Trees Data Analysis

Milica Cudina

The data set `trees` is built-in. Let's take a look at it.

```
names(trees)
```

```
## [1] "Girth"  "Height" "Volume"
```

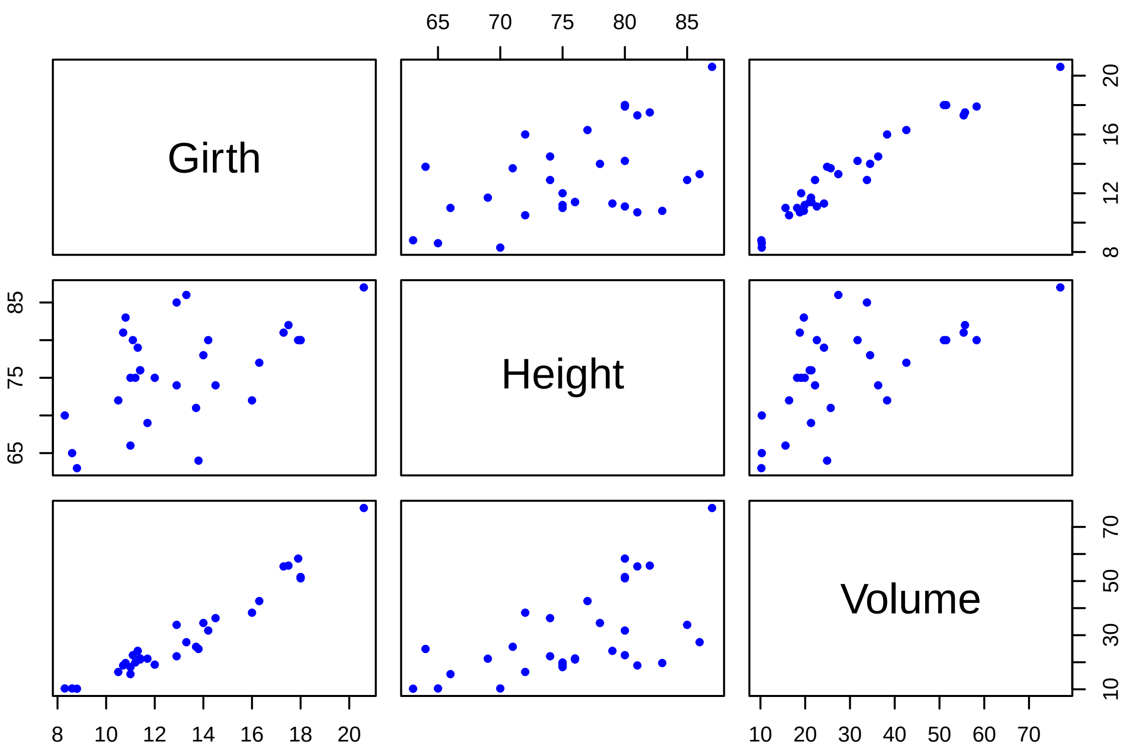
```
dim(trees)
```

```
## [1] 31  3
```

It should contain measurements of 31 cherry trees, namely, their `Girth`, `Height`, and `Volume`.

Again, we undertake a rudimentary exploratory data analysis. It's natural to be interested in pairwise interactions. So, we create an array of scatterplots with which we can visually assess the shape of the dependence and the correlations of each pair of variables.

```
plot(trees,
     col="blue", pch=20)
```

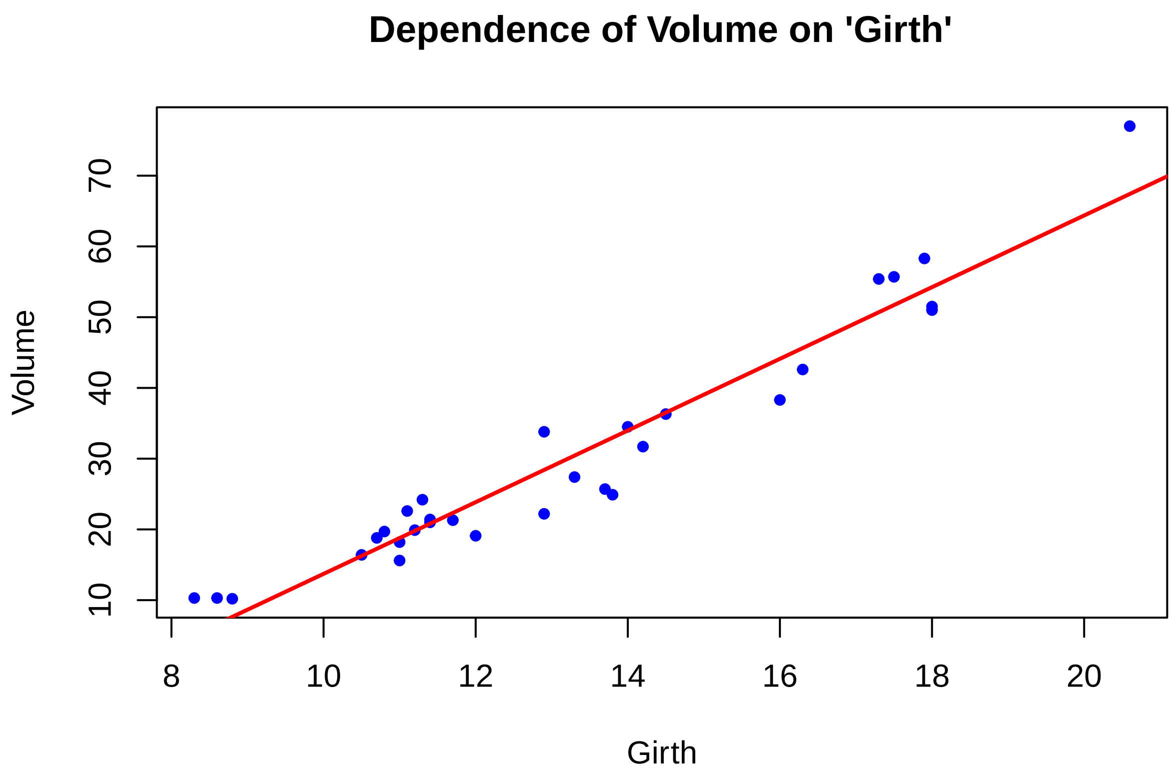


We might be interested in looking at, say, `Girth` as the explanatory and `Volume` as the response. This would be a simple linear regression.

```
lm.fit.g<-lm(Volume ~ Girth, data=trees)
summary(lm.fit.g)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 0.00000000000762 ***
## Girth         5.0659     0.2474   20.48          < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

```r
attach(trees)
plot(Girth, Volume,
     pch=20, col="blue",
     main="Dependence of Volume on 'Girth'")
abline(lm.fit.g, col="red", lwd=2)
```




```r
lm.fit.h<-lm(Volume ~ Height, data=trees)
summary(lm.fit.h)
```

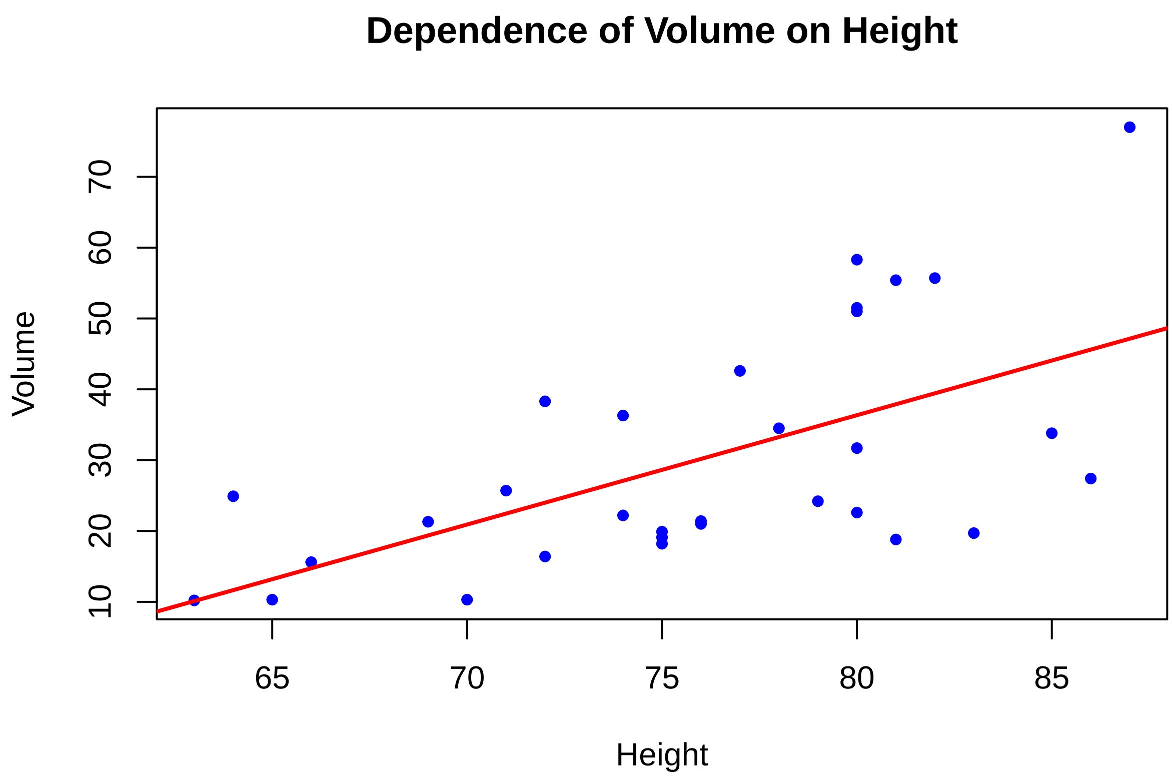
```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

```r
attach(trees)
```

```
## The following objects are masked from trees (pos = 3):
##
##     Girth, Height, Volume
```

```r
plot(Height, Volume,
     pch=20, col="blue",
     main="Dependence of Volume on Height")
abline(lm.fit.h, col="red", lwd=2)
```



Now, let's see what happens when we add `Height` as an additional explanatory variable, thus creating a multiple linear regression.

Let's compare the **coefficient of determination** R^2 for the above two fits.

For anyone who has ever seen trees, it's natural to suspect that there is a correlation between `Height` and `Girth`. Let's check

So, it might be a good idea to introduce the interaction term in our multiple linear regression.

We should take note, again, of any changes (improvements?) in the R^2 and/or the $p-$values.

Now, we can decide that we are reasonably happy, or we can go back to middle-school math and remember the formulae for volumes of cylinders. Which explanatory should we choose?