Multivariate statististics

Marie-Luce Taupin
marieluce.taupin@univ-evry.fr

Laboratoire de Mathématiques et Modélisation d'Evry (LaMME)
Université d'Evry val d'Essonne

2025-2026

1. Chapter 1: Introduction and examples
2. Chapter 2: Linear model
3. Chapter 3: Generalized linear model
4. Chapter 4: Regularized regressions
5. Chapter 5: Multiple testing

1 Chapter 1: Introduction and examples

2 Chapter 2: Linear model

3 Chapter 3: Generalized linear model

4 Chapter 4: Regularized regressions

5 Chapter 5: Multiple testing

Organisation

- 12 h de CM (ML Taupin)
- 12h de TD en R (ML Taupin)
- Installe *R* then *Rstudio*
  - https://cran.rstudio.com/
  - https://posit.co/download/rstudio-desktop/

  Those software have to be installed on your computer before the next session.

Examples

Example 1

The data consists of $n = 1429$ measurements of circumference-height pairs, obtained from a plot of 6-year-old eucalyptus trees (rotation age before cutting).
Purpose: to predict the height of a tree based on its Goal: predict the height of a tree based on its circumference.
Tool: find a relationship that links circumference to height in order to predict the height of a tree based on its circumference.

Example 2

Study of corn field yields based on the type of fertilizer used

Example 3

Study of blood pressure according to patient age

Modèle linéaire

Example 4

On patients with heart problems, the speed of blood flow (using the Doppler effect) Y in the coronary arteries was measured. We want to study the effect of two quantitative variables on this speed, namely cholesterol level T and weight P. The following data is available : for each patient i, $i = 1, \cdots, 20$, we measure their weight p_i, cholesterol level t_i, and blood flow velocity y_i.

Example 5

The effect of three malaria treatments is compared by measuring the time to parasite clearance in symptomatic patients randomly assigned to three groups.

Linear model

Exemple 6 : Prostate Cancer

The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radicalprostatectomy. The variables are log cancer volume (*lcavol*), log prostate weight (*lweight*), *age* , log of the amount of benign prostatic hyperplasia (*lbph*), seminal vesicle invasion (*svi*), log of capsular penetration (*lcp*), Gleason score (*gleason*), and percent of Gleason scores 4 or 5 (*pgg*45).

$$Y = \text{level of prostate-specific antigen},$$

and

$$\mathbf{X} = (lcavol, lweight, age, lbph, lcp, pgg45)^T.$$

Generalized linear models (GLM)

Example 7: Evans dataset

Follow-up of a cohort of 609 men over a period of 7 years. The target variable Y is "onset or non-onset of coronary heart disease." The explanatory variables are: **chd**, a dichotomous variable taking the value 1 if coronary heart disease is present, 0 otherwise; **cat**, a dichotomous variable indicating whether the catecholamine level is high (1) or not (0), **age** a continuous variable expressed in years, **chol** a continuous variable defining cholesterol level, **smk** a dichotomous variable indicating whether the subject is a smoker (1) or has ever smoked (0), **ecg** a dichotomous variable indicating the presence of an abnormal electrocardiogram (1) or not (0), **dbp** a continuous variable indicating diastolic blood pressure, **sbp** a continuous variable indicating systolic blood pressure, **hpt** a dichotomous variable indicating the presence (1) or absence (0) of high blood pressure, **ch** is cat × hpt, **cc** is cat × chl.

Generalized linear models

Example 8: birthwt data set

Associated with Low Infant Birth Weight: 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986. low (indicator of birth weight less than 2.5 kg), age (mother's age in years), lwt (mother's weight in pounds at last menstrual period), race mother's race (1 = white, 2 = black, 3 = other), smoke (smoking status during pregnancy), ptl (number of previous premature labours), ht (history of hypertension), ui (presence of uterine irritability), ftv (number of physician visits during the first trimester), et bwt (birth weight in grams).

Example 9

Standardize fertility in a country as a function of socio-economic indicators (proportion of males involved in agriculture as occupation, education beyond primary school for draftees, infant mortality, ...).

Example 10

Gain in weight of rats fed on four different diets, distinguished by amount of protein (low and high) and by source of protein (beef and cereal).

Examples

Example 11

Number of asthma-related visits to an Emergency Room as a function of air pollution indices.

Example 12

Cancer diagnosis as a function of exposure to particular chemicals.

Example 13

LTNP status of HIV-infected patients as a function of multiple SNPs genotypes.

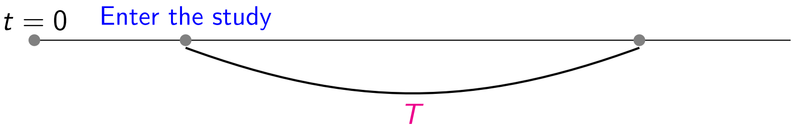
Survival data analysis

Example 14

Study of the survival time of patients with lung cancer based on age, gender, and smoking status. The variable studied is called duration T.
Survival time: the time that elapses from an initial moment (start of treatment, diagnosis, etc.) until the occurrence of a final event of interest (death of the patient, relapse, remission, cure, appearance of a tumor, etc.). This duration is not generally observed in its entirety in all individuals.

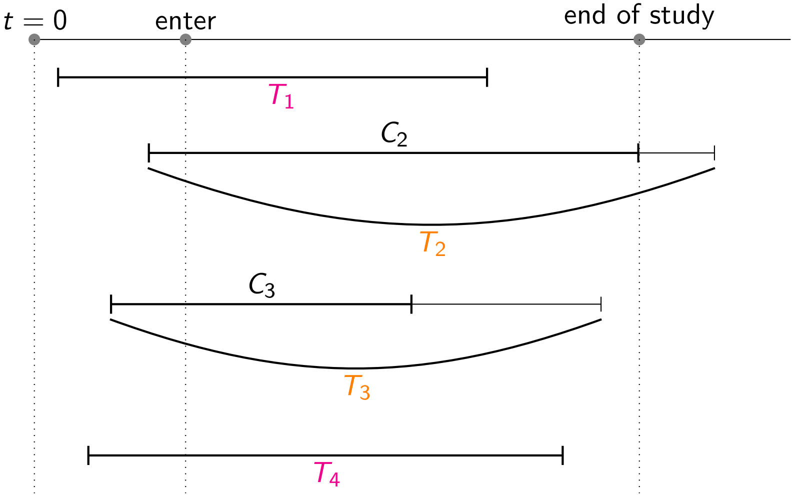
Right censoring

Let T be the duration between the enter in the study and the event of interest.



- $t = 0$: beginning of the study.
- Enter the study.
- Event of interest
- T: duration between the enter in the study and the event of interest.

Right censoring: example with 4 patients



Right censoring

In this example,
- for individual 1, the variable of interest T_1 is observed.
- for individual 2, the variable of interest T_2 is not observed. Only the variable C_2 is observed! This is a case of "administrative censoring".
- for individual 3, the variable of interest T_3 is not observed (individual lost to follow-up. Only the variable C_3 is observed!
- for individual 4, the variable of interest T_4 is observed.

Right-censoring
- Right-censoring occurs when the event of interest is not always observed.
- For censored individuals, we observe a duration shorter than the variable of interest.

Right censoring on a simulated dataset

We observe $\min(T_i, C_i, \delta_i)$

n° Patient	True duration	Censuring time	Observation	$\delta = 0$: censured
1	3	4	3	1
2	1	1	1	1
3	4 *not observed*	3	3	0
4	4 *not observed*	1	1	0
5	2 *not observed*	1	1	0
6	3	5	3	1
7	4 *not observed*	3	3	0
8	3 *not observed*	2	2	0
9	1	3	1	1
10	4 *not observed*	2	2	0

Comparing the means

Mean of observed durations: $2{=}(3+1+3+1+1+3+3+2+1+2)/10$

Mean of non censored data: $2{=}(3+1+3+1)/4$

Mean of true data $2.9{=}(3+1+4+4+2+3+4+3+1+4)/10$

Right censoring

For the individual i with $i = 1, ..., n$, let

- T_i: the true survival time
- C_i: le censored time
- δ_i: the indicator of censoring
- $\mathbf{Z}_i$: the vector of covariates

(age, sex, treatment, smoking status,...)
For each individul, we observe

$$\begin{cases} (T_i, 1, \mathbf{Z}_i) & \text{if } C_i \geq T_i \quad \text{not censored} \\ (C_i, 0, \mathbf{Z}_i) & \text{if } C_i \leq T_i \quad \text{censored} \ . \end{cases}$$

Censored datas

- Censored data cannot be considered to be the true durations (bias).
- Censored data cannot be removed (bias).

Summary

- Examples 2, 5 (Anova) and 14 are not considered in this course.
- Other examples would be considered in this course.

Variables

In all examples, we have :

- A variable of interest Y (the outcome).
- Several explanatory variables $X_i^{(1)}, \ldots, X^{(p)}$ which explain the variations of Y.
- We are looking for a relation like $Y = F_{\theta^*}(X)$ or more specifically $\mathbb{E}(Y) = F_{\theta^*}(X)$.

Objectives

Main objectives

- To describe the relationship between Y and the $X^{(j)}$'s.
- To test the significance of the relationship.
- To predict Y for new X_i values.

Illustrative example 1: eucalyptus

Eucalyptus

The data set represents $n = 1429$ mesures couples measures circumference-height, measurements obtained on a plot of 6-year-old eucalyptus trees (rotation age before cutting).
The goal is to predict the height of an eucalyptus from its circumference.
In order to do this, we have a dataset of $n = 1429$ trees with 4 measurements: the height ‘ht‘, the circumference ‘circ‘ as well as two other variables, the geographic zone ‘bloc‘ and the tree origin ‘clone‘, that we are not going to use in the beginning. Goal: predict the height of a tree based on its circumference.
Tool: find (estimate) the relationship between circumference and height in order to predict the height of a tree based on its circumference.

Illustrative example 2: birthwt

Birthwt

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986. This data frame contains the following columns: - low: indicator of birth weight less than 2.5 kg;
-age: mother's age in years;
-lwt: mother's weight in pounds at last menstrual period;
- race: mother's race (1 = white, 2 = black, 3 = other); - smoke: smoking status during pregnancy;
- ptl: number of previous premature labours;
- ht: history of hypertension;
- ui: presence of uterine irritability;
- ftv: number of physician visits during the first trimester;
- bwt: birth weight in grams.
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

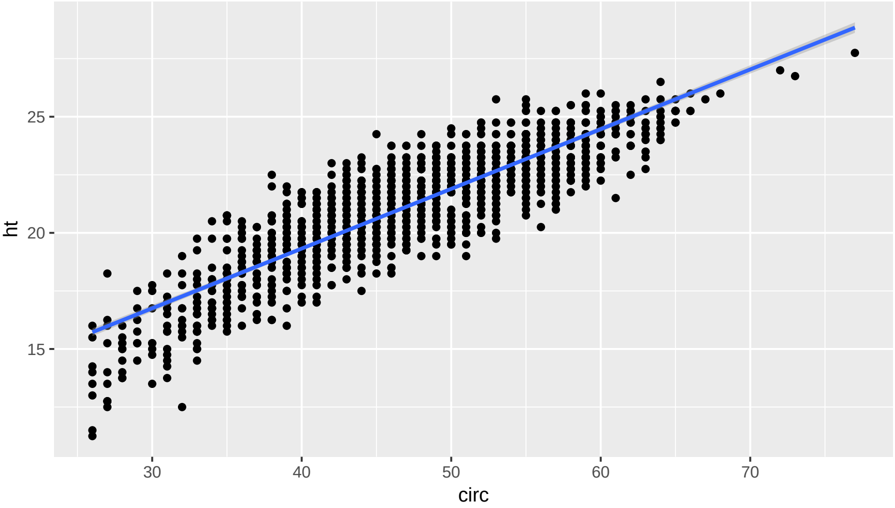
Illustrative example 3: Lung cancer

Lung cancer Lung cancer incidence in four Danish cities 1968–1971

This data set contains counts of incident lung cancer cases and population size in four neighbouring Danish cities by age group. A data frame with 24 observations on the following 4 variables: - city: a factor with levels Fredericia, Horsens, Kolding, and Vejle;
- age a factor with levels 40-54, 55-59, 60-64, 65-69, 70-74, and 75+;
- pop: a numeric vector, number of inhabitants;
- cases: a numeric vector, number of lung cancer cases.
These data were "at the center of public interest in Denmark in 1974", according to Erling Andersen's paper. The city of Fredericia has a substantial petrochemical industry in the harbour area.
Source E.B. Andersen (1977), Multiplicative Poisson models with unequal cell rates, Scandinavian Journal of Statistics, 4:153–158.

Application: eucalyptus

First graph:



Application: eucalyptus

Here for each tree $Y_i = ht_i$, $i = 1, \ldots, n$.
Consider the model $ht_i = \beta_0^* + \beta_1^* * circ_i + \varepsilon_i$.

```
##
## Call:
## lm(formula = ht ~ circ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683,Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application: eucalyptus

Comments

- The estimated regression line is given by

$$y = 9.03 + 0.25 * x.$$

- the coefficient corresponding to the slope is positive and equals 0.25.
- for a circumference of $40cm$, the predicted height is

$$ht_{pred} = 9.03 + 0.25 * 40 = 19m.$$

- for the three i, the prediction error is given by

$$ht_i - ht_{pred,i} = ht_i - (9.03 + 0.25 * circ_i).$$

Application: eucalyptus

Other (linear) models

One can try other models like

$$ht_i = \beta_0^* + \beta_1^* * \sqrt{circ_i} + \varepsilon_i,$$

or

$$ht_i = \beta_0^* + \beta_1^* * circ_i^2 + \varepsilon_i,$$

or

$$ht_i = \beta_0^* + \beta_1^* * circ_i + \beta_2^* * circ_i^2 + \varepsilon_i.$$

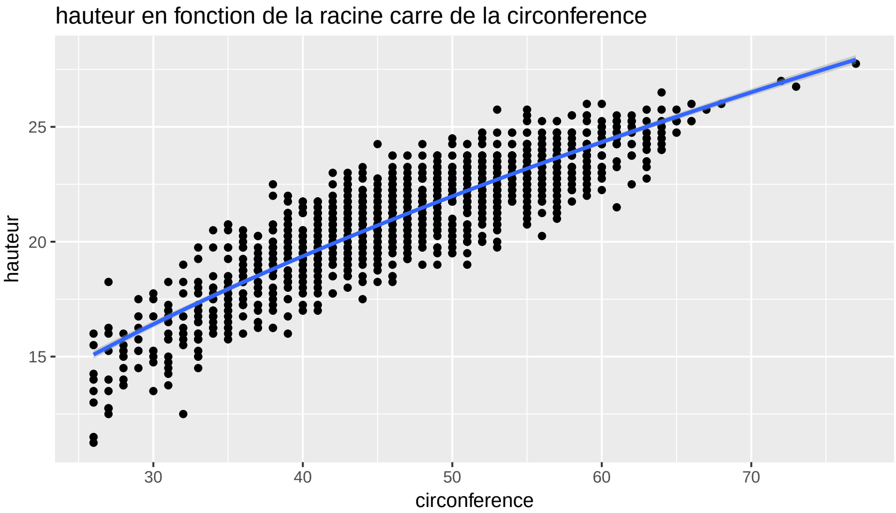
Those models are also linear models function of the square root of the circumference or function of the square of the circumference.

Application: Eucalyptus

Consider the model $ht_i = \beta_0^* + \beta_1^* * \sqrt{circ_i} + \varepsilon_i$.

```
##
## Call:
## lm(formula = ht ~ I(sqrt(circ)), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5360 -0.7249  0.0265  0.7813  3.6904
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.73036    0.33600  -8.126 9.51e-16 ***
## I(sqrt(circ))  3.49424    0.04883  71.560  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.163 on 1427 degrees of freedom
## Multiple R-squared:  0.7821,Adjusted R-squared:  0.7819
## F-statistic:  5121 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application: Eucalyptus - function of the square root

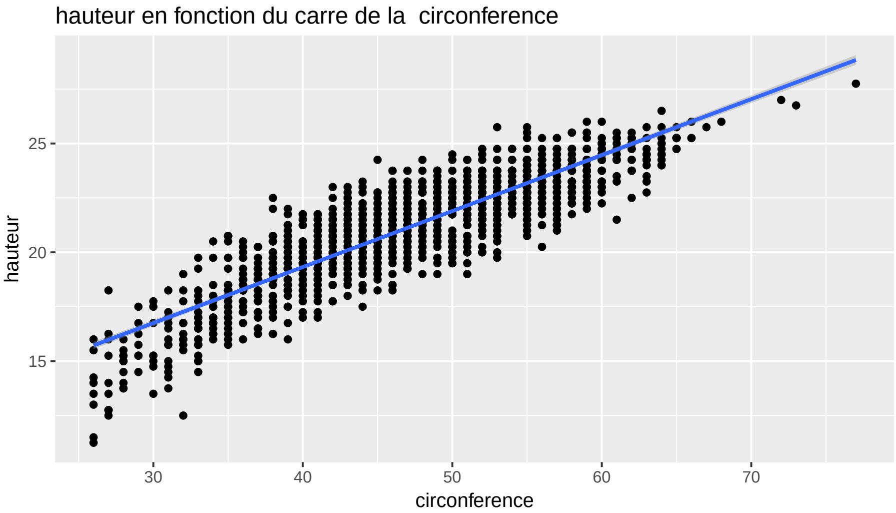


Application: Eucalyptus - function of the square

On considère le modèle linéaire $ht_i = \beta_0^* + \beta_1^* * circ_i^2 + \varepsilon_i$.

```
##
## Call:
## lm(formula = ht ~ I(circ^2), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5942 -0.8184  0.0551  0.8449  3.8080
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.504e+01  1.056e-01  142.46   <2e-16 ***
## I(circ^2)   2.667e-03  4.315e-05   61.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.299 on 1427 degrees of freedom
## Multiple R-squared:  0.7281,Adjusted R-squared:  0.7279
## F-statistic:  3821 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Application: Eucalyptus - function of the square

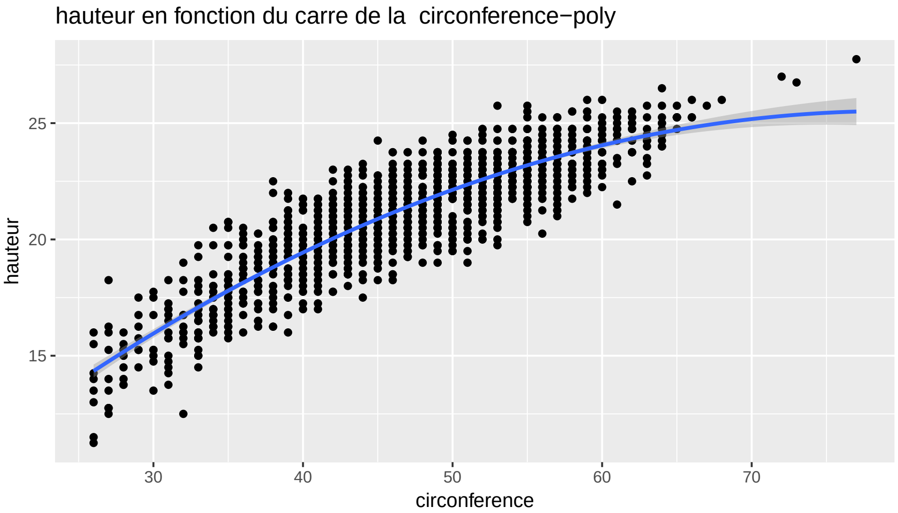


Application: eucalyptus

Consider the model $ht_i = \beta_0^* + \beta_1^* * circ_i + \beta_2^* * circ_i^2 + \varepsilon_i$.

```
reg3un<-lm(ht~circ+I(circ^2),data=euca)
summary(reg3un)

##
## Call:
## lm(formula = ht ~ circ + I(circ^2), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2140 -0.6947  0.0360  0.7732  3.6970
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8028038  0.7012035   1.145    0.252
## circ         0.6227415  0.0303984  20.486   <2e-16 ***
## I(circ^2)   -0.0039224  0.0003239 -12.110   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.142 on 1426 degrees of freedom
## Multiple R-squared:  0.7899,Adjusted R-squared:  0.7896
## F-statistic:  2681 on 2 and 1426 DF,  p-value: < 2.2e-16
```

Application: Eucalyptus - polynomial of degree 2



Application

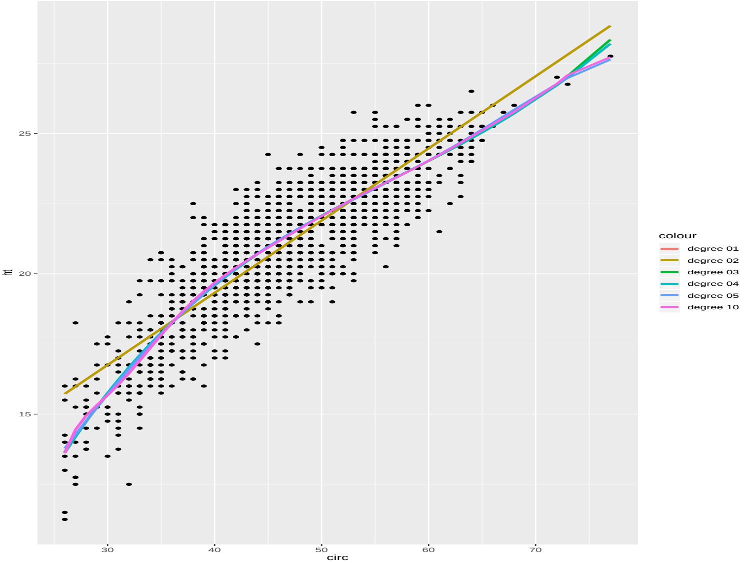
Eucalyptus- autre modèles: polynomials of degree 1, 2, 3, 4, 5 et 10

```
reg_poly1 <- lm(ht ~ poly(circ, 1, raw = TRUE), data = eucalyptus)
pred_poly1 <- predict.lm(reg_poly1, eucalyptus)
pred_poly1_df=data.frame(x=circ, pred=pred_poly1)


reg_poly2 <- lm(ht ~ poly(circ, 2, raw = TRUE), data = eucalyptus)
pred_poly2 <- predict.lm(reg_poly2, eucalyptus)
pred_poly2_df=data.frame(x=circ, pred=pred_poly2)

reg_poly3 <- lm(ht ~ poly(circ, 3, raw = TRUE), data = eucalyptus)
pred_poly3 <- predict.lm(reg_poly3, eucalyptus)
pred_poly3_df=data.frame(x=circ, pred=pred_poly3)


reg_poly4 <- lm(ht ~ poly(circ, 4, raw = TRUE), data = eucalyptus)
pred_poly4 <- predict.lm(reg_poly4, eucalyptus)
pred_poly4_df=data.frame(x=circ, pred=pred_poly4)


reg_poly5 <- lm(ht ~ poly(circ, 5, raw = TRUE), data = eucalyptus)
pred_poly5 <- predict.lm(reg_poly5, eucalyptus)
pred_poly5_df=data.frame(x=circ, pred=pred_poly5)

reg_poly10 <- lm(ht ~ poly(circ, 10, raw = TRUE), data = eucalyptus)
pred_poly10 <- predict.lm(reg_poly10, eucalyptus)
pred_poly10_df=data.frame(x=circ, pred=pred_poly10)
```

Application: eucalyptus

```
ggplot(data = eucalyptus, aes(x = circ, y = ht)) + geom_point(data = eucalyptus, aes(y = ht)) +
  geom_line(data = pred_poly1_df, aes(y = pred_poly1, color = "degree 01"),size=1.2) +
  geom_line(data = pred_poly2_df, aes(y = pred_poly2, color = "degree 02"),size=1.2) +
  geom_line(data = pred_poly3_df, aes(y = pred_poly3, color = "degree 03"),size=1.2) +
  geom_line(data = pred_poly4_df, aes(y = pred_poly4, color = "degree 04"),size=1.2) +
  geom_line(data = pred_poly5_df, aes(y = pred_poly5, color = "degree 05"),size=1.2) +
  geom_line(data = pred_poly10_df, aes(y = pred_poly10, color = "degree 10"),size=1.2)
```

Eucalyptus - polynomials of degree 1, 2, 3, 4, 5 et 10



Eucalyptus- polynomials: prediction errors

```
emp_err1 <- mean((ht - pred_poly1)^2)
emp_err2 <- mean((ht - pred_poly2)^2)
emp_err3 <- mean((ht - pred_poly3)^2)
emp_err4 <- mean((ht - pred_poly4)^2)
emp_err5 <- mean((ht - pred_poly5)^2)
emp_err10 <- mean((ht- pred_poly10)^2)
data.frame( emp_err = c(emp_err1, emp_err2, emp_err3, emp_err4, emp_err5, emp_err10))
writeLines(strwrap(paste("Degre 01:", emp_err1)))
writeLines(strwrap(paste("Degre 02:", emp_err2)))
writeLines(strwrap(paste("Degre 03:", emp_err3)))
writeLines(strwrap(paste("Degre 04:", emp_err4)))
writeLines(strwrap(paste("Degre 05:", emp_err5)))
writeLines(strwrap(paste("Degre 10:", emp_err10)))

Degre 01: 1.436028
Degre 02: 1.302107
Degre 03: 1.2752659
Degre 04: 1.2752150
Degre 05: 1.2740634
Degre 10: 1.2713997
```

Eucalyptus- polynomials: prediction errors

Denote by $\widehat{P}_n$ Ithe polynomial with coefficients estimated on the whole sample then the prediction error

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{P}_n(X_i))^2$$

is a biaised estimator of the true prediction error.

$$\mathbb{E}[Y_i - \widehat{P}_n(X_i))^2|\widehat{P}_n] \neq \mathbb{E}[Y_{n+1} - \widehat{P}_n(X_{n+1}))^2|\widehat{P}_n],$$

since $\widehat{P}_n$ depends on $(X_i, Y_i)_{i=1,\cdots,n}$.
The prediction error has to be estimated on an independant sample.
We split our sample on two sample called euca.test et euca.train. On the first we estimate $\widehat{P}_n$, and on the second set we estimate the prediction error.

Eucalyptus- polynomials: prediction errors by Cross-Validation

The prediction errors are given by

```
> writeLines(strwrap(paste("Degre 01:", errCV1)))
Degre 01: 1.44080379837949
> writeLines(strwrap(paste("Degre 02:", errCV2)))
Degre 02: 1.30841677587969
> writeLines(strwrap(paste("Degre 03:", errCV3)))
Degre 03: 1.28157598221082
> writeLines(strwrap(paste("Degre 04:", errCV4)))
Degre 04: 1.28377932201013
> writeLines(strwrap(paste("Degre 05:", errCV5)))
Degre 05: 1.28556022084155
> writeLines(strwrap(paste("Degre 10:", errCV10)))
Degre 10: 2.97506661125687
```

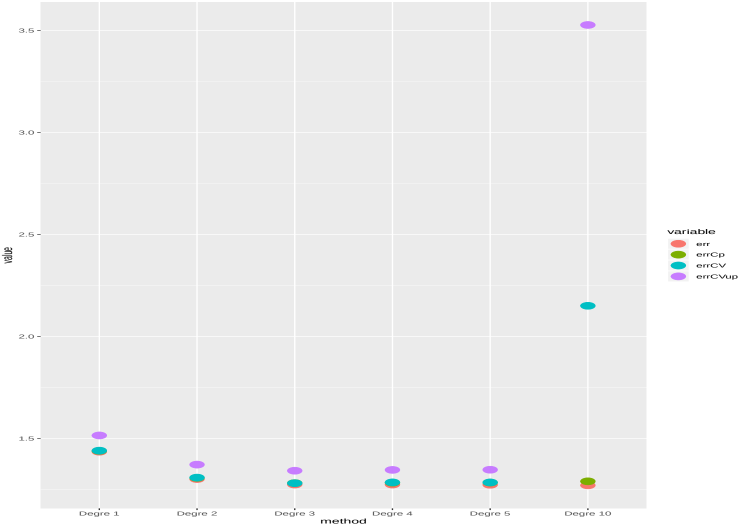
Based on estimated prediction errors, the best model seems to be the polynomial of order 3.

Eucalyptus- polynomials: other criterions

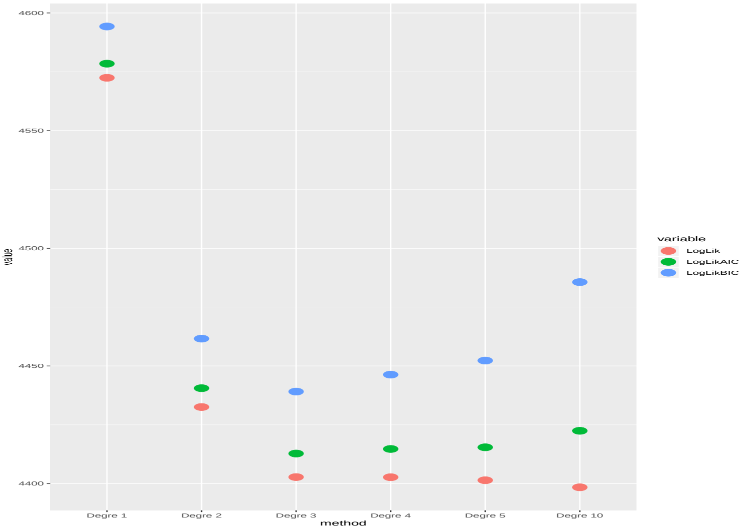
```
method      err     errCp    errCV   errCVup   LogLik LogLikAIC LogLikBIC
Degre 1 1.436028 1.440048 1.441353 1.515764 4572.454  4578.454  4594.248
Degre 2 1.302107 1.307575 1.309203 1.372948 4432.560  4440.560  4461.619
Degre 3 1.275266 1.282405 1.281815 1.342947 4402.794  4412.794  4439.118
Degre 4 1.275215 1.284139 1.285715 1.346916 4402.737  4414.737  4446.326
Degre 5 1.274063 1.284762 1.285519 1.347748 4401.446  4415.446  4452.299
Degre 10 1.271400 1.290973 2.151152 3.527410 4398.456  4422.456  4485.632
```

1. err= prediction error based on the whole sample (biaised)
2. errCp= prediction error based on Cp de Mallows
3. errCV= prediction error based on Cross-validation
4. LogLik= $-2*$log-likelihood(model)
5. LogLikAIC $=-2*$log-likelihood(model) $+2$ dim(model) (R)
6. LogLikBIC$=-2*$log-likelihood(modèle) $+2$ dim(modèle)*(log(n)).

Eucalyptus: errors comparison



Eucalyptus: AIC and BIC comparison



Eucalyptus- polynomials : comparison of nested models

```
anova(reg_poly1,reg_poly2)
anova(reg_poly1,reg_poly3)
anova(reg_poly2,reg_poly3)
anova(reg_poly3,reg_poly4)
anova(reg_poly3,reg_poly5)
anova(reg_poly4,reg_poly5)
anova(reg_poly3,reg_poly10)
```

Eucalyptus- polynomials: comparison of nested models

```
> anova(reg_poly1,reg_poly2)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 1, raw = TRUE)
Model 2: ht ~ poly(circ, 2, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1427 2052.1
2   1426 1860.7  1    191.37 146.66 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(reg_poly1,reg_poly3)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 1, raw = TRUE)
Model 2: ht ~ poly(circ, 3, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1427 2052.1
2   1425 1822.3  2    229.73 89.819 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Eucalyptus- polynomials : comparison of nested models

```
> anova(reg_poly2,reg_poly3)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 2, raw = TRUE)
Model 2: ht ~ poly(circ, 3, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1426 1860.7
2   1425 1822.3  1    38.357 29.993 5.118e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(reg_poly3,reg_poly4)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 3, raw = TRUE)
Model 2: ht ~ poly(circ, 4, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1425 1822.3
2   1424 1822.3  1  0.072734 0.0568 0.8116
```

Eucalyptus- polynomials : comparison of nested models

```
> anova(reg_poly3,reg_poly5)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 3, raw = TRUE)
Model 2: ht ~ poly(circ, 5, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1425 1822.3
2   1423 1820.6  2    1.7183 0.6715 0.5111
> anova(reg_poly4,reg_poly5)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 4, raw = TRUE)
Model 2: ht ~ poly(circ, 5, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1424 1822.3
2   1423 1820.6  1    1.6456 1.2862 0.2569
```

Eucalyptus - polynomials: comparison of nested models

```
> anova(reg_poly3,reg_poly10)
Analysis of Variance Table

Model 1: ht ~ poly(circ, 3, raw = TRUE)
Model 2: ht ~ poly(circ, 10, raw = TRUE)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1   1425 1822.3
2   1418 1816.8  7    5.5247 0.616 0.7431
```

Application - Eucalyptus

Questions

Q1 Description of models and analyze R output?

Q2 Which coefficients are significative?

Q3 Which models to choose? Which model for best prediction? Which is the smallest model?

Application - Eucalyptus

Eucalyptus - statistic answers

Q1 Parameter estimation

Q2 Statistical tests

Q3 Models comparison and variable selection

Second illustrative example: birthwt

Birthwt

The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986. This data frame contains the following columns: - low: indicator of birth weight less than 2.5 kg;
-age: mother's age in years;
-lwt: mother's weight in pounds at last menstrual period;
- race: mother's race ($1 =$ white, $2 =$ black, $3 =$ other); - smoke: smoking status during pregnancy;
- ptl: number of previous premature labours;
- ht: history of hypertension;
- ui: presence of uterine irritability;
- ftv: number of physician visits during the first trimester;
- bwt: birth weight in grams.
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.