Trasformazioni

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Indice

1 Direct Marketing

Il foglio elettronico `DirectMarketing.csv`[1] contiene informazioni su una campagna di *direct marketing* basata sulla spedizione via posta di cataloghi

```
direct <- read.csv( "DirectMarketing.csv" )
```

I dati comprendono una varietà di informazioni relative ad un campione di clienti (`Age`, `Gender`, `OwnHome`, `Married`, `Salary`). I dati comprendono anche un indicatore se il cliente viva vicino ad un negozio che commercia i prodotti pubblicizzati (`Location`), la sua propensione passata ad acquistare i prodotti pubblicizzati (`History`), il numero di cataloghi ricevuti (`Catalogs`) e la spesa effettuata nei prodotti pubblicizzati (`AmountSpent`)

```
summary(direct)
```

```
##      Age         Gender    OwnHome       Married     Location
##  Middle:508   Female:506   Own :516   Married:502   Close:710
##  Old   :205   Male  :494   Rent:484   Single :498   Far  :290
##  Young :287
##
```

[1] Il dataset è tratto da *Jank, W. (2011). Business Analytics for Managers. Springer.*

```
##
##
##      Salary          Children       History       Catalogs
##  Min.   : 10100   Min.   :0.000   High  :255   Min.   : 6.0
##  1st Qu.: 29975   1st Qu.:0.000   Low   :230   1st Qu.: 6.0
##  Median : 53700   Median :1.000   Medium:212   Median :12.0
##  Mean   : 56104   Mean   :0.934   NA's  :303   Mean   :14.7
##  3rd Qu.: 77025   3rd Qu.:2.000                3rd Qu.:18.0
##  Max.   :168800   Max.   :3.000                Max.   :24.0
##   AmountSpent
##  Min.   :  38
##  1st Qu.: 488
##  Median : 962
##  Mean   :1217
##  3rd Qu.:1688
##  Max.   :6217
```

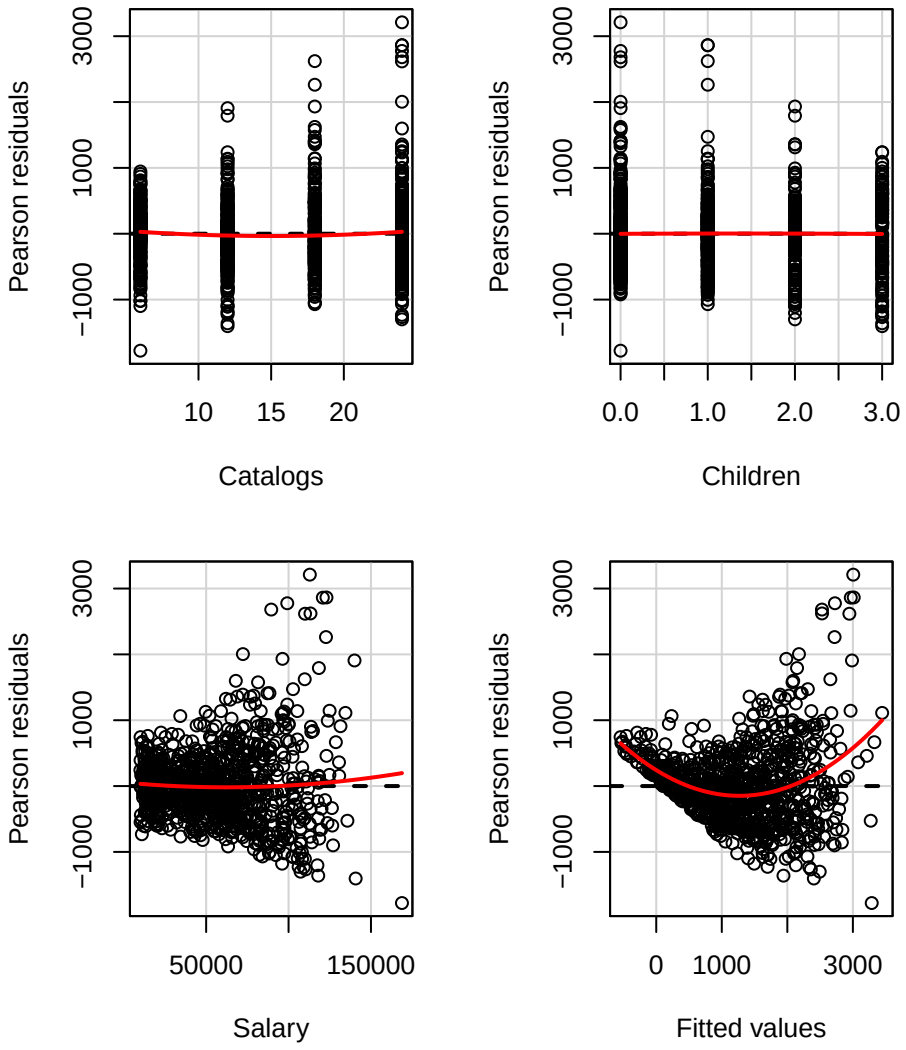
Consideriamo il modello di regressione

```
modello <- lm( AmountSpent ~ Catalogs+Children+Salary, data=direct )
summary( modello )
```

```
##
## Call:
## lm(formula = AmountSpent ~ Catalogs + Children + Salary, data = direct)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1776   -349    -39    255   3211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.43e+02   5.37e+01   -8.24  5.3e-16 ***
## Catalogs     4.77e+01   2.76e+00   17.31  < 2e-16 ***
## Children    -1.99e+02   1.71e+01  -11.63  < 2e-16 ***
## Salary       2.04e-02   5.93e-04   34.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 563 on 996 degrees of freedom
## Multiple R-squared:  0.658,Adjusted R-squared:  0.657
## F-statistic:  640 on 3 and 996 DF,  p-value: <2e-16
```

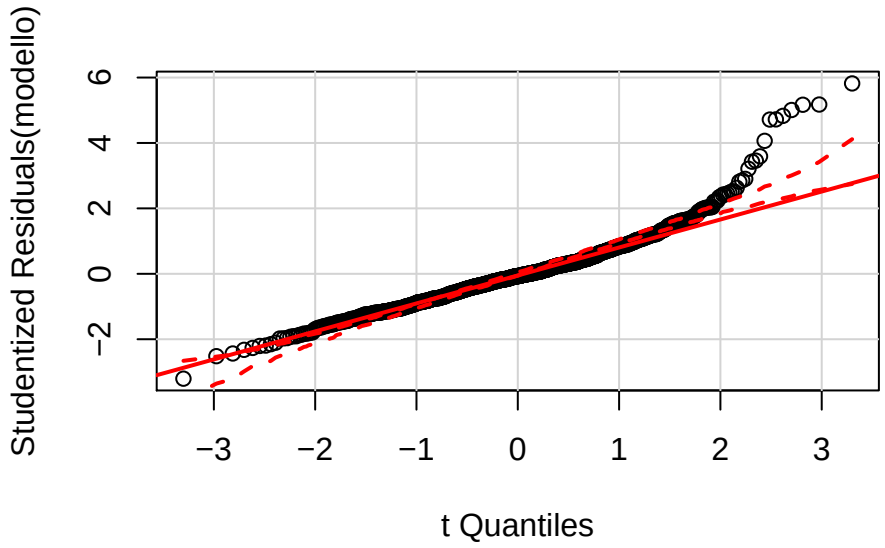
Valutiamo i residui del modello

```
library(car)
residualPlots(modello)
```




```
##            Test stat Pr(>|t|)
## Catalogs       1.600    0.110
## Children      -0.124    0.902
## Salary         1.131    0.258
## Tukey test    10.465    0.000
```

```
qqPlot( modello )
```



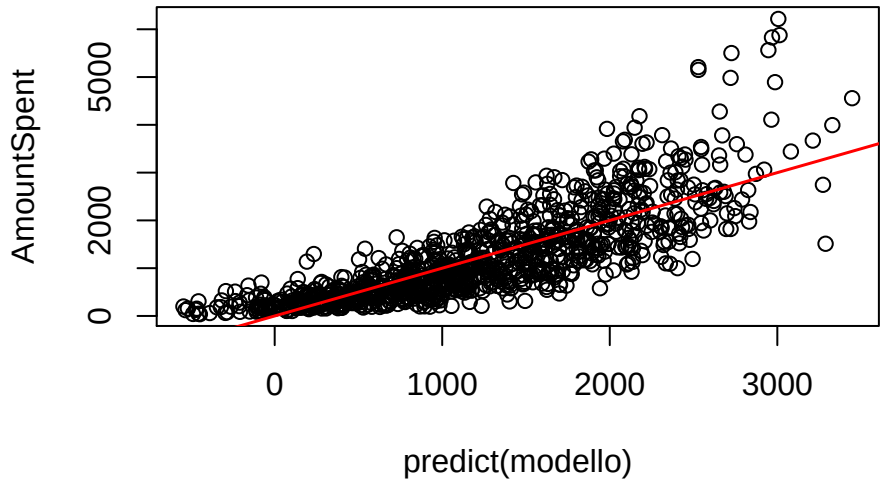
Inoltre, i valori previsti dal modello includono valori negativi!

```
summary(predict(modello))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -545     645    1160    1220    1800    3450
```

Come si può vedere anche dal grafico a dispersione fra residui e spesa

```
plot(AmountSpent~predict(modello), data=direct)
abline(0, 1, col="red", lwd=1.5)
```



Il grafico evidenzia anche una scarsa corrispondenza fra i valori della variabile risposta e i valori predetti dal modello.

2 Transformare la variabile risposta

Proviamo a modellare la variabile risposta su scala logaritmica (perché?)

```
modello2 <- lm( log(AmountSpent) ~ Catalogs+Children+Salary, data=direct )
summary( modello2 )

##
## Call:
## lm(formula = log(AmountSpent) ~ Catalogs + Children + Salary,
##     data = direct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4546 -0.3254  0.0064  0.3054  1.4346
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.26e+00   4.28e-02   123.0   <2e-16 ***
## Catalogs     4.50e-02   2.19e-03    20.5   <2e-16 ***
## Children    -2.42e-01   1.36e-02   -17.8   <2e-16 ***
## Salary       1.92e-05   4.72e-07    40.8   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.448 on 996 degrees of freedom
## Multiple R-squared:  0.739,Adjusted R-squared:  0.738
## F-statistic:  941 on 3 and 996 DF,  p-value: <2e-16
```

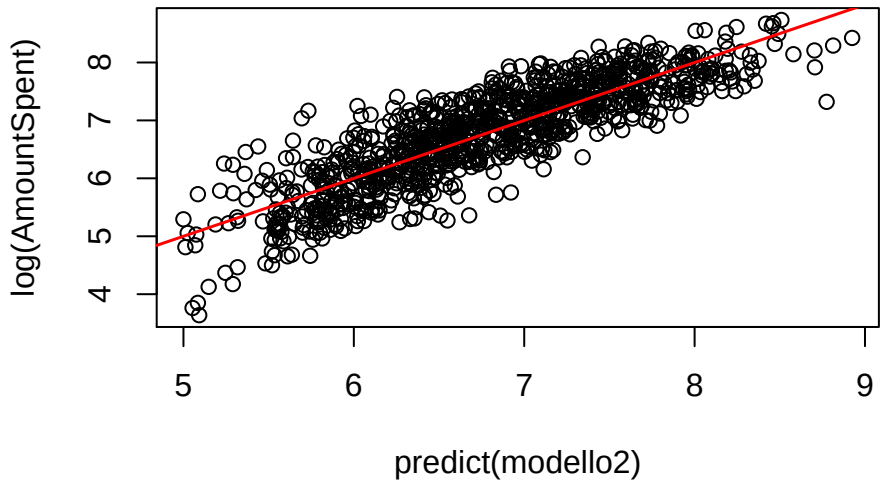
In questo modello ovviamente tutte le previsioni sono positive

```
summary( exp( predict(modello2) ) )

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     148     509     844    1160    1520    7510
```

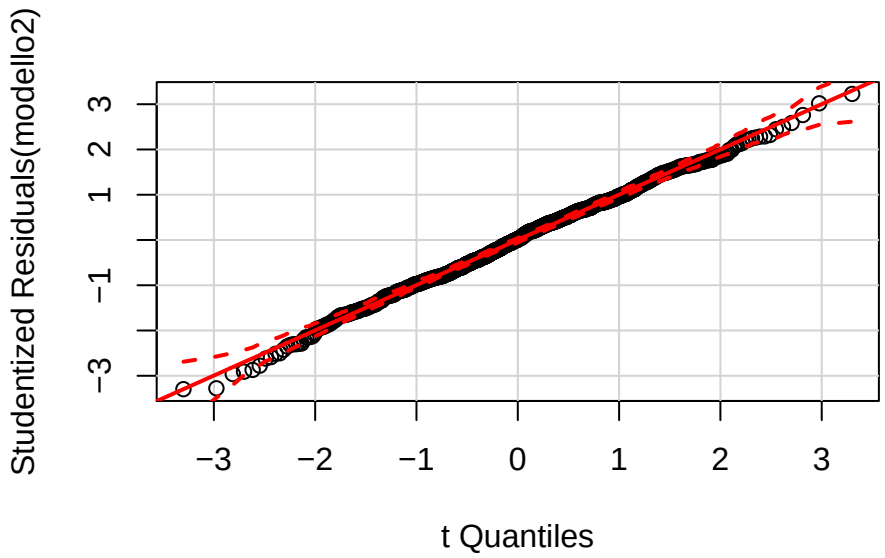
Vi è anche una buona corrispondenza fra i valori della variabile risposta su scala logaritmica e i valori predetti dal modello

```
plot(log(AmountSpent)~predict(modello2), data=direct)
abline(0, 1, col="red", lwd=1.5)
```

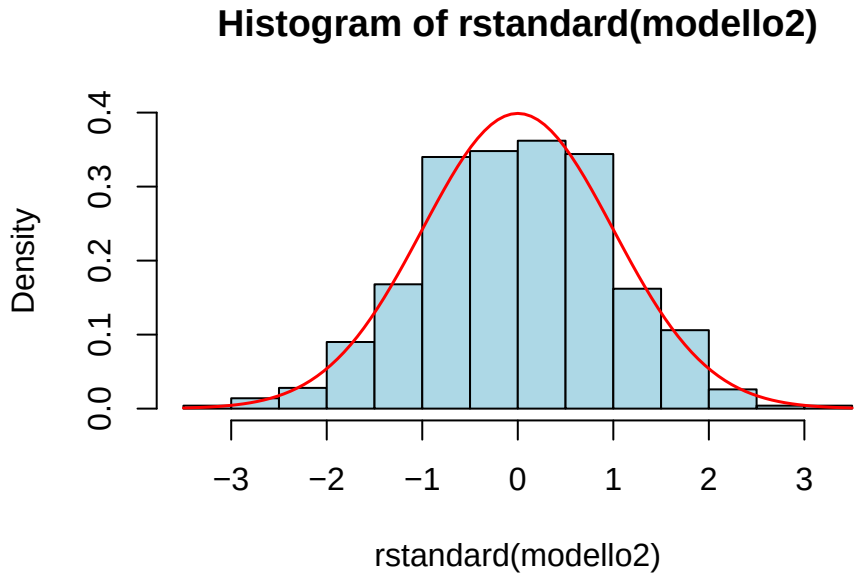


Controlliamo se il modello soddisfa le assunzioni. Iniziamo con l'assunzione di normalità

```
qqPlot(modello2)
```



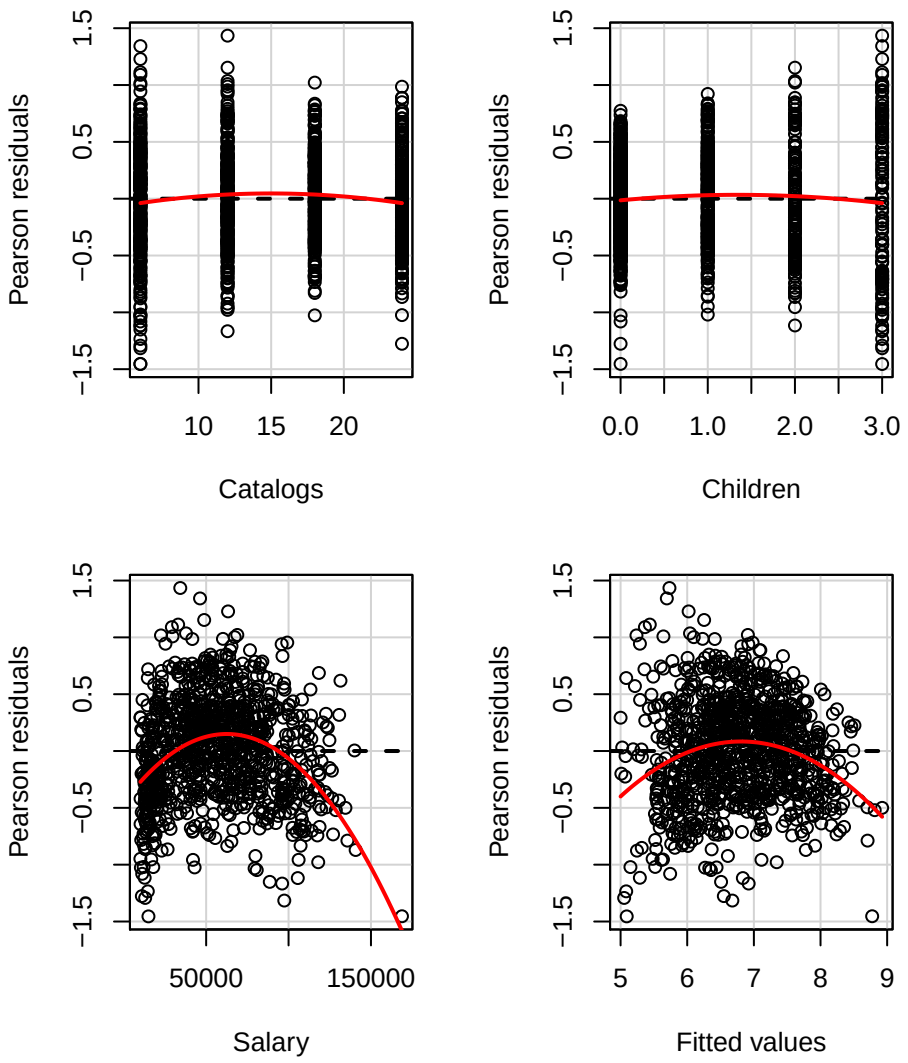

```
hist( rstandard(modello2), col="lightblue", freq=FALSE,
ylim=c( 0, 1/sqrt(2*pi) ) )
curve(dnorm(x), col="red", lwd=1.5, add=TRUE)
```



Decisamente meglio!

Però se controlliamo gli andamenti dei residui...

```
residualPlots(modello2)
```




```
##            Test stat Pr(>|t|)
## Catalogs      -2.681    0.007
## Children      -1.755    0.080
## Salary       -12.265    0.000
## Tukey test    -7.439    0.000
```

Vi è evidenza di relazioni non lineari fra i residui e il salario e fra i residui e i valori predetti.

3 Transformare i predittori

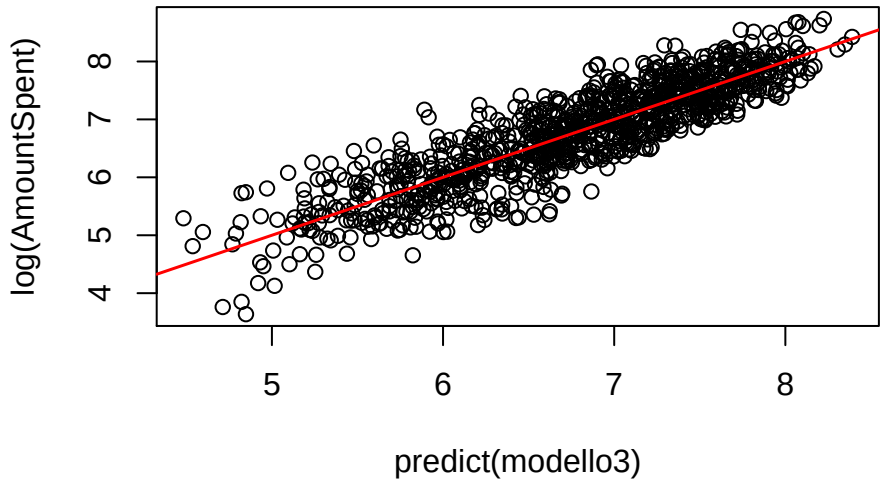
Proviamo a trasformare anche il predittore `Salary`

```
modello3 <- lm( log(AmountSpent) ~ Catalogs+Children+log(Salary),
data=direct )
summary(modello3)

##
## Call:
## lm(formula = log(AmountSpent) ~ Catalogs + Children + log(Salary),
##     data = direct)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2587 -0.2746 -0.0067  0.2758  1.2769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.83016    0.21505   -17.8   <2e-16 ***
## Catalogs     0.04321    0.00201    21.5   <2e-16 ***
## Children    -0.22897    0.01245   -18.4   <2e-16 ***
## log(Salary)  0.94700    0.02034    46.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.41 on 996 degrees of freedom
## Multiple R-squared:  0.781,Adjusted R-squared:  0.78
## F-statistic: 1.18e+03 on 3 and 996 DF,  p-value: <2e-16
```

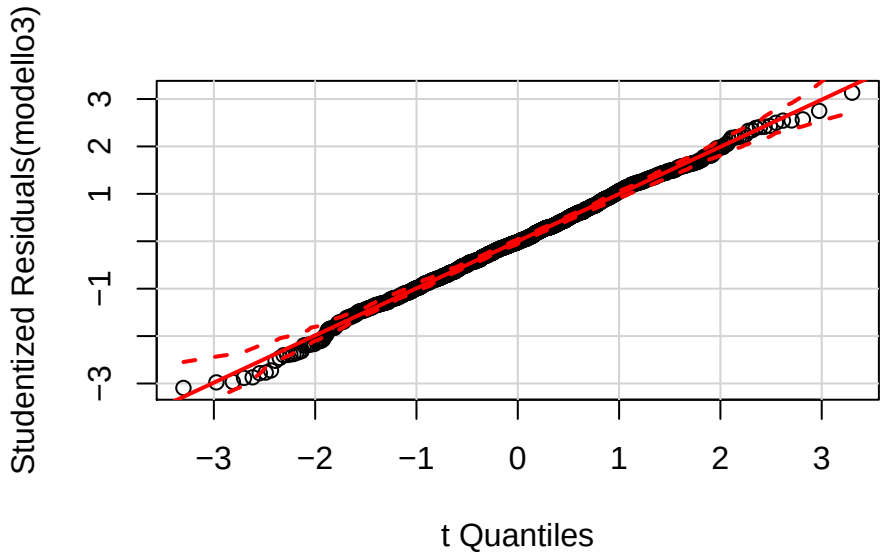
Valutiamo l'adattamento del modello

```
plot(log(AmountSpent)~predict(modello3), data=direct)
abline(0, 1, col="red", lwd=1.5)
```



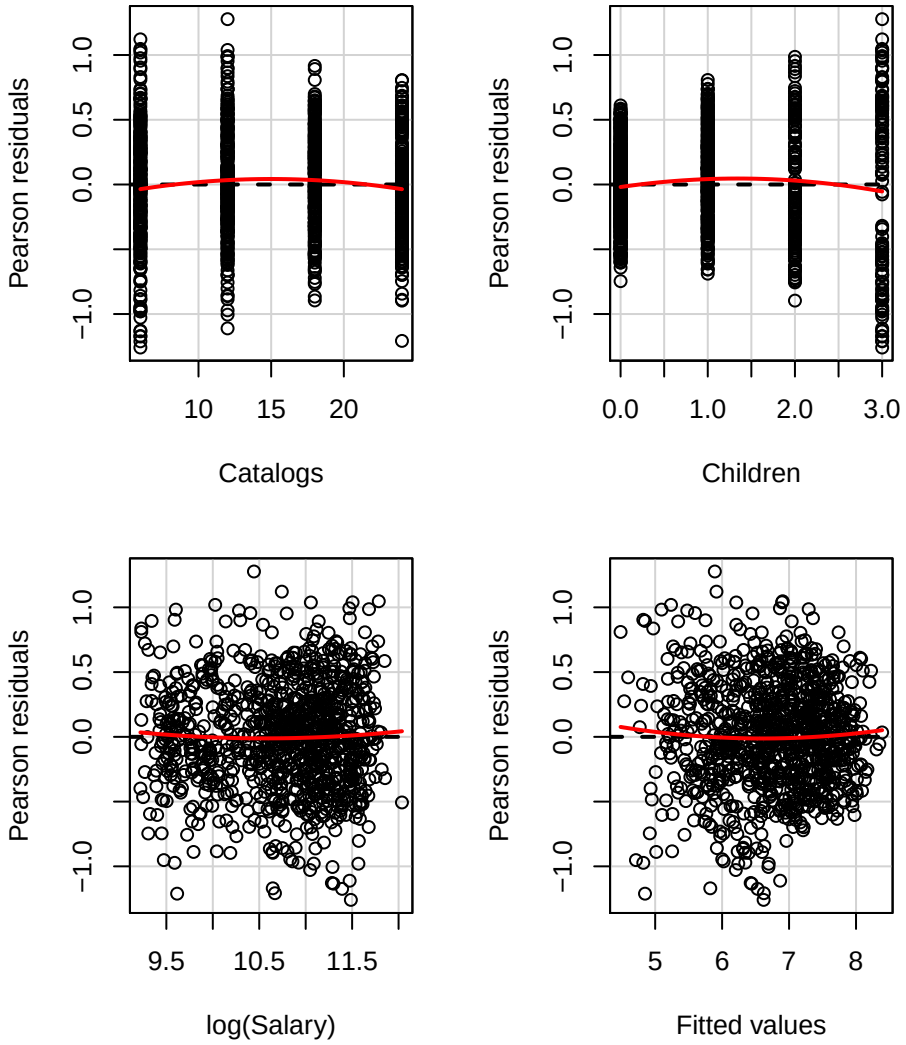
Controlliamo l'assunzione di normalità

```
qqPlot(modello3)
```



Infine, valutiamo i nuovi residui

```
residualPlots(modello3)
```




```
##             Test stat Pr(>|t|)
## Catalogs       -2.701    0.007
## Children       -2.606    0.009
## log(Salary)     0.837    0.403
## Tukey test      1.086    0.278
```

Commenti?