

Программирование и статистический анализ данных на языке R

Лекция 6 (Основы статистического анализа на языке R)



Петровский Михаил (ВМК МГУ), michael@cs.msu.su

Рассматриваемые модели

Предиктор \ Отклик	Категориальный	Непрерывный	Непрерывный и категориальный
Непрерывный	Дисперсионный анализ (ANOVA)	Регрессия наименьших квадратов (OLS Regression)	Ковариационный анализ (ANCOVA)
Категориальный	таблицы частот	Логистическая регрессия	Логистическая регрессия





Анализ таблиц частот

Цели:





- Рассчитать частоты, проценты и накопленные проценты.
- Распознать наличие ассоциаций между категориальными переменным.

Ассоциации:

- Если зависимость есть, то условное распределение целевой категориальной переменной (частоты) меняется в зависимости от значения категориального предиктора



72%	28%
72%	28%



82%	18%
60%	40%

Таблицы частот, кросстаблицы

- Одномерные

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

- Двумерные (кросстаблицы):

	column 1	column 2	...	column c
row 1	cell ₁₁	cell ₁₂	...	cell _{1c}
row 2	cell ₂₁	cell ₂₂	...	cell _{2c}
...
row r	cell _{r1}	cell _{r2}	...	cell _{rc}

Таблицы частот, таблицы сопряженности, кросстаблицы

Функция	Результат
<code>table(v1,v2,...,vn)</code>	Создается N-мерная таблица сопряженности на основе категориальных переменных v1-vn
<code>xtabs(formula,data)</code>	Создается N-мерная таблица сопряженности на основе формулы и набора данных
<code>prop.table(table,margins)</code>	Пересчет счетчиков ячеек в виде пропорций
<code>margin.table(table,margins)</code>	Суммирование статистик по строкам и столбцам

```
> table(cars$Type,cars$Origin)
```

	Asia	Europe	USA
Hybrid	3	0	0
Sedan	94	78	90
Sports	17	23	9
SUV	25	10	25
Truck	8	0	16
Wagon	11	12	7

```
> xtabs(~ Type+Origin,cars)
```

Type	Origin		
	Asia	Europe	USA
Hybrid	3	0	0
Sedan	94	78	90
Sports	17	23	9
SUV	25	10	25
Truck	8	0	16
Wagon	11	12	7

```
> prop.table(xtabs(~Type+Origin,cars))
```

Type	Origin		
	Asia	Europe	USA
Hybrid	0.007009346	0.000000000	0.000000000
Sedan	0.219626168	0.182242991	0.210280374
Sports	0.039719626	0.053738318	0.021028037
SUV	0.058411215	0.023364486	0.058411215
Truck	0.018691589	0.000000000	0.037383178
Wagon	0.025700935	0.028037383	0.016355140

```
> margin.table(xtabs(~Type+Origin,cars))
```

```
[1] 428
```

```
> margin.table(xtabs(~Type+Origin,cars),1)
```

Type						
	Hybrid	Sedan	Sports	SUV	Truck	Wagon
	3	262	49	60	24	30

```
> margin.table(xtabs(~Type+Origin,cars),2)
```

Origin		
Asia	Europe	USA
158	123	147

Таблицы частот, таблицы сопряженности, кросстаблицы

Функция	Результат
<code>addmargins(table, margins)</code>	Расчет пользовательских статистик по строкам и столбцам
<code>ftable(table)</code>	«Плоская» таблица
<code>as.table(obj)</code>	Преобразование в таблицу
<code>is.table(table)</code>	Проверка типа
<code>as.data.frame(table, row.names = NULL, responseName = "Freq", stringsAsFactors = TRUE)</code>	Преобразование таблицы во фрейм
<pre>> addmargins(xtabs(~Type+Origin, cars), FUN=median) Margins computed over dimensions in the following order: 1: Type 2: Origin Origin Type Asia Europe USA median Hybrid 3.0 0.0 0.0 0.0 Sedan 94.0 78.0 90.0 90.0 Sports 17.0 23.0 9.0 17.0 SUV 25.0 10.0 25.0 25.0 Truck 8.0 0.0 16.0 8.0 Wagon 11.0 12.0 7.0 11.0 median 14.0 11.0 12.5 12.5 > as.table(matrix(c(1,2,3,4),2)) A B A 1 3 B 2 4</pre>	<pre>> ftable(table(cars\$Type, cars\$Origin, + cars\$Cylinders)) 3 4 5 6 8 10 12 Hybrid Asia 1 2 0 0 0 0 0 Europe 0 0 0 0 0 0 0 USA 0 0 0 0 0 0 0 Sedan Asia 0 49 0 41 4 0 0 Europe 0 18 6 34 18 0 2 USA 0 29 0 45 16 0 0 Sports Asia 0 8 0 6 1 0 0 Europe 0 3 0 12 7 0 1 USA 0 0 0 2 6 1 0 SUV Asia 0 5 0 15 5 0 0 Europe 0 0 0 4 6 0 0 USA 0 2 0 11 11 1 0 Truck Asia 0 3 0 4 1 0 0 Europe 0 0 0 0 0 0 0 USA 0 3 0 5 8 0 0 Wagon Asia 0 7 0 3 1 0 0 Europe 0 4 1 4 3 0 0 USA 0 3 0 4 0 0 0</pre>

Таблица частот

```
> cars_wo_hyb <- subset(cars, Type != "Hybrid")
> table_cars <- table(cars_wo_hyb$Origin,
+                     cars_wo_hyb$Type)

> table_cars %>% margin.table()
[1] 425

> table_cars %>% margin.table(1)
```

```
Asia Europe USA
155 123 147

> table_cars %>% margin.table(2)

Sedan Sports SUV Truck Wagon
262 49 60 24 30
```

- Нулевая гипотеза:
 - Нет связи между переменными «по вертикали» и «по горизонтали»
 - В нашем случае – распределение типов кузова производимых машин не зависит от страны производителя.
- Альтернативная гипотеза:
 - Зависимость есть.
 - В нашем случае – распределение типов кузова производимых машин зависит от страны производителя

```
> table_cars
```

	Sedan	Sports	SUV	Truck	Wagon
Asia	94	17	25	8	11
Europe	78	23	10	0	12
USA	90	9	25	16	7

```
> table_cars %>% prop.table() %>% round(4)
```

	Sedan	Sports	SUV	Truck	Wagon
Asia	0.2212	0.0400	0.0588	0.0188	0.0259
Europe	0.1835	0.0541	0.0235	0.0000	0.0282
USA	0.2118	0.0212	0.0588	0.0376	0.0165

```
> table_cars %>% prop.table(1) %>% round(4)
```

	Sedan	Sports	SUV	Truck	Wagon
Asia	0.6065	0.1097	0.1613	0.0516	0.0710
Europe	0.6341	0.1870	0.0813	0.0000	0.0976
USA	0.6122	0.0612	0.1701	0.1088	0.0476

```
> table_cars %>% prop.table(2) %>% round(4)
```

	Sedan	Sports	SUV	Truck	Wagon
Asia	0.3588	0.3469	0.4167	0.3333	0.3667
Europe	0.2977	0.4694	0.1667	0.0000	0.4000
USA	0.3435	0.1837	0.4167	0.6667	0.2333

Нет зависимости

Наблюдаемые частоты = ожидаемым частотам

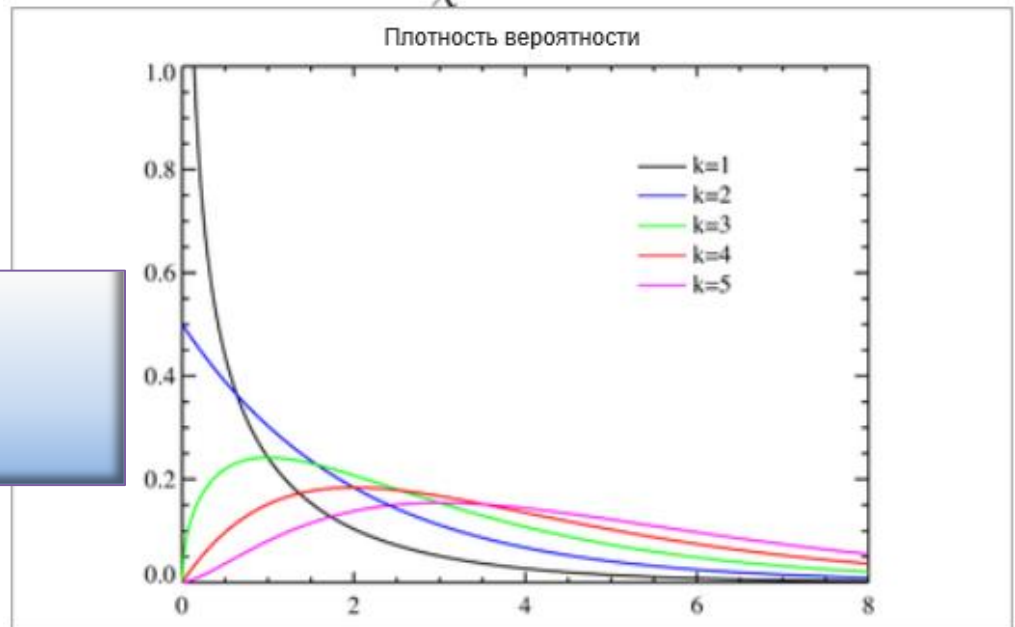
Зависимость

Наблюдаемые частоты \neq ожидаемым частотам

Ожидаемая частота = (всего по строке * всего по столбцу) /
всего по выборке²

Критерий Хи-квадрат

- Проверяет гипотезу о наличии связи
- P-value – не сила зависимости, она **не** измеряется с помощью теста!
- Зависит от размера выборки, есть ограничения на число в ячейках
- Статистика:



- Число степеней свободы = (число строк – 1) * (число столбцов - 1)

Пример

```
> TO_cars <- cars[, c("Type", "Origin")]
> TO_cars[TO_cars$Type != "Truck", "Type"] <- "Other"
> TO_cars[TO_cars$Origin != "USA", "Origin"] <- "Other"

> TO_cars <- cars[, c("Type", "Origin")]
> TO_cars[TO_cars$Type != "Sedan", "Type"] <- "Other"
> TO_cars[TO_cars$Origin != "Asia", "Origin"] <- "Other"
```

```
> chi.t <- chisq.test(table(TO_cars), correct = FALSE)
```

```
X-squared = 11.779, df = 1, p-value = 0.0005991
```

```
X-squared = 0.31255, df = 1, p-value = 0.5761
```

```
> fisher.test(table(TO_cars))
```

```
p-value = 0.001331
```

```
95 percent confidence interval:
```

```
1.627541 11.509447
```

```
sample estimates:
```

```
odds ratio
```

```
4.153281
```

```
p-value = 0.6079
```

```
95 percent confidence interval:
```

```
0.7344629 1.7079507
```

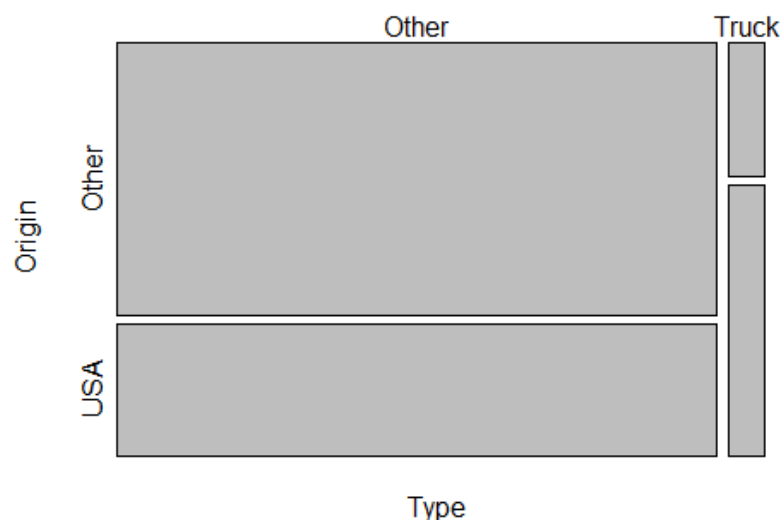
```
sample estimates:
```

```
odds ratio
```

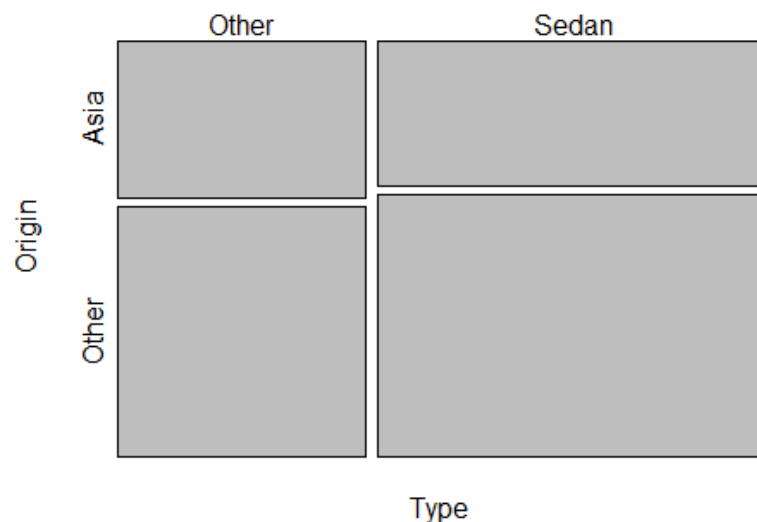
```
1.121098
```

```
> mosaicplot(chi.t$observed, cex.axis = 1)
```

Observed counts



Observed counts



Оценка «силы» ассоциации между предиктором и откликом

- **Шанс** (это не вероятность) – отношение вероятностей события к не событию:

$$Odds = \frac{p_{event}}{p_{nonevent}}$$

- **Отношение шансов** (тоже не вероятность) показывает насколько вероятнее в терминах шансов появления события в группе А (соответствующей набору значений предикторов) по сравнению с другой группой В:

$$Odds_{ratio} = \frac{odds(A)}{odds(B)}$$

Нет зависимости



Группа в **знаменателе** имеет более высокие шансы наступления события

Группа в **числителе** имеет более высокие шансы

0

1



∞

Сравнение вероятностей и шансов

	Заболеел		Total
	Да	Нет	
Прививка	60	20	80
Без прививки	90	10	100
Total	150	30	180

Всего **Заболеел** **Без**
прививки

÷

Всего исходов **Без**
прививки

Вероятность **Заболеел** **Без** прививки
=90÷100=0.9

Сравнение вероятностей и шансов

	Заболеел		Total
	Да	Нет	
Прививка	60	20	80
Без прививки	90	10	100
Total	150	30	180

Вероятность
Заболеел Без
прививки = 0.90

÷

Вероятность Не
заболеел Без
прививки = 0.10

Шанс Заболеть Без прививки =
0.90 ÷ 0.10 = 9

Без прививки шанс заболеть в 9 раз выше чем с прививкой

Сравнение вероятностей и шансов

	Заболел		Total
	Да	Нет	
Прививка	60	20	80
Без прививки	90	10	100
Total	150	30	180

$$\frac{\text{Шанс Заболеть с прививкой}=3}{\text{Шанс Заболеть Без прививки}=9}$$

$$\text{Отношение шансов} = 3 \div 9 = 0.3333$$

Шансов заболеть с прививкой в 3 раза меньше чем без

Точный тест Фишера

- Как все точные тесты рассчитывает точную вероятность получить данное или более экстремальное значение статистики для проверяемой гипотезы («суммы» зафиксированы, «ячейки» варьируем):

	Number of Responders	Number of Non-Responders	Total
Group 1	X_1	$n_1 - X_1$	n_1
Group 2	X_2	$n_2 - X_2$	n_2
Combined	$X_1 + X_2$	$N - (X_1 + X_2)$	$N = n_1 + n_2$

$$\text{prob} = \frac{\binom{n_1}{X_1} \cdot \binom{n_2}{X_2}}{\binom{N}{X_1 + X_2}} = \frac{(X_1 + X_2)! \cdot (N - X_1 - X_2)! \cdot (n_1)! \cdot (n_2)!}{N! \cdot X_1! \cdot X_2! \cdot (n_1 - X_1)! \cdot (n_2 - X_2)!}$$

- Полезен, когда не выполняется условие для аппроксимации распределения гипотезы для Хи-квадрат, например, когда маленькие или «перекошенные» выборки.
- Вычислительно затратный ...

Тест Кохрана-Мантеля-Хензеля (СМН)

Stratum	Group	Responders	Non-Responders	Total
1	1	X_{11}	$n_{11} - X_{11}$	n_{11}
	2	X_{12}	$n_{12} - X_{12}$	n_{12}
	Total	$X_{11} + X_{12}$	$N_1 - (X_{11} + X_{12})$	N_1

2	1	X_{21}	$n_{21} - X_{21}$	n_{21}
	2	X_{22}	$n_{22} - X_{22}$	n_{22}
	Total	$X_{21} + X_{22}$	$N_2 - (X_{21} + X_{22})$	N_2

⋮

k	1	X_{k1}	$n_{k1} - X_{k1}$	n_{k1}
	2	X_{k2}	$n_{k2} - X_{k2}$	n_{k2}
	Total	$X_{k1} + X_{k2}$	$N_k - (X_{k1} + X_{k2})$	N_k

$$NUM_j = \frac{X_{j1} \cdot n_{j2} - X_{j2} \cdot n_{j1}}{N_j}$$

$$DEN_j = \frac{n_{j1} \cdot n_{j2} \cdot (X_{j1} + X_{j2}) \cdot (N_j - X_{j1} - X_{j2})}{N_j^2 \cdot (N_j - 1)}$$

$$H_0: p_1 = p_2$$

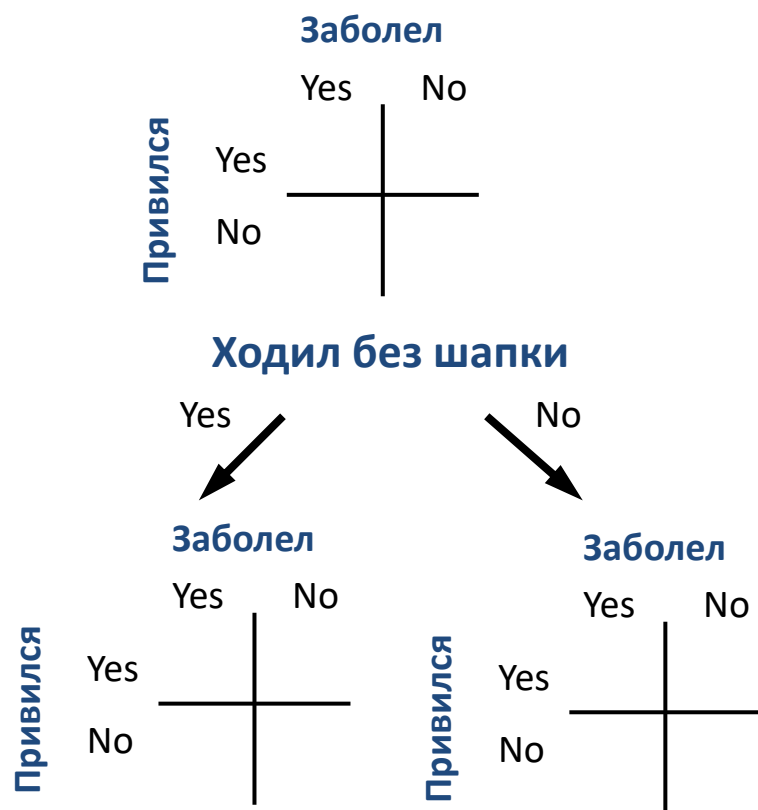
$$H_A: p_1 \neq p_2$$

$$\chi^2_{CMH} = \frac{\left(\sum_{j=1}^k NUM_j \right)^2}{\sum_{j=1}^k DEN_j}$$

Тип предиктора	Тип отклика	Тип теста СМН	Альтернативная гипотеза
Ordinal	Ordinal	1	Линейная зависимость
Nominal	Ordinal	2	Различие ранговых групповых средних
Nominal	Nominal	3	Общая зависимость

Стратифицированные таблицы частот

- Бинарный «предиктор», бинарный «отклик» и категориальная «страта» :



Кохран-Мантель-Хензель (СМН) тест:

- Строит множество таблиц 2 на 2 – по одной на каждое значение переменной страты
- Оценивает наличие связи между предиктором и откликом как взвешенное усреднение отношений шансов в стратифицированных таблицах 2 на 2 с учетом размеров страт
- СМН статистика имеет распределение Хи-квадрат (асимптотически)
- Можно использовать для поиска «скрытых» важных предикторов, сравнивая отношения шансов в стратифицированной и в не стратифицированной таблицах
- Не чувствителен к размеру страт, но чувствителен к размеру всей выборки

Пример (проверка на скрытые зависимости)

- Проверяем гипотезу I, что не в Европе тип привода влияет на то, является ли машина дорогой люксовой?

```
> to_cars <- subset(cars, (DriveTrain != "All") & (Origin != "Europe"))  
> to_cars$Expensive <- to_cars$Invoice > 50000
```

- Получаем ответ «да»:

```
> CMHtest(Expensive ~ DriveTrain, data=to_cars)  
Cochran-Mantel-Haenszel Statistics for DriveTrain by Expensive
```

	AltHypothesis	Chisq	Df	Prob
cor	Nonzero correlation	12.754	1	0.00035519
rmeans	Row mean scores differ	12.754	1	0.00035519
cmeans	Col mean scores differ	12.754	1	0.00035519
general	General association	12.754	1	0.00035519

Пример (проверка на скрытые зависимости)

- Проверяем гипотезу II, за счет стратификации проверяем, что может быть дело не в приводе, а в типе кузова, от которого зависит тип привода?
- Получаем на самом деле для разных типов кузова зависимости высокой стоимости от привода нет!

```
> mantelhaen.test(xtabs(~Expensive+DriveTrain+Type,to_cars))
```

```
Mantel-Haenszel chi-squared test with continuity correction
```

```
data:  xtabs(~Expensive + DriveTrain + Type, to_cars)
Mantel-Haenszel X-squared = 0.16378, df = 1, p-value = 0.6857
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.210783 229.817405
sample estimates:
common odds ratio
      6.96
```

Проверка наличия и силы ассоциации для счетных данных

CMHtest(table)
assocstats(table)

Измерения 3 и дальше образуют страту

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}},$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \quad \phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}.$$

```
> assocstats(xtabs(~Expensive+DriveTrain,to_cars))
```

```
              X^2 df  P(> X^2)
Likelihood Ratio 7.1093  1 0.0076686
Pearson          8.7186  1 0.0031499
```

```
Phi-Coefficient   : 0.187
Contingency Coeff.: 0.184
Cramer's V        : 0.187
```

```
> assocstats(xtabs(~EngineSize+Cylinders,cars))
```

```
              X^2 df P(> X^2)
Likelihood Ratio 871.95 246      0
Pearson          1449.10 246      0
```

```
Phi-Coefficient   : NA
Contingency Coeff.: 0.879
Cramer's V        : 0.753
```

```
> CMHtest(xtabs(~EngineSize+Cylinders,cars))
```

```
Cochran-Mantel-Haenszel Statistics for EngineSize by Cylinders
```

	AltHypothesis	Chisq	Df	Prob
cor	Nonzero correlation	341.04	1	3.7807e-76
rmeans	Row mean scores differ	394.43	41	2.5931e-59
cmeans	Col mean scores differ	363.49	6	1.9615e-75
general	General association	1445.70	246	9.1022e-169

Рассматриваемые модели

<div>Предиктор</div> <div>Отклик</div>	Категориальный	Непрерывный	Непрерывный и категориальный
Непрерывный	Дисперсионный анализ (ANOVA)	Регрессия наименьших квадратов (OLS Regression)	Ковариационный анализ (ANCOVA)
Категориальный	Логистическая регрессия и таблицы частот	Логистическая регрессия	Логистическая регрессия

Логистическая регрессия

Отклик

Подход



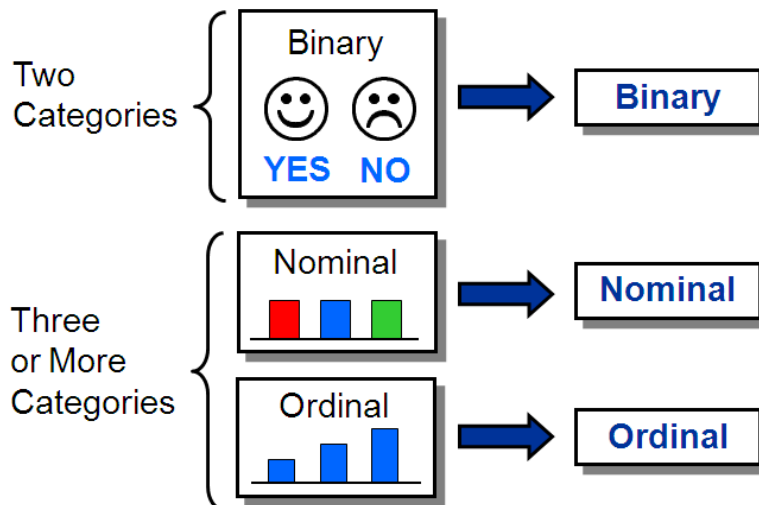
Ординальная (порядковая) регрессия моделирует для k категорий k уравнений регрессии для оценки: $Pr(Y \leq i/x)$, $i=1, \dots, k$

Категориальный случай с k категориями сводится к набору бинарных задач:

- **Каждый против базового (по умолчанию):** $k-1$ уравнений
- Каждый против всех (и голосование или более сложные схемы): k уравнений
- Каждый против каждого (и голосование или более сложные схемы) $k(k-1)/2$ уравнение
- ECOC схемы порядка $k \log(k)$ уравнений

Response Variable

Type of Logistic Regression

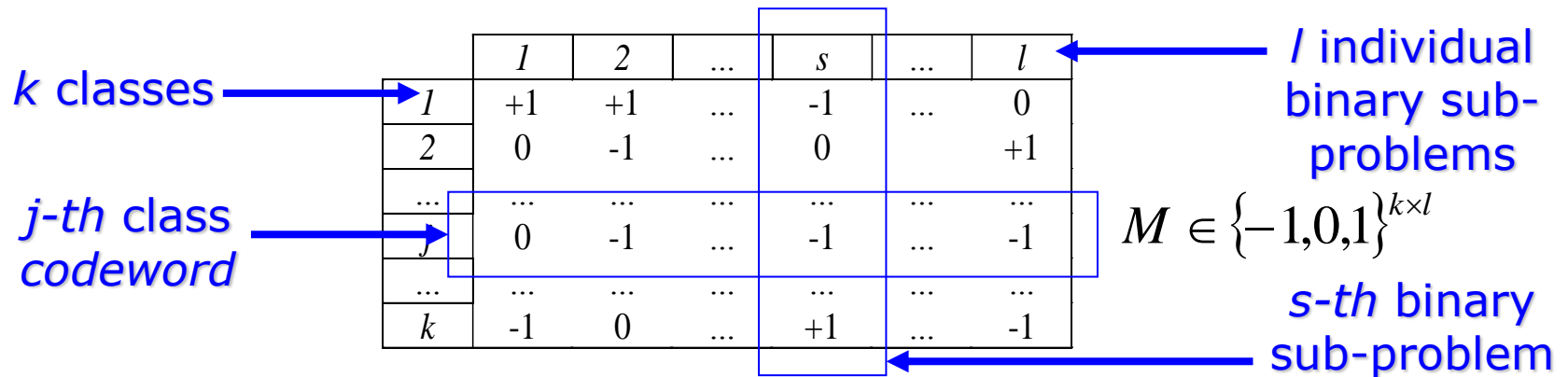


Error Correcting Output Coding (ECOC)

- Предложено в 1995 Dietterich и Bakiri
- Идея из теории информации и телекоммуникаций:
 - В телекоммуникациях: использовать избыточные коды для коррекции ошибок при передачи данных по «зашумленному» каналу
 - В машинном обучении: использовать избыточное число бинарных моделей (кодируется множество классов в супер-классы = группы) для повышения точности классификации, т.е. отклик избыточно кодируется
- Три этапа в ECOC:
 - Coding (кодирование): составление кодовой матрицы (coding matrix) и на ее основе обучающих выборок для бинарных задач
 - Learning (обучение): строятся бинарные модели
 - Decoding (декодирование): прогнозируется отклик (метка класса) на основе индивидуальных прогнозов бинарных классификаторов и кодовой матрицы.

Кодирование в ЕСОС

- Исходная задача с k классами конвертируется в l бинарных подзадач с помощью кодовой матрицы



- Каждый j -й класс имеет кодовое слово, соответствующее строке в матрице M
- Каждая s -я бинарная задача имеет 3 типа классов :
 - “positive”: $I_s^+ = \{y \mid y \in Y \wedge M(y, s) = +1\}$
 - “negative”: $I_s^- = \{y \mid y \in Y \wedge M(y, s) = -1\}$
 - “ignored”: $I_s^0 = \{y \mid y \in Y \wedge M(y, s) = 0\}$

Кодирование в ЕСОС

- “Разреженный” ЕСОС – общий случай:

- “Плотный” ЕСОС – матрица без 0
- “Каждый против всех”:

	1	2	$k-1$	k
1	+1	-1	-1	-1	-1	-1
2	-1	+1	-1	-1	-1	-1
...	-1	-1	+1	-1	-1	-1
...	-1	-1	-1	+1	-1	-1
$k-1$	-1	-1	-1	-1	+1	-1
k	-1	-1	-1	-1	-1	+1

- “Каждый против каждого”:

	1	2	...	$k-1$	k	$k+1$...	$\binom{k}{2}$
1	+1	+1	...	+1	0	0	...	0
2	-1	0	...	0	+1	+1	...	0
...	0	-1	...	0	-1	0	...	0
...	0	0	...	0	0	-1	...	0
...	0	0	...	0	0	0	...	+1
k	0	0	...	-1	0	0	...	-1

- Методы кодирования:

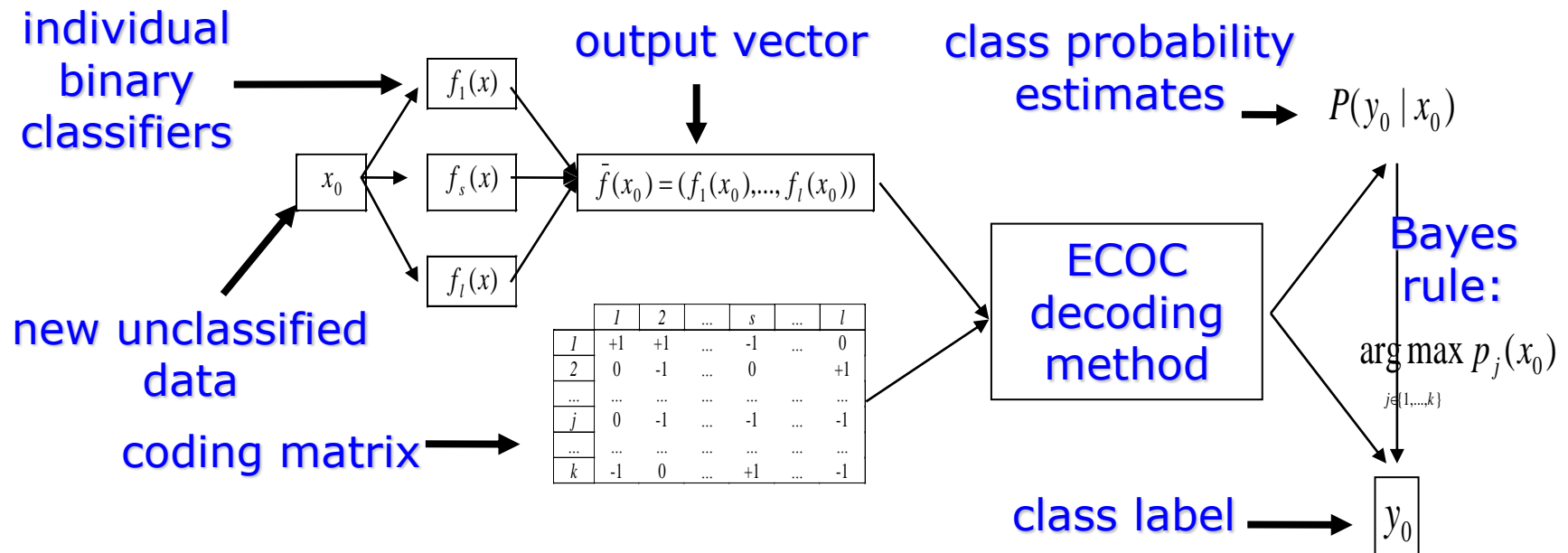
- *Алгебраическая теория кодирования* (коды Хэмминга, например)
- *Задаче-зависимое кодирование*: группы задает эксперт
- *Случайные коды*: случайные длинные «хорошо разделяемые» коды

Обучение в ЕСОС

- l бинарных задач решаются независимо:
 - s -й бинарный классификатор отделяет s -е “положительные” примеры от s -х “отрицательных, так что s -й тренировочный набор:
$$Z_s = \{(x_i, M(y_i, s)) \mid (x_i, y_i) \in Z \wedge (y_i \in I_s^- \vee y_i \in I_s^+)\} \in X \times \{-1, +1\}$$
 - Бинарный алгоритм используется для решения
 - Получаем l бинарных классификаторов (l гипотез) $f_1(x), \dots, f_l(x)$ таких, что $f_s : X \rightarrow Y_{bin}$
- Типы бинарного отклика:
 - Булевый (hard-level): $Y_{bin} = \{-1, 1\}$
 - Вещественный (soft-level): $Y_{bin} \subseteq \mathbb{R}$
 - Вероятностный : $r_s(x) = P(f_s(x) \in I_s^+ \mid f_s(x) \in I_s^+ \cup I_s^-)$

Декодирование в ЕСОС

- Процесс прогнозирования:



- Применить все бинарные классификаторы, получить вектор откликов длины l
- Применить к нему выбранный метод декодирования и получить прогноз

Декодирование в ЕСОС

- На основе расстояний:
 - Поиск ближайшего к вектору откликов кодового слова



- Используются разные метрики:

Хэмминга (hard-level):

Минковского (probabilistic):

$$d_H(\bar{f}(x), M(y)) = \sum_{s=1}^l [1 - \text{sgn}(M(y, s) f_s(x))] \quad d_{L1}(\bar{r}(x), M(y)) = \sum_{s=1}^l |M(y, s) - r_s(x)|$$

- На основе функции потерь
- С оценкой вероятности и т.п.

$$\text{Loss}(\bar{f}(x), M(r)) = \sum_{s=1}^l \text{loss}(M(y, s) f_s(x))$$

Вернемся к бинарным моделям

Почему нельзя моделировать вероятность отклика p как непрерывный отклик с помощью линейной регрессии?

OLS Reg: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

- Если целевая переменная категориальная, как представить ее в виде числовой?
- Если целевая закодирована (1=Yes and 0=No) а результат модели 0.5 или 1.1 или -0.4, что это означает?
- Если переменная имеет только два значения (или несколько), имеет ли смысл требовать постоянства дисперсии или нормальности ошибок?

Linear Prob. Model: $p_i = \beta_0 + \beta_1 X_{1i}$

- Вероятность ограничена, а линейная функция принимает любые значения.
- Принимая во внимание ограниченность вероятности, можно ли предполагать линейную связь между X и p ?
- Можно ли предполагать ошибку с постоянной дисперсией?
- Что такое наблюдаемая вероятность для конкретного наблюдения? 0 и 1?

Логистическая регрессия

Уравнение логистической регрессии:

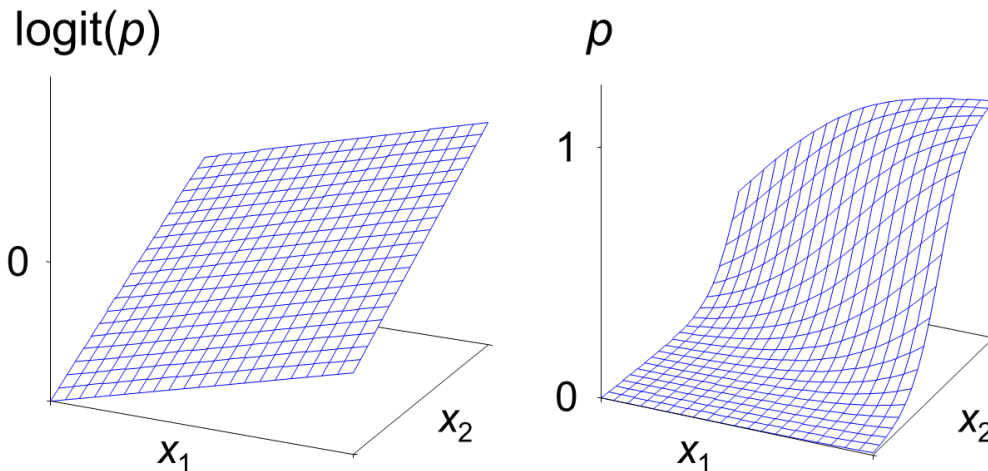
$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

вероятность

параметр

предиктор

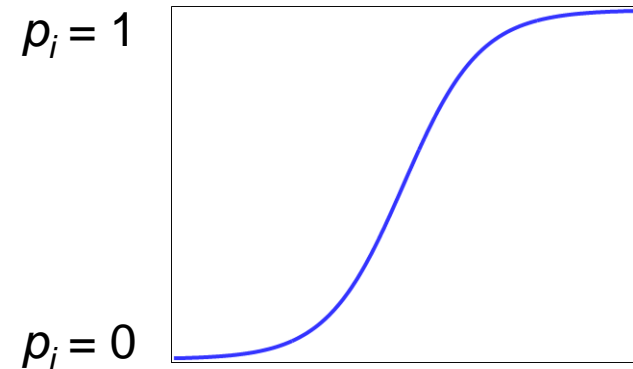
Основное предположение линейной логистической регрессии (линейная зависимость логита от предикторов):



Функция связи (логит) и обратная ей (логистическая):

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta$$

$$\Leftrightarrow p_i = \frac{1}{1+e^{-\eta}}$$



меньше $\leftarrow \eta \rightarrow$ больше

Ограничивает значение отклика

Логистическая регрессия как GLM

```
glm(...,family=binomial(link="logit"))  
wald.test(Sigma, b, ...)  
lrtest(model1,model2)
```

- Целевая задача - максимизация логарифмического правдоподобия

$$\max_{\beta} \left[\sum_{\forall y_i=1} \log(p_i) + \sum_{\forall y_i=0} \log(1-p_i) \right] \quad p_i = \frac{1}{1 + e^{-\sum_j \beta_j x_{ij} + \beta_0}}$$

β_0 = неизвестная константа регрессионного уравнения

β_k = неизвестный параметр $k^{\text{го}}$ предиктора

- Используются и другие функции связи, например, обратная от плотности нормального распределения (пробит регрессия)
- Тест Уальда (тета – MLE модели, W - асимптотически хи-квадрат):

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})}$$

- В wald.test Sigma – ковариационная матрица, b - вектор параметров модели
- lrtest – тест отношения правдоподобия вложенных моделей:

$$LR = 2(l_L - l_S) = 2 \ln \frac{L_L}{L_S},$$

Пример

```
> cars_logist <- subset(cars, DriveTrain != "All")
> cars_logist$DriveTrainEvent <- cars_logist$DriveTrain == "Front"
>
> logit_model <- glm(DriveTrainEvent ~ Invoice + EngineSize + Horsepower + Length +
+                   Weight + Cylinders + Wheelbase + MPG_City + MPG_Highway,
+                   data=cars_logist, family=binomial(link="logit"))
```

тест Вальда для отдельных
предикторов, H_0 : i -й коэф. = 0

```
> summary(logit_model)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-3.0074	-0.3662	0.2860	0.5705	2.5355

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.077e+01	4.678e+00	-2.302	0.021338 *
Invoice	-1.305e-04	2.785e-05	-4.685	2.81e-06 ***
EngineSize	-1.220e+00	5.858e-01	-2.082	0.037302 *
Horsepower	1.560e-02	6.658e-03	2.342	0.019160 *
Length	9.573e-02	2.695e-02	3.552	0.000383 ***
Weight	5.292e-03	9.079e-04	5.829	5.58e-09 ***
Cylinders	-3.625e-01	3.146e-01	-1.152	0.249277
Wheelbase	-2.778e-01	5.796e-02	-4.793	1.65e-06 ***
MPG_City	2.389e-01	1.595e-01	1.498	0.134246
MPG_Highway	2.584e-01	1.136e-01	2.274	0.022969 *

```
> glance(logit_model)
```

```
# A tibble: 1 × 8
```

	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1	420.	333	-120.	260.	298.	240.	324	334

Хи квадрат тесты, H_0 – все коэф. = 0

```
> wald.test(b = coef(logit_model),
+           Sigma = vcov(logit_model), Terms = 1:9)
```

Wald test:

Chi-squared test:

$X^2 = 72.9$, $df = 9$, $P(> X^2) = 4.1e-12$

```
> lrtest(logit_model)
```

Likelihood ratio test

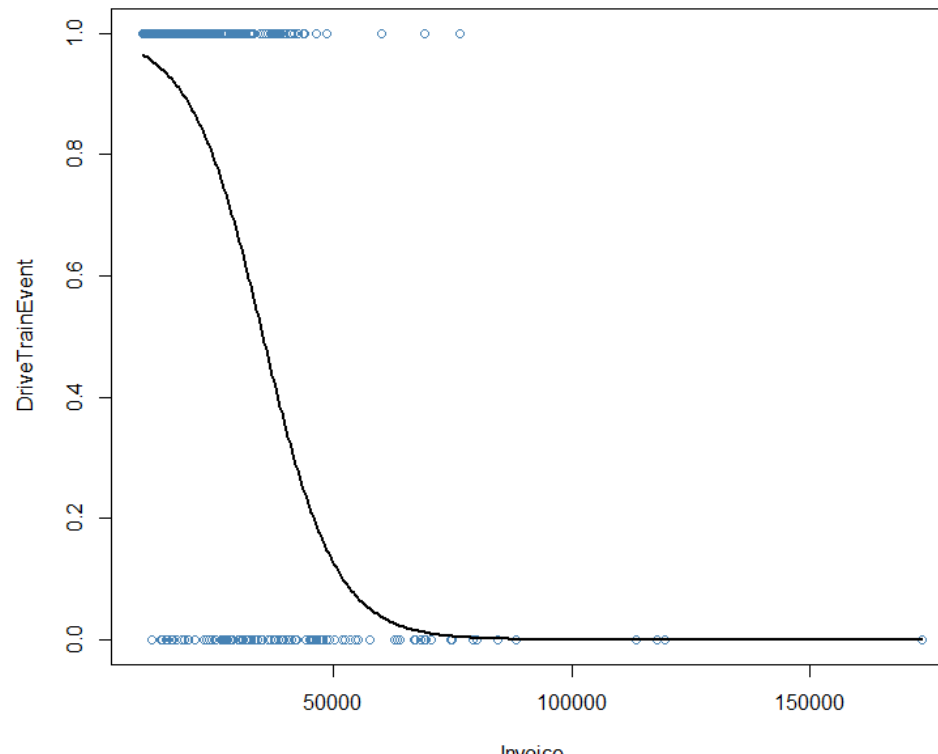
Model 1: DriveTrainEvent ~ Invoice + EngineSize + ...

Model 2: DriveTrainEvent ~ 1

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	10	-120.09			
2	1	-210.21	-9	180.24	< 2.2e-16 ***

Пример

```
> newdata <- data.frame(Invoice=seq(min(cars_logist$Invoice), max(cars_logist$Invoice), len=500),  
+   EngineSize=rep(3.09, len=500),  
+   Horsepower=rep(185.8, len=500),  
+   Length=rep(185.8, len=500),  
+   Weight=rep(3418, len=500),  
+   Cylinders=rep(5.695, len=500),  
+   Wheelbase=rep(107.4, len=500),  
+   MPG_City=rep(20.92, len=500),  
+   MPG_Highway=rep(28.06, len=500))  
> newdata$Probability = predict(logit_model, newdata, type="response")  
> plot(DriveTrainEvent ~ Invoice, data=cars_logist, col="steelblue")  
> lines(Probability ~ Invoice, newdata, lwd=2)
```



Отношение шансов

- Показывает как изменится отношение шансов при изменении i -ой переменной на 1 unit (равно exp от коэф.)

$$\text{logit}(\hat{p}) = \log(odds) = \beta_0 + \beta_i * x_i + \sum_{j \neq i} \beta_j * x_j$$

$$odds = \exp(\beta_0 + \beta_i * x_i + \sum_{j \neq i} \beta_j * x_j)$$

$$\text{logit}(\hat{p}') = \log(odds) = \beta_0 + \beta_i * (x_i + 1) + \sum_{j \neq i} \beta_j * x_j$$

$$odds' = \exp(\beta_0 + \beta_i * (x_i + 1) + \sum_{j \neq i} \beta_j * x_j)$$

$$odds \text{ ratio} = odds' / odds = \exp(\beta_i)$$

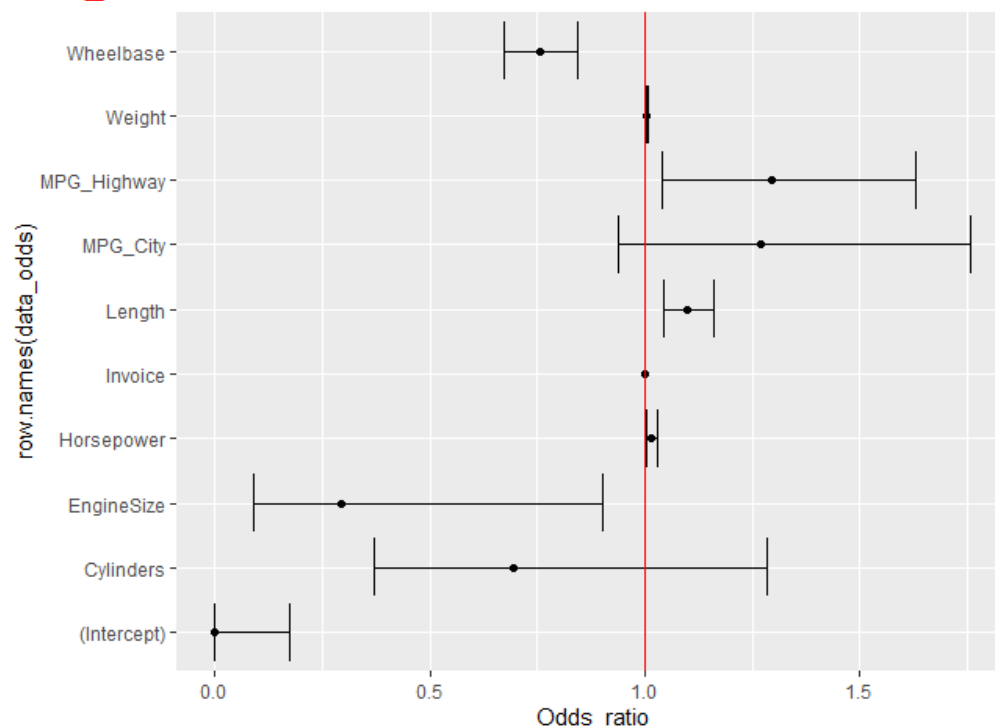
Больше 1 – отношение шансов увеличивается, если меньше, то уменьшается

Отношение шансов (пример)

```
> data_odds <- data.frame(cbind(OR = coef(logit_model)))
> data_odds <- cbind(data_odds, confint(logit_model))
> data_odds <- exp(data_odds) %>% round(3)
> names(data_odds) <- c("Odds_ratio", "Lower", "Upper")
```

```
> data_odds
```

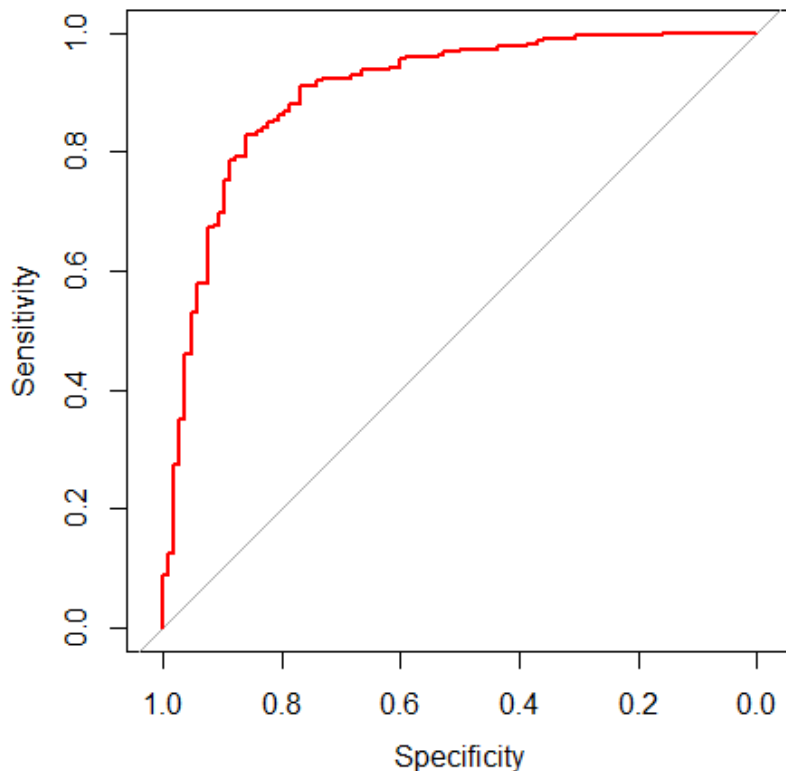
	Odds_ratio	Lower	Upper
(Intercept)	0.000	0.000	0.174
Invoice	1.000	1.000	1.000
<u>EngineSize</u>	0.295	0.090	0.901
Horsepower	1.016	1.003	1.029
<u>Length</u>	1.100	1.044	1.161
Weight	1.005	1.004	1.007
Cylinders	0.696	0.372	1.283
Wheelbase	0.757	0.673	0.845
<u>MPG_City</u>	1.270	0.939	1.757
<u>MPG_Highway</u>	1.295	1.041	1.629



```
> data_odds <- exp(data_odds) %>% round(3)
> names(data_odds) <- c("Odds_ratio", "Lower", "Upper")
>
> ggplot(data_odds, aes(y = row.names(data_odds), x = Odds_ratio)) +
+   geom_point() +
+   geom_errorbar(aes(xmin = Lower, xmax = Upper)) +
+   geom_vline(xintercept = 1.0, colour = "red")
```

Оценка модели

- На основе согласованности всевозможных пар наблюдений (правильной упорядоченности наблюдений в паре), принадлежащих разным классам.
- Чем больше процент согласованных пар тем лучше модель



```
> f1 = roc(logit_model$y ~ logit_model$fitted.values)
> plot(f1, col="red")

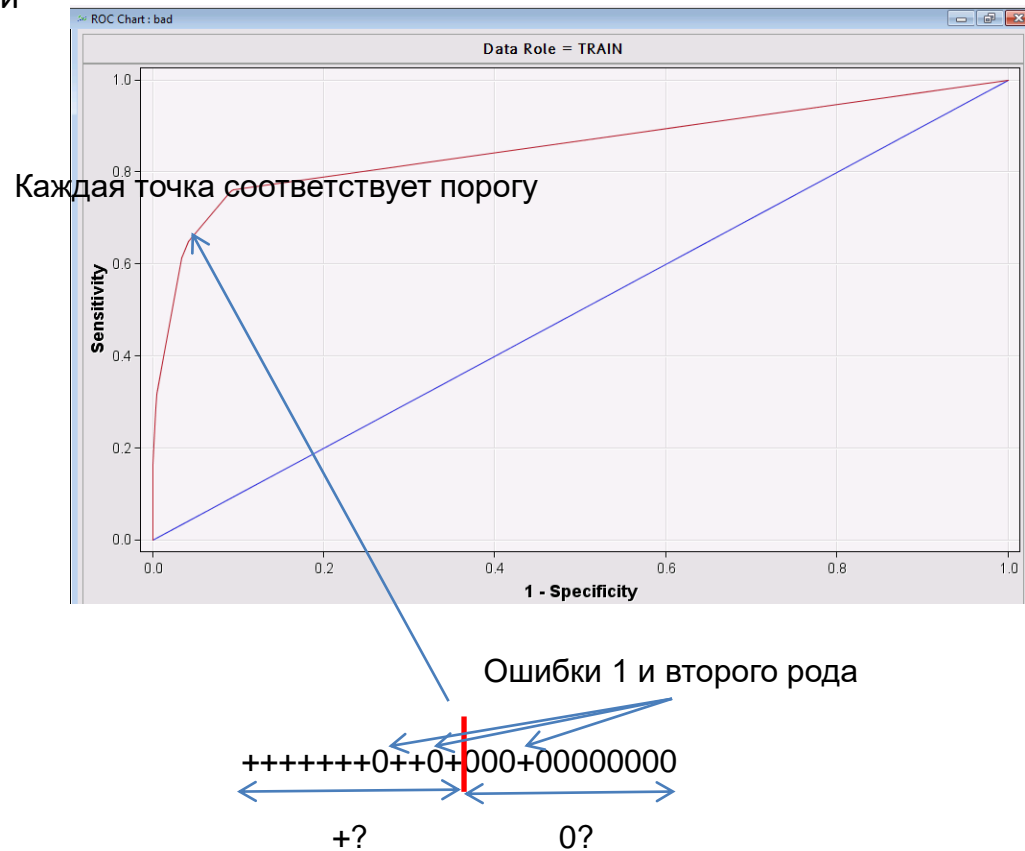
> Evaluate <- function (actuals, predictedScores){
+   res <- Concordance(actuals, predictedScores)
+   res$tau <- (res$Concordance - res$Discordance)*res$Pairs /
+             (length(actuals) * (length(actuals) - 1) / 2)
+   res$AUC = res$Concordance + 0.5*res$Tied
+   res$Somers = 2*res$AUC - 1
+   return (lapply(res, round, 5))
+ }
>
> Evaluate(logit_model$y, logit_model$fitted.values)
```

	Concordance	Discordance	Tied	Pairs	tau	AUC	Somers
1	0.90294	0.09706	0	24408	0.35371	0.90294	0.80588

roc(data, response, predictor,...)
Concordance(actuals, predictedScores)

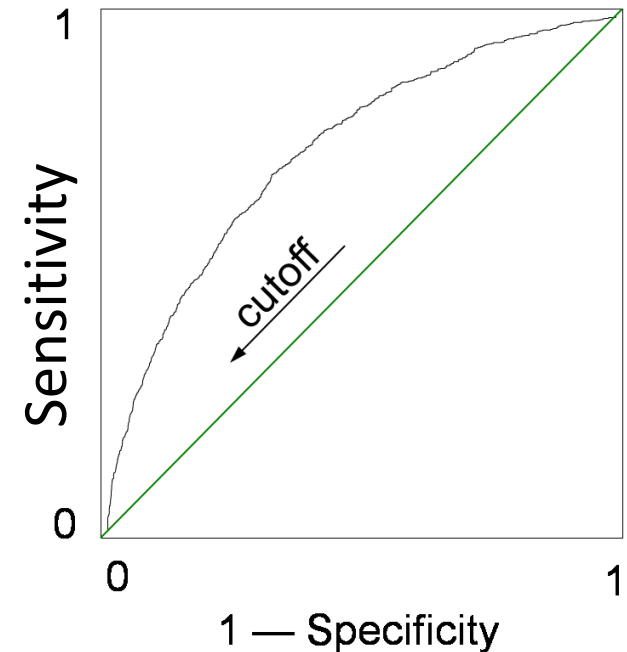
ROC кривая и AUC

- Процедура построения:
 - Сортируем (например, слева направо) набор по убыванию спрогнозированной оценки (вероятности положительного отклика)
 - Идем порогом отсечения по отсортированному набору (слева направо)
 - Для каждого положения порога считаем:
 - отношение числа положительных примеров «слева» от порога к числу всех положительных примеров – detection rate
 - отношение числа отрицательных примеров «слева» от порога к числу всех отрицательных примеров – false positive
 - Ставим точку на графике



Оценка модели

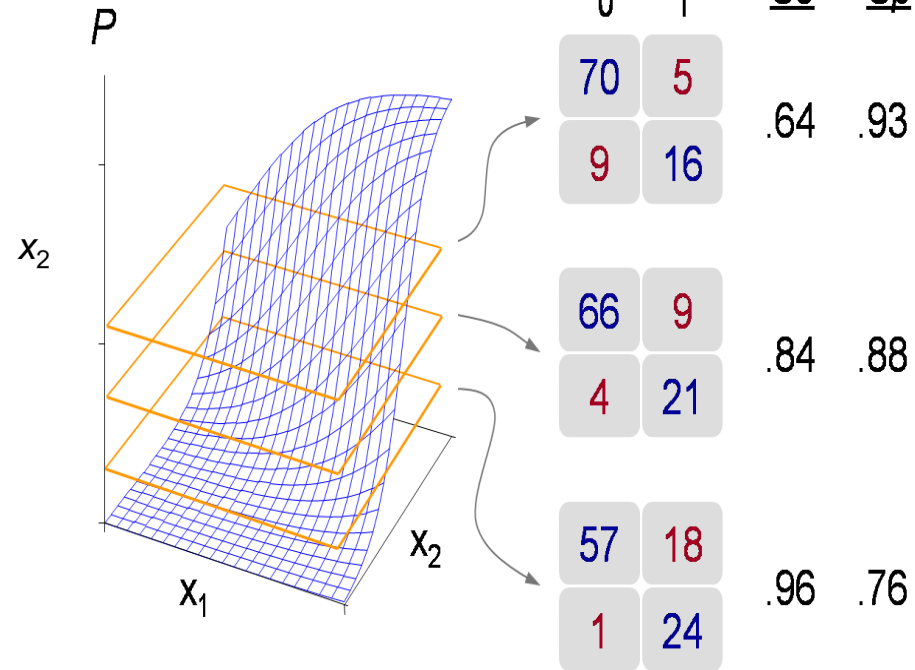
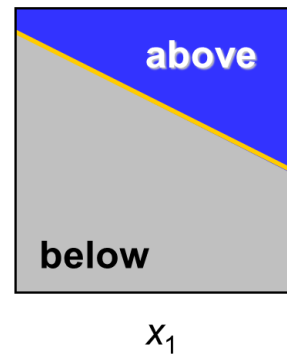
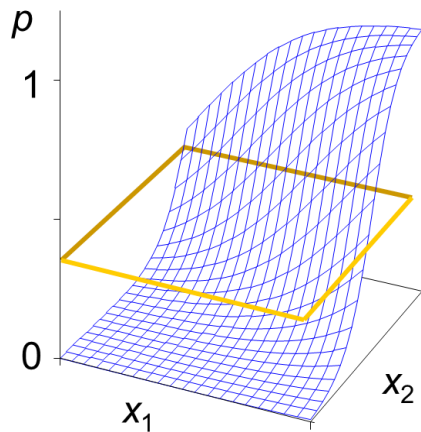
		Прогноз		
		0	1	
Реальность	0	True Negative	False Positive	Actual Negative
	1	False Negative	True Positive	Actual Positive
		Predicted Negative	Predicted Positive	



SENSITIVITY (true positive rate (TPR),
hit rate, recall)
$$TPR = TP / (TP + FN)$$

SPECIFICITY (SPC) (true negative
rate (TNR))
$$SPC = TN / (FP + TN)$$

Оценка модели



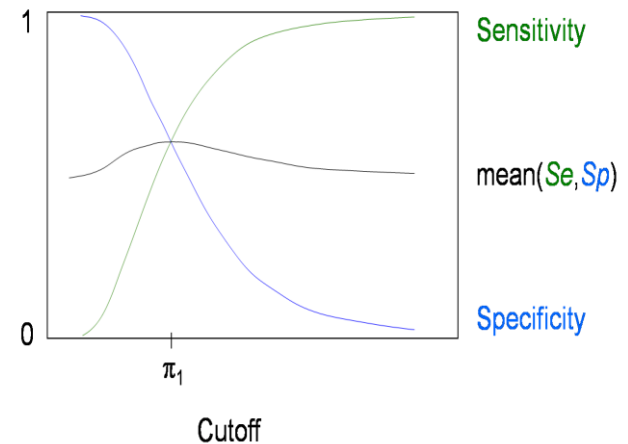
Матрица выигрыша-проигрыша:

		Decision	
		0	1
Actual Class	0	δ_{TN}	δ_{FP}
	1	δ_{FN}	δ_{TP}

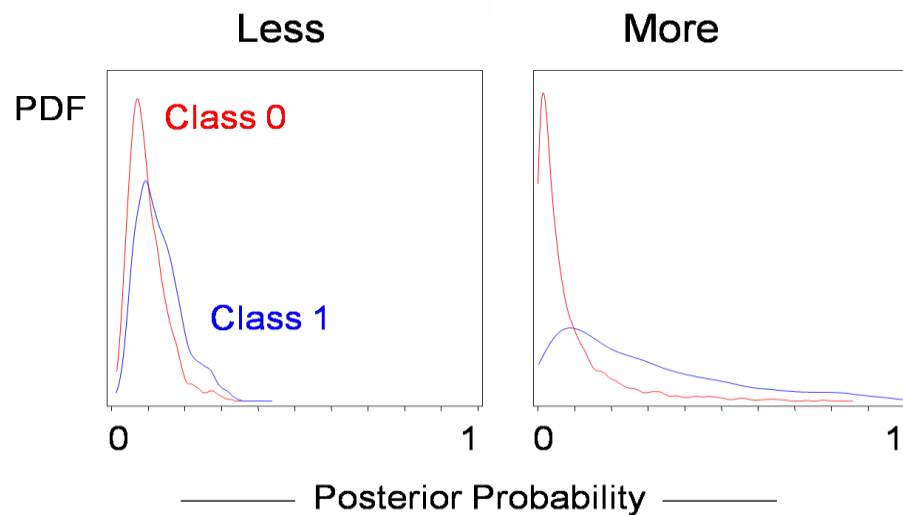
Bayes Rule:

Decision 1 if

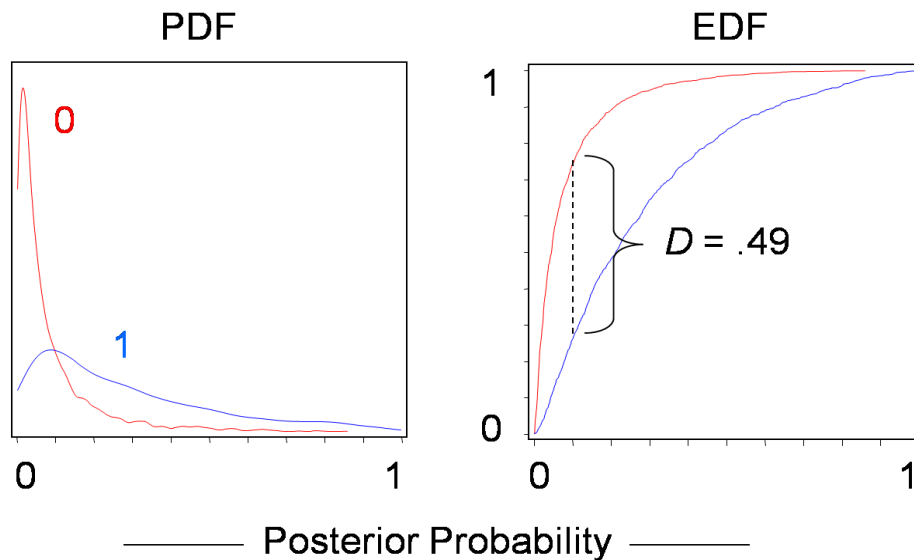
$$P > \frac{1}{1 + \left(\frac{\delta_{TP} - \delta_{FN}}{\delta_{TN} - \delta_{FP}} \right)}$$



Оптимальное разделение классов



Критерий Колмогорова-Смирнова



Схемы кодировки категориальных предикторов

- Effect coding (относительно «среднего»)

contrasts(x, contrasts = TRUE, sparse = FALSE)

<u>CLASS</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

- Reference coding (относительно «базового»)

relevel(x, ref, ...)

<u>CLASS</u>	<u>Value</u>	<u>Label</u>	<u>1</u>	<u>2</u>
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

Effect Coding: Пример

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 = Средний логит по всем категориям

β_1 = Разница между логитом для Low income и средним логитом

β_2 = разница между Medium income и средним логитом

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5363	0.1015	27.9143	<.0001
IncLevel	1	1	-0.2259	0.1481	2.3247	0.1273
IncLevel	2	1	-0.2200	0.1447	2.3111	0.1285

Reference Coding: Пример

$$\text{logit}(p) = \beta_0 + \beta_1 * D_{\text{Low income}} + \beta_2 * D_{\text{Medium income}}$$

β_0 = Логит для High

β_1 = Разница логитов между Low и High

β_2 = Разница логитов между Medium и High

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0904	0.1608	0.3159	0.5741
IncLevel	1	1	-0.6717	0.2465	7.4242	0.0064
IncLevel	2	1	-0.6659	0.2404	7.6722	0.0056

Пример

```
> cars_logist_fact <- cbind(cars_logist)
> cars_logist_fact$Type <- factor(cars_logist_fact$Type)
> cars_logist_fact$Origin <- factor(cars_logist_fact$Origin)
>
> contrasts(cars_logist_fact$Origin) <- contr.sum(3)
> cars_logist_fact$Type <- relevel(cars_logist_fact$Type, ref = "Sedan")
> f_logit_model <- glm(DriveTrainEvent ~ Invoice + EngineSize +
+                       Type + Origin + Horsepower + Length + Weight +
+                       Cylinders + Wheelbase + MPG_City + MPG_Highway,
+                       data=cars_logist_fact,
+                       family=binomial(link="logit"))
>
> summary(f_logit_model)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.023e+01	7.726e+00	-1.325	0.18527
Invoice	-9.194e-05	3.853e-05	-2.386	0.01703 *
EngineSize	-1.884e+00	8.743e-01	-2.155	0.03119 *
TypeHybrid	3.275e+00	3.470e+03	0.001	0.99925
TypeSports	-1.870e+00	8.707e-01	-2.148	0.03169 *
TypeSUV	1.917e+01	1.216e+03	0.016	0.98742
TypeTruck	-1.975e+01	1.497e+03	-0.013	0.98948
TypeWagon	-1.799e+00	7.273e-01	-2.474	0.01337 *
Origin1	4.717e-01	3.631e-01	1.299	0.19393
Origin2	-8.321e-01	5.083e-01	-1.637	0.10159
Horsepower	1.852e-02	8.403e-03	2.204	0.02753 *
Length	1.278e-01	3.559e-02	3.590	0.00033 ***
Weight	4.935e-03	1.260e-03	3.916	9.01e-05 ***
Cylinders	-5.020e-01	4.307e-01	-1.166	0.24376
Wheelbase	-3.266e-01	8.014e-02	-4.076	4.58e-05 ***
MPG_City	1.610e-01	2.219e-01	0.725	0.46818
MPG_Highway	3.725e-01	1.770e-01	2.105	0.03531 *

```
> contrasts(cars_logist_fact$Type)
Hybrid Sports SUV Truck Wagon
Sedan      0      0      0      0      0
Hybrid      1      0      0      0      0
Sports      0      1      0      0      0
SUV         0      0      1      0      0
Truck       0      0      0      1      0
Wagon       0      0      0      0      1
> contrasts(cars_logist_fact$Origin)
[,1] [,2]
Asia    1    0
Europe  0    1
USA    -1   -1
```

Пример

```
> f_odds <- data.frame(cbind(OR = coef(f_logit_model)))
> f_odds <- cbind(f_odds, confint(f_logit_model))
> f_odds <- exp(f_odds) %>% round(3)
> names(f_odds) <- c("Odds_ratio", "Lower", "Upper")
```

```
> f_odds
```

	Odds_ratio	Lower	Upper
(Intercept)	0.000000e+00	0.000	9.916800e+01
Invoice	1.000000e+00	1.000	1.000000e+00
EngineSize	1.520000e-01	0.026	8.050000e-01
TypeHybrid	2.645500e+01	0.000	Inf
TypeSports	1.540000e-01	0.027	8.360000e-01
TypeSUV	2.105504e+08	0.000	1.361270e+205
TypeTruck	0.000000e+00	0.000	1.333297e+20
TypeWagon	1.650000e-01	0.038	6.890000e-01
Origin1	1.603000e+00	0.798	3.351000e+00
Origin2	4.350000e-01	0.158	1.175000e+00
Horsepower	1.019000e+00	1.002	1.036000e+00
Length	1.136000e+00	1.061	1.221000e+00
Weight	1.005000e+00	1.003	1.008000e+00
Cylinders	6.050000e-01	0.255	1.395000e+00
Wheelbase	7.210000e-01	0.612	8.400000e-01
MPG_City	1.175000e+00	0.765	1.833000e+00
MPG_Highway	1.451000e+00	1.040	2.089000e+00

```
> Anova(f_logit_model, type = 3)
```

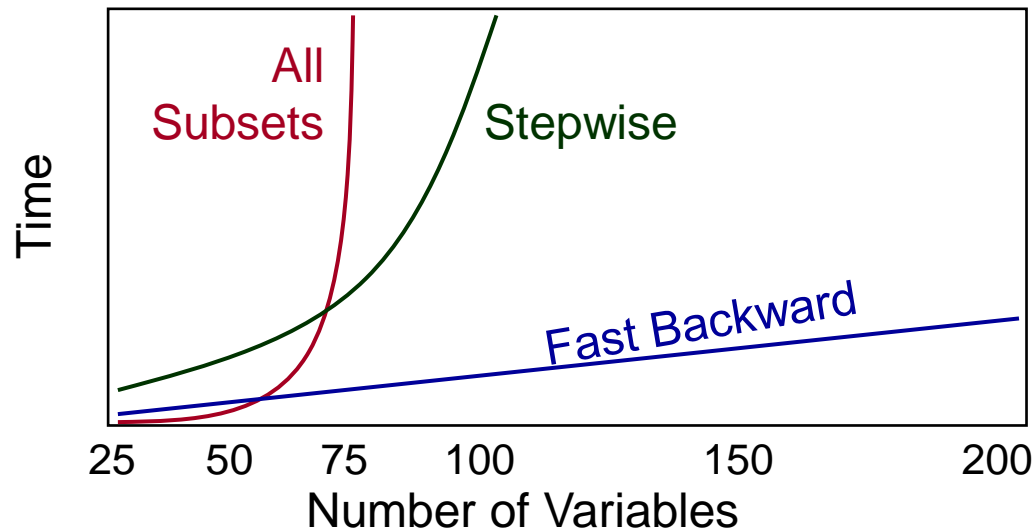
Analysis of Deviance Table (Type III tests)

Response: DriveTrainEvent

	LR	Chisq	Df	Pr(>Chisq)	
Invoice	6.850	1	0.0088623	**	
EngineSize	4.940	1	0.0262452	*	
Type	64.783	5	1.243e-12	***	
Origin	2.892	2	0.2355565		
Horsepower	5.110	1	0.0237904	*	
Length	13.587	1	0.0002278	***	
Weight	19.890	1	8.202e-06	***	
Cylinders	1.380	1	0.2400470		
Wheelbase	18.092	1	2.104e-05	***	
MPG_City	0.533	1	0.4652322		
MPG_Highway	4.837	1	0.0278627	*	

Пошаговый отбор переменных с помощью AIC и BIC для lm и glm

```
step(object, scope, direction = c("both", "backward", "forward"),  
      trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```



Отбор переменных (пример)

```
> glm_step <- glm(DriveTrainEvent ~ 1,
+                 data=na.omit(cars_logist_fact),
+                 family=binomial(link="logit"))
>
> steps <- list()
> for (i in 1:5){
+   glm_step <- step(
+     glm_step, direction = "forward",
+     scope = ~ Invoice + EngineSize + Type + Origin +
+     Horsepower + Length + Weight + Cylinders +
+     Wheelbase + MPG_City + MPG_Highway,
+     steps = 1, trace = 0)
+   roc_glm <- roc(glm_step$y ~ glm_step$fitted.values)
+   plot(roc_glm, add = i != 1, col=i)
+   steps <- cbind(steps, paste("Step", i, "(AUC:",
+                               round(roc_glm$auc, 3), ")"))
+ }
> legend("bottom", legend = steps, lwd = 3, col = 1:5)
> summary(glm_step)
```

Coefficients:

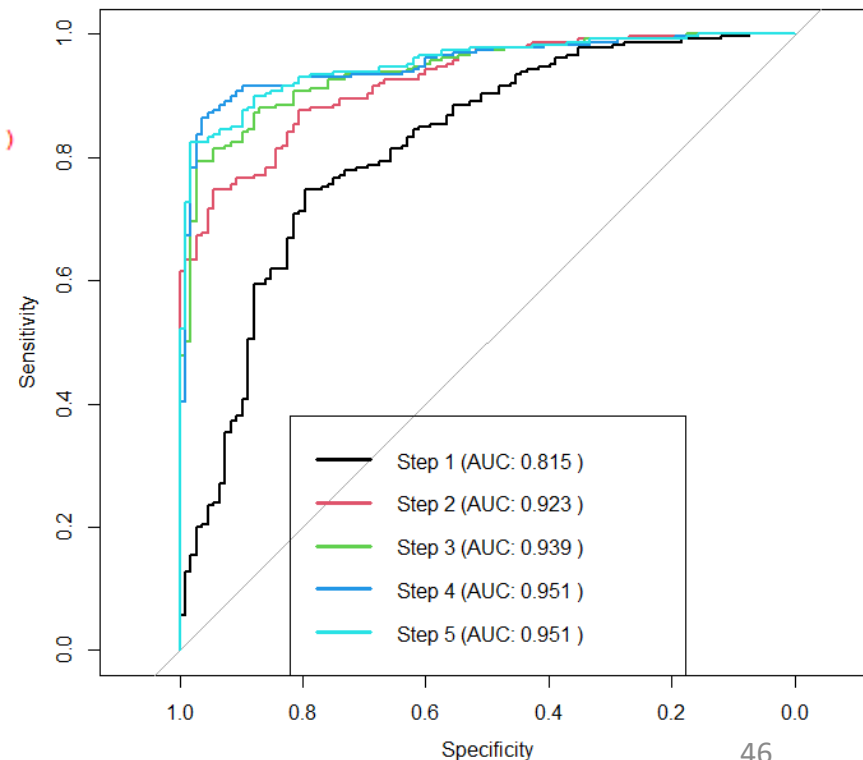
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.931e+00	3.699e+00	1.063	0.28790
Invoice	-5.680e-05	2.090e-05	-2.718	0.00657 **
TypeHybrid	1.560e+01	3.749e+03	0.004	0.99668
TypeSports	-1.863e+00	6.888e-01	-2.705	0.00684 **
TypeSUV	1.895e+01	1.235e+03	0.015	0.98776
TypeTruck	-2.176e+01	1.514e+03	-0.014	0.98853
TypeWagon	-1.205e+00	6.404e-01	-1.881	0.05994 .
Cylinders	-1.078e+00	2.497e-01	-4.318	1.57e-05 ***
Length	1.268e-01	2.808e-02	4.515	6.34e-06 ***
Wheelbase	-1.706e-01	5.765e-02	-2.959	0.00309 **

Null deviance: 420.42 on 333 degrees of freedom
 Residual deviance: 198.41 on 324 degrees of freedom
 AIC: 218.41

```
> print>Anova(glm_step, type = 3))
Analysis of Deviance Table (Type III tests)
```

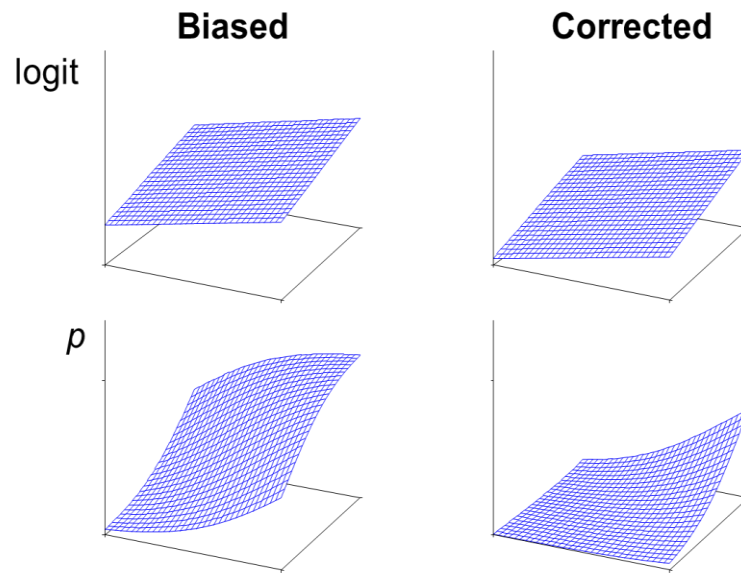
Response: DriveTrainEvent

	LR	Chisq	Df	Pr(>Chisq)
Invoice	9.567	1	0.001981	**
Type	96.989	5	< 2.2e-16	***
Cylinders	23.394	1	1.320e-06	***
Length	21.727	1	3.144e-06	***
Wheelbase	8.501	1	0.003549	**

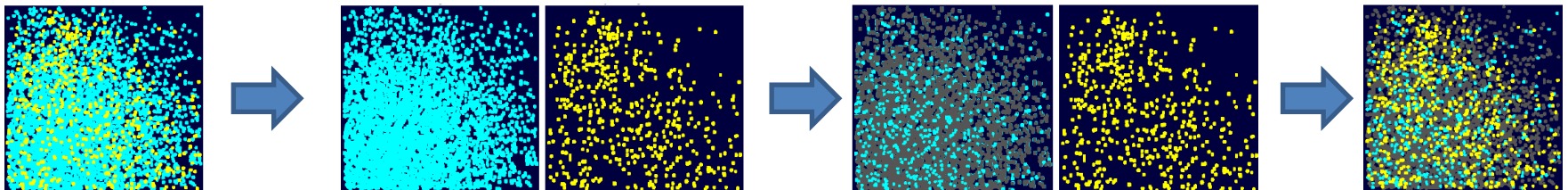


«Балансировка» выборки (oversampling)

- Порог отсечения для логистической функции:



- «Балансировка»:



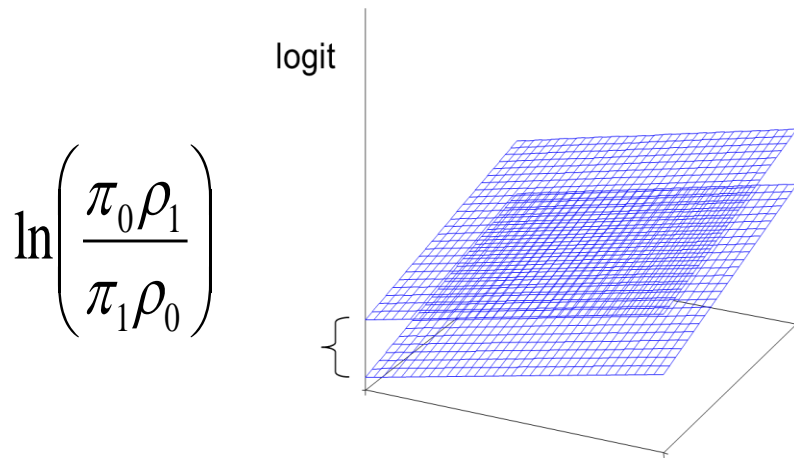
Корректировка отклика после Oversampling

Два способа корректировки

1. Включить параметр
«сдвига» в модель

`offset=X`

2. Скорректировать
вероятности на выходе
модели



π_1, π_0 - в действительности
 ρ_1, ρ_0 - в выборке

Adjusted Probability:

$$p_1^{adj} = \frac{p_1\pi_1\rho_0}{p_1\pi_1\rho_0 + (1 - p_1)\pi_0\rho_1}$$