Регрессия және болжау

Мүмкін статистикадағы ең көп таралған мақсат-сұрақтарға жауап беру: x айнымалысы байланысты Y айнымалысы бар, егер солай болса, онда бұл байланыс неде және біз оны Y болжау үшін пайдалана аламыз ба? Статистика мен деректер ғылымының бірлігі болжамның, атап айтқанда, басқа "болжау" айнымалыларының мәндеріне негізделген (мақсатты) айнымалының нәтижесін болжауға қарағанда еш жерде тығыз емес. Тағы бір маңызды өзара байланыс аномалияны анықтау саласында жатыр, мұнда деректерді талдауға және регрессиялық модельді жетілдіруге арналған регрессия диагностикасы деректердегі ерекше жазбаларды анықтау үшін пайдаланылуы мүмкін. Корреляция мен сызықтық регрессияның алғышарттары асады.

Қарапайым сызықтық регрессия

Қарапайым сызықтық регрессия немесе сызықтық жұптық регрессия бір айнымалының шамасы мен екіншісінің шамасы арасындағы байланысты модельдейді — мысалы, x ұлғайған сайын, y артады.немесе X ұлғайған сайын Y1 азаяды. Корреляция - бұл екі айнымалының қалай байланысты екенін өлшеудің тағы бір әдісі. Олардың арасындағы айырмашылық мынада: корреляция екі айнымалы арасындағы байланысты өлшейді, ал регрессия осы байланыстың табиғатын сүзеді.

Негізгі терминдер

Жауап (жауап) біз болжауға тырысатын айнымалы.

Синонимдер: тәуелді айнымалы, Y айнымалы, мақсат, нәтиже. Тәуелсіз айнымалы (тәуелсіз айнымалы) жауапты болжау үшін қолданылатын айнымалы. Синонимдер: тәуелсіз айнымалы, X-айнымалы, болжаушы, белгі, атрибут.

Жазу (record) -Жеке деректер элементі немесе жағдай үшін болжаушы мәндерден және нәтиже мәнінен тұратын Вектор.

Синонимдер: жол, жағдай, прецедент, үлгі, дана, мысал.

Қиылысу (intercept) -Регрессиялық түзудің қиылысы, яғни 0X = болғанда болжамды мән .

Синонимдер: b β, қиылысу нүктесі.

Регрессия коэффициенті (regression coefficient)

Регрессиялық түзудің көлбеуі.

Синонимдер: көлбеу, 11, B β, параметрлерді бағалау, салмақ.

Реттелген мәндер (fitted values) регрессия сызығынан алынған Оценки I Y ұпайлары.

Синоним: болжамды мағыналар.

Қалдықтар (residuals)

Бақыланатын мәндер мен орнатылған мәндер арасындағы айырмашылық.

Синоним: қателер.

Ең кіші квадраттар (жапырақ алаңдары)

Қалдық квадраттарының қосындысын азайту арқылы регрессияны реттеу әдісі.

Синонимдер: ең кіші квадраттардың әдеттегі әдісі, қарапайым тпа.

Регрессия теңдеуі

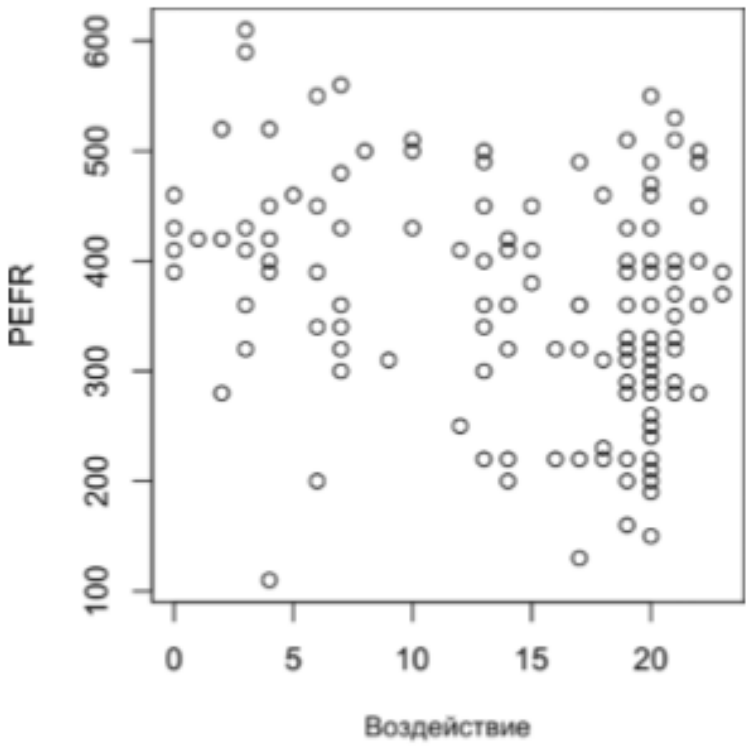
Қарапайым сызықтық регрессия X белгілі бір шамаға өзгерген кезде Y-нің қаншалықты өзгеретінін дәл бағалайды. Корреляция коэффициенті үшін X және Y айнымалылары бір-бірін алмастырады. Регрессия жағдайында біз x айнымалысынан Y айнымалысын сызықтық қатынасты (яғни түзу)пайдаланып болжауға тырысамыз:

$$Y = b_0 + b_1 X.$$

Бұл формула "Y 1 b-ге X-ге және 0 B тұрақтысына көбейтіледі"деп оқылады. 0 B еркін термині қиылысу (немесе тұрақты) деп аталады және X үшін 1B көлбеу коэффициенті екі мүше де R-де коэффициенттер ретінде көрсетіледі, дегенмен барлық жерде "коэффициент" термині көбінесе 1 b-ге арналған . Y айнымалысы жауап немесе тәуелді айнымалы деп аталады, өйткені ол X-ге тәуелді. болжаушы-болжаушы) немесе тәуелсіз айнымалы. Машиналық оқыту қауымдастығы басқа терминдерді қолдану тенденциясын көрсетеді, Y нысанасы мен X — векторы деп атайды.

4.1-Суреттегі шашырау диаграммасын қарастырыңыз., жұмысшының өкпе көлемінің көрсеткішіне (Pefr — дем шығарудың ең жоғары көлемдік жылдамдығы) қарсы мақта шаңына (экспозиция) ұшыраған жылдар санын көрсетеді. PER айнымалысы Exposure-мен қалай байланысты? Қарапайым сурет негізінде нақты бір нәрсені айту қиын.

$$Y = b_0 + b_1 X.$$



Сурет. 4.1. Мақтаның өкпе көлеміне қарсы әсері

Қарапайым сызықтық регрессия болжамды айнымалының функциясы ретінде PEFR реакциясын болжау үшін "оңтайлы" түзуді таңдауға тырысады Exposure.

$$\text{PEFR} = b_0 + b_1 \text{Exposure}.$$

R-дегі lm функциясын сызықтық регрессияны реттеу үшін пайдалануға болады.

model <- lm(PEFR ~ Exposure, data=lung)

lm сызықтық модельді (linear model) білдіреді және ~ таңбасы PER айнымалысын Exposure айнымалысы болжайтынын білдіреді.

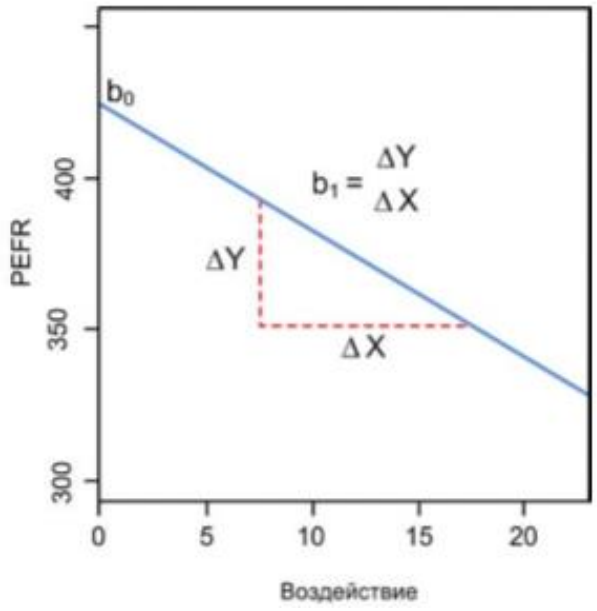
Model нысанын басып шығару келесі деректерді шығарады:

```
Call:
lm(formula = PEFR ~ Exposure, data = lung)

Coefficients:
(Intercept)     Exposure
    424.583       -4.185
```

Қиылысу немесе 0 b 424,583-ке тең және нөлдік әсер ететін PR жұмыс мәні үшін болжамды ретінде түсіндірілуі мүмкін, яғни 0 жыл. Регрессия коэффициенті немесе 1 b келесі түрде түсіндірілуі мүмкін: жұмысшы мақта шаңына ұшыраған әрбір қосымша жыл үшін жұмысшының pefr өлшеу нәтижесі 4,185 - ке азаяды .

Бұл модельдің тікелей регрессиясы суретте көрсетілген. 4.2.



Сурет. 4.2. Өкпе деректеріне сәйкес келетін регрессия үшін көлбеу және қиылысу

Орнатылған мәндер мен қалдықтар

Регрессиялық талдауда маңызды ұғымдар сәйкес мәндер мен қалдықтар болып табылады. Әдетте түзу қолда бар деректер арқылы дәл өтпейді,

сондықтан регрессия теңдеуі нақты түрде берілген қалдық мүшені қамтуы керек e-1 :

$$Y = b_0 + b_1 X + e_i.$$

Орнатылған мәндер немесе болжамды мәндер әдетте ˆ iY (Y-hat, у қалпақпен) деп белгіленеді.Олар келесі формуламен беріледі:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i.$$

Форма записи 0 ˆ b и 1 ˆ b говорит о том, что эти коэффициенты оценочные (расчет- ные) в отличие от известных (фактических).

Біз бастапқы деректерден болжамды мәндерді алып тастау арқылы ˆi e қалдықтарын есептейміз:

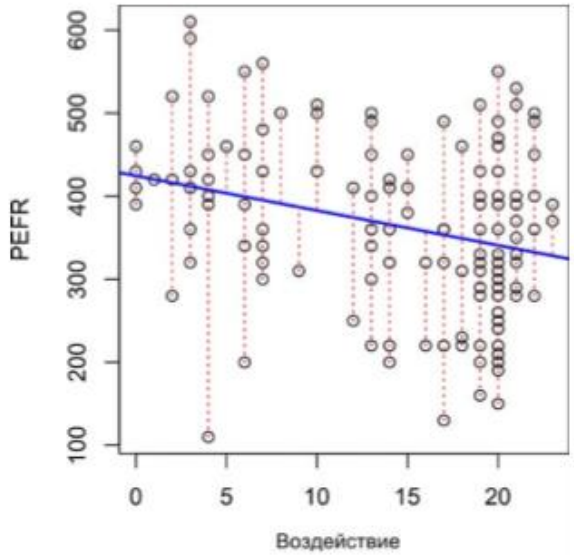
$$\hat{e}_i = Y_i - \hat{Y}_i.$$

В К predict және residuals функциясының көмегімен реттелген мәндер мен қалдықтарды алуға болады:

fitted <- predict(model)

resid <- residuals(model)

Суретте. 4.3 қалдықтар, яғни өкпе деректеріне бейімделген тікелей регрессиядан ауытқулар суреттелген. Қалдықтар-деректерден түзу сызыққа дейінгі тік нүктелі сызықтардың ұзындығы.



Сурет. 4.3. Қалдықтар, яғни тікелей регрессиядан ауытқулар
(суреттегі басқа у осінің шкаласына назар аударыңыз.
4.2, демек, түзудің басқа көлбеуі)

Ең кіші квадраттар

Модельді деректерге сәйкестендіру қалай жүзеге асырылады? Нақты байланыс болған кезде, сіз түзу сызықты қолмен ойша елестете аласыз. Іс жүзінде тікелей регрессия-бұл квадраттардың қалдық қосындысы немесе RSS (squares residual sum)деп аталатын қалдықтардың квадраттық мәндерінің қосындысын азайтатын бағалау:

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2.$$

0 b b және 1 b b ұпайлары RSS минимизациялайтын мәндер болып табылады. Квадрат қалдықтардың қосындысын азайту әдісі ең кіші квадраттардың регрессиясы немесе ең кіші квадраттардың әдеттегі әдісі (қарапайым МНҚ) деп аталады. Бұл әдіс көбінесе неміс математигі Карл Фридрих Гауссқа жатады, оны 1805 жылы алғаш рет француз математигі Ан - дре-Мари Легендре (Adrien-Marie Legendre) жариялады. Ең кіші квадраттардың регрессиясы коэффициенттерді есептеудің қарапайым формуласына әкеледі:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2};$$
$$\hat{b}_0 = \bar{Y} - \hat{b}_1 X.$$

Тарихи тұрғыдан алғанда, есептеу ыңғайлылығы регрессиядағы ең кіші квадраттар әдісін кеңінен қолданудың себептерінің бірі болып табылады.

Үлкен деректердің пайда болуымен оның есептеу жылдамдығы әлі де маңызды фактор болып табылады. Ең кіші квадраттар әдісі, орташа мән сияқты (бөлімді қараңыз. "Медиана және ұялшақ бағалар" 1-тарау), шығарындыларға сезімтал, дегенмен бұл факт кішігірім немесе орташа тапсырмаларда ғана маңызды мәселе болып табылады. Бөлімді қараңыз. "Шығарындылар" регрессиядағы шығарындыларды қарастыратын осы тарауда.

Болжау және түсіндіру (профильдеу).

Тарихи тұрғыдан алғанда, регрессияны бірінші кезекте қолдану болжамды айнымалылар мен өзгермелі нәтиже арасындағы болжамды сызықтық байланысты анықтау болды. Мақсат-байланысты түсіну және оны регрессияға сәйкес келетін мәліметтердің көмегімен түсіндіру. Бұл жағдайда басты назар Рег В регрессия теңдеуінің көлбеуінің бағалау мәніне аударылады . Эко-номистер тұтыну шығындары мен ЖІӨ өсімі арасындағы байланысты білгісі келеді. Денсаулық сақтау саласындағы Чи - новниктер қоғамды ақпараттандыру науқаны қауіпсіз жыныстық қатынас әдістерін ілгерілетуде тиімді ме, жоқ па, соны түсінгісі келуі мүмкін. Мұндай жағдайларда жеке жағдайларды болжау емес, керісінше жалпы байланысты түсіну басты назарда болады. Үлкен деректердің пайда болуымен регрессия қолда бар деректерді (яғни болжамды модель) статистикалық түсіндірудің орнына жаңа деректердің жеке нәтижелерін болжау мақсатында модель құру үшін кеңінен қолданылады. Бұл жағдайда негізгі мақсатты компоненттер ˆ Y мәндері болып табылады . Маркетингте регрессияны жарнамалық науқанның көлеміне жауап ретінде кірістің өзгеруін болжау үшін пайдалануға болады. Университеттер sat3 академиялық қабілеттерін анықтауға арналған емтихан бағалары негізінде студенттердің GPA орташа академиялық балын болжау үшін регрессияны пайдаланады. Деректерге жақсы бейімделген регрессия моделі X-дегі өзгерістер у-дегі өзгерістерге әкелетіндей етіп реттеледі, алайда, мұндай регрессия теңдеуі себептік кондиционердің бағытын дәлелдемейді. Шарттылық туралы тұжырымдар байланысты түсінудің кең контекстіне негізделуі керек. Мысалы, регрессия теңдеуі веб-жарнамадағы басу саны мен түрлендіру саны арасындағы белгілі бір байланысты көрсете алады. Біздің регрессия теңдеуі емес, маркетинг процесі туралы біліміміз бізді керісінше емес, сатылымды тудыратын жарнаманы басу деген қорытындыға әкеледі.

Қарапайым сызықтық регрессияның негізгі идеялары •

*Регрессия теңдеуі у жауап айнымалысы мен болжамды х айнымалысы арасындағы байланысты түзу түрінде модельдейді. * Регрессиялық модель сәйкес мәндер мен қалдықтарды береді — жауап болжау және болжау қателері. •*
Регрессиялық модельдерді сәйкестендіру, әдетте, ең кіші квадраттар әдісімен жүзеге асырылады. •
Регрессия болжау үшін де, статистикалық түсіндіру үшін де қолданылады.

Бірнеше сызықтық регрессия.
Бірнеше болжаушы болған кезде, берілген теңдеу оларды орналастыру үшін жай ғана кеңейеді:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e.$$

Түзудің орнына қазір бізде сызықтық модель бар - әрқайсысы арасындағы байланыс коэффициент-том және оның айнымалысы (белгісі) сызықтық.

Негізгі терминдер
Орташа квадраттық қате (Root mean squared error) орташа квадраттық регрессия қатесінің квадрат түбірі (регрессиялық модельдерді салыстыру үшін ең көп қолданылатын метрикалық көрсеткіш).
Синоним: RMSE.

Стандартты қалдық қатесі (residual standard error) орташа квадраттық қатемен бірдей, бірақ еркіндік дәрежелері үшін түзетілген.
Синоним: RSE.

R-квадрат (R-squared) 0-ден 1-ге дейінгі мәндері бар модельмен түсіндірілетін дисперсия үлесі.
Синонимдер: анықтау коэффициенті, R2.

t-статистика (t-statistic) модельдегі айнымалылардың маңыздылығын салыстыруға арналған метрикалық көрсеткіш, кез - келген болжаушы үшін регрессия коэффициентін стандартты коэффициент қатесіне бөлу нәтижесінде пайда болады.

Салмақты регрессия (салмақты регрессия) Регрессия, онда жазбалар әртүрлі салмақтарға сәйкес келеді.

Қарапайым сызықтық регрессиядан алынған барлық басқа ұғымдар, мысалы, ең кіші квадраттарға сәйкестендіру және сәйкес мәндер мен қалдықтарды анықтау, бірнеше сызықтық регрессияға дейін кеңейеді. Мысалы, орнатылған мәндер келесі формуламен беріледі:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1,i} + \hat{b}_2 X_{2,i} + ... + \hat{b}_p X_{p,i} + e.$$

Кросс-тексеру. (айқас тексеру)

Регрессияның классикалық статистикалық метрикалық көрсеткіштері (2 R , F-статистика және p-мәндер) "үлгі ішіндегі" көрсеткіштер болып табылады — олар модельді сәйкестендіру үшін пайдаланылған деректерге қолданылады. Интуитивті деңгейде сіз бастапқы деректерді модельді сәйкестендіру үшін пайдаланбай - ақ қоюдың мағынасы бар екенін көресіз, содан кейін модельді резервтелген (бір жаққа қойылған) деректерге қолдана отырып, оның қаншалықты жақсы жұмыс істейтінін көре аласыз. Әдетте, сіз деректердің едәуір бөлігін модельді сәйкестендіру үшін, ал қалған бөлігін тексеру үшін қолданасыз. Мұндай "таңдамалы емес" тексеру идеясы жаңа емес, бірақ ол үлкен деректер жиынтығы кең таралғанға дейін өзін дәлелдеген жоқ; деректердің шағын жиынтығына ие бола отырып, талдаушылар әдетте барлық қолда бар деректерді пайдаланғысы келеді және олардың негізінде ең жақсы модельге сәйкес келеді. Кейінге қалдырылған деректермен бақылау үлгісін пайдалану, дегенмен, шағын бақылау үлгісіндегі өзгергіштікке байланысты туындайтын кейбір белгісіздікке байланысты Сізді жүзге айналдырады. Егер сіз нашар деректермен басқа бақылау үлгісін алсаңыз, модель диагностикасының нәтижелері қаншалықты ерекшеленеді? Кросс-тексеру кейінге қалдырылған деректермен бақылау үлгісінің идеясын бірнеше дәйекті бақылау үлгілеріне дейін кеңейтеді. Негізгі k-блокты кросс-тексеру алгоритмі келесідей:

1. Бақылау үлгісі ретінде 1/K деректерді кейінге қалдырыңыз.
2. Модельді қалған деректерге үйрету.
3. Үлгіні 1 / k бақылау үлгісіне қолданыңыз (нәтижелерді бағалаңыз) және модель диагностикасының қажетті метрикалық көрсеткіштерін жазыңыз.
4. Алғашқы 1 / K деректерін қалпына келтіріп, келесі 1/k кейінге қалдырыңыз (бірінші рет таңдалған жазбаларды қоспағанда).
5. 2 және 3-қадамдарды қайталаңыз.
6. Әрбір жазба бақылау үлгісіне тағайындалған пайыздық үлесте пайдаланылғанша қайталаңыз.

7. Модельдің метрикалық диагностикалық көрсеткіштерін орташалаңыз немесе біріктіріңіз. Деректерді жаттығу үлгісіне және бақылау үлгісіне бөлу блоктарды бөлу (fold) деп те аталады.

Регрессияға негізделген болжам.

Деректер ғылымындағы регрессияның негізгі мақсаты-болжау. Мұны ескеру пайдалы, өйткені регрессия уақытпен тексерілген және танылған статистикалық әдіс бола отырып, болжаудан гөрі дәстүрлі түсіндірме модельдеу рөліне сәйкес келетін багажбен бірге жүреді.

Негізгі терминдер
Болжалды интервал (болжамды интервал) жеке болжамды мәннің айналасындағы белгісіздік интервалы.
Экстраполяция (экстраполяция) модельді оны сәйкестендіру үшін қолданылатын деректер ауқымынан тыс кеңейту.

Сенімді және болжамды аралықтар.

Статистикалық деректердің едәуір бөлігі өзгергіштікті (белгісіздікті) түсінуді және өлшеуді көздейді. регрессиядан шыққан кезде байланысатын T-статистикасы мен p-мәндері мұны формальды түрде шешеді, бұл кейде айнымалыны таңдау үшін пайдалы (бөлімді қараңыз. "Модель диагностикасы" бұрын осы тарауда). Неғұрлым пайдалы метрикалық көрсеткіштер - бұл сенімділік біліктері мәні регрессия мен болжау коэффициентінің айналасында орналасқан белгісіздік интервалдары. Оларды түсінудің оңай жолы — жүктеу жолағын қолдану (бөлімді қараңыз. 2-тараудың" жүктеу " жалпы жүктеу процедурасы туралы қосымша ақпаратпен). Бағдарламалық жүйелерден шығуда кездесетін ең көп таралған регрессиялық сенімділік аралықтары регрессия параметрлері (коэффициенттері) үшін сенімді аралықтар болып табылады. Әрі қарай, P болжаушылары және n жазбалары (жолдары) бар деректер жиынын пайдалана отырып, регрессия параметрлері (коэффициенттері) үшін сенімділік интервалдарын жасайтын жүктеу алгоритмі берілген: 1. Әр жолды (Мысырдан шығу айнымалысын қоса) бір "пакет" ретінде қарастырыңыз және барлық N пакеттерді қорапқа салыңыз. 2. Пакетті кездейсоқ алып тастаңыз, оның мәндерін жазып алыңыз және оны қорапқа қайтарыңыз. 3. 2-қадамды n рет қайталаңыз; енді сізде қайта таңдалған жүктеу үлгісі бар. 4. Регрессияны жүктеу үлгісіне сәйкестендіріңіз және бағалау

коэффициенттерін жазыңыз. 5. 2-4 қадамдарды 1000 рет қайталаңыз. 6. Енді сізде әр коэффициент бойынша 1000 жүктеу мәні бар; әрқайсысына сәйкес процентильдерді табыңыз (мысалы, 90% сенімділік аралығы үшін 5-ші және 95-ші).
R-де Сіз коэффициенттер үшін нақты жүктеу сенімділік интервалдарын құру үшін жүктеу функциясын қолдана аласыз немесе R - де әдеттегі шығыс болып табылатын формулаларға негізделген интервалдарды қолдана аласыз. Тұжырымдамалық мағына мен интерпретация бірдей болып қалады және деректерді талдаушылар үшін бірінші кезектегі қажеттілік болып табылмайды, өйткені бұл мәліметтер регрессия коэффициенттеріне қатысты. Деректер талдаушылары үшін болжамды у (ˆ і Ү) мәндерінің айналасындағы аралықтар үлкен қызығушылық тудырады. ˆ і у айналасындағы белгісіздік екі көзден шығады:

- сәйкес болжаушы айнымалылар мен олардың коэффициенттері қандай екендігі туралы белгісіздік (жоғарыдағы Жүктеу алгоритмін қараңыз);
- жеке деректер нүктелеріне тән қосымша қате.

Жеке деректер нүктесінің қатесін келесідей көрсетуге болады: регрессия теңдеуі қандай болғаны белгілі болса да (мысалы, сәйкестікті орындау үшін көптеген жазбалар болса), берілген болжаушы мәндер жиыны үшін нақты нәтиже мәндері өзгереді. Мысалы, әрқайсысы 8 бөлмелі, жалпы ауданы 6500 шаршы фут, 3 Ванна бөлмесі және жертөлесі бар қанша үйдің құны әртүрлі болуы мүмкін. Бұл жеке қатені реттелген мәндердің қалдықтарымен модельдеуге болады. Регрессиялық модель қатесін және жеке деректер нүктесінің қатесін модельдеуге арналған жүктеу алгоритмі келесідей болады: 1. Деректерден жүктеу үлгісін алыңыз (бұрын егжей-тегжейлі түсіндірілген). 2. Регрессияға сәйкес келу және жаңа мағынаны болжау. 3. Бастапқы регрессиядан кездейсоқ бір қалдық алыңыз, оны болжанған мәнге қосыңыз және нәтижені жазыңыз. 4. 1-3 қадамдарды 1000 рет қайталаңыз. 5. Нәтижелердің 2,5 және 97,5 пайызын табыңыз.

Регрессияны қолдана отырып болжаудың негізгі идеялары

- *Деректер ауқымынан тыс Экстраполяция қатеге әкелуі мүмкін.*
- *Сенімділік аралықтары регрессия коэффициенттерінің айналасындағы белгісіздікті анықтайды.*
- *Болжалды аралықтар жеке болжамдардағы белгісіздікті анықтайды.*

- *Бағдарламалық жасақтама жүйелерінің көпшілігі, соның ішінде R, формулаларды қолдана отырып, стандартты болжамды және сенімділік интервалдарын шығарады.*
- *Жүктеу жолағын да қолдануға болады; оның түсіндірмесі мен идеясы өзгеріссіз қалады.*