

---

## 0.1 Увод у курс

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

**Дефиниција 0.1.2.** *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 0.1.1.**

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

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

Регресиона функција је права линија ако и само ако случајни вектор  $(X, Y)^T$  има вишедимензионалну нормалну расподелу. Регресиону праву има смисла конструисати и када знамо да заједничка расподела није нормална. Тада је то права која од свих правих линија најбоље описује зависност између  $Y$  и  $X$  у смислу средњеквадратног одступања.

---

Регресиони модел се може представити у облику

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

где је  $\varepsilon$  случајна величина 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  $a$  и  $b$  одређују тако да се минимизира  $E(Y - (aX + b))^2$ .

Добија се да је

$$a = \frac{EXY - EXEY}{DX}$$
$$b = EY - aEX,$$

па се коефицијенти  $a$  и  $b$  могу оценити методом замене, односно

$$\hat{a} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\bar{S}_X^2} = \hat{\rho} \frac{\bar{S}_X}{\bar{S}_Y}$$
$$\hat{b} = \bar{Y} - \hat{a}\bar{X},$$

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

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

Да поновимо, наш главни 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?

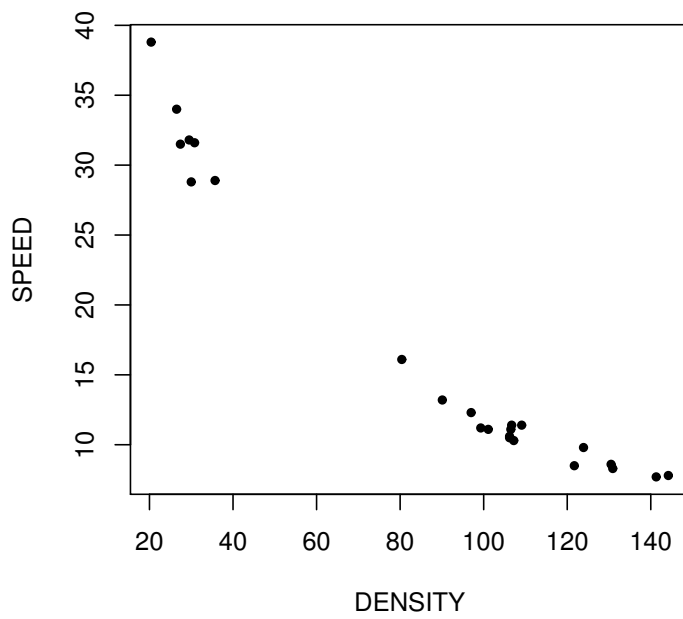
Како изгледа цео процес бирања модела демонстрираћемо на наредном примеру.

---

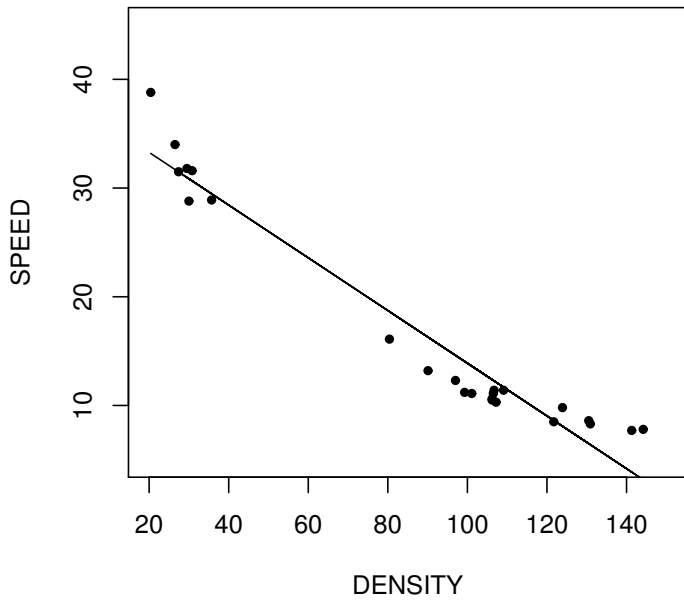
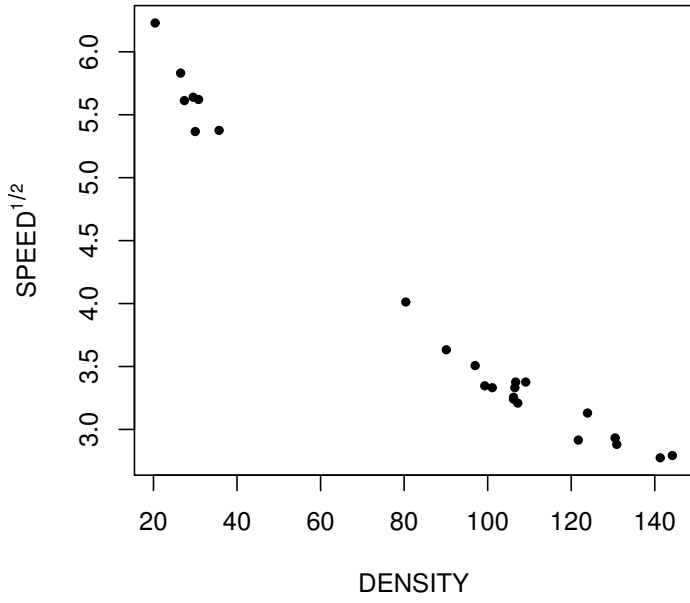
**Пример 0.1.1.** У циљу истраживања у којој мери број возила на путу утиче на брзину возила сакупљани су подаци о "густини" возила (број аутомобила у једној миљи) и просечној брзини аутомобила.

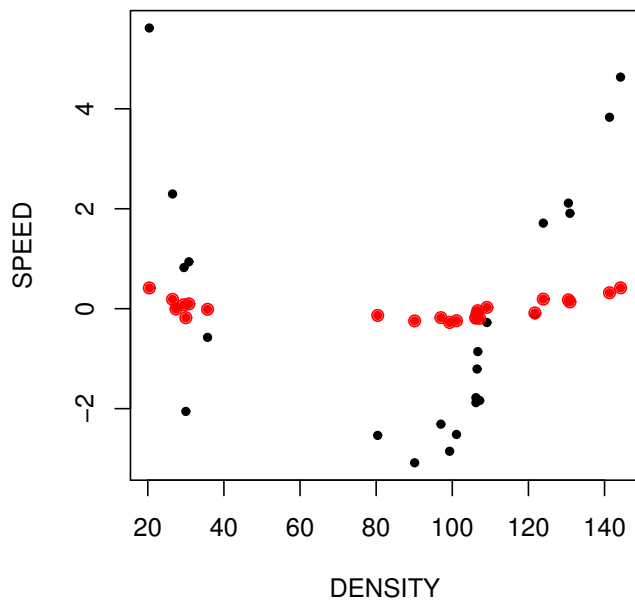
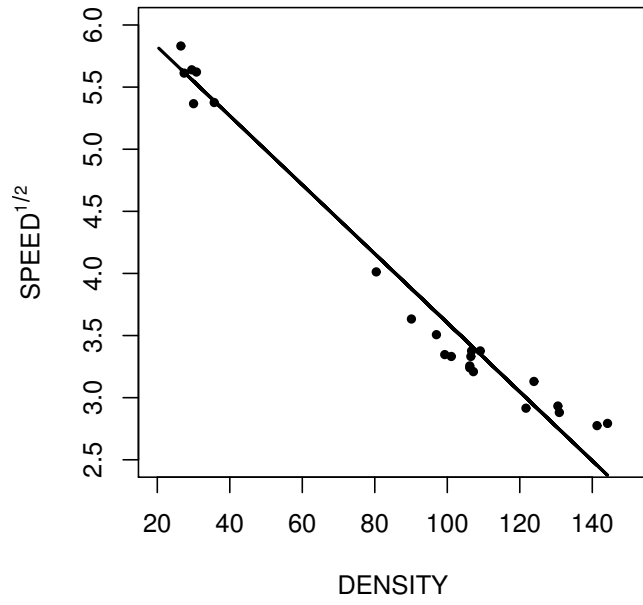
	DENSITY	SPEED
1	20.40	38.80
2	27.40	31.50
3	106.20	10.60
4	80.40	16.10
5	141.30	7.70
6	130.90	8.30
7	121.70	8.50
8	106.50	11.10
9	130.50	8.60
10	101.10	11.10
11	123.90	9.80
12	144.20	7.80
13	29.50	31.80
14	30.80	31.60
15	26.50	34.00
16	35.70	28.90
17	30.00	28.80
18	106.20	10.50
19	97.00	12.30
20	90.10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

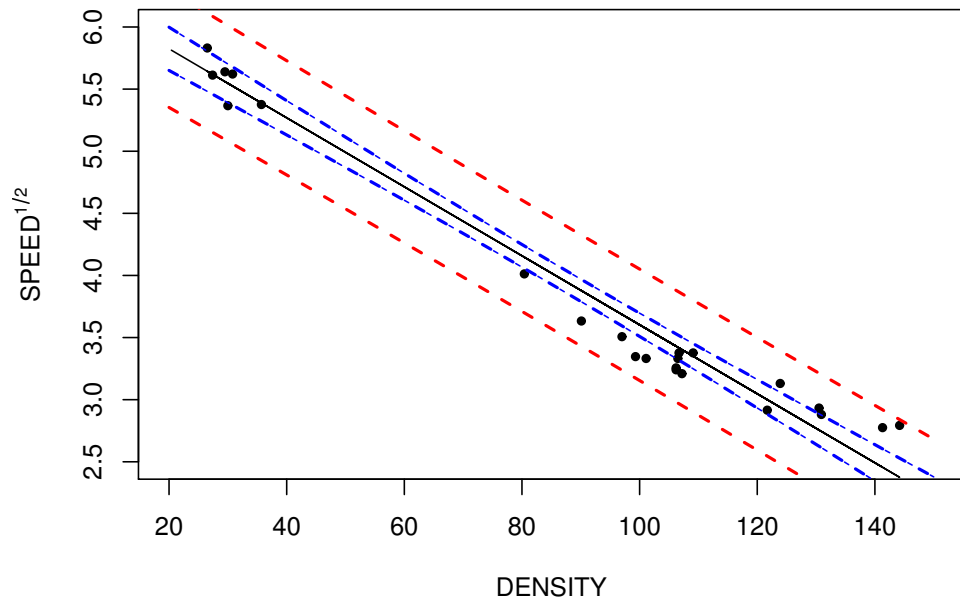
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	38.1295	1.2177	31.31	0.0000
E1.1\$DENSITY	-0.2425	0.0126	-19.22	0.0000



	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







---

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 саобраћаја. Јасно је да у модел морамо да укључимо и неки „шум“ ( $\varepsilon$ ) 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

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

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

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

Приметимо да полазни модел можемо написати у центрираном облику  $y_i = a(x_i - \bar{x}) + b + a\bar{x} + \varepsilon$ . 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  $\sum_{i=1}^n \varepsilon_i = 0$ .

У примеру 0.1.1 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 почетак,  $\sum_{i=1}^n \hat{\varepsilon}_i^2$ . Главни проблем са овом мером одступања је то што она зависи од јединице. Зато ћемо искористи сличну идеју као за увођење коефицијента корелације.



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

$SSR$  је одступање које је објашњено моделом. Зато уводимо *коэффициент 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 0.1.1 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.

## 0.2 Основни појмови из линеарне алгебре

**Дефиниција 0.2.1.** Векторски простор  $V$  је непразан скуп, затворен за сабирање и множење скаларом. Његови елементи називају се вектори.

**Дефиниција 0.2.2.** Вектори  $v_1, \dots, v_n$  су линеарно независни ако не постоји нетривијално решење једначине

$$x_1 v_1 + \dots + x_n v_n = 0.$$

**Дефиниција 0.2.3.** Вектори  $v_1$  и  $v_2$  су ортогонални ако је  $\langle v_1, v_2 \rangle = 0$ , где је са  $\langle \cdot, \cdot \rangle$  означен скаларни производ дефинисан на простору  $V$ .

Приметимо да из ортогоналности следи линеарна независност а да обрнуто не важи. На пример, посматрајмо векторе  $v_1 = (1, 0, 1)^T$  и  $v_2 = (0, 1, 1)^T$ .

**Дефиниција 0.2.4.** Квадратна матрица  $M$  је ортогонална ако је  $M^T M = I$

**Дефиниција 0.2.5.** Ранг  $n \times r$  матрице  $A$  у ознаци  $R(A)$  је максималан број линеарно независних колона(врста).

---

Из саме дефиниције следи следећи низ једнакости

$$R(A^T A) = R(AA^T) = R(A) = R(A^T).$$

**Дефиниција 0.2.6.** Ако је за неко  $\lambda \in R$ ,  $Ax = \lambda x$  онда се  $\lambda$  назива сопствена вредност матрице  $A$ , а  $x$  сопствени вектор.

Познато је да је за квадратну матрицу  $A$  је  $\lambda$  решење једначине  $\det(A - \lambda I) = 0$  и тада је  $\det A = \prod \lambda_i$ .

**Дефиниција 0.2.7.** Траг квадратне матрице  $A = [a_{ij}]$  представља збир елемената на дијагонали, односно

$$\text{tr} A = \sum_i a_{ii}.$$

За траг матрице  $A$  важе следеће једнакости (под претпоставком да је множење дефинисано):

1.  $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$ ;
2. Ако је  $A$   $n \times n$  матрица и  $P$  несингуларна матрица, онда је  $\text{tr}(P^{-1}AP) = \text{tr}(A)$ ;
3. Ако је  $A$   $n \times n$  матрица и  $M$  ортогонална матрица, онда је  $\text{tr}(M^T AM) = \text{tr}(A)$ .
4.  $\text{tr}(ABC) = \text{tr}(BCA)$

**Дефиниција 0.2.8.** Квадратна матрица  $A$  је симетрична ако је  $A = A^T$ .

Особине симетричне  $n \times n$  матрице  $A$ :

1. Постоји ортогонална матрица  $C = (c_1, \dots, c_n)$  и дијагонална матрица  $\Lambda = (\lambda_1, \dots, \lambda_n)$  (сопствене вредности матрице) таква да је  $A = C\Lambda C^T$  (спектрална декомпозиција матрице  $A$ ). Тада је  $A = \sum_{i=1}^n \lambda_i c_i c_i^T$ .
2.  $R(A)$  је број сопствених вредности различитих од нуле;
3.  $\text{tr}(A) = \sum_{i=1}^n \lambda_i$ ;
4. Ако је  $A$  несингуларна матрица онда је  $\text{tr}(A^{-1}) = \sum_{i=1}^n \lambda_i^{-1}$ ;

---

5. Постоји ортогонална трансформација  $y = M^T x$  тако да је

$$x^T A x = \sum \lambda_i y_i^2;$$

6.  $R(A + B) \leq R(A) + R(B)$

У случају несиметричне  $n \times p$  матрице  $A$  матрица ранга  $r$  постоји декомпозиција  $A = ULV$  где је  $U^T U = I = V^T V = I$  и  $L$  је несингуларна матрица ранга  $r$ .

**Дефиниција 0.2.9.** Нека је  $A$  симетрична матрица. Тада је са  $Q(x) = x^T A x$  дефинисана једна квадратна форма вектора  $x$ . Кажемо да је  $Q$  позитивно (негативно) дефинитна ако за свако  $x > 0$ ,  $Q(x) > 0$  ( $Q(x) < 0$ ). Ако се допушта једнакост онда је позитивно семи-дефинитна (негативно семи-дефинитна).

Може се показати да ако је  $A_{p \times p}$  позитивно дефинитна и  $B_{k \times p}$  матрица ранга  $k \leq p$  онда је  $BAB^T$  позитивно дефинитна.

Важи и следеће: Симетрична матрица  $A$  је позитивно дефинитна ако постоји несингуларна матрица  $P$  таква да је  $A = P^T P$

Важи и следеће

$$\max_{x \neq 0} \frac{x^T A x}{x^T x} = \lambda_{\max}, \quad (1)$$

$$\min_{x \neq 0} \frac{x^T A x}{x^T x} = \lambda_{\min}. \quad (2)$$

**Дефиниција 0.2.10.** За матрицу  $P$  за коју је  $P^2 = P$  кажемо да је идемпотента. Уколико је и симетрична онда се назива матрицом пројекције или пројектором.

Особине пројектора:

1.  $tr(P) = R(P)$ ;
2.  $P$  је позитивно семи дефинитна;
3. Нека су  $P_1$  и  $P_2$  пројектори. Ако је  $P_1 - P_2$  позитивно семи-дефинитна онда је и пројектор, као и  $P_1 P_2 = P_2 P_1 = P_2$

---

Доказ треће особине:

Из позитивне семидефинитности  $P_1 - P_2$  закључујемо да је

$$((I - P_1)y)^T(P_1 - P_2)((I - P_1)y) \geq 0.$$

Даље је  $y^T(I - P_1)^T P_1(I - P_1)y = 0$ , па је

$$((I - P_1)y)^T P_2((I - P_1)y) \leq 0.$$

Узимајући у обзир позитивну семидефинитност  $P_2$  је

$$((I - P_1)y)^T P_2((I - P_1)y) = 0,$$

односно

$$((I - P_1))^T P_2((I - P_1)) = 0 = (I - P_1)^T P_2 P_2((I - P_1)).$$

Одавде је

$$P_2(I - P_1) = 0.$$

**Дефиниција 0.2.11.** Матрица  $A$  максималног ранга има инверз  $A^{-1}$ . Уколико матрица није максималног ранга онда постоји уопштени инверз  $A^-$  за који важи  $AA^-A = A$ . Овај инверз не мора бити јединствен.

## 0.2.1 Матрично диференцирање

**Дефиниција 0.2.12.** Нека је  $X$   $n \times p$  матрица и  $f$  скаларна функција. Тада је матрично диференцирање дефинисано са

$$\frac{\partial f(X)}{\partial X} := \left( \frac{\partial f(X)}{\partial x_{ij}} \right)$$

1.  $\frac{\partial a^T x}{\partial x} = a$ ;
2.  $\frac{\partial x^T x}{\partial x} = 2x$ ;
3.  $\frac{\partial x^T A x}{\partial x} = (A + A^T)x$ ;
4.  $\frac{\partial x^T A y}{\partial x} = A y$ ;

---

## 0.3 Неке важне вишедимензионалне расподеле

### 0.3.1 Нормална расподела

**Дефиниција 0.3.1.** *Случајни вектор  $X$  има  $n$ -димензиона нормалну расподелу  $N_n(\mu, \Sigma)$  уколико је његова функција густине*

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}, \quad x \in R^n$$

где је  $\Sigma$  симетрична, позитивно дефинитна коваријациона матрица а са  $|\Sigma|$  је означена њена детерминанта.

Вишедимензионална нормална расподела има следеће лепе особине:

- Уколико  $X$  има  $N(\mu, \Sigma)$  расподелу онда  $AX+b$  има  $N(A\mu+b, A\Sigma A^T)$ ;
- Ако случајни вектор  $Z = (X^T, Y^T)^T$  има  $N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{pmatrix}\right)$  расподелу онда су маргиналне расподеле за  $X$  и  $Y$  редом  $N(\mu_x, \Sigma_x)$  и  $N(\mu_y, \Sigma_y)$ , а условне расподеле  $X|y \sim N(\mu_x + \Sigma_{xy} \Sigma_y^{-1} (y - \mu_y), \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{xy}^T)$   
 $Y|x \sim N(\mu_y + \Sigma_{xy}^T \Sigma_x^{-1} (x - \mu_x), \Sigma_y - \Sigma_{xy}^T \Sigma_x^{-1} \Sigma_{xy})$ ;
- Момент генераторна функција случајног вектора  $X$  са  $N(\mu, \Sigma)$  расподелом је  $M_X(t) = E(e^{t^T x}) = e^{\mu^T t + \frac{1}{2} t^T \Sigma t}$ ,  $t \in R$ ;
- $X$  се може представити у облику  $X = AZ + \mu$  где је  $AA^T = \Sigma$  а  $Z$  има стандардну вишедимензионалну нормалну расподелу.
- Нека  $X = (X_1, \dots, X_n)^T$  има вишедимензиону нормалну расподелу. Тада су компоненте вектора независне акко су некорелисане (коваријациона матрица је дијагонална матрица).
- $X$  има вишедимензиону нормалну расподелу акко за сваки вектор  $a$  (различит од нуле)  $a^T X$  има једnodимензионалну нормалну расподелу.

---

### 0.3.2 $\chi^2$ расподела

**Дефиниција 0.3.2.** Нека су  $X_1, \dots, X_k$  независне случајне величине са  $N(\theta_1, 1), \dots, N(\theta_k, 1)$  расподелма, редом. Тада случајна величина

$$Y = \sum_{j=1}^k X_j^2 \quad (3)$$

има  $\chi_k^2(\mu)$  расподелу, где је параметар положаја  $\mu = \sum_{j=1}^k \theta_j^2$ . Уколико је  $\mu = 0$  параметар положаја ћемо изоставити у нотацији.

**Теорема 0.3.1.** Случајна величина  $Y$  дефинисана са (3) се може представити у облику збира две независне случајне величине од којих једна има  $\chi_1^2(\mu)$  а друга  $\chi_{k-1}^2$  расподелу.

Приметимо да заправо ова теорема оправдава дефиницију  $\chi^2$  расподеле јер сугерише да расподела од  $Y$  зависи само од степени слободе и  $\mu$ .

*Доказ.* Нека је  $B = [b_{ij}]$  ортогонала матрица тако да је  $b_{1j} = \theta_j \mu^{-\frac{1}{2}}$ , за  $j = 1, 2, \dots, k$ . Нека је  $W = BX$ . Тада  $W$  има  $N(B\theta, BIB^T)$ , односно  $N(B\theta, I)$  јер је  $B$  ортогонална матрица. Одавде је јасно да су компоненте вектора  $W$  међусобно независне. Приметимо да је  $EW_1 = \sum_{j=1}^k b_{1j}\theta_j = \mu^{\frac{1}{2}}$ , и да је  $EW_i = \sum_{j=1}^k b_{ij}\theta_j = 0$ . Друга једнакост важи пошто је матрица  $B$  ортогонална.

$$Y = X^T X = W_1^2 + \sum_{i=2}^k W_i^2$$

Из ове репрезентације јасно следи тврђење теореме. □

### 0.3.3 ФишEROVA расподела

Нека  $X \sim \chi_{n_1}^2(\mu)$  и  $Y \sim \chi_{n_2}^2$  и независне су. Тада

$$\frac{\frac{X}{n_1}}{\frac{Y}{n_2}}$$

има Фишерову  $F_{n_1, n_2}(\mu)$  расподелу.

---

### 0.3.4 Студентова расподела

Нека  $X$  има нормалну  $N(\theta, 1)$  расподелу и  $Y$  има  $\chi_m^2$  расподелу и независне су. Тада

$$\frac{X}{\sqrt{\frac{Y}{m}}}$$

има Студентову  $t_m(\theta)$  расподелу, где је  $\theta$  параметар положаја.

## 0.4 Расподела квадратне форме

**Теорема 0.4.1** (Кохран). Нека су  $X_1, \dots, X_n$  независне  $\mathcal{N}(0, \sigma^2)$  случајне величине и нека је

$$\sum_{i=1}^n X_i^2 = \sum_{j=1}^k Q_j,$$

где је  $Q_j$  квадратна форма дефинисана са  $Q_j = X^T A_j X$ , за  $j = 1, 2, \dots, k$ , при чему је  $R(A_j) = r_j$ . Тада је  $\sum_{j=1}^k r_j = n$  ако и само ако

1.  $Q_1, \dots, Q_k$  су независне случајне величине и
2.  $Q_j/\sigma^2$  има  $\chi_{r_j}^2$  расподелу.

Може се формулисати и општије тврђење:

**Теорема 0.4.2** (Кохран). Нека случајни вектор  $X$  има  $\mathcal{N}(\theta, \sigma^2 I)$  расподелу и нека је

$$\sum_{i=1}^n X_i^2 = \sum_{j=1}^k Q_j,$$

где је  $Q_j$  квадратна форма дефинисана са  $Q_j = X^T A_j X$ , за  $j = 1, 2, \dots, k$ , при чему је  $R(A_j) = r_j$ . Тада је  $\sum_{j=1}^k r_j = n$  и  $\sum_{j=1}^k \mu_j = \theta^T \theta$  ако и само ако

1.  $Q_1, \dots, Q_k$  су независне случајне величине и
2.  $Q_j/\sigma^2$  има  $\chi_{r_j}^2(\mu_j)$  расподелу, при чему је  $\mu_j = \theta^T A_j \theta$ .

Доказаћемо само прву теорему. Друга се доказује аналогно. Пре него што се упустимо у доказ наводимо једну њену опште познату последицу.

**Последица 0.4.1.** Узорачка средина  $\bar{X}$  и поправљена узорачка дисперзија  $S^2$  су независне случајне величине, и  $(n-1)S^2/\sigma^2$  има  $\chi_{n-1}^2$  расподелу.

Без умањења општости можемо претпоставити да је  $\sigma^2 = 1$ .

$$(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2.$$

Одавде је

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X})^2 = X^T \left( I - \frac{J}{n} \right) X + X^T \frac{J}{n} X,$$

где је  $J$   $n \times n$  матрица чији су сви елементи јединице. Приметимо да лева страна једнакости има  $\chi_n^2$  расподелу. Даље,  $R(I - \frac{J}{n}) = n-1$  јер је, с једне стране  $R(I - \frac{J}{n}) \geq R(I) - R(\frac{J}{n})$ , а с друге, из једнакости  $(I - \frac{J}{n})1 = 0$  закључујемо да је  $R(I - \frac{J}{n}) \leq n-1$ . Како је  $R(\frac{J}{n}) = 1$  следи тврђење.

У доказу Кохранове теореме ћемо користити следећу лему.

**Лема 0.4.1.** Нека су  $x_1, \dots, x_n$  реални бројеви. Претпоставимо да се сума  $\sum_{i=1}^n x_i^2$  може представити као збир  $k$  квадратних форми  $\sum_{j=1}^k Q_j$  где је  $Q_i = x^T A_i x$  и  $R(A_i) = r_i$  за  $i = 1, \dots, k$ . Ако је  $\sum_{i=1}^k r_i = n$  тада постоји ортогонална матрица  $M$ , таква да за  $x = My$  важи

$$\begin{aligned} Q_1 &= y_1^2 + \dots + y_{r_1}^2 \\ Q_2 &= y_{r_1+1}^2 + \dots + y_{r_1+r_2}^2 \\ &\dots \\ Q_k &= y_{n-r_k+1}^2 + \dots + y_n^2 \end{aligned}$$

*Доказ.* Довољно је показати лему за  $k = 2$ . Тада је

$$Q = x^T x = x^T A_1 x + x^T A_2 x.$$

Постоји ортогонална матрица  $M$  таква да је  $M^T A_1 M = D_1$  где је  $D_1$  дијагонална матрица. Без умањења општости можемо претпоставити да су од сопствене вредности поређане тако да су  $\lambda_1, \dots, \lambda_{r_1}$  различите од нуле а остале нула. Нека је  $x = My$ . Тада је

$$x^T x = y^T M^T M y = y^T y.$$

Даље је

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^{r_1} \lambda_i y_i^2 + y^T M^T A_2 M y.$$



Одавде је

$$\sum_{i=1}^{r_1} (1 - \lambda_i) y_i^2 + \sum_{i=r_1+1}^n y_i^2 = y^T M^T A_2 M y.$$

Како је  $R(A_2) = n - r_1$  закључујемо да је  $\lambda_1 = \dots = \lambda_{r_1} = 1$  одакле следи тврђење.  $\square$

Важно је приметити да све квадратне форме које учествују у репрезентацији садрже различите  $y_i$ -ове. Независност  $Q_1, \dots, Q_k$  у тврђењу Кохранове теореме је последица овога. Да бисмо доказали Кохранову теорему потребно је још да приметимо да када применимо ортогоналну трансформацију на случајан вектор са нормалном расподелом са независним компонентама добијамо опет случајан вектор са нормалном расподелом са независним компонентама. Одавде се добијају расподеле одговарајућих квадратних форми.

Следећу теорему наводимо без доказа.

**Теорема 0.4.3.** Нека случајни вектор  $X$  има  $N(\theta, \sigma^2 I)$  расподелу и нека је  $Q_1 = X^T A_1 X$  и  $Q_2 = X^T A_2 X$ , где су  $A_1$  и  $A_2$  две симетричне матрице. Тада су  $Q_1$  и  $Q_2$  независне ако и само ако је  $A_1 A_2 = 0$ .

## 0.5 Задаци

**0.1.** Нека су  $X_1, \dots, X_n$  независне и једнако расподељене случајне величине тако да је  $EX_1 = 0$  и  $DX_1 = \sigma^2 < \infty$ . Нека је  $Y_i = X_i - \bar{X}$ , за  $i = 1, 2, \dots, n$ . Наћи коваријациону матрицу случајног вектора  $Y = (Y_1, \dots, Y_n)^T$ .

**0.2.** Нека су  $X_1, \dots, X_{n_1}$  независне са  $N(\mu_1, \sigma_1^2)$  и  $Y_1, \dots, Y_{n_2}$  независне са  $N(\mu_2, \sigma_2^2)$  расподелом. Наћи расподелу следећих статистика:

а)  $\frac{(\bar{X} - \bar{Y} - \delta)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ , где је  $\delta$  произвољна константа.

б)  $\frac{n_1(\bar{X} - \mu_1)^2}{\sigma_1^2} + \frac{n_2(\bar{Y} - \mu_2)^2}{\sigma_2^2}$

**0.3.** Нека су  $X_1, \dots, X_{n_1}$  независне са  $N(\mu_1, \sigma^2)$  и  $Y_1, \dots, Y_{n_2}$  независне са  $N(\mu_2, \sigma^2)$  расподелом. Наћи расподелу следећих статистика:

а)  $\frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{\sigma^2}$

---

**0.4.** Нека је  $(X_1, Y_1), \dots, (X_n, Y_n)$  прост случајан узорак из двовимензионалне нормалне расподеле са параметрима  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ . Одредити константу  $C$  тако да статистика

$$T = C \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sum_{i=1}^n (X_i - Y_i - \bar{X} + \bar{Y})^2}}$$

има Студентову  $t_m(\theta)$ . Изразити  $m$  и  $\theta$  у функцији од параметара расподеле и константе  $\delta$ .

**0.5.** Доказати теорему 0.4.3.

# Поглавље 1

## Линеарни модели

### 1.1 Проста линеарна регресија

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

Нека је

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

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

1. центрираност  $E\varepsilon_i = 0$ ,  $i = 1, 2, \dots, n$
2. некорелисаност  $E\varepsilon_i\varepsilon_j = 0$ ,  $i \neq j$ ;
3. хомоскедастичност  $D\varepsilon_i = \sigma^2 > 0$ ;
4.  $x_i$  и  $\varepsilon_j$  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.

$$D(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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

Дакле, оцене параметара модела су непристрасне и постојане. Видимо да у изразима за дисперзије фигурише непознати параметар  $\sigma^2$ . Зато морамо и за њега наћи одговарајућу оцену. Природно је да та оцена буде у вези са узорачком дисперзијом резидуала модела.

$$E(SSE) = E(SST) - E(SSR)$$

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

Одавде добијамо да је непристрасна оцена за  $\sigma^2$

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

Уколико се у модел уведе додатна претпоставка да је шум Гаусов, односно да  $\varepsilon_i$  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

$$\hat{Y}_0 = \hat{a}x_0 + \hat{b}.$$

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. Имајући у виду које стожерне величине користимо за прављење интервала поверења видимо да ће они бити најужи у "центру", односно кад је  $x = \bar{x}$ .

## 1.2 Вишеструка линеарна регресија

Међутим, неретко се дешава да треба узети у обзир неколико предиктора.

Претпоставимо да имамо  $p$  предиктора  $x_1, x_2, \dots, x_p$ . Тада се линеарни модел може записати у облику

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1.1)$$

или у облику

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1.2)$$

при чему је  $n$  обим узорка који имамо на располагању.

Најчешће се овај модел записује у векторском облику.

### 1.3 Матрични запис линеарног модела

Модел (1.1) се може записати у облику

$$Y = X\beta + \varepsilon,$$

где су

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$X$  се назива предиктор (независна променљива), а  $Y$  регресанд (зависна променљива). У случају модела (1.2) је

$$X = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

Матрица  $X$  се назива *дизајн матрица*.

Случајност модела потиче од случајних грешки  $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$  за које

се претпоставља центрираност, хомоскедастичност и некорелисаност. Дакле претпостављамо да важи:

1.  $E(\varepsilon) = 0$ ;
2.  $D(\varepsilon) = \sigma^2 I$ ;
3.  $X$  и  $\varepsilon$  су независни случајни вектори.

**Пример 1.3.1.** У случају прости линеарне регресије је

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \beta = \begin{pmatrix} b \\ a \end{pmatrix}.$$

**Пример 1.3.2.** Желимо да видимо какав је утицај неког фактора на средњу вредност мереног обележја  $Y$ . За почетак, претпоставимо да имамо само један фактор који има две категорије. Нека је  $x_i = 1$  уколико меримо вредност обележја  $Y$  на узорку из прве категорије. У супротном нека је  $x_i = 0$ . Даље, претпоставимо да имамо тачно  $k$  елемената узорка из прве категорије. Тада модел можемо приказати у облику.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Приметимо да је ово заправо ANOVA модел.

**Пример 1.3.3.** Понекад, на основу графичког приказа података, имамо разлога да верујемо да би било боље, уместо линеарне зависности, претпоставити квадратну зависност.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

Овај модел можемо написати на сличан начин у векторском облику. Сада је

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Најприроднији начин да се оцене коефицијенти модела  $\beta$  је метод најмањих квадрата, односно

$$\hat{\beta} := \arg \min_{b \in \mathbb{R}^{p+1}} \|Y - Xb\|^2.$$

Уколико је  $R(X) = p + 1$  онда је  $X^T X$  инвертибилна и добија се да је

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Оцењена вредност је тада

$$\hat{Y} = X(X^T X)^{-1} X^T Y = HY$$

Матрица  $H$  се назива hat matrix. Приметимо да је матрица  $H$  пројектор, тако да  $\hat{Y}$  представља ортогоналну пројекцију вектора  $Y$  на раван генерисану са  $X$ . Резидуали модела се могу приказати у облику

$$e = Y - \hat{Y} = (I - H)Y.$$

Одавде је

$$\begin{aligned} E(e) &= (I - H)E(Y) = 0 \\ Cov(e) &= \sigma^2(I - H)(I - H)^T = \sigma^2(I - H) \end{aligned}$$

Посматрајмо суму квадрата одступања од модела

$$SSE = \sum_{i=1}^n e_i^2 = e^T e = Y^T (I - H)Y = Y^T Y - Y^T H Y.$$

Означимо са  $M = (I - H) = [m_{ij}]$ . Њено очекивање је

$$\begin{aligned} E(SSE) &= E\left(\sum_{ij} m_{ij} Y_i Y_j\right) = \sum_{ij} m_{ij} E(Y_i Y_j) = \sum_{ij} m_{ij} E(\varepsilon_i \varepsilon_j) \\ &= \sum_i m_{ii} E(\varepsilon_i^2) = \sigma^2 \sum_i m_{ii} \\ &= \sigma^2 tr(M) = \sigma^2 (tr(I) - tr(H)) = \sigma^2 (n - tr((X^T X)(X^T X)^{-1})) \\ &= \sigma^2 (n - tr(I_{p+1})) = \sigma^2 (n - p - 1). \end{aligned}$$

Одавде закључујемо да је непристрасна оцена за  $\sigma^2$  управо  $(\sum_{i=1}^n e_i^2)/(n - p - 1)$ .

Резидуали и оцењене вредности су некорелисани  $\sum (Y_i - \hat{Y})(\hat{Y} - \bar{Y}) = 0$  па је

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y})^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ SSTO &= SSE + SSR \end{aligned} \tag{1.3}$$



$SSE$  је необјашњено одступање (потиче од модела),  $SSR$  објашњено одступање а  $SSTO$  укупно одступање. У матричном облику се одступања могу представити на следећи начин:

$$\begin{aligned} SSTO &= Y^T \left( I - \frac{J}{n} \right) Y, \\ SSE &= Y^T (I - H) Y \\ SSR &= Y^T \left( H - \frac{J}{n} \right) Y, \end{aligned}$$

где је матрица  $J$   $n \times n$  матрица са свим јединицама. Приликом доказа разлагања (1.3) коришћена је следећа лема.

**Лема 1.3.1.** Нека су  $\{e_i\}$  резидуали линеарног модела. Тада је

$$\sum_{i=1}^n e_i = 0.$$

*Прооф.* Претпоставимо да је  $\sum_{i=1}^n e_i = cn \neq 0$ . Тада је

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (e_i - c + c)^2 = \sum_{i=1}^n (e_i - c)^2 + nc^2 + 2c \sum_{i=1}^n (e_i - c) \geq \sum_{i=1}^n (e_i - c)^2 \\ &= \sum_{i=1}^n (Y_i - (1, x_i^T) \hat{\beta} - c)^2. \end{aligned}$$

Одавде закључујемо да вектор  $\hat{\beta}$  не минимизира суму квадрата одступања што је супротно претпоставци, па  $c$  мора бити 0.

Из методолошких разлога приказаћемо још један начин да се ово покаже. Приметимо да су резидуали ортогонални на простор генерисан са  $X$ .

$$e^T X = Y^T (I - H) X = Y^T (X - H^T X) = \mathbf{0}$$

Како је први ред матрице  $X^T$  вектор јединица, следи тврђење леме.  $\square$

Из формуле (1.3) је природно да о квалитету модела говоримо на основу количине објашњених одступања моделом односно да посматрамо коефицијент детерминације  $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$ . Видимо да је  $0 \leq R^2 \leq 1$ . У пракси је потребно да је  $R^2$  бар 0.62. Са  $R = \sqrt{R^2}$  је дефинисан вишеструки коефицијент корелације. Приметимо да кад имамо само један предиктор онда је  $R = |\rho_{xy}|$ .

Даље, важи да је  $R(I - \frac{J}{n}) = n - 1$ , као и  $R(I - H) = n - p - 1$  јер је

$$R(I - H) = \text{tr}(I - H) = n - p - 1.$$

С обзиром на то, да коефицијент детерминације увек расте са порастом броја предиктора понекад се уместо њега користи његова модификација која узима у обзир непристрасне оцене одговарајућих грешака, дефинисана са

$$R_A^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SSTO}{n-1}}.$$

У пракси је примећено да у случају малих узорака вредности  $R^2$  могу бити велике чак и кад модел није довољно квалитета. За разлику од  $R^2$ ,  $R_A^2$  може бити негативан.

$R(X)$  не мора бити једнак баш  $p + 1$ . На пример меримо температуру у цезијусима и фарехајтима, или бележимо број поена на предиспитним обавезама, на испиту, и укупан број поена. Може се десити и да је број променљивих већи или једнак од броја обсервација. Ако је  $p + 1 = n$  ради се о *сатурираном* моделу, док ако је  $p > n - 1$  кажемо да је модеј *супер-сатуриран*. У овом случају, коефицијенти модела се не могу јединствено одредити.

Најбоље је обратити пажњу у прелиминарној анализи и уклонити променљиве које ису неопходне.

Ако је  $X$  ортогонална матрица онда се оцена за  $\beta$  поклапа са вектором који би се добио када бисмо посматрали  $p + 1$  одговарајућих простих линеарних регресија.

Вратимо се сада особинама оцене коефицијената модела методом најмањих квадрата.

**Последица 1.3.1.** *Уколико  $\text{tr}((X^T X)^{-1}) \rightarrow 0$ , кад  $n \rightarrow \infty$   $\hat{\beta}$  је постојана оцена параметра  $\beta$ .*

Важна особина оцене методом најмањих квадрата је садржана у следећој теорему.

**Теорема 1.3.1.** *Нека је  $\hat{\beta}$  оцена методом најмањих квадрата и  $\tilde{\beta} = AY$  нека друга линеарна непристрасна оцена за  $\beta$ . Тада је матрица  $\text{Cov}(\tilde{\beta}|X) - \text{Cov}(\hat{\beta}|X)$  позитивно дефинитна.*

*Прооф.*

$$\begin{aligned} E(\hat{\beta}|X) &= (X^T X)^{-1} X^T E(Y|X) = (X^T X)^{-1} X^T X \beta = \beta \\ Cov(\hat{\beta}|X) &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2 \end{aligned}$$

Нека је  $B = (X^T X)^{-1} X^T$  и нека је  $A = B + C$ . Из услова непристрасности је  $E(\hat{\beta}|X) = AX\beta = \beta$ . Односно  $I = AX = (B + C)X$ . Како је  $BX = I$  закључујемо да је  $CX = 0$ .

Даље је

$$Cov(\tilde{\beta}|X) = AD(Y|X)A^T = (B + C)(B^T + C^T)\sigma^2 = Cov(\hat{\beta}|X) + BC^T + CB^T + CC^T.$$

Како је

$$\begin{aligned} CB^T &= CX(X^T X)^{-1} = 0 \\ BC^T &= (X^T X)^{-1} X^T C^T = (X^T X)^{-1} (CX)^T = 0 \end{aligned}$$

следи тврђење. □

Претпоставимо да је циљ да оценимо неку линеарну комбинацију параметара модела или више таквих линеарних комбинација, односно да оценимо  $l^T \beta$  у првом случају, односно  $(L^T \beta)$  у другом случају. Задржимо се за тренутак на првом случају. Други је аналоган. Претпоставимо да постоји непристрасна линеарна оцена  $(cY)$  оцена за  $l^T \beta$ . Тада се од свих таквих оцена оцена са најмањом дисперзијом назива најбоља линеарна непристрасна оцена и означава са BLUE (Best Linear unbiased estimate).

Слично као претходна теорема се показују и следеће теореме. Прва говори о оцени линеарне функције параметара модела  $(l^T \beta)$  а друга о оцени више линеарних функција параметара модела  $(L^T \beta)$ .

**Теорема 1.3.2** (Гаус-Марков).  $l^T \hat{\beta}$  је BLUE за  $l^T \beta$ .

**Теорема 1.3.3** (Гаус-Марков).  $L^T \hat{\beta}$  је BLUE за  $L^T \beta$ .

## 1.4 Статистичко закључивање у Гаусовом моделу

У овом поглављу претпоставићемо да грешке модела представљају Гаусов шум, односно да  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ .

### 1.4.1 Тестирање линеарне хипотезе

Желимо да тестирамо нулту хипотезу  $H_0 : C\beta = \gamma$ , где је  $C$   $m \times (p + 1)$  матрица. На пример, уколико желимо да тестирамо нулту хипотезу да нема утицаја слободног члана узећемо да је  $C = (1, 0, \dots, 0)$  и  $\gamma = 0$ . Или, уколико желимо да тестирамо нулту хипотезу да ниједан предиктор нема утицаја на  $Y$ ,  $C = (0, I_p)$  и  $\gamma = 0$ .

За тестирање наше нулте хипотезе користићемо тест количника веродостојности.

### 1.4.2 Метод максималне веродостојности

Претпоставимо да  $\varepsilon$  има  $N(0, \sigma^2 I_n)$  расподелу. Тада је функција веродостојности

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{(Y - \beta X)^T (Y - \beta X)}{2\sigma^2}}.$$

Одавде видимо да се оцена методом максималне веродостојности за  $\beta$  поклапа са оценом методом најмањих квадрата, односно  $\hat{\beta} = (X^T X)^{-1} X^T Y$ . Слично, оцена за  $\sigma^2$  је

$$\hat{\sigma}^2 = \frac{1}{n} (Y - \hat{\beta} X)^T (Y - \hat{\beta} X)$$

У претходном одељку видели смо да ова оцена није непристрасна оцена за  $\sigma^2$ .

Максимална вредност функције веродостојности је тада

$$L(\hat{\beta}, \hat{\sigma}^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{n}{2}}} e^{-\frac{1}{2n}} \sim \hat{\sigma}^{-n}$$

Како је  $\hat{\beta}$  линеарна трансформација случајног вектора са нормалном расподелом, закључујемо да

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2).$$

Да бисмо применили тест количника веродостојности неопходно је да нађемо оцену параметара уз услов  $C\beta = \gamma$ . Слично као у поставци проблема малопре, ово је еквивалентно са тражењем минимума  $S(\beta) = \frac{(Y - X\beta)^T (Y - X\beta)}{n}$  уз услов  $C\beta - \gamma = 0$  (зато што у услову не фигурише  $\sigma^2$ ). Лагранжова функција за решавање овог екстремалног проблема је

$$\mathcal{L}(\beta, a) = (Y - X\beta)^T (Y - X\beta) - a^T (C\beta - \gamma).$$

Систем нормалних једначина чије решење тражимо је

$$\begin{aligned}\frac{\partial}{\partial \beta} \mathcal{L}(\beta, a) &= -2X^T Y + 2X^T X \beta - C^T a = 0 \\ \frac{\partial}{\partial a} \mathcal{L}(\beta, a) &= \gamma - C \beta = 0.\end{aligned}$$

Добија се да је

$$\begin{aligned}\hat{\beta}_0 &= \hat{\beta} + (X X^T)^{-1} C^T [C (X X^T)^{-1} C^T]^{-1} (\gamma - C \hat{\beta}) \\ \hat{\sigma}_0^2 &= \frac{(Y - X \hat{\beta}_0)^T (Y - X \hat{\beta}_0)}{n} \\ &= \frac{1}{n} \left( (Y - X \hat{\beta})^T (Y - X \hat{\beta}) + (C \hat{\beta} - \gamma)^T [C (X X^T)^{-1} C^T]^{-1} (C \hat{\beta} - \gamma) \right)\end{aligned}$$

Максимална вредност функције веродостојности у овом случају је

$$L(\hat{\beta}_0, \hat{\sigma}_0^2) \sim \hat{\sigma}_0^{-n}$$

Претпоставимо да је  $H_1 : X \beta \neq \gamma$ . Тада је количник веродостојности

$$\begin{aligned}\lambda &= \frac{\max_{H_1} L(\beta, \sigma^2)}{\max_{H_0} L(\beta, \sigma^2)} = \frac{L(\hat{\beta}, \hat{\sigma}^2)}{L(\hat{\beta}_0, \hat{\sigma}_0^2)} = \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}} \\ &= \left( 1 + \frac{(C \hat{\beta} - \gamma)^T [C (X X^T)^{-1} C^T]^{-1} (C \hat{\beta} - \gamma)}{e^T e} \right)^{\frac{n}{2}}\end{aligned}$$

Критична област за тестирање је  $W = \{\lambda > c\}$ .

Претпоставка о нормалном моделу је јако важна за коришћење теста количника веродостојности јер се тест статистика може приказати у погодном облику као количник две независне статистике са  $\chi^2$  расподелама, односно у функцији од статистике која, под нултом хипотезом, има Фишерову расподелу.

Јасно је да  $\hat{\beta} = (X^T X)^{-1} X^T Y$  има нормалну  $N(\beta, \sigma^2 (X^T X)^{-1})$ . На основу тога имамо да

$$C \hat{\beta} - \gamma \sim N_m(C \beta - \gamma, \sigma^2 C (X^T X)^{-1} C^T).$$

Одавде, уколико важи  $H_0$  закључујемо да

$$Q = (C \hat{\beta} - \gamma)^T [C (X^T X)^{-1} C^T]^{-1} (C \hat{\beta} - \gamma) \sim \sigma^2 \chi_m^2.$$

Приметимо да се, уколико важи  $H_0$ ,  $Q$  може приказати у облику

$$\begin{aligned} Q &= (C\hat{\beta} - C\beta)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - C\beta) \\ &= (C(X^T X)^{-1} X^T \varepsilon)^T [C(X^T X)^{-1} C^T]^{-1} C(X^T X)^{-1} X^T \varepsilon \\ &= \varepsilon^T X(X^T X)^{-1} C^T [C(X^T X)^{-1} C^T]^{-1} C(X^T X)^{-1} X^T \varepsilon \\ &= \varepsilon^T P \varepsilon, \end{aligned}$$

где је  $P = X(X^T X)^{-1} C^T [C(X^T X)^{-1} C^T]^{-1} C(X^T X)^{-1} X^T$ .

Даље, применом Кохранове теореме добијамо да

$$e^T e = SSE = \varepsilon^T M \varepsilon \sim \sigma^2 \chi_{n-p-1}^2$$

Да бисмо нашли расподелу тест статистике потребно је још да покажемо да је  $PM = 0$  и применимо теорему 0.4.3. Из једнакости  $X^T M = 0$  добијамо  $PM = 0$  и закључујемо да су бројилац и имеилац у тест статистици независне случајне величине па

$$\frac{\frac{Q}{m}}{\frac{e^T e}{n-p-1}} \sim F_{m, n-p-1}. \quad (1.4)$$

Сада је могуће применити тест количника веродостојности.

**Пример 1.4.1.** Претпоставимо да тестирамо нулту хипотезу  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$  против комплементарне хипотезе. Тада је

$$L(\hat{\beta}_0, \hat{\sigma}_0^2) \sim \left( \frac{SSTO}{n} \right)^{-\frac{n}{2}}.$$

Одавде, је

$$\lambda = \left( \frac{SSTO}{SSE} \right)^{\frac{n}{2}} = \left( 1 + \frac{SSR}{SSE} \right)^{\frac{n}{2}}.$$

Применом Кохранове теореме, или теореме (0.4.3) добија се да

$$\frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F_{p, n-p-1}. \quad (1.5)$$

**Пример 1.4.2.** У случају да желимо да тестирамо хипотезу да је  $\beta_k = 0$  већ смо видели како да одаберемо матрицу  $C$ . Тада

$$\frac{Q}{\frac{e^T e^T}{n-p-1}} \sim F_{1, n-p-1}.$$

У овом посебном случају овај тест са Студентовом тест статистиком коју ћемо представити у наредном поглављу.

**Пример 1.4.3.** Посматра се зависност GPA чије вредности се налазе у интервалу  $[0,4]$  од вербалног дела SAT-а (SATV) и математичког дела (SATM). И на једном и на другом тесту могуће је остварити од 200 до 800 поена. Подаци су приказани у следећој табели

GPA	3.95	3.84	3.68	3.59	3.57	3.49	3.47	3.40	3.08
SATV	740	760	660	760	760	660	710	710	570
SATM	790	710	750	740	700	670	730	790	760

```
summary(lm(GPA~SATV+SATM))
```

```
Call: lm(formula = GPA~SATV + SATM)
```

```
Residuals:      Min       1Q   Median       3Q      Max
              -0.2055  -0.1368  -0.1090   0.1265   0.2585
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(> |t|)
(Intercept)  1.2220434   1.6127365   0.758   0.4773
SATV         0.0028666   0.0011339   2.528   0.0448 *
SATM         0.0004406   0.0017987   0.245   0.8146
```

```
Signif. codes:
```

```
0 ***      0.001 **      0.01 *      0.05 .      0.1 1
```

```
Residual standard error: 0.2038 on 6 degrees of freedom
```

```
Multiple R-squared: 0.5158,
```

```
Adjusted R-squared: 0.3544
```

```
F-statistic: 3.196 on 2 and 6 DF, p-value: 0.1135
```

У R-у је тестирање нулте хипотезе да ли су сви коефицијенти модела уз предикторе нула стандардна процедура и део уграђене функције *lm*. Наиме, вредност статистике (1.5) је 3.196. Како статистика под нултом хипотезом има  $F_{2,6}$  расподелу, добијамо да је одговарајућа *p*-вредност теста 0.1135 и самим тим не одбацујемо нулту хипотезу. Закључак је да резултати SATM-а не утичу на GPA.

### 1.4.3 Интервал поверења за $\beta$

Из особина нормалне расподеле и Кохранове теореме, за  $\hat{\sigma} = \frac{1}{n-p-1}e^T e$  је

$$\frac{(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)}{\sigma^2} \sim \chi_{p+1}^2 \quad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

и независне су. Одавде је

$$\frac{(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)}{(p+1)\hat{\sigma}^2} \sim F_{p+1, n-p-1}.$$

Зато интервали поверења представљају елипсоиде у више димензија.

Означимо са  $V_k = [(X^T X)^{-1}]_{kk}$ ,  $k = 1..p+1$ . Како су  $Y - \hat{Y}$  и  $\hat{\beta}$  некорелисани (у односу на  $X$ ) онда је и  $\hat{\sigma}^2$  (као функција од резидуала) некорелисано са  $\hat{\beta}$  а самим тим и независно (због нормалности). Даље је  $D(\hat{\beta}_k) = \sigma^2 V_k$ . Због непристрасности оцене и нормалности је

$$\frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{V_k}} \sim N(0, 1).$$

Како

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

онда

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{V_k}} \sim t_{n-p-1}.$$

Следећа лема, између осталог, је важна за коректно тумачење резултата тестирања.

**Лема 1.4.1** (Бонферонијеве неједнакости). *Нека су  $A_1, \dots, A_k$  догађаји дефинисани на истом простору вероватноће. Тада је*

$$1 - P\left(\bigcap_{j=1}^k A_j\right) \leq \sum_{j=1}^k P(A_j^c)$$

Претпоставимо да догађај  $A_j$  одговара  $j$ -тој хипотези коју тестирамо. Из ове леме закључујемо да уколико желимо да ограничимо вероватноћу грешке прве врсте за неколико хипотеза које одједном тестирамо, довољно је да ограничимо вероватноћу грешке прве врсте, за сваку појединачну хипотезу, са  $\frac{\alpha}{k}$ .

**Пример 1.4.4.** *Желимо да направимо интервале поверења за линеарне комбинације коефицијената модела  $a_1^T \beta$ ,  $a_2^T \beta, \dots, a_l^T \beta$ . Један начин је да формирамо одговарајуће елипсоиде користећи статистику (1.4). С друге стране,  $a_i^T (\hat{\beta} - \beta) \sim N(0, a_i^T (X^T X)^{-1} a_i \sigma^2)$ , односно*

$$\frac{a_i^T (\hat{\beta} - \beta)}{\sqrt{a_i^T (X^T X)^{-1} a_i \hat{\sigma}^2}} \sim t_{n-p-1}$$



Сада можемо направити одговарајуће интервале поверења као у случају прости регресије. Такође, можемо извршити и тестирање  $H_0 : a_i^T \beta = A_i$  коришћењем да, уколико важи  $H_0$ ,

$$\frac{a_i^T \hat{\beta} - A_i}{\sqrt{a_i^T (X^T X)^{-1} a_i \hat{\sigma}^2}} \sim t_{n-p-1},$$

при чему ћемо критичну област правити за ниво поверења  $\frac{\alpha}{l}$ . У примеру 1.4.3, за ниво поверења  $\alpha = 0.1$ , сматраћемо да су значајни коефицијенти код којих је добијена  $p$ -вредност мања од 0.5, односно само уз SAT.

Треба бити опрезан приликом коришћења Бонференијевој корекције. Када имамо пуно хипотеза које желимо одједном да тестирамо Бонферонијева граница, коју при томе користимо, је сувуше груба, па је боље користити елипсоиде поверења.

#### 1.4.4 Интервали предвиђања

Нека је  $x_0 = (1, x_{01}, \dots, x_{0p})^T$  "тачка" у којој вршимо предвиђање. Тада је  $\hat{Y}_0 = x_0^T \hat{\beta}$ . Одавде је

$$\begin{aligned} E(\hat{Y}_0 | X) &= x_0^T E(\hat{\beta} | X) = x_0^T \beta \\ D(\hat{Y}_0 | X) &= x_0^T (X^T X)^{-1} x_0 \sigma^2. \end{aligned}$$

Из особина нормалне расподеле закључујемо да

$$\hat{Y}_0 \sim N(x_0^T \beta, x_0^T (X^T X)^{-1} x_0 \sigma^2),$$

па резонувањем као у претходном поглављу закључујемо да

$$\frac{\hat{Y}_0 - x_0^T \beta}{\sqrt{x_0^T (X^T X)^{-1} x_0 \hat{\sigma}^2}} \sim t_{n-p-1}.$$

Сада се лако могу конструисати интервали поверења за очекивану предвиђену вредност у тачки  $x_0$ .

Нека  $Y_0$  вредност у  $x_0$ . Тада је  $Y_0 = x_0^T \beta + \varepsilon_0$ , где је  $\varepsilon_0$  независно од  $Y_1, \dots, Y_n$ . Одавде је

$$\begin{aligned} E(\hat{Y}_0 - Y_0) &= x_0^T \beta - x_0^T \beta = 0 \\ D(\hat{Y}_0 - Y_0) &= D(\hat{Y}_0) + D(Y_0) = \sigma^2(1 + x_0^T (X^T X)^{-1} x_0). \end{aligned}$$

Одавде закључујемо да

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\sigma^2(1 + x_0^T(X^T X)^{-1}x_0)}} \sim N(0, 1),$$

одакле је

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2(1 + x_0^T(X^T X)^{-1}x_0)}} \sim t_{n-1-p}.$$

### 1.4.5 Категорички предиктори

У овом одељку размотрићемо мало детаљније случај модела у коме су предиктори категорички. Посматрајмо следећи линеарни модел

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1.6)$$

$$X_{i1} = \begin{cases} 1, & \text{за } i = 1, \dots, n_1 \\ 0, & \text{за } i = n_1 + 1, \dots, n \end{cases} \quad (1.7)$$

$$(1.8)$$

где  $\varepsilon \sim N(0, \sigma^2 I)$ .

Означимо се  $\mu_1 = \beta_0$  и  $\mu_2 = \beta_0 + \beta_1$ . Тада модел (1.6) се може представити у облику

$$Y_i = \begin{cases} \mu_1 + \varepsilon_i, & \text{за } i = 1, \dots, n_1 \\ \mu_2 + \varepsilon_i, & \text{за } i = n_1 + 1, \dots, n \end{cases}$$

Тестирање нулте хипотезе  $H_0 : \mu_1 = \mu_2$ , је еквивалентно тестирању хипотезе  $H_0 : \beta_1 = 0$ . Приметимо да је ово у ствари тестирање да је средња вредност обележја  $Y$  иста у две категорије.

Наравно, предиктор не мора имати само две категорије и "кодирање" предиктора није једнозначно одређено. То у многоме зависи од циља истраживања. Претпоставимо да предиктор  $X$  има  $k$  категорија. Тада, уместо да посматрамо  $X$  ћемо увести  $k - 1$  предиктор. Једна могућност за то је следећа. Нека вредност  $(X_1, X_2, \dots, X_{k-1}) = (0, 0, \dots, 0)$  одговара првој категорији, вредност  $(X_1, \dots, X_{k-1}) = (1, 0, \dots, 0)$  одговара другој категорији, ..., вредност  $(X_1, \dots, X_{k-1}) = (0, 0, \dots, 1)$  одговара  $k$ -тој категорији. Посматрајмо модел

$$Y = X\beta + \varepsilon, ; \quad X = (1, X_1, \dots, X_{k-1})^T.$$

Сада коефицијент  $\beta_i$  представља разлику средње вредности између прве и  $(i + 1)$  категорије, па прва категорија представља референтни ниво. У следећем примеру видећемо још један начин одабира помоћних предиктора.

**Пример 1.4.5.** Услуге јавног превоза у једној држави пружају државне агенције, али и приватне, како профитне тако и непрофитне агенције. У циљу утврђивања да ли се квалитет услуге коју пружају ова три типа агенција, разликује, направљена је одговарајућа скала којом се мери квалитет услуге и извршено анкетирање. Резултати су следећи:

Државне агенције: 61.59 79.19 68.89 72.16 70.66 63.17 53.66 68.69 68.75 60.52 68.01 73.06 55.93 74.88 62.55 69.90 66.61 63.80 45.83 64.48 58.11 73.24 73.24 69.94.

Приватне непрофитне агенције: 76.77, 68.33, 72.29, 69.48, 59.26, 67.16, 71.83, 64.63, 78.31, 61.48.

Приватне агенције: 71.77, 82.92, 72.26, 71.75, 67.95, 71.90.

Увешћемо два помоћна предиктора  $X_1$  и  $X_2$  тако да пар  $(X_1, X_2) = (0, 0)$  означава државне агенције,  $(0, 1)$  приватне непрофитне агенције а пар  $(1, 1)$  приватне профитне агенције.

```
Call: lm(formula = skala ~ X1 + X2)
Residuals:
    Min       1Q   Median       3Q      Max
-20.2892  -3.7579  -0.0666   4.0008  13.0708
Coefficients:
            Estimate Std. Error t value Pr(> |t|)
(Intercept)  66.119      1.421   46.524 <2e-16 ***
           X1     4.138      3.595    1.151  0.257
           X2     2.835      2.621    1.082  0.286
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.962 on 37 degrees of freedom
Multiple R-squared:  0.1219, Adjusted R-squared:  0.0744
F-statistic: 2.567 on 2 and 37 DF, p-value: 0.09033
```

Видимо да са нивоом значајности 0.05 не можемо одбацити хипотезу да су коефицијенти  $\beta_1$  и  $\beta_2$  безначајни. У modelu је просечна вредност оцене за државне агенције узета за референтни ниво, па 66.119 представља просечну вредност оцене државних агенција. Повећања оцене средње вредности (у односу на референтни ниво) 0 за 4.138 за приватне непрофитне агенције и  $4.138 + 2.835$  за приватне профитне агенције, нису значајна.

Наравно, у моделу поред категоричких могу постојати и нумеричке променљиве. То ћемо илустровати следећим примером.

**Пример 1.4.6.** *Подаци су преузети из књиге [6] па ћемо задржати оригиналне ознаке.*

Ознака	Значење
Price	цена куће у хиљадама долара
BDR	број спаваћих соба
FLR	површина спрата
FP	број камина
RMS	број соба
ST	број олујних прозора
LOT	удаљеност од пута у стопама
TAX	годишње таксе
BTH	број купатила
CON	од цигле (1), остало (0)
GAR	0-нема, $t$ уколико је за $t$ аутомобила
CDN	да ли су потребна додатна улагања (1, за да)
X1	1 уколико је у зони $A$ , 0 иначе
X2	1 уколико је у зони $B$ , 0 иначе

Табела 1.1: Ознаке предиктора

*Подаци су сакупљени о кућама које се налазе у 3 зоне  $A, B, C$  па је припадност зони кодирана са две помоћне променљиве  $X1$  и  $X2$ .*

*Прво ћемо направити модел у коме су сви параметри једне некретнине (сви поменути предиктори) укључени у цену.*

	Price	BDR	FLR	FP	RMS	ST	LOT	TAX	BTH	CON	GAR	CDN	X1	X2
1	53	2	967	0	5	0	39	652	1.5	1	0	0	1	0
2	55	2	815	1	5	0	33	1000	1	1	2	1	1	0
3	56	3	900	0	5	1	35	897	1.5	1	1	0	1	0
4	58	3	1007	0	6	1	24	964	1.5	0	2	0	1	0
5	64	3	1100	1	7	0	50	1099	1.5	1	1.5	0	1	0
6	44	4	897	0	7	0	25	960	2	0	1	0	1	0
7	49	5	1400	0	8	0	30	678	1	0	1	1	1	0
8	70	3	2261	0	6	0	29	2700	1	0	2	0	1	0
9	72	4	1290	0	8	1	33	800	1.5	1	1.5	0	1	0
10	82	4	2104	0	9	0	40	1038	2.5	1	1	1	1	0
11	85	8	2240	1	12	1	50	1200	3	0	2	0	1	0
12	45	2	641	0	5	0	25	860	1	0	0	0	0	1
13	47	3	862	0	6	0	25	600	1	1	0	0	0	1
14	49	4	1043	0	7	0	30	676	1.5	0	0	0	0	1
15	56	4	1325	0	8	0	50	1287	1.5	0	0	0	0	1
16	60	2	782	0	5	1	25	834	1	0	0	0	0	1
17	62	3	1126	0	7	1	30	734	2	1	0	1	0	1
18	64	4	1226	0	8	0	37	551	2	0	2	0	0	1
19	66	2	929	1	5	0	30	1355	1	1	1	0	0	1
20	35	4	1137	0	7	0	25	561	1.5	0	0	0	0	0
21	38	3	743	0	6	0	25	489	1	1	0	0	0	0
22	43	3	596	0	5	0	50	752	1	0	0	0	0	0
23	46	2	803	0	5	0	27	774	1	1	0	1	0	0
24	46	2	696	0	4	0	30	440	2	1	1	0	0	0
25	50	2	691	0	6	0	30	549	1	0	2	1	0	0
26	65	3	1023	0	7	1	30	900	2	1	1	0	1	0

Табела 1.2: Подаци о некретнинама

```
m.1=lm(Price ~BDR+FLR+FP+RMS+ST+LOT+BTH+TAX+CON+GAR+
CDN+L1+L2)
summary(m.1)
```

Call:

```
lm(formula = Price ~ BDR + FLR + FP + RMS + ST + LOT + BTH +
TAX + CON + GAR + CDN + X1 + X2)
```

Residuals:

```
   Min       1Q   Median       3Q      Max
-6.3317  -1.9256   0.0523   2.0601   5.2610
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	17.3117	7.1741	2.41	0.0327	*
BDR	-5.7705	2.3091	-2.50	0.0280	*
FLR	0.0199	0.063	3.13	0.0086	**
FP	4.3614	3.6952	1.18	0.2608	
RMS	2.2832	1.8434	1.24	0.2392	
ST	9.7552	2.3916	4.08	0.0015	**
LOT	0.3223	0.1312	2.46	0.0302	*
BTH	1.3046	2.7558	0.47	0.6444	
TAX	-0.0026	0.0049	-0.54	0.5999	
CON	2.6105	2.7138	0.96	0.3551	
GAR	3.8306	1.6615	2.31	0.0398	*
CDN	-0.7758	2.6041	-0.30	0.7709	
L1	1.7716	3.0257	0.59	0.5691	
L2	6.8935	2.9073	2.37	0.0353	*

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.435 on 12 degrees of freedom

Multiple R-squared: 0.9404, Adjusted R-squared: 0.8758

F-statistic: 14.56 on 13 and 12 DF, p-value: 2.227e-05

Видимо да неки од атрибута некретности стварно утичу на цену исте. Али, уколико желимо да видимо како самостално утиче зона у којој је кућа смештена, то не можемо закључити из ове табеле. Требао би форамално тестирати нулту хипотезу да су коефицијенти уз  $X_1$  и  $X_2$  нула против алтернативе да је један од коефицијената различит од нуле. За то можемо користити тест статистику (1.4). За то тестирање коришћемо функцију апова чији су аргументи два модела, први код кога су коефицијенти  $X_1$  и  $X_2$  различити од нуле и други

у коме то нису (у нашем случају модел  $m.2$ ).

```
m.2=lm(Price BDR+FLR+FP+RMS+ST+LOT+BTH+TAX+CON+GAR+
CDN,data=E2.2)
anova(m.1,m.2)
```

*Analysis of Variance Table*

*Model 1: Price ~ BDR + FLR + FP + RMS + ST + LOT + BTH + TAX + CON + GAR + CDN + L1 + L2*

*Model 2: Price ~ BDR + FLR + FP + RMS + ST + LOT + BTH + TAX + CON + GAR + CDN*

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	14	359.02				
2	12	236.00	2	123.02	3.13	0.0807

На основу ових резултата закључујемо да са тестом нивоа значајности 0.05 нећемо одбацити нулту хипотезу, да су повећања унутар различитих зона значајна.

## 1.5 Дијагностика модела

На крају сваког моделирања треба испитати да ли су претпоставке модела задовољене. Један део модела је детерминистички а други чине случајне грешке за које смо претпоставили да важе неки услови. Сада те услове треба проверити. Подсетимо се, претпоставили смо да су случајне грешке међусобно некорелисане, центриране и једнако расподељене са  $N(0, \sigma^2)$  расподелом. Саставни део анализе модела је и да се утврди да ли су неке тачке *атлајери*-тачке које не припадају моделу и да ли те тачке утичу на модел. Тиме ћемо се бавити у овом поглављу.

Како резидуали модела представљају оцену шума модела на који се и односне све претпоставке модела, управо ће они бити кључни за проверу коректности модела.

Први корак је да се прикаже неколико карактеристичних графика који нам много говоре о претпоставкама модела.

1. **Графички приказ**  $e \sim X_i$  за  $i = 1, 2, \dots, p$ . Уколико на овим графицима уочимо да резидуали "не осцилују равномерно" око нуле већ се може приметити неки тренд који зависи од  $X$  закључујемо да претпоставка да шум не зависи од  $X$ , није задовољена.

2. **Графички приказ  $\mathbf{e}^2 \sim \mathbf{X}$**  је алтернатива претходном графику при чему се сад у обзир узима само величина шума. Уколико уочимо неки патерн који зависи од  $X$  закључујемо као малопре.
3. **Графички прикази  $\mathbf{e} \sim \hat{\mathbf{Y}}$ ,  $\mathbf{e}^2 \sim \hat{\mathbf{Y}}$  и  $|\mathbf{e}| \sim \hat{\mathbf{Y}}$**  су погодни за уочавање хетероскедачности.

Уколико је  $Cov(\varepsilon) = \sigma^2 I$  онда је  $Cov(e) = (I - H)\sigma^2$  па је методолошки исправније, уместо резидуала, посматрати стандардизоване резидуале

$$e_i^s = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}, \quad i = 1, 2, \dots, n.$$

4. **Q – Q plot стандардизованих резидуала.** Уколико важи  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$  стандардизовани резидуали ће имати приближно нормалну расподелу па се овај график може користити за проверу нормалности резидуала.

Поред овога, велике вредности резидуала могу указивати да је баш та обсервација аутлајер.

Након прелиминарне графичке анализе неке од закључака би требало и формално тестирати.

### 1.5.1 Тестови нормалности

Постоји заиста много тестова сагласности са нормалном расподелом. Неки од најпознатијих су калсични тестови засновани на емпиријској функцији расподеле као што су Колмогоров-Смирнов, Андерсон-Дарлинг ( $AD$ ), Крамер-вон Мисесов ( $CM$ ) итд, затим Шапиро Вилк ( $SW$ ) итд.

Треба водити рачуна да су неки од тестова дизајнирани за просте хипотезе и независне и једанко расподељене случајне величине па уколико се користе за сложене хипотезе треба прво оценити расподелу тест статистике под нултом хипотезом. Укратко ћемо приказати неке тестове. Претпостављамо да су сви тестови предвиђени за прост случајан узорак обима  $n$ .

**KS тест** Тест статистика је  $KS = \sup_t |F_n(t) - F_0(t)|$ . Критична област за тестирање је  $\{KS > c\}$ .

**AD тест** Тест статистика је  $AD = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dx$ . Критична област за тестирање је  $\{AD > c\}$ .



**CM тест** Тест статистика је  $CM = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dx$ . Критична област за тестирање је  $\{CM > c\}$ .

**SW тест** Тест статистика је  $W = \frac{\sum_{i=1}^n a_i X_{(i)}}{s}$  где су  $a_i$  очекиване вредности статистика поретка стандардне нормалне расподеле. Критична област за тестирање је  $\{W < c\}$ .

У пракси се најчешћ користи *SW* тест и графички приказ (*Q-Q plot*). *SW* тест је дизајниран за сложену нулту хипотезу али за независне и једнако расподељене случајне величине, тако да се ипак мора адаптирати за тестирање нормалности резидуала применом неке од буцтреп метода. У раду [3] је показано да у случају када важи да  $h_i \rightarrow 0$ , кад  $n \rightarrow \infty$ , за велике узорке, се може применити *SW* тест.

### 1.5.2 Тестирање хомоскедастичности

Приказаћемо само Бројш-Паганов тест.

Главна идеја која стоји иза овог теста је успостављање линеарне везе између квадрата оцењених резидуала  $e_i^2$  и предиктора. Онда нулта хипотеза хомоскедастичности постаје хипотеза да су сви коефицијенти уз предикторе једнаки нула. Алтернативна хипотеза је да је бар један коефицијент уз предикторе различит од нуле.

Тест статистика је  $T = nR^2$  где је  $R^2$  коефицијент детерминације за тај помоћни модел. Показано је да ако важи нулта хипотеза онда  $T$  има приближно  $\chi_p^2$ . Главна мана овог теста је што се ова расподела мења уколико резидуали нису нормално расподељени. Зато је Кроенкер (у [4]) предложио модификацију овог теста за коју је показао да је робусна на расподелу резидуала. Управо та верзија је подразумевана у *R*-овској функцији `bptest` из пакета `lmtest`.

Када се утврди хетероскедастичност можемо урадити две ствари. Прва, када дисперзија зависне променљиве зависи од очекиване вредности, је да извршемо неку трансформацију зависне променљиве. Друга могућност је да се приступи тежинској регресији.

### Трансформације променљивих

Нека је  $\psi(y)$  трансформација зависне променљиве. Означимо са  $m = E(Y)$ . Тада је

$$\begin{aligned}\psi(Y) &\approx \psi(m) + (Y - m)\psi'(m) \\ D(\psi(Y)) &\approx (\psi'(m))^2 D(Y).\end{aligned}$$

Да би стабилизовали дисперзију потребно је да је

$$\psi'(m) = \frac{c}{\sqrt{D(y)}}.$$

Ако је  $D(Y) \sim m$  онда је  $\psi(y) = \sqrt{y}$ . Ако је  $D(Y) \sim m^2$  онда је  $\psi(y) = \log y$ .

О каквој се зависности дисперзије од очекивања ради можемо видети са графика  $|e| \sim \hat{Y}$  или  $e^2 \sim \hat{Y}$ . Такође, због утицаја предиктора на дисперзију резидуала боље је уместо резидуала посматрати стандардизоване резидуале.

### 1.5.3 Тежинска регресија

Претпоставимо да је  $D(Y_i) = \frac{\sigma^2}{w_i}$  где је  $w_i$  позната константа. Тада ћемо уместо модела

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \varepsilon_i, \quad j = 1, 2, \dots, n, \quad (1.9)$$

где је  $Cov\varepsilon = \Omega\sigma^2$  ( $\Omega$  је дијагонална матрица чији су елементи  $w_i^{-1}$ ). посматрати модел

$$Y_i\sqrt{w_i} = \beta_0\sqrt{w_i} + \sum_{j=1}^p \beta_j X_{ij}\sqrt{w_i} + \varepsilon_i\sqrt{w_i}, \quad j = 1, 2, \dots, n. \quad (1.10)$$

У моделу (1.10) је очигледно задовољен услов хомоскедастичности што ћемо касније и показати.

Поставља се питање како одабрати одговарајуће тежине. Уколико је  $D(\varepsilon_i) \sim x_i^2$  онда је најбоље узети да је  $w_i = \frac{1}{x_i}$ . Уколико је за  $Y_i$  извршено  $n_i$  мерења  $D(Y_i) = \frac{\sigma^2}{n_i}$  одакле је  $w_i = n_i$ .

Оцене методом најмањих квадрата се добијају минимизирањем израза

$$\sum_{i=1}^n (Y_i\sqrt{w_i} - \beta_0\sqrt{w_i} - \sum_{j=1}^p \beta_j X_{ij}\sqrt{w_i})^2 = \sum_{i=1}^n w_i (Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2.$$

У векторском облику модел (1.10) се може представити на следећи начин. Прво,  $\Omega^{-1}$  можемо представити у облику  $\Omega^{-1} = CC^T$  где је  $C$   $n \times n$  дијагонална матрица са елементима  $\sqrt{w_i}$ . Тада је

$$CY = CX\beta + C\varepsilon,$$

и  $CovC\varepsilon = CC^T Cov\varepsilon = CC^T \Omega \sigma^2 = \sigma^2 I$ , односно услови Гаус-Маркова су задовољени па је

$$\hat{\beta} = (X^T C^T C X)^{-1} X^T C^T C Y = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y.$$

Даље је  $Cov(\hat{\beta}) = \sigma^2 (X^T \Omega^{-1} X)^{-1}$ . Оцењене вредности су  $C\hat{Y} = CX\hat{\beta}$  па је вектор резидуала новог модела 1.10

$$r = CY - CX\hat{\beta} = C(Y - \hat{Y}).$$

Непристрасна оцена за  $\sigma^2$  је  $\hat{\sigma}^2 = \frac{1}{n-p-1} r^T r$ . Пореда овога, важно је и приметити да пошто се сада може применити Гаус-Маркова теорема, за произвољан вектор  $l$ , случајна величина  $l^T \hat{\beta}$  ће имати мању дисперзију када је  $\hat{\beta}$  добијено тежинском регресијом него класичним методом најмањих квадрата (јер су обе оцене линеарне).

**Напомена 1.5.1.** Приликом коришћења софтверских пакета треба проверити да ли функције које се користе враћају резидуале почетног модела или новог, односно да ли се добија  $e$  или  $r$ .

## 1.6 Тестирање некорелисаности

До сада смо увек претпостављали да су грешке модела некорелисане. Међутим то није увек оправдана претпоставка. На пример уколико посматрамо неке податке који су уређени у времену има смисла да су  $\varepsilon_i$  и  $\varepsilon_{i+1}$  корелисани. Један од могућих модела је ауторегресивни процес првог реда, у ознаци  $AR(1)$ , односно да је

$$\varepsilon_i = a\varepsilon_{i-1} + u_i, \quad (1.11)$$

где низ  $\{u_i\}$  задовољава услове Гаус-Маркова. Оцена методом најмањих квадрата за параметар  $a$  је

$$\hat{a} = \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=1}^n e_i^2}. \quad (1.12)$$

Тада се хипотеза некорелисаности своди на  $H_0 : a = 0$ .

За тестирање се најчешће користи Дурбин-Вотсонов тест чија је статистика

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

Уколико су резидуали некорелисани тест статистика има приближну вредност 2. Вредности између 2 и 4 упућују на негативну корелисаност, а вредности између 0 и 2 на позитивну корелисаност. Расподела тест статистике под нултом хипотезом је изведена у случају да  $\{u_i\}$  имају нормалну расподелу. Ако о није испуњено онда расподела тест статистике зависи од  $X$  и добија се применом бутстрап метода.

**Напомена 1.6.1.** У  $R$ -у је овај тест имплементиран у функцији `durbinWatsonTest` из пакета `car`. Један од аргумената у функцији је и `simulate` који има вредност `TRUE` у случају да треба применити бутстрап. Аргументом `method` се бира да ли користити параметарски или непараметарски бутстрап. У првом случају се резидуали узоркују из нормалне расподеле са оцењеним параметром дисперзије, док се у другом случају реузоркују из полазног узорка резидуала.

### Генерализовани метод најмањих квадрата

Генерализовани метод најмањих квадрата је заправо уопштење тежинске регресије.

Претпоставимо да је  $Cov(\varepsilon) = \Sigma\sigma^2$  где је  $\Sigma$  симетрична, позитивно дефинитна матрица (необавезно дијагонала). Тада се аналогоно као у случају тежинске регресије, може показати да је оцена за  $\beta$  генерализованом методом најмањих квадрата дата са

$$\hat{\beta}_{GL} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

Основа извођења је да постоји ортогонална матрица  $M$  таква да је  $\Sigma = MDM^T$  где је  $D$  дијагонална матрица. Нека је  $S = M\sqrt{D}$ . Тада је

$$\begin{aligned} S^{-1}Y &= S^{-1}X\beta + S^{-1}\varepsilon \\ D(S^{-1}\varepsilon) &= \sigma^2 I. \end{aligned}$$

Одавде се закључује да се у новом линеарном моделу са променљивама  $(S^{-1}X, S^{-1}Y)$  комплетно статистичко закључивање обавља као у класичном линеарном моделу.

Када је  $\Sigma$  непознато постоје разне методе за оцену исте. О томе се може више сазнати у [6]. На пример, у случају да важи 1.11 у оцени за  $\Sigma$  ће фигурирати оцена 1.12.

Уколико је  $\hat{\Sigma}$  оцена коваријационе матрице онда је оцена коефицијената природно

$$\hat{\beta}_{GL} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y.$$

## 1.7 Још неке трансформације зависне променљиве и предиктора

### 1.7.1 Трансформације зависне променљиве

#### Бокс Коксова траснормација

Поред оних које смо напоменули, за стабилизацију дисперзије, вероватно су највише коришћене Бокс-Коксове трансформације. Ова трансформација се користи да би се добио модел са Гаусовим шумом. Трансформација је дата са

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{за } \lambda \neq 0 \\ \log Y_i & \text{за } \lambda = 0. \end{cases}$$

Поставља се питање како да одаберемо  $\lambda$ . Оценићемо га методом максималне веродостојности, под претпоставком да зависне променљиве  $Y_i^{(\lambda)}$  имају нормалну расподелу. Функција веродостојности је

$$L(\lambda, \sigma|Y) = \frac{1}{(\sqrt{2\pi}\sigma^2)^n} e^{-\frac{(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta)}{2\sigma^2}} \left(\prod_{i=1}^n Y_i\right)^{\lambda-1}.$$

Тада је

$$\log L(\lambda, \sigma|Y) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta)}{2\sigma^2} \quad (1.13)$$

$$+ (\lambda - 1) \log\left(\prod_{i=1}^n Y_i\right) \quad (1.14)$$

Оцене за  $\beta$  и  $\sigma^2$  се добијају као и до сада, а  $\lambda$  је она вредност која максимизира функцију

$$-\frac{(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta)}{2\sigma^2} + (\lambda - 1) \log\left(\prod_{i=1}^n Y_i\right).$$

С обзиром на то да модел ипак треба да буде интерпретабилан, треба узети негу смислену вредност за  $\lambda$ , а опет "довољно блиску" са оцењеном вредности.

Коришћењем Вилковске теореме добија се да  $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$  има граничну  $\chi_1^2$  расподелу па можемо извршити тестирање да ли смо одабрали добру интерпретабилну вредност.

Јасно је да се ова трансформација може примењивати само уколико је зависна променљива позитивна. Уколико то није случај, а знамо да је  $Y_i > -a$ , за неко  $a > 0$  онда се може применити трансформација

$$Y_i^{(\lambda)} = \begin{cases} \frac{(Y_i+a)^{\lambda}-1}{\lambda} & \text{за } \lambda \neq 0 \\ \log(Y_i+a) & \text{за } \lambda = 0. \end{cases}$$

Уколико је  $a$  непознато може се одредити методом максималне веродостојности максимизирањем функције (1.13). Тада  $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$  има граничну  $\chi_2^2$  расподелу.

### Мултипликативни модели

У економији се често срећу модели код којих је случајна грешка мултипликативна, односно

$$Y_i = E(Y_i|X)\varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1.15)$$

где је

$$E(Y_i|X) = A \prod_{j=1}^p X_{ij}^{\beta_j} \quad i = 1, 2, \dots, n.$$

Одавде је природна претпоставка да су случајне грешке независне од предиктора и да је  $E(\varepsilon_i) = 1$ .

Логаритмовањем (1.15) добијамо модел

$$\log(Y_i) = \log A + \sum_{j=1}^p \beta_j \log(X_{ij}) + \log(\varepsilon_i), \quad i = 1, 2, \dots, n. \quad (1.16)$$

Добили смо класичан линеарни модел. Међутим из претпоставке  $E\varepsilon_i = 1$  не можемо закључити да је  $E(\log(\varepsilon_i)) = 0$ . Шта више, то често не важи.

Уколико  $\log(\varepsilon_i)$  има  $N(m, \sigma^2)$  расподелу  $E(\varepsilon_i) = e^{m+\frac{\sigma^2}{2}}$ . Одавде добијамо да је  $E(\log(\varepsilon_i)) = m = -\frac{\sigma^2}{2}$ . Зато је модел (1.16) боље приказати у облику

$$\log(Y_i) = \left(\log A - \frac{\sigma^2}{2}\right) + \sum_{j=1}^p \beta_j \log(X_{ij}) + \log(\varepsilon_i) + \frac{\sigma^2}{2}, \quad i = 1, 2, \dots, n.$$

Како је  $D(\log(\varepsilon_i) + \frac{\sigma^2}{2}) = D(\log(\varepsilon_i)) = \sigma^2$ , услови теореме Гаус-Маркова су задовољени тако да се параметри  $\beta_1, \dots, \beta_p, \sigma^2$  и  $\beta_0 = \log A - \frac{\sigma^2}{2}$  могу оценити методом најмањих квадрата.

## 1.8 Аутлајери, тежинске и утицајне тачке

Показали смо да је  $e = (I - H)\varepsilon$ , као и да је  $Cov(e) = (I - H)\sigma^2$ . Одавде је  $D(e_i) = \sigma^2(1 - h_i)$ . Број  $h_i$  се назива моћ, тежина (*leverage*) тачке. Што је веће  $h_i$  дисперзија  $i$ -тог резидуала ће бити мања па ће права бити ближа  $Y_i$ . Може се показати да је  $h_i = (1, x_i)^T (XX^T)^{-1} (1, x_i)$ .

Приметимо још и следеће, што је  $h_i$  мање реиздуали ће се понашати више као стварна грешка. Имајући у виду да је  $tr(H) = p + 1$ , тачке за које је  $h_i > \frac{2(p+1)}{n}$  називамо *тежинским* тачкама и испитујемо њихов утицај на одређивање параметара модела.

Поред овог правила за идентификовање тежинских тачкама, у случају да имамо узорак знатно већи од броја непознатих параметара, можемо усвојити и следеће: тачке за које је  $h_i > 0.5$  сматрамо да имају велику тежину, а оне за које је  $h_i \in [0.2, 0.5]$  средње тежинским тачкама.

Поред стандардизованих резидуала

$$e_i^s = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}},$$

посматраћемо и студентизоване резидуале (externally studentized) дефинисане са

$$e_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_i}},$$

где је  $\hat{\sigma}_{(i)}$  непристрасна оцена дисперзије када се из модела избаци  $i$ -та обсервација.

*Утицајне тачке* су оне чијим изостављањем би се модел значајно променио. Може се показати да за следеће мере утицаја важи

$$DFBETA_i = \hat{\beta} - \hat{\beta}_{(i)} = \frac{(XX^T)^{-1}x_i e_i}{1 - h_i}$$

$$DFFIT_i = \hat{y}_i - \hat{y}_{i,(i)} = \frac{h_i e_i}{1 - h_i},$$

при чему је  $\hat{y}_{i,(i)}$  је прогноза  $i$ -те вредности када је из модела искључена  $i$ -та обсервација.

Тада су резудали модела у  $i$ -тој обсервацији (уочимо да се овде ради о грешци предвиђања)

$$e_{i,(i)} = y_i - \hat{y}_{i,(i)} = \frac{e_i}{1 - h_i}$$

$$D(e_{i,(i)}) = \frac{D(e_i)}{(1 - h_i)^2} = \frac{\sigma^2}{1 - h_i}.$$

Приметимо да што је  $h_i$  веће,  $e_{i,(i)}$  ће бити веће у односу на полазни резидуал  $e_i$ . Дисперзија овог резидуала се може приказати и у облику

$$D(e_{i,(i)}) = \sigma^2(1 + X_i^T(X_{(i)}X_{(i)})^{-1}X_i),$$

где је  $X_{(i)}$  дизајн матрица без  $i$ -те обсервације. Сада важи

$$\frac{e_{i,(i)}}{\frac{\hat{\sigma}_{(i)}}{\sqrt{1-h_i}}} = \frac{\frac{e_i}{1-h_i}}{\frac{\hat{\sigma}_{(i)}}{\sqrt{1-h_i}}} = e_i^* \sim t_{n-p-2}$$

Како је

$$(n-p-1)\hat{\sigma}^2 = (n-p-2)\hat{\sigma}_{(i)}^2 + \frac{e_i}{1-h_i},$$

Екстерно студентизоване резидуале можемо приказати у облику

$$e_i^* = e_i \left( \frac{n-p-2}{\hat{\sigma}^2(1-h_i)(n-p-1) - e_i^2} \right)^{\frac{1}{2}}$$

што нам омогућује да их израчунамо, не радећи регресију без изостављене обсервације, поново.

Сада се може дефинисати једноставан тест за одређивање аутлајера заснован на томе да екстерно студентизоване резудале карактерише велика апсолутна вредност. Наиме, тест статистика ће нам бити баш  $e_i^*$  која, ако је модел коректан, тј. ако  $i$ -та обсервација се уклапа у модел чак и кад је изоставимо док формирамо исти, има Студентову  $t_{n-p-2}$  расподелу. Уколико тестирамо  $k$  тачака, треба критичну област да смањимо тако да је вероватноћа сваке  $\frac{\alpha}{k}$ .

Поред наведених мера утицаја често се користи и Куково растојање. Оно је дефинисано са

$$\begin{aligned} D_i &= \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})}{(p+1)\hat{\sigma}^2} = \frac{(\hat{Y}_i - \hat{Y}_{(i)})^T (\hat{Y}_i - \hat{Y}_{(i)})}{(p+1)\hat{\sigma}^2} = \frac{e_i^2 h_i}{(p+1)\hat{\sigma}^2(1-h_i)^2} \\ &= \frac{(e_i^s)^2}{p+1} \cdot \frac{h_i}{1-h_i}. \end{aligned}$$

Видимо да је у ово растојање инкорпорирано и одступање од модела и тежина сваке тачке. Договор је да се тачке за које је Куокво растојање веће од 1 сматрају утицајним, али да треба обратити пажњу и на оне са растојањем већем од 0.5. До закључка се може доћи поређењем са квантилима Фишерове  $F_{p+1, n-p-1}$ . Све што је веће од 50% квантила се може сматрати великим растојањем.



## 1.9 Полиномијална регресија

Полиномни регресиони модели су једна врста уопштених линеарних модела. Код ових модела регресиона функција садржи квадрате, или веће степене предиктора. Један пример таквог модела би био следећи:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon.$$

Наравно, не мора бити укључен само један предиктор већ и више њих, као и чланови интеракције између њих, на пример

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2 + \beta_4 X_2 + \beta_5 X_2^2 + \varepsilon.$$

Треба имати у виду да увек треба одабрати полином што мањег степена. Постоје две стратегије, почети од полинома малог степена па додавати чланове вишег реда док се не добију ”статистички безначајни коефицијенти”, или кренути од полинома већег степена па смањивати док се не добије значајан коефицијент уз предиктор(е) највећег степена.

Мане ове врсте регресије:

1. Број променљивих значајно расте за повећавањем степена полинома;
2. Екстраполација није добра;
3. Матрица  $X^T X$  постаје ”заражена” са повећавањем степена полинома, односно рачунање инверза матрице није поуздано. О томе ће бити више речи у наредном поглављу;
4. Оцене коефицијената могу бити веома корелисане;

Неки од проблема, посебно корелисаност оцена, се могу решити центрирањем предиктора, тј. када се у модел, уместо  $X_1, \dots, X_p$  уврсте центриране променљиве  $X_1 - \bar{X}_1, \dots, X_p - \bar{X}_p$ .

## 1.10 Мултиколинеарност

До сада само претпостављали да је дизајн матрица максималног ранга, тј. да су сви предиктори линеарно независни. Наравно, у пракси се то увек не догађа. Често је да између њих не постоји баш линеарна зависност али су веома корелисани. У ове две ситуације кажемо да тада постоји проблем мултиколинеарности. Неки од разлога због ког се то дешава су следећи:

- **Превише предиктора у моделу (више од обсервација).** Овај проблем се јавља често у медицинским истраживањима у којима има премало пацијената у истраживању.
- **Непрецизна формулација модела.** Беспотребно убацивање већих степена предиктора или сабирака који се односе на њихову интеракцију. На пример, уколико имамо два предиктора  $X_1$  и  $X_2$  можда је  $X_1X_2$  непотребно убацивати у модел.
- **Убацивање у модел предиктора између којих природно постоји линеарна веза.** На пример, убацивати у модел предикторе БРУТО плата, НЕТО плата и ТАРА плата.
- **Узорак на коме се врше обсервације је условљен неким ограничењима у популацији.** Узорковање вршимо из "пот-популације" на којој су предиктори веома корелисани.

Ако између предиктора постоји линеарна зависност, матрица  $X^T X$  није инвертибилна и онда оцена за  $\beta$  није јединствена. У овом поглављу видећемо да је такву ситуацију лако детектовати, док је то код предиктора код којих постоји приближна линеарна зависност знатно теже. Велике дисперзије оцена често су један од индикатора приближне мултиколинеарности. Због тога се може десити "лажно" прихватање нулте хипотезе да коефицијенти уз предикторе нису значајни. Такође, очекује се да се, уколико се неки податак само мало промени, добију знатно различите оцене коефицијената.

Најчешћи показатељ мултиколинеарности је фактор инфлације дисперзије (*variance inflation factor*)  $VIF_j = \frac{1}{TOL_j}$ , где је  $TOL_j = 1 - R_j^2$  толеранција, а  $R_j^2$  коефицијент детерминације модела у коме је зависна променљива  $X_j$  а независне све остале. Јасно је да вредност  $VIF_j$ -а блиска јединици говори да  $X_j$  није у линеарној вези са осталим предикторима. Сматра се да проблем мултиколинеарности постоји уколико је  $VIF_j > 5$ . Једно решење проблема је да се избаце неки предиктори али се тада намеће питање који и може се направити грешка приликом њиховог избора.

## 1.11 Анализа главних компоненти

Проблем који се јавља у интерпретацији модела последица је његове осетљивости на различите мерне скале полазних предиктора. Најјед-

ноставнији пример био би кад би нпр.  $X_1$  била редовна примања у динарима, а  $X_2$  додатна месечна примања у хиљадама динара. Мерне јединице, односно скала, свакако утичу на дисперзију па се може десити да ће један предиктор доминирати првом главном компонентом. Овај проблем се може решити стандардизацијом модела. Зато у даљем тексту претпостављамо да смо стандардизацију модела извршили, тј. да се дизајн матрица  $X$  састоји од стандардизованих предиктора (можемо претпоставити и да нема слободног члана јер смо показали да се добија управо такав модел ако стандардизујемо и зависну променљиву).

Један од начина да решимо проблем мултиколинеарности је да избацимо неке променљиве. Идеја је да се полазни скуп предиктора замени неким њиховим линеарним комбинацијама ( $k \leq p$ ) који садрже скоро исту информацију као полазани скуп и да се избаце оне линеарне комбинације које не доприносе квалитету модела.

Нека је  $Z = XA$  линеарна трансформација предиктора,

$$Z_k = a_{1k}X_1 + a_{2k}X_2 + \dots + a_{pk}X_p.$$

Тада је

$$D(Z_k) = a_k^T \Sigma a_k.$$

Без умањења општости можемо претпоставити да је максимална дисперзија, уз услов да је  $|a_i| = 1$  баш  $D(Z_1)$ . Тада ћемо  $Z_1$  звати *прва главна компонента*. Нека је  $a_2$  вектор за који је  $|a_2| = 1$ ,  $Xa_2$  је ортогонално са  $Z_1$  и  $a_2^T \Sigma a_2$  је максимално могуће.  $Z_2 = Xa_2$  зваћемо *другом главном компонентом*. Поступак понављамо, при чему је свака од наредних главних компоненти ортогонална на све претходне.

**Лема 1.11.1.** *Нека је  $\Sigma$  коваријациона матрица случајног вектора  $X$ . Нека су  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$  њене сопствене вредности. Тада је  $i$ -та главна компонента дата са  $Z_i = v_i^T X$ , за  $i = 1, 2, \dots, p$ , где је  $v_i$   $i$ -ти сопствени вектор.*

**Доказ.** На предавању.

Приметимо да је тада  $D(Z_i) = \lambda_i$ , као и да је за  $i \neq j$   $Z_i$  ортогонално на  $Z_j$ , ( $Cov(Z) = V^T \Sigma V = Diag(\lambda_1, \dots, \lambda_p) = D_\lambda$ ).

**Лема 1.11.2.** *Нека је  $Z = XV$ . Тада је*

$$\sum_{i=1}^p D(X_i) = \sum_{i=1}^p D(Z_i) = \sum_{i=1}^p \lambda_i.$$

**Доказ.** Како је  $D_\lambda = V^T \Sigma V$  закључујемо да је  $\sum_{i=1}^p D(Z_i) = \text{tr}(D_\lambda) = \text{tr}(V^T \Sigma V) = \text{tr}(\Sigma)$ .

□

Последица ове леме је да ротирањем координатног система нисмо променили укупан варијабилитет система, као и да је удео објашњеног варијабилитета  $i$ -том главном компонентом  $\frac{\lambda_i}{\sum_{i=1}^p \lambda_i}$  за  $i = 1, 2, \dots, p$ . Сматра се да треба задржати онолико компоненти колико је потребно да се објасни бар 80% целокупног варијабилитета.

Ради тумачења главних компоненти згодно је види какав утицај сваки од предиктора има на  $i$ -ту главну компоненту.

**Лема 1.11.3.** *Коефицијент корелације измеу  $Z_i$  и  $X_k$  је*

$$\rho_{Z_i, X_k} = \frac{v_{ik} \sqrt{\lambda_i}}{\sqrt{D(X_k)}}.$$

Стандардизован модел  $Y = X\delta + \varepsilon$  се може приказати у облику  $Y = Z\eta + \varepsilon$  где су  $Z$  главне компоненте добијене од стандардизованих предиктора.

Тада је оцена непознатог параметра  $\eta$  добијена методом најмањих квадрата дата са

$$\hat{\eta} = (Z^T Z)^{-1} Z^T Y = \frac{1}{n-1} D_\lambda^{-1} Z^T Y.$$

Важи и

$$D(\eta_j) = \frac{\sigma^2}{(n-1)\lambda_j}.$$

Одавде се јасно види да мале сопствене вредности утичу на велике дисперзије оцена коефицијената (што није добро). Зато ћемо баш такве компоненте избацити.

Претпоставимо да смо  $r$  главних компоненти одлучили да задржимо, а преосталих  $p - r$  да избацимо. Увешћемо следећу нотацију.

$$Z = \begin{pmatrix} Z_{(r)} & Z_{(p-r)} \end{pmatrix}$$

$$\eta = (\eta_{(r)}^T \eta_{(n-r)}^T)^T$$

$Z_{(r)}$  су компоненте које задржавамо. Како је  $Z = XV$  закључујемо да је  $Z_{(r)} = XV_{(r)}$ , где је  $V_{(r)}$  је матрица која се састоји од првих  $r$  колона

матрице  $V$ , а  $V_{(p-r)}$  матрица која се састоји од преосталих колона. Нови модел се може приказати у облику

$$Y = Z_{(r)}\eta_{(r)} + \tilde{\varepsilon} = Z\eta_r + \tilde{\varepsilon},$$

где је  $\eta_r = \begin{pmatrix} \eta_{(r)} \\ \mathbf{0} \end{pmatrix}$ , или у облику

$$Y = X\delta_r + \tilde{\varepsilon},$$

где је  $\delta_r = V\eta_r$ .

Непознат параметар  $\delta$  оценићемо са  $V\hat{\eta}_r = V_{(r)}\hat{\eta}_{(r)}$ . С обзиром на то да избацујемо неке компоненте, за очекивати је да је ова оцена пристрасна.

$$\begin{aligned} E(V_{(r)}\hat{\eta}_{(r)}) &= V_{(r)}E(\hat{\eta}_{(r)}) = V_{(r)}\eta_{(r)} = V\eta_r = (V_{(r)} \ V_{(p-r)}) \begin{pmatrix} \eta_{(r)} \\ \mathbf{0} \end{pmatrix} \\ &= (V_{(r)} \ V_{(p-r)}) (V_{(r)} \ \mathbf{0})^T \delta = \begin{pmatrix} I_{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \delta \\ &= \delta - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{(p-r)} \end{pmatrix} \delta \neq \delta. \end{aligned}$$

Нека је  $\tilde{\delta}$  оцена МНК за  $\delta$ .

$$\begin{aligned} Cov(\tilde{\delta}) &= \frac{1}{n-1} \sigma^2 V D^{-1} V^T = V \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_{p-r}^{-1} \end{pmatrix} V^T \frac{\sigma^2}{n-1} \\ &= \frac{\sigma^2}{n-1} (V_{(r)} D_r^{-1} V_{(r)}^T + V_{(p-r)} D_{p-r}^{-1} V_{(p-r)}^T) \end{aligned}$$

Први сабирак представља коваријацију оцењених параметара на основу  $r$  задржаних главних компоненти а остатак, део који је нестао елиминацијом компоненти које су биле ”вишак”.

**Пример 1.11.1.** са предавања

## 1.12 Методе регуларизације

Прва метода коју ћемо описати је такозвана *назубљена регресија*.

Један од начина да се реши проблем мултиколинеарности је, као и у случају анализе главних компоненти, је да се непознат параметар  $\delta$  оцени са

$$\hat{\delta}_c = (X^T X + cI)^{-1} X^T Y. \quad (1.17)$$

Константа  $c$  је мали позитиван број који утиче, с једне стране, на пристрасност оцене, а са друге повећава стабилност оцена. Овај метод се не мора применити на центриран модел. Може се показати да је

$$\hat{\delta}_c = (c(X^T X)^{-1} + I)^{-1} \hat{\delta}$$

Испитајмо пристрасност оцене.

$$E(\hat{\delta}_c) = (X^T X + cI)^{-1} (X^T X + cI - cI) \delta = \delta - c(X X^T + cI)^{-1} \delta.$$

Као што се могло и очекивати, оцена није непристрасна. Даље је,

$$Cov(\hat{\delta}_c) = (X^T X + cI)^{-1} X^T X (X^T X + cI)^{-1} \sigma^2,$$

па је укупан варијавилитет једнак

$$\begin{aligned} tr((X^T X + cI)^{-1} X^T X (X^T X + cI)^{-1} \sigma^2) &= \sigma^2 tr((X^T X + cI)^{-2} X^T X) \\ &= \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i (1 + c\lambda_i^{-1})^2}, \end{aligned}$$

где су  $\lambda_i$  сопствене вредности матрице  $X^T X$ . То је мање од суме дисперзија појединаних компоненти оцене методом најмањих квадрата

$$\sigma^2 \sum_{j=1}^p \frac{1}{\lambda_j}.$$

Други начин да се ова оцена интерпретира је да она заправо представља оцену најмањих квадрата када за параметре постоји горње ограничење, односно решава се оптимизациони проблем

$$\min_{\delta} (y_0 - X\delta)^T (y_0 - X\delta), \quad \|\delta\|_2 \leq C,$$

за неко  $C$  које се може изразити у функцији од  $c$ . Тај проблем је еквивалентан тражењу минимума функције

$$(y_0 - X\delta)^T (y_0 - X\delta) + c\|\delta\|_2. \quad (1.18)$$

Важна особина ове методе је да се она **не може** користити за селекцију предиктора.

*Напомена:* Из израза (1.18) се може се показати да се ова оцена добија методом најмањих квадрата кад додамо у систем ”вештачке предикторе” који не утичу на вредности зависне променљиве (односно за које је вредност зависне променљиве 0).

Поред ове методе регуларизације, важна за споменути је и такозвана LASSO регуларизација (Least Absolute Shrinkage and Selection Operator). За разлику од оцене назубљеном регресијом, овде је ”казнена функција” дата  $L_1$  нормом оцене коефицијената, односно

$$\hat{\delta}_l = \arg \min_{\delta} (y_0 - X\delta)^T (y_0 - X\delta) + c \sum_{j=1}^p |\delta_j|.$$

Сада се може добити да је вредност неког од коефицијента 0, и то што је веће  $c$  више је таквих. Зато се овај метод **може** искористити за селекцију предиктора.

Поред поменутих метода регуларизације може се посматрати и

$$\hat{\delta}_l = \arg \min_{\delta} (y_0 - X\delta)^T (y_0 - X\delta) + c \sum_{j=1}^p |\delta_j|^q$$

као и

$$\hat{\delta}_l = \arg \min_{\delta} (y_0 - X\delta)^T (y_0 - X\delta) + c_1 \|\delta\|_1 + c_2 \|\delta\|_2, \quad c_1 + c_2 = 1.$$

Поставља се питање како одредити параметар  $c$ . Најчешће се то ради кросвалидацијом. (Више о томе је било речено на предавању).

**Пример 1.12.1.** *На предавању.*

## Уопштени линеарни модели

У неким ситуацијама су зависне променљиве дискретног типа. Наједноставнији пример би била променљива  $Y$  која за сваку вредност предиктора узима само две вредности: 0 или 1 (категоричка променљива са две

вредности: ДА или НЕ, ИСТИНА или НЕИСТИНА, ЗА или ПРОТИВ).  
Дакле, условна расподела предиктора је

$$Y_i|X_i : \begin{pmatrix} 0 & 1 \\ 1 - \pi(X_i) & \pi(X_i) \end{pmatrix}$$

Регресиона функција је

$$E(Y_i|X_i) = \pi(X_i)$$

Јасно је, да у овој ситуацији, линеарни модел не би био адекватан. Неки од разлога су следећи.

1. Грешке модела не могу се моделирати нормалном расподелом, или неком другом апсолутно непрекидном и симетричном око нуле.
2. Дисперзија грешака модела није константна. Важи:  $D(Y_i|X_i) = \pi(X_i)(1 - \pi(X_i)) = D(\varepsilon_i)$ .
3. С обзиром да је регресиона функција вероватноћа, треба да буде задовољено да је  $\pi(X_i) \in [0, 1]$ . За линеарну функцију то очигледно не важи.

Једно од могућих решења овог проблема је да се средња вредност зависне променљиве прво трансформише у скуп  $\mathbf{R}$  а затим моделира линеарним моделом. Трансформација свакако није јединствена. У далјем тексту ћемо размотрити неке од њих.

## Пробит регресија

Ради једноставности, за сада, претпоставимо да имамо само један предиктор. Тада је  $\pi(X_i) = \Phi(\beta_0^* + \beta_1^* X_i)$ .

Одавде је  $\Phi^{-1}(\pi(X_i)) = \beta_0^* + \beta_1^* X_i$  линеарни модел. Ова трансформација је позната под називом *пробит трансформација*.

## Логистичка регресија

Сада је  $\pi(X_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$ .

Одавде је  $F_L^{-1}(\pi(X_i)) = \log\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right) = \beta_0 + \beta_1 X_i$ . Ова трансформација је позната под називом *логит трансформација*.

Количник  $\frac{\pi(X_i)}{1 - \pi(X_i)}$  се назива *квота* (odds.)

Параметре модела оцењујемо методом максималне веродостојности. Логаритам функције веродостојности је



$$\begin{aligned} L(\beta) &= \sum_{i=1}^n (Y_i \log \left( \frac{\pi(X_i)}{1 - \pi(X_i)} \right) + \log(1 - \pi(X_i))) \\ &= \sum_{i=1}^n Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^n \log(1 + e^{\beta_0 + \beta_1 X_i}) \end{aligned}$$

Максимум ове функције одређује се нумерички. Нека су добијене оцене  $\hat{\beta}_0$  и  $\hat{\beta}_1$  за непозанте коефицијенте  $\beta_0$  и  $\beta_1$ . Одавде је оцењена регресиона функција

$$\hat{\pi}(X_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}}$$

Оцењена логит функција је

$$\hat{\lambda}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Даље, означимо са  $\hat{\pi}_i = \hat{\pi}(X_i)$ , као  $\omega_i = \hat{\pi}_i(1 - \hat{\pi}_i)$ . Тада је пондерисана средина

$$\bar{X}_\omega = \frac{\sum_{i=1}^n \omega_i X_i}{\sum_{i=1}^n \omega_i}.$$

Пондерисана сума квадратних одступања је

$$SS_\omega = \sum_{i=1}^n \omega_i (X_i - \bar{X}_\omega)^2.$$

Може се показати да су стандардна одступања оцена параметара

$$SE(\hat{\beta}_0) = \sqrt{\frac{1}{\sum_{i=1}^n \omega_i} + \frac{\bar{X}_\omega^2}{SS_\omega}}$$

$$SE(\hat{\beta}_1) = \frac{1}{\sqrt{SS_\omega}}.$$

Коваријанса оцена параметара је

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{X}_\omega}{SS_\omega}.$$

Тада је

$$\lambda(\hat{\pi}(X)) = \hat{\beta}_0 + \hat{\beta}_1 X,$$

као

$$SE(\lambda(\hat{\pi}(X))) = \sqrt{SE(\hat{\beta})^2 + 2XCov(\hat{\beta}_0, \hat{\beta}_1) + X^2SE(\hat{\beta}_1)^2}$$

и

$$SE(\hat{\pi}(X)) = \hat{\pi}(X)(1 - \hat{P}(X))SE(\lambda(\hat{\pi}(X))).$$

Сада можемо направити и интервал поверења за  $\pi(X)$ .

Тестирање значајности коефицијената се може тестирати тестом количника веродостојности. Како

$$2 \log \left( \frac{L(\hat{\beta}_0, \hat{\beta}_1)}{L(\hat{\beta}_0)} \right)$$

под нултом хипотезом има приближно  $\chi_1^2$  расподелу, можемо извршити тестирање значајности коефицијента уз предиктор на уобичајан начин.

Уколико се у моделу налази  $p$  предиктора онда тест статистика под нултом хипотезом има  $\chi_p^2$  расподелу.

Још један од начина да проверимо утицај сваке независне променљиве на посматрану зависну, као значај сваког коефицијента, односно да се процени да ли ће се избегавањем неког коефицијента изгубити на квалитету модела, је Валдов тест. Користи се Валдова тест статистика

$$Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}, \quad i = 0, 1.$$

Ова статистика, при важењу нулте хипотезе ( $\beta_i = 0$ ) има нормалну расподелу па се могу направити одговарајуће критичне области за тестирање и израчунати  $p$ -вредности тестова. Стандардно одступање оцене се може добити на следећи начин.

До сад нисмо говорили о типу предиктора. Уколико је предиктор дискретна случајна величина онда можемо проверити значајност коефицијената и на следећи начин.

Чланове узорка груписаћемо на основу вредности независне променљиве. Дакле, за свако  $X_i$  из узорка формирамо подскуп који чине они елементи узорка чија је независна компонента једнака одабраном  $X_i$ .

Нека је  $m_j$  број елемената у  $j$ -тој подгрупи посматраног узорка,  $j = 1, 2, \dots, J$ . У оквиру сваке подгрупе се може оценити условна вероватноћа  $\pi(X_j) = P\{Y = 1|X_j\}$ . Нека је  $n_j$  прој елемената у подгрупи за које је вредност зависне променљиве једнака 1. Оцена поменуте вероватноће, на основу логистичког модела је  $\hat{P}\{Y = 1|X_j\} = \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_j}}$ . Тада је

очекиван просечан број елемената из узорка чија је вредност зависне променљиве 1, једнака:

$$\hat{n}_j = m_j \hat{P}_j = m_j \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_j}}.$$

У зависности од расподељености зависне променљиве у оквиру сваке групе, као и међусобном односу група на основу те карактеристике, користе се различите статистике за проверу квалитета добијеног логистичког модела.

### Пирсонови резидуали

Пирсонов  $j$ -ти резидуал је дефинисан са

$$r_j = \frac{n_j - m_j \hat{P}_j}{\sqrt{m_j \hat{P}_j (1 - \hat{P}_j)}} = \frac{n_j - \hat{n}_j}{\sqrt{\hat{n}_j (1 - \frac{\hat{n}_j}{m_j})}}.$$

Пирсонова статистика је дефинисана са

$$C = \sum_{j=1}^J r_j^2.$$

$C$  има приближно  $\chi_{J-2}^2$ .

Квалитет Пирсонових резидуала испољава се чињеницом да је њихова очекивана вредност 0, као и да је за сваки резидуал дисперзија иста.

### Резидуали девијације

Резидуал девијације, за  $n_j - \hat{n}_j > 0$ , је дефинисан са:

$$\begin{aligned} d_j &= \sqrt{2 \left( n_j \ln \frac{n_j}{m_j \hat{P}_j} + (m_j - n_j) \ln \frac{m_j - n_j}{m_j (1 - \hat{P}_j)} \right)} \\ &= \sqrt{2 \left( n_j \ln \frac{n_j}{\hat{n}_j} + (m_j - n_j) \ln \frac{m_j - n_j}{m_j - \hat{n}_j} \right)} \end{aligned}$$

За  $n_j - \hat{n}_j < 0$  за  $j$ -ти резидуал се узима  $-d_j$ , док је једнак нули у супротном.

Када је  $n_j = 0$

$$d_j = -\sqrt{2m_j \left| \ln \frac{m_j}{m_j - \hat{n}_j} \right|},$$

док је за  $n_j = m_j$

$$d_j = \sqrt{2m_j \left| \ln \frac{m_j}{\hat{n}_j} \right|}.$$

Тест статистика је

$$D = \sum_{j=1}^J d_j^2.$$

$D$  има приближно  $\chi_{J-2}^2$  расподелу. Показује се да ови резидуали брже теже нормално распоређеној случајној променљивој, него Пирсонови резидуали.

## Лог-Вејбулова регресија

$\pi(X_i) = 1 - e^{-e^{\beta_0 + \beta_1 X_i}}$ . Одавде је  $F_G^{-1}(\pi(X_i)) = \log(-\log(1 - \pi(X_i))) = \beta_0 + \beta_1 X_1$ .

Ова трансформација, због своје асиметричности, се најчешће користи за моделовање малих и великих вероватноћа успеха. Позната је под називом *трансформација итерираног логаритма* (complementary log-log regression).

На предавању је приказано како ови модели могу искористити за класификацију зависне променљиве.

## 1.13 Уопштени линеарни модели

Ови модели се састоје од следећих компоненти:

- линеарна комбинација коефицијената модела

$$\eta_j = X_j^T \beta \quad \text{односно} \quad \eta_j = \beta_0 + \sum_{i=1}^p X_{ji} \beta_i$$

- ”линк” функције која представља трансформацију коју треба применити на функцију средње вредности зависне променљиве, да би се та трансформисана променљива могла описати линеарним моделом, односно за  $\mu_j = EY_j$  и за линк функцију  $g$  важи

$$g(\mu_j) = \eta_j$$

- расподела случајне променљиве припада експоненцијалној фамилији расподела са распршењем. Последица је да се дисперизија зависне променљиве се може представити у облику

$$D(Y_j) = CV(\mu_j).$$

Када је  $g(x) = x$ ,  $V(x) = 1$  и  $C = \sigma^2$  добијамо класичан линеарни модел. Уколико  $Y_j \sim \mathcal{B}(1, \mu_j)$  расподелу и  $g(x) = F_L^{-1}(x) = \log\left(\frac{x}{1-x}\right)$  и  $V(x) = x(1-x)$  добијамо логистичку регресију.

Веома често се у пракси јавља случај када  $Y_j \sim \mathcal{P}(\lambda_j)$ . Тада је  $\mu_j = \lambda_j$  и  $D(Y_j) = \lambda_j = \mu_j$  па је  $V(x) = x$ . Треба још да одредимо линк функцију. Приметимо да за њу треба да важи да слика  $(0, \infty)$  на  $(-\infty, \infty)$ . Зато је природан избор линк функција  $g(x) = \log(x)$ .

Нормална, биномна и Пуасонова расподела припадају експоненцијалној фамилији расподела. Експоненцијалној фамилији са распршењем (расејањем) припадају све расподеле за које се функција густине (закон расподеле) може приказати у облику:

$$f(y, \theta) = e^{\frac{c(\theta)^T T(y) - d(\theta) + S(y)}{\phi(\tau)}}$$

Параметар  $\tau$  се назива параметром *распршења*. Када је  $\phi(\tau)$  познато ради се о класичној експоненцијалној фамилији расподела.

Уколико је  $T(y) = y$  и  $c(\theta) = \theta$  кажемо да се ради о расподели у *канонском облику*. Тада је

$$\begin{aligned} EY &= -d'(\theta) = \mu \\ DY &= d''(\theta)\phi(\tau) = V(\mu)\phi(\tau). \end{aligned}$$

Најчешће је  $\phi(\tau) = a\tau$  и у случају уопштеног линеарног модела подразумевамо да је

$$f(y_i) = e^{\frac{y_i \theta_i - d(\theta_i)}{a_i \tau} + c(y_i, a_i \tau)}.$$

Непознати параметри модела се одређују, као и у случају логистичке регресије, методом максималне веродостојности.

Вратимо се одабиру линк функције. Уколико је линк функција одабрана тако да је за канонски параметар  $\theta = \eta$  онда такву функцију називамо *канонском* линк функцијом. Јасно је да је у случају логистичке регресије канонска функција баш *logit* функција.

Предност одабира канонске линк функције је што је тада  $X^T Y$  довољна статистика за  $\beta$  јер је

$$L(y, \theta) = e^{\frac{\sum_{i=1}^n y_i x_i^T \beta - d(x_i^T \beta) - S(y_i)}{\phi(\tau)}}$$

Друга предност је што се може показати да  $MLE$  постоји и да је јединствена.

$$l(y, \beta, \tau) = \sum \frac{y_i x_i^T \beta - d(x_i^T \beta)}{a_i \tau} + \sum c(a_i \tau, y_i)$$

$$\frac{\partial l(y, \beta, \tau)}{\partial \beta} = \sum \frac{x_i (y_i - d(x_i^T \beta))}{\tau a_i} = 0$$

$$\frac{\partial^2 l(y, \beta, \tau)}{\partial \beta \partial \beta^T} = - \sum \frac{x_i x_i^T d(x_i^T \beta)}{\tau a_i} \leq 0$$

### 1.13.1 Асимптотска својства $ML$ оцене

- асимптотска нормалност као последица  $MLE$  оцене у регуларном случају;
- може се искористити Валдова статистика дефинисана као у случају логистичке регресије;
- једна од мера квалитета модела је и *девијација* односно ”мера одступања претпостављеног модела од засићеног модела.” Дефинише се са  $D = 2\tau(l(y, \hat{\theta}_s) - l(y, \hat{\theta}))$  где је  $\hat{\theta}_s$  оцена непознатих параметара у засићеном моделу-моделу у коме су непознати параметри средњих вредности којих има  $n$ . У овом моделу немамо ограничење на те параметре у виду неке функције линеарног модела, као што је то у нашем претпостављеном моделу случај. Уколико је  $\tau$  познато онда се разлика девијација користи за тестирање значајности коефицијената модела. Наиме, уколико је  $H_0$  : да су неких  $k$  коефицијената у моделу 0, онда  $\frac{(D_0 - D_1)}{\tau}$  има  $\chi_{p+1-k}^2$ . Понекад се  $\frac{D}{\tau}$  назива скалираном девијацијом.

Када је  $\tau$  непознато, може се оценити са  $\hat{\tau} = \frac{1}{n-p-1} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}$  (сума квадрата Пирсонових резидуала). Тада за тестирање значајности коефицијената можемо користити статистику

$$\frac{\frac{1}{p+1-k} (D_0 - D_1)}{\hat{\tau}}$$

која уколико је нулта хипотеза тачна и  $F_{p+1-k, n-p-1}$ . Дакле, у случају просте логистичке регресије имамо два параметра и  $p = 1$ .

# Литература

- [1] J. J. Faraway. *Linear models with R*. CRC press, 2014.
- [2] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [3] J. Jurečková and J. Picek. Shapiro–wilk-type test of normality under nuisance regression and scale. *Computational Statistics & Data Analysis*, 51(10):5184–5191, 2007.
- [4] R. Koenker. A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1):107–112, 1981.
- [5] J. Neter, M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.
- [6] A. Sen and M. Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.