Testing for relationships Part 1: Correlation and regression

Aims of this practical

1. Understand when to use simple correlations and regressions

- Spearmans rank correlation
- Pearsons correlation
- Linear regression

Coding you will learn:

Tests conducted

Name	Data type	Test type
Spearman rank correlation	Two continuous (or ordinal) variables	Non parametric
Pearson's correlation	Two continuous variables	Parametric
Linear regression	Two continuous variables	Parametric

Other useful code for…

- Removing NA (missing data)
- Exploring intercorrelation with correlation matrices and plots

Relationships

The last few weeks we have been testing for differences – we are now ***testing for relationships***.

Correlation and regression are two of the simplest and most common forms of statistics. They are used to investigate *relationships* between two variables.

Correlation and regression are commonly confused with one another. The most important distinction between them is the inferences one can make from each.

- Simple linear regression between two variables implies that we think there is a *causal* relationship between the two – i.e. manipulating one variable directly results in the change of the other, e.g. light intensity and rate of photosynthesis.
- Correlations are used when we do not have an expectation of a causal relationship. However, just because there doesn't need to be a causal relationship, doesn't mean we should always correlate things. Be careful of spurious correlations: there is a very strong positive correlation between the consumption of cheese per person and the number of people who died getting tangled in their bedsheets. Are these sensible to compare? See http://www.tylervigen.com/spurious-correlations

As with any analysis using measured (continuous) variables its critical to plot and visualise your data, for correlations and regressions this is also very important. See this blog for examples of the different types of relationships and non-relationships you may encounter.

http://janhove.github.io/teaching/2016/11/21/what-correlations-look-like

(note: you can copy the script from the blog and plot the various correlations yourself)

Spearman correlation

Squirrels

Download and read into R the datafile "*grey squirrel.csv*". The data are the number of offspring produced by brother and sister grey squirrels. We want to know if there is a relationship between the number of offspring produced by siblings, because different forces can influence reproductive success (rs) in each respective sex. But does success run in the family? For this dataset I want you to;

1) State the hypothesis

2) Carry out some exploratory data analysis – are the variables normally distributed?

3) Make a scatterplot of the data, either using the **plot()** function or try it in ggplot2 (hint: use geom_point()). Label it properly. Do you think these variables are correlated?

What test to use? testing for a relationship, data are not normally distributed, no expectation of a causal relationship. So we need to use a Spearman correlation.

```
## we will use the function cor.test. But the default is to do a Pearsons correlation, so we need to add in the argument method="spearman"

cor.test (squirrel$brother.rs , squirrel$sister.rs, method="spearman")
```

```
        Spearman's rank correlation rho

data:  squirrel$brother.rs and squirrel$sister.rs
S = 76404.43, p-value = 5.966e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5415275
```

Note: Warning message - on this occasion is it OK to proceed, but you might want to google what this means and/or examine the data for how many "ties" in the data there actually are.

rho is the correlation coefficient. It describes the strength and direction of the relationship. rho ranges from -1 to +1.

- If rho is 0 there is no tendency for y to increase or decrease with x – so there is no correlation.
- The closer to either -1 or +1 the more of the variation in y is described by the variation in x – so they are strongly correlated.
- If rho is +ve, y increases when x increases – positive correlation.

- If rho is –ve y decreases when x increases – negative correlation.

So from our output we can see that there is a significant positive correlation, so we can write:

RESULTS STATEMENT: There was a significant positive correlation between the reproductive success of brother and sister grey squirrels (Spearmans; n=100, p<0.001, rho=0.54)

UN development

Download and read into R the datafile "*UN_data.csv*".

This dataset includes nine development variables in 208 countries:

- Forest: Area of forest in sq km
- infant.mort: Mortality rate of infants (per 1,000 live births)
- Fertility: Fertility rate, total (births per woman)
- Domestic_fw: Annual freshwater withdrawals, domestic (% of total freshwater withdrawal)
- CO2.emission: CO2 emissions (metric tons per capita)
- gdp: GDP per capita
- region: world regions

Start by doing some inital data exploration. This exploration should enable you to answer the following questions: [hint: you might find *na.rm=TRUE* useful]. See link for more information on missing data https://www.statmethods.net/input/missingdata.html

- What is the mean infant mortality in Asia and Europe?
- Which country has the maximum CO2 emission?

To begin with, we are interested in exploring the correlation between GDP and infant mortality. Plot these variables using a scatter plot. Do you think they look correlated? What do you notice about the scale of the plot? The data on both axes range by several orders of magnitude and so it would be most appropriate to plot them on a log scale.

```
# Take the log of gdp and infant mortality and put them into a new variable
log_gdp<-log(UN$gdp)
log_infantmort<-log(UN$infant.mort)
```

Replot your scatter plot using the new logged variables. How does this look? Label it properly.

Now, test the new variables for normality. If they are normally distributed, we can use a Pearson correlation and if not we have to use a Spearman.

Carry out a correlation test to investigate the relationship between the new logged gdp and infant mortality.

Report your results.

Pearsons correlation

US Arrests

Download and read into R the datafile "*USArrests.csv*". This dataset contains the number of arrests per 10,000 people for Murder, Assault and Armed Robbery in US states, and the percentage of the population that lives in urban areas.

We are interested in how the number of arrests for different crimes are related to the proportion of the population that lives in urban areas. Let's start with the arrests for murder.

Do some preliminary data exploration and assess each of the variables for normality. If they are not normally distributed, does transforming by taking the square root or the log of the data make it normal?

Plot a scattergraph of the proportion of the urban population and the arrests for murder. Do you think that there is a correlation?

Test for correlation and report the results. Since we are testing for a relationship between two variables that are normally distributed, we can use a Pearsons correlation. The default *cor.test* function is for a Pearsons, so you don't need to state a method argument.

Regression

Download and read into R the dataset "sperm.csv". The data are counts of sperm transferred during matings (spermnum) that were disturbed at different times (minutes) in yellow dung flies (time). We want to know what effect duration of mating has on the amount of sperm transferred to females.

We can assume that there would be a causal relationship between sperm number and the length of mating duration, so we should use a regression.

Which is the response variable? Make a histogram of the response variable – is it normally distributed? Test statistically if in doubt. *Time* needn't be normal for this test as it is the explanatory variable.

Make a scatterplot of the data, either using the **plot()** function or try it in ggplot2 (hint: use geom_point()). Make sure to label it appropriately. Are the variables on the correct axis?

Now we are going to carry out a linear regression to see if there is a significant relationship between sperm and time.

The test statistic is r-squared, **the regression coefficient.** This is a quantity describing the magnitude that the x variable explains variation in the y variable - it can range from -1 to +1.

```
# Linear regression of sperm number and time – note the order of the variables

LR1<-lm(spermnum ~ time, data=sperm)

## This gives us the intercept and slope, but no statistical test. For that we can display the linear regression results in a table using summary().

summary(LR1)
```

```
> summary(LR1) # test statistics

Call:
lm(formula = spermnum ~ time, data = sperm)

Residuals:
   Min     1Q Median     3Q    Max
-16003  -5382   -138   3217  32549

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6192.8     1605.3   3.858 0.000205 ***
time          3877.0      265.2  14.619  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7559 on 98 degrees of freedom
Multiple R-squared:  0.6856,    Adjusted R-squared:  0.6824
F-statistic: 213.7 on 1 and 98 DF,  p-value: < 2.2e-16
```

P-value for this explanatory variable (we can have more than one)

R-sq = how close the points fall to the regression line.

P-value for the whole model – this time it's the same as we only have one explanatory variable

A standard result statement would be something like…. We found that mating duration was a significant predictor of the number of spermium transferred during mating in yellow dung flies (R^2=0.682, P<0.001). You might also quote the intercept and slope of your model, we will cover these next week.

Now, add a trendline to your scatterplot using abline() and the model object (LR1)

Next week we will look at this in more detail and start using the model equation to predict values and also check the model validation. For now it is enough to understand

1) why you would use this over a correlation when examining relationships between variables
2) what statistics you should report in a results statement and
3) what the R-squared value means

For further explanation on R-squared check out the following link

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Correlation matrices and plots

Next week we will be moving onto forms of multiple regression. Often your datasets will contain multiple variables that may be correlated to one another and strong intercorrelations make it hard to identify which variables are having effects on your response variable. If you only have four or so variables, it is easy to carry out all the pairs of correlations. But it is a big problem when there are lots of variables (like more than 10 or so) – testing all the pairs would take ages!

So we use correlation plots and matrices.

Let's investigate the correlations between all the continuous variables in the UN dataset.

```
## First make a data object with just the columns 3-8 inclusive (all the continuous variables) and at the same time remove all the NAs

myUN<-na.omit(UN[,3:8])

### Then make scatterplots of each pair of variables.

pairs(myUN)
```

At a glance can you guess a numerical estimate of the correlation coefficient for any one of the cells in the matrix? Which pair is most highly correlated?

Now use the cor() function on your data object to print the correlation matrix. What are the two most highly correlated variables?

THE END – please make sure you have a clean **working** and **annotated** script before next weeks practical class.