Занятие девятое. Линейная и нелинейная регрессия. Причины и следствия

Постановка задачи

Рассмотрим набор точек (X_i, Y_i), где мы ожидаем, что Y — зависимая переменная, зависящая от X, которые называют предикторами. Задача регрессионного анализа — оценить функцию $f(x) = E(Y|X = x)$.

1) Простая линейная регрессия.

Рассмотрим случай одномерных предикторов и линейной функции $f(x) = b_0 + b_1 x$

$$Y_i = b_0 + b_1 X_i + \varepsilon_i,$$

где X_i — случайные или детерминированные величины, $E(\varepsilon_i|X_i) = 0$, $D(\varepsilon_i|X_i) = \sigma^2$ — константа, а ε_i независимы при условии X.

В таком случае нам требуется построить оценки $\hat{b}_0$ и $\hat{b}_1$ для b_0 и b_1. При каждом X_i при этом мы получаем остаток $r_i = Y_i - \hat{b}_0 - \hat{b}_1 X_i$.

Остаточной суммой квадратов или RSS называют $\sum_{i=1}^n r_i^2$.

Оценкой наименьших квадратов называют $\hat{b}_0$, $\hat{b}_1$, минимизирующие RSS.

Вопрос 1. Показать, что в нашем случае эта оценка имеет вид

$$\hat{b}_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2},\ \hat{b}_0 = \overline{Y} - \overline{X}\hat{b}_1.$$

Вопрос 2. Показать, что это ОМП в случае $\varepsilon_i|X_i \sim \mathcal{N}(0, \sigma^2)$ и найти в этом случае ОМП для σ^2.

Более принятой чем ОМП является несмещенная оценка

$$\hat{\sigma}^2 = \frac{RSS}{n-2}.$$

Оценки $\hat{b}_0$, $\hat{b}_1$ являются несмещеными, состоятельными и асимптотически нормальными с ковариационной матрицей (при условии $X_1, ..., X_n$)

$$\frac{\sigma^2}{nS_X^2}\begin{pmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix},$$

откуда нетрудно построить асимптотические доверительные интервалы для b_0, b_1. Вместе с доверительным интервалом мы получаем критерий Вальда для проверки гипотез $H_{0,i} : b_i = 0$ с общей альтернативой.

Построив линейную модель для аппроксимации данных, мы можем строить прогноз значения $Y^* = f(X^*)$ в точке X^*, то есть $\hat{Y}^* = \hat{b}_0 + \hat{b}_1 X^*$. При этом $E(\hat{Y}^*|X) = Y^*$,

$$D = D(\hat{Y}^*|X) = D\hat{b}_0 + (X^*)^2 D\hat{b}_1 + 2X^* cov(\hat{b}_0, \hat{b}_1) = \sigma^2\left(\frac{\sum_{i=1}^n (X_i - X^*)^2}{n\sum_{i=1}^n (X_i - \overline{X})^2} + 1\right).$$

Отсюда мы можем построить доверительный интервал для $Y^* = (\hat{Y}^* - z_{1-\alpha/2}\hat{D}, \hat{Y}^* + z_{1-\alpha/2}\hat{D})$, где $\hat{D}$ получается из D подстановкой $\hat{\sigma}^2$ вместо σ^2.

В R построение регрессионной модели, как мы уже видели, осуществляется с помощью $lm(x \sim y + z + w, data)$, где $data$ — данные из которых берутся столбцы y, z, w. Параметр subset позволяет строить модель на основе подмножества предикторов

У полученного объекта \$coefficients содержит полученнные коэффициенты, построить регрессионную прямую можно с помощью функции abline. Применяя к lm plot, мы получим несколько графиков, описывающих поведение остатков $r(X_i)$ (residuals) и их отклонение от нормального закона; применяя summary, получим информацию о модели, в частности, оценки, стандартные отклонения и фактический уровень значимости критерия Вальда. F-статистики, выдаваемая по модели, связана с проверкой гипотезы о соответствии линейной модели. Наконец, прогноз новых данных может быть сделан с помощью predict.

Показателем качества модели является

$$0 \le R^2 = 1 - \frac{(r,r)}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \le 1,$$

где близость R^2 к единице показывает качественное приближение.

К сожалению, линейная регрессионная модель, основанная на квадратичных отклонениях, крайне неустойчива к выбросам. Более устойчивую модель в одномерном случае представляет собой line пакета stats, основанный на подходе Тьюки, основанном на использовании медианы

Упражнение 1. Положим $x_i = i/10$, $i \le 100$, $x_{101} = 50$, $y_i \sim \mathcal{N}(x_i, 1)$, $i \le 100$, $y_{101} \sim \mathcal{N}(150, 1)$. Построить линейную регрессионную модель и модель Тьюки.

Упражнение 2. Исследовать линейную зависимость количества голосов, отданное за Бучанана от количество голосов, отданных за Буша, приведенных в PalmBeach пакета Stat2Data. Улучшится ли приближение, если предварительно прологарифмировать данные?

Последний пример показывает, что зачастую удобно изначально преобразовать часть переменных функциями, например, прологарифмировать. Одним из распространенных методов преобразований для положительной зависимой переменной является метод Box-Cox tranformation, преобразующий ее с помощью $t_\lambda(y) = (y^\lambda - 1)/\lambda$, $\lambda \ne 0$ или $\ln y$ при $\lambda = 0$, подбирая оптимальное y. В R она реализована boxcox пакета MASS.

Многомерная линейная регрессия

В случае, если X_i представляют собой векторы, мы рассматриваем $Y_i = b_1 X_{i,1} + ... + b_k X_{i,k} + \varepsilon_i$. Может показаться, что эта модель не включает в себя одномерную в силу отсутствия свободного члена, но в действительности X_i произвольны, поэтому можно рассматривать $X_{i,1} = 1$ и тем самым обеспечить наличие свободного члена. Будем считать, что $n > k$ (иначе размерность множества параметров больше числа наблюдений Y) и что матрица X имеет полный ранг (в противном случае один из столбцов выражается через остальные линейно и бесполезен в качестве параметра).

В этом случае аналогичным образом получаем

$$\hat\beta = (X^t X)^{-1} X^t Y,\ \hat\sigma^2 = \frac{1}{n-k} \sum_{i=1}^n (r,r),$$

где r как и прежде вектор остатков $r = Y - X\hat\beta$. $\hat\beta$ будет асимптотически нормальной оценкой β с условной (при условии X) матрицей ковариации $\sigma^2 (X^t X)^{-1}$, откуда мы можем построить доверительные интервалы для β_i вида $\hat\beta_i \pm \zeta_{1-\alpha/2} \hat\sigma_i$, где $\hat\sigma_i$ — диагональные элементы $(X^t X)^{-1}$, умноженные на $\hat\sigma^2$.

Аналогичным образом гипотезу $\beta_i = 0$ о том, что X_i не участвует в действии на Y, можно проверить на основе построенных интервалов.

Более мощный чем lm механизм предоставляет glm(), выдающая, среди прочего коэффициент AIC, описанный ниже.

Полезно также посмотреть на результаты применения к glm знакомой нам функции anova. Столбец Resid.Dev будет показывать дисперсию остатков по мере добавления переменной. Resid.Dev показывает дисперсию остатков, если мы аппроксимируем данные прямой, а в следующих ячейках — на сколько убывает дисперсия при добавлении новых переменных. Большое изменение дисперсии остатков показывает существенное влияние параметров.

Упражнение 3. В данных о преступности http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html содержится информация об уровне преступности в зависимости от различных факторов. Построить линейную регрессионную модель и сказать, какие параметры с вероятностью 95% имеют ненулевой коэффициент β.

Редуцируем пространство признаков

Зачастую большое количество предикторов только вносит дополнительный шум, переусложняя модель. Мы бы хотели выбрать какое-то подмножество S и делать прогноз на его основе. При этом маленькие S дадут нам большое смещение засчет недооценки, а большие S большую дисперсию засчет переоценки. Следовательно, нам понадобится мера измерения качества прогноза по подмножеству S.

Положим $\hat{Y}_i(S)$ — прогноз в точке X_i на основе регрессионной модели, построенной по S, $\hat{R}_{tr}(S) = \sum_{i=1}^{n}(\hat{Y}_i(S) - Y_i)^2$ — ошибка обучения, $R(S) = E\left(\sum_{i=1}^{n}(\hat{Y}_i(S) - Y_i^*)^2 | X_1, ..., X_n\right)$ — ошибка прогноза, где Y_i^* — независимые (при условии X_i) от Y_i значения наших зависимых переменных. Иначе говоря, ошибка прогноза — сумма средней квадратичной ошибки моего прогнозирования, если обучающий набор Y_i заменить на перегенерированный набор Y_i^*. $\hat{R}_{tr}(S)$, вообще говоря, смешенная оценка $R(S)$,

$$E(\hat{R}_{tr}(S)|X_1, ..., X_n) = R(S) - 2\sum_{i=1}^{n} cov(\hat{Y}_i(S), Y_i) < R(S),$$

где ковариация рассматривается при условии $X_1, ..., X_n$. С увеличением мощности S смещение $\hat{R}_{tr}(S)$ растет, поэтому рассматривают оценки, получающие дополнительный штраф за мощность S.
1) C_p-статистика Маллоу (Mallows) определяется как

$$\hat{R}(S) = \hat{R}_{tr}(S) + 2|S|\hat{\sigma}^2,$$

где $\hat{\sigma}^2$ — оценка дисперсии $RSS/(n-k)$ по всему множеству признаков. Минимизируя эту статистику по всем поднаборам S мы получим набор с достаточно хорошей ошибкой обучения и не слишком большим размером.
2) Перекрестная проверка (k-fold cross-validation).
В этом методе все предикторы разбиваются на l подмножеств T_i, при этом S разбиваются на S_i, и строится оценка

$$R_l = \frac{1}{l}\sum_{i=1}^{l}\sum_{j\in T_i}(Y_j - \hat{Y}_j(S \setminus S_i))^2.$$

Иначе говоря, мы поочередно исключаем всевозможные S_i из выборки и строим прогнозы для их элементов по всем остальным.
3) AIC (Akaike Information Criteria)
В случае, если ε_i заданы в параметрической модели, построим логарифмическую функцию правдоподобия (по всей выборке), найдем ее максимум l_S (т.е. возьмем ее значение в ОМП, построенной по подмножеству S) и в качестве оценки риска будем использовать $AIC = l_S - |S|$. Максимизируя его, мы получаем наиболее удачный набор
В случае нормальных данных AIC практически идентичен C_p.
4) BIC (Bayesian Information Criteria).
Этот подход вводит более жесткий штраф за размер выборки $BIC = l_S - |S|\ln n/2$, где n — размерность выборки. Соответственно, максимизируя BIC, мы будем получать более маленькие множества.

Итак, выбирая один из показателей, мы можем сравнивать по нему различные поднаборы признаков S. Остается выбрать план, как не перебирать все 2^k наборов предикторов. Наиболее простые подходы — прямой и обратный алгоритм. В случае прямого алгоритма мы начинаем с 0 признаков. Затем среди одноточечных множеств выбираем то, у которого наилучший показатель. Затем добавляем второй элемент и так далее. В случае обратного мы действуем начиная со всех k признаков.
На практике применяются обычно более сложные алгоритмы перебора подмножеств.

AIC и BIC содержатся в пакете stats, AIC(lm) подсчитывает коэффициент AIC, а AIC(lm, k= log(n)) — BIC. Оба коэффициента берутся с коэффициентом -2 по сравнению с указанным определением.

Упражнение 4. Применить к данным о преступности обратный алгоритм, основанный на AIC.

Существуют многочисленные модификации указанных подходов, в пакете bestglm функция bestglm предлагает довольно широкий спектр функций выбора наилучшего подмножества. Основной аргумент Xy работает с data.frame, чей последний столбец содержит зависимую переменную, параметр family позволяет задать семейство распределений, IC позволяет выбрать критерий (в частности, для больших размерностей более эффективен не BIC, а BICq). IC = "CV" будет работать с критериями близкими к 2, в частности, по умолчанию стоит $delete-d$ алгоритм, удаляющий поднаборы размера d и наблюдающий за ошибкой прогноза. Параметр method позволяет выбрать метод перебора подмножеств, в частности

"backward" или "forward" для упомянутых обратного и прямого алгоритма., а также "exhaustive" для более эффективного поиска.
У полученного объекта параметр BestModel содержит наилучшую из получившихся линейных регрессий.

Упражнение 5. Исследовать на том же массиве данных различные алгоритмы — AIC, BIC, BICq.

Логистическая регрессия

Предположим, что величины Y_i принимают только два значения 0 и 1, причем $P(Y_i = 1|X_1, ..., X_n)$ имеет вид

$$p_i = P(Y_i = 1) = F\left(\sum_{j=1}^{k} \beta_j X_{i,j}\right),\ F(x) = \frac{e^x}{1+e^x}.$$

то есть

$$\ln \frac{p_i}{1-p_i} = \sum_{j=1}^{k} \beta_j X_{i,j}$$

или отношение среднего числа 1 к среднему числу 0 линейно зависит от значений признаков. $F(x)$ называют логистической функцией, а полученную модель — логистической регрессией. Существует и другие виды, например, пробит регрессия, использующая нормальное распределение вместо логистического.

Для логистической регрессии подсчитывая правдоподобия

$$L(\beta_1, ..., \beta_k) = \prod_{i=1}^{n} p_i(\beta_1, ..., \beta_k)^{Y_i}(1 - p_i(\beta_1, ..., \beta_k))^{1-Y_i}$$

и находя ОМП $\hat{\beta}_1$,...,$\hat{\beta}_k$, мы получим коэффициенты логистической регрессии.

Построить логистическую регрессию в R можно с помощью glm, указав в качестве family "binomial". При прогнозировании, указав в predict параметр type='response, мы увидим не сами значения, которые разыгрываются случайно, а вероятности.

Упражнение 6. Исследовать данные о Титанике с помощью логистической регрессии и определить наиболее значимые переменные.

Нелинейная и непараметрическая регрессия

Нелинейную оценку методом наименьших квадратов можно получить подобно линейной. В этом случае

$$Y_i = f(X_i, \vec{\beta}) + \varepsilon_i,$$

где f — заданная функция, а $\vec{\beta}$ — оцениваемые параметры. Задача, опять же, сводится к задаче оптимизации и может быть решена численно. В R этот алгоритм задается тем же образом, что и линейная на основе функции nls. Формат запуска nls(x~ exp(y*c)-d*abs(z), data = Dat, start=list(c=0,d=1)), где x,y,z — столбцы Dat, а c,d — параметры модели (это указано в start).
Стоит обратить внимание, что nls не функционирует, если $\varepsilon_i = 0$.
Функция nlsLM пакета minpack.lm использует более эффективные способы численной минимизации, чем метод Ньютона, используемый в nls.

Обратите внимание на то, что, скажем, модели exp(y)~exp(a*x+b) и y~ax+b принциально отличаются, поскольку в первом случае ошибка ε прибавляется к экспоненте, а во втором случае к линейной функции.

Упражнение 7. Для данных с выборов в Палм-Бич, приведенных выше, исследовать нелинейную регрессионную модель log(y)~log(a x + b).

Другой задачей служит непараметрическая регрессия — подбор функции f(x) сравнительно общего вида, такой что $Y_i = f(X_i) + \varepsilon_i$.
smooth.spline строит аппроксимирующий сплайн (бесконечно-гладкую функцию, кусочно представляющую собой многочлен третьей степени), sm.regression пакета sm строит для f ядерную оценку (схоже

с тем, как строятся ядерные оценки для плотности). Подход principal curve также строит такого рода аппроксимирующую функцию
К сожалению, такого рода методы склонны к переобучению.

Причина и взаимосвязь

Предположим, что для ряда людей X означает был ли человек под наблюдением врача (0, если был, 1 — если нет), Y — показатель его здоровья. Предположим, что Y_i при $X_i = 0$ в среднем меньше, чем Y_i при $X_i = 1$. Означает ли это, что наблюдение врача плодотворно влияет на здоровье?
Ответ прост — предположим, что осмотр врача никак не влияет на людей, но люди сами выбирают ходить ли к врачу и те, которые выбирают ходить, изначально лучше следят за здоровьем и потому более здоровы. Тогда взаимосвязь X и Y имеется, а причинно-следственной связи X и Y нет.
Для описания математической модели свяжем с каждым человеком две величины его здоровья — C_0, если человек не будет ходить к врачу и C_1 — если будет, $Y_i = C_{X_i}$.
Назовем ассоциацией X, Y величину $A = E(Y|X = 1) - E(Y|X = 0)$, а эффектом причинности $\theta = E(_1) - E(C_0)$. Эти величины существенно отличаются, поскольку в ассоциации есть только $(C_X|X)$. Давайте, впрочем, самостоятельно выберем из популяции людей, которых подвергнуть осмотру, и сделаем это случайно (с $P(X = 0) > 0$, $P(X = 1) > 0$). Тогда любая $\theta = A$„ в частности

$$\hat{\theta} = \frac{\sum_{i:X_i=1} Y_i}{\sum_{i:X_i=1} 1} - \frac{\sum_{i:X_i=0} Y_i}{\sum_{i:X_i=0} 1}$$

будет состоятельной оценкой A. Действительно, ведь C не зависит от X, а значит $EC_1 = E(C_1|X = 1) = E(Y|X = 1)$, $EC_0 = E(_0|X = 0) = E(Y_0|X = 0)$.
Таким образом, в случайно организованном эксперименте взаимосвязь переменных уже будет означать их причинно-следственную связь.

Аналогичным образом обстоят дела, если X не бинарный, только в этом случае нам понадобится целая случайная функция $C(x)$, $Y = C(X)$. В независимом случае $\theta(x) = EC(x)$ будет совпадать с регрессионной функцией $f(x) = E(Y|X = x)$, а f мы умеем оценивать с помощью $\hat{f}(x)$.

Что же делать, если мы не можем организовать эксперимент так, чтобы Y был независим от X? Предположим, что мы сможем ввести другие переменные Z, такие что при фиксированном Z X и Y условно независимы, скажем, разобьем людей на несколько категорий одного пола, возраста, веса, жизненного уклада. Тогда можно ожидать, что нахождение под наблюдением случайно и не зависит от общего здоровья. Тогда мы можем оценивать $\theta(x)$ с помощью

$$\hat{\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x, Z_i).$$

Конечно, для утверждений такого рода (о причинности) необходимо быть уверенным, что наше разбиение на категории дает действительно независимые блоки.