BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

RELAZIONE ANALISI DATI “FELICITA’”

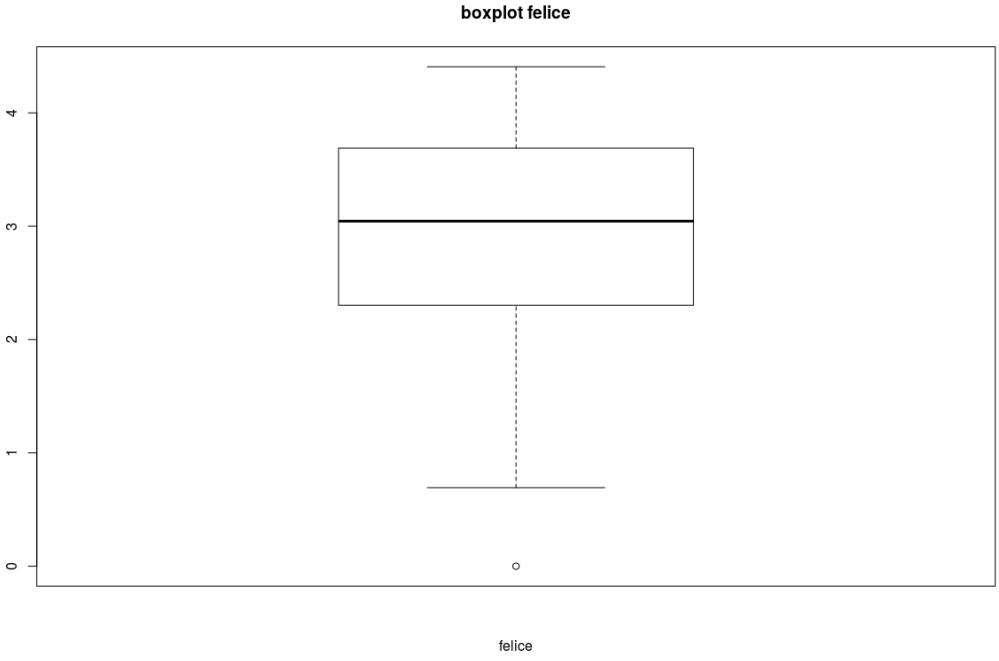
PARTE 1:

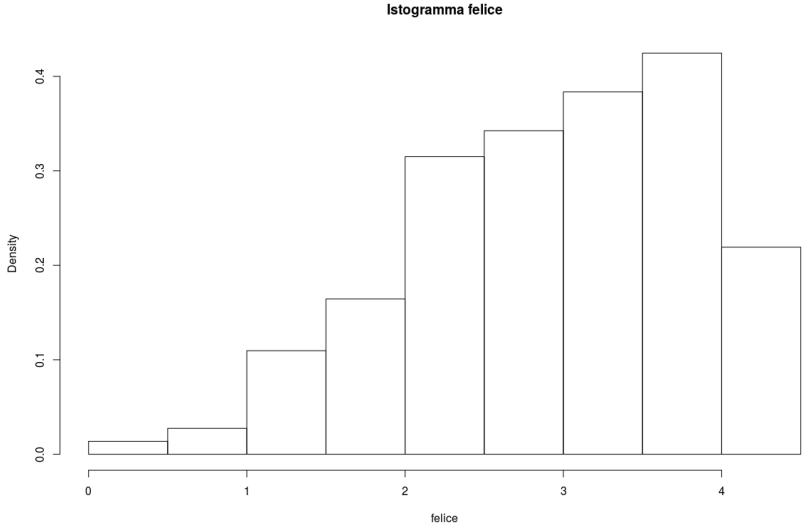
La mia analisi partira’ con ispezioni grafiche e controllo di dati mancanti. Cerchero’ poi di costruire un modello lineare e di valutare l’inserimento di qualche spline. Scegliero’ poi il modello che ha un test error minore. In finale rispondero’ alla domanda.

1) Dati mancanti e variabili qualitative considerate come fattori
Il data set dato non contiene dati mancanti, e la variabile qualitativa equatore e’ intrpretata da R come fattori

2) Ispezioni grafiche

- Distribuzione della variabile felice:

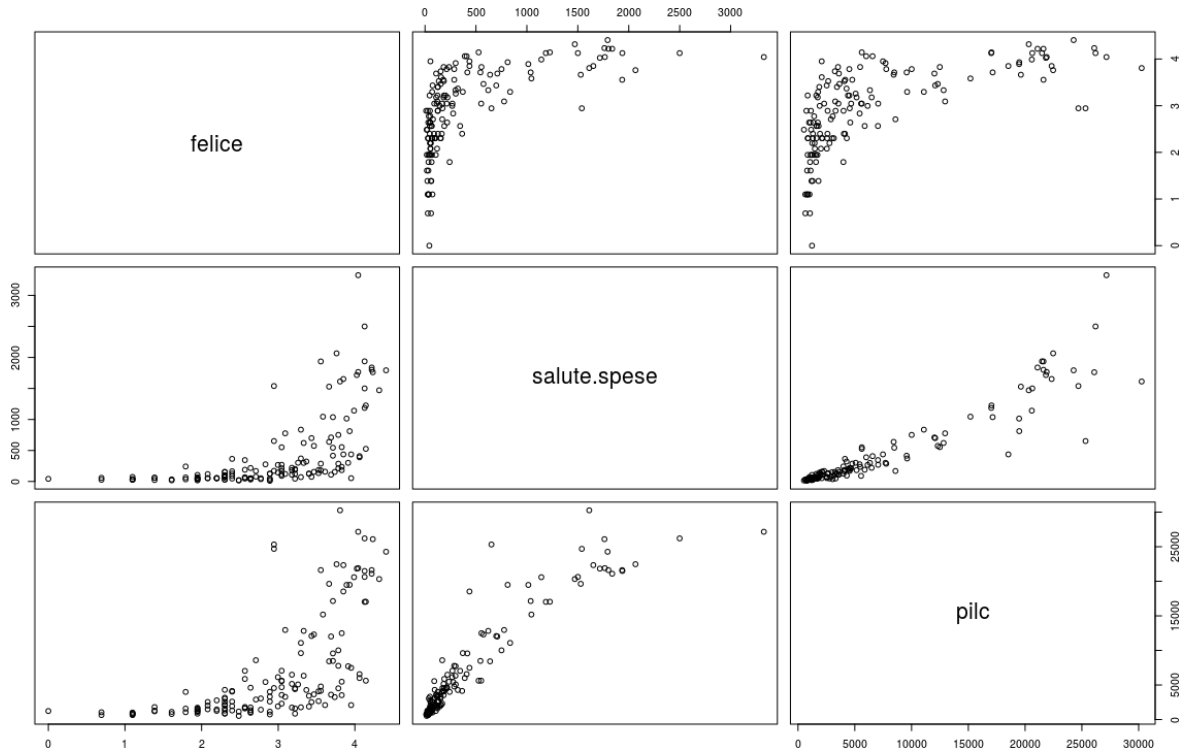




BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

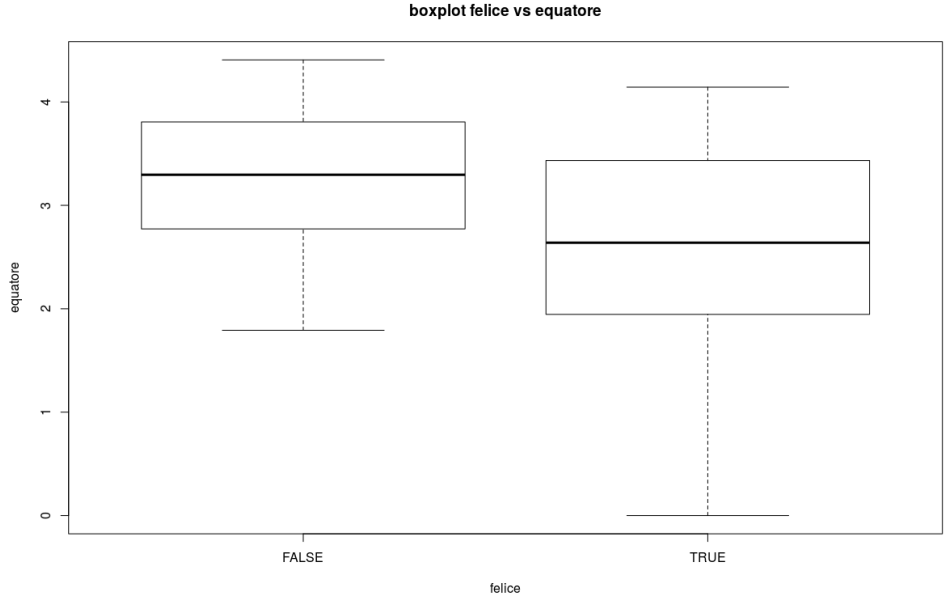
La distribuzione della variabile felice non e' normale al 100% ma una trasformata logaritmica, o radice quadrata, non aiuta a migliorare la normalita'. Lavorero' con questa distribuzione.

- Relazioni con le esplicative continue(slute.spse, pilc)



Osservo che le relazioni tra le esplicative e la variabile risposta, non sono lineari. Tuttavia la loro presenza nel modello potrebbe essere lineare. Un'altra osservazione e' la forte correlazione tra salute.spese e pilc, mi aspetto un compotamento anomalo nella stima dei coefficienti.

- Relazione tra felice equatore



BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

Osservo variabilita' dell'indice di felicita' in base alla collocazione del paese sull'equatore, mi aspetto un coefficiente significativo , di segno negativo risptto al livello base FALSE.

- Grafici interazioni tra splicative continue e esplicative qualitative

Per motivi di tempo non faro' i grafici che rilevano l'eventuale interazione, si evendenzieranno nelle stime dei coefficienti nel modello, nell'R^2 ecc..

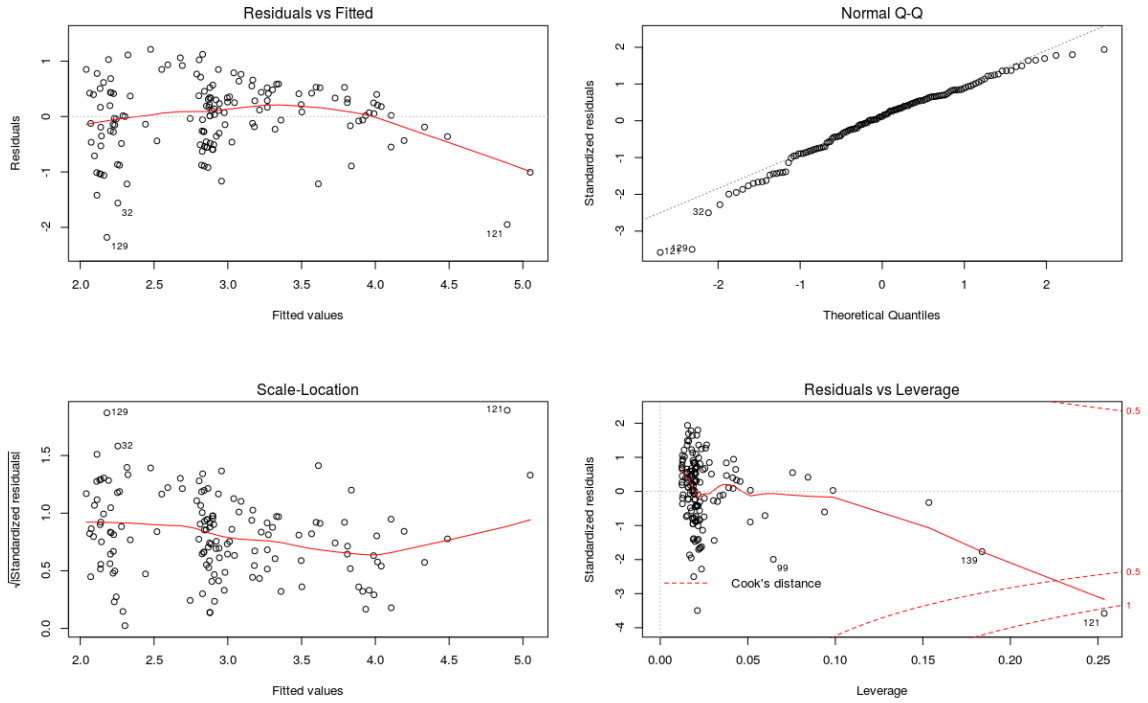
3) Modello Lineare

Dopo aver provato vari modelli lineari, sono giunti ad un modello che spiega la variabile felice in funzione di una interazione tra equatore e salute.spese. Questo modello non contiene la variabile pilc perche' era fortemente correlata a salute.spese, mi bastava quindi scegliere una tra i due ed ho scelto salute.spese in quanto il modello con salute.spese ha un andamento dei residui migliore del modello con pilc.
Il modello e':

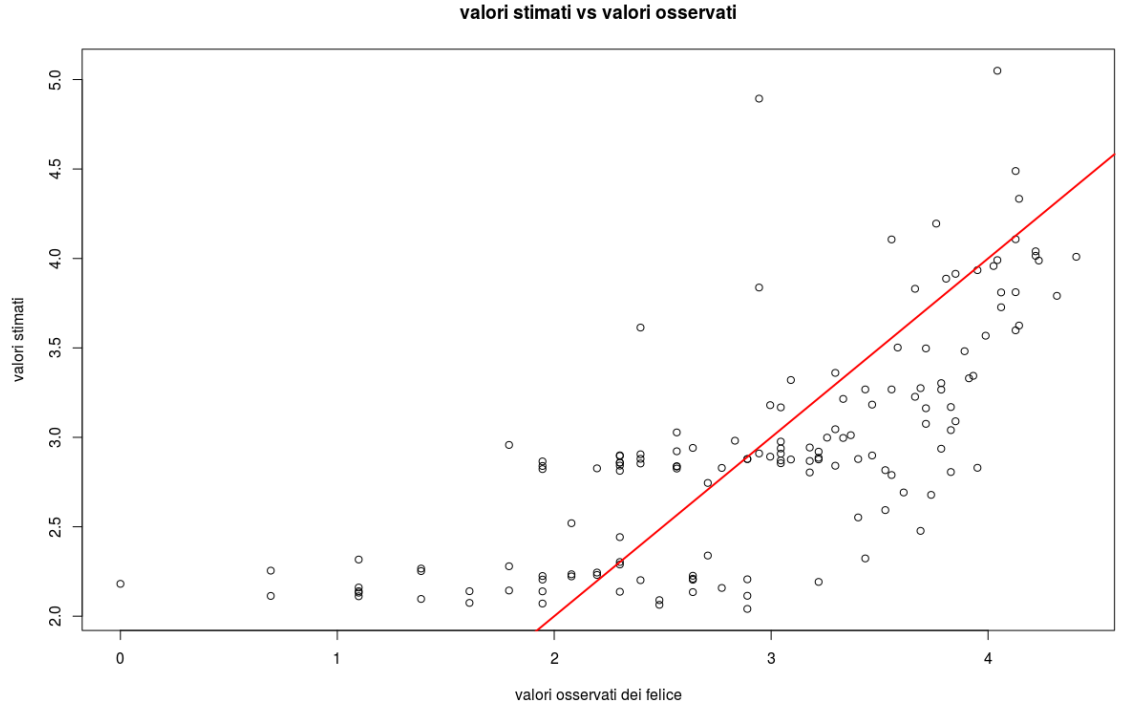
```
lm(formula = felice ~ equatore * salute.spese, data = dati1)

Residuals:
      Min       1Q   Median       3Q      Max
-2.18089 -0.36061  0.07498  0.41912  1.21191

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.794e+00  9.578e-02  29.168  < 2e-16 ***
equatoreTRUE              -8.021e-01  1.435e-01  -5.588 1.14e-07 ***
salute.spese               6.783e-04  9.812e-05   6.912 1.49e-10 ***
equatoreTRUE:salute.spese  3.772e-03  5.918e-04   6.374 2.42e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6303 on 142 degrees of freedom
Multiple R-squared:  0.5138,    Adjusted R-squared:  0.5036
F-statistic: 50.03 on 3 and 142 DF,  p-value: < 2.2e-16
```

L'andamento dei residui e' il seguente:

BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017



Dal grafico 1 e 3, si capisce bene che il modello ha bisogno di un termine polinomiale. Inoltre si nota un osservazione potenzialmente anomala secondo la distanza di cook nel grafico 4, ma forse un termine polinomiale riuscira' a cogliere questo punto. Tutti questi problemi si notano nei valori stimati:



Si nota dal grafico che i valori sono molto dispersi.
Vediamo se l'inserimento di qualche spline riesce a migliorare il modello, altrimenti valutero' se togliere qualche osservazione (quelle sospette), migliora il modello.

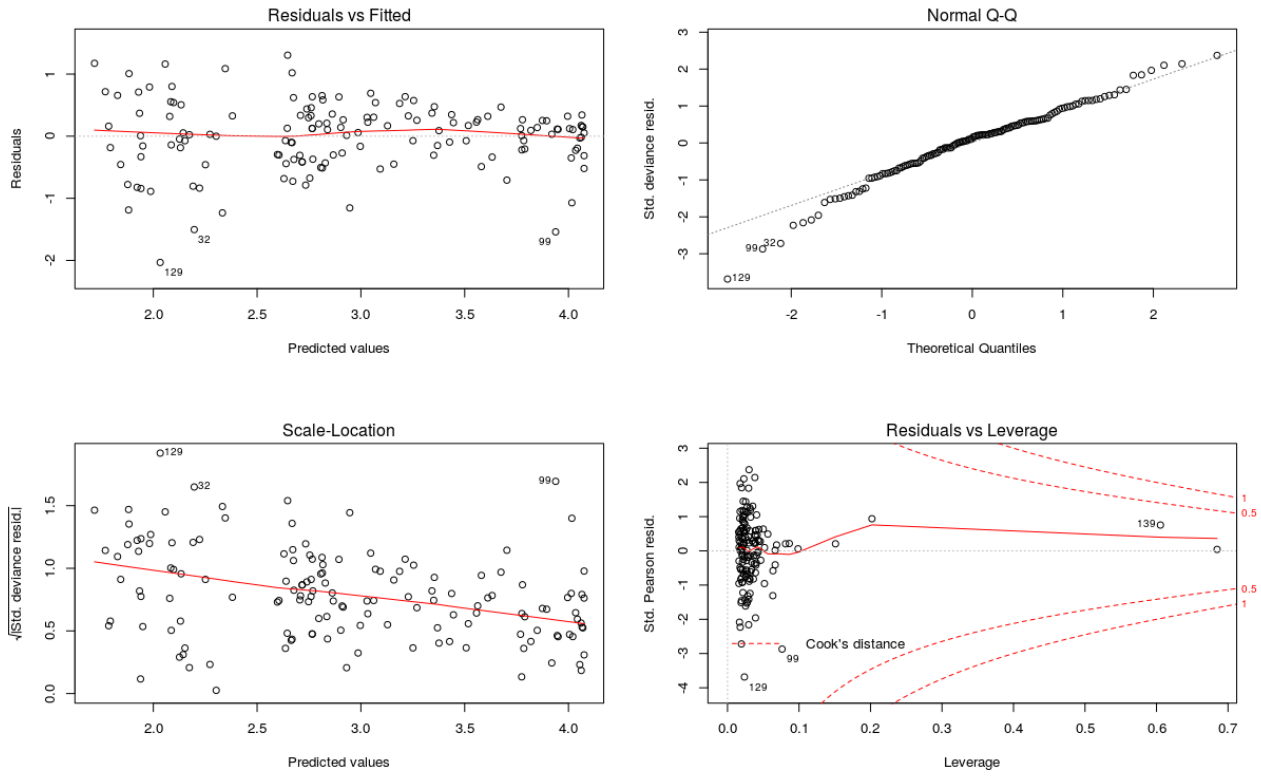
BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

3) Modello semiparametrico

Dopo aver provato vari modelli, sono giunto al seguente modello con spline quadratica:

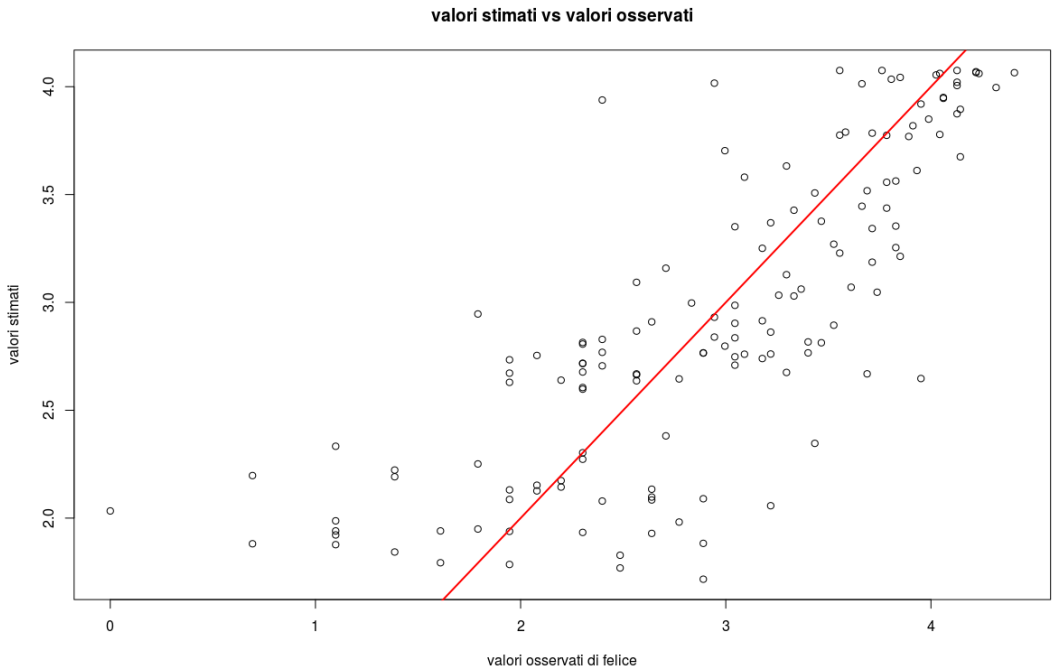
```
glm(formula = felice ~ equatore * ns(salute.spese, 2), data = dati1)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.0331  -0.3075   0.0738   0.3255   1.3042

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2.5766     0.1058  24.358  < 2e-16 ***
equatoreTRUE                            -0.8600     0.1521  -5.656 8.42e-08 ***
ns(salute.spese, 2)1                     2.6852     0.3232   8.309 7.52e-14 ***
ns(salute.spese, 2)2                     0.6255     0.4730   1.322    0.188
equatoreTRUE:ns(salute.spese, 2)1 -34.6514     8.5519  -4.052 8.38e-05 ***
equatoreTRUE:ns(salute.spese, 2)2 -91.0532    19.4320  -4.686 6.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.3118358)

    Null deviance: 116.058  on 145  degrees of freedom
Residual deviance:  43.657  on 140  degrees of freedom
AIC: 252.07

Number of Fisher Scoring iterations: 2
```

L’andamento dei residui per questo modello e’ il seguente:



BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

Ci sono miglioramenti rispetto al modello precedente: i valori considerati anomali secondo la distanza di Cook sono spariti e i residui hanno un andamento meno deterministico, anche se non sono dispersi in modo casuale. Osservo anche che la distribuzione normale teorica dei residui(grafico 2) e' piu' o meno rimasta la stessa. Guardiamo il grafico valori stimati vs valori osservati:



I punti piu' concentrati attorno la retta rispetto al modello precedente.

4) Conclusione
AIC del modello senza spline = `303.81`
AIC del modello con spline = `252.07`

Non vado avanti a stimare il test error perche' il modello con spline vincera' sempre sul modello senza spline, in quanto la relazione tra salute.spese e felice non e' lineare, e non lo diventera' se togliamo i valori sospetti mostrati nell'andamento dei residui.
Percui scelgo il modello con la spline quadratica, quello valutato nel punto 2.

BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

5) Risposta alla domanda
Riscriviamo il modello che ho scelto:

```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.0331  -0.3075   0.0738   0.3255   1.3042

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                             2.5766     0.1058  24.358  < 2e-16 ***
equatoreTRUE                           -0.8600     0.1521  -5.656 8.42e-08 ***
ns(salute.spese, 2)1                    2.6852     0.3232   8.309 7.52e-14 ***
ns(salute.spese, 2)2                    0.6255     0.4730   1.322    0.188
equatoreTRUE:ns(salute.spese, 2)1 -34.6514     8.5519  -4.052 8.38e-05 ***
equatoreTRUE:ns(salute.spese, 2)2 -91.0532    19.4320  -4.686 6.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.3118358)

    Null deviance: 116.058  on 145  degrees of freedom
Residual deviance:  43.657  on 140  degrees of freedom
AIC: 252.07

Number of Fisher Scoring iterations: 2
```

Nel mio modello la variabile pilc non c’e’ ma essendo fortemente correlata a salute spese, correlazione=0.9280356, parlare di pilc e’ come parlare di salute spese. Per rispondere alla domanda direi che un elevato pil pro capite e’ associato ad un aumento della felicita’ del paese se e solo se il paese non e’ sull’equatore

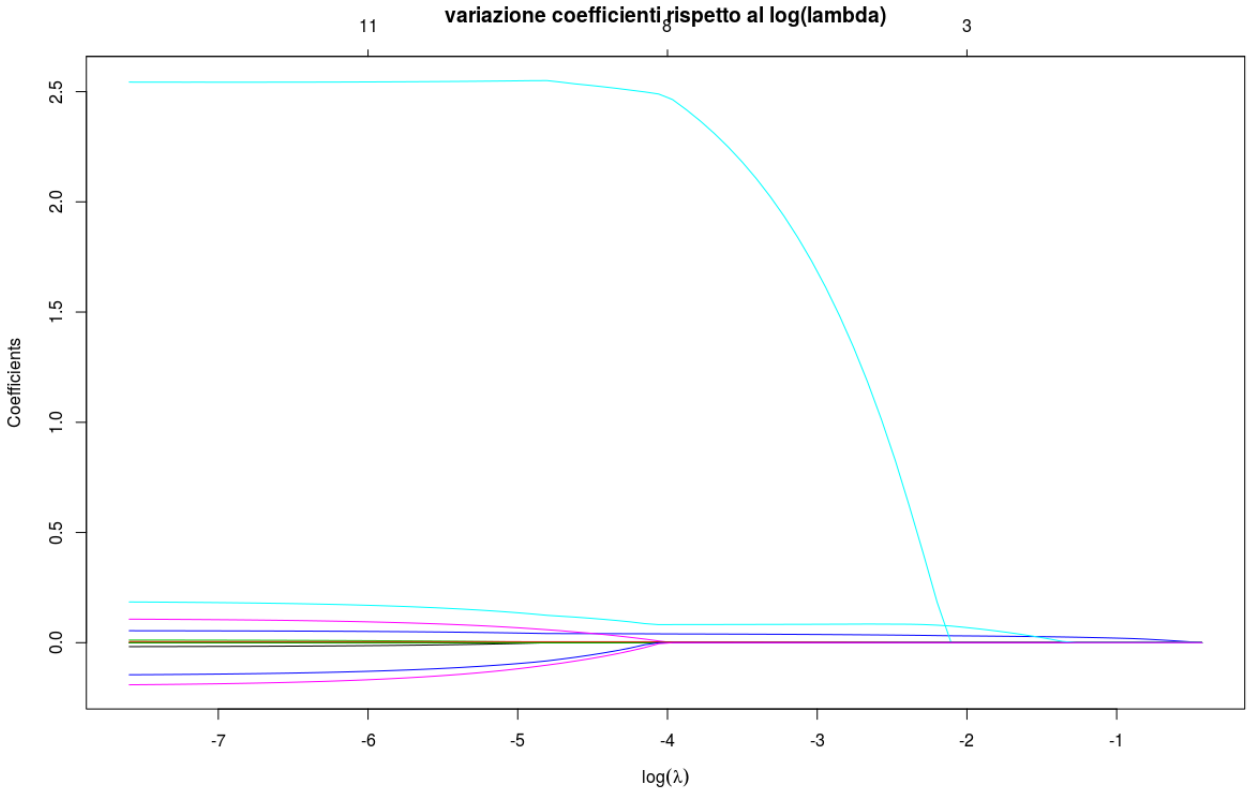
BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

PARTE 2:

Per questa parte saltero' le ispezioni grafiche in quanto faro' l'analisi con lasso per la selezione delle variabili e la stima dei cofficienti.

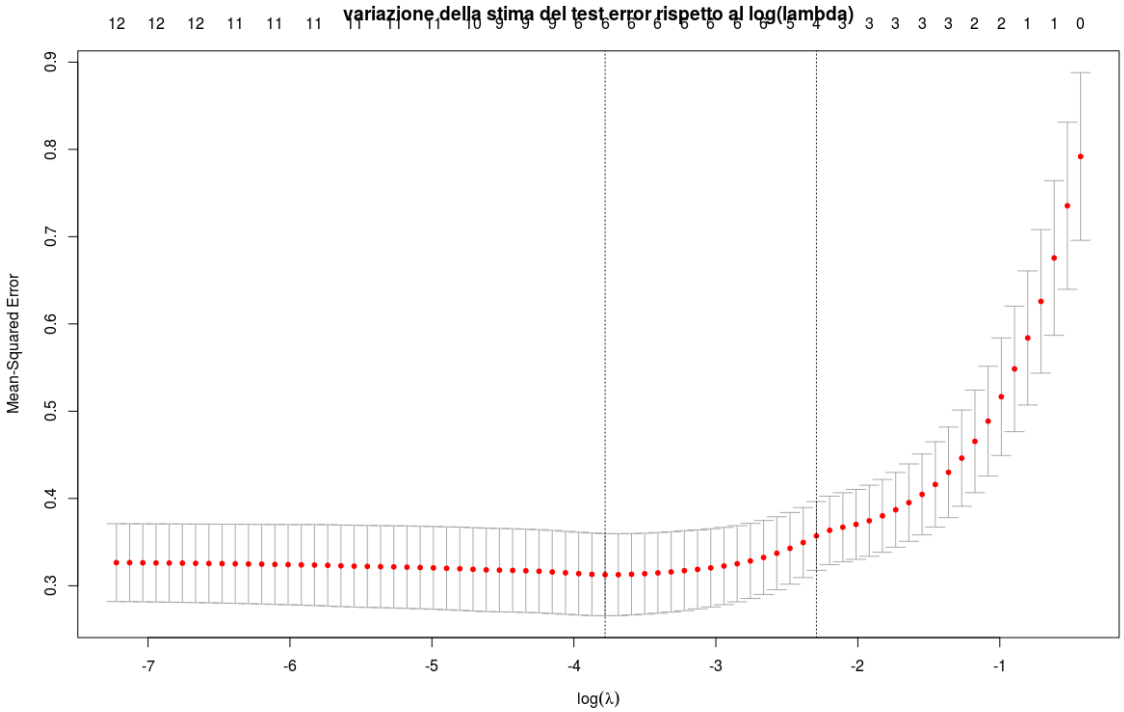
LASSO

1) Variazione dei coefficienti rispetto a lambda



2) Cross validation per trovare il lambda minimo

```
set.seed(223)
```



MSE minimo=0.312606

lambda minimo=0.02281708
lambda 1se=0.1010938

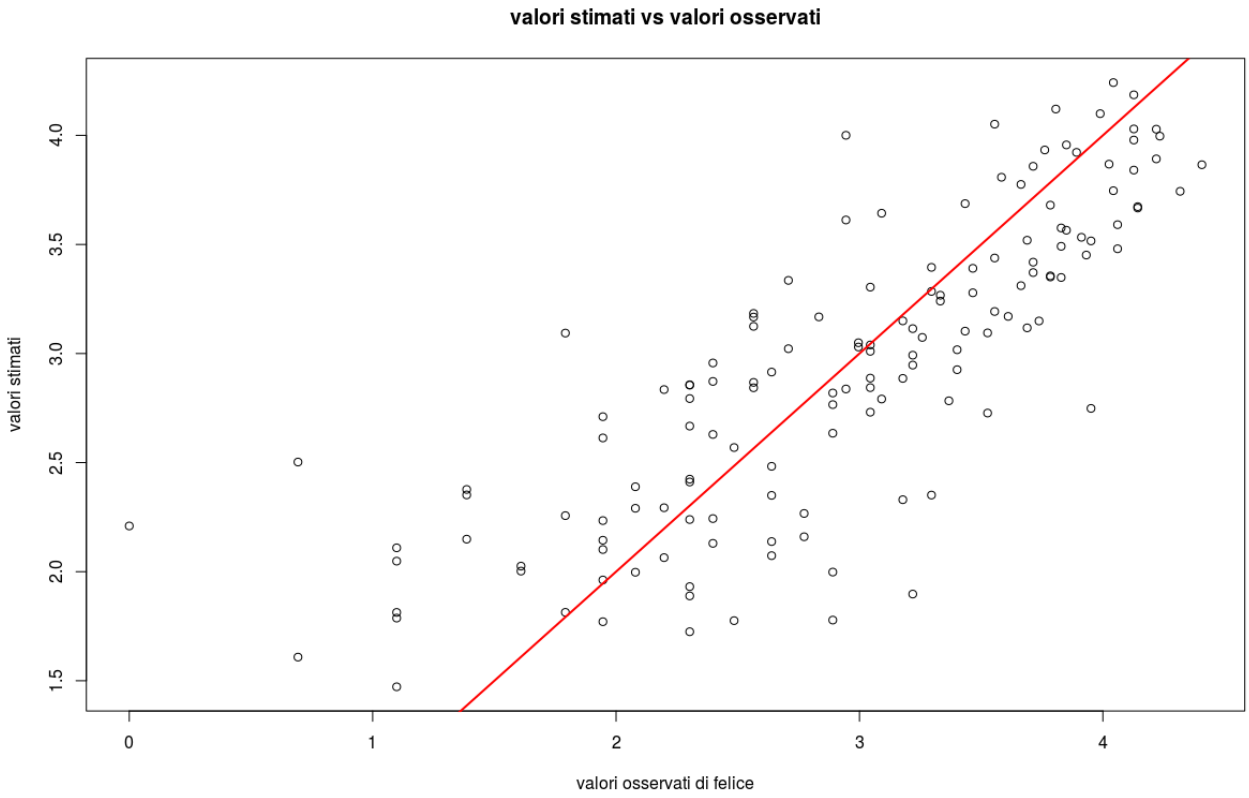
4) Modello lasso con lambda minimo

- Coefficienti

```
      13 x 1 sparse Matrix of class "dgCMatrix"
                           s0
(Intercept)   -6.006814e-01
salute.indice  .
salute.spese   5.616452e-05
istruzione     .
vita           3.958127e-02
reddito.distr  2.366501e+00
equatoreTRUE   .
popden         .
spese.pagate   3.782460e-03
pilc           2.490423e-05
banca          .
democrazia     8.252800e-02
oecdTRUE       .
```

BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

- Grafico valori stimati



Il grafico e' soddisfacente, anche se ci sono ancora dei punti distanti

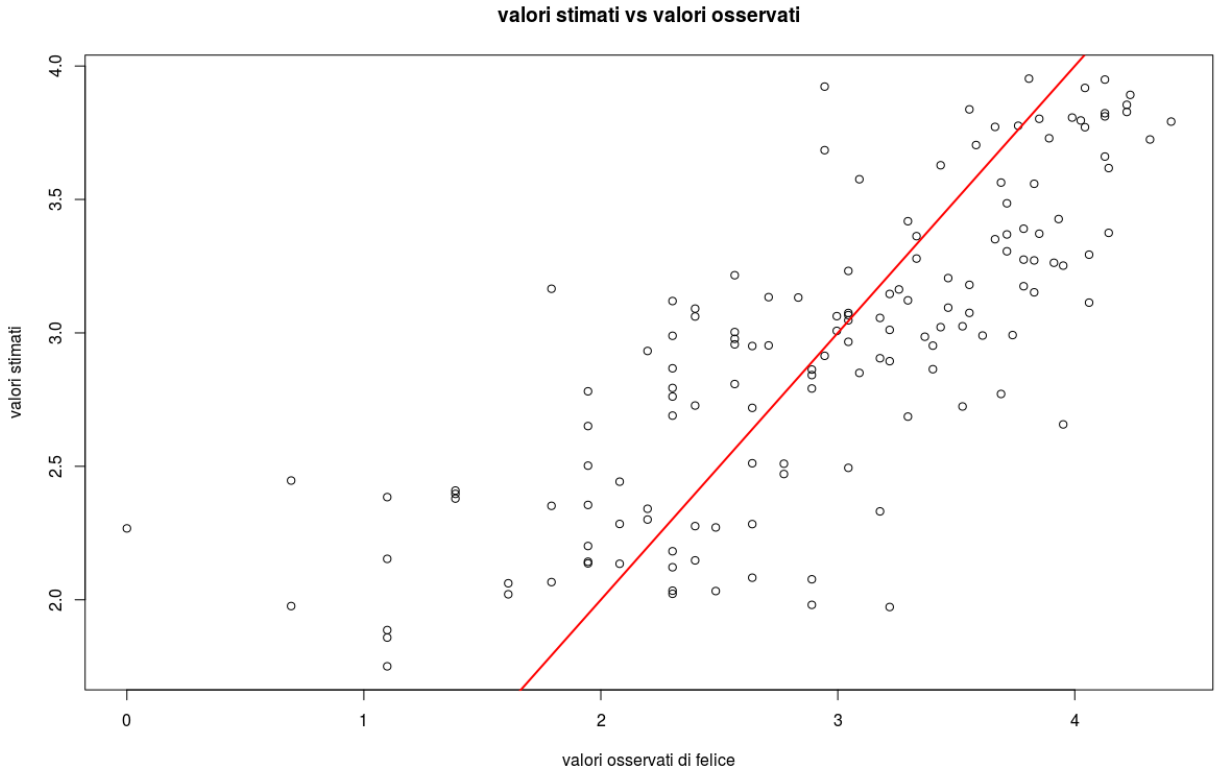
5) Modello con lambda 1se
E' un modello che da' piu' regolarita' al variare del test error

- Coefficienti

```
13 x 1 sparse Matrix of class "dgCMatrix"
                          s0
(Intercept)   -6.006814e-01
salute.indice  .
salute.spese   5.616452e-05
istruzione     .
vita           3.958127e-02
reddito.distr  2.366501e+00
equatoreTRUE   .
popden         .
spese.pagate   3.782460e-03
pilc           2.490423e-05
banca          .
democrazia     8.252800e-02
oecdTRUE       .
```

BILE EZANIN CHRISTIAN PRINCE CARLOS
16/06/2017

- Grafico valori stimati



Il risultato e' meno soddisfacente del risultato dato dal modello con lambda minimo, ma il vantaggio di essere piu' regolare al variare ditest error.