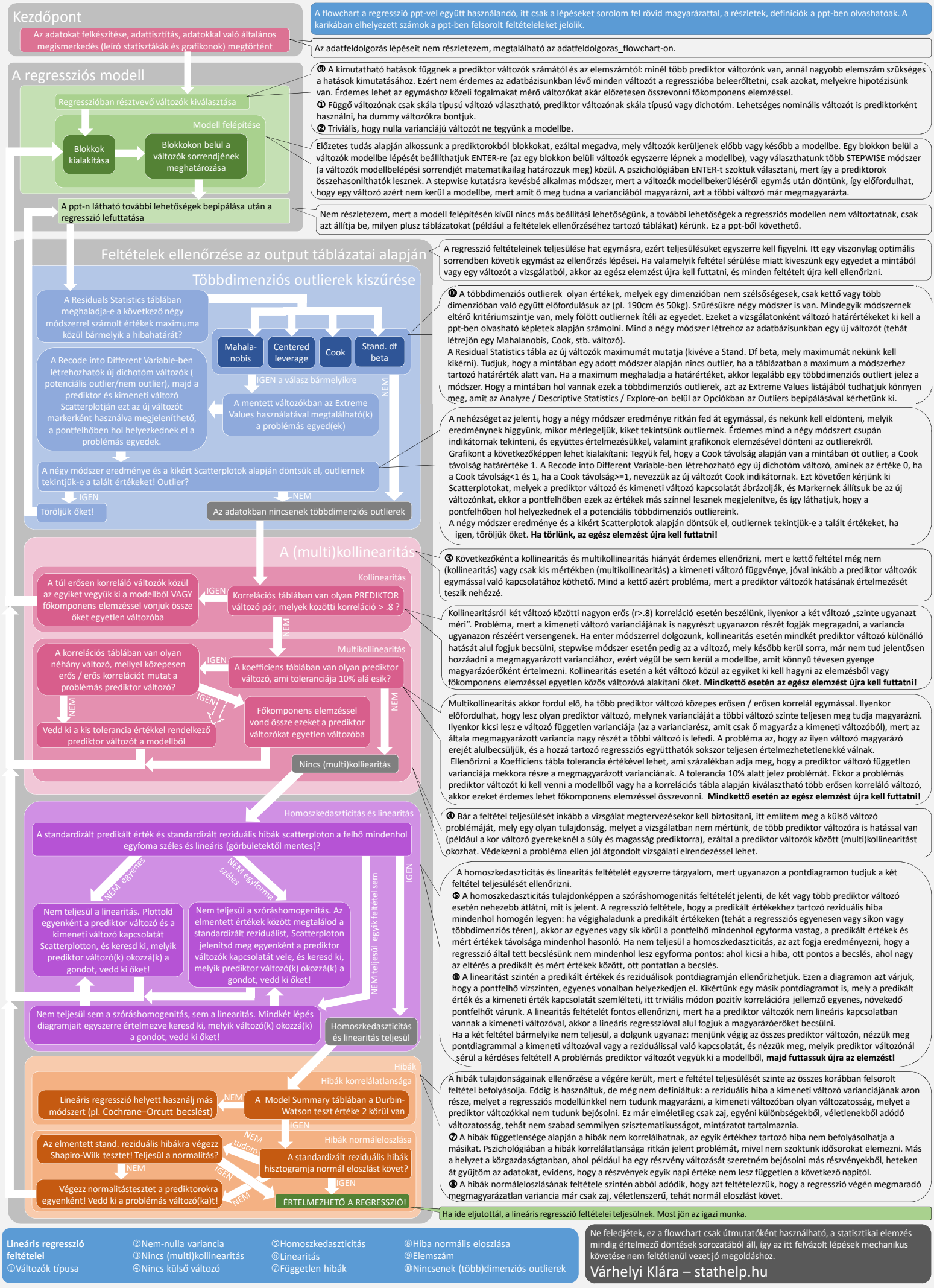
Kezdőpont
Az adatokat felkészítése, adattisztítás, adatokkal való általános megismerkedés (leíró statisztákák és grafikonok) megtörtént
A flowchart a regresszió ppt-vel együtt használandó, itt csak a lépéseket sorolom fel rövid magyarázattal, a részletek, definíciók a ppt-ben olvashatóak. A karikában elhelyezett számok a ppt-ben felsorolt feltételeleket jelölik.
Az adatfeldolgozás lépéseit nem részletezem, megtalálható az adatfeldolgozas_flowchart-on.
A regressziós modell
Regresszióban résztvevő változók kiválasztása
Modell felépítése
Blokkok kialakítása
Blokkokon belül a változók sorrendjének meghatározása
⑨ A kimutatható hatások függnek a prediktor változók számától és az elemszámtól: minél több prediktor változónk van, annál nagyobb elemszám szükséges a hatások kimutatásához. Ezért nem érdemes az adatbázisunkban lévő minden változót a regresszióba beleerőltetni, csak azokat, melyekre hipotézisünk van. Érdemes lehet az egymáshoz közeli fogalmakat mérő változókat akár előzetesen összevonni főkomponens elemzéssel.
① Függő változónak csak skála típusú változó választható, prediktor változónak skála típusú vagy dichotóm. Lehetséges nominális változót is prediktorként használni, ha dummy változókra bontjuk.
❷ Triviális, hogy nulla varianciájú változót ne tegyünk a modellbe.
Előzetes tudás alapján alkossunk a prediktorokból blokkokat, ezáltal megadva, mely változók kerüljenek előbb vagy később a modellbe. Egy blokkon belül a változók modellbe lépését beállíthatjuk ENTER-re (az egy blokkon belüli változók egyszerre lépnek a modellbe), vagy választhatunk több STEPWISE módszer (a változók modellbelépési sorrendjét matematikailag határozzuk meg) közül. A pszichológiában ENTER-t szoktuk választani, mert így a prediktorok összehasonlíthatók lesznek. A stepwise kutatásra kevésbé alkalmas módszer, mert a változók modellbekerüléséről egymás után döntünk, így előfordulhat, hogy egy változó azért nem kerül a modellbe, mert amit ő meg tudna a varianciából magyarázni, azt a többi változó már megmagyarázta.
A ppt-n látható további lehetőségek bepipálása után a regresszió lefuttatása
Nem részletezem, mert a modell felépítésén kívül nincs más beállítási lehetőségünk, a további lehetőségek a regressziós modellen nem változtatnak, csak azt állítja be, milyen plusz táblázatokat (például a feltételek ellenőrzéséhez tartozó táblákat) kérünk. Ez a ppt-ből követhető.
Feltételek ellenőrzése az output táblázatai alapján
A regresszió feltételeinek teljesülése hat egymásra, ezért teljesülésüket egyszerre kell figyelni. Itt egy viszonylag optimális sorrendben követik egymást az ellenőrzés lépései. Ha valamelyik feltétel sérülése miatt kiveszünk egy egyedet a mintából vagy egy változót a vizsgálatból, akkor az egész elemzést újra kell futtatni, és minden feltételt újra kell ellenőrizni.
Többdimenziós outlierek kiszűrése
A Residuals Statistics táblában meghaladja-e a következő négy módszerrel számolt értékek maximuma közül bármelyik a hibahatárát?
Mahala-nobis
Centered leverage
Cook
Stand. df beta
IGEN a válasz bármelyikre
NEM
A mentett változókban az Extreme Values használatával megtalálható(k) a problémás egyed(ek)
A Recode into Different Variable-ben létrehozhatók új dichotóm változók (potenciális outlier/nem outlier), majd a prediktor és kimeneti változó Scatterplotján ezt az új változót markerként használva megjeleníthető, a pontfelhőben hol helyezkednek el a problémás egyedek.
A négy módszer eredménye és a kikért Scatterplotok alapján döntsük el, outliernek tekintjük-e a talált értékeket! Outlier?
IGEN
NEM
Töröljük őket!
Az adatokban nincsenek többdimenziós outlierek
⑩ A többdimenziós outlierek olyan értékek, melyek egy dimenzióban nem szélsőségesek, csak kettő vagy több dimenzióban való együtt előfordulásuk az (pl. 190cm és 50kg). Szűrésükre négy módszer is van. Mindegyik módszernek eltérő kritériumszintje van, mely fölött outliernek ítéli az egyedet. Ezeket a vizsgálatonként változó határértékeket ki kell a ppt-ben olvasható képletek alapján számolni. Mind a négy módszer létrehoz az adatbázisunkban egy új változót (tehát létrejön egy Mahalanobis, Cook, stb. változó).
A Residual Statistics tábla az új változók maximumát mutatja (kivéve a Stand. Df beta, mely maximumát nekünk kell kikérni). Tudjuk, hogy a mintában egy adott módszer alapján nincs outlier, ha a táblázatban a maximum a módszerhez tartozó határérték alatt van. Ha a maximum meghaladja a határértéket, akkor legalább egy többdimenziós outliert jelez a módszer. Hogy a mintában hol vannak ezek a többdimenziós outlierek, azt az Extreme Values listájából tudhatjuk könnyen meg, amit az Analyze / Descriptive Statistics / Explore-on belül az Opciókban az Outliers bepipálásával kérhetünk ki.
A nehézséget az jelenti, hogy a négy módszer eredménye ritkán fed át egymással, és nekünk kell eldönteni, melyik eredménynek higgyünk, mikor mérlegeljük, kiket tekintsünk outliernek. Érdemes mind a négy módszert csupán indikátornak tekinteni, és együttes értelmezésükkel, valamint grafikonok elemzésével dönteni az outlierekről.
Grafikont a következőképpen lehet kialakítani: Tegyük fel, hogy a Cook távolság alapján van a mintában öt outlier, a Cook távolság határértéke 1. A Recode into Different Variable-ben létrehozható egy új dichotóm változó, aminek az értéke 0, ha a Cook távolság<1 és 1, ha a Cook távolság>=1, nevezzük az új változót Cook indikátornak. Ezt követően kérjünk ki Scatterplotokat, melyek a prediktor változó és kimeneti változó kapcsolatát ábrázolják, és Markernek állítsuk be az új változónkat, ekkor a pontfelhőben ezek az értékek más színnel lesznek megjelenítve, és így láthatjuk, hogy a pontfelhőben hol helyezkednek el a potenciális többdimenziós outliereink.
A négy módszer eredménye és a kikért Scatterplotok alapján döntsük el, outliernek tekintjük-e a talált értékeket, ha igen, töröljük őket. **Ha törlünk, az egész elemzést újra kell futtatni!**
A (multi)kollinearitás
❸ Következőként a kollinearitás és multikollinearitás hiányát érdemes ellenőrizni, mert e kettő feltétel még nem (kollinearitás) vagy csak kis mértékben (multikollinearitás) a kimeneti változó függvénye, jóval inkább a prediktor változók egymással való kapcsolatához köthető. Mind a kettő azért probléma, mert a prediktor változók hatásának értelmezését teszik nehézzé.
Kollinearitás
A túl erősen korreláló változók közül az egyiket vegyük ki a modellből VAGY főkomponens elemzéssel vonjuk össze őket egyetlen változóba
IGEN
Korrelációs táblában van olyan PREDIKTOR változó pár, melyek közötti korreláció > .8 ?
NEM
Kollinearitásról két változó közötti nagyon erős (r>.8) korreláció esetén beszélünk, ilyenkor a két változó „szinte ugyanazt méri”. Probléma, mert a kimeneti változó varianciájának is nagyrészt ugyanazon részét fogják megragadni, a variancia ugyanazon részéért versengenek. Ha enter módszerrel dolgozunk, kollinearitás esetén mindkét prediktor változó különálló hatását alul fogjuk becsülni, stepwise módszer esetén pedig az a változó, mely később kerül sorra, már nem tud jelentősen hozzáadni a megmagyarázott varianciához, ezért végül be sem kerül a modellbe, amit könnyű tévesen gyenge magyarázóerőként értelmezni. Kollinearitás esetén a két változó közül az egyiket ki kell hagyni az elemzésből vagy főkomponens elemzéssel egyetlen közös változóvá alakítani őket. **Mindkettő esetén az egész elemzést újra kell futtatni!**
Multikollinearitás
A korrelációs táblában van olyan néhány változó, mellyel közepesen erős / erős korrelációt mutat a problémás prediktor változó?
IGEN
A koefficiens táblában van olyan prediktor változó, ami toleranciája 10% alá esik?
IGEN!
NEM
NEM
Vedd ki a kis tolerancia értékkel rendelkező prediktor változót a modellből
Főkomponens elemzéssel vond össze ezeket a prediktor változókat egyetlen változóba
Nincs (multi)kollinearitás
Multikollinearitás akkor fordul elő, ha több prediktor változó közepes erősen / erősen korrelál egymással. Ilyenkor előfordulhat, hogy lesz olyan prediktor változó, melynek varianciáját a többi változó szinte teljesen meg tudja magyarázni. Ilyenkor kicsi lesz e változó független varianciája (az a varianciarész, amit csak ő magyaráz a kimeneti változóból), mert az általa megmagyarázott variancia nagy részét a többi változó is lefedi. A probléma az, hogy az ilyen változó magyarázó erejét alulbecsüljük, és a hozzá tartozó regressziós együtthatók sokszor teljesen értelmezhetetlenekké válnak.
Ellenőrizni a Koefficients tábla tolerancia értékével lehet, ami százalékban adja meg, hogy a prediktor változó független varianciája mekkora része a megmagyarázott varianciának. A tolerancia 10% alatt jelez problémát. Ekkor a problémás prediktor változót ki kell venni a modellből vagy ha a korrelációs tábla alapján kiválasztható több erősen korreláló változó, akkor ezeket érdemes lehet főkomponens elemzéssel összevonni. **Mindkettő esetén az egész elemzést újra kell futtatni!**
❹ Bár a feltétel teljesülését inkább a vizsgálat megtervezésekor kell biztosítani, itt említem meg a külső változó problémáját, mely egy olyan tulajdonság, melyet a vizsgálatban nem mértünk, de több prediktor változóra is hatással van (például a kor változó gyerekeknél a súly és magasság prediktorra), ezáltal a prediktor változók között (multi)kollinearitást okozhat. Védekezni a probléma ellen jól átgondolt vizsgálati elrendezéssel lehet.
Homoszkedaszticitás és linearitás
A standardizált predikált érték és standardizált reziduális hibák scatterploton a felhő mindenhol egyforma széles és lineáris (görbületektől mentes)?
NEM egyenes
NEM egyforma széles
IGEN
NEM teljesül egyik feltétel sem
Nem teljesül a linearitás. Plottold egyenként a prediktor változó és a kimeneti változó kapcsolatát Scatterplotton, és keresd ki, melyik prediktor változó(k) okozzá(k) a gondot, vedd ki őket!
Nem teljesül a szóráshomogenitás. Az elmentett értékek között megtalálod a standardizált reziduálist, Scatterploton jelenítsd meg egyenként a prediktor változók kapcsolatát vele, és keresd ki, melyik prediktor változó(k) okozzá(k) a gondot, vedd ki őket!
Nem teljesül sem a szóráshomogenitás, sem a linearitás. Mindkét lépés diagramjait egyszerre értelmezve keresd ki, melyik változó(k) okozzá(k) a gondot, vedd ki őket!
Homoszkedaszticitás és linearitás teljesül
A homoszkedaszticitás és linearitás feltételét egyszerre tárgyalom, mert ugyanazon a pontdiagramon tudjuk a két feltétel teljesülését ellenőrizni.
❺ A homoszkedaszticitás tulajdonképpen a szóráshomogenitás feltételét jelenti, de két vagy több prediktor változó esetén nehezebb átlátni, mit is jelent. A regresszió feltétele, hogy a predikált értékekhez tartozó reziduális hiba mindenhol homogén legyen: ha végighaladunk a predikált értékeken (tehát a regressziós egyenesen vagy síkon vagy többdimenziós téren), akkor az egyenes vagy sík körül a pontfelhő mindenhol egyforma vastag, a predikált értékek és mért értékek távolsága mindenhol hasonló. Ha nem teljesül a homoszkedaszticitás, az azt fogja eredményezni, hogy a regresszió által tett becslésünk nem mindenhol lesz egyforma pontos: ahol kicsi a hiba, ott pontos a becslés, ahol nagy az eltérés a predikált és mért értékek között, ott pontatlan a becslés.
❻ A linearitást szintén a predikált értékek és reziduálisok pontdiagramján ellenőrizhetjük. Ezen a diagramon azt várjuk, hogy a pontfelhő vízszinten, egyenes vonalban helyezkedjen el. Kikértünk egy másik pontdiagramot is, mely a predikált érték és a kimeneti érték kapcsolatát szemlélteti, itt triviális módon pozitív korrelációra jellemző egyenes, növekedő pontfelhőt várunk. A linearitás feltételét fontos ellenőrizni, mert ha a prediktor változók nem lineáris kapcsolatban vannak a kimeneti változóval, akkor a lineáris regresszióval alul fogjuk a magyarázóerőket becsülni.
Ha a két feltétel bármelyike nem teljesül, a dolgunk ugyanaz: menjünk végig az összes prediktor változón, nézzük meg pontdiagrammal a kimeneti változóval vagy a reziduálissal való kapcsolatát, és nézzük meg, melyik prediktor változónál sérül a kérdéses feltétel! A problémás prediktor változót vegyük ki a modellből, **majd futtassuk újra az elemzést!**
Hibák
Hibák korrelálatlansága
Lineáris regresszió helyett használj más módszert (pl. Cochrane–Orcutt becslést)
NEM
A Model Summary táblában a Durbin-Watson teszt értéke 2 körül van
IGEN
A hibák tulajdonságainak ellenőrzése a végére került, mert e feltétel teljesülését szinte az összes korábban felsorolt feltétel befolyásolja. Eddig is használtuk, de még nem definiáltuk: a reziduális hiba a kimeneti változó varianciájának azon része, melyet a regressziós modellünkkel nem tudunk magyarázni, a kimeneti változóban olyan változatosság, melyet a prediktor változókkal nem tudunk bejósolni. Ez már elméletileg csak zaj, egyéni különbségekből, véletlenekből adódó változatosság, tehát nem szabad semmilyen szisztematikusságot, mintázatot tartalmaznia.
❼ A hibák függetlensége alapján a hibák nem korrelálhatnak, az egyik értékhez tartozó hiba nem befolyásolhatja a másikat. Pszichológiában a hibák korrelálatlansága ritkán jelent problémát, mivel nem szoktunk idősorokat elemezni. Más a helyzet a közgazdaságtanban, ahol például egy részvény változását szeretném bejósolni más részvényekből, heteken át gyűjtöm az adatokat, evidens, hogy a részvények egyik napi értéke nem lesz független a következő napitól.
❽ A hibák normáleloszlásának feltétele szintén abból adódik, hogy azt feltételezzük, hogy a regresszió végén megmaradó megmagyarázatlan variancia már csak zaj, véletlenszerű, tehát normál eloszlást követ.
Hibák normáleloszlása
Az elmentett stand. reziduális hibákra végezz Shapiro-Wilk tesztet! Teljesül a normalitás?
NEM
IGEN
NEM tudom
A standardizált reziduális hibák hisztogramja normál eloszlást követ?
IGEN
Végezz normalitástesztet a prediktorokra egyenként! Vedd ki a problémás változó(ka)t!
NEM
IGEN
ÉRTELMEZHETŐ A REGRESSZIÓ!
Ha ide eljutottál, a lineáris regresszió feltételei teljesülnek. Most jön az igazi munka.
Lineáris regresszió feltételei
①Változók típusa
②Nem-nulla variancia
③Nincs (multi)kollinearitás
④Nincs külső változó
⑤Homoszkedaszticitás
⑥Linearitás
⑦Független hibák
⑧Hiba normális eloszlása
⑨Elemszám
⑩Nincsenek (több)dimenziós outlierek
Ne feledjétek, ez a flowchart csak útmutatóként használható, a statisztikai elemzés mindig értelmező döntések sorozatából áll, így az itt felvázolt lépések mechanikus követése nem feltétlenül vezet jó megoldáshoz.
Várhelyi Klára – stathelp.hu