Прикладная статистика 9. Другие виды регрессии.

Рябенко Евгений
riabenko.e@gmail.com

14 апреля 2014 г.

Постановка

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

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

Базовый метод — пуассоновская регрессия:

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

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

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

Настройка параметров

Оценка методом максимального правдоподобия:

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

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

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

$\hat{\beta}$:

- существует и единственна,
- находится методом Ньютона-Рафсона,
- является состоятельной и асимптотически эффективной оценкой β,
- 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:

$$\mathbb{D}_{ML}\left(\hat{\beta}\right) = I^{-1}\left(\hat{\beta}\right),$$

$$\hat{\beta} \overset{\text{a}}{\sim} N\left(\beta, \left(\sum_{i=1}^{n} e^{x_i^T \beta} x_i x_i^T\right)^{-1}\right) \approx$$

$$\approx N\left(\beta, \left(\sum_{i=1}^{n} e^{x_i^T \hat{\beta}} x_i x_i^T\right)^{-1}\right).$$

Доверительные интервалы

Для отдельного коэффициента β_j:

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

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

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

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

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

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

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

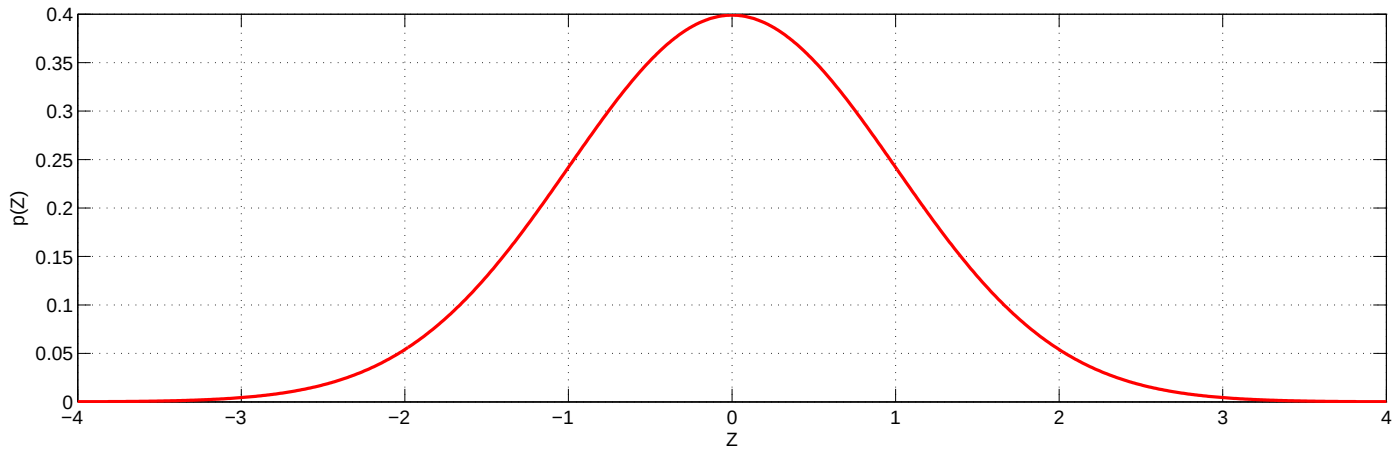
Критерий Вальда

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

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

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

$T \sim N(0,1)$ при $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - ncdf(t,0,1), & H_1 : \beta_j > 0, \\ ncdf(t,0,1), & H_1 : \beta_j < 0, \\ 2\left(1 - ncdf(|t|,0,1)\right), & H_1 : \beta_j \neq 0. \end{cases}$$

Критерий отношения правдоподобия

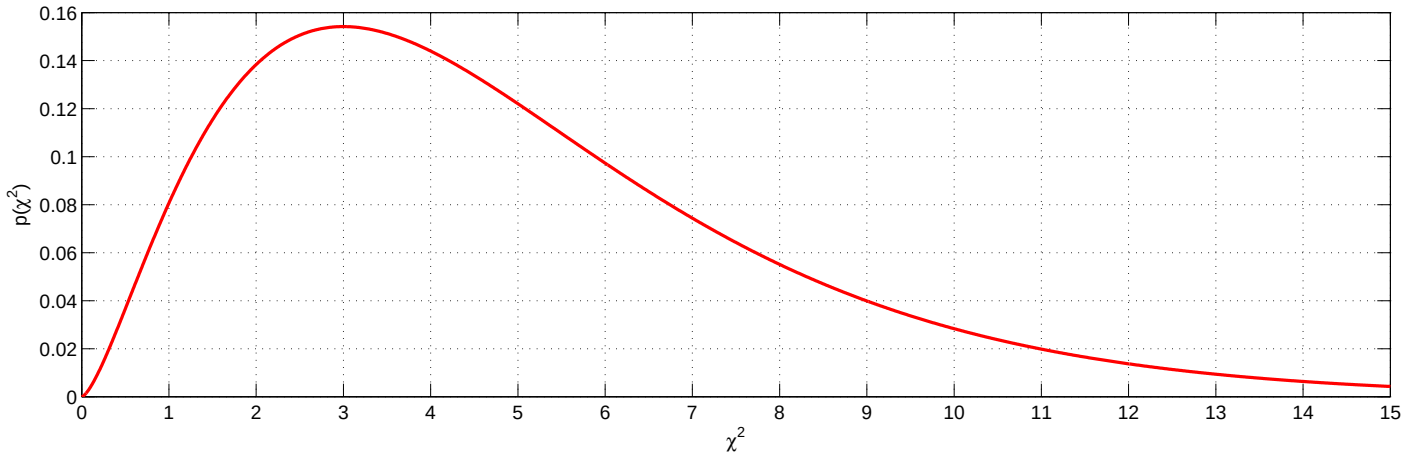
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \qquad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

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

альтернатива: $H_1 : H_0$ неверна;

статистика: $G = 2\left(L_r - L_{ur}\right);$

$G \sim \chi^2_{k_1}$ при $H_0;$



достигаемый уровень значимости:

$$p\left(g\right) = 1 - chi2cdf\left(g, k_1\right).$$

Overdispersion/underdispersion

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

- МП-оценки β 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 β.

Overdispersion/underdispersion

Overdispersion — отрицательная биномиальная модель:

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

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

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 α и β, а затем проверить гипотезу $\alpha = 0$ с помощью критерия отношения правдоподобия.

Устойчивая оценка дисперсии

Дисперсия оценки максимального квазиправдоподобия:

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

Устойчивая состоятельная оценка дисперсии, подходящая для любого вида ω:

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

Меры качества модели

Относительные:

- аномальность:

$$D = -2L,$$

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

$$D_{NB} = \sum_{i=1}^{n} \left(y_i \ln \frac{y_i}{\hat{\mu}_i} - \left(y_i + \alpha^{-1}\right) \ln \frac{y_1 + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}} \right);$$

- AIC:

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

Абсолютная:

- псевдо-R^2:

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

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:

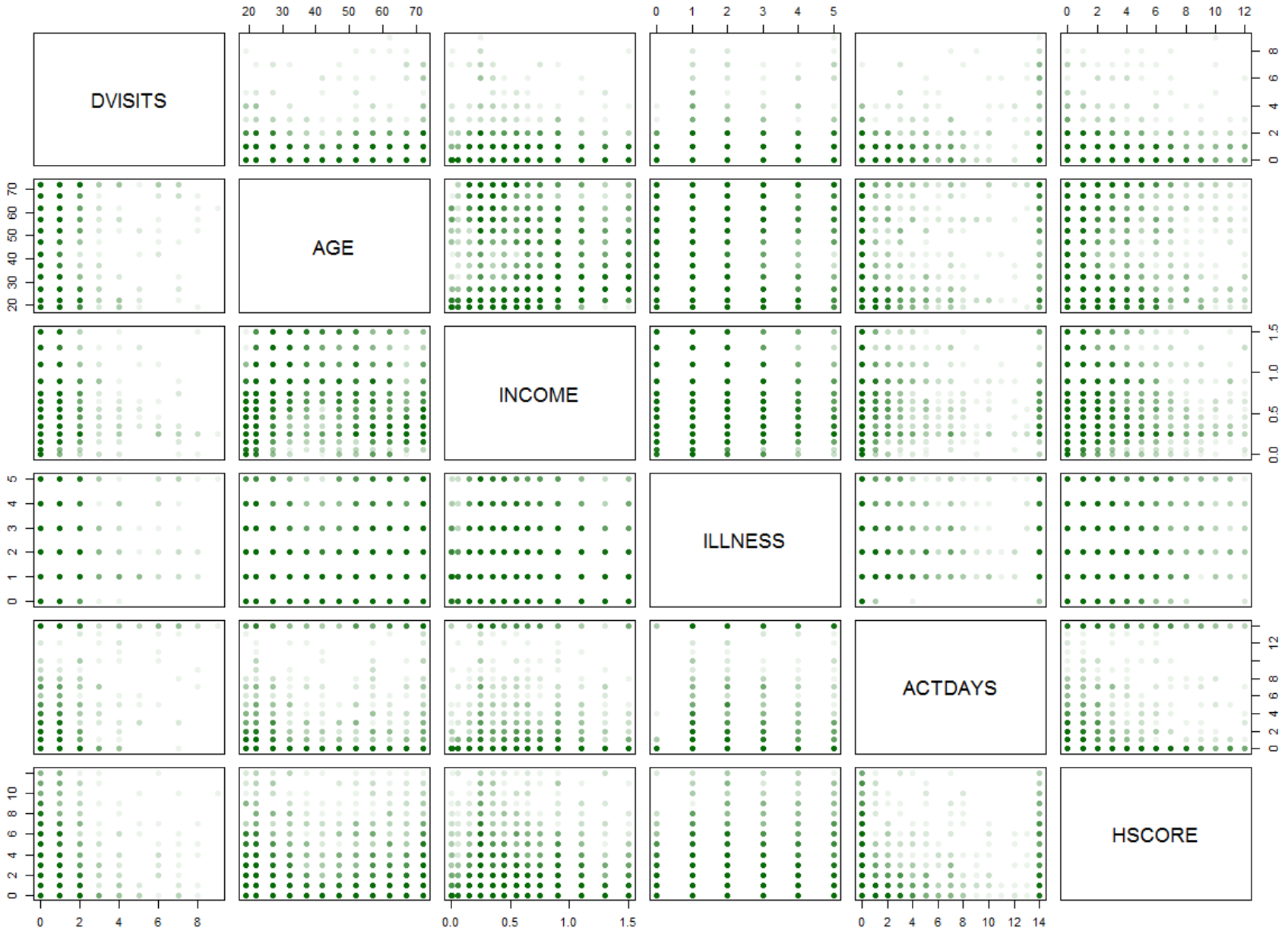
- число визитов к доктору за последние две недели;
- число заболеваний за последние две недели;
- число дней сниженной активности в связи с заболеванием или травмой за последние две недели;

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

$D = 4380.1, \quad AIC = 6736, \quad R^2_{DEV} = 0.223.$

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

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

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	7.9×10^{-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

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

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

Модель 4

Попробуем вернуть 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	9.9×10^{-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

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

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

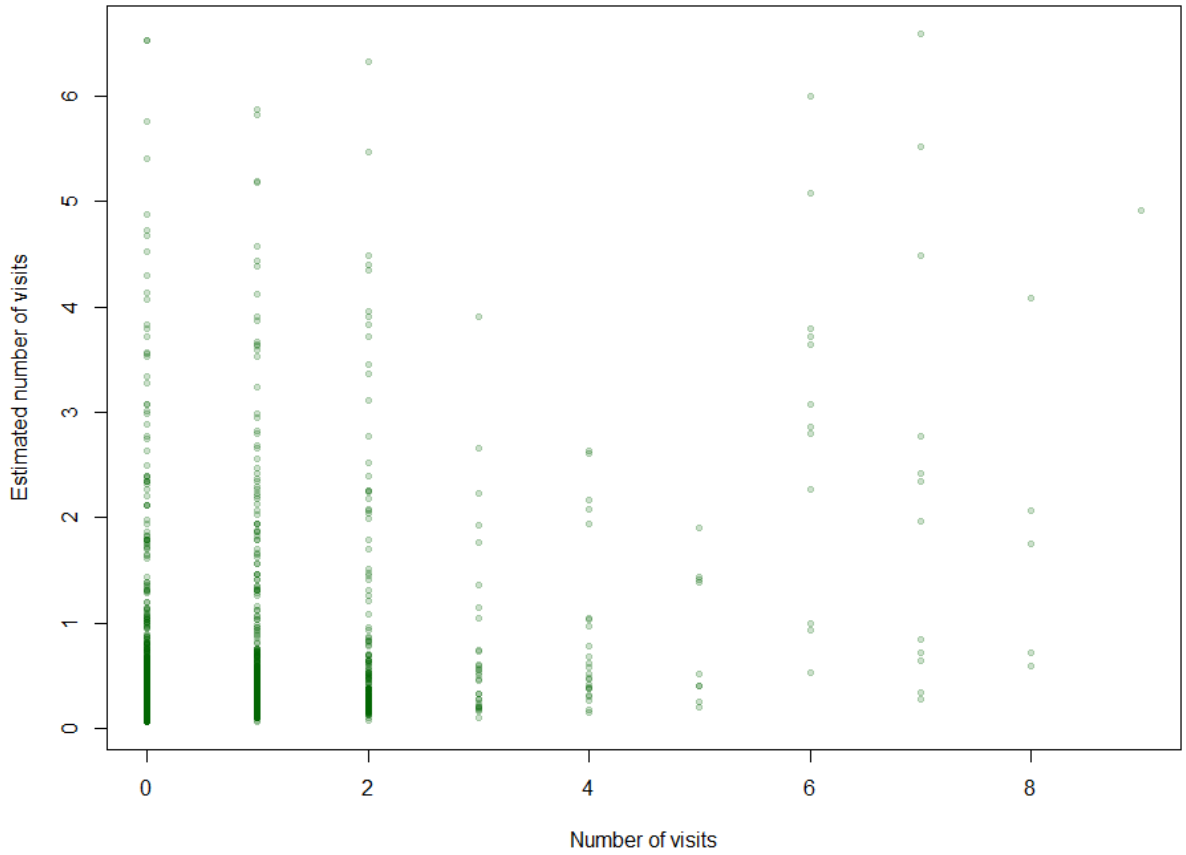
Для сравнения — модель с устойчивыми оценками дисперсии:

```
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

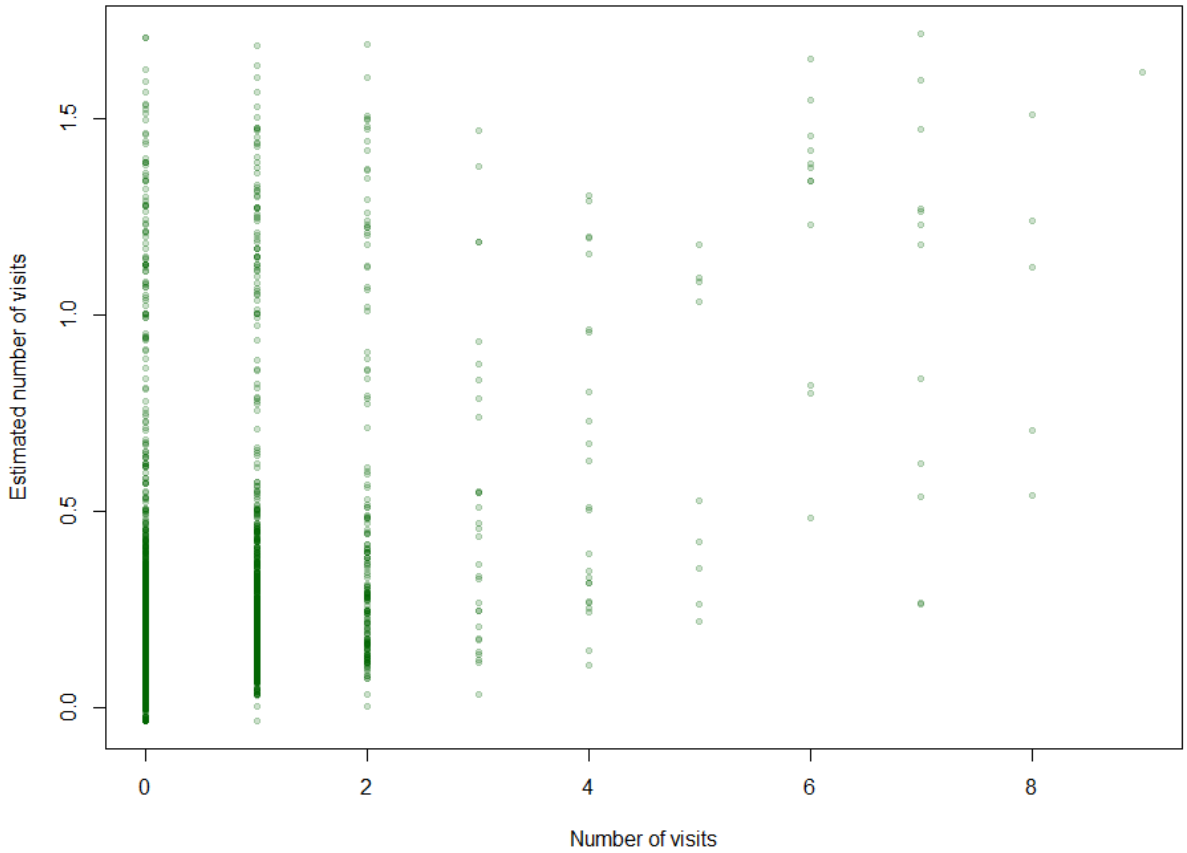
Итоговая модель построена по 5190 наблюдениям и объясняет около 23% вариации отклика.



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

Результат

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



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

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:

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

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

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

Среднеквадратичная ошибка оценки $\hat{f}$:

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

Variance-bias tradeoff

В линейной регрессии:

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

МНК-оценка

$$\hat{\beta}^{OLS} = \left(X^TX\right)^{-1}X^Ty$$

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

Если матрица X^TX плохо обусловлена, то:

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

Гребневая регрессия

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

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

Такая оценка является решением регуляризованной задачи наименьших квадратов:

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

λ — 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 β_0 не входит в регуляризатор:

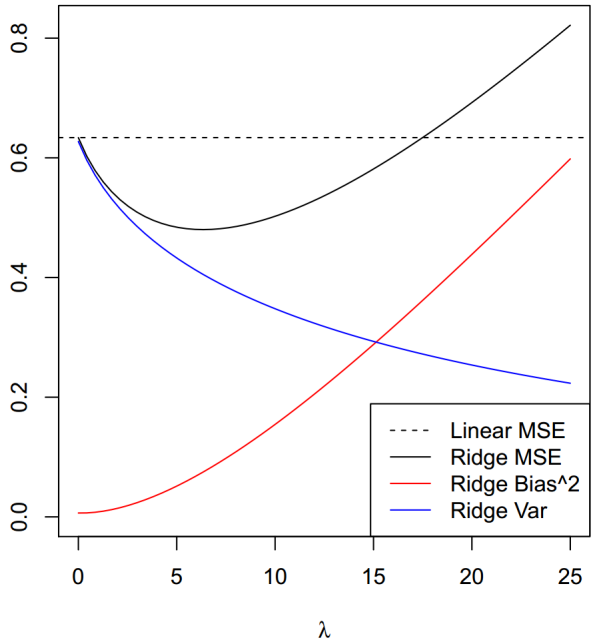
$$\left(\hat{\beta}_0, \hat{\beta}^{ridge}\right) = \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^k}{\text{argmin}} \left(\|y - \beta_0 I_n - X\beta\|_2^2 + \lambda \|\beta\|_2^2\right).$$

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

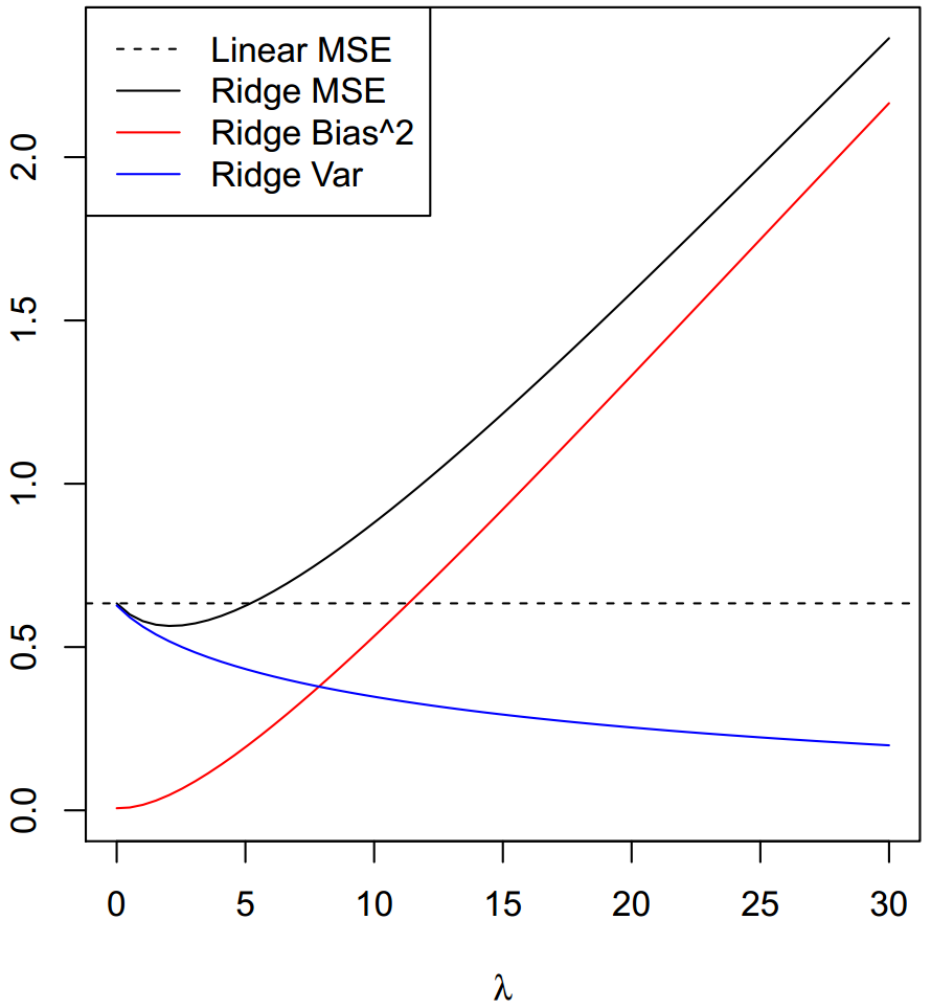
$$PE = 1 + MSE \approx 1 + 0.006 + 0.627 = 1.633.$$

Лучшая гребневая регрессия:

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

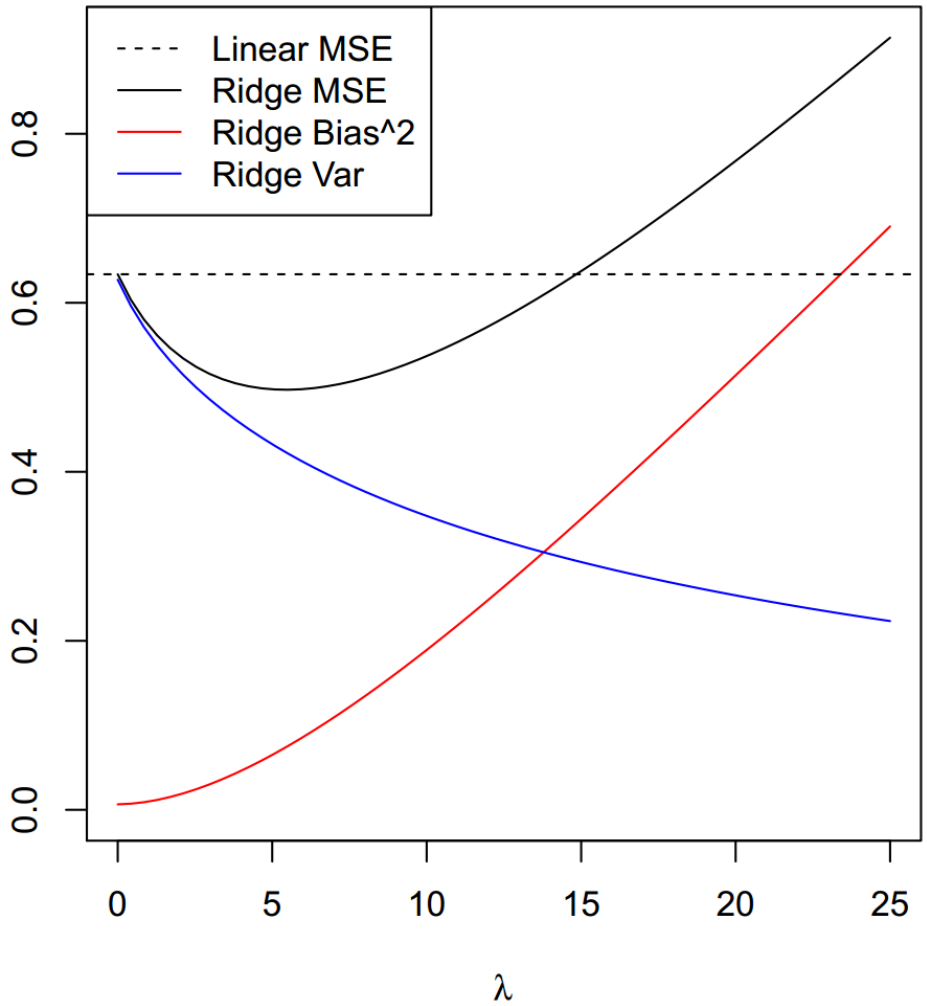
Пример 2

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

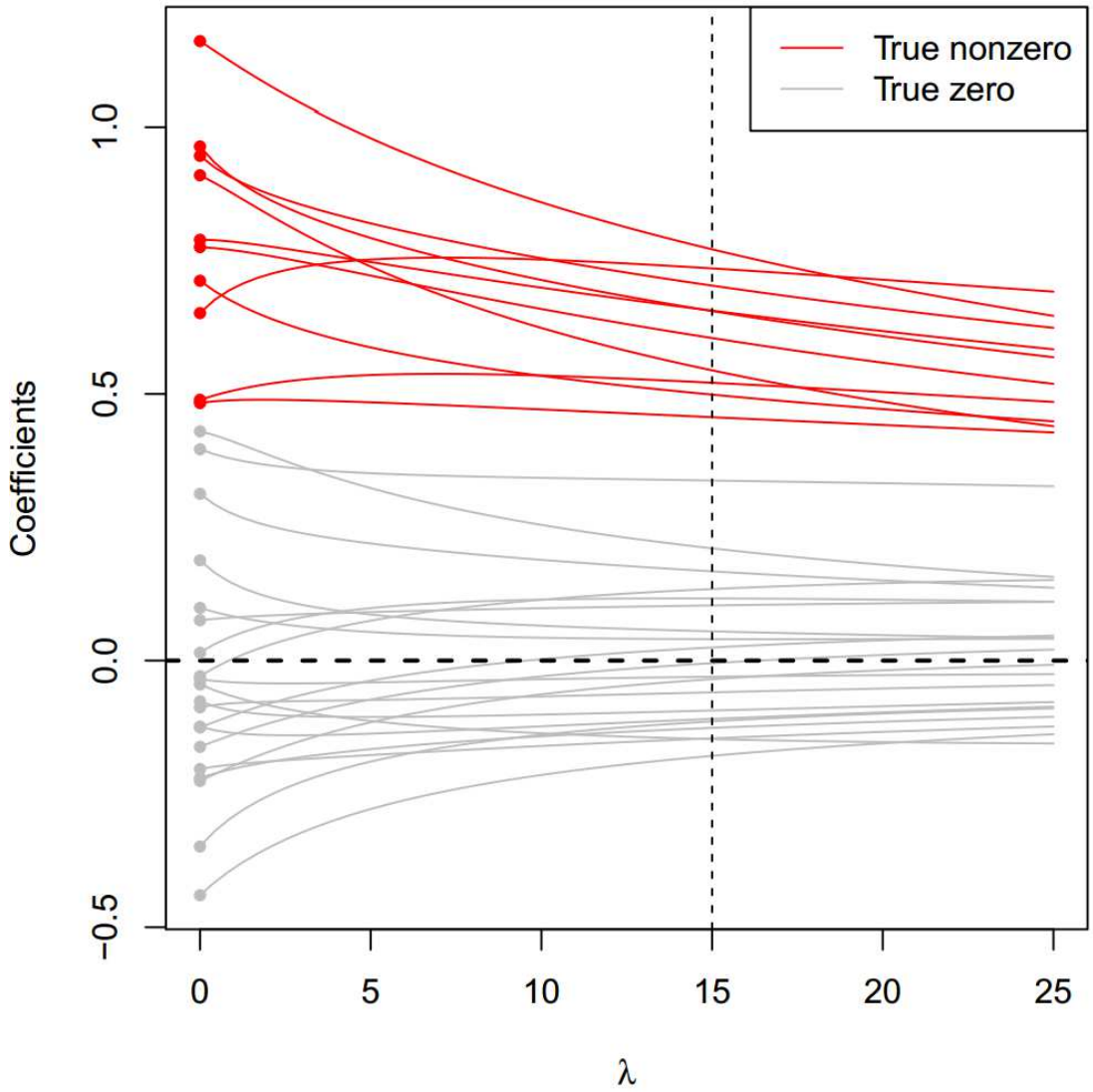


Пример 3

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 λ

Эмпирический способ: выбирается такое λ, начиная с которого коэффициенты модели меняются незначительно.

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

Лассо

Lasso (Least Absolute Selection and Shrinkage Operator):

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

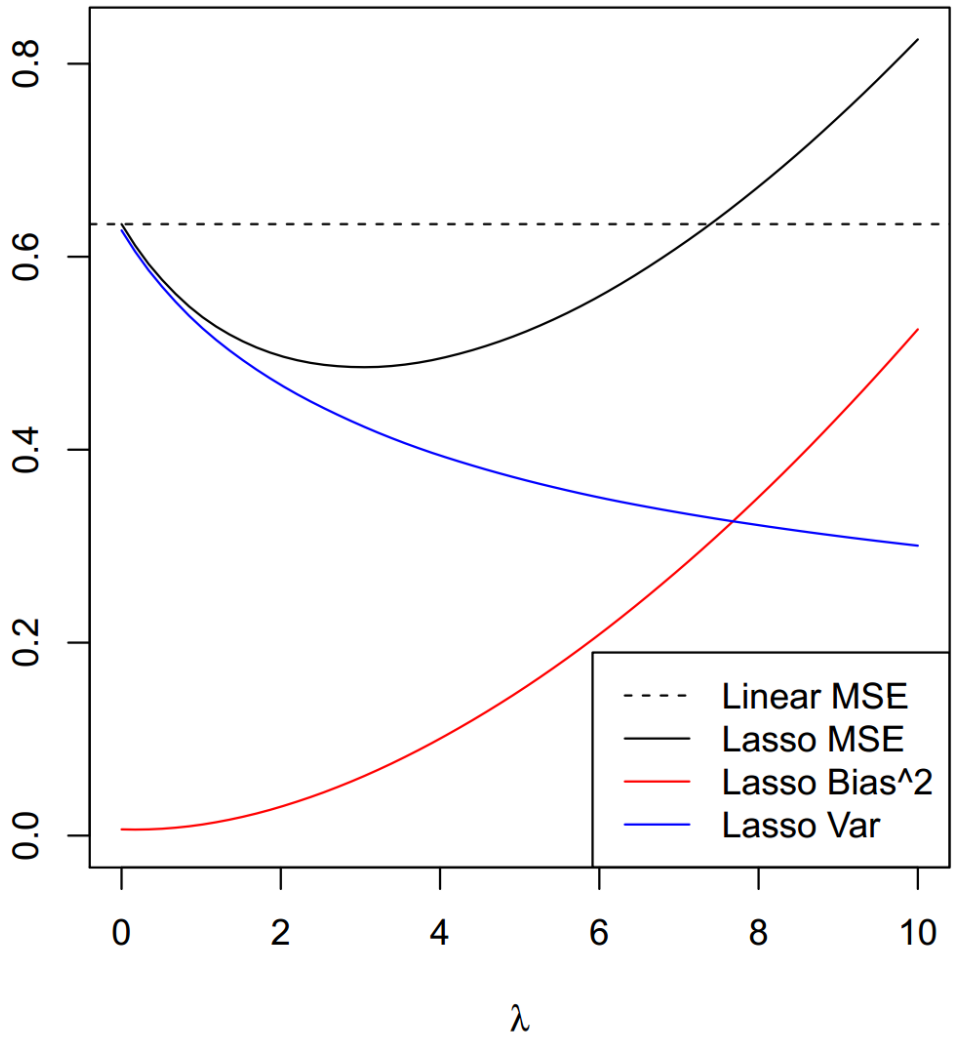
Выражения в замкнутом виде не существует.

l_1-регуляризация позволяет обнулять коэффициенты.

Всё остальные детали те же, что у гребневой регрессии: λ определяет баланс между смещением и дисперсией, β_0 не входит в регуляризатор, признаки нужно стандартизировать, итоговую модель необходимо приводить к записи в исходных величинах, λ выбирается по кросс-валидации.

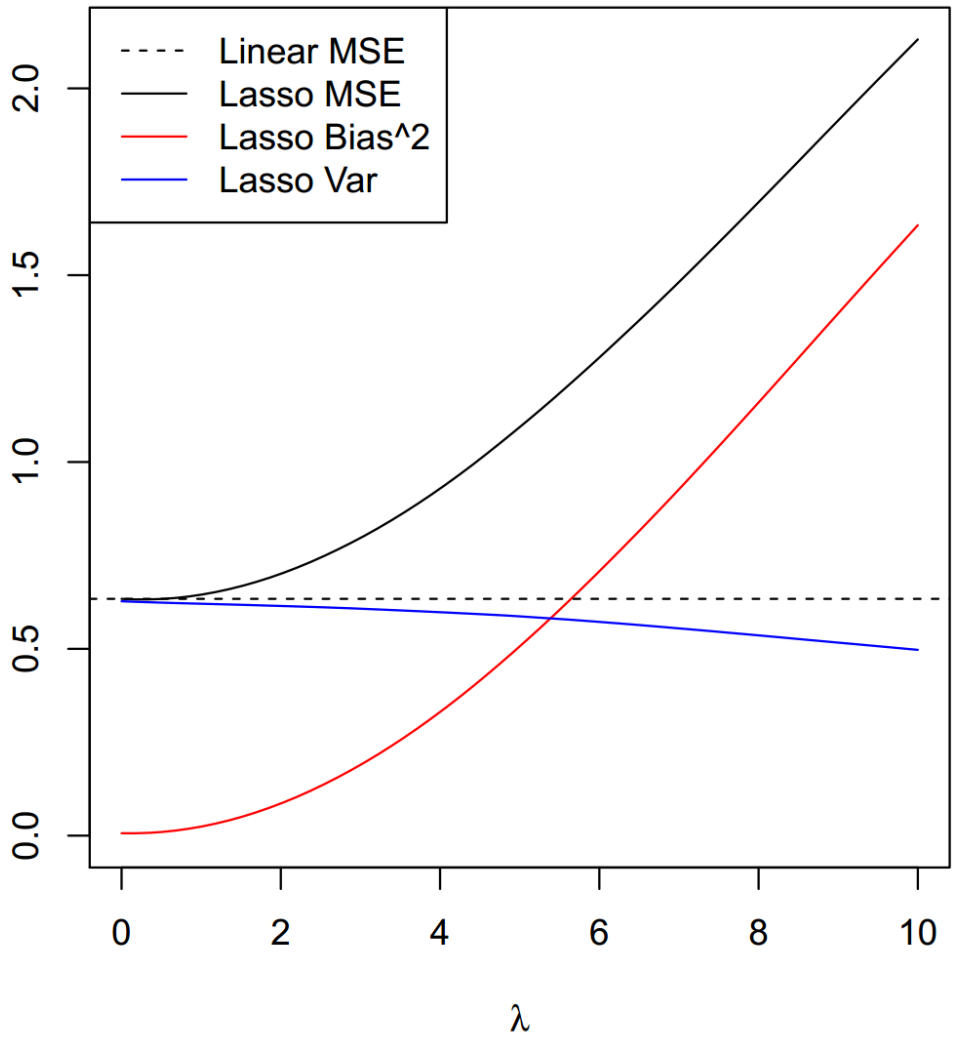
Пример 1

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



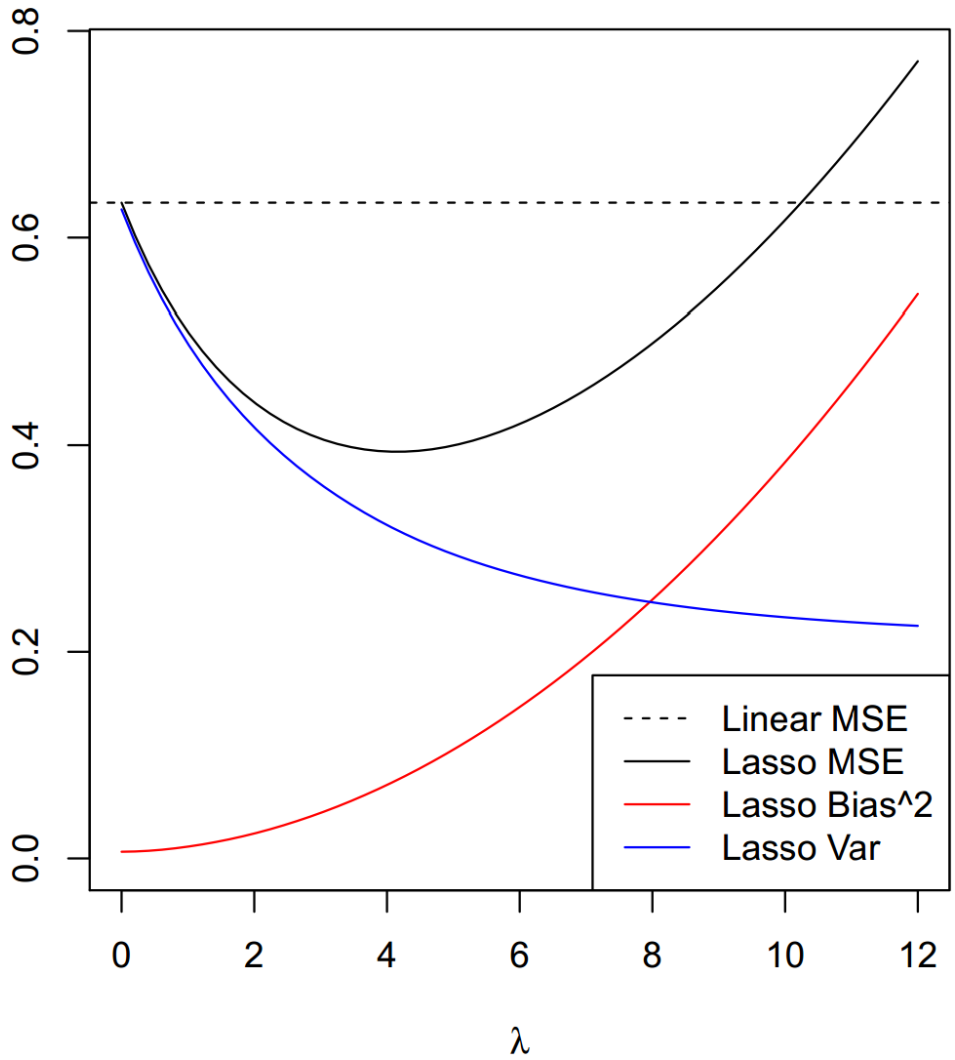
Пример 2

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 λ.
- 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:

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

Другая форма записи:

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

Лассо в аналогичном виде:

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

Роль параметров регуляризации

- При $\lambda_1 = 0$ получаем гребневую регрессию с параметром λ_2.
- При $\lambda_2 = 0$ получаем лассо с параметром λ_1.
- При $\lambda_1 = \infty$ получаем $\hat{\beta}_j^{en} = 0$.
- При $\lambda_2 = \infty$ получаем univariate soft thresholding:

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

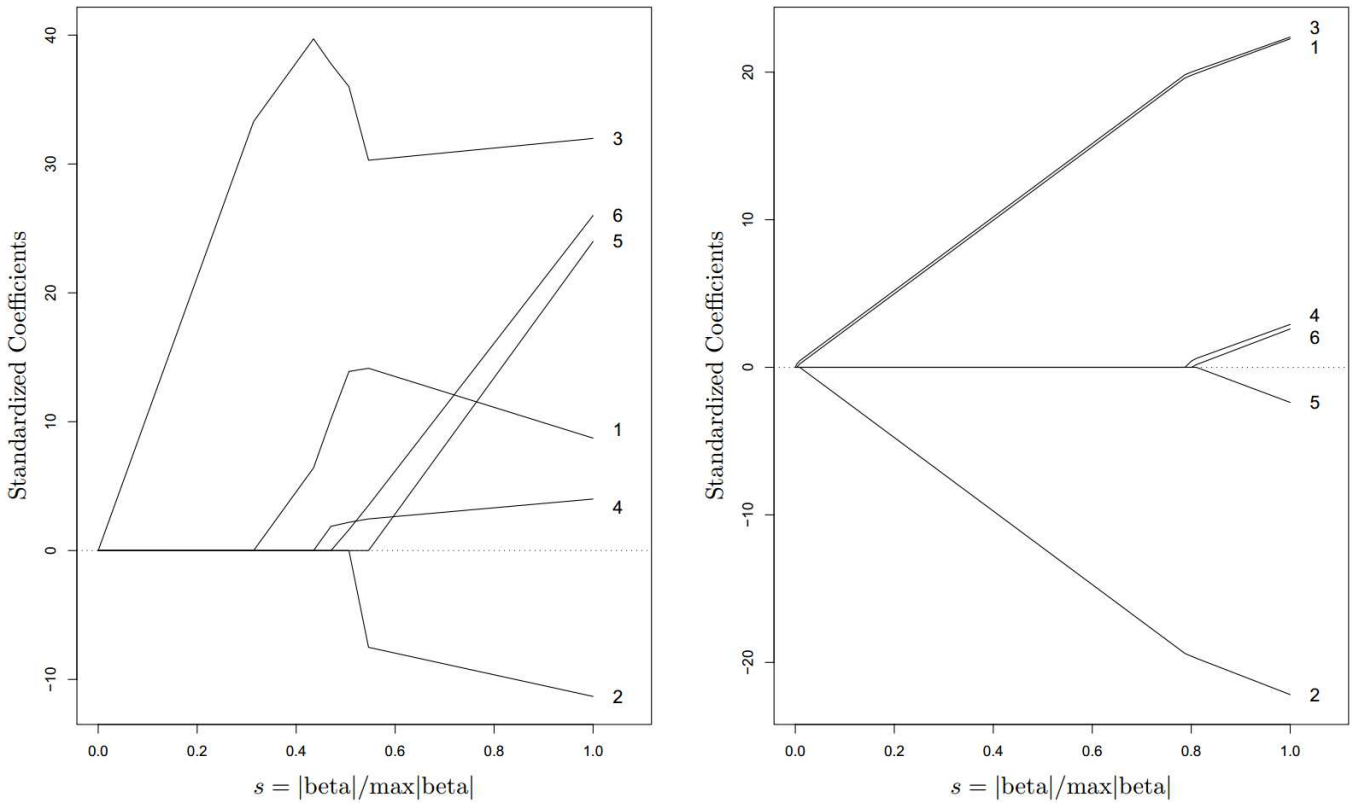
Значения параметров выбираются кросс-валидацией, λ_2 по небольшой сетке (например, $(0.0, 0.01, 0.1, 1, 10, 100)$), а λ_1 — по непрерывной кривой, получаемой алгоритмом LARS-EN.

Пример

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

$$X_1 = Z_1 + \varepsilon_1, \qquad X_2 = -Z_1 + \varepsilon_2, \qquad X_3 = Z_1 + \varepsilon_3,$$
$$X_4 = Z_2 + \varepsilon_4, \qquad X_5 = -Z_2 + \varepsilon_5, \qquad X_6 = Z_2 + \varepsilon_6,$$

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



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

Нелинейная регрессия

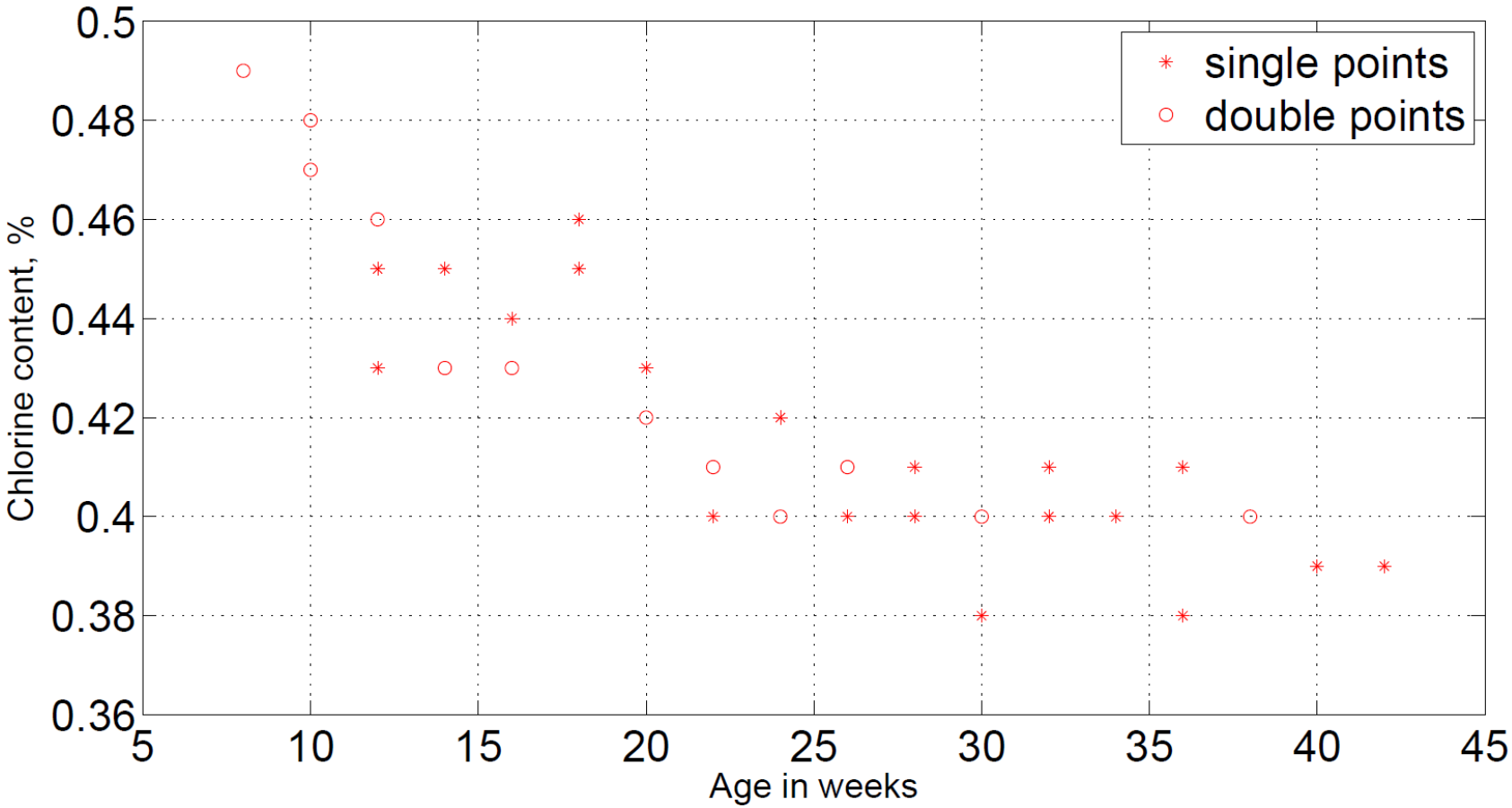
Пример: (Smith, Some reliability problems in the chemical industry, 1964) исследование компании Procter & Gamble. Исследуется продукт A, в момент производства доля свободного хлора в нём должна составлять 0.5. Известно, что со временем содержание хлора в продукте снижается. За первые 8 недель содержание хлора снизится до 0.49, но в более поздние сроки из-за влияния большого количества неконтролируемых факторов физические модели не могут достаточно точно предсказать содержание свободного хлора. Для определения закона убывания концентрации свободного хлора (y) она была измерена в 44 образцах на разных сроках хранения (x).

Была выдвинута гипотеза, что содержание хлора в продукте при $x \geq 8$ описывается уравнением вида

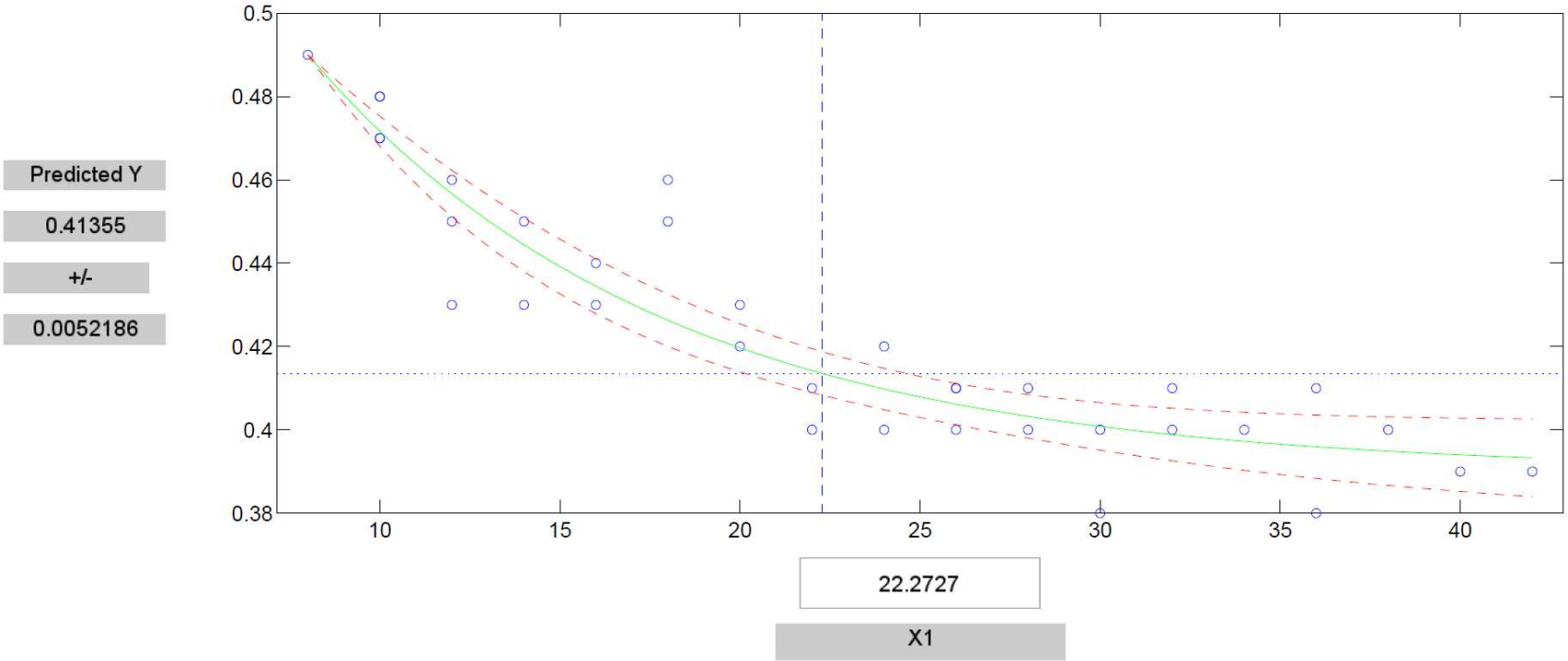
$$y = \alpha + (0.49 - \alpha)\, e^{-\beta(x-8)} + \varepsilon \equiv f\,(x) + \varepsilon.$$

Требуется оценить параметры α и β по наблюдениям $(x_i, y_i)\,, i = 1, \ldots, n.$

Данные


single points
double points
Chlorine content, %
Age in weeks
0.5
0.48
0.46
0.44
0.42
0.4
0.38
0.36
5
10
15
20
25
30
35
40
45

Модель



$\hat{\alpha} = 0.3901, \quad \hat{\beta} = 0.1016, \quad RSS = 0.00500168.$

Что дальше?

Сравнение RSS с чистой ошибкой

Чистая ошибка σ^2 — дисперсия ε, может быть оценена по повторяющимся наблюдениям.

$y_{11}, \ldots, y_{1n_1}$ — n_1 повторяющихся наблюдений при $x = x_1$,

...

$y_{m1}, \ldots, y_{mn_m}$ — n_m повторяющихся наблюдений при $x = x_m$.

$$\hat{\sigma}^2 = \frac{S_{pe}}{n_e} = \frac{\sum\limits_{j=1}^{m} \sum\limits_{u=1}^{n_j} (y_{ju} - \bar{y}_j)^2}{\sum\limits_{j=1}^{m} n_j - m}.$$

В примере $S_{pe} = 0.0024, \quad n_e = 26 \Rightarrow$

$$\frac{RSS - S_{pe}}{44 - 2 - n_e} = 0.00016,$$

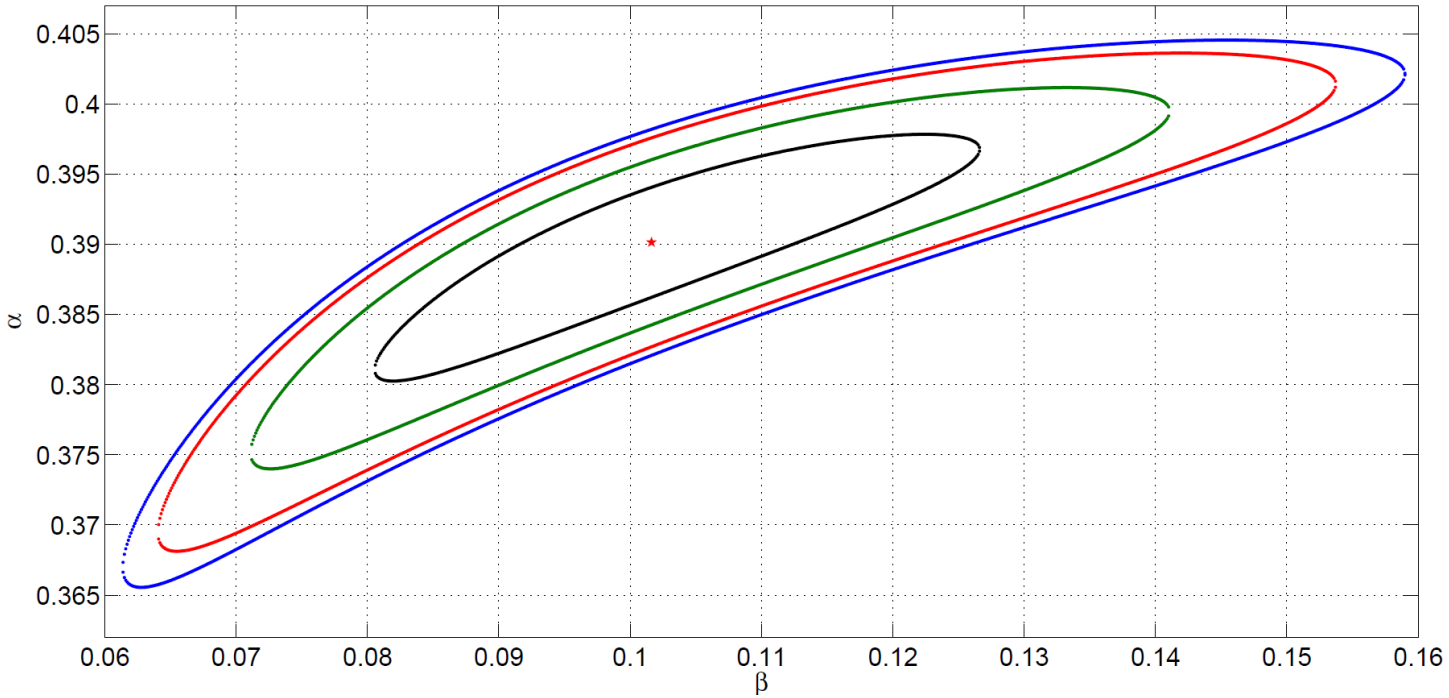
$$\frac{S_{pe}}{n_e} = 0.00009.$$

Используем критерий Фишера: $F(16, 26, 0.95) = 2.8, \quad \frac{0.00016}{0.00009} = 1.8$ — можно надеяться, что модель подобрана хорошо.

Доверительные области

Приблизительные $100\,(1-q)\,\%$ доверительные области для значений параметров α и β:

$$\sum_{i=1}^{n}\left(y_i - f\left(\alpha,\beta,x_i\right)\right)^2 = RSS\left(\hat{\alpha},\hat{\beta}\right)\times\left(1+\frac{k}{n-k}F_{1-q}\left(k,n-k\right)\right).$$



Синий контур — $q = 0.005$, красный — $q = 0.01$, зелёный — $q = 0.05$, чёрный — $q = 0.25$.

Литература

- регрессия натурального признака — Cameron, главы 2, 3, 5, 6;
- гребневая регрессия, лассо, эластичная сеть — Hastie, 3.4;
- нелинейная регрессия — Дрейпер, глава 24;

Дрейпер Н.Р., Смит Г. *Прикладной регрессионный анализ*. — М.: Издательский дом «Вильямс», 2007.
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.