L04: Regression

“All Programs have a desire to be useful.”

Master Control Program, Tron

Things we cover in this session

- Simple scatter plots
- Linear regression between two samples
- Estimating the explained variability of linear relationships

Things you need for this session

- W02-1: Reading CSV files

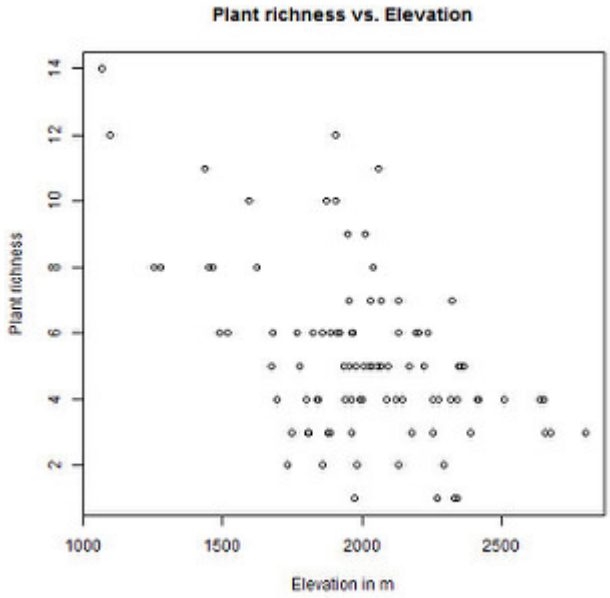
Things to take home from this session

At the end of this session you should be able to

- compute a regression between two variables using R
- derive information about the explanatory power of the linear model
- visualize two variables in a scatter plot and add a linear regression line
- visually analyze the validity of the explanatory power

Regression analysis

Scatterplots



Since linear models can only explain linear relationships and hence only make sense if such a relationship is feasible to explain the dependency between the dependent and one or more independent variables. Since eyeball analysis is very effective in estimating dependencies in the 2D space, plotting the dependent variable as a function of the independent variable should be a standard task in data analysis prior to fitting a model.

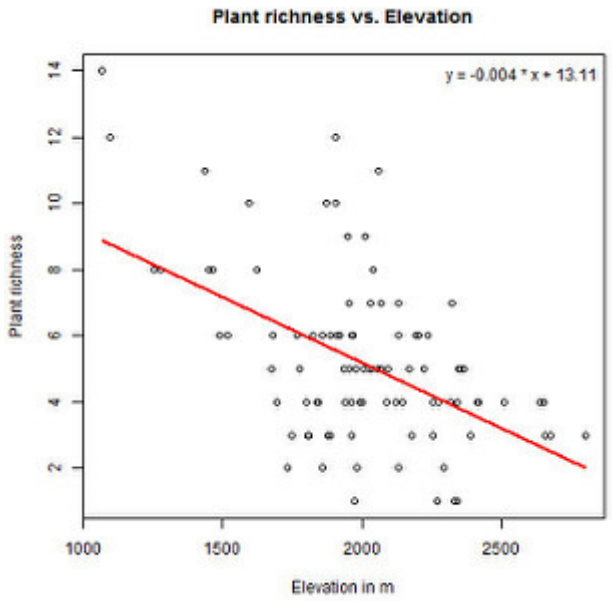
To visualize y as a function of x, a simple scatter plot can be plotted using R's plot() function:

```
''plot(x,y)''
```

Linear models

Linear regression models link a dependent variable (e.g. y) to one ore more independent variables (e.g. x) using a linear function of type

$$y = a x + b.$$



One standard function used in R for computing linear regressions is the lm() function which allows both simple and multiple linear regression. A minimalistic call of the function looks like

```
lm(y ~ x)
```

Goodness of fitted linear models

While the application of a linear model function to any distribution of x and y will result in a regression line, the goodness of the fitted model can be anything between a complete disaster and a perfect result.

To estimate this goodness, one usually first looks at the coefficient of determination, i.e. r square (R^2). In linear regression, (R^2) is generally given by

$$R^2 = 1 - \sum(y_i - \hat{y}_i)^2 / \sum(y_i - y_{mean})^2 =$$

1 - (Sum of squared differences of residuals) / (Sum of squared differences of all y)

with y_i as the observed value corresponding to x_i, $\hat{y}_i$ as the result of the fitted model for x_i, and y_{mean} as the mean value over all y_i in the sample.

Hence, R^2 shows a value range between 0 and 1 and expresses the explained variance as the difference between the unexplained variance of the model (caused by residuals different from 0) and the total variance in the y sample.

Please note that actual value of R^2 is irrelevant if it is not significant. The significance of a linear model is derived from testing the random chance that the actual slope (i.e. a) and actual intercept (i.e. b) of the linear model could also be 0.

To get R^2 and the p-value of a linear model use the summary() function applied to your linear model variable, e.g.:

```
summary(lm(y ~ x))
```

Time for practice

W04-1 Regressions

If you need more examples, have a look at C04-1 Linear model

Note on data used for illustrating analysis The analysis used for illustration on this site are based on data from a field survey of areas in the Fogo natural park in 2007 by K. Mauer. For more information, please refer to this report.