Eötvös Loránd Tudományegyetem

Természettudományi Kar

Statisztikai modellek értékelő eljárásai

Szakdolgozat

Készítette: Kovács Gergely

Matematika BSc, Matematikai elemző szakirány

Témavezető: Prőhle Tamás

Matematikai Intézet

Valószínűségelméleti és Statisztika Tanszék

Budapest

2015

Köszönetnyilvánítás

Soha ki nem fogyó hálás köszönettel tartozom édesanyámnak, aki féltő szeretetével és gondoskodásával mindenben támogatott, biztatott és ha kellett, noszogatott, valamint témavezetőmnek, Prőhle Tamásnak, akihez bármikor be tudtam menni, mindig volt hozzám egy jó szava és aki szakmailag és emberileg is mindent megtett azért, hogy ez a szakdolgozat sikerüljön.

Tartalomjegyzék

Bevezető

A természet és a társadalom jelenségeinek vizsgálata során a kutató szeretné megérteni a jelenségek mögött meghúzódó okokat és összefüggéseket, ezért ennek érdekében megfigyeléseket végez és azokról adatokat gyűjt. A valóságot azonban nem tudjuk azt felfogni, megragadni úgy, ahogy van, ezért modellek segítségével igyekszünk azt értelmezni.

Egy modell funkciója, hogy segítsen megmagyarázni egy rendszert, tanulmányozni a különböző komponenseinek a hatásait és előrejelzéseket tenni a viselkedésére.

A modellben megjelenő mennyiségeket változókra és paraméterekre oszthatjuk. A modell változói azok a mennyiségek, amelyek függetlenül mérhetők egy kísérletben. Egy modellt azért tervezünk, hogy megmagyarázzuk a változók közti kapcsolatokat. A modell paraméterei olyan állandók, amelyek a természet valamilyen inherens tulajdonságait jelölik.

A modellezés során az adatokat predikciós függvények segítségével modellezzük, az ismeretlen modellparamétereket pedig az adatokból becsüljük. Az adatok változókból állnak. A változók típusai: magyarázó (független) és magyarázandó (függő) változók. A független változók az inputokat/okokat jelenítik meg, vagy pedig az kerül megvizsgálásra, hogy a tényleges okok megegyeznek-e a feltételezett okokkal. A függő változó az outputot/hatást jeleníti meg, vagy pedig az kerül megvizsgálásra, hogy a tényleges hatás megegyezik-e a feltételezett hatással.
A modellezés célja annak vizsgálata, hogy a független változók variálásával a függő változó is variálódik-e, és ha igen, hogyan és milyen mértékben.

Egy modell helyességét a legjobban úgy lehet mérni, hogy mennyire jól tudja megmagyarázni a már ismert jelenségeket (milyen magyarázó ereje van), illetve hogy mennyire jól tudja megjósolni a még ismeretlen jelenségeket (milyen predikciós ereje van). Ha egy modell jó, akkor a predikciós érték "közel" van a tényleges értékhez. A predikciós és a tényleges érték közti különbség a predikciós hiba. A predikciós hibát úgy értelmezhetjük, hogy ez a függő változónak a független változó által nem megmagyarázott változékonysága.

Egy adott jelenséget általában nagyon sokféle modellen keresztül lehet vizsgálni, amelyek eltérnek egymástól bonyolultságban és pontosságban, ezért a modellválasztás a tudományos vizsgálódás egyik alapvető feladata.

A modellválasztás 2 fázisból áll:

Az első fázisban a kutató a számtalan szóba jövő, lehetséges modell közül kiválaszt néhányat, gyakran a háttérismeretére, előzetes tudására, intuíciójára támaszkodva, esetleg figyelembe véve egyéb szempontokat is (pl. a túlzott bonyolultság kerüléséért polinomokat használ fel, bár tudja, hogy nem az a legpontosabb). Az első fázis során a kutató által kiválasztott modelleket jelöltmodelleknek (candidate model) nevezzük.

A második fázisban pedig a statisztikai elemzésre hárul az a feladat, hogy a jelöltmodelleket értékelje, egymással összehasonlítsa, hatékonyságukat és az illeszkedés jóságát minél pontosabban mérje, annak érdekében, hogy a kutató ki tudja választani a jelöltmodellek közül a szándéka és a tudományos vizsgálat szempontjai szerint 'legjobb' modellt (vagyis azt, amelyik a legjobban leírja és megmagyarázza a kutató által vizsgált jelenséget).

1. fejezet

Modellválasztás

1.1. A modellválasztás lehetséges módszerei

A modellválasztás problémája tehát azzal a kérdéskörrel foglalkozik, hogy egy jelenség megmagyarázásának céljából összegyűjtött adatokra illeszthető számtalan statisztikai modell közül melyik a 'legjobb', amelyet a kutatónak érdemes kiválasztania, hogy azzal modellezze az adott jelenséget. E problémának a fontosságát jól mutatja az a tény, hogy milyen sokféle módszert dolgoztak ki a 'legjobb' modell fogalmának a pontos meghatározására és a modellválasztás megkönnyítésére.

Ilyen módszerek például: Akaike információs kritérium (AIC), Bayes információs kritérium (BIC), Mallows C_p, khí-négyzet teszt, az F teszt hierarchikus modellekre, minden modell értékelése (exhaustive search), lépésenkénti módszer (stepwise), a vissz- vagy előrelépő modell választó módszer, keresztvalidáció, Bayes-faktor, Bayes-féle modell átlagolás stb.

Például az Akaike információs kritériumon alapuló modellválasztási módszer egy statisztikai modell minőségét az összes többi modellhez viszonyítva becsli meg. Viszont amikor a modellezés célja az előrejelzés, vagyis meg akarják becsülni azt, hogy egy prediktív modell a gyakorlatban milyen pontosan fog működni, akkor előnyösebb validációs halmazon alapuló modellválasztási módszert használni. Mivel a még le nem zajlott jelenségekről szóló megfigyelések nem állnak rendelkezésre, ezért jön az az ötlet, hogy a rendelkezésre álló megfigyelések egy részét tekintsük úgy, mintha jövőbeli megfigyelések volnának és vizsgáljuk meg, hogy a többi megfigyelésre illesztett modell mennyire jól képes előrejelezni ezeket a jövőbelinek tekintett megfigyeléseket, vagyis értékeljük a modell teljesítményét. A modellválasztás validációs halmazon alapuló megközelítése (validation set approach, VSA) tehát nem a modelleket viszonyítja egymáshoz, hanem az adathalmazt (a megfi-

gyelések halmazát) valamilyen módszer szerint két részre osztja fel: a konstrukciós halmazra (construction set/training set) és a validációs halmazra (validation set). A konstrukciós halmazra illesztjük a modellt, majd pedig ezt az illesztett modellt a validációs halmazban lévő megfigyelések előrejelzésére (predikciójára) használjuk fel. A validációs halmazban lévő megfigyelések egy konkrét előrejelzése közben tett hibát validációs hibának, az előrejelzések várható hibáját pedig generalizációs hibának nevezzük. A generalizációs hiba alapján képet kaphatunk az adott modell generalizációs képességéről, vagyis arról a képességről, hogy mennyire jól "általánosít" új megfigyelésekre. A validációs hiba a generalizációs hiba egy becslését adja meg. A validációs hiba értékét az átlagos négyzetes hibával (mean square error, MSE) értékeljük ki, s ez az MSE-mennyiség alkalmas a modell teljesítményének mérésére. A modell teljesítményének mérése viszont arra is lehetőséget ad, hogy összehasonlítsuk alternatív modellek teljesítményét. Az a modell jobb, aminek kisebb a validációs hibája (vagyis generalizációs hibabecslése). Az alternatív modellek teljesítményének összehasonlítása pedig megteremti annak a lehetőségét, hogy az alternatív modellek közül kiválasszuk a (számunkra) optimális modellt.

A modellválasztás végső célja a "jó" generalizáció, de emellett kívánatos tulajdonság még a konzisztens választás is. Egy modellválasztás akkor konzisztens, ha 1-hez tart annak a valószínűsége, hogy a helyes és optimális modellt választjuk. Ez a két cél: a "jó" generalizáció és a konzisztens modellválasztás azonban egyszerre nem valósítható meg, vagyis egymással szemben álló döntési szabályokhoz vezet.

Sokféle módszer és szempont szerint lehet felosztani a megfigyelések halmazát. Ezeket a különböző módszerek és szempontok szerinti felosztásokat tárgyaljuk a következő fejezetben.

1.2. A keresztvalidációs módszerek fajtái

A validációs halmazon alapuló megközelítésnek (VSA-nak) két hátránya van:

1. Mivel véletlenszerű az, hogy mely megfigyelések kerülnek a konstrukciós halmazba és melyek a validációs halmazba, ezért a validációs halmazon számolt becslés nagy mértékben függ a megfigyelések felosztásától, így nagyon változékony lehet;

2. A validációs hiba hajlamos túlbecsülni a generalizációs hibát, ha a modellt a teljes adathalmazra illesztjük.

Újramintavételezési technikák használatával (keresztvalidáció, bootstrap) azonban ezek a korlátok leküzdhetők (a számítások megnövekedésének az árán). Ezeknek a módszereknek

néhány változatát mutatjuk most be.

1.2.1. Leave-one-out Cross-Validation 'egyet hagyj ki', LOO, CV(1)

CV(1)-módszer esetén a validációra egyetlen megfigyelést használunk, a többi megfigyelés a konstrukciós halmazt alkotja. Ezt megismételjük n-szer és kiszámítjuk az n becslés átlagát.

Előnyei a VSA-val szemben:

1. A torzítás sokkal kisebb (mivel a konstrukciós halmaz $n-1$ megfigyelést tartalmaz);

2. Az LOO ismétlése mindig ugyanazt az MSE-t eredményezi (mivel nincs véletlenszerűség a konstrukciós/validációs felosztásban).

Hátrányai:

1. A kiszámítása drága lehet (mivel a modellt n-szer kell illeszteni);

2. Aszimptotikusan helytelen;

3. Konzervatív (ami azt jelenti, hogy hajlamos az optimálisnál bővebb modellt választani). Ennek a problémájával a 3.4.1. fejezetben foglalkozunk. $CV(1)$.

1.2.2. Hold Out Cross-Validation 'vedd ki', HO, $CV(n_v)$

A hold out-módszer esetén is a megfigyelések halmazát két részre bontjuk: a konstrukciós halmazra és a validációs halmazra. A modellt a konstrukciós adatrész alapján illesztjük, a predikciós hibát pedig a validációs adatrészen számoljuk, mintha azok a jövőbeli értékek volnának. Ezt gyakran jelölik $CV(n_v)$-vel, ahol n_v a validációs halmaz elemszámát jelöli. Látható, hogy a $CV(1)$ módszer $CV(n_v)$ speciális esete $n_v \equiv 1$-gyel. Innen érthető a jelölés is.

1.2.3. K-fold Cross-Validation 'K-szoros', KCV, $KCV(k)$

A KF-módszer esetén az $\mathcal{A}$ adathalmazt k db, közel egyenlő méretű diszjunkt részhalmazra osztjuk fel, majd k lépésben k db különböző modellt építünk úgy, hogy mindig egy

különböző részhalmazt választunk validációs halmaznak, a többi pedig a konstrukciós halmazt alkotja, s a módszer generalizációs hibájának a k db különböző modell generalizációs hibáinak az átlagát tekintjük.

Észrevehető, hogy az LOO és a VSA egyfajta keveréke; pontosabban az LOO a KCV speciális esete $k = n$ esetben.

Előnyei:

1. Számítási gyorsaság;

2. Az LOO-nál jobb becsléseket ad a tesztelési hibára;

3. Az LOO-nak nagyobb a hibavarianciája (mivel a hibák erősen korreláltak egymással, de kisebb a torzítása);

4. A KCV $k = 5$ vagy $k = 10$ esetén olyan tesztelési hibákat ad, ami nem szenved a túlzott torzítástól vagy varianciától.

1.2.4. Monte Carlo Cross-Validation, MCCV, $MCCV(n_v)$

Annyiban különbözik az előzőtől, hogy a konstrukciós és validációs halmazra való felosztást minden alkalommal, a többi alkalomtól függetlenül, véletlenszerűen tesszük, a módszer generalizációs hibája pedig a kapott generalizációs hibák átlaga.

1.2.5. Balanced Incomplete Cross-Validation 'kiegyensúlyozott nemteljes', BICV, $BICV(n_v)$

A $BICV$-módszert az $CV(1)$ módszer aszimptotikus helytelenségének és konzervatív voltának a kijavítására fejlesztették ki. Lényege, hogy az illesztést és a validációt nem az összes különböző felosztásra, hanem csak azok egy jól megválasztott részére végezzük el. Ezt a részt az ún. 'egyensúlyi feltételekkel' választjuk ki, és a 3.4.2. fejezetben tárgyaljuk. Az approximált $CV(n_v)$ módszer ($APCV(n_v)$) a speciális esete bizonyos feltételek teljesülése esetén. Ezt a 3.4.3. fejezetben tárgyaljuk.
Az $APCV$ előnyei:

1. konzisztens;

2. kevesebb számítást igényel, mint a $BICV$ vagy az $MCCV$.

1.3. A keresztvalidációs módszer szerinti modell választás menete

A keresztvalidációs megközelítésben a modellválasztás két nagy fázisa a tervezés és az értékelés, a tervezés két nagy fázisa pedig az optimalizálás és a választás. A validációs megközelítés lényege, hogy mindhárom fázist az adathalmaz külön részein végezzük el. Ennek érdekében tehát az összes megfigyelésből álló $\mathcal{A}$ adathalmazt valamilyen módszer és szempont szerint 3 diszjunkt részre osztjuk: a $\mathcal{K}$ konstrukciós halmazra, a $\mathcal{V}$ validációs halmazra és az $\mathcal{E}$ teszthalmazra (a $\mathcal{K}$ konstrukciós halmazt és a $\mathcal{V}$ validációs halmazt együttesen a $\mathcal{T}$ tanító halmaznak nevezzük).

A $\mathcal{K}$ konstrukciós halmazban lévő megfigyelésekre több különböző modellt illesztünk (ezáltal "hangoljuk" a modellparamétereket), az illesztett modelleknek megbecsüljük a generalizációs hibáját a $\mathcal{V}$ validációs halmazban lévő megfigyelések alapján, a generalizációs hiba becsléseinek összehasonlítása alapján kiválasztjuk az optimális modellt, végül pedig ennek az optimális modellnek a teljesítményét értékeljük az $\mathcal{E}$ teszthalmaz alapján.

Az adathalmaz tanító- és teszthalmazra való felosztását azért végezzük, hogy egy olyan modellt nyerjünk, aminek nagy a generalizációs képessége és a generalizációs hibát is megbízhatóan tudjuk értékelni, a tanító halmaz konstrukciós és validációs halmazra való felosztását pedig azért végezzük, hogy a létrejövő modell generalizációs képessége a lehető legnagyobb legyen.

Ez tehát a keresztvalidáció működési elve. A dolgozat további részében arra mutatunk példákat, hogy a keresztvalidáció milyen problémák megoldásában lehet hasznos.

2. fejezet

Modellválasztás keresztvalidációval általánosan

2.1. Modellválasztás menete általánosan

Tegyük fel, hogy egy $p(x, y)$ együttes valószínűségi eloszlásból vett $\mathcal{K} = \{x(k), y(k)\}_{k=1}^{N}$ adathalmazra egy $f(x; \omega)$ függvény által leírt modellt szeretnénk illeszteni (vagyis az y értékeket az x értékek alapján akarjuk közelíteni az $f(x; \omega)$ függvény alapján, ahol az ω a modellparaméterekből álló, p dimenziós vektor). Az y érték x pontokon vett $f(.; \omega)$ általi közelítését (predikcióját, előrejelzését) $\hat{y}(k) = f(x(k); \omega)$-val jelöljük. A tényleges és a predikciós érték közötti különbséget egy ℓ veszteségfüggvénnyel mérjük és predikciós hibának nevezzük. Ilyen ℓ függvény lehet például a log-likelihood függvény vagy akár egyszerűen a négyzetes eltérés: $(\ell(y, \hat{y}) = \|y - \hat{y}\|^2)$.

Ekkor a modellt egy $C(\mathcal{K}, \omega)$ költségfüggvény minimalizálásával szokták illeszteni:

$$C(\mathcal{K}, \omega) = S(\mathcal{K}, \omega) + R(\omega) = \sum_{(x,y)\in\mathcal{K}} \ell(y, f(x; \omega)) + R(\omega)$$

Ez a költségfüggvény két tag összege: az $S(\mathcal{K}, \omega)$ tag az illesztett modell jóságát méri, az $R(\omega)$ tag pedig a választott modell bonyolultságát méri. Az $R(\omega)$ például a paraméterszám, vagy annak valamely monoton növekvő függvénye. Az illesztett modell a paramétervektor becsült értékével foglalható össze: $\hat{\omega} = \arg\min_{\omega} C(\mathcal{K}, \omega)$.

2.1.1. Definíció. *Egy $f(.; \omega)$ modell generalizációs hibájának egy (x, y) jövőbeli független megfigyelés várható veszteségét nevezzük, ami nem más, mint az $\ell(y, f(x; \omega))$ veszteségfüggvénynek az (x, y) adatokon vett, $p(x, y)$ eloszlás szerinti várható értéke. Képlettel:*

$$G(\hat{\omega}) = \mathbb{E}_{x,y}(\ell(y, \hat{y})) = \int \ell(y, \hat{y}) p(x, y) dxdy$$

2.1.2. Definíció. *Egy $f(.;\omega)$ modell átlagos generalizációs hibája nem más, mint a $G(\hat{\omega})$ értékeknek az összes lehetséges $\mathcal{K}$ halmazon vett átlaga. Képlettel:*

$$\Gamma = \mathbb{E}_{\mathcal{K}}(G(\hat{\omega})) = \int G(\hat{\omega})p(\mathcal{K})d\mathcal{K}$$

A Γ tehát egy elméleti érték, a gyakorlatban nem ismert, csak becsülni lehet. A modell optimalizálása azt jelenti, hogy a $\mathcal{V}$ validációs halmaz alapján minimalizáljuk a Γ generalizációs hiba tapasztalati becsléseit. Végül pedig az $\mathcal{E}$ teszthalmaz az eredményül kapott modell generalizációs hibájára egy torzítatlan tapasztalati becslést ad.

A modellválasztás szokványos menete, hogy előbb a $\mathcal{T}$ adatrész $\mathcal{K}$ részét felhasználva kiválasztunk egy $\hat{\omega}_{\mathcal{K}}$ paramétert. Majd ezt az $f(.,\hat{\omega}_{\mathcal{K}})$ modellt értékeljük a modell generalizációs hibájának a $\mathcal{T}$ adatrész $\mathcal{V}$ részén mutatkozó hibái alapján becsülve. Addig keresünk újabb és újabb $f(.,\hat{\omega}_{\mathcal{K}})$ modelleket, míg ez utóbbi szempont szerint optimális $\hat{\omega}$ paramétert nem találunk. Végül a $f(.,\hat{\omega})$ modell generalizációs hibáját az $\mathcal{E}$ adatrész alapján becsüljük. Ugyanis mivel a $\mathcal{V}$ adatrészt felhasználtuk a modell illesztése során, csak ez utóbbi $\mathcal{E}$ adatrész alapján becsült generalizációs hiba lehet torzítatlan becslése a tényleges generalizációs hibának.

2.2. Az optimális felosztási hányados meghatározása keresztvalidációval

A bevezetőben láthattuk, hogy a keresztvalidációs megközelítésben a modellválasztás két nagy fázisa a tervezés és az értékelés: a modellparaméterek hangolásával kiválasztjuk az optimális modellt a $\mathcal{T}$ tanító halmazon a generalizációs hiba becsléseinek segítségével, majd pedig az optimális modellt értékeljük az $\mathcal{E}$ teszthalmazon az átlagos négyzetes hiba (mean squared error, MSE) segítségével. Azt a modellt szeretnénk kiválasztani (az a modell optimális), amelyre a generalizációs hiba becslése minimális (vagyis $\omega^* = \arg\min_\omega G(\omega)$). De a kiválasztott modell és a generalizációs hiba becslése is függ attól, hogy hány megfigyelésre illesztjük a modellt, illetve hány megfigyelést különítünk el tesztelésre.

Jelöljük $\gamma \in [0,1]$-val azt, hogy a megfigyelések hányad részét különítjük el a $\mathcal{E}$ részbe (a modell értékelésére) és nevezzük ezt felosztási hányadosnak (tehát $\gamma = \frac{n_e}{n}$)! Gyakorlati okok miatt γn csak egész szám lehet (vagyis $\gamma = i/n$, ahol $i = 1, ..., n-1$). Azt a γ értéket, amelyre a generalizációs hibára adott becslés minőségének ellenőrzésére használt MSE-mennyiség minimális, optimális felosztási hányadosnak nevezzük és γ_{opt}-tal jelöljük (vagyis $\gamma_{opt} = \arg\min_\gamma MSE(\gamma)$). Ez az optimális γ érték függ attól is, hogy melyik

keresztvalidációs módszert használjuk. Ebben a fejezetben a γ_{opt} meghatározásával foglalkozunk a különböző keresztvalidációs módszerek esetén.

Jelölje $\hat{\omega}$ a $\mathcal{T}$ adatrész alapján illesztett modellparamétert, jelölje $G(\hat{\omega})$ ennek a $\hat{\omega}$ modellparaméternek a tényleges generalizációs hibáját, $\hat{G}_{HO}(\hat{\omega})$ pedig a $\hat{\omega}$ modellparaméter generalizációs hibájának az $\mathcal{E}$ adatrész alapján vett, HO-módszer szerinti becslését. Legyen ω^* az a paraméter, amelyre a modell $G(\omega^*)$ generalizációs hibája minimális. Jelölje $\mathbb{E}_{\mathcal{A}}$ a megfelelő eltérésnégyzetek várható értékét a rendelkezésre álló $\mathcal{A}$ adatok szerint.

2.2.1. A HO-módszer

A HO-módszer esetén a generalizációs hiba HO-becslése a predikciós hibák $\mathcal{E}$ halmazon vett átlaga, az MSE pedig a HO szerint becsült generalizációs hiba és az optimális modell generalizációs hibája közti eltérésnégyzet várható értéke. Képlettel:

$$\hat{G}_{HO}(\hat{\omega}) = \frac{1}{n_e} \sum_{k \in \mathcal{E}} \ell(y(k), \hat{y}(k))$$

és

$$MSE_{HO} = \mathbb{E}_{\mathcal{A}}(\hat{G}_{HO}(\hat{\omega}) - G(\omega^*))^2$$

Ez az MSE_{HO} mennyiség felbontható két tagra: egy varianciatagra: $(\hat{G}_{HO}(\hat{\omega}) - G(\hat{\omega}))^2$ és egy torzítástagra: $(G(\hat{\omega}) - G(\omega^*))^2$. A varianciatag a HO-becslés megbízhatóságát méri, a torzítástag pedig a modell fölös generalizációja. A varianciatag onnan származik, hogy a HO-módszerrel becsült modell generalizációs hibáját az $\mathcal{E}$ adatrész alapján csak becsülni tudjuk, a torzítástag pedig onnan, hogy a HO-módszerrel becsült modell generalizációs hibája nem feltétlen minimális. A γ csökkenésének hatására a varianciatag nő, a torzítástag pedig csökken. Ennek az a magyarázata, hogy az $\mathcal{E}$ méretének csökkenésére a generalizációs hiba $\hat{G}_{HO}(\hat{\omega})$ becslése romlik, az $\mathcal{E}$ növekedése viszont a $G(\hat{\omega})$ javulásával (a torzítás csökkenésével) jár.

Példaként vegyük az egyik lehető legegyszerűbb modellt!
Vizsgáljuk azt az esetet, amikor a megfigyelések $y \sim \mathcal{N}(\mu, \sigma^2)$ eloszlásúak, ismert σ szórással és ismeretlen — az adatokból becsülendő — μ várható értékkel. Ekkor az optimális modell paraméterének a generalizációja σ^2, egy tetszőleges $\dot{\omega}$ paraméter generalizációs hibája pedig $G(\hat{\omega}) = \sigma^2 + (\mu - \dot{\omega})^2$.

Ebben az esetben hosszú, ámde elemi számolássorral belátható, hogy az a γ érték, amelyre a HO-módszerrel nyert generalizációs hiba a legkisebb, a következő:

$$MSE_{HO}(\gamma) = \frac{2\sigma^4}{n\gamma}\left(1 + \frac{2}{(1-\gamma)n}\right) + \frac{3\sigma^4}{(1-\gamma)^2 n^2}$$

ahol az

$$A = -324n^2 - 144n + 8 + 12n\sqrt{3(243n^2 + 472n - 28)}$$

konstans mellett a

$$\gamma_{opt} = 1 - \frac{8}{A^{1/3}} + \frac{A^{1/3}}{6n} + \frac{2}{3nA^{1/2}} + \frac{1}{3n} \ .$$

Ebből a képletből leolvasható, hogy az $1 - \gamma_{opt} = O(n^{1/3})$, ha $n \to \infty$, vagyis az optimális felosztási hányados — meglehetősen lassan — tart az 1-hez. Ami azt jelenti, hogy ahhoz, hogy egy pontos HO-becslést nyerjünk a generalizációs hibára, az adatok zömét aszimptotikusan a validációra kell fenntartani. Ezt a következtetést az általunk elvégzett szimulációs kísérlet is megerősíti.

2.2.2. A KCV-módszer

Mint ahogy az 1.5.3. alfejezetben már utaltunk erre, a KCV-módszer az $\mathcal{A}$ adathalmazt k db közel egyenlő diszjunkt részhalmazra bontjuk ($\mathcal{A} = \cup_{j=1}^{k} \mathcal{E}_j$) és mindegyik részhalmazon értékeljük a többi ($\mathcal{T}_j = \mathcal{A} \setminus \mathcal{E}_j$) adatra illesztett modellt. Az $\mathcal{E}_j$ nélkül illesztett modellt $\hat{y}^{-j}$-vel jelöljük.

A KCV-becslés a generalizációs hiba k db becslésének az átlaga:

$$\hat{\Gamma}_{KCV} = \frac{1}{n}\sum_{j=1}^{k}\sum_{k\in\mathcal{E}_j} \ell(y(k), \hat{y}^{-j}(k))$$

Az MSE pedig a HO-hoz hasonlóan a KCV szerint becsült generalizációs hiba és az optimális modell generalizációs hiba közti eltérésnégyzet várható értéke, ami hasonlóképpen egy variancia- és torzítástagból áll. Képlettel:

$$MSE_{KCV} = \mathbb{E}_{\mathcal{A}}(\hat{\Gamma}_{KCV} - G(\omega^*))^2$$

A számítások elvégzése után az MSE értékére az alábbi képletet kapjuk:

$$MSE_{KCV}(\gamma) = \begin{cases} \frac{\sigma^4(2\gamma^3 n - 2\gamma^2 - 6n\gamma^2 + 7\gamma + 6n\gamma - 7 - 2n)}{n^2(\gamma-1)^2} \\ \frac{\sigma^4(-4n\gamma^2 - 9\gamma + 8 + 2n\gamma + 2\gamma^2 + 2\gamma^3 n)}{n^2(\gamma-1)^2\gamma} \end{cases}$$

Mivel az $MSE_{KCV}(\gamma)$ γ szerinti deriváltja pozitív minden $0 \le \gamma \le 1$ és n értékre, ezért MSE_{KCV} a $CV(1)$-módszer (LOO) esetében lesz minimális ($\gamma_{opt} = 1/n$) függetlenül az n méretétől. Ezt a következtetést megerősíti az általunk elvégzett szimulációs kísérlet is. Megfigyelhető az az érdekesség, hogy ezeknek a görbéknek a meredeksége $\gamma = 1/2$-re nemfolytonos, ami annak köszönhető, hogy ilyen γ érték körül változnak át az az átfedésben lévő konstrukciós halmazok átfedésben lévő teszthalmazokká.

2.2.3. Az MCCV-módszer

Mint ahogy az 1.5.5. alfejezetben már utaltunk erre, a MCCV-módszer újramintavételezi a teszthalmazokat úgy, hogy véletlenszerűen kiválaszt $n_e = n\gamma$ mintát teszthalmaznak, a többit pedig konstrukciós halmaznak. Ez legfeljebb $k \leq \binom{n}{n_e}$-szor ismételhető meg. Mindegyik permutáció esetén egy $\hat{\omega}_j$ paraméterekkel rendelkező $\hat{y}_j$ modellt illesztünk, aztán pedig kiszámítjuk a generalizációnak a k db teszthalmazon kiszámított tapasztalati becslését. Így kapjuk:

$$\hat{\Gamma}_{MCCV} = \frac{1}{k} \sum_{j=1}^{k} \hat{G}(\hat{\omega}_j) \ ,$$

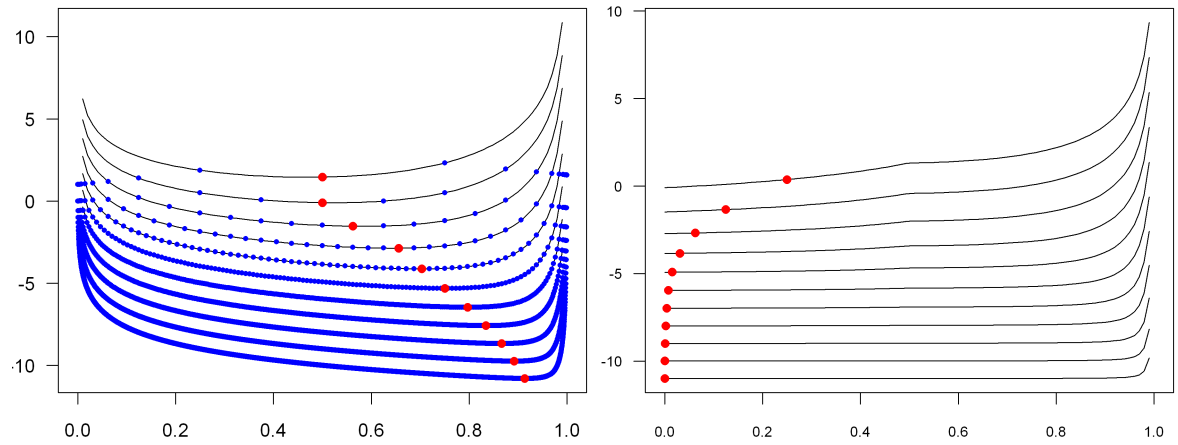
a MSE pedig ugyanaz, mint KCV esetén.

A számítások elvégzése után azt kapjuk, hogy az $MCCV$ a KCV-hez képest egyenletesen alacsonyabb (vagy egyenlő) MSE-becsléseket ad (ha k elég nagy). Mivel a KCV esetén a minimális MSE-t mindig LOO esetén érjük el, ezért ez a minőségi eredmény itt is változatlan marad.

2.3. Szimulációk

2.3.1. 1. szimuláció

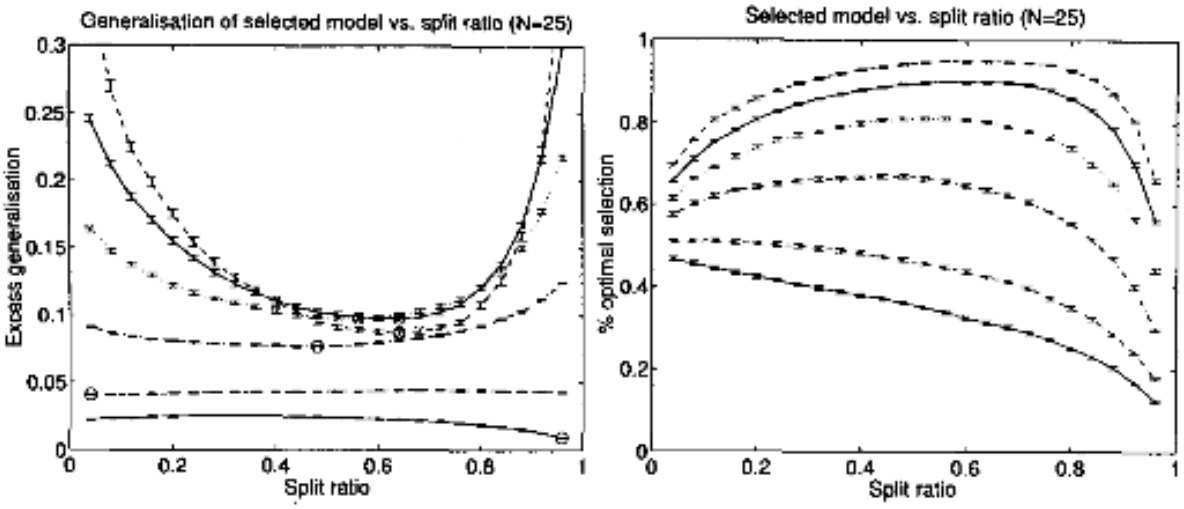
Az összes görbét átlagoljuk az adat 40000 replikációján az $n = 4, 8, 16, 32, 64$ mintaméretekre (az ábrán fentről lefelé), és γ-t ábrázoljuk MSE függvényében. Az eredmények közül a bal oldali ábra a HO-módszerre, a jobb oldali ábra pedig KCV-módszerre adja meg a kapott pontokat. A tömött karika a görbék minimumát jelzi, a hibahatárok pedig kétszeres standard szórás szerintiek.



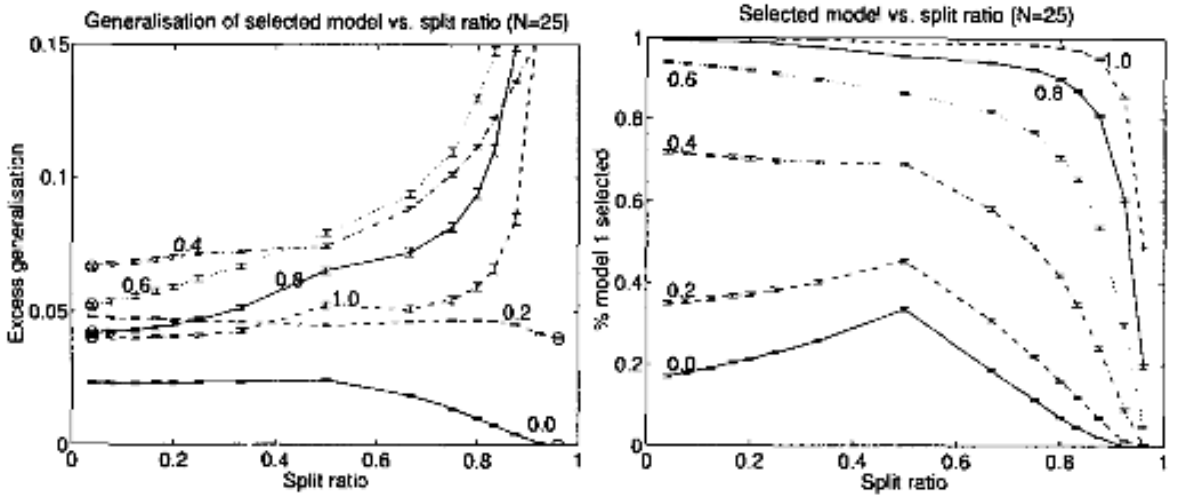
2.3.2. 2. szimuláció

A modellválasztás feladatának megkönnyítésére is elvégzünk egy szimulációs kísérletet. Most egyetlen $n = 25$ elemből álló mintát használunk. Az eredmények csak a $\theta = \mu\sqrt{n}/\sigma$ normalizált változó függvénye lesznek, és θ-t a μ értékén keresztül változtatjuk. Az n db megfigyelést az $\mathcal{N}(\mu, 1)$ eloszlásból vesszük, és most is minden eredményt átlagolunk az n megfigyelés 40000 replikációján.

A 2. ábra mutatja a HO esetén eredményül kapott modell generalizációját a μ növekvő értékeire (alul $\mu = 0$, felül $\mu = 1$ 0,2-es növekményekkel, a karika jelzi a minimumot). A bal oldali ábra a γ függvényében a HO-becslés fölös generalizációját, a jobb oldali ábra a helyesen kiválasztott modellek százalékarányát mutatja.



3. ábra: modellválasztás KCV esetén $\mu = 0$-tól 1-ig 0,2-es növekményekkel. A bal oldali ábra a γ függvényében a KCV-becslés fölös generalizációs hibáját, a jobb oldali ábra pedig a helyesen kiválasztott modellek százalékarányát mutatja.



2.4. A szimulációk értelmezése

2.4.1. 1. szimuláció

Az ábrákról leolvasható, hogy HO esetén az optimális felosztási hányados (γ_{opt}) az n növekedésével 1 felé tart (de lassan), míg KCV esetén mindig $\gamma_{opt} = \frac{1}{n}$. Ez azt jelenti, hogy minél több megfigyelésünk van, a HO esetén a megfigyeléseknek annál nagyobb hányadát kell a validációra fenntartani (és így annál kisebb hányadát az illesztésre), míg KCF esetén a megfigyelések számától függetlenül elegendő egyetlen megfigyelést félretenni a validációra, a többire illeszthetünk.

Az ábrákon az is észrevehető, hogy az n növekedésével az MSE-görbék kilaposodnak. Ez azt jelzi, hogy a közel optimális felosztási hányadosok egy széles intervallumban helyezkednek el.

2.4.2. 2. szimuláció

HO-módszer

A legnagyobb felosztási hányados ($\gamma = \frac{n-1}{n}$) a μ kis értékeire optimális. A 2. ábra jobb oldali fele jól illusztrálja azt, hogy ezekben az esetekben majdnem mindig a minimálmodellt választjuk ki, mert a minimálmodell a teljes modellnél jobb becslést ad.

Viszont $\theta = 1$-re, azaz $\mu = 0.2$-re a $\gamma = 1/n$ (vagyis az LOO) lesz optimális. Ekkor egy ún. fázisátmenet történik. A 2. ábra bal oldali felén a megfelelő görbe majdnem lapos, és annál a pontos értéknél, ahol a fázisátmenet történik, a γ nagy és kis értékei is az optimum hibahatárán belül vannak. A μ növekedésével az optimális felosztási hányados most is nagyon lassú, aszimptotikus ütemben tart az 1-hez.

A 2. ábra jobb oldali fele azt mutatja, hogy ez azért van, mert a leghelyesebb modellt adó felosztási hányados az 1-hez tart. A μ növekedésével a görbék egyre laposabbá válnak, ami a várakozásoknak megfelelően azt jelzi, hogy a γ majdnem minden választásával a helyes modellt választjuk ki, s így közel optimális generalizációt kapunk.

KCV-módszer

A felosztási hányados legnagyobb $\gamma = \frac{n-1}{n}$ értéke most is a μ kis értékeire optimális, és ebben az esetben is van egy átmenet a minimál modell és a leginkább konzisztens modell

között $\gamma_{opt} = 1/2$-re a $\mu = \sigma/\sqrt{n} = 0.2$ körül.

A KCV viszont különbözik a HO-tól abban, hogy a μ kis értékeire a leginkább konzisztens becslést nem $\gamma = 1/n$ (vagyis LOO), hanem $\gamma = 1/2$ adja! Sőt a KCV-nél van még egy átmenet: $\gamma_{opt} = 1/n$-hez (vagyis LOO-hoz) egy enyhén nagyobb értékre. Ennek a konstrukciós és a validációs halmaz közötti átfedés az oka. Ez a második átmenet akkor történik, amikor az LOO a $KCV(2)$-nél helyesebb modelleket kezd adni.

További különbségek a KCV- és a HO-módszer által adott becslések között: a KCV esetén a μ növekedésével az LOO optimális marad; a minimális fölös generalizációs hiba (minimum excess generalization error) alacsonyabb; a helyesen választott modellek aránya gyorsabb ütemben tart az 1-hez.

Megjegyzendő még, hogy az aszimptotikusan optimális felosztási hányados $1/n$, mivel a fázisátmeneti küszöbök fordítottan arányosak n-nel minden $\mu \neq 0$-ra.

$MCCV$-módszer

Az $MCCV$ eredményei most is hasonlók a KCV-éhez, s az eredményből levont kvalitatív következtetés is azonos: A γ optimáélis értéke $\gamma_{opt} = 1/n$, vagyis az LOO eljárás az optimális. Az viszont eltér a KCV-től, hogy $\gamma = 1/2$-nél nincs nemfolytonosság. Ez annak köszönhető, hogy a átlagolási stratégia jobb a közbülső felosztási hányadosokra. Tehát csak egy γ-átmenet van: az egyik szélsőértékről a másikra.

3. fejezet

Modellválasztás keresztvalidációval lineáris regresszió esetén

3.1. Modellválasztás menete a lineáris regresszió fix és véletlen modellje esetén

Legyen y magyarázó változó és x_1, ..., x_p magyarázandó változók. Legyen $(y, \underline{x})' = (y, x_1, ..., x_p)'$. A lineáris regresszió egy olyan paraméteres regressziós modell, amely feltételezi az y magyarázó változó és az x_1, ..., x_p magyarázandó változók közti (paramétereiben) lineáris kapcsolatot. A lineáris kapcsolat a következőképpen fejezhető ki:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + e,$$

ugyanez zártabb formában felírva:

$$y = X\beta + e,$$

ahol:

- n a mintanagyság (mérések/megfigyelések száma) és $p+1$ a magyarázó változók száma (konstanssal együtt);
- $y \in \mathbb{R}^{n\times 1}$ a magyarázandó változó értékére vonatkozó n megfigyelést tartalmazó $n \times 1$ méretű oszlopvektor;
- $X \in \mathbb{R}^{n\times(p+1)}$ a p magyarázó változó értékére vonatkozó n db $x_k \in \mathbb{R}^{n\times 1}$ megfigyelést tartalmazó mátrix, amit tervmátrixnak nevezünk. Ha a tervmátrixban lévő értékek a kísérlet végzője által rögzített (vagyis fix) értékek, akkor fix hatás modellről beszélünk. Ha viszont a tervmátrix a véletlentől függ, méghozzá úgy, hogy $(y, \underline{x})' = (y, x_1, ..., x_p)'$ többdimenziós normális eloszlású $(\mu_y, \underline{\mu_x})' = (\mu_y, \mu_{x_1}, ..., \mu_{x_p})'$ várható

értékkel és $\begin{pmatrix} \sigma_{y,y} & \sigma_{y,x} \\ \sigma_{x,y} & \sigma_{x,x} \end{pmatrix}$ partícionált kovariancia mátrixszal, akkor véletlen hatás modellről beszélünk;

- $\beta \in \mathbb{R}^{(p+1)\times 1}$ a modellparamétereket tartalmazó p dimenziós vektor (a paraméterek itt azokat a súlyokat jelentik, amelyekkel az egyes magyarázó változók a magyarázandó változó értékét közelítő lineáris függvényben szerepelnek);

- $e \in \mathbb{R}^{n\times 1}$ pedig a regresszió hibáit tartalmazó n elemű vektor (amely egy $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ véletlen mennyiség aktuális értéke).

A β és az e ismeretlen, ezeket az adatokból kell becsülnünk. A lineáris regresszió becslése során a β paramétervektort becsüljük a rendelkezésre álló mintából úgy, hogy az össznégyzetes hibát minimalizálja: $\hat{\beta} = \arg\min_\beta \|y - X\beta\|^2$. A legegyszerűbb becslési módszer a legkisebb négyzetek módszere. A legkisebb négyzetek módszere alapján a paraméterre adott becslés az ismert képlet szerint: $\hat{\beta} = (X^T X)^{-1} X^T y$. A paraméterre adott becslés alapján az előrejelzési érték (predikciós érték) $\hat{y} = X\hat{\beta}$ (s így $\mathbb{E}(y) = X\beta$), az előrejelzési hiba (predikciós hiba, reziduális) pedig $\hat{e} = y - X\hat{\beta}$ (s így $\mathbb{E}(e) = 0$).

Az $\hat{y}$ és az $\hat{e}$ becsült értékeket kifejezhetjük az X képterére vetítő projekciós mátrix és a képtérre merőleges komponenst előállító annihilátor segítségével (ami egyébként maga is egy projekció). Ha $P = X(X^T X)^{-1} X^T$ az X oszlopai által kifeszített térre vetítő projekciós mátrix és $M = I_n - P$ az X-re merőleges térre vetítő annihilátor mátrix, akkor

$$\hat{y} = Py,$$

$$\hat{e} = My = Me.$$

A P projekciós mátrix i-edik átlóelemét w_i-vel jelöljük.

3.2. Mintaméret meghatározása

3.2.1. Néhány eloszlás

Ezekre az eloszlásokra a mintanagyságot meghatározó képletek megértéséhez van szükség. Az egyszerűség kedvéért a meghatározásokban szereplő eloszlás jelöléseket értelmezzük úgy, mintha azok véletlen mennyiségek volnának, az adott eloszlással!

Khí-négyzet eloszlás

A khí-négyzet eloszlás nem más, mint egy k dimenziós, standard normális eloszlású pontnak az origótól vett távolságnégyzetének az eloszlása. Egyetlen paramétere, a szabadságfok azt mutatja meg, hogy hány *független*, standard normális eloszlású mennyiség négyzetösszegének az eloszlásáról van szó. Vagyis:

$$\chi_k^2 \sim \sum_{j=1}^{k} \mathcal{N}^2(0,1).$$

t-eloszlás

A t-eloszlásnak egy paramétere van, a szabadságfok. E paraméter azt mutatja, hogy mennyi a nevezőjében szereplő khí-négyzet eloszlásnak a szabadságfoka. Ugyanis:

$$t_k \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi_k^2/k}}$$

E képlet úgy értendő, hogy a t eloszlású mennyiség eloszlása olyan, mint két olyan *független* véletlen mennyiség hányadosának az eloszlása, amelyek egyrészt standard normális, másrészt χ^2 eloszlású.

F-eloszlás

Az F-eloszlásnak két paramétere van, két szabadságfok. Annak a két, *független* χ^2 eloszlásnak a szabadságfoka, amelyek a hányadosaként az F eloszlás előáll. Azaz:

$$F_{n,m} \sim \frac{\chi_n^2/n}{\chi_m^2/m}.$$

Nemcentrális khí-négyzet eloszlás

A khí-négyzet eloszlásnak a szabadságfokon kívül egy paramétere van, a nemcentralitási paraméter. E nemcentralitási paraméter értéke a meghatározásához felhasznált 1 szórású normálisok várható értékeiből képzett négyzetösszeggel egyenlő. Vagyis:

$$\chi_{n,\delta^2}^2 \sim \sum_{j=1}^{k} \mathcal{N}^2(\mu_j, 1),$$

akkor, ha a $\delta^2 = \sum_{j=1}^{k} \mu_j^2$, vagyis a véletlen pont *várható értékének* az origótól vett távolságnégyzete.

Nemcentrális t-eloszlás

A nemcentrális t-eloszlásnak két paramétere van, a szabadságfok és a nemcentralitási paraméter. Ez utóbbi paraméter azt mutatja, hogy mennyi a *számlálójában* szereplő normális eloszlás várható értéke. Ugyanis:

$$t_{k,\mu} \sim \frac{\mathcal{N}(\mu, 1)}{\sqrt{\chi^2_k/k}}.$$

Vagyis egy olyan eloszlás, amelyiknek a nevezőjében egy centrális χ^2 eloszlás van.

Nemcentrális F-eloszlás

A nemcentrális F-eloszlásnak a két szabadságfok paraméteren kívül egy paramétere van, a nemcentralitási paraméter. Ez a nemcentralitási paraméter egyenlő annak a nemcentrális χ^2 eloszlásnak a paraméterével, amelyik a számlálójában szerepel. Ugyanis:

$$F_{n,m,\delta^2} \sim \frac{\chi^2_{n,\delta^2}/n}{\chi^2_m/m}.$$

Vagyis ennek az eloszlásnak a nevezőjében is *centrális* χ^2 eloszlás van.

Mindhárom eloszlás, a nemcentrális esetet is figyelembe véve elérhető az **R** alaprendszeréhez tartozó `stats` csomag megfelelő függényei segitségével:

- a χ^2 eloszláshoz tartozó függvények:
  `dchisq(x,df,ncp)`, `pchisq(q,df,ncp)`, `qchisq(p,df,ncp)`, `rchisq(n,df,ncp)`
- a t eloszláshoz tartozó függvények:
  `dt(x,df,ncp)`, `pt(q,df,ncp)`, `qt(p,df,ncp)`, `rt(n,df,ncp)`
- az F eloszláshoz tartozó függvények:
  `df(x,df1,df2,ncp)`, `pf(q,df1,df2,ncp)`, `qf(p,df1,df2,ncp)`, `rf(n,df1,df2,ncp)`.

Itt `d`-vel kezdődnek a sűrűségfüggvények, `p`-vel az eloszlásfüggvények, `q`-val a kvantilisfüggvények és `r`-rel az adott eloszlás szerint véletlen számot generáló eljárások. Az argumentumokban az `x` az eloszlás értelmezési tartományának egy pontja, `p`$\in [0,1]$ egy valószínűség, `q` egy kvantilis, `n` a generálandó véletlen számok számossága. A `df` jelöli a szabadságfokokat, az `ncp` pedig a nemcentralitási paraméter értékét.

3.2.2. A regressziós modell mintanagyságának elméleti meghatározása a korreláció függvényében

A keresztvalidáció használható a regressziós függvény illeszkedésének tesztjeként is. Ebben az esetben a keresztvalidáció során veszünk egy második véletlen mintát és kiszámoljuk az új megfigyelt függő változó és az új magyarázó változók azon lineáris kombinációjának korrelációját, amelynek együtthatóit az eredeti minta alapján nyertünk. Az eredményül kapott korrelációt $r_c(\hat{\beta})$-vel jelöljük.

Az $r_c(\hat{\beta})$ tehát értékelése a mintából származtatott egyenlet érvényességének, és egy becslése a $\hat{\beta}$ paraméter mellett a

$$\varrho_c(\hat{\beta}) = \frac{\sigma'_{xy}\hat{\beta}}{(\sigma_{yy}\hat{\beta}'\Sigma'_{xx}\hat{\beta})^{1/2}}$$

populációs paraméternek, amit röviden ϱ_c-vel fogunk jelölni. Ha ϱ jelöli a populációs értéket, akkor tekintettel annak maximális voltára, bizonyos, hogy $\varrho_c \leq \varrho$.

Ha ϱ_c mintabeli eloszlását fel tudnánk írni, mint az n függvényét, akkor a

$$P(\varrho - \varrho_c \leq \epsilon) = \gamma$$

képlet alapján adott ϵ és γ mellett a keresett n mintanagyság, — a regressziós függvény jóságának kellő szintű meghatározásához szükséges mintaelemszám — megadható volna.

Azonban ϱ_c sűrűségfüggvényére irányuló minden eddigi kísérlet hiábavalónak bizonyult. Viszont a ϱ_c^2-é meghatározható. Belátható, hogy a fontosabb esetekben ez elégséges is. Így most ϱ_c^2 sűrűségfüggvényének meghatározásával fogunk foglalkozni.

Ha ϱ_c^2 eloszlását egy lineáris transzformációval egyszerűsítjük, akkor ϱ_c^2 kifejezhető korrelálatlan változók négyzetösszegeinek függvényeként:

$$\varrho_c^2 = \frac{W_1^2 \left\|B\right\|^2}{W_1^2 + \sum_{i=2}^{p} W_i^2} = \frac{\varrho^2}{1 + \sum_{i=2}^{p} \frac{W_i^2}{W_1^2}}$$

Ez átírható a következő alakba:

$$\varrho_c^2 = \frac{\varrho^2}{1 + \frac{\chi^2_{p-1,\delta_2=0}}{\chi^2_{1,\delta_1}}}$$

ahol a χ^2_{p-1} és χ^2_{1,δ_1} változóknak független khí-négyzet illetve nemcentrális khí-négyzet eloszlása van, ahol a nemcentralitási paraméter értéke:

$$\delta_1 = |EW_1| = |\varrho|\sqrt{\frac{n-p-2}{1-\varrho^2}}.$$

A szabadságfokokkal való szorzás és osztás után végül ezt kapjuk:

$$\varrho_c^2 = \frac{\varrho^2}{1 + \frac{p-1}{F_{1,p-1,\delta_1}}}$$

ahol $F_{1,p-1,\delta_1}$ egy nemcentrális F-eloszlás δ_1 nemcentralitási paraméterrel.

Tehát összefoglalva, a ϱ_c^2 keresett eloszlása:

$$F_{\varrho_c^2}(\lambda) = P(\varrho_c^2 \leq \lambda) = F_{1,p-1,\delta_1}\left(\lambda(p-1)/(\varrho^2 - \lambda)\right)$$

ahol a $F_{1,p-1,\delta_1}$ az 1 és $p-1$ szabadságfokú, δ_1 nemcentrális F-eloszlást eloszlásfüggvénye.

Így ϱ_c^2 eloszlása táblázatba foglalható a nemcentrális t-eloszlás nyilvános táblázatainak, statisztikai programrendszerekben fellelhető szubrutinjainak felhasználásával.

A szükséges mintaméret a random modell esetén:

$$n_r = \frac{(1-\varrho^2)\delta_1^2}{\varrho^2} + p + 2,$$

a fix modell esetén pedig:

$$n_f = \frac{(1-\varrho^2)\delta_1^2}{\varrho^2}.$$

Vagyis a szükséges mintaméret a fix modell esetén $p+2$-vel kisebb mint a véletlen modell esetén.

3.2.3. A regressziós modell mintanagyságának gyakorlati meghatározása táblázattal

Az alábbi táblázat azt mutatja, hogy az imént bemutatott képletet alkalmazva, $p = 2$ magyarázó változó mellett, adott $\varrho = .05, .1, ..., .98$ korreláció esetén az $\epsilon = .01, ..., .20$ pontosság $99\%, ..., 40\%$ valószínűséggel, hány elemű minta alapján érhető el.

		.99	.95	.90	.80	.60	.40
.05	.01	634	369	261	160	72	31
	.03	213	124	88	54	23	8
.10	.01	601	350	248	152	68	30
	.03	203	119	85	53	25	12
	.05	123	73	52	33	16	7
.25	.01	501	292	207	127	57	25
	.03	170	100	71	45	22	11
	.05	104	62	45	29	15	8
	.10	53	32	24	16	9	6
	.20	29	18	14	10	5	4
.50	.01	336	196	139	86	39	18
	.03	115	68	49	31	16	9
	.05	70	42	31	20	11	7
	.10	37	23	17	12	7	6
	.20	20	13	11	8	6	5
.75	.01	170	100	72	45	22	11
	.03	59	36	27	18	10	6
	.05	37	23	18	12	8	5
	.10	21	14	11	8	6	5
	.20	12	9	7	6	5	4
.98	.01	17	12	9	7	5	5
	.03	8	7	6	5	4	4
	.05	7	6	5	5	4	4
	.10	5	5	5	4	4	4
	.20	5	4	4	4	4	4

Példa

A fenti táblázat felhasználásával egy konkrét alkalmazás során például az alábbi típusú következtetésekre juthatunk.

Ha egy olyan regressziót vettünk, amelynél a magyarázó változók száma 2, és amelynél a $\hat{\varrho}^2$ értéke .5, akkor a korrelációnégyzet keresztvalidációval nyert becslése a valódi korreláció érték négyzetét egy 68 elemű minta esetén az alábbi táblázatba foglalt módon és mértékben közelíti:

	véletlen	fix	modell esetén
max 1% eltéréssel	$\approx .72$	$\approx .74$	
max 3%	$\approx .95$	$\approx .95$	
max 5%	$\approx .99$	$\approx .99$	valószínűséggel

Azaz például véletlen modell esetén $P(\varrho^2 - \varrho_c^2 \leq .1) \approx .72$. Tehát véletlen modellt alkalmazva, a 68 elemű minta alapján vett becslés hibája 72% valószínűséggel kisebb mint 1%.

3.3. A lineáris regresszió modell előrejelzési hibája

Ebben a fejezetben az előrejelzési hiba eloszlását adjuk meg a lineáris regresszió

$$y = X\beta + e$$

fix modellje esetén. A keresztvalidációval elvének megfelelően a modellt nem a megfigyelések teljes (n elemszámú) halmazára illesztjük, hanem csak az első néhány (n_e db) megfigyelésre, majd pedig a modellparaméterre kapott becslést felhasználva a modell illeszkedését ellenőrizzük (a többi $n_v = n - n_e$ db) megfigyelésen úgy, hogy kiszámítjuk két új (azaz két egymástól és az eddigiektől is független) megfigyelés hibabecslésének a kovarianciáját.

Jelölje az y és az X első n_e sorát y_e és X_e, utolsó n_v sorát y_v és X_v. Az első n_e db megfigyelésre illesztett modell paraméterének becslése a legkisebb négyzetek módszerével:

$$\hat{\beta}_e = X_e(X_e^T X_e)^{-1} y_e \sim \mathcal{N}(\beta, \sigma^2 (X_e^T X_e)^{-1}).$$

Nézzük meg azt, hogy az így kapott modell mennyire jól (mekkora hibával) tudja előrejelezni az újabb (egymástól és az eddigi megfigyelésektől is független) megfigyeléseket, amit a

hibabecslések kovarianciájából tudunk meghatározni.

Ha (y_j, x_j) és (y_k, x_k) két új, a fenti modellnek megfelelő, független megfigyelés (a megfelelő mérések során keletkezett e_j és e_k 0 várható értékű véletlen hibával), akkor

$$y_j = x_j\beta + e_j \text{ és } y_k = x_k\beta + e_k \ ,$$

hiba nélküli értéküknek, a rendelkezésre álló $\hat{\beta}_e$ becslés alapján vett becslése pedig:

$$\hat{y_j} = x_j\hat{\beta}_e \text{ és } \hat{y_k} = x_k\hat{\beta}_e.$$

Tehát az ezek alapján nyerhető

$$\hat{e_j} = y - \hat{y_j} \ , \ \hat{e_k} = y - \hat{y_k}$$

hibabecslések kovarianciája:

$$cov_e(\hat{e_j}, \hat{e_k}) = E((y_j - \hat{y_j})(y_k - \hat{y_k})) = E((y_j - x_j\hat{\beta}_e)(y_k - x_k\hat{\beta}_e))$$

Vonjunk ki és adjunk hozzá az első tagban $x_j\beta$-t, a második tagban $x_k\beta$-t, hogy egy négytagú összeggé alakíthassuk, aztán pedig sok minden kiessen:

$$\begin{aligned} cov_e(\hat{e_j}, \hat{e_k}) &= E((y_j - x_j\beta + x_j\beta - x_j\hat{\beta}_e)(y_k - x_k\beta + x_k\beta - x_k\hat{\beta}_e)) \\ &= E((y_j - x_j\beta)(y_k - x_k\beta)) + E((y_j - x_j\beta)(x_k\beta - x_k\hat{\beta}_e)) + \\ &+ E((x_j\beta - x_j\hat{\beta}_e)(y_k - x_k\beta)) + E((x_j\beta - x_j\hat{\beta}_e)(x_k\beta - x_k\hat{\beta}_e)) \end{aligned}$$

Az első tagban csak az y_j és az y_k függ a véletlentől. $j = k$ esetén az értéke σ^2 (mivel a megfelelő egyenlet hibatagjának varianciájáról van szó), $j \neq k$ esetén pedig a várható értéket tényezőnként lehet számolni, mivel y_j és y_k függetlenek, s mivel mindkét tényezője 0 várható értékű, ezért a szorzat várható értéke is 0.

A második tagban csak az y_j és az $\hat{\beta}_e$ függ a véletlentől. A várható értéket itt is tényezőnként lehet számolni, mivel y_j és az $\hat{\beta}_e$ a feltételek szerint függetlenek egymástól. Mivel az első tényező várható értéke 0, ezért a szorzat várható értéke is 0.

Mivel második és a harmadik tag szimmetrikus helyzetűek, ezért ez a megfontolás vonatkozik a harmadik tagra is.

A negyedik tagban csak a $\hat{\beta}_e$ függ a véletlentől, és ennek várható értéke β, eloszlásának varianciáját pedig már korábban felírtuk. Ezeket figyelembe véve a negyedik tag értéke az átalakítások után:

$$\begin{aligned} E[(x_j\beta - x_j\hat{\beta}_e)(x_k\beta - x_k\hat{\beta}_e)] &= E[x_j(\beta - \hat{\beta}_e)x_k(\beta - \hat{\beta}_e)] = E(x_j(\beta - \hat{\beta}_e)(\beta - \hat{\beta}_e)^T x_k^T) \\ &= x_j E((\beta - \hat{\beta}_e)(\beta - \hat{\beta}_e)^T)x_k^T = x_j\sigma^2(X_e^T X_e)^{-1}x_k^T \end{aligned}$$

Mivel az első három tag értéke 0, ezért a becsült hibák kovarianciájának számításánál csak a negyedik tagnak van szerepe.

Tehát ha a két újabb megfigyelés hibájának becslését az (Y_e, X_e) adatok alapján nyert $\hat{\beta}_e$ becslés alapján számoljuk, akkor a becsült hibák kovarianciája:

$$cov_e(\hat{e_j}, \hat{e_k}) = \sigma^2(\delta_{jk} + x_j(X_e^T X_e)^{-1} x_k^T)$$

Ha pedig a az utolsó n_v megfigyelés hibájának becslését az első n_e megfigyelés alapján vett $\hat{\beta}_e$ becslés alapján számoljuk, akkor a hiba eloszlása:

$$Y_v - X_v\hat{\beta}_e \sim \mathcal{N}(0, \sigma^2(I + X_v(X_e^T X_e)^{-1} X_v^T)).$$

Itt az $I\sigma^2$ az egyes megfigyelések hibáinak felel meg, a $\sigma^2 X_v(X_e^T X_e)^{-1} X_v^T$ származik a modell illesztéséből.

3.4. Változószelekció keresztvalidációval

Tekintsük most ismét a

$$y = X\beta + e$$

lineáris regresszió modellt, az előbbi jelölésnek megfelelően. A β néhány komponense strukturálisan 0 lehet (azaz nem csak az esetlegesen becsült értéke alapján statisztikailag, hanem a tényleges értéke szerint is), így a figyelembe vétele csak feleslegesen bonyolultabbá teszi a modellt. Kompaktabbá szeretnénk tenni a modellt azáltal, hogy elhagyjuk a β felesleges 0 komponenseit. Azt viszont nem tudhatjuk a lineáris regresszió elvégzése előtt, hogy van-e a β komponensei között 0, és ha igen, mennyi, ezért inkább azt csináljuk, hogy az összes lehetséges módon elhagyjuk az X és a β néhány egymásnak megfelelő komponensét (tehát a β komponensei között lehet nulla és nemnulla egyaránt), és megnézzük a komponensek elhagyásának a hatását az eredményül kapott modell teljesítményére nézve.

Legyen egyes együtthatók elhagyásával szűkített, kompaktabbá tett modell alakja:

$$y = X_\alpha \beta_\alpha + e,$$

ahol az $\alpha \subseteq \{1, ..., p\}$ egy d_α elemű index részhalmaz. Mivel $\{1, ..., p\}$-nek $2^p - 1$ db nemüres részhalmaza van, ezért elvileg $2^p - 1$ db részmodell építhető. Jelölje az $\{1, ..., p\}$ halmaz hatványhalmazát, vagyis a $\{1, ..., p\}$ nemüres részhalmazainak a halmazát $\mathcal{A}$. Jelölje az $\alpha \in \mathcal{A}$ koordináta részhalmazhoz tartozó modellt $\mathcal{M}_\alpha$. Az α számosságát jelölő d_α számot

pedig nevezzük az $\mathcal{M}_\alpha$ modell dimenziójának.

Az előbbi jelölésrendszer analógiájára legyen $P_\alpha = X_\alpha(X_\alpha^T X_\alpha)^{-1}X_\alpha^T$ (projekciós mátrix, az X_α képterére való vetítés); $w_{i\alpha}$: a P_α projekciós mátrix i-edik átlóeleme; $\hat{\beta}_\alpha$: a β_α legkisebb négyzetek módszerével vett becslése mind az n megfigyelés figyelembe vételével

---

A cél annak az α_*-gal jelölt indexhalmaznak és a hozzá tartozó $\mathcal{M}_*$ modellnek a megtalálása, amelyikre α_* a β strukturálisan nemnulla együtthatóinak indexeiből áll.
Az $\mathcal{M}_\alpha$ modellek az $\mathcal{M}_*$ optimális modell szerint 2 kategóriába sorolhatóak:

- 1. kategória: az α_*-nak legalább egy eleme nem α-beli
- 2. kategória: az α_* mindegyik eleme α-beli

Az optimális modell nyilván II. kategóriájú, hiszen az I. kategóriájú modellek mindegyike hiányos. Ugyanakkor a II. kategóriában lévő modellek az optimálist kivéve mind túl bővek. Tehát az optimális modell II. kategóriájú, ugyanakkor a legkisebb dimenziójú a II. kategóriájú modellek közül.

Ezt az optimális modellt keresztvalidációval szeretnénk meghatározni.
Ezért a keresztvalidáció korábbiakban ismertetett elvének megfelelően a rendelkezésre álló n elemszámú (y, X) adathalmazt két diszjunkt részre bontjuk, az n_c elemszámú $\mathcal{K}$ konstrukciós halmazra és az n_v elemszámú $\mathcal{V}$ validációs halmazra. Az $\mathcal{M}_\alpha$ modellt a $\mathcal{K}$ konstrukciós adatrész alapján illesztjük, a generalizációs hibát pedig a $\mathcal{V}$ adatrészen számoljuk, mintha azok a jövőbeli, azaz a z_i értékek volnának.

A keresztvalidáció azt az $\mathcal{M}_\alpha$ modellt választja ki, amelyre a generalizációs hiba $\mathcal{V}$ validációs halmazon számolt tapasztalati becslése minimális. Mivel azonban a jelen feladatban a generalizációs hiba és a generalizációs hiba becslése is függ $\alpha \in \mathcal{A}$-tól, ezért egyáltalán nem egyértelmű az, hogy pontosan mit értünk itt generalizációs hiba és generalizációs hiba becslése alatt. Ezt definiáljuk a következőkben.
Legyen $\hat{\beta}$ az (y_i, x_i) adatok alapján illesztett lineáris regressziós fix modell paraméterének becslése, és legyen z_i egy újabb megfigyelés az x_i magyarázó értékek mellett. Ekkor a z_i megfigyelés előrejelzése a $\hat{\beta}$ alapján $x_i\hat{\beta}$, a z_i megfigyelés várható értéke pedig $x_i\beta$.

3.4.1. Definíció. *Az (y_i, x_i) adatok alapján illesztett lineáris regressziós fix modell generalizációs hibájának ASPE-becslése (ASPE: average squared prediction error, átlagos*

négyzetes predikciós hiba) az összes megfigyelésnek az előrejelzett értéküktől vett négyzetes hibájának az átlaga. Képlettel:

$$\hat{G}^{ASPE}(\hat{\beta}) = \frac{1}{n}\sum_i (z_i - x_i\hat{\beta})^2$$

.

3.4.2. Definíció. *Az (y_i, x_i) adatok alapján illesztett lineáris regressziós fix modell generalizációs hibájának CESPE-becslése (CESPE: conditional expected squared prediction error, feltételes várható négyzetes predikciós hiba) az összes megfigyelés várható értékének az előrejelzett értéküktől vett négyzetes hibájának az átlaga, figyelembe véve a megfigyelések e_i hibáinak a σ^2 varianciáját is. Képlettel:*

$$\hat{G}^{CESPE}(\hat{\beta}) = \sigma^2 + \frac{1}{n}\sum_i (x_i\beta - x_i\hat{\beta})^2$$

Legyen $\Delta_{\alpha,n} = \frac{1}{n}(\beta^T X^T)M(X\beta)$ a becsült hiba.

3.4.3. Definíció. *Az (y_i, x_i) adatok alapján illesztett lineáris regressziós fix modell $\Gamma_{\alpha,n}$ átlagos generalizációs hibájának (feltétel nélküli teljes várható négyzetes predikciós hibájának, overall unconditional expected squared prediction error) a*

$$\Gamma_{\alpha,n} = \sigma^2 + \frac{1}{n}\sigma^2 d_\alpha + \Delta_{\alpha,n}$$

értéket nevezzük.

Az átlagos generalizációs hiba tehát három komponensből tevődik össze: a megfigyelés hibájából (σ^2), a modellválasztásból származó bizonytalanságból ($\frac{1}{n}\sigma^2 d_\alpha$) és a becslési hibából ($\Delta_{\alpha,n} = \frac{1}{n}(\beta^T X^T)M(X\beta)$).

A becslési hibára, valamint az átlagos generalizációs hibára vonatkozóan az alábbi megállapításokat tehetjük:

- Minden II. kategóriájú $\mathcal{M}_\alpha$ modellre $\Delta_{\alpha,n} = 0$ (és így $\Gamma_{\alpha,n} = \sigma^2 + \frac{1}{n}\sigma^2 d_\alpha$);
- Minden I. kategóriájú $\mathcal{M}_\alpha$ modellre $\Delta_{\alpha,n} > 0$ és rögzített p mellett teljesül az is, hogy a $\liminf_{n\to\infty} \Delta_{\alpha,n} > 0$;
- Ha az α-ra az $\mathcal{M}_\alpha$ egy I. kategóriájú modell és a γ-ra az $\mathcal{M}_\gamma$ egy II. kategóriájú modell, akkor a generalizációs hibára teljesül, hogy $\frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}} > 1$ minden n-re, de ez a hányados tetszőlegesen közel kerülhet 1-hez;
- Ha $\lim_{n\to\infty} \frac{\Gamma_{\alpha,n}}{\Gamma_{\gamma,n}} = 1$, akkor az $\mathcal{M}_\alpha$ és az $\mathcal{M}_\gamma$ modellek között nincs különbség a predikciós képesség tekintetében;

- A $\Gamma_{\alpha,n}/\Gamma_{\gamma,n} > 1$ egyenlőtlenség akkor és csak akkor igaz, ha $\liminf_{n\to\infty} \Delta_{\alpha,n} > 0$.

Tehát az $\mathcal{K}$ konstrukciós halmazra illesztett M_α modellt most az $\mathcal{V}$ validációs halmazon kiértékeljük a $\hat{G}^{ASPE}$ generalizációs hiba segítségével:

$$\hat{G}^{ASPE}(\hat{\beta}_\alpha) = \frac{1}{n_v} \left\| y^{\mathcal{V}} - \hat{y}_\alpha^{\mathcal{K}} \right\|^2 = \frac{1}{n_v} \left\| (I - Q_\alpha^{\mathcal{V}})^{-1} (y^{\mathcal{V}} - X_\alpha^{\mathcal{V}} \hat{\beta}_\alpha) \right\|$$

Egy $\mathcal{M}_\alpha$ modellre a $\Gamma_{\alpha,n}$ átlagos generalizációs hiba becslése a $\hat{G}^{ASPE}$ értékeknek a $\mathcal{V}$ validációs halmaz n_v méretű összes (esetleg csak némelyik) részhalmazán vett átlaga. A keresztvalidáció által kiválasztott modell pedig az az $\mathcal{M}_\alpha$ modell lesz, amelyre ez a hibabecslés minimális az $\alpha \in \mathcal{A}$ indexek közül.

Ezt a módszert 'leave-n_v-out cross validation'-nek nevezzük és $CV(n_v)$-vel rövidítjük. A következőkben e módszer három változatát mutatjuk be, a végén pedig szimulációval tapasztalatilag is megvizsgáljuk, hogy e módszerek változatai milyen jól teljesítenek az optimális modell megtalálása tekintetében.

3.4.1. A $CV(1)$ módszer

A $CV(n_v)$ változatai közül a legegyszerűbb az az eset, amikor a validációra egyetlen megfigyelést tartunk fenn (vagyis $n_v \equiv 1$). Ezt $CV(1)$-gyel jelöljük (lásd 1.2.1. fejezet).

Tehát egy $\mathcal{M}_\alpha$ modellre a $\Gamma_{\alpha,n}$ átlagos generalizációs hiba $CV(1)$-becslése a $\hat{G}^{ASPE}$ értékeknek a $\mathcal{V}$ validációs halmaz összes egyelemű részhalmazán vett átlaga. A definíció alapján az összefüggések felhasználásával a számítások elvégzése után $\Gamma_{\alpha,n}$-re ezt a képletet kapjuk:

$$\hat{\Gamma}_{\alpha,n}^{CV} = \frac{1}{n} \sum_i [(1 - w_{i\alpha})^{-1} (y_i - x_{i\alpha}^T \hat{\beta}_\alpha)]^2.$$

Belátható, hogy a $\forall \alpha \in \mathcal{A}$ $\lim_{n\to\infty} \max_{i\geq n} w_{i\alpha} = 0$ feltétel mellett

$$\hat{\Gamma}_{\alpha,n}^{CV} = \begin{cases} \Gamma_{\alpha,n} + o_p(1) & \text{ha } M_\alpha \text{ I. kategóriájú;} \\ \frac{1}{n} e^T e + \frac{2}{n}\sigma^2 d_\alpha - \frac{1}{n} e^T P_\alpha e + o_p(\frac{1}{n}) & \text{ha } M_\alpha \text{ II. kategóriájú.} \end{cases}$$

Ebből közvetlenül látható, hogy

- Mivel $\frac{1}{n} e^T e \to \sigma^2$ majdnem biztosan, a $\hat{\Gamma}_{\alpha,n}^{CV}$ konzisztens becslése $\Gamma_{\alpha,n}$-nak;
- Ha $\mathcal{M}_\alpha$ II. kategóriájú, akkor $\Gamma_{\alpha,n} \to \sigma^2$;

- Ha $n \to \infty$, akkor 0-hoz tart annak a valószínűsége, hogy a $CV(1)$ módszer által választott modell I. kategóriájú;

- Ha $n \to \infty$ és az $\mathcal{M}_*$ optimális modell nem p méretű, akkor nem tart az 1-hez annak a valószínűsége, hogy a $CV(1)$ módszer által választott modell az $\mathcal{M}_*$ optimális modell lesz (vagyis a $CV(1)$ aszimptotikusan helytelen);

- Ha az előbbi feltétel fennáll és $e \sim \mathcal{N}(0, \sigma^2 I_n)$, akkor annak a valószínűsége, hogy a $CV(1)$ módszer $\mathcal{M}_*$ helyett inkább az $\mathcal{M}_\alpha$ modellt választja: $\mathcal{P}(2k < \chi^2(k)) + o(1)$, ahol $k = d_\alpha - d_{\alpha_*}$. Nyilvánvalóan $\mathcal{P}(2k < \chi^2(k)) \neq 0$ bármely $k \geq 1$-re.

Az utolsó pontból következik a $CV(1)$ módszernek az a tulajdonsága, hogy ha az optimális modell nem p méretű, akkor a $CV(1)$ hajlamos annál bővebb modellt választani. Ezért a $CV(1)$ módszert konzervatívnak nevezzük.

Az aszimptotikus helytelenség azzal magyarázható, hogy a $CV(1)$ módszer nem képes megkülönböztetni a II. kategóriájú modelleket, ami pedig annak a következménye, hogy míg a II. kategóriájú modelleknél a modelleket megkülönböztető kifejezésben a hibatag a másik taggal azonos nagyságrendű, addig ugyanez az I. kategóriájú modelleknél kisebb nagyságrendű.

3.4.2. A $BICV(n_v)$ módszer

Az előző fejezetben láttuk, hogy a $CV(1)$ módszer aszimptotikusan helytelen és konzervatív. A $CV(1)$ módszernek ez a hiányossága kijavítható azzal, hogy nagy validációs halmazt használunk (vagyis n_v mérete nagy és n_c mérete viszonylag kicsi). Az eddigiekben a validációt elvégeztük a validációs halmaz mind az $\binom{n}{n_v}$ db részhalmazára. Azonban ha $n \to \infty$, akkor ennek az elvégzése igencsak számításigényes. Ehelyett keresünk egy olyan módszert, ami a gyakorlatban is alkalmazható nagyon nagy méretű validációs halmaz esetén is.

3.4.4. Definíció. *Válasszunk ki az $\{1, ..., n\}$ halmazból b db olyan n_v elemű részhalmazt, amelyre a következő "egyensúlyi" feltételek érvényesek:*

a) minden $i \in \{1, ..., n\}$ ugyanannyi $\mathcal{B}$-beli halmaznak az eleme;

b) minden $(i, j) \in \{1, ..., n\}^2$ pár ugyanannyi $\mathcal{B}$-beli halmazban szerepel egyszerre. Az egyensúlyi feltételeknek eleget tevő halmazok halmazát jelöljük $\mathcal{B}$-vel. Válasszuk azt a modellt, amelyre a

$$\hat{\Gamma}_{\alpha,n}^{BICV} = \frac{1}{n_v b} \sum_{\mathcal{V} \in \mathcal{B}} \left\| y^{\mathcal{V}} - \hat{y}_\alpha^{\mathcal{K}} \right\|^2$$

minimális. Az így meghatározott $\mathcal{B}$ halmaz szerinti keresztvalidációs becslést $BICV(n_v)$ módszernek (Balanced Incomplete $CV(n_v)$ Method) nevezzük és $BICV(n_v)$-vel jelöljük.

$BICV(n_v)$ esetén tehát a $\Gamma_{\alpha,n}$ becslése tehát a $\mathcal{B}$ minden n_v elemszámú részhalmazára kiszámolt b darab $\hat{G}^{ASPE}$-érték átlaga. A gyakorlatban úgy választjuk a $\mathcal{B}$ halmazt, hogy b az n lineáris függvénye, azaz $b = O(n)$ legyen.

A következő eredmény azt mutatja, hogy a $BICV(n_v)$ módszer eredménye aszimptotikusan helyes, ha $n_c \to \infty$ és $\frac{n_v}{n} \to 1$.

3.4.5. Tétel. *Ha a $\Delta_{\alpha,n}$ aszimptotikusan sem nulla, és a tervmátrixból számolt kovariancia és annak inverze véges és a sajátértékei sem tartanak nullához, továbbá teljesül, hogy a konstrukciós és a validációs magyarázó változók viselkedése az alábbi értelemben hasonló,*

$$\lim_{n\to\infty} \max_{\mathcal{V}\in\mathcal{B}} \left\| \frac{1}{n_v} \sum_{i\in\mathcal{V}} x_i x_i^T - \frac{1}{n_c} \sum_{i\in\mathcal{K}} x_i x_i^T \right\| = 0 \ .$$

és ha az n_v-t úgy választjuk meg, hogy a $\frac{n_v}{n} \to 1$ és a $n_c \to \infty$ teljesüljön,
akkor igaz, hogy:

a) ha az $\mathcal{M}_\alpha$ I. kategóriájú, akkor létezik olyan $R_n \geq 0$, hogy

$$\hat{\Gamma}_{\alpha,n}^{BICV} = \frac{1}{n} e^T e + \Delta_{\alpha,n} + o_p(1) + R_n$$

;

b) ha az $\mathcal{M}_\alpha$ II. kategóriájú, akkor

$$\hat{\Gamma}_{\alpha,n}^{BICV} = \frac{1}{n} e^T e + n_v^{-1} d_\alpha \sigma^2 + o_p(n_c^{-1})$$

;

c) ha $n \to \infty$, akkor 1-hez tart annak a valószínűsége, hogy a kiválasztott modell optimális.

Most magyarázatot adunk arra, hogy a $BICV(n_v)$ miért javít a $CV(1)$-en és hogy az n_c-t és az n_v-t miért a fenti feltételeknek megfelelően kell választani.

Az $n_c \to \infty$ feltételre a modellillesztés konzisztenciájának biztosítása miatt van szükség, ez viszont még nem ad semmilyen információt az n_c és az n_v relatív arányra

vonatkozóan. Nagy n_c-t mégsem érdemes használni, amit a következőképpen indoklunk meg: egyrészt, ha n_c-t nagynak választjuk, akkor a II. kategóriájú modellek esetén a

$$\Gamma_{\alpha,n_c} = \sigma^2 + \frac{1}{n_c}\sigma^2 d_\alpha$$

optimalizálandó célfüggvény lapos, és ezért nehéz megtalálni a Γ_{α,n_c} minimumát, másrészt pedig, minél több adatot használunk akár a modellillesztésnél, akár a validációnál, annál pontosabb eredményt kapunk. A modellillesztésnél viszont nincs szükség nagy pontosságra, hiszen az illesztés után a kiválasztott modellt előrejelzési célból úgyis újraillesztjük a teljes adathalmazon, a validációnál viszont ahhoz, hogy megbízható eredményt kapjunk, a generalizációs hibát pontosan kell tudni értékelni. Ezért érdemes nagy n_v-t és viszonylag kicsi n_c-t választani.

De önmagában az sem elég, ha nagy n_v-t és viszonylag kis n_c-t használunk, az is szükséges, hogy $\frac{n_v}{n} \to 1$ legyen. Ha $\frac{n_v}{n}$ nem tart 1-hez, akkor ugyanaz a probléma fordul elő, mint $CV(1)$ esetén: a módszer inkonzisztens lesz, vagyis nem képes megkülönböztetni a II. kategóriájú modelleket, azaz a II. kategóriájú modelleknél a modelleket megkülönböztető kifejezésben a hibatag a többi taggal azonos nagyságrendű. Bizonyos algebrai számolások elvégzése után azt kapjuk, hogy a II. kategóriájú $\mathcal{M}_\alpha$ modellek esetén

$$\hat{\Gamma}^{BICV}_{\alpha,n} = \frac{1}{n}e^T e + \frac{1}{n_c}d_\alpha\sigma^2 + \varepsilon_{\alpha,n} \ ,$$

ahol a hibatag

$$\varepsilon_{\alpha,n} = \frac{(1+n_c)d_\alpha\sigma^2}{n_c(n-1)} - \frac{1}{n}e^T P_\alpha e + o_p\left(\frac{1}{n_c}\right) \ .$$

Ha $\frac{n_v}{n} \not\to 1$, akkor a $\varepsilon_{\alpha,n}$ hibatag és a $\frac{1}{n_c}\sigma^2 d_\alpha$ tag azonos nagyságrendű, Az $\varepsilon_{\alpha,n}$ hibatag csak akkor kisebb nagyságrendű a $\frac{1}{n_c}\sigma^2 d_\alpha$ tagnál, ha $\frac{n_c}{n} \to 0$, vagyis ha $\frac{n_v}{n} \to 1$.

Végezetül nem szabad megfeledkeznünk arról sem, hogy $\hat{\Gamma}^{CV}_{\alpha,n}$ a $\Gamma_{\alpha,n-1}$-nek a becslése, nem pedig $\Gamma_{\alpha,n}$-nek, mivel $CV(1)$ az átlagos generalizációs hibát egy $n-1$ méretű minta alapján becsli meg. Hasonlóképpen $\hat{\Gamma}^{BICV}_{\alpha,n}$ a Γ_{α,n_c}-nek a becslése, nem pedig $\Gamma_{\alpha,n}$-nek, mivel $BICV(n_v)$ az átlagos generalizációs hibát n_c méretű minták alapján becsli meg. Csakhogy amíg $CV(1)$ esetén a $\Gamma_{\alpha,n-1}$ és $\Gamma_{\alpha,n}$ közti különbség aszimptotikusan elhanyagolható, addig $BICV(n_v)$ esetén a Γ_{α,n_c} és $\Gamma_{\alpha,n}$ közti különbség csak pontosan akkor nem elhanyagolható, ha az $\frac{n_c}{n}$ nem tart 1-hez.

3.4.3. Más $CV(n_v)$ módszerek

Láthattuk, hogy a $BICV(n_v)$ kijavítja a $CV(1)$ hiányosságait, jobb eredményeket ad. Ezért sokszor célszerű (volna) használni. Azonban ehhez szükség van az egyensúlyi feltéte-

leknek eleget tevő $\mathcal{B}$ halmazra. Ilyen $\mathcal{B}$ előállítása alkalmanként nehézkes, túl nagy elemszámú, vagy éppen a rendelkezésre álló tulajdonságai okán nem indokolható a használata. Ezért két olyan alternatívát mutatunk, amelynél ilyen kiegyensúlyozott $\mathcal{B}$ halmazra nincs szükség.

Monte Carlo $CV(n_v)$ módszer

3.4.6. Definíció. *Válasszunk ki véletlenszerűen (visszatevéssel vagy visszatevés nélkül) b db n_v méretű részhalmazt az $\{1, ..., n\}$ halmazból és ezen halmazok halmazát jelöljük $\mathcal{R}$-rel. Válasszuk azt a modellt, amelyre a*

$$\hat{\Gamma}_{\alpha,n}^{MCCV} = \frac{1}{n_v b} \sum_{\mathcal{V} \in \mathcal{R}} \left\| y^{\mathcal{V}} - \hat{y}_{\alpha}^{\mathcal{K}} \right\|^2$$

minimális. Az így meghatározott $\mathcal{R}$ halmaz szerinti CV-becslést Monte Carlo-módszernek nevezzük és $MCCV(n_v)$-vel jelöljük.

Más szavakkal ez a konstrukció azt jelenti, hogy az adathalmazt b-szer véletlenszerűen felosztjuk n_v méretű részhalmazokra és a felosztásokra vesszük a generalizációs hiba becsléseinek átlagát. A $MCCV(n_v)$ módszer tehát csak abban különbözik a $BICV(n_v)$ módszertől, hogy míg a $BICV(n_v)$ módszernél a validációs mintarészeket tartalmazó halmazt jól meghatározott egyensúlyi feltételek szerint választjuk ki, addig az $MCCV(n_v)$ módszernél véletlenszerűen.

Ezen módszert vizsgálva a 3.4.6. tételhez nagyon hasonló eredményeket kapunk: 1-hez tart annak a valószínűsége, hogy a kiválasztott modell optimális, azzal a feltétellel, hogy ha $n \to \infty$, akkor a $\frac{n^2}{bn_c^2} \to 0$. De mint látható, e szükséges feltétel megszorításokat ró a b-re és az n_c-re is: minél kevesebb adatot használunk fel a modellillesztéshez (n_c), annál több felosztásra van szükség (b).

Approximált $CV(n_v)$ módszer

3.4.7. Definíció. *Válasszuk azt a modellt, amelyre a*

$$\hat{\Gamma}_{\alpha,n}^{APCV} = \frac{1}{n} \left\| y - X_\alpha \hat{\beta}_\alpha \right\|^2 + \frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha} (y_i - x_{i,\alpha} \hat{\beta}_\alpha)^2$$

mennyiség minimális. A $\Gamma_{\alpha,n}$ ez utóbbi formula szerinti keresztvalidációs becslését approximált CV-módszernek nevezzük és $APCV(n_v)$-vel jelöljük..

Belátható, hogy $\hat{\Gamma}_{\alpha,n}^{APCV} = \hat{\Gamma}_{\alpha,n}^{BICV}$ abban a speciális esetben, ha a konstrukciós és a validációs adatrész kovarianciamátrixa egyenlő minden $\mathcal{V} \in \mathcal{B}$-re, vagyis

$$\frac{1}{n_v} \sum_{i \in \mathcal{V}} x_i x_i^T = \frac{1}{n_c} \sum_{i \in \mathcal{K}} x_i x_i^T.$$

Ha a 3.4.6. tételben a $\hat{\Gamma}_{\alpha,n}^{BICV}$-t helyettesítjük $\hat{\Gamma}_{\alpha,n}^{APCV}$-vel, akkor a tétel eredményei is teljesülnek, feltéve, hogy a tétel feltételei fennállnak.

Az *APCV* (analitikusan approximált CV) név onnan származik, hogy a keresztvalidációs hiba normális esetben érvényes ekvivalens alakját általánosítottuk a nemnormális eloszlások esetére is.

Megjegyzések:

- Az *APCV* előnye, hogy konzisztens és kevesebb számítást igényel, mint a *BICV* vagy az *MCCV*.
- Az *APCV* hátránya, hogy a lineáris modellekről nem könnyen általánosítható más modellekre.
- Az *APCV* teljesítménye kevésbé jó, mint az *MCCV*-é, ami azt jelzi, hogy a jó teljesítményhez az *APCV*-nek nagyobb n-et igényel, mint az *MCCV*. Ez a 3.4.4. fejezetben elvégzett szimulációból is kiderül.

3.4.4. Szimuláció

Az eddigiek interpretálására

$$y = X\beta + e$$

lineáris regressziós modellhez elvégzünk egy szimulációt

$$p = 5,\ n = 40,\ n_v = 25,\ n_c = 15$$

paraméterekkel, vagyis tekintjük ezt:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i},$$

ahol $i = 1, ..., 40$, a hibatagra: $e_i \sim \mathcal{N}(0, 1)$, az x_{ki} magyarázó változókra: $x_{1i} \equiv 1$, a többit $(k = 2, ..., 5,\ i = 1, ..., 40)$ pedig a következő táblázatból vesszük:

x2	x3	x4	x5
.3600	.5300	1.0600	.5326
1.3200	2.5200	5.7400	3.6183
.0600	.0900	.2700	.2594
.1600	.4100	.8300	1.0346
.0100	.0200	.0700	.0381
.0200	.0700	.0700	.3440
.5600	.6200	2.1200	1.4559
.9800	1.0600	2.8900	4.0182
.3200	.2000	.7600	.4600
.0100	.0000	.0700	.1540
.1500	.2500	.5000	.6516
.2400	.2800	.5900	.0611
.1100	.3500	.4000	.1922
.0800	.1300	.2800	.0931
.6100	.8500	.4900	.0538
.0300	.0300	.2300	.0199
.0600	.1100	.5000	.0419
.0200	.0800	.2500	.1093
.0400	.2400	.0800	.0328
.0000	.0200	.0400	.0797
.0900	.1800	.5900	.1855
.0200	.1600	.2400	.1572
.0200	.1100	.2100	.0998
.0500	.2400	.4300	.2804
.1100	.3900	.2900	.2879
.1800	.1100	.4300	.6810
.0400	.0900	.2300	.3242
.8500	1.3300	2.7000	2.6013
.1700	.3200	.6600	.4469
.0800	.1200	.4900	.2436
.3800	.1800	.4900	.4400
.1100	.1300	.1800	.3351
.3900	.3800	.9900	1.3979
.4300	.4600	1.4700	2.0138
.5700	1.1600	1.8200	1.9356
.1300	.0300	.0800	.1050

.0400	.0500	.1400	.2207
.1300	.1800	.2800	.0180
.2000	.9500	.4100	.1017
.0700	.0600	.1800	.096

Mivel az egész fejezetben azt vizsgáljuk, hogy mely magyarázó változók hagyhatók el (ekkor a megfelelő β_k értéke 0), ezért most az $\{x_1, ..., x_5\}$ magyarázó változók közül néhány lehetséges módon kiválasztott részhalmazra három különböző keresztvalidációs módszerrel elvégezzük a modellillesztést, ezek közül a legjobb predikciós képességű modellt választjuk ki, és megnézzük, hogy ez a modell optimális-e. A három használt módszer: $CV(1)$, $MCCV(n_v)$ (ahol $b = 2n$ a CV-ismétlések száma) és $APCV(n_v)$.

1000 szimuláció alapján az alábbi táblázat megadja a különböző esetekben mindegyik modell kiválasztásának a tapasztalati valószínűségeit.

beta			CV	MCCV	APCV
= (2, 0, 0, 4, 0)	1, 4	Optimal	.484	.934	.501
	1, 2, 4	II	.133	.025	.116
	1, 3, 4	II	.127	.026	.085
	1, 4, 5	II	.138	.012	.172
	1, 2, 3, 4	II	.049	.000	.038
	1, 2, 4, 5	II	.029	.001	.039
	1, 3, 4, 5	II	.030	.002	.037
	1, 2, 3, 4, 5	II	.009	.000	.012
=(2, 0, 0, 4, 8)	1, 4, 5	Optimal	.641	.947	.651
	1, 2, 4, 5	II	.158	.032	.161
	1, 3, 4, 5	II	.138	.020	.131
	1, 2, 3, 4, 5	II	.063	.001	.057
=(2, 9, 0, 4, 8)	1, 4, 5	I	.005	.016	.000
	1, 2, 4, 5	Optimal	.801	.965	.818
	1, 3, 4, 5	I	.005	.002	.000
	1, 2, 3, 4, 5	II	.189	.017	.182
=(2, 9, 6, 4, 8)	1, 2, 3, 5	I	.000	.002	.000
	1, 2, 4, 5	I	.000	.005	.000

1, 3, 4, 5	I	.015	.045	.001
1, 2, 3, 4, 5	Optimal	.985	.948	.999

3.4.5. A szimuláció értelmezése

1. Az optimális modell kiválasztásának a valószínűsége az $MCCV$ esetén a legnagyobb (kivéve azt az esetet, amikor a legnagyobb modell az optimális) és az $APCV$ mindegyik esetben enyhén jobban teljesít a $CV(1)$-nél.

2. I. kategóriájú (vagyis helytelen) modell kiválasztásának a valószínűsége mindegyik módszer mindegyik esetében elhanyagolható.

3. A várakozásnak megfelelően a $CV(1)$ feleslegesen nagy modelleket hajlamos kiválasztani. Az optimális modell kiválasztásának a valószínűsége az $CV(1)$ esetén nagyon alacsony (kisebb 0.5-nél). A β-nak minél több nulla komponense van, annál rosszabb a $CV(1)$ teljesítménye. Másfelől, az $MCCV$ teljesítménye stabil és a $CV(1)$-énél sokkal jobb minden olyan esetben, amikor az optimális modell nem a legnagyobb modell.

4. Az $APCV$ teljesítménye csak enyhén jobb a $CV(1)$-énél annak ellenére, hogy az $APCV$ konzisztens, a $CV(1)$ pedig inkonzisztens. Ez azt jelzi, hogy a jó teljesítményhez az $APCV$ nagyobb mintaméretet igényelhet, mint az $MCCV$.

Irodalomjegyzék

[1] J. Larsen C. Goutte, *On Optimal Data Split for Generalization and Estimation and Model Selection*, IEEE-SigProc 1999, pp. 225-234.

[2] J. Shao, *Linear Model Selection by Cross-Validation*, Journal of the American Statistical Association, Vol. 88, No. 422 (Jun., 1993), pp. 486-494.

[3] B. Efron and G. Gong *A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation*, TAS, Vol. 37, No. 1 (Feb., 1983), pp. 36-48.

[4] R. R. Picard and K. N. Berk *Data Splitting*, TAS, Vol. 44, No. 2 (May, 1990), pp. 140-147.

[5] B. Efron and R. Tibshirani *Improvements on Cross-Validation: The .632+ Bootstrap Method*, JASA, Vol. 92, No. 438 (Jun., 1997), pp. 548-560.

[6] B. M. Stone *Cross-Validatory Choice and Assessment of Statistical Predictions*, J. of Roy. Stat. Soc. Ser. B (Methodological), Vol. 36, No. 2(1974), pp. 111-147.

[7] Pröhle Tamás, *Cross-validáció és szimuláció ...*, Kézirat, 2014

[8] Colin N. Park and Arthur L. Dudycha *A Cross-Validation Approach to Sample Size Determination for Regression Models* Journal of the American Statistical Association, Vol. 69, No. 345 (Mar., 1974), pp. 214-218.