Project - Inferential Statistics

The following document will contain the Inferential Statistics project of Profession AI's Master of Science in Data Science.

Task 1 - Dataset and Import

The dataset covers 2500 newborns from 3 different hospitals.
The dataset consists of ten variables:

- Age of the mother
- Number of pregnancies sustained
- Smoking mother (0=NO, YES=1)
- No. of weeks of gestation
- Infant's weight in grams
- Length in mm of the infant
- Diameter in mm of the infant's skull
- Type of delivery: Natural or Caesarean
- Hospital: 1, 2, 3
- Sex of baby: M or F

The import of data into R was carried out using the following line of code:
data <- read.csv("newborn.csv")
which made it possible to read the entire dataset and place it within the R environment.

Task 2 - Description of the dataset and study objective

The dataset is explored using R's basic 'head' function, which allows us to display the rows and columns of the dataset. Below is an analysis of each variable in the dataset:

- Age of the mother: discrete quantitative variable
- Number of pregnancies sustained: discrete quantitative variable
- Smoking mother: coded qualitative variable (dummy)
- No. of weeks of gestation: discrete quantitative variable
- Infant weight in grams: continuous quantitative variable
- Length in mm of the newborn: continuous quantitative variable
- Diameter in mm of the infant's skull: continuous quantitative variable
- Type of delivery: qualitative variable on a nominal scale
- Hospital: coded qualitative variable (dummy)
- Gender of the newborn: qualitative variable on a nominal scale

Ultimately, there are 3 discrete quantitative variables, 3 continuous quantitative variables, 2 qualitative variables on a nominal scale and 2 qualitative variables coded (dummy).

The study aims to find out whether there is a relationship between the mother's lifestyle and the birth of the newborn. Based on this relationship, go on to predict the weight of the newborn baby at birth.

Task 3 - Descriptive dataset analysis

A descriptive data analysis was carried out on all variables in the dataset in order to understand their behaviour before moving on to the development of the statistical model.

Variable - Anni.madre

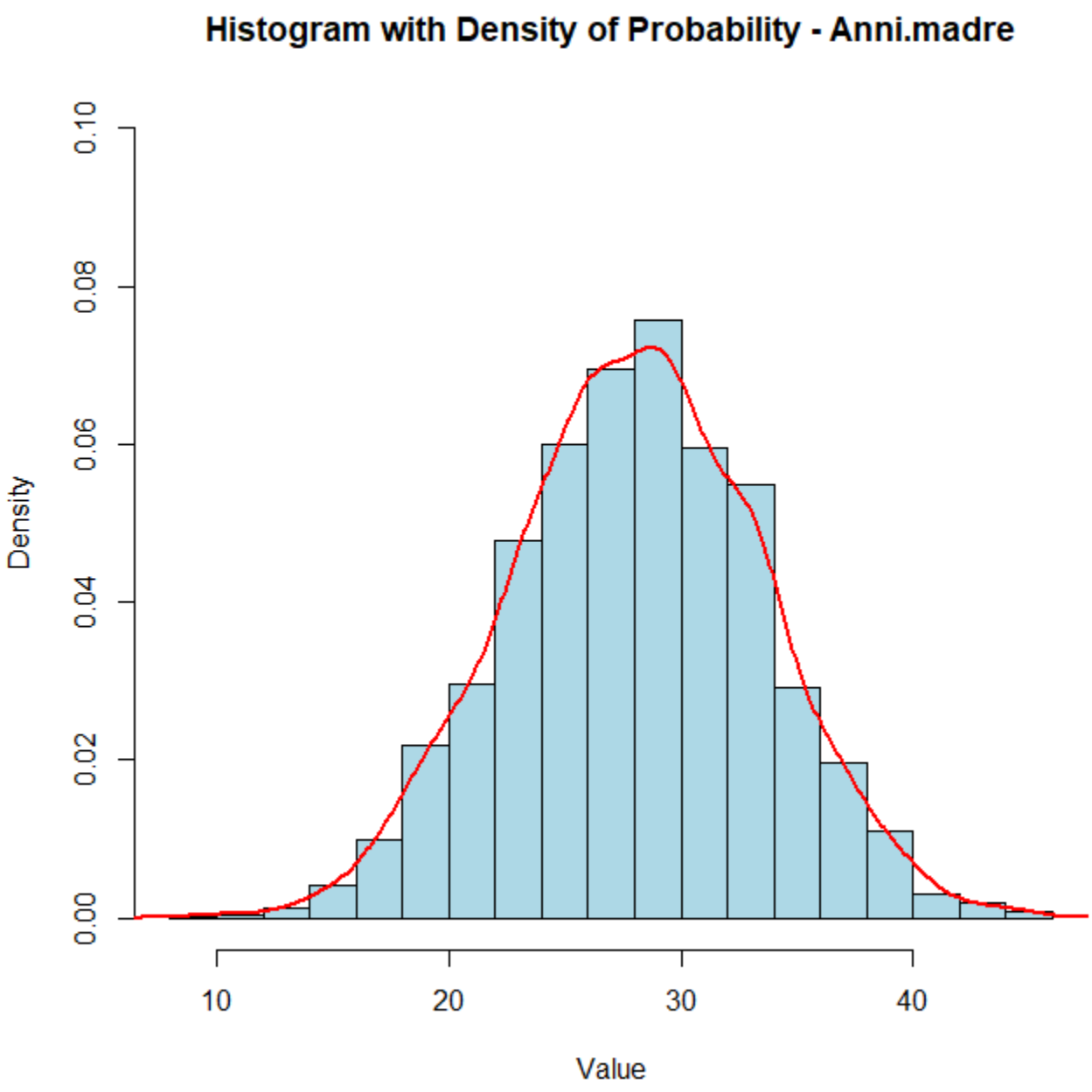
The average age of the mothers is 38 years.

The standard deviation is 5, so there is a variability on average of 5 years among the mothers surveyed in the dataset.

The probability density function shows an almost normal pattern.

The shape indices indicate a positive skew and leptokurtic distribution.

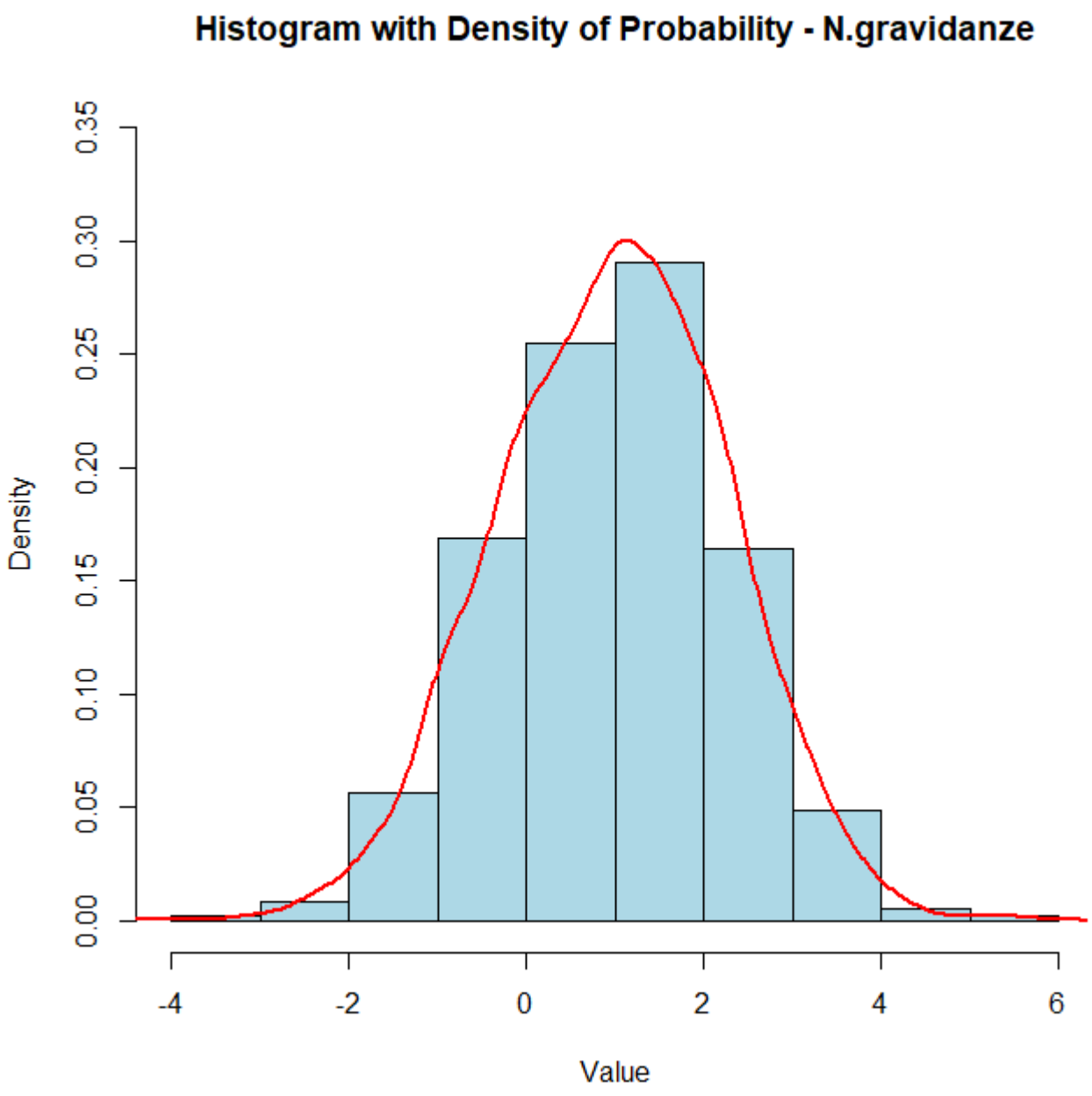
Histogram and density function of the variable Anni.madre



Variable - No. of pregnancies

The average number of pregnancies of the mothers is 1.
The standard deviation is 1, so there is a variability of 1 pregnancy on average among the mothers in the dataset.
The probability density function shows a trend not quite like a standard normal as the mean deviates from 0.
The shape indices indicate a positive skew and leptokurtic distribution.

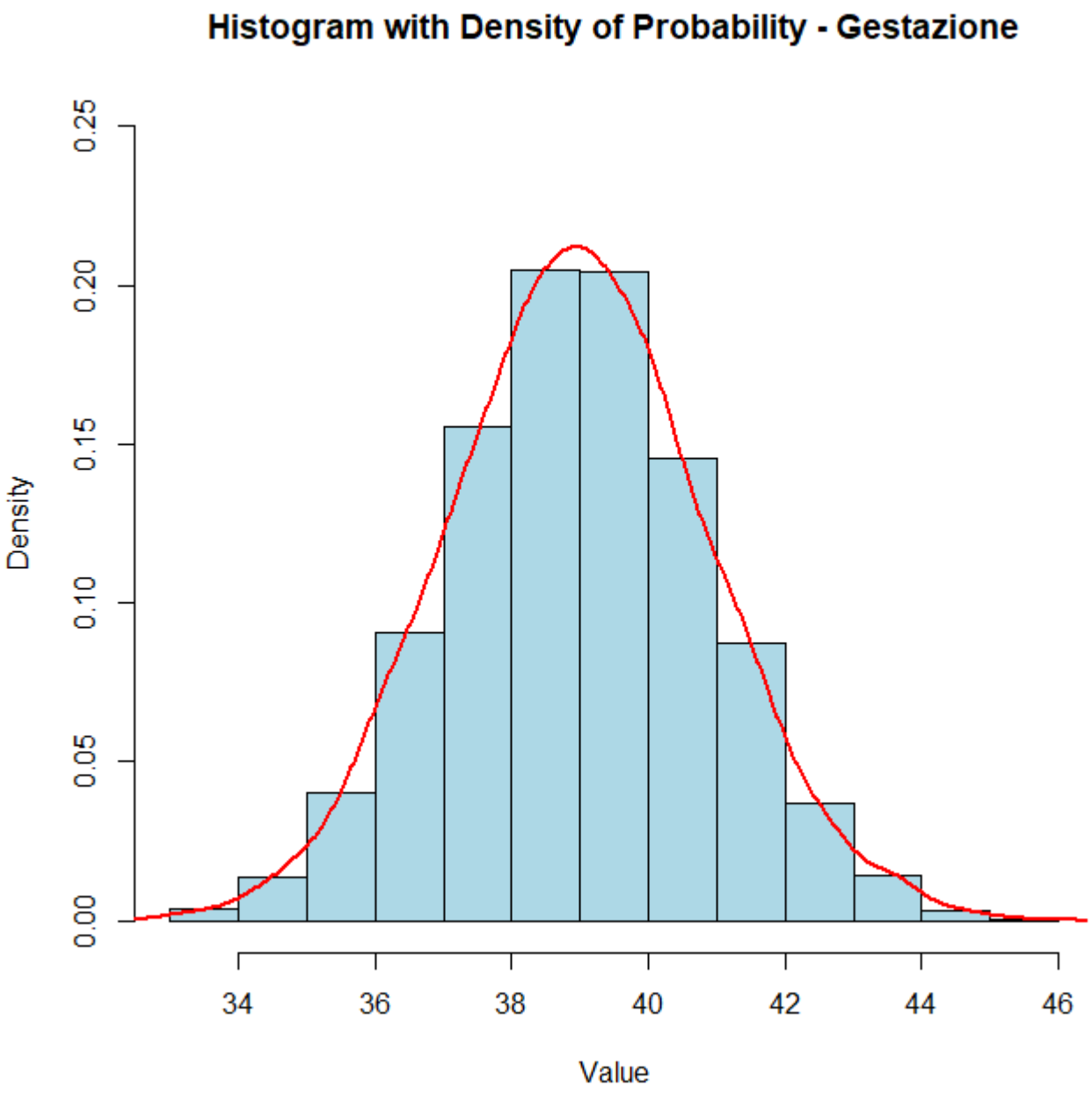
Histogram and density function of the variable No. pregnancies



Variable - Gestation

The
average weeks' gestation of the mothers is 39.
The standard deviation is 1.8, so there is a variability of 1.8 months' gestation on average among the mothers surveyed in the dataset.
The probability density function shows two dips in the centre and this could indicate that there is an outlier value at that point. The
shape indices indicate a negative
asymmetric and leptokurtic
distribution.

Histogram and density function of the variable Gestation



Variable - Weight

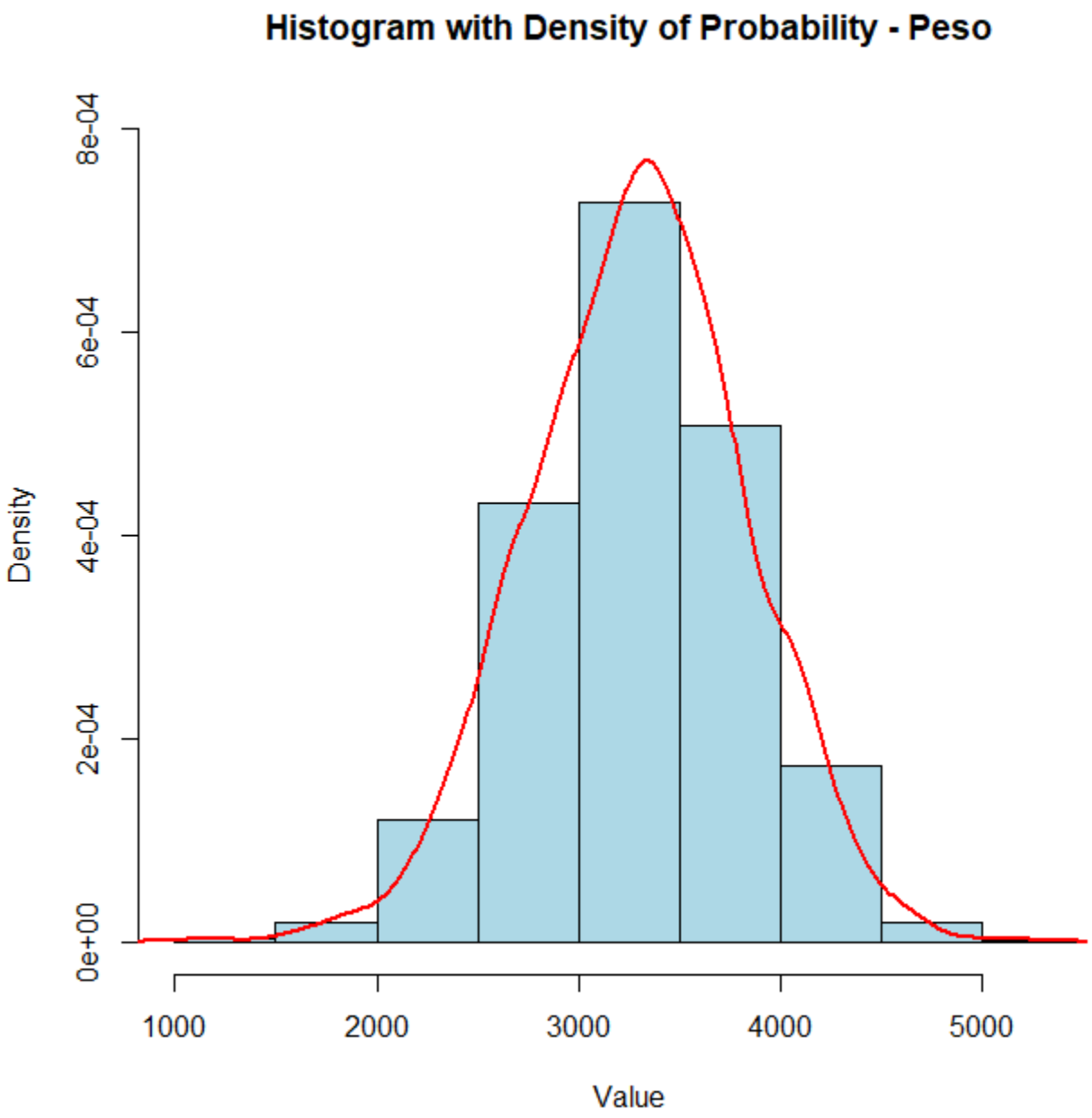
The average weight of the newborns is 3284g.

The standard deviation is 525, so there is a variability of an average of 525g in weight among the newborns surveyed in the dataset.

The probability density function shows no peculiarities that could particularly influence the study of the model.

The shape indices indicate a negative asymmetric and leptokurtic distribution.

Histogram and density function of the variable Weight



Variable - Length

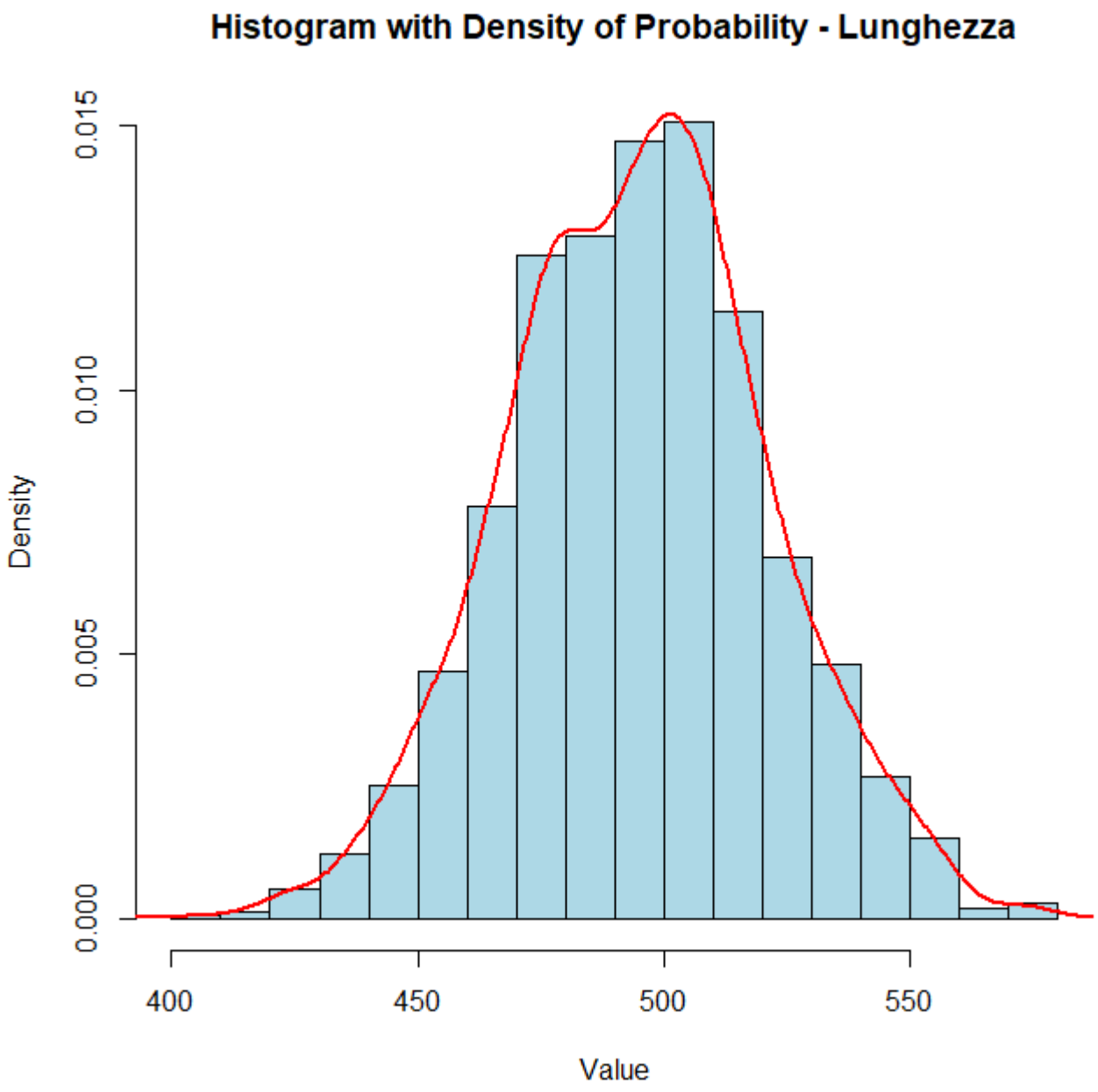
The average length of the newborns is 497.7 mm.

The standard deviation is 26, so there is a variability on average of 26mm in length among the newborns surveyed in the dataset.

The probability density function shows a very narrow wedge, and the values tend to flood much more than the average.

The shape indices indicate a negative asymmetric and leptokurtic distribution.

Variable density function Length



Variable - Skull

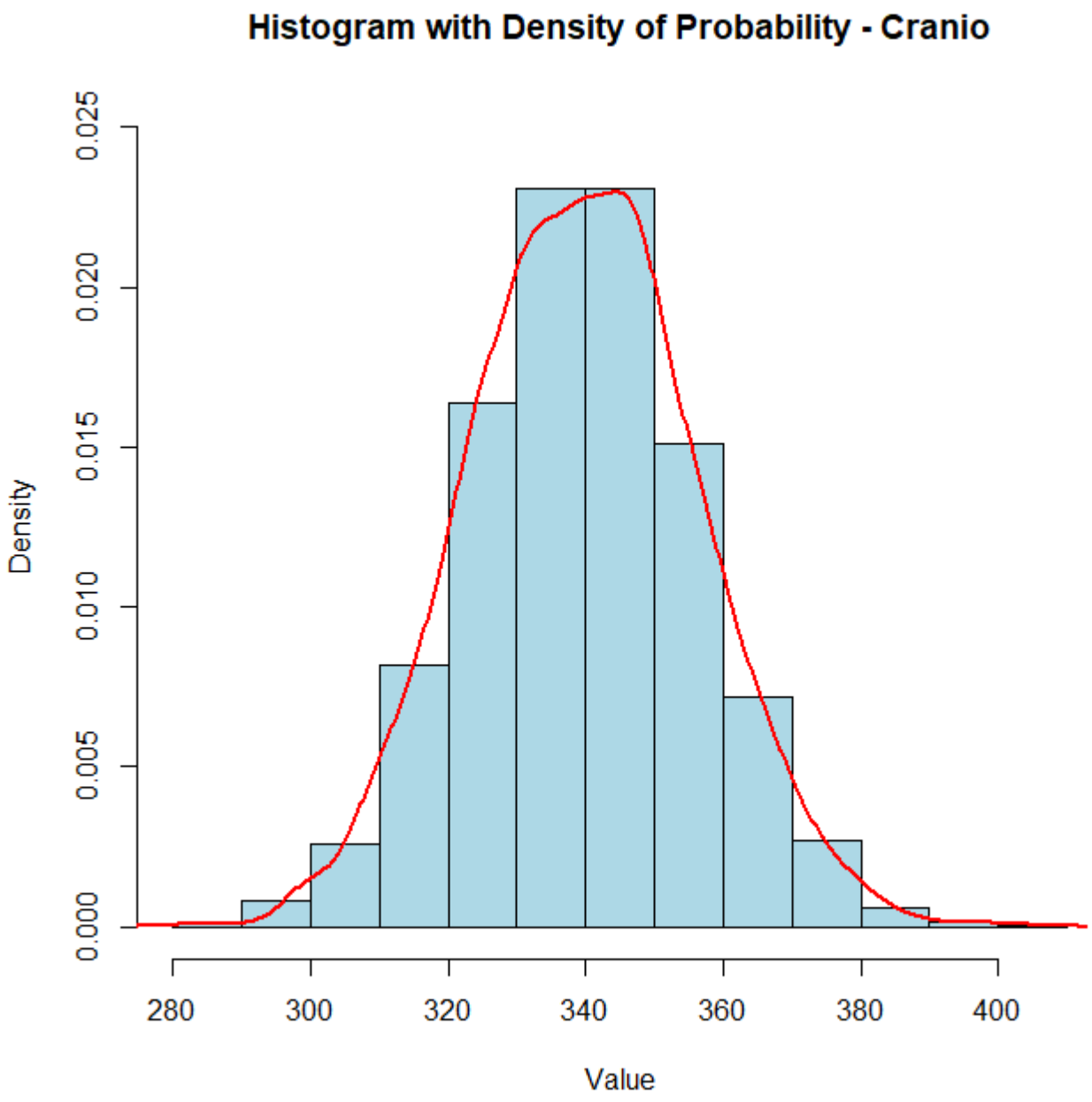
The average skull diameter of the newborns is 340mm.

The standard deviation is 16, so there is a variability of an average of 16mm in skull diameter among the infants surveyed in the dataset.

The probability density function shows a double wedge in the centre and this could indicate outliers in the dataset.

The shape indices indicate a negative asymmetric and leptokurtic distribution.

Histogram and density function of the variable Skull

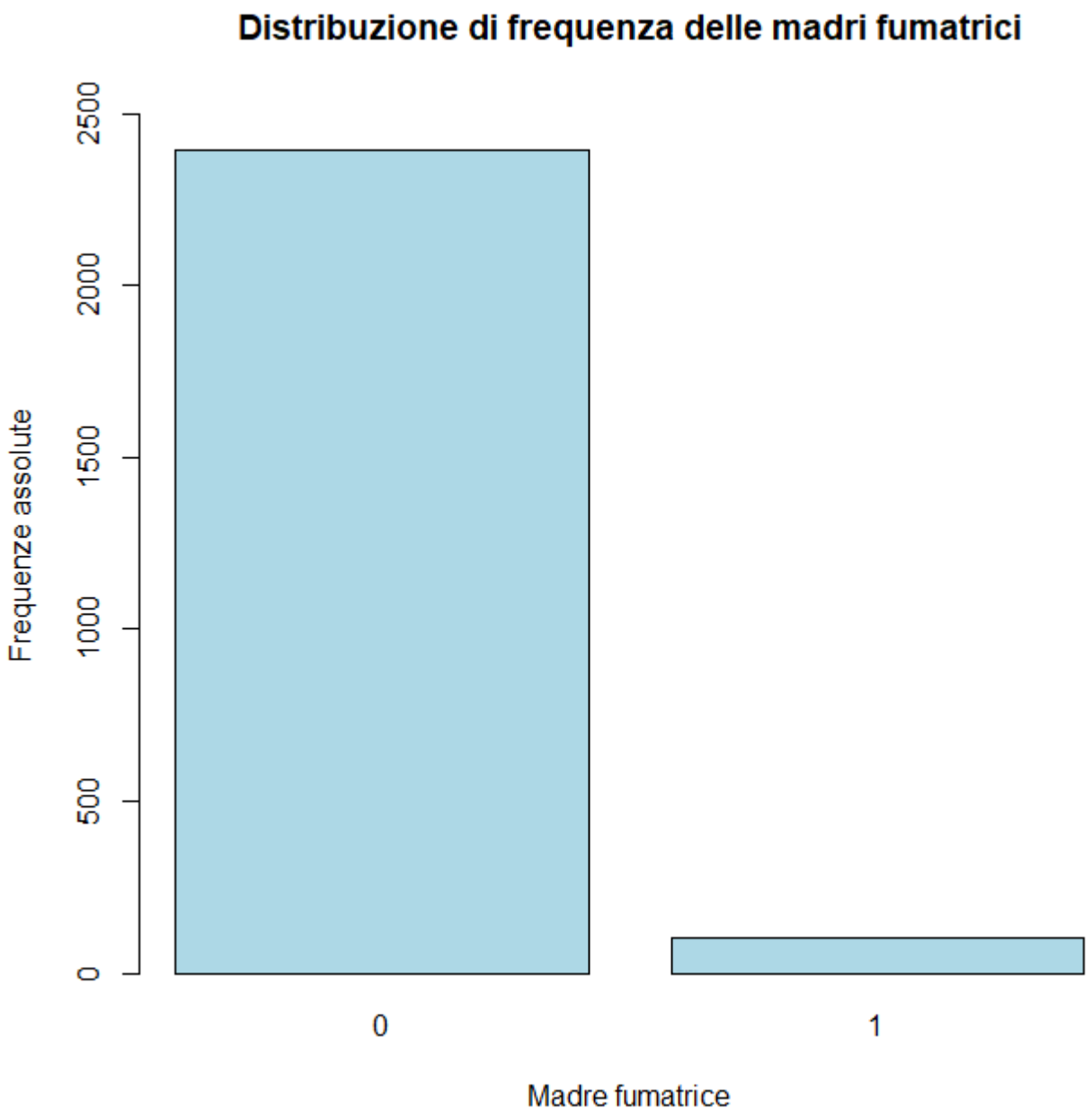


Variable - Smokers

The smoking variable being qualitative, the absolute frequencies were displayed.

The graph shows that most of the mothers surveyed in the dataset are non-smokers.

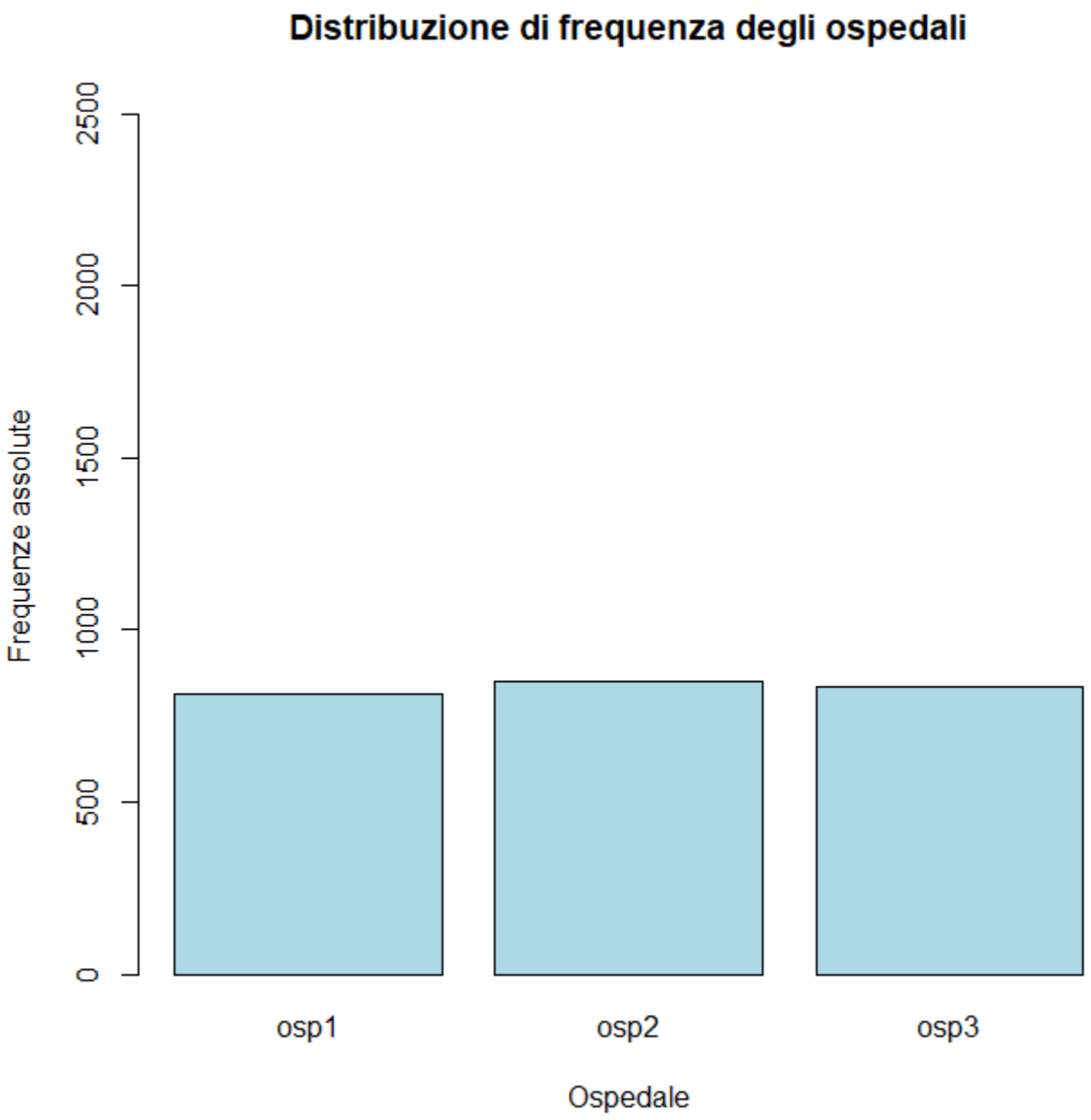
Absolute frequencies of the smoking variable



Variable - Hospital

The hospital variable being qualitative, absolute frequencies were displayed. It can be seen from the graph that the variable shows an almost equidistribution between its modes.
(> gini.index(Hospital) = 0.9998683)

Absolute frequencies of the variable hospital



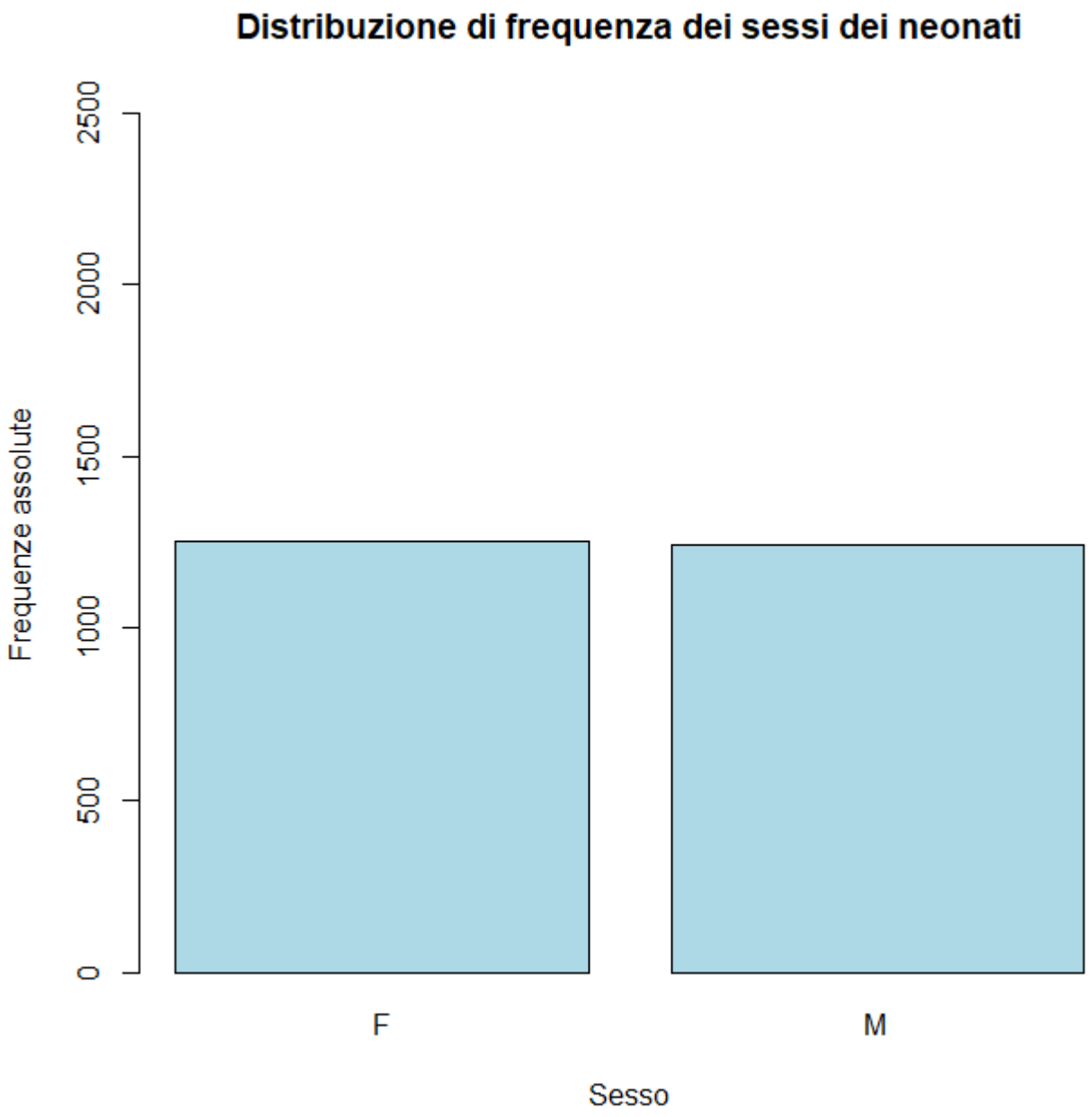
Variable - Gender

The sex variable being qualitative, the absolute frequencies were displayed.

It can be seen from the graph that the variable shows an almost equidistribution between its modes.

(> gini.index(Sex) = 0.999977)

Absolute frequencies of the variable hospital



Task 4 - Testing the hypothesis of the average weight and length of the sample in relation to the population

The population average for weight is around 3300g, while for length it is around 50cm according to ISTAT.

Having retrieved this information, we proceeded with the t-student test to test the hypothesis that the sample weight and height averages of the infants in the dataset were significantly equal to the population averages.

Tests were set by giving the variables Weight and Length as input, the alpha significance level of 0.95 was chosen, and the option to test for both tails was chosen to identify a confidence interval.

Below are the results:

- Weight
    - The p-value turns out to be greater than the chosen significance level of 5% (0.05) and for this reason the null hypothesis is not rejected. The sample mean turns out to be significantly equal to the population mean and can be stated with a confidence level of 95.

```
        One Sample t-test

data:  Peso
t = -1.516, df = 2499, p-value = 0.1296
alternative hypothesis: true mean is not equal to 3300
95 percent confidence interval:
 3263.490 3304.672
sample estimates:
mean of x
 3284.081
```

- Length
    - The p-value turns out to be lower than the chosen significance level of 5% (0.05) and for this reason the null hypothesis is rejected in favour of the alternative hypothesis. The sample mean is significantly different from the population mean.

```
        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
```

Task 5 - Statistically significant differences between males and females

It is possible to identify statistically significant differences between the two sexes, male and female, both for the variables length and weight, but also for skull size. To carry out this study, the sample was divided into two groups: males and females, in order to conduct a t-test for independent samples.

Below are the results:

- Length
    - The p-value is very close to zero and lower than the chosen significance level of 5% (0.05). For this reason, we reject the null hypothesis in favour of the alternative hypothesis, as we can state with 95% confidence that on average the length between male and female infants is

significantly different.

```
        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 not equal to 0
95 percent confidence interval:
 -11.929470  -7.876273
sample estimates:
mean in group F mean in group M
       489.7643        499.6672
```

- Weight
  - The p-value is very close to zero and lower than the chosen significance level of 5% (0.05). For this reason, the null hypothesis is rejected in favour of the alternative hypothesis, being able to state with 95% confidence that on average the weight between male and female infants is significantly different.

```
        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 not equal to 0
95 percent confidence interval:
 -287.1051 -207.0615
sample estimates:
mean in group F mean in group M
       3161.132        3408.215
```

- Skull
  - The p-value is very close to zero and lower than the chosen significance level of 5% (0.05). For this reason, the null hypothesis is rejected in favour of the alternative hypothesis, since it can be stated with 95% confidence that on average the diameter of the skull between male and female infants is significantly different.

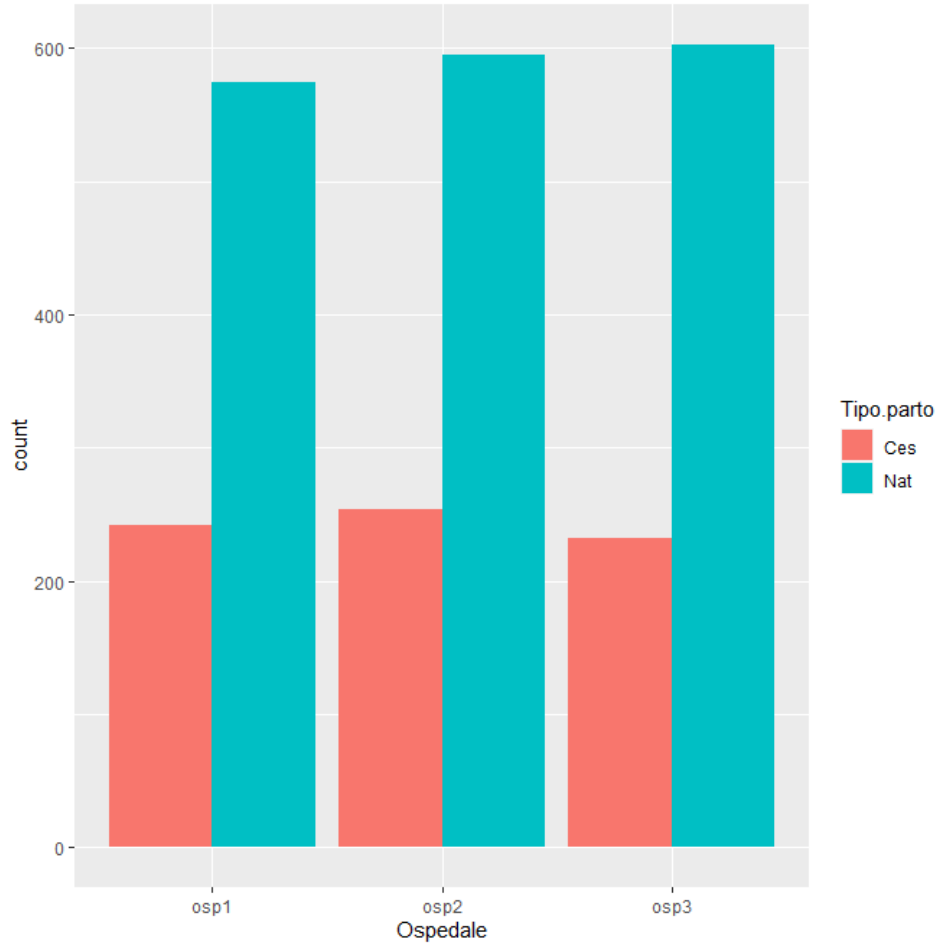
```
        Welch Two Sample t-test

data:  Cranio by Sesso
t = -7.4102, df = 2491.4, p-value = 1.718e-13
alternative hypothesis: true difference in means between group F and group M is not equal to 0
95 percent confidence interval:
 -6.089912 -3.541270
sample estimates:
mean in group F mean in group M
       337.6330        342.4486
```

Task 6 - Check in which hospital more caesarean sections are performed

In order to identify the hospital in which the most caesarean sections were performed, we opted to use the ggplot2 library and create a side-by-side bar graph in order to graphically identify in which hospital the most caesarean sections were performed.

The variable Hospital as abscissa and the variable Type.delivery as fill, and finally the position option equal to dodge to flank the bars.

Below is the result:



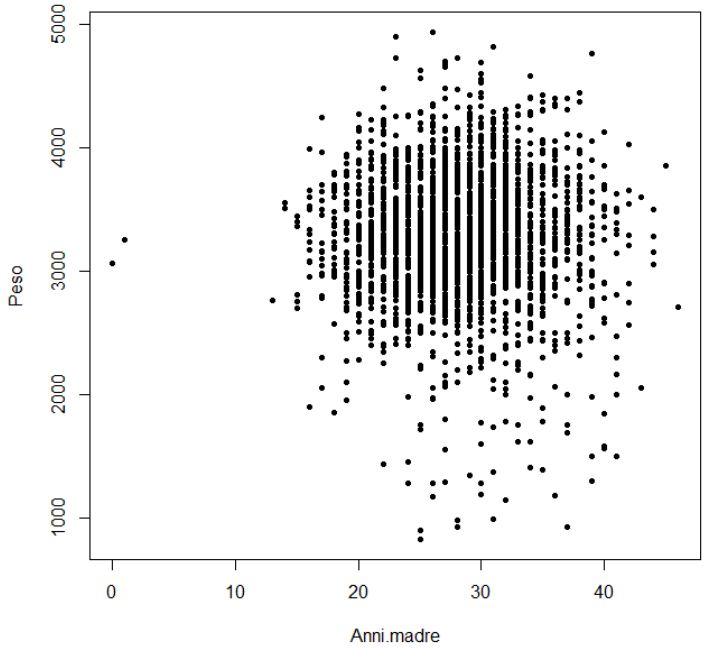
It can be said that more caesarean sections are performed in Hospital 2.

Task 7 - Multidimensional analysis: checking two-by-two relationships between variables

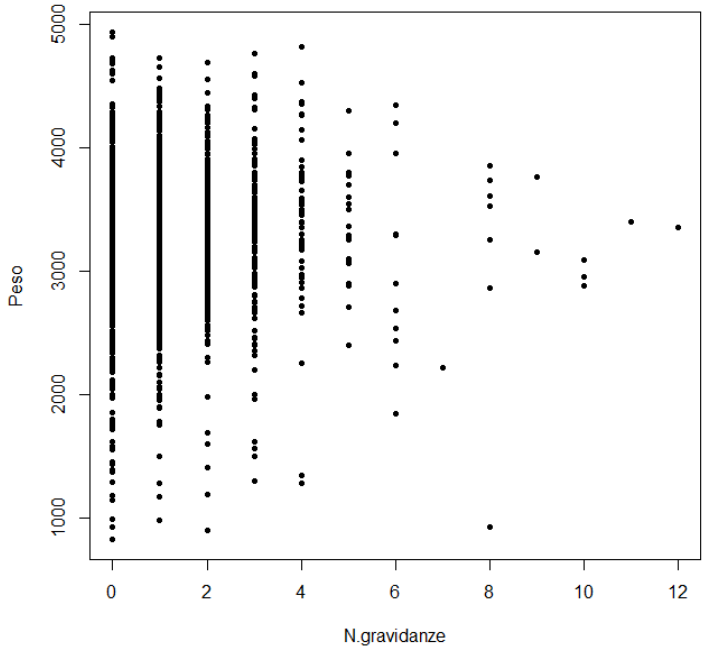
The aim of the study is to find out whether it is possible to make predictions about the weight of newborns at birth based on the lifestyle of the mothers. Before proceeding to develop the model to make the predictions, the correlations between the explanatory variables with the response variable were studied. In particular, it was deemed appropriate not to study the correlation between the variable Hospital with the response variable Weight, as this does not make any sense. Therefore, we proceeded with the correlation study between all the remaining explanatory variables and the response variable.

Below are the results:

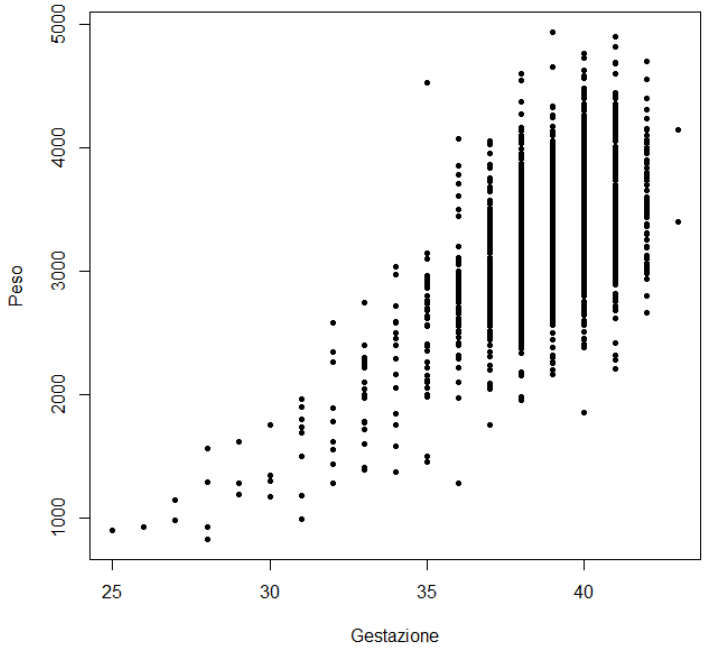
- **Mother Years** with **Weight has a** non-parametric Spearman's correlation coefficient with a p-value of 0.7648, thus exceeding the 5% significance level. This result indicates that the null hypothesis of a correlation of 0 cannot be rejected, but that the correlation is not significant or at least very weak. Furthermore, plotting the two variables in a scatterplot shows no pattern between the two variables. In particular, mother years being a time variable and in some respects important, it could be included as a control variable.



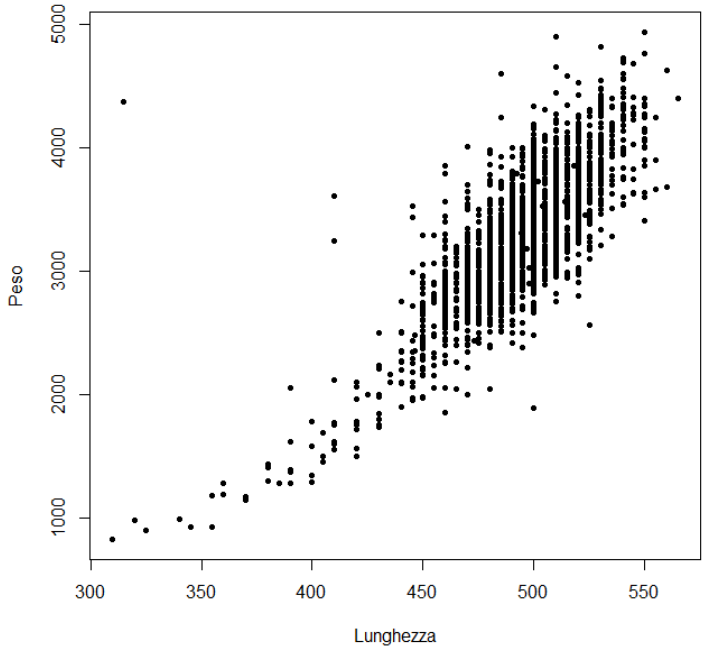
- **N.Pregnancies** with **Weight** has a non-parametric Spearman's correlation coefficient with a p-value of 0.4037, thus exceeding the 5% significance level. This result indicates that the null hypothesis of a correlation of 0 cannot be rejected, but that the correlation is not significant or at least very weak. Furthermore, plotting the two variables in a scatterplot shows no pattern between the two variables.



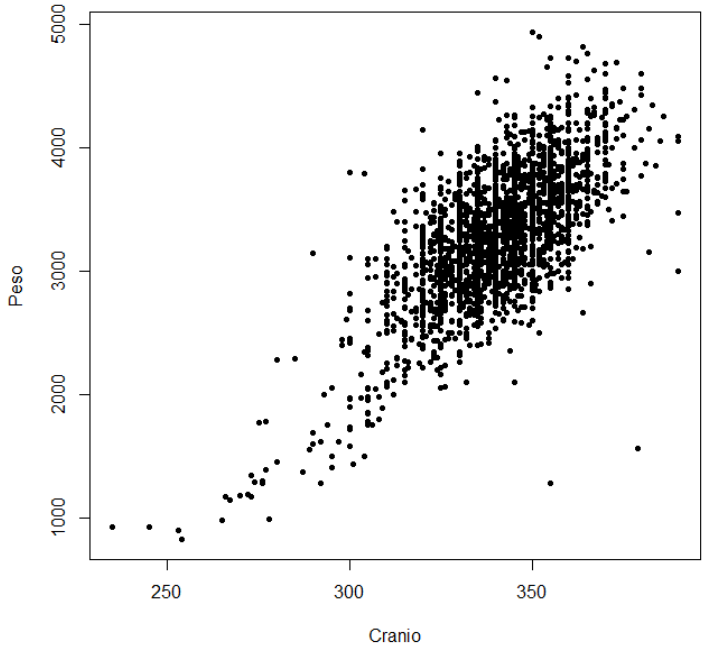
- **Gestation** with **Weight has a** non-parametric Spearman's correlation coefficient with a p-value of 2 at -16, thus well below the 5% significance level. This result indicates that the null hypothesis of a correlation of 0 can be rejected and that the correlation is significant and strong between the two variables. Furthermore, plotting the two variables in a scatterplot shows no pattern between the two variables.



- **Length** with **Weight** has a Pearson's linear correlation coefficient of 0.7960. This result indicates a strong positive linear correlation between the two variables. Furthermore, plotting the two variables in a scatterplot shows possible patterns between the two variables.



- **Skull** with **Weight** has a Pearson's linear correlation coefficient of 0.7048. This result indicates a moderate positive linear correlation between the two variables. Furthermore, plotting the two variables in a scatterplot shows possible patterns between the two variables.



- **Sex** with **Weight** results in a Pearson chi-square independence test, a p-value of 2 at -16. This result indicates that the null hypothesis of independence is rejected in favour of the alternative hypothesis of dependence between the two variables.
- **Smoking** with **Weight** results in a Pearson chi-square independence test with a p-value of 0.0995. This result indicates that the null hypothesis of independence is not rejected in favour of the alternative hypothesis of dependence between the two variables.

Task 8 - Creation of multiple linear regression model

Before proceeding with the creation of the multiple linear regression model, it was considered appropriate to study whether the response variable behaves normally or at least similar to a normal one. For this, the two shape indices Fisher and Kurtosis were studied.

Fisher's index is -0.6470 and this indicates a negative skewed distribution. The Kurtosis index is 2.031 and this indicates a leptokurtic distribution. Finally, in order to verify that the response variable behaves normally, the Shapiro-Wilk normality test was performed, which resulted in a p-value of 2 at -16. Therefore, the null hypothesis of normality is not rejected and the response variable behaves normally.

At this point, the first model was created, including all variables as required by the task.

Below are the results:

- The variable **Anni.madre,** as also found in the correlation study with the variable Weight, has a positive beta coefficient of approximately 0.9, which would mean a marginal effect on average of 0.9 of weight. The p-value is 0.430, which indicates that the variable Anni.madre has no significant effect on the response variable.
- The variable N.**pregnancies has** a positive beta coefficient of 11.2, which would mean that for every unit change in the variable N.pregnancies, there would be a change of approximately 11.2g in Weight. The p-value is 0.0157, so just below the 5% significance threshold, which is why it makes sense to include the variable within the model.
- The variable **Smokers,** as also found in the study of the correlation with the variable Weight, has a negative beta coefficient of -30, which would mean a marginal effect on average of -30 of weight. The p-value is 0.27, which indicates that the variable Smokers has no significant effect on the response variable.
- The variable **Gestation has** a positive beta coefficient of 32.56, which would mean that for every unit change in the variable Gestation, there would be a change of approximately 32.56g in Weight. The p-value is 2 at -16, thus well below the 5% significance level, which is why it makes sense to include the variable within the model.
- The variable **Length has** a positive beta coefficient of 10.29, which would mean that for every unit change in the variable Length, there would be a change of approximately 10.29g in Weight. The p-value is 2 at -16, thus well below the 5% significance level, which is why it makes sense to include the variable within the model.
- The variable **Cranio has** a positive beta coefficient of 10.47, which would mean that for every unit change in the variable Cranio, there would be a change of approximately 10.47g in Weight. The p-value is 2 at -16, thus well below the 5% significance level, which is why it makes sense to include the variable within the model.

- The variable **Type.delivery has** a positive beta coefficient of 29.52, which would mean that compared to the chosen baseline, i.e. natural childbirth, the weight of infants born by caesarean section are on average 29.52g heavier. The p-value is 0.0146, thus just below the 5% significance threshold, which is why it makes sense to include the variable within the model.
- The variable **Hospital,** despite presenting results that would indicate the possibility of further explaining the response variable, it does not make sense to include it in the model since, in real terms, birth in a hospital rather than another one cannot influence the weight of the infant.
- The variable **Sex**, has a positive beta coefficient of 77.54, this would mean that compared to the chosen baseline, i.e. male birth, the weight of male newborns is on average 77.54g more than that of females. The p-value is 5.008 at -12, thus well below the 5% significance level, which is why it makes sense to include the variable within the model.

```
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
```

With regard to the goodness of the model and thus how well it fits the data, the coefficient R square is 0.72, thus a fair model for making predictions about the weight of newborns.

Task 9 - Selecting the best model

The best model was chosen using the Step-Wise procedure. First, the car library was used to perform the Step-Wise procedure automatically with the mixed methodology. Subsequently, the Step-Wise procedure was performed, but in a backward and manual manner, eliminating those variables that were not considered significant and useful to the model.

The automatic mixed step-wise procedure produced as an output a model including all variables outside the variable Smokers, thus also including Hospital and Anni.madre. Analysing the result, the variable Hospital, although having a p-value just under the 5% threshold for the mode osp3 as a baseline compared to the others, does not make sense to keep it as it cannot influence the weight of the newborns in reality. As for Anni.madre, it too has a high p-value and therefore has no significant effect on the weight of the newborns.

For these reasons, the variables were removed, one-by-one, and the corresponding models evaluated. The models examined all have the same R-square, in particular, some vary by 0.01 compared to the others, thus a negligible and non-significant variation. The choice, therefore, of the best model was made following Occam's razor principle, which leads to selecting what is 'simplest'.

The 'simplest' model with a good compromise between explanatory variables and discrete R2 was the model including the following variables:

- No. pregnancies
- Gestation
- Sex
- Skull
- Length

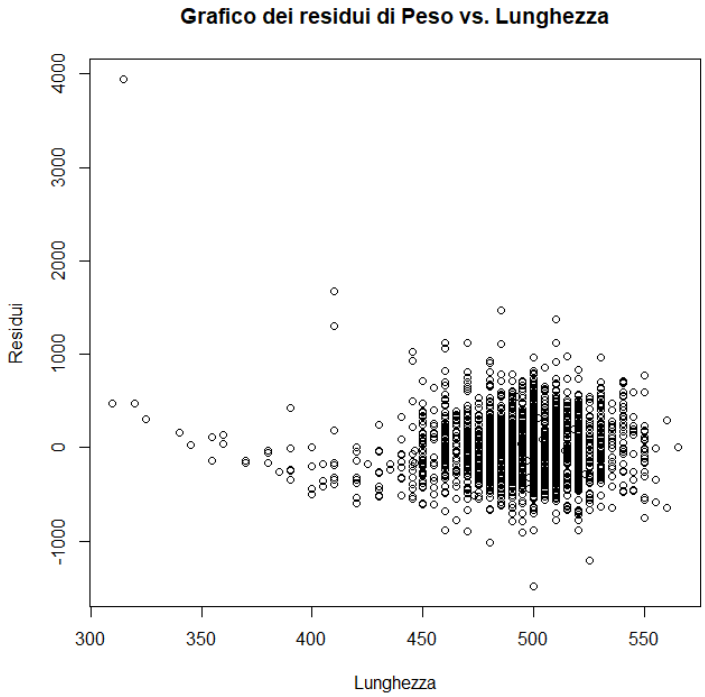
```
Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
    Sesso, data = data)

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
```

Task 10 - Study of interaction effects and non-linearity

Analysing the scatterplots of the variables Skull and Length in relation to Weight, one can suspect and hypothesise a possible non-linear correlation as the point clouds present a positive slope and in addition seem to have a kind of 'curve' in the initial part. As far as interaction effects are concerned, a significant effect is suspected between skull size and infant length.

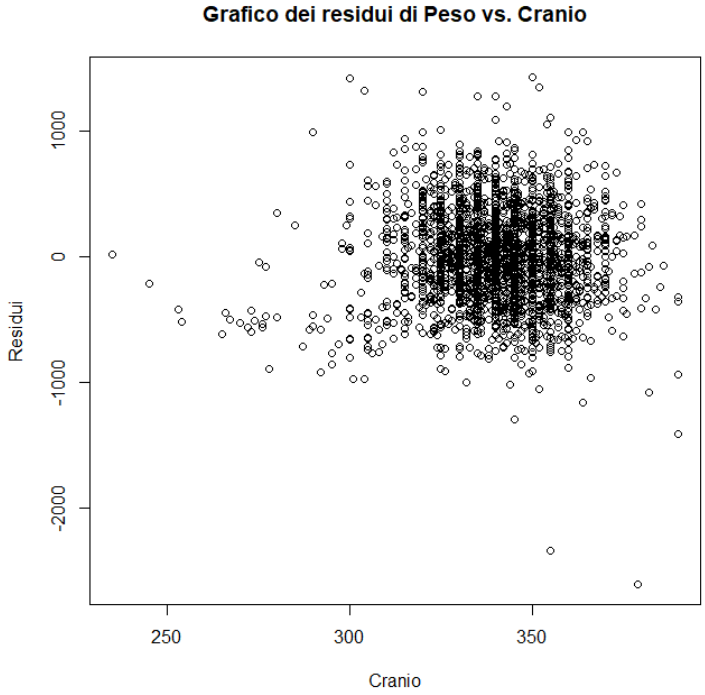
To test both hypotheses, the following steps were taken:

- The interaction effect between length and skull was studied on the basis of the fact that infants who have a moderate length and a moderate skull size are consequently heavier. For this reason, the pair of variables multiplied by each other was included in the model. The result, considering the model without interaction, an insignificant increase in the adjusted R-square of 0.002.
- The non-linearity for the variable Length was studied by trying to create a linear model between only Weight and Length, check that the residuals do not tend to disperse and create a model that takes into account the non-linear effect of the variable Length. Finally, an anova test was performed to check which of the two models tends to explain the variable Weight better. The result was that the non-linear model has a very small p-value for the F-test, this indicates a statistically significant difference between the two models in terms of the explained response variable, which is why the variable Length squared was included within the model. The adjusted R-square, with the inclusion of Length squared, had a discrete increase of 0.01.




```
Analysis of Variance Table

Model 1: Peso ~ Lunghezza
Model 2: Peso ~ Lunghezza + I(Lunghezza^2)
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   2498 252357422
2   2497 250052855  1   2304568 23.013 1.704e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The non-linearity for the variable Cranio was studied in the same way as for the variable Length, but reporting a negative result in terms of the explained response variable, and for this reason, the variable Cranio was included within the model in linear form.




```
Analysis of Variance Table

Model 1: Peso ~ Lunghezza
Model 2: Peso ~ Cranio + I(Cranio^2)
  Res.Df       RSS Df Sum of Sq F Pr(>F)
1   2498 252357422
2   2497 340614653  1 -88257230
```

In conclusion, considering Occam's razor principle and the adjusted R-square, including interaction and non-linearity effects, the goodness of the model varies very little, which is why it was considered more appropriate to include the variable Length with its non-linear effect and to keep the model with the same variables as identified in Task 9.

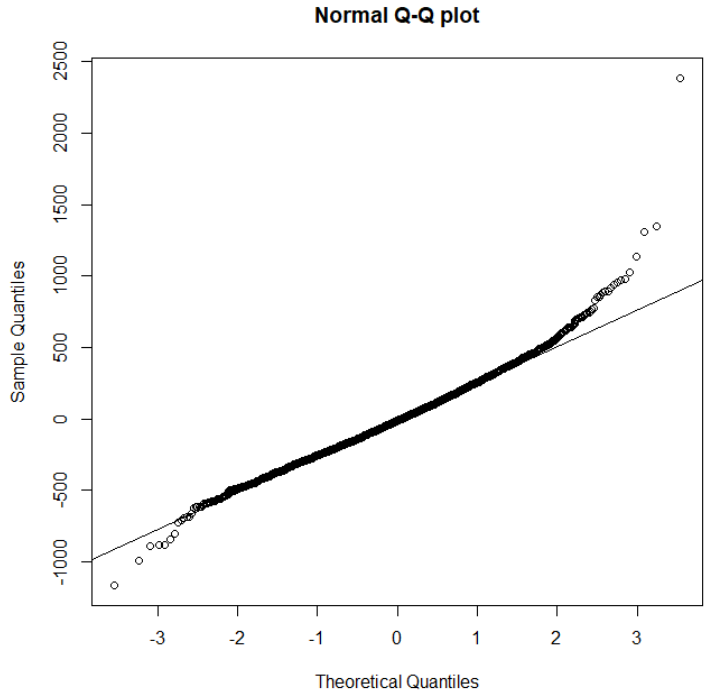
```
Call:
lm(formula = Peso ~ N.gravidanze + Gestazione + I(Lunghezza^2) +
    Cranio + Sesso, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1161.52  -179.94   -11.43   165.93  2382.49

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.333e+03  1.441e+02 -30.075  < 2e-16 ***
N.gravidanze    1.316e+01  4.296e+00   3.063  0.00222 **
Gestazione      3.476e+01  3.710e+00   9.368  < 2e-16 ***
I(Lunghezza^2)  1.082e-02  3.073e-04  35.192  < 2e-16 ***
Cranio          1.046e+01  4.210e-01  24.856  < 2e-16 ***
SessoM          7.457e+01  1.109e+01   6.721 2.22e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 271.8 on 2494 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.732
F-statistic:  1366 on 5 and 2494 DF,  p-value: < 2.2e-16
```

Task 11 - Residue Diagnostics

Residue diagnostics were performed on the final model chosen and described within the previous task, which includes the variables: N. Pregnancies, Gestation, Length squared, Skull and Sex. First, a preliminary graphical analysis was carried out, followed by the use of specific statistical tests.

Below are the results of the diagnostics:

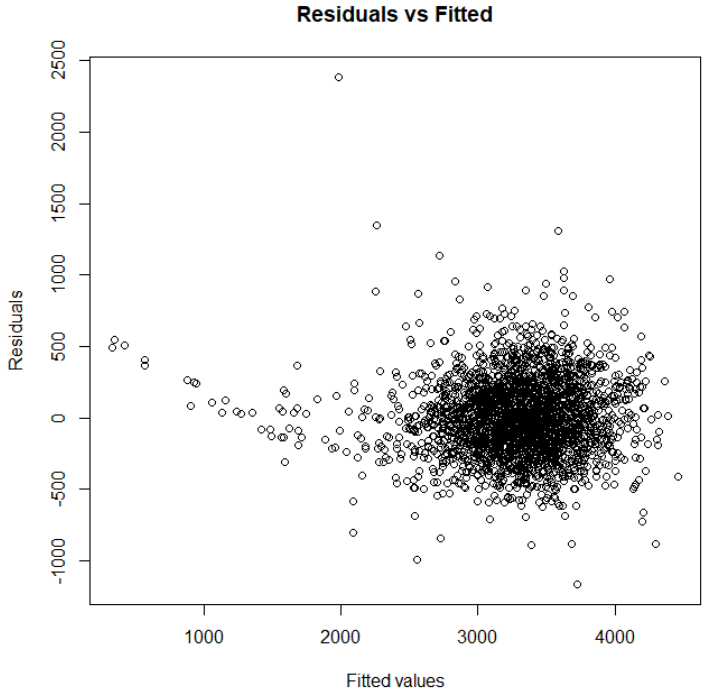
- **Shapiro-Wilk** test for residue normality
    - From the **Normal Q-Q plot, the** points lie roughly on the entire bisector of the graph; therefore, the residuals seem to behave normally.



    - P-value equal to 2 at -16, thus practically equal to 0. This result indicates that the null hypothesis of normality of the residuals is rejected.

```
        Shapiro-Wilk normality test

data:  residuals(mod_final)
W = 0.97831, p-value < 2.2e-16
```

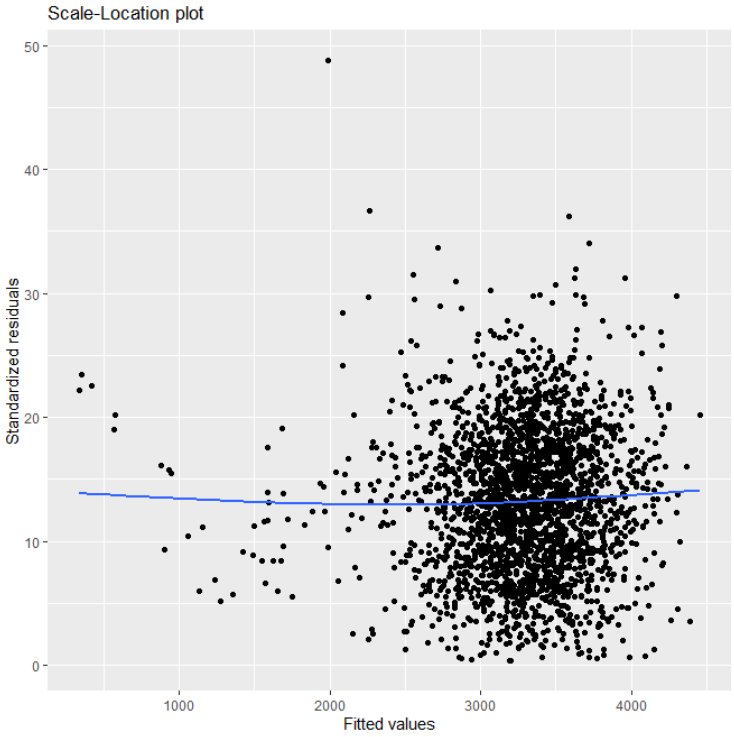
- **Breusch-Pagan** test for homoschedasticity of residuals (constant variance):
    - From the "**Residuals vs Fitted**" graph, the points seem to be scattered around the mean of 0 and thus, there would appear to be no heteroschedasticity.



- P-value equal to 6.6 at -11, thus practically equal to 0. This result indicates that the null hypothesis of homoschedasticity of the residuals is rejected.

```
        studentized Breusch-Pagan test

data:  mod_final
BP = 56.434, df = 5, p-value = 6.616e-11
```

- **Durbin-Watson** test for autocorrelation between residues:
  - The '**Scale-Location**' graph shows roughly a scattered cloud of points and thus no particular pattern indicating the non-presence of autocorrelation.

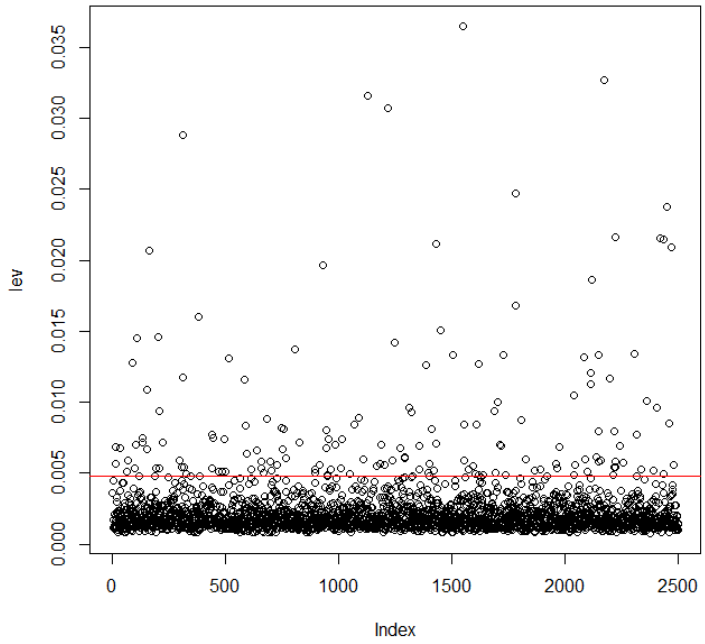


- P-value of 0.11 therefore above the 5% significance level, this indicates that the null hypothesis of independence of the residuals is not rejected. There is no autocorrelation.

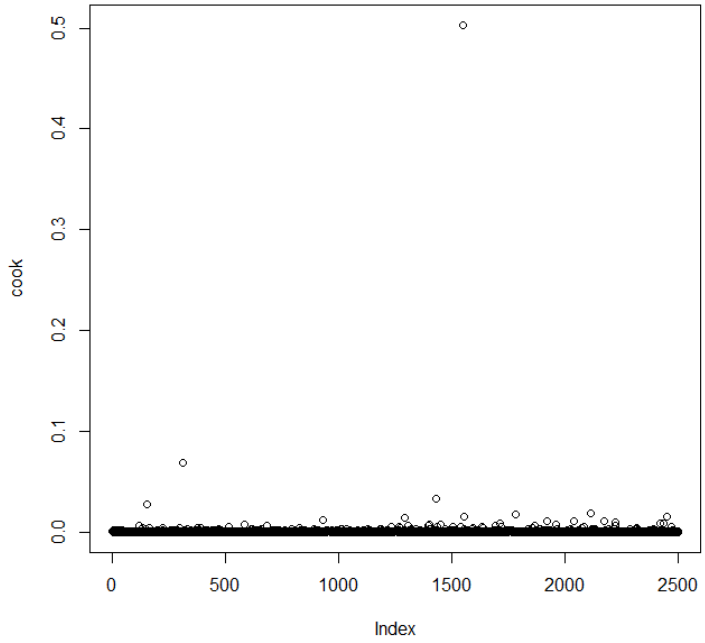
In addition, the analysis of possible leverage observations and model outliers was carried out. Below are the results:

- **Leverage**
  - The model has many leverage values, thus many unusual values in the space of the regressors.

13	15	34	61	67	89	96	101	106
0.005634869	0.006856852	0.006789690	0.005123536	0.005892043	0.012815755	0.005346708	0.006974446	0.014501447
131	134	151	155	161	189	190	204	205
0.007152901	0.007522134	0.010857487	0.006686881	0.020646654	0.004836114	0.005359739	0.014561746	0.005368505
206	220	294	305	310	312	315	328	378
0.009395842	0.007169991	0.005914088	0.005444456	0.028815836	0.011778088	0.005439498	0.004946105	0.016013227
383	440	442	445	471	486	492	497	516
0.004821786	0.005353866	0.007733654	0.007467879	0.005151434	0.005144445	0.007426840	0.005134173	0.013109804
582	587	592	614	638	656	657	666	684
0.011589393	0.008328135	0.006385449	0.005299240	0.006656372	0.005789603	0.005392553	0.005018879	0.008854742
697	702	729	748	750	757	765	805	828
0.005832300	0.005184954	0.005580713	0.008220139	0.006714156	0.008153871	0.006033821	0.013720960	0.007184570
893	895	913	928	946	947	951	956	964
0.005049721	0.005312589	0.005574439	0.019608093	0.006812509	0.008003248	0.004860534	0.007417516	0.005253396
985	1008	1014	1049	1067	1091	1106	1130	1166
0.007044236	0.005337495	0.007422811	0.004957113	0.008467863	0.008889705	0.006009987	0.031583927	0.005483590
1181	1188	1200	1219	1238	1248	1273	1291	1293
0.005680821	0.007024174	0.005616593	0.030695559	0.005928289	0.014168239	0.006776604	0.006042853	0.006142841
1311	1321	1325	1356	1357	1385	1395	1400	1402
0.009616810	0.009268424	0.004837944	0.005283307	0.006895947	0.012654315	0.005075489	0.005696087	0.004805810
1411	1420	1428	1429	1450	1505	1551	1553	1556
0.008082908	0.005215549	0.007125913	0.021165188	0.015099819	0.013327229	0.036440680	0.008454739	0.005912444
1573	1593	1606	1610	1617	1619	1639	1686	1693
0.005264832	0.005508987	0.004921121	0.008429566	0.004864169	0.012658711	0.004912025	0.009349376	0.005077909
1701	1712	1718	1727	1735	1780	1781	1809	1827
0.010000914	0.007032804	0.006966177	0.013314194	0.004846827	0.024661941	0.016791734	0.008719485	0.006016590
1868	1892	1962	1967	1977	2037	2040	2046	2086
0.005172767	0.005312439	0.005679971	0.005338521	0.006865107	0.004908959	0.010512465	0.005556061	0.013191056
2089	2098	2114	2115	2120	2140	2146	2148	2149
0.005962272	0.005117852	0.012066644	0.011267342	0.018628974	0.006163568	0.005842873	0.007960307	0.013300955
2157	2175	2200	2215	2216	2220	2221	2224	2225
0.005935064	0.032698995	0.011658863	0.004896253	0.007947271	0.005421906	0.021624108	0.005823725	0.005457911
2244	2257	2307	2317	2318	2337	2359	2391	2408
0.006928977	0.005781568	0.013414056	0.007685585	0.004824694	0.005305319	0.010063699	0.005181854	0.009637690
2422	2436	2437	2452	2458	2471	2478		
0.021535448	0.004991906	0.021483215	0.023718252	0.008487496	0.020930228	0.005592450		



- **Outliers**
  - The model has 4 outliers, i.e. 4 unusual values of the response variable. The 4 values shown have a very small p-value and this indicates that the values are significantly outliers and statistically influential for the model. This indicates a low robustness of the model.



	rstudent	unadjusted p-value	Bonferroni p
1551	9.074896	2.2497e-19	5.6243e-16
155	4.993307	6.3456e-07	1.5864e-03
1306	4.851698	1.2996e-06	3.2489e-03
1399	-4.293892	1.8228e-05	4.5571e-02

Task 12 - Description of the goodness of the model

Considering the results of the residue diagnostics, the model does not fulfil the following assumptions:

- Residue normality
- Homoschedasticity of residuals

The model has a lot of leverage values that are significantly influential as they exceed the expected threshold by a large margin, and in addition, there are also four outliers that are also significantly influential.

In conclusion, even though the model's adjusted R-square is about 0.73, it does not filter the information well by pouring it into the residuals. This makes the model discrete for making predictions, certainly not optimal.

Task 13 - Forecasting

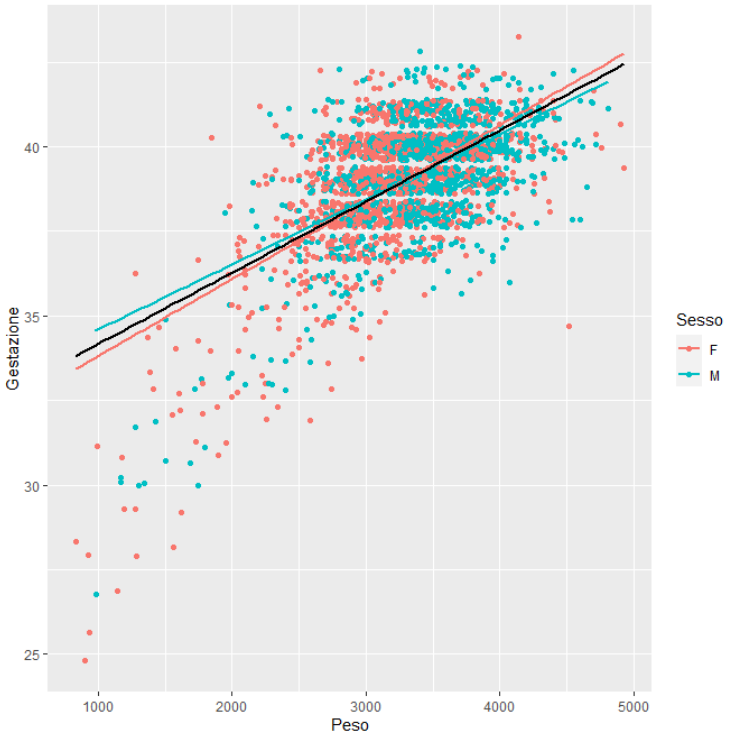
The predicted weight of a newborn at 39 weeks gestation and with a mother in her third pregnancy is: **3627g**.

Task 14 - Graphical representation of the model

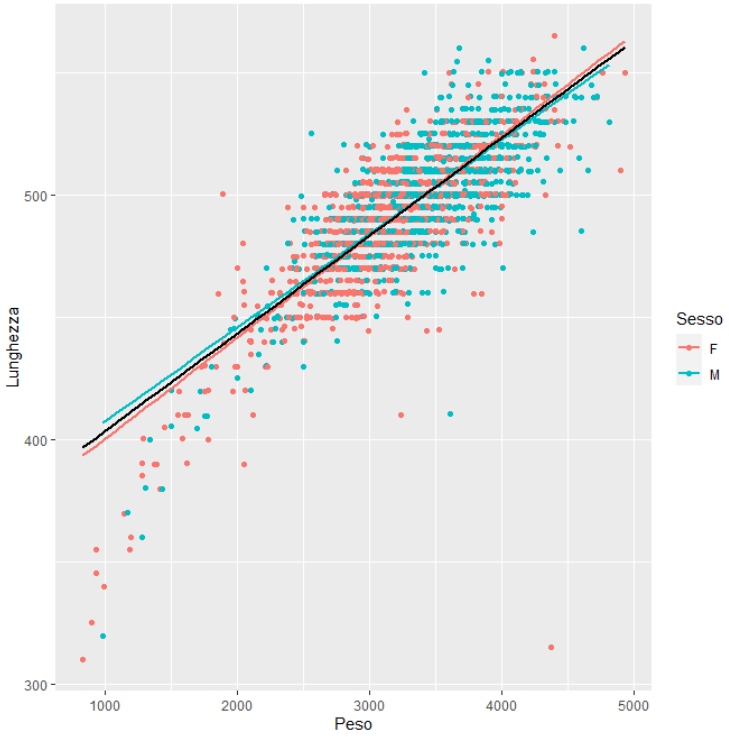
The graphical representation of the model was carried out using scatter3d and the ggplot2 library to show the relationships of the variable Weight with the other variables identified for the model.

Due to visibility problems, the scatter3ds are visible directly in R using the code linked to the project, while for the graphical representations of the regression lines, screens could also be inserted.

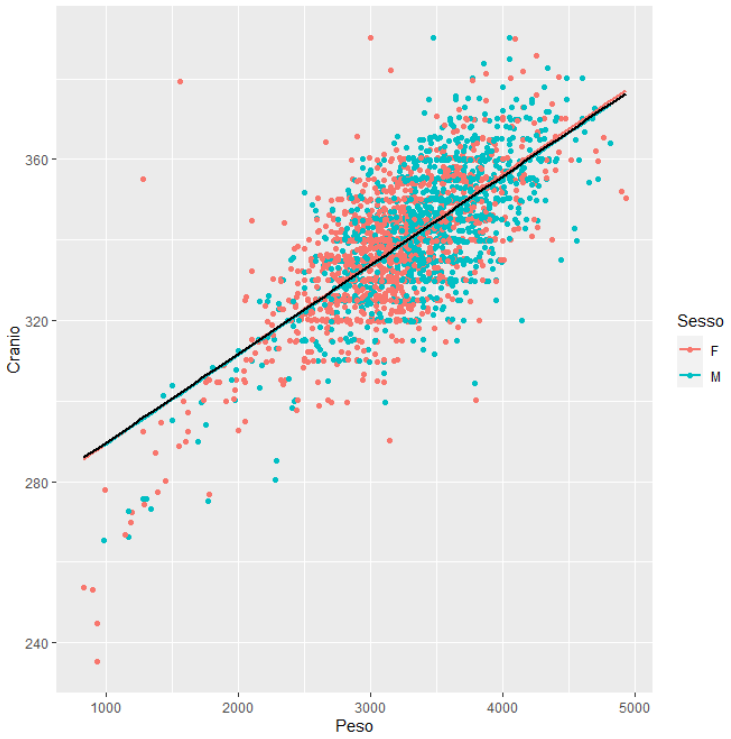
Model representation by Weight - Gestation - Gender



Model representation by Weight - Length - Gender



Model representation by Weight - Gestation - Gender


Cranio
360
320
280
240
1000
2000
3000
4000
5000
Peso
Sesso
F
M