

Тема 4. Моделиране на данни. Избор на най-подходящ модел.

1. Що е крива на модела.

Често пъти при изследването на един и друг реален феномен се започва с наблюдението и измерването му, т.е. със събирането на данни. Как могат да се използват тези данни? Един от възможните отговори на този въпрос ще дадем тази тема.

Пример 1.1. В тема 2, при изследване на модел 2 установихме, че скоростта на дъждовната капка е пропорционална на корен шести от масата, или $mv \propto \sqrt[6]{m}$, където c е неизвестна константа. За да определим тази константа, ние трябва да изберем дъждовни капки с различни маси и да измерим скоростта им. Тези данни могат да ни дадат възможност да намерим стойността на константата c .

Пример 1.2. През 1601 г. германският астроном Кеплер, е прекарал 13 години в наблюдение на относителното движение на Марс, в записването на данните и въз основа на тези данни през 1609 той е формулирал един основен закон: Всяка планета се движи по елипса, чийто фокуси съвпадат с тези на слънцето.

Ние няма да се спираме подробно на методите и начините за събиране на данни, а ще допуснем, че разполагаме с необходимите за целите ни акуратно събрани данни. Трите възможни проблема, които можем да решим, използвайки множеството от данни са:

- Съответства ли избрания модел на данните?
- Избор на най-добър модел, съответстващ на данните
- Прогнозиране на основата на дадените данни

В първите два въпроса се използва модел, за да се обясни поведението на наблюдавания обект. Тези два въпроса ще ги обединим под наименованието най-подходящ модел. Докато в третия въпрос модел не съществува, по-точно има събрана информация, която се използва, за да прогнозира поведението на обекта в някакви граници. В този случай ще казваме, че има опитен модел.

Освен това, за да се решат тези три въпроса е необходимо при първия въпрос прецизно да се дефинира понятието най-добър модел, съответстващ на данните, при втория въпрос е необходимо да даде критерий за сравняване на два избрани модела, а при третия въпрос трябва да се дадат методи за сравняване.

Да разгледаме реален обект, при който се интересуваме от взаимовръзката на два фактора, за които разполагаме с данни. При това за всяка стойност на единия единия фактор съответства стойност на другия. В този случай можем да разгледаме всички двойки данни като точки (x_k, y_k) в една координатна система. Ако тези точки лежат на една крива, то тази крива изразява връзката между двата фактора и се нарича крива на модела. Но ще отбележим, че било поради влиянието на пренебрегнати за разглеждане от нас фактори, било поради неточност в измерванията, или други фактори в реалния свят, винаги има отклонение на данните от коя да е крива. Затова ние разглеждаме абсолютното отклонение на всички точки от фиксирана крива, и целта ни е да намерим кривата на модела, като

крива, за която тези отклонения са възможно най-малки. Тази крива наричаме най-подходящата крива на модела.

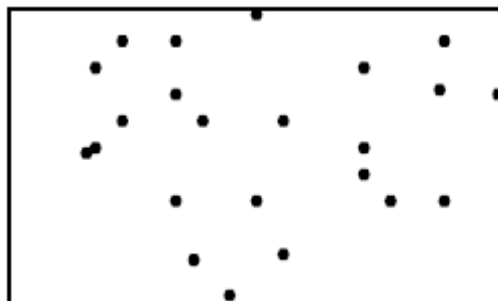


2. Избор на най-подходящата крива на модела, използвайки данни.

Често тази процедура се нарича построяване на опитен модел. Изследването се започва с внимателно анализиране на опитните данни, с които разполагаме. Има ли прои тези данни някаква тенденция? Има ли измежду данните очевидно грешни данни? Ако да, то ги пренебрегваме или правим отново измервания. Когато има очевидна тенденция в множеството от данни, ние се опитваме да намерим функцията, изразяваща тази тенденция на взаимовръзките на изследваните фактори.

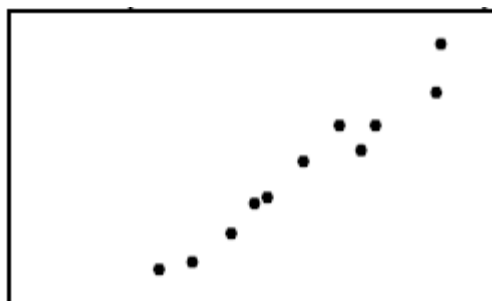
След нанасянето на дойките данни като точки, можем да наблюдаваме една от следните по-важни ситуации:

- Точките са разпръснати като облак – тогава няма подходяща крива на модела

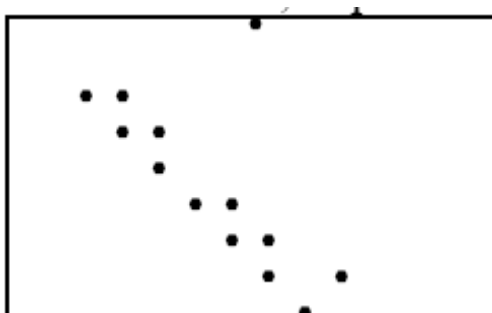


- Точките, определят някаква крива. Тази крива може да е права, крива на степенна функция (квадратична, кубична, експоненциална и пр.) или някаква друга. Ще разгледаме основните случаи:

Случай 1. Линеен случай. Данните приблизително лежат на една права.

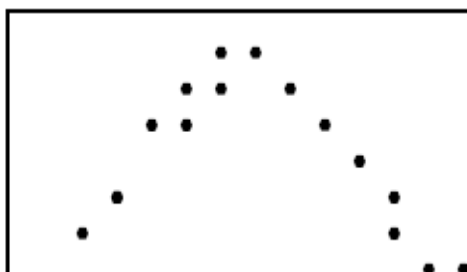


Например, на горната графика има линейна връзка между двете величини, която е положителна, нарастването на едната води и до нарастване на другата.



На горната графика има линейна връзка между двете величини, която е отрицателна, нарастването на едната води и до намаляване на другата.

Случай 2. Нелинеен случай. Данните приблизително лежат на някаква крива.



Например ако тази крива е параболата, това показва квадратна връзка между факторите, както на графика горе вляво.

В някои случаи нелинейните зависимости могат да се сведат до линейни модели чрез подходящи трансформации. Например, ако връзката е експоненциална, т.е. от вида $y = ce^{kx}$, след логаритмуване получаваме $\ln y = c + kx$, т.е. връзката между x и новата променлива $z = \ln y$ линейна (за по-подробно обяснение вж. т. 4).

Пример 2.1. Да разгледаме данните

x	1	2	3	4
y	8,1	22,1	60,1	165

След нанасяне на данните на координатна система, можем да направим предположение, че връзката между двете променливи е експоненциална. След логаритмична трансформация получаваме трансформирани данни

x	1	2	3	4
ln y	2,1	3,1	4,1	5,1

Новите данни, нанесени на координатна система показват наличието на линейна връзка между променливите $z = \ln y$ и x .

Пример 2.2. Цената на акциите на банка през периода 1993-1997 са дадени в таблицата. Коя е най-подходящата крива на модела?

година	1993	1994	1995	1996	1997
Цена p	\$8.5	\$9	\$13.17	\$21.67	\$33

Решение. След нанасяне точките на координатна система, кривата прилича на експоненциална или на част от парабола. След логаритмична трансформация получаваме

година	1993	1994	1995	1996	1997
ln p	2.14	2,2	2,58	3,08	3.5

Точките за новите данни не лежат дори приблизително на права. Затова параболата е най-подходящата крива на модела.

Ще отбележим, че при построяването на най-подходящ модел, се допускат грешки, които могат да бъдат

- Грешки от формулирането на проблема- пренебрегва се фактор, който оказва съществено влияние.
- Грешки от метода- метода за получаване аналитичния вид на модела и неговите коефициенти, е обикновен числен метод, който се характеризира с известна за метода грешка.
- Грешки от закръгляне на числата
- Грешки от измерване.

3. Аналитично построяване на най-подходящия модел.

Има различни аналитични методи за намиране коефициентите на най-подходящия модел, но ние ще разгледаме т.н. *метод на най-малките квадрати*. Идеята на този метод се състои в намирането на коефициентите така, че сумата от квадратите от отклоненията на всички точки до най-подходящата крива на модела да е минимална.

Нека са дадени данните

x	x ₁	x ₂	...	x _n
y	y ₁	y ₂	...	y _n

Да разгледаме най-елементарния случай, т.е. линеен модел от вида $y = ax + b$. Търсим коефициентите a и b така, че сумата от квадратите от отклоненията на точките от данните до кривата на модела да е минимална, т.е. функцията на две променливи

$$S = \sum_{k=1}^n (y_k - (ax_k + b))^2$$

Да имам минимум. Необходимо условие за това е коефициентите а и b да са решения на системата

$$\frac{\partial S}{\partial a} = 2 \sum_{k=1}^n (y_k - (ax_k + b))(-x_k) = 0$$

$$\frac{\partial S}{\partial b} = 2 \sum_{k=1}^n (y_k - (ax_k + b))(-1) = 0$$

или

$$a \sum_{k=1}^n x_k^2 + b \sum_{k=1}^n x_k = \sum_{k=1}^n y_k x_k$$

$$a \sum_{k=1}^n x_k + nb = \sum_{k=1}^n y_k$$

Решението на горната система е

$$a = \frac{n \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n \sum_{k=1}^n x_k^2 - (\sum_{k=1}^n x_k)^2}, b = \frac{\sum_{k=1}^n y_k - a \sum_{k=1}^n x_k}{n}$$

Ако моделът е от вида $y = Ax^m$, то намирането на коефициента А се свежда до

$$S = \sum_{k=1}^n (y_k - Ax_k^m)^2$$

минимизиране на функцията , за което необходимото условие е

$$\frac{dS}{da} = -2 \sum_{k=1}^n (y_k - Ax_k^m)(x_k^m) = 0$$

или

$$A = \frac{\sum_{k=1}^n x_k^m y_k}{\sum_{k=1}^n x_k^{2m}}.$$

Пример 3.1. Нека са събрани данните

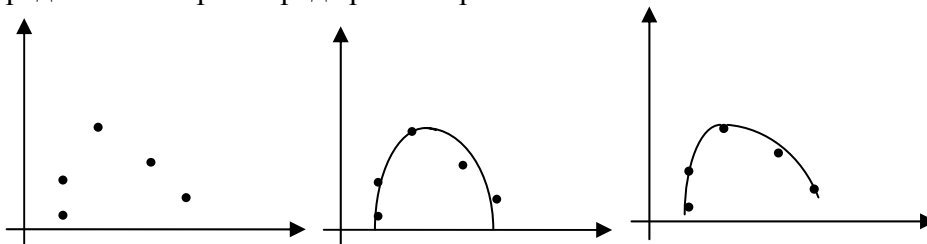
x	0,5	1	1,5	2	2,5
y	0,7	3,4	7,2	12,4	20,1

След нанасяне на данните на координатна система, правим извода, че връзката между x и y е квадратна, т.е. $y = Ax^2$. Тогава за коефициента А получаваме $A=3,1869$ или моделът е $y=3,1869x^2$.

Как можем да използваме този модел. Например, при $x=2,25$ можем да прогнозираме че $y=3,1869(2,25^2)=16,1337$.

4. Линеаризация на данни.

Ще изследваме по-подробно случая, когато данните, представени графично, не лежат приблизително на една права. В този случай, само от графиката няма възможност да се подбере най-подходящата крива. Например, да разгледаме данните на следващите графики. От графиките се вижда, че няма основание да се предпочете втората пред третата крива.



При тази ситуация е най-добре да се направи трансформация на изходните данни, която да трансформира данните в нови данни, на които графиката е приблизително права линия. Но кои са тези трансформации, и в какъв ред да ги прилагаме?

Разглеждаме следната скала от трансформации

Скала на трансформации
.....
z^2
z
$\sqrt{z}$
$\log z$
$-\frac{1}{\sqrt{z}}$
$-\frac{1}{z}$
$-\frac{1}{z^2}$
.....

Как действа всяка от трансформациите върху данните? За целта да разгледаме данните

x	1	2	3	4
y	1	4	9	16

Ще сравним относителния ефект на различните трансформации. Разглеждаме преобразованието на данните:

x	1	2	3	4
y	1	4	9	16
$z = \sqrt{y}$	1	2	3	4
$z = \log y$	0	0,6	0,95	1,2

$z = \frac{1}{\sqrt{y}}$	1	0,25	0,11	0,063
--------------------------	---	------	------	-------

При трансформацията $z = \sqrt{y}$, данните стават по близки една до друга, като това се отнася изключително много за по-големите стойности. Същото, дори в по-голяма степен се отнася и за $z = \log y$. От друга страна трансформацията $z = \frac{1}{\sqrt{y}}$ изменя

съществено поведението на данните- нарастването се променя с намаляване. Затова, за да се запази монотонността на данните, след $\log z$ в таблицата се поставят минуси, т.е. използва се трансформацията $z = -\frac{1}{\sqrt{y}}$. И така,

предпочитанието на изборът на трансформация става от таблицата, Ако изберем модел $y = ax^b$, то чертаем графика на $\ln y$ срещу $\ln x$

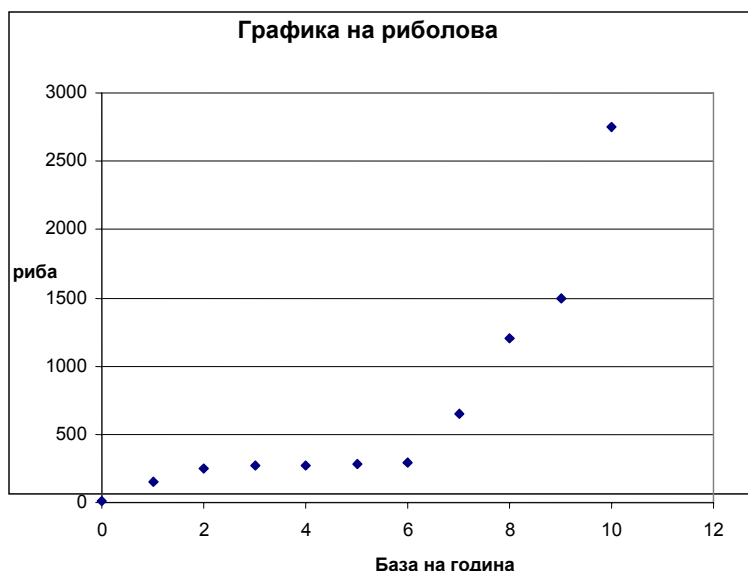
Пример 4.1. Да разгледаме данните, публикувани в пресата и събирани в продължение на 50 години за лова на риба и раци в залива Chesapeake Bay, САЩ.

Година	Риба (в тонове)	Раци (в тонове)
1940	15	100
1945	150	850
1950	250	1330
1955	275	2500
1960	270	3000
1065	280	3700
1070	290	4400
1975	650	4660
1980	1200	4800
1985	1500	4420
1990	2750	5000

Ще използваме тези данни за да изучим поотделно тенденциите при улова на риба и на раци. За целта първо променяме времето, като означаваме първата година с 0, а всяка останала данна в стълба за годините с последователни естествени числа.

База на годините	Риба (в тонове)	Раци (в тонове)
0	15	100
1	150	850
2	250	1330
3	275	2500
4	270	3000
5	280	3700
6	290	4400
7	650	4660
8	1200	4800
9	1500	4420
10	2750	5000

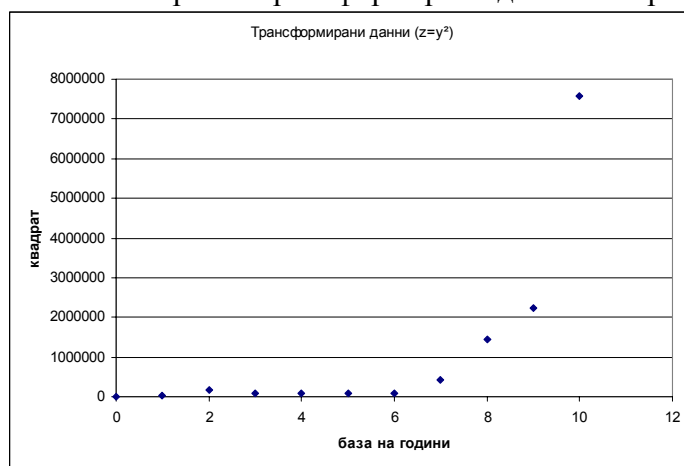
А/ **Улов на риба.** За целта нанасяме данните на графика (виж графиката по-долу).



Има тенденция на нарастване – графиката е на растяща и изпъкнала функция. Ще използваме трансформации. Можем да използваме трансформациите в таблицата нагоре по таблицата, т.е. да заместим z с z^2 , или z^3 или друга по-висока степен. Ще разгледаме например трансформацията z^2 .

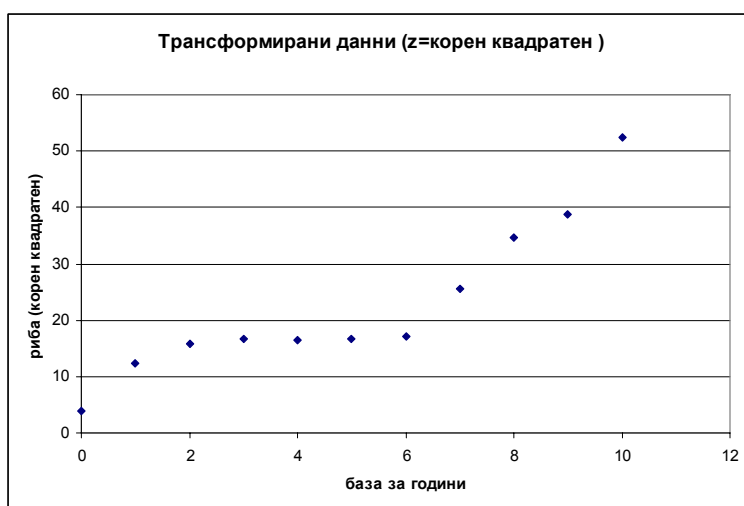
База на годините (x)	Риба (в тонове) y	Трансформирани данни ($z=y^2$)
0	15	225
1	150	22500
2	250	162500
3	275	75625
4	270	72900
5	280	78400
6	290	84100
7	650	422500
8	1200	1440000
9	1500	2240000
10	2750	7562500

Нанасяме горните трансформирани данни на графика:



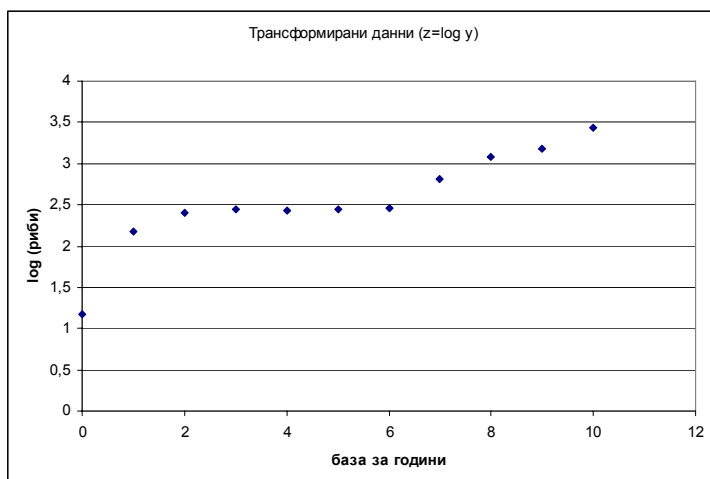
Горната графика е почти същата както и изходната, т.е. движението нагоре по стълбата на трансформациите не води в случая до успех. Затова започваме да се движим надолу. Разглеждаме трансформацията $\sqrt{z}$

База на годините (x)	Риба (в тонове) y	Трансформирани данни ($z=\sqrt{y}$)
0	15	3,873
1	150	12,2475
2	250	15,8114
3	275	16,5831
4	270	16,4317
5	280	16,7332
6	290	17,0294
7	650	25,4951
8	1200	34,6410
9	1500	38,7298
10	2750	52,4404



Горната графика отново не е особено полезна. Продължаваме по-нататък с трансформациите. Разглеждаме $\log z$

База на годините (x)	Риба (в тонове) y	Трансформирани данни (z=log y)
0	15	1,1761
1	150	2,1761
2	250	2,3979
3	275	2,4393
4	270	2,4314
5	280	2,4472
6	290	2,4624
7	650	2,8130
8	1200	3,0792
9	1500	3,1761
10	2750	3,4393



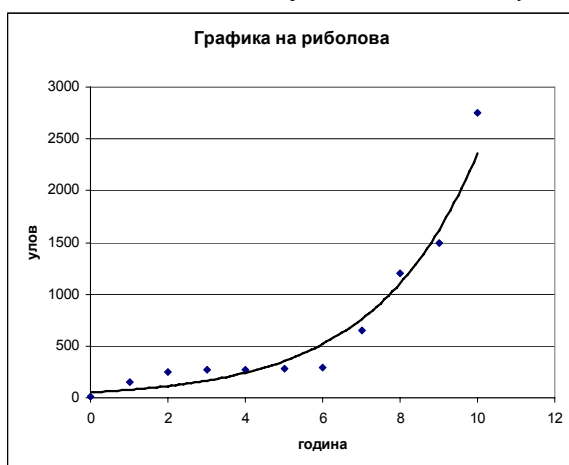
Тази графика вече е по-линейна от горнитет.е. ще използваме модела

$$\log y = mx + b$$

Използваме метода на най-малките квадрати, за да получим коефициентите

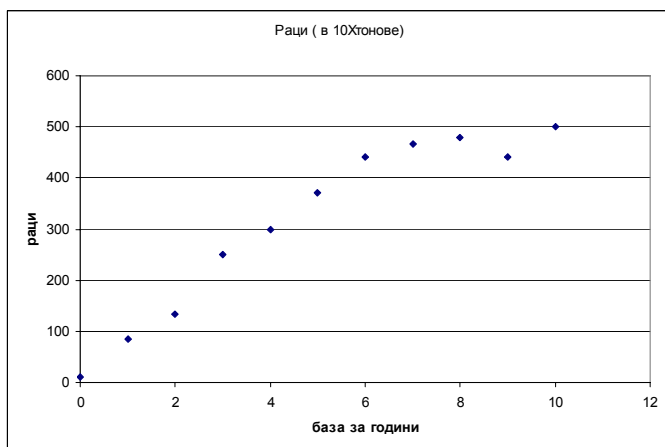
$$\log y = 0,1654x + 0,7231$$

или окончателно получаваме модела $y = 5,2857(1,4635)^x$ или



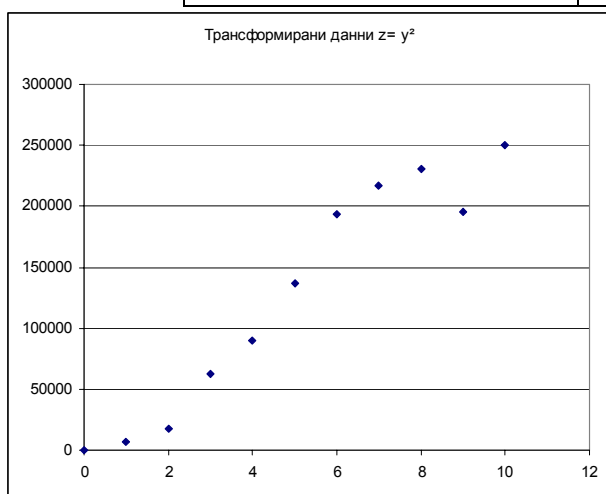
Б/ **Улов на раци.** Нанасяме на графика данните (предварително променяме и данните за рациите, като махаме по една 0, т.е. мерната единица е 10 тонове) като получаваме талицата от стойности

База на годините (x)	Раци (в тонове) (y)
0	10
1	85
2	133
3	250
4	300
5	370
6	440
7	466
8	480
9	442
10	500



Има тенденция на нарастване графиката е на растяща и вдлъбната функция. Както при изследване улова на риба, ще линеаризаме данните като използваме отначало данни от стълбата на трансформациите, движейки се нагоре, т.е. замествайки y с y^2 , или z^3 или друга по-висока степен. Например, след смяната $z = y^2$ получаваме

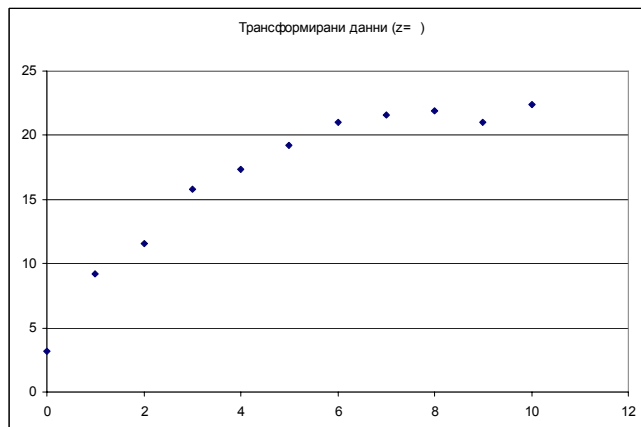
База на годините (x)	Трансформирани данни ($z = y^2$)
0	100
1	7225
2	17689
3	62500
4	90000
5	136900
6	193600
7	217156
8	230400
9	195364
10	250000



От горната графика се вижда, че трансформацията не линеаризира данните.

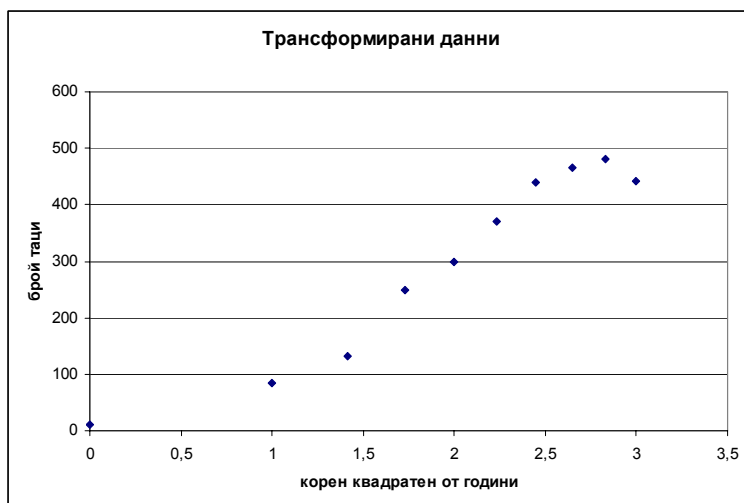
Затова продължаваме движението по стълбата , като използваме трансформацията $\sqrt{z}$. Отначало разглеждаме $z = \sqrt{y}$

База на годините (x)	Трансформирани данни ($z = \sqrt{y}$)
0	3,1623
1	9,2195
2	11,5326
3	15,81139
4	17,3205
5	19,2354
6	20,9762
7	21,5870
8	21,9089
9	21,0238
10	22,3607

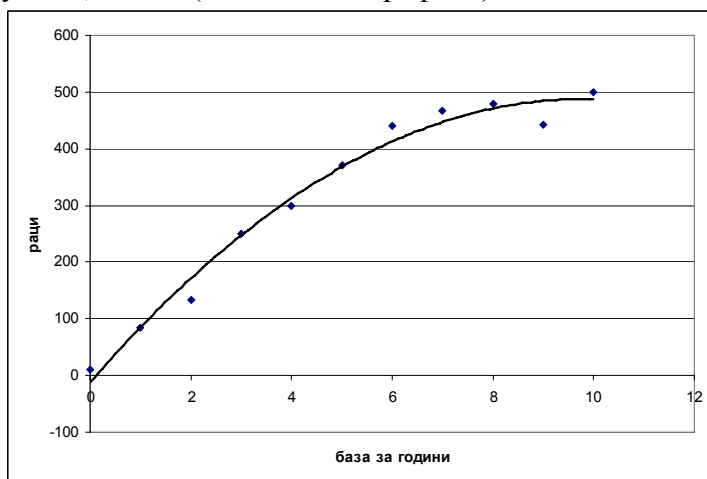


Тъй като и тази трансформация не линеаризира данните, разглеждаме $z = \sqrt{x}$

База на годините (x)	Трансформирани данни ($z = \sqrt{x}$)	Рази (в 10Хтонове)
0	0	10
1	1	85
2	1,4142	133
3	1,7321	250
4	2	300
5	2,2361	370
6	2,4495	440
7	2,6458	466
8	2,8284	480
9	3	442
10	3,1623	500



Следователно е подходящо да изберем функцията $y=k\sqrt{x}$. Използвайки метода на най-малките квадрати намираме коефициента k , т.е. получаваме функцията $y=158,344\sqrt{x}$ (вж. долната графика)



Как да използваме този модел? Можем например да прогнозираме улова през 2010 година, т.е. при $x=14$. Получаваме,

- за рибите

$$y = 5,2857(1,4635)^{14} = 1092,95 \approx 10,9 \text{ млн.кг}$$

- за раците

$$y=158,344\sqrt{14}=592,4=5,92 \text{ млн. Кг.}$$

Ще отбележим само, че моделът е по-добър за получаване на стойности, между данните (интерполация) отколкото за стойности в интервал извън данните (екстраполация).

Пример 4.2. Да разгледаме цената на акциите на банка през периода 1993-1997 са дадени в таблицата от пример 2.2 .

Година x	1993	1994	1995	1996	1997
Цена p	\$8.5	\$9	\$13.17	\$21.67	\$33

Съгласно пример 2.2 най-подходящата крива на модела е парабола, т.е. крива от вида $p=A x^2$. Тогава

$$A = \frac{1993^2 8,5 + 1994^2 9 + 1995^2 (13.17) + 1996^2 (21.67) + 1997^2 (33)}{1993^4 + 1994^4 + 1995^4 + 1996^4 + 1997^4} = 0,0000043.$$

Следователно, моделът е $p=0,0000043 x^2$. Можем да използваме този модел за прогнозиране, например, цената през 1998 година прогнозната цена е $1998^2(0,0000043)=17.16$ лв. (Ще отбележим, че за удобство, както в пример 4.1, годините могат да се кодират и по този начин се работи с по-малки цели числа).

Задача 4.3. Днес Интернет и имейлите са изключително важна съставна част в бизнес отношенията, особено в международните. Затова и изследванията в тази насока са особено важни с оглед на подобряване на тези връзки. През 2004 година, във в-к Newsweek беше публикувана следната статистическа информация

година	хора с интернет достъп (в млн.)	Дневни международни имейли (в млн.)
2001	190	80
2002	270	190
2003	350	360
2004	430	600

а/ Какъв тип функция е най-подходяща за моделиране на всяка от следните връзки

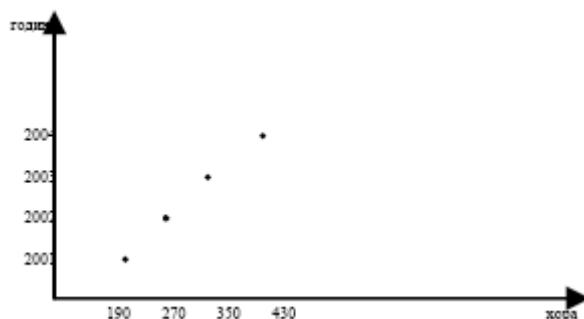
1. броят на хората с интернет достъп като функция на времето?
2. дневните международни имейли като функция на времето в години?

б/ Намерете функциите, които са най-подходящи за моделиране на горните две връзки.

в/ изследвайте броя на дневните международни имейли като функция на броя на хората с интернет достъп. Намерете подходящ модел. Какво прогнозира модела, ако броя на хората с интернет достъп е 650 000 000?

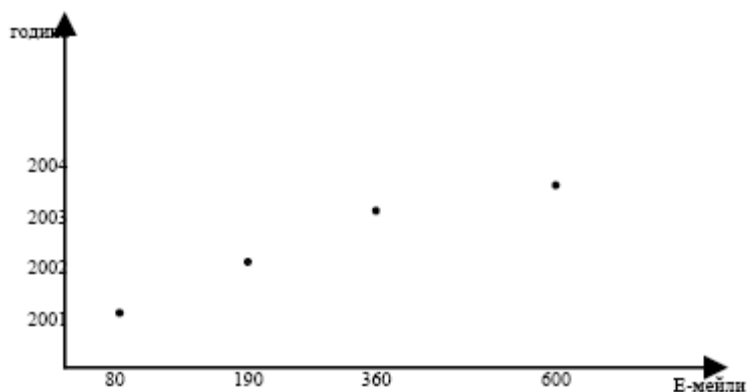
Решение.

А/ 1. Графиката е

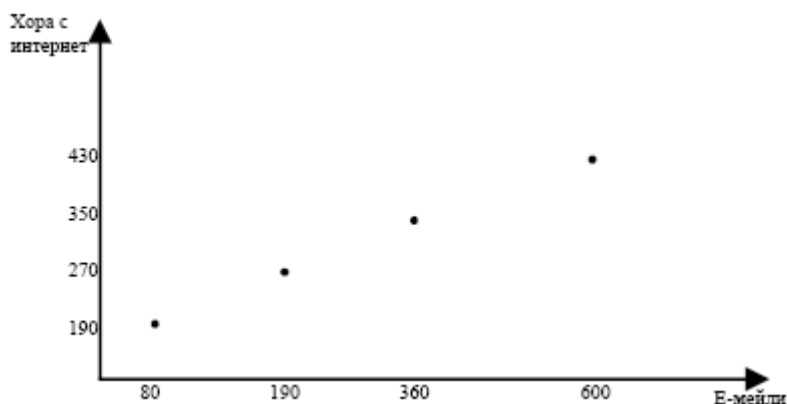


Следователно линейна функция е най-подходяща за моделиране броя на хората с интернет достъп z като функция на времето в години t , т.е. $z=at+b$.

2. Графиката е



Следователно нелинейна функция е най-подходяща за моделиране на дневните международни имейли y като функция на времето t , например $y = C \ln t$
в/ Графиката е



Следователно линейна функция е най-подходяща за моделиране на броя на дневните международни имейли y като функция на броя на хората с интернет достъп z , $y = Az + B$.

Тогава коефициентите A и B са

$$A = \frac{4(80 \cdot 190 + 190 \cdot 270 + 360 \cdot 350 + 600 \cdot 430) - (190 + 270 + 350 + 430)(80 + 190 + 360 + 600)}{4(190^2 + 270^2 + 350^2 + 430^2) - (190 + 270 + 350 + 430)^2} = 2,1625,$$

$$B = \frac{80 + 190 + 360 + 600 - 2,1625(190 + 270 + 350 + 430)}{4} = -362,875$$

или $y = 2,1625z - 362,875$. При $z = 650$ получаваме $y \approx 1043$ дневни международни имейли.