Прикладная статистика. Занятие 9. Логистическая регрессия.

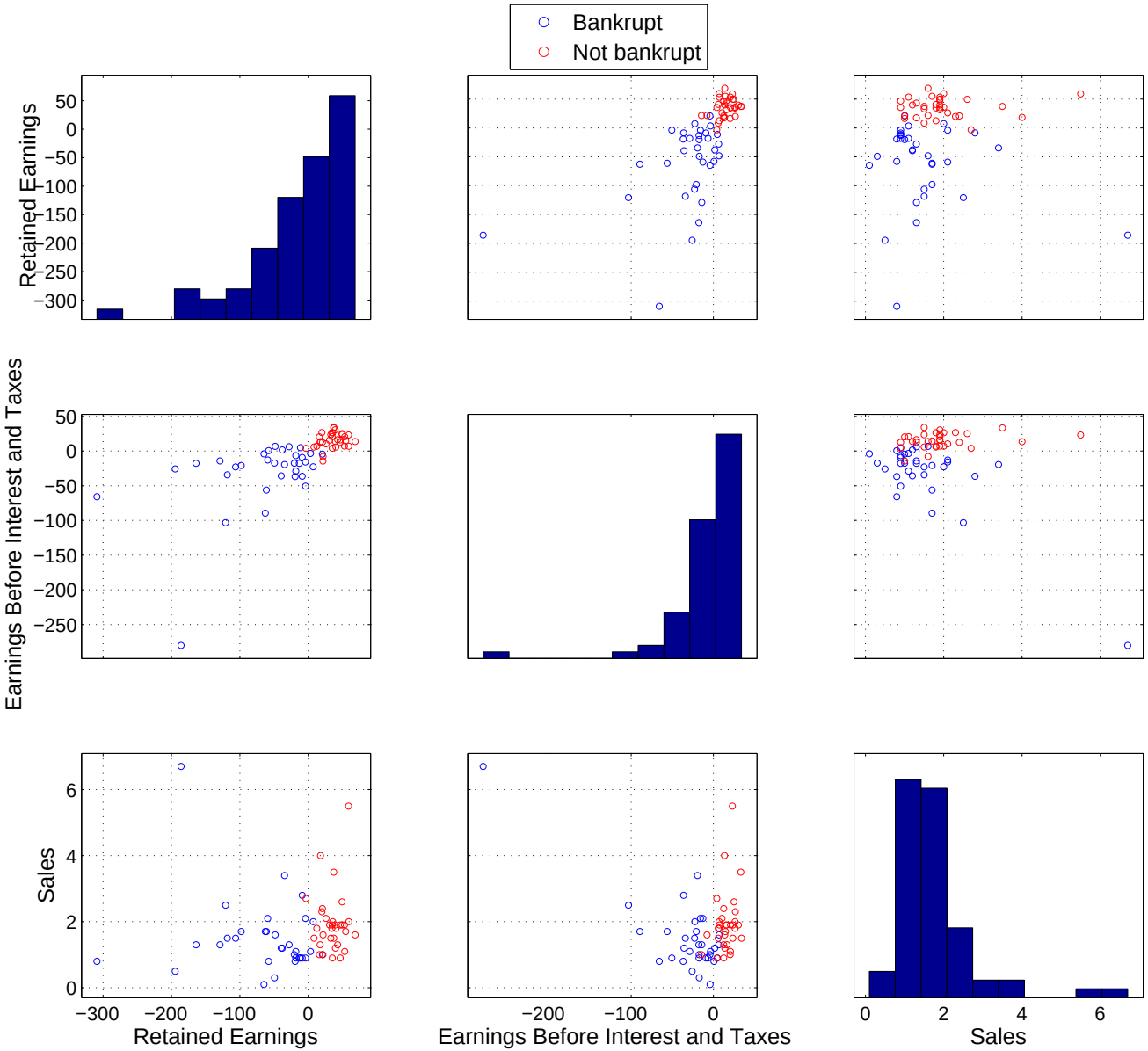
17 апреля 2012 г.

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.

Данные


Bankrupt
Not bankrupt
Retained Earnings
Earnings Before Interest and Taxes
Sales
50
0
−50
−100
−150
−200
−250
−300
6
4
2
0
−300
−200
−100
0
Retained Earnings
−200
−100
0
Earnings Before Interest and Taxes
0
2
4
6
Sales

Корреляции

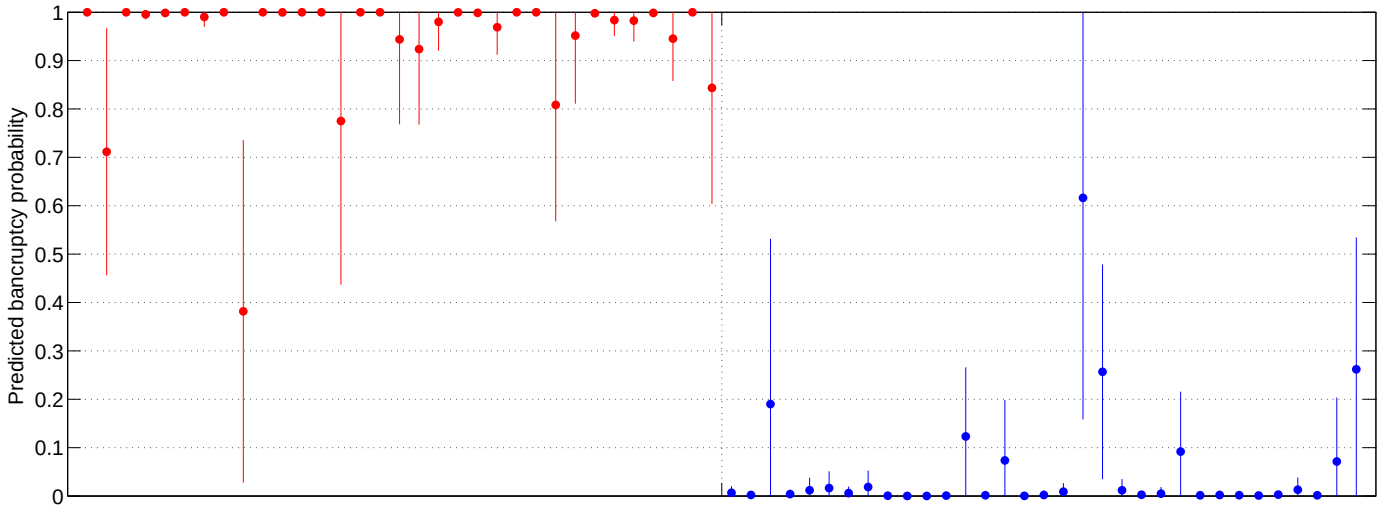
Коэффициенты корреляции Пирсона:

ρ	X_1	X_2	X_3
X_1	1.0000	0.6409	0.0467
X_2	0.6409	1.0000	-0.3501
X_3	0.0467	-0.3501	1.0000

Достигаемые уровни значимости:

p-value	X_1	X_2	X_3
X_1	0	0.0000	0.7094
X_2	0.0000	0	0.0040
X_3	0.7094	0.0040	0

Модель 1



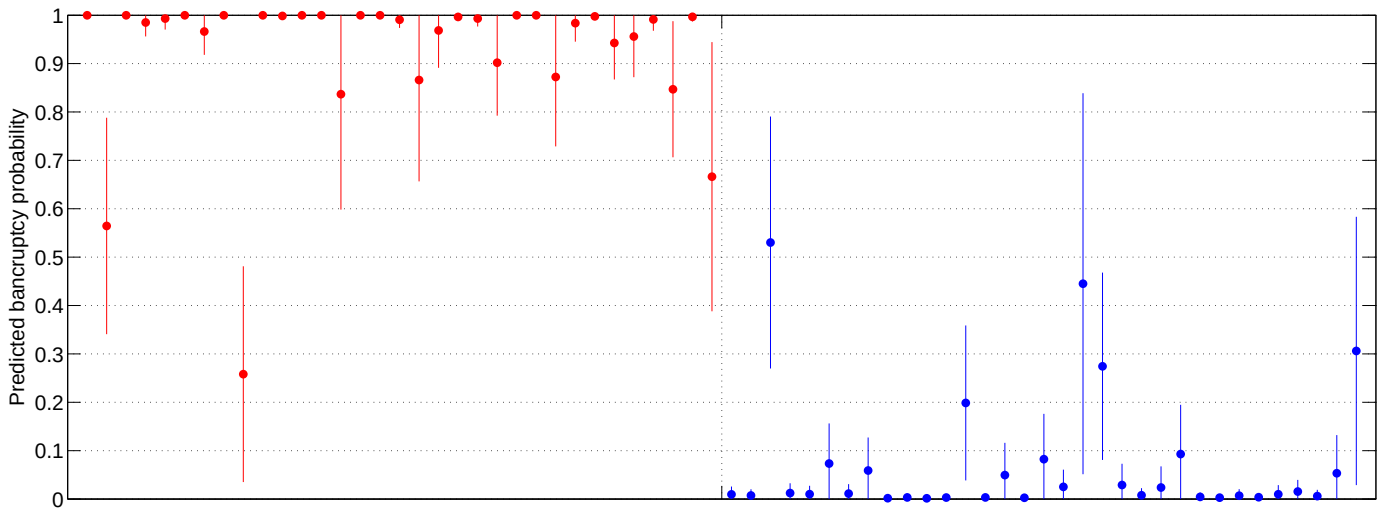
$$p = P\left(bankruptcy\right) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

$$g(x) = -10.1535 + 0.3312X_1 + 0.1809X_2 + 5.0875X_3.$$

	β_0	β_1	β_2	β_3
SE	10.8401	0.3007	0.1069	5.0821
Wald	-0.94	1.10	1.69	1.0
p	0.35	0.27	0.09	0.32

$$LL = -2.906, \quad G = 85.683, \quad p \approx 0.$$

Модель 2



$$p = P\left(bankruptcy\right) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

$$g(x) = -2.1466 + 0.1944X_1 + 1.3911X_2.$$

	β_0	β_1	β_2
SE	1.3125	0.0513	0.8166
Wald	-1.6355	3.7873	1.7034
p	0.1019	0.0002	0.0885

$$LL = -4.736, \quad G = 82.024, \quad p \approx 0.$$

Сравнение моделей

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

Модели 1 и 2: $G = -2(-4.736 + 2.906) = 3.66; \;\; p = 0.0557.$

Сравнение моделей

Переменные	AIC	BIC
$X_1 X_2 X_3$	13.81	22.57
$X_1 X_2$	15.47	22.04
$X_1 X_3$	18.12	24.69
$X_2 X_3$	33.40	39.97
X_1	19.80	24.18
X_2	34.50	38.88
X_3	92.46	96.84
Константа	93.50	95.69

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 — $p_E = 0.15, p_R = 0.20$).
- ...

Пошаговая логистическая регрессия

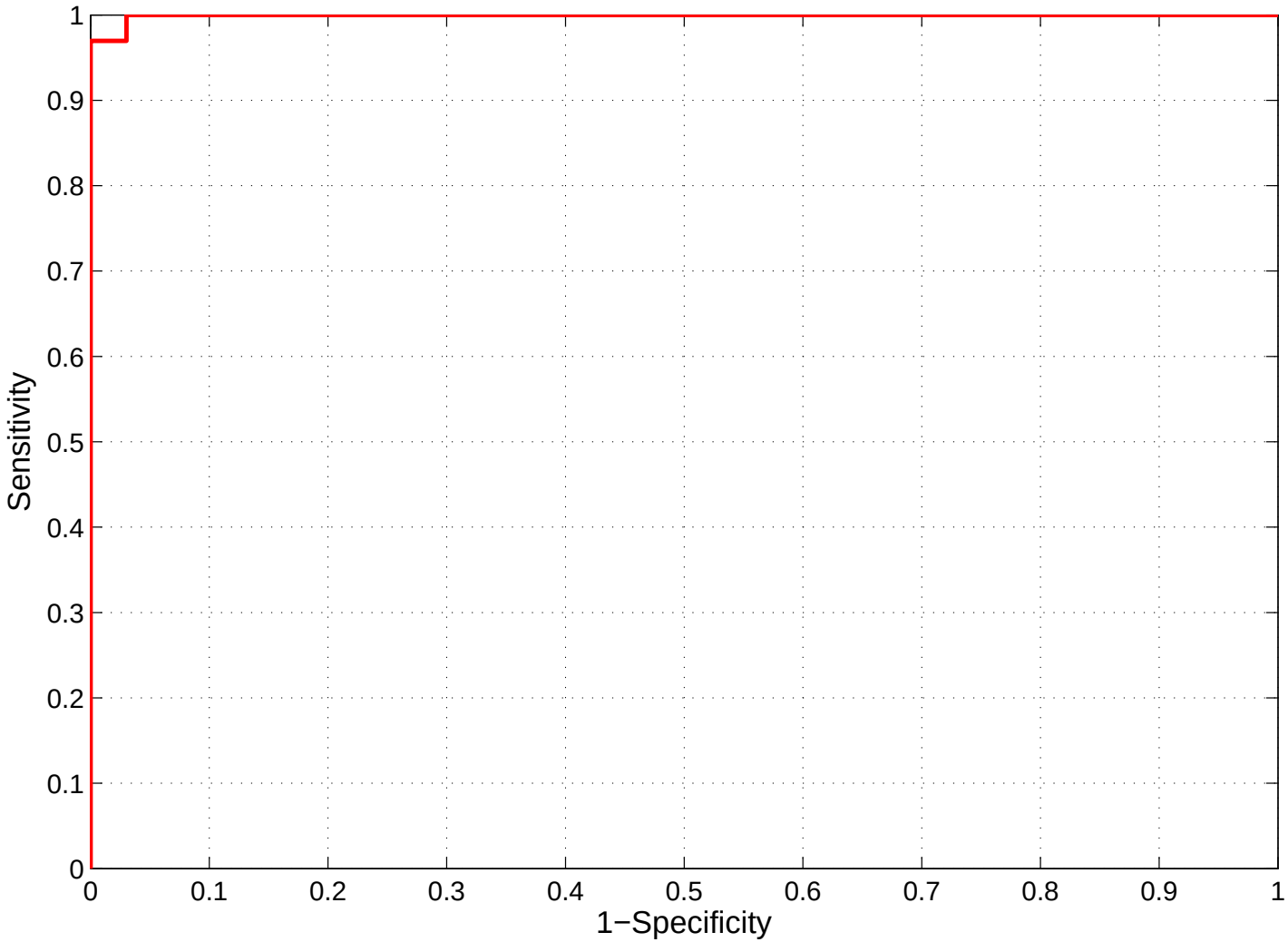
Переменные	LL	G	p
X_1	-7.902	75.692	≈ 0
X_2	-15.250	60.996	≈ 0
X_3	-44.230	3.036	0.0814
Константа	-45.748		

Переменные	LL	G	p
X_1X_2	-4.736	6.332	0.0119
X_1X_3	-6.059	3.686	0.0549
X_1	-7.902		

Переменные	LL	G	p
X_1	-7.902	6.332	0.0119
X_2	-15.250	21.028	4.5×10^{-6}
X_1X_2	-4.736		

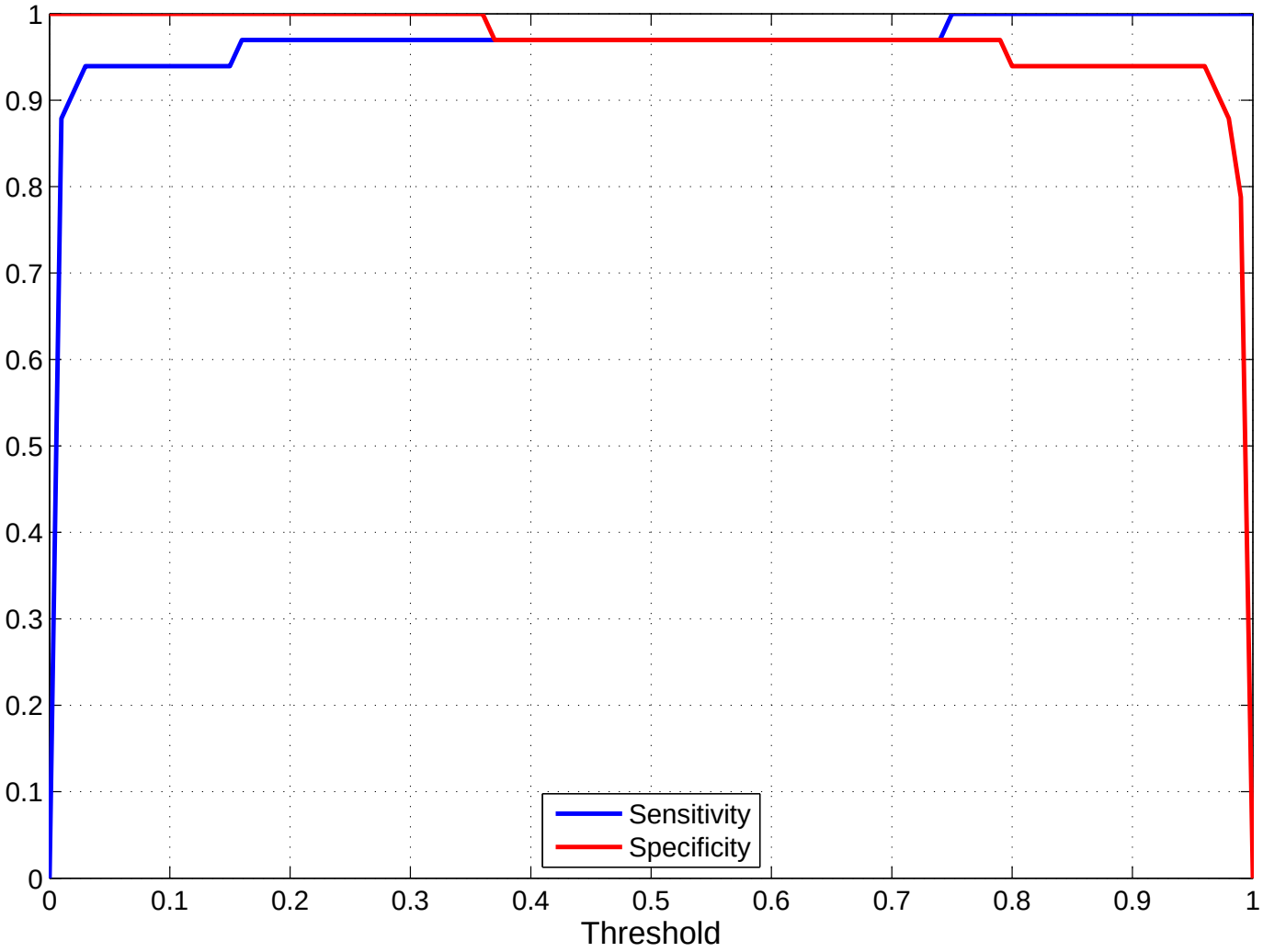
Переменные	LL	G	p
$X_1X_2X_3$	-2.907	3.658	0.0558
X_1X_2	-4.736		

Выбор порога



0.9991

Выбор порога



Прикладная статистика
Семинар 9. Логистическая регрессия.

Рябенко Евгений
riabenko.e@gmail.com