Statistical Model To Predict The Weight Of Newborns

Enrico Michelon

Contents

Introduction (Point 2)

This project concerns the creation of a statistical model to predict the weight of newborns. Our objective is to create a statistical model given the *neonati.csv* dataset that can be extended to the entire population. In particular we want to find out whether it is possible to predict the infant's weight at birth given all the other variables. In particular, one wants to study a relationship with the mother's variables, to find out whether or not these have a not have a significant effect, such as the potentially harmful effect of smoking. The length and diameter of the infant's skull are also used because they can be estimated already from ultrasound scans, but in general they could also serve as control variables.

Table 1: Dataset first rows

Anni.madre	N.gravidanze	Fumatrici	Gestazione	Peso	Lunghezza	Cranio	Tipo.parto	Ospedale	Sesso
26	0	0	42	3380	490	325	Nat	osp3	M

Anni.madre	N.gravidanze	Fumatrici	Gestazione	Peso	Lunghezza	Cranio	Tipo.parto	Ospedale	Sesso
21	2	0	39	3150	490	345	Nat	osp1	F
34	3	0	38	3640	500	375	Nat	osp2	M
28	1	0	41	3690	515	365	Nat	osp2	M
20	0	0	38	3700	480	335	Nat	osp3	F
32	0	0	40	3200	495	340	Nat	osp2	F

Dataset (Points 1, 2, 3)

The dataset is composed by 2500 samples and, studying its first rows, we can distinguish 10 variables: Anni.madre, N.gravidanze, Fumatrici, Gestazione, Peso, Lunghezza, Cranio, Tipo.parto, Ospedale and Sesso.

Anni.madre

Anni.madre is a quantitative variable on radio scale. In the dataset we have at least two outlayers, which can be found at rows 1152 and 1380, and report an age of 1 and 0, respectively. Computing position measures and standard deviation excluding those rows, we obtain:

Table 2: Position measures and standard deviation for Anni.madre

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
13	25	28	28.19	32	46	5.22

N.gravidanze

N.gravidanze is a quantitative variable on ratio scale, which represents the number of pregnancies per mother. In Table 3 position measures and standard deviation for the variable are shown. We can see that mean and standard deviation and third interquartile are around 1 (0.98, 1.28 and 1 respectively), while the maximum reaches a value of 12.

Table 3: Position measures and standard deviation for Anni.madre

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
0	0	1	0.98	1	12	1.28

We can look now at the distribution of *N.gravidanze*. From Figure 1 we can notice that it is a normal positive skewed distribution, with mean 0.98 and standard deviation of 1.28.

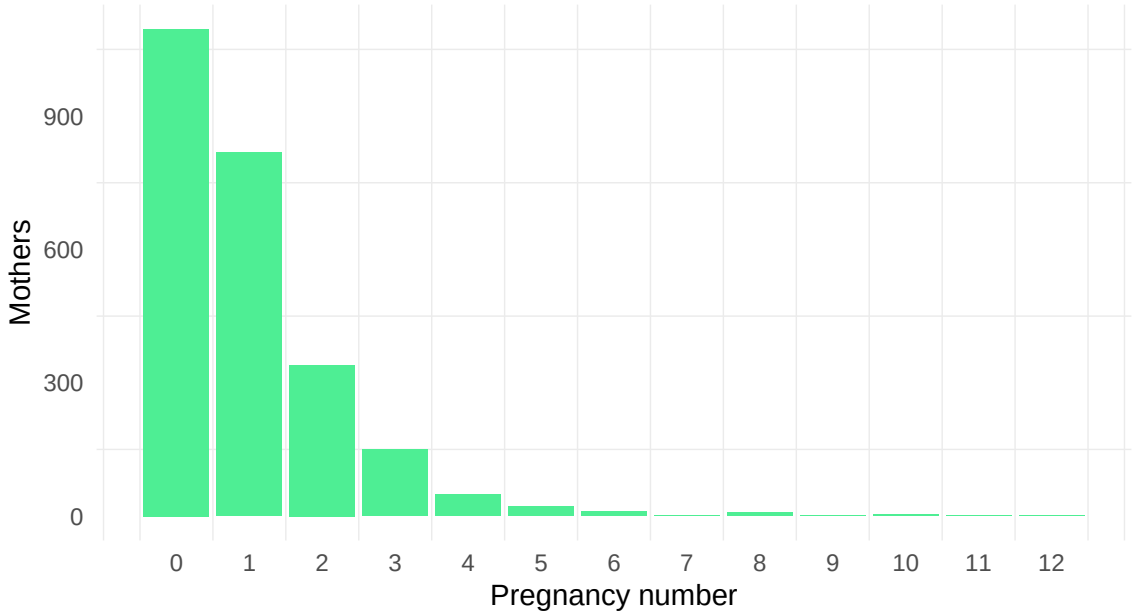


Figure 1: N.gravidanze distribution

Fumatrici

Fumatrici is a qualitative encoded variable on nominal scale, with values 0 and 1. Value 0 means that the mother is not a smoker, while mothers with "Fumatrici" value of 1 means she is a smoker.

Computing the Gini coefficient for the variable, we obtain a value of 0.16, which means that the ditribution is not much equally distributed. In Figure 2, the distribution of variable "Fumatrici" is shown.

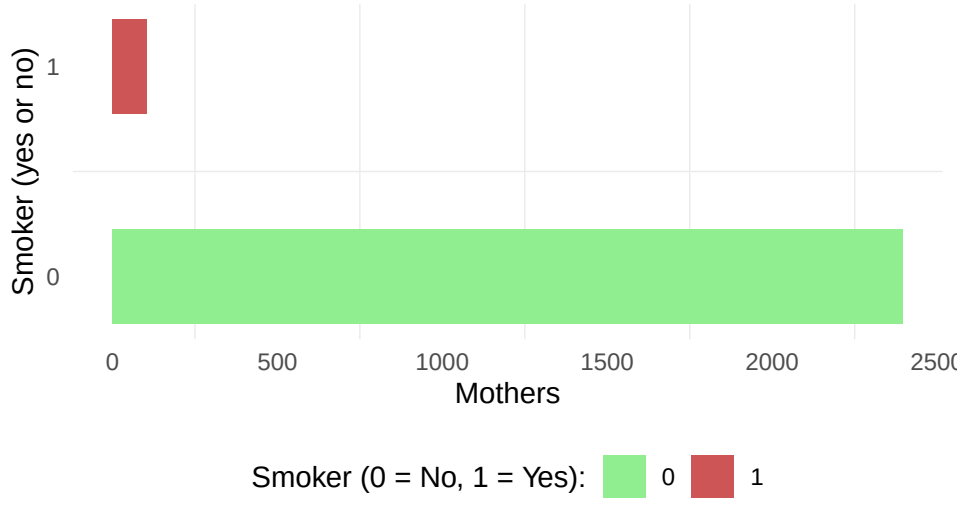


Figure 2: Smoker vs Not Smoker

Gestazione

Gestazione is a quantitative variable on ratio scale, measured in weeks, with position and standard deviation measures that can be seen in Table 5. As we expect, the mean value is around 40 weeks (38.98), with a low

standard deviation of 1.87 weeks.

Table 4: Position measures and standard deviation for Gestazione

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
25	38	39	38.98	40	43	1.87

Peso

Peso is a quantitative variable on ratio scale, and represents the weight of newborns in grams. Position measures and standard deviation are observable in Table 6.

Table 5: Position measures and standard deviation for Peso

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
830	2990	3300	3284.08	3620	4930	525.04

It is interesting to note that "Peso" has in IQR value of 630 grams, while the range is of 4100. This can been explained studying the graphic on Figure 3. The distribution is negatively skewed, with very long tails, specially on the left, making a large different between IQR (represented by red line on the graphic) and range. The distribution is not a Normal distribution, as we will see later observing the Q-Q Plot.

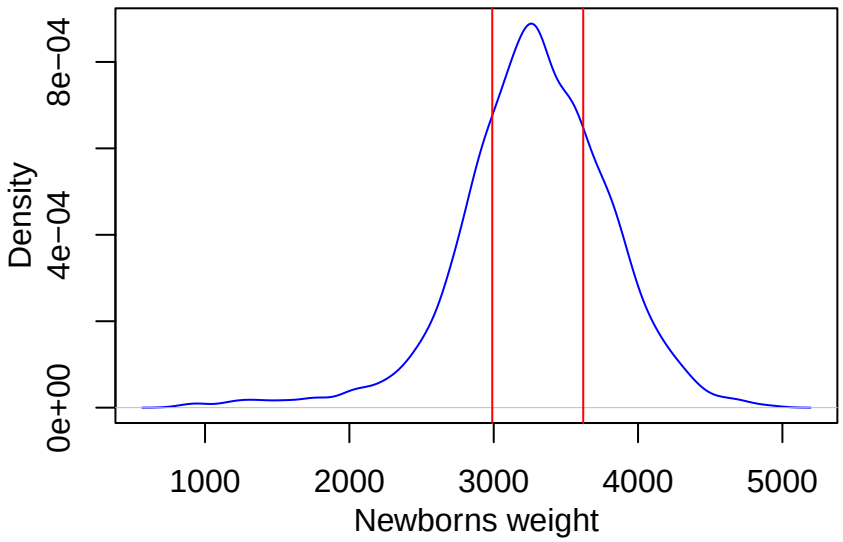


Figure 3: Peso distribution

Lunghezza

Lunghezza is a quantitative variable on ratio scale. Represents the length of the newborn in millimeters. For this variable we expect a behaviour quite similar to variable "Peso".

Table 6: Position measures and standard deviation for Lunghezza

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
310	480	500	494.69	510	565	26.32

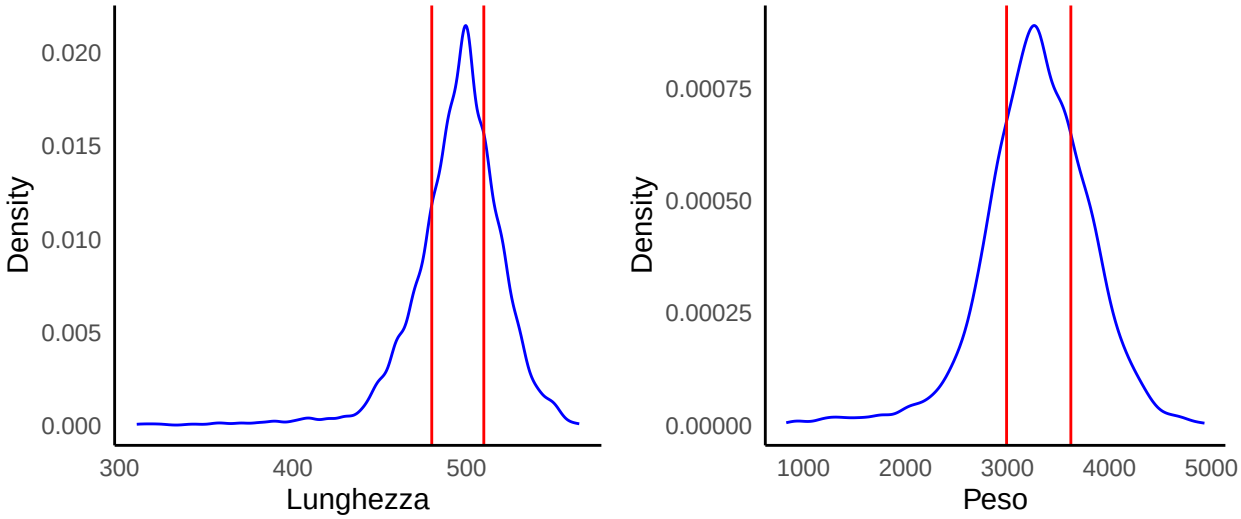


Figure 4: Lunghezza distribution vs. Peso distribution

Cranio

Cranio is a quantitative variable on ratio scale. Represents the cranium diameter of the newborn in millimeters. In Table 8, we can observe position measures and standard deviation for variable "Cranio".

Table 7: Position measures and standard deviation for Cranio

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std.dev
235	330	340	340.03	350	390	16.43

Tipo.parto

Tipo.parto is a qualitative variable on nominal scale. Represents the type of childbirth, and is encoded with "Nat" (natural) and "Ces" (caesarean). For a qualitative variable we can compute Gini coefficient, which provides a measure of heterogeneity. Gini coefficient for "Tipo.parto" is 0.83, indicating a very high heterogeneity.

Ospedale

Ospedale is a qualitative variable on nominal scale. It represents three different hospitals, and is coded in "osp 1", "osp 2" and "osp 3". The Gini coefficient is very close to 1 which means that the variable is almost heterogeneous.

Sesso

Sesso is a qualitative variable on nominal scale. Represents the gender of the newborn, and is coded with "M" for males and "F" for females. The Gini coefficient is very close to 1 which means that the variable is almost heterogeneous. In Figure 5 we can see the distribution of "Sesso".

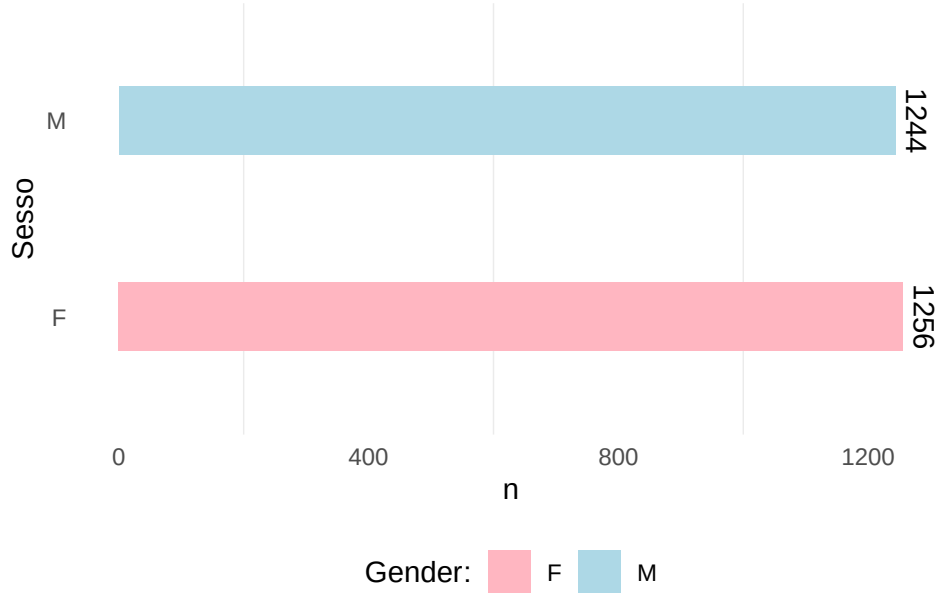


Figure 5: Gender distribution

Descriptive statistics (Points 3, 5)

We want to introduce now a short descriptive analysis of the dataset. For our porpuses, we divide frequency distribution of Anni.madre into classes, and we show the distribution in Figure 6.

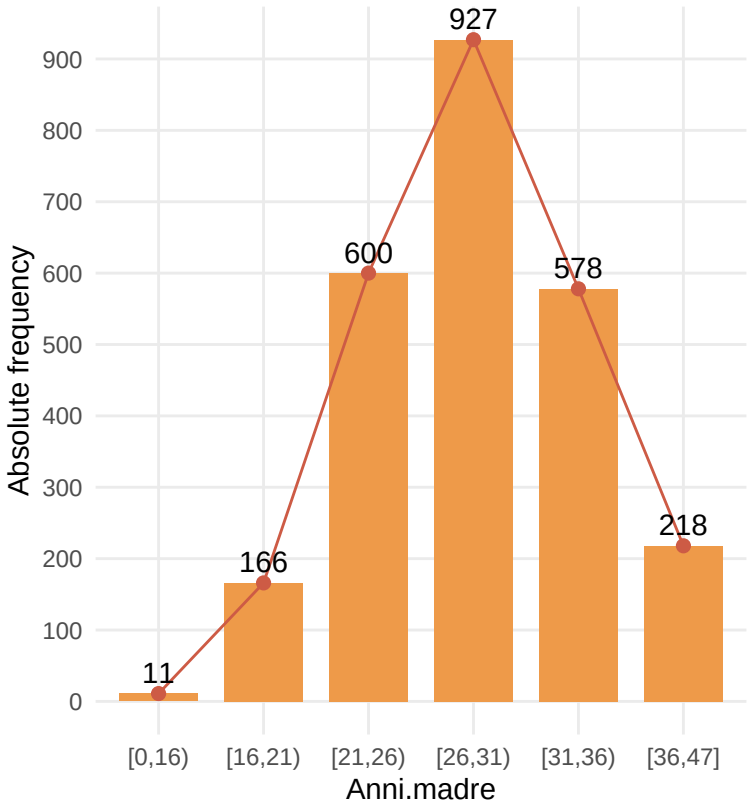


Figure 6: Frequency distribution of "Anni.madre" class

Once obtained the distribution in classes of Anni.madre, we are interested in observing the trend of number of pregnancies per class of ages.

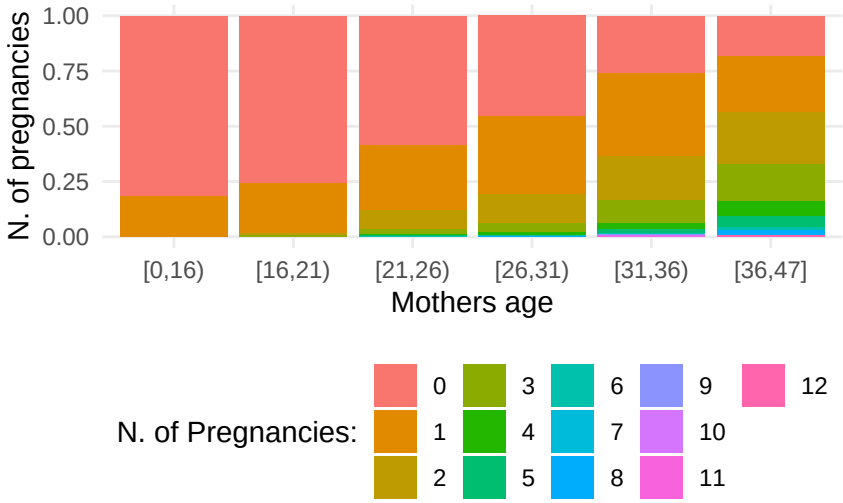


Figure 7: N. of pregnancies per class of ages

As we can see in Figure 7 as we could expect, the number of pregnancies has an increasing trend. Moreover, in the classes of ages between 31 and 35, there is still more than 25% of mothers, which are waiting the first baby. Another interesting result is obtained observing the class of age 0-15, where a consistent percentage of

mothers are waiting for the second baby. However, the latter statistic is somewhat misleading, as the sample of mothers between the ages of 0 and 15 is obviously very small. Computation of this statistic in general does not give us particular information. Therefore, we can proceed by observing the type of childbirth, subordinate to the age and the fact that mother smokes or not.

From Figure 8, we can note that until the range 26-30, there seems to be no impact for smoker mothers with respect to non smoker mothers. On the other hand, as age passing, we note that the number of caesarean births decreases if the mother is a smoker, which is really strange. However we wanted to show you this statistic, to remember that great care must be taken when analysing data. In fact, this statistic is highly altered by the fact that the number of smoking mothers, as shown in Figure 2, is only a small percentage compared to the number of non-smoking mothers.

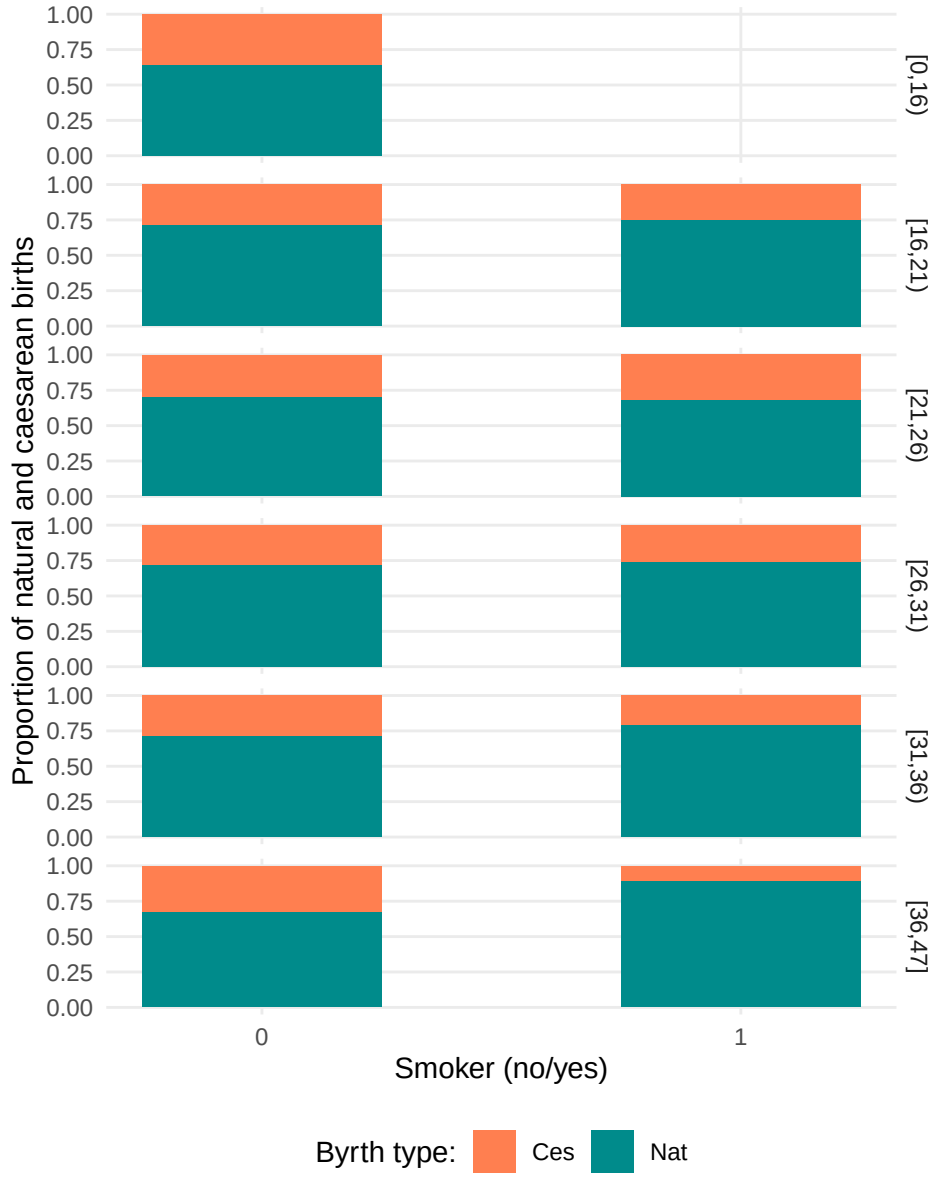


Figure 8: Birth type per class of ages, subordinates to smoker

We want now to analyse the skull circumference with respect to pregnancy and class of ages of mothers. From Figure 9, we can observe that the class of ages does not significantly affect the skull circumference,

whereas, as expected, weeks of gestation significantly affect skull size. Later on, we will analyse how and whether gender affects the size of the fetus' skull.

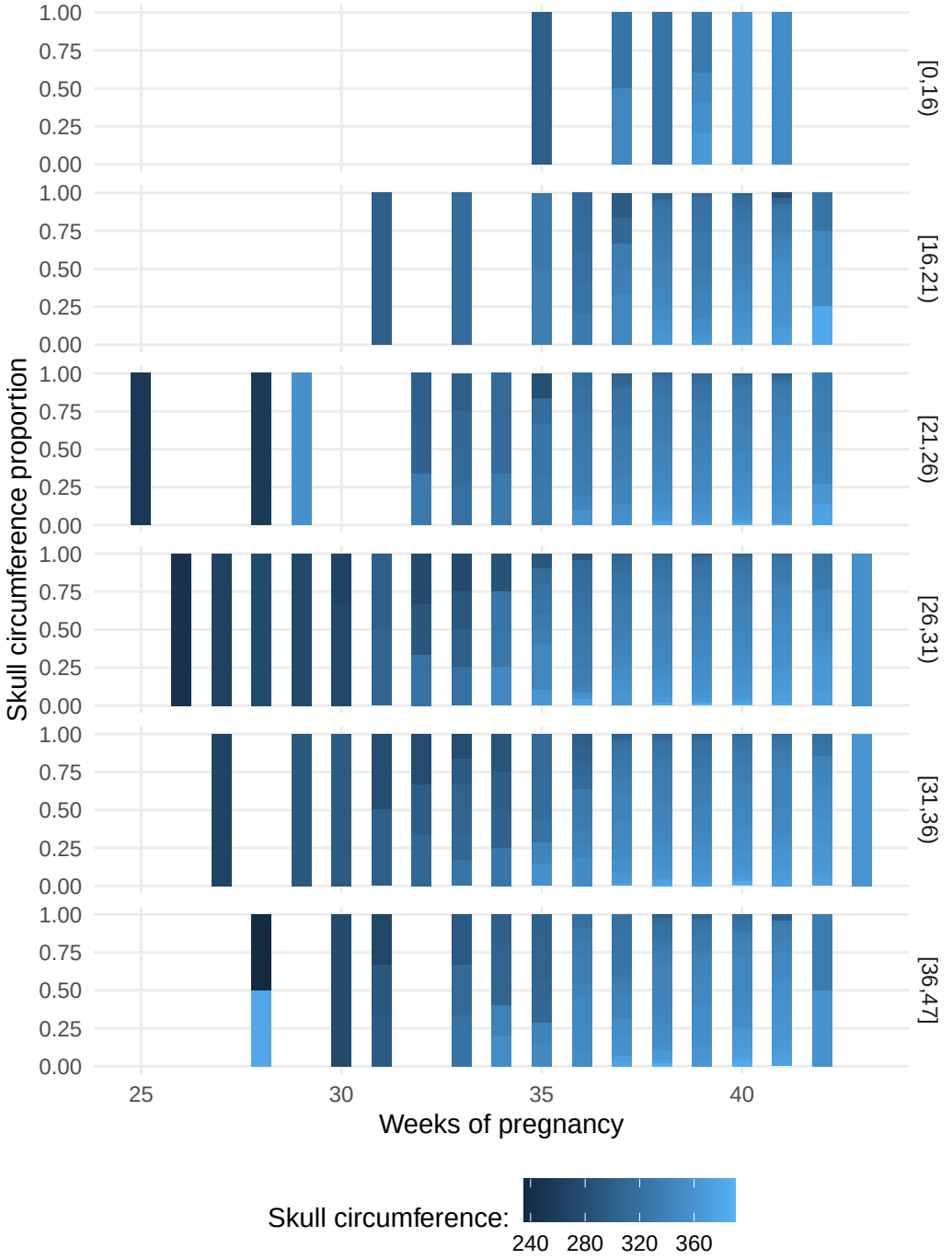


Figure 9: Skull circumference, subordinates to weeks of pregnancy and class of ages

In Figure 10, we can observe the different behaviour of weight based on the gender of the newborns. Since the statistic contained a lot of samples, we have only shown gestations in the interquartile range in the graph. Later on, we will analyse the full dataset, through a boxplot.

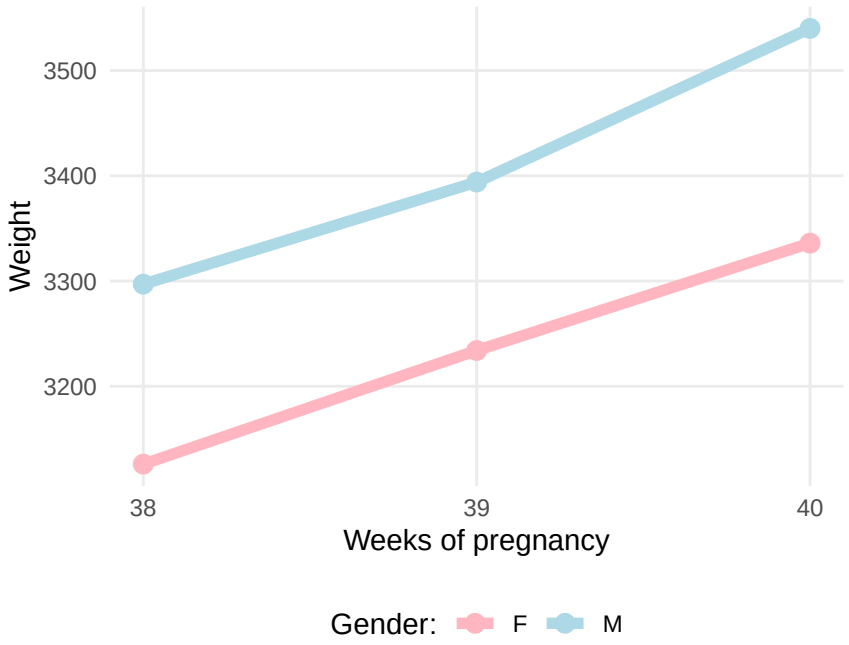


Figure 10: Weight of newborns per gender and weeks of pregnancy

In Figure 11 instead, we can observe the different behaviour of length of newborns, considering a caesarean birth or natural birth, depending on the weeks of pregnancy. This statistic has been computed with the idea to check whether the caeserean birth, has been induced by the fetus dimensions, for example in case of fetal macrosomia. Clearly, this statistic no longer gives us the reason for the choice of type of birth, however, it is interesting to note that in most cases, the length of the fetus born by caesarean birth was greater than or equal to the length of the fetus born by natural childbirth.

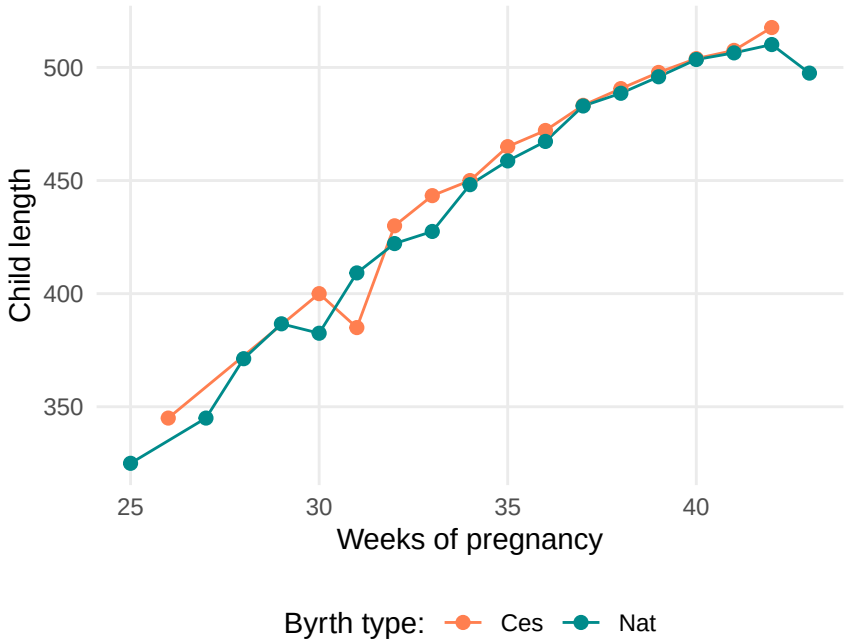


Figure 11: Length of newborns, based on birth−type and gestation

Finally in Figure 12, we can observe the differences of weight, length, skull circumference and type of childbirth, between genders. It is interesting to note the total independence between the sex of the fetus and the type of birth. On the other hand, there is a difference between gender observations concerning physical characteristics. The most significant difference can be noticed in weight of newborns.

With the support of Figure 12 and Tables 8 and 9, we can see that IQR ranges for variable "Peso", are equal, while the total range is much wider for females, meaning that outlayers are move further away from the average value. Moreover this means that also the standard deviation is higher for females with respect to males.

Table 8: Statistics for Peso (males)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
980	3150	3430	3408.22	3720	4810	493.8	570	-0.76

Table 9: Statistics for Peso (females)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
830	2900	3160	3161.13	3470	4930	526.31	570	-0.58

Differently in Tables 10 and 11, we can see that IQR ranges, standard deviation, mean and median for variable "Lunghezza", are quite similar. However, the total range is still a bit wider in females.

Table 10: Statistics for Lunghezza (males)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
320	490	500	499.67	515	560	24.04	25	-1.37

Table 11: Statistics for Lunghezza (females)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
310	480	490	489.76	505	565	27.53	25	-1.6

Finally for variable "Cranio" (Tables 12 and 13, Figure 12c), shows results comparable to the variable "Lunghezza".

Table 12: Statistics for Cranio (males)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
265	334	343	342.45	352	390	15.74	18	-0.66

Table 13: Statistics for Cranio (females)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	std-dev	IQR	skewness
235	330	340	337.63	348.25	390	16.74	18.25	-0.88

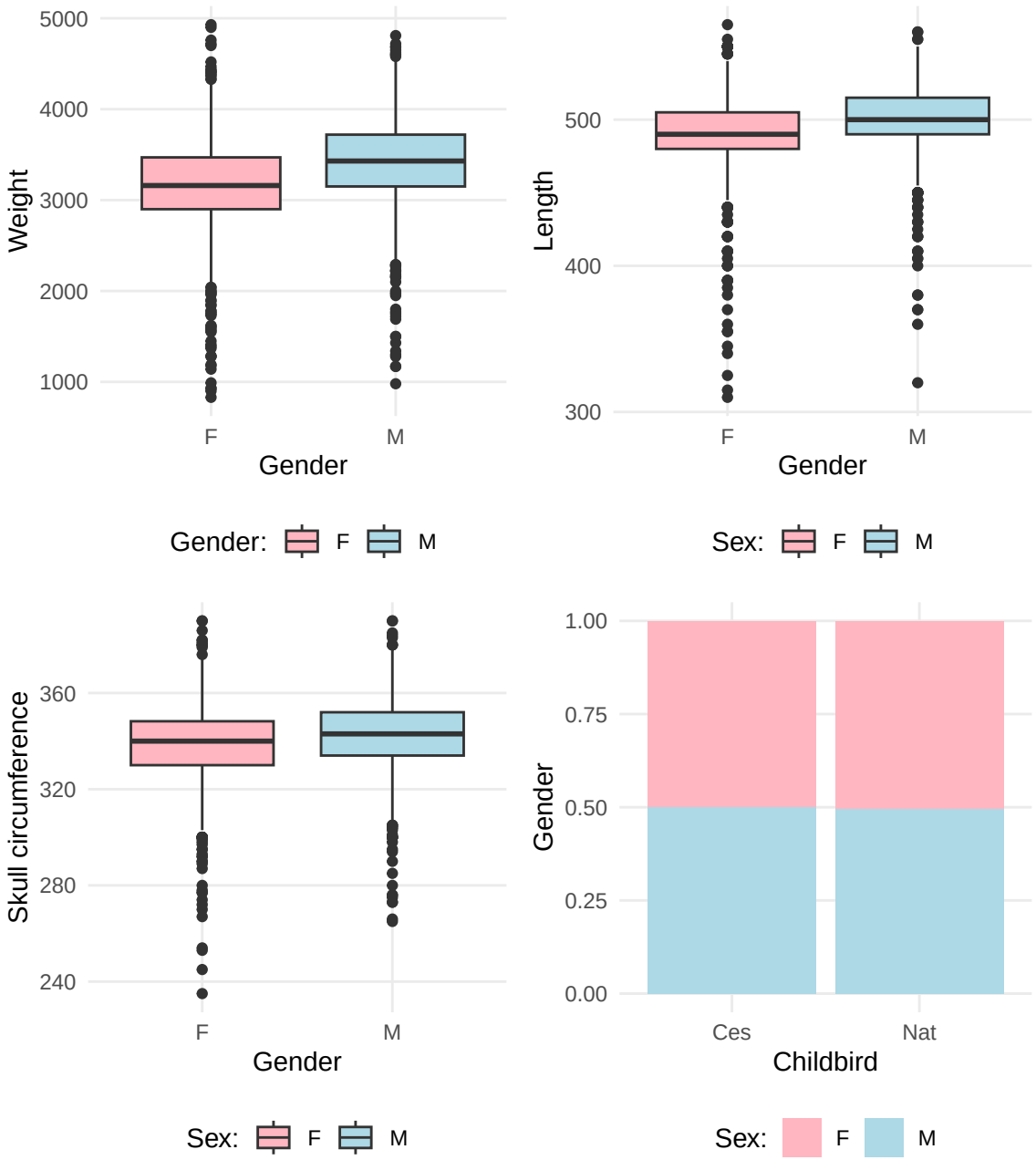


Figure 12: Gender comparisons for weight, length, skull circumference and type of childbirth

Hypothesis testing (Points 4 and 6)

We now want to understand whether the sample of infants extracted correctly represents the population. In particular we want to test the hypothesis that the mean of weight and length of newborns is the same as the population. From *ospedalebambinogesu.it* we got that mean of weight and length of population of fetus is 3300 grams and 50 cm respectively. As we said on paragraph Peso, distribution of "Peso" and "Lunghezza" are not normal. In Figure 13 we can see the Q-Q Plots for the two variables, which show a strong deviation from normality.

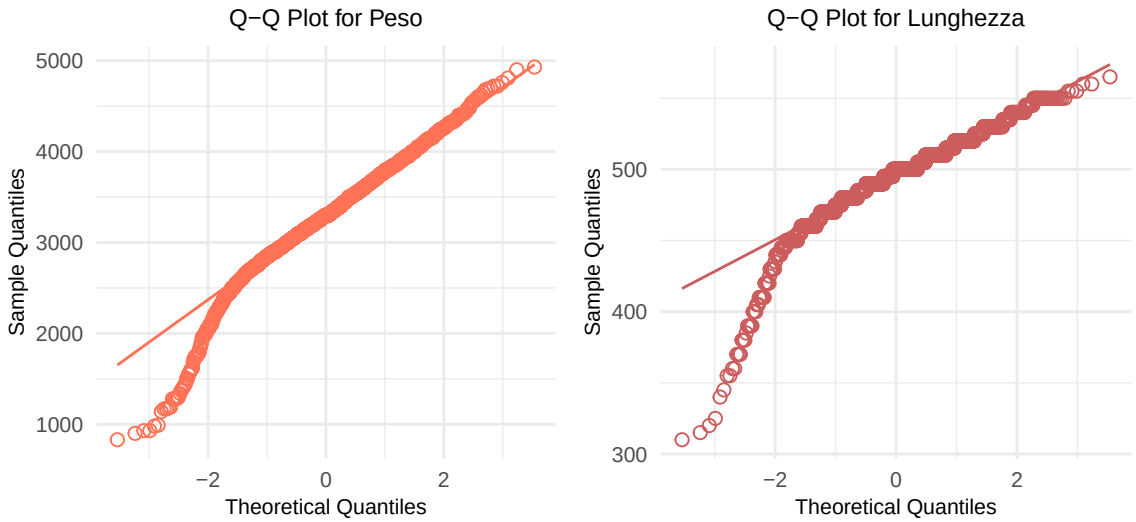


Figure 13: Q−Q Plot for Peso and Lunghezza

We decided to compute Student T-test, even if the Shapiro Wilk test for Normality, shows that we have to reject the null hyphotesis, which means that the two distributions are not normal. With T-test we also try the one sample Wilcoxon rank sum, to test the hypothesis reported before. With both test we have the same result, which is to not reject the null hypothesis for variable Peso, while to reject the null hypothesis for variable Lunghezza.

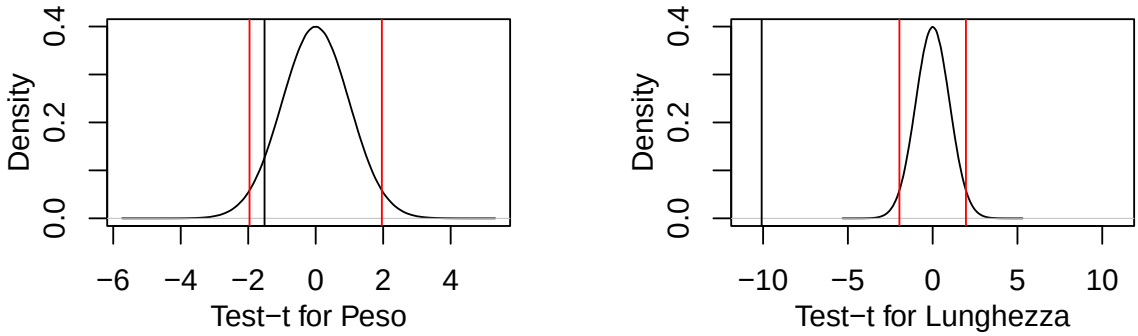


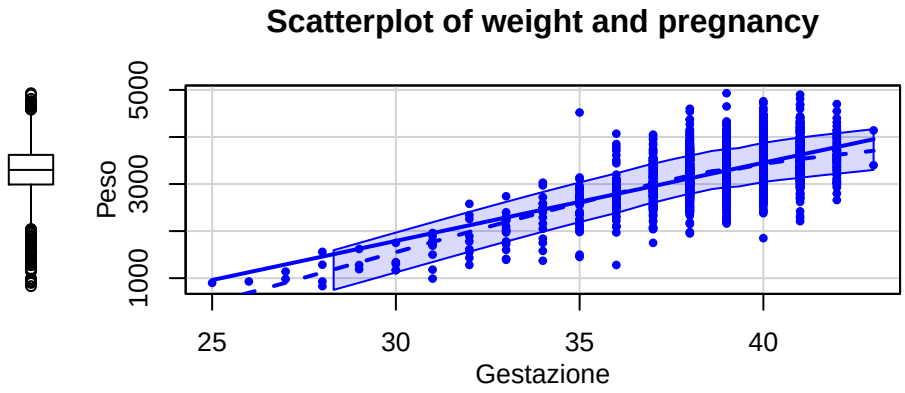
Figure 14: Decision rule for Peso and Lunghezza

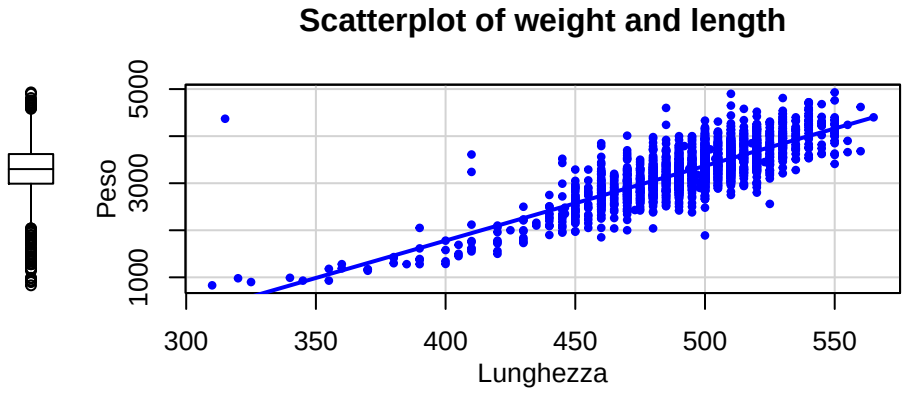
On Figure 14, we can see the decision rule for Peso and Lunghezza. In case of Peso we are inside the acceptation zone while in case of Lunghezza we are inside the rejecting zone. Moreover for variable Peso we have a P-value that is 0.1296452 which is higher than 0.05, using a Student T-test, while it is 0.9611712 using Wilcoxon. For variable Lunghezza instead, the corresponding P-value are $1.8141222 \times 10^{-23}$ and $1.2260292 \times 10^{-16}$. Even if mean value of lenght of our dataset differs from the global mean value by just 4 millimeters, the null hypothesis has been rejected. This could be caused by the low standard deviation (and consequently variance), which cause the sample to be distorted but precise, and the proof is given by another Wilcoxon test for Lunghezza, considering only measures of Lunghezza inside the interquartile range, which provide the same result.

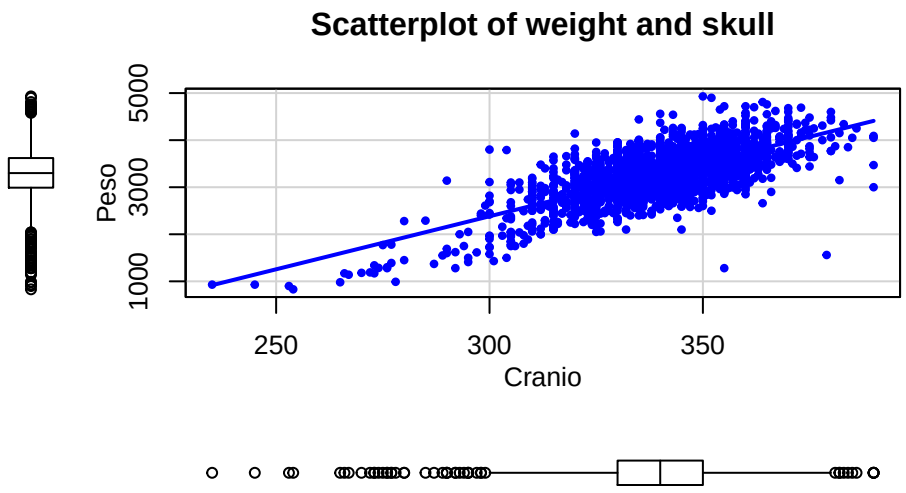
Another point we want to test is **if the percentage of caesarean section is the same for all the hospitals, or there is some hospitals in which this percentage is higher**. To do this, we use a chi-square test, which is a test for indipendence between two categorical variables. The null hypothesis is that the two variables are indipendent, while the alternative hypothesis is that they are not. The result of this test is that we can not reject the null hypotesis, obtaining a p-value of 0.58. The other test we made was a pairwise Wilcoxon test, which is a non-parametric test for multiple comparisons. The null hypothesis is that the three hospitals have the same percentage of caesarean section, while the alternative hypothesis is that they are not. The result of this test is that we can not reject the null hypotesis, obtaining a p-value of 1 comparing hospitals. This means that the percentage of caesarean section is the same for all the hospitals, and there is no hospital in which this percentage is higher than the others.

Part 2: Generalized Linear Models (Point 1)

In this part we will use the dataset `newborns_dataset` to want to build a model to predict the weight of a newborn, using the other variables of the dataset. The first thing we want to do is to investigate the relationship between the weight of a newborn and the other variables. We can investigate this relashionship both numerically, by using covariance and coefficient of linear correlation and graphically by means of scatterplots.







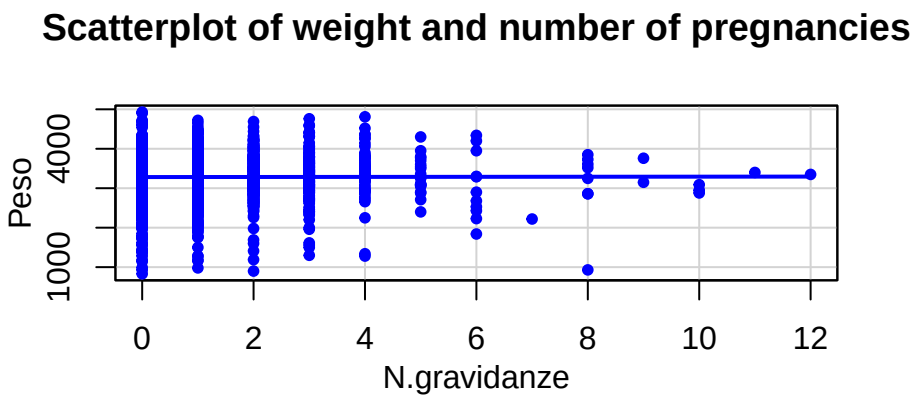


Figure 15: Scatterplots for weight, and other variables

In Figure 15 we can see the scatterplots of the weight of a newborn and the other variables. The scatterplot of the weight and pregnancy length shows a linear relationship, with coefficient of linear correlation of 0.59. The scatterplot of the weight and length of the newborn shows a linear relationship, with a coefficient of linear correlation of 0.8. Similar result we can see in the scatterplot of the weight and the skull of the newborn, where the linear correlation coefficient is 0.7. On the other hand, the scatterplot of the weight and the number of pregnancies shows an absence of relationship, with linear correlation coefficient of 0.

A final scatterplot matrix to resume the correlation between “Peso” and other variables of the dataset can be observed in Figure 16.

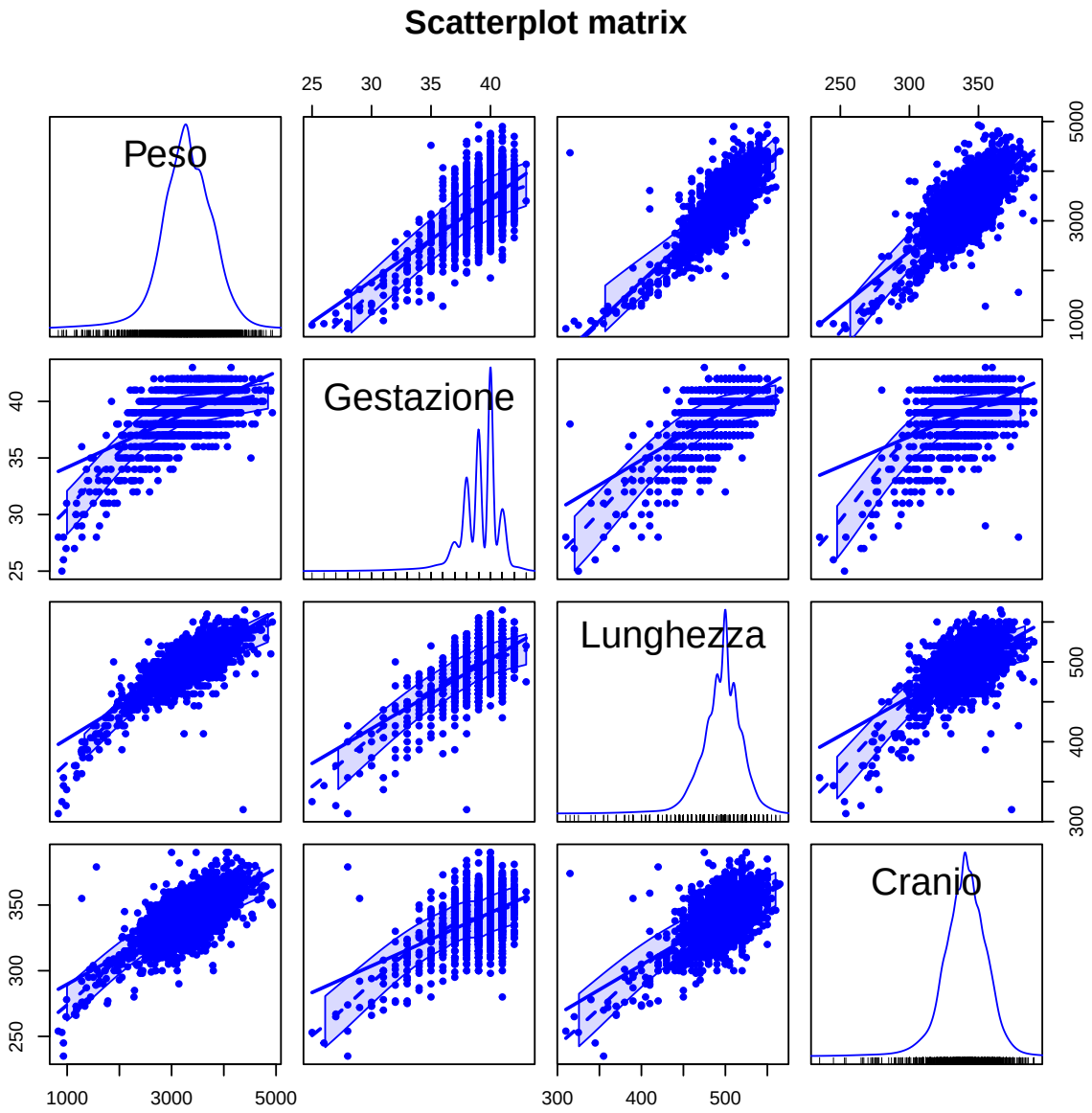


Figure 16: Scatterplot matrix for weight, and variables

These correlations and scatterplots computed, show relationships between “Peso” and other quantitative variables of the dataset. In order to investigate the relationship between “Peso” and qualitative variables, we can use `pairs()` function. Adjusting the parameter of the function, we can obtain the graphic shown in Figure 17. The figure shows a resume of the results obtained above, and the relashionship with variables “Sesso”, “Fumatrici” and “Ospedale”.

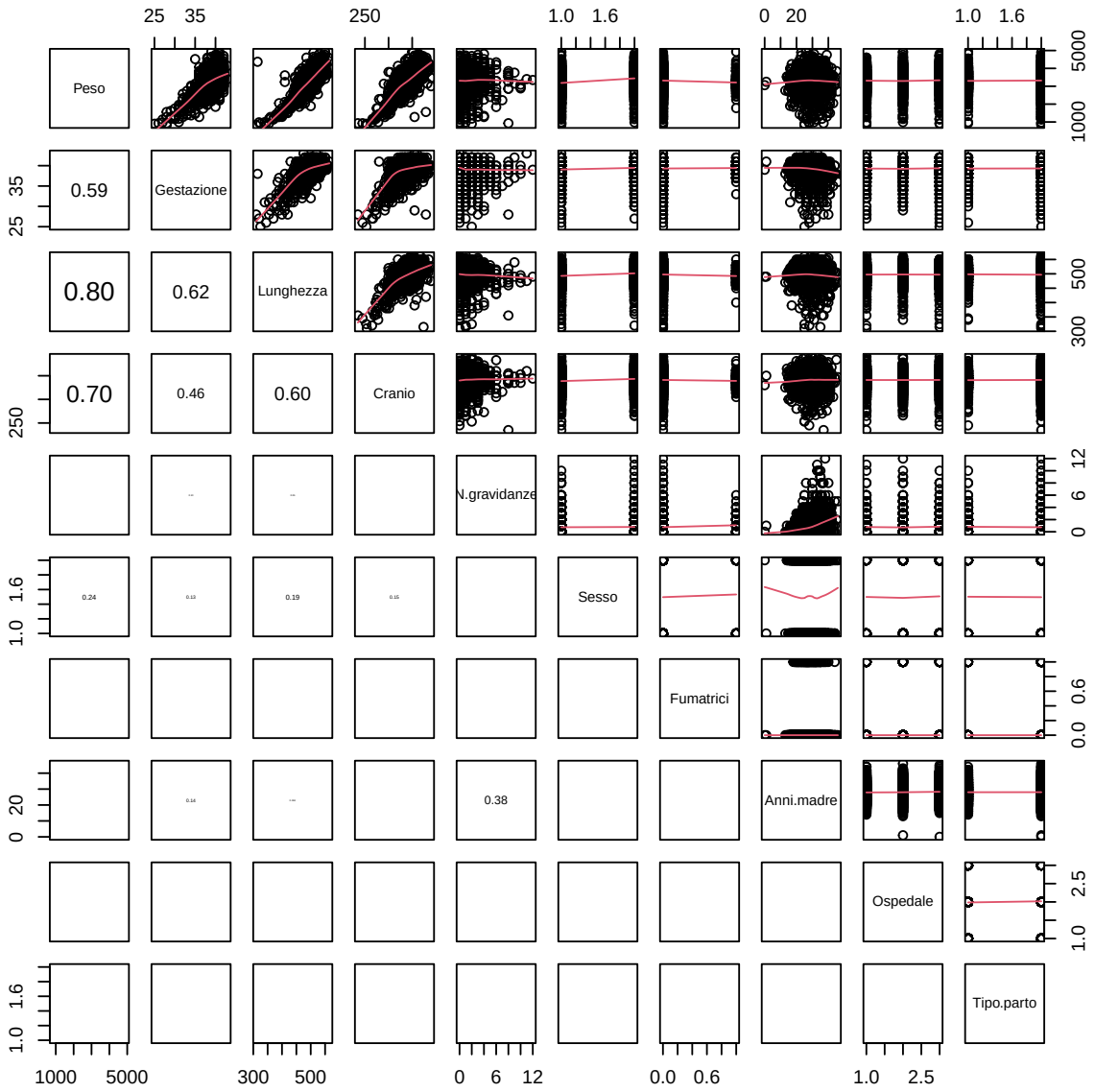


Figure 17: Correlation matrix for weight, and variables

We can see that none of the variables have a really linear relashionship with "Peso", and this could represent a problem for the linear regression model, while the variables with the highest correlation are "Cranio" and "Lunghezza", as could be expected.

Multiple linear regression model (Point 2, 3)

We want to build a multiple linear regression model to predict the weight of a newborn. The model we want to build is in

$$Peso = \beta_0 + \beta_1 \cdot Gestazione + \beta_2 \cdot Lunghezza + \beta_3 \cdot Cranio + \beta_4 \cdot N.gravidanze + \beta_5 \cdot Fumatrici + \beta_6 \cdot Sesso + \\ + \beta_7 \cdot Ospedale + \beta_8 \cdot Anni.madre + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. We can build this model by using the function `lm()` in R. The result of this model is the following:

Table 14: Coefficients of the linear regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6738.48	141.31	-47.69	0.00
Gestazione	32.57	3.82	8.53	0.00
Lunghezza	10.29	0.30	34.24	0.00
Cranio	10.47	0.43	24.58	0.00
N.gravidanze	11.27	4.66	2.42	0.02
Fumatrici	-30.16	27.54	-1.10	0.27
SessoM	77.54	11.18	6.94	0.00
Ospedaleosp2	-11.21	13.44	-0.83	0.40
Ospedaleosp3	28.10	13.50	2.08	0.04
Anni.madre	0.89	1.13	0.79	0.43
Tipo.partoNat	29.53	12.08	2.44	0.01

Gestazione, Lunghezza, Cranio, N.gravidanze, and Sesso are all statistically significant, with p-value less than 0.05. The adjusted R^2 is 0.73, which means that the model explains 72.78% of the variability of the weight of a newborn, which is not a quite good result. The F-statistic is 669.19 with p-value close to 0, which means that the model is statistically significant.

Table 15: Coefficients of the linear regression model (optimized)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6675.81	135.78	-49.17	0.00
Gestazione	31.19	3.78	8.25	0.00
Lunghezza	10.24	0.30	34.05	0.00
Cranio	10.64	0.42	25.08	0.00
SessoM	79.07	11.20	7.06	0.00
Tipo.partoNat	29.11	12.11	2.40	0.02

The model in general is quite good, but certainly not excellent for testing our study objective in a professional context. We can try to improve it by removing some variables, with the stepwise backward selection method, which allow to remove varibles with high p-value one at a time. We also could performe a mixed selection, but we prefered to simplify the model, as much as possible. We can try to remove the variable "Anni.madre", which is the variable with the highest p-value. We obtain that the adjusted R^2 is 0.73, the same same as before. We then proceed to remove variables "Ospedale" and in the following step "Fumatrici". Again we obtain that the adjusted R^2 is 0.73. We finally proceed removing "N.gravidanze", obtaining once again the same the adjusted R^2. We decide to not remove variables "Sesso" and "Tipo.parto", both because they are significant variables and because they are control variables. We can check the validity of the model by using the function `anova()`, which compares models and provides the p-value of the F-statistic. The p-value is NA, which means that the model is a bit improved. Moreover, we can check the BIC of the optimized model, which is c(7, 12), better than the BIC of the previous model, which was c(35222.59, 35241.84). Finally, we can check the VIF of the optimized model, which are all less than 5, which means that there is no multicollinearity problem.

Model interaction and no linear effects (Point 4)

We can try to improve the model by adding some interaction terms. We can add the interaction term between "Gestazione" and "Lunghezza", and the interaction term between "Gestazione" and "Cranio". The result of the model is in the following table.

Table 16: Coefficients of the linear regression model with interaction terms

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-328.28	1108.05	-0.30	0.77
Gestazione	-138.20	29.56	-4.68	0.00
Lunghezza	9.17	3.76	2.44	0.01
Cranio	-7.46	6.44	-1.16	0.25
SessoM	73.15	11.19	6.54	0.00
Tipo.partoNat	27.81	12.04	2.31	0.02
Gestazione:Lunghezza	0.03	0.10	0.34	0.73
Gestazione:Cranio	0.47	0.17	2.85	0.00

This model breaks the significance of variable "Cranio", and also the interaction between "Gestazione" and "Lunghezza", seems to be not significant. We decide to remove interaction between Gestazione and Lunghezza, and we obtain the result in the following table.

Table 17: Coefficients of the linear regression model with interaction terms (optimized)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-319.73	1107.57	-0.29	0.77
Gestazione	-138.28	29.55	-4.68	0.00
Lunghezza	10.45	0.30	34.72	0.00
Cranio	-9.31	3.48	-2.68	0.01
SessoM	73.31	11.17	6.56	0.00
Tipo.partoNat	27.82	12.04	2.31	0.02
Gestazione:Cranio	0.52	0.09	5.78	0.00

The model is improved with respect to the previous one. We can check the validity of the model by using ANOVA and BIC, which provide us a very little improvement of the model. We finally can try to add non linear effects to the model, for example a lograritmic effect of "Gestazione" and "Lunghezza". We opted for a logaritmic effect because the scatterplot of the variables seems to suggest a logaritmic effect.

Table 18: Coefficients of the linear regression model with interaction terms and non linear effects

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48328.82	8790.95	5.50	0.00
Gestazione	-398.94	106.42	-3.75	0.00
Lunghezza	44.47	3.75	11.86	0.00
Cranio	-1.02	5.83	-0.17	0.86
SessoM	72.67	10.98	6.62	0.00
Tipo.partoNat	26.99	11.83	2.28	0.02
log(Gestazione)	12833.53	2584.19	4.97	0.00
log(Lunghezza)	-16475.13	1806.40	-9.12	0.00
Gestazione:Cranio	0.30	0.15	1.99	0.05

We obtained an R^2 adjusted of 0.74, which is a bit improved with respect to the previous model. Moreover we can see that we can remove the interaction term.

Table 19: Coefficients of the linear regression model with interaction terms and non linear effects (without Gestazione:Cranio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48328.82	8790.95	5.50	0.00
Gestazione	-398.94	106.42	-3.75	0.00
Lunghezza	44.47	3.75	11.86	0.00
Cranio	-1.02	5.83	-0.17	0.86
SessoM	72.67	10.98	6.62	0.00
Tipo.partoNat	26.99	11.83	2.28	0.02
log(Gestazione)	12833.53	2584.19	4.97	0.00
log(Lunghezza)	-16475.13	1806.40	-9.12	0.00
Gestazione:Cranio	0.30	0.15	1.99	0.05

In the upper table we have removed the interaction term. We can observe that the model is improved with respect to the initial one, and it is improved also with respect to the model selected before as reference. In fact,we obtained an adjusted R^2 of 0.74, which is a bit better than the adjusted R^2 of the previous model, which was 0.73. Moreover the anova test provide a p-value very close to 0, which also means that the model is improved with respect to the previous one. Finally, the BIC of the model is lower than the BIC of the previous model, which means that this is a better model.