Matematičke metode u marketingu
Linearna regresija

Lavoslav Čaklović
PMF-MO

2016

LM

$y, x_1, \ldots, x_p$ slučajne varijable[1]. Ispitajmo model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon.$$

Želimo procijeniti koeficijente β_i. U tu svrhu mjerimo vrijednosti tih varijabli na uzorku subjekata $S = \{s_1, \ldots, s_m\}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \ i = 1, \ldots, m \qquad (1)$$

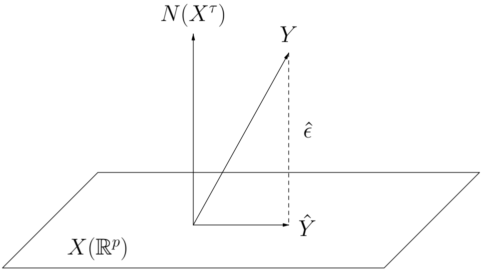
Uvodimo oznake: $Y = [y_i]^\tau \in \mathbb{R}^{m\times 1}$, $\epsilon = [\epsilon_i]^\tau \in \mathbb{R}^{m\times 1}$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mp} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

[1] y – odzivna varijabla (*response*), x_i – explanatorne varijable

Jednadžbe (1) pišemo u matričnoj formi

$$Y = X\beta + \epsilon \quad (X \text{ — dizajn matrica}).$$



$R(X)$ — prostor stupaca od X
$N(X^\tau)$ – jezgra od X^τ

$$R(X) \oplus N(X^\tau) = \mathbb{R}^m$$
$$X\beta \oplus \epsilon = Y.$$

Najbolja procjena Y pomoću stupaca matrice X u smislu najmanjih kvadrata dobije se za $\hat{\epsilon} \in N(X^\tau)$, tj. kad je $X^\tau \hat{\epsilon} = 0$. U tom slučaju

$$\hat{Y} := X\hat{\beta} \text{ najbolja procjena od } Y,$$

$$\hat{\beta} = (X^\tau X)^{-1} X^\tau Y \text{ (stupci od } X \text{ nezavisni) i vrijedi}$$

$$\hat{Y} = X\hat{\beta} = X(X^\tau X)^{-1} X^\tau Y =: HY \text{ (} H \text{ – hat projektor).}$$

Teorem (Gauss-Markov)

Neka je X realna matrica punog ranga po stupcima i

$$Y = X\beta + \epsilon$$

gdje je očekivanje $E(\epsilon) = 0$ i $Cov(\epsilon) = \sigma^2 I_{m\times m}$. Tada $\hat{\beta} = (X^\tau X)^{-1} Y$ ima majmanju kovarijancu[2] *u klasi svih linearnih nepristranih procjena od Y.*

Što raditi kad su narušeni uvjeti teorema:

– Heteroscedastičnost: Uvode se težine.

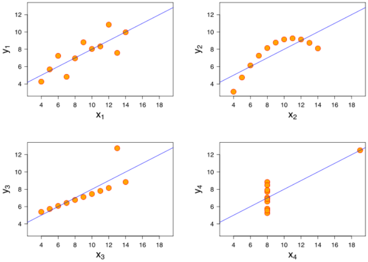
– Ako ϵ ima specifičnu distribuciju: MLE, Bayesov model

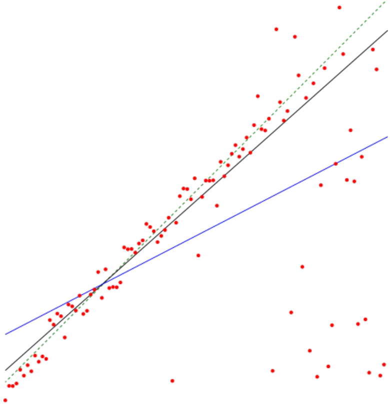
– Ako su stupci od X visoko korelirani: pristrane procjene, miješani modeli.

– Još...: vidi wiki

[2] u odnosu na konus pozitivno semdefinitnih matrica.

Neki primjeri.





Slika: Linearni (lijevo), Thiel-Sen (desno, prisutnost outliersa)

Procjene parametara ako $\epsilon \sim N(0, \sigma^2 * I)$

Očekivanje i varijanca procjene od β.

$$E(\hat{\beta}) = E((X^\tau X)^{-1} X^\tau \underbrace{Y}_{X\beta+\epsilon}) = (X^\tau X)^{-1} X^\tau E(X\beta) = \beta$$

$$Covar(\hat{\beta}) = (X^\tau X)^{-1} X^\tau * \sigma^2 I * X(X^\tau X)^{-1} = \sigma^2 (X^\tau X)^{-1}$$

spec. $\forall i$, $Var(\hat{\beta}_i) = \sigma^2 (X^\tau X)^{-1}_{ii}$.

Procjena σ^2. Zbog $\hat{\epsilon} = (Y - \hat{Y}) = (I - H)Y$ (H je ort. projektor)

$$\hat{\epsilon}^\tau \hat{\epsilon} = Y^\tau (I-H)^\tau (I-H) Y = Y^\tau (I-H) Y$$
$$Var(\hat{\epsilon}) = E(\hat{\epsilon}^\tau \hat{\epsilon}) = E(Y^\tau (I-H) Y) = (m-p)\sigma^2$$

pa je nepristrana procjena od σ^2 dana s

$$\hat{\sigma}^2 = (\hat{\epsilon}^\tau \hat{\epsilon})/(m-p).$$

Kvaliteta modela (GOF[3])

- $TSS = \sum_{i=1}^{m}(y_i - \bar{y})^2$ (T=Total, SS=square sum)
- $RSS = \|\hat{\epsilon}\|^2 = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ (R=Residual)
- $ESS = \sum_{i=1}^{m} (\hat{y}_i - \bar{y})^2$ (E=Explained)

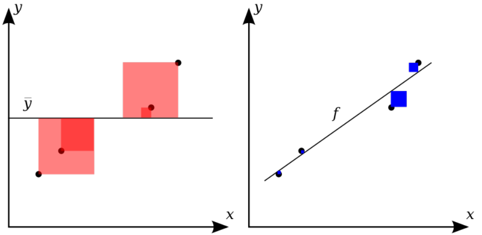
Koeficijent determinacije (bezdimenzionalan)

$$R^2 := 1 - \frac{RSS}{TSS}$$

Alternativno:

$$\hat{\sigma}^2 = (\hat{\epsilon}^T \hat{\epsilon})/(m - p).$$

U dimenzijama odzivne varijable (interpretacija).



Slika. TSS (lijevo) — RSS (desno)

[3] Goodness of fit.

Teorem

$$TSS = ESS + RSS \iff \sum_i y_i = \sum_i \hat{y}_i. \qquad (2)$$

Napomena. Desna strana u (2) vrijedi ako je $\hat{Y}$ dobiveno linearnom regresijom. Zaista: $0 = X^\tau \hat{\epsilon} = X^\tau (Y - \hat{Y}) \implies \sum_i y_i = \sum_i \hat{y}_i$ jer prvi stupac od X sadrži samo jedinice.

Dokaz teorema.

$$\begin{aligned} TSS &= \|Y - \bar{Y}\|^2 = \|Y - \hat{Y} + \hat{Y} - \bar{Y}\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2 + 2(Y - \hat{Y})^\tau (\hat{Y} - \bar{Y}) \\ &= RSS + ESS + 2(Y - \hat{Y})^\tau \hat{Y} - 2(Y - \hat{Y})^\tau Y \\ &= RSS + ESS - 2(Y - \hat{Y})^\tau \bar{Y}, \end{aligned}$$

gdje je $\bar{Y} = [\bar{y}\ \bar{y} \ldots \bar{y}]^\tau$ konstantan vektor. □

Alternativna definicija R^2

Zbog teorema, u slučaju linearne regresije

$$R^2 := \frac{ESS}{TSS}.$$

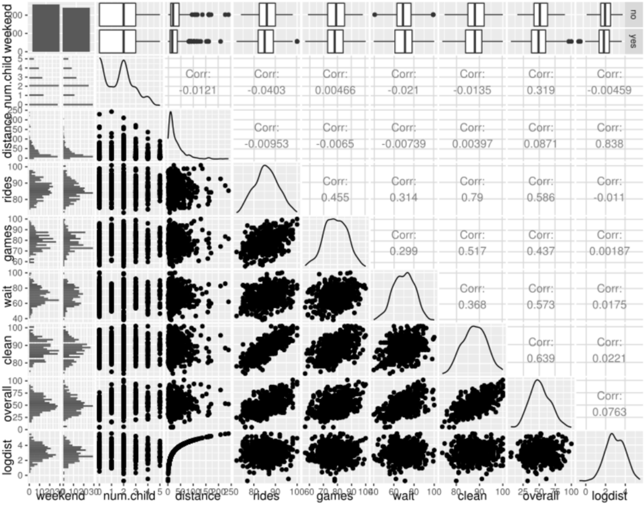
R^2 može poprimiti vrijednost izvan intervala $[0, 1]$ ako se ne radi o linearnom modelu i ovisno o tome koja se definicija prihvati.

Korigirani R^2.
– R^2 ne ukazuje na pristranost u procjeni koeficijenata i predikcije.
– Dodavanje prediktora povećava R^2 (prostor $R(X)$ raste).
– Previše prediktora u modelira šum (overfitting) — loša predikcija.

Korigirani R^2 dizajniran je tako da daje **nepristranu procjenu** populacijskog R^2 i služi za međusobno uspoređivanje modela.

'Ručno' modeliranje uz pomoć R-a

```
mfit <- lm(overall ~ rides + games + wait + clean, data)
# Kreiranje matrice X i odzivne varijable Y
X <- cbind(1,data[, c("rides","games", "wait", "clean")])
Y <- data$overall
# inverz od X^T*X
X <- as.matrix(X); xtxi <- solve(t(X) %*% X)
# računanje procjene parametara \hat{\beta}
xtxi %*% t(X) %*% Y
# procjena \sigma^2: ||reziduali||^2/(m-p)
m <- nrow(X); p <- ncol(X)
sigma.hat <- sqrt(sum(mfit$res^2)/(m-p))
# standardne greške koeficijenata
sigma.hat * sqrt(diag(xtxi))
# procjena R^2
1 - sum(mfit$res^2)/sum((Y - mean(Y)^2)
```



Slika. Početni korak analize primjera *zabavni park* s predavanja. Korelacija parova varijabli.

- Varijabla *distance* odstupa od normalnosti. Zamjenjujemo ju s *logdist*.
- Varijabla *num.child* ima 2 maksimuma. Možda uvesti novu dihotomnu varijablu *has child*.
- Var. *rides*-*logdist* imaju međusobnu zavisnost ok (elipse).
- Linearni model bi trebao biti zadovoljavajući.

Output linearne regresije podataka *zabavni park* s jednom interakcijom `wait:has.child`. Za varijablu *logdist* treba ispitati interval pouzdanosti, a *games* je na granici odbacivanja (95%). R^2 je visok.

```
Coefficients:       Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.69316    0.03684 -18.814  < 2e-16
rides                0.21264    0.03313   6.419 3.24e-10
games                0.04870    0.02394   2.034   0.0425
wait                 0.15095    0.03688   4.093 4.98e-05
clean                0.30244    0.03485   8.678  < 2e-16
logdist              0.02919    0.02027   1.440   0.1504
has.childTRUE        0.99830    0.04416  22.606  < 2e-16
wait:has.childTRUE   0.34688    0.04380   7.920 1.59e-14

Residual standard error: 0.4508 on 492 degrees of freedom
Multiple R-squared:  0.7996,Adjusted R-squared:  0.7968
```

Za daljnje učenje?

Pouzdanost:	Odrediti intervale pouzdanosti za procjenjene parametre i predikciju.
Imputacija:	Rekonstrukcija nepostojećih vrijednosti.
Interpretacija β_i:	Mjerena greška često ima uzrok u parametrima koje ne mjerimo.
Zavisnost	prediktora (uvođenje težina), korelirane greške.
Dijagnostika:	outliersi, leverage
Nelinearnost:	normaliziranje podataka, reg. splineovi

Literatura

Julian J. Faraway, Practical Regression and Anova using R, 2002
http://blog.minitab.com/blog/adventures-in-statistics/