Modello lineare:
un esempio guidato con R

October 30, 2019

I dati e l'obiettivo dell'analisi[1]

In questo esempio, si esaminano i dati della Survey of Consumer Finances (SCF), indagine condotta in USA su un campione di 275 cittadini che hanno acquistato una polizza vita e per i quali si raccolgono informazioni sul loro reddito e sulle loro caratteristiche demografiche.

i dati sono disponibili su `https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html`

Obiettivo dell'analisi è determinare quali caratteristiche influenzino la spesa per polizze vita. Per quanto riguarda le polizze temporanee in caso morte (term life insurance), l'ammontare che viene rilevato è FACE, ovvero quanto la compagnia pagherà in caso di morte dell'assicurato.

Le informazioni disponibili sono riassunte nella seguente tabella:

VARIABILE	DESCRIZIONE
AGE	Age of the survey respondent
MARSTAT	Marital status of the respondent (single/not single)
EDUCATION	Number of years of education of the survey respondent
NUMHH	Number of household members
INCOME	Annual income of the family
FACE	Amount that the company will pay in the event of the death of the named insured

Prime analisi esplorative

Col seguente comando carichiamo i dati sul workspace di R:

```
> TL <- read.csv("TL.csv", header=TRUE, sep=",", row.names=1)
```

L'oggetto `TL` contiene quindi i dati, essi possono essere visualizzati col comando:

```
> TL
```

Spesso tuttavia è utile dare un'occhiata solo alle prime righe col comando:

```
> head(TL)
```

```
  AGE    MARSTAT EDUCATION NUMHH INCOME    FACE
1  30 not single        16     3  43000   20000
2  50 not single         9     3  12000  130000
3  39 not single        16     5 120000 1500000
4  43 not single        17     4  40000   50000
5  34 not single        11     4  28000  220000
6  29 not single        16     3 100000  600000
```

Il comando

```
> str(TL)
```

[1] tratto da: Frees, E. W. (2010), *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press.

```
'data.frame':        275 obs. of  6 variables:
 $ AGE      : int  30 50 39 43 34 29 72 51 58 73 ...
 $ MARSTAT  : Factor w/ 2 levels "not single","single": 1 1 1 1 1 1 1 1 2 1 ...
 $ EDUCATION: int  16 9 16 17 11 16 17 16 14 12 ...
 $ NUMHH    : int  3 3 5 4 4 3 2 4 1 2 ...
 $ INCOME   : int  43000 12000 120000 40000 28000 100000 112000 15000 32000 25000 ...
 $ FACE     : int  20000 130000 1500000 50000 220000 600000 100000 2500000 250000 50000 ...
```

fornisce invece la struttura dell'oggetto TL.

La variabile risposta che consideriamo `FACE`, e come prima possibile variabile esplicativa usiamo solo `INCOME` che ragionevole pensare che sia legata alla prima.

Un modo per accedere direttamente alle variabili del data frame invece di scrivere per esteso `TL$nome variabile` è quello di usare

```
> attach(TL)
```

Una prima descrizione di una variabile quantitativa si ottiene usando il comando generico `summary`

```
> summary(FACE)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     800    50000   150000   747581   590000 14000000
```

```
> summary(INCOME)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     260    38000    65000   208975   120000 10000000
```

`summary` è una funzione generica di R il cui effetto dipende dal tipo di oggetto cui <U+00E8> applicata.

Le due variabili sembrano avere una forte asimmetria positiva (si vedano media e mediana, ove la seconda è di molto minore della prima) questo appare chiaramente anche dal diagramma di dispersione:

```
> plot(INCOME, FACE)
```

la grande maggioranza dei valori sono in basso a sinistra mentre solo pochi valori, molto elevati appaiono in alto. È una conseguenza della forte asimmetria, e questo maschera la relazione fra le variabili.

Costruzione del modello

Modello di regressione semplice

Come primo banale esercizio, stimiamo un modello lineare semplice di `FACE` vs `INCOME` ignorando per ora il problema dell'asimmetria delle due variabili.

$$FACE_i = \beta_0 + \beta_1 INCOME_i + \epsilon_i.$$

La funzione da utilizzare per il modello lineare è `lm()`.

```
> m0<- lm(FACE~INCOME)
```

Una sintesi dei risultati della stima del modello si può visualizzare con il comando generico `summary`:

```
> summary(m0)
```

```
Call:
lm(formula = FACE ~ INCOME)

Residuals:
     Min       1Q   Median       3Q      Max
-3069185  -628950  -551689  -167579 12976542

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.553e+05  1.019e+05   6.433 5.59e-10 ***
```

```
INCOME      4.414e-01  1.200e-01   3.677 0.000284 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1637000 on 273 degrees of freedom
Multiple R-squared:  0.04718,        Adjusted R-squared:  0.04369
F-statistic: 13.52 on 1 and 273 DF,  p-value: 0.0002843
```

Innanzitutto proviamo a dare un'interpretazione dei coefficienti stimati

- $\hat{\beta}_0$: questo parametro rappresenta il valore atteso dell'ammontare del valore della polizza quando x=0, ed <U+00E8> pari a 655300. Non ha molto senso interpretarlo in questo caso poichè y ha sempre valori positivi.
- $\hat{\beta}_1$: è positivo, <U+00E8> indica una relazione positiva fra reddito e valore delle polizze acquistate. Ci dice di quanto aumenta il valore atteso di y se ho un incremento unitario del predittore 1. Quando `INCOME` aumenta di un dollaro, la media di $FACE$ aumenta di 0.44.

La tabella ci consente di valutare agevolmente se i parametri sono *significativamente* differenti da zero. Si guardano ad esempio i $p-$values e si conclude che il parametro è significativamente diverso da 0 se essi sono molto piccoli (sotto 0.01 ad esempio) ovvero è molto bassa la probabilit<U+00E0> di ottenere valori ancora pi<U+00F9> *estremi* del parametro nell'ipotesi che questo coefficiente sia pari a 0. In particolare, $\hat{\beta}_1$ è significativo e conferma che il reddito ha un effetto sull'ammontare delle polizze acquistate e non è plausibile pensare che questo sia dovuto al caso. Si arriva alla medesima conclusione guardando i valori del test F.

Tuttavia si noti che R^2 è quasi 0. Inoltre, si può guardare ai residui (il comando generico `plot()` applicato all'oggetto prodotto da `lm()` da origine a diversi grafici dei residui).

Tali grafici confermano che il modello non è del tutto convincente. I residui non seguono un pattern regolare e non sembrano adattarsi a una gaussiana (primi due grafici). Le altre diagnostiche sui residui relativi all'azione di leva (leverage) e ai residui di Cook confermano la cattiva performance del modello.

Una possibile causa del non soddisfacente adattamento è la presenza di variabili fortemente asimmetriche. In questo caso potrebbe rivelarsi più opportuno utilizzare una trasformazione delle variabili che riduca l'asimmetria.

```
> LFACE <- log(FACE)
> LINCOME <- log(INCOME)
```

Prima di specificare una nuovo modello pu<U+00F2> essere comodo rimpiazzare nella matrice dei dati (data frame), le variabili trasformate.

```
> TL2 <- TL[, -c(5,6)]
> TL2$LINCOME <- with (TL2, LINCOME)
> TL2$LFACE <- with(TL2, LFACE)
```

Ristimiamo ora il seguente modello lineare semplice

$$\log(FACE_i) = \beta_0 + \beta_1 \log(INCOME_i) + \epsilon_i.$$

```
> m1 <- lm(LFACE~LINCOME, data=TL2)

> summary(m1)

Call:
lm(formula = LFACE ~ LINCOME, data = TL2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1967 -0.8032 -0.0018  0.8954  6.4711

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.23003    0.85985   4.920  1.5e-06 ***
LINCOME      0.69604    0.07661   9.086  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.642 on 273 degrees of freedom
Multiple R-squared:  0.2322,        Adjusted R-squared:  0.2294
F-statistic: 82.55 on 1 and 273 DF,  p-value: < 2.2e-16
```

Nell'interpretazione dei coefficienti è necessario ricordare che le variabili sono ora espresse nei logaritmi

- $\hat{\beta}_0$: Si noti che il predittore $log($`INCOME`$)$ <U+00E8> pari a 0 se `INCOME`=1. Tuttavia anche y è su scala logaritmica, e se vogliamo dare un'interpretazione dobbiamo ritrasformare in dollari $\Rightarrow$ quando `INCOME = 1` il valore atteso di `FACE` è $exp(4.23) \sim 69$ dollars. In questo caso l'interpretazione ha tuttavia più senso che nel caso del modello `m0`.
- $\hat{\beta}_1$: come ci si aspettava è positivo. Tuttavia ora quando $log($`INCOME`$)$ aumenta di 1, $\log(FACE)$ aumenta circa di 0.7. Ma se valutiamo cosa accade nelle variabili originarie allora

$$0.7 \simeq log(FACE)_{x+1} - log(FACE)_x = log\left(\frac{FACE_{x+1}}{FACE_x}\right) \Rightarrow \left(\frac{FACE_{x+1}}{FACE_x}\right) \simeq \exp(0.7) \simeq 2.01.$$

Quindi un incrementoi unitario di $log($`INCOME`$)$ corrisponde a raddoppiare l'ammontare di polizza.

Ancora una volta i parametri stimati, e in particolare $\hat{\beta}_1$ sono *significativamente* diversi da zero ($p-$values sempre molto piccoli).

Una previsione dell'ammontare medio del logaritmo della polizza per un reddito di 30000 $ si ottiene come segue

```
> predict(m1, newdata=data.frame(LINCOME=log(30000)))

       1
11.40552
```

e tornando alla scala originale:

```
> exp(11.41)

[1] 90219.42
```

Possiamo sovrapporre la retta con i coefficienti stimati sul grafico come segue:

```
> beta<- coef(m1)
> plot(LINCOME, LFACE)
> abline(beta[1], beta[2])
```

e dare uno sguardo ai residui

```
> plot(m1)
```

I grafici vanno meglio e R^2 <U+00E8> pi<U+00F9> elevato. Ci sono per<U+00F2> margini di miglioramento:

- il normal qq-plot mostra che le code della distribuzione dei residui non sembrano adattarsi alla normale
- R^2 è ancora basso in assoluto
- la linea che interpola lo scatterplot `LFACE` vs `LINCOME` non sembra soddisfacente

Verso un modello di regressione multipla

Il data frame contiene altre variabili di potenziale valore predittivo. Per esempio essere sposati o meno potrebbe avere qualche rilievo. la variabile `MARSTAT` assume solo due valori: `single` e `not single`. Possiamo tracciare lo scatterplot `LFACE` vs `LINCOME` e usare simboli diversi per le due categorie della variabile `MARSTAT`.

```
> plot(LINCOME, LFACE, pch=as.character(MARSTAT), col=as.numeric(MARSTAT))
```

Le due nuvole di punti si sovrappongono ma si vede abbastanza bene che i single tendono a avere polizze meno elevate. Si vede anche che la relazione fra reddito e polizza è meno marcata per le coppie cioè potrebbe esserci un'interazione fra `MARSTAT` e `LINCOME`. Stimiamo il modello

$$\log(FACE_i) = \beta_0 + \beta_1 \log(INCOME)_i + \beta_2 MARSTAT + \beta_3 \log(INCOME)_i \cdot MARSTAT + \epsilon_i.$$

```
> m2 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT, data=TL2)
> summary(m2)

Call:
lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT, data = TL2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2149 -0.8287  0.0696  0.9308  5.6070

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.77902    0.92550   6.244 1.64e-09 ***
LINCOME               0.57288    0.08124   7.051 1.47e-11 ***
MARSTATsingle        -7.29211    2.74216  -2.659   0.0083 **
LINCOME:MARSTATsingle  0.61244    0.25764   2.377   0.0181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.601 on 271 degrees of freedom
Multiple R-squared:  0.2756,        Adjusted R-squared:  0.2676
F-statistic: 34.36 on 3 and 271 DF,  p-value: < 2.2e-16
```

il modello stimato risulta essere:

$$\begin{aligned}
\log(\hat{FACE}_i) &= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_2 MARSTAT + \hat{\beta}_3 \log(INCOME)_i \cdot MARSTAT \\
&= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i \qquad \text{when } i \text{ is not single} \\
&= 5.7790 + 0.5729 \log(INCOME)_i \\
&= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \log(INCOME)_i \qquad \text{when } i \text{ is single} \\
&= (5.7790 - 7.2921) + (0.5729 + 0.6124) \log(INCOME)_i \\
&= -1.5131 + 1.1853 \log(INCOME)_i
\end{aligned}$$

In altre parole, i single comprano meno polizze per i redditi bassi ma se il reddito cresce l'ammontare delle polizze aumenta pi<U+00F9> che negli sposati e quindi per redditi alti i single comprano in media pi<U+00F9> polizze.

è possibile anche verificare l'ipotesi che i due modelli siano differenti

$$H_0 : \beta_2 = \beta 3 = 0 \qquad H1 : \text{ almeno uno fra} \beta_2 \text{ and } \beta_3 \text{ non è} 0$$

```
> anova(m1,m2)

Analysis of Variance Table

Model 1: LFACE ~ LINCOME
Model 2: LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    273 736.27
2    271 694.64  2    41.625 8.1195 0.0003761 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il valore molto piccolo del $p-$value suggerisce di rigettare l'ipotesi nulla che il parametro relativo sia pari a 0, si conclude quindi che i due modelli sono differenti e che il modello più ampio ha condotto a un significativo miglioramento.

Possiamo a questo punto arricchhire il modello con le altre variabili disponibili (`EDUCATION` and `NUMHH` and `AGE`) e verificare se si abbia un ulteriore miglioramento della capacità predittiva.

```
> m3 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT+NUMHH+EDUCATION+AGE, data=TL2)
> summary(m3)

Call:
lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT +
    NUMHH + EDUCATION + AGE, data = TL2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7177 -0.7904  0.1661  0.8790  4.6193

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.091007   1.004241   4.074 6.10e-05 ***
LINCOME               0.405959   0.081313   4.993 1.07e-06 ***
MARSTATsingle        -7.270812   2.588558  -2.809 0.005338 **
NUMHH                 0.253450   0.074292   3.412 0.000746 ***
EDUCATION             0.203700   0.038389   5.306 2.35e-07 ***
AGE                  -0.004639   0.007948  -0.584 0.559958
LINCOME:MARSTATsingle 0.638950   0.244258   2.616 0.009404 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.503 on 268 degrees of freedom
Multiple R-squared:  0.3687,        Adjusted R-squared:  0.3546
F-statistic: 26.09 on 6 and 268 DF,  p-value: < 2.2e-16
```

si noti che l'età non sembra aver un effetto significativo per cui la si pu<U+00F2> escludere dal modello.

```
> m4 <- lm(LFACE~LINCOME+MARSTAT+LINCOME*MARSTAT+NUMHH+EDUCATION, data=TL2)
> summary(m4)

Call:
lm(formula = LFACE ~ LINCOME + MARSTAT + LINCOME * MARSTAT +
    NUMHH + EDUCATION, data = TL2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7693 -0.7619  0.1405  0.9185  4.5721

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.85894    0.92104   4.190 3.79e-05 ***
LINCOME               0.40358    0.08111   4.976 1.16e-06 ***
MARSTATsingle        -7.21946    2.58389  -2.794 0.005580 **
NUMHH                 0.26887    0.06935   3.877 0.000133 ***
EDUCATION             0.20273    0.03831   5.292 2.51e-07 ***
LINCOME:MARSTATsingle 0.63641    0.24392   2.609 0.009587 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.501 on 269 degrees of freedom
Multiple R-squared:  0.3679,        Adjusted R-squared:  0.3562
F-statistic: 31.32 on 5 and 269 DF,  p-value: < 2.2e-16

> par(mfrow=c(2,2))
> plot(m4, c(1,2,4,5))
```

Il comando permette di ottenere ancora i grafici dei residui. Si vede che ora la situazione e molto migliorata. Tuttavia sono presenti alcuni valori anomali (quelli cui sono associati valori del residuo di Cook elevato e con effetto leva elevato). Anche se il modello potrebbe essere suscettibile di ulteriori ritocchi, esso appare ora accettabile.

Il modello finale ristimato è quindi:

$$
\begin{aligned}
\log(\hat{FACE}_i) &= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_2 MARSTAT_i + \hat{\beta}_3 \log(INCOME)_i \cdot MARSTAT_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \\
&= \hat{\beta}_0 + \hat{\beta}_1 \log(INCOME)_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \qquad \text{when } i \text{ is not single} \\
&= 3.86 + 0.40 \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i \\
&= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \log(INCOME)_i + \hat{\beta}_4 EDUCATION_i + \hat{\beta}_5 NUMHH_i \qquad \text{when } i \text{ is single} \\
&= (3.86 - 7.21) + (0.40 + 0.64) \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i \\
&= -3.35 + 1.04 \log(INCOME)_i + 0.20 EDUCATION_i + 0.27 NUMHH_i
\end{aligned}
$$