0.1 Основни појмови из линеарне алгебре

Дефиниција 0.1.1. *Векторски простор V је непразан скуп, затворен за сабирање и скаларно множење. Његови елементи називају се вектори.*

Дефиниција 0.1.2. *Вектори $v_1, ..., v_n$ су линеарно независни ако не постиј нетривијално решење једначине*

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

Дефиниција 0.1.3. *Вектори v_1 и v_2 су ортогонални ако је $< v_1, v_2 >= 0$.*

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

Дефиниција 0.1.5. *Ранг $n \times p$ 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 0.1.6. *Ако је за неко $\lambda \in R$, $Ax = \lambda x$ онда се λ назива сопствена вредност матрице A, а x сопствени вектор.*

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

Дефиниција 0.1.7. *Траг квадратне матрице $A = [a_{ij}]$ предцавља збир елемената на дијагонали, односно*

$$trA = \sum_i a_{ii}.$$

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

1. $tr(A + B) = tr(A) + tr(B)$;

2. Ако је A $n \times n$ матрица и P несингуларна матрица, онда је $tr(P^{-1}AP) = tr(A)$;

3. Ако је A $n \times n$ матрица и M ортогонална матрица, онда је $tr(M^TAM) = tr(A)$.

Дефиниција 0.1.8. *Квадратна матрица A је симетрична ако је $A = A^T$.*

Особине симетричне $n \times n$ матрице A:

1. Постоји ортогонална матрица $C = (c_1, ...c_n)$ и дијагонална матрица $\Lambda = (\lambda_1, ...\lambda_n)$ (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. $tr(A) = \sum_{i=1}^{n} \lambda_i$;

4. Ако је A несингуларна матрица онда је $tr(A^{-1}) = \sum_{i=1}^{n} \lambda_i^{-1}$;

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

$$x^TAx = \sum \lambda_i y_i^2;$$

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 $U^TU = I = V^TV = I$ и L је несингуларна матрица ранга r.

Дефиниција 0.1.9. *Неке је A симетрична матрица. Тада је са $Q(x) = x^TAx$ 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 допуста 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 акко постоји несингуларна матрица P таква да је $A = P^TP$

Важи и следеће

$$\max_{x\neq 0} \frac{x^T A x}{x^T x} = \lambda_{\max} \tag{1}$$

$$\min_{x\neq 0} \frac{x^T A x}{x^T x} = \lambda_{\min} \tag{2}$$

Дефиниција 0.1.10. *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 семидефинитна онда је и пројектор, као и $P_1P_2 = P_2P_1 = P_2$

0.1.1 Матрично диференцирање

Дефиниција 0.1.11. *Нека је X $n \times p$ матрица и f скаларна функција. Тада је матрично диференцирање дефинисано са*

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

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. $\frac{\partial x^T A y}{\partial x} = Ay$;

0.2 Неке важне вишедименизионалне расподеле

0.2.1 Нормална расподела

Дефиниција 0.2.1. *Случајни вектор X има n-димензиона нормалну расподела $N_n(\mu, \Sigma)$ уколико је његова фунцкија 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 Σ *симетрична, позитивно дефинитна коваријациона матрица а са* $|\Sigma|$ *је означена њена детерминанта.*

Вишедимензионална нормална расподела има следеће лепе особине:

- Уколико X има $N(\mu, \Sigma)$ распделу 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

  $X|y \sim N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T)$

  $Y|x \sim N(\mu_y + \Sigma_{xy}^T\Sigma_{xx}^{-1}(x - \mu_x), \Sigma_y - \Sigma_{xy}^T\Sigma_x^{-1}\Sigma_{xy})$;
- Момент генераторна функција случајног вектора X са $N(\mu, \Sigma)$ расподелом је $M_X(t) = E(e^{t^T x}) = e^{\mu^T t + \frac{1}{2}t^T \Sigma t}, \quad t \in R$;
- X се може представити у облику $X = AZ + \mu$ где је $AA^T = \Sigma$ а Z има стандардну вишедимензионалну нормалну расподелу.
- Нека $X = (X_1, ..., X_n)^T$ има вишедимензиону нормалну расподелу. Тада су компоненте вектора независне акко су некорелисане (коваријациона матрица је дијагонална матрица).

0.2.2 χ^2 расподела

Дефиниција 0.2.2. *Нека су* $X_1, ..., X_n$ *независне случајне величине са* $N(\theta_1, 1)$,...,$N(\theta_n, 1)$ *расподелма ,редом. Тада случајна величина*

$$Y = \sum_{j=1}^{k} X_j^2 \tag{3}$$

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

Приметимо да заправо ова теорема оправдава дефиницију χ^2 расподеле јер сугерише да расподела од Y зависи само од степени слободе и μ.

Доказ. Нека је $B = [b_{ij}]$ ортогонала матрица тако да је $b_{1j} = \theta_j \mu^{-\frac{1}{2}}$, за $j = 1, 2..., k$. 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. □

0.2.3 Фишерова расподела

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

$$\frac{\frac{U_1}{n_1}}{\frac{U_2}{n_2}}$$

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

0.2.4 Студентова расподела

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

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

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

0.3 Расподела квадратне форме

Теорема 0.3.1 (Кохран). *Нека су $X_1, ..., X_n$ 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 $j = 1, 2, ..., k$, при чему је $R(A_j) = r_j$. Тада је $\sum_{j=1}^{k} r_j = n$. ако и само ако

1. $Q_1, ..., Q_k$ *су независне случајне величине и*

2. Q_j/σ^2 *има* $\chi^2_{r_j}$ *расподелу.*

Може се формулисати и општије тврђење:

Теорема 0.3.2 (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* $j = 1, 2, ..., k$*, 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. $Q_1, ..., Q_k$ *су независне случајне величине и*

2. Q_j/σ^2 *има* $\chi^2_{r_j}(\mu_j)$ *расподелу, при чему је* $\mu_j = \theta^T A_j \theta$

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

Последица 0.3.1. *Узорачка средина* $\bar{X}$ *и поправљена узорачка дисперзија* S^2 *су незавине сучајне величине, и* $(n-1)S^2/\sigma^2$ *има* χ^2_{n-1} *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

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

где је J $n \times n$ матрица чији су сви елементи јединице. Приметимо да лева страна једнакости има χ^2_n 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.

Лема 0.3.1. *Нека су* $x_1, ...x_n$ *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* $R(A_i) = r_i$. *Ако је* $\sum_{i=1}^{k} r_i = n$ *тада постоји ортогонална матрица* M, *таква да за* $x = My$ *важи*

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

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_1 таква да је $M_1^T A_1 M_1 = D_1$ где је D_1 дијагонална матрица. Без умањења општости можемо претпоставити да су од сопствене вредности поређане тако да су $\lambda_1, ...\lambda_{r_1}$ различите од нуле а остале нула. Нека је $x = M_1 y$. 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

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

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

Важно је приметити да све квадратне форме које учествују у репрезентацији садрже различите y_i-ове. Независност Q_1,..., Q_k у тврдњењу 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.

Теорема 0.3.3. *Нека случајни векто* 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* $A_1 A_2 = 0$.

0.4 Задаци

0.1. Нека су $X_1, ..., X_n$ 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 $i = 1, 2, ..., n$. Наћи коваријациону матрицу случајног вектора $Y = (Y_1, ..., Y_n)^T$.

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

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

0.4. Нека је $(X_1, Y_1), ..., (X_n, Y_n)$ прост случајан узорак из дводимензионалне 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 θ у функцији од параметара расподеле и константе δ.

0.5. Доказати теорему 0.3.3.

Поглавље 1

Линеарни модели

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

Како се променом једне или више енезависних случајних променљивих 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" матхематичари 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 1.0.1. *Регресија је зависност једне случајне промењљиве од друге (или више њих). Регресиони модел је математички модел који описује ту зависност.*

Дефиниција 1.0.2. *Случајна величина $f(X) = E(Y|X)$ назива се регресиона функција, при чему X може бити вишедимензиона случајна величина.*

Следећа теорема оправдава облик функције регресије.

Теорема 1.0.1.

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

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

Доказ. □

Регресиона функција је права линија акко случајни вектор $(X, Y)^T$ има вишедимензионална 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 опицује зависност између Y и X у смислу средњеквадратног одступања.

Регресиони модел се може предцавити у облику

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

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

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

Добија се да је

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

па се коефицијенти a, 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 променљива говоримо о *контролисаној регресији.*

1.1 Матрични запис ленеарног модела

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

где су

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

X се назива предиктор (независна променљива), а Y регресанд (зависна променљива). 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 ε су независни случајни вектори.

Уколико модел допушта и слободан члан онда се најчешће модел приказује у истом облуку при чему је матрица X

$$X = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}$$

и назива се *дизајн матрица*. Најприроднији начин да се оцене коефицијенти модела β је метод најмањих квадрата, односно

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

Уколико је $R(X) = p$ 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

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

Посматрајмо суму квадрата одступања од модела

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

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

$$\begin{aligned} E(SSE) = E(\sum_{ij} m_{ij}Y_iY_j) &= \sum_{ij} m_{ij}E(Y_iY_j) = \sum_{ij} m_{ij}E(\varepsilon_i\varepsilon_j) \\ &= \sum_{i} m_{ii}E(\varepsilon_i^2) = \sigma^2 \sum_{i} m_{ii} \\ &= \sigma^2 tr(M) = \sigma^2(tr(I) - tr(H)) = \sigma^2(n - tr((X^TX)(X^TX)^{-1})) \\ &= \sigma^2(n - tr(I_{p+1})) = \sigma^2(n - p - 1). \end{aligned}$$

Одавде закључујемо да је непристрасна оцена за σ^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

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = &\sum_{i=1}^{n}(Y_i - \hat{Y})^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ SSTO = &SSE + SSR \end{aligned} \tag{1.1}$$

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

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

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

$$R(I - H) = tr(I - H) = n - p - 1.$$

Из формуле (1.1) 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 $R = |\rho_{xy}|$.

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 обзервација. 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, кеофицијенти модела се не могу јединствено одредити.

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

Приметимо да су резидуали ортогонални на простор генерисан са X.

$$e^T X = Y^T(I - H)X = Y^T(X - H^T X) = 0$$

Одавде је $e^T X u = 0$ за сваки вектор u.

Ако је X ортогонална матрица онда се оцена за β 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. Њена важна особина је садржана у следећој теореми.

Теорема 1.1.1 (Гаус-Марков). *Од свих непрестрасних линеарних оцена параметра β, оцена методом најмањих квадрата има најмању дисперзију.*

$$E(\hat{\beta}|X) = (X^TX)^{-1}X^TE(Y|X) = (X^TX)^{-1}X^TX\beta = \beta$$
$$D(\hat{\beta}|X) = (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} = (X^TX)^{-1}\sigma^2$$

Нека је $\tilde{\beta} = AY$. Треба показати да је $D(\tilde{\beta}|X) - D(\hat{\beta}|X) \geq 0$.
Нека је $B = (X^TX)^{-1}X^T$ и нека је $A = B+C$. Из услова непреистрасности је $E(\tilde{\beta}|X) = AX\beta = \beta$. Односно $I = AX = (B + C)X$. Како је $BX = I$ закључујемо да је $CX = 0$.
Даље је

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

Како је

$$CB^T = CX(X^TX)^{-1} = 0$$
$$BC^T = (X^TX)^{-1}X^TC^T = (X^TX)^{-1}(CX)^T = 0$$

следи тврђење.

Последица 1.1.1. *Уколико $tr((X^TX)^{-1}) \to 0$, кад $n \to \infty$ $\hat{\beta}$ је постојана оцена параметра β.*

Напоменимо да ћемо оцене које су најбоље линеарне непристрасне оцене означавати са BLUE (Best Linear unbiased estimate).

Слично се показују и следећа уопштења ове 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 1.1.2. *$l^T\hat{\beta}$ је BLUE за $l^T\beta$.*

Теорема 1.1.3. *$L^T\hat{\beta}$ је BLUE за $L^T\beta$.*

1.2 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 хопотезу да нема утицаја слободног члана узећемо да је $C = (1, 0, ..., 0)$ и $\gamma = 0$. Или, уколико желимо да тестирамо нулту хипотезу да ниједан предиктор нема утицаја на Y, $C = (0, I_p)$ и $\gamma = 0$.

Претпоставимо да ε има $N(0, \sigma^2I_n)$ расподелу. За тестирање наше нулте хипотезе користићемо тест количника веродостојности.

1.2.1 Метод максималне веродостојности

Претпоставимо да ε 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 β поклапа са оценом методом најмањих квадрата, односно $\hat{\beta} = (X^T X)^{-1}(X^T X)Y$. Слично, оцена за σ^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 σ^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 $S(\beta) = \frac{(Y-\beta X)^T (Y-\beta X)}{n}$ уз услов $C\beta - \gamma = 0$. 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 нормлних једначина чије решење тражимо је

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

Добија се да је

$$\hat{\beta}_0 = \hat{\beta} + (XX^T)^{-1}C^T[C(XX^T)^{-1}C^T]^{-1}(\gamma - C\hat{\beta})$$
$$\hat{\sigma}_0^2 = \frac{(Y - \hat{\beta}_0 X)^T(Y - \hat{\beta}_0 X)}{n}$$
$$= \frac{1}{n}\Big((Y - \hat{\beta}X)^T(Y - \hat{\beta}X) + (C\hat{\beta} - \gamma)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta} - \gamma)\Big)$$

Максимална вредност функције веродостојности у овом случају је

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

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

$$\lambda = \frac{\max\limits_{H_1} L(\beta, \sigma^2)}{\max\limits_{H_0} L(\beta, \sigma^2)} = \frac{L(\hat{\beta}, \hat{\sigma}^2)}{L(\hat{\beta}_0, \hat{\sigma}_0^2)} = \Big(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\Big)^{\frac{n}{2}}$$
$$= \Big(1 + \frac{(C\hat{\beta} - \gamma)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta} - \gamma)}{e^Te}\Big)^{\frac{n}{2}}$$

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

Претпоставка о нормалном моделу је јако важна за коришћење теста количника веродстојности јер се тест статистика може приказати у погодном облику као количник две независне статистике са χ^2 расподелама.

Јасно је да $\hat{\beta} = (X^TX)^{-1}XY$ има нормалну $N(\beta, \sigma^2(X^TX)^{-1})$. Одавде закључујемо да

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

Одавде, уколико важи H_0 закључујемо да

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

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

$$Q = (C\hat{\beta} - C\beta)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta} - C\beta)$$
$$= (C(X^TX)^{-1}X^T\varepsilon)^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}X^T\varepsilon$$
$$= \varepsilon^TX(X^TX)^{-1}C^T[C(X^TX)^{-1}C^T]^{-1}C(X^TX)^{-1}X^T\varepsilon$$
$$= \varepsilon^TP\varepsilon,$$

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

Даље, применом Кохранове теореме добијамо да

$$e^Te = SSE = \varepsilon^TM\varepsilon \sim \sigma^2\chi^2_{n-p-1}$$

Да бисмо нашли расподелу тест статистике потребно је још да покажемо да је $PM = 0$ и применимо теорему 0.3.3. Из једнакости $X^TM = 0$ добијамо $PM = 0$ и закључујемо да су бројилац и имеилац у тест статистици независне случајне величине па

$$\frac{\frac{Q}{m}}{\frac{e^Te}{n-p-1}} \sim F_{m,n-p-1}. \tag{1.2}$$

Пример 1.2.1. *Претпоставимо да тестирамо нулту хипотезу* $H_0 : \beta_1 = \beta_2 = \cdots \beta_p = 0$ *против комплементарне хипотезе. Тада је*

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

Одавде, је

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

Применом Кохранове теореме, или теореме (0.3.3) *добија се да*

$$\frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F_{p,n-p-1}.$$

Пример 1.2.2. *У случају да желимо да тестирамо хипотезу да је* $\beta_k = 0$ *већ смо видели како да одаберемо матрицу* C*. Тада*

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

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

1.2.2 Интервал поверења за β

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

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{\sigma^2} \sim \chi^2_{p+1} \qquad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

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 $V_k = [(X^T X)^{-1}]_{kk}$. 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). $D(\hat{\beta}_k|X) = \sigma^2 V_k$. Због непристрасности оцене и нормалности

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

Како $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}$ има χ^2_{n-p-1} онда

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

Пример 1.2.3. *Посматра се зависност GPA чије вредности се налазе у интервалу [0,4] од вербалног дела SAT−а (SATV) и матхематичког 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~SAT+SATM))

Call: lm(formula = GPA~SAT4 + 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
SAT          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 је стандардна процедура и део уграђене функције lm. На основу табеле изнад можемо закључити да резултати на SATM не утичу на GPA.

Лема 1.2.1 (Бонферонијеве неједнакости)**.** *Нека су $A_1, ..., A_k$ догађаји дефинисани на истом простору вероватноће. Тада је*

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

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 1.2.4. *Желимо да направимо интервале поверења за линеарне комбинације коефицијената модела $a_1^T\beta$, $a_2^T\beta, ..., a_l^T\beta$. Један начин је да формирамо одговарајуће елипсоиде користећи статистику* (1.2)*. С друге стране, $a_i^T(\hat{\beta} - \beta) \sim N(0, a_i^T(X^TX)^{-1}a_i\sigma^2)$, односно*

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

1.2.3 Интервали предвиђања

Нека је x_0 "тачка" у којој вршимо предвиђање. Тада је $\hat{Y}_0 = x_0^T\hat{\beta}$. Одавде је

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

Из особина нормалне расподеле закључујемо да

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

па резоновањем као у претходном поглављу закључујемо да

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

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

Нека Y_0 вредност у x_0. Тада је $Y_0 = x_0\beta + \varepsilon_0$, где је ε_0 независно од $Y_1, ..., Y_n$. Одавде је

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

Одавде закључујемо да

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

одакле је

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

1.3 Категорички предиктори

Посматрајмо следећи линеарни модел

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad i = 1, 2...n \tag{1.3}$$

$$X_{i1} = \begin{cases} 1, & \text{за } i = 1, ..., n_1 \\ 0, & \text{за } i = n_1 + 1, ..., n \end{cases} \tag{1.4}$$

$$\tag{1.5}$$

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

Означимо се $\mu_1 = \beta_0$ и $\mu_2 = \beta_0 + \beta_1$. Тада модел (1.3) се може представити у облику

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

Тестирање нулте химпотезе $H_0 : \mu_1 = \mu_2$ је еквивалентно тестирању хипотезе $H_0 : \beta_1 = 0$

Пример 1.3.1. *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 оружају 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

Увели смо помоћне променљиве L_2 *и* L_2 *тако да пар* $(L_1, L_2) = (0, 0)$ *означава државне агенције,* $(0, 1)$ *приватне непрофитне агенције а пар* $(1, 1)$ *приватне профитне агенције.*

Call: lm(formula = skala ∼L1 + L2)
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*	*****
L1	*4.138*	*3.595*	*1.151*	*0.257*	
L2	*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* β_1 *и* β_2 *безначајни. У моделу је просечна вредност оцене за државне агенције узета за референти ниво, па* 66.119 *представља просечну вредност оцене државних агенција. Повећања оцене (у односу на референти ниво) за* 4.138 *за приватној непрофитне агенције и* $4.138 + 2.835$ *за приватне профитне агенције, нису значајна.*

Наравно, у моделу поред категоричких могу постојати и нумеричке променљиве.

Подаци су преузети из књиге [3] па ћемо задржати оригиналне ознаке.

Ознака	Значење
Price	цена куће у хиљадама дулара
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, m уколико је за m аутомобила
CDN	да ли су потребна додатна улагања (1, за да)
L1	1 уколико је у зони A , 0 иначе
L2	1 уколико је у зони B, 0 иначе

Подаци су сакупљени о кућама које се налазе у 3 зоне A, B, C па је припадност зони кодирана са две помоћне променљиве $L1$ и $L2$.

	Price	BDR	FLR	FP	RMS	ST	LOT	TAX	BTH	CON	GAR	CDN	L1	L2
1	53.00	2.00	967.00	0.00	5.00	0.00	39.00	652.00	1.50	1.00	0.00	0.00	1.00	0.00
2	55.00	2.00	815.00	1.00	5.00	0.00	33.00	1000.00	1.00	1.00	2.00	1.00	1.00	0.00
3	56.00	3.00	900.00	0.00	5.00	1.00	35.00	897.00	1.50	1.00	1.00	0.00	1.00	0.00
4	58.00	3.00	1007.00	0.00	6.00	1.00	24.00	964.00	1.50	0.00	2.00	0.00	1.00	0.00
5	64.00	3.00	1100.00	1.00	7.00	0.00	50.00	1099.00	1.50	1.00	1.50	0.00	1.00	0.00
6	44.00	4.00	897.00	0.00	7.00	0.00	25.00	960.00	2.00	0.00	1.00	0.00	1.00	0.00
7	49.00	5.00	1400.00	0.00	8.00	0.00	30.00	678.00	1.00	0.00	1.00	1.00	1.00	0.00
8	70.00	3.00	2261.00	0.00	6.00	0.00	29.00	2700.00	1.00	0.00	2.00	0.00	1.00	0.00
9	72.00	4.00	1290.00	0.00	8.00	1.00	33.00	800.00	1.50	1.00	1.50	0.00	1.00	0.00
10	82.00	4.00	2104.00	0.00	9.00	0.00	40.00	1038.00	2.50	1.00	1.00	1.00	1.00	0.00
11	85.00	8.00	2240.00	1.00	12.00	1.00	50.00	1200.00	3.00	0.00	2.00	0.00	1.00	0.00
12	45.00	2.00	641.00	0.00	5.00	0.00	25.00	860.00	1.00	0.00	0.00	0.00	0.00	1.00
13	47.00	3.00	862.00	0.00	6.00	0.00	25.00	600.00	1.00	1.00	0.00	0.00	0.00	1.00
14	49.00	4.00	1043.00	0.00	7.00	0.00	30.00	676.00	1.50	0.00	0.00	0.00	0.00	1.00
15	56.00	4.00	1325.00	0.00	8.00	0.00	50.00	1287.00	1.50	0.00	0.00	0.00	0.00	1.00
16	60.00	2.00	782.00	0.00	5.00	1.00	25.00	834.00	1.00	0.00	0.00	0.00	0.00	1.00
17	62.00	3.00	1126.00	0.00	7.00	1.00	30.00	734.00	2.00	1.00	0.00	1.00	0.00	1.00
18	64.00	4.00	1226.00	0.00	8.00	0.00	37.00	551.00	2.00	0.00	2.00	0.00	0.00	1.00
19	66.00	2.00	929.00	1.00	5.00	0.00	30.00	1355.00	1.00	1.00	1.00	0.00	0.00	1.00
20	35.00	4.00	1137.00	0.00	7.00	0.00	25.00	561.00	1.50	0.00	0.00	0.00	0.00	0.00
21	38.00	3.00	743.00	0.00	6.00	0.00	25.00	489.00	1.00	1.00	0.00	0.00	0.00	0.00
22	43.00	3.00	596.00	0.00	5.00	0.00	50.00	752.00	1.00	0.00	0.00	0.00	0.00	0.00
23	46.00	2.00	803.00	0.00	5.00	0.00	27.00	774.00	1.00	1.00	0.00	1.00	0.00	0.00
24	46.00	2.00	696.00	0.00	4.00	0.00	30.00	440.00	2.00	1.00	1.00	0.00	0.00	0.00
25	50.00	2.00	691.00	0.00	6.00	0.00	30.00	549.00	1.00	0.00	2.00	1.00	0.00	0.00
26	65.00	3.00	1023.00	0.00	7.00	1.00	30.00	900.00	2.00	1.00	1.00	0.00	1.00	0.00

Пример 1.3.2. *Направићемо модел у коме су сви параметри једне некретнине укључени у цену.*

```
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 + L1 + L2)

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    0.0063      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 некретнине стварно утичу на цену исте. Али, уколико желимо да видимо како утиче зона у којој је кућа смештена, то не можемо закључити из ове табеле. Требало би форамлно тестирати нулту хипотезу да су коефицијенти уз L_1 и L_2 нула против алтернативе да је један од коефицијената различит од нуле. За то можемо користити тест статистику (1.2). *За то тестирање користићемо функцију* anova *чији су аргументи два модела,*

први код кога су коефицијенти L_1 и L_2 различити од нуле и други у коме то нису (у нашем случају модел m.1).

```
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.

1.4 Провера коректности 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.

1.4.1 Аутлајери, тежинске и утицајне тачке

Показали смо да је $e = (I - H)\varepsilon$, односно $Cove = (I - H)\sigma^2$. Јасно је да је $D(e_i) = \sigma^2(1 - h_i)$, где је $h_i = H_{ii}$. Број h_i се назаива моћ, тежина (*leverage*) тачке. Што је веће h_i дисперзија i-тог резидуал ће бити мања па ће права бити ближа Y_i. Може се показати да је $h_i = x_i^T(XX^T)^{-1}x_i$.

Приметимо још и следеће, што је h_i мање реиздуали ће се понашати више као стварна грешка. 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. Уместо резидула често ћемо посматрати стандардизоване резидуале internally studentized

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

као и студентизоване резидуале externally studentized дефинисане са

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

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

Може се показати да је

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

Тада су резудали модела у i-тој обзервацији (уочимо да је се овде ради о грешци прдвиђања)

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

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 обзервације. Сада важи

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

Како је

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

Еwтерно студентизоване резидуале можемо приказати у облику

$$e_i^\star = e_i\Big(\frac{n-p-2}{\hat{\sigma}^2(1-h_i)(n-p-1)-e_i^2}\Big)^{\frac{1}{2}}$$

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

Сада се може дефинисати једноставан тест за одређивање аутлајера заснован на томе да екстерно студентизоване резудале карактерише велика апосолутна вредност. Наиме, тест статистика ће нам бити баш $e_i^\star$ која, ако је модел коректан, тј. ако i-та обзервација 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 $\frac{\alpha}{k}$.

Утицајне тачке су оне чијим изостављањем би се модел значајно променио. Једна од мера утицаја је и Куково растојање. Оно је дефинисано са

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

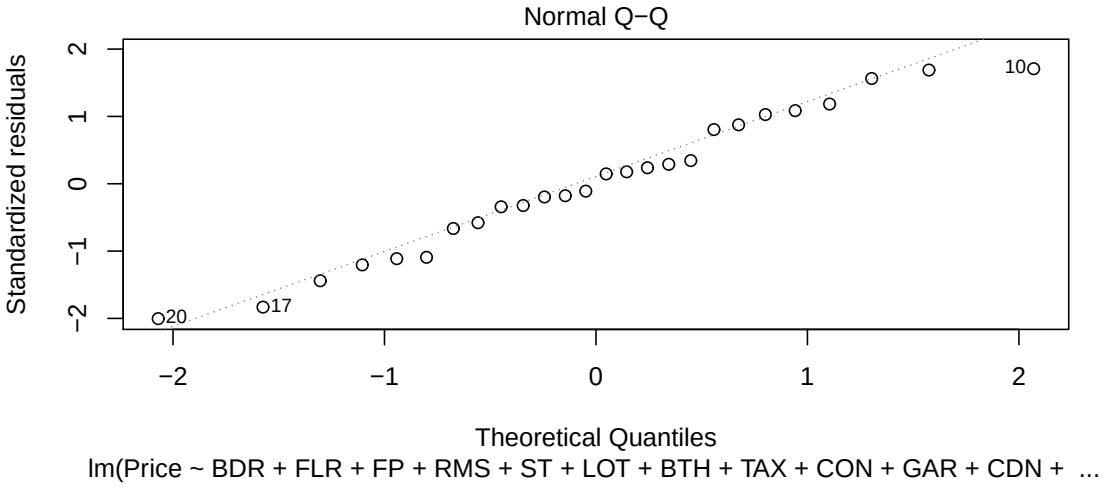
Видимо да је у ово растојање инкорпорирано и одступање од модела и тежина сваке тачке. Договор је да се тачке за које је Кукокво растојање веће од 1 сматрају утицајним, али да треба обратити пажњу и на оне са растојањем већем од 0.5. До закључка се може доћи поређењем са квантилима Фишерове $F_{p+1,n-p-1}$.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

1.4.2 Проверавање номалности

-QQplot

-тестови нормалности (неки од тестова заснован на емпиријској функцији расподеле: Колмогоров-Смирнов, Андерсон-Дарлинг (AD)), Крамер-вон Мисесов (CM) итд, Шапиро Вилк (SW)...)

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 $\{W < c\}$.

Пример 1.4.1. *Тестираћемо да ли резидуали имају приближно нормалну расподелу.*

```
shapiro.test(m.1$residuals)

        Shapiro-Wilk normality test
data:  m.1$residuals
W = 0.98193, p-value = 0.9119
```

Тестираћемо стандардизоване резидуале модела $m.1$.

```
standardizovaniReziduali=m.1$residuals/sqrt(1-
influence(m.1)$hat)/summary(m.1)$sig

ks.test(standardizovaniReziduali,"pnorm",0,1)

    One-sample Kolmogorov-Smirnov test
    data: standardizovaniReziduali
D = 0.097172, p-value = 0.9469
alternative hypothesis: two-sided


goftest::ad.test(standardizovaniReziduali,"pnorm",0,1)

Anderson-Darling test of goodness-of-fit
Null hypothesis: Normal distribution
data: standardizovaniReziduali
An = 0.27609, p-value = 0.9546


goftest:: cvm.test(standardizovaniReziduali,"pnorm",0,1)

Cramer-von Mises test of goodness-of-fit
Null hypothesis: Normal distribution
data: standardizovaniReziduali
omega2 = 0.034875, p-value = 0.9603


shapiro.test(standardizovaniReziduali)


Shapiro-Wilk normality test

data: standardizovaniReziduali
W = 0.97035, p-value = 0.6323
```

Уколико је модел коректан стандардизовани резидуали не би требало да буду корелисани и сви би требало да имају исту диспрезију. Под претпоставком да је наша претпоставка о моделу тачна стандардизовани резидуали ће имати t_{n-p-1}.

У случају да нормалност није задовољена, под неким условима, тестирања се могу радити као и до сада.

1.4.3 Провера хомоскедастичности

-Најједноставније је да се дође до закључка из графика оцењене вредности-резидуали. Постоји неколико статистичких тестова за проверу хомоскедастичности али се на томе нећемо задржавати.

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

Трансформације промењливих

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

$$f(Y) \approx f(m) + (Y - m)f'(m)$$
$$D(f(Y)) \approx (f'(m))^2 D(Y).$$

Да би стабилизовали дисперзију потребно је да је

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

односно

$$f(m) = \int \frac{c}{\sqrt{D(y)}}$$

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

1.4.4 Тежинска регресија

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

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

Поставља се питање како одабрати одговарајуће тежине. Уколико је $D(\varepsilon_i) \sim x_i$ 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 $w_i = n_i$.

1.4.5 Провера некорелисаности

Дурбин-Вотсонов тест се често користи за тестирање некорелисаности. Тест статистика је

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

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.

Генерализовани метод најмањих квадрата

Генерализовани метод најмањих квадрата је заправо уопштење тежинске регресије.

Претпоставимо да је $Cov(e) = \Sigma\sigma^2$ где је Σ симетрична, позитивно дефинитна матрица, тада је оцена за β генерализованом методом најмањих квадрата дата са

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

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

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

Када је Σ непознато постоје разне методе за оцену исте. О томе се може више сазнати у [3].

1.5 Још неке трансформације зависне променљиве и предиктора

1.5.1 Трансформације зависне променљиве

Бокс Коксова траснормација

Поред оних које смо напоменули, за стабилизацију дсиперзију, вероватно највише коришћене су Бокс-Коксове реансформације.

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

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

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

$$\log L(\lambda, \sigma|Y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{(Y^{(\lambda)}-X\beta)^T(Y^{(\lambda)}-X\beta)}{2\sigma^2} + (\lambda-1)(\prod_{i=1}^{n} Y_i) \tag{1.6}$$

Оцене за β и σ^2 се добијају као и до сада, а λ је она вредност која максимизира функцију

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

С обзиром на то да модел ипак треба да буде интерпретабилан, треба узети негу смислену вредност за λ, а опет "довољно блиску" са оцењеном вредности. Коришћењем Вилксове теореме добија се да $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$ има граничну χ_1^2 расподелу.
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 (1.6). Тада $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$ има граничну χ_2^2 расподелу.

Мултипликативни модели

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

$$Y_i = E(Y_i|X)\varepsilon_i, \quad i = 1, 2..., n, \tag{1.7}$$

где је

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

Одавде је природна претпоставка да су случајне грешке независне од предиктора и да је $E(\varepsilon_i) = 1$. Логаритмовањем (1.7) добијамо модел

$$\log(Y_i) = \log A + \sum_{j=1}^{p} \beta_j \log(X_{ij}) + \log(\varepsilon_i), \quad i = 1, 2..., n. \tag{1.8}$$

Добили смо класичан линеарни модел. Међутим из претпоставке $E\varepsilon_i$ 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 (1.8) боље приказати у облику

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

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

1.6 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 полином малог степена па додавати чланове вишег реда док се не добију "статистички незначајни коефицијенти", или кренути од полинома веће степена па смањивати док се не добије значајан коефицијент уз предиктор(е) највеће степена.

Мане ове врсте регресије:

1. Број променљивих значајно расте за повећавањем степена полинома.

2. Екстраполација није добра.

3. Матрица X^TX 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 $X_1, .., X_k$ уврсте центриране променљиве $X_1 - \bar{X}_1, .., X_k - \bar{X}_k$.

1.7 Сплајнови

1.8 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 обзервација).** 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 убацити у модел.

- **Убацивање у модел предиктора између којих природно постоји линеарна веза.** На пример, убацити у модел предикторе БРУТО плата, НЕТО плата и ТАРА плата.

- **Узорак на коме се врше обзервације је условљен неким ограничењима у популацији.** Узорковање вршимо из "потпопулације" на којој су предиктори веома корелисани.

Ако између предиктора постоји линеарна зависност, матрица X^TX није инвертибилна и онда оцена за β 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 ту може направити грешка приликом њиховог избора.

Нека су λ_j сопствене вредности матрице $X_{(s)}^T X_{(s)}$

$$\sum_{i=0}^{p} \lambda_i = tr(X_{(s)}^T X_{(s)}) = p + 1$$

Нека је $\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$

Сматра се да вредности веће од 30 треба да упуте на даље испитивање тих обзервација.

1.9 Анализа главних компоненти

Један од начина да решимо проблем мултиколинеарности је да избацмо 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. Нека је $Z = AX$ линеарна трансформација предиктора. Тада је $D(Z_i) = a_i^T \Sigma a_i$. Без умањења општости можемо претпоставити да је максимална дисперзија, уз услов да је $|a_i| = 1$ баш $D(Z_1)$. Z_1 ћемо звати *прва главна компонента*. Без умањења општости можемо претпоставити да је a_2 ред матрице A за који је $|a_2| = 1$, $a_2 X_2$ је ортогонално са Z_1 и $a_2^T \Sigma a_2$. $Z_2 = a_2^T X$ зваћемо *другом главном компонентом*. Поступак понављамо, при чему је свака од наредних главних компоненти ортогонална на све претходне.

Лема 1.9.1. *Нека је Σ коваријациона матрица случајног вектора X. Нека су $\lambda_1 \geq \lambda_2 \cdot \geq \lambda_p \geq 0$ њене сопствене вредности. Тада је i-та главна компонента дата са $Z_i = v_i^T X$, за $i = 1, 2, ..., p$, где је v_i i-ти сопствени вектор.*

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

Лема 1.9.2. *Нека је $Z = VX$. Тада је $\sum_{i=1}^{p} D(X_i) = \sum_{i=1}^{p} D(Z_i) = \sum_{i=1}^{p} \lambda_i$.*

Доказ. $D = V \Sigma V^T$
$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} D(Z_i) = tr(D) = tr(V \Sigma V^T) = tr(\Sigma V V^T) = tr(\Sigma) = \sum_{i=1}^{p} D(Z_i)$.

□

Последица ове леме је да ротирањем координатног система нисмо променили укупан варијабилитет система, као и да је удео објашњеног вариабилитета i-том главном компонентом $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ за $i = 1, 2.., p$. 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 1.9.3. *Коефицијент корелације између Z_i и X_k је*

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

Даље, како је $\Sigma = V^T D V$ закључујемо и да је $D(X_k) = \sum_{i=1}^{p} \lambda_i v_{ik}^2$.

Проблем који се јавља у интерпретацији главних компоненти последица је њихове осетљивости на различите мерне скале полазних предиктора. Најједноставнији 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 један предиктор доминирати првом главном компонентом. Овај проблем се може решити стандардизацијом модела, када се коваријациона матрица поклапа са корелационом матрицом.

Претпоставимо да се ради о стандардизованом моделу. Нека је $S_i = \frac{X_i - \bar{X}_i}{\sqrt{(n-1)\tilde{S}_i^2}}$, за $i = 1, 2, ..., p$ и $Y_{(s)} = Y - \bar{Y}$. Стандардизован модел $Y_{(s)} = S\delta + \varepsilon$ се може приказати у облику $Y_{(s)} = Z\eta + \varepsilon$ где су Z главне компоненте добијене од стандардизованих предиктора (полазни модел

се може приказати у облику $Y = S_0 + S\delta + \varepsilon$, на предавању видели да је оцена за слободан члан $\bar{Y}$ шо оправдава птретпоставку да је полазни модел еквивалентан стандардизованом).

Тада је оцена непознатог параметра η добијена методом најмањих квадрата дата са

$$\hat{\eta} = (Z^T Z)^{-1} Z^T Y_{(s)} = D^{-1} Z^T Y_{(s)}.$$

Одавде је $D(\eta_j) = \frac{\sigma^2}{\lambda_j}$. И одавде видимо да мале сопствене вредности узоркују велике дисперзије оцене коефицијената. Сада је природно да те главне компоненте избацимо. Пошто су све главне компоненте међусобно ортогоналне, избацивање једне од компоненти неће имати за последицу промену оцена других коефицијената, и оцене ће задржати особину непристрасности.

Из претходног текста може се уочити да је у случају стандардизованог модела је

$$\sum_{i=0}^{p} \lambda_i = tr(S^T S) = p,$$

Претпоставимо да смо r главних компоненти одлучили да задржимо, а преосталих $p-r$ да избацимо. Зато ћемо приказати главне компоненте приказати у облику $Z = \begin{pmatrix} Z_r \\ Z_{p-r} \end{pmatrix}$ где су Z_r задржане компоненте а Z_{r-p} избачене компоненте. Аналогно ћемо приказати и вектор коефицијената $\eta = (\eta_r^T \eta_{n-r}^T)^T$.. Како је $Z = SV$ закључујемо да је $Z_{(r)} = SV_r$ (V_r је матрица која се састоји од првих r колона матрице V). Нови модел се може приказати у облику

$$Y_{(s)} = Z_r \eta_r + \tilde{\varepsilon},$$

односно

$$Y_{(s)} = S\delta_r + \tilde{\varepsilon},$$

где је $\delta_r = V_r \eta_r$.

Непознате параметре δ оценићемо са $V_r \hat{\eta}_r$. Испитајмо непристрасност оцене.

$$\begin{aligned} E(V_r \hat{\eta}_r) &= V_r E(\hat{\eta}_r) = V_r \eta_r = (V_r \ \mathbf{0}) \begin{pmatrix} \eta_r \\ \mathbf{0} \end{pmatrix} \\ &= (V_r \ \mathbf{0}) (V_r \ \mathbf{0})^T \delta = V_r V_r^T \delta = \delta - V_{p-r} V_{p-r}^T \delta = \delta - V_{p-r} \eta_{n-r}. \end{aligned}$$

Избацивање главних компоеннти има за последицу пристрасност оцене за δ_r. Нека је $\hat{\delta}$ оцена за δ_r кад је $r = p$. Тада је

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

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

Дакле, овим поступком смањујемо дисперзију оцене коефицијената али губимо непристрасност оцене.

Осврт на детекцију мултиколинеарности:

Нека је условни број (conditional number) $\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}$

Сматра се да вредности веће од 30 треба да упуте на даље испитивање тих обзервација.

1.10 Назубљена регресија

Енглески назив за ову регресију је *Ridge regression*.

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

$$\hat{\delta}_c = (S^T S + cI)^{-1} S^T y_0. \tag{1.9}$$

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

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

Може се показати да се ова оцена добија методом најмањих квадрата кад додамо у систем "вештачке предикторе" који не утичку на вредности зависне променљиве (односно за које је вредност зависне променљиве 0). Више информација се може наћи у у [3] (стр. 257). Нађимо пристрасност оцене

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

Одговарајућа коваријациона матрица оцене је

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

Одавде се добија да је сума дисперзија појединачних компоненти

$$tr((SS^T + cI)^{-1} S^T S (SS^T + cI)^{-1} \sigma^2) = \sigma^2 tr((SS^T + cI)^{-2} S^T S)$$
$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i (1 + c\lambda_i)^2}$$

Приметимо да је то мање од суме дисперзија појединачних компоненти оцене методом најмањих квадрата $\sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$

ОДАБИР ПАРАМЕТРА c

Други начин да се овај естиматор интерпретира је он заправо представља оцену најмањих квадрата када за параметре постоји горње ограничење, односно решава се проблем

$$\min_{\delta} (y_0 - S\delta)^T (y_0 - S\delta), \;\; ||\delta|| \leq C,$$

за неко C које се може изразити у функцији од c. Даље, тај проблем је еквивалентан тражењу минимума функције

$$(y_0 - S\delta)^T (y_0 - S\delta) + c||\delta|| \tag{1.10}$$

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

1.11 LASSO

Енглески назив за ову регресију је Least Absolute Shrinkage and Selection Operator.

За разлику од оцене назубљеном регресијом оцена овом методом "казнена функција" је дата $L1$ нормом оцене коефицијената, односно

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

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.

1.12 Задаци

1.1. Доказати теореме 1.1.2 и 1.1.3.

1.2. Посматрајмо моделе

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

где је

1. $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$,
2. $Y^\star = \Gamma Y$, $X^\star = \Gamma X$, $\varepsilon^\star = \Gamma\varepsilon$.

Показати да се оцене коефицијената β на основу првог и другог модела, методом најмањих квадрата, поклапају, као и оцене дисперзије σ^2.

1.3. Показати да уколико важи алтернативна хипотеза количник (1.2) има померену Фишерову расподелу $F_{m,n-p-1}(\delta)$ и одредити δ.

1.4. Извести пормулу за h_{ij} у случају просте регресије, у функцији од X_i, X_j и S_x^2.

1.5. У пакету PASWR налази се база URLadress у којој је 30 обзервација о количини сачуваних података и времена конектованог на Интернет. Испитати да ли постоји линеарна веза између ових података. Формирати модел и проверити његову коретност. Графички представити одговарајуће 95% интервале поверења и предвиђања. Уколико је потребно избацити аутлајере из скупа обзервација.

1.6. Показати да у случају просте линеарне регресије, бар један елемент матрице $\frac{X^TX}{n}$ тежи нули кад n тежи бесконачности, као и да h_i тежи нули.

1.7. Претпоставите да имате линеарни модел код кога је слободан члан нула, а дизајн матрица је $X = I_n$.
а) Показати да је оцена назубљеном регресијом дата са

$$\hat{\beta}_i = \frac{y_i}{1+c},$$

где је c подешавајући параметар (види (1.9)).
б) Наћи LASSO оцену непознатих параметара.

Литература

[1] J. J. Faraway. *Linear models with R.* CRC press, 2014.

[2] J. Neter, M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[3] A. Sen and M. Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.