

Прикладной статистический анализ данных.  
9. Другие виды регрессии.

Рябенко Евгений  
riabenko.e@gmail.com

7 ноября 2014 г.

## Постановка

$1, \dots, n$  — объекты;  
 $x_1, \dots, x_k$  — предикторы;  
 $y$  — отклик,  $y_i \in \mathbb{N}$ .

$$\mathbb{E}(y | x_1, \dots, x_k) = \mathbb{E}(y | x) = ?$$

Базовый метод — пуассоновская регрессия:

$$f(y | x) = \frac{e^{-\mu} \mu^y}{y!},$$

$$\mu = \mathbb{E}(y | x) = e^{x^T \beta},$$

$$\omega \equiv \mathbb{D}(y | x) = e^{x^T \beta}.$$

## Настройка параметров

Оценка методом максимального правдоподобия:

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^n \left( y_i x_i^T \beta - e^{x_i^T \beta} - \ln(y_i!) \right),$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} L(\beta) \Leftrightarrow$$

$$\sum_{i=1}^n \left( y_i - e^{x_i^T \beta} \right) x_i = 0.$$

$\hat{\beta}$ :

- существует и единственна,
- находится методом Ньютона-Рафсона,
- является состоятельной и асимптотически эффективной оценкой  $\beta$ ,
- асимптотически нормальна.

## Дисперсия оценок

Для оценки дисперсии  $\hat{\beta}$  снова используется матрица вторых производных  $L(\beta)$ :

$$I(\beta) = \frac{\partial^2 L}{\partial \beta \partial \beta^T} = - \sum_{i=1}^n e^{x_i^T \beta} x_i x_i^T.$$

Из теории оценок максимума правдоподобия:

$$\begin{aligned} \mathbb{D}_{ML}(\hat{\beta}) &= I^{-1}(\hat{\beta}), \\ \hat{\beta} &\stackrel{a}{\sim} N\left(\beta, \left(\sum_{i=1}^n e^{x_i^T \beta} x_i x_i^T\right)^{-1}\right) \approx \\ &\approx N\left(\beta, \left(\sum_{i=1}^n e^{x_i^T \hat{\beta}} x_i x_i^T\right)^{-1}\right). \end{aligned}$$

## Доверительные интервалы

Для отдельного коэффициента  $\beta_j$ :

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\left(I^{-1}(\hat{\beta})\right)_{jj}}$$

Для  $\ln \mathbb{E}(y | x = x_0) = x_0^T \beta$ :

$$x_0^T \hat{\beta} \pm z_{1-\alpha/2} \sqrt{x_0^T I^{-1}(\hat{\beta}) x_0}$$

Для  $\mathbb{E}(y | x = x_0) = e^{x_0^T \beta}$ :

$$\left[ e^{x_0^T \hat{\beta} - z_{1-\alpha/2} \sqrt{x_0^T I^{-1}(\hat{\beta}) x_0}}, e^{x_0^T \hat{\beta} + z_{1-\alpha/2} \sqrt{x_0^T I^{-1}(\hat{\beta}) x_0}} \right]$$

Приближённый предсказательный интервал для  $y(x_0)$  — отклика на новом объекте  $x_0$ :

$$e^{x_0^T \hat{\beta}} \pm 2 \sqrt{e^{x_0^T \hat{\beta}}}$$

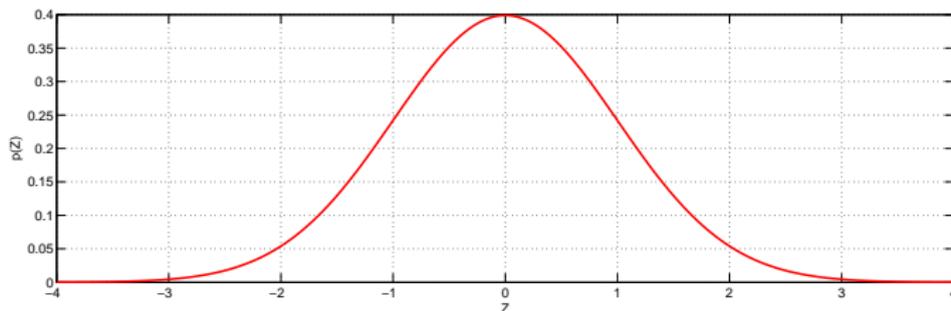
## Критерий Вальда

нулевая гипотеза:  $H_0: \beta_j = 0;$

альтернатива:  $H_1: \beta_j < \neq > 0;$

статистика:  $T = \frac{\hat{\beta}_j}{\sqrt{(I^{-1}(\hat{\beta}))_{jj}}};$

$T \sim N(0, 1)$  при  $H_0.$



## Критерий отношения правдоподобия

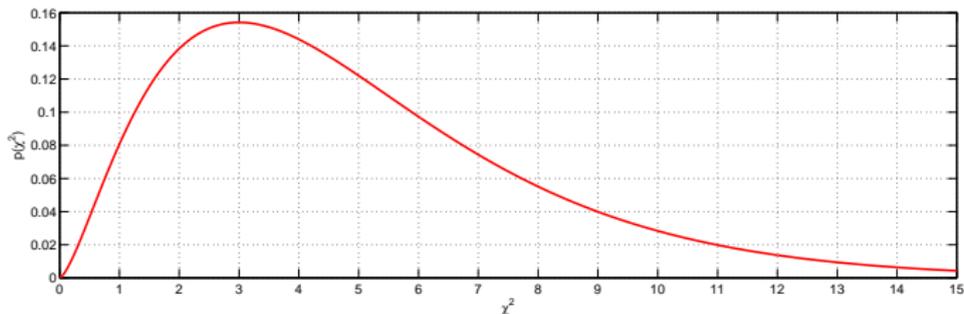
$$X_{n \times (k+1)} = \begin{pmatrix} X_1 & X_2 \\ n \times (k+1-k_1) & n \times k_1 \end{pmatrix}; \quad \beta^T_{(k+1) \times 1} = \begin{pmatrix} \beta_1^T & \beta_2^T \\ (k+1-k_1) \times 1 & k_1 \times 1 \end{pmatrix}^T;$$

нулевая гипотеза:  $H_0: \beta_2 = 0;$

альтернатива:  $H_1: H_0$  неверна;

статистика:  $G = 2(L_r - L_{ur});$

$G \sim \chi^2_{k_1}$  при  $H_0.$



## Overdispersion/underdispersion

Пуассоновская модель предполагает, что  $\omega = \mu$  (equidispersion).

- МП-оценки  $\beta$  остаются состоятельными, даже если распределение  $y|x$  не является пуассоновским — достаточно того, что модель  $\mathbb{E}(y|x)$  определена корректно.
- Оценки дисперсии  $\hat{\beta}$  и соответствующие критерии требуют верного определения и  $\mathbb{D}(y|x)$ , поэтому они дают некорректные результаты, если матожидание и дисперсия не равны.
- Предположение о равенстве матожидания и дисперсии можно проверить; если оно не выполняется, можно изменить модель. Это позволит построить корректные критерии и более эффективные оценки  $\beta$ .

## Overdispersion/underdispersion

Overdispersion — отрицательная биномиальная модель:

$$\omega(\alpha) = \mu + \alpha\mu^2,$$

$$f(y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^y.$$

Underdispersion — пороговая модель (hurdle model):

$$P(y = j) = \begin{cases} f_1(0), & j = 0, \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(j), & j > 0. \end{cases}$$

Можно построить МП-оценки для  $\alpha$  и  $\beta$ , а затем проверить гипотезу  $\alpha = 0$  с помощью критерия отношения правдоподобия.

## Устойчивая оценка дисперсии

Дисперсия оценки максимального квазиправдоподобия:

$$\mathbb{D}_{QML}(\hat{\beta}) = \left( \sum_{i=1}^n \mu_i x_i x_i^T \right)^{-1} \left( \sum_{i=1}^n \omega_i x_i x_i^T \right) \left( \sum_{i=1}^n \mu_i x_i x_i^T \right)^{-1}.$$

Устойчивая состоятельная оценка дисперсии, подходящая для любого вида  $\omega$ :

$$\mathbb{D}_R(\hat{\beta}) = \left( \sum_{i=1}^n \mu_i x_i x_i^T \right)^{-1} \left( \sum_{i=1}^n (y_i - \mu_i)^2 x_i x_i^T \right) \left( \sum_{i=1}^n \mu_i x_i x_i^T \right)^{-1}.$$

## Меры качества модели

Относительные:

- аномальность:

$$D = -2L,$$

$$D_P = \sum_{i=1}^n \left( y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right),$$

$$D_{NB} = \sum_{i=1}^n \left( y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i + \alpha^{-1}) \ln \frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}} \right);$$

- AIC:

$$AIC = -2L + 2(k + 1).$$

Абсолютная:

- псевдо- $R^2$ :

$$R_{DEV}^2 = 1 - \frac{D}{D_0},$$

$D_0$  — аномальность модели с одной константой.

## Число визитов к доктору

Cameron, Trivedi, Regression Analysis of Count Data: изучается функционирование системы здравоохранения Австралии. Для 5190 одиноких совершеннолетних граждан известны значения следующих показателей:

социоэкономические:

- возраст, лет;
- годовой доход, 10 тыс. долл.;
- индикаторы наличия страховки различных типов: частной, государственной для малоимущих, государственной для пожилых, инвалидов и ветеранов;

краткосрочные характеристики здоровья:

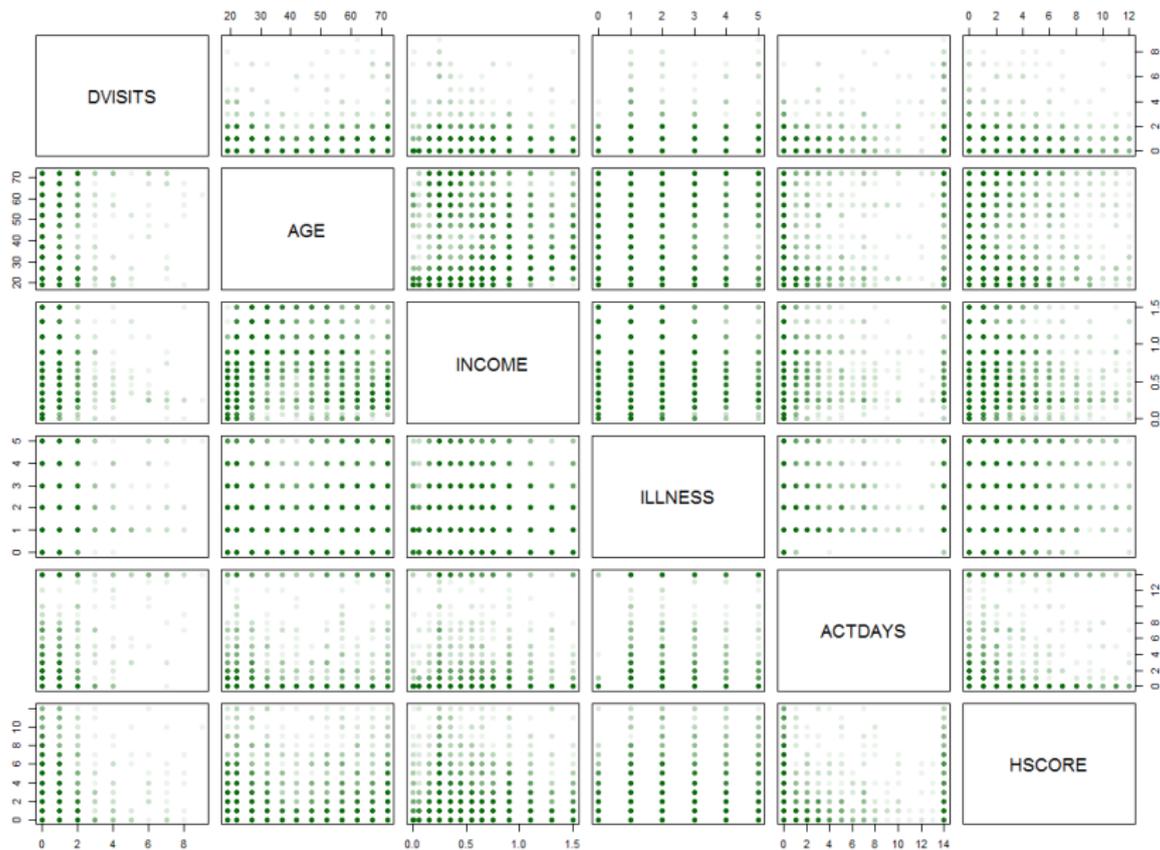
- число визитов к доктору за последние две недели;
- число заболеваний за последние две недели;
- число дней сниженной активности в связи с заболеванием или травмой за последние две недели;

долгосрочные характеристики здоровья:

- оценка состояния здоровья по опроснику Голдберга;
- индикаторы наличия хронических заболеваний, ограничивающих и не ограничивающих активность.

Построить модель числа визитов к доктору в зависимости от остальных признаков.

## Данные



## Модель 1

Стандартная пуассоновская модель:

```
lm1 <- glm(y~., family=poisson(), data=X)
summary(lm1)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.09782	0.10155	-20.66	$< 2 \times 10^{-16}$
SEX	0.15649	0.05614	2.79	0.0053
AGE	0.00279	0.00166	1.68	0.0926
INCOME	-0.18742	0.08548	-2.19	0.0283
LEVYPLUS	0.12650	0.07155	1.77	0.0771
FREEPOOR	-0.43846	0.17980	-2.44	0.0147
FREEREPA	0.08364	0.09207	0.91	0.3636
ILLNESS	0.18616	0.01826	10.19	$< 2 \times 10^{-16}$
ACTDAYS	0.12669	0.00503	25.18	$< 2 \times 10^{-16}$
HSCORE	0.03068	0.01007	3.05	0.0023
CHCOND1	0.11730	0.06655	1.76	0.0780
CHCOND2	0.15072	0.08226	1.83	0.0669

$D = 4380.1$ ,  $AIC = 6736$ ,  $R_{DEV}^2 = 0.223$ .

# Модель 1

Проверим гипотезу равенства среднего и дисперсии против альтернативы увеличенной дисперсии:

```
library(AER)
dispersiontest(lm1)
```

$$p = 3.1 \times 10^{-5}.$$

## Модель 2

Отрицательная биномиальная модель NB2:

```
library(MASS)
lm2 <- glm.nb(y~., data=X)
summary(lm2)
```

В MASS используется параметр нелинейности  $\theta = 1/\alpha$ .

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.276279	0.122946	-18.515	$< 2 \times 10^{-16}$
SEX	0.216367	0.069374	3.119	0.0018
AGE	0.003314	0.002088	1.587	0.1125
INCOME	-0.156203	0.103176	-1.514	0.1301
LEVYPLUS	0.116382	0.085417	1.363	0.1731
FREEPOOR	-0.497302	0.206882	-2.404	0.0163
FREEREPA	0.145755	0.116852	1.247	0.2123
ILLNESS	0.214960	0.024182	8.889	$< 2 \times 10^{-16}$
ACTDAYS	0.143753	0.007809	18.408	$< 2 \times 10^{-16}$
HSCORE	0.037541	0.013742	2.732	0.0063
CHCOND1	0.097890	0.078601	1.245	0.2130
CHCOND2	0.183505	0.103180	1.778	0.0754

$\hat{\alpha} = 1.08$ ,  $D = 3029.8$ ,  $AIC = 6424$ ,  $R_{DEV}^2 = 0.229$ .

## Модель 3

Сокращённая отрицательная биномиальная модель NB2:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.18241	0.06311	-34.58	$< 2 \times 10^{-16}$
SEX	0.32114	0.06505	4.94	$7.9 \times 10^{-7}$
FREEPOOR	-0.60120	0.20070	-3.00	0.0027
ILLNESS	0.24790	0.02204	11.25	$< 2 \times 10^{-16}$
ACTDAYS	0.14697	0.00712	20.65	$< 2 \times 10^{-16}$
HSCORE	0.03938	0.01338	2.94	0.0032

$\hat{\alpha} = 1.08$ ,  $D = 3051.5$ ,  $AIC = 6436$ ,  $R_{DEV}^2 = 0.223$ .

Критерия отношения правдоподобия: модель 3 значительно хуже модели 2 ( $p = 0.0006$ ).

# Модель 4

Попробуем по одному возвращать удалённые признаки. Лучшей получается модель, в которую возвращён возраст:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.39129	0.08374	-28.56	$< 2 \times 10^{-16}$
SEX	0.26141	0.06715	3.89	$9.9 \times 10^{-5}$
AGE	0.00626	0.00161	3.88	0.0001
FREEPOOR	-0.49241	0.20183	-2.44	0.0147
ILLNESS	0.22999	0.02244	10.25	$< 2 \times 10^{-16}$
ACTDAYS	0.14499	0.00716	20.26	$< 2 \times 10^{-16}$
HSCORE	0.04169	0.01345	3.10	0.0019

$$\hat{\alpha} = 1.08, \quad D = 6406.4, \quad AIC = 6422, \quad R_{DEV}^2 = 0.227.$$

Критерия отношения правдоподобия: модель 4 существенно лучше модели 3 ( $p = 9.9 \times 10^{-5}$ ) и не хуже модели 2 ( $p = 0.1207$ ).

## Модель 4

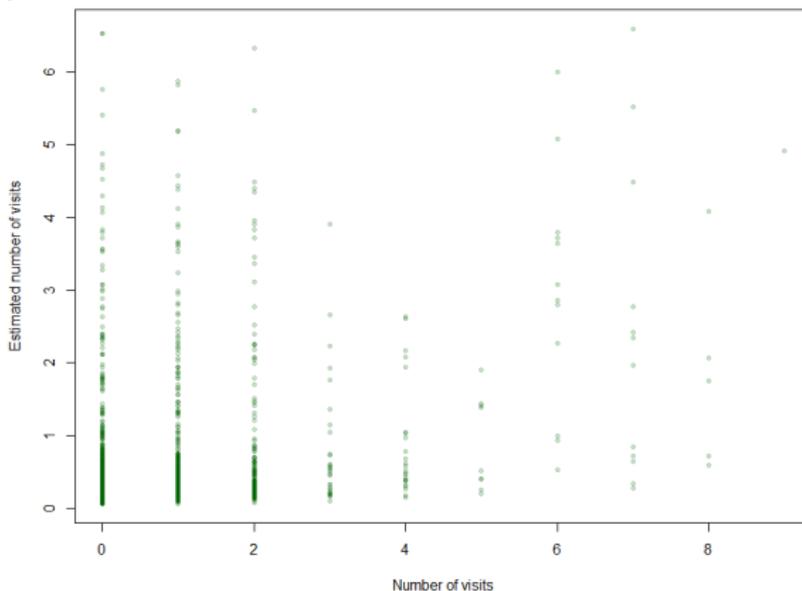
Для сравнения — модель с устойчивыми оценками дисперсии:

```
library(sandwich)
cov.robust <- vcovHC(lm3, type="HC0")
se.robust <- sqrt(diag(cov.robust))
coeffs <- coef(lm3)
t.robust <- coeffs / se.robust
summary.robust <- cbind(coeffs, se.robust, t.robust,
                        pvalue=2*(1-pnorm(abs(coeffs/se.robust))))
print(summary.robust)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.39129	0.08533	-28.02	$< 2 \times 10^{-16}$
SEX	0.26141	0.07180	3.64	0.00027
AGE	0.00626	0.00174	3.60	0.00032
FREEPOOR	-0.49241	0.26232	-1.88	0.06050
ILLNESS	0.22999	0.02076	11.08	$< 2 \times 10^{-16}$
ACTDAYS	0.14499	0.00734	19.74	$< 2 \times 10^{-16}$
HSCORE	0.04169	0.01356	3.07	0.00211

## Результат

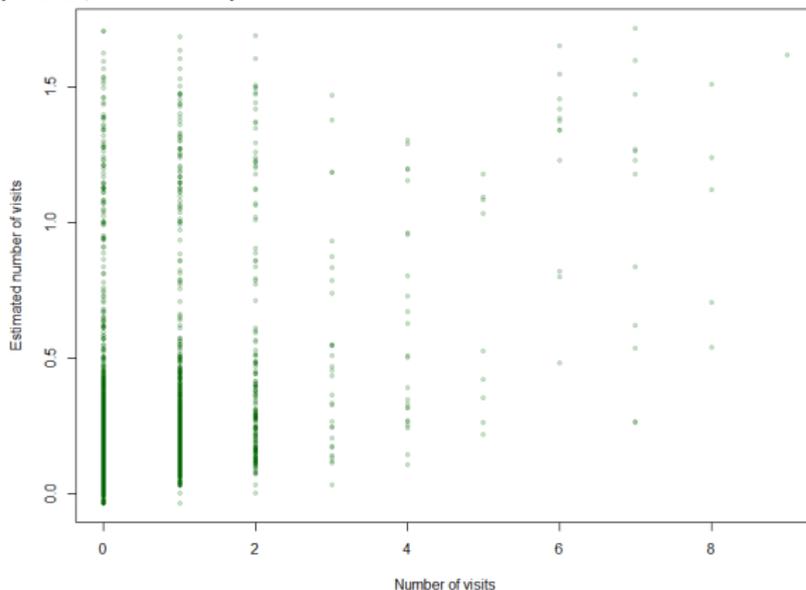
Итоговая модель построена по 5190 наблюдениям и объясняет около 23% вариации отклика.



$$\hat{\mathbb{E}}(y|x) = \exp(-2.3 + 0.26 * SEX + 0.006 * AGE - 0.49 * FREEPOOR + 0.23 * ILLNESS + 0.14 * ACTDAYS + 0.04 * HSCORE).$$

## Результат

Для сравнения — линейная модель, не учитывающая дискретность отклика  $y$  ( $R_{DEV}^2 = 0.191$ ):



$$\hat{\mathbb{E}}(y|x) = \exp(0.007 + 0.03 * SEX + 0.001 * AGE - 0.07 * FREEPOOR + 0.04 * ILLNESS + 0.04 * ACTDAYS + 0.01 * HSCORE) - 1.$$

## Результат

Модель позволяет сделать следующие выводы:

- женщины посещают врача в 1.3 раза чаще мужчин (доверительный интервал (1.14, 1.48));
- обладатели государственной страховки для малоимущих посещают врача в 1.6 раз реже (доверительный интервал (1.1, 2.4));
- каждая болезнь увеличивает число посещений врача в 1.3 раза (доверительный интервал (1.2, 1.3));
- каждый день сниженной активности увеличивает число посещений врача в 1.16 раза (доверительный интервал (1.13, 1.17));
- каждые 10 лет число посещений врача в возрастает в 1.06 раза (доверительный интервал (1.03, 1.10));
- каждый балл оценки здоровья по опроснику Голдберга увеличивает число посещений врача в 1.04 раза (доверительный интервал (1.02, 1.07)).

# Variance-bias tradeoff

Модель:

$$y = f(x) + \varepsilon, \quad \mathbb{E}\varepsilon = 0, \quad \mathbb{D}\varepsilon = \sigma^2.$$

Ошибка предсказания  $y_0$  по вектору  $x_0$  с помощью модели  $\hat{f}$ :

$$PE(\hat{f}(x_0)) = \mathbb{E}(y_0 - \hat{f}(x_0))^2 = \sigma^2 + MSE(\hat{f}(x_0)).$$

Среднеквадратичная ошибка оценки  $\hat{f}$ :

$$\begin{aligned} MSE(\hat{f}(x_0)) &= \mathbb{E}(f(x_0) - \hat{f}(x_0))^2 = \\ &= (\mathbb{E}(f(x_0)) - f(x_0))^2 + \mathbb{E}(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0))^2 = \\ &= bias^2(\hat{f}(x_0)) + variance(\hat{f}(x_0)). \end{aligned}$$

## Variance-bias tradeoff

В линейной регрессии:

$$y = x\beta + \varepsilon,$$
$$MSE(x_0\hat{\beta}) = bias^2(x_0^T\hat{\beta}) + variance(x_0^T\hat{\beta}).$$

МНК-оценка

$$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$$

является несмещённой ( $bias = 0$ ) и имеет наименьшую дисперсию среди всех несмещённых оценок ( $variance = \sigma^2 (X^T X)^{-1}$ ).

Если матрица  $X^T X$  плохо обусловлена, то:

- МНК-оценка имеет большую дисперсию;
- может возникать численная неустойчивость при обращении  $X^T X$ .

# Гребневая регрессия

Для уменьшения дисперсии оценки и повышения вычислительной устойчивости добавим к  $X^T X$  диагональную матрицу:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I_n)^{-1} X^T y.$$

Такая оценка является решением регуляризованной задачи наименьших квадратов:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2).$$

$\lambda$  — параметр регуляризации:

- при  $\lambda = 0$   $\hat{\beta}^{ridge} = \hat{\beta}^{OLS}$ , смещения нет;
- при  $\lambda = \infty$   $\hat{\beta}^{ridge} = 0$ , дисперсии нет;
- в промежутке — баланс между смещением и дисперсией.

## Важные детали

- Коэффициент  $\beta_0$  не входит в регуляризатор:

$$\left(\hat{\beta}_0, \hat{\beta}^{ridge}\right) = \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^k}{\operatorname{argmin}} \left( \|y - \beta_0 I_n - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right).$$

Если перед применением метода центрировать все признаки  $X$  (вычесть выборочное среднее), то можно положить  $\hat{\beta}_0 = \bar{y}$  и исключить  $\beta_0$  из задачи минимизации.

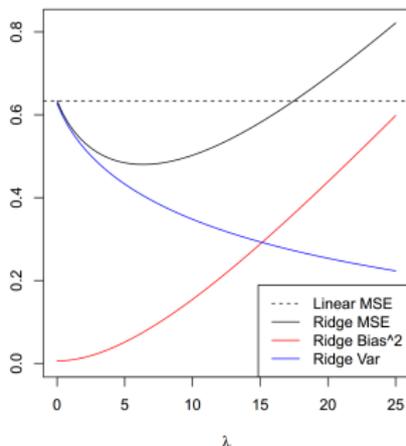
Если центрировать ещё и  $y$ , то  $\hat{\beta}_0 = 0$ .

- Штраф  $\|\beta\|_2^2 = \sum_{j=1}^k \beta_j^2$  неравномерно распределяется между признаками, если они измерены в разных шкалах. Поэтому перед применением метода признаки  $X$  дополнительно стандартизируют, чтобы выборочная дисперсия каждого равнялась единице.
- После построения модели необходимо привести её к записи в терминах исходных признаков.

## Пример 1

Пусть  $n = 50$ ,  $k = 30$ ,  $X_{ij} \sim N(0, 1)$ ,  $\sigma^2 = 1$ .

Сгенерируем  $y$  согласно линейной модели с 10 большими коэффициентами (между 0.5 и 1) и 20 маленькими (между 0 и 0.3).



Линейная регрессия:

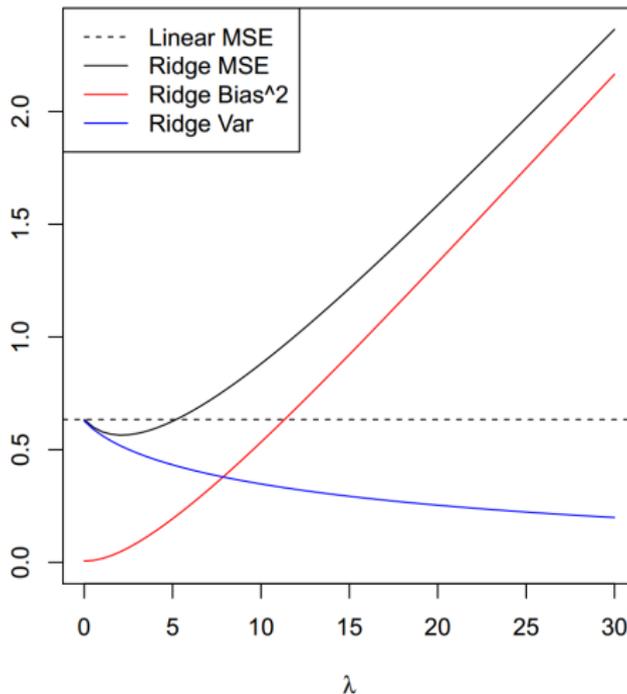
$$PE = 1 + MSE \approx 1 + 0.006 + 0.627 = 1.633.$$

Лучшая гребневая регрессия:

$$PE = 1 + MSE \approx 1 + 0.077 + 0.403 = 1.48.$$

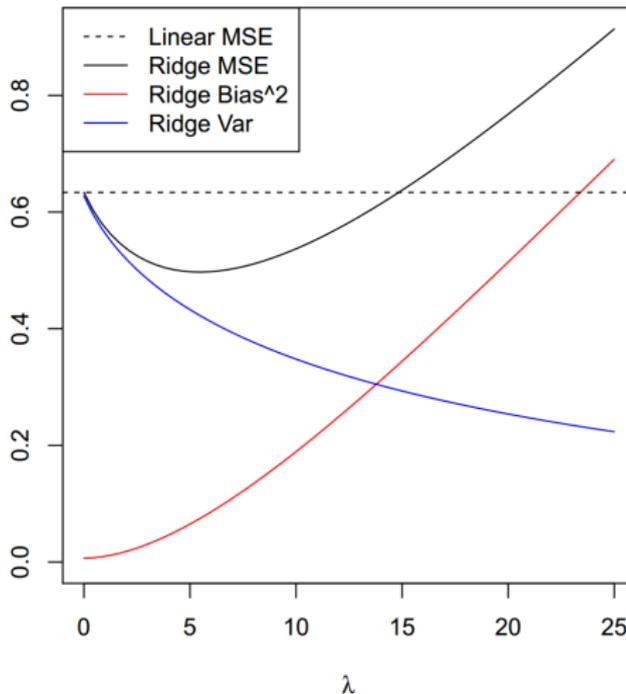
## Пример 2

Сгенерируем  $y$  согласно линейной модели с 30 большими коэффициентами (между 0.5 и 1).

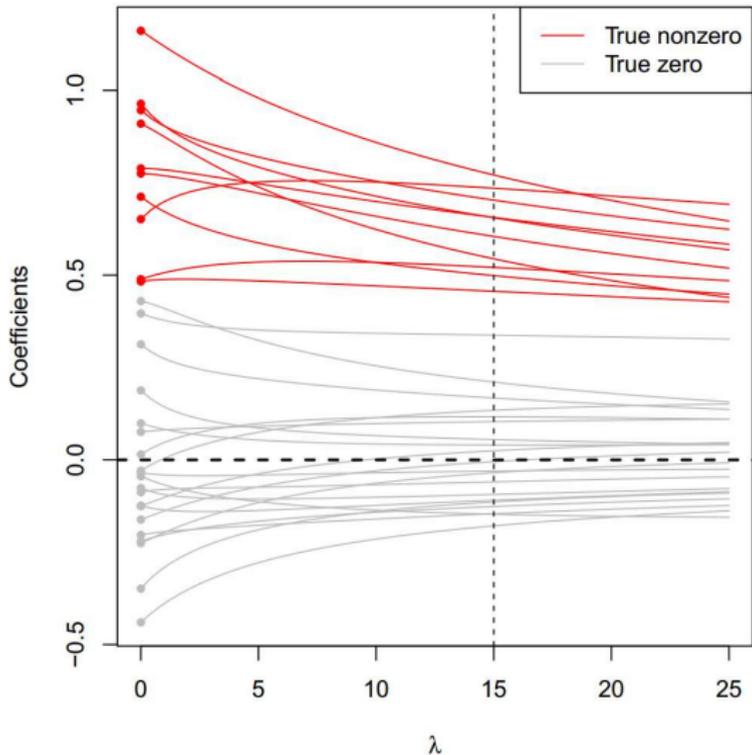


## Пример 3

Сгенерируем  $y$  согласно линейной модели с 10 большими коэффициентами (между 0.5 и 1) и 20 нулевыми.



# Пример 3



Оценки нулевых коэффициентов не становятся равны нулю, а только уменьшаются.

## Выбор $\lambda$

**Эмпирический способ:** выбирается такое  $\lambda$ , начиная с которого коэффициенты модели меняются незначительно.

**Кросс-валидация:** выбирается  $\lambda$ , доставляющее минимум средней ошибке на контроле; модель затем настраивается по полным данным.

## Лассо

Lasso (Least Absolute Selection and Shrinkage Operator):

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|_2^2 + \lambda \|\beta\|_1).$$

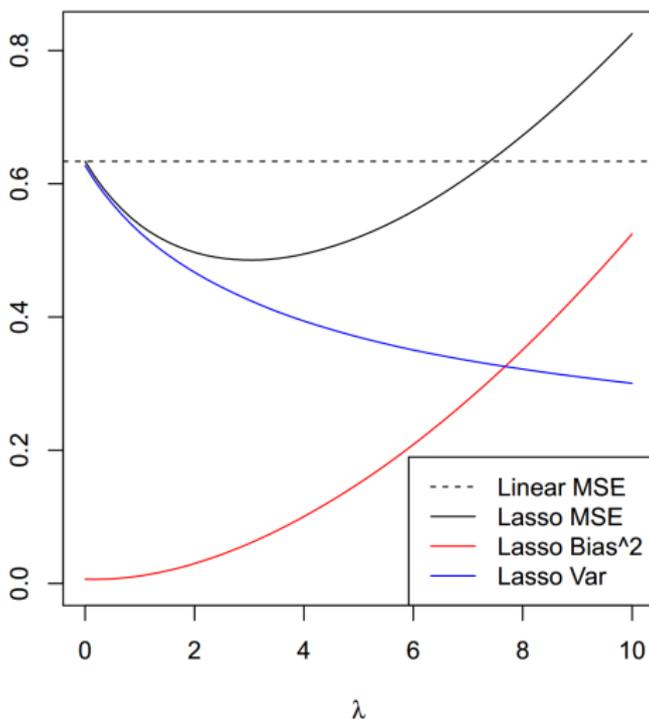
Выражения в замкнутом виде не существует.

$l_1$ -регуляризация позволяет обнулять коэффициенты.

Всё остальные детали те же, что у гребневой регрессии:  $\lambda$  определяет баланс между смещением и дисперсией,  $\beta_0$  не входит в регуляризатор, признаки нужно стандартизировать, итоговую модель необходимо приводить к записи в исходных величинах,  $\lambda$  выбирается по кросс-валидации.

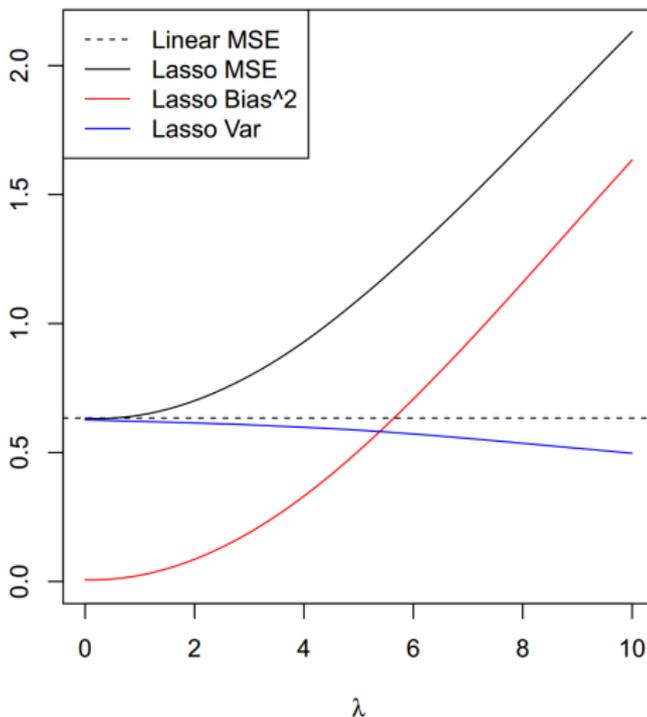
## Пример 1

10 больших коэффициентов (между 0.5 и 1) и 20 маленьких (между 0 и 0.3).



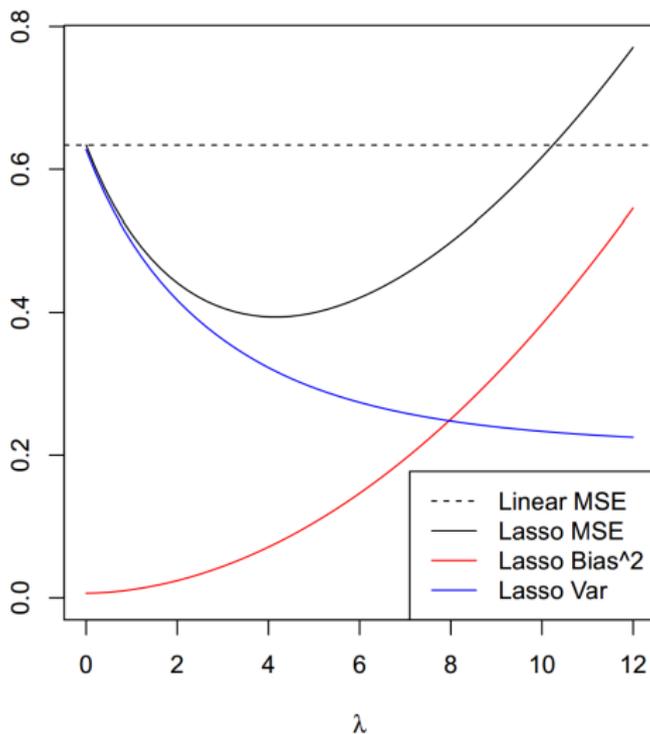
## Пример 2

30 больших коэффициентов (между 0.5 и 1).



## Пример 3

10 больших коэффициентов (между 0.5 и 1) и 20 нулевых.



## Особенности

- Алгоритм LARS позволяет получить значения коэффициентов при всех  $\lambda$ .
- Лассо можно применять даже при  $k > n$ , но получится не больше  $n$  ненулевых коэффициентов.
- При  $n > k$  и наличии высоко коррелированных предикторов лассо проигрывает по ошибке предсказания гребневой регрессии.
- Если среди признаков есть группа высоко коррелированных, лассо, как правило, отбирает только один из них.

# Эластичная сеть

Elastic net:

$$\hat{\beta}^{en} = \left(1 + \frac{\lambda_2}{n}\right) \operatorname{argmin}_{\beta} (\|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2).$$

Другая форма записи:

$$\hat{\beta}^{en} = \operatorname{argmin}_{\beta} \left( \beta^T \left( \frac{X^T X + \lambda_2 I_n}{1 + \lambda_2} \right) \beta - 2y^T X\beta + \lambda_1 \|\beta\|_1 \right).$$

Лассо в аналогичном виде:

$$\hat{\beta}^{lasso} = \operatorname{argmin}_{\beta} \left( \beta^T X^T X\beta - 2y^T X\beta + \lambda_1 \|\beta\|_1 \right).$$

## Роль параметров регуляризации

- При  $\lambda_1 = 0$  получаем гребневую регрессию с параметром  $\lambda_2$ .
- При  $\lambda_2 = 0$  получаем лассо с параметром  $\lambda_1$ .
- При  $\lambda_1 = \infty$  получаем  $\hat{\beta}_j^{en} = 0$ .
- При  $\lambda_2 = \infty$  получаем univariate soft thresholding:

$$\hat{\beta}_j^{UST} = \left( |y^T x_j| - \frac{\lambda_1}{2} \right)_+ \text{sign} (y^T x_j), \quad j = 1, \dots, k.$$

Значения параметров выбираются кросс-валидацией,  $\lambda_2$  по небольшой сетке (например, (0.0, 0.01, 0.1, 1, 10, 100)), а  $\lambda_1$  — по непрерывной кривой, получаемой алгоритмом LARS-EN.

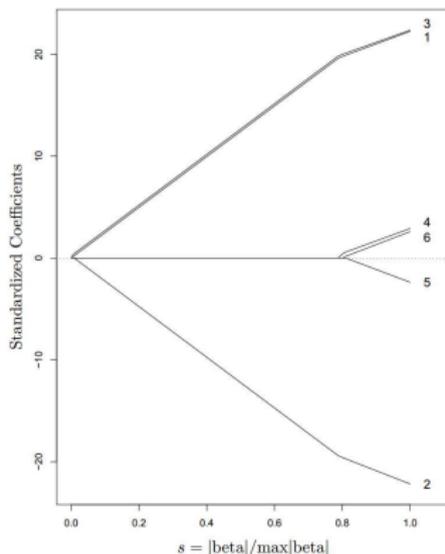
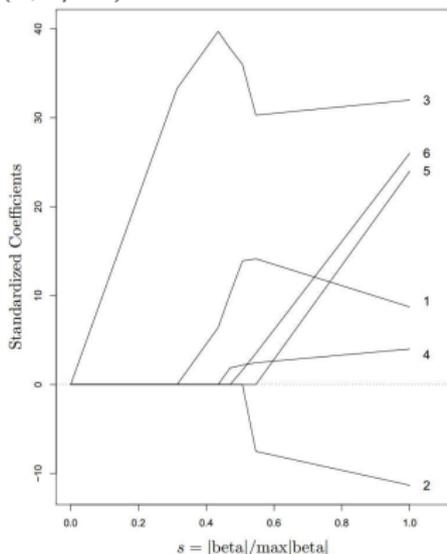
## Пример

Пусть  $Z_1, Z_2 \sim N(0, 1)$ ,  $y \sim N(Z_1 + 0.1Z_2, 1)$ ,

$$X_1 = Z_1 + \varepsilon_1, \quad X_2 = -Z_1 + \varepsilon_2, \quad X_3 = Z_1 + \varepsilon_3,$$

$$X_4 = Z_2 + \varepsilon_4, \quad X_5 = -Z_2 + \varepsilon_5, \quad X_6 = Z_2 + \varepsilon_6,$$

$\varepsilon_i \sim N(0, 1/16)$ .



Слева лассо, справа эластичная сеть с  $\lambda_2 = 0.5$ .

## Литература

- регрессия натурального признака — Cameron, главы 2, 3, 5, 6;
- гребневая регрессия, лассо, эластичная сеть — Hastie, 3.4.

Cameron C.A., Trivedi P.K. *Regression Analysis of Count Data*. — Cambridge University Press, 2013.

Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning*. — Springer, 2009.