GP modellek IV.

Fegyverneki Sándor

Miskolci Egyetem
Alkalmazott Matematikai Intézeti Tanszék
matfs@uni-miskolc.hu

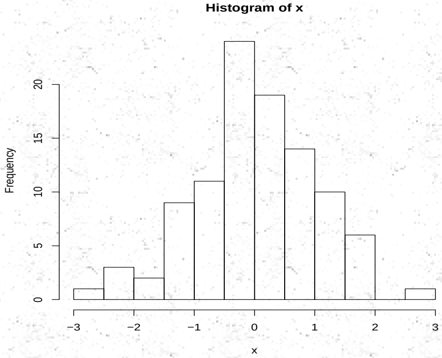
2021. március 01.

Full Screen
Search
Close

1 Bevezetés

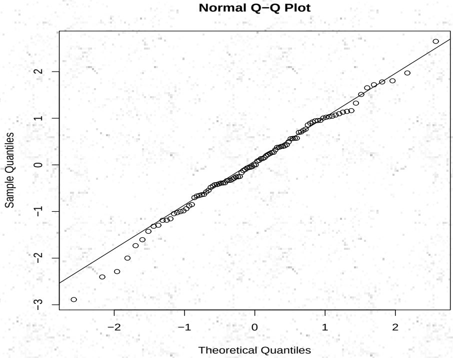
```
set.seed=139
x=rnorm(100)
hist(x)
```

Figure 1: A hisztogram



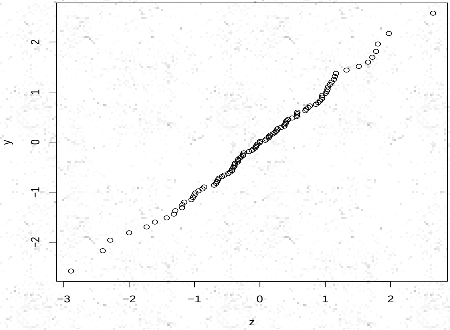
```
qqnorm(x)
qqline(x)
```

Figure 2: Pontok + egyenes



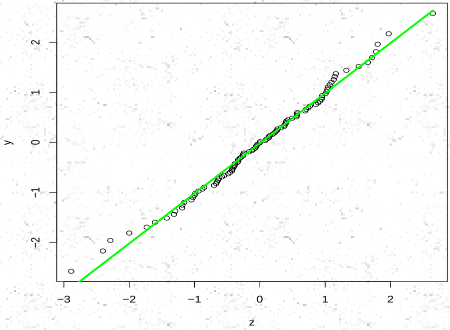
```
z=sort(x)
y=vector()
for (i in 1:length(x))
y[i]=qnorm((i-0.5)/length(x))
plot(z,y)
```

Figure 3: Pontok (saját)



```
reg.curve.yz=lm( y ~ z)
lines(predict(reg.curve.yz)~z,
col="green", lwd=3)
```

Figure 4: Pontok + egyenes (saját)



```
summary(reg.curve.yz)

Call:
lm(formula = y ~ z)

Residuals:
     Min          1Q    Median
        3Q         Max
-0.13898 -0.04780 -0.01391
0.01764  0.34996

Coefficients:
             Estimate Std. Error
             t value Pr(>|t|)
(Intercept) -0.018789   0.009242
-2.033   0.0448 *
z            1.000879   0.009347
 107.076   <2e-16 ***
```

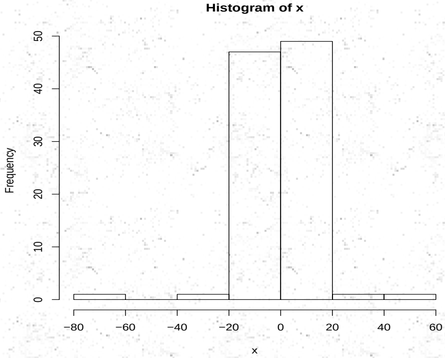
```
---
Signif. codes:  0 *** 0.001 **
0.01 * 0.05 . 0.1 ' ' 1

Residual standard error: 0.0924
on 98 degrees of freedom
Multiple R-squared:  0.9915,
Adjusted R-squared:  0.9914
F-statistic: 1.147e+04
on 1 and 98 DF,  p-value: < 2.2e-16
```


Full Screen
Search
Close

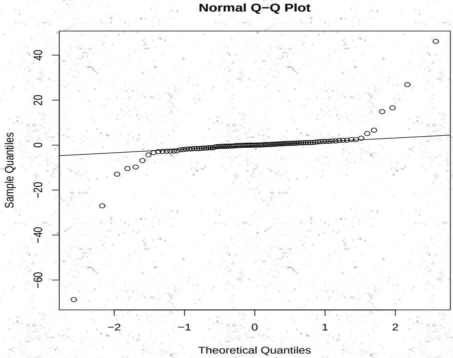
```
set.seed=139
x=rcauchy(100)
hist(x)
```

Figure 5: A hisztogram



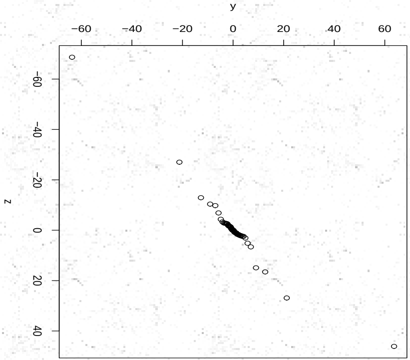
```
qqnorm(x)
qqline(x)
```

Figure 6: Pontok + egyenes



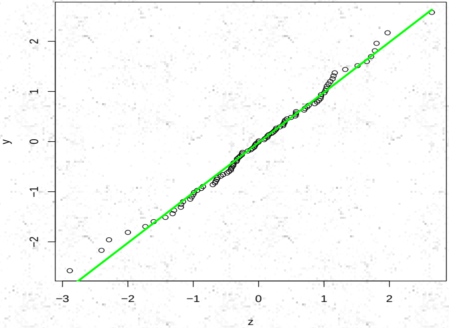
```
z=sort(x)
y=vector()
for (i in 1:length(x))
y[i]=qcauchy((i-0.5)/length(x))
plot(z,y)
```

Figure 7: Pontok (saját)



```
reg.curve.yz=lm( y ~ z)
lines(predict(reg.curve.yz)~z,
col="green", lwd=3)
```

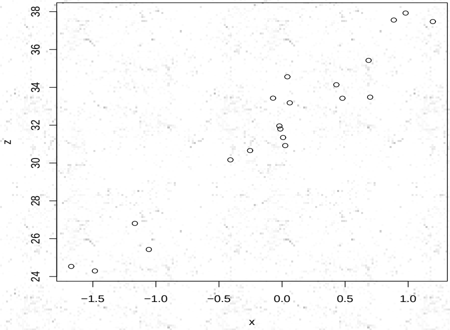
Figure 8: Pontok + egyenes (saját)



2 Regresszió

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
z=a+b*x+epsilon
plot(x,z)
```

Figure 9: Regresszió



3 Alapfeladat

Adott (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$.

Keressük azt a függvényt,

$$y = g(x),$$

amelyre

$$Q(g) = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

minimális.

Ez azt jelenti, hogy minimalizáljuk az

$$\varepsilon_i = y_i - g(x_i)$$

véletlennek tekintett rezidumok négyzetösszegét.

Full Screen
Search
Close

Általános feladat:

X és Y valószínűségi változók. Keressük azt a g valós függvényt, amelyre

$$\min_g E((Y - g(X))^2).$$

Megoldás a feltételes várható érték:

$$E(Y|X = x).$$

Diszkrét eset:

$$E(Y|X = x_i) = \sum_{j=1}^{\infty} y_j \frac{p_{ij}}{q_i}$$

Folytonos eset:

$$E(Y|X = x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) dy = m_1(x)$$

A függvényt az Y-nak az X-re vonatkozó regressziós függvényének nevezzük.

Full Screen
Search
Close

Normális eloszlás

$$(X, Y) \sim N(m_1, m_2, {\sigma_1}^2, {\sigma_2}^2, \rho).$$

ekkor

$$\begin{aligned} m_1(x) =& m_2 + \rho \frac{\sigma_2}{\sigma_1}(x - m_1), \\ m_2(y) =& m_1 + \rho \frac{\sigma_1}{\sigma_2}(y - m_2). \end{aligned}$$

(1)

Tehát a regressziós függvények egyenesek.

GP modellek
Bevezetés
Regresszió
Alapfeladat
Regressziós egyenes
Regressziós módszerek
Példák R-ben
Egy robusztus becslés
A lognormális eloszlás
A Weibull-eloszlás
ω^2-próba
Kolmogorov-próba
Vége

Page 25 of 88

Full Screen
Search
Close

4 Regressziós egyenes

A regressziós egyenes meghatározása: Adott (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$. Keressük azt az egyenest,

$$y = bx + a,$$

amelyre

$$Q(a,b) = \sum_{i=1}^{n} (y_i - bx_i - a)^2$$

minimális. Ez azt jelenti, hogy minimalizáljuk az $\varepsilon_i = y_i - bx_i - a$ véletlennek tekintett rezidumok négyzetösszegét.

Vezessük be a következő jelöléseket:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

$$Q_x = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2, \quad Q_y = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2, \quad Q_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}.$$

Ekkor a regressziós egyenes együtthatói

$$b_{yx} = \frac{Q_{xy}}{Q_x}, \quad a_{yx} = \overline{y} - b_{yx}\overline{x}.$$

A hibák, a konfidencia-intervallumok és a hipotézisvizsgálatok leírásához bevezetett jelölésekben az s a korrigált tapasztalati szórásra utal.

$$Q_{y.x} = Q_y - b_{yx}Q_{xy} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = Q(a_{yx}, b_{yx}),$$

$$r = \frac{Q_{xy}}{\sqrt{Q_x Q_y}}, \quad s_x^2 = \frac{Q_x}{n-1}, \quad s_y^2 = \frac{Q_y}{n-1}, \quad s_{xy}^2 = \frac{Q_{xy}}{n-1},$$

$$s_{y.x}^2 = \frac{Q_{y.x}}{n-2}, \quad s_{b_{yx}}^2 = \frac{s_{y.x}^2}{Q_x}, \quad s_{a_{yx}}^2 = s_{y.x}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{Q_x} \right).$$

Két egyenes összehasonlítása: Adott két minta. A mintaelemszámok n és m. A formulákban az 1 illetve 2 indexek az aktuális mintára utalnak. A b_1 és b_2 regressziós együtthatók összehasonlításához először végezzük el a következő F-próbát:

$$\frac{s^2_{y_1.x_1}}{s^2_{y_2.x_2}} > F_{(n-2,m-2;0.05)}.$$

A további lépéseinket az F-próba alapján két irányba folytathatjuk:

1. Ha az F-próba elfogad, azaz feltehetjük, hogy a szórásnégyzetek megegyeznek, akkor a

$$\hat{t} = \frac{b_1 - b_2}{\sqrt{\dfrac{Q_{y_1.x_1} + Q_{y_2.x_2}}{n+m-4}\left[\dfrac{1}{Q_{x_1}} + \dfrac{1}{Q_{x_2}}\right]}}$$

statisztika közelítőleg Student-eloszlású $n+m-4$ szabadságfokkal.

2. Ha az F-próba elutasít, azaz feltehetjük, hogy a szórásnégyzetek különböznek, akkor elég nagy minták esetén $(n > 20, m > 20)$ a

$$\hat{t} = \frac{b_1 - b_2}{\sqrt{\dfrac{s^2_{y_1.x_1}}{Q_{x_1}} + \dfrac{s^2_{y_2.x_2}}{Q_{x_2}}}}$$

statisztika közelítőleg Student-eloszlású ν szabadságfokkal, ahol

$$\nu = \frac{1}{\dfrac{c^2}{n-2} + \dfrac{(1-c)^2}{m-2}} \quad \text{és} \quad c = \frac{\dfrac{s^2_{y_1.x_1}}{Q_{x_1}}}{\dfrac{s^2_{y_1.x_1}}{Q_{x_1}} + \dfrac{s^2_{y_2.x_2}}{Q_{x_2}}}, \quad n \le m.$$

Regressziós egyenesekre vonatkozóan sokszor csupán ezt a párhuzamossági kérdést vizsgálják. Ha elfogadtuk a párhuzamossági hipotézist, akkor a regressziós együttható most már a két minta adataiból a következőképpen becsülhető: Hátra van még az egyenesek azonosságára vonatkozó döntés kérdése. Ezt további két részre kell bontanunk.

1. Ha $\overline{x}_1 \neq \overline{x}_2$, akkor a közös regressziós együttható megadható

$$\frac{\overline{y}_2 - \overline{y}_1}{\overline{x}_2 - \overline{x}_1}$$

formában is.

2. Ha $\overline{x}_1 = \overline{x}_2$, akkor az $\overline{y}_1 = \overline{y}_2$ biztosítja a regressziós egyenesek azonosságát, így azt kell megvizsgálnunk. A szokásos kétmintás t-próba alkalmas ennek eldöntésére.

Ha nem szükséges a párhuzamosság, akkor közvetlenül alkalmazható a következő F-próba:

$$\hat{F} = \frac{n+m-4}{2} \frac{Q_{y.x;T} - Q_{y_1.x_1} - Q_{y_2.x_2}}{Q_{y_1.x_1} + Q_{y_2.x_2}} > F_{(2,n+m-4;0.05)},$$

ahol $Q_{y.x;T}$ az egyesített mintára vonatkozó érték.

Full Screen

Search

Close

5 Regressziós módszerek

Elsőként tekintsük a többszörös lineáris regressziót. Ez nem más, mint az egyszerű lineáris regresszió általánosítása arra az esetre, amikor legalább két magyarázó változó van. Legyen

$$g(\xi_1, \xi_2, \dots \xi_k) = \beta_0 + \beta_1 \xi_1 + \dots + \beta_k \xi_k$$

a többszörös regresszió egyenlete, ahol $\xi_1, \xi_2, \dots \xi_k$ a k darab független (magyarázó) változó, amelyek kapcsolódnak az η (függő) változóhoz. Az adatok

$$(Y_i, X_{i1}, \dots, X_{ik}), \quad i = 1, 2, \dots, n,$$

egy $(k+1)-$dimeziós tér pontjainak tekinthető. A

$$\beta_i, \quad i = 0, 1, \dots, k,$$

Full Screen
Search
Close

a regressziós együttható. Egyenként nincs jelentésük viszont együtt fontos szerepet játszanak az η értékeinek előrejelzésében. Legyen

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \ddots & \ddots & \vdots & \ddots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}.$$

Jelölje $E(Z)$ a várható értékek mátrixát a Z mátrix elemeihez. Feltesszük, hogy

$$E(Y) = X\beta,$$

ahol $\beta^T = (\beta_0, \beta_1, \dots \beta_k)$. Legyen $\varepsilon = Y - E(Y)$. ε elemeit rezidumoknak nevezzük. Ekkor

$$Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0.$$

Ha $X^T X$ nemszinguláris, akkor a regressziós módszerek lényege a következő eredményekben foglalható össze:

1. A $b = (X^T X)^{-1} X^T Y$ a legkisebb négyzetes paraméterbecslés a β vektorra.

2. A $b = (X^T X)^{-1} X^T Y$ a β maximum-likelihood becslése, ha Y $n-$dimenziós normális eloszlású.

3. $E(b) = \beta$.

4. Legyen $cov(Y) = \sigma^2 I$, akkor $cov(b) = \sigma^2 (X^T X)^{-1}$. I az egységmátrix.

5. Ha Y normális eloszlású, akkor b $(k+1)-$dimenziós normális eloszlású.

6. Normalitást feltételezve a σ^2 maximum-likelihood becslése

$$\tilde{\sigma}^2 = \frac{Y^T [I - X(X^T X)^{-1} X^T] Y}{n}.$$

7. $E(\tilde{\sigma}^2) = (n-k-1)\sigma^2/n$, így

$$\hat{\sigma}^2 = \frac{Y^T[I - X(X^TX)^{-1}X^T]Y}{n-k-1}$$

a σ^2 torzítatlan becslése.

8. Ha Y normális eloszlású, akkor

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} = \frac{Y^T[I - X(X^TX)^{-1}X^T]Y}{\sigma^2} \sim \chi^2_{n-k-1}.$$

9. b és $\hat{\sigma}^2$ függetlenek.

10. Bármely $h^T\beta$ lineáris kombinációra (ha nincs normalitás akkor is) a h^Tb minimális szórásnégyzetű lineáris torzítatlan becslés.

Hipotézisvizsgálat: Normalitás feltételezése esetén

$$\frac{h^Tb - h^T\beta}{s_{h^Th}} \sim t_{n-k-1},$$

azaz Student-eloszlású, ahol $s_{h^T h} = \sqrt{\hat{\sigma}^2 h^T (X^T X)^{-1} h}$.
A $h^T \beta = a$ hipotétis vizsgálatára a

$$t = \frac{h^T b - a}{s_{h^T h}}$$

teszt statisztika éppen megfelelő. Sőt ez alapján $h^T \beta$ értékére konfidenciaintervallum készíthető.

Az illesztett egyenlet minőségének a vizsgálata az előzőek alapján. Néhány használatos mérőszám.

A rezidumok négyzeteinek az összege

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Az $\overline{Y}$ átlagtól való négyzetes eltérések teljes összege

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Népszerű még a négyzetes többszörös korrelációs együttható

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Erre teljesül, hogy $0 \leq R^2 \leq 1$, és azt szokás mondani, ha közel van egyhez akkor jó az illeszkedés.

Vigyázat RSS és TSS direkten nem összehasonlíthatók, hiszen az egyik esetben csak egy míg a másik esetben $(k+1)$ paramétert becsültünk.

Egy másik mérőszám sorozat az előző négyzetek átlaga (a torzítatlanság figyelembe vétele):

$$RMS = \frac{RSS}{n-k-1}, \quad TMS = \frac{TSS}{n-1}, \quad R_a^2 = 1 - \frac{RMS}{TMS},$$

ahol R_a^2 az ún. korrigált ("adjusted") R^2.

Példa:

Adottak a következő megfigyelési adatok.

$$\begin{array}{ccc} Y & X_1 & X_2 \\ 13 & 4 & 1 \\ 7 & 5 & 3 \\ 2 & 2 & 3 \\ 15 & 6 & 1 \\ 8 & 3 & 2 \end{array}.$$

Ekkor

$$Y = \begin{pmatrix} 13 \\ 7 \\ 2 \\ 15 \\ 8 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 5 & 3 \\ 1 & 2 & 3 \\ 1 & 6 & 1 \\ 1 & 3 & 2 \end{pmatrix}.$$

$$X^T X = \begin{pmatrix} 5 & 20 & 10 \\ 20 & 90 & 37 \\ 10 & 37 & 34 \end{pmatrix},$$

$$(X^T X)^{-1} = \begin{pmatrix} 5.103 & -0.710 & -1.032 \\ -0.710 & 0.129 & 0.097 \\ -1.032 & 0.097 & 0.323 \end{pmatrix},$$

$$X^T Y = \begin{pmatrix} 45 \\ 205 \\ 71 \end{pmatrix}, \quad b = \begin{pmatrix} 10.871 \\ 1.387 \\ -3.710 \end{pmatrix}, \quad \hat{\sigma}^2 = 0.419.$$

Ha meg akarjuk vizsgálni a $H_0 : \beta_1 = \beta_2$ hipotézist, ekkor a $h^T = (0, 1, -1)$ választással ez átírható a $H_0 : h^T \beta = 0$ hipotézissé. Viszont

$$s^2_{h^T h} = 0.1081, \quad \text{azaz} \quad t = \frac{h^T b}{s_{h^T h}} = \frac{5.097}{0.3288} = 15.5,$$

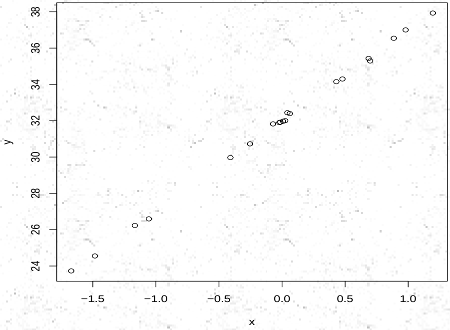
amelyből következik H_0 elutasítása, hiszen a kritikus érték 0.95-ös szinten 2.35.

6 Példák R-ben

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
y=a+b*x+0.1*epsilon
plot(x,y)
```

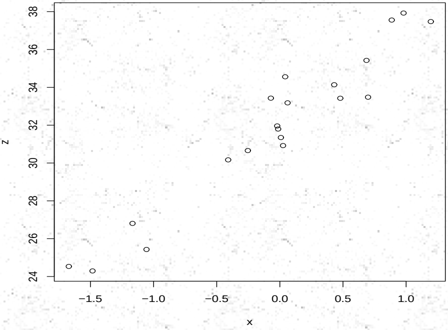

Full Screen
Search
Close

Figure 10: Regresszió



```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
z=a+b*x+epsilon
plot(x,z)
```

Figure 11: Regresszió

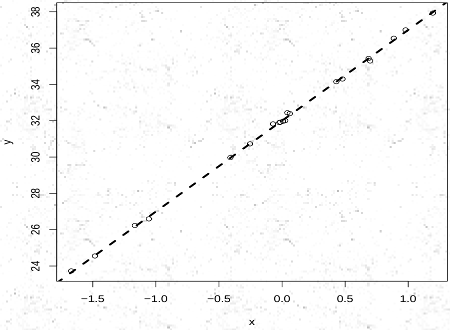


Full Screen
Search
Close

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
y=a+b*x+0.1*epsilon
plot(x,y)
abline(a,b, lty=2, lwd=3)

abline(lm(y~x), col="red", lwd=3)
```

Figure 12: Regresszió



```
summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
      Min          1Q     Median
            3Q        Max
-0.204676 -0.056521 -0.005408
       0.070688  0.226248

Coefficients:
            Estimate Std. Error
            t value Pr(>|t|)
(Intercept) 32.00867    0.02460
1301.0   <2e-16 ***
x            4.99216    0.03142
158.9   <2e-16 ***
```

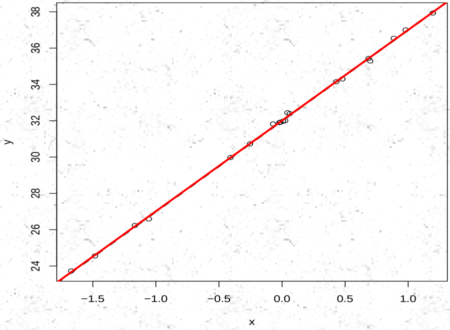
```
---
Signif. codes:  0 ’***’ 0.001
’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1099
 on 18 degrees of freedom
Multiple R-squared:  0.9993,
 Adjusted R-squared:  0.9992
F-statistic: 2.525e+04 on 1 and 18 DF,
 p-value: < 2.2e-16
```

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
z=a+b*x+epsilon
plot(x,z)
abline(a,b, lty=2, lwd=3)

abline(lm(z~x), col="red", lwd=3)
```

Figure 13: Regresszió



```
summary(lm(z~x))

Call:
lm(formula = z ~ x)

Residuals:
     Min         1Q   Median
          3Q        Max
-2.04676 -0.56521 -0.05408
0.70688  2.26248

Coefficients:
            Estimate Std. Error
            t value Pr(>|t|)
(Intercept)  32.0867     0.2460
130.42  < 2e-16 ***
x             4.9216     0.3142
15.66 6.22e-12 ***
```

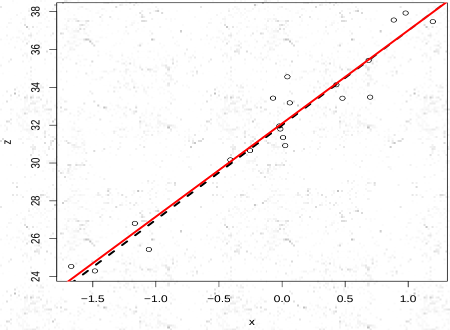
```
---
Signif. codes:  0 '***' 0.001 '**'
 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.099 on 18
 degrees of freedom
Multiple R-squared:  0.9317,
Adjusted R-squared:  0.9279
F-statistic: 245.4 on 1 and 18 DF,
p-value: 6.222e-12
```

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
y=a+b*x+epsilon
plot(x,y)
abline(a,b, lty=2, lwd=3)

abline(lm(y~x), col="red", lwd=3)
```

Figure 14: Regresszió



```
summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min          1Q    Median
        3Q         Max
-10.2338  -2.8261  -0.2704
3.5344  11.3124

Coefficients:
            Estimate Std. Error
            t value Pr(>|t|)
(Intercept)   32.434      1.230
26.366  7.8e-16 ***
x              4.608      1.571
2.934  0.00888 **
```

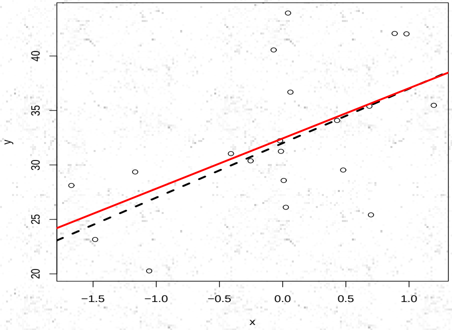
```
---
Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.497 on 18
degrees of freedom
Multiple R-squared:  0.3234,
Adjusted R-squared:  0.2859
F-statistic: 8.605 on 1 and 18 DF,
p-value: 0.008878
```

```
set.seed(38)
n=20
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
y=a+b*x+5*epsilon
plot(x,y)
abline(a,b, lty=2, lwd=3)

abline(lm(y~x), col="red", lwd=3)
```

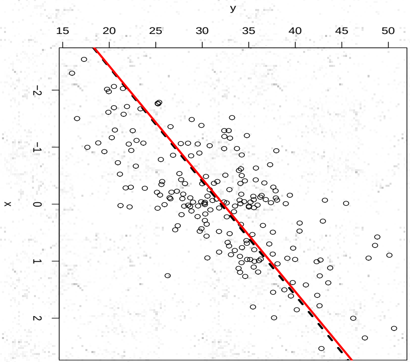
Figure 15: Regresszió



```
set.seed(38)
n=200
a=32;b=5
x=rnorm(n);epsilon=rnorm(n)
y=a+b*x+5*epsilon
plot(x,y)
abline(a,b, lty=2, lwd=3)

abline(lm(y~x), col="red", lwd=3)
```

Figure 16: Regresszió



```
summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min        1Q   Median
         3Q       Max
-12.3088  -3.5666  -0.7507
3.6359  13.6598

Coefficients:
            Estimate Std. Error
            t value Pr(>|t|)
(Intercept)  32.2465     0.3562
90.52   <2e-16 ***
x             5.0534     0.3700
13.66   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.034 on 198
 degrees of freedom
Multiple R-squared:  0.485,
Adjusted R-squared:  0.4824
F-statistic: 186.5 on 1 and 198 DF,
p-value: < 2.2e-16
```

7 Egy robusztus becslés

Két eloszlást azonos típusúnak nevezünk, ha léteznek a és $b > 0$ valós számok, amelyre

$$F(x) = G(bx + a),$$

ahol F és G a megfelelő eloszlásfüggvények.

Ez a típus definíció egy osztályozást határoz meg az eloszlások halmazán. Válasszunk ki egy osztályt és legyen ennek egy reprezentánsa F_0, ekkor az osztály egy tetszőleges elemére igaz, hogy létezik a és $b > 0$ valós szám, amelyre

$$F(x) = F_0\left(\frac{x-a}{b}\right).$$

Megjegyzés:

1. Ha a ξ valószínűségi változó eloszlásfüggvénye F_0 és az η eloszlásfüggvénye F, akkor

$$\eta = b\xi + a.$$

2. Az így adódó a illetve b értéket hely- illetve skála-paraméternek nevezzük.

A hely- és skálaparaméter becslésére vonatkozó feladatot a következőképpen fogalmazhatjuk meg:

Adott a $\xi_1, \xi_2, \ldots, \xi_n$ minta, amely elemeinek közös eloszlásfüggvénye F, amelyről tudjuk, hogy az F_0 eloszlásfüggvénnyel reprezentált osztályhoz tartozik. Határozzuk meg az F eloszlásfüggvényt, azaz azokat az a illetve $b > 0$ értékeket, amelyre teljesül

Ekkor az eloszlásfüggvények tulajdonságai alapján felírhatók a következő egyenlőségek:

$$\int_{-\infty}^{+\infty} \left(F_0 \left(\frac{x-a}{b} \right) - 0.5 \right) dF_(x) = 0$$

$$\int_{-\infty}^{+\infty} \left(F_0 \left(\frac{x-a}{b} \right) - 0.5 \right)^2 dF_(x) = \frac{1}{12}.$$

A Monte Carlo módszerek elmélete (a nagy számok erős törvénye) alapján az egyenlőségekből következik, hogy

$$\frac{1}{n}\sum_{k=1}^{n}\left(F_0\left(\frac{x-a}{b}\right)-0.5\right)\cong 0$$

$$\frac{1}{n}\sum_{k=1}^{n}\left(F_0\left(\frac{x-a}{b}\right)-0.5\right)^2\cong \frac{1}{12}.$$

A relációkat egyenlőségként tekintve kapunk egy egyenletrendszert az a és b ismeretlenekre. Az egyenletrendszert a következő algoritmussal oldhatjuk meg.

Jelölje az a illetve b közelítését (becslését) a $m-$edik iterációs lépésben $\vartheta_n^{(m)}$ illetve $s_n^{(m)}$. Ekkor legyen

$$\vartheta_n^{(0)} = med\{\xi_i\},$$

$$s_n^{(0)} = MAD\{\xi_i\},$$

Full Screen

Search

Close

ahol $med\{\xi_i\}$ a minta mediánja és

$$MAD\{\xi_i\} = med\{|\xi_i - med\{\xi_i\}|\}$$

pedig a medián abszolút eltérés.

Az algoritmus iterációs lépései a következők:

$$\vartheta_n{}^{(k+1)} = \vartheta_n{}^{(k)} + \frac{s_n{}^{(k)} \sum_{i=1}^{n} \psi\left(\frac{\xi_i - \vartheta_n{}^{(k)}}{s_n{}^{(k)}}\right)}{n}$$

$$[s_n{}^{(k+1)}]^2 = \frac{1}{n\beta} \sum_{i=1}^{n} \psi^2\left(\frac{\xi_i - \vartheta_n{}^{(k+1)}}{s_n{}^{(k)}}\right) [s_n{}^{(k)}]^2$$

Az algoritmus végrehajtásához szükséges az F_0 eloszlásfüggvény illetve helyettesítési értékeinek a kiszámítása.

A Gauss (normális) eloszlás

A mintaelemek eloszlásfüggvénye

$$F(x) = \int_{-\infty}^{x} \frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{(u-a)^2}{2b^2}\right) du,$$

és az F_0 reprezentáns eset az $a = 0$, $b = 1$.

A matematikai statisztika klasszikus elmélete alapján a paraméterek legjobb becslései az átlag (ϑ) és a korrigált tapasztalati szórás (s), azaz

$$\vartheta = \frac{\sum_{i=1}^{n} \xi_i}{n}, \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (\xi_i - \vartheta)^2}{n-1}}$$

Full Screen

Search

Close

Megjegyzés:

1. Az átlag és a korrigált tapasztalati szórás megkapható a maximum likelihood módszer alkalmazásával.

2. A Gauss-hálózat (papír) szemléletes ellenőrzési lehetőség, amely azt az állítást használja fel, hogy ha a ξ valószínűségi változó eloszlásfüggvénye F, akkor $F(\xi)$ egyenletes eloszlású a $(0,1)$ intervallumon.

Full Screen

Search

Close

8 A lognormális eloszlás

A ξ valószínűségi változó lognormális eloszlású, ha az $\eta = \ln\xi$ normális eloszlású, ahol $E(\eta) = a$ és $D^2(\eta) = b^2$. A sűrűségfüggvénye

$$f_\xi(x) = \begin{cases} \frac{1}{bx\sqrt{2\pi}} \exp\left[-\frac{1}{2b^2}(\ln x - a)^2\right], & \text{ha } x > 0, \\ 0, & \text{ha } x \le 0. \end{cases}$$

Továbbá

$$E(\xi) = \exp(a + \frac{b^2}{2})$$

és

$$D^2(\xi) = \exp(2a + 2b^2) - \exp(2a + b^2).$$

Látható, hogy a szokásos módszerek (maximum likelihood, momentum) közvetlenül nem alkalmazhatóak, ezért a problémát visszavezetjük a logaritmikus transzformációval a normális esetre.

A lognormális eloszlás aszimmetriája következtében alkalmas például bizonyos anyagtulajdonságok külső hatások miatti megváltozásának leírására. Ezt alkalmaztuk a lognormális-hálózat (papír) elkészítésekor is.

9 A Weibull-eloszlás

A Weibull-eloszlás paramétereire többféle elterjedt jelölésrendszer van. Az eltérő jelölések használatát egyértelműen magyarázza, hogy a Weibull-eloszlás igen széles körben, a legkülönfélébb tudományterületeken alkalmazzák, valamint a paramétereknek sokféle elméleti meghatározási módja is ismeretes és az egyes megoldásoknál a változók átírása jelentős egyszerűsítéseket eredményez. Sajnos a háromparaméteres esetre nincs jól használható (numerikusan kezelhető) módszer.

Mi a következőkben az

$$F_c(x) = \begin{cases} 1 - exp(-x^c), & \text{ha } x \geq 0 \\ 0, & \text{ha } x < 0 \end{cases}$$

jelölést alkalmazzuk a standard Weibull-eloszlás jelölésére. Ebből a lineáris transzformáltak eloszlása

$$F_c(\frac{x-a}{b}).$$

Tehát ez az eloszláscsalád is háromparaméteres, amelyből a c az ún. alakparaméter (típusparaméter). Viszont lényeges, hogy aszimmetrikus eloszlás.

Megjegyzés:

1. Az eloszlás $c = 1$ esetén az exponenciális eloszlást, $c = 2$ a Rayleigh eloszlás adódik, míg $c = 3.57$ közelében az eloszlás közel szimmetrikussá válik és jól közelíti a normális eloszlást. Megfelelő paraméter választással az is elérhető, hogy a Weibull- eloszlás jól közelítse a lognormális és Γ-eloszlásokat.

2. A paraméterek meghatározására számos módszer létezik, de ezek egyrészt nem robusztusak, másrészt nehezen kezelhetőek. A nehezen kezelhetőségre jó példa a momemtum módszer segítségével történő paraméter meghatározás.

Ha ξ a, b, c paraméterű Weibull-eloszlású valószínűségi változó, akkor

$$E\left((\xi-a)^k\right) = b^k \Gamma\left(1+\frac{k}{c}\right),$$

azaz

$$\mu = E(\xi)$$
$$\sigma^2 = E(\xi-\mu)^2$$
$$\alpha_3 = E\left(\left(\frac{\xi-\mu}{\sigma}\right)^3\right).$$

Mint látható, hogy habár μ, σ^2, α_3 egy adott mintából könnyen meghatározható, annak ellenére az egyenletrendszert nem könnyű megoldani, sőt nem is mindig van megoldása $(c < 1)$.

Foglalkozzunk először részesetekkel, azaz feltételezzük , hogy vagy a c vagy az a értéke ismert:

Ha c ismert, akkor az előző egyenletrendszerből az a és b értéke könnyen becsülhető, viszont a becslés nem lesz robusztus, mert a várható érték megközelítésére használt átlag nem robusztus. Ellenben jó becslést ad az eloszlásfüggvényen alapuló robusztus momentumok módszere, hiszen az eloszlásfüggvény könnyen kezelhető, s a típusparaméter ismert. S így máris van robusztus módszerünk pl. az exponenciális eloszlásra, Rayleigh eloszlásra.

Ha az a ismert, akkor a mintánk illetve az eloszlásfüggvény áttranszformálható egy másik típusú eloszlásba (Gumbel-eloszlás) az

$$\eta = \ln(\xi - a)$$

transzformációval, ekkor az eloszlásfüggvény

$$G(y) = 1 - exp\left(-exp\left(\frac{y-T}{s}\right)\right),$$

ahol

$$T = \ln b$$

a helyparaméter és

$$s = \frac{1}{c}$$

a skálaparaméter.

Ezzel nemcsak egyszerűen egy robusztus becslési lehetőséget kaptunk a Weibull-eloszlás skála- illetve alakparaméterére, hanem egyáltalán egy egyszerű lehetőséget az eddig viszonylag bonyolult mószerekkel szemben.

A következő részek a háromparaméteres esetekre vonatkoznak, azaz hogyan készíthető becslés arra az esetre, amikor mind a három paraméter ismeretlen.

MAXIMÁLIS REGRESSZIÓS EGYÜTTHATÓ MÓDSZERE

A Weibull-eloszlásfüggvényt átalakíthatjuk a következő alakra:

$$\ln(-\ln(1 - F(x)) = c\ln(x - a) - c\ln b.$$

Ez új változókat bevezetve

$$Y = BX + A$$

alakú lineáris összefüggés.

Ezen az elven alapszik a Weibull-hálózat (papír) is.

Az a helyparaméter megválasztásával és az $F(x)$ valószínűséget az

$$\frac{i-0.5}{n}$$

"medián rank" összefüggésből meghatározva a egyenlet transzformációi segítségével lineáris regresszióval határozható meg a Weibull-eloszlás másik két paramétere. A mérési pontsorozat legjobb kiegyenlítését a három paraméter esetében nem lehet meghatározni. Az a helyparaméter felvételével, megváltoztatásával lehet a legjobban illeszkedő helyzethez eljutni. Az illeszkedés vizsgálatánál több más lehetőség közül a korrelációs együttható az egyik legalkalmasabb. A feladat ebben az esetben tehát úgy változtatni a helyparamétert, hogy a korrelációs együttható maximális legyen. A helyparaméter függvényében a korrelációs együttható egy konkáv függvény, amelynek maximuma egyértelmű.

Full Screen
Search
Close

A ROBUSZTUS MÓDSZER

Hasonló a maximális regressziós együttható módszeréhez, azonban itt felhasználva a b és c paraméterekre a robusztus becslési módszert az ω^2 statisztika értékét minimalizáljuk.

Full Screen
Search
Close

10 ω^2-próba

Legyen a $\xi_1, \xi_2, \ldots, \xi_n$ mintához tartozó rendezett minta ${\xi_1}^\star, {\xi_2}^\star, \ldots, {\xi_n}^\star$. Továbbá legyen ${\xi_0}^\star = -\infty$ és ${\xi_{n+1}}^\star = +\infty$, ekkor az empirikus eloszlásfüggvény

$$F_n{}^\star(x) = \frac{k}{n}, \quad \text{ha} \quad {\xi_k}^\star < x < {\xi_{k+1}}^\star,$$

ahol $k = 0, 1, \ldots, n$.

Az $F_n{}^\star(x)$-nek a feltételezett $F(x)$-től vett eltérése mértékéül az

$$\omega_n{}^2 = n \int_{-\infty}^{+\infty} (F(x) - F_n{}^\star(x))^2 dF(x)$$

mennyiséget használják a következő Mises-Szmirnov tétel alapján.

Tétel:
Tetszőleges $F(x)$ folytonos eloszlásfüggvényű ξ valószínűségi változóra, minden $x > 0$ esetén igaz, hogy

$$\lim_{n\to\infty} P({\omega_n}^2 < x) = a_1(x),$$

ahol az $a_1(x)$ függvény nem függ ξ-től.

Megjegyzés:
1. Az ${\omega_n}^2$ statisztikát a következőképpen lehet meghatározni az $F(x)$ eloszlásfüggvény és a minta segítségével:

$${\omega_n}^2 = \frac{1}{12n} + \sum_{k=1}^{n} \left(F({\xi_k}^\star) - \frac{k-0.5}{n} \right)^2 .$$

2. Az $a_1(x)$ eloszlásfüggvénynek csak a karakterisztikus függvénye adható meg közvetlenül használható formában,

ezért a próba alkalmazásához szükséges a következő táblázat:

x	0	0.12	0.15	0.18	0.24	0.35	0.46
$a_1(x)$	0	0.50	0.60	0.70	0.80	0.90	0.95

x	0.74	0.87	1.17	1.49
$a_1(x)$	0.99	0.995	0.999	0.9998

3. *Az ω^2-próba:* Rögzítünk egy p megbízhatósági szintet és az

$$a_1(x_p) = p$$

egyenletből kiszámítjuk a megfelelő x_p kvantilis értékét illetve az előző segítségével ellenőrizzük, hogy a feltételezett $F(x)$ eloszlásfüggvény és az $F_n^\star(x)$ empirikus eloszlásfüggvényből kiszámított ω_n^2 milyen x_p-hez viszonyítva. Ha $\omega_n^2 > x_p$, akkor a feltevésünk nem fogadható el.

11 Kolmogorov-próba

A Kolmogorov-próba az előző ω^2-próbához hasonlóan az empirikus és az elméleti eloszlás eltérését vizsgálja. Felhasználva az előző szakasz jelöléseit a módszer a következő: határozzuk meg a

$$D = \sup_x |F_n{}^\star(x) - F(x)|$$

értékét. A D eloszlása az előzőhöz hasonlóan független F-től, így a D értéke alapján az illeszkedésvizsgálat könnyen eldönthető.

Köszönöm

a figyelmet!

References

[1] A. C. Allen: *Probability, Statistics and Queueing Theory, With Computer Applications*, Academic Press, New York, 2003. ISBN-13: 978-0120510504

[2] Deák I.: *Véletlenszámgenerátorok és alkalmazásaik*, Akadémiai Kiadó, Budapest, 1986.

[FR11] Fegyverneki Sándor, Raisz Péter: *Sztochasztikus modellezés*,elektronikus jegyzet, 2011, TÁMOP 4.1.2-08/1/A-2009-0001 project, https://www.uni-miskolc.hu/˜ matfs/

[FS11] Fegyverneki Sándor: *Valószínűség-számítás és matematikai statisztika*, elektronikus jegyzet, Kempelen Farkas elktronikus könyvtár, 2011,

Full Screen
Search
Close

TÁMOP 4.1.2-08/1/A-2009-0001 project, https://www.uni-miskolc.hu/~ matfs/

[FE78] W. Feller: *Bevezetés a valószínűségszámításba és alkalmazásaiba,* Műszaki Könyvkiadó, Budapest, 1978.

[3] I.M. Szobol: *A Monte-Carlo módszerek alapjai,* Műszaki Könyvkiadó, Budapest, 1981.