A multikollinearitás

Ferenci Tamás
tamas.ferenci@medstat.hu

Utoljára frissítve: 2023. május 12.

Tartalom

Tartalom

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:

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

- 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:

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

- 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:

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

- 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? $\to$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás

- Mi történik ilyenkor? $\rightarrow$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás

- Mi történik ilyenkor? $\rightarrow$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás

- Mi történik ilyenkor? $\rightarrow$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás

- Mi történik ilyenkor? $\rightarrow$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás

- Mi történik ilyenkor? $\rightarrow$ 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*:

$$\mathrm{Tol}\left(\mathrm{Ber}\right) = 1 - R^2_{\mathrm{Ber}|\mathrm{Fo}}$$

Multikollinearitás leírása

- Általában: a vizsgálat magyarázó változót mennyire magyarázza a többi magyarázó változó, tehát
$$\mathrm{Tol}\,(j) = 1 - R_j^2 = 1 - R^2_{\mathbf{x_j}|\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_{j-1}},\mathbf{x_{j+1}},\ldots,\mathbf{x_k}}$$
- Minél nagyobb R_j^2, annál kisebb a tolerancia $\to$ 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
$$\mathrm{Tol}\left(j\right) = 1 - R_j^2 = 1 - R^2_{\mathbf{X_j}|\mathbf{X_1},\mathbf{X_2},\ldots,\mathbf{X_{j-1}},\mathbf{X_{j+1}},\ldots,\mathbf{X_k}}$$
- 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:

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

- 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:

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

- 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:

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

- 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:

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

- 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):

$$\mathrm{VIF}\left(j\right) = \frac{1}{\mathrm{Tol}\left(j\right)}$$

- $\mathrm{VIF}\left(j\right) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\mathrm{VIF}\left(j\right) = 2$: 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):

$$\mathrm{VIF}\left(j\right) = \frac{1}{\mathrm{Tol}\left(j\right)}$$

- $\mathrm{VIF}\left(j\right) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\mathrm{VIF}\left(j\right) = 2$: 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):

$$\mathrm{VIF}\left(j\right) = \frac{1}{\mathrm{Tol}\left(j\right)}$$

- $\mathrm{VIF}\left(j\right) = 1$ jelentése: a fenti variancia az elvi minimum (tehát: a magyarázó változót egyáltalán nem magyarázza a többi magyarázó változó); $\mathrm{VIF}\left(j\right) = 2$: 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!