Выбор вида зависимости

Задача: выбрать из всевозможных видов моделей наилучшую.

1. Метод проб и ошибок

1. Построить различные варианты моделей (полиномы, гиперболического, экспоненциального, степенного, логарифмического типа и др.).
2. Оценить модели (найти значения всех коэффициентов модели).
3. Выбрать наилучшую из этих моделей:
   1) Используя методы проверки гипотезы о виде функции регрессии
   2) По максимальному значению множественного коэффициента корреляции с учетом количества параметров модели (если y входит в модель линейно!):

$\hat{R}^2_{y.X} = 1 - \dfrac{D\varepsilon}{Dy} = 1 - \dfrac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2}$ $\qquad$ $\hat{R}^2_{y.X} = 1 - \dfrac{|R|}{|R|_{00}}$, R – корреляционная матрица.

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

2. Метод Бокса-Кокса

Метод Бокса-Кокса – формализованная процедура подбора линеаризующего преобразования:

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

Гипотеза: существует значение λ^*, такое что

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

Замечание 1

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

$$\tilde{y}_i(\lambda) = \frac{\left(y_i + c^{(0)}\right)^\lambda - 1}{\lambda},\ \ \tilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)} + c^{(j)}\right)^\lambda - 1}{\lambda},\ \ c^{(j)} > \left|\min_{i=1,...,n} x_i^{(j)}\right|,\ \ i = 1,..,n,\ \ j = 1,...,p.$$

Замечание 2

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

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

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

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

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

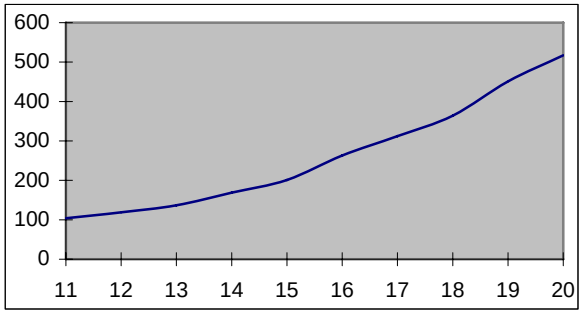
При других λ^* получаем связь каких-то степеней исходных переменных.

Оценка λ^* (решетчатая процедура)

1. Задается интервал $\lambda \in [\lambda_{\min}; \lambda_{\max}]$, часто $\lambda \in [-1; 2]$.
2. С некоторым шагом $\Delta\lambda$
   1) Вычисляются значения $\tilde{y}_i(\lambda)$ и, при необходимости, $\tilde{x}_i^{(j)}(\lambda)$;
   2) Находятся оценки $\hat{\theta}_j(\lambda)$ и множественный коэффициент корреляции $\hat{R}^2(\lambda)$.
3. Строится зависимость $\hat{R}^2(\lambda)$ и находится $\lambda^* = \arg\max_\lambda \hat{R}^2(\lambda)$.

Объем предложения акций на фондовом рынке в зависимости от цены

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



$\tilde{y}(\lambda) = \theta_0 + \theta_1 x$

x	$\tilde{y}(-1)$	$\tilde{y}(-0{,}5)$	$\tilde{y}(-0{,}1)$	$\tilde{y}(0)$	$\tilde{y}(0{,}1)$	$\tilde{y}(0{,}5)$	$\tilde{y}(1)$	$\tilde{y}(2)$
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

$\lambda = -1$, $\tilde{y} = 0{,}9814 + 0{,}000876x$, $\hat{R}^2 = 0{,}9604$.

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

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

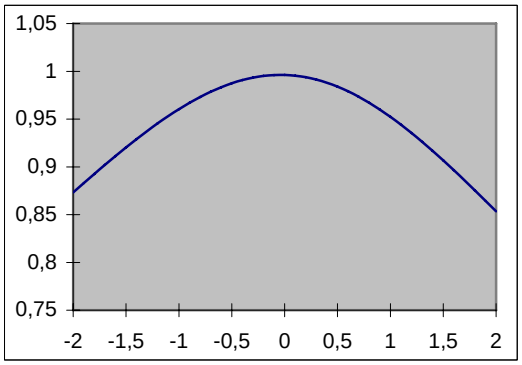
$\lambda = 0$, $\tilde{y} = 2{,}5459 + 0{,}1864x$, $\hat{R}^2 = 0{,}9962$.

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

$\lambda = 0{,}5$, $\tilde{y} = -15{,}341 + 2{,}8855x$, $\hat{R}^2 = 0{,}9840$.

$\lambda = 1$, $\tilde{y} = -457{,}53 + 46{,}467x$, $\hat{R}^2 = 0{,}9524$.

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



$\max\limits_{\lambda} \hat{R}^2(\lambda) = \hat{R}^2(0) = 0{,}9962$; с точностью до сотых $\lambda^* = -0{,}04$.

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

Замечание 1

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

Замечание 2

На некоторых практических задачах λ^* находится вне интервала $\lambda \in [-1; 2]$.

$\lambda^* = 0{,}5$ – квадратичная зависимость между исходными переменными:

$\tilde{y} = \dfrac{y^{0,5} - 1}{0{,}5} = \theta_0 + \theta_1 x$, $y^{0,5} = 1 + 0{,}5\theta_0 + 0{,}5\theta_1 x$, $\boxed{y = (1 + 0{,}5\theta_0)^2 + (1 + 0{,}5\theta_0)\theta_1 x + 0{,}25\theta_1^2 x^2}$.