Прикладная статистика. Регрессия II

Леонид Иосипой

Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

11 марта 2021

Повторение

Регрессионный анализ решает задачу выявления искаженной случайным «шумом» зависимости некоторого показателя Y от измеряемых переменных $X_1, \ldots, X_k$.

Повторение

Мы будем изучать линейную регрессию.

В линейной регрессии мы делаем предположение, что

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}, \quad i = 1, \ldots, n,$$

где

- $y_i, x_{i1}, \ldots, x_{ik}$ — отклик и значения k признаков для этого отклика (нам известные);
- $\beta_0, \beta_1, \ldots, \beta_k$ — константы, которые не зависят от номера отклика (нам неизвестные).

Задача состоит в том, чтобы оценить $\beta_0, \beta_1, \ldots, \beta_k$.

Повторение

Регрессионное равенство можно переписать в матричном виде как

$$y \approx X\beta,$$

где

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{10} = 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Здесь мы добавили в матрицу X единичный столбец, чтобы больше не думать про коэффициент β_0.

Повторение

Мы будем изучать свойства метода наименьших квадратов без использования каких-либо регуляризаторов:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_k x_{ik})^2 = \|y - X\beta\|^2 \to \min_{\beta}$$

Точное решение $\widehat{\beta}$ этой задачи известно и равно

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y.$$

Повторение

Чтобы исследовать качество решения метода наименьших квадратов, определим величину TSS (Total Sum of Squares) — разброс y относительно своего среднего:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Повторение

Оказывается, что (если в модель включен коэффициент β_0) TSS можно представить в виде суммы:

$$\mathrm{TSS} = \mathrm{RSS} + \mathrm{ESS},$$

- RSS (Residual Sum of Squares) — это сумма квадратов отклонений предсказанных y от их истинных значений:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

- ESS (Explained Sum of Squares) — это сумма квадратов отклонений среднего y от предсказанных y:

$$\mathrm{ESS} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

Повторение

По величинам RSS и ESS можно составить меру R^2, которая называется коэффициентом детерминации:

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

По сути, это доля объясненной дисперсии отклика во всей дисперсии отклика.

Повторение

Сделаем следующие предположения:

(П1) Истинная модель действительно является «зашумленной» линейной:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

для некоторых (неизвестных) коэффициентов $\beta_0, \ldots, \beta_k \in \mathbb{R}$ и некоторой случайной ошибки ε_i с $\mathbb{E}[\varepsilon_i] = 0$.

(П2) Наблюдения действительно случайны, то есть $(y_i, x_{i1}, \ldots, x_{ik})$ для $i = 1, \ldots, n$ образуют независимую выборку.

Повторение

(П3) Матрица X является матрицей полного (столбцового) ранга:

$$\operatorname{rank} X = k + 1.$$

То есть ни один из признаков не должен являться линейной комбинацией других. Поскольку среди столбцов есть константа, никакой из признаков в выборке не должен быть константой.

Повторение

Уже из этих трех предположений можно вывести, что оценки, получаемые методом наименьших квадратов, являются несмещенными и состоятельными:

$$\mathbb{E}\left[\widehat{\beta}_j\right] = \beta_j \quad \text{и} \quad \widehat{\beta}_j \xrightarrow{\mathbb{P}} \beta_j, \quad j = 0, \ldots, k.$$

Повторение

Более того, предположим еще что:

(П4) Ошибки $\varepsilon_1, \dots, \varepsilon_n$ имеют одинаковую дисперсию, которая не зависит от значений признаков (гомоскедастичность ошибок):

$$\mathrm{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n,$$

где $\sigma^2 > 0$ — неизвестный параметр.

Тогда можно показать, что дисперсия оценок, получаемых методом наименьших квадратов, является наименьшей в классе всех оценок, линейных по y (теорема Гаусса-Маркова).

То есть оценки метода наименьших квадратов являются в некотором смысле оптимальными.

Повторение

Рассмотрим еще одно предположение:

(П5) Ошибки $\varepsilon_1, \dots, \varepsilon_n$ имеют нормальное распределение

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Если выполняются (П1)-(П5), то оценки метода наименьших квадратов совпадают с оценками максимального правдоподобия.

Это означает, что оценки метода наименьших квадратов обладают всеми свойствами, которыми обладают оценки максимального правдоподобия.

Повторение

Более того, при выполнении (П1)-(П5) мы можем посчитать распределения всех случайных объектов в модели:

$$
\begin{aligned}
y &\sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right), \\
\widehat{\beta} &\sim \mathcal{N}\left(\beta, \sigma^2 (X^\top X)^{-1}\right), \\
\widehat{y} &\sim \mathcal{N}\left(X\beta, \sigma^2 X (X^\top X)^{-1} X^\top\right).
\end{aligned}
$$

Повторение

Эти факты позволяют нам построить следующие доверительные интервалы уровня доверия $1-\alpha$, $\alpha \in (0,1)$:

- для неизвестной дисперсии шума σ^2:

$$\mathbb{P}\left(\frac{\mathrm{RSS}}{c_{1-\alpha/2}} \leq \sigma^2 \leq \frac{\mathrm{RSS}}{c_{\alpha/2}}\right) = 1-\alpha,$$

  где c_α — квантиль уровня α распределения χ^2_{n-k-1}.

- для регрессионных коэффициентов $\beta_0, \ldots, \beta_k$:

$$\mathbb{P}\left(\widehat{\beta}_j - c_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}} \leq \beta_j \leq \widehat{\beta}_j + c_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}\right) = 1-\alpha,$$

  где $(X^\top X)^{-1}_{jj}$ — j,j элемент матрицы $(X^\top X)^{-1}$ и c_α — квантиль уровня α распределения T_{n-k-1}.

Повторение

Аналогично построению доверительных интервалов, можно проверить гипотезу о том, что признак j незначим, то есть что $\beta_j = 0$, $j = 0, \ldots, k$.

Критерий Стьюдента

$$\begin{aligned} \text{нулевая гипотеза:} \quad & H_0 : \ \beta_j = 0 \\ \text{альтернатива:} \quad & H_1 : \ \beta_j \neq 0 \text{ или } \beta_j > 0 \text{ или } \beta_j < 0 \\ \text{статистика:} \quad & T = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \\ \text{нулевое распределение:} \quad & T \sim T_{n-k-1} \end{aligned}$$

Повторение

Можно также проверить гипотезу о том, что сразу несколько коэффициентов β_j равны 0.

Критерий Фишера

нулевая гипотеза: $H_0: \ \beta_{j_1} = \ldots = \beta_{j_m} = 0$
для некоторых $0 \leq j_1 < \ldots < j_m \leq k$

альтернатива: $H_1: \ \beta_{j_1}, \ldots, \beta_{j_m} \neq 0$ одновременно

статистика: $F = \ldots$

нулевое распределение: $F \sim F_{m,n-k-1}$ — распределение Фишера

Повторение

Обратите внимание, что доверительные интервалы и критерии строятся в предположениях (П1)-(П5).

Если ошибки имеют разную дисперсию и/или распределены не нормально, то доверительные интервалы будут неверными!

Повторение

Есть несколько типичных ошибок, которые следует иметь в виду, применяя регрессионный анализ. Сами по себе они достаточно очевидны. Тем не менее, о них часто забывают при работе с реальными данными и в результате приходят к неверным выводам.

Существуют три вида лжи: ложь, наглая ложь и статистика.

(Марк Твен)

Повторение

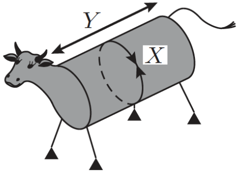
Пример

Рассмотрим в качестве отклика Z вес коровы, а в качестве предикторов — окружность ее туловища X и расстояние от хвоста до холки Y. Сравнительному анализу были подвергнуты три регрессионные модели:

(1) линейная: $Z = \theta_1 + \theta_2 X + \theta_3 Y$;

(2) степенная: $Z = \theta_1' X^{\theta_2'} Y^{\theta_3'}$;

(3) учитывающая содержательный смысл задачи $Z = \theta_0 X^2 Y$.



Повторение

Модель	По всем наблюдениям		По части наблюдений	
	$\widehat{\boldsymbol{\theta}}$	$\widehat{\sigma}$	$\widehat{\boldsymbol{\theta}}_{\text{тяж}}$	$\widehat{\sigma}_{\text{лег}}$
1	$\widehat{\theta}_1 = -984{,}7$ $\widehat{\theta}_2 = 4{,}73$ $\widehat{\theta}_3 = 4{,}70$	25,9	$\widehat{\theta}_1 = 453{,}2$ $\widehat{\theta}_2 = 0{,}62$ $\widehat{\theta}_3 = -0{,}22$	81
2	$\widehat{\theta}'_1 = 0{,}0011$ $\widehat{\theta}'_2 = 1{,}556$ $\widehat{\theta}'_3 = 1{,}018$	24,5	$\widehat{\theta}'_1 = 266{,}4$ $\widehat{\theta}'_2 = 0{,}203$ $\widehat{\theta}'_3 = -0{,}072$	79
3	$\widehat{\theta}_0 = 1{,}13 \cdot 10^{-4}$	26,6	$\widehat{\theta}_0 = 1{,}11 \cdot 10^{-4}$	28

Подробнее про пример можно посмотреть в книге Лагутина.

Спасибо за внимание!