Multiple Linear Regression (MLR) Tutorial

For this tutorial, we will learn how to fit multiple linear regression (MLR) in R. You will realize that fitting MLR is very similar to fitting SLR.

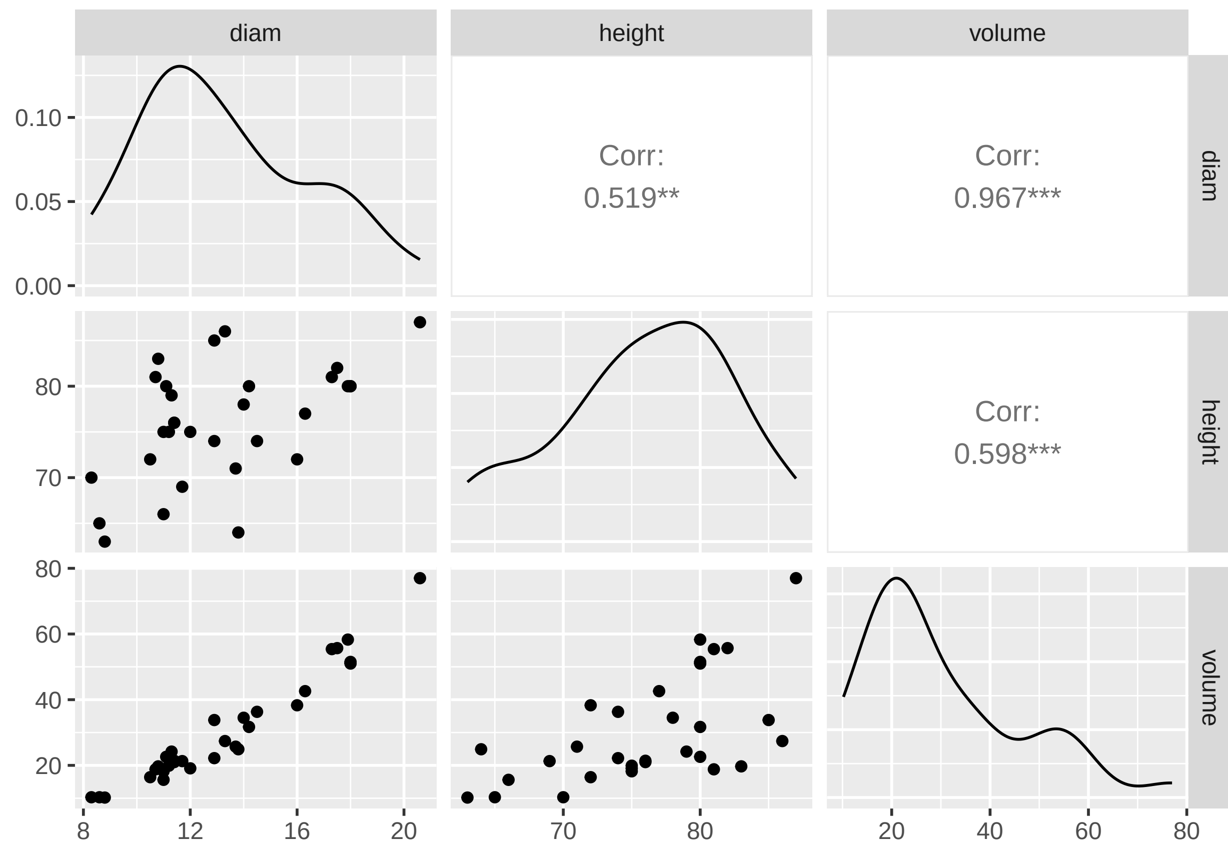
We will look at data regarding black cherry trees. The data, `cherry`, come from the `openintro` package. Researchers want to understand the relationship between the volume (in cubic feet) of these trees and their diameter (in inches, at 54 inches above ground) and height (in feet). Data come from 31 trees in the Allegheny National Forest, Pennsylvania.

```
library(openintro)
Data<-openintro::cherry
```

1. Scatterplot matrix

A scatterplot matrix is useful to create scatterplots involving more than two quantitative variables. We will use the `ggpairs()` function from the `GGally` package:

```
library(GGally)
##scatterplot matrix
GGally::ggpairs(Data)
```



A few pieces of information are presented in the output. Notice the output is displayed in a matrix format.

- The off-diagonal entries of the output give us the scatterplot and correlation between the corresponding pair of quantitative variables.
  - For example, look at the scatterplot in row 3, column 1 of the output. The corresponding label for the column is `diam` and the label for the row is `volume`. This informs us this is a scatterplot for `volume` on the vertical axis and `diam` on the horizontal axis. We see a strong positive linear association between these two variables.
  - The correlation between `volume` and `diam` is displayed in row 1, column 3. Again, notice the label for the column and row. This correlation is 0.967, which is high.
  - For practice, locate the scatterplot of `volume` and `height` and its corresponding correlation. Also locate the scatterplot of `diam` and `height` and its corresponding correlation.
- The diagonal entries display the density plot of the corresponding variable. For example, the third diagonal entry displays the density plot for `volume`. We can see that the distribution is somewhat right skewed as most trees have a volume between 10 and 40 cubic feet.

2. Fit MLR using `lm()`

To fit multiple linear regression (MLR)

```
##Fit MLR model, using + in between predictors
result<-lm(volume~diam+height, data=Data)
```

where we list the predictors after `~` with a `+` operator in between the predictors. Another way would be

```
result<-lm(volume~., data=Data)
```

The `.` after `~` informs the `lm()` function to use every column other than `volume` in the data frame as predictors.

Just like with simple linear regression (SLR) we can get relevant information using `summary()`:

```
summary(result)
```

```
##
## Call:
## lm(formula = volume ~ diam + height, data = Data)
##
## 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 2.75e-07 ***
## diam          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
```

- The estimated regression equation is $\hat{y} = -57.988 + 4.708 diam + 0.339 height$.
  - The estimated coefficient for `diam` is interpreted as: the predicted volume of a cherry tree increases by 4.708 cubic feet per inch increase in diameter, while holding height constant.

  - The estimated coefficient for `height` is interpreted as: the predicted volume of a cherry tree increases by 0.339 cubic feet per foot increase in height, while holding diameter constant.
- The R^2 is 0.948. About 94.8% of the variance in volume of cherry trees can be explained by their diameter and height.
- The residual standard error is 3.882. This estimates σ, the standard deviation of the error term.

3. Inference with MLR

Just like SLR, each coefficient is tested against a null hypothesis that $\beta_j = 0$ with a two-sided alternative. The test is significant for both coefficients, so we cannot drop either predictor from the model.

The ANOVA F statistic is 255, with a small p-value. So data supports the claim that our model is useful.

The confidence intervals for the coefficients can be found using `confint()`:

```
confint(result,level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) -75.68226247 -40.2930554
## diam          4.16683899   5.2494820
## height        0.07264863   0.6058538
```

The confidence interval for the mean response and the prediction interval for a new observation given a specific value of the predictors can also be found using `predict()`. For example, when the diameter is 10 inches and height is 80 feet:

```
newdata<-data.frame(diam=10, height=80)

predict(result, newdata, level=0.95,
        interval="confidence")
```

```
##        fit      lwr      upr
## 1 16.23404 13.36762 19.10047
```

```
predict(result, newdata, level=0.95,
        interval="prediction")
```

```
##        fit      lwr      upr
## 1 16.23404 7.781596 24.68649
```

You might realize by now we are using the same functions as we did in SLR.

Note: Obviously, all these calculations are performed and interpreted assuming the regression assumptions are met. Regression assumptions are checked in the same way as in SLR. On your own, as practice, assess the regression assumptions.