ISMÉTELT MÉRÉSES MODELLEK R-KÖRNYEZETBEN

Virág Katalin

Szegedi Tudományegyetem

Általános Orvostudományi Kar, Orvosi Fizikai és Orvosi Informatikai Intézet

A kutatás a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosító számú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt keretében zajlott. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

2014. augusztus 21.

BEVEZETÉS

Orvosi kutatások során gyakran előfordul, hogy a vizsgált változót ugyanazon a páciensen többször is megmérik időben vagy változó kísérleti körülmények között. Az így keletkezett adatok összetartozó mintákat eredményeznek, és az adatokat ismételt méréses adatoknak nevezzük. Ezen cikk az ilyen jellegű adatok statisztikai kiértékeléséhez nyújt segítséget. Konkrét példákon keresztül ismertetjük a legegyszerűbb ismételt méréses modelleket, a módszerek alkalmazhatóságának feltételeit, a statisztikai szoftver megfelelő paraméterezését, az eredmények értelmezését. *A statisztikai elemzéseket a mindenki számára ingyenesen elérhető R-környezetben végezzük, és a felhasznált kódokat mellékeljük.*

ISMÉTELT MÉRÉSES MODELLEK

Az ismételt méréses varianciaanalízis modellek alkalmazhatóságának feltétele a normalitás és a **szfericitás** *(sphericity)*. A szfericitási (cirkularitási) feltétel a kovarianciamátrix struktúrájára vonatkozó feltétel, mely szerint az összes lehetséges módon képezett különbségváltozók elméleti szórásainak azonosnak kell lennie. Biztosan teljesül, ha a függő változók elméleti szórásai megegyeznek, továbbá a páronkénti korrelációk is ugyanakkorák. A szfericitási feltétel teljesülése a **Mauchly-teszt**tel ellenőrizhető. Nem teljesülése esetére különböző korrekciós formulákat dolgoztak ki, melyek a varianciaanalízis *F*-értékének szabadságfokát lecsökkentik.

Ezen cikk az ismételt méréses modellek egy másfajta megközelítését mutatja be, az úgynevezett **kovariancia alakzat modellek**et *(covariance pattern models)* ismerteti. Az adatokra különböző kovarianciastruktúrákat illesztünk, majd a vizsgált modelleket likelihood-hányados próbák segítségével hasonlítjuk össze, és kiválasztjuk a legegyszerűbb modellt, mely jól illeszkedik az adatokhoz.

1. Egyszempontos ismételt méréses modellek

A legegyszerűbb ismételt méréses elrendezés, amikor csak az idő hatását vizsgáljuk.

1.1. Adatok

Ezt a modellt generált adatokon mutatjuk be. Egy folytonos változót vizsgálunk, melyet minden egyeden három alkalommal mértek meg. Azt szeretnénk megtudni, hogy van-e szignifikáns különbség az egyes időpontokban történt mérések átlagai között.

Az adatbázist „hosszanti" formában kell elrendezni:

- egyetlen folytonos függő változó (példánkban a „*meres*" nevű változó),
- két kategorikus változó:
    - a mérés időpontjának azonosítására („*ido*" nevű változó),
    - az egyed azonosítására („*azonosito*" nevű változó).

Az adatbázis első hét sora:

AZONOSITO	IDO	MERES
1	1	12.401225
1	2	10.964604
1	3	6.598895
2	1	9.708189
2	2	10.483301
2	3	10.8851
3	1	6.939121

1.2. A modellezés folyamata R-ben

 # A szükséges csomagok:

```
library(nlme)
library(lsmeans)
```

 # Kontrasztok beállítása

```
options(contrasts = c("contr.sum", "contr.poly"))
```

 # Adatok beolvasása, kategorikus változók faktorrá alakítása

```
adat1 <- read.csv2("adat1.csv")
adat1 <- within(adat1, {
  azonosito <- factor(azonosito)
  ido <- factor(ido)
})
```

 # Exploratív elemzések

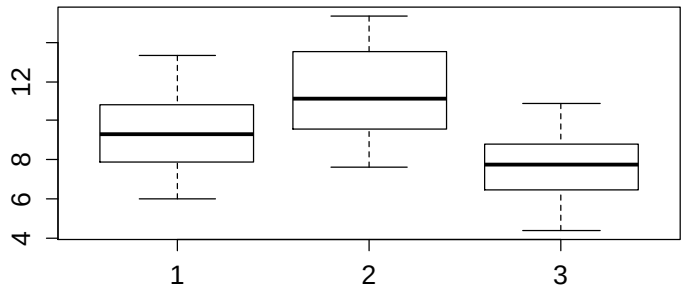
 # Adatok leíró statisztikái

```
summary(adat1)
```

```
   azonosito  ido        meres
 1       : 3   1:30   Min.   : 4.375
 2       : 3   2:30   1st Qu.: 7.670
 3       : 3   3:30   Median : 9.282
 4       : 3          Mean   : 9.463
 5       : 3          3rd Qu.:10.983
 6       : 3          Max.   :15.364
 (Other):72
```

 # Doboz ábra

```
boxplot(adat1$meres ~ adat1$ido)
```



 # Különböző kovarianciastruktúrák illesztése

 # Összetett szimmetria (Compound Symmetry) kovarianciamátrix

```
m1.cs <- lme(meres ~ ido, random = ~ 1 | azonosito, data = adat1, corr = corCompSymm(, form = ~ 1 | azonosito))
```

 # Elsőrendű autoregresszív (First-Order Autoregressive) kovarianciamátrix

```
m1.ar1 <- lme(meres ~ ido, random = ~ 1 | azonosito, data = adat1, corr = corAR1(, form = ~ 1 | azonosito))
```

 # Általános (Unstructured) kovarianciamátrix

```
m1.un <- lme(meres ~ ido, random = ~ 1 | azonosito, data = adat1, corr = corSymm(form = ~ 1 | azonosito), weights = varIdent(form = ~ 1 | ido))
```

 # A modellek összehasonlítása

 # Összetett szimmetria és elsőrendű autoregresszív modellek összehasonlítása

```
anova(m1.cs, m1.ar1)
```

```
       Model df      AIC      BIC    logLik
m1.cs      1  6 387.6723 402.4677 -187.8361
m1.ar1     2  6 387.6655 402.4610 -187.8328
```

Az összetett szimmetria és az elsőrendű autoregresszív kovarianciastruktúrájú modellek összehasonlítása esetén azt a modellt választjuk, amelyiknél az Akaike-féle információs kritérium (AIC) értéke alacsonyabb, jelen esetben az elsőrendű autoregresszív struktúrát *(m1.ar1)*. A következő lépésben a most kiválasztott *m1.ar1* modellt hasonlítjuk az általános struktúrához.

 # Elsőrendű autoregresszív és általános modellek összehasonlítása

```
anova(m1.ar1, m1.un)
```

```
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1.ar1     1  6 387.6655 402.4610 -187.8328
m1.un      2 10 392.0473 416.7064 -186.0237 1 vs 2 3.618205  0.4601
```

Az elsőrendű autoregresszív és az általános kovarianciastruktúrájú modellek összehasonlítására elvégzett likelihood-hányados próba eredménye 5%-os szinten nem szignifikáns ($p = 0{,}4601$), tehát az általános kovarianciastruktúra nem vezetett szignifikánsan jobb illeszkedéshez. Így a szűkebb, elsőrendű autoregresszív modellt *(m1.ar1)* választjuk.

 # Az idő hatás szignifikancia vizsgálata az m1.ar1 modell esetén

```
anova(m1.ar1, type = "marginal")
```

```
            numDF denDF   F-value p-value
(Intercept)     1    58 1703.1536  <.0001
ido             2    58   31.4185  <.0001
```

Az idő hatás 5%-os szinten szignifikáns ($p < 0{,}0001$), vagyis a három időpontban történt mérések átlagai közül legalább egy szignifikánsan eltér valamelyik másik időpont átlagától.

 # Páronkénti összehasonlítások az m1.ar1 modell esetén

```
lsmeans(m1.ar1, pairwise ~ ido)
```

```
$lsmeans
 ido    lsmean        SE df asymp.LCL asymp.UCL
 1    9.285104 0.3585442 NA  8.582285  9.987923
 2   11.428566 0.3585442 NA 10.725747 12.131385
 3    7.674316 0.3585442 NA  6.971497  8.377135

Confidence level used: 0.95

$contrasts
 contrast  estimate        SE df   z.ratio p.value
 1 - 2    -2.143462 0.4756893 NA -4.506012  <.0001
 1 - 3     1.610788 0.4808919 NA  3.349585  0.0023
 2 - 3     3.754250 0.4756893 NA  7.892231  <.0001

P value adjustment: tukey method for a family of 3 means
```

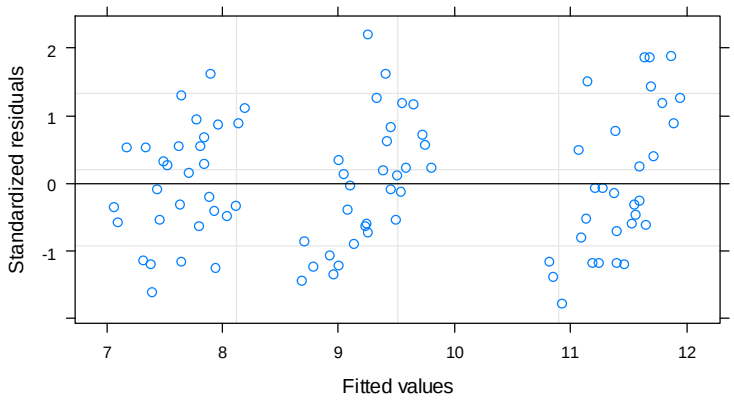
```
P values are asymptotic
```

A páronkénti hasonlítások során azt kaptuk, hogy az első időpontban történt mérés átlaga 5%-os szinten szignifikánsan eltér a második időpont átlagától ($p < 0{,}0001$) és a harmadik időpont átlagától is ($p = 0{,}0023$), valamint a második és a harmadik mérések átlagai közötti különbség is szignifikáns ($p < 0{,}0001$).

 # Az illesztett modell diagnosztikája

 # Standardizált reziduumok vs. becsült értékek

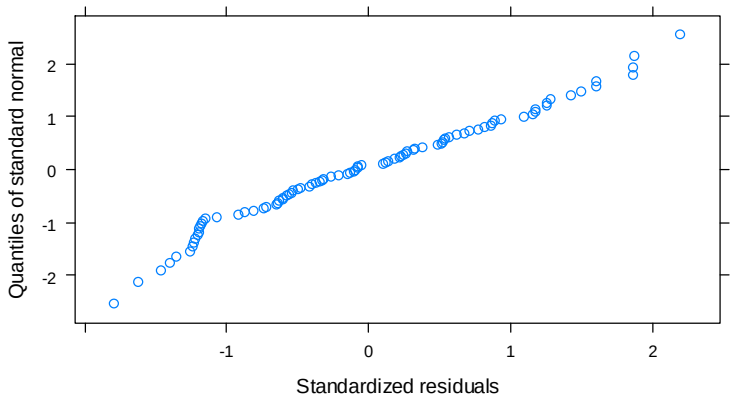
```
plot(m1.ar1)
```



A modell megfelelő illeszkedése esetén a pontok a konstans nulla körüli véletlen eltérések. Jelen esetben a modell megfelelő. Ezen az ábrán a kiugró értékeket is vizsgálhatjuk, jelenleg nincsenek.

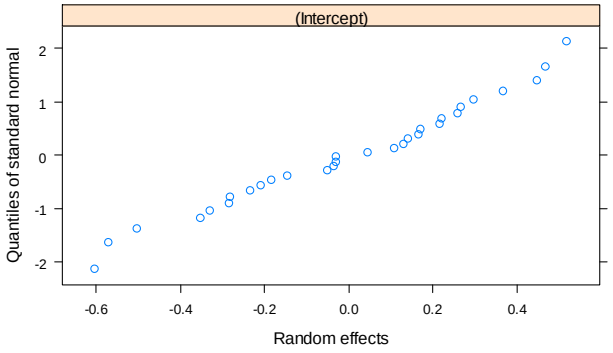
 # A hibatag normalitásának ellenőrzése

```
qqnorm(m1.ar1)
```



Normális eloszlás esetén a pontok egy egyenes közelében helyezkednek el. Jelen esetben ez teljesül.

 # A random hatás normalitásának ellenőrzése



Ez a feltétel is teljesül, a pontok egy egyenes közelében helyezkednek el.

2. Kétszempontos ismételt méréses modell

Egy csoportosító változót is bevonunk a modellbe, az idő hatáson felül azt is vizsgáljuk, hogy a csoportátlagok között van-e szignifikáns eltérés, valamint a *csoport*idő* interakciót is, vagyis azt, hogy a csoportátlagok közötti különbség függ-e attól, hogy a vizsgált változót melyik időpontban mérték.

2.1. Adatok

Prof. Dr. Hantos Zoltán vezetésével zajlik újszülöttek légzésmechanikájának követéses vizsgálata, melynek során a születést követő három nap során mérnek légzésfunkciós értékeket. Az újszülötteket a születés módja alapján két csoportra osztottuk. Vizsgáltuk a légzésfunkciós paraméterek (példánkban a légző rendszer rezisztenciájának *(R)*) időbeli változását, valamint összehasonlítottuk a császármetszéssel és hüvelyen keresztül világra jött újszülöttek különböző napokon mért légzésfunkciós értékeit.

2.2. A modellezés folyamata R-ben

 # A szükséges csomagok:

```
library(nlme)
library(lsmeans)
```

 # Kontrasztok beállítása

```
options(contrasts = c("contr.sum", "contr.poly"))
```

 # Adatok beolvasása, kategorikus változók faktorrá alakítása

```
adat2 <- read.csv2("spirometria.csv")
adat2 <- within(adat2, {
```

```
  azonosito <- factor(azonosito)
  csoport <- factor(csoport, levels = 1:2, labels = c("PVN",
"SC"))
  nap <- factor(nap)
})
```

 # Exploratív elemzések

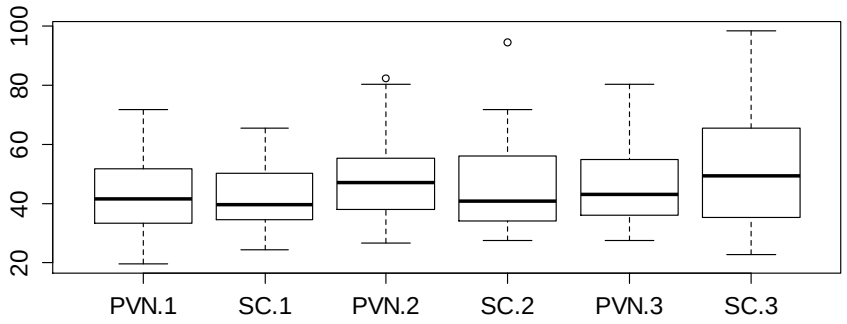
 # Adatok leíró statisztikái

```
summary(adat2)
```

```
   azonosito    csoport   nap          R
 1       :  3   PVN:93    1:61   Min.   :19.60
 2       :  3   SC:90     2:61   1st Qu.:34.22
 3       :  3             3:61   Median :43.80
 4       :  3                    Mean   :46.51
 5       :  3                    3rd Qu.:54.21
 6       :  3                    Max.   :98.23
 (Other):165
```

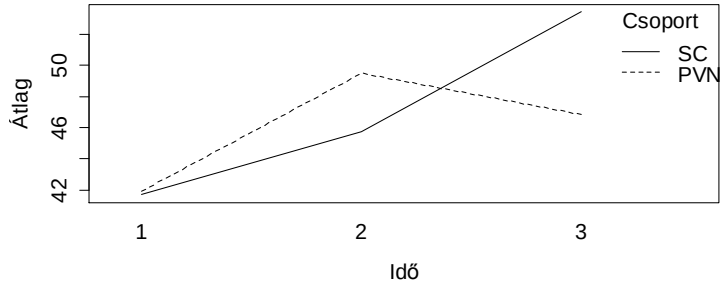
 # Doboz ábra

```
boxplot(adat2$R ~ adat2$csoport*adat2$nap)
```



 # Interakciós ábra

```
interaction.plot(adat2$nap, adat2$csoport, adat2$R, ylab =
"Átlag", xlab = "Idő", trace.label = "Csoport")
```



 # Különböző kovarianciastruktúrák illesztése

 # Összetett szimmetria (Compound Symmetry) kovarianciamátrix

```
m2.cs <- lme(R ~ csoport*nap, random = ~ 1 | azonosito, data = adat2, corr = corCompSymm(, form = ~ 1 | azonosito))
```

 # Elsőrendű autoregresszív (First-Order Autoregressive) kovarianciamátrix

```
m2.ar1 <- lme(R ~ csoport*nap, random = ~ 1 | azonosito, data = adat2, corr = corAR1(, form = ~ 1 | azonosito))
```

 # Elsőrendű autoregresszív (First-Order Autoregressive) kovarianciamátrix heterogén varianciákkal

```
m2.arh1 <- lme(R ~ csoport*nap, random = ~ 1 | azonosito, data = adat2, corr = corAR1(, form = ~ 1 | azonosito), weight = varIdent(form = ~ 1 | nap))
```

 # Általános (Unstructured) kovarianciamátrix

```
m2.un <- lme(R ~ csoport*nap, random = ~ 1 | azonosito, data = adat2, corr = corSymm(form = ~ 1 | azonosito), weights = varIdent(form = ~ 1 | nap))
```

 # A modellek összehasonlítása

 # Összetett szimmetria és elsőrendű autoregresszív modellek összehasonlítása

```
anova(m2.cs, m2.ar1)
```

```
       Model df      AIC      BIC    logLik
m2.cs      1  9 1475.929 1504.514 -728.9643
m2.ar1     2  9 1473.720 1502.305 -727.8599
```

Az összetett szimmetria és az elsőrendű autoregresszív kovarianciastruktúrájú modellek összehasonlítása során a kisebb AIC értékkel rendelkező *m2.ar1* modellt választjuk, majd ezt a heterogén varianciájú elsőrendű autoregresszív modellhez hasonlítjuk.

 # Elsőrendű autoregresszív és heterogén varianciájú elsőrendű autoregresszív modellek összehasonlítása

```
anova(m2.ar1, m2.arh1)
```

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m2.ar1      1  9 1473.720 1502.305 -727.8599
m2.arh1     2 11 1469.969 1504.907 -723.9847 1 vs 2 7.750341  0.0208
```

A likelihood-hányados próba eredménye 5%-os szinten szignifikáns (p = 0,0208), tehát a heterogén varianciájú elsőrendű autoregresszív kovarianciastruktúra

szignifikánsan jobban illeszkedik az adatokhoz. A következő lépésben ezt az *m2.arh1* modellt és az általános kovarianciastruktúrájú modellt hasonlítjuk össze.

```
# Heterogén varianciájú elsőrendű autoregresszív és általános modellek összehasonlítása
anova(m2.arh1, m2.un)
```

```
        Model df      AIC      BIC    logLik   Test   L.Ratio p-value
m2.arh1     1 11 1469.969 1504.907 -723.9847
m2.un       2 13 1473.055 1514.345 -723.5276 1 vs 2 0.9140965  0.6331
```

A likelihood-hányados próba eredménye 5%-os szignifikancia szinten nem szignifikáns (p = 0,6331), így a szűkebb *m2.arh1* modellt választjuk.

```
# Fix hatások (csoport és nap) szignifikanciavizsgálata az m2.arh1 modell esetén
anova(m2.arh1, type = "marginal")
```

```
            numDF denDF  F-value p-value
(Intercept)     1   118 857.4580  <.0001
csoport         1    59   0.0734  0.7873
nap             2   118   9.0844  0.0002
csoport:nap     2   118   3.0981  0.0488
```

A csoport*nap interakció 5%-os szinten szignifikáns (p = 0,0488), tehát az egyes napok átlagai közötti eltérések függnek attól, hogy az újszülött császármetszéssel vagy hüvelyi úton született.

```
# Páronkénti hasonlítások az m2.arh1 modell esetén
lsmeans(m2.arh1, pairwise ~ nap*csoport)
```

```
$lsmeans
 nap csoport   lsmean       SE df asymp.LCL asymp.UCL
 1   PVN     41.94128 2.259608 NA  37.51199  46.37057
 2   PVN     49.46326 2.852497 NA  43.87179  55.05473
 3   PVN     46.84985 3.069079 NA  40.83384  52.86586
 1   SC      41.70628 2.296959 NA  37.20378  46.20879
 2   SC      45.74083 2.899648 NA  40.05694  51.42473
 3   SC      53.39007 3.119811 NA  47.27461  59.50552

Confidence level used: 0.95

$contrasts
 contrast        estimate       SE df     z.ratio p.value
 1,PVN - 2,PVN  -7.521978 2.423193 NA -3.10415923  0.0235
 1,PVN - 3,PVN  -4.908570 3.010866 NA -1.63028508  0.5784
 1,PVN - 1,SC    0.234997 3.222087 NA  0.07293317  1.0000
 1,PVN - 2,SC   -3.799553 3.676110 NA -1.03357969  0.9067
 1,PVN - 3,SC  -11.448786 3.852149 NA -2.97205214  0.0352
 2,PVN - 3,PVN   2.613408 2.893524 NA  0.90319209  0.9458
 2,PVN - 1,SC    7.756975 3.662343 NA  2.11803623  0.2779
 2,PVN - 2,SC    3.722425 4.067517 NA  0.91515904  0.9428
 2,PVN - 3,SC   -3.926808 4.227287 NA -0.92891912  0.9392
 3,PVN - 1,SC    5.143567 3.833441 NA  1.34176249  0.7616
 3,PVN - 2,SC    1.109017 4.222228 NA  0.26266155  0.9998
 3,PVN - 3,SC   -6.540216 4.376353 NA -1.49444428  0.6679
 1,SC - 2,SC    -4.034550 2.463249 NA -1.63789781  0.5733
 1,SC - 3,SC   -11.683783 3.060636 NA -3.81743674  0.0019
 2,SC - 3,SC    -7.649233 2.941354 NA -2.60058194  0.0971
```

```
P value adjustment: tukey method for a family of 6 means
P values are asymptotic
```

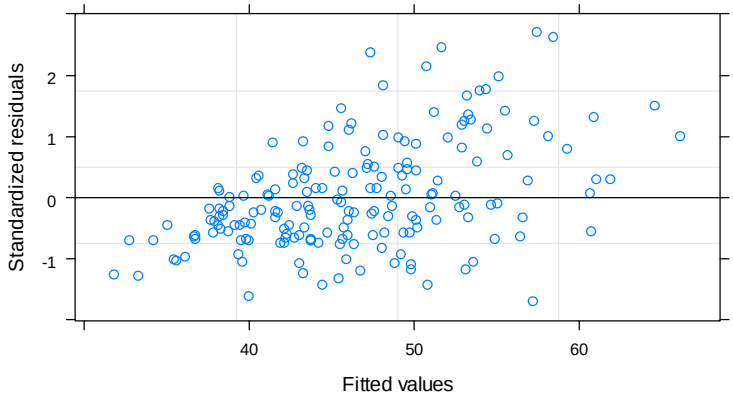
A hüvelyi úton született csoport esetén az első és a második nap átlaga különbözik szignifikánsan (p = 0,0235), a császármetszéses csoport esetén pedig az első és a harmadik nap átlaga (p = 0,0019).

Egy adott napon belül a két csoport között nem mutatható ki szignifikáns eltérés.

 # Az illesztett modell diagnosztikája

 # Standardizált reziduumok vs. becsült értékek

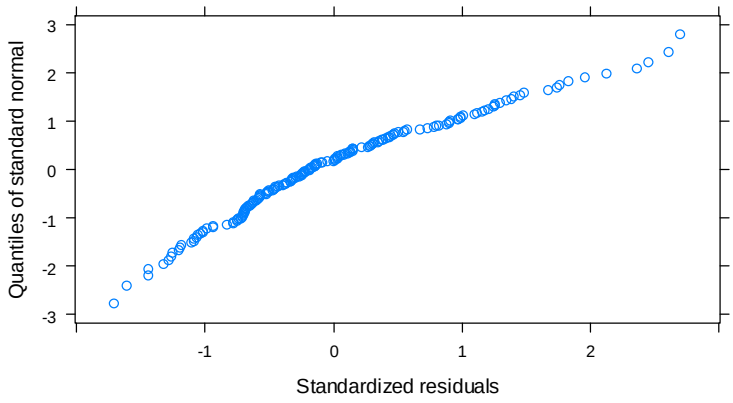
```
plot(m2.arh1)
```



Ezen az ábrán a modellilleszkedést és a kiugró értékek jelenlétét vizsgálhatjuk. A modell megfelelőnek tűnik, kiugró értékek nincsenek.

 # A hibatag normalitásának ellenőrzése

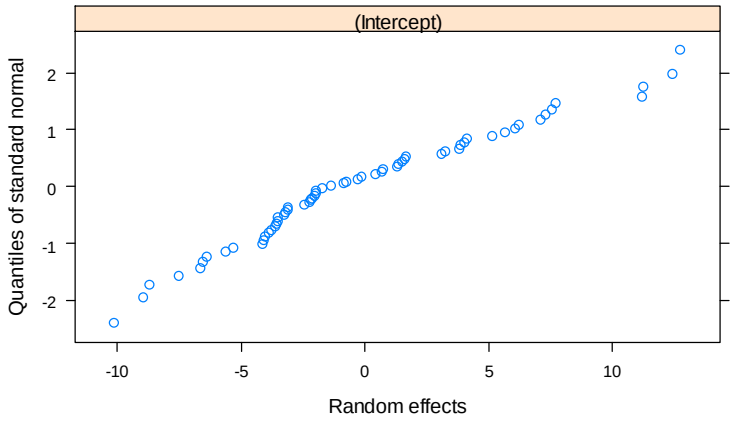
```
qqnorm(m2.arh1)
```



A pontok megközelítőleg egy egyenes mentén helyezkednek el, a hibatag normalitási feltétele teljesül.

 # A random hatás normalitásának ellenőrzése

```
qqnorm(m2.arh1, ~ ranef(.))
```



A pontok megközelítőleg egyenes mentén helyezkednek el, a véletlen hatás normalitási feltétele is teljesül.

HIVATKOZÁSOK

[1] P. Armitage (Editor), T. Colton (Editor). Encyclopedia of Biostatistics: 8-Volume Set, Second Edition. Hoboken, New Jersey, John Wiley & Sons Inc., 2005.

[2] H. Brown, R. Prescott. Applied Mixed Models in Medicine, Second Edition. John Wiley & Sons Ltd., Chichester, 2006.

[3] Data Analysis Examples. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/dae/ (accessed August 08, 2014).

[4] R 3.0.3: A Language and Environment for Statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria)

[5] A. Vargha. Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal. Budapest, Pólya Kiadó, 2000.