UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

6. Regressione semiparametrica: esempi

Francesco Pauli

DEAMS

Università di Trieste

A.A. 2018/2019

Indice

Previsione del danno

Per 1340 sinistri automobilistici con danni a persone si sono rilevati il costo del sinistro e una serie di caratteristiche relative all'assicurato o al sinistro stesso.

L'obiettivo è prevedere l'ammontare del danno sulla base delle caratteristiche, allo scopo di determinare la riserva sinistri.

(Quando accade un sinistro, l'assicuratore sa che dovrà pagare un risarcimento ma non sa ancora quanto. La determinazione dell'ammontare da liquidare può richiedere un tempo non trascurabile, specie quando vi sono danni alle persone. È allora rilevante prevedere l'ammontare, a sinistro avvenuto ma non ancora liquidato, per accantonare una cifra congrua.)

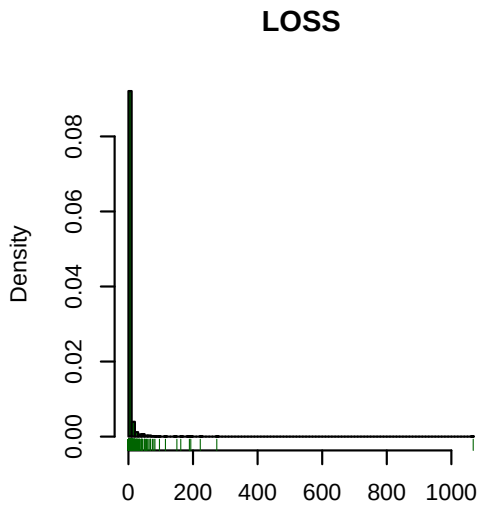
Automobile Bodily Injury Claims (autoBI)

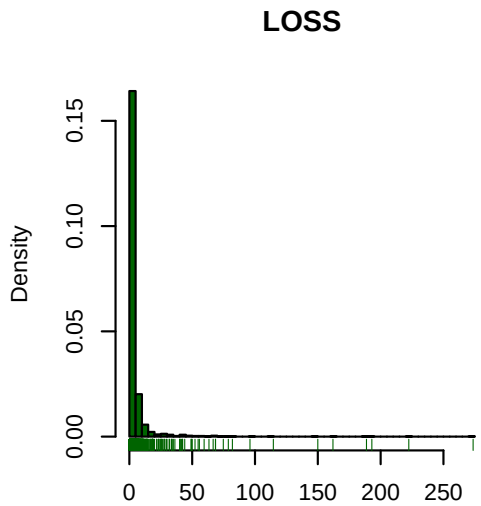
Fonte: Frees (2009).

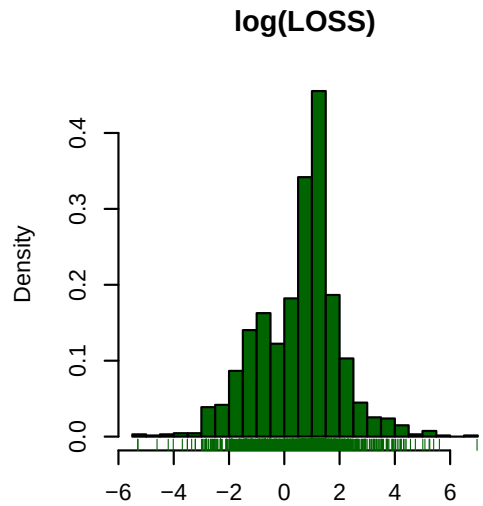
- `ATTORNEY` : se il danneggiato è assistito da un avvocato (yes/no)
- `CLMSEX` : maschio/femmina
- `MARITAL` : sposato (M)/libero (S)/vedovo (W)/divorziato (D)
- `CLMINSUR` : se il guidatore del veicolo danneggiato è assicurato (yes/no)
- `SEATBELT` : se il danneggiato indossava la cintura di sicurezza/seggiolino per bambini (yes/no)
- `CLMAGE` : età del danneggiato
- `AGECLASS` : età del danneggiato in classi: (-18] / (18,26] / (26,36] / (36,47] / (47+)
- `LOSS` : danno subito ($\times 1000$)

Distribuzione del danno (LOSS)

È di interesse costruire un modello per prevedere l'entità del danno sulla base di un campione di sinistri passati.



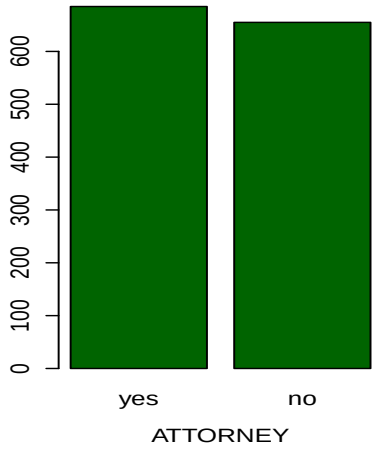


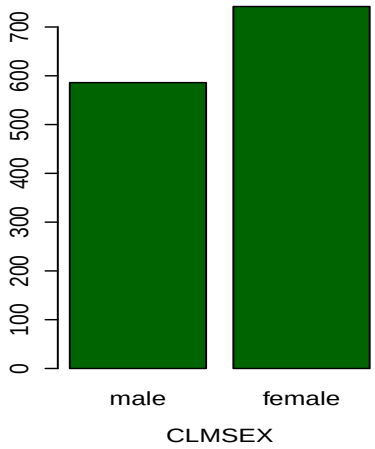


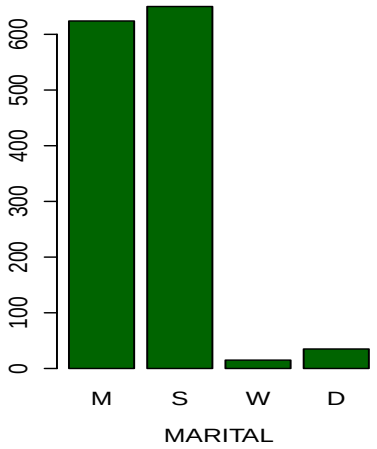
La distribuzione del danno è fortemente asimmetrica, la trasformazione logaritmica appare, almeno marginalmente, normalizzante.

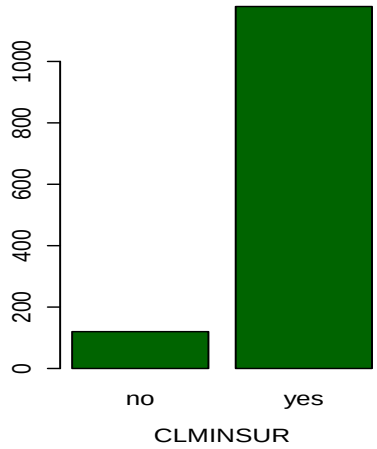
Informazioni sui sinistri

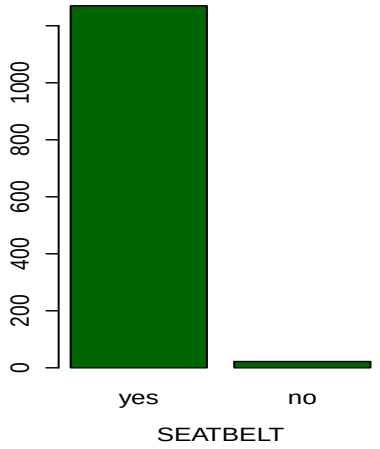
È disponibile una serie di ulteriori informazioni sui sinistri.

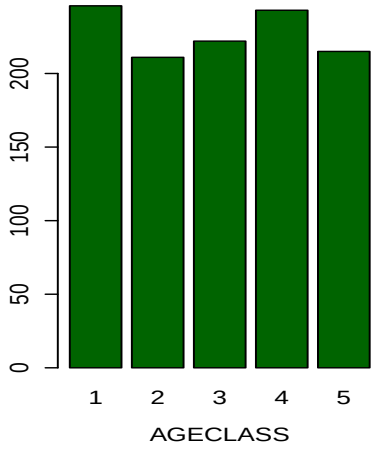


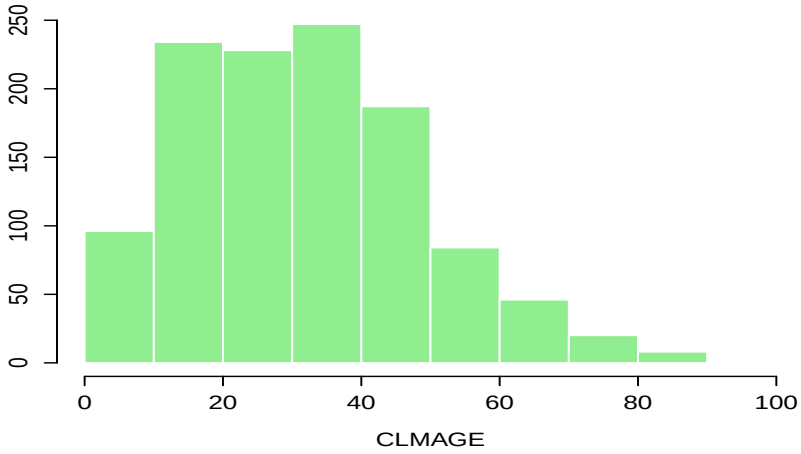






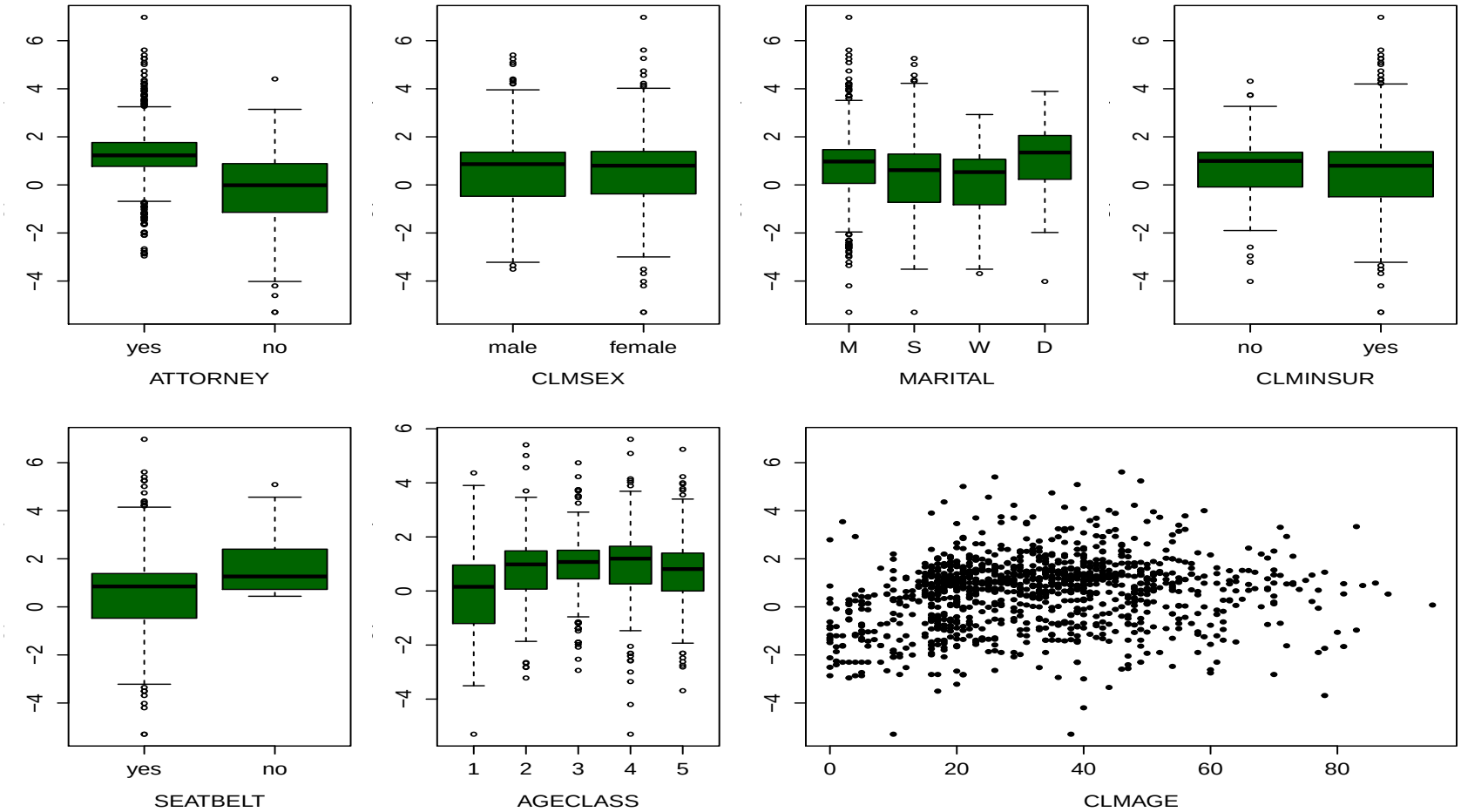






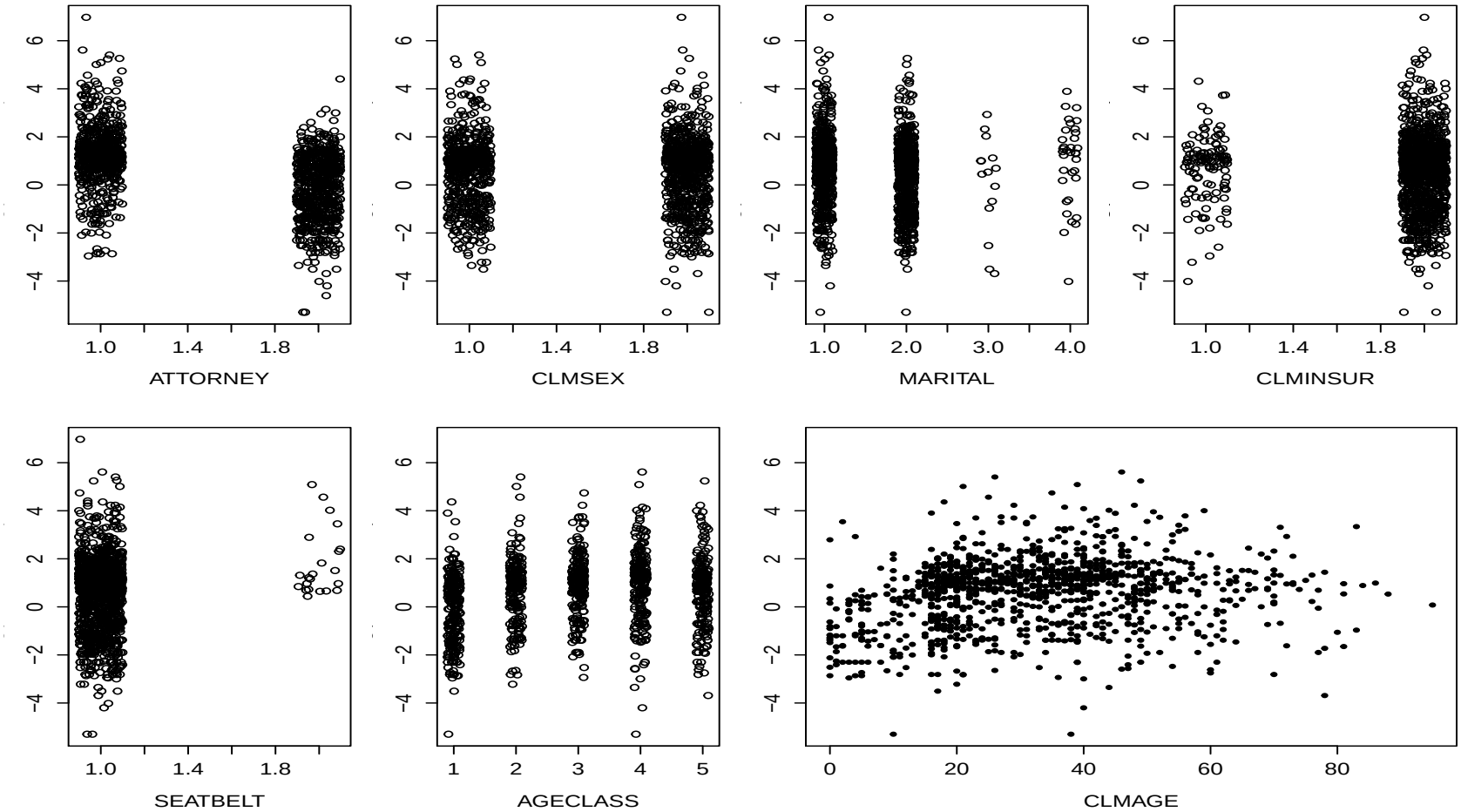
Log(LOSS) e variabili esplicative

In base all'analisi esplorativa, sussiste un legame tra il (log)danno e alcune delle esplicative.



Log(LOSS) e variabili esplicative

In base all'analisi esplorativa, sussiste un legame tra il (log)danno e alcune delle esplicative.



Dettaglio su CLMAGE

la variabile CLMAGE – di sotto indicata con z – può essere introdotta in vari modi ragionevoli nel modello

1. come un termine lineare,

$$y_i = \beta_0 + \beta z_i + \text{other covariates} + \varepsilon_i$$

2. come un termine quadratico,

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \text{other covariates} + \varepsilon_i$$

3. in versione discretizzata (opzione spesso scelta per la tariffazione)

Si ottengono risultati diversi (in termini ad esempio di previsioni) e possiamo ragionevolmente provare un numero limitato di alternative.

Dettaglio su CLMAGE

la variabile CLMAGE – di sotto indicata con z – può essere introdotta in vari modi ragionevoli nel modello

1. come un termine lineare,

$$y_i = \beta_0 + \beta z_i + \text{other covariates} + \varepsilon_i$$

2. come un termine quadratico,

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \text{other covariates} + \varepsilon_i$$

   - ma si potevano usare altre funzioni, quali

$$y_i = \beta_0 + \beta \log z_i + \text{other covariates} + \varepsilon_i$$

$$y_i = \beta_0 + \beta \sqrt{z_i} + \text{other covariates} + \varepsilon_i$$

3. in versione discretizzata (opzione spesso scelta per la tariffazione)
   - è arbitraria la scelta del numero di classi e dei confini delle classi

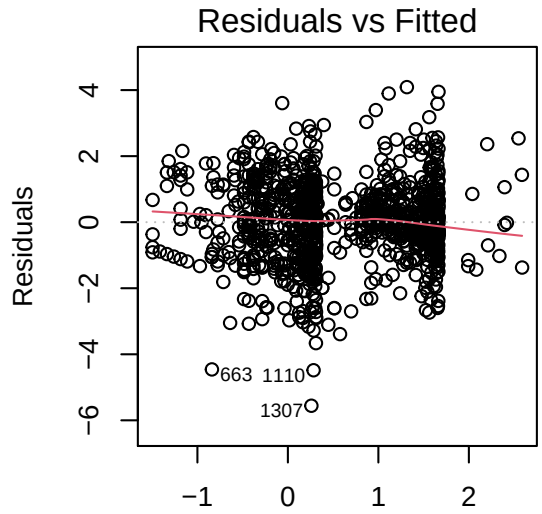
Si ottengono risultati diversi (in termini ad esempio di previsioni) e possiamo ragionevolmente provare un numero limitato di alternative.

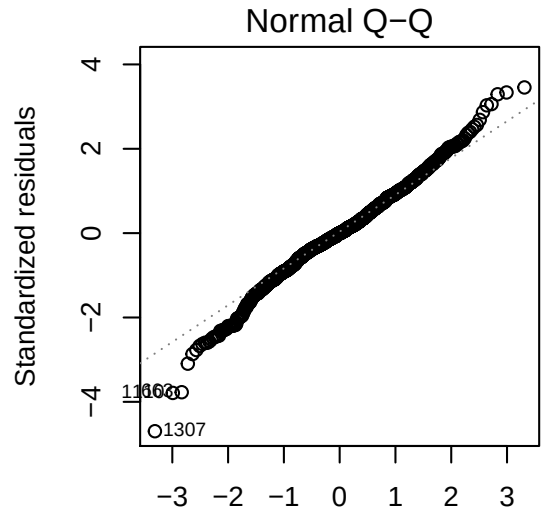
Linear model for log(LOSS): modello selezionato

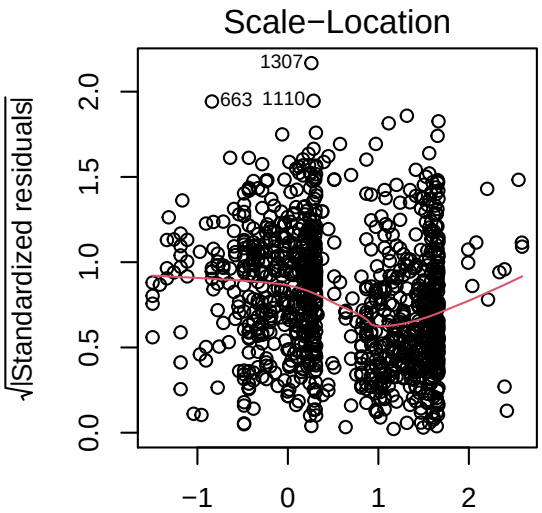
Il modello selezionato include `ATTORNEY`, `CLMAGE` (effetto quadratico) e `SEATBELT`.

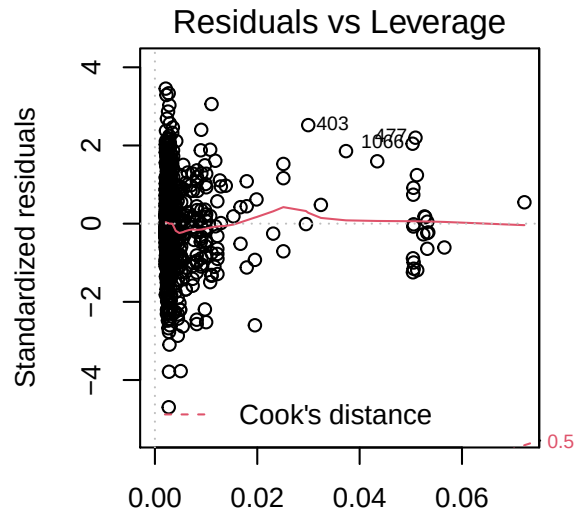
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2249	0.1376	-1.63	0.1024
ATTORNEYno	-1.3522	0.0725	-18.66	¡0.0001
CLMAGE	0.0828	0.0075	11.05	¡0.0001
I(CLMAGE^2)	-0.0009	0.0001	-9.52	¡0.0001
SEATBELTno	0.9241	0.2681	3.45	0.0006
	$s = 1.2$		R^2_{adj}=0.32	

Analisi dei residui: modello lineare per log(LOSS)









Spline per CLMAGE

Consideriamo allora un modello con una spline per CLMAGE

$$y_i = \beta_0 + f(z_i) + \texttt{other covariates} + \varepsilon_i$$

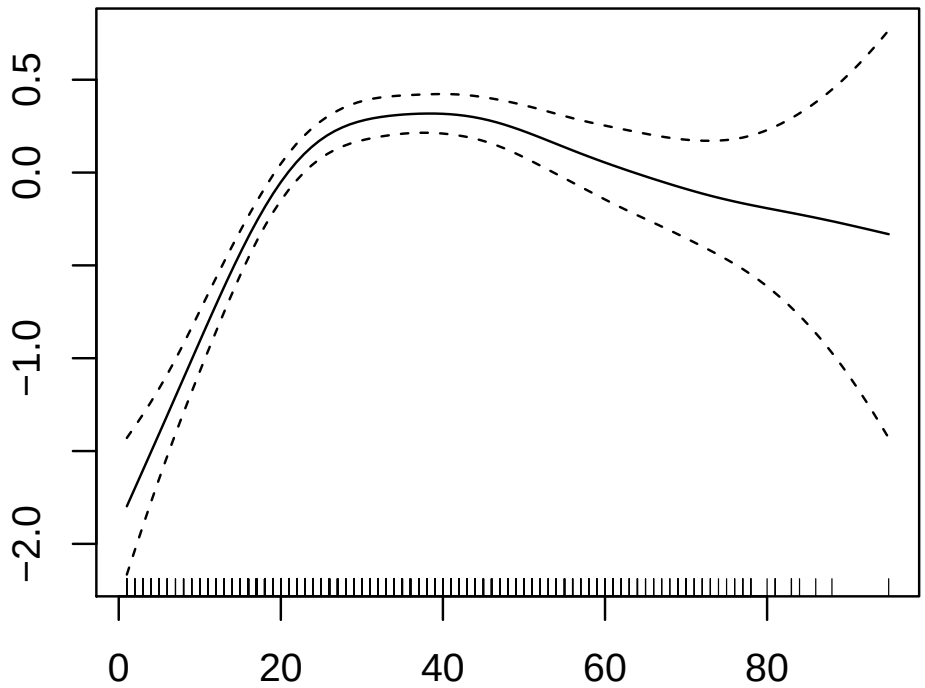
più precisamente avremo

$$y_i = \beta_0 + f(\texttt{CLMAGE}_\texttt{i}) + \beta_1 \texttt{ATTORNEYno}_\texttt{i} + \beta_2 \texttt{SEATBELTno}_\texttt{i} + \varepsilon_\texttt{i}$$

Usiamo la funzione gam() del pacchetto mgcv per la stima.

Spline per CLMAGE

```
par(mar=c(2,2,0.5,0.5))
modSopt=gam(log(LOSS) ~ s(CLMAGE) + ATTORNEY + SEATBELT,
            data = autobi,
            subset = complete.cases(autobi))
plot(modSopt)
```

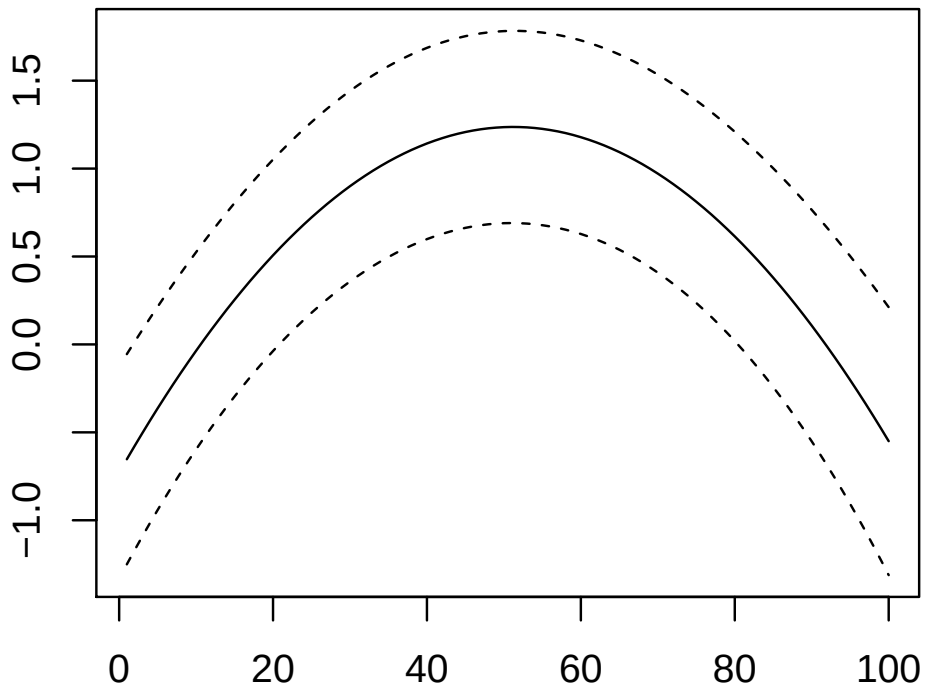


Spline per CLMAGE

```
summary(modSopt)



Family: gaussian
Link function: identity

Formula:
log(LOSS) ~ s(CLMAGE) + ATTORNEY + SEATBELT

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.26460    0.04968  25.454  < 2e-16 ***
ATTORNEYno  -1.35649    0.07179 -18.896  < 2e-16 ***
SEATBELTno   0.89874    0.26535   3.387 0.000732 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
            edf Ref.df     F p-value
s(CLMAGE) 4.563  5.597 28.42  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.335   Deviance explained = 33.9%
GCV =  1.384  Scale est. = 1.3743    n = 1078
```

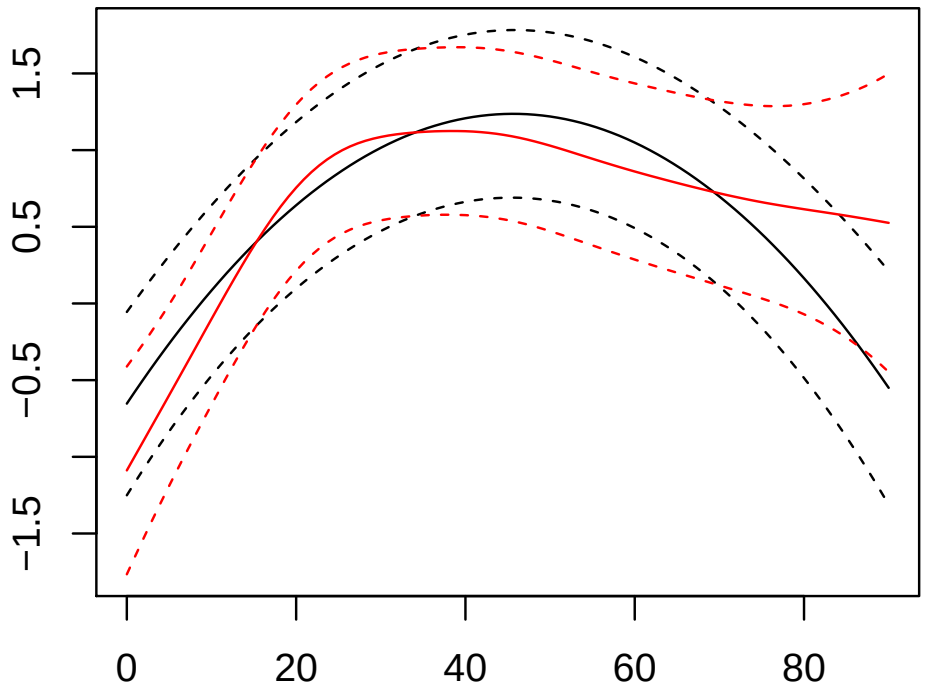
Modello quadratico per CLMAGE

```
modTrad=gam(log(LOSS) ~ CLMAGE+I(CLMAGE^2) + ATTORNEY + SEATBELT,
            data = autobi,
            subset = complete.cases(autobi))
xx=seq(0,90,length=100)
predT=predict(modTrad,se.fit=TRUE,
              newdata=data.frame(CLMAGE=xx,ATTORNEY="no",SEATBELT="no"))
matplot(cbind(predT$fit-2*predT$se.fit,predT$fit,predT$fit+2*predT$se.fit),type="l",lty=c(2,1,2),col="black")
```



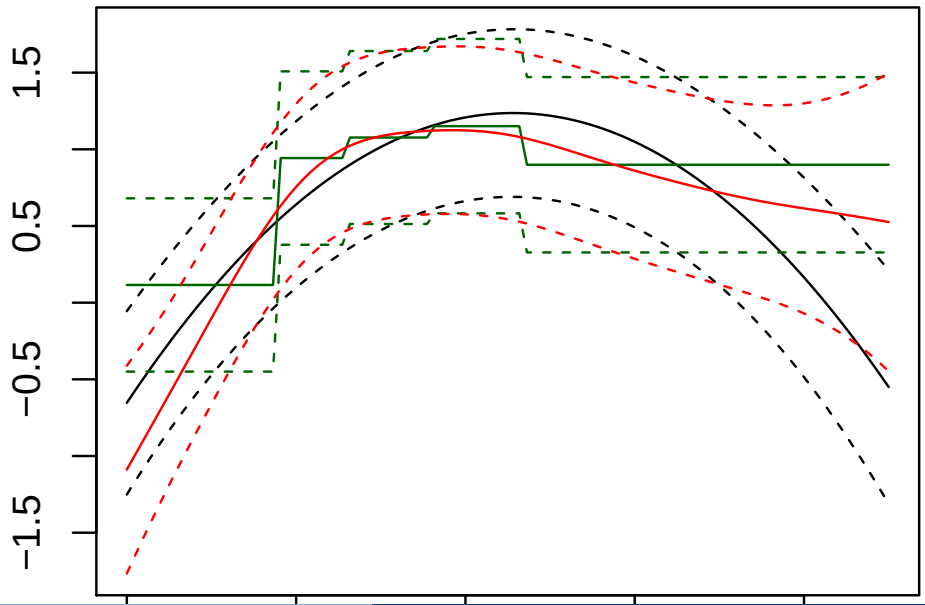
Confronto tra modelli

```
xx=seq(0,90,length=100)
predT=predict(modTrad,se.fit=TRUE,
              newdata=data.frame(CLMAGE=xx,ATTORNEY="no",SEATBELT="no"))
predS=predict(modSopt,se.fit=TRUE,
              newdata=data.frame(CLMAGE=xx,ATTORNEY="no",SEATBELT="no"))
matplot(xx,
        cbind(predT$fit-2*predT$se.fit,predT$fit,predT$fit+2*predT$se.fit,
              predS$fit-2*predS$se.fit,predS$fit,predS$fit+2*predS$se.fit),type="l",lty=c(2,1,2,2,1,2),
              col=c("black","black","black","red","red","red"),ylab="")
```



Confronto tra modelli

```
xx=seq(0,90,length=100)
xclass=as.factor(1+(xx>18)+(xx>26)+(xx>36)+(xx>47))
modClass=gam(log(LOSS) ~ AGECLASS + ATTORNEY + SEATBELT,
             data = autobi,
             subset = complete.cases(autobi))
predC=predict(modClass,se.fit=TRUE,
             newdata=data.frame(AGECLASS=xclass,ATTORNEY="no",SEATBELT="no"))
predS=predict(modSopt,se.fit=TRUE,
             newdata=data.frame(CLMAGE=xx,ATTORNEY="no",SEATBELT="no"))
matplot(xx,
        cbind(predT$fit-2*predT$se.fit,predT$fit,predT$fit+2*predT$se.fit,
              predC$fit-2*predC$se.fit,predC$fit,predC$fit+2*predC$se.fit,
              predS$fit-2*predS$se.fit,predS$fit,predS$fit+2*predS$se.fit),type="l",lty=c(2,1,2,2,1,2,2,1,2),
              col=c("black","black","black","darkgreen","darkgreen","darkgreen","red","red","red"),ylab="")
```

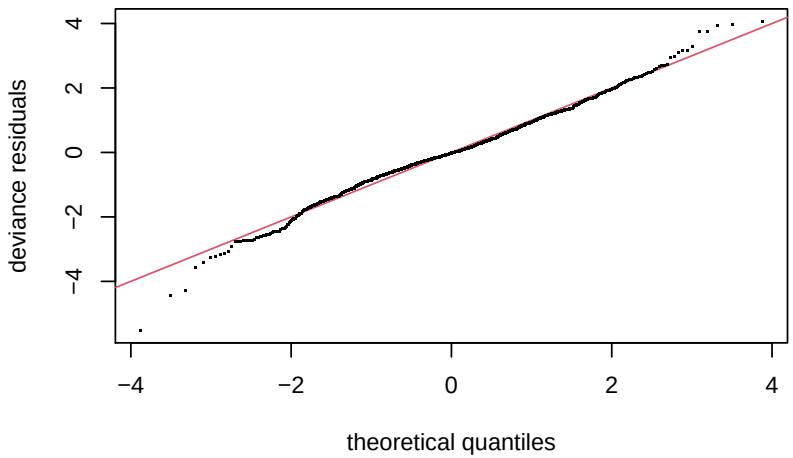


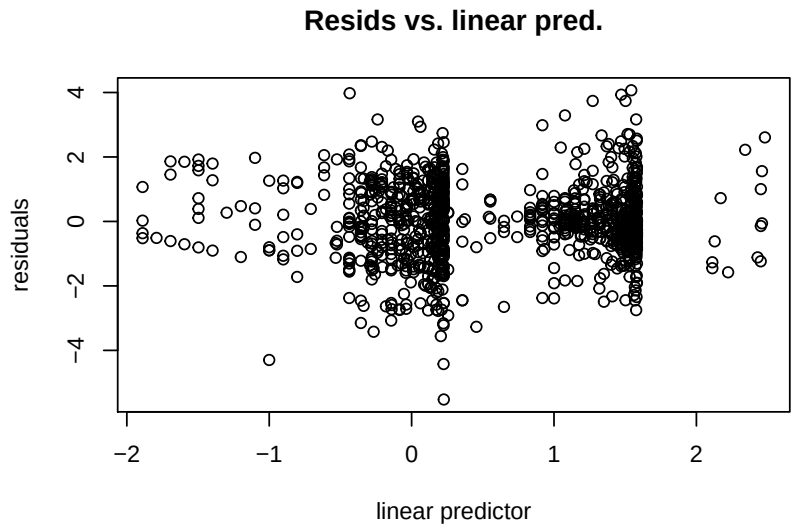
Confronto tra i modelli

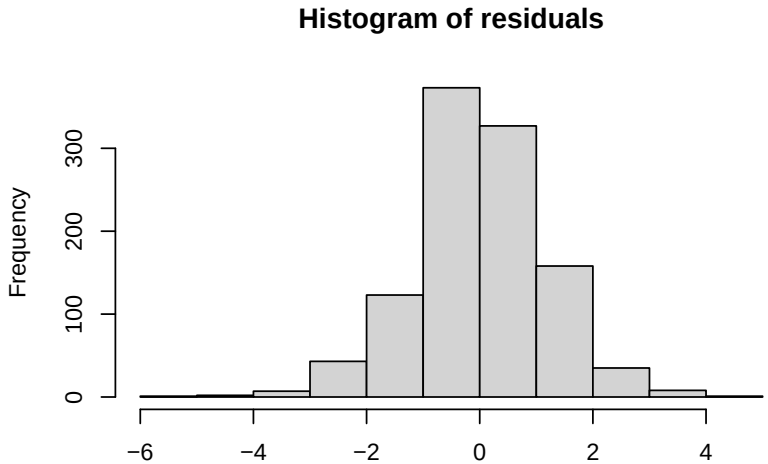
```
anova(modTrad,modSopt)

Analysis of Deviance Table

Model 1: log(LOSS) ~ CLMAGE + I(CLMAGE^2) + ATTORNEY + SEATBELT
Model 2: log(LOSS) ~ s(CLMAGE) + ATTORNEY + SEATBELT
  Resid. Df Resid. Dev     Df Deviance
1    1073.0     1507.0
2    1069.4     1471.1 3.5966   35.853

modClass$gcv.ubre

  GCV.Cp
1.446484

modTrad$gcv.ubre

  GCV.Cp
1.410998

modSopt$gcv.ubre

  GCV.Cp
1.384032
```

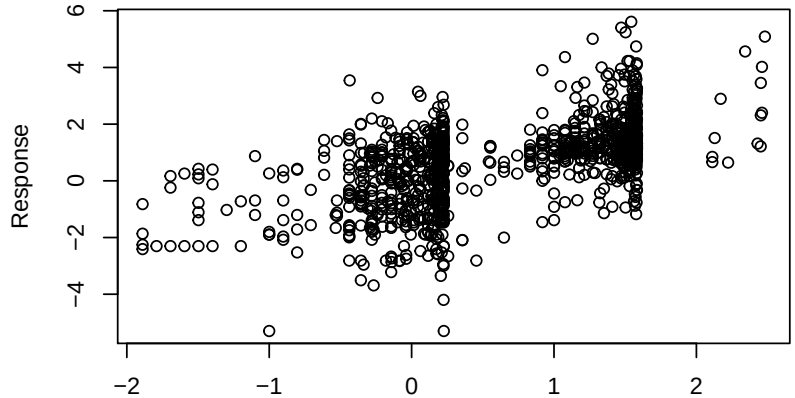
Spline per CLMAGE

```
gam.check(modSopt)
```









Indice

Assicurazione RC

I dati riguardano 67856 polizze di assicurazione auto in forza nel biennio 2004-5. Di queste, 4624 (6.8%) hanno avuto almeno un sinistro. (Fonte: De Jong and Heller (2008).)

- `exposure` ($\in [0,1]$) porzione dell'anno per cui la polizza è attiva
- `numclaims` numero di sinistri
- `claimcst0` ammontare dei sinistri (0 in assenza)
- `veh_value` valore del veicolo ($10,000)
- `veh_body` tipo di veicolo (13 valori)
- `veh_age` età del veicolo, classi: 1 (più nuovo), 2, 3, 4
- `gender` sesso del guidatore: M, F
- `area` area di residenza: A, B, C, D, E, F
- `agecat` età del guidatore: 1 (più giovane), 2, 3, 4, 5, 6

Un portafoglio di polizze danni

- Portafoglio di n rischi (polizze).
- Per ciascuna polizza i si osservano $N_i \in \{0, 1, 2, \ldots\}$ sinistri.
- Il valore di ciascun sinistro per le polizze per cui $N_i > 0$ è D_{ij}, $j = 1, \ldots, N_i$.
- Il costo dei sinistri per l'i-esima polizza è

$$L_i = \begin{cases} 0 & \text{if } N_i = 0 \\ \sum_{j=1}^{N_i} D_{ij} & \text{altrimenti} \end{cases}$$

- La perdita totale del portafoglio

$$V = \sum_{i=1}^{n} L_i.$$

- le osservazioni a disposizione sono
  - `numclaims`: N_i;
  - `claimcst0`: L_i;
  - alcune caratteristiche di ciascun rischio/polizza, $\boldsymbol{x}_i$.

Il costo dei sinistri

L'obiettivo è di modellare la relazione tra i sinistri e le caratteristiche del rischio, cioè

$$L_i|\boldsymbol{x}_i$$

in quanto questo modello, una volta stimato, potrà essere impiegato per prevedere su un insieme di nuovi rischi (polizze) con caratteristiche $\boldsymbol{x}_i'$

- il costo medio dei sinistri per polizza

$$\mathrm{E}(L_i'|\boldsymbol{x}_i')$$

  per determinare i premi.
- la probabilità che il costo totale dei sinistri ecceda una certa soglia (per decidere sulle riserve o sulla riassicurazione)
- il valore atteso di trasformazioni di V' quali

$$\min(\max(0, V - a), b)$$

  (che corrisponde ai pagamenti attesi per una riassicurazione a strati $(b - a)$ in eccesso di a.)

Modello statistico per il costo dei sinistri per polizza

- Modello per il numero di sinistri N_i
- Modello per il costo dei singoli sinistri D_{ij}
- Combinare i risultati

Modello statistico per il costo dei sinistri per polizza

- Modello per il numero di sinistri N_i
  - $\mapsto$ Modello di regressione ove la distribuzione di N_i dipende da $\boldsymbol{x}_i$, e.g.: GLM Poisson o Binomiale negativa.

- Modello per il costo dei singoli sinistri D_{ij}
  - $\mapsto$ Modello di regressione ove la distribuzione di D_{ij} dipende da $\boldsymbol{x}_i$, e.g.: GLM gamma o gaussiana inversa, oppure ML su un opportuno trasformato di D.

- Combinare i risultati

Modello statistico per il costo dei sinistri per polizza

- Modello per il numero di sinistri N_i
  - $\mapsto$ Modello di regressione ove la distribuzione di N_i dipende da $\boldsymbol{x}_i$, e.g.: GLM Poisson o Binomiale negativa.
  - $\mapsto$ con questo modello si ottiene, ad esempio, una stima di

$$\mathrm{E}(N|\boldsymbol{x})$$

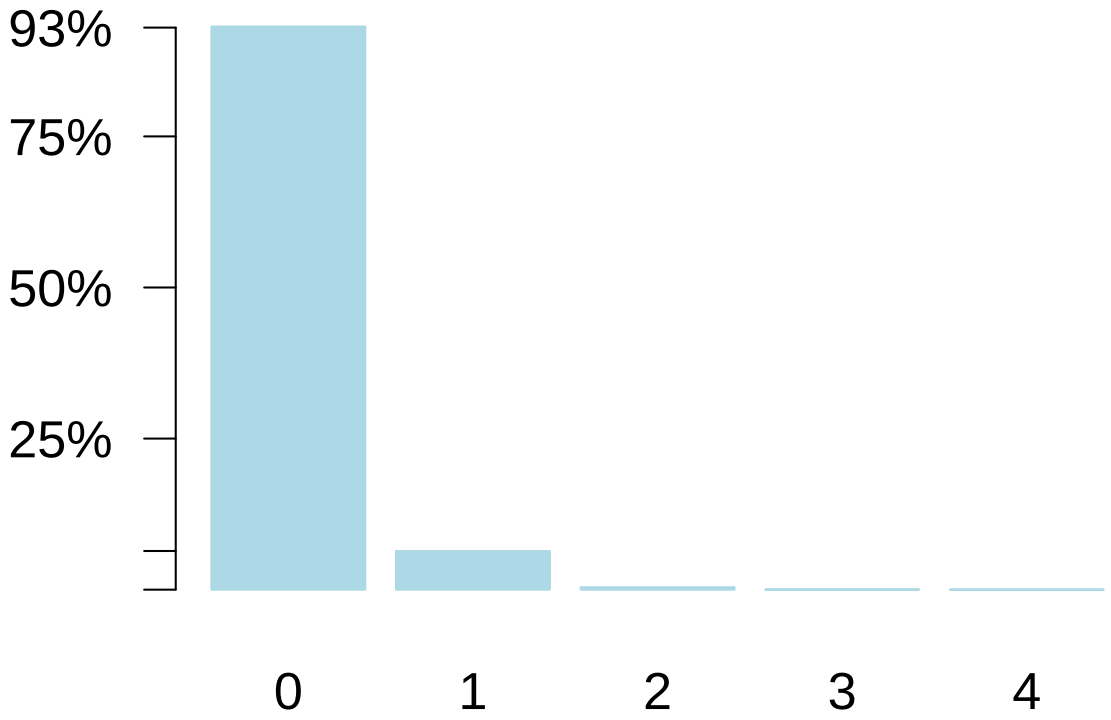
- Modello per il costo dei singoli sinistri D_{ij}
  - $\mapsto$ Modello di regressione ove la distribuzione di D_{ij} dipende da $\boldsymbol{x}_i$, e.g.: GLM gamma o gaussiana inversa, oppure ML su un opportuno trasformato di D.
  - $\mapsto$ con questo modello si ottiene, ad esempio, una stima di

$$\mathrm{E}(D|\boldsymbol{x})$$

- Combinare i risultati
  - $\mapsto$ ad es. $\boxed{\mathrm{E}(L|\boldsymbol{x}) = \mathrm{E}(N|\boldsymbol{x})\mathrm{E}(D|\boldsymbol{x})}$
  - $\mapsto$ altro.

sinistri

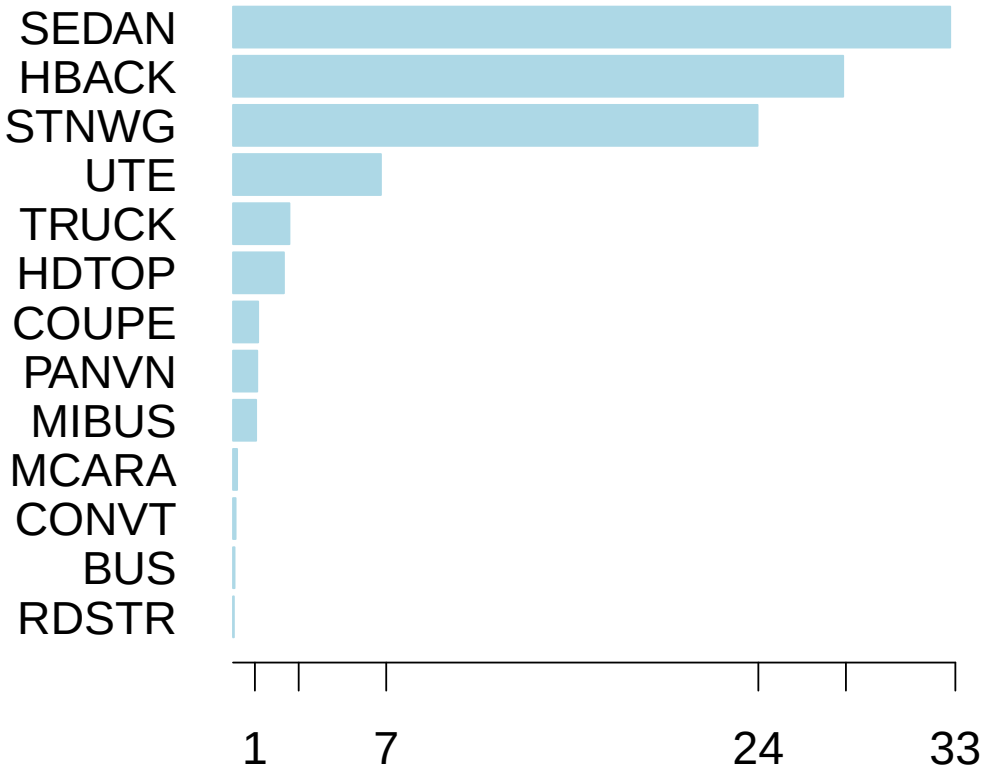
Il numero di sinistri per polizza è un processo di conteggio, assume valori da 0 a 4 nel campione.



# sinistri	Freq
0	63232
1	4333
2	271
3	18
4	2

- solo 4.3 per mille delle polizze hanno più di un sinistro;
- il numero medio di sinistri per polizza è 0.072757.

Tipo di veicolo

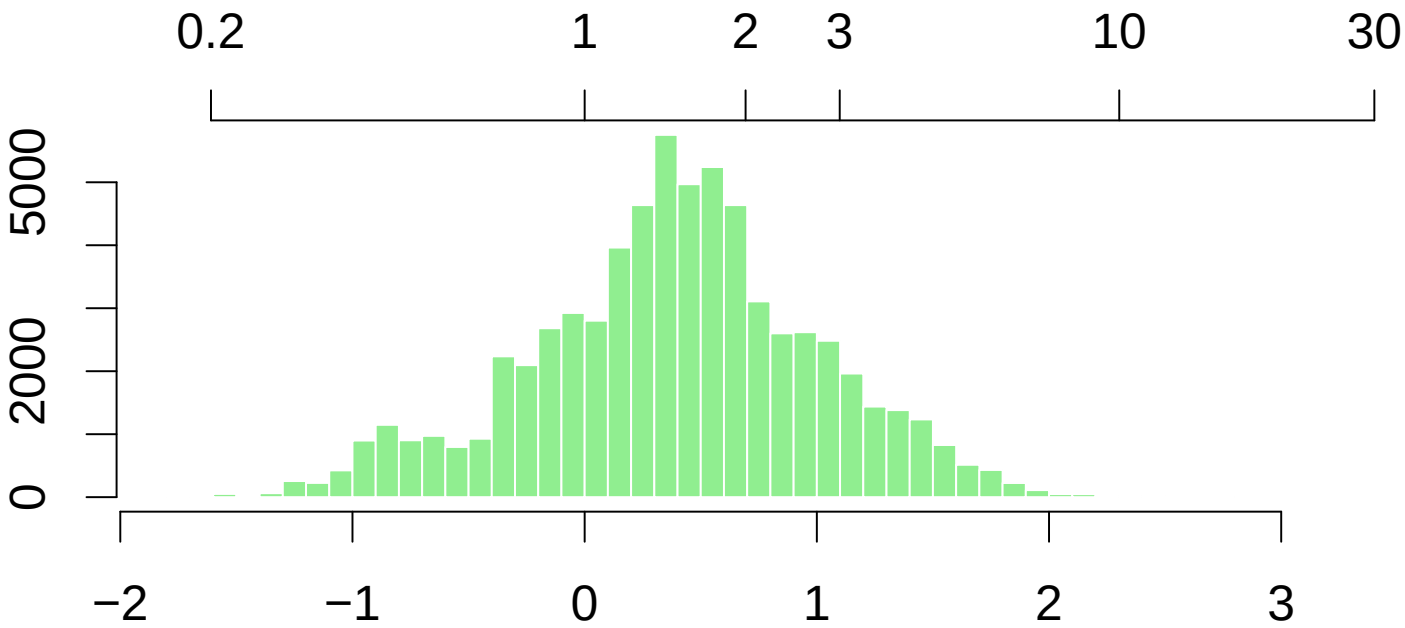


	Freq	NA
1	SEDAN	22233
2	HBACK	18915
3	STNWG	16261
4	UTE	4586
5	TRUCK	1750
6	HDTOP	1579
7	COUPE	780
8	PANVN	752
9	MIBUS	717
10	MCARA	127
11	CONVT	81
12	BUS	48
13	RDSTR	27

- Alcuni tipi sono molto particolari.
- Alcuni tipi presentano frequenze molto piccole.

Valore del veicolo

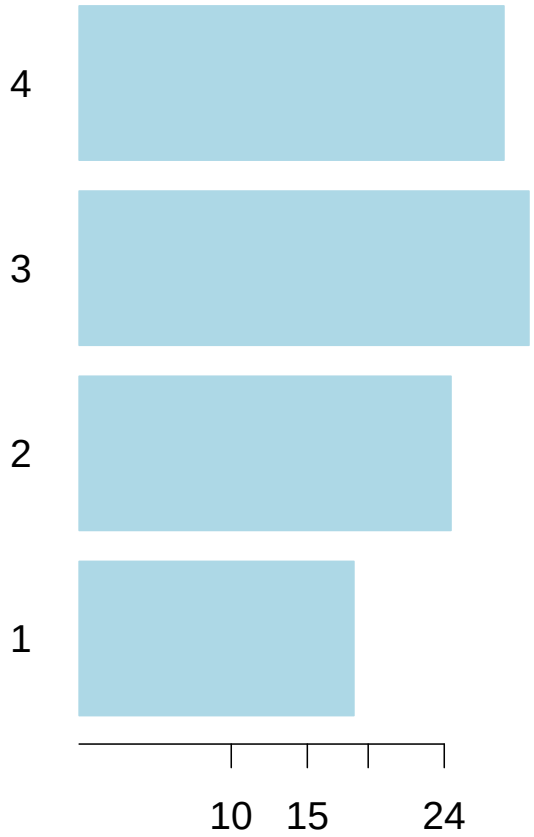
Il valore del veicolo varia da \$0 a \$345600, il valore medio è 17770.21 e la mediana è 15000



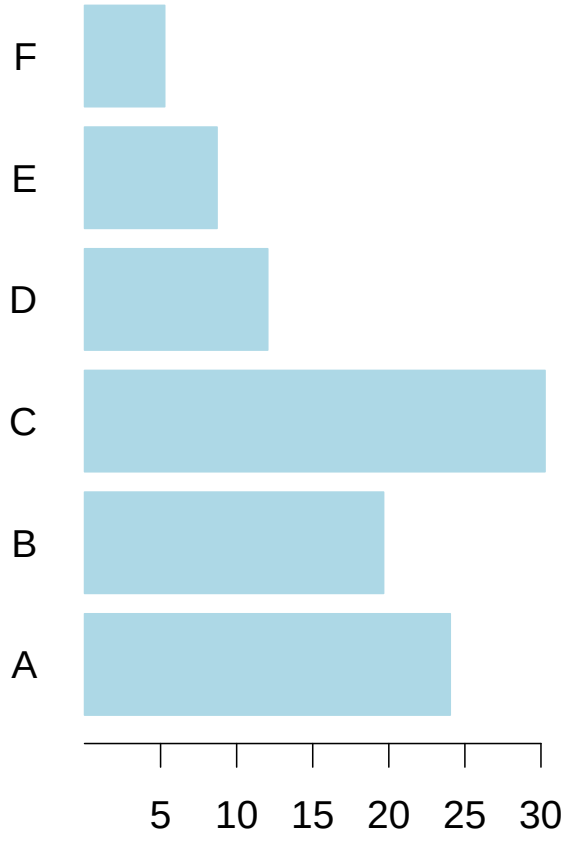
Una trasformazione logaritmica regolarizza (normalizza) la distribuzione, eliminando però 53 casi in cui la variabile assume valore 0.

Altre variabili

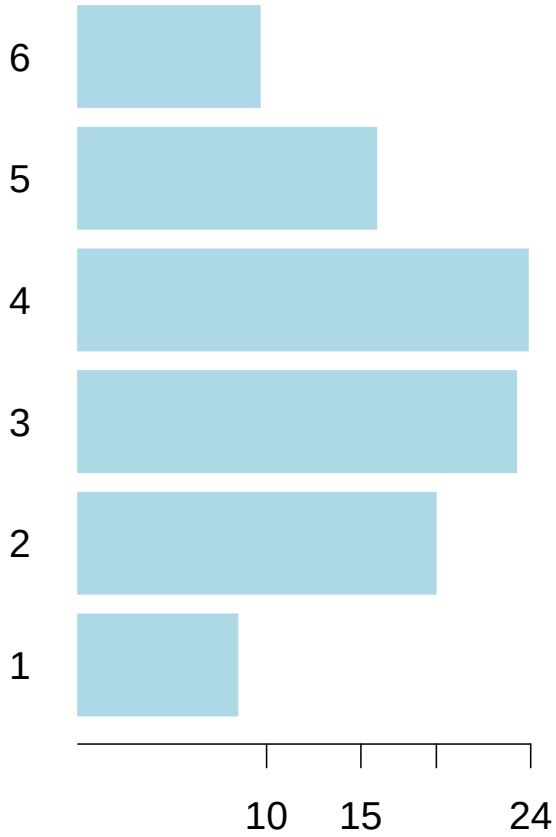
Età veicolo (categorie)



Area di residenza



Età (categorie)



GLM Poisson per il numero di sinistri

Assumiamo

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{indip.},$$

e

$$\log \lambda_i = \log e_i + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

dove

- Y_i numero di sinistri per la polizza i;
- λ_i è la sinistrosità: numero medio di sinistri per unità di tempo (un anno);
- $\boldsymbol{x}_i$ include `veh_value`, `veh_body`, `veh_age`, `gender`, `area`, `agecat` (ovvero le rilevanti variabili indicatrici per le esplicative categoriali);
- e_i è l'esposizione.
- $\boldsymbol{\beta}$ sono parametri da stimare.

Costruzione (selezione) del modello

Il modello "base" sia quello senza esplicative, ovvero

$$\log \lambda_i = \beta_0 + \log e_i$$

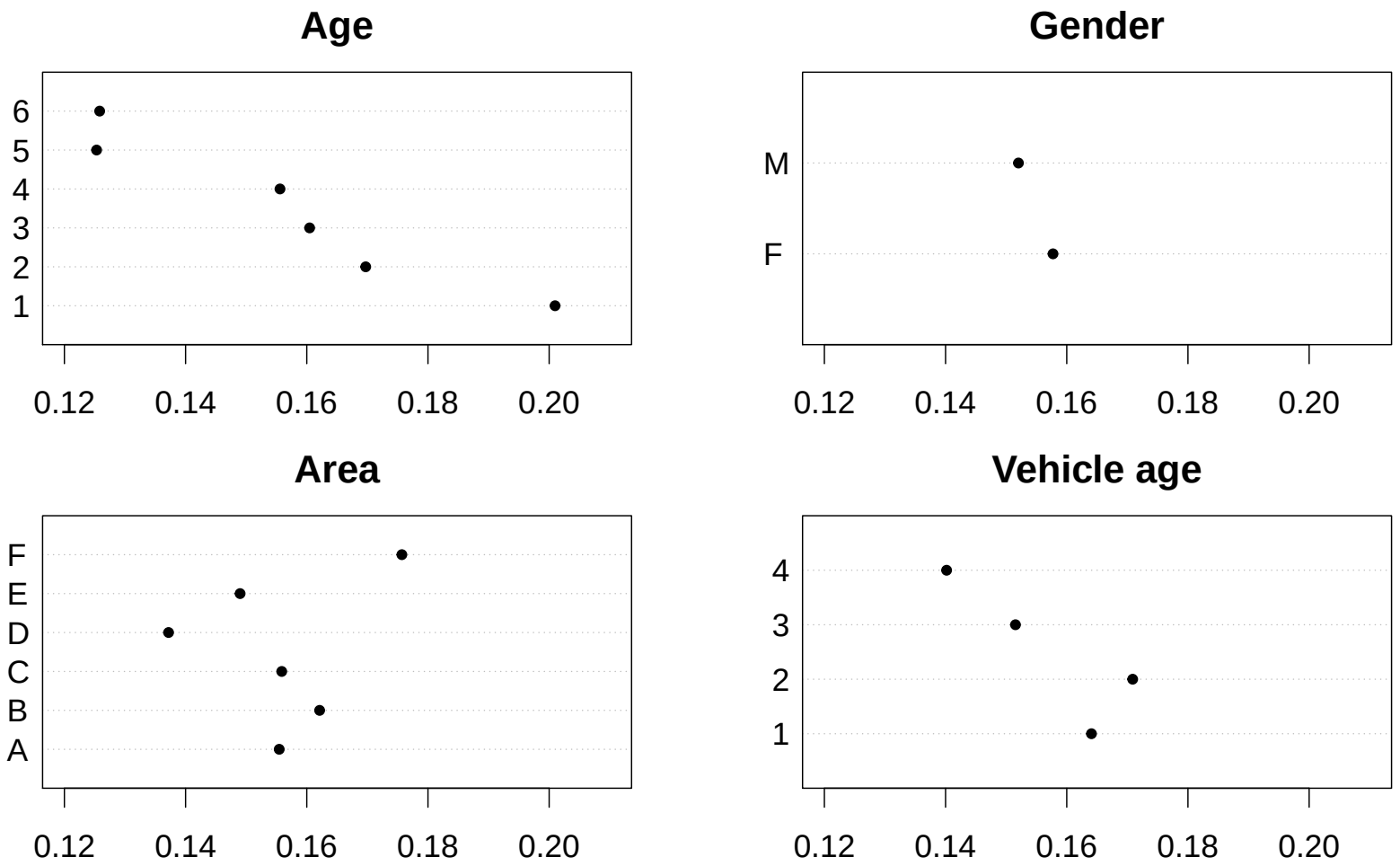
si ottiene $\widehat{\beta}_0 = -1.863$, che corrisponde a un tasso di sinistrosità annuale pari a 0.155.
Si noti che questo è pari a

$$\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} e_i}$$

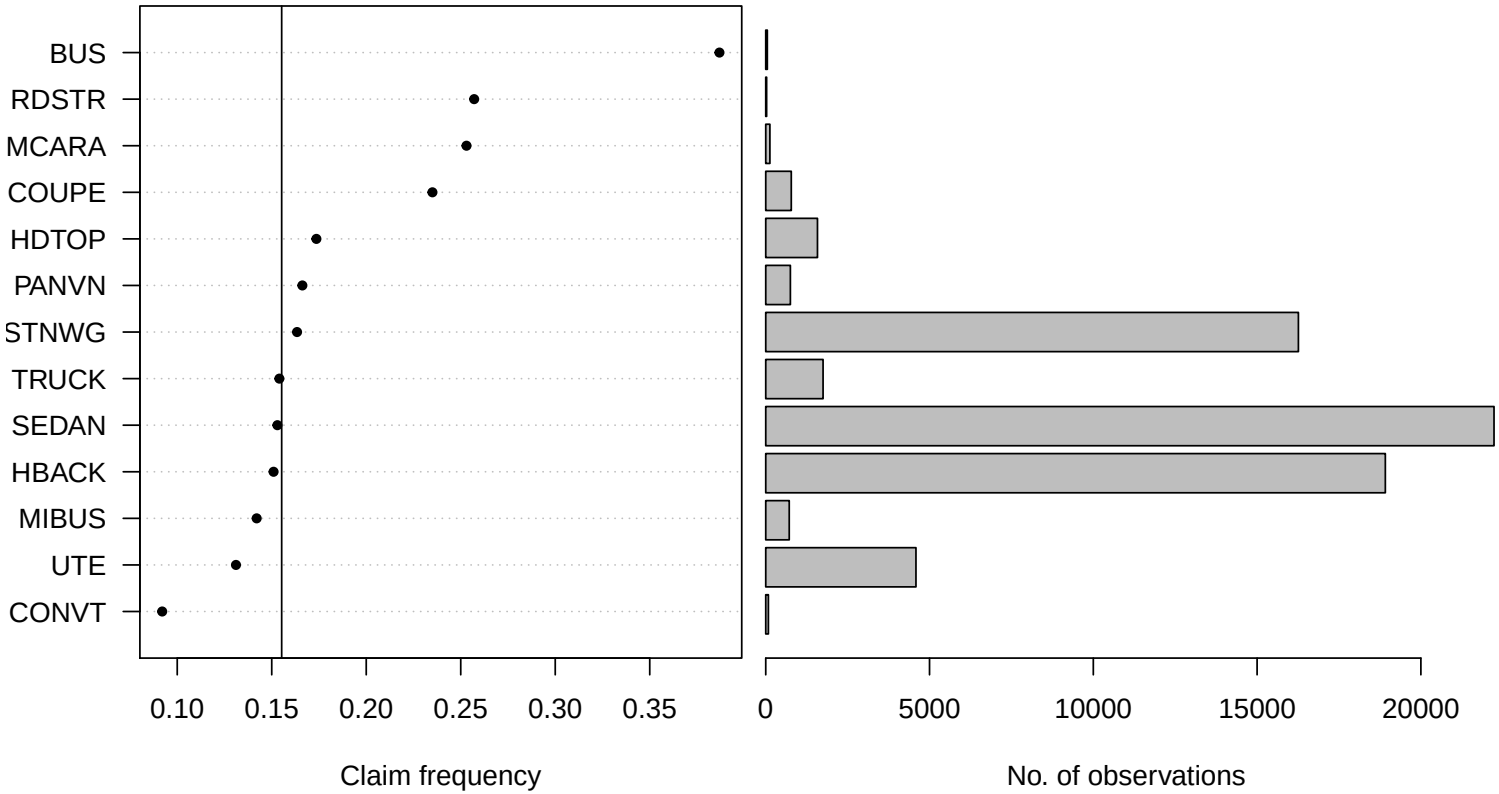
(Chiaramente maggiore del tasso grezzo di sinistrosità in quanto le esposizioni sono tutte inferiori a 1.)

Scelta di un'esplicativa (fattore di rischio)

Sinistrosità annuale per livelli dei fattori di rischio.

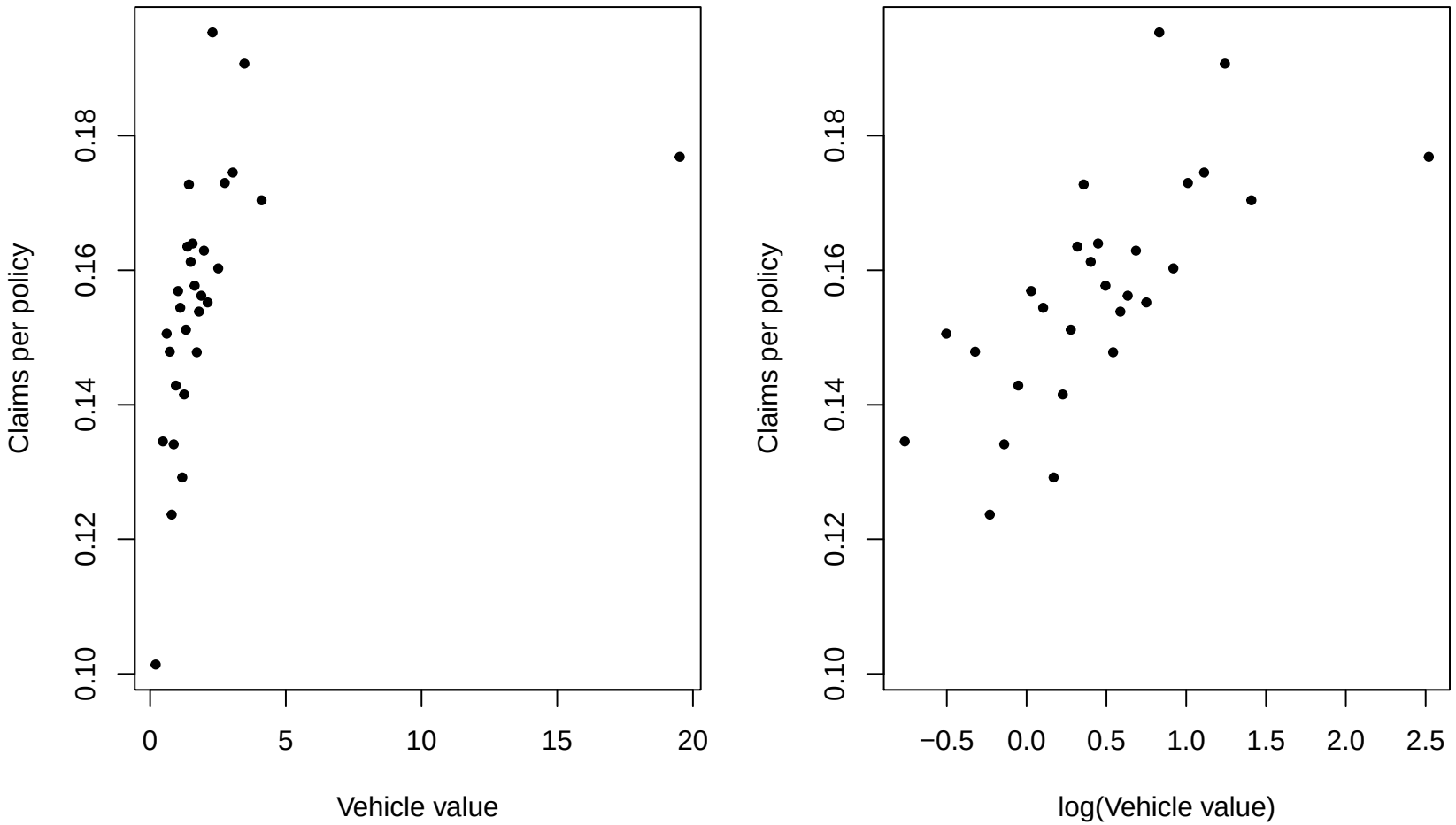


Scelta di un'esplicativa (fattore di rischio)



- i tassi variano in misura rilevante con il tipo di veicolo
- va tenuto conto che alcune classi sono poco numerose

Scelta di un'esplicativa (fattore di rischio)



Tasso di sinistrosità per classi del (logaritmo del) valore del veicolo.

Modello stimato

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6046	0.0436	-36.77	6.5653e-296
agecat2	-0.1690	0.0539	-3.14	1.7150e-03
agecat3	-0.2251	0.0524	-4.30	1.7459e-05
agecat4	-0.2560	0.0524	-4.88	1.0450e-06
agecat5	-0.4724	0.0587	-8.04	8.6762e-16
agecat6	-0.4683	0.0668	-7.01	2.4582e-12

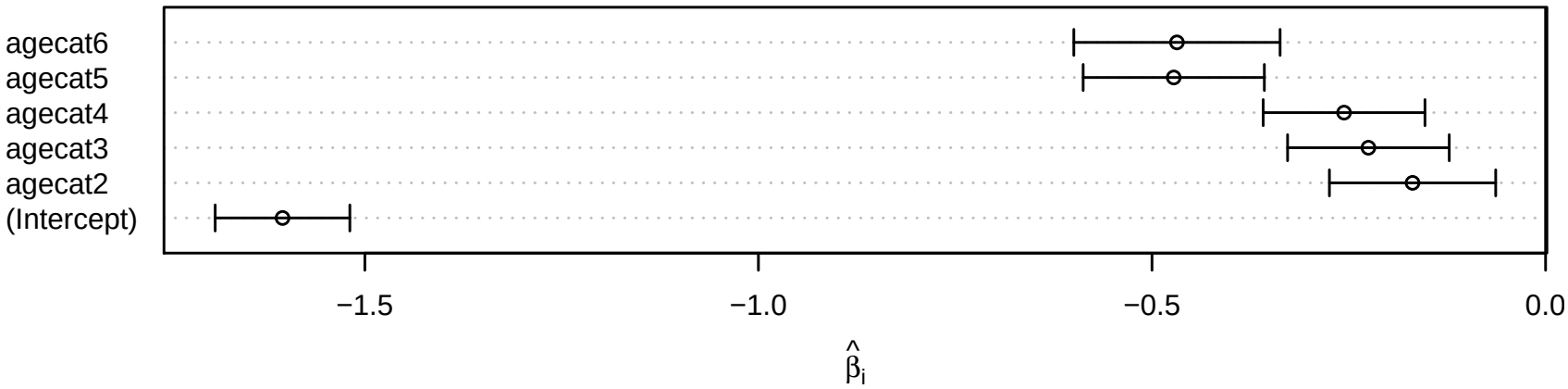
- Il tasso di sinistrosità per la cat. 1 (base) è

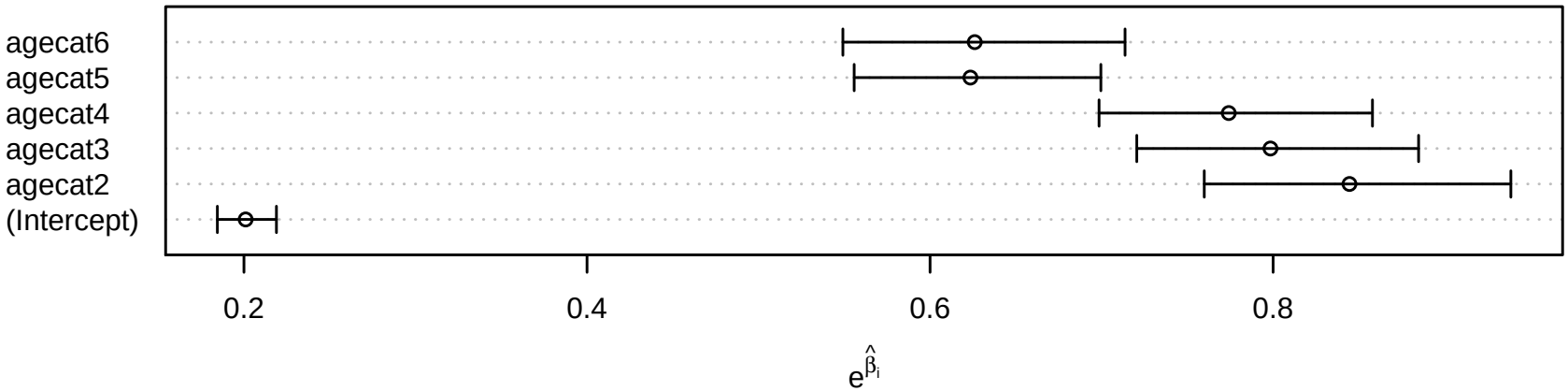
$$e^{-1.604578} = 0.2009743$$

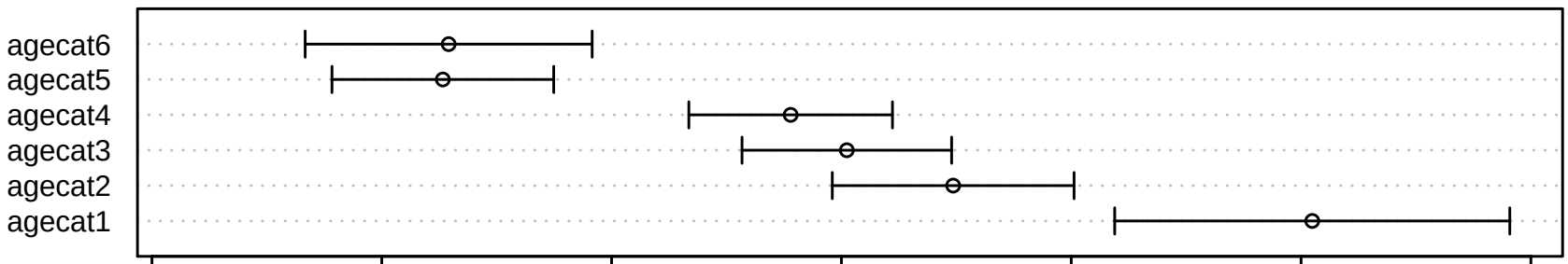
- Nelle altre categorie il tasso è inferiore: ad es. in cat. 2 è $e^{-0.1689956} = 0.8445126$ di quello nella categoria base, cioè

$$e^{-1.604578} e^{-0.1689956} = 0.1697254$$

Stime dei coefficienti







Indice

Effetto non lineare del valore del veicolo

Adottiamo un modello additivo generalizzato in cui

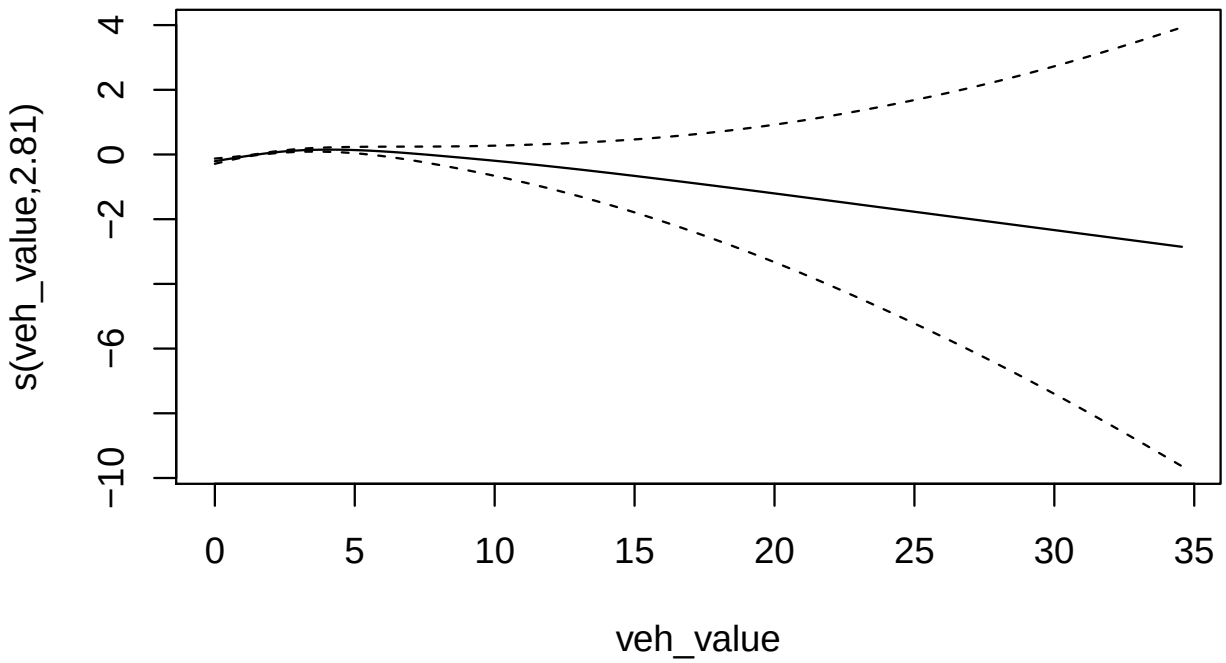
$$\log \frac{\lambda_i}{e_i} = s(v_i) + \texttt{agecat}_\texttt{i}$$

che possiamo stimare in R con

```
fitV=gam(numclaims ~ agecat + s(veh_value)+
                       offset(log(exposure)),
         data=cars,
         family=poisson(link=log))
```

Effetto non lineare del valore del veicolo:stima

```
plot(fitV)
```



Effetto non lineare del valore del veicolo:stima

```
summary(fitV)


Family: poisson
Link function: log

Formula:
numclaims ~ agecat + s(veh_value) + offset(log(exposure))

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.60935    0.04367 -36.855  < 2e-16 ***
agecat2     -0.17637    0.05391  -3.271  0.00107 **
agecat3     -0.23146    0.05241  -4.416 1.01e-05 ***
agecat4     -0.25222    0.05243  -4.810 1.51e-06 ***
agecat5     -0.47150    0.05872  -8.029 9.81e-16 ***
agecat6     -0.44550    0.06697  -6.653 2.88e-11 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(veh_value) 2.813  3.585  40.36  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0196   Deviance explained = 0.533%
UBRE = -0.62584  Scale est. = 1         n = 67856
```

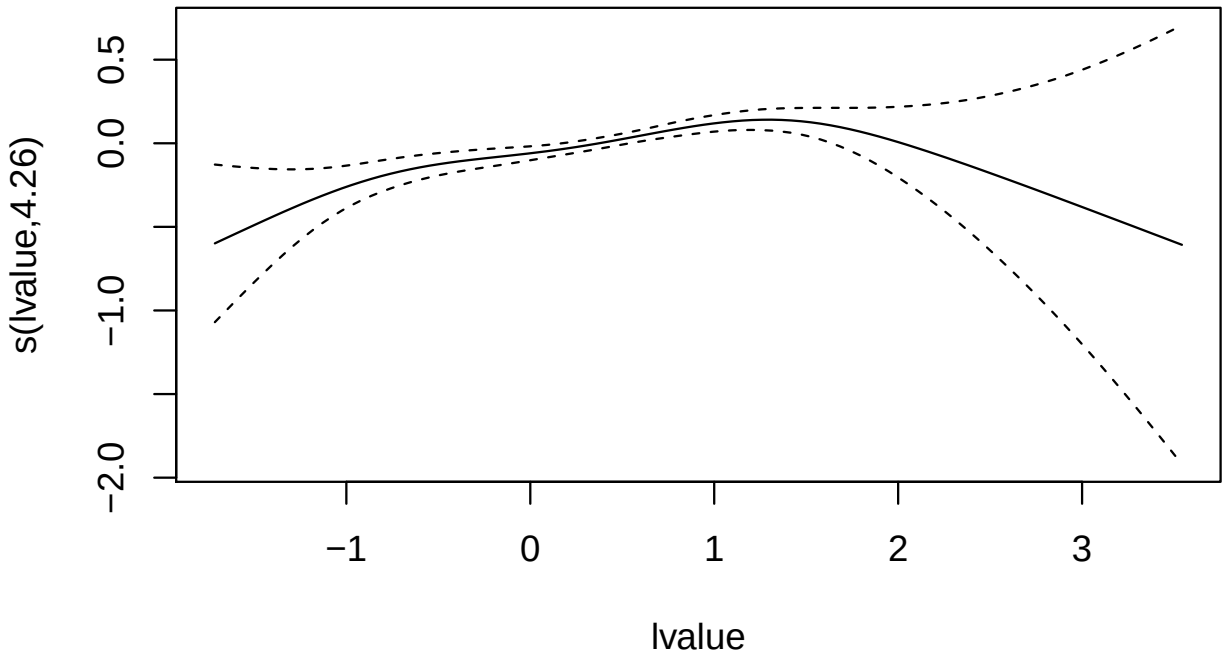
Effetto non lineare del valore del veicolo

Anche se abbiamo specificato un modello semiparametrico, conviene comunque operare la trasformazione logaritmica del valore del veicolo.

```
cars$lvalue=log(cars$veh_value)
fitLV=gam(numclaims ~ agecat + s(lvalue)+
                        offset(log(exposure)),
          data=cars[is.finite(cars$lvalue),],
          family=poisson(link=log))
```

Effetto non lineare del valore del veicolo:stima

```
plot(fitLV)
```



Effetto non lineare del valore del veicolo:stima

```
summary(fitLV)


Family: poisson
Link function: log

Formula:
numclaims ~ agecat + s(lvalue) + offset(log(exposure))

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.61289    0.04375 -36.862  < 2e-16 ***
agecat2     -0.17632    0.05398  -3.266  0.00109 **
agecat3     -0.23097    0.05249  -4.400 1.08e-05 ***
agecat4     -0.24898    0.05249  -4.744 2.10e-06 ***
agecat5     -0.46787    0.05880  -7.956 1.77e-15 ***
agecat6     -0.44054    0.06711  -6.565 5.22e-11 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
            edf Ref.df Chi.sq p-value
s(lvalue) 4.261  5.222  45.26  <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0197   Deviance explained = 0.55%
UBRE = -0.62604  Scale est. = 1         n = 67803
```

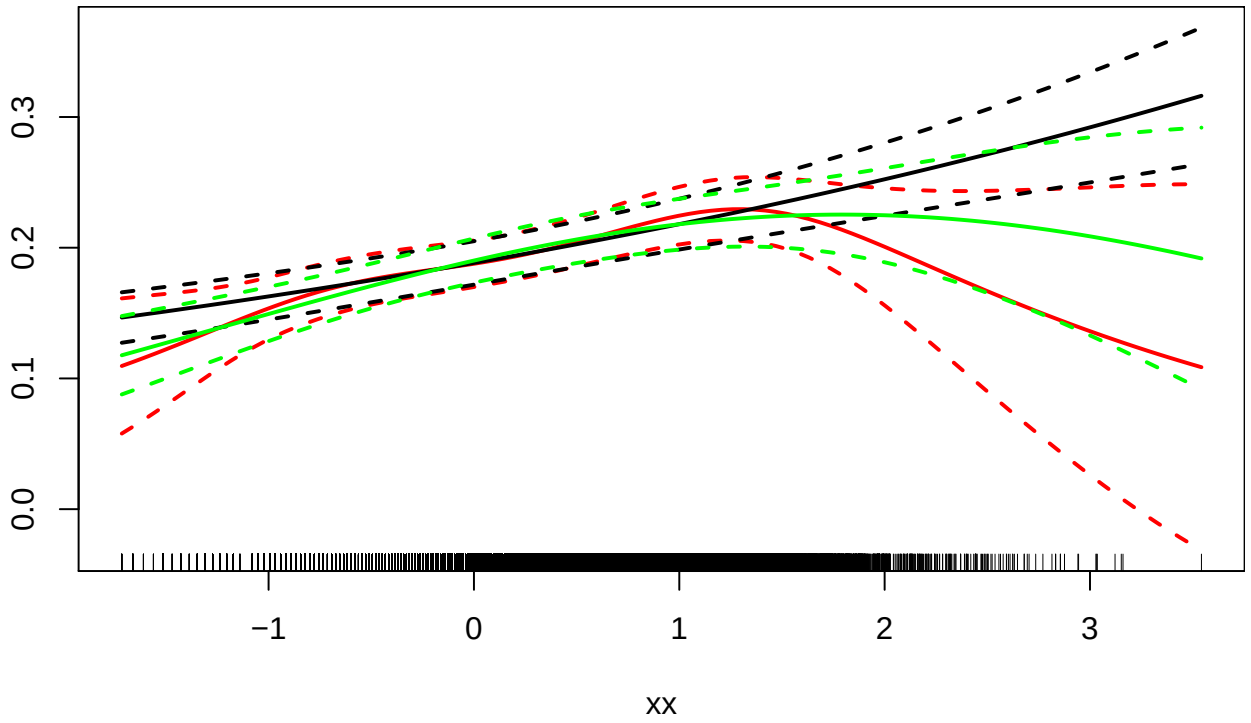
Valore veicolo: diverse opzioni

```
fitLVlin=gam(numclaims ~ agecat + lvalue
                                  +offset(log(exposure)),
             data=cars[is.finite(cars$lvalue),],
             family=poisson(link=log))
fitLVquad=gam(numclaims ~ agecat + lvalue + I(lvalue^2)
                                  +offset(log(exposure)),
             data=cars[is.finite(cars$lvalue),],
             family=poisson(link=log))
```

Valore veicolo: confronto delle opzioni

```
xx=seq(min(cars$lvalue[is.finite(cars$lvalue)],na.rm=TRUE),
       max(cars$lvalue,na.rm=TRUE),length=100)
predLV=predict(fitLV,newdata=data.frame(agecat=1,lvalue=xx,exposure=1),
               type="response",se.fit=TRUE)
predLVlin=predict(fitLVlin,newdata=data.frame(agecat=1,lvalue=xx,exposure=1),
                  type="response",se.fit=TRUE)
predLVquad=predict(fitLVquad,newdata=data.frame(agecat=1,lvalue=xx,exposure=1),
                   type="response",se.fit=TRUE)

matplot(xx,cbind(predLV$fit-1.96*predLV$se.fit,
                 predLV$fit,
                 predLV$fit+1.96*predLV$se.fit,
                 predLVlin$fit-1.96*predLVlin$se.fit,
                 predLVlin$fit,
                 predLVlin$fit+1.96*predLVlin$se.fit,
                 predLVquad$fit-1.96*predLVquad$se.fit,
                 predLVquad$fit,
                 predLVquad$fit+1.96*predLVquad$se.fit),
        col=c("red","red","red","black","black","black","green","green","green"),
        type="l",lty=c(2,1,2,2,1,2,2,1,2),lwd=2,ylab="")
rug(cars$lvalue)
```

Valore veicolo: confronto delle opzioni



Valore del veicolo: confronto con CV

Seleziono un *training set* e un *test set*

```
cars1=cars[is.finite(cars$lvalue),]
indici=sample(1:nrow(cars1),15000)
## TRAINING SET
carsTR=cars1[-indici,]
## TEST SET
carsT=cars1[indici,]
```

Valore del veicolo: confronto con CV

Stimo i modelli alternativi sul training set

```
fitLVlin=gam(numclaims ~ agecat + s(lvalue)+offset(log(exposur
             data=carsTR,
             family=poisson(link=log))
fitLVlin=gam(numclaims ~ agecat + lvalue+offset(log(exposure))
             data=carsTR,
             family=poisson(link=log))
fitLVquad=gam(numclaims ~ agecat + lvalue + I(lvalue^2)+offset
             data=carsTR,
             family=poisson(link=log))
```

Uso le stime per fare una previsione sul test set

```
predLV=predict(fitLV,newdata=carsT,type="response")
predLVlin=predict(fitLVlin,newdata=carsT,type="response")
predLVquad=predict(fitLVquad,newdata=carsT,type="response")
```

Valore del veicolo: confronto con CV

Calcolo la distorsione sul test set

```
mean(carsT$numclaims-predLV)

[1] -0.002370104

mean(carsT$numclaims-predLVquad)

[1] -0.003077286

mean(carsT$numclaims-predLVlin)

[1] -0.003122794
```

Valore del veicolo: confronto con CV

Calcolo l'errore di previsione sul test set

```
sqrt(mean((carsT$numclaims-predLV)^2))

[1] 0.2709319

sqrt(mean((carsT$numclaims-predLVquad)^2))

[1] 0.2709816

sqrt(mean((carsT$numclaims-predLVlin)^2))

[1] 0.2709965
```