



13. előadás

Matlab 7.

(Statisztika, regresszió,
mérési adatok feldolgozása)

Dr. Szörényi Miklós,
Dr. Kallós Gábor

2013–2014





Tartalom

- Statisztikai alapfogalmak
 - Populáció, hisztogram, átlag, medián, szórás, korreláció
 - Beépített támogatás a Matlabban
- Interpoláció és regresszió
- Adatsorok regressziós közelítése
 - Alapszintű illesztés (Basic fitting)
- Mérési adatok feldolgozása





Statisztikai alapfogalmak

- (Statisztikai) **populáció** ~ **alapsokaság** *population*
 - A vizsgálandó egyedek vagy objektumok adatainak az a (teljes) köre, amelyre a vizsgálat irányul, azaz amelyre következtetéseinket vonatkoztatni szeretnénk
- **Minta** *sample*
 - A vizsgálandó egyedek vagy objektumok adatainak az a köre, amelyeket ténylegesen megvizsgálunk, azaz amelyeken következtetéseink alapulnak
- **Megfigyelési egység** *observational* vagy *experimental unit*
 - A populáció, illetve a minta egy eleme, egy egyed vagy objektum adata, amelyet feljegyezzük (lehet egy ember vagy állat, egy vérminta, egyedek egy csoportja, pl. egy család, stb. adata)
- **Változó** *variable*
 - Adat, jellemző, ismérv, tulajdonság, amelyet a mintabeli egyedeken megfigyelünk, megmérünk, feljegyezzük (életkor, testtömeg, kapott kezelés típusa, időtartama, stb.)
 - A mintán megfigyelt adatokat az *adatmátrix* tartalmazza; szokásos elrendezésében minden sor egy mintavételi egységnek és minden oszlop egy változónak felel meg



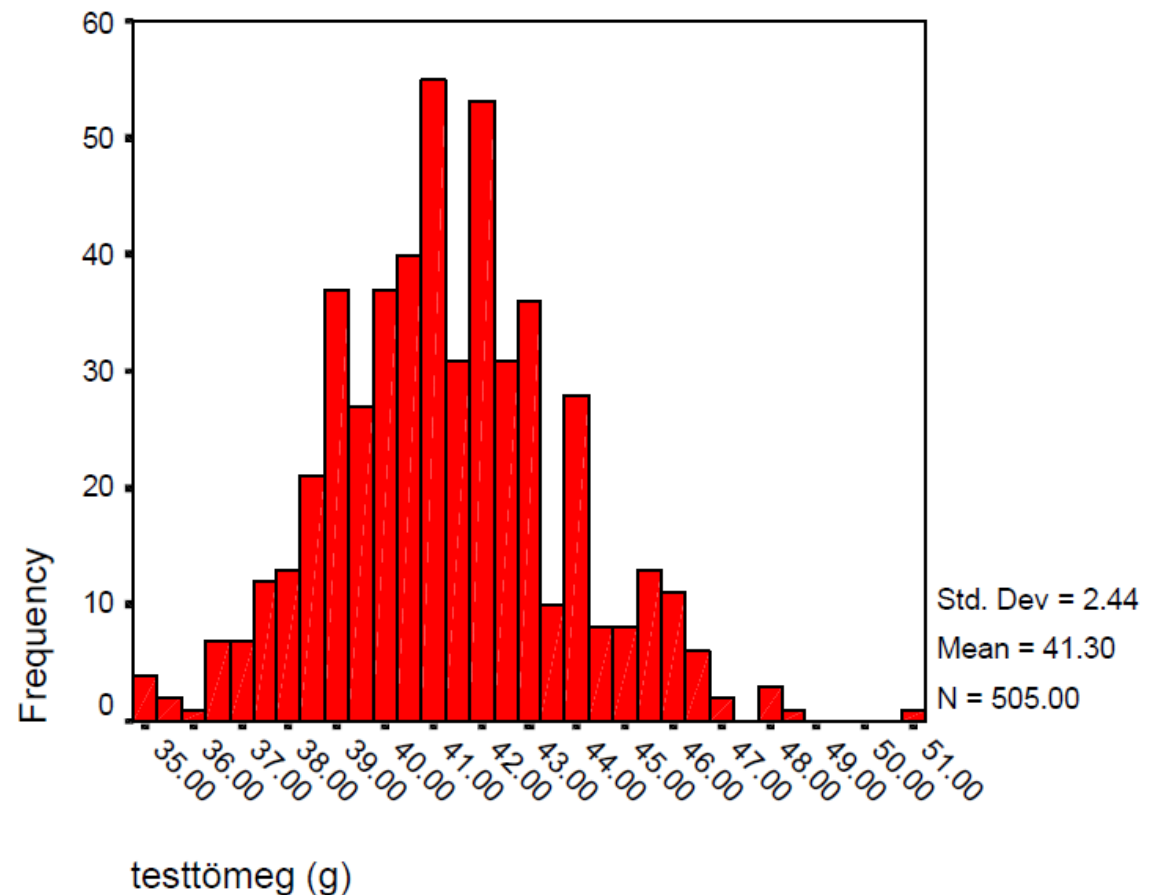


Statisztikai alapfogalmak

■ **Hisztogram** *histogram*: tapasztalati sűrűségfüggvény

- Vízszintes tengelyén: osztályintervallumok, fölötté olyan téglalapok, amelyek *területe* megegyezik a megfelelő relatív, vagy százalékos gyakorisággal
- Így a hisztogram teljes területe 1, vagy 100% lesz
- Diszkrét változó esetén a változó értékei az intervallumok közepén helyezkednek el
- A hisztogram – ha a minta elemszámát növeljük – közelíti a valószínűségi változó elméleti sűrűségfüggvényét
- (Hisztogram helyett gyakorisági poligon is rajzolható)

■ (Kumulatív hisztogram ~ tapasztalati eloszlásfüggvény)



Statisztikai alapfogalmak

Alapstatisztikák

- Az eloszlás közepére vonatkozóak: az **átlag**, a **medián** és a **módusz**

- **Átlag** *average, mean*

- Legyenek a minta elemei x_1, x_2, \dots, x_n

- Ekkor:

- Az átlag az az érték, amely a „legközelebb” van a minta elemeihez

- **A mintabeli értékek és a mintaátlag közti eltérések összege mindig 0**

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

- **Módusz** *mode*

- A leggyakrabban előforduló érték, jelölés: M_0

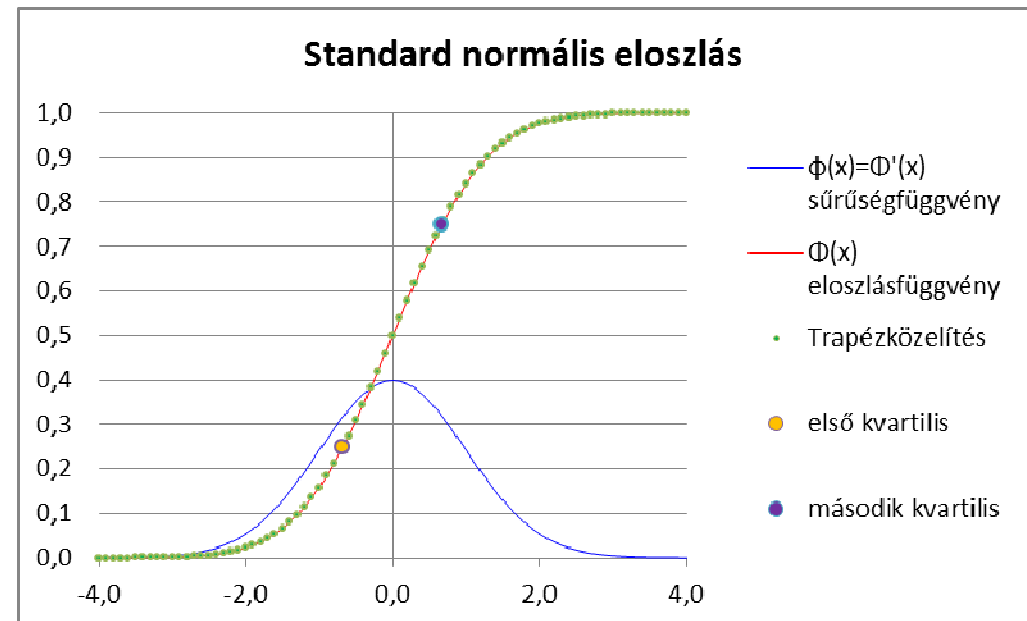
- **Medián** *median*

- Sorbarendezett adatok középső eleme (50%-os vágóérték), jelölés: M_e

- **Percentilis**: adott százalékos vágóérték

- **Kvartilis** (alsó, felső): 25, ill. 75%-os vágóérték

- Jelölés: Q_1 és Q_3 , Q_2 a medián



Statisztikai alapfogalmak

Alapstatisztikák (folyt.)

- **Tapasztalati szórás és szórásnégyzet** vagy más néven **variancia** *variance*

- A szórás a variancia négyzetgyöke (a képletben s – vagy $D(X)$ – a szórás, ennek négyzete pedig a variancia, s^2)

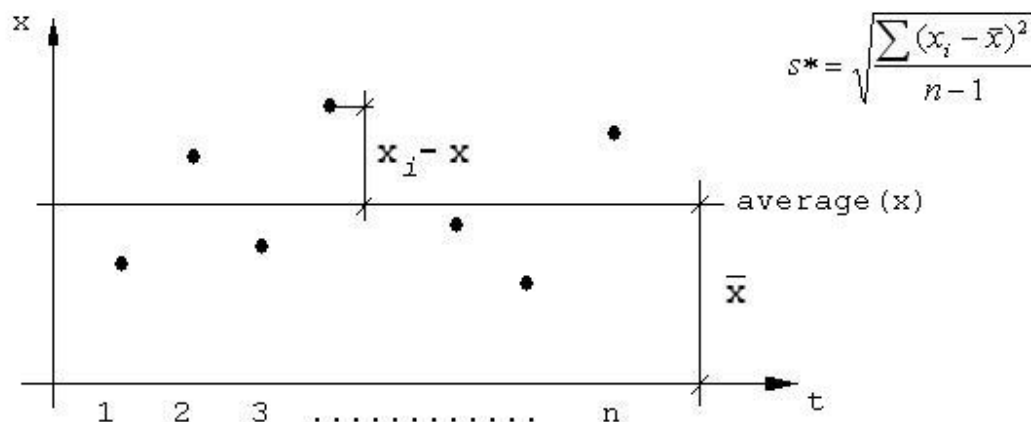
- A szórás azt mutatja meg, hogy az adataink átlagosan milyen távol helyezkednek el a számtani középtől

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad \sqrt{E((X - E(X))^2)} = \sqrt{E(X^2) - E^2(X)}$$

- Egyes esetekben – csak normális eloszlásúnak tekinthető val. változó esetén – az ún. **korrigált tapasztalati szórást** (*Standard Deviation*: SD) használjuk

- *Miért $n - 1$ -gyel osztunk: eggyel csökken a szabadsági fok (normális eloszlás)
 - A programok általában használják a korrigált szórást is (nagy n esetén alig van eltérés, csak kicsi mintaelemszám esetén van szerepe)

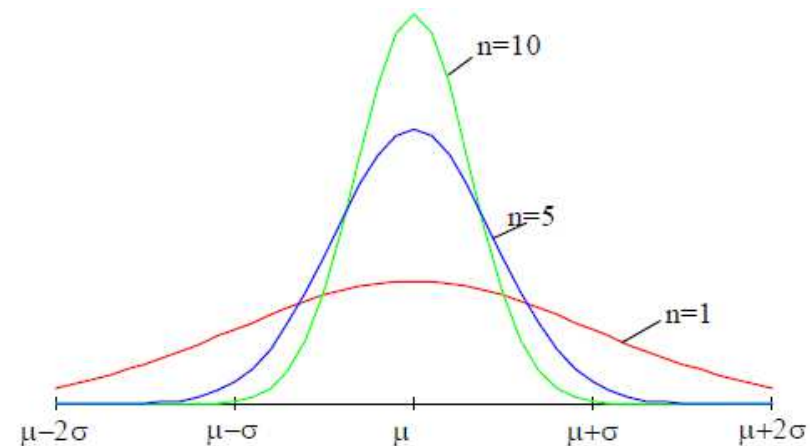
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$



Statisztikai alapfogalmak

Alapstatisztikák (folyt.)

- Ha a mintából készített hisztogram elég jól közelíti a normális görbét, akkor a normális eloszlás táblázatából kiolvasható, hogy
 - az $(\bar{x} - s, \bar{x} + s)$ intervallumban van adatunk kb. 68%-a (**kb 2/3-a**),
 - az $(\bar{x} - 2s, \bar{x} + 2s)$ intervallumban van **kb. 95%-a**,
 - az $(\bar{x} - 3s, \bar{x} + 3s)$ intervallumba pedig kb. 99,7%-uk esik (**majdnem mind**)
- **Standard hiba** (*standard error*, **SE**) teljes neve „a mintaátlag standard hibája”, azaz szórása (itt n a mintaelemszám):
$$SE(\bar{x}) = SD(X) / \sqrt{n}$$
 - Szemléletes jelentés: 100-szor több adatból 10-szer pontosabb statisztikai eredményt kapunk
- *Matematikailag bizonyítható (Centrális határeloszlás tétel), hogy függetlenül a mintaelemek eloszlásától, a minta átlagának eloszlása mindig a normális eloszláshoz tart, és az átlag várható értéke a populáció várható értékével egyezik meg
- Pl. kérdőíves felmérésnél megbecsüljük, hogy hány adat alapján lehet kellően megbízható kijelentést tenni (a korlátot a populáció mérete is befolyásolja)
 - De: egy bizonyos elemszám felett a becslés megbízhatósága már csak kevésbé javul (példa: pártszimp. felmérés Mo-on)



A mintaátlag szórásának elemszám függése



Statisztikai alapfogalmak

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\mathbb{D}(X) \mathbb{D}(Y)}$$

■ Korreláció *correlation*

- Két változó közötti kapcsolat erősségének mérőszáma („együtfutás”)
 - Pl. igaz-e, hogy ha kétszer akkora az autó tömege, akkor a fogyasztása is jóval nagyobb?
- Teljesül: $-1 \leq r \leq 1$
- 1 közeli értékek: erős kapcsolat; -1 közeli értékek: erős, de szembefutó kapcsolat; 0 közeli értékek: gyenge kapcsolat, függetlenség feltételezhető
- Ábrázolás: a pontokat összekötni nem szabad, de trendvonal húzható

■ Kovariancia *covariance*

- Szintén változók közötti függőségek mérésére; a korreláció a kovariancia skálázott változata (osztjuk a szórásokkal)

$$\text{cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}(X)) (Y - \mathbb{E}(Y))]$$

■ R-négyzet

- A korrelációs együttható négyzete, mindig nemnegatív

	A	B	C	D	E	F	G	H	I	J	K
1		Kobcm	Telj	Nyom	Gyors	Vmax	Tomeg	Hossz	Szeles	Csomag	Fogy
2	Kobcm	1									
3	Telj	0.750436	1								
4	Nyom	0.8569	0.712138	1							
5	Gyors	0.323246	0.79807	0.357642	1						
6	Vmax	0.427314	0.836641	0.454459	0.924247	1					
7	Tomeg	0.810403	0.498836	0.767362	-0.0613	0.074022	1				
8	Hossz	0.77785	0.568661	0.681689	0.11042	0.284192	0.813907	1			
9	Szeles	0.69409	0.524396	0.6329	0.108673	0.223071	0.788961	0.685854	1		
10	Csomag	0.382567	0.307708	0.363363	0.076291	0.181909	0.483343	0.482642	0.583473	1	
11	Fogy	0.587514	0.628174	0.384842	0.353058	0.30603	0.499987	0.581196	0.49378	0.265225	1
12	Ar	0.848481	0.893494	0.850877	0.558585	0.660903	0.691331	0.658447	0.674627	0.414921	0.587681
13											



Alapstatisztikák a Matlabbal

Beépített támogatás a Matlabban (fontosabb parancsok, alap Matlab)

- Egyváltozós jellemzők

Függvény	Jelentés
max	Maximális elem
min	Minimális elem
mean	Átlag (várható érték torzítatlan becslése)
median	Rendezett minta közepe
std	Tapasztalati szórás
sum	Összeg
cumsum	Kumulatív részösszegek
sort	Növekvő sorrendbe rendezés
diff	Szomszédos elemek differenciái
hist	Hisztogram (gyakoriság oszlopdiagram)

- Többváltozós jellemzők

corrcoef	Korrelációs mátrix
cov	Kovariancia mátrix





Alapstatisztikák a Matlabbal

Mintafeladat

- Az ADAT.dat fájl (oszloponként, fejléc nélkül) magyarországi városok egyes statisztikai adatait tartalmazza a következők szerint:
 - terület – a település területe
 - szja – befizetett SZJA (eFt)
 - mun_reg – regisztrált munkanélküliek aránya a 18–59 lakosságon belül (ezrelék)
 - mun_tart – tartósan munkanélküliek aránya, mint fent
 - tele_uzl – üzleti vonalak aránya az összes vonalon belül (ezrelék)
 - kocsi – gépkocsik aránya (ezrelék)
 - telefon – vezetékes telefonok aránya (ezrelék)
 - lakos – állandó lakosság
- Kérjük ki változónként (oszloponként) az adatminta következő statisztikai jellemzőit: max, min, sum, mean, std, median

```
>> szja = ADAT(:, 2);  
>> min(szja), mean(szja), median(szja)
```
- Kérjük ki az adatoszlopok hisztogramját is!

```
>> hist(ADAT(:, 6))
```
- Kérjük ki az adatoszlopok kapcsolati erősségét bemutató – szimmetrikus – korrelációs mátrixot (elemei: korrelációs együtthatók)

```
>> corrcoef(ADAT)
```
- Mely két adatoszlop között van a legerősebb kapcsolat?

```
>> abs(corrcoef(ADAT)) - eye(8)
```





Interpoláció és regresszió

- Feladat: ismerjük egy jelenség, folyamat matematikai modelljét, de annak *aktuális paraméterei ismeretlenek*, és meg kell határozni ezeket
 - (Paraméterbecslési feladat)
- Vagy: *nem ismerjük a matematikai modellt*, ekkor ismert alapfüggvények, pl. algebrai vagy trigonometrikus polinomok vagy exponenciális fv-ek kombinációit tekintjük a modell egy közelítésének
 - Ekkor a cél ezek együtthatóinak a meghatározása
- A megoldáshoz mindkét esetben a jelenséghez tartozó (összetartozó) bemeneti-kimeneti értékpárokat kell ismernünk (ezek adottak)
- Ha ezek az értékpárok *hibátlanoknak tekinthetők és számuk egyenlő a keresett paraméterek (együtthatók) számával: interpoláció*
- Ha ezek az értékpárok *hibával terheltek és számuk meghaladja a meghatározandó paraméterek (együtthatók) számát: regresszió*
 - Technikai mo.: meghatározzuk a modellfüggvényt (közelítő függvényt)
- A modellfüggvény ezután alkalmas lesz arra, hogy a bemenet/kimenet kapcsolatát olyan pontokban is megadjuk, amelyekben korábban nem ismertük
- Az interpoláció tipikus technikái
 - Polinomiális interpoláció, spline interpoláció, két- vagy többváltozós interpoláció
- A regresszió tipikus technikái
 - Lineáris regresszió, nemlineáris regresszió





Adatsorok regressziós közelítése

- A mi célunk: adatsorok közelítése adott függvénytypussal
 - (*Vagy: a modellfüggvényt is mi készítjük el)
- Ismert a műszaki-fizikai jelenség modellje és a konkrét megvalósítás kimért adatsora
 - Ez általában ténylegesen regressziós feladat
- Feladattípusok:
 - A modell paramétereinek minél jobb becslése, mert ennek műszaki-fizikai jelentése van (hővezetési tényező, rugalmassági modulus, stb.)
 - A mért pontok közötti intervallumokban is szeretnénk minél jobb becslést adni a lehetséges értékekre
 - A teljes mért intervallumon kívül is szeretnénk előrejelzést adni a további függvényértékekre (trend)
- A közelítés jóságának mérőszámai:
 - A független változó mért értékeinél a függő változó mért és becsült értékeinek korrelációs együtthatója
 - Ugyanennek a négyzete az R^2 (determinációs együttható)
 - Az eltérések szórása
- A Matlab az ilyen típusú feladatok megoldásához változatos megoldási lehetőségeket nyújt
 - Alap lehetőségek, ill. Toolboxok parancsai (pl. nlinfit, cftool)

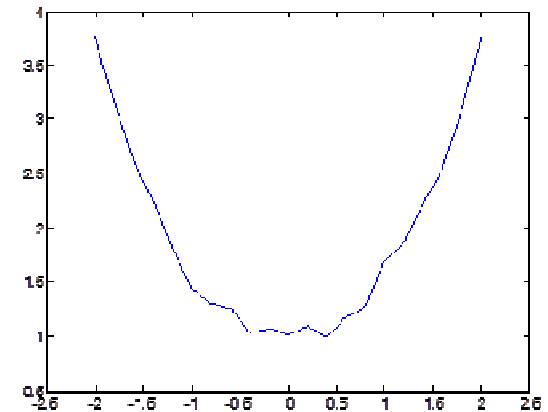


Alapszintű illesztés (Basic fitting)

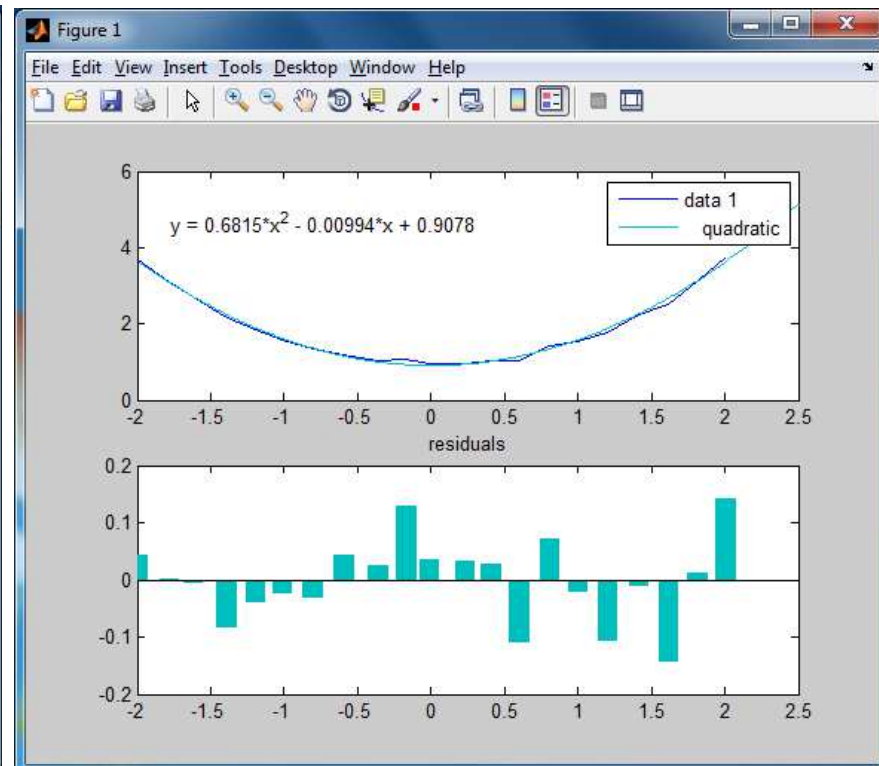
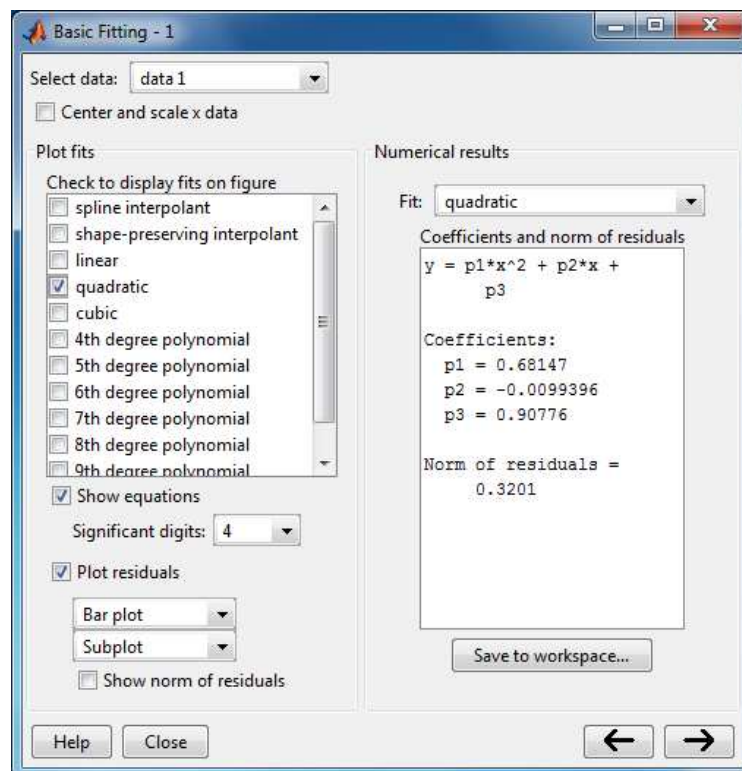
Példa: pontsorozat polinomiális közelítése

- A $\cosh(x)$ láncgörbe mentén egy véletlen pontsorozatot generálunk és kirajzoljuk:

```
>> x = (-2:0.2:2) + 0.01*randn(1,21);  
>> y = cosh(x) + 0.05*randn(1,21);  
>> plot(x,y)
```



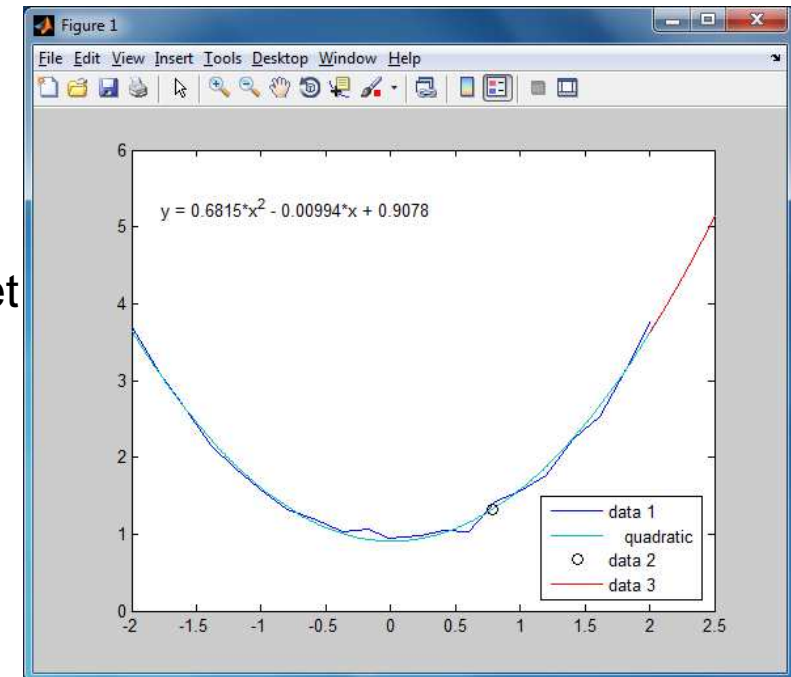
- Láthatóan parabolászerű, illesszünk hozzá másodfokú görbét
- Az ábra Tools/Basic Fitting menüjében beállítjuk a közelítés szolgáltatásait



Alapszintű illesztés (Basic fitting)

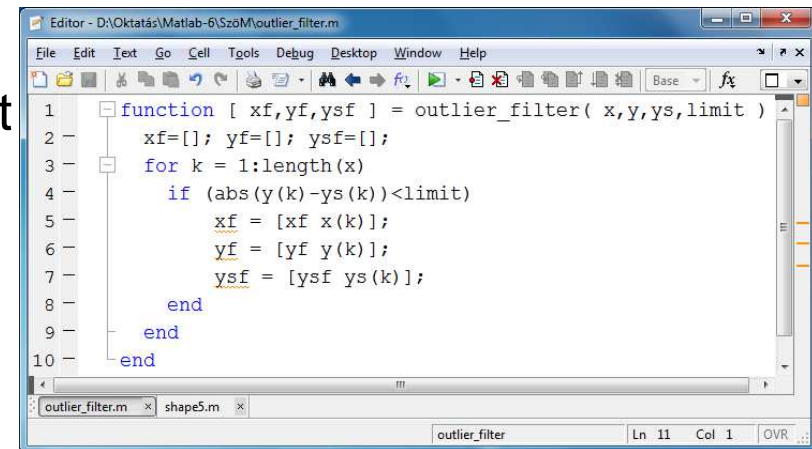
Példa (folyt.)

- Az együttthatók értékadásait bemásoljuk a parancsablakba és lefuttatjuk:
`>> p1 = 0.68147, p2 = -0.0099396,
p3 = 0.90776`
- Kiszámítjuk az alappontbeli közelítő értékeket és az R-négyzetet:
`>> y_pred = p1*x.^2+p2*x+p3;
>> R_square = ...
min(min(corrcoef(y,y_pred))).^2
R_square = 0.9939`
- Az eltérések négyzetösszege (Norm of residuals):
`>> sqrt(sum((y-y_pred).^2))
ans = 0.3201`
- Intervallumon belüli becslés és rajz:
`>> xp = pi/4, yp = p1*xp.^2 + p2*xp + p3
>> hold on, plot(xp, yp, 'ok')`
- Előrejelzés:
`>> xz = 2:0.1:2.5; yz = p1*xz.^2 + p2*xz + p3;
>> plot(xz,yz, 'r')`



Mérési adatok feldolgozása

- Mintafeladat: részletesen lásd gyakorlat
- A feladat részei
 - A megadott szövegfájl „megtisztítása”
 - Adatoszlopok betöltése, kirajzoltatás
 - Elemzés: a függvény részekre bontása a regressziós feladathoz
 - *Deriváltbecslés
 - Alapstatisztikai vizsgálat
 - Lineáris regresszió (polinomiális regresszió)
 - Simítás, szűrés
 - Saját függvénnyel (csúszóátlag)
 - *Paraméterbecslés a legkisebb négyzetek módszerével
 - Saját függvénnyel



```
function [ xf,yf,ysf ] = outlier_filter( x,y,ys,limit )
1  xf=[]; yf=[]; ysf=[];
2  for k = 1:length(x)
3      if (abs(y(k)-ys(k))<limit)
4          xf = [xf x(k)];
5          yf = [yf y(k)];
6          ysf = [ysf ys(k)];
7      end
8  end
9  end
```

