Математические основы искусственного интеллекта.
Регрессионный анализ - II

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Март 2022

План лекции

В рамках теоретико-вероятностного подхода рассматриваем систему линейно зависимых случайных величин Y и X, распределения которых известны, и описываем связь между ними в виде уравнения линейной регрессии — по сути, находим условное математическое ожидание Y по X.

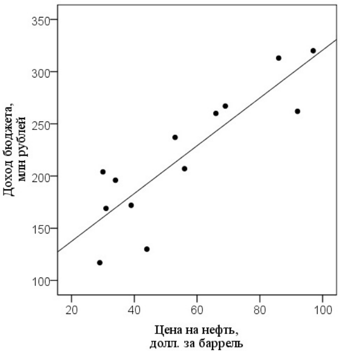
Рассматриваем систему линейно зависимых случайных величин Y и X, но распределения Y и X неизвестны, есть лишь набор из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, за X и Y. По этому набору строим выборочное уравнение линейной регрессии.

Строим множественную линейную регрессию, решаем проблемы связанные с увеличением числа предикторов.

Рассматриваем величины, связь которых близка к линейной, но таковой не является, например, $Y = X^{1.2} + \varepsilon$. Ситуация осложняется тем, что точный вид связи априори неизвестен. По набору из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, строим выборочное уравнение линейной регрессии, которая лишь упрощенно описывает нелинейную связь.

Даны наблюдения за двумя случайными величинами X и Y

X	30	69	86	56	44	97	53
Y	204	267	313	207	130	320	237



Требуется составить уравнение парной линейной регрессии:

$$\hat{y}(x) = b_1 x + b_0.$$

Если предикторов много

Очевидно, показатель, который необходимо спрогнозировать, может зависеть не от одного, а от многих факторов.

Изучается зависимость уровня артериального давления. В качестве возможных факторов, оказывающих влияние, выбраны:

1. возраст;
2. индекс массы тела;
3. индекс курения;
4. количество минут, затрачиваемых на занятия спортом в день;
5. уровень холестерина.

Включать ли факторы в модель?

Каждый фактор сам по себе лишь в незначительной степени может спрогнозировать уровень АД.

Включение в модель нескольких факторов позволяет более полно охарактеризовать пациента и, следовательно, дать более точный прогноз касательно его АД.

Стоит ли вводить все имеющиеся у исследователя факторы (независимые переменные) в модель, чтобы объяснить наблюдаемые значения зависимой величины? Например, должен ли уровень холестерина входить в модель как фактор, влияющий на уровень АД?

В общем случае ответ отрицательный — необоснованный ввод переменных в модель может ухудшить ее свойства.

Оценка точности модели

Коэффициент детерминации R^2 — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными.

Коэффициент детерминации R^2 — это единица минус доля необъясненной дисперсии (дисперсии случайной ошибки модели) в дисперсии зависимой переменной.

Коэффициент детерминации — универсальная мера зависимости одной случайной величины от множества других.

В частном случае линейной зависимости R^2 является квадратом множественного коэффициента корреляции между зависимой переменной и объясняющими переменными. В частности, для модели парной линейной регрессии R^2 равен квадрату обычного коэффициента корреляции между Y и X.

Коэффициент детерминации

Коэффициент детерминации модели зависимости случайной величины Y от фактора X определяется следующим образом:

$$R^2 = 1 - \frac{D[Y|X]}{D[Y]},$$

где $D[y]$ — дисперсия случайной величины Y, а $D[Y|X]$ — условная (по факторам X) дисперсия зависимой переменной (дисперсия ошибки модели).
В данном определении используются истинные параметры, характеризующие распределение случайных величин.

Условной дисперсией случайной величины Y относительно случайной величины X называется случайная величина

$$D[Y|X] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X] = \mathbb{E}[Y^2|X] - \mathbb{E}[Y|X]^2.$$

Выборочная оценка коэффициента детерминации

Если использовать выборочную оценку значений соответствующих дисперсий, то получим формулу для выборочного коэффициента детерминации:

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{SS_{res}/n}{SS_{tot}/n} = 1 - \frac{SS_{res}}{SS_{tot}},$$

где $SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ — сумма квадратов остатков регрессии, $y_i, \hat{y}_i$ — фактические и расчетные значения объясняемой переменной,

$SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$ — общая сумма квадратов,

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$

Чем больше — тем лучше?

Основная проблема применения (выборочного) R^2 заключается в том, что его значение увеличивается (не уменьшается) от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют.

Поэтому сравнение моделей с разным количеством факторов с помощью коэффициента детерминации, вообще говоря, некорректно. Для этих целей можно использовать альтернативные показатели.

Скорректированный коэффициент детерминации

Скорректированный коэффициент детерминации:

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-k)}{SS_{tot}/(n-1)} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)} \leqslant R^2,$$

где n — количество наблюдений, а k — количество параметров.

R^2_{adj} дает штраф за дополнительно включенные факторы

Включать ли факторы в модель?

При введении дополнительного фактора в модель доля объясненной дисперсии увеличивается, однако штрафные множители («плата» за количество факторов) тоже увеличиваются.

В результате R^2_{adj} увеличится лишь в том случае, если вновь вводимый фактор приводит к значительному росту доли объясненной дисперсии, т. е. может «объяснить» то, что не могли объяснить другие факторы.

Включать ли факторы в модель?

Суть метода пошагового отбора в следующем:

1. Рассчитывается матрица корреляций и выбирается фактор, имеющий наибольшую корреляцию с зависимой переменной.
2. К выбранному регрессору последовательно добавляются каждый из оставшихся регрессоров и вычисляются скорректированные коэффициенты детерминации для каждой из моделей. К модели присоединяется тот регрессор, который обеспечивает наибольшее значение R^2_{adj}.
3. Процесс присоединения регрессоров прекращается, когда значение R^2_{adj} становится меньше достигнутого на предыдущем шаге.

Информационные критерии

AIC — информационный критерий Акаике:

$$AIC = \frac{2k}{n} + \ln \frac{SS_{res}}{n},$$

где k — количество параметров модели.
Чем меньше значение AIC, тем модель лучше.

BIC — байесовский информационный критерий:

$$BIC = \frac{k \ln n}{n} + \ln \frac{SS_{res}}{n}.$$

BIC дает больший штраф за включение параметров в модель, чем AIC.

Модель должна быть достаточно точной, чтобы описывать закономерности, но достаточно грубой, чтобы отсекать случайные шумы.

Мультиколлинеарность

Мультиколлинеарность — наличие линейной зависимости между объясняющими переменными регрессионной модели.

Различают полную коллинеарность, которая означает наличие функциональной линейной зависимости, и частичную или просто мультиколлинеарность — наличие сильной корреляции между факторами.

Мультиколлинеарность: аналитический пример

Полная коллинеарность приводит к неопределенности параметров в линейной регрессионной модели независимо от методов оценки.

Рассмотрим пример линейной модели:

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon.$$

Пусть факторы этой модели тождественно связаны следующим образом: $x_1 = x_2 + x_3$. Тогда рассмотрим исходную линейную модель, в которой к первому коэффициенту добавим произвольное число a, а из двух других коэффициентов это же число вычтем.

Тогда имеем (без случайной ошибки):

$y = (b_1 + a)x_1 + (b_2 - a)x_2 + (b_3 - a)x_3 =$

$= b_1 x_1 + b_2 x_2 + b_3 x_3 + a(x_1 - x_2 - x_3) = b_1 x_1 + b_2 x_2 + b_3 x_3.$

Таким образом, несмотря на произвольное изменение коэффициентов модели, мы получили ту же модель.

Такая модель принципиально неидентифицируема.

Мультиколлинеарность: аналитический пример

Если рассмотреть трехмерное пространство коэффициентов, то в этом пространстве вектор истинных коэффициентов в данном случае не единственный, а представляет собой плоскость. Любая точка этой плоскости — истинный вектор коэффициентов.

Мультиколлинеарность: содержательный пример

Рассмотрим пример титрования дозы препарата у детей до 7 лет в зависимости от возраста и веса ребенка. Чем ребенок старше и чем он больше весит, тем большую дозу ему необходимо назначить.

Желая учесть оба фактора, исследователь включил (не используя методы пошагового отбора переменных!) их в модель и с помощью регрессионного анализа построил следующую модель:

$$\text{доза} = 2 \times \text{вес} + 1 \times \text{возраст}.$$

Пусть все величины в моделт обезразмерены. Коэффициенты при независимых величинах имеют содержательный смысл: при увеличении веса на 1 кг надо увеличивать дозу на 2 у. е., и с каждым следующим годом надо увеличивать дозу на 1 у. е.

Известно, что вес ребенка сильно коррелирует с его возрастом, а потому включение в модель и возраста, и веса как предикторов может (хотя это не обязательно) привести к неверной идентификации параметров.

Действительно, если вес $\approx$ возраст, то при незначительных изменениях в исходных данных программа могла построить модель вида:

$$\text{доза} = 4 \times \text{вес} - 1 \times \text{возраст}.$$

Интраоперационная кровопотеря является риском развития острого делирия в ближайшем послеоперационном периоде. Желая спрогнозировать риски и повысить точность оценок, исследователь строит модель и вводит в нее уровень гемоглобина, уровень гематокрита и число эритроцитов.

Проблема такой модели — мультиколлинеарность, так как все эти три фактора очень сильно коррелированы между собой. В результате оценки параметров окажутся неточными (или вовсе абсурдными), а качество прогноза сильно ухудшится.

Нелинейная связь величин

Для точного описания уравнения регрессии (необязательно линейной) требуется знать условный закон распределения случайной величины Y, чтобы найти ее условное математическое ожидание $M[Y|X = x]$.

В статистической практике такую информацию, как правило, не удается получить, а потому на основе наблюдаемых данных выбирают подходящую аппроксимацию функции $M[Y|X = x]$.

Связь между следующими тремя сущностями

1. истинная (вообще говоря, нелинейная) регрессия $f(x) = M[Y|X = x]$. Она была бы построена, если бы была известна плотность совместного распределения;
2. линейная аппроксимация функции регрессии $\tilde{f}(x) \approx f(x)$. Эта функция тоже могла быть построена, если бы была известна плотность совместного распределения;
3. оценка линейной аппроксимации функции регрессии $\hat{f}(x)$.

Истинная нелинейная регрессия $f(x)$

Пусть случайная величина Y связана с X равенством $Y = X^{1.2} + \varepsilon$, где X равномерно распределена на $[1, 3]$; ε — нормально распределенная случайная величина с нулевым математическим ожиданием и независящей от X дисперсией.
Истинная регрессия имеет вид: $y = f(x) = M[Y|X = x] = x^{1.2}$.

Линейная аппроксимация функции регрессии $\tilde{f}(x)$

График функции $y = f(x) = x^{1.2}$, $x \in [1,3]$, близок к линейному. Учитывая, что линейные модели предпочтительны в смысле своей простоты, разумным представляется построить аппроксимацию функции $f(x)$ в классе линейных:

$$\tilde{f}(x) = \beta_1 x + \beta_0.$$

Параметры выбираются исходя из приближения в той или иной норме, например,

$$\max_{x \in [1,3]} \left| x^{1.2} - \beta_1 x - \beta_0 \right| \xrightarrow[\beta_0,\, \beta_1]{} \min,$$

$$\int_1^3 (x^{1.2} - \beta_1 x - \beta_0)^2 \, dx \xrightarrow[\beta_0,\, \beta_1]{} \min .$$

Ошибка (систематическая), возникающая при этом, будет небольшой, а работать с функцией $\tilde{f}(x)$ может быть проще, чем с $f(x)$.

Оценка линейной аппроксимации функции регрессии $\hat{f}(x)$

В реальных задачах точный вид взаимосвязи случайных величин нам вообще неизвестен, имеется лишь конечный набор наблюдений за парой X, Y.

Расположение точек дает основание предположить линейную взаимосвязь X и Y и построить выборочное уравнение $\hat{f}(x) = b_1 x + b_0$, которое по вероятности будет сходиться к $\tilde{f}(x) = \beta_1 x + \beta_0$ при неограниченном увеличении объема выборки n.

Поскольку мы ошиблись в выборе класса функций, когда строили выборочное уравнение, получаемые оценки не будут состоятельными. То есть как бы мы ни увеличивали объем выборки, выборочная оценка $\hat{y}$ не будет сходиться к истинной функции регрессии $f(x)$.

Может ли коэффициент детерминации R^2 быть отрицательным? А может ли скорректированный коэффициент детерминации (*англ.* adjusted R^2) принимать отрицательные значения?