ТОМСКИЙ
ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

Статистическое моделирование и прогнозирование
Лекция 6. Регрессионный анализ

Семёнов Михаил Евгеньевич
к. ф.-м. н., доцент ОЭФ ИЯТШ

Томский политехнический университет
12 мая 2020 г.

Содержание

Зависимые и независимые переменные

Регрессионный анализ исследует и оценивает связь между *зависимой* или *объясняемой* переменной и *независимыми* или *объясняющими* переменными.

Зависимую переменную иногда называют *результативным признаком*, а объясняющие переменные — *предикторами, регрессорами* или *факторами*.

Обозначим зависимую переменную y, а независимые — $x_1, x_2, ..., x_k$.

- При $k = 1$ имеется только одна независимая переменная x и регрессия называется **парной**.
- При $k > 1$ имеется множество независимых переменных $x_1, x_2, \ldots, x_k$ и регрессия называется **множественной**.

Построение регрессионной модели

Рассмотрим построение простейшей регрессионной модели:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

где y — зависимая случайная переменная; x — независимая детерминированная переменная; β_0, β_1 — постоянные параметры уравнения; ε — случайная переменная, называемая также ошибкой.

Будем считать, что истинная зависимость между x и y — линейная, то есть существует некоторая зависимость $y = \beta_0 + \beta_1 x$.

Задача регрессионного анализа заключается в получении оценок коэффициентов β_0, β_1.

Для этого обычно применяют метод наименьших квадратов (МНК), а также метод максимального правдоподобия, метод наименьших модулей и.т.д.

Оценка параметров уравнения регрессии

Пусть имеется n наблюдений, тогда уравнение регрессии можно переписать в виде:

$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1,2,...,n \tag{2}$$

Будем рассматривать случайное слагаемое ε как последовательность n случайных величин: $e_1, e_2, ..., e_n$. Метод наименьших квадратов сводится к тому, чтобы получить такие оценки b_0, b_1 параметров β_0, β_1, при которых минимизируется сумма квадратов отклонений e_i фактических значений признака y_i от теоретических $\hat{y}_i = b_0 + b_1 x_i$:

$$Q_e(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \to min. \tag{3}$$

Оценка параметров уравнения регрессии

Для минимизации функции $Q_e(b_0, b_1)$ приравняем к нулю её частные производные $\frac{\partial Q_e}{\partial b_0}$ и $\frac{\partial Q_e}{\partial b_1}$:

$$\begin{cases} -2\sum_{i=1}^{n} y_i + 2nb_0 + 2b_1 \sum_{i=1}^{n} x_i = 0; \\ -2\sum_{i=1}^{n} y_i x_i + 2b_0 - 2\sum_{i=1}^{n} x_i + 2b_1 - 2\sum_{i=1}^{n} x_i^2 = 0. \end{cases}$$

После преобразований получим *систему нормальных уравнений МНК*:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i. \end{cases}$$

Оценка параметров уравнения регрессии

Решая систему нормальных уравнений, находим b_0, b_1:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ где } b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \qquad (4)$$

или в компактной форме: $b_0 = \bar{y} - b_1\bar{x}$, где $b_1 = \frac{\mathrm{cov}(x,y)}{s_x^2}$.

Коэффициент b_1 называется выборочным коэффициентом регрессии. Если независимую переменную x увеличить на единицу, то новое значение зависимой переменной $y(x+1)$ будет равно $y(x) + b_1$. Коэффициент b_0 численно равен значению результирующего признака y при нулевом значении фактора x.

Оценка качества выборочного уравнения регрессии

Уравнение выборочной регрессии имеет вид $y = b_0 + b_1 x$. Обозначим $\hat{y}_i = b_0 + b_1 x_i$ — *расчётное значение* зависимой переменной y, вычисленное при значении независимой переменной $x = x_i$.

Тогда $e_i = y_i - \hat{y}_i, i = 1,2,...,n$ — *остатки*, характеризующие отклонения наблюдаемых значений зависимой переменной от расчётных. Заметим, что полная сумма отклонений e_i будет равна нулю при любых выборочных значениях y_i и, следовательно, не может быть использована для оценки качества уравнения регрессии. Это свойство является одним из важнейших оптимизационных свойств МНК-оценок.

Оценка качества выборочного уравнения регрессии

В связи с этим при оценке качества выборочного уравнения регрессии используются следующие суммы квадратов отклонений:

$$Q_t = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ Q_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,\ Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \qquad (5)$$

где Q_t – общая сумма квадратов отклонений значений зависимой переменной от её выборочного среднего значения x_{ij}; Q_r — сумма квадратов отклонений расчетных значений зависимой переменной от ее выборочного среднего значения; Q_e -– сумма квадратов отклонений y_i, от линии регрессии, обычно называемая суммой квадратов остатков или ошибок.

Оценка качества выборочного уравнения регрессии

Величину $\sqrt{\frac{Q_e}{n-2}}$ *называют средней квадратической погрешностью* или *ошибкой* уравнения регрессии. Между приведёнными выше суммами квадратов существует связь: $Q_t = Q_r + Q_e$, которая и позволяет характеризовать качество построенного уравнения регрессии. Уравнение регрессии считается тем лучше, чем больше сумма квадратов, обусловленная регрессией Q_r, по сравнению с суммой квадратов остатков Q_e. В этом случае уравнение регрессии воспроизводит большую часть суммы квадратов отклонений зависимой переменной от её среднего значения и может быть использовано в практических приложениях.

Оценка качества выборочного уравнения регрессии

Для того чтобы формализовать это представление используется *коэффициент детерминации*:

$$R^2 = \frac{Q_r}{Q_t} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad R^2 \in [0,1] \tag{6}$$

причём, чем ближе коэффициент детерминации R^2 к единице, тем выше качество полученного уравнения регрессии. Максимальное значение коэффициента детерминации $R^2 = 1$ достигается в том случае, когда все остатки $e_i = 0$, а уравнение прямой регрессии проходит точно через все точки y_i.

Таким образом, значение коэффициента детерминации R^2 можно интерпретировать как долю общей дисперсии зависимой переменной y, которая будет объяснена (воспроизведена) с помощью уравнения регрессии.

Ограничения для случайной составляющей ε

Использование МНК накладывает ряд ограничений на поведение случайной составляющей ε уравнения регрессии: $y = \beta_0 + \beta_1 x + \varepsilon$. Обычно эти ограничения формулируются в следующем виде:

1. Математические ожидания всех случайных составляющих равны нулю: $M(\varepsilon_i) = 0$, где $i = 1,2,...,n$. Практически это условие означает, что случайная составляющая ε не вносит систематического смещения в значения зависимой переменной y;
2. Дисперсии всех случайных составляющих равны друг другу: $D(\varepsilon_i) = \sigma^2$, где $i = 1,2,...,n$. Практически это условие означает, что все наблюдаемые значения зависимой переменной y_i измерены с одинаковой точностью;

Ограничения для случайной составляющей ε

3. Различные случайные составляющие ε_i не коррелируют друг с другом: $cov(\varepsilon_i,\varepsilon_j) = 0$ при $i \neq j$, где $i,j = 1,2,...,n$. Практически это условие означает, что ошибки при различных наблюдениях независимы. Данное условие часто заменяют предположением о независимости распределения случайной составляющей ε_j и значений величины X, то есть $cov(x_i,\varepsilon_j) = 0$;
4. Случайные составляющие ε_i распределены по нормальному закону: $\varepsilon_i \sim N(0,\sigma)$, где $i = 1,2,...,n$. При выполнении этого условия уравнение регрессии называется нормальной (классической) линейной регрессионной моделью.

Условия 1–3 называют *условиями Гаусса–Маркова*, а соответствующая им теорема утверждает, что при выполнении данных условий МНК-оценки параметров уравнения регрессии будут *несмещёнными, состоятельными* и *эффективными*.

Множественный регрессионный анализ является развитием парного анализа в случае, когда зависимая переменная связана с более чем одной независимой переменной.

Модель парной регрессии даёт хороший результат в том случае, когда влиянием других факторов на объект исследования можно пренебречь.

Например, если коэффициент детерминации для построенного уравнения регрессии близок к единице: $R^2 \geqslant 0.8$. Однако в практических задачах такие ситуации являются скорее исключением, чем правилом. Поэтому модели множественной линейной регрессии имеют довольно широкое распространение.

Метод наименьших квадратов для множественной регрессии

Рассмотрим регрессионное уравнение, в котором определяется линейная связь зависимой переменой y от k независимых переменных $x_1, x_2,..., x_k$:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \tag{7}$$

Пусть проведено n наблюдений, в результате которых получены следующие эмпирические наборы данных:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i, i = 1,2,...,n. \tag{8}$$

Все использованные обозначения соответствуют по смыслу введённым ранее.

Метод наименьших квадратов для множественной регрессии

Основная задача будет заключаться в том, чтобы получить такие оценки b_i параметров β_i, где $i = 0,1,...,k$, при которых сумма квадратов отклонений e_i фактических значений признака y_i от расчётных $\hat{y}_i$ была бы минимальна:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1x_{1i} - ... - b_kx_{ki})^2 = \sum_{i=1}^{n} e_i^2 \to min \qquad (9)$$

Рассмотрим следующие векторы и матрицы:

$$\mathrm{Y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \quad \mathrm{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1k} \\ 1 & x_{21} & ... & x_{2k} \\ 1 & ... & ... & ... \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix}, \quad \mathrm{b} = \begin{pmatrix} b_1 \\ b_2 \\ ... \\ b_k \end{pmatrix}, \quad \mathrm{e} = \begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix}.$$

Столбцами матрицы X являются векторы $X_s = (x_{1s}, x_{2s},...,x_{ns})$, где $s = 0,1,...,k$, соответствующие независимым переменным $x_1, x_2,..., x_k$. Каждый элемент матрицы x_{ij} представляет собой результат i-го наблюдения для j-го признака, а первый единичный столбец соответствует значениям некоторой фиктивной переменной, используемой для большего удобства. Тогда система уравнений для определения оценок параметров линейной модели множественной регрессии $b_0, b_1,..., b_k$ в матричной форме примет вид:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}, \tag{10}$$

а подлежащая минимизации сумма квадратов отклонений

$$\sum_{i=1}^{n} e_i^2 = e^T e = (\mathbf{Y} - \mathbf{Xb})^T (\mathbf{Y} - \mathbf{Xb}) \to min \tag{11}$$

Оценка качества уравнения множественной регрессии

Как и в случае парной регрессии, качество полученного уравнения будем оценивать по той доли изменчивости зависимой переменной $\mathbf{Y}$, которая объясняется построенным уравнением. С учётом того, что $\varepsilon = \mathbf{Y}\hat{\mathbf{Y}}$ и $\varepsilon \perp \hat{\mathbf{Y}}$ запишем следующие равенства:

$$||\mathbf{Y}||^2 = \mathbf{Y}^T\mathbf{Y} = ((\mathbf{Y} - \hat{\mathbf{Y}}) + \hat{\mathbf{Y}})^T((\mathbf{Y} - \hat{\mathbf{Y}}) + \hat{\mathbf{Y}}) = ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 + ||\hat{\mathbf{Y}}||^2. \qquad (12)$$

Полученное разложение суммы квадратов в общем случае будет иметь вид: $Q_t = Q_e + Q_r$, где $Q_t = \mathbf{Y}\mathbf{Y}^T - n\bar{\mathbf{Y}}^2$ – общая сумма квадратов отклонений $\mathbf{Y}$ относительно среднего ; $Q_e = \mathbf{Y}\mathbf{Y}^T - \mathbf{b}^T\mathbf{X}^T\mathbf{Y}$ – сумма квадратов отклонений $\mathbf{Y}$, относительно расчётных значений по уравнению регрессии ; $Q_r = \mathbf{b}^T\mathbf{X}^T\mathbf{Y} - n\bar{\mathbf{Y}}^2$ – сумма квадратов отклонений расчётных значений $\hat{\mathbf{Y}}$ относительно среднего $\bar{\mathbf{Y}}$ или остаточная сумма квадратов.

Тогда коэффициент детерминации R^2, определяется так же, как и в случае парной регрессии:

$$R^2 = \frac{Q_r}{Q_t} = \frac{\mathbf{b}^T\mathbf{X}^T\mathbf{Y} - n\bar{\mathbf{Y}}^2}{\mathbf{Y}\mathbf{Y}^T - n\bar{\mathbf{Y}}^2} \tag{13}$$

Свойства коэффициента детерминации R^2 аналогичны сформулированным ранее.

Коэффициент R^2 показывает качество подгонки регрессионной модели к наблюдаемым значениям $\mathbf{Y}$.

Замечание: Величину $R = \sqrt{R^2}$ в случае множественной регрессионной модели называют ещё и коэффициентом множественной корреляции. Такой подход позволяет обобщить и распространить понятие связи на совокупности переменных.

Оценка качества уравнения множественной регрессии

Недостатком коэффициента детерминации R^2, ограничивающим его применение, является то, что при добавлении новых независимых переменных его значение всегда возрастает, хотя это и не означает улучшения качества модели как таковой. Чтобы избежать этой ситуации предлагается использовать коэффициент детерминации R_a^2, скорректированный по числу степеней свободы:

$$R_a^2 = 1 - \frac{(n-1)(1-R^2)}{(n-k-1)} = \frac{(n-1)e^T e}{(n-k-1)\mathbf{Y}^T\mathbf{Y}} \tag{14}$$

В отличие от R^2 при введении в модель новых независимых переменных скорректированный коэффициент R_a^2 может уменьшаться в том случае, когда эти переменные не оказывают существенного влияния на зависимую переменную.

Модель множественной линейной регрессии является, безусловно, весьма полезной и широко применяемой для прогнозирования количественного отклика. Считается, что наиболее эффективный путь улучшения качества регрессии – исключение незначимых коэффициентов, или, выражаясь точнее, отбор *информативного комплекса*.
Причины, по которым стоит проводить селекцию "оптимального подмножества"предикторов, Дж. Фаравэй (Faraway, 2006) видит в следующем:

1. Принцип бритвы Оккама утверждает, что из нескольких вероятных объяснений явления лучшим является самое простое. Компактная модель, из которой удалены избыточные предикторы, лучше объясняет имеющиеся данные.

2. Ненужные предикторы добавляют шум к оценке влияния других интересующих нас факторов. Иначе степени свободы (часто ограниченные) будут тратиться впустую.
3. При наличии коллинеарности некоторые переменные будут "пытаться сделать одну и ту же работу"(т.е., повторно объяснять вариацию значений зависимой переменной).
4. Если модель используется для прогнозирования, то можно сэкономить время и/или деньги, не измеряя избыточные переменные.

Алгоритмы выбора оптимального подмножества обычно основаны на последовательном "переборном"процессе, при котором многократно создаются модели с различными наборами предикторов, лучшая из которых определяется по некоторому критерию эффективности.

Постановка заданий

1 Для векторов выборочных данных:

x	2,36	2,67	2,98	3,3	3,61	3,93	4,24	4,56	4,87	5,18	5,5
y	1,12	0,46	0,19	-0,27	-0,85	-0,79	-1,17	-1,88	-1,62	-1,25	-1,04

Построить линейную модель парной регрессии и проверить её качество.

Постановка заданий

2 Для векторов выборочных данных:

y	12,2	7,6	10,4	9,9	15,7	14	12,7	10,5	15,1	10,6	15,2	17,2
x_1	4795	6962	6571	4249	9540	3488	4888	6237	2997	2990	1748	2128
x_2	69	82	87	92	23	31	55	81	65	98	100	69

Построить линейную модель множественной регрессии и проверить её качество.

3 Построить полную регрессионную модель и провести пошаговую процедуру включений с исключениями "слабых"предикторов. Сохранить скрипт в файле (*.r).

Алгоритм 1 Построение линейной модели парной регрессии и проверка ее качества

Вход: x, y – вектора выборочных данных.

```
1: x = c(2,36; 2,67; 2,98; 3,3; 3,61; 3,93; 4,24; 4,56; 4,87; 5,18; 5,5)
2: y = c(1,12; 0,46; 0,19; -0,27; -0,85; -0,79; -1,17; -1,88; -1,62; -1,25; -1,04)
3: fit = lm(y ~ x)
4: summary(fit)
```

Выход: Сводка основных результатов расчета

Результат


```
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-0.7473 -0.2662 -0.1076  0.2487  0.8165
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3786     0.5900   4.032 0.002965 **
x            -0.7700     0.1456  -5.287 0.000502 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4799 on 9 degrees of freedom
Multiple R-squared: 0.7565,        Adjusted R-squared: 0.7294
F-statistic: 27.96 on 1 and 9 DF,  p-value: 0.0005022
```

Рисунок 1 – Результат выполнения алгоритма №1

Алгоритм 2 Построение линейной модели множественной регрессии и проверка ее качества

Вход: y, x_1, x_2 – вектора выборочных данных.

1: y = c(12,2; 7,6; 10,4; 9,9; 15,7; 14; 12,7; 10,5; 15,1; 10,6; 15,2; 17,2)
2: x_1 = c(4795; 6962; 6571; 4249; 9540; 3488; 4888; 6237; 2997; 2990; 1748; 2128)
3: x_2 = c(69; 82; 87; 92; 23; 31; 55; 81; 65; 98; 100; 69)
4: fit = lm(y $\sim$ x_1+x_2)
5: summary(fit)

Выход: Сводка основных результатов расчета.

Результат


```
Call:
lm(formula = y ~ x1 + x2)
Residuals:
    Min      1Q  Median      3Q     Max
-2.9490 -1.1543 -0.2731  1.0857  2.8351
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.3218043  2.9415849   7.588 3.37e-05 ***
x1          -0.0007869  0.0003039  -2.590   0.0292 *
x2          -0.0847747  0.0281108  -3.016   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.115 on 9 degrees of freedom
Multiple R-squared: 0.5596,        Adjusted R-squared: 0.4617
F-statistic: 5.717 on 2 and 9 DF,  p-value: 0.02497
```

Рисунок 2 – Результат выполнения алгоритма №2

Алгоритм 3 Процедура исключения "слабых"предикторов

Вход: *fit* – сводка результатов расчета регрессии.
1: fit.step = step(fit, trace = 0)
2: summary (fit.step)
Выход: Сводка основных результатов расчета

Результат


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.7204080 20.5365454   2.713  0.00727 **
sizemedium   3.3401764  3.3721698   0.991  0.32316
sizesmall   10.9864235  3.7636536   2.919  0.00393 **
mxPH        -3.7829230  2.4260402  -1.559  0.12056
NO3         -1.5247600  0.4931425  -3.092  0.00228 **
NH4          0.0014816  0.0009336   1.587  0.11416
PO4         -0.0691868  0.0102496  -6.750 1.68e-10 ***
---
Residual standard error: 17.49 on 193 degrees of freedom
Multiple R-squared: 0.3494,     Adjusted R-squared: 0.3292
F-statistic: 17.27 on 6 and 193 DF,  p-value: 5.882e-16
```

Рисунок 3 – Результат выполнения алгоритма №3

Буховец, , Москалев, , Богатова, , and Бирючинская, (2010).
Статистический анализ данных в системе R. Учебное пособие.
ВГАУ, Воронеж.

Шитиков, and Мастицкий, (2017).
Классификация, регрессия и другие алгоритмы Data Mining с использованием R.