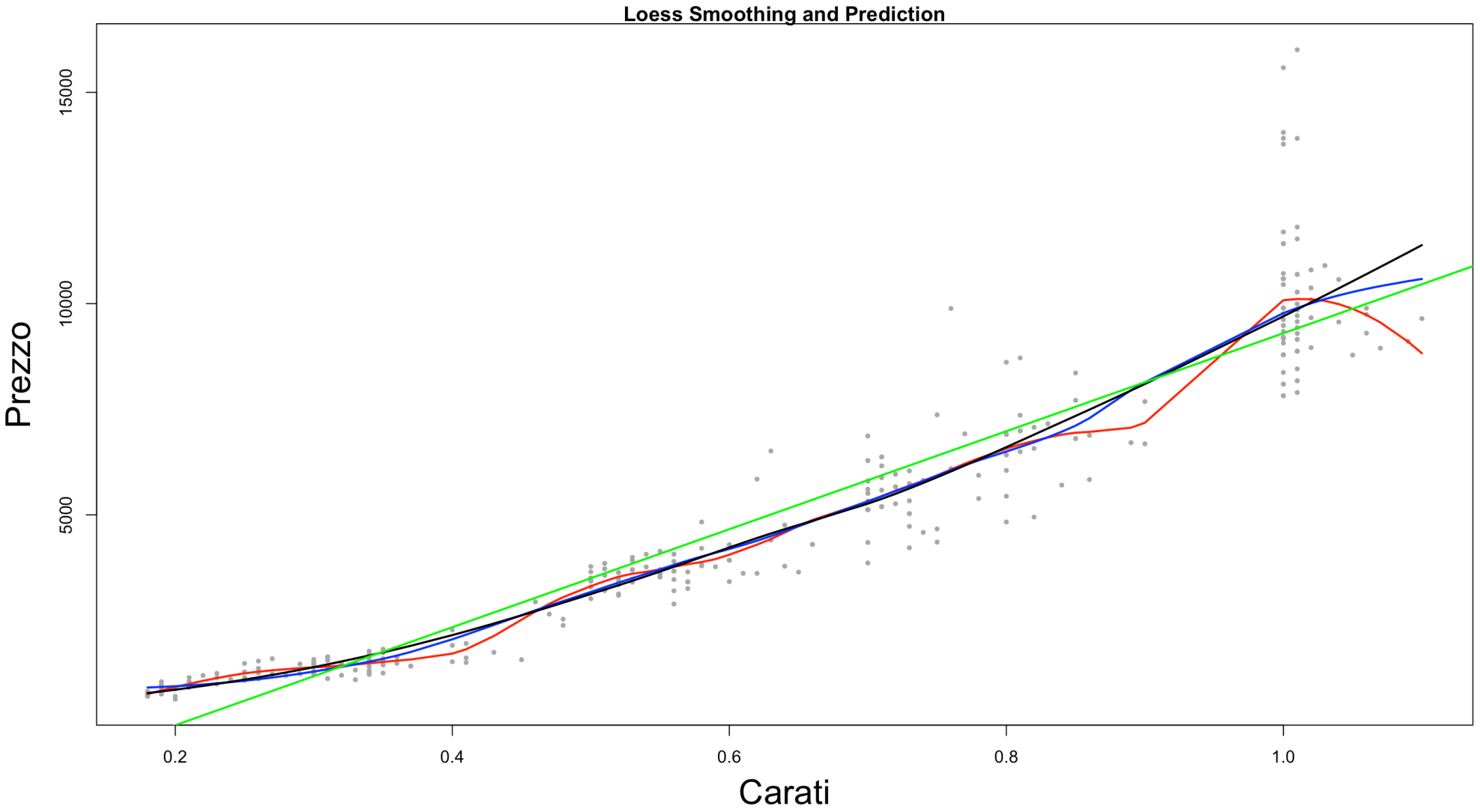
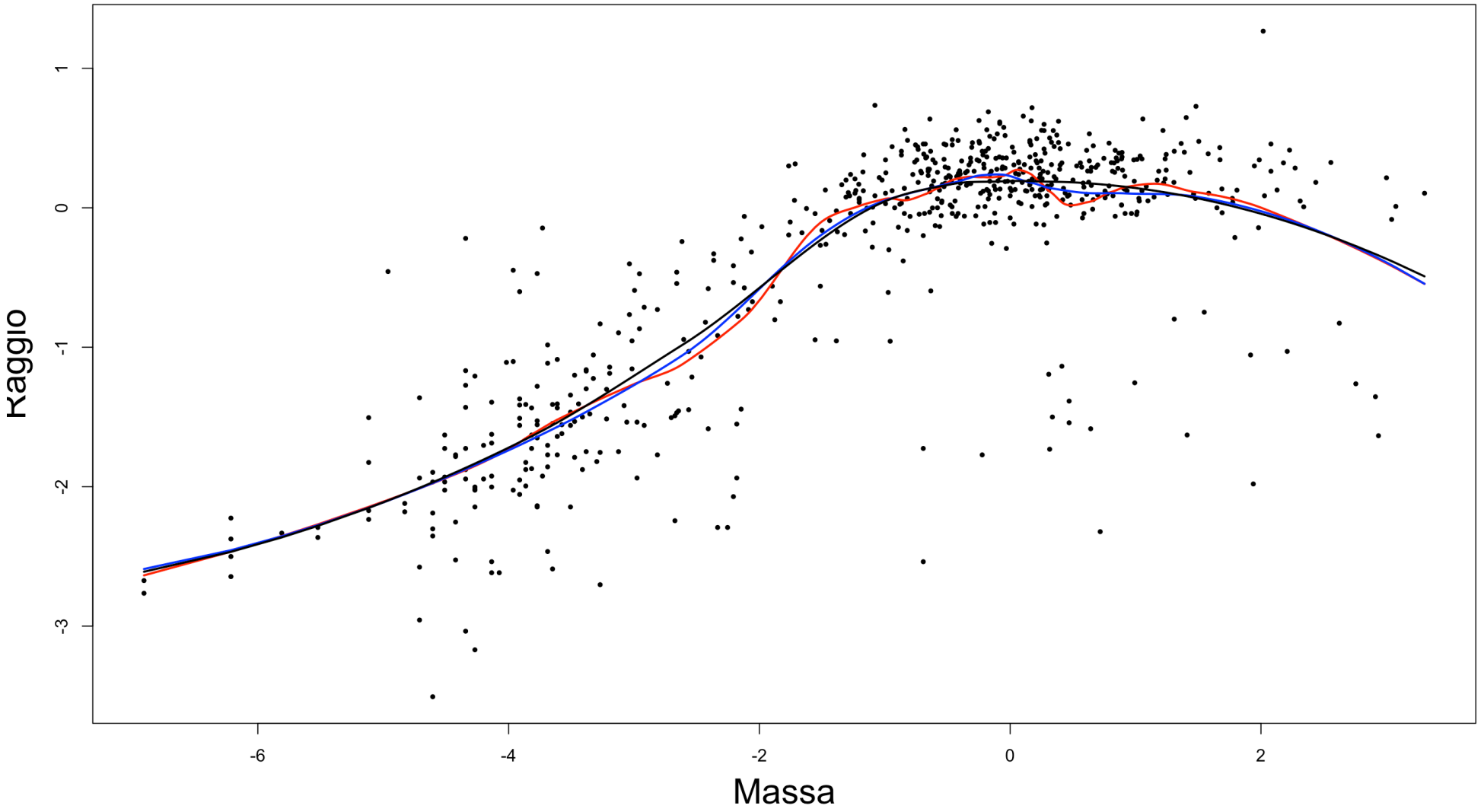
Esempio: prezzo dei diamanti



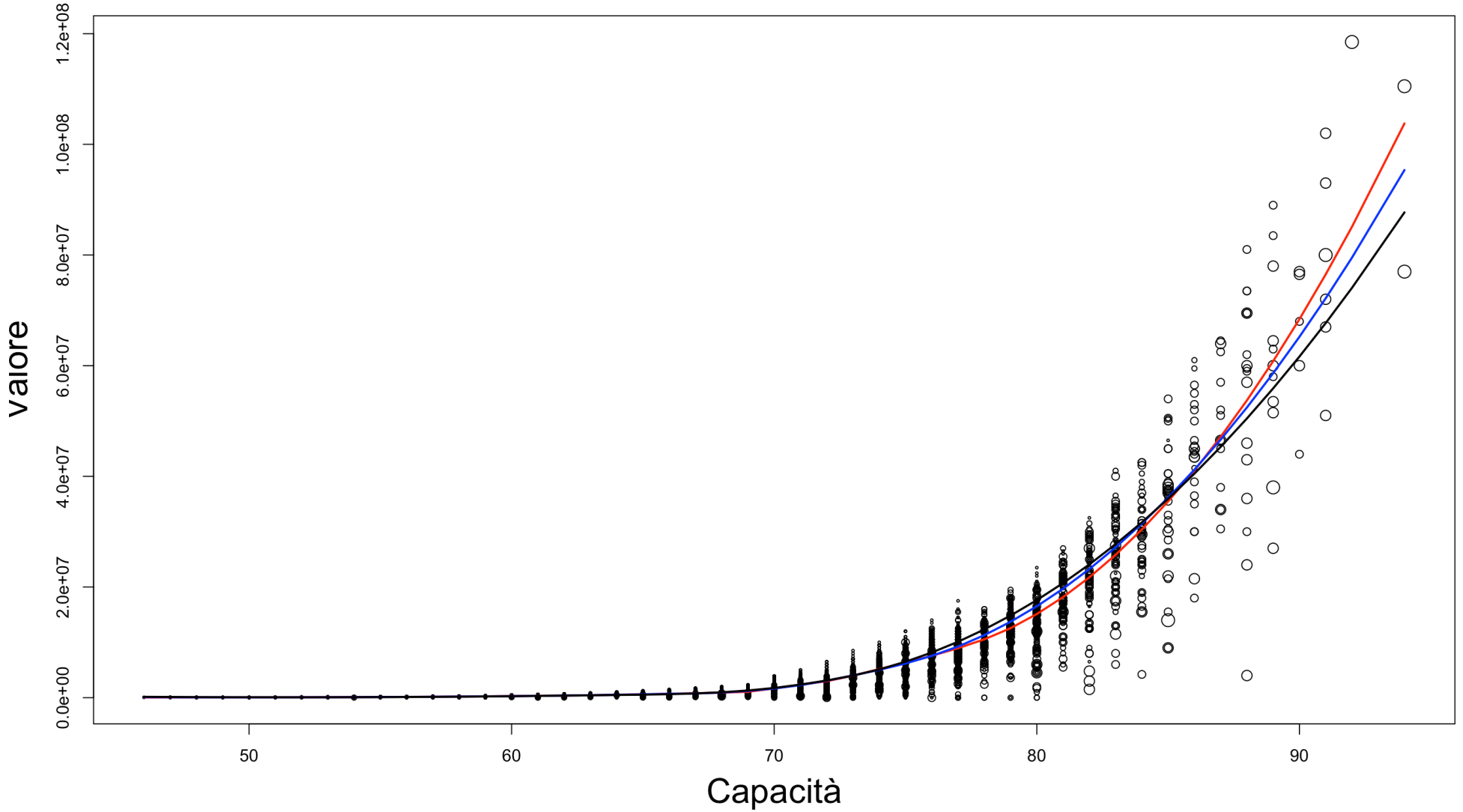
Esempio: prezzo dei diamanti

```
Diamond <- Diamond[order(Diamond$carat),]
loess25 <- loess(price ~ carat, data=Diamond, span=0.25)
loess50 <- loess(price ~ carat, data=Diamond, span=0.50)
loess75 <- loess(price ~ carat, data=Diamond, span=0.75)
smoothed25 <- predict(loess25)
smoothed50 <- predict(loess50)
smoothed75 <- predict(loess75)
plot(Diamond$carat, Diamond$price, main="Loess Smoothing and Prediction",
xlab="Carati",
ylab="Prezzo",xlim=caratlims,cex=.5,pch=19,col="darkgrey",cex.lab=2)
lines(smoothed25, x=Diamond$carat, col="red",lwd=2)
lines(smoothed50, x=Diamond$carat, col="blue",lwd=2)
lines(smoothed75, x=Diamond$carat, col="black",lwd=2)
abline(lreg,lwd=2,col="green")
```

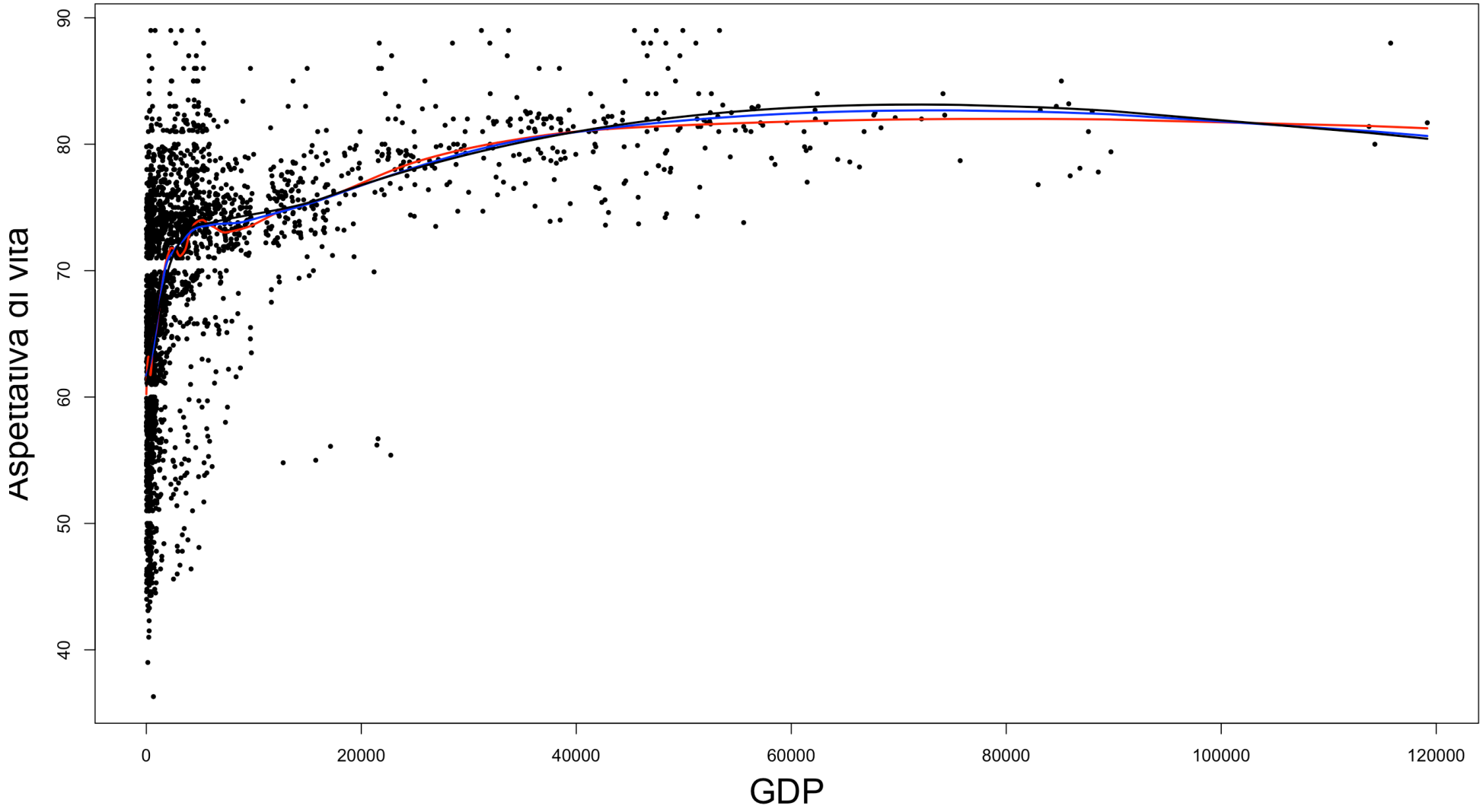
Esempio: Massa e Raggio degli esopianeti



Esempio: valore dei calciatori



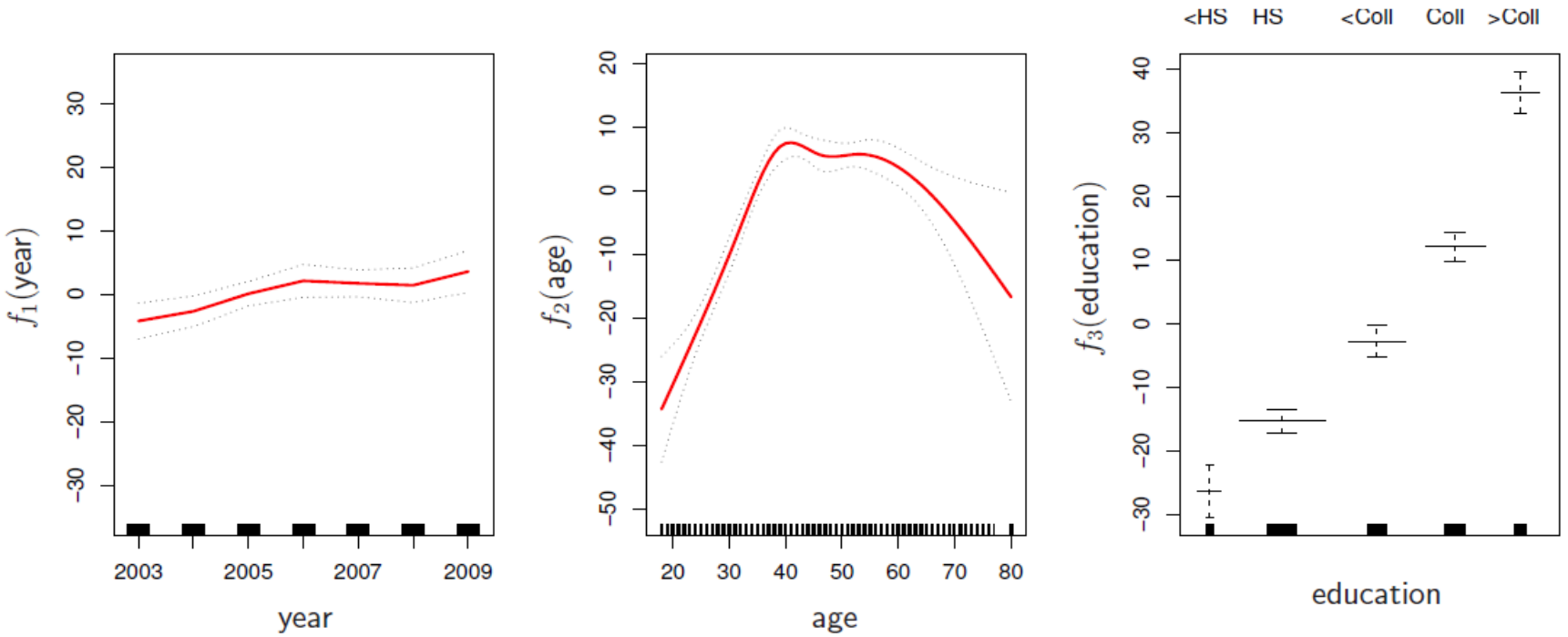
Esempio: curva di Preston



Modelli additivi generalizzati (GAM)

Questo tipo di modelli permette di gestire la non linearità nelle variabili, ma conserva la struttura additiva dei modelli lineari. Infatti, la forma funzionale del modello sarà:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$



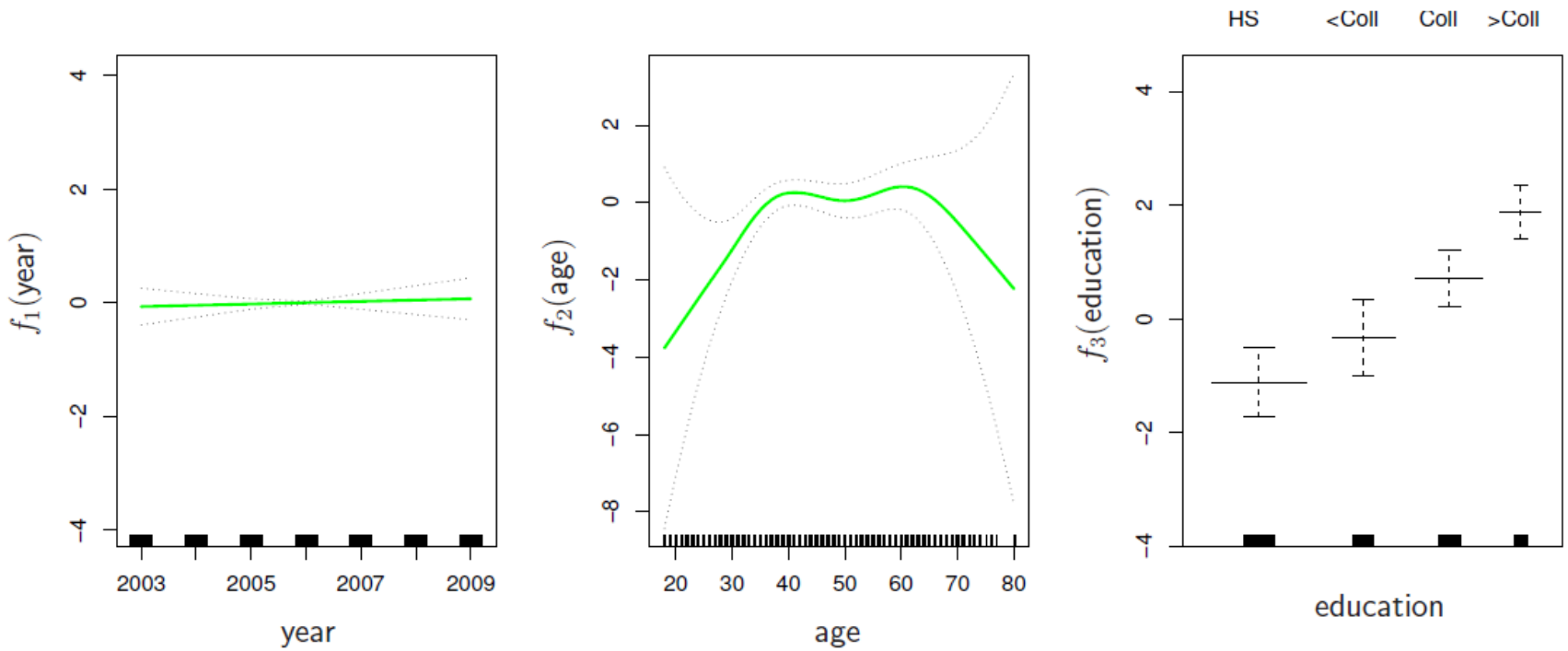
Caratteristiche del GAM

- Dove ogni termine del modello è una funzione della variabile predittiva di riferimento;
- Si possono utilizzare, per questo tipo di modello anche gli splines:
  - lm(wage~ns(year,df = 5) + ns(age,df = 5) + education);
- Ciò che più interessa sono le forme funzionali, e non i coefficienti. Il plot precedente è stato creato con la funzione plot.gam;
- Dal grafico si nota che è possibile utilizzare sia termini lineari, che termini non lineari. Per comparare i modelli si utilizza la funzione anova();
- È possibile utilizzare anche «smoothing splines» oppure regressioni locali:
  - gam(wage~s(year,df = 5) + lo(age,span = .5) + education)
- I GAMs hanno forma additiva, sebbene interazioni di piccolo ordine possono essere inclusi in maniera naturale, utilizzando forme del tipo:
  - ns(age,df = 5): ns(year, df = 5).

GAMs per classificazione

$$\log\left(\frac{\rho(x)}{1-\rho(x)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

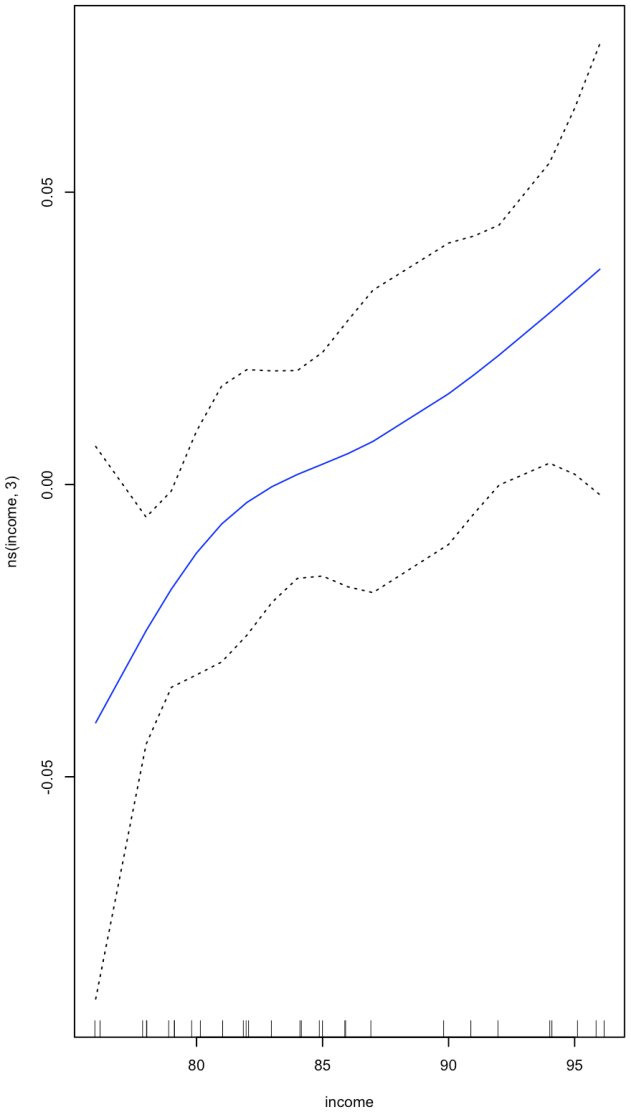
gam(I(wage > 250) ~ year + s(age, df = 5) + education, family = binomial)

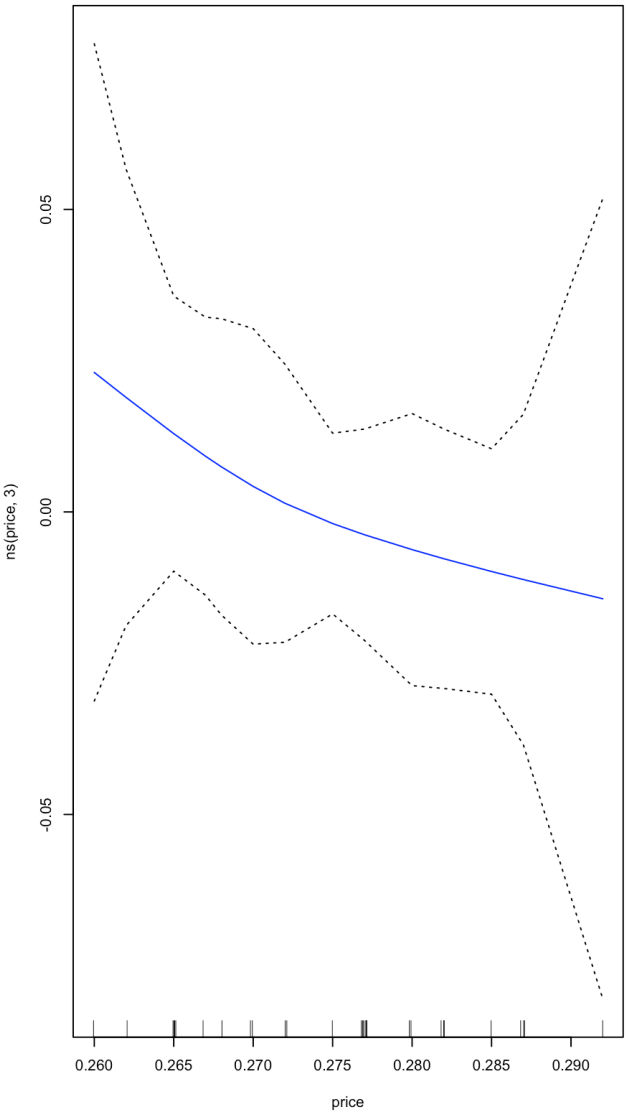


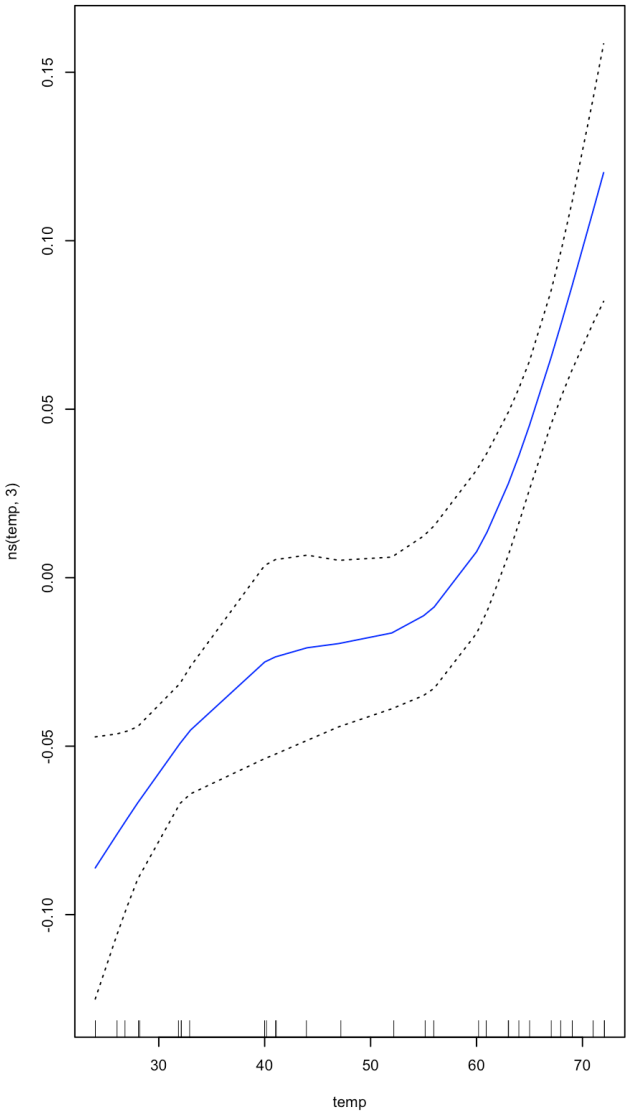
Esempio: consumo di gelati

```
gamlm <- lm(cons~ns(income,3)+ns(price,3)+ns(temp,3),data=Icecream)
library(gam)
gamice <- gam(cons~ns(income,3)+ns(price,3)+ns(temp,3),data=Icecream)
par(mfrow=c(1,3))
plot(gamice, se=TRUE,col="blue")
summary(gamice)

Null Deviance: 0.1255 on 29 degrees of freedom
Residual Deviance: 0.0247 on 20 degrees of freedom
AIC: -105.9169
Number of Local Scoring Iterations: 2
Anova for Parametric Effects
               Df   Sum Sq   Mean Sq F value    Pr(>F)
ns(income, 3)   3 0.000396 0.0001319  0.1067  0.955170
ns(price, 3)    3 0.022089 0.0073631  5.9598  0.004489 **
ns(temp, 3)     3 0.078329 0.0261098 21.1336 2.058e-06 ***
Residuals      20 0.024709 0.0012355
```

Esempio: consumo di gelati



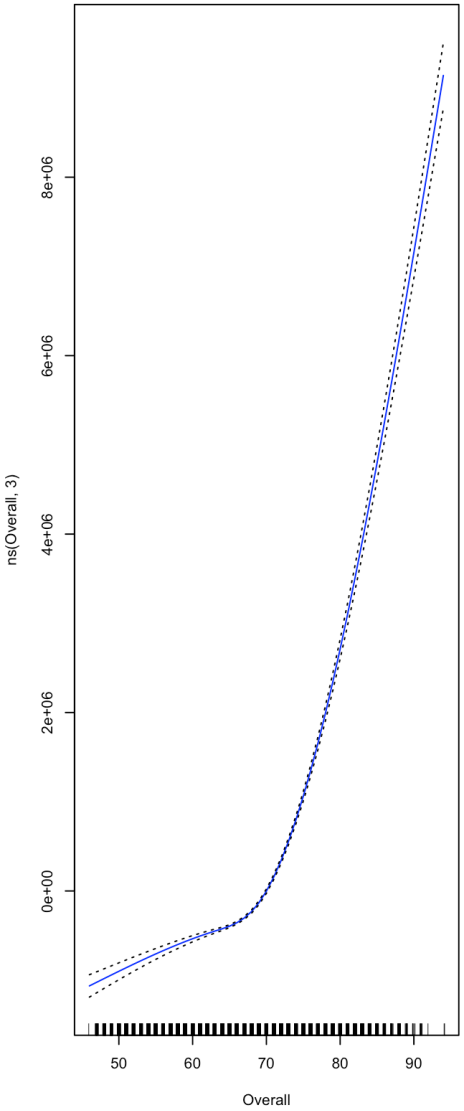


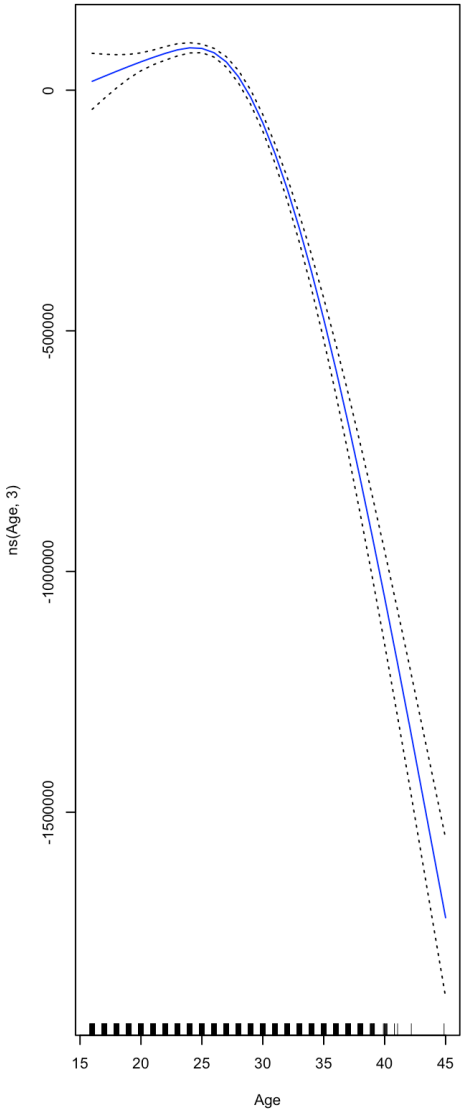


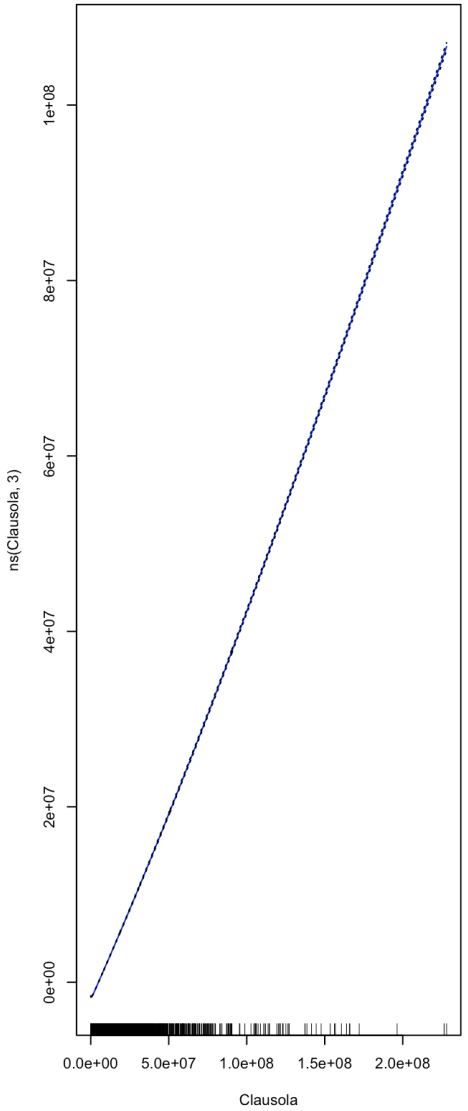
Esempio: valore dei calciatori

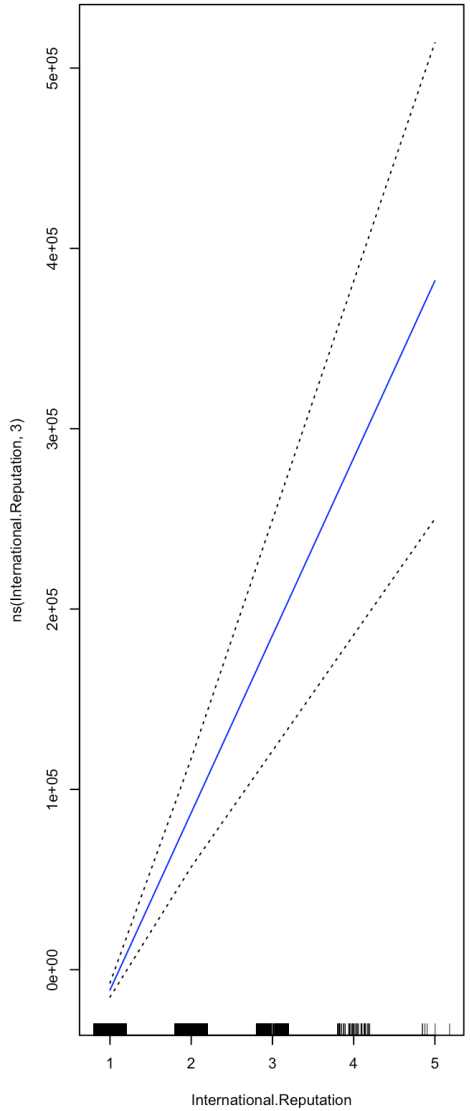
```
gamlm <- lm(Valore~ns(Overall,3)+ns(Age,3)+ns(Clausola,3)+International.Reputation,data=fifa)
gamfifa <- gam(Valore~ns(Overall,3)+ns(Age,3)+ns(Clausola,3)+International.Reputation,data=fifa)
par(mfrow=c(1,4))
plot(gamfifa, se=TRUE,col="blue")
summary(gamfifa)

                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                -338063      37924  -8.914  < 2e-16 ***
ns(Overall, 3)1             659227      68652   9.602  < 2e-16 ***
ns(Overall, 3)2            6527416     176866  36.906  < 2e-16 ***
ns(Overall, 3)3            9934548     201578  49.284  < 2e-16 ***
ns(Age, 3)1                  95676      29236   3.273  0.00107 **
ns(Age, 3)2                -871467      87457  -9.965  < 2e-16 ***
ns(Age, 3)3               -1766183      88233 -20.017  < 2e-16 ***
ns(Clausola, 3)1          30488539     177926 171.355  < 2e-16 ***
ns(Clausola, 3)2          69520402     182492 380.950  < 2e-16 ***
ns(Clausola, 3)3         108274117     228714 473.404  < 2e-16 ***
International.Reputation     98329      16991   5.787 7.29e-09 ***
Multiple R-squared:  0.9911,  Adjusted R-squared:  0.9911

Null Deviance: 5.446193e+17 on 16642 degrees of freedom
Residual Deviance: 4.842421e+15 on 16632 degrees of freedom
AIC: 486570.8 - 1564 observations deleted due to missingness
Number of Local Scoring Iterations: 2
Anova for Param. Effects    Df     Sum Sq    Mean Sq    F value    Pr(>F)
ns(Overall, 3)               3 4.6323e+17 1.5441e+17 530342.957 < 2.2e-16 ***
ns(Age, 3)                   3 7.4879e+15 2.4960e+15   8572.718 < 2.2e-16 ***
ns(Clausola, 3)              3 6.9050e+16 2.3017e+16  79053.946 < 2.2e-16 ***
International.Reputation     1 9.7509e+12 9.7509e+12     33.491 7.289e-09 ***
Residuals                16632 4.8424e+15 2.9115e+11
```

Esempio: valore dei calciatori







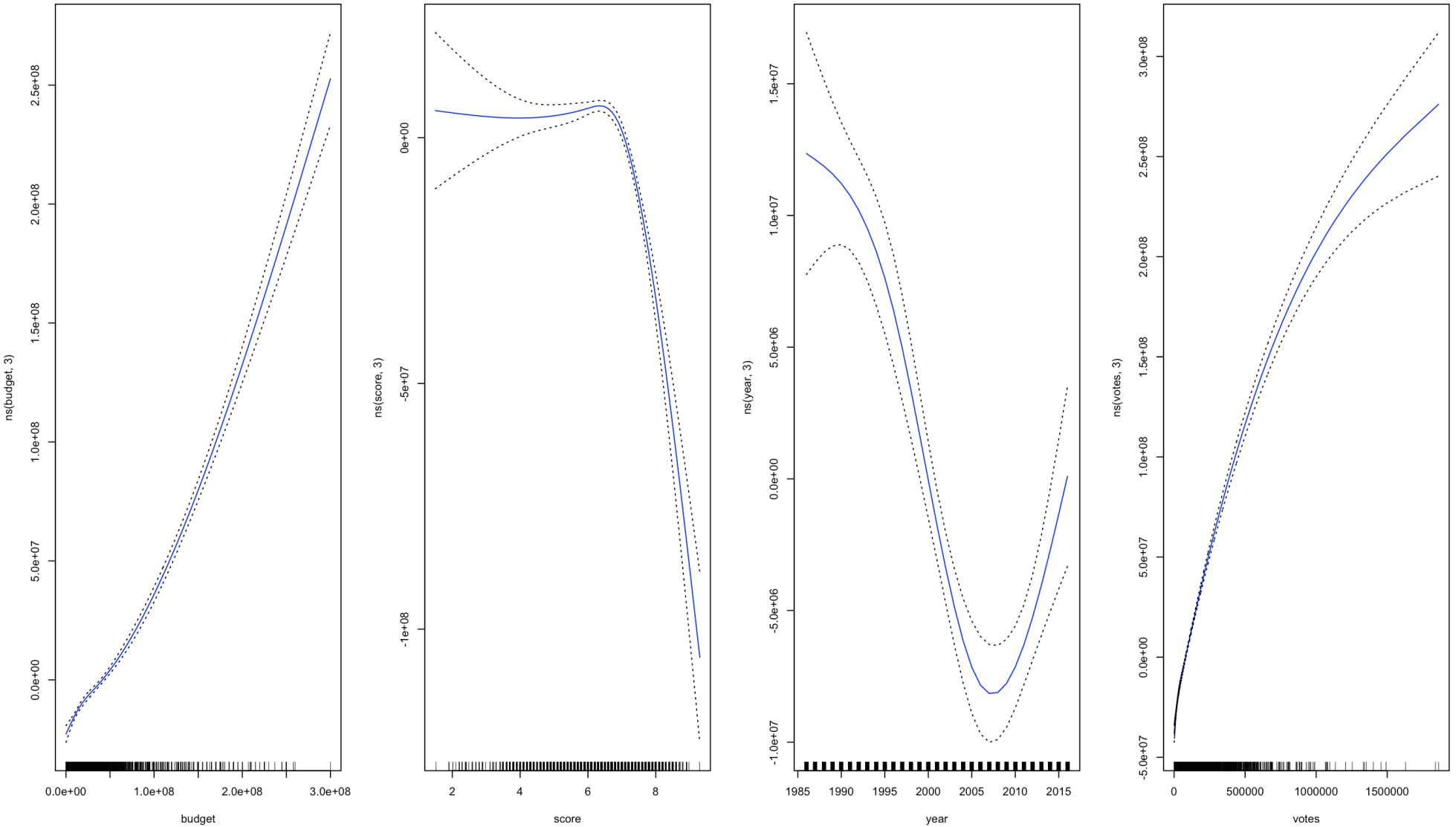


Esempio: incassi al box-office dei film

```
setwd("~/Documents/prova r pkg/lucidi del corso/movies")
revfilm <- read.csv("movies.csv")
revfilm$budget[revfilm$budget == 0] <- NA
regfilm <- lm(gross ~ ns(budget,3) + ns(score,3) + ns(year,3) + ns(votes,3),data=revfilm)
library(gam)
gamfifa <- gam(gross ~ ns(budget,3) + ns(score,3) + ns(year,3) + ns(votes,3),data=revfilm)
par(mfrow=c(1,4))
plot(gamfifa, se=TRUE,col="blue")
summary(gamfifa)
```

```
                Estimate Std. Error t value Pr(>|t|)    AIC: 175504.8
(Intercept)     -16184841   10614303  -1.525 0.127374    2182 observations deleted due to missingness
ns(budget, 5)1   11230904    2922775   3.843 0.000123 ***Number of Local Scoring Iterations: 2
ns(budget, 5)2   19233241    3379281   5.692 1.34e-08 ***Anova for Parametric Effects
ns(budget, 5)3   54903118    5510808   9.963  < 2e-16 ***                 Df     Sum Sq    Mean Sq  F value    Pr(>F)
ns(budget, 5)4  173321272    7772042  22.301  < 2e-16 ***ns(budget, 3)  3 9.5392e+18 3.1797e+18 2005.830 < 2.2e-16 ***
ns(budget, 5)5  269418666   11031485  24.423  < 2e-16 ***ns(score, 3)   3 9.3944e+17 3.1315e+17  197.538 < 2.2e-16 ***
ns(score, 5)1    14321855    9079660   1.577 0.114782    ns(year, 3)    3 3.0499e+16 1.0166e+16    6.413 0.0002484 ***
ns(score, 5)2     2955757   10052106   0.294 0.768738    ns(votes, 3)   3 2.5380e+18 8.4600e+17  533.669 < 2.2e-16 ***
ns(score, 5)3    22205538    6420680   3.458 0.000548 ***Residuals   4625 7.3318e+18 1.5853e+15
ns(score, 5)4   -85299874   22168973  -3.848 0.000121 ***
ns(score, 5)5  -210184554   15607370 -13.467  < 2e-16 ***
ns(year, 5)1    -16111672    3548985  -4.540 5.77e-06 ***
ns(year, 5)2     -7248466    4280047  -1.694 0.090419 .
ns(year, 5)3    -34660289    3293845 -10.523  < 2e-16 ***
ns(year, 5)4    -10507927    7003874  -1.500 0.133605
ns(year, 5)5     -7595549    2690119  -2.823 0.004771 **
ns(votes, 5)1    24254552    3923245   6.182 6.86e-10 ***
ns(votes, 5)2    34521291    4349362   7.937 2.58e-15 ***
ns(votes, 5)3   239802807    7673023  31.253  < 2e-16 ***
ns(votes, 5)4   346987922   13980108  24.820  < 2e-16 ***
ns(votes, 5)5   410386151   21843117  18.788  < 2e-16 ***
Multiple R-squared:  0.6482, Adjusted R-squared:  0.6467
```

Esempio: incassi al box-office dei film



Introduzione alla robustezza

La statistica parametrica, sia essa frequentista o bayesiana, ha come obiettivo primario quello di trovare procedure che, sotto un dato modello stocastico, siano ottime secondo qualche criterio. Tuttavia nulla viene detto sul comportamento di queste ultime, quando il modello ipotizzato è solo approssimativamente valido. L'obiettivo della statistica robusta è quello di predisporre strumenti per valutare la bontà delle procedure statistiche in intorni di modelli stocastici, e quindi di trovare procedure che mantengano buone proprietà anche quando il modello ipotizzato è solo un'approssimazione del "vero" modello. Vista in quest'ottica la statistica robusta può essere definita come la statistica dei modelli parametrici approssimati.

Introduzione alla robustezza

La necessità di una statistica robusta è brillantemente mostrata da Tukey nel suo studio sulla distribuzione normale contaminata. Confrontando, infatti, stimatori di posizione e di scala applicati a campioni provenienti da una popolazione normale N(0,1) e dalle popolazioni contaminate

- $(1 - \eta)N(0, 1) + \eta N(0, 9)$ Modello simmetrico (1)
- $(1 - \eta)N(0, 1) + \eta N(2, 9)$ Modello asimmetrico (2)

con $\eta \in (0, 1)$, si giunge ai sorprendenti risultati. Se per la stima della media di una normale l'utilizzo della mediana campionaria in luogo della media campionaria porta ad una perdita d'efficienza del 36% è sufficiente il 10% di contaminazione nel modello (1) ed il 6% nel modello (2) per rendere le due stime ugualmente efficienti.

Introduzione alla robustezza

η	0.0000	0.0018	0.0282	0.1006
Var(media)	1.0000	1.0140	1.2252	1.8047
Var(mediana)	1.5708	1.5745	1.6315	*1.8047*
C.V.(deviazione standard)	0.7071	0.7627	1.1725	1.3540
C.V.(media scarti assoluti)	0.7555	*0.7627*	0.8514	0.9822
C.V.(mediana scarti assoluti)	1.1664	1.1668	*1.1725*	1.1899

Tabella 1: Modello simmetrico: $(1-\eta)N(0,1)+\eta N(0,9)$

η	0.0000	0.0008	0.0115	0.0617
Var(media)	1.0000	1.0097	1.1377	1.7254
Var(mediana)	1.5708	1.5727	1.5977	*1.7254*
C.V.(deviazione standard)	0.7071	0.7611	1.1694	1.5176
C.V.(media scarti assoluti)	0.7555	*0.7611*	0.8263	0.9973
C.V.(mediana scarti assoluti)	1.1664	1.1666	*1.1694*	1.1838

Tabella 2: Modello asimmetrico: $(1-\eta)N(0,1)+\eta N(2,9)$

Introduzione alla robustezza

Per quanto riguarda la stima del parametro di scala si nota che è sufficiente lo 0.2% di contaminazione nel modello (1) e lo 0.1% nel modello (2) per rendere la media degli scarti assoluti dalla media più efficiente della devizione standard campionaria (normalmente più efficiente del 12%).

Tukey commenta cos`i i suoi risultati:

“Una tacita speranza nell'ignorare le deviazioni dai modelli ideali era che non avrebbero avuto importanza; che le procedure statistiche che erano ottimali secondo il modello rigoroso sarebbero ancora approssimativamente ottimali sotto il modello approssimativo. Sfortunatamente, si è scoperto che questa speranza era spesso drasticamente sbagliata; anche le lievi deviazioni hanno spesso effetti molto più grandi di quanto non fossero stati anticipati dalla maggior parte degli statistici.”

Outlier, leverage, influence

Un outlier è un punto che non è ben stimato dal modello. Se il residuo associato al punto i è " grande", allora possiamo considerare il punto i come un potenziale outlier. L'i.esimo residuo dovrebbe seguire una distribuzione t di Student con (n-p-1) gradi di libertà.
Nel package `car` troviamo il comando `outlierTest()` che consente di effettuare il test per individuare i valori outlier fornendo il Bonferroni p-value per il valore outlier più estremo.

Gli elementi h_{ii} sulla diagonale principale della matrice H (hat) si chiamano leverages (punti di leva).

Poiché $var(pred(y_i)) = h_i\sigma^2$ h_i è la precisione con cui il valore è stimato relativamente a σ^2. Quindi, valori piccoli di h_i indicano che lo stimatore di y_i è basato sul contributo di molte osservazioni. Invece, valori grandi di h_i ossia molto vicini a 1, implicano che $var(y_i - pred(y_i)) = (1-h_i)\sigma^2 \approx 0$ e che $pred(y_i)$ tende a essere vicino a y_i e che $pred(y_i)$ è determinato in modo predominante dalla singola osservazione y_i che quindi ha un effetto di leva importante.

Outlier, leverage, influence

Un punto con “ alto” leverage ha un residuo con varianza “piccola” (cioè la retta deve passare “ vicino” a questo punto). Un punto con “ alto” leverage è un punto “ distante”. Hoaglin e Wesh suggeriscono di segnalare come punti con un elevato effetto di leva quei punti per cui $h_i > 2p/n$. Per calcolare i punti di leverage in R possiamo usare i comandi `hat()` e `hatvalues()`.

Un punto influente (influence) è un punto che, se rimosso, produce un notevole cambiamento nella stima del modello. Un punto influente non necessariamente è un outlier e può o non può avere un leverage elevato, ma, in generale ha almeno una di queste due caratteristiche. Misure di influence sono date dai residui jacknife, dai cambiamenti nelle stime dei coefficienti di regressione e della varianza residua che si ottengono escludendo un punto dal stima e dalla distanza di Cook.

Per la distanza di Cook possiamo usare i comandi `cooks.distance()` ed `influence.measures()` nel package `stats`.

Regressione robusta

Quando nella regressione gli errori non sono distribuiti normalmente oppure si hanno molti valori outliers le stime OLS non sono buone e si deve ricorrere alla regressione robusta36. Il metodo più usato per questo tipo di regressione è quello della M-estimation introdotto da Huber. Tali stimatori possono essere considerati una generalizzazione delle stime di massima verosimiglianza. Abbiamo il modello lineare generale (con k regressori) riferito all'i.esima di n osservazioni ed espresso in notazione vettoriale:

$$y_i = X_i^T \beta + \varepsilon_i \qquad \text{e stimato da:} \qquad y_i = X_i^T \hat{\beta} + e_i$$

Il generico *M-estimator* è ottenuto minimizzando la funzione obiettivo:

$$\rho(e_i) = \sum_{i=1}^{n} \rho\left(y_i - X_i^T \hat{\beta}\right)$$

dove la funzione ρ fornisce il contributo di ogni singolo residuo alla funzione obiettivo. La funzione ρ deve godere delle seguenti proprietà:

Regressione robusta

- 1) $\rho(e) \geq 0$
  2) $\rho(0) = 0$
  3) $\rho(e) = \rho(-e)$
  4) $\rho(e_i) \geq \rho(e_j)$ se $|e_i| > |e_j|$

Ci sono diverse funzioni obiettivo che possono essere utilizzate e che soddisfano questi requisiti: se $\rho(e) = e^2$ abbiamo le stime dei minimi quadrati ordinari (OLS), se $\rho(e) = |e|$ abbiamo le stime LAD (Least absolute deviation regression), altre funzioni sono quella di Huber e la biquadratica:

Method	*Objective Function*	*Weight Function*
Least-Squares	$\rho_{\mathrm{LS}}(e) = e^2$	$w_{\mathrm{LS}}(e) = 1$
Huber	$\rho_{\mathrm{H}}(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \leq k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{H}}(e) = \begin{cases} 1 & \text{for } \lvert e\rvert \leq k \\ k/\lvert e\rvert & \text{for } \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_{\mathrm{B}}(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \leq k \\ k^2/6 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{B}}(e) = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \leq k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

Regressione robusta

Se poniamo ψ = ρ' (la derivata prima) e differenziamo la funzione obiettivo rispetto ai coefficienti $\hat{\beta}$, poniamo uguali a zero le derivate parziali, otteniamo un sistema di k+1 equazioni di stima per i coefficienti:

$$\sum_{i=1}^{n} \psi\left(y_i - X_i^T \hat{\beta}\right) X_i^T = 0$$

Se definiamo la funzione peso w(e) = ψ(e)/e e poniamo $w_i = w(e_i)$ le equazioni possono essere riscritte in questi termini:

$$\sum_{i=1}^{n} w_i \left(y_i - X_i^T \hat{\beta}\right) X_i^T = 0$$

La soluzione di tale sistema è un problema di minimi quadrati ponderati e si ottiene attraverso una procedura iterativa (IRLS=iteratively reweigheted least squares).
Un'altra tecnica di regressione robusta è quella della regressione LTS (least trimmed squares).

Regressione robusta

Un'altra tecnica di regressione robusta è quella della regressione LTS (least trimmed squares) .

In questo caso i quadrati dei residui vengono ordinati in ordine crescente:

$(e^2)_{(1)}$,$(e^2)_{(2)}$,...$(e^2)_{(n)}$

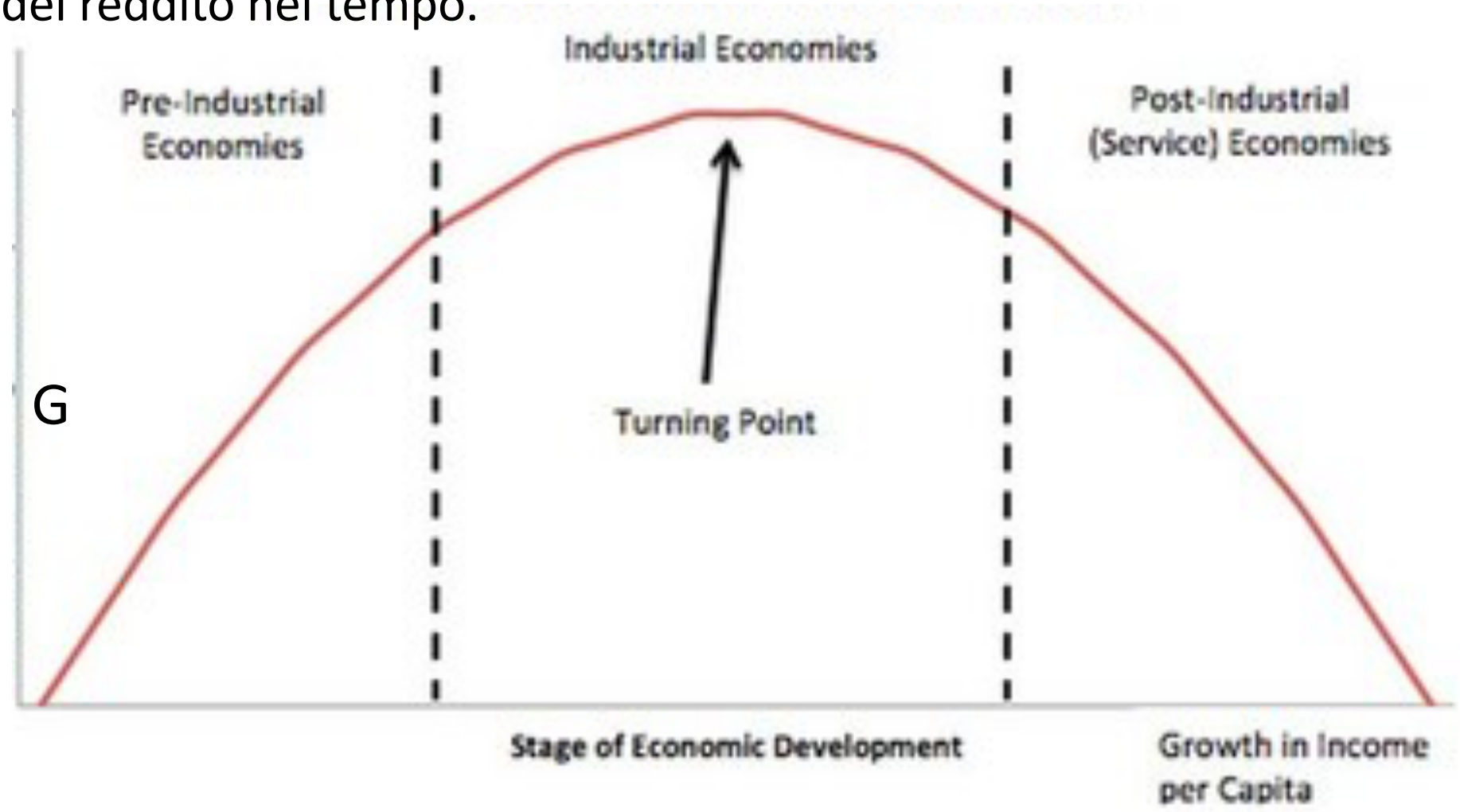
le stime LTS dei coefficienti di regressione $\hat{\beta}$ sono ottenute minimizzando la somma dei piccoli m valori dei quadrati dei residui:

$$LTS(\beta) = \min \left\{ \sum_{i=1}^{m} \left(e^2\right)_{(i)} \right\}$$

dove m = ⌊n / 2⌋ + ⌊(k + 2) / 2⌋ e ⌊ ⌋ indica l'approssimazione all'intero più piccolo.

Esempio: curva di Kuznets

La curva di Kuznets descrive l'andamento della diseguaglianza in rapporto al tasso di sviluppo, mostrando l'evoluzione della distribuzione del reddito nel tempo.



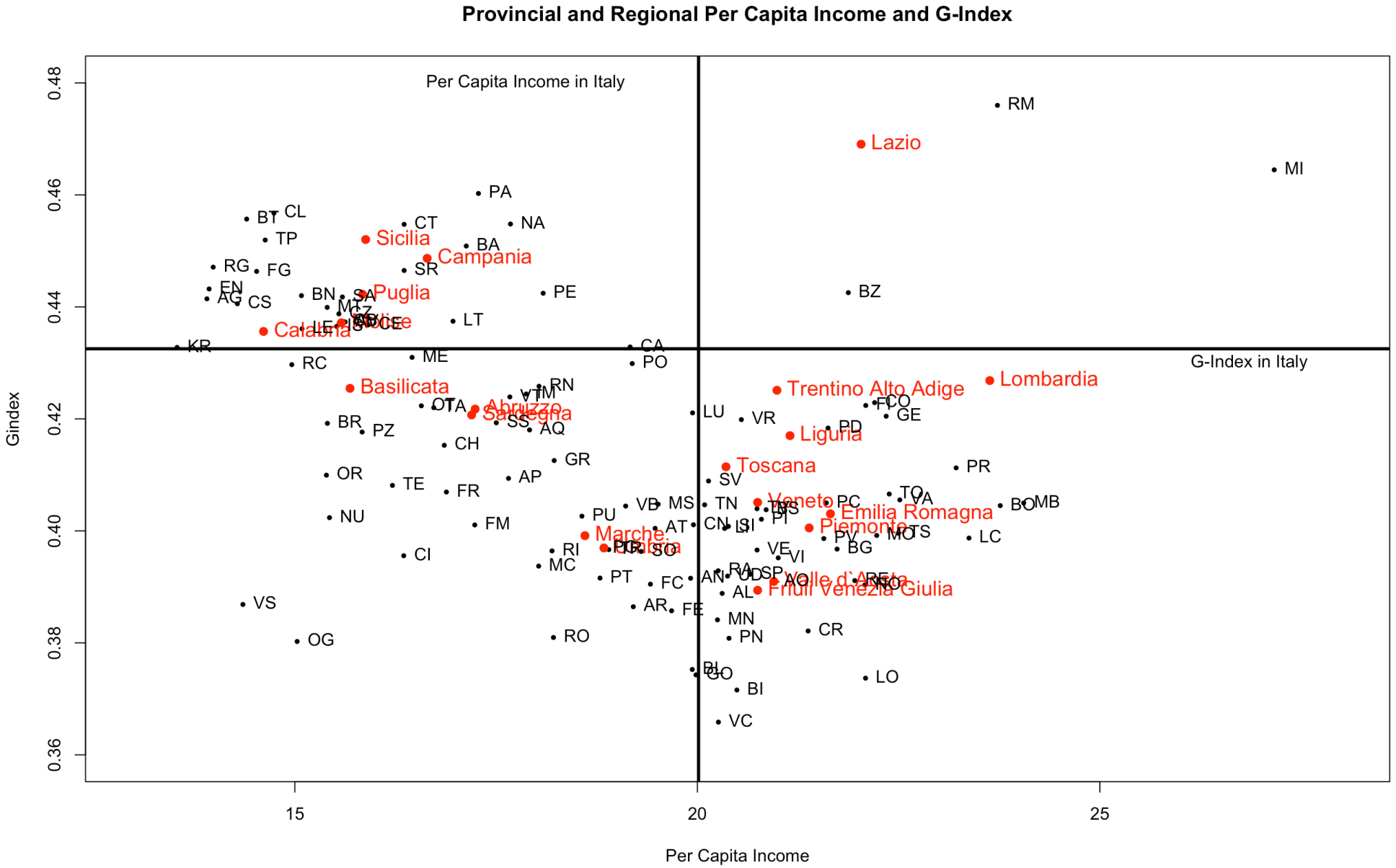
Esempio: curva di Kuznets

Sull'asse delle ascisse troviamo il Prodotto nazionale lordo pro-capite, mentre su quello delle ordinate il coefficiente di Gini. La forma della curva assomiglia ad una U rovesciata e sta appunto ad indicare che la distribuzione del reddito tende a peggiorare nella prima fase dello sviluppo (massimo incurvamento), migliorando invece in maniera costante con la transizione a un'economia di tipo industriale. Questo avviene in quanto in una prima fase la fascia di popolazione più ricca investe il proprio capitale, incrementando ulteriormente la propria ricchezza; in un secondo momento però viene colpita maggiormente dalla tassazione, con conseguente effetto redistributivo.

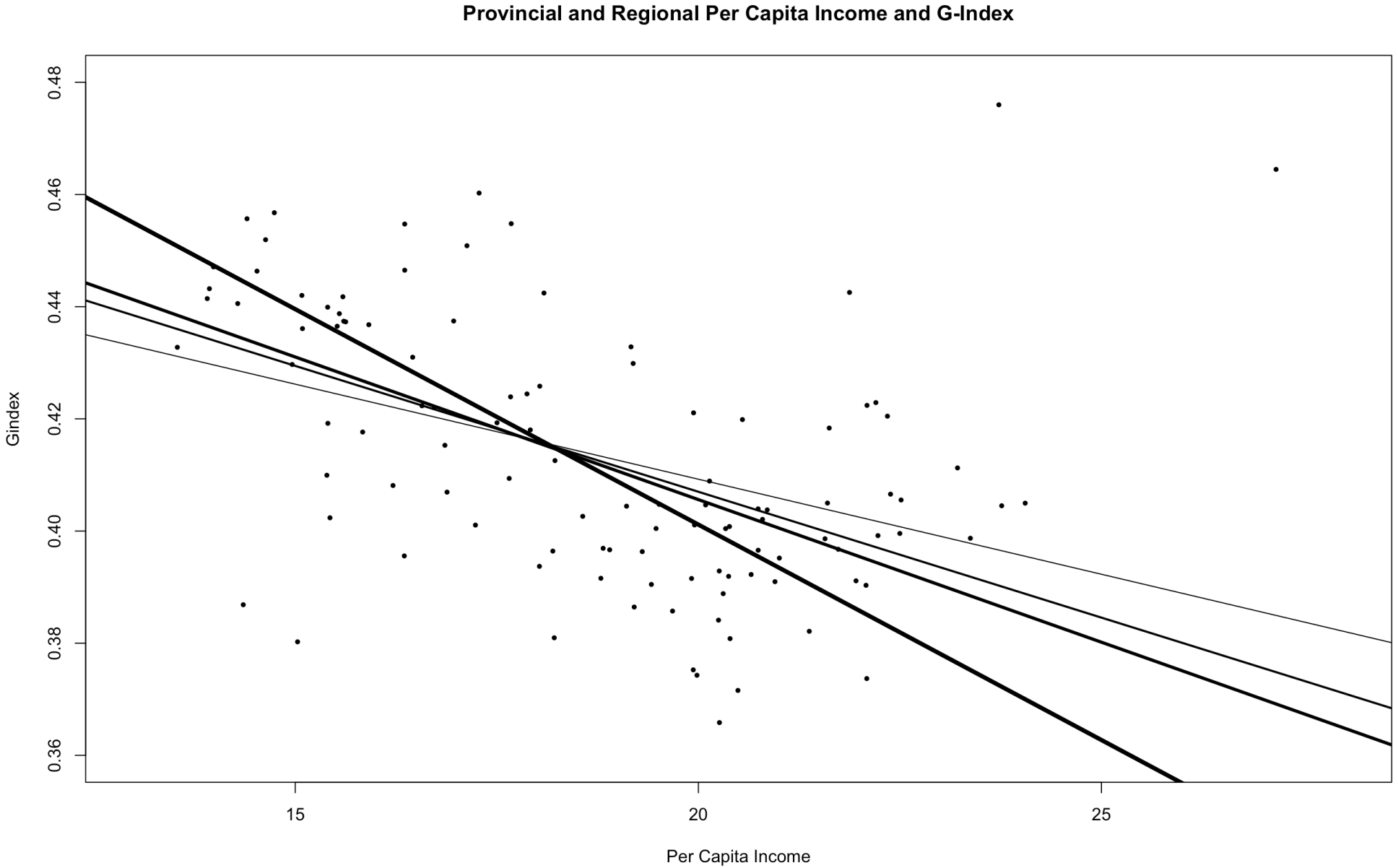
Ne esistono anche varianti ambientali.

Proviamo a vedere la cosa nello spazio (cross section) invece che nel tempo (time series), utilizzando i dati resi noti dall'agenzia delle entrate per tutti gli e per tutti i comuni italiani.

Esempio: curva di Kuznets



Esempio: curva di Kuznets

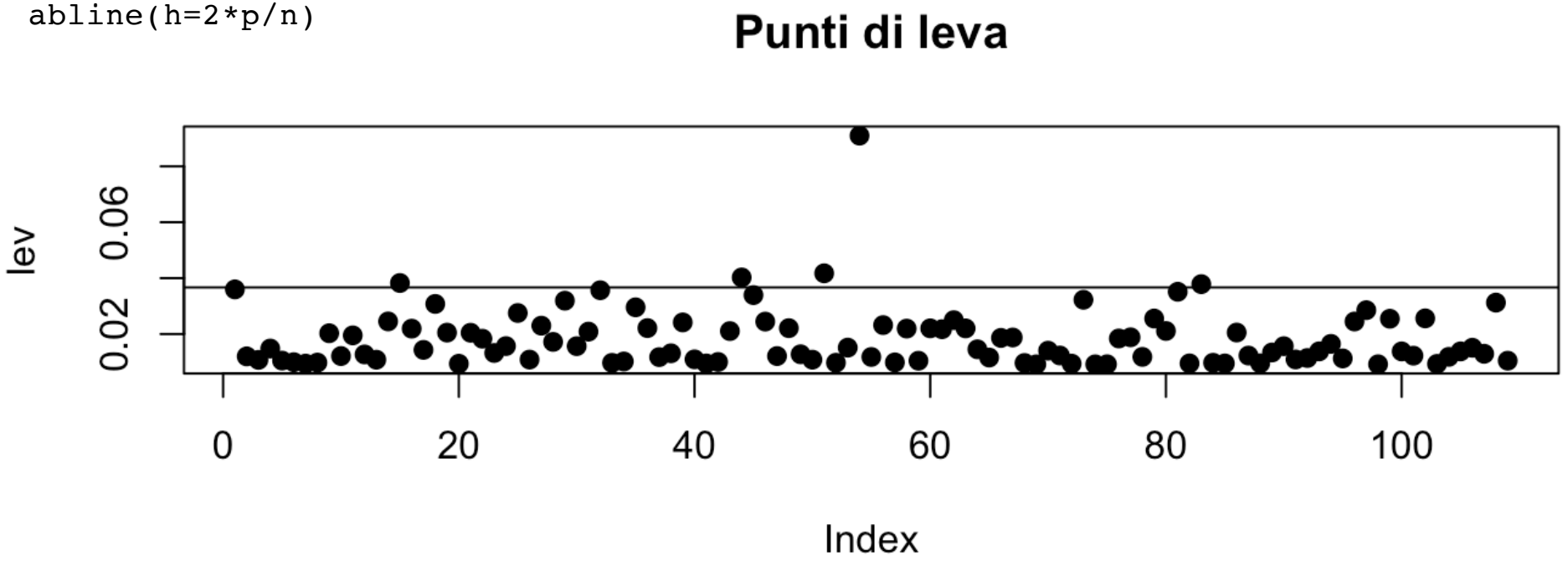


Esempio: curva di Kuznets

```
load("/Users/utente/Documents/prova r pkg/lucidi del corso/Kuznets.RData")
plot(provku$rmed,provku$gindex,cex=0.5,pch=19,xlim=c(13,28),ylim=c(0.36,0.48),xlab=
"Per Capita Income",ylab="Gindex",main="Provincial and Regional Per Capita Income
and G-Index") points(regku$rmed,regku$gindex,cex=1,pch=19,col=2)
text(regku$rmed,regku$gindex,regku$cod,pos = 4,cex=1.2,col=2)
text(provku$rmed,provku$gindex,provku$cod,pos = 4)
abline(v=itku$rmed,lw=3);abline(h=itku$gindex,lw=3)
text(16.5,0.48,"Per Capita Income in Italy",pos=4)
text(26,0.43,"G-Index in Italy",pos=4)
library(MASS);plot(provku$rmed,provku$gindex,cex=0.5,pch=19,xlim=c(13,28),ylim=c(0.
36,0.48),xlab="Per Capita Income",ylab="Gindex",main="Provincial and Regional Per
Capita Income and G-Index")
reglm   <-lm(gindex~rmed, data=provku)
regrml1 <-rlm(gindex~rmed, data=provku, psi=psi.huber)
regrml2 <-rlm(gindex~rmed, data=provku, psi=psi.bisquare)
reglts  <-lqs(gindex~rmed, data=provku, method="lts")
abline(reglm);abline(regrml1,lwd=2);abline(regrml2,lwd=3);abline(reglts,lwd=4)
confronto<-data.frame(coef(reglm),coef(regrml1),coef(regrml2),coef(reglts))
options(digits=3); Confronto
            coef.reglm. coef.regrml1. coef.regrml2. coef.reglts.
(Intercept)     0.47703       0.49677       0.50731      0.55481
rmed           -0.00339      -0.00449      -0.00509     -0.00768
library(car)
outlierTest(reglm)
                  rstudent        p-value Bonferonni p
Milano 4.07           8.93e-05      0.00973
Roma   3.94           1.49e-04      0.01622
```

Esempio: curva di Kuznets

```
library(car)
outlierTest(reglm)
 rstudent         p-value Bonferonni p
Milano 4.07            8.93e-05      0.00973
Roma   3.94            1.49e-04      0.01622

lev<-hat(model.matrix(reglm))
lev<-hatvalues(reglm)
n<-length(lev)
p<-sum(lev)
plot(lev, main="Punti di leva",cex=1,pch=19)
abline(h=2*p/n)
```



Regressione Quantilica

Motivazione

Quello che fa la curva di regressione è un grande riassunto per le medie delle distribuzioni corrispondenti all'insieme delle x. Potremmo andare oltre e calcolare diverse curve di regressione corrispondenti ai vari punti percentuali delle distribuzioni e ottenere così un'immagine più completa dell'insieme dei dati. Di solito questo non viene fatto, e così la regressione dà spesso un'immagine piuttosto incompleta. Proprio come la media fornisce un'immagine incompleta di una singola distribuzione, così la curva di regressione fornisce un'immagine corrispondente incompleta per un insieme di distribuzioni.

Mosteller and Tukey (1977)

Regressione Quantilica

Data una variabile aleatoria X, con funzione di distribuzione F, definiamo il τ-esimo quantile di X come:

$$Q_{X(\tau)} = F^{-1}{}_X(\tau) = \inf\{x \mid F(x) \geq \tau\}.$$

dal punto di vista delle densità, il τ-esimo quantile divide l'area sotto la densità in due parti: una con area τ al di sotto del τ-esimo quantile e l'altra con area 1 - τ sopra di esso.

I quantili risolvono un semplice problema di ottimizzazione:

$$\alpha^*(\tau) = \text{argmin}_\alpha \, E \, \rho_\tau(Y - \alpha)$$

Così come la media risolve:

$$\mu = \text{argmin}_m \, E(Y - m)^2$$

Regressione Quantilica

Se ε_i è l'errore di previsione del modello, OLS minimizza la somma dei quadrati degli errori. La regressione mediana, minimizza la somma dei loro valori assoluti. La regressione quantilica minimizza una somma che dà penalità asimmetriche $(1 – q)|e_i|$ per errori positivi e $q|e_i|$ per errori negativi. Sebbene il suo calcolo richieda metodi di programmazione lineare, lo stimatore di regressione quantilica è asintoticamente distribuito normalmente.

La regressione mediana è più robusta rispetto alla regressione rispetto ai minimi quadrati ed è semiparametrica in quanto evita ipotesi sulla distribuzione Gaussiana dell'errore.

Regressione Quantilica

La regressione quantilica può essere in alcune circostanze una valida alternativa alla regressione ordinaria quando non sono verificati tutti i requisiti di base per applicare gli OLS o la ML, in particolare quando si hanno degli outlier (la regressione quantilica è una tecnica robusta) e le stime dei coefficienti di regressione risultano più efficienti quando gli errori non sono distribuiti normalmente.

Definiamo come $f(y_i|x_i)$ e $F(y_i|x_i)$, rispettivamente, la funzione di densità di probabilità (pdf) e la funzione di distribuzione cumulativa (cdf) della variabile dipendente y_i condizionata alle esplicative (ausiliarie, covariate) osservate x_i. Supponiamo che per ogni unità *i* e per ogni $\tau \in (0, 1)$, esista un vettore K-dimensionale $\beta(\tau)$ tale che:

$$Q_{\mathbf{y}_i}\left(\tau \middle| \mathbf{x}_i\right) = \mathbf{x}_i^T \beta\left(\tau\right)$$

Regressione Quantilica

dove $Q_{\mathbf{y}_i}\left(\tau|\mathbf{x}_i\right)$ ≡ inf{ y_i: F(y_i|$\mathbf{x}_i$) ≥ τ } è il τ-esimo quantile condizionato di y_i.

Una stima $\hat{\beta}_{QR}\left(\tau\right)$ del vettore β seguendo l'approccio suggerito da Koenker e Bassett, si ottiene minimizzando con metodi di programmazione lineare (cioè l'algoritmo del simplesso), la funzione obiettivo:

$$L_n\left(\beta\left(\tau\right)\right) = n^{-1}\sum_{i=1}^{n} w_i\left[\left(\tau - \pi_{i\tau}\right)\left(y_i - \mathbf{x}_i^T\beta\left(\tau\right)\right)\right]$$

con $\pi_{i\tau} = I\left(y_i \le \mathbf{x}_i^T\beta\left(\tau\right)\right)$, dove $I\left(\cdot\right)$ è la funzione di indicatore e wi è il peso dell'unità i. Questo peso può essere utilizzato per tenere conto del disegno di campionamento dell'indagine, anche se di solito è considerato costante, assumendo implicitamente che il disegno di campionamento non sia informativo per l'inferenza su β.

Regressione Quantilica

Il vettore di coefficienti $\hat{\beta}_{QR}(\tau)$ si ottiene quindi minimizzando una funzione:

$$L_n\left(\beta(\tau)\right) = \sum_{i:y_i \geq X_i^T\beta}^{n} q\left|y_i - \mathbf{x}_i^T\beta(\tau)\right| + \sum_{i:y_i < X_i^T\beta}^{n} (1-q)\left|y_i - \mathbf{x}_i^T\beta(\tau)\right|$$

Questa funzione non è differenziabile ed è minimizzata tramite il metodo del simplesso, che garantisce una soluzione in un numero finito di iterazioni. Sebbene sia dimostrato che lo stimatore è asintoticamente normale con matrice di var e cov nota analiticamente, la sua espressione è difficile da stimare. Quindi degli errori standard Bootstrap vengono spesso utilizzati al posto degli errori standard analitici.

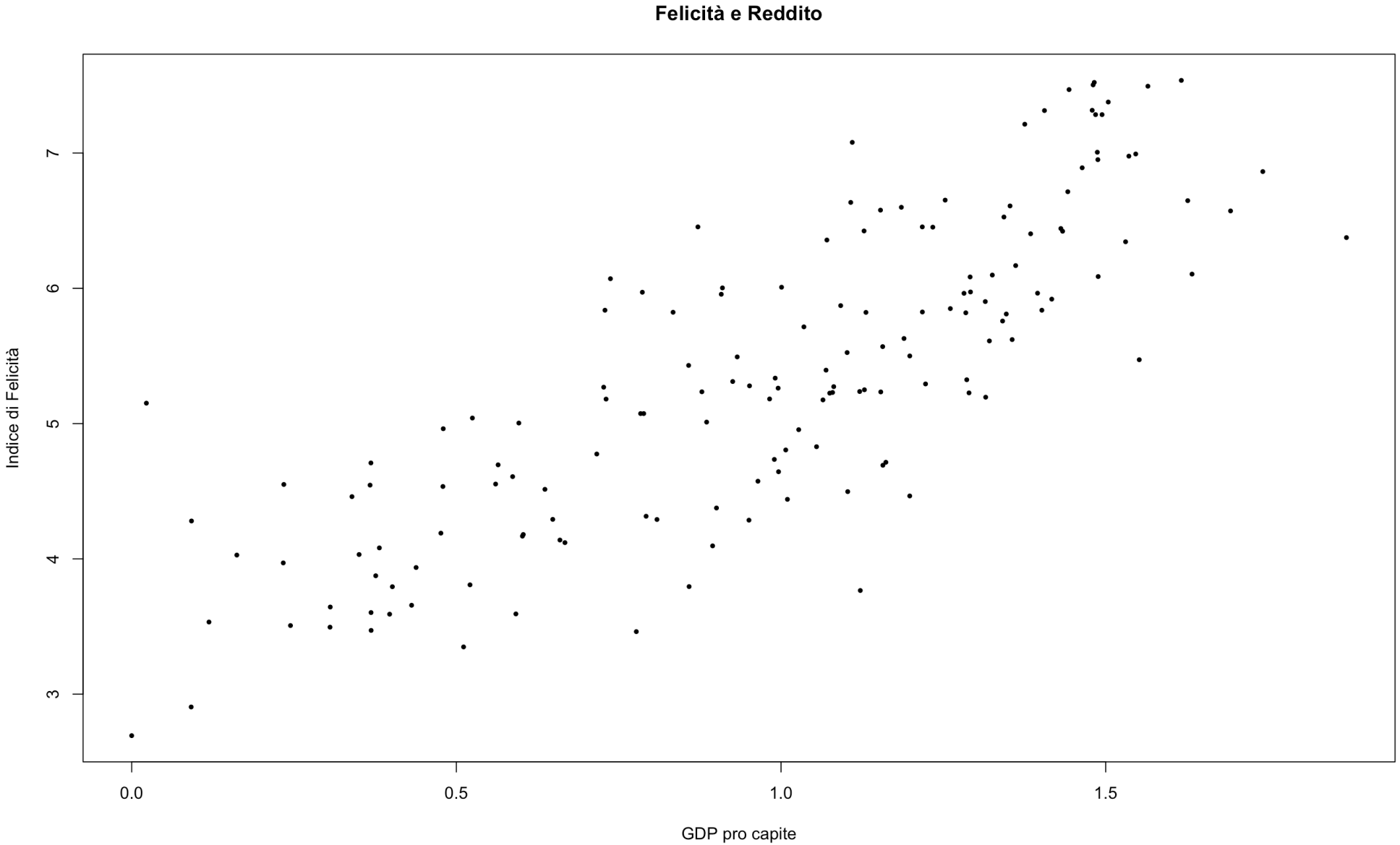
Regressione Quantilica

Si ottiene quindi un vettore di stime che, secondo i teoremi asintotici riportati da Koenker, hanno una distribuzione normale asintoticamente multivariata, con vettore medio e matrice di covarianza la cui stima presume che gli errori siano iid (Koenker e Bassett, 1978).

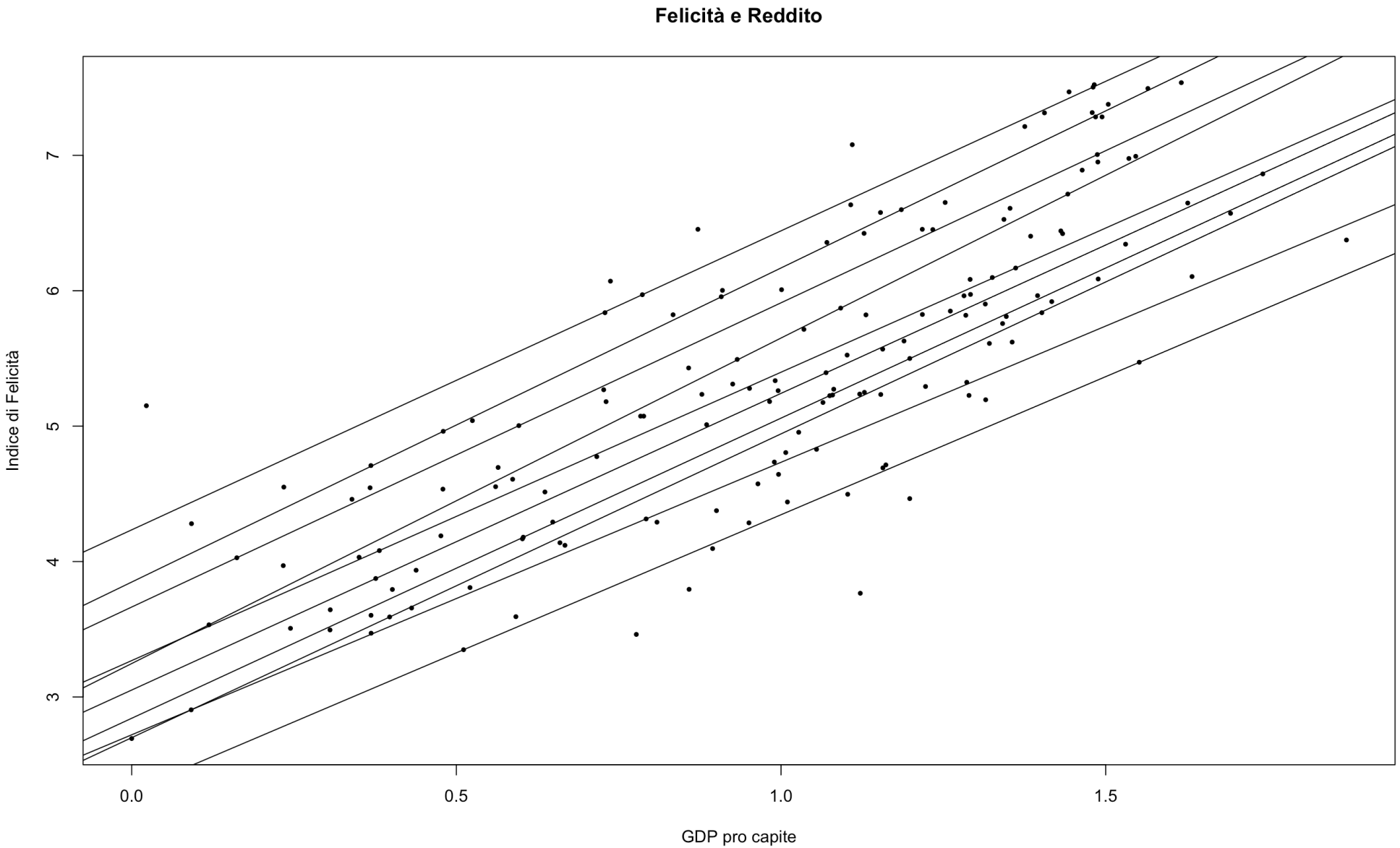
Alcuni vantaggi della regressione quantilica (QR): mentre l'OLS può essere inefficiente se gli errori sono lontani dall'ipotesi di gaussianità, QR è più robusta rispetto agli errori non normali e ai valori anomali. Il QR fornisce anche una descrizione più ricca di informazioni, permettendoci di considerare l'impatto di una covariata sull'intera distribuzione di y, non solo sulla sua media condizionata.

Inoltre, QR è invariante alle trasformazioni monotone, come log(·), quindi i quantili di h(y), una trasformata monotona di y, sono $h(Q_q(y))$, e la trasformazione inversa può essere usata per tradurre il risale a y.

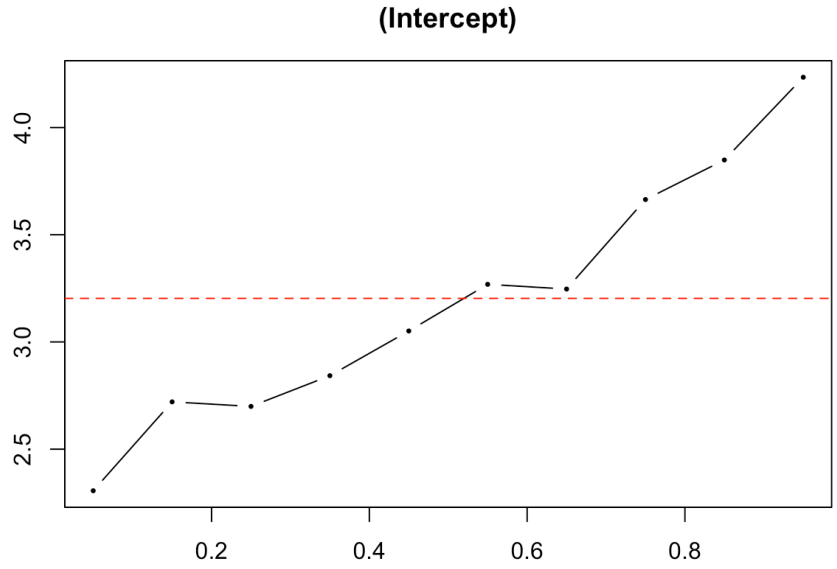
Esempio: “I soldi fanno la felicità?”

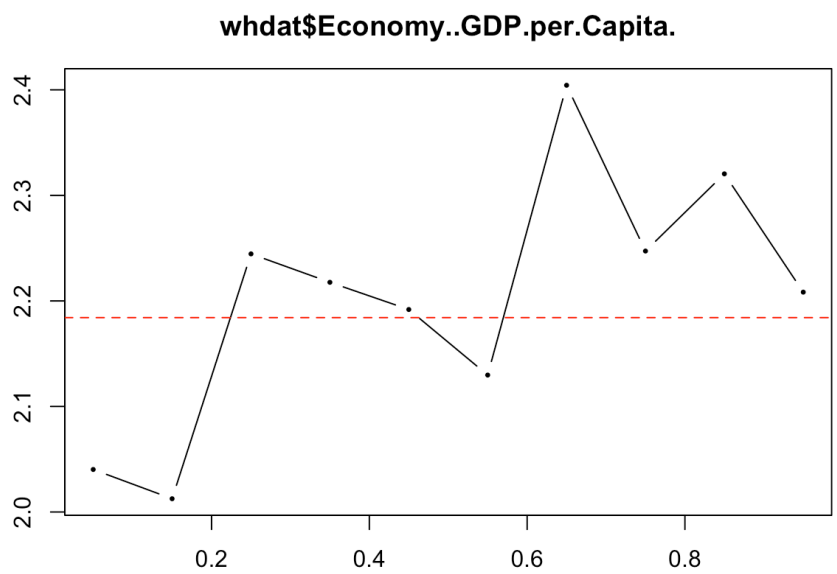


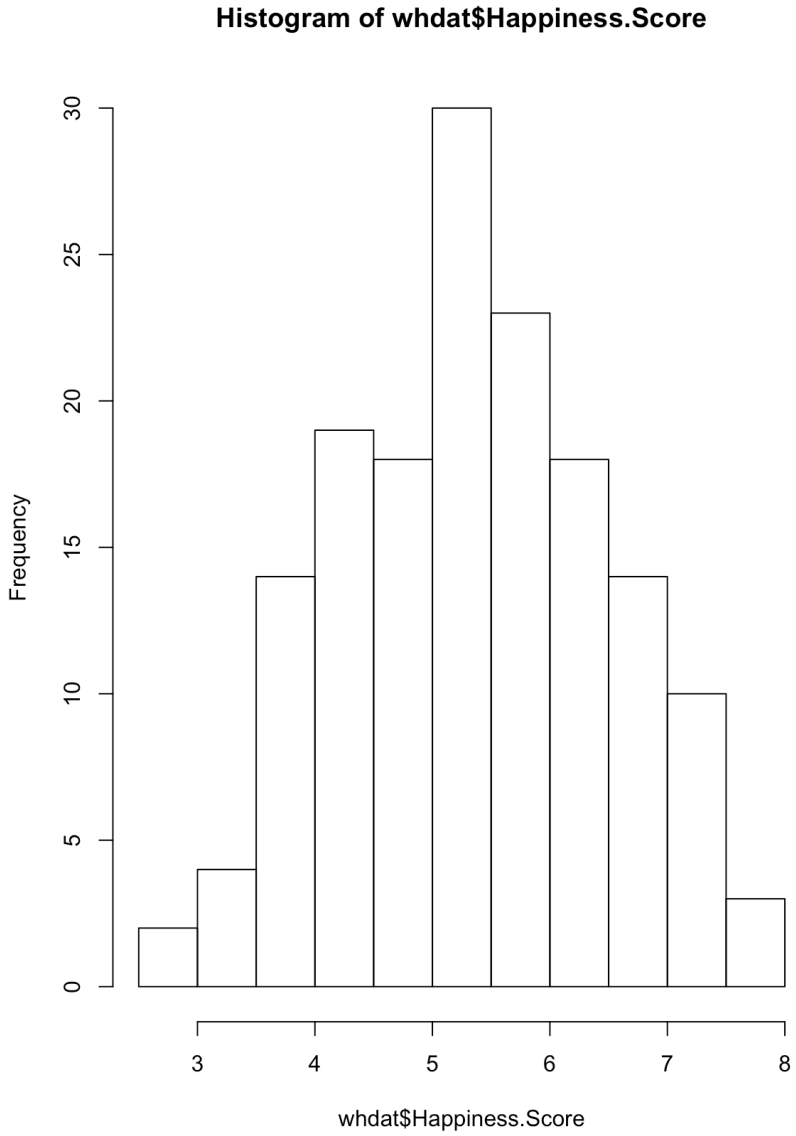
Esempio: “I soldi fanno la felicità?”



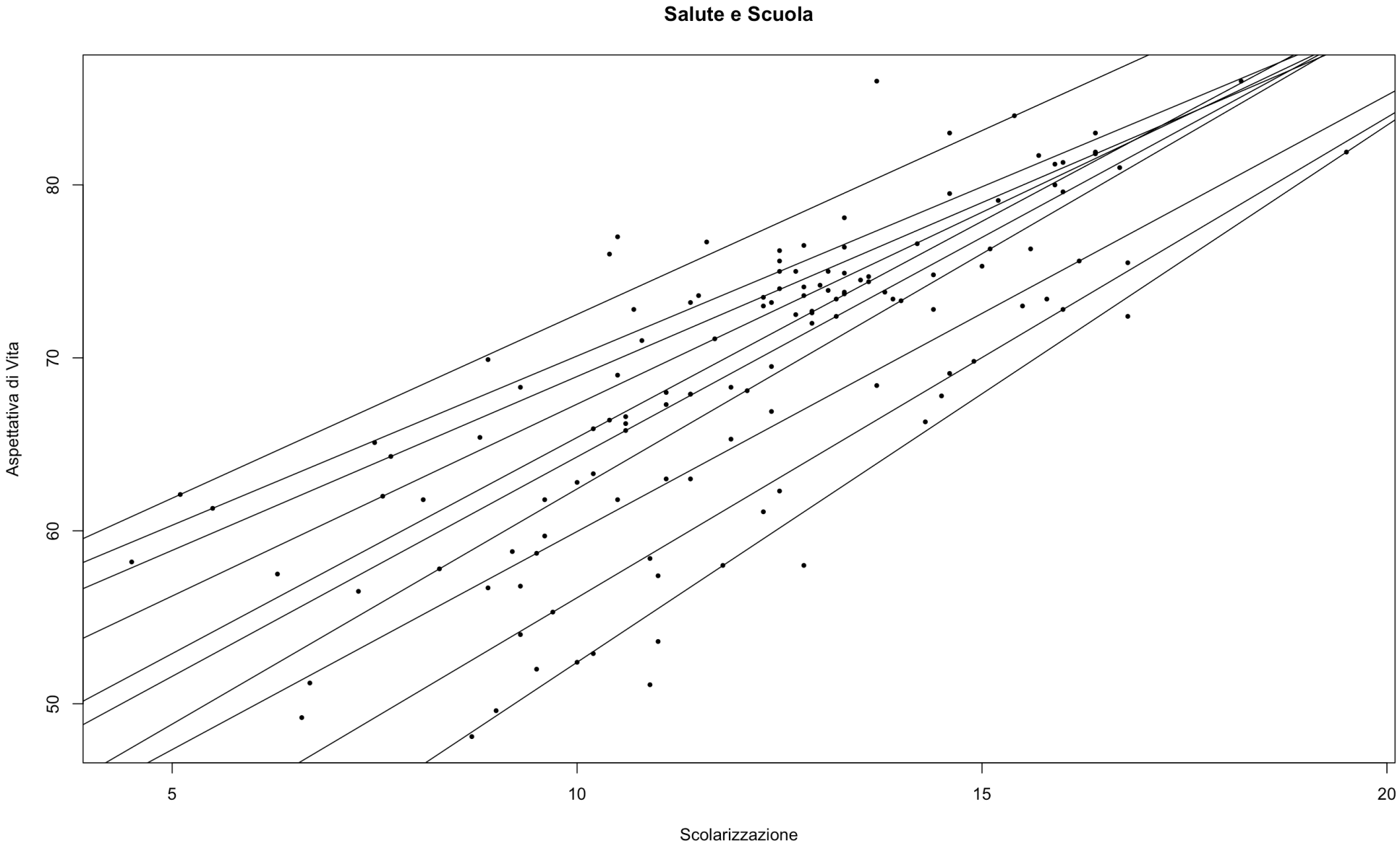
Esempio: “I soldi fanno la felicità?”







Esempio: eteroschedasticità



Esempio: eteroschedasticità