Recapitulare Regresie Liniară Simplă

Căutăm un model liniar cu eroare ireductibilă ϵ:

$$Y = f(X) + \epsilon = \beta_1 X + \beta_0 + \epsilon.$$

Estimarea unui astfel de model ia forma:

$$\widehat{Y} = \widehat{f}(X) = \widehat{\beta_1} X + \widehat{\beta_0}$$

Importanța unui predictor folosind p-values

- Vom vedea în output-ul unui model python de regresie detalii legate de p-values.
- Ele se referă la legătura dintre (coeficientul) predictor(ului) și variabila țintă
- În cele ce urmează vom detalia:

  - Ce înseamnă o valoare mică pentru acest p?

Importanța unui predictor folosind p-values

- Ce înseamnă o valoare mică pentru acest p?

1.Enunțul ipotezei nule:

H_0 : Nu există nici o relație între predictor și variabila țintă

2. Enunțul ipotezei alternative:

H_a : Există relație între predictor și variabila țintă

Importanța unui predictor folosind p-values

1. Enunțul ipotezei nule:

H_0 : Nu există nici o relație între predictor și variabila țintă

– Reformulat – coeficientul beta din modelul liniar este un 0 statistic, adică valorile nenule sunt obținute absolut aleator

2. Enunțul ipotezei alternative:

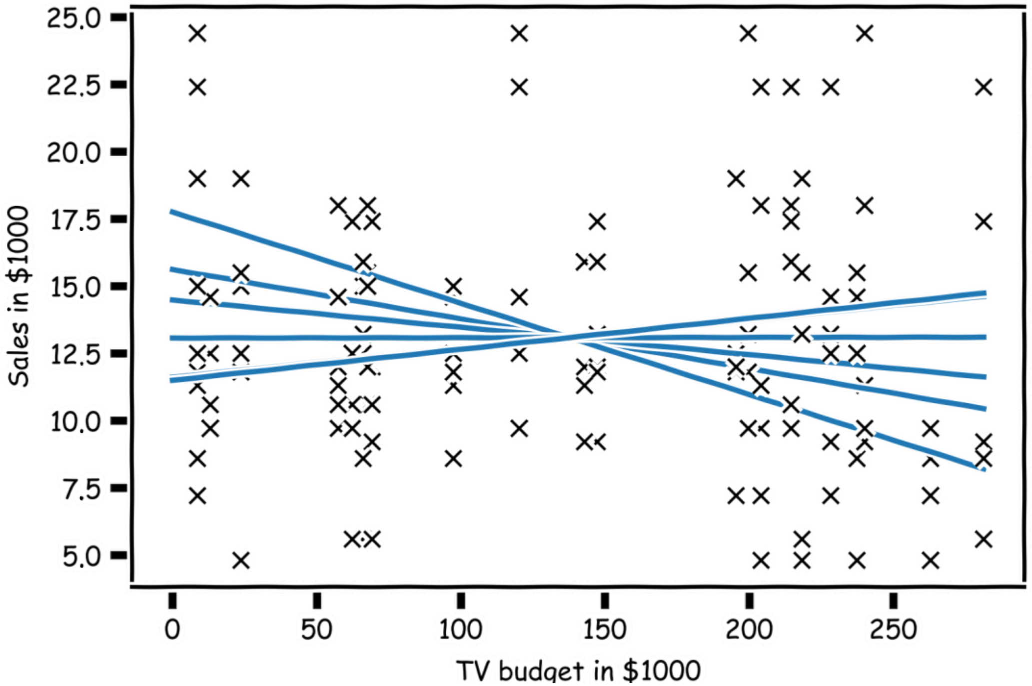
H_a : Există relație între predictor și variabila țintă

– Reformulat – coeficientul beta din modelul liniar nu este un 0 statistic

TV	sales
[illegible]	22.1
[illegible]	10.4
[illegible]	9.3
[illegible]	18.5
[illegible]	12.9
[illegible]	7.2
[illegible]	11.8
[illegible]	13.2
[illegible]	4.8
[illegible]	10.6
[illegible]	8.6
[illegible]	17.4
[illegible]	9.2
[illegible]	9.7
[illegible]	19.0
[illegible]	22.4
[illegible]	12.5
[illegible]	24.4

Eșantionarea aleatorie a datelor

Amestecăm valorile predictorului



Justificare intuitivă a faptului că ipoteza nulă înseamnă $\hat{\beta}_1 = 0$

Importanța unui predictor folosind p-values

3. Alegerea testului statistic pentru a aduna dovezi
   - Trebuie sa arătam că distribuția de valori estimate pentru coeficientul predictorului este suficient departe de o distribuție centrată in 0
   - Se folosește următorul test statistic, unde numărătorul este media estimărilor coeficientului, iar numitorul deviația standard

$$t = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

4. Pentru a folosi acest test statist trebuie eșantionate date – boostrapare => estimări coefient => medie, std

Erori Standard

Ipoteze uzuale asupra erorilor în regresia liniară:

- erorile $\epsilon_i = y_i - \hat{y}_i$ și $\epsilon_j = y_j - \hat{y}_j$ sunt necorelate, pentru $i \neq j$,
- fiecare ϵ_i are medie 0 și varianță σ_ϵ^2,

Importanța unui predictor folosind p-values

4. Dacă nu există legătură între predictor și variabila țintă, atunci (***știm de la statistică, în contextul ipotezelor asupra erorilor etc***.) valoarea t definită anterior se supune unei *t*-statistici cu n-2 grade de libertate (n dimensiuea eșantionului, 2 parametrii regresiei – medie+coeficient predictor)

Importanța unui predictor folosind p-values

Un eșantion bootstrap ne dă un *t*

Probabilitatea să obținem/observăm cu altă ocazie o valoare mai mare decât I*t*I, condiționat de ipoteza nulă, se poate calcula acum și se numește p-value.

Ce înseamnă o valoare mică pentru acest p? (mică= $p<0.05$):

- șanse mici ca valoarea mai mare decât I*t*I să fie aleatorie
- deci șanse mici ca relația dintre predictor și variabila țintă să fie aleatorie
- deci ipoteza nulă este rejectată,
- deci legătura predictorului cu variabila țintă nu este aleatorie

CONCLUZIE

O valoare mică - **$p<0.05$** pentru p înseamnă:

- **legătura predictorului cu variabila țintă nu este aleatorie**
- **Intermezzo ipynb**

REGRESIE LINIARĂ MULTIPLĂ

Regresie Liniară Multiplă

Dacă trebuie să ghiciți înălțimea cuiva, ați prefera să vi se spună

- Numai greutatea lor
- Greutatea și genul lor
- Greutatea, sexul și veniturile lor
- Greutatea, sexul, venitul și numărul

Desigur, ați dori întotdeauna cât mai multe date despre o persoană. Chiar dacă înălțimea și numărul favoritului nu pot fi strâns legate, în cel mai rău caz, puteți ignora informațiile despre numărul favorit. Vrem ca modelele noastre să poată prelua o mulțime de date pe măsură ce își fac predicțiile

.

Variabile răspuns/target vs. Variabile predictor

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

predictori

features

covariabile

$Y = y_1, \dots, y_n$

Observație

Variabila **răspuns**

Variabilă dependentă

n **observații**

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

p predictori

Regresie Liniară Multiplă

Din nou, a se potrivi acestui model înseamnă a calcula $\hat{\beta}_0, \dots, \hat{\beta}_J$ sau pentru a minimiza o funcție de pierdere; vom alege din nou **MSE** ca funcție de loss.

Având un set de observații,

$$\{(x_{1,1}, \dots, x_{1,J}, y_1), \dots (x_{n,1}, \dots, x_{n,J}, y_n)\},$$

datele și modelul pot fi exprimate în notație vectorială,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_y \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,J} \\ 1 & x_{2,1} & \dots & x_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,J} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_J \end{pmatrix},$$

Model Multiliniar, exemplu

Pentru datele noastre

$$\text{Sales} = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times Newspaper + \epsilon$$

Cu notații din algebra liniară

$$\boldsymbol{Y} = \begin{pmatrix} \boxed{Sales_1} \\ \vdots \\ Sales_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boxed{1 \quad TV_1 \quad Radio_1 \quad News_1} \\ \vdots \quad \vdots \quad \vdots \\ 1 \quad TV_n. \quad Radio_n \quad News_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \boxed{\begin{matrix} \beta_0 \\ \vdots \\ \beta_3 \end{matrix}} \end{pmatrix}$$

$$\boxed{Sales_1} = \boxed{1 \quad TV_1 \quad Radio_1 \quad News_1} \times \boxed{\begin{matrix} \beta_0 \\ \vdots \\ \beta_3 \end{matrix}}$$

Regresie Liniară Multiplă

Modelul ia o formă algebrică simplă:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Astfel, MSE poate fi exprimat în notație vectorială ca

$$\mathrm{MSE}(\beta) = \frac{1}{\mathrm{n}}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

Minimizarea MSE:

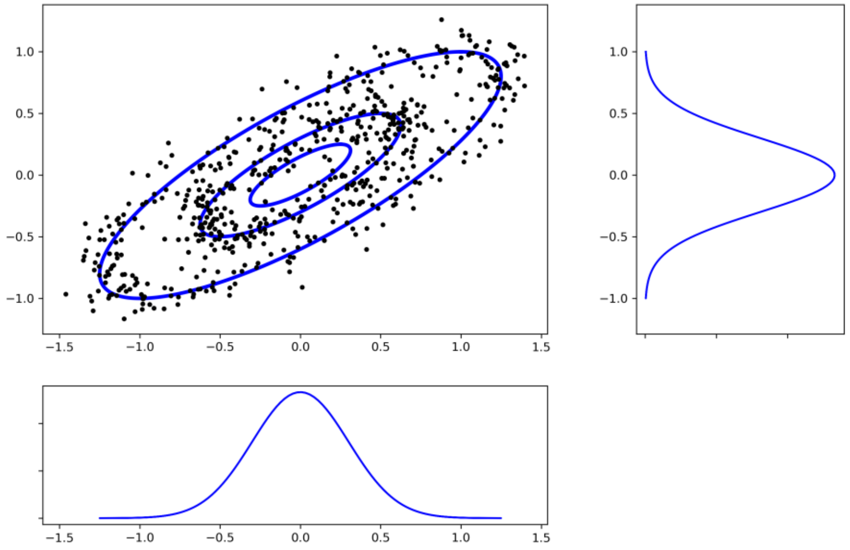
$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, \mathrm{MSE}(\boldsymbol{\beta}).$$

Standard pentru Regresie Liniară Multiplă

Ca și în cazul regresiei liniare simple, erorile standard pot fi calculate fie folosind modelarea statistică

$$SE(\beta)^2 = \sigma^2 (X^T X)^{-1}$$

fie cu **bootstrapare**



Coliniaritate

Coliniaritatea se referă la cazul în care doi sau mai mulți predictori sunt corelați.

Vom vizita din nou colinearitatea atunci când vom aborda problema de **overfitting** (supraadaptarea), dar deocamdată dorim să examinăm modul în care coliniaritatea ne afectează încrederea asupra coeficienților și, în consecință, asupra importanței acestor coeficienți.

Coliniaritate

Trei modele individuale

TV

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
6.679	0.478	13.957	2.804e-31	5.735	7.622
0.048	0.0027	17.303	1.802e-41	0.042	0.053

RADIO

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
9.567	0.553	17.279	2.133e-41	8.475	10.659
0.195	0.020	9.429	1.134e-17	0.154	0.236

NEWS

Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
11.55	0.576	20.036	1.628e-49	10.414	12.688
0.074	0.014	5.134	6.734e-07	0.0456	0.102

Un model multiliniar

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
β_0	2.602	0.332	7.820	3.176e-13	1.945	3.258
β_{TV}	0.046	0.0015	29.887	6.314e-75	0.043	0.049
β_{RADIO}	0.175	0.0094	18.576	4.297e-45	0.156	0.194
β_{NEWS}	0.013	0.028	2.338	0.0203	0.008	0.035

Găsirea unor predictori semnificativi: testarea ipotezei

Pentru verificarea semnificației coeficienților de regresie liniară:

1.Ipoteza nulă H_0 *și alternativa*:

$$H_0 : \beta_0 = \beta_1 = \ldots = \beta_J = 0 \qquad \textbf{(Null)}$$

$$H_1 : \beta_j \neq 0, \ \text{pentru cel puțin un } j \qquad \textbf{(Alternative)}$$

2. alegem *F*-stat pentru a evalua ipoteza nulă,

$$F = \frac{\text{varianță explicată}}{\text{varianță neexplicată}}$$

Finding Significant Predictors: Hypothesis Testing

3. Putem calcula *F*-stat pentru regresia liniară astfel:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/J}{\mathrm{RSS}/(n - J - 1)}, \quad \mathrm{TSS} = \sum_i (y_i - \overline{y})^2, \mathrm{RSS} = \sum_i (y_i - \widehat{y}_i)^2$$

4. Dacă $F = 1$ considerăm asta ca dovadă în sprijinul lui H_0; dacă $F > 1$, considerăm asta ca o dovadă împotriva lui H_0.

Predictori Calitativi/Categoriali

Până acum, am presupus că toate variabilele sunt **cantitative**. Dar, în practică, adesea unii predictori sunt calitativi/categoriali.

Exemplu: setul de date privind creditul conține informații despre sold, vârstă, carduri, educație, venit, limită și rating pentru un număr de clienți potențiali.

Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
14.890	3606	283	2	34	11	Male	No	Yes	Caucasian	333
106.02	6645	483	3	82	15	Female	Yes	Yes	Asian	903
104.59	7075	514	4	71	11	Male	No	No	Asian	580
148.92	9504	681	3	36	11	Female	No	No	Asian	964
55.882	4897	357	2	68	16	Male	No	Yes	Caucasian	331

Predictori Calitativi/Categoriali

Dacă predictorul ia doar două valori, atunci creăm un indicator sau o variabilă fictivă care preia două valori numerice posibile.

De exemplu, pentru sex, creăm o nouă variabilă:

$$x_i = \begin{cases} 1 & \text{dacă a } i\text{-a persoană e femeie} \\ 0 & \text{dacă a } i\text{-a persoană e bărbat} \end{cases}$$

Apoi folosim această variabilă ca predictor în ecuația de regresie.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{dacă a } i\text{-a persoană e femeie} \\ \beta_0 + \epsilon_i & \text{dacă a } i\text{-a persoană e bărbat} \end{cases}$$

Predictori Calitativi/Categoriali

Întrebare: Care este interpretarea coeficienților β_0 și β_1 în acest caz?

Predictori Calitativi/Categoriali

Întrebare: Care este interpretarea coeficienților β_0 și β_1 în acest caz?

- β_0 este soldul mediu al cardului de credit în rândul bărbaților,
- $\beta_0 + \beta_1$ este soldul mediu al cardului de credit în rândul femeilor,
- Iar β_1 diferența medie în soldul cardului de credit între femei și bărbați.

Exemplu: Calculați β_0 și β_1 pentru setul de date Credit data:

You should find $\beta_0 \sim \$509, \beta_1 \sim \$19$

Mai mult de două niveluri: One hot encoding

Adesea, predictorul calitativ ia mai mult de două valori (de exemplu, etnia în datele de credit).

În această situație, o singură variabilă fictivă nu poate reprezenta toate valorile posibile.

Creăm o variabilă dummy (falsă) suplimentară ca:

$$x_{i,1} = \begin{cases} 1 & \text{dacă a } i\text{-a persoană e asiatică} \\ 0 & \text{dacă a } i\text{-a persoană nu e asiatică} \end{cases}$$

$$x_{i,2} = \begin{cases} 1 & \text{dacă a } i\text{-a persoană e caucaziană („alb”)} \\ 0 & \text{dacă a } i\text{-a persoană nu e caucaziană} \end{cases}$$

Mai mult de două niveluri: One hot encoding

Apoi folosim aceste variabile ca predictori, ecuația de regresie devine:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{dacă a } i\text{-a persoană e asiatică} \\ \beta_0 + \beta_2 + \epsilon_i & \text{dacă a } i\text{-a persoană e caucaziană} \\ \beta_0 + \epsilon_i & \text{dacă a } i\text{-a persoană e afro-americană} \end{cases}$$

Întrebare: Care este interpretarea coeficienților $\beta_0, \beta_1, \beta_2$?

Dincolo de liniaritate

În datele publicitare, am presupus că efectul asupra vânzărilor creșterii unui mediu publicitar este independent de suma cheltuită pe celelalte medii.

Dacă presupunem un model liniar, atunci efectul mediu asupra vânzărilor unei creșteri cu o unitate a televizorului este întotdeauna β_1, indiferent de suma cheltuită la radio.

Efectul de **sinergie** sau efectul de **interacțiune** indică faptul că o creștere a bugetului radio afectează eficiența cheltuielilor TV pentru vânzări.

Dincolo de liniaritate

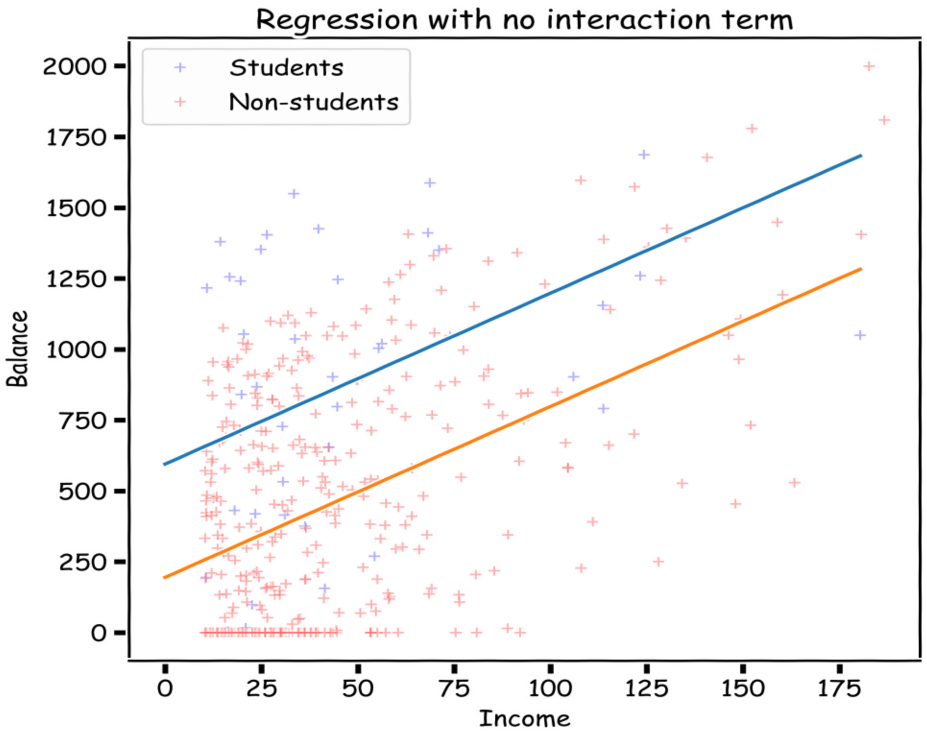
Modificăm modelul

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Adăugând termenul de interacțiune

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boxed{\beta_3 X_1 X_2} + \epsilon$$

- **Ce înseamnă asta?**



$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1) \times Income. \end{cases}$$

$$x_{Student} = \begin{cases} 0 & Balance = \beta_0 + \beta_1 \times Income. \\ 1 & Balance = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times Income \end{cases}$$

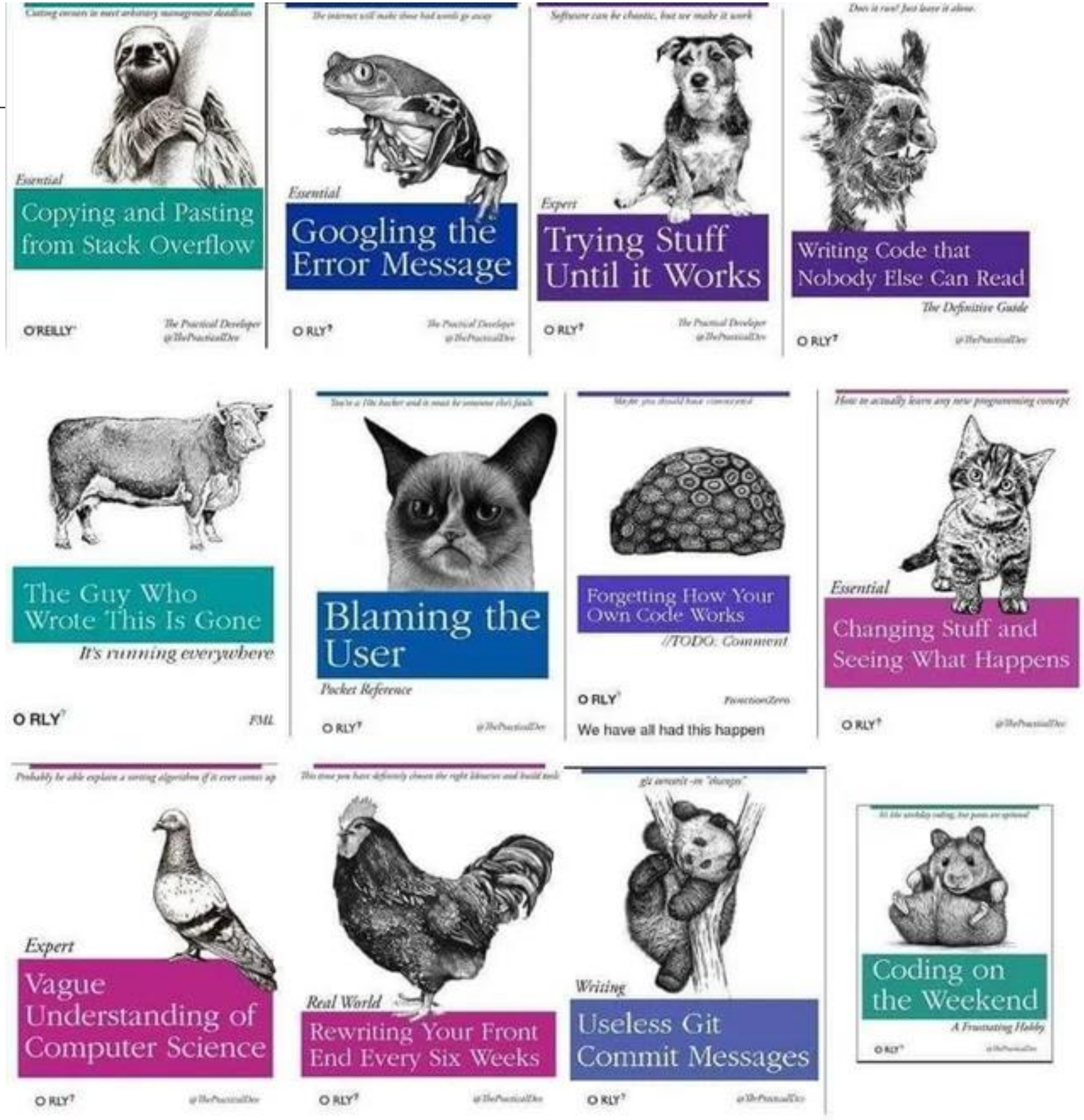
Predictori predictori predictori

Avem o mulțime de predictori!

Este asta o problemă?

- **Da**: Cost Computational
- **Da**: Overfitting/Supraadaptare

”Wait there is more …”


Essential
Copying and Pasting from Stack Overflow
O'REILLY
Essential
Googling the Error Message
O RLY?
Expert
Trying Stuff Until it Works
O RLY?
Writing Code that Nobody Else Can Read
The Definitive Guide
O RLY?
The Guy Who Wrote This Is Gone
It's running everywhere
O RLY?
FML
Blaming the User
Pocket Reference
O RLY?
Forgetting How Your Own Code Works
//TODO: Comment
O RLY?
We have all had this happen
Essential
Changing Stuff and Seeing What Happens
O RLY?
Expert
Vague Understanding of Computer Science
O RLY?
Real World
Rewriting Your Front End Every Six Weeks
O RLY?
Writing
Useless Git Commit Messages
O RLY?
Coding on the Weekend
A Frustrating Hobby

Reziduuri

Am început cu

$$y = f(x) + \epsilon$$

Am **presupus** o formă exactă pentru $f(x)$,

$$f(x) = \beta_0 + \beta_1 x,$$

apoi am estimat coeficienții $\hat{\beta}$.

Dacă nu este corectă presupunerea? Atunci:

$$f(x) = \beta_0 + \beta_1 x + \phi(x),$$

Dar noi am modelat

$$\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Atunci reziduul este:

$$r = (y - \widehat{y)} = \hat{f}(x) = \epsilon + \phi(x)$$

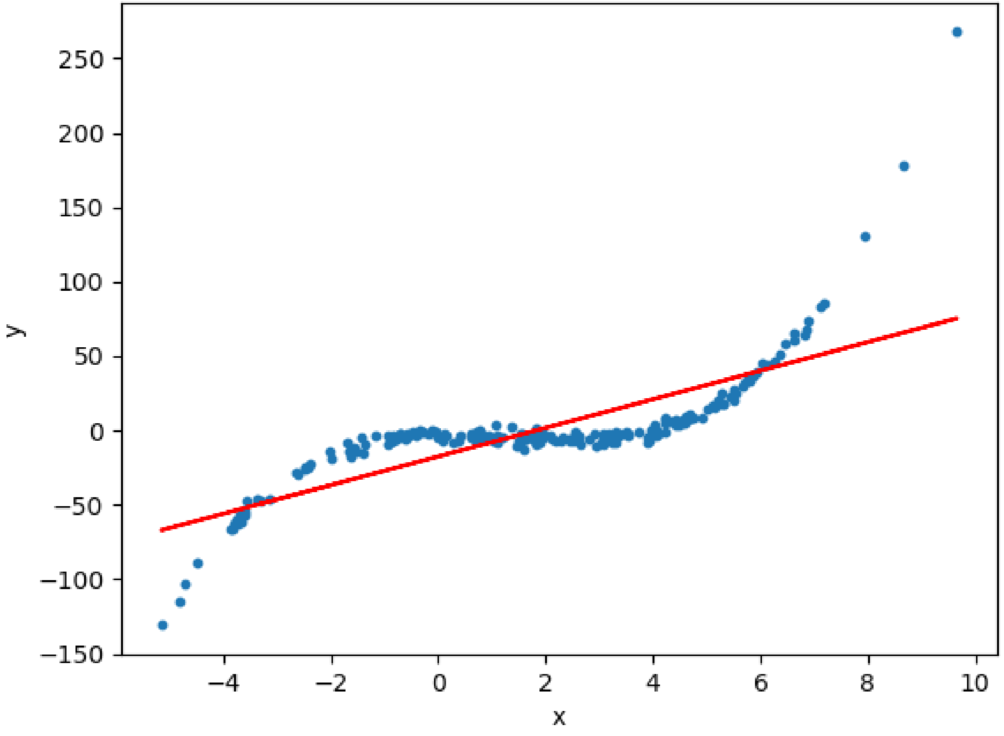
Analiza Residuală

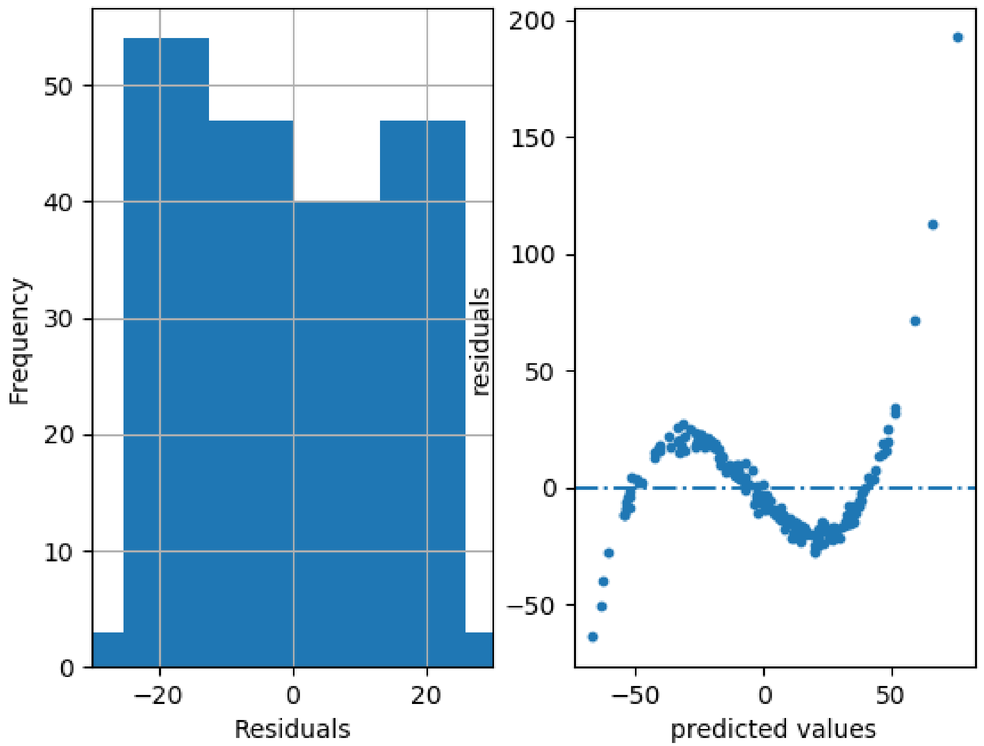
Când am estimat varianța lui ϵ, am presupus că reziuurile $r_i = y_i - \hat{y}_i$ au fost necorelate și distribuite în mod normal cu media 0 și varianță fixă.

Aceste ipoteze trebuie verificate folosind datele. În analiza reziduală, de obicei creăm două tipuri de grafice:

1. Unul cu r_i funcție de x_i sau $\hat{y}_i$. Acest lucru ne permite să comparăm distribuția zgomotului la diferite valori ale x_i.
2. o histogramă a lui r_i. Acest lucru ne permite să explorăm distribuția zgomotului independent de x_i sau $\hat{y}_i$.

Analiză Reziduală





REGRESIE POLINOMIALĂ

Regresie Polinomială

Cel mai simplu model neliniar pe care îl putem considera, pentru o variabilă răspuns *Y* și un predictor *X*, este un model polinomial de grad *M*,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_M x^M + \epsilon.$$

La fel ca în cazul regresiei liniare cu termeni încrucișați, regresia polinomială este un caz special de regresie liniară - tratăm fiecare x^m ca predictor separat. Astfe putem scrie

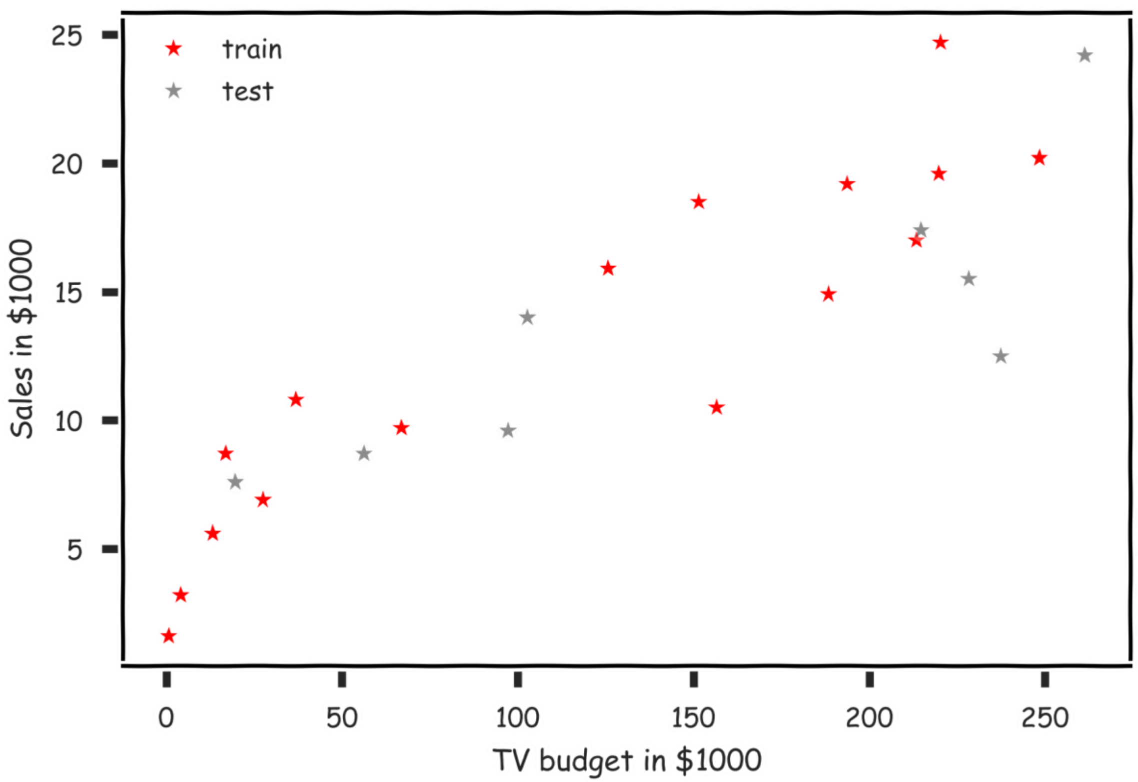
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \ldots & x_1^M \\ 1 & x_2^1 & \ldots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^M \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Regresie Polinomială

Din nou, minimizând MSE cu metode de analiză matematică obținem,

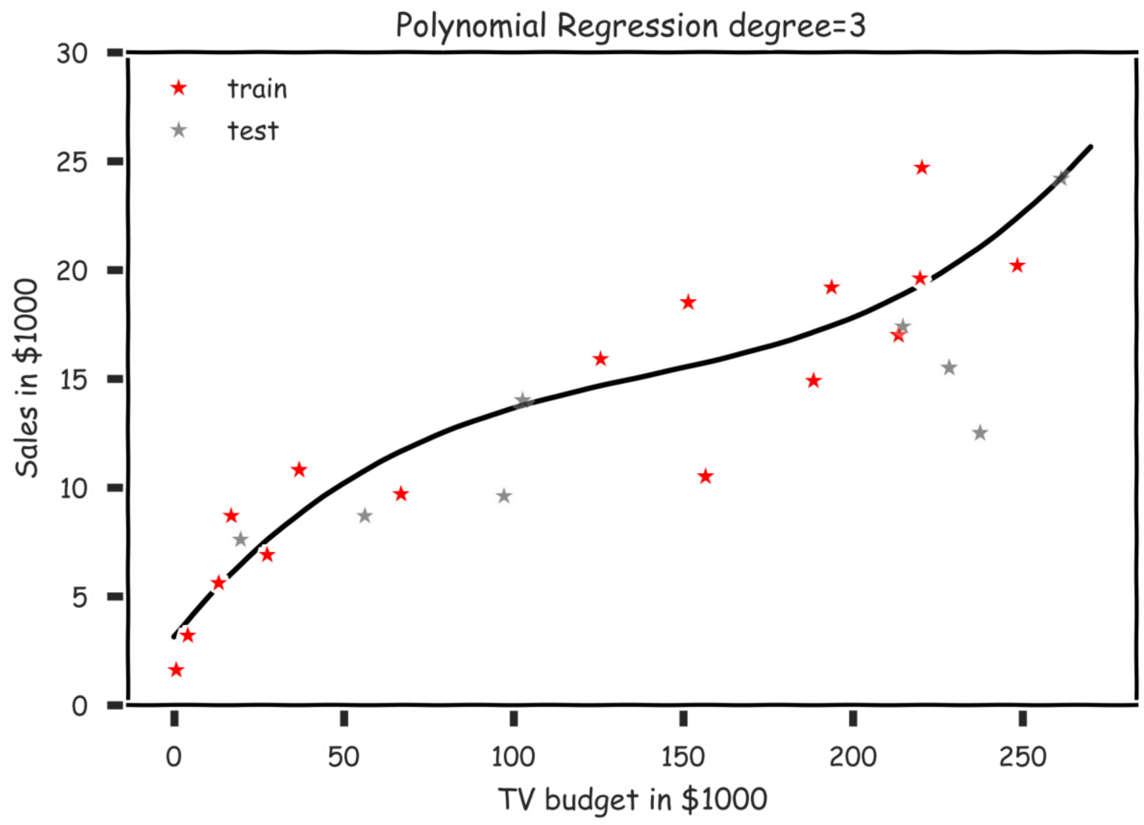
$$\widehat{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \operatorname{MSE}(\beta) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Regresie Polinomială (cont)

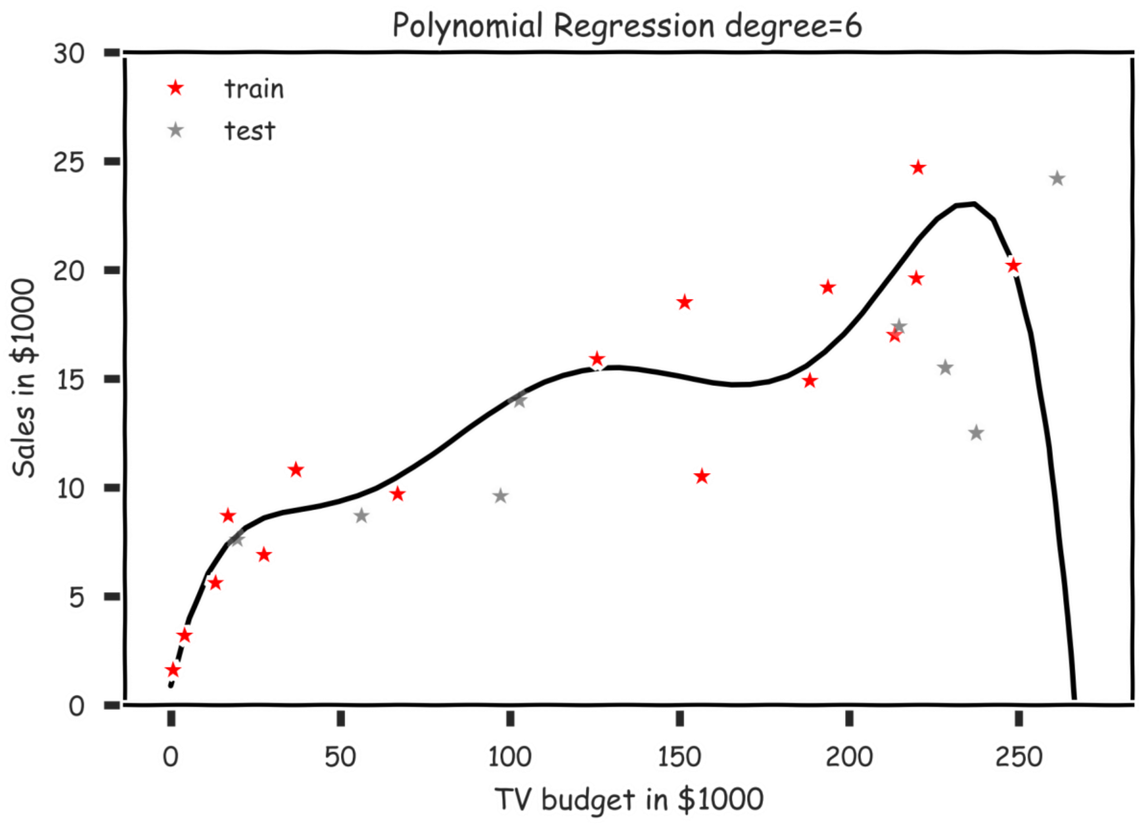


Regresie Polinomială(cont)

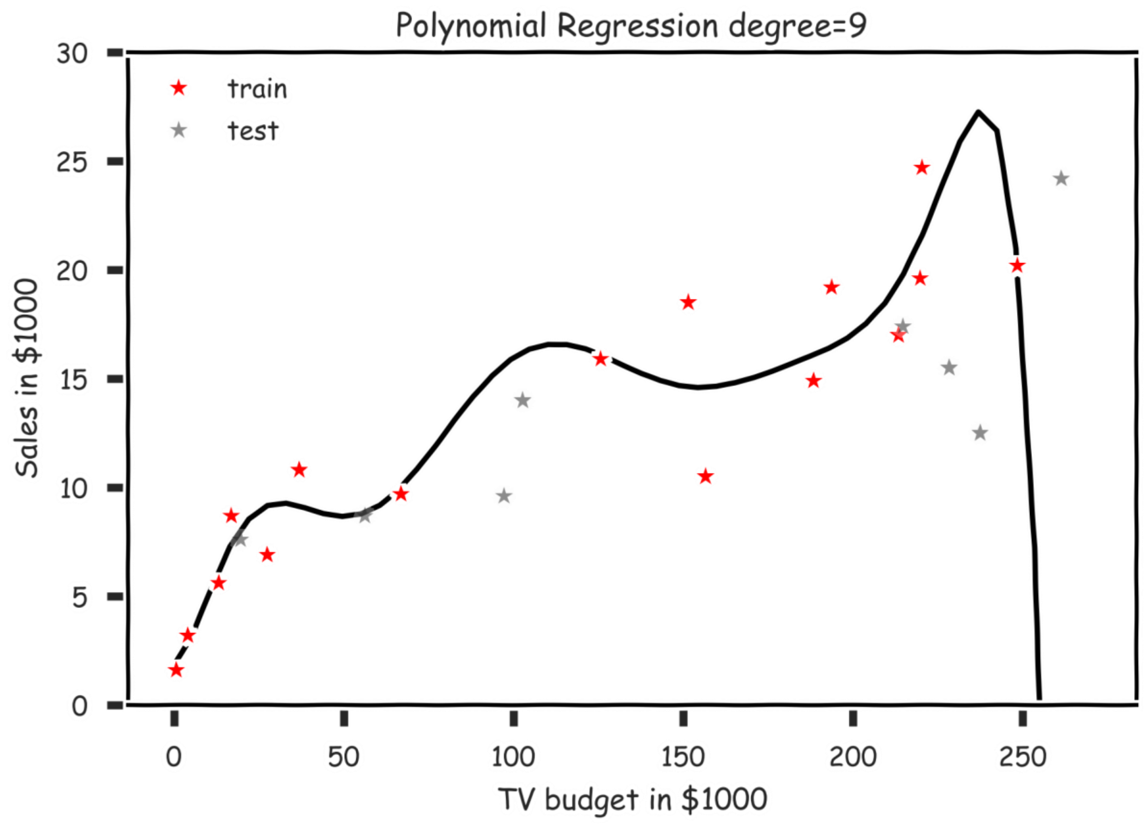
Regresie Polinomială(cont)



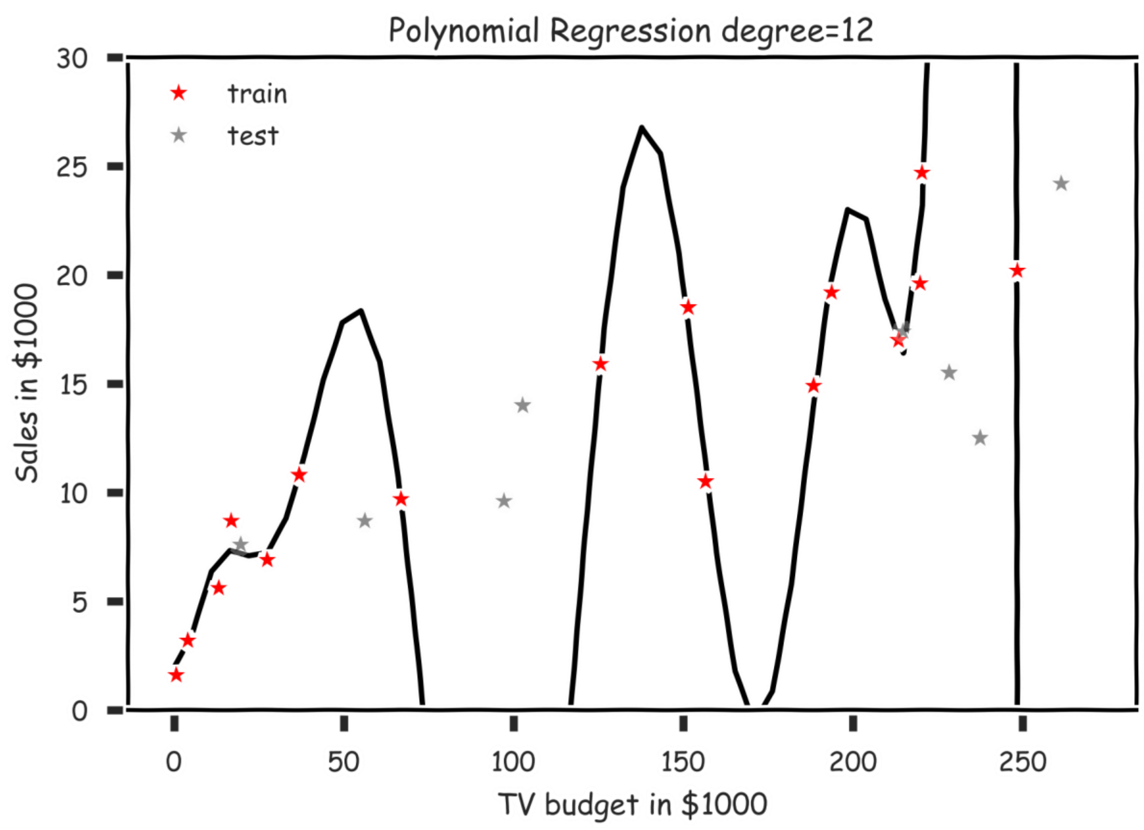
Polynomial Regression (cont)



Regresie Polinomială (cont)



Regresie Polinomială (cont)



Overfitting/Supraadaptare

În statistici, **supraadaptarea** este „producerea unei analize care corespunde prea strâns sau exact unui anumit set de date și, prin urmare, poate să nu se potrivească cu date suplimentare sau să nu prezică fiabil observațiile viitoare”

Mai multe în cele ce urmează