Statistika
(za biologe, 5.dio)

Bojan Basrak

rujan 2006

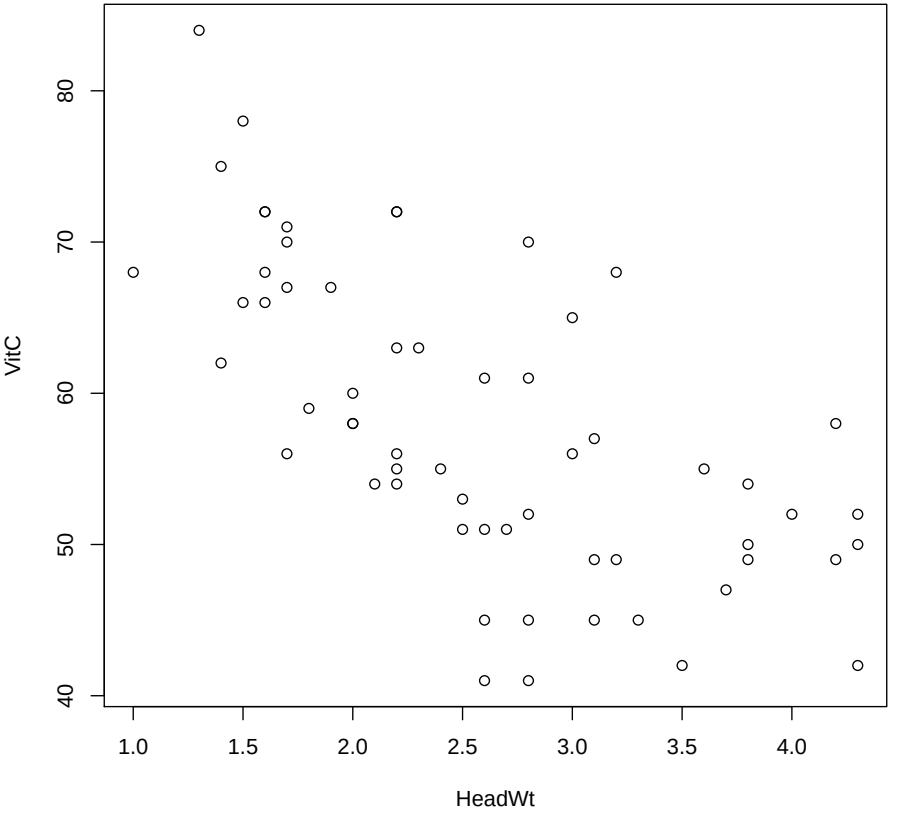
Jednostavna linearna regresija

Pretpostavljamo da za svaku od n jedinki u uzorku imamo dva numerička podatka. Tada se podaci daju reprezentirati kao niz parova numeričkih varijabli

$$(x_i, y_i),\ i = 1, \ldots, n.$$

Primjer

U jednoj studiji prikupljeni su podaci o biljkama kupusa. Za svaku od 60 biljaka izmjerena je veličina glavice u kilogramima (x_i) i sadržaj vitamina C (u nepoznatim jedinicama) (y_i).



Scatterplot jasno ukazuje da dvije varijable nisu nezavisne.

Sjetite se definicije koeficijenta korelacije za dvije sl. varijable X i Y kao

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-EX)(Y-EY)]}{\sigma_X \sigma_Y}.$$

Ako $\operatorname{cov}(X,Y) = 0$, slijedi $\operatorname{corr}(X,Y) = 0$, u tom slučaju kažemo da su X i Y nekorelirane. Npr. nezavisne sl. varijable su uvijek nekorelirane, no obrnuto ne vrijedi.

Broj $\operatorname{corr}(X,Y)$ je uvijek izmedju -1 i 1, što je bliži granicama tog intervala linearna veza izmedju sl. varijabli X i Y je jača.

Kako mi imamo samo podatke i ne znamo pravu (teorijsku) razdiobu para sl. varijabli (X, Y), mi možemo izračunati samo procjenu za koeficijent korelacije. Nju zovemo **korelacija sl. uzorka** i računamo kao

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y},$$

uz uobičajene oznake. Tako je npr.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

varijanca uzorka $x_1, \ldots, x_n$.

Kod kupusa su npr

$$\bar{x} = 2.593,\ \bar{y} = 57.95,\ s_X = 0.8823825,\ s_Y = 10.11866$$

a korelacija uzorka iznosi

$$r = -0.659892.$$

U modeliranju zavisnosti izmedju varijabli najjednostavnije je prepostaviti da postoji linearna veza izmedju njih, tj.

$$y_i = a + bx_i \, .$$

Takvu vezu zovemo determinističkom, a svi bi podaci u tom slučaju ležali točno na pravcu koji opisuje gornja jednadžba. Zbog raznih izvora varijabilnosti kod stvarno prikupljneih podataka, prihvatljivije je uvesti vjerojatnosni model za podatke

$$Y_i = a + bx_i + \varepsilon_i \, ,$$

gdje ε_i predstavlja slučajnu pogrešku.

Pitanje je kako bismo našli optimalne koeficijente pravca kroz točke podataka, recimo $\hat{a}$ i $\hat{b}$. Najčešće korištena metoda minimizira tzv. sumu kvadrata ostataka (eng. residual sum of squares–RSS ili sum of squares for error–SSE)

$$SSE = \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b}x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Ova metoda za pronalaženje optimalnih koeficijenata $\hat{a}$ i $\hat{b}$ naziva se **metoda najmanjih kvadrata**.

Napomena Termina regresija uveo je F. Galton, a R. Bošković je predložio minimizirati sumu apsolutnih pogrešaka $\sum |y_i - \hat{y}_i|$, no tada optimalne a i b nije lako naći.

Ispostavlja se da je optimalni koeficijent $\hat{b}$ zapravo

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx r\frac{s_Y}{s_X}$$

dok je

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

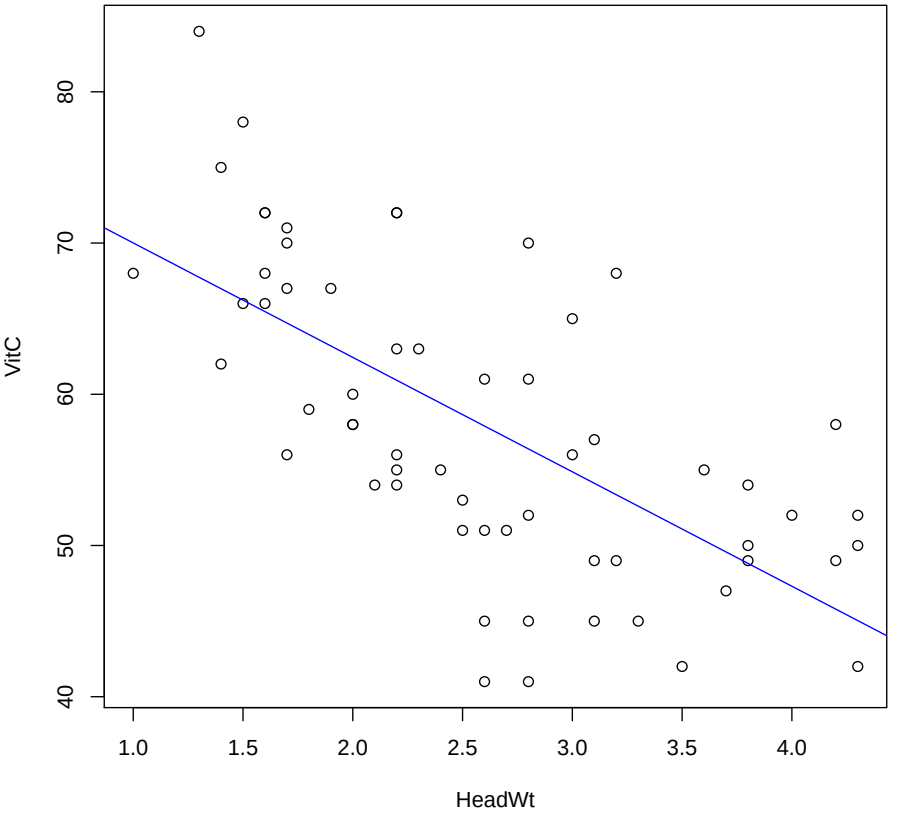
Veličine

$$\hat{y}_i = \hat{a} + \hat{b}x_i,\ i = 1, \ldots, n$$

zovemo procjenama za vrijednosti y_i, a

$$y_i - \hat{y}_i,\ i = 1, \ldots, n,$$

nazivamo ostacima.



Procjenjeni koeficijenti za podatke o kupusu su $\hat{a} = 77.57$ *i* $\hat{b} = -7.57$.

Naravno nas interesira i neizvjesnost u procjeni ovih parametara ili čak statistički test. Ako želimo npr. napraviti interval pouzdanosti $(1-\alpha)100\%$ za parametar b to možemo učiniti ako vrijede dodatni (Gauss-Markovljevi) uvjeti na slučajne greške ε_i

- i) $E(\varepsilon_i) = 0$ za sve i
- ii) $\mathrm{var}(\varepsilon_i) = \sigma^2 < \infty$ za sve i
- iii) $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ za sve $i \neq j$

Pod ovim prepostavkama $(1-\alpha)100\%$-tni interval pouzdanosti za parametar b je

$$\left(\hat{b} - t_{\alpha/2}(n-2)\frac{s}{S_X}, \hat{b} + t_{\alpha/2}(n-2)\frac{s}{S_X}\right),$$

gdje su

$$s = \frac{SSE}{n-2}$$

i

$$S_X^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)s_X^2.$$

Ako želimo testirati sljedeće dvije hipoteze

$$H_0 : b = 0 \quad \text{nasuprot} \quad H_A : b \neq 0,$$

uz nivo značajnosti $\alpha > 0$ testna statistika je

$$T = \frac{\hat{b} S_X}{s}$$

a kritično područje je

$$C_\alpha = \mathbb{R} \setminus (-t_{\alpha/2}(n-2), t_{\alpha/2}(n-2)).$$

Dakle, koristimo Studentovu t-razdiobu s $n-2$ stupnja slobode.

Ako nam je potreban i $(1-\alpha)100\%$-tni interval pouzdanosti za vrijednost Y kada znamo vrijednost kontrolirane varijable $x = x_0$, on se može dobiti kao

$$\hat{a} + \hat{b}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_X^2}}.$$

Ponekad se u softverskim paketima regresijska analiza provodi koristeći tzv. ANOVA (analysis of variance) tablice. U njima se varijabilnost podataka y_i razlaže na razne komponente.

ukupna varijabilnost	$S_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$
varijabilnost zbog regresije	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
varijabilnost zbog greške	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

U primjeru s vrstama kupusa za test $H_0 : b = 0$ dobijemo testnu statistiku $T = -6.69$, a broj stupnjeva slobode je 58. Tako da je $T \in C_{0.05}$, a i p vrijednost zanemarivo mala. Dakle odbacili bismo nul-hipotezu na nivou $\alpha = 0.05$.

Napomena Ako imamo više od dvije numeričke varijable tada možemo napraviti regresiju jedne od njih na sve ostale, takav model se naziva višestruka regresija (multiple regression).

Jednofaktorska analiza varijance

Ponekad o svakoj jedinki u uzorku, osim jednog numeričkog opažanja, imamo i jedno kvalitativno koje indicira pripadnost jedinke nekoj od kategorija ili vrsti tretmana pod kojim je izvršeno opažanje. Prva je dakle varijabla numerička, a drugu u ovom kontekstu nazivamo **faktorom**.

Analiza ovakvih podataka ovisi o samom dizajnu pokusa. Dizajn eksperimenta je izuzetno važan korak svakog istraživanja. Istraživač se mora odlučiti koje faktore želi proučavati, ali i kako će dodjeljivati različite pokusne jedinke različitim fakorima. Vrijednosti faktora su dakle razne vrste tretmana kojima podvrgavamo jedinke u pokusu, a ono što želimo tipično utvrditi je: postoji li zavisnost izmedju vrste tretmana i numeričke varijable.

Umjesto da usporedjujemo različite tretmane u parovima (npr. t-testovima) očito je efikasnije usporediti ih sve odjednom. Ako uzorke skupljamo u svakoj od k kategorija ili populacija (tj. za svaki tretman) na slučajan i nezavisan način govorimo o **potpuno randomiziranom dizajnu**.

Sve podatke sada dijelimo na k uzoraka ovisno o nivou faktora tj. vrsti tretmana

$$
\begin{aligned}
&x_{1,1}, x_{1,2}, \ldots, x_{1,n_1},\\
&x_{2,1}, x_{2,2}, \ldots, x_{2,n_2},\\
&\qquad\qquad \vdots\\
&x_{k,1}, x_{k,2}, \ldots, x_{1,n_k},
\end{aligned}
$$

gdje je $n = n_1 + n_2 + \cdots + n_k$ ukupan broj jedinki u pokusu.

Za svaki od k uzoraka možemo izračunati aritmetičku sredinu

$$\bar{x}_i = \frac{1}{n_i}(x_{i,1} + x_{i,2} + \cdots + x_{i,n_i}) = \frac{\sum_{j=1}^{n_i} x_{i,j}}{n_i}$$

odn. varijancu

$$s_i^2 = \frac{1}{n-1}(x_{i,j} - \bar{x}_i)^2,$$

$i = 1, \ldots, k$.

Ukupna aritmetička sredina podataka je

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j}.$$

Ako postoji zavisnost izmedju faktora i numeričke varijable za očekivati je da se aritmetičke sredine raznih uzoraka bitno razlikuju, no nije apriori jasno kako odrediti ono što bismo zvali bitnim razlikama.

U tu svrhu provodimo **analizu varijance** (analysis of variance–ANOVA) razlažući varijabilnost u podacima. Definiramo sumu kvadrata u odn. na tretman (sum of squares for treatments) kao

$$SST = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2,$$

te sumu kvadrata pogrešaka (sum of squares for error)

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2.$$

Na osnovu ovih veličina definiramo i srednjekvadratna odstupanja:

- MST (mean square for tretments)

$$MST = \frac{SST}{k-1}$$

- MSE (mean square for error)

$$MSE = \frac{SSE}{n-k}$$

Ako se značajan dio varijabilnosti da objasniti različitim tretmanima tada očekujemo da je MST značajno veća od MSE.

Vjerojatnosni model za podatke možemo napisati na sljedeći način

$$X_{i,j} = \mu + \beta_i + \varepsilon_{i,j},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

gdje je μ ukupno očekivanje za numeričku varijablu, β_i predstavlja efekt i-tog tretmana, a $\varepsilon_{i,j}$ su nezavisne greške za koje pretpostavljamo da sve imaju $N(0, \sigma^2)$ razdiobu.

Ako želimo utvrditi imaju li tretmani efekta na numeričku varijablu možemo testirati

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

nasuprot alternative da je bar neka od β_i različita od 0.

Uz naše prepostavke testna statistika

$$F = \frac{MST}{MSE}$$

pod hipotezom H_0 ima Fisherovu razdiobu s $(k-1, n-k)$ stupnjeva slobode ili kraće $F(k-1, n-k)$ razdiobu.

Na nivou značajnosti $\alpha > 0$ odbacujemo nul-hipotezu ako je

$$F \geq F_\alpha(k-1, n-k),$$

tj. ako je F veća od $(1-\alpha)$-kvantila odgovarajuće F razdiobe.

Cijeli račun se tipično izvodi koristeći tzv. ANOVA tablice koje imaju sljedeći oblik

izvor varijabilnosti	st. slob.	suma kvad.	sredjekv. ods.	F stat.
tretman	$k-1$	SST	MST	MST/MSE
greška	$n-k$	SSE	MSE	
ukupno	$n-1$	SS	s_X^2	

Ovdje je SS jednostavna suma kvadrata (sum of squares) tj.

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2 = (n-1)s_X^2.$$

Uočite da vrijedi

$$SS = SST + SSE.$$

Postoje i mnoge druge ANOVA verzije. No postavlja se i pitanje ako je metoda izvedena pod prepostavkom da su greške normalno distribuirane koliko je ona robusna na odstupanja od ove prepostavke, Ispostavlja se da je ANOVA nažalost osjetljiva na slučaj nejednakih varijanci, no taj je efekt mali u slučaju jednako velikih uzoraka za razne faktore (p. 574. McClave et al).