UNIVERSITATEA DIN BUCUREȘTI
VIRTUTE ET SAPIENTIA

Statistică și Metode de Cercetare Cantitativă în Psihologie și Științe Cognitive

Regresia Lineara

George Gunnesch-Luca

10.04.2024

Introducere în Regresia Lineara

Ce este Regresia?

Analizele de regresie sunt un set de tehnici statistice care permit evaluarea relației dintre o variabilă dependentă (DV) și mai multe variabile independente (IVs).

De exemplu, este capacitatea de citire în clasele primare (DV) legată de mai multe IV-uri precum dezvoltarea perceptivă, dezvoltarea motorie și vârsta?

De ce să utilizăm Regresia?

1. **Predicție**: ajustarea unui model predictiv la un set de date observate, apoi folosind acel model pentru a face predicții despre un rezultat dintr-un nou set de variabile explicative;
2. **Explicație**: ajustarea unui model pentru a explica relațiile dintre un set de variabile.

- Variabilă explicativă (x): expunere, predictor, variabilă independentă
- Variabilă dependentă (y): rezultat, răspuns
- Regresie liniară univariată (aka, simplă): o singură variabilă explicativă
- Regresie multivariată (multiplă): mai multe variabile explicative

În cazul regresiei liniare

Caracteristicile Regresiei Liniare

- Variabila dependentă *trebuie* să fie continuă, însă,
- Variabilele explicative pot fi continue sau categoriale

Scopul regresiei

Descrierea și Explicarea Cantitativă a Relațiilor

- Direcția relației
- Puterea relației
  Exemplu: Cum afectează efortul anterior al unui student nota lor la examen?

Estimarea (Predicția) Valorilor Necunoscute ale Variabilei Dependente

- Predicția variabilei dependente la un anumit nivel al variabilei independente
  Exemplu: Ce notă va obține un student cu efort de studiu mediu/ridicat/sub medie?

Factori ce Afectează Regresia

- Liniaritatea relației (non-liniaritate -> subestimare)
- Grupe extreme (-> supraestimare)
- Valori extreme (outliers) (-> distorsiune)
- Interval restricționat (-> subestimare)
- Subgrupe eterogene
- Abaterea de la distribuția normală bivariată

Modele de Regresie Liniară

Modelele de regresie liniară sunt un tip specific de modelare a datelor. Modelul este considerat liniar pentru că ajustăm o linie dreaptă "exact în mijlocul" datelor.

Linia "explică" datele: pentru fiecare valoare X, oferă o valoare Y prezisă.

O linie poate fi descrisă folosind

$$y = b_0 + b_1 x$$

unde y este valoarea prezisă (criteriul), $b_1 x$ este panta liniei, b_0 este interceptul pe Y când $x = 0$, iar x reprezintă valoarea predictorului.

Modele de Regresie Liniară

O regresie indică un "trend" în date bazat pe o linie de regresie, și este ideală pentru a face predicții: odată ce avem o linie, putem face predicții pentru valoarea Y (criteriul) pentru fiecare valoare X (predictorul).

Modelele liniare sunt utilizate în mod comun pentru două scopuri: predicția și descrierea trendului. Un trend se referă de obicei la panta liniei. În loc să spunem "descrierea unui trend", se poate spune de asemenea "explicarea unui criteriu bazat pe o funcție a predictorului."

Modele de Regresie Liniară

Într-o relație perfect liniară, toate punctele se află pe o linie. Totuși, acest lucru este rar cazul în datele din lumea reală. Pentru a estima forța relației liniare, trebuie să găsim o linie care reprezintă suficient de bine datele. Această linie se numește linia de regresie.

Cum trasam linia

Metoda OLS

- Linia de regresie este "linia cea mai potrivită" daca minimizează suma patratelor diferentelor dintre valorile observate si cele prezise (estimate).

Pentru o reprezentare grafică și o explicație detaliată citiți:

https://mlu-explain.github.io/linear-regression/

Metoda Celor Mai Mici Pătrate (OLS)

- OLS este o metodă fundamentală pentru ajustarea unei linii de regresie.
- **Principiu:** Minimizarea sumei diferențelor pătrate dintre valorile observate și cele prezise de linie.
- **Proces:**
  1. Calcularea reziduurilor (diferențele dintre valorile observate și cele prezise).
  2. Ridicarea la patrat a reziduurilor și adunarea acestora.
  3. Alegerea parametrilor liniei (panta, interceptul) care minimizează această sumă.

Înțelegerea Liniei de Regresie

Linia de regresie este o linie dreaptă care reprezintă relația dintre o variabilă predictor (X) și o variabilă rezultat (Y).

În regresia liniară simplă, linia de regresie poate fi descrisă folosind formula:

$y = \beta_0 + \beta_1 \, \mathrm{X} + \epsilon$

unde:

- Y este variabila rezultat
- X este variabila predictor
- β_0 este interceptul (valoarea lui Y când $\mathrm{X} = 0$)
- β_1 este panta (schimbarea în Y pentru o creștere cu o unitate în X)
- ϵ este termenul de eroare (diferența între valorile observate și cele prezise ale lui Y)

Explorează conceptul

https://argoshare.is.ed.ac.uk/simple_regression

Exemplu

$y = \beta_0 + \beta_1 \, \mathrm{X} + \epsilon$

Să presupunem că suntem interesați să înțelegem cum numărul de ore de muncă ale unui student înainte de un examen afectează performanța acestuia. Ipotezăm că munca în plus va duce la note mai bune la examen. Colectăm datele și trasăm linia folosind metoda OLS. Parametrii estimati sunt (exemplu imaginar):

- Formula de regresie: valoarea medie așteptată a lui y = 4 + 0.27 * Numărul de ore de muncă
- Intercept: 4 (scorul de performanță prezis când nu s-a efectuat nicio muncă)
- Panta: 0.27 (creșterea prezisă a scorului de performanță pentru fiecare oră suplimentară de muncă)

Coeficienți de Regresie Standardizați vs. Nestandardizați

Coeficienți Nestandardizați (B)

- Unități originale de măsurare
- Pot fi interpretați direct
- Utili pentru compararea impactului predictorilor în cadrul aceluiași model

Coeficienți Standardizați (Beta)

- Fără unitate (deviații standard)
- Permite comparația între modele diferite și predictorii cu unități diferite
- Utili pentru identificarea predictorilor cei mai influenți

Diferențe Cheie între Coeficienții de Regresie Standardizați și Nestandardizați

1. **Unități**: Coeficienții nestandardizați păstrează unitățile originale, în timp ce coeficienții standardizați sunt fără unitate.
2. **Interpretare**: Coeficienții nestandardizați sunt mai ușor de interpretat, în timp ce coeficienții standardizați sunt mai buni pentru compararea predictorilor cu unități diferite.
3. **Comparare**: Coeficienții nestandardizați sunt folosiți pentru compararea predictorilor în cadrul unui model, în timp ce coeficienții standardizați sunt folosiți pentru compararea predictorilor între modele.

Interpretarea Coeficienților de Regresie Standardizați vs. Nestandardizați în Regresia Simplă

Coeficienți Nestandardizați (b)

- Interpretare directă în unitățile originale
- Exemplu: Dacă $b = 3$, pentru fiecare creștere cu 1 unitate a lui X, rezultatul Y crește cu 3 unități.

Coeficienți Standardizați (β)

- Interpretare în termeni de deviații standard
- Exemplu: Dacă $\beta = 0.5$, pentru fiecare creștere cu 1 unitate a lui X, rezultatul Y crește cu 0.5DS.

R-pătrat (Coeficientul de Determinare)

- Reprezintă proporția variației în variabila dependentă (Y) explicată de variabila independentă (X)
- Variază de la 0 la 1
- Valorile mai mari ale R-pătrat indică o relație mai puternică între variabile

Interpretarea R-pătrat

- Exemplu: Dacă R-pătrat = 0.6, 60% din varianța lui Y este explicată de X.
- Un R-pătrat mai mare indică o potrivire mai bună a modelului la date
- În regresia simplă, R-pătrat este egal cu pătratul coeficientului de corelație (R)

R-pătrat Ajustat

- Ține cont de numărul de predictorii din model
- Se ajustează pentru includerea predictorilor suplimentari care s-ar putea să nu îmbunătățească puterea explicativă a modelului
- Mai util în regresia multiplă, dar poate fi folosit și în regresia simplă în scopuri comparative

Interpretarea R-pătrat Ajustat

- Valorile mai apropiate de 1 indică o potrivire mai bună a modelului
- O diferență mai mică între R-pătrat și R-pătrat ajustat sugerează că predictorii adăugați sunt utili
- În regresia simplă, R-pătrat ajustat este adesea foarte apropiat de R-pătrat, deoarece există doar un predictor

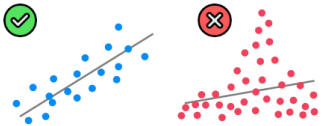
Asumptiile Regresiei Liniare

1. **Liniaritate**: Există o relație liniară între variabilele independente și variabila dependentă.
2. **Independență**: Observațiile sunt independente una de alta.
3. **Homoscedasticitate**: Varianța termenilor de eroare este constantă la toate nivelurile variabilelor independente.
4. **Normalitate**: Termenii de eroare sunt distribuiți normal.
5. **Fără multicolinearitate**: În regresia multiplă, variabilele independente nu sunt foarte corelate între ele.

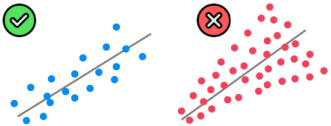
- Încălcarea acestor presupuneri poate duce la estimări inexacte sau părtinitoare.
- Este important să verificăm și să abordăm aceste presupuneri atunci când efectuăm o analiză de regresie liniară.

Asumptiile Regresiei Liniare

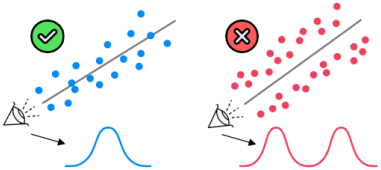
1. Linearity
(Linear relationship between Y and each X)



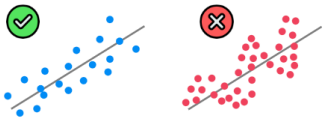
2. Homoscedasticity
(Equal variance)



3. Multivariate Normality
(Normality of error distribution)



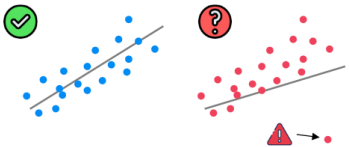
4. Independence
(of observations. Includes “no autocorrelation”)



5. Lack of Multicollinearity
(Predictors are not correlated with each other)

$X_1 \not\sim X_2$ $X_1 \sim X_2$

6. The Outlier Check
(This is not an assumption, but an “extra”)



Asumptiile Regresiei Liniare: Explicație Detaliată

1. **Liniaritate**
   - Relația dintre variabilele independente și variabila dependentă este liniară.
   - Verificare cu ajutorul graficelor de dispersie sau a graficelor reziduurilor.
   - Abordarea încălcărilor cu transformări ale datelor sau modele neliniare.
2. **Independență**
   - Observațiile sunt independente una de alta.
   - Adesea presupus în eșantioane aleatoare sau experimente.
   - Verificare cu testul Durbin-Watson pentru date de tip serie temporală.
   - Abordarea încălcărilor cu modele alternative (de exemplu, modele de serie temporală).

Asumptiile Regresiei Liniare: Explicație Detaliată

3. **Homoscedasticitate**
   - Varianța termenilor de eroare este constantă la toate nivelurile variabilelor independente.
   - Verificare cu graficele reziduurilor.
   - Abordarea încălcărilor cu metode de regresie ponderată prin cel mai mic pătrat, transformări ale datelor sau metode robuste de regresie.
4. **Normalitate**
   - Termenii de eroare sunt distribuiți normal.
   - Verificare cu histograme, grafice Q-Q sau teste de normalitate (de exemplu, testul Shapiro-Wilk).
   - Abordarea încălcărilor cu transformări ale datelor sau metode robuste de regresie.

Asumptiile Regresiei Liniare: Explicație Detaliată

5. **Multicolinearitate**
   - În regresia multiplă, variabilele independente nu au voie sa fie foarte corelate între ele.
   - Verificare cu coeficienții de corelație sau Factorul de Inflație al Variantei (VIF).
   - Abordarea încălcărilor prin eliminarea sau combinarea variabilelor foarte corelate, sau utilizând tehnici de reducere a dimensionalității (de exemplu, PCA).

Introducere în Regresia Multiplă

- **Recenzie**: Regresia liniară simplă modelează relația dintre un singur predictor și o variabilă de răspuns
- **Scop**: Extinderea acestui concept pentru a include mai mulți predictor, permițând o înțelegere mai cuprinzătoare a relațiilor și îmbunătățirea acurateței predicției
- **Beneficii**: Evaluarea impactului mai multor factori asupra unei variabile de răspuns, controlul variabilelor și construirea unor modele mai robuste și mai precise

Recapitulare Regresie Liniară Simplă

- **Model**: $y = \beta_0 + \beta_1 X + \epsilon$
- **Variabile**:
  - y: variabila dependentă (răspuns)
  - X: variabila independentă (predictor)
  - β_0: interceptul
  - β_1: coeficientul de regresie
  - ϵ: termenul de eroare
- **Scop**: Modelarea relației liniare dintre y și X

Regresia Multiplă: O Extensie

- **Model**: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$
- **Variabile**:
  - y: variabila dependentă (răspuns)
  - $X_1, X_2, \ldots, X_n$: variabilele independente (predictori)
  - β_0: interceptul
  - $\beta_1, \beta_2, \ldots, \beta_n$: coeficienții de regresie
  - ϵ: termenul de eroare
- **Scop**: Modelarea relației liniare dintre y și mai mulți predictor $(X_1, X_2, \ldots, X_n)$

Regresia Multiplă

Introducere în Controlul Variabilelor

Obiectiv: Izolarea efectului unor predictorii specifici prin controlul altor variabile. **Concept Cheie**: Variabilele de control sunt acelea care ar putea afecta atât variabila independentă (IV), cât și variabila dependentă (DV) și ar putea confunda rezultatele. **Importanță**: Ajută la înțelegerea relației reale între IV și DV prin eliminarea zgomotelor sau a efectelor confundatoare.

Cum Funcționează Controlul Variabilelor

Proces Pas cu Pas

1. Identificare Variabile: Determinați care variabile (IVs și Dvs) ar putea fi corelate și necesită analiză.
2. Selectare Variabile de Control: Alegeți care variabile ar putea confunda relația dintre IV-ul principal și DV.
3. Colectare Date: Asigurați-vă că datele includ toate variabilele relevante (principale și de control).

Implementarea Controlului în Regresie

Folosind Teorema Frisch-Waugh-Lovell

- Regresați Variabilele de Control (Z) pe IV (X): Obțineți reziduurile. Acestea reprezintă partea din X care nu este explicată de Z.
- Regresați Variabilele de Control (Z) pe DV (Y): Obțineți reziduurile. Acestea reprezintă partea din Y care nu este explicată de Z.
- Regresați Reziduurile lui X pe Reziduurile lui Y: Panta acestei regresii oferă efectul lui X asupra lui Y, controlând pentru Z.

Citiți în detaliu, capitolul 4.5 https://theeffectbook.net/ch-DescribingRelationships.html#conditional-conditional-means-a.k.a.-controlling-for-a-variable

Interpretarea Coeficienților de Regresie

Pentru predictorii continui!

- β_0 **(Intercept)**: Valoarea așteptată a lui y când toți predictorii sunt zero
- $\beta_1, \beta_2, \ldots, \beta_n$ **(Coeficienți)**: Schimbarea așteptată în y pentru o creștere cu o unitate a predictorului corespunzător, menținând constanți toți ceilalți predictorii!

Menținerea Constantă a Predictorilor în Regresia Multiplă

- Importanța izolării efectului unic al fiecărui predictor
- Interpretarea coeficienților presupune că ceilalți predictorii rămân constanți
- Nu este o manipulare explicită a datelor, ci un cadru conceptual pentru înțelegerea coeficienților

Cadru Conceptual: Menținerea Predictorilor Constant

- Permite o evaluare precisă a contribuției unice a fiecărui predictor
- Ajută la înțelegerea relațiilor dintre predictorii și variabila dependentă
- Nu se face o manipulare directă a datelor; presupunerile sunt făcute în timpul interpretării coeficienților

Menținerea Constantă a Predictorilor și Variabilele de Control

- Menținerea constantă a predictorilor: Cadru de interpretare pentru înțelegerea contribuțiilor unice ale predictorilor în regresia multiplă
- Variabilele de control: Variabile incluse în model pentru a contabiliza factorii confundatori potențiali

Scopul Variabilelor de Control

- Îmbunătățirea acurateței și fiabilității estimărilor de regresie
- Contabilizarea factorilor confundatori potențiali care ar putea afecta relația dintre predictorii și variabila dependentă
- Eliminarea parțială a efectelor variabilelor de control pentru a înțelege mai bine relația dintre predictorii de interes și variabila dependentă

Adăugarea Blocurilor de Control în Regresia Multiplă

- Blocuri de control: Gruparea variabilelor de control în blocuri separate pe baza naturii sau sursei lor
- Scop: Contabilizarea sistematică a factorilor confundatori potențiali și îmbunătățirea interpretabilității modelului
- Pași:
  1. Identificarea predictorilor primari de interes și a variabilelor de control potențiale
  2. Gruparea variabilelor de control în blocuri
  3. Adăugarea blocurilor de control secvențial în modelul de regresie și evaluarea schimbărilor coeficienților predictorilor primari

Beneficiile Blocurilor de Control în Regresia Multiplă

- Înțelegerea îmbunătățită a relațiilor dintre predictorii și variabila dependentă
- Identificarea factorilor confundatori potențiali
- Abordare sistematică a adăugării variabilelor de control în model
- Îmbunătățirea interpretabilității modelului

Exemplu de Regresie Multiplă: Predicția Scorurilor la Examen Bazată pe Orele de Somn și Orele de Studiu

Să presupunem că dorim să explorăm mai departe factorii care afectează performanța studenților la examene. Pe lângă numărul de ore de muncă, acum luăm în considerare și numărul de ore de somn ca predictor potențial al scorurilor la examen.

Scorul așteptat la examen $= 2 + 0.25 \times$ Numărul de ore de somn $+ 0.5 \times$ Numărul de ore de studiu

- Intercept (2): Acesta reprezintă scorul prezis pentru un student care nici nu a dormit, nici nu a studiat. Este scorul de bază în cele mai adverse condiții.
- Panta pentru Somn (0.25): Acest coeficient sugerează că fiecare oră suplimentară de somn este asociată cu o creștere de 0.25 puncte a scorului la examen, menținând constante orele de studiu.
- Panta pentru Orele de Studiu (0.5): Acest coeficient indică faptul că fiecare oră suplimentară de studiu este așteptată să crească scorul la examen cu 0.5 puncte, menținând constante orele de somn.

Modele Aditive vs. Multiplicative

Coeficient	Model Aditiv	Model de Interacțiune
β_0 (Intercept)	Același lucru	Y când atât consumul de cafea, cât și fumatul $= 0$. Rezultatul baseline pentru nefumătorii care nu consumă cafea.
β_1 (Cafea)	Schimb. în Y pentru schimb. cu 1u. în consumul de cafea, presupunând că efectul este constant indif. de stat. de fumător.	Schimbarea în Y pentru o creștere de 1 unitate în consumul de cafea *specific* când fumatul $= 0$. Pentru fumători, efectul cafelei este modificat de termenul de interacțiune (β_3).
β_2 (Fumat)	**Diferența** în Y datorată fumatului, presupunând că efectul fumatului este constant la diferite niveluri de consum de cafea.	Efectul fumatului asupra lui Y când consumul de cafea=0. Este independent de cafea, cu excepția cazului în care este luat în considerare termenul de interacțiune (β_3), care mod. ef. fumatului în funcție de consumul de cafea.
β_3 (Interacțiunea Cafea:Fumat)	Nu este aplicabil.	Efectul *suplimentar* asupra lui Y pentru o creștere de 1 unitate în consumul de cafea pentru fumători **peste** ce este prezis de β_1. Modificiarea efectului cafelei de catre fumat.