R COURSE

VITOR SOUSA, LAURENT EXCOFFIER

Day 3, Linear models

Linear models with R

In this session, we are going to examine in details a simple linear model

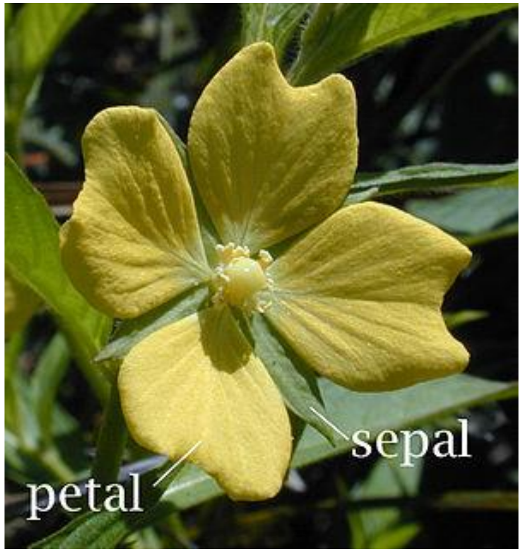
- Linear regression

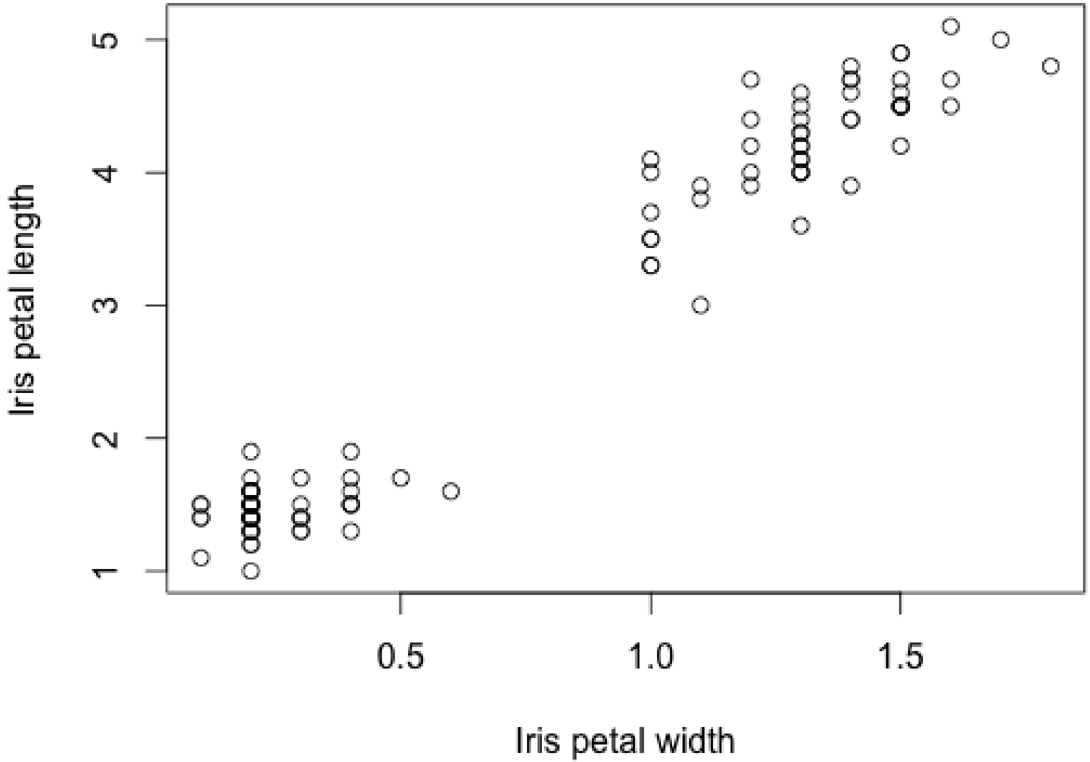
At the end of this document you will also find information on additional basic linear models, but we will not go over them here.

- One factor ANOVA (one way ANOVA)
- Two factor ANOVA (two way ANOVA)
- ANCOVA

Linear regression

Linear regression is a statistical method aiming at modeling a **linear relationship between two quantitative variables**, usually for making predictions.
The variable to be explained is sometimes called the **response or dependent variable**. The other variable is sometimes called the **explanatory or independent variable**.
For instance we have the following relationship between petal width and length in iris





Linear regression

Suppose that we have a given variable Y that we want to predict from an explanatory variable X.
We can write a simple linear model as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The ε variable represents noise or stochastic errors (e.g. in measurements), and the parameters β_0 and β_1 are unknown.

We usually want to estimate these parameters from *n* observations of couples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and a given observation y_i can be predicted as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

This is our model, and we estimate the parameters $(\beta_0; \beta_1)$ from the data.

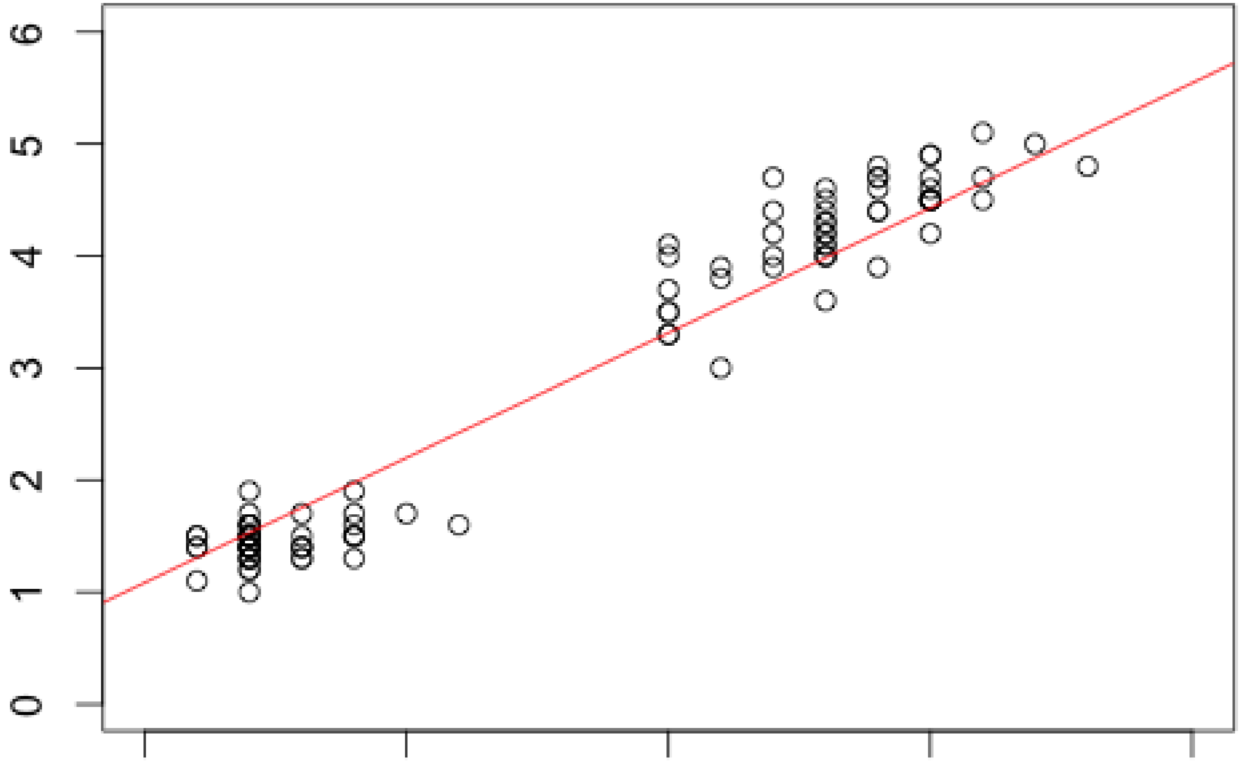
Regression line

b_0 corresponds to the y-axis intercept at *x*=0 and b_1 represents the slope of the **regression line** given by

$$f(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

which is plotted below in red. The predicted values of *y* are given by

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

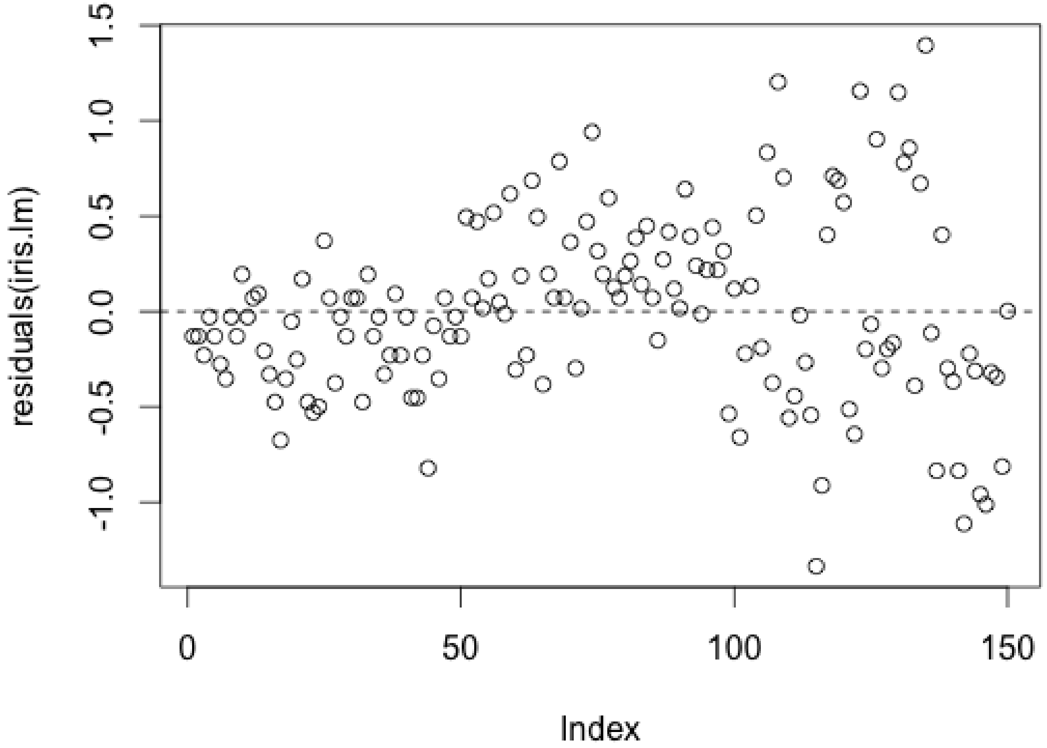


Residuals

The **residuals** ε_i represent the difference between the observed and predicted values of the Y dependent variable.
They are given by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Their examination allows one to check the adjustment of the model to the data and to see if there are any aberrant points.



Formulae

R is using a specific notation to describe relationships between variables. These are called formulae. They use the tilde sign "~" like in

$$Y \sim X$$

which describes here a linear relationship between the dependent variable *Y* and the explanatory variable *X*.
This is a compact notation that R is using instead of

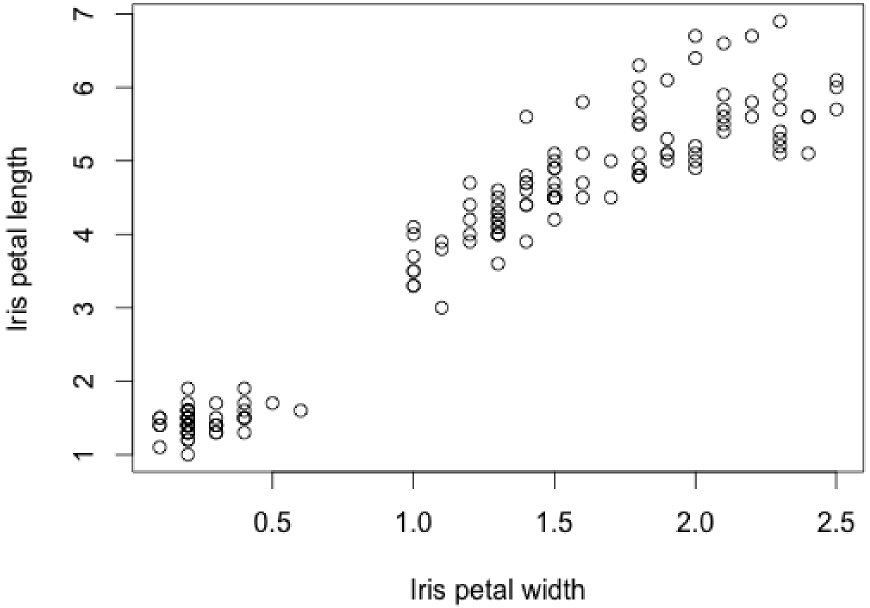
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

One can use formulae to describe more complex relationships between variables as well as more complex models, which can be linear or non-linear, with or without interaction terms between variables.

Let's examine the use of formulae and linear regression using the iris data set

Relation between iris petal width and length

```
data(iris)
str(iris)
'data.frame':150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 
1 1 1 1 ...
plot(iris$Petal.Width,iris$Petal.Length, xlab="Iris petal width",
  ylab="Iris petal length" )
```



Relation between iris petal width and length

A simple linear regression of a dependent variable Y on an explanatory variable X (abbreviated as: regression of Y on X) is simply computed in R with the **lm function**, which expects a formula as input.

If we want to regress iris petal length on iris petal width we can simply state:

```
iris.lm=lm(Petal.Length ~ Petal.Width, data=iris)
```

which is a slightly clearer notation than

```
iris.lm=lm(iris$Petal.Length ~ iris$Petal.Width)
```

Let’s have a look at the returned results found in iris.lm

```
names(iris.lm)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"       "call"          "terms"         "model"
```

Examining the estimated parameters

```
summary(iris.lm)

Call:
lm(formula = Petal.Length ~ Petal.Width, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.33542 -0.30347 -0.02955  0.25776  1.39453 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,  Adjusted R-squared:  0.9266 
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Examining the significance of the estimated parameters with the **summary** function

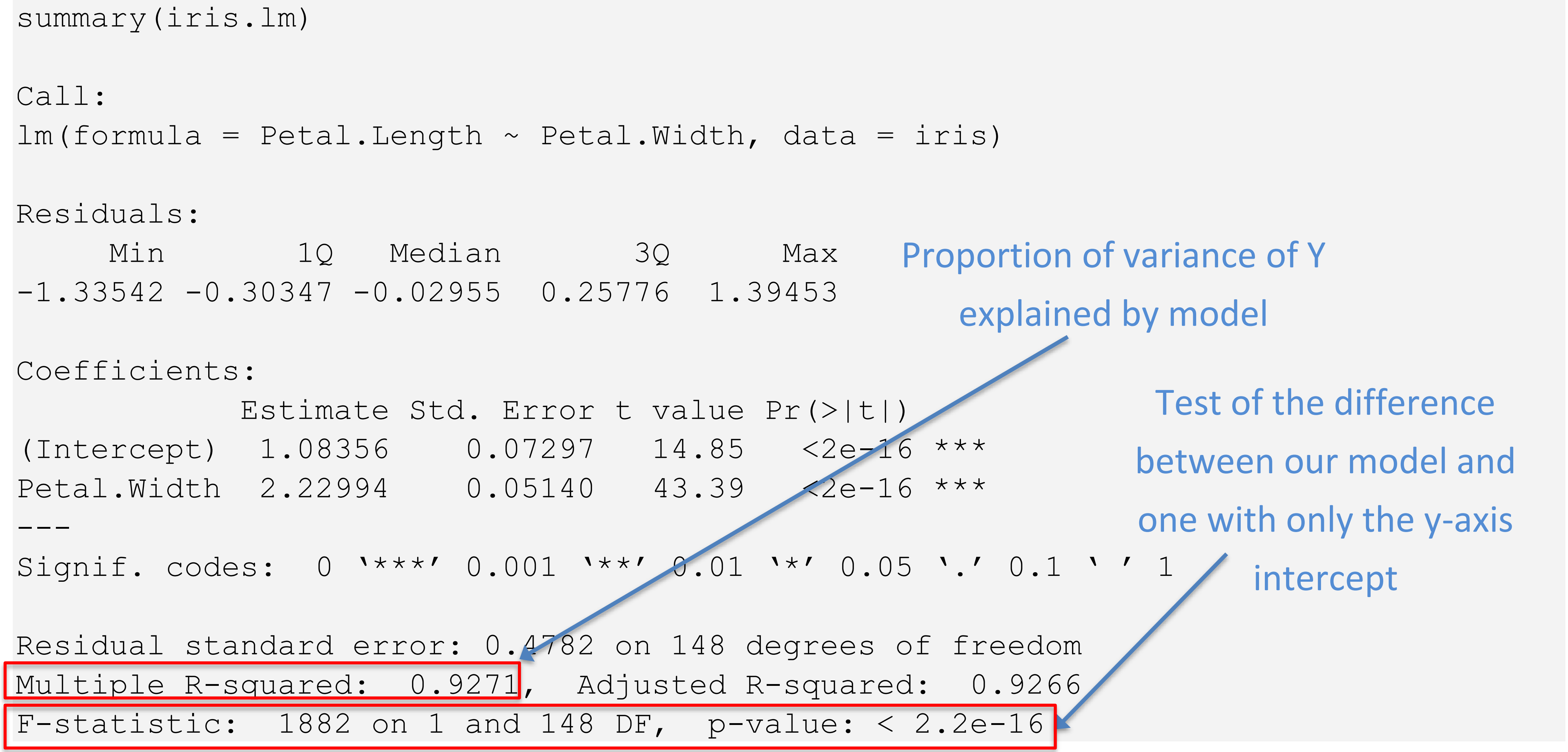
```
summary(iris.lm)

Call:
lm(formula = Petal.Length ~ Petal.Width, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.33542 -0.30347 -0.02955  0.25776  1.39453 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,	Adjusted R-squared:  0.9266 
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Intercept significantly different from zero

Slope significantly different from zero

There is thus a significant relationship between petal width and length!

Examining the estimated parameters



```
summary(iris.lm)

Call:
lm(formula = Petal.Length ~ Petal.Width, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33542 -0.30347 -0.02955  0.25776  1.39453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,  Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Proportion of variance of Y explained by model

Test of the difference between our model and one with only the y-axis intercept

Plotting the regression line

The estimated coefficients $(\beta_0;\beta_1)$ are given by

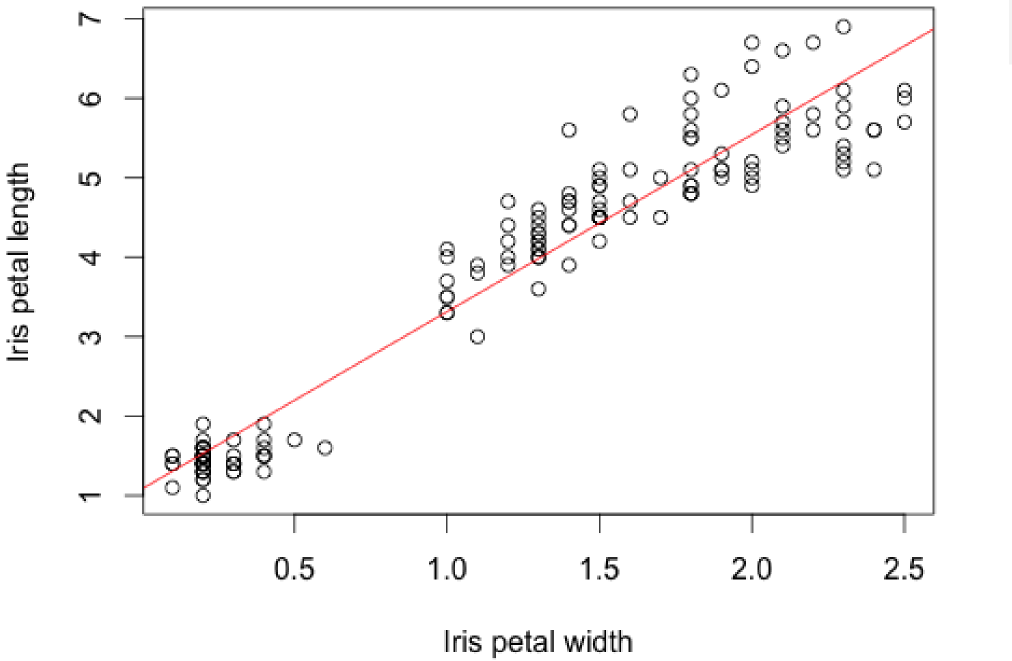
```
iris.lm$coefficients
(Intercept) Petal.Width
   1.083558    2.229940
```

The first column is β_0 (intercept) and the second is β_1 (slope)

You can directly plot the regression line on top of the scatter plot with abline, which knows how to handle the coefficients of an lm analysis

```
plot(iris$Petal.Width,iris$Petal.Length, xlab="Iris petal width",
     ylab="Iris petal length" )
abline(coef(iris.lm), col="red")
```

(note that coef(iris.lm) is equivalent to iris.lm$coefficients)

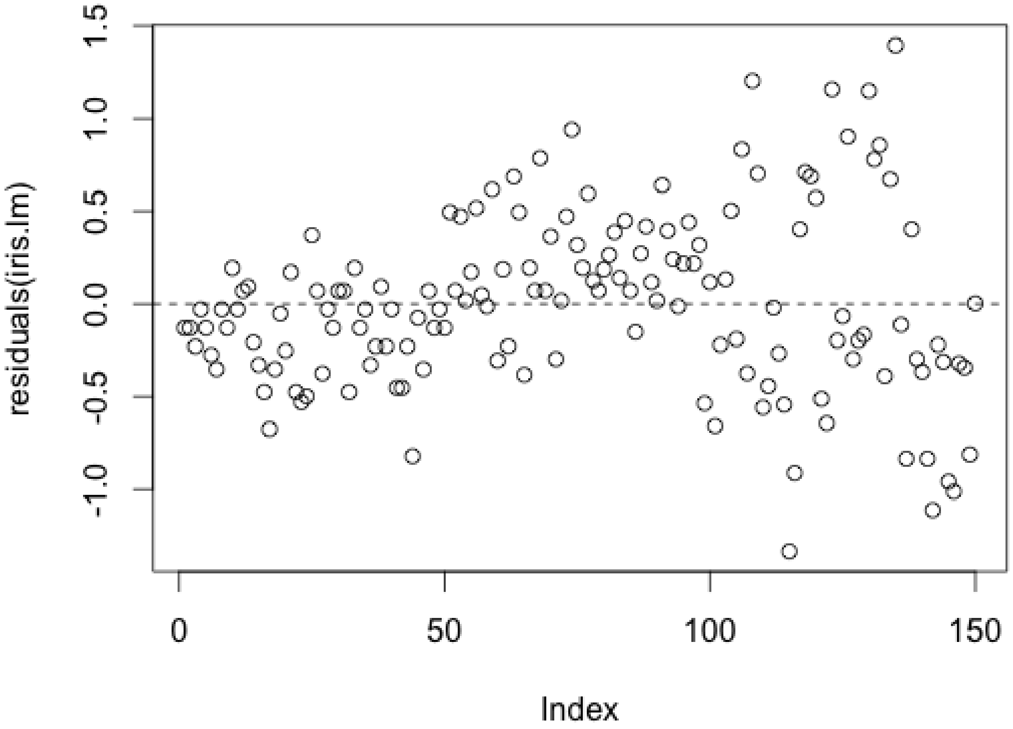


Plotting the residuals

We can easily plot the residuals of the regression analysis

```
plot(residuals(iris.lm)); abline(h=0, lty=2)
```

We see that the variance of the residuals increases for larger indices (heteroscedasticity)



Note that the sum of residuals should be zero by definition

But

```
sum(residuals(iris.lm))
[1] -2.94556e-15
```

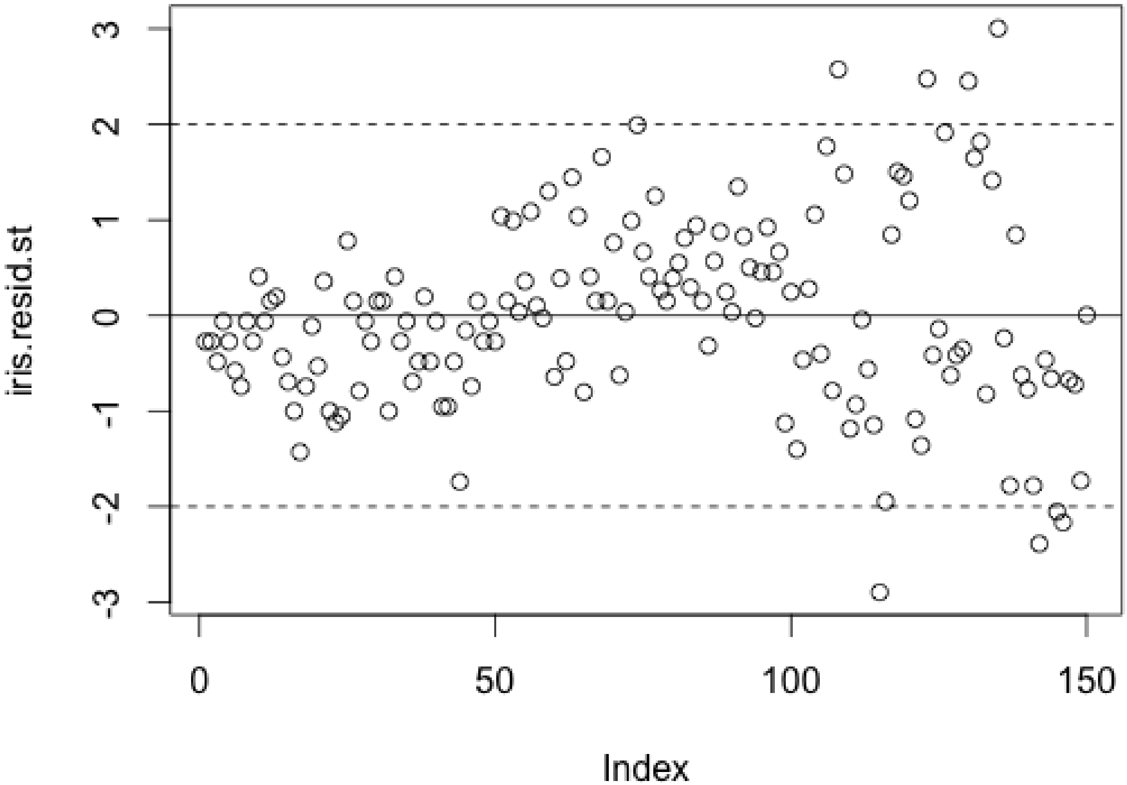
(not zero due to rounding errors in R)

Plotting the residuals (advanced)

We can **standardize the residuals** with the **rstudent** function and plot the limits of a 95% CI for the residuals.

```
iris.resid.st=rstudent(iris.lm)
plot(iris.resid.st)
abline(h=c(2,0,-2), lty=c(2,1,2))
```

We have about 8/150 points (5.33%) outside the 95% CI when we were expecting 5%.



Making predictions

Since we have built a model, we could use this model to be able to predict the petal length of a given flower with known petal width

Suppose we have a flower with petal width of 0.8 cm, what would be its expected length?

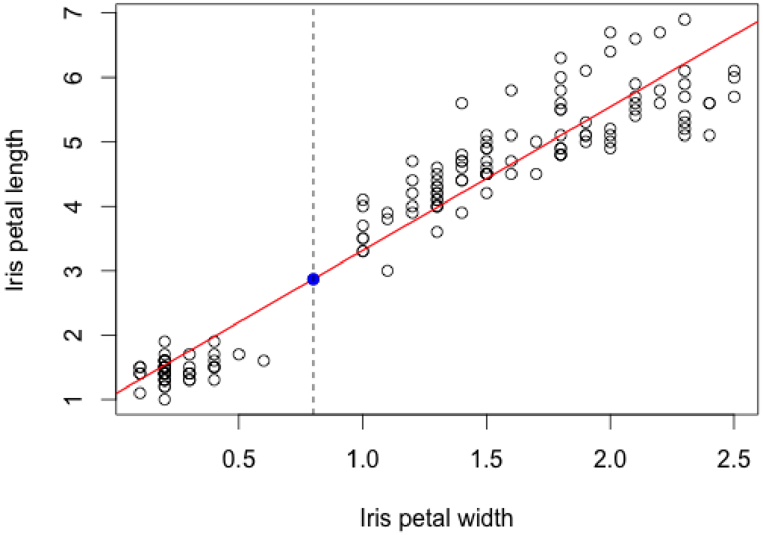
Making predictions

Since we have built a model, we could use this model to be able to predict the petal length of a given flower with known petal width

Suppose we have a flower with petal width of 0.8 cm, what would be its expected length?

We can use the **predict** function for this, but it needs the new X value to be given as a **data frame** similar to that analysed with lm

```
newXval=as.data.frame(0.8)
colnames(newXval)="Petal.Width"
predict(iris.lm, newXval)
       1
2.86751
```



Check…

```
abline(v=0.8,lty=2)
points(0.8,predict(iris.lm, newXval), pch=19, col="blue")
```

Confidence intervals (advanced notion)

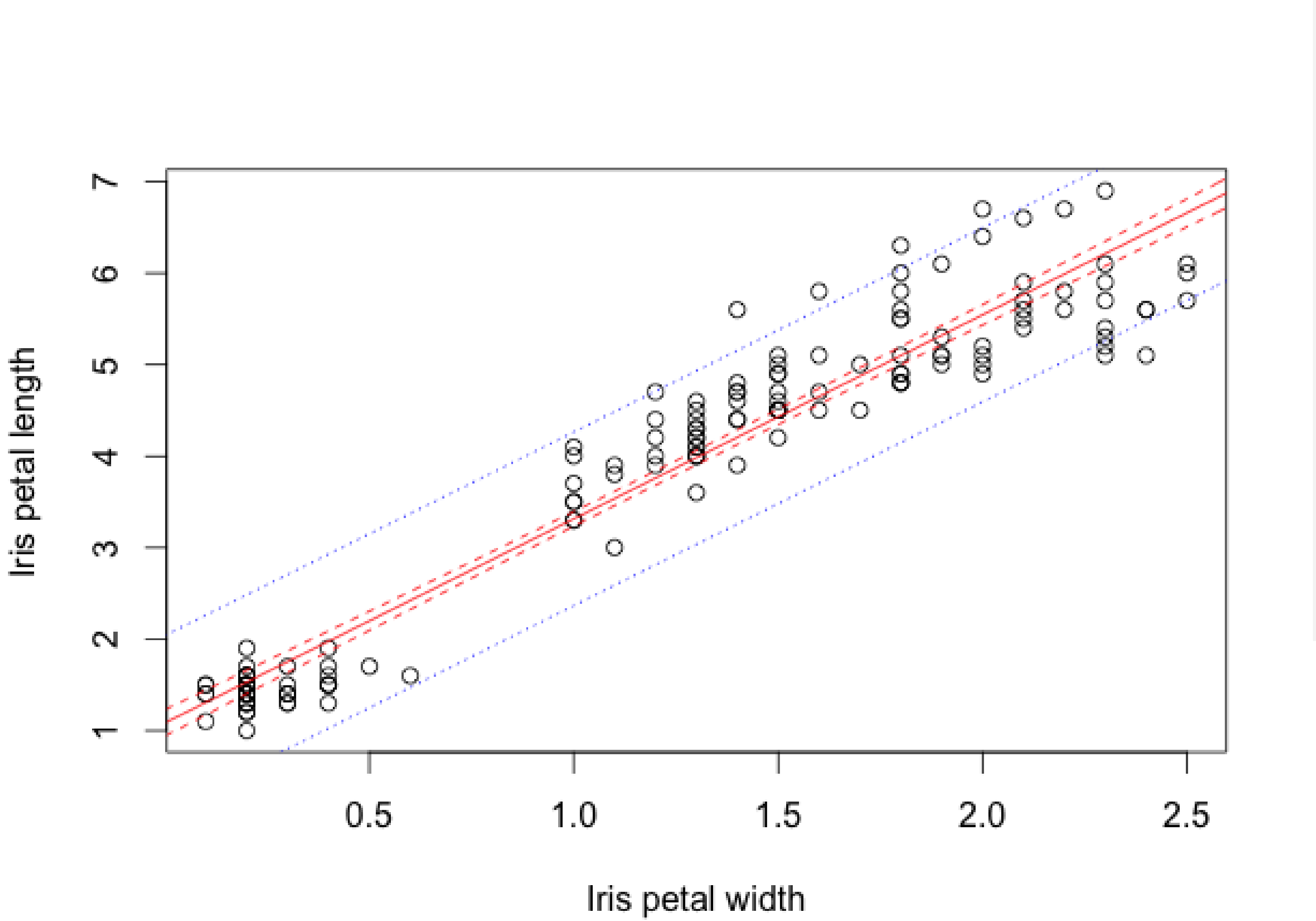
We can also use the predict function to get confidence intervals

```
#Building vector of x values
xvals=seq(0, 2.6, length.out=101)
xvals=as.data.frame(xvals)
colnames(xvals)="Petal.Width"

#Calculating CI for y values and regression line
CI95.yVal=predict(iris.lm, xvals, interval="pred", level=0.95)
CI95.regLine=predict(iris.lm, xvals, interval="conf", level=0.95)

#Plotting all points and lines
plot(iris$Petal.Width, iris$Petal.Length,
     xlab="Iris petal width",ylab="Iris petal length")
matlines(xvals, CI95.yVal, lty=c(0,3,3), col="blue")
matlines(xvals, CI95.regLine, lty=c(1,2,2), col="red")
```

Confidence intervals

We can also use the predict function to get confidence intervals

```
#Building vector of x values
xvals=seq(0, 2.6, length.out=101)
xvals=as.data.fra
colnames(xvals)="

#Calculating CI fo
CI95.yVal=predict
CI95.regLine=pred

#PLotting all poi
plot(iris$Petal.Wi
     xlab="Iris pe
matlines(xvals, CI
matlines(xvals, CI
```



One factor ANalysis Of Variance (ANOVA)

In a one factor ANOVA, one studies the relationship between a **quantitative variable**, say Y, and **a categorical (qualitative) variable**, say A.

Even though name suggests that we are analyzing the variance, this analysis **compares the means of Y for the different A categories** (taking into account their variances)

It can thus be considered as an **extension of a t-test to more than two categories**.

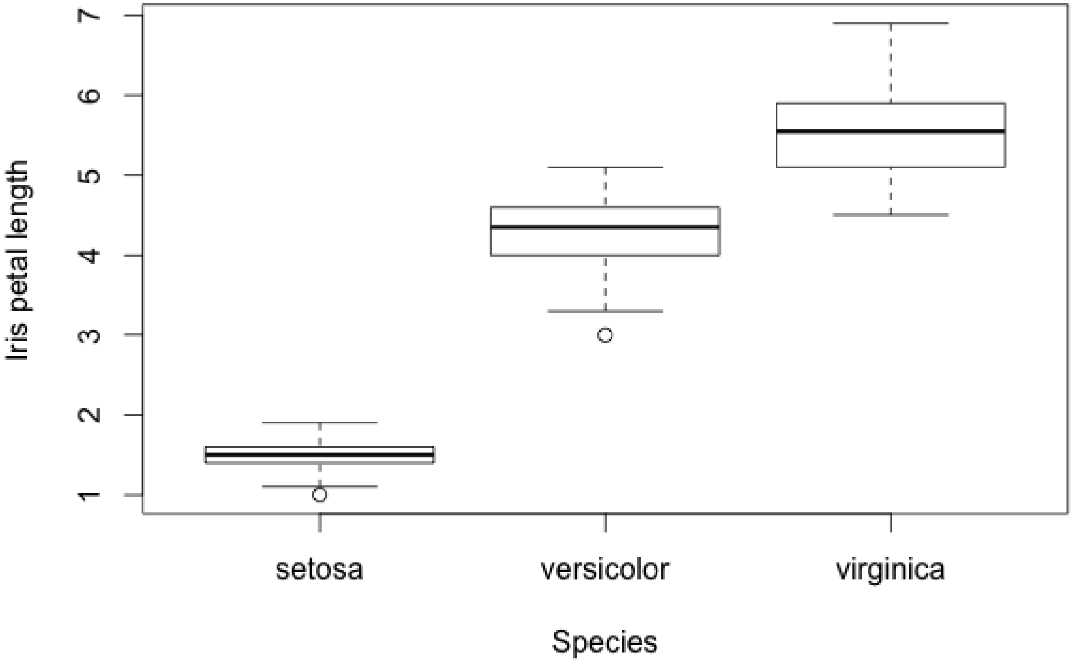
One factor ANalysis Of Variance (ANOVA)

For instance, taking the iris data set, we could want to compare the petal lengths among the different iris species.

```
boxplot(Petal.Length ~ Species, data=iris, ylab="Iris petal length",
        xlab="Species")
```

As suspected before, there are obvious differences between the three species for petal length

An ANOVA formalizes such a test



ANOVA with lm

ANOVAs are extremely simply performed in R. One uses the same syntax as for linear regression.

```
iris.anova=lm(Petal.Length ~ Species, data=iris)
```

lm automatically performs the ANOVA when the formula describes a quantitative variable as a function of a categorical variable.

ANOVA with lm

ANOVAs are extremely simply performed in R. One uses the same syntax as for linear regression.

```
iris.anova=lm(Petal.Length ~ Species, data=iris)
```

lm automatically performs the ANOVA when the formula describes a quantitative variable as a function of a categorical variable.

The results of the analysis can be visualized by the **anova** helper function

```
anova(iris.anova)
Analysis of Variance Table

Response: Petal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 437.10 218.551  1180.2 < 2.2e-16 ***
Residuals 147  27.22   0.185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA with lm

ANOVAs are extremely simply performed in R. One uses the same syntax as for linear regression.

```
iris.anova=lm(Petal.Length ~ Species, data=iris)
```

lm automatically performs the ANOVA when the formula describes a quantitative variable as a function of (a) categorical variable(s).

The results of the analysis can be visualized with the anova helper function

```
anova(iris.anova)
Analysis of Variance Table

Response: Petal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 437.10 218.551  1180.2 < 2.2e-16 ***
Residuals 147  27.22   0.185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Probability that the observed means of the different species are as different by chance alone

ANOVA as a linear model

Actually an ANOVA analysis can be expressed as a linear model where a given observation *j* of the variable Y in group *i* can be expressed as

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

with:

μ : global mean effect

α_i : additional effect of being in group i

ε_{ij} : residual for observation j

This model is contrasted to the model

$$y_{ij} = \mu + \varepsilon_{ij}$$

where it is assumed that there is no effect of the modality, or in other words that all groups have the same mean.

ANOVA estimated parameters

The parameters of the model can be examined with the **summary** function

```
summary(iris.anova)

Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max 
-1.260 -0.258  0.038  0.240  1.348 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,	Adjusted R-squared:  0.9406 
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

ANOVA estimated parameters

The parameters of the model can be examined with the summary function

```
summary(iris.anova)

Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max
-1.260 -0.258  0.038  0.240  1.348

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,  Adjusted R-squared:  0.9406
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

μ

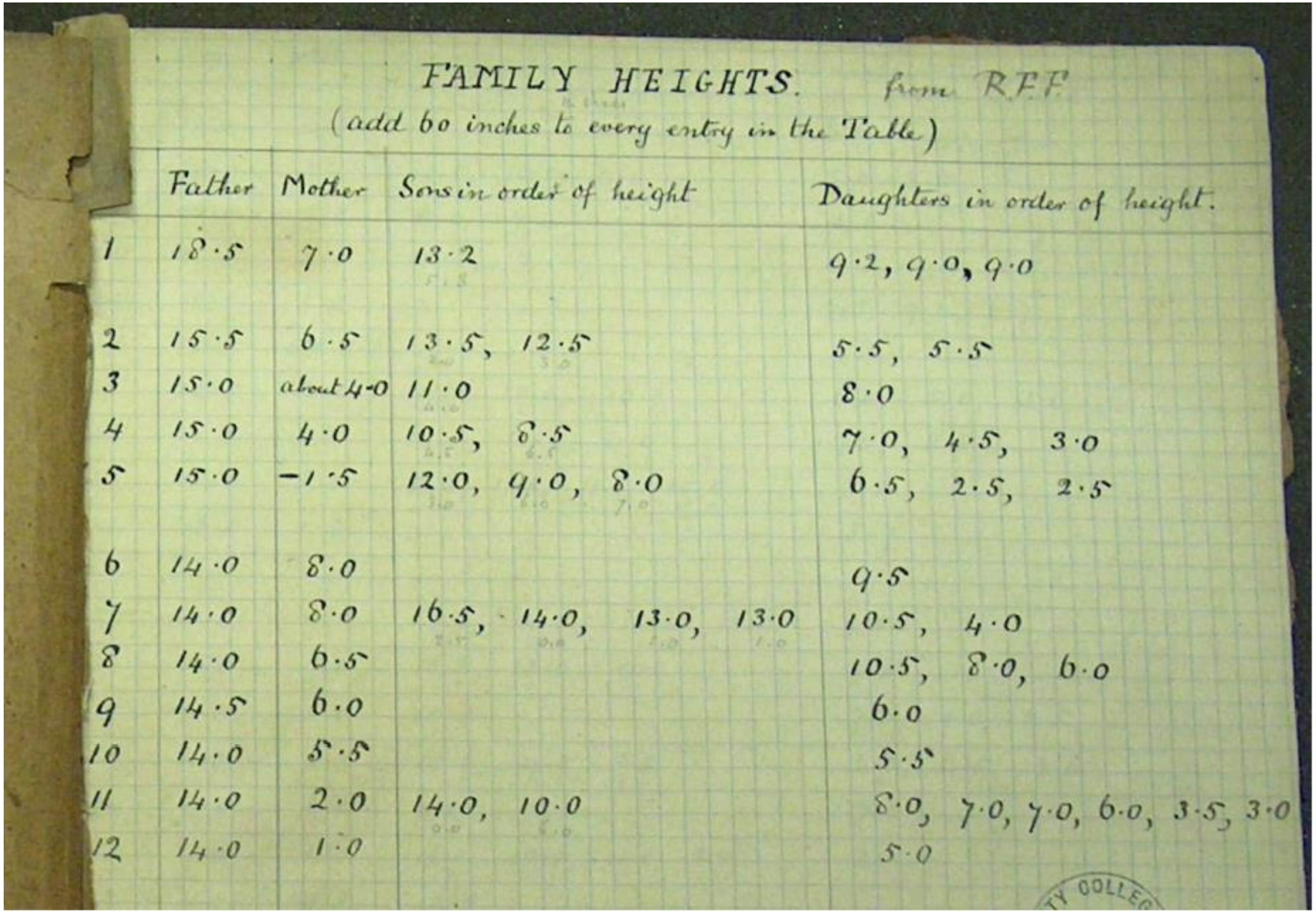
$\alpha_{versicolor}$

$\alpha_{virginica}$

Here the first group (*I. setosa*) is taken as a reference and $\alpha_{setosa} = 0$. It implies that the mean petal length of *I. setosa* is μ

Probability of simpler model

$y_{ij} = \mu + e_{ij}$



FAMILY HEIGHTS. from RFF

(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5
6	14·0	8·0		9·5
7	14·0	8·0	16·5, 14·0, 13·0, 13·0	10·5, 4·0
8	14·0	6·5		10·5, 8·0, 6·0
9	14·5	6·0		6·0
10	14·0	5·5		5·5
11	14·0	2·0	14·0, 10·0	8·0, 7·0, 7·0, 6·0, 3·5, 3·0
12	14·0	1·0		5·0

Figure 1. *Photograph of the entries for the first 12 families listed in Galton's notebook. Published with the permission of the Director of Library Services of University College London.*

Hanley (2004) The American Statistician

Advanced linear models

(for your own information, but not examined in this course)

Multi-factor ANOVA with interaction

In this analysis, one wants to model one quantitative variable (Y) as a function of several qualitative variables (A, B,…). Here we shall consider only two factors A and B.

In this case, we need to take into account the separate effects of the variables A and B, as well as their joint effects (interaction effects)

The model we will consider can thus be written as

effect of A effect of B

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

mean effect interaction residual

R modeling of 2-factor ANOVA

The previous model can be simply specified in R with the following formula

Y ~ A*B

An equivalent and perhaps more explicit formulation would be

Y ~ A+B+A:B

Where A:B specifies explicitly the interaction term to be tested

If one does not want to test the interaction (assumes that factors act additively only), then one can use the following formula

Y ~ A+B

or equivalently Y ~ A*B-A:B

Example of 2-factor ANOVA with interaction

We shall analyze the MASS::cabbages data set, which reports data from a cabbage field experiment.

Load the MASS library and the cabbages data set

```
require(MASS)
data(cabbages)
```

Look at its structure

```
str(cabbages)
'data.frame':60 obs. of  4 variables:
 $ Cult   : Factor w/ 2 levels "c39","c52": 1 1 1 1 1 1 1 1 1 1 ...
 $ Date   : Factor w/ 3 levels "d16","d20","d21": 1 1 1 1 1 1 1 1 1 1 ...
 $ HeadWt: num  2.5 2.2 3.1 4.3 2.5 4.3 3.8 4.3 1.7 3.1 ...
 $ VitC  : int  51 55 45 42 53 50 50 52 56 49 ..
```

Cult: 2 two cabbage cultivars (plants selected for some desirable characteristics)

Date: 3 planting dates

HeadWt: Cabbage head weight (kg)

VitC: amount of ascorbic acid (Vit C) received (information discarded here)

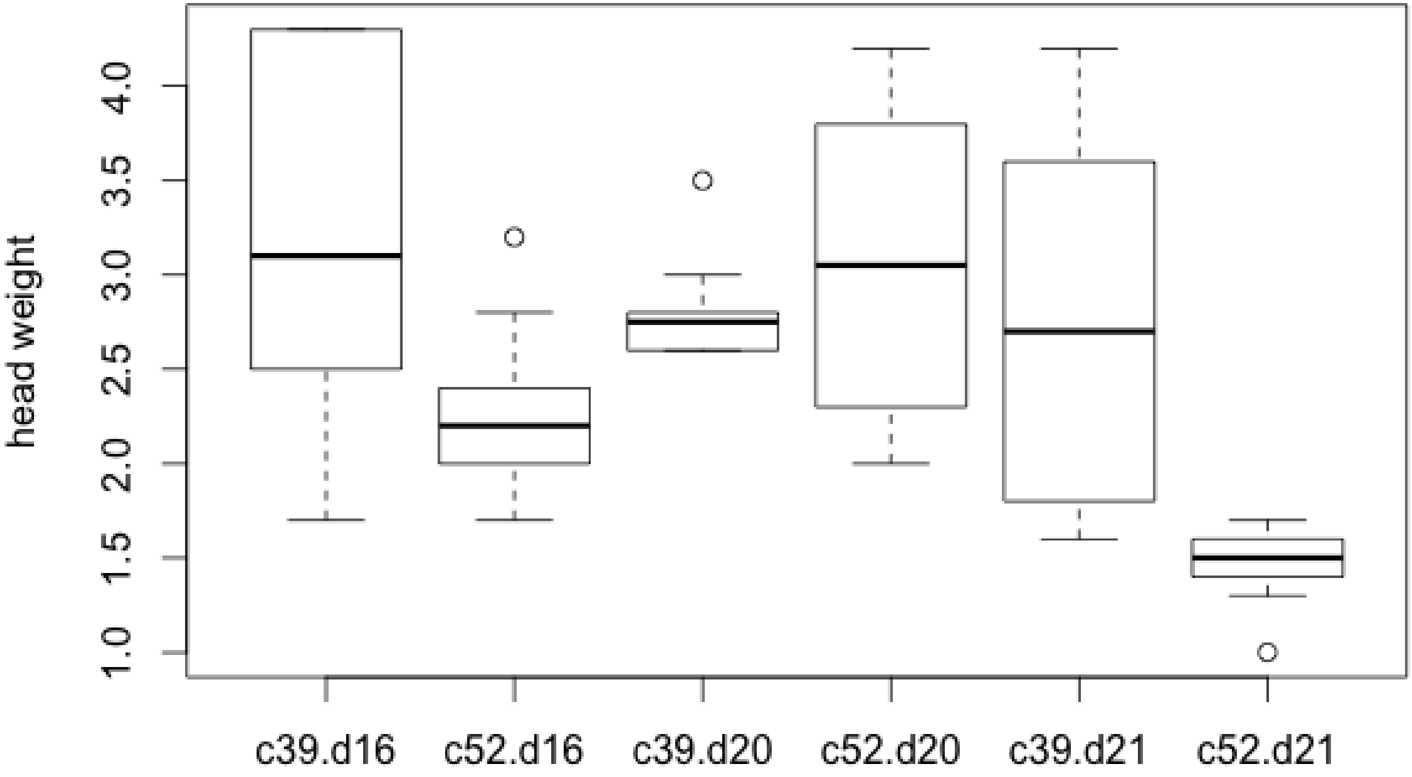
Question: How does the head weight depend on Cult and Date?

Look at the data

We can also use a formula in boxplot to specify we would like to plot all types of weight for all combinations of Cult and Date

```
boxplot(HeadWt ~ Cult*Date, data=cabbages)
```

The mean weights of the different combinations are quite different

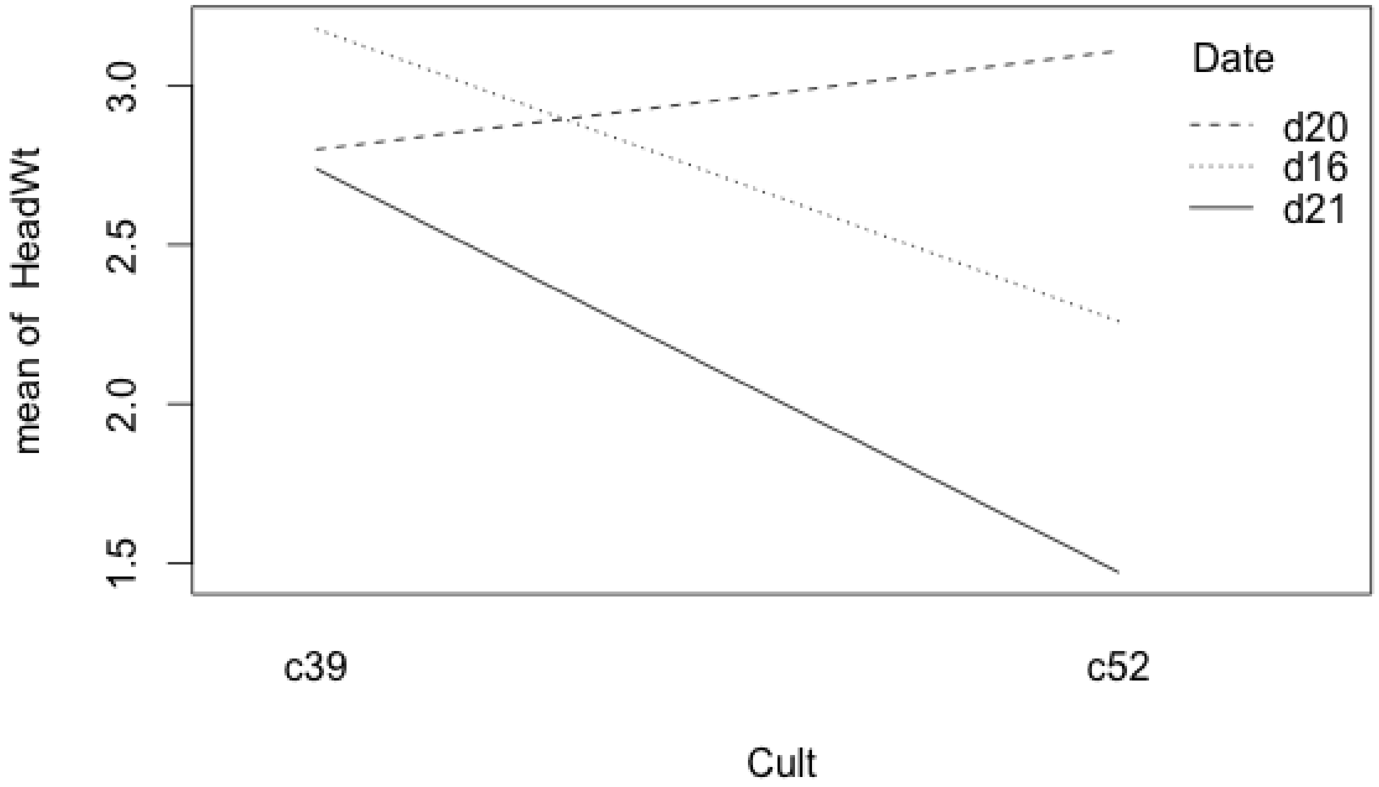


Look at the data

Use an **interaction plot** to visualize how the means of the different combinations of factors differ

```
with(cabbages, interaction.plot(Cult, Date, HeadWt))
```

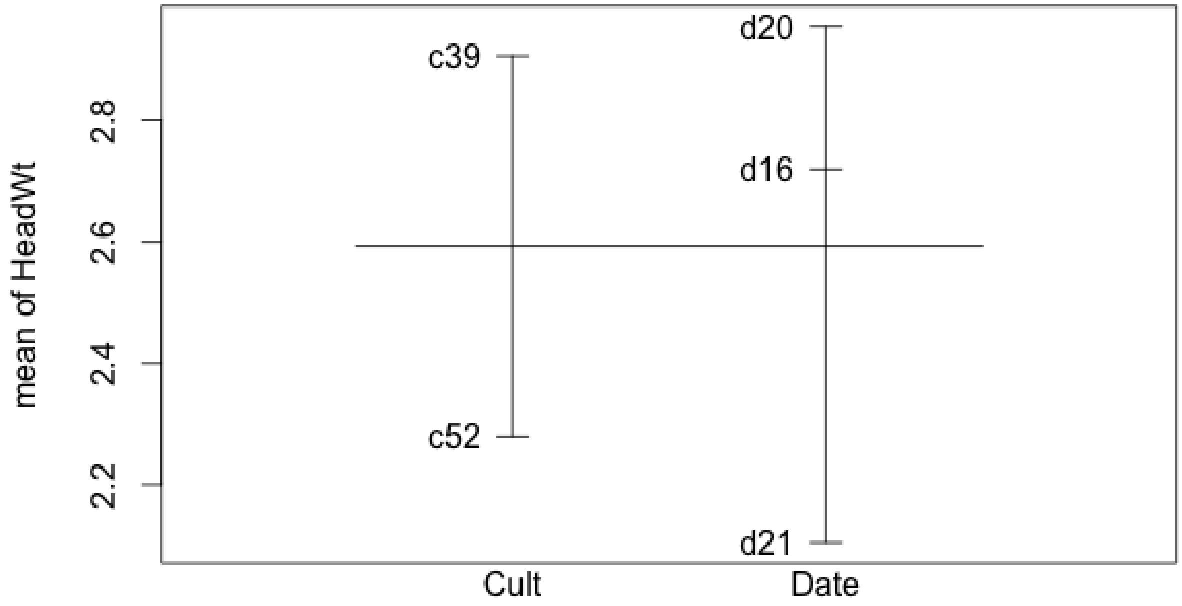
Note the use of the **with** function, and the order of the variables (explanatory first, and dependent last)



Look at the data

Finally, another way to plot differences in means of single categories is to use the **plot.design** function

```
plot.design(cabbages)
```



ANOVA results

Let’s perform our 2 factor ANOVA (2-way ANOVA):

```
>cab.anova=lm(HeadWt ~ Cult*Date, data=cabbages)
>anova(cab.anova)
Analysis of Variance Table

Response: HeadWt
          Df  Sum Sq Mean Sq F value    Pr(>F)
Cult       1  5.8907  5.8907 12.4969 0.0008451 ***
Date       2  7.7063  3.8532  8.1744 0.0007920 ***
Cult:Date  2  6.8863  3.4432  7.3046 0.0015571 **
Residuals 54 25.4540  0.4714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA results

Let’s perform our 2 factor ANOVA:

```
>cab.anova=lm(HeadWt ~ Cult*Date, data=cabbages)
>anova(cab.anova)
Analysis of Variance Table

Response: HeadWt
          Df  Sum Sq Mean Sq F value    Pr(>F)
Cult       1  5.8907  5.8907 12.4969 0.0008451 ***
Date       2  7.7063  3.8532  8.1744 0.0007920 ***
Cult:Date  2  6.8863  3.4432  7.3046 0.0015571 **
Residuals 54 25.4540  0.4714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interaction term is significant: It implies that the weight depends on the interaction between the cultivar and the planting date

Distinguishing between models

We have seen that all our factors were significant, but we might want to compare the fit of this model with that of simpler ones, for instance let's compare

HeadWt ~ Cult * Date

with

HeadWt ~ Cult + Date

```
hw1=lm(HeadWt ~ Cult*Date, data=cabbages)
hw2=lm(HeadWt ~ Cult+Date, data=cabbages)
#Performs an analysis of variance on associated model residuals
anova(hw2,hw1)
Analysis of Variance Table

Model 1: HeadWt ~ Cult + Date
Model 2: HeadWt ~ Cult * Date
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     56 32.340
2     54 25.454  2    6.8863 7.3046 0.001557 **
```

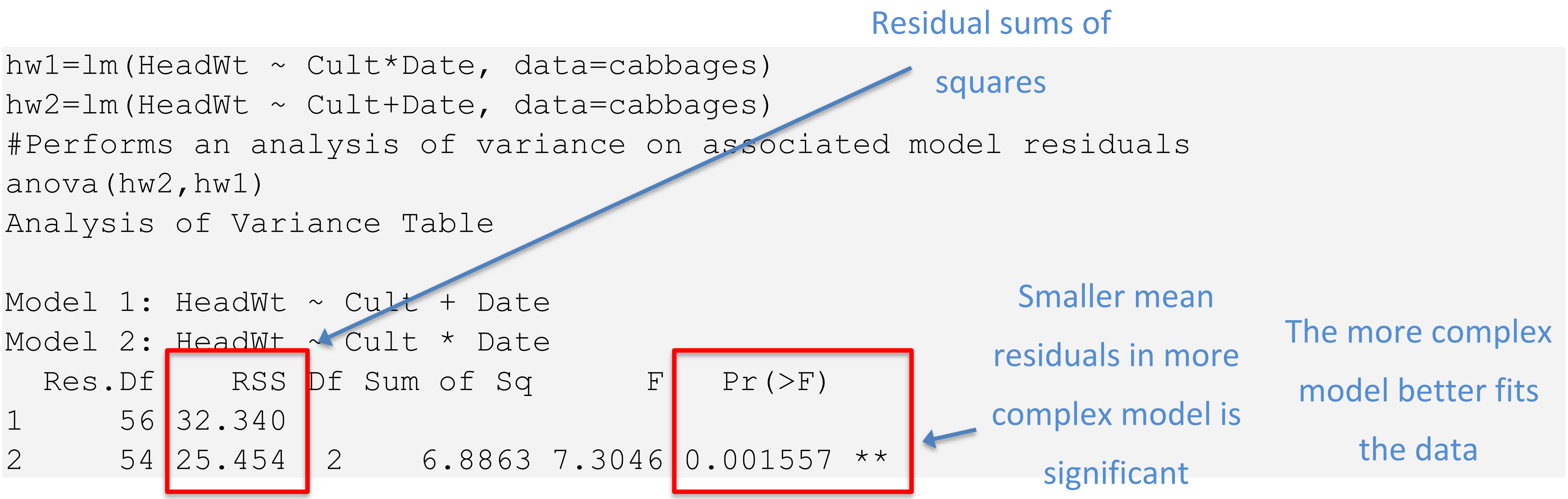
Distinguishing between models

We have seen that all our factors were significant, but we might want to compare the fit of this model with that of simpler ones, for instance let’s compare

HeadWt ~ Cult * Date

with

HeadWt ~ Cult + Date



```
hw1=lm(HeadWt ~ Cult*Date, data=cabbages)
hw2=lm(HeadWt ~ Cult+Date, data=cabbages)
#Performs an analysis of variance on associated model residuals
anova(hw2,hw1)
Analysis of Variance Table

Model 1: HeadWt ~ Cult + Date
Model 2: HeadWt ~ Cult * Date
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     56 32.340
2     54 25.454  2    6.8863 7.3046 0.001557 **
```

Residual sums of squares

Smaller mean residuals in more complex model is significant

The more complex model better fits the data

Analysis of covariance - ANCOVA

ANCOVA is about the analysis of **a quantitative variable** to be **explained by quantitative and qualitative variables**.

The simplest model would be that of a quantitative variables Y to be explained by another quantitative variable X and a categorical variable A with *I* categories.

Since the relationship between Y and X may depend on the category of A, one performs separate regressions for each of the *I* categories.

The model for the *i*-th category would thus look like

$$y_{i,j} = \alpha_i + \gamma_i \, x_{i,j} + \varepsilon_{i,j}$$

Analysis of covariance - ANCOVA

ANCOVA is about the analysis of a quantitative variable to be explained by quantitative and qualitative variables.

The simplest model would be that of a quantitative variables Y to be explained by another quantitative variable X and a categorical variable A with *I* categories.

Since the relationship between Y and X may depend on the category of A, one performs separate regressions for each of the *I* categories.

The model for the *i*-th category would thus look like

$$y_{i,j} = \alpha_i + \gamma_i \, x_{i,j} + \varepsilon_{i,j}$$

Effect of category i (intercept)

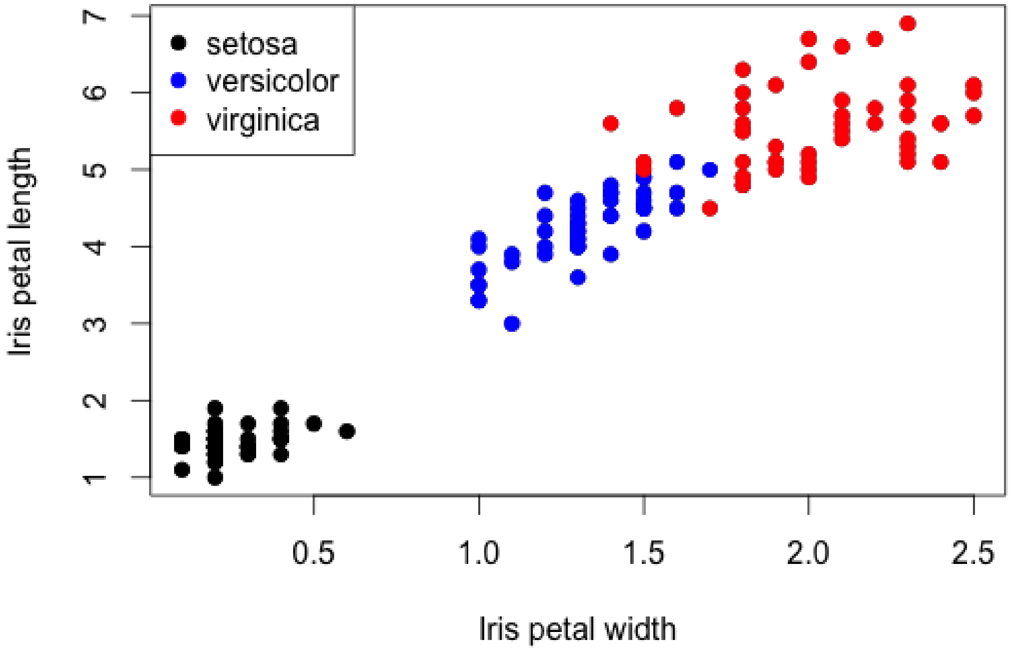
slope effect for category i

residuals

ANCOVA applied example

Let’s consider the iris example again.

We have previously analyzed the relationship between petal length and petal width, by pooling data from the three species, but the three species have almost non-overlapping distributions.



We could thus **ask whether the relationship between these two quantitative variables differs between species**.

Defining the ANCOVA model in R

We would thus perform an ANCOVA to test if the relationship between the size and the shape of the petals is different in the three species

This is straightforward in R as we just have to analyze the model

Petal.Length ~ Petal.Width*Species

R will understand that we need to do an ANCOVA since we want to explain a quantitative variable by both a quantitative and a qualitative variable

Performing the ANCOVA

We can again use the lm function to perform the ANCOVA

```
iris.ancova=lm(Petal.Length ~ Petal.Width*Species, data=iris)
anova(iris.ancova)

Analysis of Variance Table

Response: Petal.Length
                     Df  Sum Sq Mean Sq   F value    Pr(>F)    
Petal.Width           1  430.48  430.48 3294.5561 < 2.2e-16 ***
Species               2   13.01    6.51   49.7891 < 2.2e-16 ***
Petal.Width:Species   2    2.02    1.01    7.7213 0.0006525 ***
Residuals           144   18.82    0.13
```

Performing the ANCOVA

We can again use the lm function to perform the ANCOVA

```
iris.ancova=lm(Petal.Length ~ Petal.Width*Species, data=iris)
anova(iris.ancova)

Analysis of Variance Table

Response: Petal.Length
                     Df  Sum Sq Mean Sq   F value     Pr(>F)
Petal.Width           1  430.48  430.48 3294.5561 < 2.2e-16 ***
Species               2   13.01    6.51   49.7891 < 2.2e-16 ***
Petal.Width:Species   2    2.02    1.01    7.7213 0.0006525 ***
Residuals           144   18.82    0.13
```

All effects are significant:
Petal length depends on i) petal width, ii) species, and iii) the interaction between species and width

Looking at estimated parameters

The interpretation of the summary is not obvious and requires some explanations

```
summary(iris.ancova)
Call:
lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)
Residuals:
     Min       1Q   Median       3Q      Max
-0.84099 -0.19343 -0.03686  0.16314  1.17065
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.3276     0.1309  10.139  < 2e-16 ***
Petal.Width                     0.5465     0.4900   1.115   0.2666
Speciesversicolor               0.4537     0.3737   1.214   0.2267
Speciesvirginica                2.9131     0.4060   7.175 3.53e-11 ***
Petal.Width:Speciesversicolor   1.3228     0.5552   2.382   0.0185 *
Petal.Width:Speciesvirginica    0.1008     0.5248   0.192   0.8480
---
Residual standard error: 0.3615 on 144 degrees of freedom
Multiple R-squared:  0.9595,  Adjusted R-squared:  0.9581
F-statistic: 681.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

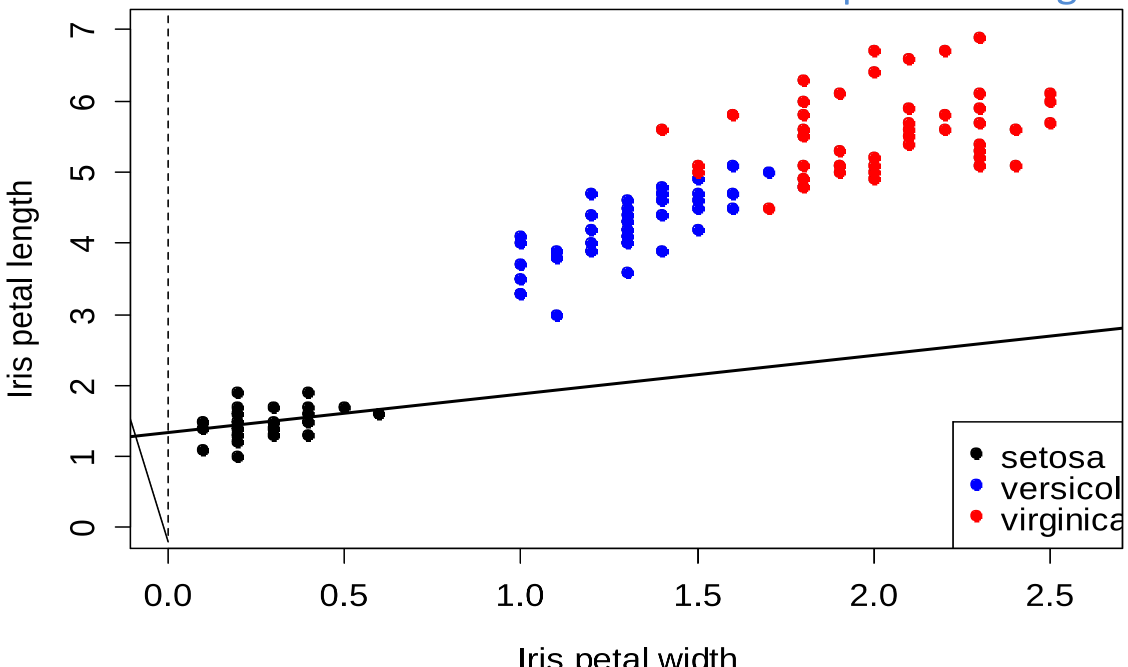
Looking at estimated parameters

Let’s concentrate on the coefficients:

Intercept of the regression for 1st category (setosa) = α_{setosa}

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.3276     0.1309  10.139  < 2e-16 ***
Petal.Width                     0.5465     0.4900   1.115   0.2666    
Speciesversicolor               0.4537     0.3737   1.214   0.2267    
Speciesvirginica                2.9131     0.4060   7.175 3.53e-11 ***
Petal.Width:Speciesversicolor   1.3228     0.5552   2.382   0.0185 *  
Petal.Width:Speciesvirginica    0.1008     0.5248   0.192   0.8480    
```

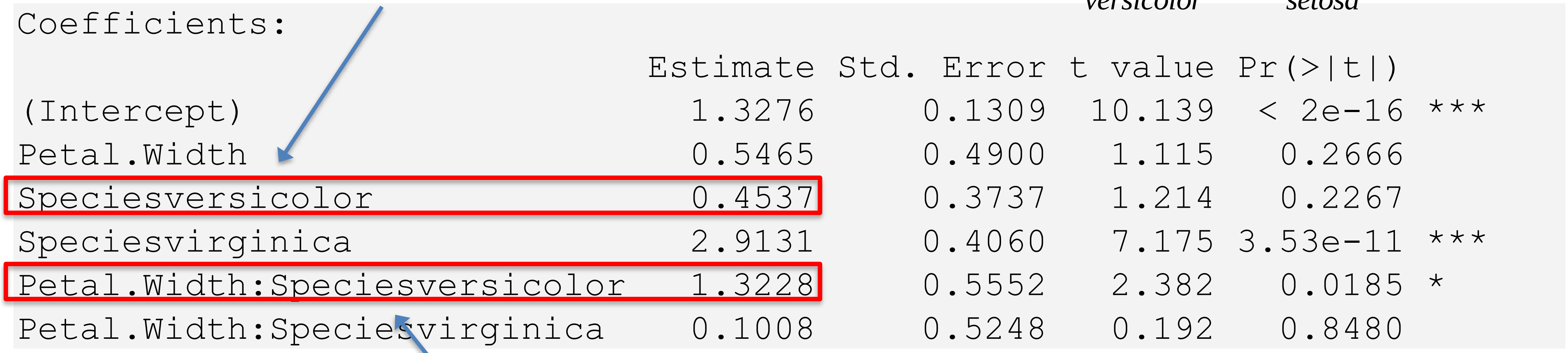
Slope of the regression for 1st category (setosa) = γ_{setosa}



Looking at estimated parameters

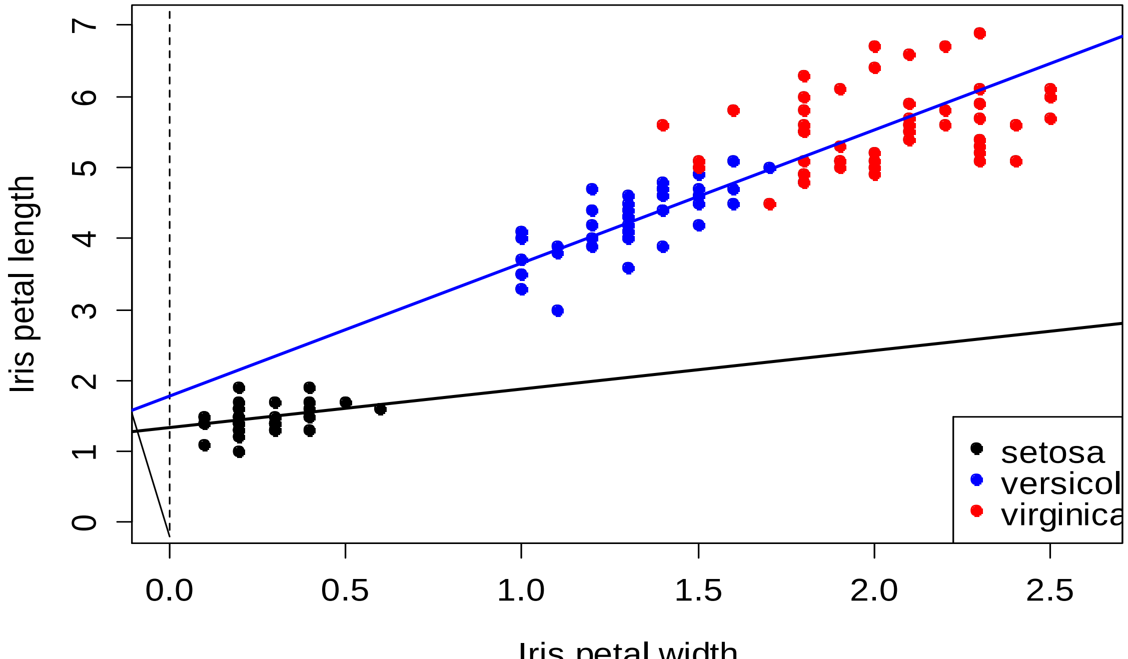
Let’s concentrate on the coefficients:

Difference in intercept between versicolor and setosa = $\alpha_{versicolor} - \alpha_{setosa}$



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.3276     0.1309  10.139  < 2e-16 ***
Petal.Width                     0.5465     0.4900   1.115   0.2666    
Speciesversicolor               0.4537     0.3737   1.214   0.2267    
Speciesvirginica                2.9131     0.4060   7.175 3.53e-11 ***
Petal.Width:Speciesversicolor   1.3228     0.5552   2.382   0.0185 *  
Petal.Width:Speciesvirginica    0.1008     0.5248   0.192   0.8480    
```

Difference in slope between versicolor and setosa = $\gamma_{versicolor} - \gamma_{setosa}$



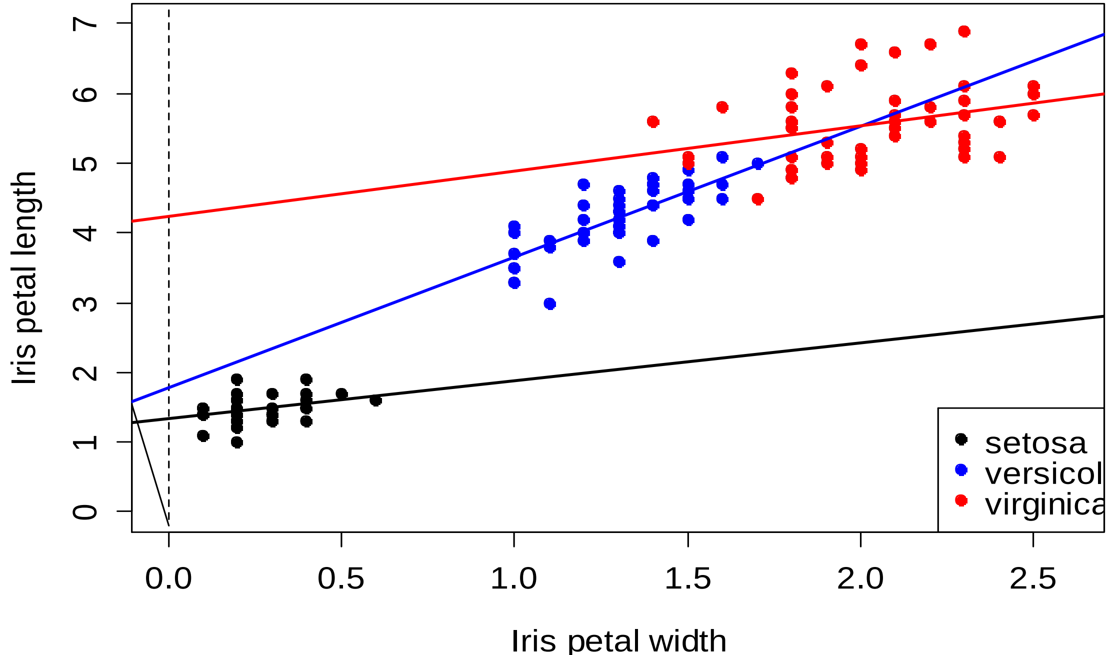
Looking at estimated parameters

Let’s concentrate on the coefficients:

Difference in intercept between virginica and setosa = $\alpha_{virginica} - \alpha_{setosa}$

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.3276     0.1309  10.139  < 2e-16 ***
Petal.Width                    0.5465     0.4900   1.115   0.2666
Speciesversicolor              0.4537     0.3737   1.214   0.2267
Speciesvirginica               2.9131     0.4060   7.175 3.53e-11 ***
Petal.Width:Speciesversicolor  1.3228     0.5552   2.382   0.0185 *
Petal.Width:Speciesvirginica   0.1008     0.5248   0.192   0.8480
```

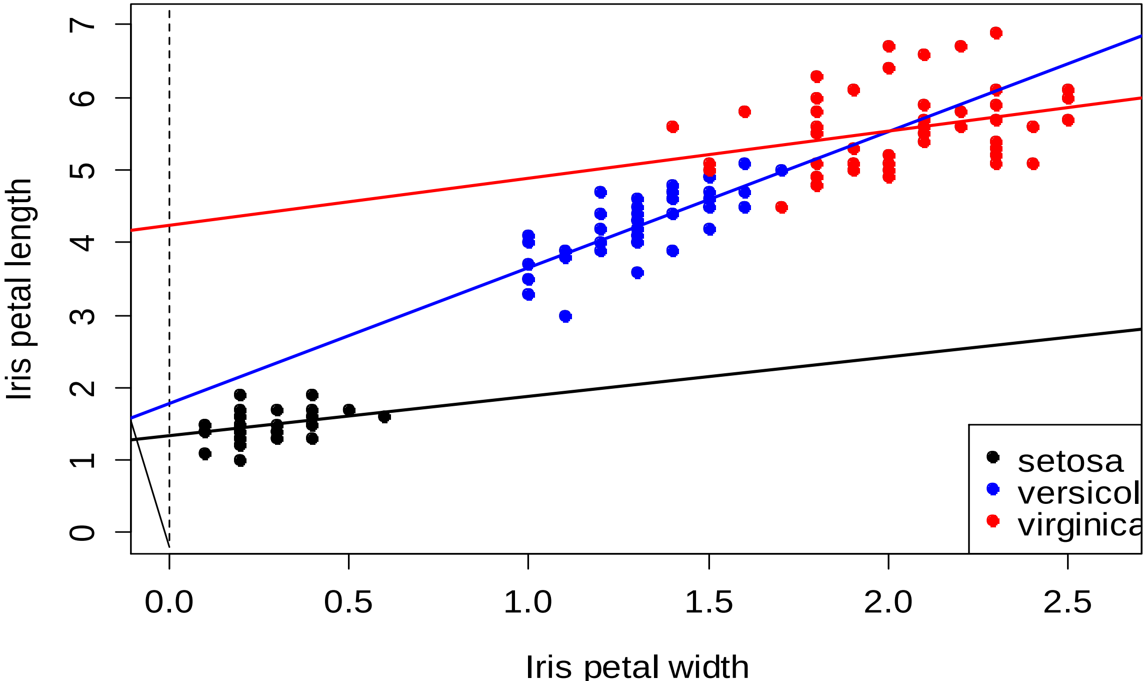
Difference in slope between virginica and setosa = $\gamma_{virginica} - \gamma_{setosa}$



Interpretation of ANCOVA results

Our ANCOVA analysis of the iris data show the following results:

1) Petal length significantly depends on petal width, on the species, and on the interaction between petal width and the species.
2) The relationships between petal length and width differs between species
   a. In **setosa**, length does not depend on width (slope is not significant)
   b. In **versicolor**, the intercept is the same as in **setosa** but the slope is different
   c. In **virginica**, the intercept is not the same as in **setosa** but the slope is the same: **virginica** petals are significantly longer than in **setosa** for the same width

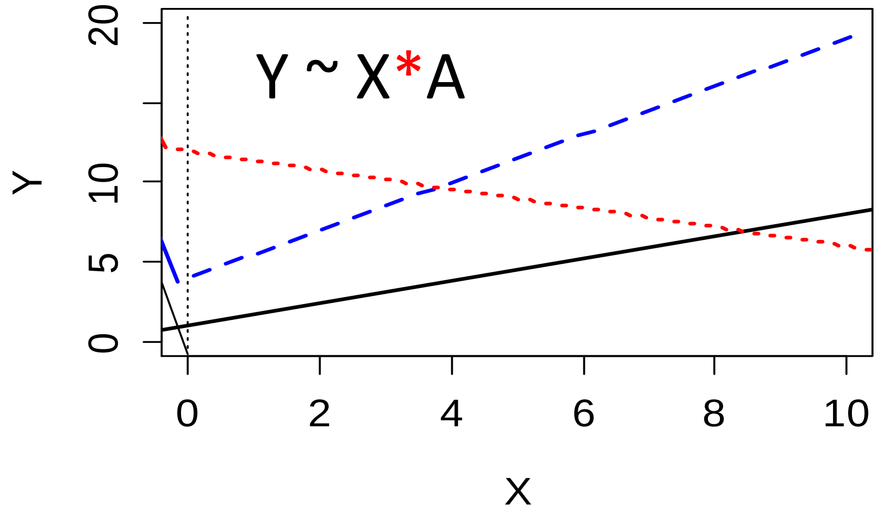


Exercice

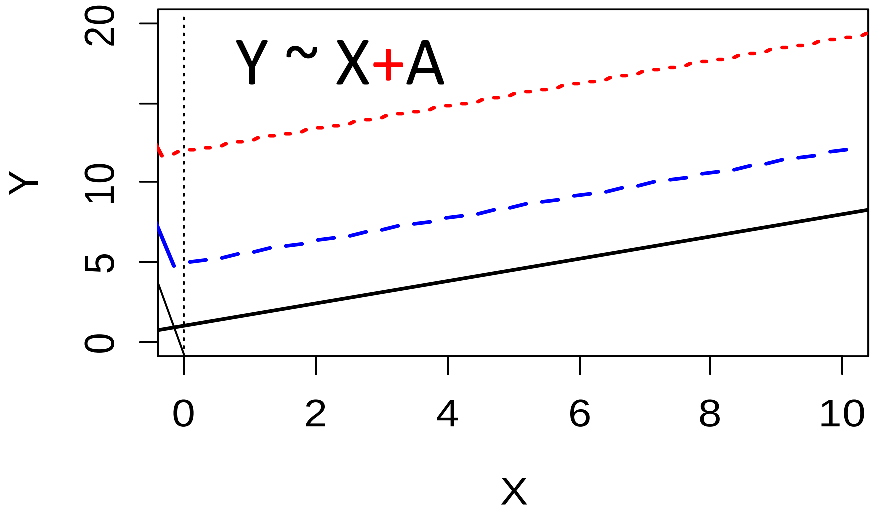
Try to perform the ancova analysis on the iris data set and plot the regression lines on top of the plot as shown in the former slide

Comparing models

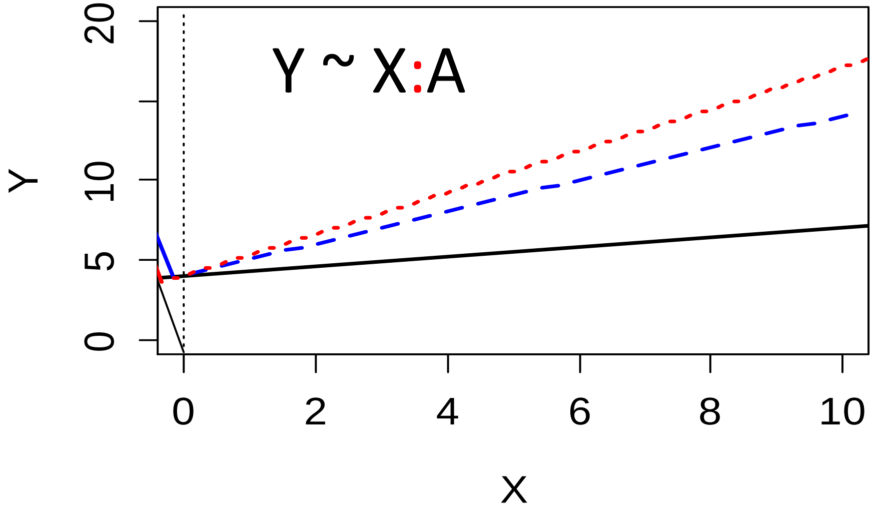
Different intercepts, different slopes



Same slopes, different intercepts



Same intercepts, different slopes



Comparing models

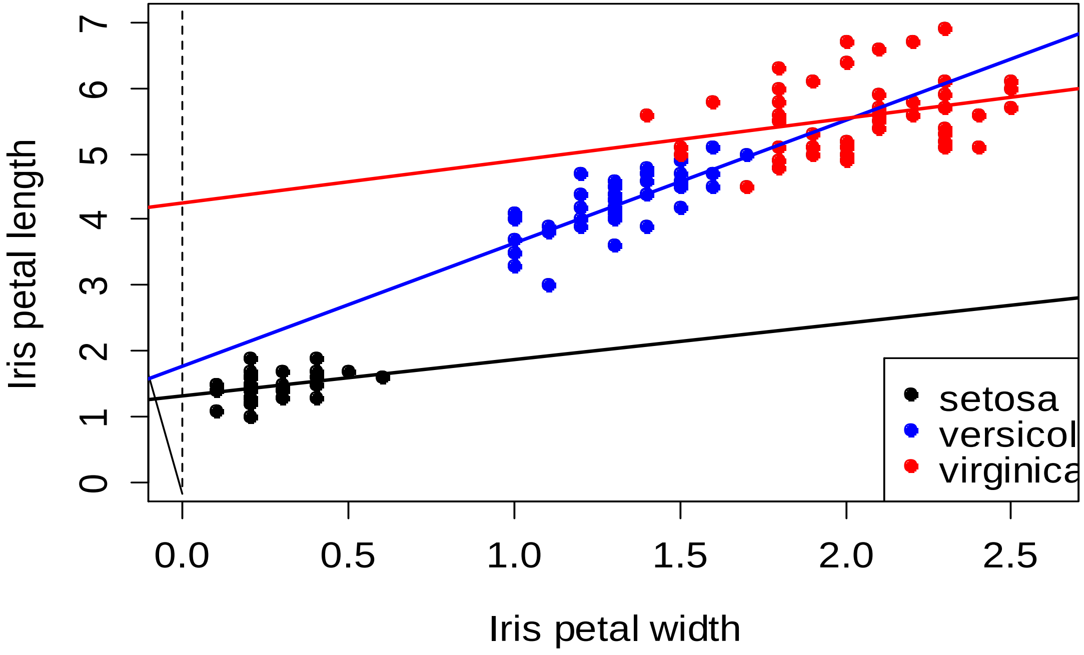
We can build the different iris models in R very simply

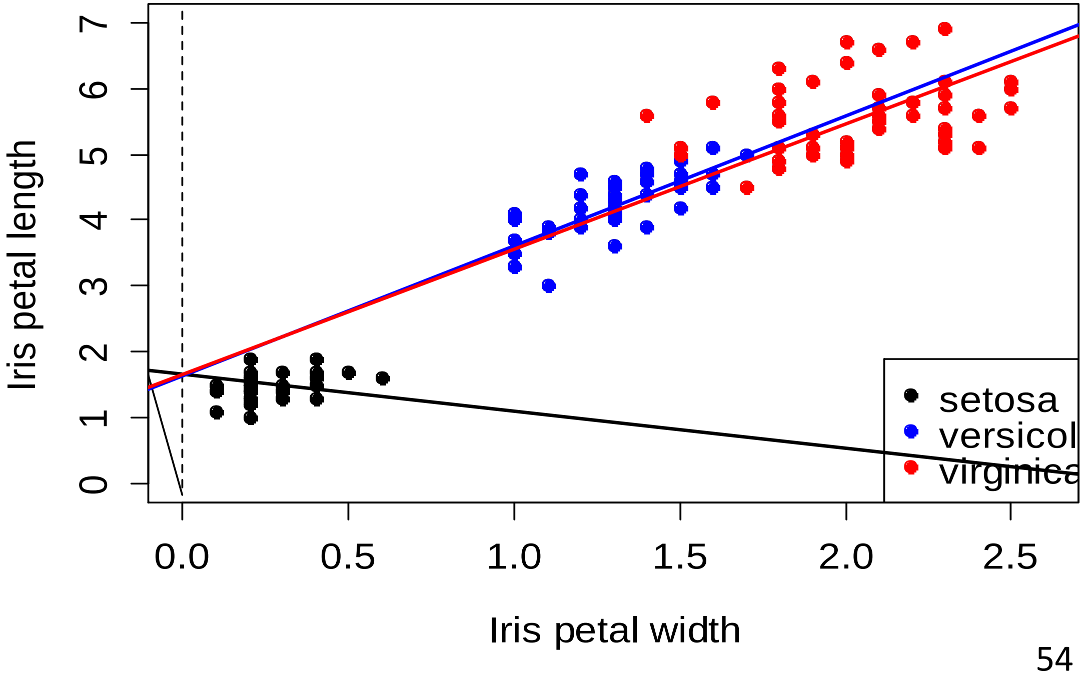
```
full_model=lm(Petal.Length ~ Petal.Width*Species, data=iris)

same_slope=lm(Petal.Length ~ Petal.Width+Species, data=iris)

same_intercept=lm(Petal.Length ~ Petal.Width:Species, data=iris)
```

and we can perform model comparison by an ANOVA on the residuals

Comparing full model and model with same intercept

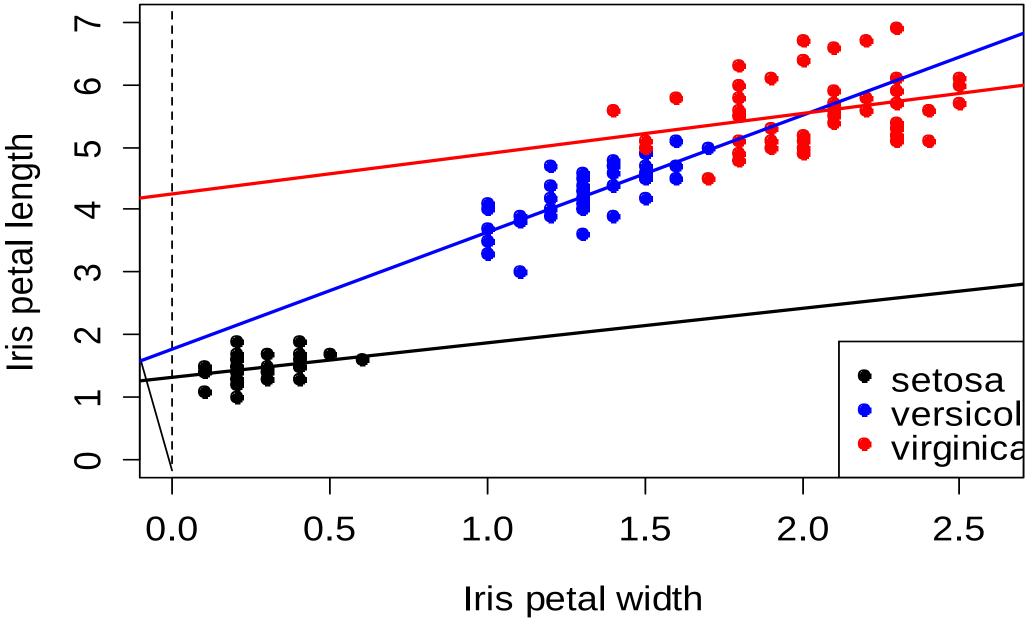
```
anova(same_intercept, full_model)
Analysis of Variance Table

Model 1: Petal.Length ~ Petal.Width:Species
Model 2: Petal.Length ~ Petal.Width * Species
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    146 25.563
2    144 18.816  2    6.7474 25.82 2.614e-10 ***
```

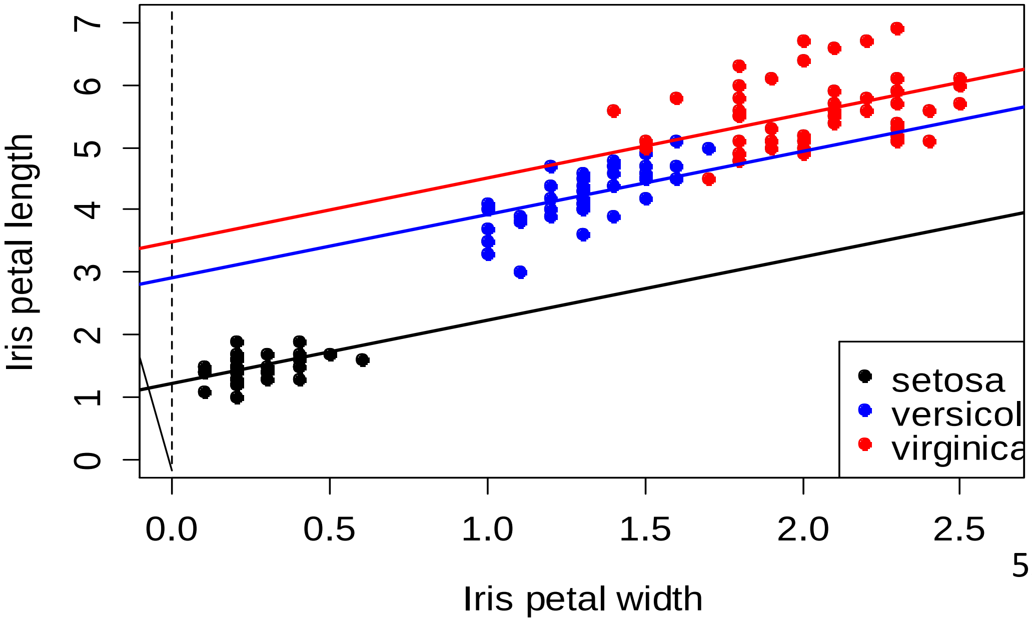




Comparing full model and model with same slope

```
anova(same_slope, full_model)
Analysis of Variance Table

Model 1: Petal.Length ~ Petal.Width + Species
Model 2: Petal.Length ~ Petal.Width * Species
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    146 20.833
2    144 18.816  2    2.0178 7.7213 0.0006525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Iris petal length
Iris petal width
setosa
versicol
virginica


Iris petal length
Iris petal width
setosa
versicol
virginica