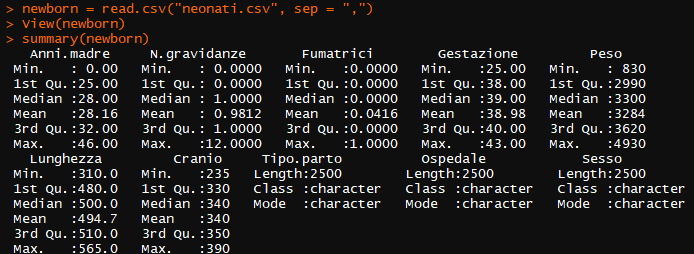
STATISTICAL MODEL TO PREDICT NEWBORN WEIGHT

1. **Loading the “neonati.csv” dataset and visualize it:**
2. **Describe the dataset composition and the variable types:**

The dataset is composed of 2500 observations of 10 variables. Their classification is the following:

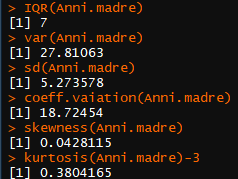
* **“*Anni.madre*”**: quantitative continuous on ratio scale;
* **“*N.gravidanze*”**: quantitative discrete on ratio scale;
* **“*Fumatrici*”**: qualitative nominal, codified in numbers;
* **“*Gestazione*”**: quantitative continuous on ratio scale;
* **“*Peso*”**: quantitative continuous on ratio scale;
* **“*Lunghezza*”**: quantitative continuous on ratio scale;
* **“*Cranio*”**: quantitative continuous on ratio scale;
* **“*Tipo.parto*”**: qualitative nominal;
* **“*Ospedale*”**: qualitative nominal;
* **“*Sesso*”**: qualitative nominal;

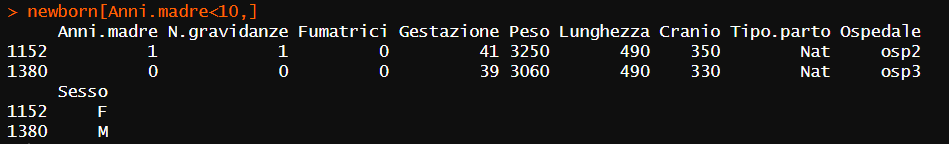
The *summary()* command gives back a first look at the position indexes of the dataset’s variables. R software considers three of them as qualitative nominal (“*Tipo.parto*”, “*Ospedale*”, and “*Sesso*”) and the other seven as qualitative ones. However, “Fumatrici” is a qualitative variable numerically codified (dummy variable) as already specified.

1. **Brief Descriptive analysis of “neonati.csv” dataset variables:**

* A graph of a normal distribution

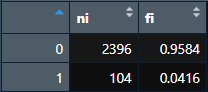
  Description automatically generated“*Anni.madre*”: plotting the density distribution of the variable shows how there are some outliers (2) close to the origin of the axis. For biological reasons, these are probably incorrect values (it is impossible to become a mother before puberty). Moreover, the graph and the shape indexes, display how the variable distribution is leptokurtic, and slightly asymmetric positively.

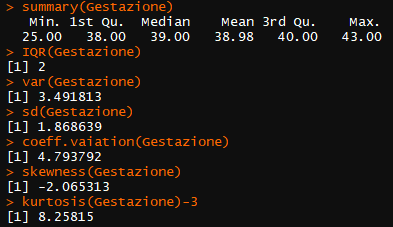


Looking at the two incorrect values, it is plausible they result from a wrong typing of age data. Indeed, the remaining values related to the observations 1152 and 1380 are completely credible. Correcting them is an option but it would be hard to guess the actual age values. Thus, these two observations will still be considered unless other issues emerge in the analysis subsequent steps.

“*N.gravidanze*”: many women included in the study had no previous pregnancies (43%), and around 10 % had already had more than two childbirths.



* “*Fumatrici*”: the frequency table shows how the majority of women who carried a pregnancy are non-smokers (“0”, 95,84 %), whereas only a small percentage smoked during maternity (“1”, 4,16 %).



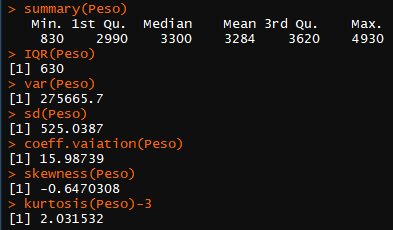
* “*Gestazione*”: its distribution does not follow a normal-like shape as shown by the below graph and confirmed by the shape indexes.

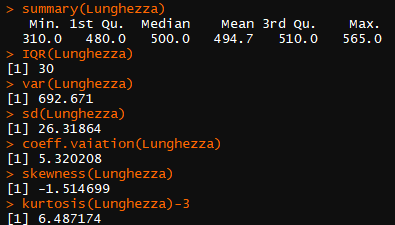
A graph of a person with a red line

Description automatically generated

* “*Peso*”: it has a leptokurtic distribution and presents a negative asymmetry.

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

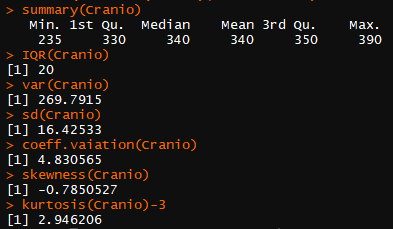


* “*Lunghezza*”: again, the distribution is leptokurtic and negatively asymmetric.

Immagine che contiene testo, diagramma, linea, Diagramma

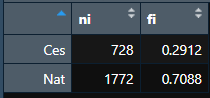
Descrizione generata automaticamente

* “*Cranio*”: the distribution presents a small left asymmetry and it is leptokurtic.

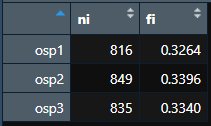
Immagine che contiene testo, diagramma, Diagramma, linea

Descrizione generata automaticamente

* “*Tipo.parto*”:

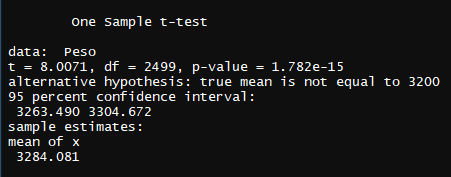


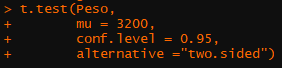
* “*Ospedale*”:

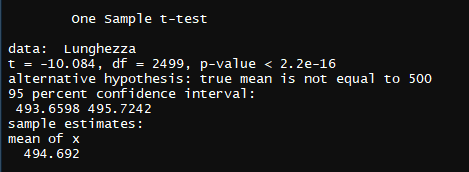


* “*Sesso*”:

1. **Test if “Peso” & “Lunghezza” means differ from the real population data:**

As previously emerged, neither “Peso” nor “Lunghezza” variables are normally distributed. However, the Central-limit Theorem states that for large n, the distribution of a sample estimator (like the sum or the mean) approximates to a Gaussian curve independently of the variable population distribution. Therefore, considering a sample size of 2500, parametric tests will be adopted.





The worldwide mean weight and height of a newborn with full gestation are equal to 3,2 kg and 50 cm. Both significantly differ from the sample mean values (3284 g & 494.7 mm).

1. **Test for the same variables and the remaining ones (when the comparison is worth) the difference between males and females:**

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente



Considering the variable “*Peso*” the Welch Two Sample t-test reports a statistically significant difference between sexes.

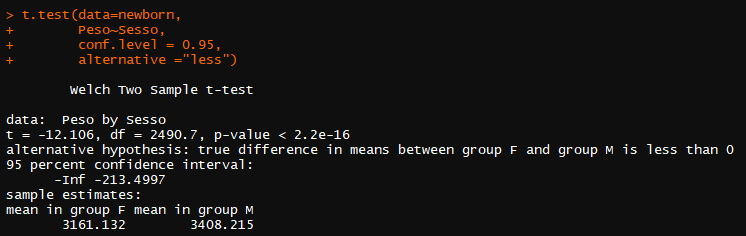
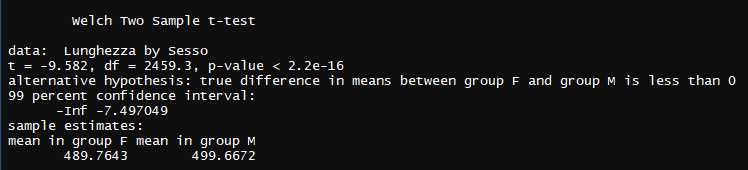


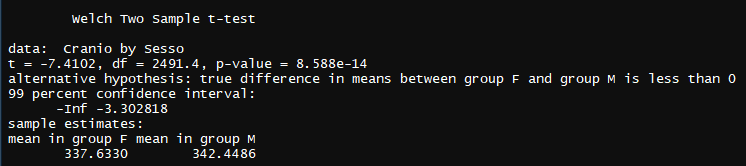
Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

On the other hand, also “*Lunghezza*”, “*Cranio*”, and “*Gestazione*” t-test results point out statistically significant differences between males and females, with males having higher sample means (p-values = 2.2e-16, 8.588e-14, and 1.228e-11).

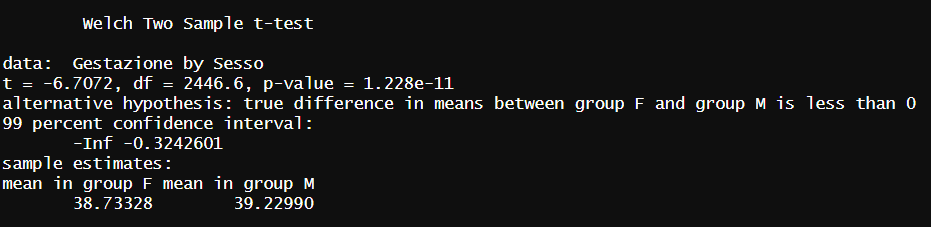
Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

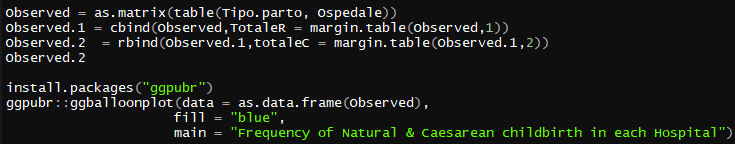


**A graph with a red and blue rectangle

Description automatically generated**

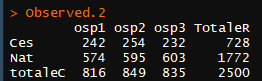


1. **Are some hospitals performing more Caesarean deliveries than others?**

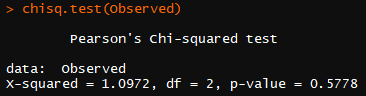
****Firstly, with R help is possible to visualize the contingency table with the marginal frequencies of the two qualitative variables, together with the balloon plot:

**A graph of a number of dots

Description automatically generated**

****

To understand if some hospitals are more likely than others to practice caesarean section, the Chi-squared test might come to help. Indeed, it is useful for assessing if there is an association between the variables “*Ospedale*” and “*Tipo.parto*” or whether the frequency of delivery is not affected by the hospital (independence).

****

The result of Pearson’s Chi-squared test shows how there is no significant difference in delivery type considering the three hospitals contained within the dataset (p-value > 0.5) establishing how these two variables are independent of each other.

**MULTIDIMENSIONAL ANALYSIS**

1. **Investigate the relationship between the Response variable and Predictors**

This project goal is to unveil if the newborn weight can be predicted using the remaining variables of the dataset “newborn”. The first step in this complex process consists of evaluating the relationship between the Response (“*Peso*”) and the Predictors. Moreover, it is also important to assess what relation occurs between Response variables:

Immagine che contiene parole crociate, testo, quadrato, bianco e nero

Descrizione generata automaticamente

The upper panel furnishes some useful information:

* The highest correlated covariates with newborns’ weight are “*Gestazione*”(ρ = 0.59), “*Lunghezza*” (ρ = 0.80), and “*Cranio*” (ρ = 0.70) which is foreseeable. Indeed, it is likely for babies that have a longer gestation time to weigh more at childbirth. In the same way, infants that weigh more will also have a wider skull diameter or they will be higher, on average;
* Immagine che contiene testo, diagramma, linea

  Descrizione generata automaticamenteFor the same reason, these three explanatory variables are highly correlated with each other (“*Cranio*” vs “*Lunghezza*”, ρ = 0.60; “*Cranio*” vs “*Gestazione*”, ρ = 0.46; “*Gestazione*” vs “*Lunghezza*”, ρ = 0.62). Thus, if all are included in the regression model they might give multicollinearity-related issues with subsequent instability of β coefficients;Immagine che contiene testo, diagramma, schermata

  Descrizione generata automaticamenteImmagine che contiene testo, diagramma, schermata, Parallelo

  Descrizione generata automaticamente
* For what concerns the quantitative variables, “*Peso*” does not seem to be associated with them. On the other side, “*Anni.madre*” has a moderate correlation with “*N.gravidanze*” (ρ = 0.38).

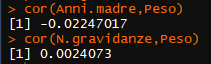


Immagine che contiene testo, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, diagramma, schermata, Parallelo

Descrizione generata automaticamente

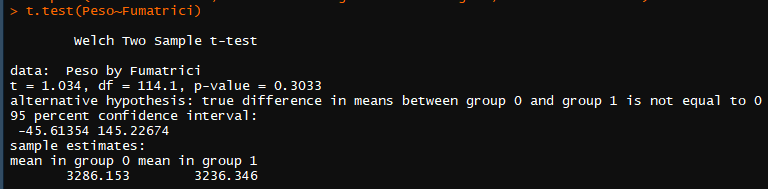
* Knowing that an infant’s weight at birth significantly varies considering its sex, let’s examine if other qualitative variables might be impacting weight prediction:

Immagine che contiene testo, diagramma, schermata, Rettangolo

Descrizione generata automaticamenteNo statistically meaningful differences occur when infants’ weight is considered in smoker vs non-smoker mothers or in differnet delivery types.



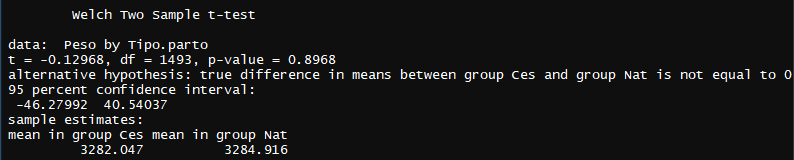
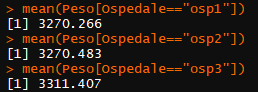




Immagine che contiene testo, diagramma, schermata, Parallelo

Descrizione generata automaticamente



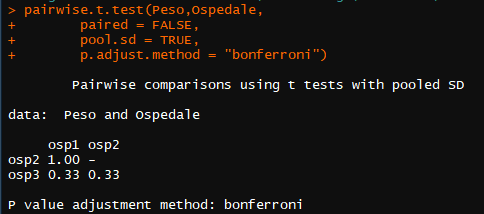


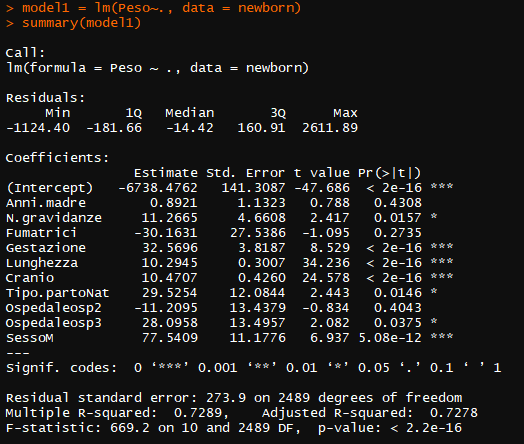
Immagine che contiene testo, diagramma, Disegno tecnico, Piano

Descrizione generata automaticamente

There is no relevant difference in weight also among hospitals. The Bonferroni corrected p-value relative to the difference between hospitals 1 and 2 is even 1.00 (due to R software correction). Indeed, looking at mean values per hospital the two are almost identical numbers, differing just for the decimal components.

1. **Create a Multivariate Linear Regression model containing all the variables of the dataset**

Building the regression model by taking into account all the possible Predictors, produces the β coefficients (and respective p-values) shown below. It can be noticed how “*Anni.madre*”, “*Fumatrici*”, and “*Ospedaleosp2*” (dummy variable that works together with “*Ospedaleosp3*” and is nothing less than the result of “*Ospedale*” transformation) have non-significative β coefficients, with respective p-values of 0.4383, 0.2735, and 0.4043. Therefore, they do not give a precious contribution to the weight estimation and will be removed from the model. These considerations are in line with the correlation coefficients previously elaborated.

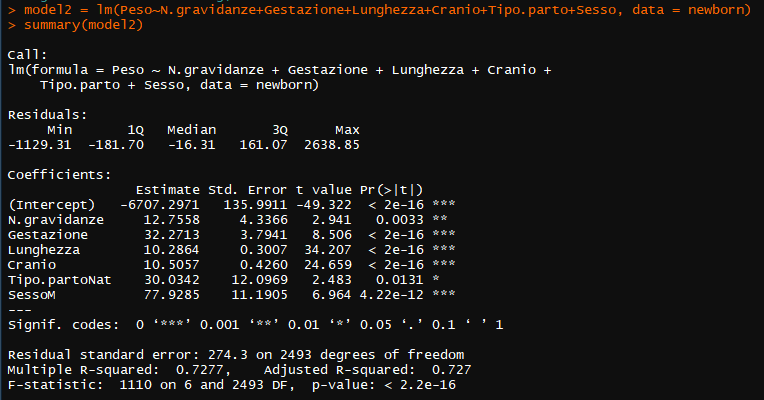


On the other hand, all the other regression coefficients have statistically significant p-values:

* “*N.gravidanze*”, “*Ospedaleosp3*”, and “*Tipo.partoNat*” p-values ~ 0.01;
* “*Gestazione*”, “*Lunghezza*”, “*Cranio*”, and “*SessoM*”, (adding 32.56, 10.29, 10.47, and 77.54 grams of weight for each predictor unit variation respectively) with p-values close to 0, indicating how meaningful they are in foreseeing the Response value;

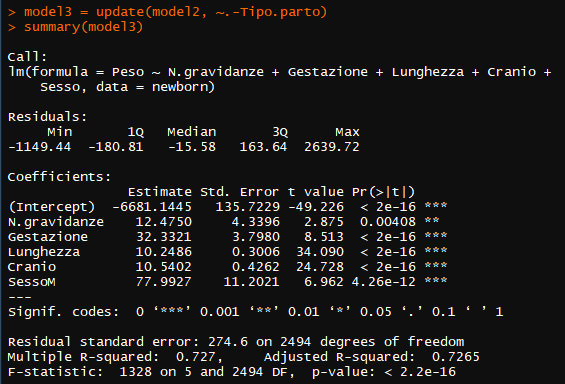
Overall, the model explains around 72 % of the Outcome variable total variability (R2 = 0.7289 and adjusted R2 = 0.7278).

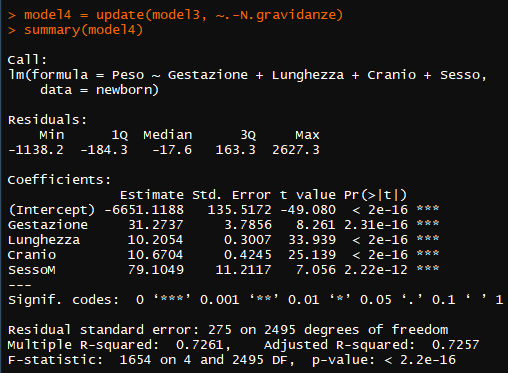
1. **Find the “best” model using the known selection parameters:**

A first update of the model, can be performed by excluding the non-statistically significant regressors:

After this update, “*N.gravidanze*” β-coefficient acquires significance compared to the previous elaborations (differently from what occurred for the correlation coefficient analysis). However, the adjusted R2 does not change much (0.7278 vs 0.7270).

A further perfecting step consists of maintaining exclusively the regressors with the smallest p-values according to the parsimony principle (not including anything more within the model unless is it essential) and evaluating what happens to the model overall:





The new adjusted R2 are respectively 0.7265 (model3) and 0.7257 (model4) after having removed two regressors. Thus model3 seems the more promising between the two.

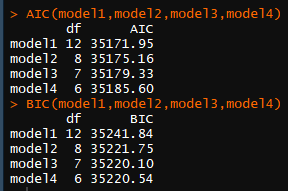
As noticed before, three out of five variables included in the 3rd model3 present high correlation coefficients between each other. This might be a problem, 1) because it would not be possible to estimate the effect of each predictor on the response independently; 2) the β-coefficient estimations would widely change depending on which other independent variable is included in the model. Given that, the Variance Inflation Factor (VIF) is a powerful index that helps in understanding if some variables should be excluded (VIF > 5):

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

No VIF among the ones calculated exceeds 5, therefore it seems there are not multicollinearity-related issues.

Through the computation of Akaike Information Criterion (AIC) and Baesian Information Criterion (BIC) it is possible to select the best model. Theoretically, the best situation occurs when both variance and bias are minimized producing a model that fits pretty well to current data and it is also capable of adapting when further observations are added (bias-variance trade-off).

Model1 has the lowest AIC and model3 has the lowest BIC. AIC tends to prefer more overfitted models compared to BIC, and the process of selection so far has been done following the parsimony principle. Maintaining this criterion of judgment, model3 can be considered better than model4, having a slightly smaller BIC (35220.10 vs 35220.54).

Anyway, even through AIC evaluation, model3 remains the most suitable choice (35179.33 vs 35185.60).

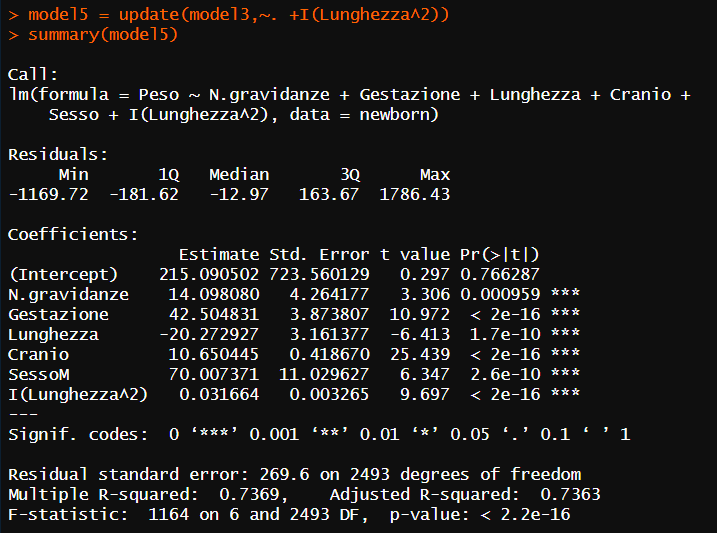
1. **Non-linear effects or Interaction terms**

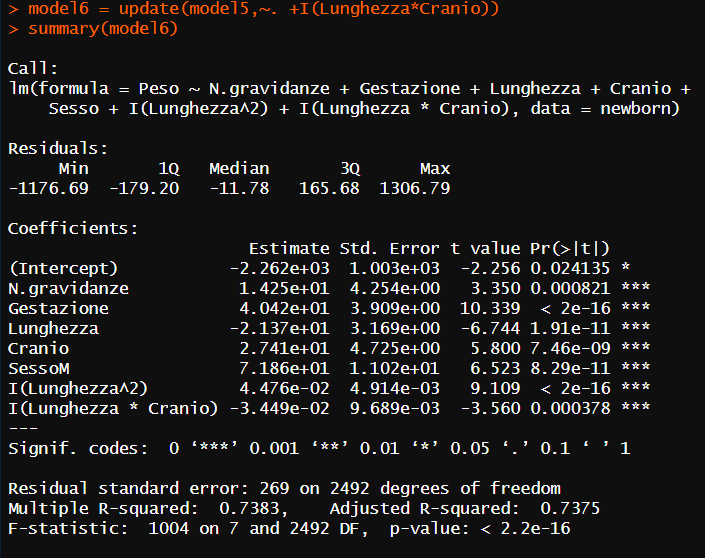


* Immagine che contiene testo, schermata

  Descrizione generata automaticamenteNon-linear effects: The figure above is the result of plotting “*Gestazione*”, “*Cranio*”, and “*Lunghezza*” versus the dependent variable. With the addition of *geom\_smooth()* command, it is possible to create a line approximating the trend of the scatter plot, making it easier to visually spot non-linear correlations. None of the three lines generated has a perfect linear trend. Specifically, “*Lunghezza*” red line shape might indicate a quadratic relationship with “*Peso*”. Indeed, there is one observation with low height and suspicious high weight whose effect bends the curve upward. On the other side, these three variables either indicate the time in which the fetus develops (gestation time) or the proportions of the infant’s body at delivery (length or skull diameter) as previously pointed out. It is expected that the more time the fetus spends in their mother’s womb, the heavier it will be at delivery, as it is likely that the body measurements will grow more or less proportionally to each other. For these reasons, non-linear correlations seem unlikely.

Nevertheless, adding the non-linear effects of “*Lunghezza”, “Cranio”,* and *“Gestazione”* within the model results in an overall increase of the adjusted R2 only when the quadratic term of the newborn’s length is included (0.7363 of model5 versus 0.7265 of model3).



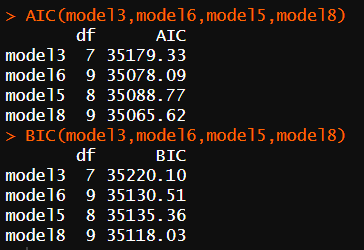
* Interaction terms: the model built consists of five independent variables: “*N.gravidanze*”, “*Gestazione*”, “*Lunghezza*”, “*Cranio*”, and “*Sesso*”. Verifying the possible interactions between regressors negatively influences the p-values associated with the single-term coefficients in the most of cases except for *Lunghezza\*Cranio* and *Gestazione\*Lunghezza* interaction terms. However, the first interaction does not lead to an increment in the Response variable explained variability (same adjusted R2), therefore it will not be considered further.

On the other hand, the second term contained in the model8 brings a more consistent increment in the adjusted R2 (0.7375 vs 0.7388).

A computer screen shot of a program

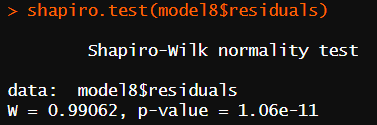
Description automatically generated

Importantly, both AIC and BIC confirm the superiority of model8 when compared with previous ones.



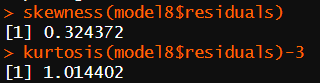
1. **Residual Analysis**

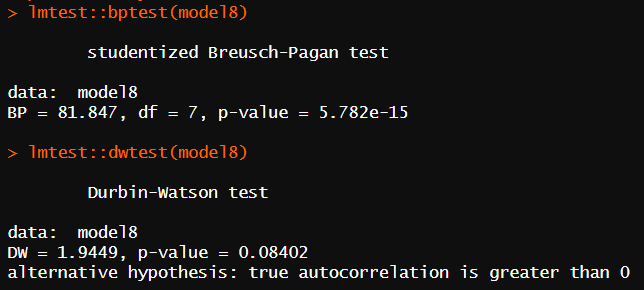
This step of the analysis is crucial to 1) control if residuals are normally distributed: this is necessary to ensure that predictions performed based on the model will be accurate; 2) make sure that the erratic part of the model does not contain any information that could “escape” from the deterministic portion and weaken its statistical power: there should not be any relievable pattern; 3) Analyse if some outliers or leverage values (extreme Response or Predictors observations) might impact on the regression model adequacy.



A graph of a normal distribution

Description automatically generated

The Shapiro-Wilk test for normality states that residual distribution is different from a normal one (p-value < 1.06e-11). Indeed, the curve is leptokurtic even though the residual population is not far from a Gaussian shape (the Shapiro-Wilk test is pretty sensitive to variation from the normal distribution). However, this element may contribute negatively to the model's adequacy by impacting β-coefficients hypothesis testing validity. The Breusch-Pagan test performs the homoscedasticity check. Homoscedasticity or homogeneity of variance, is an important assumption to be met for ensuring an accurate prediction across the whole range of the model. In this case, homoscedasticity is violated (p-value = 5.782e-15) and prediction accuracy might be affected.

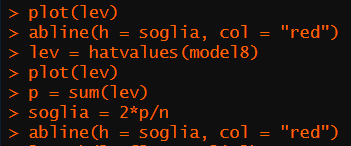
The Durbin-Watson test on the other side, controls whether residuals are correlated. If a certain level of autocorrelation is present, it means that there is some hidden pattern in the erratic part that has not been explained by the model itself. In this case, the test is not significative, indicating non-correlated residuals (p-value = 0.08402).

A group of graphs showing different types of data

Description automatically generatedTo look deeper at the model and understand if something needs to be ameliorated, the below plot comes as a useful tool:

* The upper-left frame plots the Residuals versus the Fitted values and it is useful to visualize linearity & independence: it is fundamental for residuals mean to match the zero across the whole length of the x-axis otherwise predictions detach from the actual reported value. Moreover, any trends besides a random distribution around zero-mean would indicate the violation of the independence assumption and an incomplete or inadequate model building. In this case, residuals are randomly scattered around zero even though there is not a complete overlapping of the mean (red line) with the x = 0 line;
* The upper-right Q-Q plot checks for the normality of residuals: the perfectly normal population would lay on the bisector line. As stated by the Shapiro-Wilk test, this is not the case;
* The lower-left graph evaluates the homogeneity of variance & independence. If the homoscedasticity assumption is violated, β coefficients are less precise, and therefore p-value estimates are incorrect (smaller) leading to false predictions. Again, there is a slight upward inclination on both extremes;
* Lower-right panel shows Residuals versus Leverage, and it is needed to spot potentially influential values (present among outliers or high leverage values that are respectively extreme outcome or predictor variable values). The inclusion of influential values might modify the results of the regression model. Cook’s distance is a quantitative parameter that helps understand if the model contains some influential values. In this case, the observation number 1551 has a Cook’s distance higher than 1, therefore it is conditioning the model to some extent.

To better analyze the influential-cases-related issues, it is possible to examine high-leverage values:



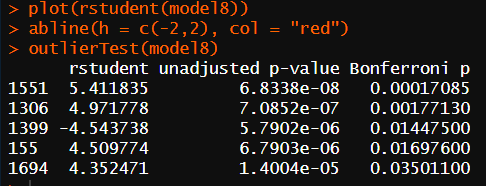
Finding 134 high-leverage observations present in model number 8.

A graph with numbers and lines

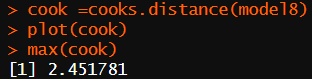
Description automatically generated with medium confidence

On the other side, outliers can be examined:

A graph showing a number of dots

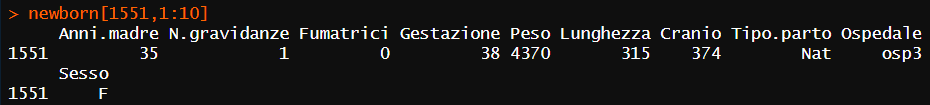
Description automatically generatedIn this case, five values are reported by R (among which there is the value 1551), and way more are visually displayed;

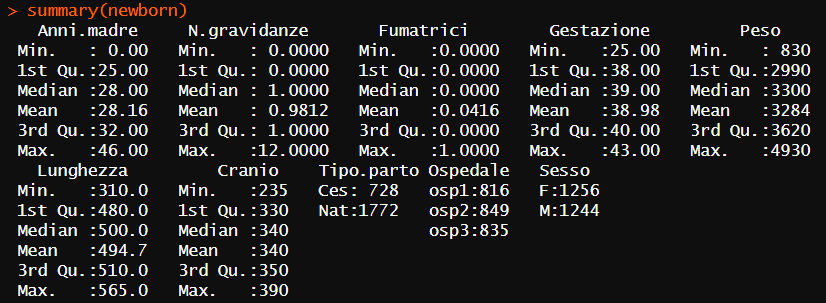
The Cook’s distance is the most concise way for influential values identification:

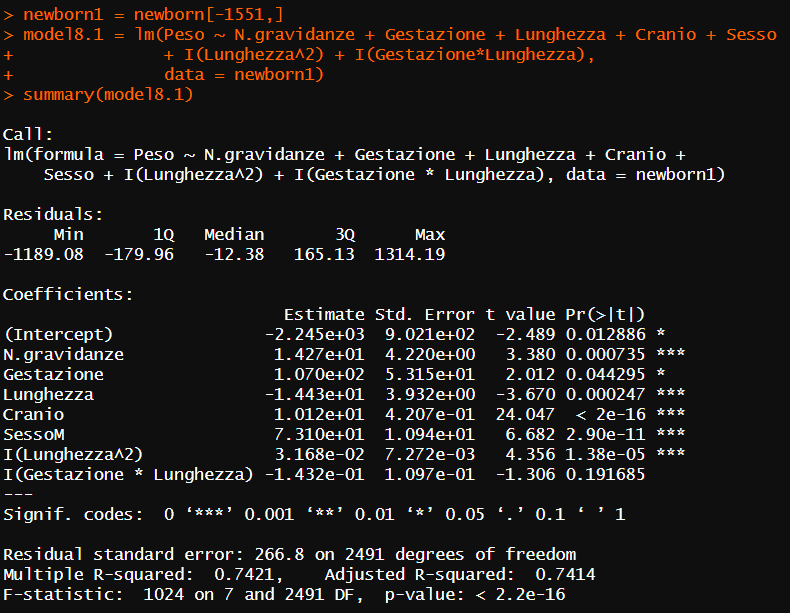
R identifies one observation with a value of 2.451, mirroring the result of the Residuals versus Leverage plot which identified observation 1551 with a Cook’s distance higher than 1.

A graph with numbers and lines

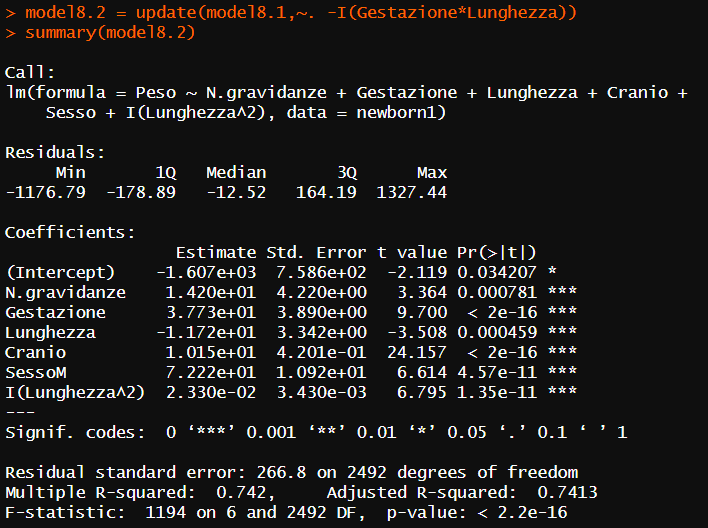
Description automatically generated with medium confidence

This observation refers to a baby girl with a weight and a cranium diameter above the 3rd quartile (3620 grams and 350 millimeters), both close to their maximum value (4370 in 4930 grams and 374 millimeters in 390). The same child has an uncommon height for these measurements (315 millimeters) which is close to the minimum value for that measure (310.0 millimeters).

This specific case has uncommon relationships among weight, height, and cranium diameter, very different than the average. The reasons behind this might be diverse: a mistake in data recording or even the presence of a pathological condition. Eliminating observation 1551 might benefit the overall model predictive capacity.

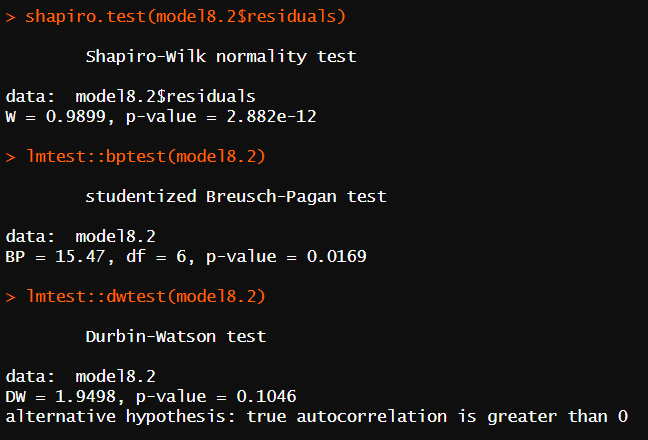
1. **How good is the model for making predictions?**

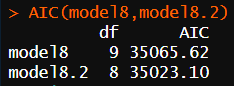
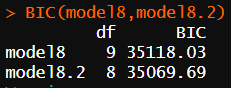
Model8.1 has an adjusted R2 of 0.7414 vs 0.7388 of model8, increasing the quantity of Outcome variability explained by the independent variables. However, the p-values of the interaction term between *Gestazione* and *Lunghezza* is not significant anymore, therefore, it can be deleted:



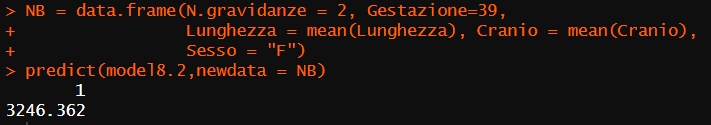
Overall, the new model8.2 presents a slightly smaller adjusted R2 than model8.1 (0.7413 versus 0.7414). Moreover, all the terms included have statistically significant p-values.

Model 8.2 still presents a slight level of heteroscedasticity, despite being lower than previous ones (Breush-Pagan test p-value of 0.0169 versus 5.782e-15of model8).

Both AIC and BIC confirm the superiority of model8.2. Usually, AIC and BIC decrease when observations are added, because the more the sample grows, the more it gets similar to the reference population. In this case, despite an observation was removed, both criteria consequentially decrease, confirming the improvement in respect to its predecessors.



1. **Predict the weight of a baby girl knowing that she will be born at 39 weeks of gestation, by a mother that is having her third gestation:**

****

1. **Graphical representations of the simplified model**

Through the *scatterplot3d()* command of the *scatterplot3d()* package is possible to visualize the model. However, representing six variables within a two-dimensional space can not be performed without losing some information. Therefore, to elaborate a graphical representation as trustworthy as possible, four variables (the response and three predictors) are examined together:

1. A graph of a graph with dots

   Description automatically generated with medium confidence
2. A graph with dots and lines

   Description automatically generated
3. A graph with a line graph and a diagram

   Description automatically generated with medium confidence
4. A graph of a graph with a line graph

   Description automatically generated with medium confidence