"Essentially, all models are wrong, but some are useful"
George E.P. Box

Како се променом једне или више независних случајних промењливих мења вредност зависне случајне величине? Како одредити аналитичко-математички облик одговарајуће везе? Одговор на ова, као и на низ других питања даје нам управо регресија. Овај курс биће посвећен линеарној регресији. Идеје које се овде користе могу послужити и приликом анализирања других типова регресије.

Први записи о методи најмањих квадрата могу се наћи у радовима Лежандра и Гауса, почетком 19. века. Они су овај метод користили за одређивање орбита небеских тела око Сунца. Са речју "регресија" математичари су се први пут сусрели у раду Ф. Галтона, *Regression toward mediocrity in hereditary stature* из 1855. године. Он је дошао до закључка да синови веома високих очева нису тако високи. Иако је Галтон разлог за то пронашао у генетици, његов пример иницирао је проучавање ове теме од стране статистичара и тако почиње развој ове веома значајне статистичке области.

Дефиниција 0.0.1. *Регресија је зависност једне случајне промењливе од друге (или више њих). Регресиони модел је математички модел који описује ту зависност.*

Дефиниција 0.0.2. *Случајна величина $f(X) = E(Y|X)$ назива се регресиона функција, при чему X може бити вишедимензиона случајна величина.*

Следећа теорема оправдава облик функције регресије.

Теорема 0.0.1 (2)**.**

$$E(Y - E(Y|X))^2 \leq E(Y - g(X))^2$$

за сваку функцију $g(X)$, уз претпоставку да постоји математичко очкевиање на десној страни неједнакости.

Доказ.

$$\begin{aligned} E(Y - g(X))^2 &= E(Y - E(Y|X) + E(Y|X) - g(X))^2 \\ &= E\Big(E(Y - E(Y|X) + E(Y|X) - g(X))^2|X\Big) \\ &= E\Big(E(Y - E(Y|X)^2|X + E(E(Y|X) - g(X))^2)|X\Big) \\ &\geq E(Y - E(Y|X)^2). \end{aligned}$$

□

Претходну неједнакост можемо да посматрамо и из геометријског угла јер $E(Y|X)$ ће представљати пројекцију Y на раван одређену случајним вектором X.

Регресиони модел се може представити у облику

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

где је ε случајна промењлива независна од X, најчешће са нормалном $\mathcal{N}(0, \sigma^2)$ расподелом.

Уколико из нпр. графичког приказа зависности (X, Y) имамо разлога да претпоставимо да је $f(X) = aX + b$ онда се коефицијенти a, b одређују тако да се минимизира $E(Y - (aX + b))^2$.

Добија се да је

$$\begin{aligned} a &= \frac{EXY - EXEY}{DX} \\ b &= EY - aEX, \end{aligned}$$

па се коефицијенти a, b могу оценити нпр. методом замене, односно

$$\begin{aligned} \hat{a} &= \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\bar{S}_X^2} = \hat{\rho}\frac{\bar{S}_X}{\bar{S}_Y} \\ \hat{b} &= \bar{Y} - \hat{a}\bar{X}. \end{aligned}$$

Уколико претпоставимо да X није случајна променљива говоримо о *контролисаној регресији*.

Имајући у циду саму дефиницију регресионе функције, од сада па надаље можемо претпоставити да се ради о контролисаној регресији.

Дакле, наш циљ ће бити да оценимо функцију $f(x)$ при чему ћемо претпоставити неку зависност до на непознате параметре (параметарски приступ). Овом проблему се свакако може приступити и непараметарски и о томе ће бити више речи следеће године.

Да поновимо, главни наш задатак у овом курсу је да одговоримо на следећа питања.

- Каква је веза између различитих обележја?
- Када одредимо облик модела како да оценимо његове параметре?
- Који су модели "допустиви" и у ком смислу?
- Како да испитамо квалитет модела?

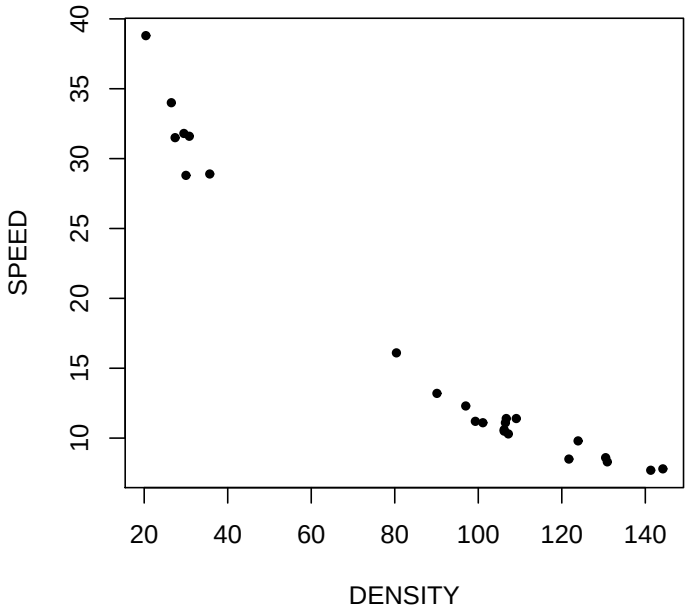
Понекад се за ову проблематику користи термин *статистичко учење* јер желимо да, користећи доступне податке, "научимо" модел.

Како изгледа цео процес бирања модела демонстрираћемо на наведеном примеру.

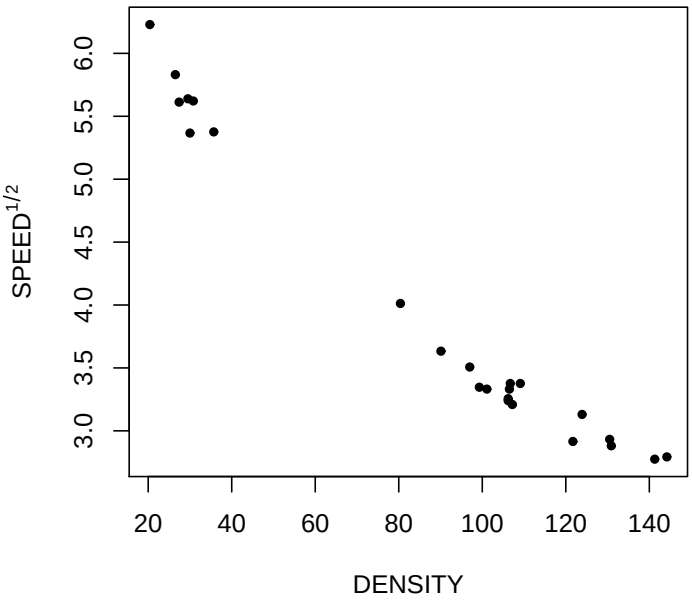
Пример 0.0.1. *У циљу истраживања у којој мери број возила на путу утиче на брзину возила сакупљани су подаци о "густини" возила (број аутомобила у једној миљи) и просечној брзини аутомобила.*

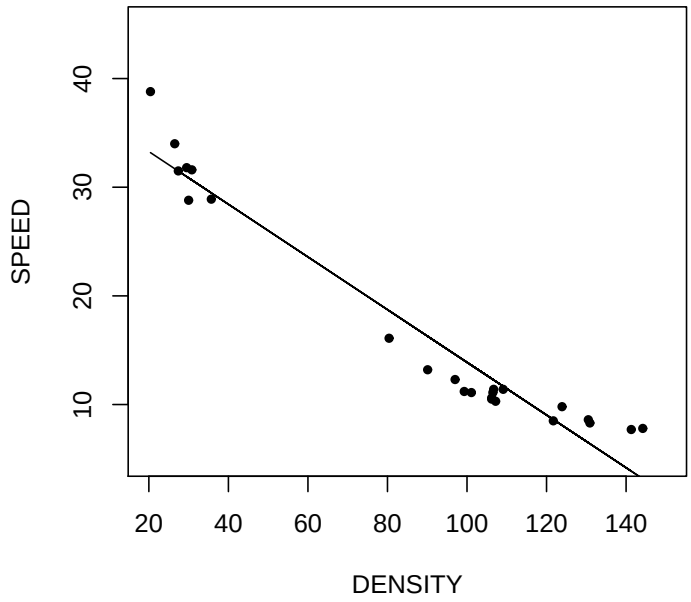
	DENSITY	SPEED
1	20.40	38.80
2	27.40	31.50
3	106.20	10.60
4	80.40	16.10
5	141.30	7.70
6	130.90	8.30
7	121.70	8.50
8	106.50	11.10
9	130.50	8.60
10	101.10	11.10
11	123.90	9.80
12	144.20	7.80
13	29.50	31.80
14	30.80	31.60
15	26.50	34.00
16	35.70	28.90
17	30.00	28.80
18	106.20	10.50
19	97.00	12.30
20	90.10	13.20
21	106.70	11.40
22	99.30	11.20
23	107.20	10.30
24	109.10	11.40

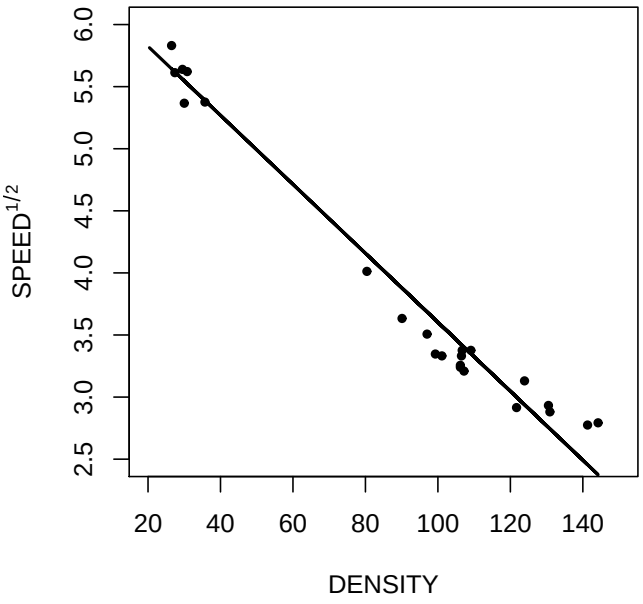
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.1295	1.2177	31.31	0.0000
E1.1$DENSITY	-0.2425	0.0126	-19.22	0.0000

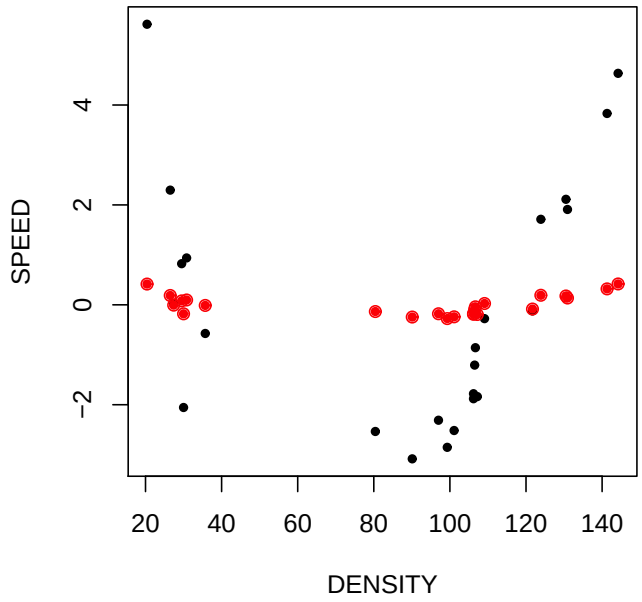


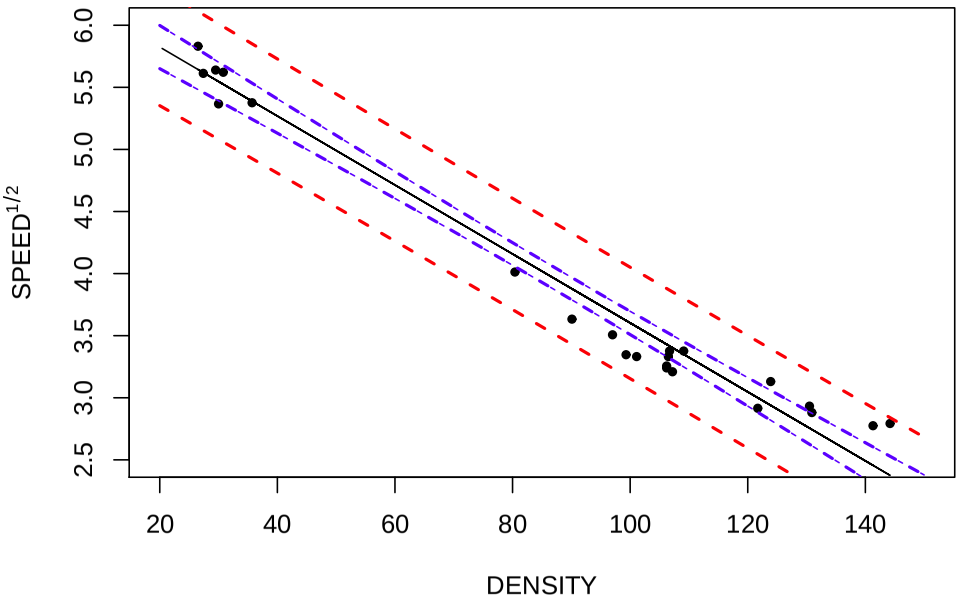
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.3797	0.1028	62.09	0.0000
E1.1$DENSITY	-0.0278	0.0011	-26.09	0.0000


SPEED$^{1/2}$
3.0
3.5
4.0
4.5
5.0
5.5
6.0
20
40
60
80
100
120
140
DENSITY


SPEED
10
20
30
40
20
40
60
80
100
120
140
DENSITY


SPEED$^{1/2}$
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
20
40
60
80
100
120
140
DENSITY


SPEED
−2
0
2
4
20
40
60
80
100
120
140
DENSITY


SPEED1/2
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
20
40
60
80
100
120
140
DENSITY

Први корак је свакако да графички представимо податке и да уочимо неки облик зависности (ако постоји).

Са првог графика можемо закључити да са повећањем густине саобраћаја опада брзина истог, што је сасвим очекиван закључак. Најједноставнији модел који би могао да опише податке је линеарна веза, односно $y = ax + b + \varepsilon$, где је y просечна брзина аутомобила а x густина собраћаја. Јасно је да у модел морамо да укључимо и неки "шум" (ε) који би оправдао то што тачке на графику нису све колинеарне. Неке природне особине које тај шум треба да задовољава је да је 'мали', да је "центриран" око нуле, да не зависи од x и y итд. Шум заправо представља грешку модела.

Један од најпопуларнијих, најједноставнијих и слободно можемо рећи основних метода за оцену параметара модела је метод најмањих квадрата. Идеја је да параметре оценимо оним вредностима који минимизирају суму квадратних одступања оцењене од праве вредности, односно

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2.$$

Добијамо да су тражени $\hat{a}$ и $\hat{b}$

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

Означимо са $e_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + b)$

Приметимо да полазни модел можемо написати у центрираном облику $y_i = a(x_i - \bar{x}) + b + a\bar{x} + \varepsilon_i$. Испоставља се да је овај облик погоднији за прогнозирање јер $\hat{y}_i = \hat{a}(x_i - \bar{x}) + \bar{y}$.

Још је важно да се примети да је $\sum_{i=1}^{n} e_i = 0$.

У (0.0.1) једначине правих које се добијају у првом, односно другом моделу су $y = -0.24x + 38.13$, односно $y = -0.028x + 6.38$.

Када је модел добар одступања оцењених вредности од правих (резидуали) су мали. Зато је природно за меру квалитета модела узети, за початак, $\sum_{i=1}^{n} e_i^2$. Главни проблем са овом мером одступања је зато што она зависи од јединице. Зато ћемо искористи сличну идеју као за уводјење коефицијента корелације.

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

SSR је одступање које је објашњено моделом. Зато уводимо *коефицијент детерминације* R^2 као меру квалитета модела.

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

Јасно је да ако бисмо имали перфектан модел онда би $R^2 = 1$.

У примеру 0.0.1 за први модел се добија да је $R^2 = 0.94$ док је за други модел $R^2 = 0.98$.

Може се показати да је $R = |\rho_{xy}|$.

Напомена: не треба увек (само) користити R^2 као меру квалитета модела. О томе ће бити више речи у остатку курса.

Поглавље 1

Линеарни модели

1.1 Проста линеарна регресија

У уводном поглављу смо претпостављали да имамо један предиктор. Такав модел се назива *прост линеарни регресиони модел*. Видели смо нека лепа својства која модел поседује при чему нисмо наводили које претпоставке модел треба да зодовољава да бисмо га уопште разматрали.

Нека је

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, 2, ..., n$$

прост линеарни модел, при чему шум задовољава следеће услове Гаус-Маркова

1. центрираност $E\varepsilon_i = 0, \quad i = 1, 2, ..., n$
2. некорелисаност $E\varepsilon_i\varepsilon_j = 0, \ i \neq j$;
3. хомоскедастичност $D\varepsilon_i = \sigma^2 > 0$;
4. x_i и ε_j су независни за свако i, j.

Тада за оцене добијене методом најмањих квадрата

$$\begin{aligned}\hat{a} &= \frac{\sum_{i=1}^n Y_i x_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n Y_i (x_i - \bar{x})}{nS_x^2} \\ \hat{b} &= \bar{y} - \hat{a}\bar{x} = \frac{\sum_{i=1}^n Y_i (S_x^2 - \bar{x}(x_i - \bar{x}))}{nS_x^2},\end{aligned}$$

важе следећа својства:

1. $E\hat{a} = 0$ и $E\hat{b} = 0$;

2. $$D(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$D(\hat{b}) = \frac{\sigma^2}{n}\Big(1 + \frac{\bar{x}}{S_x^2}\Big);$$

Дакле, оцене параметара модела су непристрасне и постојане. Видимо да у изразима за дисперзије вигуришe непознати параметар σ^2. Зато морамо и за њега наћи одговарајућу оцену.

$$\begin{aligned}
E(SSE) &= E(SST) - E(SSR)\\
E(SST) &= E(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2)\\
&= \sum_{i=1}^{n}(\sigma^2 + (aX_i + b)^2) - \frac{1}{n}(\sum_{i=1}^{n} EY_i^2 + 2\sum_{1\le i<j<n} EY_iEY_j)\\
&= (n\sigma^2 + \sum_{i=1}^{n}((aX_i + b)^2))\frac{(n-1)}{n} - \frac{2}{n}(\sum_{1\le i<j\le n}(aX_i + b)(aX_j + b))\\
&= \sigma^2(n-1) + nS_x^2a^2\\
E(SSR) &= \sum_{i=1}^{n}(x_i - \bar{x})^2E\hat{a}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2(a^2 + \frac{\sigma^2}{nS_x^2})\\
&= nS_x^2(a^2 + \frac{\sigma^2}{nS_x^2})\\
E(SSE) &= \sigma^2(n-2)
\end{aligned}$$

Одавде добијамо да је непристрасна оцена за σ^2

$$\hat{\sigma}^2 = \frac{SSE}{n-2}.$$

Уколико се у модел уведе додатна претпоставка да је шум Гаусов, односно да ε_i има нормалну $\mathcal{N}(0, \sigma^2)$, добијене оцене имају многа друга лепа својства. Прво, приметимо да су оцене за a и b линеарне комбинације независних случајних величина са нормалним расподелама па и

саме оцене имају нормалне $\mathcal{N}(a, D\hat{a})$ и $\mathcal{N}(b, D\hat{b})$ расподеле. У наставку курса ћемо показати да $\hat{a}$ и $\hat{b}$ су независне од $\hat{\sigma}^2$ па закључујемо да

$$\frac{\hat{a} - a}{\frac{\hat{\sigma}}{\sqrt{nS_x^2}}} \sim t_{n-2}$$
$$\frac{\hat{b} - b}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}}} \sim t_{n-2}$$

Сада се могу правити интервали поверења за a и b и тестирати хипотезе у вези са њиховим параметрима. Приметимо да $H_0 : a = 0$ заправо значи да утицај предиктора није значајан.

Прогнозирана вредност зависне променљиве Y_0 и средње вредности зависне променљиве EY_0 у тачки x_0 је

$$\hat{Y}_0 = \hat{a}x_i + \hat{b}.$$

Користећи исте аргументе као до сада, можемо показати да

$$\frac{\hat{Y}_0 - EY_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nS_x^2}}} \sim t_{n-2}$$
$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nS_x^2}}} \sim t_{n-2}.$$

Сада можемо правити интервале предвиђања.