A statistical model to predict the weight of newborns

After loading the dataset and using an attach(), we go on to study the data in detail with the head() function to display the first few rows and, most importantly, with the summary() function to examine the position indices.

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

There are several continuous quantitative variables, including Anni.madre, N.pregnancies, Gestation, Weight, Length and Skull, as well as categorical variables such as Smokers, Gender, Hospital and Type of delivery.

Let us now take a closer look at the variable Anni.madre, which is a continuous quantitative variable. Considering the context of the study, the minimum value of the variable, 0, does not make sense. Similarly, too low an age is unlikely for a mother.

To address this issue, we will proceed by sorting the first values of Anni.madre in ascending order in order to better observe the presence of any implausible values.

```
var_cresc <- sort(neonati$Anni.madre)
head(var_cresc, 5)
```

```
[1]  0  1 13 14 14
```

There are only two problematic values, 0 and 1, in the variable Anni.madre. Since these are only two cases out of 2500 (about 0.08% of the total), we will be able to exclude them smoothly from the dataset or, alternatively, replace the offending values with the mean. In this case we will proceed with the substitution.

Once we have performed the substitution of the two values, we check the indices for all noncategorical variables.

	Anni.madre	Cranio	Fumatrici	Gestazione	Lunghezza	N.gravidanze	Peso
Mean	28.18614892	340.02922338	0.04163331	38.97958367	494.69575661	0.98158527	3284.18414732
Std.Dev	5.21720609	16.42946924	0.19978977	1.86895026	26.32884653	1.28094893	525.22937427
Min	13.00000000	235.00000000	0.00000000	25.00000000	310.00000000	0.00000000	830.00000000
Q1	25.00000000	330.00000000	0.00000000	38.00000000	480.00000000	0.00000000	2990.00000000
Median	28.00000000	340.00000000	0.00000000	39.00000000	500.00000000	1.00000000	3300.00000000
Q3	32.00000000	350.00000000	0.00000000	40.00000000	510.00000000	1.00000000	3620.00000000
Max	46.00000000	390.00000000	1.00000000	43.00000000	565.00000000	12.00000000	4930.00000000
MAD	4.44780000	14.82600000	0.00000000	1.48260000	22.23900000	1.48260000	459.60600000
IQR	7.00000000	20.00000000	0.00000000	2.00000000	30.00000000	1.00000000	630.00000000
CV	0.18509822	0.04831782	4.79879667	0.04794690	0.05322230	1.30497978	0.15992690
Skewness	0.15097173	-0.78461925	4.58665318	-2.06389091	-1.51366518	2.51190318	-0.64701485
SE.Skewness	0.04898001	0.04898001	0.04898001	0.04898001	0.04898001	0.04898001	0.04898001
Kurtosis	-0.10792299	2.94011162	19.04501197	8.24650594	6.47334115	10.97043487	2.02472766
N.Valid	2498.00000000	2498.00000000	2498.00000000	2498.00000000	2498.00000000	2498.00000000	2498.00000000
Pct.Valid	100.00000000	100.00000000	100.00000000	100.00000000	100.00000000	100.00000000	100.00000000

The minimum of Anni.Madre is now 13, a sensible value for the nature of the variable. By studying Skewness and Curtosis we can assume that:

- For Anni.madre, skewness close to zero suggests a rather skewed distribution, while kurtosis slightly below zero indicates lighter tails than a normal distribution. The coefficient of variation is rather low, indicating low variability of the data relative to the mean.
- For the Skull variable, slightly negative skewness suggests a skewed distribution toward the left tail, while positive kurtosis indicates heavier tails than a normal distribution. The coefficient of variation is extremely low.
- For Gestation, rather negative skewness indicates a skewed distribution with longer tails to the left, while positive kurtosis suggests heavier tails than a normal distribution. Again, the coefficient of variation is low, indicating low variability of the data from the mean.
- By Length, negative skewness suggests an asymmetric distribution with longer tails toward the left, while positive kurtosis indicates heavier tails.
- For N.Pregnancy, extremely positive skewness indicates a strongly skewed distribution with a very long tail to the right, while very high kurtosis suggests heavier tails than a normal distribution. In addition, high coefficient of variation indicates large variability in the data.
- Regarding the variable Weight, slightly positive skewness suggests a slightly skewed distribution toward the left, while positive kurtosis indicates heavier tails than a normal distribution. The coefficient of variation indicates low variability relative to the mean."

Let us turn to the frequency tables of the variables Smokers, Type.delivery, Hospital and Sex, which were transformed into percentage proportions.

```
table(Fumatrici)
freq_fum <- prop.table(table(Fumatrici))*100
freq_fum
table(Tipo.parto)
freq_parto <- prop.table(table(Tipo.parto))*100
freq_parto
table(Ospedale)
freq_osp <- prop.table(table(Ospedale))*100
freq_osp
table(Sesso)
freq_sesso <- prop.table(table(Sesso))*100
freq_sesso
```

Non-smokers (0) make up almost all of our sample, with smoking mothers(1) settling at just 4 percent of the total.

```
Fumatrici
   0    1
2394  104
```

```
Fumatrici
        0         1
95.836669  4.163331
```

Natural births (nearly 71%) are much more common than Caesares, 29% of our observations.

```
Tipo.parto
 Ces  Nat
 728 1770
```

```
Tipo.parto
     Ces      Nat
29.14331 70.85669
```

The births observed in our study are equivalently divided among the three hospitals involved.

```
Ospedale
osp1 osp2 osp3
 816  848  834
```

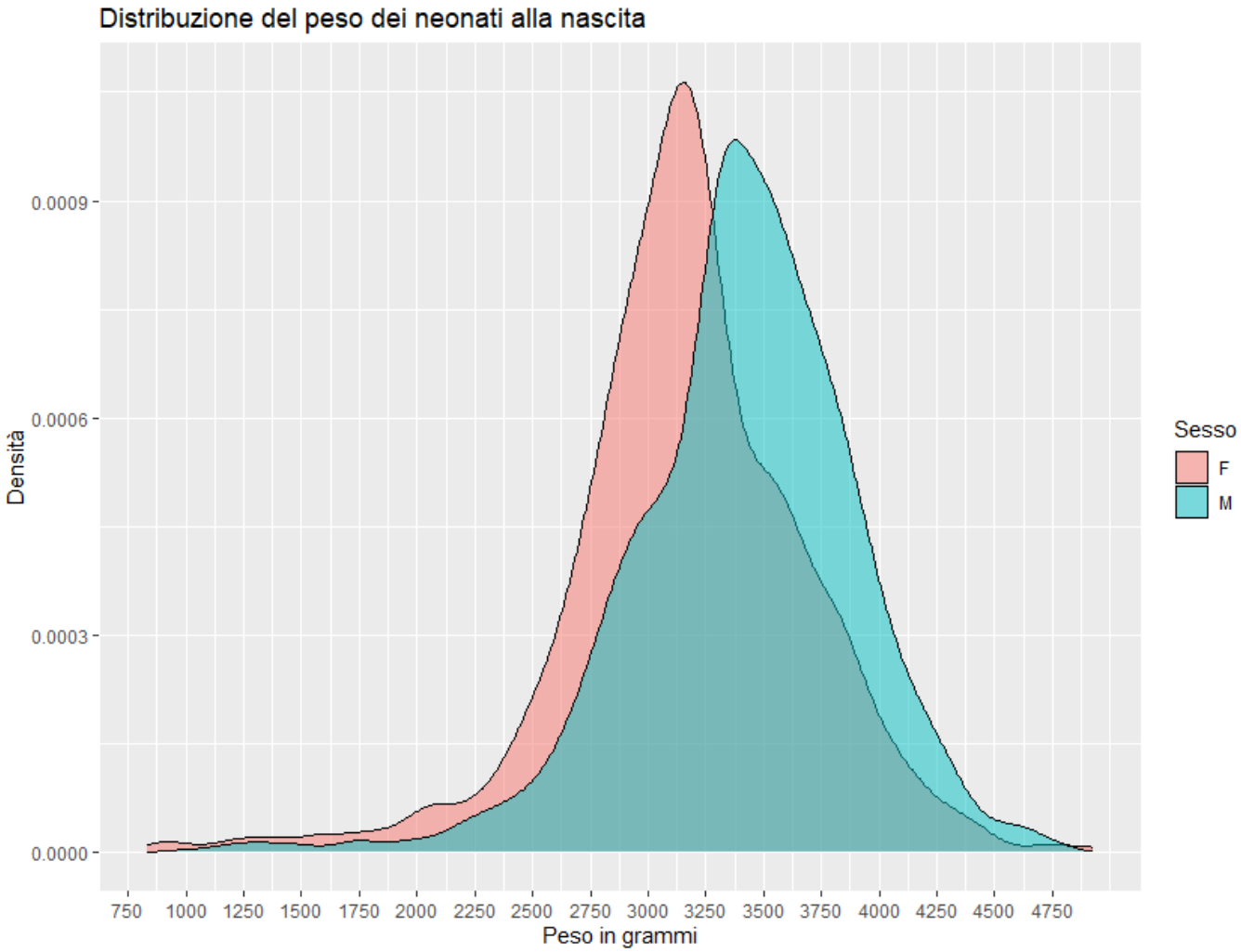
```
Ospedale
    osp1     osp2     osp3
32.66613 33.94716 33.38671
```

We also find that there are no significant differences in the number of males and females.

```
Sesso
   F    M
1255 1243
```

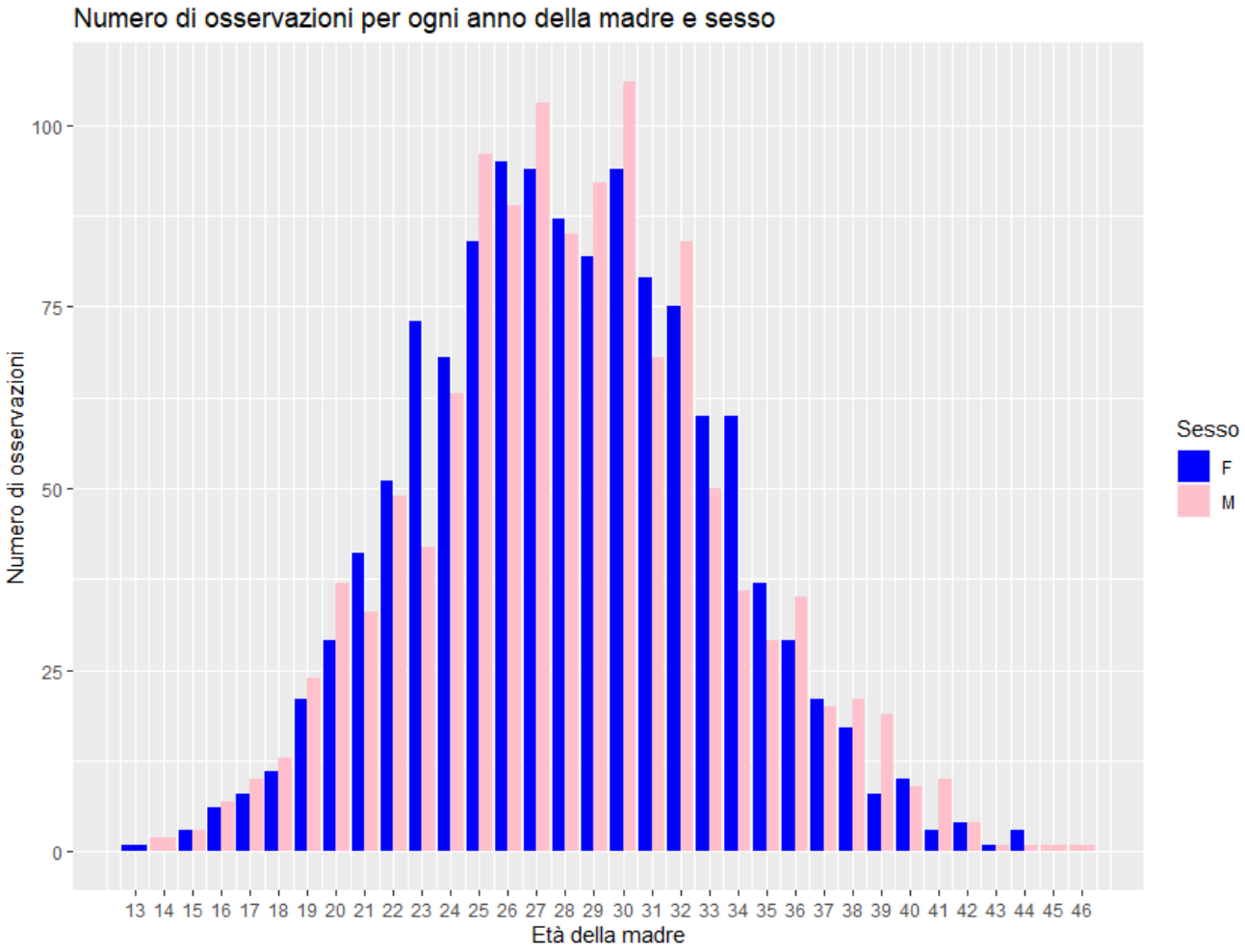
```
Sesso
       F        M
50.24019 49.75981
```

Study of graphs



In the weight density graph of infants by sex, it is evident that the female sample shows a higher peak value than the male sample. In addition, the weight distribution for females appears more uniform than that of males, which shows a more pronounced tail to the right. This suggests a greater concentration of observations near the maximum weight in male infants.

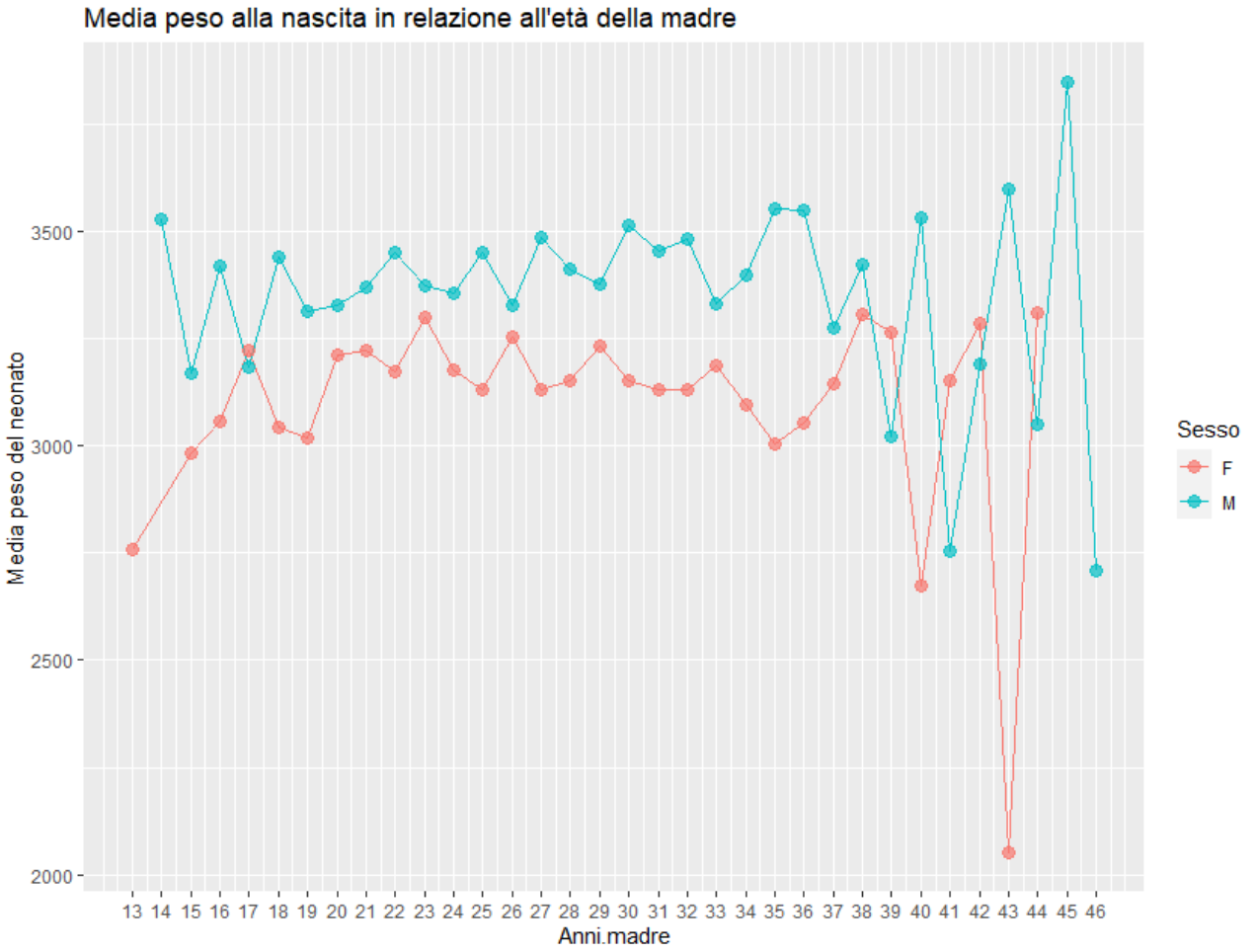
We also note that the female sample shows a leaner distribution in the middle part, indicating slightly less variability around the peak.



The histogram shows a common 'range' for the age of mothers during childbirth, between 25 and 32 years. We note that at age 33 there is a negative peak followed by a steady decline. This graph provides useful insights into the most common fertility periods.

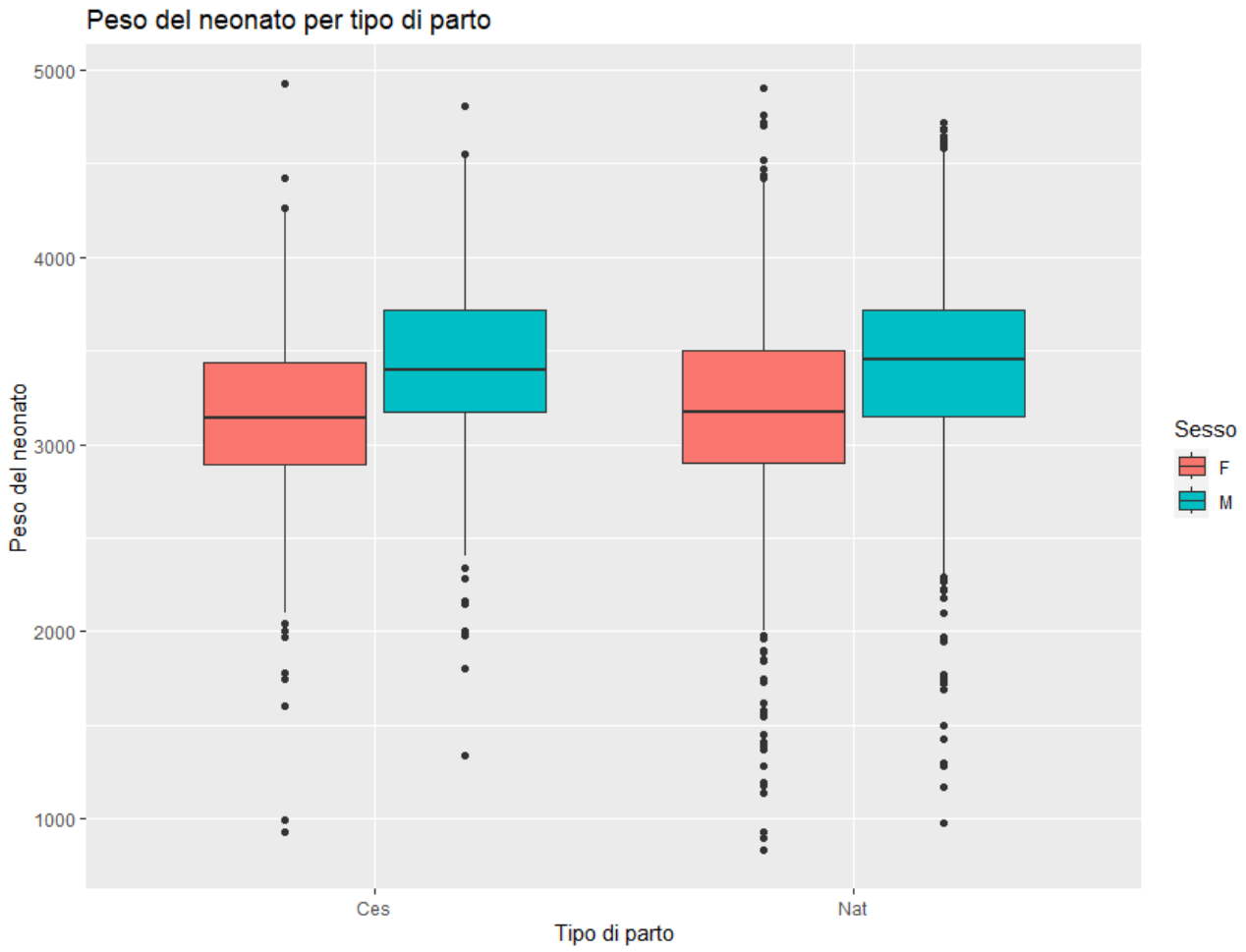
No significant differences seem to emerge regarding the sex of the infants. Although there are some isolated peaks in both cases, such as a slight concentration of female infants in 30-year-old mothers and a peak of male infants in 23-year-old mothers, no particular trends beyond normality are observed.

It is also important to consider the possible impact of the mother's socioeconomic conditions on the common 'range' of ages found. Although outside the scope of our medical study, this factor could significantly influence the results and thus deserves consideration.



In the line graph examining the weight of infants in relation to the age of the mother and the sex of the child, we noted that the number of infants born to mothers over 40 or under 20 is small, which affects the visualization of average weight. Therefore, to obtain a more accurate view, we focus on values with a larger number of observations.

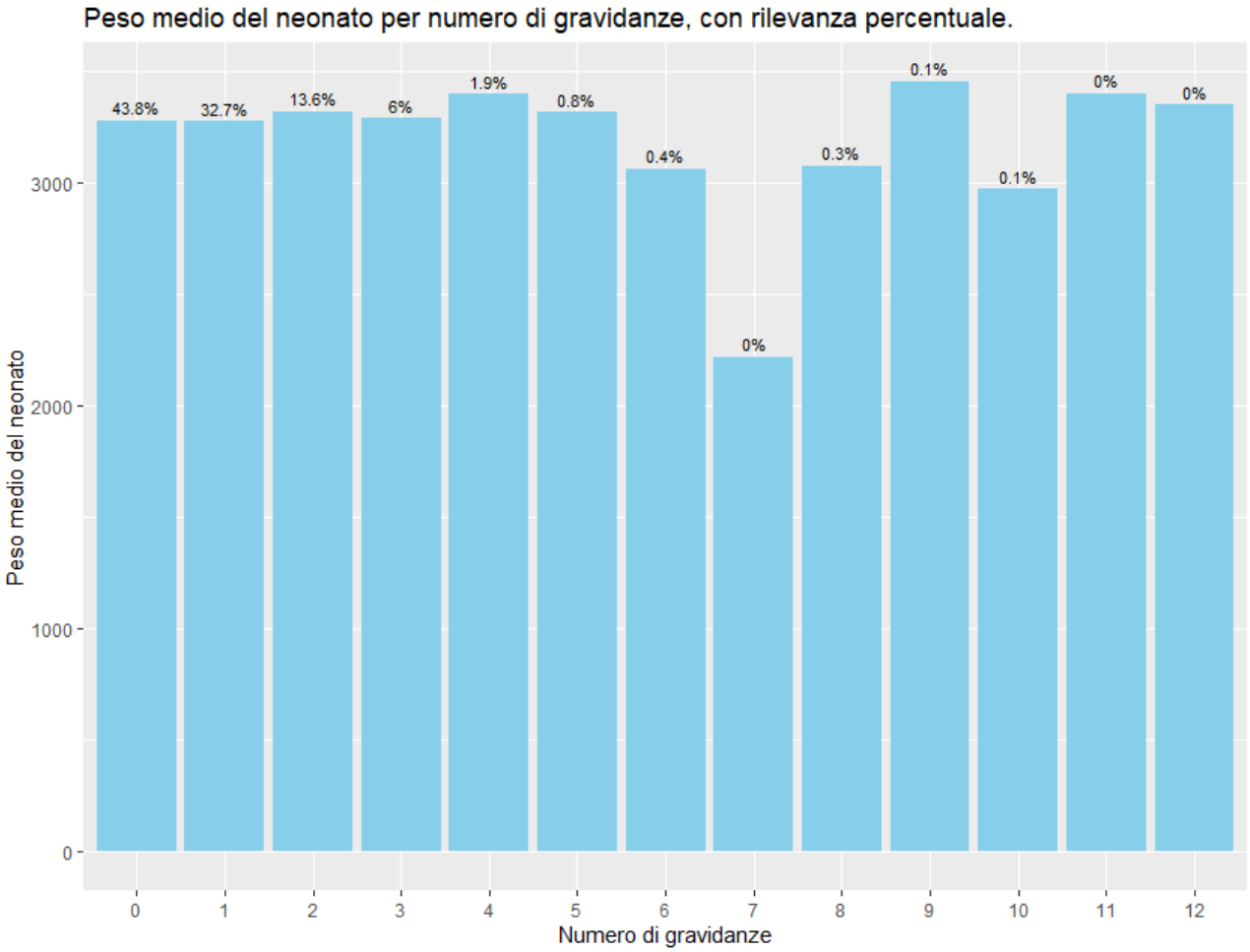
It is evident that male infants have a higher average weight than females, with a difference of about 500 grams at the point with the greatest disparity between the two sexes. However, excluding this extreme point, the variations in weight between the two sexes and between the different age groups appear to be approaching each other. In the middle part of the graph, which contains the largest number of significant observations, we note that the variations in weight between males and females are within a range of about 250 grams. Specifically, males tend to have an average weight between 3250 and 3500 grams, while for females the average weight ranges between 3000 and 3250 grams.



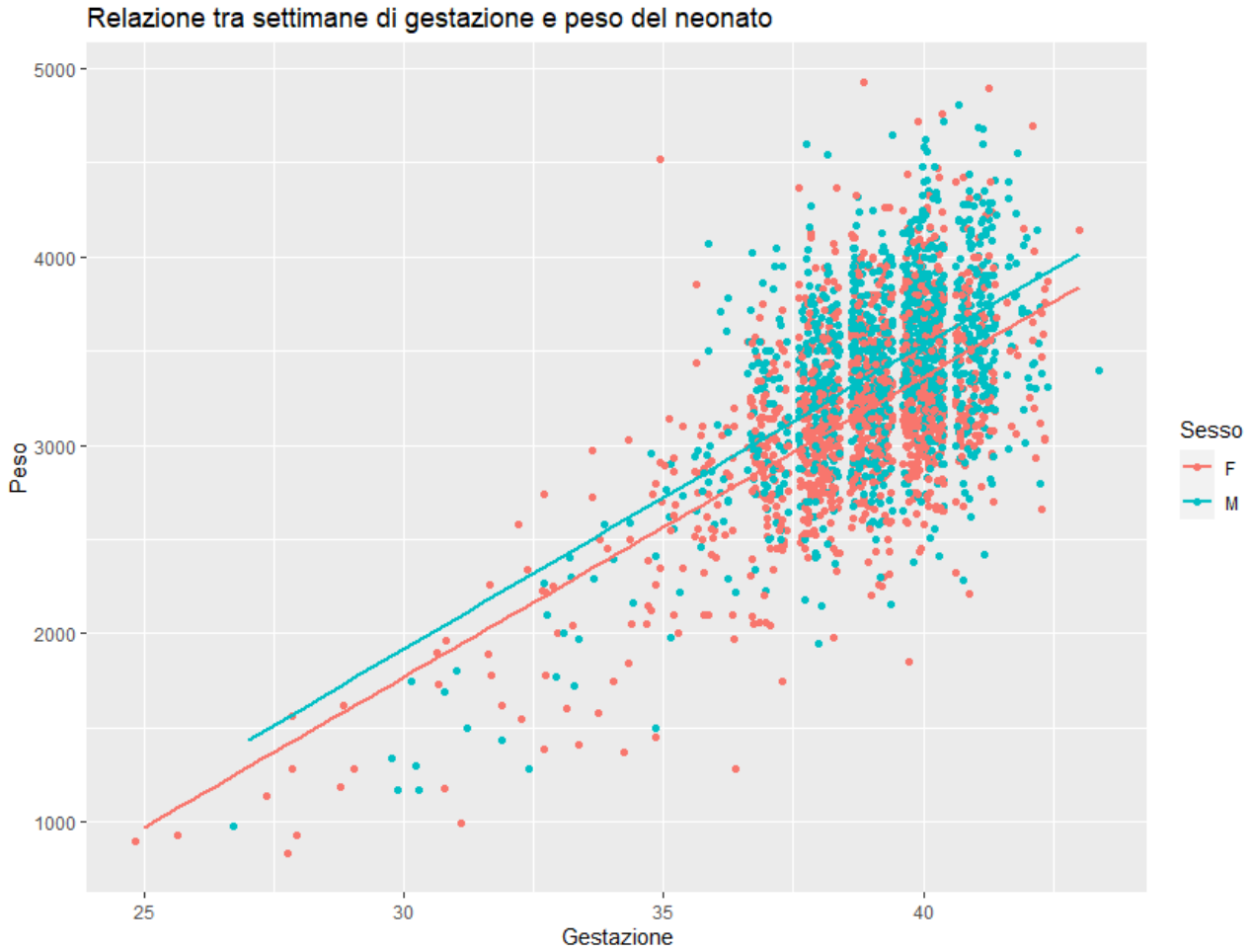
The boxplot shows no significant differences in weight in relation to type of delivery, thus suggesting no correlation between mode of delivery and infant weight, as well as for the sex of the unborn child. The distributions of weights appear to be balanced across modes of delivery.

However, it is interesting to note a detail already observed in the previous line graph: the range of weight difference by sex, here expressed in terms of interquartile range rather than mean weight. It can be seen that 50 percent of female infants weigh between 2900 and 3400 grams, while for male infants the variation is between 3200 and 3700 grams.

A further observation concerns the outliers for natural childbirth, which appear to be more numerous than the outliers for cesarean deliveries.

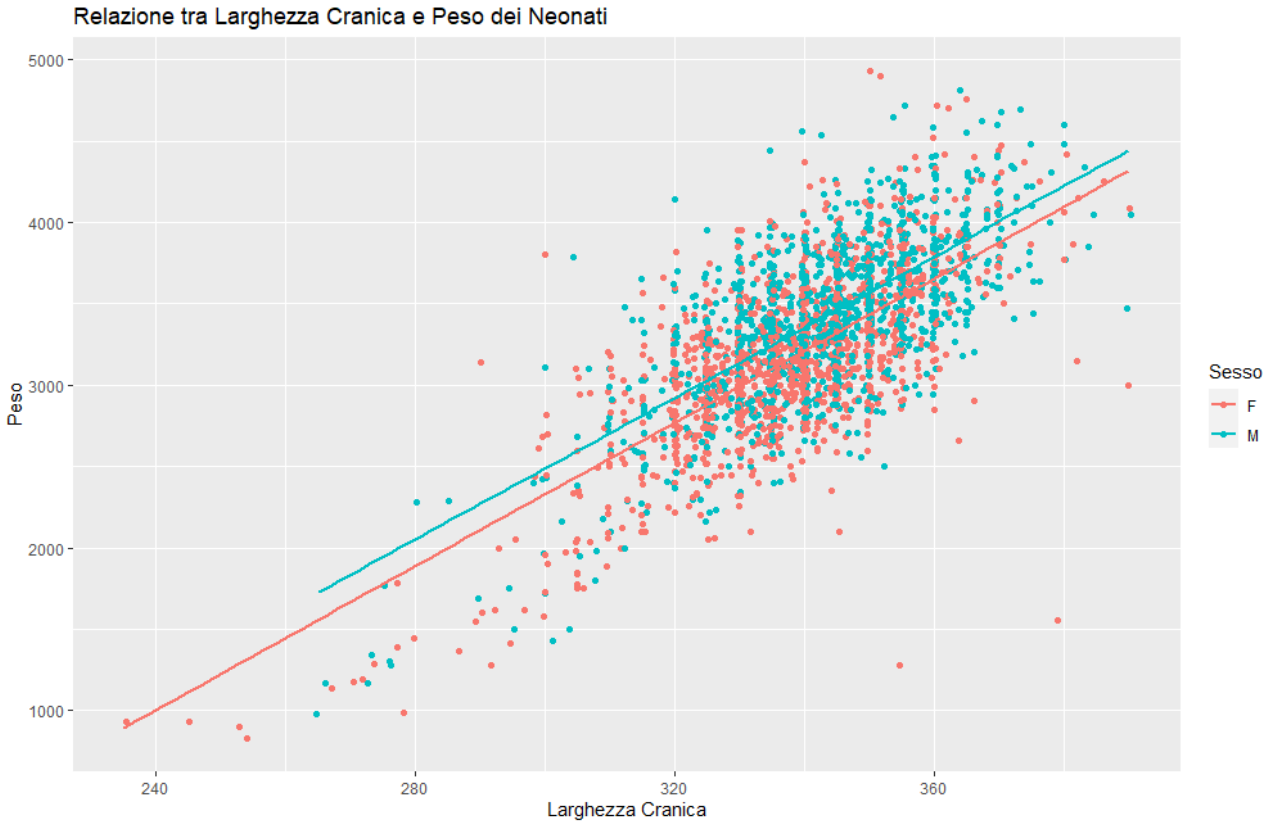


In the graph, we are examining whether there is a correlation between the number of pregnancies and average infant weight. The values are generally similar except for those who had seven pregnancies. This disparity can be attributed to the limited number of observations for this group, which does not allow for an in-depth analysis. We have included the percentage of significance of each number of pregnancies, indicating how frequently it appears in our study and its statistical significance. We immediately note that pregnancies 0, 1, 2 and 3 make up about 96 percent of the total, and that among them, variations in mean infant weight are minimal.



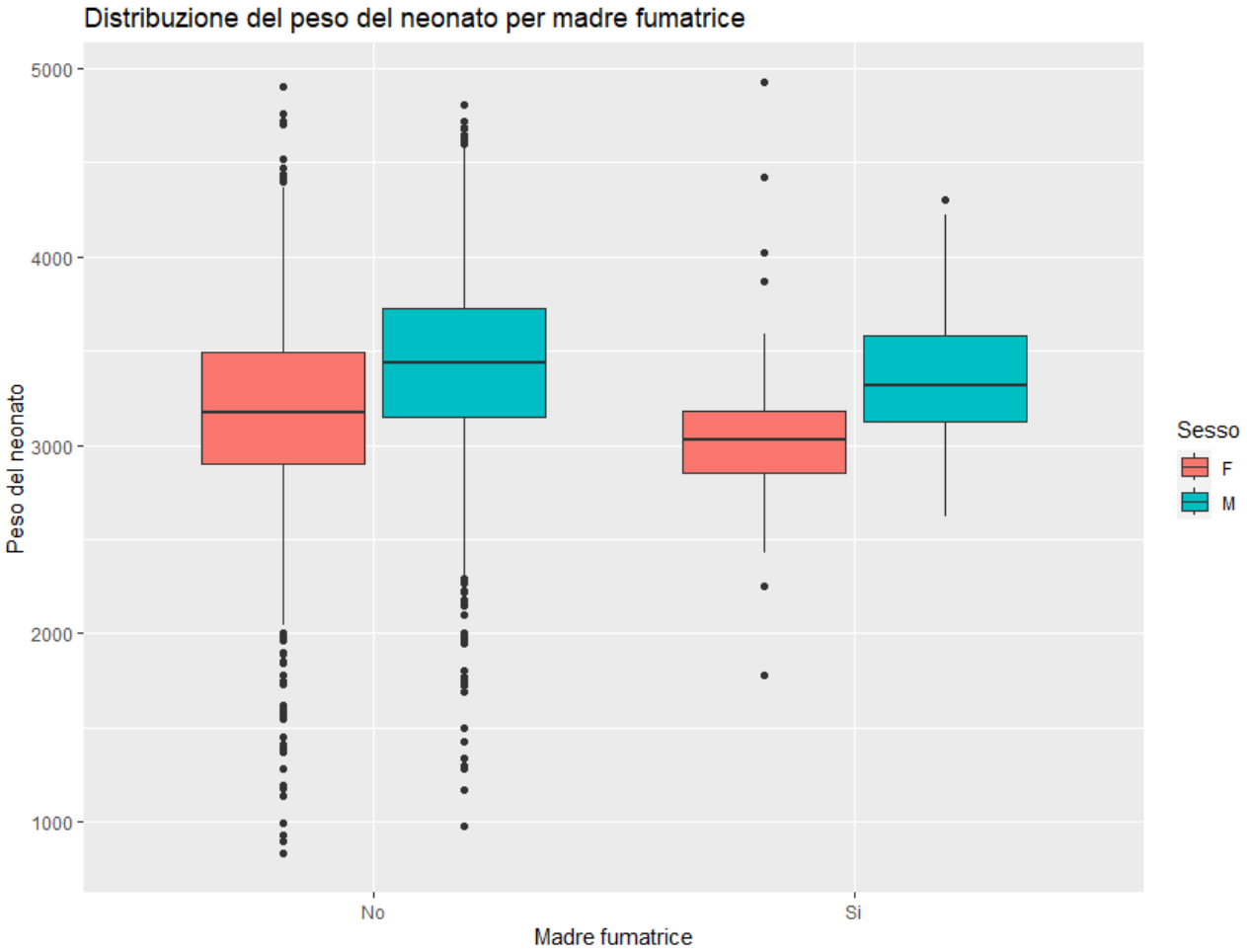
In the scatter plot analyzing the influence of weeks of gestation on infant weight, divided by sex, a positive correlation emerges. It is observed that infant weight tends to increase as the gestation period increases, as evidenced by the concentration of many observations between weeks 35 and 40. No particular patterns seem to emerge with respect to sex, with an equal distribution between the two groups. The only peculiarity, however, due to two individual observations, is the minimum gestation values in two female case histories, which lengthen the initial part of the F line.

It is important to note that early gestations show a clear decline in average infant weight; most observations between 25 and 35 weeks fall below the regression line, indicating lower weight than expected.



In the scatter plot analyzing the correlation between Cranial Width and infant weight, a clear positive correlation is observed. Skull width tends to increase as the infant's weight increases, as evidenced by the regression line that spans a compact range from about 320 to 360 for skull width and from about 2500 to 4000 grams for weight.

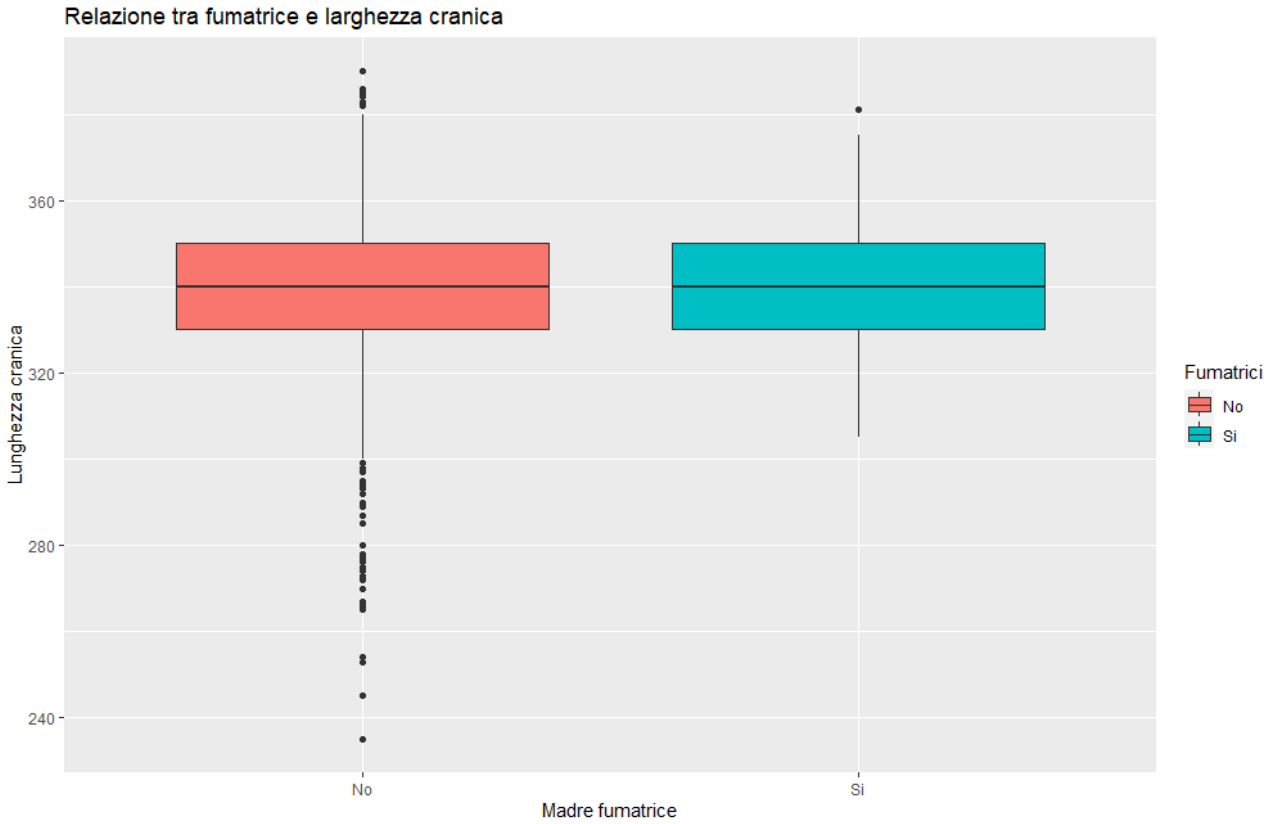
On the left side of the graph, below the regression lines, there are some observations, indicating the presence of infants with lower-than-average weight relative to their cranial width. Also, it is noted that the regression line for females is longer on the left than that for males, suggesting greater variation in cranial width for females in infants with lower weight.



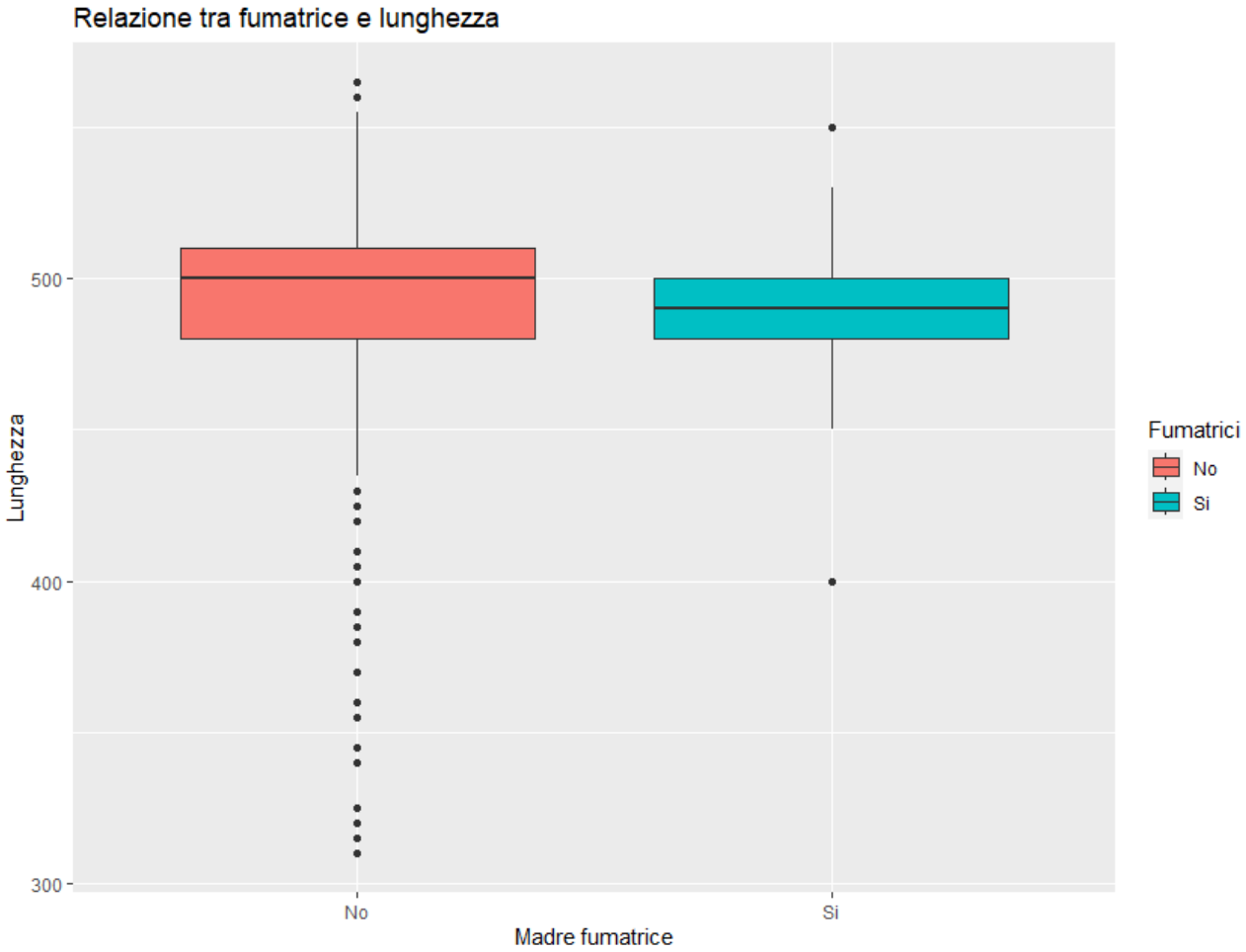
In this boxplot, we examine the relationship between infant weight, infant sex, and smoking status of the mother. The boxes, representing data in the interquartile range, are significantly larger for nonsmoking mothers than for smoking mothers, indicating higher average weights. This difference is particularly evident in the female sample.

We note a large number of outliers in smoking mothers, which could be attributed to their lower frequency in the study sample. In fact, non-smoking mothers make up 96 percent of the sample, while smokers make up only 4 percent. Consequently, there are fewer outliers in smoking mothers.

Finally, we can infer that given the length of the whiskers in the boxes for nonsmoking mothers, the data follow a normal distribution. Although there is a slight difference for smoking mothers, it is not that significant.

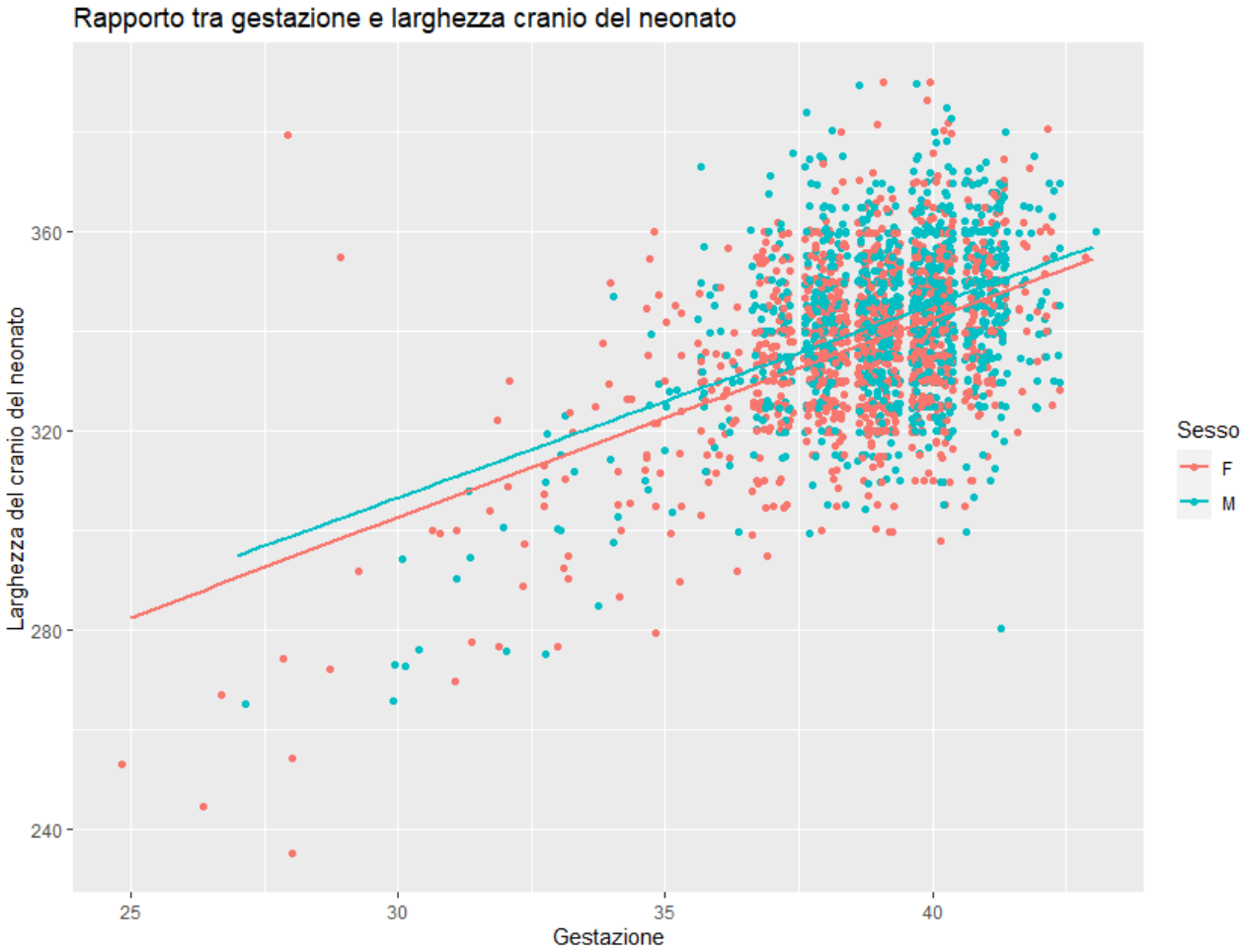


No substantial differences appear to emerge between the head widths of infants with smoking and nonsmoking mothers. The only notable peculiarity is the presence of numerous outliers for nonsmokers. This phenomenon could be attributable to the variety and quantity of data examined.



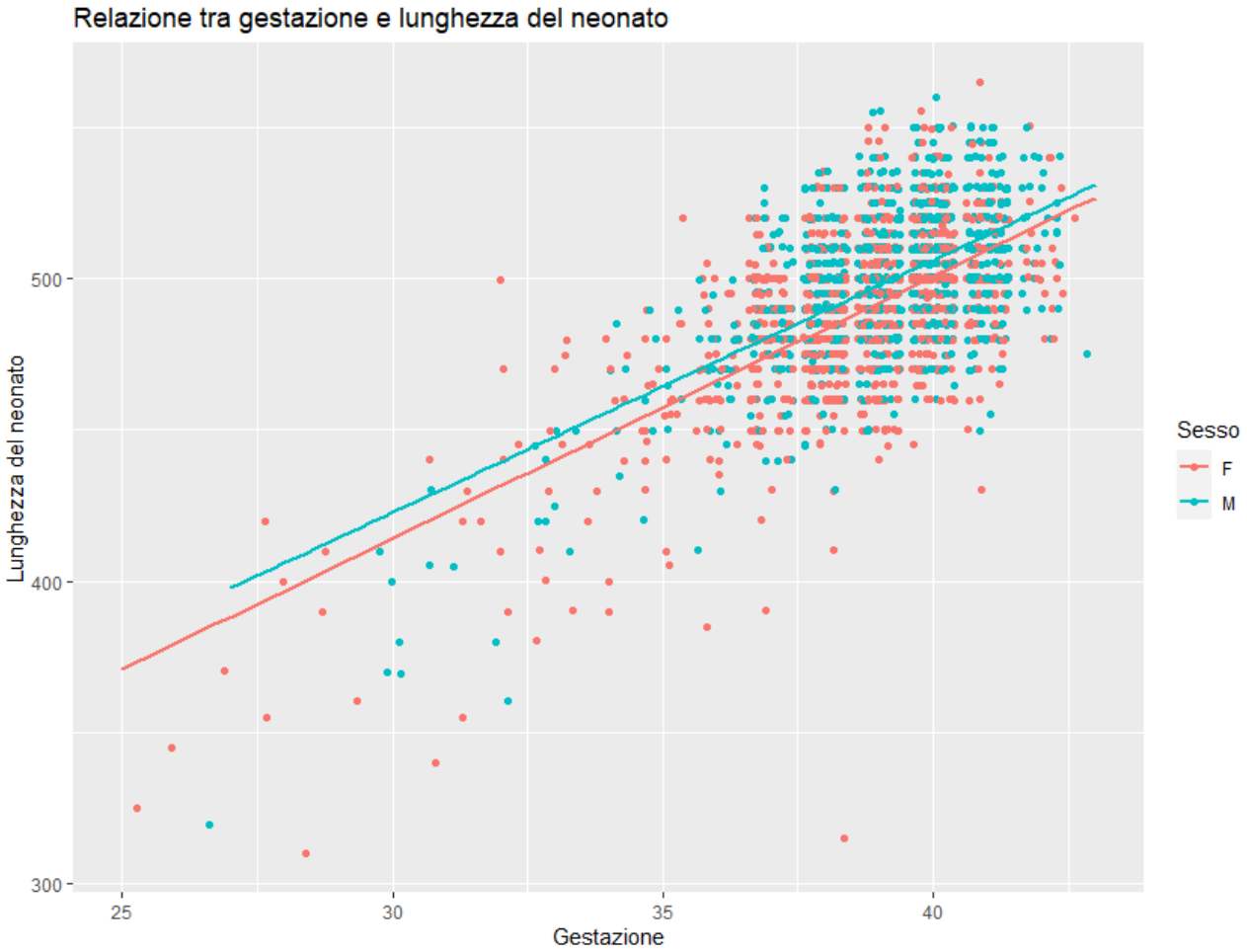
In the boxplot for nonsmoking mothers, the box is nearly twice as large as for smokers, with a fairly higher median. However, the median appears slightly skewed within the interquartile range, indicating a negative skewed distribution. This suggests that most observations are concentrated toward higher values. In contrast, in the boxplot of smoking mothers, the box seems to indicate a more symmetrical distribution, albeit with lower median values.

As observed previously in other boxplots, outliers are present mainly in the observations of nonsmoking mothers, particularly concentrated in the lower part of the graph.



We immediately notice a positive correlation between weeks of gestation and infant skull width. Most observations are concentrated in the range between 35 and 40 weeks of gestation, with skull width between 320 and 360 mm.

On the left side of the graph, below the regression lines, are some important observations. These data show that as weeks of gestation decreases, the correlation with skull width also decreases. This suggests that in infants born prematurely there may be variation in the relationship between weeks of gestation and skull size, which could have significant implications for their health and development.



Also in this case study, we observe a positive correlation between weeks of gestation and infant length at birth. Most observations are concentrated in the range of 35 to 40 weeks' gestation, with infant length between 450 and 550 mm.

It is important to dwell on a previously observed peculiarity: as weeks of gestation decrease, the correlation with infant length also decreases. This phenomenon could have significant implications for the health and development of prematurely born infants, underscoring the importance of carefully considering prematurity in newborn growth assessments.

Various hypotheses

Let's test some hypotheses, for example:

- Is the average weight and length of this sample of infants significantly equal to those of the population?

We begin by choosing the appropriate test, using a Shapiro-test to assay the normality of weight distributions.

```
> shapiro.test(Peso)

        Shapiro-Wilk normality test

data:  Peso
W = 0.97068, p-value < 2.2e-16

> shapiro.test((Peso[Sesso=="F"]))

        Shapiro-Wilk normality test

data:  (Peso[Sesso == "F"])
W = 0.96293, p-value < 2.2e-16

> shapiro.test((Peso[Sesso=="M"]))

        Shapiro-Wilk normality test

data:  (Peso[Sesso == "M"])
W = 0.96637, p-value = 2.225e-16
```

With a value of W very close to one but an extremely low p-value (< 2.2e-16) we reject the null hypothesis that the data are normally distributed. Therefore, we will use a Wilcoxon test.

We start by saving as variables the averages of our interest concerning Population, taking the data from clinical studies taken from the web. Since weight averages between males and females differ slightly, as opposed to length, we will test by sex in addition to the single test.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso
V = 1162385, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 3400
```

In this case the extremely low p-value leads us to reject the null hypothesis, and conclude that there is a significant statistical difference between the mean weight of our sample and that of the population.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso[Sesso == "F"]
V = 267680, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 3300
```

As in the previous case, we are inclined to reject the null hypothesis, and conclude that there is a significant statistical difference between your averages.

```
        Wilcoxon signed rank test with continuity correction

data:  Peso[Sesso == "M"]
V = 355593, p-value = 0.1086
alternative hypothesis: true location is not equal to 3450
```

Testing the mean weight for males brings us a p-value above the 0.05% significance level, so we do not have sufficient evidence to reject the null hypothesis. There is no statistically significant difference between the mean weight for males in the sample and the mean for the population used as a comparison.

We continue with the variable Length, and some Shapiro-tests to assay its normality.

```
> shapiro.test(Lunghezza)

        Shapiro-Wilk normality test

data:  Lunghezza
W = 0.90944, p-value < 2.2e-16

> shapiro.test(Lunghezza[Sesso=="F"])

        Shapiro-Wilk normality test

data:  Lunghezza[Sesso == "F"]
W = 0.8996, p-value < 2.2e-16

> shapiro.test(Lunghezza[Sesso=="M"])

        Shapiro-Wilk normality test

data:  Lunghezza[Sesso == "M"]
W = 0.92026, p-value < 2.2e-16
```

Such low p-values lead us to reject the hypothesis of normality of the distribution. Therefore, we will use Wilcoxon tests.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza
V = 875939, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 500
```

The two averages, Sample Length and Population Length, show a statistically significant difference.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza[Sesso == "F"]
V = 160462, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 500
```

Also for the mean inherent in the Length of Female Infants we can see, given the very low p-value, a statistically significant difference. Last we go on to study the Male Infants.

```
        Wilcoxon signed rank test with continuity correction

data:  Lunghezza[Sesso == "M"]
V = 280003, p-value = 0.1414
alternative hypothesis: true location is not equal to 500
```

In this case we do not notice statistical evidence to suggest a significant difference between the sample mean and the population mean value.

We finish by studying the variable Skull, first performing Shapiro tests to assay its normality.

```
> shapiro.test(Cranio)

        Shapiro-Wilk normality test

data:  Cranio
W = 0.96358, p-value < 2.2e-16

> shapiro.test(Cranio[Sesso=="F"])

        Shapiro-Wilk normality test

data:  Cranio[Sesso == "F"]
W = 0.95547, p-value < 2.2e-16

> shapiro.test(Cranio[Sesso=="M"])

        Shapiro-Wilk normality test

data:  Cranio[Sesso == "M"]
W = 0.97038, p-value = 2.901e-15
```

The extremely low p-values lead us to reject the normality hypothesis. Again, we will use the Wilcoxon test.

```
> wilcox_cranio

        Wilcoxon signed rank test with continuity correction

data:  Cranio
V = 469692, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350

> wilcox_cranio_femmine

        Wilcoxon signed rank test with continuity correction

data:  Cranio[Sesso == "F"]
V = 83761, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350

> wilcox_cranio_maschi

        Wilcoxon signed rank test with continuity correction

data:  Cranio[Sesso == "M"]
V = 154592, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 350
```

In all three cases we have p-values below the 0.05% significance level, so we can reject the null hypothesis. There is a statistically significant difference between the mean weight for males in the sample and the mean for the population used as a comparison.

We now test the hypothesis that more cesarean sections are performed in some hospitals. We create a table and use Pearson's Chi-square test.

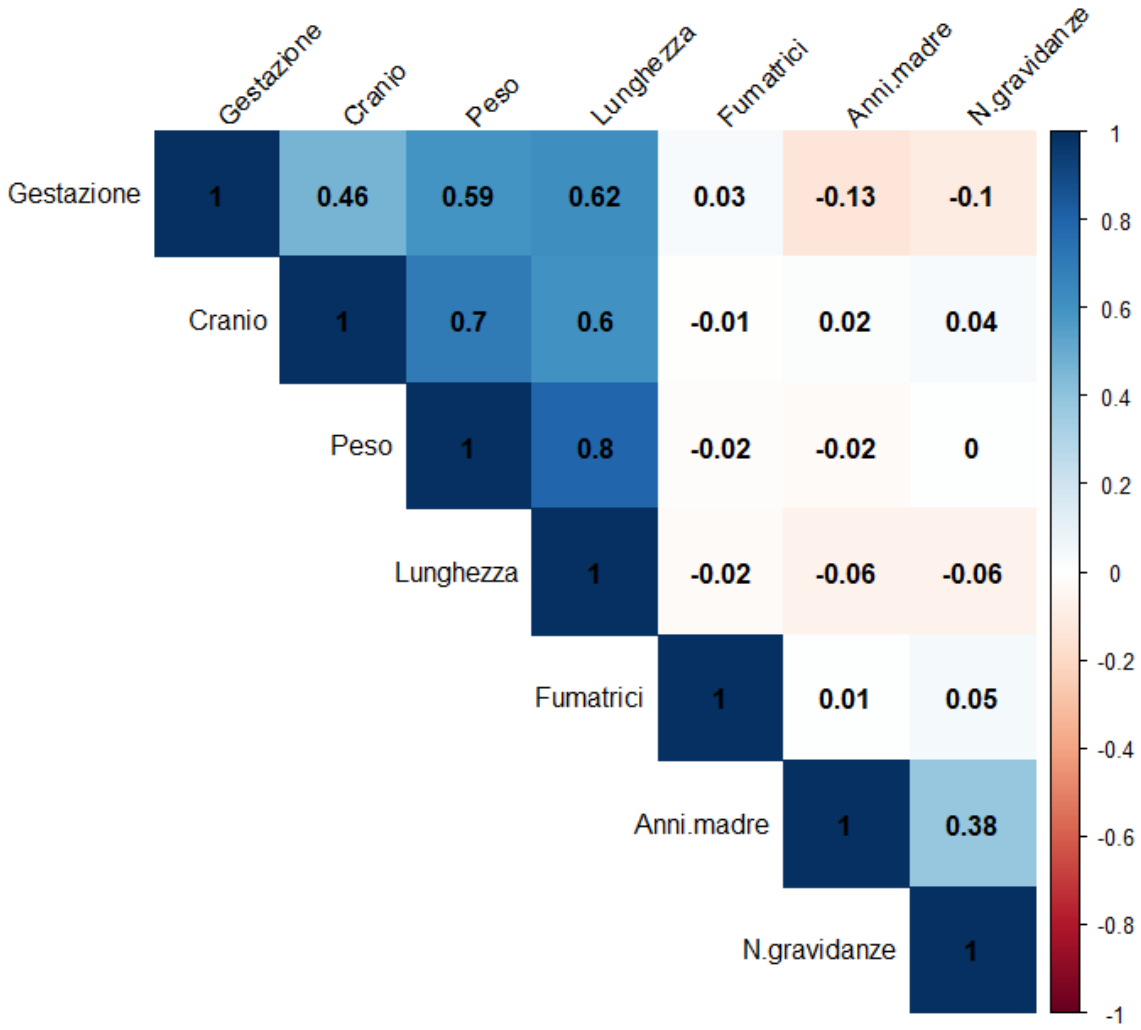
	Tipo.parto	osp1	osp2	osp3
1	Ces	242	254	232
2	Nat	574	594	602

```
> chi_square

        Pearson's Chi-squared test

data:  cont_table2[-1]
X-squared = 1.083, df = 2, p-value = 0.5819
```

Since the p-value is greater than the 5% significance level, we cannot reject the null hypothesis. Therefore, we cannot assert that there is a significant association between the type of delivery (natural or cesarean) and the hospital where the delivery takes place.

Multidimensional analysis

We investigate the relationships between the variables, focusing primarily on the variable Weight. To do this we will use a correlation matrix, a heatmap.



One of the strongest correlations is between Weight and Gestation, which shows us that as the gestation period increases, the weight of the newborn tends to increase.

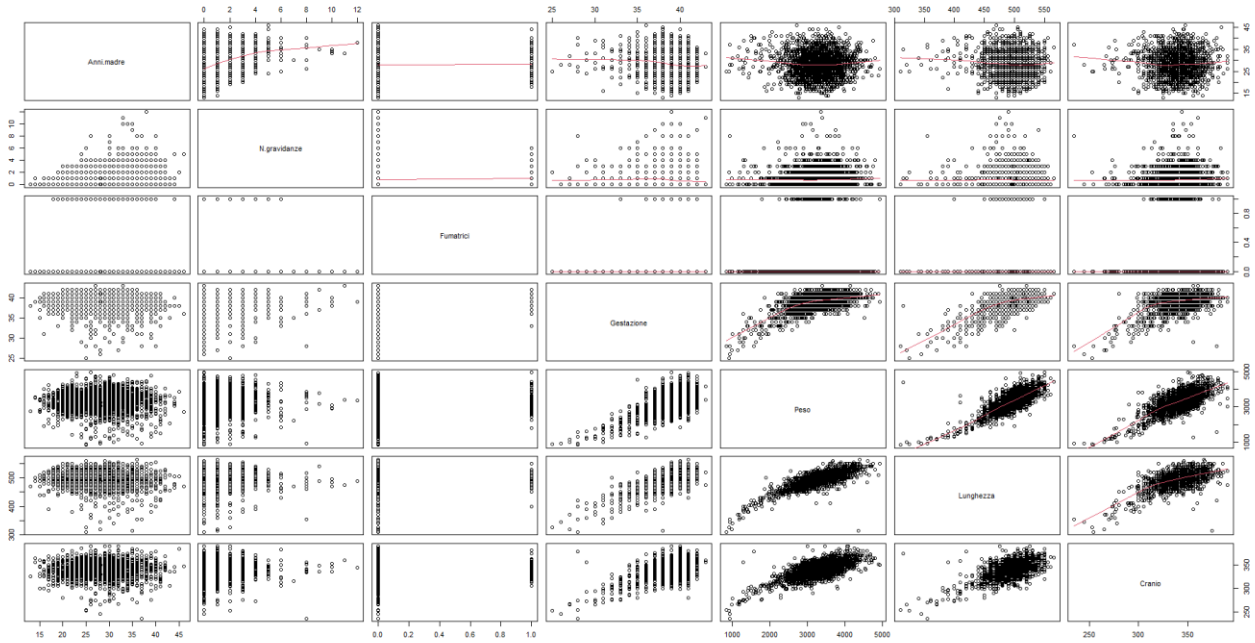
There is also a very similar positive correlation between Gestation and Length, with infants born after longer gestations demonstrating greater length.

Weight and Length are also strongly correlated, as heavier infants tend to be longer. This assumption is also present in the correlation between Weight and Skull, which is strongly positive: larger infants obviously tend to have larger skulls.

There appears to be virtually no correlation, positive or negative, between Smokers and the other variables. This suggests to us the absence of directional evidence of the impact of smoking on the length, weight, gestation period, or skull diameter of the infant.

Almost the same can be said for the variable Mother Years, which shows the highest coefficient with N.pregnancies, suggesting a modest relationship between the mother's age and the number of pregnancies she has had.

Several very weak correlations, close to zero, are also present. One of interest is that between Gestation and Mother Years, which suggests to us that older mothers may tend to have slightly shorter gestation periods.



Given the size of the graph I recommend viewing it on R. Some nonlinear effects are present, especially between Length and Skull, Gestation and Length, and Gestation and Skull.

We create a regression model with all variables.

```
mod1 <- lm(Peso ~ . , data=neonati)
summary(mod1)
```

```
> summary(mod1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-1123.3  -181.2   -14.6   160.7  2612.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6735.1400   141.3974 -47.633  < 2e-16 ***
Anni.madre        0.7975     1.1463   0.696   0.4867
N.gravidanze     11.4130     4.6665   2.446   0.0145 *
Fumatrici       -30.1567    27.5396  -1.095   0.2736
Gestazione       32.5262     3.8179   8.519  < 2e-16 ***
Lunghezza        10.2951     0.3007  34.237  < 2e-16 ***
Cranio           10.4725     0.4261  24.580  < 2e-16 ***
Tipo.partoNat    29.5025    12.0848   2.441   0.0147 *
Ospedaleosp2    -11.2217    13.4388  -0.835   0.4038
Ospedaleosp3     28.0985    13.4972   2.082   0.0375 *
SessoM           77.5473    11.1779   6.938 5.07e-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: 669.1 on 10 and 2489 DF,  p-value: < 2.2e-16
```

The model has a good overall R-square, but there are several variables that do not contribute significantly with the model, given their p-values above or near the 0.05 threshold. We will then proceed to stepwise eliminate these values, and create as many models.

```
mod2 <- update(mod1,~.-Anni.madre)
summary(mod2)
anova(mod2,mod1)
BIC(mod2,mod1)
car::vif(mod2)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6708.1065   135.9394 -49.346  < 2e-16 ***
N.gravidanze     12.6085     4.3381   2.906  0.00369 **
Fumatrici       -30.3092    27.5359  -1.101  0.27113
Gestazione       32.2501     3.7968   8.494  < 2e-16 ***
Lunghezza        10.2944     0.3007  34.239  < 2e-16 ***
Cranio           10.4876     0.4255  24.651  < 2e-16 ***
Tipo.partoNat    29.5351    12.0834   2.444  0.01458 *
Ospedaleosp2    -11.0816    13.4359  -0.825  0.40957
Ospedaleosp3     28.3660    13.4903   2.103  0.03559 *
SessoM           77.6205    11.1763   6.945 4.81e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 273.9 on 2490 degrees of freedom
Multiple R-squared:  0.7288,    Adjusted R-squared:  0.7278
F-statistic: 743.6 on 9 and 2490 DF,  p-value: < 2.2e-16
```

By removing Anni.madre the adjusted r-square did not vary, but N.gravidities seems to be more significant. We use an ANOVA test to see if removing the variable significantly helps the model.

```
> anova(mod2,mod1)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Ospedale + Sesso
Model 2: Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
    Cranio + Tipo.parto + Ospedale + Sesso
  Res.Df       RSS Df Sum of Sq     F Pr(>F)
1   2490 186809099
2   2489 186772779  1     36321 0.484 0.4867
```

The RSS varies hardly at all, while the F-value is rather low, indicating that the variable Anni.madre does not contribute significantly to the model. The p-value above the threshold of 0.05 suggests to us that the addition of the variable does not improve the predictive model.

We use the BIC model to compare the two models.

```
> BIC(mod2,mod1)
     df      BIC
mod2 11 35234.64
mod1 12 35241.97
```

With a lower BIC, mod2 remains preferable to mod1, which, despite including an additional parameter, does not improve the model enough to justify the addition.

Finally, we perform a VIF test, an index that measures whether multicollinearity is present.

```
> car::vif(mod2)
                GVIF Df GVIF^(1/(2*Df))
N.gravidanze 1.027985  1        1.013896
Fumatrici    1.007346  1        1.003666
Gestazione   1.676688  1        1.294870
Lunghezza    2.085755  1        1.444214
Cranio       1.626661  1        1.275406
Tipo.parto   1.004240  1        1.002118
Ospedale     1.003421  2        1.000854
Sesso        1.040558  1        1.020077
```

There are no VIFs greater than 5, so the test can be said to have passed.

We update the model by removing the Hospital variable.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6708.074    135.984 -49.330  < 2e-16 ***
N.gravidanze   13.012      4.342   2.997  0.00276 **
Fumatrici     -31.759     27.570  -1.152  0.24946
Gestazione     32.541      3.801   8.561  < 2e-16 ***
Lunghezza      10.272      0.301  34.129  < 2e-16 ***
Cranio         10.501      0.426  24.648  < 2e-16 ***
Tipo.partoNat  30.296     12.098   2.504  0.01234 *
SessoM         78.114     11.191   6.980 3.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.3 on 2492 degrees of freedom
Multiple R-squared:  0.7278,    Adjusted R-squared:  0.7271
F-statistic:   952 on 7 and 2492 DF,  p-value: < 2.2e-16
```

The r framework has decreased very little, and we have a simpler model.

```
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Ospedale + Sesso
  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1   2492 187501837
2   2490 186809099  2    692738 4.6168 0.009969 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Although the p-value indicates that the Hospital variable is significant, we saw in the descriptive part of our project how the average weight does not vary by hospital facility.

We create a new model (mod4) by extracting the variable Type.parto.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6681.6714   135.7178 -49.232  < 2e-16 ***
N.gravidanze   12.7185     4.3450   2.927  0.00345 **
Fumatrici     -30.4634    27.5948  -1.104  0.26972
Gestazione     32.5914     3.8051   8.565  < 2e-16 ***
Lunghezza      10.2341     0.3009  34.011  < 2e-16 ***
Cranio         10.5359     0.4262  24.718  < 2e-16 ***
SessoM         78.1713    11.2028   6.978 3.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.6 on 2493 degrees of freedom
Multiple R-squared:  0.7271,    Adjusted R-squared:  0.7265
F-statistic:  1107 on 6 and 2493 DF,  p-value: < 2.2e-16
```

```
> anova(mod4,mod3)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Tipo.parto + Sesso
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1   2493 187973654
2   2492 187501837  1    471817 6.2707 0.01234 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The adjusted r framework remained virtually unchanged, and the model is simpler. In addition, in the descriptive part of the project, we could see graphically how there was no difference in the mean weight of infants due to delivery type. Although the ANOVA test indicates that the variable Type.delivery is statistically significant the result is much closer to the 0.05 threshold than the other variables; I therefore decide to exclude it from the model, preferring to continue with a simpler one with more statistically robust variables.

In addition, there is also a slight improvement in BIC and no multicollinearity among the variables.

```
> BIC(mod4,mod3,mod2,mod1)
     df      BIC
mod4  8 35226.70
mod3  9 35228.24
mod2 11 35234.64
mod1 12 35241.97
```

```
> car::vif(mod4)
N.gravidanze    Fumatrici   Gestazione    Lunghezza       Cranio       Sesso
    1.026120     1.006607     1.675575     2.078644     1.624603     1.040271
```

Although Smokers is a non-significant variable for the time being, it has been retained as a control variable. Let us try to exclude it to see if this leads to a significant improvement in the model.

```
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
```

```
> anova(mod4nofum, mod4)
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1   2494 188065546
2   2493 187973654  1     91892 1.2187 0.2697
> BIC(mod4nofum,mod4,mod3,mod2,mod1)
          df      BIC
mod4nofum  7 35220.10
mod4       8 35226.70
mod3       9 35228.24
mod2      11 35234.64
mod1      12 35241.97
```

The adjusted r square did not vary. According to the ANOVA test, the variable Smokers does not significantly improve the model's ability to predict infants' weight. In addition, the BIC without this variable improved significantly.

Nevertheless, I decide to keep it as a control variable since it is a clinical study, and also taking into account the results obtained in the study boxplot graphs, in which a lower average weight for infants born to smoking mothers was clear.

I proceed by creating several models that consider nonlinear interactions and effects. For clarity of the project I will not report all the models here; they can be found in the .R file. One of the best turns out to be this one:

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

Residuals:
     Min       1Q   Median       3Q      Max
-1170.35  -181.95   -11.83   162.98  1785.71

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    198.983315 723.822749   0.275 0.783411
N.gravidanze    14.285516   4.269659   3.346 0.000833 ***
Fumatrici      -23.907419  27.104975  -0.882 0.377845
Gestazione      42.685197   3.879372  11.003  < 2e-16 ***
Cranio          10.646792   0.418709  25.428  < 2e-16 ***
Lunghezza      -20.214842   3.162204  -6.393 1.94e-10 ***
I(Lunghezza^2)   0.031592   0.003267   9.671  < 2e-16 ***
SessoM          70.165729  11.031579   6.360 2.39e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269.6 on 2492 degrees of freedom
Multiple R-squared:  0.737,     Adjusted R-squared:  0.7363
F-statistic: 997.6 on 7 and 2492 DF,  p-value: < 2.2e-16
```

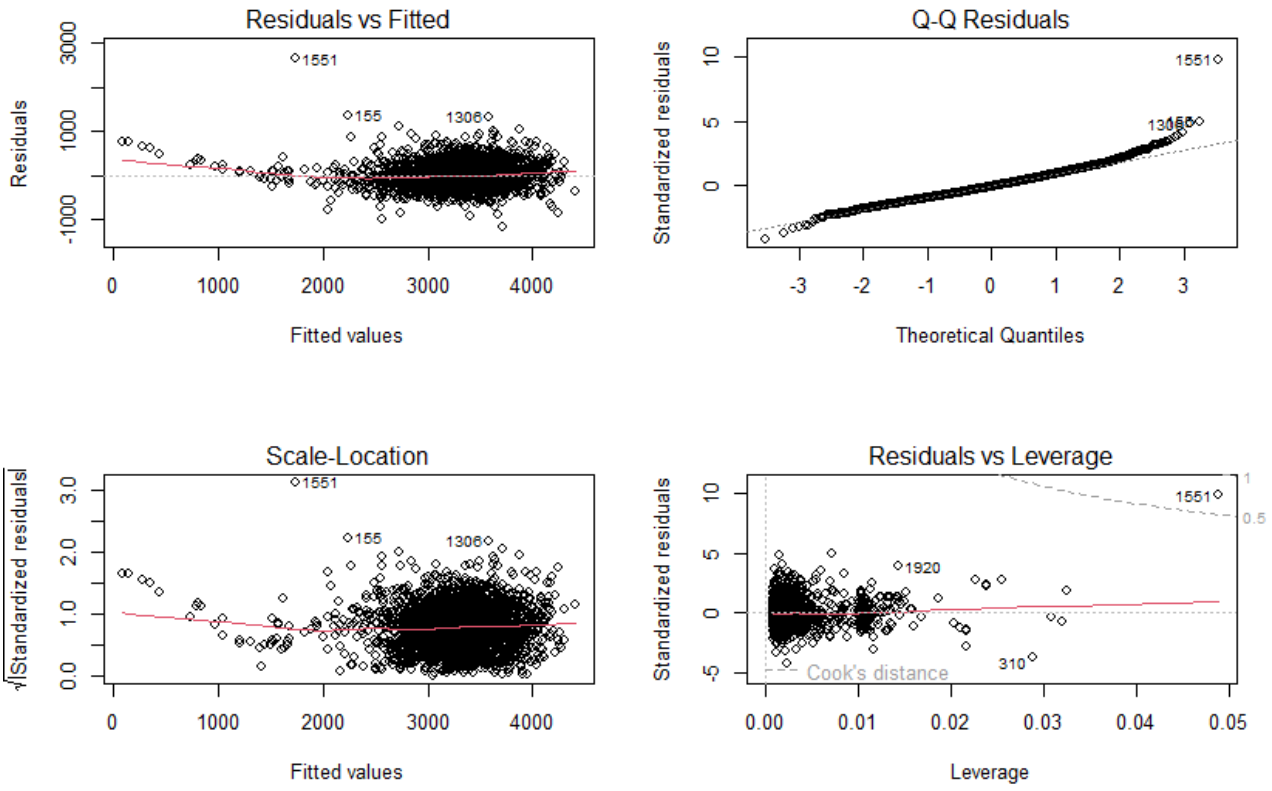
```
Analysis of Variance Table

Model 1: Peso ~ N.gravidanze + Fumatrici + Gestazione + Cranio + Lunghezza +
    I(Lunghezza^2) + Sesso
Model 2: Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
    Sesso
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   2492 181173497
2   2493 187973654 -1  -6800157 93.535 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model has a better adjusted r square than mod4, but it is also more complex. With the ANOVA test we see how the more complex model provides a better fit to the data in a statistically significant way.

```
> BIC(mod_non_lin2,mod4)
             df      BIC
mod_non_lin2  9 35142.41
mod4          8 35226.70
> car::vif(mod_non_lin2)
  N.gravidanze      Fumatrici     Gestazione
      1.027600       1.007237       1.806315
        Cranio      Lunghezza I(Lunghezza^2)
      1.625823     238.081896     230.152216
         Sesso
      1.046162
```

The BIC of mod_not_lin2 is certainly better than that of mod4, but the VIF test shows us some issues. Length and its square exhibit strong multicollinearity. Therefore, we will prefer mod4.

Let's analyze the residuals.



Clear heteroschedasticity is present, with the variance of the residuals not constant. Studying the Q-Q we note a slight positive tail, but for the most part it indicates a normal distribution.

Even with Scale-Location we notice Heteroschedasticity. Finally for Residuals vs Leverage we can note observation 1551, above the 0.5 threshold of cook distance.

```
        Shapiro-Wilk normality test

data:  residuals(mod4)
W = 0.9741, p-value < 2.2e-16
```

According to Shapiro-wilk we reject the null hypothesis that the residuals follow a normal distribution.

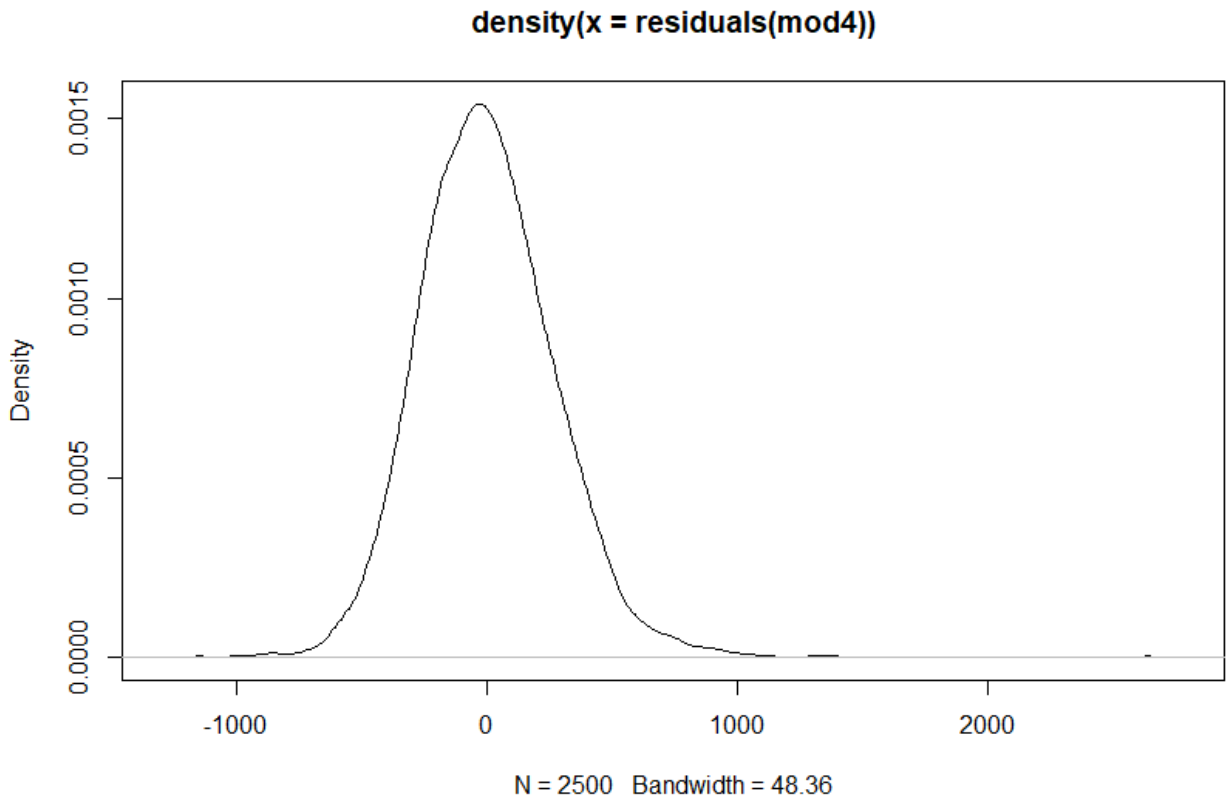
```
        studentized Breusch-Pagan test

data:  mod4
BP = 89.798, df = 6, p-value < 2.2e-16
```

The test suggests to us a heteroschedasticity of the residuals, so the variance of the residuals is not constant.

```
        Durbin-Watson test

data:  mod4
DW = 1.9542, p-value = 0.126
alternative hypothesis: true autocorrelation is greater than 0
```

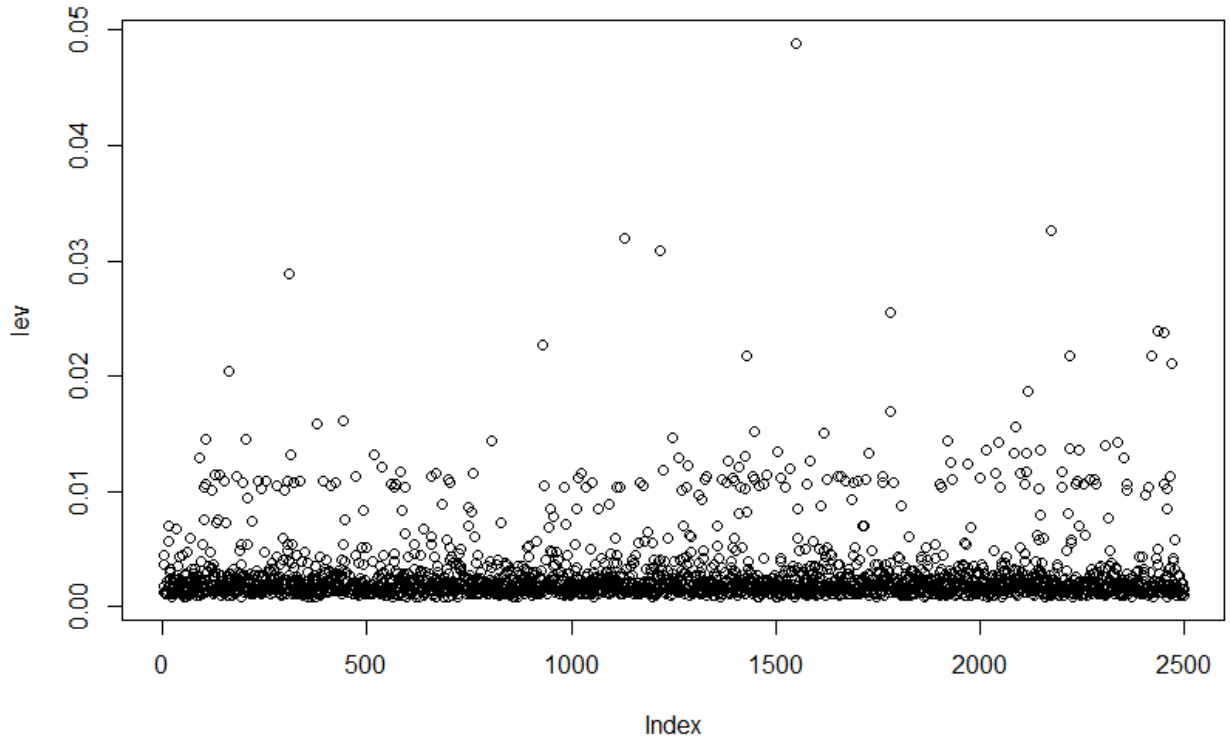
The Durbin-Watson test indicates the absence of autocorrelation among the model residuals.

Let us go on to study the density plot of the model residuals.



It turns out to be in a normal distribution, although tails are present especially on the right side of the density plot, stretched for a single observation (the 1551).

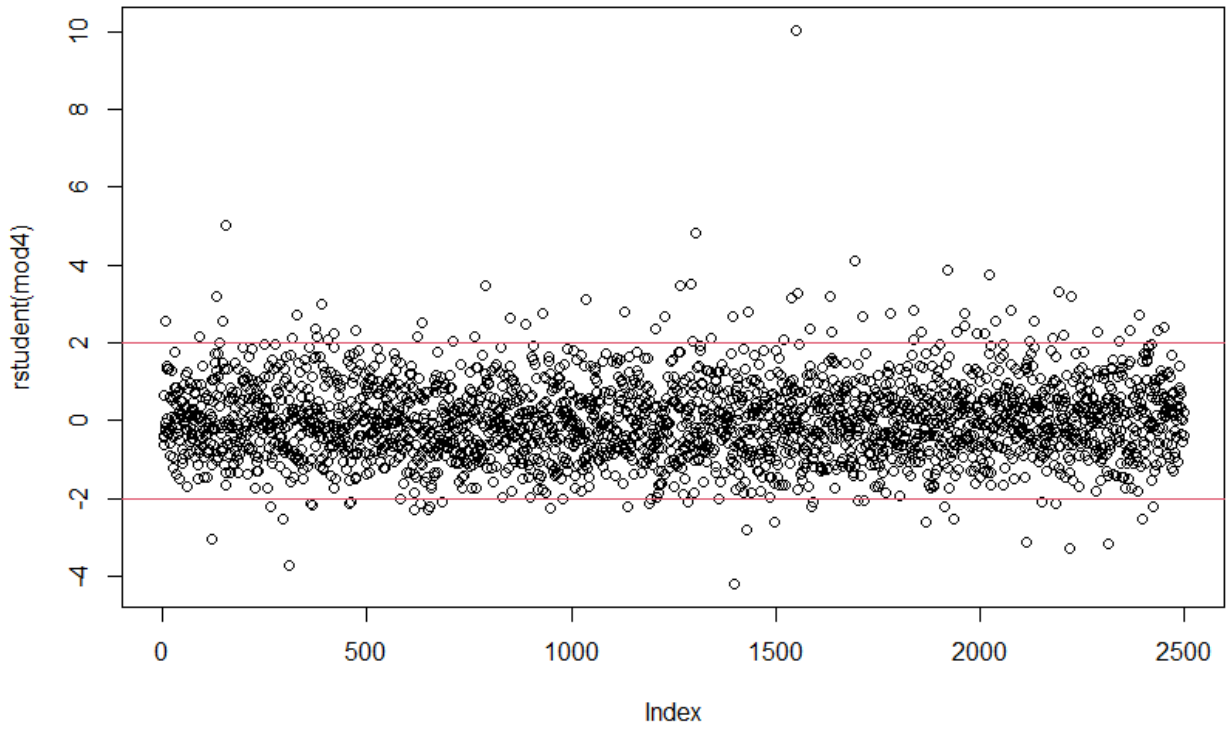
Let's go to study leverage.



There are no values above the threshold, which, moreover, is not even displayed in the plot. For confirmation, let us recall the leverage above the threshold, which should be zero.

```
> lev[lev>soglia]
named numeric(0)
```

Let us turn to the outliers.




```
> car::outlierTest(mod4)
      rstudent unadjusted p-value Bonferroni p
1551 10.039719         2.8060e-23   7.0149e-20
155   5.022108         5.4723e-07   1.3681e-03
1306  4.823102         1.4986e-06   3.7465e-03
>
```

Observation 1551 has a very high residual, significantly different from zero, and a very low Bonferroni. It is therefore an outlier.

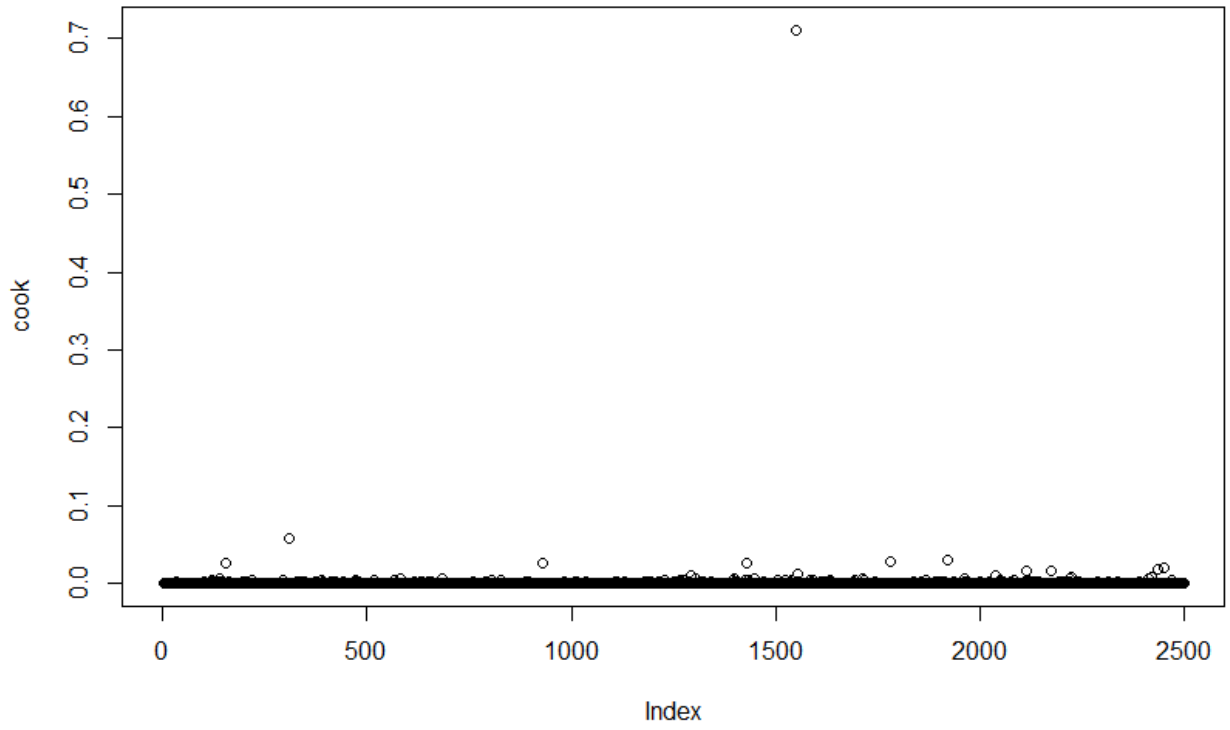
For observations 155 and 1306 we also have rather high residuals, and they turn out to be non-zero. Although the Bonferroni is not as low as for observation 1551 they too can be classified as outliers.

Let us study them in more detail.

	Anni.madre	N.gravidanze	Fumatrici	Gestazione	Peso	Lunghezza	Cranio	Tipo.parto	Ospedale	Sesso
1551	35	1	0	38	4370	315	374	Nat	osp3	F
155	30	0	0	36	3610	410	330	Nat	osp1	M
1306	23	0	0	41	4900	510	352	Nat	osp2	F

1306 shows a weight of 4900 grams and a length of 510mm, high values for an infant. In the other two observations these values are also different from the average, which could be the reason for their recognition as outliers. Skull does not seem very far from the median values, as do Anni.mother and N.pregnancies.

We also study the cook distance by a graph, and recall the maximum value.



The maximum distance is 0.7, quite high.

Let us study the problematic observations individually.

```
> print(righe_interessate)
     Anni.madre N.gravidanze Fumatrici Gestazione Peso Lunghezza Cranio Tipo.parto Ospedale Sesso
1551         35            1         0         38 4370       315    374        Nat     osp3     F
155          30            0         0         36 3610       410    330        Nat     osp1     M
1306         23            0         0         41 4900       510    352        Nat     osp2     F
310          40            3         0         28 1560       420    379        Nat     osp3     F
```

Although they present data above the mean, these are not physiologically incorrect values. Later we will study a new model without observation 1551 and observation 310 to clarify any variations.

Forecast mod4.

Let us now attempt a weight prediction using model 4, for a female infant whose mother has already had three pregnancies and is at 39 weeks gestation. We will make the prediction for both smoking and nonsmoking mother.

```
nuovi_neonatifumSI <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 1,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
  )

previsionifumSI <- predict(mod4, newdata=nuovi_neonatifumSI)
```

The model calculates that the newborn should weigh 3242.302 grams. Let's turn to the model for nonsmoking mother.

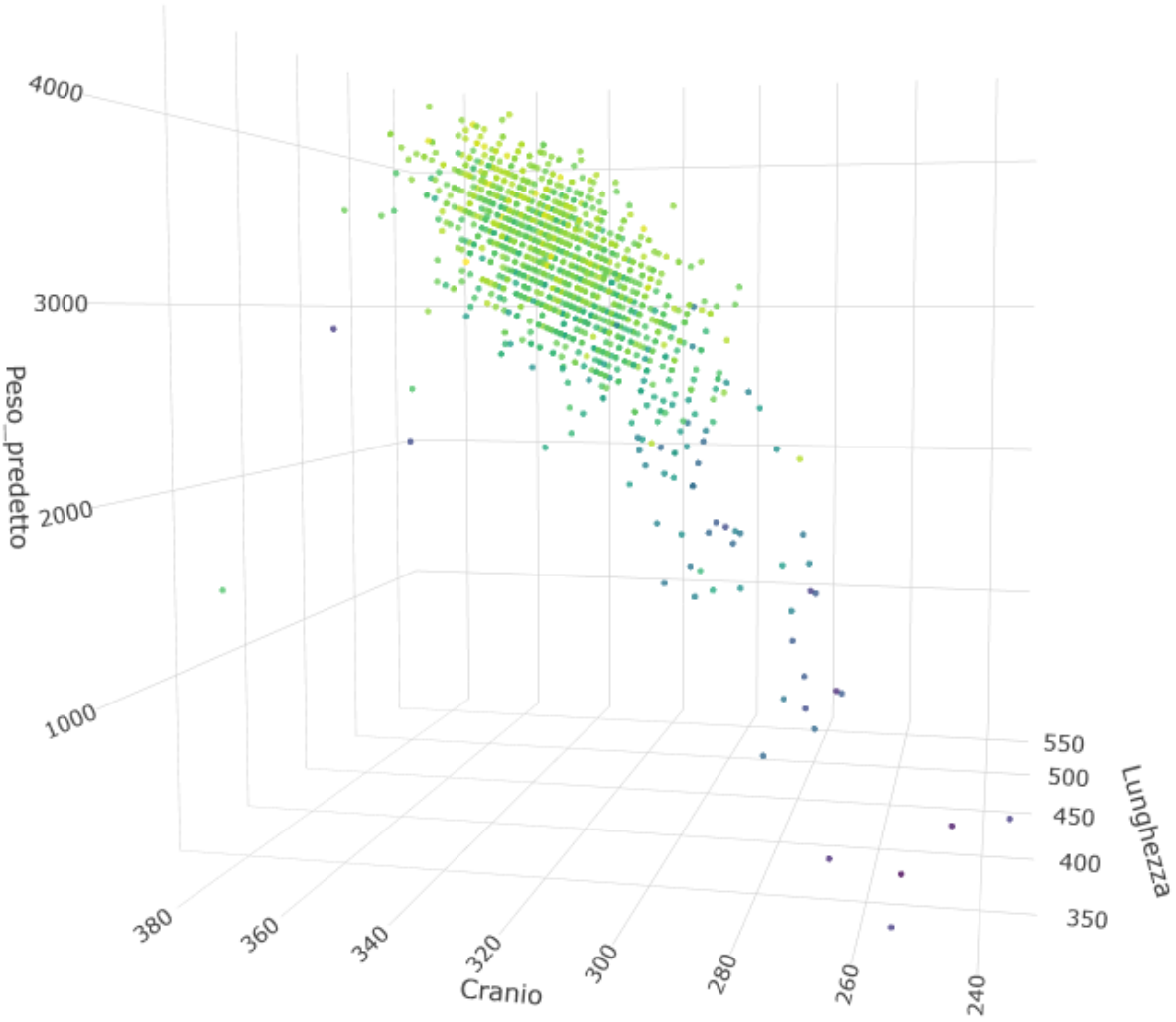
```
nuovi_neonatifumNO <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 0,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
)

previsionifumNO <- predict(mod4, newdata=nuovi_neonatifumNO)
```

In this case the model reports us an expected weight of 3272.765 grams.

To have a graphical representation of this model would be impossible, given the many variables used. Let us attempt to construct a model with three variables, representing them with a plot3D. To do this we will choose those with lower p-values.

```
mod_simpl <- lm(Peso ~ Lunghezza + Cranio, data = neonati)
summary(mod_simpl)
library(plotly)
df_plot <- neonati
df_plot$Peso_predetto <- predict(mod_simpl, newdata = neonati)

# Creare il grafico
plot_ly(data = df_plot, x = ~Lunghezza, y = ~Cranio, z = ~Peso_predetto,
                type = 'scatter3d', mode = 'markers',
                marker = list(size = 2, color = ~Gestazione, colorscale='Viridis', opacity = 0.8))
```

A stamp follows, but since it is a 3D plot I recommend viewing it via the R-file.


4000
3000
2000
1000
Peso_predetto
380
360
340
320
300
280
260
240
Cranio
550
500
450
400
350
Lunghezza

Forecast mod4plus

Previously we saw how two particular observations, 1551 and 310, were beyond Cook's 0.5 threshold. Let us try to build a new model excluding them.

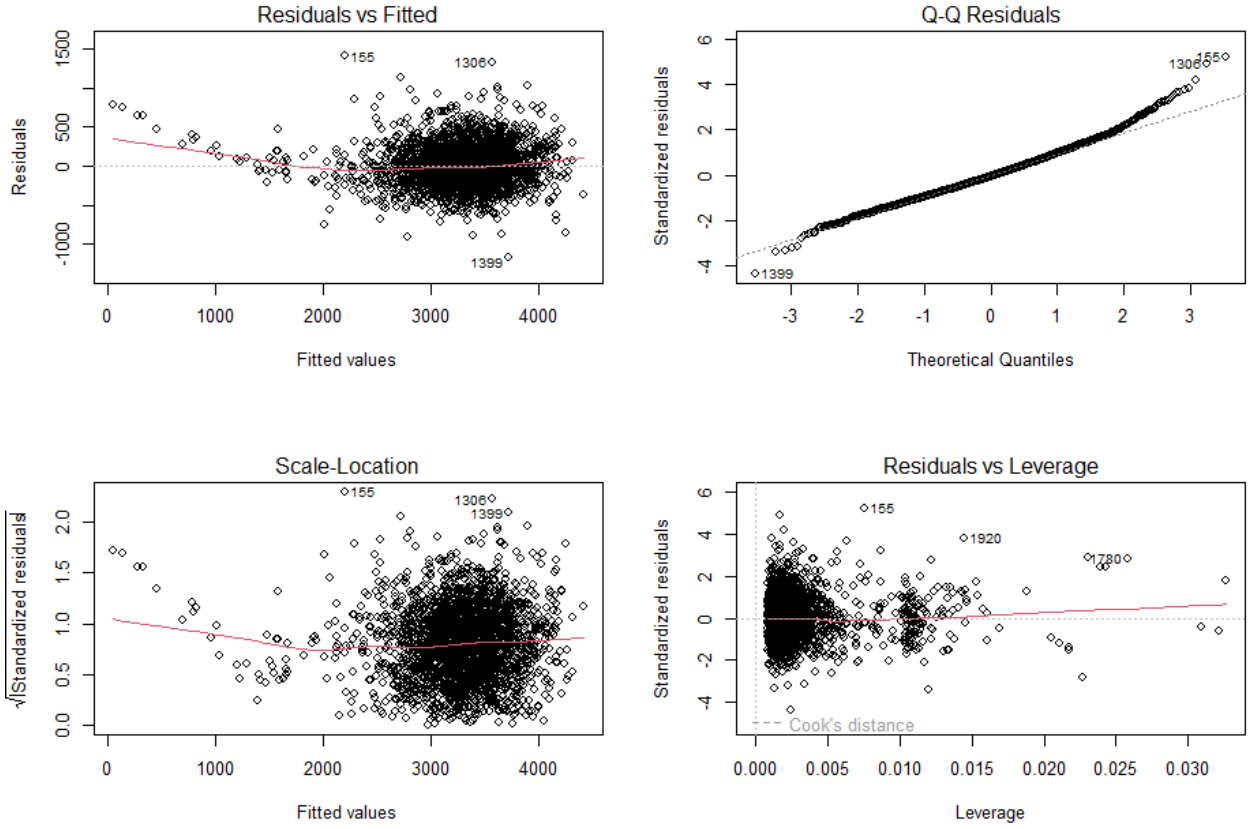
```
neonati2 <- neonati[-c(310, 1551), ]

mod4plus <- lm(formula = Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza +
                 Cranio + Sesso, data = neonati2)
summary(mod4plus)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1168.77  -179.86   -14.73   161.76  1402.96

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6662.9185   132.8843 -50.141  < 2e-16 ***
N.gravidanze   13.4937     4.2509   3.174  0.00152 **
Fumatrici     -27.2839    26.9948  -1.011  0.31225
Gestazione     28.2164     3.7585   7.507 8.36e-14 ***
Lunghezza      10.8386     0.3014  35.955  < 2e-16 ***
Cranio         10.0986     0.4245  23.789  < 2e-16 ***
SessoM         77.8064    10.9593   7.100 1.63e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 268.6 on 2491 degrees of freedom
Multiple R-squared:  0.7375,    Adjusted R-squared:  0.7369
F-statistic:  1167 on 6 and 2491 DF,  p-value: < 2.2e-16
```

Compared with the previous model, the indices of the residuals all change considerably, except for the third quartile. Some degrees of significance also change, but without affecting the model (we note an increase in Smokers, Gestation and SexM).

Adjusted R-squared benefits from a slight increase, from 0.7265 to 0.7369.

We continue with the analysis of the residuals of the new model.



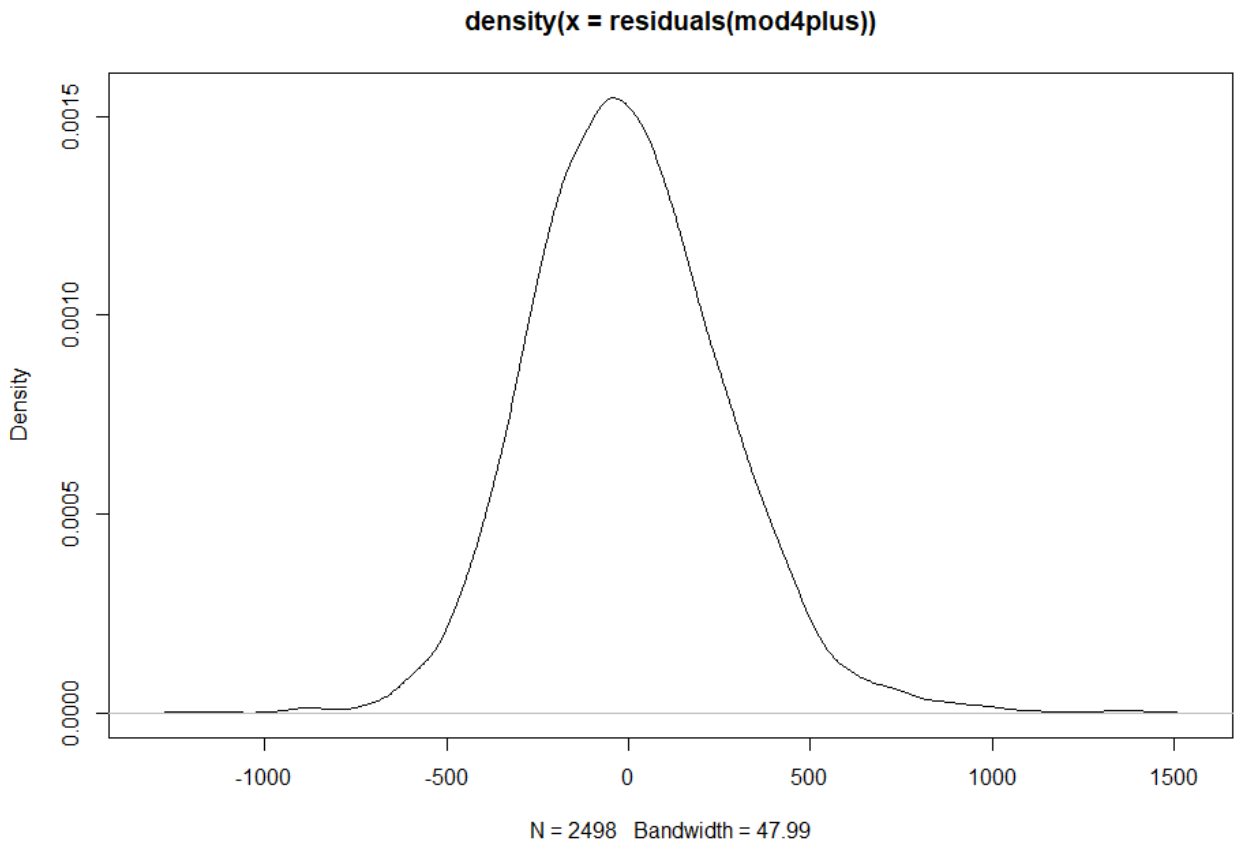
Heteroschedasticity is still present, but with this model we have no observations that cross Cook's distance thresholds.

```
        Shapiro-Wilk normality test

data:  residuals(mod4plus)
W = 0.98884, p-value = 4.641e-13

> lmtest::bptest(mod4plus)

        studentized Breusch-Pagan test

data:  mod4plus
BP = 6.3139, df = 6, p-value = 0.389

> lmtest::dwtest(mod4plus)

        Durbin-Watson test

data:  mod4plus
DW = 1.9584, p-value = 0.1493
alternative hypothesis: true autocorrelation is greater than 0
```

The Shapiro test returns data that are yes better, but that do not allow us to accept the null hypothesis. Therefore, we confirm that the residuals do not follow a normal distribution.

The Breusch-pagan test has varied considerably; we now have insufficient evidence to reject the null hypothesis of homoschedasticity. We can say that the variance of the residuals is now constant.

The Durbin-Watson test shows no notable variation, confirming the absence of autocorrelation among the model residuals.

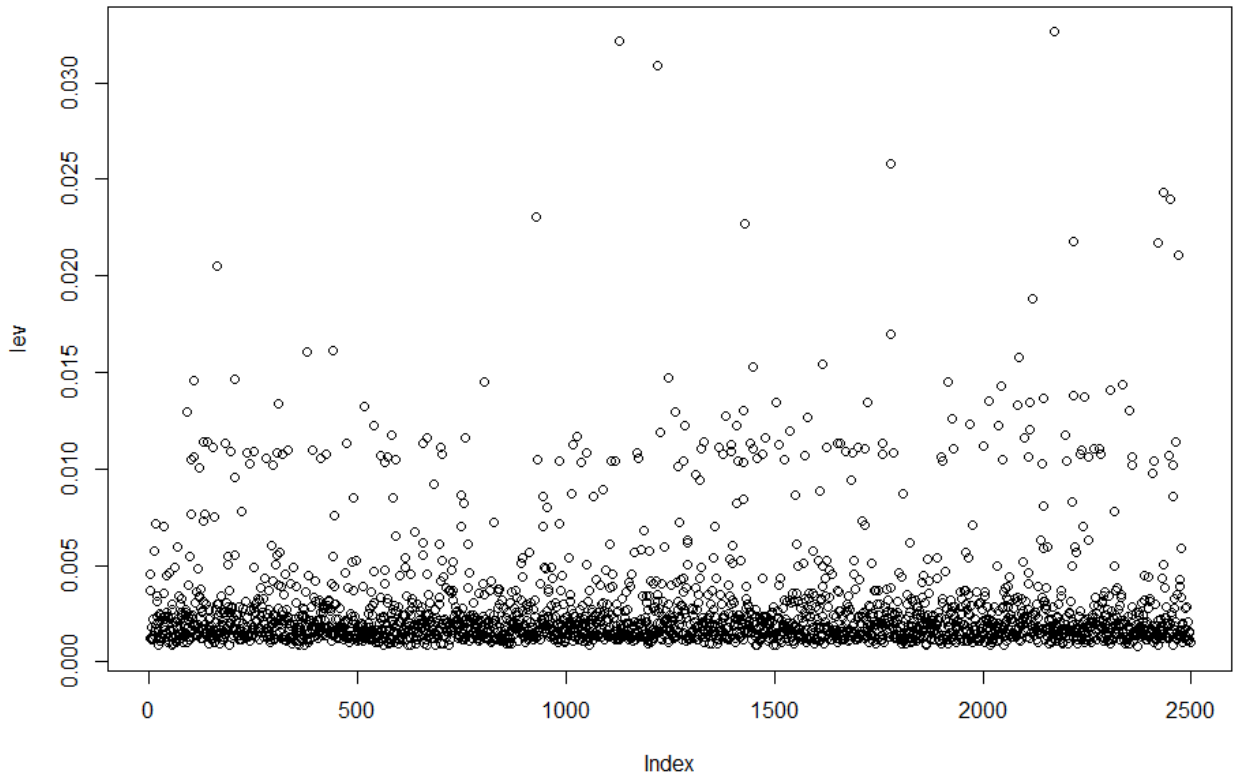


The density graph now shows a smaller tail on the right, given the absence of observation 1551. Other than that, no change in trend is shown, with long tails on the left and right of the graph.

```
#leverage
lev <- hatvalues(mod4plus)
plot(lev)
p <- sum(lev)
soglia=2*p/n
abline(h=soglia,col=2)
lev[lev>soglia]
```

Checking the clean model leverage, we can once again see its total absence. The threshold is not part of the visualization.

```
> lev[lev>soglia]
named numeric(0)
>
```



Regarding outliers, there are three, which have been observed previously.

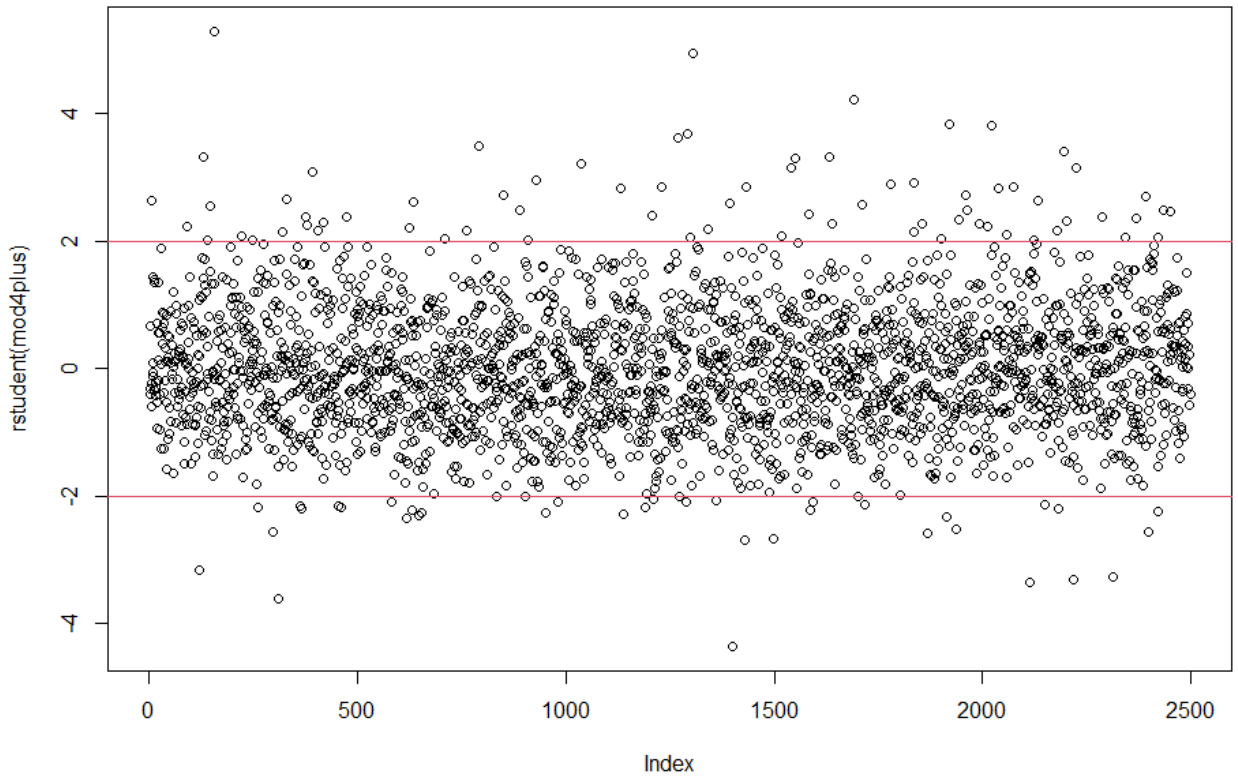
```
#outliers
plot(rstudent(mod4plus))
abline(h=c(-2,2), col=2)
car::outlierTest(mod4plus)
```

```
> car::outlierTest(mod4plus)
      rstudent unadjusted p-value Bonferroni p
155   5.282337         1.3859e-07   0.00034634
1306  4.938830         8.3797e-07   0.00209410
1399 -4.353267         1.3954e-05   0.03487100
```

Observation 155 has a high residual, with a rather low p-value and an equally low Bonferroni p-value, leading us to classify it as an outlier.

Observation 1306 also has a high residual and a low Bonferroni p-value, suggesting that it is an outlier.

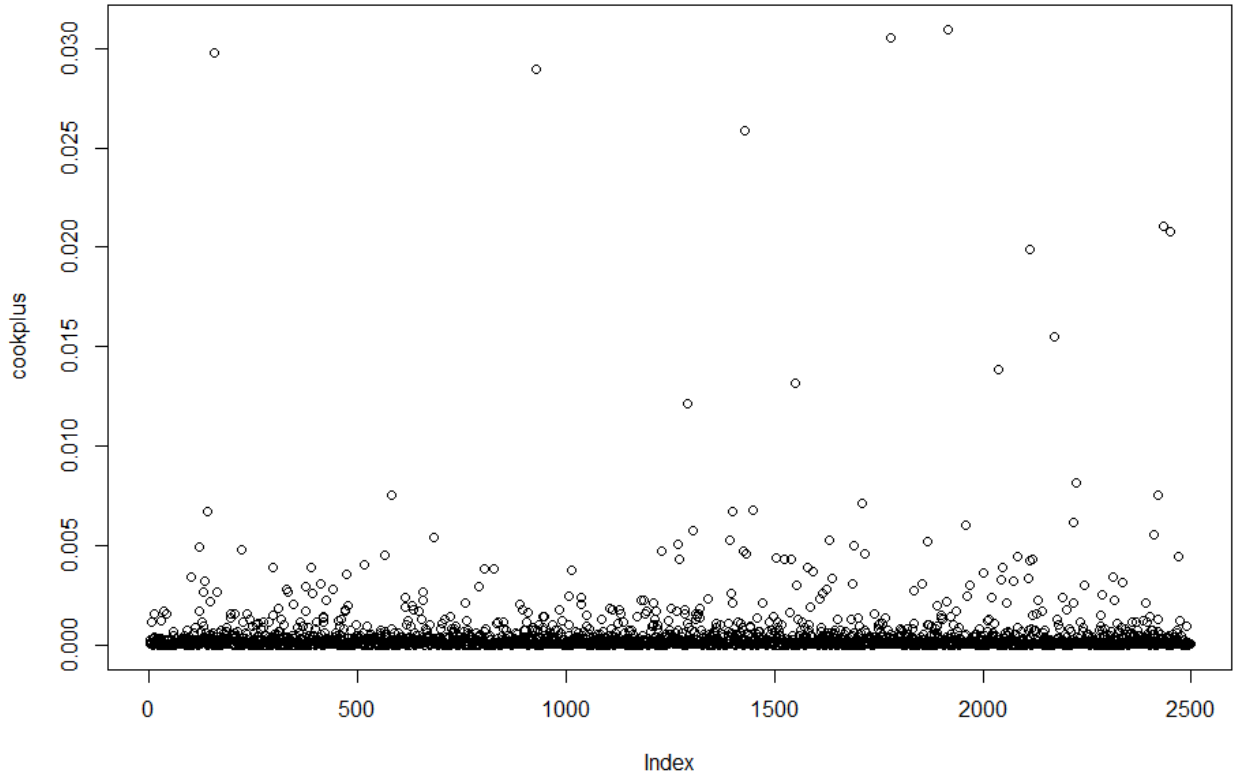
Observation 1399 has a negative residual, and a low Bonferroni p-value, higher than the other two observations but still below the threshold. Again we regard the observation as a significant outlier.



```
#distanza di cook
cookplus <-cooks.distance(mod4plus)
plot(cookplus)
max(cookplus)
```

Further check with cook distance, to see if there are any values that go beyond the threshold. We perform the check both with a plot and by calling up the maximum value, which in this case will be 0.03093833.

The plot confirms that there are no more values above the threshold



Now that the model is clean let us attempt a new prediction, again considering values for a female infant with a mother at 39 weeks gestation and in her third pregnancy, both smoker and nonsmoker.

```
plus_neonatifumSI <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 1,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
)

plus_previsionifumSI <- predict(mod4plus, newdata=plus_neonatifumSI)
```

The predicted weight for a newborn born to a smoking mother is 3246.302 grams, as in the previous model.

```
plus_previsionifumSI <- predict(mod4plus, newdata=plus_neonatifumSI)

plus_neonatifumNO <- data.frame(
  N.gravidanze = 3,
  Gestazione = 39,
  Fumatrici = 0,
  Sesso = "F",
  Lunghezza = mean(Lunghezza),
  Cranio = mean(Cranio)
  
)

plus_previsionifumNO <- predict(mod4plus, newdata=plus_neonatifumNO)
```

The predicted weight for a newborn born to a nonsmoking mother is 3273.586 grams.

We now proceed to construct a more simplified model that can be represented with a plot3D.

```
mod_simpl_plus <- lm(Peso ~ Lunghezza + Cranio, data = neonati2)
summary(mod_simpl_plus)
library(plotly)
df_plot2 <- neonati2
df_plot2$Peso_predetto <- predict(mod_simpl_plus, newdata = neonati2)

# Creare il grafico
plot_ly(data = df_plot2, x = ~Lunghezza, y = ~Cranio, z = ~Peso_predetto,
        type = 'scatter3d', mode = 'markers',
        marker = list(size = 2, color = ~Gestazione, colorscale='Viridis', opacity = 0.8))
```

Again I invite you to view the plot from the R-file, here I will attach a printout for a first view.