**Real word data case:**

In this practical assignment, we will examine a real-world case involving fetal growth data from the 18th to the 42nd week of pregnancy. The main objective is to gain hands-on experience in all phases of statistical modeling, encompassing preliminary data analysis, data cleansing, modeling, and validation. You will be provided with a dataset comprising 6000 samples for analysis. All the scripts used in the analysis, along with their accompanying explanations, must be submitted as part of the designated task. Additionally, a detailed PDF report covering descriptive analysis, modeling, and validation is expected. The work will be presented and defended publicly for evaluation.

**Data:**

The database was simulated using the "syn" function of the "synthpop" R library with information from 9113 births. From this dataset, a random sample size of 6000 was chosen, which includes maternal variables such as age, height, weight, and parity. Additionally, information on the sex and fetus weight was collected both at delivery and through ultrasound scans. Data preprocessing is required, and consideration must be given to remove repetitions and handle outlier data.

**Objective:**

The main objective is to build a model that can predict the percentile fetus weight at different gestational ages. To achieve this goal, a predictive model for fetus weight growth and confidence intervals, dependent on gestational age, must be provided. Percentiles 1%, 3%, 10%, 90%, 97% and 99% growth charts should be estimated.

**Model Building:**

Various techniques, including linear regression, generalized additive models, or mixed linear models, can be explored for predicting fetus weight percentile 50. It is also essential to estimate the variance in order to construct confidence intervals.

Example

Assuming a polynomial regression model, the prediction of percentile 50, depending on the gestational week x, can be expressed as follows:

Percentile50<-function(x){exp(0.3590+0.04959\*7\*x-0.00007852\*(7\*x)^2)}

Confidence percentile intervals can be constructed using the coefficient of variation of Pearson of weight, which is calculated as the standard deviation divided by the mean.

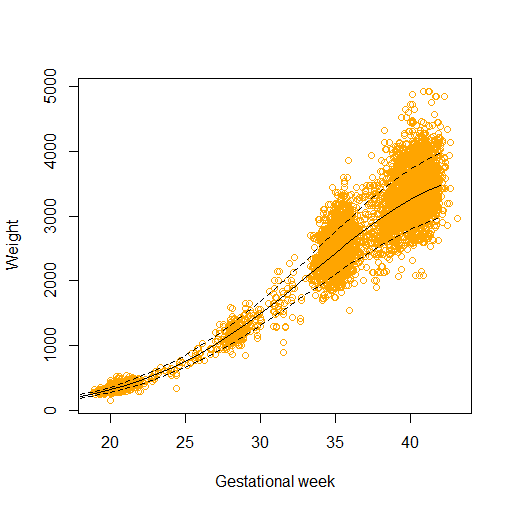
For Percentile 10 and 90, the function can be defined as:

Percentile10<-function(x){exp(0.3590+0.04959\*7\*x-0.00007852\*(7\*x)^2)\*(1+{if(x>=37) 0.1112588 else if(x<37&x>=34) 0.1055413 else if (x<34&x>22) 0.09876804 else 0.09810124} \*qnorm(0.10))}

Percentile90<-function(x){exp(0.3590+0.04959\*7\*x-0.00007852\*(7\*x)^2)\*(1+{if(x>=37) 0.1112588 else if(x<37&x>=34) 0.1055413 else if (x<34&x>22) 0.09876804 else 0.09810124} \*qnorm(0.90))}

In this example, coefficient of variation of Pearson is not assumed as constant, but it was calculated as constant stratified in the intervals of gestational age weeks (18,22), (22,34), (34,37), (37,42). Hadlock growth models considered coefficient of variation of Pearson as constant by gestational age week. In this work, you can assume as constant, estimated by week or considered it as depending on gestational age.

In the Figure below can be seen the percentile 10, 50, and 90 growth curves plotted over the simulated data, showing their adjustment on data.



**Validation**

To assess the performance of the percentile growth curves, it is necessary to estimate the rate of weights that fall above or below a percentile curve. For instance, it was expected that a 90th percentile would result in a rate of 10% of fetus weights above the curve.

**Evaluation**

The task will be evaluated based on: Data preliminary analysis (20%), Weight growth model (40%), Percentile confidence intervals (20%), Validation (20%).