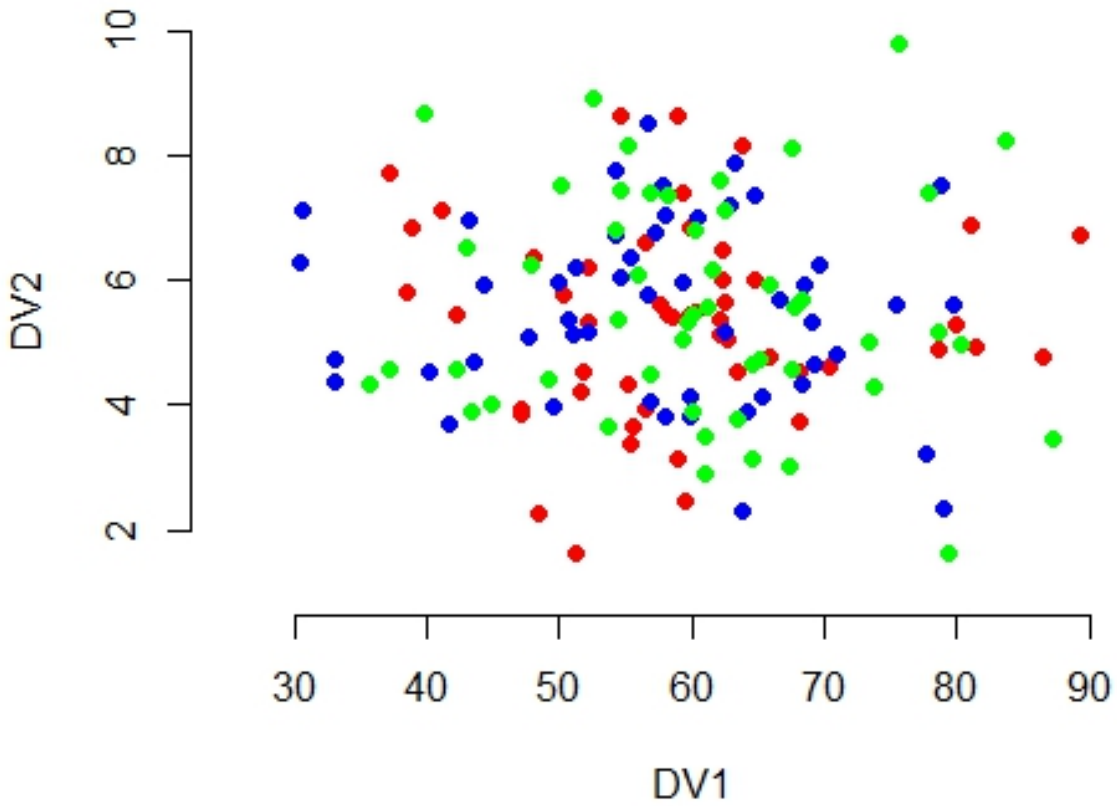
Scenario 1: A complete random case:



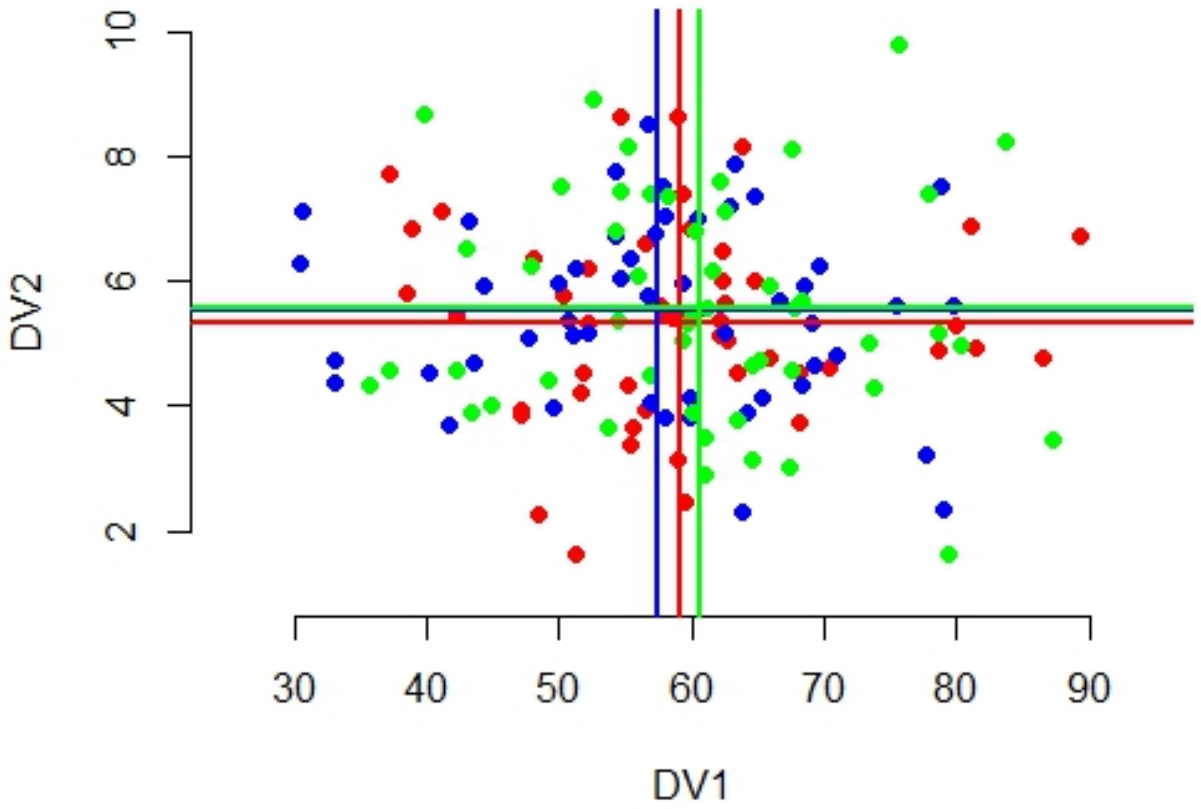
...and their mean values:



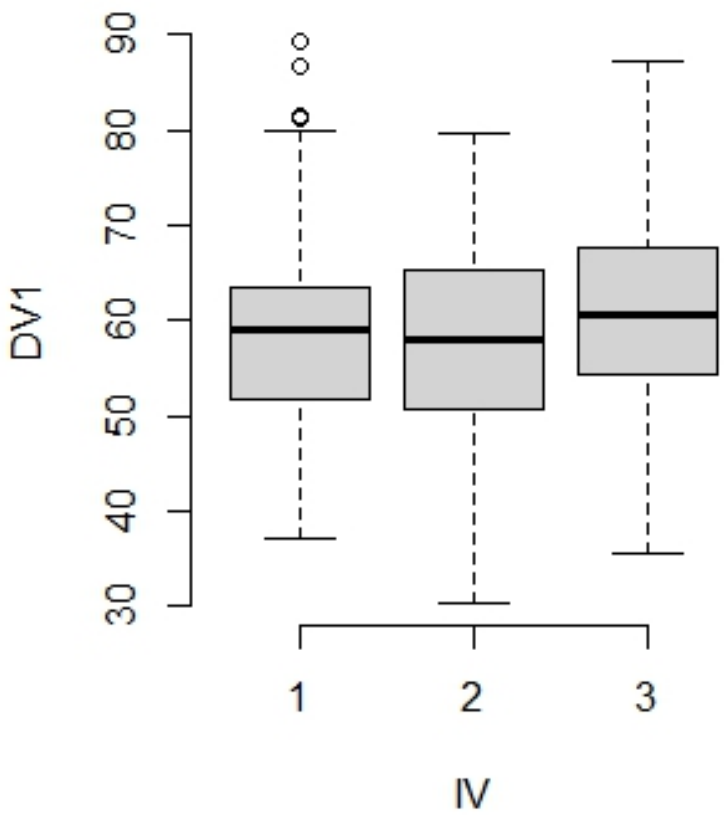
Plot the data per factor level (three cars: red/blue/green):

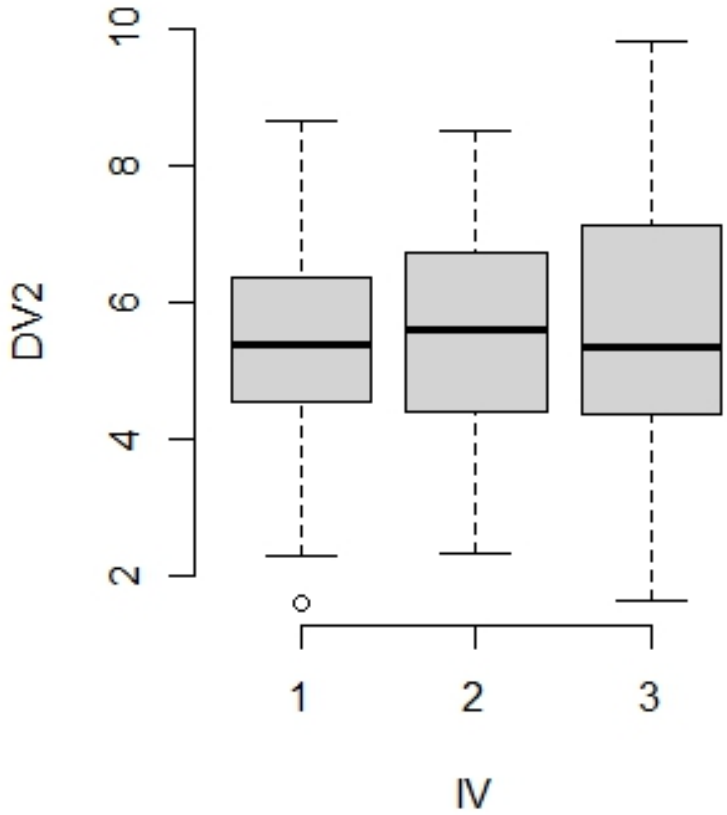


Get the means per factor level:



Check the boxplots:





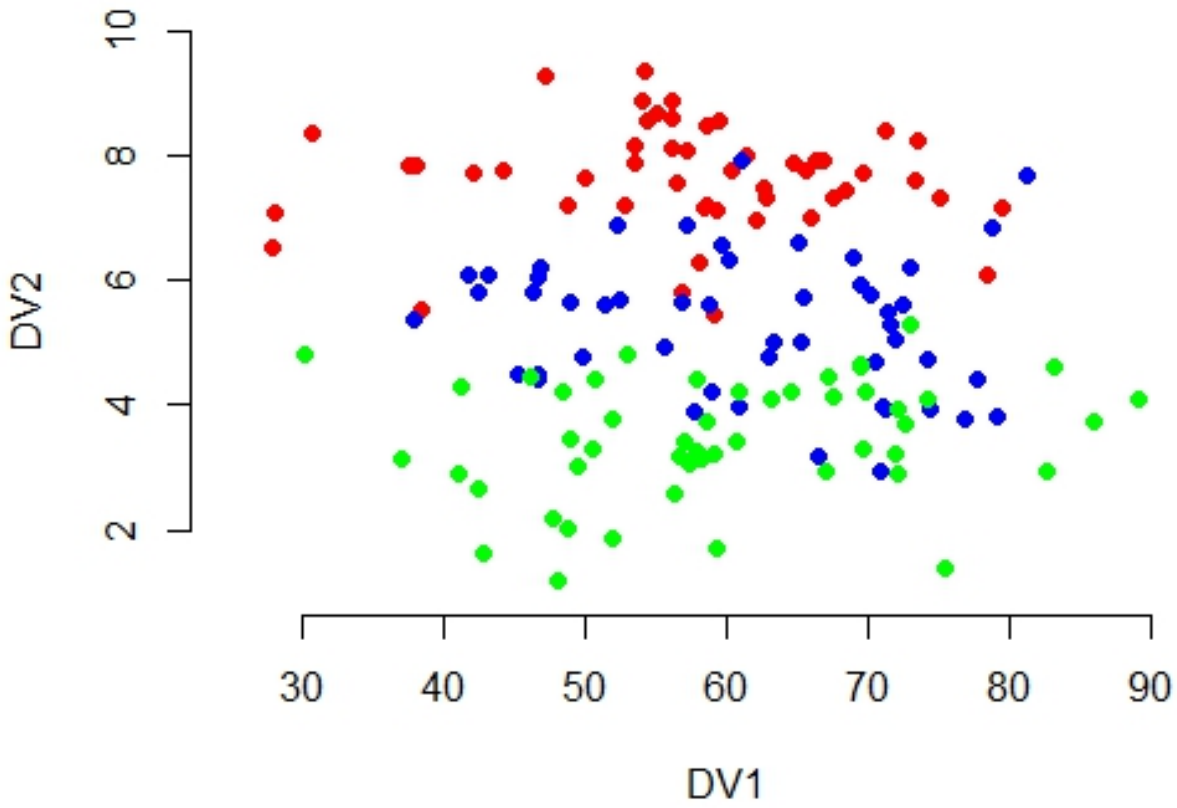
MANOVA:

```
> m1 = manova(W~d1$group)
> summary(m1,test='Wilks')
           Df   Wilks approx F num Df den Df Pr(>F)
d1$group    2 0.96524   1.3029      4    292  0.269
Residuals 147
> summary.aov(m1)
 Response 1 :
             Df  Sum Sq Mean Sq F value Pr(>F)
d1$group      2   289.3  144.64  0.8772 0.4181
Residuals   147 24237.3  164.88

 Response 2 :
             Df  Sum Sq Mean Sq F value Pr(>F)
d1$group      2   6.523  3.2615  1.7269 0.1814
Residuals   147 277.626  1.8886
```

Scenario 1: MANOVA output

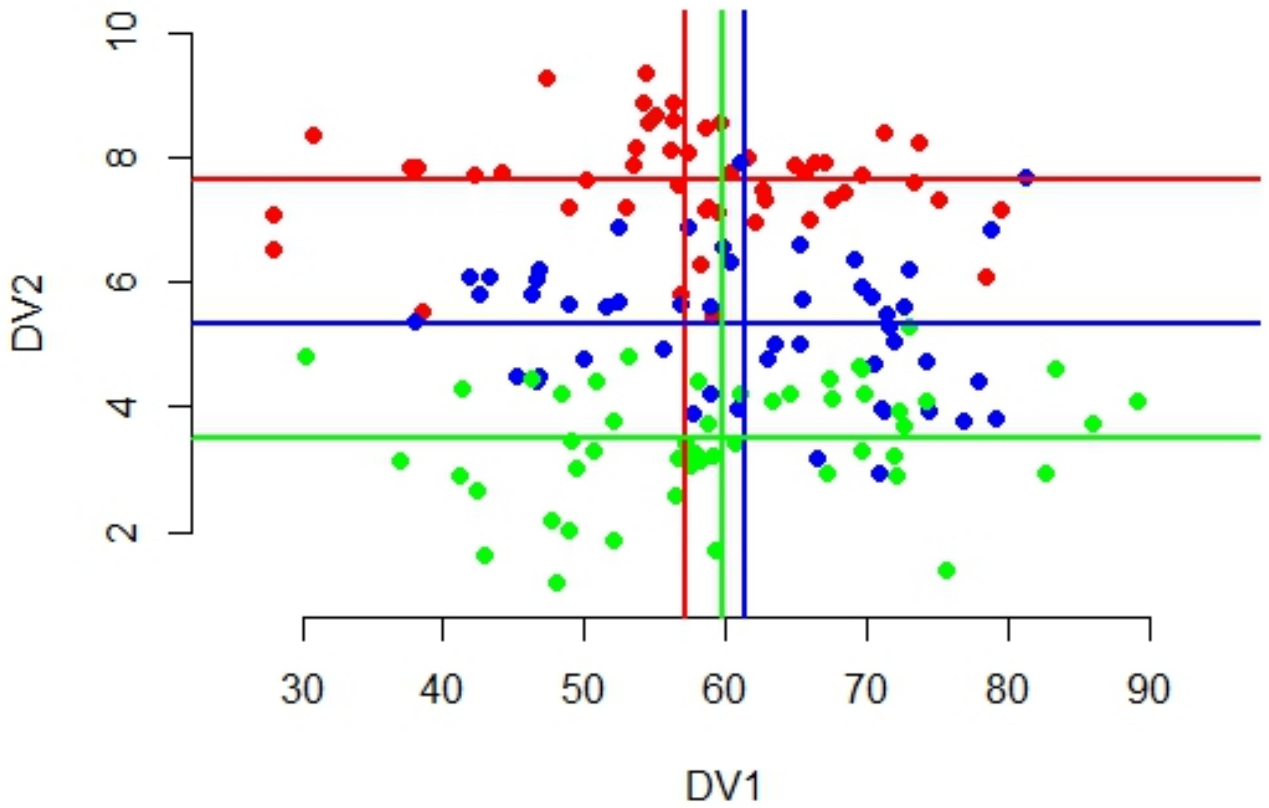
- P-value is not significant in this case, as expected.
- We can't reject the null hypothesis that groups A, B and C have the same centroid in the Performance-Enjoyment space

Scenario 2: Differences in DV2:

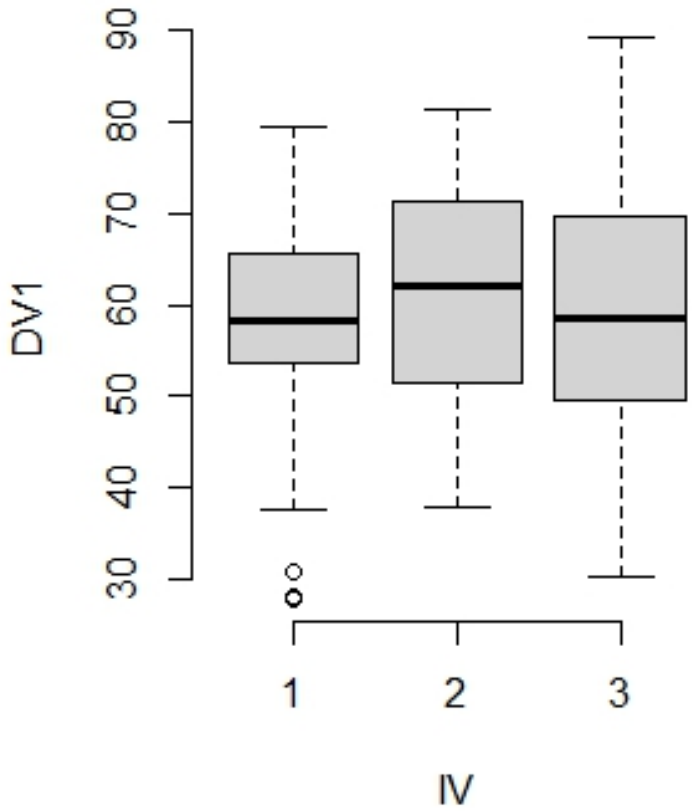


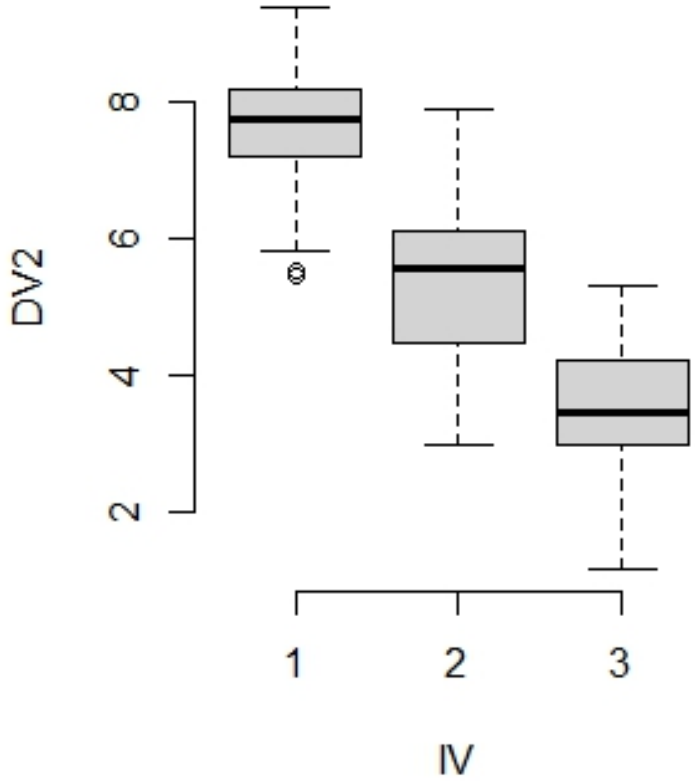
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Plot data with means:



Boxplots:





Analysis in R:

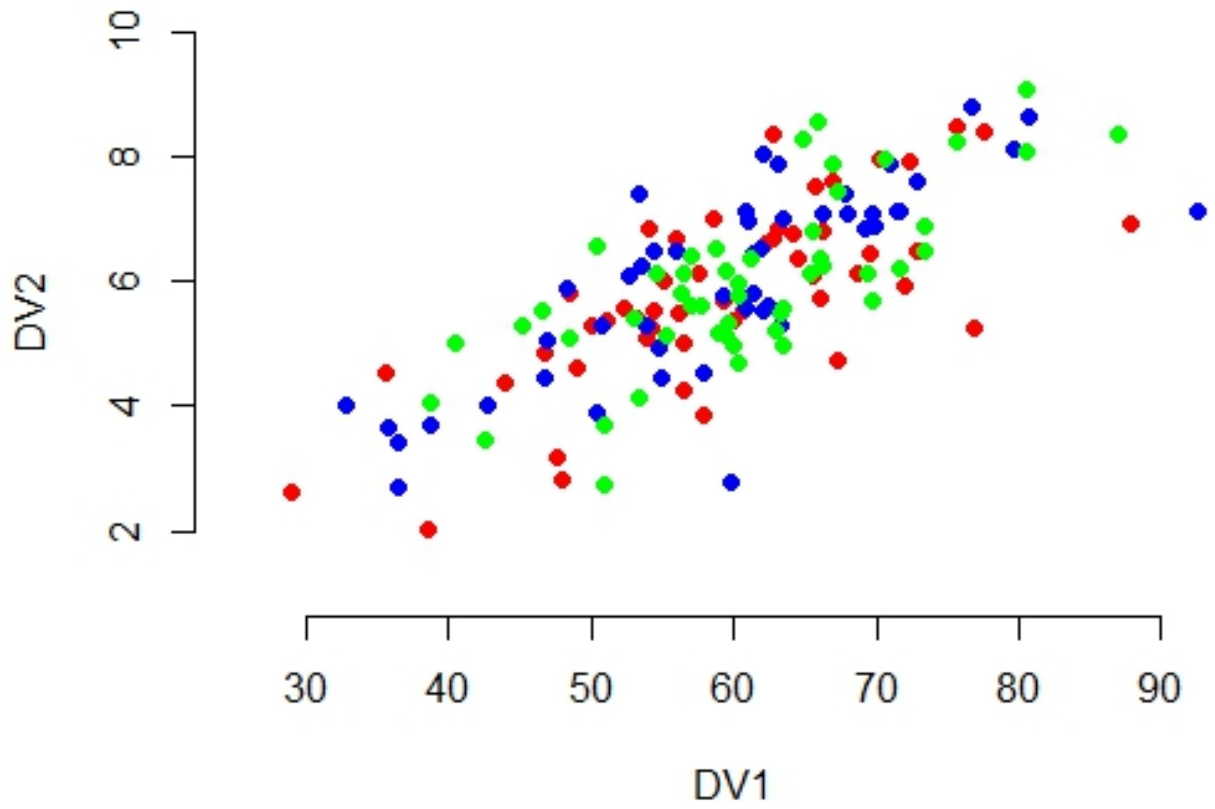
```
> summary(m1,test='Wilks')
           Df   Wilks approx F num Df den Df    Pr(>F)
d1$group    2 0.24352    74.93      4    292 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.aov(m1)
 Response 1 :
             Df  Sum Sq Mean Sq F value Pr(>F)
d1$group      2   460.3  230.16  1.5207  0.222
Residuals   147 22248.1  151.35

 Response 2 :
             Df Sum Sq Mean Sq F value    Pr(>F)
d1$group      2 437.45 218.726  224.02 < 2.2e-16 ***
Residuals   147 143.53   0.976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Scenario 2: MANOVA output

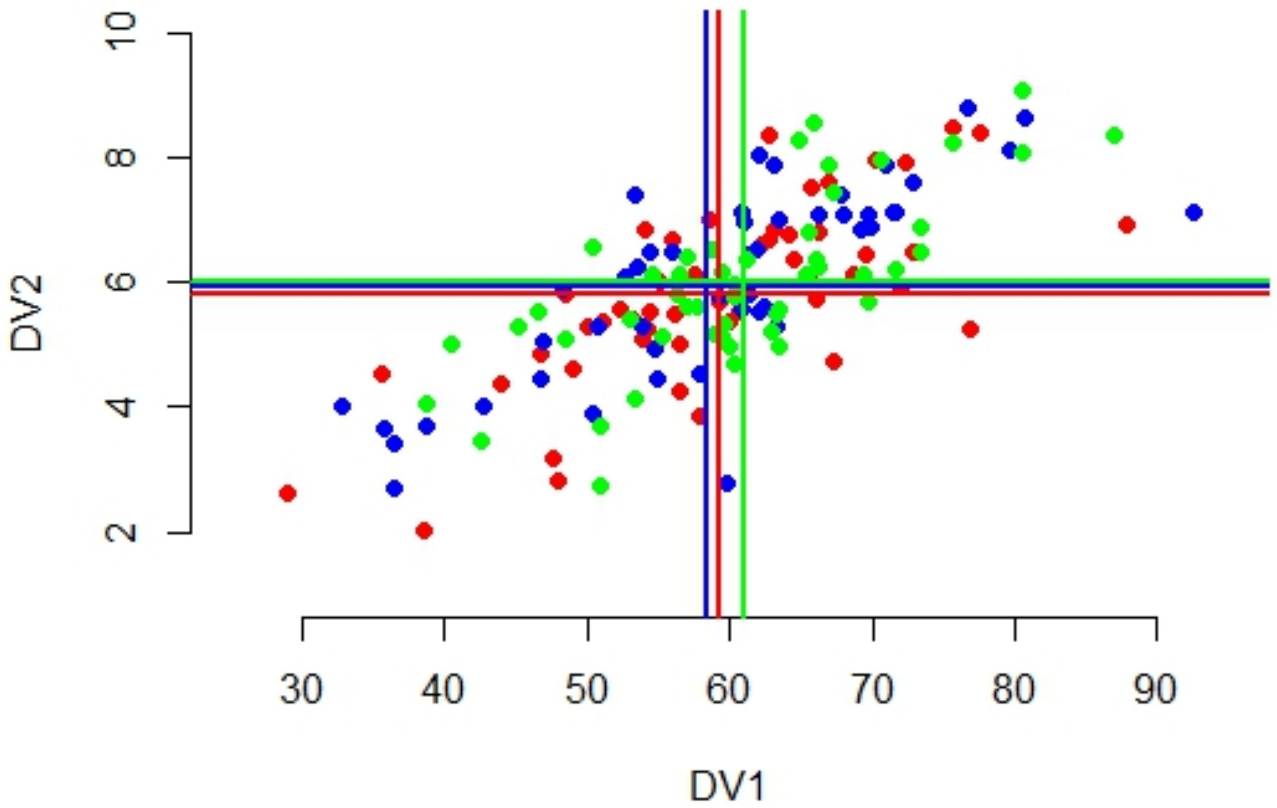
- We can reject the null hypothesis that the three groups share the same centroid in DVs space
- From MANOVA we know there are differences, but we do not know *how* they are different
- A univariate analysis confirms that there are differences in DV2 (Enjoyment) but not in DV1 (performance)

Scenario 3: DV1 and DV2 are related:

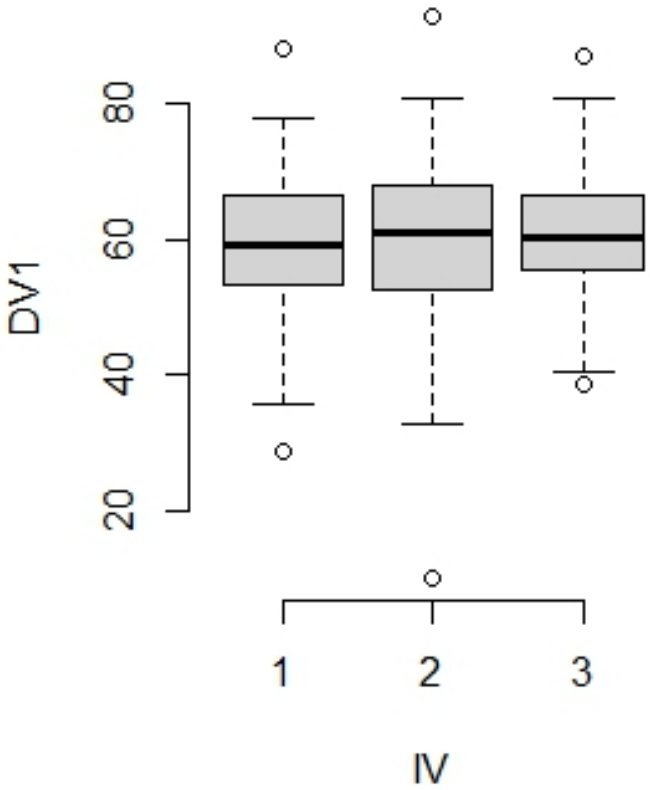


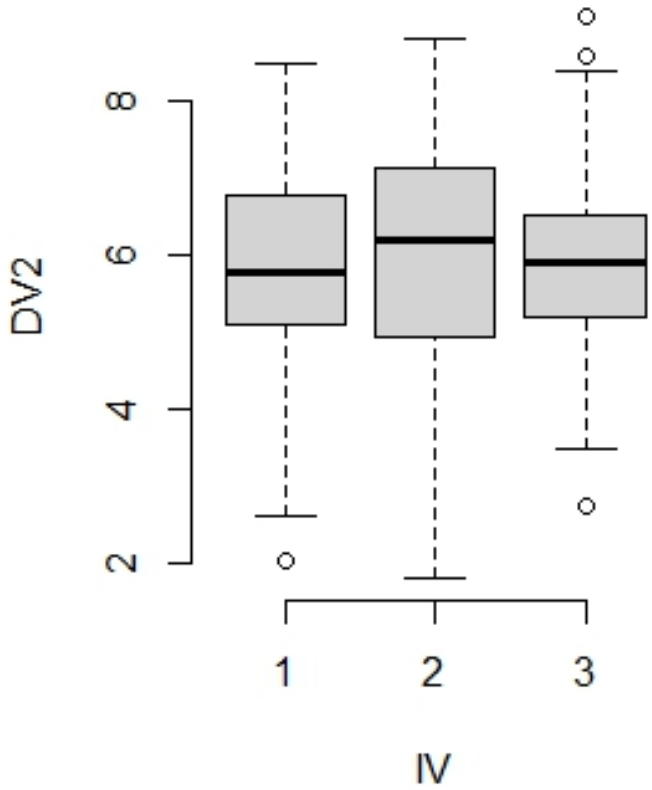
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Plot with means:



Boxplots:





HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

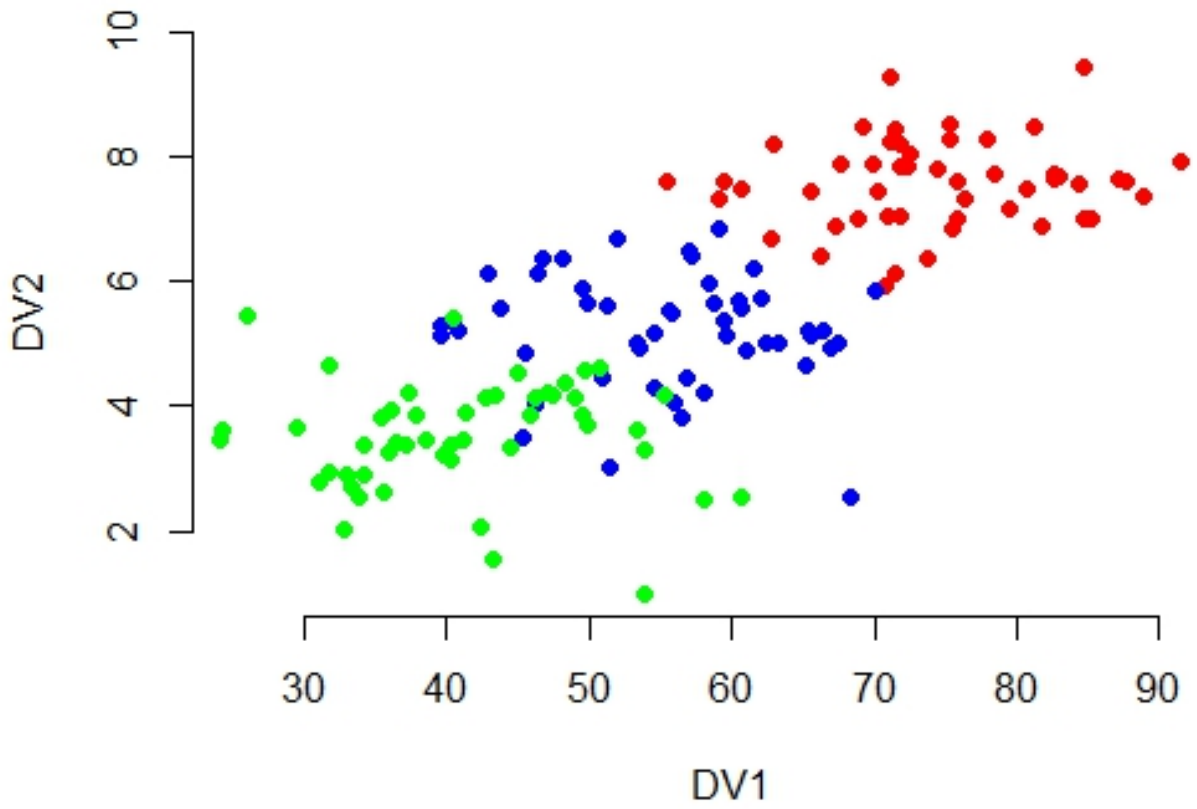
Analysis in R:

```
> summary(m1,test='Wilks')
           Df Wilks approx F num Df den Df Pr(>F)
d1$group    2 0.982  0.66593      4    292 0.6161
Residuals 147
> summary.aov(m1)
 Response 1 :
             Df  Sum Sq Mean Sq F value Pr(>F)
d1$group      2   170.5  85.248   0.601 0.5496
Residuals   147 20851.5 141.847

 Response 2 :
             Df Sum Sq Mean Sq F value Pr(>F)
d1$group      2   0.99  0.4973  0.2264 0.7977
Residuals   147 322.87  2.1964
```

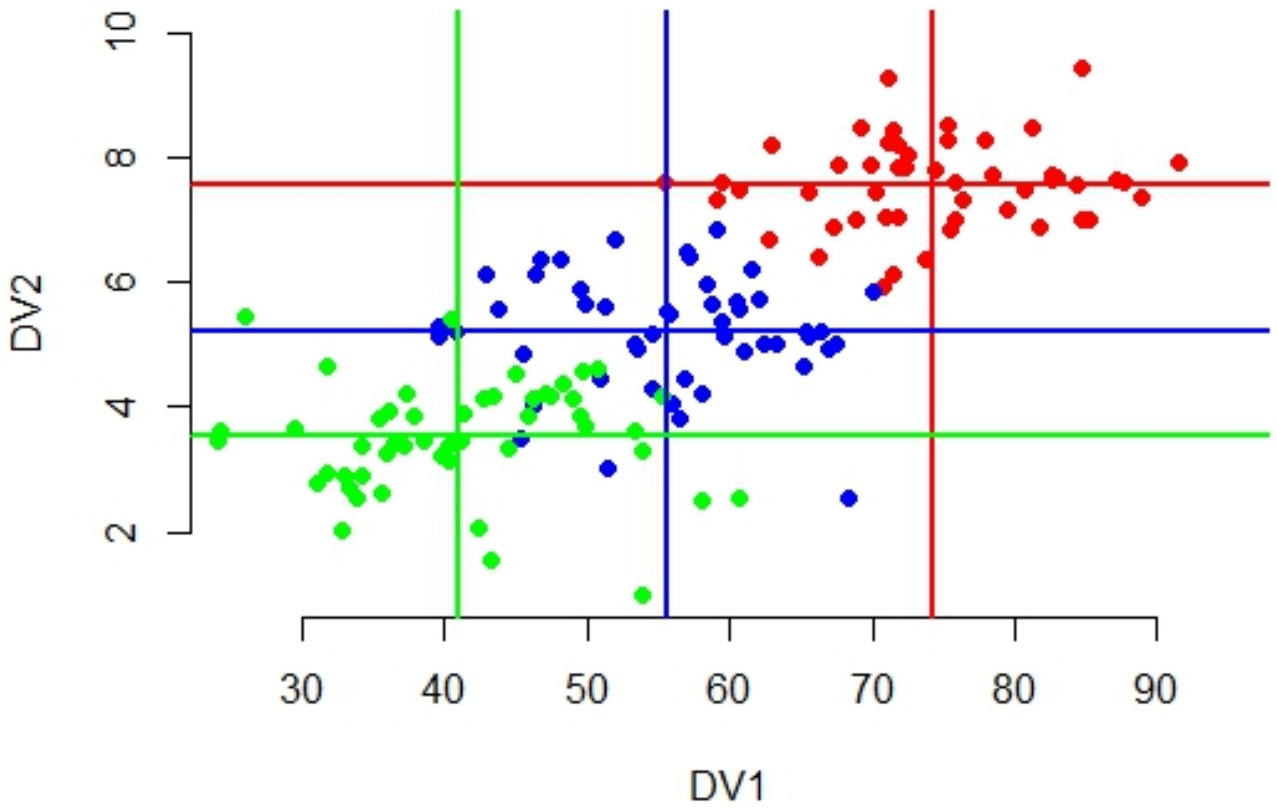
Scenario 3: MANOVA output

- The fact that DV1 and DV2 are related doesn't change the fact that they have similar centroid in the DVs space
- As such, the MANOVA results are not significant

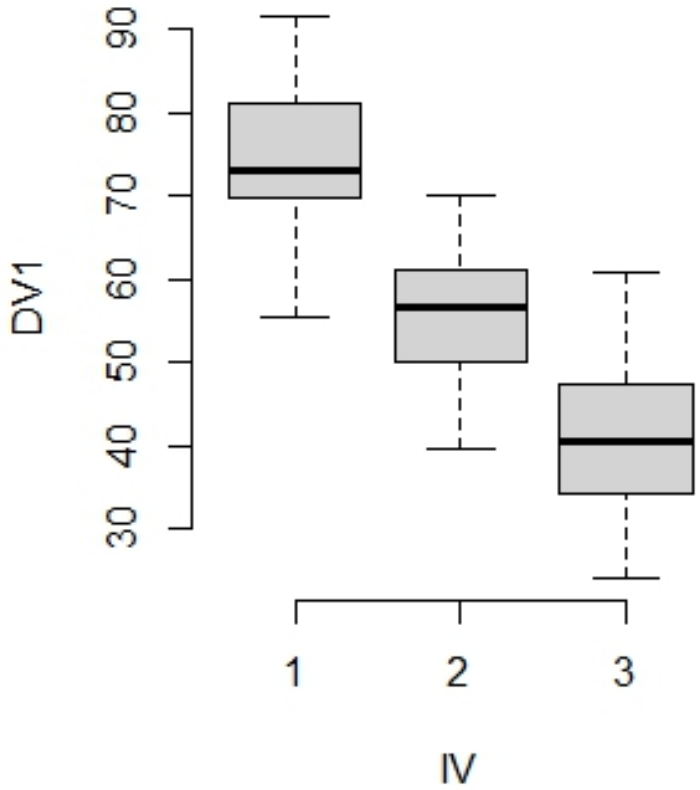
Scenario 4: Differences in both DV1 and DV2 across levels:

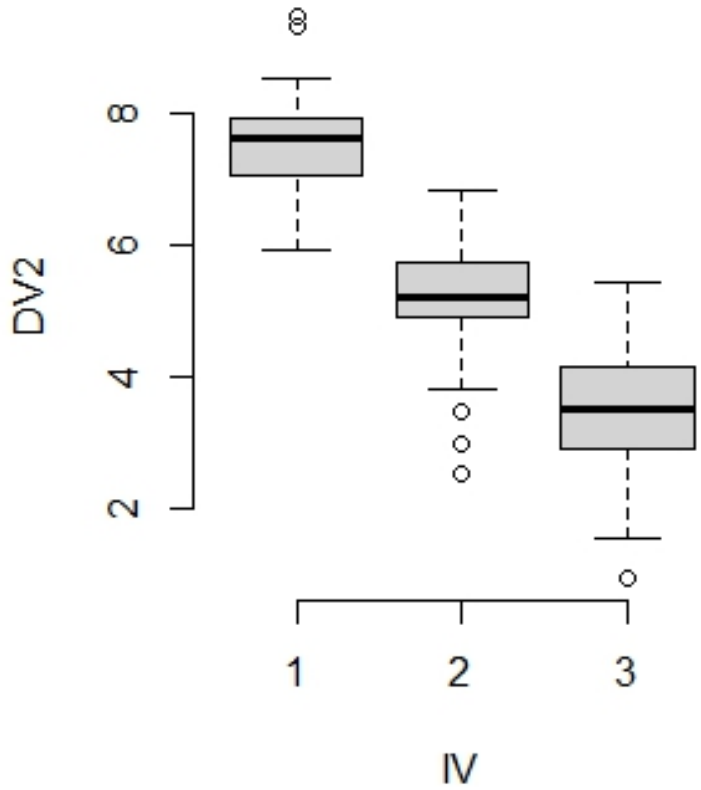


Plot with means:



Boxplots:





HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Analysis in R:

```
> summary(m1,test='Wilks')
           Df   Wilks approx F num Df den Df    Pr(>F)
d1$group    2 0.12861   130.56      4    292 < 2.2e-16 ***
Residuals 147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.aov(m1)
 Response 1 :
            Df Sum Sq Mean Sq F value    Pr(>F)
d1$group     2  27808 13903.8   201.2 < 2.2e-16 ***
Residuals  147  10158    69.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
            Df Sum Sq Mean Sq F value    Pr(>F)
d1$group     2 417.12 208.559  296.79 < 2.2e-16 ***
Residuals  147 103.30   0.703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

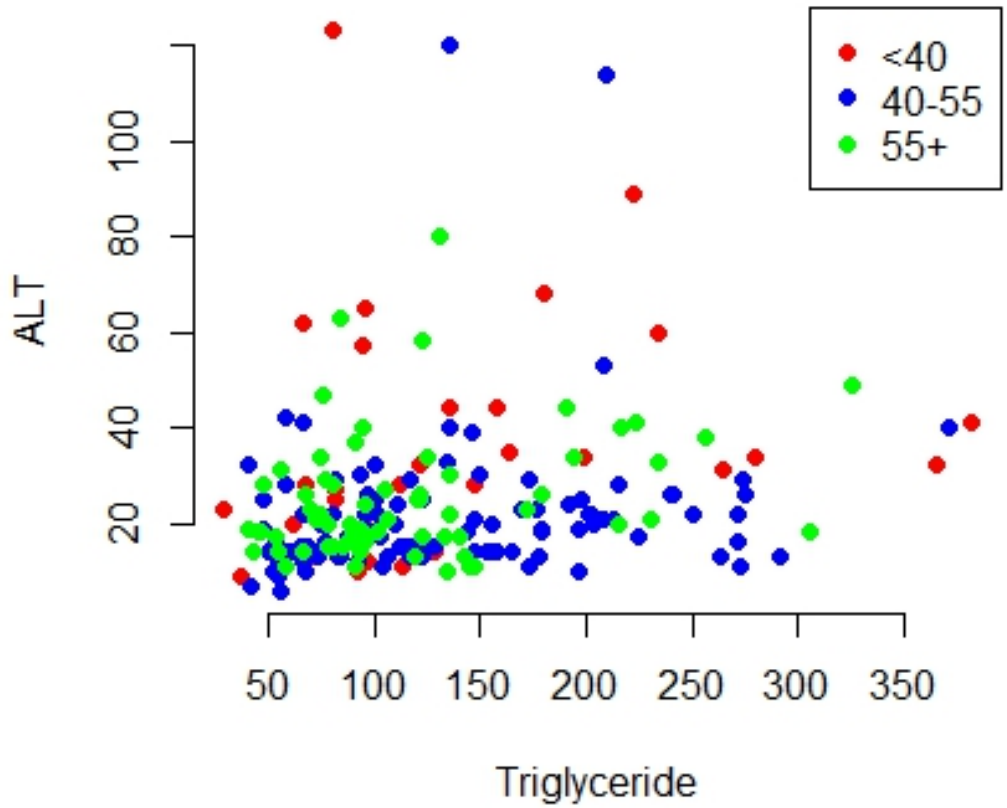
Scenario 4: MANOVA output

- The overall MANOVA model is highly significant, because the three groups occupy different parts in the DVs space
- The univariate analyses confirms that there are between groups differences in both DV1 and DV2

Generalization

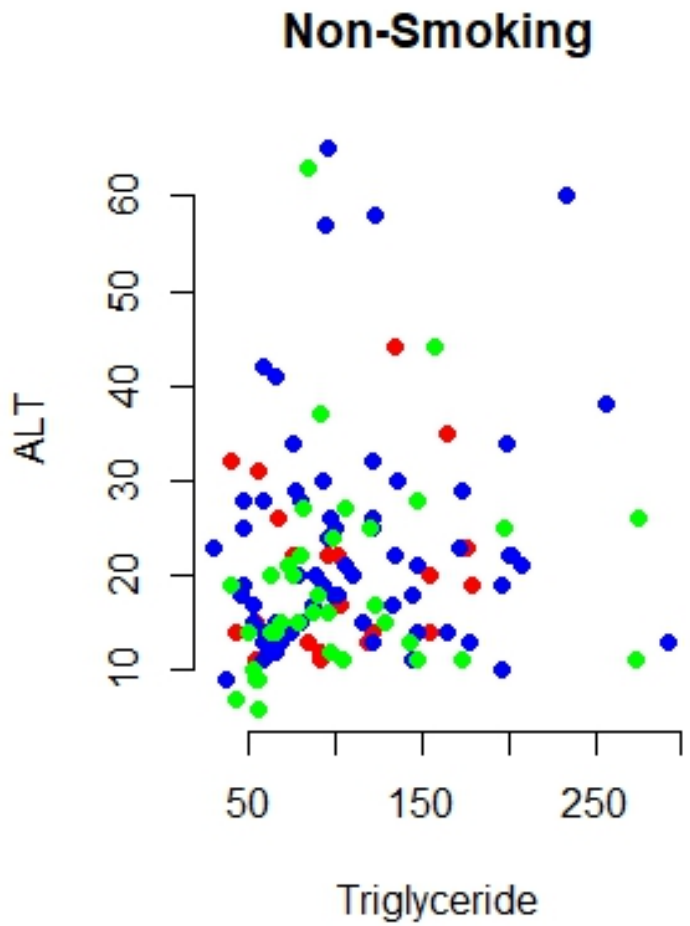
- This was a very simple example with two DVs and one IV with three levels
- In such cases the results are easy to visualize, and from these visualizations we can get all the information we need
- We can have more DVs, which means that we move to ¿2 dimensions in the DVs space
- If the IV has more levels or additional IVs are considered in the model, then things can become even more complicated
- In more complex scenarios, interactions between IVs are also of interest

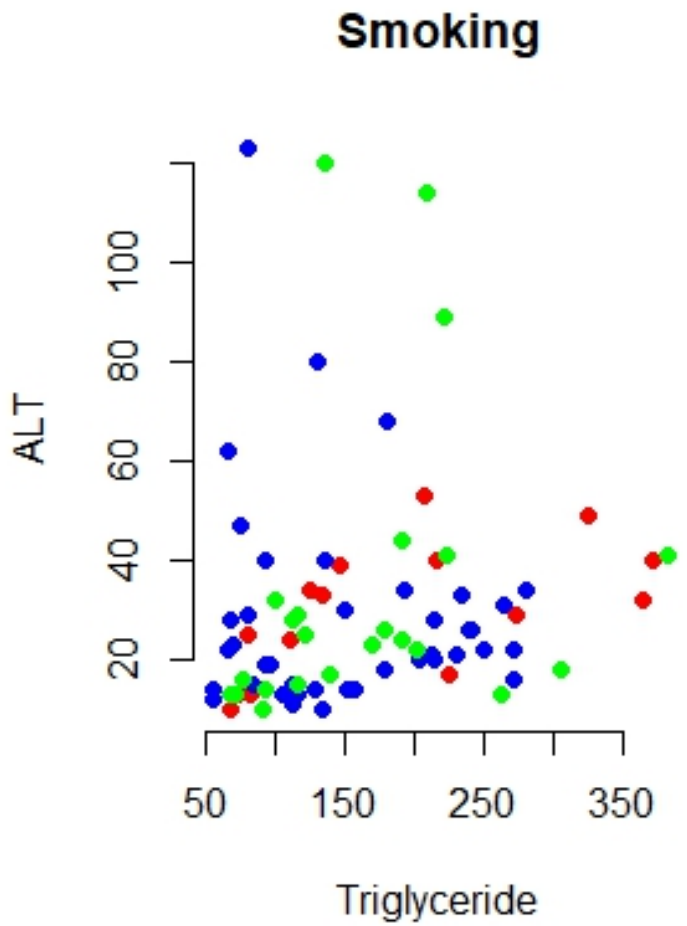
Let's go back to the Smoking example. Assume: DV1=ALT, DV2=Triglycerides, IV=Age Group (3 levels)



```
> m1 = manova(W~as.factor(smokew$age2))
> summary(m1,test='Wilks')
                         Df   Wilks approx F num Df den Df    Pr(>F)
as.factor(smokew$age2)    2 0.90218    5.176      4    392 0.0004504 ***
Residuals               197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.aov(m1)
 Response 1 :
                         Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smokew$age2)    2   5391 2695.37  9.4059 0.0001255 ***
Residuals               197  56453  286.56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
                         Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(smokew$age2)    2   14379  7189.4  1.4073 0.2473
Residuals               197 1006441  5108.8
```

Smoke Example: DV1=ALT, DV2=Triglyceride, IV1=Age Group (3 levels) and IV2=Smoking Status (2 levels)





```
> m1 = manova(W~as.factor(smokew$age2)*as.factor(smokew$smoking))
> summary(m1,test='Wilks')
                                                Df   Wilks approx F num Df den Df    Pr(>F)
as.factor(smokew$age2)                           2 0.89598   5.4478      4    386 0.0002821 ***
as.factor(smokew$smoking)                        1 0.84304  17.9665      2    193 6.989e-08 ***
as.factor(smokew$age2):as.factor(smokew$smoking)  2 0.98942   0.5146      4    386 0.7250322
Residuals                                       194
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(m1)
 Response 1 :
                                                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smokew$age2)                           2   5391 2695.37  9.8081 8.755e-05 ***
as.factor(smokew$smoking)                        1   2859 2859.42 10.4050  0.001475 **
as.factor(smokew$age2):as.factor(smokew$smoking)  2    280  139.83  0.5088  0.602008
Residuals                                       194  53314  274.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
                                                Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smokew$age2)                           2  14379    7189  1.6051    0.2035
as.factor(smokew$smoking)                        1 132680  132680 29.6228 1.573e-07 ***
as.factor(smokew$age2):as.factor(smokew$smoking)  2   4836    2418  0.5398    0.5837
Residuals                                       194 868925    4479
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> m1 = manova(W~as.factor(smokew$age2)+as.factor(smokew$smoking))
> summary(m1,test='Wilks')
                          Df   Wilks approx F num Df den Df    Pr(>F)
as.factor(smokew$age2)     2 0.89648   5.4755      4    390 0.0002682 ***
as.factor(smokew$smoking)  1 0.84382  18.0463      2    195 6.445e-08 ***
Residuals                196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(m1)
 Response 1 :
                          Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smokew$age2)     2   5391 2695.37  9.8575 8.334e-05 ***
as.factor(smokew$smoking)  1   2859 2859.42 10.4574  0.001434 **
Residuals                196  53593  273.43
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
                          Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(smokew$age2)     2  14379    7189  1.6127     0.202
as.factor(smokew$smoking)  1 132680  132680 29.7625 1.462e-07 ***
Residuals                196 873760    4458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```