

Biometry. Lecture 18

Alexey Shipunov

Minot State University

April 13, 2016



1 2D statistics

- Strength and sign of relation: correlation
- Anatomy of relation: regression



```
> setwd("<working folder>")  
or  
"Change dir"  
in menu!
```

On Mac, be sure that startup option is working: `getwd()`
(`getwd()` checks if R is in working folder, `dir()` checks the folder
content)



2D statistics

Stength and sign of relation: correlation



Covariance and correlation

- It is always interesting to know, **how much** are two random variables change together. Covariance show that but it is not easy to interpret.
- **Correlation coefficient** is a normalized version of covariance and therefore widely used as a measure of correlation. If correlation is close to 1 or -1 , it is high.
- Therefore, correlation coefficient will show the strength of relation



Features of correlation coefficient

- Correlation is a measure of **linear** relation. If relation is non-linear, correlation could be small or even zero. To check the linearity, it is recommended to make a `plot()` of two variables (scatterplot).
- Correlation may be positive or negative (from -1 to 1). If you need a sign-less measure, you may use determination coefficient = correlation coefficient²
- Correlation will only show that relation exists and has some strength, it will not show any other details about relation. For example, if correlation between A and B is high, it could mean that:
 - A depends on B
 - B depends on A
 - A and B depends on each other
 - A and B both independently depend on C and have nothing in common



Calculation of correlation coefficient

```
> cor(5:15, 7:17)
> cor(5:15, c(7:16, 23))
> cor(5:15, c(7:16, 2))
> cor(5:15, 17:7)
> cor(trees)
```

`cor()` function works with vectors or tables (matrices and data frames). If NAs are present, one may use option `use="complete.obs"` (better) or `use="pairwise.complete.obs"`



Non-parametric correlation

By default, `cor()` calculates parametric Pearson's correlation coefficient, it is possible to specify non-parametric (Spearman or Kendall) coefficients.

```
> cor(5:15, 7:17, method="spearman")
```



Visualization of correlation

```
> cor(longley)
> symnum(cor(longley))
> # install.packages("ellipse")
> library(ellipse)
> plotcorr(cor(longley), type="lower")
> pairs(cor(longley))
> source("http://ashipunov.info/r/pleiad.r")
> Pleiad(cor(longley)) # requires pleiad.r
> library(lattice)
> levelplot(cor(longley))
```



Correlation tests

- The alternative hypotheses for these tests is that correlation differs from zero
- There are both parametric and non-parametric tests



Correlation tests

```
> with(trees, cor.test(Girth, Volume))  
> with(trees, cor.test(Girth, Height, method="spearman"))
```



2D statistics

Anatomy of relation: regression



Idea of regression

- The basic regression analysis will apply the linear model for data
- It will study not only if variables are associates and not only the strength of association but also the shape of association (law of association)



Regression formula

- The simplest is a linear regression,

$$m = b_0 + b_1 \times x,$$

where m is a **predicted value**, x is an **independent variable** and b_0 and b_1 are coefficients (so-called **intersect** and **slope**).

- In other terms, linear regression is
response = intersect + slope * influence
- In R model formula language, it is simply
response ~ influence



Analysis of the regression model

- If y is a real response, then error of model

$$E = y - m$$

- If σ^2 are variances of m and y , then

$$R^2 = 1 - \sigma_m^2 / \sigma_y^2,$$

- In a background, R^2 is similar to coefficient of determination



Test for the regression

- To test if regression model is correct, the Fisher test is normally applied
- Null hypothesis for Fisher test is that a model is not reliable



Regression example: women data

```
> lm.women <- lm(weight ~ height, data = women)
> plot(weight ~ height, data = women, main="",
+ xlab="Height (feet)", ylab="Weight (pounds)")
> grid()
> abline(lm.women, col="red")
> source("http://ashipunov.info/r/asmisc.r") # for Cladd()
> Cladd(lm.women, data=thuesen)
```



Analysis of the regression assumptions

```
> summary(lm.women)
```

For help, use `?summary.lm` because `summary()` is generic.



Analysis of analysis

- Resulted model: $\text{weight} = -87.51667 + 3.45 * \text{height}$
- Maximum deviations from model are -1.7333 and 3.1167 pounds
- Almost half of residuals are between first and third quartiles
- All coefficients are significant
- Adjusted R-squared is close to 1 (very high!)
- The overall p-value is much less than 0.05 therefore the model is reliable
- There are 1 and 13 degrees of freedom (for columns and for rows)



Diagnostic plots for regression

```
> plot(lm.women) # Hit "Enter" to change plot
```

- “Residuals vs. Fitted”: checks **linearity**, the best is flat line
- “Normal Q-Q”: checks residuals for normal distribution, if they are not normal then our regression is not linear
- “Scale-Location”: checks the trend in dispersion, should be random
- “Residuals vs. Leverage & Cook’s distance”: checks the most influential observations
- Use `?plot.lm` for help.



Regression training

For *Cochlearia* data, tell:

- Which two measurement characters are most correlated?
- Is this correlation reliable?
- Is the linear model based on these characters reliable?



Finishing...

Save your commands!

`(savehistory(<today's date>.r)` or File -> Save as... on
Mac)



Summary: most important commands

- `cor()` —calculates correlation coefficients
- `cor.test()` —run correlation tests
- `lm()` —estimate the linear regression



For Further Reading



A. Shipunov.

Biometry [Electronic resource].

2012—onwards.

Mode of access:

http://ashipunov.info/shipunov/school/biol_240



A. Shipunov, and many others.

Visual statistics. Use R!

2016—onwards.

Mode of access: http://ashipunov.info/shipunov/school/biol_240/en/visual_statistics.pdf

