

A hazarrendez.sav állományban fogjuk lineáris regresszióval becsülni a házak eladási árát az adatfile-ban szereplő többi változó segítségével.

**Először olvassátok el figyelmesen a „Ház adatok elérhetősége” résznél a változók leírását!** A hazarrendez.sav úgy készült, hogy töröltem az age változót, mert túl sok eset hiányzott, néhány hiányzó adatos esetet is töröltem, véletlent generáltam és aszerint sorba rendeztem az adatokat. A nen, cust, cor változók bináris változók, melyek rendre azt kódolják, hogy északi fekvésű-e, egyedi építésű-e, sarkon van-e a ház.

### **Olvassátok el az ötödik gyakorlat anyagához írt regressziós összefoglalót!**

Megszokott dolog becselőmodell építésnél, hogy az adathalmaz egyik részén kerül felépítésre a modell (tanuló rész), míg a másik részén vizsgáljuk meg a teljesítményét (tesztelő rész). Ezzel elkerülhető a túltanulás jelensége (vagyis az, hogy általános törvényszerűségek helyett az adatokban való zajra tanuljunk rá). **Ennek megfelelően az első részfeladat az, hogy szorítkozzatok az 1-98 közötti esetekre!** Ezt a Data/Select cases-en belül a Based on time or case range-el tudjátok megtenni. Fontos, hogy ne töröljétek a nem kiválasztott eseteket, csak húzzátok át őket, vagyis alul a filtered legyen bepipálva. Az első 98 esetet fogjuk modellépítésre használni, míg a hátsó 20 eseten teszteljük az illeszkedést (mivel véletlen szerint sorba rendeztem, ezért mindez úgy tekinthető, hogy véletlenül választottuk ki a 20 tesztelésre szánt esetet).

**Ezt követően kérjétek lineáris regressziót, úgy, hogy dependent variable legyen a price, independent variable pedig az összes többi változó leszámítva a mesterségesen létrejött filter változót! Milyen kiírt statisztika segít eldönteni, hogy milyen jó az illeszkedés? Mennyi ennek a statisztikának az értéke?**

Fontos, hogy a lineáris regresszió legjobban folytonos változókkal működik. Ilyennek tekinthető az sqft, a tax, és egy kis jóindulattal a features változó. Mindazonáltal bináris változók is szerepelhetnek benne (a következő órán majd lesz arról szó, hogy hogyan lehet egyéb kategórikus változót szerepeltetni a modellben, például 3 értékűt). Például a nen bináris változó szerepeltetése lehetővé teszi, hogy a becselőmodellben különbözzön a város északi részén lévő házak és a többi ház konstansa.

A fenténél tovább mehetünk. **A Transform/Compute variable paranccsal hozzátok létre a nen és az sqft változók szorzatát! Majd futtassátok le a regressziót úgy, hogy az új változó is szerepeljen az independent résznél!** Ez a változó hozzáadást úgy nevezik, hogy interakciós lehetőséggel bővítettük a modellt. Az értelme pedig az, hogy így megengedjük azt, hogy az északi és a többi részen levő lakásoknál más legyen az sqft változó konstansa. **Mit látunk, javult az új változó bevonásával a modellünk? Nézzétek meg a modellt, az együtthatók megfelelnek a várakozásainknak? Nézzétek meg a Beta oszlopot is! Mit mondhatunk melyik a legfontosabb változó a Beta oszlop alapján? Melyik a legkevésbé fontos?** Figyelem, a beta oszlopban az abszolút nagyság a mérvadó. A Beta oszlop mellett látjátok azon hipotézisvizsgálatok p értéket, amelyeknek az a nullhipotézise, hogy a megfelelő együttható 0. **90%-on dolgozva melyik az az egy változó, amelyhez tartozó együttható 0-nak tekinthető?**

**Ezt követően futtassátok le a regressziót újra, de úgy, hogy az egyetlen 0-nak tekinthető együttthatójú változót hagyjátok ki! Az így kapott modellt fogjuk véglegesnek tekinteni. Romlott a modellünk a nem jelentős változó kihagyásával?**

Eljutottunk egy elég jó, felesleges változót nem tartalmazó modellhez. **A Transform/Compute variable segítségével számoljuk ki a modell által szolgáltatott becslést** (megjegyzem, hogy elvileg ezt lehetne automatikusan csinálni a lineáris regresszió sav almenüjével, de nekem nem működött, egész pontosan teljesen irreális számokat kaptam)! Egyszerűen be kell írni, hogy kapott konstans + együtttható1\*változó1+együtttható2\*változó2+stb. Fontos lesz a későbbiek szempontjából, hogy a regressziót az első 98 eset alapján készítette a program, most a becslést kiszámolja a teljes adathalmazra.

**Kérjetekek korrelációt a most számolt becslés változó és a price változó között!** A kapott korrelációnak elvileg meg kell egyeznie a lineáris regressziónál kiírt R statisztikával.

**Végül nézzük meg a 99-108 esetek illeszkedését!** A fenti technikával tudunk R statisztikát is számolni a 99-108 esetekre vonatkozólag, amiből egy számmal tudjuk jellemezni, hogy mennyire is jó az illeszkedés. Ezt úgy kivitelezük, hogy a Data/Select case-ben fókuszáljatok a 99-108 közötti esetekre! Majd kérjetekek újra korrelációt a price és a becslés változó között! Mit látunk?

Megoldásképpen az adatfájlt, egy táblázatokkal, magyarázó szöveggel ellátott pdf fájlt és a command syntacsot várom.

Jó munkát kívánok! Bármilyen kérdésetek van, forduljatok hozzám bizalommal!

Megjegyzés: Több interakciós változót is szerepeltethettünk volna a modellben. Nem tettük, azért, mert így is összetett a házi. A másik ok pedig az, hogy nem jó ha a modellépítésre használt 98 esethez képest túl sok változót használunk. Azt megtehettük volna, hogy definiálunk még interakciós változókat, és vagy kézzel, vagy automatikus módszerrel megkerestük volna azt a néhányat a sok változóból, ami jó illeszkedő modellt ad. Sajnos a PSPP jelenleg még nem tud automatikusan modellt építeni, de az SPSS-ben vannak ilyen lehetőségek.