A statisztikai modellek alapjai

Ferenci Tamás
tamas.ferenci@medstat.hu
http://www.medstat.hu/
https://www.youtube.com/c/FerenciTamas

Utoljára frissítve: 2022. június 30.

A statisztikai modellek alkalmazása

- A statisztika modellek rengeteg módon vezethetőek be
- Én most úgy fogom tekinteni, mint egy eszközt a confounding kezelésére
- Nézzünk meg először egy – szimulált – példát!

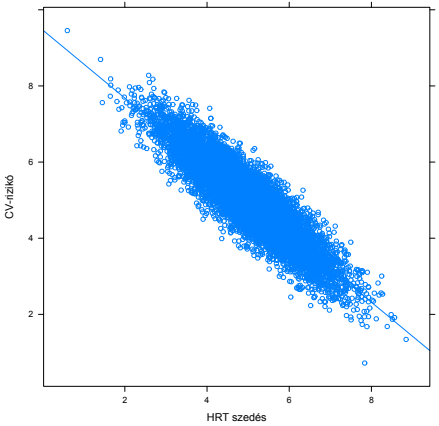
A statisztikai modellek alkalmazása

- A statisztika modellek rengeteg módon vezethetőek be
- Én most úgy fogom tekinteni, mint egy eszközt a confounding kezelésére
- Nézzünk meg először egy – szimulált – példát!

A statisztikai modellek alkalmazása

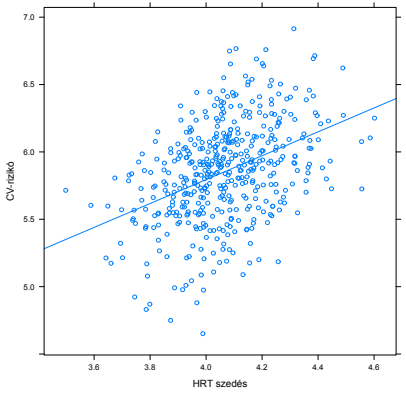
- A statisztika modellek rengeteg módon vezethetőek be
- Én most úgy fogom tekinteni, mint egy eszközt a confounding kezelésére
- Nézzünk meg először egy – szimulált – példát!

A confounding alaphelyzete



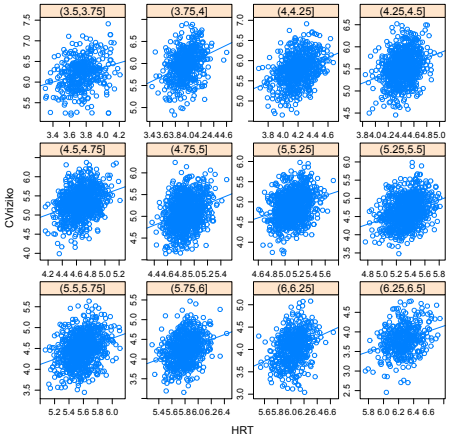
Első kezelési lehetőség: rétegzés

A confounder szerint bontsuk meg – rétegezzük – a vizsgálatot; például 4 körüli (3,9 és 4,1) közötti SES-nél az összefüggés:



Első kezelési lehetőség: rétegzés

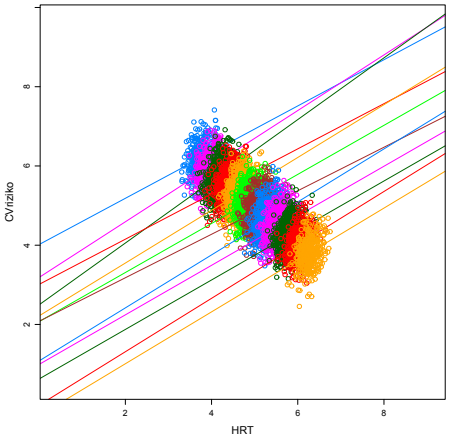
Az összes ilyen együtt:



A confounding illusztrációja

Ez mellesleg a confounding jelenségét is jól illusztrálja! Még térlátás-igényesebb megoldás: ugyanez 3D-ben...

A confounding illusztrációja és a rétegzés



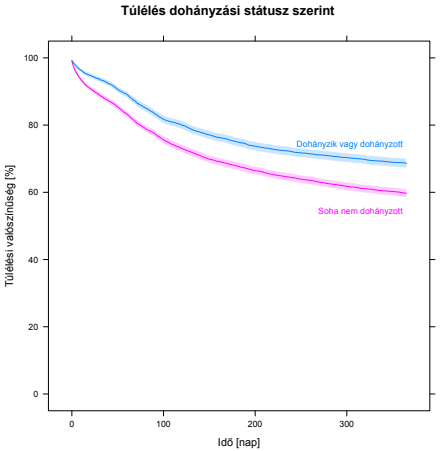
Rétegzés a gyakorlatban

Igazából a Simpson-paradoxon is ez:

	Nyílt feltárás	Perkután eljárás
Kőátmérő $<$ 2cm	93% (81/87)	87% (234/270)
Kőátmérő $\geq$ 2cm	73% (192/263)	69% (55/80)
Összességében	78% (273/350)	83% (289/350)

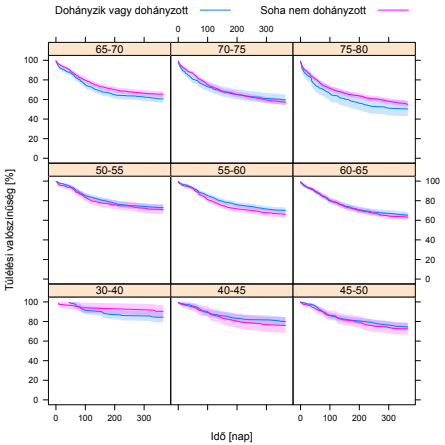
Rétegzés a gyakorlatban

Végrehajtható a szívinfarktusos esetben is:



Rétegzés a gyakorlatban

Végrehajtható a szívinfarktusos esetben is:



Mi a baj a rétegzéssel?

- Noha ez sem tökéletes (nem lehet „adott” SES-nél vagy adott életkornál vizsgálódni, össze kell vonni egy tartományt, egybemosva az ottani viszonyokat), egy confoundernél még tulajdonképpen különösebb gond nélkül alkalmazható
- De mi van akkor, ha több potenciális confounder van…?
- A rétegek („cellák”) száma *kombinatorikusan* nő, egy idő után kezelhetetlen lesz (gondoljunk a császármetszés vs. T1DM példájára)!
- (Nem is egyszerűen olyan értelemben, hogy sok lesz, hanem, hogy a mintaszámhoz *képest* lesz sok!)
- Valamint: nehezen értelmezhető az eredmény (hogyan mondunk valamilyen kezelhető, kompakt információt abból, hogy a 12 rétegben mi az összefüggés?)

Mi a baj a rétegzéssel?

- Noha ez sem tökéletes (nem lehet „adott” SES-nél vagy adott életkornál vizsgálódni, össze kell vonni egy tartományt, egybemosva az ottani viszonyokat), egy confoundernél még tulajdonképpen különösebb gond nélkül alkalmazható
- De mi van akkor, ha több potenciális confounder van…?
- A rétegek („cellák”) száma *kombinatorikusan* nő, egy idő után kezelhetetlen lesz (gondoljunk a császármetszés vs. T1DM példájára)!
- (Nem is egyszerűen olyan értelemben, hogy sok lesz, hanem, hogy a mintaszámhoz *képest* lesz sok!)
- Valamint: nehezen értelmezhető az eredmény (hogyan mondunk valamilyen kezelhető, kompakt információt abból, hogy a 12 rétegben mi az összefüggés?)

Mi a baj a rétegzéssel?

- Noha ez sem tökéletes (nem lehet „adott” SES-nél vagy adott életkornál vizsgálódni, össze kell vonni egy tartományt, egybemosva az ottani viszonyokat), egy confoundernél még tulajdonképpen különösebb gond nélkül alkalmazható
- De mi van akkor, ha több potenciális confounder van…?
- A rétegek („cellák”) száma *kombinatorikusan* nő, egy idő után kezelhetetlen lesz (gondoljunk a császármetszés vs. T1DM példájára)!
- (Nem is egyszerűen olyan értelemben, hogy sok lesz, hanem, hogy a mintaszámhoz *képest* lesz sok!)
- Valamint: nehezen értelmezhető az eredmény (hogyan mondunk valamilyen kezelhető, kompakt információt abból, hogy a 12 rétegben mi az összefüggés?)

Mi a baj a rétegzéssel?

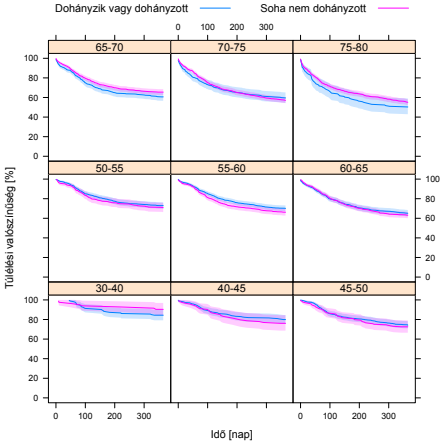
- Noha ez sem tökéletes (nem lehet „adott” SES-nél vagy adott életkornál vizsgálódni, össze kell vonni egy tartományt, egybemosva az ottani viszonyokat), egy confoundernél még tulajdonképpen különösebb gond nélkül alkalmazható
- De mi van akkor, ha több potenciális confounder van…?
- A rétegek („cellák”) száma *kombinatorikusan* nő, egy idő után kezelhetetlen lesz (gondoljunk a császármetszés vs. T1DM példájára)!
- (Nem is egyszerűen olyan értelemben, hogy sok lesz, hanem, hogy a mintamérethez *képest* lesz sok!)
- Valamint: nehezen értelmezhető az eredmény (hogyan mondunk valamilyen kezelhető, kompakt információt abból, hogy a 12 rétegben mi az összefüggés?)

Mi a baj a rétegzéssel?

- Noha ez sem tökéletes (nem lehet „adott” SES-nél vagy adott életkornál vizsgálódni, össze kell vonni egy tartományt, egybemosva az ottani viszonyokat), egy confoundernél még tulajdonképpen különösebb gond nélkül alkalmazható
- De mi van akkor, ha több potenciális confounder van…?
- A rétegek („cellák”) száma *kombinatorikusan* nő, egy idő után kezelhetetlen lesz (gondoljunk a császármetszés vs. T1DM példájára)!
- (Nem is egyszerűen olyan értelemben, hogy sok lesz, hanem, hogy a mintaérethez *képest* lesz sok!)
- Valamint: nehezen értelmezhető az eredmény (hogyan mondunk valamilyen kezelhető, kompakt információt abból, hogy a 12 rétegben mi az összefüggés?)

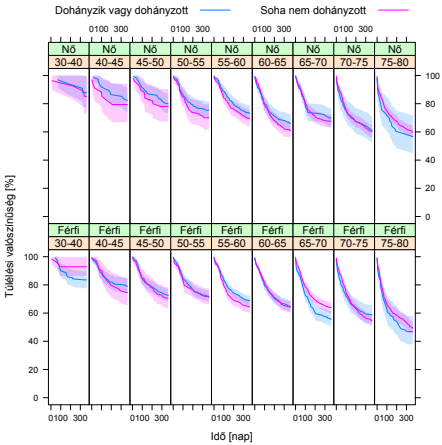
Egy gyakorlati példa minderre

Végrehajtható a szívinfarktusos esetben is:



Egy gyakorlati példa minderre

De a rétegek száma egy idő után kezd nagyon elszaladni...



Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\text{CVrizikó} = \beta_0 + \beta_{\text{SES}}\text{SES} + \beta_{\text{HRT}}\text{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\text{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\text{CVriziko} = \beta_0 + \beta_{\text{SES}}\text{SES} + \beta_{\text{HRT}}\text{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\text{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\text{CVriziko} = \beta_0 + \beta_{\text{SES}}\text{SES} + \beta_{\text{HRT}}\text{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\text{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\mathrm{CVriziko} = \beta_0 + \beta_{\mathrm{SES}}\mathrm{SES} + \beta_{\mathrm{HRT}}\mathrm{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\mathrm{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\mathrm{CVriziko} = \beta_0 + \beta_{\mathrm{SES}}\mathrm{SES} + \beta_{\mathrm{HRT}}\mathrm{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\mathrm{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

Egy megoldási lehetőség

- A rétegzés előnye, hogy *semmilyen* feltevéssel nem kell élnünk a változók kapcsolatáról
- De ha ez már nem működik, akkor egyszerűsíteni kell a helyzetet: éljünk valamilyen feltételezéssel arról, hogy hogyan függenek össze ezek a változók!
- (Lényegében: redukáljuk a paramétereink számát)
- Például: lineárisan függ a CV-rizikó a SES-től, és a HRT-szedéstől is:

$$\mathrm{CVriziko} = \beta_0 + \beta_{\mathrm{SES}}\mathrm{SES} + \beta_{\mathrm{HRT}}\mathrm{HRT} + u$$

- Itt β_{HRT} a HRT-szedés szocioökonómiai státusztól *tisztított* hatása lesz (gondoljuk végig a ceteris paribus értelmezést)
- Valóban, az előbbi szimulált példán $\widehat{\beta_{\mathrm{HRT}}} = 0{,}89$, míg ha a SES-t nem raktuk volna bele, akkor $-0{,}90$

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

A regressziós modell

- Ezt hívjuk regressziós modellnek
- Szétválogatja az egyes tényezők hatásait
- Több változós is teljesen hasonlóan kezelhető
- De a dolog nincs ingyen: beépítettük a feltevéseket, amiknek teljesülnie kell, hogy hiteles képet kapjunk
- Például linearitás, nincs interakció

McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005 Jul;62(7):500-6, 472.

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Folytonos eredményváltozó: lineáris regresszió (esettanulmány: CAP és HS)

- Egyváltozós vizsgálatok
- Confounding
- Többváltozós regresszió
- β-k és értelmezésük, confounding elleni védekezés
- Nemlinearitás (spline-nal) és tesztelése
- Vizualizáció (forest plot és teljes tartományos)
- Modellszelekció és csapdái (nem prespecifikált modellek, automatikus modellszelekció)
- Túlilleszkedés
- Modell (belső) validálása és kalibrálása: split-sample (hold-out sample), keresztvalidáció, bootstrap
- Regularizált (penalizált) regresszió

Kategoriális eredményváltozó: logisztikus regresszió (esettanulmány: antiarritmiás szerek paradox hatása)

Mint az előbbiek +

- Potenciális confounderekhez szükséges paraméterek számának csökkentése (blinded to the outcome!)
- OR-k és értelmezésük
- Cut-off, szenzitivitás, specificitás
- ROC-görbe, AUC
- Bináris logisztikus regresszió kiterjesztése több kategóriára (multinomiális és ordinális logisztikus regresszió)

Kategoriális eredményváltozó: logisztikus regresszió (esettanulmány: antiarritmiás szerek paradox hatása)

Mint az előbbiek +

- Potenciális confounderekhez szükséges paraméterek számának csökkentése (blinded to the outcome!)
- OR-k és értelmezésük
- Cut-off, szenzitivitás, specificitás
- ROC-görbe, AUC
- Bináris logisztikus regresszió kiterjesztése több kategóriára (multinomiális és ordinális logisztikus regresszió)

Kategoriális eredményváltozó: logisztikus regresszió (esettanulmány: antiarritmiás szerek paradox hatása)

Mint az előbbiek +

- Potenciális confounderekhez szükséges paraméterek számának csökkentése (blinded to the outcome!)
- OR-k és értelmezésük
- Cut-off, szenzitivitás, specificitás
- ROC-görbe, AUC
- Bináris logisztikus regresszió kiterjesztése több kategóriára (multinomiális és ordinális logisztikus regresszió)

Kategoriális eredményváltozó: logisztikus regresszió (esettanulmány: antiarritmiás szerek paradox hatása)

Mint az előbbiek +

- Potenciális confounderekhez szükséges paraméterek számának csökkentése (blinded to the outcome!)
- OR-k és értelmezésük
- Cut-off, szenzitivitás, specificitás
- ROC-görbe, AUC
- Bináris logisztikus regresszió kiterjesztése több kategóriára (multinomiális és ordinális logisztikus regresszió)

Kategoriális eredményváltozó: logisztikus regresszió (esettanulmány: antiarritmiás szerek paradox hatása)

Mint az előbbiek +

- Potenciális confounderekhez szükséges paraméterek számának csökkentése (blinded to the outcome!)
- OR-k és értelmezésük
- Cut-off, szenzitivitás, specificitás
- ROC-görbe, AUC
- Bináris logisztikus regresszió kiterjesztése több kategóriára (multinomiális és ordinális logisztikus regresszió)

Time-to-event eredményváltozó: Cox-regresszió (esettanulmány: culprit ér szerepe AMI utáni túlélésben)

Mint az előbbiek +

- HR-k és értelmezésük
- Proporcionalitási feltevés

Time-to-event eredményváltozó: Cox-regresszió (esettanulmány: culprit ér szerepe AMI utáni túlélésben)

Mint az előbbiek +

- HR-k és értelmezésük
- Proporcionalitási feltevés