Эконометрика-1

Филатов Александр Юрьевич

(Главный научный сотрудник, доцент ШЭМ ДВФУ)

alexander.filatov@gmail.com

http://vk.com/alexander.filatov, **http://vk.com/baikalreadings**

Лекции 5.1-5.2

Нелинейные модели.

Логит- и пробит-модели

Нелинейные модели, поддающиеся непосредственной линеаризации

Часто зависимость между y и регрессором x может носить нелинейный или даже немонотонный характер. При этом для их оценивания нужен тот же самый инструментарий, включая функцию ЛИНЕЙН в Excel.

Полиномиальные зависимости

$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \varepsilon.$

Вводим дополнительные переменные $x^{(2)}=x^2$, $x^{(3)}=x^3$,…, оцениваем модель $y = \theta_0 + \theta_1 x + \theta_2 x^{(2)} + \theta_3 x^{(3)} + \ldots + \varepsilon$ обычным МНК.

Замечание 1. Схема работает для любого числа нелинейно воздействующих на результат переменных.

Замечание 2. Самый простой способ учесть немонотонное воздействие фактора, эффект насыщения и т.д.

Замечание 3. Не следует прибегать к высоким степеням. Линейный член – рост, квадратичный – ускорение, кубичный – ???

Темпы роста инфляции стали сокращаться.

Гиперболические зависимости

Гипербола смещенная по вертикали.

$$y = \theta_0 + \frac{\theta_1}{x} + \varepsilon, \quad x \in (0; +\infty).$$

Случай A. $\theta_0 > 0, \theta_1 > 0.$

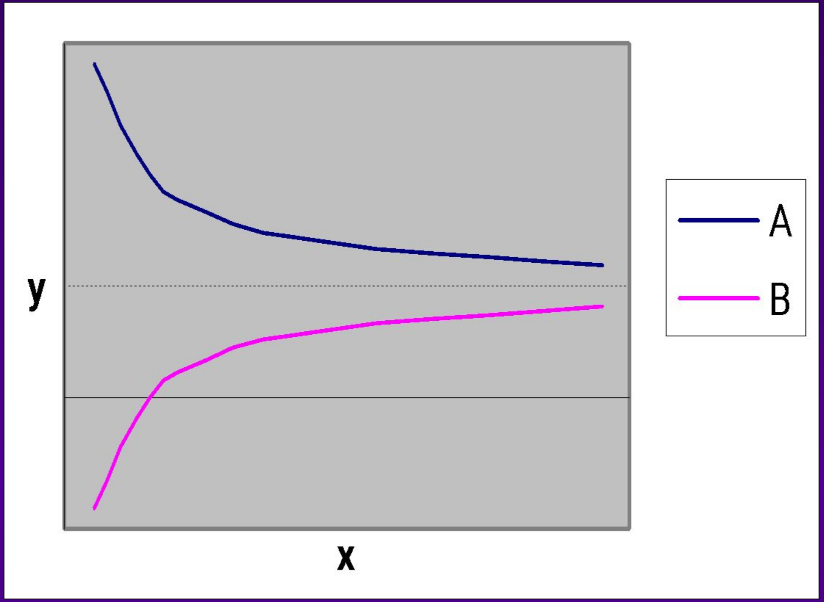
Случай B. $\theta_0 > 0, \theta_1 < 0.$

Случай C. $\theta_0 < 0, \theta_1 > 0.$

Вертикальная асимптота $x = 0$.

Горизонтальная асимптота $y = \theta_0$.

Замена $\tilde{x}_i = 1/x_i$, модель $y = \theta_0 + \theta_1 \tilde{x} + \varepsilon.$



Гипербола смещенная по двум осям. Величина x_0 задана.

$$y = \theta_0 + \frac{\theta_1}{x - x_0} + \varepsilon, \quad x \in (x_0; +\infty).$$

Наиболее типичный случай $\theta_0 < 0, \theta_1 > 0, x_0 < 0$

Вертикальная асимптота $x = x_0$.

Горизонтальная асимптота $y = \theta_0$.

Замена $\tilde{x}_i = 1/(x_i - x_0)$, модель $y = \theta_0 + \theta_1 \tilde{x} + \varepsilon.$

Экспоненциальные зависимости

Постоянный темп относительного прироста во времени.

5% / год = 132 раза / век, 10% / год = 13781 раз / век.

$y = \theta_0 e^{\theta_1 x + \varepsilon}, \; y \in (0; +\infty),$

θ_0 – начальный уровень,

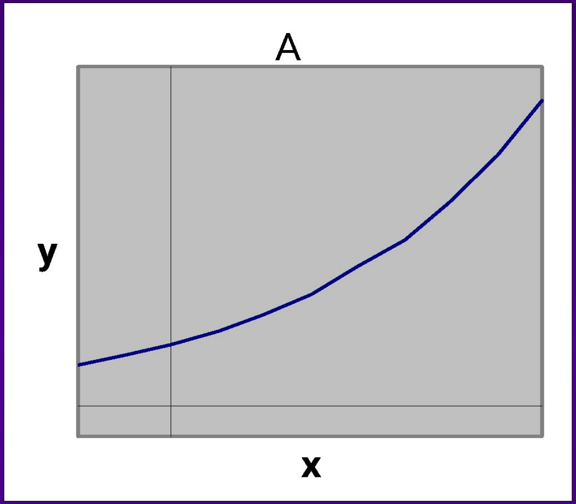
θ_1 – темп прироста.

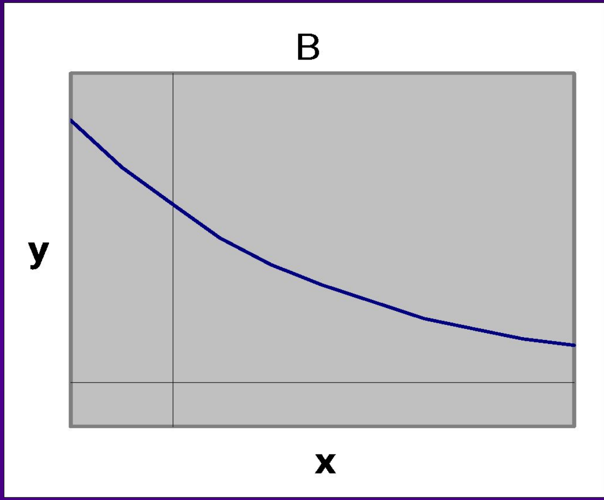
Случай А. $\theta_1 > 0$ – рост.

Случай В. $\theta_1 < 0$ - спад.

Замена $\tilde{y}_i = \ln y_i, \; \tilde{\theta}_0 = \ln \theta_0.$

Модель $\tilde{y} = \tilde{\theta}_0 + \theta_1 x + \varepsilon.$





Логистическая зависимость.

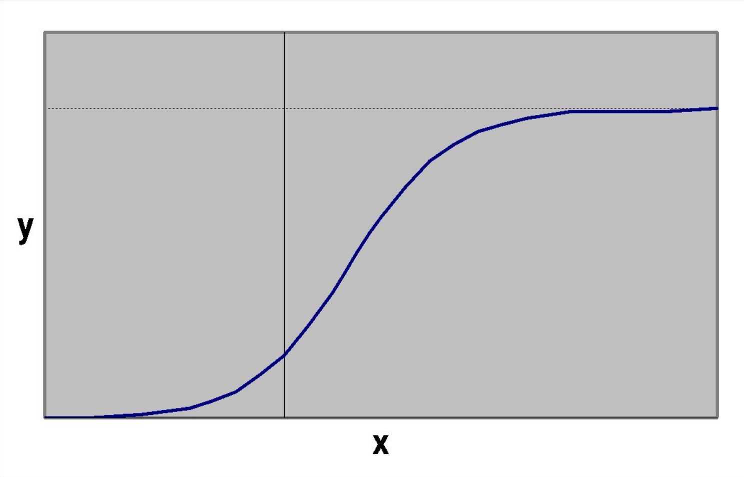
$$y = \frac{1}{\theta_0 + \theta_1 e^{-x} + \varepsilon}, \; x \in (-\infty; +\infty).$$

Горизонтальные асимптоты $y = 0$ и $y = 1/\theta_0$.

Пересечение оси в точке $y = 1/(\theta_0+\theta_1)$

Замена $\tilde{x}_i = e^{-x_i}, \; \tilde{y}_i = 1/y_i$

Модель $\tilde{y} = \theta_0 + \theta_1 \tilde{x} + \varepsilon$



Логарифмические зависимости

Самая медленно растущая из неограниченных функций.

Обратная функция к экспоненте.

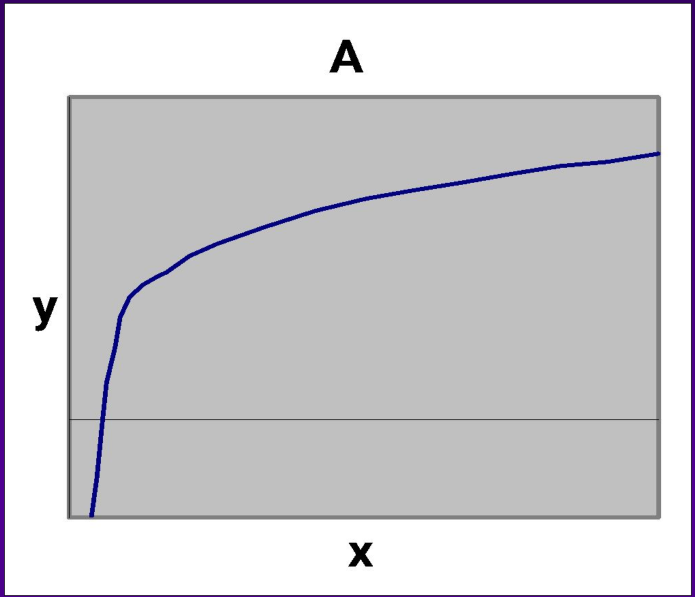
$y = \theta_0 + \theta_1 \ln x + \varepsilon, \;\; x \in (0; +\infty),$

Случай A. $\theta_1 > 0$ – неограниченный рост.

Случай B. $\theta_1 < 0$ – неограниченный спад.

Замена $\tilde{x}_i = \ln x_i$.

Модель $\tilde{y} = \theta_0 + \theta_1 \tilde{x} + \varepsilon$.

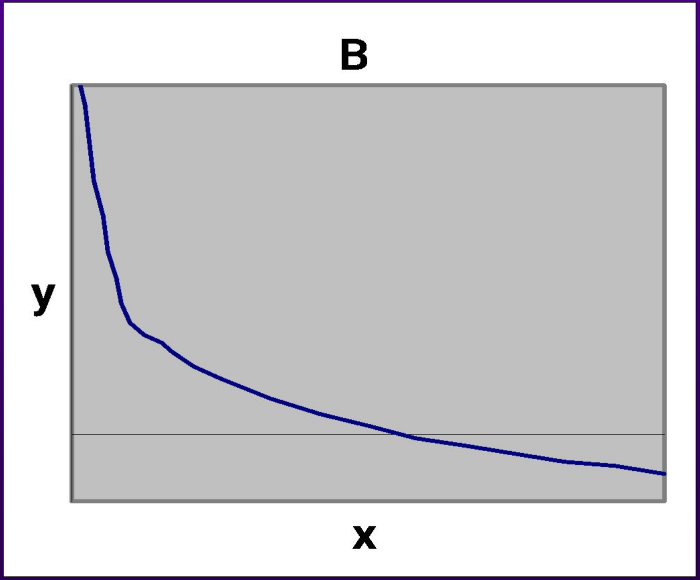


Модификация:

Так же, как и для гиперболической зависимости, возможен горизонтальный сдвиг на заранее зафиксированную величину x_0.

$y = \theta_0 + \theta_1 \ln(x - x_0) + \varepsilon, \;\; x \in (0; +\infty),$

$\tilde{x}_i = \ln(x_i - x_0). \quad \tilde{y} = \theta_0 + \theta_1 \tilde{x} + \varepsilon.$



Степенные зависимости

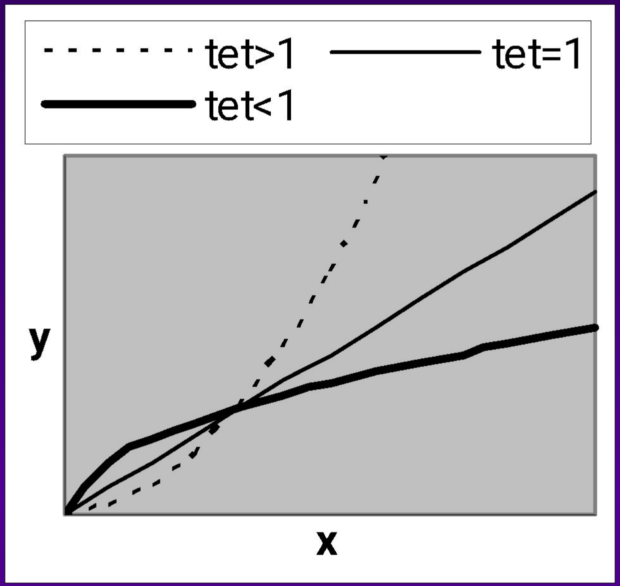
Функция с постоянной эластичностью.
Возможна множественная регрессия.

$$y = \theta_0 \left(x^{(1)}\right)^{\theta_1} \ldots \left(x^{(p)}\right)^{\theta_p} e^{\varepsilon}, \quad x^{(j)} \in (0; +\infty), \quad j = 1, \ldots, p.$$

θ_j – эластичности y по $x^{(j)}$.

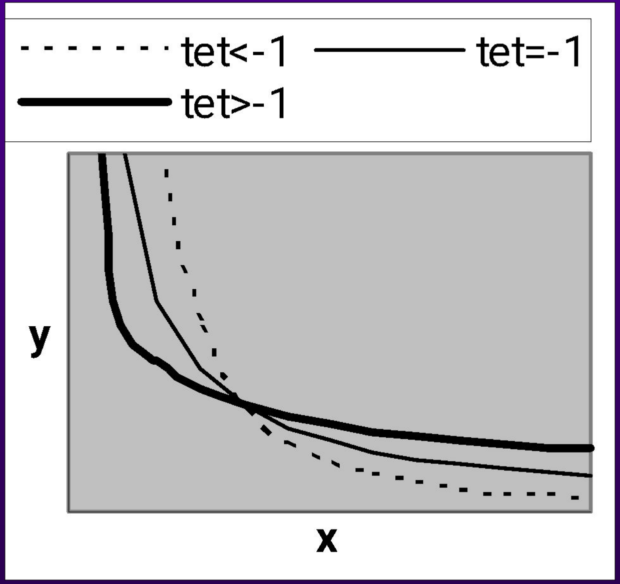
Замена $\tilde{y}_i = \ln y_i$, $\tilde{x}_i^{(j)} = \ln x_i^{(j)}$, $\tilde{\theta}_0 = \ln \theta_0$.

Модель $\tilde{y} = \tilde{\theta}_0 + \theta_1 \tilde{x}^{(1)} + \ldots + \theta_p \tilde{x}^{(p)} + \varepsilon$.



Различие монотонных функций:

Экспонента – линейна в логарифмических координатах, растет быстрее всех функций.
Степень – линейна в двойных логарифмических координатах, растет быстрее полинома.
Логарифм – растет медленнее всех функций!



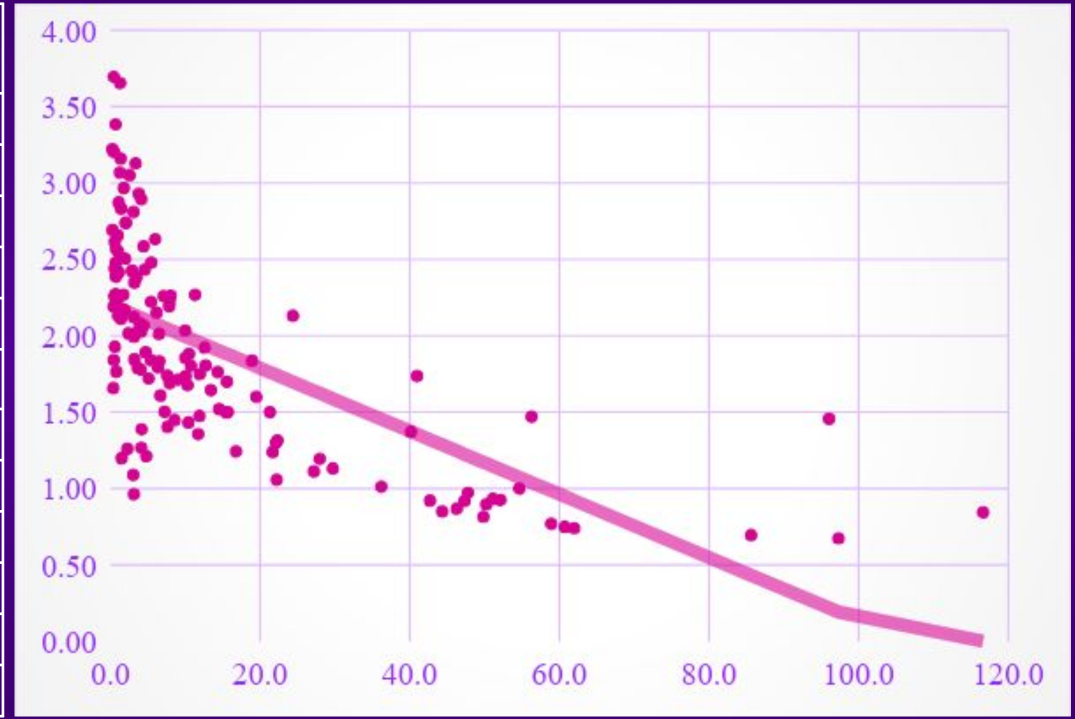
Закон Зипфа:

Многие экономические показатели (размеры городов, фирм, доходы богатых людей и т.д. распределены по степенном закону!

Численный пример

Взаимосвязь реального обменного курса y (во сколько раз цены в стране ниже, чем в США) и среднедушевого ВВП x по 138 странам за 2014 г.

Страна	x, GDP	y, RER
Норвегия	97 300	0,674
Швейцария	85 616	0,695
Австралия	61 979	0,741
США	54 629	1,000
Япония	36 194	1,012
Корея	27 970	1,194
Россия	12 736	1,805
Беларусь	8 040	2,262
Китай	7 590	1,740
Таджикистан	1 114	2,415
Гамбия	441	3,695
Малави	255	3,221



$\hat{y} = 2{,}20 - 0{,}021x, \quad \hat{R}^2 = 0{,}429,$

$\hat{y} = 2{,}48 - 0{,}339\ln x, \quad \hat{R}^2 = 0{,}555,$

$\hat{y} = 1{,}66 + 0{,}498/x, \quad \hat{R}^2 = 0{,}266,$

$\hat{y} = 2{,}49x^{-0{,}200}, \quad \hat{R}^2 = 0{,}498,$

$\hat{y} = 2{,}16e^{-0{,}013x}, \quad \hat{R}^2 = 0{,}488.$

Выбор вида зависимости. Метод проб и ошибок

Задача: выбрать из всевозможных видов моделей наилучшую.

1. Построить различные варианты моделей (полиномиальные, гиперболические, экспоненциальные, логарифмические, степенные и т.д.).
2. Оценить модели (найти значения всех коэффициентов).
3. Выбрать наилучшую из моделей, учитывая значение коэффициента детерминации, а также число оцениваемых параметров.

Двухкритериальная задача: $R^2_{y.X} = 1 - \frac{D\varepsilon}{Dy} \to \max, \ k \to \min$.

Максимизируем значение коэффициента детерминации, одновременно сокращая число оцениваемых параметров модели.

Замечание 1. Иногда максимизируют несмещенную оценку R^2:

$$\hat{R}^{2*}_{y.X} = 1 - \left(1 - \hat{R}^2_{y.X}\right)\frac{n-1}{n-p-1}$$

Замечание 2. Если y входит в модель линейно **(только в этом случае!)**, можем использовать оценку R^2, которую дает функция ЛИНЕЙН.

Замечание 3. Можем учесть значимость регрессоров и другие факторы.

Метод Бокса-Кокса

Метод Бокса-Кокса – автоматическая процедура подбора линеаризующего преобразования:

$$\tilde{y}_i(\lambda) = \frac{y_i^\lambda - 1}{\lambda}, \quad \tilde{x}_i^{(j)}(\lambda) = \frac{\left(\tilde{x}_i^{(j)}\right)^\lambda - 1}{\lambda}, \quad i = 1, \dots, n, \;\; j = 1, \dots, p.$$

Гипотеза: существует значение λ^*, такое что

$$\tilde{y}_i(\lambda^*) = \theta_0 + \theta_1 \tilde{x}_i^{(1)}(\lambda^*) + \dots + \theta_p \tilde{x}_i^{(p)}(\lambda^*) + \varepsilon_i$$

или

$$\tilde{y}_i(\lambda^*) = \theta_0 + \theta_1 x_i^{(1)} + \dots + \theta_p x_i^{(p)} + \varepsilon_i.$$

Замечание. Преобразования применяются исключительно к положительным переменным. Если для некоторой переменной имеются отрицательные значения, осуществляется сдвиг:

$$\tilde{y}_i(\lambda) = \frac{\left(y_i + c^{(0)}\right)^\lambda - 1}{\lambda}, \quad \tilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)} + c^{(j)}\right)^\lambda - 1}{\lambda},$$

$$c^{(j)} > \left| \min_{i=1,\dots,n} x_i^{(j)} \right|, \quad i = 1, \dots, n, \;\; j = 1, \dots, p.$$

Виды зависимостей в методе Бокса-Кокса

$\lambda^* = 1$ – линейная зависимость y от $x^{(1)},\ldots,x^{(p)}$.

$\lambda^* = 0$ – степенная или экспоненциальная зависимость y от $x^{(1)},\ldots,x^{(p)}$:

$$\tilde{y}_i(0) = \lim_{\lambda\to 0}\frac{y_i^{\lambda}-1}{\lambda} = \ln y_i,$$

$$\tilde{x}_i^{(j)}(0) = \lim_{\lambda\to 0}\frac{\left(x_i^{(j)}\right)^{\lambda}-1}{\lambda} = \ln x_i^{(j)}, \quad i = 1,\ldots,n, \quad j = 1,\ldots,p.$$

$$\ln\hat{y} = \theta_0 + \theta_1\ln x^{(1)} + \ldots + \theta_p\ln x^{(p)}, \quad \hat{y} = e^{\theta_0}\left(x^{(1)}\right)^{\theta_1}\ldots\left(x^{(p)}\right)^{\theta_p}$$ или

$$\ln\hat{y} = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}, \quad \hat{y} = e^{\theta_0}e^{\theta_1 x^{(1)}+\ldots+\theta_p x^{(p)}}.$$

При других значениях λ^* получаем связь некоторых степеней исходных переменных:

$\lambda^* = 0{,}5, \quad \tilde{y}(0{,}5) = \frac{y^{0,5}-1}{0{,}5} = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)},$

$$\hat{y} = \left(0{,}5\left(\theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}\right) + 1\right)^2,$$

Оценивание λ^*. Решетчатая процедура

1. Задаем интервал $\lambda^* \in [\lambda_{\min}; \lambda_{\max}]$, часто $\lambda^* \in [-1; 2]$.
2. С некоторым шагом $\Delta\lambda$
   1) Вычисляем значения $\tilde{y}_i(\lambda)$ и, при необходимости $\tilde{x}_i^{(j)}(\lambda)$;
   2) Находим оценки коэффициентов $\theta_j(\lambda)$ линейной регрессии
   $\tilde{y}(\lambda) = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}$ или
   $\hat{\tilde{y}}(\lambda) = \theta_0 + \theta_1 \tilde{x}^{(1)}(\lambda) + \ldots + \theta_p \tilde{x}^{(p)}(\lambda)$;
   3) Вычисляем коэффициент детерминации $\hat{R}^2(\lambda)$.
3. Строим зависимость $\hat{R}^2(\lambda)$ и находим $\lambda^* = \arg\max\limits_{\lambda} \hat{R}^2(\lambda)$.
4. Переходим в исходные координаты $y(x)$ и строим прогноз.

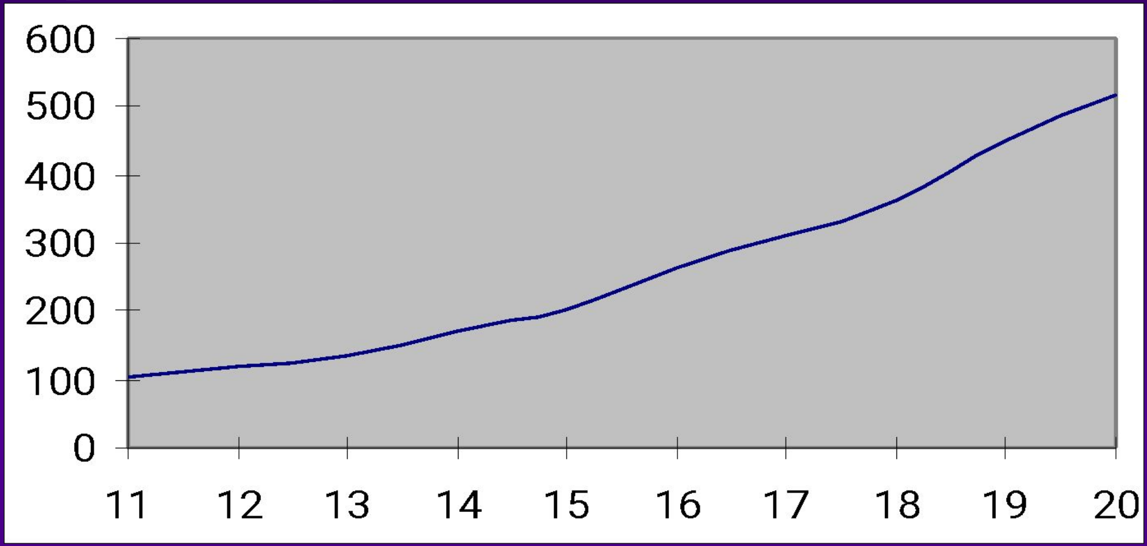
Замечание 1. При практической реализации решетчатой процедуры можно сначала оценить значение λ^* достаточно грубо, используя то, что при $\lambda \in (-\infty; \lambda^*)$ $\hat{R}^2(\lambda)$ монотонно возрастает, а при $\lambda \in (\lambda^*; +\infty)$ – убывает. Можно также использовать методы одномерной оптимизации.

Замечание 2. На некоторых практических задачах оптимальное значение λ^* находится вне интервала $[-1; 2]$.

Численный пример

Объем предложения акций на фондовом рынке y в зависимости от цены x

x, $	y, тыс.шт.
11	104
12	119
13	137
14	169
15	201
16	263
17	312
18	364
19	451
20	517



x	$y(-1)$	$y(-0,5)$	$y(-0,1)$	$y(0)$	$y(0,1)$	$y(0,5)$	$y(1)$	$y(2)$
11	0,9904	1,8039	3,7151	4,6444	5,9112	18,396	103	5408
12	0,9916	1,8167	3,7992	4,7791	6,127	19,817	118	7080
13	0,9927	1,8291	3,886	4,92	6,3558	21,409	136	9384
14	0,9941	1,8462	4,013	5,1299	6,7028	24	168	14280
15	0,9950	1,8589	4,1159	5,3033	6,9949	26,355	200	20200
16	0,9962	1,8767	4,272	5,5722	7,458	30,435	262	34584
17	0,9968	1,8868	4,369	5,743	7,7589	33,327	311	48672
18	0,9973	1,8952	4,4551	5,8972	8,0348	36,158	363	66248
19	0,9978	1,9058	4,5727	6,1115	8,4254	40,474	450	101700
20	0,9981	1,912	4,6463	6,248	8,6788	43,475	516	133644

Численный пример

Объем предложения акций на фондовом рынке *y* в зависимости от цены *x*

$\lambda = -1, \quad \tilde{y} = 0{,}9814 + 0{,}00088x, \quad \hat{R}^2 = 0{,}9604,$

$\lambda = -0{,}5, \quad \tilde{y} = 1{,}6689 + 0{,}0125x, \quad \hat{R}^2 = 0{,}9875,$

$\lambda = -0{,}1, \quad \tilde{y} = 2{,}5062 + 0{,}1083x, \quad \hat{R}^2 = 0{,}9961,$

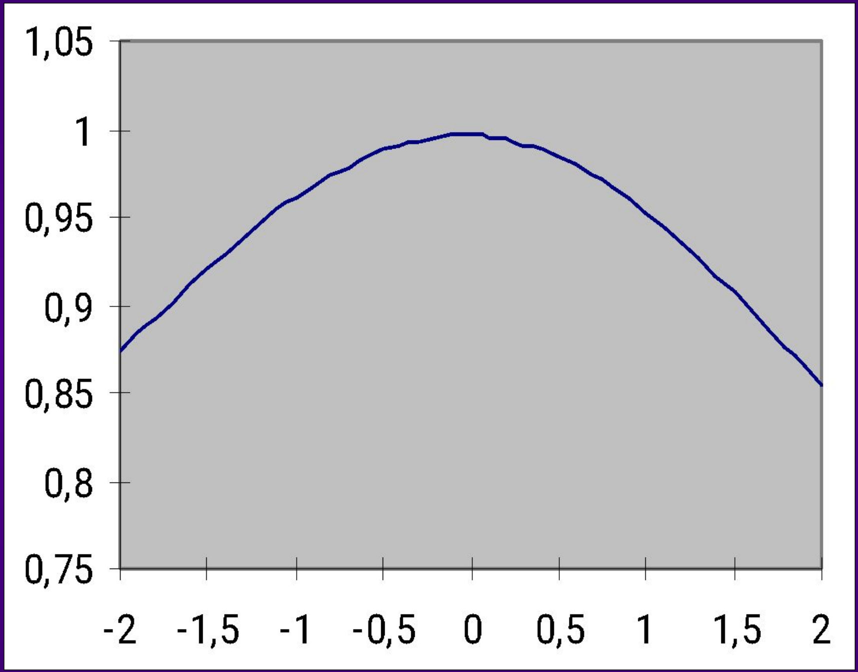
$\lambda = 0, \quad \tilde{y} = 2{,}5459 + 0{,}1864x, \quad \hat{R}^2 = 0{,}9962,$

$\lambda = 0{,}1, \quad \tilde{y} = 2{,}2638 + 0{,}3214x, \quad \hat{R}^2 = 0{,}9955,$

$\lambda = 0{,}5, \quad \tilde{y} = -15{,}34 + 2{,}8855x, \quad \hat{R}^2 = 0{,}9840,$

$\lambda = 1, \quad \tilde{y} = -457{,}5 + 46{,}467x, \quad \hat{R}^2 = 0{,}9524,$

$\lambda = 2, \quad \tilde{y} = -164270 + 13445x, \quad \hat{R}^2 = 0{,}8535.$



$$\max_{\lambda} \hat{R}^2(\lambda) = \hat{R}^2(0) = 0{,}9962,$$

$$\tilde{y}(0) = \ln y = 2{,}5459 + 0{,}1864x, \qquad \hat{y} = e^{2{,}5459 + 0{,}1864x} = 12{,}755e^{0{,}1864x}.$$

Бинарные результирующие показатели

$$x^{(1)}, x^{(2)}, \ldots, x^{(p)} \quad \Rightarrow \quad y = \begin{cases} 0 \\ 1 \end{cases}$$

Возраст, образование, стаж, желаемая зарплата $\Rightarrow$ безработный $= \begin{cases} 0 \\ 1 \end{cases}$

Применение обычной линейной регрессии $y = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}$

1. Проблема интерпретации результирующего показателя

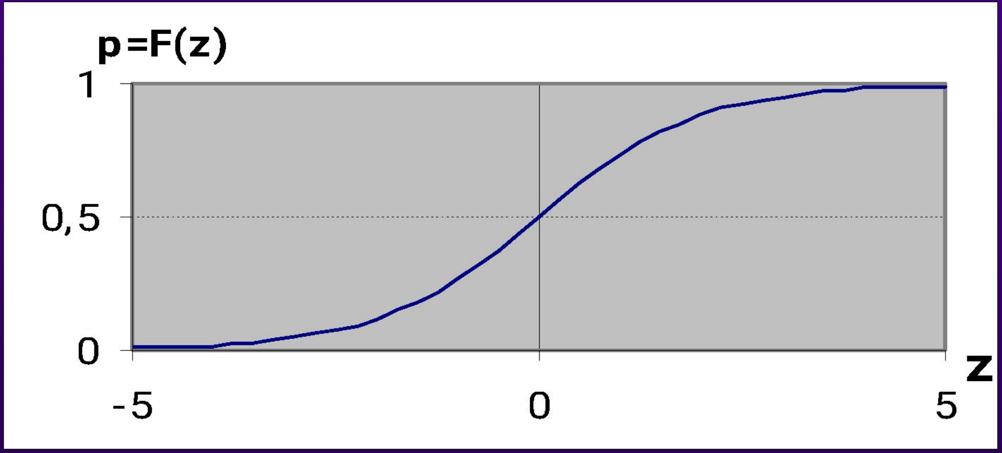
   Решение: можно интерпретировать как вероятность.

2. Прогнозируемая вероятность выходит за пределы отрезка [0; 1].

   Решение: необходимо подобрать преобразование F, переводящее интервал $(-\infty; +\infty)$ в [0; 1].

Требуемые свойства:

1. $F(z)$ – монотонно возрастает.
2. $F(z) \in [0; 1]$.
3. $F(z) \to 0$ при $z \to -\infty$.
4. $F(z) \to 1$ при $z \to +\infty$.



Логит- и пробит-модели

Наибольшее распространение получили две модели, основанные на следующих преобразующих функциях:

1. Логит-модель

Использует логистическую функцию

$$P(y_i = 1 : X_i) = \frac{e^{z_i}}{1 + e^{z_i}} = \frac{e^{X_i \Theta}}{1 + e^{X_i \Theta}} = \Lambda(X_i \Theta) = \Lambda(z_i)$$

2. Пробит-модель:

Использует функцию распределения нормального стандартного закона.

$$P(y_i = 1 : X_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_i \Theta} e^{-t^2/2} dt = \Phi(X_i \Theta) = \Phi(z_i)$$

Замечание:

Обе функции симметричны относительно $z = 0$ и в целом очень похожи по конфигурации.

Пробит-модель демонстрирует чуть более быстрое стремление вероятности к нулю и единице при высоких и низких значениях z.

Практическая реализация Логит- и пробит-моделей

Нужно сделать переход от индивидуальных наблюдений к частотам, поэтому для оценивания необходимы повторяющиеся исходные данные.

Вариант 1. Несколько наблюдений для каждого значения x.

Вариант 2. Несколько наблюдений для каждого интервала группировки:

$$\left.\begin{array}{l} y_{11};\ y_{12};\ldots;\ y_{1n_1};\quad X_{11}=X_{12}=\ldots=X_{1n_1}=X_1 \\ y_{21};\ y_{22};\ldots;\ y_{2n_2};\quad X_{21}=X_{22}=\ldots=X_{2n_2}=X_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{N1};y_{N2};\ldots;y_{Nn_N};\quad X_{N1}=X_{N2}=\ldots=X_{Nn_N}=X_N \end{array}\right\} \Rightarrow\ p_j=\frac{\sum_{i=1}^{n_j} y_{ji}}{n_j},\ \ j=1,\ldots,N$$

– относительная частота появления единиц для j-значения X

Логит-модель: $p=\dfrac{e^z}{1+e^z},\quad z=\ln\dfrac{p}{1-p}.$

Пробит-модель: $p=НОРМСТРАСП(z),\quad z=НОРМСТОБР(p).$

$z=\theta_0+\theta_1 x^{(1)}+\ldots+\theta_p x^{(p)}$, находим МНК-оценки коэффициентов $\hat\theta_0,\hat\theta_1,\ldots,\hat\theta_p$, при необходимости учитываем гетероскедастичность $D\varepsilon_j=F_j\left(1-F_j\right)\big/\left(n_j f_j^2\right)$.

Численный пример. Логит-модель

Доля владельцев автомобилей *p* в зависимости от среднедушевого дохода (*x*., тыс. руб./мес.)

x	*p*	*z*	*z^*	*p^*	*ε^*	*x*	*p*	*z*	*z^*	*p^*	*ε^*
10	0,2	-1,386	-1,478	0,186	0,014	110	0,6	0,405	0,542	0,632	-0,032
20	0,1	-2,197	-1,276	0,218	-0,118	120	0,8	1,386	0,744	0,678	0,122
30	0,2	-1,386	-1,074	0,255	-0,055	130	0,7	0,847	0,946	0,720	-0,020
40	0,3	-0,847	-0,872	0,295	0,005	140	0,8	1,386	1,148	0,759	0,041
50	0,2	-1,386	-0,670	0,339	-0,139	150	0,8	1,386	1,350	0,794	0,006
60	0,6	0,405	-0,468	0,385	0,215	160	0,9	2,197	1,552	0,825	0,075
70	0,4	-0,405	-0,266	0,434	-0,034	170	0,7	0,847	1,754	0,852	-0,152
80	0,8	1,386	-0,064	0,484	0,316	180	0,8	1,386	1,956	0,876	-0,076
90	0,5	0,000	0,138	0,535	-0,035	190	0,9	2,197	2,158	0,896	0,004
100	0,6	0,405	0,340	0,584	0,016	200	0,9	2,197	2,360	0,914	-0,014

$$z_i = \ln\frac{p_i}{1-p_i},\quad \hat{z}_i = -1{,}680 + 0{,}020x_i,\quad \hat{p}_i = \frac{e^{z_i}}{1+e^{z_i}},\quad \hat{K}_d = 1-\frac{D\varepsilon}{Dy} = 1-\frac{0{,}0118}{0{,}0679} = 0{,}826.$$

Численный пример. Пробит-модель

Доля владельцев автомобилей *p* в зависимости от среднедушевого дохода (*x*., тыс. руб./мес.)

x	*p*	*z*	*z*^	*p*^	*ε*^
10	0,2	-0,842	-0,879	0,190	0,010
20	0,1	-1,282	-0,759	0,224	-0,124
30	0,2	-0,842	-0,639	0,262	-0,062
40	0,3	-0,524	-0,518	0,302	-0,002
50	0,2	-0,842	-0,398	0,345	-0,145
60	0,6	0,253	-0,278	0,391	0,209
70	0,4	-0,253	-0,157	0,438	-0,038
80	0,8	0,842	-0,037	0,485	0,315
90	0,5	0,000	0,083	0,533	-0,033
100	0,6	0,253	0,204	0,581	0,019

x	*p*	*z*	*z*^	*p*^	*ε*^
110	0,6	0,253	0,324	0,627	-0,027
120	0,8	0,842	0,444	0,672	0,128
130	0,7	0,524	0,565	0,714	-0,014
140	0,8	0,842	0,685	0,753	0,047
150	0,8	0,842	0,805	0,790	0,010
160	0,9	1,282	0,926	0,823	0,077
170	0,7	0,524	1,046	0,852	-0,152
180	0,8	0,842	1,166	0,878	-0,078
190	0,9	1,282	1,287	0,901	-0,001
200	0,9	1,282	1,407	0,920	-0,020

$$z_i = НОРМСТОБР(p_i),\quad \hat{z}_i = -1{,}000 + 0{,}012x_i,\quad p_i = НОРМСТРАСП(z_i),$$

$$\hat{K}_d = 1 - \frac{D\varepsilon}{Dy} = 1 - \frac{0{,}0120}{0{,}0679} = 0{,}823.$$

Спасибо за внимание!

alexander.filatov@gmail.com
http://vk.com/alexander.filatov, **http://vk.com/baikalreadings**