

Név: Unger Tamás István

Neptun: FTD1YJ

Web: <http://maxwell.sze.hu/~ungert>

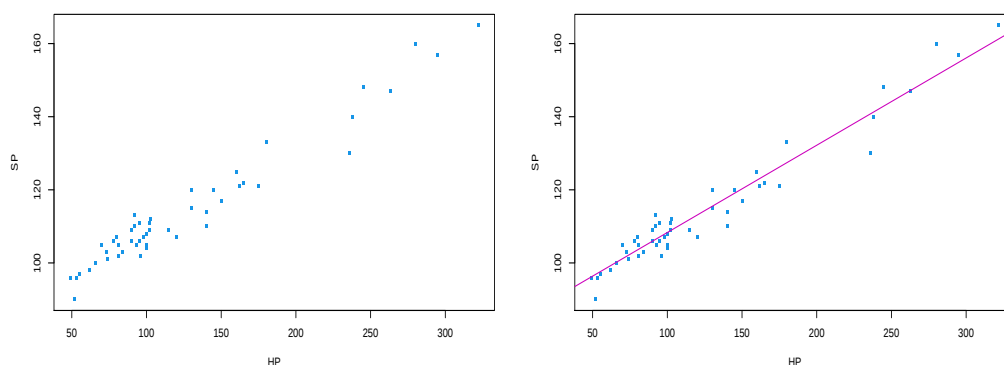
ALKALMAZOTT STATISZTIKA, BEADANDÓ FELADATOK

1. A "UScars.txt" adatsor a '80-as években az amerikai piacon forgalmazott néhány autótípus műszaki adatait tartalmazza. A változók:

- MODEL: a típus neve;
- COUNTRY: gyártási ország;
- VOL: utastér térfogata (köbláb);
- HP: teljesítmény (lóerő);
- MPG: gazdaságosság (mértföld/gallon);
- SP: végsebesség (mértföld/óra);
- WT: tömeg (100 font).

a. Ábrázolja az SP változót a HP függvényeként. A grafikon alapján milyen irányú kapcsolat tapasztalható a két változó között? Rajzolja be a regressziós egyenest is! A grafikon alapján mennyire jól illeszkednek a megfigyelések az egyeneshez?

A végsebesség a lóerő függvényében, valamint a regressziós egyenes az 1. ábrán látható.



1. ábra. A végsebesség a lóerő függvényében és a regressziós egyenes

Azt látjuk, hogy a két változó között szisztematikus, pozitív lineáris kapcsolat van. A grafikon alapján azt is látjuk, hogy a megfigyelések jól illeszkednek a berajzolt regressziós egyeneshez. A grafikon alapján azt tudjuk mondani, hogy a megfigyelések többsége az egyenes közelében, de természetesen nem pont az egyenesen található.

b. Végezzen lineáris regressziót a HP és SP változókon, fejezze ki a végsebességet a teljesítmény függvényeként, írja fel a regressziós egyenes egyenletét. Adja meg a négyzetes korrelációs együtthatót is! Mi a négyzetes korrelációs együttható jelentése? Most a négyzetes korrelációs együttható alátámasztja az **a.** pontban kapott eredményeket, vagy éppen ellentmond azoknak?

A lineáris regresszió-analízis négy tizedesjegy pontossággal a következő eredményt adja:

```
> lm(Uscars$SP~Uscars$HP)

Call:
lm(formula = Uscars$SP ~ Uscars$HP)

Coefficients:
(Intercept)      Uscars$HP 
    84.4541         0.2387
```

A regressziós egyenes képlete tehát a következő:

$$Y = 84,4541 + 0,2387X,$$

ezért a végsebesség a teljesítmény függvényeként:

$$\text{végsebesség} = 84,4541 + 0,2387 \cdot \text{teljesítmény},$$

A négyzetes korrelációs együtthatót kétféleképpen is meghatározhatjuk, ugyanarra az eredményre jutunk:

```
> (cov(Uscars$SP,Uscars$HP)*cov(Uscars$SP,Uscars$HP))
/(var(Uscars$SP)*var(Uscars$HP))
[1] 0.9342096
> cor(Uscars$SP,Uscars$HP)*cor(Uscars$SP,Uscars$HP)
[1] 0.9342096
```

A négyzetes korrelációs együttható megadja, hogy a végsebesség varianciájának mekkora hányadát magyarázza a teljesítmény varianciája. Ez az arány ebben az esetben 0,9342096 ($\approx 93,42096$ százalék), ami szoros kapcsolatra utal éppen úgy, ahogy arra észszerűen számítottunk.

c. Végezzen többszörös lineáris regressziót, írja fel az SP változót a HP és a WT függvényeként! Írja fel a regressziós egyenes egyenletét, továbbá adja meg a többszörös determinációs együtthatót! Mi a többszörös determinációs együttható jelentése? Ebben a feladatban milyen következtetést vonhatunk le a többszörös determinációs együttható alapján?

Futtassuk le a többszörös lineáris regresszió-analízist:

```
> print(summary(lm(Uscars$SP~Uscars$HP+as.numeric(Uscars$WT)),digits=1))

Call:
lm(formula = Uscars$SP ~ Uscars$HP + as.numeric(Uscars$WT))

Residuals:
    Min       1Q   Median       3Q      Max
```

-3.887 -1.006 0.158 1.046 3.173

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	96.419959	0.829768	116.20	<2e-16 ***
Uscars\$HP	0.322843	0.005987	53.92	<2e-16 ***
as.numeric(Uscars\$WT)	-0.705857	0.041801	-16.89	<2e-16 ***

Residual standard error: 1.698 on 79 degrees of freedom

Multiple R-squared: 0.9857, Adjusted R-squared: 0.9854

F-statistic: 2728 on 2 and 79 DF, p-value: < 2.2e-16

Látjuk, hogy mindkét változó hatása szignifikáns. Írjuk fel az SP változót a HP és a WT függvényeként:

$$SP = 96,419959 + 0,322843 \cdot HP - 0,705857 \cdot WT.$$

A többszörös determinációs együtthatót a következőképpen határozzuk meg:

```
> modell<-summary(lm(Uscars$SP~Uscars$HP+as.numeric(Uscars$WT)),digits=1)
> print(modell$r.squared,digits=4)
[1] 0.9857
```

Azt látjuk tehát, hogy $R^2 = 0,9857$. Tekintettel arra, hogy ez tulajdonképpen a négyzetes korrelációs együttható kiterjesztése több változóra, az 1-közeli érték szoros kapcsolatot mutat a végsebesség és a teljesítmény, valamint a végsebesség és a tömeg között. Világos, hogy minél nagyobb teljesítményű és minél kisebb tömegű egy jármű, annál nagyobb végsebességet tud elérni.

d. Hasonlítsa össze a b. illetve a c. pontban kapott regressziós modellt! Javított szignifikáns módon az illeszkedésen az a módosítás, hogy a modellbe belevettük a WT változót?

Végezzük el a kívánt variancia-analízist a következőképpen:

```
> g<-lm(Uscars$SP~Uscars$HP)
> h<-lm(Uscars$SP~Uscars$HP+as.numeric(Uscars$WT))
> print(anova(g,h),digits=3)
Analysis of Variance Table

Model 1: Uscars$SP ~ Uscars$HP
Model 2: Uscars$SP ~ Uscars$HP + as.numeric(Uscars$WT)
  Res.Df  RSS Df Sum of Sq  F Pr(>F)
1      80 1050
2      79  228  1      822 285 <2e-16 ***
---
```

Az RSS (csökkenő) értékéből azt látjuk, hogy szignifikáns módon javított az illeszkedésen az a módosítás, hogy a modellt kibővítettük a WT változóval.

2. Ebben a feladatban végig a "Uscars.txt" adatsor numerikus (szám értékű) változóival dolgozzon!

a. Írja fel az öt változó kovarianciamátrixát, és adja meg a mátrix sajátértékeit!

A kovarianciamátrixot a következőképpen kapjuk meg:

```
> Uscars_num <- read.table("UScars_num.txt", header=TRUE,
sep="\t", na.strings="NA", dec=".", strip.white=TRUE)
> cov<-cov(Uscars_num)
> print(cov,digits=3)
      VOL      HP      MPG      SP      WT
VOL 491.3   96.4  -81.7  -13.4   69.5
HP   96.4 3230.9 -449.2 771.2 385.1
MPG -81.7 -449.2 100.1 -96.7 -73.7
SP  -13.4 771.2  -96.7 197.1 77.5
WT   69.5 385.1  -73.7 77.5 66.3
```

Most pedig meghatározzuk a sajátértékeket és a sajátvektorokat:

```
> eigen<-eigen(cov)
> print(eigen,digits=3)
eigen() decomposition
$values
[1] 3528.025 508.026 36.595 12.080 0.937

$vectors
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] -0.0356 0.9774 0.207 -0.0216 -0.00916
[2,] -0.9567 -0.0492 0.076 0.1738 -0.21526
[3,] 0.1351 -0.1418 0.762 0.5915 0.17847
[4,] -0.2280 -0.0904 0.351 -0.6202 0.65706
[5,] -0.1151 0.1186 -0.498 0.4845 0.70000
```

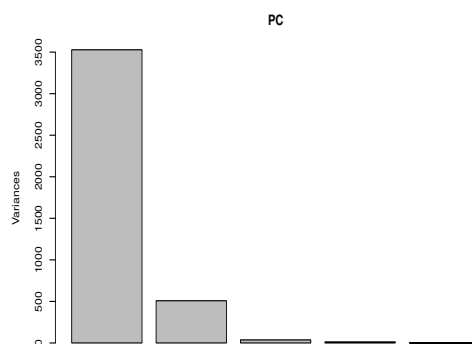
b. Végezzen főkomponens-analízist az öt változón! A kapott főkomponensek külön-külön milyen arányban magyarázzák a teljes adatsor varianciáját? Ezek az arányok milyen kapcsolatban állnak az a. pontban kapott sajátértékekkel? Hány főkomponensre van szükség, ha a teljes variancia 95 százalékát meg akarjuk magyarázni?

Végezzük el a főkomponens-analízist a következő módon:

```
> PC<-summary(prcomp(~Uscars_num$VOL+
Uscars_num$HP+Uscars_num$MPG+Uscars_num$SP+Uscars_num$WT))
> print(PC)
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5
Standard deviation	59.3972	22.5394	6.04939	3.47560	0.96809
Proportion of Variance	0.8635	0.1243	0.00896	0.00296	0.00023
Cumulative Proportion	0.8635	0.9879	0.99681	0.99977	1.00000

Az első főkomponens az összvariancia 86,35 százalékát magyarázza, a második mindössze a 12,43 százalékát, vagyis az első két variancia összesen 98,79 százalékot magyaráz, ahogyan az az alsó sorban is látszik. A további főkomponensek kevesebb, mint 1%-ot magyaráznak, és az öt főkomponens együtt a teljes varianciát magyarázza. Ez azt jelenti,



2. ábra. Oszlopdiagram (sajátértékek)

hogy két főkomponensre van szükségünk akkor, ha a teljes variancia 95%-át meg akarjuk magyarázni.

Jelölje a kapott sajátértékeket rendre $\lambda_1, \lambda_2, \dots, \lambda_5$, az ebben a feladatban kapott arányokat pedig rendre $\alpha_1, \alpha_2, \dots, \alpha_5$. A következő figyelhető meg:

$$\frac{\lambda_1}{\sum_{i=1}^5 \lambda_i} = \alpha_1, \quad \frac{\lambda_2}{\sum_{i=1}^5 \lambda_i} = \alpha_2, \dots, \quad \frac{\lambda_5}{\sum_{i=1}^5 \lambda_i} = \alpha_5.$$

c. Írja fel az első főkomponenst az adatsorban található öt változó segítségével! Az első

főkomponenst az öt változó segítségével a következőképpen írható fel:

```
> mean(Uscars_num$VOL)
[1] 98.80488
> mean(Uscars_num$HP)
[1] 117.1341
> mean(Uscars_num$MPG)
[1] 33.78171
> mean(Uscars_num$SP)
[1] 112.4146
> mean(Uscars_num$WT)
[1] 30.91463

> print(PC$rotation,digits=3)
      PC1      PC2      PC3      PC4      PC5
Uscars_num$VOL -0.0356 -0.9774  0.207 -0.0216  0.00916
Uscars_num$HP  -0.9567  0.0492  0.076  0.1738  0.21526
Uscars_num$MPG  0.1351  0.1418  0.762  0.5915 -0.17847
Uscars_num$SP  -0.2280  0.0904  0.351 -0.6202 -0.65706
Uscars_num$WT  -0.1151 -0.1186 -0.498  0.4845 -0.70000
```

ezért az első főkomponens:

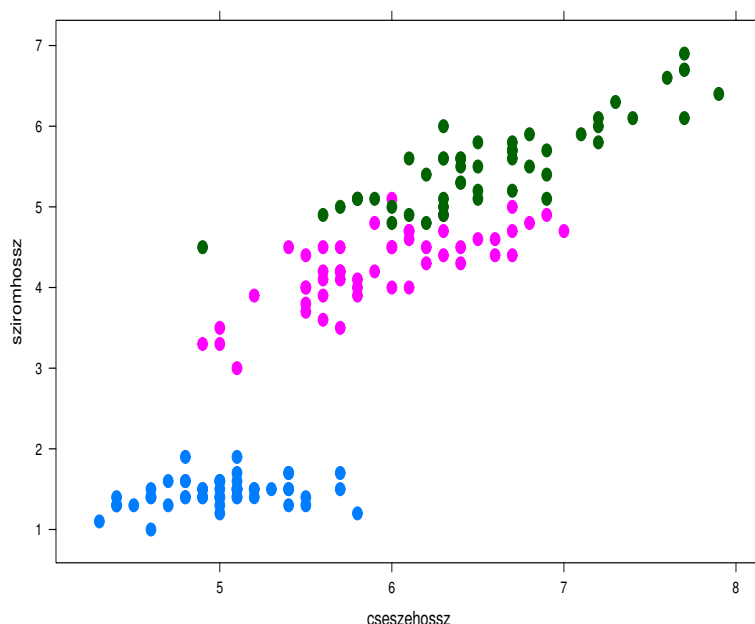
$$-0,0356 (\text{VOL} - 98,80488) - 0,9567 (\text{MP} - 117,1341) + 0,1351 (\text{MPG} - 33,78171) - 0,2280 (\text{SP} - 112,4146) - 0,1151 (\text{WT} - 30,91463).$$

3. Az "iriszadat.txt" adatsor írisz (nőszirm) virágokról tartalmaz megfigyeléseket. A minta három faj 50-50 növényét tartalmazza. A változók:

- **cseszehossz**, **cseszszel**: a csészelevél hosszúsága illetve szélessége (cm),
- **sziromhossz**, **sziromszel**: a sziromlevél hosszúsága illetve szélessége (cm),
- **faj**, **fajkod**: melyik fajból származik.

a. Ábrázolja grafikonon a "sziromhossz" változót a "cseszehossz" függvényeként! Ábrázolja különböző színnel a különböző fajhoz tartozó egyedeket! A grafikon alapján mely fajok különülnek el egymástól, és melyek nem?

A grafikont a 3. ábra mutatja. Kékkel a **setosa**, rózsaszínnel a **versicolor**, zölddel pedig a **virginica** faj megfigyelései láthatók. Azt látjuk, hogy a **setosa** faj elkülönül



3. ábra. A sziromhossz a cseszehossz függvényeként

a másik kettőtől, de a **versicolor** és a **virginica** fajok nem különülnek el érdemben egymástól.

b. Végezzen lineáris diszkriminancia-analízist az első négy változón! Adja meg az első diszkriminanciafüggvény egyenletét!

A következőképpen járunk el:

```
> man<-manova(cbind(cseszehossz,cseszszel,sziromhossz,sziromszel)
~faj, data=iriszadat)
> DA<-summary(man, test="Wilks")
> print(DA, digits=3)
      Df      Wilks approx F num Df den Df      Pr(>F)
```

```
faj          2 0.023439   199.15      8   288 < 2.2e-16 ***
Residuals 147
---
```

Tovább haladva:

```
> library(MASS)
> diszkr<-lda(faj~cseszehossz+cseszeszel+szioromhossz+szioromszel,
  data=iriszadat)
> print(diszkr, digits=3)
Call:
lda(faj ~ cseszehossz + cseszeszel + szioromhossz + szioromszel,
  data = iriszadat)
```

```
Prior probabilities of groups:
      setosa versicolor virginica
      0.333      0.333      0.333
```

Group means:

	cseszehossz	cseszeszel	szioromhossz	szioromszel
setosa	5.01	3.43	1.46	0.246
versicolor	5.94	2.77	4.26	1.326
virginica	6.59	2.97	5.55	2.026

Coefficients of linear discriminants:

	LD1	LD2
cseszehossz	0.829	0.0241
cseszeszel	1.534	2.1645
szioromhossz	-2.201	-0.9319
szioromszel	-2.810	2.8392

Proportion of trace:

	LD1	LD2
	0.9912	0.0088

Az első diszkriminanciafüggvény ezek alapján a következő:

$$Z = 0,829 \cdot \text{cseszehossz} + 1,534 \cdot \text{cseszeszel} - 2,201 \cdot \text{szioromhossz} - 2,810 \cdot \text{szioromszel}.$$

c. Találunk egy nősziorom növényt, amely biztosan a három faj valamelyikébe tartozik, de nem tudjuk, hogy melyikbe. A növény esetében a négy numerikus mennyiség értéke: 5, 4, 2, 1. A diszkriminancia-analízis segítségével adjon előrejelzést arra, hogy ez a növény melyik fajba tartozik!

A következőképpen járunk el:

```
> newdata<-rbind(c(5,4,2,1))
> dimnames(newdata)<-list(NULL,
  c("cseszehossz","cseszeszel","szioromhossz","szioromszel"))
> newdata<-data.frame(newdata)
> pred<-predict(diszkr,newdata=newdata)
```

```
> print(pred$posterior,digits=3)
      setosa versicolor virginica
1         1    1.84e-11  3.58e-25
```

Az eredmény alapján azt mondhatjuk, hogy ez a növény nagy valószínűséggel a **setosa** fajba tartozik.