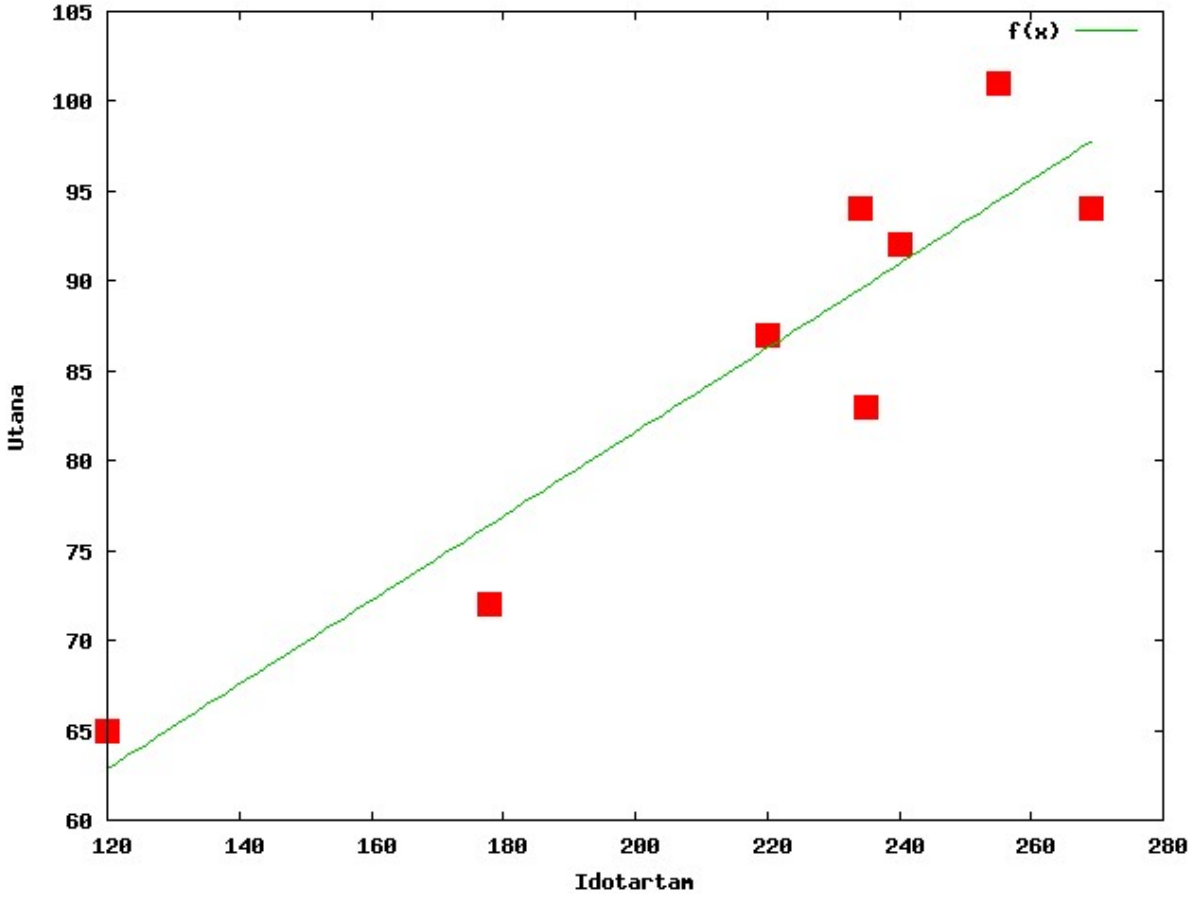
10. Előadás
Korreláció és regresszió

10-1. fejezet
Áttekintés

Old Faithful Geyser (Yellowstone)

- http://www.nps.gov/yell/tours/livecams/oldfaithful/OFVChours.htm
- A kitörések néhány adata percekben ill. méterekben
- **10-1. Táblázat:**

Időtartam	240	120	178	234	235	269	255	220
Előző	98	90	92	98	93	105	81	108
Következő	92	65	72	94	83	94	101	87
Magasság	42	33	38	36	42	36	38	45


f(x)
105
100
95
90
85
80
75
70
65
60
Utana
120
140
160
180
200
220
240
260
280
Idotartam

Áttekintés

Ebben a fejezetben bevezetjük a korreláció fogalmát, amelynek segítségével összefüggést lehet keresni két valószínűségi változó között, és bizonyos esetekben az egyik változó értékének ismeretében a másik értékére lehet következtetni.

Olyan mintákkal fogunk foglalkozni, ahol a minta adatok párokba vannak rendezve.

10-2. fejezet
Korreláció

Kulcsfogalmak

Ebben a fejezetben bevezetjük a lineáris korrelációs együttható *r* fogalmát, ami két véletlen változó közti kapcsolat erősségét számszerűen méri.

Mivel a korrelációs együttható könnyen kiszámítható, ezért itt főleg a fogalom megértésére koncentrálunk.

Definíció

Két változó között korreláció lép fel, ha az egyik a másikkal valamilyen módon kapcsolatban van.

Definíció

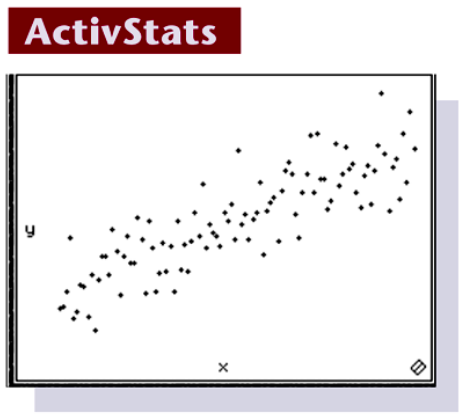
A lineáris korrelációs együttható r méri a lineáris kapcsolat erősségét egy x és y párokból álló minta értékei között.

Az adatok feltárása

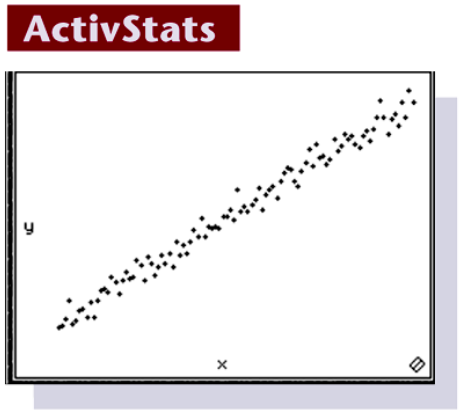
Gyakran felfedezhetünk kapcsolatot két változó között a szórásdiagram segítségével.

A következő 10-2. ábra néhány különböző tulajdonságokkal rendelkező szórásdiagramot mutat be.

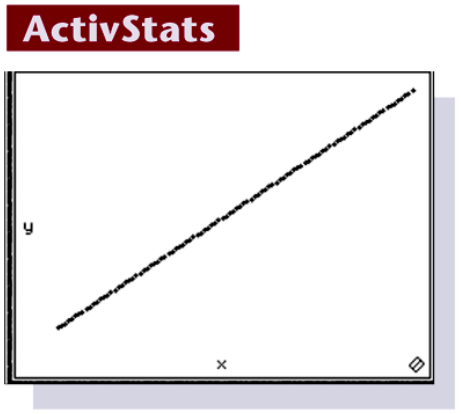
Szórásdiagramok párosított adatokra



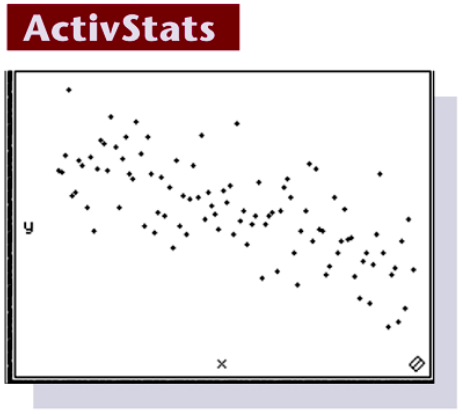
(a) Positive correlation:
$r = 0.851$



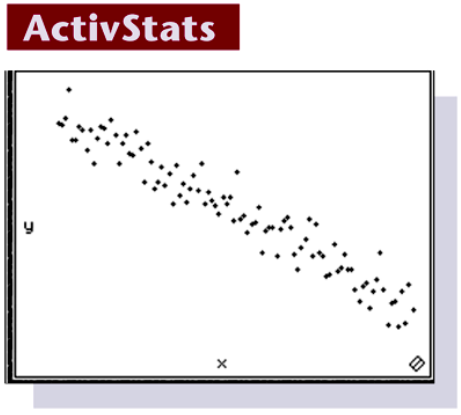
(b) Positive correlation:
$r = 0.991$



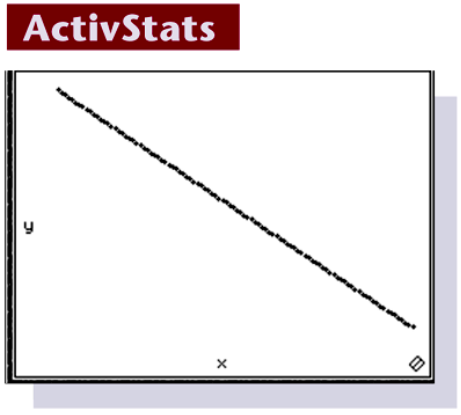
(c) Perfect positive correlation:
$r = 1$



(d) Negative correlation:
$r = -0.702$



(e) Negative correlation:
$r = -0.965$

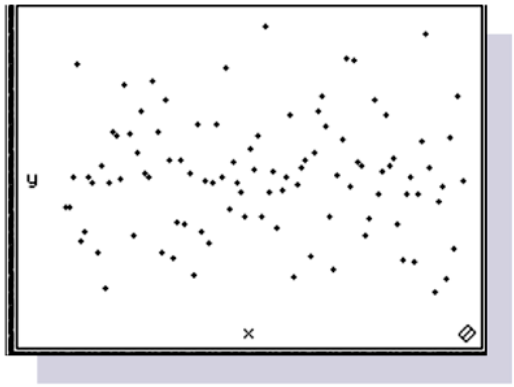


(f) Perfect negative correlation:
$r = -1$

10-2. ábra

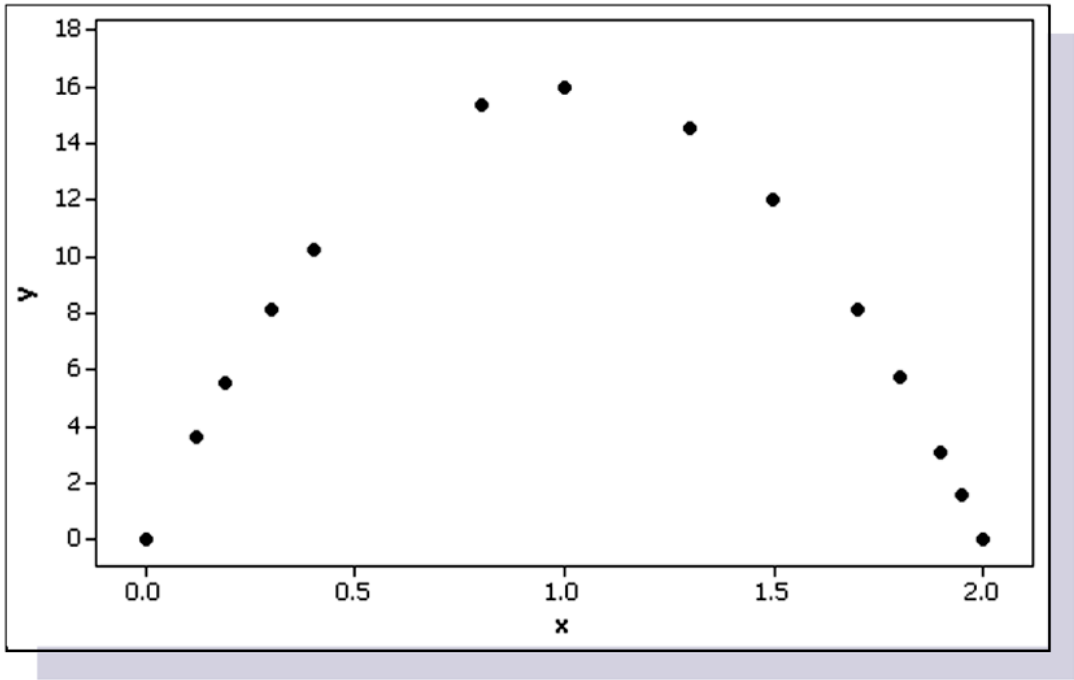
Szórásdiagramok párosított adatokra

ActivStats



(g) No correlation: $r = 0$

Minitab



(h) Nonlinear relationship: $r = -0.087$

10-2. ábra

Követelmények

1. Az (x, y) párokból álló adatok **véletlen** független minta adatok.

2. Vizuálisan meg kell győződnünk arról, hogy az adatok nagyjából egyenest alkotnak. (Nem determinisztikusak vagy más bonyolultabb alakjuk van.)

3. Az outliereket el kell távolítani, amennyiben meggyőződtünk arról, hogy hibásak voltak. Az *r* értékét ki kell számítani az outlierekkel együtt és azok nélkül. Meg kell nézni, mekkora a hatásuk.

Jelölések

n	az adatpárok száma
Σ	az adott értékek összegzése
Σx	az x értékek összege
Σx^2	minden x értéket négyzetre kell emelni és utána összeadni
$(\Sigma x)^2$	először össze kell adni az x értékeket, majd az eredményt négyzetre kell emelni
Σxy	minden x értéket meg kell szorozni a párjának y értékével, majd a szorzatokat összeadni
r	a minta lineáris korrelációs együtthatója.
ρ	a populáció lineáris korrelációs együtthatója.

Képletek

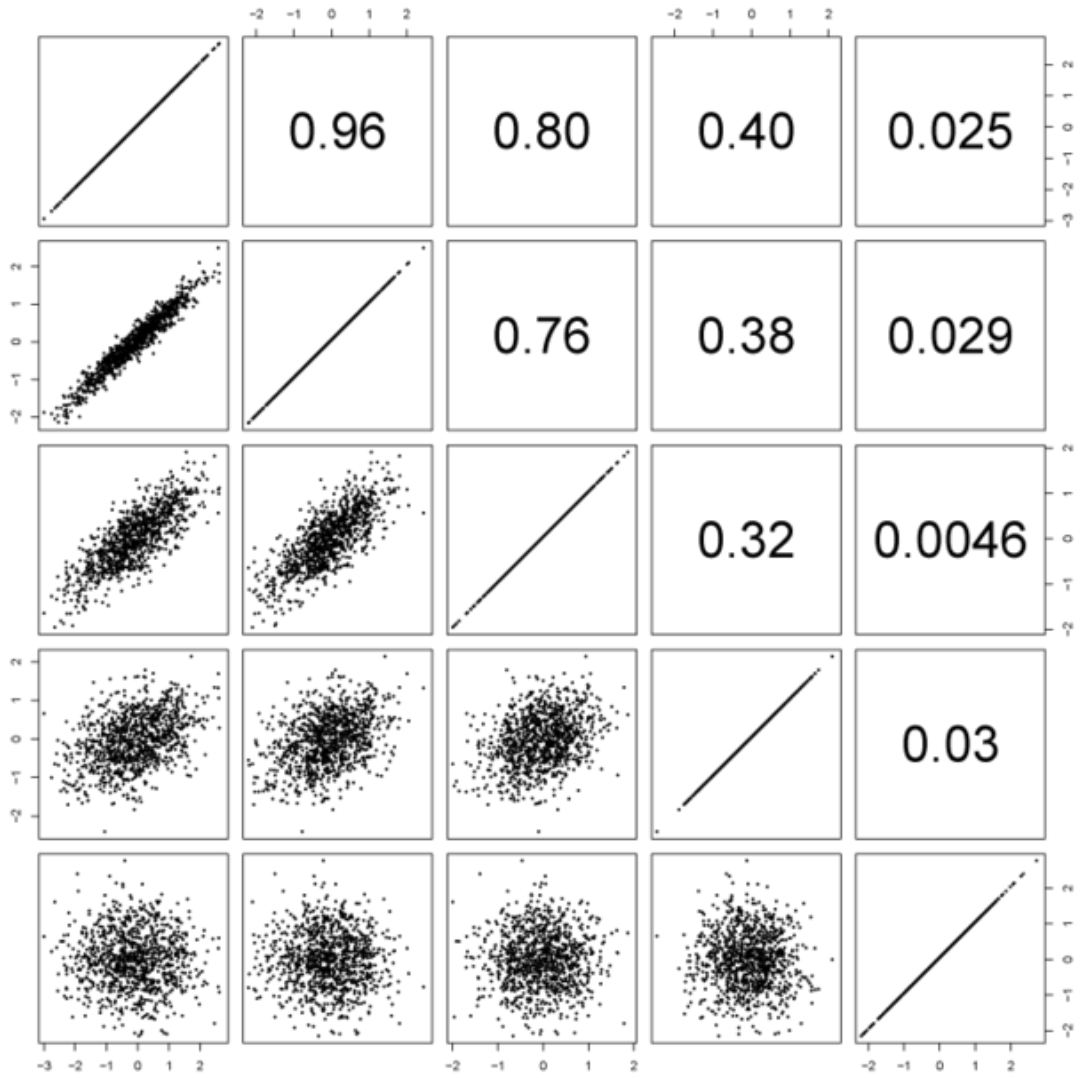
Az *r* lineáris korrelációs együttható méri a lineáris kapcsolat erősségét a minta adatpárok tagjai között (x és y között!).

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

10-1. képlet

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

1000 normális eloszlású adatpár különböző r értékekkel



r interpretálása

Táblázat: **Ha az *r* abszolút értéke nagyobb, mint a következő táblázatban, akkor arra következtetünk, hogy van lineáris korreláció.**

Critical Values for the Correlation Coefficient		
Number of Points	95% Confidence	99% Confidence
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708

Példa: *r* kiszámítása

Az alábbi egyszerű véletlen mintaadatokat használva számítsuk ki *r* értékét.

Adatok:

x	3	1	3	5
y	5	8	6	4

Table 10-2 Finding Statistics Used to Calculate *r*

	x	y	$x \cdot y$	x^2	y^2
	3	5	15	9	25
	1	8	8	1	64
	3	6	18	9	36
	5	4	20	25	16
Total	**12**	**23**	**61**	**44**	**141**
	↑	↑	↑	↑	↑
	Σx	Σy	Σxy	Σx^2	Σy^2

Példa: folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2}\ \sqrt{4(141) - (23)^2}}$$

$$r = \frac{-32}{33.466} = -0.956$$

Példa: folyt.

Adott r = - 0.956, ha 0.05-ös szignifikancia szintet használunk, akkor arra jutunk, hogy van lineáris kapcsolat x és y között, mivel r abszolút értéke meghaladja a 0.950-ös kritikus értéket. Azonban, ha a 0.01-es szignifikancia szintet használjuk, akkor nem jutunk arra, hogy lineáris kapcsolat van, mert r abszolút értéke nem haladja meg a 0.990-es kritikus értéket.

Példa: Old Faithful

A 10-1. táblázat adatait használva, keressük meg a lineáris korrelációs együttható értékét *r*, majd ellenőrizzük, hogy van-e szignifikáns lineáris kapcsolat a változók között.

Ugyanúgy számolva, mint előbb $r = 0.926$ adódik.

A táblázatban az $n = 8$ adatpont esetét keressük ki. Az $\alpha = 0.05$-höz tartozó értéket leolvasva, 0.707 kritikus értéket kapunk. Mivel $r = 0.926$, abszolút értéke több mint 0.707, úgy döntünk, hogy van lineáris kapcsolat a kitörések hossza és az utánuk következő várakozási idők között.

A lineáris korrelációs együttható tulajdonságai

1. $-1 \leq r \leq 1$
2. Az r értéke nem változik, ha bármelyik változónak megváltoztatjuk a mértékegységét.
3. Az r értékét nem befolyásolja az x és y felcserélése.
4. r méri a lineáris kapcsolat erősségét.

Interpretáció: Megmagyarázott variabilitás

Az r^2 érték mondja meg, hogy y variabilitásának hányad részét magyarázza az x és y közti lineáris kapcsolat.

Példa: Old Faithful

A kitörés után eltelő idő ingadozásának mekkora részét magyarázza meg a kitörés időtartamának ingadozása?

r = 0.926, akkor r^2 = 0.857.

Azt mondhatjuk, hogy 0.857-ed részét (vagy 86%-át) magyarázza meg a kitörések után eltelő idő ingadozásának a kitörés hosszával való lineáris kapcsolata. Ez azt is jelenti, hogy a kitörések után eltelő idő hosszának 14%-ára nem ad magyarázatot a kitörések hossza.

Szokásos hibák a korrelációval kapcsolatban

1. **Oksági összefüggés: Hibás azt állítani, hogy a korreláció oksági kapcsolatot jelent.**

2. **Átlagolás: Az átlagolás elnyomja az az eredeti adatokban meglévő ingadozásokat, ami módosítja a korrelációs együtthatót.**

3. **Linearitás: Lehetséges, hogy van valamilyen kapcsolat x és y között, még akkor is, ha nincs köztük lineáris korreláció.**

Formális hipotézis tesztelés

- **Szeretnénk meghatározni, hogy van-e szignifikáns lineáris kapcsolat két változó között.**
- **Legyen a null és alternatív hipotézis:**

H_0: $\rho = 0$ **(nincs szignifikáns lin. korreláció)**

H_1: $\rho \neq 0$ **(szignifikáns lin. korreláció)**

Teszt statisztika

Teszt statisztika:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

A transzformáció után t statisztika!

Kritikus értékek:

Megegyezik az n-2 szabadsági fokú Student *t* statisztikával!

Összefoglalás

Ebben a fejezetben megvitattuk a:

- **Korrelációt.**
- **A lineáris korrelációs együtthatót.**
- **A feltételeket .**
- **Az interpretációt.**
- **Formális hipotézis tesztelést.**

10-3. fejezet
Regresszió

Kulcsfogalmak

A legfontosabb ebben a fejezetben, hogy meghatározzuk azt az egyenest, és azt az egyenletet, ami legjobban reprezentálja a változók közti kapcsolatot.

Az egyenest regressziós egyenesnek nevezik és az egyenletet regressziós egyenletnek.

Regresszió

A regressziós egyenlet az x változó (független változó, prediktor változó vagy magyarázó változó) , és az y változó (függő változó vagy válasz változó vagy magyarázott változó) közötti kapcsolatot adja meg.

A tipikus lineáris kapcsolatot $\hat{y} = mx + b$, *vagy* az $\hat{y} = b_0 + b_1x$, formában fejezzük ki, ahol b_0 az y-tengelymetszet és b_1 a meredekség.

Feltételek

1. **Az adatpárok (x, y) véletlen minta adatok.**

2. **Vizuális vizsgálattal arra jutunk, hogy a szórásdiagram egy egyeneshez hasonló.**

3. **Ki kell hagyni azokat az outliereket, amik hibák miatt vannak jelen.**

Definíciók

❖ Regressziós egyenlet

Az adatpárok egy halmaza esetén a regressziós egyenlet:

$$\hat{y} = b_0 + b_1 x$$

algebrailag leírja a kapcsolatot a két változó között.

❖ Regressziós egyenes

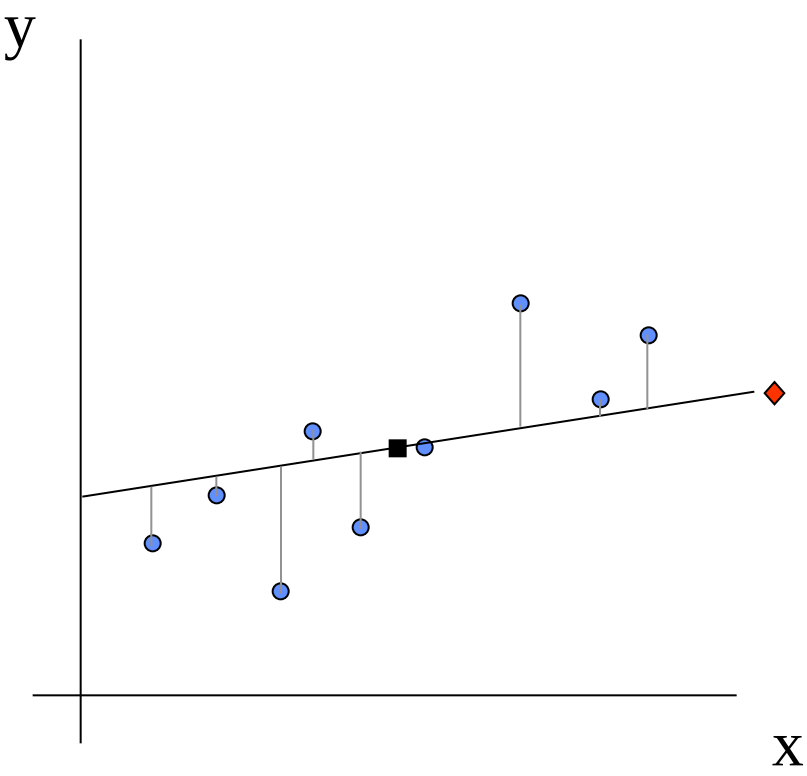
A regressziós egyenes (vagy legjobban illő egyenes, vagy a négyzetesen legjobb egyenes) a regressziós egyenlet gráfja.

Jelölések

	Populáció paraméter	Minta becslés
y-tengelymetszet	β_0	b_0
Meredekség	β_1	b_1
Egyenlet	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Speciális tulajdonság

A regressziós egyenes illik legjobban az adatokhoz.



A legkisebb négyzetek módszere

Keressük azt az egyenest, aminél a reziduumok négyzetének összege a lehető legkisebb:

$$\mathrm{F}(\mathrm{b}_0, b_1) = \sum (y - b_1 x - b_0)^2$$

Megkeressük azokat a paramétereket, amelyek mellett a fenti összeg a legkisebb:

$$\frac{\partial F(b_0, b_1)}{\partial b_0} = 0 \qquad \frac{\partial F(b_0, b_1)}{\partial b_1} = 0$$

folyt.

Bontsuk fel a négyzetet:

$$F(b_0, b_1) = \sum y^2 + b_1^2 \sum x^2 + nb_0^2 - 2b_1 \sum xy - 2b_0 \sum y + 2b_1 b_0 \sum x$$

Végezzük el az egyik deriválást:

$$0 = \frac{\partial F(b_0, b_1)}{\partial b_0} = 2(nb_0 - \sum y + b_1 \sum x)$$

Fejezzük ki az egyik paramétert:

$$b_0 = \frac{1}{n} \sum y - b_1 \frac{1}{n} \sum x = \bar{y} - b_1 \bar{x}$$

folyt.

Végezzük el a másik deriválást is:

$$0 = \frac{\partial F(b_0,b_1)}{\partial b_1} = 2(b_1 \sum x^2 - \sum xy + b_0 \sum x)$$

Oldjuk meg:

$$b_1 \sum x^2 = \sum xy - b_0 \sum x =$$
$$= \sum xy - (\sum x)(\sum y)/n + b_1 (\sum x)^2/n$$

$$\rightarrow \qquad b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

A b_0 és b_1 képletei

10-2. képlet $$b_1 = \frac{n(\Sigma xy) - (\Sigma x)\,(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$ **(meredekség)**

10-3. képlet $b_0 = \bar{y} - b_1 \bar{x}$ **(*y* tengelymetszet)**

A regressziós egyenes kiszámítása

Adatok:

x	3	1	3	5
y	5	8	6	4

A 10-2. fejezetben ezeket az adatokat használva kiszámítottuk a korrelációs együtthatót $r = -0.956$. Határozzuk meg a regressziós egyenest!

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b_1 = \frac{4(61) - (12)(23)}{4(44) - (12)^2}$$

$$b_1 = \frac{-32}{32} = -1$$

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

$$b_0 = \overline{y} - b_1 \overline{x}$$

$$5.75 - (-1)(3) = 8.75$$

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$

$\Sigma x = 12$

$\Sigma y = 23$

$\Sigma x^2 = 44$

$\Sigma y^2 = 141$

$\Sigma xy = 61$

A kiszámított regressziós egyenlet:

$$\hat{y} = 8.75 - 1x$$

Példa: Old Faithful

A 10-1. táblázat alapján, számítsuk ki a regressziós egyenest.

Ugyanazokat a lépéseket végigcsinálva, mint az előbb, kapjuk b_1 = 0.234 és b_0 = 34.8. Így a regressziós egyenlet:

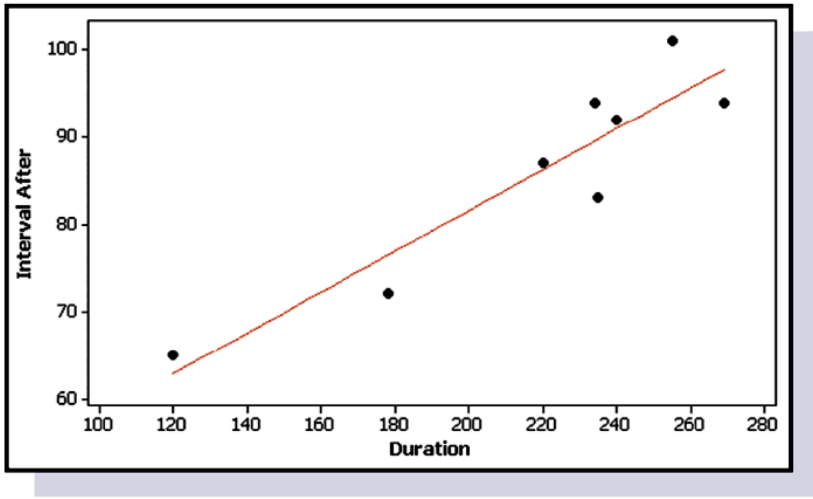
$$\hat{y} = 34.8 + 0.234x$$

Példa: Old Faithful - folyt

The regression equation is
Interval After = 34.8 + 0.234 Duration

Predictor	Coef	SE Coef	T	P
Constant	34.770	8.732	3.98	0.007
Duration	0.23406	0.03908	5.99	0.001

S = 4.97392 R-Sq = 85.7% R-Sq(adj) = 83.3%



Predikciók

Az y értékének becslése az x adott értékére alapozva ...

1. **Ha nem tudunk semmilyen kapcsolatról x és y között, akkor a legjobb predikció y értékére $\overline{y}$.**
2. **Ha van ismert lineáris kapcsolat, akkor a legjobb predikció, ha a regressziós egyenletbe behelyettesítjük x értékét és kiszámítjuk hozzá az y értékét.**

Példa: Old Faithful

A 10-1. táblázat alapján azt találtuk, hogy a regressziós egyenlet $\hat{y} = 34.8 + 0.234x$. Feltéve, hogy az utolsó kitörés hossza $x = 180$ másodperc volt, keressük meg a legjobb becslést y-ra, azaz a következő kitörésig eltelő időre.

$$\hat{y} = 34.8 + 0.234x$$

$$34.8 + 0.234(180) = 76.9 \text{ perc}$$

Az előrejelzett idő 76.9 perc.

Definíciók

- **Marginális változás**

  A marginális változás az a mennyiség, amennyit a változó változik, miközben a másikat egy egységnyivel megváltoztatjuk.

- **Outlier**

  Egy outlier egy olyan pont, ami a többitől messze esik.

- **Torzító pont**

  Egy torzító pont erősen befolyásolja a regressziós egyenes elhelyezkedését.

Definíciók

Reziduum

A reziduum egy (x, y) adatpár esetén , az $(y - \hat{y})$ különbség a megfigyelt y minta érték és a regressziós egyenes által adott y érték között.

reziduum = megfigyelt y – prediktált $y = y - \hat{y}$

Definíciók

- **Legkisebb négyzetek tulajdonság**

  Egy egyenes rendelkezik a legkisebb négyzetek tulajdonsággal ha a reziduumok négyzeteinek összege a lehető legkisebb.

- **Reziduális diagram**

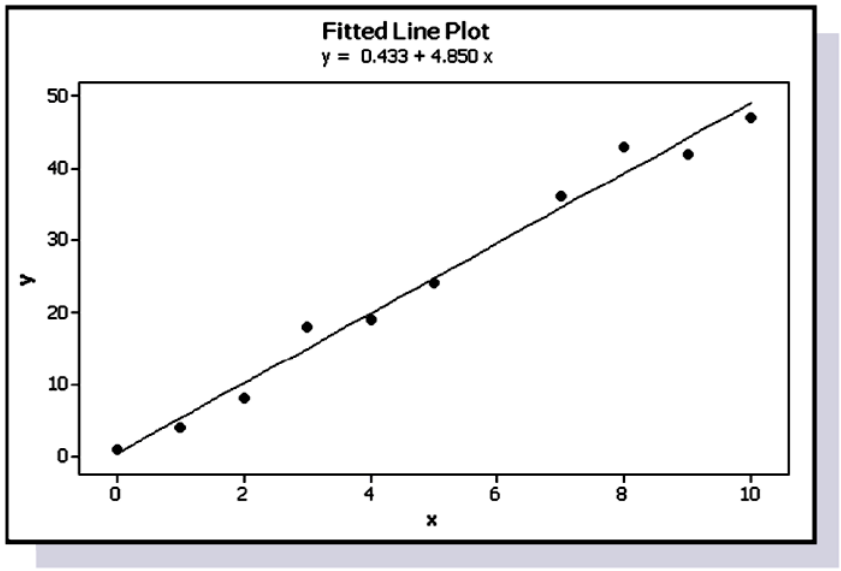
  Az (x, y) értékekből képzett szórásdiagramban az y-koordinátát az $y - \hat{y}$ reziduummal helyettesítjük. A reziduális diagram az $(x, y - \hat{y})$ pontpárokból áll.

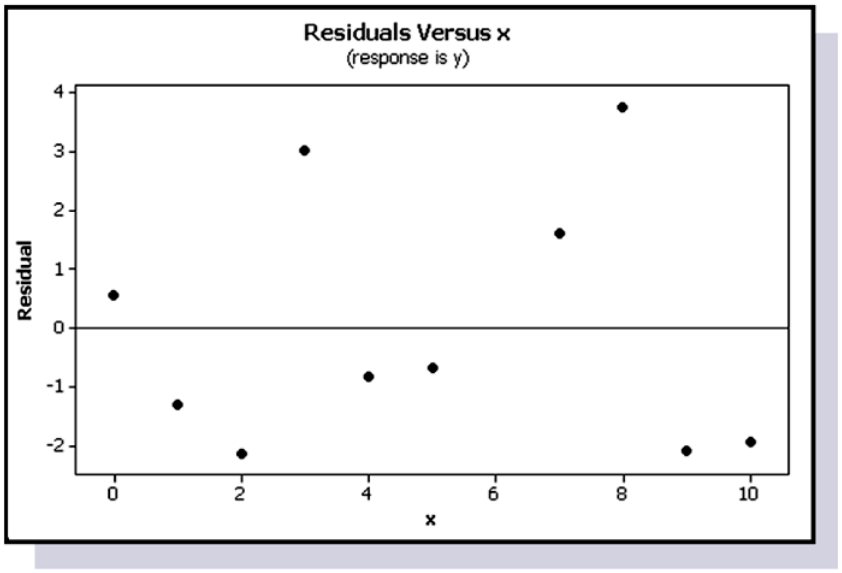
Reziduális diagram

Ha a reziduális diagram nem mutat semmilyen szabályosságot vagy alakzatot, akkor a regressziós egyenlet jól reprezentálja a két változó közti kapcsolatot.

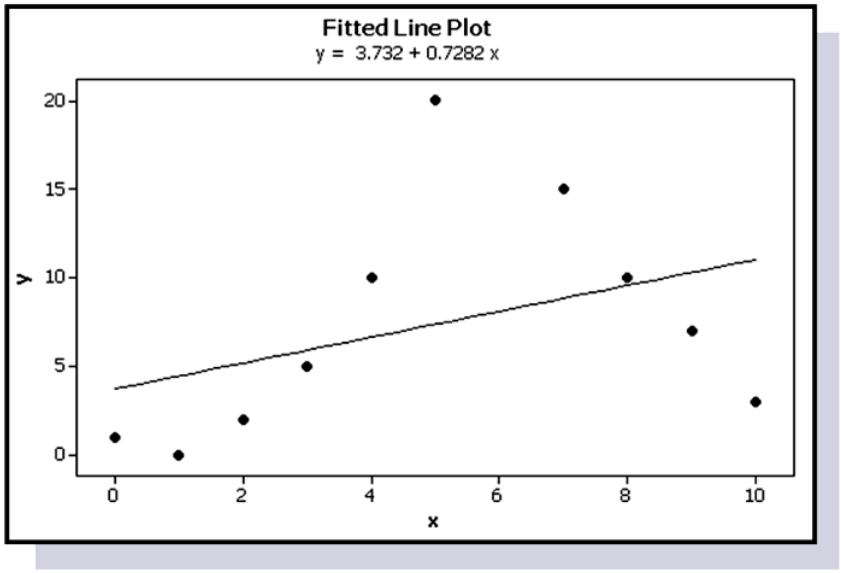
Ha a reziduális diagram valamilyen szabályos mintázatot mutat, akkor a regressziós egyenlet nem jó reprezentáció.

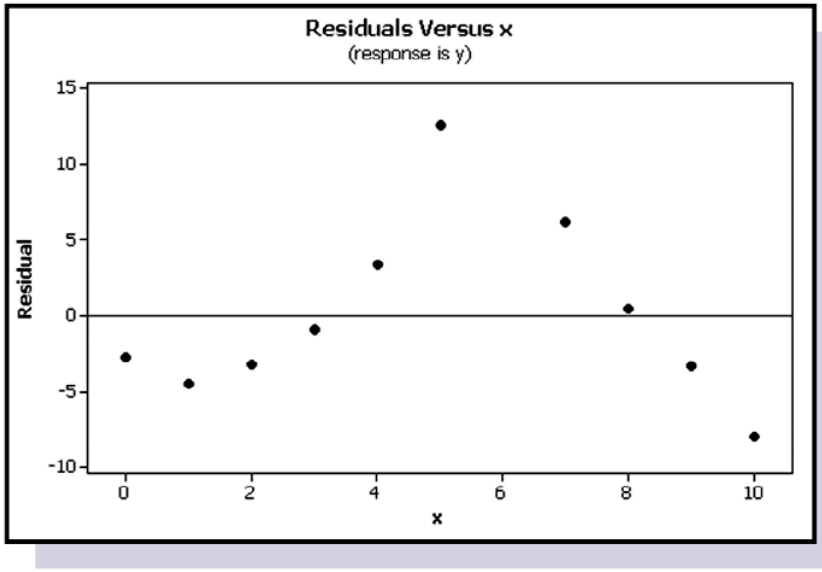
Reziduális diagram



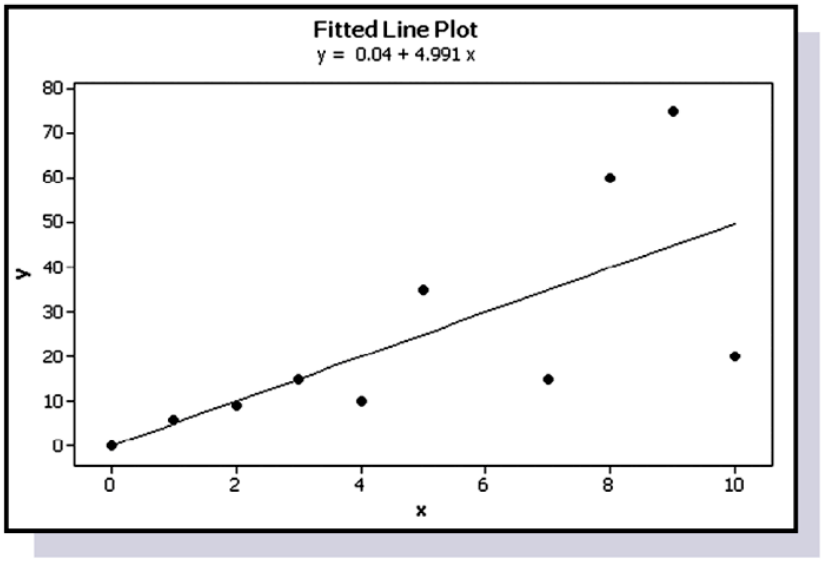


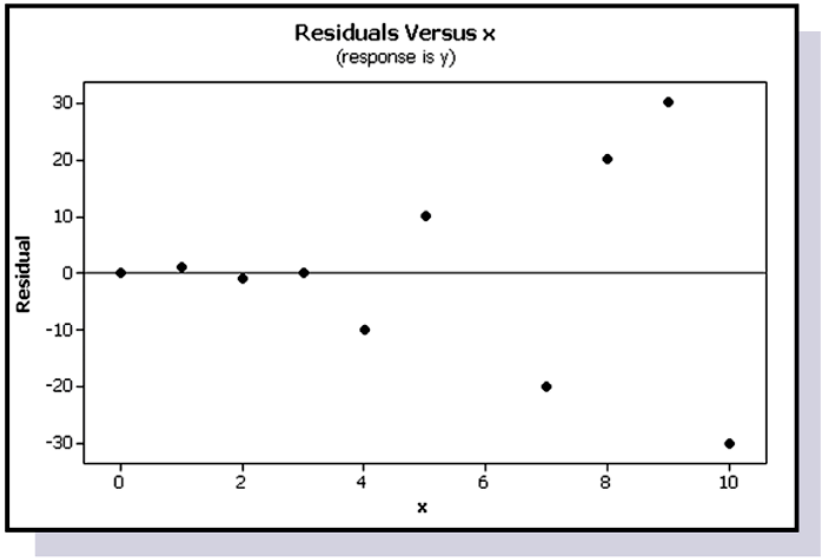
Reziduális diagram





Reziduális diagram





Összefoglalás

Ebben a fejezetben megvitattuk:

- **A regresszió alapjait.**
- **A regressziós egyenes előrejelzésre való használatát.**
- **A regressziós egyenlet interpretálását.**
- **Outlier-eket**
- **Reziduumokat és a legkisebb négyzeteket.**
- **Reziduális diagramokat.**

10-4. fejezet
Variabilitás és predikciós intervallum

Kulcsfogalmak

Ebben a fejezetben a predikciós intervallum megkonstruálásnak módszerét tekintjük át, ami az y értékének egy intervallum becslése.

Definíció

Teljes deviancia (eltérés)

A teljes deviancia az (x, y) pont párra vonatkozóan az a függőles $y - \bar{y}$ távolság ami az (x, y) pont és a minta átlagon $\bar{y}$ keresztül húzott vízszintes vonal között van.

Definíció

Magyarázott deviancia

A magyarázott deviancia az a függőleges távolság, ami a becsült $\hat{y}$-érték $\hat{y} - \bar{y}$ távolsága a minta átlagától.

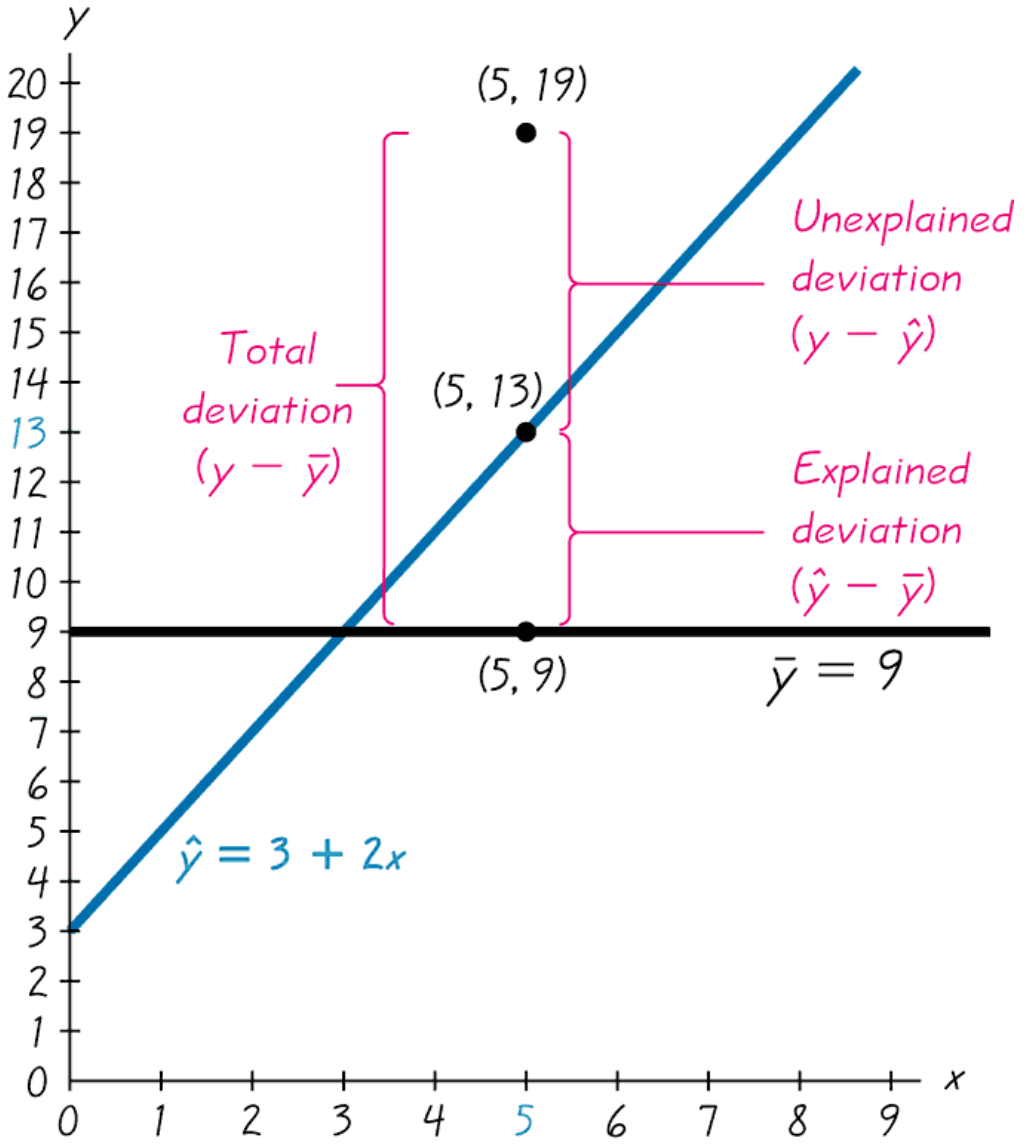
Definíció

Nem magyarázott deviancia

A nem magyarázott (reziduális) deviancia az $y - \hat{y}$ eltérés, ami a becsült és az igazi y érték különbsége. (Reziduumnak neveztük10-3.-ban.)

Nem magyarázott, magyarázott és teljes deviancia

10-9. ábra



Összefüggések

(teljes deviancia) = (magyarázott) + (nem magyarázott)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(teljes eltérésnégyzetösszeg) = (magyarázott) + (nem magyarázott)

$$\Sigma (y - \bar{y})^2 = \Sigma (\hat{y} - \bar{y})^2 + \Sigma (y - \hat{y})^2$$

10-4. képlet

Definíció

Determinációs együttható

az y variabilitásának az a része, amit a regressziós egyenes megmagyaráz.

$$r^2 = \frac{\text{magyarázott eltérésnégyzetösszeg.}}{\text{teljes eltérésnégyzetösszeg}}$$

Az r^2 értéke a variblitásnak az a hányada, amit az x és y közti lineáris kapcsolat megmagyaráz

Néhány mellékszámítás:

$$s_y^2 = \frac{\sum (y-\bar{y})^2}{n-1} \qquad b_1 = \frac{\overline{xy}-\bar{x}\bar{y}}{s_x^2}$$

$$\sum (\hat{y}-\bar{y})^2 = \sum (b_1 x + b_0 - \bar{y})^2 = b_1^2 \sum (x-\bar{x})^2 =$$

$$= b_1^2 s_x^2 (n-1) = (n-1)\frac{(\overline{xy}-\bar{x}\bar{y})^2}{s_x^2}$$

$$r^2 = \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} = \left[\frac{\overline{xy}-\bar{x}\bar{y}}{s_x s_y}\right]^2$$

Ez ugyanaz mint a lin. korr. együttható.

Definíció

A becslés hibájának szórása

A becslés hibájának szórása, $s_{e,}$ a mérőszáma a minta megfigyelt y értékei és a regressziós egyenes eltérésének.

A becslés hibájának szórása

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

vagy

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

Példa: Old Faithful

A 10-1 táblázat adatait használva határozzuk meg a becslés hibájának szórását.

$n = 8$

$\Sigma y^2 = 60{,}204$

$\Sigma y = 688$

$\Sigma xy = 154{,}378$

$b_0 = 34.7698041$

$b_1 = 0.2340614319$

$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

$$s_e = \sqrt{\frac{60{,}204 - (34.7698041)(688) - (0.2340614319)(154{,}378)}{8 - 2}}$$

$$= 4.973916052$$

A regressziós paraméterek konfidencia intervallumai

$$\hat{b}_1 - t_{\alpha/2} s_1 < \beta_1 < \hat{b}_1 + t_{\alpha/2} s_1$$

$$\hat{b}_0 - t_{\alpha/2} s_0 < \beta_0 < \hat{b}_0 + t_{\alpha/2} s_0$$

$$s_1^2 = \frac{s_e^2 \sum x^2}{n \sum x^2 - (\sum x)^2}$$

$$s_0^2 = \frac{n s_e^2}{n \sum x^2 - (\sum x)^2}$$

A becslési intervallum egyes y értékekre vonatkozóan

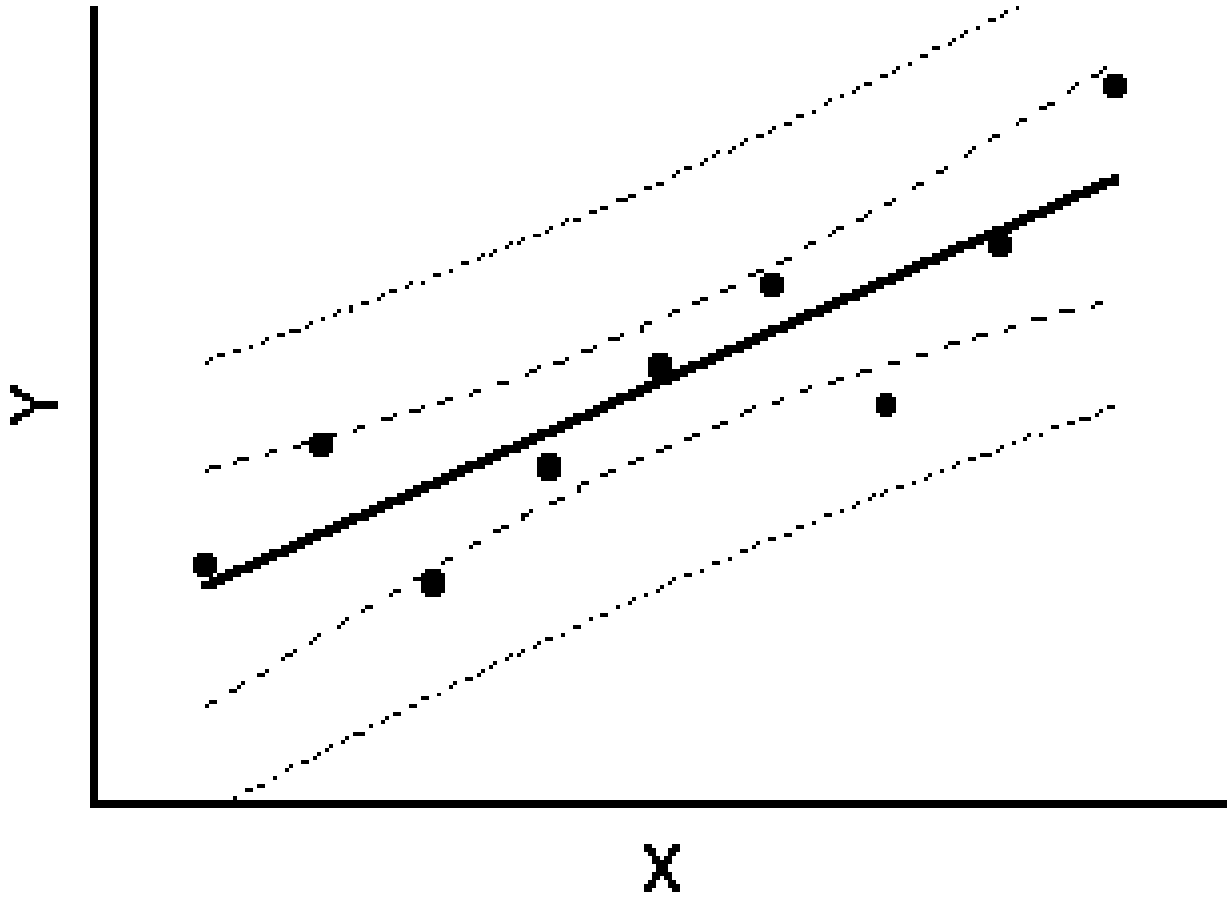
$$\hat{y} - E < y < \hat{y} + E$$

ahol

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

x_0 **az** x **megadott értéke** $t_{\alpha/2}$ **-nek** $n - 2$ **szabadsági foka van**

Predikciós és konfidencia intervallumok



Példa: Old Faithful

Az 10-1 táblázat adataihoz illesztett egyenes alapján azt találtuk, hogy a 180 sec. hosszúságú kitörés után a legközelebbi kitörés idejére adott becslés 76.9 perc. Adjuk meg a 95%-os becslés intervallumot ehhez az értékhez!

$$E = t_{\alpha/2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \overline{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

$$E = (2.447)(4.973916052)\sqrt{1 + \frac{1}{8} + \frac{8(180 - 218.875)^2}{8(399{,}451) - (1751)^2}}$$

$E = 13.4$ **(kerekítve)**

Példa: Old Faithful - folyt

$$\hat{y} - E < y < \hat{y} + E$$

$$76.9 - 13.4 < y < 76.9 + 13.4$$

$$63.5 < y < 90.3$$

Összefoglalás

Ebben a fejezetben foglalkoztunk:

- **Magyarázott és nem magyarázott devianciával.**
- **A determinációs együtthatóval.**
- **A hiba szórásával.**
- **A becslési intervallumokkal.**

10-5. fejezet
Többszörös regresszió

Kulcsfogalmak

Ebben a fejezetben a több mint két változó közti lineáris kapcsolatok elemzési módszerét vizsgáljuk meg.

Három kulcs elemre koncentrálunk:

1. A többszörös regressziós egyenletre.

2. Az adjusztált R^2 értékeire.

3. A P-értékre.

Definíció

Többszörös regressziós egyenlet

Lineáris kapcsolat a válasz változó y és a kettő vagy több prediktor változó között ($x_1, x_2, x_3 \ldots, x_k$)

Általános alakja:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k.$$

Jelölés

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k$$

(Az általános alakja a becsült regressziós egyenletnek)

n = minta méret

k = a prediktor változók száma

$\hat{y}$ = az y becsült értéke

$x_1, x_2, x_3 \ldots, x_k$ a prediktor változók

Jelölések- folyt

$ß_0$ = az y tengelymetszet, azaz az y értéke, amikor minden prediktor változó 0.

b_0 = becslése $ß_0$ –nak a minta alapján

$ß_1, ß_2, ß_3 \ldots, ß_k$ együtthatók a független változók előtt $x_1, x_2, x_3 \ldots, x_k$

$b_1, b_2, b_3 \ldots, b_k$ a mintabecslései az együtthatóknak $ß_1, ß_2, ß_3 \ldots, ß_k$

Példa: Old Faithful

A 10-1. táblázat alapján keressük meg a többszörös regressziós egyenletet, ahol a válasz változó (y) a kitörés után eltelő idő, és a prediktor változók (x) a kitörés hossza és magassága.

Az együtthatók megkeresését számítógépes csomagok (pl. Excel) végzik …

Példa: Old Faithful - folyt

The regression equation is
Interval After = 45.1 + 0.245 Duration - 0.098 Height

Predictor	Coef	SE Coef	T	P
Constant	45.10	19.41	2.32	0.068
Duration	0.24464	0.04486	5.45	0.003
Height	-0.0983	0.1623	-0.61	0.571

S = 5.25937 **R-Sq = 86.7%** **R-Sq(adj) = 81.3%**

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	2	897.69	448.85	16.23	**0.007**
Residual Error	5	138.31	27.66		
Total	7	1036.00			

Példa: Old Faithful - folyt

Eredmény:

Utána = 45.1 + 0.245 időtartam – 0.098 magasság

Vagy:

$$y = 45.1 + 0.245\, x_1 - 0.098 x_2$$

Definíció

- **Többszörös determinációs együttható**

  A többszörös determinációs együttható R^2 annak a mérőszáma, hogy mennyire illik a többszörös regressziós egyenlet a mintaadatokhoz.

- **Korrigált többszörös determinációs együttható**

  A korrigált többszörös determinációs együttható az előző R^2 olyan korrekciója, amely figyelembe veszi a változók számát és a minta méretét is.

Korrigált R^2

$$\text{Korrigált } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

10-6. képlet

ahol n = minta elemszáma

k = a független (x) változók száma

A legjobb többszörös regressziós egyenlet megkeresése

1. **Használd a józan eszedet arra, hogy kiválaszd a fontos és a nem fontos változókat.**
2. **Vedd figyelembe a *P*-értéket. Válassz olyan egyenletet, aminek nagy a szignifikanciája a számítógép által által adott *P*-értékek szerint.**
3. **Használd a nagy korrigált R^2 –tel rendelkező egyenleteket és csak kevés változót vegyél be.**
   - **Ha egy újabb prediktor változót veszel be és a korrigált R^2 nem növekszik lényegesen.**
   - **Adott számú prediktor (x) változó használata esetén használd a legnagyobb korrigált R^2-et adó változókat.**
   - **Hogy kidobáljuk a felesleges (x) változókat, amelyeknek nincs nagy hatásuk y-ra, segíthet a változók közti lineáris korrelációs együttható ismerete.**

Összefoglalás

Ebben a fejezetben megviattuk:

- **A többszörös regresszió egyenleteit.**
- **Korrigált R^2-et.**
- **A legjobb többszörös regressziós egyenlet megkeresését.**

10-6. fejezet
Modellezés

Kulcsfogalmak

Ebben a fejezetben bemutatjuk annak a részleteit, hogyan illeszthetünk matematikai modellt az adatainkhoz.

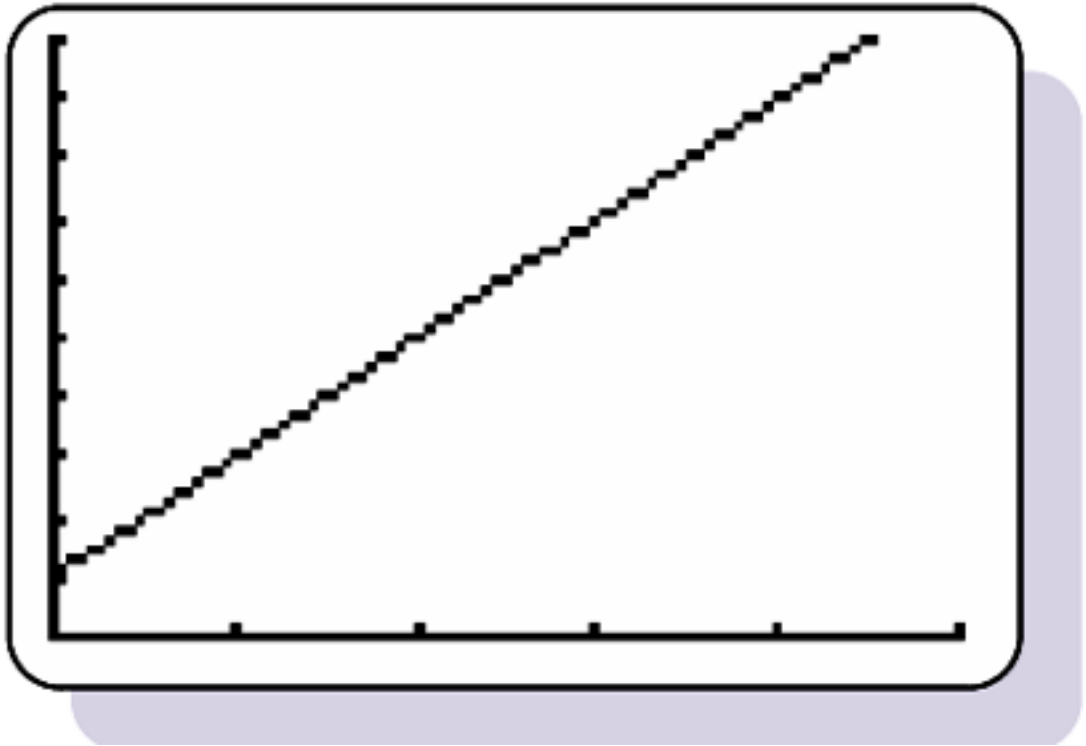
Ezt a folyamatot nemlineáris regressziónak is nevezik.

Példák

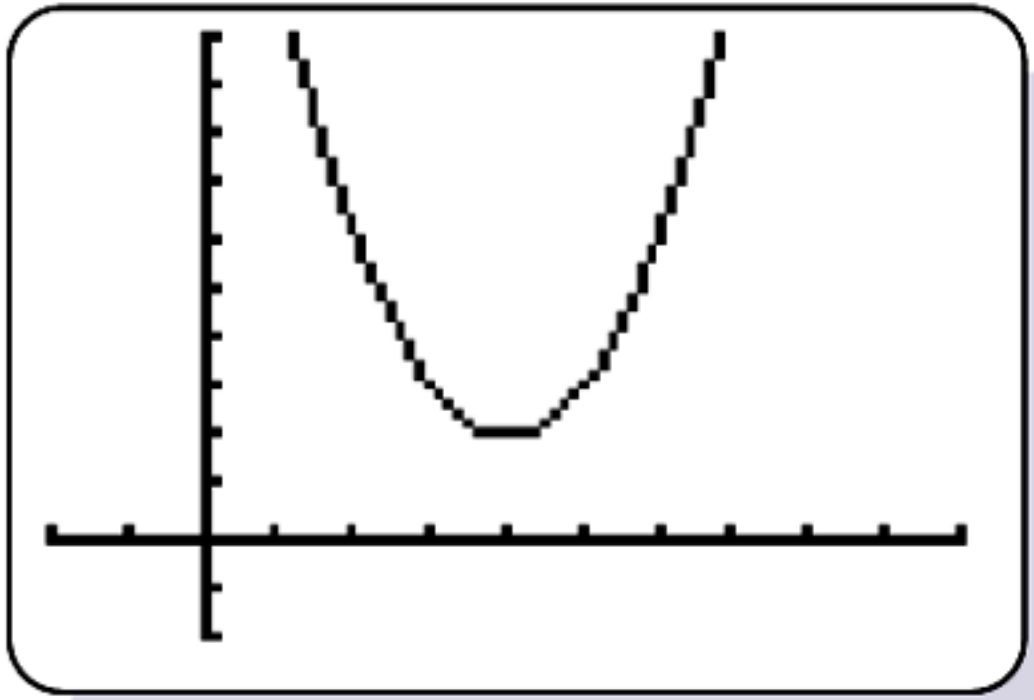
- **Lineáris:** $y = a + bx$
- **Kvadratikus:** $y = ax^2 + bx + c$
- **Logaritmikus:** $y = a + b \ln x$
- **Exponenciális:** $y = ab^x$
- **Hatvány:** $y = ax^b$

Illusztrációk:

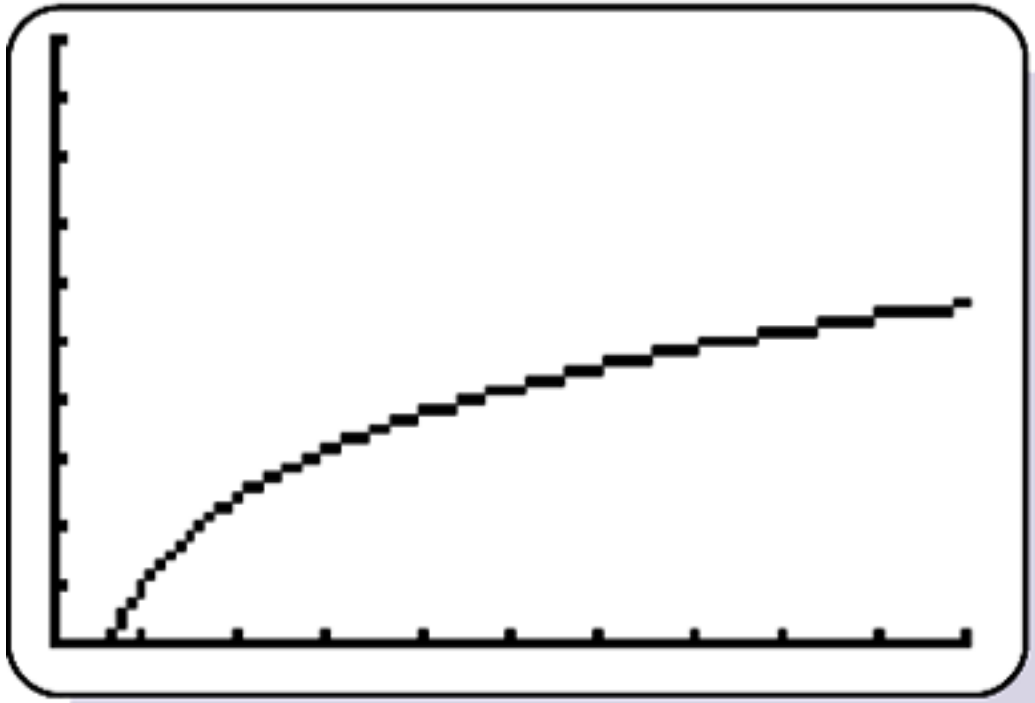
Linear: $y = 1 + 2x$



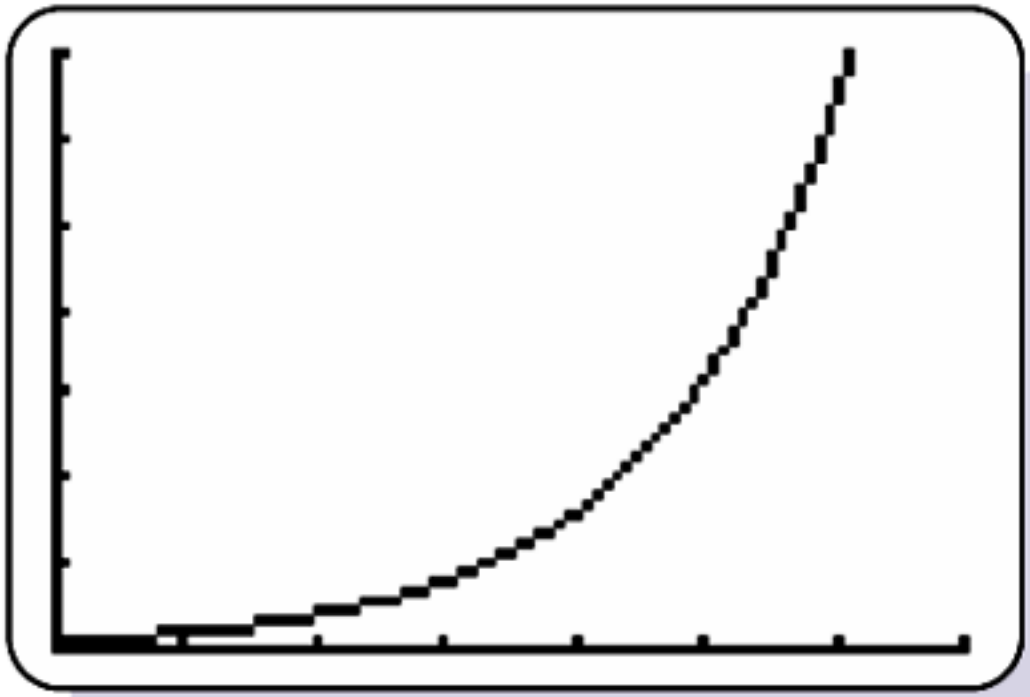
Quadratic: $y = 2x^2 - 8x + 9$



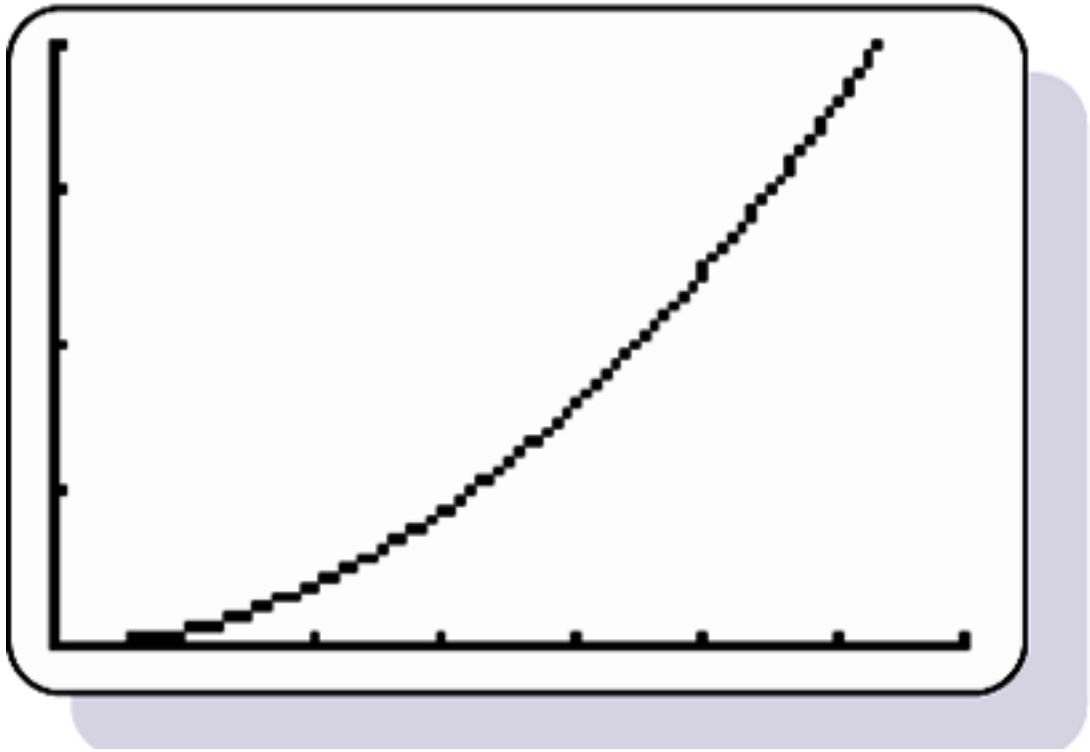
Logarithmic: $y = 1 + 2\ln x$



Exponential: $y = 2^x$



Power: $y = x^2$



Nemlineáris visszavezetése lineárisra

- **1, Polinom illesztés: visszavezethető lineárisra**

$$x_1 = x, x_2 = x^2, x_3 = x^3$$

- **2, Transzformációval visszavezethetők: exponenciális, hatvány**

$$y = a\exp(-bx) \rightarrow \log y = \log a - bx$$

$$y = ax^b \rightarrow \log y = \log a - b\log x$$

folyt

- **3, Nemlineáris függvény illesztése:**

$$min \sum (y - f(x, p_1, p_2, \ldots))^2$$

A jó modell (illesztő függvény) megkeresése

- **Keresd az adathalmazban a szabályosságot:** **Nézegesd az ábrát és próbáld meg kitalálni, milyen függvényt követnek az adatok.**
- **Számítsd ki R^2 -et és keress olyan függvényeket, amelyek minél nagyobb R^2 –et adnak, mivel ez azt jelenti, hogy azok jobban illenek az adatokhoz.**
- **Gondolkozz:** **Zárd ki a nem realisztikus modelleket, melyek hibás következtetésekre vezetnek.**

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A nemlineáris regressziót.**
- **Néhány jó tanácsot.**