

## Esempio: i dati *Orange Juice (OJ)*

2024-05-21

Si considerino i dati, raccolti nel 1990, relativi alle vendite di succo d'arancia (OJ) della catena di alimentari Dominick di Chicago (fonte: Taddy M, Business Data science, Mc Graw Hill, 2019). Il dataset include i costi (`prices`) e le vendite (`sales`) settimanali per tre brand di OJ denominati Tropicana, Minute Maid e Dominick's (`brand`), venduti in 83 punti vendita; la variabile `feat` è una variabile *dummy* che presenta il valore 1 se il prodotto è stato pubblicizzato nella settimana in cui sono state registrate le vendite (mediante flyer o in negozio), 0 altrimenti.

```
oj <- read.csv("oj.csv")
head(oj)
```

```
##   sales price   brand feat
## 1  8256  3.87 tropicana  0
## 2  6144  3.87 tropicana  0
## 3  3840  3.87 tropicana  0
## 4  8000  3.87 tropicana  0
## 5  8896  3.87 tropicana  0
## 6  7168  3.87 tropicana  0
```

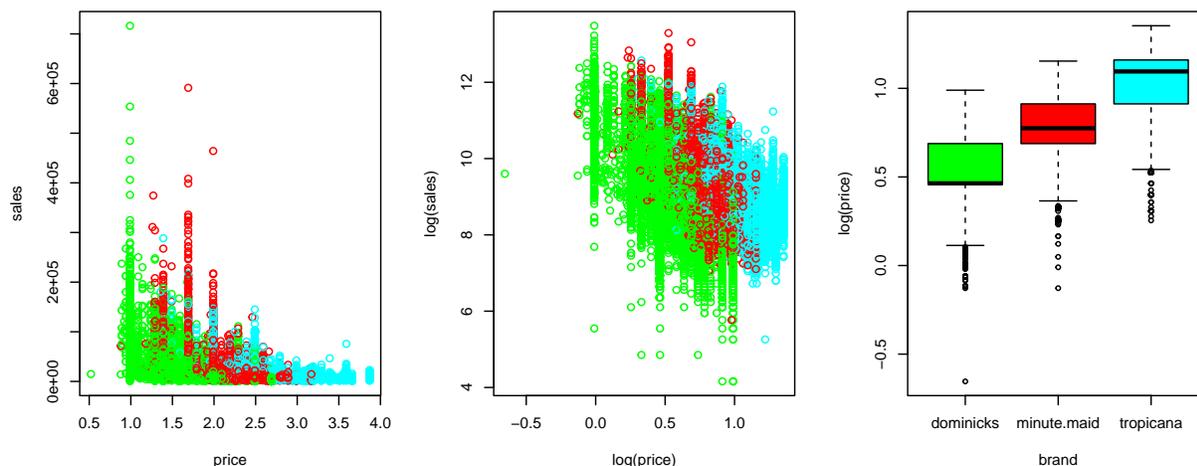
Una volta letti i dati, codifichiamo il `brand` e la variabile `feat` come variabili categoriali:

```
oj$brand<-as.factor(oj$brand)
levels(oj$brand)
```

```
## [1] "dominicks" "minute.maid" "tropicana"
```

```
oj$feat<-as.factor(oj$feat)
```

```
bcol <- c("green","red","cyan")
par(mfrow=c(1,3))
plot(sales ~ price, data=oj, col=bcol[oj$brand])
plot(log(sales) ~ log(price), data=oj, col=bcol[oj$brand])
plot(log(price) ~ brand, data=oj, col=bcol)
```



Osserviamo che il modello lineare può essere adottato anche quando la linearità delle relazioni tra la variabile risposta e le esplicative è presente tra opportune trasformazioni delle stesse.

Ad esempio, se si adotta la trasformazione data dal logaritmo naturale per la variabile risposta il modello diventa *log-lineare*:

$$\log(Y) = \beta_1 + \beta_2 X + \varepsilon$$

In tal caso l'interpretazione della stima  $\hat{\beta}_2$  è che un incremento unitario di  $X$  produce in media una variazione di  $\hat{\beta}_2$  unità su  $\log(Y)$ .

In termini di  $Y$  (la variabile non trasformata), si ha un effetto *moltiplicativo*: ogni incremento di una unità di  $X$  porta ad un valore medio della variabile risposta pari a  $ye^{\beta_2}$ . Infatti se  $x$  diventa  $x + 1$ , allora  $Y' = e^{\beta_1 + \beta_2(x+1)} = e^{\beta_1 + \beta_2 x} e^{\beta_2} = Y e^{\beta_2}$ .

In questo caso, grafici mostrano che sia ragionevole considerare una trasformazione logaritmica per entrambe le variabili continue **price** e **sales**.

Si considera quindi un modello *log-log* della forma

$$\log(Y) = \beta_1 + \beta_2 \log(X) + \varepsilon$$

In questo caso l'interpretazione della stima  $\hat{\beta}_2$  è quella della variazione percentuale su  $Y$  per un incremento dell'1% di  $X$  (il valore atteso del rapporto  $y'/y$  per un incremento di  $x$  dell'1% è  $(1.01)^{\beta_2} \approx 1 + 0.01\beta_2$ ).

Stimiamo quindi innanzitutto un modello di regressione lineare semplice utilizzando  $\log(\text{sales})$  come variabile risposta e  $\log(\text{price})$  come esplicative:

$$\log(\text{sales}) = \beta_1 + \beta_2 \log(\text{price}) + \varepsilon$$

```
m1 <- lm(log(sales) ~ log(price), data=oj)
summary(m1) # coef, tests, fit

##
## Call:
## lm(formula = log(sales) ~ log(price), data = oj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -5.0441 -0.5853 -0.0330 0.5756 3.7264
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.42342  0.01535  679.04  <2e-16 ***
## log(price)  -1.60131  0.01836  -87.22  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9071 on 28945 degrees of freedom
## Multiple R-squared:  0.2081, Adjusted R-squared:  0.2081
## F-statistic: 7608 on 1 and 28945 DF, p-value: < 2.2e-16
```

Le stime dei coefficienti sono  $\hat{\beta}_1 \approx 10.42$  e  $\hat{\beta}_2 \approx -1.60$

```
coef(m1)
```

```
## (Intercept) log(price)
## 10.423422 -1.601307
```

e indicano che, indipendentemente dal brand,  $\log(\text{sales})$  diminuisce all'aumentare di  $\log(\text{prices})$ . In particolare, un incremento dell'1% di `prices` porta a moltiplicare  $y = \text{sales}$  di un fattore pari a  $e^{\hat{\beta}_2 \log(1.01)} = e^{-1.6 \log(1.01)} = 0.9842$ , da cui si deduce che ad un aumento dell'1% di `prices` corrisponde ad una diminuzione delle vendite pari all'1.58%.

Per introdurre l'effetto del brand sull'intercetta del modello si può stimare un modello che include `brand` come variabile esplicativa. Ciò equivale a ritenere che, a parità di prezzo, il valore atteso delle vendite sarà diverso a seconda del brand:

```
m2 <- lm(log(sales) ~ log(price) + brand, data=oj)
summary(m2)
```

```
##
## Call:
## lm(formula = log(sales) ~ log(price) + brand, data = oj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.3152 -0.5246 -0.0502  0.4929  3.5088
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.82882  0.01453  745.04  <2e-16 ***
## log(price)  -3.13869  0.02293 -136.89  <2e-16 ***
## brandminute.maid 0.87017  0.01293  67.32  <2e-16 ***
## brandtropicana  1.52994  0.01631  93.81  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7935 on 28943 degrees of freedom
## Multiple R-squared:  0.3941, Adjusted R-squared:  0.394
## F-statistic: 6275 on 3 and 28943 DF, p-value: < 2.2e-16
```

Dal modello si ottiene  $\hat{\beta}_2 = -3.14$ , pertanto le vendite calano di circa il 3% per ogni incremento dell'1% del prezzo. In secondo luogo, si osservi che il modello restituisce la stima di due coefficienti per i brand `Minute Maid` e `Tropicana`, ma non per il terzo brand. Ciò è dovuto al fatto che, essendo `brand` una variabile con tre modalità, vengono create due variabili indicatrici  $z_i$  (`brandminute.maid`) e  $w_i$  (`brandtropicana`) tali che,

per  $i = 1, \dots, n$ ,

$$z_i = \begin{cases} 1, & \text{brand} = \text{Minute Maid} \\ 0, & \text{altrimenti} \end{cases}$$

e

$$w_i = \begin{cases} 1, & \text{brand} = \text{Tropicana} \\ 0, & \text{altrimenti} \end{cases}$$

Per tre modalità, sono sufficienti due variabili indicatrici. Infatti, quando  $z_i = w_i = 0$ , allora  $\text{brand} = \text{Dominick's}$ .

Quanto detto può essere facilmente verificato ottenendo la matrice ( $n \times p$ ) di regressione o *design matrix* del modello,  $\mathbf{X}$ , le cui colonne sono un vettore di dimensione  $n$  con tutti elementi pari a 1 e i vettori delle  $p - 1$  esplicative incluse nel modello (nella notazione matriciale,  $y = \mathbf{X}\beta + \varepsilon$ ).

```
x <- model.matrix( ~ log(price) + brand, data=oj)
head(x)
```

```
## (Intercept) log(price) brandminute.maid brandtropicana
## 1          1  1.353255                0                1
## 2          1  1.353255                0                1
## 3          1  1.353255                0                1
## 4          1  1.353255                0                1
## 5          1  1.353255                0                1
## 6          1  1.353255                0                1
```

Il modello ha quindi la forma

$$\log(Y_i) = \beta_1 + \beta_2 \log(x_i) + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$$

I coefficienti che vengono riportati nella sintesi del modello relativi alla variabile `brand` sono quindi da interpretare in relazione alla categoria di riferimento, qui data dal brand `Dominick's`, che appare per primo digitando il comando

```
levels(oj$brand)
```

```
## [1] "dominicks" "minute.maid" "tropicana"
```

Dalle stime sotto riportate, si deduce che, a parità di prezzo, il brand `Tropicana` realizza vendite maggiori di `Minute Maid`, che a sua volta ha in media vendite superiori al brand `Dominick's`. Si ricordi che il coefficiente di una variabile dummy rappresenta la differenza tra il valore medio della variabile risposta quando la variabile dummy vale 1 e quando vale 0.

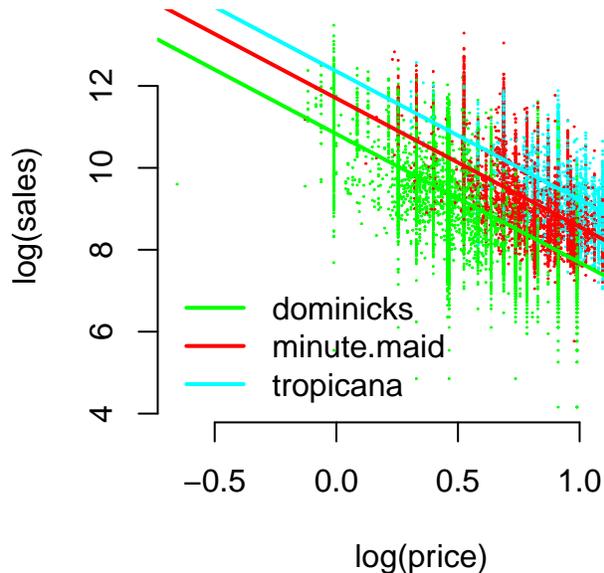
```
coef(m2)
```

```
## (Intercept)          log(price) brandminute.maid brandtropicana
## 10.8288216      -3.1386914          0.8701747          1.5299428
```

Per il brand `Dominick's` si ha  $E(\log(Y)|x) = \hat{\beta}_1 + \hat{\beta}_2 \log(x)$  e per il brand `Minute Maid` si ha  $E(\log(Y)|x) = \hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_2 \log(x)$ ; si evince quindi che  $\hat{\beta}_3$  è la differenza tra la stima della media di  $\log(\text{sales})$  per `minute.maid` e quella per `dominick's`:

```
betav <- coef(m2)
plot(log(sales) ~ log(price), data=oj, col=bcol[oj$brand],
     cex=.1, pch=20, bty="n")
abline(a=betav[1], b=betav[2], col=bcol[1], lwd=2)
```

```
abline(a=betav[1]+betav[3], b=betav[2], col=bcol[2], lwd=2)
abline(a=betav[1]+betav[4], b=betav[2], col=bcol[3], lwd=2)
legend("bottomleft", bty="n", lwd=2, col=bcol, legend=levels(oj$brand))
```



```
newdata<-data.frame(price=rep(4,3),
                    brand=factor(c("tropicana","minute.maid","dominicks"),
                                levels=levels(oj$brand)))
predict(m2, newdata=newdata) ## predicted log units moved
```

```
##      1      2      3
## 8.007614 7.347846 6.477671
```

```
exp(predict(m2, newdata=newdata)) ## predicted # of units moved
```

```
##      1      2      3
## 3003.7420 1552.8481 650.4545
```

Sebbene il modello  $m_2$  consenta di ottenere una intercetta diversa per ogni brand, esso assume che la pendenza della retta sia la stessa nei tre casi; un modello più realistico potrebbe essere ottenuto includendo l'interazione tra  $\log(\text{price})$  e  $\text{brand}$ , che consente di verificare se il brand ha effetto sulla relazione tra il prezzo e le vendite:

$$\log(Y_i) = \beta_1 + \beta_2 x'_i + \beta_3 z_i + \beta_4 w_i + \beta_5 x'_i z_i + \beta_6 x'_i w_i + \varepsilon_i$$

dove  $x'_i = \log(x_i)$ .

```
## Modello con Interazioni
m3 <- lm(log(sales) ~ log(price)*brand, data=oj)
summary(m3)
```

```
##
```

```

## Call:
## lm(formula = log(sales) ~ log(price) * brand, data = oj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.4434 -0.5232 -0.0494  0.4884  3.4901
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.95468    0.02070  529.136 <2e-16 ***
## log(price)     -3.37753    0.03619  -93.322 <2e-16 ***
## brandminute.maid  0.88825    0.04155   21.376 <2e-16 ***
## brandtropicana  0.96239    0.04645   20.719 <2e-16 ***
## log(price):brandminute.maid  0.05679    0.05729    0.991  0.322
## log(price):brandtropicana  0.66576    0.05352   12.439 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7911 on 28941 degrees of freedom
## Multiple R-squared:  0.3978, Adjusted R-squared:  0.3977
## F-statistic: 3823 on 5 and 28941 DF, p-value: < 2.2e-16

```

Nel modello  $m_3$ , per ogni brand si ottiene una retta con intercetta e coefficiente angolare differente.

Come prima, la modalità di riferimento per il brand è Dominick's: per ogni incremento dell'1% del prezzo, la diminuzione percentuale attesa delle vendite del brand Tropicana è

$$\log(\text{price}) + \log(\text{price}) : \text{brandtropicana} = -3.37753 + 0.66576 = -2.71$$

ed è minore rispetto a quella di Minute maid, pari a

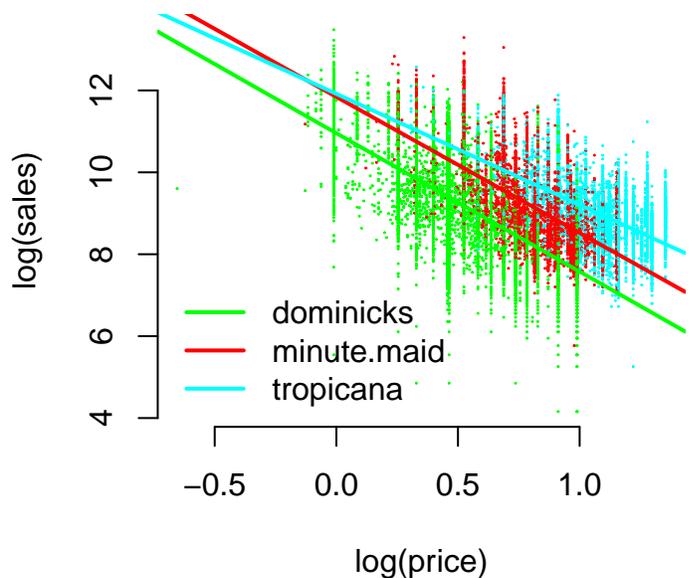
$$\log(\text{price}) + \log(\text{price}) : \text{brandminute.maid} = -3.37753 + 0.05679 = -3.32$$

e di Dominick's, pari a  $\log(\text{price}) = -3.38$ .

```

betav<-coef(m3)
plot(log(sales) ~ log(price), data=oj, col=bcol[oj$brand], cex=.1, pch=20, bty="n")
abline(a=betav[1], b=betav[2], col=bcol[1], lwd=2)
abline(a=betav[1]+betav[3], b=betav[2]+betav[5], col=bcol[2], lwd=2)
abline(a=betav[1]+betav[4], b=betav[2]+betav[6], col=bcol[3], lwd=2)
legend("bottomleft", bty="n", lwd=2, col=bcol, legend=levels(oj$brand))

```



Infine, è ragionevole supporre che la pubblicità di un prodotto influenzi le vendite, per cui stimiamo un modello che tenga conto della variabile `feat`:

$$\log(Y_i) = \beta_1 + \beta_2 x'_i + \beta_3 z_i + \beta_4 w_i + \beta_5 x'_i z_i + \beta_6 x'_i w_i + \beta_7 \delta_i + \varepsilon_i$$

dove  $x'_i = \log(x_i)$  e  $\delta_i = 1$  se `feat=1` e 0 altrimenti.

```
levels(oj$feat)<-c("no", "si")
m4 <- lm(log(sales) ~ log(price)*brand + feat, data=oj)
summary(m4)
```

```
##
## Call:
## lm(formula = log(sales) ~ log(price) * brand + feat, data = oj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.9524 -0.4408 -0.0036  0.4231  3.2363
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.46852    0.01932  541.843 < 2e-16 ***
## log(price)     -2.89261    0.03279  -88.227 < 2e-16 ***
## brandminute.maid  0.51271    0.03734   13.732 < 2e-16 ***
## brandtropicana  0.62865    0.04163   15.100 < 2e-16 ***
## featsi         0.89728    0.01043   86.029 < 2e-16 ***
## log(price):brandminute.maid  0.33426    0.05123    6.525 6.92e-11 ***
## log(price):brandtropicana  0.82914    0.04780   17.346 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  
## Residual standard error: 0.706 on 28940 degrees of freedom  
## Multiple R-squared:  0.5204, Adjusted R-squared:  0.5203  
## F-statistic: 5234 on 6 and 28940 DF,  p-value: < 2.2e-16
```