Теоретични основи на линейната регресия.

1. Понятие за регресия

Регресията търси връзки между променливите. Например можете да наблюдавате няколко служители на някоя компания и да се опитате да разберете как заплатите им зависят от техните характеристики, като например опит, ниво на образование, роля, град на работа и т.н.

Това е регресионен проблем, при който данните, свързани с всеки служител, представляват едно наблюдение. Презумпцията е, че опитът, образованието, ролята и градът са независими характеристики, а заплатата зависи от тях.

По същия начин можете да се опитате да установите математическата зависимост на цените на жилищата от площта, броя на спалните, разстоянието до центъра на града и т.н.

Като цяло при регресионния анализ се разглежда някакво явление, което ви интересува, и разполагате с определен брой наблюдения. Всяко наблюдение има две или повече характеристики. Следвайки предположението, че поне една от характеристиките зависи от останалите, се опитвате да установите връзка между тях.

С други думи, трябва да намерите функция, която достатъчно добре съпоставя някои характеристики или променливи с други.

Зависимите характеристики се наричат зависими променливи, изходи или отговори. Независимите характеристики се наричат независими променливи, входове, регресори или предиктори.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Задачите за регресия обикновено имат една непрекъсната и неограничена зависима променлива. Входящите данни обаче могат да бъдат непрекъснати, дискретни или дори категорични данни, като пол, националност или марка.

Обичайна практика е изходните данни да се означават с y, а входните с x. Ако има две или повече независими променливи, тогава те могат да бъдат представени като вектор $\mathbf{x} = (x_1, ..., x_r)$, където r е броят на входовете.

Кога е необходима регресия?

Обикновено се нуждаете от регресия, за да отговорите на въпроса дали и как едно явление влияе на друго или как са свързани няколко променливи. Например можете да я използвате, за да определите дали и до каква степен опитът или полът влияят върху заплатите.

Регресията е полезна и когато искате да прогнозирате даден отговор, като използвате нов набор от предиктори. Например можете да се опитате да прогнозирате потреблението на електроенергия от дадено домакинство за следващия час, като вземете предвид външната температура, времето на деня и броя на жителите в това домакинство.

Регресията се използва в много различни области, включително в икономиката, компютърните науки и социалните науки. Нейното значение нараства с всеки изминал ден с наличието на големи количества данни и с нарастващото осъзнаване на практическата стойност на данните.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

2. Линейна регресия

Линейната регресия е вероятно една от най-важните и широко използвани техники за регресия. Тя е сред най-простите регресионни методи. Едно от основните ѝ предимства е лесното интерпретиране на резултатите.

Формулиране на проблема

Когато прилагате линейна регресия на някаква зависима променлива y върху набор от независими променливи $\mathbf{x} = (x_1, ..., x_r)$, където r е броят на предикторите, приемате, че между y и $\mathbf{x}$ има линейна зависимост: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. Това уравнение е уравнението на регресията. $\beta_0, \beta_1, ..., \beta_r$ са регресионните коефициенти, а ε е случайната грешка.

Линейната регресия изчислява оценките на регресионните коефициенти или просто предсказаните тегла, означени с $b_0, b_1, ..., b_r$. Тези оценители определят оценената регресионна функция $(\mathbf{x}) = b_0 + b_1 x_1 + \cdots + b_r x_r$. Тази функция трябва да улавя достатъчно добре зависимостите между входовете и изхода.

Оценената или прогнозирана реакция, $(\mathbf{x}_i)$, за всяко наблюдение $i = 1, ..., n$, трябва да бъде възможно най-близка до съответната действителна реакция y_i. Разликите $y_i - (\mathbf{x}_i)$ за всички наблюдения $i = 1, ..., n$, се наричат остатъци. Регресията е свързана с определяне на най-добрите прогнозни тегла - т.е. теглата, съответстващи на най-малките остатъци.

За да получите най-добрите тегла, обикновено минимизирате сумата на квадратните остатъци (SSR) за всички наблюдения $i = 1, ..., n$: $SSR = \Sigma_i(y_i - f(\mathbf{x}_i))^2$. Този подход се нарича метод на обикновените най-малки квадрати.

Изпълнение на регресията

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Варирането на действителните отговори y_i, $i = 1, ..., n$, се дължи отчасти на зависимостта от предикторите $\mathbf{x}_i$. Съществува обаче и допълнителна присъща дисперсия на изхода.

Коефициентът на детерминация, означаван като R^2, ви казва каква част от вариацията в y може да бъде обяснена чрез зависимостта от $\mathbf{x}$, като се използва конкретният регресионен модел. По-голямото R^2 показва по-добро прилягане и означава, че моделът може по-добре да обясни вариацията на продукцията при различни входове.

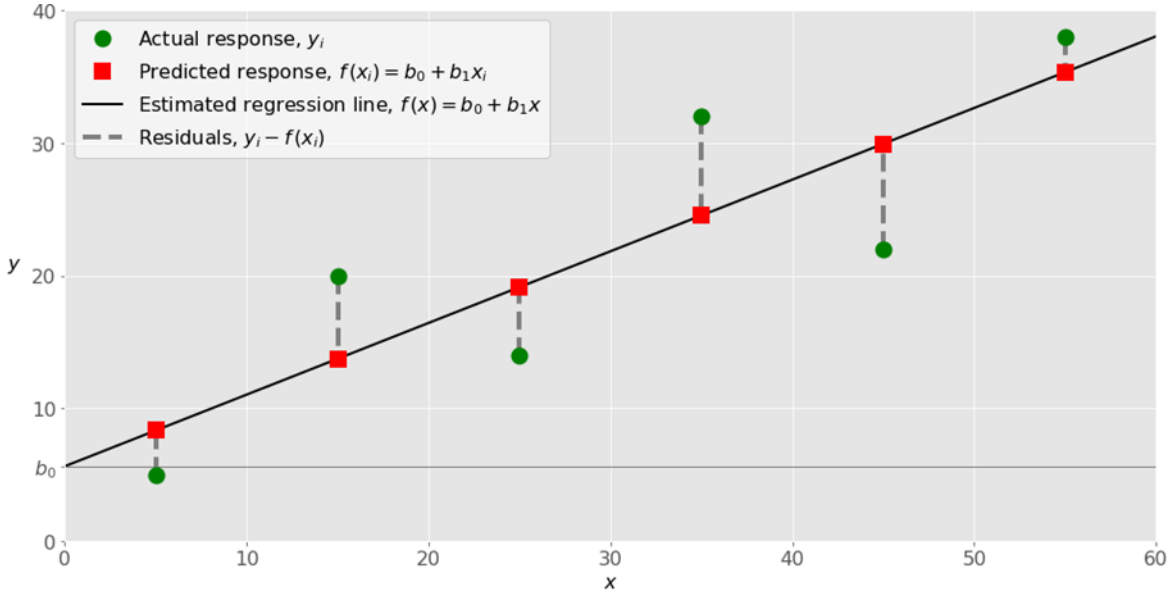
Стойността $R^2 = 1$ съответства на SSR = 0. Това е перфектното напасване, тъй като стойностите на предсказаните и действителните отговори напълно съвпадат една с друга.

3. Проста линейна регресия

Простата или едновариантна линейна регресия е най-простият случай на линейна регресия, тъй като има една единствена независима променлива, $\mathbf{x} = x$.

Следващата фигура илюстрира простата линейна регресия:

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.



Пример за проста линейна регресия

Когато прилагате проста линейна регресия, обикновено започвате с даден набор от двойки вход-изход (x-y). Тези двойки са вашите наблюдения, показани като зелени кръгове на фигурата. Например най-лявото наблюдение има вход $x = 5$ и действителен изход, или отговор, $y = 5$. Следващото наблюдение има $x = 15$ и $y = 20$ и т.н.

Оценената регресионна функция, представена с черната линия, има уравнението $(x) = b_0 + b_1 x$. Вашата цел е да изчислите оптималните стойности на прогнозните тегла b_0 и b_1, които минимизират SSR, и да определите оценената регресионна функция.

Стойността на b_0, наричана още прехващане, показва точката, в която оценената регресионна линия пресича оста y. Това е стойността на оценения отговор (x) за $x = 0$. Стойността на b_1 определя наклона на оценената регресионна линия.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Прогнозираните отговори, показани като червени квадрати, са точките върху регресионната линия, които съответстват на входните стойности. Например, за входните стойности $x = 5$, предсказаният отговор е (5) = 8,33, който представлява най-левият червен квадрат.

Вертикалните прекъснати сиви линии представляват остатъците, които могат да бъдат изчислени като y_i - ($\mathbf{x}_i$) = y_i - b_0 - $b_1 x_i$ за i = 1, ..., n. Това са разстоянията между зелените кръгчета и червените квадратчета. Когато прилагате линейна регресия, всъщност се опитвате да минимизирате тези разстояния и да направите червените квадрати възможно най-близки до предварително зададените зелени кръгове.

4. Множествена линейна регресия

Множествената или многовариантната линейна регресия е случай на линейна регресия с две или повече независими променливи.

Ако има само две независими променливи, тогава оценената регресионна функция е $(x_1, x_2) = b_0 + b_1 x_1 + b_2 x_2$. Тя представлява регресионна равнина в триизмерно пространство. Целта на регресията е да се определят стойностите на теглата b_0, b_1 и b_2 така, че тази равнина да е възможно най-близка до действителните отговори, като същевременно дава минимален SSR.

Случаят с повече от две независими променливи е подобен, но по-общ. Оценената регресионна функция е $(x_1, ..., x_r) = b_0 + b_1 x_1 + \cdots + b_r x_r$ и има $r + 1$ тегла, които трябва да се определят, когато броят на входовете е r.

5. Полиномна регресия

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Можете да разглеждате полиномната регресия като обобщен случай на линейната регресия. Предполага се полиномна зависимост между изхода и входовете и съответно полиномна оценена регресионна функция.

С други думи, в допълнение към линейните членове като b_1x_1, вашата регресионна функция f може да включва нелинейни членове като $b_2{x_1}^2$, $b_3{x_1}^3$ или дори $b_4x_1x_2$, $b_5{x_1}^2x_2$.

Най-простият пример за полиномна регресия има една независима променлива, а оценената регресионна функция е полином от втора степен: $(x) = b_0 + b_1x + b_2x^2$.

Сега си спомнете, че искате да изчислите b_0, b_1 и b_2, за да минимизирате SSR. Това са вашите неизвестни!

Като имате предвид това, сравнете предишната регресионна функция с функцията $(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$, използвана за линейна регресия. Те изглеждат много сходни и двете са линейни функции на неизвестните b_0, b_1 и b_2. Ето защо можете да решите задачата за полиномна регресия като линейна задача, като членът x^2 се разглежда като входна променлива.

В случая на две променливи и полином от втора степен регресионната функция има този вид: $(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 + b_3{x_1}^2 + b_4x_1x_2 + b_5{x_2}^2$.

Процедурата за решаване на задачата е идентична с тази в предишния случай. Прилага се линейна регресия за пет входни данни: x_1, x_2, ${x_1}^2$, x_1x_2 и ${x_2}^2$. В резултат на регресията получавате стойностите на шестте тежести, които минимизират SSR: b_0, b_1, b_2, b_3, b_4 и b_5.

Разбира се, има и по-общи проблеми, но това би трябвало да е достатъчно, за да илюстрира въпроса.

Недостатъчно и прекомерно приспособяване

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Един много важен въпрос, който може да възникне при прилагането на полиномна регресия, е свързан с избора на оптималната степен на полиномната регресионна функция.

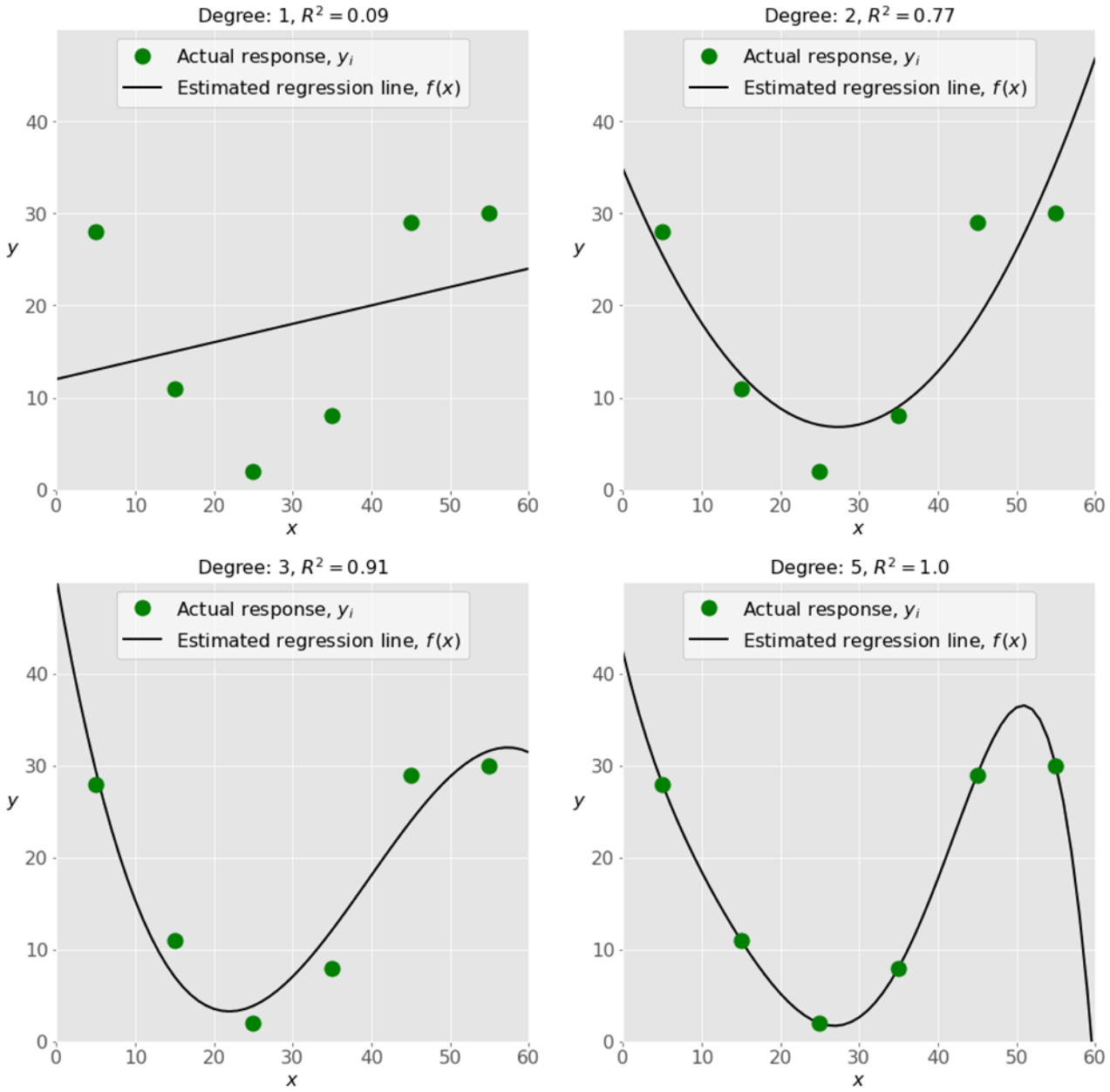
Няма ясно правило за това. То зависи от конкретния случай. Трябва обаче да сте наясно с два проблема, които могат да последват избора на степента: недостатъчно приспособяване и прекомерно приспособяване.

Недостатъчно приспособяване се получава, когато моделът не може да улови точно зависимостите между данните, обикновено като следствие от собствената си простота. То често дава нисък RI при известни данни и лоши възможности за обобщаване, когато се прилага с нови данни.

Прекаленото приспособяване се случва, когато моделът изучава както зависимостите от данните, така и случайните колебания. С други думи, моделът научава твърде добре съществуващите данни. Сложните модели, които имат много характеристики или термини, често са склонни към overfitting. Когато се прилагат към познати данни, такива модели обикновено дават висок RI. Често обаче те не обобщават добре и имат значително по-нисък RI, когато се използват с нови данни.

Следващата фигура илюстрира недостатъчно добре, добре и прекалено добре пригодените модели:

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.



Графиката в горния ляв ъгъл показва линия на линейна регресия, която има ниско R^2. Може също така да е важно, че правата линия не може да вземе

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

предвид факта, че действителният отговор нараства, когато x се отдалечава от двадесет и пет към нула. Това вероятно е пример за недостатъчно оборудване.

Графиката в горния десен ъгъл илюстрира полиномна регресия със степен, равна на две. В този случай това може да е оптималната степен за моделиране на тези данни. Моделът има стойност R^2, която е задоволителна в много случаи и показва добре тенденциите.

Графиката в долния ляв ъгъл представя полиномна регресия със степен, равна на три. Стойността на R^2 е по-висока, отколкото в предходните случаи. Този модел се държи по-добре с известни данни от предишните. Въпреки това показва някои признаци на пренастройване, особено за входните стойности, близки до шестдесет, където линията започва да намалява, въпреки че действителните данни не показват това.

И накрая, на диаграмата в долния десен ъгъл можете да видите идеалното прилягане: шест точки и линията на полинома на степен пет (или по-висока) добив $R^2 = 1$. Всеки действителен отговор се равнява на съответната си прогноза.

В някои ситуации това може да е точно това, което търсите. В много случаи обаче това е прекалено вграден модел. Вероятно е да има лошо поведение с невидяни данни, особено с входове, по-големи от петдесет.

Например, той приема, без никакви доказателства, че има значителен спад в отговорите за x над петдесет и че y достига нула за x близо до шестдесет. Подобно поведение е следствие от прекомерни усилия за изучаване и адаптиране на съществуващите данни.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.