Un Modello Statistico per Prevedere il Peso dei Neonati

1) Importo il database con la funzione read.csv

```
dati <- read.csv("neonati.csv",
                 stringsAsFactors = T)
```

2) Il dataset contiene 2500 nascite ognuna composta da dieci variabili, vediamole nel dettaglio:

- Anni.madre: quantitativa continua, indica gli anni della madre al momento del parto.
- N.gravidanze: quantitativa discreta, indica eventuali gravidanze avute in precedenza.
- Fumatrici: categorica binaria, indica se la madre è una fumatrice o meno.
- Gestazione: quantitativa continua, indica il periodo di gestazione espresso in settimane.
- Peso: quantitativa continua, indica il peso del neonato espresso in grammi.
- Lunghezza: quantitativa continua, indica la lunghezza del neonato in mm.
- Cranio: quantitativa continua, indica il diametro in mm del cranio del neonato.
- Tipo parto: categorica binaria, parto naturale o cesareo.
- Ospedale: categorica nominale, indica l'ospedale in cui è avvenuto il parto.
- Sesso: categorica binaria, indica il sesso del nascituro.

Obiettivo dello studio è quello di usare tutte le variabili in nostro possesso per creare un modello in grado di prevedere, indicando specifici parametri, il peso di un neonato alla nascita. In particolare si vuole indagare se le variabili della madre abbiano un effetto significativo su quelle del nascituro.

3) Indaghiamo brevemente alcune variabili degne di nota, partiamo con gli anni della madre:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moments)

attach(dati)

dati %>%
  summarise(media = mean(Anni.madre),
            sd = sd(Anni.madre),
            asimmetria = skewness(Anni.madre),
            curtosi = kurtosis(Anni.madre)-3)
```

```
##    media       sd asimmetria   curtosi
## 1 28.164 5.273578  0.0428115 0.3804165
```

L'età media del campione è di 28 anni, con una variazione standard di 5 anni, abbiamo una distribuzione leggermente asimmetrica positiva leptocurtica.

```
dati %>%
  summarise(media = mean(N.gravidanze),
            sd = sd(N.gravidanze),
            asimmetria = skewness(N.gravidanze),
            curtosi = kurtosis(N.gravidanze)-3)
```

```
##    media       sd asimmetria  curtosi
## 1 0.9812 1.280587   2.514254 10.98941
```

In media le donne prese in esame hanno già avuto una gravidanza in precedenza, abbiamo una deviazione standard di 1 circa con una distribuzione asimmetrica positiva leptocurtica.

```
dati %>%
  summarise(media = mean(Gestazione),
            sd = sd(Gestazione),
            asimmetria = skewness(Gestazione),
            curtosi = kurtosis(Gestazione)-3)
```

```
##     media       sd asimmetria curtosi
## 1 38.9804 1.868639  -2.065313 8.25815
```

Le gravidanze prese in esame hanno una durata di 39 settimane con una deviazione standard di 2. Abbiamo una distribuzione asimmetrica negativa leptocurtica.

```
dati %>%
  summarise(media = mean(Peso),
            sd = sd(Peso),
            asimmetria = skewness(Peso),
            curtosi = kurtosis(Peso)-3)
```

```
##      media       sd asimmetria  curtosi
## 1 3284.081 525.0387 -0.6470308 2.031532
```

I neonati presi in esame hanno un peso medio di 3.3 kg, con una deviazione standard di 525g. Abbiamo una distribuzione asimmetrica negativa leptocurtica.

```
dati %>%
  summarise(media = mean(Lunghezza),
            sd = sd(Lunghezza),
            asimmetria = skewness(Lunghezza),
            curtosi = kurtosis(Lunghezza)-3)
```

```
##     media       sd asimmetria  curtosi
## 1 494.692 26.31864  -1.514699 6.487174
```

I neonati presi in esame sono lunghi in media 495mm con una deviazione standard di 26mm. Abbiamo una distribuzione asimmetrica negativa leptocurtica.

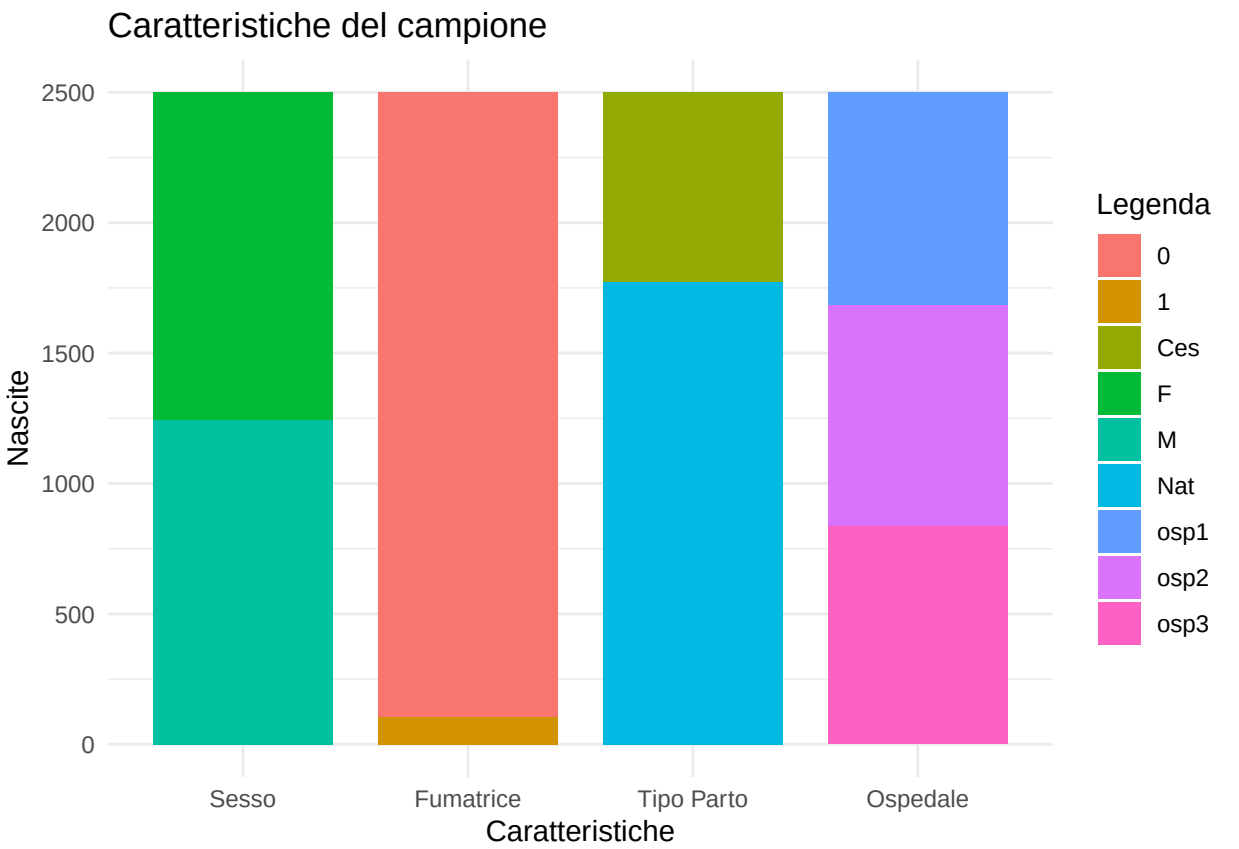
```
dati %>%
  summarise(media = mean(Cranio),
            sd = sd(Cranio),
            asimmetria = skewness(Cranio),
            curtosi = kurtosis(Cranio)-3)
```

```
##      media       sd asimmetria  curtosi
## 1 340.0292 16.42533 -0.7850527 2.946206
```

Il diametro del cranio del campione preso in esame è di 340mm con una deviazione standard di 16.5mm. Abbiamo una distribuzione leggermente asimmetrica negativa leptocurtica.

Proseguiamo con una rappresentazione grafica per rendere più chiaro a colpo d'occhio la composizione del nostro campione.

```
library(ggplot2)
df <- data.frame(
  Sesso = factor(Sesso),
  Fumatrice = factor(Fumatrici),
  TipoParto = factor(Tipo.parto),
  Ospedale = factor(Ospedale)
)

ggplot(df) +
  geom_bar(aes(x = factor(1), fill = Sesso),position = "stack", width = 0.8) +
  geom_bar(aes(x = factor(2), fill = Fumatrice), position = "stack", width = 0.8) +
  geom_bar(aes(x = factor(3), fill = TipoParto), position = "stack", width = 0.8) +
  geom_bar(aes(x = factor(4), fill = Ospedale), position = "stack", width = 0.8) +
  scale_x_discrete(labels = c("Sesso", "Fumatrice", "Tipo Parto", "Ospedale")) +
  labs(title = "Caratteristiche del campione",
       x = "Caratteristiche",
       y = "Nascite") +
  guides(fill = guide_legend(title = "Legenda"))+
  theme_minimal()
```



Dal grafico notiamo che:

- I neonati di sesso maschile e femminile sono presenti in egual quantità all'interno del campione.
- La maggior parte delle mamme sono non fumatrici (condizione codificata con 0).
- Poco più dei due terzi dei nascituri sono nati da un parto naturale.
- Le nascite sono equidistribuite fra i tre ospedali.

4) Per saggiare l'ipotesi che la media di peso e lunghezza del nostro campione di neonati sia uguale a quello della popolazione eseguiamo uno t test.

Iniziamo con il peso. In media i neonati alla nascita hanno un peso di 3.3kg. (FONTE: https://www.ospedalebambinogesu.it/da-0-a-30-giorni-come-si-presenta-e-come-cresce-80012/) Avremo quindi un'ipotesi nulla che sosterrà una significativa uguaglianza fra le medie, e un'ipotesi alternativa che sosterrà una differenza fra le medie.

Vado quindi ad eseguire il test t:

```
t.test(Peso, mu = 3300, conf.level = 0.95, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  Peso
## t = -1.516, df = 2499, p-value = 0.1296
## alternative hypothesis: true mean is not equal to 3300
## 95 percent confidence interval:
```

```
##  3263.490 3304.672
## sample estimates:
## mean of x
##  3284.081
```

Con un P-value di 0.13 non rifiutiamo l'ipotesi nulla. Le due medie non presentano differenze significative.

Passiamo ora alla lunghezza. Sempre secondo secondo la fonte citata in precedenza alla nascita i neonati presentano una lunghezza di 50cm. Avremo quindi un'ipotesi nulla che sosterrà una significativa uguaglianza fra le medie, e un'ipotesi alternativa che sosterrà una differenza fra le medie.

Vado quindi ad eseguire il test t:

```
t.test(Lunghezza, mu = 500, conf.level = 0.95, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  Lunghezza
## t = -10.084, df = 2499, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  493.6598 495.7242
## sample estimates:
## mean of x
##   494.692
```

Il test restituisce un P-value molto piccolo, suggerendoci quindi di rifiutare l'ipotesi nulla in quanto le due medie presentano differenze significative.

NOTA: La differenza di lunghezza fra la media della popolazione e quella del nostro campione è di soli 5mm, non rilevante per lo studio che stiamo effettuando. Non rifiutiamo quindi l'ipotesi che la media di peso e di lunghezza del nostro campione è significativamente uguale a quella della popolazione.

5) Verifichiamo se ci sono differenze significative fra i due sessi. Iniziamo con il peso:

```
t.test(Peso~Sesso)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Sesso
## t = -12.106, df = 2490.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -287.1051 -207.0615
## sample estimates:
## mean in group F mean in group M
##        3161.132        3408.215
```

Il test ci restituisce un p-value molto basso, rifiutiamo quindi l'ipotesi nulla di uguaglianza delle medie. Notiamo infatti nel nostro campione i neonati maschi sono in media più pesanti di 247g rispetto alle femmine.

Eseguiamo lo stesso test per la lunghezza:

```r
t.test(Lunghezza~Sesso)
```

```
##
##  Welch Two Sample t-test
##
## data:  Lunghezza by Sesso
## t = -9.582, df = 2459.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -11.929470  -7.876273
## sample estimates:
## mean in group F mean in group M
##        489.7643        499.6672
```

Anche in questo caso otteniamo un p-value molto basso e rifiutiamo quindi l'ipotesi nulla di uguaglianza fra le medie. I neonati maschi sono lunghi in media 10mm in più rispetto alle femmine.

Infine consideriamo la circonferenza del cranio:

```r
t.test(Cranio~Sesso)
```

```
##
##  Welch Two Sample t-test
##
## data:  Cranio by Sesso
## t = -7.4102, df = 2491.4, p-value = 1.718e-13
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -6.089912 -3.541270
## sample estimates:
## mean in group F mean in group M
##        337.6330        342.4486
```

Anche qui otteniamo un p-value molto piccolo e rifiutiamo l'ipotesi nulla di uguaglianza fra le medie. In media la circonferenza del cranio dei neonati maschi è più grande di 5mm rispetto alle femmine.

In conclusione possiamo dire che in media i neonati di sesso femminile hanno misure più piccole rispetto a quelli di sesso maschile. Notiamo che il nostro campione differisce dalla popolazione solo per il peso. Di norma infatti la differenza di peso fra maschi e femmine è di 150g in più per i primi, mentre nel nostro campione abbiamo osservato una differenza di 247g. Per quanto riguarda la lunghezza, nella popolazione non si riscontrano particolari differenze di lunghezza fra maschi e femmine, e nel nostro campione ce ne sono di minime con una differenza di 10mm fra maschi e femmine.

6) Verifichiamo se c'è un ospedale in cui vengono fatti più parti cesarei. Comincio convertendo in variabili numeriche binarie "Nat" e "Ces", d'ora in poi 0 sarà da intendersi "Naturale" e 1 "Cesareo".

```r
dati <- dati %>%
  mutate(Tipo.parto = ifelse(Tipo.parto == "Nat", 0, 1))
```

Creo poi una tabella di contingenza dove saranno indicati il numero di parti naturali e cesarei per ogni ospedale.

```r
tabella <- table(Tipo.parto, Ospedale)
tabella
```

```
##           Ospedale
## Tipo.parto osp1 osp2 osp3
##        Ces  242  254  232
##        Nat  574  595  603
```

Già dalla tabella è possibile notare che le cifre non variano significativamente ma continuiamo con la nostra indagine, andiamo ad eseguire un test chi quadrato per verificare se c'è qualche associazione fra l'ospedale e il tipo di parto.

```r
chisq.test(tabella)
```

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

Con un p-value di 0.58 non rifiutiamo l'ipotesi nulla di indipendenza, non c'è quindi motivo di pensare che ci sia un ospedale in cui vengano eseguiti significativamente più parti cesarei.

ANALISI MULTIDIMENSIONALE

1) Andiamo ora ad indagare le relazioni fra le variabili del nostro dataset. Verifichiamo prima se la nostra variabile risposta peso si distribuisce come una normale.

```r
skewness(Peso)
```

```
## [1] -0.6470308
```

```r
kurtosis(Peso)-3
```

```
## [1] 2.031532
```

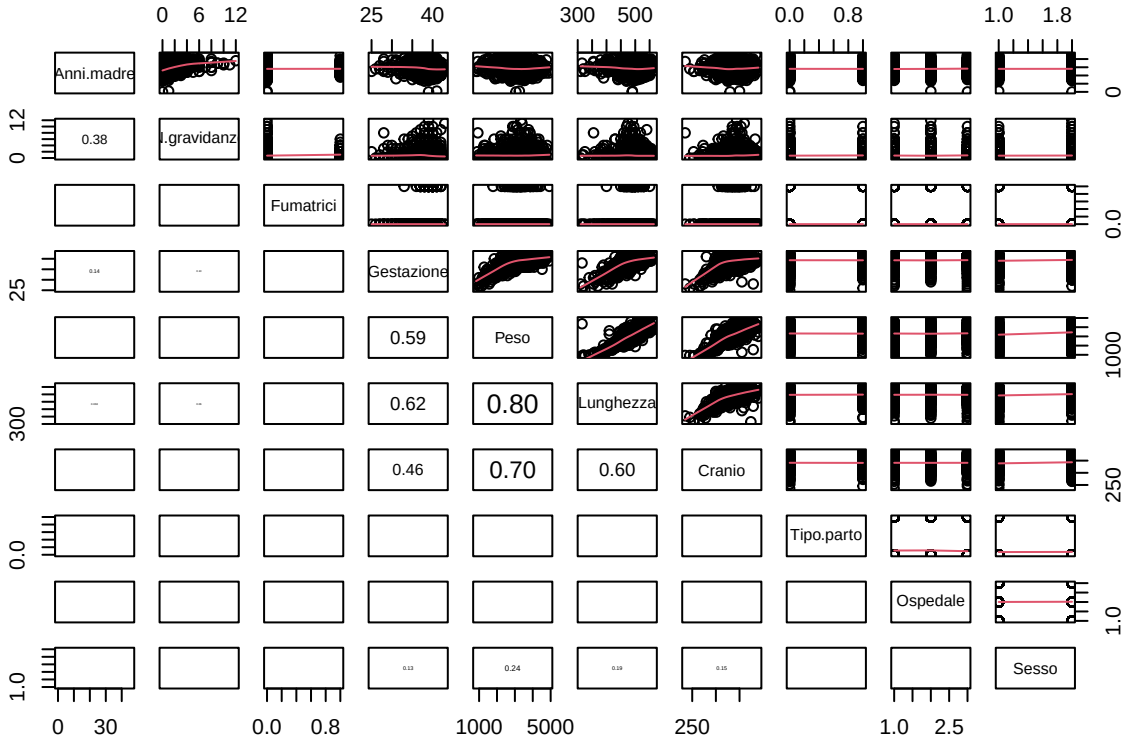
```r
shapiro.test(Peso)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Peso
## W = 0.97066, p-value < 2.2e-16
```

Lo Shapiro test ci restituisce un p-value molto basso, rifiutiamo quindi l'ipotesi nulla di normalità. Controllando i valori di asimmetria e di curtosi notiamo che ci troviamo di fronte ad una distribuzione leggermente asimmetrica negativa leptocurtica. Potremmo avere qualche problema in seguito con l'analisi dei residui.

Andiamo ora a vedere graficamente la correlazione fra le variabili:

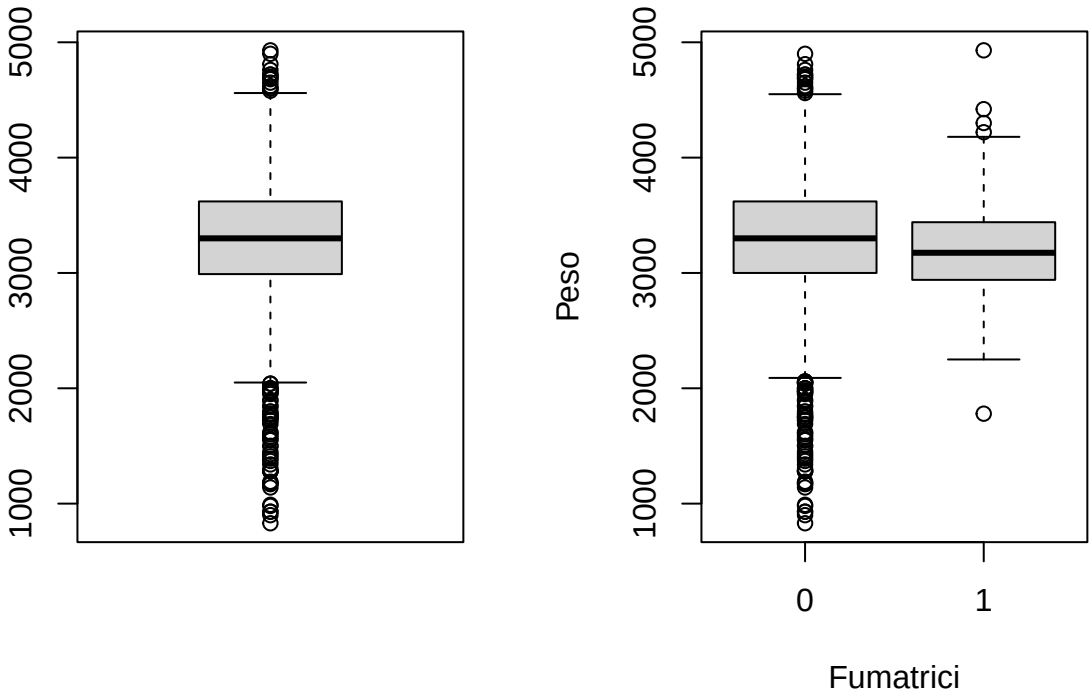
```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(dati, upper.panel = panel.smooth, lower.panel = panel.cor)
```



Dal grafico notiamo che le variabili più correlate alla variabile peso sono (in ordine crescente): Gestazione, Cranio e Lunghezza. Notiamo anche una correlazione della variabile Lunghezza con le variabili Cranio e Gestazione, cosa che potrebbe portare a problemi di multicollinearità ma che indagheremo in seguito.

Indaghiamo adesso la correlazione fra Peso e le due variabili qualitative Sesso e Fumatrici. Partiamo con quest'ultima.

```
par(mfrow=c(1,2))
boxplot(Peso)
boxplot(Peso~Fumatrici)
```



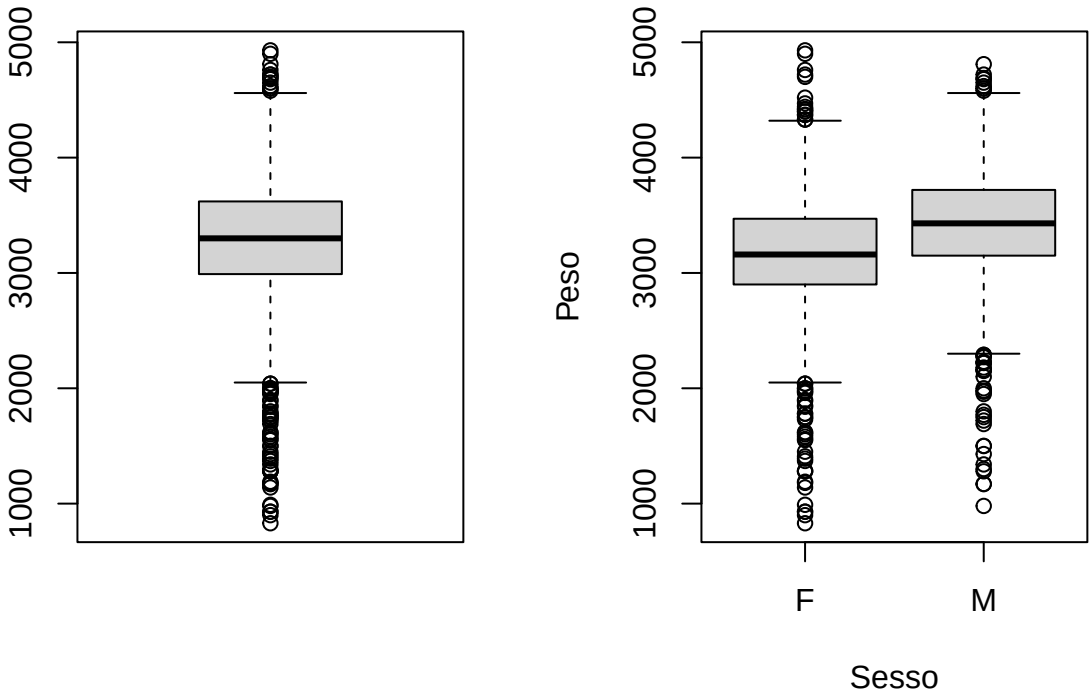

```
t.test(Peso~Fumatrici)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Fumatrici
## t = 1.034, df = 114.1, p-value = 0.3033
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -45.61354 145.22674
## sample estimates:
## mean in group 0 mean in group 1
##        3286.153        3236.346
```

Sia dai boxplot che dal test T vediamo che non ci sono differenze significative di peso fra bambini nati da mamme fumatrici e non. Risultato dato probabilmente dal fatto che le mamme fumatrici hanno smesso di fumare una volta saputo di essere incinte.

Passiamo adesso alla variabile Sesso:

```
par(mfrow=c(1,2))
boxplot(Peso)
boxplot(Peso~Sesso)
```




```
t.test(Peso~Sesso)
```

```
##
##  Welch Two Sample t-test
##
## data:  Peso by Sesso
## t = -12.106, df = 2490.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -287.1051 -207.0615
## sample estimates:
## mean in group F mean in group M
##        3161.132        3408.215
```

Come osservato già in precedenza nel punto 5 ci sono differenze significative di peso fra i due sessi.

Possiamo quindi dire che la variabile Peso è correlata alle variabili: Gestazione, Lunghezza, Cranio e Sesso.

2) Andiamo a creare il nostro primo modello con all'interno tutte le variabili del database.

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

```
##
```

```
## Call:
## lm(formula = Peso ~ ., data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1124.40  -181.66   -14.42   160.91  2611.89
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6708.9507   140.9304 -47.605  < 2e-16 ***
## Anni.madre       0.8921     1.1323   0.788   0.4308
## N.gravidanze    11.2665     4.6608   2.417   0.0157 *
## Fumatrici      -30.1631    27.5386  -1.095   0.2735
## Gestazione      32.5696     3.8187   8.529  < 2e-16 ***
## Lunghezza       10.2945     0.3007  34.236  < 2e-16 ***
## Cranio          10.4707     0.4260  24.578  < 2e-16 ***
## Tipo.parto     -29.5254    12.0844  -2.443   0.0146 *
## Ospedaleosp2   -11.2095    13.4379  -0.834   0.4043
## Ospedaleosp3    28.0958    13.4957   2.082   0.0375 *
## SessoM          77.5409    11.1776   6.937 5.08e-12 ***
## ---
## 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: 669.2 on 10 and 2489 DF,  p-value: < 2.2e-16
```

Concentriamoci sulle varibili che presentano più asterischi e quindi significatività maggiore. Per ogni settimana di gestazione si guadagnano 32.5g, per ogni mm di lunghezza 10g e per ogni mm di larghezza del cranio 10.5g. Tenendo fissate le altre variabili si rileva un peso medio di 77.5g in più rispetto alle femmine. Con un R-quadro aggiustato di quasi 0.73 possiamo considerare mod1 come un buon modello, ci sono però margini di miglioramento.

3) Perfezioniamo il modello attraverso la procedura Stepwise, in questo caso la applichiamo automaticamente attraverso la funzione stepAIC.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
stepwise.mod <- stepAIC(mod1,
                        direction = "both",
                        k=log(2500))
```

```
## Start:  AIC=28139.32
## Peso ~ Anni.madre + N.gravidanze + Fumatrici + Gestazione + Lunghezza +
##     Cranio + Tipo.parto + Ospedale + Sesso
```

```
##
##                Df Sum of Sq       RSS   AIC
## - Anni.madre    1     46578 186809099 28132
## - Fumatrici     1     90019 186852540 28133
## - Ospedale      2    685979 187448501 28133
## - N.gravidanze  1    438452 187200974 28137
## - Tipo.parto    1    447929 187210450 28138
## <none>                      186762521 28139
## - Sesso         1   3611021 190373542 28179
## - Gestazione    1   5458403 192220925 28204
## - Cranio        1  45326172 232088693 28675
## - Lunghezza     1  87951062 274713583 29096
##
## Step:  AIC=28132.12
## Peso ~ N.gravidanze + Fumatrici + Gestazione + Lunghezza + Cranio +
##     Tipo.parto + Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Fumatrici     1     90897 186899996 28126
## - Ospedale      2    692738 187501837 28126
## - Tipo.parto    1    448222 187257321 28130
## <none>                      186809099 28132
## - N.gravidanze  1    633756 187442855 28133
## + Anni.madre    1     46578 186762521 28139
## - Sesso         1   3618736 190427835 28172
## - Gestazione    1   5412879 192221978 28196
## - Cranio        1  45588236 232397335 28670
## - Lunghezza     1  87950050 274759149 29089
##
## Step:  AIC=28125.51
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Ospedale + Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Ospedale      2    701680 187601677 28119
## - Tipo.parto    1    440684 187340680 28124
## <none>                      186899996 28126
## - N.gravidanze  1    610840 187510837 28126
## + Fumatrici     1     90897 186809099 28132
## + Anni.madre    1     47456 186852540 28133
## - Sesso         1   3602797 190502794 28165
## - Gestazione    1   5346781 192246777 28188
## - Cranio        1  45632149 232532146 28664
## - Lunghezza     1  88355030 275255027 29086
##
## Step:  AIC=28119.23
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Tipo.parto +
##     Sesso
##
##                Df Sum of Sq       RSS   AIC
## - Tipo.parto    1    463870 188065546 28118
## <none>                      187601677 28119
## - N.gravidanze  1    651066 188252743 28120
## + Ospedale      2    701680 186899996 28126
```

```
## + Fumatrici     1     99840 187501837 28126
## + Anni.madre    1     54392 187547285 28126
## - Sesso         1   3649259 191250936 28160
## - Gestazione    1   5444109 193045786 28183
## - Cranio        1  45758101 233359778 28657
## - Lunghezza     1  88054432 275656108 29074
##
## Step:  AIC=28117.58
## Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
##
##                Df Sum of Sq       RSS   AIC
## <none>                      188065546 28118
## - N.gravidanze  1    623141 188688687 28118
## + Tipo.parto    1    463870 187601677 28119
## + Ospedale      2    724866 187340680 28124
## + Fumatrici     1     91892 187973654 28124
## + Anni.madre    1     54816 188010731 28125
## - Sesso         1   3655292 191720838 28158
## - Gestazione    1   5464853 193530399 28181
## - Cranio        1  46108583 234174130 28658
## - Lunghezza     1  87632762 275698308 29066
```

Vediamo che sono state eliminate le variabili meno significative: Anni.madre, Fumatrici, Tipo.parto, Ospedale.

Diamo uno sguardo più dettagliato:

```
summary(stepwise.mod)
```

```
##
## Call:
## lm(formula = Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio +
##     Sesso, data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1149.44  -180.81   -15.58   163.64  2639.72
##
## 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
```

I coefficienti non hanno subito cambiamenti significativi e R-quadro aggiustato è rimasto a 0.72

In un'ottica di ulteriore semplificazione proviamo a togliere la variabile meno significativa del modello, N.gravidanze.

```
mod2 <- update(stepwise.mod, ~. - N.gravidanze)
summary(mod2)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso,
##     data = dati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1138.2  -184.3   -17.6   163.3  2627.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6651.1188   135.5172 -49.080  < 2e-16 ***
## Gestazione     31.2737     3.7856   8.261 2.31e-16 ***
## Lunghezza      10.2054     0.3007  33.939  < 2e-16 ***
## Cranio         10.6704     0.4245  25.139  < 2e-16 ***
## SessoM         79.1049    11.2117   7.056 2.22e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275 on 2495 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7257
## F-statistic:  1654 on 4 and 2495 DF,  p-value: < 2.2e-16
```

Ancora una volta vediamo che i coefficienti non subiscono variazioni significative e che il nostro R-quadro aggiustato rimane a 0.72.

Andiamo ora a confrontare il modello stepwise con mod2. Partiamo con il test anova:

```
anova(stepwise.mod, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: Peso ~ N.gravidanze + Gestazione + Lunghezza + Cranio + Sesso
## Model 2: Peso ~ Gestazione + Lunghezza + Cranio + Sesso
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1   2494 188065546
## 2   2495 188688687 -1   -623141 8.2637 0.004079 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con un p-value di 0.004 non vediamo un aumento significativo di varianza spiegata quando si aggiunge la variabile N.gravidanze al modello.

Eseguiamo ora il BIC:

```
BIC(stepwise.mod, mod2)
```

```
##              df      BIC
## stepwise.mod  7 35220.10
## mod2          6 35220.54
```

Visto il valore leggermente inferiore del modello stepwise si potrebbe propendere per lui tuttavia, applicando il principio del rasoio di Occam, a parità di validità delle soluzioni sceglieremo quella più semplice. Sceglieremo quindi mod2 per il numero inferiore di regressori.

Infine calcoliamo i vif per verificare l'assenza di multicollinearità fra le variabili del nostro modello.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif(mod2)
```

```
## Gestazione  Lunghezza     Cranio      Sesso
##   1.653502   2.069517   1.606131   1.038813
```

Con dei vif inferiori a 5 possiamo dire che non c'è multicollinearità fra le variabili.

4) Verifichiamo se è possibile considerare interazioni fra i regressori, creiamo un terzo modello aggiungendo l'effetto di interazione fra gestazione, lunghezza e cranio

```
mod3 <- lm(Peso~Gestazione*Lunghezza*Cranio+Sesso)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione * Lunghezza * Cranio + Sesso)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1141.99  -181.89   -16.29   162.74  2582.65
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.625e+04  8.315e+03   4.360 1.36e-05 ***
## Gestazione                 -1.185e+03  2.359e+02  -5.025 5.40e-07 ***
## Lunghezza                  -7.288e+01  1.958e+01  -3.723 0.000201 ***
## Cranio                     -1.238e+02  2.740e+01  -4.517 6.56e-06 ***
## SessoM                      7.393e+01  1.116e+01   6.627 4.19e-11 ***
## Gestazione:Lunghezza        2.355e+00  5.276e-01   4.464 8.41e-06 ***
## Gestazione:Cranio           3.779e+00  7.540e-01   5.012 5.76e-07 ***
```

```
## Lunghezza:Cranio             2.595e-01  6.198e-02   4.187 2.93e-05 ***
## Gestazione:Lunghezza:Cranio -7.293e-03  1.642e-03  -4.441 9.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.3 on 2491 degrees of freedom
## Multiple R-squared:  0.7319, Adjusted R-squared:  0.7311
## F-statistic: 850.3 on 8 and 2491 DF,  p-value: < 2.2e-16
```

Nonostante il nostro R-quadro salga e la significativià sia alta per tutti i regressori, notiamo dei coefficienti preoccupanti, scartiamo quindi questo modello.

Proviamo adesso mettendo in interazione Gestazione e Lunghezza:

```
mod3 <- lm(Peso~Gestazione*Lunghezza+Cranio+Sesso)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione * Lunghezza + Cranio + Sesso)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1122.0  -181.5   -14.0   166.3  2640.2
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.030e+03  9.216e+02  -2.202  0.02774 *
## Gestazione           -9.318e+01  2.484e+01  -3.751  0.00018 ***
## Lunghezza             2.870e-02  2.030e+00   0.014  0.98872
## Cranio                1.089e+01  4.246e-01  25.651  < 2e-16 ***
## SessoM                7.353e+01  1.121e+01   6.559 6.57e-11 ***
## Gestazione:Lunghezza  2.688e-01  5.302e-02   5.069 4.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.7 on 2494 degrees of freedom
## Multiple R-squared:  0.7289, Adjusted R-squared:  0.7283
## F-statistic:  1341 on 5 and 2494 DF,  p-value: < 2.2e-16
```

In questo caso vediamo che la variabile lunghezza perde completamente di significatività e anche qui i coefficienti assumono valori strani, scartiamo anche questo caso.

Proviamo adesso con Lunghezza*Cranio:

```
mod3 <- lm(Peso~Gestazione+Lunghezza*Cranio+Sesso)
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza * Cranio + Sesso)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1139.05  -181.44   -15.46   163.32  2850.10
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.839e+03  1.019e+03  -1.804   0.0714 .
## Gestazione        3.692e+01  3.951e+00   9.344  < 2e-16 ***
## Lunghezza        -2.013e-01  2.205e+00  -0.091   0.9273
## Cranio           -4.409e+00  3.194e+00  -1.380   0.1676
## SessoM            7.449e+01  1.121e+01   6.648 3.64e-11 ***
## Lunghezza:Cranio  3.113e-02  6.537e-03   4.763 2.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.8 on 2494 degrees of freedom
## Multiple R-squared:  0.7286, Adjusted R-squared:  0.728
## F-statistic:  1339 on 5 and 2494 DF,  p-value: < 2.2e-16
```

Ancora peggio rispetto a prima, qui oltre ai coefficienti preoccupanti abbiamo anche la completa perdita di significatività di Lunghezza e Cranio.

Verifichiamo ora la presenza di effetti non lineari. Eseguiremo la stessa operazione su Gestazione, Lunghezza e Cranio.

```
mod3 <- update(mod2,~. + I(Gestazione^2))
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     I(Gestazione^2), data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1132.84  -181.45   -15.99   162.62  2649.78
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4643.6094   899.4989  -5.162 2.63e-07 ***
## Gestazione        -80.7957    49.7863  -1.623   0.1047
## Lunghezza          10.3087     0.3039  33.920  < 2e-16 ***
## Cranio             10.7663     0.4262  25.259  < 2e-16 ***
## SessoM             76.9359    11.2436   6.843 9.75e-12 ***
## I(Gestazione^2)     1.4960     0.6627   2.258   0.0241 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 274.8 on 2494 degrees of freedom
## Multiple R-squared:  0.7267, Adjusted R-squared:  0.7261
## F-statistic:  1326 on 5 and 2494 DF,  p-value: < 2.2e-16
```

Vediamo che il coefficiente di Gestazione viene compromesso e che Gestazione^2 ha significatività bassissima, scartiamo.

```
mod3 <- update(mod2,~. + I(Lunghezza^2))
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     I(Lunghezza^2), data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1156.75  -182.00   -12.52   166.67  1783.83
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    156.673587 724.783055   0.216    0.829
## Gestazione      41.174410   3.860512  10.666  < 2e-16 ***
## Lunghezza      -19.913124   3.165788  -6.290 3.73e-10 ***
## Cranio          10.795854   0.417182  25.878  < 2e-16 ***
## SessoM          71.369249  11.043854   6.462 1.24e-10 ***
## I(Lunghezza^2)   0.031241   0.003269   9.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270.2 on 2494 degrees of freedom
## Multiple R-squared:  0.7358, Adjusted R-squared:  0.7352
## F-statistic:  1389 on 5 and 2494 DF,  p-value: < 2.2e-16
```

Qui abbiamo un aumento di R-quadro e una significatività molto alta per Lunghezza^2, tuttavia il coefficiente di Lunghezza viene totalmente stravolto in negativo, scartiamo anche questo.

```
mod3 <- update(mod2,~. + I(Cranio^2))
summary(mod3)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso +
##     I(Cranio^2), data = dati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1127.21  -183.53   -14.88   163.98  2610.62
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.70852 1153.12221   0.067    0.947
## Gestazione    37.73144    3.91778   9.631  < 2e-16 ***
## Lunghezza     10.44511    0.30147  34.647  < 2e-16 ***
## Cranio       -31.41831    7.17690  -4.378 1.25e-05 ***
## SessoM        74.30205   11.16712   6.654 3.50e-11 ***
## I(Cranio^2)    0.06226    0.01060   5.875 4.80e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

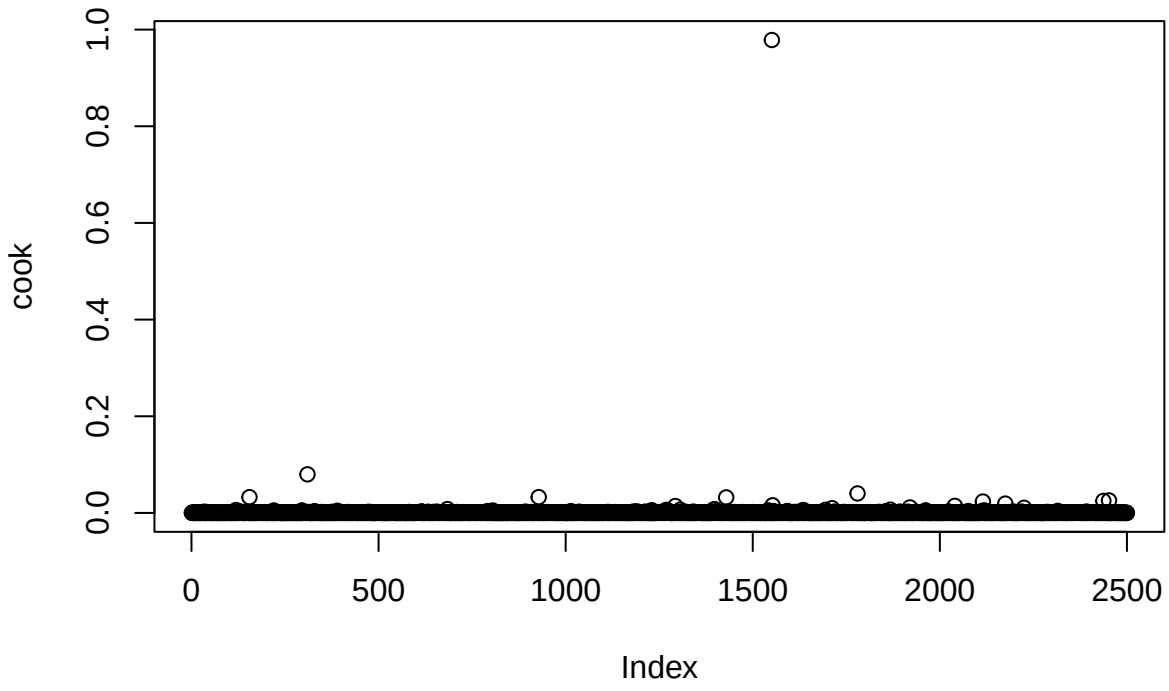
```
##
## Residual standard error: 273.2 on 2494 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7293
## F-statistic:  1347 on 5 and 2494 DF,  p-value: < 2.2e-16
```

In questo caso Cranioˆ2 ha una significatività molto alta ma i coefficienti di Cranio vengono totalmente stravolti, scartiamo anche quest'ultimo.

In conclusione quindi non è possibile considerare interazioni o effetti non lineari. Il nostro modello di riferimento rimane quindi mod2.

5) Procediamo ora con un analisi dei residui di mod2. Calcoliamo la distanza di Cook in modo da vedere se sono presenti valori che potrebbero influenzare il nostro modello.

```
cook <- cooks.distance(mod2)
plot(cook)
```




```
max(cook)
```

```
## [1] 0.9783883
```

Abbiamo un valore oltre la soglia di avvertimento di 0.5 e pericolosamente vicino alla soglia di pericolo 1. Andiamo quindi ad individuarlo per poi eliminarlo dal nostro dataset.

Calcoliamo gli outliers:

```r
outlierTest(mod2)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 1551 9.986149         4.7193e-23   1.1798e-19
## 155  4.951276         7.8654e-07   1.9663e-03
## 1306 4.781188         1.8440e-06   4.6100e-03
```

Incrociando il risultato ottenuto con il grafico precedente identifichiamo il nostro outlier problematico, l'osservazione 1551.

Andiamo a creare un nuovo dataset eliminando questa osservazione:

```r
dati.nuovi <- dati[-1551, ]
```

Ora che abbiamo eliminato il dato problematico creiamo un nuovo modello basandoci sui nuovi dati:

```r
mod4 <- lm(Peso ~Gestazione+Lunghezza+Cranio+Sesso, dati.nuovi)
summary(mod4)
```

```
##
## Call:
## lm(formula = Peso ~ Gestazione + Lunghezza + Cranio + Sesso,
##     data = dati.nuovi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1153.79  -181.99   -14.94   163.85  1391.36
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6651.7241   132.9131 -50.046  < 2e-16 ***
## Gestazione     28.4861     3.7233   7.651 2.83e-14 ***
## Lunghezza      10.8440     0.3018  35.935  < 2e-16 ***
## Cranio         10.0591     0.4208  23.906  < 2e-16 ***
## SessoM         79.3075    10.9963   7.212 7.27e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.7 on 2494 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7358
## F-statistic:  1740 on 4 and 2494 DF,  p-value: < 2.2e-16
```

Siamo passati da un R-quadro aggiustato di 0.7257 di mod2 a un 0.7358 di mod4, abbiamo guadagnato un po' più di accuratezza e siamo contenti.

Eseguiamo ora i test per verificare la presenza di omoschedasticità fra i resisui, la loro non correlazione e se si distribuiscono secondo una distribuzione normale.

Eseguiamo prima il test di omoschedasticità di Breusch-Pagan:

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(mod4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod4
## BP = 9.3632, df = 4, p-value = 0.05263
```

Non rifiutiamo l'ipotesi nulla di omoschedasticità, la varianza è costante.

Vediamo ora se i residui sono autocorrelati con il test Darwin-Watson:

```
dwtest(mod4)
```

```
##
##  Durbin-Watson test
##
## data:  mod4
## DW = 1.956, p-value = 0.1354
## alternative hypothesis: true autocorrelation is greater than 0
```

Non rifiutiamo l'ipotesi nulla di non autocorrelazione.

E infine chiudiamo con uno Shapiro test per verificare se i residui seguono una distribuzione normale:

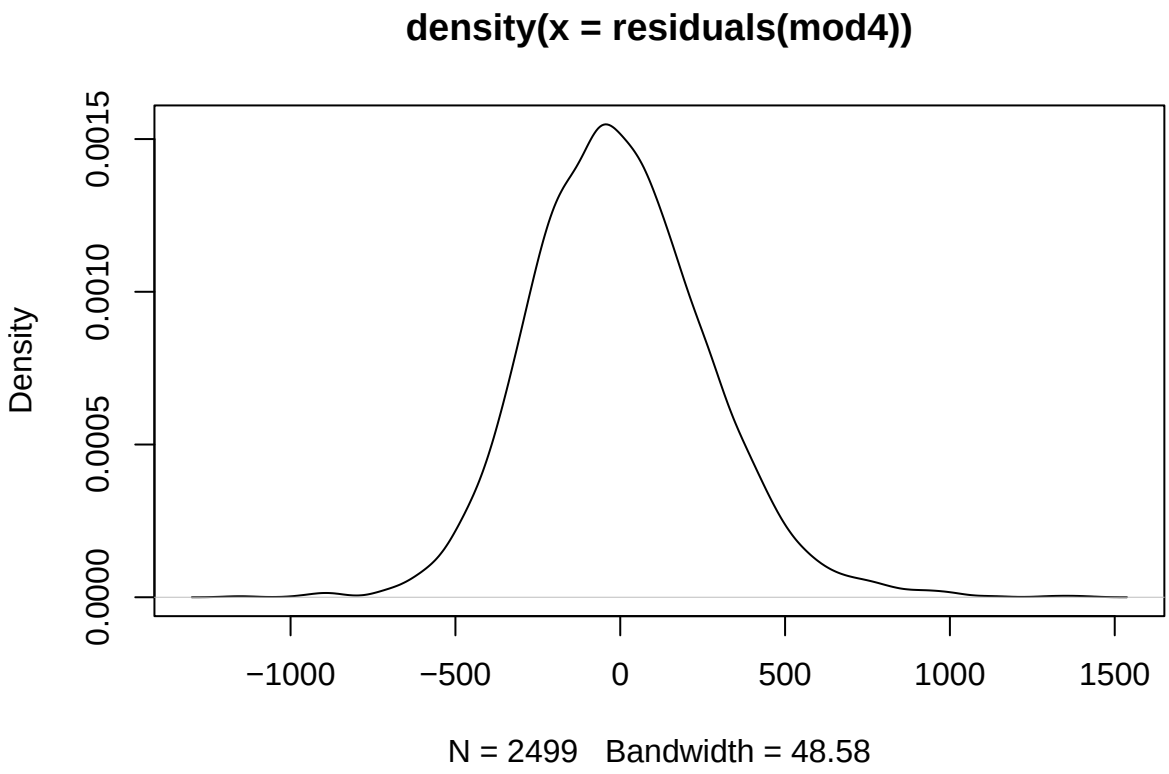
```
shapiro.test(residuals(mod4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod4)
## W = 0.98864, p-value = 3.278e-13
```

Rifiutiamo l'ipotesi nulla, i residui non sono distribuiti secondo una normale.

6) Con un R-quadro aggiustato di 0.73 ci assicura una buona accuratezza delle nostre previsioni. I test eseguiti sui residui dimostrano l'assenza di eteroschedasticità e di autocorrelazione e, anche se il nostro Shapiro test è fallito, vediamo che il grafico dei residui non si discosta di molto da una normale.

```
plot(density(residuals(mod4)))
```



Nel complesso quindi possiamo considerare il nostro mod4 come un buon modello per fare previsioni.

7) Mettiamolo quindi subito alla prova eseguiamo una previsione per calcolare il peso di una neonata, la mamma è alla terza gravidanza (ignoreremo questo dato), e partorirà alla 39esima settimana. Non abbiamo misure dell'ecografia e quindi useremo la media femminile di Lunghezza e Cranio.

```
media_lunghezza_f <- round(mean(Lunghezza[Sesso=="F"]))
media_cranio_f <- round(mean(Cranio[Sesso=="F"]))

nuovi.dati <- data.frame(
  Gestazione = 39,
  Lunghezza = media_lunghezza_f,
  Cranio = media_cranio_f,
  Sesso = "F"
)
predizioni <- predict(mod4, newdata = nuovi.dati)
predizioni
```

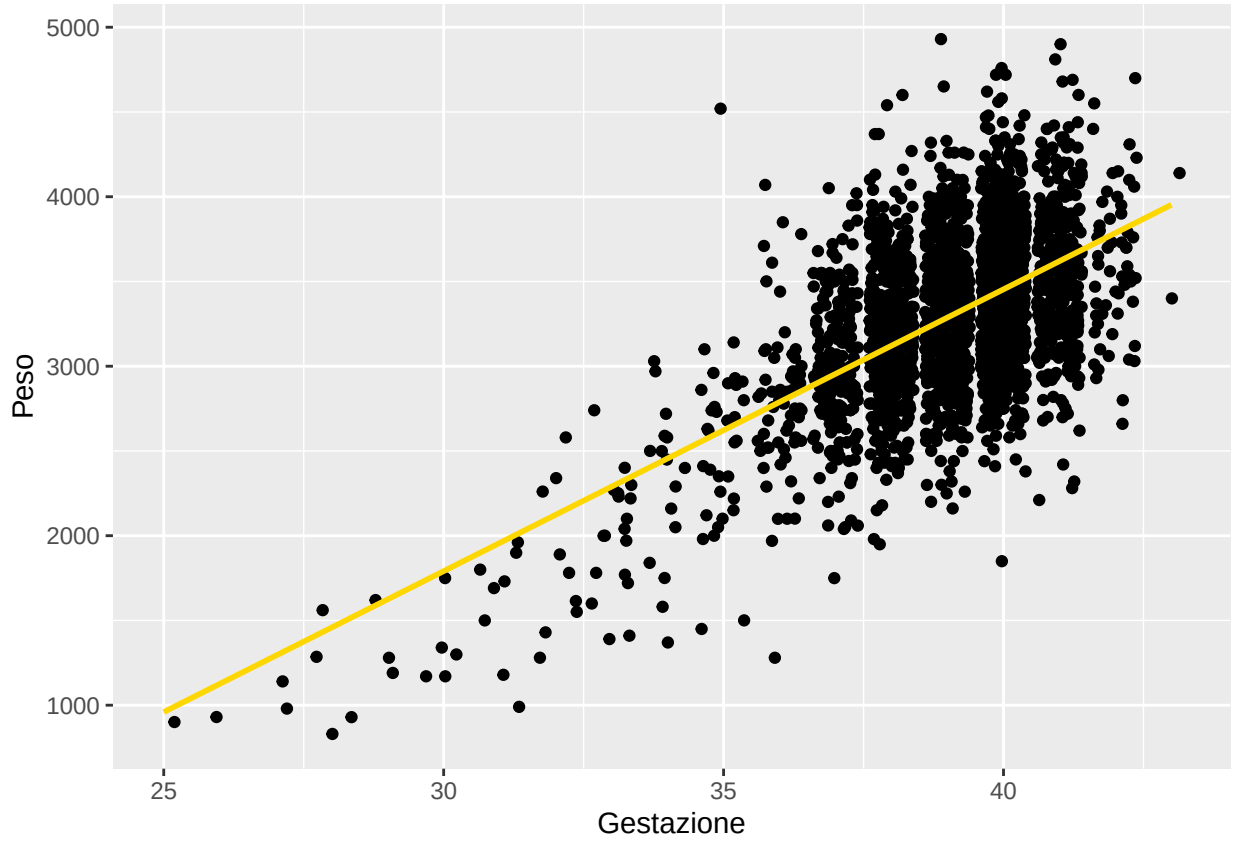
```
##        1
## 3172.779
```

Ecco fatto! La neonata dovrebbe pesare 3.172kg, dato che possiamo considerare corretto in quanto in linea con la media della popolazione.

8) Visualizziamo adesso il modello attraverso qualche rappresentazione grafica. Visualizziamo la correlazione fra le settimane di gestazione e il peso:

```
library(ggplot2)
ggplot(dati = dati.nuovi)+
  geom_point(aes(x=Gestazione,
                 y=Peso), position = "jitter")+
  geom_smooth(aes(x=Gestazione,
                 y=Peso), col = "gold", se = F, method = "lm")
```

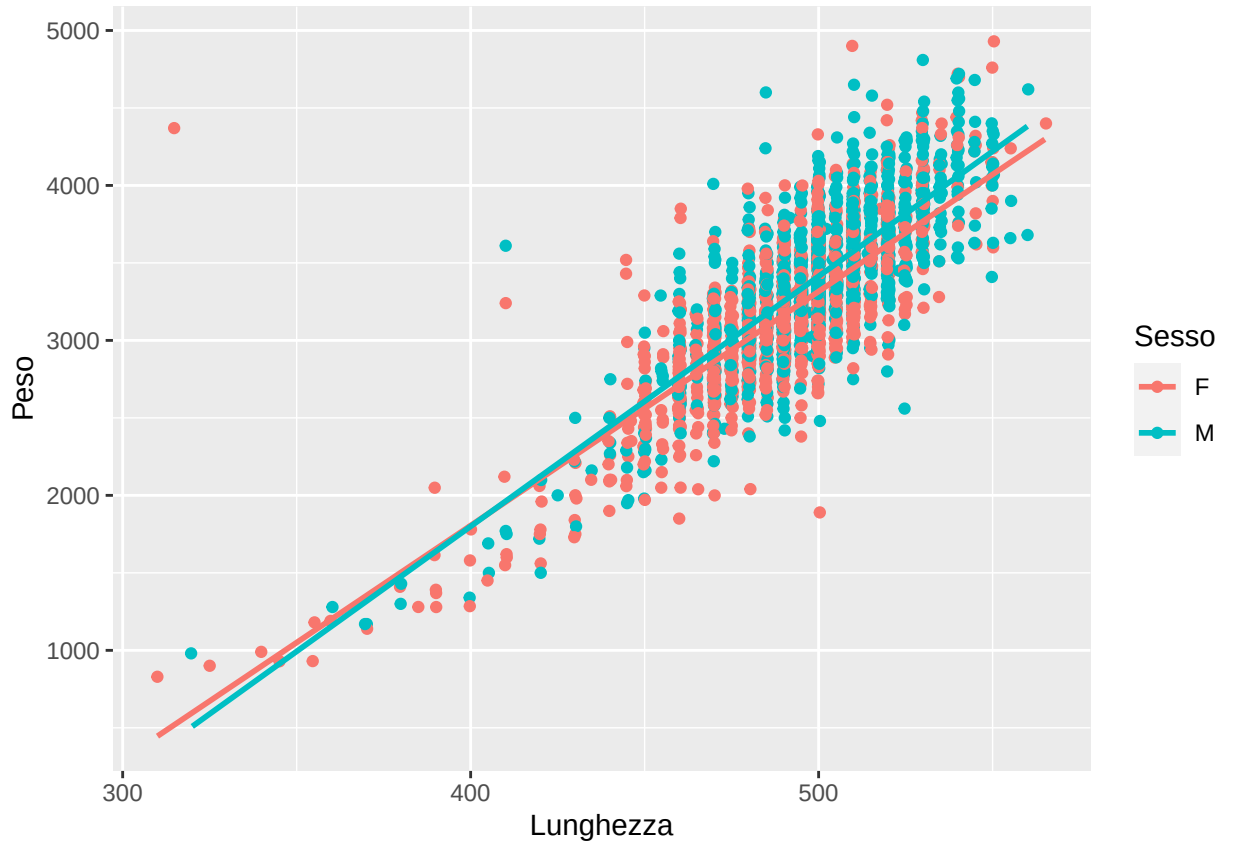
```
## `geom_smooth()` using formula = 'y ~ x'
```



Visualizziamo la correlazione fra Lunghezza e Peso:

```
ggplot(dati = dati.nuovi)+
  geom_point(aes(x=Lunghezza,
                 y=Peso,
                 col= Sesso), position = "jitter")+
  geom_smooth(aes(x=Lunghezza,
                 y=Peso,
                 col = Sesso), se = F, method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

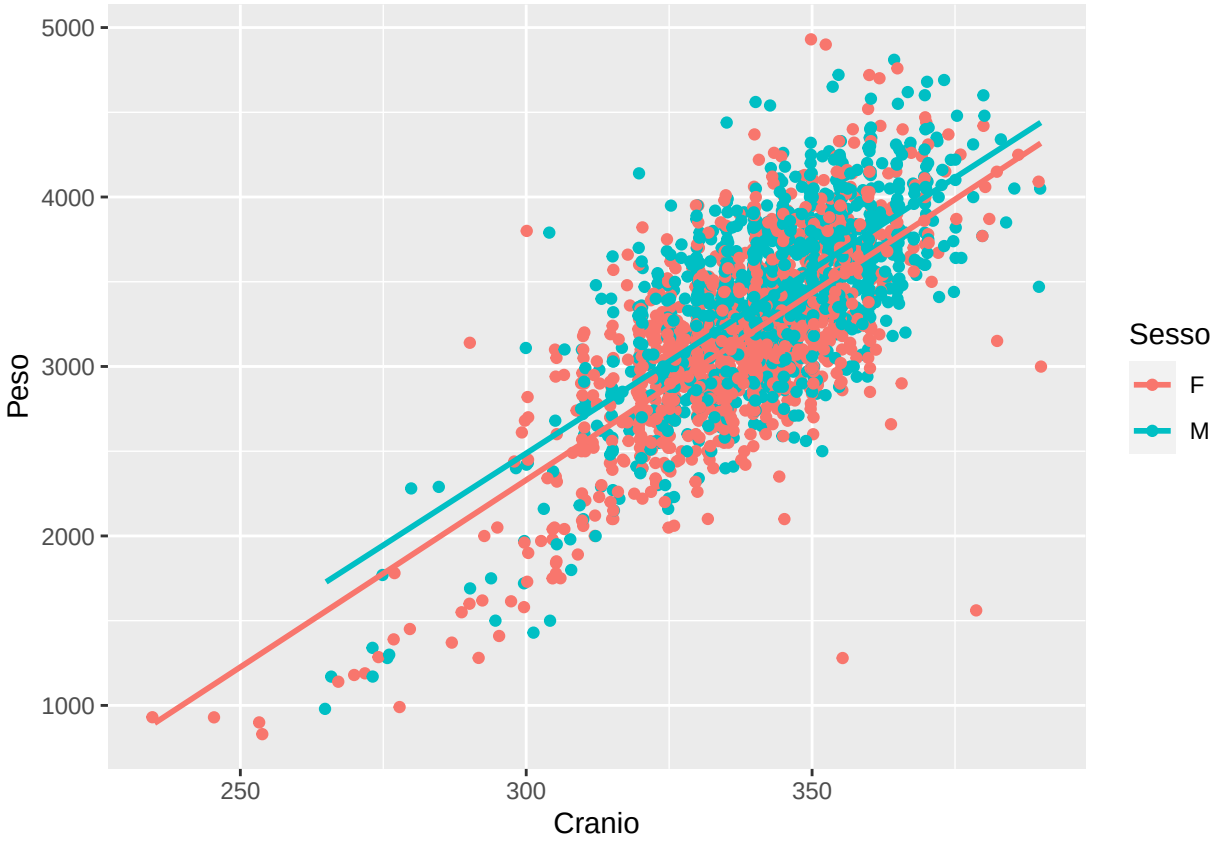


Vediamo che all'aumentare della lunghezza naturalmente aumenta anche il peso, con con differenze che si accentuano fra i due sessi verso la fine della gravidanza, con i maschi più lunghi e più pesanti.

Infine visualizziamo la correlazione fra Cranio e Peso:

```
ggplot(dati = dati.nuovi)+
  geom_point(aes(x=Cranio,
                 y=Peso,
                 col= Sesso), position = "jitter")+
  geom_smooth(aes(x=Cranio,
                 y=Peso,
                 col = Sesso), se = F, method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Come per la lunghezza abbiamo valori differenti fra i due sessi, con i maschi che presentano un diametro del cranio più grande e un conseguente peso maggiore.