

Regresioni i thjeshtë linear

Një model i udhëzuar i vlerave të përshtatura, pjerrësive dhe ndryshueshmërisë reziduale

Ratiomera Statistics

Përmbajtja

Përmbledhja e faqes së kursit	2
Vija e përshtatur	2
Inferenca dhe diagnostikimi	2
Shtrirja e përfundimit	3
Referencë e zgjeruar	4
Qëllimi dhe bazat	4
Idetë kryesore	4
Udhëzuesi i formulave	5
Si lexohet figura shpjeguese	7
Lista e kontrollit për interpretim	8
Si lidhet kjo temë me të tjerat	8

Përmbledhja e faqes së kursit

Regresioni i thjeshtë linear modelon mesataren e një rezultati sasior nga një parashikues. Nuk e parashikon saktë çdo rast. Vija e përshtatur përshkruan mesataren e kushtëzuar, ndërsa çdo rezidual regjistron dallimin mes vlerës së vërtetuar dhe asaj të përshtatur për një rast.

Vija e përshtatur

Modeli i popullatës është

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

ndërsa vija e përshtatur e kampionit është

$$\hat{Y}_i = b_0 + b_1 X_i.$$

Tabela 1: Çdo madhësi e regresionit i përgjigjet një pjese të ndryshme të pyetjes për modelin e përshtatur.

Madhësia	Kuptimi
b_0	Rezultati mesatar i përshtatur kur $X = 0$; interpretoje vetëm kur zeroja ka kuptim dhe mbështetet nga të dhënat
b_1	Ndryshimi i përshtatur në rezultatin mesatar për rritje prej një njësie në X
$\hat{Y}_i$	Rezultati mesatar i përshtatur në vlerën e parashikuesit për rastin i
$e_i = Y_i - \hat{Y}_i$	Reziduali vertikal; vlera pozitive do të thotë se rezultati i vërtetuar gjendet mbi vijën
R^2	Përqindja e ndryshueshmërisë së rezultatit në kampion që përfaqësohet nga vija e përshtatur
Gabimi standard i rezidualëve	Ndryshueshmëria tipike e rezultatit që mbetet rreth vijës së përshtatur, në njësitë e rezultatit

Pjerrësia lidhet drejtpërdrejt me Temën 4:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}.$$

Kovarianca jep lëvizjen e përbashkët, pjesëtimi me variancën e parashikuesit e shndërron atë në njësi rezultati për njësi parashikuesi, ndërsa forma e korrelacionit tregon si rikthehet lidhja pa njësi në shkallët fillestare. Në regresionin e thjeshtë me prerje, $R^2 = r^2$.

Inferenca dhe diagnostikimi

Intervali i besimit dhe testi për pjerrësinë përdorin

$$t = \frac{b_1 - 0}{SE(b_1)}, \quad df = n - 2.$$

Para se t'i besosh këtij përfundimi inferencial, kontrollo nëse një vijë e drejtë është përmbledhje e përshtatshme, nëse shpërndarja e rezidualëve është afërsisht e qëndrueshme, nëse shpërndarja e gabimeve përputhet me modelin dhe nëse rastet janë të pavarura sipas dizajnit. Diagrami i rezidualëve kundrejt vlerave të përshtatura kontrollon formën dhe shpërndarjen. Diagrami normal

Q-Q kontrollon trajtën e përgjithshme të gabimeve. Vlera leverage dhe distanca e Cook-ut i dallojnë rastet që duhen shqyrtuar, jo rastet që duhen fshirë automatikisht.

Shtrirja e përfundimit

- Interpolimi qëndron brenda intervalit të mbështetur të parashikuesit; ekstrapolimi shkon përtej tij dhe kërkon supozime më të forta.
- Parashikimi për një individ kërkon më shumë pasiguri se intervali për rezultatin mesatar në të njëjtën vlerë të parashikuesit.
- Një vlerë e vogël p nuk provon shkakësi, matje të saktë apo parashikim të mirë në një popullatë të re.
- Një R^2 i madh nuk provon se modeli është përcaktuar saktë.

Tema 6 e përdor idenë e rezidualit dy herë: i përshtat të dyja ndryshoret kryesore për një ndryshore të tretë dhe korrelohet pjesët që mbeten. Kjo e bën korrelacionin e pjesshëm hapin e natyrshëm pas vijës me një parashikues.

Referencë e zgjeruar

Qëllimi dhe bazat

Regresioni i thjeshtë linear përshkruan si ndryshon mesatarja e kushtëzuar e një ndryshoreje numerike të rezultatit nëpër vlerat e një ndryshoreje parashikuese. **Ndryshorja e rezultatit**, e shënuar me Y , është ndryshorja që modeli synon të shpjegojë ose parashikojë. **Ndryshorja parashikuese**, e shënuar me X , jep informacionin shpjegues. «I thjeshtë» do të thotë se modeli përmban një ndryshore parashikuese, ndërsa «linear» do të thotë se mesatarja e modeluar e Y ndryshon përgjatë një vije të drejtë kur ndryshon X .

Çdo rast i vrojtuar ka një vlerë parashikuese x_i dhe një vlerë rezultati y_i . Modeli jep një vlerë të përshtatur $\hat{y}_i$, pra nivelin e rezultatit që parashikon vija e përshtatur te vlera parashikuese e atij rasti. Dallimi mes rezultatit të vrojtuar dhe atij të përshtatur është **reziduali** e_i . Një rezidual pozitiv e vendos pikën mbi vijë; një rezidual negativ e vendos poshtë saj. Rezidualët e ruajnë njësinë e rezultatit dhe tregojnë pjesën e secilës vlerë të vrojtuar që vija e përshtatur nuk e riprodhoi.

Regresioni lidhet ngushtë me kovariancën dhe korrelacionin, por rolet e tij janë të ndryshme. Korrelacioni i trajton dy ndryshoret në mënyrë simetrike dhe përmbledh lidhjen lineare të standardizuar. Regresioni u cakton role të dallueshme ndryshores parashikuese dhe asaj të rezultatit, ruan njësinë e rezultatit dhe jep një ekuacion për vlerat e përshtatura. Prandaj ndërrimi i X me Y prodhon një problem tjetër regresioni, edhe pse korrelacioni i tyre mbetet i njëjtë.

Përbërësi	Kuptimi	Njësia
Konstanta b_0	Rezultati mesatar i përshtatur kur $X = 0$	Njësi të rezultatit
Pjerrësia b_1	Ndryshimi i përshtatur i rezultatit për një rritje prej një njësie në X	Njësi rezultati për njësi parashikuese
Vlera e përshtatur $\hat{y}_i$	Pika në vijën e përshtatur te x_i	Njësi të rezultatit
Reziduali e_i	Rezultati i vrojtuar minus rezultatit e përshtatur	Njësi të rezultatit

Idetë kryesore

Konstanta është matematikisht e nevojshme për ta pozicionuar vijën, por interpretimi i saj përmbajtësor varet nga fakti nëse zeroja ka kuptim dhe përfaqësohet nga të dhënat. Nëse vlerat e vrojtuar të ndryshores parashikuese janë larg zeros, konstanta është ekstrapolim. Në atë situatë, raportoje si koeficient modeli pa i dhënë një interpretim bazë nga bota reale, të cilin diapazoni i vrojtuar nuk mund ta mbështesë.

Pjerrësia është koeficienti qendror. Pjerrësia pozitive do të thotë se rezultati mesatar i përshtatur rritet kur rritet ndryshorja parashikuese. Pjerrësia negative do të thotë se ai ulet. Madhësia duhet lexuar me njësitë e të dyja ndryshoreve. Pjerrësia dy do të thotë dy njësi rezultati për një njësi parashikuese, jo korrelacion prej dy. Pjerrësia përshkruan një model mesatar të kushtëzuar dhe jo një ndryshim të garantuar për çdo rast.

Metoda e zakonshme e katrorëve më të vegjël zgjedh konstantën dhe pjerrësinë që minimizojnë shumën e rezidualëve të ngritura në katror. Ngritja në katror pengon anulimin e rezidualëve pozitive me ato negative dhe u jep peshë më të madhe mospërputhjeve më të mëdha. Kur modeli përmban një konstantë, vija e përshtatur kalon nëpër $(\bar{x}, \bar{y})$. Atëherë shuma e rezidualëve është afërsisht zero, duke lënë mënjatë rrumbullakimin numerik.

Veçoria diagnostikuese	Modeli i dëshiruar	Shqetësimi që sugjeron një model i dukshëm
Rezidualët kundrejt vlerave të përshtatura	Brez pa strukturë rreth zeros	Lakim, shpërhapje që ndryshon ose strukturë e lënë jashtë
Shpërhapja e rezidualëve	Afërsisht e njëjtë nëpër vlerat e përshtatura	Variancë e kushtëzuar jo e pandryshueshme
Krahasimi kuantil i rezidualëve	Model afërsisht i drejtë kur përdoret inferenca me gabime normale	Devijime të forta në bishta ose reziduale të pazakonta
Leva dhe ndikimi	Asnjë rast i vetëm nuk e mbizotëron vijën e përshtatur	Një rast me pozicion të pazakontë parashikues mund t'i ndryshojë fort koeficientët

R^2 e krahason ndryshueshmërinë reziduale pas përshtatjes së vijës me ndryshueshmërinë e përgjithshme rreth mesatares së rezultatit. Për një model të zakonshëm me konstantë, ai shtrihet nga zero në një. Një vlerë më e madhe do të thotë se vija e përshtatur përfaqëson më shumë ndryshueshmëri të kampionit në Y , por ajo nuk vërteton shkakësi, nuk garanton parashikime të sakta për individët dhe nuk dëshmon se forma e modelit është e përshtatshme. Një R^2 i lartë mund të shoqërohet me një model sistematik të rezidualëve.

Inferenca për pjerrësinë pyet nëse një lidhje lineare e popullatës pajtohet me zeron sipas modelit. Intervali i besimit tregon cilat vlera të pjerrësisë pajtohen me vlerësimin dhe gabimin e tij standard. Supozimet lidhen me marrëdhënien e kushtëzuar: strukturë lineare e mesatares, vrojtme të pavarura, variancë e përshtatshme e rezidualëve dhe një shpërndarje reziduale e mjaftueshme për inferencën e synuar. Vetë ndryshorja parashikuese nuk ka pse të jetë e shpërndarë normalisht.

Parashikimi i një pike kërkon kujdes. Vlera e përshtatur vlerëson rezultatin mesatar të kushtëzuar të një vlerë e zgjedhur parashikuese; ajo nuk garanton rezultatin e një individi. Parashikimi brenda diapazonit të vrojtuar është interpolim. Parashikimi përtej atij diapazoni është ekstrapolim dhe mbështetet në vazhdimin e patestuar të vijës së përshtatur, prandaj diapazoni i pambështetur duhet identifikuar shprehimisht.

Udhëzuesi i formulave

Modeli i popullatës e ndan një vijë sistematike nga një term individual gabimi:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 dhe β_1 janë koeficiente të popullatës. Gabimi ε_i përfaqëson dallimin mes rastit i dhe mesatares së kushtëzuar të popullatës. Pas përshtatjes së të dhënave të kampionit, vija e vlerësuar është:

$$\hat{y}_i = b_0 + b_1 x_i$$

Reziduali e krahason vlerën e vrojtuar me këtë vlerë të përshtatur:

$$e_i = y_i - \hat{y}_i$$

Pjerrësia e katrorëve më të vegjël mund të shkruhet duke përdorur prodhimet e kryqëzuara të paraqitura të kovarianca:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Pastaj konstanta e pozicionon vijën nëpër dy mesataret e kampionit:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Koeficienti i përcaktimit është pjesa e ndryshueshmërisë së përgjithshme të rezultatit që përfaqësohet nga modeli i përshtatur:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Me një ndryshore parashikuese dhe një konstantë, $R^2 = r_{xy}^2$. Pjerrësia mund të shprehet edhe si $b_1 = r_{xy}(s_y/s_x)$, gjë që tregon si përkthehet lidhja e standardizuar përsëri në njësitë fillestare të ndryshoreve.

Vija e përshtatur e ndan secilin devijim të vërtetuar në një përbërës të modelit dhe një përbërës rezidual:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Për një model OLS me konstantë, ngritja në katror dhe mbledhja nëpër të gjitha rastet japin

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{error}},$$

ku

$$SS_{\text{total}} = \sum_i (y_i - \bar{y})^2, \quad SS_{\text{model}} = \sum_i (\hat{y}_i - \bar{y})^2, \quad SS_{\text{error}} = \sum_i (y_i - \hat{y}_i)^2.$$

Ndarja e saktë e katrorëve është rezultat i modelit OLS për kampionin e plotë. Mos i ngri në katror tri largësitë me shenjë për një rast të vetëm duke pritur që i njëjti identitet të vlejë rast pas rasti.

Gabimi standard i rezidualëve vlerëson shpërhapjen tipike të gabimit në njësitë e rezultatit:

$$s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}.$$

Emëruesi përdor $n-2$, sepse vija vlerëson një konstantë dhe një pjerrësi. Kjo madhësi ndryshon nga gabimi standard i pjerrësisë, i cili mat pasigurinë e b_1 nëpër kampione të përsëritura hipotetike:

$$SE(b_1) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Për $H_0 : \beta_1 = 0$, testi i koeficientit dhe intervali përkatës i dyanshëm janë

$$t = \frac{b_1}{SE(b_1)}, \quad df = n - 2,$$

$$b_1 \pm t_{1-\alpha/2, n-2} SE(b_1).$$

Nëse intervali përkatës e përjashton zeron, testi përkatës i dyanshëm e hedh poshtë hipotezën zero të pjerrësisë së barabartë me zero. Përfshirja e zëros do të thotë se hipoteza zero nuk hidhet poshtë, jo se pjerrësia e popullatës është vërtetuar se është saktësisht zero.

Tabela e modelit i pjesëton shumën e katrorëve të modelit dhe të gabimit me shkallët e tyre të lirisë:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}.$$

Me një ndryshore parashikuese, testi i përgjithshëm i modelit dhe testi i dyanshëm i pjerrësisë bëjnë të njëjtën pyetje zero. Prandaj, nën të njëjtin model, $F = t^2$ dhe vlerat e tyre p përputhen.

Pyetja e parashikimit	Kuptimi	Kujdesi i nevojshëm
Vlera e përshtatur të x_0	Mesatarja e kushtëzuar e vlerësuar $\hat{y} = b_0 + b_1 x_0$	Nuk është rezultat i garantuar për një rast
Interpolimi	x_0 gjendet brenda diapazonit të vrojtuar të ndryshores parashikuese	Forma e modelit dhe diagnostika vazhdojnë të kenë rëndësi
Ekstrapolimi	x_0 gjendet jashtë diapazonit të vrojtuar të ndryshores parashikuese	Marrëdhënia mund të jetë ndryshe atje ku nuk u vrojtuan të dhëna

Cilësia e matjes i përket interpretimit. Sipas modelit klasik të gabimit në matjen e ndryshores parashikuese, të zhvilluar në materialin e dhënë, zhurma në X zakonisht e tërheq pjerrësinë e regresionit të thjeshtë drejt zeros. Ky dobësim mund ta bëjë një marrëdhënie themelore të duket më e dobët. Ai nuk nënkupton se çdo proces i mundshëm i gabimit në matje krijon të njëjtën anshmëri.

Si lexohet figura shpjeguese

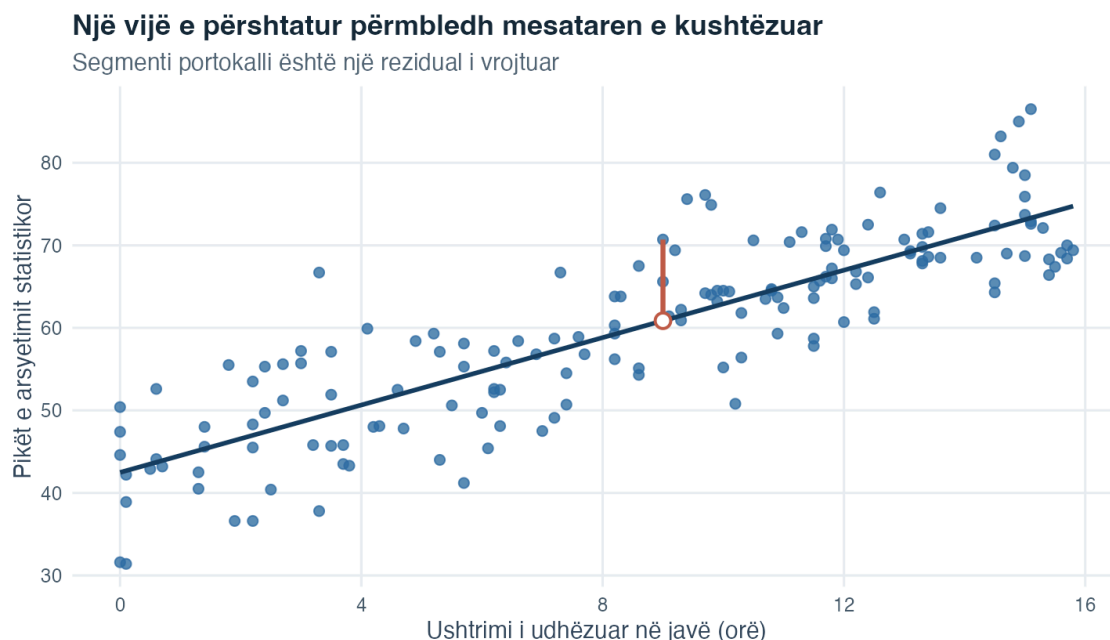


Figura 1: Diagram shpërndarjeje i orëve të praktikës së udhëzuar dhe pikëzimeve të arsyetit statistikor, me një vijë të përshtatur në rritje dhe një segment vertikal portokalli të rezidualit.

Boshti horizontal jep orët javore të praktikës së udhëzuar, ndërsa boshti vertikal jep pikëzimet e arsyetit statistikor. Çdo pikë blu është një rast i simuluar. Vija e errët ngrihet, prandaj në këtë grup të dhënash pikëzimi mesatar i përshtatur është më i lartë se vlerat më të mëdha të praktikës. Vija përshkruan modelin mesatar të modeluar. Pikat individuale mbeten të shpërndara sipër dhe poshtë saj, duke të kujtuar se informacioni parashikues nuk e përcakton rezultatin e çdo personi.

Afërsisht te nëntë orë praktikë, rrethi bosh shënon pikëzimin e përshtatur mbi vijë. Pika e vrojtuar për atë rast është më lart. Segmenti vertikal portokalli është reziduali: pikëzimi i vrojtuar minus pikëzimin e përshtatur. Drejtimi i tij është pozitiv dhe gjatësia matet në pikë të pikëzimit. Katrorët më të vegjël e bëjnë këtë krahasim për çdo pikë dhe zgjedhin vijën me shumën më të vogël të gjatësive të rezidualeve të ngritura në katror.

Grafiku mbështet një pohim për lidhjen lineare në të dhënat e simuluar. Ai nuk tregon se praktika shtesë shkaktoi pikëzime më të larta. Përgatitja paraprake, përzgjedhja, formati i ushtrimeve ose

ndryshore të tjera mund të lidhen me të dyja. Ai gjithashtu nuk tregon nëse supozimet për rezidualet qëndrojnë; duhen grafikë diagnostikues të veçantë për lakimin, ndryshimin e variancës, rezidualet e pazakonta dhe ndikimin.

Lista e kontrollit për interpretim

Emërto ndryshoren e rezultatit dhe ndryshoren parashikuese dhe trego njësitë e tyre. Shqyrto shpërndarjet dhe diagramin e shpërndarjes. Sigurohu se një përmbledhje në vijë të drejtë është e përshtatshme nëpër diapazonin e vrojtuar. Raporto ekuacionin e përshtatur dhe ktheje pjerrësinë në një fjali të plotë që përmban të dyja njësitë. Interpreto konstantën vetëm nëse vlera e saj referuese ka kuptim. Raporto R^2 si ndryshueshmëri të kampionit që përfaqësohet nga modeli, jo si përqindje shkakore.

Shqyrto diagnostikën e rezidualeve dhe ndikimit para se të mbështetesh te inferenca. Raporto vlerësimin e pjerrësisë, gabimin standard, intervalin e besimit, statistikën e testit, shkallët e lirisë dhe vlerën p kur janë të rëndësishme. Dallo një mesatare të kushtëzuar të vlerësuar nga një rezultat individual i garantuar dhe identifikoj ekstrapolimin. Përshkruaj rezultatet e simuluar si të simuluar dhe mbaje gjuhën e lidhjes të ndarë nga gjuha shkakore.

Si lidhet kjo temë me të tjerat

Ky model e bën konkrete lidhjen me kovariancën dhe korrelacionin. Kovarianca siguron numëruesin e pjerrësisë, varianca e ndryshores parashikuese siguron emëruesin dhe korrelacioni e standardizon të njëjtën përputhje lineare. Regresioni shton drejtim dhe njësi: ai pyet si ndryshon mesatarja e përshtatur e një rezultati të zgjedhur nëpër vlerat e ndryshores parashikuese.

Hapat vijues trajtojnë një kufizim qendror të vijës me një ndryshore parashikuese. Një ndryshore e tretë mund të shpjegojë një pjesë të lidhjes së vrojtuar. Korrelacioni i pjesshëm i heq nga të dyja ndryshoret qendrore përbërësit linearë që lidhen me atë ndryshore të tretë dhe korrelojnë pjesët që mbeten. Regresioni i shumëfishtë vendos disa ndryshore parashikuese në një model, kështu që secili koeficient përshkruan një lidhje të kushtëzuar ndërsa të tjerat mbahen të pandryshuara. Ideja e rezidualit e paraqitur këtu bëhet gjuha e përbashkët për të dyja zgjerimet.

ID-ja e dokumentit: topic-05-simple-linear-regression-summary-sq

Shënime 1 nga 3

Regresioni i thjeshtë linear | Tema 5

Shënime 2 nga 3

Regresioni i thjeshtë linear | Tema 5

Shënime 3 nga 3

Regresioni i thjeshtë linear | Tema 5