

# Тестовое задание IQMen

12-14 февраля 2015г.

исполнитель: Игорь Колодкин

## Содержание

1. Описание тестовых задач
2. Основные выводы и результаты
3. Решение задачи №1
  - Выявление статистической связи
4. Решение задачи №2
  - Обзор данных
  - Очистка и предобработка данных
  - Построение статистических моделей
    - Построение признакового пространства
    - Обучение модели
    - Оценка модели
  - Выбор лучшей модели
5. Решение задачи №3

# 1. Описание тестовых задач

Задача №1. Выяснить, от каких данных зависят столбцы K, L, M, N.

Задача №2. Построить модель данных любыми способами (приветствуется построение нескольких моделей, основанных на различных методиках)

Задача №3. По построенным моделям предсказать значения зависимых переменных для данных, расположенных на вкладке Test

## 2. Основные выводы и результаты

Задача №1.

Столбец K не зависит ни от одного из столбца A-J.

Столбец L не зависит ни от одного из столбцов A-J.

Столбец M имеет статистическую связь со столбцами A и B и фактором J.

Столбец N имеет статистическую связь со столбцами B, C, D.

Задача №2.

1. Построена регрессионная модель для столбца M:

$$m \sim 1 + a + b + a*b + a^2 + b^2.$$

R-squaared = 0.99, RMSE =0.04.

2. Построен бинарный классификатор на основе логистической регрессии по меткам столбца N. Средняя ошибка классификации(cv 10kfold) = 0.24, F1score = 0.78.

## 3. Решение задачи №1

### Столбец K

Случайная величина(СВ) **K** не зависит ни от одного числового признака A-I (случайные величины A-I). Это видно из верхних графиков рисунка 1.

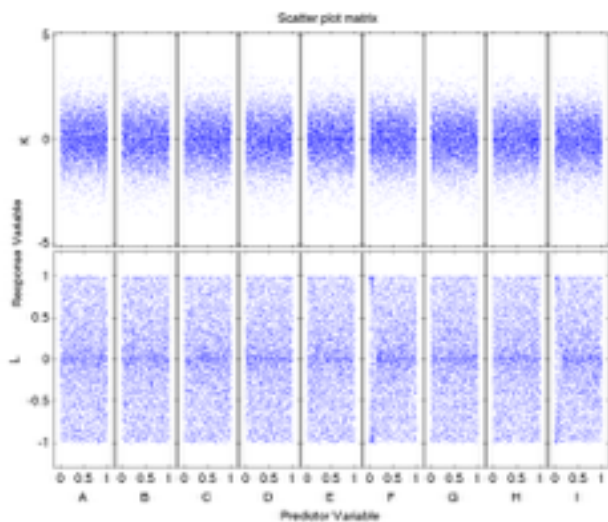


Рис. 1. Графики зависимости СВ K и L от числовых независимых СВ A,B,...,I

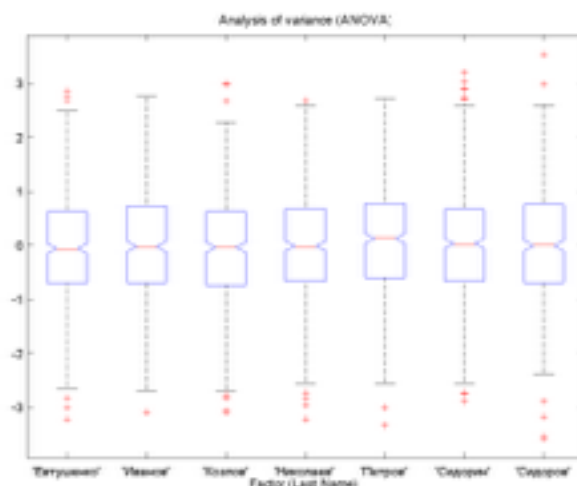


Рис. 2. Boxplot СВ K для семи категорий СВ J

Случайная величина К имеет нормальное распределение согласно критерию Хи-квадрат ( $p\text{-value} = 0,2$ ). Качественный признак J принимает одно из семи значений. Используя его в качестве фактора, проверено условие равенства дисперсий факторных подвыборок СВ К. Выполненные условия позволяют провести однофакторный дисперсионный анализ. Его результат показал, что СВ К также не зависит от качественного признака J. Результат проиллюстрирован на рисунке 2.

**Заключение:** Столбец К не зависит ни от одного из столбца А-J.

## Столбец L

На основе нижних графиков рисунка 1 можно сделать вывод, что СВ L не зависит ни от одного численного признака А-I (случайные величины А-I). Определено, что СВ L имеет ненормальное распределение. Зависимость СВ L от качественного признака J изображена на рисунке 3. Очевидно, что столбец L не зависит от признака J.

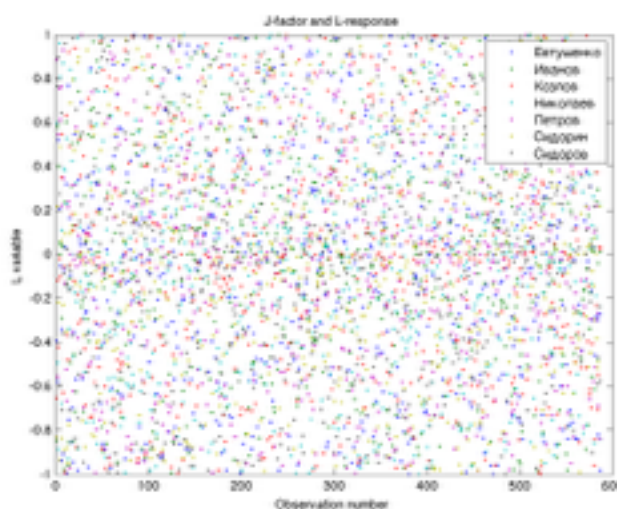


Рис. 3. Диаграмма рассеяния СВ L. По оси x - номер наблюдения, цветом обозначено значение признака J.

**Заключение:** Столбец L не зависит ни от одного из столбцов А-J.

## Столбец M

На диаграммах рисунка 3 показаны зависимости СВ M и количественных СВ А-I, на которых цветом обозначен качественный признак J. Вероятнее,

СВ М имеет статистическую зависимость с СВ А и В, причем значение

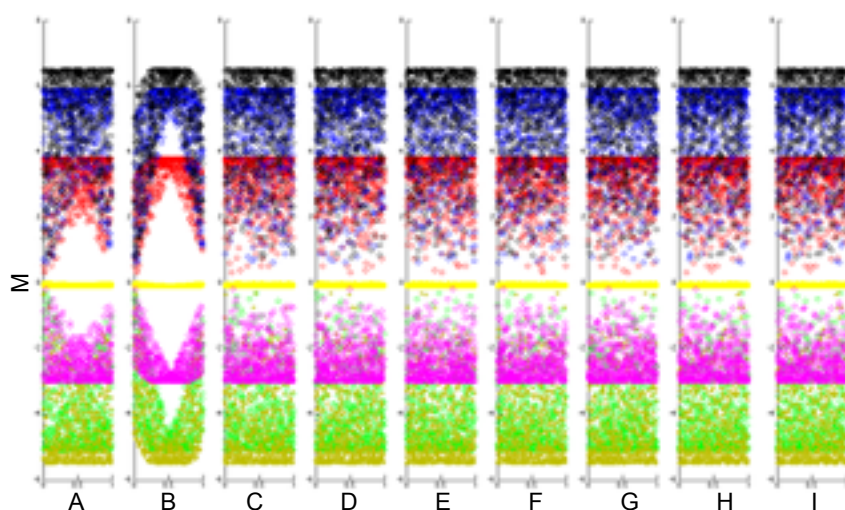


Рис. 3. Диаграммы рассеяния СВ М и независимых СВ А-І.

признака J разделяет выборку на подвыборки, в которых выделяется функциональная зависимость.

**Заключение:** Столбец М зависит от столбцов А и В и фактора J.

## Столбец N

Значения столбца N принимают значения 0 или 1. Образовав два класса, сравним распределения значений их признаков. На рисунке 4 представлены распределения значений признаков, красным цветом обозначен класс 0, синим — класс 1.

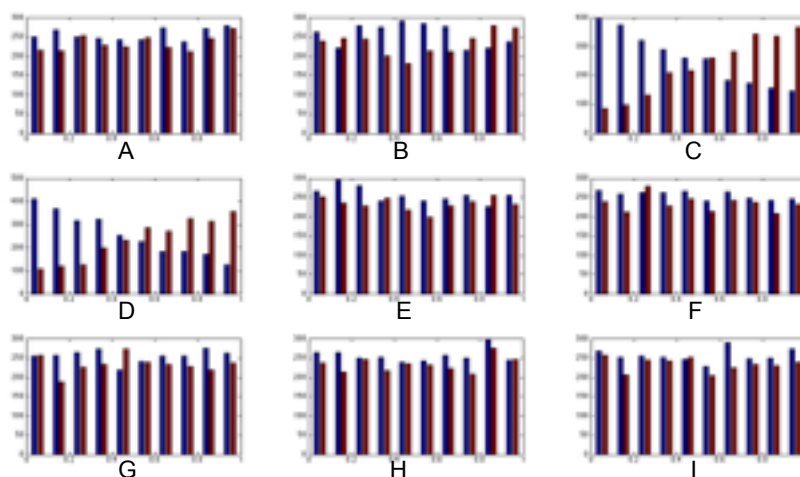


Рис. 4. Гистограммы СВ А-І для классов N=0(красные) и N=1(синие)

Заметны различия у признаков В, С, D, что говорит о их статистической связи с СВ N. Зависимость N от J не обнаружена.

**Заключение:** Столбец N имеет статистическую связь от столбцов В, С, D.



## 4. Решение задачи №2

В одной из работ (“Statistical Modeling : The Two Cultures”) Лео Брейман рассматривает два принципиально разных взгляда на построения статистических моделей. Схема на рис. отражает эти подходы.

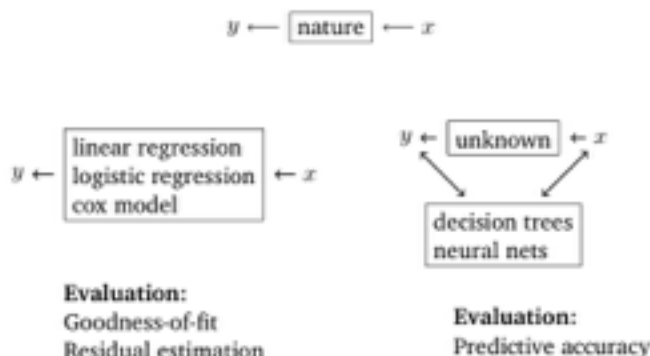


Figure 1.1: The top part of the figure is the way Breiman pictures the world as a model taking input  $x$  and outputting  $y$ . In the bottom part, the two approaches to data analysis are sketched, on the left data modelling and on the right algorithmic modelling. In algorithmic modelling, we do not try to infer nature's true model.

Для решения каждой из поставленных задач, я решил попробовать оба подхода.

### 4.1 Построение модели для М

#### I. Обзор данных

Перед нами задача построения регрессионной модели. Из графиков на рис. 3 очевидно, что качественный признак  $J$  определяет вид зависимости случайной величины  $M$  от признаков  $A$  и  $B$ . Отсюда можно предположить, что один из наилучших вариантов решения сводится к построению нескольких регрессионных моделей. Для каждой категории признака  $J$  строится регрессионная модель на основе признаков  $A$  и  $B$ .

#### II. Очистка и предобработка данных

Не требуется

#### III. Построение статистических моделей

##### A. Построение регрессионной модели отражающей природу данных

Из графиков видно, что фактор  $J$  определяет вид статистической связи случайной величины  $M$  от  $A$  и  $B$ . Отсюда можно предположить, что модель данных будет

состоять из семи регрессионных моделей, каждая из которых соответствует одному из значений фактора. Более того, несложно заметить квадратичную зависимость.

Для каждого значения  $J$  строится регрессионная модель вида:

$M_j(c, a, b) = c_0 + c_1*a + c_2*b + c_3*a*b + c_4*a*a + c_5*b*b + e$ , где  $j=1,...,N_j$ ,  $c_i$  - параметры модели.

### *Построение признакового пространства*

Сформированы следующие обучающие выборки для 7 регрессионных моделей.

J	Объем выборки	Среднее СВ М	СКО СВ М
'Евтушенко'	641	3,10	0,75
'Иванов'	602	-4,08	1,07
'Козлов'	627	4,76	1,25
'Николаев'	622	5,09	1,39
'Петров'	1185	-0,08	0,02
'Сидорин'	636	-4,49	1,09
'Сидоров'	587	-2,42	0,57

### *Анализ моделей*

В таблицу занесены коэффициенты моделей.

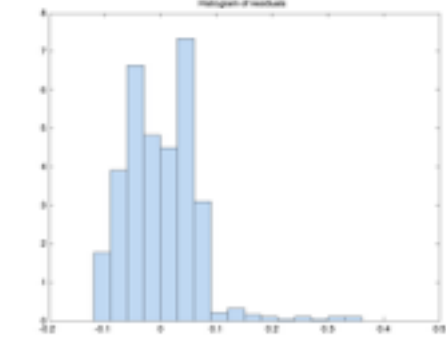
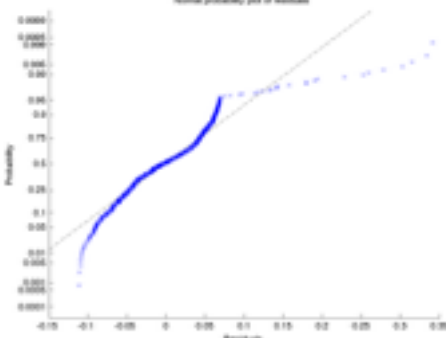
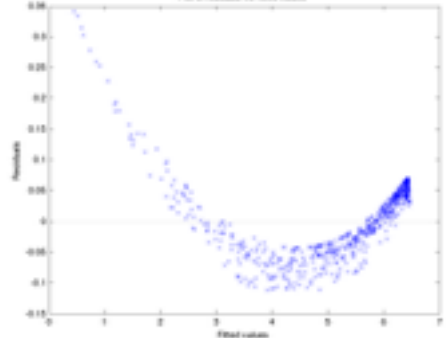
J	$c_0$	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$
'Евтушенко'	-0,35	5,11	10,61	-6,69	-1,55	-6,80
'Иванов'	0,44	-6,86	-14,12	8,92	2,11	9,04
'Козлов'	-0,51	7,89	16,38	-10,32	-2,39	-10,49
'Николаев'	-0,59	8,82	18,06	-11,45	-2,72	-11,55
'Петров'	0,01	-0,14	-0,28	0,18	0,04	0,18
'Сидорин'	0,52	-7,46	-15,39	9,71	2,29	9,86
'Сидоров'	0,30	-4,09	-8,40	5,32	1,25	5,37

Все признаки являются значимыми по критерию Стьюдента.

Оценка моделей

J	Root Mean Squared Error	R-squared	<i>p-value</i>
'Евтушенко'	0,036	0,998	0
'Иванов'	0,054	0,998	0
'Козлов'	0,057	0,998	0
'Николаев'	0,065	0,998	0
'Петров'	0,001	0,998	0
'Сидорин'	0,050	0,998	0
'Сидоров'	0,026	0,998	0
среднее	0,041	0,998	

По оценкам моделей следует, что вариативность данных на 100% процентов объясняется построенными моделями. Модели являются точными. В кросс-валидационной проверке обобщающей способности моделей нет необходимости.

	Анализ остатков модели для одной из модели.
	
	Наблюдаемая зависимость остатков говорит, что модель нельзя назвать адекватной.



## В. Построение алгоритмической модели

Альтернативный вариант построения модели возможен при использовании таких алгоритмов как регрессионные решающие деревья, нейронные сети и др. В этом конкретном примере деревья и сети не представляют интерес. В качестве примера можно использовать модификацию простой регрессинной модели — stepwise regression или шаговая регрессия с жадным добавлением регрессоров.

Мы можем предположить, что у нас нет возможность понаблюдать зависимости характерные в этой задаче, и поэтому строить модель основе все признаков A-I.

Не забудем про J, поэтому модель данных будет также состоять из семи регрессионных моделей, каждая из которых соответствует одному из значений фактора J.

Цель пошаговой регрессии будет состоит в отборе из всех A-I переменных тех, которые вносят наибольший вклад в вариацию зависимой переменной M. Этот процесс выполняет автоматизированная процедура, которая вводит или выводит регрессоры из уравнения регрессии по очереди, основываясь на серии t-тестов.

Начальная модель - константа. Максимальная - линейная по параметрам регрессионная модель в форме полинома третьей степени.

### Анализ моделей

J	Formula
'Евтушенко'	$y \sim 1 + a*b + a^2 + b^2 + (a^2)*b + a*(b^2) + a^3 + b^3'$
'Иванов'	$y \sim 1 + a*b + a*d + a^2 + b^2 + (a^2)*b + (a^2)*d + a*(b^2) + b^3'$
'Козлов'	$y \sim 1 + b + b^2'$
'Николаев'	$y \sim 1 + a*b + a^2 + b^2 + (a^2)*b + a*(b^2) + a^3 + b^3'$
'Петров'	$y \sim 1 + a*b + a^2 + b^2 + (a^2)*b + a*(b^2) + a^3 + b^3'$
'Сидорин'	$y \sim 1 + a*b + a^2 + b^2 + (a^2)*b + a*(b^2) + b^3'$
'Сидоров'	$y \sim 1 + f + a*b + a^2 + b^2 + (a^2)*b + a*(b^2) + b^3'$

## Оценка моделей

J	Root Mean Squared Error	R-squared	<i>p-value</i>
'Евтушенко'	0,036	0,998	0
'Иванов'	0,054	0,998	0
'Козлов'	0,057	0,998	0
'Николаев'	0,065	0,998	0
'Петров'	0,001	0,998	0
'Сидорин'	0,050	0,998	0
'Сидоров'	0,026	0,998	0

Получили более сложную модель с такими же показателями качества. Модели можно упростить отбросив регрессоры с наименьшими по абсолютным значениям коэффициентами и, возможно, прийти к квадратичной форме.

## IV. Выбор лучшей модели

Выбор падает на первую модель, отражающую природу данных.

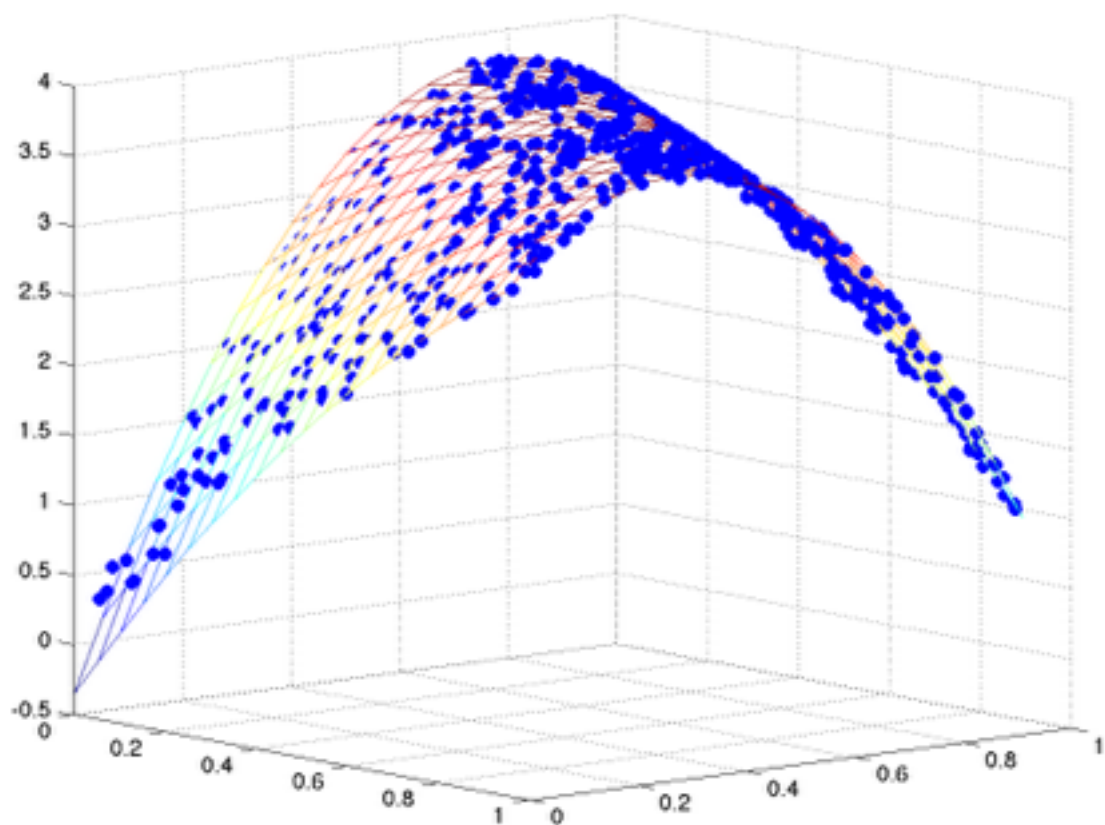
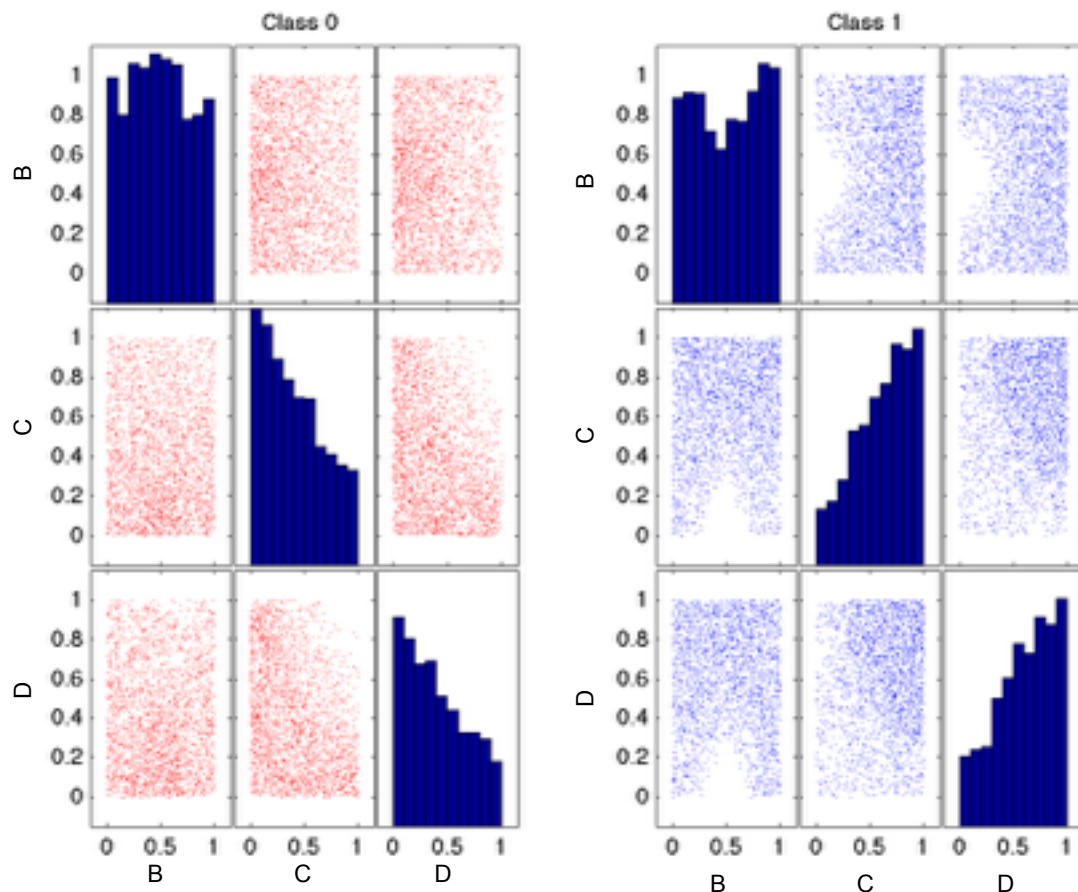


Рис. 5. Регрессия СВ М по А и В, при J = 'Евтушенко'

## Построение модели для N

### I. Обзор данных

На графиках представлены диаграммы рассеяния двух классов для признаков B, C, D. Заметны различия в распределении значений. Классы линейно не разделимы.



### II. Очистка и предобработка данных

Не требуется

### III. Построение статистических моделей

#### A. Построение модели отражающей природу данных

Классификатор на основе дискриминантного анализа.

##### *Метод построения модели классификатора*

Необходимо выбрать тип дискриминантного анализа: линейный или кваддпачичный. В случае линейного подобрать оптимальный коэффициент регуляризации.

Признаки B, C, D. Метки класса - {0, 1}.

### Анализ модели

Среди возможных вариантов был выбран **квадратичный дискриминантный анализ**:  
Classification Discriminant

- PredictorNames: {'B' 'C' 'D'}
- ResponseName: 'N'
- ClassNames: [0 1]
- NumObservations: 4900
- DiscrimType: 'quadratic'

### Оценка модели

Средняя ошибка классификации на всей тренировочной выборке (MCR)= 0.2398

*Кроссвалидационное тестирование.*

При 10 kfold:

средняя кроссвалидационная ошибка классификации = 0.2407,

среднеквадратическое отклонение ошибок = 0.0038

	Предсказанный Класс 0	Предсказанный Класс 1	сумма
Класс 0	2122	441	2563
Класс 1	734	1603	2337
сумма	2856	2044	

$\text{precision} = 2122/2856 = 0.74$

$\text{recall} = 2122/2563 = 0.82$

$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) = 0.78$

## В. Построение алгоритмической модели

### Логистическая регрессия

Коэффициенты линейной модели будут определять разделяющую гиперплоскость для двух классов.

### Метод построения модели

В качестве признакового описания выбраны столбцы A-I. Метки класса - {0, 1}.

Метод построения заключается в настройке параметров модели, такие как вид функции распределения данных, вид связывающей функции и линейная форма. Линейная форма может быть простой линейной:  $y \sim c_0 + c_1 * a + \dots + c_9 * i$ , или быть полиномом n-ой степени. Для определения лучшей формы предложен следующий подход:

1. Инициализировать начальную форму вида  $y \sim c_0 + c_1 * c + c_2 * d$ . Так как на графиках видно, что классы отличаются в распределениях C и D.

2. Используя алгоритм Stepwise Regression наращивать форму, добавляя регрессоры из полной квадратичной формы 9-ти переменных A-I.

### Анализ модели

В результате предложенного подхода получена следующая модель:

$$\log(y) \sim 1 + c + d + c*d + c^2 + d^2$$

	Оценка коэф.	Стандартная ошибка	t-статистика	p-value
1	-2,71	0,16	-17,01	0,00
c	2,90	0,38	7,56	0,00
d	2,43	0,38	6,44	0,00
c*d	0,69	0,30	2,28	0,02
c^2	-1,73	0,31	-5,66	0,00
d^2	-1,41	0,30	-4,70	0,00

Анализ же простейшей линейной модели,

$$y \sim c_0 + c_1*a + \dots + c_9*i,$$

показывает отсутствие связи с рядом признаков и наличие связи у признаков входящих в начальную форму.

	Оценка коэф.	Стандартная ошибка	t- статистика	p-value
1	-2,37	0,12	-20,21	0,00
a	0,04	0,07	0,56	0,58
b	0,12	0,07	1,69	0,09
c	1,40	0,08	18,48	0,00
d	1,32	0,07	17,69	0,00
e	0,04	0,07	0,61	0,54
f	0,01	0,07	0,17	0,86
g	-0,01	0,07	-0,12	0,91
h	0,04	0,07	0,56	0,58
i	-0,02	0,07	-0,27	0,79

### Оценка модели

Средняя ошибка классификации на всей тренировочной выборке (MCR) = 0.2404

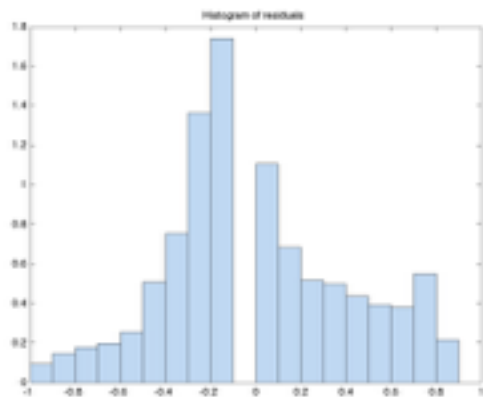
Кроссвалидационное тестирование.

При 10 kfold:

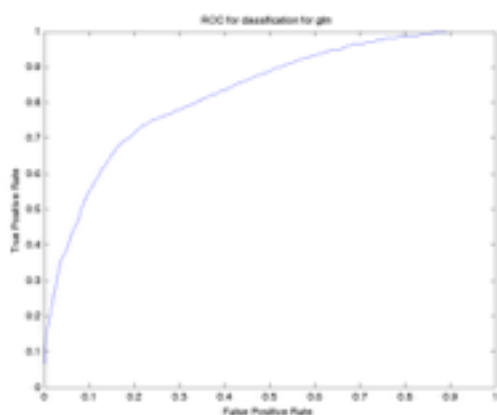
средняя кроссвалидационная ошибка классификации = 0.2406,

среднеквадратическое отклонение ошибок = 0.012

	F1	precision	recall
Класс 0	0.7839	0.7397	0.8338



Гистограмма ошибок классификации



ROC-кривая классификации. AUC = 0,83

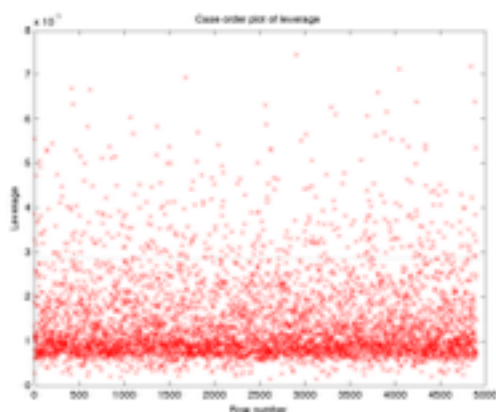


Диаграмма рассеяния ошибок.

## IV. Выбор лучшей модели

Однослойная нейронная сеть с 200 нейронами показала результат хуже.

Качество классификации логистической модели и дискриминантного анализа одинаковое. Думаю, похожие результаты буду у других алгоритмов классификации, что связано напрямую с самой обучающей выборкой, где классы линейно неразделимы и пересекаются.

Выбор падает на модели логистической регрессии, т.к. она позволяет одновременно провести анализ значимости регрессоров.