CLMM: Cumulative Link Mixed Models

Ordinális skálákra alkalmazható modellek

Az ordinális skálák jellemzői:

- az adatok sorrendezhetők,
- az egyes adatpontok közötti távolságok nem feltétlenül azonosak,
- nem folytonosak (nem vehetnek fel akármilyen számértéket),
- lehet rájuk mediánt és interkvartilis (fél)tartományt számolni, de átlagot és szórást nem.

Ezért az átlagon (és szóráson) alapuló t-próba és a lineáris regresszión alapuló varianciaanalízis és a lineáris kevert modellek nem alkalmazhatóak az ordinális adatokra.
Ekvivalens nemparametrikus próbák: Wilcoxon-próba (egymintás és páros t-próba), Mann-Whitney próba (kétmintás független t-próba), Kruskal-Wallis és Friedman-próba (független mintás és ismételt méréses ANOVA).

Likert-skála

Likert-skála: sorrendezett, pontokhoz kötött megítélési fokok.
Fajtái:

- minden egyes ítélet meg van szövegezve, pl.: *1: egyáltalán nem értek egyet, 2: nem tudok vagy nem akarok válaszolni, 3: teljesen egyetértek*,
- a skála két végpontja van meghatározva, a közte levő értékek megítélése a résztvevőre van bízva,
- a skála középpontja 0 (semleges válasz), a két végpont negatív és pozitív,
- a skála állhat páros vagy páratlan számú pontból. Páratlan előnye: a remélt medián egybeesik egy valóban létező pontszámmal. Páros előnye: a résztvevő nem nyomkodhatja mindig a középsőt, ha nem tud vagy nem akar pontozni.

Likert-skála megítélése

Bár víziló, de nem igazi ló – legyen a ló az ordinális skála szimbóluma.

Vagyis: elvben ordinális, mert diszkrét, nem vehet fel akármilyen értéket. Egy ötpontos skálán (1: minimum, 5: maximum) nem adhatunk $\sqrt{2}$ pontot egy kérdésre.

De: ha a pontszámok ekvidisztánsak, vagyis az 1–2 és a 4–5 pont között (vagy bármely más pontszám között) azonosnak tekinthetjük a távolságot, akkor el lehet fogadni parametrikus skálaként. Ebben az esetben az 1,41 pontos átlagot is tudnunk kell értelmezni (nem teljesen rossz egy megítélendő mondat, de nagyon rossz).

Javaslat: ha a Likert-skála megfelel a fenti követelményeknek, használjunk parametrikus modelleket, mert pontosabb becslést adnak.

Forrás: Sullivan, Gail M. and Artino, Anthony R. 2013: Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education* 5 (4), 541–542.

Reziduálisokkal szembeni elvárások LMM esetén

Hasznos olvasnivaló: Bross, Fabian 2019: Using mixed effect models to analyze acceptability rating data. `fabianbross.de/mixedmodels.pdf`

Lineáris kevert modellek reziduálisainak előfeltételei:

- lineáris eloszlás (nagyjából illeszthetőek legyenek egy egyenesre): `qqnorm(resid(modellem))`,
- homoszkedaszticitás, vagyis a variancia ne különbözzön az alacsonyabb és magasabb x-értékek mentén: `plot(modellem)`. Elvárás: az eloszlás ne mutasson semmilyen mintázatot, egyenletes felhőként jelenjen meg a modellezett regressziós egyenes körül.

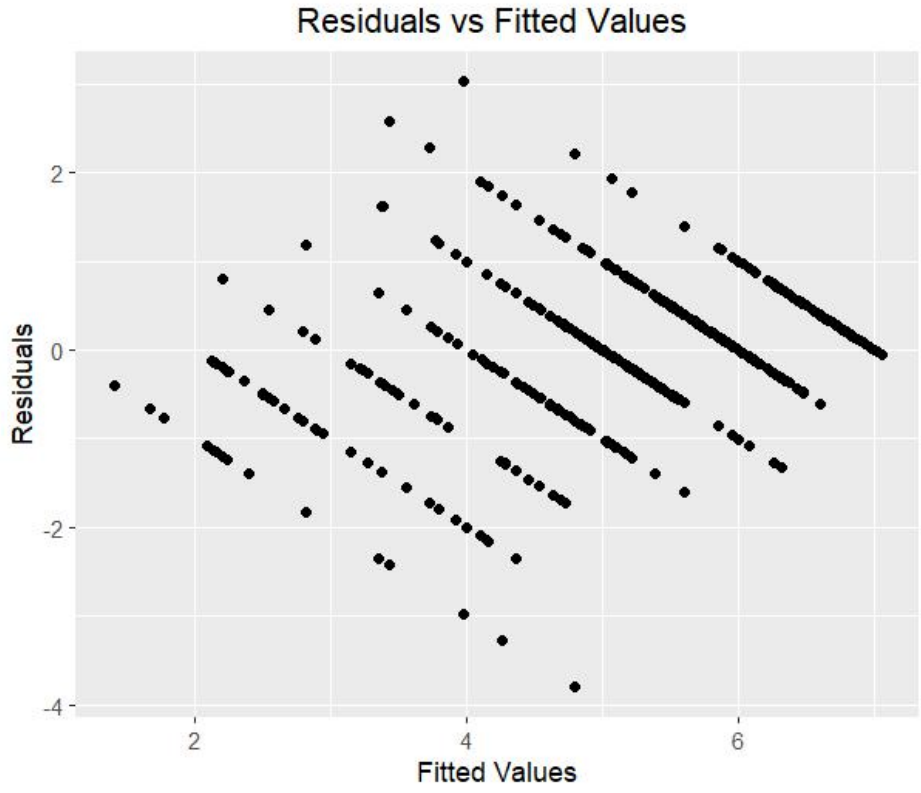
Mintázatok, amik nem mutatnak lineáris eloszlást:

`https://statsnotebook.io/blog/analysis/linearity_homoscedasticity/`

Lehetséges problémák:

- növekvő értéknél növekvő variancia a reziduálisokban,
- növekvő értéknél csökkenő variancia a reziduálisokban,
- valamilyen mintázat, pl. parabola, csíkok, csoportosulás stb.

Példa csíkokra:



Probléma: a diszkrét skála értékei természetüknél fogva csíkokba rendeződnek.

Forrás:
https://stats.stackexchange.com/questions/371805/understanding-residual-plots-and-how-to-transform-my-data-to-not-violate-the-nor

Cumulative Link (Mixed) Models

Ha a skála nem tekinthető parametrikusnak, vagy nem vagyunk benne biztosak, használjunk Cumulative Link (Mixed) Modelt. Tehát lehet személyenként egy adatunk (CLM) vagy több, azaz ismételt méréses dizájnunk (CLMM).

Kumulatív gyakoriság: sorrendbe állított értékeknél a *legalább* egy adott kategóriához tartozó értékek gyakorisága. Pl.: „hány ember olvasott el *legalább x* könyvet egy hónapban?

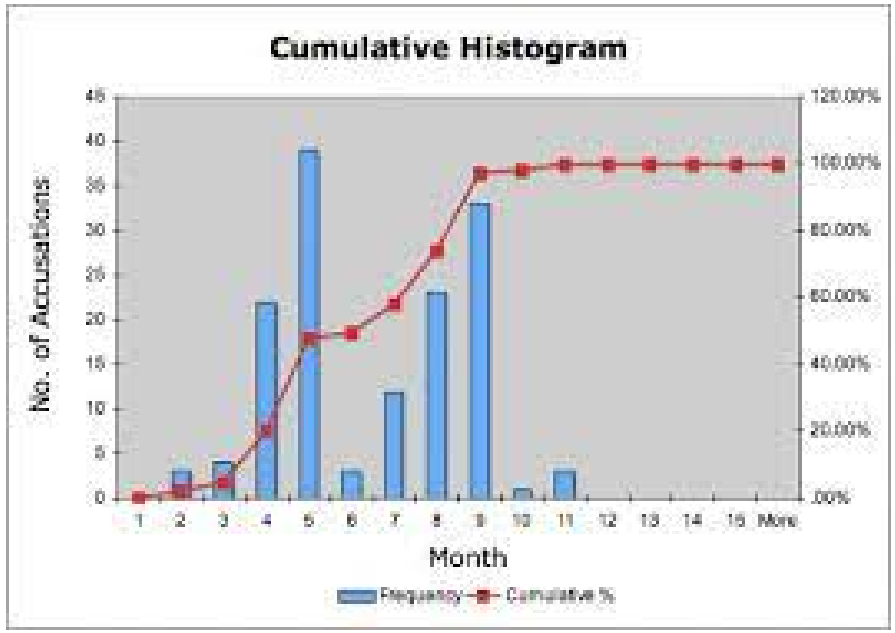
Number of books read in a month	Frequency	Cumulative Frequency
2	1	**1**
3	3	1 + 3 = **4**
4	5	4 + 5 = **9**
5	2	9 + 2 = **11**
6	1	11 + 1 = **12**

Forrás: https://study.com/learn/lesson/cumulative-frequency-table.html

Kumulatív relatív gyakoriság

Megadható százalékban vagy egy 0 és 1 közötti értékként.

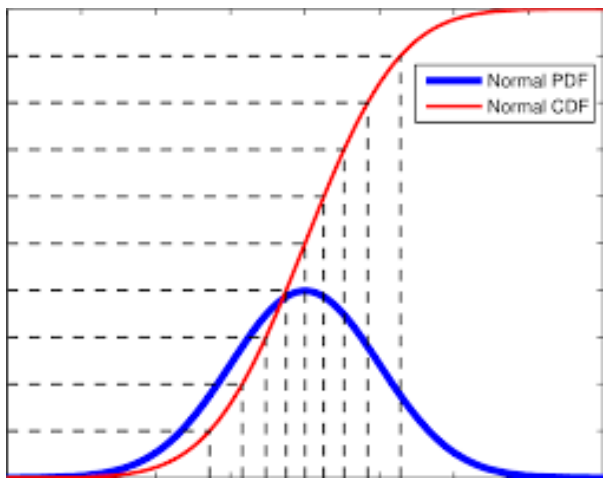
Modellezés:



Ideális esetben a gyakoriság monoton növekvő. Ha stagnál, mert kimaradnak értékek, össze lehet vonni osztályokat.

Forrás: https://www2.tulane.edu/∼salem/Cumulative%20Histogram.html

Feltételezés: létezik egy **látens folytonos** y változó, ami az egyes osztályok közötti értékeket is felveheti (pl. 5 pontos skálán 3,14-et).
A cumulative link model nem vár lineáris összefüggést, mint a lineáris regresszión alapuló modellek. Helyette: a sűrűségfüggvénnyel ábrázolt eloszlásból kiindulva számolja ki az eloszlásfüggvényt, vagyis ezzel modellezi a Likert-skála értékeinek gyakoriságát.
Normális eloszlás esetén a sűrűségfüggvény és az eloszlásfüggvény viszonya:



Ez a szigmoidgörbe vagy S-görbe, a normál eloszlás sűrűségfüggvényének integráltja.

Jellemzők

CLM és CLMM futtatása: `ordinal` csomagban a `clm` vagy `clmm()` függvénnyel. A szintaxis megegyezik az `lmer()` függvényével.

FONTOS: a függő változónak faktornak kell lennie, nem lehet numerikus (ellenőrizni kell a `class()` függvénnyel).

Hátrányok LMM-hez képest: (1) sokáig tart, míg lefut, (2) körülményesebb hatásnagyságot, azaz r^2-et számolni, (3) kevésbé ismert, kevesebb irodalom.

Paraméterállítási lehetőségek: `threshold = c("flexible", "symmetric", "symmetric2", "equidistant")`. Ha tudjuk, hogy a skálánk ekvidisztáns, ezt állítsuk be. Default: flexible.

Példa: borok keserűsége

A csomagba beépített `wine` adatsor fehérborok keserűségérzetére adott pontszámokat tartalmaz 1 és 5 között 9 kóstoló személytől.

Fix hatások: préselési hőmérséklet és kontaktus a héjjal, random intercept: kóstoló személy.

Először futtassunk lineáris regressziós modellt, hogy a reziduálisokat meg tudjuk szemlélni:

```
h.lm = lm(rating~temp+contact, data = wine)
plot(h.lm)
```

Futtassunk Cumulative Link Modellt random hatások nélkül:

```
h = clm(rating~temp+contact, data = wine)
summary(h)
```

Ellenőrizzük, hogy a hőfok és a héjon áztatás ideje interakciót mutat-e:

```
h = clm(rating~temp*contact, data = wine)
```

Hasonlítsuk össze a két modellt az `anova()` függvénnyel.

A p-értékek mellé megkapjuk a skála pontszámai közötti küszöbértékeket. Ezek a **mögöttes folytonos** y értékeinek kategóriákat elválasztó küszöbértékei.

Mivel szigmoidgörbére modelleztük az eloszlást, ami negatív értékeket is felvehet, a modell küszöbértéke lehet negatív és a maximális pontjánál nagyobb is.

Modellszelekció, post-hoc teszt

Most állítsunk kevert modellt a kóstoló személyt random hatásként figyelembe véve, random slope, majd random intercept modellel. Ezeket szintén vessük össze az `anova()` függvénnyel. Válasszuk ki a maximálisan szükséges modellt.

Random hatás szükségességének tesztelése: CLMM és CLM modellkimenetének összehasonlítása.

```
h2 = clm(rating~temp+contact, data = wine)
```

Post-hoc teszt az `emmeans` csomag telepítése után:

```
emmeans(h,specs = pairwise~temp+contact,adjust="tukey")
```

pseudo R-squared null-modellel összehasonlítva az `rcompanion` csomag `nagelkerke()` függvényével

```
h.null = clmm(rating~1+(1|judge), data = wine)
nagelkerke(fit = h, null = h.null)
```