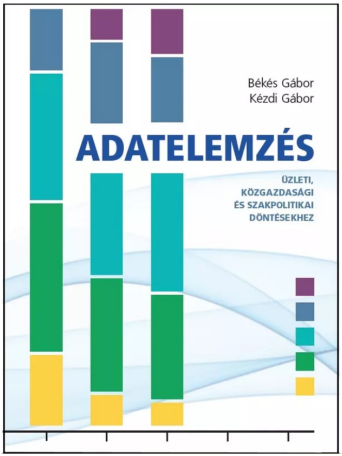
10. hét: Többváltozós regresszió

A prezentációt készítette: Tőkés László

Bevezetés az empirikus közgazdaságtanba

2025/26/1

Prezentáció a Békés-Kézdi-féle adatelemzés könyvhöz



- Alinea Kiadó, 2024
- gabors-data-analysis.com
  - Adatok és kódok elérhetőek itt: gabors-data-analysis.com/data-and-code/
- Prezentáció a 10. fejezethez

Bevezetés

- Emlékeztető: 4. házi feladat → korreláció szálláshelyek ára és központtól vett távolsága között

Bevezetés

- Emlékeztető: 4. házi feladat $\rightarrow$ korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$

Bevezetés

- Emlékeztető: 4. házi feladat $\rightarrow$ korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?

Bevezetés

- Emlékeztető: 4. házi feladat $\rightarrow$ korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?
  - Valószínűleg igen. De csak *részben*.

Bevezetés

- Emlékeztető: 4. házi feladat → korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?
  - Valószínűleg igen. De csak *részben*.
- A szálláshelyek ára a minősítésüktől is függ: $price^E = -17 + 43 \cdot stars$

Bevezetés

- Emlékeztető: 4. házi feladat $\rightarrow$ korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?
  - Valószínűleg igen. De csak *részben*.
- A szálláshelyek ára a minősítésüktől is függ: $price^E = -17 + 43 \cdot stars$
- Azt is látjuk továbbá, hogy a közelebb lévő szálláshelyek jellemzően jobbak is: $Corr(distance, stars) = -0,177$.

Bevezetés

- Emlékeztető: 4. házi feladat $\rightarrow$ korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?
  - Valószínűleg igen. De csak *részben*.
- A szálláshelyek ára a minősítésüktől is függ: $price^E = -17 + 43 \cdot stars$
- Azt is látjuk továbbá, hogy a közelebb lévő szálláshelyek jellemzően jobbak is: $Corr(distance, stars) = -0,177$.
- Mi a baj tehát az első regresszióval?

Bevezetés

- Emlékeztető: 4. házi feladat → korreláció szálláshelyek ára és központtól vett távolsága között
- Az ár függ a távolságtól: $price^E = 151 - 12 \cdot distance$
  - A közelség *miatt* drágábbak a szálláshelyek?
  - Valószínűleg igen. De csak *részben*.
- A szálláshelyek ára a minősítésüktől is függ: $price^E = -17 + 43 \cdot stars$
- Azt is látjuk továbbá, hogy a közelebb lévő szálláshelyek jellemzően jobbak is: $Corr(distance, stars) = -0,177$.
- Mi a baj tehát az első regresszióval?
  - Megoldás: többváltozós regresszió.

Motiváció

- Hogyan tudnánk felhasználni az adatokat arra, hogy megtaláljuk azokat a szállodákat, amelyek az *összes jellemzőjükhöz* képest olcsónak számítanak?

Bevezetés

Többváltozós lineáris regresszió: mikor és miért?

Három okból végezhetünk az egyváltozós helyett többváltozós regresszióelemzést:

- Több kapcsolati mintázatot szeretnénk feltárni.

Többváltozós lineáris regresszió: mikor és miért?

Három okból végezhetünk az egyváltozós helyett többváltozós regresszióelemzést:

- Több kapcsolati mintázatot szeretnénk feltárni.
- Minél jobb predikciót szeretnénk készíteni $\rightarrow$ szeretnénk a szóródás nagyobb részét megmagyarázni.

Többváltozós lineáris regresszió: mikor és miért?

Három okból végezhetünk az egyváltozós helyett többváltozós regresszióelemzést:

- Több kapcsolati mintázatot szeretnénk feltárni.
- Minél jobb predikciót szeretnénk készíteni $\rightarrow$ szeretnénk a szóródás nagyobb részét megmagyarázni.
- Szeretnénk ok-okozati kapcsolatot feltárni $\rightarrow$ szeretnénk jobban összehasonlítható megfigyeléseket összehasonlítani.

Többváltozós lineáris regresszió: alapok

- A többváltozós regresszióelemzés y átlagát egynél több x változó függvényeként tárja fel: $y^E = f(x_1, x_2, ...)$.

Többváltozós lineáris regresszió: alapok

- A többváltozós regresszióelemzés y átlagát egynél több x változó függvényeként tárja fel: $y^E = f(x_1, x_2, ...)$.
- A többváltozós lineáris regresszió y átlagát a magyarázó változók lineáris függvényeként specifikálja.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Többváltozós lineáris regresszió: alapok

- A többváltozós regresszióelemzés y átlagát egynél több x változó függvényeként tárja fel: $y^E = f(x_1, x_2, ...)$.
- A többváltozós lineáris regresszió y átlagát a magyarázó változók lineáris függvényeként specifikálja.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- Alapja: y különböző értékeit x különböző értékei szerint és z hasonló értékei mellett hasonlítjuk össze (4. hét).
  - további feltételes összevetés; feltételes összehasonlítás

Két magyarázó változó esete

Két magyarázó változó esete

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Két magyarázó változó esete

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- β_1 megmutatja, hogy átlagosan mennyivel nagyobb y értéke azoknál a megfigyeléseknél, amelyeknél x_1 egy egységgel nagyobb, míg x_2 *azonos értékű*.

Két magyarázó változó esete

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- β_1 megmutatja, hogy átlagosan mennyivel nagyobb y értéke azoknál a megfigyeléseknél, amelyeknél x_1 egy egységgel nagyobb, míg x_2 *azonos értékű*.
- β_2 megmutatja, hogy átlagosan mennyivel nagyobb y értéke azoknál a megfigyeléseknél, amelyeknél x_2 egy egységgel nagyobb, míg x_1 *azonos értékű*.

Két magyarázó változó esete

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- β_1 megmutatja, hogy átlagosan mennyivel nagyobb y értéke azoknál a megfigyeléseknél, amelyeknél x_1 egy egységgel nagyobb, míg x_2 *azonos értékű*.
- β_2 megmutatja, hogy átlagosan mennyivel nagyobb y értéke azoknál a megfigyeléseknél, amelyeknél x_2 egy egységgel nagyobb, míg x_1 *azonos értékű*.
- Összehasonlítja azokat a megfigyeléseket, amelyek az egyik magyarázó változóban azonosak annak érdekében, hogy megmutassa a másik magyarázó változóhoz tartozó különbségeket.

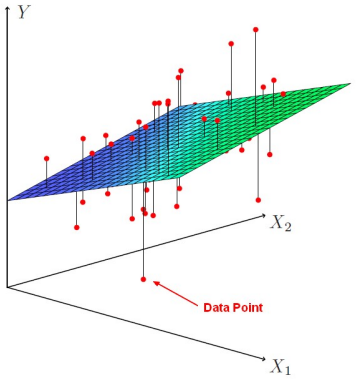
Két magyarázó változó esete

A két magyarázó változó esete vizuálisan egy lineáris sík illesztését jelenti:

- Még mindig a hibák négyzetösszegét minimalizáljuk:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{N} (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2$$

- K változó esetén egy K dimenziós lineáris síkot illesztünk.



Kép forrása: Datacadamia.

Feladat: 1. feladat

Töltse le OSF-ről, majd nyissa meg a `hotels-vienna.csv` adattáblát, és hajtsa végre a következőket:

1) Futtasson egy regressziót, ahol az árat magyarázza a távolsággal és az értékeléssel! Értelmezze a kapott együtthatókat!

A kihagyott változók miatti torzítás

Többváltozós vs. egyváltozós regresszió

Hasonlítsuk össze az egyváltozós regresszió meredekségi együtthatóját (β) a többváltozóséval (β_1):

$$\text{Egyváltozós: } y^E = \alpha + \beta x_1$$

$$\text{Többváltozós: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Többváltozós vs. egyváltozós regresszió

Hasonlítsuk össze az egyváltozós regresszió meredekségi együtthatóját (β) a többváltozóséval (β_1):

$$\text{Egyváltozós: } y^E = \alpha + \beta x_1$$

$$\text{Többváltozós: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β és β_1 összehasonlításához futtassuk x_2 regresszióját x_1-en ($x - x$ regreszió):

$$x_2^E = \gamma + \delta x_1$$

Többváltozós vs. egyváltozós regresszió

Hasonlítsuk össze az egyváltozós regresszió meredekségi együtthatóját (β) a többváltozóséval (β_1):

$$\text{Egyváltozós: } y^E = \alpha + \beta x_1$$

$$\text{Többváltozós: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β és β_1 összehasonlításához futtassuk x_2 regresszióját x_1-en ($x - x$ regreszió):

$$x_2^E = \gamma + \delta x_1$$

Helyettesítsük ezt be a többváltozós regresszióba:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2(\gamma + \delta x_1) = \beta_0 + \beta_2\gamma + (\beta_1 + \beta_2\delta)x_1$$

Többváltozós vs. egyváltozós regresszió

Hasonlítsuk össze az egyváltozós regresszió meredekségi együtthatóját (β) a többváltozóséval (β_1):

$$\text{Egyváltozós: } y^E = \alpha + \beta x_1$$

$$\text{Többváltozós: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

β és β_1 összehasonlításához futtassuk x_2 regresszióját x_1-en ($x - x$ regreszió):

$$x_2^E = \gamma + \delta x_1$$

Helyettesítsük ezt be a többváltozós regresszióba:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2(\gamma + \delta x_1) = \beta_0 + \beta_2\gamma + (\beta_1 + \beta_2\delta)x_1$$

Látható, hogy:

$$\beta = \beta_1 + \delta\beta_2$$

Különbség a meredekségekben: torzítás

Tehát:

- Egyváltozós regresszió: $y^E = \alpha + \beta x_1$
- Többváltozós regresszió: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Vagyis:

- x_1 együtthatója: $\beta = \beta_1 + \delta\beta_2$

Különbség a meredekségekben: torzítás

Tehát:

- Egyváltozós regresszió: $y^E = \alpha + \beta x_1$
- Többváltozós regresszió: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Vagyis:

- x_1 együtthatója: $\beta = \beta_1 + \delta\beta_2$

- Az egyváltozós regresszióban az x_1 meredeksége eltér a többváltozós regresszióbeli meredekségétől. Két kivétel:

Különbség a meredekségekben: torzítás

Tehát:

- Egyváltozós regresszió: $y^E = \alpha + \beta x_1$
- Többváltozós regresszió: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Vagyis:

- x_1 együtthatója: $\beta = \beta_1 + \delta\beta_2$

- Az egyváltozós regresszióban az x_1 meredeksége eltér a többváltozós regresszióbeli meredekségétől. Két kivétel:
  - x_1 és x_2 korrelálatlan ($\delta = 0$)

Különbség a meredekségekben: torzítás

Tehát:

- Egyváltozós regresszió: $y^E = \alpha + \beta x_1$
- Többváltozós regresszió: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Vagyis:

- x_1 együtthatója: $\beta = \beta_1 + \delta\beta_2$

- Az egyváltozós regresszióban az x_1 meredeksége eltér a többváltozós regresszióbeli meredekségétől. Két kivétel:
  - x_1 és x_2 korrelálatlan ($\delta = 0$)
  - a többváltozós regresszióban x_2 meredeksége nulla ($\beta_2 = 0$).

Különbség a meredekségekben: torzítás

Tehát:

- Egyváltozós regresszió: $y^E = \alpha + \beta x_1$
- Többváltozós regresszió: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Vagyis:

- x_1 együtthatója: $\beta = \beta_1 + \delta\beta_2$

- Az egyváltozós regresszióban az x_1 meredeksége eltér a többváltozós regresszióbeli meredekségétől. Két kivétel:
  - x_1 és x_2 korrelálatlan ($\delta = 0$)
  - a többváltozós regresszióban x_2 meredeksége nulla ($\beta_2 = 0$).
- Azt mondjuk, hogy a β együttható torzított.

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.
- Ha x_1 és x_2 korrelált, akkor számít, hogy olyan megfigyeléseket hasonlítunk-e össze, amelyek ugyanolyan, vagy eltérő x_2 értékkel rendelkeznek.

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.
- Ha x_1 és x_2 korrelált, akkor számít, hogy olyan megfigyeléseket hasonlítunk-e össze, amelyek ugyanolyan, vagy eltérő x_2 értékkel rendelkeznek.
  - Például, pozitív korreláció esetén a nagyobb x_1 értékek nagyobb x_2 értékeket is jelentenek.

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.
- Ha x_1 és x_2 korrelált, akkor számít, hogy olyan megfigyeléseket hasonlítunk-e össze, amelyek ugyanolyan, vagy eltérő x_2 értékkel rendelkeznek.
  - Például, pozitív korreláció esetén a nagyobb x_1 értékek nagyobb x_2 értékeket is jelentenek.
    - Ennek megfelelően tehát a megfelelő y-beli különbségek tehát nemcsak az x_1-beli különbségekből, hanem az x_2-beli különbségekből is adódhatnak.

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.
- Ha x_1 és x_2 korrelált, akkor számít, hogy olyan megfigyeléseket hasonlítunk-e össze, amelyek ugyanolyan, vagy eltérő x_2 értékkel rendelkeznek.
  - Például, pozitív korreláció esetén a nagyobb x_1 értékek nagyobb x_2 értékeket is jelentenek.
    - Ennek megfelelően tehát a megfelelő y-beli különbségek tehát nemcsak az x_1-beli különbségekből, hanem az x_2-beli különbségekből is adódhatnak.
- Példa: női bérhátrány - iparág

Különbség a meredekségekben: miből fakad a torzítás?

- Az egyváltozós regresszióban figyelmen kívül hagyjuk az x_2-beli különbségeket, és azokat a megfigyeléseket hasonlítjuk össze, amelyek különböző x_1 értékkel rendelkeznek.
- Ha x_1 és x_2 korrelált, akkor számít, hogy olyan megfigyeléseket hasonlítunk-e össze, amelyek ugyanolyan, vagy eltérő x_2 értékkel rendelkeznek.
  - Például, pozitív korreláció esetén a nagyobb x_1 értékek nagyobb x_2 értékeket is jelentenek.
    - Ennek megfelelően tehát a megfelelő y-beli különbségek tehát nemcsak az x_1-beli különbségekből, hanem az x_2-beli különbségekből is adódhatnak.
- Példa: női bérhátrány - iparág
- Az egyébként fontos x_2 figyelmen kívül hagyása *kihagyott változók miatti torzításhoz* vezet.

A kihagyott változó okozta torzítás: a megoldás

- Ha egy együttható értékére vagyunk kíváncsiak, a kihagyott változók fontosak:

A kihagyott változó okozta torzítás: a megoldás

- Ha egy együttható értékére vagyunk kíváncsiak, a kihagyott változók fontosak:
  - Ha van egy mérőszám/változó x_2-re, akkor használjuk és a probléma megoldva.

A kihagyott változó okozta torzítás: a megoldás

- Ha egy együttható értékére vagyunk kíváncsiak, a kihagyott változók fontosak:
  - Ha van egy mérőszám/változó x_2-re, akkor használjuk és a probléma megoldva.
  - Ha nincs mérőszám/változó x_2-re, akkor...
    - gondolkozzunk és "spekuláljunk"!
    - A 'valódi' paraméter a becsültnél vajon kisebb vagy nagyobb?

A kihagyott változó okozta torzítás: a megoldás (folyt.)

- Legyen az elméleti összefüggés: $y^E = \beta_0 + \beta_1 x + \beta_2 z$.

A kihagyott változó okozta torzítás: a megoldás (folyt.)

- Legyen az elméleti összefüggés: $y^E = \beta_0 + \beta_1 x + \beta_2 z$.
- Probléma: z-re nincs változónk, ezért a becsült modell: $y^E = \hat{\beta}_0 + \hat{\beta}_1 x$.

A kihagyott változó okozta torzítás: a megoldás (folyt.)

- Legyen az elméleti összefüggés: $y^E = \beta_0 + \beta_1 x + \beta_2 z$.
- Probléma: z-re nincs változónk, ezért a becsült modell: $y^E = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Tudjuk, hogy z fontos változó! Honnan tudjuk?

A kihagyott változó okozta torzítás: a megoldás (folyt.)

- Legyen az elméleti összefüggés: $y^E = \beta_0 + \beta_1 x + \beta_2 z$.
- Probléma: z-re nincs változónk, ezért a becsült modell: $y^E = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Tudjuk, hogy z fontos változó! Honnan tudjuk?
- Azt is tudjuk, hogy $Corr(x, z) \neq 0$. Honnan tudjuk?

A kihagyott változó okozta torzítás: a megoldás (folyt.)

- Legyen az elméleti összefüggés: $y^E = \beta_0 + \beta_1 x + \beta_2 z$.
- Probléma: z-re nincs változónk, ezért a becsült modell: $y^E = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Tudjuk, hogy z fontos változó! Honnan tudjuk?
- Azt is tudjuk, hogy $Corr(x, z) \neq 0$. Honnan tudjuk?
- A torzítás *irányát* meg tudjuk határozni:

	$Corr(x,z) > 0$	$Corr(x,z) < 0$
$\beta_2 > 0$	Pozitív torzítás	Negatív torzítás
$\beta_2 < 0$	Negatív torzítás	Pozitív torzítás

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!
3) Mit gondol, milyen kapcsolat van az ár és a csillagok száma között? Miért?

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!
3) Mit gondol, milyen kapcsolat van az ár és a csillagok száma között? Miért?
4) Mit gondol, milyen kapcsolat van a csillagok száma és a távolság között? Miért?

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!
3) Mit gondol, milyen kapcsolat van az ár és a csillagok száma között? Miért?
4) Mit gondol, milyen kapcsolat van a csillagok száma és a távolság között? Miért?
5) Ellenőrizze az adatokon az előző két pontra adott válaszát!

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!
3) Mit gondol, milyen kapcsolat van az ár és a csillagok száma között? Miért?
4) Mit gondol, milyen kapcsolat van a csillagok száma és a távolság között? Miért?
5) Ellenőrizze az adatokon az előző két pontra adott válaszát!
6) Az előzőek alapján mit gondol, hogyan változik a 2)-es ponthoz képest a távolság becsült együtthatója, ha a csillagok számát is betesszük a bal oldalra?

Feladat: 1. feladat (folyt.)

2) Futtasson egy egyváltozós regressziót, ahol az árat magyarázza a távolsággal. Értelmezze az együtthatókat!
3) Mit gondol, milyen kapcsolat van az ár és a csillagok száma között? Miért?
4) Mit gondol, milyen kapcsolat van a csillagok száma és a távolság között? Miért?
5) Ellenőrizze az adatokon az előző két pontra adott válaszát!
6) Az előzőek alapján mit gondol, hogyan változik a 2)-es ponthoz képest a távolság becsült együtthatója, ha a csillagok számát is betesszük a bal oldalra?
7) Becsülje meg a kérdéses többváltozós regressziót és értelmezze az eredményeket!

A többváltozós regresszió terminológiája

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség
  - kontrollálunk x_2-re

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség
  - kontrollálunk x_2-re
  - x_2-t rögzítve

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség
  - kontrollálunk x_2-re
  - x_2-t rögzítve
  - x_2-t változatlannak feltételezve

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség
  - kontrollálunk x_2-re
  - x_2-t rögzítve
  - x_2-t változatlannak feltételezve
- Amikor a többváltozós regresszióban az x_1 változóra összpontosítunk, akkor x_2-t kovariánsnak, vagy összemosó változónak is nevezzük.

Többváltozós regresszió: terminológia

- Az y változó különbsége az x_1 szerint, feltéve, hogy x_2 változatlan:
  - x_2 rögzítése melletti feltételes különbség
  - kontrollálunk x_2-re
  - x_2-t rögzítve
  - x_2-t változatlannak feltételezve
- Amikor a többváltozós regresszióban az x_1 változóra összpontosítunk, akkor x_2-t kovariánsnak, vagy összemosó változónak is nevezzük.
  - Ha az x_1 meredeksége más lesz, amikor x_2-t kihagyjuk, akkor azt mondjuk, hogy x_2 összemossa az y és az x_1 közötti kapcsolatot.

A konfidenciaintervallum

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.
  - Közepe.

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.
  - Közepe.
  - Kiszámolása.

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.
  - Közepe.
  - Kiszámolása.
- A standard hiba viszont más! Az egyszerű képlet:

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$$

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.
  - Közepe.
  - Kiszámolása.
- A standard hiba viszont más! Az egyszerű képlet:

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}}$$

  - Az SE kicsi, ha: a maradéktag szórása kicsi; a minta nagy; x_1 szórása nagy.

A konfidenciaintervallum

- Számos dologban ugyanaz, mint az egyváltozós regresszió esetében:
  - Jelentése.
  - Közepe.
  - Kiszámolása.
- A standard hiba viszont más! Az egyszerű képlet:

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}}$$

  - Az SE kicsi, ha: a maradéktag szórása kicsi; a minta nagy; x_1 szórása nagy.
  - Minél nagyobb az x_1 és x_2 közötti korreláció (R_1), annál nagyobb $\hat{\beta}_1$ standard hibája.

Multikollinearitás

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.
  - Az együtthatókat nem lehet kiszámolni (az egyiket kidobja a szoftver).

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.
  - Az együtthatókat nem lehet kiszámolni (az egyiket kidobja a szoftver).

- A magyarázó változók közötti erős, de nem tökéletes korrelációt *multikollinearitásnak* nevezzük:

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.
  - Az együtthatókat nem lehet kiszámolni (az egyiket kidobja a szoftver).

- A magyarázó változók közötti erős, de nem tökéletes korrelációt *multikollinearitásnak* nevezzük:
  - Ilyenkor a meredekségi együtthatók és standard hibáik meghatározhatóak, de:
    - a standard hibák nagyok lehetnek.

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.
  - Az együtthatókat nem lehet kiszámolni (az egyiket kidobja a szoftver).

- A magyarázó változók közötti erős, de nem tökéletes korrelációt *multikollinearitásnak* nevezzük:
  - Ilyenkor a meredekségi együtthatók és standard hibáik meghatározhatóak, de:
    - a standard hibák nagyok lehetnek.
    - a β értékét nem befolyásolja.

Magyarázó változók kollinearitása

A β standard hibája tehát: $SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}}$

- Tökéletes kollinearitásról beszélünk, amikor x_1 és x_2 között 1 vagy -1 a lineáris korrelációs együttható (azaz, egymás lineáris függvényei):
  - Egyik együttható SE-je sem létezik.
  - Az együtthatókat nem lehet kiszámolni (az egyiket kidobja a szoftver).

- A magyarázó változók közötti erős, de nem tökéletes korrelációt *multikollinearitásnak* nevezzük:
  - Ilyenkor a meredekségi együtthatók és standard hibáik meghatározhatóak, de:
    - a standard hibák nagyok lehetnek.
    - a β értékét nem befolyásolja.
  - Megoldási lehetőségek: egyik kidobása; a kettő változó kombinálása.

Hipotézistesztelés

Hipotézistesztelés

- Az egyes β-kra vonatkozó hipotézistesztek ugyanazok, mint az egyváltozós esetben.

Hipotézistesztelés

- Az egyes β-kra vonatkozó hipotézistesztek ugyanazok, mint az egyváltozós esetben.
- *Együttes hipotézistesztelés*: olyan nullhipotézis, amely egynél több regressziós együtthatóra vonatkozó állítást tartalmaz.

Hipotézistesztelés

- Az egyes β-kra vonatkozó hipotézistesztek ugyanazok, mint az egyváltozós esetben.
- *Együttes hipotézistesztelés*: olyan nullhipotézis, amely egynél több regressziós együtthatóra vonatkozó állítást tartalmaz.
- Tesztelhetjük, hogy a regresszióban *az összes meredekségi együttható értéke nulla-e*. $\rightarrow$ "globális F-teszt"
  - A teljes modell magyarázóerejét vizsgálja.

Feladat: 1. feladat (folyt.)

8) Értelmezze a 7)-es pontban becsült modell globális F-tesztjét!

Több magyarázó változó

Több magyarázó változó

- Több magyarázó változó bevonása egy egyértelmű kiterjesztés:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

Több magyarázó változó

- Több magyarázó változó bevonása egy egyértelmű kiterjesztés:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

- x_1 együtthatójának értelmezése: y értéke az adatokban átlagosan β_1 egységgel nagyobb azoknál a megfigyeléseknél, amelyeknél az x_1 egy egységgel nagyobb, de a többi x változó értéke azonos.

Több magyarázó változó

- Több magyarázó változó bevonása egy egyértelmű kiterjesztés:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...$$

- x_1 együtthatójának értelmezése: y értéke az adatokban átlagosan β_1 egységgel nagyobb azoknál a megfigyeléseknél, amelyeknél az x_1 egy egységgel nagyobb, de a többi x változó értéke azonos.
- Az (egyszerű) SE képlet:

$$SE(\hat{\beta}_k) = \frac{Std[e]}{\sqrt{n} Std[x_k] \sqrt{1 - R_k^2}}$$

  - SE kicsi, amikor R_k^2 kicsi
    - x_k *összes többi* x-en futtatott regressziójának R^2-e.

Kvalitatív magyarázó változó

Ismétlés: kvantitatív és kvalitatív változók

- Az adatok többféleképpen születhetnek, és a változóink az adott jelenség *minőségét* vagy *mennyiségét* írhatják le.

Ismétlés: kvantitatív és kvalitatív változók

- Az adatok többféleképpen születhetnek, és a változóink az adott jelenség *minőségét* vagy *mennyiségét* írhatják le.
- A kvantitatív változók számokként születnek.
- A kvalitatív változók nem számokként születnek, a megfigyeléseket verbálisan jellemzik és sajátos az értelmezésük (adott kategóriához tartozás).

Ismétlés: kvantitatív és kvalitatív változók

- Az adatok többféleképpen születhetnek, és a változóink az adott jelenség *minőségét* vagy *mennyiségét* írhatják le.
- A kvantitatív változók számokként születnek.
- A kvalitatív változók nem számokként születnek, a megfigyeléseket verbálisan jellemzik és sajátos az értelmezésük (adott kategóriához tartozás).
  - Számokat rendelünk hozzájuk *önkényesen*.

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.
- Hogyan vonjuk be ezeket a regressziós modellbe?

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.
- Hogyan vonjuk be ezeket a regressziós modellbe?
  - Ami eleve bináris, azt alakítsuk dummy változóvá és tegyük be.

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.
- Hogyan vonjuk be ezeket a regressziós modellbe?
  - Ami eleve bináris, azt alakítsuk dummy változóvá és tegyük be.
  - A kettőnél több értékű kvalitatív változóból hozzunk létre minden kategóriára dummy változókat és azokat tegyük be.

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.
- Hogyan vonjuk be ezeket a regressziós modellbe?
  - Ami eleve bináris, azt alakítsuk dummy változóvá és tegyük be.
  - A kettőnél több értékű kvalitatív változóból hozzunk létre minden kategóriára dummy változókat és azokat tegyük be.
    - Pontosabban: ha k különböző értéke van, akkor $k-1$ dummy változót tegyünk a jobb oldalra.

Kvalitatív változók használata: alapötlet

- A regresszióba bevonhatunk bináris és egyéb kvalitatív magyarázó változókat is.
- Hogyan vonjuk be ezeket a regressziós modellbe?
  - Ami eleve bináris, azt alakítsuk dummy változóvá és tegyük be.
  - A kettőnél több értékű kvalitatív változóból hozzunk létre minden kategóriára dummy változókat és azokat tegyük be.
    - Pontosabban: ha k különböző értéke van, akkor $k-1$ dummy változót tegyünk a jobb oldalra.
    - A kihagyott kategóriát nevezzük *bázis*nak, vagy *referenciakategória*nak, vagy *referenciacsoport*nak.

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Értelmezés:
  - β_0:

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Értelmezés:
  - β_0: y átlagát mutatja meg, ha minden x értéke nulla ($x_m = 0$ és $x_h = 0$), vagyis:

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Értelmezés:
  - β_0: y átlagát mutatja meg, ha minden x értéke nulla ($x_m = 0$ és $x_h = 0$), vagyis:
    - y átlagát mutatja meg a referenciakategóriában, $x = low$ esetében.

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Értelmezés:
  - β_0: y átlagát mutatja meg, ha minden x értéke nulla ($x_m = 0$ és $x_h = 0$), vagyis:
    - y átlagát mutatja meg a referenciakategóriában, $x = low$ esetében.
  - β_1 az y átlagában lévő különbséget mutatja az $x = medium$ és $x = low$ megfigyelések között.

Kvalitatív változók használata: Példa

- x egy kategorikus változó három értékkel: *low*, *medium* és *high*.
- Az x_m bináris változó jelezze, ha $x = medium$, x_h pedig jelezze, ha $x = high$.
- Az $x = low$ esetet nem vonjuk be a regresszióba (*referenciakategória*).

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Értelmezés:
  - β_0: y átlagát mutatja meg, ha minden x értéke nulla ($x_m = 0$ és $x_h = 0$), vagyis:
    - y átlagát mutatja meg a referenciakategóriában, $x = low$ esetében.
  - β_1 az y átlagában lévő különbséget mutatja az $x = medium$ és $x = low$ megfigyelések között.
  - β_2 az y átlagában lévő különbséget mutatja az $x = high$ és $x = low$ megfigyelések között.

Feladat: 1. feladat (folyt.)

9) Futtasson egy regressziót, ahol a bal oldalon az ár van, a jobb oldalon pedig egy dummy változó, amely 1, ha a szálláshely a központtól maximum 2 mérföldre van és 0 egyébként. Értelmezze az eredményeket!

Feladat: 1. feladat (folyt.)

9) Futtasson egy regressziót, ahol a bal oldalon az ár van, a jobb oldalon pedig egy dummy változó, amely 1, ha a szálláshely a központtól maximum 2 mérföldre van és 0 egyébként. Értelmezze az eredményeket!
10) Futtasson egy regressziót, ahol a bal oldalon az ár van, a jobb oldalon pedig az *accommodation_type* változó. Értelmezze az eredményeket!

Interakciók

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...
    - ... de kifejezetten utáljuk a töltöttkáposzta tetején.

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...
    - ... de kifejezetten utáljuk a töltöttkáposzta tetején.
  - Az y és x közötti kapcsolat eltérhet egy harmadik változó, z értékei szerint.

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...
    - ... de kifejezetten utáljuk a töltöttkáposzta tetején.
  - Az y és x közötti kapcsolat eltérhet egy harmadik változó, z értékei szerint.
  - Példa: A munkahelyi tapasztalattól pozitívan függ a bér, de ez a pozitív hatás a nőknél erősebb.

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...
    - ... de kifejezetten utáljuk a töltöttkáposzta tetején.
  - Az y és x közötti kapcsolat eltérhet egy harmadik változó, z értékei szerint.
  - Példa: A munkahelyi tapasztalattól pozitívan függ a bér, de ez a pozitív hatás a nőknél erősebb.
- Másképp: a megfigyelések egyes *részhalmazaiban* más-más általános mintázatot figyelhetünk meg.

Interakció: bevezetés

- Amikor egy kvalitatív változó különböző kategóriáira bináris változókat vonunk be a regresszióba, akkor az y átlagos különbségeit tárjuk fel.
  - Példa: bér - nő dummy
- Sokszor azonban egy változó befolyásolja két másik változó *közötti* kapcsolati mintázatot:
  - A csokiöntetet szeretjük a vaníliafagyin...
    - ... de kifejezetten utáljuk a töltöttkáposzta tetején.
  - Az y és x közötti kapcsolat eltérhet egy harmadik változó, z értékei szerint.
  - Példa: A munkahelyi tapasztalattól pozitívan függ a bér, de ez a pozitív hatás a nőknél erősebb.
- Másképp: a megfigyelések egyes *részhalmazaiban* más-más általános mintázatot figyelhetünk meg.
- Orvostudományi példa: egy *hatásmoderáló változó* csökkentheti/erősítheti egy gyógyszer emberekre gyakorolt hatását.

Interakció: fontosabb esetek

- interakció = attól függ

Interakció: fontosabb esetek

- interakció = attól függ
- Alapeset: x_1 folytonos, D pedig bináris.
- Több bináris változó esete: x_1 folytonos, D_1 és D_2 pedig bináris.
- Folytonos változók esete: x_1 és x_2 is folytonos.

Interakció: alapeset

- Regresszió két magyarázó változóval: x_1 folytonos, D pedig bináris, amely két csoportra bontja az adatokat (pl. női vagy férfi munkavállalók).

Interakció: alapeset

- Regresszió két magyarázó változóval: x_1 folytonos, D pedig bináris, amely két csoportra bontja az adatokat (pl. női vagy férfi munkavállalók).
- Azt szeretnénk megtudni, van-e különbség az y átlaga és az x_1 *közötti kapcsolatban* a $D = 1$, és a $D = 0$ megfigyelések között.

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- Interakciós eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- Interakciós eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$
  - Hogy néznek ki a prediktált értékek?

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- Interakciós eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- Interakciós eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- A két csoport önálló regressziói és a megfigyeléseket összevonó, de egy interakciós tagot tartalmazó regresszió pontosan ugyanazokat az együtthatóbecsléseket adják.

Interakció: alapeset (folyt.)

- Interakció nélküli eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- Interakciós eset: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$
  - Hogy néznek ki a prediktált értékek?
  - Hogy tudnánk lerajzolni?
- A két csoport önálló regressziói és a megfigyeléseket összevonó, de egy interakciós tagot tartalmazó regresszió pontosan ugyanazokat az együtthatóbecsléseket adják.
  - Az interakciót tartalmazó, összevont regresszió azonban lehetővé teszi, hogy közvetlenül teszteljük, azonosak-e a meredekségek.

Interakció: több bináris változó esete

- Általánosíthatunk három (vagy több) csoportra (legyen D_1, D_2 bináris és x egy folytonos változó):

$$y^E = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \beta_4 (D_1 \times x) + \beta_5 (D_2 \times x)$$

- Általánosságban, K darab kategóriával

$$y^E = \beta_0 + \beta_1 x + \sum_{k=2}^{K} \beta_k D_{k-1} + \beta_{K+k} (D_{k-1} \times x)$$

Interakció: két folytonos változó esete

- Ugyanaz a modell, két folytonos változóval, x_1 and x_2:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Példa: céges adat
  - y a bevétel változása, x_1 a globális kereslet változása, x_2 a cég pénzügyi állapota.
  - Az interakció megmutatja, hogy a kereslet csökkenése csökkenti a bevételt, de az erős mérleggel rendelkező cégek esetében kevésbé.

Interakció: két folytonos változó esete

- Ugyanaz a modell, két folytonos változóval, x_1 and x_2:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Példa: céges adat
  - y a bevétel változása, x_1 a globális kereslet változása, x_2 a cég pénzügyi állapota.
  - Az interakció megmutatja, hogy a kereslet csökkenése csökkenti a bevételt, de az erős mérleggel rendelkező cégek esetében kevésbé.

- Pontos értelmezés: deriválással.

Feladat: 1. feladat (folyt.)

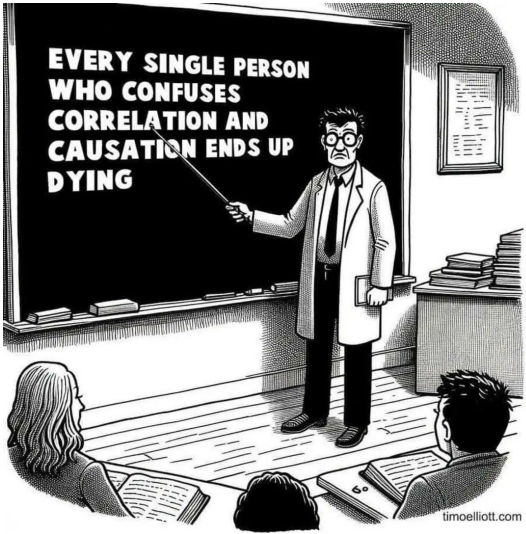
11) Futtasson egy regressziót, ahol a bal oldalon az ár, a jobb oldalon pedig 9)-es pontban létrehozott dummy, az *offer* dummy és azok interakciója van. Készítsen predikciót és értelmezze az eredményeket!

Feladat: 1. feladat (folyt.)

11) Futtasson egy regressziót, ahol a bal oldalon az ár, a jobb oldalon pedig 9)-es pontban létrehozott dummy, az *offer* dummy és azok interakciója van. Készítsen predikciót és értelmezze az eredményeket!

12) Futtasson egy regressziót, ahol a bal oldalon az ár, a jobb oldalon pedig a 9)-es pontban létrehozott dummy, a *stars* változó és azok interakciója van. Értelmezze az eredményeket!

Regresszió és okság


EVERY SINGLE PERSON WHO CONFUSES CORRELATION AND CAUSATION ENDS UP DYING
timoelliott.com

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.
  - De az jellemzően nincs, ami van: megfigyeléses adat.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.
  - De az jellemzően nincs, ami van: megfigyeléses adat.
- Más változókra történő kontrollálással közelebb kerülhetünk ahhoz, hogy hasonló dolgokat hasonlítsunk össze még megfigyeléses adatoknál is.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.
  - De az jellemzően nincs, ami van: megfigyeléses adat.
- Más változókra történő kontrollálással közelebb kerülhetünk ahhoz, hogy hasonló dolgokat hasonlítsunk össze még megfigyeléses adatoknál is.
  - De a közelebb kerülés nem egyenlő az eléréssel.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.
  - De az jellemzően nincs, ami van: megfigyeléses adat.
- Más változókra történő kontrollálással közelebb kerülhetünk ahhoz, hogy hasonló dolgokat hasonlítsunk össze még megfigyeléses adatoknál is.
  - De a közelebb kerülés nem egyenlő az eléréssel.
- Elvileg ezen javíthatunk, ha rögzítjük az *összes* lehetséges összemosó tényezőt: olyan változókat, amelyek egyszerre befolyásolnák y-t és x_1-et, az oksági változót.
  - Ceteris paribus = az *összes* ilyen fontos változóra történő kontrollálás.

Oksági elemzés többváltozós regresszióval

- Az egyik fő érv a többváltozós regresszióval történő becslés mellett az oksági értelmezéshez történő közelebb kerülés.
- Ami az oksági elemzéshez jó lenne: kísérleti adatok.
  - De az jellemzően nincs, ami van: megfigyeléses adat.
- Más változókra történő kontrollálással közelebb kerülhetünk ahhoz, hogy hasonló dolgokat hasonlítsunk össze még megfigyeléses adatoknál is.
  - De a közelebb kerülés nem egyenlő az eléréssel.
- Elvileg ezen javíthatunk, ha rögzítjük az *összes* lehetséges összemosó tényezőt: olyan változókat, amelyek egyszerre befolyásolnák y-t és x_1-et, az oksági változót.
  - Ceteris paribus = az *összes* ilyen fontos változóra történő kontrollálás.
- A valóságban mindenre kontrollálni lehetetlen.

Predikció

Predikció többváltozós regresszióval

- Az egyik ok, amiért többváltozós regressziót becslünk, az a *predikció*.

Predikció többváltozós regresszióval

- Az egyik ok, amiért többváltozós regressziót becslünk, az a *predikció*.
  - Keressük az y_j függő változó legjobb becslését egy adott j *célmegfigyelésre*.

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + ...$$

Predikció többváltozós regresszióval

- Az egyik ok, amiért többváltozós regressziót becslünk, az a *predikció*.
  - Keressük az y_j függő változó legjobb becslését egy adott j *célmegfigyelésre*.

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + ...$$

- Ha a cél a predikció, akkor azt akarjuk, hogy a regresszió a lehető legjobb illeszkedést eredményezze.
  - 'Jó illeszkedés' a j célmegfigyelést reprezentáló általános mintázathoz.

Predikció többváltozós regresszióval

- Az egyik ok, amiért többváltozós regressziót becslünk, az a *predikció*.
  - Keressük az y_j függő változó legjobb becslését egy adott j *célmegfigyelésre*.

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + ...$$

- Ha a cél a predikció, akkor azt akarjuk, hogy a regresszió a lehető legjobb illeszkedést eredményezze.
  - 'Jó illeszkedés' a j célmegfigyelést reprezentáló általános mintázathoz.
- Gyakori veszély az adatok *túlillesztése*: olyan mintázatokat találunk az adatokban, amelyek az általános mintázatban nem igazak.

Predikció többváltozós regresszióval

- Az egyik ok, amiért többváltozós regressziót becslünk, az a *predikció*.
  - Keressük az y_j függő változó legjobb becslését egy adott j *célmegfigyelésre*.

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + ...$$

- Ha a cél a predikció, akkor azt akarjuk, hogy a regresszió a lehető legjobb illeszkedést eredményezze.
  - 'Jó illeszkedés' a j célmegfigyelést reprezentáló általános mintázathoz.
- Gyakori veszély az adatok *túlillesztése*: olyan mintázatokat találunk az adatokban, amelyek az általános mintázatban nem igazak.
  - Az R^2 jó kiindulópont, de nem cél annak a maximalizálása.

A többváltozós regresszió illeszkedésének vizualizációja

A többváltozós regresszió illeszkedésének vizualizációja

- Az $\hat{y} - y$ *diagram* vízszintes tengelyén ábrázoljuk az $\hat{y}$-ot, függőleges tengelyén pedig az y-t.

A többváltozós regresszió illeszkedésének vizualizációja

- Az $\hat{y} - y$ *diagram* vízszintes tengelyén ábrázoljuk az $\hat{y}$-ot, függőleges tengelyén pedig az y-t.
  - A diagramon a 45 fokos egyenes és a körülötte lévő pontdiagram szerepel.

A többváltozós regresszió illeszkedésének vizualizációja

- Az $\hat{y} - y$ *diagram* vízszintes tengelyén ábrázoljuk az $\hat{y}$-ot, függőleges tengelyén pedig az y-t.
  - A diagramon a 45 fokos egyenes és a körülötte lévő pontdiagram szerepel.
- Az egyenes körüli pontdiagram azt mutatja, mennyire térnek el y tényleges értékei az prediktált értékeitől, $\hat{y}$-tól.

A többváltozós regresszió illeszkedésének vizualizációja

- Az $\hat{y} - y$ *diagram* vízszintes tengelyén ábrázoljuk az $\hat{y}$-ot, függőleges tengelyén pedig az y-t.
  - A diagramon a 45 fokos egyenes és a körülötte lévő pontdiagram szerepel.
- Az egyenes körüli pontdiagram azt mutatja, mennyire térnek el y tényleges értékei az prediktált értékeitől, $\hat{y}$-tól.
  - A 45 fokos egyenestől jobbra eső megfigyeléseket túl magasra becsüljük ($\hat{y} > y$).
  - A 45 fokos egyenestől balra eső megfigyeléseket túl alacsonyra becsüljük ($\hat{y} < y$).

Feladat: 1. feladat (folyt.)

13) Futtasson egy regressziót, ahol a bal oldalon az ln(ár), a jobb oldalon pedig a távolság, vendégértékelés és csillagok száma van. Készítsen $\hat{y} - y$ diagramot és keresse meg a legjobb ajánlatot!

Összefoglalás

Összefoglaló tanulságok

- A többváltozós regresszió egy több x változót tartalmazó lineáris modell.
- Bevonhatunk bináris változókat és interakciókat is.
- A többváltozós regresszió közelebb vihet minket az oksági értelmezéshez és segíthet jobb predikciók előállításában.

Házi feladat

A házi feladatban ismét a vendégmunkásokkal szembeni bérdiszkriminációt elemezzük, a múlt heti házi feladatot folytatva. Töltse le OSF-ről, majd nyissa meg a `morg-2014-emp.csv` adattáblát, és hajtsa végre a következőket:

1) Számolja ki az órabért és annak természetes alapú logaritmusát!
2) Tartsa meg azon embereket, akik vagy (i) külföldön születtek és nem amerikai állampolgárok (Foreign Born, Not a US Citizen), vagy (ii) az USA-ban születtek (Native, Born In US). Hozzon létre egy *native* nevű dummy változót, amely 0 az előbbiekre, és 1 az utóbbiakra.

A következő oldalon folytatódik!

Házi feladat (folyt.)

3) Vizsgáljuk ismét a "Driver/sales workers and truck drivers" foglalkozási kategóriát (occ2012=9130), a többieket távolítsa el a mintából!
4) A regressziós modellben a következő változókat fogja használni: órabér, a 2)-es pontban létrehozott dummy változó, a munkavállaló kora (age), a munkavállaló neme (sex), valamint a munkavállaló iskolai végzettsége (grade92). Készítsen ezekről a változókról egy táblázatot, amelyben szerepel minden változó esetén (i) annak pontos tartalma, (ii) hogy a változó kvalitatív vagy kvantitatív, és (iii) egy értelmes középértékmutató, valamint ha értelmezhető, akkor egy szóródási mutató. Az adatok részletes dokumentációja itt elérhető, nézze meg a használandó változókat!

A következő oldalon folytatódik!

Házi feladat (folyt.)

5) Futtasson egy többváltozós regressziót, amelyben bal oldalon a log órabér van, jobb oldalon pedig a következő változók: a 2)-es pontban létrehozott dummy változó, a munkavállaló kora és annak négyzete, a munkavállaló neme, valamint a munkavállaló iskolai végzettsége. Figyeljen, hogy minden változót értelmes módon tegyen be a modellbe! Az eredménytáblát másolja be a Moodle-be feltöltendő dokumentumba.
6) Értelmezze a modell fontosabb eredményeit: R^2, α, β-k és azok szignifikanciája.

A következő oldalon folytatódik!

Házi feladat (folyt.)

7) Szeretnénk ellenőrizni, hogy vajon különbözik-e a nők és férfiak esetében a "vendégmunkás-bérhátrány". Bővítse úgy az előző modellt, hogy válaszolni tudjon a kérdésre és válaszoljon is! (Hogyan bővítette a modellt és milyen eredményt kapott?)

A következő oldalon folytatódik!

Házi feladat (folyt.)

8) Gondolkozzon el, hogy milyen olyan fontos változók maradhattak ki a modellből, amelyek kihagyása torzítja a *native* nevű dummy változó becsült együtthatóját! Mondjon két példát és írja le azt is, hogy milyen irányba és miért torzítja az adott változó kihagyása a *native* dummy β-ját!