Termini di ordine superiore

Statistica Applicata
Corso di Laurea in Informatica

cristiano.varin@unive.it

Installiamo il pacchetto `Ecdat` che contiene il dataset che analizzeremo

```r
install.packages("Ecdat")
library(Ecdat)
```

Il dataset si chiama `Diamond`

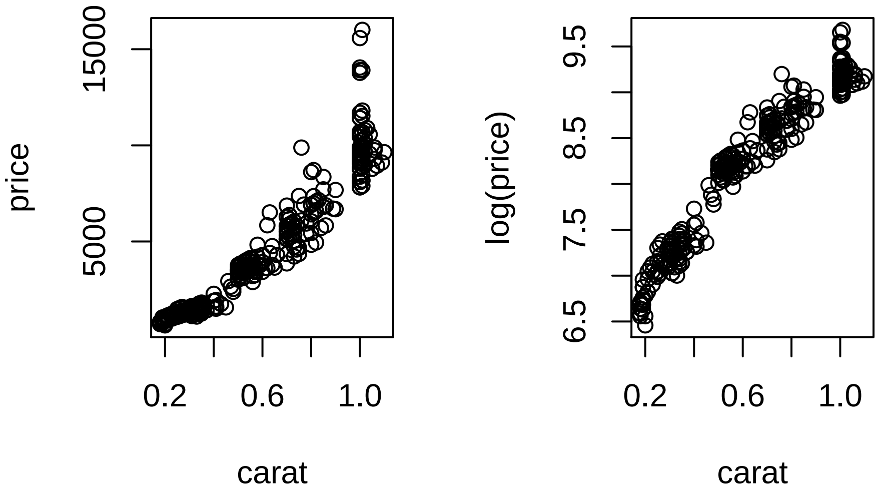
```r
data(Diamond)
```

Il dataset riguarda la relazione fra il prezzo dei diamanti (`price`), il loro peso (`carat`), il loro colore (`colour`), la loro limpidezza (`clarity`) e l'ente che ha certificato la pietra (`certification`)

```r
help(Diamond)
```

Iniziamo con qualche grafico per valutare la relazione fra prezzo e peso

```r
par(mfrow=c(1,2))
plot(price~carat, data=Diamond)
plot(log(price)~carat, data=Diamond)
```



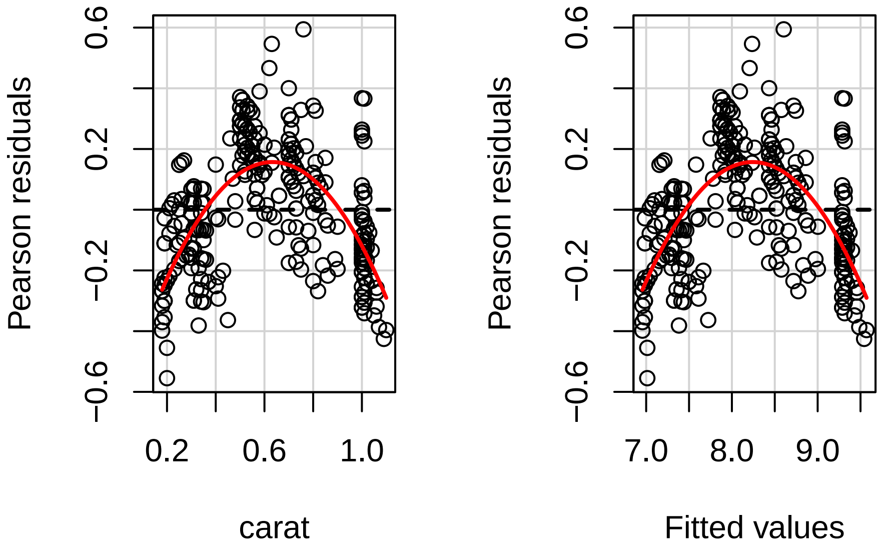
Stimiamo un primo modello di regressione su scala logaritmica

```
modelloA <- lm(log(price)~carat, data=Diamond)
summary(modelloA)

##
## Call:
## lm(formula = log(price) ~ carat, data = Diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5549 -0.1627 -0.0087  0.1552  0.5943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4449     0.0294   219.4   <2e-16 ***
## carat         2.8416     0.0426    66.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.207 on 306 degrees of freedom
## Multiple R-squared:  0.936,Adjusted R-squared:  0.935
## F-statistic: 4.44e+03 on 1 and 306 DF,  p-value: <2e-16
```

Controlliamo i residui

```
library(car)
residualPlots(modelloA)
```



```
##            Test stat Pr(>|t|)
## carat         -15.46        0
## Tukey test    -15.46        0
```

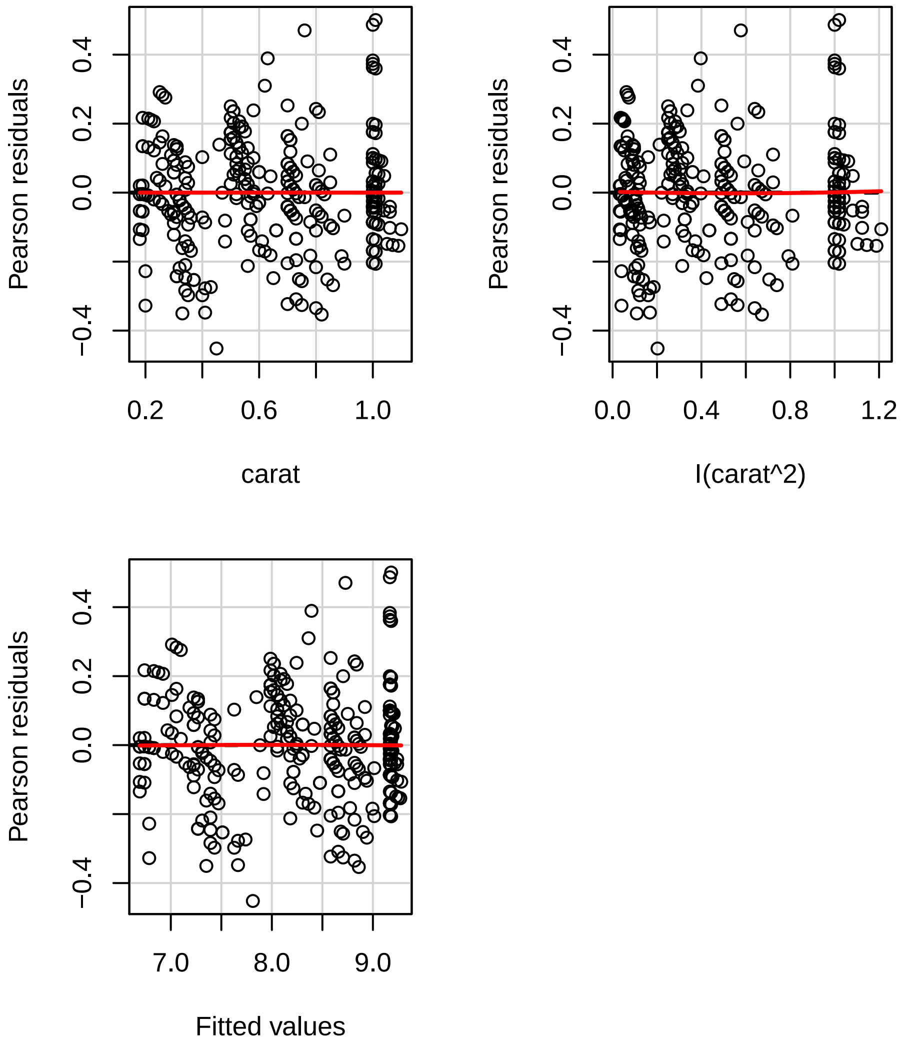
Proviamo ora un modello polinomiale del secondo ordine

```
modelloB <- lm(log(price)~carat+I(carat^2), data=Diamond)
summary(modelloB)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2), data = Diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4519 -0.0886 -0.0044  0.0969  0.5004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7806     0.0483   119.7   <2e-16 ***
## carat         5.4368     0.1709    31.8   <2e-16 ***
## I(carat^2)   -2.0501     0.1326  -15.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.155 on 305 degrees of freedom
## Multiple R-squared:  0.964,Adjusted R-squared:  0.964
## F-statistic: 4.07e+03 on 2 and 305 DF,  p-value: <2e-16
```

Controlliamo i nuovi residui

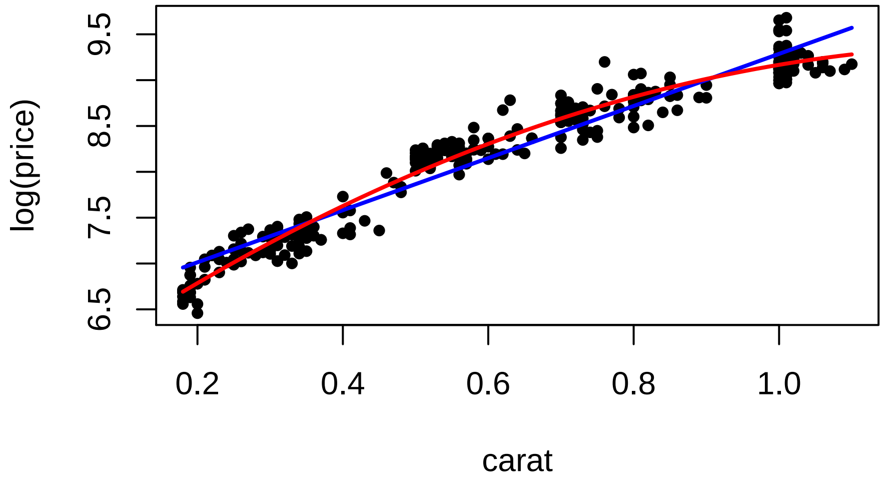
```
residualPlots(modelloB)
```



```
##             Test stat Pr(>|t|)
## carat           0.612    0.541
## I(carat^2)      0.494    0.622
## Tukey test     -0.234    0.815
```

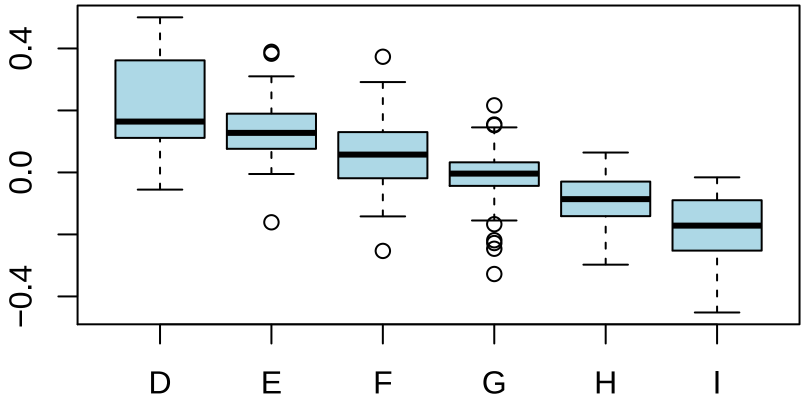
Ora confrontiamo l'adattamento ai dati osservati nei due modelli

```
plot(log(price)~carat, data=Diamond, pch=20)
dati.nuovi <- with(Diamond, seq(min(carat), max(carat), length=50) )
previsioniA <- predict( modelloA, newdata=data.frame( carat=dati.nuovi ) )
lines( dati.nuovi, previsioniA, col="blue", lwd=2 )
previsioniB <- predict( modelloB, newdata=data.frame( carat=dati.nuovi ) )
lines( dati.nuovi, previsioniB, col="red", lwd=2 )
```

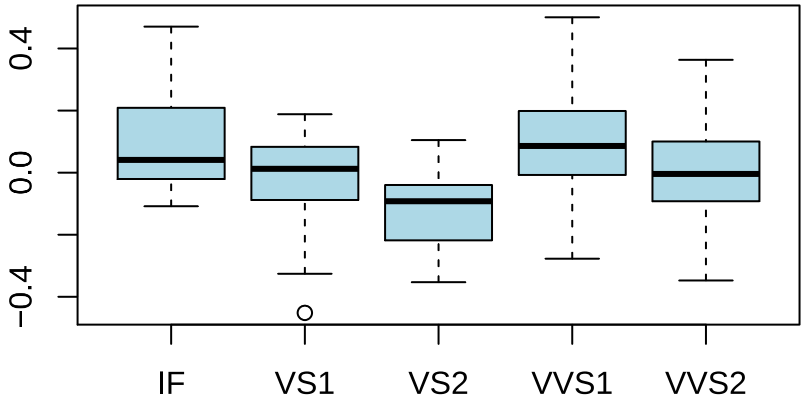


Proviamo ad estendere il modello con le altre variabili

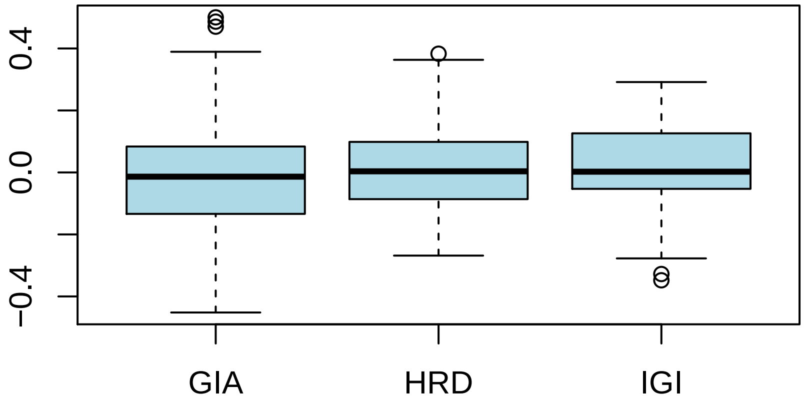
```
resB <- residuals(modelloB)
boxplot(resB~colour, data=Diamond, col="lightblue")
```



```
boxplot(resB~clarity, data=Diamond, col="lightblue")
```



```
boxplot(resB~certification, data=Diamond, col="lightblue")
```



Proviamo aggiungendo una alla volta le tre variabili

```
modelloC <- update(modelloB, .~.+colour)
summary(modelloC)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + colour, data = Diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30465 -0.06140  0.00351  0.06702  0.28783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.9754     0.0419  142.49  < 2e-16 ***
## carat         5.4038     0.1206   44.82  < 2e-16 ***
## I(carat^2)   -1.9832     0.0937  -21.16  < 2e-16 ***
## colourE      -0.0723     0.0319   -2.27    0.024 *
## colourF      -0.1432     0.0298   -4.80  2.5e-06 ***
## colourG      -0.2138     0.0304   -7.03  1.4e-11 ***
## colourH      -0.3048     0.0307   -9.94  < 2e-16 ***
## colourI      -0.3989     0.0323  -12.37  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.109 on 300 degrees of freedom
## Multiple R-squared:  0.983,Adjusted R-squared:  0.982
## F-statistic: 2.41e+03 on 7 and 300 DF,  p-value: <2e-16
```

```
modelloD <- update(modelloB, .~.+clarity)
summary(modelloD)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + clarity, data = Diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4167 -0.0914  0.0081  0.0898  0.3815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.8100     0.0423  137.26  < 2e-16 ***
## carat         5.6343     0.1602   35.16  < 2e-16 ***
## I(carat^2)   -2.1303     0.1218  -17.49  < 2e-16 ***
## clarityVS1   -0.1373     0.0277   -4.96  1.2e-06 ***
## clarityVS2   -0.2537     0.0307   -8.28  4.2e-15 ***
## clarityVVS1  -0.0281     0.0302   -0.93     0.35
## clarityVVS2  -0.1242     0.0281   -4.41  1.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.136 on 301 degrees of freedom
## Multiple R-squared:  0.973,Adjusted R-squared:  0.972
## F-statistic: 1.79e+03 on 6 and 301 DF,  p-value: <2e-16
```

```
modelloE <- update(modelloB, .~.+certification)
summary(modelloE)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + certification,
##     data = Diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4253 -0.0908 -0.0085  0.0928  0.5193
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.6857     0.0635   89.56   <2e-16 ***
## carat               5.6501     0.1984   28.47   <2e-16 ***
## I(carat^2)         -2.1867     0.1454  -15.04   <2e-16 ***
## certificationHRD    0.0311     0.0221    1.41    0.161
## certificationIGI    0.0678     0.0271    2.51    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.154 on 303 degrees of freedom
## Multiple R-squared:  0.965,Adjusted R-squared:  0.964
## F-statistic: 2.07e+03 on 4 and 303 DF,  p-value: <2e-16
```

Il miglior modello è quello con l'aggiunta di `colour`. Proviamo ora ad aggiungere a questo un altro predittore

```
modelloF <- update(modelloC, .~.+clarity)
summary(modelloF)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + colour + clarity,
##     data = Diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15017 -0.04058 -0.00793  0.04528  0.14465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0372     0.0231  261.30  < 2e-16 ***
## carat          5.7441     0.0709   81.06  < 2e-16 ***
## I(carat^2)    -2.1500     0.0539  -39.92  < 2e-16 ***
## colourE       -0.0795     0.0174   -4.56  7.6e-06 ***
## colourF       -0.1572     0.0164   -9.60  < 2e-16 ***
## colourG       -0.2461     0.0168  -14.67  < 2e-16 ***
## colourH       -0.3385     0.0170  -19.93  < 2e-16 ***
## colourI       -0.4428     0.0178  -24.84  < 2e-16 ***
## clarityVS1    -0.2336     0.0125  -18.74  < 2e-16 ***
## clarityVS2    -0.3098     0.0136  -22.86  < 2e-16 ***
## clarityVVS1   -0.0899     0.0134   -6.69  1.1e-10 ***
## clarityVVS2  -0.1718     0.0125  -13.79  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0595 on 296 degrees of freedom
## Multiple R-squared:  0.995,Adjusted R-squared:  0.995
## F-statistic: 5.2e+03 on 11 and 296 DF,  p-value: <2e-16
```

```
modelloG <- update(modelloC, .~.+certification)
summary(modelloG)
```

```
##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + colour + certification,
##     data = Diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31334 -0.07063 -0.00238  0.06412  0.31287
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.8660     0.0489  119.91  < 2e-16 ***
## carat               5.6675     0.1353   41.90  < 2e-16 ***
## I(carat^2)         -2.1534     0.0994  -21.67  < 2e-16 ***
## colourE            -0.0763     0.0307   -2.49   0.0134 *
## colourF            -0.1524     0.0288   -5.29  2.3e-07 ***
## colourG            -0.2272     0.0294   -7.72  1.8e-13 ***
## colourH            -0.3135     0.0296  -10.57  < 2e-16 ***
## colourI            -0.4082     0.0311  -13.13  < 2e-16 ***
## certificationHRD    0.0425     0.0152    2.80   0.0055 **
## certificationIGI    0.0855     0.0185    4.62  5.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.105 on 298 degrees of freedom
## Multiple R-squared:  0.984,Adjusted R-squared:  0.983
## F-statistic: 2.03e+03 on 9 and 298 DF,  p-value: <2e-16
```

Il modello F è il migliore, ora proviamo ad aggiungere anche `certification`

```
modelloH <- update(modelloF, .~.+certification)
summary(modelloH)

##
## Call:
## lm(formula = log(price) ~ carat + I(carat^2) + colour + clarity +
##     certification, data = Diamond)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15411 -0.04120 -0.00911  0.04543  0.14158
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.07535    0.02920  208.05  < 2e-16 ***
## carat               5.67062    0.07928   71.52  < 2e-16 ***
## I(carat^2)         -2.10292    0.05802  -36.24  < 2e-16 ***
## colourE            -0.07925    0.01739   -4.56  7.6e-06 ***
## colourF            -0.15599    0.01633   -9.55  < 2e-16 ***
## colourG            -0.24503    0.01673  -14.64  < 2e-16 ***
## colourH            -0.33910    0.01697  -19.98  < 2e-16 ***
## colourI            -0.44261    0.01774  -24.95  < 2e-16 ***
## clarityVS1         -0.24447    0.01336  -18.30  < 2e-16 ***
## clarityVS2         -0.32018    0.01428  -22.42  < 2e-16 ***
## clarityVVS1        -0.09401    0.01357   -6.93  2.7e-11 ***
## clarityVVS2        -0.17670    0.01259  -14.03  < 2e-16 ***
## certificationHRD -0.00622    0.00894   -0.70    0.487
## certificationIGI -0.02541    0.01154   -2.20    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0592 on 294 degrees of freedom
## Multiple R-squared:  0.995,Adjusted R-squared:  0.995
## F-statistic: 4.45e+03 on 13 and 294 DF,  p-value: <2e-16
```

Tabella dell'analisi della devianza

```
anova(modelloH)

## Analysis of Variance Table
##
## Response: log(price)
##                 Df Sum Sq Mean Sq  F value Pr(>F)    
## carat            1  190.5   190.5 54339.61 <2e-16 ***
## I(carat^2)       1    5.8     5.8  1644.51 <2e-16 ***
## colour           5    3.8     0.8   217.31 <2e-16 ***
## clarity          4    2.5     0.6   178.54 <2e-16 ***
## certification    2    0.0     0.0     2.45  0.088 .  
## Residuals      294    1.0     0.0                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```