Computer Lab Practical n. 3

Today's task

Examples of use of multiple linear regression.

How to start

• Access the course web page on Moodle (/Biology). Download and save the three data files cherry.dat, temperature.dat and oliveoil.dat. (To download these, you may need to click on the right button of your mouse.)

How to predict the volume of wood

• Load the data of the `cherry.dat` file.

```
cherry <- read.table("~/cherry.dat", header = TRUE)
```

• Graphically inspect the data. You will see three variables measured on 31 black cherry trees: the girth (circumference) of the trunk (Girth, in inches), the height (Height, in feet) and the volume of wood obtained from the cut trees (Volume, in cubic feet). We want to use this data set to formulate an equation to predict the volume of wood, which is only known after cutting and machining, from the other two variables, which can be easily measured before cutting. Equations of this kind have various applications, from deciding how many and which trees to cut, to fixing the price of a forest.

• Let us start by exploring the relationships among the three variables.

```
plot(cherry)
```

Obtain the matrix of correlation coefficients with pairwise p-values:

```
library(RcmdrMisc)
rcorr.adjust(cherry, type="pearson")
```

Your comments?

• Let us try with a multiple linear regression model of the type

$$Volume = \beta_0 + \beta_1 Girth + \beta_2 Height + error$$

Fit a linear model:

```
fit1 <- lm(Volume ~ Girth + Height, data=cherry)
summary(fit1)
```

Your comments? Are the regression coefficients of the two explanatory variables statistically significant? What do they tell us?

- Check in your output for `Multiple R-squared:  0.948`.

What does this value tell us?

- The model has a good ability to predict. *Can we improve it?*

Your middle school math teacher probably taught you that solid volumes are proportional to (base area) $\times$ (height). The base area in turn is linked to the square of the circumference. Let us hence try with the following model:

$$\text{Volume} = \beta_0 + \beta_1 \text{Girth} + \beta_2 \text{Height} + \beta_3 (\text{Girth}^2 \times \text{Height}) + \text{error}$$

Let us create a new variable.

```
cherry$GGH <- cherry$Girth^2 * cherry$Height
```

Inspect the data to verify that the new variable was correctly defined. Now, re-fit the model. The model "formula" now must be $\text{Volume} \sim \text{Girth} + \text{Height} + \text{GGH}$.

```
fit2 <- lm(Volume ~ Girth + Height + GGH, data=cherry)
summary(fit2)
```

Was it useful to include the product into the model? Can we simplify the model?

- *Question: would log transforming the data help us to reproduce one of the math formulae seen in middle school for volumes?*

Temperature, age, heart rate and sex

The temperature.dat data has four variables: temperatura = temperature, anni = age, battito = heart rate, sesso = sex. The purpose of the analysis is to assess whether temperature varies with heart rate, sex and age.

- Load the data of temperature.dat file you downloaded.

```
temp_data <- read.table("temperature.dat", header = TRUE,
                        stringsAsFactors=TRUE)
```

- Produce a numerical summary of the data.

```
summary(temp_data)
```

What do you notice?

• Obtain the histogram of temperature.

```
hist(temp_data $temperatura, xlab="temperature",
     main="Histogram of temperature", breaks=10,xlim = c(35,40))
```

• To assess whether/how each single variable impacts on temperature try the following:
— for sex, compare the boxplots of temperatura using sesso as grouping variable.

```
boxplot(temperatura ~ sesso, data=temp_data,col= c("lightpink","lightblue"))
```

The statistical significance can be assessed with Student's t test or with Wilcoxon's two-sample test.

```
t.test(temperatura~sesso, alternative="two.sided", conf.level=.95,
       var.equal=FALSE, data=temp_data)
wilcox.test(temperatura ~ sesso, alternative="two.sided", data=temp_data)
```

— for the two "numerical" variables, that is, battito and anni, check the pairs plots

```
pairs(temp_data[,-1],col="blue")
```

Calculate the corresponding correlation coefficients and assess their statistical significance following the same steps than specified in the previous example.
Your comments?

• Let us now try and see what happens if we account for all three variables simultaneously by using a multiple linear regression model of the type

$$temperatura = \beta_0 + \beta_1 battito + \beta_2 anni + \beta_3 \overline{sesso} + errore$$

where $\overline{sesso} = 1$ for male and $\overline{sesso} = 0$ for female subjects.

```
fit_mul <- lm(temperatura ~ ., data=temp_data)
summary(fit_mul)
```

Your comments?

A taste of residual analysis

• It is common practice (besides being good practice) to have a look at the "residuals" of the model to identify possible errors in the specification of the model itself. The "residuals" are nothing but the "estimates" of the error components of the model. For the latter case study,

$$residuals = temperature - \left(\hat{\beta}_0 + \hat{\beta}_1 battito + \hat{\beta}_2 anni + \hat{\beta}_3 sesso\right)$$

where $\hat{\beta}$'s ("the β hat's") are the estimates of the regression coefficients we calculated at the previous step.
• Save the residuals in our dataset by creating a new variable called residuals.mul

```
temp_data$residuals.mul <- residuals(fit_mul)
head(temp_data)
```

• Standard checks consist in verifying whether there are outlying observations (a boxplot of the residuals can be useful) and in drawing the residuals against the explanatory variables to identify any possible relationship not yet explained by the model. In the latter case, there seems to be nothing worthwhile reporting. For example, try and plot the residuals against heart rate (battito).

```
boxplot(temp_data$residuals_mul)
plot(residuals.mul ~ battito, data = temp_data)
```

The figure doesn't suggest any relationship between the residuals and the heart rate. We can conclude that the relationship between temperature and heart rate is linear given that "the whole relationship" is explained by the linear model we fitted.

• To see a different case where the residuals suggest a change, do the following:

– estimate the linear model

$$temperatura = \beta_0 + \beta_1 battito + errore$$

that is, the model with the only covariate battito.

```
fit_simple <- lm(temperatura ~ battito, data = temp_data)
```

– let R compute the residuals of this model and check whether they are uncorrelated with age (anni) by using both graphical and numerical inspection.

```
temp_data$residuals_simple <- residuals(fit_simple)
plot(residuals_simple ~ anni, data = temp_data)
cor.test(temp_data$anni, temp_data$residuals_simple, alternative="two.sided",
         method="pearson")
```

Pay attention not to use the old residuals which are uncorrelated with anni.

What do you notice?

Olive oil

• Load the data of the `oliveoil.dat` file.

```
oliveoil <- read.table(" /oliveoil.dat", header = TRUE)
```

• Inspect the data. There are 16 samples of olive oil whose acidity was measured together with further 4 physico-chemical variables (peroxide, K232, K270, DK). We want to assess whether these four variables influence acidity.

• Start off by graphically representing the data using pairs plots ("scatterplot matrix"), estimate the correlations present among the variables and verify their statistical significance.

- Then formulate and fit a multiple linear regression model with “acidity” as response variable and the remaining variables as covariates.
- Comment upon the results.