STATISTICAL MODEL TO PREDICT NEWBORN WEIGHT

1. Loading the "neonati.csv" dataset and visualize it:

Filter

	Anni.madre	N.gravidanze	Fumatrici	Gestazione	Peso	Lunghezza	Cranio	Tipo.parto	Ospedale	Sesso
1	26	0	0	42	3380	490	325	Nat	osp3	M
2	21	2	0	39	3150	490	345	Nat	osp1	F
3	34	3	0	38	3640	500	375	Nat	osp2	M
4	28	1	0	41	3690	515	365	Nat	osp2	M
5	20	0	0	38	3700	480	335	Nat	osp3	F
6	32	0	0	40	3200	495	340	Nat	osp2	F
7	26	1	0	39	3100	480	345	Nat	osp3	F
8	25	0	0	40	3580	510	349	Nat	osp1	M

2. Describe the dataset composition and the variable types:

The dataset is composed of 2500 observations of 10 variables. Their classification is the following:

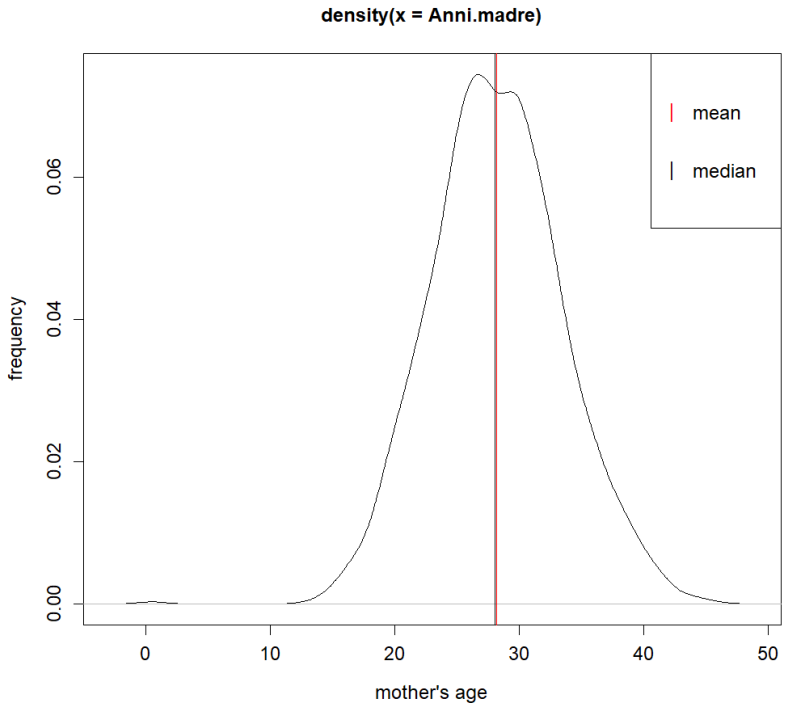
- ***"Anni.madre"***: quantitative continuous on ratio scale;
- ***"N.gravidanze"***: quantitative discrete on ratio scale;
- ***"Fumatrici"***: qualitative nominal, codified in numbers;
- ***"Gestazione"***: quantitative continuous on ratio scale;
- ***"Peso"***: quantitative continuous on ratio scale;
- ***"Lunghezza"***: quantitative continuous on ratio scale;
- ***"Cranio"***: quantitative continuous on ratio scale;
- ***"Tipo.parto"***: qualitative nominal;
- ***"Ospedale"***: qualitative nominal;
- ***"Sesso"***: qualitative nominal;

The *summary()* command gives back a first look at the position indexes of the dataset's variables. R software considers three of them as qualitative nominal ("*Tipo.parto*", "*Ospedale*", and "*Sesso*") and the other seven as qualitative ones. However, "Fumatrici" is a qualitative variable numerically codified (dummy variable) as already specified.

```
> newborn = read.csv("neonati.csv", sep = ",")
> View(newborn)
> summary(newborn)
   Anni.madre     N.gravidanze       Fumatrici        Gestazione         Peso
 Min.   : 0.00   Min.   : 0.0000   Min.   :0.0000   Min.   :25.00   Min.   : 830
 1st Qu.:25.00   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:38.00   1st Qu.:2990
 Median :28.00   Median : 1.0000   Median :0.0000   Median :39.00   Median :3300
 Mean   :28.16   Mean   : 0.9812   Mean   :0.0416   Mean   :38.98   Mean   :3284
 3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.:40.00   3rd Qu.:3620
 Max.   :46.00   Max.   :12.0000   Max.   :1.0000   Max.   :43.00   Max.   :4930
   Lunghezza        Cranio     Tipo.parto         Ospedale           Sesso
 Min.   :310.0   Min.   :235   Length:2500        Length:2500        Length:2500
 1st Qu.:480.0   1st Qu.:330   Class :character   Class :character   Class :character
 Median :500.0   Median :340   Mode  :character   Mode  :character   Mode  :character
 Mean   :494.7   Mean   :340
 3rd Qu.:510.0   3rd Qu.:350
 Max.   :565.0   Max.   :390
```

3. Brief Descriptive analysis of "neonati.csv" dataset variables:

- "*Anni.madre*": plotting the density distribution of the variable shows how there are some outliers (2) close to the origin of the axis. For biological reasons, these are probably incorrect values (it is impossible to become a mother before puberty). Moreover, the graph and the shape indexes, display how the variable distribution is leptokurtic, and slightly asymmetric positively.




```
> summary(Anni.madre)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   25.00   28.00   28.16   32.00   46.00
> IQR(Anni.madre)
[1] 7
> var(Anni.madre)
[1] 27.81063
> sd(Anni.madre)
[1] 5.273578
> coeff.vaiation(Anni.madre)
[1] 18.72454
> skewness(Anni.madre)
[1] 0.0428115
> kurtosis(Anni.madre)-3
[1] 0.3804165
```

Looking at the two incorrect values, it is plausible they result from a wrong typing of age data. Indeed, the remaining values related to the observations 1152 and 1380 are completely credible. Correcting them is an option but it would be hard to guess the actual age values. Thus, these two observations will still be considered unless other issues emerge in the analysis subsequent steps.

```
> newborn[Anni.madre<10,]
     Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto Ospedale
1152          1             1         0         41 3250       490    350        Nat      osp2
1380          0             0         0         39 3060       490    330        Nat      osp3
     Sesso
1152     F
1380     M
```

"*N.gravidanze*": many women included in the study had no previous pregnancies (43%), and around 10 % had already had more than two childbirths.

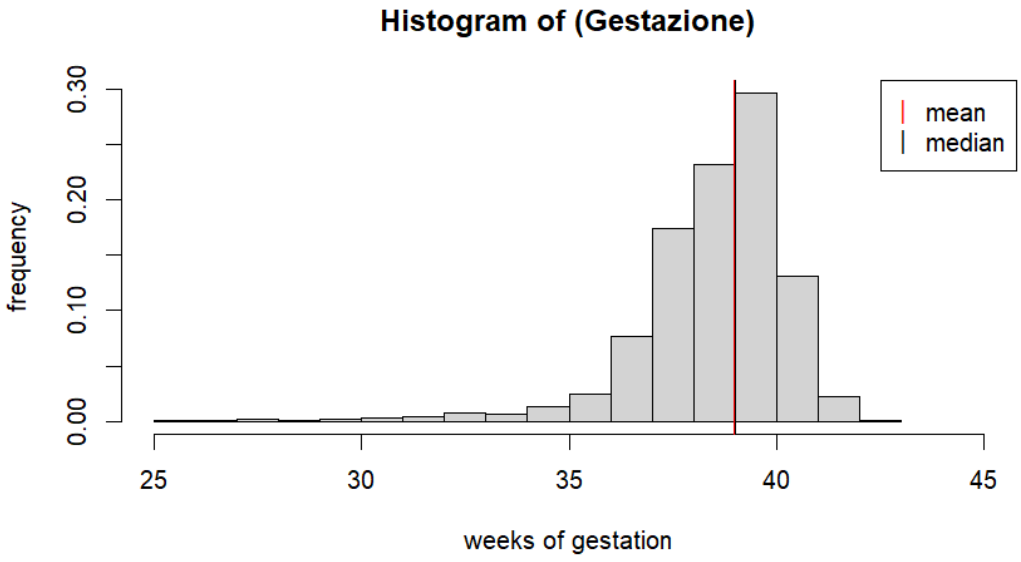
	ni	fi	Ni	Fi
0	1096	0.4384	1096	0.4384
1	818	0.3272	1914	0.7656
2	340	0.1360	2254	0.9016
3	150	0.0600	2404	0.9616
4	48	0.0192	2452	0.9808
5	21	0.0084	2473	0.9892
6	11	0.0044	2484	0.9936
7	1	0.0004	2485	0.9940
8	8	0.0032	2493	0.9972
9	2	0.0008	2495	0.9980
10	3	0.0012	2498	0.9992
11	1	0.0004	2499	0.9996
12	1	0.0004	2500	1.0000

- "*Fumatrici*": the frequency table shows how the majority of women who carried a pregnancy are non-smokers ("0", 95,84 %), whereas only a small percentage smoked during maternity ("1", 4,16 %).

	ni	fi
0	2396	0.9584
1	104	0.0416

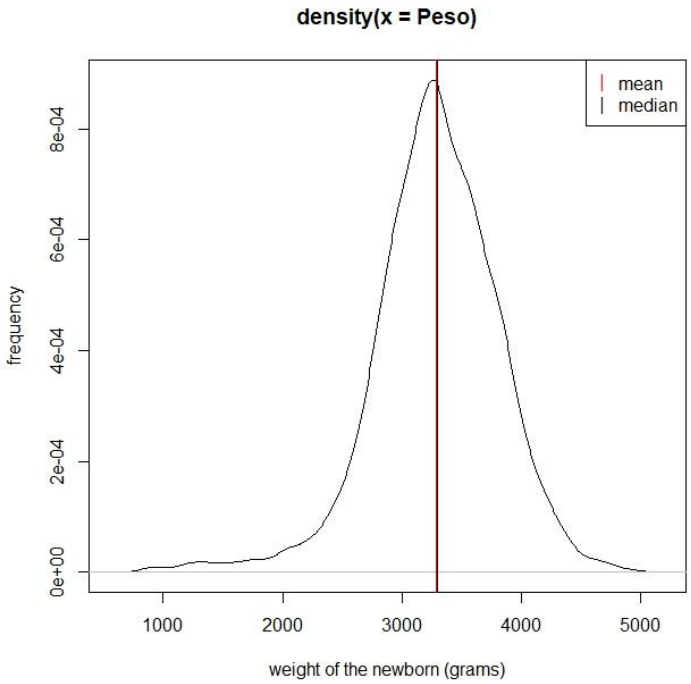
- "*Gestazione*": its distribution does not follow a normal-like shape as shown by the below graph and confirmed by the shape indexes.

```
> summary(Gestazione)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00   38.00   39.00   38.98   40.00   43.00
> IQR(Gestazione)
[1] 2
> var(Gestazione)
[1] 3.491813
> sd(Gestazione)
[1] 1.868639
> coeff.vaiation(Gestazione)
[1] 4.793792
> skewness(Gestazione)
[1] -2.065313
> kurtosis(Gestazione)-3
[1] 8.25815
```



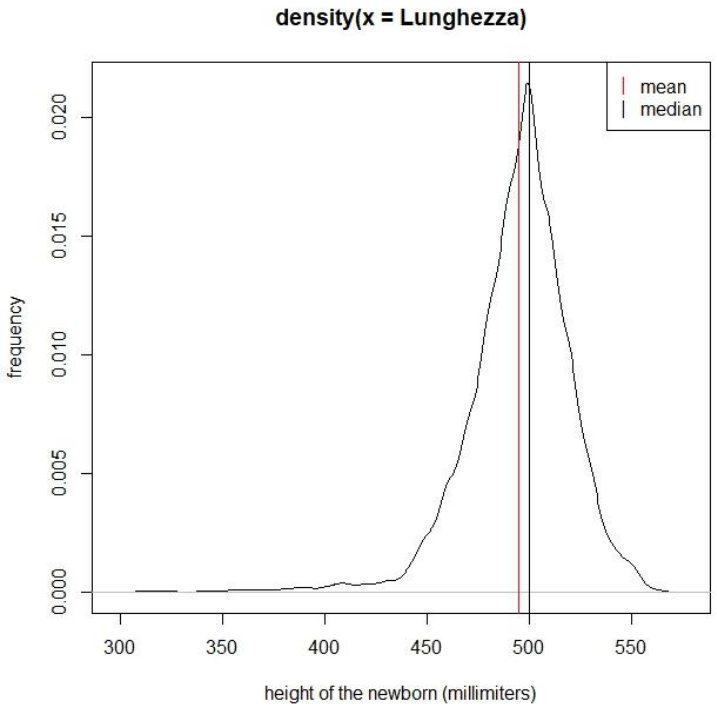
- "*Peso*": it has a leptokurtic distribution and presents a negative asymmetry.

```
> summary(Peso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    830    2990    3300    3284    3620    4930
> IQR(Peso)
[1] 630
> var(Peso)
[1] 275665.7
> sd(Peso)
[1] 525.0387
> coeff.vaiation(Peso)
[1] 15.98739
> skewness(Peso)
[1] -0.6470308
> kurtosis(Peso)-3
[1] 2.031532
```



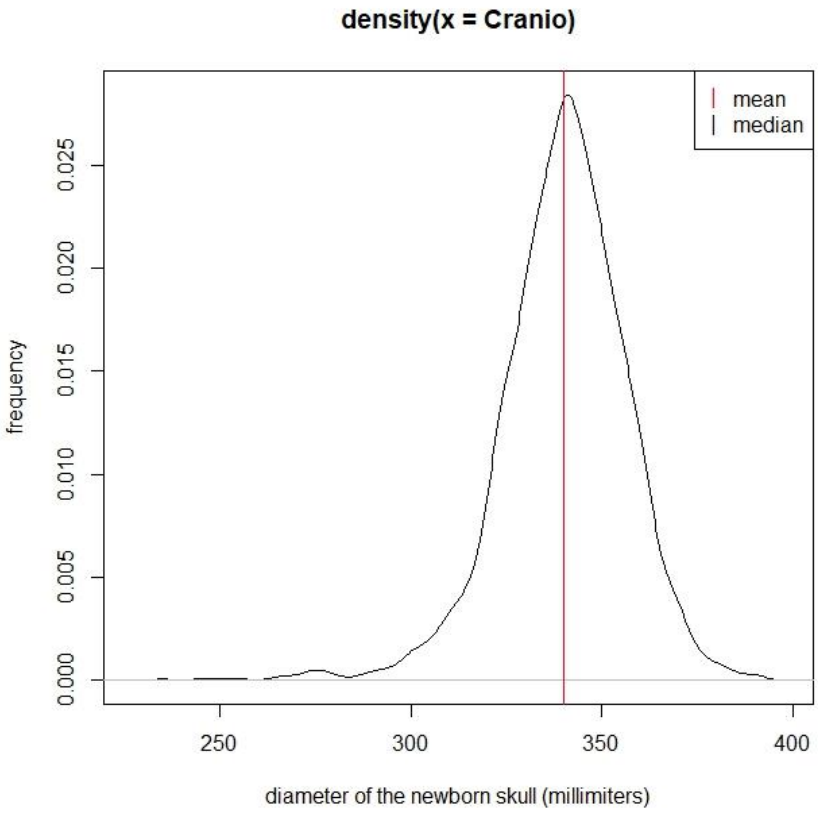
- "*Lunghezza*": again, the distribution is leptokurtic and negatively asymmetric.

```
> summary(Lunghezza)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  310.0   480.0   500.0   494.7   510.0   565.0
> IQR(Lunghezza)
[1] 30
> var(Lunghezza)
[1] 692.671
> sd(Lunghezza)
[1] 26.31864
> coeff.vaiation(Lunghezza)
[1] 5.320208
> skewness(Lunghezza)
[1] -1.514699
> kurtosis(Lunghezza)-3
[1] 6.487174
```



- "*Cranio*": the distribution presents a small left asymmetry and it is leptokurtic.

```
> summary(Cranio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    235     330     340     340     350     390
> IQR(Cranio)
[1] 20
> var(Cranio)
[1] 269.7915
> sd(Cranio)
[1] 16.42533
> coeff.vaiation(Cranio)
[1] 4.830565
> skewness(Cranio)
[1] -0.7850527
> kurtosis(Cranio)-3
[1] 2.946206
```



- "*Tipo.parto*":

	ni	fi
Ces	728	0.2912
Nat	1772	0.7088

- "*Ospedale*":

	ni	fi
osp1	816	0.3264
osp2	849	0.3396
osp3	835	0.3340

- "*Sesso*":

	ni	fi
F	1256	0.5024
M	1244	0.4976

4. **Test if “Peso” & “Lunghezza” means differ from the real population data:**

As previously emerged, neither “Peso” nor “Lunghezza” variables are normally distributed. However, the Central-limit Theorem states that for large n, the distribution of a sample estimator (like the sum or the mean) approximates to a Gaussian curve independently of the variable population distribution. Therefore, considering a sample size of 2500, parametric tests will be adopted.

```
> t.test(Peso,
+        mu = 3200,
+        conf.level = 0.95,
+        alternative ="two.sided")
```

```
        One Sample t-test

data:  Peso
t = 8.0071, df = 2499, p-value = 1.782e-15
alternative hypothesis: true mean is not equal to 3200
95 percent confidence interval:
 3263.490 3304.672
sample estimates:
mean of x
 3284.081
```

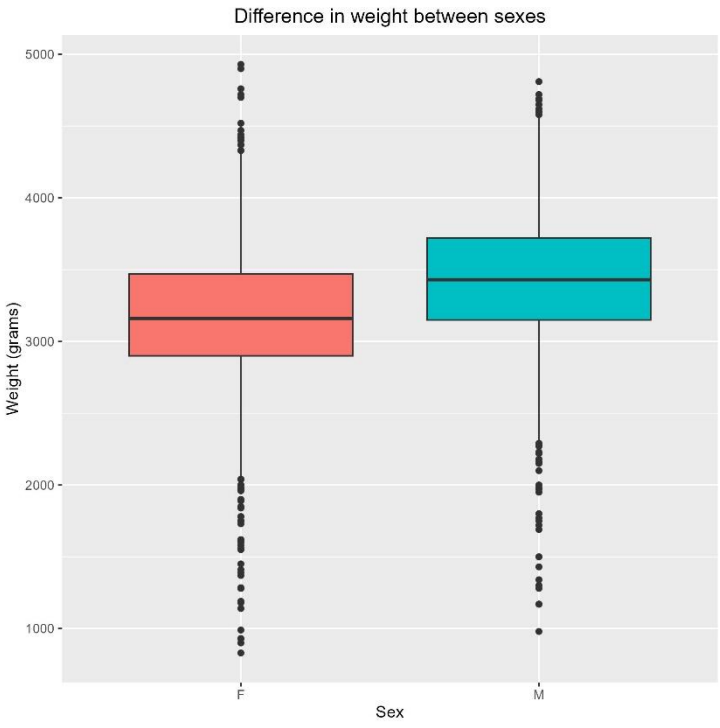
```
        One Sample t-test

data:  Lunghezza
t = -10.084, df = 2499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 493.6598 495.7242
sample estimates:
mean of x
  494.692
```

The worldwide mean weight and height of a newborn with full gestation are equal to 3,2 kg and 50 cm. Both significantly differ from the sample mean values (3284 g & 494.7 mm).

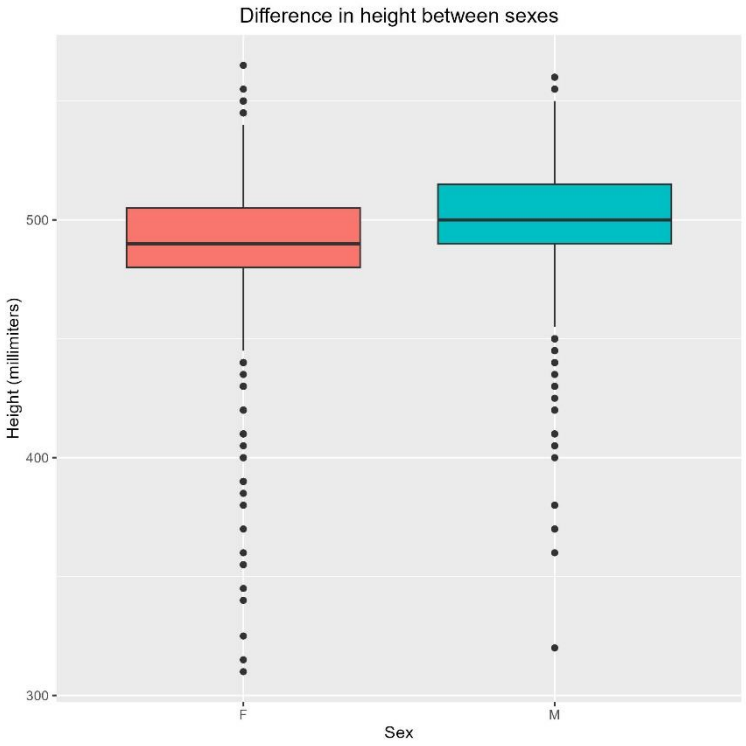
5. **Test for the same variables and the remaining ones (when the comparison is worth) the difference between males and females:**

```
> mean(Peso[Sesso=="M"])
[1] 3408.215
> mean(Peso[Sesso=="F"])
[1] 3161.132
```

Considering the variable “*Peso*” the Welch Two Sample t-test reports a statistically significant difference between sexes.

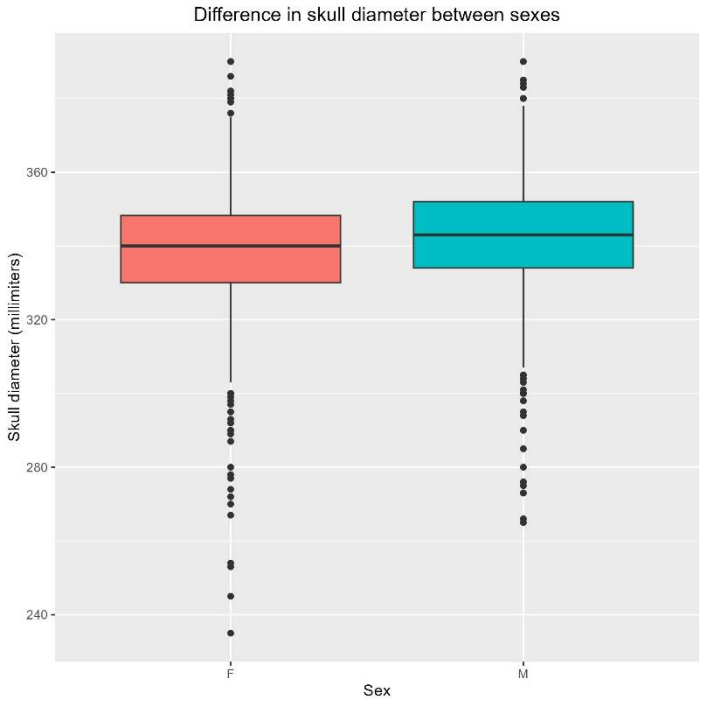



```
> t.test(data=newborn,
+        Peso~Sesso,
+        conf.level = 0.95,
+        alternative ="less")

        Welch Two Sample t-test

data:  Peso by Sesso
t = -12.106, df = 2490.7, p-value < 2.2e-16
alternative hypothesis: true difference in means between group F and group M is less than 0
95 percent confidence interval:
      -Inf -213.4997
sample estimates:
mean in group F mean in group M
       3161.132        3408.215
```

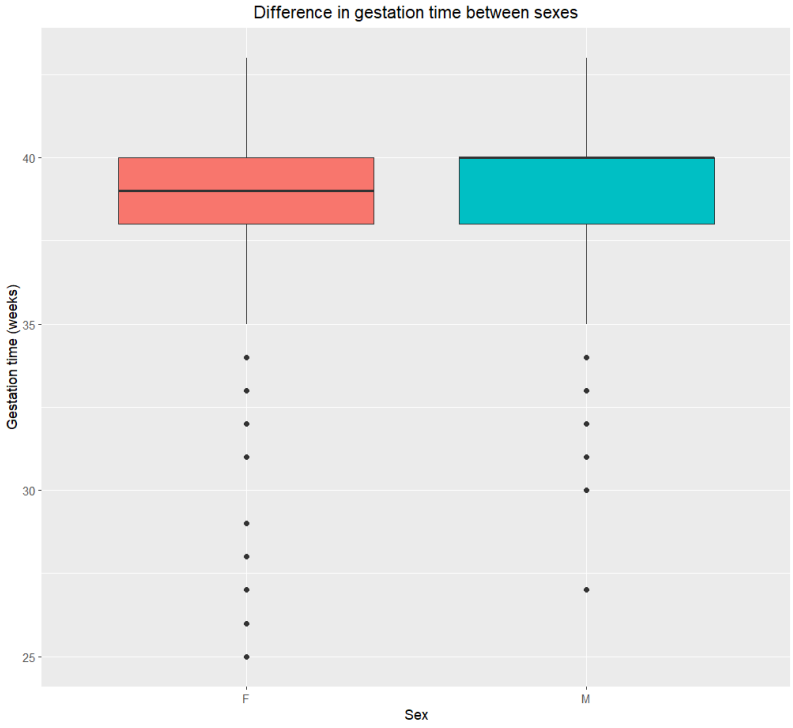


On the other hand, also "*Lunghezza*", "*Cranio*", and "*Gestazione*" t-test results point out statistically significant differences between males and females, with males having higher sample means (p-values = 2.2e-16, 8.588e-14, and 1.228e-11).

```
        Welch Two Sample t-test

data:  Lunghezza by Sesso
t = -9.582, df = 2459.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group F and group M is less than 0
99 percent confidence interval:
      -Inf -7.497049
sample estimates:
mean in group F mean in group M
       489.7643        499.6672
```




```
        Welch Two Sample t-test

data:  Cranio by Sesso
t = -7.4102, df = 2491.4, p-value = 8.588e-14
alternative hypothesis: true difference in means between group F and group M is less than 0
99 percent confidence interval:
      -Inf -3.302818
sample estimates:
mean in group F mean in group M
       337.6330        342.4486
```

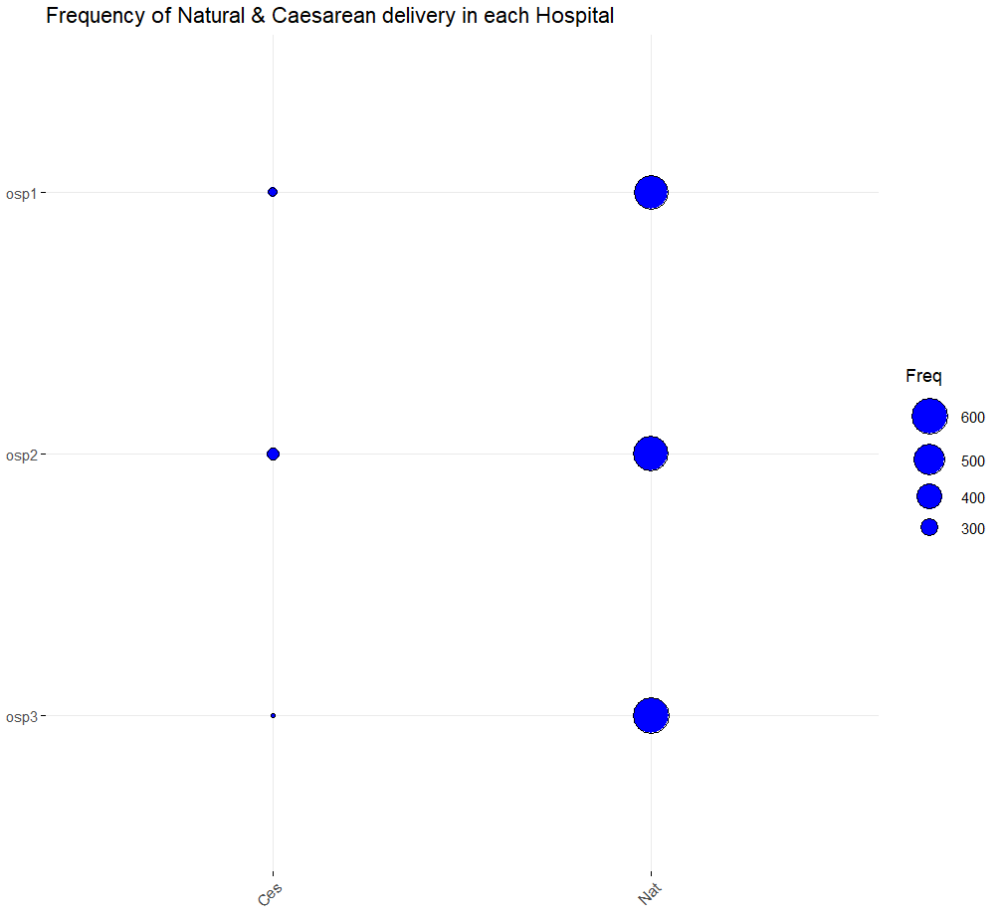



```
        Welch Two Sample t-test

data:  Gestazione by Sesso
t = -6.7072, df = 2446.6, p-value = 1.228e-11
alternative hypothesis: true difference in means between group F and group M is less than 0
99 percent confidence interval:
       -Inf -0.3242601
sample estimates:
mean in group F mean in group M
       38.73328        39.22990
```

6. Are some hospitals performing more Caesarean deliveries than others?

Firstly, with R help is possible to visualize the contingency table with the marginal frequencies of the two qualitative variables, together with the balloon plot:

```
Observed = as.matrix(table(Tipo.parto, Ospedale))
Observed.1 = cbind(Observed,TotaleR = margin.table(Observed,1))
Observed.2  = rbind(Observed.1,totaleC = margin.table(Observed.1,2))
Observed.2

install.packages("ggpubr")
ggpubr::ggballoonplot(data = as.data.frame(Observed),
                      fill = "blue",
                      main = "Frequency of Natural & Caesarean childbirth in each Hospital")
```




```
> Observed.2
         osp1 osp2 osp3 TotaleR
Ces       242  254  232    728
Nat       574  595  603   1772
totaleC   816  849  835   2500
```

To understand if some hospitals are more likely than others to practice caesarean section, the Chi-squared test might come to help. Indeed, it is useful for assessing if there is an association between the variables "*Ospedale*" and "*Tipo.parto*" or whether the frequency of delivery is not affected by the hospital (independence).

```
> chisq.test(Observed)

        Pearson's Chi-squared test

data:  Observed
X-squared = 1.0972, df = 2, p-value = 0.5778
```

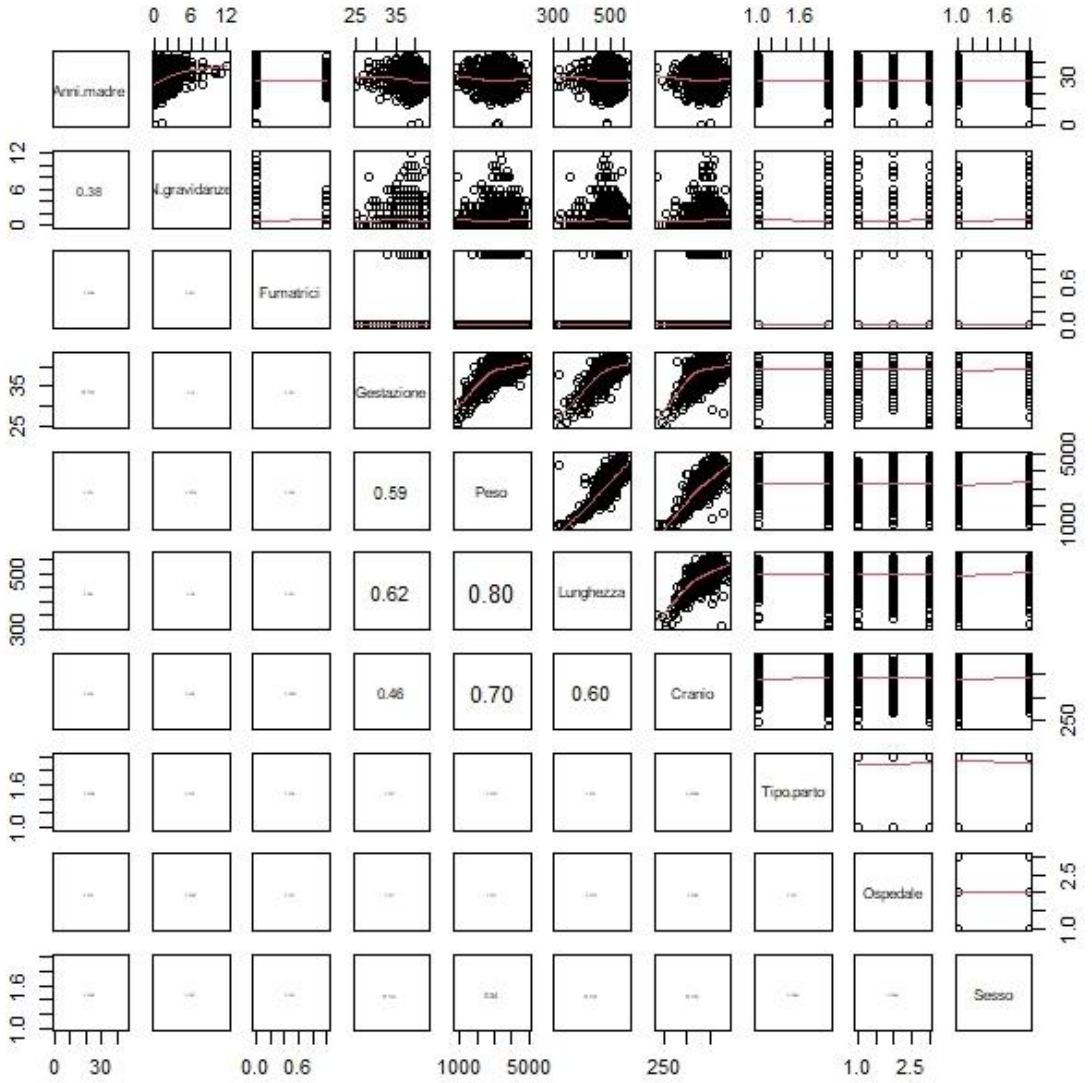
The result of Pearson's Chi-squared test shows how there is no significant difference in delivery type considering the three hospitals contained within the dataset (p-value > 0.5) establishing how these two variables are independent of each other.

MULTIDIMENSIONAL ANALYSIS

1) Investigate the relationship between the Response variable and Predictors

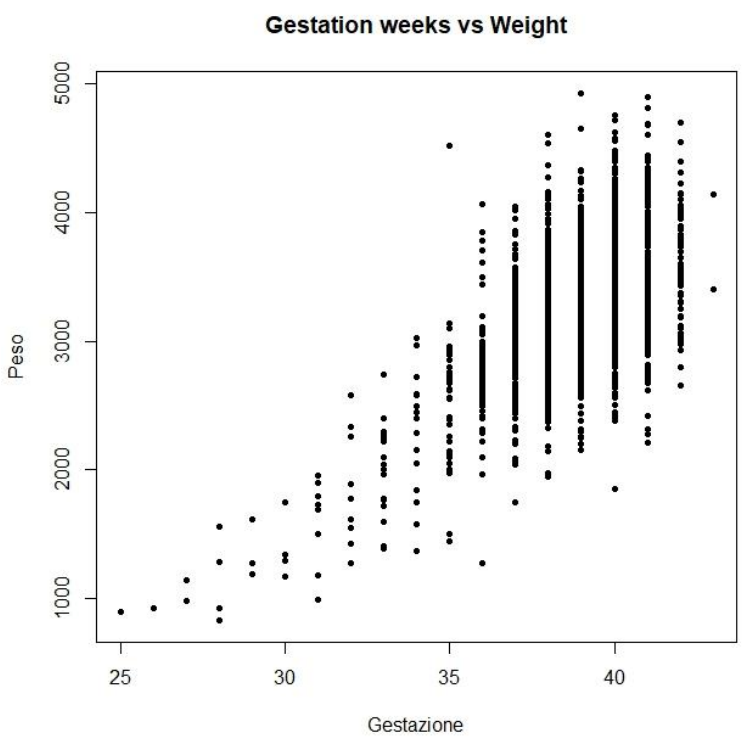
This project goal is to unveil if the newborn weight can be predicted using the remaining variables of the dataset "newborn". The first step in this complex process consists of evaluating the relationship between the Response ("*Peso*") and the Predictors. Moreover, it is also important to assess what relation occurs between Response variables:

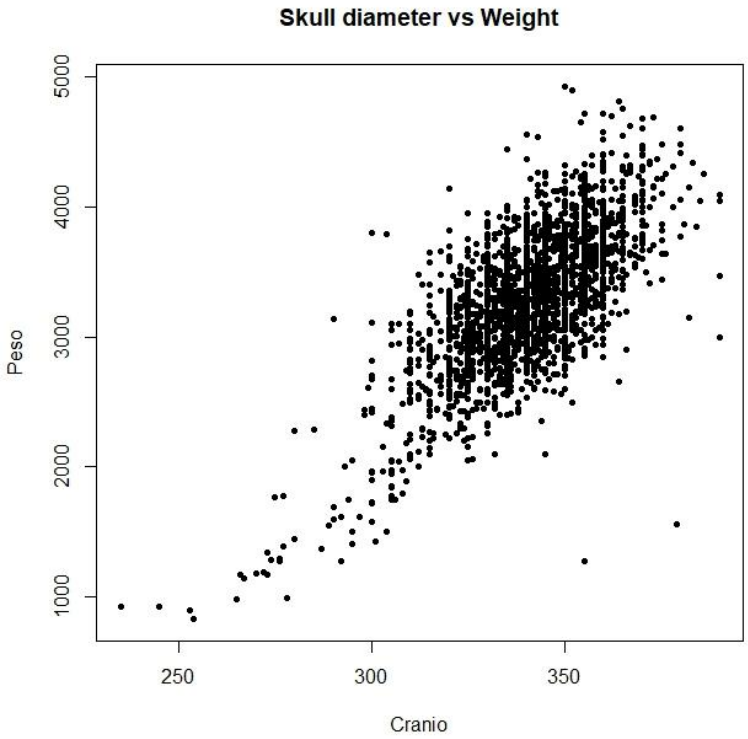
```
> pairs(newborn, upper.panel = panel.smooth,lower.panel = panel.cor)
```

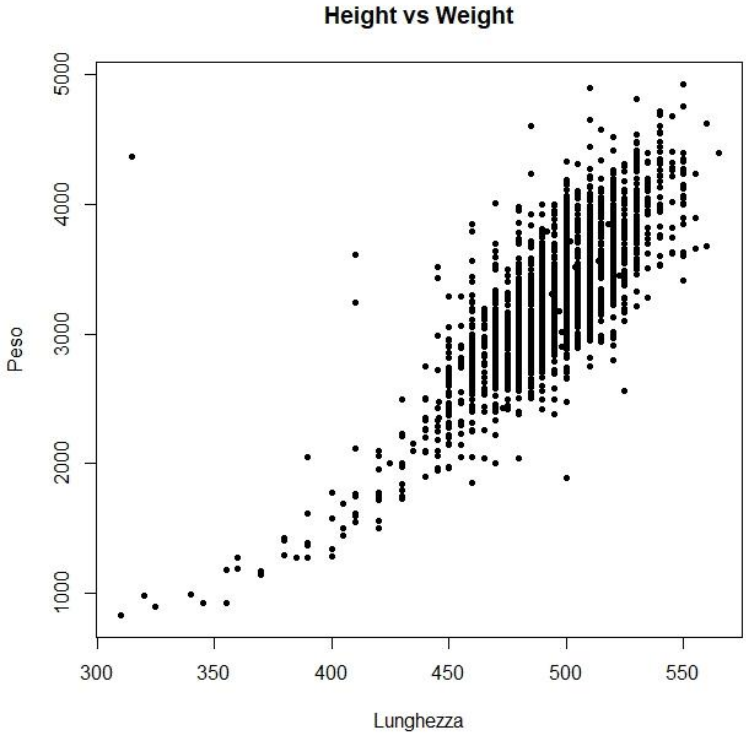


The upper panel furnishes some useful information:

- The highest correlated covariates with newborns' weight are "*Gestazione*"(ρ = 0.59), "*Lunghezza*" (ρ = 0.80), and "*Cranio*" (ρ = 0.70) which is foreseeable. Indeed, it is likely for babies that have a longer gestation time to weigh more at childbirth. In the same way, infants that weigh more will also have a wider skull diameter or they will be higher, on average;



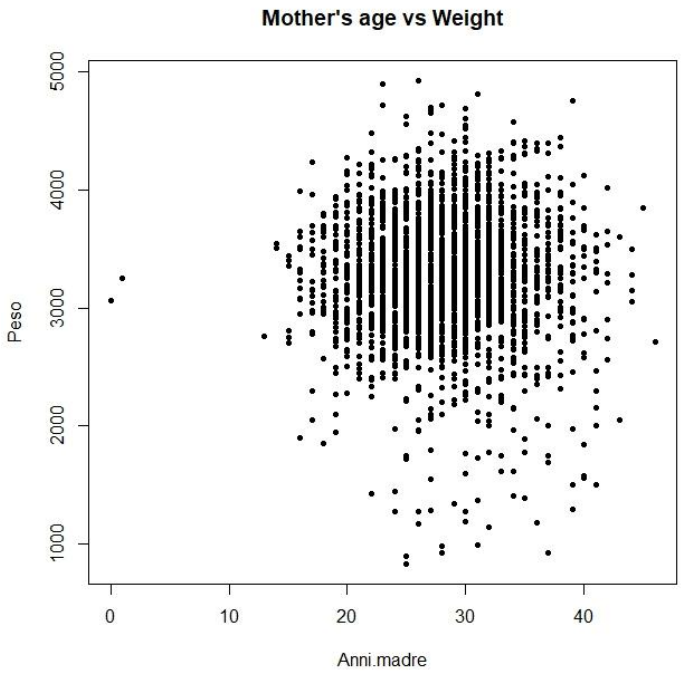


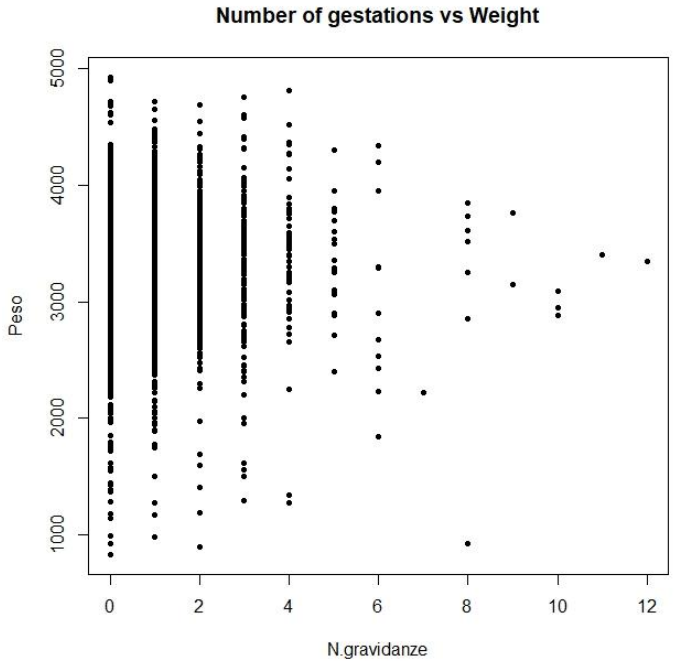


- For the same reason, these three explanatory variables are highly correlated with each other ("*Cranio*" vs "*Lunghezza*", ρ = 0.60; "*Cranio*" vs "*Gestazione*", ρ = 0.46; "*Gestazione*" vs "*Lunghezza*", ρ = 0.62). Thus, if all are included in the regression model they might give multicollinearity-related issues with subsequent instability of β coefficients;

- For what concerns the quantitative variables, "*Peso*" does not seem to be associated with them. On the other side, "*Anni.madre*" has a moderate correlation with "*N.gravidanze*" (ρ = 0.38).

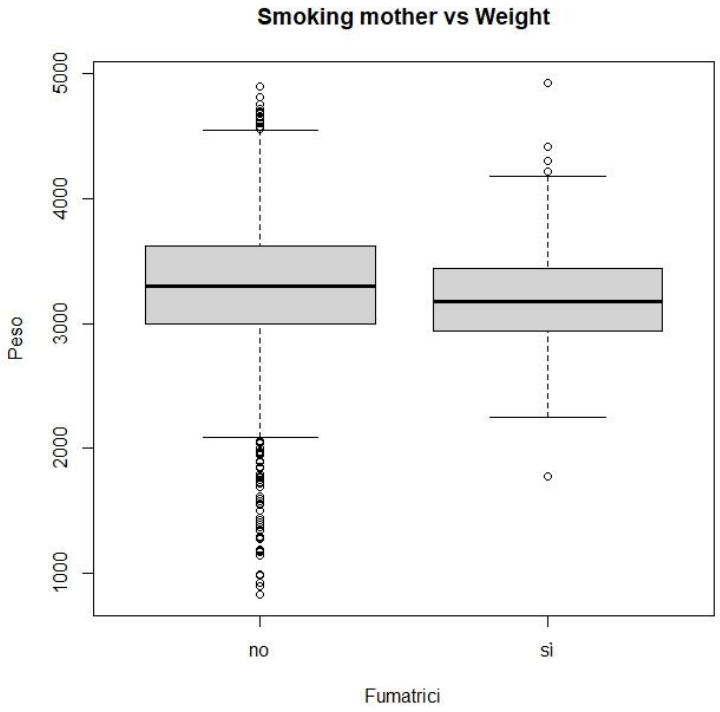
```
> cor(Anni.madre,Peso)
[1] -0.02247017
> cor(N.gravidanze,Peso)
[1] 0.0024073
```





- Knowing that an infant's weight at birth significantly varies considering its sex, let's examine if other qualitative variables might be impacting weight prediction:

```
> t.test(Peso~Fumatrici)

        Welch Two Sample t-test

data:  Peso by Fumatrici
t = 1.034, df = 114.1, p-value = 0.3033
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -45.61354 145.22674
sample estimates:
mean in group 0 mean in group 1
       3286.153        3236.346
```

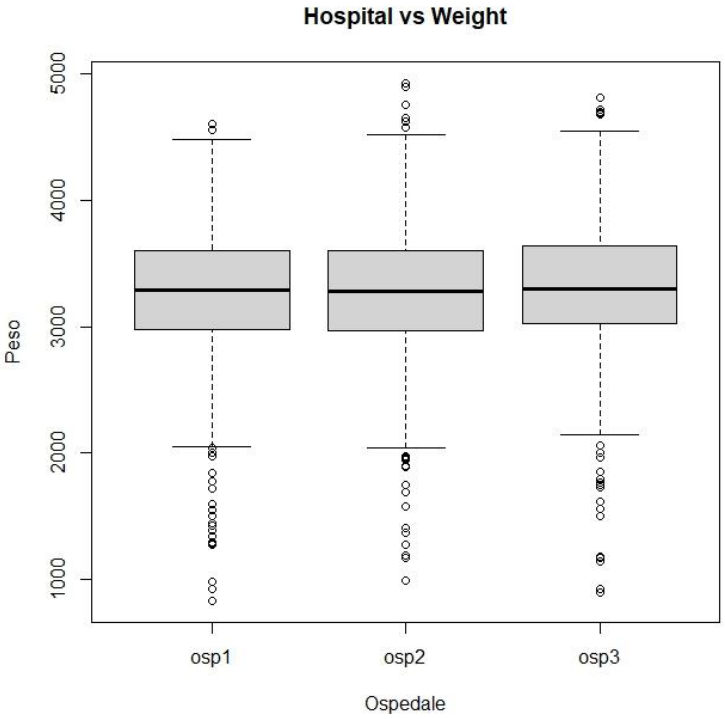


No statistically meaningful differences occur when infants' weight is considered in smoker vs non-smoker mothers or in differnet delivery types.

```
> mean(Peso[Fumatrici==0])
[1] 3286.153
> mean(Peso[Fumatrici==1])
[1] 3236.346
```

```
        Welch Two Sample t-test

data:  Peso by Tipo.parto
t = -0.12968, df = 1493, p-value = 0.8968
alternative hypothesis: true difference in means between group Ces and group Nat is not equal to 0
95 percent confidence interval:
 -46.27992  40.54037
sample estimates:
mean in group Ces mean in group Nat
         3282.047          3284.916
```


```
> mean(Peso[Tipo.parto=="Nat"])
[1] 3284.916
> mean(Peso[Tipo.parto=="Ces"])
[1] 3282.047
```




```
> pairwise.t.test(Peso,Ospedale,
+         paired = FALSE,
+         pool.sd = TRUE,
+         p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  Peso and Ospedale

     osp1 osp2
osp2 1.00 -
osp3 0.33 0.33

P value adjustment method: bonferroni
```

```
> mean(Peso[Ospedale=="osp1"])
[1] 3270.266
> mean(Peso[Ospedale=="osp2"])
[1] 3270.483
> mean(Peso[Ospedale=="osp3"])
[1] 3311.407
```

There is no relevant difference in weight also among hospitals. The Bonferroni corrected p-value relative to the difference between hospitals 1 and 2 is even 1.00 (due to R software correction). Indeed, looking at mean values per hospital the two are almost identical numbers, differing just for the decimal components.

2) Create a Multivariate Linear Regression model containing all the variables of the dataset

Building the regression model by taking into account all the possible Predictors, produces the β coefficients (and respective p-values) shown below. It can be noticed how "*Anni.madre*", "*Fumatrici*", and "*Ospedaleosp2*" (dummy variable that works together with "*Ospedaleosp3*" and is nothing less than the result of "*Ospedale*" transformation) have non-significative β coefficients, with respective p-values of 0.4383, 0.2735, and 0.4043. Therefore, they do not give a precious contribution to the weight estimation and will be removed from the model. These considerations are in line with the correlation coefficients previously elaborated.

```
> model1 = lm(Peso~., data = newborn)
> summary(model1)

Call:
lm(formula = Peso ~ ., data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1124.40  -181.66   -14.42   160.91  2611.89

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6738.4762   141.3087 -47.686  < 2e-16 ***
Anni.madre        0.8921     1.1323   0.788   0.4308
N.gravidanze     11.2665     4.6608   2.417   0.0157 *
Fumatrici       -30.1631    27.5386  -1.095   0.2735
Gestazione       32.5696     3.8187   8.529  < 2e-16 ***
Lunghezza        10.2945     0.3007  34.236  < 2e-16 ***
Cranio           10.4707     0.4260  24.578  < 2e-16 ***
Tipo.partoNat    29.5254    12.0844   2.443   0.0146 *
Ospedaleosp2    -11.2095    13.4379  -0.834   0.4043
Ospedaleosp3     28.0958    13.4957   2.082   0.0375 *
SessoM           77.5409    11.1776   6.937 5.08e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 273.9 on 2489 degrees of freedom
Multiple R-squared:  0.7289,    Adjusted R-squared:  0.7278
F-statistic: 669.2 on 10 and 2489 DF,  p-value: < 2.2e-16
```

On the other hand, all the other regression coefficients have statistically significant p-values:

- "*N.gravidanze*", "*Ospedaleosp3*", and "*Tipo.partoNat*" p-values ~ 0.01;
- "*Gestazione*", "*Lunghezza*", "*Cranio*", and "*SessoM*", (adding 32.56, 10.29, 10.47, and 77.54 grams of weight for each predictor unit variation respectively) with p-values close to 0, indicating how meaningful they are in foreseeing the Response value;

Overall, the model explains around 72 % of the Outcome variable total variability ($R^2 = 0.7289$ and adjusted $R^2 = 0.7278$).

3) Find the "best" model using the known selection parameters:

A first update of the model, can be performed by excluding the non-statistically significant regressors:

```
> model2 = lm(Peso~N.gravidanze+Gestazione+Lunghezza+Cranio+Tipo.parto+Sesso, data = newborn)
> summary(model2)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Sesso, data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1129.31  -181.70   -16.31   161.07  2638.85

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6707.2971   135.9911 -49.322  < 2e-16 ***
N.gravidanze    12.7558     4.3366   2.941   0.0033 **
Gestazione      32.2713     3.7941   8.506  < 2e-16 ***
Lunghezza       10.2864     0.3007  34.207  < 2e-16 ***
Cranio          10.5057     0.4260  24.659  < 2e-16 ***
Tipo.partoNat   30.0342    12.0969   2.483   0.0131 *
SessoM          77.9285    11.1905   6.964 4.22e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274.3 on 2493 degrees of freedom
Multiple R-squared:  0.7277,    Adjusted R-squared:  0.727
F-statistic:  1110 on 6 and 2493 DF,  p-value: < 2.2e-16
```

After this update, "*N.gravidanze*" β-coefficient acquires significance compared to the previous elaborations (differently from what occurred for the correlation coefficient analysis). However, the adjusted R^2 does not change much (0.7278 vs 0.7270).

A further perfecting step consists of maintaining exclusively the regressors with the smallest p-values according to the parsimony principle (not including anything more within the model unless is it essential) and evaluating what happens to the model overall:

```
> model3 = update(model2, ~.-Tipo.parto)
> summary(model3)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso, data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1149.44  -180.81   -15.58   163.64  2639.72

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6681.1445   135.7229 -49.226  < 2e-16 ***
N.gravidanze    12.4750     4.3396   2.875  0.00408 **
Gestazione      32.3321     3.7980   8.513  < 2e-16 ***
Lunghezza       10.2486     0.3006  34.090  < 2e-16 ***
Cranio          10.5402     0.4262  24.728  < 2e-16 ***
SessoM          77.9927    11.2021   6.962 4.26e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274.6 on 2494 degrees of freedom
Multiple R-squared:  0.727,     Adjusted R-squared:  0.7265
F-statistic:  1328 on 5 and 2494 DF,  p-value: < 2.2e-16
```

```
> model4 = update(model3, ~.-N.gravidanze)
> summary(model4)

Call:
lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso,
    data = newborn)

Residuals:
    Min      1Q  Median      3Q     Max
-1138.2  -184.3   -17.6   163.3  2627.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6651.1188   135.5172 -49.080  < 2e-16 ***
Gestazione     31.2737     3.7856   8.261 2.31e-16 ***
Lunghezza      10.2054     0.3007  33.939  < 2e-16 ***
Cranio         10.6704     0.4245  25.139  < 2e-16 ***
SessoM         79.1049    11.2117   7.056 2.22e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 275 on 2495 degrees of freedom
Multiple R-squared:  0.7261,    Adjusted R-squared:  0.7257
F-statistic:  1654 on 4 and 2495 DF,  p-value: < 2.2e-16
```

The new adjusted R^2 are respectively 0.7265 (model3) and 0.7257 (model4) after having removed two regressors. Thus model3 seems the more promising between the two.

As noticed before, three out of five variables included in the 3rd model3 present high correlation coefficients between each other. This might be a problem, 1) because it would not be possible to estimate the effect of each predictor on the response independently; 2) the β-coefficient estimations would widely change depending on which other independent variable is included in the model. Given that, the Variance Inflation Factor (VIF) is a powerful index that helps in understanding if some variables should be excluded (VIF > 5):

```
> vif(model3)
N.gravidanze   Gestazione    Lunghezza       Cranio        Sesso
    1.023475     1.669189     2.074689     1.624465     1.040054
```

No VIF among the ones calculated exceeds 5, therefore it seems there are not multicollinearity-related issues.

Through the computation of Akaike Information Criterion (AIC) and Baesian Information Criterion (BIC) it is possible to select the best model. Theoretically, the best situation occurs when both variance and bias are minimized producing a model that fits pretty well to current data and it is also capable of adapting when further observations are added (bias-variance trade-off).

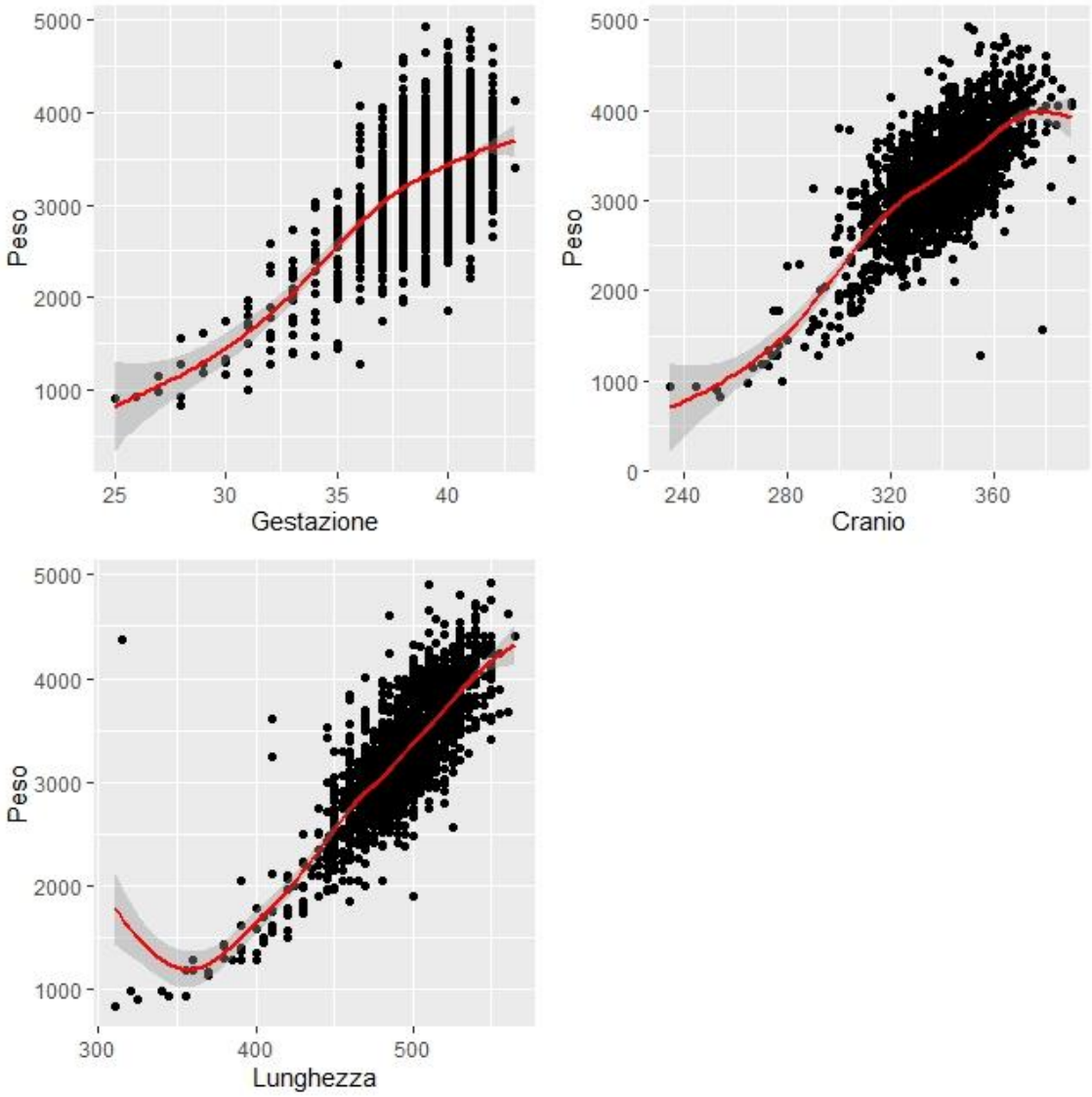
```
> AIC(model1,model2,model3,model4)
       df      AIC
model1 12 35171.95
model2  8 35175.16
model3  7 35179.33
model4  6 35185.60
> BIC(model1,model2,model3,model4)
       df      BIC
model1 12 35241.84
model2  8 35221.75
model3  7 35220.10
model4  6 35220.54
```

Model1 has the lowest AIC and model3 has the lowest BIC. AIC tends to prefer more overfitted models compared to BIC, and the process of selection so far has been done following the parsimony principle. Maintaining this criterion of judgment, model3 can be considered better than model4, having a slightly smaller BIC (35220.10 vs 35220.54).

Anyway, even through AIC evaluation, model3 remains the most suitable choice (35179.33 vs 35185.60).

4) Non-linear effects or Interaction terms

```
> gest = ggplot(newborn)+ geom_point(aes(Gestazione,Peso))+
+   geom_smooth(aes(Gestazione,Peso), col = "red")
```



- Non-linear effects: The figure above is the result of plotting "*Gestazione*", "*Cranio*", and "*Lunghezza*" versus the dependent variable. With the addition of *geom_smooth()* command, it is possible to create a line approximating the trend of the scatter plot, making it easier to visually spot non-linear correlations. None of the three lines generated has a perfect linear trend. Specifically, "*Lunghezza*" red line shape might indicate a quadratic relationship with "*Peso*". Indeed, there is one observation with low height and suspicious high weight whose effect bends the curve upward. On the other side, these three variables either indicate the time in which the fetus develops (gestation time) or the proportions of the infant's body at delivery (length or skull diameter) as previously pointed out. It is expected that the more time the fetus spends in their mother's womb, the heavier it will be at delivery, as it is likely that the body measurements will grow more or less proportionally to each other. For these reasons, non-linear correlations seem unlikely.

Nevertheless, adding the non-linear effects of *"Lunghezza", "Cranio",* and *"Gestazione"* within the model results in an overall increase of the adjusted R^2 only when the quadratic term of the newborn's length is included (0.7363 of model5 versus 0.7265 of model3).

```
> model5 = update(model3,~. +I(Lunghezza^2))
> summary(model5)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso + I(Lunghezza^2), data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1169.72  -181.62   -12.97   163.67  1786.43

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    215.090502 723.560129   0.297 0.766287
N.gravidanze    14.098080   4.264177   3.306 0.000959 ***
Gestazione      42.504831   3.873807  10.972  < 2e-16 ***
Lunghezza      -20.272927   3.161377  -6.413  1.7e-10 ***
Cranio          10.650445   0.418670  25.439  < 2e-16 ***
SessoM          70.007371  11.029627   6.347  2.6e-10 ***
I(Lunghezza^2)   0.031664   0.003265   9.697  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 269.6 on 2493 degrees of freedom
Multiple R-squared:  0.7369,    Adjusted R-squared:  0.7363
F-statistic:  1164 on 6 and 2493 DF,  p-value: < 2.2e-16
```

- Interaction terms: the model built consists of five independent variables: "*N.gravidanze*", "*Gestazione*", "*Lunghezza*", "*Cranio*", and "*Sesso*". Verifying the possible interactions between regressors negatively influences the p-values associated with the single-term coefficients in the most of cases except for *Lunghezza*Cranio* and *Gestazione*Lunghezza* interaction terms. However, the first interaction does not lead to an increment in the Response variable explained variability (same adjusted R^2), therefore it will not be considered further.

```
> model6 = update(model5,~. +I(Lunghezza*Cranio))
> summary(model6)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso + I(Lunghezza^2) + I(Lunghezza * Cranio), data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1176.69  -179.20   -11.78   165.68  1306.79

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.262e+03  1.003e+03  -2.256 0.024135 *
N.gravidanze         1.425e+01  4.254e+00   3.350 0.000821 ***
Gestazione           4.042e+01  3.909e+00  10.339  < 2e-16 ***
Lunghezza           -2.137e+01  3.169e+00  -6.744 1.91e-11 ***
Cranio               2.741e+01  4.725e+00   5.800 7.46e-09 ***
SessoM               7.186e+01  1.102e+01   6.523 8.29e-11 ***
I(Lunghezza^2)       4.476e-02  4.914e-03   9.109  < 2e-16 ***
I(Lunghezza * Cranio) -3.449e-02  9.689e-03  -3.560 0.000378 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 269 on 2492 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7375
F-statistic:  1004 on 7 and 2492 DF,  p-value: < 2.2e-16
```

On the other hand, the second term contained in the model8 brings a more consistent increment in the adjusted R^2 (0.7375 vs 0.7388).

```
> model8 = update(model5,~. +I(Gestazione*Lunghezza))
> summary(model8)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso + I(Lunghezza^2) + I(Gestazione * Lunghezza), data = newborn)

Residuals:
     Min       1Q   Median       3Q      Max
-1212.30  -181.42   -11.57   163.68  1326.28

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.542e+03  9.056e+02  -2.807 0.005044 **
N.gravidanze              1.436e+01  4.244e+00   3.384 0.000726 ***
Gestazione                2.655e+02  4.459e+01   5.954 2.98e-09 ***
Lunghezza                -2.585e+01  3.337e+00  -7.748 1.35e-14 ***
Cranio                    1.036e+01  4.207e-01  24.628  < 2e-16 ***
SessoM                    7.368e+01  1.100e+01   6.698 2.61e-11 ***
I(Lunghezza^2)            5.572e-02  5.790e-03   9.624  < 2e-16 ***
I(Gestazione * Lunghezza) -4.652e-01  9.266e-02  -5.020 5.53e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268.3 on 2492 degrees of freedom
Multiple R-squared:  0.7396,    Adjusted R-squared:  0.7388
F-statistic:  1011 on 7 and 2492 DF,  p-value: < 2.2e-16
```

Importantly, both AIC and BIC confirm the superiority of model8 when compared with previous ones.

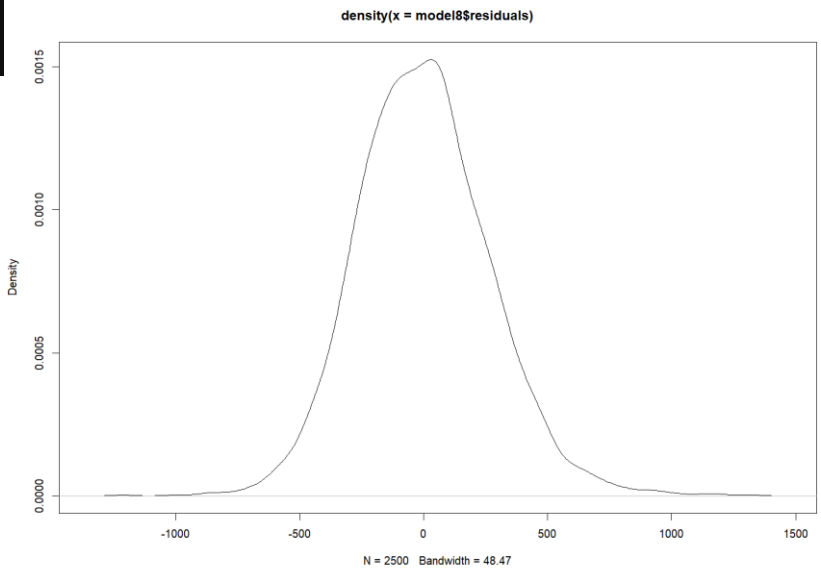
```
> AIC(model3,model6,model5,model8)
       df      AIC
model3  7 35179.33
model6  9 35078.09
model5  8 35088.77
model8  9 35065.62
> BIC(model3,model6,model5,model8)
       df      BIC
model3  7 35220.10
model6  9 35130.51
model5  8 35135.36
model8  9 35118.03
```

5) Residual Analysis

This step of the analysis is crucial to 1) control if residuals are normally distributed: this is necessary to ensure that predictions performed based on the model will be accurate; 2) make sure that the erratic part of the model does not contain any information that could “escape” from the deterministic portion and weaken its statistical power: there should not be any relievable pattern; 3) Analyse if some outliers or leverage values (extreme Response or Predictors observations) might impact on the regression model adequacy.

```
> shapiro.test(model8$residuals)

        Shapiro-Wilk normality test

data:  model8$residuals
W = 0.99062, p-value = 1.06e-11
```

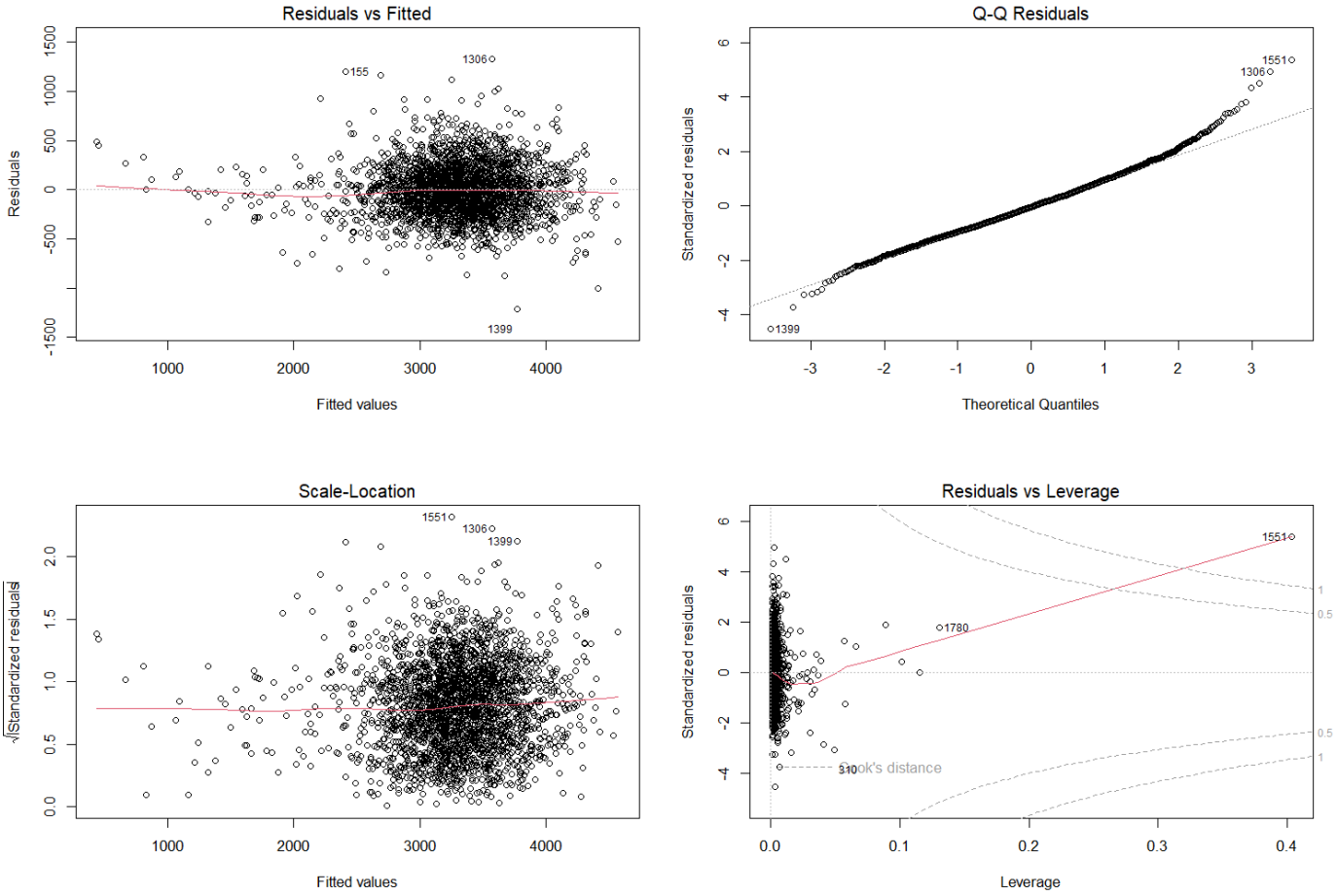
```
> skewness(model8$residuals)
[1] 0.324372
> kurtosis(model8$residuals)-3
[1] 1.014402
```



The Shapiro-Wilk test for normality states that residual distribution is different from a normal one (p-value < 1.06e-11). Indeed, the curve is leptokurtic even though the residual population is not far from a Gaussian shape (the Shapiro-Wilk test is pretty sensitive to variation from the normal distribution). However, this element may contribute negatively to the model's adequacy by impacting β-coefficients hypothesis testing validity. The Breusch-Pagan test performs the homoscedasticity check. Homoscedasticity or homogeneity of variance, is an important assumption to be met for ensuring an accurate prediction across the whole range of the model. In this case, homoscedasticity is violated (p-value = 5.782e-15) and prediction accuracy might be affected. The Durbin-Watson test on the other side, controls whether residuals are correlated. If a certain level of autocorrelation is present, it means that there is some hidden pattern in the erratic part that has not been explained by the model itself. In this case, the test is not significative, indicating non-correlated residuals (p-value = 0.08402).

```
> lmtest::bptest(model8)

        studentized Breusch-Pagan test

data:  model8
BP = 81.847, df = 7, p-value = 5.782e-15

> lmtest::dwtest(model8)

        Durbin-Watson test

data:  model8
DW = 1.9449, p-value = 0.08402
alternative hypothesis: true autocorrelation is greater than 0
```

To look deeper at the model and understand if something needs to be ameliorated, the below plot comes as a useful tool:



- The upper-left frame plots the Residuals versus the Fitted values and it is useful to visualize linearity & independence: it is fundamental for residuals mean to match the zero across the whole length of the x-axis otherwise predictions detach from the actual reported value. Moreover, any trends besides a random distribution around zero-mean would indicate the violation of the independence assumption and an incomplete or inadequate model building. In this case, residuals are randomly scattered around zero even though there is not a complete overlapping of the mean (red line) with the x = 0 line;

- The upper-right Q-Q plot checks for the normality of residuals: the perfectly normal population would lay on the bisector line. As stated by the Shapiro-Wilk test, this is not the case;

- The lower-left graph evaluates the homogeneity of variance & independence. If the homoscedasticity assumption is violated, β coefficients are less precise, and therefore p-value estimates are incorrect (smaller) leading to false predictions. Again, there is a slight upward inclination on both extremes;

- Lower-right panel shows Residuals versus Leverage, and it is needed to spot potentially influential values (present among outliers or high leverage values that are respectively

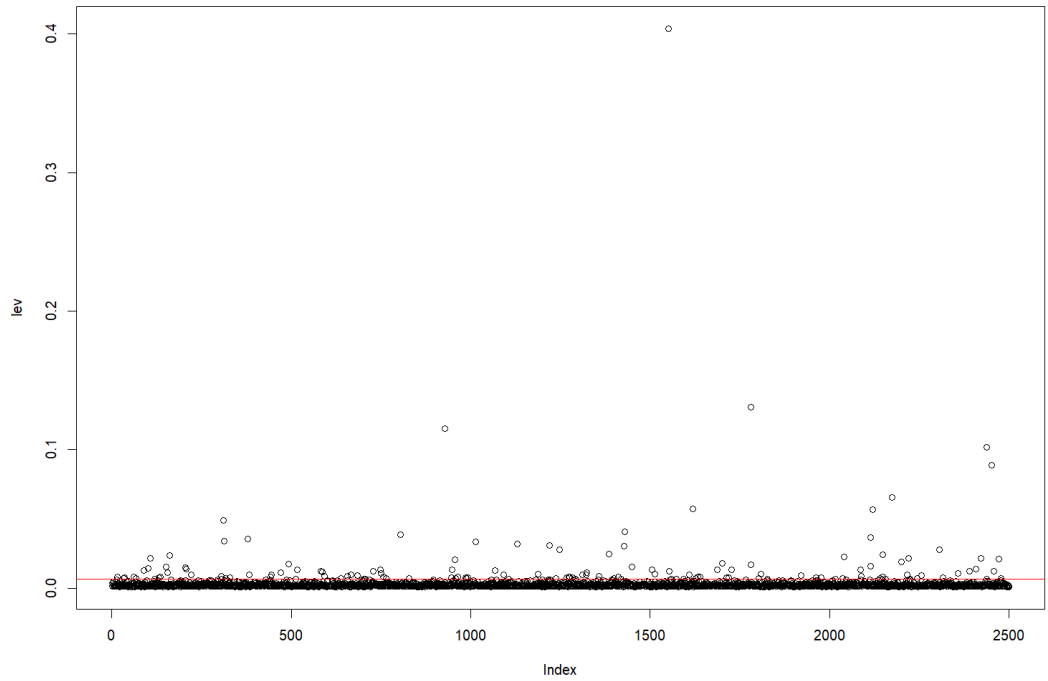
extreme outcome or predictor variable values). The inclusion of influential values might modify the results of the regression model. Cook's distance is a quantitative parameter that helps understand if the model contains some influential values. In this case, the observation number 1551 has a Cook's distance higher than 1, therefore it is conditioning the model to some extent.

To better analyze the influential-cases-related issues, it is possible to examine high-leverage values:

```
> plot(lev)
> abline(h = soglia, col = "red")
> lev = hatvalues(model8)
> plot(lev)
> p = sum(lev)
> soglia = 2*p/n
> abline(h = soglia, col = "red")
```

```
> length(lev[lev>soglia])
[1] 134
```

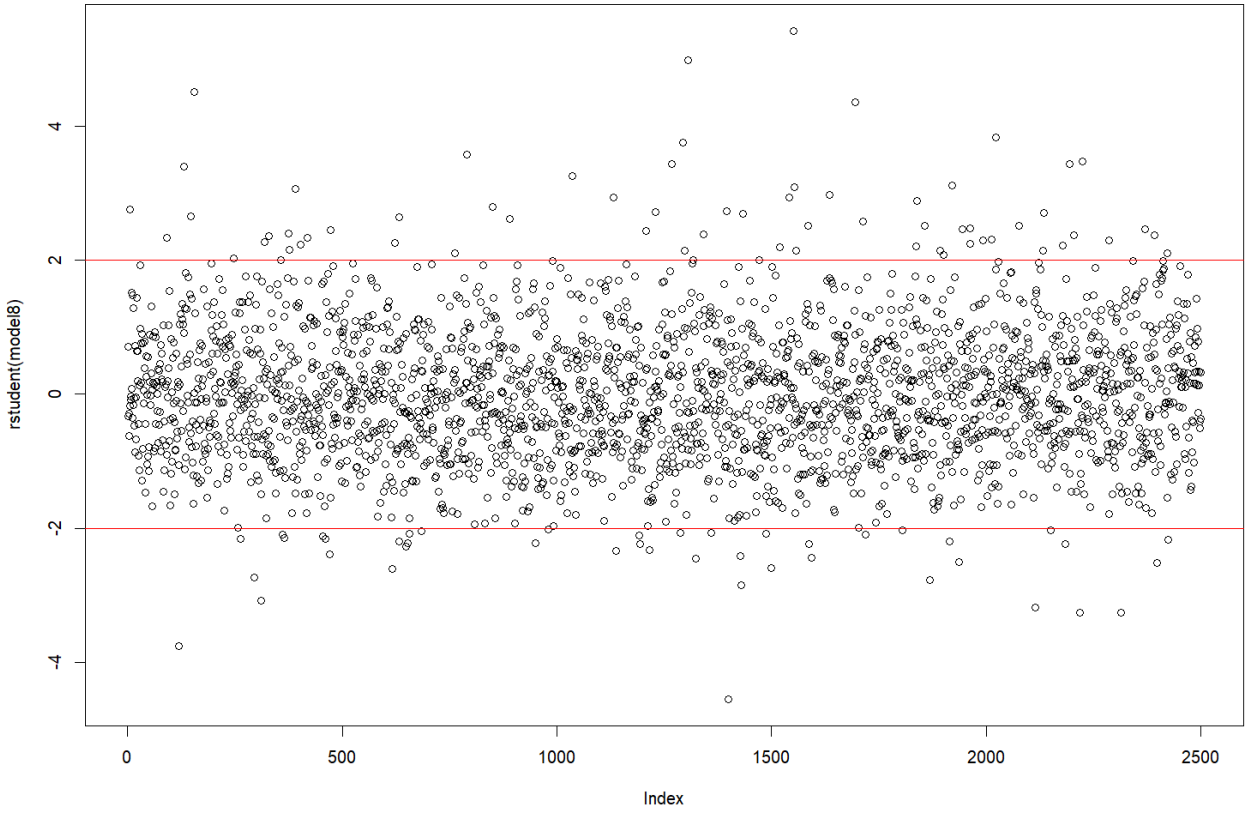
Finding 134 high-leverage observations present in model number 8.



On the other side, outliers can be examined:

```
> plot(rstudent(model8))
> abline(h = c(-2,2), col = "red")
> outlierTest(model8)
      rstudent unadjusted p-value Bonferroni p
1551  5.411835         6.8338e-08   0.00017085
1306  4.971778         7.0852e-07   0.00177130
1399 -4.543738         5.7902e-06   0.01447500
155   4.509774         6.7903e-06   0.01697600
1694  4.352471         1.4004e-05   0.03501100
```

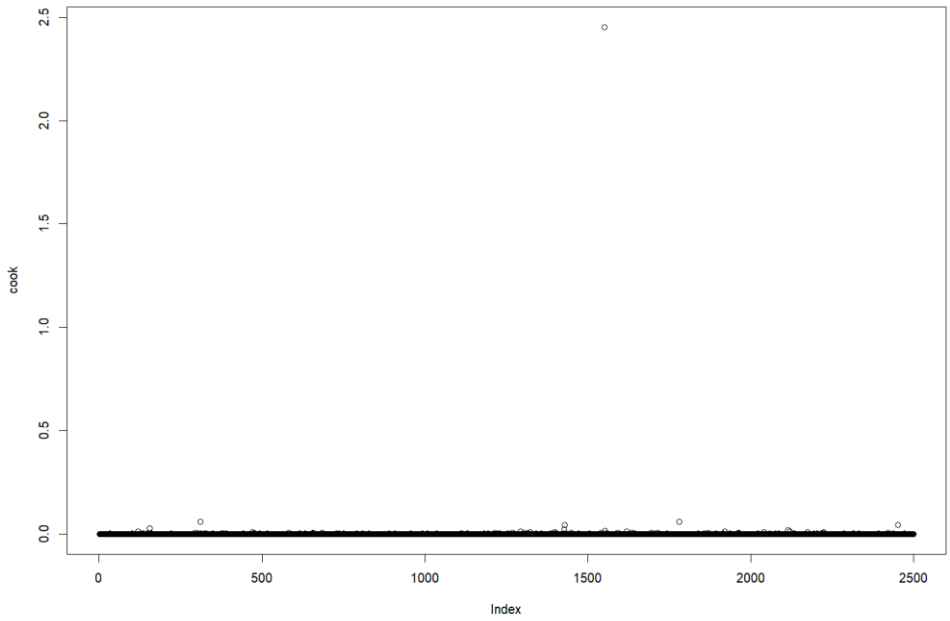
In this case, five values are reported by R (among which there is the value 1551), and way more are visually displayed;



The Cook's distance is the most concise way for influential values identification:

```
> cook =cooks.distance(model8)
> plot(cook)
> max(cook)
[1] 2.451781
```

R identifies one observation with a value of 2.451, mirroring the result of the Residuals versus Leverage plot which identified observation 1551 with a Cook's distance higher than 1.



This observation refers to a baby girl with a weight and a cranium diameter above the 3rd quartile (3620 grams and 350 millimeters), both close to their maximum value (4370 in 4930 grams and 374 millimeters in 390). The same child has an uncommon height for these measurements (315 millimeters) which is close to the minimum value for that measure (310.0 millimeters).

```
> newborn[1551,1:10]
     Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto Ospedale
1551         35            1         0         38 4370       315    374        Nat     osp3
     Sesso
1551     F
```

This specific case has uncommon relationships among weight, height, and cranium diameter, very different than the average. The reasons behind this might be diverse: a mistake in data recording or even the presence of a pathological condition. Eliminating observation 1551 might benefit the overall model predictive capacity.

```
> summary(newborn)
   Anni.madre     N.gravidanze       Fumatrici       Gestazione         Peso
 Min.   : 0.00   Min.   : 0.0000   Min.   :0.0000   Min.   :25.00   Min.   : 830
 1st Qu.:25.00   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:38.00   1st Qu.:2990
 Median :28.00   Median : 1.0000   Median :0.0000   Median :39.00   Median :3300
 Mean   :28.16   Mean   : 0.9812   Mean   :0.0416   Mean   :38.98   Mean   :3284
 3rd Qu.:32.00   3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.:40.00   3rd Qu.:3620
 Max.   :46.00   Max.   :12.0000   Max.   :1.0000   Max.   :43.00   Max.   :4930
   Lunghezza        Cranio     Tipo.parto Ospedale   Sesso
 Min.   :310.0   Min.   :235   Ces: 728   osp1:816   F:1256
 1st Qu.:480.0   1st Qu.:330   Nat:1772   osp2:849   M:1244
 Median :500.0   Median :340              osp3:835
 Mean   :494.7   Mean   :340
 3rd Qu.:510.0   3rd Qu.:350
 Max.   :565.0   Max.   :390
```

6) **How good is the model for making predictions?**

```
> newborn1 = newborn[-1551,]
> model8.1 = lm(Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
+               + I(Lunghezza^2) + I(Gestazione*Lunghezza),
+               data = newborn1)
> summary(model8.1)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso + I(Lunghezza^2) + I(Gestazione * Lunghezza), data = newborn1)

Residuals:
     Min       1Q   Median       3Q      Max
-1189.08  -179.96   -12.38   165.13  1314.19

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.245e+03  9.021e+02  -2.489 0.012886 *
N.gravidanze             1.427e+01  4.220e+00   3.380 0.000735 ***
Gestazione               1.070e+02  5.315e+01   2.012 0.044295 *
Lunghezza               -1.443e+01  3.932e+00  -3.670 0.000247 ***
Cranio                   1.012e+01  4.207e-01  24.047  < 2e-16 ***
SessoM                   7.310e+01  1.094e+01   6.682 2.90e-11 ***
I(Lunghezza^2)           3.168e-02  7.272e-03   4.356 1.38e-05 ***
I(Gestazione * Lunghezza) -1.432e-01  1.097e-01  -1.306 0.191685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 266.8 on 2491 degrees of freedom
Multiple R-squared:  0.7421,    Adjusted R-squared:  0.7414
F-statistic:  1024 on 7 and 2491 DF,  p-value: < 2.2e-16
```

Model8.1 has an adjusted R^2 of 0.7414 vs 0.7388 of model8, increasing the quantity of Outcome variability explained by the independent variables. However, the p-values of the interaction term between *Gestazione* and *Lunghezza* is not significant anymore, therefore, it can be deleted:

```
> model8.2 = update(model8.1,~. -I(Gestazione*Lunghezza))
> summary(model8.2)

Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso + I(Lunghezza^2), data = newborn1)

Residuals:
     Min       1Q   Median       3Q      Max
-1176.79  -178.89   -12.52   164.19  1327.44

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.607e+03  7.586e+02  -2.119 0.034207 *
N.gravidanze   1.420e+01  4.220e+00   3.364 0.000781 ***
Gestazione     3.773e+01  3.890e+00   9.700  < 2e-16 ***
Lunghezza     -1.172e+01  3.342e+00  -3.508 0.000459 ***
Cranio         1.015e+01  4.201e-01  24.157  < 2e-16 ***
SessoM         7.222e+01  1.092e+01   6.614 4.57e-11 ***
I(Lunghezza^2) 2.330e-02  3.430e-03   6.795 1.35e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 266.8 on 2492 degrees of freedom
Multiple R-squared:  0.742,     Adjusted R-squared:  0.7413
F-statistic:  1194 on 6 and 2492 DF,  p-value: < 2.2e-16
```

Overall, the new model8.2 presents a slightly smaller adjusted R^2 than model8.1 (0.7413 versus 0.7414). Moreover, all the terms included have statistically significant p-values.

Model 8.2 still presents a slight level of heteroscedasticity, despite being lower than previous ones (Breush-Pagan test p-value of 0.0169 versus 5.782e-15of model8).

```
> shapiro.test(model8.2$residuals)

        Shapiro-Wilk normality test

data:  model8.2$residuals
W = 0.9899, p-value = 2.882e-12

> lmtest::bptest(model8.2)

        studentized Breusch-Pagan test

data:  model8.2
BP = 15.47, df = 6, p-value = 0.0169

> lmtest::dwtest(model8.2)

        Durbin-Watson test

data:  model8.2
DW = 1.9498, p-value = 0.1046
alternative hypothesis: true autocorrelation is greater than 0
```

Both AIC and BIC confirm the superiority of model8.2. Usually, AIC and BIC decrease when observations are added, because the more the sample grows, the more it gets similar to the reference population. In this case, despite an observation was removed, both criteria consequentially decrease, confirming the improvement in respect to its predecessors.

```
> AIC(model8,model8.2)
          df      AIC
model8     9 35065.62
model8.2   8 35023.10
```

```
> BIC(model8,model8.2)
          df      BIC
model8     9 35118.03
model8.2   8 35069.69
```

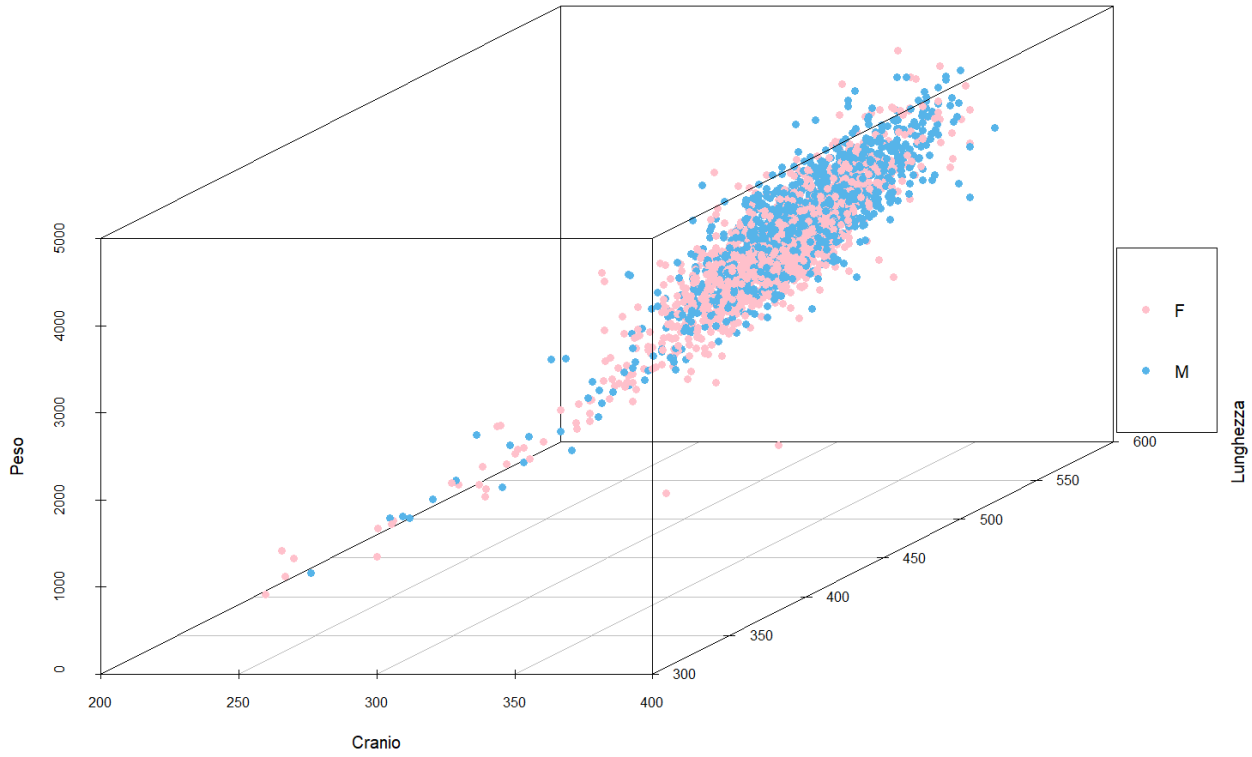
7) Predict the weight of a baby girl knowing that she will be born at 39 weeks of gestation, by a mother that is having her third gestation:

```
> NB = data.frame(N.gravidanze = 2, Gestazione=39,
+                 Lunghezza = mean(Lunghezza), Cranio = mean(Cranio),
+                 Sesso = "F")
> predict(model8.2,newdata = NB)
       1
3246.362
```

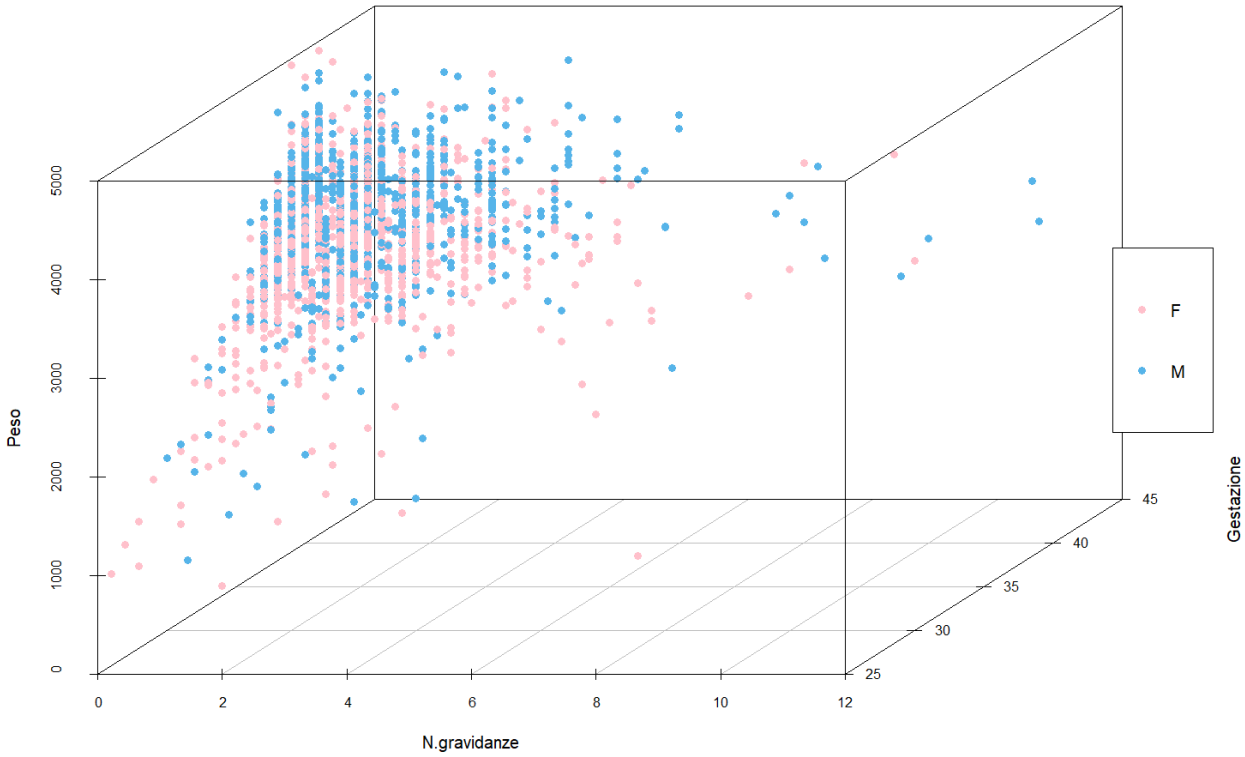
8) Graphical representations of the simplified model

Through the *scatterplot3d()* command of the *scatterplot3d()* package is possible to visualize the model. However, representing six variables within a two-dimensional space can not be performed without losing some information. Therefore, to elaborate a graphical representation as trustworthy as possible, four variables (the response and three predictors) are examined together:

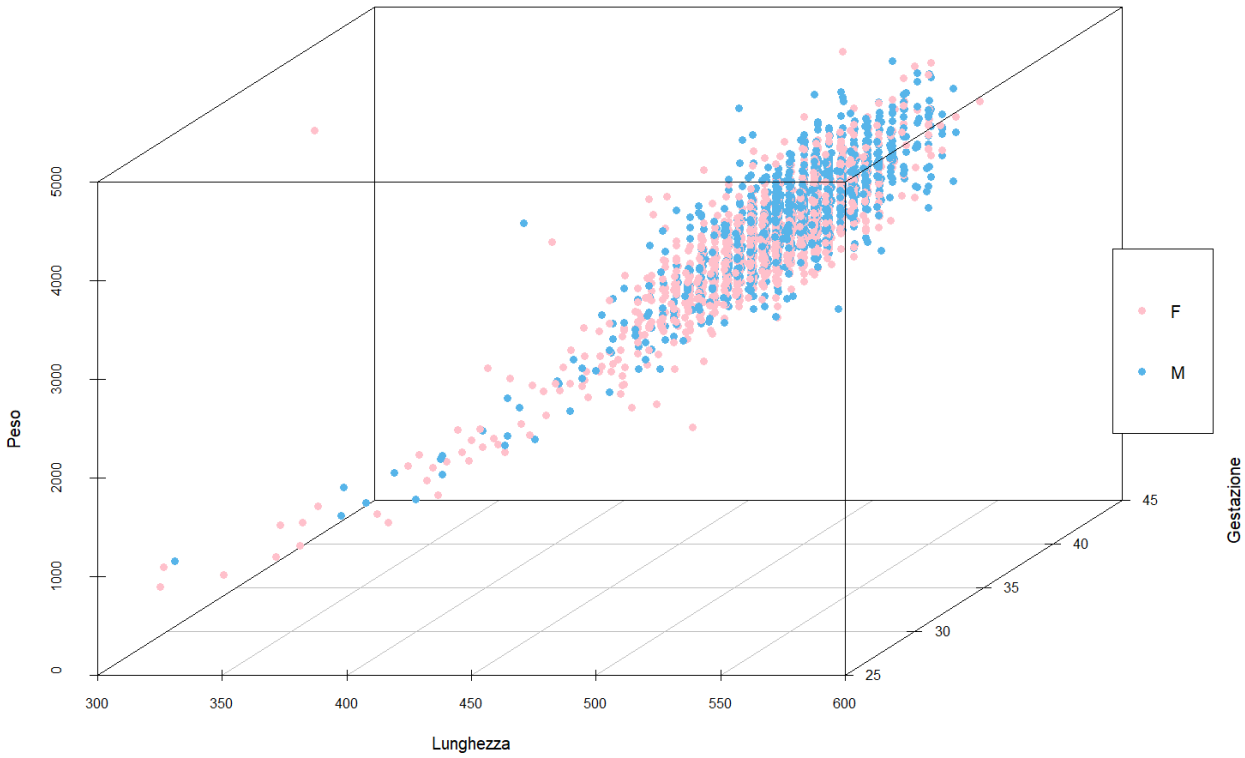
1)



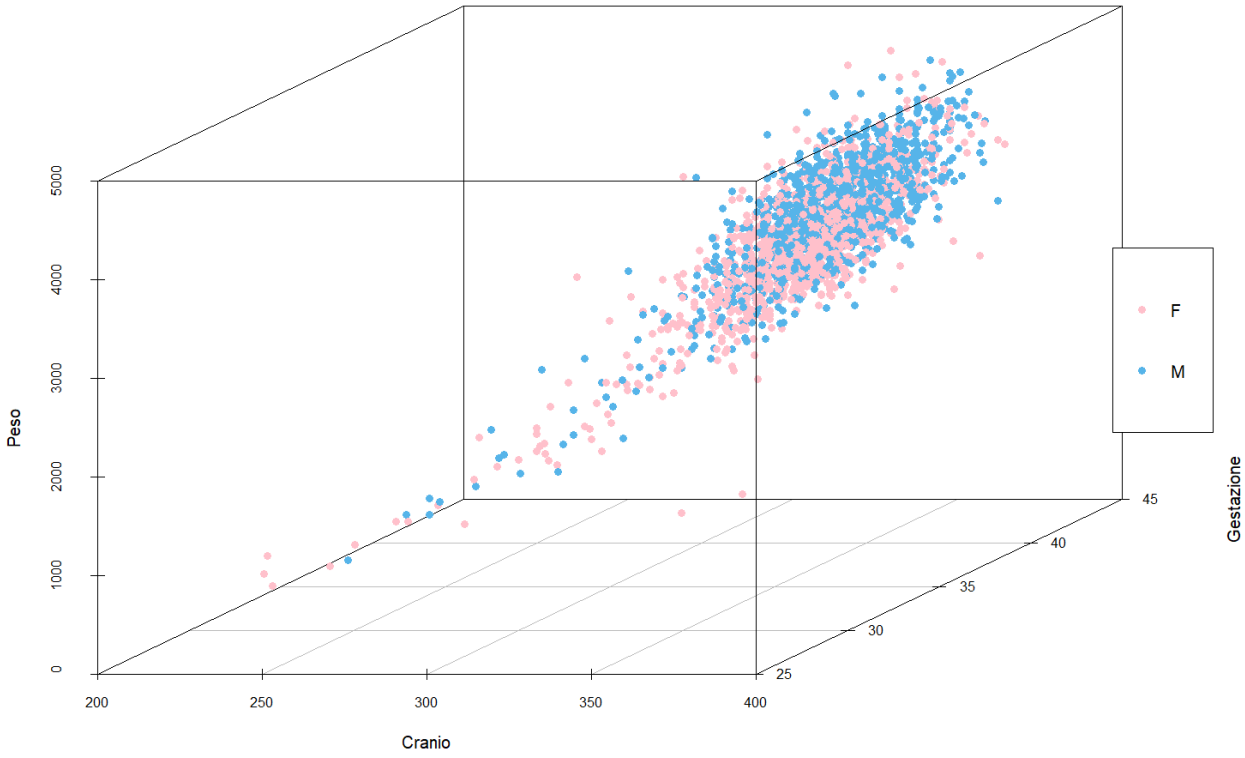
2)


Peso
0
1000
2000
3000
4000
5000
0
2
4
6
8
10
12
N.gravidanze
25
30
35
40
45
Gestazione
F
M

3)


Peso
0
1000
2000
3000
4000
5000
300
350
400
450
500
550
600
Lunghezza
25
30
35
40
45
Gestazione
F
M

4)


Peso
0
1000
2000
3000
4000
5000
Cranio
200
250
300
350
400
Gestazione
25
30
35
40
45
F
M