Linear Models, ANOVA, and ANCOVA

Emmanuel Paradis

Institut Pertanian Bogor

November 5, 2012

Three typical examples of biological data sets:

1. Measures of yield of peas on 24 plots with application of nitrogen (N), phosphorus (P), and/or potassium (K). The plots were distributed on 6 blocks of 4:

```
   block N P K yield
1      1 0 1 1  49.5
2      1 1 1 0  62.8
3      1 0 0 0  46.8
...
23     6 0 1 1  53.2
24     6 0 0 0  56.0
```

2. Morphometric measurements on 200 individual of the crab *Leptograpsus variegatus*. Five measures, sex and colour:

```
    colour sex   FL   RW   CL   CW   BD
1        B   M  8.1  6.7 16.1 19.0  7.0
2        B   M  8.8  7.7 18.1 20.8  7.4
3        B   M  9.2  7.8 19.0 22.4  7.7
...
```

3. Survival times of 33 patients with leukemia with respect to a treatment and white cell counts:

```
   While cells counts Treatment Surv. time
1               2300   present         65
2                750   present        156
3               4300   present        100
...
```

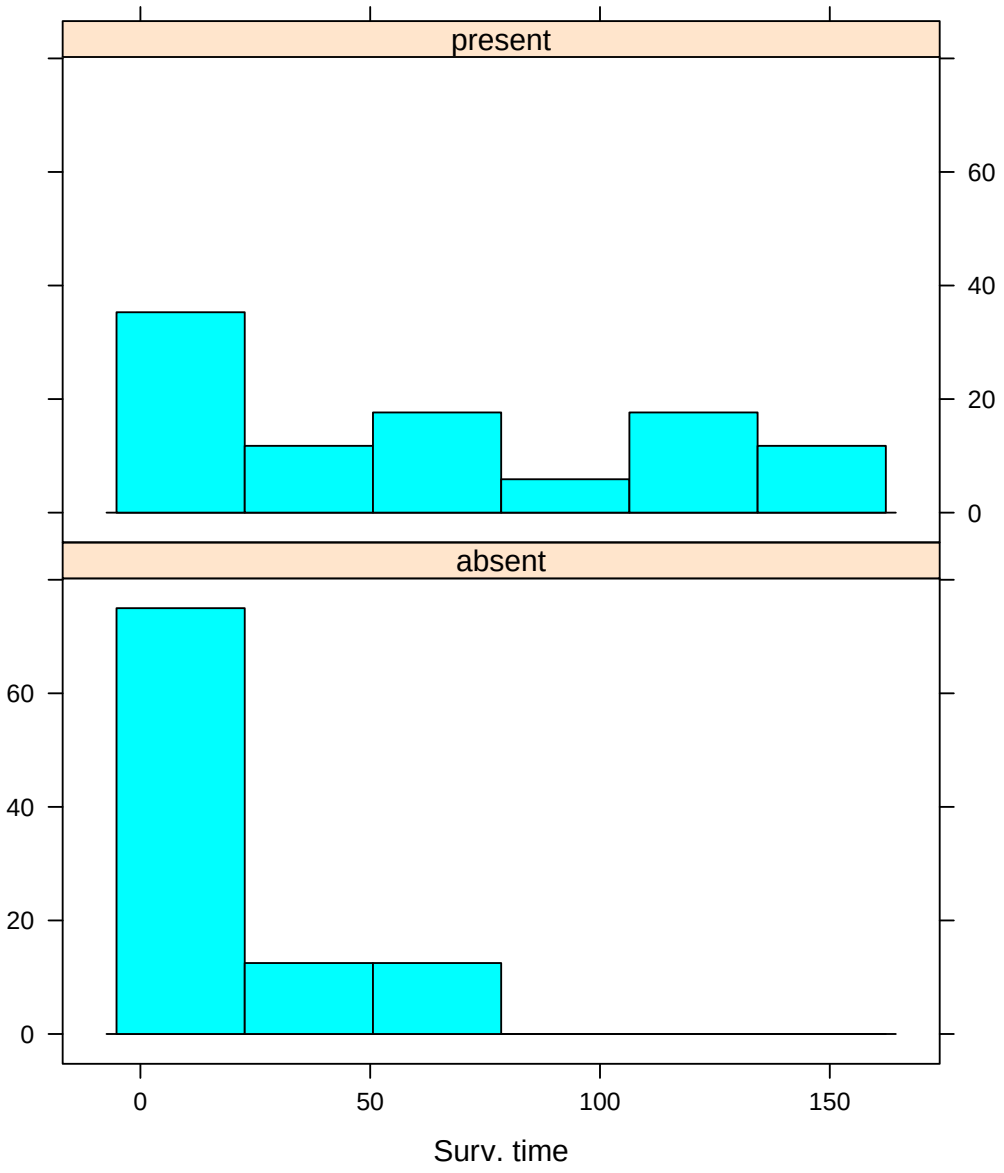
A very general question in biology is: explaining variation in a quantity with respect to one or several variables.

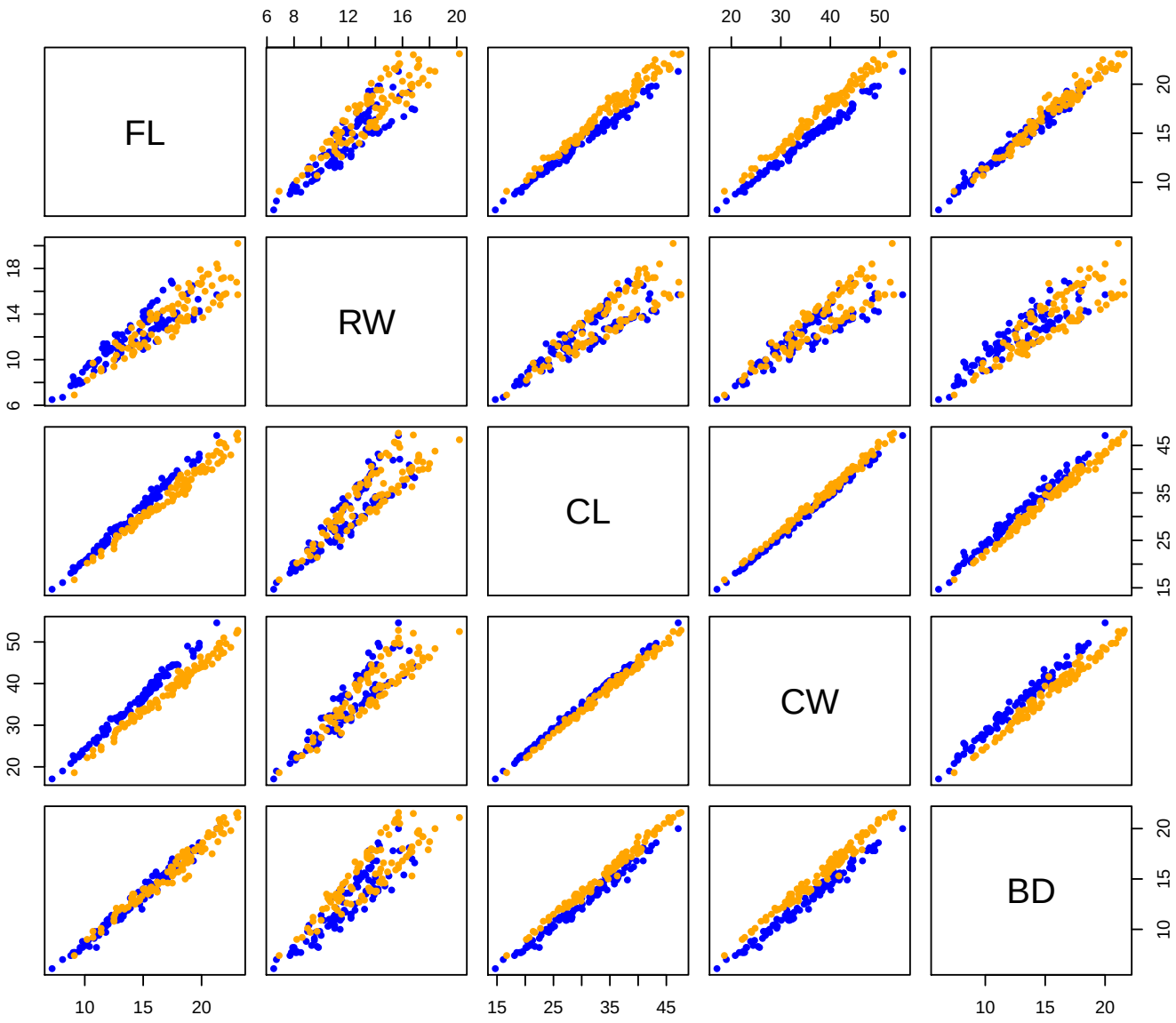
Suppose for a moment that the quantity we are studying is completely determined by one or two variables: then prediction is easy and testing hypothesis is simple. . .

A very general question in biology is: explaining variation in a quantity with respect to one or several variables.

Suppose for a moment that the quantity we are studying is completely determined by one or two variables: then prediction is easy and testing hypothesis is simple. . .

. . . but in reality, it's almost never the case.


present
absent
60
40
20
0
60
40
20
0
0
50
100
150
Surv. time


FL
RW
CL
CW
BD
6 8 12 16 20
20 30 40 50
10 15 20
6 10 14 18
15 25 35 45
20 30 40 50
10 15 20
10 15 20
15 25 35 45
10 15 20

To start, let us consider a very simple situation: we want to compare two samples and want to test the hypothesis that they originate from the same *statistical population*.

A statistical population is characterized by its *distribution*.

To start, let us consider a very simple situation: we want to compare two samples and want to test the hypothesis that they originate from the same *statistical population*.

A statistical population is characterized by its *distribution*.

Taking two samples from the same statistical population will lead to differences that are due to chance.

To start, let us consider a very simple situation: we want to compare two samples and want to test the hypothesis that they originate from the same *statistical population*.

A statistical population is characterized by its *distribution*.

Taking two samples from the same statistical population will lead to differences that are due to chance.

William Gosset (1876–1937), better known as “Student”, invented a test to compare the means of two samples: the t-test. The null hypothesis (H_0) is that both samples come from the same population.

To start, let us consider a very simple situation: we want to compare two samples and want to test the hypothesis that they originate from the same *statistical population*.

A statistical population is characterized by its *distribution*.

Taking two samples from the same statistical population will lead to differences that are due to chance.

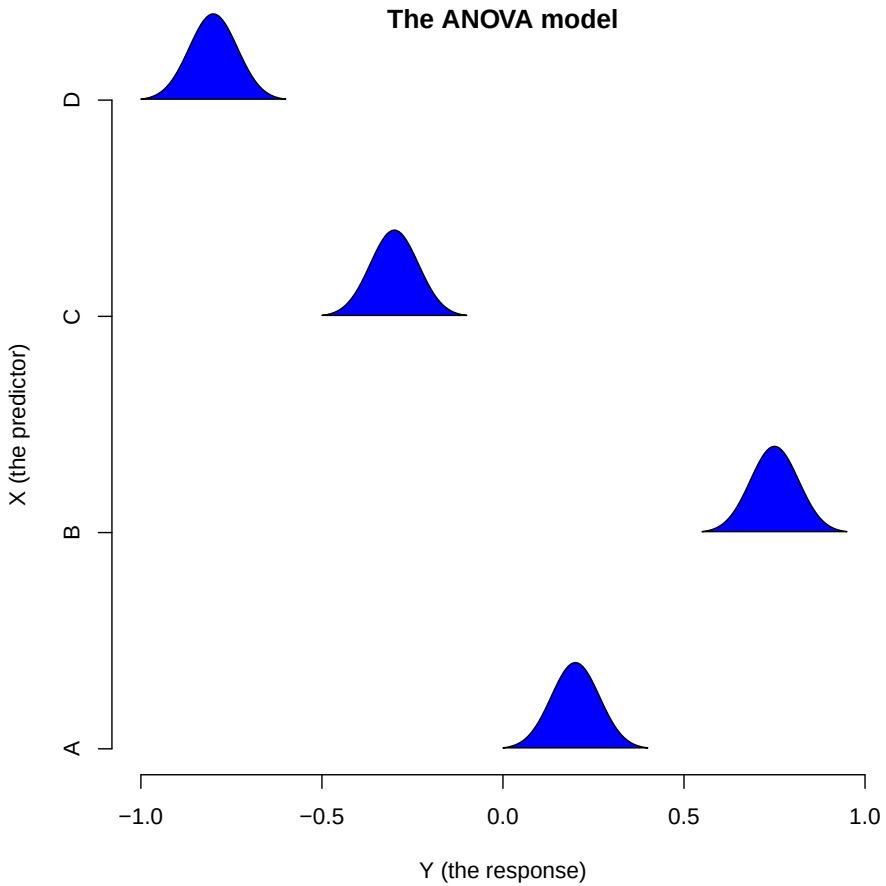
William Gosset (1876–1937), better known as “Student”, invented a test to compare the means of two samples: the t-test. The null hypothesis (H_0) is that both samples come from the same population.

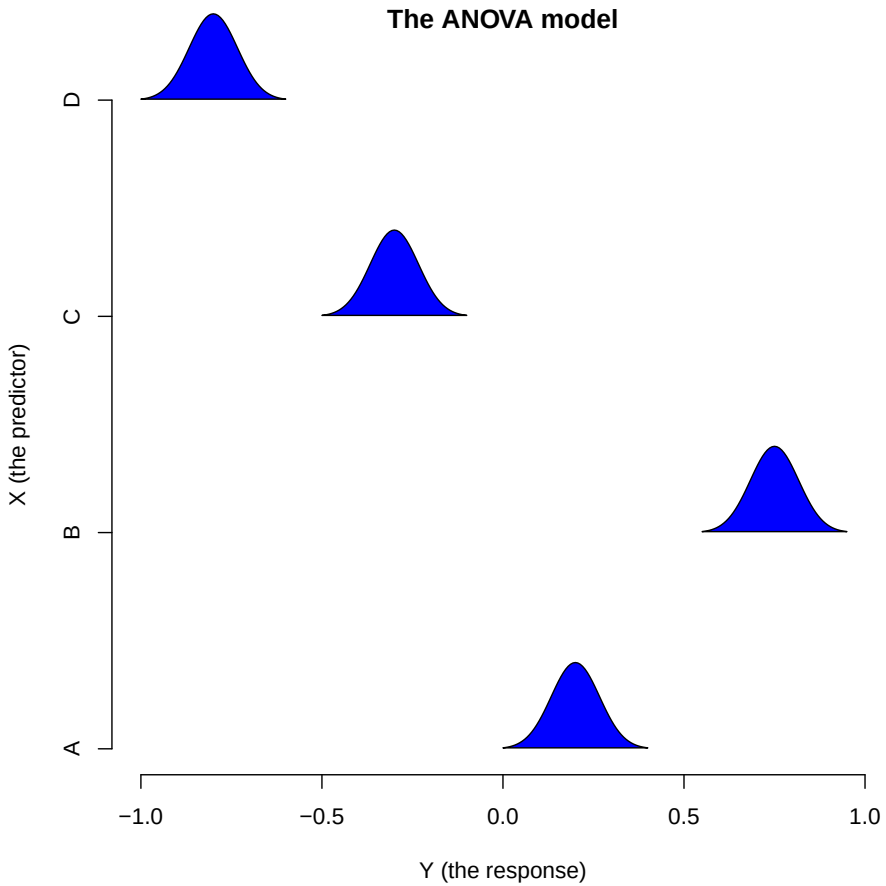
What if there are more than two samples? This is the analysis of variance invented by R. A. Fisher (1890–1962).

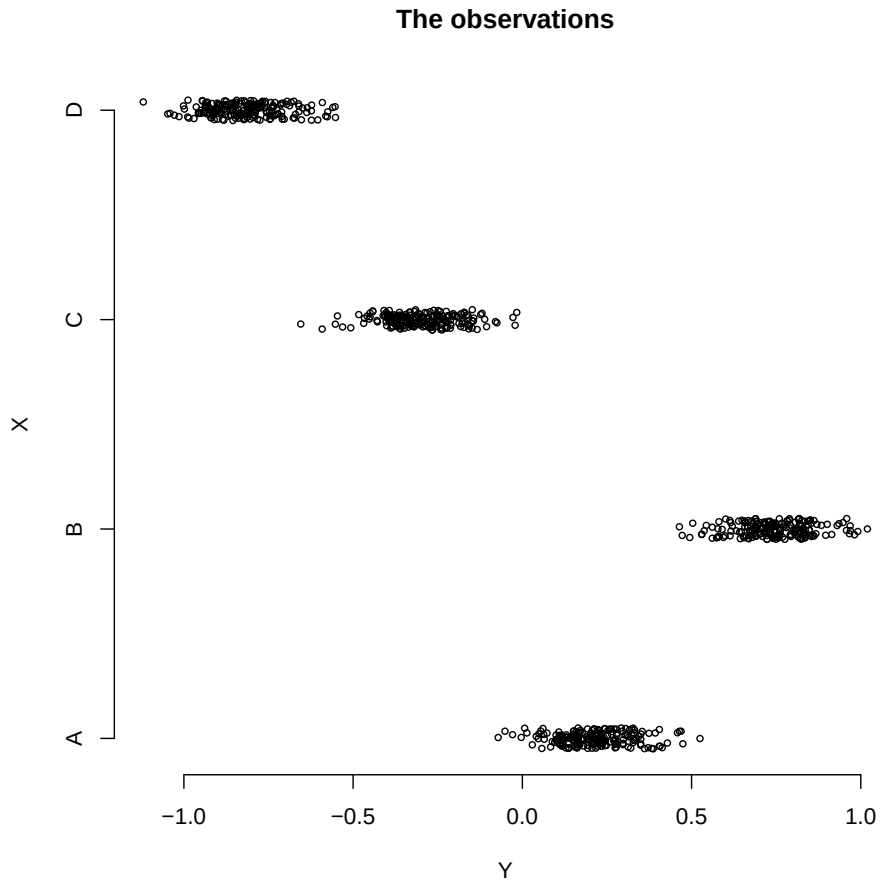
Consider a case with four samples: the ANOVA assumes that each sample follows a normal distribution with means μ_1, μ_2, μ_3, and μ_4.

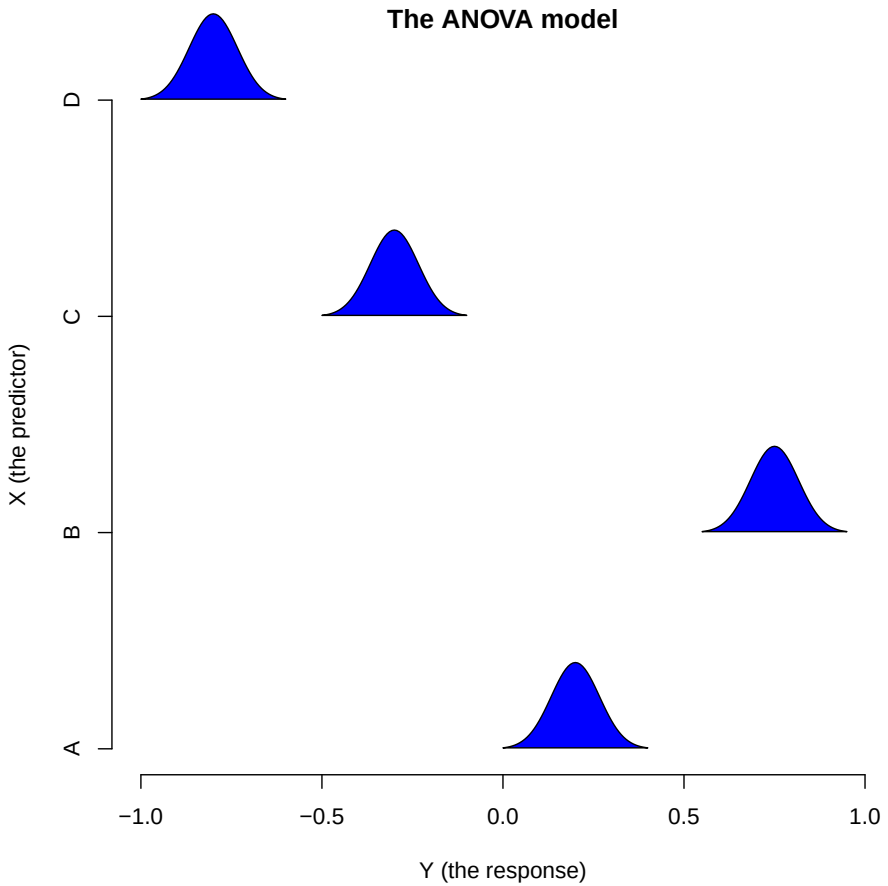
The observations: $x_{1i} \sim \mathcal{N}(\mu_1, \sigma^2)$, $x_{2i} \sim \mathcal{N}(\mu_2, \sigma^2)$, etc.

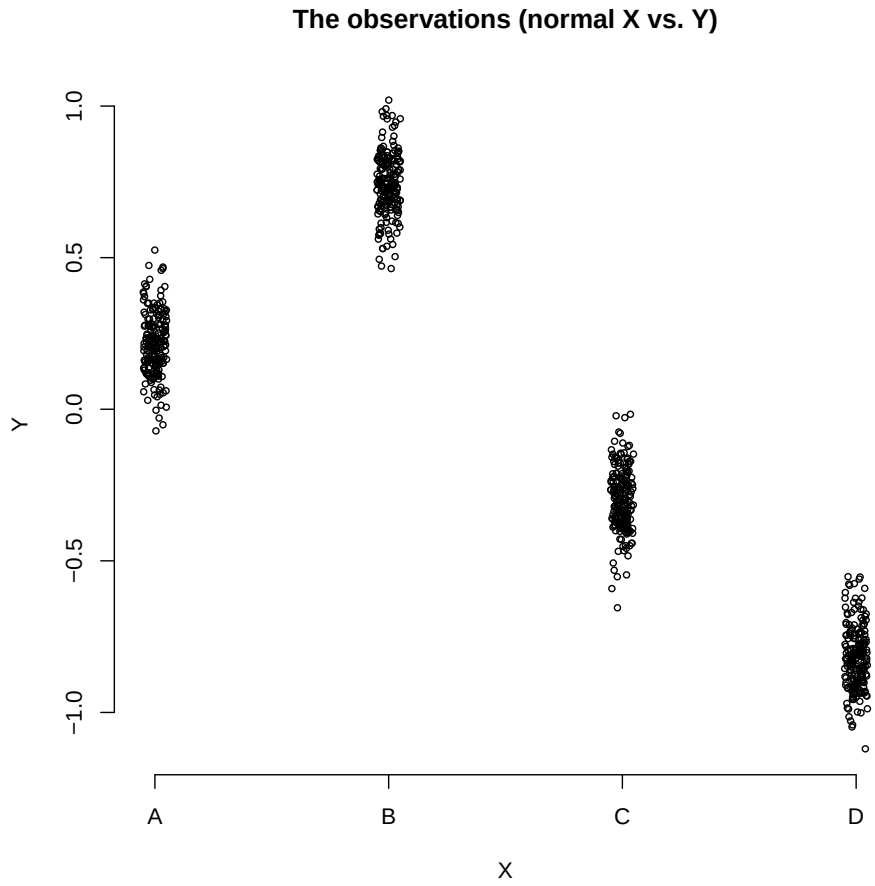
H_0: all four means are equal


The ANOVA model
D
C
B
A
X (the predictor)
−1.0
−0.5
0.0
0.5
1.0
Y (the response)


The ANOVA model
X (the predictor)
D
C
B
A
−1.0
−0.5
0.0
0.5
1.0
Y (the response)


The observations
X
D
C
B
A
−1.0
−0.5
0.0
0.5
1.0
Y


The ANOVA model
X (the predictor)
D
C
B
A
−1.0
−0.5
0.0
0.5
1.0
Y (the response)


The observations (normal X vs. Y)
Y
1.0
0.5
0.0
−0.5
−1.0
A
B
C
D
X

Similar assumptions are made in a linear regression:

$$y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Similar assumptions are made in a linear regression:

$$y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

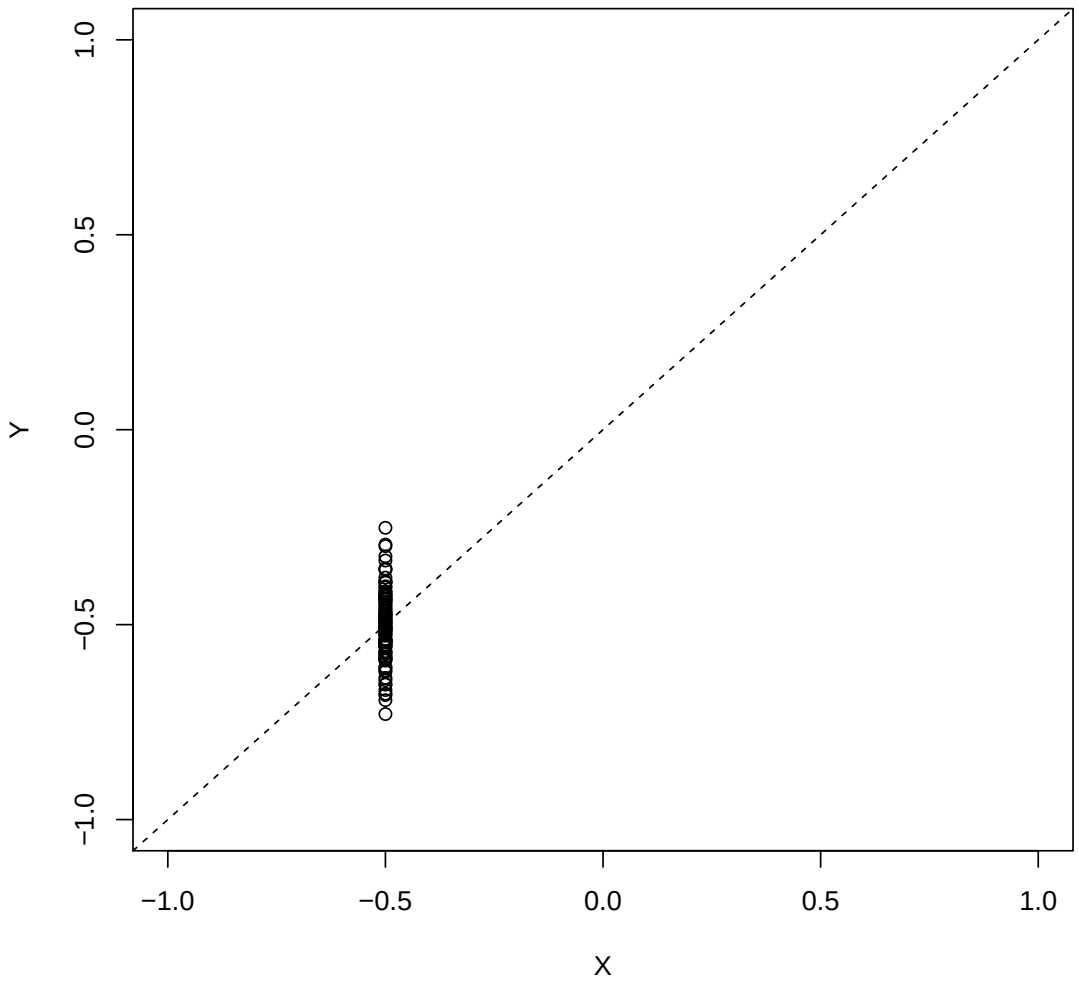
It means that for a given value of x: $y_i \sim \mathcal{N}(\bar{y}, \sigma^2)$ with $\bar{y} = \beta x + \alpha$.

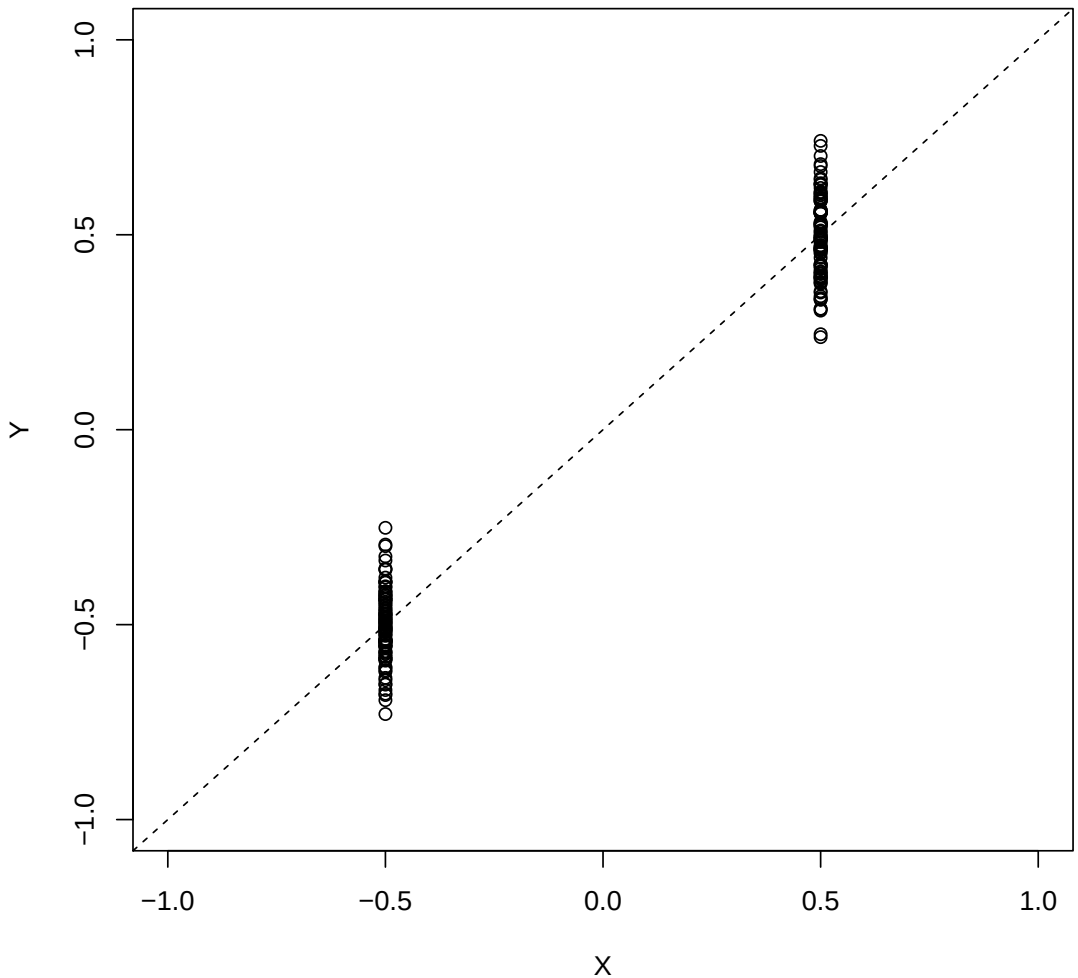
Similar assumptions are made in a linear regression:

$$y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

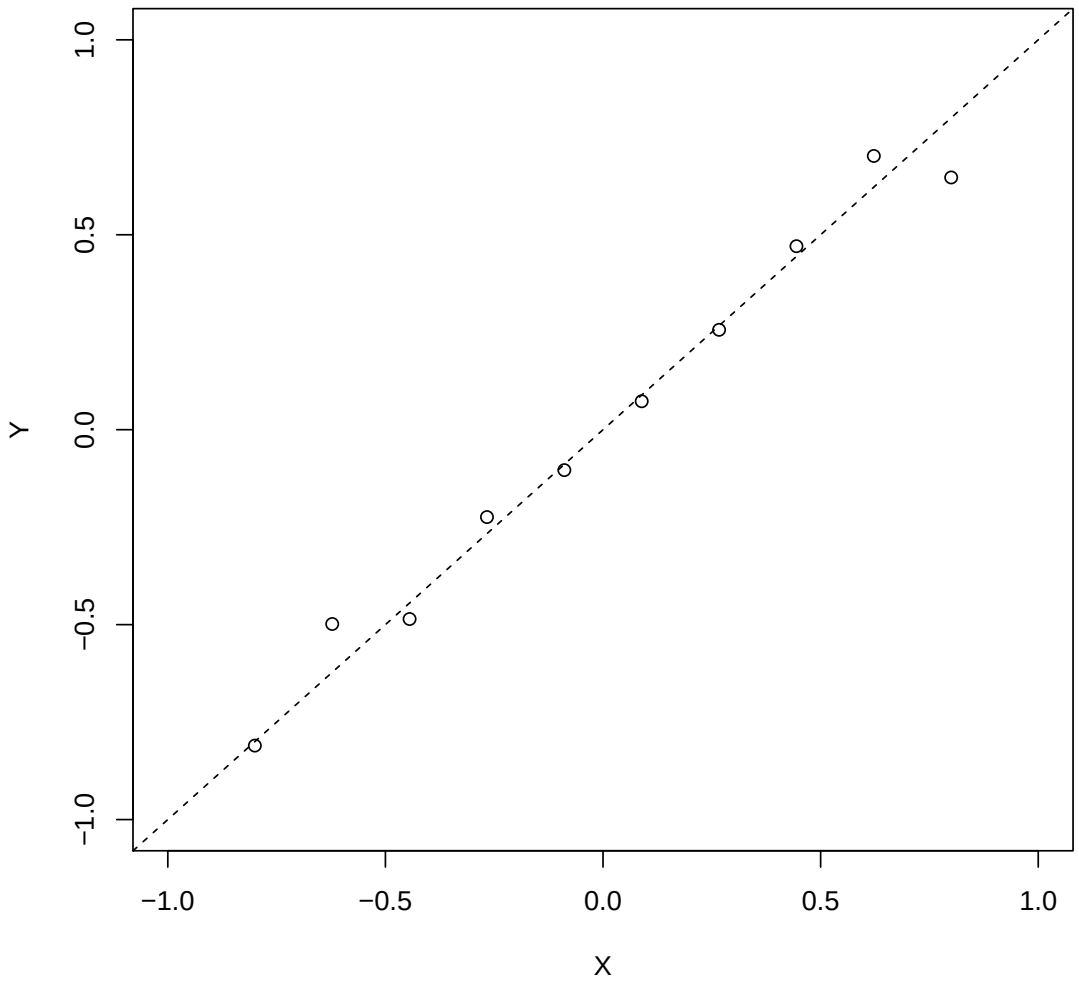
It means that for a given value of x: $y_i \sim \mathcal{N}(\bar{y}, \sigma^2)$ with $\bar{y} = \beta x + \alpha$.

Suppose we do many observations of y for this value of $x = -0.5$:


1.0
0.5
0.0
−0.5
−1.0
Y
−1.0
−0.5
0.0
0.5
1.0
X



Suppose we do now many observations of y for $x = 0.5$.



In reality, we often don't have so many points, but the assumptions of the linear still hold.

The ANOVA and the linear regression models are the same, the difference is that the predictor x is discrete or continuous.

How to use a categorical (discrete) variable in this formulation?

The ANOVA and the linear regression models are the same, the difference is that the predictor x is discrete or continuous.

How to use a categorical (discrete) variable in this formulation? The answer is provided by the *contrasts*: a coding of categories into numerical variable(s).

Consider a variable with two categories: colour (blue/red). This variable is replaced by a numeric variable taking the values 0 (blue) or 1 (red).

Blue $\rightarrow z = 0$
Red $\rightarrow z = 1$

The ANOVA and the linear regression models are the same, the difference is that the predictor x is discrete or continuous.

How to use a categorical (discrete) variable in this formulation? The answer is provided by the *contrasts*: a coding of categories into numerical variable(s).

Consider a variable with two categories: colour (blue/red). This variable is replaced by a numeric variable taking the values 0 (blue) or 1 (red).

Blue $\rightarrow z = 0$
Red $\rightarrow z = 1$

We then fit the linear model $y = \beta z + \alpha$ which takes two forms:

Blue $\rightarrow y = \alpha$ Red $\rightarrow y = \beta + \alpha$

Suppose there are three categories (blue, red, green): we substitute by two numeric variables taking the values 0 or 1:

$$y = \beta_1 z_1 + \beta_2 z_2 + \alpha$$

Suppose there are three categories (blue, red, green): we substitute by two numeric variables taking the values 0 or 1:

$$y = \beta_1 z_1 + \beta_2 z_2 + \alpha$$

	z_1	z_2
Blue	0	0
Red	1	0
Green	0	1

Suppose there are three categories (blue, red, green): we substitute by two numeric variables taking the values 0 or 1:

$$y = \beta_1 z_1 + \beta_2 z_2 + \alpha$$

	z_1	z_2	
Blue	0	0	$y = \alpha$
Red	1	0	$y = \beta_1 + \alpha$
Green	0	1	$y = \beta_2 + \alpha$

Suppose there are three categories (blue, red, green): we substitute by two numeric variables taking the values 0 or 1:

$$y = \beta_1 z_1 + \beta_2 z_2 + \alpha$$

	z_1	z_2	
Blue	0	0	$y = \alpha$
Red	1	0	$y = \beta_1 + \alpha$
Green	0	1	$y = \beta_2 + \alpha$

For a variable with n categories, $n - 1$ variables $0/1$ are made.

There are two advantages in this approach:

- No need to consider special cases seperately (unbalanced samples, etc.) which require special formulae when doing sums of squares (SS) decomposition.

There are two advantages in this approach:

- No need to consider special cases seperately (unbalanced samples, etc.) which require special formulae when doing sums of squares (SS) decomposition.

- This makes a synthesis of several methods that were traditionaly seen as distinct: simple and multiple regressions, ANOVA and ANCOVA in all its designs (one- or multiple-factor, hierarchical, etc.)

$y = \beta x + \alpha$	Linear regression
$y = \beta z + \alpha$	Analysis of variance (ANOVA)
$y = \beta_1 x + \beta_2 z + \alpha$	Analysis of covariance (ANCOVA)

In all cases, the model is fitted by minimizing the sums of squares around the mean predicted by the model: $\sum_i (y_i - \bar{y}_i)^2$.

Interactions Among Variables

Two cases: continous $\times$ categorical, categorical $\times$ categorical

1. Continous $\times$ categorical.

To code this interaction, a new variable is made with the product of x and z:

$$y = \beta_1 x + \beta_2 z + \beta_3 (xz) + \alpha$$

Interactions Among Variables

Two cases: continous $\times$ categorical, categorical $\times$ categorical

1. Continous $\times$ categorical.

To code this interaction, a new variable is made with the product of x and z:

$$y = \beta_1 x + \beta_2 z + \beta_3 (xz) + \alpha$$

Blue: $y = \beta_1 x + \alpha$
Red: $y = \beta_1 x + \beta_2 + \beta_3 x + \alpha$

Interactions Among Variables

Two cases: continous $\times$ categorical, categorical $\times$ categorical

1. Continous $\times$ categorical.

To code this interaction, a new variable is made with the product of x and z:

$$y = \beta_1 x + \beta_2 z + \beta_3 (xz) + \alpha$$

Blue: $y = \beta_1 x + \alpha$
Red: $y = \beta_1 x + \beta_2 + \beta_3 x + \alpha = (\beta_1 + \beta_3)x + \beta_2 + \alpha$

If no interaction ($\beta_3 = 0$), the slope is the same for both categories.

Interactions Among Variables

Two cases: continous $\times$ categorical, categorical $\times$ categorical

1. Continous $\times$ categorical.

To code this interaction, a new variable is made with the product of x and z:

$$y = \beta_1 x + \beta_2 z + \beta_3 (xz) + \alpha$$

Blue: $y = \beta_1 x + \alpha$
Red: $y = \beta_1 x + \beta_2 + \beta_3 x + \alpha = (\beta_1 + \beta_3)x + \beta_2 + \alpha$

If no interaction ($\beta_3 = 0$), the slope is the same for both categories.

For n categories, $n - 1$ new variables will be made to code the interaction.

2. Categorical $\times$ categorical

New variables are made with the products of all the possible combinations 2 by 2 among the numeric codings of the two variables.

Male $\rightarrow z_1 = 0$
Female $\rightarrow z_1 = 1$

Blue $\rightarrow z_2 = 0$
Red $\rightarrow z_2 = 1$

$$y = \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 z_2) + \alpha$$

2. Categorical $\times$ categorical

New variables are made with the products of all the possible combinations 2 by 2 among the numeric codings of the two variables.

Male $\rightarrow z_1 = 0$
Female $\rightarrow z_1 = 1$

Blue $\rightarrow z_2 = 0$
Red $\rightarrow z_2 = 1$

$$y = \beta_1 z_1 + \beta_2 z_2 + \beta_3 (z_1 z_2) + \alpha$$

Male	Blue	$y = \alpha$
	Red	$y = \beta_2 + \alpha$
Female	Blue	$y = \beta_1 + \alpha$
	Red	$y = \beta_1 + \beta_2 + \beta_3 + \alpha$

If no interaction ($\beta_3 = 0$):

Male	Blue	$y = \alpha$
	Red	$y = \beta_2 + \alpha$
Female	Blue	$y = \beta_1 + \alpha$
	Red	$y = \beta_1 + \beta_2 + \alpha$

The “contrast” between ‘Blue’ and ‘Red’ is the same for ‘Male’ ‘Female’ (and vice-versa).

If no interaction ($\beta_3 = 0$):

Male	Blue	$y = \alpha$
	Red	$y = \beta_2 + \alpha$
Female	Blue	$y = \beta_1 + \alpha$
	Red	$y = \beta_1 + \beta_2 + \alpha$

The “contrast” between ‘Blue’ and ‘Red’ is the same for ‘Male’ ‘Female’ (and vice-versa).

For the case of two variables with respectively n_1 and n_2 categories $(n_1 - 1)(n_2 - 1)$ new variables $0/1$ will be made.

If no interaction ($\beta_3 = 0$):

Male	Blue	$y = \alpha$
	Red	$y = \beta_2 + \alpha$
Female	Blue	$y = \beta_1 + \alpha$
	Red	$y = \beta_1 + \beta_2 + \alpha$

The “contrast” between ‘Blue’ and ‘Red’ is the same for ‘Male’ ‘Female’ (and vice-versa).

For the case of two variables with respectively n_1 and n_2 categories $(n_1 - 1)(n_2 - 1)$ new variables $0/1$ will be made.

For interactions of higher orders (between three variables or more) the combinations 3 by 3, 4 by 4, and so on, are used.

If no interaction ($\beta_3 = 0$):

Male	Blue	$y = \alpha$
	Red	$y = \beta_2 + \alpha$
Female	Blue	$y = \beta_1 + \alpha$
	Red	$y = \beta_1 + \beta_2 + \alpha$

The “contrast” between ‘Blue’ and ‘Red’ is the same for ‘Male’ ‘Female’ (and vice-versa).

For the case of two variables with respectively n_1 and n_2 categories $(n_1 - 1)(n_2 - 1)$ new variables $0/1$ will be made.

For interactions of higher orders (between three variables or more) the combinations 3 by 3, 4 by 4, and so on, are used.

⚠ Interactions require a lot of data to be detected and estimated correcly.

Linear Models With R

The model is specified with a *formula*:

`y ~ x1 + x2`	additive effects
`y ~ x1 * x2`	additive effects and interaction
	identical to `y ~ x1 + x2 + x1:x2`

Linear Models With R

The model is specified with a *formula*:

`y ~ x1 + x2`	additive effects
`y ~ x1 * x2`	additive effects and interaction
	identical to `y ~ x1 + x2 + x1:x2`

The model is fitted with the function `lm` (or sometimes `aov`), e.g.:

```
lm(y ~ x1 + x2)
summary(lm(y ~ x1 + x2))
summary(lm(y ~ x1 + x2, data = DF))
```

Tests of Effects

What is the difference between *effect* and *parameter*?

- For a continuous predictor, there is one parameter (aka coefficient).
- For a categorical predictor with n categories, there are $n - 1$ parameters. When testing the statistical effect of such a predictor, we test for the significance of all parameters linked to this predictor. This is done with the function `anova`

Tests of Effects

What is the difference between *effect* and *parameter*?

- For a continuous predictor, there is one parameter (aka coefficient).
- For a categorical predictor with n categories, there are $n - 1$ parameters. When testing the statistical effect of such a predictor, we test for the significance of all parameters linked to this predictor. This is done with the function `anova`

```
res.lm  <-  lm(....
res.aov <- aov(....
```

1. `summary(res.aov)`: ANOVA table (= tests of the effets $\sim F$)

   `summary(res.lm)`: tests of the parameters ($\sim t$)

1. `summary(res.aov)`: ANOVA table (= tests of the effets $\sim F$)

   `summary(res.lm)`: tests of the parameters ($\sim t$)

2. `anova`

   ANOVA table by including the effects in the order of the formula (type I ANOVA).

   (a) `anova(res.lm)` and `summary(res.aov)` are identical.

   (b) The order of the variables in the formula is important if there are several categorical predictors and the design is *unbalanced* (can be checked with `table`).

3. `drop1`: tests each effect individually *vs.* the full model (type II ANOVA).

3. `drop1`: tests each effect individually *vs.* the full model (type II ANOVA).

4. `add1` tests one or several additional effects.

   Ex.: if the initial model does not include interactions: `add1(res, ~.^2)` tests the addition of each interaction.

3. `drop1`: tests each effect individually *vs.* the full model (type II ANOVA).

4. `add1` tests one or several additional effects.

   Ex.: if the initial model does not include interactions: `add1(res, ~.^2)` tests the addition of each interaction.

5. `predict` calculates the values predicted by the model.

⚠ To get help on these functions: `?summary.lm`, `?anova.lm`, `?add1.lm`, ?drop1.lm, ?predict.lm.

Models can be compared only if they are fitted to the same vector of responses:

- `y ~ x` and `log(y) ~ x` *cannot* be compared!
- `y ~ x` and `y ~ x + z` will not be fitted to the same data if `z` has missing data (`NA`) and not `x`.

An Application

```
> library(MASS)
> data(leuk)
> names(leuk)
[1] "wbc"  "ag"   "time"
```

It is *always* crucial to do graphical exploratory analyses before fitting the models. Some examples of graphics here could be:

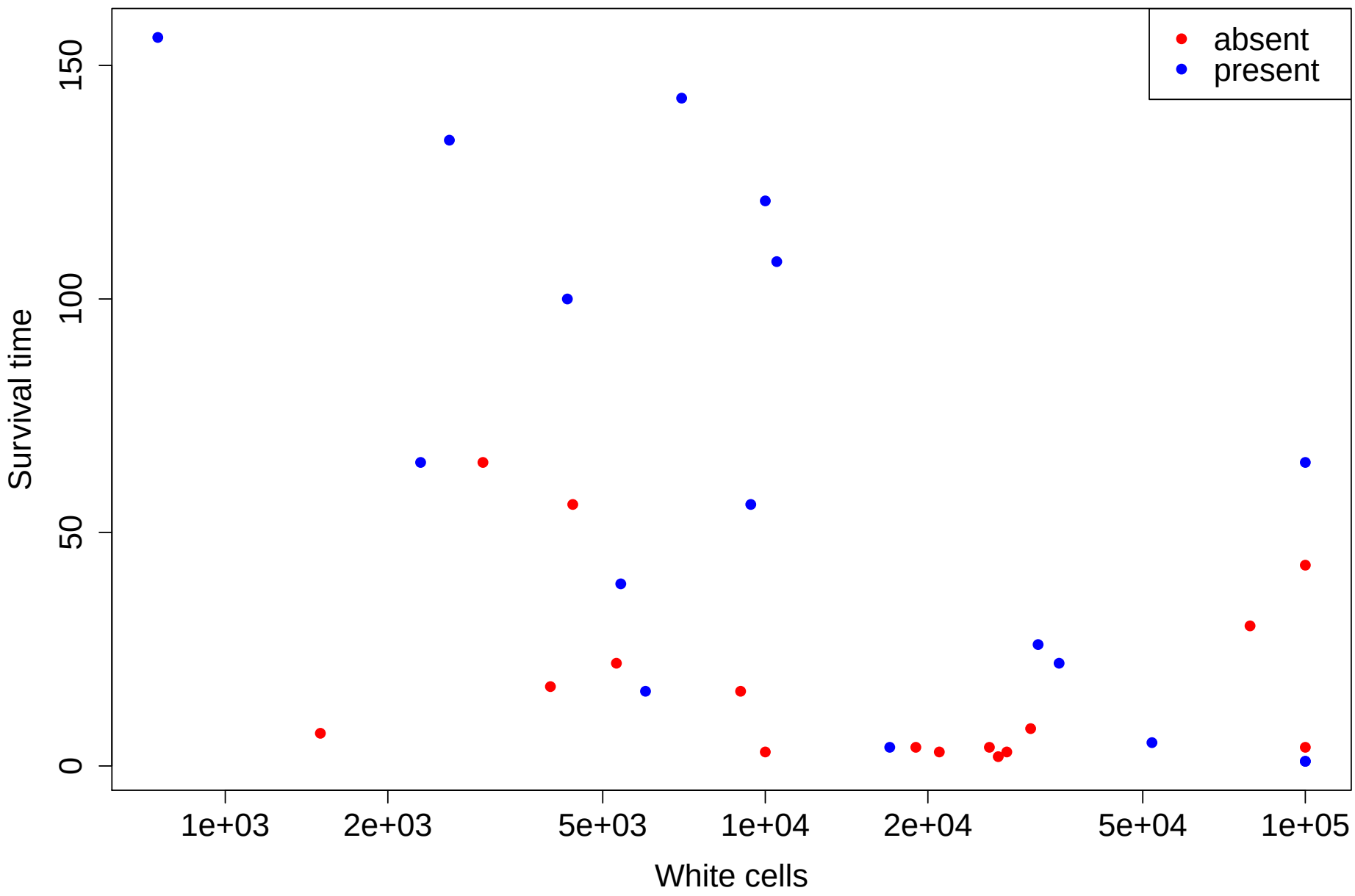
```
> plot(leuk$wbc, leuk$time)
> plot(leuk$wbc, leuk$time, log = "x")
> plot(leuk$ag, leuk$time)
```

```
> mod.leuk <- lm(time ~ log(wbc) * ag, data = leuk)
```

```
> summary(mod.leuk)

Call:
lm(formula = time ~ ag * log(wbc), data = leuk)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.400 -13.776  -7.617  20.805  65.588 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          55.065     64.171   0.858  0.39787
agpresent           251.391     83.887   2.997  0.00554
log(wbc)             -3.859      6.615  -0.583  0.56419
agpresent:log(wbc)  -22.011      8.711  -2.527  0.01722
```

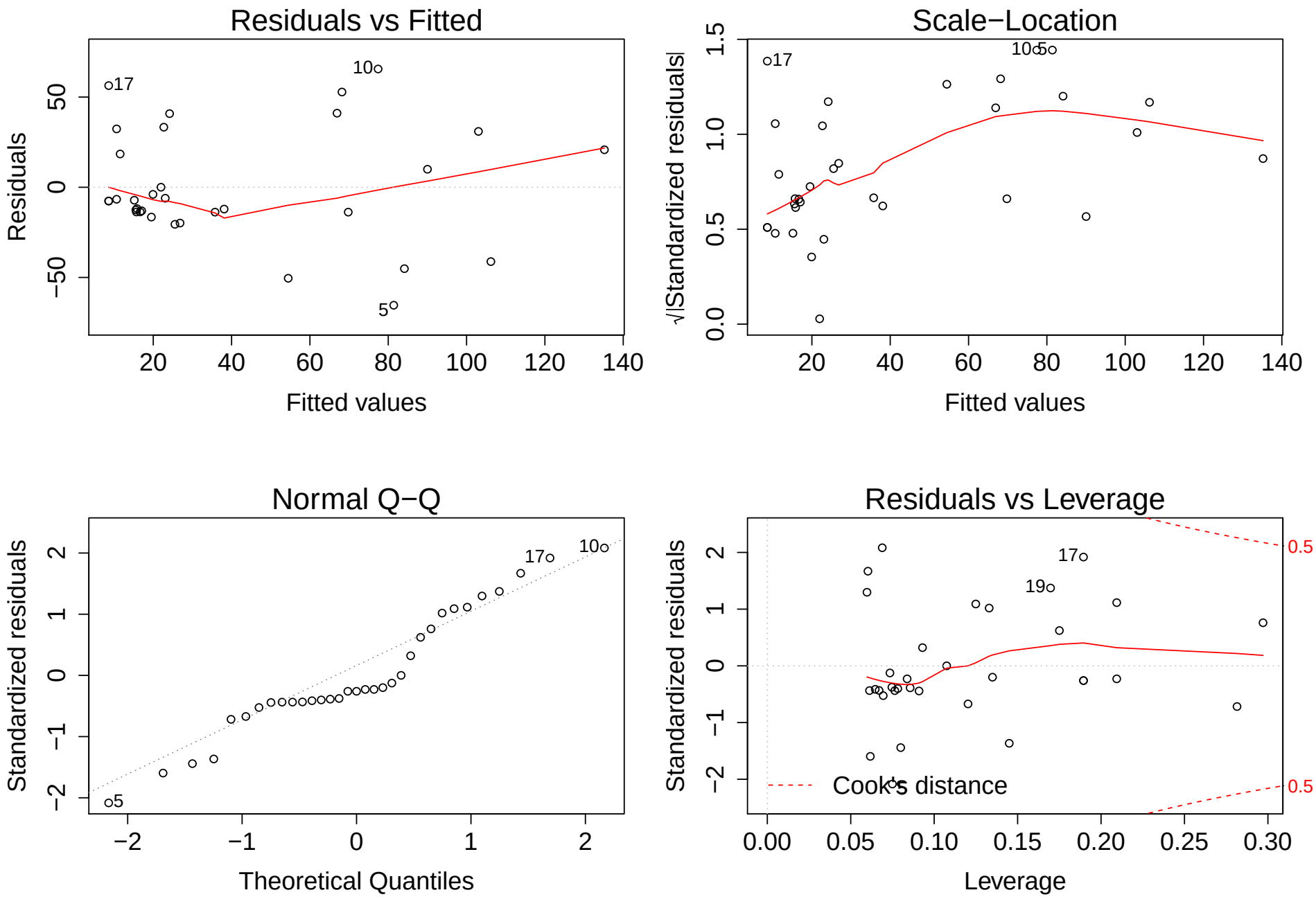
```
Residual standard error: 32.64 on 29 degrees of freedom
Multiple R-squared: 0.5574,Adjusted R-squared: 0.5116
F-statistic: 12.18 on 3 and 29 DF,  p-value: 2.482e-05

> anova(mod.leuk)
Analysis of Variance Table

Response: time
            Df  Sum Sq Mean Sq F value    Pr(>F)
ag           1 16346.3 16346.3 15.3459 0.0005004
log(wbc)     1 15758.6 15758.6 14.7942 0.0006062
ag:log(wbc)  1  6801.9  6801.9  6.3856 0.0172151
Residuals   29 30890.6  1065.2
```

```
> plot(leuk$wbc, leuk$time, log = "x", col = c("red",
      "blue")[leuk$ag], pch = 19, xlab = "White cells",
      ylab = "Survival time")
> legend("topright", legend = levels(leuk$ag),
    col = c("red", "blue"), pch = 19)
```


absent
present
Survival time
0
50
100
150
1e+03
2e+03
5e+03
1e+04
2e+04
5e+04
1e+05
White cells

Regression diagnostics

```
> par(mfcol = c(2, 2))
> plot(mod.leuk)
```


Residuals vs Fitted
Residuals
Fitted values
17
10
5
Scale−Location
√|Standardized residuals|
Fitted values
17
10
5
Normal Q−Q
Standardized residuals
Theoretical Quantiles
17
10
5
Residuals vs Leverage
Standardized residuals
Leverage
17
19
5
Cook's distance
0.5
0.5

1. Values predicted by the model $\widehat{y}_i$ (as x) and residuals r_i (as y).

1. Values predicted by the model $\widehat{y}_i$ (as x) and residuals r_i (as y).

2. Predicted values (as x) and square root of standardized residuals, for the ith observation:

$$e_i = r_i \Big/ \left(\widehat{\sigma}\sqrt{1 - h_{ii}}\right)$$

   because the r_i's are not independant and of the same variance (h_{ii}: variance of r_i).

1. Values predicted by the model $\widehat{y}_i$ (as x) and residuals r_i (as y).

2. Predicted values (as x) and square root of standardized residuals, for the ith observation:

$$e_i = r_i \Big/ \left(\widehat{\sigma}\sqrt{1 - h_{ii}}\right)$$

   because the r_i's are not independant and of the same variance (h_{ii}: variance of r_i).

3. Since $e_i \sim \mathcal{N}(0, 1)$, the plot of the values predicted by the the normal distribution of the observed one must be on the line $x = y$.

1. Values predicted by the model $\widehat{y}_i$ (as x) and residuals r_i (as y).

2. Predicted values (as x) and square root of standardized residuals, for the ith observation:

$$e_i = r_i \Big/ \left(\widehat{\sigma}\sqrt{1 - h_{ii}}\right)$$

   because the r_i's are not independant and of the same variance (h_{ii}: variance of r_i).

3. Since $e_i \sim \mathcal{N}(0, 1)$, the plot of the values predicted by the the normal distribution of the observed one must be on the line $x = y$.

4. *leverage* = h_{ii}, measures the influence (leverage effect) of each observation on the regression.

Two important points about linear models:

- The assumptions of normality is on the residuals, *not* on the variables.
  ⚠ Do not test the normality of the data before doing a regression.

Two important points about linear models:

- The assumptions of normality is on the residuals, *not* on the variables.
  ⚠ Do not test the normality of the data before doing a regression.
- The (possible) log-transformation of the variables is not to normalize them, but to handle non-linear relationships.

Two important points about linear models:

- The assumptions of normality is on the residuals, *not* on the variables.
  ⚠ Do not test the normality of the data before doing a regression.
- The (possible) log-transformation of the variables is not to normalize them, but to handle non-linear relationships.

Terima kasih