Analysis of variance
Correlation
Linear regression

Quentin Giai Gianetto\textsuperscript{1,2}, Emeline Perthame\textsuperscript{1}

\textsuperscript{1} Bioinformatics and Biostatistics HUB – C3BI
\textsuperscript{2} Proteomic Platform – Mass Spectrometry for Biology Unit
Analysis of variance
Reminder about the last course

Experiment n° 1

- Several plates with purple treatment, measure of the viral load in each plate

- Several plates with yellow treatment, measure of the viral load in each plate

- I am interested in studying the difference of viral load between the 2 treatments

- Statistical test: comparison of 2 means

- Tools:
  - If « N » is large: Student’s t-test (parametric)
  - If « N » is small: Mann-Whitney-Wilcoxon test (non parametric)
Analysis of variance (ANOVA)

Experiment n° 2

• Several plates with purple treatment, measure of the viral load in each plate

• Several plates with yellow treatment, measure of the viral load in each plate

• Several plates with black treatment, measure of the viral load in each plate

• I am interested in studying the difference of viral load between the 3 treatments

• Statistical test: comparison of 3 means (or more!) simultaneously

• Tools:
  o If « N » is large: ANOVA (parametric)
  o If « N » is small: Kruskall-Wallis test (non parametric)
Analysis of variance (ANOVA)

Different measures of the variability

• Intensity values can be compared to the mean of their own group (or condition)
• Intensity values can be compared to the global mean of all the data
• The mean in each group can be compared to the global mean.
Different measures of the variability

- **Within-group variance**: measures the dispersion of the observations in each group
- **Between-group variance**: measures the dispersion of the means between groups
- **Total variance** = Within-group variance + Between-group variance
Analysis of variance (ANOVA)

Within- and between- group variances are the same:

If all *means* of the groups are equal, the ratio of the *estimated* between-variance and the *estimated* within-variance follows a Fisher distribution.
Analysis of variance (ANOVA)

Hypothesis testing

$H_0$: « All means of the groups are equal »

(« $m_{c1} = m_{c2} = m_{c3}$ »)

$H_1$: « At least one of the mean is not equal to the others »

(« $m_{c1} \neq m_{c2}$ or $m_{c1} \neq m_{c3}$ or $m_{c2} \neq m_{c3}$ »)

Assumptions and limits of the test

• Normality and homoskedasticity: it is assumed, under the null hypothesis, that the samples are drawn from the same population and follow the same Gaussian distribution.

• Independence among samples: it is assumed that each analyzed sample is independent of other samples
Analysis of variance (ANOVA): non-parametric version

**Assumptions and limits of the test**

- **Normality and homoskedasticity:** it is assumed, under the null hypothesis, that the samples are drawn from the same population and follow the same Gaussian distribution.

- **Independence among samples:** it is assumed that each analyzed sample is independent of other samples.

If the sample size is small or if normality is not checked, use the **Kruskall-Wallis** test:

```r
kruskal.test()
``` function in R
**Hypothesis testing**

\[ H_0: \text{"All means of the groups are equal" } \left( \text{"} m_{c1} = m_{c2} = m_{c3} \text{"} \right) \]

\[ H_1: \text{"At least one of the means is not equal to the others" } \left( \text{"} m_{c1} \neq m_{c2} \text{ or } m_{c1} \neq m_{c3} \text{ or } m_{c2} \neq m_{c3} \text{"} \right) \]

Imagine the ANOVA test concludes all means are not equal.

**Question:** where are the differences?
Studying the « contrasts »

**Hypothesis testing**

- **$H_0$:** « All means of the groups are equal » (« $m_{c1} = m_{c2} = m_{c3}$ »)
- **$H_1$:** « At least one of the means is not equal to the others »
  (« $m_{c1} \neq m_{c2}$ or $m_{c1} \neq m_{c3}$ or $m_{c2} \neq m_{c3}$ »)

Imagine the ANOVA test concludes all means are not equal.

**Question :** where are the differences?

**Tukey's HSD (honest significant difference) test**

- **Better than using multiple t-tests:** Tukey’s HSD test adjusts the p-values for multiple testing, so that the family-wise error rate is controlled (probability to get at least one false positive among the family of tests performed).
Analysis of variance (parametric)

1. Import the `pigs` dataset in R using `read.table()` function. Be careful with the separator, it is "," so add the argument `sep=","` in your line of code.

2. Plot the distribution of the `weight` against the `diet` (use `boxplot(y ~ x, data=my_data)` and replace the fields by your values).

3. Using `lm` function, make an ANOVA which explains the `weight` with the `diet`. Interpret the result.

   ```r
   aov.out <- lm(weight ~ feed, data = pigs)
   anova(aov.out)
   ```

4. For testing constraints, install `lsmeans` package and use

   ```r
   lsmeans(aov.out, specs = pairwise ~ feed, adjust ="tukey")
   ```

which compares the mean of weights between the 4 diets. Interpret the result.
Analysis of variance (parametric)

1. Import the `flies` dataset in R using `read.table()` function.

2. Plot the distribution of the `density` against the `layer`.

3. Using `kruskal.test` function, make an ANOVA which explains the `density` with the `layer`. Interpret the result.

4. For testing constraints, use

   `pairwise.wilcox.test(flies$density, flies$layer)`

   which compares the mean of density between the 3 layers. Interpret the result.
Interactions in ANOVA

**Example**
I study the impact of 2 treatments at 2 time points on expression of a gene G.

There could be no interaction between the treatment and the time, meaning that the effect of treatment is the same at both time points.
Interactions in ANOVA

*But ….*
There could be an interaction between the treatment and the time, meaning that the effect of treatment depends on the time.

The treatment effect is *smaller* at time I than at time II.
Interactions in ANOVA

**But ....**
There could be an interaction between the treatment and the time, meaning that the effect of treatment depends on the time.

The treatment effect is **larger** at time I than at time II.
And worse ....
There could be an interaction between the treatment and the time, meaning that the effect of treatment depends on the time.

At time I, the treatment effect is positive. At time II, the treatment effect is negative.
Interactions in ANOVA in R

• Need to account for this potential interaction when writing the model in R

```r
## No interaction in the model
aov.out = aov(expression ~ time + treatment, data=mydata)

## Add an interaction between treatment and time
aov.out = aov(expression ~ time + treatment + time:treatment, data=mydata)
```

Warning (vital) importance of replicates in order to make feasible the estimation of interactions ➔ see Experimental design course in few days!
Analysis of variance (ANOVA)

1. Install and load the library `RcmdrMisc` and `FactoMineR`

2. Plot an interaction plot using `plotMeans()` which represents the calcium concentration according to the gender and the hormone status (replace the corresponding fields in `plotMeans(response, factor1, factor2)`)

3. Using `lm` function, make an ANOVA which explains the calcium concentration with the gender, the hormone status and an interaction between them

5. Use `lsmeans(aov.out3, pairwise ~ hormone + gender, adjust = "tukey")` to make all comparisons

6. Estimate the model without interaction for comparison
Conclusion

• ANOVA is useful to
  o Compare several means simultaneously
  o Account for more complex designs with several factors

• Non parametric counterpart of ANOVA
  o Kruskall-Wallis and pairwise Wilcoxon tests

• Be careful with interactions for complex designs

• Always make plots, boxplots and interaction plots to represent the data
Linear relationship between two continuous variables
Correlation coefficient: context

<table>
<thead>
<tr>
<th></th>
<th>Cytokin1</th>
<th>Cytokin2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.94</td>
<td>10.71</td>
</tr>
<tr>
<td>2</td>
<td>9.84</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>8.53</td>
<td>9.96</td>
</tr>
<tr>
<td>4</td>
<td>9.52</td>
<td>9.32</td>
</tr>
<tr>
<td>5</td>
<td>10.42</td>
<td>9.68</td>
</tr>
<tr>
<td>6</td>
<td>11.36</td>
<td>10.06</td>
</tr>
<tr>
<td>7</td>
<td>9.90</td>
<td>9.41</td>
</tr>
<tr>
<td>8</td>
<td>10.39</td>
<td>10.53</td>
</tr>
<tr>
<td>9</td>
<td>9.95</td>
<td>8.48</td>
</tr>
<tr>
<td>10</td>
<td>8.62</td>
<td>10.31</td>
</tr>
<tr>
<td>11</td>
<td>9.59</td>
<td>8.46</td>
</tr>
<tr>
<td>12</td>
<td>9.61</td>
<td>9.70</td>
</tr>
<tr>
<td>13</td>
<td>9.94</td>
<td>9.47</td>
</tr>
<tr>
<td>14</td>
<td>11.10</td>
<td>9.35</td>
</tr>
<tr>
<td>15</td>
<td>10.76</td>
<td>9.94</td>
</tr>
<tr>
<td>16</td>
<td>9.84</td>
<td>8.09</td>
</tr>
<tr>
<td>17</td>
<td>9.75</td>
<td>11.18</td>
</tr>
<tr>
<td>18</td>
<td>10.70</td>
<td>8.34</td>
</tr>
<tr>
<td>19</td>
<td>10.56</td>
<td>9.54</td>
</tr>
<tr>
<td>20</td>
<td>9.31</td>
<td>8.88</td>
</tr>
</tbody>
</table>

- Is there a relationship between **Cytokin 1** and **Cytokin 2**?
Is there a relationship between Cytokine 1 and Cytokine 2?

<table>
<thead>
<tr>
<th></th>
<th>Cytokine 1</th>
<th>Cytokine 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.94</td>
<td>10.71</td>
</tr>
<tr>
<td>2</td>
<td>9.84</td>
<td>9.93</td>
</tr>
<tr>
<td>3</td>
<td>8.53</td>
<td>9.96</td>
</tr>
<tr>
<td>4</td>
<td>9.52</td>
<td>9.32</td>
</tr>
<tr>
<td>5</td>
<td>10.42</td>
<td>9.68</td>
</tr>
<tr>
<td>6</td>
<td>11.36</td>
<td>10.06</td>
</tr>
<tr>
<td>7</td>
<td>9.90</td>
<td>9.41</td>
</tr>
<tr>
<td>8</td>
<td>10.39</td>
<td>10.53</td>
</tr>
<tr>
<td>9</td>
<td>9.95</td>
<td>8.48</td>
</tr>
<tr>
<td>10</td>
<td>8.62</td>
<td>10.31</td>
</tr>
<tr>
<td>11</td>
<td>9.59</td>
<td>8.46</td>
</tr>
<tr>
<td>12</td>
<td>9.61</td>
<td>9.70</td>
</tr>
<tr>
<td>13</td>
<td>9.94</td>
<td>9.47</td>
</tr>
<tr>
<td>14</td>
<td>11.10</td>
<td>9.35</td>
</tr>
<tr>
<td>15</td>
<td>10.76</td>
<td>9.94</td>
</tr>
<tr>
<td>16</td>
<td>9.84</td>
<td>8.09</td>
</tr>
<tr>
<td>17</td>
<td>9.75</td>
<td>11.18</td>
</tr>
<tr>
<td>18</td>
<td>10.70</td>
<td>8.34</td>
</tr>
<tr>
<td>19</td>
<td>10.56</td>
<td>9.54</td>
</tr>
<tr>
<td>20</td>
<td>9.31</td>
<td>8.88</td>
</tr>
</tbody>
</table>
• Is there a relationship between Cytokine 3 and Cytokine 4?

<table>
<thead>
<tr>
<th>Cytokine 3</th>
<th>Cytokine 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.94</td>
</tr>
<tr>
<td>2</td>
<td>9.84</td>
</tr>
<tr>
<td>3</td>
<td>8.53</td>
</tr>
<tr>
<td>4</td>
<td>9.52</td>
</tr>
<tr>
<td>5</td>
<td>10.42</td>
</tr>
<tr>
<td>6</td>
<td>11.36</td>
</tr>
<tr>
<td>7</td>
<td>9.90</td>
</tr>
<tr>
<td>8</td>
<td>10.39</td>
</tr>
<tr>
<td>9</td>
<td>9.95</td>
</tr>
<tr>
<td>10</td>
<td>8.62</td>
</tr>
<tr>
<td>11</td>
<td>9.59</td>
</tr>
<tr>
<td>12</td>
<td>9.61</td>
</tr>
<tr>
<td>13</td>
<td>9.94</td>
</tr>
<tr>
<td>14</td>
<td>11.10</td>
</tr>
<tr>
<td>15</td>
<td>10.76</td>
</tr>
<tr>
<td>16</td>
<td>9.84</td>
</tr>
<tr>
<td>17</td>
<td>9.75</td>
</tr>
<tr>
<td>18</td>
<td>10.70</td>
</tr>
<tr>
<td>19</td>
<td>10.56</td>
</tr>
<tr>
<td>20</td>
<td>9.31</td>
</tr>
</tbody>
</table>
Correlation coefficient: context

- Is there a relationship between Cytokine 3 and Cytokine 4?

<table>
<thead>
<tr>
<th>Cytokine 3</th>
<th>Cytokine 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.94</td>
</tr>
<tr>
<td>2</td>
<td>9.84</td>
</tr>
<tr>
<td>3</td>
<td>8.53</td>
</tr>
<tr>
<td>4</td>
<td>9.52</td>
</tr>
<tr>
<td>5</td>
<td>10.42</td>
</tr>
<tr>
<td>6</td>
<td>11.36</td>
</tr>
<tr>
<td>7</td>
<td>9.90</td>
</tr>
<tr>
<td>8</td>
<td>10.39</td>
</tr>
<tr>
<td>9</td>
<td>9.95</td>
</tr>
<tr>
<td>10</td>
<td>8.62</td>
</tr>
<tr>
<td>11</td>
<td>9.59</td>
</tr>
<tr>
<td>12</td>
<td>9.61</td>
</tr>
<tr>
<td>13</td>
<td>9.94</td>
</tr>
<tr>
<td>14</td>
<td>11.10</td>
</tr>
<tr>
<td>15</td>
<td>10.76</td>
</tr>
<tr>
<td>16</td>
<td>9.84</td>
</tr>
<tr>
<td>17</td>
<td>9.75</td>
</tr>
<tr>
<td>18</td>
<td>10.70</td>
</tr>
<tr>
<td>19</td>
<td>10.56</td>
</tr>
<tr>
<td>20</td>
<td>9.31</td>
</tr>
</tbody>
</table>

How to summarize in a value this relationship?
As heights of the parents deviated from the average height (that is as they became taller or shorter than the average adult), their children tended to be less extreme in height: the heights of the children regressed to the average height of an adult.

“Regression towards Mediocrity in Hereditary Stature.” – Francis Galton (1886)
First step: the covariance

On a bivariate sample \((x, y)\) of size \(n\) and means equal to \(m_x\) and \(m_y\)

\[
cov(x, y) = \frac{(x_1 - m_x)(y_1 - m_y) + \cdots + (x_n - m_x)(y_n - m_y)}{n - 1}
\]

also denoted:

\[
cov(x, y) = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)
\]
And finally: the Pearson correlation coefficient

\[ \rho = \text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{v}(x)}\sqrt{\text{v}(y)}} \]

✓ Normalized covariance:
Pearson Correlation Coefficient

- The correlation coefficient is always between -1 and +1
  - \( \rho = 1 \) perfect positive relation \( \rightarrow y = a \times x + b \)
  - \( \rho = -1 \) perfect negative relation \( \rightarrow y = -a \times x + b \)
  - \( \rho = 0 \) no linear relation between \( x \) and \( y \)
The four quadrants

$\rho > 0$: the viral load increases when the dose of treatment increases.
The four quadrants

$\rho < 0$ : the viral load decreases when the dose of treatment increases
Exercise: link between birth weight and abdominal diameter

Steps:

1. Import the « BirthWeight.csv » dataset in R using `read.table()` function. Be careful with the separator, it is ";" so add the argument `sep=";"` in your line of code

2. Plot the birth weight « bw » in function of the abdominal diameter « ad »

3. Compute the Pearson correlation coefficient, using `cor` function
Correlation with cautious – 1

Correlation on this dataset: 0.0008

Compute the correlation coefficient between $W$ and $Y$, and also $W^2$ and $Y$. 
Correlation with cautious – 2

Correlation on the whole dataset: -0.08
Correlation on the blue dots: -0.88
Correlation with cautious – 3

✓ A regular situation in biological data: having a lot of zero values

> cor(y, x)
[1] 0.3564926

> cor(y[x > 0], x[x>0])
[1] 0.01611771
Correlation with cautious – 3

✓ A regular situation in biological data: having a lot of zero values

```r
> cor(y,x)
[1] 0.3564926
```

```r
> cor(y[x > 0], x[x > 0])
[1] 0.01611771
```

Note: if you have missing values (NA), you can use:

```r
> cor(y, x, na.rm=TRUE)
```
What can we measure with correlation?

✓ Link between variables
✓ Link between responses at different times
✓ Link between paired individuals
✓ also...
✓ Link between a situation of interest (mutation, disease, radiation etc.) and a control
Test your intuition:
http://www.guessthecorrelation.com

source: http://www.guessthecorrelation.com
Hypothesis testing and confidence interval of a linear correlation coefficient

✓ Biological question
  o At the population level, is the birth weight linearly correlated to abdominal diameter?

✓ We observed $R = 0.87$, is this correlation significant?
  o One solution: Hypothesis testing of the correlation coefficient

✓ What confidence can I put in my observed value of correlation?
  o One solution: Confidence interval for the correlation coefficient
Hypothesis testing of a linear correlation coefficient

✓ Hypothesis testing
  o $H_0: \rho = 0$ (no linear relation between the two continuous variables)
  o $H_1: \rho \neq 0$ (linear relation between the two continuous variables)
Hypothesis testing of a linear correlation coefficient

✓ Hypothesis testing
  o $H_0: \rho = 0$ (no linear relation between the two continuous variables)
  o $H_1: \rho \neq 0$ (linear relation between the two continuous variables)

✓ Computation of test statistics

$$T = \frac{\rho \sqrt{n-2}}{\sqrt{1-\rho^2}}$$

✓ Decision making (with R: `cor.test()` function)

  o If $|T| \leq t_{n-2} \left(1 - \frac{\alpha}{2}\right)$: do not reject $H_0$
  o If $|T| \geq t_{n-2} \left(1 - \frac{\alpha}{2}\right)$: reject $H_0$
Exercise: link between birth weight and abdominal diameter

Steps:

1. Come back to the « BirthWeight.csv » dataset, and test the significance of the computed correlation coefficients

2. Give a confidence interval for the correlation coefficient
Exercise: link between birth weight and abdominal diameter

```r
> cor.test(dta$bw, dta$ad)
```

Pearson's product-moment correlation

```
data:  dta$bw and dta$ad
t = 18.347, df = 105, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
  0.8189566 0.9117873
sample estimates:
cor
  0.8730657
```

P-value of the test

Recall of the alternative hypothesis

Confidence interval for $\rho$

Value of the correlation coefficient
Spurious correlation and regression fallacy

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

> cor.test(c(109,102,102,98,85,95,96,98,123,94,102),c(2,2,2,3,1,1,2,3,4,1,4))

Pearson's product-moment correlation

data: c(109, 102, 102, 98, 85, 95, 96, 98, 123, 94, 102) and c(2, 2, 2, 3, 1, 1, 2, 3, 4, 1, 4)  
t = 2.6785, df = 9, p-value = 0.02527  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
0.1101273 0.9045101  
sample estimates:  
cor  
0.6660043
Spurious correlation and regression fallacy

Number of people who drowned by falling into a pool correlates with Films Nicolas Cage appeared in

> cor.test(c(109, 102, 102, 98, 85, 95, 96, 98, 123, 94, 102), c(2, 2, 2, 3, 1, 1, 2, 3, 4, 1, 4))

Pearson's product-moment correlation

data: c(109, 102, 102, 98, 85, 95, 96, 98, 123, 94, 102) and c(2, 2, 2, 3, 1, 1, 2, 3, 4, 1, 4)
t = 2.6785, df = 9, p-value = 0.02527
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.1101273 0.9045101
sample estimates:
cor
0.6660043

Hello, I am in 6 new films in 2017!
Spurious correlation and regression fallacy


Citing articles = 48, I.F. = 55.873 (2014)

“Conclusions
Chocolate consumption enhances cognitive function, which is a sine qua non for winning the Nobel Prize, and it closely correlates with the number of Nobel laureates in each country.”
Exercise: Scatterplot and correlation matrix

Steps:

1. Coming back to the « BirthWeight.csv » dataset, use `plot()` to display a scatterplot of all the variables of the dataset

2. Compute the correlation matrix between all the variables of the dataset using `cor()`

3. Install and load the `psych` package

4. Compute the matrix of p-values to know if each correlation is significant with the `corr.test()` function of the `psych` package

5. Install and load the R package `corrplot`

6. Use `corrplot()` and `corrplot.mixed()` to display correlation matrices
Exercise: Make nice correlation matrices with corrplot

Steps:

1. Install and load the R package `corrplot`

2. Use `corrplot()` and `corrplot.mixed()` to display correlation matrices
Exercise: Make nice correlation matrices with corrplot

More options: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
Some remarks

✓ Correlation is different from causation

Be careful of confounding factors
Linear regression
Intents of the course

✓ Understand how linear regression works and how R can be used for this purpose.
✓ Being able to understand and plot this kind of graph with R:

✓ Being able to use linear regression to select a subset of relevant variables from a set of predictors
The « population »

Data population are generated by a parametric model

SAMPLING (SIMULATION)

Data Sample

The sample we observe
Statistical inference

The « population »

Data population are generated by a parametric model

Data Sample

The sample we observe

INERENCE

Estimating parameters of a model from a data sample using relevant assumptions

Regression
Linear regression models are a special case of regression models

Linear regression models are a special case of regression models

Figure 9.2 All (or most) of statistics. The labels in parentheses (non-normal errors and nonlinearity) imply restricted cases: (non-normal errors) means exponential family (e.g., binomial or Poisson) distributions, while (nonlinearity) means nonlinearities with an invertible linearizing transformation. Models to the right of the gray dashed line involve multiple levels or types of variability; see Chapter 10.
Why using a linear regression?

- How to show there is a significant effect of “pack_year” on “tumor_size”?
- How to model mathematically this influence?
Why using a linear regression?

• How to show there is a significant effect of “pack_year” on “tumor_size”?

• How to model mathematically this influence?

REGRESSION model:

\[ \text{tumor size}_i = f(\text{pack year}_i) + \text{error}_i \]
Why using a linear regression?

• How to show there is a significant effect of “pack_year” on “tumor_size”?

• How to model mathematically this influence?

LINEAR regression model:

\[
tumor\_size_i = a + b \times pack\_year_i + error_i
\]
Why using a linear regression?

• How to show there is a significant effect of “pack_year” on “tumor_size”?

• How to model mathematically this influence?

**LINEAR regression model:**

\[ \text{tumor\_size}_i = \alpha + b \times \text{pack\_year}_i + \text{error}_i \]

The « intercept »

The coefficient associated to « pack\_year »

<table>
<thead>
<tr>
<th>PREDECTOR</th>
<th>RESPONSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pack_year</td>
<td>tumor_size</td>
</tr>
<tr>
<td>12</td>
<td>23.6</td>
</tr>
<tr>
<td>9</td>
<td>20.8</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>19.1</td>
</tr>
<tr>
<td>5</td>
<td>18.2</td>
</tr>
<tr>
<td>13</td>
<td>25.3</td>
</tr>
<tr>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>24</td>
<td>33.1</td>
</tr>
<tr>
<td>16</td>
<td>25.3</td>
</tr>
<tr>
<td>18</td>
<td>28.1</td>
</tr>
<tr>
<td>21</td>
<td>30.7</td>
</tr>
<tr>
<td>5</td>
<td>15.4</td>
</tr>
<tr>
<td>10</td>
<td>22.9</td>
</tr>
</tbody>
</table>
Simple linear regression with R

**Steps:**

1. Import the «`lung_cancer.csv`» dataset in R with

   ```r
   read.table("path to your file/lung_cancer.csv",header=TRUE,sep=" ")
   ```

2. Plot the response «`tumor_size`» in function of the predictor «`pack_year`»

3. Make a Pearson correlation coefficient analysis to know if a linear relationship can be assumed

4. Use `lm()` to fit the linear model to explain «`tumor_size`» in function of the predictor «`pack_year`» and plot the regression line with `abline()`

Remark: If `mod` is your output of the `lm()` function, use `summary(mod)` to print out the summary of this output
How are the coefficients estimated?

\[ y_i = a + b \times x_i + \varepsilon_i \]

\( \hat{a} \) and \( \hat{b} \) are fitted by minimizing the squares of the error terms:

\[
\left( \hat{a}, \hat{b} \right) = \text{arg min}_{(a,b)} \sum_{i=0}^{n} (y_i - a - b \times x_i)^2 = \text{arg min}_{(a,b)} \sum_{i=0}^{n} (\varepsilon_i)^2
\]
How are the coefficients estimated?

Tumor size of patients in function of their pack-year smoking history

(number of packs of cigarettes smoked per day) × (number of years the person has smoked)
How are the coefficients estimated?
How are the coefficients estimated?

Tumor size of patients in function of their pack-year smoking history

(number of packs of cigarettes smoked per day) × (number of years the person has smoked)
How are the coefficients estimated?
How are the coefficients estimated?

The $i^{th}$ coefficient of vector $\text{tumor\_size}$

$y_i = a + b \times x_i + \varepsilon_i$

The $i^{th}$ coefficient of vector $\text{pack\_year}$

$\hat{a}$ and $\hat{b}$ are fitted by minimizing the squares of the error terms:

What is equivalent to cancel the derivatives with respect to $a$ and $b$:

\[
\begin{align*}
\frac{d}{da} \sum_{i=0}^{n} (y_i - a - b \times x_i)^2 &= 0 \\
\frac{d}{db} \sum_{i=0}^{n} (y_i - a - b \times x_i)^2 &= 0
\end{align*}
\]

\[
\begin{align*}
\hat{b} &= \frac{\sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n} (x_i - \bar{x})^2} \\
\hat{a} &= \bar{y} - \hat{b} \bar{x}
\end{align*}
\]

what R uses
What are the assumptions of a standard linear regression?

Response = $a + b \times \text{Predictor1} + c \times \text{Predictor2} + \ldots + \text{error}$

1. The « response » variable is a **continuous** variable.
2. There are **more observations than predictor variables**.
3. There is a **linear relationship** between the response and the predictor(s) (or explanatory variable(s)).
4. The **errors** are **independently distributed**.
5. The **errors** are **identically distributed** following a **Normal law centered on 0**.
6. The **errors** have a constant variance (**homoskedasticity**).
7. No **multicollinearity** between predictors if they are several.
1. The « response » variable is a **continuous** variable

**Discrete variables:**

Take only certain values along an interval.

Integers between 0 and 8:
0, 1, 2, 3, 4, 5, 6, 7, 8.

**Continuous variables:**

Take any value at any point along an interval.

Any real value between 0 and 8:
0.123, 1.908, 2.463, 3.013, 4.972, 5.839, 6.998, 7.638, ....
2. There are more observations than predictor variables

Response = \( a + b \times \text{Predictor1} + c \times \text{Predictor2} + \ldots \) + error

With \( n \) observations and \( p \) predictor variables:

\[
\begin{align*}
y_1 &= a + b \times x_{1,1} + \cdots + z \times x_{1,p} + \varepsilon_1 \\
y_2 &= a + b \times x_{2,1} + \cdots + z \times x_{2,p} + \varepsilon_2 \\
&\vdots \\
y_i &= a + b \times x_{i,1} + \cdots + z \times x_{i,p} + \varepsilon_i \\
&\vdots \\
y_n &= a + b \times x_{n,1} + \cdots + z \times x_{n,p} + \varepsilon_n
\end{align*}
\]

\( n > p \)
3. There is a **linear relationship** between the response and the predictor(s) (or explanatory variable(s))

\[ \text{Response} = a + b \times \text{Predictor} + \text{Error} \]
3. There is a **linear relationship** between the response and the predictor(s) (or explanatory variable(s))

\[
\text{Response} = a + b \times \text{Predictor} + \text{Error}
\]
4. The errors are **independently distributed**
4. The **errors** are *independently distributed*

« If we know error terms, we cannot predict other error terms »
5. The errors are **identically distributed following a Normal law centered on 0**.

![Diagram of response vs predictor with a normal distribution centered on 0.](image)

*Same distribution for all errors*
5. The errors are **identically distributed** following a Normal law centered on 0

**Useful trick**

1) Plot an histogram or a qqplot of your response to graphically check the normality of the variable (if you have enough observations...)

2) Use transformations to get closer to Normal distribution (ex: log, sqrt, power transform...)

3) Plot an histogram or a qqplot of the transformed variable to check that the transformation is correct
5. The errors are **identically distributed following a Normal law centered on 0**

**Useful trick**

1) Plot an histogram or a qqplot of your response to graphically check the normality of the variable (if you have enough observations...)

2) Use transformations to get closer to Normal distribution (ex: log, sqrt, power transform...)

3) Plot an histogram or a qqplot of the transformed variable to check that the transformation is correct
Reminder about quantiles

**Definition**

The $q^{th}$ theoretical quantile corresponds to an actual value $Q$ which separates the universe in two sub-groups such that:

- $q\%$ of the observed values are lower than $Q$
- $(1 - q)\%$ are greater than $Q$
Study of the normality of the tumor size

Steps:

1. Use `qqnorm()` function to study the normality of the tumor size

2. Use `qqline()` function to add a line as a landmark

Do not hesitate to custom your plot by changing the color (col=...), the type of point (pch=...), the title (main = ...)
6. The **errors** have a constant variance (homoskedasticity)
6. The **errors** have a constant variance *(homoskedasticity)*

- **Homoskedasticity**
  - Constant variance of the errors
  - Mean = 0

- **Heteroskedasticity**
  - Non-constant variance of the errors
7. No **multicollinearity** between predictors if they are several

- **Collinearity:**
  
  \[ \text{Predictor1} = a + b \times \text{Predictor2} \]

- **Multicollinearity:**

  \[ \text{Predictor1} = a + b \times \text{Predictor2} + c \times \text{Predictor3} \]

- **No multicollinearity:**

  Each predictor is not equal to a linear combination of other predictors.

---

*If there is multicollinearity, then R cannot estimate all the model coefficients and some NA coefficients will appear for predictors equal to a linear combination of others by using the `lm()` function.*
The `summary()` function

```r
> sf = summary(fit)
> sf

Call:
`lm(formula = tumor_size ~ pack_year)`

Residuals:
       Min      1Q  Median      3Q     Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.9274     0.7503   18.56 9.27e-12 ***
pack_year    0.8072     0.0543   14.87 2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared: 0.9364, Adjusted R-squared: 0.9322
F-statistic: 221 on 1 and 15 DF, p-value: 2.202e-10
```
The `summary()` function

```r
> sf = summary(fit)
> sf

Call:
lm(formula = tumor_size ~ pack_year)

Residuals:
  Min     1Q   Median     3Q    Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640

Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)    
(Intercept) 13.9274     0.7503   18.56 9.27e-12 ***
pack_year   0.8072     0.0543   14.87 2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared: 0.9364,  Adjusted R-squared: 0.9322
F-statistic: 221 on 1 and 15 DF,  p-value: 2.202e-10

The quartiles of the distribution of residuals

The summary() function
The `summary()` function

```r
> sf = summary(fit)
> sf

Call:
lm(formula = tumor_size ~ pack_year)

Residuals:
    Min      1Q  Median     3Q    Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.9274     0.7503   18.56  9.27e-12 ***
pack_year     0.8072     0.0543   14.87  2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared: 0.9364, Adjusted R-squared: 0.9322
F-statistic: 221 on 1 and 15 DF, p-value: 2.202e-10

The quartiles of the distribution of residuals

Two results of Student (or Wald) tests to know if each coefficient is significantly not null. This can be assumed if Pr(>|t|) is small enough.
The `summary()` function

```r
> sf = summary(fit)
> sf

Call:
  lm(formula = tumor_size ~ pack_year)

Residuals:
    Min     1Q  Median    3Q     Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept) 13.9274     0.7503   18.56   9.27e-12 ***
pack_year   0.8072     0.0543   14.87   2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared: 0.9364, Adjusted R-squared: 0.9322
F-statistic: 221 on 1 and 15 DF, p-value: 2.202e-10
```

The quartiles of the distribution of residuals

Two results of Student (or Wald) tests to know if each coefficient is significantly not null. This can be assumed if \( \Pr(>|t|) \) is small enough.

An unbiased estimate of the standard deviation of the distribution of residuals
The `summary()` function

> sf = summary(fit)
> sf

Call:
`lm(formula = tumor_size ~ pack_year)`

Residuals:
```
   Min     1Q  Median     3Q    Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.9274     0.7503   18.56 9.27e-12 ***
pack_year   0.8072     0.0543   14.87 2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared: 0.9364, Adjusted R-squared: 0.9322
F-statistic: 221 on 1 and 15 DF, p-value: 2.202e-10

The quartiles of the distribution of residuals

Two results of Student (or Wald) tests to know if each coefficient is significantly not null. This can be assumed if Pr(>|t|) is small enough.

An unbiased estimate of the standard deviation of the distribution of residuals

The adjusted multiple correlation coefficient $R^2$
The `summary()` function

```r
> sf = summary(fit)
> sf

Call:
  lm(formula = tumor_size ~ pack_year)

Residuals:
     Min      1Q  Median       3Q      Max
-2.5636 -0.4780 -0.1794  0.8075  2.5640

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 13.9274     0.7503   18.56 9.27e-12 ***
pack_year   0.8072     0.0543   14.87 2.20e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.332 on 15 degrees of freedom
Multiple R-squared:  0.9364, Adjusted R-squared:  0.9322
F-statistic: 221 on 1 and 15 DF,  p-value: 2.202e-10
```

The quartiles of the distribution of residuals

Two results of Student (or Wald) tests to know if each coefficient is significantly not null. This can be assumed if Pr(>|t|) is small enough.

An unbiased estimate of the standard deviation of the distribution of residuals

The adjusted multiple correlation coefficient $R^2$

The p-value of the Fisher test $R^2 = 0$ vs $R^2 \neq 0$
About the coefficient of determination $R^2$ (or multiple correlation coefficient)

Total Sum of Squares (TSS) = Explained Sum of Squares (ESS) + Residuals Sum of Squares (RSS)

\[
\sum_{i} (y_i - \bar{y})^2 = \sum_{i} (\hat{y}_i - \bar{y})^2 + \sum_{i} (y_i - \hat{y}_i)^2
\]
About the coefficient of determination $R^2$ (or multiple correlation coefficient)

Total Sum of Squares (TSS)

$\sum_i (y_i - \bar{y})^2$

Response variability

Explained Sum of Squares (ESS)

$\sum_i (\hat{y}_i - \bar{y})^2$

Variability explained by the predicted values using the fitted regression model

Residuals Sum of Squares (RSS)

$\sum_i (y_i - \hat{y}_i)^2$

Variability not explained by the predicted values
About the coefficient of determination $R^2$ (or multiple correlation coefficient)

Total Sum of Squares (TSS)

Explained Sum of Squares (ESS)

Residuals Sum of Squares (RSS)

\[
\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2
\]

\[
R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}
\]
How to interpret the $R^2$ coefficient?
Quizz about the coefficient of determination $R^2$ (or multiple correlation coefficient)

- How to interpret the $R^2$ coefficient?
  The $R^2$ evolves between 0 and 1. The closer to 1 it is, the better the model represents the observed response.
Quizz about the coefficient of determination $R^2$ (or multiple correlation coefficient)

• How to interpret the $R^2$ coefficient?
  The $R^2$ evolves between 0 and 1. The closer to 1 it is, the better the model represents the observed response.

• What is the relation between $R^2$ and the Pearson correlation coefficient?
Quizz about the coefficient of determination $R^2$ (or multiple correlation coefficient)

• **How to interpret the $R^2$ coefficient?**
  The $R^2$ evolves between 0 and 1. The closer to 1 it is, the better the model represents the observed response.

• **What is the relation between $R^2$ and the Pearson correlation coefficient?**
  The $R^2$ is equal to the square of the Pearson correlation coefficient if there is just one explanatory variable. But it can be used with several explanatory variables!
• **How to interpret the R² coefficient ?**
  The R² evolves between 0 and 1. The closer to 1 it is, the better the model represents the observed response.

• **What is the relation between R² and the Pearson correlation coefficient ?**
  The R² is equal to the square of the Pearson correlation coefficient if there is just one explanatory variable. But it can be used with several explanatory variables !

• **What does R² = 0.93 mean ?**
Quizz about the coefficient of determination $R^2$ (or multiple correlation coefficient)

• **How to interpret the $R^2$ coefficient?**
  The $R^2$ evolves between 0 and 1. The closer to 1 it is, the better the model represents the observed response.

• **What is the relation between $R^2$ and the Pearson correlation coefficient?**
  The $R^2$ is equal to the square of the Pearson correlation coefficient if there is just one explanatory variable. But it can be used with several explanatory variables!

• **What does $R^2 = 0.93$ mean?**
  It means the « pack_year » variable explains around 93% of the variability of the « tumor_size » variable with the fitted regression model.
About the ADJUSTED coefficient of determination $R^2$

- The « adjusted » $R^2$ is a ratio of unbiased estimates of variances instead of simple sum of squares.

$$\text{Adj. } R^2 = 1 - \frac{RSS/(n - p)}{TSS/(n - 1)}$$
About the ADJUSTED coefficient of determination $R^2$

- The « adjusted » $R^2$ is a ratio of unbiased estimates of variances instead of simple sum of squares.

$$Adj. R^2 = 1 - \frac{RSS/(n - p)}{TSS/(n - 1)}$$

- It is more relevant than the simple $R^2$: each time when you add a predictor variable, the simple $R^2$ always either increase or stays constant, but not the adjusted $R^2$: it can decrease. The adjusted $R^2$ will increase only if the added predictor has some interest.
About the ADJUSTED coefficient of determination $R^2$

• The « adjusted » $R^2$ is a ratio of unbiased estimates of variances instead of simple sum of squares.

$$Adj. R^2 = 1 - \frac{RSS/(n - p)}{TSS/(n - 1)}$$

• It is more relevant than the simple $R^2$: each time when you add a predictor variable, the simple $R^2$ always either increase or stays constant, but not the adjusted $R^2$: it can decrease. The adjusted $R^2$ will increase only if the added predictor has some interest.

• Check the $F$-stat p-value to determine if the adjusted $R^2$ is significantly not null! If it is null, the fitted model is not able to explain the variability of the response variable.
R² and confidence bands

1. Extract the adjusted R² and the p-value of the hypothesis testing « R²=0 » vs « R²≠ 0 ». What is your conclusion?

2. Use the predict() function to extract predicted values. What are these values?

3. On a plot with observed data, add a line with the predicted values (use lines())
Fitting confidence bands

- `interval="prediction"` : the confidence bands concerns the distribution of points generated from the regression model

- `interval="confidence"` : the confidence bands concerns the estimation of the regression line

- To plot the confidence bands, you have to order the data :
  ```
  lines(pack_year[order(pack_year)],cb[order(pack_year),2])
  ```
Highlighting outliers

1. Insert two additional (factitious) outliers in your dataset:

```r
lung.2 <- lung[,c("tumor_size","pack_year")]] # to not modify the true dataset
myoutliers = data.frame(pack_year = c(7,40),tumor_size = c(30,35))
lung.2 <- rbind(lung.2,myoutliers)
```

2. Make a scatter plot to visualize these outliers and fit the linear model

3. Use `plot(mod)` to highlight these outliers
Highlighting outliers

> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:

Large residuals (= far from the regression line)
Highlighting outliers

Empirical quantiles of the residuals divided by their standard deviation

Quantiles associated to outliers does not correspond to theoretical quantiles (= Normality assumption not respected )

Quantiles of a standard Normal distribution N(0,1)
Highlighting outliers

> plot(fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:

Square root of residuals divided by their standard deviation
Highlighting outliers

> plot(fit)

A point superior to the Cook’s distance of 1 has to be considered as an « influential » point.

When it is superior to 0.5, it is not classified as « influential » but it has to be examined carefully.
Trick: Using functions of predictors allows making linear nonlinear relations
Trick: Using functions of predictors allows making linear nonlinear relations.
1. Import the « lungA » dataset in your R session:
   `read.table("path to your file/lungA.csv",header=TRUE,sep="","`)`

2. Make a scatterplot with the variable « tumor_size », « age », « pack_year », « pollution »

3. Display the correlation matrix between these variables (using `corrplot()` from package `corrplot`)

4. Use the `lm()` function to explain the tumor_size in function of the age, pack_year and pollution variables. Interpret the result
Tricks when using the \texttt{lm()} function

- \texttt{lm(tumor_size~., data=data_reg)}: allows explaining the « tumor_size » variable with all the variables of the dataset « data_reg »

- \texttt{lm(tumor_size~.-pollution, data=data_reg)}: allows explaining the « tumor_size » variable with all the variables of the dataset « data_reg » except the « pollution » variable

- \texttt{lm(tumor_size~pack_year*gender, data=data_reg)}: allows to consider interaction factors, i.e.:

  \[
  \text{tumor\_size} = a + b \times \text{pack\_year} + c \times \text{gender} + d \times \text{gender} \times \text{pack\_year} + \text{Error}
  \]

- \texttt{lm(tumor_size~I(pack\_year^2), data=data_reg)}: the \texttt{I()} allows to write operations as in the R console
Selecting predictors with stepwise regression

- **Problem:** a large set of candidate predictor variables. Some are redundant and do not provide much information in relation to others.

- **Principle of Occam’s Razor (or Parcimonious principle):**
  
  “among several plausible explanations, the simplest is the best”

  [Image of William of Ockham]

  William of Ockham (Gulielmus Occamus; 1287 – 1347)

- **Goal:** to choose a small subset of predictors so that the resulting regression model is simple and have a good predictive ability.
Stepwise regression

- **How it works:** Enter and remove predictors in a stepwise manner until a specified criterion is minimized (or maximized).

- **Which criteria can be used?**
  - Akaike or Bayesian Information Criteria (penalized likelihood)
  - Adjusted R² (be careful to not use the simple R²: it would always choose the largest possible model)
  - Others (Mallow’s Cp, PRESS, etc.)
### (very short) Summary of the course

<table>
<thead>
<tr>
<th>Data types</th>
<th>Example</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 continuous variables</td>
<td>Size of the tumor and number of packs smoked by year</td>
<td>Correlation coefficient + transformation if needed (e.g. log)</td>
</tr>
<tr>
<td>2 continuous Gaussian variables</td>
<td>Size of the tumor and number of packs smoked by year</td>
<td>Simple linear regression + transformation if the relation is not linear (e.g. log)</td>
</tr>
<tr>
<td>Several continuous explaining variables + 1 continuous response</td>
<td>Number of packs smoked by year, alcohol consumption, age, sex ... + size of the tumor</td>
<td>Multiple linear regression + transformation if needed (e.g. log) (+ variable selection to reduce the variables to the most relevant ones)</td>
</tr>
<tr>
<td>One or more explaining factors + 1 continuous response</td>
<td>Strain type, treatment, day of the experiment + viral load</td>
<td>ANOVA + transformation if needed (e.g. log)</td>
</tr>
</tbody>
</table>

Reminder about transformations such as log: they are very useful on biological data to make a relation « more » linear or to get data « closer » to a Gaussian distribution (we will see it during the project)
Some bibliography to go further

• Nice tutorial: http://tutorials.iq.harvard.edu/R/Rstatistics/Rstatistics.html

• R package **car** contains some useful functions for regression