

I dati delle automobili

Data Mining

CLAMSES - University of Milano-Bicocca

Aldo Solari

Riferimenti bibliografici

- AS §2.1.1, §2.1.2

I dati delle automobili si riferiscono alle caratteristiche di 203 modelli di automobili importati negli USA nel 1985

I dati originali sono disponibili qui:

`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
autos`

Questi dati sono stati elaborati convertendo le unità di misura, eliminando alcune variabili originarie, correggendo alcuni nomi di marche, etc.

Obiettivo: prevedere il consumo di carburante (o, equivalentemente, la distanza percorsa per unità di carburante) in funzione di determinate caratteristiche di un'automobile

- brand : manufacturer (factor, 22 levels), casa produttrice (fattore, 22 livelli)
- fuel : type of engine fuel (factor, 2 levels: diesel, gasoline), tipo di alimentazione del motore (fattore, 2 livelli)
- aspiration : type of engine aspiration (factor, 2 levels: standard, turbo), tipo di aspirazione del motore (fattore, 2 livelli)
- bodystyle : type of body style (factor, 5 levels: hardtop, wagon, sedan, hatchback, convertible), tipo di carrozzeria (fattore, 5 livelli)
- drive.wheels : type of drive wheels (factor, 3 levels: 4wd, fwd, rwd), tipo di trazione (fattore, 3 livelli)
- engine.location : location of engine (factor, 2 levels: front, rear), posizione del motore (fattore, 2 livelli)
- wheel.base : distance between axes (cm), distanza tra gli assi (cm)
- length : length (cm), lunghezza (cm)
- width : width (cm), larghezza (cm)
- height : height (cm), altezza (cm)
- curb.weight : weight (kg), peso (kg)
- engine size : engine size (l), cilindrata (l)
- compression.ratio : compression ratio, rapporto di compressione
- HP : horsepower, cavalli motore
- peak.rot : number of peak revolutions per minute, numero di giri massimi del motore al minuto
- **city.distance** : city distance covered (km/l), percorrenza urbana (km/l)
- highway.distance : highway distance (km/l), percorrenza extra urbana (km/l)
- n.cylinders : number of cylinders, numero di cilindri

I dati delle automobili si riferiscono alle caratteristiche di 203 modelli di automobili importati negli USA nel 1985

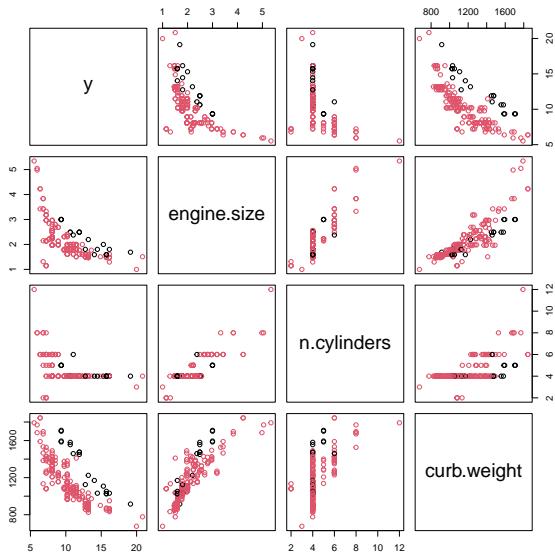
I dati originali sono disponibili qui:

`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
autos`

Questi dati sono stati elaborati convertendo le unità di misura, eliminando alcune variabili originarie, correggendo alcuni nomi di marche, etc.

Obiettivo: prevedere il consumo di carburante (o, equivalentemente, la distanza percorsa per unità di carburante) in funzione di determinate caratteristiche di un'automobile

Dataset ridotto



Istogramma e stima della densità

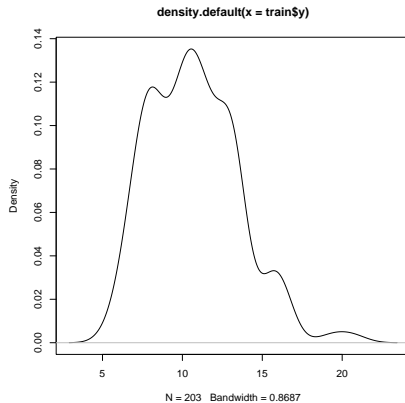
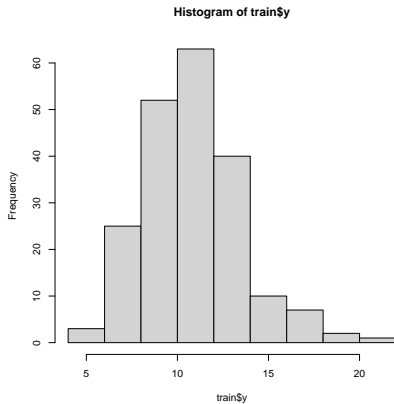
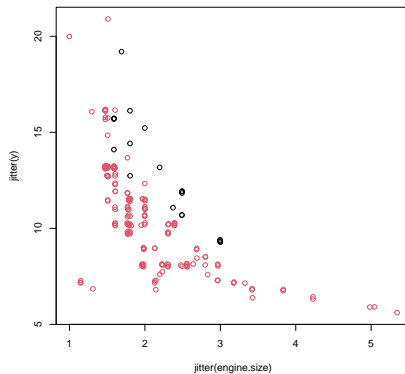
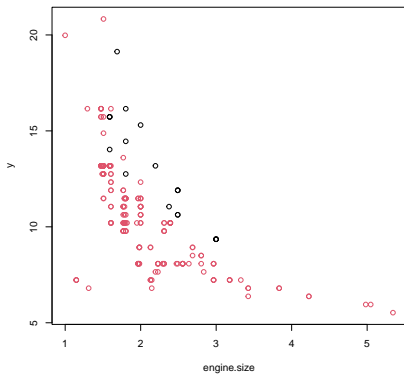
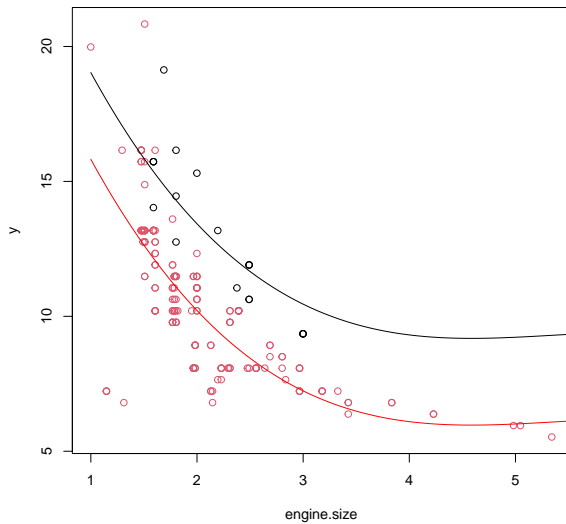


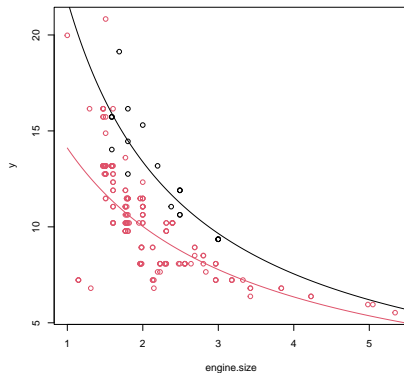
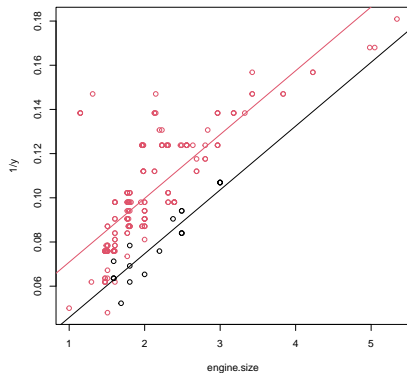
Diagramma di dispersione (jittered) $y \sim \text{engine.size}$



Modello 1

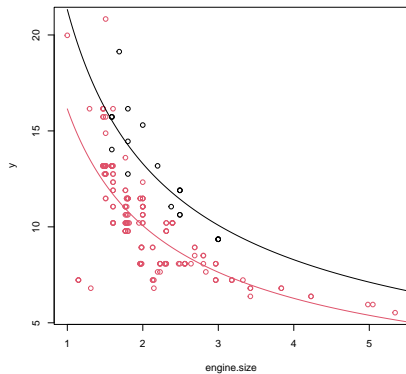
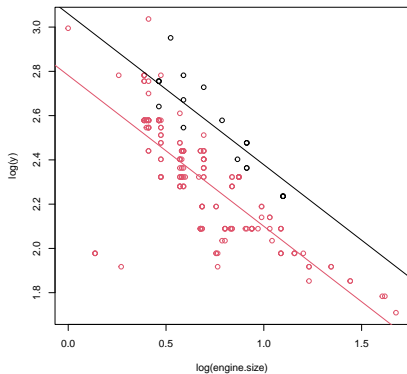


Modello 2



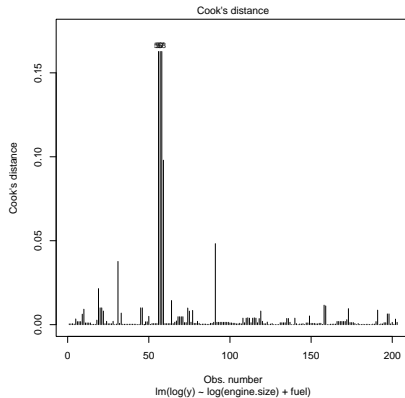
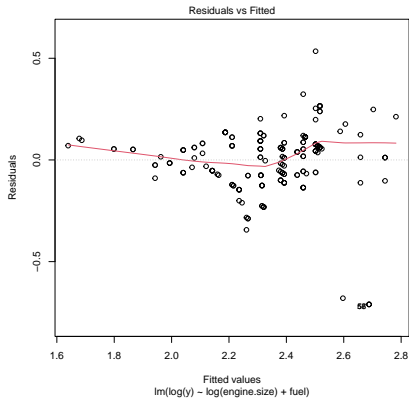
$$1/y = \beta_1 + \beta_2 \text{engine.size} + \beta_3 I\{\text{fuel} = \text{gas}\} + \varepsilon$$

Modello 3



$$\log(y) = \beta_1 + \beta_2 \log(\text{engine.size}) + \beta_3 I\{\text{fuel} = \text{gas}\} + \varepsilon$$

Bontà di adattamento



```
train[which(cooks.distance(fit3) > .07), ]
```

	y	engine.size	n.cylinders	curb.weight	fuel
31	20.832	1.5076	4	777.0	gas
154	11.479	1.5076	4	1410.7	gas

```
table(train$n.cylinders)
```

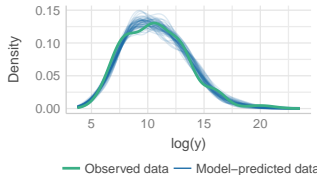
2	3	4	5	6	8	12
4	1	157	11	24	5	1

Modello 3 (aggiornato)

$$\begin{aligned}\log(y) = & \beta_1 + \beta_2 \log(\text{engine.size}) + \beta_3 I\{\text{fuel} = \text{gas}\} + \\ & + \beta_4 \log(\text{curb.weight}) + \beta_5 I\{\text{n.cylinders} = 2\} + \varepsilon\end{aligned}$$

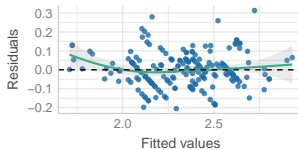
Posterior Predictive Check

Model-predicted lines should resemble observed data li



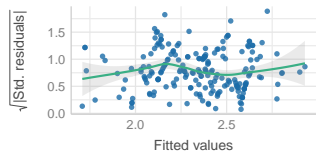
Linearity

Reference line should be flat and horizontal



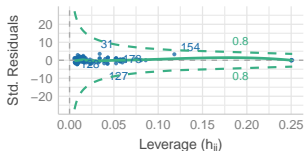
Homogeneity of Variance

Reference line should be flat and horizontal



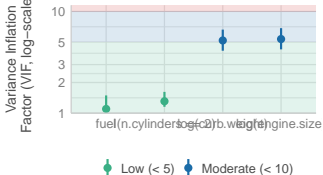
Influential Observations

Points should be inside the contour lines



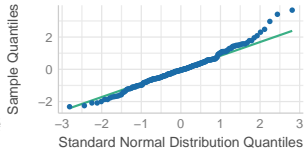
Collinearity

High collinearity (VIF) may inflate parameter uncertainty



Normality of Residuals

Dots should fall along the line



```
fit1 = update(fit1, . ~ . + log(curb.weight)
+ I(n.cylinders==2), train)
mean(resid(fit1)^2)
```

```
[1] 1.07706
```

```
fit2 = update(fit2, . ~ . + log(curb.weight)
+ I(n.cylinders==2), train)
mean((train$y - 1/fitted(fit2))^2)
```

```
[1] 1.143411
```

```
fit3 = update(fit3, . ~ . + log(curb.weight)
+ I(n.cylinders==2), train)
mean((train$y - exp(fitted(fit3)))^2)
```

```
[1] 1.016788
```