Université Grenoble Alpes

TP : Linear regression on trees

A sawmill wants to determine the height of a tree in order to predict its raw material requirements. They only have data on the circumference of the trees. You are going to create a model to predict the height of a tree based on its circumference. The sawmill has a data set which contains the characteristics for 201 Parisian spruce trees.

Let us import the file arbres-tot.csv which contains 201 data on Parisian spruces (source: https://opendata.paris.fr/explore/dataset/les-arbres/table/). The two variables of interests are:

- circumference of the tree (in cm),
- height: height of the tree (in m).

First, load the file and keep only the trees that have a height not equal to 0 (data cleaning)

```
myData=read.table(file="arbres-tot.csv",sep=";",skip=3,header=TRUE)
myData=myData[myData$X10!=0,]
```

1 Simple regression

We represent the point cloud (circ,height) :

```
circ=myData$X70
height=myData$X10

ggplot(myData,aes(x=circ,y=height))+ geom_point()+
xlab("circ")+
ylab("height")
```

We notice that the point cloud is not very far from a line, we launch the simple linear regression:

```
simple_reg <- lm(height~circ,data=myData)
```

R gives us a lot of informations :

```
names(simple_reg)
```

- coefficients : estimation of the parameters $\hat{\beta}_j$
- fitted.values : estimated values $\hat{y}_i$
- residuals : $e_i = y_i - \hat{y}_i$
- df.residual : number of freedom of the residuals $(n-2)$

You can have a look at the anova output:

```
anova(simple_reg)
```

We obtain the coefficient of determination as well as the parameters and their significance tests:

```
summary(simple_reg)
```

- We see that the coefficient of determination is about 0.65, which is not very high
- We reject the nullity of the parameters at the 5% test level.

It is possible to draw the regression line :

```
ggplot(myData,aes(x=circ,y=height))+ geom_point()+
  stat_smooth(method="lm",se=FALSE)+ xlab("circ")+
  ylab("height")
```

For hypotheses checking, we can have a look at the following graphs:

- Evaluation of the hypothesis of residuals independence

  The presence of an auto-correlation can be highlighted by a lag plot.

  ```
  acf(residuals(simple_reg))
  ```

- Evaluation of the hypothesis of residuals normality

  This hypothesis can be evaluated graphically using a QQplot. If the residuals are well distributed along the line shown on the plot, then the normality hypothesis is accepted. On the contrary, if they deviate from it, then the normality hypothesis is rejected.

  ```
  plot(simple_reg,2)
  ```

- Evaluation of the hypothesis of residuals homogeneity

  Here again, this hypothesis can be checked visually, by looking at the residuals, where no structure should appear. Various plots can be used. The "fitted" values correspond to the responses predicted by the model, for the observed values of the predictor variable.

  ```
  plot(simple_reg$residuals)
  plot(simple_reg,3)
  plot(simple_reg,1)
  ```

- Detection of outliers. The Cook's distance is a commonly used estimate of the influence of a data point. In a practical ordinary least squares analysis, It indicatew influential data points that are particularly worth checking for validity.

  ```
  plot(simple_reg,4)
  ```

Finally, we can use the model to predict what would be the height of a tree from its circumference, with the associated confidence interval. As for exemple:

```
predict(simple_reg data.frame(circ=10),interval="prediction")
```

The confidence intervals can hold assuming

- the linear model holds true
- either the errors in the regression are normally distributed
- or the number of observations is sufficiently large so that the actual distribution of the estimators can be approximated using the central limit theorem

2 Multivariate regression

Add a column to the sample. Name it circ_sqrt and fill it with the square root of the circumference of each tree.

```
myData$circ_sqrt <- sqrt(myData$X70)
```

Perform the multivariate linear regression of height on the basis of:

- of circumference ;

- of circ_sqrt.

```
multi_reg <- lm(height~circ+circ_sqrt,data=myData)
summary(multi_reg)
```

Analyse the significance of the parameters, and remove any non-significant parameters. The variable circ is not significant at the 5% test level, so we remove it:

```
multi_reg_2 <- lm(height~circ_sqrt,data=myData)
summary(multi_reg_2)
```

Give and interpret the coefficient of determination of the model finally retained.

We reject the nullity of the parameters at the 5% test level.

Write the obtained model.

It is possible to draw the regression "curve":

```
circ_pred <- seq(0,175,len=1000)
height_pred <- multi_reg_2$coefficients[1]+multi_reg_2$coefficients[2]*sqrt(circ_pred)
fct_reg <- data.frame(circ_pred=circ_pred,height_pred=height_pred)
ggplot()+
  geom_point(data=myData,aes(x=circ,y=height))
  + geom_line(data=fct_reg,aes(x=circ_pred,y=height_pred),col="blue")
  + stat_smooth(method="lm",se=FALSE)+
  xlab("circ")+
  ylab("height")
```

Try the model

$$heigth = \beta_1 circ + \beta_2 circ^2 + \beta_3$$

Draw the obtained model.

Compare the two previous model.