Универзитет у Београду

Математички факултет

Мастер рад

Алтернативне методе регресије

Ментор:

Др Јелена Јоцковић

Студент:

Андреа Стојић

Бр. индекса 1160/15

Београд, септембар 2016.

Садржај:

1. Увод

Појам регресије се везује за утврђивање међусобних односа између две или више појава. Може нас, на пример, интересовати зависност између времена проведеног у спремању испита и добијене оцене на испиту, зарада запослених и њиховог образовања, каматне стопе и понуде новца... Како бисмо утврдили да ли су и у којој мери ове појаве зависне, правимо регресиони модел. Регресиони модел нам може послужити за прогнозирање и предвиђање појава. Тај модел може да буде линеарни и нелинеарни. Ми ћемо се бавити само линеарним моделима.

Линеарни регресиони модел је модел код кога је регресиона функција линеарна. Постоје једноструки линеарни регресиони модели, помоћу којих се приказује зависност између две променљиве, и вишеструки линеарни регресиони модели, помоћу којих се приказује зависност између већег броја променљивих.

Једноструки линеарни регресиони модел се представља функцијом

$$Y = \alpha + \beta X + e$$

где је Y зависна променљива, X независна или објашњавајућа променљива, α и β непознати параметри а e случајна грешка односно резидуал.

Вишеструки линеарни модел се представља функцијом

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + e$$

где је Y зависна променљива, $X_1, \dots, X_n (n \geq 2)$ независне или објашњавајуће променљиве, $\beta_0, \beta_1, \dots, \beta_n$ непознати параметри а e случајна грешка.

За случајну грешку e се претпоставља да има очекивање 0. Она представља разлику између оцењене вредности регресионе функције и њене стварне вредности.

Претпоставимо да подаци прате линеарни тренд, односно да могу бити представљени линеарном функцијом $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + e$. Наш задатак је да нађемо оцене $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n$ непознатих параметара. Постоји велики број метода помоћу којих се могу оценити непознати параметри а неке од њих ће бити представљене у овом раду. Када се оцене непознати параметри добија се регресиона функција $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_n X_n$, која представља линеарни модел. Добар модел мора добро да представља податке и да буде што једноставнији. Понекад, неке од променљивих модела немају утицај на ефикасност модела, односно њиховим изостављањем из модела се не губи на квалитету модела. Како бисмо испитали да ли нека од објашњавајућих променљивих има утицај на зависну променљиву тестираћемо хипотезу $\beta_1 = \beta_2 = \cdots = \beta_n = 0$. Уколико дођемо до закључка да је један или више коефицијената $\beta_i, 1 \leq i \leq n$, једнак нули, одговарајуће променљиве ћемо избацити из модела и тако добити једнако добар али једноставнији модел.

Најпознатија метода за оцењивање непознатих параметара је метода најмањих квадрата. Међутим, постоје случајеви када се не препоручује коришћење ове методе, односно када ова метода не даје добре оцене непознатих параметара. У оваквим случајевима се користе алтернативне методе регресије:

1. Ако постоје подаци који значајно одступају од осталих података у узорку, аутлајери, најбоље резултате дају такозване робусне методе регресије. У овом раду ће бити представљене две овакве методе, Метода најмањих апсолутних одступања, у делу 3 и М-регресија, у делу 4.
2. Када случајне грешке немају нормалну расподелу препоручује се коришћење Непараметарске регресије која ће бити описана у делу 5.
3. Уколико имамо неке информације о подацима из претходних експеримената најбоље је користити Бајесову регресију, описану у делу 6.
4. Уколико постоји корелација између објашњавајућих променљивих препоручује се коришћење Риџ регресије описане у делу 7.

Кораци у моделирању одређених података су: конструкција модела, тестирање адекватности модела, модификовање модела уколико се установи да он није адекватан и упрошћавање модела избацивањем објашњавајућих променљивих које немају утицај на зависну променљиву.

Један од начина за тестирање адекватности модела је израчунавање коефицијента детерминације односно вредности R^2,

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{Total\ SS(corrected\ for\ mean)}$$

Вредност R^2 се налази између 0 и 1. Што је R^2 ближе броју 1 то је модел бољи. Код једноструког линеарног модела важи $R^2 = r^2$, где је r корелација између променљивих X и Y.

-Порекло речи регресија

Франсис Галтон (Francis Galton) је 1877.године, у Енглеској, представио рад „Типични закони наслеђа“, у коме је изложио концепт регресије. Он је открио везу између величине зрна грашка родитељске биљке и величине зрна грашка биљке потомка. Установио је да је ова веза приближно линеарна. Такође је утврдио да величина зрна „регресира“ ка средњој вредности. Овај феномен је назвао „регресија ка медиокритету“.

У даљем раду је дат опис шест метода регресије и њихова илустрација на одговарајућим примерима.

2. Метода најмањих квадрата

Метода најмањих квадрата је регресиона метода која се најчешће користи. Ову методу су открили Карл Фридрих Гаус[1] (Carl Friedrich Gauss) у Немачкој, око 1795. године и независно од њега Адриен Мери Лежандр[2] (Adrien Marie Legendre) у Француској, око 1805. године. Прва примена ове методе била је у астрономији и геодезији.

2.1 Оцењивање регресионе праве

Претпоставимо да нам је дат узорак или база података која садржи две променљиве чију зависност желимо да испитамо. Претпоставимо да је обим узорка једнак n, односно да имамо n опсервација . Једну опсервацију ових двеју променљивих посматраћемо као једну тачку. Идеја је да пронађемо регресиону праву која ће најбоље одговарати тачкама датим у узорку. Како бисмо донели одлуку о томе колико добро регресиона права $\hat{Y} = \hat{\alpha} + \hat{\beta}x$ описује наше податке , можемо да посматрамо величину резидуала $\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i), 1 \leq i \leq n$. Резидуали $\hat{e}_i$ представљају вертикално растојање одређене тачке из узорка од регресионе праве. Ми желимо да изаберемо $\hat{\alpha}$ и $\hat{\beta}$ тако да резидуали буду што мањи. Код методе најмањих квадрата укупна величина резидуала се мери помоћу израза $\sum_{i=1}^{n} e_i^2$. Оцене параметара α и β представљају они бројеви $\hat{\alpha}$ и $\hat{\beta}$ за које је $\sum e_i^2$ најмања, односно они $\hat{\alpha}$ и $\hat{\beta}$ за које је сума квадрата резидуала најмања. Формуле за добијање оцена $\hat{\alpha}$ и $\hat{\beta}$ су:

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad (2.1)$$

где $\bar{x}$ и $\bar{y}$ представљају редом средње вредности бројева x_i и y_i.

Формула $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ може бити написана као $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$. Одавде видимо да регресиона права садржи тачку $(\bar{x}, \bar{y})$. Формула за $\hat{\beta}$ може бити написана као

$$\hat{\beta} = \sum w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right), \qquad w_i = \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}.$$

Приметимо да $\frac{y_i - \bar{y}}{x_i - \bar{x}}$ представља нагиб праве која садржи тачке $(\bar{x}, \bar{y})$ и (x_i, y_i). Одатле видимо да нагиб $\hat{\beta}$ представља неку врсту средње вредности нагиба $\frac{y_i - \bar{y}}{x_i - \bar{x}}$. Прецизније, $\hat{\beta}$ представља пондерисану средњу вредност. Приметимо да су пондери w_i ненегативне вредности и да је њихова сума једнака 1.

---

1 Карл Фридрих Гаус (1777-1855) био је немачки математичар и научник који је дао значајан допринос у многим пољима, укључујући торију бројева, анализу, диференцијалну геометрију, геодезију, електростатику, астрономију и оптику. Сматра се једним од најутицајнијих математичара у историји.

2 Адриен Мери Лежандр (1752-1833) био је француски математичар који је дао значајан допринос математици. Добро познати Легендрови полиноми и Легендрова трансформација.

Претпоставимо да су a и b кандидати за оцене $\hat{\alpha}$ и $\hat{\beta}$. Сума квадрата резидуала је тада $\sum(y_i - a - bx_i)^2$. Посматрајући тај израз као функцију од a и b можемо наћи њен минимум, тако што ћемо наћи парцијалне изводе по a и b а затим их изједначити са 0. Решења тих једначина су $a = \hat{\alpha}$ и $b = \hat{\beta}$.

2.2 Тестирање хипотезе $\boldsymbol{\beta = 0}$

Уколико тачке из узорка представимо на графику, можемо видети да ли су наше променљиве X и Y зависне. Ако тачке прате линеарни тренд, можемо да претпоставимо да су променљиве зависне. Да бисмо испитали у којој мери су ове две променљиве зависне можемо да упоредимо модел $Y = \alpha + \beta X + e$ са моделом $Y = \alpha + e$. Ако први модел одговара подацима много боље него други модел, онда можемо да закључимо да су променљиве X и Y у великој мери зависне. Ову зависност можемо да испитамо и тестирањем хипотезе $\beta = 0$.

Приликом тестирања узећемо да нам је нулта хипотеза $\beta = 0$ а алтернативна $\beta \neq 0$.

Прво ћемо да израчунамо $\hat{\beta}$ помоћу формуле (2.1) а затим ћемо израчунати

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}} \tag{2.2}$$

Након тога ћемо да израчунамо оцене стандардног одступања за $\hat{\beta}$ тако што ћемо да заменимо $\hat{\sigma}$ са σ у формули

$$SD(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{2.3}$$

где је σ стандардно одступање популације грешака. Тест статистика је

$$|t| = \frac{|\hat{\beta}|}{ocena.SD(\hat{\beta})} \tag{2.4}$$

где је $ocena.SD(\hat{\beta})$ оцена стандардног одступања параметра $\hat{\beta}$.

(Користимо апсолутну вредност јер ћемо касније говорити о $t = \frac{|\hat{\beta}|}{ocena.SD(\hat{\beta})}$.)

p-вредност теста се добија из t расподеле са $n - 2$ степени слободе. p-вредност се рачуна као вероватноћа да апсолутна вредност случајне променљиве са овом расподелом буде већа или једнака са вредношћу $|t|$ добијеној помоћу формуле (2.4).

Најрелевантнија информација о β је оцена $\hat{\beta}$. Вредност оцене $\hat{\beta}$ би требало да укаже на то да ли је β једнако 0 или није. Ако је $\hat{\beta}$ удаљено од нуле онда скоро можемо да будемо сигурни да је $\beta \neq 0$.

Видимо да тест хипотезе $\beta = 0$ може бити заснован на тест статистици $|t|$ из (2.4). Веома велика вредност за $|t|$ нам говори да је $\hat{\beta}$ много више удаљено од 0 него што би било очекивано да буде ако је $\beta = 0$.

Посматрајмо формулу (2.3). Она нам указује на то да је $SD(\hat{\beta})$ мање када је σ мање, односно када је варијабилност случајних грешака мања. Ово има смисла јер ми можемо да оценимо β прецизније уколико су нам случајне грешке мање. Такође, $SD(\hat{\beta})$ је мање када је $\sum(x_i - \bar{x})^2$ веће, односно када су вредности x_i више распршене.

-Оцењивање вредности σ

Природна оцена стандардног одступања популације грешака је стандардно одступање резидуала $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$. Ова оцена је $\sqrt{\frac{\sum \hat{e}_i^2}{n-2}}$, али уместо да користимо овај израз, ми ћемо га мало модификовати и користити формулу (2.3). Модификовам израз има добру особину да је $\hat{\sigma}^2$ непристрасна оцена параметра σ^2.

Делилац $n - 2$ у формули (2.2) понекад зовемо и број степени слободе оцене $\hat{\sigma}^2$. Број 2 у изразу $n - 2$ значи да треба да оценимо два параметра α и β у циљу формирања резидуала $\hat{e}_i$.

-p-вредност

Рекли смо већ да је веома велика вредност за $|t|$ доказ да је $\beta \neq 0$ зато што уколико је $\beta = 0$ таква вредност за $|t|$ се тешко може добити. Другим речима, у зависности од тога колико је мало вероватна добијена вредност за $|t|$, када је $\beta = 0$, ми можемо да измеримо јачину доказа против хипотезе $\beta = 0$ које добијамо из посматраног узорка.

Како бисмо дефинисали p-вредност теста морамо направити разлику између два начина посматрања вредности $|t|$ дату изразом (2.4). Први начин посматрања вредности $|t|$ је да посматрамо $|t|$ као вредност добијену заменом посматраних вредности y_i у формулу (посматрана вредност $|t|$). Дрги начин је да посматрамо вредности y_i, а самим тим и вредност $|t|$, као случајне променљиве. p-вредност теста је вероватноћа, под претпоставко да је $\beta = 0$, да је случајна променљива $|t|$ већа или једнака него посматрана променљива $|t|$. Претпоставимо да је добијени модел тачан под претпоставком да је $\beta = 0$ и претпоставимо да сеодређени експеримент понавља бесконачан број пута. За сваки експеримент може да се израчуна тест статистика $|t|$. p-вредност је аритметичка средина ове замишљене колекције вредности $|t|$ која би била једнако велика или већа од стварне вредности $|t|$.

Ако је p-вредност јако мала, онда је мало вероватно да је $\beta = 0$, тако да можемо да закључимо да је $\beta \neq 0$. У супротном ћемо закључити да је $\beta = 0$. Оставићемо особи која врши експеримент да одлучи када је p-вредност јако мала. Већина статистичара сматра да је 0.01 јако мала вредност а неки сматрају чак и да је 0.1 јако мала вредност.

Када је p-вредност јако мала и када закључимо да је $\beta \neq 0$ онда смо сигурни у наш закључак, а када закључимо да је $\beta = 0$ ми нисмо сигурни да је тај закључак тачан али нам само тај закључак обезбеђује адекватан модел.

-Расподела за t

Да бисмо израчунали p-вредност морамо знати расподелу за $|t|$ или за t, када је $\beta = 0$. Тест статистика t је добијена из узорка, односно помоћу вредности x_i и y_i. Случајност вредности t потиче од случајности вредности y_i, јер су вредности x_i константе које нису случајне. Приметимо да расподела за y_i није у потпуности позната, чак и ако претпоставимо да је $\beta = 0$, јер је тада $y_i = \alpha + e_i$, α је непознати параметар, а за стандардно одступање случајних грешака, σ, се не може претпоставити да је познато, а самим тим и расподела за e_i није у потпуности позната. Посматрајући формулу за t можемо да приметимо да она зависи од y_i само у делу $y_i - \bar{y}$, а да се параметар α поништава у овим разликама. Штавише, параметар σ се поништава у количнику $\frac{|\hat{\beta}|}{ocena.SD(\hat{\beta})}$. Дакле, ако назначимо облик расподеле случајних грешака, ми можемо да сматрамо познатом расподелу за t без знања о стандардном одступању грешака e_i.

Ако назначимо да је облик расподеле нормалан (у облику звона), односно ако претпоставимо да ће нам одговарати нормалан линеарни регресиони модел, онда добијамо да је расподела за t, под условом да је $\beta = 0$, t-расподела са $n - 2$ степени слободе. (Степени слободе су повезани са оценом $\hat{\sigma}$ параметра σ.) Чак и када расподела случајних грешака није нормална, када је $\beta = 0$, расподела тест статистике t је близу t-расподеле са $n - 2$ степени слободе, под условом да је величина узорка n довољно велика.

2.3 Проверавање нормалности.

Тестови и оцене у методи најмањих квадрата дају оптималне резултате уколико се може претпоставити да случајне грешке имају нормалну расподелу. Уколико се не може претпоставити нормалност случајних грешака онда процедуре методе најмањих квадрата дају валидне резултате али далеко су од оптималних. Размотримо тестирање хипотезе $\beta = 0$ и претпоставимо да случајне грешке немају нормалну расподелу. Тест методе најмањих квадрата је тада и даље валидан (под условом да је обезбеђен велики обим узорка и да одступање од нормалне расподеле није превише велико), у смислу да је израчуната p-вредност приближно слична стварној p-вредности. Међутим постоје многи други тестови који дају боље резултате када је $\beta \neq 0$.

Уколико установимо да не можемо да претпоставимо нормалност, постоје два начина да решимо тај проблем. Први начин је да покушамо да модификујемо узорак, док је други начин да користимо неку алтернативну методу регресије која не захтева нормалност случајних грешака.

Постоји велики број графика и тестова за тестирање нормалности случајних грешака, али ми ћемо представити само график нормалне вероватноће. Стандардизовани резидуали су поређани у растући поредак и представљени су у графику заједно са њиховим оочекиваним вредностима у случају да они представљају резидуале n независних случајних променљивих са

стандардизованом нормалном расподелом. Ако је претпоставка о нормалносто тачна онда би график требао да буде приближно линеаран.

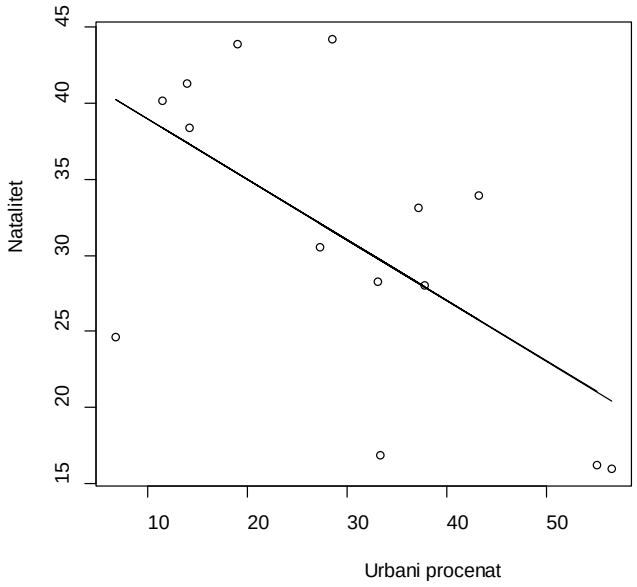
Пример 2.1 Наталитет

Посматрајмо табелу 2.1. У табели је дато 14 држава из Северне и Средње Америке, које имају преко милион становника (1985. година). У табели је дат наталитет сваке државе (број рођења у години на хиљаду људи) и урбани проценат (проценат људи који живе у градовима са преко 100000 становника).

Табела 2.1 Узорак „Наталитет“.

Држава	**Наталитет** (Y)	**Урбани проценат** (X)
Канада	16.2	55.0
Костарика	30.5	27.3
Куба	16.9	33.3
Доминиканска република	33.1	37.1
Ел Салвадор	40.2	11.5
Гватемала	38.4	14.2
Хаити	41.3	13.9
Хондурас	43.9	19.0
Јамајка	28.3	33.1
Мексико	33.9	43.2
Никарагва	44.2	28.5
Панама	28.0	37.7
Тринидад и Тобаго	24.6	6.8
САД	16.0	56.5

На слици 2.1 дат је график помоћу ког су представљене ове две променљиве и регресиона права добијена методом најмањих квадрата. Тачка на графику, која представља Тринидад и Тобаго се издваја од других тачака, али лако се уочава образац по коме можемо закључити да подаци имају опадајући тренд. Из тог разлога се чини разумним претпоставка да ће линеарни регресиони модел одговарати датим подацима.



Слика 2.1 Регресиона права узорка „Наталитет“, добијена методом најмањих квадрата.

Претпоставимо да модел

$$Y = \alpha + \beta X + e \tag{2.5}$$

одговара датим подацима. Наш задатак је да оценимо параметре α и β.

Убацујући податке из табеле (2.1) у формуле (2.1) добијамо оцене $\hat{\alpha} = 42.9905$ и $\hat{\beta} = 0.3989$. Одавде закључујемо да је оцењена регресиона права $\hat{Y} = 42.9905 + 0.3989X$.

Даље ћемо да тестирамо хипотезу $\beta = 0$. Број степени слободе је $12\ (14 - 2)$.

Како бисмо израчунали оцену за стандардно одступање оцене $\hat{\beta}$ као и вредност $|t|$, направили смо табелу 2.2.

Табела 2.2 Израчунавања коришћена за тестирање хипотезе $\beta = 0$.

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
55.0	16.2	21.05	-4.85	23.52	25.21	635.54
27.3	30.5	32.10	-1.60	2.56	-2.49	6.21
33.3	16.9	29.71	-12.81	164.10	3.51	12.3
37.1	33.1	28.19	4.91	24.11	7.31	53.44
11.5	40.2	38.40	1.80	3.24	-18.29	334.52
14.2	38.4	37.33	1.07	1.14	-15.59	243.05
13.9	41.3	37.45	3.85	14.82	-15.89	252.49
19.0	43.9	35.41	8.49	72.08	-10.79	116.42
33.1	28.3	29.79	-1.49	2.22	3.31	10.96
43.2	33.9	25.76	8.14	66.26	13.41	179.83
28.5	44.2	31.62	12.58	158.26	-1.29	1.66
37.7	28.0	27.95	0.05	0.0025	7.91	65.57
6.8	24.6	40.28	-15.68	245.86	-22.99	528.54
56.5	16.0	20.45	-4.45	19.80	26.71	713.42

Даље рачунамо оцену

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{797.97}{12}} = 8{,}154.$$

Како бисмо израчунали одступање оцене $\hat{\beta}$ од стварне вредности праметра β, заменићемо добијену вредност за $\hat{\sigma}$ у формулу (2.3)

$$SD(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{8{,}154}{\sqrt{797.97}} = \frac{8.154}{56.13} = 0.1452.$$

Даље је

$$|t| = \frac{|\hat{\beta}|}{est.SD(\hat{\beta})} = \frac{0.3989}{0.1452} = 2.7472.$$

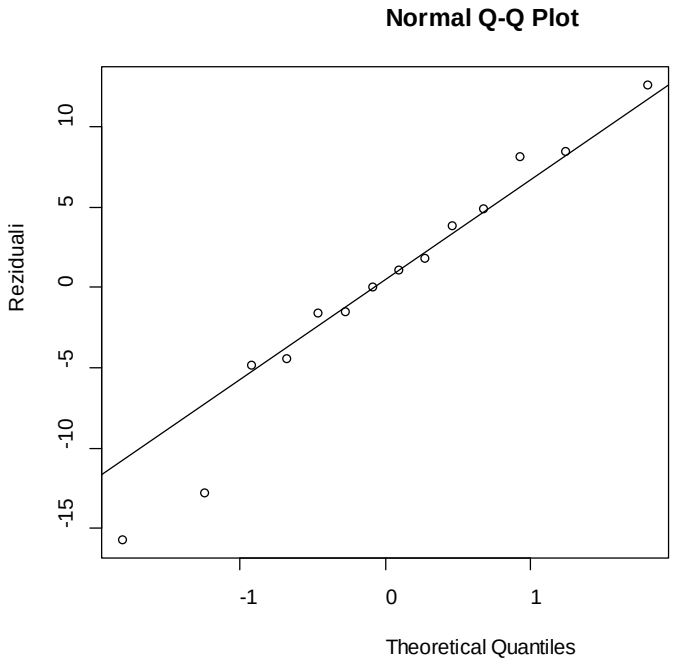
Сада ћемо да користимо таблицу за t-расподелу, како бисмо нашли p-вредност. Пошто је наш број степени слободе једнак 12, гледаћемо 12. врсту у таблици. У овој врсти нам се налазе бројеви $1.356, 1.782, 2.179, 3.055, 4.318$. Добијена вредност за $|t|$ се налази између бројева 2.179 и 3.055 и ближа је броју 3.055. Број 3.055 се налази у колони која је означена са 0.01. Ово значи да је вероватноћа, када је $\beta = 0$, да случајна променљива $|t|$ већа или једнака са бројем 3.055 једнака 0.01, односно

$$P(|t| \geq 3.055) = 0.01.$$

Можемо да кажемо да је p-вредност теста приближно једнака 0.01. Одавде закључујемо да је јако мала вероватноћа да је $\beta = 0$, односно да су променљиве наталитет и урбани проценат зависне.

-Тестирање нормалности

График за тестирање нормалности резидуала дат је на слици 2.2. На слици видимо да је график приближно линеаран па можемо претпоставити да резидуали имају приближно нормалну расподелу. Када график није линеаран, можемо трансформисати узорак тако да добијемо нормалност резидуала.



Слика 2.2 Тестирање нормалности резидуала за узорак „Наталитет“.

Пример 2.1 у R-у.

Прво унесемо променљиве X и Y.

> X<-c(55.0,27.3,33.3,37.1,11.5,14.2,13.9,19.0,33.1,43.2,28.5,37.7,6.8,56.5)

> Y<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9,28.3,33.9,44.2,28.0,24.6,16.0)

Затим помоћу функције „lm“ конструишемо модел.

> mnk<-lm(Y~X)

Помоћу функције „$summary$“ добијамо све потребне информације о моделу.

```
> summary(mnk)

Call:

lm(formula = Y ~ X)

Residuals:

    Min      1Q  Median      3Q     Max

-15.6782  -3.7413   0.5601   4.6440  12.5772

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  42.9905     4.8454   8.872 1.28e-06 ***

X            -0.3989     0.1453  -2.746   0.0177 *

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.154 on 12 degrees of freedom

Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3347

F-statistic:  7.54 on 1 and 12 DF,  p-value: 0.01774
```

Видимо да је добијена регресиона права $Y = 42.9905 - 0.3989X$, стандардно одступање 8.154 и p-вредност 0.01774. Вредност R^2 је 0.3347, па закључујемо да модел не описује податке на задовољавајући начин.

Тестирање нормалности резидуала се врши помоћу Q-Q плота.

```
> qqnorm(residuals(mnk),ylab="Reziduali")

> qqline(residuals(mnk))
```

На овај начин добијамо слику 2.2 на којој видимо да резидуали имају приближно нормалну расподелу.

Уколико из узорка избацимо аутлајер Тринидад и Тобаго, добијамо модел који ће се разликовати од добијеног модела. Одатле можемо да закључимо да је за овај пример боље користити неку другу методу, која није осетљива на аутлајере, на пример методу најмањих апсолутних одступања, описану у делу 3. или M-регресију описану у делу 4.

```
> X1<-c(55.0,27.3,33.3,37.1,11.5,14.2,13.9,19.0,33.1,43.2,28.5,37.7,56.5)

> Y1<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9,28.3,33.9,44.2,28.0,16.0)

> mnk1<-lm(Y1~X1)

> summary(mnk1)

Call:

lm(formula = Y1 ~ X1)


Residuals:

   Min     1Q  Median     3Q    Max

-13.753  -2.534  -1.910   4.534  10.911

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  48.9427     4.4691  10.951 2.96e-07 ***

X1           -0.5492     0.1293  -4.248  0.00137 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.57 on 11 degrees of freedom

Multiple R-squared:  0.6213,    Adjusted R-squared:  0.5868

F-statistic: 18.04 on 1 and 11 DF,  p-value: 0.001371
```

Добијена регресиона права је $Y = 48.9427 - 0.5492X$. Вредност R^2 је 0.5868 што је доста боље од вредности R^2 претходног модела.

2.4 Вишеструка регресија

Претпоставимо да нашим подацима одговара вишеструки линеарни регресиони модел

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e.$$

Наш задатак је да нађемо вектор $\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ који минимизира суму $\sum_{i=1}^{n} \hat{e}_i^2$.

Означимо са

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

где је k број објашњавајућих променљивих, а n број опсервација. Модел тада можемо написати у облику

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}.$$

Вектор оцена се добија помоћу формуле

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

где је $\boldsymbol{X}'$ транспонована матрица матрице $\boldsymbol{X}$.

2.5 Тестирање хипотезе $\boldsymbol{\beta_{q+1}} = \cdots = \boldsymbol{\beta_k} = \boldsymbol{0}$.

Уколико желимо да испитамо да ли упрошћени модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + e$ једнако добро описује податке као неупрошћени модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e$, где је $k > q$, тестираћемо хипотезу $\beta_{q+1} = \cdots = \beta_k = 0$.

Нека SKR представља суму квадрата резидуала модела. Можемо да упоредимо упрошћени модел са неупрошћеним моделом тако што ћемо упоредити $SKR_{uprošćen}$ са $SKR_{neuprošćen}$. Нека је $\hat{\sigma}^2$ непристрасна оцена параметра σ^2, дата једнакошћу

$$\hat{\sigma}^2 = \frac{SKR_{neuprošćen}}{n-k-1}.$$

Тест статистика за тестирање хипотрезе $\beta_{q+1} = \cdots = \beta_k = 0$ је

$$F = \frac{SKR_{uprošćen} - SKR_{neuprošćen}}{(k-q)\hat{\sigma}^2}.$$

Под претпоставком да случајне грешке имају нормалну расподелу, тест статистика F има F-расподелу са $k-q$ и $n-k-1$ степени слободе.

Пример 2.2

Дата нам је база података, која је укључена у програмски пакет R, о броју врста корњача на различитим острвима Галапагоса. Имамо укупно 30 острва и 7 променљивих. Променљиве су:

Species (Y)-број врста корњача пронађен на одговарајућем острву,

Endemics (X_1)-број ендемских врста корњача на одговарајућем острву,

Area (X_2)-област острва (у квадратним километрима),

Elevation (X_3)-највећа надморска висина на острву (у метрима),

Nearest (X_4)-удаљеност од најближег острва (у километрима),

Scruz (X_5)-удаљеност до острва Санта Круз (у километрима),

Adjacent (X_6)-област суседног острва (у квадратним километрима).

Ми желимо да испитамо зависност променљиве Species од осталих променљивих. Модел правимо помоћу функције „lm“.

Прво ћемо да учитамо базу података која се налази у пакету faraway.

```
> library(faraway)
> data(gala)
> gala
```

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	23	25.09	346	0.6	0.6	1.84
Bartolome	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamano	2	1	0.05	77	1.9	1.9	903.82
Daphne.Major	18	11	0.34	119	8.0	8.0	1.84
Daphne.Minor	24	0	0.08	93	6.0	12.0	0.34
Darwin	10	7	2.33	168	34.1	290.2	2.85
Eden	8	4	0.03	71	0.4	0.4	17.95
Enderby	2	2	0.18	112	2.6	50.2	0.10
Espanola	97	26	58.27	198	1.1	88.3	0.57
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32
Gardner1	58	17	0.57	49	1.1	93.1	58.27
Gardner2	5	4	0.78	227	4.6	62.2	0.21
Genovesa	40	19	17.35	76	47.4	92.2	129.49

Isabela	347	89	4669.32	1707	0.7	28.1	634.49
Marchena	51	23	129.49	343	29.1	85.9	59.56
Onslow	2	2	0.01	25	3.3	45.9	0.10
Pinta	104	37	59.56	777	29.1	119.6	129.49
Pinzon	108	33	17.95	458	10.7	10.7	0.03
Las.Plazas	12	9	0.23	94	0.5	0.6	25.09
Rabida	70	30	4.89	367	4.4	24.4	572.33
SanCristobal	280	65	551.62	716	45.2	66.6	0.57
SanSalvador	237	81	572.33	906	0.2	19.8	4.89
SantaCruz	444	95	903.82	864	0.6	0.0	0.52
SantaFe	62	28	24.08	259	16.5	16.5	0.52
SantaMaria	285	73	170.92	640	2.6	49.2	0.10
Seymour	44	16	1.84	147	0.6	9.6	25.09
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

Затим правимо модел и податке о моделу добијамо помоћу функције „summary“.

```
> model<-lm(Species~Endemics+Area+Elevation+Nearest+Scruz+Adjacent,data=gala)
> summary(model)

Call:
lm(formula = Species ~ Endemics + Area + Elevation + Nearest +
    Scruz + Adjacent, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-68.219 -10.225   1.830   9.557  71.090
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-15.337942	9.423550	-1.628	0.117	
Endemics	4.393654	0.481203	9.131	4.13e-09	***
Area	0.013258	0.011403	1.163	0.257	
Elevation	-0.047537	0.047596	-0.999	0.328	
Nearest	-0.101460	0.500871	-0.203	0.841	
Scruz	0.008256	0.105884	0.078	0.939	
Adjacent	0.001811	0.011879	0.152	0.880	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.96 on 23 degrees of freedom

Multiple R-squared: 0.9494, Adjusted R-squared: 0.9362

F-statistic: 71.88 on 6 and 23 DF, p-value: 9.674e-14

Видимо да је модел који смо добили

$$Y = -15.337942 + 4.393654X_1 + 0.013258X_2 - 0.047537X_3 - 0.101460X_4 - 0.008256X_5 + 0.001811X_6$$

Вредност R^2 је 0.9494 што је близу броја 1 па можемо да закључимо да модел добро одговара подацима. F статистика има вредност 71.88 са 6 и 23 степени слободе.

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generalizations. Least squares and alternatives.1995.*

[3] Julian J.Fraway. Linear models with R,2005.

[4] Norman R.Draper, Harry Smith. *Applied regression analysis.* University of Wisconsin, 1998.

[5] Ronald Christensen. Department of mathematics and statistics. University of New Mexico. *Plane answers to complex questions: The theory of linear models.2001.*

3. Метода најмањих апсолутних одступања

Метода најмањих апсолутних одступања уведена је скоро 50 година пре Методе најмањих квадрата, а увео ју је Руђер Бошковић[1] 1757. године. Након што је Пјер Симон Лаплас[2] усвојио методу 30 година касније, она је била повремено примењивана, док убрзо није пала у сенку Методе најмањих квадрата.

3.1 Одређивање регресионе праве

Прво ћемо претпоставити да подацима које имамо одговара линеарни регресиони модел $Y = \alpha + \beta X + e$. Наш задатак је да пронађемо оцене $\hat{\alpha}$ и $\hat{\beta}$, параметара α и β.

Код методе најмањих квадрата, оцене $\hat{\alpha}$ и $\hat{\beta}$ су изабране тако да сума квадрата резидуала $\sum_{i=1}^{n} \hat{e}_i^2$, $1 \leq i \leq n$, буде најмања могућа. Код методе најмањих апсолутних одступања, оцене $\hat{\alpha}$ и $\hat{\beta}$ су изабране тако да сума апсолутних вредности резидуала $\sum_{i=1}^{n} |\hat{e}_i|$, буде најмања могућа. Односно, код методе најмањих апсолутних одступања, оцене $\hat{\alpha}$ и $\hat{\beta}$ су вредности a и b за које је вредност израза

$$\sum |y_i - (a + bx_i)| \tag{3.1}$$

минимална.

Разлику $y_i - (a + bx_i)$ зовемо одступање тачке (x_i, y_i) од праве $\hat{Y} = a + bX$, а кад су a и b замењене оцењеним вредностима користи се термин резидуал.

Циљ је проналажење праве која највернији представља одређене податке. Не постоје формуле које би одговарале методи најмањих апсолутних одступања, уместо тога представљамо алгоритам за израчунавање тражених оцена. За почетак ћемо претпоставити да наши подаци немају особине нејединствености и дегенеративности, о којима ћемо говорити у следећем одељку. Ове особине, које узрокују техничке проблеме у алгоритму, се не јављају често у пракси.

Главни део алгоритма је процедура, која за дату тачку (x_0, y_0), проналази најбољу праву која садржи ту тачку, од свих правих које је садрже. Ова процедура се користи заједно са чињеницом да у овој методи регресиона права пролази кроз две дате тачке. Почињемо са једном датом тачком, на пример тачком (x_1, y_1) и проналазимо најбољу праву која садржи ту тачку. Добијена права садржи још једну тачку, реиндексирањем је можемо означити са (x_2, y_2). Даље, проналазимо најбољу праву која садржи тачку (x_2, y_2). Ова права садржи још једну тачку, означимо је са (x_3, y_3). Даље, проналазимо најбољу праву која садржи тачку (x_3, y_3) и тако даље. Како настављамо са коришћењем нашег алгоритма, праве које добијамо постају све боље. У једном од корака ћемо добити праву која је идентична са правом из претходног корака. Та права је онда регресиона права по методи најмањих апсолутних одступања.

---

[1] Руђер Бошковић (1711-1787) био је српски физичар, астроном, математичар и дипломата, бавио се још и оптиком, поезијом и архитектуром.

[2] Пјер Симон Лаплас (1749-1827) био је француски математичар и астроном.

Треба бјаснити на који начин, од свих правих које садрже дату тачку (x_0, y_0), бирамо најбољу праву. За сваку тачку (x_i, y_i) израчунајмо нагиб $\frac{(y_i - y_0)}{(x_i - x_0)}$, праве која садржи тачке (x_0, y_0) и (x_i, y_i). Ако је $x_i = x_0$ за неко i, нагиб неће бити дефинисан, али такве тачке можемо да занемаримо. Даље ћемо да реиндексирамо тачке тако да добијемо

$$\frac{(y_1 - y_0)}{(x_1 - x_0)} \leq \frac{(y_2 - y_0)}{(x_2 - x_0)} \leq \cdots \leq \frac{(y_n - y_0)}{(x_n - x_0)}.$$

Означимо са $T = \sum_{i=1}^{n} |x_i - x_0|$. Нађимо индекс k који задовољава следеће неједнакости

$$|x_1 - x_0| + \cdots + |x_{k-1} - x_0| < \frac{1}{2}T$$

$$|x_1 - x_0| + \cdots + |x_{k-1} - x_0| + |x_k - x_0| > \tfrac{1}{2}T \qquad (3.2)$$

Најбоља права која садржи тачку (x_0, y_0) је права $\hat{Y} = \alpha^* + \beta^* X$, где је

$$\beta^* = \frac{y_k - y_0}{x_k - x_0}$$

$$\alpha^* = y_0 - \beta^* x_0 \qquad (3.3)$$

Најбитнија чињеница у овом алгоритму је да регресиона права садржи две од датих тачака. Да бисмо видели зашто то важи, замислимо праву $y = a + bx$, нацртану на графику на коме смо представили тачке из базе података. Апсолутно одступање тачке из узорка је уствари најкраће растојање тачке од праве на графику.

Вредност израза (3.1) је сума ових апсолутних одступања. Претпоставимо да изабрана права не садржи ни једну тачку из узорка. Ако праву померимо на горе за неко мало растојање, рецимо ε, онда ће свако апсолутно одступање или да се повећа за ε, или да се смањи за ε, у зависности од тога да ли се тачка налази изнад или испод праве на графику. Вредност израза (3.1) можемо смањити, тако што ћемо праву померити на горе, уколико има више тачака изнад те праве, односно на доле уколико има више тачака испод те праве. Праву ћемо померити док она не „сретне“ једну тачку из узорка.

Ако права садржи тачно једну тачку из узорка, онда је можемо ротирати око те тачке у смеру кретања казаљке на сату или у супротном смеру, док та права не сретне другу тачку из узорка. Апсолутно одступање тачке која припада датој правој остаје нула а остала апсолутна одступања ће се повећати или смањити за одређени променљиви износ. Ротирањем праве у смеру кретања казаљке на сату или у супротном смеру, док права не сретне другу тачку, можемо смањивати вредност суме апсолутних одступања. Ово нам указује на то да од свих правих треба да бирамо оне које садрже две тачке из узорка.

Од свих правих које садрже тачку (x_0, y_0), ми хоћемо да пронађемо праву која минимизира вредност израза (3.1). Да би права $\hat{Y} = a + bX$ садржала тачку (x_0, y_0) мора да важи $y_0 = a + bx_0$, односно $a = y_0 - bx_0$. Ово оправдава другу једнакост (3.3). Одступање $y_i - (a + bx_i)$ сада може бити записано као $(y_i - y_0) - b(x_i - x_0)$. Ми сада желимо да нађемо вредност за променљиву b, која минимизира вредност израза

$$\sum_{i=1}^{n}(y_i - y_0) - b(x_i - x_0). \tag{3.4}$$

Посматрајмо израз (3.4) као функцију од b. Да бисмо нашли минимум ове функције морамо наћи први извод ове функције. Функција $|t|$ није диференцијабилна за $t = 0$, диференцијабилна је за $t \neq 0$.

Ово нам дозвољава да функцију (3.4) диференцирамо, за све вредности променљиве b, осим за ону вредност која задовољава $(y_i - y_0) - b(x_i - x_0) = 0$, односно за $b = \frac{y_i - y_0}{x_i - x_0}$. Неједнакости (3.2) представљају стања када је извод функције (3.4) негативан за $b < \beta^*$ и позитиван за $b > \beta^*$. Заједно са чињеницом да је функција (3.4) непрекидна чак и у тачкама у којима није диференцијабилна, ово имплицира да је функција (3.4) опадајућа за $b < \beta^*$ односно растућа за $b > \beta^*$, што даље имплицира да је $b = \beta^*$ минимум наше функције. Ово објашњава прву једнакост (3.3).

3.2 Нејединственост и дегенеративност

Алгоритам за проналажење регресионе праве ће у већини случајева дати резултате али у одређеним случајевима ћемо се сусрести са особинама нејединствености и дегенеративности. Нејединственост значи да постоји више од једне најбоље праве која садржи дату тачку. Дегенеративност значи да најбоља права која садржи дату тачку садржи још две или више тачака из узорка. Уколико дође до једног од ова два проблема, алгоритам може да се врти у круг, не налазећи регресиону праву, или може да пронађе праву која није стварно решење.

Могућност појаве ових проблема повећава се уколико се у изразу (3.2) добије једнакост или уколико је нагиб $\beta^* = \frac{y_k - y_0}{x_k - x_0}$ једнак изразу $(y_{k-1} - y_0)/(x_{k-1} - x_0)$ или изразу $(y_{k+1} - y_0)/(x_{k+1} - x_0)$. У оваквим примерима ћемо користити други алгоритам, који представља поједностављење полазног и у коме нема проблема нејединствености и дегенеративности.

Описани алгоритам заиста можемо поједноставити али ће бити потребно много више израчунавања. Познато је да регресиона права у овој методи садржи две тачке из узорка. Одатле закључујемо да праву можемо да тражимо међу свим правама које садрже две тачке из узорка. При том се неке од ових правих могу подударати. Можемо да израчунамо суму апсолутних одступања (3.1) за сваку од тих правих и изаберемо ону (или оне) код које је то одступање најмање. Изводљивост овог алгоритма зависи од обима узорка јер у општем случају треба радити проверу за $\binom{n}{2}$ правих.

Овај алгоритам нема особину дегенеративности а у случају појаве нејединствености, када добијемо више од једне најбоље праве, ми можемо изабрати једну произвољно.

3.3 Тестирање хипотезе $\boldsymbol{\beta = 0}$

Сада ћемо описати како проверавамо да ли β може бити 0. Дакле тестирамо хипотезу $H_0(\beta = 0)$. Методом најмањих апсолутних одступања смо израчунали оцене $\hat{\alpha}$ и $\hat{\beta}$ и резидуале $\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$. Нека је $m = n - 2$, број резидуала који су различити од нуле. Поређајмо ове резидуале у растући поредак тј. формирајмо одговарајући варијациони низ. Означимо са $\hat{e}_1$ резидуал који има најмању вредност, са $\hat{e}_2$ резидуал који је следећи по величини, ..., и $\hat{e}_m$ резидуал са највећом вредношћу.

Узмимо да је k_1 цео број, који је најближи вредности израза $\frac{m+1}{2} - \sqrt{m}$ и узмимо да је k_2 цео број, који је најближи вредности израза $\frac{m+1}{2} + \sqrt{m}$. Израчунајмо сада

$$\hat{\tau} = \frac{\sqrt{m}[\hat{e}_{(k_2)} - \hat{e}_{(k_1)}]}{4}, \tag{3.5}$$

а затим израчунајмо

$$ocena.SD(\hat{\beta}) = \frac{\hat{\tau}}{\sqrt{\sum(x_i - \bar{x})^2}}. \tag{3.6}$$

Тест статистика је

$$|t| = \frac{|\hat{\beta}|}{ocena.SD(\hat{\beta})} \tag{3.7}$$

и при нултој хипотези има Студентову расподелу са $n - 2$ степени слободе и њену реалну вредност означавамо такође са $|t|$.

p −вредност теста израчунаћемо као $P[|T| \geq |t|]$ где T представља случајну променљиву која има t- расподелу са $n - 2$ степени слободе.

Вредност $\hat{\tau}$, дата изразом (3.5) је оцена параметра τ који у методи најмањих апсолутних одступања има улогу аналогну улози параметра σ у методи најмањих квадрата. Стандардно одступање оцене параметра β у методи најмањих квадрата $\hat{\beta}_{MNK}$ је $\sigma/\sqrt{\sum(x_i - \bar{x})^2}$ а стандардно одступање оцене $\hat{\beta}_{MNAO}$, добијене методом најмањих апсолутних одступања, је $\tau/\sqrt{\sum(x_i - \bar{x})^2}$ (Тачност оцене се повећава са повећањем обима узорка). Одавде закључујемо да количник τ/σ одређује која од ове две методе је боља за оцену нагиба регресионе праве.

И τ и σ представљају вредности случајних грешака. τ је једнак $1/(2\theta)$, где је θ медијана густина расподеле случајних грешака. Када користимо регресију углавном претпостављамо да је медијана случајних грешака једнака нули. Ако је расподела случајних грешака симетрична онда се средња вредност и медијана подударају. Ако је σ велики број , онда су грешке широко распршене,

па је густина расподеле има малу вредност близу медијане, отуда је θ мали број, што имплицира да је τ велики број.

Грубо речено, важи да је τ велико када је σ велико и обрнуто, али тачан количник τ/σ зависи од облика расподеле случајних грешака. Ако грешке имају нормалну расподелу, онда ће $\frac{\tau}{\sigma} = 1.253 > 1$, па ће због тога за узорке великог обима, метода најмањих апсолутних одступања бити мање тачна од методе најмањих квадрата. Ако грешке имају Лапласову расподелу (која је такође симетрична, али има дебље репове од нормалне расподеле) онда ће бити $\frac{\tau}{\sigma} = 0.707 < 1$.

У наведеном тесту, при хипотези $\beta = 0$, случајна променљива t има t-расподелу са $n - 2$ степени слободе. Одавде следи да, када је n велики број и $\beta = 0$, случајна променљива t има приближно стандардну нормалну расподелу. Ово је оправдано јер за велике вредности n, t-расподела је јако слична стандардној нормалној расподели. Када је n мали број, свакако треба користити t-расподелу уместо стандардне нормалне расподеле.

Пример 3.1

На пример „Наталитет“, из дела 2., примењујемо и Методу најмањих апсолутних одступања. У табели 2.1, на страни 9, су нам дати подаци. Као први корак пронаћи ћемо најбољу праву која садржи тачку из табеле која представља Канаду (55.0,16.2). Да бисмо ово урадили, морамо да формирамо нагибе за праве одређене овом и осталим тачкама из узорка $\frac{y_i - 16.2}{x_i - 55.0}$. Нагиби за 13 држава приказани су у табели 3.1 и поређани су у растућем поретку.

Табела 3.1 Израчунавања коришћена за проналажење најбоље праве кроз тачку која представља Канаду.

Држава	$\frac{y_i - 16.2}{x_i - 55.0}$	$\lvert x_i - 55.0 \rvert$	$\sum \lvert x_i - 55.0 \rvert$
Мексико	-1.5000	11.8	11.8
Никарагва	-1.0566	26.5	38.3
Доминиканска република	-0.9441	17.9	56.2
Хондурас	-0.7694	36.0	92.2
Панама	-0.6821	17.3	109.5
Хаити	-0.6107	41.1	150.6
Јамајка	-0.5525	21.9	172.5
Ел Салвадор	-0.5517	43.5	216.0
Гватемала	-0.5447	40.8	256.8
Костарика	-0.5162	27.7	284.5
Тринидад и	-0.1743	48.20	332.7

Тобаго			
САД	-0.1333	1.5	334.2
Куба	-0.0323	21.7	355.9

Израчунајмо прво $\mathrm{T} = \sum|x_i - 55.0| = 355.9$, а затим $\frac{\mathrm{T}}{2} = 177{,}95$. У табели пронађимо прву државу чија кумулативна сума већа од ове вредности. Из табеле видимо да је то држава Ел Салвадор. Дакле $\beta^* = -0.5517$ а $\alpha^* = 16.2 - (-0.5517)(55.0) = 46.54$. Заправо, у овом кораку ми не морамо да рачунамо β^* и α^*, довољно нам је да знамо да најбоља права, која садржи тачку која представља Канаду, садржи и тачку која представља Ел Салвадор. Следећи корак је да пронађемо најбољу праву која садржи тачку која представља Ел Салвадор. Формирамо нагибе $\frac{y_i - 40.2}{x_i - 11.5}$ и направимо табелу сличну табели 3.1. Ако бисмо направили такву табелу видели бисмо да је укупна кумулативна сума 265.5 а да је прва држава чија је кумулативна сума већа од 265.5/2=132.75, Сједињене Америчке Државе. Дакле најбоља права, која садржи тачку која представља Ел Салвадор, садржи и тачку која представља САД.

Даље тражимо најбољу праву која садржи тачку која представља САД. Прво формирамо табелу и на исти начин налазимо да најбоља права, која садржи тачку која представља САД, садржи и тачку која представља Ел Салвадор. Ово је иста права коју смо добили у претходном кораку и овим се наш алгоритам зауставља. Пронашли смо да је регресиона права она права која садржи тачке Ел Салвадор и САД. Њен нагиб је $\hat{\beta} = \frac{40.2 - 16.0}{11.5 - 56.5} = -0.5378$ а $\hat{\alpha} = 40.2 - (-0.5378)(11.5) = 46.38$, дакле регресиона права је $\hat{Y} = 46.38 - 0.5378X$.

-Тестирање хипотезе $\beta = 0$

У табели 3.2 дати су резидуали који су различити од нуле. Како је $m = n - 2 = 12$ и $\frac{m+1}{2} - \sqrt{m} = \frac{13}{2} - \sqrt{12} = 3.04$ имамо да је $k_1 = 3$. Слично налазимо $k_2 = 10$. Одавде је

$\hat{\tau} = \sqrt{12}\left[\hat{e}_{(10)} - \hat{e}_{(3)}\right]/4$. У табели 3.2 видимо да је $\hat{e}_{(3)} = -1.203$ и $\hat{e}_{(10)} = 7.733$. Одатле је $\hat{\tau} = 7.739$.

Израчунајмо даље $\sum(x_i - \bar{x})^2 = 3151$. Па је $ocena.SD\left(\hat{\beta}\right) = \frac{7.739}{\sqrt{3151}} = 0.1379$, и $|t| = \frac{|-0.5378|}{0.1379} = 3.900$. Да бисмо израчунали p-вредност користићемо t расподелу са 12 степени слободе. Из t-табеле, видимо да је p-вредност између 0.001 и 0.01. Закључујемо да је нагиб регресионе праве различит од нуле (тј.одбацујемо хипотезу H_0 са прагом значајности $\alpha > 0{,}01$).

Табела 3.2 Резидуали који су различити од нуле, добијени методом најмањих апсолутних одступања, поређани у растући поредак.

i	$\hat{e}_{(i)}$
1	-18.128
2	-11.576
3	-1.203
4	-0.607
5	-0.348
6	-0.284
7	1.890
8	2.391
9	6.667
10	7.733
11	10.748
12	13.142

Пример 3.1 у R-у

Прво морамо да учитамо пакет „quantreg“.

```
> library(quantreg)
```

Затим дефинишемо променљиве и правимо модел помоћу функције „rq”.

```
> X<-c(55.0,27.3,33.3,37.1,11.5,14.2,13.9,19.0,33.1,43.2,28.5,37.7,6.8,56.5)
```

```
> Y<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9,28.3,33.9,44.2,28.0,24.6,16.0)
```

```
> mnao<-rq(Y~X)
```

Даље, помоћу функције „summary“ добијамо све потребне информације о моделу.

```
> summary(mnao)
```

```
Call: rq(formula = Y ~ X)
```

```
tau: [1] 0.5
```

Coefficients:

coefficients lower bd upper bd

(Intercept) 46.38444 29.76220 53.22936

X -0.53778 -0.58461 0.17729

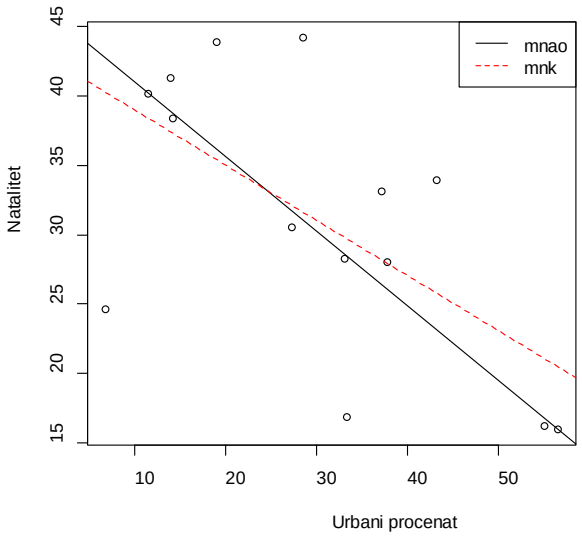
Warning message:

In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) : Solution may be nonunique

Добијена регресиона права је $Y = 46.3844 - 0.53778X$. Видимо да нам и програмски пакет R даје упозорење да решење може бити нејединствено.

Затим ћемо да нацртамо праву добијену методом најмањих апсолутних одступања и праву добијену методом најмањих квадрата на истом графику.

```
> plot(Y~X,xlab="Urbani procenat",ylab="Natalitet")
> abline(mnao)
> abline(lm(Y~X),col=2,lty=2)
> legend("topright",legend=c("mnao","mnk"),col=1:2,lty=1:2)
```



Слика 2.1 Регресионе праве за узорак „Наталитет“ добијене методом најмањих апсолутних одступања и методом најмањих квадрата.

3.4 Пример вишеструке регресије

Сада ћемо да применимо методу најмањих апсолутних одступања на анализирање учесталости пожара у одређеним стамбеним четвртима. Ми желимо да видимо на који начин је учесталост пожара повезана са следећим особинама ових четврти: старост кућа које се налазе у тим четвртима, учесталост провала (крађа) и примања породица које живе у овим четвртима.

Узорак, дат табелом 3.3, садржи податке из 47 четврти које се претежно налазе у Чикагу, а подаци су из 1975. године. У табели су дати редни бројеви четврти и пет променљивих. Променљива „Пожар“ представља број пожара на 1000 стамбених јединица у одређеној четврти. Променљива „Старост“ представља проценат стамбених јединица у одговарајућој четврти, саграђених пре 1940. године. Променљива „Крађе“ представља број крађа (провала) на 1000 стамбених објеката. Променљива „Приход“ представља просечан приход фамилија које живе у одређеној области.

У неким ранијим истраживањима закључено је да области 7 и 24 много одступају од осталих података и да их треба избрисати из модела. Даље, проблем са великом разликом у одступањима је решен тако што је додата колона „$\log(\text{Пожар})$“.

Модел који посматрамо је модел

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

где је $Y = \log(\text{Пожар})$, $X_1 = \text{Старост}$, $X_2 = \text{Крађе}$, и $X_3 = \text{Приход}$. Сада ћемо да применимо методу најмањих апсолутних одступања како бисмо оценили и тестирали коефицијенте у овом моделу.

Табела 3.3 Узорак „Пожар“.

Област	Пожар	$\boldsymbol{log}(\text{пожар})$	Старост	Крађе	Приход
1	6.2	1.825	0.604	29	11.744
2	9.5	2.251	0.765	44	9.323
3	10.5	2.351	0.735	36	9.948
4	7.7	2.041	0.669	37	10.656
5	8.6	2.152	0.814	53	9.730
6	34.1	3.529	0.526	68	8.231
7	11.0	2.398	0.426	75	21.480
8	6.9	1.932	0.785	18	11.104
9	7.3	1.988	0.901	31	10.694
10	15.1	2.715	0.898	25	9.631
11	29.1	3.371	0.827	34	7.995
12	2.2	0.788	0.402	14	13.722
13	5.7	1.740	0.279	11	16.250

14	2.0	0.693	0.077	11	13.686
15	2.5	0.916	0.638	22	12.405
16	3.0	1.099	0.512	17	12.198
17	5.4	1.686	0.851	27	11.600
18	2.2	0.788	0.444	9	12.765
19	7.2	1.974	0.842	29	11.084
20	15.1	2.715	0.898	30	10.510
21	16.5	2.803	0.727	40	9.784
22	18.4	2.912	0.729	32	7.342
23	36.2	3.589	0.631	41	6.565
24	39.7	3.681	0.830	147	7.459
25	18.5	2.918	0.783	22	8.014
26	23.3	3.148	0.790	29	8.177
27	12.2	2.501	0.480	46	8.212
28	5.6	1.723	0.715	23	11.230
29	21.8	3.082	0.731	4	8.330
30	21.6	3.073	0.650	31	5.583
31	9.0	2.197	0.754	39	8.564
32	3.6	1.281	0.208	15	12.102
33	5.0	1.609	0.618	32	11.876
34	28.6	3.353	0.781	27	9.742
35	17.4	2.856	0.686	32	7.520
36	11.3	2.425	0.734	34	7.388
37	3.4	1.224	0.020	17	13.842
38	11.9	2.477	0.570	46	11.040
39	10.5	2.351	0.559	42	10.332
40	10.7	2.370	0.675	43	10.908
41	10.8	2.380	0.580	34	11.156
42	4.8	1.569	0.152	19	13.323
43	10.4	2.342	0.408	25	12.960
44	15.6	2.747	0.578	28	11.260
45	7.0	1.946	0.114	3	10.080
46	7.1	1.960	0.492	23	11.428
47	4.9	1.589	0.466	27	13.731

3.4.1 Оцена регресионих коефицијената

Оцене коефицијената $\beta_0, \beta_1, \beta_2, \beta_3$ бирамо тако да сума апсолутних вредности резидуала , $\sum|\hat{e}_i|$, буде најмања могућа. То значи да су $\beta_0, \beta_1, \beta_2, \beta_3$ вредности b_0, b_1, b_2, b_3 за које израз

$$\sum|y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3})| \tag{3.7}$$

има најмању вредност.

Не постоје формуле које проналазе ове вредности $\beta_0, \beta_1, \beta_2, \beta_3$, али ми ћемо описати алгоритам помоћу кога можемо да их пронађемо. Алгоритам претпоставља да узорак није склон појави нејединствености и дегенеративности.

Користићемо векторско представљање параметара:

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{и} \quad \boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \end{bmatrix}$$

Тада сума апсолутних одступања (3.7) може бити записана у следећем облику

$$\sum|y_i - \boldsymbol{b}'\boldsymbol{x_i}| \tag{3.8}$$

Ми желимо да пронађемо вектор $\boldsymbol{b}$ који минимизира вредност израза (3.8).

Алгоритам за вишеструку регресију је, као и алгоритам за једноструку регресију, итеративан. Почињемо са вектором $\boldsymbol{b}$, затим проналазимо бољи вектор (вектор за који је вредност израза (3.8) мања) , и тако даље, све док не пронађемо најбољи вектор $\widehat{\boldsymbol{\beta}}$. У сваком кораку, ако имамо вектор $\boldsymbol{b}$, проналазимо бољи вектор $\boldsymbol{b}^*$, тако што прво пронађемо одговарајући вектор смера $\boldsymbol{d}$ а затим проналазимо вредност t за коју је вектор $\boldsymbol{b}^* = \boldsymbol{b} + t\boldsymbol{d}$.

Стога нам треба поступак који проналази вредност t која минимизира вредност израза

$$\sum|y_i - (\boldsymbol{b} + t\boldsymbol{d})'\boldsymbol{x_i}| \tag{3.9}$$

Ако означимо са $z_i = y_i - \boldsymbol{b}'\boldsymbol{x_i}$ и $w_i = \boldsymbol{d}'\boldsymbol{x_i}$, онда добијамо израз

$$\sum|z_i - t w_i| \tag{3.10}$$

Овим смо добили исти проблем као и проблем који се јавио код проналажења вредности b која минимизира вредност израза (3.4), за који већ знамо решење. Узећемо количнике z_i/w_i, поређати их у растући поредак, реиндексирати све z_i-ове и w_i-ове и пронаћи индекс k који задовољава изразе

$$|w_1| + |w_2| + \cdots + |w_{k-1}| < \frac{1}{2}T,$$

$$|w_1| + |w_2| + \cdots + |w_{k-1}| + |w_k| > \frac{1}{2}T,$$

где је $T = \sum|w_i|$. Минимизирајућа вредност за t је z_k/w_k.

У сваком кораку, алгоритам разматра четири вектора d_1, d_2, d_3, d_4. Ова четири вектора уствари представљају 8 смерова, јер за сваки вектор $\boldsymbol{d_j}$ имамо позитиван и негативан смер. Од свих смерова најбољи је онај за који вредност израза (3.9) најбрже опада у тачки $t = 0$. Да бисмо видели колико израз (3.9) стрмо опада, израчунаћемо први извод у тачки $t = 0$. Десни извод у тачки $t = 0$ је $W_- + W_0 - W_+$, где је W_- сума свих $|w_i|$ за које је количник z_i/w_i негативан, W_0 је сума свих $|w_i|$ за које је $z_i = 0$, и W_+ сума свих $|w_i|$ за које је количник z_i/w_i позитиван. Израчунаћемо вредности извода за свих осам смерова и изабрати најпогоднији смер, односно онај смер код кога је вредност извода негативна и најмања. Ако су вредности свих извода позитивне, онда је тренутни вектор $\boldsymbol{b}$ најбољи вектор $\widehat{\boldsymbol{\beta}}$ и алгоритам се зауставља.

Код једноструке регресије, регресиона права је садржавала две тачке из базе. Слично, код вишеструке регресије са p случајних променљивих, регресиона једнакост је задовољена за $p + 1$ тачку из узорка. У нашем примеру $p = 3$, па је $p + 1 = 4$. Одавде можемо да закључимо да би разуман почетак алгоритма био да изаберемо 4 тачке из базе, на пример тачке индексиране са $i = 1,2,3,4$, и да узмемо да је вектор $\boldsymbol{b}$ на почетном кораку одређен са $y_i = \boldsymbol{b}'\boldsymbol{x_i}$, за $i = 1,2,3,4$. У матричној нотацији, ове четири једнакости можемо записати као $\boldsymbol{Ab} = \boldsymbol{c}$, где је

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{x_1}' \\ \boldsymbol{x_2}' \\ \boldsymbol{x_3}' \\ \boldsymbol{x_4}' \end{bmatrix} \quad \text{и} \quad \boldsymbol{c} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Одавде добијамо да је $\boldsymbol{b} = \boldsymbol{A^{-1}c}$. Узмимо да су почетне вредности вектора d_1, d_2, d_3, d_4 баш четири колоне матрице $\boldsymbol{A^{-1}}$.

У сваком кораку имамо тренутни вектор $\boldsymbol{b}$ такав да је задовољена регресиона једнакост за четири тачке из базе, на пример за тачке индексиране са i_1, i_2, i_3, i_4. Нека је

$$\boldsymbol{A} = \begin{bmatrix} x_{i1}' \\ x_{i2}' \\ x_{i3}' \\ x_{i4}' \end{bmatrix}.$$

Тренутни вектори смера су четири колоне матрице $\boldsymbol{A^{-1}}$.

Као што је и описано изнад, треба да израчунамо износе осам извода и да изаберемо смер код кога је ова вредност најнегативнија. Претпоставимо да је то смер $\boldsymbol{d_3}$.

Вектор који је бољи од вектора $\boldsymbol{b}$ се добија као $\boldsymbol{b}^* = \boldsymbol{b} + \boldsymbol{t}^*\boldsymbol{d_3}$, где је $\boldsymbol{t}^*$ вредност од t која минимизира (3.9). Приметимо да важи $t^* = \frac{y_k - \boldsymbol{b}'\boldsymbol{x_k}}{\boldsymbol{d_3}'\boldsymbol{x_k}}$ за неко k. Заменимо трећу врсту (јер је трећи вектор смера коришћен) матрице $\boldsymbol{A}$ са $\boldsymbol{x_k}'$. Означимо ову нову матрицу са $\boldsymbol{A}^*$.

Нови комплет вектора смера су сада четири колоне матрице $\boldsymbol{A^{*-1}}$. Итерација се понавља док не дођемо до корака где су сви изводи позитивни.

Да бисмо применили описани поступак за вишеструку регресију на узорак „Пожар“, изаберимо прво четири области из табеле 3.3, рецимо области 1,2,3 и 4. Коришћењем података за те четири области формирајмо матрице

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0.604 & 29 & 11.744 \\ 1 & 0.765 & 44 & 9.323 \\ 1 & 0.735 & 36 & 9.948 \\ 1 & 0.669 & 37 & 10.656 \end{bmatrix} \quad \text{и} \quad \boldsymbol{c} = \begin{bmatrix} 1.825 \\ 2.251 \\ 2.351 \\ 2.041 \end{bmatrix}.$$

Почетни вектор је

$$\boldsymbol{b} = \boldsymbol{A^{-1}c} = \begin{bmatrix} 47.95 \\ -23.26 \\ -0.1161 \\ -2.443 \end{bmatrix}.$$

А почетна матрица вектора смера је

$$\boldsymbol{A^{-1}} = \begin{bmatrix} -284.9 & -308.3 & 157.8 & 436.4 \\ 164.5 & 176.6 & -79.71 & -261.3 \\ 0.5233 & 0.6744 & -0.4656 & -0.7321 \\ 14.59 & 15.51 & -8.185 & -21.91 \end{bmatrix}.$$

Сада желимо да пронађемо вектор који је бољи од вектора $\boldsymbol{b}$. Да бисмо то урадили нађимо вредност десног првог извода израза (3.9) за 8 смерова, 4 су представљена са 4 колоне матрице $\boldsymbol{A^{-1}}$ а преостала 4 представљају њихове супротне векторе. Означимо прву колону матрице $\boldsymbol{A^{-1}}$ са $\boldsymbol{d_1}$. Израчунавања за добијање извода у тачки $\boldsymbol{d_1}$ су приказана у табели 3.4, при чему су као што је већ напоменуто квартови 7 и 24 изостављени из модела.

Табела 3.4 Израчунавања коришћена за добијање десног извода израза (3.9) за $\boldsymbol{t = 0}$.

$\boldsymbol{i}$	$\boldsymbol{z_i}$	$\boldsymbol{w_i}$	знак$(\frac{\boldsymbol{z_i}}{\boldsymbol{w_i}})$	$\lvert\boldsymbol{w_i}\rvert$
1	0	1	0	1
2	0	0	*	0
3	0	0	*	0
4	0	0	*	0
5	3.08	18.71	+	18.71
6	-4.16	-42.68	+	42.68
8	1.48	15.67	+	15.67
9	4.74	35.57	+	35.57
10	2.10	16.43	+	16.43

11	-1.84	-14.42	+	14.42
12	-2.65	-11.21	+	11.21
13	1.27	3.88	+	3.88
14	-10.74	-66.75	+	66.75
15	0.68	12.57	+	12.57
16	-3.15	-13.78	+	13.78
17	5.02	38.48	+	38.48
18	-4.59	-20.88	+	20.88
19	4.07	30.51	+	30.51
20	4.83	31.87	+	31.87
21	0.33	-1.62	-	1.62
22	-6.41	-41.11	+	41.11
23	-8.87	-63.85	+	63.85
25	-4.67	-27.65	+	27.65
26	-3.06	-20.46	+	20.46
27	-8.86	-62.04	+	62.04
28	0.53	8.62	+	8.62
29	-7.03	-41.01	+	41.01
30	-12.50	-80.29	+	80.29
31	-2.74	-15.50	+	15.50
32	-10.51	-66.23	+	66.23
33	0.78	6.80	+	6.80
34	0.52	-0.15	-	0.15
35	-7.03	-45.58	+	45.58
36	-6.44	-38.57	+	38.57
37	-10.46	-70.71	+-	70.71
38	0.12	-5.97	+	5.97
39	-2.46	-20.20	+	20.20
40	1.78	7.80	+	7.80
41	-0.86	-8.91	+	8.91
42	-8.08	-55.53	+	55.53
43	-1.54	-15.58	+	15.58
44	-0.98	-10.86	+	10.86
45	-18.36	-117.48	+	117.48
46	-3.94	-25.17	+	25.17
47	1.18	6.25	+	6.25

Приметимо да је $z_i = y_i - \boldsymbol{b}'\boldsymbol{x_i}$ и $w_i = \boldsymbol{d_1}'\boldsymbol{x_i}$. На пример

$$z_5 = 2.152 - [47.93 \quad -23.26 \quad -0.1161 \quad -2.443]\begin{bmatrix}1\\0.814\\53\\9.730\end{bmatrix} = 3.08$$

и

$$w_5 = [-284.9 \quad 164.5 \quad 0.5233 \quad 14.59]\begin{bmatrix}1\\0.814\\53\\9.730\end{bmatrix} = 18.70$$

(Разлика између 18.70 и 18.71 је због заокруживања).

Ми смо изабрали почетни вектор $\boldsymbol{b}$ такав да задовољава једнакост $y_i = \boldsymbol{b}'x_i$ за $i = 1,2,3,4$. Због овога важи $z_1 = z_2 = z_3 = z_4$. i-та врста матрице $\boldsymbol{A}$ је $\boldsymbol{x_i}'$ а прва колона матрице $\boldsymbol{A^{-1}}$ је $\boldsymbol{d_1}$, па се у матрици производа $\boldsymbol{AA^{-1}}$, на месту $(i, 1)$, налази $\boldsymbol{x_i}'\boldsymbol{d_1}$ што је баш w_i. Имамо да је $w_1 = 1$, $w_2 = w_3 = w_4 = 0$.

Након израчунавања z_i и w_i , одређујемо знак количника z_i/w_i. Ако је $w_i = 0$ овај количник није дефинисан, али ово није ни битно јер такве тачке из базе података не доприносе својом вредношћу укупној вредности извода, јер је $|w_i| = 0$. Сада ћемо да саберемо све $|w_i|$ за које је количник z_i/w_i или негативан или нула, и сабраћемо све $|w_i|$ за које је количник z_i/w_i позитиван. Одавде добијамо да извод, за смер d_1, износи -1221.

Када смо израчунали вредност извода за $\boldsymbol{d_1}$, можемо на једноставнији начин да израчунамо вредност извода за $-\boldsymbol{d_1}$. Табела ће се само мало променити. Вредности за z_i и $|w_i|$ ће остати исте, али ће се знаци од w_i променити, па ће се и знаци израза z_i/w_i променити, осим тамо где је $z_i = 0$. Постоје четири тачке из узорка код којих је $z_i = 0$, једна код које је $w_i = 1$, а код остале три је $w_i = 0$. Одавде следи да је извод за смер $-\boldsymbol{d_1}$, $-(-1221 - 1) + 1 = 1223$.

Изводи за четири колоне матрице $\boldsymbol{A^{-1}}$ и за њихове супротне векторе, су $-1221, 1223, -1323, 1325, 654, -652, 1903$ и -1901 . Најнегативнији од ових је -1901 у смеру $-\boldsymbol{d_4}$. Даље тражимо бољи вектор облика $\boldsymbol{b} + t\boldsymbol{d_4}$.

Нека је $z_i = y_i - \boldsymbol{b}`\boldsymbol{x_i}$ и $w_i = \boldsymbol{d_4}`\boldsymbol{x_i}$. Вредност за t која даје најбољи вектор је нагиб најбоље праве која садржи тачку $(0,0)$, а која одговара подацима (w_i, z_i), $i = 1,2, \dots, 47, i \neq 7,24$. Алгоритам који се овде користи смо већ објаснили. Применом овог алгоритма налазимо да најбоља права, која садржи тачку $(0,0)$, садржи и тачку (w_{14}, z_{14}). Због тога су четврте врсте матрица $\boldsymbol{A}$ и $\boldsymbol{c}$ замењене подацима из четврти 14. Односно у следећем кораку имамо

$$\boldsymbol{A} = \begin{bmatrix}1 & 0.604 & 29 & 11.744\\ 1 & 0.765 & 44 & 9.323\\ 1 & 0.735 & 36 & 9.948\\ 1 & 0.077 & 11 & 13.686\end{bmatrix} \quad \text{и} \quad \boldsymbol{c} = \begin{bmatrix}1.825\\2.251\\2.351\\0.693\end{bmatrix}.$$

У сваком кораку, четири тачке одређују тренутну регресиону једнакост. Њих некада зовемо и базне тачке. У сваком кораку, једну од базних тачака замењујемо тачком која не припада бази. Поступак почињемо са тачкама 1, 2, 3 и 4 као иницијалном базом.

У првом кораку смо тачку 4 заменили тачком 14. Како се алгоритам наставља, тачка 3 се замењује тачком 29, 2 тачком 26, 1 тачком 23, 26 тачком 10, 29 тачком 39, 10 тачком 19, 14 тачком 45, и 23 тачком 37. Након овог последњег корака, наша база се састоји од тачака 37, 19, 39 и 45. Када смо израчунали изводе за 8 смерова, нашли смо да су сви они позитивни и алгоритам се зауставља. Регресиона једнакост је одређена тачкама 37, 19, 39 и 45:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 0.020 & 17 & 13.842 \\ 1 & 0.842 & 29 & 11.084 \\ 1 & 0.559 & 42 & 10.332 \\ 1 & 0.114 & 3 & 10.080 \end{bmatrix}^{-1} \begin{bmatrix} 1.224 \\ 1.974 \\ 2.351 \\ 1.946 \end{bmatrix} = \begin{bmatrix} 4.362 \\ -0.09098 \\ 0.01299 \\ -0.2425 \end{bmatrix}$$

$$\hat{Y} = 4.362 - 0.09098X_1 + 0.01299X_2 - 0.2425X_3.$$

3.4.2 Нејединственост и дегенеративност

Вишеструка регресија најмањих апсолутних одступања користи обичну регресију најмањих апсолутних одступања за проналажење најбољег вектора $\boldsymbol{b}$ за одређени смер. У првом делу смо објаснили како нејединственост и дегенеративност могу утицати на процедуру. Сличан проблем се јавља када више од једног смера даје најнегативнију вредност извода или када је најмања вредност извода једнака нули.

Без обзира на појаву неједнствености или дегенеративности у алгоритму, ако алгоритам дође до корака у коме су изводи позитивни за свих $2(p+1)$ смерова, онда је тренутни вектор оцена онај који узимамо за оптимално решење. Ако алгоритам дође до корака у ком су изводи или позитивни или нула у свим смеровима, онда је тренутни вектор оцена онај који узимамо, а остали вектори могу да се нађу у правцима са изводом једнаким 0.

У неким случајевима ће се десити да се алгоритам врти у круг, не стижући никад у корак у коме су сви изводи ненегативни. Постоји неколико решења за овај проблем. Теоретски, како је за вектор оцене (или бар један од њих, у случају неједнствености) познато да садржи бар $p+1$ тачака, ми можемо формирати све могуће подскупове од $p+1$ тачака, који одређују одговарајући вектор оцене ($\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$, као што је описано), оценимо суму апсолутних одступања за сваки вектор и онај са најмањом сумом ћемо узети за тражени вектор. Међутим, за велике базе података, ово захтева јако велики број израчунавања. За узорак „Пожар“, са $n = 45$ и $p = 3$, било би 148995 могућих подскупова тачака.

Много изводљивија стратегија била би да поновимо алгоритам тако што ћемо узети други вектор да нам буде почетни вектор оцена. Још једна стратегија била би да генеришемо неколико веома малих случајних бројева и да их додамо нашим променљивим. Ово би требало да елиминише кружење и вектор оцена који се на овај начин добија би требао да буде и тражени вектор оцена. Ово може бити проверено испитивањем да ли су изводи у свих $2(p+1)$ смерова ненегативни.

Дегенеративност понекад доводи до ситуације у којој је вредност за t, која минимизира вредност израза (3.9) једнака $t^* = 0$. Ако дође до овога ми можемо узети другачији вектор $\boldsymbol{d}$, тако да водимо рачуна о томе да извод у новом смеру $\boldsymbol{d}$ буде такође негативан.

3.4.3 Тестирање хипотезе $\boldsymbol{\beta_{q+1} = \cdots = \beta_p = 0}$

Размотримо општи линеарни регресиони модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$. Тест-статистика за тестирање хипотезе $\beta_{q+1} = \cdots = \beta_p = 0$ код методе најмањих квадрата је

$$F_{\text{MNK}} = \frac{SKR_{uprošćen} - SKR_{neuprošćen}}{(p-q)\hat{\sigma}^2}$$

где је SKR сума квадрата резидуала, $SKR = \sum_{i=1}^{n} \hat{e}_i^2$. Упрошћен модел је модел $Y = \beta_0 + e$, односно модел који не садржи промењљиве $X_1, \ldots, X_p$. Слична тест-статистика се користи и код методе најмањих апсолутних одступања:

$$F_{MNAO} = \frac{SAR_{uprošćen} - SAR_{neuprošćen}}{(p-q)(\frac{\hat{\tau}}{2})} \tag{3.11}$$

где је SAR сума апсолутних вредности резидуала, $SAR = \sum_{i=1}^{n} |\hat{e}_i|$. Оцена $\hat{\tau}$ дата је у (3.5) где је $m = n - (p+1)$.

Када за резидуале претпостављамо да имају нормалну расподелу и када је нулта хипотеза тачна, тест статистика F_{MNK} има F расподелу, без обзира на обим узорка n. Када расподела случајних грешака није позната и када је нулта хипотеза тачна, онда F_{MNK} приближно има F расподелу, али под условом да је обим узорка n велики број. Ово исто важи и за тест-статистику F_{MNAO}. За велике вредности n ми можемо израчунати приближну p-вредност теста, као $P[F \geq F_{MNAO}]$, где је F случајна промењљива која има F расподелу са $p-q$ и $n-p-1$ степени слободе.

Апроксимација p-вредности је побољшана следећом модификацијом. Израчунаћемо p-вредност као $P[G \geq (p-q)(1-(p-q)/n)F_{MNAO}]$, где је G случајна променљива која има χ^2 расподелу са $p-q$ степени слободе. За јако велике вредности n, ово је скоро исто као и $P[F \geq F_{MNAO}]$, јер $1-(p-q)/n \approx 1$ и $(p-q)F$ са $n = \infty$ има исту расподелу као G.

Можемо да применимо сада тест који смо описали, са $p = 3$ и $q = 0$, на базу „Пожар“, како бисмо тестирали $\beta_1 = \beta_2 = \beta_3 = 0$, односно како бисмо тестирали да ли промењљиве Старост, Крађе, и Приход имају значајан утицај на промењљиву Пожар.

Треба да израчунамо F_{MNAO}. Да бисмо ово урадили, прво морамо да израчунамо резидуале $\hat{e}_i = y_i - (4.362 - 0.0909x_{i1} + 0.01299x_{i2} - 0.2425x_{i3})$ који су дати у табели (3.5). Сума њихових апсолутних вредности је $SAR_{neuprošćen} = 15.78$. $m = 41$ (број резидуала који су различити од нуле) је искоришћено за израчунавање $\hat{\tau}$. Како је $\frac{41+1}{2} - \sqrt{41} = 14.60$, имамо $k_1 = 15$. Слично, $k_2 = 27$. Из табеле (3.5) можемо да видимо да је $\hat{e}_{(15)} = -0.1492$ и $\hat{e}_{(27)} = 0.2488$. Одавде је $\hat{\tau} = \frac{\sqrt{41}(0.2488+0.1492)}{4} = 0.6371$. Упрошћен модел је $Y = \beta_0 + e$. Оцена овог модела је једноставно

узорачка медијана $\overline{y}$, што за базу „Пожар“ износи 2.251. Добијамо $SAR_{uprošćen} = \sum|y_i - 2.251| = 27.29$. Даље је $F_{MNAO} = \frac{27.29-15.78}{3\left(\frac{0.6371}{2}\right)} = 12.04$.

Да бисмо добили p-вредност, прво израчунамо $(p-q)\left(1-\frac{p-q}{n}\right)F_{MNAO} = 3\left(1-\frac{3}{45}\right)(12.04) = 33.71$. p-вредност је приближно $P[G \geq 33.71]$, где G има χ^2 расподелу са 3 степена слободе. Из табеле са χ^2 расподелом, можемо да видимо да је $P[G \geq 16.29] = 0.001$. Видимо да је p-вредност мања од 0.001. Закључак је да бар неке од променљивих имају значајан утицај на променљиву Пожар.

Табела 3.5 Резидуали регресије методе најмањих апсолутних одступања за узорак „Пожар“, дати у растућем поретку.

-0.6652	-0.3219	0.0000	0.2488	0.7547
-0.5547	-0.2786	0.0000	0.2850	0.8047
-0.5261	-0.2323	0.0000	0.3277	0.8355
-0.5202	-0.1565	0.0000	0.3347	1.0740
-0.4857	-0.1492	0.0008	0.3439	1.2015
-0.4790	-0.1361	0.1006	0.3604	
-0.4647	-0.1012	0.1155	0.4457	
-0.4233	-0.0356	0.1563	0.4642	
-0.3914	-0.0188	0.2052	0.5815	
-0.3520	-0.0106	0.2468	0.5939	

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generaliyations. Least squares and alternatives.1995.*

[3] Julian J.Fraway. Linear models with R,2005.

[4] N.H.Bingham, John M.Fry.*Regression. Linear models in statistics.* 2010.

[5] Hill, R.W., and P.W.Holland. *Two robust alternatives to least squares regression*. Journal of the American Statistical Association. 1977.

4. M регресија

M-регресија (M-оцена) и Метода најмањих апсолутних одступања спадају у такозване робусне методе. Статистичка процедура се назива робусном ако показује добре резултате и на подацима који немају искључиво нормалну расподелу. Питер Хубер[1] (Peter Huber) је увео M-регресију 1964.године.

4.1 Оцена регресионих коефицијената

Код методе најмањих квадрата коефицијенти $\hat{\alpha}$ и $\hat{\beta}$ су изабрани тако да $\sum_{i=1}^{n} \hat{e}_i^2$ буде најмања могућа. Код методе најмањих апсолутних одступања, они су изабрани тако да $\sum_{i=1}^{n} |\hat{e}_i|$ буде најмања могућа. Код M-оцене, ова идеја је уопштена и $\hat{\alpha}$ и $\hat{\beta}$ су изабрани тако да $\sum_{i=1}^{n} \rho(\hat{e}_i)$ буде најмања могућа, где је $\rho(e)$ нека функција од e. Метода најмањих квадрата и метода најмањих апсолутних одступања могу бити схваћене као специјални случајеви M-оцене, где је $\rho(e) = e^2$ и $\rho(e) = |e|$.

M-оцена коју ћемо описати у овом делу, зове се Хуберова M-оцена и код ње је функција $\rho(e)$, функција која је комбинација функција e^2 и $|e|$. Највећа предност методе најмањих апсолутних одступања над методом најмањих квадрата је то што она није осетљива на аутлајере. Где немамо аутлајера боље се показује метода најмањих квадрата. Ми ћемо покушати да искомбинујемо предности обе ове методе, дефинисањем функције $\rho(e)$ на следећи начин. Ставићемо да је $\rho(e)$ једнако e^2, када је e близу нуле, а да је $\rho(e)$ једнако $|e|$ (или бар слично са $|e|$), када је e далеко од нуле.

На пример, $\rho(e)$ можемо да дефинишемо на следећи начин

$$\rho(e) = \begin{cases} e^2 & \text{ако је } -k \le e \le k \\ 2k|e| - k^2 & \text{ако је } e < -k \text{ или } k < e \end{cases}. \tag{4.1}$$

Хубер је предложио да буде $k = 1.5\hat{\sigma}$, где је $\hat{\sigma}$ оцена стандардног одступања σ за популацију случајних грешака. Уместо $|e|$ коришћена је функција $2k|e| - k^2$, да би функција $\rho(e)$ била глатка.

Да бисмо оценили σ, користићемо $\hat{\sigma} = 1.483MAO$, где је MAO медијана апсолутних одступања од $|\hat{e}_i|$. Чинилац 1.483 је изабран тако да $\hat{\sigma}$ буде добра оцена параметра σ и у случају када су случајне грешке нормално расподељене.

Хуберове M-оцене $\hat{\alpha}$ и $\hat{\beta}$ су вредности a и b које минимизирају вредност суме

$$\sum_{i=1}^{n} \rho(y_i - (a + bx_i)). \tag{4.2}$$

Приметимо да a и b, осим што се директно појављују у изразу (4.2), појављују се индиректно и у дефиницији функције ρ.

---

(1) Питер Хубер (рођен 1934.године) је швајцарски статистичар.

За почетне вредности α и β узећемо оцене добијене методом најмањих квадрата. Њих користимо да израчунамо одступање и оцену за σ. Такође ћемо их користити како бисмо добили још боље вредности за α и β, што ће бити описано у следећем пасусу. Ове, сада побољшане вредности за α и β ћемо користити за добијање нових одступања и побољшаних вредности за σ. Даље се те нове вредности за σ користе за добијање побољшаних вредности за α и β. Алгоритам се на овај начин понавља све до тренутка када су вредности побољшаних оцена исте као и претходне.

Како бисмо били мало прецизнији, узмимо a^0 и b^0 као тренутне оцене за α и β. Израчунајмо одступања $y_i - (a^0 + b^0 x_i)$ и помоћу њих израчунајмо $\hat{\sigma}^0 = 1.483MAO$. Сада ћемо да прилагодимо вредности за y да бисмо елиминисали велика одступања. Одступање вредности y_i од тренутне регресионе праве је $e_i^0 = y_i - (a^0 + b^0 x_i)$. Одавде је $y_i = a^0 + b^0 x_i + e_i^0$. Дефинишимо сада $y_i^* = a^0 + b^0 x_i + e_i^*$, где је вредност e_i^* добијена скраћивањем e_i^0 тако да ниједно од одступања није веће од апсолутне вредности израза $1.5\hat{\sigma}^0$. Одавде следи да је $e_i^* = e_i^0$ (отуда $y_i^* = y_i$) када је e_i^0 између $-1.5\hat{\sigma}^0$ и $1.5\hat{\sigma}^0$, $e_i^* = -1.5\hat{\sigma}^0$ када је e_i^* мање од $-1.5\hat{\sigma}^0$, а $e_i^* = 1.5\hat{\sigma}^0$ када је e_i^* веће од $1.5\hat{\sigma}^0$. Узмимо даље да су побољшане оцене за α и β, оцене добијене методом најмањих квадрата помоћу $y_1^*, \dots, y_n^*$.

Иако алгоритам делује оправдано, није у потпуности јасно како он минимизира вредност израза (4.2). Ако је $\hat{\sigma}$ фиксирано, ми можемо минимизирати израз (4.2) тако што ћемо наћи први извод од (4.2) и изједначити га са нулом. На овај начин добијамо две једначине са две непознате a и b:

$$\sum_{i=1}^{n} \rho'(y_i - (a + bx_i)) = 0$$

$$\sum_{i=1}^{n} x_i \rho'(y_i - (a + bx_i)) = 0 \qquad (4.3)$$

Приметимо да извод $\rho'(e)$ има исту вредност $-3\hat{\sigma}$ за све e мање или једнаке од $-1.5\hat{\sigma}$, а исту вредност $3\hat{\sigma}$ за све e веће од $1.5\hat{\sigma}$. Због тога ће решења система (4.3) остати иста ако одступања $e_i = y_i - (a + bx_i)$ заменимо њиховим скраћеним вредностима e_i^*, где је $e_i^* = e_i$ када је e_i између -1.5$\hat{\sigma}$ и $1.5\hat{\sigma}$, $e_i^* = -1.5\hat{\sigma}$ када је e_i мање од $-1.5\hat{\sigma}$, а $e_i^* = 1.5\hat{\sigma}$ када је e_i веће од $1.5\hat{\sigma}$. Односно, y_i могу бити замењени побољшаним вредностима $y_i^* = a_i + bx_i + e_i^*$ а да тиме не променимо решења система (4.3). Као резултат овога добијамо $\rho\big(y_i^* - (a_i + bx_i)\big) = [y_i^* - (a_i + bx_i)]^2$. Минимизирањем суме $\sum_{i=1}^{n}[y_i^* - (a_i + bx_i)]^2$ добијамо оцене добијене методом најмањих квадрата, а коришћењем побољшаних вредности y_i^*.

4.2 Тестирање хипотезе $\boldsymbol{\beta = 0}$

Део 4.3.2 описује како да тестирамо хипотезу $\beta_{q+1} = \cdots = \beta_p = 0$ у вишеструком линеарном регресионом моделу. Тестирање хипотезе $\beta = 0$ схватићемо као специјалан случај , где је $p = 1$ и $q = 0$. Тест статистика је F_M дата формулом (4.6). Приближна p-вредност може бити израчуната као $P[F > F_M]$, где је F случајна променљива која има F расподелу са 1 и $n - 2$ степени слободе или као $P[|t| \geq |t_M|]$, где је $|t_M| = \sqrt{F_M}$ и t представља случајну променљиву која има t расподелу са $n - 2$ степени слободе.

Пример 4.1

Као и претходне две методе, и М-регресију ћемо демонстрирати на истом примеру. Подаци су нам дати у табели 2.1, на страни 9.

За почетне вредности параметара α и β ћемо узети оцене које смо добили методом најмањих квадрата, $a^0 = 42.9905$ и $b^0 = -0.3989$. Израчунаћемо вредности $\hat{y}_i^0 = a^0 + b^0 x_i$ и одступања $e_i^0 = y_i - \hat{y}_i^0$, као у табели 4.1. Медијана апсолутних одступања $|e_i^0|$ је $MAO = 4.67$. Па је $\hat{\sigma}^0 = 1.483 \cdot 4.67 = 6.92561$ и $1.5\hat{\sigma}^0 = 10.388415$. Сада ћемо да смањимо одступања e_i^0 да добијемо e_i^*. Два од ових одступања су мања од $1.5\hat{\sigma}^0$, то су $e_3^0 = -12.78$ и $e_{13}^0 = -15.68$. Ова два су повећана на $e_3^* = e_{13}^* = -10.388415$. Једно одступање је веће од $-1.5\hat{\sigma}^0$, то је одступање $e_{11}^0 = 12.6$. Оно је смањено на $e_8^* = 10.388415$. Прилагођене y вредности су добијене као $y_i^* = \hat{y}_i^0 + e_i^*$. Приметимо да је $y_i^* = y_i$, за све i осим за $i = 3, 11$ и 13.

Ако, у табели (2.1), заменимо y вредности прилагођеним y вредностима (вредностима из последње колоне табеле 4.1), затим на те вредности применимо методу најмањих квадрата, добићемо нове оцене $a^0 = 44.4291$ и $b^0 = -0.4340$. Овим смо завршили прву итерацију алгоритма.

Табела 4.1 Израчунавања у првом кораку алгоритма за добијање М-оцена, за узорак „Наталитет“.

y_i	$\hat{y}_i^0$	e_i^0	e_i^*	y_i^*
16.2	21.0	-4.8	-4.8	16.2
30.5	32.08	-1.58	-1.58	30.5
16.9	29.68	-12.78	-10.39	19.29
33.1	28.16	4.94	4.94	33.1
40.2	38.40	1.8	1.8	40.2
38.4	37.32	1.08	1.08	38.4
41.3	37.44	3.86	3.86	41.3
43.9	35.40	8.5	8.5	43.9
28.3	29.76	-1.46	-1.46	28.3
33.9	25.72	8.18	8.18	33.9
44.2	31.60	12.6	10.39	41.98
28.0	27.92	0.08	0.08	28.0
24.6	40.28	-15.68	-10.39	29.89
16.0	20.40	-4.4	-4.4	16.0

За другу итерацију морамо да направимо табелу сличну табели 4.1. Прва колона остаје иста. Нове оцене a^0 и b^0 се сада користе за израчунавање y вредности $\hat{y}_i^0$ за другу колону. Након што

израчунамо и последњу колону у табели, користимо методу најмањих квадрата да добијемо нове вредности $a^0 = 45.0360$ и $b^0 = -0.4471$. Итерација се понавља све док оцене не почну да конвергирају ка неком броју. Након седам итерација, разлика између добијене оцене и претходне оцене је мања од 10^{-4} и за α и за β. Одавде добијамо да је $\hat{\alpha} = 46.3309$ и $\hat{\beta} = -0.4836$.

Пример 4.1 у R-у

Прво уносимо променљиве.

```
> X<-c(55.0,27.3,33.3,37.1,11.5,14.2,13.9,19.0,33.1,43.2,28.5,37.7,6.8,56.5)

> Y<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9,28.3,33.9,44.2,28.0,24.6,16.0)
```

Затим учитавамо пакет „MASS“ у коме нам се налазе потребне функције.

```
> library(MASS)
```

Модел правимо помоћу функције „rlm“ а затим потребне информације о моделу добијамо помоћу функције „summary“.

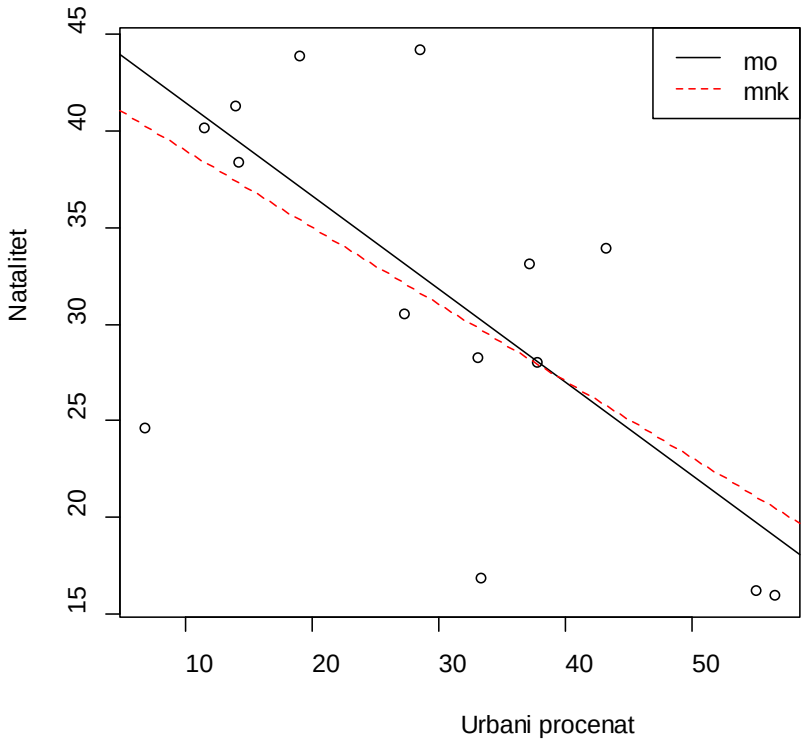
```
> mo<-rlm(Y~X)

> summary(rlm)

Error in object[[i]] : object of type 'closure' is not subsettable

> summary(mo)


Call: rlm(formula = Y ~ X)

Residuals:

    Min      1Q  Median      3Q     Max

-18.4425  -2.9133  -0.8167   3.9555  11.6514


Coefficients:

            Value   Std. Error t value

(Intercept) 46.3309  4.7765     9.6997

X           -0.4836  0.1432    -3.3771
```

Residual standard error: 4.851 on 12 degrees of freedom

Добијена регресиона права је $Y = 46.3309 - 0.4836X$.

Сада ћемо да представимо добијену праву на графику заједно са правом добијеном помоћу методе најмањих квадрата.

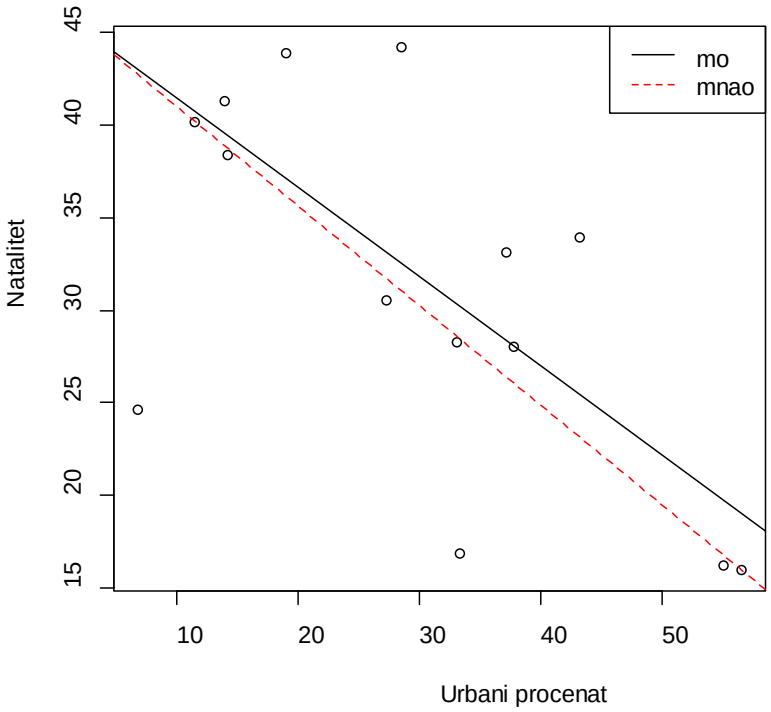
```
> plot(Y~X,xlab="Urbani procenat",ylab="Natalitet")
> abline(mo)
> abline(lm(Y~X),col=2,lty=2)
> legend("topright",legend=c("mo","mnk"),col=1:2,lty=1:2)
```



Слика 4.1 Регресионе праве за узорак „Наталитет“ добијене М-оценом и методом најмањих квадрата.

Како бисмо упоредили оцене регресионих правих добијених методом најмањих апсолутних одступања и М-регресијом, на истом графику ћемо представити ове две праве.

```
> plot(Y~X,xlab="Urbani procenat",ylab="Natalitet")

> abline(mo)

> abline(rq(Y~X),col=2,lty=2)

> legend("topright",legend=c("mo","mnao"),col=1:2,lty=1:2)
```



Слика 4.2 Регресионе праве за узорак „Наталитет“ добијене М-оценом и методом најмањих апсолутних одступања.

4.3 Пример вишеструке регресије

Посматрајмо сада узорак „Аеробик фитнес“, дат у табели 4.2. Променљиве у овом узорку су стопа потрошње кисеоника (у милилитрима, по килограму тежине, у минуту); старост (у годинама); тежина (у килограмима); време (у минутима) потребно да се истрчи 1,5 миља; пулс (број откуцаја срца у минуту) за време одмарања; и пулс по завршетку трчања. Ми можемо користити ове податке да изразимо стопу потрошње кисеоника као функцију зависну од осталих променљивих. Линеарни регресиони модел је $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$.

Табела 4.2 Узорак „Аеробик фитнес“.

Идентификациони број особе	Потрошња Кисеоника (Y)	Старост (X_1)	Тежина (X_2)	Време трчања (X_3)	Пулс у мировању (X_4)	Пулс након трчања (X_5)
1	44.609	44	89.47	11.37	62	178
2	45.313	40	75.07	10.07	62	185
3	54.297	44	85.84	8.65	45	156
4	59.571	42	68.15	8.17	40	166
5	49.874	38	89.02	9.22	55	178
6	44.811	47	77.45	11.63	58	176
7	45.681	40	75.98	11.95	70	176
8	49.091	43	81.19	10.85	64	162
9	39.442	44	81.42	13.08	63	174
10	60.055	38	81.87	8.63	48	170
11	50.541	44	73.03	10.13	45	168
12	37.388	45	87.66	14.03	56	186
13	44.754	45	66.45	11.12	51	176
14	47.273	47	79.15	10.60	47	162
15	51.855	54	83.12	10.33	50	166
16	49.156	49	81.42	8.95	44	180
17	40.836	51	69.63	10.95	57	168
18	46.672	51	77.91	10.00	48	162
19	46.774	48	91.63	10.25	48	162
20	50.388	49	73.37	10.08	76	168
21	39.407	57	73.37	12.63	58	174
22	46.080	54	79.38	11.17	62	156
23	45.441	52	76.32	9.63	48	164
24	54.625	50	70.87	8.92	48	146
25	45.118	51	67.25	11.08	48	172

26	39.203	54	91.63	12.88	44	168
27	45.790	51	73.71	10.47	59	186
28	50.545	57	59.08	9.93	49	148
29	48.673	49	76.32	9.40	56	186
30	47.920	48	61.24	11.50	52	170
31	47.467	52	82.78	10.50	53	170

4.3.1 Оцењивање регресионих коефицијената

Процедура за проналажење М-оцена код вишеструке регресије представља генерализовање процедуре једноструке регресије, описане у делу 4.1. Хуберове М-оцене $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, су вредности $b_0, b_1, \dots, b_p$ које минимизирају суму

$$\sum_{i=1}^{n} \rho(y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip})) \qquad (4.4)$$

где је $\rho(e)$ функција дефинисана са (4.1), p је број објашњавајућих променљивих. Згодније је да користимо векторску нотацију. Нека је

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \quad \text{и} \quad \boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}.$$

Вектор $\widehat{\boldsymbol{\beta}}$ код Хуберове М-оцене је дефинисан да буде вектор који минимизира суму $\sum_{i=1}^{n} \rho(y_i - \boldsymbol{b}`\boldsymbol{x_i})$.

Алгоритам

Вектор регресионих коефицијената, означен са $\boldsymbol{\beta}$, је оцењен помоћу вектора оцена добијених методом најмањих квадрата. Овај почетни вектор $\boldsymbol{\beta}$ користимо да израчунамо одступања и оцену за σ. Њих даље користимо, као што ћемо описати у следећем пасусу, како бисмо добили побољшану вредност за $\boldsymbol{\beta}$. Алгоритам се наставља на овај начин све до тренутка када је побољшана вредност $\boldsymbol{\beta}$ једнака (или приближно једнака), са претходном вредношћу.

Како бисмо били конкретнији, у сваком кораку алгоритма, означимо са $\boldsymbol{b_0}$ тренутну вредност од $\boldsymbol{\beta}$. Израчунајмо одступања $y_i - (\boldsymbol{b^0})'\boldsymbol{x_i}$, па помоћу њих израчунајмо $\hat{\sigma} = 1.483MAO$. Даље ћемо да прилагодимо вредности за y, како бисмо се ослободили великих одступања. Одступање y_i од тренутне регресионе праве је $e_i^0 = y_i - (\boldsymbol{b^0})'\boldsymbol{x_i}$. Сада дефинишимо $y_i^* = (\boldsymbol{b^0})'\boldsymbol{x_i} + e_i^*$, где је e_i^* прилагођена вредност одступања, добијена тако што смо e_i^0 модификовали, тако да ниједно одступање не буде веће од апсолутне вредности израза $1.5\hat{\sigma}$. На крају ћемо узети да је побољшана вредност вектора $\boldsymbol{\beta}$, оцена добијена методом најмањих квадрата помоћу прилагођених вредности $y_1^*, \dots, y_n^*$.

За почетни вектор регресионих коефицијената, узећемо вектор добијен методом најмањих квадрата, што је $\boldsymbol{b^0} = (116.0, -0.2802, -0.05063, -2.743, -0.01224, -0.1279)$. Сада израчунајмо одговарајуће y вредности $\hat{y}_i^0 = (b^0)'x_i$ и одступања $e_i^0 = y_i - \hat{y}_i^0$, као у табели 4.3. Медијана апсолутних одступања $\left|e_i^0\right|$ је $MAO = 1.229$. Па је $\hat{\sigma} = (1.483)(1.229) = 1.822$ и $1.5\hat{\sigma} = 2.733$. Сада ћемо да коригујемо одступања e_i^0 како бисмо добили e_i^*. Три одступања су већа од $1.5\hat{\sigma}$, то су $e_4^0 = 2.873$, $e_{10}^0 = 4.802$, и $e_{15}^0 = 5.324$. Ова одступања су смањена на $e_4^* = e_{10}^* = e_{15}^* = 2.733$. Три одступања су мања од $-1.5\hat{\sigma}$, то су $e_2^0 = -3.685$, $e_{17}^0 = -5.177$ и $e_{23}^0 = -4.195$. Ова одступања су повећана на $e_2^* = e_{17}^* = e_{23}^* = -2.733$. Сада су прилагођене y вредности добијене као $y_i^* = \hat{y}_i^0 + e_i^*$. Приметимо да је $y_i^* = y_i$ за све i осим за $i = 2,4,10,15,17$ и 23.

Табела 4.3 Израчунавања у првој итерацији алгоритма за добијање М-оцена за узорак „Аеробик фитнес“.

Посматрана y вредност y_i	**Прилагођена y вредност** $\hat{y}_i^0$	**Одступање** e_i^0	**Скраћено одступање** e_i^*	**Коригована вредност** y_i^*
44.609	44.478	0.131	0.131	44.609
45.313	48.998	-3.685	-2.733	46.265
54.297	55.143	-0.846	-0.846	54.297
59.571	56.698	2.873	2.733	59.431
49.874	52.164	-2.290	-2.290	49.874
44.811	43.838	0.973	0.973	44.811
45.681	44.849	0.832	0.832	45.681
49.091	48.625	0.466	0.466	49.091
39.442	40.695	-1.253	-1.253	39.442
60.055	55.253	4.802	2.733	57.986
50.541	50.198	0.343	0.343	50.541
37.388	36.044	1.344	1.344	37.388
44.754	46.439	-1.685	-1.685	44.754
47.273	48.502	-1.229	-1.229	47.273
51.855	46.531	5.324	2.733	49.264
49.156	50.086	-0.930	-0.930	49.156
40.836	46.013	-5.177	-2.733	43.280
46.672	49.077	-2.405	-2.405	46.672
46.774	48.537	-1.763	-1.763	46.774
50.388	48.537	1.851	1.851	50.388
39.407	38.755	0.652	0.652	39.407
46.080	45.549	0.531	0.531	46.080
45.441	49.636	-4.195	-2.733	46.903

54.625	54.722	-0.097	-0.097	54.625
45.118	45.376	-0.258	-0.258	45.118
39.203	38.925	0.278	0.278	39.203
45.790	44.796	0.994	0.994	45.790
50.545	50.319	0.226	0.226	50.545
48.673	48.196	0.477	0.477	48.673
47.920	45.576	2.344	2.344	47.920
47.467	46.094	1.373	1.373	47.467

Вектор почетних вредности је добијен применом методе најмањих квадрата на податке из табеле 4.2. Ако у тој табели заменимо y вредности, вредностима из последње колоне табеле 4.3 и применимо методу најмањих квадрата, добићемо нови вектор $\boldsymbol{b^0} = (115.7, -0.2721, -0.07352, -2.694, -0.00059, -0.1245)$. Овим завршавамо прву итерацију алгоритма.

За другу итерацију ћемо користити нови вектор $\boldsymbol{b^0}$ да израчунамо нове вредности за $\hat{y}_i^0$. Направићемо табелу сличну табели 4.3. Из те табеле узећемо вредности из последње колоне и израчунати нови вектор $\boldsymbol{b^0} = (114.8, -0.2646, -0.07890, -2.678, 0.00601, -0.1228)$. Настављамо итерације све док не добијемо да оцене конвергирају ка неким вредностима. Након 18 итерација, добићемо да су нове вредности оцена веће од претходних за мање од 10^{-4}. Одавде добијамо да је $\widehat{\boldsymbol{\beta}} = (113.1, -0.2489, -0.07718, -2.654, 0.01475, -0.1216)$, па је регресиона једнакост

$$Y = 113.1 - 0.2489X_1 - 0.07718X_2 - 2.654X_3 + 0.01475X_4 - 0.1216X_5 \qquad (4.5)$$

4.3.2 Тестирање хипотезе $\boldsymbol{\beta_{q+1} = \cdots = \beta_p = 0}$

Тест-статистика за тестирање хипотезе $\beta_{q+1} = \cdots = \beta_p = 0$ код методе најмањих квадрата, за модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ је

$$F_{MNK} = \frac{SKR_{uprošćen} - SKR_{neuprošćen}}{(p-q)\hat{\sigma}^2_{MNK}}$$

где је SKR сума квадрата резидуала, $SKR = \sum_{i=1}^{n} \hat{e}_i^2$, а $\hat{\sigma}^2_{MNK} = \sum_{i=1}^{n} \hat{e}_i^2/(n-p-1)$. Резидуали из $SKR_{uprošćen}$ и $SKR_{neuprošćen}$ су израчунати применом методе најмањих квадрата, на, респективно, упрошћен модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q$ и неупрошћен модел $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ **.** Приликом израчунавања $\hat{\sigma}^2_{MNK}$ коришћени су резидуали из неупрошћеног модела.

Слична тест статистика се користи и код М-регресије:

$$F_M = \frac{STR_{uprošćen} - STR_{neuprošćen}}{(p-q)\widehat{\lambda}} \qquad (4.6)$$

где је STR сума трансформисаних резидуала , $STR = \sum_{i=1}^{n} \rho(e_i)$, и $\hat{\lambda} = (n/m)\sum_{i=1}^{n} \hat{e}_i^{*2}/(n-p-1)$. Функција $\rho(e)$ је уведена у делу (4.1) а вредност $\hat{e}_i^*$ је добијена смањивањем вредности $\hat{e}_i$, као што је раније описано. Оцена за $\hat{\sigma}$ је $1.483MAO$. Цео број m је број резидуала $\hat{e}_i$, којима није потребно смањивање, односно оних за које је $|\hat{e}_i| \leq 1.5\hat{\sigma}$. Резидуали у $STR_{uprošćen}$ и $STR_{neuprošćen}$ су израчунати применом М-регресије на, респективно, упрошћен модел и неупрошћен модел. Процедура за проналажење оцена код упрошћеног модела се мало разликује од процедуре описане у претходном делу. Током итерација потребних за добијање М-оцене регресионих коефицијената у упрошћеном моделу, задржавамо вредност за $\hat{\sigma}$, добијену из неупрошћеног модела. У израчунавању $\hat{\lambda}$ користимо резидуале из неупрошћеног модела. Приближна p-вредност теста је израчуната на исти начин као и код методе најмањих квадрата, као $P[F \geq F_M]$, где је F случајна променљива која има F расподелу са $p-q$ и $n-p-1$ степени слободе.

Сада ћемо да тестирамо да ли променљиве које представљају пулс имају значајан допринос једнакости (4.5); односно тестираћемо хипотезу $\beta_4 = \beta_5 = 0$. Имамо да је $p = 5$ и $q = 3$.

Да бисмо изарачунали F_M, прво морамо наћи резидуале $\hat{e}_i = y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{x_i}$, где је $\widehat{\boldsymbol{\beta}}$ вектор М-оцена. Медијана апсолутних резидуала $|\hat{e}_i|$ је $MAD = 0.9046$, па је $\hat{\sigma} = (1.483)(0.9046) = 1.341$ и $1.5\hat{\sigma} = (1.5)(1.341) = 2.012$. За 24 резидуала који се налазе између -2.012 и 2.012, $\rho(\hat{e}_i) = \hat{e}_i^2$, а за осталих 7 резидуала, $\rho(\hat{e}_i) = 4.024|\hat{e}_i| - 4.049$. Сабирањем свих вредности $\rho(\hat{e}_i)$, добијамо $STR_{neuprošćen} = 114.7$. Смањивањем 4 резидуала који су испод -2.012 и 3 резидуала који су изнад 2.012, добијамо смањене вредности резидуала $\hat{e}_i^*$. Сума њихових квадрата је 50.44. Сада је $\hat{\lambda} = \frac{\left(\frac{31}{24}\right)(50.44)}{25} = 2.606$.

Упрошћен модел код кога је $\beta_4 = \beta_5 = 0$ је $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$. Хуберове М-оцене су вредности b_0, b_1, b_2, b_3, које минимизирају функцију $\sum \rho(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3}))$. Иста вредност $\hat{\sigma} = 1.341$ израчуната из неупрошћеног модела, се користи непромењена кроз све итерације алгоритма. Вредност k из дефиниције функције ρ, дате у (4.1), остаје константна, $k = 1.5\hat{\sigma} = (1.5)(1.341) = 2.012$.

Вектор М-оцена у упрошћеном моделу је $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (95.07, -0.1844, -0.08861, -3.015)$, па су резидуали у упрошћеном моделу $\hat{e}_i = y_i - (95.07 - 0.1844x_{i1} - 0.08861x_{i2} - 3.015x_{i3})$. За 21 резидуал између -2.012 и 2.012, $\rho(\hat{e}_i) = \hat{e}_i^2$, а за осталих 10, $\rho(\hat{e}_i) = 4.024|\hat{e}_i| - 4.049$. Сабирањем свих 31 вредности $\rho(\hat{e}_i)$, добијамо $STR_{neuprošćen} = 145.8$. Даље $F_M = \frac{145.8-114.7}{[(5-3)(2.606)]} = 5.964$.

Приближна p-вредност теста је $P[F \geq 5.964]$, где је F случајна променљива која има F расподелу са 2 и 25 степени слободе. Из табеле за F расподелу можемо да видимо да је p-вредност између 0.001 и 0.01. Одавде закључујемо да једна или обе променљиве које представљају пулс имају значајан утицај на потрошњу кисеоника.

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generaliyations. Least squares and alternatives.1995.*

[3] Julian J.Fraway. Linear models with R,2005.

[4] N.H.Bingham, John M.Fry.*Regression. Linear models in statistics.* 2010.

[5] Hill, R.W., and P.W.Holland. *Two robust alternatives to least squares regression*. Journal of the American Statistical Association. 1977.

5. Непараметарска регресија

Параметарска статистичка процедура подразумева да случајне грешке у узорку имају одређени тип расподеле. Метода најмањих квадрата је параметарска процедура јер приликом њене примене захтевамо да случајне грешке имају нормалну расподелу. Робусне процедуре, као што је М-регресија захтевају да случајне грешке имају расподелу сличну нормалној расподели. Непараметарске процедуре дају добре резултате за било коју расподелу случајних грешака. Непараметарске процедуре су засноване на идеји коришћења рангова бројева уместо самих бројева.

Најранија појава статистичке анализе засноване на ранговима је Галтоново упоређивање висина двеју типова биљака из 1876. године. Развој непараметарских ранг метода је почео у 1960-им и 1970-им годинама.

5.1 Оцењивање регресионе праве

Најразумнији начин проналажења нагиба регресионе праве је да пронађемо нагибе правих које садрже по две тачке из узорка, а затим да нађемо неку врсту средње вредности или медијане тих нагиба. Нагиб праве која садржи тачке (x_i, y_i) и (x_j, y_j) је $b_{ij} = \big(y_i - y_j\big)/(x_i - x_j)$. (Занемарићемо парове за које је $x_i = x_j$ јер је нагиб који одговара таквој правој недефинисан.) Формула за оцену параметра β је формула (2.1) али може бити представљена и као пондерисани просек нагиба b_{ij}. Специјално, $\hat{\beta}_{MNK} = \sum w_{ij} b_{ij}$, где је $w_{ij} = (x_i - x_j)^2 / \sum_{i=1}^{n} (x_i - x_j)^2$. (Имамо $n(n-2)/2$ парова целих бројева i и j за које важи $1 \leq i < j \leq n$.) Други разуман начин проналажења нагиба регресионе праве је узимање пондерисане медијане нагиба правих које садрже по две тачке из узорка.

Пондерисане медијане

Присетимо се појма медијане одређеног скупа бројева. Медијану скупа бројева добијамо тако што све бројеве поређамо у неопадајући поредак а затим, уколико имамо непаран број бројева, узимамо број који се налази у средини, а уколико имамо паран број бројева, узимамо аритметичку средину бројева који се налазе у средини.

Пондерисана медијана низа бројева x_i са пондерима w_i се добија на следећи начин: Прво поређамо бројеве x_i у неопадајући поредак. Мењајући индексе тим бројевима добијамо $x_1 \leq x_2 \leq \cdots \leq x_n$. Пондери су ненегативне величине и у збиру дају 1. Наћи ћемо индекс такав да важи

$$w_1 + w_2 + \cdots + w_{k-1} < 0.5$$

$$w_1 + w_2 + \cdots + w_{k-1} + w_k > 0.5 \qquad (5.1)$$

Тада је x_k пондерисана медијана. (Понекад се деси да постоји индекс k такав да је $w_1 + w_2 + \cdots + w_{k-1} = 0.5$. Тада је $(x_{k-1} + x_k)/2$ пондерисана медијана.) Претпоставимо да су нам дати бројеви $203, 235, 47, 219$ и 156 са пондерима $0.1, 0.1, 0.4, 0.1$ и 0.3 редом. Ови бројеви

поређани у неопадајућем поретку су $47, 156, 203, 219, 235$. Пондерисана медијана је број $x_2 = 156$ јер је $w_1 = 0.4 < 0.5$ и $w_1 + w_2 = 0.4 + 0.3 = 0.7 > 0.5$.

Када су сви пондери једнаки онда је $w_i = \frac{1}{n}$ па је пондерисана медијана заправо обична медијана.

Нагиб β^* у изразу (3.3), који представља нагиб праве која садржи тачку (x_0, y_0) која минимизира суму апсолутних одступања, може бити представљен као пондерисана медијана нагиба $b_i = (y_i - y_0)(x_i - x_0)$ правих које садрже тачку (x_i, y_i) из узорка и дату тачку (x_0, y_0), са пондерима који су пропорционални растојању $|x_i - x_0|$.

Непараметарске оцене параметара α и β.

Нека је $\hat{\beta}$ пондерисана медијана нагиба $b_{ij} = (y_i - y_j)/(x_i - x_j)$ чији су пондери пропорционални растојању x координата одређених двају тачака, односно $w_{ij} = |x_i - x_j| / \sum|x_i - x_j|$. Нека је $\hat{\alpha}$ обична медијана разлика $y_i - \hat{\beta}x_i$. Приметимо да су у моделу $y_i = \alpha + \beta x_i + e_i$ разлике $y_i - \beta x_i$ центриране око α, па је ова оцена параметра α разумна.

Оцене параметара α и β треба изабрати тако да резидуали $\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$ буду мали. Један начин мерења њихове величине је помоћу средњих вредности пондерисаних сума апсолутних вредности резидуала, $\sum_{i=1}^{n} w_i|\hat{e}_i|$. Пондери w_i би требало да буду ненегативни. У методи најмањих квадрата, параметре $\hat{\alpha}$ и $\hat{\beta}$ бирамо тако да они минимизирају пондерисану суму са пондерима $w_i = |\hat{e}_i|$. У методи најмањих апсолутних одступања минимизирамо пондерисану суму са пондерима $w_i = 1$. Још један начин бирања пондера је да узмемо да су $w_i = rang(|\hat{e}_i|)$ (најмањој вредности $|\hat{e}_i|$ додељен је најмањи ранг, 1). Ово ће ограничити утицај великих резидуала у већој мери него што то чини метода најмањих квадрата, обзиром да $rang(|\hat{e}_i|)$ не може бити веће од n, док $|\hat{e}_i|$ може бити произвољно велико, али у мањој мери него што то чини метода најмањих апсолутних одступања. Ми ћемо користити процедуру која је слична овој.

Уместо да изаберемо оцене које минимизирају суму $\sum_{i=1}^{n} rang(|\hat{e}_i|)|\hat{e}_i|$ ми ћемо изабрати оцене које минимизирају суму

$$\sum_{i=1}^{n} \left[rang(\hat{e}_i) - \frac{n+1}{n}\right] \hat{e}_i. \qquad (5.2)$$

Обе суме дају исте оцене, кад популација случајних грешака има симетричну расподелу и када је обим узорка n велики. Од суме (5.2) се очекује да да боље резултате када расподела случајних грешака није симетрична.

Касније ћемо видети да непараметарска оцена параметра $\hat{\beta}$, која је представљена као пондерисана медијана, може такође бити окарактерисана као вредност b, која минимизира (5.2), где је $\hat{e}_i = y_i - (a + bx_i)$. Али прво ћемо објаснити везу између ове две суме.

Ако знамо да је расподела случајних грешака приближно симетрична, онда можемо очекивати да $rang(|\hat{e}_i|)$ буде приближно једнак са $2\left|rang(\hat{e}_i) - \frac{1}{2}(n+1)\right|$. Ово важи јер је $\frac{1}{2}(n+1)$ ранг

медијане резидуала, за који очекујемо да буде 0, а такође и зато што би апсолутне вредности негативних резидуала требало да буду равномерно распршене, са позитивним вредностима. На пример, претпоставимо да су резидуали $-23, -18, -11, 2, 16, 19, 29$. Вредности за $rang(|\hat{e}_i|)$ су редом $6, 4, 2, 1, 3, 5, 7$. Вредности за $2\left|rang(\hat{e}_i) - \frac{1}{2}(n+1)\right|$ су редом $6, 4, 2, 0, 2, 4, 6$. Приметимо да дељење свих пондера бројем 2 неће утицати на минимизирање пондерисане суме $\sum w_i|\hat{e}_i|$. Одавде закључујемо да је коришћење пондера $w_i = rang(|\hat{e}_i|)$ исто што и коришћење пондера $w_i = \left|rang(\hat{e}_i) - \frac{1}{2}(n+1)\right|$.

Ако је медијана резидуала једнака 0, онда негативни резидуали имају ранг који је мањи од $\frac{1}{2}(n+1)$ а позитивни резидуали имају ранг који је већи од $\frac{1}{2}(n+1)$. Одавде следи да је $rang(\hat{e}_i) - \frac{1}{2}(n+1)$ истог знака као и $\hat{e}_i$, па $\left|rang(\hat{e}_i) - \frac{1}{2}(n+1)\right| |\hat{e}_i| = \left[rang(\hat{e}_i) - \frac{1}{2}(n+1)\right]\hat{e}_i$. На пример, претпоставимо да су резидуали $-23, -18, -11, 0, 16, 19, 29$. Вредности израза $rang(\hat{e}_i) - \frac{1}{2}(n+1)$ за ове резидуале су $-3, -2, -1, 0, 1, 2, 3$. Приметимо да је $|-3||-23| = (-3)(-23)$, … , $|0||0| = (0)(0)$, … , $|3||29| = (3)(29)$. Због тога, ако је расподела грешака приближно симетрична, односно ако се очекује да је медијана резидуала једнака 0, онда би пондерисана сума $\sum w_i|\hat{e}_i|$ са пондерима $w_i = \left|rang(\hat{e}_i) - \frac{1}{2}(n+1)\right|$ требала да буде једнака суми (5.2).

5.2 Тестирање хипотезе $\boldsymbol{\beta = 0}$

Да бисмо тестирали хипотезу $\beta = 0$ користићемо тест статистику

$$|t| = \frac{|U|}{SD(U)}$$

где је

$$U = \sum_{i=1}^{n}\left[rang(y_i) - \frac{n+1}{2}\right]x_i$$

и

$$SD(U) = \sqrt{\frac{n(n+1)}{12}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

p-вредност се рачуна као $P[|Z| \geq |t|]$ где је Z случајна променљива која има стандардну нормалну расподелу.

Тест статистике које се користе у методи најмањих квадрата и у методи најмањих апсолутних одступања имају облик $|t| = \frac{|\widehat{\beta}|}{est.SD(\widehat{\beta})}$ и засноване су на чињеници да, ако је $\beta = 0$, онда је врло

вероватно да је и $\hat{\beta}$ близу нуле. Ова чињеница важи за све разумне оцене параметра β, укључујући и непараметарску оцену описану у овом поглављу, али код непараметарске оцене је тешко добити добру оцену за $SD(\hat{\beta})$. Тест статистика (6.5) је заснована на чињеници да, када је $\beta = 0$, онда је врло вероватно да је U близу нуле. Због тога, ако је $|t|$ велико, односно ако је U далеко од нуле у односу на величину $SD(U)$, онда закључујемо да је $\beta \neq 0$.

Очекујемо да је U близу нуле када је $\beta = 0$, јер је тада очекивање од U једнако 0. Када је $\beta = 0$ имамо да је $y_i = \alpha + e_i$, па се може сматрати да су опсервације $y_1, y_2, \dots, y_n$ изабране независно једна од друге, из исте популације. Одавде следи да су очекивања од $rang(y_i)$ једнака, за свако i. Обзиром да се вредности $rang(y_i)$ морају додати вредностима $\frac{n(n+1)}{2}$, очекивање сваке вредности $rang(y_i)$ мора бити $\frac{1}{2}(n+1)$. Пошто $\left[rang(y_i) - \frac{n+1}{2}\right]$ има очекивање 0 за свако i, онда и U има очекивање 0.

Ако је $\beta = 0$, онда би непараметарска оцена $\hat{\beta}$ требала да буде близу 0, што значи да би минимум функције (6.3) требао да буде близу 0. Близу минимума функције њен нагиб је близу 0. Због тога би, ако је $\beta = 0$, нагиб функције (6.3) требао да буде близу 0 за $b = 0$. Приметимо да је – U нагиб функције (6.3) у интервалу који садржи тачку $b = 0$. Ово је још један разлог зашто је очекивање од U близу 0, када је $\beta = 0$.

p-вредност

Када је обим узорка n велики, познато је да $t = \frac{U}{SD(U)}$ има приближно стандардну нормалну расподелу, ако је нулта хипотеза $\beta = 0$ тачна. Дакле, што је обим узорка већи, то смо сигурнији у тачност израчунате p-вредности.

Сличност са тестом методе најмањих квадрата

Присетимо се да се у тесту методе најмањих квадрата користи $t_{MNK} = \frac{\hat{\beta}_{MNK}}{ocena.SD(\hat{\beta}_{MNK})}$ где је $\hat{\beta}_{MNK} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$ и $ocena.SD(\hat{\beta}_{MNK}) = \hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Тада је

$$t_{MNK} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

У непараметарски тесту се користи $t_{NP} = U/SD(U)$. Нека r_i представља $rang(y_i)$. Средња вредност ранга је $\bar{r} = \frac{1}{2}(n+1)$, па је $U = \sum_{i=1}^{n}(r_i - \bar{r})x_i$. Користећи чињеницу да је $\sum_{i=1}^{n}(r_i - \bar{r}) = 0$ можемо записати $U = \sum_{i=1}^{n}(x_i - \bar{x})(r_i - \bar{r})$ и одатле имамо

$$t_{NP} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(r_i - \bar{r})}{\sqrt{\frac{n(n+1)}{12}\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Ова формула је слична као и формула за t_{MNK}, само уместо y_i имамо r_i и уместо $\hat{\sigma}$ имамо $\sqrt{\frac{n(n+1)}{12}}$. $\hat{\sigma}$ је оцена стандардног одступања свих вредности y_i, а $\sqrt{\frac{n(n+1)}{12}}$ је једнако узорачком стандардном одступању вредности r_i.

На овај начин добијамо још једну могућност да изразимо t_{NP} као производ вредности $\sqrt{n-1}$ и корелационог коефицијента између вредности x_i и r_i.

Пример 5.1

На истом примеру ћемо демонстрирати коришћење непараметарске регресије у R-у. У табели 2.1 су нам дати подаци о наталитету и природном прираштају.

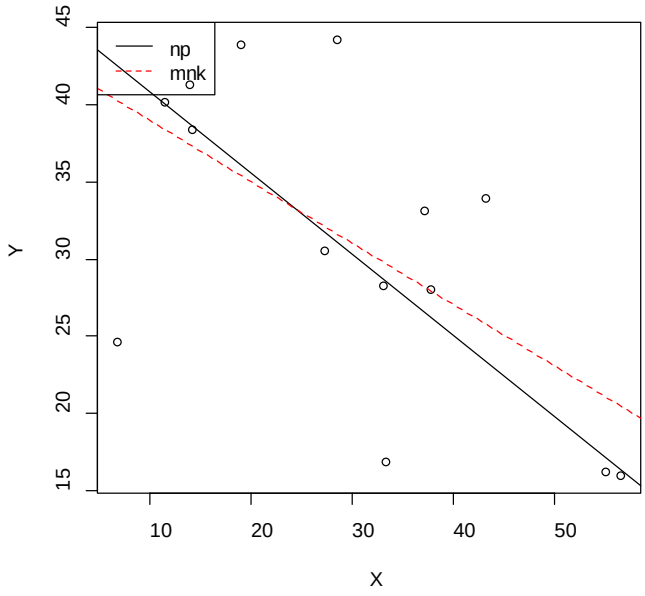
Прво морамо да инсталирамо пакет „Rfit“. Главна функција овог пакета је функција “rfit“.

```
> X<-c(55.0,27.3,33.3,37.1,11.5,14.2,13.9,19.0,33.1,43.2,28.5,37.7,6.8,56.5)

> Y<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9,28.3,33.9,44.2,28.0,24.6,16.0)

> library(Rfit)

> nr<-rfit(Y~X)

> summary(nr)

Call:

rfit.default(formula = Y ~ X)


Coefficients:

            Estimate Std. Error t.value   p.value

(Intercept) 46.05448    4.63626  9.9336 3.849e-07 ***

X           -0.52564    0.12882 -4.0803  0.001525 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Multiple R-squared (Robust): 0.443325

Reduction in Dispersion Test: 9.55656 p-value: 0.00934
```

Видимо да су нам оцењени коефицијенти $\hat{\alpha} = 46.05448$ и $\hat{\beta} = -0.52564$.

Сада ћемо да нацртамо график оцењене регресионе праве добијен непараметарском регресијом и график регресионе праве добијен методом најмањих квадрата.

```
plot(Y~X)

> abline(nr)

> abline(lm(Y~X), col=2, lty=2)

> legend("topleft",legend=c("np","mnk"),col=1:2,lty=1:2)
```



Слика 5.1 Оцењене регресионе праве за узорак „Наталитет“,добијене методом најмањих квадрата и непараметарском регресијом.

Пример 5.2

Имамо базу података, укључену у програмски пакет R, која садржи узорак из Канадског пописа становништва, из 1971.године, о просечној заради мушкараца истог образовања. У бази је дато 205 опсервација за две променљиве: логаритам зараде сваког мушкарца(logwage) и његов број година(age).

Кренућемо са параметарским моделом добијеним методом најмањих квадрата, како бисмо упоредили ове две методе.

Прво учитавамо пакет „np“ и базу података „cps71“.

```
> library("np")
> data("cps71")
> model.par <- lm(logwage ~ age + I(age^2), data = cps71)
> summary(model.par)

Call:
lm(formula = logwage ~ age + I(age^2), data = cps71)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4041 -0.1711  0.0884  0.3182  1.3940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.0419773  0.4559986  22.022  < 2e-16 ***
age          0.1731310  0.0238317   7.265 7.96e-12 ***
I(age^2)    -0.0019771  0.0002898  -6.822 1.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5608 on 202 degrees of freedom
Multiple R-squared:  0.2308,    Adjusted R-squared:  0.2232
F-statistic:  30.3 on 2 and 202 DF,  p-value: 3.103e-12
```

Сада ћемо наћи модел непараметарском регресијом.

```
>  model.np <- npreg(logwage ~ age,
```

```
+ regtype = "ll",
+  bwmethod = "cv.aic",
+ gradients = TRUE,
+  data = cps71)
> summary(model.np)

Regression Data: 205 training points, in 1 variable(s)
                age
Bandwidth(s): 2.805308

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 0.5215268
R-squared: 0.3251639

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Упоређујући вредности R^2 ова два модела, можемо закључити да је други модел бољи, јер је његова R^2 вредност 0.3251639.

Даље ћемо испитати да ли је $\beta = 0$, користећи тест значајности $npsigtest$.

```
> npsigtest(model.np)

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
age (1)
```

age

Bandwidth(s): 2.805308

Individual Significance Tests

P Value:

*age < 2.22e-16 ****

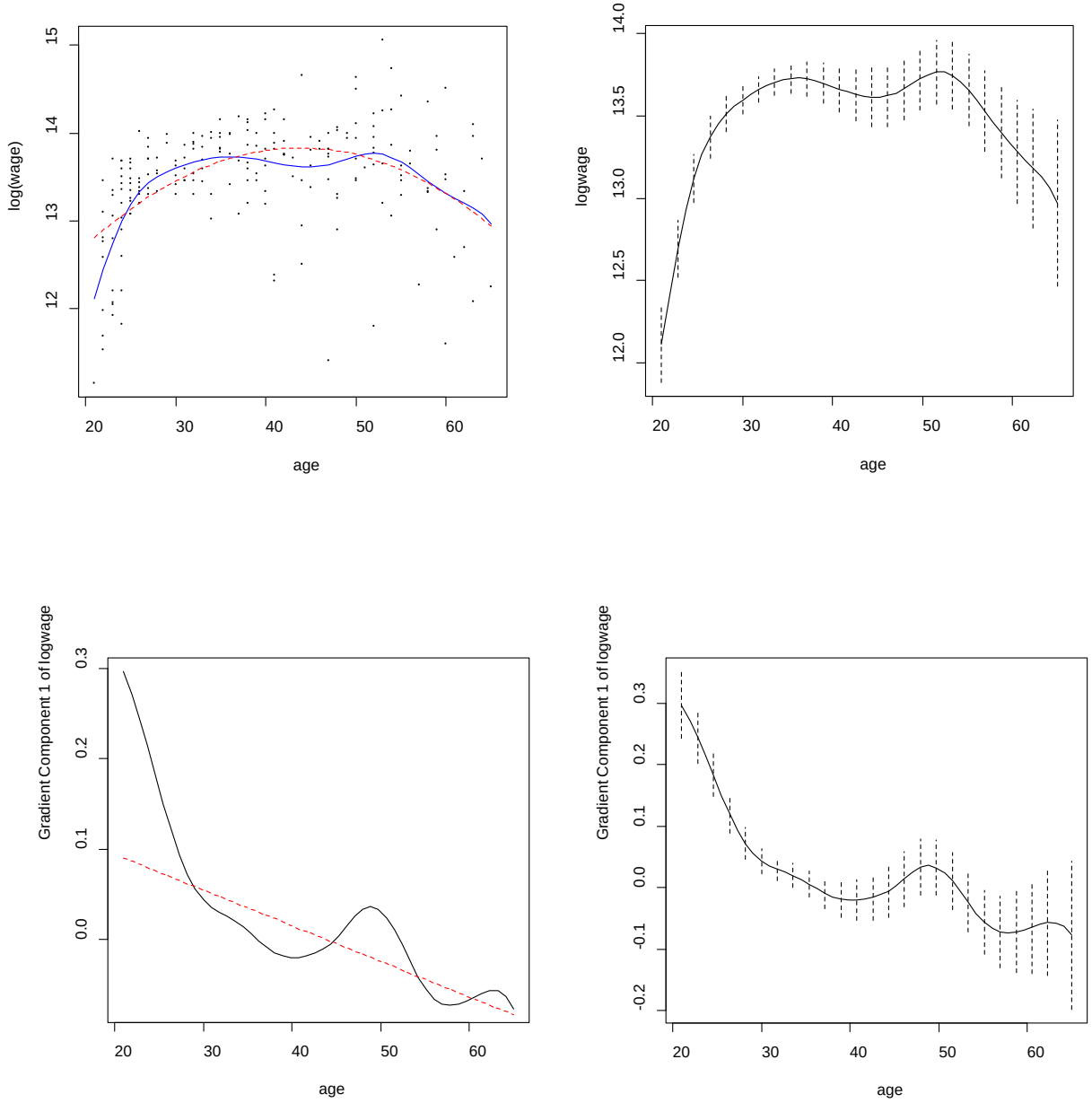
---

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

Одавде видимо да је променљива „age“ значајна, односно да одбацујемо хипотезу $\beta = 0$.

Даље вршимо графичко упоређивање параметарског и непараметарског модела.

```
> plot(cps71$age, cps71$logwage, xlab = "age", ylab = "log(wage)", cex=.1)
> lines(cps71$age, fitted(model.np), lty = 1, col = "blue")
> lines(cps71$age, fitted(model.par), lty = 2, col = " red")
> plot(model.np, plot.errors.method = "asymptotic")
> plot(model.np, gradients = TRUE)
> lines(cps71$age, coef(model.par)[2]+2*cps71$age*coef(model.par)[3],
+  lty = 2,
+ col = "red")
> plot(model.np, gradients = TRUE, plot.errors.method = "asymptotic")
```



Слика 5.2. На слици горе лево представљене су параметарска и непараметарска оцена регресионе праве за базу података cps71. На слици доле лево представљене су параметарска и непараметарска оцена нагиба. На сликама десно су представљене непараметарске оцене регресионе праве и нагиба заједно са њиховим границама варијабилности.

5.3 Вишеструка регресија

Претпоставимо да нашим подацима одговара вишеструки линеарни регресиони модел

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e.$$

Непараметарске оцене $\hat{\beta}_1, \ldots, \hat{\beta}_k$ су вредности $b_1, \ldots, b_k$ које минимизирају суму

$$\sum \left[rang\left(y_i - (b_1 x_{i1} + \cdots + b_k x_{ik})\right) - \frac{n+1}{2} \right] \times (y_i - (b_1 x_{i1} + \cdots + b_k x_{ik})) \tag{5.3}$$

Непараметарска оцена $\hat{\beta}_0$ се добија као медијана разлика $y_i - (\hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})$.

Алгоритам за минимизирање суме (5.3) дат је у књизи Alternative methods of regression, на страни 123.

5.3.1 Тестирање хипотезе $\boldsymbol{\beta_{q+1} = \cdots = \beta_k = 0}$

Непараметарски тест је аналоган тесту методе најмањих квадрата.

Означимо са SPR суму пондерисаних резидуала

$$\sum \left[rang(\hat{e}_i - \frac{1}{2}(n+1)) \right] \hat{e}_i.$$

Непараметарска тест статистика је

$$F_{NP} = \frac{SPR_{uprošćen} - SPR_{neuprošćen}}{(k-q)c\hat{\tau}}$$

где је $c = (n+1)/\sqrt{48}$ а $\hat{\tau}$ је дато формулом (2.6).

Како бисмо израчунали $\hat{\tau}$, узмимо резидуале из неупрошћеног модела и формирајмо средње вредности $A_{ij} = (\hat{e}_i + \hat{e}_j)/2$ за $1 \le i \le j \le n$. Затим поређајмо ових $N = n(n+1)/2$ бројева у растући поредак: $A_{(1)} \le A_{(2)} \le \cdots \le A_{(N)}$. Нека је $a = n(n+1)/4$, $b = \sqrt{n(n+1)(2n+1)/24}$, r_1 =најближи цео број броју $\frac{1}{2} + a - (1.645)b$, r_2 =најближи цео број броју $\frac{1}{2} + a + (1.645)b$ а $f = \sqrt{n/[n-(k+1)]}$. Важи

$$\tau = f \frac{\sqrt{n}\left[A_{r_2} - A_{r_1}\right]}{2(1.645)}.$$

Као и код теста у методи најмањих квадрата, приближна p-вредност се рачуна као вероватноћа

$$P[F \ge F_{NP}]$$

где је F случајна променљива која има F расподелу са $k-q$ и $n-k-1$ степени слободе.

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] W.H.Chang, J.W.McKean, J.D.Naranjo, and S.J.Sheather. *High-breakdown rank regression.*

[3] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generaliyations. Least squares and alternatives.* 1995.

[4] K.S.Crimin, A.Abebe, and J.W.McKean*. Robust general linear models and graphics via a user interface.*

6. Бајесова регресија

Бајесов приступ статистичкој анализи је другачији од уобичајеног класичног приступа. Код класичног приступа, једини извор информација представља узорак (база података). Код Бајесовог приступа, за добијање оцене или тестирање параметара се, поред узорка, користе и информације добијене из ранијих искустава.

Главну формулу, која се користи за инкорпорирање ранијих знања у статистичку анализу, формулисао је Томас Бајес[1] (Thomas Bayes) око 1760.године. Ову методу је на регресиону анализу први применио Харолд Џефриз[2] (Sir Harold Jeffreys) 1939.године.

6.1 Бајесов приступ (Бајесова анализа)

Бајесов приступ се може применити на различите статистичке проблеме. Да бисмо анализирали одређене податке, прво што морамо урадити је проналажење одговарајућег модела за те податке. Нека $\boldsymbol{y}$ представља вектор који садржи одређене податке и нека $\boldsymbol{\theta}$ представља вектор непознатих параметара модела. (Код једноструке регресије $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, $\boldsymbol{\theta} = (\alpha, \beta, \sigma)$.) Треба да оценимо неке од параметара вектора $\boldsymbol{\theta}$, користећи вектор $\boldsymbol{y}$.

Вектор $\boldsymbol{y}$, посматран као случајна променљива, има одређену расподелу за сваку фиксирану вредност вектора параметара $\boldsymbol{\theta}$. (У једноструком линеарном моделу, за фиксиране вредности α, β и σ , y_i има нормалну расподелу са очекивањем $\alpha + \beta x_i$ и варијансом σ^2 и све вредности y_i су међусобно независне.) Расподела вектора $\boldsymbol{y}$ за фиксирану вредност $\boldsymbol{\theta}$ се назива условна расподела (расподела за $\boldsymbol{y}$ под условом $\boldsymbol{\theta}$).

Код Бајесовог приступа треба прецизирати и расподелу за $\boldsymbol{\theta}$. Пре него што узмемо у обзир податке из узорка, треба да размотримо шта знамо о вредностима параметара и да ово знање искористимо за проналажење расподеле за вектор $\boldsymbol{\theta}$. Расподела вектора $\boldsymbol{\theta}$ се назива априорна расподела. Следећи корак је проналажење расподеле за $\boldsymbol{\theta}$ под условом $\boldsymbol{y}$, помоћу априорне расподеле параметра $\boldsymbol{\theta}$ и расподеле за $\boldsymbol{y}$ под условом $\boldsymbol{\theta}$.

Бајесова формула изгледа овако:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = cf(\boldsymbol{\theta}) \cdot f(\boldsymbol{y}|\boldsymbol{\theta})$$

где је

$f(\boldsymbol{\theta}|\boldsymbol{y})$-расподела вектора $\boldsymbol{\theta}$ под условом $\boldsymbol{y}$,

$f(\boldsymbol{\theta})$-априорна расподела вектора $\boldsymbol{\theta}$,

---

(1) Томас Бајес (1701-1761) био је енглески статистичар и филозоф који је познат по теореми која је добила његово име.

(2) Харолд Џефриз (1891-1989) био је енглески математичар, статистичар, геофизичар и астроном. Његова књига „Теорија вероватноће“ је дала значајан допринос откривању Бајесовог погледа на вероватноћу.

$f(\boldsymbol{y}|\boldsymbol{\theta})$-расподела вектора $\boldsymbol{y}$ под условом $\boldsymbol{\theta}$,

c-израз који не зависи од $\boldsymbol{\theta}$.

6.2 Оцењивање регресионе праве

Прво морамо одредити расподелу вектора података. Најчешће коришћена расподела је нормална расподела. Претпоставимо да нашим подацима одговара линеарни регресиони модел. Тада, уколико су нам дате вредности α, β и σ, вектор $\boldsymbol{y} = (y_1, \dots, y_n)$ има мултиваријантну нормалну расподелу (Случајни вектор има k-варијантну расподелу ако свака линеарна комбинација његових k компоненти има нормалну расподелу).

Бајесов метод подразумева да одредимо заједничку априорну расподелу параметара α, β и σ. Теоретски, свака расподела може бити изабрана као априорна расподела, али је у пракси боље изабрати расподелу која се лакше комбинује са расподелом вектора $\boldsymbol{y} = (y_1, \dots, y_n)$**.** У даљем тексту ће бити представљене две априорне расподеле које могу да се користе.

6.2.1 Коришћење неинформативних априорних информација

Априорне информације и сазнања, која особа има о параметрима, зависе искључиво од те особе. Претпоставимо да немамо никакве информације и сазнања о параметрима. Такође, претпоставимо да немамо идеју о томе које вредности могу одговарати параметрима α, β и σ, осим што знамо да σ мора бити позитивно.

Када немамо никакву информацију о параметрима најбоље је, за густину расподеле случајног вектора (α, β и σ), узети функцију

$$f(\alpha, \beta, \sigma) = \frac{1}{\sigma}. \tag{6.1}$$

Ово значи да, пре него што погледамо у узорак, ми знамо да су све вредности за α, β и $log\sigma$ једнако вероватне. Међутим, ова функција не може бити густина расподеле, јер њен интеграл за $\alpha, \beta \in \mathbb{R}$ и $\sigma > 0$ није једнак 1 већ ∞. Дакле, израз (6.1) није добар избор за расподелу случајних параметара.

Бајесове оцене

Бајесове оцене параметара α и β су узете тако да буду очекивања за α и β при услову да $\boldsymbol{\theta}$ има апостериорну расподелу. Ове оцене ће бити исте као и оцене добијене методом најмањих квадрата. Доказ ове чињенице се може наћи у књизи „Alternative methods of regression“, у делу 7.7, на страни 154.

6.2.2 Коришћење конјугованих априорних расподела

Претпоставимо да имамо априорне информације о параметрима. Конјугована априорна расподела је она расподела, која у комбинацији са расподелом вектора узорка даје апостериорну расподелу, која има исти облик као и априорна расподела. Када узорак има мултиваријантну

нормалну расподелу, ако допустимо да условна расподела параметара (α, β) за све вредности σ, буде биваријантна нормална расподела и допустимо да расподела параметра $\frac{1}{\sigma^2}$ буде Гама расподела, онда то представља коњуговану априорну расподелу за вектор (α, β, σ). Користећи конјуговану априорну расподелу моћи ћемо да добијемо формуле за израчунавање очекивања за α и β у апостериорној расподели.

Квантификовање априорне информације

Одлука о томе која ће се априорна расподела, из фамилије конјугованих расподела, узети, се не доноси на основу априорне информације, већ се узима најпогоднија априорна расподела. Априорну информацију користимо тек пошто изаберемо конјуговану расподелу. На пример, ми можемо квантификовати априорну информацију тако што ћемо наћи очекивања, стандардна одступања и корелацију параметара α и β , при датом σ, и очекивање и стандардно одступање величине $\frac{1}{\sigma^2}$. Уколико само желимо да израчунамо оцене за α и β, као у овом поглављу, није потребно наћи очекивање и стандардно одступање величине $\frac{1}{\sigma^2}$. Уместо да преводимо априорну информацију у расподелу за α и β, било би лакше да ту информацију преведемо у расподелу за μ и β, где је $\mu = \alpha + \beta x_m$ а x_m је „средња“ x вредност.

Бајесове оцене.

Нека је $\mu = \alpha + \beta\bar{x}$, односно нека је x_m средња вредност свих вредности x у узорку. Претпоставимо да смо, засновано на претходној информацији, одредили априорну расподелу код које су, при датој вредности за σ, очекивања за μ и β редом e_μ и e_β, а њихова стандардна одступања редом $c_\mu\sigma$ и $c_\beta\sigma$. Нека су априорне расподеле за μ и β независне. Бајесове оцене, дате на основу очекивања за α и β, при апостериорној расподели, су:

$$\hat{\alpha} = \hat{\mu} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \left\{\frac{c_\beta^{-2}}{c_\beta^{-2}+\sum(x_i-\bar{x})^2}\right\} e_\beta + \left\{\frac{\sum(x_i-\bar{x})^2}{c_\beta^{-2}+\sum(x_i-\bar{x})^2}\right\} \hat{\beta}_{MNK} \tag{6.2}$$

где је

$$\hat{\mu} = \left\{\frac{c_\mu^{-2}}{c_\mu^{-2}+n}\right\} e_\mu + \left\{\frac{n}{c_\mu^{-2}+n}\right\} \bar{y}$$

а $\hat{\beta}_{MNK}$ је оцена израчуната методом најмањих квадрата на истом узорку.

Ове формуле показују како се користи претходно знање на тренутном узорку. Бајесова оцена за β је пондрерисани просек од e_β, априорна очекивана вредност од β добијена на основу претходног знања, а $\hat{\beta}_{MNK}$ је оцена за β добијена методом најмањих квадрата на тренутном узорку. Бајесова оцена параметра μ је сличан пондерисани просек априорне очекиване вредности засноване на претходном знању и оцена добијена методом најмањих квадрата на тренутном узорку.

6.3 Тестирање хипотезе $\boldsymbol{\beta = 0}$

Апостериорне вероватноће и p-вредност

Обично приликом тестирања хипотезе израчунавамо p-вредност. Лако се може помислити да p-вредност представља вероватноћу да је нулта хипотеза тачна. p-вредност се може посматрати као вероватноћа, али вероватноћа једног другог догађаја. Претпоставимо да се један експеримент понавља више пута. Нека је D^* догађај да подаци из поновљеног експеримента дају тест статистику која је исте величине или чак већа од тест статистике коју дају подаци добијени из почетног експеримента. Нека је H нулта хипотеза. p-вредност је вероватноћа $P(D^*|H)$, односно вероватноћа да би тест статистика у поновљеном експерименту била једнако екстремна као и тренутна посматрана тест статистика, под условом да је нулта хипотеза тачна. Апостериорна вероватноћа нулте хипотезе је вероватноћа $P(H|D)$, где је D догађај да су подаци из поновљеног експеримента једнаки као и тренутни посматрани подаци. Апостериорна вероватноћа и p-вредност се могу представити као условне вероватноће, у којима фигуришу нулта хипотеза H и подаци D.

Априорна расподела

Ниједна од априорних расподела, које смо користили у претходном делу за добијање Бајесових оцена, није погодна за тестирање хипотезе $\beta = 0$. Ове расподеле шире вероватноћу непрекидно преко скупа параметара и морају доделити вероватноћу нула сваком подскупу који је мање димензије. Приметимо да скуп параметара $\{(\alpha, \beta, \sigma)$: где је α било који број, $\beta = 0, \sigma > 0\}$ који је описан помоћу нулте хипотезе има две димензије, односно да представља раван, а скуп параметара $\{(\alpha, \beta, \sigma)$: где су α и β било који бројеви а $\sigma > 0\}$ има три димензије. Дакле, непрекидна априорна расподела скупа параметара мора доделити априорну вероватноћу нула нултој хипотези, што имплицира да апостериорна расподела нулте хипотезе мора такође бити нула. Са таквом априорном расподелом ми увек можемо закључити да је $\beta \neq 0$, без обзира на податке. Дакле, треба нам априорна расподела која додељује позитивну вероватноћу нултој хипотези.

Овде узимамо у обзир само ситуације у којима није доступно никакво знање о параметрима. Формулисаћемо априорну расподелу која се чини као добар избор у оваквој ситуацији. Уместо уобичајеног модела $y_i = \alpha + \beta x_i + e_i$ узећемо модел

$$y_i = \mu + \beta(x_i - \bar{x}) + e_i$$

где је $\mu = \alpha + \beta\bar{x}$.

Разлог коришћења овог модела уместо уобичајеног модела је тај што је претпоставку да су μ и β независни лакше оправдати него претпоставку да су α и β независни. Ова независност поједностављује презентацију априорне расподеле.

Нулта и алтернативна хипотеза одговарају, редом, скуповима
$H_0 = \{(\mu, \beta, \sigma)$: где је μ било који број, $\beta = 0,\ \sigma > 0\}$ и
$H_a = \{(\mu, \beta, \sigma)$: где је μ било који број, $\beta \neq 0, \sigma > 0\}$.

Можемо да користимо следећу априорну расподелу:

$$P(H_0) = \frac{1}{2}, \qquad P(H_a) = \frac{1}{2} \tag{6.3а}$$

$$f(\mu, \sigma | H_0) = \frac{1}{\sigma} \tag{6.3б}$$

$$f(\mu, \sigma | H_a) = \frac{1}{\sigma} \tag{6.3ц}$$

$$f(\beta | \mu, \sigma, H_a) = ce^{-\beta^2/(2\vartheta)} \tag{6.3д}$$

где је $c = 1/(\sqrt{2\pi\vartheta})$ а $\vartheta = n\sigma^2 / \sum(x_i - \bar{x})^2$.

У одсуству априорне информације било би погодно доделити једнаке априорне вероватноће нултој и алтернативној хипотези, као у (6.3а). Густина расподеле (6.3б) је слична густини расподеле (6.1).

Било би разумно изабрати неинформативну расподелу и за скуп H_a. Чињеница да H_a, поред μ и σ, укључује и β, компликује процес. Сетимо се да је (6.1) неодговарајућа густина расподеле јер њен интеграл није 1 већ ∞. Било би еквивалентно узети $\frac{c}{\sigma}$ за густину расподеле, где је c било која позитивна константа. Ако за оба сета H_0 и H_a изаберемо неинформативне расподеле сличне (6.1), њихове густине расподела ће бити, респективно, $f(\mu, \sigma | H_0) = c_0/\sigma$ и $f(\mu, \beta, \sigma | H_a) = c_a/\sigma$, али, обзиром да су то расподеле различитих сетова (чак имају и различите димензије), не постоји разлог да изаберемо да је c_0 једнако са c_a. Нажалост апостериорна вероватноћа нулте хипотезе ће много зависити од релативних димензија величина c_0 и c_a, а не постоји неинформативни начин да одаберемо ове величине.

Дакле, морамо да пронађемо други начин за проналажење неинформативне расподеле за H_a. Ако за тренутак занемаримо постојање параметра β, било би разумно изабрати исте неинформативне расподеле за μ и σ при претпоставци да је тачна нулта хипотеза као и при претпоставци да је тачна алтернативна хипотеза. Ово је урађено у (6.3ц). Остало нам је да продискутујемо (6.3д).

Густина расподеле у (6.3д) је густина нормалне расподеле са очекивањем 0 и дисперзијом ϑ. Изабрали смо нормалну расподелу јер је најпогоднија. Такође у сличном али једноставнијем тестирању смо закључили да облик расподеле није пресудан. Што се тиче очекивања, имамо да је 0 претпостављена вредност за β, па је разумно узети алтернативне вредности које се налази у близини централне вредности, која је 0.

Размотримо сада избор параметра ϑ. Можемо покушати да добијемо неинформативну расподелу тако што ћемо узети јако велики број за ϑ, па ће се вероватноћа простирати у широком спектру и

густина расподеле ће бити скоро константна. Међутим, ако пустимо да ϑ тежи ка ∞, апостериорна расподела нулте хипотезе ће бити 1, независно од података које посматрамо. Поново смо наишли на проблем приликом формулисања неинформативне априорне расподеле.

Разлог овог проблема је разлика између ситуације коју имамо приликом оцењивања и ситуације коју имамо код тестирања. Када оцењујемо параметар β, понекад можемо бити у ситуацији да немамо баш никакву информацију о вредности тог параметра, али када тестирамо хипотезу $\beta = 0$, онда нам чињеница да имамо претпостављену једну одређену вредност 0, говори да имамо неко знање о вредности параметра β. Како бисмо објаснили шта значи веома мала количина априорне информације, претпоставимо да је наше знање еквивалентно количини информације у једној опсервацији. Тиме што су нам познате информације о једној опсервацији, не угрожавамо валидност тестирања хипотезе $\beta = 0$, јер нам информација о једној опсервацији не говори ништа о томе да ли је $\beta = 0$ или није. Удаљеност једне опсервације од нуле се једнако добро може приписати случајној грешци као и вредности параметра β која није једнака 0, јер једна опсервација не даје никакву информацију о томе колика је случајна грешка.

Коју вредност треба да узмемо за варијансу параметра β са априорном расподелом, такву да она одговара познатој опсервацији?

Можемо да користимо концепт Фишерове информације. Ако претпоставимо да узорак $y_1, \dots, y_n$ има нормалну расподелу и да је дата вредност параметра σ, онда је количина Фишерове информације о параметру β у узорку једнака прецизности оцене параметра β добијене методом најмањих квадрата, односно оцене $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$. Што прецизније оценимо параметар β, то имамо више информације о њему. Можемо рећи да је количина Фишерове информације о параметру β, садржана у једној опсервацији, једнака n-том делу информације садржане у целом узорку, односно $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n\sigma^2}$. Изједначавајући информацију са прецизношћу, ми узимамо да то буде прецизност априорне расподеле параметра β, под условом σ, када претпостављамо да је алтернативна хипотеза тачна.

-Опис теста.

Бајесов тест хипотезе $\beta = 0$ можемо извршити тако што ћемо израчунати апостериорну вероватноћу нулте хипотезе, засновану на априорној расподели, описаној у (6.3). Апостериорна вероватноћа може бити изражена као узорачка корелација између x и y

$$P(H_0|y) = \frac{1}{1 + \frac{1}{\sqrt{g}}} \tag{6.4}$$

где је

$$g = (n+1)\left[1 - \left(\frac{n}{n+1}\right) r^2\right]^{n-1}$$

а r је коефицијент корелације

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}.$$

Апостериорна вероватноћа такође може бити изражена помоћу тест статистике t_{MNK} методе најмањих квадрата, јер је

$$r^2=\frac{1}{1+\frac{n-2}{t^2}}.$$

Може се видети да је формула за апостериорну вероватноћу оправдана помоћу њене зависности од узорачког коефицијента корелације r. Видимо да је $P(H_0|y)$ опадајућа функција по $|r|$. Ово има смисла, јер што је веће $|r|$ то је корелација између x и y већа, па је и веза између њих јача, стога има више доказа против хипотезе $\beta=0$ а такође закључујемо да би апостериорна вероватноћа хипотезе H_0 требала да буде мања.

Размотримо случај када је $r=0$. Тада следи да не постоји зависност између x и y, што одговара случају $\beta=0$, и апостериорна вероватноћа хипотезе H_0 би требала да буде приближно 1. Заиста, када је $r=0$, важи $P(H_0|y)=1-(1+\sqrt{n+1})$. За $n\geq 5$ то је веће од 0.70. Када је $r=\pm 1$ тада је зависност између x и y велика, што одговара случају $\beta\neq 0$, и апостериорна вероватноћа хипотезе H_0 би требало да буде приближно 0. Када је $r=\pm 1$, онда је $P(H_0|y)=1/(1+(\sqrt{n-1})^{n-2})$. За $n\geq 5$ то је мање од 0.07. Извођење формуле (6.4) захтева употребу компликованих интеграла.

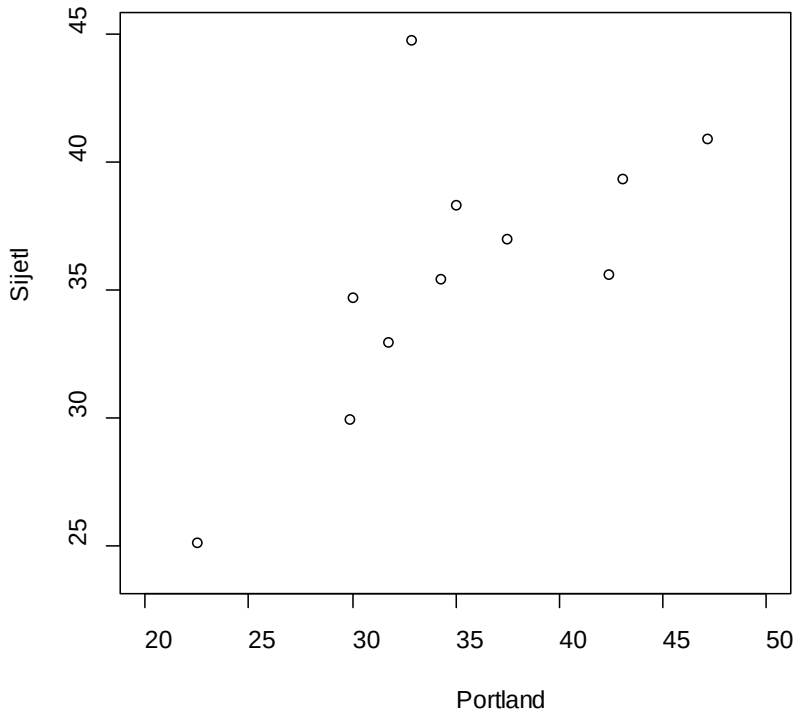
Пример 6.1

Због временских образаца на северозападу Сједињених Америчких Држава количина падавина у Сијетлу је условљена количином падавина у Портланду. У табели 6.1 су дате годишње количине падавина у овим градовима, од 1980. до 1990. године. На слици 6.1 представљено је 11 тачака чија x координата представља количину падавина у Портланду, а y координата представља количину падавина у Сијетлу за одређену годину. На основу положаја тачака на графику, можемо да закључимо да ће линеарни модел $Y=\alpha+\beta X+e$ одговарати датим подацима. Ми желимо да изразимо годишњу количину падавина у Сијетлу као линеарну функцију од годишње количине падавина у Портланду.

Табела 6.1 Узорак „Количина падавина“.

Година	**Количина падавина У Сијетлу (Y)**	**Количина падавина У Портланду (X)**
1980	35.60	42.41
1981	35.40	34.29
1982	39.32	43.04
1983	40.93	47.19
1984	36.99	37.50
1985	25.13	22.48

1986	38.34	35.04
1987	29.93	29.91
1988	32.98	31.72
1989	34.69	30.05
1990	44.75	32.86



Слика 6.1 График годишњих падавина у Портланду и Сијетлу.

Уколико немамо никакво предзнање о годишњим количинама падавина у ова два града, модел који тражимо биће исти као модел добијен методом најмањих квадрата. Примењујући методу најмањих квадрата на узорак „количина падавина“ из табеле, добијамо регресиону праву $\hat{Y} = 18.03 + 0.5063X$.

Претпоставимо да имамо резултате регресионе анализе података о годишњим количинама падавина у Сијетлу и Портланду од 1950. до 1979. године. Означићемо ове резултате индексом „0“. Бројеви који су нам потребни за априорну расподелу су $\hat{\alpha}_0 = 5.513$, $\widehat{\beta}_0 = 0.8961$, $SD(\hat{\alpha}_0) = 1.140\sigma$, $SD(\hat{\beta}_0) = 0.02986\sigma$ и корелација $Corr(\hat{\alpha}_0, \hat{\beta}_0) = -0.9871$. (Ови бројеви не одређују у потпуности априорну расподелу али су они довољни за оцену параметара α и β.) Нека је $\mu = \alpha + 35.14\beta$, где је 35.14 средња вредност количине падавина у Портланду за тренутни узорак. Даље је $\hat{\mu}_0 = 5.513 + 35.14 \cdot 0.8961 = 37.00$ и $SD(\hat{\mu}_0) = 0.1977\sigma$. Број 0.1977 је

квадратни корен од $1.140^2 + 35.14^2 \cdot 0.02986^2 + 2 \cdot 25.14 \cdot 1.140 \cdot 0.02986 \cdot (-0.9871)$. Да бисмо одредили априорну расподелу параметара рећи ћемо да μ има очекивање 37.00 и стандардно одступање 0.1977σ (зависи од σ) и рећи ћемо да β има очекивање 0.8961 и стандардно одступање 0.02986σ. Претпоставимо да су априорне расподеле параметара μ и β независне. Приметимо да оцене $\hat{\mu}_0$ и $\hat{\beta}_0$ нису у потпуности независне јер аритметичка средина x-вредности из претходног узорка није тачно 35.14.

Даље ћемо из тренутног узорка израчунати $\bar{y} = 35.82$, $\hat{\beta}_{MNK} = 0.5063$, и $\sum(x_i - \bar{x})^2 = 497.2$. Сада можемо користити формуле (6.2) како бисмо добили Бајесове оцене за α и β.

$$\frac{0.02986^{-2}}{0.02986^{-2} + 497.2} = 0.6929$$

$$\hat{\beta} = 0.6929 \cdot 0.8961 + (1 - 0.6929) \cdot 0.5063 = 0.7764$$

$$\frac{0.1977^{-2}}{0.1977^{-2} + 11} = 0.6993$$

$$\hat{\mu} = 0.6993 \cdot 37.00 + (1 - 0.6993) \cdot 35.82 = 36.65$$

$$\hat{\alpha} = 36.65 - 0.7764 \cdot 35.14 = 9.363$$

Закључујемо да је Бајесова оцена регресионе праве $\hat{Y} = 9.363 + 0.7764X$.

Тестирање хипотезе $\beta = 0$

Одредимо сада да ли постоји значајна зависност између годишње количине падавина у Портланду и у Сијетлу. Коефицијент корелације је $r = 0.6695$, па је $g = 12 \cdot \left[1 - \frac{11}{12} \cdot 0.6695^2\right]^{10} = 0.06044$ а апостериорна вероватноћа нулте хипотезе је $\frac{1}{1+\frac{1}{\sqrt{0.06044}}} = 0.1973$. Према томе, ако узмемо да је априорна расподела дефинисана као у (6.3), односно ако претпоставимо да нулта хипотеза има 50% шансе да буде тачна, а затим ако искомбинујемо ту вероватноћу са нашим подацима, добијамо да нулта хипотеза има 20% шансе да буде тачна.

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Broeming, L.D. *Bayesian Analysis of Linear Models*. Marcel Dekker, New Zork, 1985.

[3] Ronald Christensen. Department of mathematics and statistics. University of New Mexico. *Plane answers to complex questions: The theory of linear models.2001.*

7. Риџ регресија

Риџ регресија (Ridge regression) је уведена 1962. године од стране Артура Хоерла(Arthur Hoerl) у једном чланку у часопису о хемијском инжењерингу. У случајевима када постоји корелација између објашњавајућих променљивих (мултиколинеарност) метода најмањих квадрата даје оцене које су непристрасне али су њихове варијансе велике па оне могу доста да се разликују од њихове праве вредности. Мултиколинеарност може да изазове неприродно повећање стандардних грешака регресионих коефицијената, добијање нетачне p-вредности теста и да смањи предвидљивост модела. Артур Хоерл је осмислио метод који ће давати боље резултате у оваквим случајевима. Уколико се ради о једнострукој регресији, обзиром да ту имамо само једну објашњавајућу променљиву, проблем не постоји. Риџ регресија се примењује искључиво на вишеструку регресију.

Постоји неколико начина за утврђивање мултиколинеарности. Неки од њих су:

1. Испитивање корелационе матрице коефицијената може открити велику корелацију међу паровима променљивих.
2. За сваку променљиву треба израчунати R_i^2. Ако је вредност R_i^2 близу јединице то нам указује на постојање проблема.
3. Налажење сопствених вредности матрице $X'X$, где је λ_1 највећа сопствена вредност. Релативно мала сопствена вредност указује на проблем. Број $k = \sqrt{\frac{\lambda_1}{\lambda_p}}$ ћемо назвати условни број. Сматра се да је $k \geq 30$ велико. И остале условне бројеве $\sqrt{\frac{\lambda_1}{\lambda_i}}$ треба размотрити, јер они указују на то да постоји више од једне линеарне комбинације.

7.1 Оцењивање регресионе праве

Стандардизација

Први корак у Риџ регресији је стандардизација објашњавајуће променљиве. Како бисмо стандардизовали променљиву X, узећемо да је $z_i = (x_i - \bar{x})/s_x$ где је $\bar{x}$ средња вредност бројева x_i а s_x стандардно одступање величине x_i

Модел са нестандардизованим променљивим ћемо означити са $Y = \alpha + \beta X + e$, док ћемо модел са стандардизованим променљивим означити са $Y = \mu + \gamma Z + e$.

Риџ оцене

Оцене параметара μ и γ се методом најмањих квадрата добијају помоћу формула (2.1). Пошто су вредности z_i стандардизоване, имамо да је $\bar{z} = 0$, па формуле постају

$$\hat{\gamma}_{MNK} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i^2}$$

$$\hat{\mu}_{MNK} = \bar{y}$$

За Риџ оцене такође узимамо да је $\hat{\mu} = \bar{y}$ и

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i^2 + k} \tag{7.1}$$

где је $k = \hat{\sigma}^2_{MNK}/\hat{\gamma}^2_{MNK}$ а $\hat{\sigma}_{MNK}$ је оцена параметра σ добијена методом најмањих квадрата, помоћу формула (2.2).

Средњеквадратна грешка

Пре него што објаснимо како формула (7.1) даје тачније резултате него $\hat{\gamma}_{MNK}$, треба да прецизирамо шта подразумевамо под „тачније резултате“. $\hat{\gamma}$ је тачна оцена ако, посматрајући је као случајну променљиву, она има сличну вредност као γ, односно ако је разлика између $\hat{\gamma}$ и γ мала. Најлакши начин за мерење ове „тачности“ је да за оцену $\hat{\gamma}$ пронађемо њену средњеквадратну грешку $SKG(\hat{\gamma}) = E[(\hat{\gamma} - \gamma)^2]$. Што је мања средњеквадратна грешка то је тачност оцене већа.

Важи и

$$SKG(\hat{\gamma}) = Var(\hat{\gamma}) + [E(\hat{\gamma}) - \gamma]^2 \tag{7.2}$$

Уколико знамо да оцена има малу средњеквадратну грешку, можемо да закључимо и да она има малу варијансу и пристрасност.

Оцена $\hat{\gamma}_{MNK}$, добијена методом најмањих квадрата је непристрасна, односно важи $E(\hat{\gamma}_{MNK}) = \gamma$. Риџ оцена је пристрасна али има малу варијансу. Претпоставка је да повећање квадрата пристрасности није веће од смањења варијансе, односно да оцена добијена Риџ регресијом има малу средњеквадратну грешку.

Риџ регресију можемо да схватимо као покушај побољшања тачности оцене добијене методом најмањих квадрата, тако што ћемо је „смањити“. Како бисмо смањили оцену $\hat{\gamma}_{MNK}$, помножићемо је неким бројем c који се налази између 0 и 1, и на тај начин добијамо $\hat{\gamma} = c\hat{\gamma}_{MNK}$. Средњеквадратна грешка оцене $\hat{\gamma}_{MNK}$ је једнака њеној варијанси, $\vartheta = \sigma^2/\sum z_i^2$ (из формуле (2.3)), јер је њена пристрасност једнака 0. Средњеквадратна грешка вредности $\hat{\gamma} = c\hat{\gamma}_{MNK}$ је

$$SKG(c\hat{y}_{MNK}) = c^2\vartheta + (c-1)^2\gamma^2 \tag{7.3}$$

Уколико нађемо извод по променљивој c, изједначимо га са 0 и решимо ту једначину по c, наћи ћемо да се најмања средњеквадратна грешка добија за $c = \gamma^2/(\gamma^2 + \vartheta)$. Ако у овом изразу заменимо непознате параметре γ и σ њиховим оценама $\hat{\gamma}_{MNK}$ и $\hat{\sigma}_{MNK}$ добићемо оцену $\hat{c}$. Даље, можемо да нађемо да је $\hat{c} = \sum z_i^2/(\sum z_i^2 + \hat{\sigma}^2_{MNK}/\hat{\gamma}^2_{MNK})$ и да се $\hat{c}\hat{\gamma}_{MNK}$ поклапа са Риџ оценом (7.1).

Формула (7.3) важи само ако је c број који није случајан. Обзиром да смо ми изабрали $\hat{c}$ као случајан број (добијен је помоћу случајних бројева $\hat{\gamma}_{MNK}$ и $\hat{\sigma}_{MNK}$), биће тешко одредити средњеквадратну грешку оцене $\hat{\gamma} = \hat{c}\hat{\gamma}_{MNK}$ и она, у суштини, зависи од облика расподеле случајних грешака. Ипак, аргумент дат у претходном делу бар сугерише да је оцена $\hat{\gamma}$ тачнија од оцене $\hat{\gamma}_{MNK}$. Међутим, уочено је да је оцена $\hat{\gamma}$ тачнија од оцене $\hat{\gamma}_{MNK}$ само када је γ близу 0. За једноструку регресију се не препоручује коришћење Риџ регресије.

7.2 Вишеструка регресија

Стандардизација

Пре него што кренемо на Риџ оцењивање, морамо да стандардизујемо објашњавајуће променљиве. При израчунавању Риџ оцена све објашњавајуће променљиве се третирају на исти начин. Због тога, потребно је стандардизовати објашњавајуће променљиве тако да буде могуће међусобно поређење њихових јединица мерења.

Неки статистичари предлажу стандардизацију променљивих и приликом уппотребе методе најмањих квадрата.

Оцењивање регресионих коефицијената

Модел у којем нисмо стандардизовали променљиве ћемо представити као

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$$

док ћемо модел у којем смо стандардизовали променљиве представити као

$$y_i = \mu + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip} + e_i$$

У матричној нотацији модел ће бити

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z\gamma} + \boldsymbol{e}$$

где је

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix}, \qquad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

Оцене параметара μ и $\boldsymbol{\gamma}$ добијене методом најмањих квадрата су

$$\hat{\mu}_{MNK} = \bar{y}$$

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{MNK}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-\mathbf{1}}\boldsymbol{Z}'\boldsymbol{y}$$

За Риџ оцене параметара ћемо узети

$$\hat{\mu} = \bar{y}$$

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y} \tag{7.4}$$

где је $k = p\hat{\sigma}^2_{MNK}/\|\widehat{\boldsymbol{\gamma}}_{\boldsymbol{MNK}}\|^2$, $\hat{\sigma}^2_{MNK}$ је оцена параметра σ^2 добијена методом најмањих квадрата, $\boldsymbol{Z}$ је матрица димензије $n \times p$, а $\boldsymbol{I}$ је јединична матрица димензије $p \times p$. (За вектор $\boldsymbol{\vartheta}$, израз $\|\boldsymbol{\vartheta}\|$ означава дужину вектора, односно $\|\boldsymbol{\vartheta}\| = \sqrt{\sum \boldsymbol{\vartheta}_i^2} = \sqrt{\boldsymbol{\vartheta}'\boldsymbol{\vartheta}}$.)

Фактор повећања варијансе

Колинеарна веза која укључује више променљивих неће обавезно бити откривена помоћу корелације парова променљивих. Боља индикација за постојање корелације између променљивих је фактор повећања варијансе (FPV). Фактор повећања варијансе једне објашњавајуће променљиве X_j је мера колико је променљива X_j блиска осталим променљивим. Односно важи

$$FPV_j = \frac{1}{1 - R_j^2}$$

где је R_j^2 коефицијент помоћу кога доносимо одлуку о томе да ли је променљива X_j зависна од осталих променљивих. Ако променљива X_j може да се представи као линеарна комбинација осталих променљивих онда је $R_j^2 = 1$ и $FPV_j = \infty$. Ако је променљива X_j у потпуности некорелисана са осталим променљивим онда је $R_j^2 = 0$ и $FPV_j = 1$.

Постоји директна зависност између фактора повећања варијансе променљиве X_j и стандардног одступања оцене њеног регресионог коефицијента, добијене методом најмањих квадрата:

$$SD(\hat{\beta}_j) = \frac{\sqrt{FPV_j}}{\sqrt{n-1}}\left(\frac{\sigma}{s_j}\right)$$

где је $\hat{\beta}_j = \hat{\beta}_{MNKj}$ а s_j стандардно одступање вредности X_j. Величина вредности $SD(\hat{\beta}_j)$ зависи од три фактора: фактора $\sqrt{FPV_j}$, који зависи од зависности променљиве X_j са осталим променљивим, фактора $\frac{\sigma}{s_j}$, који зависи од варијације случајних грешака, које зависе од варијације мерења величине X_j, и фактора $\frac{1}{\sqrt{n-1}}$, који зависи од величине узорка.

Узрок нетачности оцене $\hat{\beta}_j$ може бити колинеарност променљиве X_j са осталим променљивим, или сувише велике случајне грешке, или јако мали обим узорка. Израз „фактор повећања

варијансе“ је настао због чињенице да је $Var(\hat{\beta}_j)$, FPV_j пута већа него што би била, ако би X_j била некорелисана са осталим објашњавајућим променљивим.

Матрична формулација проблема мултиколинеарности

У случају матрице објашњавајућих променљивих $\boldsymbol{Z}$, мултиколинеарност значи да нека од колона матрице $\boldsymbol{Z}$ представља приближно линеарну комбинацију осталих колона. Одавде следи да је матрица $\boldsymbol{Z}'\boldsymbol{Z}$, за коју морамо да нађемо инверзну матрицу како бисмо израчунали оцену параметра $\boldsymbol{\gamma}$ помоћу методе најмањих квадрата, приближно јединична матрица. Тражење инверзне матрице јединичној матрици је исто као и тражење реципрочне вредности броју 0. Тражење инверзне матрице матрици која је приближно јединична је исто као и тражење реципрочне вредности броја који је јако мали.

Варијанса оцене $\hat{\gamma}_{MNKj}$ добијена методом најмањих квадрата је једнака σ^2 пута j-та дијагонална вредност матрице $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Матрица $\boldsymbol{Z}'\boldsymbol{Z}$ је приближно јединична вероватно због постојања великих варијанси неких оцена добијених методом најмањих квадрата. Овде видимо како мултиколинеарност утиче на појаву нетачних регресионих оцена.

Риџ регресија трансформише матрицу $\boldsymbol{Z}'\boldsymbol{Z}$, тако да она не буде приближно јединична матрица. Матрица $\boldsymbol{Z}'\boldsymbol{Z}$ је модификована тако да она буде слична матрици која би се добила у случају да не постоји мултиколинеарност.

Матрица $\boldsymbol{Z}'\boldsymbol{Z}$ је производ броја $n-1$ и узорачке корелационе матрице објашњавајућих променљивих. Како бисмо ово потврдили, приметимо да се у j-тој врсти и k-тој колони матрице $\boldsymbol{Z}'\boldsymbol{Z}$ налази израз $\sum_i z_{ij}z_{ik}$. Обзиром да је $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, где су $\bar{x}_j$ и s_j узорачка средина и стандардно одступање посматраних вредности променљиве X_j, у j-тој врсти и k-тој колони матрице $\boldsymbol{Z}'\boldsymbol{Z}$ налази се израз $\sum_i(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/(s_j s_k)$. Присетимо се да је $s_{jk} = \sum_i(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)/(n-1)$ узорачка коваријација променљивих X_j и X_k, па се у j-тој врсти и k-тој колони матрице $Z'Z$ налази израз $(n-1)s_{jk}/(s_j s_k)$. По дефиницији, $s_{jk}/(s_j s_k)$ представља узорачку корелацију између променљивих X_j и X_k.

У најбољем случају, када објашњавајуће променљиве нису међусобно корелисане, узорачка корелациона матрица је јединична матрица $\boldsymbol{I}$, па је $\boldsymbol{Z}'\boldsymbol{Z} = (n-1)\boldsymbol{I}$. Када постоји мултиколинеарност међу променљивим, ми можемо трансформисати матрицу $\boldsymbol{Z}'\boldsymbol{Z}$ тако што ћемо јој додати $k\boldsymbol{I}$. На тај начин добијамо израз (7.4).

Шта статистичари кажу о Риџ регресији?

Постоје многе контроверзе везане за Риџ регресију. У овом делу ћемо навести коментаре аутора неколико књига.

Neter, Wasserman, and Kutner (1983):

„Оцене добијене Риџ регресијом су стабилне, у смислу да нису осетљиве на мале модификације узорка. Насупрот томе, оцене добијене методом најмањих квадрата су јако нестабилне при променама узорка као и при постојању мултиколинеарности међу обајашњавајућим променљивим. Такође, Риџ регресиона функција повремено обезбеђује добра предвиђања нових опсервација које се разликују од опсервација на којима смо формирали регресиону функцију. Метода најмањих квадрата у тим случајевима даје незавидне резултате.“

„Главно ограничење Риџ регресије представља то што закључци ове процедуре нису применљиви и што тачна дистрибутивна својства нису позната.“

Raymond H.Myers (1990):

„Риџ регресија је, иако контроверзна, једна од најпопуларнијих метода за оцењивање параметара код којих се јавља проблем мултиколинеарности. Процедуре описане у овом делу спадају у категорију пристрасних техника оцењивања. Оне су засноване на следећој идеји: иако обична метода најмањих квадрата даје непристрасне оцене и обезбеђује минималну варијансу свих параметара који се оцењују, не постоји горња граница за варијансу и појава мултиколинеарности може да проузрокује настанак високе варијансе. Као резултат тога, можемо да приметимо да, уколико постоји мултиколинеарност, плаћа се висока цена постизања непристрасних оцена применом методе најмањих квадрата. Пристрасна оцена постиже значајно смањење варијансе праћено повећањем стабилности регресионих коефицијената.“

Draper and Smith (1981):

„Из ове дискусије можемо закључити да је употреба Риџ регресије савршено разумна у случајевима код којих се верује да су велике вредности за параметар бета нереалне из практичног угла гледања. Међутим, треба увидети да је избор вредности k еквивалентан личном мишљењу појединца, који врши моделирање, о томе колика је заправо вредност параметра бета. У случајевима у којима онај који врши моделирање, не може да прихвати чињеницу о рестрикцији параметара бета, ова метода је потпуно неприкладна.“

„Уопштено, ми смо против индискриминисане употребе Риџ регресије осим у случајевима када су њена ограничења захвална.“

Пример 7.1

У табели 8.2 дато је 13 смеша цемента. Свака смеша је направљена од четири састојка: трикалцијум алуминат (X_1), трикалцијум силикат (X_2), трикалцијум алумино ферит (X_3), и дикалцијум силикат (X_4). У табели су приказани проценти ових састојака у свакој од четири смеше цемента. Вршен је експеримент о томе колико релативне количине ових састојака утичу на топлоту (Y), која се постиже приликом очвршћавања цемента. Топлота је изражена у калоријама по граму цемента.

Табела 7.1 Узорак „Смеша цемента“.

Број смеше	**Топлота (Y)**	**Састојак 1 (X_1)**	**Састојак 2 (X_2)**	**Састојак 3 (X_3)**	**Састојак 4 (X_4)**
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Пре него што кренемо на Риџ оцењивање, морамо да стандардизујемо објашњавајуће променљиве. Стандардизоване променљиве у узорку „Смеша цемента“ дате су у табели 7.2.

Помоћу стандардизованих података из табеле (7.2) ћемо прво наћи оцену методом најмањих квадрата

$$\hat{\mu}_{MNK} = 95.42, \qquad \widehat{\boldsymbol{\gamma}}_{\boldsymbol{MNK}} = \begin{bmatrix} 9.124 \\ 7.937 \\ 0.653 \\ -2.412 \end{bmatrix}, \qquad \text{и} \qquad \hat{\sigma}_{MNK} = 2.466$$

Даље ћемо израчунати $\|\widehat{\boldsymbol{\gamma}}_{\boldsymbol{MNK}}\|^2 = (9.124)^2 + (7.937)^2 + (0.653)^2 + (-2.412)^2 = 152.5$ и $k = \frac{4(2.446)^2}{152.5} = 0.1569$. Сада ћемо да израчунамо израз (7.4)

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} 7.644 \\ 4.667 \\ -0.910 \\ -5.835 \end{bmatrix}.$$

Регресиона права ће имати облик $\hat{Y} = 95.42 + 7.644Z_1 + 4.667Z_2 - 0.910Z_3 - 5.835Z_4$.

Табела 7.2 Стандардизовани узорак „Смеша цемента“.

Број смеше	**Топлота (Y)**	**Састојак 1 стандардизован (Z_1)**	**Састојак 2 стандардизован (Z_2)**	**Састојак 3 стандардизован (Z_3)**	**Састојак 4 стандардизован (Z_4)**
1	78.5	-0.0785	-1.4237	-0.9007	1.7923
2	74.3	-1.0985	-1.2309	0.5044	1.3144
3	104.3	0.6015	0.5042	-0.5885	-0.5974
4	87.6	0.6015	-1.1024	-0.5885	1.0156
5	95.9	-0.0785	0.2472	-0.9007	0.1792
6	109.2	0.6015	0.4400	-0.4323	-0.4779
7	102.7	-0.7585	1.4682	0.8167	-1.4338
8	72.5	-1.0985	-1.1024	1.5973	0.8364
9	93.1	-0.9285	0.3757	0.9728	-0.4779
10	115.9	2.3015	-0.0742	-1.2130	-0.2390
11	83.8	-0.985	-0.5240	1.7534	0.2390
12	113.3	0.6015	1.1469	-0.4323	-1.0754
13	109.4	0.4315	1.2754	-0.5885	-1.0754

-Мултиколинеарност

Скуп променљивих је колинеаран ако једна од њих представља линеарну комбинацију осталих променљивих. Скуп променљивих је приближно колинеаран ако једна од њих представља приближно линеарну комбинацију осталих променљивих. Корелација између сваке две од објашњавајућих променљивих у узорку „Смеша цемента“ дата је табелом (7.3). Постоји веома висока негативна корелација, -0.973, између променљивих X_2 и X_4. Ако погледамо узорак можемо видети разлог за ову појаву. Укупан проценат компонената X_2 и X_4 је око 77%, па важи да је X_4 приближно $77 - X_2$. Одатле закључујемо да је тешко направити разлику између ефеката променљивих X_2 и X_4. На пример, четири највеће вредности за X_4, које су $60, 52, 47$ и 44, се јављају у смеши која постиже температуру мање од просечне, која је 95.4. Одатле бисмо могли да закључимо да велике количине компоненте 4 доприносе смањењу постигнуте количине топлоте. Видимо да се 4 највеће вредности компоненте 4 јављају у смеши заједно са 4 најмање вредности компоненте 2, што одговара високој негативној корелацији. Дакле, можда висок садржај

компоненте 4 није одговоран за мању постигнуту температуру, већ је то низак садржај компоненте 2. На овом примеру видимо да нам висок ниво корелације између две променљиве отежава да установимо утицај сваке појединачне променљиве.

За узорак „Смеша цемента“ вредности FPV за променљиве X_1, X_2, X_3 и X_4 су $38.5, 254.4, 46.9$ и 282.5. За променљиву X_j постоји проблем мултиколинеарности ако је R_j^2 блиско јединици, односно ако је FPV_j велики број. Компјутерски програм Minitab даје упозорење ако је FPV_j веће од 100.

Табела 7.3 Корелација између објашњавајућих променљивих у узорку „Смеша цемента“.

Пар променљивих	**Корелација**
X_1, X_2	0.229
X_1, X_3	-0.824
X_1, X_4	-0.245
X_2, X_3	-0.139
X_2, X_4	-0.973
X_3, X_4	0.030

Литература коришћена у овом поглављу:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generalizations. Least squares and alternatives.* 1995.

[3] Julian J.Fraway. Linear models with R,2005.

[4] Ronald Christensen. Department of mathematics and statistics. University of New Mexico*. Plane answers to complex questions: The theory of linear models.* 2001.

[5] Hoerl A., Aplications of ridge analysis to regression problems. Chemical Engineering Progress. 1962.

8. Закључак

Метода најмањих квадрата је оптимална за коришћење у случајевима када популација грешака има нормалну расподелу. Метода најмањих апсолутних одступања је ефикасна у случајевима у којима расподела случајних грешака има тешке репове. Хуберова М-регресија и непараметарска регресија заснована на ранговима су ефикасне када расподела грешака има тешке репове и обе дају резултате приближно добре као и метода најмањих квадрата, када су случајне грешке нормално расподељене. Бајесова регресија укључује претходна знања о подацима у регресиону анализу. Риџова регресија се користи уколико постоји мултиколинеарност међу објашњавајућим променљивим. Још неке од метода које се могу користити су:

1. Робусне методе: Л-оцена, Р-оцена, Оцене са високом тачком прелома
2. Регресија главних компоненти
3. Еколошка регресија
4. Lasso регресија
5. Оцене максималне веродостојности.

Препоручује се да на једном скупу података искористимо више метода регресије како бисмо били сигурни у тачност резултата. Ако на истим подацима, користећи неколико различитих метода регресије, добијемо сличне резултате, можемо бити сигурни у тачност наших закључака. Ако добијемо значајна одступања у резултатима треба да, међу подацима, пронађемо узрок тих одступања. Што више метода регресије познајемо имамо већи избор за прављење модела. Све методе представљене у овом раду су ефикасне и на нама је да изаберемо у којој ситуацији ћемо користити коју методу.

Литература:

[1] David Birkes, Yadolah Dodge. *Alternative methods of regression*,1993.

[2] Rao, Tautenburg, Shalbh, Heumann. *Linear models and generalizations. Least squares and alternatives.1995.*

[3] Julian J.Fraway. Linear models with R,2005.

[4] Norman R.Draper, Harry Smith. *Applied regression analysis.* University of Wisconsin, 1998.

[5] Ronald Christensen. Department of mathematics and statistics. University of New Mexico*. Plane answers to complex questions: The theory of linear models.2001.*

[6] N.H.Bingham, John M.Fry.*Regression. Linear models in statistics.* 2010.

[7] Hill, R.W., and P.W.Holland. *Two robust alternatives to least squares regression*. Journal of the American Statistical Association. 1977.

[8] W.H.Chang, J.W.McKean, J.D.Naranjo, and S.J.Sheather. *High-breakdown rank regression.*

[9] K.S.Crimin, A.Abebe, and J.W.McKean*. Robust general linear models and graphics via a user interface.*

[10] Broeming, L.D. *Bayesian Analysis of Linear Models*. Marcel Dekker, New Zork, 1985.

[11] Hoerl A., Aplications of ridge analysis to regression problems. Chemical Engineering Progress. 1962.