Статистический анализ данных.MADE
7. Регрессионный анализ

Михаил Хальман
khalman.m@yandex.com

12 декабря 2019 г.

Постановка задачи линейной регрессии

$1, \dots, n$ — объекты;
$x_1, \dots, x_k, y$ — признаки, значения которых измеряются на объектах;
$x_1, \dots, x_k$ — объясняющие переменные (предикторы, регрессоры, факторы, признаки);
y — зависимая переменная, отклик.

Хотим найти такую функцию f, что $y \approx f(x_1, \dots, x_k)$;

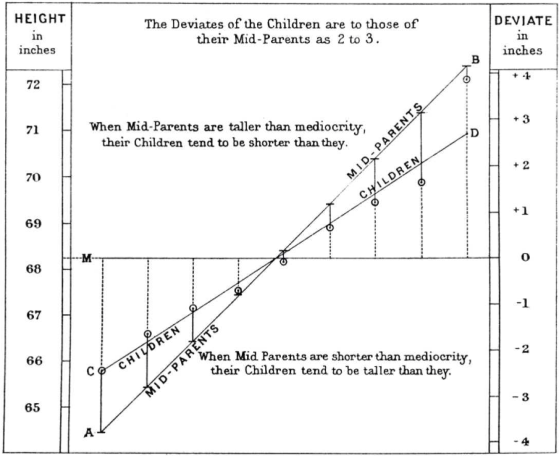
$$\underset{f}{\operatorname{argmin}}\, \mathbb{E}\left(y - f(x_1, \dots, x_k)\right)^2 = \mathbb{E}\left(y \,|\, x_1, \dots, x_k\right).$$

$\mathbb{E}\left(y \,|\, x_1, \dots, x_k\right) = f(x_1, \dots, x_k)$ — модель регрессии;
$\mathbb{E}\left(y \,|\, x_1, \dots, x_k\right) = \beta_0 + \sum\limits_{j=1}^{k} \beta_j x_j$ — модель линейной регрессии.

Здесь и далее $n > k$ ($n \gg k$).

Первое появление

Впервые такая постановка появляется в работе Гальтона 1885 г. «Регрессия к середине в наследственности роста».



$$y - \bar{y} \approx \frac{2}{3}\left(x - \bar{x}\right).$$

Метод наименьших квадратов (МНК)

Матричные обозначения:

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

Метод наименьших квадратов:

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

$$\|y - X\beta\|_2^2 \to \min_{\beta};$$

$$2X^T \left(y - X\beta \right) = 0,$$

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T y,$$

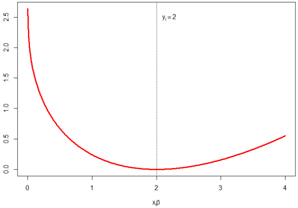
$$\hat{y} = X \left(X^T X \right)^{-1} X^T y.$$

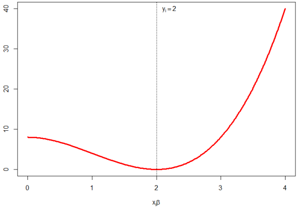
Метод наименьших квадратов (МНК)

МНК в линейной регрессии даёт выборочную оценку линейной аппроксимации условного матожидания $\mathbb{E}\left(y\,|x\right)$
Кроме $\|\cdot\|_2^2$ это делает любая дивергенция Брегмана:

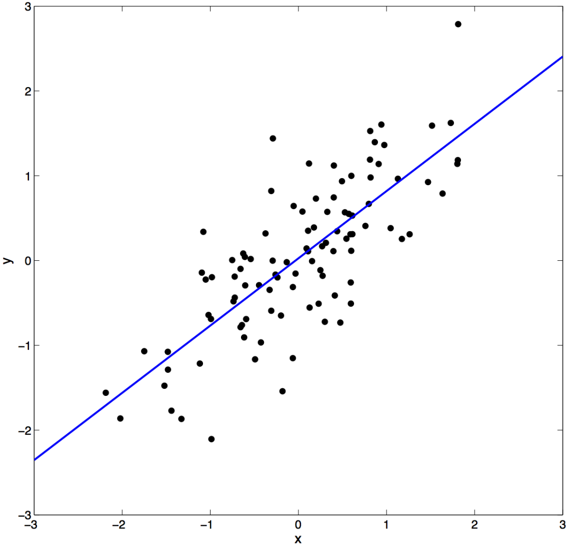
$$D\left(y, X\beta\right) = \sum_{i=1}^{n} \left(\phi\left(y_i\right) - \phi\left(x_i\beta\right) - \phi'\left(x_i\beta\right)\left(y_i - x_i\beta\right)\right),$$

где ϕ — произвольная непрерывно дифференцируемая выпуклая функция.

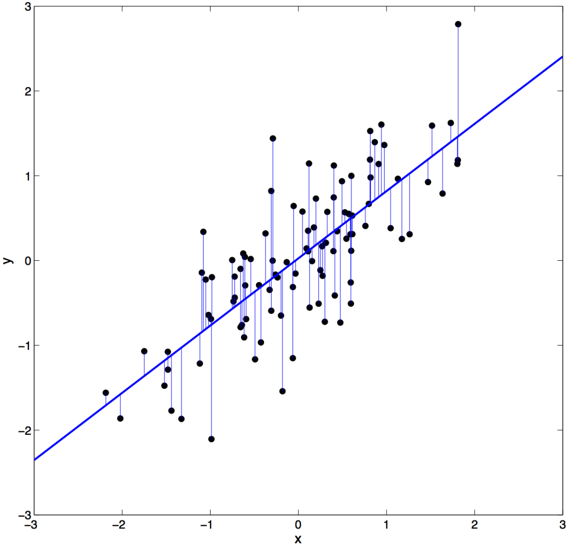




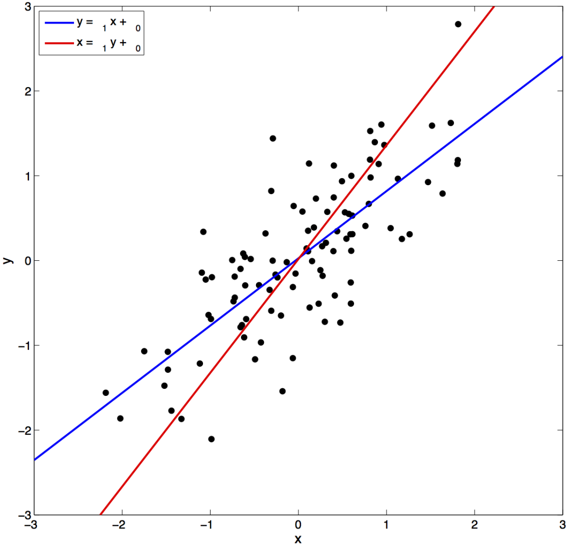
Инверсия задачи регрессии



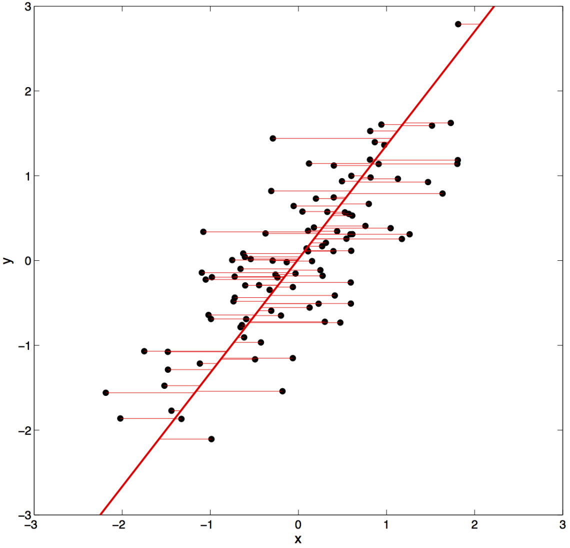
Инверсия задачи регрессии



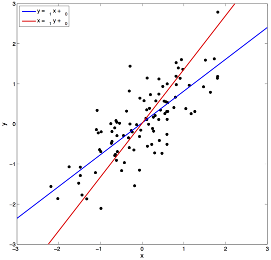
Инверсия задачи регрессии



Инверсия задачи регрессии



Инверсия задачи регрессии



- Две прямые пересекаются в точке $(\bar{x}, \bar{y})$.
- Выборочный коэффициент корреляции между x и y — среднее геометрическое $\hat{\theta}_1$ и $\hat{\phi}_1$:

$$\hat{\theta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}, \quad \hat{\phi}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2},$$

$$\hat{r}_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}.$$

Goodness-of-fit

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

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

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

$$TSS = ESS + RSS.$$

Коэффициент детерминации:

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

$R^2 = r^2_{y\hat{y}}$ — квадрат коэффициента множественной корреляции y с X.

Приведённый коэффициент детерминации

Стандартный коэффициент детерминации всегда увеличивается при добавлении регрессоров в модель, поэтому для отбора признаков его использовать нельзя.

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

$$R_a^2 = \frac{ESS/(n-k-1)}{TSS/(n-1)} = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1}.$$

Предположения модели

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \dots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке ($\operatorname{rank} X = k + 1$).
4. Случайность ошибок: $\mathbb{E}\left(\varepsilon \,|X\right) = 0$.

В предположениях (1-4) МНК-оценки коэффициентов β являются несмещёнными:

$$\mathbb{E}\hat{\beta}_j = \beta_j,\ j = 0, \dots, k,$$

и состоятельными:

$$\forall\gamma > 0 \lim_{n\to\infty} P\left(\left|\beta_j - \hat{\beta}_j\right| < \gamma\right) = 1,\ j = 0, \dots, k.$$

Предположения модели

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \dots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке ($\operatorname{rank} X = k + 1$).
4. Случайность ошибок: $\mathbb{E}\left(\varepsilon \,|\, X\right) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}\left(\varepsilon \,|\, X\right) = \sigma^2$.

(предположения Гаусса-Маркова).

Теорема Гаусса-Маркова: в предположениях (1-5) МНК-оценки имеют наименьшую дисперсию в классе оценок β, линейных по y.

Неправильное определение модели

Недоопределение: если зависимая переменная определяется моделью

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

а вместо этого используется модель

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

то МНК-оценки $\hat\beta_0, \dots, \hat\beta_{j-1}, \hat\beta_{j+1}, \dots, \hat\beta_k$ являются смещёнными и несостоятельными оценками $\beta_0, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k$.

Переопределение: если признак x_j не влияет на y, т. е. $\beta_j = 0$, то МНК-оценка $\hat\beta$ остаётся несмещённой состоятельной оценкой β, но дисперсия её возрастает.

Дисперсия $\hat{\beta}_j$

В предположениях (1-5) дисперсии МНК-оценок коэффициентов β задаются следующим образом:

$$\mathbb{D}\left(\hat{\beta}_j \middle| X\right) = \frac{\sigma^2}{TSS_j\left(1 - R_j^2\right)},$$

где $TSS_j = \sum\limits_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$, R_j^2 — коэффициент детерминации при регрессии x_j на все остальные признаки из X.

- Чем больше дисперсия ошибки σ^2, тем больше дисперсия оценки $\hat{\beta}_j$.
- Чем больше вариация значений признака x_j в выборке, тем меньше дисперсия оценки $\hat{\beta}_j$.
- Чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия оценки $\hat{\beta}_j$.

Дисперсия $\hat{\beta}_j$

$R_j^2 < 1$ по предположению (3); тем не менее, может быть $R_j^2 \approx 1$.

В матричном виде:

$$\mathbb{D}\left(\hat{\beta}\middle| X\right) = \sigma^2 \left(X^T X\right)^{-1}.$$

Если столбцы X почти линейно зависимы, то матрица $X^T X$ плохо обусловлена, и дисперсия оценок $\hat{\beta}_j$ велика.

Близкая к линейной зависимость между двумя или более признаками x_j называется **мультиколлинеарностью**.
Проблема мультиколлинеарности решается с помощью отбора признаков или использования регуляризаторов.

Бинарные признаки

Если x_j принимает только два значения, то они кодируются нулём и единицей. Например, если x_j — пол испытуемого, то можно задать $x_j = [\text{пол} = \text{мужской}]$.

Механизм построения регрессии не меняется.

Категориальные признаки

Как кодировать дискретные признаки x_j, принимающие более двух значений?
Пусть y — средний уровень заработной платы, x — тип должности (рабочий / инженер / управляющий). Допустим, мы закодировали эти должности следующим образом:

Тип должности	x
рабочий	1
инженер	2
управляющий	3

и построили регрессию $y = \beta_0 + \beta_1 x$. Тогда для рабочего, инженера и управляющего ожидаемые средние уровни заработной платы определяются следующим образом:

$$\begin{aligned} y_{bc} &= \beta_0 + \beta_1, \\ y_{pr} &= \beta_0 + 2\beta_1, \\ y_{wc} &= \beta_0 + 3\beta_1. \end{aligned}$$

Согласно построенной модели, разница в средних уровнях заработной платы рабочего и инженера в точности равна разнице между зарплатами инженера и управляющего.

Фиктивные переменные

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	
Тип должности	x_1	x_2	x_1	x_2
рабочий	0	0	1	0
инженер	1	0	0	1
управляющий	0	1	-1	-1

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

- 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.

Вопросы

1. Как найти доверительные интервалы для β_j и проверить гипотезу $H_0\colon \beta_j = 0$?
2. Как найти доверительный интервал для значений отклика на новом объекте $y(x_0)$?
3. Как проверить адекватность построенной модели?

Предположение о нормальности ошибок

6. Нормальность ошибок: $\varepsilon \,|X \sim N\left(0, \sigma^2\right)$.

Эквивалентная запись: $y \,|X \sim N\left(X\beta, \sigma^2\right)$.

- В предположениях (1-6) МНК-оценки совпадают с оценками максимального правдоподобия.

ММП:

$$p\left(\varepsilon_i\right) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma}\varepsilon_i^2},$$

$$\ln \prod_{i=1}^{n} p\left(\varepsilon_i\right) \to \max_{\beta},$$

$$\sum_{i=1}^{n} \left(-\frac{1}{2}\ln\left(2\pi\sigma\right) - \frac{1}{2\sigma}\varepsilon_i^2\right) \to \max_{\beta},$$

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

Предположение о нормальности ошибок

- МНК-оценки $\hat{\beta}$ имеют наименьшую дисперсию среди всех несмещённых оценок β.
- МНК-оценки $\hat{\beta}$ имеют нормальное распределение $N\left(\beta, \sigma^2\left(X^TX\right)^{-1}\right)$.
- Несмещённой оценкой σ^2 является

$$\hat{\sigma}^2 = \frac{1}{n-k-1} RSS;$$

  кроме того, $\frac{1}{\sigma^2} RSS \sim \chi^2_{n-k-1}$.
- $\forall c \in \mathbb{R}^{k+1}$

$$\frac{c^T\left(\beta - \hat{\beta}\right)}{\hat{\sigma}\sqrt{c^T\left(X^TX\right)^{-1}c}} \sim St(n-k-1).$$

Доверительные и предсказательные интервалы

$100(1-\alpha)\%$ доверительный интервал для σ:

$$\sqrt{\frac{RSS}{\chi^2_{n-k-1,1-\alpha/2}}} \leq \sigma \leq \sqrt{\frac{RSS}{\chi^2_{n-k-1,\alpha/2}}}.$$

Возьмём $c = \left(0 \ldots 0 \underset{j}{1} 0 \ldots 0\right)$; $100(1-\alpha)\%$ доверительный интервал для β_j:

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}.$$

Для нового объекта x_0 возьмём $c = x_0$; $100(1-\alpha)\%$ доверительный интервал для $\mathbb{E}\left(y \mid x = x_0\right) = x_0^T\beta$:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{x_0^T\left(X^TX\right)^{-1}x_0}.$$

Чтобы построить предсказательный интервал для $y\left(x_0\right) = x_0^T\beta + \varepsilon\left(x_0\right)$, учтём ещё дисперсию ошибки:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x_0^T\left(X^TX\right)^{-1}x_0}.$$

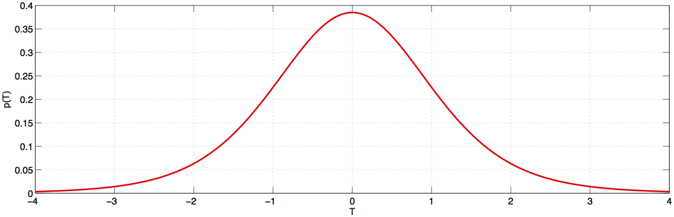
t-критерий Стьюдента

нулевая гипотеза: $H_0\colon \beta_j = 0$

альтернатива: $H_1\colon \beta_j <\neq> 0$

статистика: $T = \frac{\hat{\beta}_j}{\sqrt{\frac{RSS}{n-k-1}\left(X^TX\right)^{-1}_{jj}}}$

нулевое распределение: $St(n-k-1)$



t-критерий Стьюдента

Пример: имеется 12 испытуемых, x — результат прохождения испытуемым составного теста скорости реакции, y — результат его теста на симулятора транспортного средства. Проведение составного теста значительно проще и требует меньших затрат, поэтому ставится задача предсказания y по x, для чего строится линейная регрессия согласно модели

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Значима ли переменная x для предсказания y?

$H_0: \beta_1 = 0.$
$H_1: \beta_1 \neq 0 \Rightarrow p = 2.2021 \times 10^{-5}.$

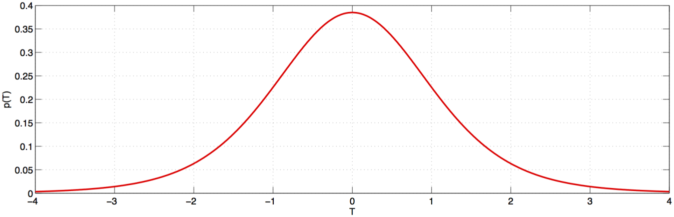
t-критерий Стьюдента

нулевая гипотеза: $H_0\colon \beta_j = a$

альтернатива: $H_1\colon \beta_j <\neq> a$

статистика: $T = \frac{\hat{\beta}_j - a}{\sqrt{\frac{RSS}{n-k-1}\left(X^TX\right)^{-1}_{jj}}}$

нулевое распределение: $St(n-k-1)$



t-критерий Стьюдента

Пример: по выборке из 506 жилых районов, расположенных в пригородах Бостона, строится модель средней цены на жильё следующего вида:

$$\ln price = \beta_0 + \beta_1 \ln nox + \beta_2 \ln dist + \beta_3 rooms + \beta_4 stratio + \varepsilon,$$

где nox — содержание в воздухе двуокиси азота, dis — взвешенное среднее расстояние от жилого района до пяти основных мест трудоустройства, $rooms$ — среднее число комнат в доме жилого района, $stratio$ — среднее отношения числа студентов к числу учителей в школах района.
Коэффициент β_1 имеет смысл эластичности цены по признаку nox.
По экономическим соображениям интерес представляет гипотеза о том, что эластичность равна -1.

$H_0\colon \beta_1 = -1.$
$H_1\colon \beta_1 \neq -1 \Rightarrow p = 0.6945.$

Критерий Фишера

$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

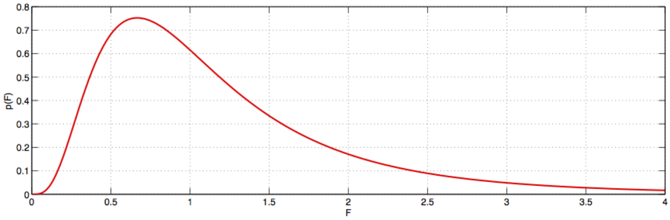
нулевая гипотеза: $H_0\colon \beta_2 = 0$

альтернатива: $H_1\colon H_0$ неверна

статистика: $RSS_r = \|y - X_1\beta_1\|_2^2, \quad RSS_{ur} = \|y - X\beta\|_2^2,$

$$F = \frac{(RSS_r - RSS_{ur})/k_1}{RSS_{ur}/(n-k-1)}$$

нулевое распределение: $F(k_1, n-k-1)$



Критерий Фишера

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.
Зависит ли вес ребёнка при рождении от уровня образования родителей?

$H_0: \beta_4 = \beta_5 = 0.$
$H_1: H_0$ неверна $\Rightarrow p = 0.2421.$

Связь между критериями Фишера и Стьюдента

Если $k_1 = 1$, критерий Фишера эквивалентен критерию Стьюдента для двусторонней альтернативы.

Иногда критерий Фишера отвергает гипотезу о незначимости признаков X_2, а критерий Стьюдента не признаёт значимым ни один из них. Возможные объяснения:

- отдельные признаки из X_2 недостаточно хорошо объясняют y, но совокупный эффект значим;
- признаки в X_2 мультиколлинеарны.

Иногда критерия Фишера не отвергает гипотезу о незначимости признаков X_2, а критерий Стьюдента признаёт значимыми некоторые из них. Возможные объяснения:

- незначимые признаки в X_2 маскируют влияние значимых;
- значимость отдельных признаков в X_2 — результат множественной проверки гипотез.

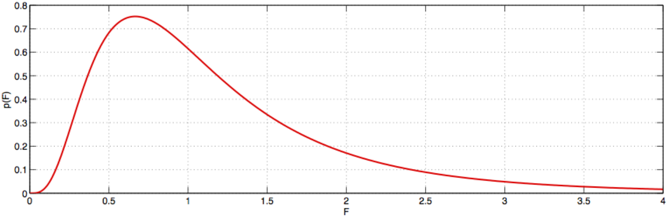
Критерий Фишера

нулевая гипотеза: $H_0\colon \beta_1 = \dots = \beta_k = 0$

альтернатива: $H_1\colon H_0$ неверна

статистика: $F = \frac{R^2/k}{\left(1-R^2\right)/(n-k-1)}$

нулевое распределение: $F(k, n-k-1)$



Критерий Фишера

Пример: имеет ли вообще смысл модель веса ребёнка при рождении, рассмотренная выше?

$H_0 : \beta_1 = \cdots = \beta_5 = 0.$
$H_1 : H_0$ неверна $\Rightarrow p = 6.0331 \times 10^{-9}.$

Сравнение невложенных моделей

Пример: имеются две модели:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{1}$$

$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \varepsilon. \tag{2}$$

Как понять, какая из них лучше?

Критерий Давидсона-Маккиннона

Пусть $\hat{y}$ — оценка отклика по первой модели, $\hat{\hat{y}}$ — по второй. Подставим эти оценки как признаки в чужие модели:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \hat{\hat{y}} + \varepsilon,$$
$$y = \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \gamma_3 \hat{y} + \varepsilon.$$

При помощи критерия Стьюдента проверим

$$H_{01}\colon \beta_3 = 0, \;\; H_{11}\colon \beta_3 \neq 0,$$
$$H_{02}\colon \gamma_3 = 0, \;\; H_{12}\colon \gamma_3 \neq 0.$$

H_{01} \ H_{02}	Принята	Отвергнута
Принята	Обе модели хороши	Модель (1) значимо лучше
Отвергнута	Модель (2) значимо лучше	Обе модели плохи

Пошаговая регрессия

- **Шаг 0**. Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается F-статистика каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения $p_E = 0.05$.
- **Шаг 1**. Рассчитывается F-статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e1}. Аналогично принимается решение о включении X_{e2}.
- **Шаг 2**. Если была добавлена переменная X_{e2}, возможно, X_{e1} уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение $p_R = 0.1$.
- . . .

Значимость категориальных предикторов

Категориальный предиктор, кодируемый несколькими фиктивными переменными, необходимо включать или исключать целиком. Значимость соответствующих фиктивных переменных лучше проверять в совокупности.

В случае, когда по отдельности какие-то фиктивные переменные не значимы, допустимо объединять уровни категориального предиктора, основываясь на интерпретации.

Проверка предположений Гаусса-Маркова

1. Линейность отклика: $y = X\beta + \varepsilon$.
2. Случайность выборки: наблюдения $(x_i, y_i), i = 1, \ldots, n$ независимы.
3. Полнота ранга: ни один из признаков не является константой или линейной комбинацией других признаков ни в популяции, ни в выборке $(\operatorname{rank} X = k + 1)$.
4. Случайность ошибок: $\mathbb{E}\left(\varepsilon \,|X\right) = 0$.
5. Гомоскедастичность ошибок: дисперсия ошибки не зависит от значений признаков: $\mathbb{D}\left(\varepsilon \,|X\right) = \sigma^2$.

Проверка предположений Гаусса-Маркова

- Предположения (1-2) проверить нельзя.
- Предположение (3) легко проверяется, без его выполнения построить модель вообще невозможно.
- Предположения (4-6) об ошибке ε необходимо проверять.

Оценивать ошибку ε будем при помощи **остатков**:

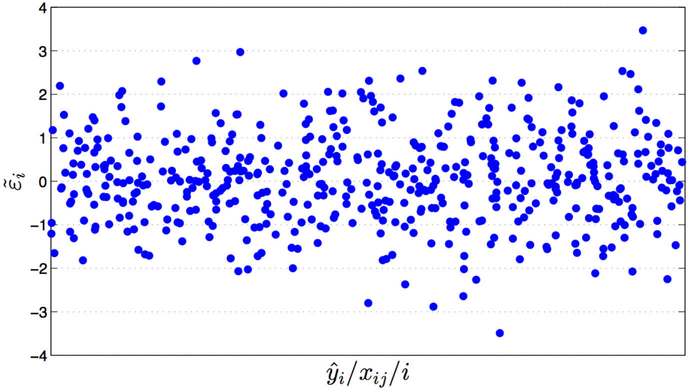
$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \ i = 1, \ldots, n.$$

Стандартизированные остатки:

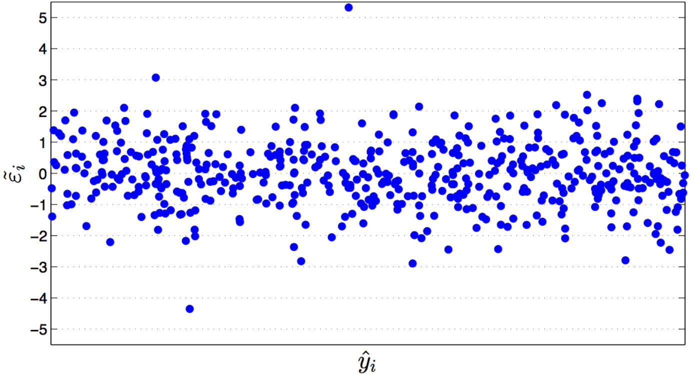
$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}}, \ i = 1, \ldots, n.$$

Визуальный анализ

Строятся графики зависимости $\tilde{\varepsilon}_i$ от $\hat{y}_i, \quad x_{ij}, j = 1, \ldots, k, \quad i.$

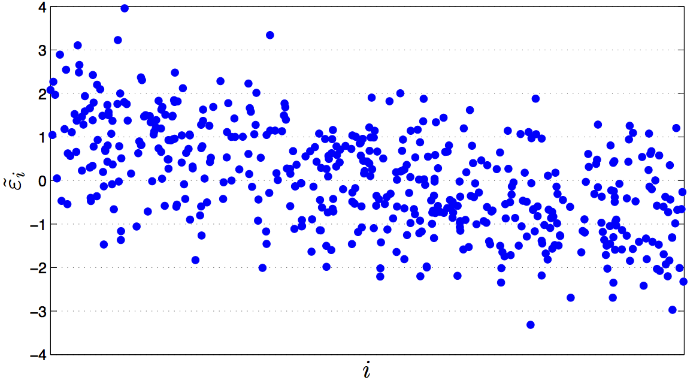


Визуальный анализ



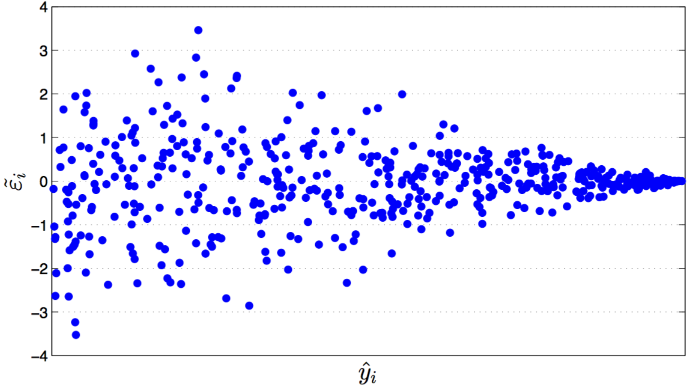
Возможно, присутствуют выбросы

Визуальный анализ



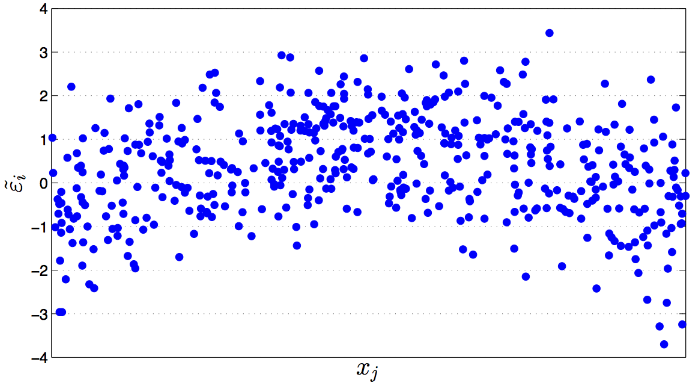
В данных имеется тренд

Визуальный анализ



Гетероскедастичность

Визуальный анализ



Стоит добавить квадрат признака x_j

Формальные критерии

- Проверка нормальности — критерий Шапиро-Уилка, qqplot.
- Проверка несмещённости: если остатки нормальны — критерий Стьюдента, нет — непараметрический критерий.
- Проверка гомоскедастичности: критерий Бройша-Пагана.

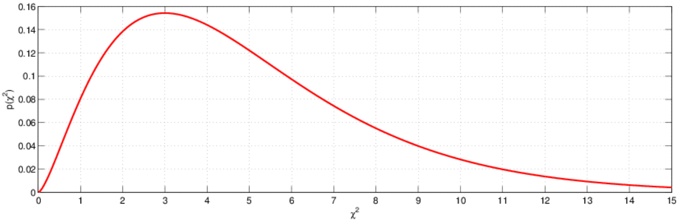
Критерий Бройша-Пагана

нулевая гипотеза: $H_0\colon \mathbb{D}\varepsilon_i = \sigma^2$

альтернатива: $H_1\colon H_0$ неверна

статистика: $LM = nR^2_{\hat{\varepsilon}^2}$, $R^2_{\hat{\varepsilon}^2}$ — коэффициент детерминации при регрессии квадратов остатков на признаки

нулевое распределение: χ^2_k



Гетероскедастичность

Гетероскедастичность может быть следствием недоопределения модели.

Последствия гетероскедастичности:

- МНК-оценки β и R^2 остаются несмещёнными и состоятельными
- нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для σ и β (независимо от объёма выборки)

Варианты:

- переопределить модель, добавить признаки, преобразовать отклик
- использовать модифицированные оценки дисперсии коэффициентов

Преобразование Бокса-Кокса

Пусть значения отклика $y_1, \ldots, y_n$ положительны. Если $\frac{\max y_i}{\min y_i} > 10$, стоит рассмотреть возможность преобразования y. В каком виде его искать?

Часто полезно рассмотреть преобразования вида y^λ, но оно не имеет смысла при $\lambda = 0$.
Вместо него можно рассмотреть семейство преобразований

$$W = \begin{cases} \left(y^\lambda - 1\right)/\lambda, & \lambda \neq 0, \\ \ln y, & \lambda = 0. \end{cases}$$

но оно сильно варьируется по λ.
Вместо него можно рассмотреть семейство преобразований

$$V = \begin{cases} \left(y^\lambda - 1\right)/\left(\lambda \dot{y}^{\lambda-1}\right), & \lambda \neq 0, \\ \dot{y} \ln y, & \lambda = 0, \end{cases}$$

где $\dot{y} = (y_1 y_2 \ldots y_n)^{1/n}$ — среднее геометрическое наблюдений отклика.

Метод Бокса-Кокса

Процесс подбора λ:

1. выбирается набор значений λ в некотором интервале, например, $(-2, 2)$;
2. для каждого значения λ выполняется преобразование отклика V, строится регрессия V на X, вычисляется остаточная сумма квадратов $RSS(\lambda)$;
3. строится график зависимости $RSS(\lambda)$ от λ, по нему выбирается оптимальное значение λ;
4. выбирается ближайшее к оптимальному удобное значение λ (например, целое или полуцелое);
5. строится окончательная регрессионная модель с откликом y^λ или $\ln y$.

Доверительный интервал для λ определяется как пересечение кривой $RSS\left(\lambda\right)$ с линией уровня $\min\limits_{\lambda} RSS\left(\lambda\right) \cdot e^{\chi^2_{1,1-\alpha}/n}$. Если он содержит единицу, возможно, не стоит выполнять преобразование.

Устойчивая оценка дисперсии Уайта

Если не удаётся избавиться от гетероскедастичности, при анализе моделей (дальше) можно использовать устойчивые оценки дисперсии.
White's heteroscedasticity-consistent estimator (HCE):

$$\mathbb{D}\left(\hat{\beta}\middle| X\right) = \left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Асимптотика устойчивой оценки:

$$\sqrt{n}\left(\beta - \hat{\beta}\right) \xrightarrow{d} N\left(0, \Omega\right),$$

$$\hat{\Omega} = n\left(X^T X\right)^{-1} \left(X^T \operatorname{diag}\left(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\right) X\right) \left(X^T X\right)^{-1}.$$

Другие устойчивые оценки дисперсии

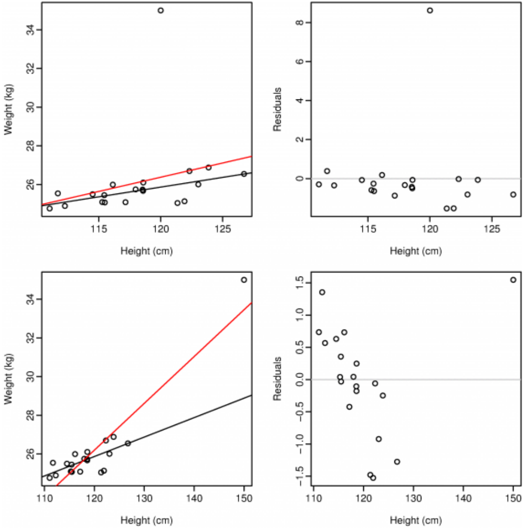
Элементы диагональной матрицы могут задаваться разными способами:

const	$\hat{\sigma}^2$
HC0	$\hat{\varepsilon}_i^2$
HC1	$\frac{n}{n-k}\hat{\varepsilon}_i^2$
HC2	$\frac{\hat{\varepsilon}_i^2}{1-h_i}$
HC3	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}$
HC4	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^{\min\left(4,\frac{nh_i}{k}\right)}}$

const — случай гомоскедастичной ошибки,
HC0 — оценка Уайта,
HC1–HC3 — модификации МакКиннона-Уайта,
HC4 — модификация Крибари-Нето.
В python: model.get$_{r}obustcov_{r}esults()$

Расстояние Кука

Регрессия сильно подстраивается под далеко стоящие наблюдения.



Расстояние Кука

Расстояние Кука — мера воздействия i-го наблюдения на регрессионное уравнение:

$$D_i = \frac{\sum\limits_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{RSS\,(k+1)} = \frac{\hat{\varepsilon}_i^2}{RSS\,(k+1)} \frac{h_i}{(1-h_i)^2},$$

$\hat{y}_{j(i)}$ — предсказания модели, настроенной по наблюдениям $1,\ldots,i-1,i+1,\ldots,n$, для наблюдения j;
h_i — диагональный элемент матрицы $H = X\left(X^T X\right)^{-1} X^T$ (hat matrix).

Варианты порога на D_i:

- $D_i = 1$;
- $D_i = 4/n$;
- $D_i = 3\bar{D}$;
- визуально по графику зависимости D_i от $\hat{y}_i$.

Обобщённая линейная модель

$1, \ldots, n$ — объекты;
$x_1, \ldots, x_k$ — предикторы;
y — отклик;

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}; \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix};$$

регрессионная модель:

$$\mathbb{E}\left(y \,|\, X\right) \equiv \mu = f\left(x_1, \ldots, x_k\right);$$

линейная регрессионная модель:

$$\mu = X\beta;$$

обобщённая линейная регрессионная модель (GLM):

$$g\left(\mu\right) = X\beta, \quad \mu = g^{-1}\left(X\beta\right),$$

$g\left(x\right)$ — связующая функция — позволяет ограничить диапазон предсказываемых для μ значений.

Обобщённая линейная модель

В обычной линейной модели используется предположение о нормальности отклика:

$$y\,|X \sim N\left(X\beta, \sigma^2\right).$$

В обобщённой линейной модели распределение y берётся из экспоненциального семейства:

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

	$Pois(\lambda)$	$Bin(N, p)$	$N\left(\mu, \sigma^2\right)$
$a(\phi)$	1	1	σ^2
$b(\theta)$	e^θ	$n \ln\left(1 + e^\theta\right)$	$\theta^2/2$
$c(y, \phi)$	$\ln y!$	$\ln C_n^y$	$\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$
$g(x)$	$\ln x$	$\ln \frac{x}{1-x}$	x
$g^{-1}(x)$	$e^x \in [0, \infty)$	$\frac{e^x}{1+e^x} \in [0, 1]$	$x \in \mathbb{R}$

Оценка параметров GLM

$\hat{\beta}$:

- оценивается методом максимального правдоподобия;
- существует и единственна,
- находится численно
  - методом Ньютона-Рафсона (Newton-Raphson method)
  - методом оценок Фишера (Fisher scoring method)
- состоятельна, асимптотически эффективна, асимптотически нормальна.

Итерационный процесс вычисления $\hat{\beta}$ может не сойтись, если k слишком велико относительно n.

$$\mathbb{D}\hat{\beta} = I^{-1}\left(\hat{\beta}\right),$$

$I\left(\beta\right) \in \mathbb{R}^{(k+1)\times(k+1)}$ — информационная матрица Фишера — матрица вторых производных логарифма правдоподобия $L\left(\beta\right)$.

Доверительные интервалы

Для отдельного коэффициента β_j:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}\left(\hat{\beta}\right)\right)_{jj}}.$$

Для $g\left(\mathbb{E}\left(y\,|x_0\right)\right)$ — преобразованного матожидания отклика на новом объекте x_0:

$$x_0^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}.$$

Для матожидания отклика на новом объекте x_0:

$$\left[g^{-1}\left(x_0^T\hat{\beta} - z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}\right), g^{-1}\left(x_0^T\hat{\beta} + z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}\right)\right].$$

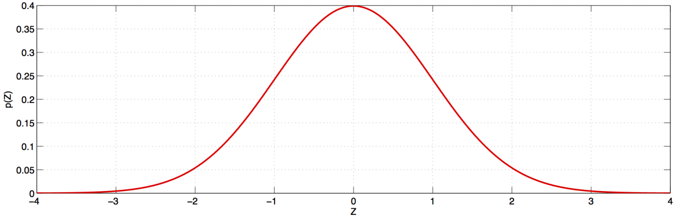
Критерий Вальда

нулевая гипотеза: $H_0: \beta_j = 0$

альтернатива: $H_1: \beta_j <\neq> 0$

статистика: $T = \frac{\hat{\beta}_j}{\sqrt{\left(I^{-1}(\hat{\beta})\right)_{jj}}}$

нулевое распределение: $N(0,1)$



Критерий отношения правдоподобия

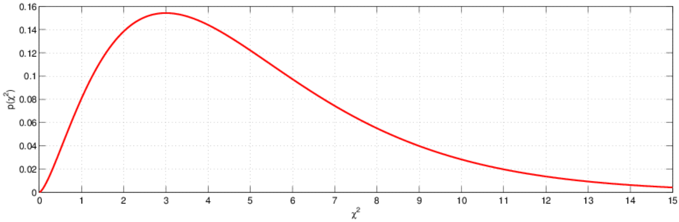
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2} \right); \quad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T} \right)^T;$$

нулевая гипотеза: $H_0\colon \beta_2 = 0$

альтернатива: $H_1\colon H_0$ неверна

статистика: $G = 2\left(L_r - L_{ur}\right)$

нулевое распределение: $\chi^2_{k_1}$



Связь между критериями Вальда и отношения правдоподобия

При $k_1 = 1$ критерии Вальда и отношения правдоподобия не эквивалентны, в отличие от случая линейной регрессии, когда в этом случае достигаемые уровни значимости критериев Стьюдента и Фишера совпадают.

При больших n разница между критериями невелика, но в случае, когда их показания расходятся, рекомендуется смотреть на результат критерия отношения правдоподобия.

Меры качества моделей

Остаточная аномальность (residual deviance):

$$D_{res} = 2(L_{sat} - L_{fit})$$

Где L_{sat} – насыщенная (saturated) модель, имеющая число параметров равное числу объектов.
Аномальность — аналог RSS в линейной регрессии; при добавлении признаков она не может убывать.

Для сравнения моделей с разным числом признаков можно использовать **информационные критерии**:

1. AIC — информационный критерий Акаике:

$$AIC = -2L + 2\,(k+1)\,;$$

2. $AICc$ — он же с поправкой на случай небольшого размера выборки;

$$AICc = -2L + \frac{2k\,(k+1)}{n-k-1};$$

3. BIC (SIC) — байесовский (Шварца) информационный критерий:

$$BIC = -2L + \ln n\,(k+1)\,.$$

Меры качества моделей

- $AIC < AICc$
- $BIC > AIC$ при $n \geq 8$
- выбор модели по BIC приводит к состоятельным оценкам — с ростом n вероятность выбора верного подмножества признаков стремится к 1
- минимизация AIC асимптотически даёт модель с наименьшей среднеквадратичной ошибкой предсказания
- модели со значением информационного критерия на расстоянии двух единиц от значения лучшей модели можно считать неотличимыми от лучшей

Постановка

Задача: оценить влияние одного или нескольких признаков на наступление какого-либо события и оценить его вероятность.

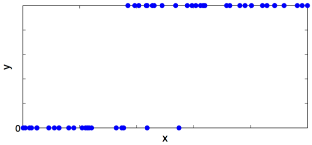
$1, \ldots, n$ — объекты;
$x_1, \ldots, x_k$ — предикторы;
y — отклик, $y_i \in \{0, 1\}$.

Хотим найти такой вектор β, что

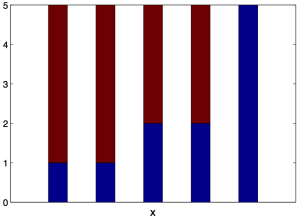
$$\mu = \mathbb{E}\left(y \,|\, X\right) = P\left(y = 1 \,|\, X\right) \equiv \pi(x) \approx X\beta.$$

Примеры

Неповторяемый эксперимент со случайными уровнями фактора: построение кривой спроса, x_i — цена товара, y_i — согласие купить товар.



Повторяемый эксперимент с фиксированными уровнями фактора: разработка пестицидов, x_i — доза пестицида, y_i — смерть вредителя.



$\Longrightarrow$ логистическая регрессия может также использоваться для моделирования $y \in [0, 1]$.

Параметризация

Линейная регрессия:

$$\pi(x) = \beta_0 + \beta_1 x + \varepsilon.$$

- Оценка вероятности может выходить за $[0, 1]$.
- В линейной регрессии $y = \mathbb{E}(y|x) + \varepsilon$, и МНК-оценка β хороша, когда $\varepsilon \sim N(0, \sigma)$. Здесь же, если $y = \pi(x) + \varepsilon$, то $\varepsilon = 1 - \pi(x)$ или $\varepsilon = \pi(x)$, и МНК-оценка будет плохой.

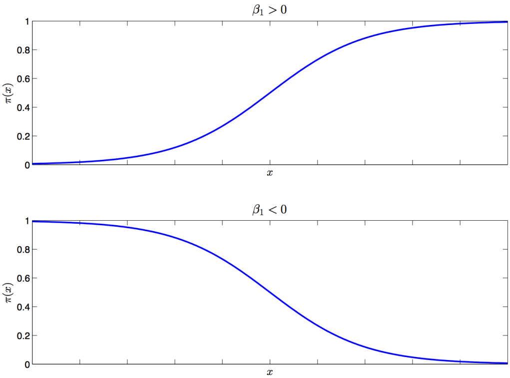
Параметризация

Логит:

$$g(x) = g(\pi(x)) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x + \varepsilon,$$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

Параметризация



- $\hat{\pi}(x) = g^{-1}(\beta_0 + \beta_1 x)$ принимает значения из $[0, 1]$;
- изменения на краях диапазона значений x приводят к меньшим изменениям $\pi(x)$: x — годовой доход, y — покупка автомобиля,

$$\pi(10\,000\,000 + 200\,000) - \pi(10\,000\,000) < \pi(500\,000 + 200\,000) - \pi(500\,000).$$

Отношение шансов

Пусть $y \sim Ber(p)$, тогда шансы (odds) события $y = 1$:

$$ODDS = \frac{p}{1-p}.$$

Если $y_1 \sim Ber(p_1)$, $y_2 \sim Ber(p_2)$, то отношение шансов (odds ratio) события $y_1 = 1$ по сравнению с событием $y_2 = 1$:

$$OR = \frac{p_1 / (1-p_1)}{p_2 / (1-p_2)}.$$

Серд. заболевания \ Возраст	≥ 55	< 55
есть	21	22
нет	6	51

$$OR = \frac{21/6}{22/51} \approx 8.1.$$

Роль коэффициентов логистической регрессии

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad \Longleftrightarrow \quad \frac{p}{1-p} = e^{\hat{\beta}_0} \left(e^{\hat{\beta}_1}\right)^x$$

Пусть $x = [\text{возраст} \geq 55]$, $y = \left[\text{есть сердечные заболевания}\right]$. По $\hat{\beta}_1$ легко оценить отношение шансов получения заболевания пожилыми людьми:

$$\widehat{OR} = e^{\hat{\beta}_1}.$$

Пусть $x = \text{возраст}$, $y = \left[\text{есть сердечные заболевания}\right]$. $e^{\hat{\beta}_1}$ имеет смысл мультипликативного прироста риска получения заболевания при увеличении возраста на 1 год.

Настройка параметров

ММП:

$$\begin{aligned}
P\left(x_i, 1\right) &= \pi\left(x_i\right), \\
P\left(x_i, 0\right) &= 1 - \pi\left(x_i\right), \\
l\left(\beta\right) &= \prod_{i=1}^{n} \pi\left(x_i\right)^{y_i} \left(1 - \pi\left(x_i\right)\right)^{1-y_i}, \\
L\left(\beta\right) &= \ln l\left(\beta\right) = \sum_{i=1}^{n} \left(y_i \ln \pi\left(x_i\right) + \left(1 - y_i\right) \ln\left(1 - \pi\left(x_i\right)\right)\right), \\
\hat{\beta} &= \beta L\left(\beta\right).
\end{aligned}$$

Информационная матрица Фишера:

$$I\left(\hat{\beta}\right) = X^T V X,$$

$$V = \operatorname{diag}\left(\hat{\pi}\left(x_i\right)\left(1 - \hat{\pi}\left(x_i\right)\right)\right).$$

Проблемы настройки параметров

Если матрица X вырождена, некоторые коэффициенты модели не будут определены.

Если наблюдения $y = 0$ и $y = 1$ линейно разделимы в пространстве X, то:

- в теории коэффициенты бесконечно возрастают
- на практике коэффициенты и их дисперсии получаются большими, а почти все вероятности в обучающей выборке близки к 0 или 1.

Можно использовать регуляризацию Фирта. Функция меток исходной модели для коэффициента β_j:

$$\sum_{i=1}^{n} \left(y_i - \pi\left(x_i\right)\right) x_{ij}.$$

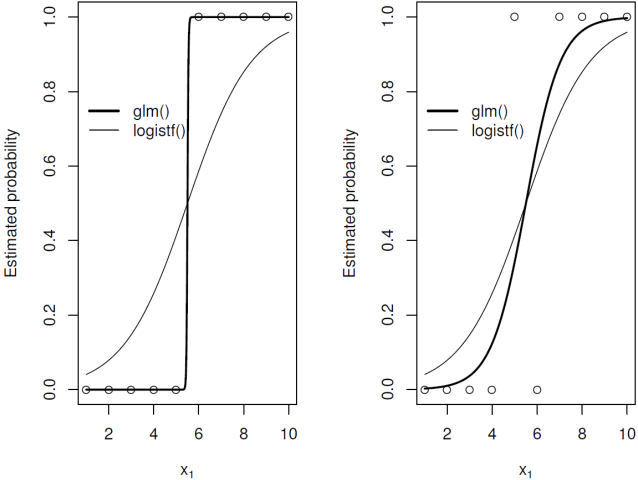
Регуляризованная версия:

$$\sum_{i=1}^{n} \left(y_i - \pi\left(x_i\right) + h_i\left(0.5 - \pi\left(x_i\right)\right)\right) x_{ij},$$

h_i — диагональный элемент hat matrix:

$$H = V^{1/2} X \left(X^T V X\right)^{-1} X^T V^{1/2}.$$

Проблемы настройки параметров



Доверительные интервалы

Для отдельного коэффициента β_j:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}\left(\hat{\beta}\right)\right)_{jj}}.$$

Для $g(x_0)$ — логита нового объекта x_0:

$$x_0^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}.$$

Для вероятности $y=1$ при $x=x_0$:

$$\left[\frac{e^{x_0\hat{\beta}-z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}{1+e^{x_0\hat{\beta}-z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}, \frac{e^{x_0\hat{\beta}+z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}{1+e^{x_0\hat{\beta}+z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}\right].$$

Линейность логита

Проверка линейности логита по признакам — аналог визуального анализа остатков в обычной линейной регрессии.

Методы анализа линейности логита:

- сглаженные диаграммы рассеяния;
- дробные полиномы.

Сглаженные диаграммы рассеяния (smoothed scatterplots)

Рассмотрим оценку логита, полученную ядерным сглаживанием по x_j:

$$\bar{y}_{sm}\left(x_{ji}\right) = \frac{\sum\limits_{l=1}^{n} y_i K\left(\frac{x_{ji}-x_{li}}{h}\right)}{\sum\limits_{l=1}^{n} K\left(\frac{x_{ji}-x_{li}}{h}\right)},$$

$$\bar{l}_{sm}\left(x_{ji}\right) = \ln \frac{\bar{y}_{sm}\left(x_{ji}\right)}{1-\bar{y}_{sm}\left(x_{ji}\right)}.$$

График функции $\bar{l}_{sm}\left(x_j\right)$ должен быть похож на прямую.

Дробные полиномы (fractional polynomials)

Если логит нелинеен по признаку, можно попробовать добавлять в модель его осмысленные степени и проверять их значимость.
В автоматическом режиме это можно делать с помощью дробных полиномов.

1. Настраиваются модели с заменой x_j на допустимые степени признака x_j, например, из множества $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Выбирается степень, максимизирующая правдоподобие.
2. Настраиваются модели с заменой x_j на двухкомпонентный полином x_j вида $\beta_{j1} x_j^{p_1} + \beta_{j2} x_j^{p_2}, \quad p_1, p_2 \in S$ (если $p_1 = p_2$, то берётся $\beta_{j1} x_j^{p_1} + \beta_{j2} x_j^{p_1} \ln x_j$). Выбираются степени, максимизирующие правдоподобие.
3. Если модель с полиномом второй степени значимо не лучше, чем линейная, используется линейная модель.
4. Если модель с полиномом второй степени значимо не лучше, чем с полиномом первой степени, используется модель с полиномом первой степени, иначе — с полиномом второй.

Содержательный отбор признаков

1. Если признаков достаточно много (например, больше 10), желательно сделать их предварительный отбор, основанный на значимости в однофакторной логистической регрессии. Для дальнейшего рассмотрения остаются признаки, достигаемый уровень значимости которых не превышает 0.25.
2. Строится многомерная модель, включающая все отобранные на шаге 1 признаки. Проверяется значимость каждого признака, удаляется небольшая группа незначимых признаков. Новая модель сравнивается со старой с помощью критерия отношения правдоподобия.
3. К признакам модели, полученной в результате циклического применения шагов 2 и 3, по одному добавляются удалённые признаки. Если какой-то из них становится значимым, он вносится обратно в модель.

Содержательный отбор признаков

4. Для непрерывных признаков полученной модели проверяется линейность логита. В случае обнаружения нелинейности признаки заменяются на соответствующие полиномы.
5. Исследуется возможность добавления в полученную модель взаимодействий факторов. Добавляются значимые интерпретируемые взаимодействия.
6. Проверяется адекватность финальной модели: близость y и $\hat{y}$; малость вклада наблюдений (x_i, y_i) на каждом объекте i в $\hat{y}$.

Порог классификации

Как по $\pi(x)$ оценить y?

$$y = [\pi(x) \geq p_0].$$

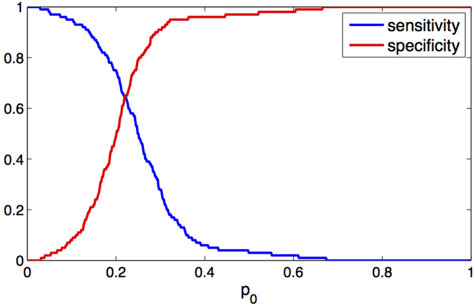
Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

Порог классификации

Пример: эффективность терапии для наркозависимых, $p_0 = 0.5$:

$\hat{y}$ \ y	1	0
1	16	11
0	131	417

Чувствительность: $\frac{16}{16+131} \approx 10.9\%$.
Специфичность: $\frac{417}{11+417} \approx 97.4\%$.



Выбросы

Остатки Пирсона:

$$r_i = \frac{y_i - \hat{\pi}\left(x_i\right)}{\sqrt{\hat{\pi}\left(x_i\right)\left(1 - \hat{\pi}\left(x_i\right)\right)}}.$$

Аналог расстояния Кука:

$$\Delta\hat{\beta}_i = \frac{r_i^2 h_i}{\left(1 - h_i\right)^2}.$$

Требования к решению задачи методом линейной регрессии

- визуализация данных, анализ распределения признаков (оценка необходимости трансформации), оценка наличия выбросов;
- оценка необходимости преобразования отклика и его поиск методом Бокса-Кокса;
- визуальный анализ остатков;
- проверка гипотез об остатках: нормальность, несмещённость, гомоскедастичность;
- отбор признаков с учётом множественной проверки гипотез и возможной гетероскедастичности;
- анализ необходимости добавления взаимодействий и квадратов признаков;
- расчёт расстояний Кука, возможное удаление выбросов, обновление модели;
- выводы.

Требования к решению задачи методом логистической регрессии

- визуализация данных, оценка наличия выбросов, анализ таблиц сопряжённости по категориальным признакам;
- содержательный отбор признаков: выбор наилучшей линейной модели, оценка линейности непрерывных признаков по логиту, анализ необходимости добавления взаимодействий, проверка адекватности финальной модели (анализ влиятельных наблюдений, классификация);
- выводы.

Литература

- линейная регрессия в целом — Wooldridge (много примеров, без матричной алгебры);
- критерий Давидсона-Маккиннона (Davidson-MacKinnon test) — Davidson;
- множественная оценка значимости коэффициентов — Bretz, 4.4;
- преобразование Бокса-Кокса (Box-Cox transformation) — Дрейпер, гл. 14;
- устойчивые оценки дисперсии — White, MacKinnon, Cribari-Neto;
- расстояние Кука (Cook's distance) — Cook.

Дрейпер Н.Р., Смит Г. *Прикладной регрессионный анализ*, 2007.

Кобзарь А.И. *Прикладная математическая статистика*, 2006.

Bretz F., Hothorn T., Westfall P. *Multiple Comparisons Using R*, 2010.

Cook D.R., Weisberg S. *Residuals and influence in regression*, 1982.

Литература

Cribari-Neto F. (2004). *Asymptotic inference under heteroskedasticity of unknown form*. Computational Statistics & Data Analysis, 45(2), 215–233.

Davidson R., MacKinnon J. (1981). *Several Tests for Model Specification in the Presence of Alternative Hypotheses*. Econometrica, 49, 781-793.

Freedman D.A. *A Note on Screening Regression Equations*. The American Statistician, 37(2), 152-155.

MacKinnon J., White H. (1985). *Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties*. Journal of Econometrics, 29, 305–325.

White H. (1980). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Econometrica: Journal of the Econometric Society, 48(4), 817–838.

Wooldridge J. *Introductory Econometrics: A Modern Approach*, 2016.