Прикладная статистика. Занятие 11. Логистическая регрессия.

10 мая 2011 г.

Относительный риск (odds ratio)

Относительный риск — величина, показывающая, во сколько раз больше или меньше шансы реализации определённого значения переменной отклика при наличии фактора риска.

$$OR = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}.$$

Пример: курение и воспалительные заболевания органов таза.

Курение	Число больных	Число здоровых	Всего
Да	77	123	200
Нет	54	171	225
Всего	131	294	425

Уравнение логистической регрессии:

$$p = P\left(diseased\right) = \frac{e^{\beta_0+\beta_1\cdot\text{курение}}}{1+e^{\beta_0+\beta_1\cdot\text{курение}}}, \quad \hat{\beta}_0 = -1.1527, \hat{\beta}_1 = 0.6843;$$

$$\widehat{OR} = e^{\hat{\beta}_1} \approx 1.98.$$

Интерпретация: шанс найти больного среди курящих почти в два раза выше.

Доверительный интервал для относительного риска

$$CI_{100(1-\alpha)\%} = \left(e^{\beta_1} - z_{1-\alpha/2} \times \hat{SE}(\hat{\beta}_1); e^{\beta_1} + z_{1-\alpha/2} \times \hat{SE}(\hat{\beta}_1)\right)$$

$\hat{SE}(\hat{\beta}_1)$ оценивается по диагонали матрицы вторых производных функции максимального правдоподобия, и, кроме того, выдаётся любым статистическим пакетом вместе со значением β_1.

$CI_{95\%} = (1.56; 2.40)$.

Фиктивные переменные

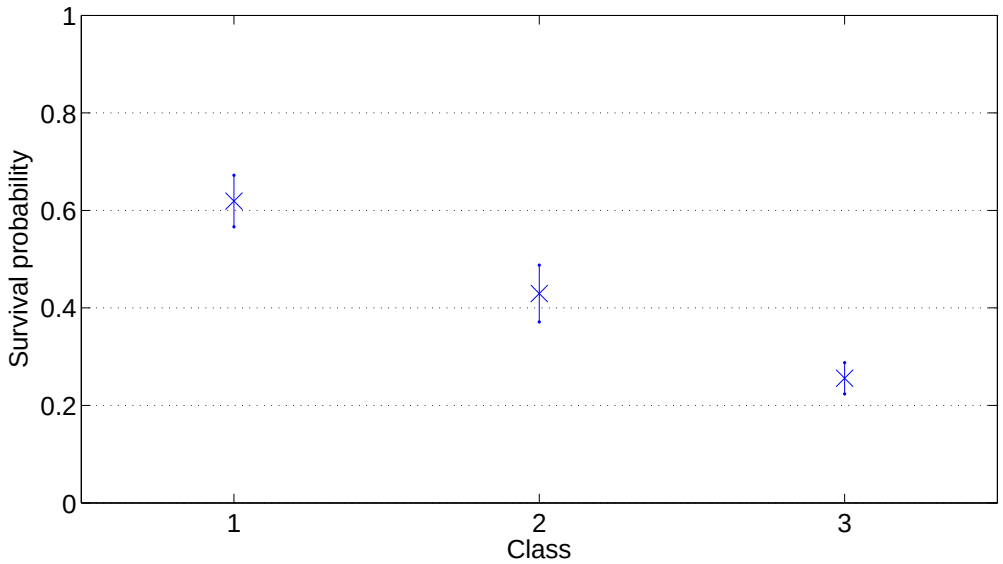
Имеются данные по выживаемости 1308 пассажиров Титаника, о каждом из них известен класс билета.

Выжил	Класс			Всего
	Первый	Второй	Третий	
Да	200	119	181	500
Нет	123	158	527	808
Всего	323	277	508	1308

Категориальный предиктор сводится к набору бинарных фиктивных переменных (dummy variables):

Класс	Фиктивные переменные	
	C_1	C_2
Первый	1	1
Второй	1	0
Третий	0	1

Пассажиры Титаника



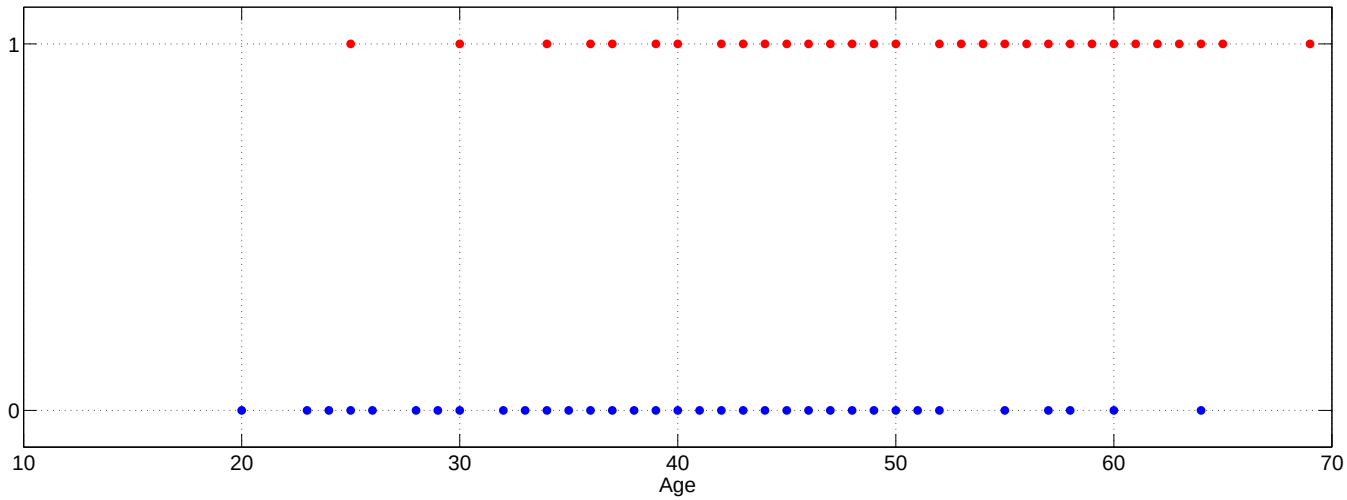
$$p = P\left(survived\right) = \frac{e^{-1.8383+0.7696C_1+1.5548C_2}}{1+e^{-1.8383+0.7696C_1+1.5548C_2}},$$

$$\hat{OR}(Class1, Class2) = e^{0.7696} \approx 2.16,$$

$$\hat{OR}(Class1, Class3) = e^{1.5548} \approx 4.73.$$

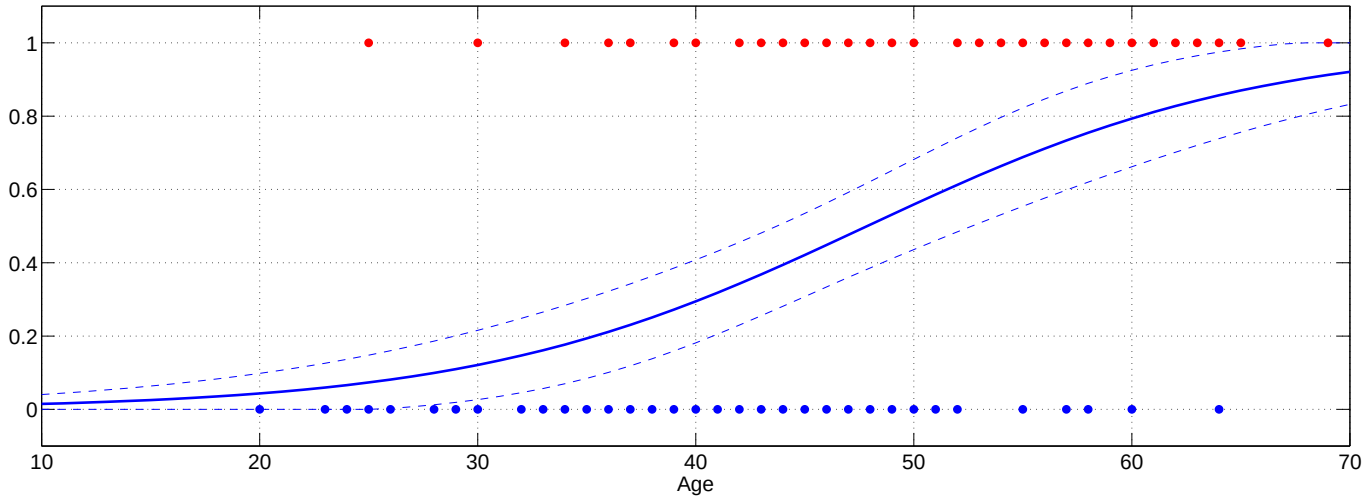
Ишемическая болезнь сердца

По 100 испытуемым известны возраст и наличие ишемической болезни сердца.



$$p = P\left(diseased\right) = \frac{e^{-5.3095+0.1109\times Age}}{1+e^{-5.3095+0.1109\times Age}}.$$

Ишемическая болезнь сердца



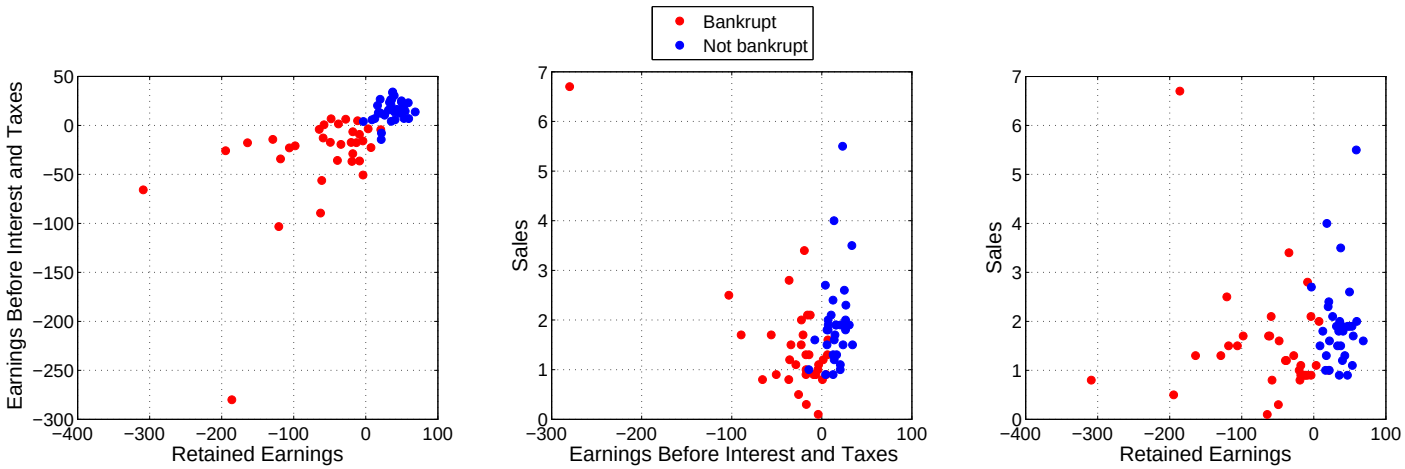
Для определения относительного риска необходимо задать дополнительно константу c и сравнивать отношения шансов для значений предиктора, отстоящих друг от друга на c.

$$\hat{OR}(c, \hat{\beta}_1) = e^{c\hat{\beta}_1};$$

$\hat{OR}(10, \hat{\beta}_1) = e^{10 \cdot 0.1109} \approx 3.03$ — каждый 10 лет риск получить ишемическую болезнь сердца увеличивается более чем втрое.

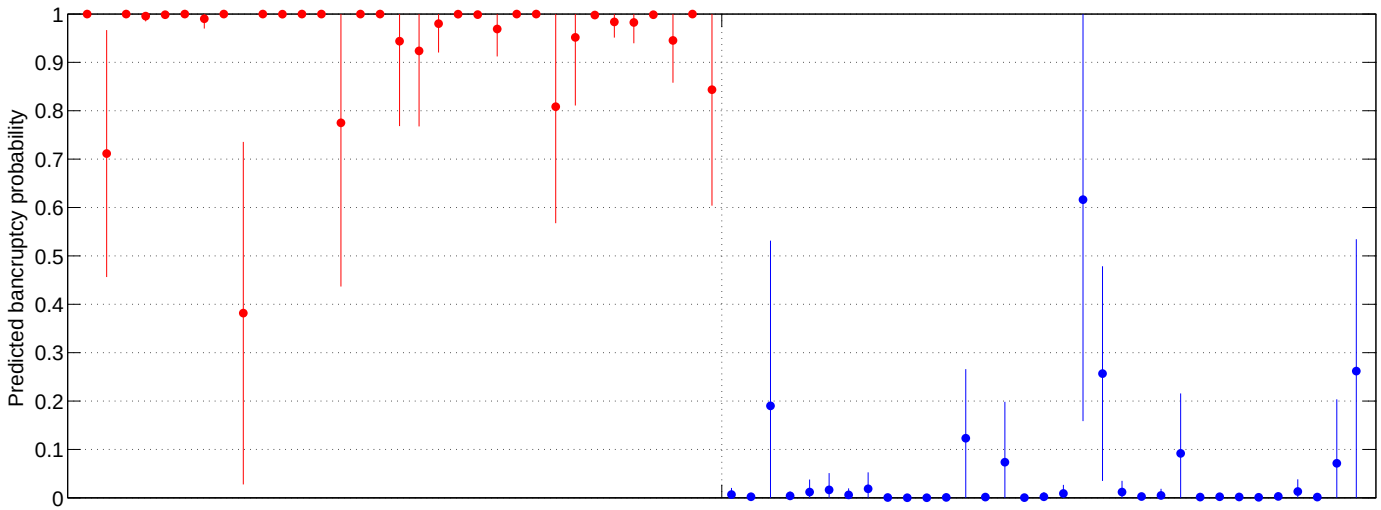
Постановка задачи

Для 66 фирм измерены следующие показатели: отношение полученной прибыли к активам, отношение дохода до вычета прибыли и уплаты процентов к активам, отношение продаж к активам. Известно, что половина этих фирм была признана банкротом в течение двух лет после измерений.



Построить функцию, оценивающую вероятность банкротства.

Результат-1

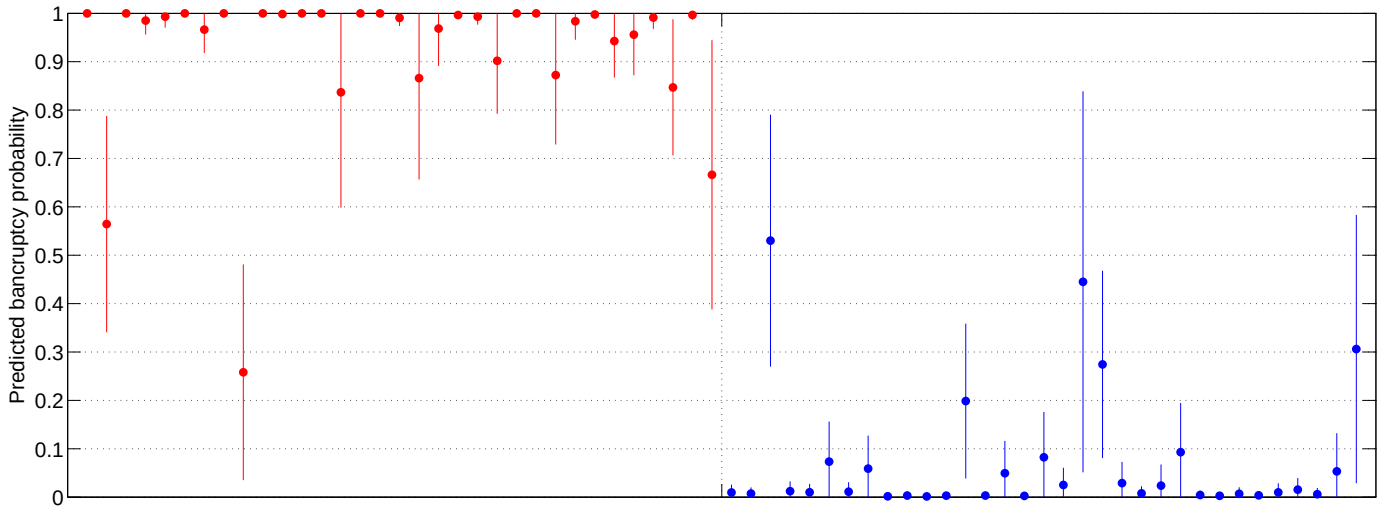


$$p = P\left(bankruptcy\right) = \frac{e^{2.3630-0.0916X_1-0.1010X_2-1.3744X_3}}{1+e^{2.3630-0.0916X_1-0.1010X_2-1.3744X_3}}.$$

	β_0	β_1	β_2	β_3
t	1.7511	-3.7629	-2.5454	-1.7233
p	0.0799	0.0002	0.0109	0.0848

$$LL = \sum_{i=1}^{n} \left(y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i))\right) = -439.3457.$$

Результат-2



$$p = P\left(bankruptcy\right) = \frac{e^{0.2119-0.0792X_1-0.0882X_2}}{1+e^{0.2119-0.0792X_1-0.0882X_2}}.$$

	β_0	β_1	β_2
t	0.4508	-3.8960	-2.5899
p	0.6522	0.0001	0.0096

$$LL = \sum_{i=1}^{n} \left(y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i))\right) = -359.5306.$$

Сравнение моделей

Таблицы классификации при отсечении по вероятности 0.5.

Модель 1:

	Предсказано банкротство	Не предсказано банкротство
Обанкротились	32	1
Не обанкротились	1	32

Модель 2:

	Предсказано банкротство	Не предсказано банкротство
Обанкротились	32	1
Не обанкротились	1	32

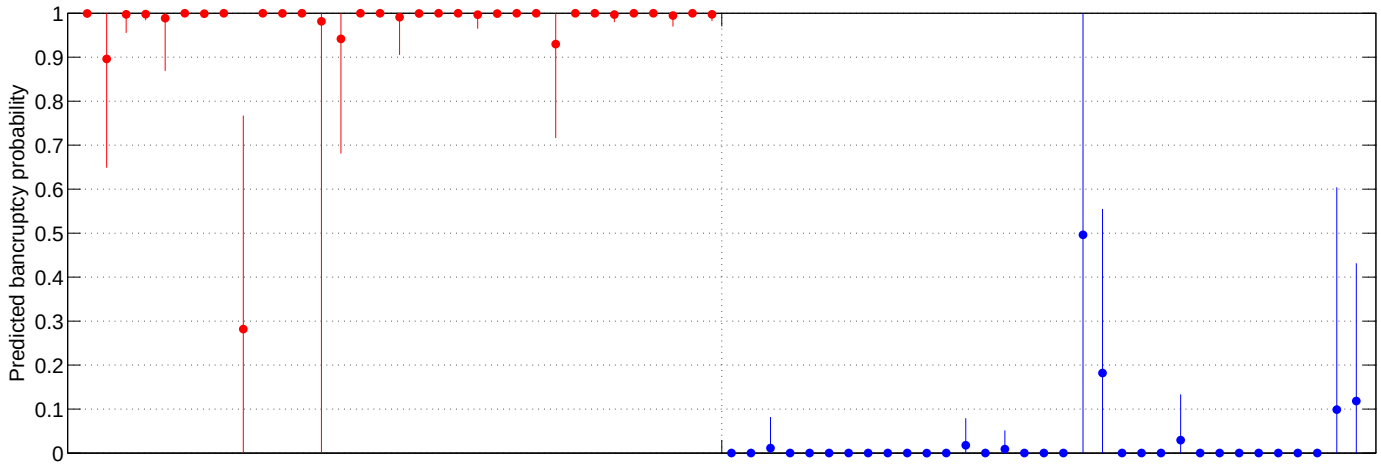
Сравнение моделей

$$G = -2\left(LL_{reduced} - LL_{full}\right).$$

Для расчёта значимости каждой построенной модели вычисляется статистика G, где в качестве редуцированной модели берётся чистая константа. Для такой редуцированной модели
$\beta_0 = \ln(\sum_{i=1}^{n} y_i / \left(\sum_{i=1}^{n}(1 - y_i)\right))$.
При справедливости гипотезы $H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad G \sim \chi^2_p$.
Модель 1 и константа: $G = -2(-45.7477 + 439.3457) = -787.196; \quad p \approx 0.$
Модель 2 и константа: $G = -2(-45.7477 + 359.5306) = -627.5658; \quad p \approx 0.$

Вложенные модели сравниваются аналогично, p — число исключённых из редуцированной модели переменных.
Модели 1 и 2: $= G = -2(-359.5306 + 439.3457) = -159.6302; \quad p \approx 0.$

Взаимодействия



$$p = P\left(bankruptcy\right) = \frac{e^{7.0376-0.2391X_1-0.1336X_2-4.0086X_3-0.0033X_1X_2+0.0274X_1X_3-0}}{1+e^{7.0376-0.2391X_1-0.1336X_2-4.0086X_3-0.0033X_1X_2+0.0274X_1X_3}}$$

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
t	1.9254	-1.7636	0.5737	-1.8581	-2.2130	0.3342	-1.1485
p	0.0542	0.0778	0.5662	0.0632	0.0269	0.7382	0.2508

$$LL = \sum_{i=1}^{n} \left(y_i \ln(\pi(x_i)) + (1-y_i)\ln(1-\pi(x_i))\right) = -796.1910.$$

Пошаговая логистическая регрессия

- **Шаг 0.** Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается G-статистика каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e_1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения p_E (рекомендуется брать не 0.05, а 0.15–0.20).
- **Шаг 1.** Рассчитывается G-статистика и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e_1}. Аналогично принимается решение о включении X_{e_2}.
- **Шаг 2.** Если была добавлена переменная X_{e_2}, возможно, X_{e_1} 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 (если нет боязни построить избыточную модель, можно взять порог 0.9; более строгий вариант — $p_E = 0.15, p_R = 0.20$).
- ...

Прикладная статистика
Семинар 11. Логистическая регрессия.

Рябенко Евгений
riabenko.e@gmail.com