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"
```

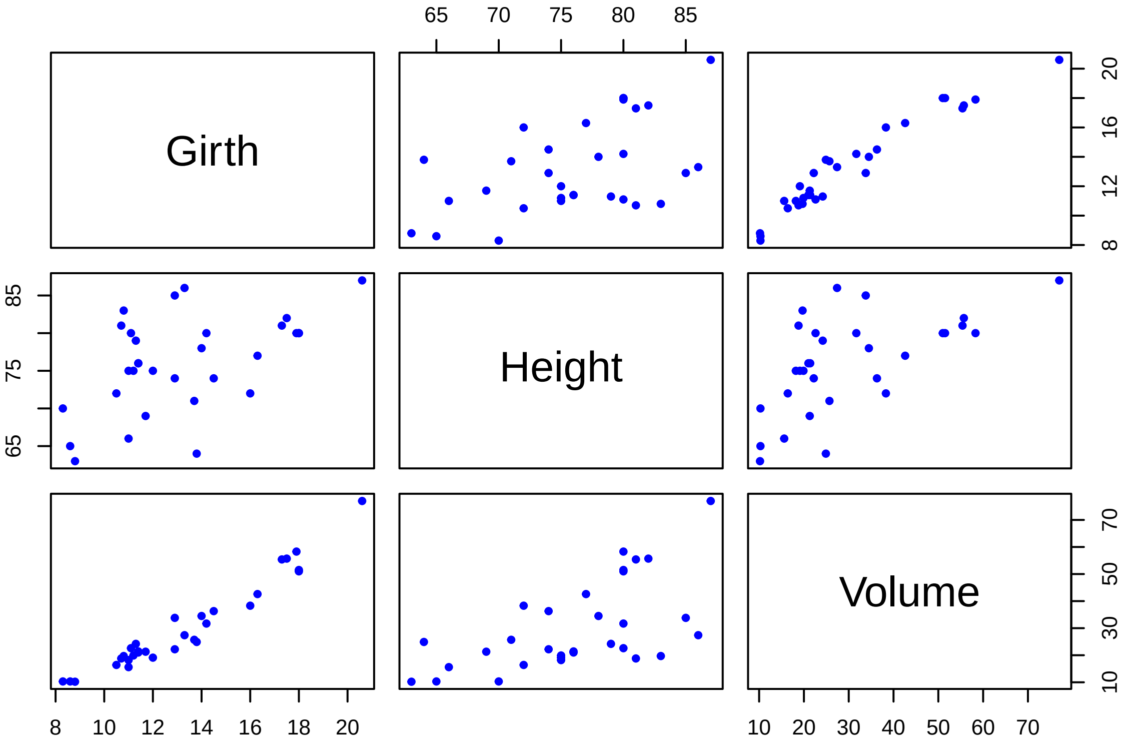
```
dim(trees)
```

```
## [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)
```



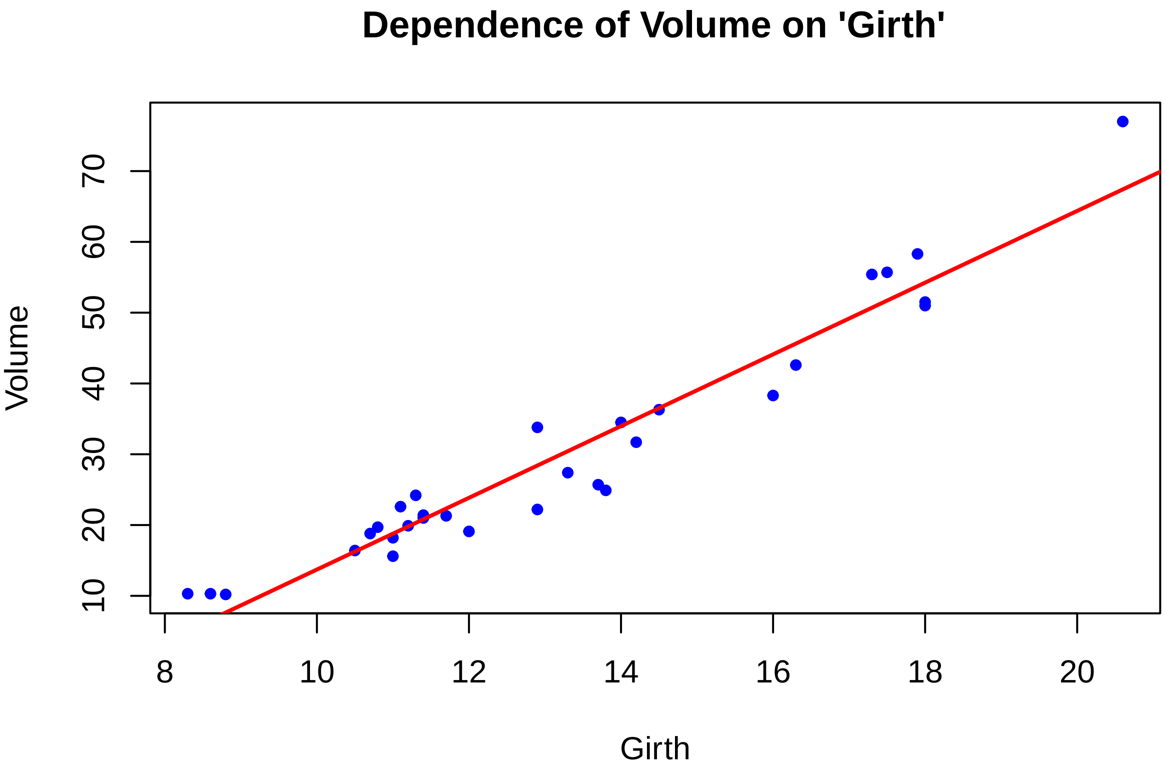
We might be interested in looking at, say, `Girth` as the explanatory and `Volume` as the response. This would be a simple linear regression.

```
attach(trees)
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
```

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

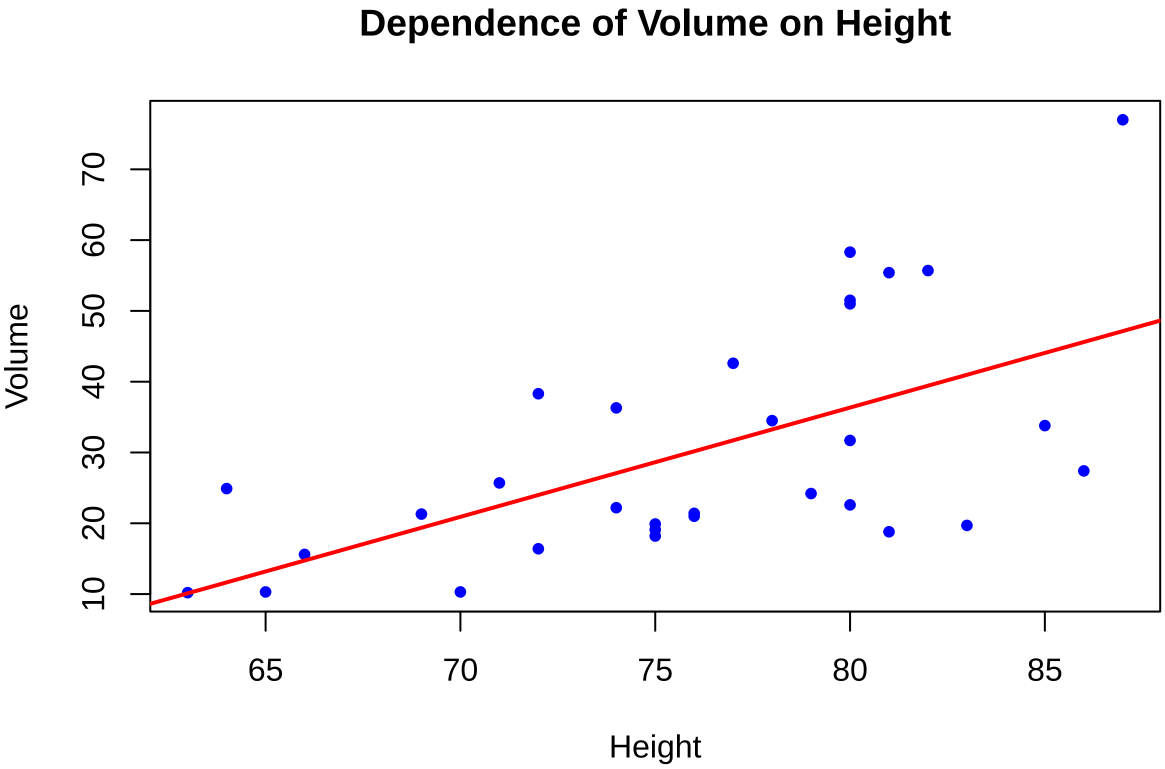


```
lm.fit.h<-lm(Volume ~ Height, data=trees)
summary(lm.fit.h)
```

```
##
```

```
## 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
```

```
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.

```
lm.fit.m<-lm(Volume ~ Girth + Height)
summary(lm.fit.m)
```

```
##
## Call:
```

```
## lm(formula = Volume ~ Girth + Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 0.000000275 ***
## Girth         4.7082     0.2643  17.816     < 2e-16 ***
## Height        0.3393     0.1302   2.607      0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

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

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

```
cor(Height, Girth)
```

```
## [1] 0.5192801
```

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

```
lm.fit.mi<-lm(Volume ~ Height*Girth)
summary(lm.fit.mi)
```

```
##
## Call:
## lm(formula = Volume ~ Height * Girth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5821 -1.0673  0.3026  1.5641  4.6649
##
## Coefficients:
##              Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  69.39632   23.83575   2.911    0.00713 **
## Height       -1.29708    0.30984  -4.186    0.00027 ***
## Girth        -5.85585    1.92134  -3.048    0.00511 **
## Height:Girth  0.13465    0.02438   5.524 0.00000748 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.709 on 27 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9728
## F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

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?

```r
lm.fit.geom<-lm(Volume ~ 0 + I(Girth^2):Height)
summary(lm.fit.geom)
```

```
##
## Call:
## lm(formula = Volume ~ 0 + I(Girth^2):Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6696 -1.0832 -0.3341  1.6045  4.2944
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## I(Girth^2):Height 0.00210810 0.00002722   77.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 30 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9949
## F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```

With the hierarchy principle:

```r
lm.fit.geom.h<-lm(Volume ~ I(Girth^2)*Height)
summary(lm.fit.geom.h)
```

```
##
## Call:
## lm(formula = Volume ~ I(Girth^2) * Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9548 -1.0501 -0.1482  1.7188  4.2102
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.0512340 11.9831440  -0.171   0.8654
## I(Girth^2)        -0.0043818  0.0705431  -0.062   0.9509
## Height             0.0267203  0.1565299   0.171   0.8657
## I(Girth^2):Height  0.0021616  0.0008789   2.459   0.0206 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.577 on 27 degrees of freedom
## Multiple R-squared:  0.9779, Adjusted R-squared:  0.9754
## F-statistic:   398 on 3 and 27 DF,  p-value: < 2.2e-16
```