W03D

Benjamin Huang

2024-09-12

#trees Model Summary

Below are histograms for each variable in the built-in data set trees.

##Diameter, Height and Volume for Black Cherry Trees

###Description

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is erroneously labelled Girth in the data. It is measured at 4 ft 6 in above the ground.

###Usage

trees

###Format

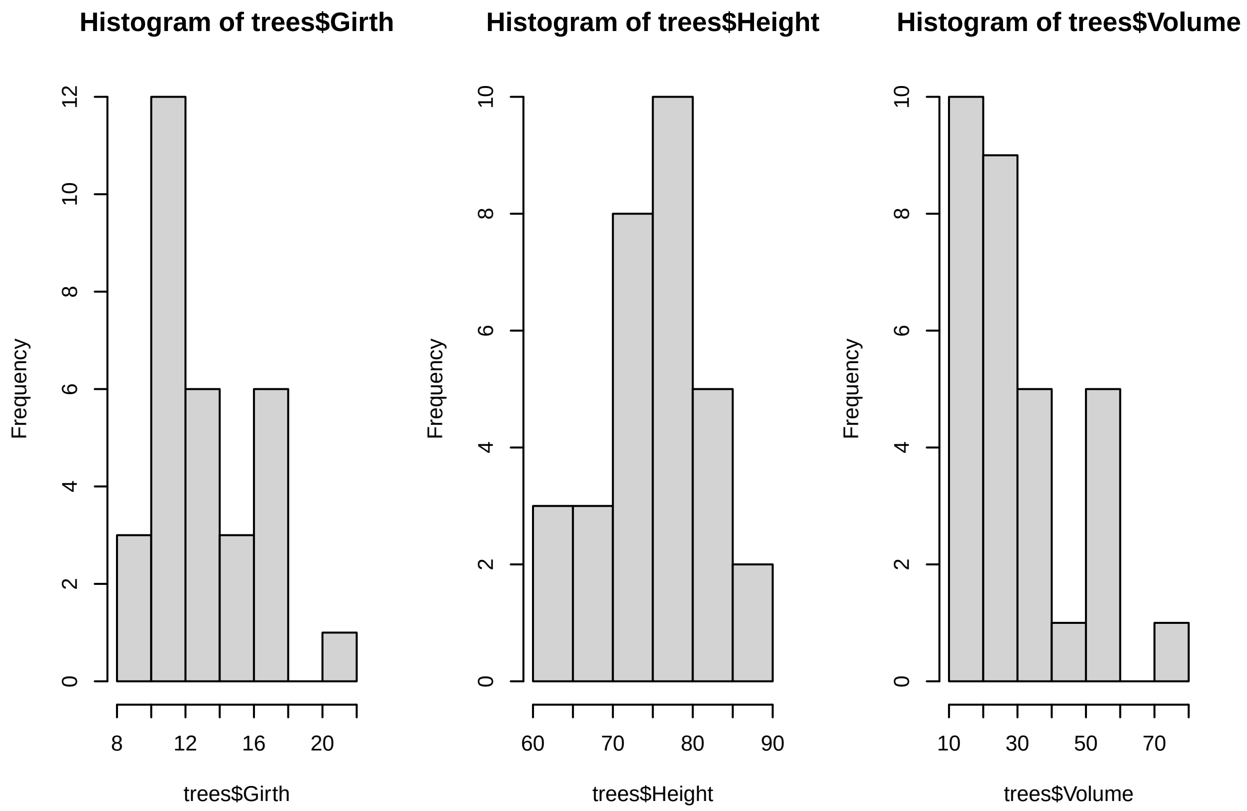
A data frame with 31 observations on 3 variables.

[,1] Girth; numeric; tree diameter (rather than girth, actually) in inches

[,2] Height; numeric; height in ft

[,3] Volume; numeric; volume of timber in cubic ft

```r
par(mfrow = c(1, 3))
hist(trees$Girth); hist(trees$Height); hist(trees$Volume)
```

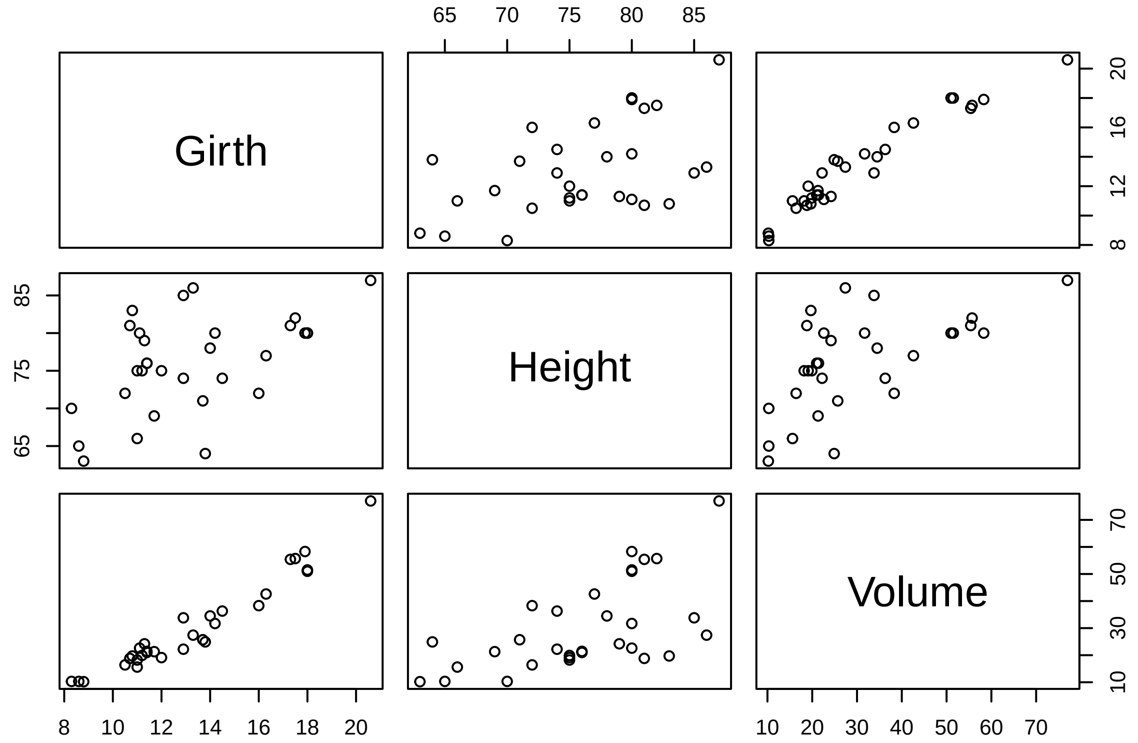


Q1. Which variable is closest to being normally distributed?

Answer: None of them

Next are plots of the columns against each other.

```r
plot(trees)
```



Q2. Which pair has the highest correlation?

Answer: Girth and Volume

Below is a summary for a regression using Volume as the y and Girth as the x.

Note the “t value and”Pr(>|t|)” columns. In lecture, we called these the t-ratio and the p-value. The null hypothesis is that the coefficient is 0.

Next the these columns are three asterisks, ***. In the row below, you can see the significance level that these coefficients exceed. Both exceed a 0.1% significance level, giving confidence that neither is zero.

```r
lm <- lm(Volume~Girth, data = trees)
summary(lm)
```

```
##
## 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 7.62e-12 ***
## 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
```

Q3. What is the equation of the fitted regression line?

Answer:

$Volume = 5.0659 \times Girth - 36.9435$

Q4. How is the "t value column calculated?

Answer:

$\frac{Coefficient\ Estimate}{Std.\ Error}$

Q5. Explain what "< 2e-16" means in the output above.

Answer:

This means the value (a p-value in this case), is less that 2*10^(-16) = 0.000…0002. A very small value, close to zero.

Let's compare this regression to a regression of Volume on Height, which is below.

Note that the Intercept loses significance in this regression.

```
lm2 <- lm(Volume~Height, data = trees)
summary(lm2)
```

```
##
## 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
```

Q6. What is the p-value corresponding to $H_o : \beta_1 = 0$? Answer: $p = 0.000378$

We can find the percentiles from a t-distribution using the qt(p, df) function.

The pt(t, df) function gives the corresponding probability for a t-value.

```
qt(0.005835 / 2, 29) #quantile of the t distribution
```

```
## [1] -2.97621
```

```
pt(-2.976, 29)*2 #probability level of the t distribution
```

```
## [1] 0.005838043
```

Q7. Explain why there is "/2" in qt and "*2" outside pt.

Answer: It's performing two-tail test.

Q8. Find the approximate and actual 95% confidence interval for β_1 for lm2.

```
#Answer
#Approximation
1.5433 - 2*0.3839; 1.5433 +2*0.3839
```

```
## [1] 0.7755
```

```
## [1] 2.3111
```

```
#Actual
1.5433 + qt(.025, 29)*0.3839; 1.5433 + qt(.975, 29)*0.3839
```

```
## [1] 0.7581363
```

```
## [1] 2.328464
```

This can also be done with the confint() function.

```
confint(lm2, "Height", level = 0.95) #confidence interval
```

```
##           2.5 %   97.5 %
## Height 0.758249 2.328451
```

Q9. Find the 99.9% confidnece interval for the intercept.

```
#Answer
confint(lm2, "(Intercept)", level = 0.999)
```

```
##                 0.05 % 99.95 %
## (Intercept) -194.2458 19.9986
```

To create a prediction interval, we need to define a new data set that contains value of Height.

This is done below.

```r
newdata = data.frame(Height = mean(trees$Height))
predict.lm(lm2, newdata = newdata, interval = "prediction", level = .95)
```

```
##        fit      lwr     upr
## 1 30.17097 2.332636 58.0093
```