Analysis and Modeling to Predict Newborns Weight

Luca Tripodi

October 2023

Descriptive Analysis

1 Dataset Introduction

The dataset under consideration contains 2500 observations concerning newborns characteristics, information about their mothers and birth circumstances.
The research objective is the prediction of the newborn weight by assessing the mother background (age, number of previous pregnancies, smoking status, gestational period), examining some information about the newborn obtained through echography (length, cranial diameter, gender) and taking into account certain circumstantial data (birth hospital and delivery method).

2 Variables Types

Below the description of the variable types for each observation:

- **"Anni.madre"**: numeric continuous variable, denoting the age of the mother in years.
- **"N.gravidanze"**: numeric discrete variable, denoting the number of previous pregnancies.
- **"Fumatrici"**: categorical variable on nominal scale (control variable), denoting whether the mother is a smoker (0 = no, 1 = yes).
- **"Gestazione"**: numeric continuous variable, denoting the duration of pregnancy in weeks.
- **"Peso"**: numeric continuous variable (response variable), denoting the newborn weight in grams.
- **"Lunghezza"**: numeric continuous variable, denoting the newborn length in millimeters.
- **"Cranio"**: numeric continuous variable, denoting the newborn cranium diameter in millimeters.
- **"Tipo.parto"**: categorical variable on nominal scale (control variable), denoting the type of delivery (natural or cesarean).
- **"Ospedale"**: categorical variable on nominal scale (control variable), denoting the hospital of birth (1, 2 or 3).
- **"Sesso"**: categorical variable on nominal scale (control variable), denoting the newborn gender (male or female).

3 Descriptive Statistical Summary

Below some tables and plots are shown to summarize the data distributions for each variable, along with some observations. For information on calculations please refer to the attached R script.

Position Indexes

	Min	1st Qu	Median	Mean	3rd Qu	Max
Anni.madre	0	25	28	28.16	32	46
N.gravidanze	0	0	1	0.98	1	12
Gestazione	25	38	39	38.98	40	43
Peso	830	2990	3300	3284	3620	4930
Lunghezza	310	480	500	494.7	510.0	565.0
Cranio	235	330	340	340	350	390

Dispersion Indexes

	Variance	Std Dev	Coefficient of Variation
Anni.madre	27.81	5.27	18.72
N.gravidanze	1.64	1.28	130.51
Gestazione	3.49	1.87	4.79
Peso	275665	525	15.99
Lunghezza	693	26	5.32
Cranio	270	16	4.83

Heterogenity and Shape Indexes

	Normalized Gini Index	Fisher Index	Kurtosis Index
Anni.madre	0.97	0.04	3.38
N.gravidanze	0.73	2.51	13.99
Gestazione	0.85	-2.07	11.26
Peso	1	-0.65	5.03
Lunghezza	0.94	-1.51	9.49
Cranio	0.97	-0.79	5.95

Counts for Categorical Variables

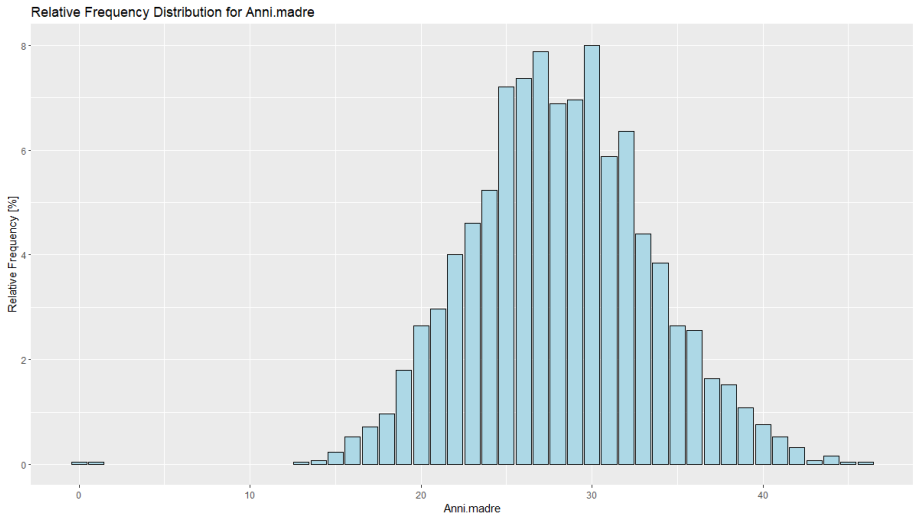
	Ces	Nat
Tipo.parto	728	1772

	F	M
Sesso	1256	1244

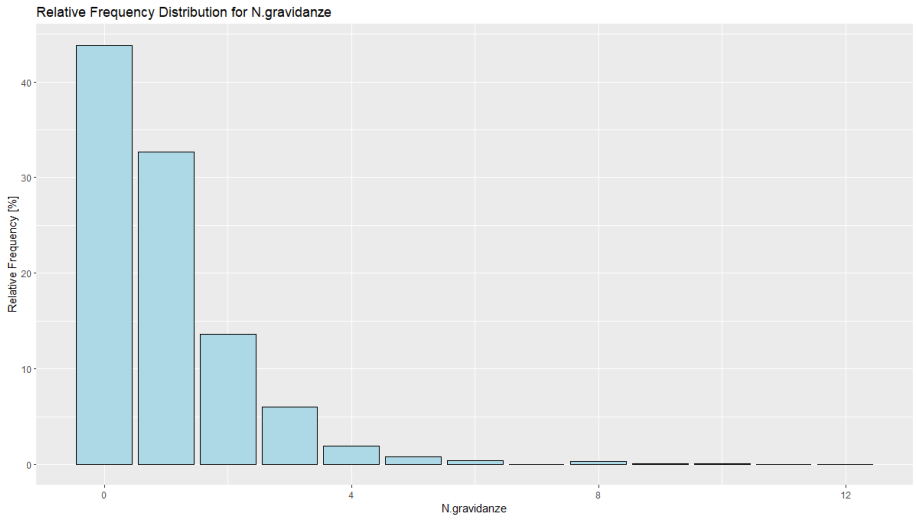
	0	1
Fumatrici	2396	104

	osp1	osp2	osp2
Ospedale	816	849	835

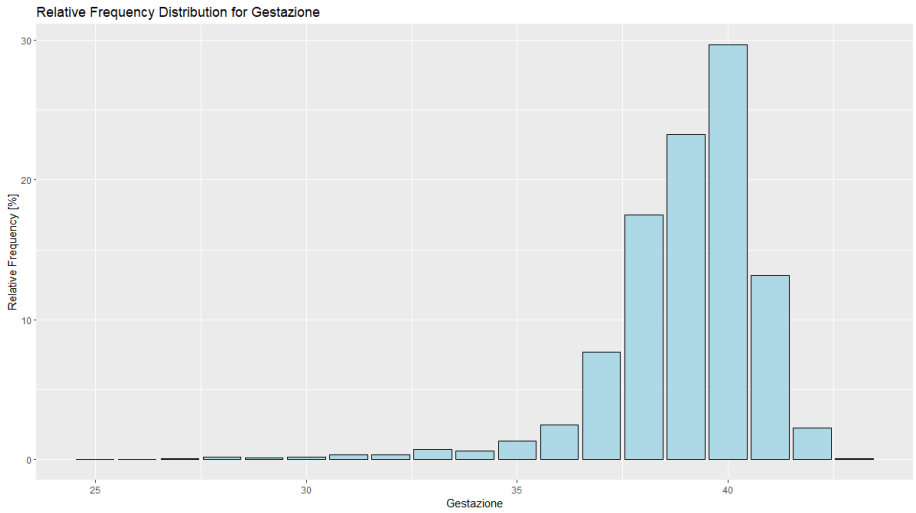
As indicated by the minimum values observed for the mother age there are a few cases where the reported maternal age at delivery is unrealistically low, at 0 or 1 years. Such values are likely errors in data entry or may result from missing information regarding maternal age.



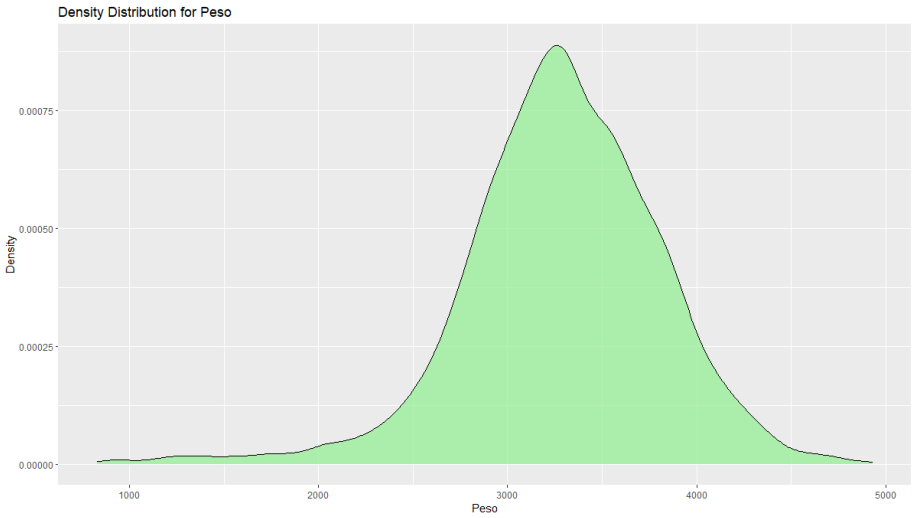
As naturally expected, the number of previous pregnancies exhibits a strong left skew, with the majority of observations corresponding to mothers who are giving birth for the first time.

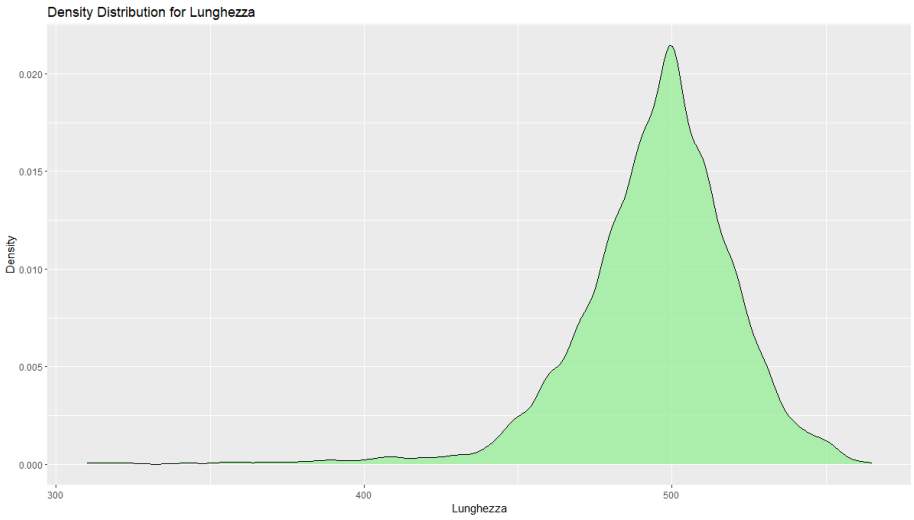


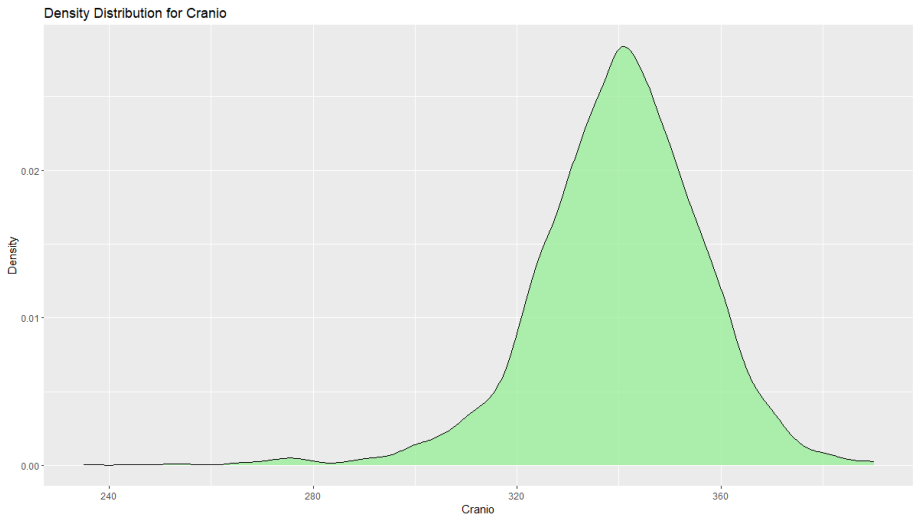
The gestation period exhibits a prolonged left tail for lower values compared to the central indexes. This observation could be attributed to the fact that deviations from the expected date of birth are often associated with complications necessitating premature deliveries.



Regarding the density distribution of weight, length, and cranial diameter, there appears to be a similar long left tail. Given that these measurements are all related to the size of the newborn, it is plausible to attribute this pattern to reasons similar to those for the gestation period.



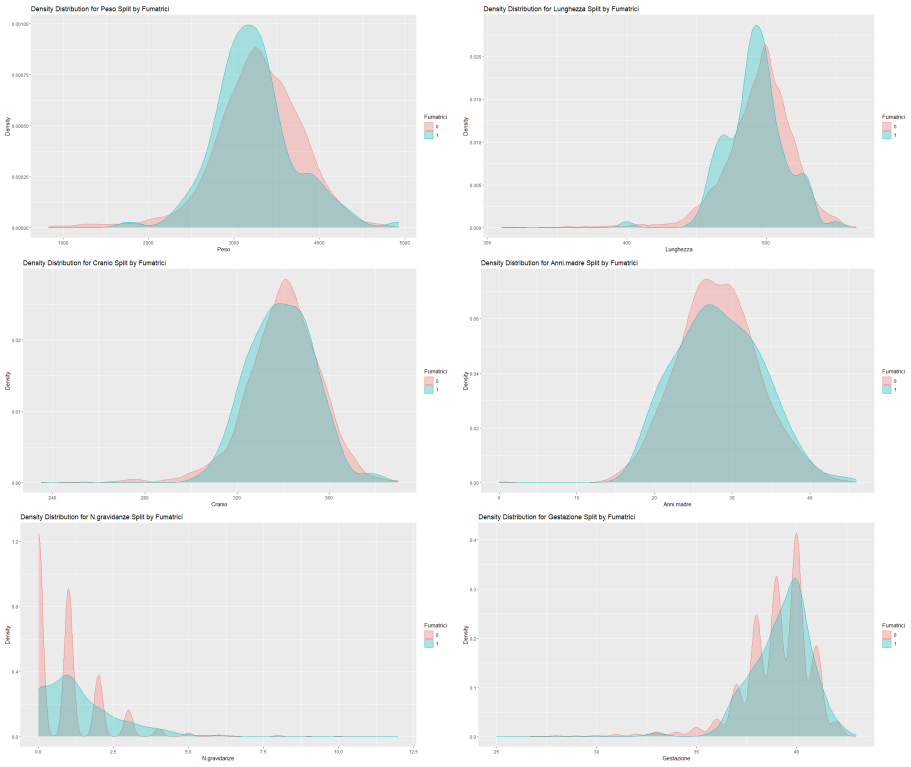




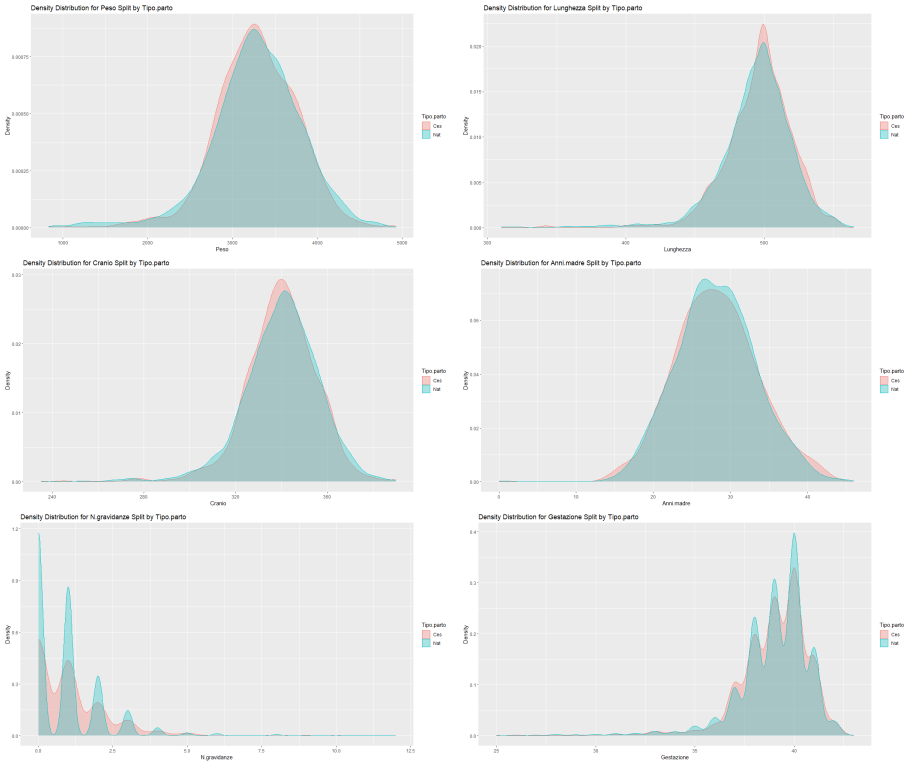
3.1 Density Split by Control Variables

Density plots have been generated for each control variable. It's important to note that the density plots for the number of pregnancies and gestation weeks display a wavy pattern due to the fact that they are treated as continuous variables, causing the density trend to dip between consecutive discrete values. This is a result of the common process applied to all density combinations for ease of data exploration.

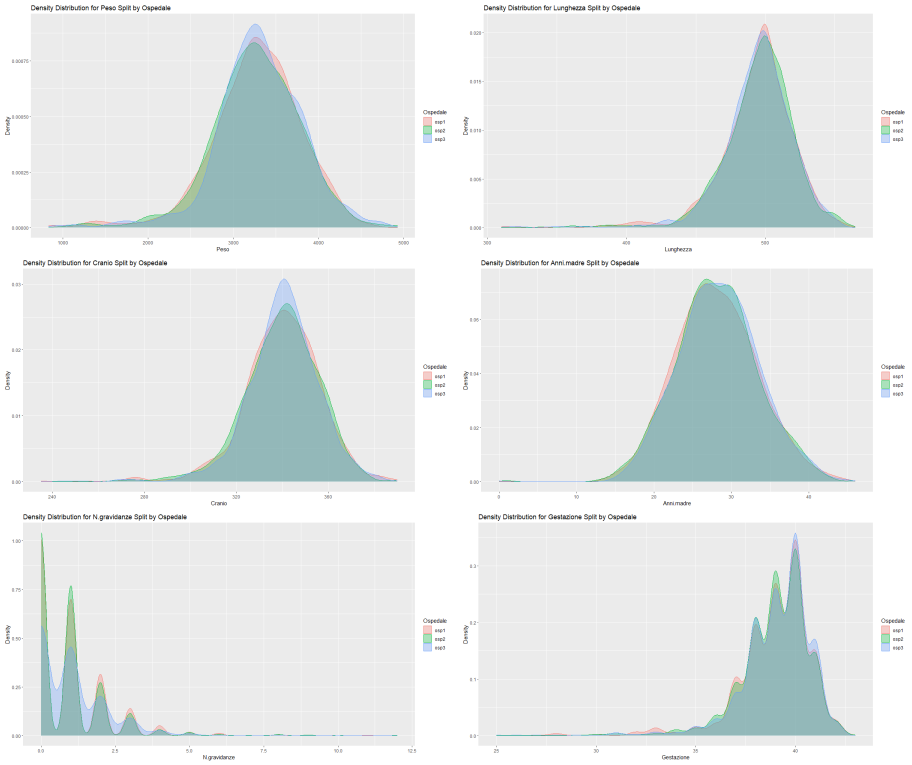
When observing the distribution between mothers who smoke and those who do not, the density shapes appear quite different. However, it's important to note that this difference could be influenced by the significant disparity in the number of observations between the two categories. There are 2396 non-smokers and only 104 smokers, making it challenging to establish a homogeneous density for the latter group. Nevertheless, it is particularly interesting to investigate whether the average weight significantly differs between the two groups.



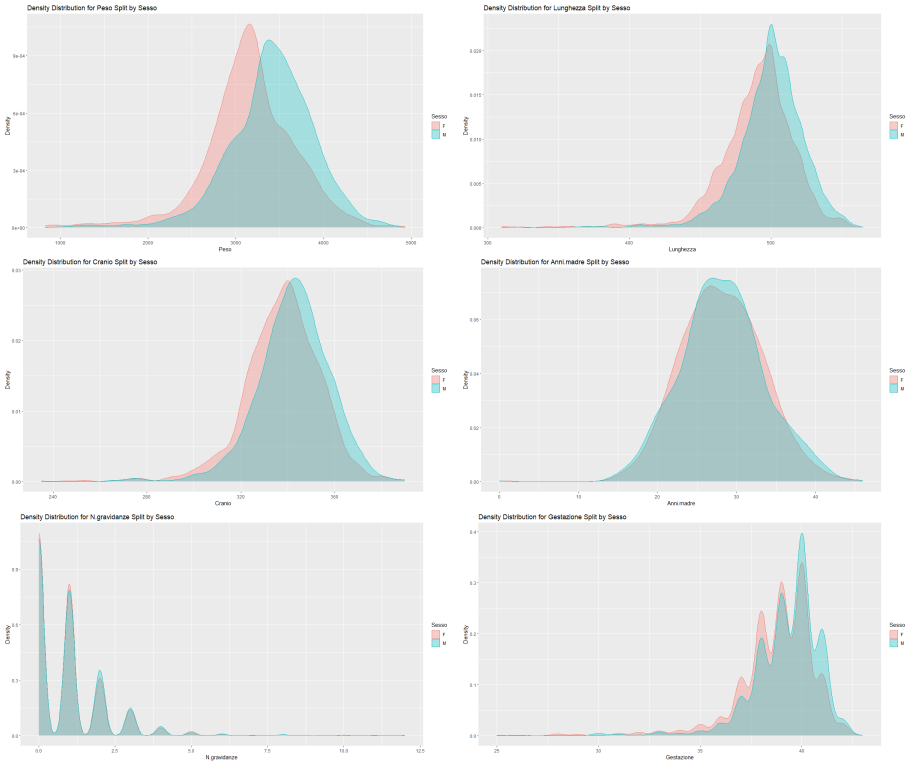
When observing the plots comparing natural and cesarean deliveries, there don't appear to be any significant differences.



When observing the plots comparing the three hospitals, there don't appear to be any significant differences.



When observing the plots comparing the two genders, it is quite evident that the male weight distribution is shifted to the right compared to the female distribution, meaning that average weight is higher. This suggests that there may be a statistically significant difference between the genders of the newborns.



4 Standard Values for Weight and Length

The objective is to assess if the average neonatal weight and length calculated for this dataset sample significantly differ from those measured for the entire population. Since the standard deviation related to average neonatal weight and length in the population is unknown, the Student's t-test can be appropriately applied to check if the mean values in the sample could reasonably be considered equal to those of the population. As average weight and length values, the information have been retrieved from World Health Organization data (weight tables and length tables). Since the average values are divided by gender, the mean global value have been calculated supposing the same number of males and females.

4.1 Weight

```
One Sample t-test
data:  Peso
t = -0.46846, df = 2499, p-value = 0.6395
alternative hypothesis: true mean is not equal to 3289
95 percent confidence interval:
  3263.490 3304.672
sample estimates:
  mean of x
3284.081
```

With a 95% confidence interval and a p-value of 0.64, there is inadequate statistical evidence to reject the null hypothesis. The estimated sample mean weight of 3284.1 falls within the confidence interval [3263.5, 3305.7], making it plausible to consider it consistent with the null hypothesis, which assert that sample weight is significantly the same as the population one.

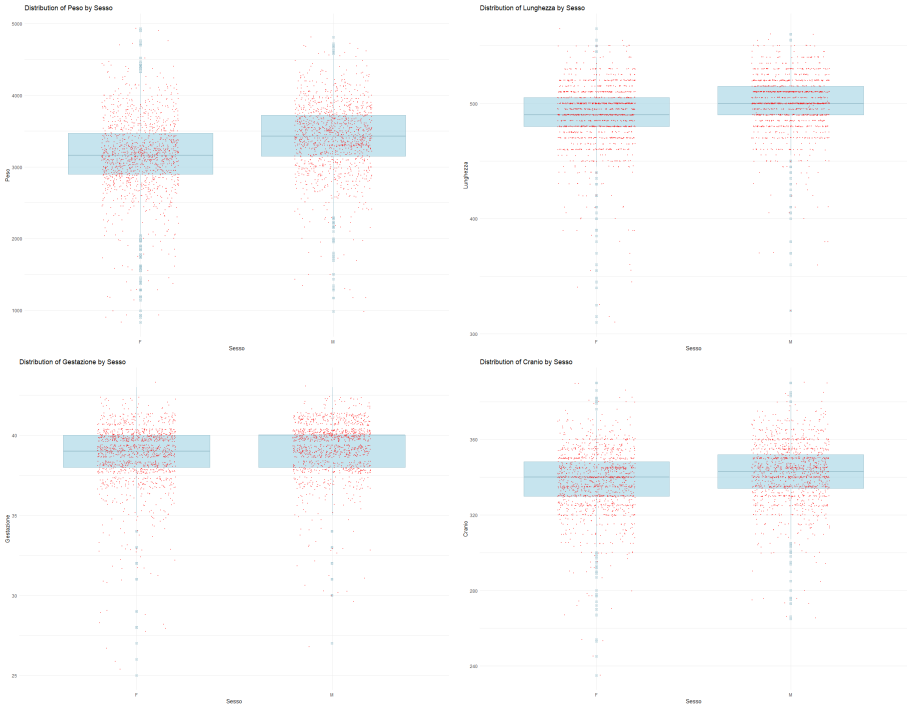
4.2 Length

```
One Sample t-test
data:  Lunghezza
t = -0.58514, df = 2499, p-value = 0.5585
alternative hypothesis: true mean is not equal to 495
95 percent confidence interval:
  493.6598 495.7242
sample estimates:
  mean of x
494.692
```

With a 95% confidence interval and a p-value of 0.56, there is inadequate statistical evidence to reject the null hypothesis. The estimated sample mean length of 494.7 falls within the confidence interval [493.7, 495.7], making it plausible to consider it consistent with the null hypothesis, which asserts that the sample length is not significantly different from the population length.

5 Gender Differences

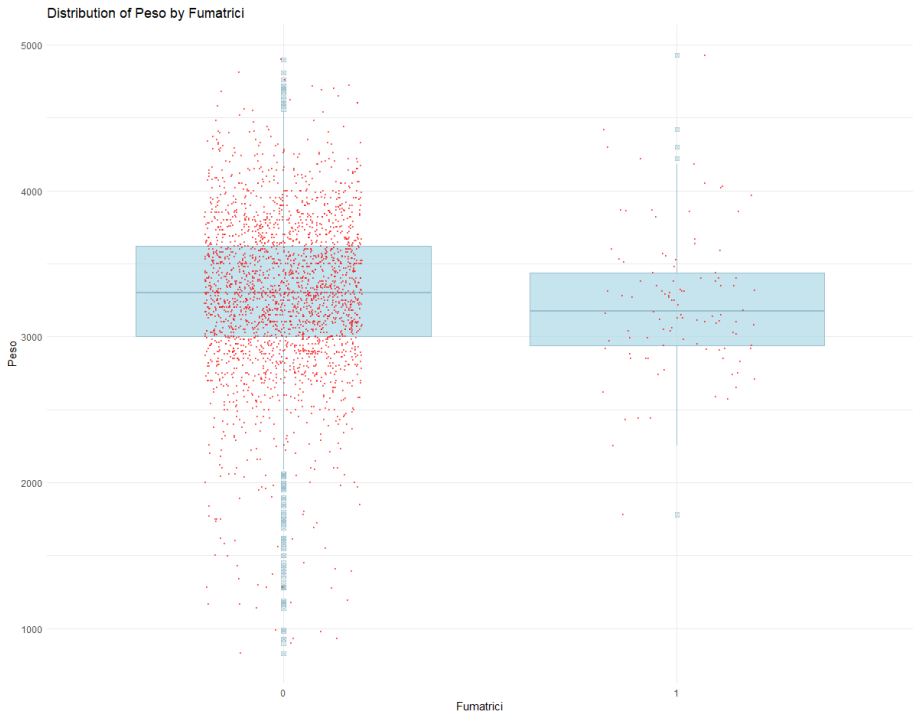
For some variables, it is of interest to investigate whether there are any significant differences between the two genders. Below the box plots representing the distribution of weight, length, gestation period and cranium diameter, separated by male and female observations.



Statistically significant differences have been observed for variables related to newborn size (weight, length, and cranium), with p-values very close to 0. Additionally, a substantial statistical difference has been detected for the gestation period (p-value = 2.5e-11). The result is somewhat surprising because a significant difference in gestation period based on the baby's gender would not be expected. Since the absolute difference is relatively small (a few days), it could be attributed to the gestation period being discretized in weeks. Alternatively, this dataset may be influenced by specific cases or undisclosed variables and environmental conditions, making it challenging to explain the difference. In conclusion, this result should be interpreted with caution.

5.1 Smokers Difference

Since the weight density plot suggested the possibility of a difference between smoker and non-smoker mothers, it is interesting to investigate whether the mean weight differs statistically between these two categories.




```
Welch Two Sample t-test
data:  weight_smoker_no and weight_smoker_yes
t = 1.034, df = 114.1, p-value = 0.3033
alternative hypothesis: true difference in means is
   not equal to 0
95 percent confidence interval:
  -45.61354 145.22674
sample estimates:
  mean of x mean of y
3286.153  3236.346
```

With a 95% confidence interval and a p-value of 0.3, there is insufficient statistical evidence to reject the null hypothesis. Although the plots suggest that newborns from smoking mothers may weigh less than others, it is possible that with a larger sample size, this difference could become more statistically significant.

6 Delivery Type Relationship by Hospital

To assess whether there is a significant relationship between the delivery type and the hospital where the birth took place, a contingency table is constructed for these two variables.

Contingency Table - Delivery Type by Hospitals

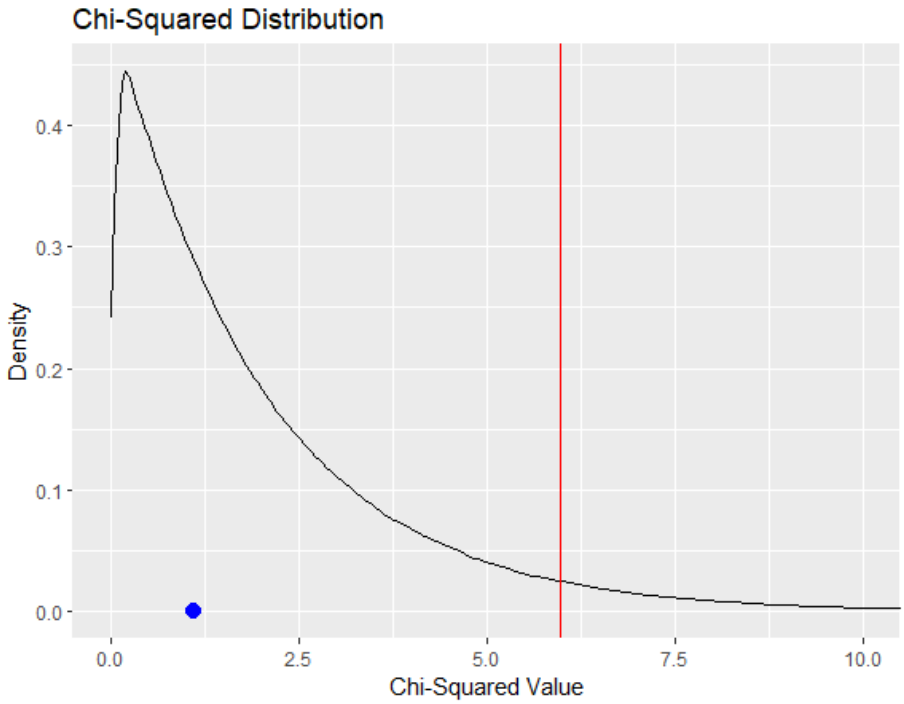
Hospital	Natural Births Sum	Cesarean Births Sum
osp1	574	242
osp2	595	254
osp3	603	232

Subsequently, observed frequencies are compared to expected frequencies using a Chi-squared test. The null hypothesis assumes that the variables are independent.

```
Pearson's Chi-squared test
data:  table_observed
X-squared = 1.0972, df = 2, p-value = 0.5778
```

The chi-squared test yields a high p-value, signifying a lack of substantial statistical evidence to support a connection between delivery types and hospitals.

This outcome is effectively visualized through the chi-squared distribution plot. A vertical red line denotes the critical chi-squared value corresponding to a significance level *alpha* of 0.05. The blue dot, representing the calculated chi-squared value, falls well outside the rejection region, reinforcing the absence of a significant relationship.



Multidimensional Analysis

0 Correlation Analysis

Before proceeding with the construction of a model to discover which variables interact significantly with the response variable, it is necessary to study the correlations between all pairs of variables in order to identify those that are more concordant or discordant with each other.

Correlation Matrix

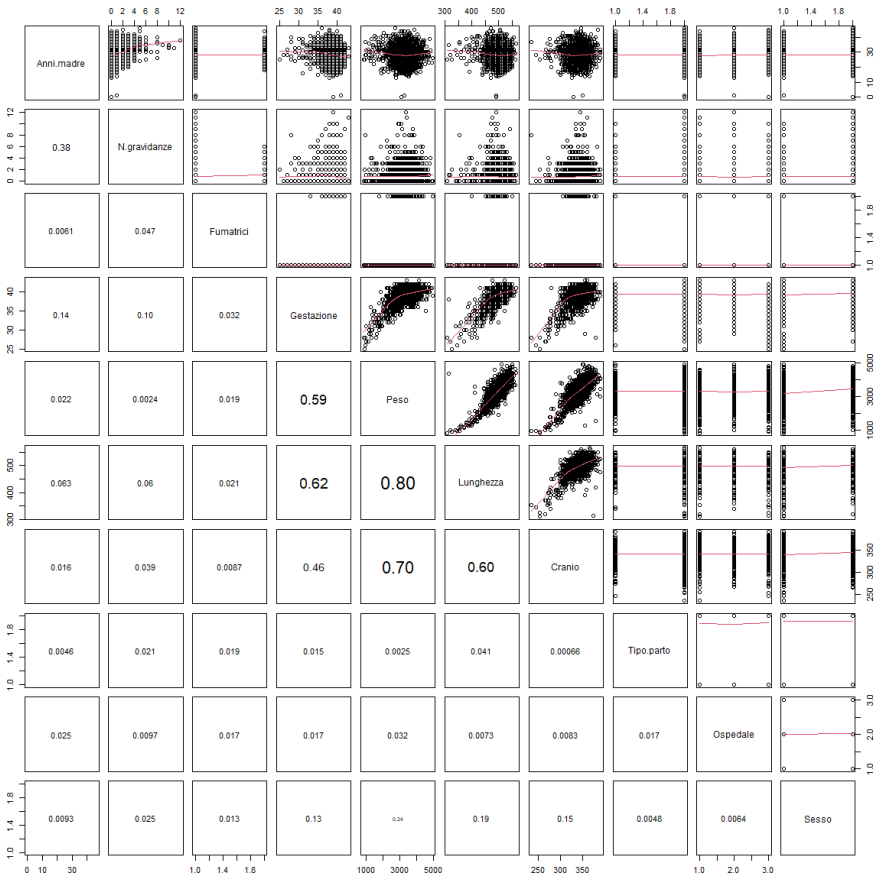
	Anni.madre	N.gravidanze	Gestazione	Peso	Lunghezza	Cranio
Anni.madre	1.00	0.38	-0.14	-0.02	-0.06	0.02
N.gravidanze	0.38	1.00	-0.10	0.00	-0.06	0.04
Gestazione	-0.14	-0.10	1.00	0.59	0.62	0.46
Peso	-0.02	0.00	0.59	1.00	0.80	0.70
Lunghezza	-0.06	-0.06	0.62	0.80	1.00	0.60
Cranio	0.02	0.04	0.46	0.70	0.60	1.00

Based on the results from the generated correlation matrix, below the considerations:

- **Relations with response variable *Peso***: The weight of newborns appears to have a moderate positive correlation with gestational period, newborn length and cranial dimensions. However, it's important to note that correlation does not necessarily imply a direct causal relationship between these variables. Regarding the correlation between weight and gestational period, it can be hypothesized that a shorter gestational period may lead to lighter newborns, as premature births could affect birth weight. On the other hand, it seems plausible that newborns with greater length and larger head dimensions have a higher birth weight.

- **Relations with response variable**: There is a slight positive correlation between the mother's age and the number of pregnancies, plausibly due to the logical fact that more time (and so at older age) is required for additional pregnancies. It can be noticed that gestational period, length and cranial diameter are moderately positively correlated. The correlation coefficients suggest that on average longer gestational periods are associated with increased newborn length and cranial diameter. This implies that a shorter gestational period might lead to less developed newborns in terms of these measurements.

Correlation Matrix with Scatterplots



1 Multiple Linear Regression Model

Upon a preliminary examination of the response variable, a weight distribution with a negative skewness (-0.65) and a leptokurtic shape with a positive kurtosis (2.03) is evident compared to a normal distribution. Conducting a Shapiro-Wilk test to assess whether the weight follows a normal distribution reveals that, despite the expectation of a normal distribution in the population, this sample exhibits a non-normal distribution. This significant deviation from normality can introduce distortions in regression coefficients, impact the normal distribution of residuals, and lead to less accurate predictions.

Constructing a model that incorporates all variables, the following results are obtained:

```
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 *
Fumatrici1      -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 ***
---

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
```

Below the considerations regarding the explanatory variables:

- *The Most Significant*: The gestation period, newborn length and cranium diameter have a substantial impact on the model, as indicated by their p-values close to 0. These variables show a strong influence and their increases appear to significantly contribute to an increase in the newborn's weight. Newborn gender also appears to be a significant factor in predicting the weight. The coefficient suggests that on average a male newborn contributes to a weight increase of 77.5 units compared to a female newborn when all other variables are held constant. This was expected based on test result on mean weights comparison by genders.

- *The Less Significant*: The number of pregnancies appears to have a slight but significant positive influence on weight increase. It is also noteworthy that on average, a natural birth may positively affect the estimated weight. Conversely, it can be hypothesized that a cesarean birth might be performed under critical conditions where the gestation does not reach full term, which significantly impacts the weight. Furthermore, it seems that being born in Hospital 2 or Hospital 3 has a slightly significant effect on weight, either negative or positive, respectively. Without additional data or analysis, it could be assumed that there might be demographic and geographical reasons for the difference in contribution.

2 Model Improvement

To improve the model, the path of simplification was followed, which involves an overall reduction of the explanatory variables to retain only the most significant ones, aiming for a good balance between accuracy and simplicity. Below the incremental model updates, the related adjusted R-squared and a brief description (please refer to attached R script for details):

- Model 2 ($R^2_{adj} = 0.727$): Removal of variables that were considered not significant due to their high p-values:

$$-Anni.madre - Fumatrici - Ospedale$$

- Model 3 ($R^2_{adj} = 0.726$): Removal of less significant variables:

$$-N.gravidanze - Tipo.parto$$

- Model 4 ($R^2_{adj} = 0.599$) and Model 5 ($R^2_{adj} = 0.656$): Compared to the previous model, it is important to verify whether the selected explanatory variables exhibit multicollinearity. The variance inflation factors (VIF) are calculated to assess this possible impact. The results indicate that none of the VIF values exceeds the threshold of 5, which suggests that there is no strong correlation and consequently no significant distortion in the regression coefficients of the model. When attempting to construct two additional models by removing either length or cranial diameter, the resulting models show decreased accuracy compared to the previous one.

- Model 5 ($R^2_{adj} = 0.728$): An automatic stepwise procedure was employed to create a new model, selecting the Akaike Information Criterion, which tends to favor models with more parameters. The resulting model is very similar to Model 2 but includes the additional complexity of the hospital variable.

- Model 5 ($R^2_{adj} = 0.726$): By once again employing the automatic stepwise procedure, this time with a preference for the Bayesian Information Criterion, which penalizes overly parameterized models, the exact same Model 3 is obtained.

Executing an ANOVA test, so analyzing the variance on best created models according to their adjusted R-squared values, the following results are obtained:

```
1                Res.Df       RSS  Df  Sum of Sq      F     Pr(>F)
2  Model 1         2489 186762521
3  Model 2         2493 187601677  -4   -839155 2.7959 0.0247927 *
4  Model 3         2495 188688687  -2  -1087011 7.2433 0.0007301
       ***
5  Model 6         2491 186899996   4   1788691 5.9595  9.03e-05
       ***
```

In light of similar results regarding the sum of squared residuals and further evaluation of goodness of fit through Akaike and Bayesian evaluation criteria, which still report very similar results, Model 3 is considered the best choice for data fitting. This model shows a better balance between accuracy and simplicity.

3 Not Linear Relations

Based on the model that includes all conceivable explanatory variables, an exploration of potential nonlinear effects in weight prediction is conducted. It is hypothesized that weight, which could be associated with newborn volume, might exhibit a cubic relationship with dimensional variables, including length and cranial diameter. To investigate this possible interaction, incremental nonlinear contributions are introduced and the related outcomes are assessed:

- Model 1.1 ($R^2_{adj} = 0.731$): Added cubic relationship with cranial diameter.
- Model 1.2 ($R^2_{adj} = 0.737$): Added cubic relationship with length.
- Model 1.3 ($R^2_{adj} = 0.739$): Added cross-relationship between cranial diameter and length.
- Model 1.4 ($R^2_{adj} = 0.734$): Applying automatic stepwise procedure using Akaike Information Criterion, below predictor variables selected:

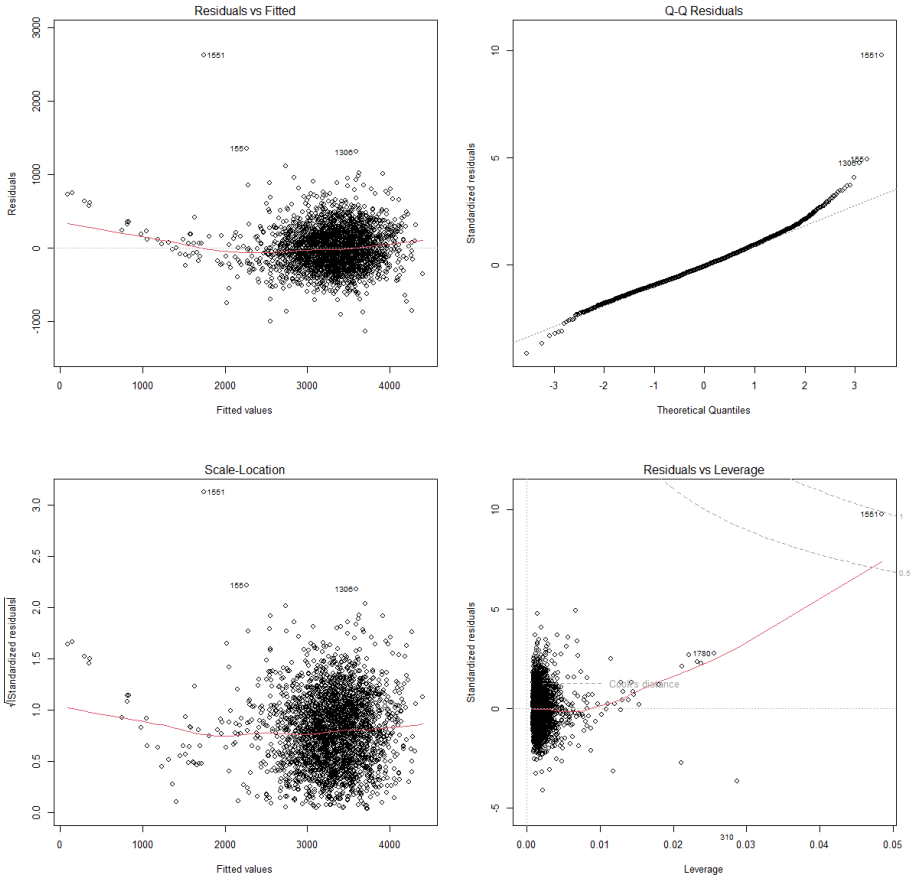
$$Gestazione + Cranio + Sesso + Lunghezza^3$$

Evaluating the ANOVA test, Model 1.3 has the lowest residual sum of squares and the highest explained sum of squares, but it is also the most complex model. To favor simplicity with fewer predictor variables, considering the similar results about Information Criteria and the related adjusted R-squared (0.734), Model 1.4 could be a preferred choice for modeling the data.

4 Residuals Diagnostics

To complete the analysis of Model 3 that was previously evaluated, it is necessary to perform a diagnostic of its residuals to ensure they adhere to the required assumptions.

Residuals Diagnostics



Below consideration on residuals scatterplots:

- **Residuals vs Fitted**: Relating the estimates obtained from the model to their respective residuals, the plot shows data points moderately scattered randomly around the mean of 0. Especially for lower estimated values, there is a slight curve indicating that some of the information hasn't been correctly filtered out by the predictors and has manifested in the residuals.

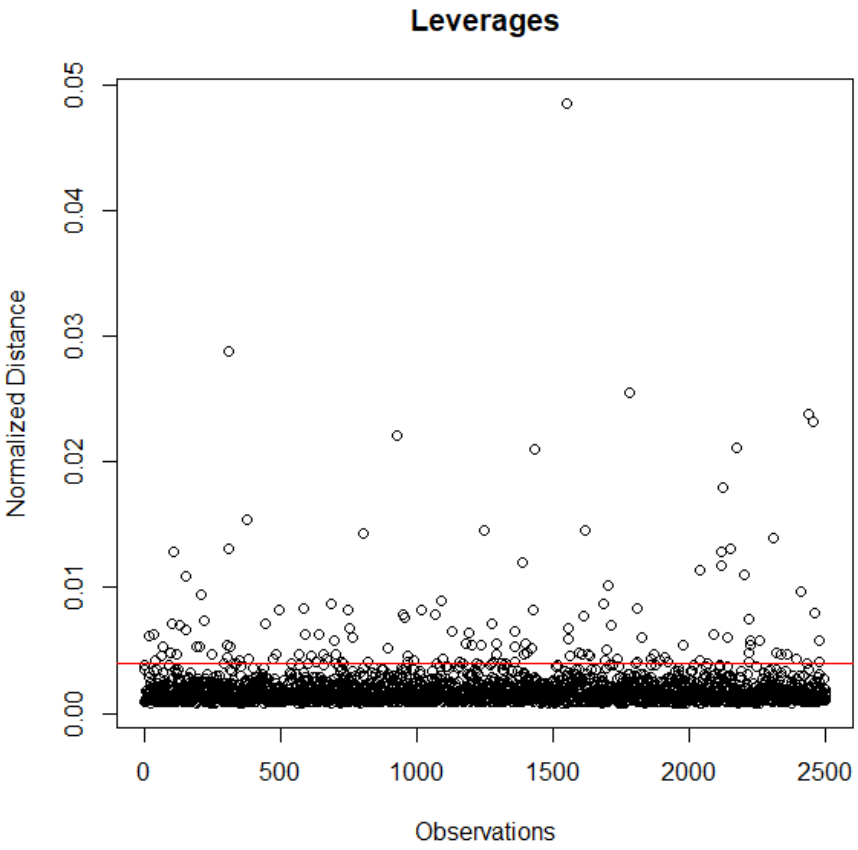
- **Q-Q Residuals**: When relating standardized residuals to theoretical quantiles of a normal distribution, the points are distributed around the bisector for a significant portion of the central region of the plot, except in the outer regions where they diverge. This indicates that the residuals mostly follow a normal distribution, except for lowset or highest values of the response variable, where the model does not fit accurately and nonlinear relationships may be present.

- **Scale-Location**: When relating the square root of standardized residuals to the model's estimates, the plot displays a random scatter of points around a horizontal line for the upper part. This indicates that in that region the variance among residuals is constant, indicating homoscedasticity. However for lower values different variances are observed.

- **Residuals vs Leverage**: By relating the standardized residuals to the leverages, all values fall within the Cook's distance except for observation 1551. Therefore it is necessary to assess how this influential observation impacts the regression coefficients and consequently the whole model estimation.

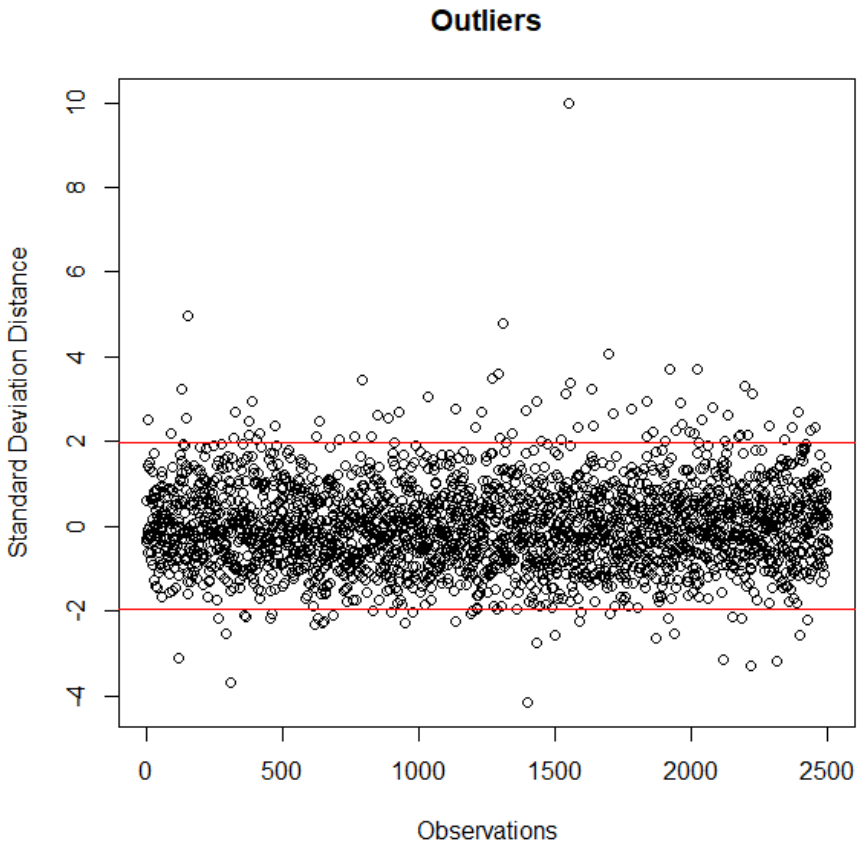
4.1 Leverages

After calculating all normalized distances between observed and fitted values and establishing the derived threshold, it is determined that there are 135 observations with high leverage values that deviate significantly from the rest of the observations in the predictor space.



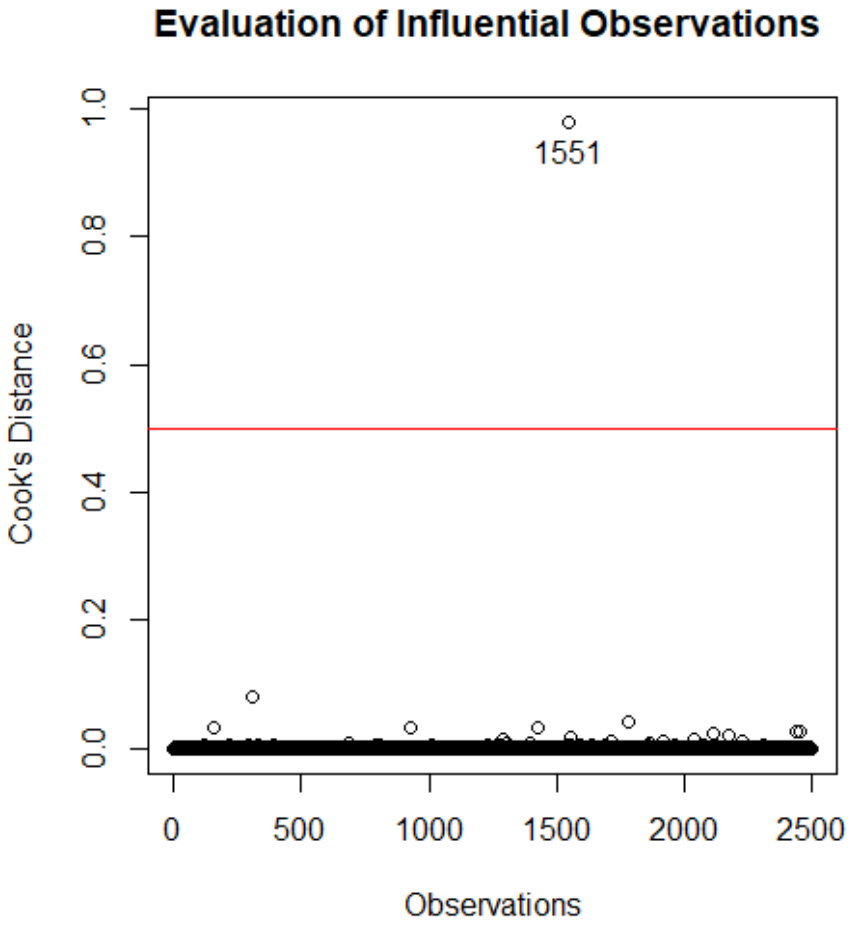
4.2 Outliers

After calculating all standard deviation distances between observed and predicted values and establishing the derived threshold, it is determined that there are 112 observations considered as outliers if compared to the rest of the observations. However, when applying the Bonferroni correction which considers p-values, the count of anomalous observations is reduced to 3, one of which is the observation 1551 previously identified.



4.3 Cook's Distance

Evaluating the Cook's distance, the observation 1551 stands out with a value very close to 1, indicating its strong influence which could impact the estimation of regression coefficients. However, by restructuring Model 3 by removing the influential observation, the adjusted R-squared value remains practically unchanged.



4.4 Residuals Tests

Finally, canonical tests are performed to check the residual requirements:

- **Normality - Shapiro-Wilk Test**: With a test result of 0.9742 and a p-value close to 0, it appears that the residuals do not follow a normal distribution.
- **Zero Mean**: Residuals mean is equal to -9.39e-15, so quite respecting the assumption of mean equal to 0.
- **Homoscedasticity - Breusch-Pagan Test**: With a test result of 89.148 and a p-value close to 0, it appears that residuals variance is not constant.
- **Independence - Durbin-Watson Test**: With a test result of 1.9557 and a p-value equal to 0.1337, it is accepted the hypothesis of independence and therefore no autocorrelation among residuals.

5 Model Evaluation

To assess the goodness of fit of the models, the mean square error (MSE) is calculated among the best-performing ones:

	Model	MSE
1	model_1_all_var	74705
2	model_2_only_significant	75041
3	model_3_best_significant_fixed	75475
4	model_1.4_not_linear	73238
5	model_6_stepwise_AIC	74760

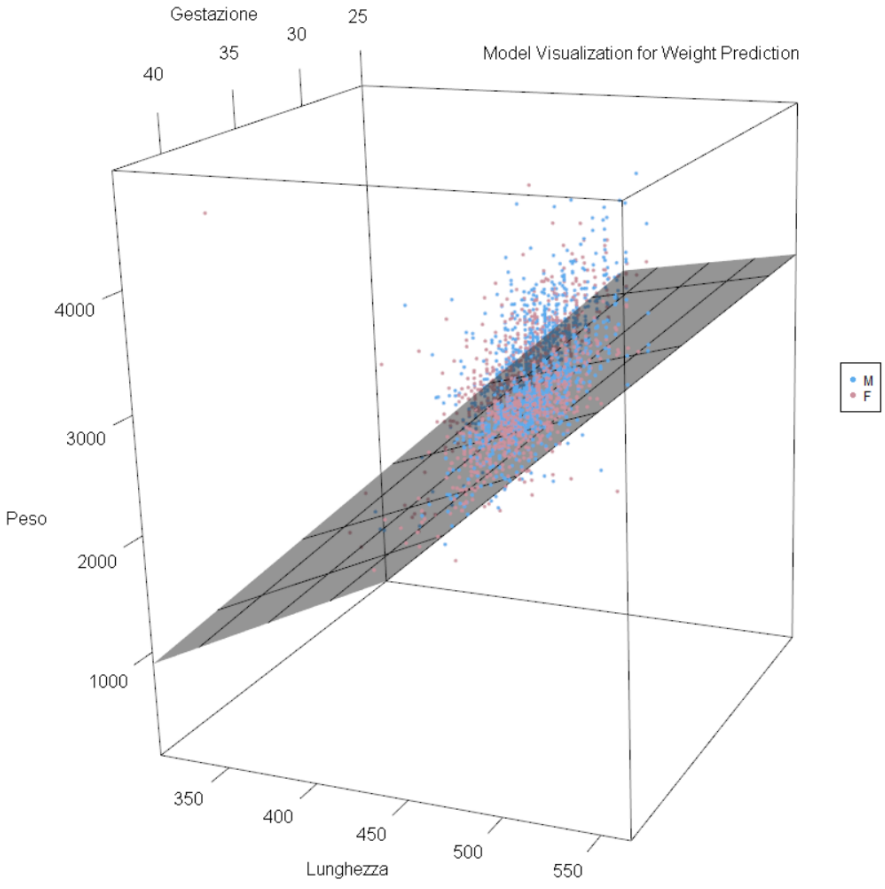
When comparing the MSE among the best models, it's evident that the results are quite similar. However, when considering the MSE values in relation to the variance of the response variable equal to 275665, it becomes apparent that all these models explain a significant portion of the response variable.

6 Model Prediction

To make a weight prediction for mothers on their third pregnancy who gave birth in the 39^{th} week of gestation, a model was constructed using these two variables as predictors. The prediction was then compared to the averages of observations in the sample dataset that shared the same values for the number of pregnancies and gestation period. The obtained prediction of a weight of 3340 grams for this specific case appears reasonable as it falls in line with the central values of the data. However, it's important to note that the model constructed using only two explanatory variables has a relatively low adjusted R-squared value (0.354). This suggests that the model may have limited accuracy in making predictions across a broader range of cases.

7 Model Visualization

To create a 3D visualization for weight prediction, a simplified model incorporating only gestation period and newborn length was developed. The visualization includes a color mapping based on gender, which subtly suggests that higher weights are more frequent in males, as already evident from the data.


Model Visualization for Weight Prediction
Gestazione
40
35
30
25
Peso
4000
3000
2000
1000
Lunghezza
350
400
450
500
550
M
F