Регрессиядағы факторлық айнымалылар.

Регрессия теңдеуінің интерпретациясы.

Регрессия диагностикасы. Көпмүшелік және сплайндық регрессия.

Регрессиядағы факторлық айнымалылар

Категориялық айнымалылар деп аталатын факторлық айнымалылар дискретті мәндердің шекті санын алады. Мысалы, несиенің мақсаты" қарызды шоғырландыру"," үйлену тойы"," автокөлік " және т.б. болуы мүмкін. екілік (иә/жоқ) айнымалы, сонымен қатар индикатор айнымалысы деп аталады, факторлық айнымалының ерекше жағдайы. Регрессия сандық деректерді енгізуді талап етеді, сондықтан факторлық айнымалыларды модельде қолдануға болатындай етіп қайта кодтау қажет. Жалпы қабылданған тәсіл айнымалыны екілік жалған айнымалылар жиынтығына түрлендіруден тұрады.

Негізгі терминдер

Жалған айнымалылар (dummy variables) регрессияда және басқа модельдерде қолдану үшін факторлық деректерді қайта кодтау арқылы алынған 0-1 форматындағы екілік айнымалылар.

Анықтамалық кодтау (анықтамалық кодтау) статистикада қолданылатын жалпы қабылданған кодтау түрі, онда фактордың бір деңгейі сілтеме ретінде пайдаланылады, ал басқа факторлар осы деңгейге сәйкес келеді. Синоним: біріктірілген кодтау.

Бір белсенді күйді кодтаушы (h hot encoding) Машиналық оқыту қауымдастығында жалпы қабылданған кодтау түрі факторлардың барлық деңгейлерін сақтайды. Машиналық оқытудың белгілі бір алгоритмдерінде кеңінен қолданылады; сонымен қатар, бұл әдіс бірнеше сызықтық регрессияға сәйкес келмейді. Ауытқуларды кодтау (deviation encoding) әр деңгейді анықтамалық деңгеймен емес, жалпы орташа деңгеймен салыстыратын кодтау түрі. Синонимдер: қосындылардың контрасттары, эффектілерді кодтау, маргиналды кодтау.

Көп деңгейлі факторлық айнымалылар

Кейбір факторлық айнымалылар екілік жалған айнымалылардың үлкен санын шығара алады — Пошта индекстері факторлық айнымалылардың үлгісі болып табылады және АҚШ - та 43 мың пошта индексі бар. Мұндай жағдайларда деректерді, сондай-ақ болжамды айнымалылар мен нәтиже

арасындағы байланыстарды санаттарда пайдалы ақпараттың бар-жоғын анықтау пайдалы. Егер солай болса, онда барлық факторларды сақтаудың мағынасы бар ма, әлде деңгейлерді біріктіру керек пе, соны шешу керек. Кинг округінде үй сатумен байланысты 82 пошта индексі бар:

table(house$ZipCode)

9800 89118 98001 98002 98003 98004 98005 98006 98007 98008 98010 98011
1 1 358 180 241 293 133 460 112 291 56 163 98014 98019 98022
98023 98024 98027 98028 98029 98030 98031 98032 98033 85 242 188 455
31 366 252 475 263 308 121 517 98034 98038 98039 98040 98042 98043
98045 98047 98050 98051 98052 98053 575 788 47 244 641 1 222 48
7 32 614 499 98055 98056 98057 98058 98059 98065 98068 98070 98072
98074 98075 98077 332 402 4 420 513 430 1 89 245 502 388
204

98092 98102 98103 98105 98106 98107 98108 98109 98112 98113 98115 98116
289 106 671 313 361 296 155 149 357 1 620 364 98117 98118
98119 98122 98125 98126 98133 98136 98144 98146 98148 98155 619 492
260 380 409 473 465 310 332 287 40 358 98166 98168 98177 98178
98188 98198 98199 98224 98288 98354 193 332 216 266 101 225 393
3 4 9

Zipcode айнымалысы ерекше мәнге ие, өйткені ол үйдің құнына әсер ететін орналасу эффектісі үшін эрзат болып табылады. Барлық деңгейлерді қосу үшін 81 коэффициент қажет, бұл 81 еркіндік дәрежесіне сәйкес келеді. Бастапқы house_lm моделінде тек 5 еркіндік дәрежесі бар (бөлімді қараңыз. Осы тараудың басында" модель диагностикасы"). Сонымен қатар, бірнеше пошталық индекстер тек бір сатылымға ие. Кейбір тапсырмаларда пошта индексін субметрополиялық географиялық аймаққа сәйкес келетін алғашқы екі немесе үш цифрмен біріктіруге болады. Кинг округі үшін барлық дерлік сатылымдар 980xx немесе 981xx индекстерінде болады, сондықтан бұл көмектеспейді. Балама тәсіл-пошта индекстерін сату бағасы сияқты басқа айнымалыға сәйкес топтастыру. Егер сіз бастапқы модельдің қалдықтарын пайдаланып Пошта индекстері бар топтар құрсаңыз, одан да жақсы болады. Төмендегі код үзіндісінде dplyr осы 82 пошта индексін house_lm регрессиясының қалдығының медианасына негізделген бес топқа біріктіреді:

zip_groups <- house %>%

mutate(resid = residuals(house_lm)) %>%

```
group_by(ZipCode) %>%

summarize(med_resid = median(resid),

cnt = n()) %>%

arrange(med_resid) %>%

 mutate(cum_cnt = cumsum(cnt),

ZipGroup = ntile(cum_cnt, 5))

house <- house %>%

left_join(select(zip_groups, ZipCode, ZipGroup), by='ZipCode')
```

Медианалық қалдықтар әр пошта индексі үшін есептеледі және tile функциясы медианалық сұрыпталған пошта индекстерін бес топқа бөлу үшін қолданылады. Бөлімді қараңыз. "Бұрмаланған айнымалылар" осы тарауда оның бастапқы фитингті жақсартатын регрессия теңдеуінде термин ретінде қалай қолданылатыны туралы мысал келтірілген. Регрессияның сәйкестігін бағдарлауға көмектесу үшін қалдықтарды пайдалану принципі модельдеу процесінің негізгі қадамы болып табылады

Реттік факторлық айнымалылар

Кейбір факторлық айнымалылар фактор деңгейлерін көрсетеді; олар қатарлы факторлық айнымалылар немесе реттік категориялық айнымалылар деп аталады. Мысалы, несие деңгейі A, B, C және т.б. болуы мүмкін — әр деңгей алдыңғы деңгейге қарағанда үлкен тәуекелге ие. Реттік факторлық айнымалыларды әдетте сандық мәндерге түрлендіруге және сол күйінде пайдалануға болады. Мысалы, BldgGrade айнымалысы реттік факторлық айнымалы болып табылады. Оның деңгейлерінің бірнеше түрі кестеде келтірілген. 4.1. Бұл деңгейлер белгілі бір мәнге ие болғанымен, оның сандық мәндері жоғары деңгейдегі үйлерге сәйкес төменнен жоғары қарай реттелген. House_lm регрессия моделінде, оның реттелуі бөлімде орындалды.

Осы тараудың басында" бірнеше сызықтық регрессия", BldgGrade сандық айнымалы ретінде қарастырылды. Реттік факторларды сандық айнымалы ретінде қарастыру реттіліктегі ақпаратты сақтайды, ол факторға айналдыру кезінде жоғалады.

Кесте 4.1. Типтік деректер форматы

Мағынасы	Сипаттама

1	Бюджеті төмен
2	Орташадан төмен
5	Лайықты
10	Өте жақсы
12	Сәнді
15	Особняк

Регрессиядағы факторлық айнымалыларға арналған негізгі идеялар •

*Факторлық айнымалыларды регрессияда қолдану үшін сандық айнымалыларға түрлендіру қажет. * Факторлық айнымалыны P әр түрлі мәндермен кодтаудың жалпы қабылданған әдісі оларды 1 p - жалған айнымалыларды қолдану арқылы ұсынудан тұрады. •*

Көп деңгейлі факторлық айнымалы, тіпті өте үлкен деректер жиынтығында да, деңгейлері аз айнымалыға біріктіруді қажет етуі мүмкін. •

Кейбір факторлардың реттелген деңгейлері бар және оларды бір сандық айнымалы ретінде ұсынуға болады.

Регрессия теңдеуін түсіндіру

Деректер ғылымында регрессияның ең маңызды қолданылуы тәуелді айнымалыны (нәтижені) болжау болып табылады. Алайда кейбір жағдайларда болжаушылар мен нәтиже арасындағы байланыстың табиғатын түсіну үшін теңдеудің өзімен тікелей танысу үлкен рөл атқаруы мүмкін. Бұл бөлімде регрессия теңдеуін зерттеуге және оны түсіндіруге қатысты нұсқаулар берілген

Негізгі терминдер

Корреляцияланған болжамды айнымалылар (корреляцияланған болжаушы айнымалылар) болжамды айнымалылар жоғары корреляцияланған кезде жеке коэффициенттерді түсіндіру қиын.

Мультиколлинеарлық (multicollinearity) болжамды айнымалылар мінсіз немесе мінсіз корреляцияға ие болған кезде регрессия тұрақсыз болуы мүмкін немесе сіз оны санай алмайсыз. Синоним: коллинеарлық.

Бұрмаланған айнымалылар (confounding variables) маңызды болжаушы болып табылады, ол оны өткізіп жіберген кезде регрессия теңдеуіндегі

ойдан шығарылған байланыстарға әкеледі. Синоним: шатастыратын айнымалылар.

Негізгі әсерлер (main effects) болжамды айнымалы мен басқа айнымалыларға тәуелді емес нәтиже айнымалысы арасындағы байланыс.

Өзара әрекеттесулер (өзара әрекеттесулер) екі немесе бірнеше болжаушылар мен жауап арасындағы өзара тәуелді байланыс.

Болжамдарды тексеру: регрессияны диагностикалау

Түсіндірме модельдеуде (яғни зерттеу контекстінде) бұрын айтылған метрикалық көрсеткіштерге қосымша әртүрлі қадамдар қабылданады (бөлімді қараңыз. "Осы тараудың басында "модель диагностикасы") модельдің деректерге қаншалықты сәйкес келетінін диагностикалау үшін. Көпшілігі модельдің негізінде жатқан болжамдарды тексере алатын қалдықтарды талдауға негізделген. Бұл қадамдар болжамды дәлдік мәселесін тікелей шешпейді, бірақ олар болжамды орнату туралы пайдалы түсініктер бере алады.

Негізгі терминдер

Стандартталған қалдықтар (стандартталған резидуалдар) стандартты қалдықтар қатесіне бөлінген қалдықтар.

Шығарындылар (Outliers) деректердің қалған бөлігінен (немесе болжамды нәтижеден) алыс орналасқан жазбалар (немесе нәтиже мәндері).

Әсер етуші мән (influential value) регрессия теңдеуінде болуы немесе болмауы үлкен мәнге ие мән немесе жазба.

Иық (leverage) бір жазбаның рег - Ресей теңдеуіне әсер ету дәрежесі. Синонимдер: hat мәні, проекциялық матрицадағы диагональ.

Қалыптан тыс қалдықтар (қалыпты емес резидуалдар) қалыптан тыс бөлінген қалдықтар регрессияның кейбір техникалық шарттарын жоюы мүмкін; сонымен бірге олар әдетте деректер ғылымында алаңдаушылық туғызбайды.

Гетероскедастика (heteroskedasticity) кейбір нәтиже диапазондары дисперсиясы жоғары қалдықтарды көрсететін жағдай (бұл теңдеуде жоқ болжаушы туралы айтуы мүмкін).

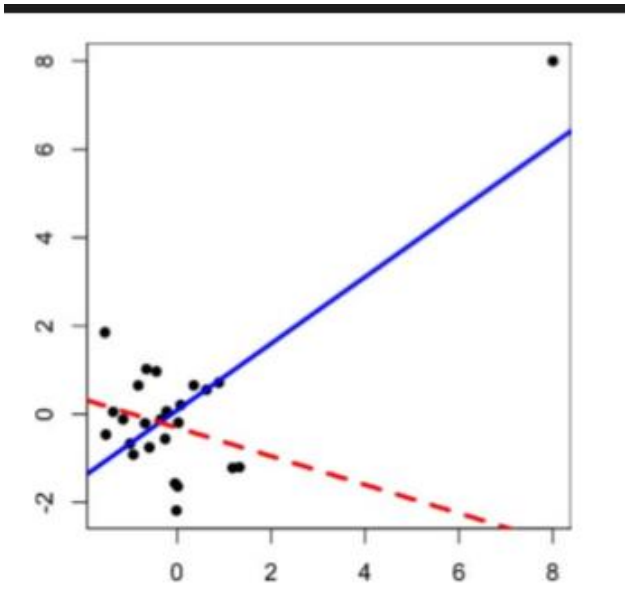
Жеке қалдық графиктері (partial residual plots) нәтиже айнымалысы мен жалғыз болжаушы арасындағы байланысты анықтауға арналған диагностикалық график. Синоним: қосылған айнымалылары бар график.

Шығарындылар

Жалпы айтқанда, шекті мән, лақтыру деп аталады, бұл басқа бақылаулардың көпшілігінен алыс болатын мән. Орталық ставка мен өзгергіштік бағаларын алу үшін шығарындыларды басқару керек сияқты (бөлімді қараңыз. "Орталық позицияны бағалау "және" өзгергіштікті бағалау " 1-тарау), шығарындылар регресс модельдерімен проблемалар тудыруы мүмкін. Регрессияда Эжекция-бұл нақты у мәні болжанған мәннен алыс болатын жазба. Шығарындыларды стандартталған қалдықты тексеру арқылы анықтауға болады, яғни. стандартты қалдық қатесіне бөлінген қалдық. Шығарындыларды шығарындылардан бөлетін статистикалық теория жоқ. Оның орнына бақылаудың деректердің негізгі бөлігінен қаншалықты алыс болуы керек екендігі туралы (ерікті) ережелер бар, бірақ оны Эжекция деп атауға болады. Мысалы, қорап диаграммасында шығарындылар қораптың шекарасынан тым жоғары немесе тым төмен деректер нүктелері болып табылады (бөлімді қараңыз. 1-тараудың" процентильдер және қорап диаграммалары"), мұндағы " тым " "1,5-тен асатын шаманы білдіреді. квартиль аралығына көбейту". Регрессияда стандартталған қалдықтың метрикалық көрсеткіші, әдетте, жазбаның таңдау санатына жатпайтынын анықтау үшін қолданылады. Стандартталған қалдықтарды "тікелей регрессиядан стандартты қателер саны"деп түсіндіруге болады.

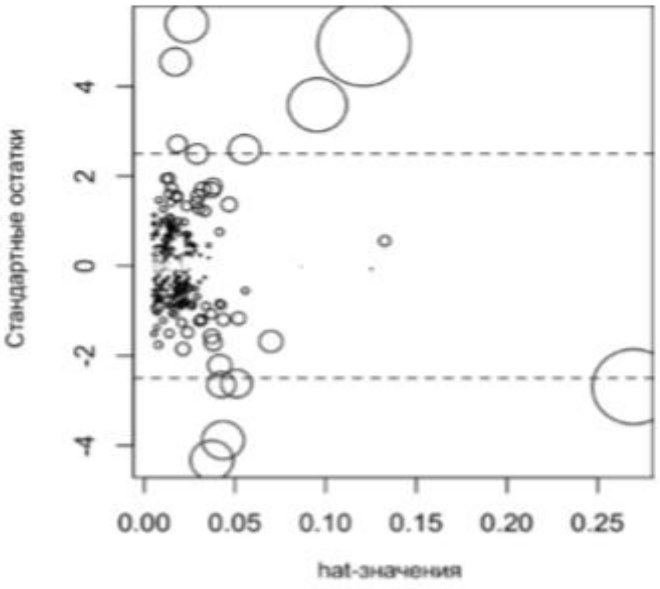
Ықпалды мәндер

Регрессия теңдеуін айтарлықтай өзгертетін мән ықпалды бақылау деп аталады. Регрессияда бұл мән үлкен қалдықпен байланысты болмауы керек. Мысал ретінде суреттегі тікелей регрессияны қарастырайық. 4.5. Нүктелі сызық барлық деректермен регрессияға сәйкес келеді, ал қалың сызық оң жақ жоғарғы бұрыштағы нүктемен регрессияға сәйкес келеді. Әрине, бұл деректер мәні регрессияға үлкен әсер етеді, бірақ ол үлкен шығарылыммен байланысты емес (толық регрессиядан алыс). Бұл деректер мәні регрессияда күшті иыққа (leverage) ие деп саналады. Стандартталған қалдықтардан басқа (бөлімді қараңыз. "Осы тараудың басында "шығарындылар"), статистиктер регрессияға бір жазбаның әсерін анықтау үшін бірнеше метрикалық көрсеткіштерді әзірледі. Жалпы қабылданған иық өлшемі-hat мәні; $2(P+1)/n$ жоғары мәндер жоғары иық деректерінің мәні туралы айтады.



4.5. -Сурет. Регрессиядағы ықпалды деректер нүктесінің мысалы

Тағы бір метрикалық көрсеткіш — Куктың қашықтығы (Cook ' s distance), ол әсерді иық пен қалдық мөлшерінің тіркесімі ретінде сипаттайды. Негізгі ереже-Куктың қашықтығы 4/(n-P- 1) - ден асатын болса, бақылаудың әсері жоғары



4.6. Сурет. Қандай бақылаулардың жоғары әсер ететінін анықтау кестесі

Әсер ету графигі немесе көпіршікті график стандартталған қалдықтарды, hat мәнін және Кук қашықтығын бір графикке біріктіреді. Суретте. 4.6 Кинг округінің тұрғын үй қоры туралы мәліметтер үшін әсер ету кестесі

көрсетілген. Бұл график R-дегі төмендегі код үзіндісінің көмегімен жасалуы мүмкін.

```
std_resid <- rstandard(lm_98105)

cooks_D <- cooks.distance(lm_98105)

hat_values <- hatvalues(lm_98105)

plot(hat_values, std_resid, cex=10*sqrt(cooks_D))

abline(h=c(-2.5, 2.5), lty=2)
```

Шамасы, берілген регрессияға үлкен әсер ететін бірнеше деректер нүктелері бар. Кук қашықтығын cooks функциясы арқылы есептеуге болады.қашықтық, және сіз диагностикалық көрсеткішті есептеу үшін hatvalues пайдалана аласыз. Hat мәндері x осінде, қалдықтары y осінде және нүктелік өлшемдер Кук қашықтығының мәнімен байланысты.

Кестеде. 4.2 регрессияны толық мәліметтер жиынтығымен және өте ықпалды деректер нүктелерімен салыстыру келтірілген. Bathrooms үшін регрессия коэффициенті айтарлықтай өзгереді

Кесте 4.2. Регрессия коэффициенттерін толық деректермен және жойылған ықпалды деректермен салыстыру

	Оригинал	Влиятельные убраны
(пересечение)	–772 550	–647 137
SqFtTotLiving	210	230
SqFtLot	39	33
Bathrooms	2282	–16132
Bedrooms	–26320	–22 888
BldgGrade	130 000	114 871

Болашақ деректерді сенімді түрде болжайтын регрессияны сәйкестендіру үшін ықпалды бақылауларды анықтау тек кішірек деректер жиынында пайдалы. Көптеген жазбалармен байланысты регрессиялар үшін кез-келген бақылаудың орнатылған теңдеуге шекті әсер ету үшін жеткілікті салмақ түсіруі екіталай (бірақ регрессия әлі де үлкен шығарындыларға ие болуы мүмкін). Аномалияны анықтау мақсатында ықпалды бақылауларды анықтау өте пайдалы болуы мүмкін.

Гетероскедастика, аномалия және корреляциялық қателер

Статистика қалдықтарды бөлуге көп көңіл бөледі. Кәдімгі ең кіші квадраттар екенін көрсетіңіз орын ауыстырылмаған және кейбір жағдайларда ха - рактер туралы кең ауқымды болжамдарға негізделген бағалаудың жалғыз" оңтайлы " критерийлері болып табылады. Бұл көптеген тапсырмаларда деректер талдаушылары қалдықтардың таралу сипаты туралы көп уайымдаудың қажеті жоқ дегенді білдіреді. Қалдықтардың таралуы негізінен болжамды дәлдікпен айналысатын деректер талдаушылары үшін минималды мәнге ие ресми статистикалық қорытындының (гипотезаны тексеру және P мәні) дұрыстығын растау үшін маңызды.

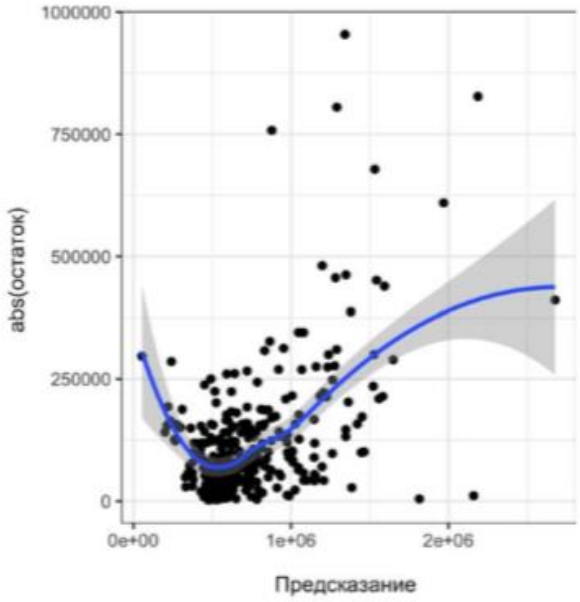
Ресми тұжырымның толық болуы үшін қалдықтар қалыпты түрде бөлінеді, бірдей дисперсияға ие және тәуелсіз деген болжам қабылданады. Деректер ғылымының талдаушылары үшін қызығушылық тудыруы мүмкін салалардың бірі - қалдықтардың табиғаты туралы болжамдарға негізделген болжамды мәндер үшін сенімділік аралықтарын стандартты есептеу (бөлімді қараңыз. Осы тараудың басында" сенімділік және болжау аралықтары").

Гетероскедастика-бұл болжамды мәндер диапазонында тұрақты қалдық дисперсияның болмауы. Басқаша айтқанда, диапазонның кейбір бөліктері үшін қателер басқаларға қарағанда көбірек. Ggplot2 бағдарламалық пакетінде қалдықтарды талдауға арналған бірнеше ыңғайлы құралдар бар.

Төмендегі код үзіндісі бөлімге орнатылған lm_98105 регрессиясының болжамды мәндеріне қарсы абсолютті қалдықтары бар графикті шығарады. Осы тараудың басында" шығарындылар".

```
df <- data.frame(

resid = residuals(lm_98105),

pred = predict(lm_98105))

ggplot(df, aes(pred, abs(resid))) +

geom_point() +

geom_smooth()
```

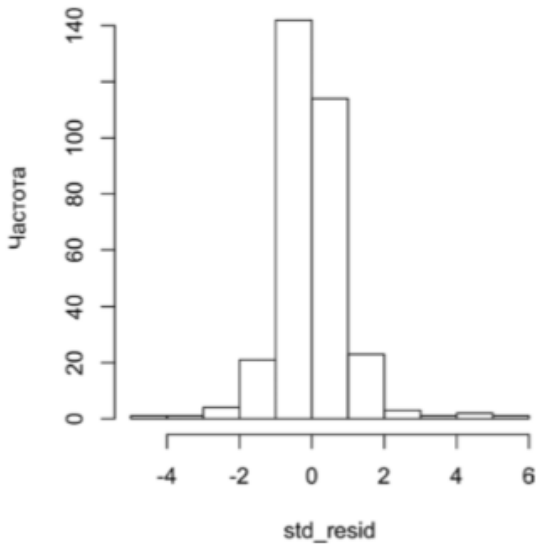
4.7-суретте. нәтиже кестесі ұсынылған. Geom_smooth көмегімен абсолютті қалдықтардан тегістелген қисық сызықты қолдану өте оңай. Бұл функция шашырау диаграммасындағы X және y осьтеріндегі айнымалылар арасындағы байланысты бағалау үшін визуалды тегістелген қисық құру үшін loess әдісін шақырады (осы тарауда "шашырау диаграммалары үшін тегістеу" кірістіруін қараңыз). Қалдықтардың дисперсиясы құны жоғары үйлер үшін өсетіні анық, бірақ болып табылады сонымен қатар құны төмен үйлер үшін үлкен. Бұл график lm_98105 регрессиясының гетероскедастикалық қателіктері бар екенін көрсетеді.



4.7. -сурет. Қалдықтардың абсолютті мәнінің графигі және болжамды мәндер

Суретте. 4.8 lm_98105 регрессиясы үшін стандартталған қалдықтардың гистограммасы берілген. Оның таралуы шұңқырлардың таралуына қарағанда ұзын құйрықтарға ие және үлкен қалдықтардың бо - леясына қарай орташа асимметрияны көрсетеді. Статистика мамандары қателіктер тәуелсіз деген болжамды да тексере алады. Бұл әсіресе ұзақ уақыт бойы жиналатын деректерге қатысты. Дурбин — Уотсон статистикасы (Дурбин — Уотсон) уақыт сериясының деректерімен конъюгацияланған регрессияда айтарлықтай автокорреляцияның бар-жоғын анықтау үшін пайдаланылуы мүмкін. Регрессия таралу сипаты туралы болжамдардың бірін бұзуы мүмкін болса да, бұл бізге қамқорлық жасауы керек пе? Көп жағдайда деректер ғылымында қызығушылықтың негізгі объектісі болжамды дәлдік болып табылады, сондықтан гетероскедастиканың қандай да бір талдауы зиян тигізбейді. Сіз деректерде сіздің моделіңіз қамтымаған сигналдың қандай да бір түрін таба

аласыз. Алайда, ресми статистикалық қорытындының (p - мәні, F-статистикасы және т. б.) дұрыстығын растау үшін тарату сипаты туралы болжамдарды қанағаттандыру деректерді талдаушы үшін қандай да бір ерекше маңыздылықты білдірмейді



Сурет. 4.8. Тұрғын үй қорының регрессиясынан қалған қалдықтардың гистограммасы

Жеке қалдық графиктері және сызықтық емес

Жеке қалдық Графиктері-бұл болжаушы мен нәтиже арасындағы байланысты қаншалықты жақсы есептелген сәйкестік түсіндіретінін визуализациялау тәсілі. Шығарындыларды анықтаумен қатар, бұл деректер талдаушылары үшін ең маңызды диагностикалық көрсеткіш болуы мүмкін. Жеке қалдық графигінің негізгі идеясы барлық басқа болжамды айнымалыларды ескере отырып, болжамды айнымалы мен жауап арасындағы байланысты оқшаулау болып табылады. Жеке қалдық "синтетикалық нәтиже" мәні ретінде бір болжаушыға негізделген болжамды толық регрессия теңдеуінің нақты қалдығымен біріктіру арқылы ұсынылуы мүмкін. I x болжаушысы үшін жеке қалдық-бұл жалпы қалдық және I X байланысты регрессия мүшесі

Частный остаток= Остаток + biX i

мұндағы ˆ i b-регрессияның бағалау коэффициенті. R-дегі predict функциясы возіі регрессиясының жеке мүшелерін қайтару мүмкіндігіне ие

terms <- predict(lm_98105, type='terms')

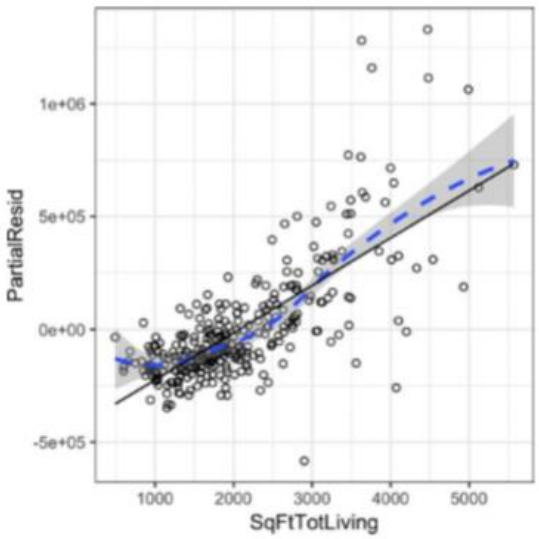
partial_resid <- resid(lm_98105) + terms

Жеке қалдықтар графигі x осінде і X және У осінде жеке қалдықтарды көрсетеді.

Ggplot2 бағдарламалық жасақтамасын пайдалану жеке қалдықтарды тегістеуді жеңілдетеді.

```
df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'], Terms = terms[, 'SqFtTotLiving'],

PartialResid = partial_resid[, 'SqFtTotLiving'])

ggplot(df, aes(SqFtTotLiving, PartialResid)) +

geom_point(shape=1) +

scale_shape(solid = FALSE) +

geom_smooth(linetype=2) +

geom_line(aes(SqFtTotLiving, Terms))
```

Алынған график суретте көрсетілген. 4.9. Жеке қалдық-бұл үлесті бағалау-иә, sqfttotliving сату бағасына қосады. Sqfttotliving пен сату бағасы арасындағы байланыс сызықтық емес екенін оңай көруге болады. Тікелей регрессия 1000 шаршы футтан аз үйлердің сату бағасын төмендетеді және 2000 және 3000 шаршы фут арасындағы үйлердің бағасын асыра бағалайды. 4000 шаршы футтан жоғары бұл үйлер үшін қорытынды жасау үшін тым аз деректер нүктелері бар.



Сурет. 4.9. SqFtTotLiving айнымалысы үшін жеке қалдық графигі

Бұл жағдайда бұл сызықтық емес мағынасы бар: шағын үйге 500 фут қосу үлкен үйге 500 фут қосудан әлдеқайда үлкен айырмашылыққа ие. Бұл sqfttotliving үшін қарапайым сызықтық мүшенің орнына сызықтық емес мүшені қарастыру керек екенін көрсетеді

Болжамдарды тексеруге арналған негізгі идеялар •

Шығарындылар шағын деректер жиынында қиындықтар тудыруы мүмкін болса да, шығарындыларға басты қызығушылық деректер мәселелерін анықтау немесе ауытқуларды локализациялау болып табылады. •

Жалғыз жазбалар (регрессиялық шығарындыларды қоса) шағын деректермен регрессия теңдеуіне үлкен әсер етуі мүмкін, бірақ бұл әсер үлкен деректерде қайталанады. •

Егер регрессиялық модель ресми қорытынды жасау үшін қолданылса (р мәні және т.б.), онда қалдықтардың таралу сипаты туралы белгілі бір болжамдарды салыстыру қажет. Жалпы алғанда, қалдықтарды бөлу деректер ғылымында маңызды емес. •

Жеке қалдық графигі регрессияның әрбір мүшесі үшін сәйкестікті сапалы диагностикалау үшін пайдаланылуы мүмкін, мүмкін балама модельдің ерекшелігі.

Сызықтық емес регрессия

Жауап пен болжамды айнымалы арасындағы байланыс міндетті түрде сызықтық емес. Препараттың дозасына жауап көбінесе сызықты емес: дозаны екі есе көбейту әдетте екі есе жауап бермейді. Өнімге деген сұраныс ақша жұмсау маркетингінің сызықтық функциясы емес,өйткені белгілі бір уақытта сұраныс қанағаттандырылуы мүмкін. Осы сызықтық емес әсерлерді алу үшін регрессияны кеңейтудің бірнеше жолы бар

Негізгі терминдер

Параболалық регрессия (polynomial regression) регрессияға көпмүшелік терминдерді (квадраттар, текшелер және т.б.) қосады. Синоним: көпмүшелік регрессия.

Сплайндық регрессия (spline regression) көпмүшелік сегменттер сериясымен тегіс қисыққа сәйкес келеді. Түйіндер (түйіндер) сплайн сегменттерін бөлетін мәндер.

Жалпыланған аддитивті модельдер (generalized additive models) автоматтандырылған түйін таңдауы бар Сплайндық модельдер.

Параболалық регрессия

Параболалық немесе көпмүшелік регрессия регрессия теңдеуінің құрамына көпмүшелік терминдерді қосумен байланысты. Параболалық регрессияны қолдану іс жүзінде 1815 жылы джергонн (Гергонне) мақаласында регрессияның өзін дамытудан басталды. мысалы, Y реакциясы мен x болжаушысы арасындағы квадраттық регрессия келесі формада болады:

$$Y = b_0 + b_1 X + b_2 X^2 + e.$$

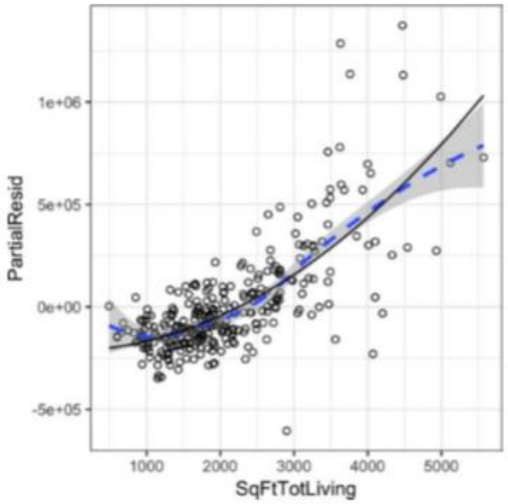
Параболалық регрессияны Poly функциясы арқылы реттеуге болады. Мысалы, төменде келтірілген код үзіндісі Sqfttotliving үшін квадраттық көпмүшені Кинг округінің тұрғын үй қоры туралы мәліметтермен сәйкестендіреді:

```
lm(AdjSalePrice ~ poly(SqFtTotLiving, 2) +
SqFtLot +
BldgGrade + Bathrooms + Bedrooms,
data=house_98105)
```

```
Call: lm(formula = AdjSalePrice ~ poly(SqFtTotLiving, 2) +
SqFtLot +     BldgGrade + Bathrooms + Bedrooms, data = house_98105)
Coefficients:
        (Intercept)  poly(SqFtTotLiving, 2)1
-402530.47                 3271519.49 poly(SqFtTotLiving, 2)2                 SqFtLot
776934.02                       32.56                BldgGrade                Bathrooms
135717.06                   -1435.12                Bedrooms                -9191.94
```

Енді екі коэффициент sqfttotliving-пен байланысты: біреуі сызықтық мүше үшін, екіншісі квадраттық мүше үшін (екінші дәрежелі мүше). Жеке қалдықтар кестесі (бөлімді қараңыз. "Жеке қалдық графиктері және сызықтық емес" осы тараудың басында) регрессия теңдеуіндегі кейбір қисықтық туралы айтады, sqfttotliving-пен байланысты. Орнатылған сызық

сызықтық фитингпен салыстырғанда жеке қалдықтардың тегістелгеніне дәлірек сәйкес келеді (келесі бөлімді қараңыз) (сурет. 4.10).



Сурет. 4.10. SqFtTotLiving айнымалысына сәйкес келетін параболалық регрессия (қатты сызық) және тегістелген (қалың сызық; сплайндар туралы келесі бөлімді қараңыз)

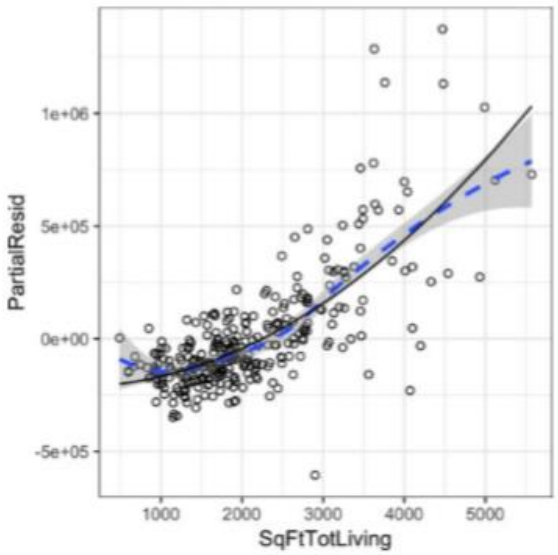
Сплайн регрессиясы

Параболалық регрессия сызықтық емес байланыстағы қисықтықтың белгілі бір мөлшерін ғана алады. Текше квадрат көпмүшелік сияқты жоғары дәрежелі терминдерді қосу көбінесе регрессия теңдеуінде қажетсіз "толқындарға" әкеледі. Сызықтық емес байланыстарды модельдеудің балама және жиі жоғары тәсілі-сплайндарды қолдану. Сплайндар бекітілген нүктелер арасында тегіс интерполяция әдісін ұсынады. Сплайндарды бастапқыда сызғыштар тегіс қисық сызу үшін қолданған, атап айтқанда кеме жасау және ұшақ жасау.

Сплайндар "үйректер" деп аталатын шайнектердің көмегімен ағаштың жұқа бөлігін деформациялау арқылы жасалды (сурет. 4.11).

4.11. - Сурет. Сплайндар бастапқыда Деформацияланатын ағаш пен "үйректер" негізінде жасалған және қисықтарды сәйкестендіру үшін сызу құралы ретінде пайдаланылған. Боб Перридің рұқсатымен фотосурет (Боб Перри)

Коэффициент тікелей мәнге ие сызықтық мүшеден айырмашылығы, сплайн мүшесінің коэффициенттері түсіндірілмейді. Оның орнына сплайнға сәйкестіктің табиғатын анықтау үшін визуалды дисплейді пайдалану пайдалы. Суретте. 4.12 жеке қалдықтардың регрессияға тәуелділігінің графигі көрсетілген. Көпмүшелік модельден айырмашылығы, сплайн моделі тегістелгенге әлдеқайда жақын, бұл сплайндардың икемділігін көрсетеді. Бұл жағдайда сызық деректерге әлдеқайда жақын орналасқан. Бұл сплайн регрессиясы жақсы үлгі екенін білдіре ме? Міндетті емес. Экономикалық тұрғыдан алғанда, өте кішкентай үйлердің (ауданы 1000 шаршы футтан аз) үйлерге қарағанда қымбатырақ болуы мағынасы жоқ - бұл өлшем. Бұл айнымалыны бұрмалайтын артефакт болуы мүмкін



Сурет. 4.12. Sqfttotliving (қатты сызық) айнымалысы үшін тегістелгенге (нүктелі сызық)сәйкес келетін сплайн регрессиясы

Жалпыланған аддитивті модельдер

Априорлық білімге немесе регрессияның диагностикалық көрсеткіштерін тексеруге байланысты жауап пен болжамды айнымалы арасындағы сызықтық емес байланысқа күдіктендіңіз делік. Көпмүшелік мүшелер байланысқа түсу үшін жеткілікті икемді болмауы мүмкін, ал сплайндық мүшелер түйіндерді анықтауды қажет етеді. Жалпыланған аддитивті модельдер (generalized additive models, GAM) — бұл сплайндық регрессияны автоматты түрде сәйкестендіруге арналған арнайы әдіс. Gem бағдарламалық пакеті K тұрғын үй қорының деректеріне GEM моделін сәйкестендіру үшін пайдаланылуы мүмкін

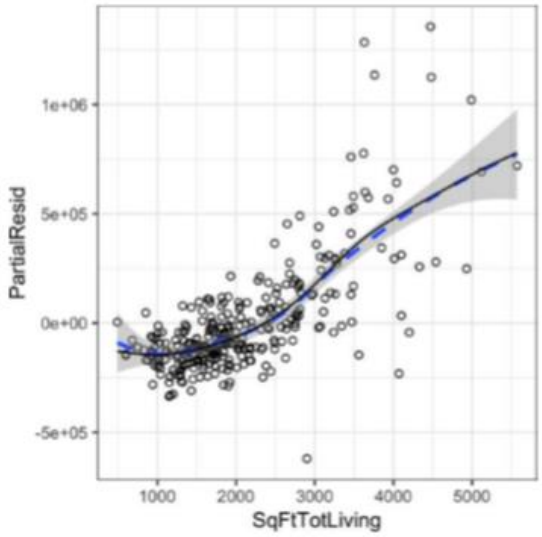
library(mgcv)

lm_gam <- gam(AdjSalePrice ~ s(SqFtTotLiving) + SqFtLot +

Bathrooms + Bedrooms + BldgGrade,

data=house_98105)

S (sqfttotliving) мүшесі GAM функцияларына теңдеудің сплайн мүшесі үшін "ең жақсы" түйіндерді табуды айтады (сурет. 4.13).



4.13. -Сурет. Sqfttotliving айнымалысына сәйкес келетін GAM регрессиясы (қатты сызық) тегістелген (нүктелі сызық)

Сызықтық емес регрессияның негізгі идеялары •

Регрессиядағы шығарындылар-бұл үлкен қалдықтары бар жазбалар. •

Мультиколлинеарлық регрессия теңдеуіне сәйкес келетін сандық тұрақсыздықты тудыруы мүмкін. •

Бұрмаланған айнымалы - бұл мо-Делиден алынып тасталған және ойдан шығарылған байланыстармен регрессия теңдеуіне әкелуі мүмкін маңызды болжаушы. •

Екі айнымалының өзара әрекеттесуін сипаттайтын теңдеу мүшесі, егер бір айнымалының әсері екіншісінің деңгейіне байланысты болса, қажет. •

Параболалық регрессия болжаушылар мен нәтиже айнымалысы арасындағы сызықтық емес байланыстарды реттей алады. •

Сплайндар-бұл түйіндерге бекітілген тізбектелген көпмүшелік сегменттер сериясы. •

Жалпыланған аддитивті модельдер (GAM) сплайндардағы түйіндерді анықтау процесін автоматтандырады.