Mašinsko učenje

Regresija

Nemanja Maček

Sadržaj

Uvodne napomene

Šta je regresija?

- U velikom broju eksperimenata uočava se veza između dve ili više promenljivih.
- Na primer,
  - može se uočiti veza između stepena erozije i količine padavina,
  - može se uočiti veza između širine reke i maksimalnog godišnjeg proticaja,
  - može se uočiti veza između broja rođene dece po ženi, nivoa obrazovanja i vrste zaposlenja, itd.

Uvodne napomene

Šta je regresija?

- Ako posmatramo obeležja $X_1, X_2, \dots, X_p$ i Y, tada tražimo funkciju $\varphi(x_1, x_2, \dots, x_p)$ za koju će biti:

$$Y \approx \varphi(X_1, X_2, \dots, X_p)$$

- Funkcija φ se bira tako da srednje-kvadratno odstupanje $E\left(Y - \varphi(X_1, X_2, \dots, X_p)\right)^2$ bude najmanje.
- Funkcija φ se zove regresija Y po $X_1, X_2, \dots, X_p$.

Jednostruka regresija

Šta je jednostruka regresija?

- Najjednostavnija regresija je jednostruka regresija kada se posmatraju dve promenljive X i Y.
- Traži se funkcija φ takva da je $Y \approx \varphi(X)$.
- Iz populacije se izdvaja realizovani uzorak $\big((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\big)$ i predstavlja se u Dekartovoj ravni.
- Na osnovu dobijenog dijagrama, koji se zove dijagram rasturanja, bira se familija funkcija sa kojom će se raditi, npr:
  - $y = ax + b$,
  - $y = ae^{bx}$,
  - $y = a \ln x + b$.
- Na kraju se određuju vrednosti parametara regresije.

Jednostruka regresija

Jednostruka linearna regresija.

- Ako je zavisnost φ između promenljivih X i Y linearna, tj. oblika $Y = aX + b$, tada kažemo da je φ jednostruka linearna regresija.
- Jednostruka linearna regresija može biti:
  - prve vrste, ukoliko obeležje Y zavisi od slučajne promenljive X,
  - druge vrste, ukoliko obeležje Y zavisi od neslučajne (kontrolisane) promenljive X.

Jednostruka linearna regresija prve vrste.

- Zavisnost je oblika $Y = aX + b$.
- Parametri a i b su nepoznati parametri koji se određuju iz uslova da je srednje-kvadratno odstupanje $E(Y - aX - b)^2$ najmanje.
- Ocene parametara su:

$$\hat{a} = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}$$

$$\hat{b} = \bar{Y}_n - \hat{a}\bar{X}_n$$

- Vrednost $\hat{Y}_i$ dobijena kao $\hat{Y}_i = \hat{a}\hat{X}_i + \hat{b}$ je procena vrednosti Y_i na osnovu vrednosti X_i.

Jednostruka linearna regresija prve vrste.

- Greška ocenjivanja se definiše na sledeći način:

$$S^2_{Y-\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- Greška ocenjivanja sadrži informaciju o tome koliko izabrana funkcija dobro aproksimira zavisnost obeležja Y po X.
- U slučaju da je uzorak mali, greška ocenjivanja se definiše na sledeći način:

$$S^2_{Y-\hat{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

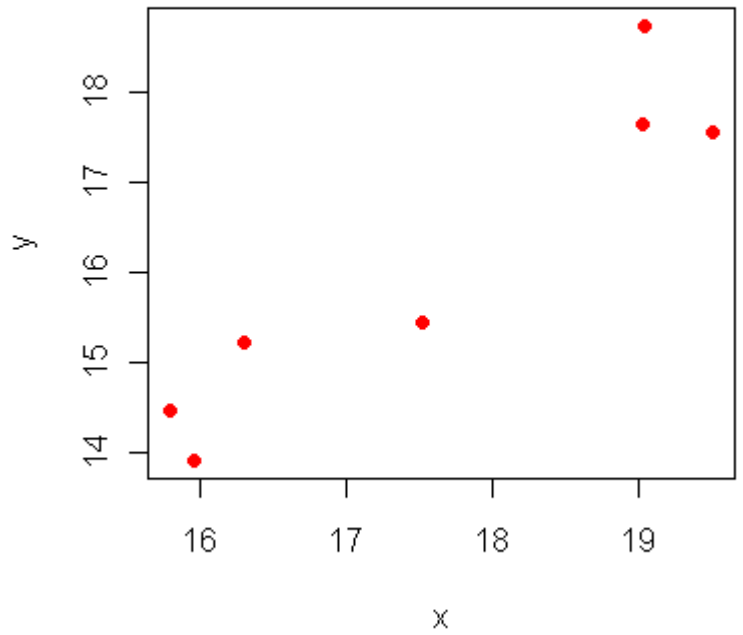
Primer.

- Na osnovu merenja pretprolećnog minimalnog srednjemesečnog nivoa podzemnih voda (X), i srednjeg godišnjeg nivoa (Y) u godinama 1952-1958. dobijeno je:

godina	1952	1953	1954	1955	1956	1957	1958
X	19,51	19,02	19,04	15,80	17,52	15,96	16,31
Y	17,57	17,66	18,74	14,48	15,44	13,92	15,22

- Odrediti pravu linearne regresije Y po X i na osnovu nje prognozirati Y za izmerenu vrednost x=16,99 u 1959. godini. Izračunati grešku ocenjivanja.

Primer – dijagram rasturanja.



Primer – rešenje.

- Koeficijenti prave linearne regresije su:

$$\hat{a} = \frac{2005{,}18 - \frac{1}{7} \times 123{,}16 \times 113{,}03}{2182{,}25 - \frac{1}{7} \times 123{,}16^2} = 1{,}076$$

$$\hat{b} = \frac{113{,}03}{7} - 1{,}076 \times \frac{123{,}16}{7} = -2{,}784$$

- Tražena prava linearne regresije je: $\hat{Y} = 1{,}076 \times X - 2{,}784$.
- Procenjujemo da je u 1959. godini srednji godišnji nivo bio:

$$y(16{,}99) = 1{,}076 \times 16{,}99 - 2{,}784 = 15{,}50$$

Primer – rešenje.

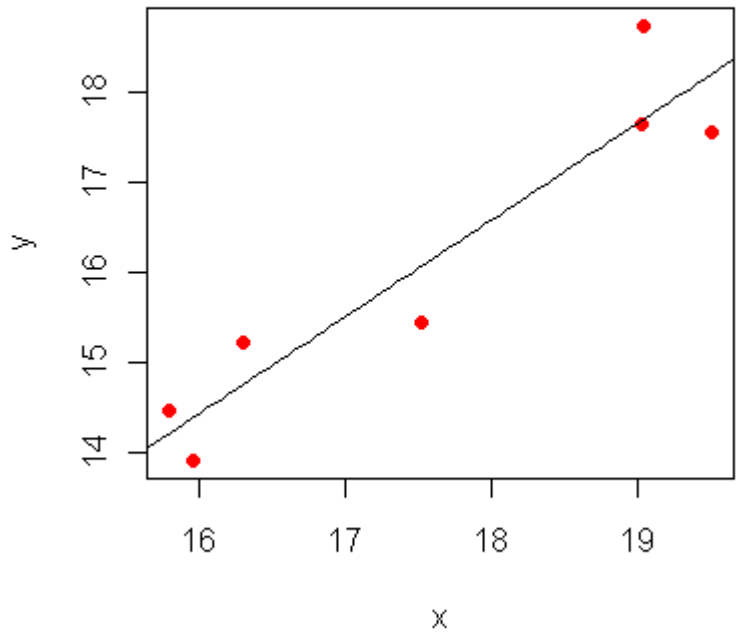
- Greška ocenjivanja je:

$$S^2_{Y-\hat{Y}} = \frac{2{,}3795}{6} = 0{,}3966$$

- Pošto je greška mala u odnosu na vrednosti obeležja Y, zaključujemo da smo dobro aproksimirali vrednosti obeležja Y po X.

Jednostruka regresija

Primer – grafik prave linearne regresije.



Jednostruka linearna regresija prve vrste.

- Može se posmatrati i linearna regresija X po Y.
- Ona je oblika $X = cY + d$, gde su c i d nepoznati parametri koji se određuju iz uslova da je srednje-kvadratno odstupanje $E(X - cY - d)^2$ najmanje.
- Ocene parametara su:

$$\hat{c} = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sum Y_i^2 - \frac{1}{n}(\sum Y_i)^2}$$

$$\hat{d} = \bar{X}_n - \hat{c}\bar{Y}_n$$

Jednostruka linearna regresija prve vrste.

- Greška ocenjivanja se definiše na sledeći način:

$$S_{X-\hat{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2$$

- U slučaju da je uzorak mali, greška ocenjivanja se definiše na sledeći način:

$$S_{X-\hat{X}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2$$

Jednostruka regresija

Jednostruka linearna regresija druge vrste.

- Zavisnost je oblika $Y = aX + b + \varepsilon$, gde je ε slučajna promenljiva koja se najčešće identifikuje kao greška merenja.

Nelinearni modeli zavisnosti

Šta su nelinearni modeli zavisnosti?

- Nelinearni modeli zavisnosti su modeli koji jednostavnim transformacijama mogu da se svedu na linearne modele.
- Posmatraćemo:
  - model linearan po parametrima,
  - stepeni model,
  - eksponencijalni model.
- Nelinearni modeli mogu biti prve i druge vrste.

Model linearan po parametrima.

- Model linearan po parametrima prve vrste je oblika

$$Y = ag(X) + b$$

- gde je g unapred poznata funkcija od slučajne promenljive X, a a i b nepoznati parametri.
- Model se smenom $Z = g(X)$ svodi na linearan model prve vrste $Y = aZ + b$.
- Nepoznati parametri a i b se ocenjuju na osnovu uzorka $\big((Z_1, Y_1), \dots, (Z_n, Y_n)\big)$ koji je dobijen transformacijom $Z_i = g(X_i), i = 1, \dots, n$ od početnog uzorka $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$.

Model linearan po parametrima.

- Ocene parametara a i b regresionog modela $Y = aZ + b$ su:

$$\hat{a} = \frac{\sum Z_i Y_i - \frac{1}{n}\sum Z_i \sum Y_i}{\sum Z_i^2 - \frac{1}{n}(\sum Z_i)^2}$$

$$\hat{b} = \bar{Y}_n - \hat{a}\bar{Z}_n$$

- U tom slučaju početni regresioni model je oblika:

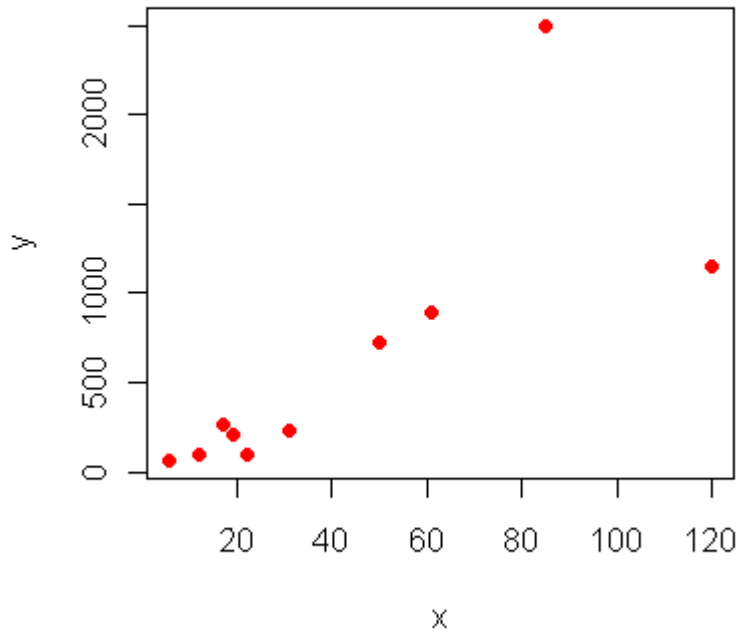
$$\hat{Y} = \hat{a}g(X) + \hat{b}$$

Primer.

- Odrediti regresionu krivu $Y = a \log X + b$ veze između širine reke Y i maksimalog godišnjeg proticaja X (u m^3/sec) na osnovu uzorka od deset reka:

maks. proticaj	5,7	17	22	31	50	61	85	120	12	19
širina reke	63	260	92	230	720	890	2500	1150	93	210

Primer – dijagram rasturanja.



Primer – rešenje.

- Realizovane vrednosti ocena $\hat{a}$ i $\hat{b}$ su redom:

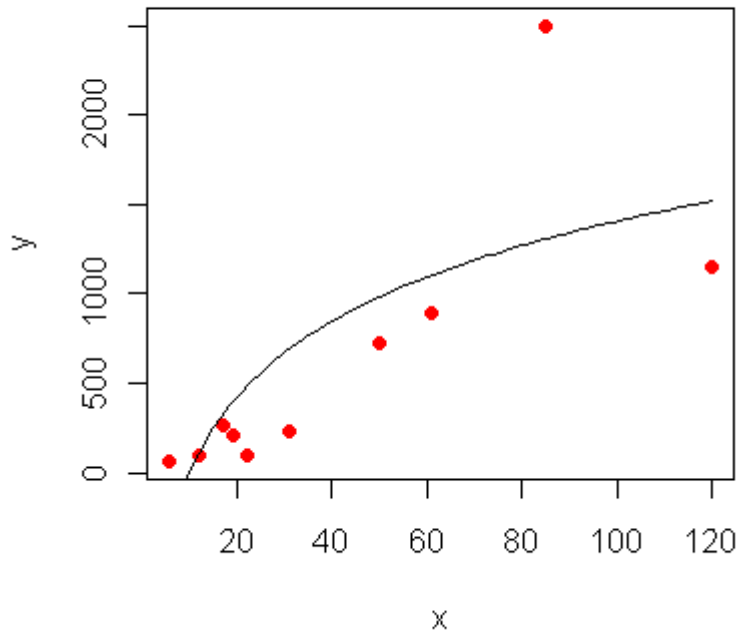
$$\hat{a} = \frac{25857{,}48 - \frac{1}{10} \times 33{,}781 \times 6208}{122{,}108 - \frac{1}{10} \times 33{,}781^2} = 611{,}360$$

$$\hat{b} = \frac{6208}{10} - 611{,}360 \times \frac{33{,}781}{10} = -1407{,}754$$

- Regresiona prava između obeležja Y i Z je oblika: $\hat{Y} = 611{,}360 \times Z - 1407{,}754$.
- Regresiona kriva koja opisuje vezu između širine reke i maksimalnog godišnjeg proticaja je:

$$\hat{Y} = 611{,}360 \log X - 1407{,}754$$

Primer – grafik regresione krive.



Stepeni model.

- Stepeni model prve vrste je oblika

$$Y = bX^a$$

- gde su a i b nepoznati parametri.
- Model se smenama $Z = \ln X$, $W = \ln Y$ i $c = \ln b$ svodi na linearan model prve vrste, odnosno na $W = aZ + c$.
- Nepoznati parametri a i c se ocenjuju na osnovu uzorka $\big((Z_1, W_1), \dots, (Z_n, W_n)\big)$ koji je dobijen transformacijama $Z_i = \ln X_i\,, i = 1, \dots, n$ i $W_i = \ln Y_i\,, i = 1, \dots, n$ od početnog uzorka $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$.

Stepeni model.

- Ocene parametara a i c regresionog modela $W = aZ + c$ su:

$$\hat{a} = \frac{\sum Z_i W_i - \frac{1}{n}\sum Z_i \sum W_i}{\sum Z_i^2 - \frac{1}{n}(\sum Z_i)^2}$$

$$\hat{c} = \bar{W}_n - \hat{a}\bar{Z}_n$$

- Parametar b ocenjujemo kao $\hat{b} = e^{\hat{c}}$.
- U tom slučaju početni regresioni model je oblika:

$$\hat{Y} = \hat{b}X^{\hat{a}}$$

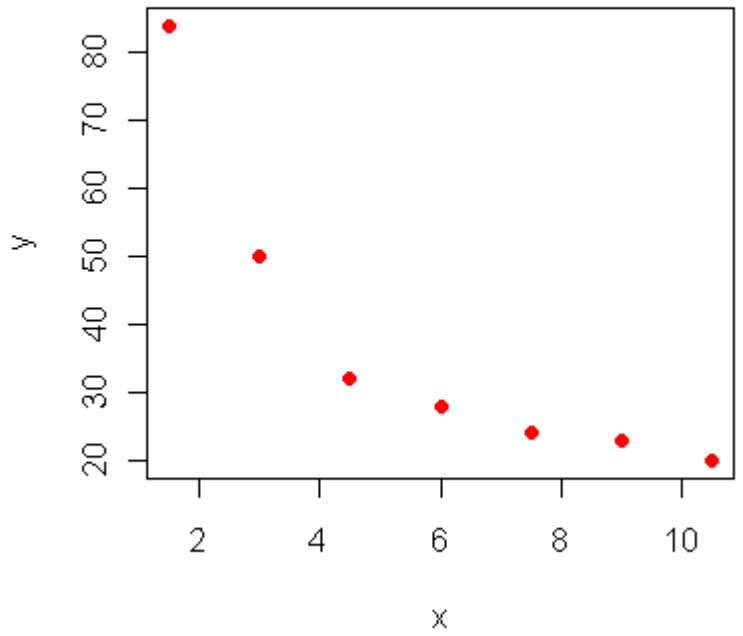
Primer.

- Na obali zaliva se ispituje vlažnost mulja (u gramima vode na 100 grama suve materije). Iz jedne bušotine je dobijeno:

dubina (X)	1,5	3	4,5	6	7,5	9	10,5
vlažnost (Y)	84	50	32	28	24	23	25

- Odredite regresionu krivu $Y = bX^a + \varepsilon$.

Primer – dijagram rasturanja.



Primer – rešenje.

- Realizovane vrednosti ocena nepoznatih paramaetara su:

$$\hat{a} = \frac{37{,}614 - \frac{1}{7} \times 11{,}363 \times 24{,}450}{21{,}259 - \frac{1}{7} \times 11{,}363^2} = -0{,}738$$

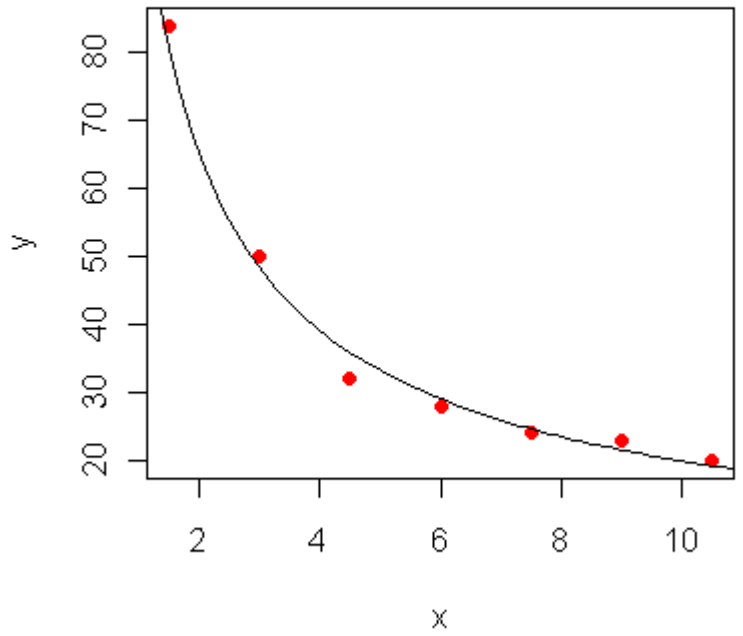
$$\hat{c} = \frac{24{,}450}{7} + 0{,}738 \times \frac{11{,}363}{7} = 4{,}961$$

$$\hat{b} = e^{4{,}961}$$

Primer – rešenje.

- Regresiona prava između obeležja W i Z je oblika: $\widehat{W} = -0{,}738 \times Z + 4{,}691$.
- Regresiona kriva koja opisuje vezu između dubine i vlažnosti mulja je: $\hat{Y} = 108{,}962X^{-0{,}738} + \varepsilon$.

Primer – grafik regresione krive.



Eksponencijalni model.

- Eksponencijalni model prve vrste je oblika

$$Y = be^{aX}$$

- gde su a i b nepoznati parametri.
- Model se smenama $W = \ln Y$ i $c = \ln b$ svodi na linearan model prve vrste $W = aX + c$.
- Nepoznati parametri a i c se ocenjuju na osnovu uzorka $\big((X_1, W_1), \dots, (X_n, W_n)\big)$ koji je dobijen transformacijom $W_i = \ln Y_i\,, i = 1, \dots, n$ od početnog uzorka $\big((X_1, Y_1), \dots, (X_n, Y_n)\big)$.

Eksponencijalni model.

- Ocene parametara a i c regresionog modela $W = aX + c$ su:

$$\hat{a} = \frac{\sum X_i W_i - \frac{1}{n}\sum X_i \sum W_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}$$

$$\hat{c} = \overline{W}_n - \hat{a}\overline{X}_n$$

- Parametar b ocenjujemo kao $\hat{b} = e^{\hat{c}}$.
- U tom slučaju početni regresioni model je oblika:

$$\hat{Y} = \hat{b}e^{\hat{a}X}$$

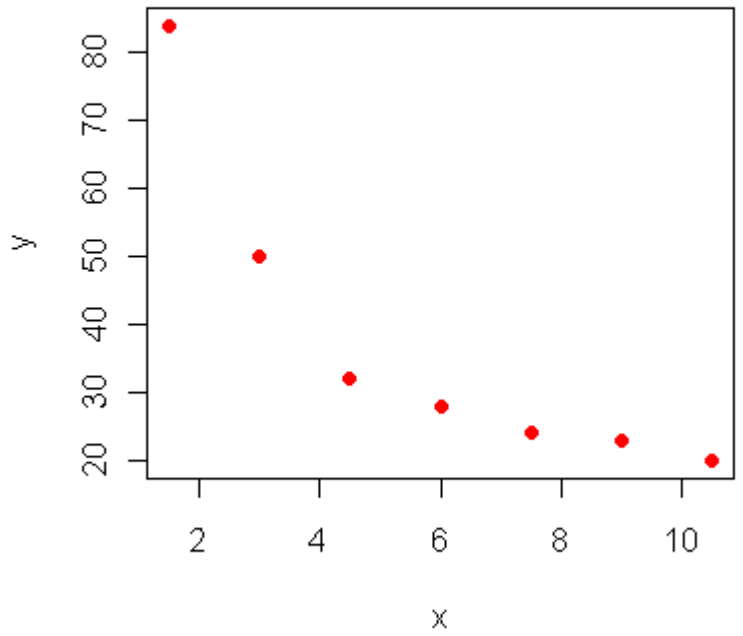
Primer.

- Na obali zaliva se ispituje vlažnost mulja (u gramima vode na 100 grama suve materije). Iz jedne bušotine je dobijeno:

dubina (X)	1,5	3	4,5	6	7,5	9	10,5
vlažnost (Y)	84	50	32	28	24	23	25

- Odredite regresionu krivu $Y = be^{aX} + \varepsilon$.

Primer – dijagram rasturanja.



Primer – rešenje.

- Realizovane vrednosti ocena nepoznatih paramaetara su:

$$\hat{a} = \frac{137{,}8 - \frac{1}{7} \times 42 \times 24{,}450}{315 - \frac{1}{7} \times 42^2} = -0{,}146$$

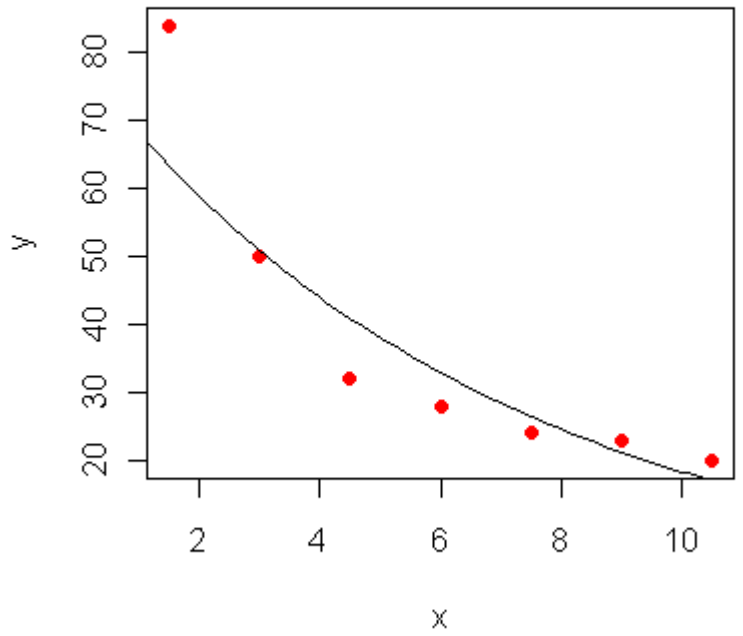
$$\hat{c} = \frac{24{,}450}{7} + 0{,}146 \times \frac{42}{7} = 4{,}369$$

$$\hat{b} = e^{4{,}366} = 78{,}965$$

Primer – rešenje.

- Regresiona prava između obeležja W i X je oblika: $\widehat{W} = -0{,}146 \times X + 4{,}369 + \varepsilon$.
- Regresiona kriva koja opisuje vezu između dubine i vlažnosti mulja je: $\widehat{Y} = 78{,}965e^{-0{,}146X} + \varepsilon$.

Primer – grafik regresione krive.



Završne napomene

- Sama reč regresija znači vraćanje unatrag.
- Naziv regresija za ovakve modele potiče od statitičara Sir Fransis Galtona, (1822 - 1911) koji je prvi upotebio date modele za ispitivanje bioloških i psiholoških pojava.
- Galton je ispitivao vezu između visina roditelja i dece – ustanovio je da roditelji sa natprosečnom visinom teže da imaju decu koja su takođe viša od proseka, ali ne toliko visoka kao roditelji. Ta činjenica važi i za decu izuzetno niskih roditelja – deca su taođe bila niska ali ne toliko niska kao roditelji. Visine su se „vraćale natrag“, regresirale prema proseku.
- Galton je ovu činjenicu nazvao „regresija prema proseku“, a naziv je ostao i za metodu.

Literatura

- Beleške pripremljene prema knjizi – Luis Pedro Coelho, Willi Richert (2015): „Building Machine Learning Systems with Python, Second Edition“. Packt Publishing.
- Dodatni materijali korišćeni prilikom pripreme beleški:
  - „Regresija“, predavanja dostupna na sajtu Prirodno matematičkog fakulteta, Univerziteta u Nišu.

Hvala na pažnji

Pitanja su dobrodošla.