Методы современной прикладной статистики
8. Регрессионный анализ.

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Пусть есть n объектов, на которых мы наблюдаем значения признаков (объясняющих переменных, предикторов) $X_1, \dots, X_k$ и отклика (зависимой переменной) Y.

Хотим найти такую функцию $f(x_1, \dots, x_k)$, что на всех объектах $Y \approx f(X_1, \dots, X_k)$. По теореме о наилучшем квадратичном прогнозе,

$$arg\,\min_f \mathrm{E}(Y - f(X_1, \dots, X_n))^2 = E(Y | X_1, \dots, X_n),$$

однако на практике найти оценку $E(Y|X_1, \dots, X_n)$ по n наблюдениям отклика Y и признаков $X_1, \dots, X_k$ оказывается не так просто.

Поэтому рассмотрим сначала модель линейной регрессии

$$E(Y|X_1, \dots, X_n) = \beta_0 + \sum_j \beta_j X_j.$$

Обозначения

Обозначим $Y = (y_1, \dots, y_n)^T$, $X = \|x_{ij}\|, 1 \le i \le n$, $0 \le j \le k$, – наблюдения отклика Y и признаков $(X_0, X_1, \dots, X_k)$ соответственно на n объектах, причем $X_0 \equiv 1$, т.е. $x_{i0} = 1$.

Тогда регрессионная задача переформулируется так: найти по Y и X оценку вектора параметров $\beta = (\beta_0, \dots, \beta_k)^T$ в модели

$$Y = X\beta + \varepsilon,$$

где ε – вектор ошибок с нулевым средним.

Метод наименьших квадратов

Наиболее популярный метод решения регрессионной задачи - это метод наименьших квадратов, в котором оценка есть

$$\widehat{\beta} = arg \min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \beta_j x_{ij} \right)^2 .$$

Из курса статистики мы знаем, что $\widehat{\beta} = (X^T X)^{-1} X^T Y$, а оценка вектора откликов Y равна $\widehat{Y} = X\widehat{\beta}$.

Предположения Гаусса-Маркова

1) Линейность отклика: $Y = X\beta + \varepsilon$.

2) Случайность выборки: наблюдения $(x_{i1}, \ldots, x_{ik}, y_i)$ есть независимые реализации вектора $(X_1, \ldots, X_k, Y)$.

3) Полнота ранга: $rank(X) = k + 1$.

4) Случайность ошибки: $\mathsf{E}\varepsilon = 0$.

5) Гомоскедастичность: $\mathsf{D}\varepsilon = \sigma^2 I_n$, где $I_n = diag\{1, \ldots, 1\}$.

Теорема Гаусса-Маркова.
В предположениях 1)-5) оценка $\widehat{\beta}$ несмещенная и имеет наименьшую дисперсию среди несмещенных линейных по Y оценок параметра β.

Кроме того, в рамках предположений Гаусса-Маркова $\widehat{\beta}$ – состоятельная оценка вектора β и $\mathsf{D}\widehat{\beta} = \sigma^2 (X^T X)^{-1}$.

Понятно, что если признак является бинарным, т.е. принимает только 2 значения, то эти значения можно кодировать 0 и 1. Что делать, если признак принимает m значений?

Пусть y – средний уровень зарплаты служащих, а x – тип должности (рабочий, инженер или управляющий). Закодируем рабочих как 1, инженеров – 2, управляющих – 3, и построим регрессию $y = \beta_0 + \beta_1 x$. Тогда для указанных должностей получаем следующие модели:

$$y_{\text{раб.}} = \beta_0 + \beta_1, \quad y_{\text{инж.}} = \beta_0 + 2\beta_1, \quad y_{\text{упр.}} = \beta_0 + 3\beta_1.$$

Т.е., согласно нашей модели, разница в средней зарплате рабочего и инженера в точности равна разнице между зарплатами инженера и управляющего.

Категориальные признаки

Верный способ использования категориальных признаков в регрессии – введение бинарных фиктивных переменных (dummy variables). Пусть признак x_j принимает m различных значений, тогда для его кодирования необходима $m - 1$ фиктивная переменная.

Способы кодирования:

	Dummy		Deviation	
Должность	x_1	x_2	x_1	x_2
Рабочий	0	0	1	0
Инженер	1	0	0	1
Управляющий	0	1	-1	-1

Категориальные признаки

Рассмотрим новую модель с учетом таких кодировок категориального признака

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- При dummy-кодировании коэффициенты β_1, β_2 оценивают среднюю разницу в уровнях зарплат инженера и управляющего с рабочим.
- При deviation-кодировании коэффициенты β_1, β_2 оценивают среднюю разницу в уровнях зарплат рабочего и инженера со средним по всем должностям.

1. Все ли признаки значимы (т.е. проверить гипотезу $H_0 : \beta_j = 0$)?
2. Как предсказать значение отклика на новом объекте x_{n+1} (т.е. найти доверительный интервал для отклика)?
3. Как проверить адекватность (качество) полученной модели?

Далее будем работать в рамках гауссовской линейной модели, т.е. считаем, что $\varepsilon \sim N(0, \sigma^2 I_n)$. В этом случае $\widehat{\beta}$ – оптимальная оценка вектора параметров β.

Значимость признака

Проверим гипотезу $H_0 : \beta_j = 0$ о значимости j-того признака. Мы знаем, что $\widehat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$, отсюда $\forall c \in \mathbb{R}^k$ $c^T(\widehat{\beta} - \beta) \sim N(0, \sigma^2 c^T (X^T X)^{-1} c)$.

Оценкой для параметра σ^2 в линейной модели служит

$$\widehat{\sigma}^2 = \frac{1}{n-k-1} \|Y - X\widehat{\beta}\|^2 = \frac{1}{n-k-1} \sum_{i=1}^{k} \left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j \right)^2 .$$

Используя $\widehat{\sigma}^2$, получаем

$$\frac{c^T(\widehat{\beta} - \beta)}{\widehat{\sigma}\sqrt{c^T (X^T X)^{-1} c}} \sim St(n-k-1). \qquad (1)$$

Для проверки гипотезы $H_0 : \beta_j = 0$ выберем c_j – вектор с единицей на j-том месте и остальными нулями, тогда

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{a_{jj}}} \sim St(n-k-1),$$

где $a_{jj} = \left((X^T X)^{-1}\right)_{jj}$, откуда получаем доверительный интервал для j-того коэффициента линейной регрессии:

$$\left(\widehat{\beta}_j - t_{1-\frac{\alpha}{2}}\widehat{\sigma}\sqrt{a_{jj}}; \widehat{\beta}_j + t_{1-\frac{\alpha}{2}}\widehat{\sigma}\sqrt{a_{jj}}\right)$$

где $t_{1-\alpha/2}$ – $(1-\alpha/2)$-квантиль распределения $St(n-k-1)$. Таким образом, если 0 не попадает в этот доверительный интервал, то отвергаем гипотезу H_0 и признаем признак значимым для нашей модели.

Коэффициент детерминации

Прежде чем проверять гипотезу о значимости нескольких признаков, введем необходимые обозначения:

$$TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2 \text{ (Total Sum of Squares)};$$

$$ESS = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \text{ (Explained Sum of Squares)};$$

$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \text{ (Residual Sum of Squares)};$$

В случае, если мы пользуемся методом наименьших квадратов и один из признаков – константа, верна формула

$$TSS = ESS + RSS.$$

Кроме того, $\widehat{\sigma}^2 = RSS/(n-k-1)$.

Коэффициент детерминации

Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

По сути, R^2 – это квадрат коэффициента множественной корреляции (Пирсона) Y с X.

Логично, что чем меньше RSS, точнее, отношение $\frac{RSS}{TSS}$, тем лучше мы оценили Y_i и тем лучше качество нашей модели. Однако не всё так просто, как мы увидим далее.

Кроме того, $R^2 = 1 - \frac{RSS}{TSS}$ может оказаться отрицательным в случае модели без константного признака, что говорит о крайней неадекватности модели, потому что даже константа $(= \overline{y})$ приближает данные лучше.

Значимость нескольких признаков

Проверим гипотезу $H_0: \beta_1 = \ldots = \beta_m$, $m \leq k$, о значимости нескольких признаков. Пусть RSS_{H_0} построена по модели, в которой отсутствуют признаки $X_1, \ldots, X_m$, т.е. $\widehat{Y} = (X_{H_0}^T X_{H_0})^{-1} X_{H_0} Y$, где X_{H_0} – матрица признаков, откуда выбросили столбцы с номерами $1, \ldots, m$. Если H_0 верна, то

$$\frac{RSS_{H_0} - RSS}{RSS} \cdot \frac{n-k-1}{m} \sim F_{m,n-k-1}.$$

Этот критерий называется критерием Фишера. Для проверки гипотезы $H_0: \beta_1 = \ldots = \beta_k = 0$ (т.е. что все неконстантные признаки не влияют на модель) можно использовать такой критерий: при верной H_0

$$\frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \sim F_{k,n-k-1}.$$

Связь между критериями Фишера и Стьюдента

1) Если $m = 1$, критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы.

2) Иногда критерий Фишера отвергает гипотезу о незначимости признаков $\mathbb{X} = (X_1, \ldots, X_m)$, а критерий Стьюдента не признаёт значимым ни один из них. Возможные объяснения:

- отдельные признаки из $\mathbb{X}$ недостаточно хорошо объясняют Y, но совокупный эффект значим;
- признаки в $\mathbb{X}$ мультиколлинеарны.

Связь между критериями Фишера и Стьюдента

Иногда критерий Фишера не отвергает гипотезу о незначимости признаков $\mathbb{X}$, а критерий Стьюдента признаёт значимыми некоторые из них. Возможные объяснения:

- незначимые признаки в $\mathbb{X}$ маскируют влияние значимых;
- незначимость признаков $\mathbb{X}$ – результат множественной проверки гипотез.

Пример: для веса ребёнка при рождении имеется следующая модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon,$$

где *cigs* – среднее число сигарет, выкуривавшихся матерью за один день беременности, *parity* – номер ребёнка у матери, *inc* – среднемесячный доход семьи, *med* – длительность в годах получения образования матерью, *fed* – отцом.

Данные имеются для 1191 детей.

1) Зависит ли вес ребёнка при рождении от уровня образования родителей? Для проверки этого рассмотрим $H_0 : \beta_4 = \beta_5 = 0$ против альтернативы $H_1 : H_0$ *неверна*. В такой постановке задачи с помощью критерия Фишера получаем $p - value = 0.242$, т.е. два этих признака можно отбросить.

2) Имеет ли вообще смысл модель веса ребёнка при рождении, рассмотренная выше? Рассмотрим $H_0 : \beta_1 = \ldots = \beta_5 = 0$ против альтернативы $H_1 : H_0$ *неверна*. В этом случае p-значение критерия Фишера равно $6.033 \cdot 10^{-9}$, т.е. модель следует признать значимой.

ДИ для отклика на новом объекте

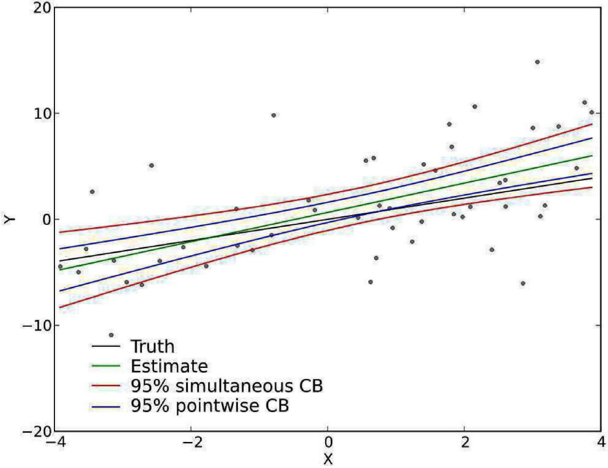
Значение отклика на новом наблюдении x_0 мы можем предсказать как $y(x_0) = x_0^T \beta + \varepsilon(x_0)$. Предположим, что $\varepsilon(x_0) \sim N(0, \sigma^2)$ – так же, как и предыдущие ошибки в нашей модели.

Для получения доверительного интервала для отклика $y(x_0)$ воспользуемся формулой (1) и подставим x_0 вместо вектора c. Получаем следующий доверительный интервал

$$\left(x_0^T \widehat{\beta} - t_{1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{D+1}; x_0^T \widehat{\beta} + t_{1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{D+1} \right),$$

где, как и ранее, $t_{1-\alpha/2}$ – $(1-\alpha/2)$-квантиль распределения $St(n-k-1)$, а $D = x_0^T (X^T X)^{-1} x_0$.

Если рассматривать полученные доверительные границы как функцию от x_0, то можно получить т.н. доверительную ленту для отклика.



Адекватность модели и feature selection

Видим (по графику, по критериям, о которых позже, по R^2), что модель плохо описывает данные. Возможные варианты действий:

1) Замена переменной: попробовать к признакам из матрицы X или к Y применить какие-нибудь функции. Возможные варианты функций: $\ln x$, $e^{\alpha x}$, x^{α} и так далее. При этом стоит проверить, что новые остатки ε_i останутся после замены переменной независимыми и, желательно, нормально распределенными.

2) Добавление переменных вида X_i^k. Может привести к переобучению, т.е. нестабильности решения на новых объектах.

3) Удаление переменных с помощью различных процедур.

Информационные критерии

Прежде чем переходить к методам добавления и удаления признаков, обсудим методы оценки качества модели.

Критерий Акаике

$$AIC = 2k + n\left(\ln \frac{RSS}{n} + 1\right).$$

Критерий Шварца

$$BIC = 2k \ln n + n\left(\ln \frac{RSS}{n} + 1\right).$$

Выбирается та модель, у которой AIC или BIC меньше. В случае, если сравниваем модели с разным количеством признаков, предпочтительно использовать BIC.

Информационные критерии не являются статистическими критериями проверки гипотез, это просто числовые показатели качества модели.

Сравнение невложенных моделей

Пусть имеются 2 невложенные модели, скажем,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \qquad (1)$$

$$y = \gamma_0 + \gamma_1 \ln x_1 + \gamma_2 \ln x_2 + \varepsilon, \qquad (2)$$

Как определить, какая из них лучше? Пусть $\widehat{y}$ – оценка отклика по первой модели, $\widetilde{y}$ – по второй. Подставим эти оценки как признаки в “чужие” модели:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \widetilde{y} + \varepsilon,$$

$$y = \gamma_0 + \gamma_1 \ln x_1 + \gamma_2 \ln x_2 + \gamma_3 \widehat{y} + \varepsilon,$$

Критерий Давидсона-Маккиннона

При помощи критерия Стьюдента проверим гипотезы

$$H_{01} : \beta_3 = 0, \quad H_{11} : \beta_3 \neq 0,$$

$$H_{02} : \gamma_3 = 0, \quad H_{12} : \gamma_3 \neq 0.$$

	H_{01} принята	H_{01} отвергнута
H_{02} принята	Обе модели хороши	Модель (2) значимо лучше
H_{02} отвергнута	Модель (1) значимо лучше	Обе модели плохи

Неправильное определение модели

1) **Недоопределение:** если зависимая переменная определяется моделью

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k + \varepsilon,$$

а вместо этого используется модель

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k + \varepsilon,$$

то МНК-оценки $\widehat{\beta}_0, \ldots, \widehat{\beta}_{j-1}, \widehat{\beta}_{j+1}, \ldots, \widehat{\beta}_k$ являются смещёнными и несостоятельными оценками $\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k$ соответственно.

2) **Переопределение:** если признак x_j не влияет на y, т.е. $\beta_j = 0$, то МНК-оценка $\widehat{\beta}$ остаётся несмещённой состоятельной оценкой вектора параметров β, но дисперсия её возрастает.

Критерий RESET Рэмси

Критерий в рамках линейной гауссовской модели проверяет предположение $E\varepsilon_i = 0 \ \forall i$. Это условие эквивалентно тому, что оценка отклика является несмещенной. С помощью этого критерия можно выявить:

1) наличие пропущенных, значимых для регрессии факторов;

2) неправильную функциональную форму регрессии;

3) наличие корреляции между факторами (мультиколлинеарность).

Правда, после отклонения гипотезы критерия нельзя с уверенностью сказать, что именно мы выявили.

Критерий RESET Рэмси

Пусть $\widehat{y}_i$ – оценки отклика в рассматриваемой модели линейной регрессии методом наименьших квадратов.

Рассмотрим вспомогательную модель

$$y_i = \sum_{j=0}^{k} \beta_j x_{ij} + \gamma(\widehat{y}_i)^2 + \eta_i, \; i = 1, \ldots, n,$$

где η_i – ошибки с нулевым матожиданием, и проверим гипотезу $H_0 : \gamma = 0$. Вычислим оценки откликов по новой модели (с помощью МНК): $\widetilde{y}_i$, посчитаем по ним RSS_{new}. Тогда при верной гипотезе H_0

$$F = \frac{RSS - RSS_{new}}{RSS/(n-k-1)} \approx F_{1,n-k-1}.$$

Отбор признаков (feature selection)

Для отбора признаков в линейной модели можно использовать следующие методы:

1) приведенный коэффициент детерминации;

2) одномерный отбор признаков;

3) пошаговая регрессия;

4) методы понижения размерности (например, метод главных компонент);

5) отбор на основе важности признаков (например, ExtraTreesClassifier или RandomForestClassifier).

Приведенный коэффициент детерминации

1) Стандартный коэффициент детерминации R^2 всегда увеличивается при добавлении новых признаков в модель, поэтому для отбора признаков его использовать нельзя.

Для сравнения моделей, содержащих разное число признаков, можно использовать приведённый коэффициент детерминации:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

Одномерный отбор признаков

Метод заключается в том, чтобы найти признаки, имеющие наиболее выраженную взаимосвязь с откликом, с помощью статистических тестов.

В частности, можно воспользоваться коэффициентами корреляции, критерием Стьюдента для проверки значимости линейной модели. В случае, если и отклик, и признак дикретны, можно воспользоваться критерием хи-квадрат для таблиц сопряженности.

С помощью функции SelectKBest (или руками) отбирается m признаков, имеющие наибольшую взаимосвязь с откликом (т.е. наименьшие p-значения в задаче проверки гипотез об отсутствии влияния).

Пошаговая регрессия

Выберем 2 пороговых уровня $p_{in} < p_{out}$, например, $p_{in} = 0.05$ и $p_{out} = 0.1$.

Шаг 0. Настраиваем модель только с константой, а также все модели с одной переменной. Выбирается модель с наименьшим уровнем значимости (пусть это модель с признаком X_{e1}), например, по критерию Фишера (см. слайд 14). Если этот уровень значимости меньше p_{in}, то включаем признак в модель.

Шаг 2N-1. Если есть признак, при добавлении которого критерий Фишера дает p-value $< p_{in}$, то добавляем в модель признак, на котором достигается наименьшее p-value.

Шаг 2N. Если есть признак, при удалении которого критерий Фишера дает p-value $> p_{out}$, то удаляем признак, на котором достигается наибольшее p-value.

Недостатки метода:

1) метод пошаговой регрессии не позволяет выводить оптимальные уравнения регрессии с точки зрения получения наибольшего коэффициента детерминации R^2 для данного количества признаков;

2) p-value зависят от результатов предшествующих тестов, что усложняет их интерпретацию;

3) метод несовместим с процедурами множественной проверки гипотез;

4) тесты являются смещенными, так как проводятся на одних и тех же данных.

Однако метод дает хорошие результаты в ситуации, когда количество признаков достаточно велико.

Finita!