Lineáris regressziószámítás 2.

Orlovits Zsanett

2018. szeptember 24.
2018. október 1.

Modellszelekció

A modellszelekció feladata nem más, mint a magyarázó változók körének optimális kiválasztása.

- Eddig egyetlen minősítőjét láttuk a modellnek: az R^2 mutatót.
- Kérdés: új változó felvétele a modellbe változtatja-e R^2 értékét?
- Az világos, hogy ekkor R^2 értéke csökkeni biztosan nem fog, hiszen "jobban" magyarázzuk Y-t, azaz egy nagyobb változókészleten minimalizáljuk a veszteségfüggvényt. Ezzel együtt viszont nő a modell bonyolultága, ami nem mindig jó!
- Tehát, ha R^2-tel jellemezzük a modellünket, akkor mindig az összes potenciális magyarázó változó felhasználása lesz a legjobb döntés.
- A valóságban azonban ez korántsem biztos!
- Mert R^2 a minta jó leírását adja, de mi a sokaságot akarjuk megragadni.

Probléma

- Fontos lesz számunkra a modell ún. általánosító-képessége, azaz hogy mennyire jól tud a mintán kívüli világról is számot adni.
- Erre a feladatra viszont az R^2 nem a legszerencsésebb mutató, hiszen ez a minta jó "megjegyzését" adja, de nekünk ennél több kell!
- Persze közben R^2-tel is törődni kell, hiszen a kellő információt ki kell szednünk a mintából ahhoz, hogy a sokaságról jól tudjunk nyilatkozni.
- Kompromisszumos megoldás kell: tanuljunk a mintából, hogy jól jellemezhessük a sokaságot, de csak annyit, ami még nem rontja a modell általánosító képességét. (Ha "túltanulom" a mintát, túlságosan specifikus lesz a modell, ami a sokaság jellemzésében már helytelen információkkal szolgál majd.)

Egy lehetséges megoldás: modellszelekciós mutató

A korábbi R^2 mutató olyan módosítása, mely figyelembe veszi a modell változóinak számát is, és meghatározható vele az optimális magyarázó-változók köre $\Rightarrow$ **korrigált determinációs együttható**

$$\bar{R}^2 = 1 - \frac{ESS(n-2)}{TSS(n-k)}$$

- Büntetjük a magyarázó változók számának növelését.
- Könnyen látható , hogy $\bar{R}^2 \leq R^2$, azaz 1-nél biztosan kisebb ez is, de vigyázat, lehet negatív is!
- A gyakorlatban heurisztikus stratégiákat használunk (forward, backward és stepwise szelekciós módszerek), hogy ne kelljen az összes 2^k kombinációt tesztelni. (automatikus modellszelekció)

Egy másik lehetséges megoldás: információs kritériumok

Módosított veszteségfüggvény használata (avagy információs kritériumok), melyek egyszerre büntetik a magyarázó változók nagy számát és a nagy hibát, a kettő közt egyensúlyt keresve:

$$V_{SC}(b) = \ln \frac{e(b)^T e(b)}{n} + \frac{k}{n} \ln n$$

$$V_{AIC}(b) = \ln \frac{e(b)^T e(b)}{n} + \frac{2k}{n}$$

$$V_{HQC}(b) = \ln \frac{e(b)^T e(b)}{n} + \frac{2k}{n} \ln(\ln n)$$

Itt a veszteségfüggvények minimalizálásával párhuzamosan k optimális értéke is keresendő!

Harmadik megoldás: modellszelekciós tesztek – WALD

- Egyidejűleg több változó elhagyását is tudja tesztelni.
- Két modell között döntünk: egy bővebb (U-unrestricted) és egy szűkebb (R-restricted) között

$$U: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+m} X_{q+m} + e_U$$
$$R: Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + e_R$$

- Beágyazott (nested) modellszelekció: a szűkebb modell minden változója benne van a bővebb modellben.
- $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+m} = 0$, tehát a megadott m db változó még összességében sem bír lényeges magyarázó erővel.
- Elhagyhatóak anélkül, hogy a modell lényegesen romlana.

Harmadik megoldás: modellszelekciós tesztek – WALD

- A próba:

$$F = \frac{(ESS_R - ESS_U)/m}{ESS_U/(n-q-m)} = \frac{(R^2_U - R^2_R)/m}{(1-R^2_U)/(n-q-m)} \sim \mathcal{F}_{m,n-q-m}$$

- Speciálisan $m = 1$ esetén ez nem más, mint egy t-teszt, de vigyázat, a Wald-teszt nem ekvivalens a t-próba m-szeri elvégzésével!
- Példánkban ha a Félszoba és Emelet változók elhagyását teszteljük, akkor az eredmények:

Null hypothesis: the regression parameters are zero for the variables
– Felszoba, DeliTaj

Test statistic: $F(2, 1398) = 7,63531$, p-value $0,00050351$

A modellfeltevések feloldása

Modellfeltevések

Emlékezzünk arra, hogy kezdetben az alábbi feltevésekkel éltünk:

- adatmátrix teljes oszloprangú, azaz nincs lineáris kapcsolat az oszlopai közt (kollinearitás)
- a hibák függetlenek a magyarázó változóktól (exogenitás)
- a hibák azonos eloszlásból származnak, azaz szórásuk megegyezik (homoszkedaszticitás)
- a hibák korrelálatlan sorozatot alkotnak (autokorrelálatlanság)
- lineáris modellstruktúra

Ezen feltételek fennállásának, illetve sérüléseinek következményeivel az ún. modellspecifikáció kérdésköre foglalkozik. Ezt fogjuk most mi is részleteiben vizsgálni.

Specifikációs hibák

- A magyarázó- és eredményváltozók kiválasztásának alapja: szakirányú elmélet, mögöttes viselkedés ismerete, múltbeli tapasztalatok.
- Mivel a valós összefüggéseket nem ismerhetjük pontosan, így nyilván az ökonometriai modellek hibákat tartalmaznak.
- Specifikációs hiba: helytelen modellfelírás, akár a változók megválasztását, akár a függvényformák vagy az eltérésváltozók (u_t) struktúráját tekintetve.
- Az alábbi esetekkel foglalkozunk: kollinearitás kérdése, lényeges változó kihagyása, felesleges változó szerepeltetése - exogenitás; az eltérésváltozó struktúrájának vizsgálata - homoszkedaszticitás és autokorrelálatlanság; valamint a megfelelő függvényforma kiválasztása - linearitás.
- Most a struktúrális hatás érdekel bennünket, azaz hogy ezek hogyan hatnak a modell belső struktúrájára.

Modellspecifikáció 1.

Multikollinearitás

Példa: $k = 3$ eset

A háromváltozós modell becslése az alábbi alakban írható fel:

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2,t} + \hat{\beta}_3 X_{3,t} + e_t, \quad t = 1, \ldots, n.$$

Átlagolva a fenti egyenletet adódik, hogy

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3.$$

Kivonva egymásból a kettőt kapjuk, hogy

$$\underbrace{Y_t - \bar{Y}}_{=:\xi_t} = \hat{\beta}_2 \underbrace{(X_{2,t} - \bar{X}_2)}_{=:\eta_{2,t}} + \hat{\beta}_3 \underbrace{(X_{3,t} - \bar{X}_3)}_{=:\eta_{3,t}} + e_t, \quad t = 1, \ldots, n,$$

azaz a konstans tagot kiküszöböltük a modellből. Ekkor

$$\det(X^T X) = \Big(\sum_{i=1}^{n} \eta_{2,t}^2\Big)\Big(\sum_{i=1}^{n} \eta_{3,t}^2\Big) - \Big(\sum_{i=1}^{n} \eta_{2,t}\eta_{3,t}\Big)^2 =$$

$$= \Big(\sum_{i=1}^{n} \eta_{2,t}^2\Big)\Big(\sum_{i=1}^{n} \eta_{3,t}^2\Big)(1 - r_{23}^2), \quad \text{ahol } r_{23} = \mathrm{corr}(X_2, X_3).$$

Példa: $k = 3$ eset

Mivel $D^2(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$, így adódik, hogy

$$D^2(\hat{\beta}_2) = \frac{\sigma^2 \sum_{i=1}^n \eta_{3,t}^2}{(\sum_{i=1}^n \eta_{2,t}^2)(\sum_{i=1}^n \eta_{3,t}^2)(1 - r_{23}^2)} = \frac{\sigma^2}{(\sum_{i=1}^n \eta_{2,t}^2)(1 - r_{23}^2)}$$

$$D^2(\hat{\beta}_3) = \frac{\sigma^2 \sum_{i=1}^n \eta_{2,t}^2}{(\sum_{i=1}^n \eta_{2,t}^2)(\sum_{i=1}^n \eta_{3,t}^2)(1 - r_{23}^2)} = \frac{\sigma^2}{(\sum_{i=1}^n \eta_{3,t}^2)(1 - r_{23}^2)}$$

Azaz ha $|r_{23}| \simeq 1$, azaz X_2 és X_3 között szoros a lineáris kapcsolat, akkor a paraméterek szórása nagy lesz.

Gyakorlati példaként gondoljunk arra, amikor pl. $X_2 =$ bér, $X_3 =$ létszám egy adott vállalat esetén.

Mi történik? A magyarázó változók egymást is magyarázzák, tehát a modell minősége romlik. Ez jelenik meg a szórások növekedésében.

Multikollinearitás

Azt a jelenség, amikor a magyarázó változók lineáris kapcsolatban vannak egymással, multikollinearitásnak nevezzük. Bár nem tökéletesen precíz, de a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást. Az ennek megfelelő mérőszám az ún. tolerancia:

$$\mathrm{Tol}_j = 1 - R_j^2 = 1 - R^2_{X_j|X_2,X_3,\dots,X_{j-1},X_{j+1},\dots,X_k},$$

azaz hogy a vizsgált magyarázó változót mennyire magyarázza a többi magyarázó változó. **Azaz minél kisebb a tolerancia, annál kevesebb többletinformációt hoz a változó a modellbe a többi változó mellett.** Ekkor

$$D^2(\hat{\beta}_j) = \frac{ESS/(n-k)}{(n-1)D^2(X_j)} \cdot \frac{1}{\mathrm{Tol}_j}$$

azaz, ha a tolerancia romlik (csökken), akkor a becsült paraméter szórása nő!

Multikollinearitás

Variancia infláció tényező:

$$\mathrm{VIF}_j = \frac{1}{\mathsf{Tol}_j}$$

$\mathrm{VIF}_j = 1$ esetén a magyarázó változót egyáltalán nem magyarázza a többi, és pl. $\mathrm{VIF}_j = 2$ esetén a mintavételi variancia megduplázódott pusztán a multikollinearitás miatt. A példánkra vetítve:

Variance Inflation Factors

Terulet	3,683
Terasz	1,241
Szoba	3,512
Felszoba	1,249
Furdoszoba	1,632
Emelet	1,020
DeliTaj	1,038

Modellspecifikáció 2.

Exogenitás

Specifikációs hibák - erős exogenitás

Elsőként az ún. **erős exogenitás** kérdésével foglalkoznunk, azaz a hibák és a magyarázó változók (várható értékben való) függetlenségére vonatkozó feltétellel.

Sérülésének tipikus esetei:

- **lényeges magyarázó változó kihagyása**
- mérési hiba a magyarázó változónál (pl. zajjal terhelten tudjuk csak megfigyelni)
- szimultaneitás (többegyenletes modelleknél)

Jelen kurzus keretei közé csak az **első eset** vizsgálata fér.

1. Lényeges változó kihagyása - Példa

Tegyük fel, hogy a valódi modell

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

alakú, de mi az

$$Y_t = \beta_1 + \beta_2 X_{2t} + v_t$$

modellt becsültük. (Azaz a $\beta_3 = 0$ feltételezéssel éltünk, de ez hibás.) Ekkor könnyen látható, hogy $v_t = \beta_3 X_{3t} + u_t$, melyre

$$\mathrm{E} v_t = \beta_3 X_{3t} \neq 0,$$

tehát v_t megsérti a modell feltételeit. Továbbá az is megmutatható, hogy ha a $\beta_3 = 0$ feltételezés hibás, akkor az összes többi együttható becslése torzított lesz, azaz az előrejelzések is torzítottak lesznek és a tesztek is hatástalanok.

2. Változó bevonásának hatása a modellre - Példa

Tekintsük a következő feladatot:

- Eredményváltozó: (Y) a háztartások kiadása
- Magyarázóváltozó: (X_J) a jövedelem
- Magyarázóváltozó: (X_L) a család létszáma

Tegyük fel, hogy modellünket kétféle módon becsültük, és az alábbi eredményeket kaptuk:

$$\hat{Y}_t = 339,746 + 0,637 \cdot X_J, \qquad \bar{R} = 0,5369$$

$$\hat{Y}_t = 283,172 + 0,617 \cdot X_J + 34,17 \cdot X_L, \qquad \bar{R} = 0.5386$$

Miért változtak meg az együtthatók?

2. Változó bevonásának hatása a modellre - Példa

- Tegyük fel, hogy a bővebb, második modell írja le a valóságos helyzetet (ez persze csak feltételezés, hiszen a valóságot nem ismerhetjük).
- Lehet-e ez alapján következtetni az első modellbeli jövedelem együtthatóra?
- Nézzük meg, hogy van-e a két változó között kapcsolat, azaz illesszünk ezekre is egy regressziót!

Az eredmény:

$$\hat{X}_L = 1,655 + 0,000598 \cdot X_J, \quad \bar{R} = 0,2359,$$

azaz a jövedelem nem csak a kiadásra hat sztochasztikusan, hanem összefügg a család létszámával is!

2. Változó bevonásának hatása a modellre - Példa

Tehát azt kaptuk, hogy a szűkebb modell együtthatói előállíthatók a bővebb modell és a harmadik regresszió segítségével, mégpedig

$$339,746 = 283,172 + 34,17 \cdot 1,655$$

és

$$0,637 = 0,617 + 34,17 \cdot 0,000598$$

formában. Mi ennek az oka?

- A bővebb modellben a jövedelem **direkt hatás**a szerepel csak, mert itt a család létszámát állandó értéken tartjuk, ezért nincs jelentősége a köztük lévő kapcsolatnak.
- A szűkebb modellben viszont a jövedelem egységnyi növekedése a létszámot is növeli tendenciájában, a növekvő létszám viszont önmagában is növeli a kiadást. Ez az ún. **indirekt hatás**.

2. Változó bevonásának hatása a modellre - Példa

- A szűkebb modellben nem tudjuk izolálni a létszám hatását, a jövedelem változásával a létszám is változik, míg a bővebb modellben a jövedelem növekedése vagy csökkenése nem jár a létszám megváltozásával.
- Tehát a szűkebb modellben a kihagyott változón keresztül terjedő hatások is beépülnek az együtthatókba.
- A becsült paraméterekre gyakorolt hatás nehezen következtethető ki ebből, mert az indirekt hatástól függ a torzítás iránya és nagysága. (A kihagyott változónak önmagában is van hatása az eredményváltozóra, és korrelál a bent lévő változóval is.)
- Ebből adódóan a torzítás bármilyen előjelű lehet attól függően, hogy ez a két tényező milyen előjelű.

Modellspecifikáció 3-4.

Heteroszkedaszticitás és autokorrelálatlanság

Specifikációs hibák - homoszkedaszticitás és autokorrelálatlanság

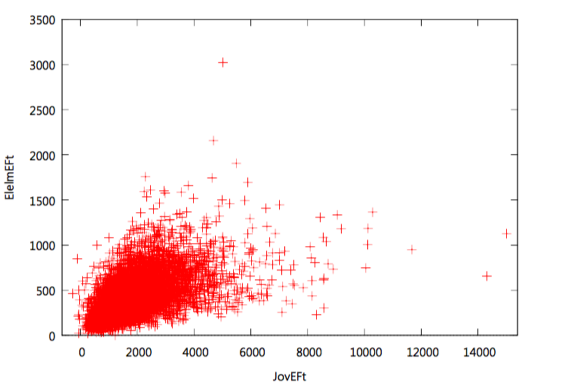
A következőkben vizsgáljuk meg, hogy mit is jelent a **homoszkedaszticitás és autokorrelálatlanság** feltételek megsértése.

- A **homoszkedaszticitás**i feltétel azt kötötte ki, hogy a hibák különböző megfigyelésekhez tartozó szórása állandó legyen, azaz nem függ attól, hogy melyik megfigyelésről van szó, avagy a becsült értékek szóródása a tényleges körül állandó.
- Az **autokorrelálatlanság** tartalma az, hogy a különböző megfigyelésekhez tartozó hibák korrelálatlanok egymással
- **Együtt a kettő úgy fogalmazható meg, hogy a hibák kovariancia mátrixa $\sigma^2 I$ alakú, ahol σ a közös szórás, I pedig a megfelelő méretű egységmátrix.**

Specifikációs hibák - homoszkedaszticitás és autokorrelálatlanság

Az autokorrelálatlanság megsértését az idősorok elemzésénél vizsgáljuk, mert keresztmetszeti adatok független mintavételezésénél ennek megsértése elvileg lehetetlen.

Homoszkedaszticitás megsértése - **heteroszkedaszticitás**



Specifikációs hibák - homoszkedaszticitás és autokorrelálatlanság

A **heteroszkedaszticitás** (változó szórás) lehetséges okai:

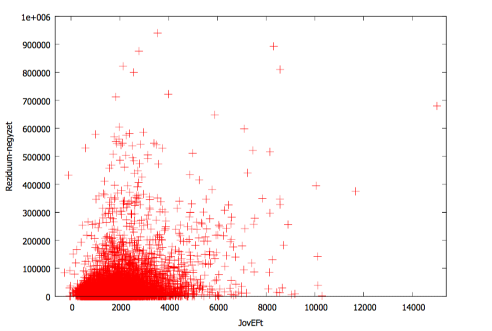
- A jelenség természetes velejárója (ld. a fenti példát)
- Csoportosított adatok használata - ha háztartásonként átlagoljuk mind a jövedelmet, mind a kiadást, akkor a csoportosított adatokban a szórásnégyzet σ^2/n_i alakú (n_i az i-dik háztartás létszáma), ami nyilván nem lesz konstans háztartásról háztartásra különböző létszámok miatt

Ebben az esetben az OLS ugyan működik, de minden hasznos tulajdonsága (hatásosság, teszt-statisztikák eloszlása, stb...) sérül, így használata nem javasolt.

Az első kérdés most az, hogy hogyan is lehet kimutatni a heteroszkedaszticitás jelenségét az adatokban.

Specifikációs hibák - heteroszkedaszticitás

1. lehetőség - reziduum-négyzetek grafikus ábrázolása a különféle magyarázó változókkal (vagy az eredményváltozóval) szemben:



Előzetesen persze nem tudjuk, hogy melyik változó felel a jelenségért, ezért kell többet is vizsgálni.

2. lehetőség - Analitikus próbák (LM próbák, White teszt)

Specifikációs hibák - heteroszkedaszticitás

A heterokedaszticitás kezelése:

- Modellspecifikáció megváltoztatása - szakmai ismeretet igényel, nem univerzális (pl. logaritmálás, deflálás - népességről áttérés népsűrűségre, stb...)
- **Általánosított legkisebb négyzetek módszere (GLS)** - a teljes modellt újrabecsüli. Ennek részleteit fogjuk most megvizsgálni.

GLS-becslés

A modell most is $y = X\beta + u$ alakú, és tegyük fel, hogy

$$u \sim \mathcal{N}(\underline{0}, \sigma \cdot \Omega^{1/2}),$$

ahol

- $\Omega > 0$ ismert kovariancia mátrix,
- nem feltétlenül diagonális, és ha az, a diagonális elemek akkor sem feltétlenül egyenlők,
- azaz a modell heteroszkedasztikus

Mi történik ekkor az OLS becsléssel? A torzítatlanság és a konzisztencia nem romlik el, de már nem lesz hatásos a becslés, azaz nem ez lesz a legkisebb szórású becslése a paramétereknek. A becsült standard hibák is torzítottak lesznek, így a tesztek érvényüket vesztik!

GLS-becslés

Ω szimmetria tulajdonsága és pozitív definitsége miatt Ω^{-1} is szimmetrikus pozitív definit mátrix, így létezik olyan P nemszinguláris mátrix, melyre $\Omega^{-1} = P^T P$. Szorozzuk végig ezzel a P mátrixszal balról a modell-egyenletet. Ekkor

$$Py = PX\beta + Pu,$$

és legyen $Py = \tilde{y}$, $PX = \tilde{X}$ és $Pu = \tilde{u}$. Könnyen látszik, hogy ekkor

$$\mathrm{E}(\tilde{u}\tilde{u}^T) = P\mathrm{E}(uu^T)P^T = P(\sigma^2\Omega)P^T = \sigma^2 P(P^T P)^{-1}P^T = \sigma^2 I,$$

tehát a transzformált modell már homoszkedasztikus, így működnek a korábbi becsléseink. Azaz

$$\hat{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y} = (X^T P^T P X)^{-1} X^T P^T P y =$$

$$= (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y.$$

Modellspecifikáció 5.

Nemlinearitás

Nemlinearitás - alapgondolatok

- Az élet általában nem lineáris.
- Mégis, a lineáris modellek sokszor nem térnek el (nagyon) a valóságtól, ráadásul matematikailag sokkal könnyebben kezelhetők, mint a nemlineáris modellek.
- Az esetek többségében persze ez csak egy közelítés lesz, de ha "ügyesek" vagyunk, akkor ez nem nagy baj.
- Érdemes vizsgálni a modellek érvényességi határait.

Nemlinearitás fajtái 1.

Változójában nemlineáris modell:

- Továbbra is fennáll a modell-struktúra, azaz csak alapműveleteket végzünk a paraméterek és a változók közt.
- Az OLS szempontjából mindegy, hogy egy változó esetlegesen egy másik valamilyen transzformáltjaként áll elő.
- Ide tartozik a kvadratikus és polinomiális hatás, a logaritmikus és exponenciális hatás, stb...
- Minden korábbi becslés és technika ugyanúgy működik, mint korábban.
- De, vigyázni kell, **paramétereiben lineáris marad a modell**!
- Az interakció szintén változóbeli nemlinearitás, ezt a jelenséget vizsgáljuk most meg részletesen.

A marginális hatás

- **A marginális hatás:** a magyarázó változó egységnyi növelésének hatására mennyit változik az eredményváltozó
- Mértékegységekkel nem kell törődni.
- Matematikai definíció:

$$\frac{\partial Y}{\partial X_j}$$

- Ha a modellünk az

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u,$$

akkor

$$\frac{\partial Y}{\partial X_j} = \beta_j,$$

azaz a **marginális hatás és a becsült regressziós együttható kvázi szinonimák**! (Persze csak akkor, ha a modell minden korábbi feltételnek eleget tesz!)

Interakció

- Eddigi modellünkben a marginális hatások a többi változó szintjétől függetlenül állandóak voltak. De!
- Térjünk vissza a korábbi példához: igaz-e az, hogy adott egységnyi pluszjövedelem a család létszámától függetlenül azonos többletkiadást jelent? Nyilván nem...
- Ilyenkor azt modjuk, hogy a két változó között **interakció** van, azaz az egyik marginális hatásának nagyságát befolyásolja a másik szintje.
- Tehát a **kapcsolat a marginális hatás és a szint között van**, nem pedig a marginális hatások közt avagy a szintek közt külön-külön!

Interakció

Tegyük fel, hogy a kapcsolat lineáris, azaz az egyik változó szintje lineárisan hat a másik változó marginális hatására. Például

$$(\beta_J + \beta_{JL} X_{letszam}) X_{jovedelem},$$

ahol β_{JL} az interakció hatását kifejező együttható. Ekkor behelyezve ezt a regresszióba adódik, hogy

$$\begin{aligned} Y &= \beta_1 + (\beta_J + \beta_{JL} X_{letszam}) X_{jovedelem} + \beta_L X_{letszam} + u = \\ &= \beta_1 + \beta_J X_{jovedelem} + (\beta_L + \beta_{JL} X_{jovedelem}) X_{letszam} + u \end{aligned}$$

azaz az interakció szükségképpen **szimmetrikus**, tehát **a hatás kölcsönös**! Ezért írható helyette egyszerűbben a

$$\beta_J X_{jovedelem} + \beta_L X_{letszam} + \beta_{JL} X_{jovedelem} X_{letszam}$$

összefüggés is, a marginális hatás ebből már adódik.

Marginális hatás interakciók esetén

Ha **interakció van a modellben**, például az l-dik és m-dik tag közt, akkor az l-dik marginális hatása

$$
\begin{aligned}
\frac{\partial Y}{\partial X_l} =& \frac{\partial}{\partial X_l}(\beta_1 + \beta_2 X_2 + \ldots + \\
&+ \beta_l X_l + \ldots + \beta_m X_m + \ldots + \beta_k X_k + \beta_{lm} X_l X_m + u) = \\
&= \beta_l + \beta_{lm} X_m,
\end{aligned}
$$

azaz az egyik változó szerinti marginális hatás tényleg a másik változó szintjétől függ!

Nemlinearitás fajtái 2.

Paramétereiben nemlineáris modell:

- Megsérti a lineáris kombináció struktúráját, hiszen a paraméter nem csak együtthatóként szerepel a modellben.
- Ez már nem becsülhető az OLS módszerrel, mert az eredményváltozó nem állítható elő mátrixműveletekkel.
- Nemlineáris legkisebb négyzetek (NLS) módszerével becsülhetőek ezek a modellek, de ezek általában nehezen, vagy csak lokálisan megoldható feladatok. (Iteratív algoritmusokat használunk, de ezeknek gond lehet a konvergenciájával, stabilitásával és az egyértelműségével is.)
- **Kezelése: algebrai linerizációval és az OLS használatával történik.** (Ez persze nem mindig működik, de a gyakorlatban fontos esetekben azért általában használható.)

Log-Log modell

- Kedvelt a termelési és keresleti függvények becslésénél, pl. a jól ismert **Cobb-Douglas termelési függvény** is ilyen:

$$Y = \beta_1 L^{\beta_L} K^{\beta_K} u$$

ahol Y a kibocsátás, L a munka, K a tőke felhasználása.

- Linearizálás: vegyük mindkét oldal logaritmusát. Ekkor

$$\log Y = \log \beta_1 + \beta_L \log L + \beta_K \log K + u'$$

Log-Lin modell

- Tipikusan **jövedelmek, bérek alakulásának modellezése** a tanulással vagy tapasztalatszerzéssel töltött évek függvényében.
- A modell:
$$Y = e^{\beta_1 + \beta_2 X + u}$$
azaz
$$\log Y = \beta_1 + \beta_2 X + u$$
- Tehát az eredményváltozót logaritmizálva a magyarázóváltozók maradnak szintben.

Lin-Log és reciprok modellek

- Lin-Log modell: tipikusan a **termés és a megművelt terület nagysága**, vagy a terület és a kínálati ár összefüggésének jellemzésére szolgál.

$$Y = \beta_1 + \beta_2 \log X + u$$

- Reciprok modell: tipikusan a **keresleti modellek** ilyenek. Itt

$$Y = \beta_1 + \frac{\beta_2}{X} + u$$

De hogyan is dönthetünk okosan és egyszerűen a különböző függvényformák közt?

Specifikációs tesztek - Ramsey RESET teszt

- **RESET - Regression Specification Error Test** - a regresszió specifikációs hibájának tesztje
- Ez a modellspecifikáció általános tesztje.
- A magyarázó változók körét bővíti ki az eredeti regresszió becsült eredményváltozójának magasabb hatványaival, majd erre a kibővített modellre végez Wald-féle F-próbát az új változók elhagyását tesztelve.
- A nullhipotézis elutasításakor azt kapjuk, hogy az új változók legalább egyikének hatása van a modellre, azaz ez a specifikációs hibának a jele.
- Általában vizsgálja, hogy egy adott specifikáció jó-e, de ha nemleges választ ad, akkor sajnos nem derül ki, hogy hol a hiba.