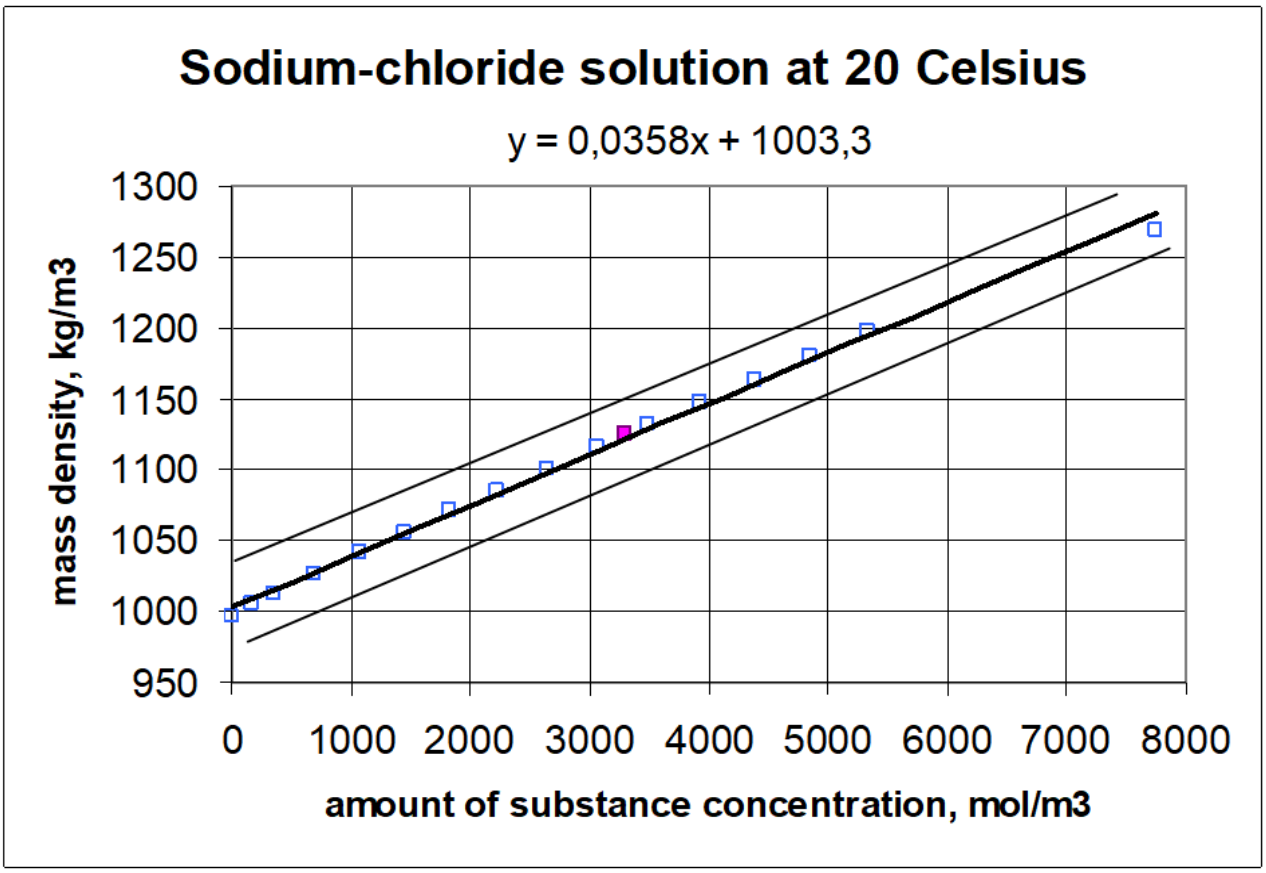
Regresszió számítása

A regresszió más néven a görbe illesztési feladatok közé tartozik. Ilyen módszert használunk, ha a független változó (egy, vagy több) feltételezésünk szerint kapcsolatban áll a függő változó értékével. A matematikai statisztika általában a folytonos függvények halmazán értelmezi az efféle műveleteket. A fizika tantárgy oktatásánál ennél fontosabbnak tartjuk annak feltételezését, hogy a független és a függő változó között fizikai törvénnyel megmagyarázható ok-okozati kapcsolat van.

Megfordítva a kérdést: ismeretlen összefüggést vizsgálva regressziós számítás segítségével kíséreljük meg megkeresni, hogy vajon létezik-e, és milyen kapcsolat a mérési eredményeink között. Ez a módszer tévutakra is vezethet, ugyanis esetleg összefüggést mutat ki olyan változók között, ami szakmai szempontból helytelen. Ilyen lehet, ha diszkrét értékekre kísérelünk meg folytonos függvényt illeszteni. Ilyen lehet, ha egy összefüggést fizikai törvény alapján le lehet írni, de nem túlságosan jól, és létezik ugyan a vizsgált jelenségre jobb illeszkedés is, de az a fizikai törvényeknek ellentmondó összefüggés.

Az elsőéves fizika oktatásánál arra helyezzük a súlyt, hogy a hallgatók tanulják meg a méréseket elvégezni, és eredményeiket szakszerűen kiértékelni. Ehhez engedményeket kell tennünk, elsősorban a túlságosan kevés mérési adat miatt, másrészt a hallgatóink tapasztalansága miatt. Például ilyenek a következő esetek.



Az ábra közepén a piros pont az eloszlás súlypontját jelzi. Baloldalt a híg oldatok tartományában kisebb a sűrűség (kék kocka) a regresszió (fekete vonal) által jelzett értéknél, a súlypontnál nagyobb; utána elég jól illeszkedik, a telített oldatnál is (5408 mol/m^3) de ismét kisebb a tiszta kristályos nátrium-kloridnál (7746 mol/m^3). A konfidencia sávot a szükségesnél szélesebbre rajzoltuk. A tengelymetszetnek a tiszta víz sűrűségét kellene adnia (997 kg/m^3) az 1003,3 helyett.

Az oldatok összetételi aránya és sűrűsége között lineáris kapcsolatot feltételezünk. Ez azzal jár, hogy elhanyagoljuk a komponensek közti kölcsönhatást, amely változó értékű, ha eltérő összetételi arányokat vizsgálunk. Ilyen jelenség a moláris térfogatcsökkenés, V_M. Különösképp, ha ionos oldatot használunk (nátrium-klorid).

A répacukor törésmutatója és az összetételi arány mérésénél elég szokványosak a rendszeres hibák. Ilyen például az, hogy nem távolították el az előző oldatot a műszerről, s emiatt a következő oldat összetétele már megváltozik. A pipetta pontatlan használata miatt az előzőleg mért oldat egy kis része belekerül a következő oldatba, emiatt a következő hallgatói csoport már nem az előzőleg pontosan kimért oldatból végez mérést. A számítás eredménye tartalmaz ugyan hibát, de az összefüggés szemléltetését és számítását gyakorolni lehet ezzel az egyszerűsített módszerrel. (Hasonló hibák jelentkeznek a felületi feszültség és a viszkozitás mérésénél, de ezekhez nem kérjük a regressziós számítást.)

A híg oldatok fagyáscsökkenését hallgatóink nem tudják mérni. Ehhez ugyanis csaknem végtelen hígítású oldatokat kellene mérnünk. Tapasztalataink szerint ehhez az elsőévesek felkészültsége – és a rendelkezésre álló eszközök – nem elégségesek. Ezért a méréseket olyan oldatokkal végezzük, amelyek mesze nem tekinthetőek híg oldatnak, az eredmények azonban lehetővé teszik a kiértékeléssel kapcsolatos számítások gyakorlását. Példaképpen: a répacukor fagyási görbéje az eutektikus pontig legfeljebb mínusz kilenc fokot adhat eredményül. Ezeket az adatokat a hallgatók sokszor a víz fagyáspontja feletti értéknek mérték, plusz négy-öt-hat fokot, s ez képtelenség. A konyhasóoldat fagyási görbéjét mérve ugyanekkora hibát kapnak, azonban a só eutektikus pontja sokkal alacsonyabb: mínusz 21 fok. Tapasztalatok szerint – csak példaképpen – a mínusz tizenöt fokos fagyáspontú oldatokat mínusz nyolc-tíz fokosnak szokták mérni. Ez is tartalmaz ugyan mérési hibát, de a fizikai törvényekkel szöges ellentétben nem áll. Hiszen fagyáspontcsökkenésre számítottunk, és az is történt.

Hasonlóképpen a víz forrási nyomása és hőmérséklete közötti kapcsolat mérésekor az atmoszférikus nyomásra vonatkoztatva száz fok helyett a hallgatók általában 98,5 fokot szoktak mérni. Ismerjük ennek az okát: a pontos méréshez ki kellene zárnunk a levegő jelenlétét, ami bonyolultabb felszerelést és hosszabb időt feltételez. Ez nem áll rendelkezésre. Ezért a nyomásmérőről az össznyomást olvassuk le, és úgy számolunk vele, mintha ez volna a vízgőz parciális nyomása. A számítás tartalmaz ugyan hibát, de az effajta számítások elvégzéséhez modellként használható.

Az elektromos és termikus jelenségek összefüggésének mérésénél (termisztor és termoelem) esetén leginkább a felhasznált műszerek méréshatára, vagy nem kielégítő felbontóképessége (skálaosztás) okoz mérési hibát. A mérési eredményeken ennek ellenére gyakorolni lehet a kiértékelés módszereit. Természetesen gyakoroltathatnánk a hiba feltárásának módszereit is. Ez azonban meghaladja az elsőéves oktatás lehetőségeit; ilyen mérésekre csak harmadévben nyílik lehetőség. Ez a *Méréselmélet* tantárgy tematikájába tartozik.

A regressziós számítások fajtái

- Lineáris regresszió
- Görbeillesztés
- Többváltozós regresszió

Az elsőéves fizika tantárgy a görbeillesztés problémáját úgy oldja meg, hogy a megfelelő transzformációk elvégzése után olyan értékkészlethez jutunk, amelyről a fizikai törvényszerűségek ismeretében feltételezhetjük, hogy az lineáris, tehát a hallgatóknak csak a lineáris regressziós számítások menetével kell tisztában lenniük.

Több mérési gyakorlat hasonló matematikai modellre épül. Ez az Arrhenius egyenlet.

Az Arrhenius egyenletet eredetileg kémiai reakciósebességek leírására dolgozták ki. A huszadik században ismerték fel, hogy vannak olyan *fizikai* jelenségek, amelyek hasonló egyenletekkel fejezhetőek ki. Ezeknél a független változó a hőmérséklet reciprok értéke, a függő változó viszont az exponenciális egyenlet kitevőjébe, annak is a nevezőjébe kerül. Ezzel a jelenséggel találkozhatnak elsőéveseink a viszkozitási mérésnél; a forráspont és a termisztor mérésénél.

Az értékkészleten a függvénytranszformáció elvégzése után alkalmazhatóvá válik a lineáris regressziós számítási módszer.

Ehhez táblázatba rendezzük a mérési eredményeket. Az összetartozó x és y azonos indexet viselnek (a képletekben i betűt). Legyen ***a*** a meredekség jele és ***b*** a tengelymetszeté. ***n*** a mérések száma.

Az eljárás teljes képletei:

meredekség $a=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$ tengelymetszet $b=\frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$

Látható, hogy ismétlődő kifejezések találhatóak a fenti képletekben, méghozzá ezeknek van megjelenítésre érdemes jelentésük, elsőként $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ a két változó átlaga. $s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ $s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ a két változó szórásnégyzete.

A kovariancia a változóknak az átlaguktól való eltéréseiknek szorzata: $cov_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ A fentiek ismertében könnyebb az eddig még nem ismert értékeket kiszámítani. A korrelációs együttható: $r=\frac{cov_{xy}}{s_x s_y}$ A regressziós együttható (vagyis az egyenes meredeksége): $a=r\frac{s_y}{s_x}$ (itt most a szórás szerepel, nem a négyzete). A regresszió ellentéte kiszámítható a két szórás cseréjével. Ez a fordított regressziós egyenes meredeksége; ahol a függő és független változó helyet cserél. A fizikában ennek nincs jelentősége, azért, mert az okot és az okozatot nem cserélhetjük fel egymással.

Végül a tengelymetszet más módon is kifejezhető: $b=\bar{y}-\bar{x}a$

Számítási példák

Számítások a sóoldatok sűrűségének példájával.

Példák az EXCEL felhasználásával. Középen a mért adatok láthatóak, mellettük ezek négyzete. Az alsó sorban a sorozat értékeinek összege szerepel.

négyzete	sűrűség	tömegkoncentráció	
	kg/m^3	kg/m3	
1024144	1012	20	400
1081600	1040	63	3969
1144900	1070	100	10000
1254400	1120	180	32400
1345600	1160	250	62500
1392400	1180	256	65536
7243044	6582	869	174805

Például a sűrűségmérések összege 6582 kg/m³, a hozzájuk tartozó értékek négyzetének összege 7243044 (kg/m³)²

Tekintettel arra, hogy a sűrűségnek is és a tömegkoncentrációnak is a görög ró betű a jele, itt ettől eltérünk: a sűrűség most az y (B3:B8), a tömegkoncentráció pedig az x (C3:C8) jelet kapja. Az átlagok egyszerűen számíthatóak (a mértékegység jelölését most elhagyjuk): **FIZLABOR**

$\bar{x}=\frac{6582}{5}=1097$ =ÁTLAG(b3:b8) =1097, $\bar{y}=\frac{869}{6}=148{,}83$ =ÁTLAG(C3:C8)=144,83

Ezt az adatpárt az eloszlás súlypontjának nevezzük. Értékének a regresszió estén közvetlen értelme nincs. Laboratóriumi mérésnél végeredményként jelölni téves és hibás!

A variancia és a szórás hasonló sorrendben:

=VAR(B3:B8)=4518, =SZÓRÁS(B3:B8)=62,22 (sűrűség)

=VAR(C3:C8)=9788,8, =SZÓRÁS(C3:C8)=98,94 (tömegkoncentráció) **FIZLABOR**

sűrűség	eltérés	tömegkoncentráció		
kg/m^3		kg/m3		
1012	-85	20	-124,833	10610,83
1040	-57	63	-81,8333	4664,5
1070	-27	100	-44,8333	1210,5
1120	23	180	35,16667	808,8333
1160	63	250	105,1667	6625,5
1180	83	256	111,1667	9226,833
1097		144,8333		6629,4

A kovariancia számításához széthúztuk a táblázatot. Mindkét változó mellett ott áll az átlagtól való eltérése (az átlag a legalsó sorba került). Az oszlopok betűjele balról-jobbra A, B, C, D, E. Az E oszlopba került a változóknak páronként az átlagtól való eltérésének szorzata. Összegüket elosztjuk a szabadsági fokkal, ez került az E9 cellába: =SZUM(E3:E8)/5=6629,4

Ellenőrizzük; az EXCEL ezt önállóan is ki tudja számítani ezt: =KOVARIANCIA.M(A3:A8;C3:C8) = 6629,4 (a KOVARIANCIA.S függvény az értékek darabszámával számol, a KOVARIANCIA.M viszont a szabadsági fokkal)

A kovariancia értékének ismertében számíthatjuk a korrelációs együtthatót:

=6629,4/(62,22*98,94) = 0,9968

A korrelációs együttható négyzetét determinációs együtthatónak nevezik: 0,9968² = 0,9937

Most már csak egy lépés a regressziós együttható, a =0,9968*62,22/98,94=0,6772

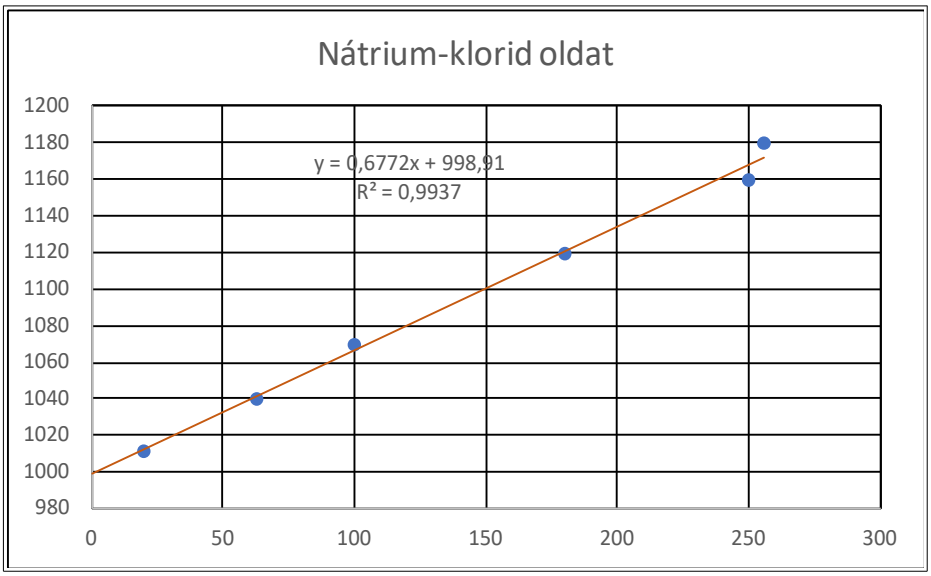
Nem szabad elfelejteni, ennek van mértékegysége. Ezt úgy kapjuk, hogy a függő változó mértékegységét elosztjuk a független változó mértékegységével, jelenleg kg/m³ / kg/m³ =1.

A tengelymetszet értékének számítása: $b=\bar{y}-\bar{x}\,a$ számértékekkel:

b= 1097-144,83×0,6772=998,914 **FIZLABOR**

Ennek mértékegysége is van, hiszen a sűrűséget ábrázoltuk: kg/m³

Most már megtekinthetjük, hogyan ábrázolja ezt az EXCEL. Független változó: a tömeg-koncentráció, függő változó: az oldat sűrűsége. Külön kérésre egyenest is illeszt, és adatait ráírja az ábrára (itt: meredekség, tengelymetszet, alul pedig a determinációs együttható).



A számunkra szükséges adatokat több különböző módon ki lehet számítani. Feltételezve, hogy a függő változó a B oszlopban van, a független változó a C oszlopban, a meredekség: =MEREDEKSÉG(B3:B8;C3:C8)=0,67723, a tengelymetszet:

=METSZ(B3:B8;C3:C8)=998,914

További lehetőséget kínál a lineáris illesztés. Ehhez tömbképletet kell beírnunk a következőképp. A mi esetünkben egyváltozós regressziót számítunk, ezért elég kijelölni egy olyan területet, amely két oszlopból és öt sorból áll. Írjuk be a parancssort: =LIN.ILL(B3:B8;C3:C8;IGAZ;IGAZ), de ekkor *még ne nyomjuk meg az ENTER billentyűt!* Ügyeljünk a változók közti pontosvesszőre. Az első logikai változó, amelyet beírtunk (IGAZ), azt jelzi, hogy szükségünk van a tengelymetszet értékére. A második azt jelenti, hogy a statisztikai adatokat is kérjük. Csak, ha mindez megvan, akkor nyomjuk meg **egyszerre** a *shift-control-enter* billentyűket. Ez a tömbképet beírásának módja.

0,677232	998,9143
0,026912	4,593563
0,993723	5,953904
633,254	4
22448,2	141,7959

Az első sor tartalmazza a meredekség és a tengelymetszet értékét.

A második sorban áll e két adatnak a szórása. Például a meredekség értéke: 0,677232±0,026912; a tengelymetszet értéke: 998,9143±4,593563 (bizony, eszerint nem lehetünk biztosak abban, hogy az egyenes a tiszta víz sűrűségének értékéhez irányul).

A harmadik sorban áll a determinációs együttható, mellette a függő változó szórása.

A negyedik sorban az F-próba, valamint a szabadsági fok értékét látjuk.

Az ötödik sor tartalmazza a számított és a mért értékek eltérésének négyzetösszegét. Az első a szóráselemzés értelmezése szerint az, amelyet a lineáris regresszió megmagyaráz (ssreg), a másik az, amelyet **nem** magyaráz meg – tehát a hiba (ssresid, reziduum, más néven maradék).

Számítás az Analysis Toolpak segítségével

ÖSSZESÍTŐ TÁBLA						
Regressziós statisztika						
r értéke	0,996857					
r-négyzet	0,993723					
Korrigált r	0,992154					
Standard h	5,953904					
Megfigyele	6					
VARIANCIAANALÍZIS						
	df	*SS*	*MS*	*F*	*ignifkanciája*	
Regresszic	1	22448,2	22448,2	633,254	1,48E-05	
Maradék	4	141,7959	35,44897			
Összesen	5	22590				
	Koeffiensek	*andard hib*	*t érték*	*p-érték*	*Alsó 95%*	*Felső 95%*
Tengelym	998,9143	4,593563	217,4596	2,68E-09	986,1605	1011,668
tömegkon	0,677232	0,026912	25,16454	1,48E-05	0,602512	0,751952

A táblázat több részének magyarázata megtalálható a szóráselemzés című fejezetben. Itt most a következőkre figyeljünk: a *Regresszióra* vonatkozó SS értéke sokkal nagyobb, mint a *Maradék* (reziduum). Ez azt jelenti, hogy a lineáris regressziós közelítés rendkívül jól illeszkedik. A két utolsó sorban áll, ami bennünket érdekel, a *Tengelymetszet;* alatta a *tömegkoncentráció* kifejezés a regressziós egyenes meredekségére utal. A *standard hiba* az *egyszeres* szórásra vonatkozik. Mindkettőre kiszámították a Student T-próba értékét is és a hozzá tartozó valószínűséget. Tőle jobbra a 95%-os valószínűségi szinthez tartozó konfidencia intervallum látható. Ennek értelmében a tengelymetszet 95% valószínűséggel becsülhető, hogy a 986,1605 és az 1011,668 tartományba esik;

a meredekség 95% valószínűséggel becsülhető, hogy a 0,602512 és a 0,751952 közé esik.

A regressziós egyenestől való eltérés

Ez az eltérés a konfidencia sáv szélessége. Előző ábránkon szükségtelenül szélesre rajzoltuk, hogy ne takarja el a mérési adatok pontjait (azon az ábrán az anyagmennyiség-koncentráció a független változó). Ezúttal a valóban használatos értékét számítjuk ki. A fent leírt módon számítjuk a regressziós egyenes adatait, majd valamennyi független változóhoz kiszámítjuk azt az értéket, amelyet a regressziós közelítésből kaptunk, és összehasonlítjuk a valóságosan mért eredménnyel. Az első mért pont például 1012, a számított érték: =$B$9*C3+$B$10=1012,459 (a C3 cellában a tömegkoncentráció értéke áll, 20 kg/m³

Számadatokkal. 0,677232×20+998,9143=1012,459

Számítjuk valamennyi eltérés négyzetét, majd azok összegét. Az első adatnál például: =(B3-A3)² numerikusan: (1012-1012,459)² = 0,21058

Ezeket az eltéréseket összegezzük, elosztjuk a szabadsági fokkal, majd kiszámítjuk a négyzetgyökét. A számítás tökéletesen megfelel a korrigált empirikus szórás folyamatának.

=GYÖK((SZUM(D3:D8)/4))=5,953904 – tessék megfigyelni, a LIN.ILL táblázatában ugyanezt az értéket látjuk a harmadik sorban. Most már magyarázattal is szolgálhatunk. A mért és számított adatok közötti négyzetes eltérés értéke az egész adathalmazra vonatkozóan ennyi. Ekkora a konfidencia sáv szélessége, ha az egyszeres szórással számolunk. A gyakorlatban inkább a szórás háromszorosát vesszük; ha az adatok normál eloszlást követnének, a szignifikancia szint 95% lenne. Tehát az adatok egy olyan sávban volnának, amelynél a sűrűség-értékek nem térnének el a számított értéktől jobban, mint 3×5,953904 =17,86171 kg/m³ (nézzük meg a kiinduló adatok táblázatát: egyik eltérés sem ekkora).

Számítási lehetőségek más szoftver felhasználásával

Más egyéb szoftver természetesen máshogy oldja meg ugyanazokat a feladatokat. Mi most megkeressük a fenti mintapéldák megfelelőit a SCILAB program alkalmazásával.

```
filename = fullfile( 'sooldat.txt');

rr= csvRead(filename,";" ); // ezúttal a mezőelválsztó a pntosvessző

//sorok és oszlopok mérete

nr = size(rr, "r")

nc = size(rr, "c")

// sur = sűrűség, osz = összetételi arány (aktuálisan tömegkoncentráció)

sur = rr(:,1)

osz = rr(:,2)

msur= mean(su)  //mean = átlag

mosz= mean(osz)

nsur = size(sur,”r”)  //number = elemek száma

nosz=size(osz,”r”)

C = cov( sur, osz)

C  =

 4518.     6629.4

 6629.4    9788.9667
```

Mint látjuk, a beépített kovariancia függvény a varianciákat és a covarianciákat írja ki – pontosabban: egy kovariancia mátrixot. Természetesen az egyik változónak a másikra vonatkoztatott kovarianciája egyenlő az ellenkezőjével (másiknak az egyikre)

A korrelációs együttható közvetlenül is kiszámítható:

rho = correl (sur, osz)

rho =

0.9968566

Természetesen a determinációs együttható is kiszámítható. A **det** nem használható,, mert az a *determináns*; a SCILAB egyik függvényének a neve

determ = correl (sur, osz)^2

determ =

0.9937231

A korrelációs együtthatóhoz vesszük a C mátrixnak akár az 1,2, akár a 2,1 indexű tagját:

rhoo = C(1,2)/(sosz*ssur)

rhoo =

0.9968566

Természetesen van regressziós függvénye is a SCILABnek. Sajnos, azonban a használatához nem oszlopvektorok, hanem sorvektorok szükségesek. A *mátrix transzponálást* hívjuk segítségül. Ehhez az *aposztrof* jelet kell beírnunk a változó azonosítója mellé. Például az eredeti oszlopvektor:

osz =

20.

63.

100.

180.

250.

256.

Most felcseréljük az oszlopokat sorokra:

osz'

ans =

20. 63. 100. 180. 250. 256.

Vagy beírhatjuk közvetlenül a transzponált vektort a regresszió képletébe:

[a, b, sig] = reglin (osz', sur')

sig =

4.8613422

b =

998.91425

a =

0.6772318

s íme, megvan a szignifikancia szint, a tengelymetszet és a meredekség is.

Megjegyzés: a kovariancia mátrix, más néven *szórásmátrix* négyzetes elrendezésű és rendezett formában tartalmazza az átlagtól való eltéréseket

$$\lambda_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) \qquad \lambda_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\lambda_{yx}=\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x}) \qquad \lambda_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})$$

$$C=\begin{bmatrix}\lambda_{xx} & \lambda_{xy}\\ \lambda_{yx} & \lambda_{yy}\end{bmatrix}$$ (mátrixként rendezett formában)

Az ábrát meg is nézhetjük. Pontok:

plot(osz, sur, 'db','Linewidth',3)

regressziós egyenes:

plot(osz, am*osz+b,'r','Linewidth',3)

xlabel("tömegkoncentráció", 'Fontsize',4)

ylabel("sűrűség",'Fontsize',4)