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
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.

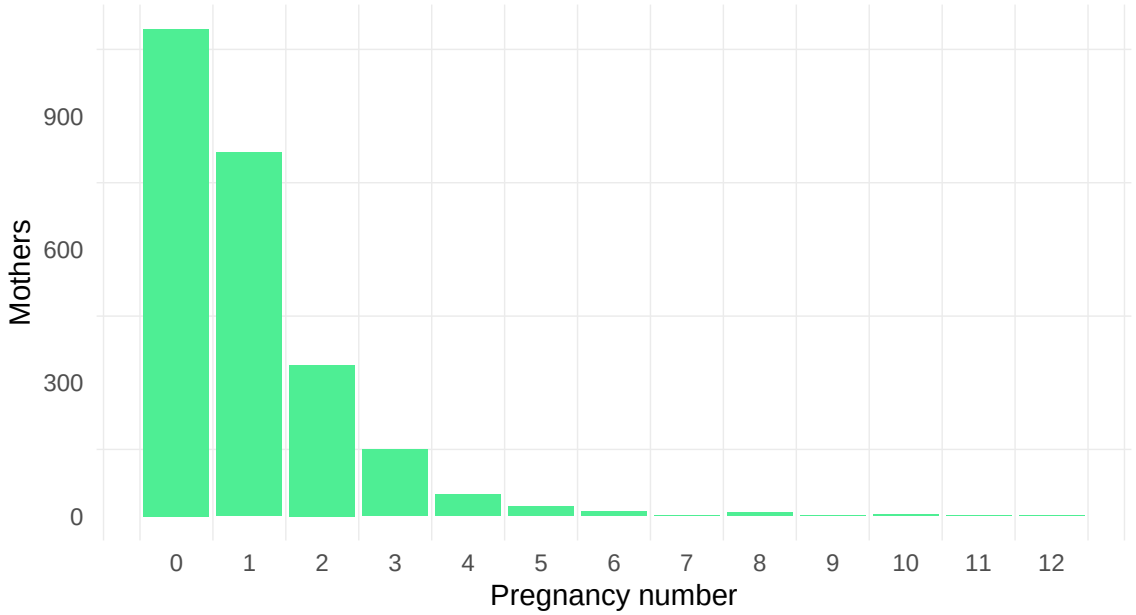

Figure 1: N.gravidanze distribution

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.

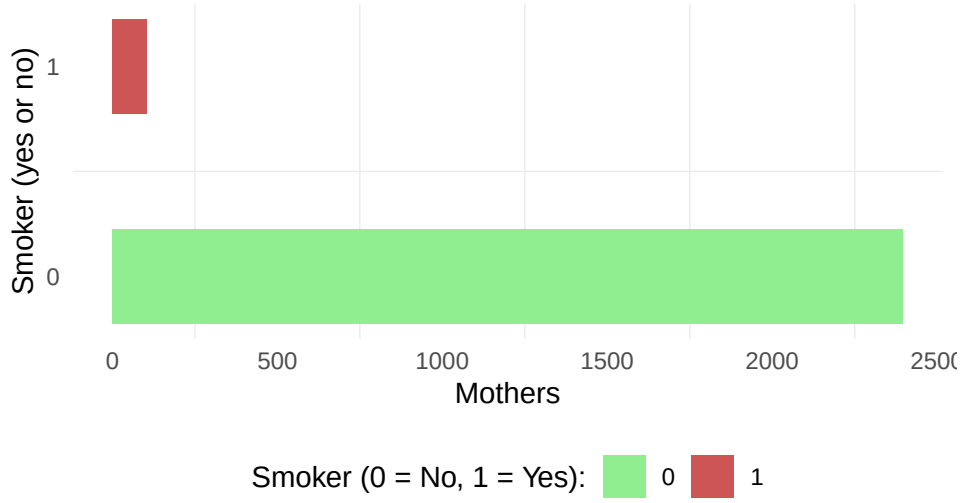

Figure 2: Smoker vs Not Smoker

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 a Normal distribution, 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.

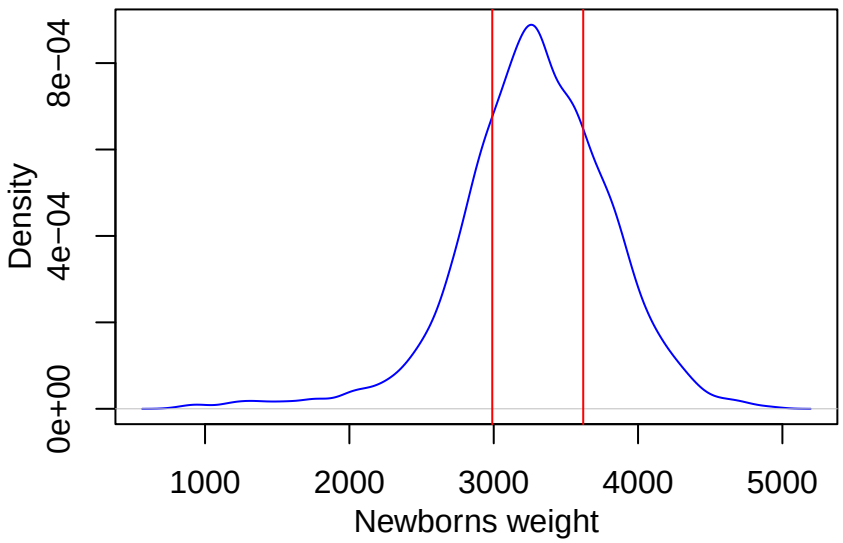

Figure 3: Peso distribution

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

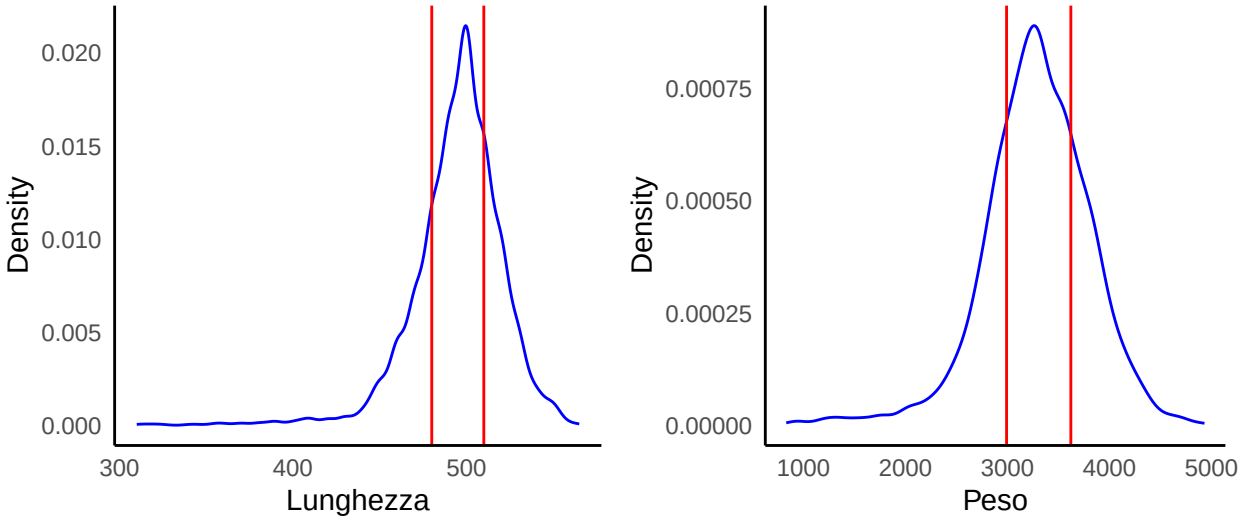

Figure 4: Lunghezza distribution vs. Peso distribution

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 sex 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".

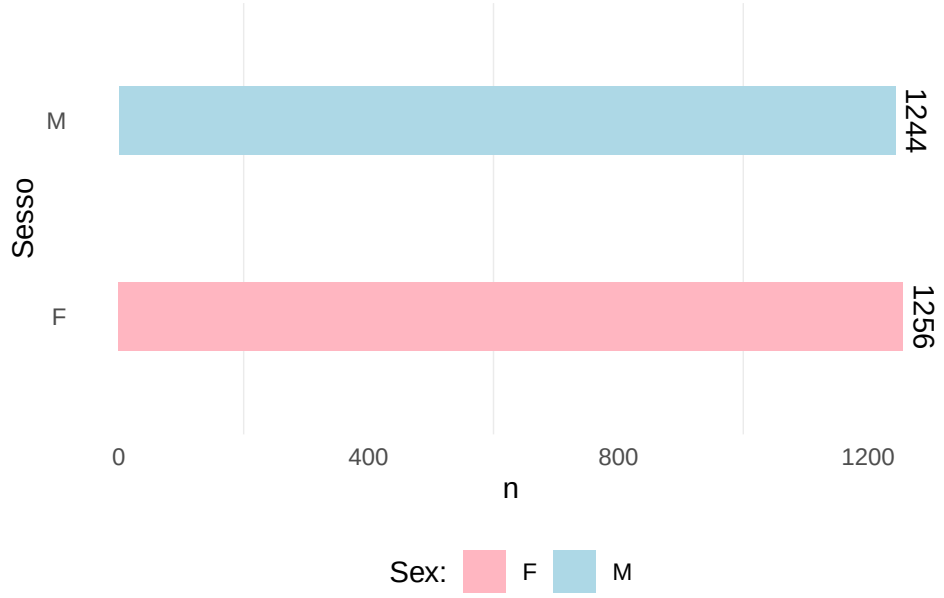

Figure 5: Sex distribution

Descriptive statistics

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.

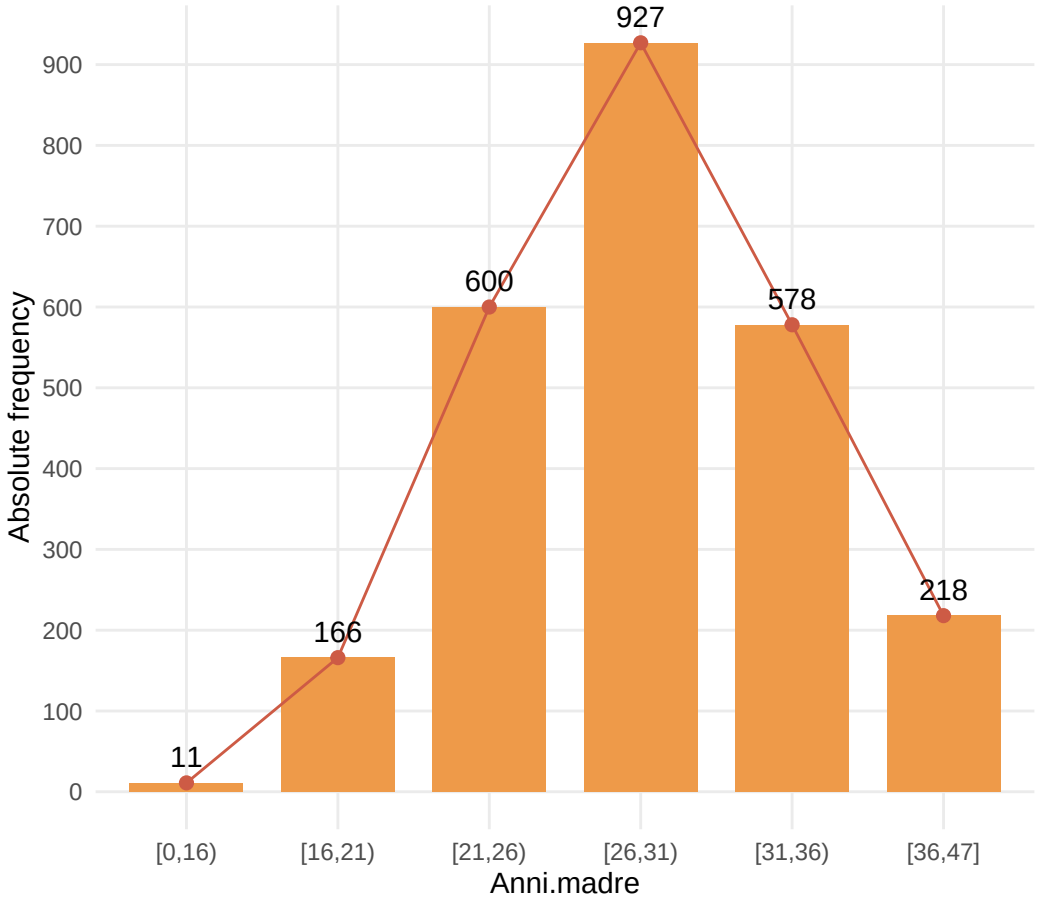

Figure 6: Frequency distribution of "Anni.madre" class

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

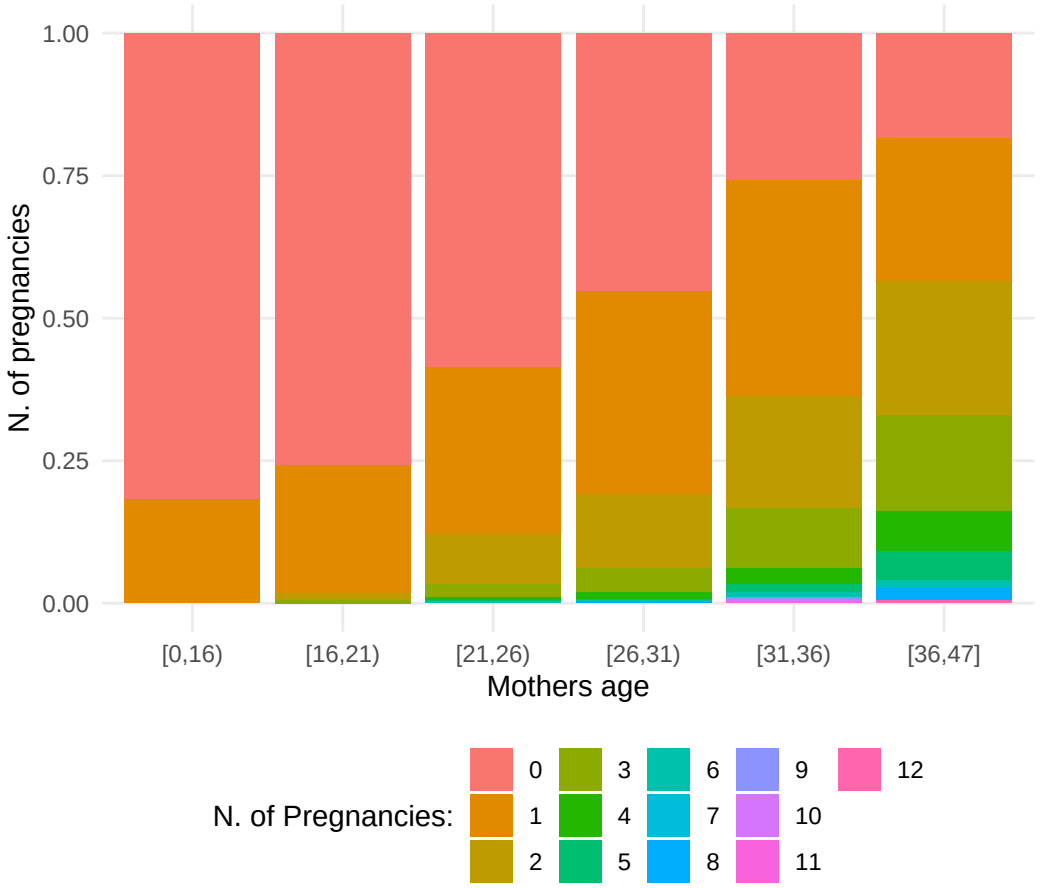

Figure 7: N. of pregnancies per class of ages

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. 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 the figure, is only a small percentage compared to the number of non-smoking mothers.

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