

A multikollinearitás

Ferenci Tamás
`tamas.ferenci@medstat.hu`

Utoljára frissítve: 2023. május 9.

Tartalom

1 Multikollinearitás

Tartalom

1 Multikollinearitás

A magyarázó változók körében rejlő egyéb probléma-lehetőségek

- Van egy másik oka is annak, hogy túl sok magyarázó változó használata miért lehet problémás: az, hogy a magyarázó változók a tipikus gyakorlati esetekben egymást is magyarázzák, vannak közöttük lineáris kapcsolatok
- Ezt a következő egyszerű példán mutatjuk be:

$$Y = \beta_0 + \beta_B \text{Ber} + \beta_F \text{Fo} + u,$$

- Tegyük most fel (nyilván nem igaz ilyen erősen, de nem teljesen elrugaszkodott), hogy a Bér-hez képest a Fő hozzáadása már felesleges, mégpedig azért mert „nem hordoz további információt” (ugyanazt írja le más szemszögből), mi mégis bevonjuk a modellünkbe

A magyarázó változók körében rejlő egyéb probléma-lehetőségek

- Van egy másik oka is annak, hogy túl sok magyarázó változó használata miért lehet problémás: az, hogy a magyarázó változók a tipikus gyakorlati esetekben egymást is magyarázzák, vannak közöttük lineáris kapcsolatok
- Ezt a következő egyszerű példán mutatjuk be:

$$Y = \beta_0 + \beta_B \text{Ber} + \beta_F \text{Fo} + u,$$

- Tegyük most fel (nyilván nem igaz ilyen erősen, de nem teljesen elrugaszkodott), hogy a Bér-hez képest a Fő hozzáadása már felesleges, mégpedig azért mert „nem hordoz további információt” (ugyanazt írja le más szemszögből), mi mégis bevonjuk a modellünkbe

A magyarázó változók körében rejlő egyéb probléma-lehetőségek

- Van egy másik oka is annak, hogy túl sok magyarázó változó használata miért lehet problémás: az, hogy a magyarázó változók a tipikus gyakorlati esetekben egymást is magyarázzák, vannak közöttük lineáris kapcsolatok
- Ezt a következő egyszerű példán mutatjuk be:

$$Y = \beta_0 + \beta_B \text{Ber} + \beta_F \text{Fo} + u,$$

- Tegyük most fel (nyilván nem igaz ilyen erősen, de nem teljesen elrugaszkodott), hogy a Bér-hez képest a Fő hozzáadása már felesleges, mégpedig azért mert „nem hordoz további információt” (ugyanazt írja le más szemszögből), mi mégis bevonjuk a modellünkbe

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becsülhetőség
- Vigyázat: *együtt* becsülhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becslhetőség
- Vigyázat: *együtt* becslhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becsléhetőség
- Vigyázat: *együtt* becslhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becslhetőség
- Vigyázat: *együtt* becslhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becslhetőség
- Vigyázat: *együtt* becslhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás

- Mi történik ilyenkor? → a magyarázó változók egymást is magyarázni fogják
- Egyre rosszabb a becslhetőség
- Vigyázat: *együtt* becslhetőek, csak külön-külön nem – a probléma épp az, hogy csak nagyon bizonytalanul lesznek elkülöníthetőek a hatások!
- Ez a *multikollinearitás*: az a jelenség, hogy a magyarázó változók lineáris kapcsolatban vannak egymással
- Bár nem tökéletesen precíz, de ezt a gyakorlatban azzal jellemezzük, hogy mennyire magyarázzák egymást
- Ennek megfelelő mérőszám az ún. *tolerancia*:

$$\text{Tol (Ber)} = 1 - R_{\text{Ber}|\text{Fo}}^2$$

Multikollinearitás leírása

- Általában: a vizsgálat magyarázó változót mennyire magyarázza a többi magyarázó változó, tehát

$$\text{Tol}(j) = 1 - R_j^2 = 1 - R_{\mathbf{x}_j | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k}^2$$

- Minél nagyobb R_j^2 , annál kisebb a tolerancia \rightarrow intuitíve: annál kevesebb többletinformációt hoz be ez a változó a modellbe a többi magyarázó változó mellett

Multikollinearitás leírása

- Általában: a vizsgálat magyarázó változót mennyire magyarázza a többi magyarázó változó, tehát

$$\text{Tol}(j) = 1 - R_j^2 = 1 - R_{\mathbf{x}_j | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k}^2$$

- Minél nagyobb R_j^2 , annál kisebb a tolerancia \rightarrow intuitíve: annál kevesebb többletinformációt hoz be ez a változó a modellbe a többi magyarázó változó mellett

Multikollinearitás hatása

- Írjuk most fel egy már bent levő változó koefficiensének mintavételi varianciáját:

$$\mathbb{D}^2 \left(\hat{\beta}_j \right) = \frac{ESS / (n - k - 1)}{(n - 1) \mathbb{D}^2 (X_j)} \cdot \frac{1}{\text{Tol}(j)} = \frac{\widehat{\sigma^2}}{(n - 1) \mathbb{D}^2 (X_j)} \cdot \frac{1}{\text{Tol}(j)}$$

- Látszik, hogy egy magyarázó változó koefficiensének a mintavételi varianciája c. p. *nő*, ahogy a tolerancia romlik (csökken); elvi minimum erre a varianciára a tolerancia = 1-nél
- Itt a c.p.-t úgy képzeljük el, mintha tudnánk *csak* a multikollinearitást változtatni
- De figyelem: a multikollinearitás, bármilyen közel is van 1-hez az R_j^2 , *nem* megsértése a standard modellfeltevéseknek (hacsak nem egzakt)

Multikollinearitás hatása

- Írjuk most fel egy már bent levő változó koefficiensének mintavételi varianciáját:

$$\mathbb{D}^2 \left(\hat{\beta}_j \right) = \frac{ESS / (n - k - 1)}{(n - 1) \mathbb{D}^2 (X_j)} \cdot \frac{1}{\text{Tol}(j)} = \frac{\widehat{\sigma^2}}{(n - 1) \mathbb{D}^2 (X_j)} \cdot \frac{1}{\text{Tol}(j)}$$

- Látszik, hogy egy magyarázó változó koefficiensének a mintavételi varianciája c. p. *nő*, ahogy a tolerancia romlik (csökken); elvi minimum erre a varianciára a tolerancia = 1-nél
- Itt a c.p.-t úgy képzeljük el, mintha tudnánk *csak* a multikollinearitást változtatni
- De figyelem: a multikollinearitás, bármilyen közel is van 1-hez az R_j^2 , *nem* megsértése a standard modellfeltevéseknek (hacsak nem egzakt)

Multikollinearitás hatása

- Írjuk most fel egy már bent levő változó koefficiensének mintavételi varianciáját:

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

- Látszik, hogy egy magyarázó változó koefficiensének a mintavételi varianciája c. p. *nő*, ahogy a tolerancia romlik (csökken); elvi minimum erre a varianciára a tolerancia = 1-nél
- Itt a c.p.-t úgy képzeljük el, mintha tudnánk *csak* a multikollinearitást változtatni
- De figyelme: a multikollinearitás, bármilyen közel is van 1-hez az R_j^2 , *nem* megsértése a standard modellfeltevéseknek (hacsak nem egzakt)

Multikollinearitás hatása

- Írjuk most fel egy már bent levő változó koefficiensének mintavételi varianciáját:

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

- Látszik, hogy egy magyarázó változó koefficiensének a mintavételi varianciája c. p. *nő*, ahogy a tolerancia romlik (csökken); elvi minimum erre a varianciára a tolerancia = 1-nél
- Itt a c.p.-t úgy képzeljük el, mintha tudnánk *csak* a multikollinearitást változtatni
- De figyelem: a multikollinearitás, bármilyen közel is van 1-hez az R_j^2 , *nem* megsértése a standard modellfeltevéseknek (hacsak nem egzakt)

A multikollinearitás mérése

- Bevezetjük a variancia infláló tényezőt (VIF):

$$\text{VIF}(j) = \frac{1}{\text{Tol}(j)}$$

- $\text{VIF}(j) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\text{VIF}(j) = 2$: a mintavételi variancia megduplázódott *pusztán a multikollinearitás miatt* (tehát amiatt, hogy a magyarázó változók egymást is magyarázzák) *ahhoz képest* mintha nem lenne multikollinearitás stb.
- A használatával kapcsolatban vannak bizonyos fenntartások!

A multikollinearitás mérése

- Bevezetjük a variancia infláló tényezőt (VIF):

$$\text{VIF}(j) = \frac{1}{\text{Tol}(j)}$$

- $\text{VIF}(j) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\text{VIF}(j) = 2$: a mintavételi variancia megduplázódott *pusztán a multikollinearitás miatt* (tehát amiatt, hogy a magyarázó változók egymást is magyarázzák) *ahhoz képest* mintha nem lenne multikollinearitás stb.
- A használatával kapcsolatban vannak bizonyos fenntartások!

A multikollinearitás mérése

- Bevezetjük a variancia infláló tényezőt (VIF):

$$\text{VIF}(j) = \frac{1}{\text{Tol}(j)}$$

- $\text{VIF}(j) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\text{VIF}(j) = 2$: a mintavételi variancia megduplázódott *pusztán a multikollinearitás miatt* (tehát amiatt, hogy a magyarázó változók egymást is magyarázzák) *ahhoz képest* mintha nem lenne multikollinearitás stb.
- A használatával kapcsolatban vannak bizonyos fenntartások!