Лекция 10. Мультиколлинеарность. Робастные регрессионные модели. Логит-регрессия

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Содержание

Ридж-регрессия

Рассмотрим следующая модель наблюдений, связывающую значения некоторого наблюдаемого показателя y и объясняющих переменных $x = (x_1, \ldots, x_k)^T$:

$$Y = X\beta + \varepsilon, \tag{1}$$

где $Y = (y_1, \ldots, y_n)^T$ — наблюдения, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^T$ — вектор неизвестных параметров, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ — вектор ненаблюдаемых случайных компонент,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

— матрица порядка $n \times (k+1)$, x_{ij} - значения объясняющих факторов, n — объем наблюдений.

Предположим наблюдается мультиколлинеарность столбцов матрицы X и как следствие плохая обусловленность матрицы $X^T X$ и неустойчивость оценок коэффициентов регрессии. Оценки могут иметь, например, неправильный знак или иметь значения, которые намного превосходят те, которые приемлемы из физических или практических соображений.

Критерием плохой обсуловленности является высокая величина отношения $\lambda_{max}/\lambda_{min}$ максимального и минимального собственных чисел матрицы $X^T X$, — называемого показателем обусловленности. Это соотношение также позволяет судить о степени серьезности проблем мультиколлинеарности: показатель обусловленности в пределах от 10 до 100 свидетельствует об умеренной коллинеарности, свыше 1000 — об очень серьезной коллинеарности.

Наиболее детальным показателем наличия проблем, связанных с мультиколлинеарностью, является коэффициент увеличения дисперсии, определяемый для каждой переменной как

$$VIF(\beta_j) = \frac{1}{1 - R_j^2},$$

где R_j^2 — коэффициент множественной детерминации в регрессии X_j на прочие X, т.е. уравнения регрессии

$$x_j = c_0 + c_1 x_1 + \ldots + c_{j-1} x_{j-1} + c_{j+1} x_{j+1} + \ldots + c_k x_k, \quad j = 1, \ldots, m$$

О мультиколлинеарности будет свидетельствовать VIF от 4 и выше хотя бы для одного j.

Если фактор x_j имеет небольшой разброс значений, то вектор X_j будет коррелировать с вектором X_0. Для того, чтобы обойти данную проблему стандартизируем факторы и отклик. А именно факторы центрируем и нормируем, а отклик центрируем:

$$z_{ij} = \frac{x_{ij} - \overline{x}}{\sqrt{\frac{1}{n}\sum_{m=1}^{n}(x_{mj} - \overline{x})^2}}, \quad \overline{x} = \frac{1}{n}\sum_{m=1}^{n} x_{mj}$$

$$Y' = Y - \overline{y}X_0, \quad X_0 = (1, 1, \ldots, 1)^T.$$

В результате стандартизации перейдем от модели (1) к модели

$$Y' = Z\beta' + \varepsilon, \tag{2}$$

где $Y' = (y'_1, \ldots, y'_n)^T$ — вектор центрированных наблюдений, $\beta' = (\beta'_1, \ldots, \beta'_k)^T$ — вектор неизвестных параметров,

$$Z = \begin{pmatrix} z_{11} & z_{12} & \ldots & z_{1k} \\ z_{21} & z_{22} & \ldots & z_{2k} \\ \ldots & \ldots & \ldots & \\ z_{n1} & z_{n2} & \ldots & z_{nk} \end{pmatrix}$$

— матрица порядка $n \times k$, z_{ij} - значения нормированных центрированных факторов.

Параметры модели (2) связаны с параметрами исходной модели (1) следующими соотношениями

$$\beta_j = \frac{\beta'_j}{\frac{1}{n}\sum_{m=1}^{n}(x_{mj} - \overline{x})^2}, \quad j = 1, \ldots, k.$$

Сдедовательно, оценки $\hat{\beta}$ неизвестных параметров β исходной модели (1) могут быть выражены через оценки $\hat{\beta}'$ модели (2)

$$\hat{\beta}_j = \frac{\hat{\beta}'_j}{\frac{1}{n}\sum_{m=1}^{n}(x_{mj} - \overline{x})^2}, \quad j = 1, \ldots, k.$$

$$\hat{\beta}_0 = \overline{y} - \sum_{j=1}^{n} \hat{\beta}_j \overline{x}_j.$$

Оценки метода наименьших квадратов $\hat{\beta}'$ неизвестных параметров β' модели (2) могут быть получены по формуле

$$\hat{\beta}' = (Z^T Z)^{-1} Z^T Y'. \tag{3}$$

в случае обратимости матрицы $Z^T Z$.

Одним из методов, позволяющих решить проблему мультиколлинеарности, является ридж-регрессия (гребневая регрессия). Идея подхода состоит в том, чтобы попытаться найти оценку, минимизирующую среднквадратическое отклонение оценки

$$\hat{\beta}' = \arg\min_{\hat{\beta}' \in B} E(\hat{\beta}' - \beta')^2,$$

где B — более широкий класс, чем класс несмещенных линейных оценок.
В рамках такого подхода матрицу $Z^T Z$ можно регуляризовать, или сделать "более обратимой"путем добавления заведомо регулярной матрицы Г размерности $k \times k$. Тогда в качестве минимизируемого критерия имеем

$$(Y' - Z\beta')^T (Y' - Z\beta') + (\Gamma\beta')^T (\Gamma\beta'), \tag{4}$$

Оценки метода ридж-регрессии $\hat{\beta'}_{ridge}$ неизвестных параметров β' будут иметь вид:

$$\hat{\beta'}_{ridge} = (Z^T Z + \Gamma^T \Gamma)^{-1} Z^T Y'. \tag{5}$$

Часто матрицу Γ берут равной $\sqrt{\lambda} E_k$, $\lambda > 0$. В этом случае оценки $\hat{\beta}'_{ridge}$ неизвестных параметров β' принимают вид:

$$\hat{\beta}'_{ridge} = (Z^T Z + \lambda E_k)^{-1} Z^T Y. \tag{6}$$

Замечание 1

1. *Если* $\lambda \to 0$, $\hat{\beta}'_{ridge} \to \hat{\beta}'_{mnk}$.
2. *Если* $\lambda \to \infty$, $\hat{\beta}'_{ridge} \to 0$.

Для любой матрицы Z матрица $Z^T Z + \lambda E_k$, $\lambda > 0$, обращаема, следовательно, всегда существует единственное решение $\hat{\beta}'_{ridge}$.

Можно выразить ридж-оценки через МНК-оценки

$$\hat{\beta}'_{ridge} = (E_k + \lambda(Z^T Z)^{-1})^{-1} \hat{\beta}'_{mnk} = Q\hat{\beta}'_{mnk},$$

так что ридж-оценки оказываются линейными комбинациями МНК-оценок.
Математическое ожидание оценок метода ридж-регресии

$$E\hat{\beta}'_{ridge} = (E_k - \lambda W)\beta',$$

где $W = (Z^T Z + \lambda E_k)^{-1}$.
Таким образом, оценки метода ридж-регрессии являются смещенными оценками параметров β со смещением

$$Bias(\hat{\beta}'_{ridge}) = -\lambda W \beta'.$$

Ковариационная матрица оценок ридж-регрессии

$$D\hat{\beta}'_{ridge} = \sigma^2 WZ^T ZW.$$

С ростом λ дисперсия оценок уменьшается, однако их смещение увеличивается.

Можно показать, что средний квадрат ошибки для ридж-оценок равен

$$MSE = E(\hat{\beta}'_{ridge} - \beta')^T(\hat{\beta}'_{ridge} - \beta') = \\ = \sigma^2 tr\{Q(Z^T Z)^{-1} Q^T\} + \beta'^T (Q - E_k)^T (Q - E_k)\beta'. \quad (7)$$

Теорема 1

Существует $\lambda^* > 0$ *такое, что*

$$E(\hat{\beta}'_{ridge} - \beta')^T(\hat{\beta}'_{ridge} - \beta') < E(\hat{\beta}'_{mnk} - \beta')^T(\hat{\beta}'_{mnk} - \beta')$$

Хотя величина $\lambda^* > 0$ существует, нет способа, позволяющего при решении конкретной практической задачи убедиться, что перед нами значение, которому отвечает величина среднего кварата ошибки, меньшая, чем средний квадрат ошибки МНК-оценок.

Робастные регресионные модели

Помимо проблемы мультиколлинеарности можно столкнуться с проблемой наличия выбросов в наблюдениях, т.е. редких но больших по величине значений факторов или отклика. Выбросы могут оказать сильное влияние на оценки параметров регрессионной модели, полученных методом наименьших квадратов.
Робастность оценок параметров линейной регрессионной модели может быть обеспечена различными способами, рассмотрим некоторые из них.

Рассмотрим множественную линейную регрессионную модель

$$Y = X\beta + \varepsilon, \tag{8}$$

M-оценки

Метод заключается в выборе функции ρ такой, что

- ρ обладает свойством симметрии;
- ρ неотрицательна;
- ρ монотонно неубывающая.

M-оценки неизвестных параметров β находят из условия

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - X_i^T \beta).$$

Если ρ дифференцируема $\psi = \rho'$, то

$$\sum_{i=1}^{n} \psi(y_i - X_i^T \beta) X_i = 0.$$

Функцию ψ называют функцией влияния.

Примеры функций ρ

Type	$\rho(x)$	$\psi(x)$
L_2	$x^2/2$	x
L_1	$\lvert x\rvert$	$Sgn(x)$
L_1 - L_2	$2(\sqrt{1+x^2/2}-1)$	$\frac{x}{\sqrt{1+x^2/2}}$
Huber *If* $\lvert x\rvert \le k$ *If* $\lvert x\rvert \ge k$	$x^2/2$ $k(\lvert x\rvert - \frac{k}{2})$	x $kSgn(x)$
L_p	$\frac{\lvert x\rvert^v}{v}$	$Sgn(x)\lvert x\rvert^{v-1}$
Cauchy	$\frac{c^2}{2}\log(1+(x/c)^2)$	$\frac{x}{1+(x/c)^2}$
German-Maclure	$\frac{x^2/2}{1+x^2}$	$\frac{x^2}{(1+x^2)^2}$
Welsch	$\frac{c^2}{2}[1-\exp(-(x/c)^2)]$	$x\exp(-(x/c)^2)$

Устойчивые M-оценки должны иметь:

- ограниченную функцию влияния,
- единственную точку минимума,

L_2 оценки не устойчивы, так как функция влияния не ограничена.
L_1 — оценки неизвестных параметров необязательно будут единственными.
Для функции Хубера рекомендуется выбирать $k = 1.345$. В этом случае имеем 95% асимпотическую эффективность оценок для нормального стандартного распределения.
Функция Коши не гарантирует единственного решения. Можно получить неверные решения. При $c = 2.3849$ имеем 95% асимпотическую эффективность оценок для нормального стандартного распределения.
Функции МкКлуре и Велча имеют такие же недостатки, что и функция Коши.

LMS и LTS

Метод наименьшей медианы квадратов (LMS)

$$\hat{\beta} = \arg\min_{\beta} \operatorname{med}\{(y_1 - X_1^T\beta)^2, \ldots, (y_n - X_n^T\beta)^2\}.$$

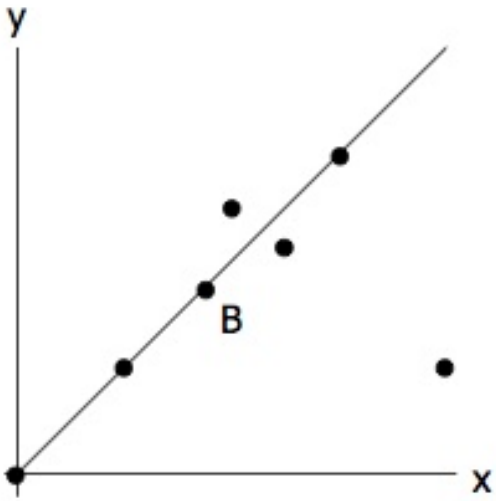
Метод наименьших усеченных квадратов (LMS)

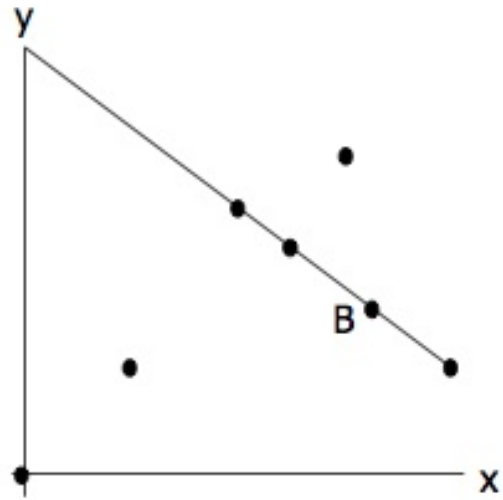
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (r^2)_{i:n},$$

где $(r^2)_{i:n}$ — i-ый наименьший квадрат остатка в сортированной по возрастанию последовательности квадратов остатков

$$(r^2)_{1:n} \leq (r^2)_{2:n} \leq \ldots \leq (r^2)_{n:n}.$$

Оценки LMS и LTS будут устойчивы к выбросам, но не стабильны. Кроме того, они не эффективны, если данные нормально распределены.





Сглаживание данных при помощи метода скользящей медианы

Другой способ построения робастных оценок регресии основан на предварительном сглаживании данных при помощи метода скользящей медианы [6] и последующем использовании МНК-оценок, рассчитанных по сглаженным данным.
Исходные наблюдения (x_i, y_i), $i = 1, \ldots, n$, преобразуют в следующие $(x_i, \tilde{y}_i)$, $i = m+1, \ldots, n-m$, где

$$\tilde{y}_i = \operatorname{med}\{y_{i-m}, y_{i-m+1}, \ldots, y_{i+m}\}, \quad i = m+1, \ldots, n-m,$$

представляет собой выборочную медиану, построенную по $2m+1$ последовательным значениям отклика $\{y_{i-m}, y_{i-m+1}, \ldots, y_{i+m}\}$, $m \geq 1$.
Выбор m равным ожилаемому числу выбросов в данных гарантирует робастность процедуры сглаживания.
Если исходные данные монотонны, то и сглаженные данные совпадают с исходными $\tilde{y}_i = y_i$, $i = m+1, \ldots, n-m$.
Метод скользящей медианы, как правило, применяется при анализе временных рядов.

Бинарная регрессия

Дискретные данные

Пусть зависимые переменные принимают дискретные значения, выражающие какие-либо качественные признаки. Объясняющие переменные могут быть как дискретными, так и непрерывными. Выделим несколько классов задач, в которых зависимые переменные принимают дискретные значения:

1. Переменные — это решения «да» (1) или «нет» (0), т. е. выбор одной из двух альтернатив.
   Если имеется только две альтернативы, то результат наблюдения обычно описывается переменной, называемой бинарной. В общем случае при наличии k альтернатив результат выбора можно представить переменной, принимающей значения $1, \ldots, k$. В этих случаях соответствующую переменную называют *номинальной*.
2. Переменные — ранги. Соответствующая переменная называется *порядковой, ординальной* или *ранговой*.
3. Переменная — количественная целочисленная характеристика.

Модель линейной вероятности

Пусть имеется выборка объема n наблюдений (x_i, y_i), $i = 1, \ldots, n$, где $x_i^T = (1, x_{i1}, \ldots, x_{ik})^T$, y_i — зависимая переменная, которая может принимать только два значения: ноль и единица. Рассмотрим стандартную модель линейной регрессии:

$$y_i = \beta^T x_i + \varepsilon_i, \tag{9}$$

где $\beta^T = (\beta_0, \beta_1, \ldots, \beta_k)$ — вектор неизвестных параметров, $\beta \in \mathbb{R}^k$, ε_i — случайная компонента. В предположениях регрессионного анализа считается, что случайная компонента подчиняется нормальному закону распределения с нулевым математическим ожиданием. Учитывая это, получаем, что

$$Ey_i = \beta^T x_i.$$

Так как y_i принимает значения 0 или 1, то для математического ожидания y_i имеем равенство:

$$Ey_i = 1 \cdot P\{y_i = 1\} + 0 \cdot P\{y_i = 0\} = P\{y_i = 1\}. \quad (10)$$

Таким образом, получаем равенство:

$$P\{y_i = 1\} = \beta^T X_i. \quad (11)$$

которое дало название модели линейной вероятности (linear probability model).

Следует отметить некоторые недостатки этой модели, которые не позволяют успешно применять метод наименьших квадратов для оценивания параметров β и построения прогнозов. Из (9) следует, что компонента ε_i в каждом наблюдении может принимать только два значения: $(1 - \beta^T x_i)$ с вероятностью $P\{y_i = 1\}$ и $(-\beta^T x_i)$ с вероятностью $1 - P\{y_i = 1\}$. Это, в частности, не позволяет считать случайную компоненту нормально распределенной случайной величиной или, подчиняющейся распределению, близкому к нормальному.

Проверим выполнение условия из первой группы предположений регрессионного анализа о равенстве дисперсий различных наблюдений. Вычислим дисперсию компоненты:

$$D\varepsilon_i = \beta^T x_i(1 - \beta^T x_i).$$

Получается, что дисперсия компоненты ε_i зависит от x_i. Известно, что оценка параметров β, полученная обычным методом наименьших квадратов, в этом случае не является эффективной.
Еще одним серьезным недостатком модели линейной вероятности является тот факт, что прогнозные значения $\hat{y}_i = \hat{\beta}^T x_i$, т. е. прогнозные значения вероятности $P\{y_i = 1\}$, могут лежать вне отрезка $[0, 1]$ (здесь $\hat{\beta}$ — оценка параметра β, полученная методом наименьших квадратов).

Логит и пробит модели бинарного выбора

Откажемся от предположения о линейной зависимости вероятности $P\{y_i = 1\}$ от β. Предположим, что

$$P\{y_i = 1\} = F(\beta^T x_i), \tag{12}$$

где $F(x)$ — некоторая функция, область значений которой лежит в отрезке $[0, 1]$.
Наиболее часто в качестве функции $F(x)$ используют:

1. Функцию стандартного нормального распределения

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz,$$

   в этом случае модель принято называть *пробит моделью*.
2. Функцию логистического распределения

$$\Lambda(u) = \frac{e^u}{1 + e^u}, \tag{13}$$

   тогда модель принято называть *логит моделью*.

Оценивание параметров в логит и пробит моделях

Для нахождения оценок параметров β обычно используют метод максимального правдоподобия, предполагая, что наблюдения $y_1, \ldots, y_n$ независимы. Так как y_i может принимать значения 0 или 1, то функция правдоподобия примет следующий вид:

$$L(y_1, \ldots, y_n) = \prod_{i:y_i=0} (1 - F(\beta^T x_i)) \prod_{i:y_i=1} F(\beta^T x_i). \quad (14)$$

Нетрудно заметить, что

$$L(y_1, \ldots, y_n) = \prod_{i=1}^{n} F^{y_i}(\beta^T x_i)(1 - F(\beta^T x_i))^{1-y_i}.$$

Рассмотрим логарифмическую функцию правдоподобия:

$$\ln L(y_1, \ldots, y_n) = \sum_{i=1}^{n} \Big(y_i \ln F(\beta^T x_i) + (1 - y_i) \ln(1 - F(\beta^T x_i)) \Big). \quad (15)$$

Дифференцируя равенство (15) по вектору β, получаем уравнение правдоподобия, записанное в векторной форме:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i f(\beta^T x_i)}{F(\beta^T x_i)} - \frac{(1-y_i) f(\beta^T x_i)}{1 - F(\beta^T x_i)} \right) x_i = 0, \qquad (16)$$

где $f(x)$ — плотность распределения, соответствующая функции $F(x)$.

Можно показать, что для пробит и логит моделей логарифмическая функция правдоподобия (15) является вогнутой по β функцией и, значит, решение уравнения (16) дает оценку максимального правдоподобия параметра β [3].

Для логит модели уравнение (16) можно существенно упростить, если воспользоваться тождеством $\Lambda'(u) = \Lambda(u)(1 - \Lambda(u))$:

$$\sum_{i=1}^{n}(y_i - \Lambda(\beta^T x_i))x_i = 0. \tag{17}$$

Гессиан для логит модели имеет следующий вид:

$$H = \frac{\partial^2 \ln L}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n}\Lambda(\beta^T x_i)(1 - \Lambda(\beta^T x_i))x_i x_i^T. \tag{18}$$

Заметим также, что гессиан в этом случае отрицательно определен [3], т. е. логарифмическая функция правдоподобия вогнута.

Для пробит модели логарифмическую функцию правдоподобия (15) можно записать в следующем виде:

$$\ln L = \sum_{i:y_i=0} \ln(1 - \Phi(\beta^T x_i)) + \sum_{i:y_i=1} \ln(\Phi(\beta^T x_i)). \tag{19}$$

Тогда условие (16) будет следующим:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i:y_i=0} \frac{-\varphi(\beta^T x_i)}{1 - \Phi(\beta^T x_i)} x_i + \sum_{i:y_i=1} \frac{\varphi(\beta^T x_i)}{\Phi(\beta^T x_i)} x_i,$$

где $\varphi(x) = \Phi'(x)$. Учитывая, что нормальное распределение, как и логистическое, симметрично, $1 - \Phi(\beta^T x) = \Phi(-\beta^T x)$, получаем:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{q_i \varphi(\beta^T x_i)}{\Phi(q_i \beta^T x_i)} x_i = \sum_{i=0}^{n} \lambda_i x_i = 0, \tag{20}$$

где $q_i = 2y_i - 1$, $\lambda_i = q_i\varphi(\beta^T x_i)/\Phi(q_i\beta^T x_i)$.
Для вычисления гессиана в модели пробит анализа будем использовать свойство стандартного нормального распределения: $d\varphi(u)/du = -u\varphi(u)$. Тогда для пробит модели получим следующее выражение для гессиана:

$$H = \frac{\partial^2 \ln L}{\partial\beta^T \partial\beta} = -\sum_{i=1}^{n} \lambda_i(\lambda_i + \beta^T x_i)x_i x_i^T. \qquad (21)$$

Эта матрица также отрицательно определена [1].

Уравнения правдоподобия (17) и (20) являются системой нелинейных (относительно β) уравнений, аналитическое решение которой невозможно найти в явном виде в общем случае, поэтому при ее решении приходится прибегать к численным методам.

Проверка гипотез о значимости параметров логит и пробит моделей бинарного выбора

Для логит и пробит моделей проверка гипотез о наличии ограничений на коэффициенты, в частности, гипотез о значимости одного или группы коэффициентов, может проводиться с помощью любого из трех критериев — Вальда, отношения правдоподобия, множителей Лагранжа [7], [1].
Рассмотрим нулевую гипотезу в виде системы уравнений:

$$H_0: \quad Q\beta = r, \tag{22}$$

где $\beta^T = (\beta_0, \beta_1, \ldots, \beta_k)$, Q — матрица констант, q — число строк матрицы Q, r — вектор констант, которые формируются определенным образом в зависимости от того, какую гипотезу необходимо проверить.
Альтернативная гипотеза $H_1 : Q\beta \neq r$.

Например, рассмотрим пробит модель
$P\{y = 1\} = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$. Для проверки нулевой гипотезы $H_0 : \beta_1 = 0$ система уравнений (22) примет следующий вид:

$$\begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix}.$$

Для проверки гипотезы $\beta_1 = \beta_2 = 0$ система уравнений (22) примет следующий вид:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Критерий Вальда

Пусть мы нашли оценку максимального правдоподобия $\hat{\beta}$ для неизвестного параметра β, и $\hat{V}(\hat{\beta})$ — состоятельная оценка для асимптотической ковариационной матрицы $V(\hat{\beta})$. Статистика критерия Вальда выглядит следующим образом:

$$W = (Q\hat{\beta} - r)^T (Q\hat{V}(\hat{\beta})Q^T)^{-1}(Q\hat{\beta} - r). \tag{23}$$

При справедливости нулевой гипотезы статистика (23) асимптотически подчиняется распределению χ^2 с числом степеней свободы, равным количеству тестируемых параметров, т. е. равным q [7].
Если численное значение статистики W попадет в критическую область $(\chi^2_{1-\alpha,q}; \infty)$, где $\chi^2_{1-\alpha,q}$ — квантиль уровня $1-\alpha$ распределения χ^2 с q степенями свободы, то нулевая гипотеза H_0 отвергается, в противном случае нет оснований ее отвергнуть при уровне значимости приближенно равном α.
Критерий Вальда носит асимптотический характер, и, поэтому, уровень значимости критерия должен быть близок к α при больших объемах наблюдений.

Критерий отношения правдоподобия

Для проверки адекватности пробит и логит моделей бинарных регрессий рассмотрим критерий, основанный на сравнении значений функции правдоподобия в случае, когда максимизация проводится по всем неизвестным параметрам, и при условии, что $Q\beta = r$.
Пусть
$\ln L_1$ — максимальное значение логарифмической функции правдоподобия (15) при условии, что максимизация производится по всем параметрам β без ограничений на параметры;
$\ln L_0$ — максимальное значение логарифмической функции правдоподобия (15) при условии, что $Q\beta = r$.
Очевидно, что $\ln L_1 \geqslant \ln L_0$.
Чем больше разность между значениями функций, тем более оправдано использование регрессионной пробит или логит модели.

Статистика отношения правдоподобия выглядит следующим образом:

$$LR = 2(\ln L_1 - \ln L_0), \tag{24}$$

которая при справедливости нулевой гипотезы асимптотически подчиняется распределению χ^2 с числом степеней свободы, равным q [7], [1].

Для принятия статистического решения находим значение функции правдоподобия $\ln L_1$ в точке $\hat{\beta}$, которая является оценкой максимального правдоподобия для неизвестного параметра β в задаче без ограничений, и $\ln L_0$.

Если численное значение статистики (24) попадет в критическую область $(\chi^2_{1-\alpha,q}; \infty)$, где $\chi^2_{1-\alpha,q}$ — квантиль уровня $1-\alpha$ распределения χ^2 с q степенями свободы, то нулевая гипотеза H_0 отвергается, в противном случае нет оснований ее отвергнуть при уровне значимости приближенно равном α.

Подробное описание критерия множителей Лагранжа для проверки гипотезы $H_0 : Q\beta = r$ можно найти в [7], [1], [3].

Меры адекватности моделей бинарной регрессии

В настоящее время предложено большое количество мер адекватности для моделей бинарной регрессии [1], [3], [5], приведем некоторые из них.

Сумма квадратов остатков SSR вычисляется по формуле:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{F}_i)^2, \tag{25}$$

где $\hat{F}_i = F(\hat{\beta}^T x_i)$. Использование этой меры не может быть математически строго обосновано, поскольку модели бинарной регрессии не удовлетворяют условию равенства дисперсий [2].

Взвешенная сумма квадратов WSSR для моделей бинарной регрессии может быть вычислена по формуле:

$$WSSR = \sum_{i=1}^{n} \frac{(y_i - \hat{F}_i)^2}{\hat{F}_i(1 - \hat{F}_i)}. \tag{26}$$

Как утверждается в работе [2], критерий (26) более предпочтителен, чем критерий (25).

B. Efron предложил аналог R^2 следующего вида [6]:

$$R^2_{Ef} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{F}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{27}$$

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

Квадратичный коэффициент корреляции SCC вычисляется по формуле:

$$SCC = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})\hat{F}_i\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{F}_i - \bar{F})^2}, \tag{28}$$

где $\bar{F} = \sum_{i=1}^{n} \hat{F}_i/n$.

Существует еще одна мера адекватности моделей бинарной регрессии [4], [2]:

$$R^2_{BL} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i\hat{F}_i + (1-y_i)(1-\hat{F}_i)\right), \tag{29}$$

которая представляет собой среднюю вероятность правильного предсказания в соответствии с полученным правилом.
Существуют меры адекватности моделей бинарной регрессии, основанные на сравнении значений функции правдоподобия при различных ограничениях. Например, D. MacFadden предложил индекс отношения правдоподобия следующего вида [4]:

$$LRI = 1 - \frac{\ln L(\hat{\beta})}{\ln L_0}, \tag{30}$$

где $\ln L_0$ — максимальное значение логарифмической функции правдоподобия при $\beta_1 = \cdots = \beta_k = 0$.

Замечание 2

Различные скалярные меры адекватности моделей бинарной регрессии дают различные результаты [2]. Оценка максимального правдоподобия, на которой основаны все выше перечисленные скалярные меры адекватности для моделей бинарной регрессии, не выбирается из условия максимизации критерия адекватности, в отличие от классической модели линейной регрессии (коэффициенты регрессии, найденные методом наименьших квадратов, максимизируют коэффициент детерминации R^2). В случае бинарной регрессии оценка максимального правдоподобия $\hat{\beta}$ максимизирует совместную плотность распределения наблюдаемых случайных величин. Возникает вопрос для исследователя: выбрать лучшую оценку параметров при возможно низком уровне достоверного прогноза или получить наилучшую оценку параметров, максимизирующую выбранную скалярную меру адекватности модели, которая чаще всего не будет являться оценкой максимального правдоподобия?

ROC-кривая и AUC

Рассмотрим таблицу сопряженности

	Фактически	
Модель	положительно	отрицательно
положительно	*TP*	*FP*
отрицательно	*FN*	*TN*

TP (True Positives) — верно классифицированные положительные примеры;
TN (True Negatives) — верно классифицированные отрицательные примеры;
FN (False Negatives) — положительные примеры, классифицированные как отрицательные (ошибка I рода);
FP (False Positives) — отрицательные примеры, классифицированные как положительные (ошибка II рода).
Что является положительным событием, а что – отрицательным, зависит от конкретной задачи.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями, выраженными в процентах:
Доля истинно положительных примеров (True Positives Rate):

$$TPR = \frac{TP}{TP + FN}$$

Доля ложно положительных примеров (False Positives Rate):

$$FPR = \frac{FP}{TN + FP}$$

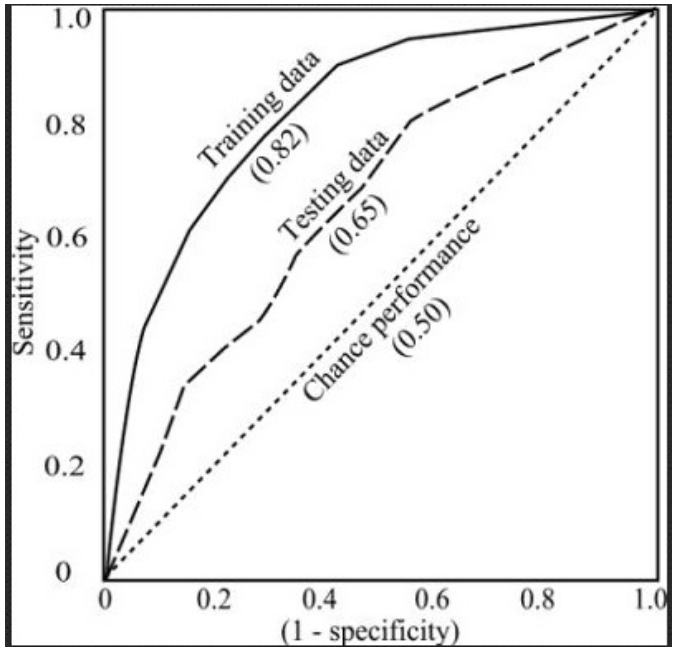
Чувствительность (Sensitivity) – это и есть доля истинно положительных случаев:

$$Se = TPR = \frac{TP}{TP + FN}$$

Специфичность (Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$Sp = 1 - FPR = \frac{TN}{TN + FP}$$

ROC-кривая — кривая зависимости количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров для каждого значения порога отсечения, которое меняется от 0 до 1.



AUC — площадь под графиком ROC-кривой. $AUC \in (0.5;\ 1)$. *AUC* можно интерпретировать как вероятность того, что случайно взятый объект "1"и случайно взятый объект "0"будут отранжированы в правильном порядке.

Литература

- Дрейпер Н., Смит Г.Прикладной регрессионный анализ
- Amemiya T.Qualitive Response Models: A Survey, Journal of Economic Literature, 1981, Vol. XIX, pp. 1483-1536.
- Amemiya T.Advanced Econometrics, Cambridge: Harvard University Press, 1985.
- Ben-Akiva M., Lerman S.Discrete choice analysis, The MIT Press, Cambridge Massachusetts, 1985.
- Berndt E., Hall B., Hall R., Hausman J.Estimation and Inference in Nonlinear Structural Models, // Annals of Economic and Social Measurement, 1974, Vol. 3, 653–665.
- Efron B.Regression and ANOVA with Zero-One Data: Measures of Residual Variation, Journal of American Statistical Association, 1978, Vol. 73, pp. 113-121.

- Engle R. F.Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Intriligator, M. D.; and Griliches, Z. Handbook of Econometrics. II. Elsevier, 1983, pp. 796–801
- Greene W. H.Econometric Analysis, 5th edition, New Jearsey: Pearson Education, 2003
- Kay R., Little S.Assessing the Fit of the Logistic Model: A Case Study of Children with Haemolytic Uraemic Syndrome, Applied Statistics, 35, 1986, pp. 16–30.
- Long J. S.Regression models for categorial and limited dependent variables, Thousand Oaks: Sage Publ., 1997.
- MacFadden D.The Measurement of Urban Travel Demand // Journal of Public Economics, 3, 1978, pp. 303-328.
- Maddala G. S.Introduction to Econometrics, 2nd ed., Macmillan, 1992.
- Тьюки Д.Анализ результатов наблюдений. Разведочный анализ. М.:Мир, 1981