

Analisi_finale

Introduzione

Di seguito riportiamo la lista delle librerie utilizzate nell'elaborazione del dataset.

```
library(lme4)
library(tidyverse)
library(RColorBrewer)
library(lmerTest)
library(gamlss)
library(ggcorrplot)
library(fitdistrplus)
```

Descrizione e caricamento del dataset

Il dataset RiceFarms è stato importato attraverso il seguente codice e inoltre vengono presentate le prime 5 righe del dataset:

```
load("DataRegression2025_unical.RData")
data=data.frame(RiceFarms)
# Trasformo la variabile in fattore in modo da avere una rappresentazione
# corretta del dataset
data$id = as.factor(data$id)
attach(data)
head(data)
```

```
##      id  size status varieties bimas seed urea phosphate pesticide pseed purea
## 1 101001 3.000  owner    mixed mixed  90  900      80      6000      80    75
## 2 101001 2.000  owner    trad mixed  40  600       0      3000      70    75
## 3 101001 1.000  owner    high mixed 100  700     150      5000     140    70
## 4 101001 2.000  owner    high mixed  60  600     100      5000      90    70
## 5 101001 3.572  share    high  no 105  400     400     10200     350    80
## 6 101001 3.572  share    high  no 105  400     400     10200     250    80
##  pphosph hiredlabor famlabor totlabor  wage goutput noutput price
## 1      75      2875      40      2915  68.49  7980  6800  60
## 2      75      2110      45      2155  60.09  4083  3500  60
## 3      70       980      95      1075  51.99  2650  2242  65
## 4      70      2081      10      2091  56.98  4500  3750  70
## 5      80      3889       1      3889 152.03 16300 13584 120
## 6      80      3519       1      3519 154.49 17424 14520 140
##      region
## 1 wargabinangun
## 2 wargabinangun
## 3 wargabinangun
## 4 wargabinangun
## 5 wargabinangun
## 6 wargabinangun
```

Prima dell'analisi è stata utilizzata la funzione `which(is.na(data))` per vedere se ci fossero eventuali valori mancanti. Il dataset non presenta valori mancanti. Di seguito possiamo osservare il tipo di variabili presenti nel dataset:

```
str(data)

## 'data.frame':    1026 obs. of  20 variables:
## $ id           : Factor w/ 171 levels "101001","101017",...: 1 1 1 1 1 1 2 2 2 ...
## $ size         : num  3 2 1 2 3.57 ...
## $ status       : Factor w/ 3 levels "owner","share",...: 1 1 1 1 2 2 3 3 3 1 ...
## $ varieties    : Factor w/ 3 levels "trad","high",...: 3 1 2 2 2 2 1 2 2 3 ...
## $ bimas        : Factor w/ 3 levels "no","yes","mixed": 3 3 3 3 1 1 3 3 3 1 ...
## $ seed         : int   90 40 100 60 105 105 50 20 15 7 ...
## $ urea         : int   900 600 700 600 400 400 120 100 150 50 ...
## $ phosphate    : int    80 0 150 100 400 400 0 0 50 0 ...
## $ pesticide    : int   6000 3000 5000 5000 10200 10200 0 0 900 0 ...
## $ pseed        : num    80 70 140 90 350 250 60 50 130 150 ...
## $ purea        : num    75 75 70 70 80 80 75 75 70 70 ...
## $ pphosph      : num    75 75 70 70 80 80 75 75 70 70 ...
## $ hiredlabor   : int   2875 2110 980 2081 3889 3519 670 805 380 40 ...
## $ famlabor     : int    40 45 95 10 1 1 140 50 80 69 ...
## $ totlabor     : int   2915 2155 1075 2091 3889 3519 810 855 460 109 ...
## $ wage         : num    68.5 60.1 52 57 152 ...
## $ goutput      : int   7980 4083 2650 4500 16300 17424 3840 2800 950 240 ...
## $ noutput      : int   6800 3500 2242 3750 13584 14520 3200 2400 800 200 ...
## $ price        : num    60 60 65 70 120 140 60 50 62 60 ...
## $ region       : Factor w/ 6 levels "wargabinangun",...: 1 1 1 1 1 1 1 1 1 1 ...
```

Descrizioni variabili

Vediamo una breve descrizione delle variabili presenti nel dataset: - id: identificativo univoco dell'azienda agricola.

- CARATTERISTICHE DEL TERRENO: SUDDIVISE IN ETTARI COLTIVATI E PROPRIETÀ DEL TERRENO
 - size: area totale coltivata a riso (in ettari).
 - status: stato della terra coltivata, che può essere:
 - * owner: agricoltori proprietari o affittuari (non mezzadri).
 - * share: mezzadri.
 - * mixed: combinazione delle due categorie precedenti.
 - varieties: tipo di varietà di riso coltivate:
 - * trad: varietà tradizionali.
 - * high: varietà ad alta resa.
 - * mixed: combinazione delle due varietà.
 - bimas: partecipazione al programma di intensificazione BIMAS:
 - * no: non partecipante.
 - * yes: partecipante.
 - * mixed: solo una parte del terreno è registrata nel programma.
- FATTORI DI INPUT PRODUTTIVI: COSTO DELLE MATERIE PRIME E TIPOLOGIA
 - seed: quantità di semi utilizzati (kg).
 - urea: quantità di urea utilizzata (kg).
 - phosphate: quantità di fosfato utilizzata (kg).
 - pesticide: costo dei pesticidi (in Rupiah).
 - pseed: prezzo del seme (in Rupiah per kg).
 - purea: prezzo dell'urea (in Rupiah per kg).
 - pphosph: prezzo del fosfato (in Rupiah per kg).

- INPUT: COSTI DEL PERSONALE E ORE DI LAVORO
 - hiredlabor: ore di lavoro salariato.
 - famlabor: ore di lavoro familiare.
 - totlabor: totale ore di lavoro (escludendo il raccolto).
 - wage: salario della manodopera (in Rupiah per ora).
- PRODUZIONE LORDA E NETTA
 - goutput: produzione lorda di riso (kg).
 - noutput: produzione netta di riso, calcolata sottraendo il costo del raccolto dalla produzione lorda.
 - price: prezzo del riso grezzo (in Rupiah per kg).
- AREE GEOGRAFICHE IN CUI OPERANO LE AZIENDE:
 - region: area geografica di appartenenza dell'azienda agricola, tra:
 - * wargabinangun
 - * langan
 - * gunungwangi
 - * malausma
 - * sukaambit
 - * ciwangi

Analisi preliminare

In questo paragrafo analizziamo le variabili presenti nel dataset. Questa analisi ci permette di comprendere meglio la distribuzione dei dati e di valutare il loro andamento rispetto alla variabile target (price), ovvero il prezzo, fondamentale per la costruzione del nostro modello.

Attraverso la funzione summary possiamo osservare che il dataset è di tipo multilivello. In quanto, i dati relativi a ciascuna azienda sono stati osservati su diversi cicli di produzione. Dalla variabile price possiamo osservare come questi presentino un'asimmetria positiva con una coda molto lunga, lasciando intendere una forte variabilità della distribuzione.

```
summary(data)
```

```
##           id           size           status  varieties           bimas
## 101001 :    6   Min.      :0.0100  owner:736   trad :682   no   :779
## 101017 :    6   1st Qu.:0.1430  share: 79   high :294   yes  : 85
## 101026 :    6   Median :0.2860  mixed:211  mixed: 50  mixed:162
## 101035 :    6   Mean    :0.4316
## 101056 :    6   3rd Qu.:0.5000
## 101057 :    6   Max.     :5.3220
## (Other):990
##           seed           urea           phosphate           pesticide
## Min.      :    1.00   Min.      :    1.00   Min.      :    0.00   Min.      :    0
## 1st Qu.:    5.00   1st Qu.:   25.00   1st Qu.:    8.00   1st Qu.:    0
## Median :   10.00   Median :   60.00   Median :   20.00   Median :    0
## Mean     :   18.21   Mean     :   95.44   Mean     :   33.73   Mean     :   595
## 3rd Qu.:   20.00   3rd Qu.:  100.00   3rd Qu.:   50.00   3rd Qu.:   265
## Max.     : 1250.00   Max.     : 1250.00   Max.     :   700.00   Max.     : 62600
##
##           pseed           purea           pphosph           hiredlabor
## Min.      :   40.0   Min.      :   50.00   Min.      :   60.00   Min.      :    1
## 1st Qu.:   70.0   1st Qu.:   70.00   1st Qu.:   70.00   1st Qu.:   36
## Median :   81.0   Median :   80.00   Median :   80.00   Median :   112
## Mean     :  112.1   Mean     :   78.98   Mean     :   79.57   Mean     :   237
## 3rd Qu.:  150.0   3rd Qu.:   85.00   3rd Qu.:   85.00   3rd Qu.:   260
## Max.     :  375.0   Max.     :  100.00   Max.     :  120.00   Max.     :  4536
##
```

```
##      famlabor      totlabor      wage      goutput
## Min.   :   1.0   Min.   : 17.0   Min.   : 30.00   Min.   : 42.0
## 1st Qu.: 69.0   1st Qu.: 144.0   1st Qu.: 49.38   1st Qu.: 420.0
## Median : 111.0   Median : 252.0   Median : 57.14   Median : 886.5
## Mean   : 151.5   Mean   : 388.4   Mean   : 80.42   Mean   : 1405.2
## 3rd Qu.: 185.0   3rd Qu.: 435.0   3rd Qu.:128.75   3rd Qu.: 1606.0
## Max.   :1526.0   Max.   :4774.0   Max.   :175.35   Max.   :20960.0
##
##      noutput      price      region
## Min.   : 42   Min.   : 50.00   wargabinangun:114
## 1st Qu.: 380   1st Qu.: 60.50   langan         :144
## Median : 800   Median : 75.00   gunungwangi    :222
## Mean   : 1241   Mean   : 90.96   malausma       :198
## 3rd Qu.: 1444   3rd Qu.:120.00   sukaambit      :132
## Max.   :17610   Max.   :190.00   ciwangi        :216
##
```

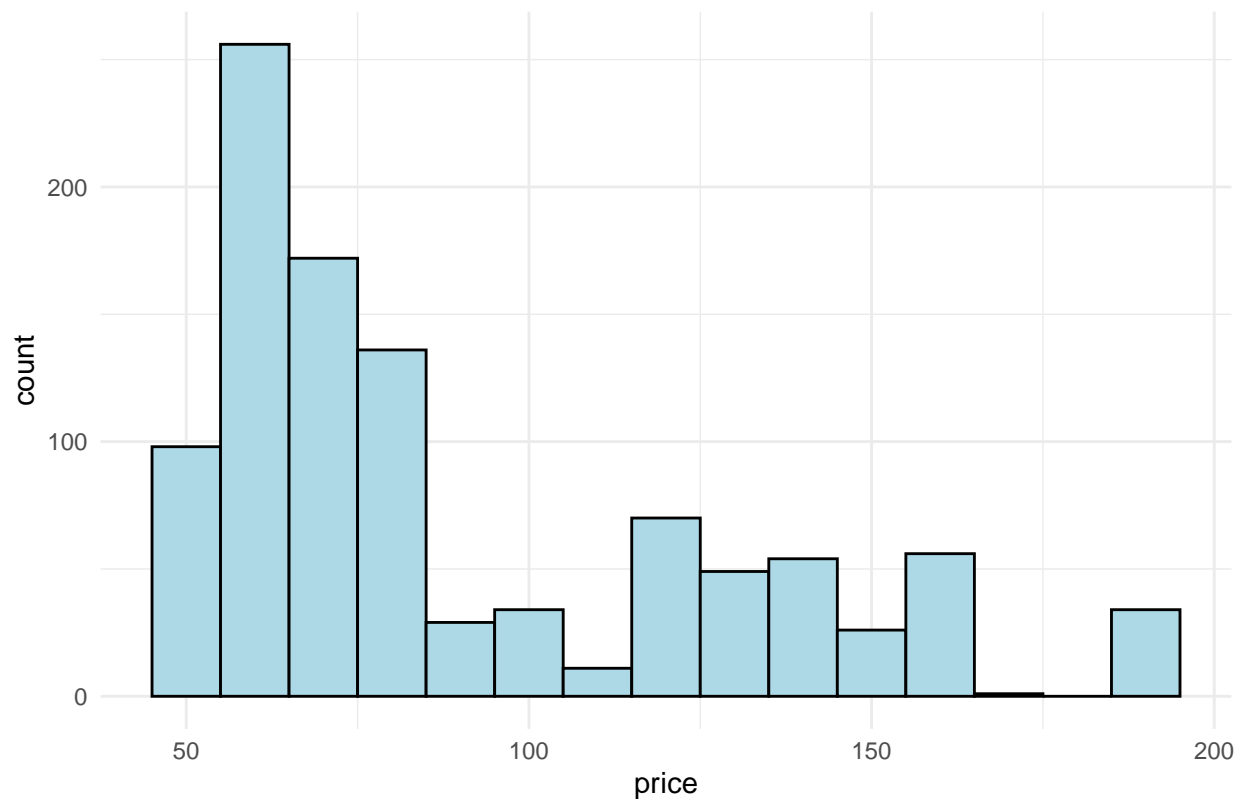
```
print(paste('Deviazione standard di price:', sd(price)))
```

```
## [1] "Deviazione standard di price: 37.4950096631271"
```

Nel seguente grafico è riportato l'istogramma della variabile price. Questo ci suggerisce il tipo distribuzione che potrebbe assumere la variabile prezzo. Tuttavia, questa non sembra ben definita dato che i valori sembrano concentrarsi intorno a due valori distinti di prezzo.

```
ggplot(data, aes(x = price)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribuzione dei prezzi")
```

Distribuzione dei prezzi

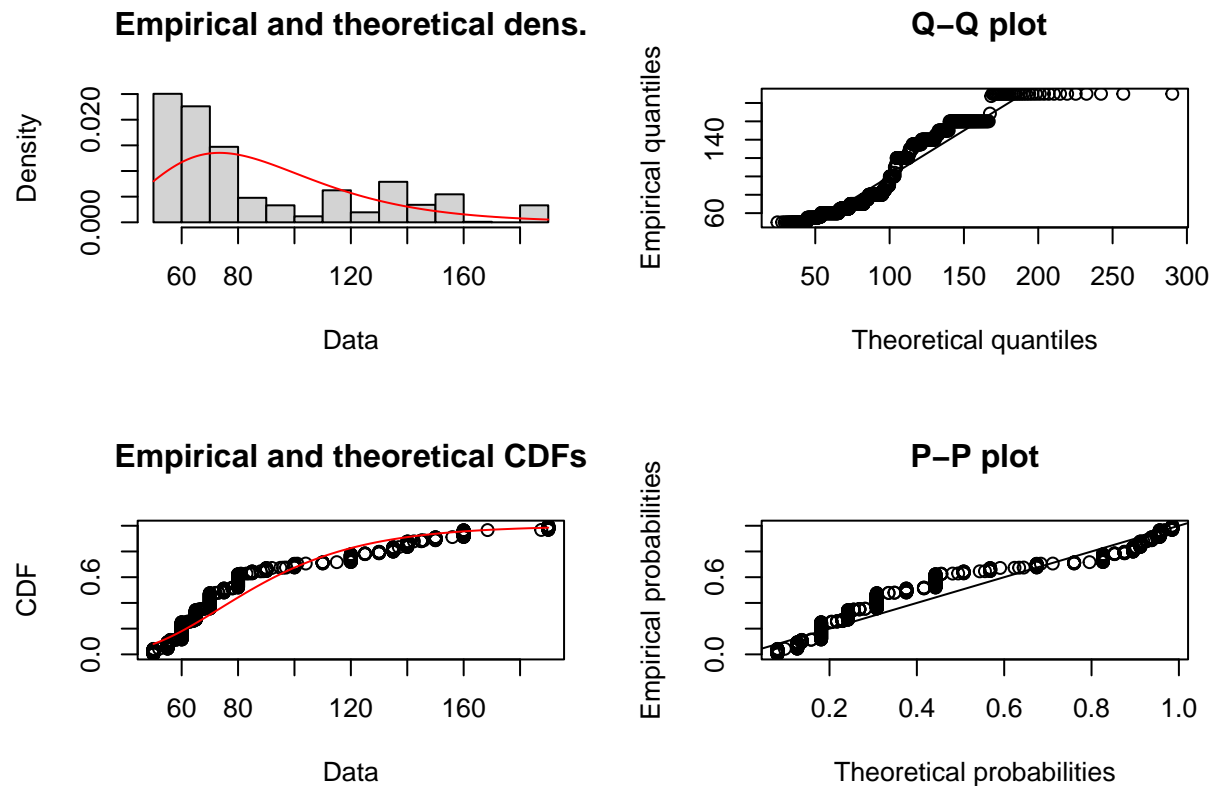


```
# Fitting diretto alla lognormale
fit <- fitdlist(price, "lnorm")
#fit <- fitdlist(price*goutput, "gamma")
```

```
# Mostra i parametri
summary(fit)
```

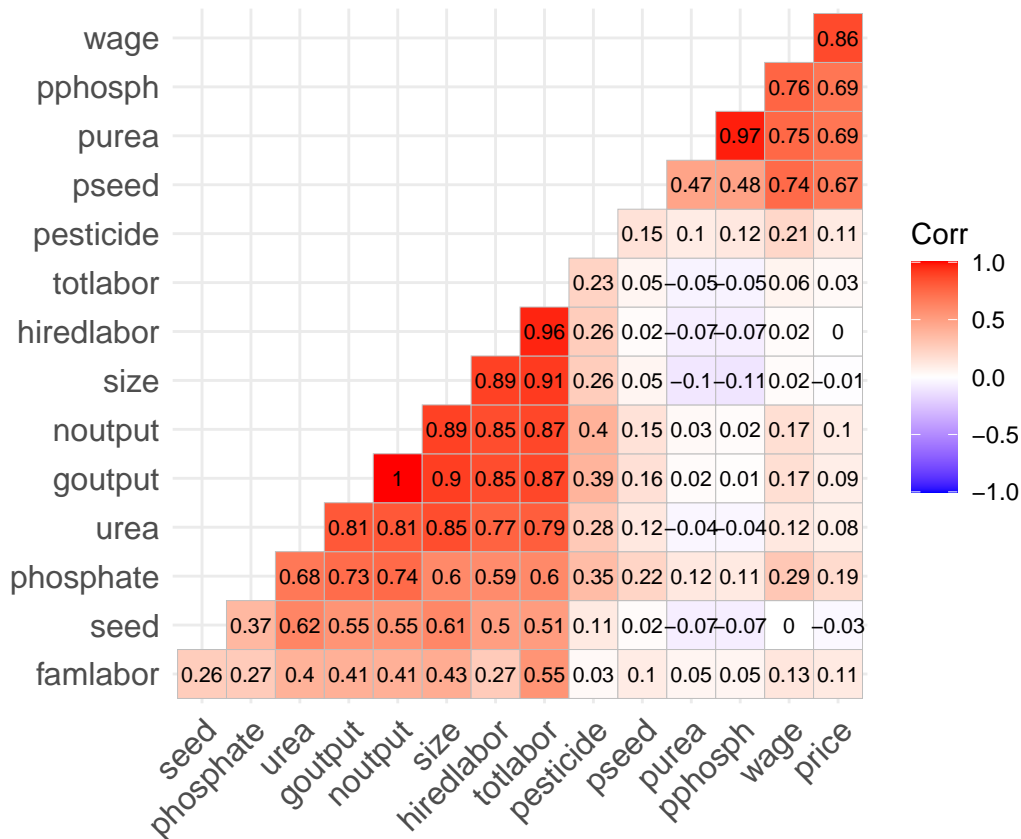
```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## meanlog 4.435941 0.011686625
## sdlog   0.374337 0.008263427
## Loglikelihood: -4998.96   AIC:  10001.92   BIC:  10011.79
## Correlation matrix:
##      meanlog sdlog
## meanlog      1      0
## sdlog        0      1
```

```
plot(fit)
```



Riportiamo di seguito la matrice di correlazione fra le variabili contenute nel dataset. Essendo presenti numerose colonne, questa matrice offre una visione chiara delle potenziali relazioni esistenti tra di esse. Si può notare come molte variabili siano fortemente correlate tra loro, e non solo con la variabile price. Gran parte di queste correlazioni risultano di facile interpretazione, come nel caso delle variabili goutput e noutput, che indicano rispettivamente la produzione lorda di riso in kg e la produzione netta, calcolata sottraendo il costo del raccolto dalla produzione lorda.

```
matrix_corrplot = round(cor(select_if(data, is.numeric), method="pearson"),4)
ggcorrplot(matrix_corrplot, hc.order=T, type="lower", lab=T, lab_size = 2.7)
```



Nel grafico riportiamo il valore della variabile price (sull'asse delle ordinate) in funzione del salario orario della manodopera (sull'asse delle ascisse), colorando i punti in base alla regione di appartenenza. L'obiettivo era indagare se esistesse una correlazione tra il prezzo e il salario dei lavoratori, variabile al variare delle regioni. Dal grafico emerge la presenza di due gruppi distinti: in uno sembrano concentrarsi salari e prezzi elevati, mentre nell'altro salari e prezzi risultano nettamente inferiori. È importante precisare che, dal dataset non si riesce a motivare la presenza di questi due gruppi così distinti.

```
ggplot(data, aes(x = wage, y = price, color = region)) +
  geom_point() +
  labs(
    x = "Wage",
    y = "Price",
    title = "Scatter Plot: Wage vs Price"
  ) +
  theme_minimal()
```

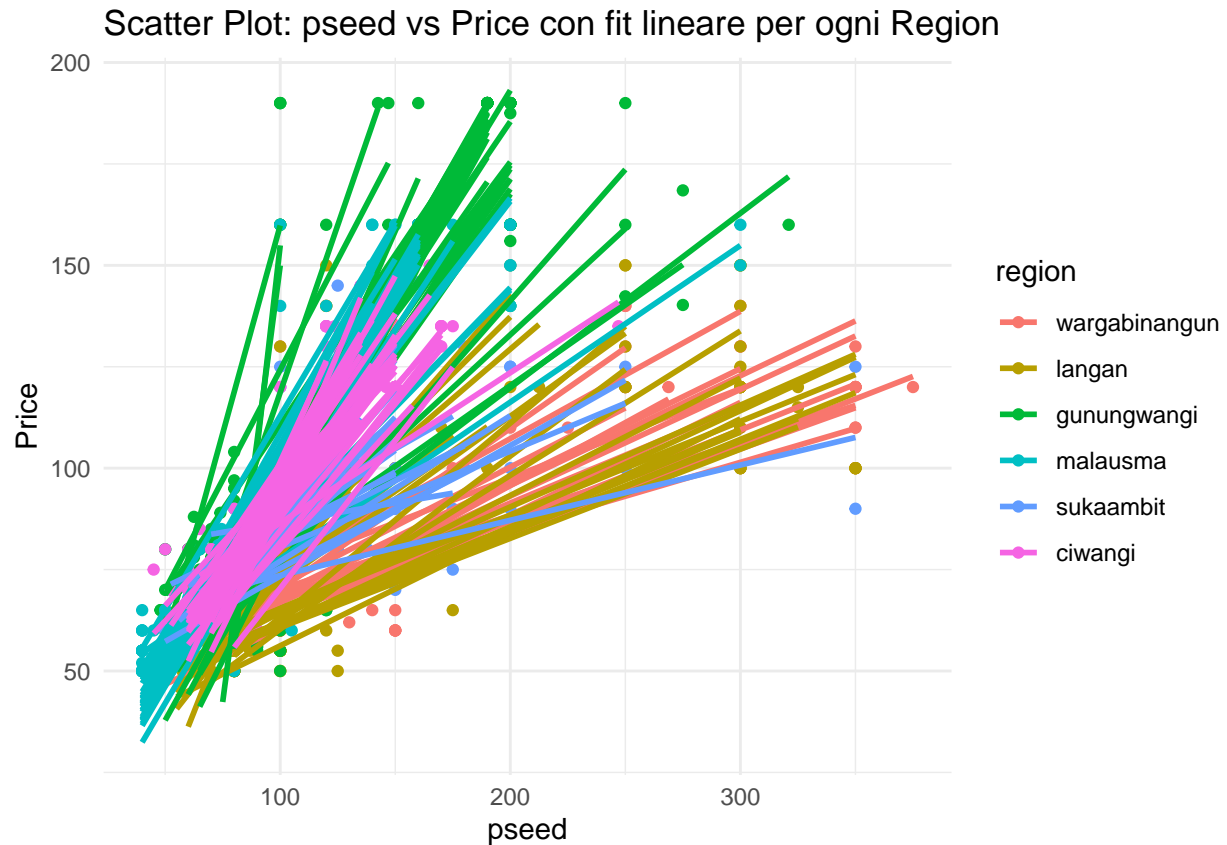
Scatter Plot: Wage vs Price



Nel grafico osserviamo la relazione tra il prezzo del riso (asse delle ordinate) e il prezzo dei semi (asse delle ascisse). I dati sono stati raggruppati per id e suddivisi per regione, evidenziati da colori differenti in base alla regione, e per ciascun gruppo è stata tracciata una retta di regressione lineare. Da tale analisi emerge che, per ogni variazione unitaria del prezzo dei semi, il corrispondente cambiamento nel prezzo del riso varia a seconda della regione di appartenenza.

```
ggplot(data, aes(x = pseed, y = price, group = id, color = region)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "pseed",
    y = "Price",
    title = "Scatter Plot: pseed vs Price con fit lineare per ogni Region"
  ) +
  theme_minimal()
```

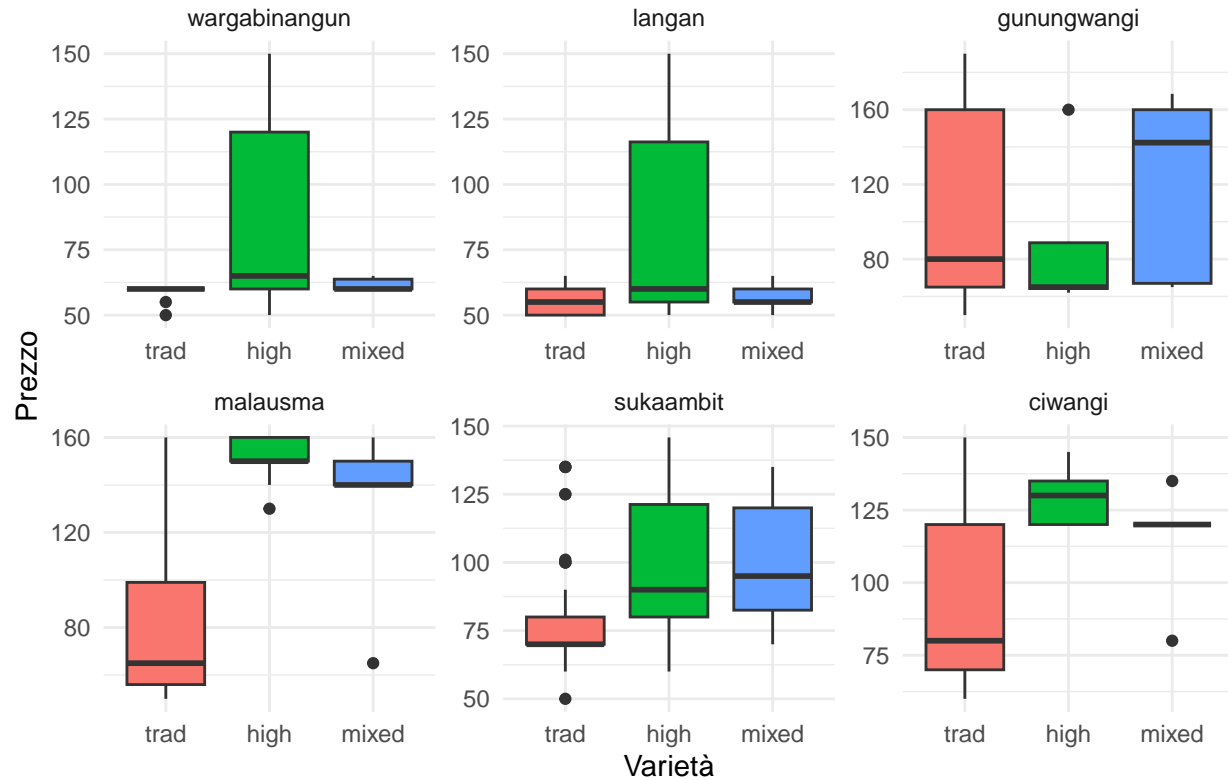
```
## `geom_smooth()` using formula = 'y ~ x'
```

Nei grafici seguenti analizziamo se, nelle diverse regioni, vengano coltivate le stesse varietà di riso e in quale misura. Questa analisi risulta particolarmente interessante poiché varietà differenti di riso sono associate a prezzi diversi. Dal grafico si evince che vi sia molta differenza fra le regioni in termini di varietà di riso coltivate e di prezzo a cui queste vengono vendute.

```
ggplot(data, aes(x = varieties, y = price, fill = varieties)) +
  geom_boxplot() +
  # Facet per regione, scales="free" permette ad ogni grafico
  # di avere le proprie scale
  facet_wrap(~ region, scales = "free") +
  labs(title = "Boxplot dei Prezzi per Varietà per Regione",
        x = "Varietà", y = "Prezzo") +
  theme_minimal() +
  theme(legend.position = "none")
```

Boxplot dei Prezzi per Varietà per Regione

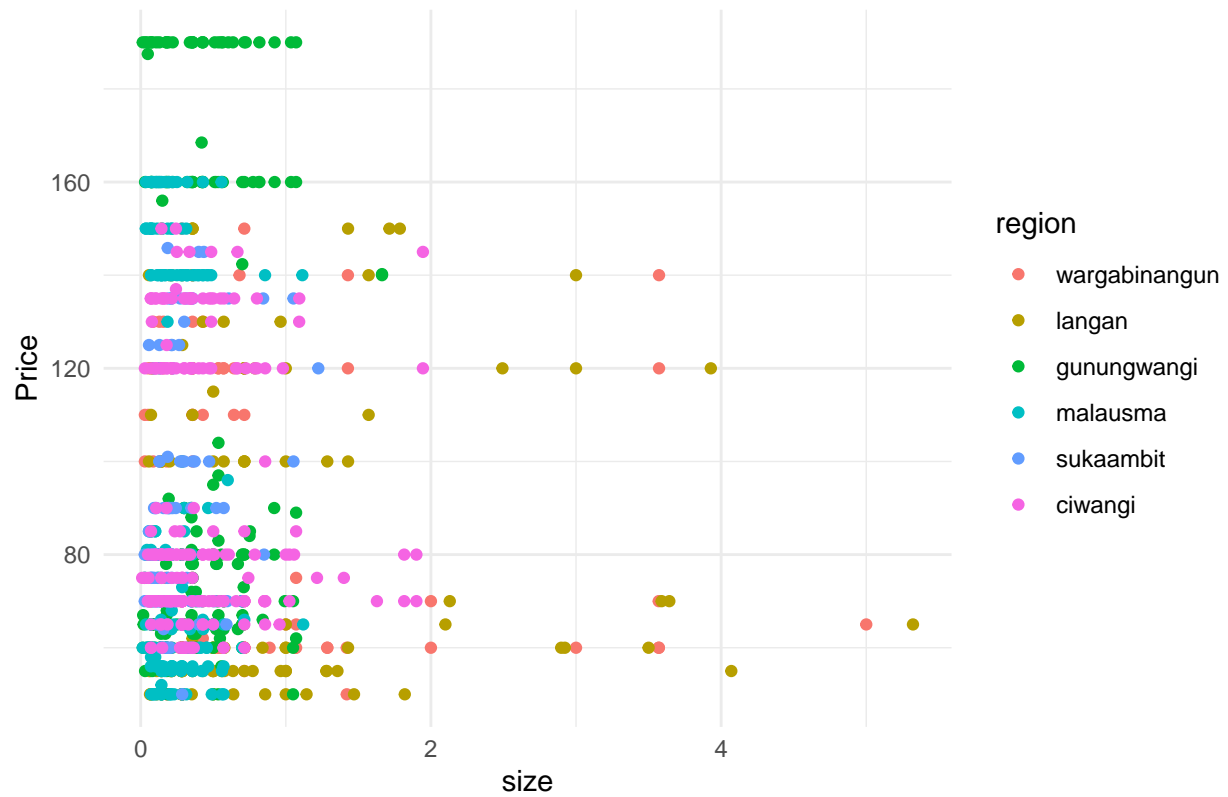


```
# Ripristino il layout di default
par(mfrow = c(1, 1))
```

In questi due grafici si evince l'assenza di una correlazione sia tra la quantità di semi utilizzata e il prezzo del riso, sia tra la dimensione del campo di coltivazione e il prezzo del riso.

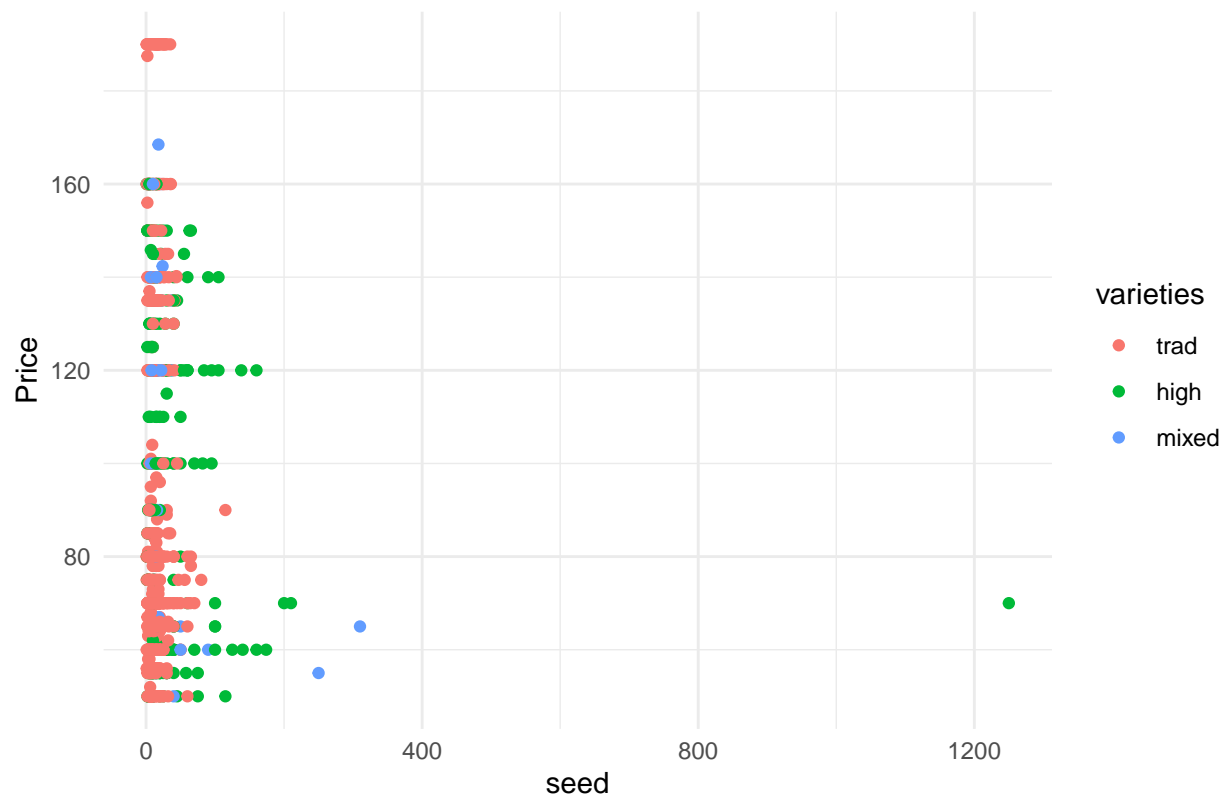
```
ggplot(data, aes(x = size, y = price, color = region)) +
  geom_point() +
  labs(
    x = "size",
    y = "Price",
    title = "Scatter Plot: size vs Price con fit lineare per ogni Region"
  ) +
  theme_minimal()
```

Scatter Plot: size vs Price con fit lineare per ogni Region



```
ggplot(data, aes(x = seed, y = price, color = varieties)) +
  geom_point() +
  labs(
    x = "seed",
    y = "Price",
    title = "Scatter Plot: seed vs Price con fit lineare per ogni Region"
  ) +
  theme_minimal()
```

Scatter Plot: seed vs Price con fit lineare per ogni Region



Analisi

Effetti casuali sulla variabile ID

Terminata l'esplorazione del dataset, ci siamo concentrati sul capire quali fossero i fattori più significativi nella determinazione del prezzo del riso per le diverse aziende. In particolare si è utilizzato un modello di regressione lineare a effetti casuali. In questo modello l'attribuzione degli effetti casuali è stata legata alle diverse aziende distinte per ID. Inoltre, per semplicità inizialmente si sono utilizzate solo alcune variabili quali: pseed, purea e wage.

```
model.1<-lmer(price~1 + pseed + purea + wage + (1|id), data=data)
summary(model.1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: price ~ 1 + pseed + purea + wage + (1 | id)
## Data: data
##
## REML criterion at convergence: 8909.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.9023 -0.5226 -0.0273  0.5245  3.3565
##
## Random effects:
## Groups   Name                Variance Std.Dev.
```

```
## id      (Intercept)  57.39    7.575
## Residual          301.39   17.361
## Number of obs: 1026, groups: id, 171
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept) -11.65000    6.90401 1000.20917  -1.687   0.0918 .
## pseed        0.08490    0.01386 1021.94005   6.128 1.27e-09 ***
## purea        0.58837    0.10097  993.99271   5.827 7.62e-09 ***
## wage         0.57977    0.02708  989.06337  21.410 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) pseed  purea
## pseed -0.197
## purea -0.981  0.160
## wage  0.605 -0.674 -0.666
```

Si può notare che nel primo modello le variabili inserite risultino significative tranne per l'intercetta. In quanto sia un prezzo negativo non risulta ragionevole, sia il suo valore di p-value risulta troppo elevato.

Nel blocco di codice seguente sono stati implementati due modelli, entrambi senza intercetta, in linea con le considerazioni precedenti. In particolare, sono state aggiunte ulteriori variabili, tra cui la varieties. Poiché quest'ultima è di tipo categorico, è stato eseguito un test AIC per valutare se la sua inclusione migliorasse il modello o meno. Il test ha evidenziato un miglioramento con la presenza della variabile. Tuttavia, nel summary si osserva che il p-value associato agli effetti casuali è molto elevato, indicando che tali effetti non sono statisticamente significativi e dunque da escludere.

```
print("-----Model 2-----")

## [1] "-----Model 2-----"

model.2<-lmer(price~0 + pseed + purea + wage +varieties + bimas + (1|id), data=data)
model.2B<-lmer(price~0 + pseed + purea + wage + bimas + (1|id), data=data)
AIC(model.2, model.2B)

##              df      AIC
## model.2    10 8824.764
## model.2B    8 8897.666

summary(model.2)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: price ~ 0 + pseed + purea + wage + varieties + bimas + (1 | id)
##      Data: data
##
## REML criterion at convergence: 8804.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.9714 -0.6178 -0.0057  0.5293  3.3002
##
## Random effects:
##  Groups   Name      Variance Std.Dev.
## id      (Intercept)  13.18     3.63
```

```
## Residual          301.75   17.37
## Number of obs: 1026, groups: id, 171
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## pseed          0.10446   0.01384 1012.73597   7.549 9.78e-14 ***
## purea          0.53331   0.09778 1017.97983   5.454 6.18e-08 ***
## wage          0.58541   0.02685 1015.10668  21.803 < 2e-16 ***
## varietiestrad  -4.09048   6.67448 1017.98996  -0.613 0.54011
## varietieshigh -17.69857   6.77927 1006.16565  -2.611 0.00917 **
## varietiesmixed -11.32341   7.20448 1014.90466  -1.572 0.11633
## bimasyes       -10.04658   2.11680  603.24959  -4.746 2.59e-06 ***
## bimasmixed     -4.88847   1.66888  626.91129  -2.929 0.00352 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          pseed purea wage vrtstr vrtshg vrtsmx bimsys
## purea      0.158
## wage     -0.634 -0.663
## varietistrd -0.192 -0.978 0.583
## varietishgh -0.254 -0.959 0.586 0.978
## varietismxd -0.193 -0.914 0.531 0.931 0.924
## bimasyes    0.009 0.025 -0.070 -0.045 -0.025 -0.037
## bimasmixed  0.043 -0.062 0.178 -0.034 -0.046 -0.043 0.110

ranova(model.2)

## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## price ~ pseed + purea + wage + varieties + bimas + (1 | id) - 1
##          npar logLik   AIC   LRT Df Pr(>Chisq)
## <none>      10 -4402.4 8824.8
## (1 | id)     9 -4404.0 8826.1 3.3013 1 0.06922 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effetti casuali sulla variabile region

```
regioni_id <- unique(data$region)
regione_mappa <- setNames(seq_along(regioni_id), regioni_id)

data_2 <- data %>%
  mutate(Regione_id = recode(region, !!!regione_mappa))
#data_2$Regione_id = as.factor(data_2$Regione_id)
attach(data_2)
```

```
## I seguenti oggetti sono mascherati da data:
##
## bimas, famlabor, goutput, hiredlabor, id, noutput, pesticide,
## phosphate, pphosph, price, pseed, purea, region, seed, size,
## status, totlabor, urea, varieties, wage
```

```
head(data_2)
```

```
##           id  size status varieties bimas seed urea phosphate pesticide pseed purea
## 1 101001 3.000  owner    mixed mixed   90  900      80      6000    80    75
## 2 101001 2.000  owner    trad mixed   40  600       0      3000    70    75
## 3 101001 1.000  owner    high mixed  100  700     150      5000   140    70
## 4 101001 2.000  owner    high mixed   60  600     100      5000    90    70
## 5 101001 3.572 share    high  no  105  400     400     10200   350    80
## 6 101001 3.572 share    high  no  105  400     400     10200   250    80
##   pphosph hiredlabor famlabor totlabor   wage goutput noutput price
## 1      75         2875      40    2915  68.49   7980   6800    60
## 2      75         2110      45    2155  60.09   4083   3500    60
## 3      70          980      95    1075  51.99   2650   2242    65
## 4      70         2081      10    2091  56.98   4500   3750    70
## 5      80         3889       1    3889 152.03  16300  13584   120
## 6      80         3519       1    3519 154.49  17424  14520   140
##           region Regione_id
## 1 wargabinangun      1
## 2 wargabinangun      1
## 3 wargabinangun      1
## 4 wargabinangun      1
## 5 wargabinangun      1
## 6 wargabinangun      1
```

```
print("-----Model 2 BIS-----")
```

```
## [1] "-----Model 2 BIS-----"
```

```
model.2B<-lmer(price~0 + pseed + purea + wage + varieties + bimas + pesticide + urea + phosphate + (1|
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
model.2B_2<-lmer(price~0 + pseed + purea + wage +bimas+ pesticide + urea + phosphate + (1|Regione_id), c
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
AIC(model.2B, model.2B_2) # Il modello con la variabile varieties ha un AIC migliore
```

```
##           df      AIC
## model.2B    13 8688.274
## model.2B_2  11 8692.400
```

```
summary(model.2B)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: price ~ 0 + pseed + purea + wage + varieties + bimas + pesticide +
##          urea + phosphate + (1 | Regione_id)
## Data: data_2
##
## REML criterion at convergence: 8662.3
```

```

##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.00168 -0.62588  0.07293  0.61224  3.07464
##
## Random effects:
##   Groups      Name      Variance Std.Dev.
## Regione_id (Intercept)  92.01    9.592
## Residual                260.30   16.134
## Number of obs: 1026, groups: Regione_id, 6
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## pseed          9.482e-02  1.281e-02  1.012e+03   7.402 2.81e-13 ***
## purea          6.062e-01  9.481e-02  1.014e+03   6.394 2.46e-10 ***
## wage          5.780e-01  2.566e-02  1.011e+03  22.524 < 2e-16 ***
## varietiestrad -1.438e+01  7.522e+00  6.082e+01  -1.911 0.060690 .
## varietieshigh -1.633e+01  7.800e+00  6.848e+01  -2.093 0.040056 *
## varietiesmixed -1.670e+01  7.981e+00  7.573e+01  -2.093 0.039732 *
## bimasyes      -7.592e+00  2.047e+00  1.014e+03  -3.710 0.000219 ***
## bimasmixed    -4.031e+00  1.539e+00  1.011e+03  -2.620 0.008927 **
## pesticide     -4.893e-04  1.875e-04  1.010e+03  -2.609 0.009202 **
## urea           2.508e-02  5.890e-03  1.014e+03   4.258 2.25e-05 ***
## phosphate     -5.086e-02  1.570e-02  1.012e+03  -3.240 0.001233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      pseed purea wage vrtstr vrtshg vrtsmx bimsys bmsmxd pestcd
## purea      0.105
## wage      -0.605 -0.648
## varietistrd -0.125 -0.835  0.496
## varietishgh -0.155 -0.828  0.482  0.972
## varietismxd -0.124 -0.803  0.458  0.950  0.945
## bimasyes   -0.037  0.034 -0.016 -0.037 -0.036 -0.043
## bimasmixed  0.025 -0.101  0.204  0.014  0.022  0.012  0.145
## pesticide   0.044  0.023 -0.091 -0.001 -0.015  0.008  0.002  0.018
## urea        0.074  0.057 -0.050 -0.080 -0.080 -0.084 -0.090 -0.173 -0.058
## phosphate  -0.038  0.076 -0.135 -0.044 -0.055 -0.051 -0.024 -0.003 -0.179
##      urea
## purea
## wage
## varietistrd
## varietishgh
## varietismxd
## bimasyes
## bimasmixed
## pesticide
## urea
## phosphate -0.645
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling

```



```
ranova(model.2B)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## price ~ pseed + purea + wage + varieties + bimas + pesticide + urea + phosphate + (1 | Regione_id) -
##               npar logLik    AIC    LRT Df Pr(>Chisq)
## <none>           13 -4331.1 8688.3
## (1 | Regione_id)  12 -4406.1 8836.3 149.99  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAMLSS

```
mod_ga<- gamlss(price~ 1+ pseed + purea + wage + varieties + bimas + pesticide + goutput + noutput + (1
```

```
## GAMLSS-RS iteration 1: Global Deviance = 8316.628
## GAMLSS-RS iteration 2: Global Deviance = 8316.628
```

```
summary(mod_ga)
```

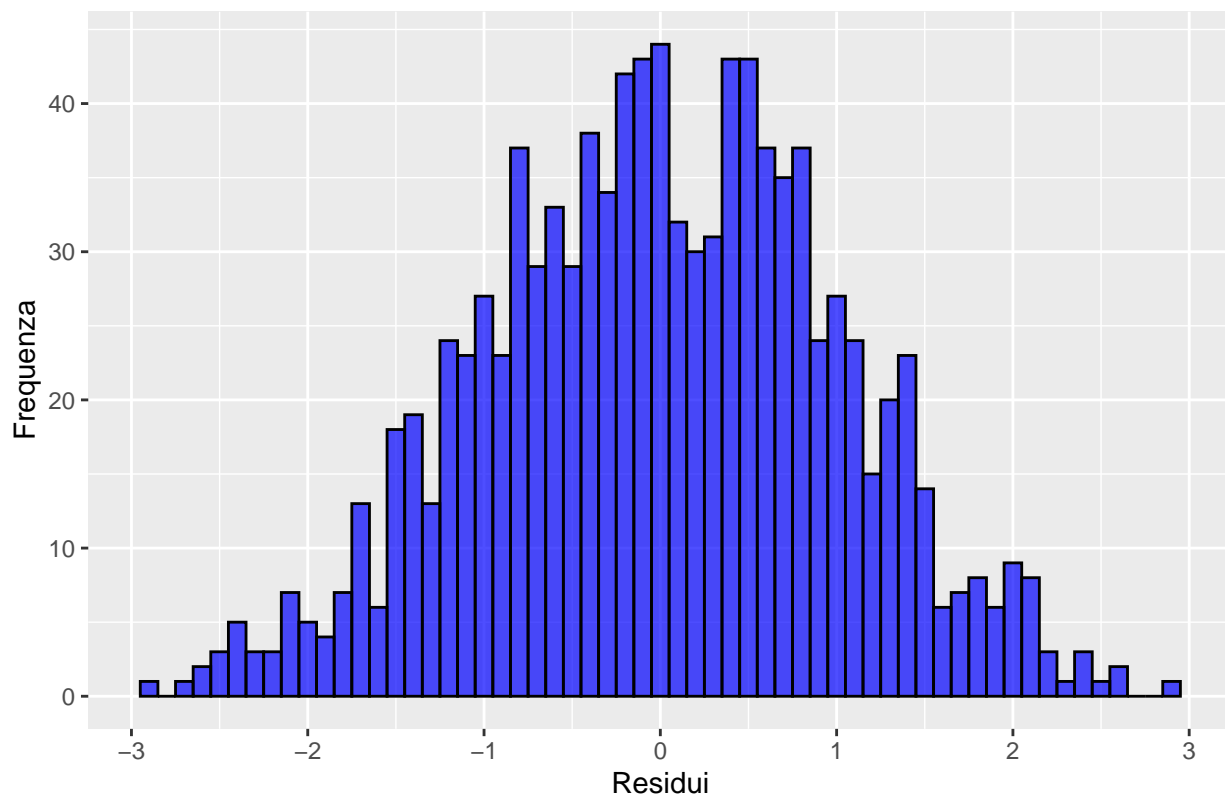
```
## *****
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = price ~ 1 + pseed + purea + wage +
##      varieties + bimas + pesticide + goutput + noutput +
##      (1 | Regione_id), family = GA, data = data_2)
##
## Fitting method: RS()
##
## -----
## Mu link function:  log
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.588e+00  6.256e-02  57.344 < 2e-16 ***
## pseed         1.171e-03  1.338e-04   8.752 < 2e-16 ***
## purea         3.562e-03  9.203e-04   3.870 0.000116 ***
## wage          6.163e-03  2.600e-04  23.702 < 2e-16 ***
## varietieshigh -1.309e-01  1.336e-02  -9.796 < 2e-16 ***
## varietiesmixed -7.527e-02  2.445e-02  -3.078 0.002138 **
## bimasyes      -8.011e-02  1.933e-02  -4.145 3.69e-05 ***
## bimasmixed    -6.317e-02  1.534e-02  -4.118 4.14e-05 ***
## pesticide     -6.046e-06  1.908e-06  -3.169 0.001578 **
## goutput       -1.733e-04  6.301e-05  -2.750 0.006060 **
## noutput        2.085e-04  7.392e-05   2.820 0.004899 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.80644    0.02198  -82.19 <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit: 1026
## Degrees of Freedom for the fit: 12
##      Residual Deg. of Freedom: 1014
##                      at cycle: 2
##
## Global Deviance:      8316.628
##           AIC:        8340.628
##           SBC:        8399.829
## *****
```

```
res=residuals(mod_ga)
```

```
res=residuals(mod_ga)
ggplot(data.frame(res), aes(x = res)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Istogramma dei residui", x = "Residui", y = "Frequenza")
```

Istogramma dei residui

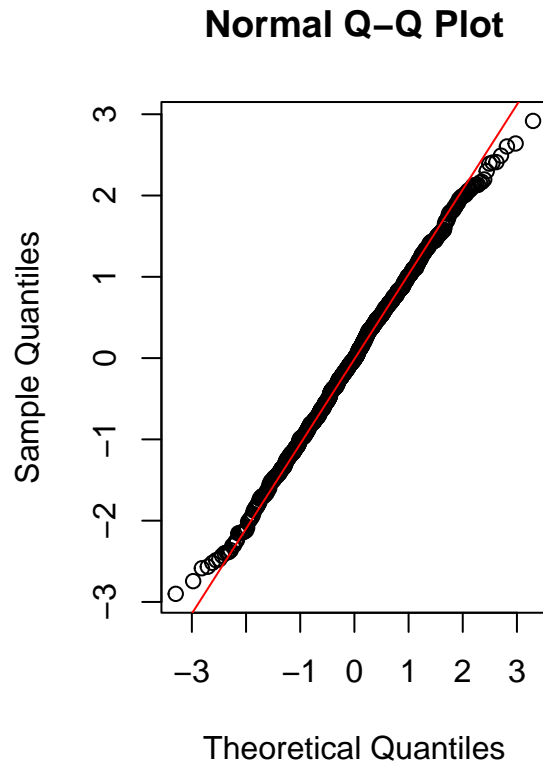


```
par(mfrow=c(1,2))
```

```
qqnorm(res)
qqline(res, col = "red")
```

```
shapiro.test(res)
```

```
##
## Shapiro-Wilk normality test
##
## data:  res
## W = 0.99806, p-value = 0.2895
```



```
mod_bis_ga <- gamlss(price ~ 1 + pseed + goutput + noutput + pphosph + varieties + bimas, random = ~1

## GAMLSS-RS iteration 1: Global Deviance = 8774.749
## GAMLSS-RS iteration 2: Global Deviance = 8774.749

mod_bis_ga_2 <- gamlss(price ~ 1 + pseed + goutput + noutput + pphosph + bimas, random = ~1 | region, f

## GAMLSS-RS iteration 1: Global Deviance = 8854.446
## GAMLSS-RS iteration 2: Global Deviance = 8854.446

AIC(mod_bis_ga, mod_bis_ga_2) # Il modello con varieties è migliore

##           df      AIC
## mod_bis_ga  10 8794.749
## mod_bis_ga_2  8 8870.446

summary(mod_ga)

## *****
## Family:  c("GA", "Gamma")
##
## Call:   gamlss(formula = price ~ 1 + pseed + purea + wage +
##               varieties + bimas + pesticide + goutput + noutput +
```

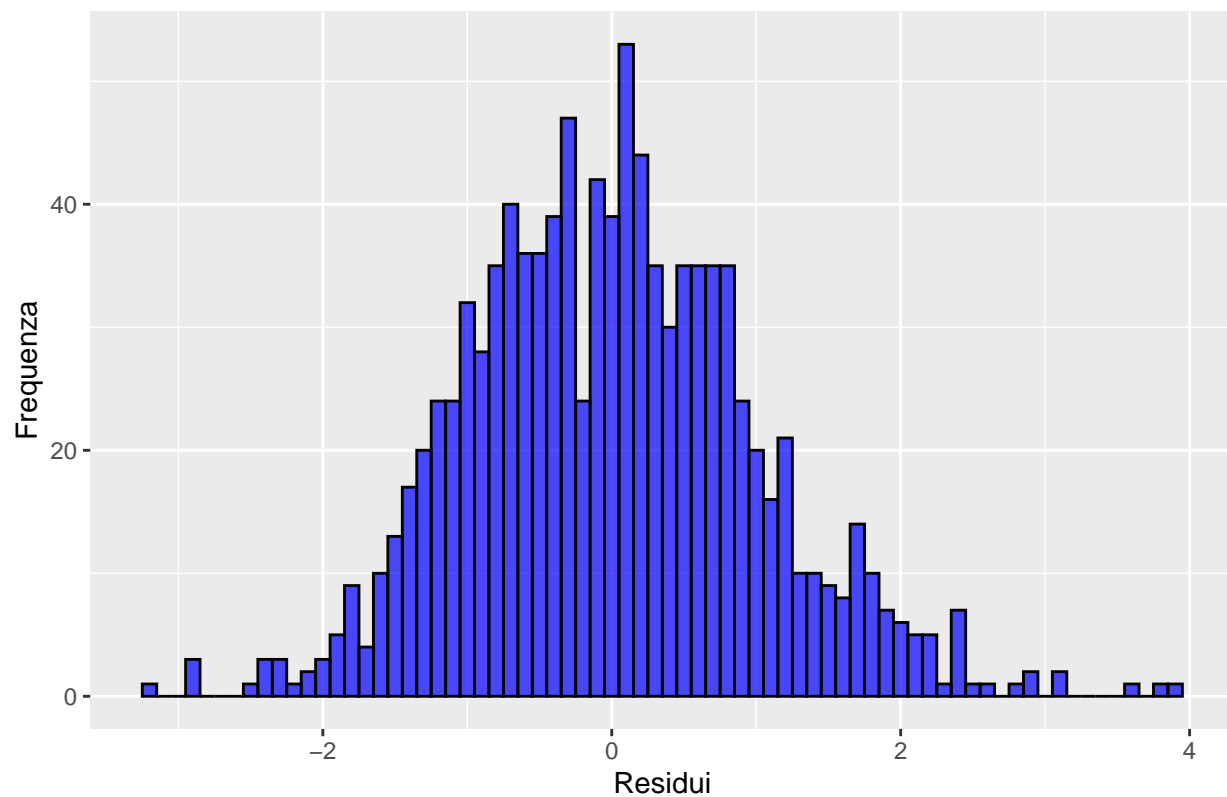
```

##      (1 | Regione_id), family = GA, data = data_2)
##
## Fitting method: RS()
##
## -----
## Mu link function:  log
## Mu Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.588e+00  6.256e-02  57.344 < 2e-16 ***
## pseed          1.171e-03  1.338e-04   8.752 < 2e-16 ***
## purea          3.562e-03  9.203e-04   3.870 0.000116 ***
## wage           6.163e-03  2.600e-04  23.702 < 2e-16 ***
## varietieshigh -1.309e-01  1.336e-02  -9.796 < 2e-16 ***
## varietiesmixed -7.527e-02  2.445e-02  -3.078 0.002138 **
## bimasyes       -8.011e-02  1.933e-02  -4.145 3.69e-05 ***
## bimasmixed     -6.317e-02  1.534e-02  -4.118 4.14e-05 ***
## pesticide      -6.046e-06  1.908e-06  -3.169 0.001578 **
## goutput        -1.733e-04  6.301e-05  -2.750 0.006060 **
## noutput         2.085e-04  7.392e-05   2.820 0.004899 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## Sigma link function:  log
## Sigma Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.80644    0.02198  -82.19 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit:  1026
## Degrees of Freedom for the fit:  12
##      Residual Deg. of Freedom:  1014
##      at cycle:  2
##
## Global Deviance:      8316.628
##      AIC:      8340.628
##      SBC:      8399.829
## *****

res_ga=residuals(mod_bis_ga)
ggplot(data.frame(res_ga), aes(x = res_ga)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Istogramma dei residui", x = "Residui", y = "Frequenza")

```

Istogramma dei residui



```
par(mfrow=c(1,2))
```

```
qqnorm(res_ga)
```

```
qqline(res_ga, col = "red")
```

```
shapiro.test(res_ga)
```

```
##
```

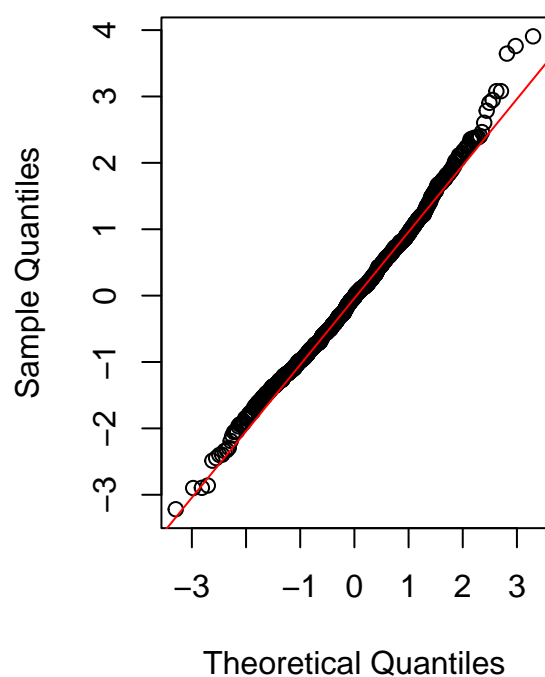
```
## Shapiro-Wilk normality test
```

```
##
```

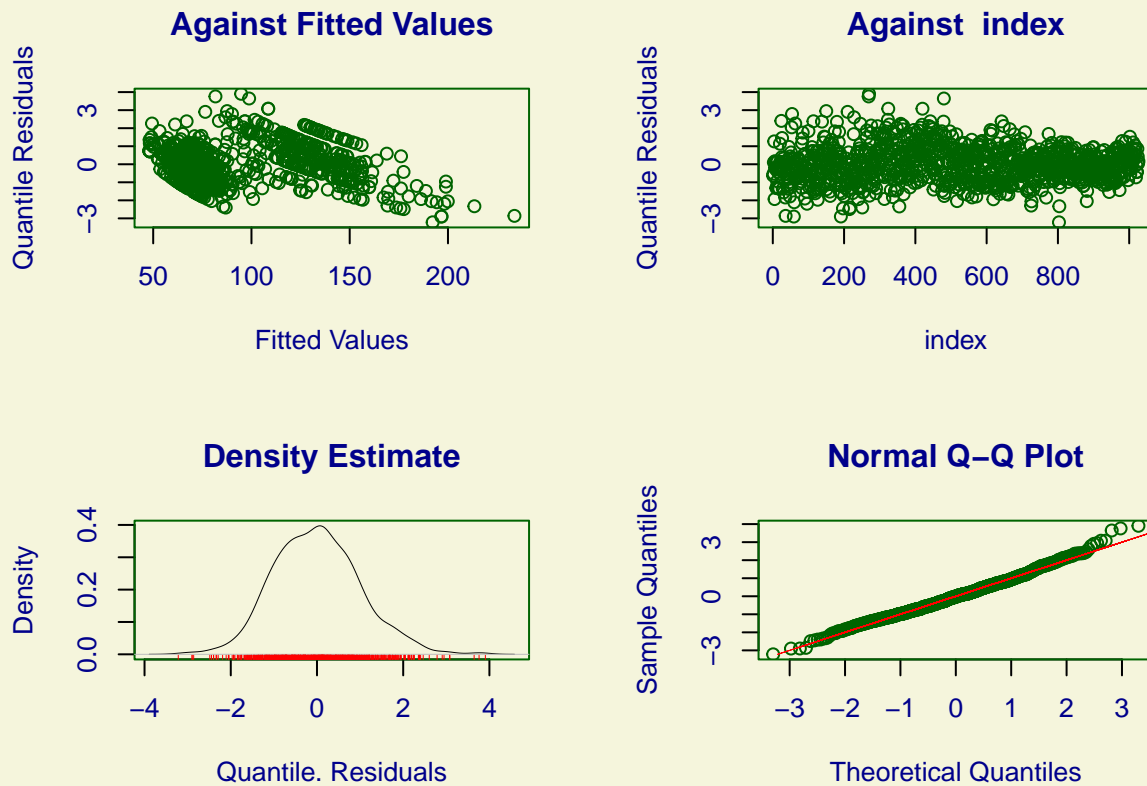
```
## data: res_ga
```

```
## W = 0.9937, p-value = 0.0002519
```

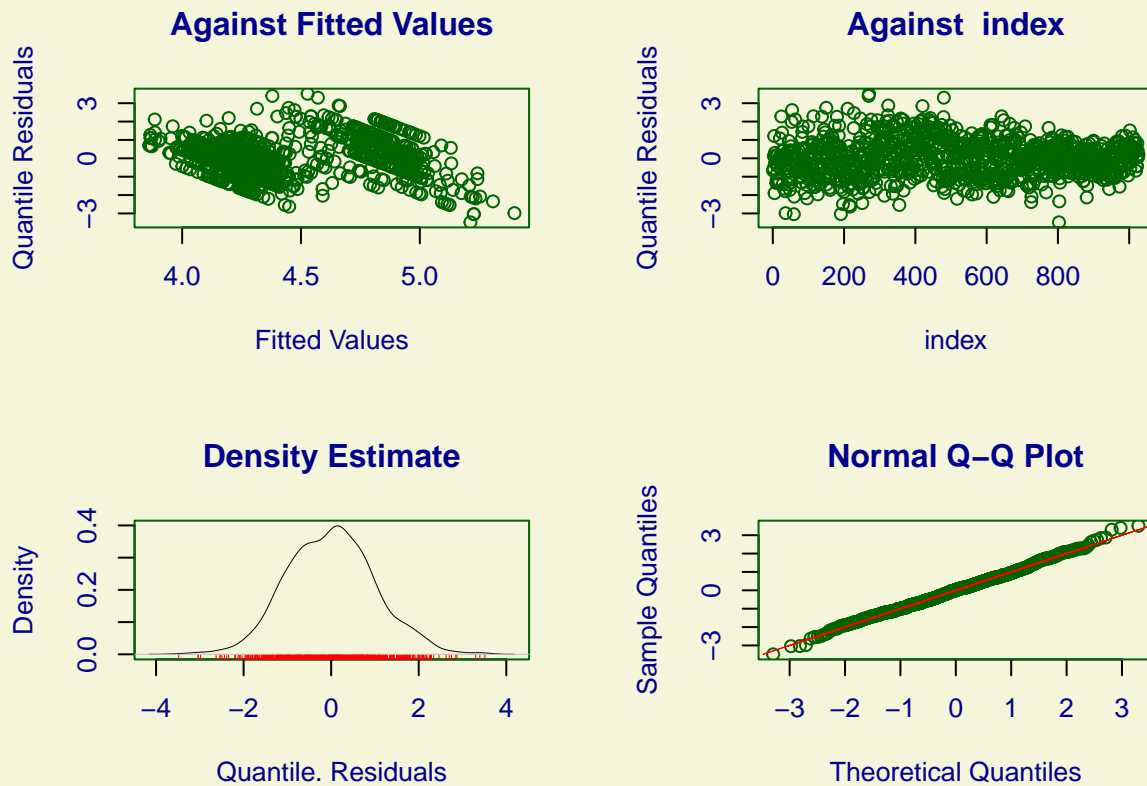
Normal Q-Q Plot



```
mod_logno <- gamlss(price ~ 1 + pseed + goutput + noutput + pphosph + varieties + bimas, random = ~1 |  
## GAMLSS-RS iteration 1: Global Deviance = 8760.353  
## GAMLSS-RS iteration 2: Global Deviance = 8760.353  
a=AIC(mod_bis_ga,mod_logno)  
b=BIC(mod_bis_ga,mod_logno)  
plot(mod_bis_ga, which = 1, main = "Gamma - Residui normalizzati")
```



```
## *****
##      Summary of the Quantile Residuals
##              mean   = -0.0003283216
##              variance = 1.000928
##              coef. of skewness = 0.2907995
##              coef. of kurtosis = 3.430588
## Filliben correlation coefficient = 0.9968034
## *****
plot(mod_logno, which = 1, main = "LOGNO - Residui normalizzati")
```



```
## *****
##      Summary of the Quantile Residuals
##              mean   = 2.650948e-16
##              variance = 1.000976
##              coef. of skewness = 0.1221622
##              coef. of kurtosis = 3.225733
## Filliben correlation coefficient = 0.9986982
## *****

cbind(a[2],b[2])

##              AIC      BIC
## mod_logno 8780.353 8844.083
## mod_bis_ga 8794.749 8829.688

print("mod_bis_ga")

## [1] "mod_bis_ga"

summary(mod_bis_ga)

## *****
## Family:  c("GA", "Gamma")
##
## Call:  gamlss(formula = price ~ 1 + pseed + goutput + noutput +
##      pphosph + varieties + bimas, family = GA, data = data_2,
##      random = ~1 | region)
##
```



```

## Fitting method: RS()
##
## -----
## Mu link function: log
## Mu Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.780e+00  5.991e-02  46.407 < 2e-16 ***
## pseed        3.150e-03  1.341e-04  23.489 < 2e-16 ***
## goutput      -2.884e-04  7.689e-05  -3.751 0.000186 ***
## noutput       3.555e-04  8.984e-05   3.957 8.12e-05 ***
## pphosph       1.710e-02  8.190e-04  20.875 < 2e-16 ***
## varietieshigh -1.511e-01  1.643e-02  -9.192 < 2e-16 ***
## varietiesmixed -5.757e-02  3.047e-02  -1.890 0.059096 .
## bimasyes      -6.444e-02  2.409e-02  -2.674 0.007609 **
## bimasmixed    -1.324e-01  1.873e-02  -7.069 2.90e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## Sigma link function: log
## Sigma Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.58568    0.02192  -72.33 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit: 1026
## Degrees of Freedom for the fit: 10
##      Residual Deg. of Freedom: 1016
##      at cycle: 2
##
## Global Deviance:      8774.749
##      AIC:      8794.749
##      SBC:      8844.083
## *****

print("mod_logno")

## [1] "mod_logno"

summary(mod_logno)

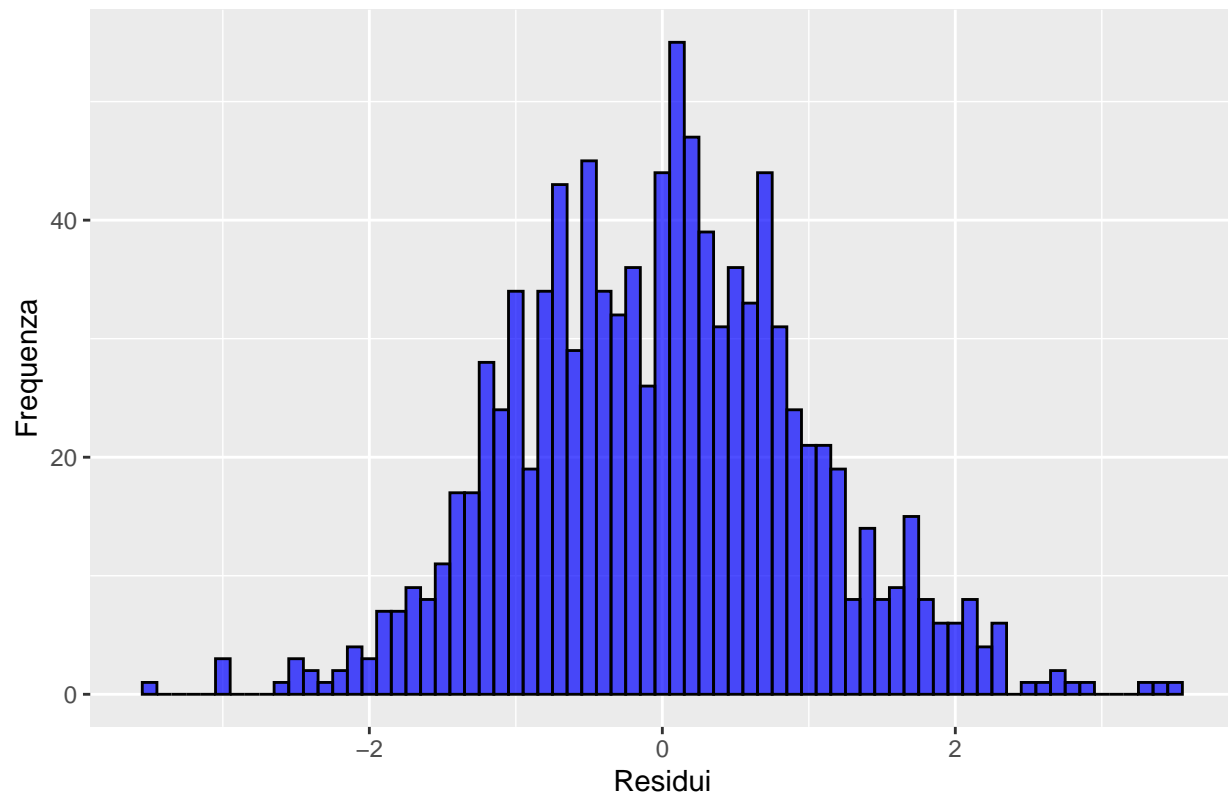
## *****
## Family: c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = price ~ 1 + pseed + goutput + noutput +
##      pphosph + varieties + bimas, family = LOGNO, data = data_2,
##      random = ~1 | region)
##
## Fitting method: RS()
##
## -----
## Mu link function: identity
## Mu Coefficients:

```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.787e+00  5.987e-02  46.554 < 2e-16 ***
## pseed          3.030e-03  1.255e-04  24.143 < 2e-16 ***
## goutput        -2.986e-04  7.609e-05  -3.924 9.29e-05 ***
## noutput         3.678e-04  8.886e-05   4.139 3.78e-05 ***
## pphosph         1.688e-02  8.107e-04  20.820 < 2e-16 ***
## varietieshigh  -1.512e-01  1.668e-02  -9.068 < 2e-16 ***
## varietiesmixed -5.549e-02  3.040e-02  -1.825  0.0683 .
## bimasyes       -5.695e-02  2.415e-02  -2.358  0.0185 *
## bimasmixed     -1.263e-01  1.864e-02  -6.779 2.05e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## Sigma link function:  log
## Sigma Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.58570    0.02208  -71.83  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
## No. of observations in the fit:  1026
## Degrees of Freedom for the fit:  10
##      Residual Deg. of Freedom:  1016
##               at cycle:  2
##
## Global Deviance:      8760.353
##           AIC:        8780.353
##           SBC:        8829.688
## *****
```

```
res_logno=residuals(mod_logno)
ggplot(data.frame(res_logno), aes(x = res_logno)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Istogramma dei residui", x = "Residui", y = "Frequenza")
```

Istogramma dei residui

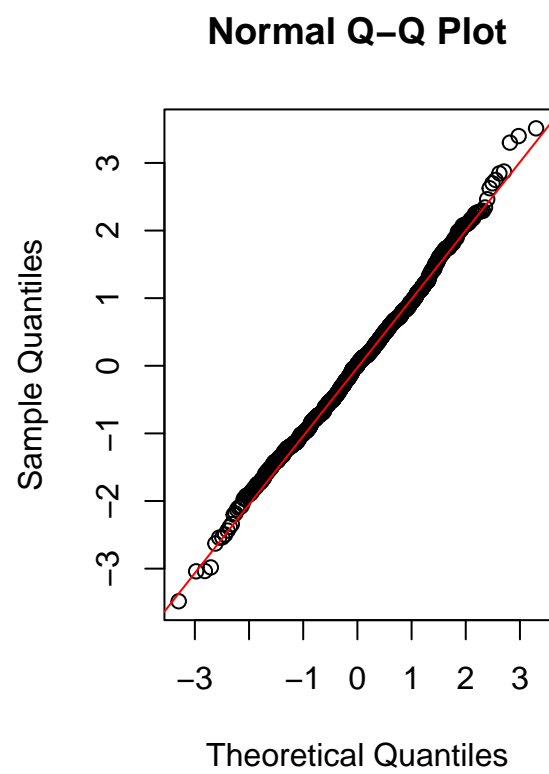


```
par(mfrow=c(1,2))

qqnorm(res_logno)
qqline(res_logno, col = "red")

shapiro.test(res_logno)

##
##  Shapiro-Wilk normality test
##
## data:  res_logno
## W = 0.99745, p-value = 0.1074
```



Conclusioni