LSM

Bojana Milošević

9/30/2019

Osnovna literatura i obaveze

- beleške
- *Linear models with R*, J.J.Faraway
- *Regression analyses: theory, methods, and aplications*, A. Sen, M. Srivastava
- *Regression Models: Methods and Applications*, L. Fahrmeir, T. Kneib, S. Lang, B. Marx

Obaveze

- seminarski
- ispit

Čime ćemo se baviti?

- Kakva je veza između različitih obeležja?

PARMETARSKI PRISTUP

- Kada odredimo oblik modela kako da ocenimo njegove parametre? (frekvencionistički pristup)
  - linearni modeli (prosta, višestuka linearna regresija)
  - uopšteni linearni modeli (logistićka, Puasonova regresija. . .)
- Koji su modeli "dopustivi" i u kom smislu?
- Kako da ispitamo kvalitet modela?

Šta posle?

- Sve što budemo radili se može sagledati i Bajesovim pristupom
- NEPARAMETARSKI PRISTUP
- ...

Regresiona funkcija

- Želimo da modeliramo zavisnost Y od nekih poznatih promenljivih X (može biti vektor). Y je zavisna promenljiva a X nezavisna. Funkcija $f(X)$ koja je "najbliža" Y je regresiona funkcija

$$f(X) = E(Y|X)$$

- Treba pronaći odgovarajući oblik funkcije f
- Aditivni modeli

$$Y = f(X) + \varepsilon$$

gde je ε greška koja, ako znamo X, predstavlja jedinu slučajnu komponentu u modelu. Treba da važi $E(\varepsilon) = 0$.

Linearni regresioni model

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- linearna funkcija po parametrima $\beta_0, \beta_1, ..., \beta_p$

  - $p = 1$ prosta linearna regresija

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Prosta linearna regresija

Polazimo od uzorka $(x_1, y_1), ..., (x_n, y_n)$ uz pretpostavku da svaki par se može opisati prostim linearnim regresionim modelom, odnosno

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

uz uslove za $\{\varepsilon_i\}$:

1. NJR (nezavisne i jednako raspodeljene) sluchajne veličine koje ne zavise od $\{x_i\}$
2. $E(\varepsilon_i) = 0$
3. $D(\varepsilon_i) = \sigma^2 < \infty$
4. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Ocenjivanje parametara

- Metod najmanjih kvadrata (minimizira se srednjekvadratno odstupanje stvarne vrednosti i vrednosti ocenjene modelom)

$$\min S(\beta_0, \beta_1) = \min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (1)$$

- Dobija se

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{a}\bar{x}.$$

- Metod maksimalne verodostojnosti
  - Svodi se na optimizacioni problem (1)

Osobine ocena

- nepristrasnost

$$E(\hat{\beta}_0) = \beta_0 \quad E(\hat{\beta}_1) = \beta_1$$

- postojanost

$$D(\hat{\beta}_1) = \frac{\sigma^2}{nS_x^2}$$

$$D(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{S_x^2}\right)$$

Osobine ocena

Ukoliko važi uslov (4):

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right)$
- $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}}{S_x^2}\right)\right)$
- Ocene možemo standardizovati tako da imaju standardnu raspodelu
- PROBLEM:

Osobine ocena

Ukoliko važi uslov (4):

- $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right)$
- $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}}{S_x^2}\right)\right)$
- Ocene možemo standardizovati tako da imaju standardnu raspodelu
- PROBLEM: Ne znamo σ^2

Osobine ocena

- $e_i = y_i - \hat{y}_i$ reziduali modela
- $\sum_i e_i = 0$
- Nepristrasna ocena za σ^2 je $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
- $\frac{(n-2)\hat{\sigma}^2}{\hat{\sigma}^2} \sim \chi^2_{n-2}$
- $$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{nS_x^2}}} \sim t_{n-2}$$
- $$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2}{n}\left(1 + \frac{\bar{x}}{S_x^2}\right)}} \sim t_{n-2}$$
- Možemo konstruisati intervale poverenja za ocene, testirati hipoteze o njihovim vrednostima itd.

Ocenjena vrednost zavisne promenljive

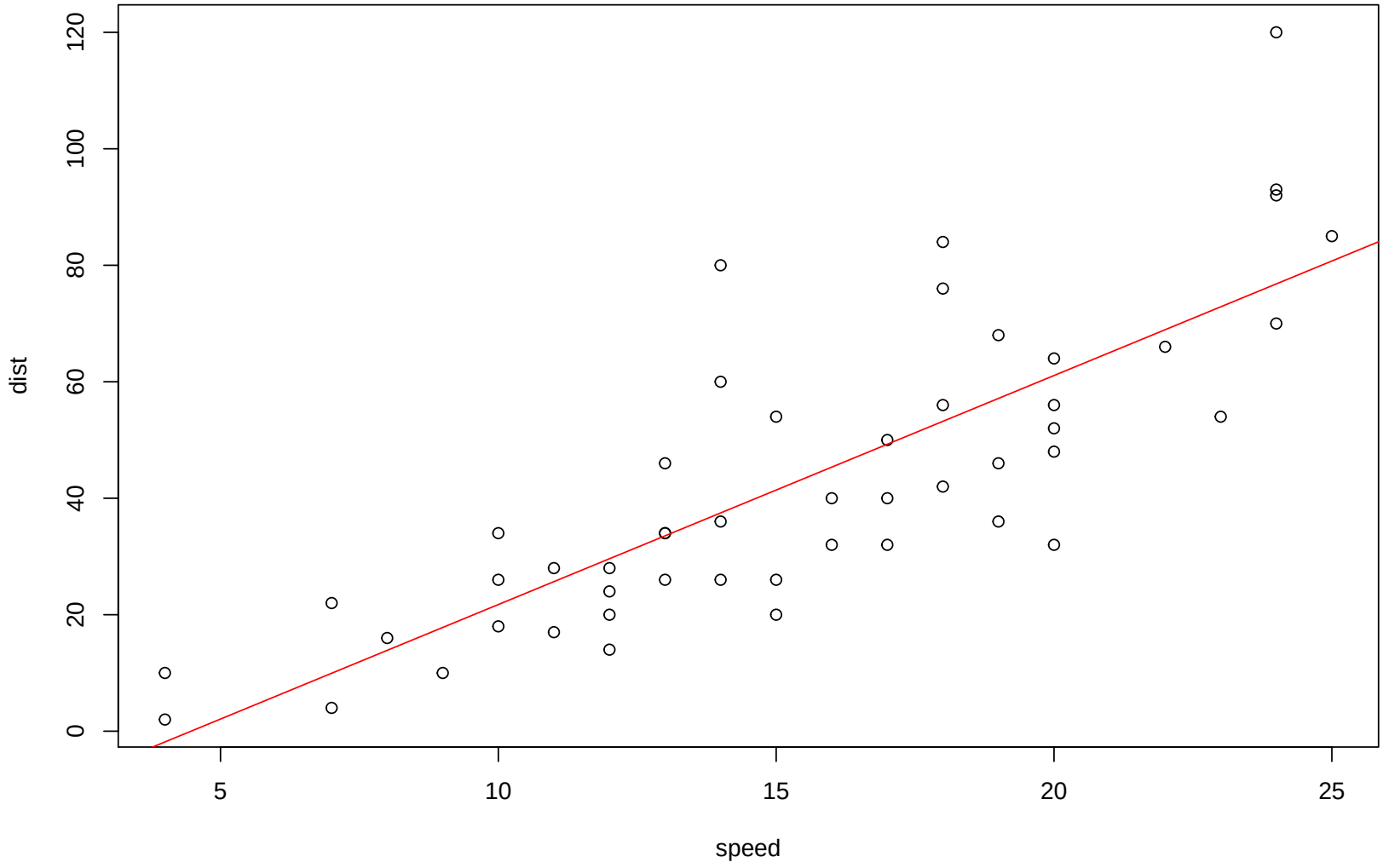
- Ocenjena vrednost za srednju vrednost zavisne promenljivu kada prediktor uzima vrednost x_0 je $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Kako - može se prikazati kao linearna kombinacija $\{y_i\}$
- $$\frac{\hat{Y}_0 - EY_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nS_x^2}}} \sim t_{n-2}$$
- Iz $Y_0 = \beta_0 + \beta x_0 + \varepsilon_0$ dobijamo da je ocenjena vrednost $\hat{\hat{Y}}_0$ ista kao ocenjena vrednost odgovarajuće srednje vrednosti pa c1e odgovarajuc1i interval poverenja biti širi, odnosno

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nS_x^2}}} \sim t_{n-2}$$

Primer

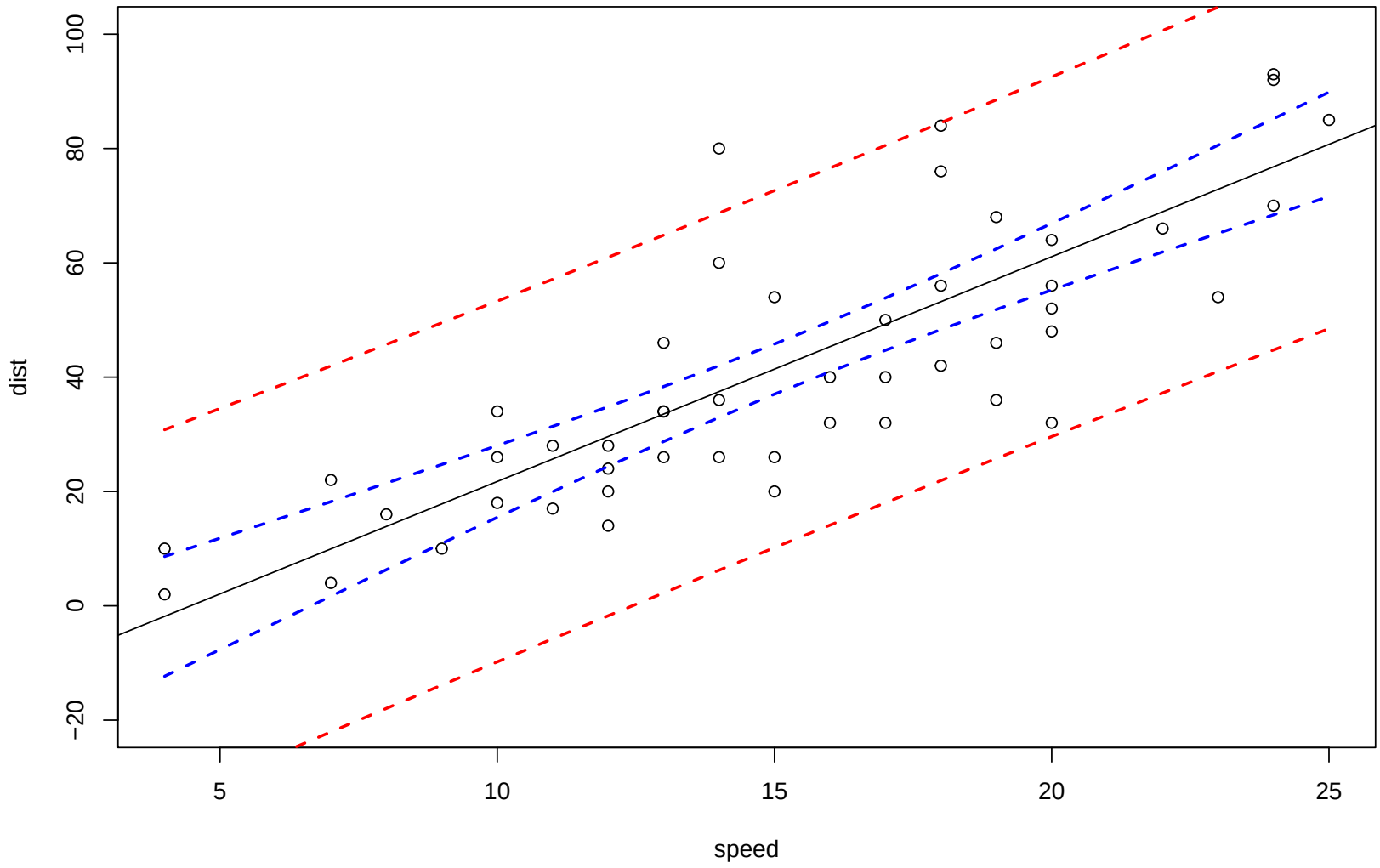
```
##    speed dist
## 1      4    2
## 2      4   10
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
## 8     10   26
## 9     10   34
## 10    11   17
```

Primer



$$\hat{a} = 3.93 \quad \hat{b} = -17.59$$

Intervali poverenja



Koliko je model dobar?

$$SSE = \sum_{i=1}^{n} e_i^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSTO = SSE + SSR$$

$$R^2 = 1 - \frac{SSE}{SSTO}$$

Primer

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```