HÁZI FELADAT 2 - JELLEMZŐ HIBÁK JAVÍTÁSA ÉS EGYÉB JAVASLATOK

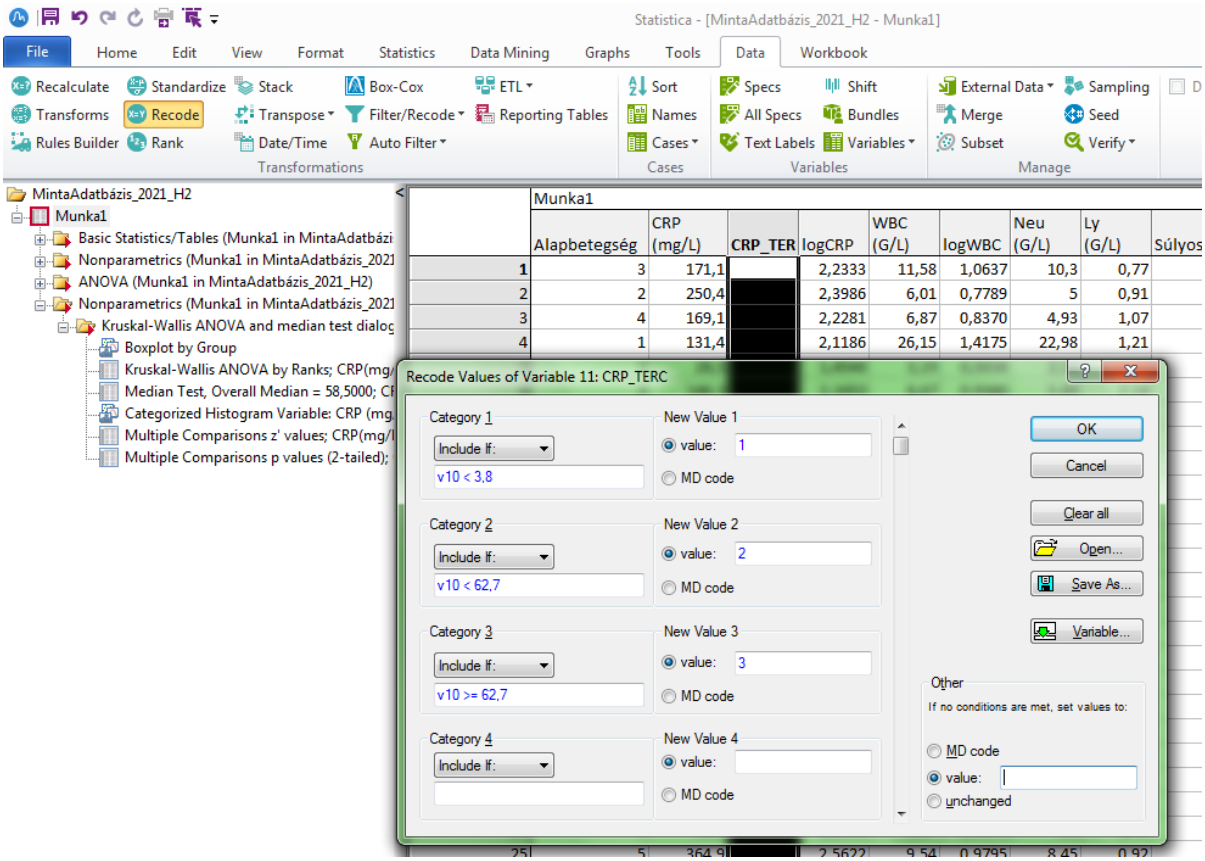
1. Feladat

Leíró statisztika:

- Nem normál eloszlású folytonos változó esetén akkor járunk el helyesen, ha az átlag és a szórás helyett a mediánt és a kvartilisokat (és esetleg a minimumot-maximumot) tüntetjük fel
- Diszkrét változó esetén az egyes kategóriákba tartozó esetek számát és százalékos arányát kell feltüntetni (*Basic statistics / Frequency tables*)

Kategorikus változó létrehozása:

- Először kiszámítjuk a megfelelő vágópontokat (itt: tercilisek) a már ismert módon (*Basic statistics / Descriptive statistics / Percentile boundaries/ First: 33, Second: 67)*
- Ez után új változót hozunk létre (pl: *jobbklikk / Add Variables)*
- Lehetséges megoldás, hogy itt függvény formájában megadjuk az új változó értékeit: =iif(v10<3.8; 1; iif(v10<62.7; 2; iif(v10>=62.7; 3; "")))
- Ennél egyszerűbb és átláthatóbb, ha kijelöljük az új változót, majd a *Data / Recode* funkció segítségével kódoljuk be a kategóriáinkat. A további kategóriákba csak a megmaradó esetekből válogat, azaz jelen esetben a 2-es értéket csak azoknak a 62,7-nél kisebb értékeknek adja, amelyek nem kerültek a 1-es kategóriába, azaz nem kisebbek 3,8-nál, stb.):

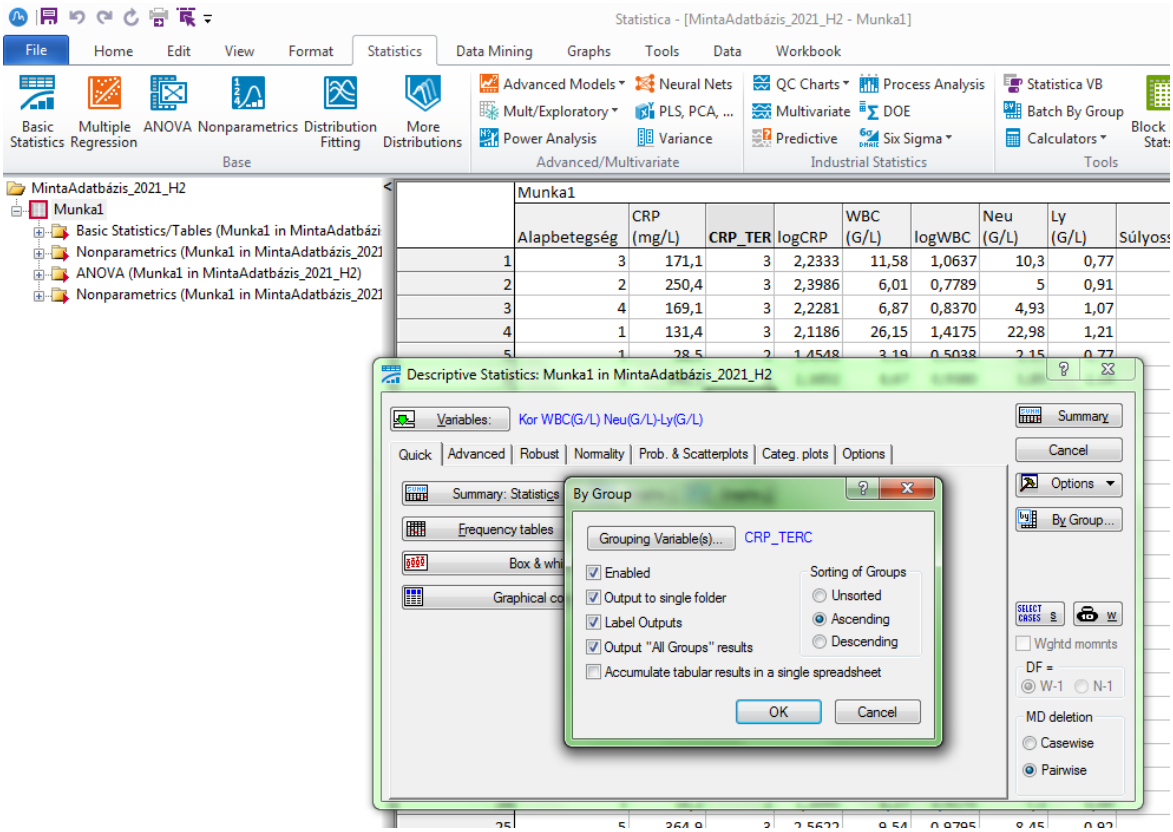


Hiányzó értékek

Ha valaki más módon végezte a kódolást, előfordulhat, hogy a hiányzó értékekkel rendelkező eseteket besorolta valamelyik csoportba. Ezt itt nem tekintettem hibának, de alapvetően érdemes erre is gondolni, mert más esetekben érdemi hibaforrás lehet.

Leíró statisztika kategorikus bontásban

A megfelelő (!) leíró statisztika elvégzésekor a *By Group* gombra kattintva tudjuk megadni a csoportosításhoz használt kategorikus változót:



2. feladat

A kontingenciatábla elkészítése, a Chi-négyzet-teszt elvégzése és az eredmények értékelése szinte mindenkinek jól sikerült.

3. feladat

Logisztikus regresszó:

Ahhoz, hogy egy adott független változónak (itt valamilyen alapbetegség megléte) a függő változóra (ITO felvétel) gyakorolt hatását önmagában logisztikus regresszióval megvizsgáljuk, először egyváltozós modelleket érdemes készíteni, vagyis amiben csak a kérdéses változó (alapbetegség) szerepel független változóként.

Többváltozós modellben az egyes független változók hatását egymásra adjusztálva tudjuk megvizsgálni. Ennek is van létjogosultsága, szerepe, szóval, ha valaki ezt végezte el az előbbi helyett, ugyanúgy elfogadtam.

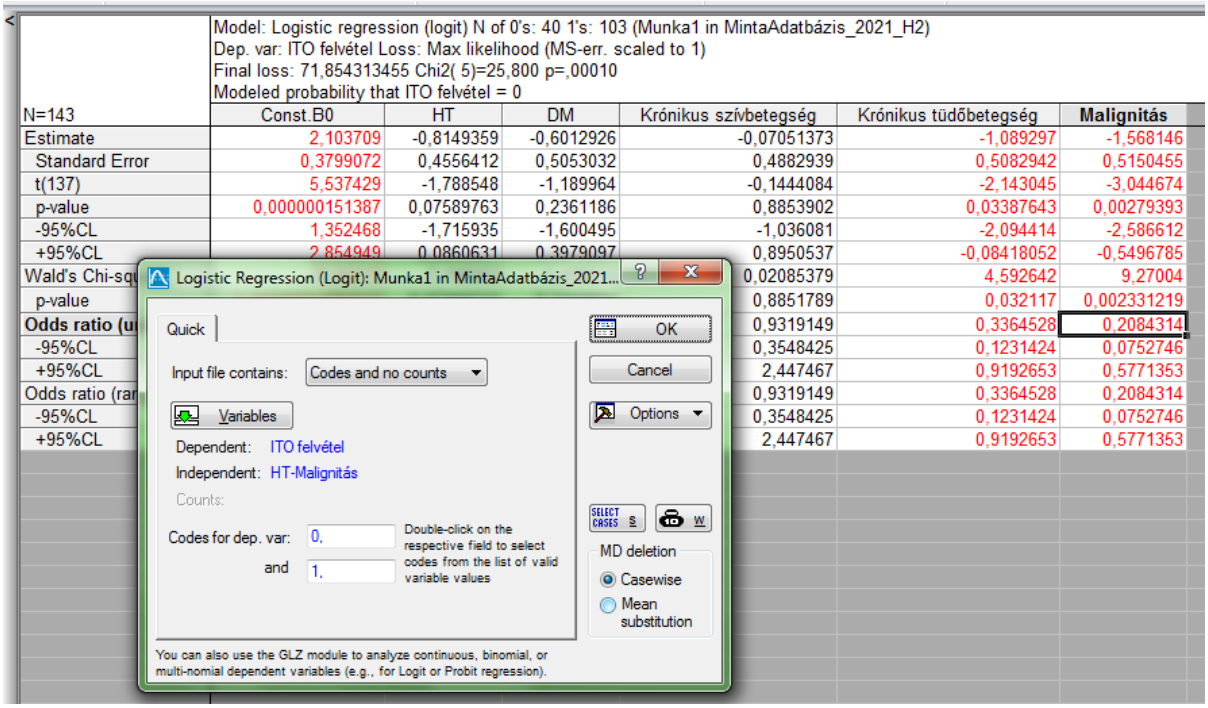
Azonban arra vigyázni kell, hogy csak olyan változókat vizsgáljunk együtt, egy modellben, amelyek között nincs erős összefüggés. Mivel pl. az "alapbetegség" változó az összes többi változó függvénye (összege), ezért ez nem vizsgálható a többivel együtt.

Az esélyhányados meghatározásánál hibaforrás lehet, ha fordítva kódoljuk a csoportokat. Ekkor ugyanis az ITO-ra nem kerülés esélyére vonatkozó eredményt kapunk (ezt sajnos csak utólag, a modell leírásánál olvashatjuk, hogy: *Modeled probability that ITO felvétel = 0)*, ami alacsonyabb lesz 1-nél (a helyes megoldás reciproka) eggyel több alapbetegség esetén.

Model: Logistic regression (logit) N of 0's: 40 1's: 103 (Munka1 in MintaAdatbázis_2021_H2)
Dep. var: ITO felvétel Loss: Max likelihood (MS-err. scaled to 1)
Final loss: 71,854313455 Chi2(5)=25,800 p=,00010
Modeled probability that ITO felvétel = 0

N=143	Const.B0	HT	DM	Krónikus szívbetegség	Krónikus tüdőbetegség	**Malignitás**
Estimate	2,103709	-0,8149359	-0,6012926	-0,07051373	-1,089297	-1,568146
Standard Error	0,3799072	0,4556412	0,5053032	0,4882939	0,5082942	0,5150455
t(137)	5,537429	-1,788548	-1,189964	-0,1444084	-2,143045	-3,044674
p-value	0,000000151387	0,07589763	0,2361186	0,8853902	0,03387643	0,00279393
-95%CL	1,352468	-1,715935	-1,600495	-1,036081	-2,094414	-2,586612
+95%CL	2,854949	0,0860631	0,3979097	0,8950537	-0,08418052	-0,5496785
Wald's Chi-square	30,66312	3,198903	1,416014	0,02085379	4,592642	9,27004
p-value	0,00000003093523	0,07369713	0,2340694	0,8851789	0,032117	0,002331219
Odds ratio (unit ch)	8,196513	0,4426677	0,5481027	0,9319149	0,3364528	0,2084314
-95%CL	3,866959	0,1797955	0,2017966	0,3548425	0,1231424	0,0752746
+95%CL	17,37355	1,089875	1,48871	2,447467	0,9192653	0,5771353
Odds ratio (range)		0,4426677	0,5481027	0,9319149	0,3364528	0,2084314
-95%CL		0,1797955	0,2017966	0,3548425	0,1231424	0,0752746
+95%CL		1,089875	1,48871	2,447467	0,9192653	0,5771353

Ha ezt látjuk, végezzük el ismét az analízist úgy, hogy az 0-s kódot írjuk az első helyre:

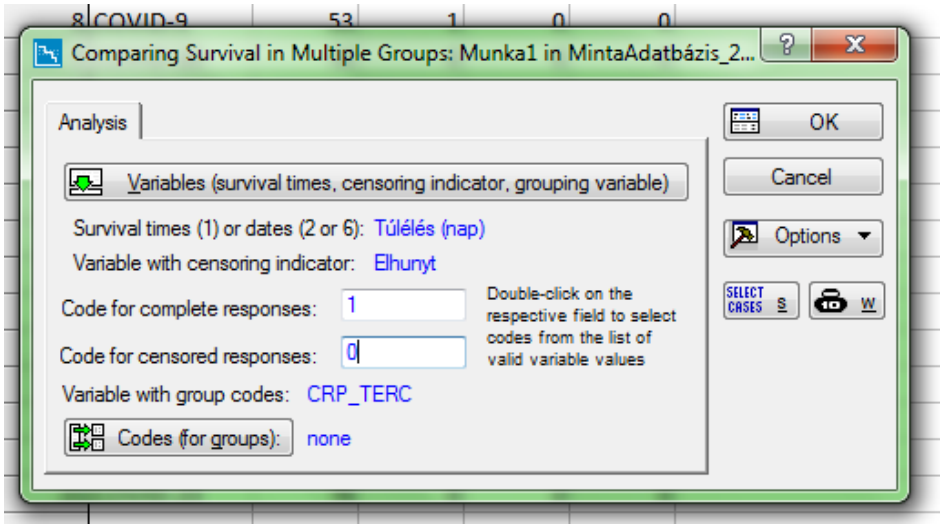


4. feladat

Kaplan-Meier görbék

Lehetséges útvonal: *Advanced models / Survival / Comparing multiple samples* (ekkor egy ábrán tudjuk az összes csoport görbéit ábrázolni…)

A görbék készítésekor lehetséges hibaforrás, ha felcseréljük a 'complete' és 'censored' eseményekhez tartozó kódokat. Helyesen:



Általános:

A szöveges válaszban (hacsak nem erre irányul konkrét kérdés) nem a statisztikai fogalmak és eredmények elméleti jelentését kell megmagyarázni, hanem az a cél, hogy úgy foglaljuk össze, fogalmazzuk meg az eredményeinket, ahogy egy cikk szöveges részében leírnánk azokat.