

Exercises for the course
“An introduction to R”

Exercise session Linear regression: Thursday, March 12 2015

Exercise 1: The built-in data set `trees` provides the measurement of diameter (`Girth`), height and volume of black cherry trees.

The measures are expressed in inches for `Girth`, feet for `Height` and cubic feet for `Volume`. For the purpose of this exercise we will focus on `Girth` and `Height`.

Write a function `convert` that converts the values into the metric system. Your function should take as argument the value to convert and the unit it was expressed in (inches or feet). Think about what your function should control for. Here is how it should work:

```
> convert(8.3,"inches")
[1] 0.21082
> convert(8.3,"feet")
[1] 2.52984
> convert(8.3,"foot")
Error in convert(8.3, "foot") : the unit was not "inches" or "feet"
```

Then use your function to convert the values and test if there is a significant correlation between the two variables.

Formulate your answer. Is there a difference when you use the non-converted values?

Write a linear model to predict the `Girth` of a tree knowing the other two variables. Is the effect of

the two variables significant? Check that your model is ok (use `plot()`).

What is the equation to be used to make a prediction using your model? Use it to predict the `Girth` of a tree of `Height` 82 feet and `Volume` 30 cubic feet. If you have time at the end of the session, try

to improve your model (transform variables ...).

Exercise 7.2: Set the seed to 1234 to get the same picture. Define two vectors `x <- seq(from=0,to=5,by=0.1)` and `noisyline <- 4*x + rnorm(length(x))`. Explain the variable `noisyline` with a linear model of the independent variable `x`. Use the command `lm()` for this. Denote the returned object as `regr`. What is the Anova table of this regression? Read off the intercept and the slope of the fitted line with `coef()`. Extract the p-values of the intercept and of the slope from `summary(regr)`. What is the fraction of the total variation in `noisyline` that is explained by the regression? This fraction is called 'r squared' and is printed by `summary(regr)` under 'Multiple R-squared'. Check that this value is equal to $(\text{cor}(x, \text{noisyline}))^2$. In order to visualize the linear model, plot `noisyline` against `x`. Add the regression line with `abline(regr)`. Next calculate confidence intervals of the fitted parameters with `confint()` applied to `regr`. Denote the object returned by this command as `cf`. Enter `cf` to view it. The two columns of `cf` define two lines. Add these two lines to the plot with line type "dashed". Your result should resemble the following figure.

